Development of topological tools for the analysis of biological data

THÈSE N^{O} 8835 (2018)

PRÉSENTÉE LE 29 OCTOBRE 2018

À LA FACULTÉ DES SCIENCES DE LA VIE

UNITÉ DU PROF. BRISKEN

PROGRAMME DOCTORAL EN APPROCHES MOLÉCULAIRES DU VIVANT

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Rachel JEITZINER

acceptée sur proposition du jury:

Prof. E. Oricchio, présidente du jury
Prof. C. Brisken, Prof. K. Hess Bellwald, directrices de thèse
Prof. J. Brodzki, rapporteur
Prof. O. C. Lingjaerde, rapporteur
Dr J. Rougemont, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2018

I am among those who think that science has great beauty.
— Marie Curie

To my parents and my family without whom this thesis would not have seen the light of day.

Acknowledgements

I would to start by expressing my gratitude towards Prof. Cathrin Brisken for having supervised my work, for her availability, her scientific input that I grew to appreciate, and her patience when writing an article, as this is my greatest weakness and is where I needed most guidance. Introducing one-to-one meetings enabled us to not only have scientific discussions, but to build up a personal relationship.
Prof. Kathryn Hess, my co-director managed from the first day to pass her enthusiasm on to me. Her rigor and mathematical experience are a great source of motivation. I would like to thank her for her availability, weekly meetings, fairness and encouragement.
I thank both Cathrin and Kathryn for giving me the opportunity to travel to numerous conferences around the world. Both of them showed me through their commitment that women have their place in science and are therefore a great source of inspiration.
Jacques Rougemont welcomed me into the bioinformatics community for which I am deeply thankful. He manages to be part of several orchestra next to his work, which was a stimulation to continue playing my instrument.
I would like to warmly thank all the members of the jury for kindly accepting to evaluate my work.
I thank my colleagues in the mathematics building, Jérôme, Martina Rovelli, Dimitri, Kay, Marc, Varvara, Martina Scolamiero, Gard, Senja, Sophie, Magda, Kate, Lyne it was great fun to do maths, organise a conference, find banana thefts, travel and share rooms at conferences, but most of all to groove together.
Then, I thank my colleagues in the bioinformatics building, Fabrice, Sara, Marion, Alexandre, Flavia, Delphine for lunch breaks around bioinformatics and way more. A special thanks to Evarist and Julien, who answered all my bioinformatics related questions and to Elena for being the best office mate.
Endless thanks to my dear colleagues in the Brisken Lab for their help, especially Valentina who actively participated in my lab meetings and invested time to read parts of my thesis, but more importantly for gossiping, and Dalya, Mélanie, Valérian, Duje, Stéphanie who became true friends, Csaba for the blood measurements, Georgios for the experiments with RANKL and progesterone, and sharing with me insights into biology and genes, Marie for the work with mice on my attempt to validate some results. Kannu and Stéphanie for the help and teaching of mice dissection. Laura for her patience in showing me how to do qPCRs, secretaries of the Brisken lab who not only have an ear for administrative issues, and to all the all current and past lab members.
I thank Steve Oudot and Mathieu Carrière for their collaboration and hard work on the theoretical part of the method.
Furthermore, I thank members of the Centre Integrative Genomic at UNIL, as well as Bastien

Acknowledgements

Mangeat and his team at EPFL for doing the sequencing of many datasets I analysed.
Also thank you to the CHUV for their collaboration to provide mammoplasty tissues through the years.
I thank my mentor Pierre Gönczy, with his strong voice he had the words to put me back into track at numerous moments.
Friends from EPFL and from conferences abroad thanks for the coffee breaks and talks.
Un grand merci à mes amies de toujours les tricoteuses, Emilie et Camille, sans vous mes semaines auraient été bien plus fades. Je vous remercie pour votre soutien.
Merci à ma future belle famille pour leur accueil chaleureux.
Mein grösster Dank geht an meine Eltern und meine Familie, ohne euch wäre dieses Doktorat nie zu Stande gekommen. Ihr habt mich nicht nur finanziell unterstützt, sondern besonders auch moralisch, da ihr immer ein offenes Ohr hattet und mich mit Lachen, Jassen, Wandern und vieles mehr auf andere Gedanken gebracht habt.
Adrien Marcone, je te remercie pour chaque jour passé avec toi, pour ton amour et soutien à travers les années. Je me réjouis avec impatience de la suite, et du petit miracle qu'on va bientôt rencontrer.

Jongny, 5 Juin 2018 R. J.

Preface

I take the opportunity of this preface to share my personal view on how to handle transdisciplinary projects; this thesis is an exceptional case as it regroups three different areas of study : firstly, mathematics, more specifically the field of topology, which is my field of expertise. Secondly, bioinformatics, which, during my four years of PhD, came to be the second field in which I am most comfortable. Finally, biology, which continues every day to surprise me in how rich the field is. My thesis comprises all of these three fields and I believe that it is important that it can be read by interested readers from any of these three areas of study. For this reason, introductions are needed in each of these fields, which can be skipped without loss of information, if the reader is already familiar with the basics. Some notions in each field, such as "topology" or "connected components", are assumed to be known. For this reason, I recommend the reader to read [1] for a basic course in topology and [2] for basic definitions on cells. The way I approached this thesis was by realising early on where my weaknesses and my strengths were. I discovered that the years of PhD needed to be divided according to the different fields and that I had to start by learning more about the mathematical aspects of topological data analysis, before developing bioinformatics skills to write my own method. Lastly, learn techniques of biology in order to validate my findings. Throughout this whole process I had to remember that the field in which one is least comfortable with, should never be neglected. Therefore, immersions into seminars and meetings on a regular basis were crucial. At first, the meetings were hard to grasp and there were probably a lot of question arising that could be asked afterwards. The more determination one shows, the easier it gets to understand them. Throughout this whole thesis, I kept regular meetings with my advisor in molecular science as well as with my co-advisor in topology. It was important for me to benefit from their expertise, but it was equally important for them to be informed on how the process of learning and working was moving forwards, even though it is not in their field of expertise. Moreover, the only way to see if you truly understood what you have been working on, is to explain it to somebody from outside of your area of studies. In molecular biology, one is often confronted with histochemistry, a technique which consists in viewing a piece of tissue at a given point. This method is widely used in my lab and I only fully started to grasp what can be seen on these sections, when I took a course about it. It is therefore important to follow small introductory courses to techniques used in the lab. The last important point for me is to not isolate yourself and meet with experts, regularly. If you physically sit and eat with bioinformaticians on a daily basis the conversations arising are different than when you eat with mathematicians. Last advice: Get immersed and start learning, but do not forget the fun!

Lausanne, 22 January 2018 R. J.

Abstract

There is a growing need for unbiased clustering algorithms, ideally automated to analyze complex data sets. Topological data analysis (TDA) has been used to approach this problem. This recent field of mathematics discerns characteristic features of a space without relying on probabilistic approaches. It provides robust qualitative and quantitative assessments of the structure of data. Mapper, an algorithm of TDA, showed increased power over standard methods for complex data and overcame problems of noise. However, it relies on the selection of several parameters and is not well suited for small datasets. To overcome these problems, we have developed a topology-based clustering algorithm called Two-Tier Mapper (TTMap) to detect subgroups in global gene expression datasets and to identify their distinguishing features in a two groups comparison. First, TTMap discerns and adjusts for highly variable features in the control group and identifies outliers. Second, in order to obtain an individual appreciation of the differences with respect to the control group, a profile of deviation is computed for each test sample. Test samples are clustered according to two tiers creating a global and local network using a new topological algorithm based on Mapper, where all the parameters are carefully chosen or data-driven, avoiding any user induced bias. These choices render the algorithm theoretically stable. In particular when sample sizes are small, TTMap outperforms existing clustering methods in finding relevant subgroups, in stability on synthetic and biological datasets and in revealing more gene expression changes. Datasets from different sources can readily be combined into one analysis. Thus, TTMap can extract information from highly variable biological samples, and since an individual profile of deviation is established, it has potential for personalized medicine. The algorithm was developed as an open source R package deposited at the Bioconductor.
Furthermore, two additional applications of topology were developed in order to find differences in gene expression through the menstrual cycle and cyclical patterns in gene expression related to hormone response.

Key words: Mapper, two-tier cover, topology, topological data analysis, extended persistent homology, clustering, gene expression, parameter-free, Bioconductor R package, estrous and menstrual cycle, progesterone, RANKL

Zusammenfassung

Es besteht ein wachsender Bedarf an unvoreingenommenen Clustering-Algorithmen, die idealerweise automatisiert wurden, womit man komplexe Datensätze analysieren kann. Lösungen für dieses Problem wurden in den letzten zehn Jahren mit topologischer Datenanalyse (TDA) vorgeschlagen. Dieses neue Gebiet der Mathematik verschafft relevante Merkmale eines Raumes, ohne sich auf probabilistische Ansätze zu verlassen. Es trägt zu robusten qualitativen und quantitativen Bewertungen über die Struktur der Daten bei. Diese Domäne besteht aus Algorithmen, die sichfür die Analyse von komplexen Daten eignen und die Lärmprobleme überwinden können. Mapper, ein Algorithmus von TDA, zeigte gegenüber Standardmethoden eine gesteigerte Aussagekraft. Er ist jedoch abhängig von der Auswahl von mehreren Parametern und is nicht geignet für kleine Datensätze. Um diese Probleme zu überwinden, haben wir einen Topologie-basierten Clustering-Algorithmus namens Two-Tier Mapper (TTMap) zum Erkennen von Untergruppen in globalen Genexpressionsdatensätzen entwickelt und zur Identifizierung ihrer Unterscheidungsmerkmale, in einem Vergleich zweier Gruppen, der Kontroll und der Testgruppe. Zuerst erkennt und korrigiert TTMap für stark variable Eigenschaften in der Kontrollgruppe und identifiziert Sonderfälle. Zweitens, um eine individuelle Wertschätzung von Unterschieden in Bezug auf die Kontrollgruppe zu erhalten, wird ein Abweichungsprofil für jede Testprobe berechnet. Testproben werden nach zwei Ebenen mit einem neuen topologischen Algorithmus basierend auf Mapper geclustert, wo alle Parameter sorgfältig ausgewählt wurden oder datengetrieben sind, um eine Voreingenommenheit vom Benutzer zu vermeiden. Diese Auswahl macht den Algorithmus theoretisch stabil. Durch die zwei Ebenen entsteht ein globales und lokales Netzwerk. Insbesondere bei kleinen Datensätzen, übertrifft TTMap aktuelle Clustering-Methoden in der Suche nach relevanten Untergruppen und in der Stabilität von synthetischen und biologischen Datensätzen und enthüllt bisher unentdeckte Veränderungen der Genexpression. Datensätze von mehreren Quellen können leicht zu einer Analyse kombiniert werden. TTMap kann Informationen aus sehr variablen Daten extrahieren un aufgrund der Tatsache, dass ein individuelles Abweichungsprofil vorliegt, hat TTMap das Potential für personalisierte Medizin nützlich zu sein. Der Algorithmus wurde als Open-Source-R-Paket in Bioconductor entwickelt.
Darüber hinaus wurden zwei weitere Ideen der Anwendung der Topologie entwickelt, um Unterschiede in Genexpression in menschliches Brustgewebe durch den Menstruationszyklus zu finden und zyklische Muster in der Genexpression zu finden, die in Zusammenhang mit einer Hormonreaktion stehen.

Stichwörter: Mapper, zweistufige Bedeckung, Topologie, topologische Analysis von Daten, erweiterte beständige Homologie, clustering, Genexpression, ohne Parameter, Bioconductor R Paket, Menstruationszyklus und Östruszyklus, Progesteron, RANKL.

Résumé

Afin d'analyser le nombre croissant d'ensembles complexes de données, il est devenu nécessaire de recourir à des algorithmes de regroupement automatisés qui ne soient pas biaisés. Au cours de la dernière décennie, l'analyse topologique de données (TDA) s'est imposée comme l'un d'eux. Ce nouveau domaine des mathématiques permet une analyse non probabiliste, qualitative et quantitative de la structure des données au moyen d'algorithmes adaptés à l'analyse de données complexes et capables de résoudre des problèmes de bruit. Mapper, l'un des ces algorithmes, a montré une efficacité supérieure par rapport aux méthodes standards d'analyse de données. Cependant, il dépend de nombreux paramètres et n'est pas adapté au petits ensembles de données. Afin de parer ces inconvénients, nous avons développé un algorithme de regroupement basé sur la TDA, appelé Two-Tier Mapper (TTMap), permettant la détection de sous-groupes au sein d'ensembles de données d'expression globale de gènes ainsi que l'identification de leurs caractéristiques distinctives dans une comparaison entre deux groupes : le contrôle et le groupe test. Dans un premier temps, TTMap détecte et corrige les propriétés à forte variabilité dans le groupe de contrôle et remarque les valeurs aberrantes. Ensuite, un profil de variation est établi pour chaque échantillon du test, de manière à obtenir une appréciation individuelle des différences par rapport au groupe de contrôle. Les échantillons du test sont alors regroupés sur deux niveaux via un nouvel algorithme topologique basé sur Mapper, dans lequel tous les paramètres sont soigneusement sélectionnés ou déduits des données pour éviter les biais liés à l'utilisateur. Cette sélection rend l'algorithme théoriquement stable. Les deux niveaux créent un réseau de sous-groupes locaux et globaux. TTMap surpasse les méthodes de regroupement actuelles sur des ensembles de données synthétiques et biologiques, en particulier sur des petits ensembles de données en effectuant une meilleure classification des sous-groupes, par sa stabilité et par la découverte de variation d'expression de gènes, jusqu'ici non détectés. Les données provenant de sources multiples peuvent facilement être combinées en une seule analyse. TTMap est capable d'extraire des informations à partir de données extrêmement variables et comme TTMap produit un profil individuel de déviation par rapport au groupe de contrôle, il possède un potentiel d'utilisation dans la médecine personnalisée. L'algorithme a été développé en tant que package R open-source dans Bioconductor.
De plus, deux idées d'application de la topologie à la détection des différences dans l'expression de gènes du tissu mammaire humain au cours du cycle menstruel ainsi qu'à la détermination de modèles cycliques dans l'expression de ces gènes qui seraient associés à une réponse hormonale.

Mots clefs : Mapper, recouvrement à deux niveaux, topologie, analyse topologique des données, homologie persistente étendue, regroupement, expression de gènes, sans paramètres, Bioconductor R package, cycle menstruel ou oestral, progesterone, RANKL.

Contents

1 Background

1.1 Introduction to Topological Data Analysis

Topology is a field of mathematics that was born through concepts of geometry and set theory. It is sometimes considered as the unifying field of mathematics because it naturally appears in many branches of mathematics. An example illustrating this are spaces provided with a distance, which represent a particular subtype of topological spaces and appear naturally in geometry, or in metric analysis [1]. In topology, concepts of geometry have been redefined without using or introducing a notion of distance. This recent field of mathematics studies the shape of objects under appropriate deformations. Especially in terms of shape recognition, the goal is to identify a standing person and the same individual running as "continuously deformable one into the other" or being indistinguishable [3]. In this context, continuously deformable means that objects are allowed to be stretched and bent, but not torn apart nor glued. Moreover, objects do not need or depend on a coordinate system. This has the consequence that topology is insensitive to scale, or homothetic transformations, or rotations. Since spaces are studied under continuous deformations, the notion of distance becomes more coarse: the actual value of a distance between two points is not needed to be known, but an understanding of the global proximity is. Topology not only studies how to deform an object into another one and whether or not such deformations exist but also tries to find invariants, i.e., properties of an object that would remain unchanged through continuous deformations, by using algebra. These invariants are then used to distinguish spaces that are topologically the same in a more efficient way, i.e. calculable or computable, than trying to find an explicit deformation of one object into another one. As they are topological notions, these invariants are not depending on geometric coordinates of the space, and represent descriptions of a space that differentiates it from another. They can be of the following form: this object is composed of two separate "parts", called the **connected components**, possesses one hole in the structure and one 2-dimensional cavity. These summaries of the objects have been defined and proved to remain unchanged through these continuous deformations in the last centuries. In the 90s, these theoretical invariants started to become applied to data and helped distinguish different shapes. The invariants were called **shape descriptors** and started a new field of mathematics, called **Topological Data Analysis** (TDA) [4], [5], [6], [7]. Topological data analysis comprises two parts that sometimes overlap : complex network analysis methods consisting as well of dimension reduction algorithm, and **persistent homology** (PH) [8]: both of these aspects are discussed in this thesis (Fig. 1.1).

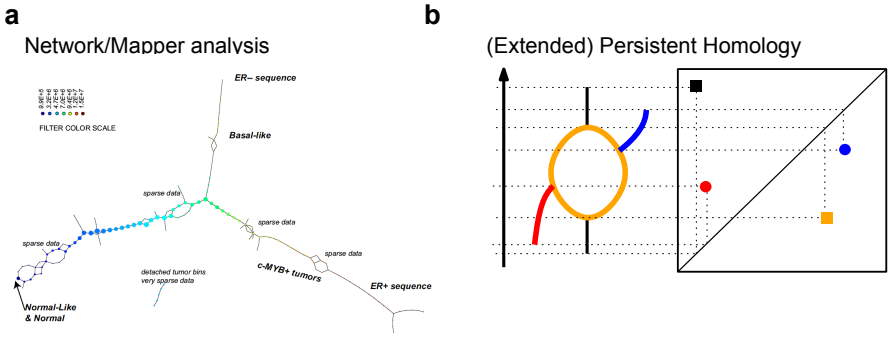


Figure 1.1: The main branches of topological data analysis. **(a)** Complex Network analysis such as the Mapper algorithm described in the following section 1.2 (picture adapted from [9]) and **(b)** Persistent as well as extended persistent homology summarising topological features (section 1.3).

1.2 The Mapper algorithm

1.2.1 Introduction to the Mapper algorithm

In 2011, the first breakthrough in the "network" analysis part of TDA applied an algorithm called **Mapper** [10] to breast cancer data revealing a hidden subgroup among breast cancer patients displaying common characteristics: no metastases, excellent survival and alteration in expression in a common group of genes undetectable by standard algorithms [9]. The Mapper algorithm was first described in a paper in 2007 [10], then explained a second time within a general theoretical TDA framework in [8] and finally made accessible to a large audience in 2013 [11], where a visual example illustrates the functioning of Mapper (example 1.2.38) and where real data examples proved the broad application of this algorithm. Two reviews, written for a large audience, summarise the theoretical aspects and usefulness of Mapper [12], and its application in the fields of genomics [13].

Mapper describes a space or a **point cloud**, i.e. a finite set X of points, usually these points are in $\mathbb{R}^n$ and n is large, as a complex network and was developed as a computational approximation of a **Reeb graph** [14] (see section 1.2.6) which is a recapitulation of a space in a certain "direction", determined by a function (Fig. 1.2). A filled donut for example equipped with the height function, is represented by a circle with two branches or a multigraph (Fig. 1.2). As observed in this example of the donut, the important topological structure of the initial space, i.e. the hole of the donut, is still kept and only the complexity of the space has been reduced. Mapper graphs and Reeb graphs are therefore simplifications of spaces that enable dimension reduction given further input/information on the space, a function, that would enable more precise decomposition of the space, called **clustering**. Implementation of such algorithms can be found in the software **R** with the package Topological data analysis of Fasy *et al.* [15] or C++ and Python implementation from Maria *et al.* [16] and finally a private company, Ayasdi, [17] whose cofounder is Gunnar Carlsson.

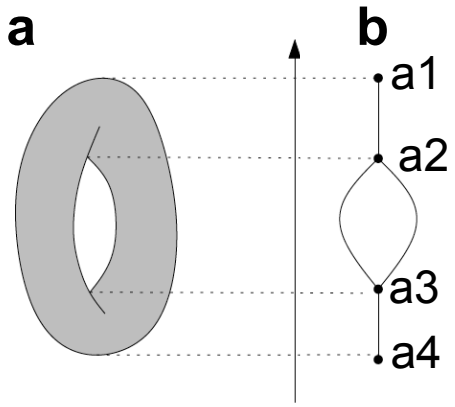


Figure 1.2: Illustration of the Reeb Graph. **(a)** A donut with the height function and **(b)** its Reeb Graph.

This enhanced clustering method based on algebraic topology considers high-dimensional datasets as point clouds and transforms them into networks; the nodes are clusters of samples, which are linked when they contain common samples [10]. As topology is insensitive to scale and small deformations, it is useful for the analysis of highly variable and noisy data and

reveals patterns not detected with standard tools [11], [12], [13]. It is used broadly to reduce dimensions and to recognize patterns in datasets as diverse as voting preferences, interactions of basketball players across games [11], and nanoporous material [18]. Mapper has been applied to analyze large biological datasets, such as global gene expression profiles of breast cancer samples [9] or temporal single-cell RNA-seq data [19]. In an approach called **Progression Analysis of Disease** (PAD)[9] consisting of a pre-processing step called Disease-Specific Genomic Analysis (DSGA) followed by Mapper, global gene-expression data are processed statistically and subsequently analyzed with Mapper [9], [20], [21], [13].
The Mapper algorithm is dependent on three parameters; a **filter function**, a **cover** of the codomain of the function and a clustering algorithm on the domain that is generally determined by a **distance** on the points in the domain, and a cutoff **parameter of closeness**. It has been shown that small changes in the choices of the cover of the codomain can affect the output and therefore make the method unstable [12]. Finding the best parameters usually requires trial-and-error strategies, until the output provides the most insight from the user perspective is reached [12]. This results in unusual choices that can be used only for limited numbers of data sets or even solely for the particular data set analysed. Despite frequently ending up with not much informative structure in the output for the user, another proposed strategy is given by estimating the distribution of the data points and then calculating a range of parameters for which the output is stable [22]. Recently advances towards a statistically well-founded version of Mapper has been made [23]. This strongly depends on the sampling of the data and the regularity of the filter function. A last approach proposed by Dey *et al.* [24],[25] is called multiscale Mapper and studies Mapper through multiple scales or ranges of a parameter, which they describe how best to choose in order to reach stability results. No application of this method have been found yet and theoretical results are still being established such as the relationship between Reeb graphs and multiscale Mapper [24]. We aim to develop a new method depending on the Mapper algorithm with optimized parameter selection for gene expression analysis. Therefore, we start by introducing the mathematics behind this method in the following sections.

Intuition 1.2.1. The particularity of Mapper and Reeb graphs is that they take as input a function of interest that helps describing or distinguishing spaces. An example is given by the distinction between $\times$ and $+$; as they are the same objects upon rotation, they are not distinguishable in the sense of standard topology and geometry. If one considers the function of height from top to bottom, then $\times$ starts with two pieces that merge into one at the middle and separate again, so the "flattened" $\times$ with the height function still is $\times$, whereas $+$ is at each level represented by one connected component, since the middle bar is flattened. Hence, the Reeb graph of $+$ is given by a line. Therefore, $\times$ and $+$ can be distinguished by their Reeb graphs.

1.2.2 Metric spaces and equivalence relations

In application the dataset that needs to be analysed is considered as a point cloud. In order to have a notion of proximity between elements of a point cloud, one defines a distance between every pair of elements of the point cloud. Taken together, those distances, if they verify certain conditions, are called a metric on the point cloud.

Intuition 1.2.2. An example of a distance on the point cloud X corresponding to the cities in Switzerland is for each two points/cities a and b in X given by the shortest time it takes by

car to travel from point a to point b.

Definition 1.2.3. A **metric space** (X, d) is a set X equipped with a function $d : X \times X \to \mathbb{R}_+$, which is called the distance, such that for any $x, y, z \in X$,

- $d(x, y) \geq 0$, and $d(x, y) = 0$ if and only if $x = y$
- $d(x, y) = d(y, x)$, and
- $d(x, y) \leq d(x, z) + d(z, y)$.

If the last point is not verified, it is called a semi-metric.

In order to define the Mapper algorithm, we need the formalism of an equivalence relation, which gives a binary relation between elements of a set X, or a point cloud.

Intuition 1.2.4. If the space X considered is the students of a class, then an example of an equivalence relation is the relation between the students given by "has the same birthday as".

Definition 1.2.5. A binary relation $\sim$ on a set X is called an **equivalence relation** if and only if it verifies for all x, y, z in X

$x \sim x$

$x \sim y$ if and only if $y \sim x$, and

$x \sim y$ and $y \sim z$ then $x \sim z$.

1.2.3 Simplices

Simplices are polyhedra built out of points, segments, triangles, tetrahedra, etc., put together in such a way that they are glued along common edges or faces.
We will introduce all the necessary formalism and illustrate it through examples.

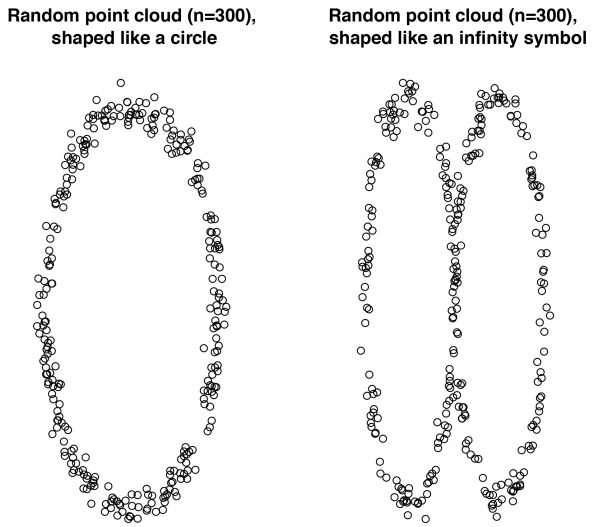


Figure 1.3: Illustration of two point clouds that topological tools are able to distinguish, on the right a circle shaped point cloud and on the left an infinity shaped point cloud.

Intuition 1.2.6. In order to distiguish the point clouds in Fig. 1.3, we would like to glue some segments between certain vertices in order to obtain the circle or the two circles that appear to the eye. These segments can not randomly be added, therefore we need to have an idea as to where to put them, and find an algorithm that would perform this task. In this first subsection, we are describing a formalism for segments, triangles, tetrahedra, etc.

Definition 1.2.7. A set of points $\{v_0, \dots, v_m\}$ in $\mathbb{R}^n$ is said to be **geometrically independent** if for any $\lambda_0, \dots, \lambda_m \in \mathbb{R}$ such that $\sum_{i=0}^m \lambda_i v_i = 0$ and $\sum_{i=0}^m \lambda_i = 0$, then $\lambda_i = 0$ for each $i = 0, \dots, m$.

Remark 1.2.8. This definition is equivalent to saying that the vectors $v_0 - v_1, v_0 - v_2, \dots, v_0 - v_m$ are linearly independent. Namely, if we have $\mu_1, \dots, \mu_m \in R$ such that $\sum_{i=1}^m \mu_i(v_0 - v_i) = 0$, then $\mu_i = 0$ for each $i = 1, \dots, m$.

We recall that a set $A \subset R^m$ is convex if, for two points of A, the line segment joining the points is still contained in A.

Lemma 1.2.9. *The intersection of convex sets is convex.*

Definition 1.2.10. The **convex hull** of $A \subseteq \mathbb{R}^m$ is the intersection of the convex sets containing A.

Proposition 1.2.11. *Take a set of geometrically independent points $\{v_0, \dots, v_m\}$ in $\mathbb{R}^n$. The **convex hull** of the set $\{v_0, \dots, v_m\}$ is equal to*

$$\sigma(v_0, \dots, v_m) = \left\{ v \in \mathbb{R}^n \mid v = \sum_{i=0}^m \lambda_i v_i \text{ where } \sum_{i=0}^m \lambda_i = 1, \text{ and } 0 \leq \lambda_i, \text{ for each } i = 0, \dots, m \right\}.$$

Definition 1.2.12. Take a set of geometrically independent points $\{v_0, \dots, v_m\}$ in $\mathbb{R}^n$. The **simplex** spanned by $v_0, \dots, v_m$ is the convex hull of the set $\{v_0, \dots, v_m\}$. The points $v_0, \dots, v_m$ are the **vertices** of the simplex $\sigma(v_0, \dots, v_m)$.

The dimension of the simplex spanned by $m + 1$ geometrically independent points is m, and we call it an **m-simplex**.
A **face** of an m-simplex σ is the simplex spanned by any subset of the sets of the $m + 1$ vertices $\{v_0, \dots, v_m\}$ of σ.

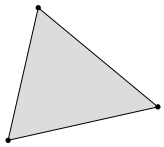
Remark 1.2.13. The fact that the dimension of a simplex spanned by $v_0, \dots, v_m$ is m and not $m + 1$ follows from the fact that $\{v_0, \dots, v_m\}$ spans a m-dimensional subspace of $\mathbb{R}^n$.

Here are some examples to understand these new notions.

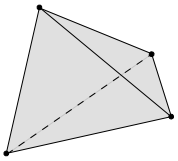
Example 1.2.14. In $\mathbb{R}^3$, the largest number of vertices a simplex can have is 4, as we cannot have more than 3 linearly independant vectors in $\mathbb{R}^3$. We construct simplices of dimension $-1, 0, 1, 2, 3$:

- ▷ (-1)-simplex : no vertices : $\emptyset$, the empty set.
- ▷ 0-simplex : one vertex : •, a point.
- ▷ 1-simplex : two vertices $\{v_0, v_1\}$, thus the linear combinations of the vertices as in Definition 1.2.12 gives the points $tv_0 + (1-t)v_1$ with $0 \leq t \leq 1$, which is the well known parametrisation for the segment between v_0 and v_1 : •——•

▷ 2-simplex : three vertices, so the combinations of the vertices as in Definition 1.2.12 give a full triangle :



▷ 3-simplex : here we have four vertices, and the combinations of these four vertices give an entire tetrahedron :



Definition 1.2.15. The n**-th standard simplex** is the n-simplex, $\Delta[n]$, spanned by $\sigma = \{e_1, \ldots, e_{n+1}\}$, where

$$(e_i)_j = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{otherwise.} \end{cases}$$

Remark 1.2.16. Hence, with the definition 1.2.12, we can see that the n-th standard simplex can be written,

$$\Delta[n] = \{(t_1, \ldots, t_{n+1}) \mid \sum_{i=1}^{n+1} t_i = 1, t_i \geq 0\}.$$

This is a useful example of an n-simplex.

1.2.4 Simplicial and abstract simplicial complexes

After the definition of a simplex, we can define what a simplicial complex is, an easier to grasp definition as its "abstract" version, called abstract simplicial complex.

Definition 1.2.17. A **simplicial complex** is a finite collection Δ of simplicies that verifies the two following conditions :

1. if $\sigma \in \Delta$ and $\tau \subseteq \sigma$ then $\tau \in \Delta$. Namely, if $\sigma \in \Delta$, then any face of σ has to be in Δ;
2. if $\sigma, \tau \in \Delta$, then $\sigma \cap \tau$ has to be a face of both σ and τ.

Definition 1.2.18. An **abstract simplicial complex** is a pair (V, Σ) where V is a finite set and $\Sigma = \coprod \Sigma_k$ is a set of ordered subsets $\Sigma_k \neq \emptyset$ of V such that if $\sigma \in \Sigma$, then for every $\tau \subseteq \sigma$, we have $\tau \in \Sigma$, and if $A \in \Sigma_k$, then $|A| = k + 1$.

The definition is aptly named, as it is an abstract object in the sense that if we want to visualise an abstract simplicial complex, we need to define its geometric realization. To construct the latter, we will look at simplices associated to the Σ_k, taking their vertices to be the associated points of V. We glue them together along vertices that correspond to the same point in V.

Definition 1.2.19. Let K be a simplicial complex (abstract or not). A subcollection L of simplices from K which in turn forms a simplicial complex is called **a subcomplex** of K. In other words, if a simplex σ is in L, then all of its faces in K are also present in L.

- We will write $\Delta[k]_\sigma$ to emphasize the fact that the space $\Delta[k]$ corresponds to the set $\sigma = (v_0, \dots, v_k) \in \Sigma_k$.
- For any $i = 0, \dots, k$, the map δ_i is defined by

$$\begin{aligned} \delta_i : \quad \Sigma_k \quad &\longrightarrow \quad \Sigma_{k-1} \\ (v_0, \dots, v_k) &\longmapsto (v_0, \dots, v_{i-1}, v_{i+1}, \dots, v_{k+1}), \end{aligned}$$

- For any $i = 0, \dots, k-1$, the map d_i is

$$\begin{aligned} d_i : \Delta[k-1] &\longrightarrow \Delta[k] \\ (x_0, \dots, x_{k-1}) &\longmapsto (x_0, \dots, \underbrace{0}_{i^{th}}, \dots, x_{k-1}), \end{aligned}$$

Definition 1.2.20. The **geometric realization** of the abstract simplicial complex (V, Σ), written $|(V, \Sigma)|$, is

$$|(V,\Sigma)| = \coprod_k \coprod_{\Sigma_k} \Delta[k]_\sigma \Big/ \Big(x \sim d_i x,\ \forall x \in \Delta[k-1]_{\delta_i \sigma} \Big) \tag{1.1}$$

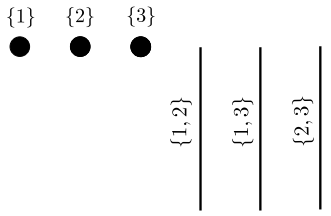
where $d_i x \in \Delta[k]_\sigma$, equipped with the quotient topology.

To better understand this complicated definition, some examples are given.

Examples 1.2.21. 1. Take $V = \{1, 2, 3\}$ and $\Sigma = \{(1), (2), (3), (1, 2), (1, 3), (2, 3)\}$. We get

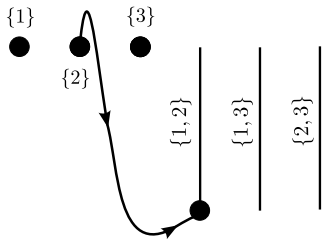
$$|(V,\Sigma)| = \begin{pmatrix} \Delta[0]_{(1)} \coprod \Delta[0]_{(2)} \coprod \Delta[0]_{(3)} \coprod \\ \Delta[1]_{(1,2)} \coprod \Delta[1]_{(1,3)} \coprod \Delta[1]_{(2,3)} \end{pmatrix} / \sim$$

and here $\delta_0((1, 2)) = (2)$, $\delta_1((1, 2)) = (1)$, etc. Thus, we have a list of 0- and 1-simplices :

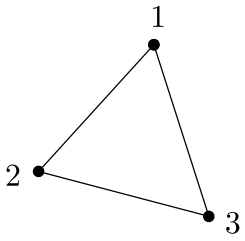


(1.2)

To patch them together, we apply the equivalence relation in (1.1). For example, take $\sigma = (1, 2) \in \Sigma_1$. So $k = 1$, and take $i = 0$. Then $\Delta[0]_{\delta_0 \sigma} = \Delta[0]_{(2)}$ which is the second point in (1.2). And $d_0(\Delta[k-1]_{\delta_0 \sigma})$ is the extremity of the edge $(1, 2)$, corresponding to (2). So we patch the point (0-simplex) corresponding to (2) to the extremity associated to 2 of the edge (1-simplex) $(1, 2)$. We get :



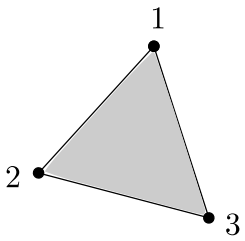
And we do the same for all $\sigma \in \Sigma$, and we get the geometric realization of (V, Σ) :



2. If we keep the same V but just add $(1, 2, 3)$ to Σ, namely

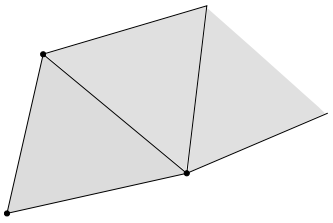
$$\Sigma' = \Sigma \cup \{(1, 2, 3)\}.$$

We have one additional 2-simplex in this example. To construct the geometric realization $|(V, \Sigma')|$ we paste each 1-simplex namely each edge, of the 2-simplex to the edge corresponding to the 1-simplex of Σ : $(1, 2), (1, 3), (2, 3)$. That gives :

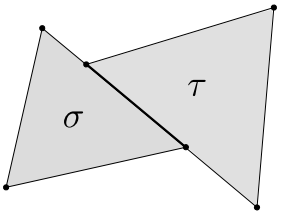


Examples 1.2.22. Here are some examples that are not realizations of abstract simpicial complexes.

1. An edge and two vertices are missing in the following diagram.

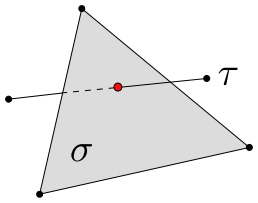


2. We have that $\sigma \cap \tau$, which is the thicker segment in the following diagram, is neither an edge of σ, nor one of τ.



Thus, the above picture is not a realization of an abstract simplicial complex.

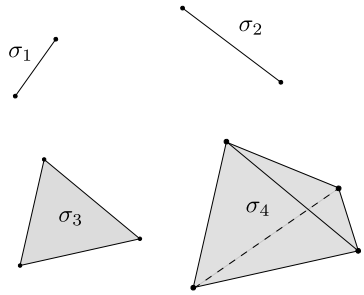
3. Here, $\sigma \cap \tau$ is the red point that is a vertex of neither σ, nor of τ :



This is not a realization of an abstract simplicial complex.

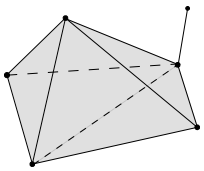
Here are some other examples of realization of abstract simplicial complexes.

1. We can take disjoint simplices for the collection :



Namely we have $\{\sigma_1, \sigma_2, \sigma_3, \sigma_4\}$ and all their faces.

2. Take two 3-simplices and one 1-simplex as follows :



This is a realization of an abstract simplicial complex as in particular all their intersections are faces of each of the two complexes.

Definition 1.2.23. A **simplicial function** is a map between two abstract complexes (V, Σ) and (V', Σ'), is a map $f : V \to V'$ verifying that if $(v_0, \ldots, v_n) \in \Sigma$, then $(f(v_0), \ldots, f(v_n)) \in \Sigma'$.

Remark 1.2.24. We can construct the category, **Csim**, whose objects are the abstract simplicial complexes, and the maps $f \in Map(\mathbf{Csim}, \mathbf{Csim})$ are the simplicial functions defined in the previous definition 1.2.23.

The realization of an abstract simplicial complexes defines a functor from the category **Csim** of abstract simplicial complexes to the category of topological spaces, once we have defined what the realisation of a simplicial function is:

Definition 1.2.25. The **realization of a simplicial function** f is a continuous extension of f to the realization of the abstract complexes, that is a map $|f| : |(V_1, \Sigma_1)| \to |(V_2, \Sigma_2)|$ such that the vertices of a simplex are always sent to vertices in the image. Namely, for $x = \sum_{i=0}^{n} \lambda_i v_i$, we have

$$|f|(x) = \sum_{i=0}^{n} \lambda_i f(v_i).$$

1.2.5 Nerve

Intuition 1.2.26. Since we defined what an abstract simplicial complex is, we have a formalism for the "links" between points. We need to know when to add such connections between points in a point cloud. Therefore, we will define an object that will be essential in the Mapper method : the nerve of a cover. The definition of a cover of a space is again central for the Mapper algorithm. This formalises the idea of dividing the space into subparts.

Definition 1.2.27. A cover of a space X is a family of sets U_α, $\alpha \in A$ such that $X = \cup_{\alpha \in A} U_\alpha$.

Examples 1.2.28. We will illustrate this concept by studying several covers of the circle $\mathbb{S}^1$. The first one in the Fig. 1.4 (a) is the cover with the northern hemisphere (in red) and the southern hemisphere (in blue). Another example (Fig. 1.4b) illustrates three subspaces that

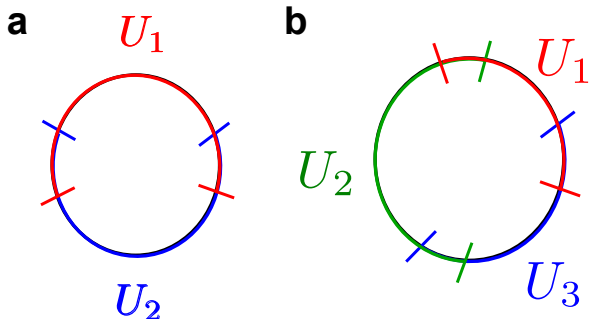


Figure 1.4: Covers of $\mathbb{S}^1$. **(a)** A first way to cover $\mathbb{S}^1$ using hemispheres. **(b)** A second way using three subspaces.

cover $\mathbb{S}^1$. These two examples clearly highlight the existence of a variety of ways to cover the same space.

Definition 1.2.29. Let $\mathscr{U} = \{U_\alpha\}_{\alpha \in A}$ be a cover of a space X. The **nerve** of the cover $\mathscr{U}$, denoted $\mathscr{N}(U)$, is the geometric realization of the abstract simplicial complex (V, Σ), where

$$V = A,$$
$$\Sigma = \coprod_{n \geq 0} \{(\alpha_0, \ldots, \alpha_n) \subseteq A \mid U_{\alpha_0} \cap \ldots \cap U_{\alpha_n} \neq \emptyset, \text{ and } i \neq j \Rightarrow \alpha_i \neq \alpha_j\}.$$

We then have $\Sigma_k = \{(\alpha_0, \ldots, \alpha_k) \subseteq A \mid U_{\alpha_0} \cap \ldots \cap U_{\alpha_k} \neq \emptyset, \text{ and } i \neq j \Rightarrow \alpha_i \neq \alpha_j\}$. If $\sigma \in \Sigma$, there exists a k such that $\sigma \in \Sigma_k$, and saying that $\tau \subseteq \sigma$ means that $\tau \in \Sigma_j$ for $j \leq k$, and its geometric realization is a subspace of σ's realization, as a polyhedron in $\mathbb{R}^k$.

Examples 1.2.30. 1. Take $X = \mathbb{S}^1$, the circle. Take the cover $\mathscr{U} = \{U_1, U_2\}$ as in Fig. 1.4a.

We have $A = \{1, 2\}$, therefore, for a subset of A we have three possibilities :

- $\{1\}$: which gives a 0-simplex : •,
- $\{2\}$: which gives a 0-simplex : •,
- $\{1, 2\}$: which gives, as $U_1 \cap U_2 \neq \emptyset$, a 1-simplex : •——•

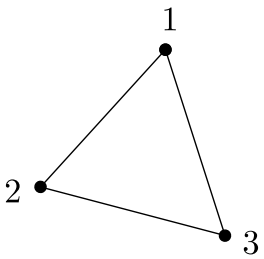
And so we get the following geometric realisation for the nerve $N(\mathscr{U})$:

2. Keep $X = \mathbb{S}^1$. Take the cover $\mathscr{U} = \{U_1, U_2, U_3\}$ in Fig. 1.4b.

We have $A = \{1, 2, 3\}$. For subsets of A we have

- $\{1\}, \{2\}, \{3\}$: each one gives a point,
- $\{1, 2\}, \{1, 3\}, \{2, 3\}$, : each one gives an edge as $U_i \cap U_j \neq \emptyset$ for any $i, j = 1, 2, 3$,
- $\{1, 2, 3\}$: which gives nothing since $U_1 \cap U_2 \cap U_3 = \emptyset$.

This time we have three points with three edges, and then the nerve $N(\mathscr{U})$ is given by :



Remark 1.2.31. We see that the second cover of $\mathbb{S}^1$ is better than the first one as the nerve $N(\mathscr{U})$ has the homotopy type (or the same "shape") of $\mathbb{S}^1$. We thus rather recover $\mathbb{S}^1$ with this latter cover.

Homotopy equivalence

Two continuous maps f_0, $f_1 : X \to Y$ are said to be homotopic if there exists a continuous map $H : X \times [0, 1] \to Y$ such that for any $x \in X$, $H(x, 0) = f_0(x)$ and $H(x, 1) = f_1(x)$. Let X and Y be two topological spaces if there exists $f : X \to Y$ and $g : Y \to X$ such that $f \circ g$ and $g \circ f$ are homotopic to the identity maps of Y and X respectively, then X and Y are said to be homotopy equivalent. Let X be a topological space. If X is homotopy equivalent to a point ($Y = \{\star\}$), then X is said to be contractible.

The nerve theorem
If the cover is well-chosen one can prove a strong relation between X and the nerve of the cover.

Theorem 1.2.32. *Let topological space X and let $U = (U_i)_{i\in I}$ be a cover by open sets such that if $C = \cap_{i\in J} U_i$, $J \subseteq I$, then C is either empty or contractible. Then, X and the nerve of $\{U_i | i \in I\}$ are homotopy equivalent.*

Example 1.2.33. In the first example 1.2.28, the intersection between U_1 and U_2 is not empty and not contractible (it is homotopy equivalent to two points). The second example verifies the hypothesis of the theorem 1.2.32, and the nerve is given by a triangle which is homotopy equivalent to the circle.

1.2.6 Reeb Graphs

Intuition 1.2.34. As already mentioned in the introduction section 1.1, Reeb graphs [14] are used as approximation of spaces and should be thought of as a thin version of a space (Fig. 1.2). The Mapper algorithm is an approximation of the Reeb graph, since the Reeb graph is not computable.

Given a topological space X and a continuous function $f : X \to \mathbb{R}$, we define the equivalence relation $\sim_f$ between points of X by:

$$x \sim_f y \iff f(x) = f(y) \text{ and } x, y \text{ belong to the same connected component of } f^{-1}(f(x)) = f^{-1}(f(y)).$$

The *Reeb graph* [14], denoted by $\mathrm{R}_f(X)$, is the quotient space $X/\sim_f$.
As f is constant on equivalence classes, there is an induced map $\tilde{f} : \mathrm{R}_f(X) \to \mathbb{R}$ such that $f = \tilde{f} \circ \pi$, where π is the quotient map $X \to \mathrm{R}_f(X)$.

1.2.7 The Mapper algorithm

Using a function $f : X \to Z$, called **the filter function**, a special type of cover of X is created from any cover of Z, by pulling back along f. This is needed here as the clustering in the Mapper algorithm is applied to the pullback of a cover of the space Z.

Definition 1.2.35. Let $f : X \to Z$ be a function and let $\mathscr{U} = \{U_\alpha\}_{\alpha\in A}$ be a cover of the space Z. The **pullback** of the cover $\mathscr{U}$ is given by $f^{-1}(\mathscr{U}) = \{f^{-1}(U_\alpha)\}_{\alpha\in A}$.

Algorithm 1: The Mapper Algorithm

Input: A data set or point cloud X with a metric or a dissimilarity measure between data points d, a filter function $f : X \to \mathbb{R}$ (or $\mathbb{R}^d$), a cover $\mathscr{U}$ of $f(X)$ and a closeness parameter ε.
Method: For each $U \in \mathscr{U}$, decompose $f^{-1}(U)$ into clusters $C_{U,1}, \ldots, C_{U,n_U}$, using an algorithm that partitions data into clusters such as single-linkage clustering with parameters d and ε (See Algorithm 2, section 1.2.8).
Compute the nerve of the cover of X defined by the sets $C_{U,1}, \ldots, C_{U,n_U}$, $\forall U \in \mathscr{U}$.
Output: a simplicial complex: for every $U \in \mathscr{U}$, we define
- vertices $v_{U,i}$ for each cluster $C_{U,i}$, where $i = 1, \ldots, n_U$,
- k-simplex between $v_{U_0,i_0}, \ldots v_{U_k,i_k}$ if and only if $C_{U_0,i_0} \cap \cdots \cap C_{U_k,i_k} \neq \emptyset$, where $U_j \in \mathscr{U}$ and $i_j \in \{1, \ldots, n_{U_j}\}$, for every $j = 1, \ldots, k$

Remark 1.2.36. It is therefore clear that if the cover of the codomain has at most two opens that overlap, no 2-simplex will be drawn. This means that in that case the algorithm only displays vertices and edges and is therefore a graph.

1.2.8 Clustering algorithms used in the Mapper algorithm

Numerous clustering algorithms exist, but only single-linkage [26], complete linkage [26] and average linkage clustering [26] are implemented algorithms for the Mapper algorithm as they are popular and density-based. We focus here on the single-linkage algorithm as it is the most frequently used clustering algorithm in applications of Mapper and the one implemented in the "Two-tier Mapper", the method we will define in chapter 2 and is the popular choice of the general use of the Mapper algorithm.

Algorithm 2: single-linkage clustering

Input: A data set or point cloud X with a metric or a dissimilarity measure between data points d, a parameter of closeness ε, and an open set of X.
Method: Define the relation $\sim_1$, by $x \sim_1 y$ if and only if $d(x, y) \leq \varepsilon$. This becomes an equivalence relation $\sim$ under transitive closure in U of $\sim_1$, i.e $x \sim y$ if and only if there exists $x_1, \dots x_n \in U$ such that $d(x_i, x_{i+1}) \leq \varepsilon$, for $i \in 0, \dots, n$, where $x_0 = x$ and $x_n = y, n \in \mathbb{N}$.
Let U be an open of X, we define the clusters to be the equivalence classes of $\sim$: $[U_1], \dots, [U_{n_U}]$.
Output: a partition of U into clusters $[U_1], \dots, [U_{n_U}]$.

Example 1.2.37. The single-linkage clusters can be obtained by drawing the ε-neighbourhood graph, i.e. linking every two points X and Y that verify $d(X, Y) \leq \varepsilon$ with an edge (in red in Fig. 1.5) on the point cloud and then extracting from this graph the connected components. The equivalence classes are drawn in this example as discs with size corresponding to the number of samples in the cluster. This representation is employed as well in the Mapper algorithm. In the following Fig. 1.5, the point cloud is considered with the euclidean distance and the epsilon parameter is chosen (bottom right). This results in an ε-neighbourhood graph (red edges) which, by taking the connected components, gives the single-linkage clusters.

1.2.9 The choices of parameters

Since Mapper has a wide range of applications, from the analysis of basketball players interaction during games [11] to the classification of nanoporous materials [18], there are also a wide range of choices for the different parameters. Software implementing Mapper allows one to choose some of the parameters to different degrees [27], [15], some can freely be chosen and some are chosen among few. In the following sections, we want to explain each parameter and show this variety of choices, mostly linked to gene expression analysis, to illustrate and guide the user towards analysing data using Mapper. The method developed in this thesis uses Mapper as well, but with specific parameters that are different from the standard ones (described in this section). Therefore, the choice of those parameters for that new algorithm will be defined in section 2.3.

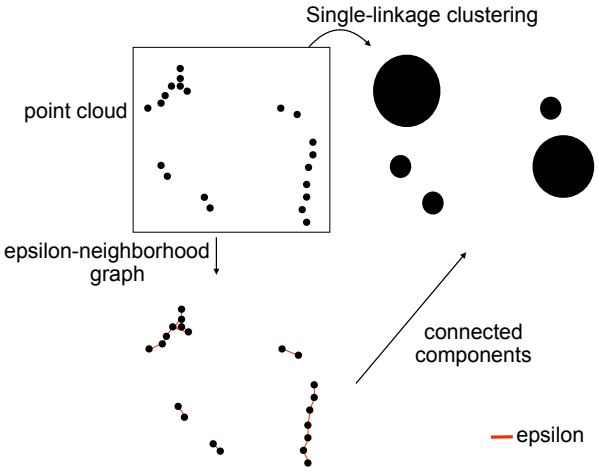


Figure 1.5: Example of the single-linkage clustering algorithm. Starting from a point cloud in the box, single-linkage clustering is the same as drawing the ε-neighbourhood graph and then taking the connected components.

The distance

The distance is used to segregate the data and is hence chosen in order to provide the user with information about proximity in the dataset. In the case of Mapper, it is enough to work with a similarity measure, not necessary verifying the triangle inequality. The choice of the similarity measure influences highly the clusters formed, as two samples could be close using one distance, but far away according to another. It needs to be chosen according to the problem and the interpretation of the output will be different. Several similarity measures have been implemented and are proposed.

The correlation distance is used in many applications to expression datasets [19], [9], [13]. Even though it is not often used, euclidean distance or variance normalized euclidean distances can be the metric of choice as it is the most common segregation method in hierarchical clustering and the default distance in the software **R** [26]. It has been used to separate basketball players [11] for instance.
Other popular distances include : special types of topological measures [18], L_p-metric (and usually p is chosen to be 2) [8].

The filter function

This function is used to gain insight into the data and should therefore be problem dependent. The filter function is often also used to color code the clusters obtained in order to know the average value of the filter function for samples in that cluster. In the different fields of application, different functions have been proposed.
Often, dimension reduction algorithms such as Multidimensional Scaling (MDS) or Principal Component Analysis (PCA) are used to generate a filter function, i.e. a filter function could be given by the first MDS or PCA component or even a function with codomain in two or more dimensions is given by taking more than one principal component. Two-dimensional MDS of the 5000, respectively 4600 genes with most variance was used, depending on the experiment, in the analysis of scRNA-seq data [19]. This dimension reduction is useful in order to

screen for topological features in single-cell RNA-seq, but in different regions of the MDS plots.

In the analysis of breast cancer data [9], the chosen parameter resulted in the method PAD and the following filter function f was chosen :

$$f : \mathsf{T} \to \mathbb{R} : T_k \mapsto (\sum_{l \in S} \mid (Dc.T_k)_l \mid^l)^{p/l},$$

where T_k is a tumour sample that was decomposed into two parts using linear regression a normal component $Nc.T_k$ and a disease component $Dc.T_k$, for every $k = 1, \ldots, S$, where S is the number of samples in the group of patients with a certain disease. The two new parameters p and l were chosen between 1 to 5 and 1 to 10 respectively [9] and have been chosen to be 4 and 2 in the breast cancer cohort [9] and unspecified values in another article using PAD [20].

Other popular filter functions include : singular value decomposition in the case of observing the interactions between basketball players or the voting preferences of the US House of Representatives [11], L-Infinity Centrality and Event Death, which is a two-dimensional function, in the case of breast cancer gene expression data [11] or the neighbourhood Lens in the case of analysis of nanoporous material [18].

The cover of the codomain of the filter function

The cover, described in definition 1.2.27, of the codomain of the filter function is the most important choice as it will determine the refinement at which the data is studied. An example of the impact of the refinement can be seen in examples 1.2.28, where two different types of covers result in graphs with topologically distinct nerves (Fig. 1.4 a, b).

The most commonly applied type of cover of the codomain consists of a number P of intervals with a percentage of overlap G between two consecutive intervals [27], [15], [19], [8]. It is sometimes referred to as the gain (% of overlap= 1-1/gain) and the resolution, corresponding to the diameter of an interval [23], or sometimes the minimimum length of an interval [23], and yet other times the number of intervals [27]. If the codomain of the filter function is $\mathbb{R}^n$ the same idea is extended to n dimensions: a certain number of hyperrectangles cover the space overlapping two-by-two with a certain percentage. Hence, the user inputs the number of intervals and the % of overlap per dimension. As examples, 15 intervals, with 80% overlap were chosen in the breast cancer data [9]. This data has been re-analysed with a two-dimensional filter function where in the "L-Infinity Centrality" dimension the resolution is 70 and the gain is 3, and in the "Event Death" dimension 30 and 3 respectively.
In the single-cell RNAseq analysis 26×26 and 62×62 rectangular patches were considered for the two datasets in the paper [19]. To avoid sampling-density biases, Rizvi *et al.* choose the size of the patches such that the number of cells in each row or column of patches is the same [19]. The overlap between patches was 66% but not in all the figures of the paper [19].

The epsilon parameter

The parameter ε is used as a cutoff for the single-linkage clustering algorithm determining when two samples are close enough in terms of the chosen distance.

In Ayasdi and [8], the parameter ε of closeness is described as a parameter that should depend on the previously determined number of intervals and is not available for the user to choose. A high number of intervals should give a small value for ε whereas a small number should allow for a high value of ε [27]. One method to determine ε is to search for a gap in the histogram of the number of merges of the dendrogram generated using the distance (Fig. 1.6).

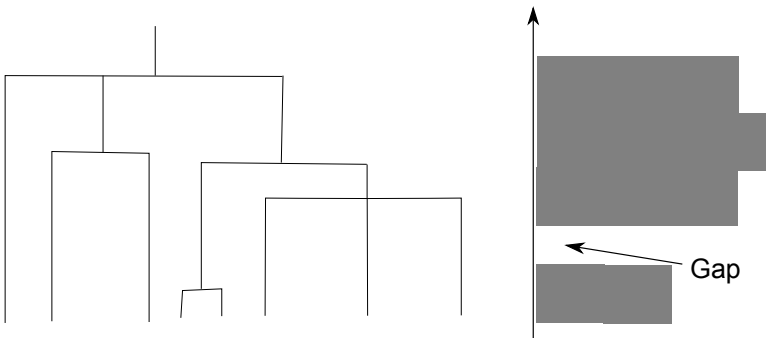


Figure 1.6: Illustration of the ε parameter selection using a dendrogram representation (left) of the distance between samples and a histogram of the merges (right).

1.2.10 Example of the Mapper algorithm

Example 1.2.38. Starting from a point cloud on an object (Fig. 1.7a), in this example of the Mapper Algorithm, a hand, we want to reduce the complexity of the space and gain insight with respect to a function of interest.

The function of interest or filter function of this example is the distance to the palm, color-coded from blue, close to the palm, to red, tip of the fingers (Fig. 1.7b). One then chooses a cover, a distance and an ε parameter of closeness that are given here by 6 intervals, with 50% overlap, the three-dimensional euclidean distance and ε (Fig. 1.7c). In each of the pre-images of the intervals, i.e. on the points with function value inside that interval, single-linkage clustering is performed with the distance and ε as inputs. It gives in each intervals the connected components of the ε-neighbourhood graph. These connected components are depicted as circles that are linked if the clusters share a point (Fig. 1.7d). As the intervals are overlapping there are always points in the intersection of two consecutive intervals and one obtains a link between the components in one interval and the next if they share a common point. If epsilon is too big, then the fingers all merge into one connected component and the resolution of the topology of the object is lost. If epsilon is too small, isolated points start to appear. One would have more than 5 fingers, and the topology of the starting object would be lost as well. An equilibrium needs to be found and parameters well chosen.

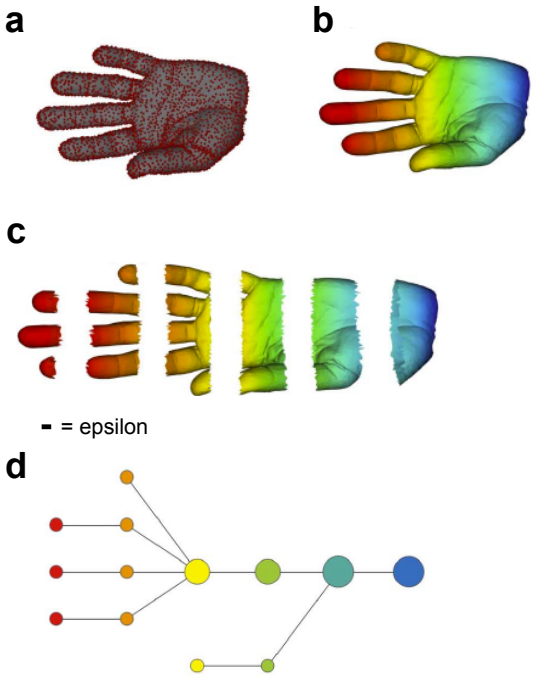


Figure 1.7: Example of the Mapper algorithm (adapted from [11]). **(a)** A point cloud with **(b)** a filter function associated to the point cloud which is the distance to the palm represented as a color code. **(c)** Dividing the color code space into overlapping bins results in an overlapping cover of the point cloud and choosing a parameter of closeness ε is then resulting in a **(d)** Mapper output.

1.3 Persistent and extended persistent homology

1.3.1 Introduction to persistent homology

Another method of topological data analysis is called Persistent Homology (PH) [28]. This notion emerged in three different research groups (Italy, Colorado, Duke) at almost the same time.

In Italy, the group of Patrizio Frosini, Massimo Ferri, and Claudia Landi started the field of size theory, which uses PH to give a well-structured mathematical answer to problems in shape recognition[4], [5], [7]. The aim is to capture information about the shape of the studied space by using a function, called a **size function**. Having characterized the shape of an object by summarising its topological features in a 2D plot [6] and having defined a distance between these 2D summaries, they are able to classify spaces by shape. Application of this theory ranges from the recognition of the sign language alphabet[29] to the study of the evolution of cyclones [30]. In 2011, this group in Italy managed to enhance the size theory by proving a theorem stating that whenever two functions on a sphere are given, and certain criteria verified, then there exists a reparametrisation of the sphere bringing the two functions close [31]. In the later chapter 6.1, we will discuss how this is useful in the research on hormones.

In her doctoral work in Colorado, V. Robins studied fractal sets using PH [32]. Her analyses of approximations on spaces led to remarkable results in material science on classification of crystalline patterns [33] or X-ray CT images of porous and granular materials [34].

Finally, at Duke and Stanford, developments of the concept of alpha shapes [35] resulted in algorithms for PH computations. Edelsbrunner et al. [36] analysed the Klein bottle [37] and was popularised with image analysis in [8] and [38].

PH is an algorithm that takes as input nested families, which are referred to as a filtration, of higher dimensional generalizations of graphs, called **simplicial complexes** [35], [39]. Many methods for converting point cloud datasets into simplicial complexes have been developed [8], [40], [35], [41]. Several different approaches to building filtrations have been proposed. The most popular and common one is the filtration arising from a function f from the space or dataset studied to $\mathbb{R}$, given by taking sublevel sets [31]: i.e., at a time i every point that f maps to a value less than or equal to i is inside that "level" of the filtration (Fig. 1.8).

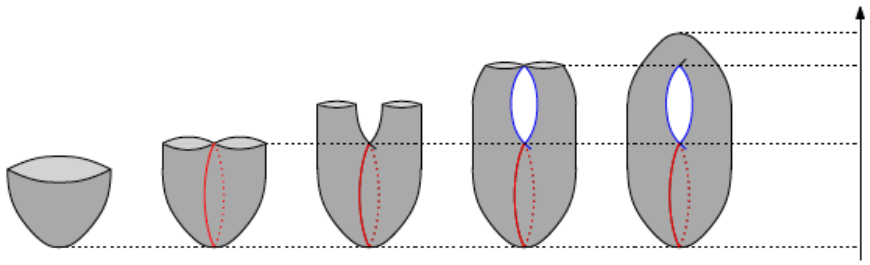


Figure 1.8: Illustration of a filtration, representing a nested family of spaces. On the left the smallest subpart and on the right, the full space. The function used to generate that filtration is the height function depicted on the far right (Adapted from [23]).

Given such a filtration of the object of study, i.e., growing subparts of the space, one keeps track of the time points of the filtration at which a certain topological feature appears , called the birth, and disappears, called the death of the feature. This is then summarised in a 2-dimensional plot [8]. This enables us to see which features persist a long time and which

features are only transiently apparent. These 2D summaries are therefore called **Persistent Diagrams** (PD) . In the beginning of the 21st century, natural distances on these diagrams were introduced [42], [43], in order to compare and distinguish the PD of two spaces. Statistics were then defined on PD in order to establish rigorous mathematical foundation for calculations [44]. Statistics were also conjointly used with PD to form new methods [45] giving rise to a new way to classify spaces. Application of persistent homology ranged from areas such as neural networks [46], [47], to histochemistry scoring of breast cancer pathological slides [48], passing by fingerprint classification [49] or even differentiating handwritten letters [8].
In the following sections, we provide a more in-depth description of the concepts sketched above.

1.3.2 Homology

Homology is a concept of algebraic topology used to detect and characterize topological features. Any k-dimensional hole is represented by an element of the vector space H_k. Therefore, H_0 measures connected components of a space, H_1 the 1-dimensional holes or loops, H_2 the 2-dimensional holes or voids, etc ... We can then know how many k-dimensional holes a space has. For less heavy notation, we will choose a field that we fix $F = \mathbb{Z}_2 = \{[0], [1]\}$, the field of two elements verifying $[1] + [1] = [0]$. We refer the reader to [50] and [51] for further concise introductions to homology.

Example 1.3.1. In the following example (Fig. 1.9), the space, a torus, is connected and has two one-dimensional holes or loops (blue, Fig. 1.9), and one two-dimensional hole (green, Fig. 1.9).

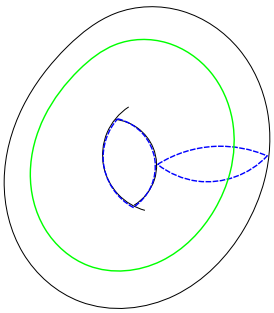


Figure 1.9: Illustration of a topological space, a torus, with its one-dimensional holes in blue and the two-dimensional holes in green.

The vector space of k-Chains

Definition 1.3.2. Let K be a simplicial complex with a finite number of simplices in each dimenison. A **k-chain** in K with coefficients in F is defined as a formal sum of k-simplices in K, i.e.

$$\sum_{i=1}^{p} \varepsilon_i \sigma_i,$$

where σ_i is a k-simplex of K and $\varepsilon_i \in F$ for all $i = 1, \ldots, p$, and p is the number of k-simplices in K.
The spaces of all k-chains is denoted $C_k(K)$, where the field is suppressed from the notation.

Remark 1.3.3. $C_k(K)$ is a vector space as there is an addition and a multiplication by a scalar defined on this space. If $c, c' \in C_k(K)$, then there exist $\varepsilon_i, \varepsilon'_i \in F$, for $i = 1, \dots, p$ such that $c = \sum_{i=1}^{p} \varepsilon_i \sigma_i$ and $c' = \sum_{i=1}^{p} \varepsilon'_i \sigma_i$. We define the sum of the two k-chains by : $c + c' = \sum_{i=1}^{p} (\varepsilon_i + \varepsilon'_i)\sigma_i$. The multiplication by a scalar $\lambda \in F$ is defined by $\lambda \cdot c = \sum_{i=1}^{p} \lambda \cdot \varepsilon_i \sigma_i$.

The boundary operator, its kernel and image

Definition 1.3.4.

- Let $\sigma = [v_0, \dots, v_k]$ be a k-simplex. The **boundary of** σ is the $(k-1)$-chain $\partial_k(\sigma) = \sum_{i=0}^{k} (-1)^i [v_0, \dots, v_{i-1}, v_{i+1}, \dots, v_k]$. In the field we have chosen $[-1] = [1]$ and therefore $\partial_k(\sigma) = \sum_{i=0}^{k} [v_0, \dots, v_{i-1}, v_{i+1}, \dots, v_k]$.

- The **boundary operator** on $C_k(K)$ is a linear extension of the boundary operator defined on individual simplices. Let $c = \sum_{i=1}^{p} \varepsilon_i \sigma_i \in C_k(K)$ be a k-chain. Then, $\partial_k(c) = \sum_{i=1}^{p} \varepsilon_i \partial_k(\sigma_i)$.

- The **kernel of the boundary operator** $Z_k(K) = \{c \in C_k(K) \mid \partial_k(c) = 0\}$ is called the space of k-cycles of K.

- The **image** $B_k(K) = \{c \in C_k(K) \mid \exists c' \in C_{k+1}(K), \partial_{k+1}(c') = c\}$ is called the space of k-boundaries of K.

The boundary operators verify the following composition property: $\partial_{k-1} \circ \partial_k = 0$ for any $k > 1$. Therefore, any k-boundary is a k-cycle and there are inclusions $B_k(K) \subseteq Z_k(K) \subseteq C_k(K)$.

Example 1.3.5. In the following example (Fig. 1.10), let $F = \mathbb{Z}_2$, then $\sigma = [v_0, v_1, v_2]$ is a 2-simplex and $\partial_2(\sigma) = [v_1, v_2] + [v_0, v_2] + [v_0, v_1]$, which is the formal sum of the 1-simplices that form the empty triangle and corresponds to our intuition of a boundary.

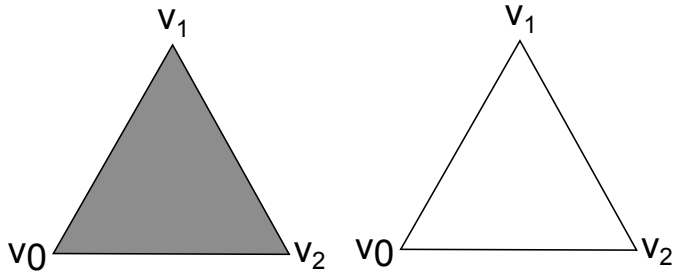


Figure 1.10: Illustration of a 2-dimensional simplex (left) and its boundary (right).

Simplicial homology groups and Betti number

The k-th (simplicial) homology group of K is the quotient vector space

$$H_k(K) = Z_k(K)/B_k(K).$$

The k-th Betti number, named after Enrico Betti, of K is the dimension of the vector space $H_k(K)$, given by $\beta_k(K) = \dim(H_k(K))$.

Relative homology

The same idea defining homology and homology groups is used to define another homology theory called the relative homology. For that, let X be a space and $A \subseteq X$ a subspace

of X. Let $C_k(X, A)$ be the quotient group $C_k(X)/C_k(A)$. This implies that chains in A are trivial in $C_k(X, A)$. The boundary map $\partial_k : C_k(X) \to C_{k-1}(X)$ induces a well defined "relative" boundary map $\partial_k : C_k(X, A) \to C_{k-1}(X, A)$. Indeed, it can easily be verified that ∂_k sends $C_k(A)$ to $C_{k-1}(A)$. We define the kernel and the image of the relative boundary operator ∂_k in the same way as in the previous paragraphs. The relation $\partial_{k-1} \circ \partial_k = 0$ for any $k > 1$ still holds. Therefore, any k-boundary is a k-cycle and there are inclusions $B_k(X, A) \subseteq Z_k(X, A) \subseteq C_k(X, A)$. The **relative homology groups** $H_k(X, A)$ are defined by $Z_k(X, A)/B_k(X, A)$.

1.3.3 Persistent homology

Persistent homology was developed to cope with noise in data, distinguish relevant feature of spaces, and describe particular topological feature of a space at different resolutions. When using persistent homology, the evolution of homology groups across nested families of simplicial complexes is studied, detecting features that are apparent at different scales and calculating how long these topological features *persist*. Persistent homology is often applied to point clouds, as they can easily be turned into simplicial complexes.

Intuition 1.3.6. The intuition one should have for this section is that persistent homology machinery will be able to distinguish for instance a circle-shaped point cloud and an ∞-shaped point cloud (see Fig. 1.3), as they have respectively one one-dimensional hole and two one-dimensional holes that are apparent.
Despite the fact that the human eye directly recognize a circle-shaped point cloud in Fig. 1.3, we need to have a formalism to determine when this should be considered to be a circle. The idea is to link the points according to a chosen criterion, and decide when a feature can be considered to be real rather than noise. There are several ways to proceed, and we will discuss hereafter the ones that are computationally the most efficient.

Čech and Vietoris-Rips complexes

Definition 1.3.7. Let $d : Z \times Z \to R$ be a metric or a similarity measure on a topological space Z.
The **open balls** $B_\varepsilon(z)$, describing a basis of the metric topology, are defined for any $\varepsilon > 0$ and $z \in Z$ as

$$B_\varepsilon(z) = \{y \in Z \mid d(z, y) < \varepsilon\}.$$

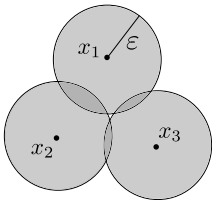
For any $\varepsilon > 0$, the family of sets $\{B_\varepsilon(z)\}_{z \in Z}$ covers the space Z as each $z \in Z$ is contained in at least $B_\varepsilon(z)$.
Let $V \subseteq Z$ be a subset such that $\cup_{z \in V} B_\varepsilon(z) = Z$. The nerve (see definition 1.2.29) of this cover is the **Čech complex** attached to V and ε, denoted $\check{C}_\varepsilon(V)$

Definition 1.3.8. For a metric space Z with distance function d and $\varepsilon > 0$, the **Vietoris-Rips complex**, written $VR(Z, \varepsilon)$, is the simplicial complex with vertices the elements in Z and where a set of points $\sigma = \{z_1, \ldots, z_n\}$ spans a k-simplex if and only if $d(z_i, z_j) \leq \varepsilon$ for any $z_i, z_j \in \sigma$.

Remark 1.3.9.
- The Vietoris-Rips complex is computationally less demanding than the Čech complex and therefore often preferred in applications.

- Other complexes such as witness complexes [40] or alpha complexes [35] are not defined here. They have been applied however in finding patterns in natural images [40].

Example 1.3.10. Suppose $Z = \{x_1, x_2, x_3\}$, as a subset of $\mathbb{R}^2$. For $\varepsilon > 0$ take $B_\varepsilon(x_i)$ for each x_i. That gives the following cover of Z :



To construct the Čech complex, we will take the nerve of $\mathscr{U} = \{B_\varepsilon(x_1), B_\varepsilon(x_2), B_\varepsilon(x_3)\}$. That means the vertex set is Z and as $U_i \cap U_j \neq 0$ for any $i, j = 1, 2, 3$, and $U_1 \cap U_2 \cap U_3 = 0$ we get that the nerve of $\mathscr{U}$ is a union of three 1-simplices :

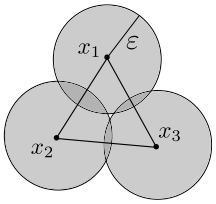


Figure 1.11: Čech complex of $Z = \{x_1, x_2, x_3\}$

To compute the Vietoris-Rips complex, the points of Z are again the vertices. Taking the cover $\mathscr{V} = \{V_1, V_2, V_3\}$ as in the following figure, we get that $d(x_i, x_j) \leq \varepsilon$ for all $i = 1, 2, 3$, and $\{x_1, x_2\}, \{x_1, x_3\}$ and $\{x_2, x_3\}$ span 1-simplices but this time as well $\{x_1, x_2, x_3\}$ spans a 2-simplex.

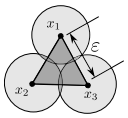


Figure 1.12: Vietoris-Rips complex of $Z = \{x_1, x_2, x_3\}$

Remark 1.3.11. We used the same scale in Fig. 1.11 and 1.12, to illustrate the difference between the two complexes, which is why Fig. 1.12 is smaller.

Level sets, sublevel sets and superlevel sets

In order to get insight into a topological space X, we use an associated real-valued function f on X and consider subsets of the space X.

Definition 1.3.12. Let X be a topological space and $f : X \to \mathbb{R}$ a real-valued function.

- The **level set** of X at a parameter $t \in \mathbb{R}$ is $f^{-1}(t)$ and denoted X^t.
- The **sublevel set** of X at a parameter $\alpha \in \mathbb{R}$ is $f^{-1}(-\infty, \alpha)$ and is denoted $X^{(-\infty,\alpha)}$.
- The **superlevel set** of X at a parameter $\alpha \in \mathbb{R}$ is $f^{-1}[\alpha, \infty)$ and is denoted $X^{[\alpha,\infty)}$.

Example 1.3.13. An illustration of these notion can be found in figure 1.13.

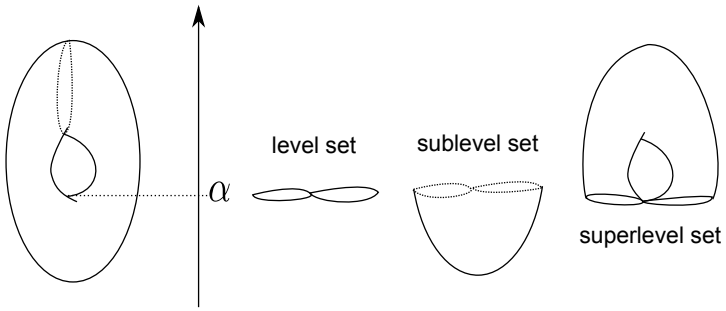


Figure 1.13: Illustration of a topological space, a torus, together with a real-valued function, the height function, and the level set, sublevel set, superlevel set at α

Filtration

We first need to introduce what it means for a simplicial complex to be nested in another simplicial complex.

Definition 1.3.14. Let K be a simplicial complex. A subcollection L of simplices from K is called a **subcomplex** of K if L also forms a simplicial complex. Therefore, if a simplex σ is in L, then all of its faces in K are also present in L. The vertex set of L can be smaller than that of K.

Filtrations built out of a point cloud

The families of Vietoris-Rips complexes $(Rips_\varepsilon(X))_{\varepsilon\geq 0}$ and Čech complexes $(\check{C}ech_\varepsilon(X))_{\varepsilon\geq 0}$ (see section 1.3.3) define interesting filtrations on point clouds. One starts with the point cloud when ε is equal to 0 (except if one allows the distance to be 0 between different points) and with increasing ε simplices are progressively added, until the complex with all the points linked and all the possible simplices added is obtained.

Filtration by sublevel sets

Let K be a simplicial complex and V its vertex set. Moreover, let $f : V \to \mathbb{R}$ be a function defined on the vertices. This function can be extended to all simplices of K by $f([v_0, \ldots, v_k]) = \max_{i=1,\ldots,k} f(v_i)$ for any simplex $\sigma = [v_0, \ldots, v_k]$. The sublevel set filtration of f is defined by $K_r = \{\sigma \in K | f(\sigma) \leq r\}$. The superlevel set filtration of f is defined in section 1.3.4.

Remark 1.3.15. In practice, all the filtrations considered are built on finite sets and are indeed finite, whence even if the index set is infinite (ε and r), the Rips or Čech filtration will only change at a finite number of values of ε, and the same holds for the sublevel set filtration K_r which only changes for a finite number of values of r. They are therefore easy to handle from an algorithmic point of view.

Persistence modules

Persistence modules are explained in a complete and precise way in [39] and [8] and with a more algebraic/categorical point of view in [52].

Definition 1.3.16. Let T be a subset of the real numbers $\mathbb{R}$. A **persistence module** $\mathbb{V}$ over T is an indexed family of vector spaces $(V_r \mid r \in T)$ together with linear maps $\nu_r^s : V_r \to V_s$,

for every $r \leq s$ which satisfy the composition law $\nu_s^t \circ \nu_r^s = \nu_r^t$ whenever $r \leq s \leq t$, and where ν_r^r is the identity map on V_r.

Remark 1.3.17. Let K be a simplicial complex and let K_r, $r \in \mathbb{R}$ be a filtration of K (as defined for instance through sublevel sets or through complexes). For every $r < r' \in \mathbb{R}$, the elements of the filtration K_r and $K_{r'}$ verify $K_r \subseteq K_{r'}$. This inclusion, denoted δ_r^s, verifies that $\delta_s^t \circ \delta_r^s = \delta_r^t$ for every $r < s < t$. This inclusion induces linear maps between the homology vector spaces $\bar{\delta}_r^s : H_k(K_r) \to H_k(K_s)$, also verifying $\bar{\delta}_s^t \circ \bar{\delta}_r^s = \bar{\delta}_r^t$ for every $r < s < t$.
Therefore, the homology groups on elements of a filtration, together with the maps $\bar{\delta}$ define a persistence module, which is the object of study of persistent homology.

Barcodes and Persistence diagram
Recall the formal definition of an interval, which corresponds to our intuition : An interval in $(\mathbb{R}, \leq)$ is a subset $I \subseteq \mathbb{R}$ such that if $i, k \in I$ and if there is a $j \in \mathbb{R}$ such that $i \leq j \leq k$, then $j \in I$.

Definition 1.3.18. Let I be an interval in $\mathbb{R}$. A special persistence module is the **interval module** k_I defined by assigning to each element $i \in I$ the vector space k and the zero vector space to elements in $\mathbb{R} \setminus I$. Moreover, if $i, j \in I$ and $i \leq j$, the maps ν_i^j are the identity map and otherwise are the zero map.

Since interval modules are defined completely by the interval where non 0 spaces appear one can represent interval modules as bars. They are denoted $\mathbb{I}(b, d)$, where b is the infimum and d the supremum of the interval.
In many cases, a persistence module $\mathbb{V}$ can be decomposed into a direct sum of intervals modules, in which case it can be shown that the decomposition is unique up to reordering of the intervals (see [53]). Therefore, one summarises the persistence module as a collection of bars (and the order does not matter) called the **persistence barcode** of $\mathbb{V}$. Another representation called the **persistence diagram** represents each interval, $\mathbb{I}(b, d)$, by a point (b, d) in $\mathbb{R}^2$, corresponding to topological feature to which the diagonal $\Delta = \{(x, x) \mid x \in \mathbb{R}\}$ is added with infinite multiplicity. Points that are close to the diagonal often represent noise in the data, as illustrated in example 1.14.
Let us see in which cases a decomposition of a persistence module as a direct sum of intervals exists.

Decomposition of persistence modules
We highlight here two situations in which persistence modules are decomposable. More details can be found in [37], [39], [35].

Theorem 1.3.19. *(Can be found in [37], [53]) Let $\mathbb{V}$ be a persistence module indexed by a subset T of $\mathbb{R}$. If T is a finite set or if all the vector spaces V_r are finite-dimensional and T is locally finite, then $\mathbb{V}$ is decomposable as a direct sum of interval modules.*

The persistent homology of filtrations of finite simplicial complexes verifies the above conditions. Therefore, the persistence diagrams of such filtrations are always well-defined.

Definition 1.3.20. A persistence module $\mathbb{V}$ indexed by $T \subseteq \mathbb{R}$ is **q-tame** if for any $r < s$ in T, the rank of the linear map $\nu_r^s : V_r \to V_s$ is finite.

Theorem 1.3.21. *(Can be found in [54], [53]). If $\mathbb{V}$ is a q-tame persistence module, then $\mathbb{V}$ has a well-defined persistence diagram.*

Examples 1.3.22. We show three examples of persistence diagrams corresponding to the three most common types: one using a function $f : [0, 1] \to \mathbb{R}$ and sublevelsets (Fig. 1.14), one using a height function on a topological space (Fig. 1.15) and the last one after creating a filtration using the Rips complex on a point cloud (Fig. 1.16).

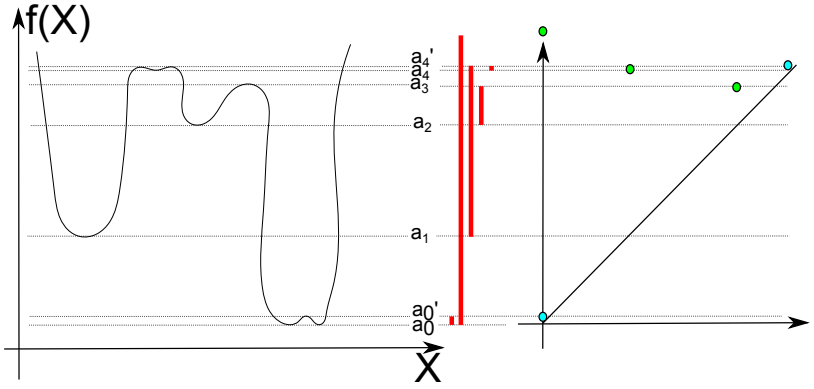


Figure 1.14: The persistence diagram (on the right) of a function $f : X \to \mathbb{R}$ (on the left) together with its barcode (in the middle). Blue points in the diagram represent noisy feature as they are close to the diagonal, whereas green points represent relevant features. As $H_q = 0$ for $q \geq 1$, only H_0 points are displayed.

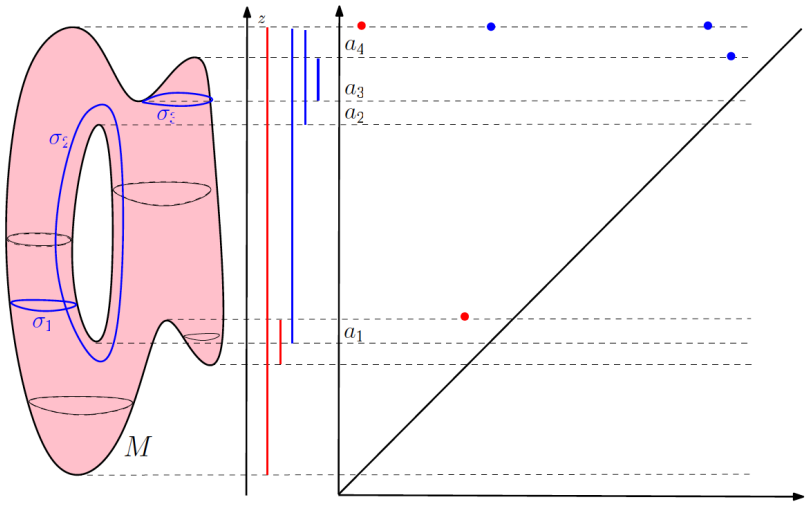


Figure 1.15: The persistence diagram of the height function on a topological space, H_0 points/bars in red, H_1 points/bars in blue. (Illustration from [12])

1.3.4 Extended persistent homology

This subsection was written in collaboration with M. Carrière and was inspired by [23].
Let X be a topological space and $f : X \to \mathbb{R}$ a real-valued function on X. The sublevel sets $\{X^{(-\infty,\alpha]}\}_{\alpha\in\mathbb{R}}$ of f define a filtration, i.e. $X^{(-\infty,\alpha]} \subseteq X^{(-\infty,\beta]}$ for all $\alpha \leq \beta \in \mathbb{R}$. The superlevel sets $\{X^{[\alpha,+\infty)}\}_{\alpha\in\mathbb{R}}$ of f are nested as well, though in the opposite direction:

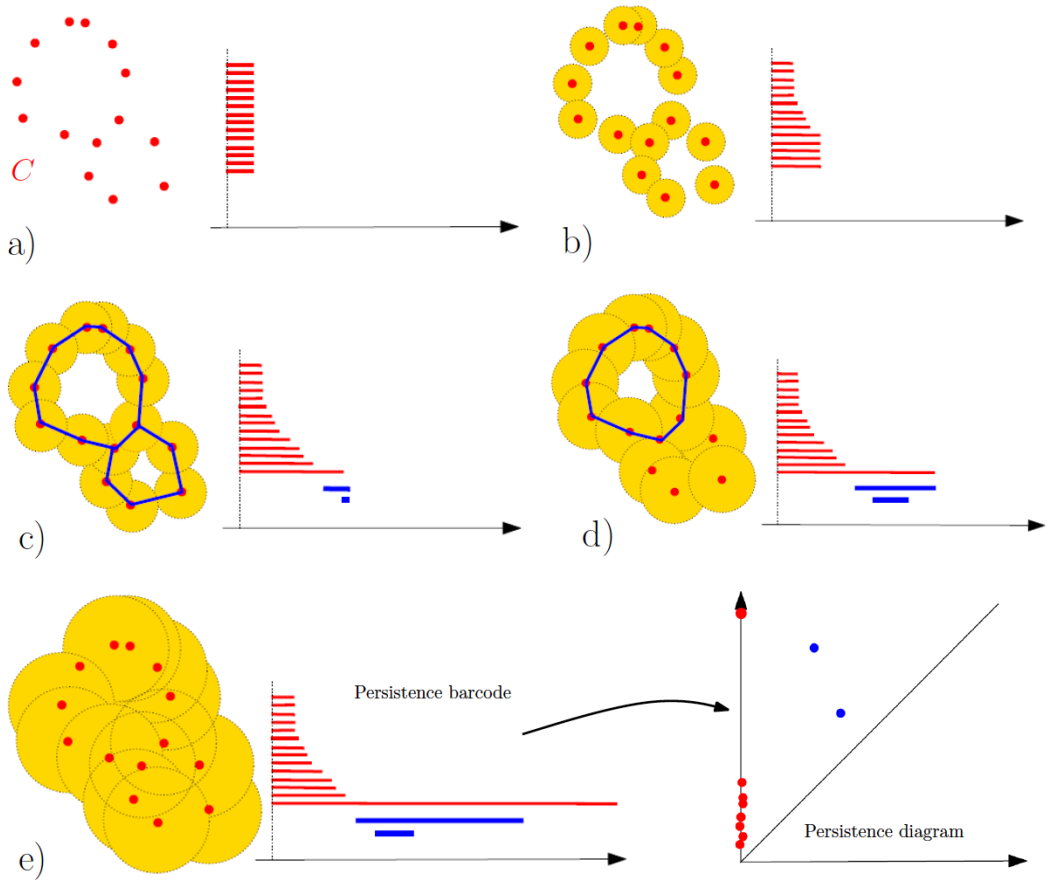


Figure 1.16: The persistence diagram of a filtration arising from a Rips complex with growing ε, H_0 points/bars in red, H_1 points/bars in blue. (Illustration from [12])

$X^{[\alpha,+\infty)} \supseteq X^{[\beta,+\infty)}$ for all $\alpha \leq \beta \in \mathbb{R}$. By reversing the real line, this nested family can be turned into a filtration. Indeed, let $R^{\text{op}} = \{\tilde{x} \mid x \in \mathbb{R}\}$, ordered by $\tilde{x} \leq \tilde{y} \Leftrightarrow x \geq y$. By indexing the family of superlevel sets by R^{op}, we obtain a filtration: $\{X^{[\tilde{\alpha},+\infty)}\}_{\tilde{\alpha} \in R^{\text{op}}}$, with $X^{[\tilde{\alpha},+\infty)} \subseteq X^{[\tilde{\beta},+\infty)}$ for all $\tilde{\alpha} \leq \tilde{\beta} \in R^{\text{op}}$.

Combining the two filtrations at infinity defines extended persistence: each superlevel set $X^{[\tilde{\alpha},+\infty)}$ is replaced by the pair of spaces $(X, X^{[\tilde{\alpha},+\infty)})$ in the second filtration. The filtration property is maintained since $(X, X^{[\tilde{\alpha},+\infty)}) \subseteq (X, X^{[\tilde{\beta},+\infty)})$ for all $\tilde{\alpha} \leq \tilde{\beta} \in R^{\text{op}}$. Then, let $\mathbb{R}_{\text{Ext}} = \mathbb{R} \cup \{+\infty\} \cup R^{\text{op}}$, where the order is completed by $\alpha < +\infty < \tilde{\beta}$ for all $\alpha \in \mathbb{R}$ and $\tilde{\beta} \in R^{\text{op}}$. One can show that there is an isomorphism between this poset and $(\mathbb{R}, \leq)$. The **extended filtration** of f over $\mathbb{R}_{\text{Ext}}$ is defined as follows:

$$\begin{aligned} F_\alpha &= X^{(-\infty,\alpha]} && \text{for } \alpha \in \mathbb{R} \\ F_{+\infty} &= X \equiv (X, \emptyset) \\ F_{\tilde{\alpha}} &= (X, X^{[\tilde{\alpha},+\infty)}) && \text{for } \tilde{\alpha} \in R^{\text{op}}, \end{aligned}$$

where the space X has been identified with the pair of spaces $(X, \emptyset)$. The filtration is well-defined since we have $X^{(-\infty,\alpha]} \subseteq X \equiv (X, \emptyset) \subseteq (X, X^{[\tilde{\beta},+\infty)})$ for all $\alpha \in \mathbb{R}$ and $\tilde{\beta} \in R^{\text{op}}$. The

ordinary part of the filtration is the subfamily $\{F_\alpha\}_{\alpha\in\mathbb{R}}$, and the subfamily $\{F_{\tilde{\alpha}}\}_{\tilde{\alpha}\in R^{\mathrm{op}}}$ is called the **relative** part. See Fig. 1.17 for an illustration.

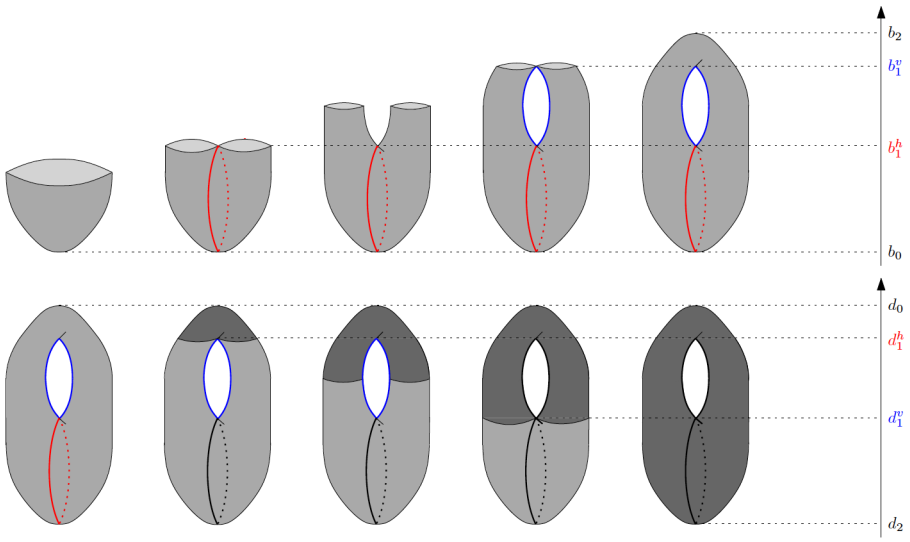


Figure 1.17: The extended filtration of the height function on a topological space, the torus. The ordinary part of the filtration is displayed on the upper row, while the lower row shows the relative part (Taken from [23]).

The homology of the elements of this filtration is then computed, giving rise to the **extended persistence module** $\mathbb{V}$ of f:

$$\begin{aligned} V_\alpha &= H_*(F_\alpha) = H_*(X^{(-\infty,\alpha]}) && \text{for } \alpha \in \mathbb{R} \\ V_{+\infty} &= H_*(F_{+\infty}) = H_*(X) \cong H_*(X,\emptyset) \\ V_{\tilde{\alpha}} &= H_*(F_{\tilde{\alpha}}) = H_*(X, X^{[\tilde{\alpha},+\infty)}) && \text{for } \tilde{\alpha} \in R^{\mathrm{op}}, \end{aligned}$$

where the linear maps between the spaces are induced by the inclusions in the extended filtration and where $H_*(X,A)$ represents the relative homology (see section 1.3.2).
The question of existence and uniqueness of a decomposition into interval modules in the extended case requires the definition of Morse-type functions for which such a decomposition exists.

Morse-type functions

Intuition 1.3.23. **Morse-type** functions are generalizations of the classical Morse functions. Some properties are shared, but the function is not required to be differentiable nor defined over a smooth manifold. Morse function are used to study particular points on a manifold, such as for instance maxima, minima and saddle points, whereas Morse-type functions are a generalisation of those concepts to any topological space with the particularity that the points studied segregate the space into distinct parts that can be studied almost independently.

Definition 1.3.24. Let f be a continuous real-valued function on a topological space X. The function f is **of Morse-type** if the following conditions hold.

(i) There exists real numbers $a_1 < \cdots < a_n \in \mathbb{R}$, with $n \in \mathbb{N}$, called the **critical values**,

such that over every open interval $(a_0 = -\infty, a_1), ..., (a_i, a_{i+1}), ..., (a_n, a_{n+1} = +\infty)$ there is a compact and locally connected space Y_i and a homeomorphism $\mu_i : Y_i \times (a_i, a_{i+1}) \to X^{(a_i,a_{i+1})}$ such that $\forall i = 0, ..., n, f|_{X^{(a_i,a_{i+1})}} = \pi_2 \circ \mu_i^{-1}$, where π_2 is the projection onto the second factor;

(ii) For all $i = 1, ..., n-1$, the function μ_i extends to a continuous function $\bar{\mu}_i : Y_i \times [a_i, a_{i+1}] \to X^{[a_i,a_{i+1}]}$ – similarly μ_0 extends to $\bar{\mu}_0 : Y_0 \times (-\infty, a_1] \to X^{(-\infty,a_1]}$ and μ_n extends to $\bar{\mu}_n : Y_n \times [a_n, +\infty) \to X^{[a_n,+\infty)}$;

(iii) Each level set X^t has a finitely-generated homology.

Remark 1.3.25. Morse functions are known to be of Morse-type, therefore we can say that they generalise Morse functions [55], while the converse is not true. Indeed, as previously mentioned, Morse-type functions are not required to be differentiable, and their domain does not have to be a smooth manifold nor even a manifold at all. It is also possible to find Morse-type functions on manifolds that are not Morse, such as the Gaussian curvature on a torus.

Example 1.3.26. In Fig. 1.18, we illustrate a Morse-type function, the height function of a topological space (on the right). On the left, we illustrate the decomposition of each $X^{(a_i,a_{i+1})}$ into a homeomorphic space $Y_i \times (a_i, a_{i+1})$, $i = 0, \ldots, 3$.

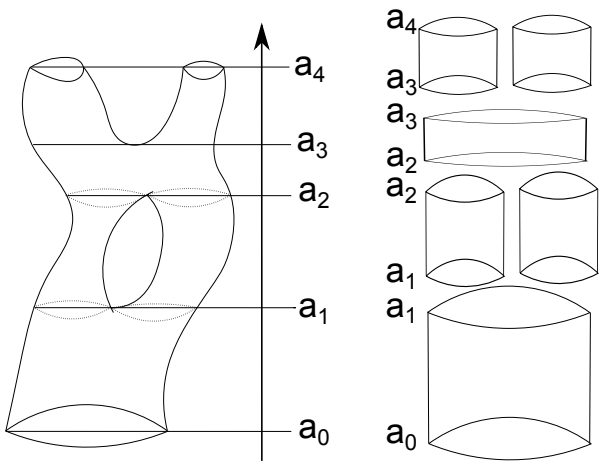


Figure 1.18: A topological space together with a Morse-type function (the height function) and the decomposition into spaces $Y_i \times (a_i, a_{i+1})$ homeomorphic to $X^{(a_i,a_{i+1})}$ explaining why it is a Morse-type function.

The function f is asked to be of Morse-type, since in that case the Reeb graph is a multigraph [56], whose nodes are in one-to-one correspondence with the connected components of the critical level sets of f. This multigraph can then be equipped with a metric by assigning the length $l(v_i, v_j) = |f(v_i) - f(v_j)|$ to each edge (v_i, v_j).

Theorem 1.3.27. *(Can be found in Chazal et al. [53]) The extended persistence module of a Morse-type function can be decomposed as a finite direct sum of half-open* ***interval modules****:*

$$\mathbb{V} \simeq \bigoplus_{k=1}^{n} \mathbb{I}[b_k, d_k),$$

where each summand $\mathbb{I}[b_k, d_k)$ is made of copies of the field of coefficients at each index $\alpha \in [b_k, d_k)$, and of copies of the zero space elsewhere, the maps between copies of the field being identities.

The lifespans of **homological features** (for example the connected components, the holes, the voids, etc.) within the filtration are represented by the summands. More precisely, the **birth time** b_k and **death time** d_k of the feature are given by the endpoints of the interval. There exists a well-defined analogue to the persistence diagrams (see section 1.3.3) called the **extended persistence diagram** of f, denoted $\mathrm{Dg}\,(f)$. Endpoints of each of the intervals in the decomposition are plotted as coordinates in $\mathbb{R}^2$. The distinction between ordinary and relative parts of the filtration leads to a classification of the points in $\mathrm{Dg}\,(f)$ in the following way.

- Points whose coordinates both belong to $\mathbb{R}$ are called **ordinary points**; they correspond to homological features being born and then dying in the ordinary part of the filtration, denoted $\mathrm{Ord}(f)$.
- Points whose coordinates both belong to R^{op} are called **relative points**; they correspond to homological features being born and then dying in the relative part of the filtration, denoted $\mathrm{Rel}(f)$
- Points whose abscissa belongs to $\mathbb{R}$ and whose ordinate belongs to R^{op} are called **extended points**; they correspond to homological features being born in the ordinary part and then dying in the relative part of the filtration, denoted $\mathrm{Ext}(f)$.

The extended persistence diagram is illustrated in the following picture (Fig. 1.19).

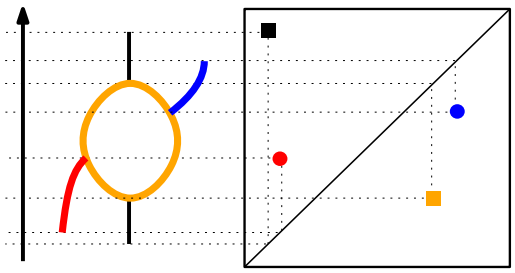


Figure 1.19: Illustration of an example of correspondences between topological features of a graph (branches, holes, components) and points in its corresponding extended persistence diagram. The ordinary persistence is unable to detect the blue upwards branch.

Ordinary points lie strictly above the diagonal $\Delta = \{(x, x) \mid x \in \mathbb{R}\}$ and relative points lie strictly below Δ, while extended points can be located anywhere, including on Δ, e.g. connected components that lie inside a single critical level – see section 1.2.6. It is common to decompose $\mathrm{Dg}\,(f)$ according to this classification:

$$\mathrm{Dg}\,(f) = \mathrm{Ord}(f) \sqcup \mathrm{Rel}(f) \sqcup \mathrm{Ext}^+(f) \sqcup \mathrm{Ext}^-(f),$$

where by convention $\mathrm{Ext}^+(f)$ are extended points above the diagonal and includes the extended points located on the diagonal Δ.

Comparing extended persistence diagrams

The definition of a distance between extended persistence diagrams is based on partial matchings.

Definition 1.3.28. Given two persistence diagrams D, D', we define a **partial matching** between D and D' as a subset Γ of $D \times D'$ such that:

$$\forall p \in D, \text{ there is at most one } p' \in D' \text{ such that } (p, p') \in \Gamma,$$

$$\forall p' \in D', \text{ there is at most one } p \in D \text{ such that } (p, p') \in \Gamma.$$

Γ is required to match points of the same type (ordinary, relative, extended) and of the same homological dimension *only*. In other words, a partial matching is an injective function from a subset of D to a subset of D' respecting certain type-matching conditions.

One defines a **cost of a matching** Γ which defines how good the matching is (i.e. the closest possible points have been matched).

Definition 1.3.29. The **cost** of Γ is:

$$\text{cost}(\Gamma) = \max\left\{\max_{p \in D} \ \delta_D(p), \ \max_{p' \in D'} \ \delta_{D'}(p')\right\},$$

where

$$\delta_D(p) = \|p - p'\|_\infty \text{ if } \exists p' \in D' \text{ s.t. } (p, p') \in \Gamma \text{ and } \delta_D(p) = d_\infty(p, \Delta) = \inf_{q \in \Delta} \|p - q\|_\infty \text{ otherwise,}$$

$$\delta_{D'}(p') = \|p - p'\|_\infty \text{ if } \exists p \in D \text{ s.t. } (p, p') \in \Gamma \text{ and } \delta_{D'}(p') = d_\infty(p', \Delta) = \inf_{q \in \Delta} \|p' - q\|_\infty \text{ otherwise.}$$

Now, the distance between two persistence diagrams is given by the matching with the smallest possible cost.

Definition 1.3.30. Let D, D' be two extended persistence diagrams. The **bottleneck distance** between D and D' is:

$$d_b^\infty(D, D') = \inf_{\Gamma} \ cost(\Gamma),$$

where Γ ranges over all partial matchings between D and D'.

The bottleneck distance d_{b}^∞ is only a pseudo-metric, not a metric, because points lying on Δ can be left unmatched at no cost.

Stability

One can prove that extended persistence diagrams of Morse-type functions with respect to the bottleneck distance are stable.

Theorem 1.3.31 (Stability [57])**.** *For any Morse-type functions $f, g : X \to \mathbb{R}$,*

$$d_{\text{b}}^\infty(\text{Dg}(f), \text{Dg}(g)) \leq \|f - g\|_\infty = sup_{x \in X} \mid f(x) - g(x) \mid .$$

Moreover, as pointed out in [57], the theorem can be strengthened to apply to each subdiagram Ord, Ext^+, Ext^-, Rel and to each homological dimension individually.

Reeb Graphs and the extended persistence of $\tilde{f}$.
Recall that in the definition of the Reeb graphs of a function f, there is an induced map $\tilde{f} : \mathrm{R}_f(X) \to \mathbb{R}$ such that $f = \tilde{f} \circ \pi$, where π is the quotient map $X \to \mathrm{R}_f(X)$.
One has an easy interpretation of $\mathrm{Dg}\,(\tilde{f})$ in terms of the structure of $\mathrm{R}_f(X)$. We refer the reader to [58] and the references therein for a full description as well as formal definitions and statements.
If the Reeb graph is oriented vertically so that $\tilde{f}$ is the height function, each connected component of the graph can be seen as a trunk with multiple branches (some oriented upwards, others oriented downwards) and holes. Then, one has the following correspondences (Fig. 1.19), where the **vertical span** of a feature is the span of its image by $\tilde{f}$.

- The vertical spans of the trunks are the points in $\mathrm{Ext}_0^+(\tilde{f})$.
- The vertical spans of the branches that are oriented downwards are the points in $\mathrm{Ord}_0(\tilde{f})$.
- The vertical spans of the branches that are oriented upwards are the points in $\mathrm{Rel}_1(\tilde{f})$.
- The vertical spans of the holes are the points in $\mathrm{Ext}_1^-(\tilde{f})$.

The rest of the diagram of $\tilde{f}$ is empty. These correspondences provide a dictionary to read off the structure of the Reeb graph from the persistence diagram of the induced map $\tilde{f}$. Note that it is a bag-of-features type descriptor, taking an inventory of all the features (trunks, branches, holes) together with their vertical spans, but leaving aside the actual layout of the features. As a consequence, it is an incomplete descriptor: two Reeb graphs with the same persistence diagram may not be isomorphic, as illustrated in Fig. 1.20.

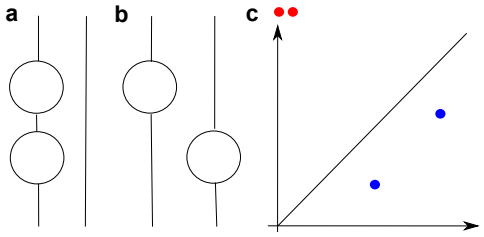


Figure 1.20: **(a)**, **(b)** Two Reeb graphs with the same set of features but not the same layout, having the same **(c)** persistence diagram.

This connection can be rephrased in terms of the extended persistence of f and of its induced map $\tilde{f}$, following the intuition that the 1-dimensional persistence diagram of $\tilde{f}$ is a subset of that of f, and that the missing points correspond either to the inessential or horizontal 1-dimensional homology generators, or to the homology generators in dimension 2 and above:

Theorem 1.3.32. *(Proof can be found in [23]) Let X be a topological space and $f : X \to \mathbb{R}$ a function of Morse-type. Then,* $\mathrm{Dg}\,(\tilde{f}) \subseteq \mathrm{Dg}\,(f)$. *Furthermore:*

$$\mathrm{Dg}_0(\tilde{f}) = \mathrm{Dg}_0(f)$$
$$\mathrm{Dg}_1(\tilde{f}) = \mathrm{Dg}_1(f) \setminus (\mathrm{Ext}_1^+(f) \cup \mathrm{Ord}_1(f))$$
$$\mathrm{Dg}_p(\tilde{f}) = \emptyset \text{ if } p \geq 2$$

1.4 Introduction to sequencing methods

1.4.1 The early days of sequencing

In each cell of the body lies the genetic material inherited by our ancestors, called the **deoxyribonucleic acid** (DNA) . The DNA is made out of a chain of **nucleotides**, which are the molecules that constitute the alphabet of the DNA and are characterized by three distinctive chemical sub-units: a five-carbon sugar molecule, a nitrogenous base (which can be Adenosine (A), Guanosine (G), Thymine (T), Cytosine (C)) and one phosphate group. To understand the role of each cell at a certain moment in time, looking at the DNA, and more specifically at its sub-entities called **genes**, is not sufficient as it is similar across the different cell types. Each of these genes, constituted of coding parts, **exons**, and non-coding parts, **introns**, exerts a specific role in a tissue. To know which of these genes are currently transcribed from DNA to ribonucleic acid (RNA) which is the macromolecule that conveys the genetic information, one measures the expression of **messenger RNA** (mRNA), which is an RNA molecule essential in transcription. mRNA changes tremendously from cell to cell or organism to organism and provides insights into what proteins are soon to be made. This step is called transcription. The proteins that are accessible to the cell can be measured (translation) and represents therefore what the fate of the latter might be. Hence three tiers are available to a biologist. First, DNA sequencing (the genomic tier) reflecting somatic mutations, i.e. mutations in subpopulations of cells of the body, happening for instance in cancer cells, or germinal mutations, i.e. mutations inherited by our ancestors. Then, a second level with mRNA sequencing to measure current or upcoming activity of the cells (transcriptome tier). Finally, protein sequencing measuring currently available proteins to the cell. In 2017, DNA sequencing celebrated its 40th anniversary [59]. Due to their much smaller sequence, proteins and mRNA were sequenced a decade before DNA [59]. This was achieved in the 1950s by sequencing the protein sequence of insulin [60], followed by many proteins afterwards, revealing that each protein has a different sequence and that this might vary from species to species and even from individual to individual. In the early 1960, the same principle than for sequencing proteins was used to sequence mRNA, i.e. fragments of RNA obtained by using enzymes that cut the RNA were separated by chromatography and electrophoresis. Fragments were deciphered and the sequence deduced. Of note, five people worked during three years with one gram of pure material, isolated from 140kg of yeast, to determine the 76 nucleotides of the first RNA sequence, of alanine transfer RNA [61].

Through the years, several strategies for sequencing mRNA have been developed, and three of them are discussed in the following sections, ordered by when they first emerged : **microarray** (section 1.4.2), **RNA-seq** (section 1.4.3) and **single-cell RNA-seq** (section 1.4.4).

1.4.2 Microarray

Microarray is the oldest sequencing technique of the three and is based on light emission. This micro-technological process is looking for complementary targets, i.e. the nucleotides A, C, T and G are complementary targets of T, G, A and C respectively [62] and when the complementary sequence and the target sequence bind they will form a hybrid which emits light. That is the concept behind microarray technology. Single stranded DNA (ssDNA) fragments formed by synthetic oligonucleotides, which are synthetic sequences of letters A, C, T and G, are fixed on the surface of a microarray chip. This DNA microarray chip is a grid

on a substrate (Fig. 1.21a), which is usually made out of glass or silicon, and each element of the grid contains thousands of copies of identical probes of ssDNA fragments that are waiting to link to their complementary sequence. These ssDNA fragments correspond to one gene per grid element (Fig. 1.21b). On one chip thousands of genes can be assessed simultaneously. During sample preparation complementary DNA (cDNA) , the complementary sequence of the mRNA, is obtained from the sample of interest and is fluorescently labelled. Therefore, if a part of the grid emits light it ensures that mRNA of this gene was present in the sample. The output is a picture of more or less bright spots (Fig. 1.21c) and can be transformed using the software **R** into a matrix (Fig. 1.21d), where the rows correspond to a gene, the columns to the different samples assessed and the value is given by the intensity observed for that gene in $\log_2$-scale.

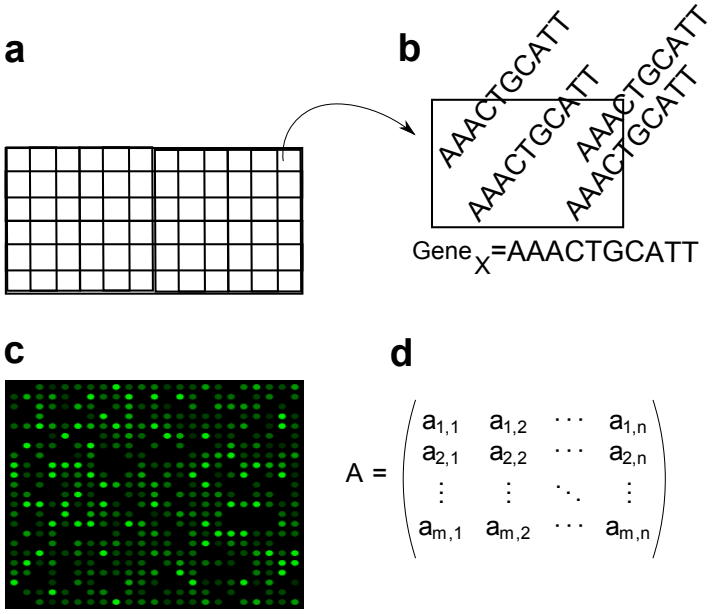


Figure 1.21: Scheme of microarray process **(a)** A schematic DNA microarray chip which is a grid **(b)** each grid element contains the sequence of the same gene. **(c)** the image of intensities produced by the microarray procedure, **(d)** translation of the .CEL files, which is the file type obtained from the sequencing step, into a matrix of values of intensity in $\log_2$ scale. The rows represent the genes and the columns represent the samples.

The brighter one spot on the picture the more the gene concerned in that spot is expressed. The values for each gene are given in $\log_2$-scale, and was shown to be normally distributed [63]. This microarray method is robust and a consortium decided on a standardised way to exchange and communicate microarray data in order to easily reproduce the results (Minimum information about a microarray experiment, MIAME) [64].

1.4.3 RNA-seq

RNA-seq emerged recently, and is more precise as it works with providing the actual nucleotides, which is the alphabet constituting the DNA, of the sequenced fragment. After isolating the mRNA from the target sample, a fragmentation step splits the mRNA into smaller pieces (Fig. 1.22a), called **reads** [65]. This step is necessary since the sequencing machines, called also

sequencer, can only read a limited length of fragments that depends only on the engine used and chosen parameters and which is shorter than most RNA transcripts. RNA fragments are then converted into double stranded DNA as this is more stable and enables easy amplification. Adapters, which are small sequences mostly repeats of the same letters, are added to the reads in order for the machine to recognize the fragments. This chemical process is not completely reliable, and some sequences might be lost as they did not bind to an adapter (Fig. 1.22b). This would also allow the user to sequence different samples at the same time since different adapters can be used for each sample. The fragments that have adapters are amplified through a process called **polymerase chain reaction** (PCR) in order to enrich for those fragments. After a quality control check, the DNA fragments are linked to a grid vertically. The sequencer processes several millions reads simultaneously, nucleotide per nucleotide (Fig. 1.22c, example for 4 reads). It uses fluorescently labelled nucleotides, called **probes**, that can bind to their corresponding nucleotide. Each probe A, C, T or G has a different color and binds to the first base of each of the reads. A picture is taken and the sequencer translates each base according to the observed color into A, C, T or G (Fig. 1.22c). The colors are washed, and the process starts again for the second nucleotide until the fragment is sequenced. The picture might not be clear for all the million spots as there might be brighter spots that the machine assesses with less certainty as to which base correspond to that site (Fig. 1.22d, light blue spot). Therefore each base receives a quality score to understand the certainty of the called nucleotide.

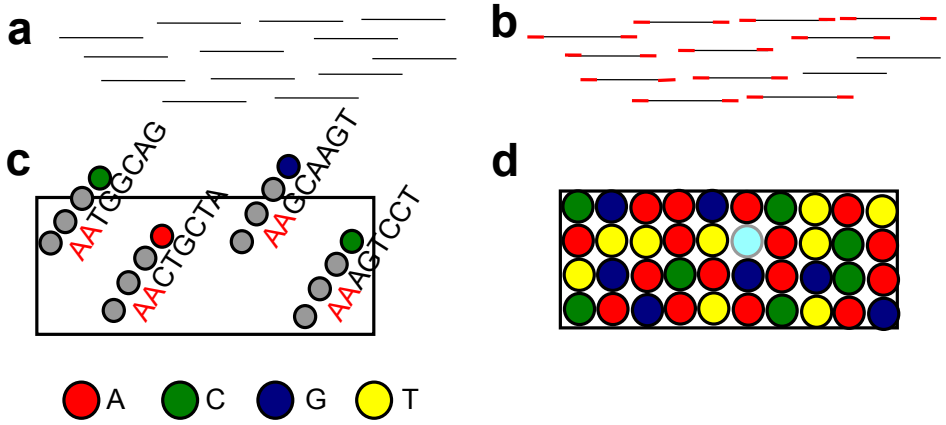


Figure 1.22: Scheme of RNA-seq analysis from RNA to the sequences of the RNA. **(a)** RNA is fragmented and converted into double stranded DNA **(b)** adapters are added (red), and the fragments are PCR amplified, **(c)** fluorescently labelled probes linking to the corresponding nucleotides enables the sequencer to translate each base into A, C, T or G. **(d)** A possible problem that arises with a spot that is less bright.

These fragments, or **reads**, can then be mapped to an existing reference genome, such as the human genome GRCh38 or a generated genome (Fig. 1.23a) [66]. They are allowed to map with mismatches between fragments and reference genome that can then reveal somatic mutations, consistent mismatches in a certain nucleotide compared to a reference genome (Fig. 1.23b). This step is achieved in an operating system such as UNIX using tools such as tophat [67], bowtie [68], hisat [69], samtools [70] among others. A count of abundance of fragments on a gene locus, consisting of the sum of the number of counted reads per exon, which is the

part of the gene that is conserved after transcription, of the gene, reveals the expression of that gene (Fig. 1.23c). Cufflinks [71], Picard [72], Kallisto [73], featureCounts [74] and others perform this step in UNIX. After this, using the software **R** or online tools such as ASAP [75], gene expressions are filtered and normalised. They can be visualised using PCA plots, segregate using hierarchical clustering and heatmaps and subgroups can be discovered by visual inspection or using specific algorithms (Fig. 1.23d). Genes are then interrogated on how they differ in subgroups, called **differential expession analysis** that are either known previous to the analysis or found using the previous step. In **R**, this can be achieved using voom [76], edgeR [77], DESeq2 [78], limma [79] or using online tools such as ASAP [75] and is visualised with Venn diagrams (Fig. 1.23e). Analysis of the differentially expressed genes is performed in order to obtain a global view on the changes using pathway analysis tools in R such as : GSEA [80], topGO [81], topAnat [82] or using online tools such as ASAP [75], GO [83], Panther [84], STRING [85] among others (Fig. 1.23f).

This sequencing technique is more precise and reveals more insights than microarray, such as somatic mutation, variants and isoforms of genes and many others [86]. Similar techniques are used to analyse RNA-seq and microarray data, even though as one might expect RNA-seq data is mainly discrete as a total number of fragments per gene is counted and therefore is usually an integer, whereas microarray is considered continuous. Techniques have been developed to normalise RNA-seq data in such a way that it resembles microarray data [87]. However, the various normalisations provide different lists of significant genes and understanding the right normalisation to use is troublesome. Moreover, numerous methods to perform differential expression analysis exist and have distinct abilities to find true positives in different situations, depending on the sample size, the number of 0s or missing values, and first attempts to decipher where a certain method should be preferred are just emerging [88].

1.4.4 Single-cell RNA-seq

In the last years, single-cell sequencing has been popularised. This method is similar to RNA-seq but gives the information of the expression of genes per cell. Changes from cell to cell can therefore be observed, e.g. a neuronal cell can be compared to a cell from the skin, and their different role questioned. This has the drawback that the dataset one obtains is extremely sparse since at a given time point in a given cell most genes are not expressed. In a data set [89], one cell displayed 87% of the genes without expression (0 value) and most of the cells have more than 60 % of the genes with no expression (Fig. 1.24a). This is rendering analysis difficult as standard RNA-seq analyses (Fig.1.24b) make the assumption that there is no such bias towards one value [88]. Several new techniques have been developed to analyse single-cell RNA-seq, among which ASAP [75] and [90], but the field is still evolving and new statistical methods developed for this type of analyses. Existing methods are not satisfactory yet in terms of reproducibility and robustness [91].

In conclusion, microarray is slowely being replaced by RNA-seq as the information provided by the latter is deeper. Single-cell RNA-seq is the newest technique and in comparison to RNA-seq provides gene expression information at the cell level rather than the tissue or mixture of cells level.

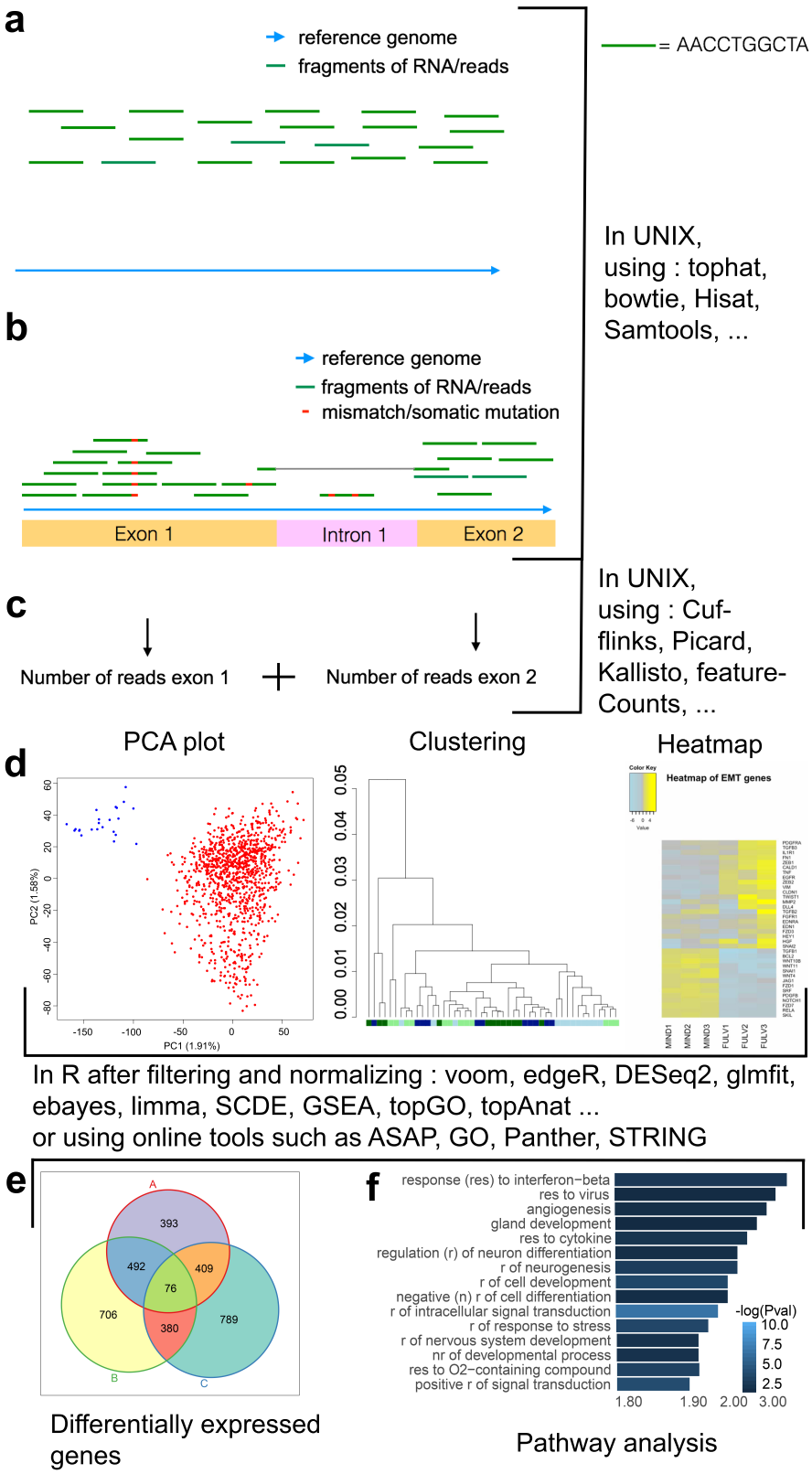


Figure 1.23: Scheme of RNA-seq analysis from the results obtained by the sequencer to pathway analysis (*continued on next page*).

Figure 1.23: Scheme of RNA-seq analysis from the results obtained by the sequencer to pathway analysis. **(a)** Per sample, the sequencer provides a FASTQ-type of file comprising the information per fragments or reads of different letters A, C, T or G found and a quality score linked to each base. These are aligned to a reference genome that can either be downloaded or generated **(b)** fragments are mapped to the reference genome with mismatches (as in exon1 and intron1, red), reflecting somatic mutations (exon1, red). Splicings over exon junctions, i.e. one part of the fragment is on one exon the other on the following exon, are possible (grey) this step is done in UNIX using tools such as tophat, bowtie, hisat, samtools, etc. **(c)** the counting of fragments mapped to a region is achieved by adding up the number of read counted per exon in UNIX using Cufflinks, Picard, Kallisto, featureCounts among others. **(d)** In **R** or using online tools such as ASAP, after filtering and normalizing one can visualise the data as a PCA plot (left), segregate the data using hierarchical clustering (middle) and heatmaps (right), subgroups can be discovered by visual inspection or using grouping algorithms (e.g. red and blue points in PCA plot) **(e)** Differential expression is performed in R to compare different groups (obtained for instance in **(d)**) using: voom, edgeR, DESeq2, limma or using online tools such as ASAP and is visualised with Venn diagrams **(f)** Analysis of the differentially expressed genes using pathway analysis tools in R such as : GSEA, topGO, topAnat or using online tools such as ASAP, GO, Panther, STRING.

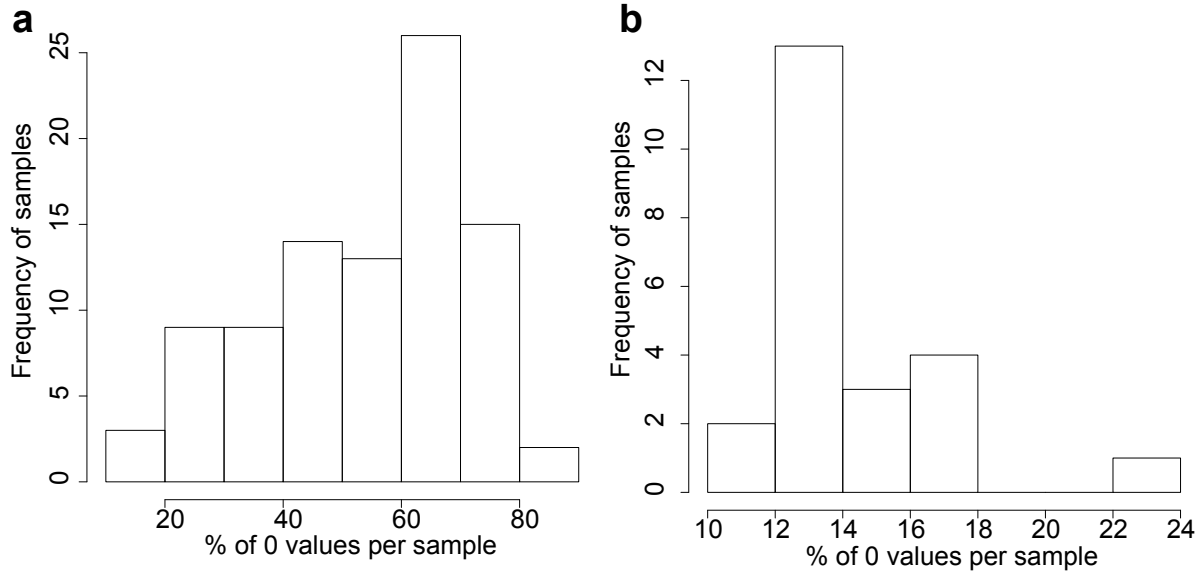


Figure 1.24: Histogram illustrating the percentage of genes having no expression **(a)** per cell in single-cell RNA-seq data [89] and **(b)** per sample in RNA-seq bulk data.

1.5 Introduction to clustering methods

Clustering algorithms are used to segregate data D consisting of M objects in n dimensions into subparts that share a certain similarity. Their main usages consists of deriving knowledge from the data by sorting it according to a measure of similarity and rules of segregation, and in a second time they are used for class predictions. In literature, clustering algorithms are sometimes referred to as knowledge discovery in databases [92]. There exists two main types of clustering algorithms : partitioning and hierarchical algorithms (Fig. 1.25).

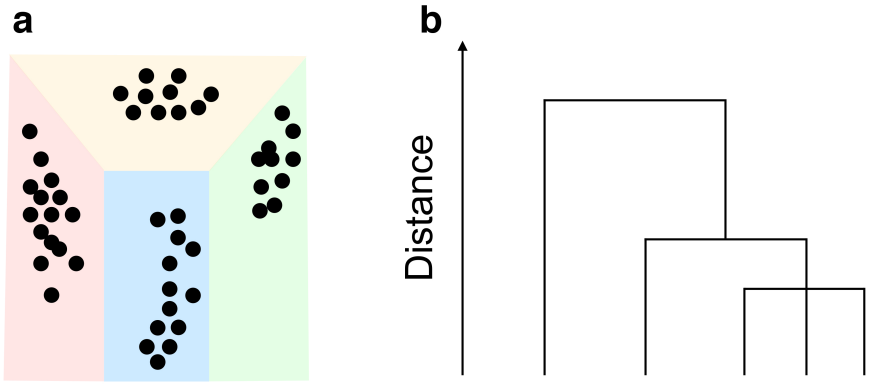


Figure 1.25: The main branches of clustering algorithms. **(a)** Partitioning algorithm, shown with a Voronoï diagram [26], a partitioning of the plane in subsets represented by different colors based on distance to points. **(b)** Hierarchical algorithms shown as a dendrogram.

The partitioning algorithms are creating K subparts out of the data D, and these algorithms are often further subdivided into (1) convex partitioning into K partitions, which is restrictive of the shape of the cluster, (2) density-based approaches, (3) grid-based methods and (4) model-based approaches [93].
The hierarchical algorithms create, as the name suggests, a hierarchical decomposition of the data into subparts starting from one cluster and ending at M clusters. This segregation is often represented by dendrograms, which are tree like structures that separate in an iterative way the data D into smaller subsets until each subset consists of one sample [26] (Fig. 1.25b). This algorithm is appreciated as it gives a full representation of the iteration at which samples merge. It is therefore not required to determine a number K of subsets, however the user can choose a height at which to cut the tree to obtain a partitioning. This height is called a termination condition (TC) [92]. There is few guidance on choosing an appropriated TC and depends on the users needs. One algorithm proposes an automatically determined TC for hierarchical algorithm, called Ejcluster [94], which has the disadvantage of being, however, computationally expensive.
We will make a brief introduction to the clustering methods to which the method developed in this work was compared to. We selected k-means [95] and density-based spatial clustering of application with noise (DBSCAN) for the analysis of real-data as these are two popular clustering algorithms and are representing different subsets of partitioning algorithms. Both of them required selection of several parameters, which usually are difficult to choose, and were optimised for solving the given real data problem. DBSCAN was also selected as it displays similarities with the algorithm developed in this work. We analysed *in silico* generated data

(section 3.2), and we compared it to a model-based clustering approach, which is yet a different subpart of partitioning clustering, called Mclust [96], since this method did not need any parameter selection and was therefore unbiased and comparable to the developed algorithm.

1.5.1 Introduction to k-means

K-means falls into the category of convex partitionings and is a popular clustering method that separates M points in n dimensions (Fig. 1.26a) into K clusters, so that the sum of squares within a cluster is minimized [95]. It is therefore an extremely powerful tool if a pre-existing intuition, obtained for instance through a PCA plot, dictates the likely number of clusters in the dataset. As a second input k-means needs the initial centers $L_1, \ldots, L_K$ (Fig. 1.26a, blue and green). If this input is missing, the **R** implementation of k-means randomly subselects in the data K points that serve as initial centers [97]. Each point in the point cloud, using the euclidean distance, is matched to the closest initial center, forming K clusters (Fig. 1.26b, each color corresponds to one group). Then, new centers are determined by finding the centroïd of the points in each cluster (Fig. 1.26c). A centroïd is a point which minimizes the sum of squares of the distances between the points in the cluster and the centroïd. The previous steps are then repeated until clusters remain unchanged from one step to the next. This algorithm is extremely sensitive to the number of clusters K and the initial centers $L_1, \ldots, L_K$. Fig. 1.26d illustrates the outputs of the algorithm on the same dataset with two different sets of initial centers.

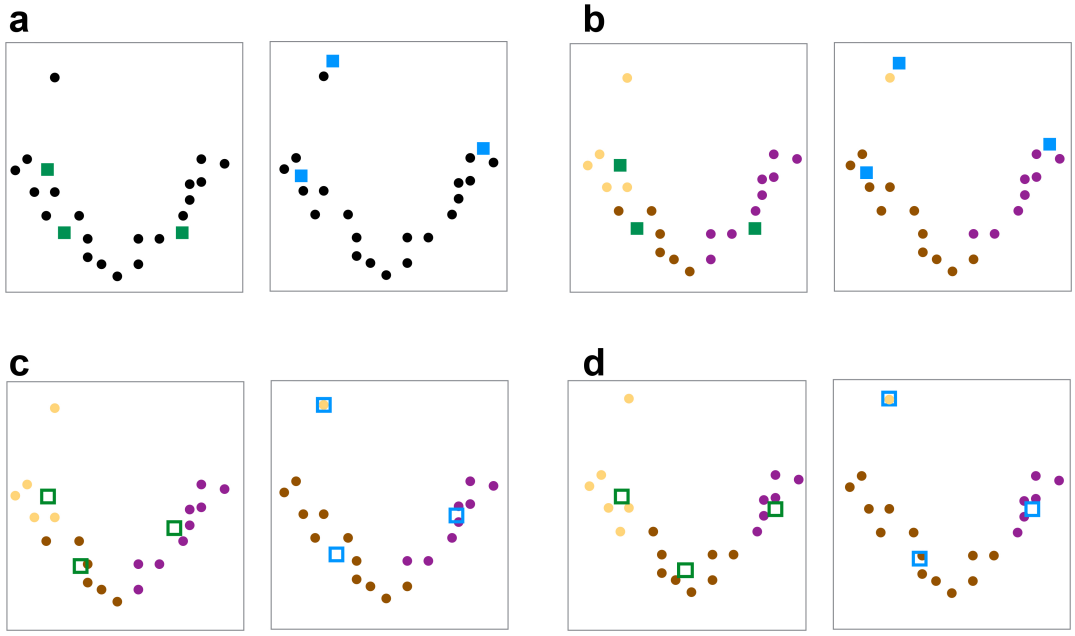


Figure 1.26: Explanation of k-means illustrating the influence of the initial centers on the clusters obtained. **(a)** The point cloud, in black, and the three initial centers, represented with filled blue and green scares, corresponding to two different initial inputs. **(b)** K clusters are being created where K corresponds to the number of initial centers, represented in orange, brown, and purple (for the green centers left, and the blue centers right). **(c)** The new centers are given by the centroïds of the clusters and are represented with unfilled scares (green, left and blue, right), the two first step are then repeated until the centers are again the same, and **(d)** the final clusters are obtained with the initial blue (left) and initial green centers (right).

1.5.2 Introduction to Density-Based Spatial Clustering of Application with Noise (DBSCAN)

DBSCAN falls into the category of density-based approaches and is therefore different from k-means as the clusters are based on a density approach [92]. Looking at the examples in Fig. 1.27a and b, the eye unambiguously defines the notion of a cluster by points that are densely packed. Noise points are defined as isolated islands of points. DBSCAN formalises this intuition by defining the concept of noise and of density. The first parameter of DBSCAN requires the definition of the concept of noise, by determining what is the minimum number of samples that are close enough to compose a core point, called *minPts*. Hence, if *minPts* is chosen to be 3 then a minimum of three points need to be close to a point p so that p is not considered as noise. The clusters are then build around such core points. The second parameter ε formalises the notion of closeness: two points are close whenever their distance is smaller than ε. The algorithm defines the notion of density. Each cluster contains at least one core point and non core-points can be part of a cluster only if they are ε-close to a core point. The notion of density is highly dependent on the parameter *minPts* as this defines the number of points required to lie in a close neighbourhood. As a non-example, if *minPts* is 3, then a two-by-two close chain (as in Fig. 1.27c, grey chain) is not a cluster and these points remain isolated.

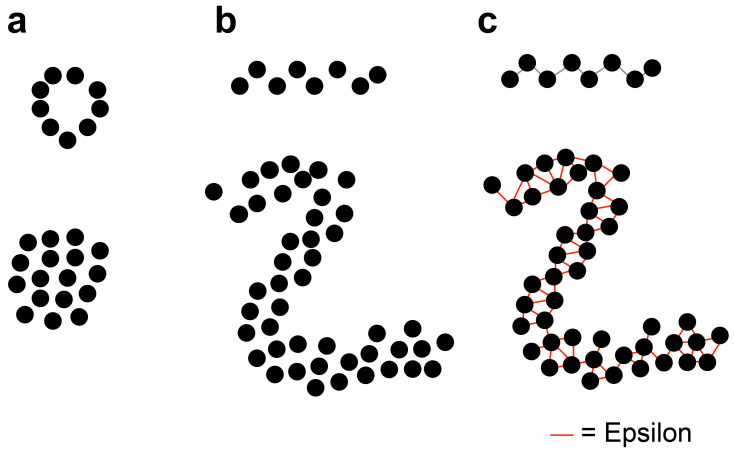


Figure 1.27: Explanation of DBSCAN. Example of two point clouds **(a)** and **(b)** where the notion of cluster is visually apparent. **(c)** In red we connected everything which is densely close, the size of ε is indicated by a red line at the bottom of the picture and *minPts* is chosen to be three. The grey chain illustrates points that are two-by-two close but would not form a cluster.

The parameter ε can be estimated using a visual inspection of a plot. Indeed, for a given k, k-dist is a function from the database that one starts with to the real numbers, where each point is mapped to the distance from its k-th nearest neighbour. One orders the points according to descending k-dist value and plots the ordered points on the x-axis and their k-dist values on the y-axis. This enables an assessment of the density distribution in the data. The threshold is chosen as the k-dist value of the first point in the first "valley" of this k-dist graph (Fig. 1.28). The parameter k can be chosen as *minPts* in this case.

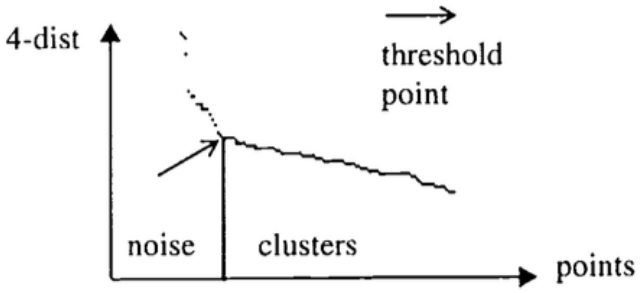


Figure 1.28: The choice of ε using the visual inspection of a k-dist graph in order to understand what is noisy points and where are the clusters. (Illustration adapted from [92].

1.5.3 Introduction to Mclust

The software Mclust in **R** stands for model-based clustering and is part of the model based approaches of clustering. Mclust models the data using a hereafter defined process called finite Gaussian mixture models (GMM) with different covariance structures and different numbers of mixture components (Fig. 1.29). It is updated regularly according to optimized ways to estimate the different parameters that are describe here, further reading can be found in [98], [99], [100] and [101].

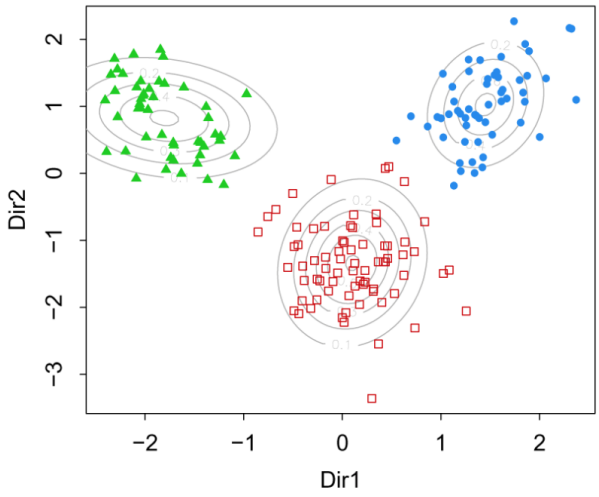


Figure 1.29: Example of application of Mclust with Gaussian Mixture Models for three clusters each having a normal distribution (illustration found in [101]).

Let $x = \{x_1, x_2, \ldots, x_n\}$ be n independent identically distributed observations of a sample. The distribution of every observation is a probability density function given by a finite mixture of models of G components, which takes the following form

$$f(x_i; \psi) = \sum_{k=1}^{G} \pi_k f_k(x_i; \theta_k), \tag{1.3}$$

where $\pi_k > 0$, $\sum_{k=1}^{G} \pi_k = 1$ and $\psi = \{\pi_1, \ldots, \pi_{G-1}, \theta_1, \ldots, \theta_G\}$ are the parameters of the mixture model (usually unknown and are needed to be estimated, chosen or given by the

data). Moreover, $f_k(x_i, \theta_k)$ is the k-th component density for observation x_i with parameter vector θ_k. Since Mclust estimates the mixture model using GMM, it assumes that f_k are Gaussian densities, for every $k = 1, \ldots, G$ parametrized by its mean μ_k and a covariance matrix Σ_k. Then, $(\pi_1, \ldots, \pi_{G-1})$ are the weigths that are reflecting the probability that an observation belongs to the k-th component, and G is the number of mixture components.
Hence, we have fitted a model to our data under the knowledge of our sample points. This needs to be optimized, i.e. the best fit must be found, and therefore, we need to maximize the likelihood function given by $\mathscr{L}(\psi; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \psi)$.
One uses the log-likelihood function to find the maximum of the previous equation to simplify the problem : $\mathscr{L}(\psi; x_1, \ldots, x_n) = \sum_{i=1}^{n} \log(f(x_i; \psi))$. However, as this is still too complicated to estimate, the Mclust algorithm uses an algorithm called the expectation-maximisation (EM) algorithm [102], [103] to compute an estimator of the maximum of this likelihood of the finite mixture model, which represents an approximation.
As said previously if each component follows a Gaussian distribution, it is called a Gaussian Mixture Model, which can be restate by $f_k(x; \theta_k) \sim \mathcal{N}(\mu_k, \Sigma_k)$. The G clusters are estimated or assumed to be ellipsoidal in this model, centered at the mean vector μ_k, and with other geometric features, such as volume, shape and orientation, determined by the covariance matrix Σ_k. As an example, if $\Sigma_k = \lambda I$, where I is the identity matrix, then all the clusters are spherical and of the same size. Moreover, the nearer the points are to the mean the more dense they are.
GMM uses a decomposition of the covariances matrices by means of an eigen-decomposition of the form $\Sigma_k = \lambda_k D_k A_k D_k^T$, where A_k is a diagonal matrix specifying the shape of the density of the points, called density contours, with $\det(A_k) = 1$, and D_k is an orthogonal matrix which determines the orientation of the corresponding ellipsoid [104], [105], finally λ_k is a scalar controlling the volume of the ellipsoid.
In Fig. 1.30, one can observe the effect of the density contours of the ellipsoid A_k, the orientation of the ellipsoid, D_k, and the volume of the ellipsoid, λ_k, being equal (E) or variable (V) across the groups.

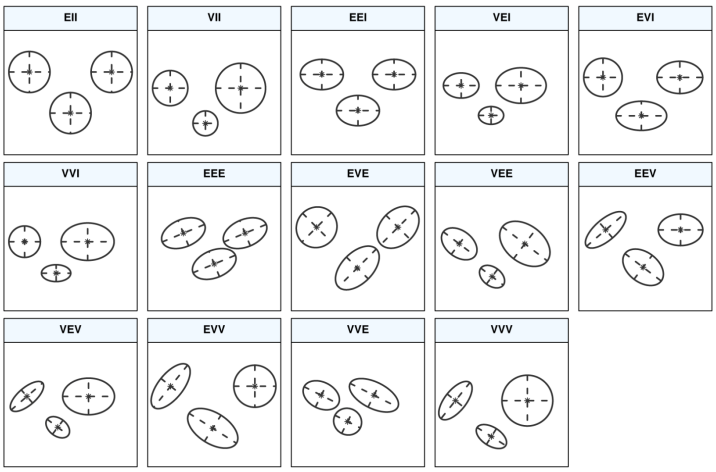


Figure 1.30: Gaussian Mixture Model, 14 possibilities of the combination of the volume, the density contours, and the orientation of the ellipsoid being equal (E) or variable (V) across the groups (illustration found in [101]).

1.6 Role of hormones in the breast development and cancer with a focus on progesterone

(Adapted from: "Progesterone and Overlooked Endocrine Pathways in Breast Cancer Pathogenesis", C. Brisken, K. Hess and R. Jeitziner, Endocrinology, 2015, 156 (10): 3442-3450 [106].)

1.6.1 Breast development

The breast is a unique organ in that it develops primarily after birth, under the control of hormones (Fig. 1.31) [107]. These can be sex hormones produced by the ovaries or the testes, such as **estrogen**, **progesterone** or testosterone or hormones secreted by the pituitary axis such as prolactin. A rudimentary **ductal system** present at birth begins to unfold during puberty and gains in complexity during adulthood with recurrent hormone stimulation during menstrual/estrous cycles. During pregnancy, ductal complexity increases further and finally secretory structures of saccular shape, called alveoli, bud all over the ductal system. Its embryonic-like state after birth makes the breast exquisitely plastic and particularly susceptible to carcinogenesis.

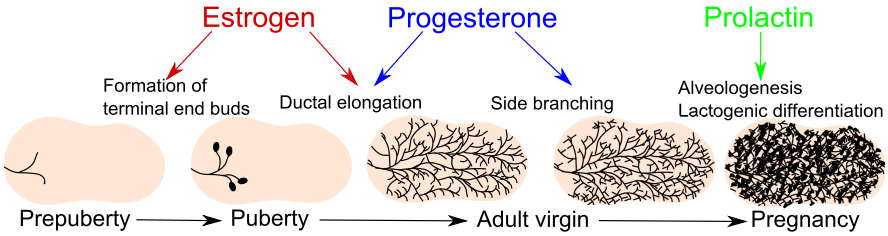


Figure 1.31: Mammary gland development in the mouse. Schematic representation of distinct stages of postnatal mammary gland development. In the pubertal mammary gland, terminal end buds appear at the tips of the ducts triggered by ovarian estrogens, which require epithelial estrogen receptor. The ducts elongate and bifurcate until the edges of the fat pad are reached, which coincides with sexual maturity. Repeated stimulation with progesterone, as occurs during estrous cycles, results in the formation of side branches, which bud from preexisting ducts at a 90° angle. Side branch formation is blocked in progesterone receptor $^{-/-}$ mammary epithelia. Ductal complexity continues to increase during the first half of pregnancy. In the last third of pregnancy, secretory structures of saccular shape, alveoli, sprout all over the ductal system and differentiate into milk-producing units under the control of prolactin receptor signaling.

The mouse mammary gland has served as a model to study gene function in vivo and to genetically dissect gene function in development. A large number of mouse mutant strains are available, and tissue recombination experiments allow one to generate epithelial specific mutants [107]. This approach has revealed that mammary epithelial intrinsic **estrogen receptor** (ER)-α signaling is required for pubertal ductal elongation [108]. **Progesterone Receptor** (PR) is essential in the mammary **epithelium** for side branching and alveologenesis [109], whereas the epithelial prolactin receptor is required for alveologenesis and milk secretion (Fig. 1.31) [110].

On the one hand, different hormone receptor signaling pathways are limiting at distinct developmental stages. On the other hand, the mammary epithelium responds differently to a hormonal stimulus depending on its developmental stage. Hormone ablation and replacement

experiments have shown that 17-β-estradiol induces cell proliferation specifically in pubertal [111] but not in adult mammary glands. In the adult, i.e., more than 8-week-old, female mouse, 17-β-estradiol pretreatment induces the expression of PR [112], whereas subsequent stimulation with progesterone triggers proliferation [113]. Hence, in the adult female PR signaling is the major stimulus of cell proliferation.
These findings among others display the importance of hormones and their receptors throughout development of the breast.

1.6.2 Human menstrual cycle and mice estrous cycle

The anatomy of the human breast with its 15–25 ducts that each give rise to a lobe containing multiple terminal ductal lobular units and 2 distinct stromal compartments, the intralobular and interlobular stroma, is more complex than that of the mouse mammary gland, which has a single stem ductal tree embedded in a homogeneous fatty stroma (Fig. 1.32).

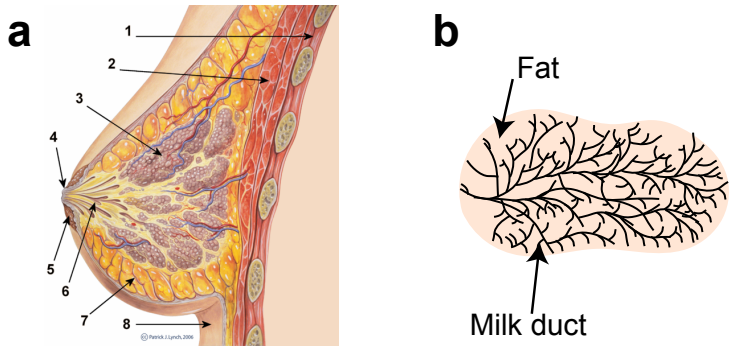


Figure 1.32: **(a)** Human breast scheme, 1. Chest wall, 2. Muscles, 3. Lobules and terminal ductal lobular units, 4. Nipple, 5. Areola, 6. Milk duct, 7. Fat cells, 8. Skin (Adapted original illustration from Patrick J. Lynch, can be found in [114]) **(b)** mouse breast scheme.

Nevertheless, in terms of hormonal regulation, there seem to be substantial similarities across species. In most mammals, the ovaries first secrete estrogens in response to increased secretion of gonadotropins, and sexual maturity coincides with the establishment of cyclic peaks of ovarian progesterone secretion. Progesterone levels increase after ovulation when the body anticipates pregnancy, and continue to rise when pregnancy is established.
The human menstrual cycle lasts around 28 days and is divided in two phases of equal length the **follicular** and the **luteal phase**, separated by ovulation and the latter being characterized by the peak of progesterone (Fig. 1.33). Mice have similar exogenous hormonal exposures and lesser genetical variations limiting drastically the variation encountered between samples. However, the major difference is that the mouse estrous cycle (EC) lasts about 4-5 days as opposed to 28 days for humans. Mice EC consists of four stages: proestrus, estrus, metestrus and diestrus (Fig. 1.34). These are determined using vaginal cytology, where the cell types that are present in the smears of mice are used to group them in: **proestrous**, if the smears consists predominantly of nucleated epithelial cells, **estrus**, with anucleated cornified cells, **metestrus**, if it consists of the three types of cell, cornified leukocytes, and nucleated epithelial cells, and **diestrus**, if it consists predominantly of leucocytes (Supplementary Fig. S7).

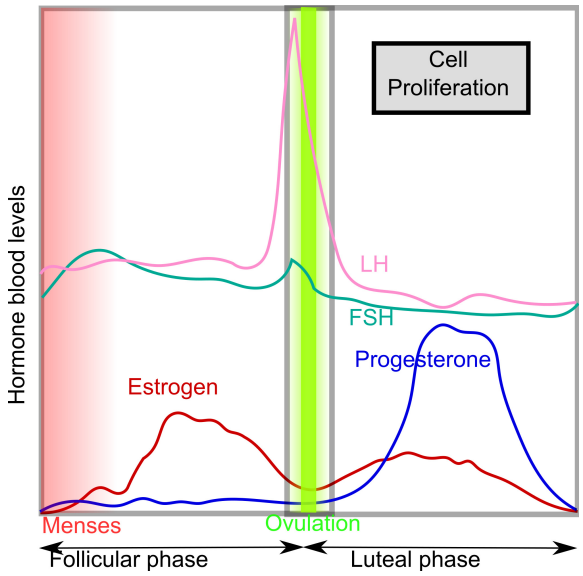


Figure 1.33: The human menstrual cycle. Graph showing serum levels of the major fluctuating hormones across a menstrual cycle. Note that progesterone levels peak during luteal phase. Estrogen reaches its maximum levels during follicular phase; a smaller peak follows during luteal phase. Cell proliferation is observed in the breast epithelium in the luteal phase and positively correlated with serum progesterone levels.

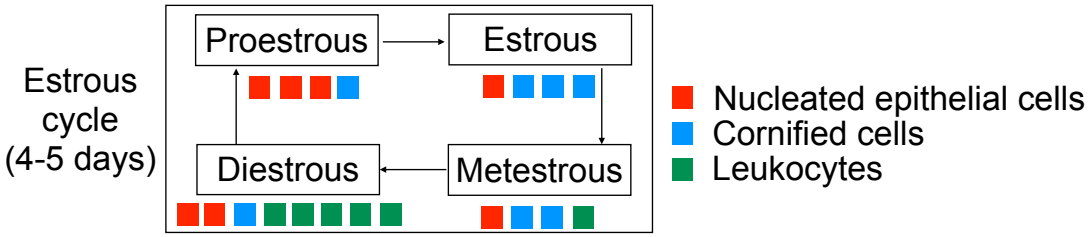


Figure 1.34: Scheme explaining the estrous cycle. Presences of the different cell types at different stages of the estrous cycle: when predominantly nucleated epithelial cells are present, it is called proestrous, estrus is when anucleated cornified cells are mainly present, metestrus smears consists of the three types of cell, cornified leukocytes, and nucleated epithelial cells, and diestrus, if it consists predominantly of leucocytes

Pathologists observe proliferative activity in the breast epithelium during the luteal phase, when progesterone levels peak (Fig. 1.33) [115],[116], suggesting that mouse and human mammary epithelia may indeed be similarly regulated, at least with regards to hormonal control of cell proliferation. This was shown as well using the proliferation marker KI67, which was higher in the luteal phase and upon progesterone stimulation giving further evidence that progesterone signaling triggers proliferation during this phase of the cycle. This is even further enhanced when women are younger than 35 years old, i.e., when the cycle is still fully functional [117]. Recently developed ex vivo models of the human breast have shown that progesterone elicits cell proliferation [118], [119]. Of note, the dog, a species with particularly long luteal phase, is especially prone to mammary carcinoma [120].

1.6.3 Organisation of the mammary epithelium

The mammary epithelium is bilayered: the inner layer of luminal cells is surrounded by a meshwork of elongated myoepithelial cells, which are in close contact with the basal membrane. Luminal cells touch the lumen and are frequently opposed to the cells that are summarized under the term "basal cells": the subluminal, myoepithelial, progenitor, and stem cells. Between 30% and 50% of the luminal cells express ER in the adult female, whether rodent or human [121], [122]. Because PR is an ER target gene, it is coexpressed in the same cells, although evidence has emerged that, at least in the human breast, PR is also independently expressed [123], [124]. In the adult mammary epithelium, most cell proliferation occurs in the luminal compartment, but few of the proliferating cells express ER and PR [121]. When mammary epithelial cells that are PR deficient (genetically PR-/-) are grafted on their own to cleared mammary fat pads, they hardly proliferate in adult hosts. However, when the PR-/- mammary epithelial cells are intermingled with PR wild-type mammary epithelial cells in a 1 to 10 ratio, they proliferate and contribute to all aspects of mammary gland development in the context of the resulting chimeric epithelia [109], indicating that PR signaling can occur in a paracrine fashion. The same applies to ER-/- mammary epithelial cells, which, when grafted on their own, fail to proliferate at all, but which contribute to all aspects of mammary gland development in the context of chimeric epithelia [108]. This motivates us to name the cells expressing ER and PR "sensor cells" [125], because they relay the systemic signal to local partners by emitting paracrine signals (Fig. 1.35).

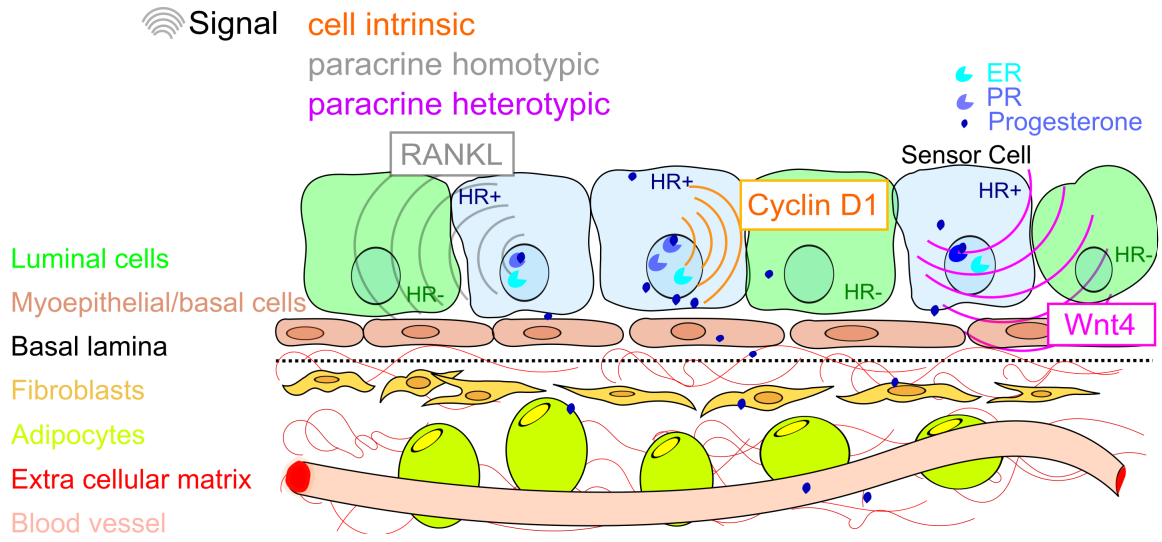


Figure 1.35: Signaling downstream of progesterone. Schematic representation of the bilayered mammary epithelium and the intra- and intercellular signaling activated by progesterone. An inner, luminal layer is surrounded by myoepithelial/basal cells, which are in contact with the basal lamina. Progesterone binds its receptor in a subset of hormone receptor + luminal cells, the sensor cells (light blue). In certain PR+ cells, it induces cell proliferation by a Cyclin-D1-dependent mechanism (cellintrinsic signaling). It induces a tumor necrosis factor, called RANKL, which elicits cell proliferation in neighboring hormone receptor − cells (paracrine homotypic) and wnt4, which acts on myoepithelial cells (paracrine heterotypic) and increases stem cell activity.

1.6.4 Cell proliferation mechanism of progesterone in the mammary epithelium

When adult female mice are hormonally ablated and subsequently pretreated with estrogens, progesterone induces cell proliferation in 2 waves. During the first 24 hours, PR+ cells proliferate, whereas proliferation of PR- cells is observed subsequently [126]. The first, small wave of cell-intrinsic proliferation requires **Cyclin-D1**; whether this relates to its cell cycle and/or its transcription-related functions is unclear. Support for such scenarios can be found in observations from the PR+ breast cancer cell line T47D, which reveal both that PR and Cyclin-D1 interact physically and are found in transcription complexes that bind to DNA and that down-modulation of Cyclin-D1 expression blocks PR-B-induced gene transcription [127]. The second wave of cell proliferation, induced by a paracrine mechanism, is larger and relies on a tumor necrosis family member, receptor activator of nuclear factor κB ligand (**RANKL**). Progesterone increases RANKL mRNA expression by a posttranscriptional mechanism stabilizing the mRNA [118]. RANKL protein is detected exclusively in PR+ cells [126], [128]. Whether RANKL itself acts as mitogen, or removes a growth-inhibitory signal, or acts through a more complex loop involving other cell types, possibly infiltrating immune cells known to express the receptor, awaits further clarification. Although individual cells in luminal and abluminal locations express the cognate receptor RANK [129], it remains to carry out costainings to determine whether RANK+ cells are actively cycling epithelial cells.

1.6.5 Stem cell activation of progesterone in the mammary gland

During luteal phase, stem cells are likely to be activated in anticipation of the cell number expansion of pregnancy. Stem cells have been studied by 2 major approaches, one entailing fluorescence-activated cell sorting, to enrich for cells with the ability to reconstitute mammary glands divested of their endogenous epithelium, and the other lineage tracing. Stem cells as defined by the first approach are located in the basal layer and express high levels of integrin β1 and α6 [130], [131] and have been shown to expand in response to hormone stimulation [132], [133]. Lineage-tracing experiments indicated, however, that most postnatal cell proliferation derives from luminally restricted stem cells [134], [135]. To assess the role of PR signaling in stem cell function comprehensively, we resorted to serial transplantation. Mammary epithelium can reconstitute up to 7 transplant cycles [136]. When we compared PR-/- and PR wild-type epithelia by serially transplanting them in contralateral glands, PR wild-type only slightly decreased in fat pad reconstitution over 4 generations, but PR-/- failed to reconstitute at the third generation, indicating that PR signaling is required to expand the stem cell pool during puberty and in adult life [137].

Wnt signaling is important to adult stem cells in many tissues, including the mammary gland [138]. Wnt family member 4 (**Wnt4**) emerges as central activator of mammary epithelial stem cells and of their niche(s). On the one hand, it may act directly on bipotent or basally restricted stem cells located in the basal layer, which can be identified based on expression of the protein c receptor, itself a Wnt target gene [139]. In this, Wnt4 is helped by a membrane protein expressed in hormone receptor negative cells, R-spondin 1, which enhances canonical Wnt signaling and is itself induced by hormone stimulation [140]. On the other hand, wnt4 may act directly, possibly via noncanonical Wnt signaling and/or indirectly via distinct paracrine signals on luminally restricted stem cells. A potential paracrine mediator is growth hormone, which can be synthesized in the breast epithelium and has been implicated in progesterone-

induced activation of stem cells in the human breast, in work inspired by observations on dogs [141].

1.6.6 Relevance of rodent work on the mammary gland to human research

Two lines of work suggest that at least some of the findings in rodent models are of relevance to humans. First, work with a novel ex vivo model for the human breast consisting of tissue microstructures isolated from fresh reduction mammoplasty specimens that remain responsive to hormones, has shown that progesterone triggers cell proliferation in the adult human breast tissue and that it induces the expression of RANKL and WNT4 transcripts [118], [119]. Second, next-generation whole transcriptome sequencing was used to analyze global gene expression in the breast epithelium from 20 premenopausal women, who were not affected by breast cancer and donated breast tissue to the Susan G. Komen for the Cure Tissue Bank (http://komentissuebank.iu.edu/), and who were carefully staged for the menstrual cycle [142]. This study revealed 255 genes that are differentially expressed between follicular and luteal phase, with 221 increased in luteal phase; in functional terms these genes related to cell cycle and mitosis, and DNA damage and repair, as also observed in vitro [143], [144]. Interestingly, this unbiased approach identified 3 paracrine factors: RANKL, WNT4, and epiregulin [142].

1.6.7 Hormonal risk factor for breast cancer

Breast cancer affects 1 in 8 women in Western countries [145]. The disease is heterogeneous: more than 20 distinct histopathological subtypes are recognized [146]. Of clinical relevance are tumor grade and tumor stage, as well as classification according to ER-α and progesterone receptor status, as assessed by **immunohistochemistry** (IHC), and erb-b2 receptor tyrosine kinase 2 called also more commonly HER2 overexpression due to amplification, as determined by IHC and fluorescent in situ hybridization. Five major molecular breast cancer subtypes were discerned by global gene expression profiling and largely correspond to IHC subtype, with luminal A representing ER+, of low grade and low Ki67 index; Luminal B, ER+ of higher grade and proliferative index; HER2 being HER2+ by IHC and either ER+ or ER-; and the "basal-like," which are dubbed triple negative because they do not express any of the 3 receptors. The last is a heterogeneous group that contains further subtypes [147]. More than 2 thirds of all breast cancers are luminal, i.e., ER+, and differ in biology and clinical course from HER2+ and basal-like tumors [148]. Tamoxifen is a selective ER modulator, which was introduced over 40 years ago and has dramatically increased survival of ER+ breast cancer patients [149]. ER signaling can now also be inhibited by pure ER antagonists, such as fulvestrant, or indirectly by aromatase inhibitors, which are the mainstay in the therapy of most postmenopausal breast cancer patients. Although most ER+ tumors express ER in at least 90% of the cells, some cancers have lower percentages of ER+ tumor cells, but are still classified as ER+ as long as at least 1% of the tumor cells express ER [150]. The ER signaling pathway has long been a major focus of research in the breast cancer field; although it is of premier importance in the therapy of ER+ breast cancer, other hormonal factors are increasingly considered to play an important role in the pathogenesis of the disease [151], [152], [153]. Endogenous hormones, in particular progesterone, impinge on the breast and their role in tumor development.

Ovariectomy [154] was shown to benefit individual breast cancer patients more than 100 years ago. Epidemiological studies revealed that breast cancer risk increases with the number of

menstrual cycles a woman experiences in her lifetime [155]: early menarche, late menopause, and short menstrual cycles all increase risk [156]. Based on breast cancer statistics from the seventies that were not confounded by hormone replacement therapy (HRT) , Pike *et al.* [155] calculated that if it were not for menopause there would be 6 times as many cases of breast cancer [155], [157]. More recently, it was shown that the risk related to menstrual cycles applies to all subtypes of breast cancer [158]. Young age at first pregnancy has a protective effect [159], [160]; more detailed data from the Nurses Health Study indicates that this applies to hormone-receptor-positive, more specifically PR+, breast cancers [158]. The protective effects of early pregnancy rely on a number of factors: lower levels of growth hormones [161] and prolactin [162] after a first pregnancy, changes in stem cell numbers and biology, changes in p53 functional status [163], and differences in the proliferative response have all been implicated [164].

The findings in section 1.6.4 suggest that PR signaling and its downstream effectors activate biological processes, such as cell proliferation and stem cell activation, that may account for the tumor-promoting effects of recurrent menstrual cycles. The same mechanisms may be activated when exogenous progestins are administered, as in the context of HRT and oral contraception. Large women's health studies revealed that breast cancer risk related to HRT increases when an estrogenic compound is combined with progestin [165], [166], [167], whereas estrogens on their own can have protective effects [168]. Indeed, since HRT was discontinued, breast cancer incidence has diminished [169]. Similarly, women who are currently on oral contraception, most of which consists of ethinyl estradiol and a progestin, have a 24% increased risk of getting breast cancer, which decreases once they stop taking the pill [170]. However, PR signaling itself is context-dependent, and not all PR signaling is tumor promoting. Pregnancies have a protective effect early in life with a 50% reduction in lifetime risk of breast cancer before the age of 20 [159]. However, they bring on extremely high levels of progesterone, with serum progesterone reaching 180 ng/mL in the third trimester, compared with 8–33 ng/mL in luteal phase and 0.1–0.8 ng/mL in follicular phase. Thus, the biological effects of progesterone may depend on the dose, the duration of the stimulus, the presence of concomitant high levels of 17-β-estradiol and other hormones, as well as on the woman's age.

A third ovarian hormone, testosterone, fluctuates to some extent during the menstrual cycle with a modest peak 3 days before the Luteinizing hormone (LH) peak [171], [172]. Interestingly, testosterone was reported to be the only hormone, the blood levels of which correlated with breast cancer risk in women with regular menstrual cycles [173]. Whether cyclic activities of this hormone contribute to the risk associated with menstrual cycles needs to be explored. The role of this hormone in tumorigenesis is complex and dependent on the ER status of the tumor, as reviewed in Ref. [174].

A number of other hormones impinge on the basic regulatory network controlled by the ovarian hormones [175]. They may serve to fine-tune the system or have distinct functions. In this context, an extensive study of normal human breast samples is of interest. It revealed 7 subsets of hormone receptor positive cells, all of which are luminal in the human breast: ER+, androgen receptor (AR)+, vitamin D receptor (VDR)+, ER+AR+, ER+VDR+, AR+VDR+, and ER+AR+VDR+. Other hormone receptors that were tested, including thyroid hormone receptor-α, thyroid hormone receptor-β, parathyroid hormone 1 receptor, oxytocin receptor, various somatostatin receptors, RARα, RARβ, RXRα, and RXRβ, did not show a bimodal expression pattern [176]. It will be of interest to see whether these populations of hormone

receptor positive cells are conserved across species and whether the distinct receptor expression patterns characterize distinct cell types with specific biological function.

1.6.8 Tumor promoting action of Progesterone

Based on the above, we propose a model of menstrual cycle effects on breast carcinogenesis (Fig. 1.36), in which the repeated activation of PR signaling during luteal phase may be tumor promoting. Some of the effects of progesterone are cell-intrinsic, but many biological responses rely on paracrine signaling that can be homotypic, i.e., to neighboring luminal cells, or heterotypic, i.e., to the myoepithelium and possibly to stromal cell types.

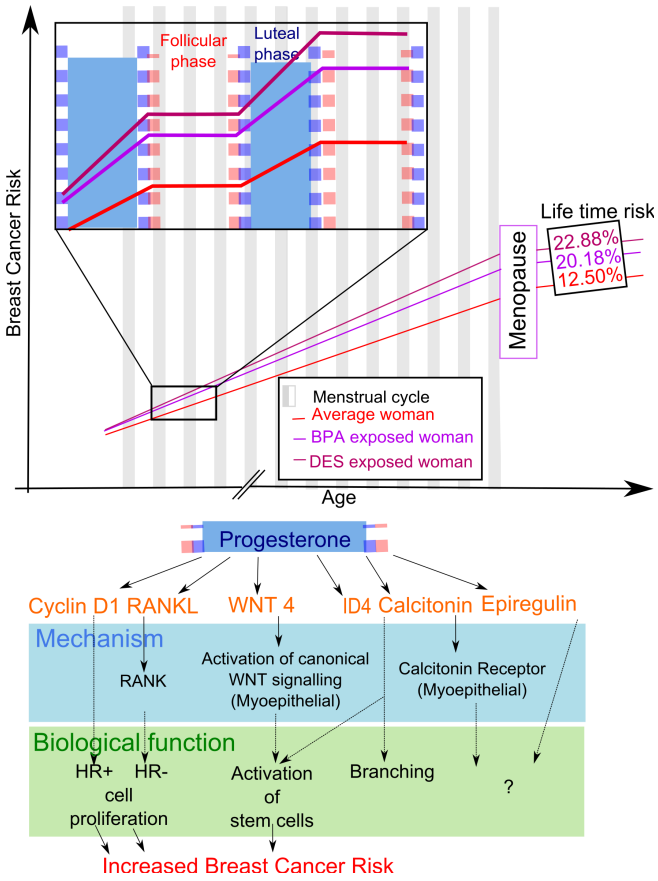


Figure 1.36: Graph showing breast cancer risk plotted over a woman's age, depending on whether or not she was exposed to DES or BPA. With each menstrual cycle, breast cancer risk increases through progesterone-induced events during luteal phase. The model proposes that perinatal exposure to endocrine disruptors increases the sensitivity of the breast to progesterone and hence increases the slope of the curve (top panel and inset). Various factors such as RANKL, WNT4, epiregulin, Cyclin-D1, ID4, and calcitonin, which act through distinct mechanisms and have been shown to have distinct biologically functions, have been implicated in the biological response to progesterone that may be amplified due to perinatal exposure.

Tumor-promoting effects of progesterone are also observed in rodents, where chemically (7,12-Dimethylbenzanthracene)-induced carcinogenesis is enhanced by progesterone/progestin administration [177], [178]. In support of this model, pharmacologically or genetically blocking RANKL delayed tumorigenesis [129], [179]. Interestingly, RANKL inhibition was not effective

anymore once the tumor was fully established [129], suggesting that the PR/RANKL axis is important specifically early in the pathogenesis of mammary carcinomas.

Similarly, the Wnt signaling pathway may promote tumorigenesis. In the mouse, wnt1, a wnt4 cousin, was long identified as an oncogene by cloning of the frequent insertion site of the oncogenic mouse mammary tumor virus [180]. Ectopic expression of Wnt1 in the mammary epithelium results in highly penetrant widespread hyperplasia and, ultimately, tumors [181], consistent with an early tumor-promoting effect that may rely largely on indirect and niche-related effects. In T47D cells, Wnt1 is a PR-B target and induces matrix metalloproteases to shed epidermal growth factor receptor ligands that transactivate the epidermal growth factor receptor [182].

1.6.9 Introduction to Receptor Activator of Nuclear factor κB ligand

In 1997, receptor activator of nuclear factor κB ligand (RANKL) was discovered along with its cellular receptor RANK while researchers were sequencing cDNAs from a human bone marrow derived myeloid dendritic cell cDNA library [183]. Two research groups independently found osteoprotegerin (the bone protector, OPG), a decoy receptor for RANKL, by using mice models lacking the cDNA for OPG [184], [185]. By limiting osteoclastic bone resorption, it appeared to protect the skeleton from bone resorption. Both groups then used expression cloning and OPG as a probe, to identify its ligand that they called OPG ligand and osteoclast differentiation factor, respectively [186], [187]. This ligand was identical to RANKL that was previously discovered [183]. Since then, RANKL was considered to be involved in osteoclastogenesis and T cell activation and later was shown to be involved in the differentiation, activation and function of osteoclasts [188]. It was identified as part of the Tumour Necrosis Factor alpha family and exerts its biological functions by binding to RANK, the complex RANK/RANKL activates nuclear factor κB (NF-κB), which is a transcription factor for immune-related genes. It is also a key regulator of inflammation, innate immunity, cell survival and differentiation [188].

RANKL displayed important functions for lactational hyperplasia of mammary epithelial cells and milk production [189]. Moreover, in lactating mice RANKL was required for the development of lobulo-alveolar structures [189]. The defect in mammary gland female mice lacking RANKL was characterized by enhanced apoptosis and failures for the cells to proliferate [189]. Through genetic studies in the mouse mammary gland, it has been shown that RANKL is important as a paracrine mediator of progesterone-induced proliferation in the adult mouse mammary gland [190]. This conclusion was based on the observation that RANKL localizes to progesterone receptor positive (PR+) cells just next to cells that are actively replicating their DNA [190]. In cancer cells, several studies demonstrated that increased RANK signaling contributes to breast carcinogenesis by interfering with mammary cell commitment [191], [192], [193]. This was shown in cells treated with the progestin medroxyprogesterone acetate, a widely used hormonal contraceptive and hormone replacement therapy compounds, which massively increases the level of RANKL on mammary epithelial tissue [191], [192]. Pharmacological inhibition of RANKL lowered tumor progression and metastasis [191], [192] in mice and if the same effect can be observed in human breast cancer is still questioned.

The study of breast cancer has been hampered by the difficulty in culturing cells from patients (both cancer and normal) that retain hormone receptor expression. To circumvent

this difficulty, a way of culturing breast tissue microstructures from healthy donors was developed that was shown to retain hormone receptor activity [118]. This technique was used on healthy human cells that were treated with promegestone, called **R5020**, a PR agonist, resulting in RANKL induction [118]. RANKL was also demonstrated to be sufficient to induce cell proliferation and was required for R5020-induced proliferation [118]. The findings were validated in vivo, where RANKL protein expression in the breast epithelium correlated with serum progesterone levels. The ligand was expressed in a subset of luminal cells that express PR. Systemic inhibition of RANKL signaling by intravenous injection of recombinant osteoprotegerin, blocked the induced proliferation in the mouse mammary epithelium [118].
Recall that breast cancer risk is increased with increasing number of menstrual cycle, and is increased while taking hormonal contraceptives (see section 1.6). This is partially associated to the high proliferation of cells in the luteal phase of a menstrual cycle. Blocking or reducing this proliferation might be beneficial for breast cancer prevention. Therefore, RANKL was proposed as a possible target for breast cancer prevention and maybe treatment as inhibiting RANKL reduces the proliferation of healthy cells and because of its role in reducing mice breast cancer progression. However, this hypothesis should be strenghten by further work using human data. For the treatment of bone diseases, there exists an inhibitor for RANKL, called denosumab, which is in use since June 2010, when it was approved by the U.S. Food and Drug Administration for treatment of postmenopausal women at high risk for fracture and for prevention of skeletal-related events in patients with bone metastases. Breast cancer patients could benefit from denosumab.
Hence, whether the findings in the mouse model can be translated to humans and whether denosumab will be of use to prevent breast cancer or treat the disease is an urgent question. It is therefore of clinical interest to further investigate the molecular functions of RANKL in the human mammary gland.

1.7 Aim : Development of an unbiased topological tool overcoming problems linked to variable data

(This section is Adapted from: "Two-Tier Mapper: a user-independent clustering method for global gene expression analysis based on topology", R. Jeitziner, M. Carrière, J. Rougemont, S. Oudot, K. Hess, and C. Brisken, 2017, arXiv: 1801.01841 [194], submitted to Bioinformatics.)
Large datasets, specially coming from high-throughput sequencing are generated at an exponentially increasing pace in biology and medicine [195],[196],[197], while the development of tools to analyze them lags behind. Unbiased clustering methods are needed to analyze these growing numbers of complex data sets. Challenges are posed by the variability of biological, in particular clinical, samples, data acquisition at different times and on different platforms, and the necessity to compare measurements at different stages of the life cycle of an individual patient. Moreover, reliability of patient generated metadata (using a questionnaire for instance) and classification in subtypes requires a particular attention. Statistical methods require large sample numbers to determine the distribution of the data and to extract statistically significant features [198]. The choice of normalization can be arbitrary and may affect the outcome of the analysis [198].
Most clustering methods including k-means [95], hierarchical clustering [26], PAD [9] and Mapper [10] require large sample sizes [199], [26], [200] and depend on parameters, which are chosen by the users and may affect the outcome [201]. To ensure that minor perturbations of the dataset do not alter clusters, the methods applied should be stable [26] [201].
We aim to address the following challenges by developing a topological tool that :

- is user-independent, parameters should be chosen by the data or strong default, such as to reduce the user-induced bias.
- is stable.
- is not harshly affected by normalisations.
- should not change when not all the data is included (removal of samples/of rows).
- is able to overcome problems encountered when facing variable human data, with a perspective towards personalised medicine.
- compares two groups.
- generates a visual clustering with new insights.
- gives for each sample an individual appreciation for how close it is to the compared group.
- is written as a Bioconductor program R such that it gets a broad audience.
- produces an interpretable output.
- works for a broad range of data, but specially for : Microarray/RNAseq and maybe single-cell RNAseq.

To address these challenges, we have developed a topology-based clustering tool called Two-Tier Mapper (TTMap) for enhanced analysis of global gene expression datasets.

2 Two-tier Mapper (TTMap)

(This whole chapter is Adapted from: "Two-Tier Mapper: a user-independent clustering method for global gene expression analysis based on topology", R. Jeitziner, M. Carrière, J. Rougemont, S. Oudot, K. Hess, and C. Brisken, 2017, arXiv: 1801.01841 [194], submitted to Bioinformatics.)

2.1 Overview of Two-tier Mapper

Each global gene expression profile is represented as a high-dimensional vector in $\mathbb{R}^n$ with n the number of genes, features or probes. The input (Fig. 2.1, green) of TTMap is given by two matrices in log-2 scale, one for the control samples, N the other for the test samples T. Batches are defined as groups of samples sharing a source of variation, such as experimental date, technical platform for data acquisition, date or site of sequencing, or biological differences, such as mouse strain, patient age, or other.

TTMap consists of two independent parts; the Hyperrectangle Deviation Assessment (HDA) and the Global-to-Local Mapper (GLMap) (Fig. 2.1). HDA characterizes the control group, N; it adjusts for outliers and generates the "corrected control group", $\bar{\mathsf{N}}$, which is the reference for calculating the deviation of each individual test vector. GLMap uses the Mapper algorithm [10] with the following parameters: a two-tier cover $\mathscr{I}$, the mismatch distance d_M, computed with the previously calculated deviations, the closeness parameter ε, which is data-driven, and a special filter function f, which provides a gradient of proximity to the corrected control group. Through the filter function the two-tier cover detects global and local similarities in the deviation patterns allowing it to capture the structure of the test group based on relatedness of samples. The test samples are clustered according to the shape of their deviation from the control; each cluster is represented by a sphere the size of which reflects the number of samples it comprises (Fig. 2.1). The extent of deviation of individual clusters from the corrected control group translates into a color-code as well as the arrangement from left to right (Fig. 2.1). A subsequent analysis of the commonly changed features in a cluster discerns the differentially expressed genes (Fig. 2.1).

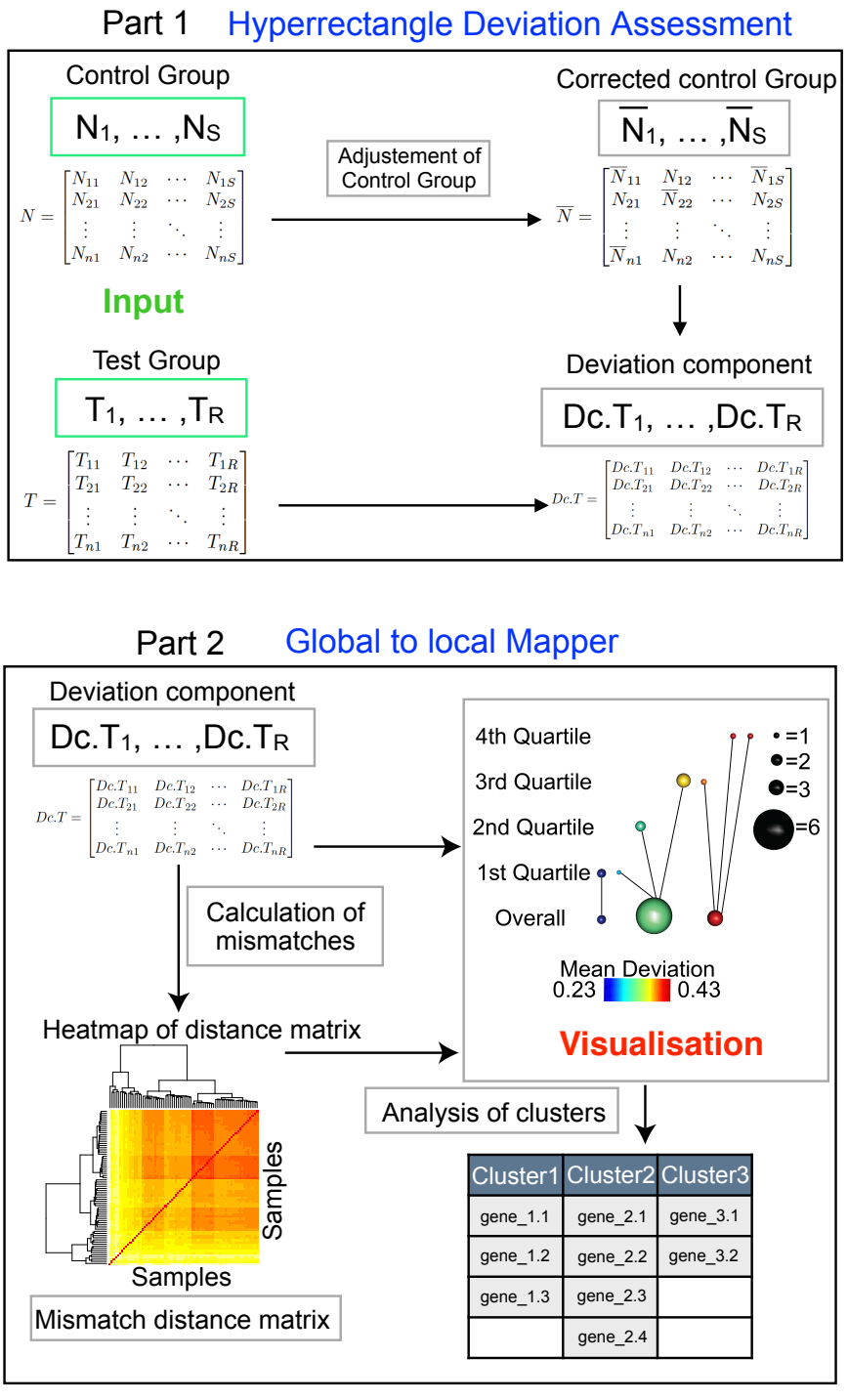


Figure 2.1: Schematic overview of TTMap. The inputs (green) are given by two gene expression matrices, the control (N) and the test group (T), rows represent genes and columns samples. In Part 1, TTMap adjusts the control group for outlier values ($\bar{N}_*$), feature by feature. It calculates deviation from this corrected control group for individual samples in the test group ($Dc.T_*$). In Part 2, TTMap computes a similarity measure, the mismatch distance (represented as a heatmap) using the deviation components. The Mapper [10] algorithm is used with a two-tier cover to generate a visual representation of the clustering creating a network of global clusters (Overall) and local clusters (1st, 2nd, 3rd, 4th quartile of a filter function). It takes as inputs the mismatch distance and the deviation components.

2.2 Hyperrectangle deviation assessment (HDA)

2.2.1 Data preprocessing

Prior to the analysis, the collected data are log-transformed and grouped into two separate tables, where columns are samples and rows are features from:

- a group N, called the **normal (or control) group** the elements of which are denoted $N_1, \ldots, N_S$, where S is the number of collected samples in this group.
- a group T, called the **test group**, the elements of which are denoted $T_1, \ldots, T_R$, where R is the number of collected samples in this group.

The number of features measured (e.g., number of genes expression levels of which were determined) in each sample is written n. Thus, each element in group T and group N is a vector in $\mathbb{R}^n$.
If different numbers of features have been measured for groups T and N, then HDA considers only the features measured across both data groups.

2.2.2 Generation of a hyperrectangle of values in the control group

HDA compares the value of each feature of each control sample N to the values of that feature of all the other control samples in the same batch N (Fig. 2.1, "adjustement of control group"). If the absolute value of the difference between a given value and the median of the values of all the other samples is more than e, the value is considered an outlier and replaced by Not a Number (NA). The e parameter is computed using the variances of all the genes to accommodate for the variability of the dataset (see section 3.2.8).
Hyperrectangle deviation assessment first analyses the vectors in the control group N; for each sample in N feature by feature, i.e. probe by probe or gene by gene, the algorithm compares the measured value to the values of all the other measurements of that feature in samples from the same batch in group N (Fig. 2.1, "adjustement of control group"). If the absolute value of the difference between a given value and the median of the values of all the other samples is more than e, the value is considered an outlier and replaced by Not a Number (NA). The e parameter is computed using the variances of all the genes to accommodate for the variability of the dataset by the 90^{th} percentile of the standard deviations for every feature multiplied by $\frac{2}{\sqrt{S}}$, where S is the number of samples in the control group. If this 90^{th} percentile is small, the user can also change and choose to take the e parameter to be 1 for instance in order to remove only highly variable features or to be a huge number to skip this step. The user can identify highly variable features of the control group by examining the numbers of replaced values for each feature (Fig. 2.2a). A barplot showing the number of adjusted values per sample helps identify outliers in the control group (Fig.2.2b, Fig. 3.10b, Supplementary Fig. S14c and d). Thus, HDA creates a matrix that describes the range of expression values expected in group N corrected for outliers. The (k, j)-coefficient of this matrix of the corrected control group, $(\overline{N}_k)_j$, which corresponds to the j^{th} feature of sample k, is computed by:

$$((\overline{N})_k)_j = \begin{cases} NA & \text{if } |(N_k)_j - \text{median}_{i \in \mathscr{I}(N_k), i \neq k}(N_i)_j| \geq e \\ (N_k)_j & \text{otherwise.} \end{cases}$$

Here, $(N_i)_j$ denotes the value of the expression of gene j in sample i, and $\mathscr{I}(N_k) \subseteq \{1, ..., S\}$ is the set of indices of control samples in the batch containing N_k. Each NA is replaced by

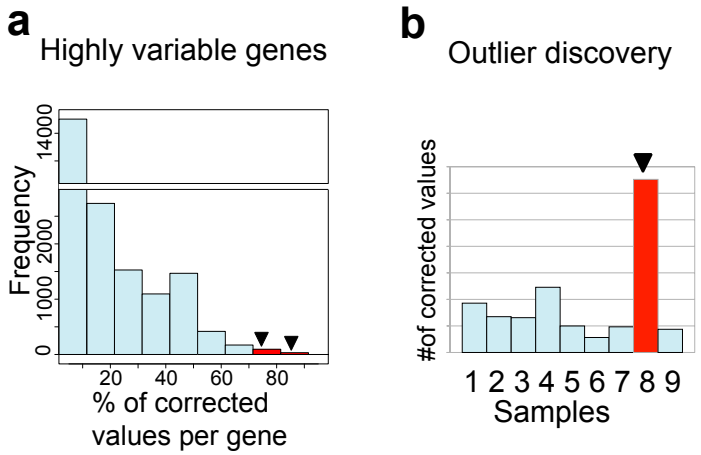


Figure 2.2: Possible outputs after the first part of TTMap **(a)** histogram representing the frequency of features per percentage of outliers and **(b)** a barplot of the number of outliers per sample in the control group to enable the discovery of highly variable genes or samples (red, arrow).

the median of its batch, in order not to affect the next steps.

This step of the method is obtained using the function *control_adjustement*() in R and can be summarized by the following table.

Algorithm 3: Control Adjustment

Input: Two matrices corresponding to the control group, and the test group, where in the rows are the genes or features, in the columns are the samples,
the e parameter (not mandatory), a method to replace the NA that will be generated in this step (not mandatory).
Method: If the e parameter is missing it is computed by the 90^{th} percentile of the standard deviations for every feature multiplied by $\frac{2}{\sqrt{S}}$, where S is the number of samples.
For each value in the control group, if the absolute value of the difference between a given value and the median of the values of all the other samples is more than e, this value is replaced by NA.
The number of NAs per row and per column is computed.
The NAs are replaced by the median of all the other values or by an method given in the inputs.
Output: 5 files are created:
- The normal matrix with only common features with the test matrix. This file is only created if the two matrices in the input have different rows.
- The test matrix with only common features with the normal matrix. This file is only created if the two matrices in the input have different rows.
- A pdf showing a plot of the mean (X axis) against the variances (Y axis) of each feature.
- A pdf showing a plot of the mean (X axis) against the variances (Y axis) of each feature after correction of the control group.
- The number of outliers per row.
- The number of outliers per column.
Moreover, an outlier corrected control matrix is created.

Each feature has a range of values, in which control measurements are expected, for sample T_k and gene j given by

$$B_j^k = \Big[\min_{i\in\mathscr{I}(T_k)}(\overline{N}_i)_j, \max_{i\in\mathscr{I}(T_k)}(\overline{N}_i)_j\Big],$$

where $\mathscr{I}(T_k)$ is the set of indices of control samples in the batch containing T_k (Fig. 2.3). For each batch, these normal ranges determine a hyperrectangle in n-dimensional space $B_k = B_1^k \times \cdots \times B_n^k$ (Fig. 2.3: example with $n = 2$).

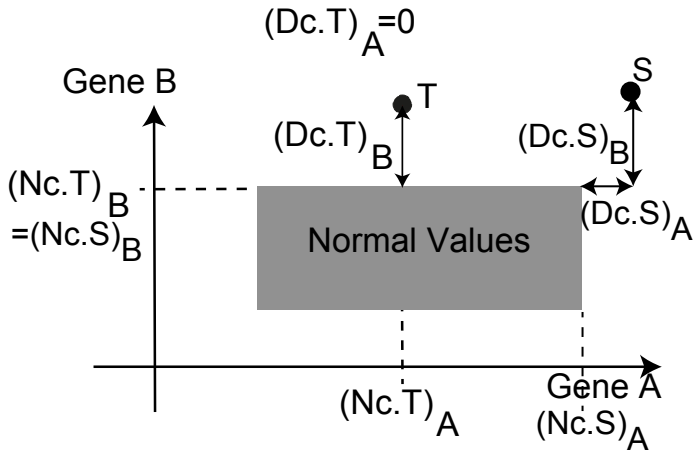


Figure 2.3: Deviation component calculation. Scheme of two test sample T and S together with their deviation components $Dc.T = (Dc.T_A, Dc.T_B)$, $Dc.S = (Dc.S_A, Dc.S_B)$ and normal component $Nc.T = (Nc.T_A, Nc.T_B)$, $Nc.S = (Nc.S_A, Nc.S_B)$ from the hyperrectangle (box) of normal values, example for $n = 2$ genes A and B

2.2.3 Deviation component calculation from the hyperrectangle

Each test sample T_k is decomposed as $T_k = Nc.T_k + Dc.T_k$, where $Nc.T_k$ is the **normal component**, which is its projection onto the hyperrectangle B_k and hence is the closest point to T_k inside B_k (Fig. 2.3) and the **deviation component** ($Dc.T_k$), which is the remainder of the projection (Fig. 2.3).

More precisely, for each test sample T_k and feature j, HDA computes

$$\bar{x}_j^k \in \Big[\min_{i\in\mathscr{I}(T_k)}(\overline{N}_i)_j, \max_{i\in\mathscr{I}(T_k)}(\overline{N}_i)_j\Big],$$

such that

$$|(T_k)_j - \bar{x}_j^k| \le |(T_k)_j - x|$$

for all

$$x \in \Big[\min_{i\in\mathscr{I}(T_k)}(\overline{N}_i)_j, \max_{i\in\mathscr{I}(T_k)}(\overline{N}_i)_j\Big].$$

Then,

$$(Nc.T_k)_j = \bar{x}_j^k \quad \text{for all } 1 \le j \le n$$

and

$$(Dc.T_k)_j = (T_k)_j - (Nc.T_k)_j \quad \text{for all } 1 \le j \le n.$$

This part is obtained using the function *hyperrectangle_deviation_assessment()* in **R** and can be summarized by the algorithm.

Algorithm 4: Hyperrectangle deviation assessment

Input: Two matrices corresponding to the control group, and the test group, where in the rows are the genes or features, in the columns are the samples, batches.
Method: In each batch in the control group, find the minimum and the maximum per gene or feature.
Define a hyperrectangle made out of the product of all the intervals of minimum and maximum calculated previously.
Each test sample is projected onto the hyperrectangle.
The projection and the remainder of the projection are calculated.
Output: 3 output files are created :
-The matrix of the remainder of the projection for each test sample (Dc).
-The matrix of the normal components for each test sample (Nc).
- The corrected control used (usually gotten through Algorithm 3).

2.3 Global-to-Local Mapper (GLMap)

The second step of TTMap first calculates distances and provides a visualization of these distances and relations in the dataset, using the Mapper algorithm (section 1.2.2). As explained in section 1.2.9, different parameters need to be chosen: a filter function, a distance or a similarity measure, a cover and a closeness parameter ε.

2.3.1 The distance

The default similarity measure in GLMap is the *mismatch distance*, d_M given by a sum of mismatches, where a mismatch between two samples X and Y in a gene is a gene that is differentially expressed in opposite direction for X and Y as measured by the deviation component (Fig. 2.4, $n = 1$). The deviation must be bigger than α to avoid counting noise as mismatch. The **mismatch distance**, or sum of mismatches is defined as follows (Fig. 2.4),

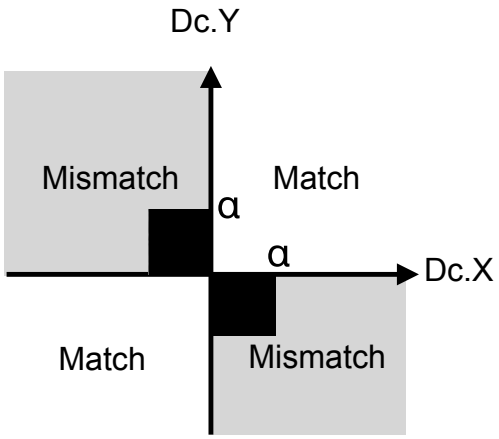


Figure 2.4: Scheme defining a match and a mismatch between two deviations components (Dc) of test samples X and Y with cutoff α to remove noise close to 0, $n = 1$. The mismatch distance between two samples is the sum of mismatches through all the genes.

for a fixed $\alpha \geq 0$

$$d_M(X,Y) = \sum_{i=1}^{n} d_m((Dc.X)_i, (Dc.Y)_i), \text{ where}$$

$$d_m(x,y) = \begin{cases} 0 & \text{if } \operatorname{sign}(x) = \operatorname{sign}(y), \\ 1 & \text{if } \operatorname{sign}(x) \neq \operatorname{sign}(y) \\ & \text{and } |x| \text{ or } |y| \geq \alpha \\ \frac{|x-y|}{8\alpha n} & \text{otherwise,} \end{cases}.$$

where

$$\operatorname{sign} : \mathbb{R} \to \{-1, 0, 1\}, x \mapsto \begin{cases} 0 & \text{if } x = 0, \\ 1 & \text{if } x > 0, \\ -1 & \text{if } x < 0. \end{cases}$$

For the theory to work and without impinging on the practical results we will consider a slightly modified version of the mismatch distance on our datasets defined by $d^*(X,Y) = d_M(X,Y) + d_{\bar{E}}(X,Y)$, where $d_{\bar{E}}(X,Y)$ is the bounded euclidean distance by 1/4 (see 2.4.2). If features measured are gene expression values, then the default value does not need to be changed and is set to $\alpha = 1$, corresponding to a 2-fold-change, which is a standard cut-off for gene expression.

Remark 2.3.1. This is the default distance, which is recommended for gene expression and verifies the hypothesis of the theoretical stability theorems. However, any distance matrix can be imputed. Alternative distances, such as correlation distance, Euclidean distance, useful when there is no control group, and complete mismatch distance, a stringent version of the mismatch distance defined above, are implemented in GLMap and can be selected by the user. Of note, in those cases the parameter ε described hereafter in the subsection 2.3.4 needs to be adapted and has no appropriate default value. The mismatch distance is appropriate for gene expression data, since it captures deviation of samples from the control values with the same orientation, regardless of the magnitude of deviation.

2.3.2 The filter function

Furthermore, GLMap uses a **filter function**, given by properties of interest of the samples. It can be chosen by the user to take into account relevant variables, such as the age of the patients in a cohort. The default filter function in GLMap, called **total absolute deviation** and denoted τ, measures the overall deviation of a test vector from the control, i.e.,

$$\tau : \mathsf{T} \to \mathbb{R} : T_k \mapsto \sum_{l \in S} \mid (Dc.T_k)_l \mid,$$

where S is a subset of features, determined by the user, the default being to select all features, and T is the set of test vectors, which is a subset of $\mathbb{R}^n$.

Remark 2.3.2. This is the default filter function. If any other metadata or function would be of interest to do local clustering, it can be imputed by a vector of the same length as the number of samples.

2.3.3 The cover of the codomain of the filter function

Let $\operatorname{Im}\tau$ denote the **image of τ with multiplicity**, i.e.,

$$\operatorname{Im}\tau = \{(\tau(X), \sigma) \mid X \in \mathsf{T},\ \sigma \in \{1, \dots, \operatorname{mult}(X)\}\} \subseteq \mathbb{R} \times \mathbb{N},$$

with the lexicographic order, where $\operatorname{mult}(X) = \operatorname{card}(\tau^{-1}(\tau(X)))$ is the multiplicity of $\tau(X)$ and for any $0 \leq a < b \leq 100$, let

$$q_{[a,b[} = \pi_1\Big(\Big\{y \in \operatorname{Im}\tau \mid \operatorname{quantile}_a(\operatorname{Im}\tau) \leq y < \operatorname{quantile}_b(\operatorname{Im}\tau)\Big\}\Big),$$

where π_1 is the natural projection on the first component, and $\operatorname{quantile}_a(\operatorname{Im}\tau)$ is the a-th quantile of the ordered values in $\operatorname{Im}\tau$.

The chosen cover of the image without multiplicity ($\{\tau(X) \mid X \in \mathsf{T}\}$ is given by

$$\mathfrak{I} = \{\pi_1 \operatorname{Im} \tau, q_{[0,25[}, q_{[25,50[}, q_{[50,75[}, q_{[75,100]}\}.$$

2.3.4 The epsilon parameter

Intuition 2.3.3. The ε parameter is estimated by the data. We use probabilities to determine, given the variability of the control group, what is the expected amount of mismatches between two samples that would be distributed as the controls. This expected amount, called ε, will give us the cutoff. If the two samples are distributed as the control then we expect less than ε mismatches between those samples. And if both should be clustered together, they are distributed as the controls in almost all but N genes, the significant genes, where $n >> N$ and is therefore negligible, and these two samples are going in the same direction (positive or negative) in terms of deviations for those N genes and will not contribute to the total mismatches. The other $n - N$ genes have the same distribution as the control and therefore it implies that the mismatch distance of these two samples should be lower than ε.

Assuming that the two vectors X and Y follow the same normal distribution $\mathscr{N}(\mu_i, \sigma_i^2)$ for feature i, the parameter ε is estimated using the data. Feature by feature the probability to be a mismatch is calculated. Let X_k be the random vector representing the gene expressions of a sample T_k. Therefore, let

$p_{1j,k} = P(X_{kj} < \min_{i \in \mathscr{I}(T_k)} (\overline{N}_i)_j)$, the probability to be underexpressed compared to normal values

$p_{2j,k} = P(X_{kj} > \max_{i \in \mathscr{I}(T_k)} (\overline{N}_i)_j)$, the probability to be overexpressed compared to normal values

$p_{3j,k} = P(\min_{i \in \mathscr{I}(T_k)} (\overline{N}_i)_j < X_{kj} < \max_{i \in \mathscr{I}(T_k)} (\overline{N}_i)_j)$, the probability to be inside the normal range

Then, we define

$$p^{\alpha}_{1j,k} = P(X_k < \min_{i \in \mathscr{I}(T_k)} (\overline{N}_i)_j - \alpha),$$

$$p^{\alpha}_{2j,k} = P(X_k > \max_{i \in \mathscr{I}(T_k)} (\overline{N}_i)_j + \alpha).$$

Hence the probability $(P^l_k)_j$ of a mismatch between the j-th gene of (X_k, X_l) is equal to : $((p_{3j,k} + p_{1j,k}) \cdot p^{\alpha}_{2j,l}) + (p_{3j,k} + p_{2j,k}) \cdot p^{\alpha}_{1j,l}) + ((p_{3j,l} + p_{1j,l}) \cdot p^{\alpha}_{2j,k}) + (p_{3j,l} + p_{2j,l}) \cdot p^{\alpha}_{1j,k}) - p^{\alpha}_{1j,k} \cdot p^{\alpha}_{2j,l} - p^{\alpha}_{2j,k} \cdot p^{\alpha}_{1j,l}$, where for example $((p_{3j,k} + p_{1j,k}) \cdot p^{\alpha}_{2j,l})$ would represent the probability that X_k for gene j is either as the control ($p_{3j,k}$) or lower than the control ($p_{1j,k}$) whereas X_l is marginally (more than alpha) higher than the control ($p^{\alpha}_{2j,l}$), and so it represents a mismatch. Using Chen-Stein's theorem, it is known that if $n >> S$, and if the probabilities accumulated around 0 as is the case for gene expression data, then the sum over all genes of mismatches follows a Poisson distribution with mean $\sum_{j=1}^{n} (P^l_k)_j$. This in turn allows one to determine how significant the number of mismatches between X and Y is, if both vectors follow the same distribution. Hence, ε is given by $P(\sum_{j=1}^{n} (P^l_k)_j < \varepsilon) = \beta$, which can be obtained from the quantiles of a Poisson law. Thus, samples are linked if the number of mismatches between them is less than ε, which is the β% confidence threshold of mismatches for samples following the same distribution.
If ε is chosen such that $P(\sum_{j=1}^{n} (P^l_k)_j < \varepsilon) = 0.025$, it means that only in 2.5% of the cases

if X_k and X_l are distributed in the same way, they would have such a small number of mismatches and therefore it is certain that X_k and X_l must be clustered together. In the same way, if ε is chosen such that $P(\sum_{j=1}^{n}(P_k^l)_j < \varepsilon) = 0.975$, it means that only in 97.5% of the cases if X_k and X_l are distributed in the same way, they would have such a high number of mismatches and therefore it is certain that X_k and X_l must be separated. The user can therefore choose either to cluster samples together only if one is sure that samples should be clustered together (0.025) or choose to separate samples only if one is sure that samples need to be separated or finally the user has the option to put another value for parameter ε, when the % of mismatches to be expected is already known.

2.3.5 The algorithm Global-to-local Mapper

This part is based on the Mapper algorithm (see section 1.2.2) [11] and can be obtained using the function *ttmap()* in **R** and can be summarized by the algorithm.

Algorithm 5: Global-to-local Mapper

Input: A matrix corresponding to the deviation components.
Optional arguments: distance matrix, filter function, ε parameter.
Method: In default mode, GLMap applies the Mapper algorithm (Algorithm 1) to the quadruple given by
- the mismatch distance d_{M} or a distance received as input,
- a closeness parameter ε, computed from the data, see 2.3.4, which depends on the variance in the control group, or is entered as input,
- the total absolute deviation τ or another filter function given as input,
- the cover of $\mathrm{Im}\,\tau$ given by $\mathfrak{I}$ representing a global tier, the full $\mathrm{Im}\,\tau$,

and the local tiers, the quartiles of the function.
Output: 3 output files are created :
- the distance matrix.
- a visual representation of the clustering, giving subgroups in the test samples, that can be zoomed in and out.
- description of the clusters of the visual representation, with information.

This means that GLMap performs single-linkage clustering (see section 1.2.2) with distance d_M and parameter of closeness ε to every $\tau^{-1}(U)$ such that $U \in \mathfrak{I}$.

- all of T, giving the connected components $\{C_{01}, \ldots, C_{0l(0)}\}$ of the graph G_ε defined by the vertex set $\{T_k\}$ and the edge set $\{(T_a, T_b)$ s.t. $d_M(T_a, T_b) < \varepsilon\}$ and then to
- the pre-image with respect to τ of each of the quantiles $q_{0,25}, q_{25,50}, q_{50,75}$, and $q_{75,100}$, which gives the connected components $\{C_{i1}, \ldots, C_{il(i)}\}$ of the subgraph $G_\varepsilon(i) = \tau^{-1}(I_i)$, where $I_i \in \mathfrak{I}$.

Two connected components C_{ij} and C_{kl} are represented as spheres with diameters increasing with the number of samples in each component. The spheres are connected by an edge whenever $C_{ij} \cap C_{kl} \neq \emptyset$, i.e. the algorithm links clusters that share samples as every sample is assessed twice for connectivity, once globally and once within its quartile, links are formed between local and global structures, enabling the discovery of subgroups based on the filter

function of the global clusters (Fig. 2.1, Part2).

The color of a sphere in the output figure of the method (see example in section 3.3, Fig. 3.8a) is determined by the average of the values of the filter function applied to the samples in the bin. A legend for the color code is provided at the bottom of the output figure, for the size of the balls on the right, and for the different tiers on the left, i.e. the overall clustering and the clustering in the different quartiles, (Fig. 2.1, Part2). A list of the differentially expressed genes per cluster is provided by genes that are all deviating into the same direction or that are as the control, and the deviation is larger than α.

2.3.6 Global-to-local Mapper through a toy example

Example 2.3.4. We recall example 1.2.37. We associate to it a function given by a color code The Mapper algorithm (see section 1.2.2) performs a clustering algorithm on the elements of the pullback of the chosen cover. In this case the pullback of the chosen cover is given by

- all of T, giving the connected components $\{C_{01}, \ldots, C_{0l(0)}\}$ of the graph G_ε defined by the vertex set $\{T_k\}$ and the edge set $\{(T_a, T_b) \text{ s.t. } d_M(T_a, T_b) < \varepsilon\}$ and then to
- the pre-image with respect to τ of each of the quantiles $q_{0,25}, q_{25,50}, q_{50,75}$, and $q_{75,100}$, which gives the connected components $\{C_{i1}, \ldots, C_{il(i)}\}$ of the subgraph $G_\varepsilon(i) = \tau^{-1}(I_i)$, where $I_i \in \mathfrak{I}$.

So one cluster is obtained per connected component in the ε-neighbourhood graph, this corresponds to the classical single-linkage clustering algorithm and we color-code each connected component shown by a sphere of the size of the number of samples by average amount of the filter function (Fig. 2.5a-d top). Then, single-linkage clustering is performed on the subset given by the points in one quartile, e.g. the lower quartile (Fig. 2.5a), the 2nd quartile (Fig. 2.5b), the 3rd quartile (Fig. 2.5c), the higher quartile (Fig. 2.5d). The Mapper algorithm adds higher order simplices if clusters share a sample. In this case the only possible overlap is between the global and the local structure, as each sample will only be in one quartile of the filter function, which only gives the possibility to have 1-simplices, i.e. the edges. These will be added to understand to which global structure the local structure belongs to (Fig. 2.5e). Once this output is spatially reordered, we obtain the TTMap output (Fig.2.5f).

Remark 2.3.5. An example to illustrate the utility would be to imagine the function to be the age of the patient in a cohort, from dark blue, the young patients to dark red the old patients. Only having the information of the global cluster, we would see (Fig. 2.5a-d top), that there are two cluster with average age of the samples a "light green" value. However, only the decomposition in terms of quantiles reflects that one cluster was composed only of "light green"-aged patients, whereas the other sample is constituted of one younger patient and one older patient. This sheds light into the composition of the global clusters.

Remark 2.3.6. As the global clusters are unaffected by the filter function since they are only affected by the distance and the epsilon parameter, if the filter function is changed only the clusters in the quartiles are changed. This gives a basis of comparison for all the different filter functions.

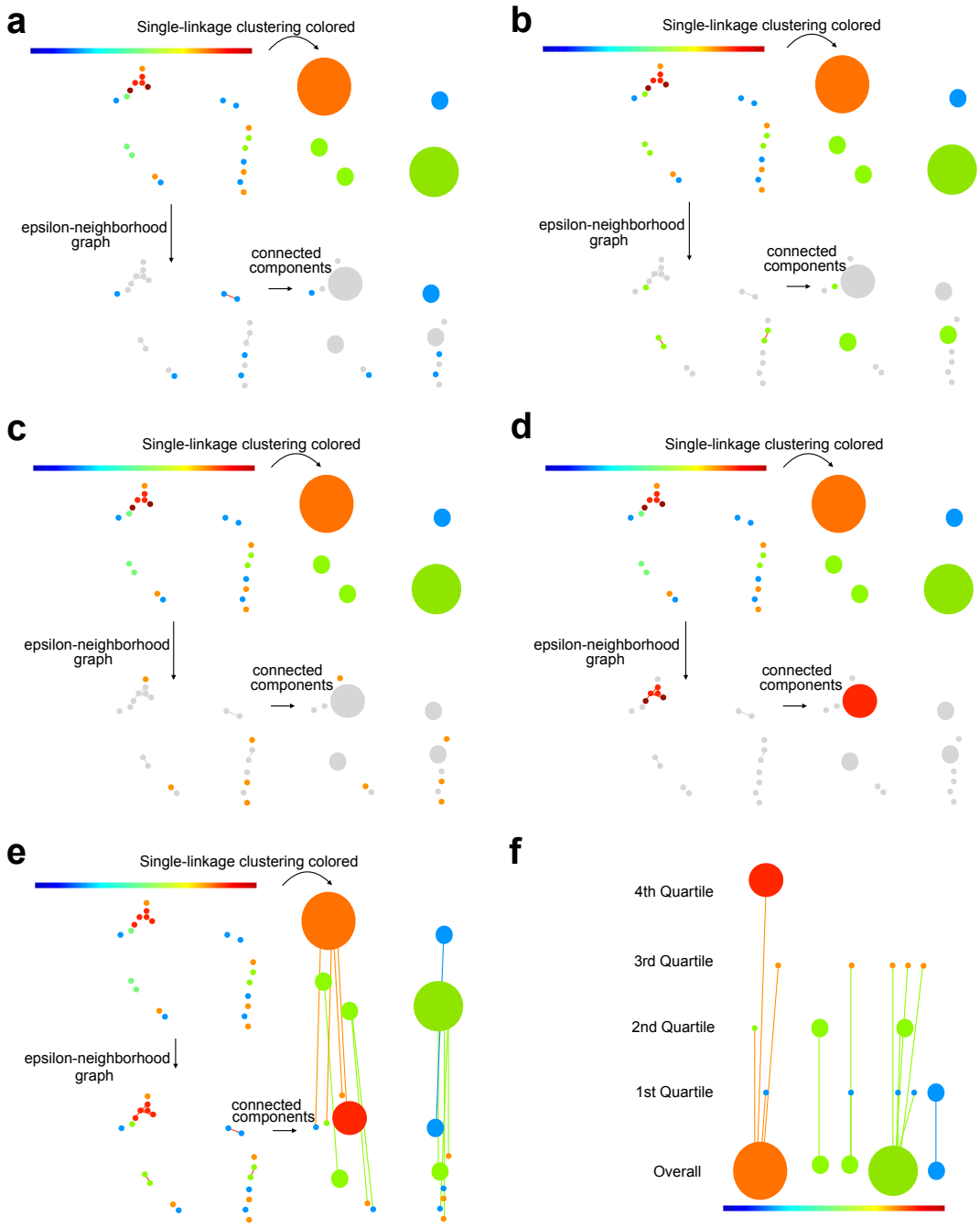


Figure 2.5: Example of TTMap on a point cloud with an associated filter function, given by the colors on the points. The data with the single linkage clustering associated to it, the cluster obtained are colored by the average amount of the filter function. The data is subselected for the **(a)** lower quartile points, from dark blue to aquamarine, **(b)** 2nd quartile points, from light blue to light green, **(c)** 3rd quartile points, from yellow to orange, **(d)** higher quartile points, from red to dark red, then the ε-neighbourhood graph of this selected subset is computed and from these graphs the connected components are taken. This corresponds to perform the single-linkage clustering algorithm on the subset of points only. **(e)** The Mapper algorithm generates links between the global and the local spheres. **(f)** Once it is rearrange spatially and annotated we obtain the TTMap output.

2.4 Theoretical aspects

To assess the stability of TTMap theoretically, we studied the effects of modifications of the source space, of the filter function and of approximations with a point cloud on its output. The absence of a natural distance on the outputs of TTMap precludes direct assessment of the stability of the TTMap graphs. Therefore, we summarized the information contained in the TTMap graphs as a diagram in $\mathbb{R}^2$, similar to a persistence diagram (PD) [202], where there is a natural distance d that generalizes the distance on PD, enabling a comparison of TTMap graphs.
The PDs summarize the topological features of the data such as connected components, holes, branches, and dots. We supplemented PDs with links between the "local" features and the connected components or global clusters, forming a descriptor, denoted $DM(X, f, \mathfrak{I})$, for a space X and a filter function $f : X \to \mathbb{R}$ that verifies mild regularity conditions. In terms of these enriched PDs, we establish the following theorems, stated informally here and precisely in Theorems 2.4.2, 2.4.4, 2.4.5, 2.4.6, respectively.

- *Completeness* The descriptor is complete: the information contained in the graph of $TTMap(X, f, \mathfrak{I})$ can be recovered from the diagram $DM(X, f, \mathfrak{I})$.
- *Stability with respect to changes of the filter function* If the filter function f is perturbed, the distance between the diagrams of f and of its perturbation is not greater than the amount of perturbation.
- *Stability with respect to perturbations of the domain* If the starting space X is perturbed, then the distance between the diagrams of X and of its perturbation depends linearly on the amount of perturbation.
- *Stability with respect to point cloud approximations*

  If data points are sampled on a space X, then the difference between the diagrams associated to X and to the δ-neighborhood graph built on the point cloud is less than a value depending on δ.

Thus, the three stability theorems state that the method is stable upon modifications of the source space, of the filter function, and upon approximations with a point cloud.

In this section, all functions are assumed to be of Morse type, whose definition and example can be found in section 1.3.4 [203]. It is a mild assumption as a wide variety of function verify to be of Morse type, especially the filter function chosen in the algorithm TTMap verifies to be of Morse type, which will be proven in section 2.4.2 along with the verification of all assumptions made in the stability theorems.
This assumption is technical and needed since in that case we have a well defined extended persistence diagram (see section 1.3.4, theorem 1.3.27). It assures that the mathematical objects dealt with are well-defined.

2.4.1 Generalized structure of TTMap

Let X be a topological space and let $f : X \to \mathbb{R}$ be a Morse-type function. Consider a family of pairwise disjoint intervals of $\mathbb{R}$ with non-empty interiors, such that the union of all the

intervals is still an interval. Add $\mathbb{R}$ to this family and call the result $\mathscr{I}$. Considering the class of Morse-type pairs (X, f) such that $\mathscr{I}$ is a cover of $\mathrm{Im}(f)$, our aim is to study the structure of $\mathrm{M}(X, f, \mathscr{I})$ and its stability with respect to perturbations of (X, f) within this class. Note that, $TTMap(Dc.T, \tau, \mathscr{I})$ is a special case of $\mathrm{M}(\mathrm{P}, \mathrm{f}, \mathscr{I})$, where P is given by Dc.T and X is the corresponding underlying support, f is given by τ and $\mathscr{I}$ is given by the quantiles q_{ab} and the real line and δ is given by the parameter ε (section 2.3).

Definition 2.4.1. We define the following descriptor for $\mathrm{M}(X, f, \mathscr{I})$:

$$\mathrm{DM}(X, f, \mathscr{I}) := (\mathrm{Dg}\,(\tilde{f}), \varphi, \{\Delta_I\}_{I \in \mathscr{I}}),$$

where:

- $\mathrm{Dg}\,(\tilde{f})$ is the extended persistence diagram of the Reeb graph of X using the induced function $\tilde{f}$ (see section 1.3.4).
- $\varphi : \mathrm{Dg}\,(\tilde{f}) \to \mathrm{Ext}_0^+(\tilde{f})$ maps a persistence pair (i.e. (a, b) where a and b are the birth and the death time respectively of a topological feature to the connected component of X to which its corresponding feature belongs,
- $\Delta_I = \{(x, x) \mid x \in I\}$ is the diagonal subset of $I \times I$.

Intuitively, $\mathrm{M}(X, f, \mathscr{I})$ can be reconstructed from $\mathrm{DM}(X, f, \mathscr{I})$ in 3 steps (Fig. 2.6a, b, c and d):

1. Create one super-node per point in $\mathrm{Ext}_0^+(\tilde{f})$.
2. For each interval $I \in \mathscr{I}$, create one node per point $(x, y) \in \mathrm{Dg}\,(\tilde{f})$ such that I is contained entirely in the lifespan of (x, y), which is materialized in the descriptor $\mathrm{DM}(X, f, \mathscr{I})$ by the fact that the line segment $\Delta_{(x,y)}$ bounded by the horizontal and vertical projections of (x, y) onto the diagonal Δ contains Δ_I. If $(x, y) \in \mathrm{Ord}_0(\tilde{f}) \cup \mathrm{Rel}_1(\tilde{f}) \cup \mathrm{Ext}_0^+(\tilde{f})$ then create a vertex also if I contains x. If $(x, y) \in \mathrm{Ext}_0^+(\tilde{f})$ then create a vertex also if I contains y.
3. Draw the links prescribed by φ between the super-nodes and the rest of the nodes.

Theorem 2.4.2. Completeness. $\mathrm{DM}(X, f, \mathscr{I})$ *is a complete descriptor of* $\mathrm{M}(X, f, \mathscr{I})$*, i.e. the Mapper graph can be drawn from the information obtained by the descriptor.*

Proof. At any level $\alpha \in \mathbb{R}$, the following equality holds:

$$\begin{aligned} &\#\left\{C \ : \ C \text{ is a connected component of } \tilde{f}^{-1}(\{\alpha\})\right\} \ = \\ &\#\left\{(x, y) \in \mathrm{Dg}\,(\tilde{f}) \ : \ \alpha \in \text{lifespan}\,(x, y)\right\}, \end{aligned} \tag{2.1}$$

where:

$$\text{lifespan}\,(x, y) \ = \ \begin{cases} [x, y] & \text{if } (x, y) \in \mathrm{Ext}_0^+(\tilde{f}) \\ (y, x) & \text{if } (x, y) \in \mathrm{Ext}_1^-(\tilde{f}) \\ [x, y) & \text{if } (x, y) \in \mathrm{Ord}_0(\tilde{f}) \\ (y, x] & \text{if } (x, y) \in \mathrm{Rel}_1(\tilde{f}) \end{cases}$$

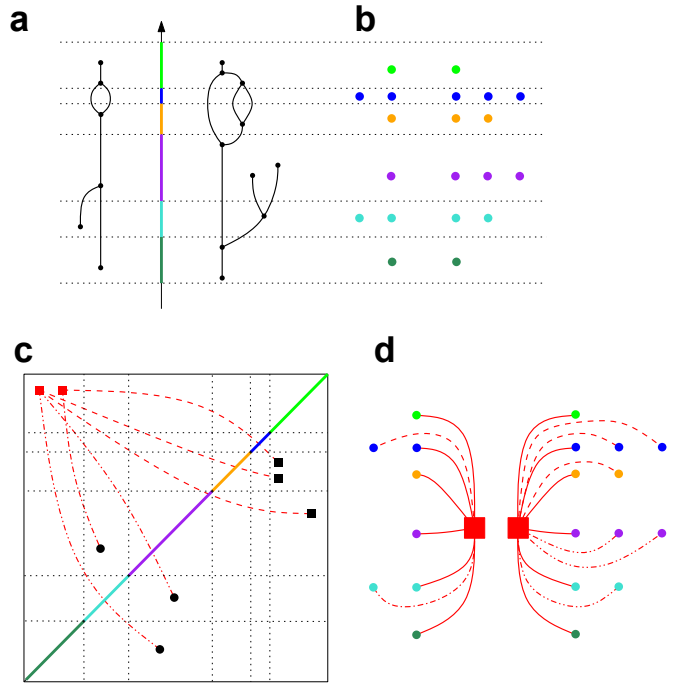


Figure 2.6: **(a)** The Reeb graph of an object X composed out of two components with a function f, the height function, and its Mapper **(b)** computed with a cover of $\mathrm{Im}(f)$ with disjoint intervals represented by different colors. **(c)** By adding $\mathbb{R}$ to this cover and calling it $\mathscr{I}$, the descriptor $\mathrm{DM}(X, f, \mathscr{I})$ is calculated and **(d)** the Mapper can be retrieved from the descriptor in **(c)**.

Indeed, let $\alpha \in \mathbb{R}$. Assume for simplicity that $\alpha \notin \mathrm{Crit}(f)$ (if $\alpha \in \mathrm{Crit}(f)$ then the same analysis holds with the extra technicality that the type of each interval endpoint, open or closed, must be taken into account). Define the following quadrants (Fig. 2.7):

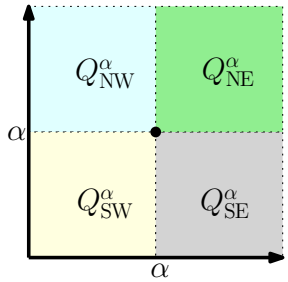


Figure 2.7: Plot of the various Q^α_* in the plane.

$$
\begin{aligned}
Q^\alpha_{\mathrm{NW}} &= \{(x,y) \in \mathbb{R}^2 \ : \ x \leq \alpha \text{ and } y \geq \alpha\} \\
Q^\alpha_{\mathrm{NE}} &= \{(x,y) \in \mathbb{R}^2 \ : \ x \geq \alpha \text{ and } y \geq \alpha\} \\
Q^\alpha_{\mathrm{SW}} &= \{(x,y) \in \mathbb{R}^2 \ : \ x \leq \alpha \text{ and } y \leq \alpha\} \\
Q^\alpha_{\mathrm{SE}} &= \{(x,y) \in \mathbb{R}^2 \ : \ x \geq \alpha \text{ and } y \leq \alpha\}
\end{aligned}
$$

Since points in $\mathrm{Ord}_0(\tilde{f})$ and $\mathrm{Ext}^+_0(\tilde{f})$ are located above the diagonal and points in $\mathrm{Ext}^-_1(\tilde{f})$

and $\mathrm{Rel}_1(\tilde{f})$ are located below, proving Equation (2.1) amounts to showing that

$$\begin{aligned}\dim\left(H_0\left(\tilde{f}^{-1}(\{\alpha\})\right)\right) &= |\mathrm{Ord}_0(\tilde{f}) \cap Q^{\alpha}_{\mathrm{NW}}| \\ &+ |\mathrm{Ext}^+_0(\tilde{f}) \cap Q^{\alpha}_{\mathrm{NW}}| + |\mathrm{Ext}^-_1(\tilde{f}) \cap Q^{\alpha}_{\mathrm{SE}}| + |\mathrm{Rel}_1(\tilde{f}) \cap Q^{\alpha}_{\mathrm{SE}}|.\end{aligned} \tag{2.2}$$

For this the Mayer-Vietoris theorem is used with spaces $A = \tilde{f}^{-1}((-\infty, \alpha])$, $B = \tilde{f}^{-1}([\alpha, +\infty))$, $A \cap B = \tilde{f}^{-1}(\{\alpha\})$, and $A \cup B = \mathrm{R}_f(X)$. This theorem can be used because the Morse-type condition implies that A, B are deformation retracts of neighborhoods A', B' in $\mathrm{R}_f(X)$ with $A' \cap B'$ deformation retracting onto $A \cap B$. Hence, the following sequence is exact:

$$\begin{aligned}H_2(\mathrm{R}_f(X)) &\xrightarrow{\partial_2} H_1\left(\tilde{f}^{-1}(\{\alpha\})\right) \xrightarrow{\varphi} \overbrace{H_1\left(\tilde{f}^{-1}((-\infty,\alpha])\right) \oplus H_1\left(\tilde{f}^{-1}([\alpha,+\infty))\right)}^{K_1} \\ &\xrightarrow{\psi} H_1(\mathrm{R}_f(X)) \xrightarrow{\partial_1} H_0\left(\tilde{f}^{-1}(\{\alpha\})\right) \xrightarrow{\zeta} \underbrace{H_0\left(\tilde{f}^{-1}((-\infty,\alpha])\right) \oplus H_0\left(\tilde{f}^{-1}([\alpha,+\infty))\right)}_{K_0} \\ &\xrightarrow{\xi} H_0(\mathrm{R}_f(X)) \xrightarrow{\partial_0} 0\end{aligned}$$

To be more specific, exactness gives the following relations:

$$\mathrm{Im}(\partial_2) = \ker(\varphi) \tag{2.3}$$
$$\mathrm{Im}(\partial_1) = \ker(\zeta) \tag{2.4}$$
$$\mathrm{Im}(\varphi) = \ker(\psi) \tag{2.5}$$
$$\mathrm{Im}(\zeta) = \ker(\xi) \tag{2.6}$$
$$\mathrm{Im}(\psi) = \ker(\partial_1) \tag{2.7}$$
$$\mathrm{Im}(\xi) = \ker(\partial_0) \tag{2.8}$$

It follows from (2.8) and from [58] that

$$\dim(\mathrm{Im}(\xi)) = \dim(\ker(\partial_0)) = \dim(H_0(\mathrm{R}_f(X))) = |\mathrm{Ext}^+_0(\tilde{f})|. \tag{2.9}$$

Moreover, according to Theorem 2.9 in [204], we have $H_p(\mathrm{R}_f(X)) = 0$ for any $p \geq 2$. Using equation (2.3), it follows that $\mathrm{Im}(\partial_2) = 0 = \ker(\varphi)$, hence

$$0 = \dim\left(H_1\left(\tilde{f}^{-1}(\{\alpha\})\right)\right) = \dim(\ker(\varphi)) + \dim(\mathrm{Im}(\varphi)) = \dim(\mathrm{Im}(\varphi)). \tag{2.10}$$

Using equations (2.4) to (2.10) and Theorem 2.5 in [204], the following equalities hold:

$$\begin{aligned}\dim\left(H_0\left(\tilde{f}^{-1}(\{\alpha\})\right)\right) &= \dim(\ker(\zeta)) + \dim(\mathrm{Im}(\zeta)) \\ &= \dim(\mathrm{Im}(\partial_1)) + \dim(\ker(\xi)) \\ &= \dim(H_1(\mathrm{R}_f(X))) - \dim(\ker(\partial_1)) + \dim(\ker(\xi)) \\ &= |\mathrm{Ext}^-_1(\tilde{f})| - \dim(\mathrm{Im}(\psi)) + \dim(\ker(\xi)) \\ &= |\mathrm{Ext}^-_1(\tilde{f})| - \dim(K_1) + \dim(\ker(\psi)) + \dim(\ker(\xi)) \\ &= |\mathrm{Ext}^-_1(\tilde{f})| - \dim(K_1) + \dim(\mathrm{Im}(\varphi)) + \dim(\ker(\xi)) \\ &= |\mathrm{Ext}^-_1(\tilde{f})| - \dim(K_1) + \dim(\ker(\xi)) \\ &= |\mathrm{Ext}^-_1(\tilde{f})| - \dim(K_1) + \dim(K_0) - \dim(\mathrm{Im}(\xi)) \\ &= |\mathrm{Ext}^-_1(\tilde{f})| - \dim(K_1) + \dim(K_0) - |\mathrm{Ext}^+_0(\tilde{f})|\end{aligned}$$

It remains to compute $\dim(K_1)$ and $\dim(K_0)$. Using the correspondence between connected components and branches of $\mathrm{R}_f(X)$ and points of $\mathrm{Dg}\,(\tilde{f})$ [58], it holds that

$$\begin{aligned}\dim(K_1) &= \dim\left(H_1\left(\tilde{f}^{-1}((-\infty,\alpha])\right)\right) + \dim\left(H_1\left(\tilde{f}^{-1}([\alpha,+\infty))\right)\right)\\ &= |\mathrm{Ext}_1^-(\tilde{f}) \cap Q^\alpha_{\mathrm{SW}}| + |\mathrm{Ext}_1^-(\tilde{f}) \cap Q^\alpha_{\mathrm{NE}}| \end{aligned} \tag{2.11}$$

and

$$\begin{aligned}\dim(K_0) &= \dim\left(H_0\left(\tilde{f}^{-1}((-\infty,\alpha])\right)\right) + \dim\left(H_0\left(\tilde{f}^{-1}([\alpha,+\infty))\right)\right)\\ &= |\mathrm{Ord}_0(\tilde{f}) \cap Q^\alpha_{\mathrm{NW}}| + |\mathrm{Ext}_0^+(\tilde{f}) \cap (Q^\alpha_{\mathrm{NW}} \cup Q^\alpha_{\mathrm{SW}})|\\ &+ |\mathrm{Rel}_1(\tilde{f}) \cap Q^\alpha_{\mathrm{SE}}| + |\mathrm{Ext}_0^+(\tilde{f}) \cap (Q^\alpha_{\mathrm{NW}} \cup Q^\alpha_{\mathrm{NE}})|. \end{aligned} \tag{2.12}$$

Combining these results, we obtain

$$\begin{aligned}\dim\left(H_0\left(\tilde{f}^{-1}\{\alpha\}\right)\right) = {} & |\mathrm{Ext}_1^-(\tilde{f})| - |\mathrm{Ext}_1^-(\tilde{f}) \cap Q^\alpha_{\mathrm{SW}}| - |\mathrm{Ext}_1^-(\tilde{f}) \cap Q^\alpha_{\mathrm{NE}}| + |\mathrm{Ord}_0(\tilde{f}) \cap Q^\alpha_{\mathrm{NW}}|\\ & + |\mathrm{Rel}_1(\tilde{f}) \cap Q^\alpha_{\mathrm{SE}}| + |\mathrm{Ext}_0^+(\tilde{f}) \cap (Q^\alpha_{\mathrm{NW}} \cup Q^\alpha_{\mathrm{SW}})|\\ & + |\mathrm{Ext}_0^+(\tilde{f}) \cap (Q^\alpha_{\mathrm{NW}} \cup Q^\alpha_{\mathrm{NE}})| - |\mathrm{Ext}_0^+(\tilde{f})|\\ = {} & |\mathrm{Ext}_1^-(\tilde{f}) \cap Q^\alpha_{\mathrm{SE}}| + |\mathrm{Ord}_0(\tilde{f}) \cap Q^\alpha_{\mathrm{NW}}| + |\mathrm{Rel}_1(\tilde{f}) \cap Q^\alpha_{\mathrm{SE}}|\\ & + |\mathrm{Ext}_0^+(\tilde{f}) \cap Q^\alpha_{\mathrm{NW}}|,\end{aligned}$$

which gives (2.2) and thus proves Equation (2.1).
The theorem is proved using the three steps of the reconstruction scheme detailed before the statement 2.4.2.
According to the one-to-one correspondence between the connected components of $\mathrm{R}_f(X)$ and the points of $\mathrm{Ext}_0^+(\tilde{f})$, Step 1 ensures that there are as many super-nodes as there are connected components in $\mathrm{R}_f(X)$.
Equation (2.1) can be extended to intervals at no cost to prove that the number of vertices created in Step 2 and the number of nodes in $\mathrm{M}(X, f, \mathscr{I})$ (apart from the super-nodes) is the same.
Finally, each node v of $\mathrm{M}(X, f, \mathscr{I})$ corresponds to some connected component of the preimage $f^{-1}(I)$ of some interval $I \in \mathscr{I}$. That connected component lies entirely in some connected component X_i of X, therefore v gets connected to the super-node corresponding to X_i in $\mathrm{M}(X, f, \mathscr{I})$. This is the only type of connection that matters for $\mathrm{M}(X, f, \mathscr{I})$, since every pair of intervals other than $\mathbb{R}$ in $\mathscr{I}$ has an empty intersection. Since the connected component corresponding to v belongs to at least one feature of $\mathrm{R}_f(X)$, or equivalently one persistence pair of $\mathrm{Dg}\,(\tilde{f})$, this proves that the links prescribed by φ in Step 3 and the ones of $\mathrm{M}(X, f, \mathscr{I})$ are the same. □

This result states that whenever two descriptors are the same, their corresponding TTMap graphs must also be the same and therefore for what follows the results are shown in terms of diagrams.

Stability theorems

Note that $\{\Delta_I\}_{I\in\mathscr{I}}$ induces the grid $(\mathrm{End}(\mathscr{I} \setminus \mathbb{R}) \times \mathbb{R}) \cup (\mathbb{R} \times \mathrm{End}(\mathscr{I} \setminus \mathbb{R}))$, (Fig. 2.6 c). Intuitively, the distances of the points of $\mathrm{Dg}\,(\tilde{f})$ to this grid give the amount of perturbation

allowed to preserve the structure of $\mathrm{M}(X, f, \mathscr{I})$. Reciprocally, for a given amount of perturbation ε, drawing a square of radius ε around each diagram point allows us to see which diagram points may change grid cells and how the structure of $\mathrm{M}(X, f, \mathscr{I})$ is impacted.

Definition 2.4.3. Let f, g be two Morse-type functions defined on topological spaces X, Y. The **descriptor distance** between $\mathrm{DM}(X, f, \mathscr{I})$ and $\mathrm{DM}(Y, g, \mathscr{I})$ is:

$$d(\mathrm{DM}(X, f, \mathscr{I}), \mathrm{DM}(Y, g, \mathscr{I})) = \inf_{\Gamma} \ cost(\Gamma),$$

where Γ ranges over all partial matchings between $\mathrm{Dg}\,(\tilde{f})$ and $\mathrm{Dg}\,(\tilde{g})$ such that $(p, p') \in \Gamma \Rightarrow (\varphi(p), \varphi(p')) \in \Gamma$.

This definition is illustrated in the following figure (Fig. 2.8)

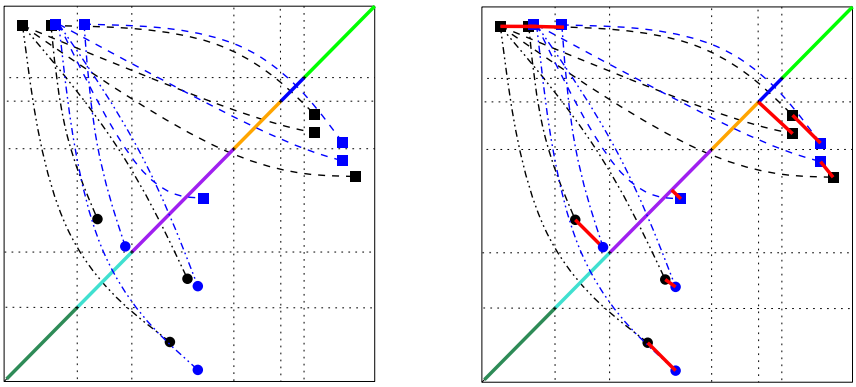


Figure 2.8: Illustration of the descriptor distance on two extended persistence diagrams between two diagrams in blue and black.

Theorem 2.4.4. Stability with respect to changes of the filter function. *For any Morse-type functions* $f, g : X \to \mathbb{R}$*:*

$$d(\mathrm{DM}(X, f, \mathscr{I}),\ \mathrm{DM}(X, g, \mathscr{I})) \leq \|f - g\|_\infty.$$

Proof. Decompose X into its various connected components: $X = X_1 \sqcup X_2 \sqcup ... \sqcup X_n$, and let $f_i := f|_{X_i} : X_i \to \mathbb{R}$ and $g_i := g|_{X_i} : X_i \to \mathbb{R}$. Note that $\mathrm{Dg}\,(f) = \mathrm{Dg}\,(f_i) \sqcup ... \sqcup \mathrm{Dg}\,(f_n)$, and similarly for g and the induced maps $\tilde{f}$ and $\tilde{g}$. Thus, one can build a matching Γ that preserves connected components by taking any matching for each pair of subdiagrams $\mathrm{Dg}\,(f_i), \mathrm{Dg}\,(g_i)$. For instance, let us take for each pair $\mathrm{Dg}\,(f_i), \mathrm{Dg}\,(g_i)$ the matching achieving $d(\mathrm{DM}(X_i, f_i, \mathscr{I}),\ \mathrm{DM}(X_i, g_i, \mathscr{I}))$. Call it Γ_i, and let $\Gamma = \bigcup_i \Gamma_i$. Hence, the following

inequalities hold:

$$
\begin{aligned}
& d(\mathrm{DM}(X, f, \mathscr{I}),\ \mathrm{DM}(X, g, \mathscr{I})) \leq \mathrm{cost}(\Gamma) \\
& \leq \max_{i \in \{1,\ldots,n\}} \mathrm{cost}(\Gamma_i) \\
& = \max_{i \in \{1,\ldots,n\}} d(\mathrm{DM}(X_i, f_i, \mathscr{I}),\ \mathrm{DM}(X_i, g_i, \mathscr{I})) \\
& = \max_{i \in \{1,\ldots,n\}} d_{\mathrm{b}}^{\infty}(\mathrm{Dg}\,(\tilde{f}_i), \mathrm{Dg}\,(\tilde{g}_i)) \text{ since } X_i \text{ is connected} \\
& \leq \max_{i \in \{1,\ldots,n\}} \|\tilde{f}_i - \tilde{g}_i\|_{\infty} \text{ by the stability theorem [205]} \\
& = \|\tilde{f} - \tilde{g}\|_{\infty} \\
& = \|f - g\|_{\infty} \text{ since the quotient maps } \tilde{f} \text{ and } \tilde{g} \text{ preserve function values.}
\end{aligned}
$$

□

Theorem 2.4.5. Stability with respect to perturbations of the domain. *Let X and Y be two compact Riemannian manifolds or length spaces with curvature bounded above. Denote by $\rho(X)$ and $\rho(Y)$ their respective convexity radii. Let $f : X \to \mathbb{R}$ and $g : Y \to \mathbb{R}$ be Lipschitz-continuous Morse-type functions, with Lipschitz constants c_f and c_g respectively. Assume $\mathrm{d_{GH}}(X, Y) \leq \frac{1}{20} \min\{\rho(X), \rho(Y)\}$. Then, for any correspondence $C \in \mathscr{C}(X, Y)$ such that $\varepsilon_{\mathfrak{m}}(C) < \frac{1}{10} \min(\rho(X), \rho(Y))$,*

$$d(\mathrm{DM}(X, f, \mathscr{I}),\ \mathrm{DM}(Y, g, \mathscr{I})) \leq (9(c_f + c_g) + \min\{c_f, c_g\})\varepsilon_{\mathfrak{m}}(C) + \varepsilon_{\mathfrak{f}}(C),$$

where $\varepsilon_{\mathfrak{m}}(C) = sup_{(x,y),(x',y') \in C} \mid d_X(x, x') - d_Y(y, y') \mid$ and $\varepsilon_{\mathfrak{f}}(C) = sup_{(x,y) \in C} \mid f(x) - g(y) \mid$ are the distance distortion and the functional distortion [206].

Proof. If there is a one-to-one matching between the connected components of X and Y induced by the correspondence achieving $\mathrm{d_{GH}}(X, Y)$, then the proof follows the same line as the proof of Theorem 2.4.4. The only difference in the proof is the use of Theorem 3.4 in [207] instead of the stability theorem [205]. If such a one-to-one matching does not exist, $\mathrm{d_{GH}}(X, Y)$ is infinite and so is $\varepsilon_{\mathfrak{m}}(C)$, hence

$$d(\mathrm{DM}(X, f, \mathscr{I}),\ \mathrm{DM}(Y, g, \mathscr{I})) \leq (9(c_f + c_g) + \min\{c_f, c_g\})\varepsilon_{\mathfrak{m}}(C) + \varepsilon_{\mathfrak{f}}(C),$$

still holds.

□

Theorem 2.4.6. Stability with respect to point cloud approximations. *Let X be a submanifold of $\mathbb{R}^d$ with positive reach $r(X)$ (largest number such that any point at distance less than $r(X)$ from X has a unique nearest point on X) and convexity radius $\rho(X)$. Let $f : X \to \mathbb{R}$ be a Lipschitz-continuous Morse-type function, with Lipschitz constant c. Let $P \subseteq X$ be such that every point of X lies within distance ε of P, for some $\varepsilon < \min\{r(X)/16,\ \rho(X)/16,\ s/8c\}$, where $s > 0$ is the minimum distance of the points of $\mathrm{Ext}_1(f)$ to the diagonal Δ. Let $\delta \in [4\varepsilon,\ \min\{r(X)/4,\ \rho(X)/4,\ s/2c\})$, and $G_\delta(P)$ be the δ-neighborhood graph built on top of P with parameter δ. Then, the following inequality holds:*

$$d\left(\mathrm{DM}(X, f, \mathscr{I}),\ \mathrm{DM}(G_\delta(P), \hat{f}, \mathscr{I}))\right) \leq 2c\delta,$$

where $\hat{f}$ is the piecewise linear interpolation of f along the edges of $G_\delta(P)$ [208].

Proof. Let $C_X = \min\{\|x - x'\|_d : x, x' \text{ do not belong to the same connected component of } X\}$, and let $x, x' \in X$ be two points achieving C_X. Let $y = \frac{1}{2}(x + x') \in \mathbb{R}^d$. Then $\|x - y\|_d \geq r(X)$ since y belongs to the medial axis of X. Hence, $C_X = 2\|x - y\|_d \geq 2r(X)$. Since $\delta < \frac{1}{4}r(X) < C_X$, it follows that X and $\mathrm{Rips}_\delta^1(P)$ have the same number of connected components. Then, the proof of Theorem 2.4.6 follows the same line as the proof of Theorem 2.4.4. The only difference in the proof is the use of Theorem 7.5 in [207] instead of the stability theorem [205]. □

2.4.2 Hypotheses verification for TTMap

In order to use those theorems, one needs to verify their hypotheses. Hence, the topology induced by the distance d^* should verify that it is equivalent to the euclidean distance to be able to use the last theorem. Moreover, the function need to be Lipschitz in order to use the theorems 2.4.5 and 2.4.6. Lastly, f needs to be of Morse-type in order to use all of the theorems of stability (2.4.2, 2.4.4, 2.4.5, 2.4.6).

For that, we will proceed in several steps : Let x, y be two deviation components, whence $x, y \in \mathbb{R}^n$. Then, $d^*(x, y) = d_M(x, y) + \bar{d_E}(x, y)$, where $\bar{d_E}(x, y)$ is the Euclidean distance bounded by $1/4$ and $d_M(x, y)$ is given by $d_M(x, y) = \sum_{i=1}^n d_{m_i}(x_i, y_i)$, where

$$d_{m_i}(x_i, y_i) = \begin{cases} 0 & \text{if } \mathrm{sign}(x_i) = \mathrm{sign}(y_i), \\ 1 & \text{if } \mathrm{sign}(x_i) \neq \mathrm{sign}(y_i) \\ & \text{and } |x_i| \text{ or } |y_i| \geq \alpha \\ \frac{|x_i - y_i|}{8\alpha n} & \text{otherwise} \end{cases} \tag{2.13}$$

We observe that even if all the values are noise smaller than α (around 0), then the $d(x, y) < 1/2$, and therefore not perturbing the results if we replace ε by $\varepsilon + 1/2$ in the corresponding section. We will prove that with this distance

1. we can define a topology.

2. we verify that $(\mathbb{R}^n, \mathscr{T}_{d^*}) = (\mathbb{R}^n, \mathscr{T}_{\bar{E}})$, which is the topology with the bounded euclidean distance and which is known to be the same as $(\mathbb{R}^n, \mathscr{T}_E)$, the standard topology with the euclidean distance.

3. the following function is Lipschitz:

$$\begin{array}{rcl} f : (\mathbb{R}^n, \mathscr{T}_{d^*}) & \to & (\mathbb{R}, \mathscr{T}_E) \\ x = (x_1, \ldots, x_n) & \mapsto & \sum_{i=1}^n |x_i|. \end{array}$$

4. the function f is Morse-type.

1. Let us show that $\{B_{d^*}(x, \varepsilon) \mid \varepsilon > 0, x \in \mathbb{R}^n\}$ defines a base of a topology. Indeed, for every $x \in \mathbb{R}^n$ and $x \in B_{d^*}(x, 1/2)$ so the first axiom is verified. Secondly, let $x, y \in \mathbb{R}^n$ be two vectors and δ and ε two real numbers then, let

$$t \in B_{d^*}(x, \delta) \cap B_{d^*}(y, \varepsilon).$$

Let

$$\nu = \min \Bigg\{ \begin{aligned} &\min\{|t_i| \mid t_i \neq 0, i = 1, \ldots, n\}, \\ &\min\{|\alpha - |t_i|| \mid t_i \neq \alpha, i = 1, \ldots, n\}, \\ &1/4, \delta - d^*(x,t), \varepsilon - d^*(y,t) \end{aligned} \Bigg\} > 0.$$

We want to show that $B_{d^*}(t,\nu) \subseteq B_{d^*}(x,\delta) \cap B_{d^*}(y,\varepsilon)$. The proof is the same for y and x just replacing ε by δ. Let us therefore focus on showing that $B_{d^*}(t,\nu) \subseteq B_{d^*}(x,\delta)$.

Let $z \in B_{d^*}(t,\nu)$. Hence, $d^*(z,t) < \nu$ and therefore

$$d_M(z,t) = \sum_{i=1}^{n} d_{m_i}(z_i,t_i) < \nu$$

since $d_{m_i}(z_i,t_i) \geq 0$ this means that $d_{m_i}(z_i,t_i) < \nu$ for every $i = 1, \ldots, n$.

Lemma (A) : If $t_i \neq 0$, then $z_i \neq 0$ and $\text{sign}(z_i) = \text{sign}(t_i)$.

Proof. Since, $d^*(z,t) \leq \nu$, then $d_{\bar{E}}(z,t) \leq \nu$.

As $\nu < 1/4$, $d_{\bar{E}}(z,t) = d_E(z,t)$ and hence $\sum_{i=1} |z_i - t_i| < \nu$. This implies that $|z_i - t_i| < \nu$ for every $i = 1, \ldots, n$.

Moreover, since $|t_i| \neq 0$, we have that$|z_i - t_i| < \nu \leq |t_i|$. Therefore, $z_i \neq 0$ because otherwise we get $|t_i| < |t_i|$, which is a contradiction. Moreover if $t_i > 0$ and $z_i < 0$, then $|z_i - t_i| = t_i - z_i = |t_i| + |z_i|$, since z_i is negative. This can not be strictly smaller than $|t_i|$ otherwise we get $|z_i| < 0$ which is a contradiction.
Similarly if $t_i < 0$ and $z_i > 0$, then $|z_i - t_i| = z_i - t_i = |z_i| + |t_i|$, which can not be strictly smaller than $|t_i|$. Therefore, z_i and t_i must have the same signature.

□

Lemma (B) : If $|t_i| \neq \alpha$ either $|t_i|$ and $|z_i|$ are $> \alpha$ or $|t_i|$ and $|z_i| < \alpha$.

Proof. By the above argument, we know that $|z_i - t_i| < \nu$ and $\nu \leq| \alpha - |t_i| |$. Let us suppose $|t_i| > \alpha$ then $| \alpha - |t_i| | = |t_i| - \alpha$, and $|z_i - t_i| < |t_i| - \alpha$ implies $-|z_i - t_i| > -|t_i| + \alpha$, which results in

$$|z_i| \geq |t_i| - |z_i - t_i| > |t_i| - |t_i| + \alpha = \alpha.$$

Hence, $|z_i| > \alpha$.

Let us suppose $|t_i| < \alpha$ then $| \alpha - |t_i| | = \alpha - |t_i|$, and

$$|z_i| \leq |z_i - t_i| + |t_i| < \alpha - |t_i| + |t_i| = \alpha$$

and hence $|z_i| < \alpha$. □

Let us enumerate the cases :

- $H = \{i \in \{1, \ldots, n\} \mid |t_i| = \alpha\}$.
- $I = \{i \in \{1, \ldots, n\} \mid i \notin H,\ \text{sign}(t_i) = 0\}$.
- $J = \{i \in \{1, \ldots, n\} \mid i \notin H \cup I,\ \text{sign}(x_i) = \text{sign}(t_i)\}$.
- $K = \{i \in \{1, \ldots, n\} \mid i \notin H \cup I,\ \text{sign}(x_i) \neq \text{sign}(t_i), |t_i| < \alpha\}$.
- $L = \{i \in \{1, \ldots, n\} \mid i \notin H \cup I,\ \text{sign}(x_i) \neq \text{sign}(t_i), |t_i| \geq \alpha\}$.

Let us calculate,

$$d_M(x, z) = \sum_{i=1}^{n} d_{m_i}(x_i, z_i) = \sum_{i \in H} d_{m_i}(x_i, z_i) + \sum_{i \in I} d_{m_i}(x_i, z_i) + \sum_{i \in J} d_{m_i}(x_i, z_i) + \sum_{i \in K} d_{m_i}(x_i, z_i) + \sum_{i \in L} d_{m_i}(x_i, z_i)$$

- For $i \in H$, there are two cases:
  - If $d_{m_i}(x_i, t_i) = 0$, then by Lemma A, we have that $\text{sign}(t_i) = \text{sign}(z_i)$ and therefore $\text{sign}(z_i) = \text{sign}(x_i)$, and therefore $d_{m_i}(x_i, z_i) = d_{m_i}(x_i, t_i) = 0$.
  - If $d_{m_i}(x_i, t_i) = 1$, then $d_{m_i}(x_i, z_i) < d_{m_i}(x_i, t_i) = 1$.
- For $i \in I$ since $t_i = 0$ there are several scenarios :
  - $|x_i| \geq \alpha$: in this case either z_i and x_i have the same signature and then $d_{m_i}(x_i, z_i) = 0 < d_{m_i}(x_i, t_i)$ or the have opposite signatures and then $d_{m_i}(x_i, z_i) = 1 = d_{m_i}(x_i, t_i)$. In both cases $d_{m_i}(x_i, z_i) \leq d_{m_i}(x_i, t_i)$.
  - $0 < |x_i| < \alpha$: if $\text{sign}(z_i) = \text{sign}(x_i)$, then $d_{m_i}(x_i, z_i) = 0 < d_{m_i}(x_i, t_i)$, otherwise by Lemma B as $t_i < \alpha$ we have that z_i is smaller than α as well and hence $\text{sign}(z_i) \neq \text{sign}(x_i)$ then $d_{m_i}(x_i, z_i) = \frac{|x_i - z_i|}{8n\alpha} = \frac{|x_i|}{8n\alpha} + \frac{|z_i|}{8n\alpha} = d_{m_i}(x_i, t_i) + d_{m_i}(t_i, z_i)$.
  - $|x_i| = 0$ then $t_i = x_i$ and $d_{m_i}(x_i, z_i) = d_{m_i}(t_i, z_i)$
- For $i \in J$ since $\text{sign}(t_i) = \text{sign}(x_i)$ and from Lemma A, we know that $\text{sign}(t_i) = \text{sign}(z_i)$. Therefore, $d_{m_i}(x_i, z_i) = d_{m_i}(x_i, t_i) = 0$.
- For $i \in K$, then $|t_i| < \alpha$, and we know from Lemma B. that this implies $|z_i| < \alpha$ as well. We again have two cases here :
  - $|x_i| \geq \alpha$, $d_{m_i}(x_i, z_i) = 1 = d_{m_i}(x_i, t_i)$.
  - $|x_i| < \alpha$, then $d_{m_i}(x_i, z_i) = \frac{|x_i - z_i|}{8n\alpha} \leq \frac{|x_i - t_i|}{8n\alpha} + \frac{|z_i - t_i|}{8n\alpha} = d_{m_i}(x_i, t_i) + \frac{|t_i - z_i|}{8n\alpha} = d_{m_i}(x_i, t_i) + d_{m_i}(t_i, z_i)$.
- For $i \in L$, since $|t_i| \geq \alpha$, we know from Lemma B that $|z_i| \geq \alpha$ as well, which implies that $d_{m_i}(x_i, t_i) = 1 = d_{m_i}(x_i, z_i)$.

Put together we have that

$$\begin{aligned}
d_M(x,z) =& \sum_{i=1}^{n} d_{m_i}(x_i,z_i) \\
=& \sum_{i\in H} d_{m_i}(x_i,z_i) + \sum_{i\in I} d_{m_i}(x_i,z_i) + \sum_{i\in J} d_{m_i}(x_i,z_i) + \sum_{i\in K} d_{m_i}(x_i,z_i) + \sum_{i\in L} d_{m_i}(x_i,z_i) \\
\leq& \sum_{i\in H} d_{m_i}(x_i,t_i) + \sum_{i\in I} d_{m_i}(x_i,t_i) + \sum_{i\in J} d_{m_i}(x_i,t_i) + d_{m_i}(t_i,z_i) + \sum_{i\in K} d_{m_i}(x_i,t_i) \\
&+ d_{m_i}(t_i,z_i) + \sum_{i\in L} d_{m_i}(x_i,t_i) \\
\leq& d_M(x,t) + d_M(t,z)
\end{aligned}$$

Hence, $d^*(x,z) = d_M(x,z) + d_{\bar{E}}(x,z) \leq d_M(x,t) + d_{\bar{E}}(x,t) + d_M(t,z) + d_{\bar{E}}(t,z) = d^*(x,t) + d^*(t,z) \leq d^*(x,t) + \delta - d^*(x,t) = \delta$.

2. " $\supseteq$ " Let $\varepsilon > 0$ and let $x \in \mathbb{R}^n$ if $\delta = \varepsilon > 0$ then

$$B_{d^*}(x,\delta) \subseteq B_{\bar{E}}(x,\varepsilon).$$

Indeed, if $y \in B_{d^*}(x,\delta)$, then $d^*(x,y) < \delta$ and hence, $d_{\bar{E}}(x,y) \leq d_M(x,y) + d_{\bar{E}}(x,y)$, since $d_M(x,y) \geq 0$ for every x, y and hence $d_{\bar{E}}(x,y) \leq d^*(x,y) < \delta = \varepsilon$. Therefore, $y \in B_{\bar{E}}(x,\varepsilon)$.
" $\subseteq$ " Let $\varepsilon > 0$ and let $x \in \mathbb{R}^n$ if $\delta = \min(\alpha/2, 1/4, \varepsilon/(\frac{1}{8\alpha} + 1)) > 0$ then

$$B_{\bar{E}}(x,\delta) \subseteq B_{d^*}(x,\varepsilon).$$

Indeed, if $y \in B_{\bar{E}}(x,\delta)$, then $d_{\bar{E}}(x,y) < \delta$ and since $\delta < \alpha/2$, and $\delta < 1/4$, then for every $i \in \{1,\ldots,n\}$ either $\text{sign}(x_i) = \text{sign}(y_i)$ or $\text{sign}(x_i) \neq \text{sign}(y_i)$ and both $|x_i|$ and $|y_i|$ are less than or equal to α.
We use the fact that $d_{\bar{E}}(x,y) < 1/4$ implies $d_{\bar{E}}(x,y) = d_E(x,y)$. Then, $d_{\bar{E}}(x,y) < \alpha/2$ implies that $\sum_{i=1}^{n} |x_i - y_i| < \alpha/2$ and therefore $|x_i - y_i| < \alpha/2$.
If $\text{sign}(x_i) \neq \text{sign}(y_i)$, then either $x_i > 0$ and $y_i < 0$, implying that $|x_i - y_i| = x_i - y_i > x_i = |x_i|$ and therefore $|x_i| < \alpha/2$. This in turn implies that $|y_i| < |x_i - y_i| + |x_i| < \alpha/2 + \alpha/2 = \alpha$, or $y_i > 0$ and $x_i < 0$ which with the same reasoning shows that $|x_i|$ and $|y_i|$ are smaller than α. Coming back to the original problem, we obtain either $d^*(x,y) = d_{\bar{E}}(x,y)$ when $\text{sign}(x_i) = \text{sign}(y_i)$ or

$$d^*(x,y) = d_M(x,y) + d_{\bar{E}}(x,y) = \sum_{i\in I} \frac{|x_i - y_i|}{8\alpha n} + d_{\bar{E}}(x,y),$$

and since the L^1 norm is bounded by $\sqrt{n}$ times the L^2 norm, it is clear that

$$\frac{|x_i - y_i|}{8\alpha n} \leq \frac{\sqrt{n}}{8\alpha n} \cdot d_{\bar{E}}(x,y) \leq \frac{1}{8\alpha} \cdot d_{\bar{E}}(x,y).$$

Therefore,

$$d^*(x,y) \leq \frac{1}{8\alpha} \cdot \delta + \delta \leq \varepsilon,$$

where $I = \{i \in \{1, \ldots, n\} \mid \text{sign}(x_i) \neq \text{sign}(y_i) \text{ and } |x_i| \text{ and } |y_i| \leq \alpha\}$. Therefore, $y \in B_{d^*}(x, \varepsilon)$.

3. f is Lipschitz since
$$f : (\mathbb{R}^n, \mathscr{T}_{d^*}) = (\mathbb{R}^n, \mathscr{T}_E) \to (\mathbb{R}, \mathscr{T}_E).$$
and hence
$$d(f(x), f(y)) = \mid f(x) - f(y) \mid = \mid \sum_{i=1}^{n} |x_i| - \sum_{i=1}^{n} |y_i| \mid \leq \sum_{i=1}^{n} \mid x_i - y_i \mid \leq d_E(x, y).$$

4. It is clearly of Morse-Type, since $f : (\mathbb{R}^n, \mathscr{T}_{d^*}) = (\mathbb{R}^n, \mathscr{T}_E) \to (\mathbb{R}, \mathscr{T}_E)$ is the L^1-norm. Each interval in $\mathbb{R}$ has as pre-image a void thickened diamond in $\mathbb{R}^n$, which is compact and locally connected. Since the thickening is given by the length of the interval, it is then straightforward to obtain the needed homeomorphism and conclude that it is of Morse-type.

2.5 Implementation

TTMap was implemented as an open-source **R** package in Bioconductor (https://www.bioconductor.org/packages/devel/bioc/html/TTMap.html).
To install this package, start **R** and enter:
source("https://bioconductor.org/biocLite.R")
biocLite("TTMap")

The reference manual, explaining all the functions, and the "vignettes" guiding the user through TTMap with an example, are available on the above mentioned webpage and in the Annexe of this thesis, Appendix A. Newest updates can be found on https://github.com/jeitziner/TTMap.

3 Applications of Two-tier Mapper

3.1 Introduction to the different sections

TTMap has been applied throughout the thesis on various types of datasets, from metabolic data, to neuronal spike data, passing by single-cell RNA-seq data and of course our main focus was on RNA-seq and microarrays. Each of these datasets has a different history and needs to be put in context. We used TTMap in each cases to answer specific biological question and compared it to standard methods. This chapter is therefore extremely various in content, and each section will be structured with an introduction to the problem possibly followed by a biological question, and then the results of TTMap are shown and discussed in the context of that experiment.

In order to understand and validate the functioning of TTMap and compare it to other methods, we started by applying it to *in silico* data sets that we generated (in **R**) (see section 3.2) and afterwards applied it to two known biological experiments reflecting the two major types of datasets in gene expression which are microarrays (for drosophila data in section 3.3) and RNA-seq (for the data on the estrous cycle of mice section 3.4). These two datasets have been studied more extensively than the others. These three sections 3.2, 3.3, 3.4 have been adapted from : *"Two-Tier Mapper: a user-independent clustering method for global gene expression analysis based on topology", R. Jeitziner, M. Carrière, J. Rougemont, S. Oudot, K. Hess, and C. Brisken, 2017, arXiv: 1801.01841 [194], submitted to Bioinformatics.)*

3.2 *In silico* validation

TTMap was tested on simulated data that mimics a situation for which standard methods are weak, i.e., small sample size (n<20). Moreover, we generated differences arising from the same genes in the subgroups, deviating in opposite directions. Control samples in group C and test samples are generated each with $10'000$ features, where the test group is composed of two subgroups TA and TB that need to be found by TTMap and the method to which we compared it called Mclust [96]. The subgroups TA and TB have the same mean per gene as the mean in the control group, except for m genes for which the mean is Δ times higher for TA, respectively lower for TB. The m genes are true positives, whereas all the other features are true negatives. The accuracy of the method is estimated by simulating at least 30 datasets per condition and calculating the percentage of times TTMap finds the right subgroups, establishing the clustering power of this method. Since TTMap is an analytical workflow, we also assessed its performance in finding the genes that are differentially expressed.

Goal. *We want to assess the performance of TTMap in finding two subgroups TA and TB on in silico generated data. Depending on the variance in the control group C, we evaluate the performance of the algorithm until the noise exceeds the signal. TTMap also finds differentially expressed genes and we will therefore determine the number of true positives and negatives according to the variance. We want to assess also the running time of the algorithm, as well as the accuracy of TTMap depending on the sample size or the subgroup sizes. We will describe how the parameter selections affect the ouput of TTMap, justifying our choices.*

3.2.1 Synthesised data generation

Since microarray gene expression data, is modelled as a normal distribution ([63]), the simulated data has been generated as follows. For a fixed natural number m less than $10'000$, K random lists of $10'000$ real numbers each are generated, where half of them $C_1, \dots C_{K/2}$ are the $K/2$ controls and the other half is divided by two $TA_1, TA_2, \dots, TA_{K/4}$, and $TB_1, TB_2, \dots, TB_{K/4}$, representing the test samples, each with $10'000$ genes. The subgroups TA and TB have a mean per gene that is Δ times higher, respectively lower than the mean of the control in m genes. Hence,

$$(C_1)_i, \dots, (C_{K/2})_i \in \mathcal{N}(\mu, \sigma^2)$$

for all $1 \leq i \leq 10,000$, and

$$(TA_1)_i, (TA_2)_i, \dots (TA_{K/4})_i, (TB_1)_i, (TB_2)_i, \dots, (TB_{K/4})_i \in \mathcal{N}(\mu, \sigma^2)$$

for all $1 \leq i \leq 10,000 - m$, while

$$(TA_1)_i, (TA_2)_i, \dots, (TA_{K/4})_i \in \mathcal{N}(\mu + \Delta, \sigma^2),$$

and

$$(TB_1)_i, (TB_2)_i, \dots, (TB_{K/4})_i \in \mathcal{N}(\mu - \Delta, \sigma^2)$$

for all $10,000 - m < i \leq 10,000$ and K is either 12, 200 or 400. The parameter μ is equal to 4 (the outputs do not change if it is another value), but Δ and σ vary in the different subsections hereafter and are therefore made precise there.

3.2.2 The performance of TTMap as a clustering method

The performance of TTMap was assessed, with the parameter ε given by the lowest 2.5 percentile (Fig. 3.1a) or the highest 2.5 percentile of the distribution of the distance d_M between two random variables (Fig. 3.1b). The variance σ^2 ranged from 0.01 to 1 in order to measure the accuracy of TTMap in situation ranging from low variance to high variance. The number of significant features m in the test cases were 50, 100, 1000, and 5000, i.e., 0.5, 1, 10, and 50% of all the features, respectively. When $\Delta = 2$, TTMap performed 100 % correctly when the variance in the control group was in the biologically encountered range [63] (Fig. 3.1a, b, pink shade), where $\sigma^2 < 0.3$ (Fig. 3.1a). For variances between 0.4 and 0.8 and for 0.5% and 1% of significant features respectively, the method could no longer distinguish between noise and signal ($\Delta = 2$) and classified all the samples as different. When ε is chosen in the higher 2.5 percentile (Fig. 3.1b), the method was less good than the lower 2.5 percentile when the variances are low (below 0.5), but much better for higher variances (greater than 0.5). Moreover, the higher the number of significant features, the better TTMap performs in finding the two subgroups. Performance also improved when Δ increased (Fig. 3.1a, Supplementary Fig. S8a).

In contrast, a standard clustering tool Mclust[96] that similarly to TTMap does not need any parameter selection, was unable to find the right groups (Fig. 3.1a, black line). This is in line with Mclust learning from the data, and hence requires a large enough sample size to be able to perform properly.

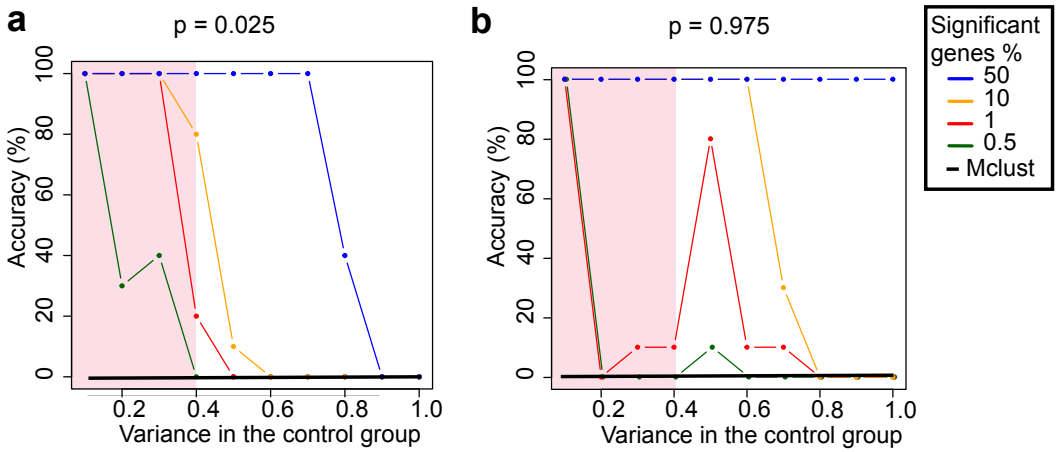


Figure 3.1: *In silico* validation of TTMap. Plot showing the accuracy of TTMap in percentage of times it correctly identifies subgroups of an *in silico* dataset over a range of different variances, $N > 30$. Individual curves were established for different percentages of significant genes. The accuracy of Mclust on the same dataset is shown in black, **(a)** using epsilon with probability 0.025 and **(b)** 0.975.

3.2.3 The running time of TTMap

The running time of TTMap follows a quadratic curve (Fig. 3.2a, blue curve), reflecting the fact that the algorithm is $\mathcal{O}(n^2)$ (Fig. 3.2). This is due to the distance function (the mismatch distance), which needs the computation of the distance for each pair of two points. The algorithm could be further fastened by optimizing this function. All the other parts of TTMap are growing almost linearly $\mathcal{O}(n)$ (for the control adjustment and HDA) and are fast (Fig. 3.2b, c) or even almost uniformly $\mathcal{O}(1)$ (for the global-to-local Mapper once the distance

has been calculated) (Fig. 3.2d). Moreover, on the *in silico* datasets tested, the running time of Mclust is 45 times longer than that of TTMap (3.8 minutes versus 5 seconds, respectively on $K = 12$).

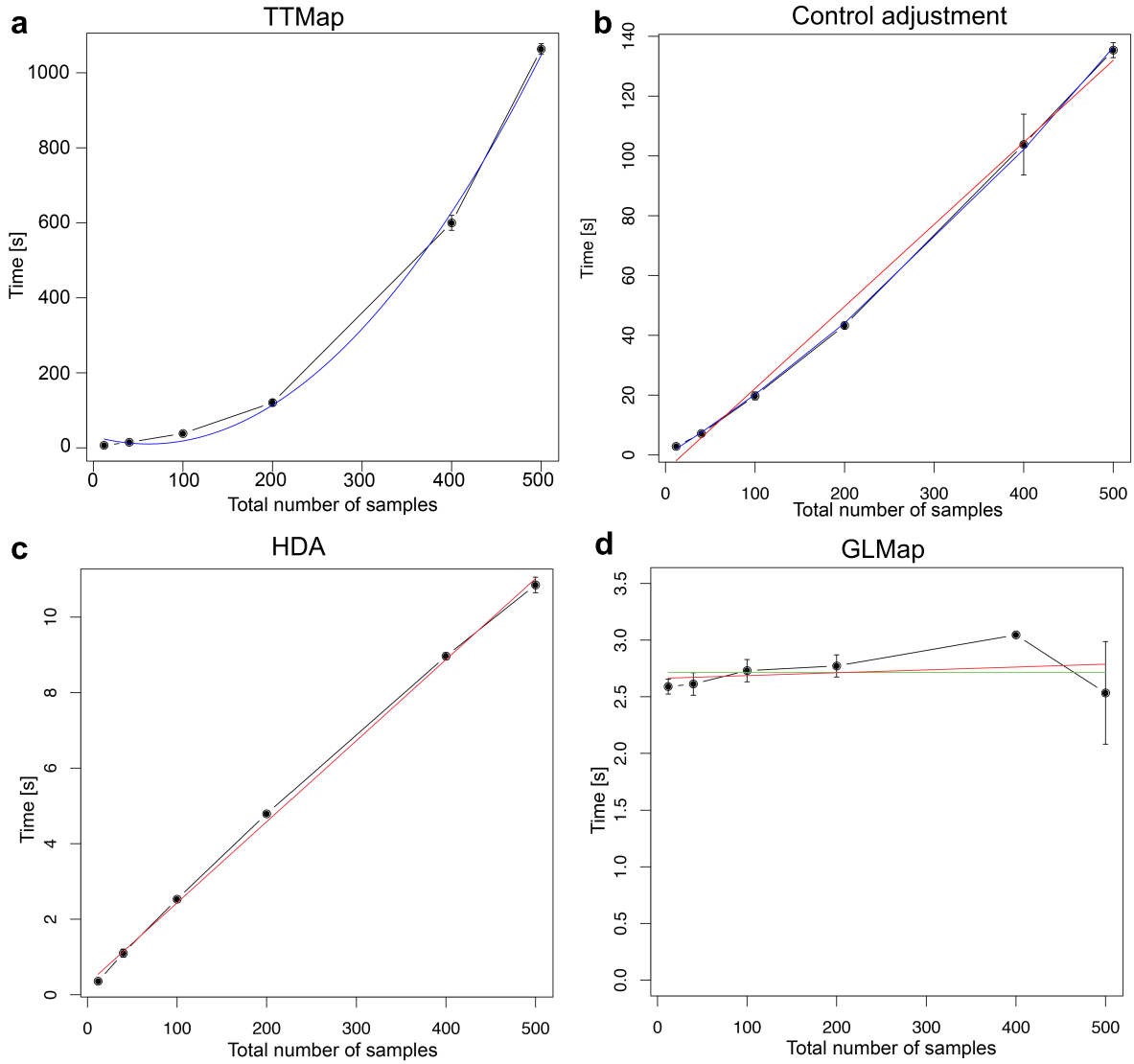


Figure 3.2: Time usage of TTMap. **(a)** Time in seconds used on *in silico* datasets of different sample size (black) with fitted quadratic curve (blue) of the full TTMap pipeline ($N = 10$, per sample size) **(b)** of only the control adjustement with fitted quadratic curve (blue) and with fitted linear curve (red) ($N = 3$, per sample size) **(c)** of only the hyperrectangle deviation assessment with fitted linear curve (red) ($N = 3$, per sample size) **(d)** of only the global-to-local Mapper after the distance has been calculated with fitted uniform curve (green) ($N = 3$, per sample size).

3.2.4 HDA and GLMap are both essential

To assess whether the accuracy of TTMap relies solely on HDA or on GLMap, we applied Mclust to the data obtained after HDA, i.e., the deviation components. The accuracy of Mclust in detecting the subgroups improved from 0 % to 20% on average (Fig. 3.3). Thus, the accuracy of Mclust improved but did not reach the level of accuracy of TTMap.

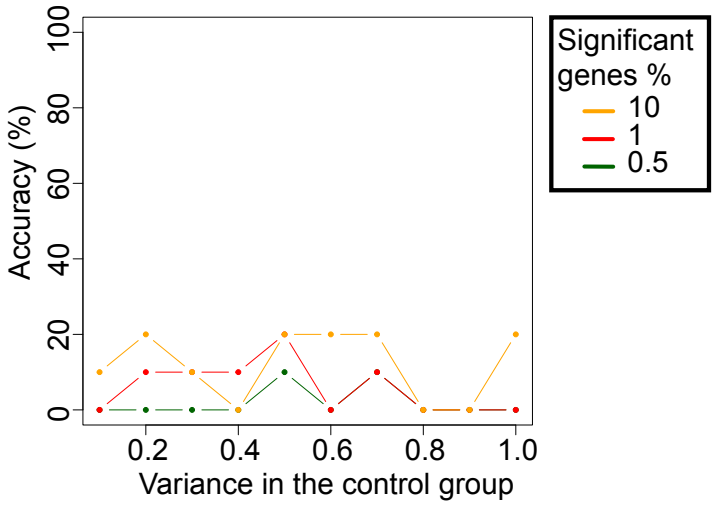


Figure 3.3: Plot showing the accuracy of Mclust on the deviation components with $N = 10$ per condition.

3.2.5 The performance of TTMap as a differential expression method in finding true positives and true negatives

To assess the performance of TTMap with regards to the genes determining a cluster, the numbers of true positives and of true negatives were computed with $\Delta = 2$, whenever the right groups are found. In datasets with low variance ($\sigma^2 < 0.5$) in the control group, TTMap found close to 100% of the true positives and true negatives (Fig. 3.4a, b). Since the samples

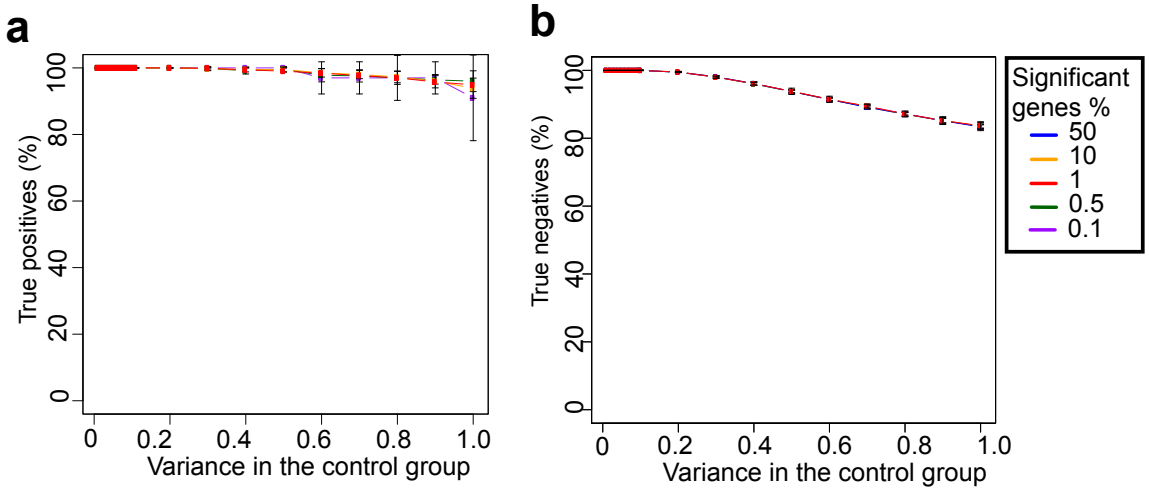


Figure 3.4: Plot showing the percentage of **(a)** true positives **(b)** true negatives when the right groups are found with $N > 30$ per condition.

in TA and TB have the same differentially expressed features but expressed in opposite directions, the moderated t-test did not detect any true positives. Even when the right groups are provided it poorly discovered the true positives in the subgroups, due to low sample size (Fig. 3.5). Together with the observation that the moderated t-test finds close to 100 % of true negatives, this suggests that the standard method is more likely to detect no significant

genes in such a situation, and is therefore dominated by TTMap.

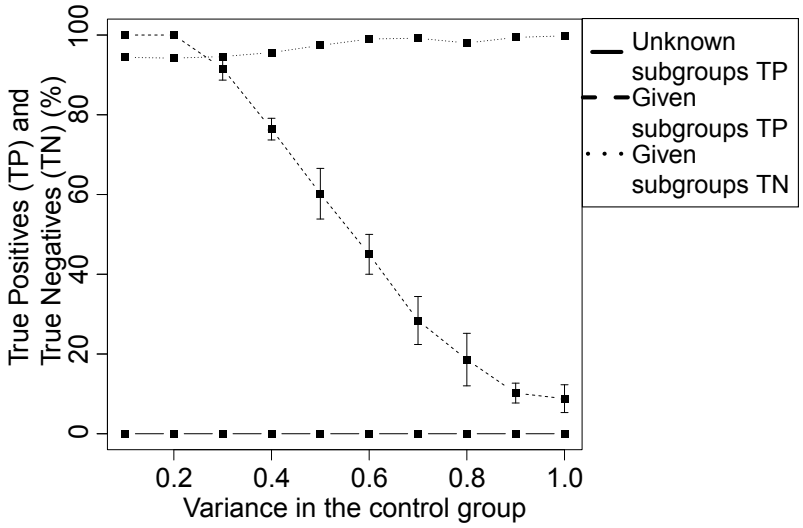


Figure 3.5: Plot showing the true positives (TP) and true negatives (TN) using moderated-t-test when the correct groups are given and when they are unknown.

3.2.6 The performance of TTMap on different sample sizes

TTMap was assessed on larger datasets as well consisting of 100 or 200 simulated samples. The method performed similarly at finding the right subgroups as in the case of small datasets (Supplementary Fig. S8c). In particular, for small variances ($\sigma^2 = 0 - 0.3$) the method's accuracy is above 98%, though it decreases for higher variances. Different sizes of subgroups TA and TB were generated, i.e., two samples in TA against four in TB and one in TA against five in TB respectively. Even if one of the subgroups is composed only of a single sample, the method accurately (more than 98% of accuracy for small variances) distinguishes the outlier sample from the rest of the samples (Supplementary Fig. S8b). Hence, TTMap is neither affected by the size of the point cloud nor by the sizes of the subgroups.

3.2.7 The output of TTMap upon changes of the mean of the features

TTMap is calculating deviation components from the control group, and therefore the output of TTMap does not change depending on the given average in the control group.

3.2.8 The parameter e

The effect of taking e as described in this thesis compared to selecting a standard confidence interval modifies considerably the values that are considered normal (Supplementary Fig. S9). Choosing $e = 1$ (corresponding to a 2 fold change, as the data is log transformed) or e as the 90-th percentile are the most reasonable choices as they are still reflecting the variability found in the samples.

3.2.9 The estimation of the parameter ε

The parameter ε is estimated using probabilities (see 2.3.4). As the data consists of a high number of genes $n >> S$, where S is the number of samples, two possibilities arise : either using Chen-Stein's theorem [209] and therefore estimate the amount of mismatches expected

by chance using a poisson approximation (see section 2.3.1) or use the Bernoulli law of large numbers stating that a normal approximation can be used to solve the problem. We estimated on simulated data which is the best approach and conlcuded that both work similarly (Supplementary Fig. S10). However, Chen-Stein's theorem is adaptable to the setting of not independent variables, which can be the case for gene expression analysis if a certain group of genes are dependent on each other. We therefore chose Chen-Stein as an approximation to ε knowing that TTMap can possibly be adapted to the case of known dependencies of genes [209].

3.3 Comparison of TTMap to standard clustering tools on biological data using the Fly atlas.

To further validate TTMap, we compared it to established clustering methods, k-means [95] and DBSCAN [92], on well-characterized biological data using the flyatlas (www.flyatlas.org). This dataset comprises microarray-based RNA expression profiles from 33 different drosophila tissues pooled from 50 male and 50 female flies or third instar feeding or wandering larvae, all in four replicates (Table 3.1). The 132 samples were compared to four replicate samples from the "whole adult fly" serving as control group N.

Name	Abbr.	Name	Abbr.
Adult Accessory gland	A	Adult Ovary	O
Adult Brain	B	Larval Feeding fat body	Qf
Adult Carcass	C	Larval Feeding Carcass	Cf
Adult Crop	R	Adult Salivary Gland	S
Adult Heart	D	Adult Spermatheca Mated 2	K3
Adult Eye	E	Larvae Wandering Tubules	Lw
Larval Feeding Hind Gut	Gf	Adult Testes	T
Adult Hind Gut	G	Adult Thoracic Muscle	V
Adult Head	H	Adult Trachea	X
Larval Feeding Mid Gut	Mf	Adult Thoracoabdominal ganglion	U
Larval Feeding Salivary Gland	Sf	Larval Feeding CNS	Nf
Adult Spermatheca Mated	K	Larval Wandering fat body	Qw
Adult Spermatheca Virgin	K2	Adult Wings	P
Adult Mid Gut	M	Whole Larvae Feeding	F
Adult Ejaculatory Duct	Z	Larval Feeding Trachea	Xf
Larval Feeding Malpighian Tubule	Lf	5th Passage Drosophila S2 Cells	Y
		Adult fat body	Q

Table 3.1: Legend used for the fly dataset, Abbr. = Abbreviation, CNS = Central Nervous System

Goal. *Using TTMap, tested in this section on microarray data, we interrogate how far each organ is from the whole fly globally on all genes, i.e. which organs are transcriptomically the closest and which one the furthest away from the whole fly. We might discover organs that overall have the same genes that change, but their extent of change is different, by looking at the distribution of the clusters in the quartiles. The dataset will also be used to assess the stability of the clusters.*

Remark 3.3.1. Even though the control group is small since it is only composed of 4 samples, the test group is large 132 samples, and the dataset represents a multiple comparison, where the multiple groups correspond to the different organs.

3.3.1 Comparison of TTMap to DBSCAN and k-means

For the standard methods parameters were chosen as to maximize their performance; k in k-means was set to 33, corresponding to the number of distinct tissues and *minPts* in DBSCAN was set to 4, reflecting the four replicates, providing an advantage to the methods, as it reduces the different possibilities of clusters generated. The ε parameter of DBSCAN was chosen according to guidelines [92] that we explained in section 1.5.2 (Fig. 3.6).

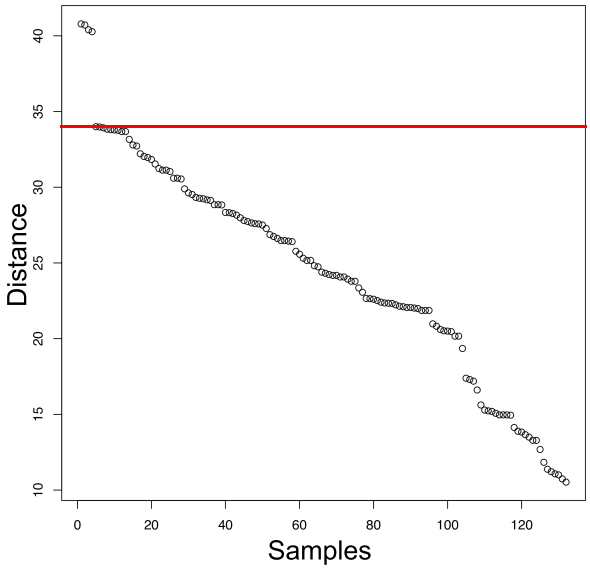


Figure 3.6: Determining the parameter ε of DBSCAN using guidelines of [92]. Points are displayed with the distance to their k-th nearest neighbor, where $k = 4$, the red line was chosen by visual inspection and shows one possible choice of ε as there is a gap in the distance between the previous points and the following points.

DBSCAN and k-means clustered 20 and 15 tissues, respectively, uniquely with their four replicates, i.e., clusters that consisted of four replicated samples (Fig. 3.7a, Supplementary Tables). TTMap, even though not provided with any parameter, clustered 21 tissues uniquely (Fig. 3.7a, Supplementary Tables).
To compare the stability of the different methods, the data were first quantile-normalized. Rand Index (RI), a measure of similarity between two clusterings [210], was 0.990 and 0.97, and 0.999 for DBSCAN, k-means and TTMap, respectively (Fig. 3.7a). However, quantile-normalization increased the number of uniquely clustering tissues to 21 with DBSCAN, to 22 with TTMap and decreased to 10 with k-means (Supplementary Tables, Fig. 3.7a). Next, we randomly selected 50% of the genes for re-clustering of the quantile-normalized data. DBSCAN's performance dropped to 12 (RI=0.86) due to the difficulty in finding the right ε parameter. K-means found 13 uniquely clustering organs (RI= 0.97). TTMap detected 20 uniquely clustering tissues under both conditions (RI=0.995) (Fig. 3.7b). Thus, TTMap is the most stable method upon normalization and random subselection and detects the maximum number of uniquely clustering organs.

3.3.2 Gained insights using TTMap

Overall, TTMap formed 32 global clusters (Fig. 3.8). The gene expression profiles of whole larvae (F) (Table 3.1) deviated the least (Fig. 3.8, cluster 1) and testis (T) and brain (B) the most from the whole adult fly that were considered as controls as indicated by the color code as well as their positions from left to right (Fig. 3.8, cluster 31 and 32).
Four clusters comprised samples from more than one tissue, while 6 clusters contained fewer than 4 replicates, and one cluster comprised four samples not all from the same tissue. The largest cluster (Fig. 3.8, cluster 16) contained the 4 replicates of virgin (K) and mated spermatacea (K2), as well as 3 replicates of the spermatacea redone (K3) along with a single

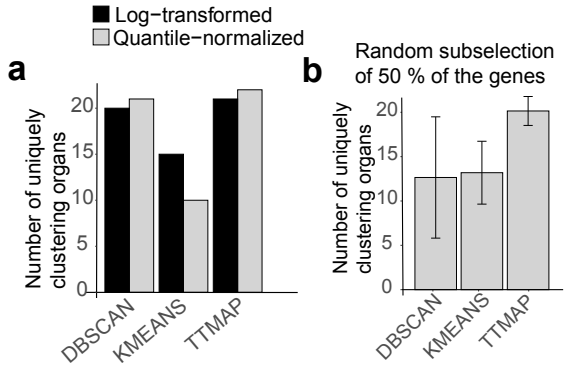


Figure 3.7: Stability of TTMap and comparison to standard clustering methods k-means, DBSCAN on the flyatlas. **(a)** Barplot representing the number of uniquely clustering organs, observed by clusters of four replicate samples, on log-transformed data and on quantile-normalised data using DBSCAN, k-means and TTMap. **(b)** Barplot representing the number of uniquely clustering organs when data are randomly subselected for 50% of the genes on quantile-normalised data using DBSCAN, k-means and TTMap.

replicate of the adult thoracic muscle (V). Interestingly, the 4^{th} replicate of K3 clustered with the 3 replicates of V, cluster 30, suggesting a labelling mistake, which may explain that standard tools (moderated t-test) revealed $< 10\%$ of the genes detected by TTMap (Supplementary Fig. S11a, b). On average, 84 % of the genes were found by both standard statistics and TTMap. However, specifically in the case of organ replicates that did not uniquely cluster with replicates of the same organ only, TTMap showed more power to detect significant genes.

Fat bodies from wandering and feeding larvae (Qw and Qf) clustered together globally (Fig. 3.8, cluster 13). Local Mapping using the filter function revealed that three of the four Qf replicates were in the 3^{rd} quartile (Fig. 3.8, cluster 64), and three Qw samples were in the 1^{st} quartile (Fig. 3.8, cluster 36, 43).

This separation in the quartiles shows that the fat bodies of Qw and Qf share differentially expressed genes, but their expression levels deviate to different extents. It is in line with the fat body having the same role in both developmental states, with an enhanced function when the larvae are constantly feeding compared to when they are wandering. On the other hand, cluster 23 comprised tubules from wandering and feeding larvae (Lw and Lf), which fell into the same quartiles because they not only share the shape of deviation, but also their extent of deviation. Interestingly, the heterogeneous cluster 2 comprises four replicates of the adult carcass (C), consisting of everything that is left of the thorax and abdomen after the gut and sexual tracts have been removed, and one replicate of the adult trachea (X). These tissues are anatomically close and technically difficult to dissect, hence cross contamination is a likely problem. In line with this hypothesis, the other trachea replicates were in nearby groups 3 and 5. An outlier from the fat body (Q) was identified as cluster 10, while the 3 other replicates clustered together much further away in terms of amount of deviation in cluster 20. An identical situation was noted for the larval trachea (Xf) found in cluster 15 and 19 (Fig. 3.8). Thus, the two-part clustering of TTMap adds information and provides additional biological insights.

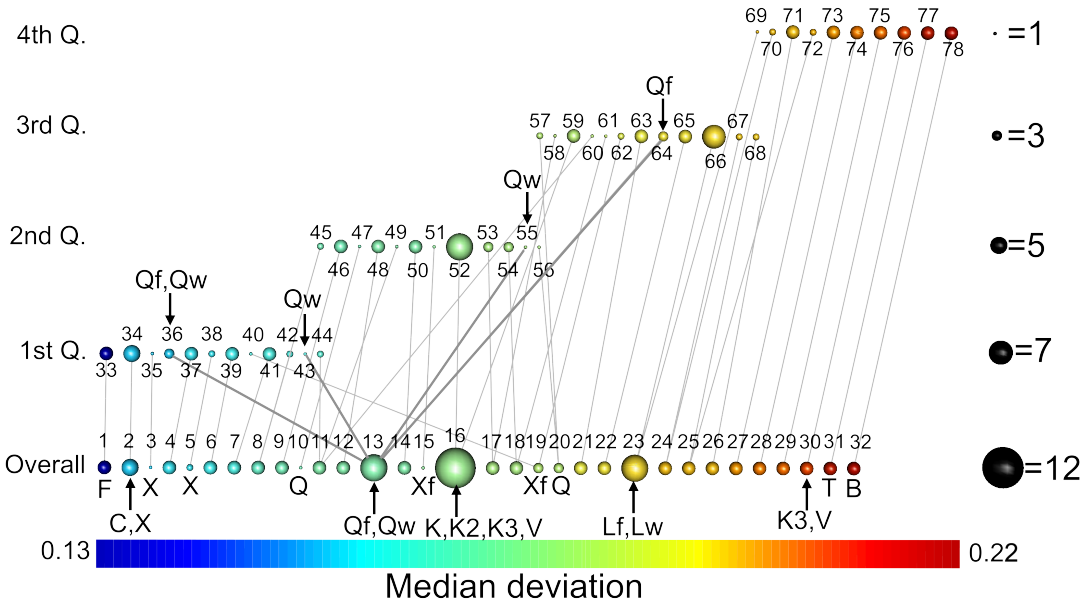


Figure 3.8: TTMap characterizes deviations of gene expression in different fly organs from whole fly tissues flyatlas. Output of TTMap showing the global clusters (Overall) and local clusters (1st, 2nd, 3rd, 4th Quartiles of the amount of deviation function) with its links to the global clusters. The size of the sphere corresponds to the number of samples in the cluster, the color the average amount of deviation. The number above the sphere identifies the clusters and the letters indicate organs inside a cluster (C: carcass, X: adult trachea, Xf: larvae trachea, Q: fat body, K: spermatacea virgin, K2: spermatacea mated and K3: spermatacea virgin (redone), V: adult toracic muscle, Qw: fat body of the wandering larvea, Qf: fat body of the feeding larvea, Lw: malpighian tubule of the wandering larvea, Lf: malpighian tubule of the feeding larvea, F: whole larvea, T: Testes, B: Brain). Outliers are the adult trachea (X) in clusters 3, 5, the larvae trachea (Xf) in cluster 15, as well as the fat body (Q) in cluster 10.

3.3.3 The impact of the choices of parameters

TTMap was tested here using Normal approximation and Poisson approximation for the ε parameter (section 2.3.4). In both cases, ε was estimated to be at the minimum of detection of the software **R** (and therefore set to 0). This low value is due to the control group having a small variation (90-th percentile of variance = 0.005, Fig. 3.9, red line). Hence, the estimation

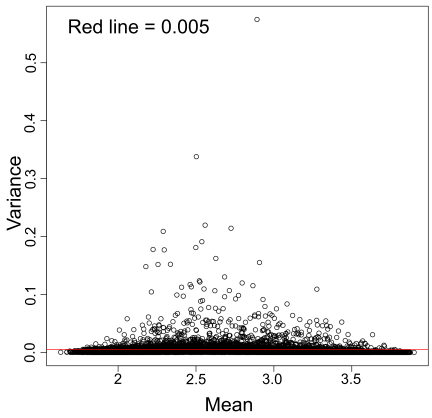


Figure 3.9: Mean against variance plot with red line representing the 90-th percentile of the variance. Each point represents a gene with its mean and variance across control samples.

of ε was independent of the choice of distribution (Poisson or Normal). Moreover, two choices are possible for the p-value in the probability estimation of ε, either 0.025 or 0.975 (see section 2.3.4). Again, no difference was observed choosing one or the other. The other parameter that could change the output is e, which when set to 1 no gene expression of the control group was considered to be an outlier. In case e is chosen by default, around 2000 genes per sample in the control group are changed even though only slightly. This difference and adjustment again, did not impact the output of TTMap, but was interesting to notice.

3.3.4 Direct comparison with DBSCAN

DBSCAN and TTMap are similiar in that they both function with a linkage algorithm (single-linkage for TTMap, section 1.2.7, a variation of it with a minimum amount of links to be connected for DBSCAN , section 1.5.2). It is therefore clear that the result on the data is quite similar in terms of uniquely clustering organs (Fig. 3.7a). The difference in uniquely clustering organs can therefore be listed : Brain and thoracoabdominal ganglion cluster together for DBSCAN and are separated for TTMap in uniquely clustering organs. TTMap in turn separates the samples from the trachea, which in DBSCAN are grouped but as outliers (counted here within the 20 uniquely clustering organs). As already mentioned, trachea is difficult to dissect and therefore is likely to have varying profiles from sample to sample. We did not find a justification for the grouping of Brain and thoracoabdominal ganglion found with DBSCAN.

3.3.5 Comparison with hierarchical clustering and PCA

We tested also other algorithms such as hierarchical clustering on raw data (Supplementary Fig. S12) or on the mismatch distance (Supplementary Fig. S13a, Supplementary tables) and PCA (Supplementary Fig. S13b) where we observed no clear separation in groups of 4. The best selection for the parameters of hierarchical clustering, the TC or the cutoff at a certain level, in terms of uniquely clustering organs, chosen by visual inspection of the plot, revealed 21 uniquely clustering organs on the mismatch distance. We tested the changes of performing PCA on the vectors after HDA (Supplementary Fig. S13c), and observed no visible improvement in the clusters, reflecting once more that both steps of TTMap are necessary in order to obtain increased information.

3.3.6 Literature search on genes found only by TTMap

TTMap revealed that larvae wandering fat body and larvae feeding fat body the same genes were differentially expressed overall but to different extents. Specific subgroups were obtained in each quartiles (Fig. 3.8, clusters 36, 43, 55, 64). We searched the literature for the genes specific to the subgroups of larvae wandering fat body and larvae feeding fat body in order to find which ones are related to the distinct stages of the larvae in the fat body.

We discovered that the gene "shd" (shade), only found with TTMap, is highly upregulated in larvae wandering fat body but not in larvae feeding fat body. In the fat body, this gene is a biological timer and regulates pupation timing in Drosophila melanogaster, i.e., this gene, expressed in the fat body, needs to be upregulated in order for the larvae wandering to turn into a pupa [211].

On the other side, highly upregulated genes in the larvae feeding fat body comprises specific genes such as "br" (broad) and "pgant8" that have been linked in several organs to the

development of organs in different stages of the larvae [212], however no literature exist on the fat body. Our data suggests that these genes might have the same role in the fat body than in these organs and could therefore be potential candidate markers of these different larvae stages. These results would need further validation.

3.3.7 Data availability

These drosophila Affymetrix array data files were downloaded from GEO accession no GSE7763.

3.4 Estrous-cycle-related gene expression changes in murine mammary glands

We challenged TTMap by asking it to identify subtle gene expression changes as they occur in a complex organ related to cyclic alterations in hormone levels. For this purpose, we studied RNA-seq data from intact mammary glands from C57BL/6 and BALB/c females, collected in different phases of the estrous cycle –proestrous (P), estrous (E), and diestrous (D)– based on the prevalence of different cell types in their vaginal smears (see section 1.6.2) ($n = 12$) [213].

Goal. *We want to assess TTMap's performance on RNA-seq data with batch effects corresponding to the strains of mice, which are C57BL/6 and BALB/c. TTMap will determine how far each sample of a certain phase of the estrous cycle is from another phase and if there are samples which overall have the same genes that change, but their extent is different.*

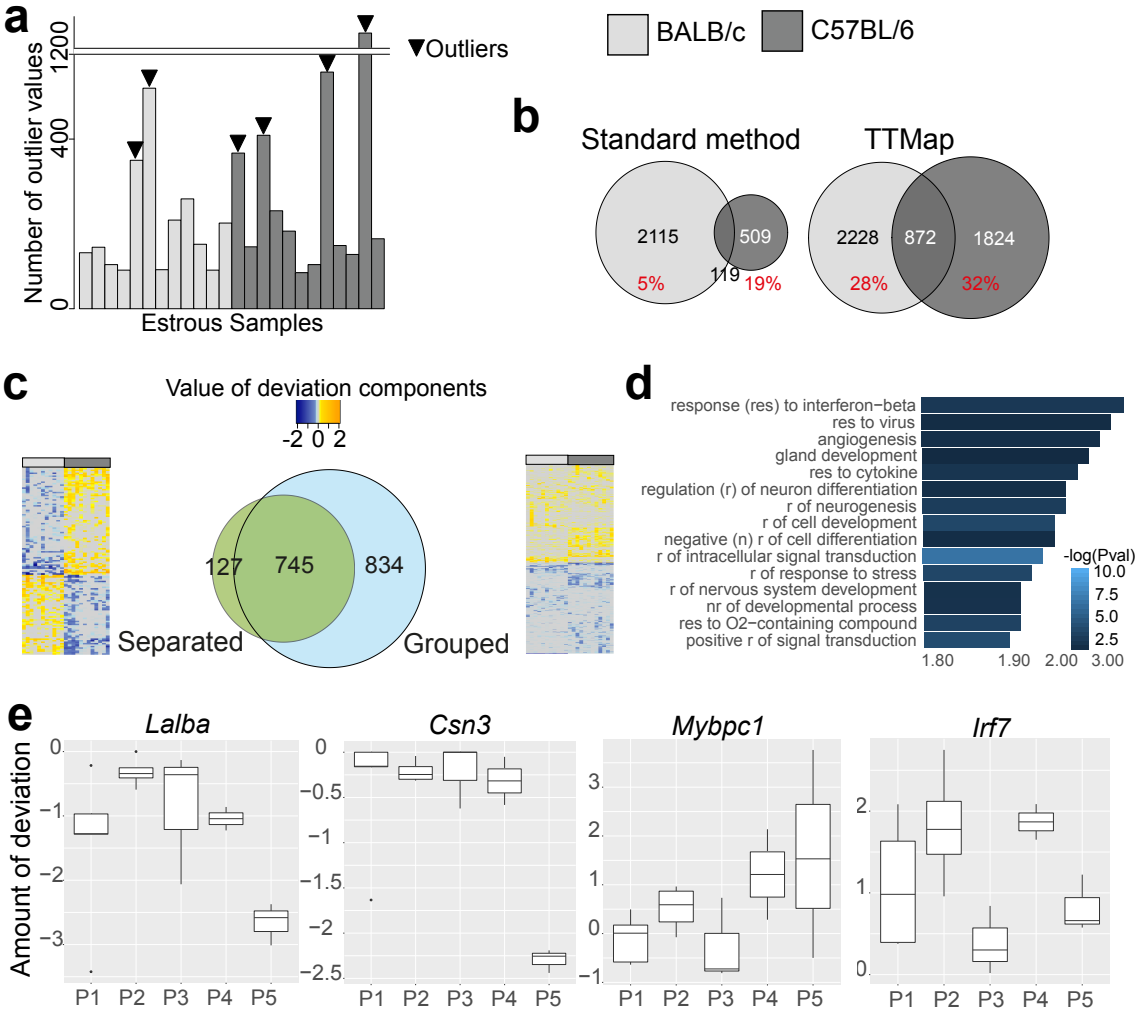


Figure 3.10: Estrous cycle related gene expression changes in the mammary glands of C57BL/6 and BALB/c mice; estrous vs proestrous phase. *(continued on the next page).*

Principal component analysis grouped samples according to strain (Supplementary Fig. S14a); and standard analysis was performed on each strain separately [213] leading to the identification

Figure 3.10: Estrous cycle related gene expression changes in the mammary glands of C57BL/6 and BALB/c mice; estrous vs proestrous phase. **(a)** Barplot representing the number of outlier values in each of the sample of the control group (estrous phase). Samples with high number of outlier values and remain isolated during clustering when E is the test group are identified as outliers (arrowheads). **(b)** Venn diagrams of the genes differentially expressed between E vs P using standard analysis tools and TTMap on BALB/c compared to C57BL/6 analyzed separately. In red, the fraction of common over total number of significant genes per strain. **(c)** Venn diagrams of the common differentially expressed genes when the analysis is done separately on the two mouse strains (Separated) or with the two mouse strains combined into one analysis (Grouped) using TTMap comparing E vs P. Adjacent heatmaps of the deviation components illustrate the reason why the genes were missed; on the separated analysis deviations are going into opposite direction, in the grouped analysis the genes deviate in the same direction, but to different extent. **(d)** Panther pathway analysis [84] of significant genes identified by TTMap in the comparison E vs P shown by Fold Change (FC) enrichment of the pathway with *-log(Pval)* as a color code. Fifteen most increased pathways are shown. **(e)** Boxplots representing the deviation component values in the identified subgroups of P (P1-P5) by TTMap ordered by amount of deviation compared to the estrous samples (controls) of the genes *Lalba*, *Csn3*, *Mybpc1* and *Irf7*.

of differentially expressed genes with a false discovery rate (FDR) < 0.05 and a low fold change (FC) of $|FC| \geq 1.2$ [213] (Fig. 3.10b, Supplementary Fig. S15b, S16b).

We considered each of the 3 cycle phases as the control group in TTMap. The number of outliers was 6/24 for estrous (E) (Fig. 3.10a, arrowheads), 4/23 for diestrous (D), and 4/23 for proestrous (P) (Supplementary Fig. S15a, S16a). TTMap increased the number of significant genes by a factor of 1.38 in the comparison E vs P in BALB/c and 4.29 in C57BL/6 (Fig. 3.10b). Moreover, a 1.08 and 5.29-fold increase in the number of significant genes in D vs P and E vs D, respectively, was observed in BALB/c, and a 2.2 and 2.83-fold increase in C57BL/6 in these two comparisons, respectively (Supplementary Fig. S15b, S16b). The overlap of significant genes between the two strains changed with TTMap compared to the standard analysis [213]. For E vs P, a consistent increase from 5 to 28 % in BALB/c and 19 to 32 % in C57BL/6 was observed (Fig. 3.10b). For D vs P, it increased from 18 to 36 % in BALB/c and decreased from 47 % to 45 % in C57BL/6 (Supplementary Fig. S15b). In E vs D, an increase from 0 % to 20% was found for both strains (Supplementary Fig. S16b).

Next, TTMap considered a strain as a batch (Fig. 3.10c). This grouped comparison increased the number of common genes 1.81-fold (Fig. 3.10c, Venn diagram) for E vs P and 1.72- and 2.23-fold for D vs P and E vs D, respectively (Supplementary Fig. S15c, S16c) over the common genes from the separate analyses with TTMap. The significant genes comprised $>$ 85% of the genes identified by separate analysis (Fig. 3.10c, Venn diagram). Heatmaps of the deviation components showed that the genes missed by the grouped analysis were differentially expressed in different phases of the cycle in BALB/c and in C57BL/6 mice but in opposite directions (Fig. 3.10c, Supplementary Fig. S15c, S16c, heatmaps on the left), whereas genes missed by separate analysis deviated in the same direction from the control in both strains, but did so to different extents and had therefore failed to reach significance in one of the strains (Fig. 3.10c, Supplementary Fig. S15c, S16c, heatmaps on the right).

Bioinformatic analysis of the genes revealed by the grouped analysis of E vs P using pathway analysis [84] revealed " angiogenesis" (FC $= 2.81$, $p < 2.23E-02$) and "gland development" (FC=2.44, $p < 4.73E-02$) as important terms (Fig. 3.10d) missed with standard tools, and "positive regulation of tumour necrosis factor (TNF) superfamily cytokine production" (FC=4.26, $p < 4.28E-03$) in D vs P (Supplementary Fig. S15d) when TNFα expression

was shown to vary through the human menstrual cycle [214]. Genes in E vs D were related to immune and inflammatory responses terms (Supplementary Fig. S16d). Using the filter function to determine the extent of deviation from the control group, TTMap orders subgroups within each phase. For P, P1 is closest and P5 furthest from the control (E) (Supplementary Fig. S14b). Among the significant genes in these subgroups are genes whose expression was previously shown to vary through the human menstrual cycle [142], [215] (Fig. 3.10e, Supplementary Fig. S15e, S16e), such as *Mybpc1*, a progesterone target gene [215], and the milk protein coding genes *Lalba*, *Csn3*, all missed with standard tools. These genes deviate significantly only in subgroups of P (Fig. 3.10e). In contrast, the normalized expression levels of *Irf7*, a gene detected by standard tools, were at least 1.2-fold higher in all 5 P subgroups, as reflected by the deviation components, compared to E (Fig. 3.10e). Biologically, estrous cycle phases are continuous rather than discrete subgroups (see section 1.6.2), TTMap maps samples in-between 2 phases by providing information about the overall closeness to control, as in the case of P1. These results were further validated by Kyoto Encyclopedia of Genes and Genomes (KEGG) and Panther pathway analysis of the genes that are differentially expressed between these phases; we discovered that P1, even though already having downregulated pathways that are common to the five proestrous subgroups, such as Fatty acid metabolism, p=0.0025, it had not yet upregulated major pathways like the oxytocin and calcium signaling pathway. Moreover, fluctuations in hormone signaling are reflected in P4 which revealed GO molecular pathways such as "response to hormone" (p=0.0144), "lactation" (p=0.0152), "response to steroid hormone" (p=0.0179), "cellular response to hormone stimulus" (p=0.0186) and "response to progesterone" (p=0.0186) (Supplementary Tables). Thus, TTMap by stratifying further on degree of deviation from normal characterizes the phases of the estrous cycle and reflects the underlying cyclic biology better than the standard tools and provides more information and additional insights into estrous-cycle-related gene expression changes.

3.4.1 Multiple comparison analysis

To observe the effect of HDA on the dataset, we performed PCA plots after HDA was performed in the comparisons E vs (P and D) (Supplementary Fig. S17a), P vs (E and D) (Supplementary Fig. S17b), and D vs (P and E) (Supplementary Fig. S17c). We notice that the samples cluster more by cycle stage and less by strains than in the starting PCA (Supplementary Supplementary Fig. S14a). Then, we assessed the effect of the filter function by looking in the multiple comparison a barplot of the values of the total absolute amount of deviation, the chosen filter function, in the different comparisons (Supplementary Fig. S18a, b and c), which revealed clear outliers and that E and D are both further away to P (Supplementary Fig. S18c, minimum deviation of 500) than to each other (Supplementary Fig. S18a, b, minimium deviation of 100 and 200). We performed hierarchical clustering on Z-score-normalised values per batch (per strain) of all the samples. This normalisation did not improve the grouping of samples which grouped samples according to their strain (Supplementary Fig. S19a, b and c). A Venn diagram compares the results of the analysis separately on E vs (D and P) on BALB/c only C57BL/6 only to the analysis with batches considered as strains. The same result as in the simple analysis was found, since few genes were not included when the analysis is done with all the samples (Supplementary Fig. S20).

3.4.2 Data availability

This mouse data was kindly provided by A. Snijders and colleagues [213].

3.5 Progesterone and R5020 action on human breast tissue

3.5.1 Understanding the role of progesterone compared to R5020

Progesterone receptor signaling (section 1.6) has been linked to proliferation [118], [117] and the potential tumorigenic role of progesterone has been described [106]. Understanding what are the molecular changes induced by progesterone inside the normal breast will shed light onto the mechanism by which progesterone is inducing proliferation. R5020 has been used for years in research as a synthetic progesterone [216]. This stable PR agonist, in a similar way than natural progesterone (P4), induces proliferation on cells treated with the compound [118], [117].

Goal. *The goal is to find genes that are controlled by progesterone in the human breast epithelium and whether or not these are induced as well by R5020. Then, we assess what are the differences in action of R5020 and progesterone on those cells.*

3.5.2 Experimental design

In order to evaluate the action of progesterone on gene expression of healthy human epithelial cells, the recently developed *ex vivo* model technique on mammoplasty cells was chosen as a model [118], obtained by a collaboration with the hospital in Lausanne the Centre Hospitalier Universitaire Vaudois (CHUV) , to ensure that hormonal pathway are still active. Tissues from the same patient were stimulated with P4 or with the synthetic progestin R5020 or with vehicle.

Patients were asked to provide information on their reproductive history (e.g. contraceptives, parity, last menses). At the moment of the mammoplasty surgery, blood was collected and later analyzed to determine the actual phase of the cycle (Table 3.2).

Sample number	Number of annotation	Age	P4 Level nmol/ L
Sample 220	15	35	<0.3
Sample 221	16	38	7.1
Sample 229	19	17	0.7
Sample 243	20	54	0.6

Table 3.2: Patient information

The tissue microstructures obtained following mechanical and enzymatic tissue dissociation of mammoplasty specimens [118] were exposed to vehicle or R5020 (20 nM) or P4 (20 nM) for 14 hours. Subsequently, tissue microstructures were dissociated and immunodepleted for immune cells, fibroblasts, and endothelial cells with a cocktail of anti-CD45, anti-FAP, and anti-CD31 antibodies. Cells were labelled with antibodies against Epithelial Cell Adhesion Molecule (EpCAM) (clone HEA-125) to enrich for the luminal cell population and Common Acute Lymphoblastic Leukemia Antigen (CD10/CALLA) for myoepithelial cells (Clone SS2/36) [217]. The EpCam+ cells were collected and processed for RNA analysis.

Illumina Eland_v2e protocol was used to generate RNA-seq data for each tissue in the condition control, progesterone treated, R5020 treated. A total of 4 different mammoplasties were used (Table 3.2).

3.5.3 Standard analysis

After using tophat (version-2.0.11, maximum 2 multimaps, not allowing for novel junctions or indels) and featureCounts (version-subread-1.6.0) for aligning and counting the reads respectively, we used the limma pipeline [79] for RNA-seq analysis with the Voom normalisation [76] after trimmed-mean of M values (TMM) normalisation [218], using the design with groups given by treatment (control, R5020 and progesterone) and batches given by the mammoplasty (15,16,19,20).

PCA plot did not group samples according to batches (mammoplasties) after TMM normalisation (Fig. 3.11a), but once the data was renormalised with Voom samples cluster according to mammoplasty (Fig. 3.11b). This is expected as human variability (from mammoplasty to mammoplasty) should be larger than any treatment inferred on the cells [219].

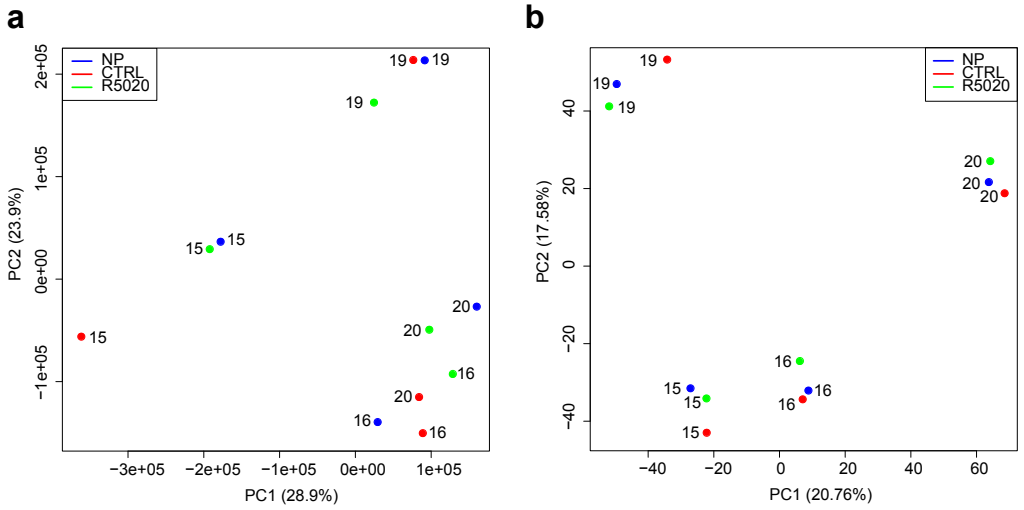


Figure 3.11: Standard clustering algorithms of gene expression profiles from four mammoplasty (15, 16, 19, 20) treated with vehicle R5020 or P4. **(a)** PCA of RNAseq samples after TMM normalisation colored by treatment and annotated by patient number **(b)** PCA of RNAseq samples after Voom normalisation performed on TMM normalised samples.

We performed differential expression analysis of R5020 samples vs control (with batch correction) and progesterone vs control (with batch correction) and R5020 vs progesterone (with batch correction) and corrected the p-values according to the Benjamini-Hochberg method [220] for adjustment of p-values. We found respectively 89, 65, and 0 significant genes by adjusted p-values (Fig. 3.12a). Noticing that many genes were common in the treatment with progesterone and the treatment with R5020, we overlapped the two list of significant genes and found that most genes are shared (Fig. 3.12a). This was still the case when looking at the significant genes only by p-value (Fig. 3.12b).

Among other known progesterone receptor target genes, *WNT4* and *RANKL* (p-value = 0.003, and 0.0008 respectively for progesterone and 0.001 and 0.009 respectively for R5020), but not adjusted p-value, we found *CXCL13* (adj.p-value= $0.0006, 6.08E-05$) , *MYBPC1* (adj.p-value= $0.0002, 6.08E-05$) [221], [142], which are positive controls and validate the induction of progesterone and R5020. As a negative control, we observed downregulation of *PGR* (p value= 0.003, 0.003) after treatment with both progesterone and R5020 [222]. *GREB1*, found to be significantly upregulated (adj.p-value= 0.049, 0.049), was recently shown

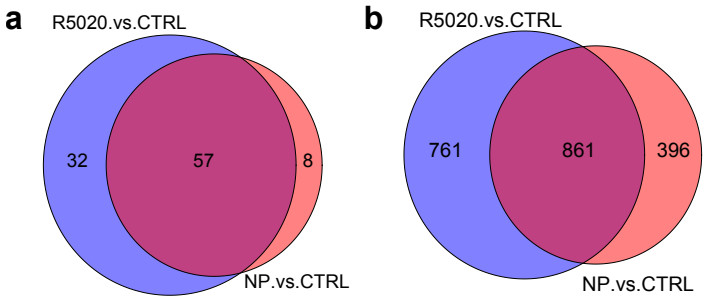


Figure 3.12: Venn diagram of differentially expressed genes in the comparison between R5020 and the control treated samples and progesterone and control treated samples **(a)** adjusted p-value selected genes **(b)** p-value selected genes.

to be a progesterone responsive gene [223] in healthy human endometrial stromal cell. This is the first evidence that this also applies to breast epithelial cells.
Standard analysis using PCA revealed that the samples cluster by mammoplasties, but since the effect of mammoplasty is larger than the treatment inferred to the cells, we do not know if this clustering reflects an individual way of reacting to treatments or if this suggests that each mammoplasty has different "baseline" levels of genes.
Then, standard differential expression analysis revealed that the effect of progesterone and R5020 is mostly similar on the cells.

3.5.4 The analysis using TTMap

We used TTMap with batches corresponding to the different mammoplasties. The controls were given by the vehicle treated samples. TTMap subgroups the treated samples with R5020 and progesterone according to their pattern of deviation compared to control and color code the obtained clusters (represented by spheres) by the amount of deviation compared to control (blue, low deviation and red, high deviation).
Four clusters formed upon TTMap analysis (Fig. 3.13) in the overall situation, each cluster corresponding to a mammoplasty sample, and comprising the treatment with R5020 and the treatment with progesterone in one cluster. Hence, TTMap's output suggest that R5020 and progesterone induce the same transcriptomic alterations on EpCam+ breast cells of mammoplasties but each mammoplasty has its unique changes induced by the two compound. Moreover, the different local tiers (the quartiles) of TTMap revealed that for each of the spheres except the most extreme one, the natural progesterone was always classified in a lower quartile than its R5020 counterpart, per mammoplasty. This suggests that R5020 and progesterone imply the same gene expression changes on the microstructures, but the extent of deviation of the gene expression changes is more enhanced with R5020, probably illustrating that R5020 is a more potent compound [118] (Fig. 3.13).
The closest sample to control is from the gene expression profile of mammoplasty number 16 (Fig. 3.13), which had the highest level of progesterone (Table 3.2). The furthest sample from control is tissue extracted from mammoplasty number 19 (Fig. 3.13), which was from the youngest patient (Table 3.2). This underlines once more the fact that each patient has a unique way of reacting to a treatment [224] and in particular patient history is crucial in order to predict the effect of progesterone.
An important improvement over the standard methods concerns the expression of *Cyclin-D1*

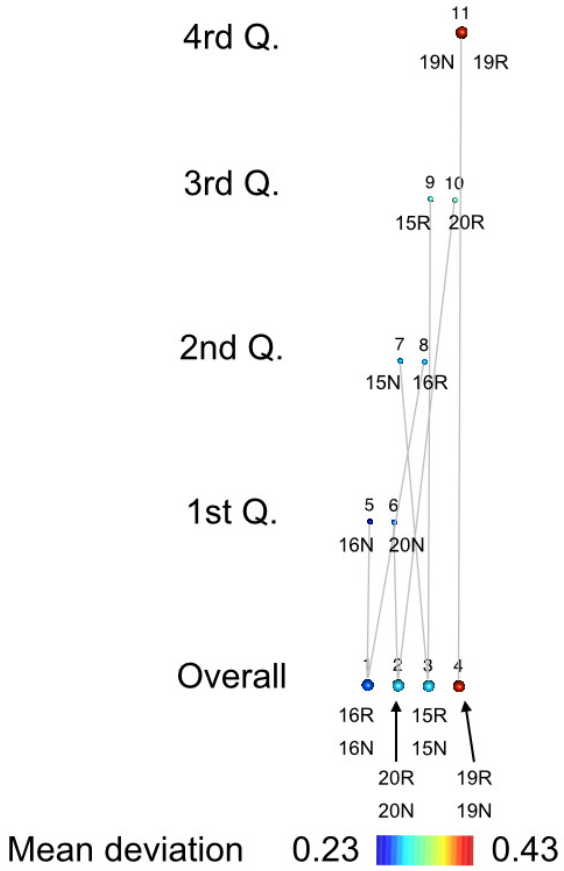


Figure 3.13: TTMap characterize the exogenous action of Progesterone (N) and R5020 (R) on microtissues obtained from mammoplasty numbered 15, 16, 19 and 20.

that was not significant with standard analysis, but was consistently downregulated in the cluster of the samples 19, and not in the three others. Similarly, *FOXA1* was at the border or about to of being significant ($p = 0.09$) but was enriched in cluster 19, and others not. A more precise decomposition of the changes occuring in each mammoplasty could only be appreciated using deviation components of TTMap.

We conclude that TTMap is producing an output with much richer structure than standard methods such as PCA, since we revealed the samples with smallest and highest reaction to treatment. Furthermore, we discovered that even though the transcriptomic changes induced by R5020 and progesterone are the same, R5020 induces the variations to a higher extent. This also reflects that a careful attention should be made when delineating analogies between R5020 and P4 as one is a more potent compound than the other one.

Based on our findings with TTMap, we hypothesised that high level of progesterone makes the cell less responsive to progesterone whereas tissue structures from young patients are more responsive to treatment. We also discovered significant genes known to be downstream of PR signaling, such as *Cyclin-D1*, which was shown to interact with PR in breast cancer cells [127]. However, this gene was only significantly changed in subsets of human breast samples, reflecting the heterogeneity of human expression profiles, only obtained through the individual profiles of deviations obtained by TTMap.

3.6 RANKL stimulation of human breast specimens

The receptor activator of nuclear factor κB ligand (RANKL), a downstream target gene of progesterone receptor signaling (section 1.6.9), plays an important role in mediating cell proliferation in the mouse mammary [191], [192] and human breast tissue [118]. The mechanism through which this cytokine induced proliferation is unknown.

Goal. *To uncover events downstream of RANKL signaling, we investigated the effect of exogenous RANKL on mammoplasty cells from healthy donors.*

3.6.1 Experimental design

To assess the effect of RANKL on gene expression in the normal human breast, we generated tissue microstructures ex-vivo [118] from reduction mammoplasty surgery patients obtained by the CHUV.

Patients were asked to provide information on their reproductive history (contraceptives, parity, last menses). At the moment of the surgery blood was also collected and later analyzed to determine the levels of progesterone (Table 3.3). Through the menstrual cycle, lower serum progesterone level are found in the follicular phase and determined as less than 4 nmol/l (see section 1.6.2) and higher serum progesterone level correspond to luteal phase and are more than 4 nmol/l [225], [226].

Name	Age	Oral Contraceptives	P4 Level nmol/ L
Sample 210	22	No	1
Sample 211	21	No	20
Sample 212	42	Unknown Pill	18
Sample 215	38	No	2
Sample 217	37	NA	1
Sample 230	17	NA	0.7

Table 3.3: Patient information

The tissue microstructures obtained following mechanical and enzymatic tissue dissociation of mammoplasty specimens [118] from six patients between 17 and 42-year-old who underwent mammoplasty surgery were exposed to vehicle or rRANKL (1μ g/ml) for 14 hours. Subsequently, tissue microstructures were dissociated and immunodepleted for immune cells, fibroblasts, and endothelial cells with a cocktail of anti-CD45, anti-FAP, and anti-CD31 antibodies. Cells were labelled with antibodies against Epithelial Cell Adhesion Molecule (EpCAM) (clone HEA-125) to enrich for the luminal cell population and Common Acute Lymphoblastic Leukemia Antigen (CD10/CALLA) for myoepithelial cells (Clone SS2/36) [217]. The EpCam+ cells were collected and processed for RNA analysis.

Affymetrix Human Gene 10 st v1 probeset protocol (microarray protocol, see section 1.4) was performed on EpCAM+ cells treated or not with RANKL.

3.6.2 Standard analysis

Standard moderated t-test paired for tissue mammoplasty and corrected using Bonferroni correction method revealed only 11 significantly differentially expressed genes (by p-value and none by adjusted p-value) between treated samples with RANKL and tissue treated with vehicle. Hierarchical clustering did not match all the pairs (Fig. 3.14a) and PCA plot did not

group samples according to mammoplasties (Fig. 3.14b). Even after taking only the 100 most variable genes (by leading fold change) samples did not group according to mammoplasty (Supplementary Fig. S21a and b). Thus, standard clustering algorithms and analysis tool could not reveal any relevant information on the action of RANKL on human microstructures.

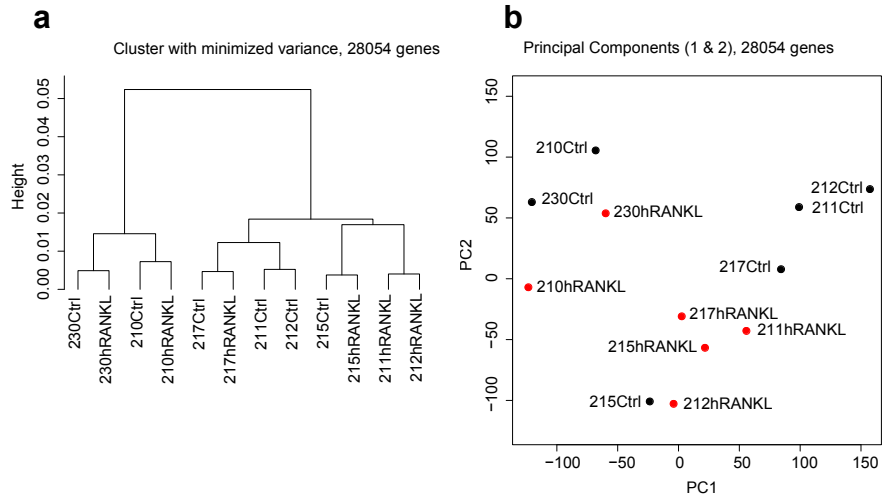


Figure 3.14: Standard clustering tools on gene expression profile of six samples (210, 211, 212, 215, 217, 230) treated with human RANKL (hRANKL, red) or with vehicle (Ctrl, black). **(a)** Hierarchical clustering on the full dataset **(b)** PCA plot on the full dataset.

3.6.3 The analysis using TTMap

We used TTMap taking the untreated samples as the control group and correcting for the batches that are the different mammoplasties. Two clusters formed upon TTMap analysis (Fig. 3.15): a cluster of four samples with low average total amount of deviation (filter function, blue color code on Fig. 3.15), and a cluster of 2 samples with high deviation showing clear changes occurring upon treatment (orange color code). Standard moderated t-test was then applied on the two subgroups discovered by TTMap. The analysis of the first subgroup confirmed the observation on the TTMap graph that there are no significant gene expression changes in that subgroup, whereas in the second subgroup 280 significant genes could be found, among which several genes linked to lactation such as *LALBA*, *CSN1S1*, *CSN2* and *CSN3* (Fig. 3.16a), in line with known function of RANKL in lactation [189]. Further analysis of the subgroups revealed a clear correlation of progesterone levels in the blood of the patient at the time of surgery and the response to treatment with RANKL (Table 3.3). The cluster showing low deviation reflected by the color code (first cluster, blue, Fig. 3.15) is composed of patients with low serum progesterone, while the two samples displaying high deviation (second cluster, orange, Fig. 3.15) were from patients with high serum progesterone levels at the time of surgery, as assessed by mass spectrometry.

As luteal phase (see section 1.6.2) corresponds with high levels of progesterone, progesterone alone is not the only factor correlating with the composition of the clusters (other correlation are for instance luteal phase, and, even though not assessed, also prolactin levels are higher during the luteal phase [227] than the follicular phase).

When RANKL was added to cells derived from women with high progesterone, we observed

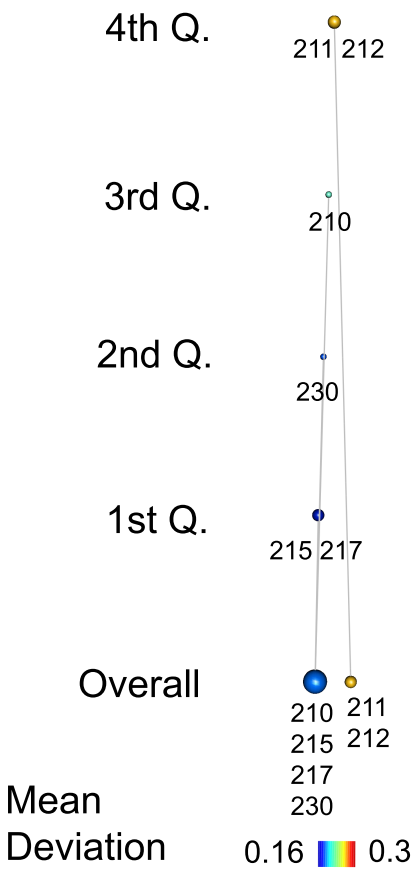


Figure 3.15: TTMap characterize the exogenous action of RANKL in mammoplasty tissue. Two major clusters are apparent in the overall group, a blue color-coded cluster made out of samples 210, 215, 217 and 230 and an orange color-coded group with two samples 211 and 212.

the downregulation of several genes linked to progesterone receptor signaling, among which *DIO2*, *CXCL13*, *MYBPC1*, *EREG*(Fig. 3.16b).

To give more strength to the observation that progesterone receptor signaling genes are downregulated, we overlapped the list of downregulated genes by RANKL to the significant genes that are upregulated in gene expression profile of luteal phase premenopausal epithelial enriched cells of women compared to gene expression profile of women in the follicular phase in I. Pardo's dataset *et al.* [142] where progesterone receptor signaling is more active and found a significant enrichment ($p = 0.0035$, hypergeometrical test). More strickingly, genes that were upregulated by RANKL in the presence of progesterone were significantly enriched in the genes downregulated during the luteal phase in I. Pardo's dataset *et al.* [142] ($p = 2.5 \cdot 10^{-8}$, hypergeometrical test). Only two genes were found to be regulated in the same direction when overlapping these gene lists and hence there was no significant enrichment ($p = 0.110$).

By pathway analysis [84] of the genes modulated by RANKL, we found significant enrichment mostly in immune related terms (e.g. negative regulation of innate immune response ($p = 2.16E-04$), or interferon-gamma related terms (e.g. cellular response to interferon-gamma, $p = 2.38E-09$, but found as well proliferation terms (e.g. regulation of epithelial cell proliferation, $p = 8.79E-04$), with genes such as *CTSL2*, *FGFBP1*, *PTN*, *CAV1*, *MMP12*,

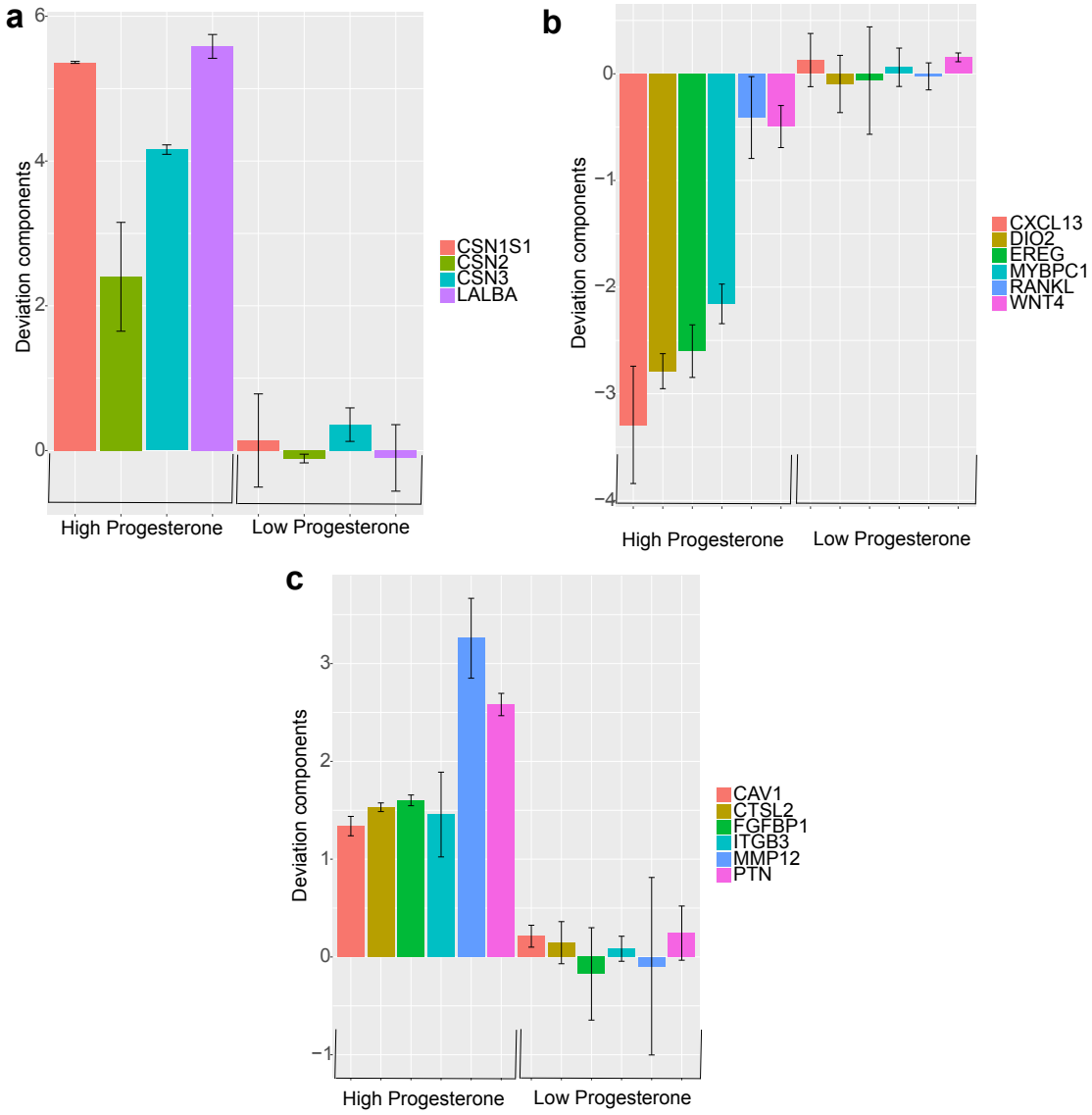


Figure 3.16: Deviation components of genes, shown as average with standard deviation in the samples with high level of progesterone and in the samples with low level of progesterone of **(a)** lactation associated genes *LALBA*, *CSN1S1*, *CSN2* and *CSN3*, **(b)** progesterone receptor target genes *DIO2*, *CXCL13*, *MYBPC1*, *EREG*, *RANKL*, *WNT4*, and **(c)** proliferation associated genes *CTSL2*, *FGFBP1*, *PTN*, *CAV1*, *MMP12*, *ITGB3*.

ITGB3 (Fig. 3.16c), in-line with RANKL increasing proliferation on the cells [118] and hints towards molecular mechanism through which this process is happening (Supplementary Fig. S22).

We conclude that TTMap unravelled a novel action of RANKL, missed by standard tools, which seems to be dependent on the presence of progesterone in the breast or the breast to have been in luteal phase. In that case, a pregnancy-like program is induced with genes such as *LALBA*, *CSN1S1*, *CSN2* and *CSN3*, confirming the implication of RANKL with lactation [189], as well as a proliferation program (*CTSL2*, *FGFBP1*, *PTN*, *CAV1*, *MMP12*, *ITGB3*), also in-line with the literature [118]. Moreover, RANKL counteracts progesterone/luteal

induced signatures, which is a novel role of RANKL that needs further validation.

3.6.4 Comparing RANKL's to progesterone's action on breast epithelium

In the two last sections, we analysed the action of RANKL on breast epithelial cells (see 3.6) with TTMap which is dependent on the presence of progesterone, and the action of progesterone (as well as R5020) on breast epithelial cells as well (see 3.5). In this section, we want to compare the two datasets and observe the common or diverging genes and pathways. As genes were only significant in the context of high progesterone with the treatment of RANKL, its action will refer to the effect of RANKL in this context.

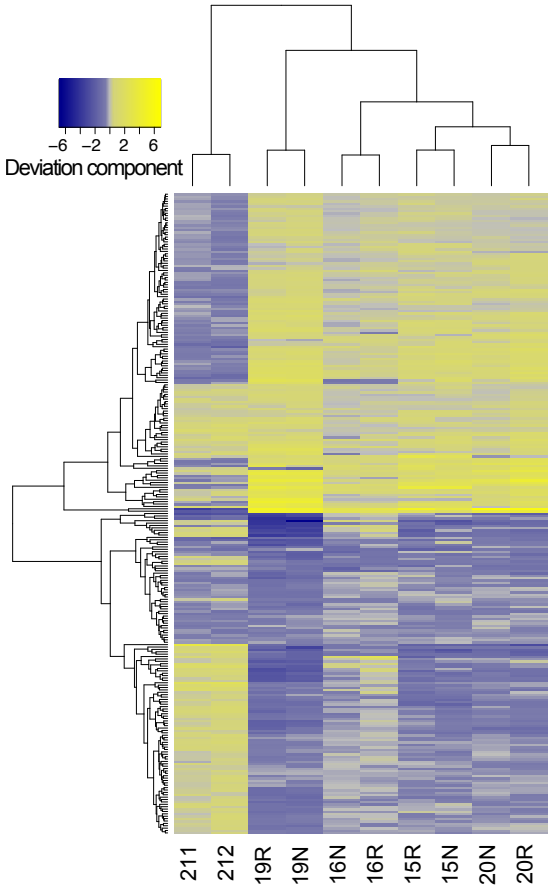


Figure 3.17: Heatmap of the deviation component of the common significant features between the RANKL experiment looking at its action in the two samples (211,212) with high progesterone and the experiment searching for progesterone action (N) and R5020 action (R) in 4 samples (15, 16, 19, 20).

We observe that the deviation component of the common significant genes in the experiments studying the effect of RANKL and progesterone are mostly going into opposite direction, as reflected by the color-code in the heatmap of the deviation components (Fig. 3.17).
We specifically analysed the pathways of the commonly changed genes and found that pathways are linked to apoptosis (adjusted p-value = 0.005) and cell death (adjusted p-value = 0.005) (Supplementary Fig. S23) , with genes such as *BMF*, *DUSP6* and *TP53INP1* which are strongly downregulated upon RANKL and upon progesterone (Fig. 3.18a). There is therefore a possibility that RANKL and progesterone communicate to block important apoptotic

processes. Additionally, the pathway of gland morphogenesis was significant with genes that are commonly downregulated such as *PGR*, *PTHLH* and *TBX3* and genes that are commonly upregulated such as *TGFA*, *CAV1* and *TGM2* (Fig. 3.18b).

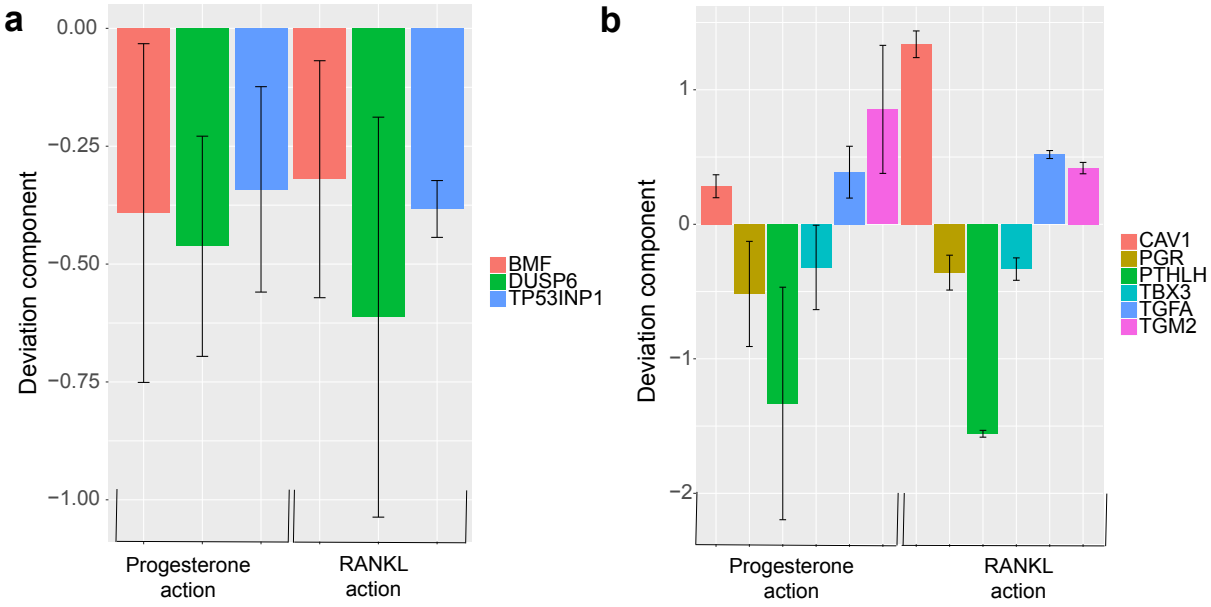


Figure 3.18: Deviation components of genes, deviating in the same direction upon RANKL stimulation or progesterone stimulation, shown as average with standard deviation linked to **(a)** apoptosis and cell death, with genes such as *BMF*, *DUSP6* and *TP53INP1* and **(b)** gland morphogenesis, with genes such as *PGR*, *PTHLH*, *TBX3*, *TGFA*, *CAV1* and *TGM2*.

We also analysed the differentially expressed pathways, which are increased upon RANKL and decreased upon progesterone. One of the top hits is "tissue development" (adjusted p-value = 0.00165) with genes linked to Transforming Growth Factor beta signaling (*TGFB2*, *TGFBI*) (Fig. 3.19a). Then, we found terms linked to signal transduction and cell communication reflecting the fact that RANKL is acting in a paracrine manner, and might hint towards possible co-players that can be validated (Supplementary Fig. S24).

Genes that are significantly decreased when RANKL is added and increased when progesterone is added, were linked to progesterone receptor signaling with genes such as *RANKL*, *WNT4*, *CXCL13* and *MYBPC1* (Fig. 3.19b). Moreover, genes linked to estrogen receptor signaling and prolactin receptor signaling are also significantly changing into opposite directions (Fig. 3.19c, d).

Pathways linked to all the significant genes varying in opposite direction upon RANKL and progesterone is revealing "epithelium development" (adjusted p-value = 0.02), "branching morphogenesis of an epithelial tube" (adjusted p-value = 0.03) and "epithelial cell differentiation" (adjusted p-value = 0.03) (Supplementary Fig. S25). These findings suggest that RANKL is counteracting progesterone receptor signaling on the cells and this process, probably, happens by regulating epithelium development. Further validation of these results will shed light onto the question and a first attempt is made in section 3.6.5.

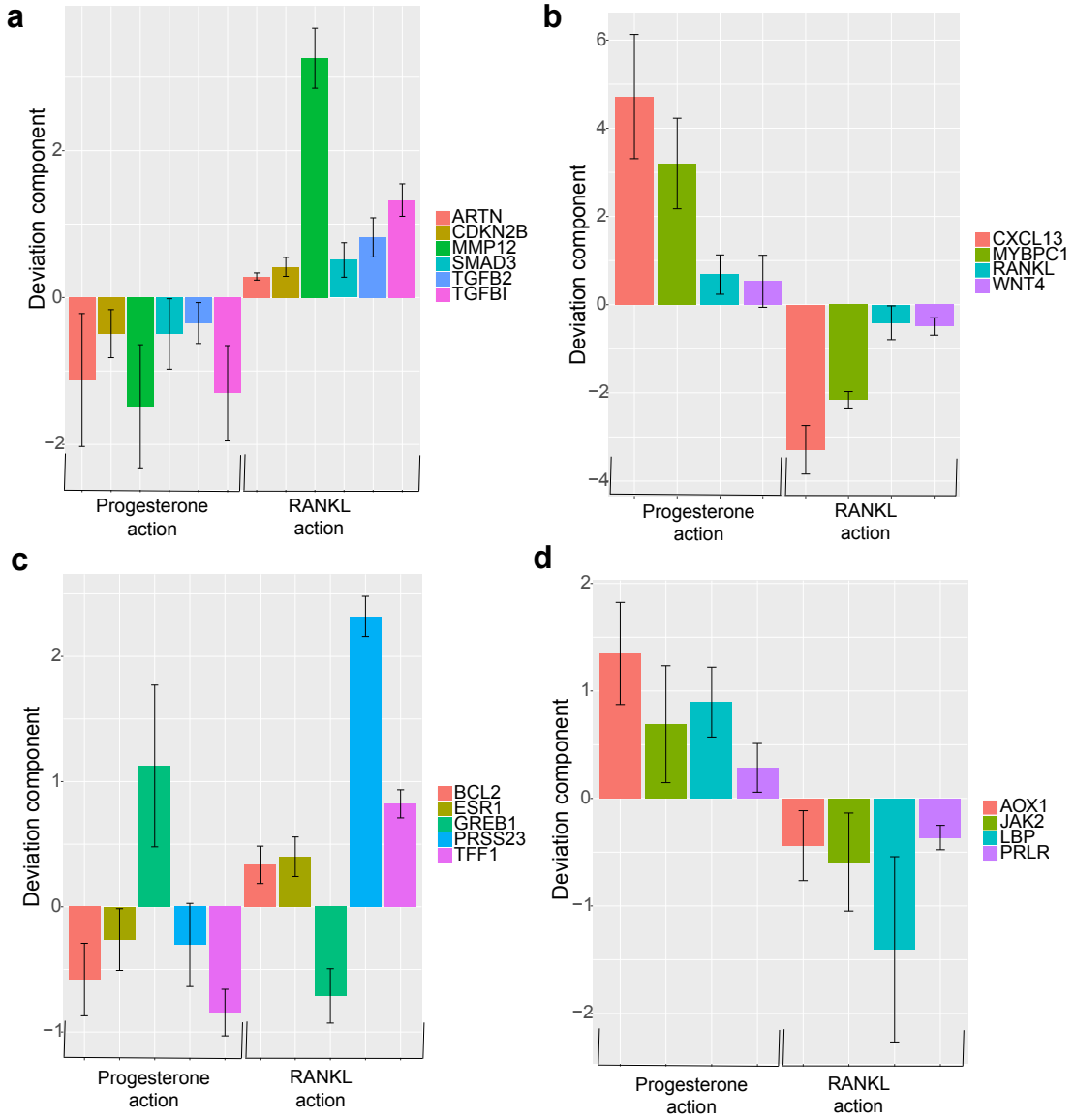


Figure 3.19: Deviation components of genes, deviating in the same direction upon RANKL stimulation or progesterone stimulation, shown as average with standard deviation linked to **(a)** TGF-β-signaling with *ARTN*, *CDKN2B*, *MMP12*, *SMAD3*, *TGFB2* and *TGFBI*, **(b)** progesterone receptor signaling with *RANKL*, *WNT4*, *CXCL13* and *MYBPC1*, **(c)** estrogen receptor signaling with *BCL2*, *GREB1*, *ESR1*, *TFF1* and *PRSS23* and **(d)** prolactin receptor signaling with *AOX1*, *JAK2*, *LBP* and *PRLR*.

3.6.5 Preliminary attempt to validate results obtained by TTMap

In order to test the hypothesis that progesterone needs to be present in the cells for RANKL to exert molecular changes on the mRNA of the cells, we injected human breast cells extracted from mammoplasty surgeries into the milk duct of recipient mice following the protocol described in [228], [217]. Half of the mice received a pellet of progesterone with a dosage of 20 nM for 14 days, mimicking luteal phase like exposure, while the other half was left untreated. The mammary glands of those animals were extracted, the cells were mouse depleted (to enrich for human cells) and stimulated overnight with RANKL or with vehicle. Hence, the four

conditions per mammoplasty injected into mice were control, control + RANKL, progesterone treated (P4), P4 + RANKL.
These four conditions should mimic the previous dataset in that we have the "follicular phase"-like condition with the control mice either treated or untreated with RANKL, then the "luteal phase"-like condition with the mice treated for 14 days with progesterone. We used two mammoplasties (number 319, 324) and several mice to reduce the variability from mice to mice. and previously validated human specific primers were used for qPCR. The patient information of these mammoplasties are missing and therefore crucial information to help us conclude on the results is unavailable. Steroid measurements were performed and the results are summarised in Table 3.4.

Mammoplasty number	Estrogen levels (pg/ml)	Progesterone levels (ng/ml)
319	<LOD	1.02
324	37.38	0.08

Table 3.4: Steroid levels in blood retreived from mammoplasties samples 319 and 324, LOD = Limit of detection.

We will hence be able to validate also some genes obtained in section 3.5 as we have the condition control and progesterone.

Prolactin pathway
Key genes of the downstream pathway of prolactin receptor signaling such as *PRLR* and *JAK2*, among others, were shown by TTMap to be downregulated upon RANKL in the background of high progesterone in the microarray data (see section 3.6) , whereas these were shown to be upregulated when cells were treated with progesterone or R5020 (Table 3.5). When progesterone was low in the patient of the mammoplasty at the time of surgery, the levels of *PRLR* and *JAK2* inconsistently vary and often times the variation is low (Table 3.6).

Gene Symbol	Description	Deviation component average upon RANKL in progesterone high samples	Deviation component average upon R5020 or progesterone
PRLR	Prolactin receptor	-0.36	0.28
JAK2	Janus kinase 2	-0.59	0.72

Table 3.5: Prolactin signaling in the experiments on the effect of RANKL (see section 3.6) on human cells and the effect of progesterone or R5020 (see section 3.5), average value for deviation components obtained by TTMap are shown for *PRLR* and *JAK2*.

Gene Symbol			Deviation component			
	211 (High P4)	212 (High P4)	210	215	217	230
PRLR	-0.28	-0.44	0.06	0.02	0.14	-0.26
JAK2	-0.27	-0.91	0.53	-0.38	-0.15	-0.42

Table 3.6: Prolactin signaling in the experiment of the effect of RANKL on 6 treated mammoplasties; 2 with high level of progesterone (P4) (211, 212) and 4 with low level (210, 215, 217, 230). Individual deviation components values are displayed for *PRLR* and *JAK2*.

We then proceed with the validation of *PRLR* by qPCR, and found in each tested mammoplasty

a slightly different result; while in mammoplasty 319 RANKL is downregulating *PRLR* in control cells and in treated cells with progesterone (Fig. 3.20a), in mammoplasty 324 RANKL increases slightly the levels of *PRLR* in control cells and downregulates them in the treated case.(Fig. 3.20b) Therefore, the effect of RANKL on treated cells with progesterone revealed a downregulation of *PRLR* in both cases, confirming the findings of the microarray. Overall, however the trend seems to go along with the findings of the microarray, where RANKL has no significant effect when progesterone is not present whereas when progesterone is present the levels of *PRLR* are downregulated (Fig. 3.20c, d). A third sample with available patient information would shed light into this validation.

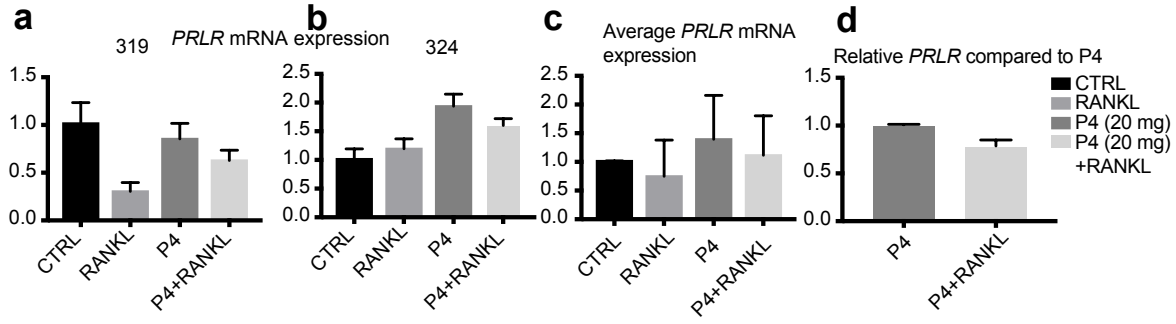


Figure 3.20: Validation of Prolactin signaling genes: *PRLR*. **(a)**, **(b)** *PRLR* on individual mammoplasties injected intraductally into recipient mice and kept for 14 days after they established and treated with control or progesterone, cells were taken out and treated with RANKL or control. Therefore, 4 conditions (control, RANKL, P4 and P4+RANKL) are found on mammoplasties **(a)** number 319 **(b)** number 324. **(c)** Average of two independent experiments of the 4 conditions and **(d)** *PRLR* levels of P4 + RANKL cells relative and compared to P4 only levels.

We then proceed with the validation of *JAK2* by qPCR, and found consistent results in both mammoplasties: RANKL is downregulating *JAK2* in treated cells with progesterone (Fig. 3.21a, b), and RANKL has no effect on the levels of *JAK2* in control cells. Therefore, the effect of RANKL on treated cells with progesterone revealed a downregulation of *JAK2* in both cases, confirming the findings of the microarray. Treatment with progesterone in both cases increased the level of *JAK2* (Fig. 3.21a, b) Overall, the trend seems to go along with the findings of the microarray, where RANKL has no significant effect on *JAK2* mRNA when progesterone is not present whereas when progesterone is present the levels of *JAK2* are downregulated (Fig. 3.21c, d). Of note, the increase in *JAK2* upon progesterone treatment was significant. A third sample with available patient information would shed light into this validation.

Estrogen pathway

Key genes downstream of estrogen receptor signaling pathway such as *ESR1*, *TFF1* [229] and *PRSS23* [230] were shown by TTMap to be upregulated upon RANKL in the background of high progesterone in the RNAseq data, whereas these were shown to be downregulated when cells were treated with progesterone or R5020 (Table 3.7). *GREB1*, an estrogen receptor target gene [229] in the breast and recently shown progesterone receptor target gene in the endometrium [231], showed an opposite pattern, it displayed an overexpression when treated with progesterone which was countered in the case of RANKL stimulation in the presence of

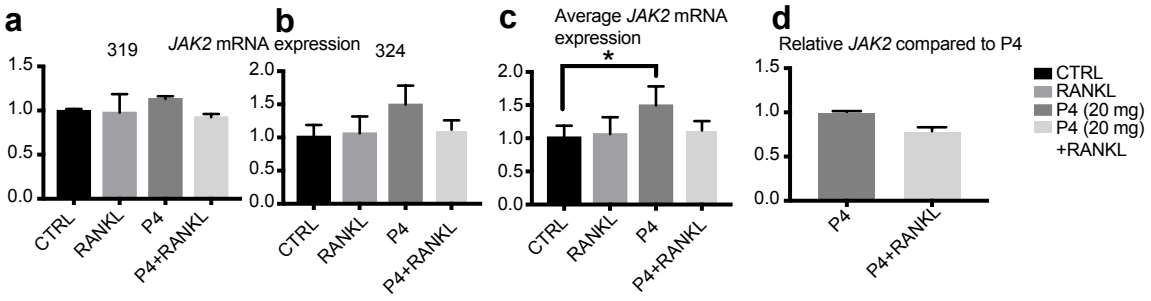


Figure 3.21: Validation of Prolactin signaling genes: *JAK2*. **(a)**, **(b)** *JAK2* mRNA measured by qPCR on individual mammoplasties injected intraductally into recipient mice and kept for 14 days after they established and treated with control or progesterone, cells were taken out and treated with RANKL or control. Therefore, 4 conditions (control, RANKL, P4 and P4+RANKL) are found on mammoplasties **(a)** number 319 **(b)** number 324. **(c)** Average of two independent experiments of the 4 conditions and **(d)** *JAK2* mRNA levels of P4 + RANKL cells relative and compared to P4 only levels.

high progesterone (Table 3.8).
When progesterone was low in the patient of the mammoplasty at the time of surgery, the levels of estrogen receptor target gene varied inconsistently and often times the variation is low (Table 3.8).

Gene Symbol	Description	Deviation component average upon RANKL in progesterone high samples	Deviation component average upon R5020 or progesterone
ESR1	Estrogen receptor (ER) 1	0.40	-0.28
TFF1	Trefoil factor 1	0.82	-0.81
PRSS23	Serine protease 23	2.32	-0.31
GREB1	Growth regulating ER binding 1	-0.71	1.34

Table 3.7: Estrogen signaling in the experiments on the effect of RANKL (see section 3.6) on human cells and the effect of progesterone or R5020 (see section 3.5), average value for deviation components obtained by TTMap are shown for *ESR1*, *TFF1*, *PRSS23* and *GREB1*.

Gene Symbol			Deviation component			
	211 (High P4)	212 (High P4)	210	215	217	230
ESR1	0.29	0.51	-0.11	-0.06	0.14	-0.22
TFF1	0.90	0.74	-0.24	0.01	0.23	0.04
PRSS23	2.21	2.43	-0.24	0.04	0.11	0.30
GREB1	-0.56	-0.87	-0.19	-0.11	0.24	-0.22

Table 3.8: Estrogen signaling in the experiment of RANKL action on 6 treated mammoplasties cells; 2 with high level of progesterone (P4) (211, 212) and 4 with low level (210, 215, 217, 230). Individual deviation components values are displayed for *ESR1*, *TFF1*, *PRSS23* and *GREB1*.

The first and most stricking difference was noted in the *ESR1* validation by qPCR as it depends strongly on the background of the mammaplosty treated. While the mammoplasty 319 with low level of estrogen strongly downregulated *ESR1* in all three cases (treated with

RANKL, with RANKL and P4, and with P4) (Fig. 3.22a), in the mammopasty 324, *ESR1* was increased upon progesterone treatment and further increased upon addition of RANKL, whereas no changes were observed upon RANKL alone (Fig. 3.22b).

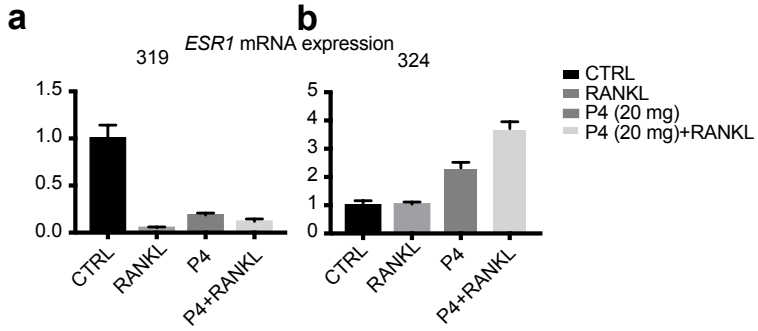


Figure 3.22: Validation of Estrogen signaling gene *ESR1*. **(a)**, **(b)** on individual mammoplasties injected intraductally into recipient mice and kept for 14 days after they established and treated with control or progesterone, cells were taken out and treated with RANKL or control. Therefore, 4 conditions (control, RANKL, P4 and P4+RANKL) are found on mammoplasties **(a)** number 319 **(b)** number 324.

Both mammoplasty cells showed a significant increase upon progesterone treatment of *TFF1* mRNA by qPCR (Fig. 3.23a, b). This is opposite to the results found in the RNA-seq. The further treatment of RANKL induced the desired changes in mammoplasty 319 but not in mammoplasty 324 (Fig. 3.23a, b). As in the microarray study, no changes were induced upon RANKL treatment alone (Fig. 3.23a, b).

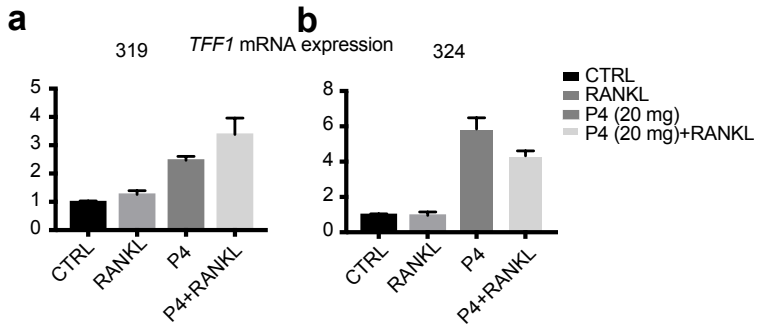


Figure 3.23: Validation of Estrogen signaling gene *TFF1*. **(a)**, **(b)** on individual mammoplasties injected intraductally into recipient mice and kept for 14 days after they established and treated with control or progesterone, cells were taken out and treated with RANKL or control. Therefore, 4 conditions (control, RANKL, P4 and P4+RANKL) are found on mammoplasties **(a)** number 319 **(b)** number 324.

Also *PRSS23* showed opposite results to the RNA-seq and the microarray (Fig. 3.24a, b). However, RANKL is counteracting and even abrogating (mammoplasty 319) the induced expression of *PRSS23* with progesterone (Fig. 3.24a, b).
Both mammoplasty cells showed a significant increase upon progesterone treatment of *GREB1* mRNA by qPCR confirming the results of the RNA-seq (Fig. 3.25a, b). Only mammoplasty 324 showed the desired changes upon further treatment with RANKL (after progesterone exposure) (Fig. 3.25b). Moreover, unlike the results of the microarray study, increased levels

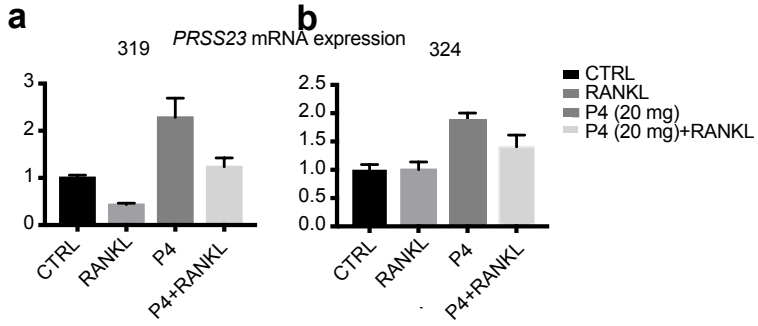


Figure 3.24: Validation of Estrogen signaling gene *PRSS23*. **(a)**, **(b)** on individual mammoplasties injected intraductally into recipient mice and kept for 14 days after they established and treated with control or progesterone, cells were taken out and treated with RANKL or control. Therefore, 4 conditions (control, RANKL, P4 and P4+RANKL) are found on mammoplasties **(a)** number 319 **(b)** number 324.

of mRNA are observed upon RANKL treatment alone (Fig. 3.25a, b). Averages confirmed those results (Fig. 3.25c, d).

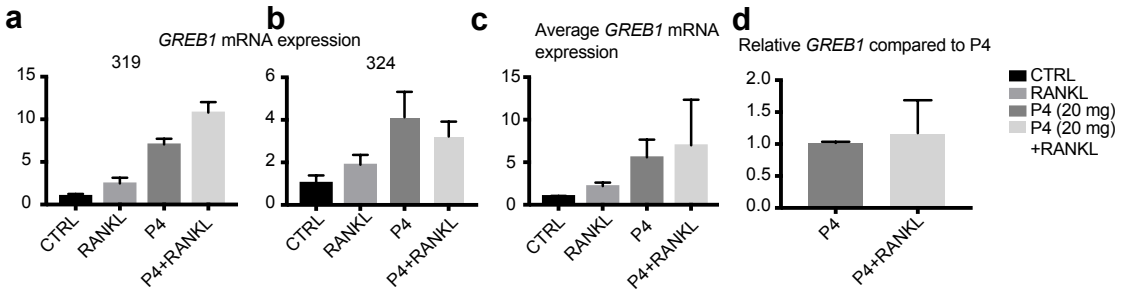


Figure 3.25: Validation of Estrogen signaling gene *GREB1*. **(a)**, **(b)** on individual mammoplasties injected intraductally into recipient mice and kept for 14 days after they established and treated with control or progesterone, cells were taken out and treated with RANKL or control. Therefore, 4 conditions (control, RANKL, P4 and P4+RANKL) are found on mammoplasties (a) number 319 **(b)** number 324.**(c)** Average of two independent experiments of the 4 conditions and **(d)** *GREB1* mRNA levels of P4 + RANKL cells relative and compared to P4 only levels.

Progesterone pathway

Key downstream target genes of progesterone receptor pathway such as *RANKL* [142], [106] *WNT4* [142], [106], *MYBPC1* [215], *CXCL13* [142] were shown by TTMap to be downregulated upon RANKL in the background of high progesterone in the RNAseq data, whereas these were shown to be upregulated when cells were treated with progesterone or R5020, in line with them being progesterone receptor target genes (Table 3.9). *PGR* showed a downregulation in both cases (Table 3.10), which follows the expected tendency [232].
When progesterone was low in the patient of the mammoplasty at the time of surgery, the levels of progesterone receptor target gene varied inconsistently and often times the variation is low (Table 3.10).

Gene Symbol	Description	Deviation component average upon RANKL in progesterone high samples	Deviation component average upon R5020 or progesterone
PGR	Progesterone receptor	-0.36	-0.65
RANKL	Receptor activator of nuclear factor kappa B ligand	-0.41	1.01
WNT4	Wingless-type family member 4	-0.49	0.72
MYBPC1	Myosin binding protein C slow type 1	-2.16	3.00
CXCL13	C-X-C motif chemokine ligand 13	-3.29	4.43

Table 3.9: Progesterone signaling in the experiments on the effect of RANKL (see section 3.6) on human cells and the effect of progesterone or R5020 (see section 3.5), average value for deviation components obtained by TTMap are shown for *PGR*, *RANKL*, *WNT4*, *MYBPC1* and *CXCL13*.

Gene Symbol			Deviation component			
	211 (High P)	212 (High P)	210	215	217	230
PGR	-0.45	-0.27	-0.34	-0.13	0.20	0.20
RANKL	-0.68	-0.14	-0.07	-0.18	0.09	0.06
WNT4	-0.63	-0.36	0.13	0.12	0.20	0.17
MYBPC1	-2.29	-2.03	0.03	-0.11	0.01	0.31
CXCL13	-3.68	-2.90	0.35	0.21	0.18	-0.23

Table 3.10: Progesterone signaling in the experiment of the effect of RANKL on 6 treated mammoplasties; 2 with high level of progesterone (P4) (211, 212) and 4 with low level (210, 215, 217, 230). Individual deviation components values are displayed for *PGR*, *RANKL*, *WNT4*, *MYBPC1* and *CXCL13*.

The qPCRs in the two mammoplasty for progesterone receptor (*PGR*) showed different patterns and inconsistent with RNA-seq (Fig. 3.26a, b).

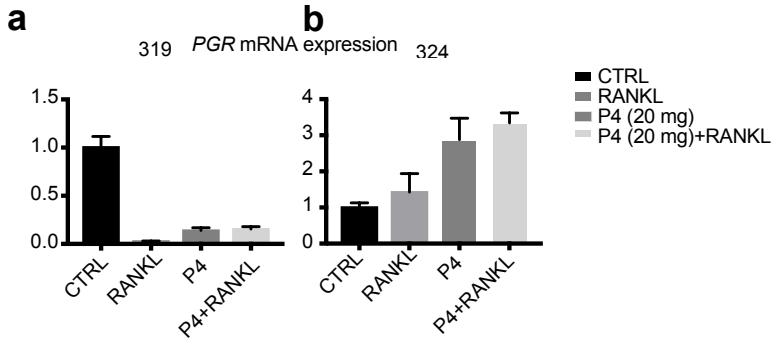


Figure 3.26: Validation of Progesterone signaling genes: *PGR*. **(a)**, **(b)** *PGR* mRNA measured by qPCR on individual mammoplasties injected intraductally into recipient mice and kept for 14 days after they established and treated with control or progesterone, cells were taken out and treated with RANKL or control. Therefore, 4 conditions (control, RANKL, P4 and P4+RANKL) are found on mammoplasties **(a)** number 319 **(b)** number 324.

RANKL also known as RANKL showed a consistent increase upon progesterone treatment in both mammoplasties (Fig. 3.27a, b), further treatment of RANKL did not show the expect result as an increase in RANKL was observed (Fig. 3.27a, b).

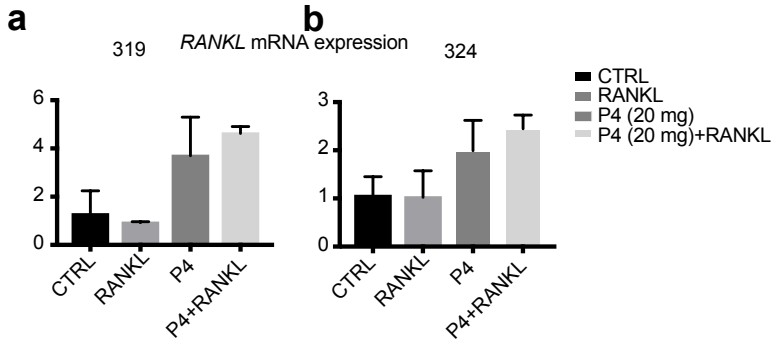


Figure 3.27: Validation of Progesterone signaling genes: *TNFSF11*. **(a)**, **(b)** *TNFSF11* mRNA measured by qPCR on individual mammoplasties injected intraductally into recipient mice and kept for 14 days after they established and treated with control or progesterone, cells were taken out and treated with RANKL or control. Therefore, 4 conditions (control, RANKL, P4 and P4+RANKL) are found on mammoplasties **(a)** number 319 **(b)** number 324.

We then proceed with the validation of *WNT4* by qPCR, which confirmed our findings; *WNT4* mRNA is upregulated in treated cells with progesterone compared to control (Fig. 3.28a, b). Moreover, treatment of RANKL in cells treated with progesterone compared to progesterone-treated cells alone showed a decrease (Fig. 3.28d). RANKL treatment on cells without progesterone treatment showed no changes (Fig. 3.28a, b). Averages of both mammoplasties show that *WNT4* mRNA follows the results found in sections 3.6 an 3.5 (Fig. 3.28c, d), although they do not reach significance yet. A third sample with available patient information would shed light into this validation.

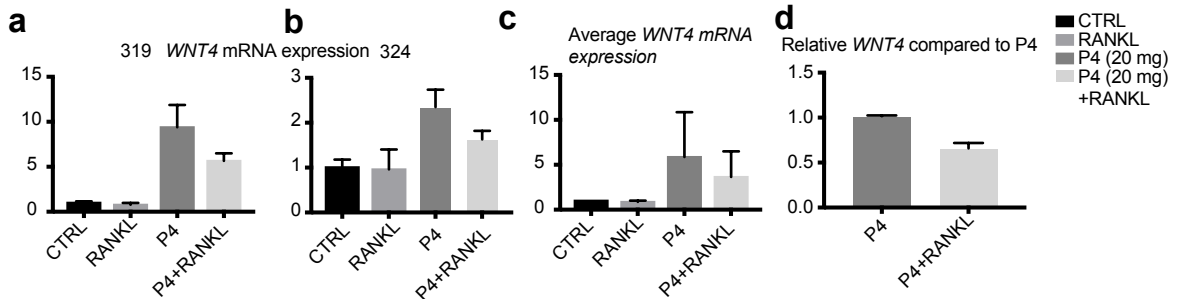


Figure 3.28: Validation of Progesterone signaling genes: *WNT4*. **(a)**, **(b)** *WNT4* mRNA measured by qPCR on individual mammoplasties injected intraductally into recipient mice and kept for 14 days after they established and treated with control or progesterone, cells were taken out and treated with RANKL or control. Therefore, 4 conditions (control, RANKL, P4 and P4+RANKL) are found on mammoplasties **(a)** number 319 **(b)** number 324. **(c)** Average of two independent experiments of the 4 conditions and **(d)** *WNT4* mRNA levels of P4 + RANKL cells relative and compared to P4 only levels.

For *MYBPC1*, only mammoplasty 324 validated the results of the mammoplasty and RNA-

seq (Fig. 3.29b). Whereas in mammoplasty 319, progesterone levels are not affecting the expression of *MYBPC1*, in mammoplasty 324 it induces the expression of *MYBPC1* (Fig. 3.29a, b). The further treatment of RANKL on those cells showed a decrease of *MYBPC1* mRNA in 324 (Fig. 3.29b). Even when averaging the results of both mammoplasty the result are going along with the previous results, but does not reach significance yet (Fig. 3.29c, d). A third sample with available patient information would shed light into this validation.

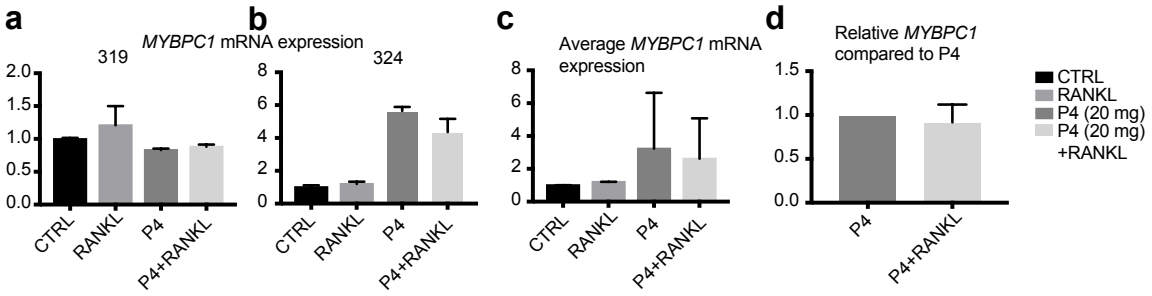


Figure 3.29: Validation of Progesterone signaling genes: *MYBPC1*. **(a)**, **(b)** *MYBPC1* mRNA measured by qPCR on individual mammoplasties injected intraductally into recipient mice and kept for 14 days after they established and treated with control or progesterone, cells were taken out and treated with RANKL or control. Therefore, 4 conditions (control, RANKL, P4 and P4+RANKL) are found on mammoplasties **(a)** number 319 **(b)** number 324. **(c)** Average of two independent experiments of the 4 conditions and **(d)** *MYBPC1* mRNA levels of P4 + RANKL cells relative and compared to P4 only levels.

While both mammoplasty were displaying a positive response of *CXCL13* mRNA upon progesterone treatments (Fig. 3.30a, b), only mammoplasty number 319 (Fig. 3.30a) reproduced the previously observed results on RANKL, i.e. a reduction of the levels of *CXCL13* upon treatment of RANKL in progesterone induced cells. This might be explained by the difference in the two mammoplasty of basal level in control cells (average cycle of 32, compared to 38).

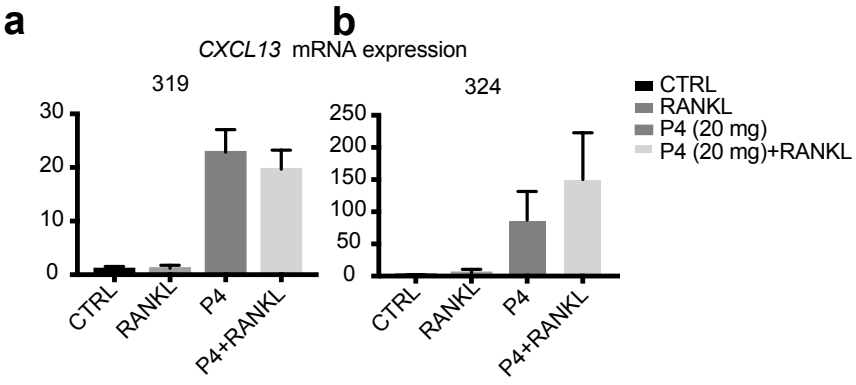


Figure 3.30: Validation of Progesterone signaling genes: *CXCL13*. **(a)**, **(b)** *CXCL13* mRNA measured by qPCR on individual mammoplasties injected intraductally into recipient mice and kept for 14 days after they established and treated with control or progesterone, cells were taken out and treated with RANKL or control. Therefore, 4 conditions (control, RANKL, P4 and P4+RANKL) are found on mammoplasties **(a)** number 319 **(b)** number 324.

Conclusions on the validation

As shown by TTMap (see section 3.6), RANKL mostly shows no effect on mRNA on the tested genes in cells that have not been previously exposed to progesterone specially on *JAK2*, *RANKL*, *WNT4*, *CXCL13* and *TFF1* mRNA levels.

For the effect of RANKL on the cells pre-treated with progesterone, no significant results are observed when grouping the two mammoplasty reflecting that each mammoplasty has a different manner to react to treatments as also observed in [224].

JAK2 and *WNT4* show reproducible result compared to the microarray experiment (see section 3.6) and the RNA-seq experiment (see section 3.5). Therefore, *WNT4* is induced by progesterone signaling, whereas RANKL reduces this increase and *JAK2* mRNA increases upon progesterone signaling and is abrogated upon RANKL treatment.

Other genes show either validation of the findings in mammoplasty 319 (*CXCL13*) or 324 (*MYBPC1*,*GREB1*) urging the need to add more samples for validation.

3.7 Gene expression changes in the breast epithelium during the menstrual cycle

The Susan G. Komen for the Cure Tissue Bank at the IU Simon Cancer Center was established as a resource of breast tissue samples from healthy women volunteering to donate a biopsy of their breast in order for researcher to gain knowledge on the biology and developmental genetics of the normal mammary gland [233]. A better understanding of the normal biology of the breast might distinguish the early events triggering breast cancer and therefore implicate an improved prevention of breast cancer [234].
A first step towards a characterization of the normal biology of the breast is to understand the impact of the fluctuation in hormonal levels throughout the menstrual cycle on the molecular level of genes. Therefore, 20 biopsies of healthy premenopausal women were selected based on abundance of epithelial tissue on hematoxylin and eosin-stained sections of the Formalin-Fixed and Paraffin-Embedded tissue in order to facilitate laser capture microdissection [142]. This technique permits the dissection and capture of cells and was used in order to enrich for epithelial cells for a more informed view on the changes through the menstrual cycle among only epithelial cells and for eliminating the bias of cell type content. Hence, microdissected epithelial parts were captured and mRNA was extracted from these samples and sequenced using RNA-seq.
At the day of surgery, women are guided through a questionnaire on their reproductive history (e.g. last menses, length of cycle, parity) [142]. Blood is drawn from the patient to assess the hormonal levels (estrogen, progesterone and LH) simultaneously to biopsy retrieval. Therefore, the 20 samples could be classify into follicular (9 women) and luteal phase (5 women) according to the last menses and confirmed by progesterone levels. Moreover, six women using hormonal contraception at the time of donation were also included.

Goal. *The major goal of this study is to find molecular differences between breast tissue from follicular and luteal phase women and understand the impact of hormonal contraceptives on healthy premenopausal breast epithelial cells. Considering the patients in follicular phase currently not under hormonal contraceptives ($n = 9$) as the control samples, we assess with TTMap how far each gene expression profile from patient from the luteal phase and each profile from women under contraceptives are to the follicular phase. Also this dataset enables us to question if the molecular profile of women under contraceptives is similar to non-takers. Lastly, TTMap permits the classification of women from closest to furthest to the control and we can question if there is a correlation between this deviation and hormone levels.*

As can be seen in the MDS plot (Fig. 3.31) there are 3 batches, two that were composed of 10 samples and 9 samples respectively and one batch that consisted only of a point (sample 19). The first part of TTMap also highlighted this sample as an outlier (Fig. 3.32).
The only African American (AA) woman in the cohort was revealed by TTMap as the woman with the overall gene expression profile which is the furthest away from the control (Fig. 3.33). Evidence from breast cancer studies suggest that breast tissue from AA are extremely different from the general population. Indeed, epidemiological studies show that AA women are prone to particular type of breast cancer [235] and different links between breast cancer risk and hormonal history are described [236]. Moreover, AA are more likely to be diagnosed with triple negative breast cancer which is less common in the general population [237].

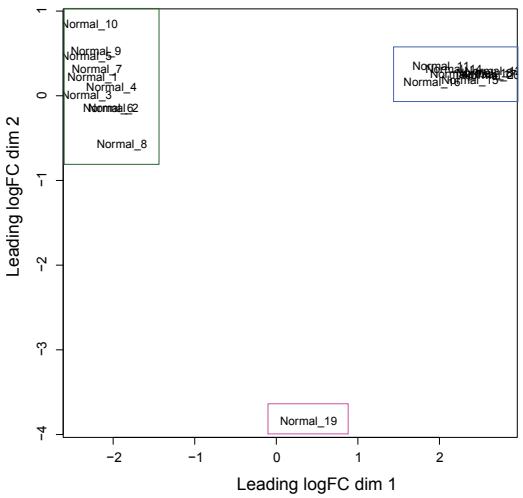


Figure 3.31: MDS plot of the expression profiles obtained from extracted RNA from breast biopsies of 20 healthy premenopausal human donors, batches are represented with different colors.

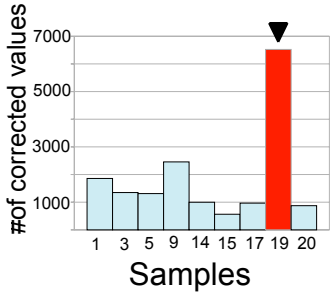


Figure 3.32: Control adjustment part of TTMap on the data from [142]. One sample showed a 5-fold increase in the number of genes considered as outliers (red, sample 19).

Apart from outliers (sample 2, 8 and 10), TTMap formed 2 major clusters in the overall level. The first group, which is the closest to the control samples is composed of samples 11, 12, 13, 16, 18. This biggest cluster could be separated into the quartiles of the function and revealed that sample 12, 13 and 18 were the closest to control (Fig. 3.33, 1st. Q.). Sample 12 was in luteal phase, but still had low level of progesterone and samples 13 and 18 are under oral contraceptives with the same type of progestin.

The second subgroup consists of the sample 4, 6 and 7. Sample 4 and 6 are gene expression profiles originating from women taking oral contraceptives which contain however different type of progestin. These both had a high deviation compared to follicular profiles.

These two different groups have well defined genes that change compared to follicular phase comprising all the 255 genes found by standard tools [142]. We observed that inside each batch (Samples 1 to 10 / 11 to 18 and 20), excluding women taking hormonal contraceptives, gene expression profile of women are ordered by amount of progesterone in the blood, i.e. the higher the progesterone the higher the deviation compared to control (Fig. 3.34). This was not the case for estradiol levels, menstrual cycle day, nor age. We can therefore suggest that the biggest change inferred on the breast cells through the menstrual cycle is driven by

progesterone levels.

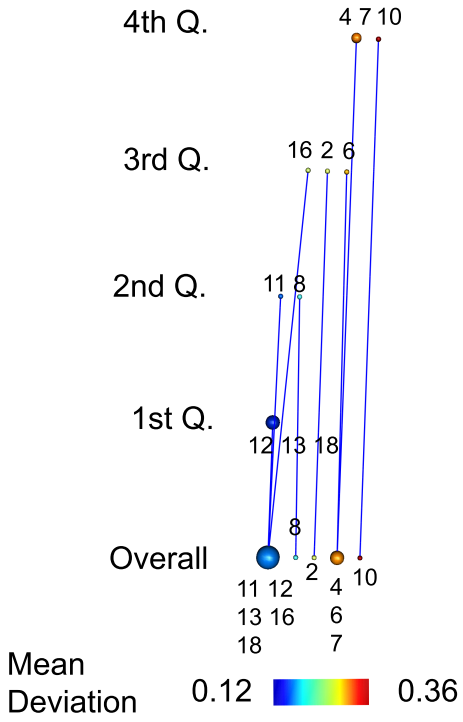


Figure 3.33: TTMap on the expression profiles obtained from extracted RNA from breast biopsies of 20 healthy premenopausal human donors where follicular-phased women are considered as controls, Q. = Quartile.

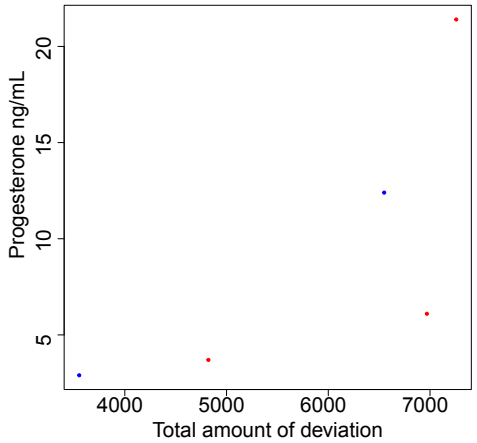


Figure 3.34: Correlation plot between the total absolute amount of deviation of the expression profile of the five luteal-phased women on the x-axis and the progesterone level (ng/mL) on the y-axis, colors represent different batches.

Remark 3.7.1. Several other dataset were generated in the lab that could be used to complete this one. First of all, the microarray dataset from section 3.6 comprises two luteal phase samples and four follicular phase samples. They were sequenced for the experiment explained

in that section, but can be used to assess the differences, in EpCAM+ cells only, in luteal against follicular-phased gene expression profiles. The difference between these two datasets is also that these cells were kept *ex-vivo* as microstructures. Both were extracted from healthy premenopausal women donors. Over 700 significant genes were found.
The second dataset consists of laser microdissected samples from 3 follicular and 3 luteal phase samples and RNA-seq was performed on the epithelial enriched parts. This resembles most closely the dataset mentioned above and was generated for the same purpose. Only 21 genes reached significance by adjusted p-value, among which the known progesterone receptor target genes *RANKL*, *DIO2*, *CXCL13* [221], [142], [118].

Lasty, aged-matched women with low level of progesterone ($n = 8$) and high level of progesterone ($n = 8$) were selected from our biobank and RNA was extracted from the full tissue, from the sorted luminal and from myoepithelial cells. Among the 8 women with low level of progesterone, 2 were categorised into follicular phase and the 6 others either are under hormonal contraception or did not provide the answer to the question of hormonal contraception. This dataset was generated in order to determine changes in different human cells when progesterone is high compared to when progesterone is low in aged-matched samples. With standard analysis comparing the two follicular phased sample to the 8 luteal phased samples in the luminal cells (that would correspond more closely to the setting above) no significant genes were found by adjusted p-value reflecting the fact that there are only 2 samples in one of the two groups. However, three milk proteins *CSN1S1*, *CSN3*, *LALBA* and a progesterone target gene *CXCL13* were the highest in terms of fold changes compared to follicular phase.
These datasets were analysed at the beginning of the thesis when TTMap was not developed yet, and therefore only standard analyses were used to find significant genes.

3.8 Applying TTMap to other types of datasets

In the previous chapters, we have appreciated applications of TTMap to RNA-seq data sets (section 3.4, 3.7, 3.5) and to microarray data (section 3.3, 3.6). In this section, TTMap is applied to various types of data to uncover generalisation of application of the method. We applied TTMap to neuronal data spike, i.e. measurements of spiking of neurons in brains of rats and mice 3.35, and to metabolomic data 3.8.2 of patients with chronic depression and early psychosis.

We are currently also testing an extension of TTMap to single-cell RNA-seq analysis, which due to the nature of the data, which is sparse (see section 1.4.4), needs modifications in HDA (see 2.2) and in the mismatch distance calculation.

3.8.1 TTMap on neurological data with the Human Brain Project

To understand which neurons of the brain are spiking upon certain impulses, electrodes have been placed on mice and rats in specific brain regions. This enables to measure average and maximum spike height, spike interval length, spike numbers, among others. Brain gray matter is divided into 6 layers (L1-L6) and different cell types have been assessed (Pyramidal cells (PC), dopaminergic (DA) neurons, fast-spiking (FS) neurons, stem cells (IN), Amygdala cells (Amyg)).

Dopaminergic neurons were considered as controls and the others cell types were considered as test samples. Batches correspond to mice and rats cells. Data has been Z-scored and missing values were considered as 0.

Goal. *We use TTMap to characterize the spiking pattern of the different neuronal cells across mice and rats.*

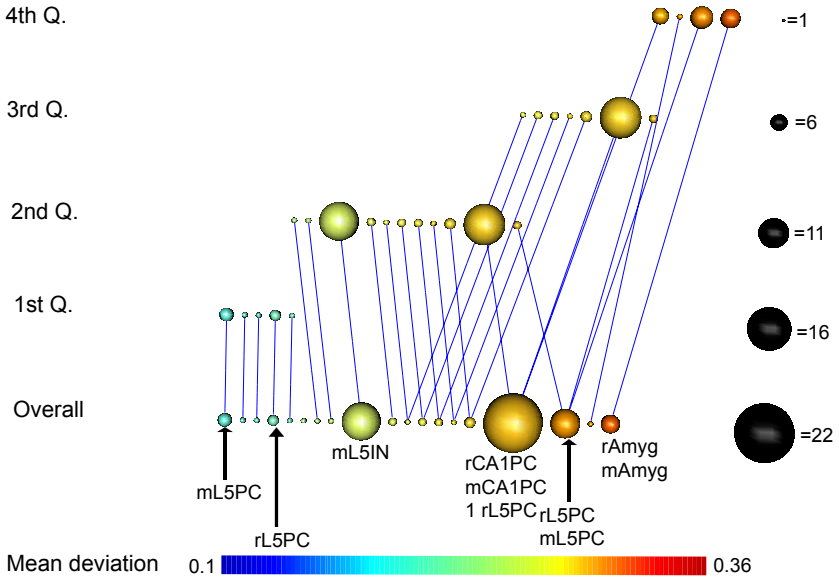


Figure 3.35: TTMap on brain spike measurements of different cell types compared to dopaminergic cells. Several subgroups are constituted of the same cell types of mice and rats (L5PC = pyramidal cells in layer 5, L5IN= stem cells in layer 5, CA1PC = pyramidal cells in layer CA1, Amyg = Amygdala cells). Outliers, i.e. groups of single samples, are not shown.

TTMap characterizes several well-defined groups that match cell types disregarding the animal of origin (rats and mice) (Fig. 3.35). This was not discerned by standard tools as the data has many NAs (Supplementary Fig. S26). These subgroups display similar differences in certain measured parameters compared to dopaminergic cells. Analysis of those features permits a better characterization of the variation compared to dopaminergic cells. L5PC cells divided into two distinct groups. Further analysis of these subgroups might reveal a subclassification in terms of spiking pattern of those cells. This dataset had a consequent amount of missing value which often constitutes a problem for standard data analysis. As TTMap operates with ranges of control values, NAs did not impinge on the results.

Data availability
This dataset was generated by Rodrigo De Campos Perin and colleagues at the EPFL, who provided it to us. For more information contact : rodrigo.perin@epfl.ch.

3.8.2 TTMap on metabolic data

The metabolome of healthy individual, of patients with chronic depression (C) and of patients with early psychosis (EP) events have been assessed in order to detect markers distinguishing the two stages. C patients and EP patients are often difficult to distinguish [238]. Therefore, it is of considerable clinical interest to discover hints for a more informed classification of these stages. The metabolome of each patient was derived from their blood using liquid chromatography-mass spectrometry.

Goal. *We use TTMap in order to understand if a subtyping of the metabolome of patients with early psychosis or chronic depression can be made and if these profiles can be ordered in how close they are to the metabolome of healthy individual.*

TTMap showed a core cluster of samples with early psychosis and chronic depression and several outliers, all EP patients (Fig. 3.36). The distribution of the samples in the quartiles revealed that all the chronic patients were in the lower quartile, and therefore had a lower total amount of deviation compared to healthy individuals, than EP, with the exception of one sample. Therefore, chronic and early psychosis patients have similar metabolic changes, but chronically depressed patients are closer to the control samples. Moreover, outliers identified in the overall clustering reflect EP patients with metabolic changes that are significantly higher and different than other samples, a precise analysis of those metabolites would shed light into the disease as they are specific to the early psychotic state. The only chronic patient who was in a higher quartile was younger (Fig. 3.36, red box). TTMap showed that young patients of EP and C between 20 and 30 years old were all assigned to higher quartiles, and therefore deviate more compared to control (Fig. 3.37, red box), showing that the metabolic changes are more noticeable in younger patients, but are the same than for older patients as they cluster together in the overall clustering.

Data availability
This dataset was generated by Margot Fournier and colleagues at the CHUV, who provided it to us. For more information contact : margot.fournier@chuv.ch.

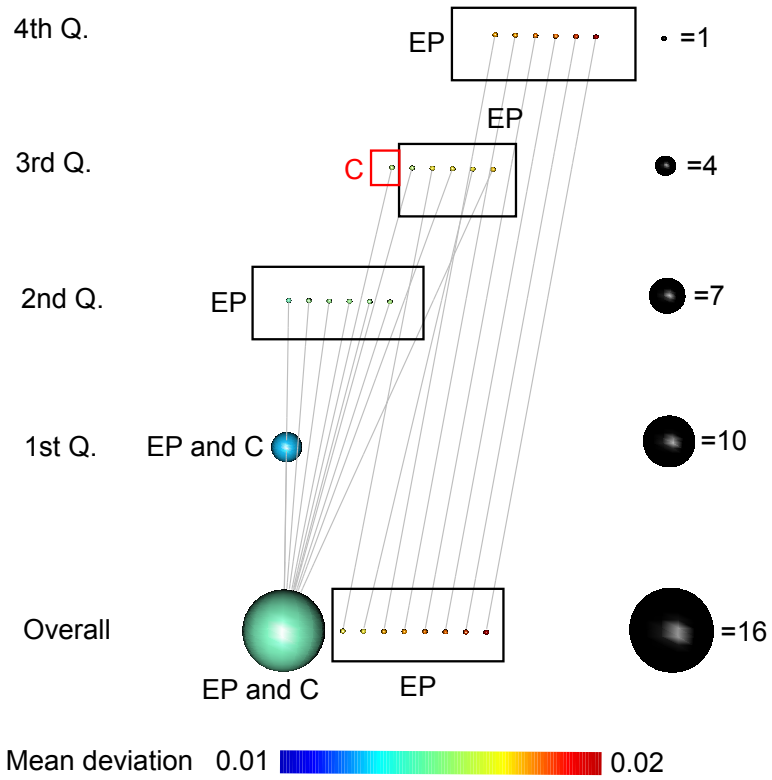


Figure 3.36: Output of TTMap comparing metabolic data obtained from blood of healthy individuals and patients with psychotic events, either patients with early psychotic events (EP) or chronically depressed patients (C). Q. = Quartile.

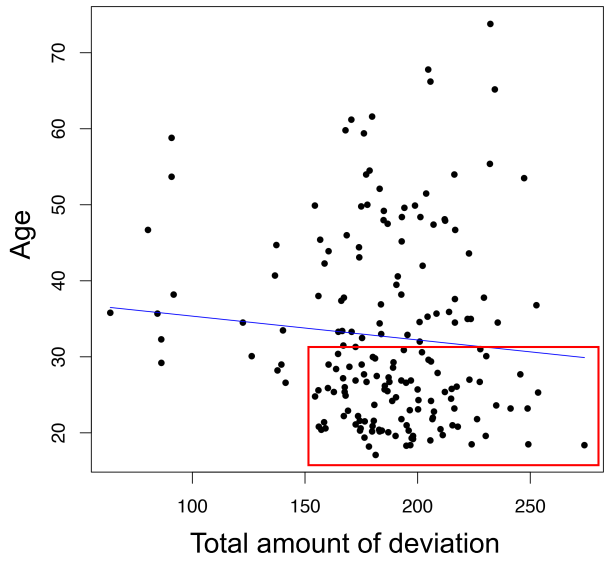


Figure 3.37: Correlation plot of total amount of deviation compared to age, blue line is a linear fit, red box represents samples between 20 and 30 years old.

4 Discussion

4.1 Discussion

(This section is Adapted from: "Two-Tier Mapper: a user-independent clustering method for global gene expression analysis based on topology", R. Jeitziner, M. Carrière, J. Rougemont, S. Oudot, K. Hess, and C. Brisken, 2017, arXiv: 1801.01841 [194], submitted to Bioinformatics.)

We have developed a topology-based clustering tool, Two-Tier Mapper (TTMap) that calculates and relates individual deviation from a given control group. TTMap outperforms existing clustering tools especially when dealing with small sample numbers.

We validated the method by analysing *in silico* data (section 3.2), several biological data sets, including microarray data from Drosophila data and human data respectively (sections 3.3 and 3.6), RNA-seq data from mice and human data (sections 3.4, 3.5 and 3.7), metabolic data (section 3.8.2) and on neuronal spike data (section 3.35), and by proving theoretical aspects of the stability of the method (section 2.4).

TTMap outperforms existing clustering tools, with particular strength for small sample numbers. TTMap identifies subgroups in the dataset, even reliably discriminating subgroups composed of a single sample. It should therefore be used as a first-line analysis tool on the dataset to reveal outliers within the control and the test group and clusters with well-defined associated gene signature. The gained insights on the data guide further statistical analyses, applied in a second phase.

Thanks to the two-tier cover, the algorithm is theoretically stable, as expressed precisely in three stability theorems. This cover not only provides the global clusters in an unbiased manner, but provides additional local information using a filter function that yields deeper insights into the composition of the clusters. Having a control group enabled us to define a new topological type of distance on the samples leading to an enhanced view on the data. TTMap characterizes the control group by finding outlier values and samples. This is useful as it can improve standard statistical methods for differential expression analysis.

TTMap provides an ordering of the test samples reflecting their proximity to the control samples. This reveals possible samples that are in-between the two compared groups by samples that are extremely close to control, and samples that are outliers since they are the furthest to the control group.

Popular clustering method as well as Mapper depend on parameters that are challenging to select [92], [95]. Our improved and extended version of Mapper includes an optimized parameter selection, making it user-independent for global gene expression analysis, and performs well independently of sample sizes.

TTMap offers the advantage that it can group analyses with batches into one analysis instead of analysing separately the data in each batch and grouping the results afterwards. This process strengthen the subgroup discovery, which is not obtained according to batches but according to shape of deviation reflecting the biological changes.

TTMap does justice to biological complexity and detects significant subgroups within a cluster. This was illustrated by the clustering of data using TTMap on different strains of mice in the estrous cycle, considered as different batches, which demonstrated the existence of samples that are in-between two phases, and revealed subgroups that reflect possible alterations of hormone levels, as they have differentially expressed genes known to vary along the human menstrual cycle [142] or are under control of progesterone [215]. These genes are invisible to standard tools, since they are significantly expressed only in those subphases of the estrous cycle. Hence, a reclassification of the estrous cycle into more than 3 phases might be necessary

in order to reflect hormonal changes as well as physiological changes in vaginal cytology.

TTMap revealed that RANKL (section 3.6), a protein linked to lactation in the mammary gland [189] and which is under the control of progesterone receptor signaling [118] (see section 1.6.9) exerts significant effect on mammoplasty tissue only when the tissue is obtained from patients in the luteal phase of the cycle, i.e. having high levels of progesterone. This suggests that careful attention should be paid when stimulating human cells with RANKL as the results vary depending on the hormonal history of the patients.

TTMap demonstrated that R5020, a synthetic progestin, has the same role than progesterone on epithelial cells, but to a higher extent, demonstrating that R5020 is a more potent compound than natural progesterone.

TTMap showed that the level of progesterone of the patient at the time of surgery influences the response of the epithelial cells to the treatment of progesterone and R5020, since a lower extent of deviation of gene expression profiles was observed in patient with higher levels of progesterone.

Moreover, the extent of deviation of gene expression profiles from young patients treated with progesterone or R5020 is higher than from older women. Therefore, cells from younger patients are more responsive, whereas cells from women with high level of progesterone are less responsive to treatment with progesterone and R5020. Progesterone induces proliferation on epithelial cells [126], [117], and increases breast cancer risk. Hormonal contraceptives as they are composed with a synthetic progestin, might act in a similar way than R5020. Our results would provide an explanation to the observed increase in breast cancer risk among young women taking oral contraceptives [170], as TTMap showed that R5020 had a higher impact on the cells provided by a mammoplasty from a young woman. We also provided first-line of evidence that GREB1 is a progesterone-responsive gene in human breast epithelial cells which was also shown to be the case in human endometrial stromal cells [231]. This findings could be reproduced by qPCR, but need further replicates to reach significance. As in the stimulation with RANKL, each mammoplasty showed a different pattern of deviation, and we realised that experiments dealing with mammoplasty tissues should be carefully interrogated for patient history in order to draw the right conclusions.

Furthermore, combining these two analyses, we discovered that the action of RANKL on these particular cells having been exposed to luteal levels of progesterone antagonises the action of progesterone on human breast tissue (section 3.6.4). The commonly changed features revealed a link with apoptosis reflecting that RANKL and progesterone might act together for blocking apoptosis in healthy epithelial cells. Both RANKL and progesterone action increased genes linked to proliferation, but not the same genes, suggesting that they might use complementary pathways to induce proliferation. This sheds light into the interplay of RANKL and progesterone and could not be found using standard analyses.

We appreciated in several sections (3.2.4, 3.3.5, 3.4.1) that both steps, HDA and GLMap, of TTMap are compelling to discern desired insights on the data.

The method is available as a freely downloadable library "TTMap" in Bioconductor, enabling widespread application of this useful tool.

4.1.1 Discussion on the differences between TTMap, PAD and to the standard Mapper algorithm

The main differences between TTMap and the standard use of Mapper is the choice of parameters and the data pre-treatment. Because of the parameter selection, each of the two version of Mapper are producing a distinct output and should be interpreted in a different way. If parameters are well-chosen, then the insights produced by the original way of using the Mapper algorithm (OMap) [10], [8], [11], i.e. the algorithm with the parameters selected in the way described in section 1.2.9, reflect the topology with regards to a filter function chosen appropriately of the underlying space on which the point cloud lies. The TTMap algorithm in turn describes one topological aspect, the connected components, locally and globally. Hence, both algorithms could be used in a complementary manner. To understand more precisely the advantages of using one against the other, here is a summary of the points on which they differ. More precisely, progression analysis of disease (PAD) [9] has been described as an optimized version of OMap for the analysis of gene expression dataset, hence our focus will be on comparing the two methods.

4.1.2 Disease-specific-genomic-analysis compared to hyperrectangle deviation assessment

Method

The first part of PAD uses several times linear regressions (LR). These LR processes are highly affected by outliers [199] and are not suitable for small datasets. Therefore, in TTMap we decided to replace this step by the hyperrectangle deviation assessment.

Outputs

The resulting "disease components" of PAD, which are comparable to TTMap's deviation components, are difficult to be interpreted due to the LR; a +1 does not mean 2-fold increase in expression, it can be an artefact of LR processes. In TTMap, a deviation of -1 for a certain feature is precisely reflecting a 2-fold decrease, since the data is in $\log_2$-scale.

Parameter selection

Both PAD and TTMap have a parameter to select for this part: the Wold's invariant for PAD and e for TTMap. They are chosen by visual inspection of the Wolds plot that sometimes is equivocal for PAD or by the data using the variances for TTMap.

4.1.3 Original use of Mapper compared to Global-to-Local Mapper

The selection of a cover

In OMap, the chosen cover of the real-line requires selection of parameters and no guidance is provided towards these choices. It is given by a number of intervals with a certain percentage of overlap (section 1.2.9). The selection of those parameter depend on the given problem [19], [239], [9], [13], [240]. Furthermore, it has not been demonstrated that the chosen parameters for a given problem are general enough for similar data. For instance, for temporal single-cell RNA-seq [19] several different parameters were chosen for the same type of data. Some attempts to guide the choice of parameters for the cover, i.e. the number of intervals and the percentage of overlap, have recently been made [241]. It is however unclear if this methodology is suitable for small data set since it is stated that the sample size needs to be large enough

in order to prove that the right parameters are chosen [241].
In TTMap, two covers are considered at the same time, a local one and a global one. Instead of tracking changes across a full range of covers as in [24] with the Multiscale Mapper, here only the local and global features are considered and directly related. TTMap is the first application of Mapper with a two-tier cover.

The selection of a distance
A new distance (see section 2.3.1), called the mismatch distance, was introduced in TTMap, similar to the Hamming distance [242]. This is useful in the analysis of gene expression, where the extent of deviation from the control is less important than the pattern of deviation. Any other distance can be entered, which enables a lot of flexibility. In some versions of the OMap, the distance can also be freely chosen [243], others have restricted choices [27].

The parameter of closeness ε
Linked to the distance is the cutoff parameter ε, which in OMap remains mysterious as it is not stated as a number in the applications, it can not be freely chosen [27]. It is linked to the number of intervals chosen for the cover [27], or is given as a number that determines the final number of clusters [243] introducing a huge user-bias. In TTMap (section 2.3.4), ε was optimised for gene expression data and represents significant subgroups in the data as it uses probabilities, but can also be chosen freely by the user to accommodate specific needs.

Significant features
TTMap detects subgroups in the dataset and then determines the features distinguishing them, which requires further work with OMap (subgroups need to be selected by visual inspection of the graph and further tests such as Kolmogorov-Smirnov tests [240] are applied).

Theoretical stability
On the theoretical side (see section 2.4), the OMap studies the topological features distinguishing two objects [8] and persistence diagrams [202] are extensively used as descriptors of those Mapper outputs. By focussing on the connected components only, but at different levels, and by adding to the extended persistence diagram the additional information of links between local and global features we defined a new type of descriptors of the TTMap graphs (Definition 2.4.1). These new descriptors are enlarging the field of persistence diagrams. The results on the persistence diagrams in [204] were adapted to the setting of this method, and unlike the OMap, the descriptors of TTMap are complete, ensuring the stability of not exclusively the descriptors but also of TTMap, which confirmed the strength of the method. Since the stability of TTMap is assessed theoretically and practically, we proved that it is a reliable clustering tool [26].

4.1.4 Future developments and outlook
As implemented here, the one-dimensional real filter function takes into account only one specific aspect of refinement. To further enhance the method, one can filter by any metadata, such as categorical information and numerical data. This flexibility enables the user also to interrogate the data in various ways. All outputs can be compared, as the global clusters are

independent of the chosen filter function. However, we imagined a generalisation of TTMap to accept as input filter functions $f : X \to R^n$ to enable to view all these outputs in one picture. In this situation, keeping the notations of 2.3.3, the chosen cover of the image of f without multiplicity is given by

$$\mathfrak{I} = \{\pi_1 \circ \operatorname{Im} f, \tilde{\pi}_1^{-1} q^{\tilde{\pi}_1 \circ f}_{[0,25[}, q^{\tilde{\pi}_1 \circ f}_{[25,50[}, q^{\tilde{\pi}_1 \circ f}_{[50,75[}, q^{\tilde{\pi}_1 \circ f}_{[75,100]}, \ldots, \tilde{\pi}_n^{-1} q^{\tilde{\pi}_n \circ f}_{[0,25[}, q^{\tilde{\pi}_n \circ f}_{[25,50[}, q^{\tilde{\pi}_n \circ f}_{[50,75[}, q^{\tilde{\pi}_n \circ f}_{[75,100]}\},$$

where $\tilde{\pi}_j : R^n \to R$ is the projection onto the j-th component, $\operatorname{Im} f$ denote the *image* of f *with multiplicity*, i.e.,

$$\operatorname{Im} f = \{(f(X), \sigma) \mid X \in \mathsf{T},\ \sigma \in \{1, \ldots, \operatorname{mult}(X)\}\} \subseteq \mathbb{R} \times \mathbb{N},$$

with the lexicographic order, where $\operatorname{mult}(X) = \operatorname{card}(f^{-1}(f(X)))$ is the multiplicity of $f(X)$ and for any $0 \leq a < b \leq 100$, let

$$q^r_{[a,b[} = \pi_1\Big(\Big\{y \in \operatorname{Im} r \mid \operatorname{quantile}_a(\operatorname{Im} r) \leq y < \operatorname{quantile}_b(\operatorname{Im} r)\Big\}\Big),$$

where $\pi_1 : \mathbb{R} \times \mathbb{N} \to \mathbb{R}$ is the natural projection on the first component, and $\operatorname{quantile}_a(\operatorname{Im} r)$ is the a-th quantile of the ordered values in $\operatorname{Im} r$.

This cover of $\mathbb{R}^n$ (Fig. 4.1) is a repetition in each dimension of the cover in one dimension, where all the other dimensions are fixed. Since the global tier is added to the cover of $\mathbb{R}^n$, each one-dimensional plot obtained by TTMap is linked to the global tier and linked with each local tier of each dimension. The idea is to draw a hotel-like structure, where in each floor one finds the i-th projection of the function. If the function is all the metadata available on the patients (e.g. one dimension is the age, the other the cancer type, then family history, then height, then BMI, ...), then at the i-th floor one finds the i-th metadata segregated into quartiles. (Work currently done by a master student Martino Milani, EPFL).

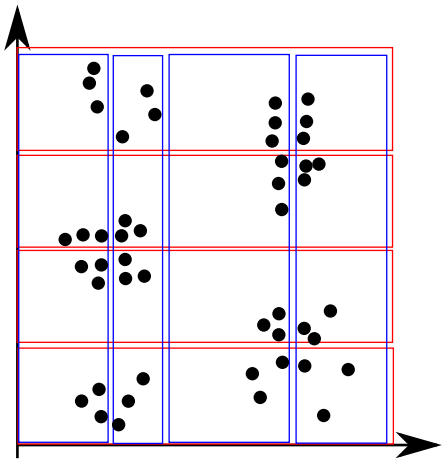


Figure 4.1: Proposition of a cover $\mathbb{R}^n$, with $n = 2$ for a generalisation of TTMap.

Another valuable ressource is the future development of a webpage which would include the current version of TTMap. All the steps would be entered via choices and clicks online. Moreover, by clicking on each sphere obtained on the TTMap graph a circular plot of the

metadata linked to that sphere would be available reflecting meta-data.
In section 2.3.4, we explained the selection of the parameter ε using the Chen-Stein method on identically and independently distributed variables. Since the theory also exists for dependent variables and since this might be highly applicable to gene expression data sets as genes work within a pathway or a network, hence are not independent variables, a possible future project would be to calculate ε using Chen-Stein's method on dependent variables [209]. The obstacle being to know prior to the analysis or estimate with the data which are the dependent variables.
A cluster in TTMap does not reflect precisely the distance between the samples. Another version of TTMap could be envisioned that upon a click on a sphere or with a closer view into a sphere one would visualize the Vietoris-Rips complex drawn on top of the N points in that cluster and each link is supplemented with a weight depending on the proximity of two samples either written next to the edge or represented as more or less thick edges (Fig. 4.2).

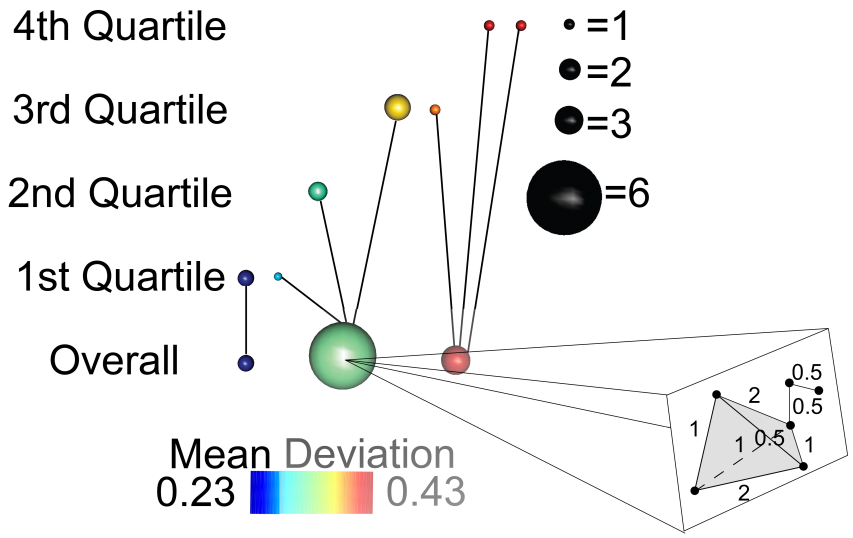


Figure 4.2: Extension of TTMap, with a zoom on a sphere that displays a Vietoris-Rips simplicial complex with weights.

The TTMap algorithm might be strengthen by replacing the single-linkage clustering step by other more powerful clustering algorithms such as DBSCAN. The situation of an outlier point between two clusters C_1 and C_2 could then avoid grouping the two clusters C_1 and C_2 together (Fig. 4.3).
TTMap produces individual profile of deviation from the control for each sample and relates it to other samples. This, together with its ability to account for batch effects, make it a promising tool for personalized medicine, where increasingly complex individual patient data need to be analyzed and related to other samples.

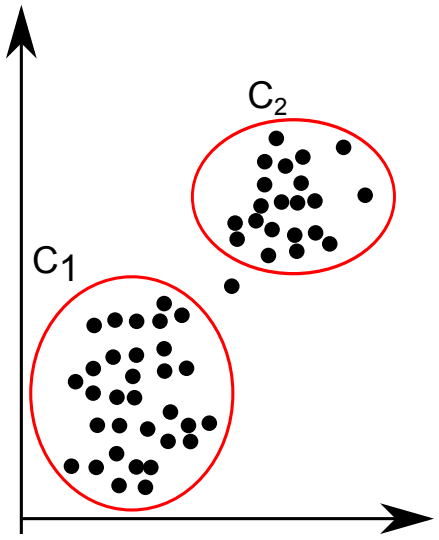


Figure 4.3: Plot of a point cloud with two clusters (C_1 and C_2) linked with one outlier point.

5 Conclusion

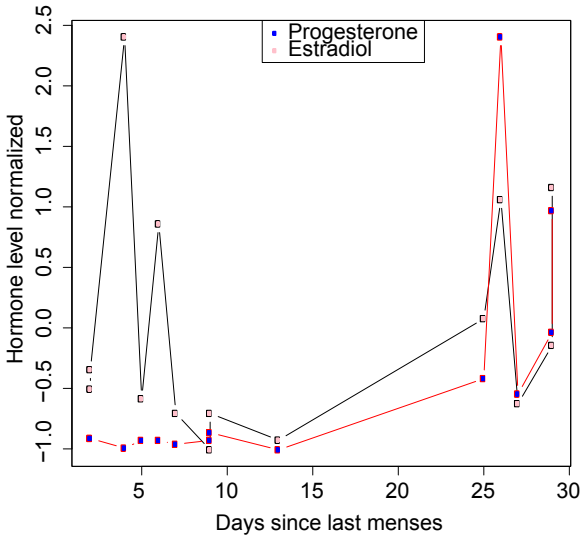
We have developed a new method for the analysis of global gene expression profile based on topology, called Two-Tier Mapper (TTMap).
All the parameters have been optimized in order to render this algorithm user-independent. This choice of parameters induced the theoretically stability of the algorithm, confirming the reliability of TTMap.
TTMap, with default parameters, is a two groups comparison that provides an individual profile of deviation, enabling precise decomposition of each sample. The method produces a visual output with subgroups and outliers easily identifiable. It associates to each cluster the significant features distinguishing it from the others.
TTMap defines a new category of clustering algorithm since it is at the intersection of the two major subtypes of clustering methods: the partitioning clusterings, as it provides subgroups and the hierarchical clusterings, as it gives a two-tier decomposition of the data, i.e. a local tier and a global tier.
TTMap provided previously unravelled insights into a wide variety of datasets ranging from microarray and RNA-seq on mice data, Drosophila data and on human data to metabolic data or even neuronal spike data. We also validated this approach on *in silico* data. We showed that it outperforms standard clustering algorithms in terms of subgroups discovery and stability. The complex biology of the mice estrous cycle could be further deciphered using TTMap, where subgroups of the different phases of the estrous cycle were found and related to hormonal action. New insights on the action and interplay of RANKL and progesterone on human epithelial cells were gained using TTMap that could not be obtained with standard methods.
We generate TTMap as an open source **R** package in Bioconductor to enable its broad usage.

6 Supplementary data: Ongoing studies

In this chapter, we report ongoing work. For that, we will introduce two further topological methods to analyse data, more specifically useful in the discovery of cyclical data, i.e. data that after a certain time returns back to its original value.
In section 6.1, we illustrate a possible application of multidimensional persistence to the study of hormone-responsive genes, with first implementations. It is based on the work of Frosini and Landi [31] that we explain in Appendix B.0.1.
The second application (section 6.2) of topology is using an adaptation of the method described by Arsuaga *et al.* [244], [245] to data from our lab.
Other ongoing work consists of 1) extending TTMap to single-cell RNA-seq data and apply it and 2) using TTMap to detect subgroups in breast cancer data obtained through a collaboration with Lund, Sweden, which provided us with 1079 breast cancer RNA-seq profiles to which was added 20 healthy mammoplasty RNA-seq profiles. The healthy tissues were provided from our biobank and shipped to Sweden for sequencing in order to reduce the bias due to the place of sequencing.

6.1 New method to determine hormone responsive genes using multidimensional persistence

Estrogen and progesterone are two major hormones fluctuating through the human menstrual cycle of premenopausal women, as can be observed in the graph of hormone levels (Supplementary Fig. S1) using the dataset of I. Pardo *et al.* [142].

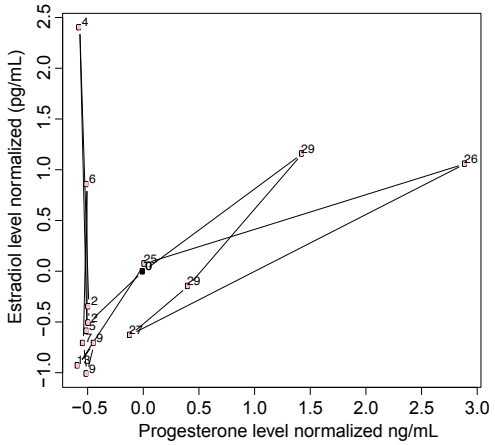


Supplementary Figure S1: Hormonal fluctuation through the menstrual cycle from dataset [142], data reported from Mass Spectrometry of blood from 14 women through the menstrual cycle with self reported last day of menses, progesterone blue, estradiol pink.

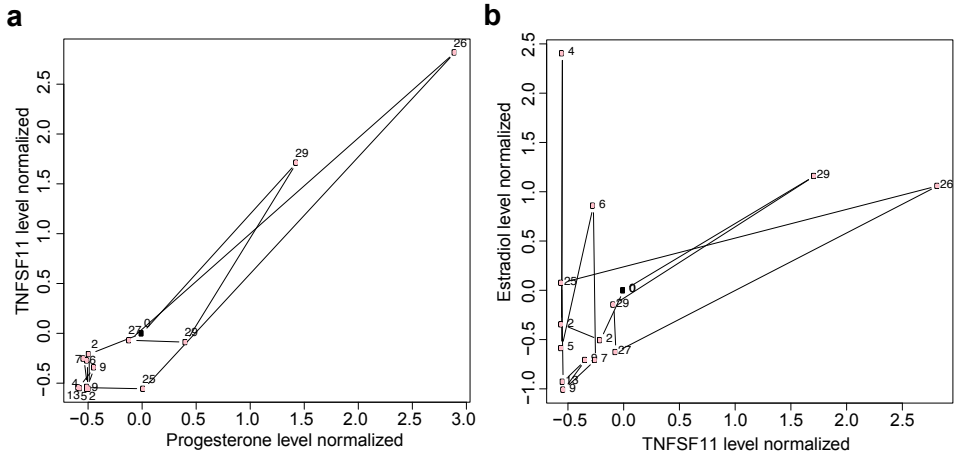
To discover a list of genes responsive to either progesterone or estrogen, as well as other measured hormones, and gain molecular insights into their respective roles in cells of the breast, we developed the following method.

Let $f : [0, 28] \to \mathbb{R}^2 : t \mapsto (P_t, E_t)$ be the function representing for each day t of the cycle the corresponding measured level of progesterone and estrogen, where $f(0) = f(28)$ (Supplementary Fig. S2). This function needs to be compared to the two functions $e_G : [0, 28] \to \mathbb{R}^2 : t \mapsto (P_t, G_t)$ and $p_G : [0, 28] \to \mathbb{R}^2 : t \mapsto (G_t, E_t)$, where G_t is the level of expression of a gene G measured for instance by RNA-seq or microarrays on day t of the cycle. The functions f, e_G and p_G can be seen as functions from S^1 to $\mathbb{R}^2$, and we want to determine whether there exists a diffeomorphism $h_e : S^1 \to S^1$ such that $|| f - e_G \circ h_e ||_\infty = 0$, which would imply that G is an estrogen responsive gene. The diffeomorphism (see Appendix B.0.1) is needed as it is unclear when the response is occurring, the same day or with a delay. The same question will be asked for p_G, i.e. find $h_p : S^1 \to S^1$ such that $|| f - e_G \circ h_p ||_\infty = 0$. An answer to that problem can be found using multidimensional persistence (the theory of multidimensional persistence and comparison of size functions is described in Appendix B.0.1, specifically theorem B.0.38), which reduces the problem to the calculation of homology groups and their ranks. We wrote a program in **R** that, specifically for gene expression and menstrual cycle data, calculates these ranks, i.e. $\text{rank}(H_0^{\cdot,\cdot}(\tau))$, where τ is a function $\tau : [0, 28] \to \mathbb{R}^2$.

It is of note that these ranks are computable, since we have a finite sampling of the points representing the functions e_G and p_G.
We are currently testing the method on the dataset generated by I. Pardo *et al.* [142], with positive controls *RANKL* and *WNT4* as they are known progesterone target genes. Plots of e_G (Supplementary Fig. S3a) and p_G (Supplementary Fig. S3b) for *RANKL* illustrate the fact that *RANKL* is closer to progesterone as p_G more closely resembles the function f in terms of shape (Supplementary Fig. S2). Indeed, e_G (Supplementary Fig. S3, left top corner) is missing the peaks observed on the day 4 and 6 (Supplementary Fig. S2, left top corner), whereas p_G displays them (Supplementary Fig. S3 b, left top corner). This is a first indication that the method should be strong enough to distinguish estrogen-responsive genes and progesterone-responsive genes, as RANKL was shown to be a progesterone-responsive gene [118].



Supplementary Figure S2: Function f representing for each day of the cycle (number above each point) the corresponding measured levels of progesterone (x-axis) and estrogen (y-axis) from dataset [142].



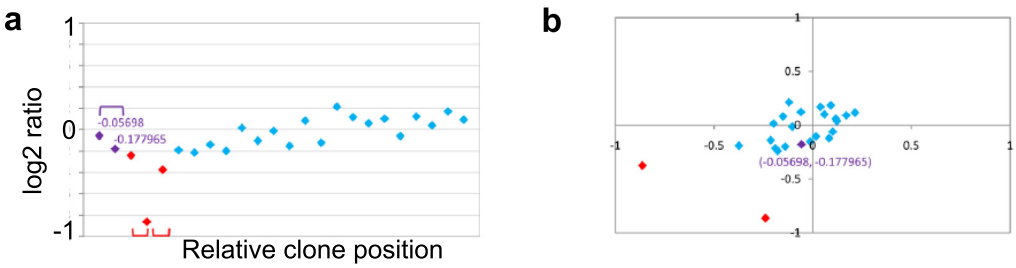
Supplementary Figure S3: Function e_G and p_G, where G is *RANKL*, representing for each day of the cycle (number above each point) the corresponding measured levels of **(a)** progesterone (x-axis) and the expression level of the gene G (y-axis), and **(b)** the expression level of the gene G (x-axis) and estrogen levels (y-axis) from the dataset [142].

6.2 Adaptation of a homological method to the analysis of gene expression data using TTMap or DSGA profiles

Inspired by the methods described in the articles of J. Arsuaga *et al.* [244],[245] that assessed the topological complexity of a comparative genomic hybridization (CGH) profile, we describe an adaptation of this method to the analysis of microarray data, which should be applicable as well to RNA-seq data. Since there is no direct comparison to a reference genome as is done with CGH profiles, the method developed for CGH could not be applied as is to microarray raw tables. Instead, we studied the vectors obtained from the DSGA method or the deviation components from TTMap, since they provide vectors resembling CGH profiles: low expression for genes that are only slightly altered with respect to a group of healthy or control vectors, and high levels for interesting genes.

Definition 6.2.1. For one sample, suppose there are m genes with corresponding values $\{x_i\}_{i=1}^m$ ordered by the localisation of the gene inside the chromosomes, order from 1-22 and X. A **point cloud generated with a sliding window** of the points $\{x_i\}_{i=1}^m$ of length n is the definition of m points in $\mathbb{R}^n$, v_i, where $i = 1, \ldots, m$, given by $v_i = (x_i, x_{i+1}, \ldots, x_{i+(n-1)})$, and $x_i = x_{i \bmod m}$, for all $i > m$.

This means, for example with $n = 2$, that the first two consecutive points will give rise to one point in $\mathbb{R}^2$, then the window moves from one gene and gene two and three give rise to a point in $\mathbb{R}^2$, and so on (Supplementary Fig. S4).



Supplementary Figure S4: **(a)** CGH profile and explanation of sliding window by two purple values and three red ones that are found in the **(b)** point cloud generated using a sliding windows of length $n = 2$ (Illustration adapted from [245]).

To determine the complexity of this point cloud in $\mathbb{R}^n$, Arsuaga *et al.* [244],[245] constructed the Vietoris-Rips complex with parameter $\varepsilon > 0$ (see definition 1.3.8), where the vertices are the points of the point cloud, and the distance between the points given by the euclidean distance between points.

Remark 6.2.2. This means a 1-simplex is constructed between two points in $\mathbb{R}^n$ whenever the Euclidean distance between the points is smaller than ε, a 2-simplex whenever the Euclidean distance between any two points of a set of three points is smaller than ε, and so on.

The Betti numbers (section 1.3.2) of those complexes are calculated to determine the complexity of the expression profile.

6.2.1 Application to the comparison of gene expression profiles from breast tissue through the menstrual cycle

In order to assess molecular changes through the menstrual cycle, gene expression using microarrays of mRNA extracted from luminal epithelial sorted mammoplasty cells was measured. Each sample was classified into luteal phase (5 samples) or follicular phase (2 samples) if the woman answered with "NO" the question of usage of hormonal contraceptives (Table S1).

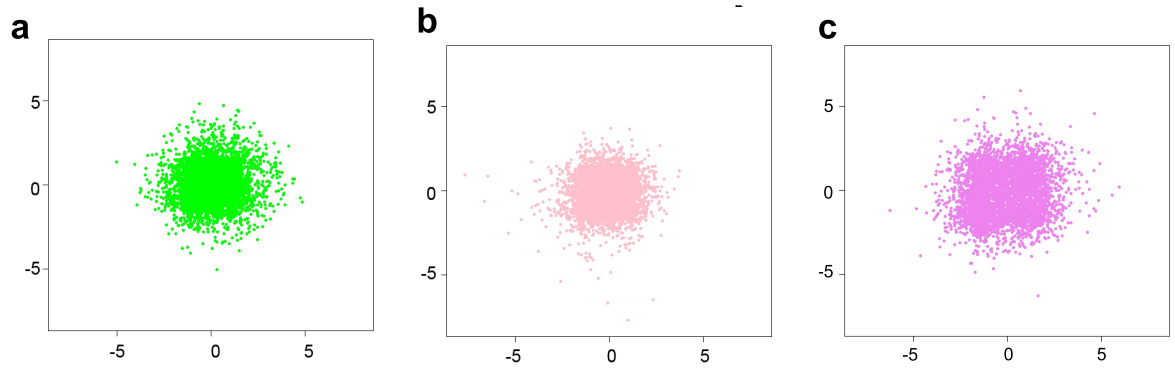
Number	Progesterone Level ng/mL	Age	Follicular or Luteal	Hormonal Contraception
90	0	41	-	Desarex 15 years
80	0.4	31	-	Marvelon
69	0.4	35	F	-
29	0.5	36	-	Minulet, during 10 years
16	0.6	28	-	Various, 12 years
64	0.7	38	?	NA
37	1.1	34	-	Yasmine, 6 months
5	1.1	45	F	-
63	1.6	41	-	Yes
27	14.8	32	-	Yes, during 10 years
46	17.4	29	L	-
22	18.6	39	L	-
51	19.4	35	L	-
83	23.2	43	-	Yes, 10 years
39	26.3	44	L	-
70	32.1	35	?	NA
11	37.2	33	L	-

Table S1: Data set of mammoplasty samples whose RNA was extracted from luminal sorted cells and assessed by microarray with available information on the patients: mammoplasty number, progesterone level at the time of surgery, age, classification in follicular or luteal phase, hormonal contraceptives information.

We generated point clouds using sliding windows of length 2 of the obtained DSGA profiles comparing luteal (Supplementary Fig. S5a, b) and profiles from women under hormonal contraceptives (Supplementary Fig. S5c) to the two follicular-phased expression profiles, representing the healthy state model.

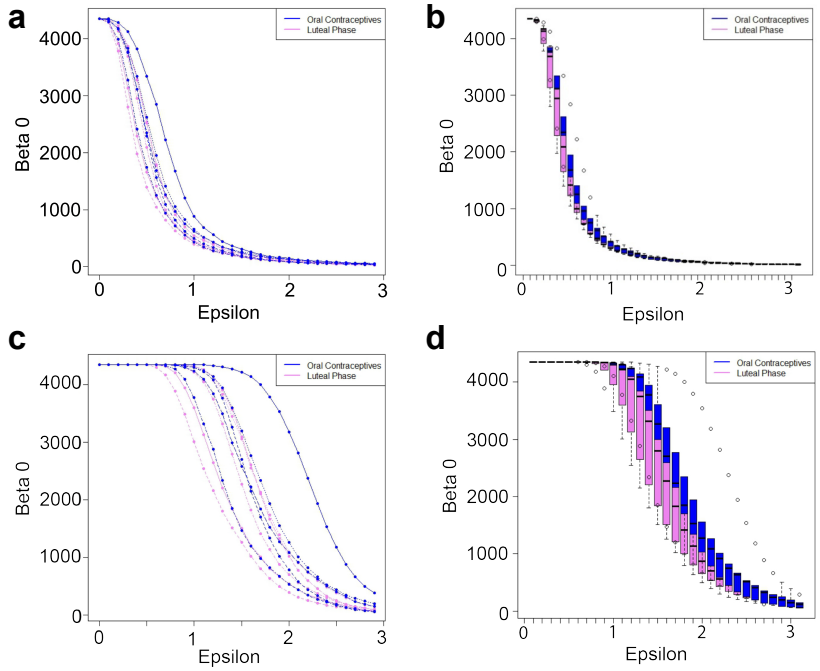
We observed differences between luteal phase and oral contraceptives by visual inspection of the plots (Supplementary Fig. S5a, b, c). We then calculated the barcodes of the Vietoris-Rips complexes to assess if the difference is statistically significant.

Every women taking oral contraceptive displays a different pattern of β_0 in their profile, which corresponds to homological features of H_0. While some have a complex gene profile with many holes in many dimensions; others are close to some profile of women in the luteal phase. Therefore comparing the group of hormonal contraceptives to the group of luteal phase women did not reach significance when comparing the Betti numbers at a given ε. This did not change by considering a sliding window either of length $n = 3$ (Supplementary Fig. S6a, b) or of length $n = 10$ (Supplementary Fig. S6c, d). The method was applied as well on only



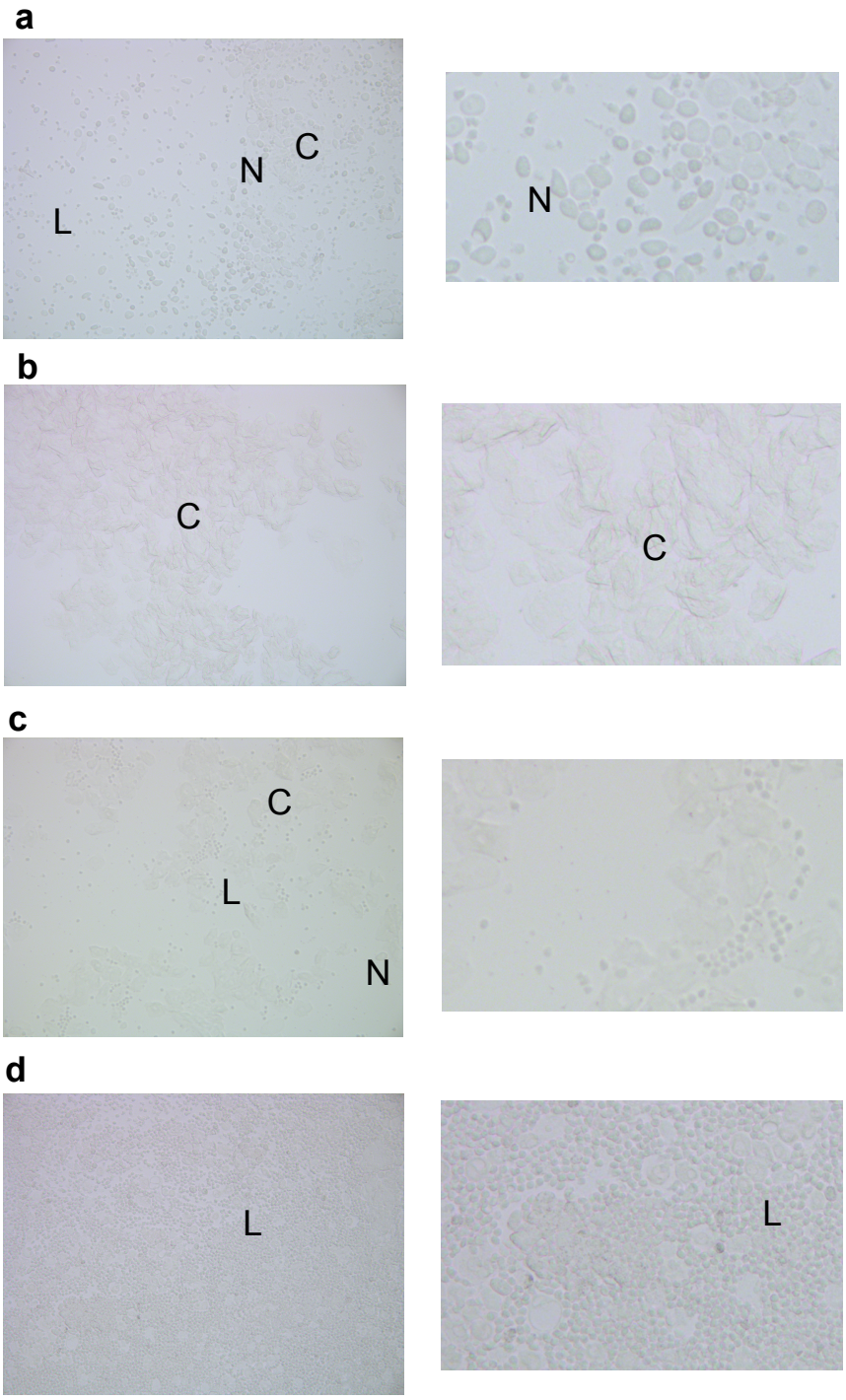
Supplementary Figure S5: Plot of the sliding window of length 2 of the vector obtained after DSGA method, with women in the follicular representing the healthy state model of **(a)** mammoplasty 39, beginning of luteal phase, **(b)** mammoplasty 46, end of luteal phase and **(c)** mammoplasty 80, taking oral contraceptives.

one chromosome, but the results did not improve (data not shown). This method should be repeated with the deviation components from TTMap tin order to determine if the results are improved.

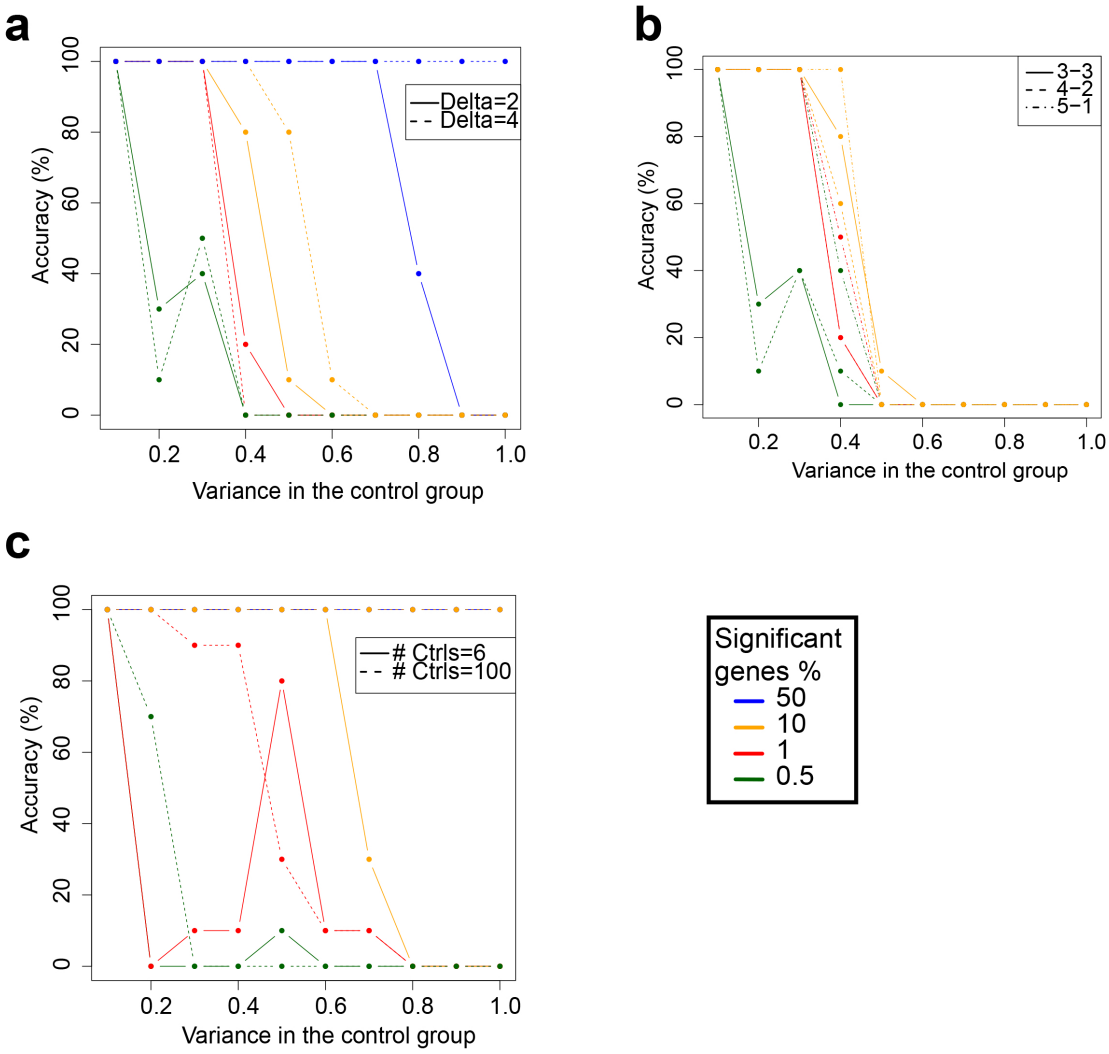


Supplementary Figure S6: **(a)** The 0-th dimensional Betti number, β_0, is computed for a sliding window of length $n = 3$, women using contraceptives in blue and luteal phase women in pink. **(b)** Boxplot of the values in **(a)** grouped by women using contraceptives (blue) or women in the luteal phase (pink). **(c)** The 0-th dimensional Betti number, β_0, is computed for a sliding window of length $n = 10$, women using contraceptives in blue and luteal phase women in pink. **(d)** Boxplot of the values in **(c)** grouped by women using contraceptives (blue) or women in the luteal phase (pink).

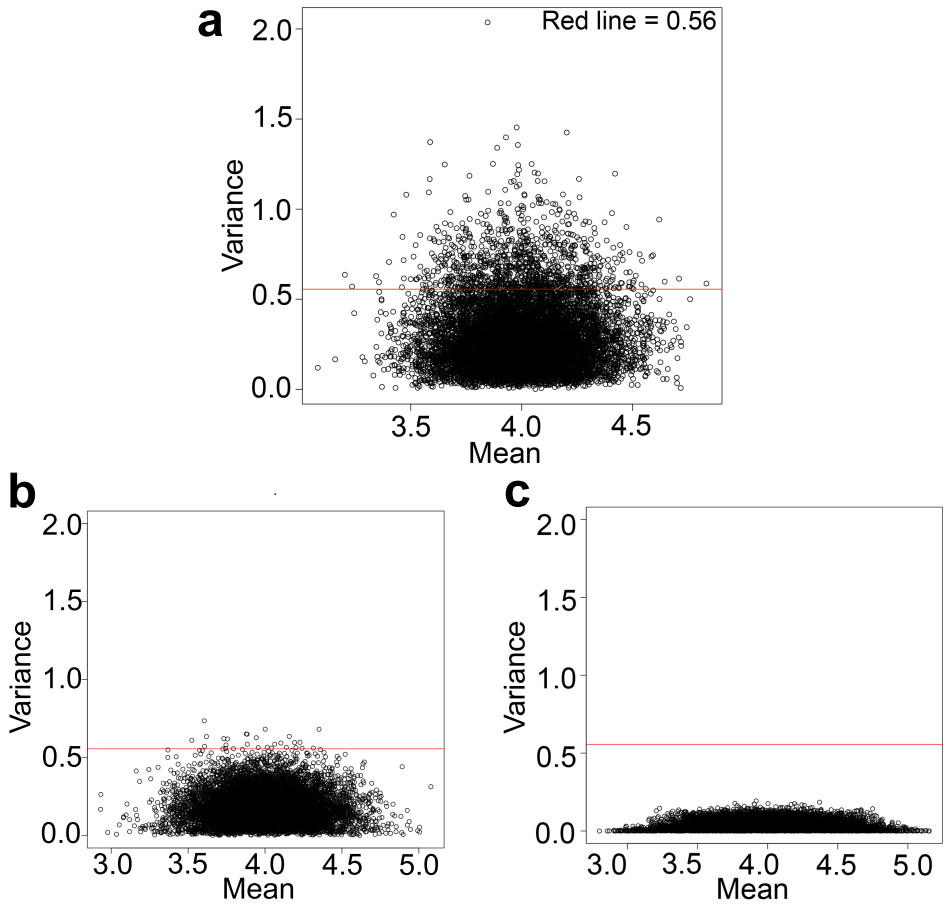
6.3 Supplementary figures



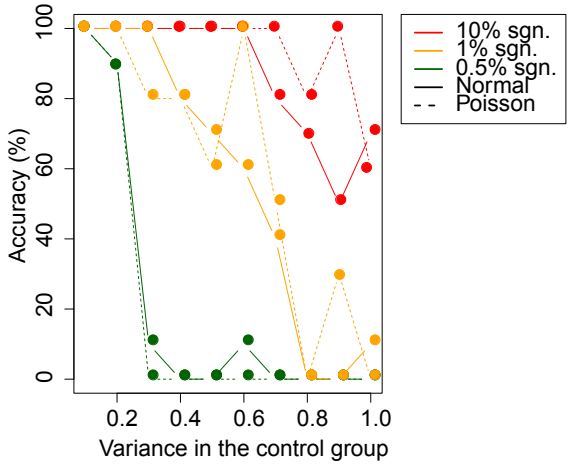
Supplementary Figure S7: Vaginal cytology through the estrous cycle at 20X and with a zoom to visualise more clearly the cell types. **(a)** In proestrous, with predominantly nucleated epithelial cells (N). **(b)** In estrous with cornified cells (C). **(c)** In metestrous, all three cell types are apparent: N, C and leukocytes (L). **(d)** In diestrous, the predominant cells are L and they fill the picture.



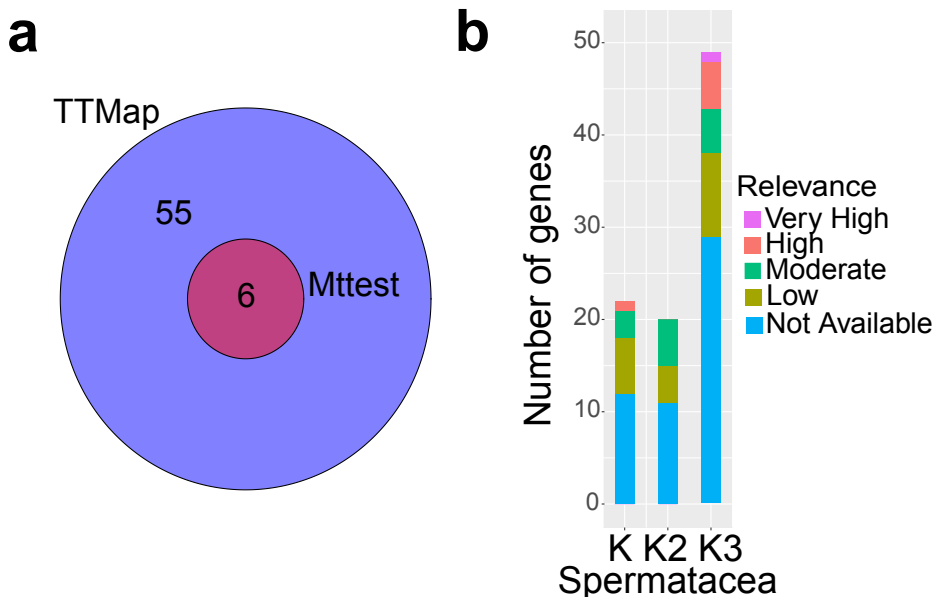
Supplementary Figure S8: *In silico* valdiation. **(a)** Accuracy plot when increasing delta. **(b)** Accuracy plot when the subgroups have different sizes. **(c)** Accuracy plot when increasing the number of samples in the control group.



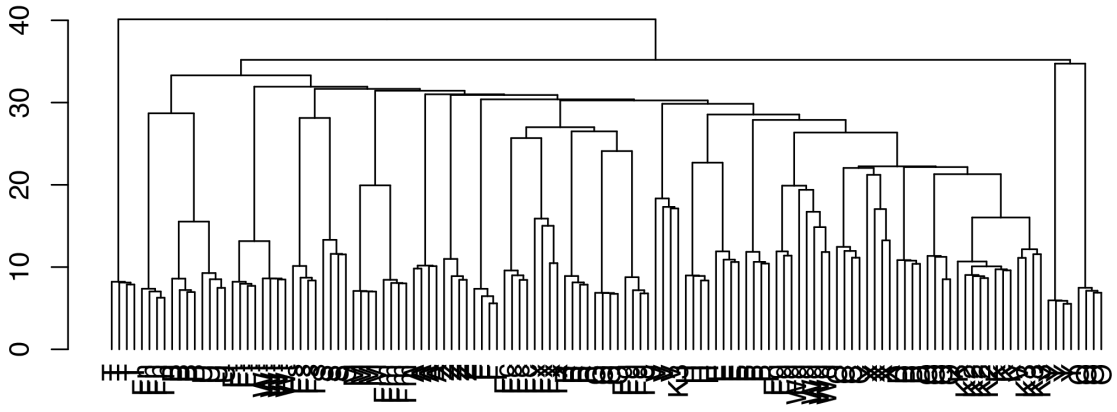
Supplementary Figure S9: Effect of changing the parameter e **(a)** Dot plot showing the effect of different ways to select an outlier on HDA Original data **(b)** After TTMap's outlier correction **(c)** using a confidence interval correction. Red line = 0.56 represents the 90-th percentile of the variance.



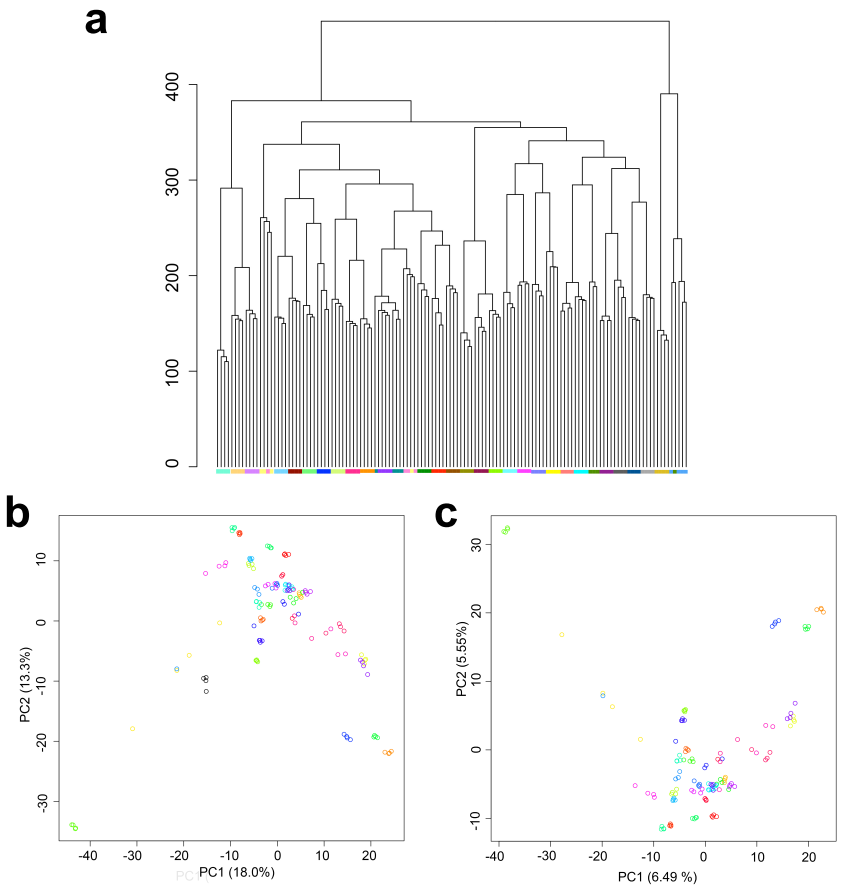
Supplementary Figure S10: Effect of changing the parameter e **(a)** Dot plot showing the effect of different ways to select an outlier on HDA Original data **(b)** After TTMap's outlier correction **(c)** using a confidence interval correction. Red line = 0.56 represents the 90-th percentile of the variance.



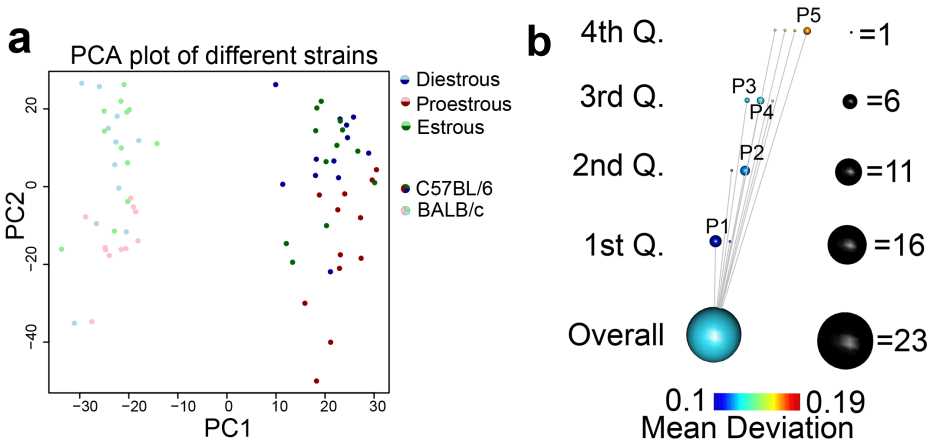
Supplementary Figure S11: Significant genes of the Spermatacea (K). **(a)** Venn diagramm of significant genes of K with TTMap (purple area) and with moderated-t-test (Mttest, red area). **(b)** Barplot showing the relevance of the genes missed by Mttest on K, K2 and K3.



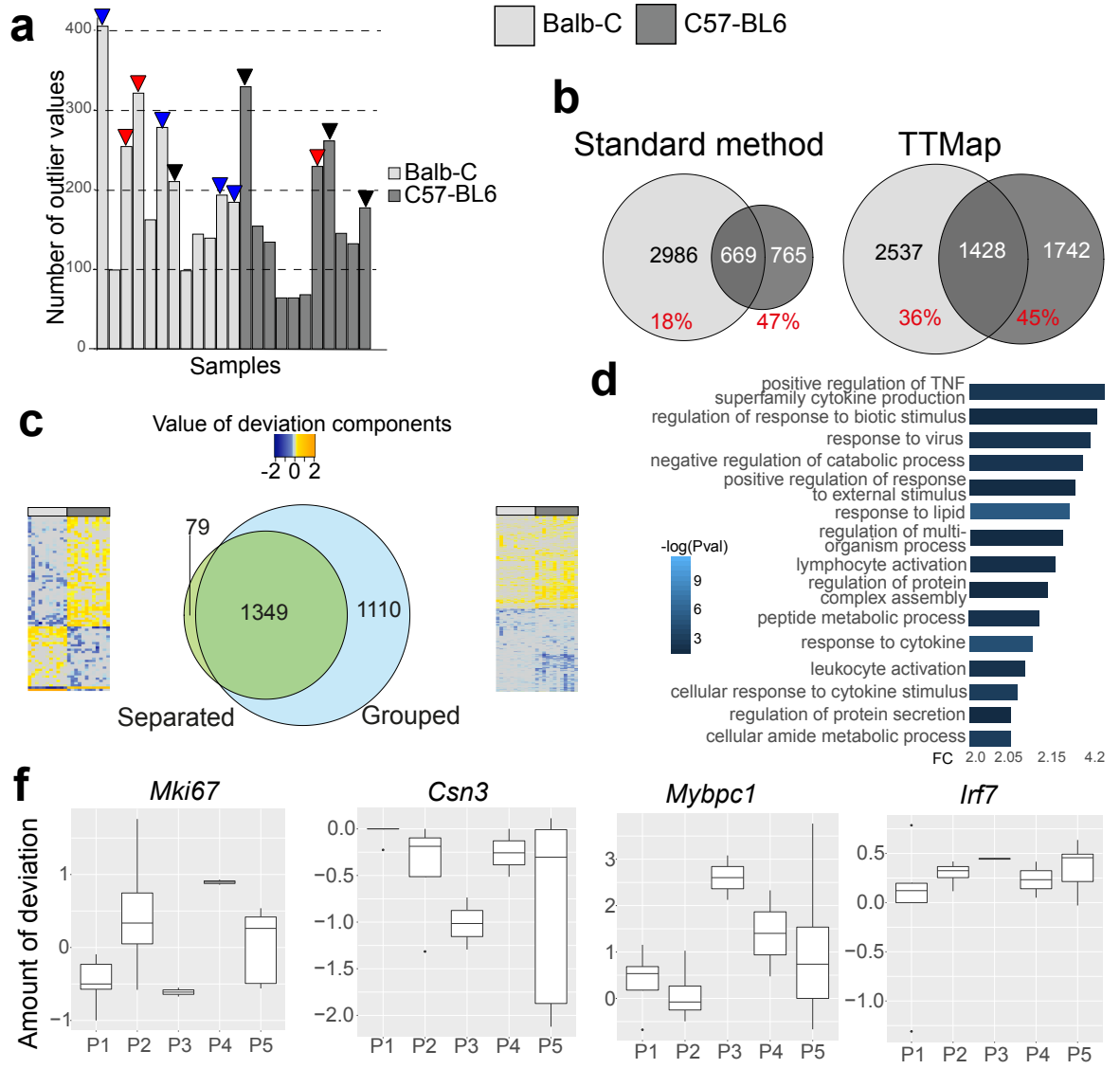
Supplementary Figure S12: Hierarchical clustering using euclidean distance of the log-transformed data from the flyatlas.



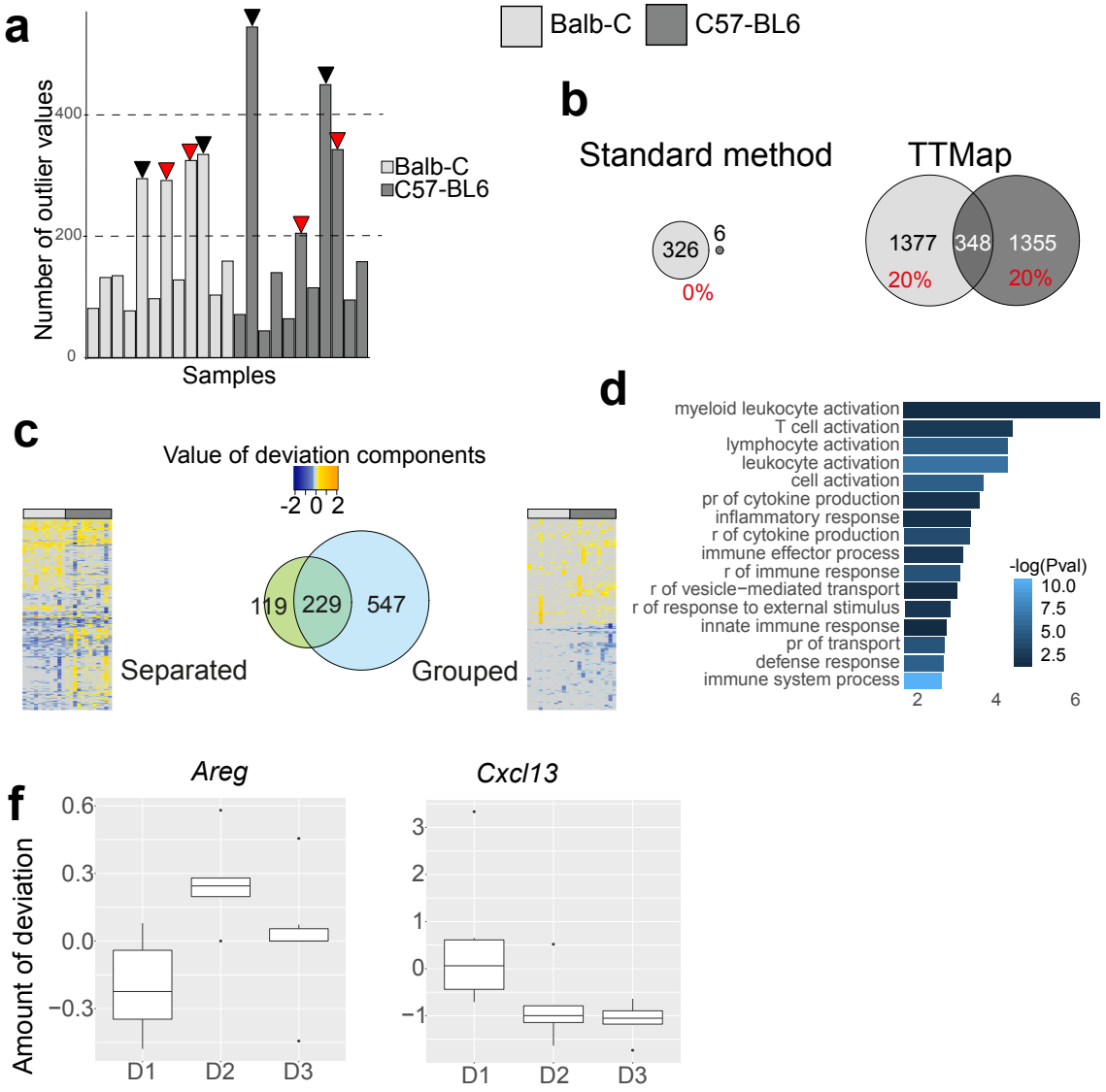
Supplementary Figure S13: Hierarchical clustering using euclidean distance and PCA on the fly atlas, a color represents one organ. **(a)** Hierarchical clustering on the mismatch distance **(b)** PCA on the log transformed data **(c)** PCA on the deviation components, after the HDA step.



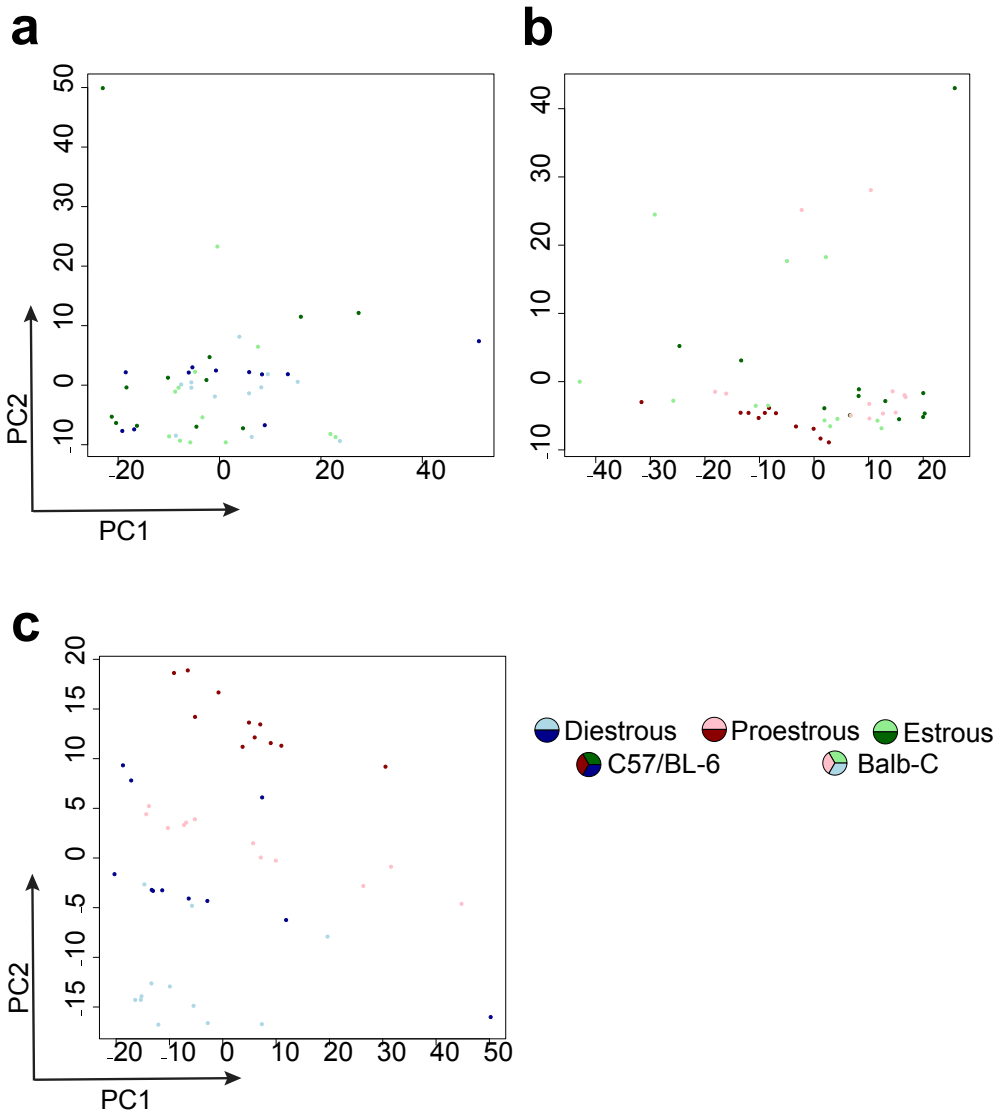
Supplementary Figure S14: Clusters obtained with PCA and with TTMap of the data in section 3.4. **(a)** PCA plot of RNA-seq profiles of mammary glands from BALB/c (light) and C57/BL6 mice (dark), in different phases of the estrous cycle. **(b)** Output of TTMap with global and local clusters showing the different subgroups of P (P1, P2, P3, P4, P5) and outliers (single points) color-coded and ordered by average amount of filter function which is the proximity to control (estrous).



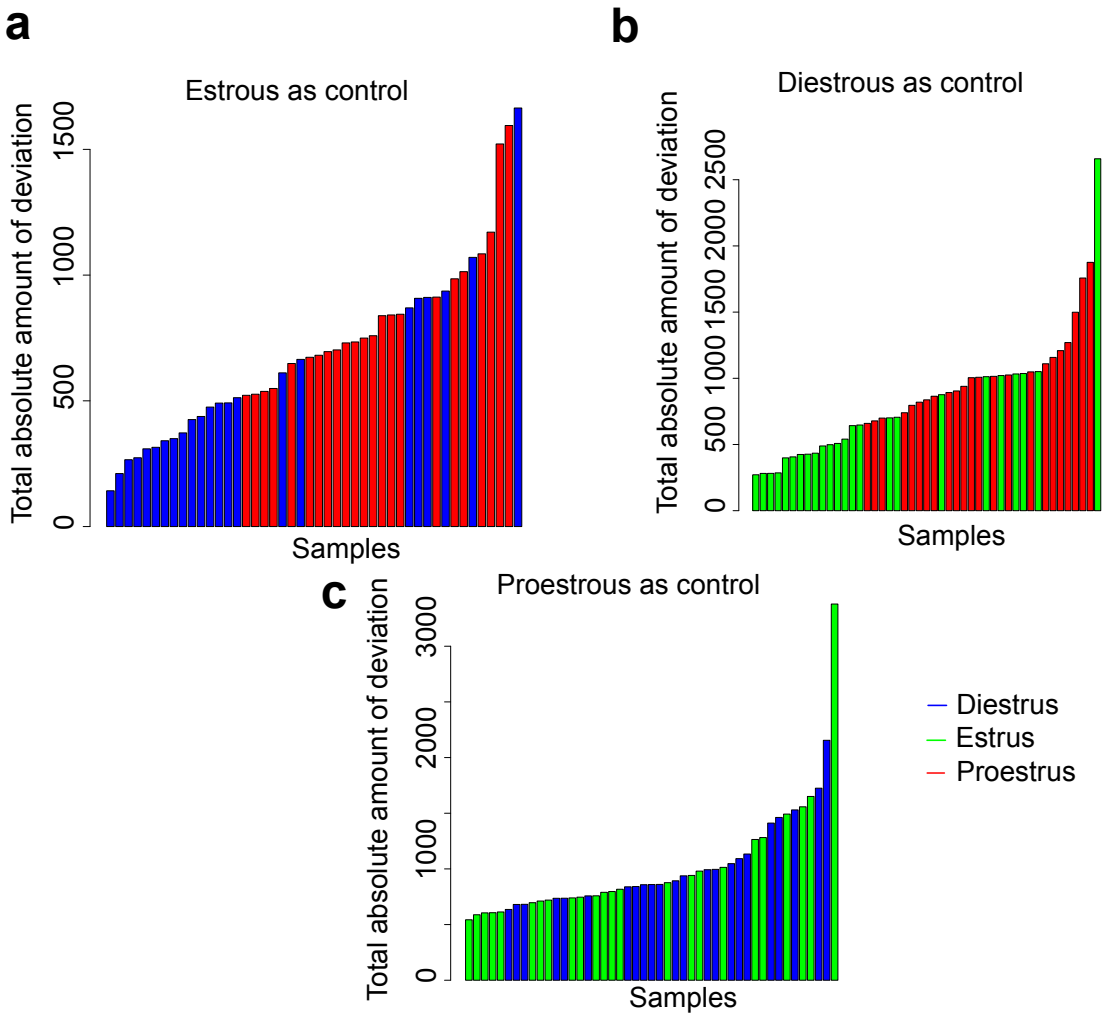
Supplementary Figure S15: Estrous cycle related gene expression changes in the mammary glands of C57BL/6 and BALB/c mice; diestrous (D) vs proestrous (P) phase. **(a)** Barplot representing the number of outlier values in the control group (proestrous phase). Samples with a high number of outlier values and remain isolated during clustering when P is the test group are identified as outliers (black arrowhead) or as highly variable samples that are forming one group (red and blue arrowheads). **(b)** Venn diagrams of the genes differentially expressed between D vs P identified with standard tools and TTMap on BALB/c compared to C57BL/6 analyzed separately. In red, the fraction of common significant genes per strain (% over total number of significant genes). **(c)** Venn diagrams of the common differentially expressed genes when the analysis is done separately on the two mouse strains (Separated) or with the two mouse strains combined into one analysis (Grouped) using TTMap comparing D vs P. Adjacent heatmaps of the deviation components illustrate why the genes were missed; while on the separated analysis deviations are going into opposite direction, in the grouped analysis the genes deviate in the same direction, but to different extent.. **(d)** Panther pathway analysis [84] of significant genes identified by TTMap in the comparison D vs P shown by Fold Change (FC) enrichment of the pathway with *-log(Pval)* as a color code. Fifteen most increased pathways are shown *p = positive, n= negative, r=regulation.* **(e)** Boxplots representing the deviation component values in the identified subgroups of P (P1, P2, P3, P4, P5) by TTMap ordered by amount of deviation compared to the diestrous samples (controls) of the genes *Mki67*, *Csn3*, *Mybpc1* and *Irf7*.



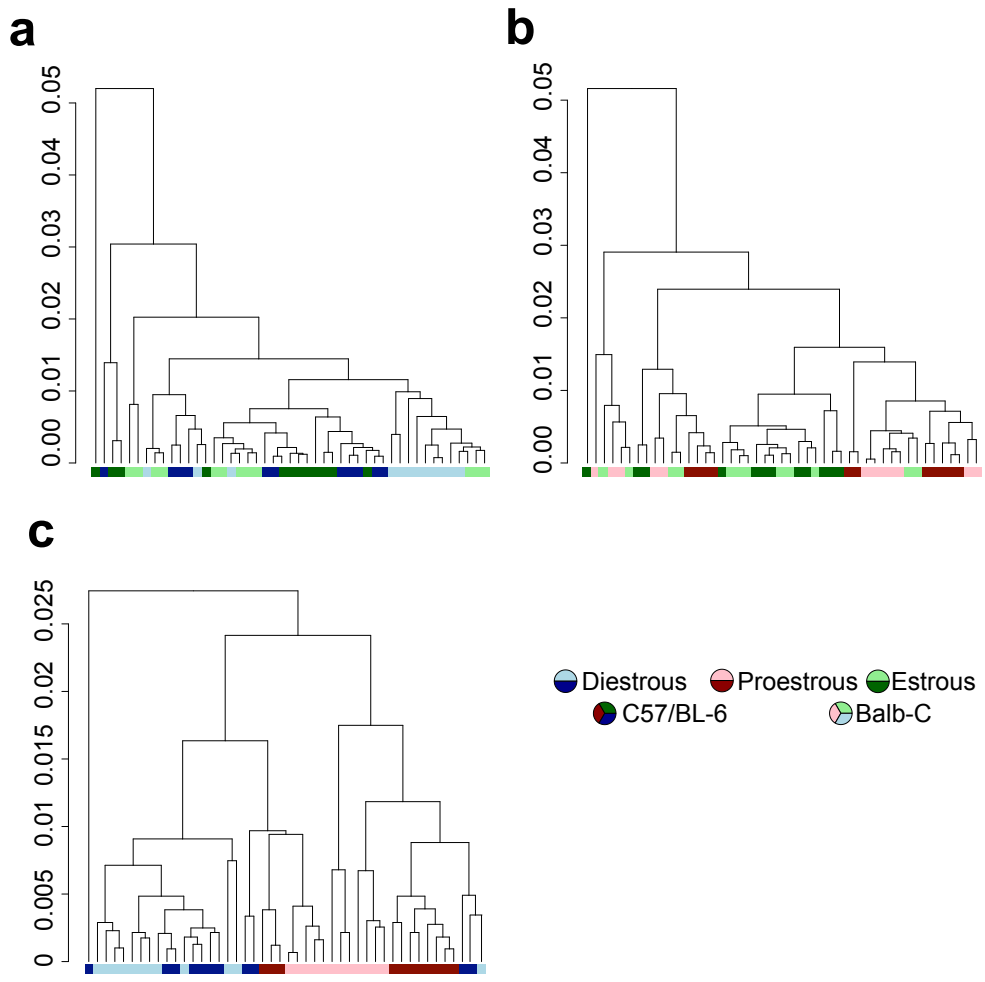
Supplementary Figure S16: Estrous cycle related gene expression changes in the mammary glands of C57BL/6 and BALB/c mice; estrous (E) vs diestrous (D) phase. **(a)** Barplot representing the number of outlier values in the control group (diestrous phase). Samples with high number of outlier values and remain isolated during clustering when D is the test group are identified as outliers (black arrowhead) or as highly variable samples that are forming one group (red arrowheads). **(b)** Venn diagrams of the genes differentially expressed between E vs D using standard analysis tools and TTMap on BALB/c compared to C57BL/6 analyzed separately. In red, the fraction of common significant genes per strain (% over total number of significant genes). **(c)** Venn diagrams of the common differentially expressed genes when the analysis is done separately on the two mouse strains (Separated) or with the two mouse strains combined into one analysis (Grouped) using TTMap comapring E vs D. Adjacent heatmaps of the deviation components illustrate why the genes were missed; while on the separated analysis deviations are going into opposite direction, in the grouped analysis the genes deviate in the same direction, but to different extent. **(d)** Panther pathway analysis [84] of significant genes identified by TTMap in the comparison E vs D shown by Fold Change (FC) enrichment of the pathway with *-log(Pval)* as a color code. Fifteen most increased pathways are shown *p = positive, n= negative, r=regulation.* **(e)** Boxplots representing the deviation component values in the identified subgroups of D (D1, D2, D3) by TTMap ordered by amount of deviation compared to the controls samples in estrous, of the genes *Areg*, *Cxcl13*.



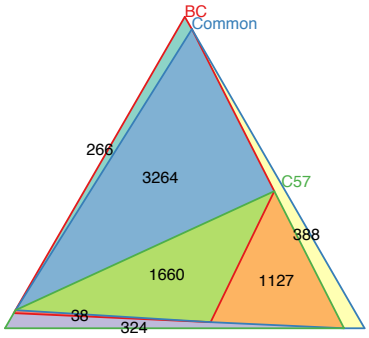
Supplementary Figure S17: PCA plot after the hyperrectangle deviation assessment step of TTMap for the multiple comparison **(a)** P vs E and D **(b)** D vs E and P and **(c)** E vs P and D



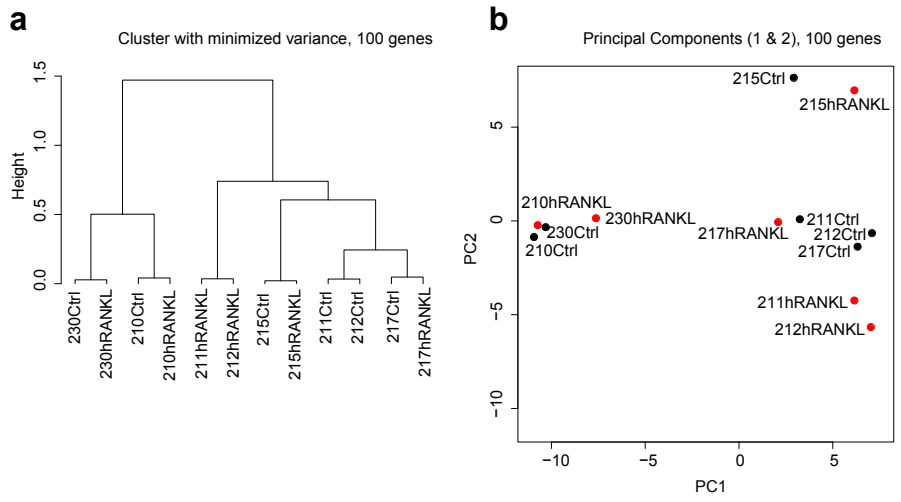
Supplementary Figure S18: The output of the filter function as a barplot for the comparison **(a)** E vs D and P **(b)** D vs E and P and **(c)** P vs E and D.



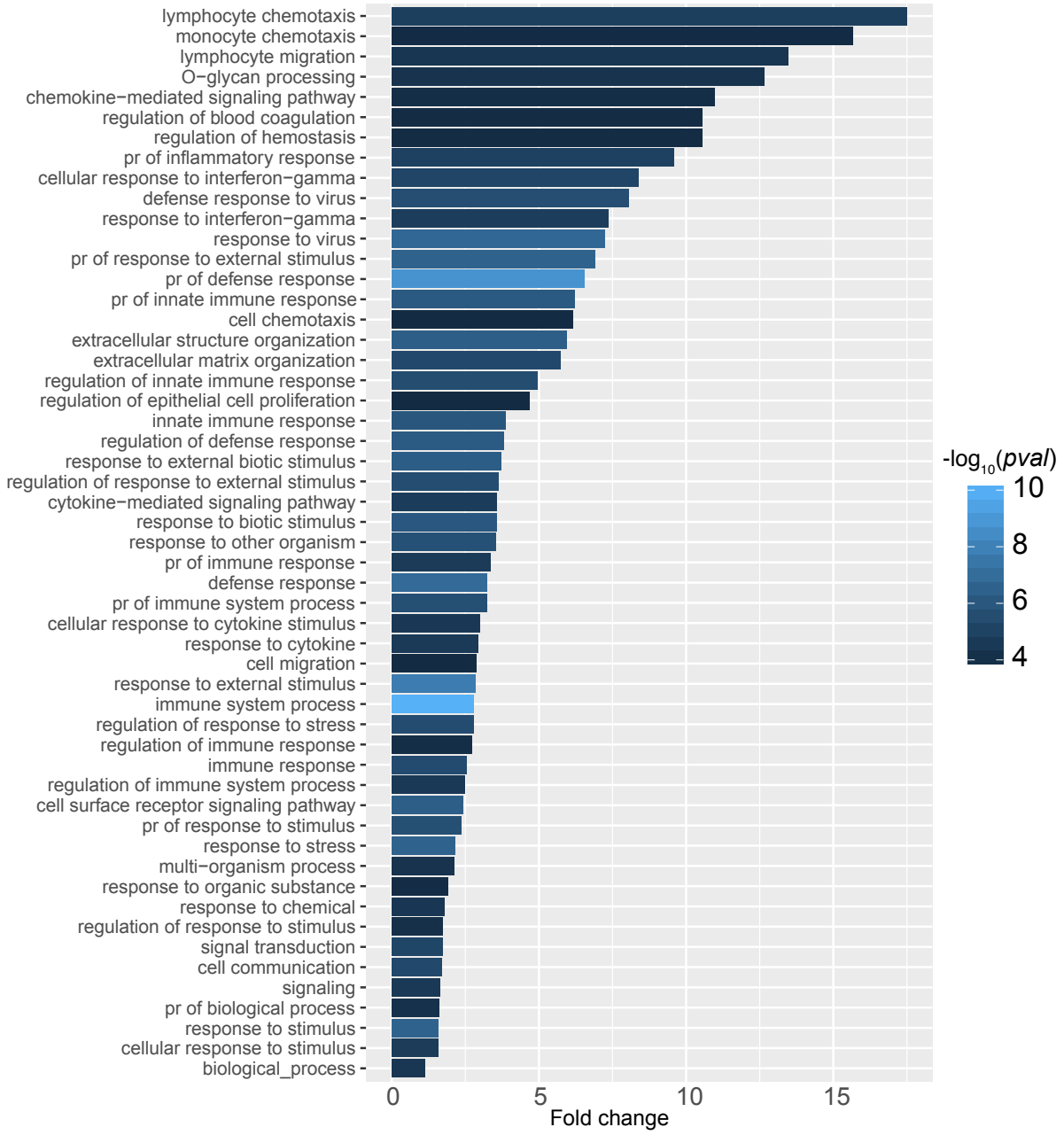
Supplementary Figure S19: Hierarchical clustering on Z-score normalised values per strains of samples from **(a)** E and D when P is the control **(b)** P and E when D is the control and **(c)** P and D when E is the control.



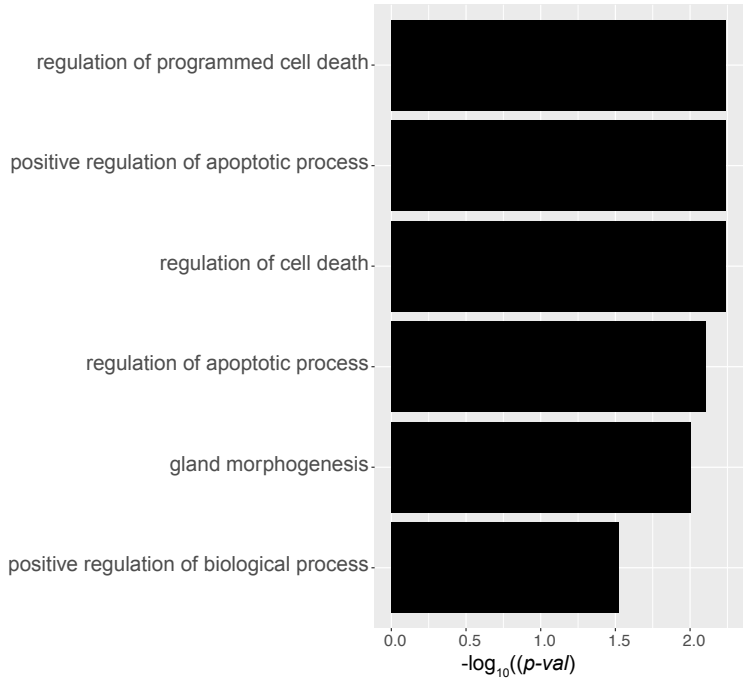
Supplementary Figure S20: Triangular Venn diagram showing the overlap of the multiple comparison of Estrous vs Diestrous and Proestrous analysed only on BALB/c samples (BC), only on C57BL/6 samples (C57) or all together.



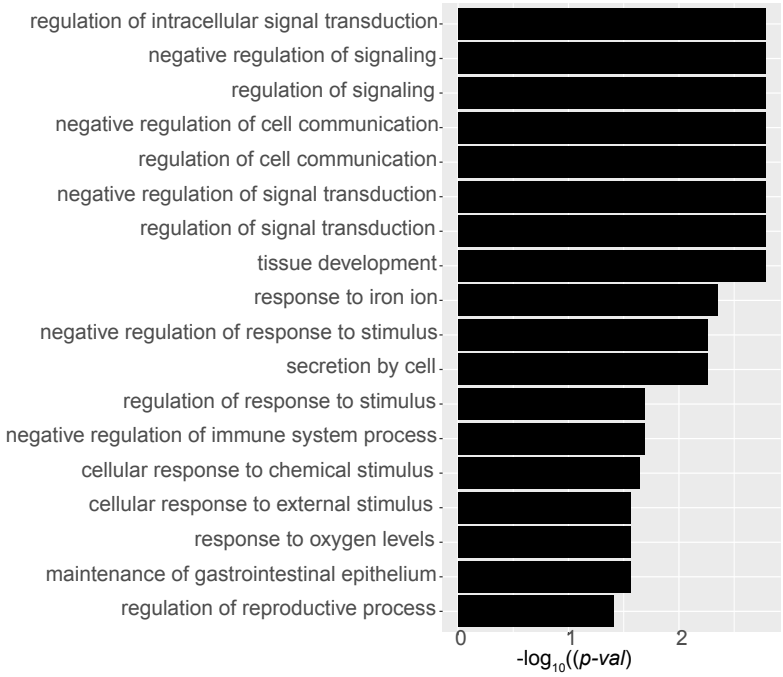
Supplementary Figure S21: Standard clustering tools on gene expression profile of six samples (210, 211, 212, 215, 217, 230) treated with human RANKL (hRANKL, red) or with vehicle (Ctrl, black). **(a)** Hierarchical clustering on the 100 most variable genes **(b)** PCA plot on the 100 most variable genes



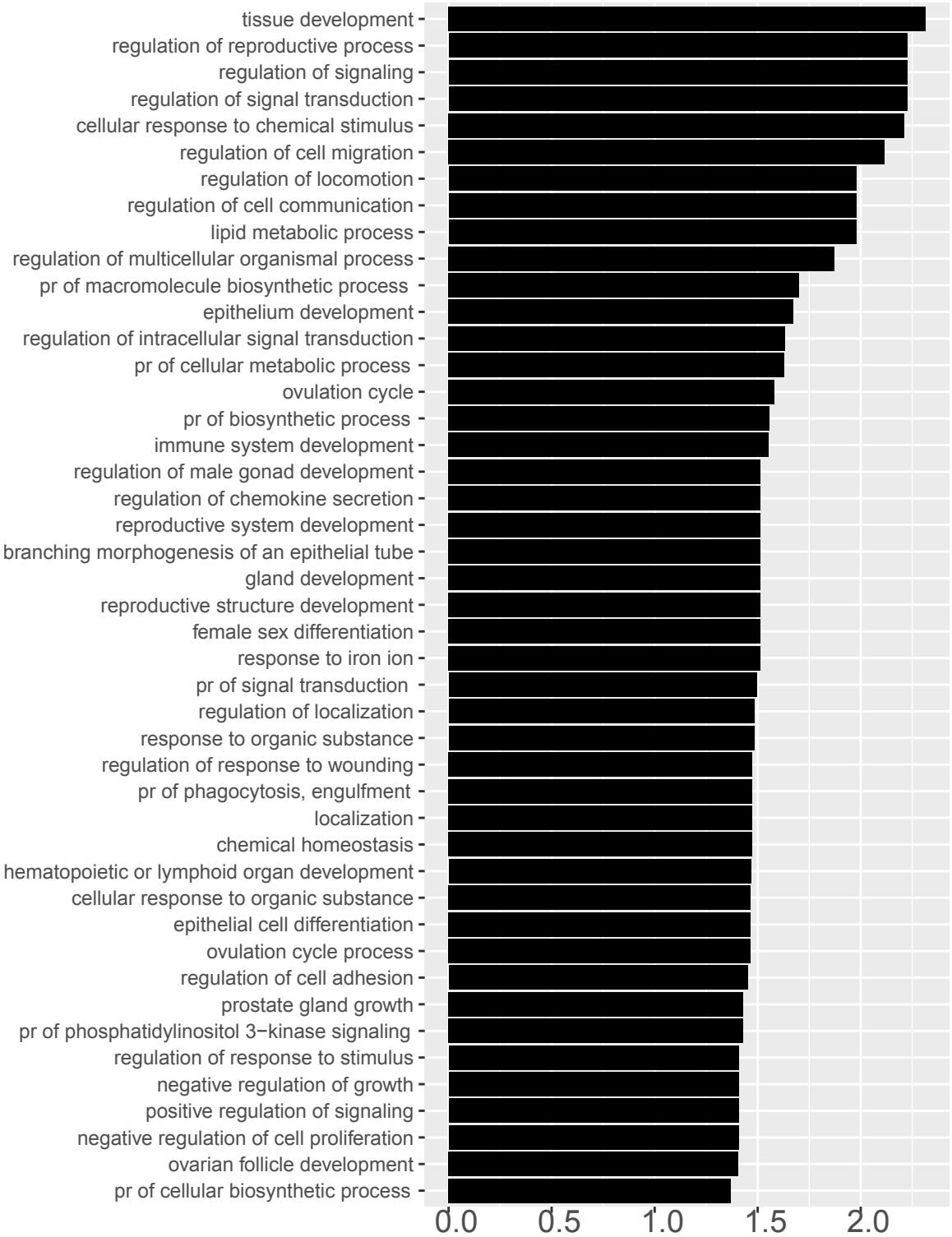
Supplementary Figure S22: Panther [84] analysis of significant genes upon treatment with RANKL in the context of mammoplasties having been exposed to high level of progesterone.



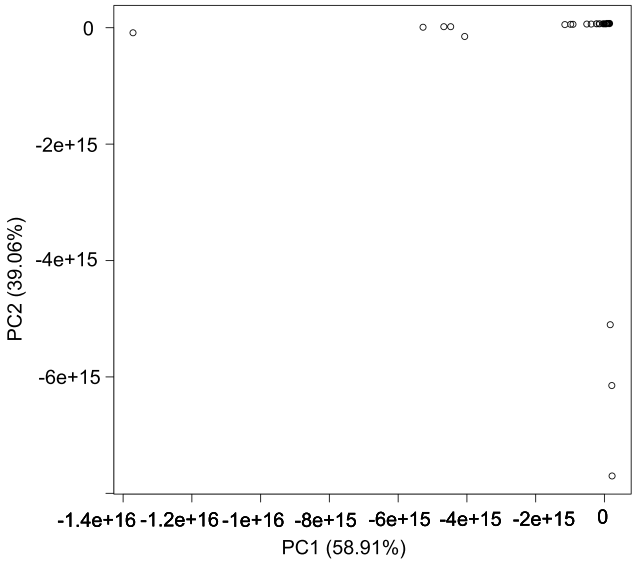
Supplementary Figure S23: Panther [84] analysis of the common significant genes varying in the same direction upon treatment with RANKL and upon treatment with progesterone.



Supplementary Figure S24: Panther [84] analysis of the common significant genes upon treatment with RANKL and upon treatment with progesterone but increased upon RANKL and decreased upon progesterone.



Supplementary Figure S25: Panther [84] analysis of the common significant genes varying in opposite direction upon treatment with RANKL and upon treatment with progesterone, *pr= positive regulation.*



Supplementary Figure S26: PCA of brain spike measurements of different cell types.

A Appendix

A.1 Appendix A.1

I contributed to the work: Cathrin Brisken, Kathryn Hess and Rachel Jeitziner, Progesterone and Overlooked Endocrine Pathways in Breast Cancer Pathogenesis, *Endocrinology*, 156(10):3442-3450, 2015 [106].
My work consisted in illustrating the different facts in a concise manner and write their legend. First, an illustration describing the evolution of the mammary gland in the mouse. Second, a scheme of the fluctuation of the hormones through the menstrual cycle. Then, a scheme of the signaling downstream of progesterone. Finally, a description of the risk factor and how they evolve through time, focussing on the role of the menopause and the different endocrine discruptors.
This work has been adapted and presented in Chapter 1 section 1.6.1.

A.2 Appendix A.2

I participated in the work: George Sflomos, Valerian Dormoy, Tauno Metsalu, Rachel Jeitziner, Laura Battista, Valentina Scabia, Wassim Raffoul, Jean-Francois Delaloye, Assya Treboux, Maryse Fiche, Jaak Vilo, Ayyakkannu Ayyanan, and Cathrin Brisken, A Preclinical Model for ERα-Positive Breast Cancer Points to the Epithelial Microenvironment as Determinant of Luminal Phenotype and Hormone Response, *Cancer Cell*, 29(3):407-422, 2016 [228].
I wrote several bioinformatic scripts for the analysis of data in this article and gave my mathematical support for numerous analysis. I therefore analysed RNA-seq data, comparing mouse intraductally injected MCF7 cells to fat pad cells with or without treatment with Fulvestrant as well as patient derived xenografts with or without treatment. Kegg, Gene ontology and metacore analysis were needed to confirm if the signatures obtained up on treatment are relevant. I drew a graph of the radiance of the metastases found in each organ *post mortem*. Then, I analysed RNA-seq data comparing cells treated or not with the compound Fulvestrant and drew a heatmap. Moreover, I drew boxplots reflecting the difference between the IHC staining of a breast biopsy and the biopsy of the cells injected into mice.

A.3 Appendix A.3

During the fourth year I was the main author of the work entitled "Two-Tier Mapper: a user-independent clustering method for global gene expression analysis based on topology" submitted as an original article to *Bioinformatics* in 2018, from the following authors : Rachel

Jeitziner, Mathieu Carrière, Jacques Rougemont, Steve Oudot, Kathryn Hess, and Cathrin Brisken and available on arXiv, *arXiv: 1801.01841* [194].
This work aimed at presenting the method that was developed during this thesis. It had been characterized on two different datasets.
This work has been adapted and presented in Chapter 2 and Chapter 3 as well as in Chapter 1 section 1.7.

A.4 Appendix A.4

I was the only author of the reference manual inside the R package "Two-Tier Mapper" [246]. This paper explains all the functions developed in **R** for the packages "TTMap" with working examples for each function.

A.5 Appendix A.5

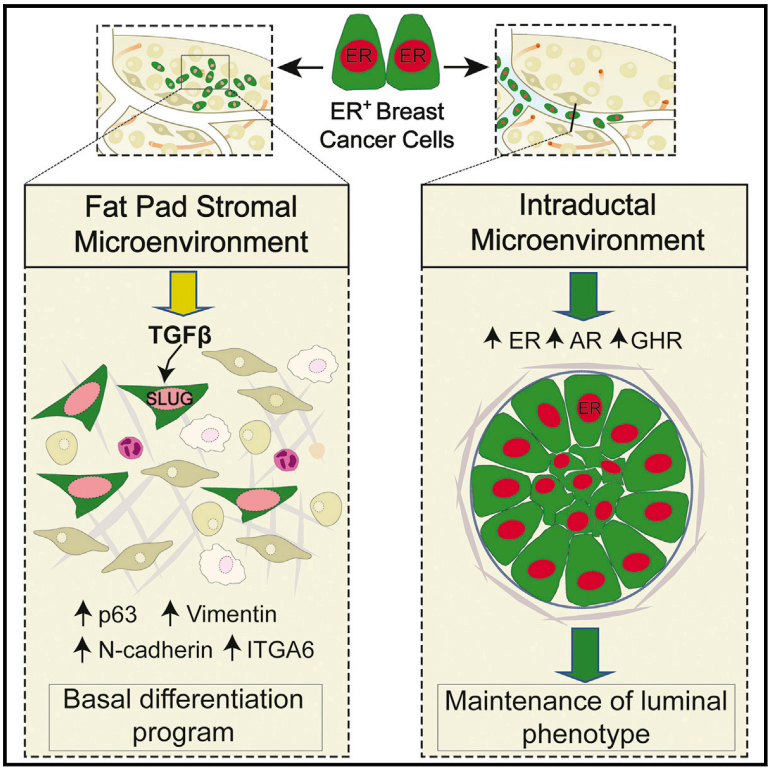
I was the only author of the user guide inside the R package "Two-Tier Mapper: a user-independent clustering method for global gene expression analysis based on topology" [247]. This paper explains through an example in **R** how to use the package "TTMap".

Article

Cancer Cell

A Preclinical Model for ERα-Positive Breast Cancer Points to the Epithelial Microenvironment as Determinant of Luminal Phenotype and Hormone Response

Graphical Abstract



Authors

George Sflomos, Valerian Dormoy, Tauno Metsalu, ..., Jaak Vilo, Ayyakkannu Ayyanan, Cathrin Brisken

Correspondence

cathrin.brisken@epfl.ch

In Brief

Sflomos et al. show that engrafting human estrogen receptor α-positive breast tumors into mouse milk ducts, in contrast to mammary fat pads, efficiently generates retransplantable xenografts that mimic the original tumors. They identify differential induction of SLUG by these microenvironments as a key factor.

Highlights

- Tissue microenvironment is critical for the growth of ER+ breast cancer cells
- Mammary stroma induces TGFβ/SLUG signaling and basal differentiation in MCF7 cells
- Mouse milk ducts enable physiological growth of ER+ breast cancer cells
- Mouse intraductal ER+ PDXs are robust, retransplantable, and predictive

Accession Numbers

GSE68694
GSE74608

Sflomos et al., 2016, Cancer Cell *29*, 407–422
March 14, 2016
http://dx.doi.org/10.1016/j.ccell.2016.02.002

CellPress

Cancer Cell
Article

A Preclinical Model for ERα-Positive Breast Cancer Points to the Epithelial Microenvironment as Determinant of Luminal Phenotype and Hormone Response

George Sflomos,[1] Valerian Dormoy,[1] Tauno Metsalu,[2] Rachel Jeitziner,[1] Laura Battista,[1] Valentina Scabia,[1] Wassim Raffoul,[3] Jean-Francois Delaloye,[3] Assya Treboux,[3] Maryse Fiche,[3] Jaak Vilo,[2] Ayyakkannu Ayyanan,[1] and Cathrin Brisken[1,*]

[1]ISREC – Swiss Institute for Experimental Cancer Research, School of Life Sciences, Ecole polytechnique fédérale de Lausanne (EPFL), SV2.832 Station 19, 1015 Lausanne, Switzerland
[2]Institute of Computer Science, University of Tartu, Liivi 2, Tartu 50409, Estonia
[3]Lausanne University Hospital, 1011 Lausanne, Switzerland
*Correspondence: cathrin.brisken@epfl.ch
http://dx.doi.org/10.1016/j.ccell.2016.02.002

SUMMARY

Seventy-five percent of breast cancers are estrogen receptor α positive (ER$^+$). Research on these tumors is hampered by lack of adequate in vivo models; cell line xenografts require non-physiological hormone supplements, and patient-derived xenografts (PDXs) are hard to establish. We show that the traditional grafting of ER$^+$ tumor cells into mammary fat pads induces TGFβ/SLUG signaling and basal differentiation when they require low SLUG levels to grow in vivo. Grafting into the milk ducts suppresses SLUG; ER$^+$ tumor cells develop, like their clinical counterparts, in the presence of physiological hormone levels. Intraductal ER$^+$ PDXs are retransplantable, predictive, and appear genomically stable. The model provides opportunities for translational research and the study of physiologically relevant hormone action in breast carcinogenesis.

INTRODUCTION

About 90% of potential oncology drugs fail in clinical trials (Arrowsmith, 2011; Hait, 2010), in part because the preclinical models used to test them do not adequately reflect their clinical counterparts. Breast cancer is the leading cause of cancer-related death among women worldwide. While there are some preclinical models, there is a paucity of in vivo models for the estrogen receptor α-positive (ER$^+$) subtypes, which represent more than 75% of all cases (Hidalgo et al., 2014). The lack of a clinically relevant model hampers progress in understanding how hormones, increasingly recognized as important factors in breast carcinogenesis, impinge on disease progression and therapy.

Many cell lines reflecting different breast cancer subtypes have been established. In those that can grow as xenografts, a million or more cells must be injected either subcutaneously or into the mammary fat pad of immune-compromised mice; the resulting tumors grow much faster than their human counterparts (Zhang et al., 2013). Cell lines derived from the most frequent specific histological subtypes, the ER$^+$ lobular carcinomas, do not grow in vivo at all (Guiu et al., 2014; Sikora et al., 2014). The few ER$^+$ cell lines that grow as xenografts depend on exogenous 17β-estradiol (E2) (Vargo-Gogola and Rosen, 2007). This results in serum E2 levels equivalent to mid-menstrual cycle levels in premenopausal women (100–400 pg/ml) (Kratz et al., 2004) whereas most ER$^+$ breast cancers occur in postmenopausal women with E2 levels <18 pg/ml. The hormonal treatment has detrimental effects on the E2-sensitive urogenital tracts of female mice, which some investigators have bypassed using male mice, circumstances that may further reduce the clinical relevance (Clinchy et al., 2000). Finally, the injection of tumor

Significance

A high percentage of potential oncology drugs fail in clinical trials, partly because preclinical models used to test them are inadequate. Breast cancer is the leading cause of cancer-related death among women worldwide, but we lack appropriate in vivo models for the ER$^+$ subtypes, which represent more than 75% of all cases. We address these issues by xenografting tumor cells to their site of origin, the milk ducts. All ER$^+$ cell lines and patient-derived xenografts grow mimicking their clinical counterparts. Disease progresses with invasion and metastasis, which become amenable to study. The action of hormones, important in breast carcinogenesis, can now be studied in a relevant context. Importantly, this model opens opportunities for development and evaluation of therapies.

cells into adipose tissue relates poorly to the human disease where it may take many years for tumor cells originating from the milk ducts to invade the stroma and select for metastatic cells (DeRose et al., 2011).

Patient-derived xenografts (PDXs) mimic the human disease more accurately (Hidalgo et al., 2014) but they are difficult to establish from ER^+ tumors, with a 2.5% engraftment rate in a series of 423 ER^+ tumors grafted into immune-compromised mice (Cottu et al., 2012). Genetically engineered mouse models (GEMMs) have been developed, in which the entire tumorigenic process including metastasis can be studied (Weinberg, 2011), but few of these produce ER^+ tumors (Zhang et al., 2013). To our knowledge, *Stat1*$^{-/-}$ mice are currently the only GEMM with consistent ERα expression in the majority of tumor cells and with functional E2 dependence, i.e. decreased growth upon ovariectomy (Chan et al., 2012).

Here, we address these concerns and the need for a model to study hormone response in vivo in clinically relevant settings.

RESULTS

Intraductal Growth of Breast Cancer Cell Lines

The mouse intraductal (MIND) model, in which cells are injected into the mouse milk duct system, was initially developed for studying ductal carcinomas in situ (DCIS) (Behbod et al., 2009; Valdez et al., 2011). To test the hypothesis that mouse milk ducts offer a supportive microenvironment for human breast cancer cells in the presence of physiological hormone levels, we obtained breast cancer cell lines of different molecular subtypes (Neve et al., 2006) (Table S1). After infection with DsRed and luciferase2 expressing lentivirus, between 50,000 and 100,000 cells were injected into the thoracic and inguinal mammary glands of adult female SCID/Beige mice through the teat, creating a MIND xenograft (Behbod et al., 2009) (Figure 1A). All cell lines grew without hormone supplements with engraftment rates between 30% and 100% with the exception of MDAMB231 cells, which grew only in 1 out of 26 grafts (Figure 1B). The findings included ER^+ cell lines, such as the most widely studied MCF7 (Lee et al., 2015), HCC1428, ZR751, and MDAMB134VI, which is derived from a lobular carcinoma and does not seem to have been established in vivo previously (Logan et al., 2015), as well as the androgen receptor $(AR)^+$ MDAMB453 (Figures 1B, S1A, and S1B), which usually requires exogenous 5α-dihydrotestosterone (Ni et al., 2011). In vivo monitoring of engrafted mice by luminescence showed that the ER^+ cell lines grow exponentially (Figure 1C). The initial signal detected from intraductally injected MDAMB231 cells dropped to background levels within a week (Figures S1C and S1D).

The basal-like cell lines BT20 and HCC1806 gave rise to palpable tumors within 3 and 8 weeks, respectively. The engrafted ER^+ cell lines merely dilated the milk ducts. MCF7 and T47D cells caused focal distensions (Figures 1D and 1E), and BT474 cells extensively dilated the milk ducts (Figure 1F). We used primate-specific Alu repeats to unequivocally identify human cells (Schmid and Deininger, 1975) (Figures S1E and S1F). BT20 and HCC1806 were highly invasive, whereas the luminal cell lines expanded predominantly within the ducts (Figure S1G). All xenografts preserved histopathological features of their clinical counterparts (Figure 1G); MCF7 cells showed moderate nuclear pleomorphism and tubular differentiation (“gland in the gland”), T47D cells were poorly differentiated, and the $HER2^+$ BT474 cells gave rise to DCIS-like structures with marked nuclear pleomorphism, solid architecture, and central necrosis, termed “comedo necrosis” frequently associated with $HER2^+$ DCIS (Bane, 2013). HCC1806 formed keratin pearls characteristic of the rare basaloid breast carcinoma from which they are derived (Volk-Draper et al., 2012) (Figure 1G).

Fat pad (FP) xenografts have high proliferative indices, irrespective of ER status and molecular subtype. In contrast, in the MIND grafts, Ki67 indices were lower in luminal (MCF7, BT474, and T47D) than in basal-like (BT20 and HCC1806) tumors with 23%, 36%, and 23% versus 89% and 77%, respectively (Figure 1H). Thus, the ER^+ MIND xenografts resemble their clinical counterparts both histopathologically and with respect to tumor kinetics. The extensive intraductal growth is reminiscent of the prolonged DCIS state of many luminal breast cancers in the clinic (Sgroi, 2010).

MCF7-MIND versus MCF7-FP

To discern the impact of the engraftment site on disease characteristics, we compared MCF7-MIND with MCF7 grafted to the FP (MCF7-FP). MCF7-FP gave rise to large, highly vascularized tumors within 4–6 weeks (Figure 2A). The MCF7-MINDs became palpable at 5 months after injection. Macroscopically, milk ducts distended by MCF7 cells appeared as white lines in a barely enlarged mammary FP (Figure 2B, arrows). This was reflected in lower growth rates of MCF7-MIND versus MCF7-FP (Figure 2C). CD31 immunohistochemistry (IHC) indicated high endothelial cell density in MCF7-FP with an average of 76 units/cm^2 (Figures 2D and 2E), whereas in the MCF7-MIND an average of 31 units/cm^2 was found selectively around ducts distended by tumor cells (Figures 2E and 2F). The Ki67 index of MCF7-FP was 82% compared with 23% in MCF7-MIND (Figures 2G–2I), which is close to that of human ER^+ in situ and invasive breast cancers, known to have lower proliferative indices than triple-negative (TN) tumors (Fiche et al., 2000). The apoptotic index measured by cleaved CK18 was 50 times higher in MCF7-FP than in MCF7-MIND (Figures S2A and S2B). Invasive breast carcinomas often show a desmoplastic reaction involving collagen deposits and accounting for the characteristic “hardness” upon palpation. Picrosirius red, which stains type I and III collagens, revealed few dispersed fibers in MCF7-FP tumors, in line with their soft consistency (Figure 2J). In MCF7-MIND, collagen fibers accumulated around the ducts (Figure 2K); some invasive areas showed higher collagen content reminiscent of desmoplasia seen in human breast cancers (Figure 2L).

Microcalcifications are a common clinical characteristic of DCIS and are typically detected by mammography (Cox et al., 2012; Hofvind et al., 2011). They were absent from MCF7-FP (Figure 2M), HCC1806-MIND, and BT20-MIND (Figure S2C) but present in MCF7- and BT474-MIND as assessed by H&E staining, mammography (Figures 2N and S2C), or micro-computed tomography (Figure S2D).

Tumor Progression in the MCF7-MIND Model

Dispersed tumor cells were detected in the stroma by Alu in situ hybridization (Alu-ISH) 12 weeks after intraductal injection (Figure 3A). H&E staining revealed invasion (Figure 3B) and

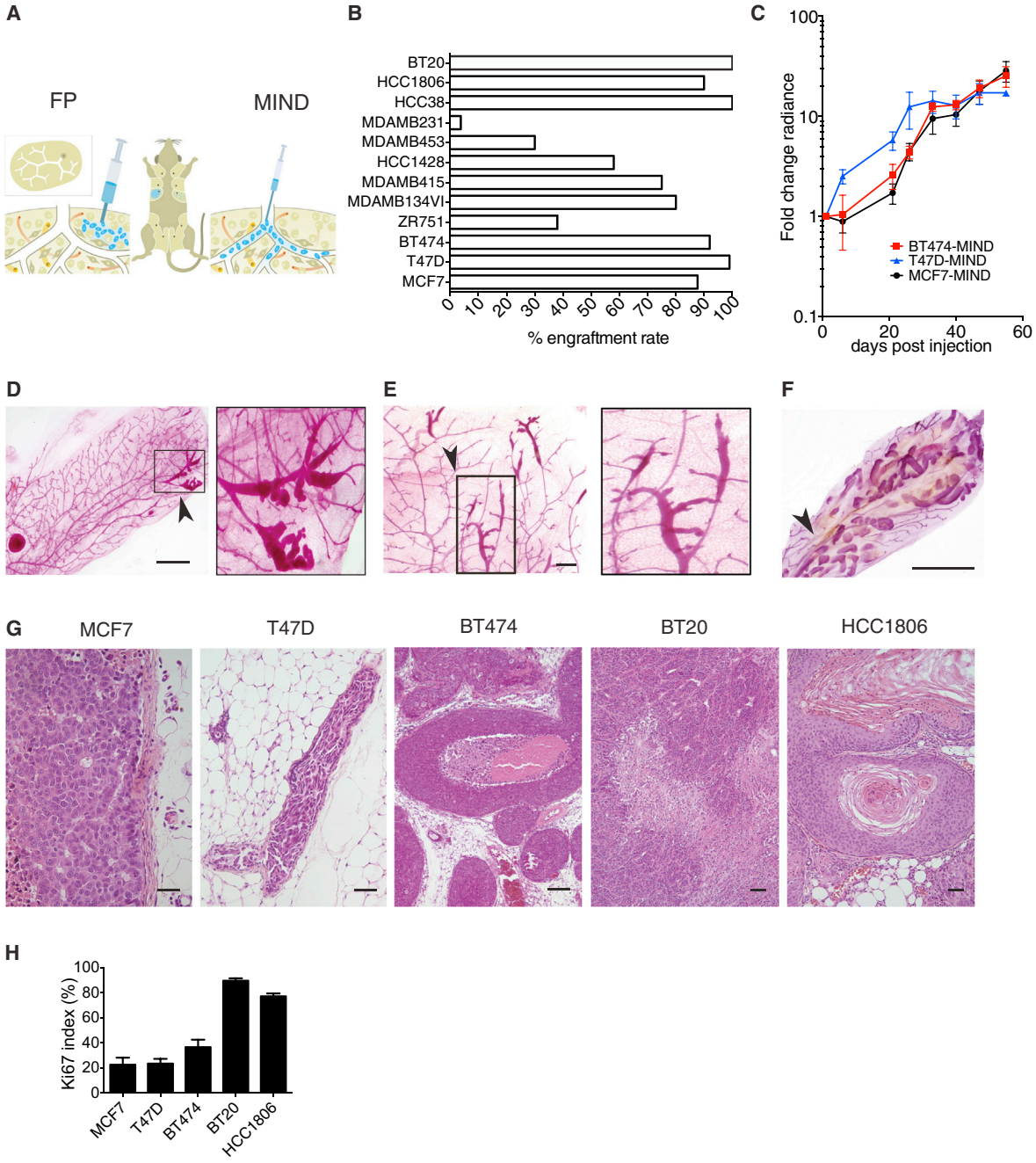


Figure 1. Intraductal Growth of Human Breast Cancer Cell Lines

(A) Scheme of the two xenograft approaches: tumor cells are injected either into the mammary fat pad (FP) or intraductally, via the teat (MIND).

(B) Bar graph showing MIND engraftment rates of 12 breast cancer cell lines representing distinct molecular subtypes. Tumor growth was assessed by bioluminescence and whole-mount analysis (number of analyzed glands 60 $\geq$ n $\geq$ 6).

(C) Tumor growth of ER$^+$ MCF7-, BT474-, and T47D-MINDs assessed by bioluminescence. Shown are means ± SEM.

(D–F) Whole-mount stereo micrographs of representative mammary glands (n $\geq$ 3) 8 weeks after intraductal injection of 5 × 10^4 MCF7 (D), T47D (E), or BT474 (F) cells. Arrowheads point to areas of intraductal growth. Scale bars: 2.5 mm (D), 0.5 mm (E), 5 mm (F).

(G) H&E-stained sections of different MINDs. Scale bars: 50 μm (MCF7, T47D); 200 μm (BT474); 100 μm (BT20); 25 μm (HCC1806).

(H) Bar graph showing Ki67 index of MCF7-, T47D-, BT474-, BT20-, and HCC1806-MIND 8 weeks after injection and 4 weeks for BT20 due to humane reasons. Data are shown as means ± SD.

See also Figure S1 and Table S1.

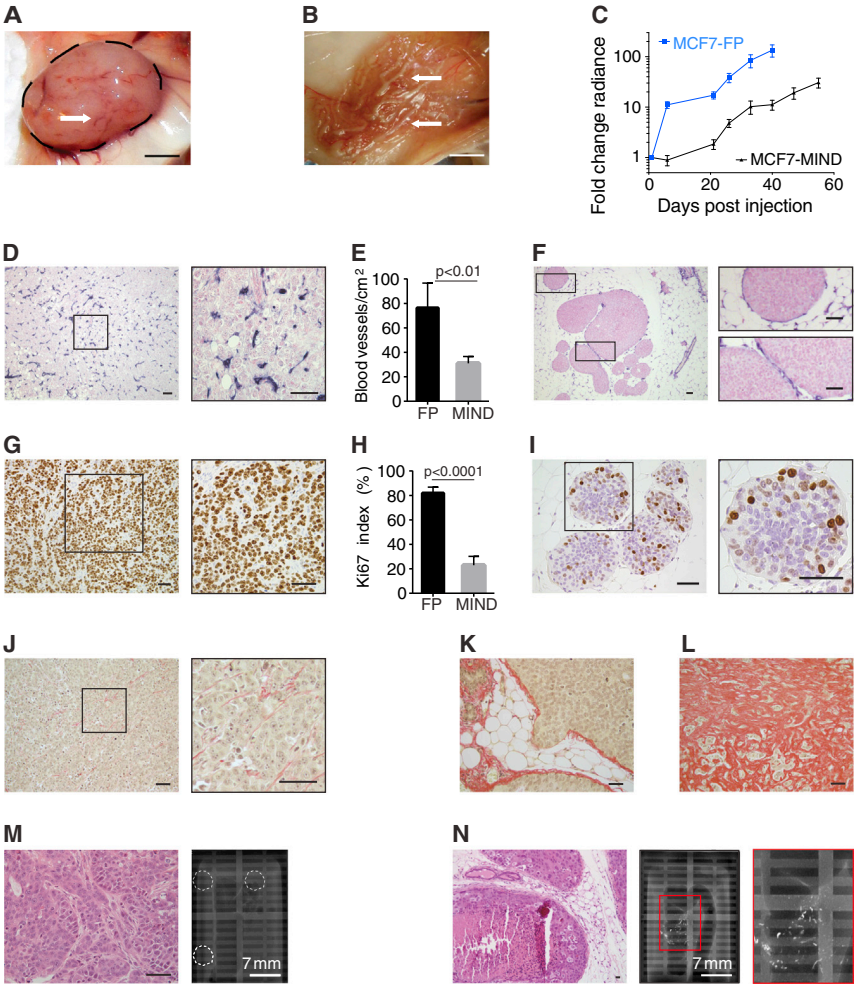


Figure 2. MCF7-MIND versus MCF7-FP

(A) Representative photograph of inguinal mammary gland 8 weeks after FP injection with 1×10^5 MCF7 cells. The arrow indicates blood vessels. Scale bar, 2 mm.

(B) Representative photograph of inguinal mammary gland with MCF7-MIND 20 weeks after injection of 5×10^4 MCF7 cells. Arrows indicate ducts engorged with tumor cells and appear white. Scale bar, 2 mm.

(C) Growth of MCF7-MIND and -FP assessed by radiance. Shown are means ± SEM.

(D–F) CD31 IHC of MCF7-FP (D) and MCF7-MIND (F), and quantification of vessel density in both models (E).

(G–I) Ki67 IHC on MCF7-FP (G) and MCF7-MIND (I), and their quantification (H).

(J–L) Picrosirius red-stained histological sections from MCF7-FP (J), and MCF7-MIND beginning invasion (K) and invasive (L).

(M and N) H&E staining of mammary tissue (left) and mammographs (right) of paraffin-embedded mammary glands, 6 weeks after FP injection (M) and 5 months after MIND injection (N). Image on the right (N) shows higher magnification of boxed area marked in the adjacent lower-magnification image. Dotted lines highlight where tumor samples were embedded.

Graphs represent means ± SD, p values by Student's t test. Scale bars, 50 μm (D, F, G, I, J–N). See also Figure S2.

intravasation of tumor cells (Figure 3C). Three to 6 months after injection, Alu-ISH showed human cells in the lungs (Figure 3D) that expressed ERα (Figure 3E). Bioluminescence imaging of organs resected post mortem revealed metastatic cells in multiple organs, the number of which increased over time (Figures 3F and 3G). The most frequent sites of metastasis were bones, lungs, and brain followed by the liver, pancreas, and kidney (Figures 3F and 3G). Thus, MCF7-MIND xenografts recapitulate the tumor progression of their clinical counterpart (Figure 3H) a finding that extended to ZR751-, BT474-, T47D-, and HCC1428-MIND (Figure 3I). The sensitive bioluminescence approach also detected lung metastases in the MCF7-FP, but few brain and no bone metastases (Figure 3I). Thus, the MIND model improves the physiological relevance of luminal breast cancer xenografts.

Response to Endocrine Therapy

The selective ER modulator tamoxifen, the selective ER downregulator fulvestrant, and aromatase inhibitors are mainstays in endocrine therapy of ER^+ tumors (Howell et al., 2004). To test whether MCF7-MIND is endocrine responsive and thereby evaluate its utility as a preclinical model for drug testing, we treated mice 4 weeks after cell injection with tamoxifen, fulvestrant, or solvent (Figure 4A). The treatments inhibited significantly tumor growth as measured by in vivo luminescence after 14 days (Figure 4B). In controls, 20% ± 0.8% of the cells were $Ki67^+$; tamoxifen decreased the Ki67 index to 8.4% ± 5% (Figure 4C) and induced cleaved CK18 in 18% of the tumor cells (Figures 4D and 4E), indicating that both decreased cell proliferation and increased apoptosis contributed to reduced tumor growth rates.

To mimic the use of fulvestrant in the advanced metastatic setting, we initiated treatment when metastatic disease was present (Figure 4F). A 2-month treatment decreased tumor burden as measured by in vivo luminescence (Figure 4G), ex vivo DsRed signal (Figure 4H), and ductal width (Figure 4I). Fulvestrant, which targets ER for degradation (Osborne et al., 2004), abrogated expression of both ER and its target, the progesterone receptor (PR) (Figure 4J). Postmortem analysis showed lung and brain metastases in control mice but not in fulvestrant-treated mice (Figure 4K).

Finally, postmenopausal patients with ER^+ tumors are frequently treated with aromatase inhibitors to achieve further estrogen depletion. As mice have more restricted aromatase expression than humans in non-ovarian tissue (Chow et al., 2009), we used ovariectomy to deplete E2 levels in MCF7-MIND bearing mice. All the control mice had to be euthanized within 8 months because of tumor burden, whereas 60% of the ovariectomized females were still alive after a year (Figure 4L). Thus, MCF7-MIND xenografts can be used as a model to study different settings of endocrine therapy in luminal breast cancer.

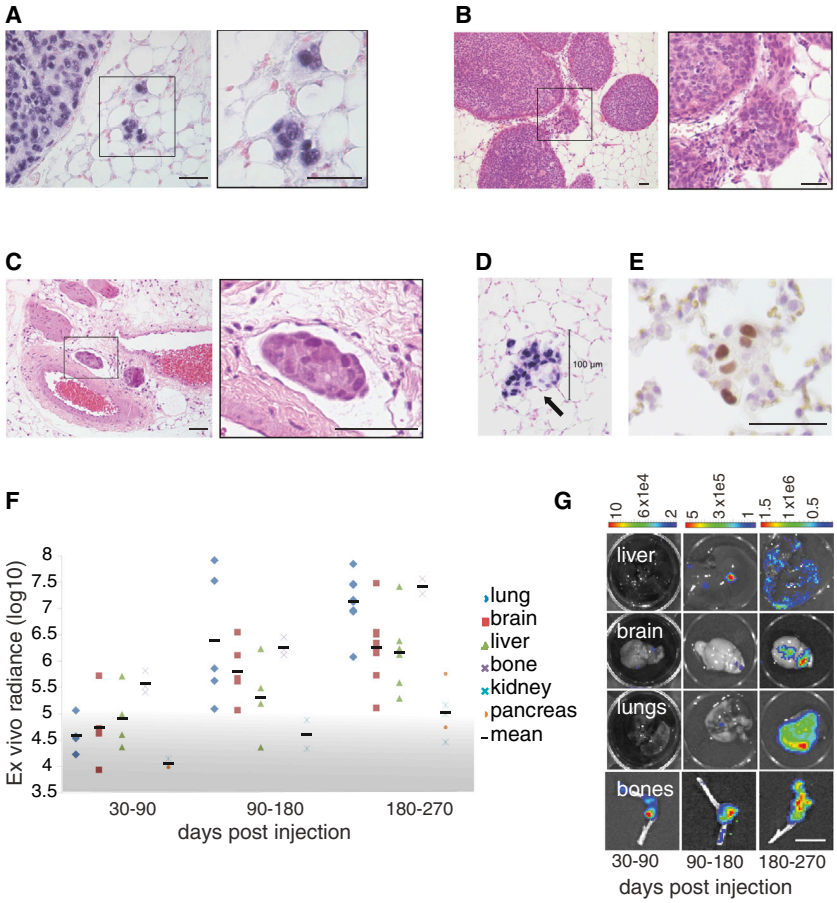


Figure 3. Hallmarks of Tumor Progression in the MCF7-MIND Model

(A–C) H&E-stained sections of MCF7-MIND 3 months after injection showing individual tumor cells that have invaded the stroma (A), small invasive focus next to in situ carcinoma (B), and disseminated tumor cells in a vessel, which is probably a lymph vessel (C). Images on the right show higher magnification of boxed areas marked in the adjacent lower-magnification image. Scale bars: 50 µm (A and C), 100 µm (B).

(D) Alu-ISH of a lung section 5 months after intraductal injection of MCF7 cells showing tumor cell colony (arrow).

(E) ERα IHC of lung micrometastases. Scale bar, 50 µm.

(F) Ex vivo bioluminescence from metastatic cells in different organs plotted over time of analysis; values $\leq 10 \times 10^5$ were considered as background (graded gray shading).

(G) Representative images of ex vivo luminescence showing MCF7-MIND cells in lungs, brain, liver, and bones dissected at different times after injection. Scale bar, 1.5 cm.

(H) Scheme summarizing the hallmarks of tumorigenesis in MCF7-MINDs over time.

(I) Summary of metastases from the ER$^+$ ZR751-, BT474-, T47D-, and HCC1428-MINDs and MCF7-FP, 5–37 weeks post injection; mice were euthanized and bioluminescence was measured in various organs.

I

xenograft	cell line characteristics			time in mice (days)	organs positive by IVIS			
	ER	PR	HER2		lung	brain	liver	bone
MIND:								
ZR751	+	neg	neg	88	3/5	0/5	1/5	5/5
ZR751	+	neg	neg	150	4/4	1/4	4/4	0/4
BT474	+	+	+	175	2/3	3/3	3/3	-
BT474	+	+	+	215	1/1	1/1	1/1	-
BT474	+	+	+	260	1/1	1/1	1/1	-
T47D	+	+	neg	45	2/2	1/2	2/2	-
HCC1428	+	+	neg	45	0/1	0/1	0/1	1/1
FP:								
MCF7	+	+	neg	45	10/10	2/10	0/4	0/4

To further assess the utility of the model for translational research and to mimic the clinical settings where patients are treated with endocrine therapy for long periods or until relapse, we treated MCF7-MIND bearing tumors for 3 months with fulvestrant or solvent. MIND bearing mammary glands were dissociated to single cells, and tumor cells were separated from mouse cells by fluorescence-activated cell sorting (FACS) based on DsRed expression. Their transcriptome was analyzed by RNA sequencing (Table S2). We identified 4,497 differentially expressed protein coding genes (logFC >2, $p < 0.05$) with 1,924 increased and 2,573 decreased upon endocrine treatment (Figures 4M and Table S3). Consistent with fulvestrant abrogating ER protein expression, Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis showed decreased expression of genes involved in ER signaling (Figure S3A). MetaCore analysis for biomarkers revealed "Breast Neoplasms" and "Breast Diseases" as the two top significant signatures, indicating clinical relevance (Figure S3B). Both MetaCore maps (Figure S3C) and network (Figure S3D) analyses revealed epithelial to mesenchymal transition (EMT) as the second most significant signature. Consistently, when 32 established EMT genes were used to interrogate the data the samples clustered into control and fulvestrant-treated groups (Figure 4N). This corresponds to what is observed in clinical samples where residual tumor cells surviving endocrine therapy are enriched for tumor-initiating cells with EMT features

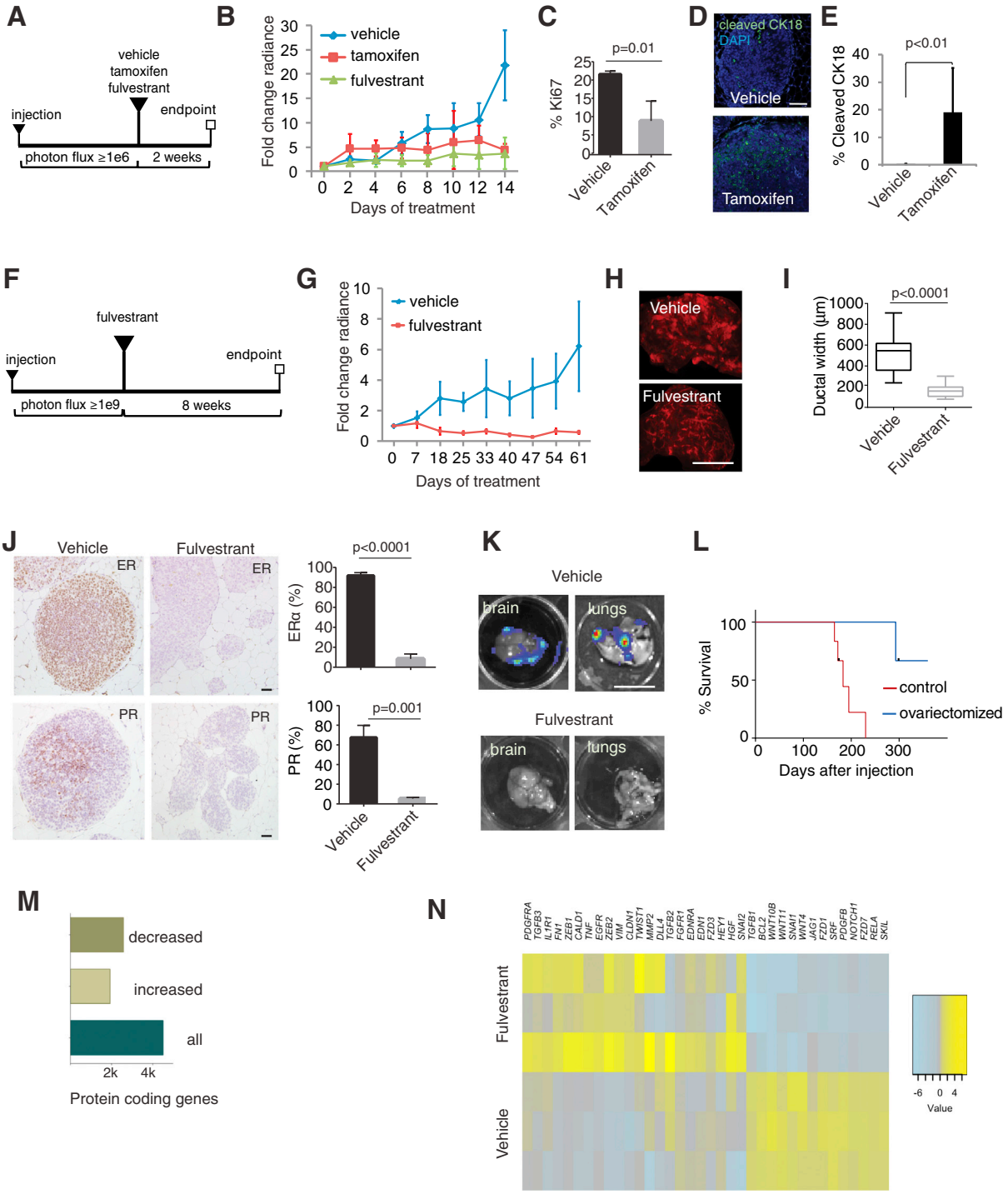


Figure 4. Response of MCF7-MIND to Endocrine Therapy

(A) Experimental scheme for short-term endocrine treatment: 4 weeks after injection of MCF7 when radiance $\geq 10 \times 10^6$, a 14-day-treatment with tamoxifen, fulvestrant, or vehicle was initiated.

(B) Graph showing tumor growth based on bioluminescence (n = 3). Statistical significance for the difference in fold-change radiance between treatment and control groups, $p < 0.02$ calculated by unpaired Student's t test, was reached after 14 days of treatment.

(C) Ki67 index of vehicle and tamoxifen-treated tumors.

(D and E) Immunofluorescence (D) and quantification (E) for cleaved CK18, an alternative marker of apoptosis adapted for cells, like MCF7, that do not express cleaved caspase-3 (Janicke, 2009), on vehicle and tamoxifen-treated tumors. Scale bar, 50 µm.

(legend continued on next page)

(Creighton et al., 2009), and suggests that the model may serve to identify biomarkers.

Molecular Signatures of MIND versus FP Model

To gain additional insights into the molecular mechanisms underlying the biological differences between the two models, we analyzed global gene expression of FACS-sorted MCF7:DsRed/luc2 cells that had been grown as MIND or FP using Affymetrix U133 Plus 2.0 arrays. Strikingly, 3,249 genes were differentially expressed between the two sites ($\geq$2-fold; $p < 0.05$) (Figure S4A and Table S4). Principal component analysis (PCA) and the PAM50 gene expression classifier for intrinsic subtype classification (Parker et al., 2009) were used to compare the models with a panel of 48 breast cancer cell lines (Neve et al., 2006). MCF7-MIND clustered with the luminal and MCF7-FP with the basal-like breast cancer cell lines (Figure 5A). In comparison with clinical breast tumor samples profiled with the same Affymetrix microarray platform (Guedj et al., 2012) and PAM50, the MCF7-MIND clustered with luminal B and MCF7-FP fell outside any tumor subtype cluster (Figure 5B).

To assess whether the microenvironment at the site of engraftment influences other breast cancer cells, we also profiled two basal-like cell lines, BT20 and HCC1806. Strikingly, not a single gene was differentially expressed between the two sites (Table S5), and both lines clustered with the basal-like cell lines and patient tumors (Figures 5A and 5B). Thus, the molecular signature of MCF7-MIND but not MCF7-FP resembles their clinical counterparts, and the in vivo observation that the intraductal microenvironment specifically favors tumor cells of the luminal type is corroborated at the molecular level.

To address the factors underlying the different phenotypes, we analyzed the most significantly changed genes bioinformatically. KEGG and REACTOME functional enrichment analysis revealed eight and 13, respectively, distinct terms enriched in the MCF7-FP (Figures 5C and 5D). Consistent with the low Ki67 index in MCF7-MIND, several genes related to cell proliferation and cell cycle *E2F1*, *MCM2*, *MKI67*, *MYBL2*, *BUB1*, *PLK1*, *CCNE1*, *CCND1*, and *CCNB1* were among the most differentially expressed genes, with higher levels in the MCF7-FP (Perou et al., 1999) (Tables S4, S5, S6, and S7). ECM components, focal adhesions, gap junction trafficking, and gap junction regulation as well as synthesis and oligomerization of connexins and transport of connexins to the plasma membrane were predicted to be affected, indicating differential regulation of gap junctions (Goodenough and Paul, 2009). By contrast, in MCF7-MIND only one term was enriched in either analysis, the Hippo signaling pathway and CXCR4/7 and CXCL12 (Figures 5C and 5D), respectively. Both of these have been implicated in breast cancer metastasis (Lamar et al., 2012; Muller et al., 2001), suggesting that the high propensity of MCF7-MINDs to metastasize may relate to their activation. Using the gene ontology term signaling pathways, interferon-, cytokine-mediated-, and vitamin D-receptor signaling pathways were found to be upregulated in MCF7-FP (Table S6).

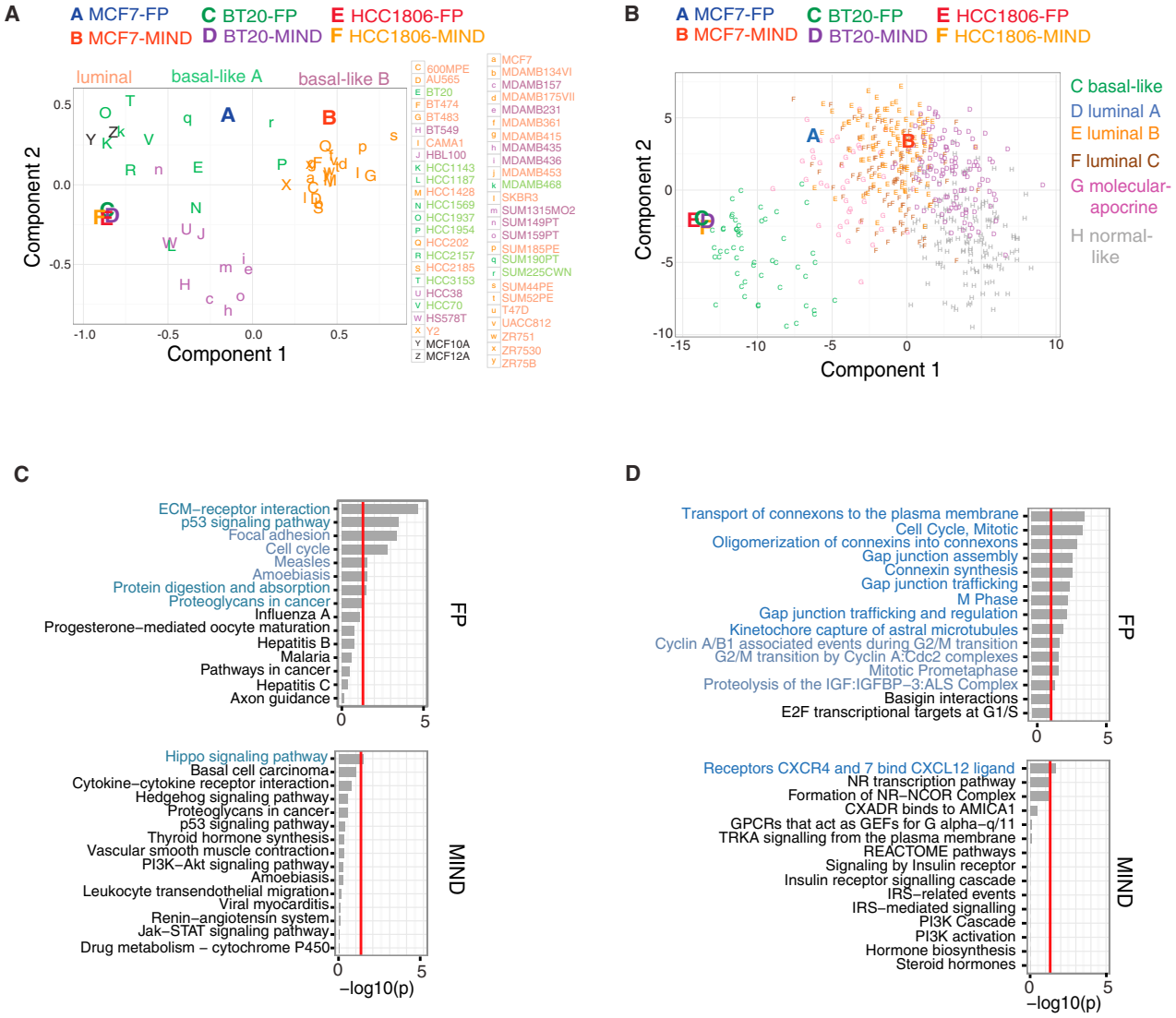
The ability of ER^+ cell lines to grow as MIND xenografts without exogenous hormones was unexpected, and led us to assess the expression of hormone receptors and receptors of downstream signaling pathways (Table S7). The receptors with known roles in mammary gland development upregulated in MCF7-MIND were those for growth hormone, androgen, E2, aryl hydrocarbon, and glucocorticoids. PR and the prolactin receptor were expressed at comparable levels in both microenvironments whereas insulin, fibroblast growth factor, and activin A receptors showed increased expression in MCF7-FP (Figure 5E). The increased ER and AR protein levels are confirmed (Figures 5F and 5G). Other ER^+ cell lines similarly showed increased ER but not PR expression in the MIND setting (Figures S4B–S4K).

The Role of SLUG in Maintaining Luminal Cell Phenotype

MCF7-FP cluster with basal-like cell lines, and the basal markers *CK5*, *CK6*, *TP63*, *S100A4*, *SNAI2* (SLUG), *VIM* (vimentin), and *ANXA1* (annexin A1) (de Graauw et al., 2010; Liu et al., 2013) were among the 50 most significantly enriched genes in MCF7-FP (Table S4), suggesting that the FP microenvironment induces a basal/EMT-like state. Gene set enrichment analysis revealed that EMT-related genes were enriched in MCF7-FP (Figure 6A). SLUG, vimentin, and annexin A1 proteins were readily detected in cultured MCF10A cells, which are basal cells, and in MCF7-FP but not in MCF7 cells in vitro nor in MCF7-MIND (Figure 6B). Similarly, Caveolin-1, ITGA6, and p63 were increased in MCF7-FP versus MCF7-MIND (Figure 6C). p63, a transcription factor important in maintaining basal cell fate (Yalcin-Ozuysal et al., 2010) is expressed in a subset of MCF7-MIND cells (Figure 6D).

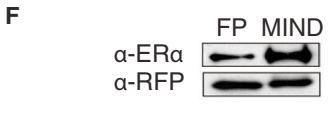
To directly assess whether the intraductal environment can induce a basal to luminal transition, we isolated MCF7-FP by FACS, reinjected the cells intraductally, and harvested them from the intraductal site on day 1 and day 20 after injection. The transcript levels of the luminally expressed *ESR1* and *AR* increased 5.1- and 11.3-fold, respectively (Figure 6E), whereas

(F) Experimental scheme for long-term fulvestrant treatment: Six weeks after injection of MCF7DsRed/luc2, when radiance $\geq 10 \times 10^9$, 60-day-treatment with fulvestrant was initiated ($n \geq 3$ per group).
(G) Graph showing tumor growth measured by radiance. Statistical significance ($p = 0.014$) by Mann-Whitney U test was reached at 18 days of treatment.
(H) Fluorescence stereomicroscopy of mammary glands with MCF7DsRed/luc2 treated with vehicle or fulvestrant; note fulvestrant-treated gland was exposed four times longer than control gland. Scale bar, 1 mm.
(I) Box plot showing ductal width in glands from control and fulvestrant-treated animals. Horizontal lines outside the box depict minimum and maximum values, upper and lower borders of the box represent lower and upper quartiles, and line inside the box identifies the median.
(J) ER- and PR-IHC on glands from mice treated with vehicle or fulvestrant, and histograms showing percentage of ER^+ and PR^+ cells. Scale bars, 20 μm.
(K) Bioluminescence images of lungs and brains isolated from mice after treatment with vehicle ($n = 3$) or fulvestrant ($n = 3$). Scale bar, 1 cm.
(L) Kaplan-Meier plot showing survival of females ovariectomized (blue) or sham-operated (red) 20 weeks after injection with MCF7-MIND ($n = 5$); $p < 0.05$ by log-rank (Mantel-Cox) test.
(M) Bar plot showing protein coding genes, expression levels of which were altered in MCF7-MIND by fulvestrant treatment.
(N) Heatmap of EMT-related genes in MCF7-MIND fulvestrant-treated and controls shown median-centered and log-scaled. Data are shown as means ± SD. See also Figure S3 and Tables S2 and S3.



E

Gene	FC	p value
GHR, growth hormone receptor	11.16	2.02E-10
NR3C1, glucocorticoid receptor	13.20	1.39E-08
IGF1R, insulin-like growth factor 1 receptor	-2.51	1.82E-07
AR, androgen receptor	4.72	2.81E-07
ESR1, estrogen receptor 1	2.73	2.16E-06
AHR, aryl hydrocarbon receptor	2.60	1.01E-05
FGFR3, fibroblast growth factor receptor 3	-1.81	2.54E-05
ACVR1, activin A receptor, type I	-1.89	2.61E-05
VDR, vitamin D receptor	-	ns
PRLR, prolactin receptor	-	ns
CSF2RB, colony stimulating factor 2 receptor	-	ns
PGR, progesterone receptor	-	ns
THRA, thyroid hormone receptor, alpha	-	ns



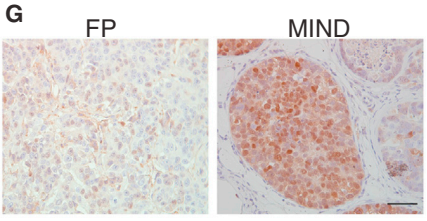


Figure 5. Molecular Signatures of MCF7-MIND versus MCF7-FP

(A) Global gene expression profiles of FACS-sorted cells derived from indicated xenografts compared with breast cancer cell lines grown in vitro by PCA using PAM50 classifier genes. First (x axis) and second (y axis) principal components are shown. Colors indicate subtypes: orange, luminal; green, basal-like A; magenta, basal-like B.

(legend continued on next page)

basal markers such as cytokeratins 6A and 14, vimentin, N-cadherin, and the transcription factors SLUG and ΔNp63 decreased up to 95% (CK14) within a day (Figure 6F). At the protein level, ER was upregulated between days 4 and 8 (Figures 6G and 6H). Thus, the intraductal environment suppresses the basal differentiation induced by the FP and promotes expression of luminal genes.

To test the basal transcription factors functionally, we ectopically expressed ΔNp63 and SLUG in MCF7 cells, and examined the effects on growth in MIND and FP. Ectopic expression of either protein was compatible with FP growth, but ectopic SLUG expression abrogated growth in MIND (Figure 6I). Next, we asked whether inhibition of SLUG expression enables MCF7 cells to grow in the FP without exogenous E2. Indeed, MCF7-shSLUG cells survived and grew (Figure 6J), suggesting that SLUG abrogates luminal features required for in vivo growth. As SLUG is a key effector of transforming growth factor β (TGFβ)/SMAD3 signaling (Xue et al., 2014), we asked whether TGFβ signaling is activated in the MCF7-FP. MetaCore network analysis showed that TGFβ signaling is increased at the transcriptional level in the MCF7-FP versus MCF7-MIND (Figure S5). Biochemically, increased phosphorylation of specifically SMAD3, not SMAD2 (Figure 6K), was detected in MCF7-FP. Of interest, another SMAD3-specific TGFβ target gene *SERPINE1* (Dennler et al., 1998) was increased 24.9-fold in the FP (Table S4).

Physiological and Clinical Relevance of the Intraductal Approach

Our findings that the FP microenvironment induces SLUG expression in MCF7 cells, which inhibited their growth, raised the question whether the difficulties experienced in establishing ER^+ PDXs are related to the engraftment site. To test this, we obtained tumor tissue from ten patients with ER^+ breast cancer, seven no special type (NST) and three lobular, and one with a TN breast cancer (Figure 7A). Single-cell suspensions were prepared from tumor tissues, lentivirally transduced with GFP and luciferase for subsequent tracing, and, depending on the number of tumor cells obtained, injected into 6–23 glands of 2–11 mice. All 11 tumors established xenografts (Figure 7A). In vivo tumor growth followed a biphasic growth pattern with a decrease in slope at around 10 weeks (Figure 7B). The ER^+ tumors were followed for up to 1 year in their recipients; the TN tumor cells grew more rapidly and recipients had to be euthanized by 30 weeks after injection. The presence of GFP-expressing human cancer cells was confirmed 12–40 weeks after injection by fluorescence stereomicroscopy (Figure 7C) and subsequent whole mounting, which revealed focally dilated milk ducts (Figure 7D). Alu-ISH confirmed the identity of human cells (Figure 7E), and H&E staining revealed that MIND-PDXs share morphological features of the patient tumors (Figure S6A). Most growth was in situ but invasive areas were identified (Figure S6A). The MIND-PDXs resembled the patient tumors with regard to ER and PR status (Figure S6B); the Ki67 index was frequently lower in the MIND-PDXs (Figures 7F–7H), which may relate to the fact that it is mostly established on in situ components in the PDXs whereas clinically it is assessed on the invasive parts.

An unresolved paradox in breast cancer research is the observation that primary cells from normal breast epithelium are more easily established in culture than are tumor cells (Hines et al., 2015). To assess whether the MIND approach reflects the biological properties of transformed and normal cells, we grafted cells from reduction mammoplasties intraductally. All four patient samples established themselves and proliferated (Figure S6C), but grew at lower rates than the tumor cells ($p < 0.05$) and plateaued at levels that are 100-fold lower than those reached by the tumor cells (Figure 7I). Individuals with mutations in *BRCA1* are at increased risk for breast cancer and have a larger progenitor cell compartment (Lim et al., 2009; Molyneux et al., 2010). Cells from three patients who had *BRCA1* mutations and underwent prophylactic mammectomy (Figure S6D) were engrafted and showed a trend to grow faster than the cells from control individuals (Figure 7I), further supporting the biological relevance of the MIND-PDXs.

We followed engrafted animals for up to 13 months and detected evidence of metastasis in all ten ER^+ and the TN MIND-PDXs, but not *BRCA1* nor normal cell grafts (Figure 7J). As observed in breast cancer patients, ER^+ PDXs frequently metastasized to brain (7 of 17) and bone (12 of 17), but rarely to liver or lungs (1 of 17) (Figure 7J).

Toward Personalized Clinical Models

Personalized medicine requires that cancer cells from individual patients be tested for response to therapy. Hence, we treated mice engrafted with TN PDX with doxorubicin and cyclophosphamide for 4 weeks similarly to patients, who receive four cycles of this combined chemotherapy. Tumor growth was inhibited (Figure 8A) and tumor shrinkage was evident upon stereoscopic inspection of the engrafted glands (Figure 8B). The GFP-labeled tumor cells were readily detected in distended ducts of the control mice, but fluorescence was sparse in the treated animals (Figure 8B). Postmortem radiance showed metastases in brain and bones of control but not of treated animals (Figure 8C). Mice bearing five different ER^+ PDX-MINDs received endocrine therapy with fulvestrant for at least 4 weeks. Tumor growth decreased in four cases; only a lobular carcinoma with *ERBB2* amplification, a genetic alteration associated with resistance to endocrine therapy, did not respond (Figure 8D). Thus, PDX-MINDs respond to therapy just as in the clinics.

All five ER^+ PDXs tested re-engrafted with an average 91% success rate, superior to the initial 76% (Figure S7A). Thus, ER^+ tumors, including lobular carcinomas, can readily be

(B) PCA of global gene expression profiles of patient samples and of cells derived from indicated xenografts. Color-coded letters indicate breast cancer subtypes.
(C and D) KEGG (C) or REACTOME (D) pathway analyses performed on genes upregulated in MCF7-FP (upper panel) and MCF7-MIND (lower panel). Top 15 groups based on p values are shown. Red line p value cutoff = 0.05, x axis −log10 of the p value. Pathways that are altered shown in blue, $p < 0.05$.
(E) Summary of differentially expressed receptors involved in mammary gland development. Fold change (FC) reflects gene expression of MIND/FP.
(F) ERα and red fluorescent protein (RFP) immunoblot of MCF7-FP and MCF7-MIND xenografts.
(G) AR IHC on histological sections of MCF7-FP and MCF7-MIND. Scale bar, 50 μm.
See also Figure S4 and Tables S4, S5, S6, and S7.

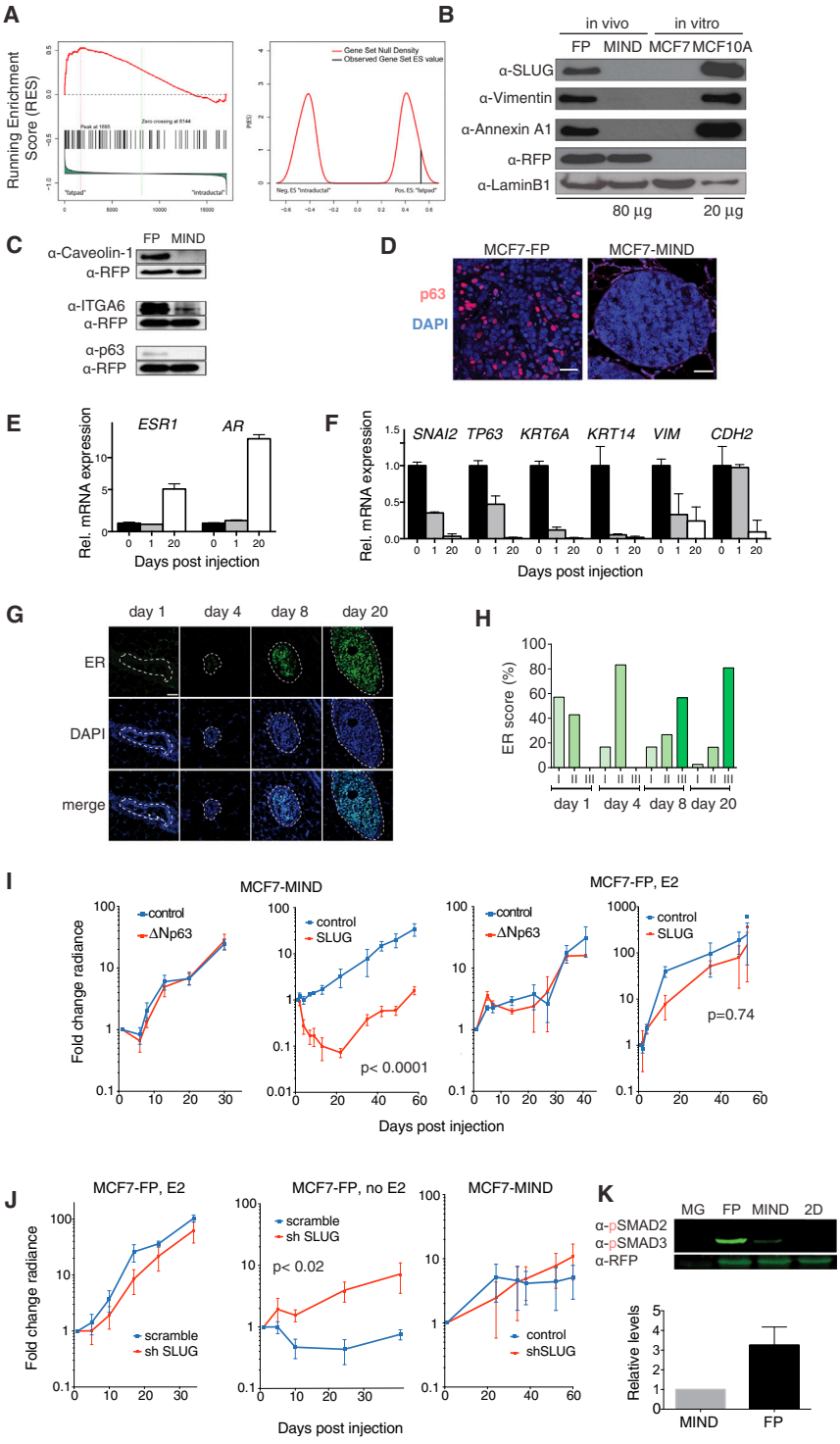


Figure 6. Microenvironment Affects Luminal Breast Cancer Cells through SLUG

(A) Gene set enrichment analysis showing over-representation of EMT category among genes differentially expressed between MCF7-FP and -MIND. High values on the left part of the red line show the enrichment with genes over-expressed in the FP (total number of genes 17,067). FDR shows the corrected p value (q value) adjusted for gene set size. NES denotes normalized enrichment score. Empirical null distribution of enrichment score calculated by randomly shuffling sample labels 1,000 times.

(B) Immunoblot of proteins encoded by selected differentially expressed genes from MCF7-MIND, MCF7-FP, and MCF7 and MCF10A growing in 2D in vitro. For MCF10A, 4-fold less protein lysate was loaded.

(C) Immunoblot analyses of selected proteins from MCF7-FP and MCF7-MIND.

(D) p63 IF of MCF7-MIND and MCF7-FP 1 month after injection counterstained with DAPI. Scale bars, 50 µm.

(E) Bar plot showing relative *ESR1* and *AR* mRNA expression normalized to *TBP1* mRNA in FACS-sorted MCF7-FP cells at different times after intraductal injection.

(F) Bar plot showing relative mRNA expression of various basal markers normalized to *TBP1* mRNA in FACS-sorted MCF7-FP cells at different times after intraductal injection.

(G) IF micrographs of mammary glands engrafted with FACS-sorted MCF7-FP cells at different times after injection. Dashed outlines highlight perimeter of cross-sectioned milk duct. Scale bar, 100 µm.

(H) Quantification of ER positivity based on signal intensity.

(I) Graphs showing bioluminescence of MCF7-FP or -MIND xenografts stably expressing luc2 and either GFP only or GFP together with ΔNp63 or SLUG after contralateral injection.

(J) Graph showing bioluminescence signal of xenografts of MCF7-FP and MCF7-MIND stably expressing luc2 and either GFP scramble or GFP shSLUG. MCF7-FP was tested both in the presence (E2) and absence (no E2) of exogenous E2. Graphs in (I) and (J) show means ± SEM. Statistical significance was determined by Mann-Whitney U test.

(K) Immunoblot analysis of pSMAD2 and pSMAD3 in control mouse mammary glands (MG), MCF7-FP, MCF7-MIND, and MCF7 cells growing in 2D; RFP loading control and quantification of the pSMAD3 level. Data in bar plots are shown as means ± SD.

See also Figure S5.

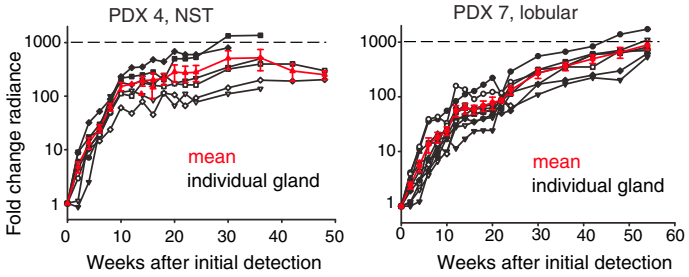
established as MIND. The TN PDX reached transplant generation 10 within 2 years (Figure S7B).

To assess whether tumor cells preserve their genomic characteristics when they grow as PDX-MIND, we sequenced 52 commonly mutated cancer genes (Table S8) using DNA isolated from ten paraffin-embedded tumors and respective PDXs. Analysis of MCF7- and MDAMB453-MINDs revealed the expected *PIK3CA* E545K and *PIK3CA* H1047R mutations and *TP53* P33R polymorphism (Figure 8E). Mutations and/or polymorphisms in patient samples were frequent in *TP53* (100%), *PIK3CA* (80%), and *KDR* (20%); individual tumors had *EGFR*, *FGFR2*, *SMAD4*, *KRAS*, *ATM*, *AKT1*, and *SMARCB1* mutations.

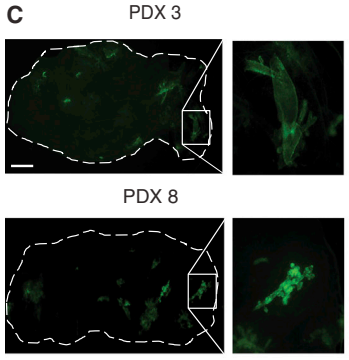
A

patient	age	tumor type	grade	pTNM	subtype	engraftment rate (%)
1	65	NST	3	pT2pN1a	Luminal B	38
2	51	NST	2	pT2pN1a	Luminal B	68
3	53	NST	1	pT2pN1a	Luminal A	83
4	58	NST	2	pT1cpN1a	Luminal B	90
5	87	NST	2	pT1cpN0	Luminal A	17
6	75	lobular	2	pT2pN3	Luminal A	100
7	69	lobular	3	pT2pN0	Luminal B	100
8	31	NST	3	pT2pN1mi	Luminal B	78
9	78	lobular	2	pT1cpN0	Luminal A	87
10	39	medullary	3	pT2pN0	Basal-like	87
11	39	NST	2	pT1cpN0	Luminal B	70

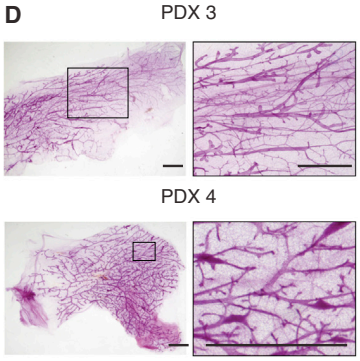
B



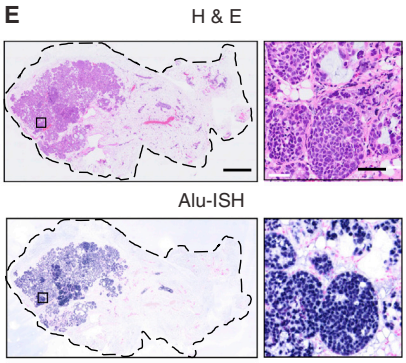
C



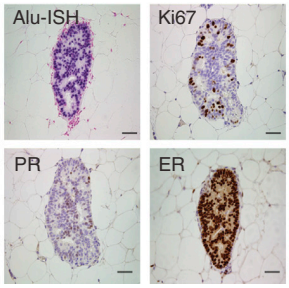
D



E



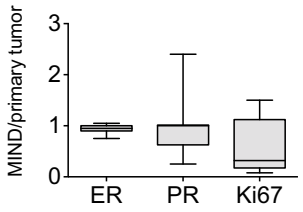
F



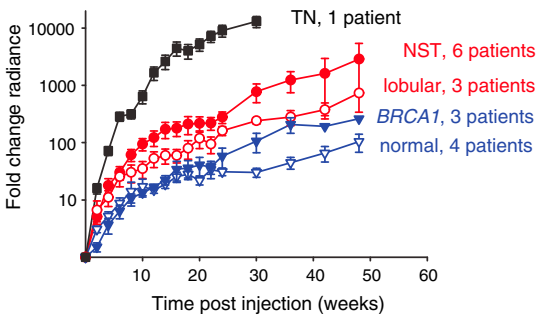
G

Patient tumor				PDX-MIND		
patient	ER	PR	Ki67	ER	PR	Ki67
1	100	0	90	95	0	30
2	100	10	25	90	0-100	5
3	100	90	17	100	95	5
4	100	5	29	100	12	35
5	95	30	20	100	28	30
6	100	100	16	92	25	5
7	100	0	26	75	0	2
8	100	60	80	90	40	60
9	100	80	10	95	40	1
10	0	0	>90	0	0	98

H



I



J

patient	tumor characteristics			time in mice (weeks)	organs positive by IVIS			
	ER	PR	HER2		bones	brain	lungs	liver
1	pos.	neg.	neg.	60	0/2	1/2	0/2	0/2
2	pos.	pos.	neg.	60	1/2	2/2	0/2	0/2
3	pos.	pos.	neg.	50	1/2	1/2	0/2	0/2
4	pos.	pos.	neg.	45	2/2	1/2	0/2	0/2
6	pos.	pos.	neg.	40	3/3	0/3	0/3	1/3
7	pos.	neg.	pos.	60	2/2	1/2	0/2	0/2
8	pos.	pos.	neg.	15	1/2	0/2	0/2	0/2
8	pos.	pos.	neg.	30	0/1	1/1	1/1	0/1
9	pos.	pos.	neg.	24	1/1	0/1	0/1	0/1
10	neg.	neg.	neg.	10	3/6	1/6	2/6	1/6
10	neg.	neg.	neg.	30	1/6	0/6	2/6	0/6

Figure 7. ER$^+$ MIND-PDXs

(A) Summary of the characteristics of patient tumors and the MINDs derived from them.

(B) Graphs showing the radiance of PDX-MINDs, no specific type (NST) or lobular, in individual glands (black) and mean thereof ± SEM (red).

(legend continued on next page)

The results were concordant in all ten pairs. Only three mutations present in tumor samples at low allele frequencies (*SMAD4*, *SMARCB1*, and *PIK3CA*) were not detected in the respective PDXs, suggesting dilution or loss of tumor subpopulations upon grafting. No de novo mutations were detected in any of the PDXs. Thus, PDX-MINDs do not acquire additional mutations in critical cancer genes and appear genomically stable.

DISCUSSION

The MIND approach addresses a long-standing need for better preclinical models of ER^+ breast cancer, and shows that the intraductal microenvironment enables ER^+ breast cancer cells to grow in vivo and to recapitulate the human disease. It offers several advantages over existing preclinical models. First, the proliferative indices are relatively low, with 23%–35% Ki67 for ER^+ cell lines. Second, tumors grow with systemic E2 levels of 10–60 pg/ml, comparable with those of postmenopausal women (<59 pg/ml) in whom most ER^+ breast cancers occur, so that mechanisms of endocrine resistance can now be studied in the context of appropriate E2 levels (Yue et al., 1994). Furthermore, it obviates the deleterious effects of excess E2, such as urinary retention, cystitis, hydronephrosis, and renal failure, which limited the utility of traditional xenografts (Gakhar et al., 2009; Levin-Allerhand et al., 2003; Pearse et al., 2009). The clinical relevance of the MCF7-MIND model is reflected at the molecular level in gene expression signatures similar to those of clinical samples. It remains to be tested whether the utility of the model extends to other hormone-responsive cancers, such as ovarian and thyroid carcinomas, and to selectively established adenomatous versus squamous lung carcinomas.

A potential drawback of MCF7-MIND as a preclinical model is the required immune suppression. The immune system is important in tumorigenesis (de Visser et al., 2006) and may affect the outcome of therapy. Its impact may differ between tumor subtypes, and its role in the luminal cancers is poorly defined (Kroemer et al., 2015). Future studies should extend the MIND model to mice with a humanized immune system (Kalscheuer et al., 2012).

The model offers opportunities to study breast cancer progression. The critical transition from in situ to invasive disease and spontaneous metastasis to relevant sites are now amenable to mechanistic studies when previous work relied on injection of a large number of tumor cells into the circulation or specific organ sites (Minn et al., 2005; Wang et al., 2015).

The distinct microenvironments dramatically alter gene expression in luminal tumor cells. The stroma bestows EMT-like changes on MCF7 cells and induces a basal differentiation program with high-level expression of SLUG. The intraductal microenvironment induces expression of ER and other hormone and growth factor receptors important in mammary gland development. It will be interesting to determine whether the observation that the hippo and the CXCR4/7 CXCL12 pathways, which are enriched in MCF7-MIND, is generalizable and functionally relevant to the metastatic behavior of the tumor cells. The genes modulated by fulvestrant show little overlap with established in vitro targets (Patani et al., 2014), but treatment duration and analytical platform differed. Of interest, we noticed that some genes among the 800 most differentially expressed genes, such as calpain 8 (*CAPN8*), heparanase (*HPSE*), and sphingomyelin phosphodiesterase 3 (*SMPD3*), were identified as in vivo E2 targets in the bovine breast with roles in ECM turnover and signaling (Li et al., 2006).

The finding that ER^+ PDXs grow readily in the mouse milk ducts suggests that the differences in hormone levels, the lack of human stroma, and human specific paracrine factors previously held responsible for the low engraftment rates are not so important (Rong et al., 1992; Utama et al., 2006). As some of the tumors proliferate less as PDX-MIND, we cannot exclude that some of the above factors may be important for these particular tumors. However, the observation that MIND tumors show two distinct growth rates point to the possibility that the differences in cell proliferation relate to the time of analysis. The ease with which primary tumor cells can now be grown in vivo opens exciting perspectives for translational research and personalized breast cancer therapy.

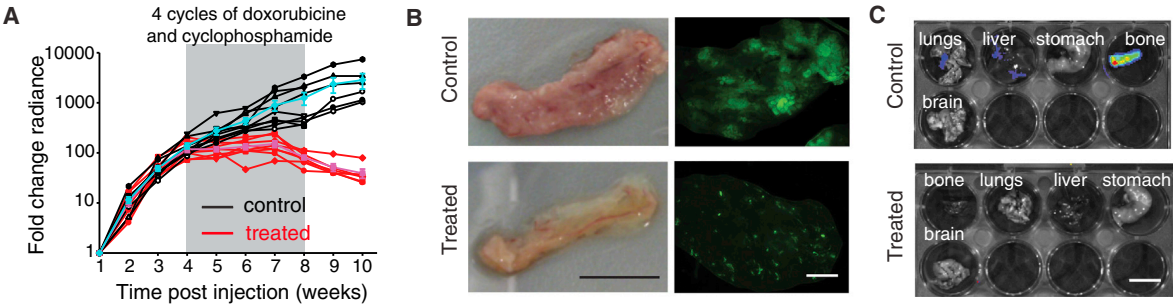
EXPERIMENTAL PROCEDURES

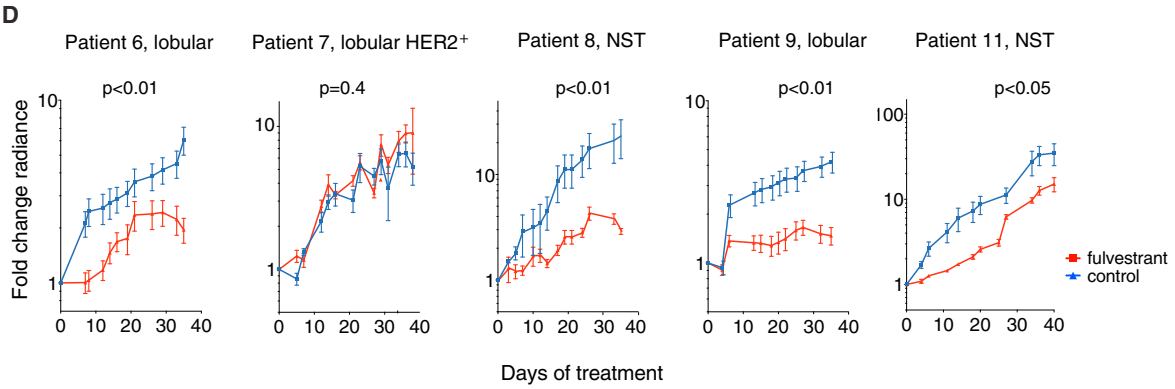
The details of cell culture, immunofluorescence, immunoblotting, qRT-PCR, and Alu-FISH are included in Supplemental Experimental Procedures.

Clinical Samples

The Commission cantonale d'éthique de la recherche sur l'être humain approved the studies (45-05 and 72-04), and informed consent was obtained from all subjects. Normal breast tissue was obtained from women undergoing reduction mammoplasties with no previous history of breast cancer, as described by Tanos et al. (2013), and freshly resected tumor material of pinhead size was obtained from the pathologist. Human tissue was mechanically dissociated, digested overnight at 37°C with 10 mg/ml collagenase A (11088793001; Roche) in DMEM/F-12 (11039-021; Gibco) supplemented with 1% penicillin-streptomycin (15070-063; Thermo Fisher Scientific) and

(C) Fluorescence stereo micrographs of inguinal mammary gland 20 weeks after injection of PDXs (patients 3 and 8). Dashed outlines highlight perimeter of the engrafted mammary gland. Scale bar, 3 mm.
(D) Stereo micrographs of whole-mounted mammary glands 20 weeks after injection of primary cancer cells from patients 3 and 4. Scale bars, 2 mm.
(E) Overview and blow-up of adjacent sections stained by H&E and Alu-ISH from PDX-MIND derived from tumor in patient 1. Dashed outlines highlight perimeter of the engrafted mammary gland. Scale bars, 2 mm and 50 μm.
(F) Alu-ISH and Ki67-, ER-, and PR-IHC on histological sections of MIND derived from the tumor in patient 4. Scale bar, 50 μm.
(G) Summary of ER, PR, and Ki67 status in patient tumors and corresponding PDX-MINDs.
(H) Box plot showing range of ratios of ER, PR, and Ki67 expression. For patients 1, 7, and 10, the ratios were corrected to 1 when patient tumor and PDX presented the value of 0 for % PR^+ or ER^+ cells. Horizontal lines outside the box depict minimum and maximum values, upper and lower borders of the box represent lower and upper quartiles, and line inside the box identifies the median.
(I) Mean radiance of MIND-PDXs of different tumor types or breast epithelial cells derived from normal donors or *BRCA1* mutation carriers. Curves represent means ± SEM of measurements performed on multiple samples.
(J) Summary of the metastatic spread in clinical relevant organs measured by ex vivo luminescence at indicated times after PDX-MIND engraftment.
See also Figure S6.





E

No.	Patient-PDXs ID	Gene	Chr	AA variant	Substitution	Position	AlleleFreq
1	T1	*PIK3CA*	3	p.I391M	c.A1173G	exon3	32.43%
		TP53	17	p.P33R	c.C98G	exon3	87.7%
	PDX-MIND 1	*PIK3CA*	3	p.I391M	c.A1173G	exon3	23.24%
		TP53	17	p.P33R	c.C98G	exon3	89.05%
2	T2	*EGFR*	7	p.S123T	c.C368T	exon3	17.05%
		TP53	17	p.P33R	c.C98G	exon3	95.45%
	PDX-MIND 2	*EGFR*	7	p.S123T	c.C368T	exon3	20.01%
		TP53	17	p.P33R	c.C98G	exon3	96.78%
3	T3	*PIK3CA*	3	p.E545K	c.G1633A	exon10	21.2%
		FGFR2	10	p.N437K	c.T1311A	exon9	22.83%
		TP53	17	p.P33R	c.C98G	exon3	67.36%
	PDX-MIND 3	*PIK3CA*	3	p.E545K	c.G1633A	exon10	46.83%
		FGFR2	10	p.N437K	c.T1311A	exon9	49.12%
		TP53	17	p.P33R	c.C98G	exon3	62.18%
4	T4	*PIK3CA*	3	p.Q546K	c.C1636A	exon10	26.42%
		PIK3CA	3	p.N1044K	c.T3132A	exon21	24.78%
		TP53	17	p.P33R	c.C98G	exon3	86.4%
		SMAD4	18	p.R361C	c.C1081T	exon9	16.5%
	PDX-MIND 4	*PIK3CA*	3	p.Q546K	c.C1636A	exon10	48.59%
		TP53	17	p.P33R	c.C98G	exon3	90.82%
5	T6	*PIK3CA*	3	p.H1047R	c.A3140G	exon21	24.25%
		KIT	4	p.V473M	c.G1417A	exon9	6.94%
		KDR	4	p.Q472H	c.A1416T	exon11	37.47%
		KRAS	12	p.G12V	c.G35T	exon2	27.71%
		TP53	17	p.P33R	c.C98G	exon3	80.06%
	PDX-MIND 6	*PIK3CA*	3	p.H1047R	c.A3140G	exon21	1.43%
		KIT	4	p.V473M	c.G1417A	exon9	**
		KDR	4	p.Q472H	c.A1416T	exon11	50.55%
		KRAS	12	p.G12V	c.G35T	exon2	4.37%
		TP53	17	p.P33R	c.C98G	exon3	66.67%
	MCF7	*PIK3CA*	3	p.E545K	c.G1633A	exon10	35.19%

No.	Patient-PDXs ID	Gene	Chr	AA variant	Substitution	Position	AlleleFreq
6	T7	*PIK3CA*	3	p.Q546R	c.A1637G	exon10	**
		TP53	17	p.P33R	c.C98G	exon3	76.57%
	PDX-MIND 7	*PIK3CA*	3	p.Q546R	c.A1637G	exon10	17%
		TP53	17	p.P33R	c.C98G	exon3	65.77%
7	T8	*PIK3CA*	3	p.N107S	c.A320G	exon2	33.24%
		PIK3CA	3	p.E545K	c.G1633A	exon10	18.15%
		KDR	4	p.Q472H	c.A1416T	exon11	54.65%
		TP53	17	p.P33R	c.C98G	exon3	42.25%
		SMARCB1	22	p.W385X	c.G1154A	exon9	5.37%
	PDX-MIND 8	*PIK3CA*	3	p.N107S	c.A320G	exon2	47%
		PIK3CA	3	p.E545K	c.G1633A	exon10	19.77%
		KDR	4	p.Q472H	c.A1416T	exon11	53.48%
		TP53	17	p.P33R	c.C98G	exon3	**
8	T9	*PIK3CA*	3	p.N1044K	c.T3132A	exon21	22.82%
		ATM	11	p.P604S	c.C1810T	exon12	45.39%
		TP53	17	p.P33R	c.C98G	exon3	98.34%
	PDX-MIND 9	*PIK3CA*	3	p.N1044K	c.T3132A	exon21	17.04%
		ATM	11	p.P604S	c.C1810T	exon12	39.56%
		TP53	17	p.P33R	c.C98G	exon3	97.68%
9	T10	*PIK3CA*	3	p.I391M	c.A1173G	exon7	7.6%
		TP53	17	p.P87fs	c.259_265del	exon2	57.03%
		TP53	17	p.P33R	c.C98G	exon3	90.71%
	PDX-MIND 10	*PIK3CA*	3	p.I391M	c.A1173G	exon7	**
		TP53	17	p.P87fs	c.259_265del	exon2	99.51%
		TP53	17	p.P33R	c.C98G	exon3	85.84%
10	T11	*AKT1*	14	p.D46E	c.C138A	exon3	88.4%
		TP53	17	p.P33R	c.C98G	exon3	97.71%
	PDX-MIND 11	*AKT1*	14	p.D46E	c.C138A	exon3	99.54%
		TP53	17	p.P33R	c.C98G	exon3	97.29%
	MDAMB453	*PIK3CA*	3	p.H1047R	c.A3140G	exon21	62.28%
		TP53	17	p.P33R	c.C98G	exon3	98.56%

Figure 8. Clinical Relevance of MIND-PDXs

(A) Bioluminescence of TN PDX-MIND from patient 10 treated with doxorubicin and cyclophosphamide for 4 weeks (gray area) and control. Each black and red line represents one gland. Blue and purple lines represent the mean ± SEM of control and treated glands, respectively ($n \geq 3$).

(legend continued on next page)

1% fungizone (cat. #15290-018; Thermo Fisher) in continuous agitation (40 rpm) as described by Sflomos et al. (2015). Samples were rinsed and erythrocytes lysed with Red Blood Cell Lysis Buffer (R7757; Sigma) and dissociated to single cells with 0.25% Gibco Trypsin-EDTA (15400-054; Thermo Fisher) for 2 min. Trypsin was inactivated with PBS/2% calf serum (CS) followed by incubation with 5 μg/ml DNase (1284932; Roche) in L-15 medium (11415; Gibco) at 37°C for 2 min. 2% CS in PBS was added, and the cells were filtered through a 70-μm pore size filter (cat. #352350; BD Falcon) and counted. Primary tumor cells were transduced with bifunctional reporter fusion gene ffLuc2/EGFP lentivirus (GFP-luc2) under control of the cytomegalovirus promoter. Lentiviral spin infection was performed at 25°C for 2.5 hr at 2,500 rpm as described by Yalcin-Ozuysal et al. (2010).

Animal Experiments

Animal experiments were performed in accordance with protocols approved by the Service de la Consommation et des Affaires Vétérinaires of Canton de Vaud. SCID/beige and NOD.Cg-*Prkdc*^{scid} *Il2rg*^{tm1Wjl}/SzJ mice (NSG) were purchased from Charles River and Jackson Laboratories, respectively. Mice were anesthetized by intraperitoneal injection with 10 mg/kg xylazine and 90 mg/kg ketamine (Graeub). Intraductal injections of single-cell suspensions were performed as described by Behbod et al. (2009) but without surgically opening the mouse. Engrafted mammary glands were harvested 4–32 weeks after intraductal injections and 2–6 weeks after FP injections, fixed in 4% paraformaldehyde for IHC or snap-frozen for RNA and protein isolation. Mammary gland whole mounts were prepared as described by Ayyanan et al. (2011). Stereomicrographs were acquired with an M205 FA (Leica).

ACCESSION NUMBERS

The Gene Expression Omnibus accession numbers for the transcriptomics data reported in this study are GEO: GSE68694 and GSE74608.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and eight tables and can be found with this article online at http://dx.doi.org/10.1016/j.ccell.2016.02.002.

AUTHOR CONTRIBUTIONS

Conceptualization, G.S., M.F., C.B.; Formal Analysis, T.M., R. J., J.V.; Investigation, G.S., V.D., A.A., V.S., L.B.; Resources, W.R., J.D., A.T., M.F., Writing, G.S., C.B; Funding Acquisition, C.B.

ACKNOWLEDGMENTS

We thank E. Anderson and J. Rougemont for advice, M. Wirth, J. Dessimoz, O. Burri, S. Leuba, and the EPFL core facilities for technical assistance, B. Bisig and E. Missiaglia for the mutation analysis, S. Cagnet and V. Simanis for reading of the manuscript, D. Lepori for the mammography, R. de Hoogt, S. Vidic (Janssen Pharmaceuticals, Beerse, Belgium), W. van Weerden (Erasmus University, Rotterdam) and S.A. Mani (M.D. Anderson, Houston) for lentiviral vectors, and E. Hill (AstraZeneca Pharmaceuticals) for providing fulvestrant (ICI 182780). Microarray and RNA-sequencing data were generated and analyzed at the Genomic Technologies Facility of the University of Lausanne. The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking (grant agreement no. 115188) for the PREDECT consortium (www.predect.eu) resources composed of financial contributions from EU-FP7 and EFPIA companies in kind contribution. The Web address of the Innovative Medicines Initiative is http://www.imi.europa.eu/. V.D. and V.S. were supported by the Swiss Cancer Research foundation, and R.J. and L.B. by the SNF.

Received: May 4, 2015
Revised: November 16, 2015
Accepted: February 8, 2016
Published: March 3, 2016

REFERENCES

Arrowsmith, J. (2011). Trial watch: phase III and submission failures: 2007-2010. Nat. Rev. Drug Discov. *10*, 87.

Ayyanan, A., Laribi, O., Schuepbach-Mallepell, S., Schrick, C., Gutierrez, M., Tanos, T., Lefebvre, G., Rougemont, J., Yalcin-Ozuysal, O., and Brisken, C. (2011). Perinatal exposure to bisphenol A increases adult mammary gland progesterone response and cell number. Mol. Endocrinol. *25*, 1915–1923.

Bane, A. (2013). Ductal carcinoma in situ: what the pathologist needs to know and why. Int. J. Breast Cancer *2013*, 914053.

Behbod, F., Kittrell, F.S., LaMarca, H., Edwards, D., Kerbawy, S., Heestand, J.C., Young, E., Mukhopadhyay, P., Yeh, H.W., Allred, D.C., et al. (2009). An intraductal human-in-mouse transplantation model mimics the subtypes of ductal carcinoma in situ. Breast Cancer Res. *11*, R66.

Chan, S.R., Vermi, W., Luo, J., Lucini, L., Rickert, C., Fowler, A.M., Lonardi, S., Arthur, C., Young, L.J., Levy, D.E., et al. (2012). STAT1-deficient mice spontaneously develop estrogen receptor alpha-positive luminal mammary carcinomas. Breast Cancer Res. *14*, R16.

Chow, J.D., Simpson, E.R., and Boon, W.C. (2009). Alternative 5′-untranslated first exons of the mouse Cyp19A1 (aromatase) gene. J. Steroid Biochem. Mol. Biol. *115*, 115–125.

Clinchy, B., Gazdar, A., Rabinovsky, R., Yefenof, E., Gordon, B., and Vitetta, E.S. (2000). The growth and metastasis of human, HER-2/neu-overexpressing tumor cell lines in male SCID mice. Breast Cancer Res. Treat. *61*, 217–228.

Cottu, P., Marangoni, E., Assayag, F., de Cremoux, P., Vincent-Salomon, A., Guyader, C., de Plater, L., Elbaz, C., Karboul, N., Fontaine, J.J., et al. (2012). Modeling of response to endocrine therapy in a panel of human luminal breast cancer xenografts. Breast Cancer Res. Treat. *133*, 595–606.

Cox, R.F., Hernandez-Santana, A., Ramdass, S., McMahon, G., Harmey, J.H., and Morgan, M.P. (2012). Microcalcifications in breast cancer: novel insights into the molecular mechanism and functional consequence of mammary mineralisation. Br. J. Cancer *106*, 525–537.

Creighton, C.J., Li, X., Landis, M., Dixon, J.M., Neumeister, V.M., Sjolund, A., Rimm, D.L., Wong, H., Rodriguez, A., Herschkowitz, J.I., et al. (2009). Residual breast cancers after conventional therapy display mesenchymal as well as tumor-initiating features. Proc. Natl. Acad. Sci. USA *106*, 13820–13825.

de Graauw, M., van Miltenburg, M.H., Schmidt, M.K., Pont, C., Lalai, R., Kartopawiro, J., Pardali, E., Le Devedec, S.E., Smit, V.T., van der Wal, A., et al. (2010). Annexin A1 regulates TGF-beta signaling and promotes metastasis formation of basal-like breast cancer cells. Proc. Natl. Acad. Sci. USA *107*, 6340–6345.

de Visser, K.E., Eichten, A., and Coussens, L.M. (2006). Paradoxical roles of the immune system during cancer development. Nat. Rev. Cancer *6*, 24–37.

(B) Representative light (left) and epifluorescence stereoscopic (right) images of inguinal mammary glands from xenografted mice treated with chemotherapy or solvent. Scale bars, 1 cm (left) and 0.2 mm (right).
(C) Representative ex vivo luminescence images of indicated organs from engrafted mice treated with chemotherapy or solvent. Scale bar, 1.5 cm.
(D) Response to endocrine therapy in 5 ER$^+$ PDXs measured by radiance at and after the onset of fulvestrant treatment. Lines represent means ± SEM of control and fulvestrant-treated glands. Statistical significance was determined by Mann-Whitney U test.
(E) Mutational repertoire of primary tumors and matched PDX as well as MCF7- and MDAMB453-MINDs. For each patient's tumor-PDX pair, concordant mutations are highlighted by green in the box on the right, whereas mutations detected only in patient tumors are highlighted by yellow. Double asterisks denote that visual inspection of the sequence data identified mutations also in the PDX, as detailed in Supplemental Experimental Procedures.
See also Figure S7 and Table S8.

Dennler, S., Itoh, S., Vivien, D., ten Dijke, P., Huet, S., and Gauthier, J.M. (1998). Direct binding of Smad3 and Smad4 to critical TGF beta-inducible elements in the promoter of human plasminogen activator inhibitor-type 1 gene. EMBO J. *17*, 3091–3100.

DeRose, Y.S., Wang, G., Lin, Y.C., Bernard, P.S., Buys, S.S., Ebbert, M.T., Factor, R., Matsen, C., Milash, B.A., Nelson, E., et al. (2011). Tumor grafts derived from women with breast cancer authentically reflect tumor pathology, growth, metastasis and disease outcomes. Nat. Med. *17*, 1514–1520.

Fiche, M., Avet-Loiseau, H., Maugard, C.M., Sagan, C., Heymann, M.F., Leblanc, M., Classe, J.M., Fumoleau, P., Dravet, F., Mahe, M., and Dutrillaux, B. (2000). Gene amplifications detected by fluorescence in situ hybridization in pure intraductal breast carcinomas: relation to morphology, cell proliferation and expression of breast cancer-related genes. Int. J. Cancer *89*, 403–410.

Gakhar, G., Wight-Carter, M., Andrews, G., Olson, S., and Nguyen, T.A. (2009). Hydronephrosis and urine retention in estrogen-implanted athymic nude mice. Vet. Pathol. *46*, 505–508.

Goodenough, D.A., and Paul, D.L. (2009). Gap junctions. Cold Spring Harb. Perspect. Biol. *1*, a002576.

Guedj, M., Marisa, L., de Reynies, A., Orsetti, B., Schiappa, R., Bibeau, F., MacGrogan, G., Lerebours, F., Finetti, P., Longy, M., et al. (2012). A refined molecular taxonomy of breast cancer. Oncogene *31*, 1196–1206.

Guiu, S., Wolfer, A., Jacot, W., Fumoleau, P., Romieu, G., Bonnetain, F., and Fiche, M. (2014). Invasive lobular breast cancer and its variants: how special are they for systemic therapy decisions? Crit. Rev. Oncol. Hematol. *92*, 235–257.

Hait, W.N. (2010). Anticancer drug development: the grand challenges. Nat. Rev. Drug Discov. *9*, 253–254.

Hidalgo, M., Amant, F., Biankin, A.V., Budinska, E., Byrne, A.T., Caldas, C., Clarke, R.B., de Jong, S., Jonkers, J., Maelandsmo, G.M., et al. (2014). Patient-derived xenograft models: an emerging platform for translational cancer research. Cancer Discov. *4*, 998–1013.

Hines, W.C., Yaswen, P., and Bissell, M.J. (2015). Modelling breast cancer requires identification and correction of a critical cell lineage-dependent transduction bias. Nat. Commun. *6*, 6927.

Hofvind, S., Iversen, B.F., Eriksen, L., Styr, B.M., Kjellevold, K., and Kurz, K.D. (2011). Mammographic morphology and distribution of calcifications in ductal carcinoma in situ diagnosed in organized screening. Acta Radiol. *52*, 481–487.

Howell, A., Robertson, J.F., Abram, P., Lichinitser, M.R., Elledge, R., Bajetta, E., Watanabe, T., Morris, C., Webster, A., Dimery, I., and Osborne, C.K. (2004). Comparison of fulvestrant versus tamoxifen for the treatment of advanced breast cancer in postmenopausal women previously untreated with endocrine therapy: a multinational, double-blind, randomized trial. J. Clin. Oncol. *22*, 1605–1613.

Janicke, R.U. (2009). MCF-7 breast carcinoma cells do not express caspase-3. Breast Cancer Res. Treat. *117*, 219–221.

Kalscheuer, H., Danzl, N., Onoe, T., Faust, T., Winchester, R., Goland, R., Greenberg, E., Spitzer, T.R., Savage, D.G., Tahara, H., et al. (2012). A model for personalized in vivo analysis of human immune responsiveness. Sci. Transl. Med. *4*, 125ra130.

Kratz, A., Ferraro, M., Sluss, P.M., and Lewandrowski, K.B. (2004). Case records of the Massachusetts General Hospital. Weekly clinicopathological exercises. Laboratory reference values. N. Engl. J. Med. *351*, 1548–1563.

Kroemer, G., Senovilla, L., Galluzzi, L., Andre, F., and Zitvogel, L. (2015). Natural and therapy-induced immunosurveillance in breast cancer. Nat. Med. *21*, 1128–1138.

Lamar, J.M., Stern, P., Liu, H., Schindler, J.W., Jiang, Z.G., and Hynes, R.O. (2012). The Hippo pathway target, YAP, promotes metastasis through its TEAD-interaction domain. Proc. Natl. Acad. Sci. USA *109*, E2441–E2450.

Lee, A.V., Oesterreich, S., and Davidson, N.E. (2015). MCF-7 cells—changing the course of breast cancer research and care for 45 years. J. Natl. Cancer Inst. *107*, djv073.

Levin-Allerhand, J.A., Sokol, K., and Smith, J.D. (2003). Safe and effective method for chronic 17beta-estradiol administration to mice. Contemp. Top. Lab. Anim. Sci. *42*, 33–35.

Li, R.W., Meyer, M.J., Van Tassell, C.P., Sonstegard, T.S., Connor, E.E., Van Amburgh, M.E., Boisclair, Y.R., and Capuco, A.V. (2006). Identification of estrogen-responsive genes in the parenchyma and fat pad of the bovine mammary gland by microarray analysis. Physiol. Genomics *27*, 42–53.

Lim, E., Vaillant, F., Wu, D., Forrest, N.C., Pal, B., Hart, A.H., Asselin-Labat, M.L., Gyorki, D.E., Ward, T., Partanen, A., et al. (2009). Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. Nat. Med. *15*, 907–913.

Liu, T., Zhang, X., Shang, M., Zhang, Y., Xia, B., Niu, M., Liu, Y., and Pang, D. (2013). Dysregulated expression of Slug, vimentin, and E-cadherin correlates with poor clinical outcome in patients with basal-like breast cancer. J. Surg. Oncol. *107*, 188–194.

Logan, G.J., Dabbs, D.J., Lucas, P.C., Jankowitz, R.C., Brown, D.D., Clark, B.Z., Oesterreich, S., and McAuliffe, P.F. (2015). Molecular drivers of lobular carcinoma in situ. Breast Cancer Res. *17*, 76.

Minn, A.J., Gupta, G.P., Siegel, P.M., Bos, P.D., Shu, W., Giri, D.D., Viale, A., Olshen, A.B., Gerald, W.L., and Massague, J. (2005). Genes that mediate breast cancer metastasis to lung. Nature *436*, 518–524.

Molyneux, G., Geyer, F.C., Magnay, F.A., McCarthy, A., Kendrick, H., Natrajan, R., Mackay, A., Grigoriadis, A., Tutt, A., Ashworth, A., et al. (2010). BRCA1 basal-like breast cancers originate from luminal epithelial progenitors and not from basal stem cells. Cell Stem Cell *7*, 403–417.

Muller, A., Homey, B., Soto, H., Ge, N., Catron, D., Buchanan, M.E., McClanahan, T., Murphy, E., Yuan, W., Wagner, S.N., et al. (2001). Involvement of chemokine receptors in breast cancer metastasis. Nature *410*, 50–56.

Neve, R.M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F.L., Fevr, T., Clark, L., Bayani, N., Coppe, J.P., Tong, F., et al. (2006). A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. Cancer Cell *10*, 515–527.

Ni, M., Chen, Y., Lim, E., Wimberly, H., Bailey, S.T., Imai, Y., Rimm, D.L., Liu, X.S., and Brown, M. (2011). Targeting androgen receptor in estrogen receptor-negative breast cancer. Cancer Cell *20*, 119–131.

Osborne, C.K., Wakeling, A., and Nicholson, R.I. (2004). Fulvestrant: an oestrogen receptor antagonist with a novel mechanism of action. Br. J. Cancer *90* (*Suppl 1*), S2–S6.

Parker, J.S., Mullins, M., Cheang, M.C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. J. Clin. Oncol. *27*, 1160–1167.

Patani, N., Dunbier, A.K., Anderson, H., Ghazoui, Z., Ribas, R., Anderson, E., Gao, Q., A'Hern, R., Mackay, A., Lindemann, J., et al. (2014). Differences in the transcriptional response to fulvestrant and estrogen deprivation in ER-positive breast cancer. Clin. Cancer Res. *20*, 3962–3973.

Pearse, G., Frith, J., Randall, K.J., and Klinowska, T. (2009). Urinary retention and cystitis associated with subcutaneous estradiol pellets in female nude mice. Toxicol. Pathol. *37*, 227–234.

Perou, C.M., Jeffrey, S.S., van de Rijn, M., Rees, C.A., Eisen, M.B., Ross, D.T., Pergamenschikov, A., Williams, C.F., Zhu, S.X., Lee, J.C., et al. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc. Natl. Acad. Sci. USA *96*, 9212–9217.

Rong, S., Bodescot, M., Blair, D., Dunn, J., Nakamura, T., Mizuno, K., Park, M., Chan, A., Aaronson, S., and Vande Woude, G.F. (1992). Tumorigenicity of the met proto-oncogene and the gene for hepatocyte growth factor. Mol. Cell Biol. *12*, 5152–5158.

Schmid, C.W., and Deininger, P.L. (1975). Sequence organization of the human genome. Cell *6*, 345–358.

Sflomos, G., Shamsheddin, M., and Brisken, C. (2015). An ex vivo model to study hormone action in the human breast. J. Vis. Exp. e52436.

Sgroi, D.C. (2010). Preinvasive breast cancer. Ann. Rev. Pathol. *5*, 193–221.

Sikora, M.J., Cooper, K.L., Bahreini, A., Luthra, S., Wang, G., Chandran, U.R., Davidson, N.E., Dabbs, D.J., Welm, A.L., and Oesterreich, S. (2014). Invasive

lobular carcinoma cell lines are characterized by unique estrogen-mediated gene expression patterns and altered tamoxifen response. Cancer Res. *74*, 1463–1474.

Tanos, T., Sflomos, G., Echeverria, P.C., Ayyanan, A., Gutierrez, M., Delaloye, J.-F., Raffoul, W., Fiche, M., Dougall, W., Schneider, P., et al. (2013). Progesterone/RANKL is a major regulatory axis in the human breast. Sci. Transl. Med. *5*, 182ra155.

Utama, F.E., LeBaron, M.J., Neilson, L.M., Sultan, A.S., Parlow, A.F., Wagner, K.U., and Rui, H. (2006). Human prolactin receptors are insensitive to mouse prolactin: implications for xenotransplant modeling of human breast cancer in mice. J. Endocrinol. *188*, 589–601.

Valdez, K.E., Fan, F., Smith, W., Allred, D.C., Medina, D., and Behbod, F. (2011). Human primary ductal carcinoma in situ (DCIS) subtype-specific pathology is preserved in a mouse intraductal (MIND) xenograft model. J. Pathol. *225*, 565–573.

Vargo-Gogola, T., and Rosen, J.M. (2007). Modelling breast cancer: one size does not fit all. Nat. Rev. Cancer *7*, 659–672.

Volk-Draper, L.D., Rajput, S., Hall, K.L., Wilber, A., and Ran, S. (2012). Novel model for basaloid triple-negative breast cancer: behavior in vivo and response to therapy. Neoplasia *14*, 926–942.

Wang, H., Yu, C., Gao, X., Welte, T., Muscarella, A.M., Tian, L., Zhao, H., Zhao, Z., Du, S., Tao, J., et al. (2015). The osteogenic niche promotes early-stage bone colonization of disseminated breast cancer cells. Cancer Cell *27*, 193–210.

Weinberg, R. (2011). Robert Weinberg. Nat. Biotech. *29*, 192.

Xue, J., Lin, X., Chiu, W.T., Chen, Y.H., Yu, G., Liu, M., Feng, X.H., Sawaya, R., Medema, R.H., Hung, M.C., and Huang, S. (2014). Sustained activation of SMAD3/SMAD4 by FOXM1 promotes TGF-beta-dependent cancer metastasis. J. Clin. Invest. *124*, 564–579.

Yalcin-Ozuysal, O., Fiche, M., Guitierrez, M., Wagner, K.U., Raffoul, W., and Brisken, C. (2010). Antagonistic roles of Notch and p63 in controlling mammary epithelial cell fates. Cell Death Differ. *17*, 1600–1619.

Yue, W., Zhou, D., Chen, S., and Brodie, A. (1994). A new nude mouse model for postmenopausal breast cancer using MCF-7 cells transfected with the human aromatase gene. Cancer Res. *54*, 5092–5095.

Zhang, X., Claerhout, S., Prat, A., Dobrolecki, L.E., Petrovic, I., Lai, Q., Landis, M.D., Wiechmann, L., Schiff, R., Giuliano, M., et al. (2013). A renewable tissue resource of phenotypically stable, biologically and ethnically diverse, patient-derived human breast cancer xenograft models. Cancer Res. *73*, 4885–4897.

Package 'TTMap'

April 23, 2018

Type Package

Title Two-Tier Mapper: a clustering tool based on topological data analysis

Version 1.1.0

Date 2017-07-12

Author Rachel Jeitziner

Maintainer Rachel Jeitziner <rachel.jeitziner@epfl.ch>

Description
TTMap is a clustering method that groups together samples with the same deviation in comparison to a control group. It is specially useful when the data is small. It is parameter free.

License GPL-2

Suggests BiocStyle, airway

Depends rgl, colorRamps

Imports grDevices,graphics,stats,utils, methods, SummarizedExperiment, Biobase

biocViews Software, Microarray, DifferentialExpression, MultipleComparison, Clustering, Classification

R topics documented:

TTMap-package *Two-Tier Mapper: a clustering tool based on topological data analysis*

Description

TTMap is a clustering method that groups together samples with the same deviation in comparison to a control group. It is specially useful when the data is small. It is parameter free.

Details

The DESCRIPTION file: TTMap/DESCRIPTION Version 1.0

Author(s)

Rachel Jeitziner Maintainer: Rachel Jeitziner <rachel.jeitziner@epfl.ch>

References

R. Jeitziner et al., TTMap, 2018, DOI:arXiv:1801.01841

See Also

rgl, colorRamps

Examples

```
#to be found in \code{\link[TTMap]{ttmap_sgn_genes}}
```

calcul_e *Calculation of the value of epsilon*

Description

Calculation of the value of epsilon

Usage

```
calcul_e(dd5, pvalcutoff = 0.95, tt1, alpha = 1, S =
colnames(tt1$Normal.mat))
calcul_e_single(dd5, pvalcutoff = 0.95, tt1, alpha = 1, S =
colnames(tt1$Normal.mat))
```

Arguments

dd5	distance matrix as created by generate_mismatch_distance
pvalcutoff	cutoff of 0.05 percent (default) or less
tt1	output of control_adjustment
alpha	a cutoff value for the FC between the group of control and the disease group
S	subset of columns to be considered

Value

al number representing the cutoff to choose for the relatedness with dd5

Author(s)

Rachel Jeitziner

See Also

control_adjustment, hyperrectangle_deviation_assessment, ttmap_sgn_genes, generate_mismatch_distanc

Examples

```
##--
library(airway)
data(airway)
airway <- airway[rowSums(assay(airway))>80,]
assay(airway) <- log(assay(airway)+1,2)
ALPHA <- 1
the_experiment <- TTMap::make_matrices(airway,
seq_len(4), seq_len(4) + 4,
rownames(airway), rownames(airway))
TTMAP_part1prime <-TTMap::control_adjustment(
normal.pcl = the_experiment$CTRL,
tumor.pcl = the_experiment$TEST,
normalname = "The_healthy_controls",
dataname = "Effect_of_cancer",
org.directory = tempdir(), e = 0, P = 1.1, B = 0);
Kprime <- 4;
TTMAP_part1_hda <-
TTMap::hyperrectangle_deviation_assessment(x =
TTMAP_part1prime,
k = Kprime,dataname = "Effect_of_cancer",
normalname = "The_healthy_controls");
annot <- c(paste(colnames(
the_experiment$TEST[,-(seq_len(3))]), "Dis", sep = "."),
paste(colnames(the_experiment$CTRL[,
-seq_len(3)]), "Dis", sep = "."))
dd5_sgn_only <-TTMap::generate_mismatch_distance(
TTMAP_part1_hda,
select=rownames(TTMAP_part1_hda$Dc.Dmat), alpha = ALPHA)
e <- TTMap::calcul_e(dd5_sgn_only, 0.95, TTMAP_part1prime, 1)
```

---

control_adjustment *Calculates a corrected control group, discovers outliers in it.*

Description

control_adjustment function finds outliers in the control group and removes them

Usage

```
control_adjustment(normal.pcl, tumor.pcl, normalname, dataname,
org.directory = "", A = 1, e = 0, meth = 0, P = 1.1, B = 0)
```

Arguments

normal.pcl	the control matrix with annotation as obtained by $CTRL from make_matrices
tumor.pcl	the disease/test data matrix with annotation as obtained by $TEST from make_matrices
normalname	A name for the corrected control files
dataname	the name of the project
org.directory	where the outputs should be saved
A	integer if A=0 then the difference to the median is calculated otherwise the difference to the mean.
e	integer giving how far to the median an outlier is at least
meth	value or method that defines how to replace outliers, default is set to replace by the median
P	if more than P percent of features are outliers the feature is removed, by default all are kept
B	Batch vector a vector for normal and test samples with a same number corresponding to a same batch

Details

control_adjustment calculates a corrected control group, discovers outliers in it.

Value

Several files are created

paste(org.directory,normalname,".normMesh",sep = "")
: The normal matrix with only common features with the test matrix. This file is only created if the two have different rows

paste(org.directory,dataname,".normMesh",sep = "")
: The test matrix with only common features with the normal matrix. This file is only created if the two have different rows.

mean_vs_variance.pdf
: A pdf showing a plot of the mean (X axis) against the variances (Y axis) of each feature

mean_vs_variance_after_correction.pdf
: A pdf showing a plot of the mean (X axis) against the variances (Y axis) of each feature after correction of the control group

na_numbers_per_row.txt
: number of outliers per row

na_numbers_per_col.txt
: number of outliers per column

And values of ttmap_part1_ctrl_adj

e	Selected criteria for what is an outlier
tag.pcl	Annotation of features, ID of features and weight
Normal.mat	The control matrix without annotation and only with the common rows with Disease.mat
Disease.mat	The test/disease matrix without annotation and only with the common rows with Disease.mat

flat.Nmat	A list $mat being the corrected control matrix $m a record of the different numbers of removed genes per sample
record	numbers recording the number of columns in Disease.mat and Normal.mat
B	The batch vector B introduced in the begining
U1	The different batches in Normal.mat
U2	The different batches in Disease.mat

Author(s)

Rachel Jeitziner

See Also

hyperrectangle_deviation_assessment, ttmap ttmap_sgn_genes

Examples

```
##--
library(airway)
data(airway)
airway <- airway[rowSums(assay(airway))>80,]
assay(airway) <- log(assay(airway)+1,2)
ALPHA <- 1
the_experiment <- TTMap::make_matrices(airway,
seq_len(4), seq_len(4) + 4,
rownames(airway), rownames(airway))
TTMAP_part1prime <-TTMap::control_adjustment(
normal.pcl = the_experiment$CTRL,
tumor.pcl = the_experiment$TEST,
normalname = "The_healthy_controls",
dataname = "Effect_of_cancer",
org.directory = tempdir(), e = 0, P = 1.1, B = 0);
```

---

generate_correlation *Generates different distance matrices*

Description

Single cell complete mismatch distance, single cell complete mismatch distance with a parameter of cutoff, mismatch distance, correlation distance, p-value of correlation test distance and euclidean distance.

Usage

```
generate_single_cell_complete_mismatch(ttmap_part1_hda,
select, alpha = 1)
generate_single_cell_mismatch_with_parameter(ttmap_part1_hda,
select, alpha = 1)
generate_correlation(ttmap_part1_hda, select)
generate_euclidean(ttmap_part1_hda, select)
generate_mismatch_distance(ttmap_part1_hda, select, alpha = 1)
generate_p_val_correlation(ttmap_part1_hda, select)
```

Arguments

ttmap_part1_hda	an object given back by hyperrectangle_deviation_assessment
select	A sublist of rownames of ttmap_part1_hda$Dc.Dmat
alpha	A real number corresponding to a cutoff

Details

If one is interested only in clustering samples according to a list of genes belonging to a certain pathway, then this list is provided to the parameter select. Alpha is a cutoff for deviations that should be considered as noise, for gene expression data such as normalised RNA-seq or microarrays for instance a cutoff of 1, corresponding to a two fold change is being chosen.

Value

Distance matrix

Author(s)

Rachel Jeitziner

Examples

```
ttmap_part1_hda <- list()
ttmap_part1_hda$Dc.Dmat <- matrix(c(-1, 2, 0, -4, 5, 6), nrow = 2)
rownames(ttmap_part1_hda$Dc.Dmat) <- c("Gene1", "Gene2")
colnames(ttmap_part1_hda$Dc.Dmat) <- c("A", "B", "C")
dd <- TTMap::generate_mismatch_distance(ttmap_part1_hda, select =
rownames(ttmap_part1_hda$Dc.Dmat))
dd <- TTMap::generate_euclidean(ttmap_part1_hda, select =
rownames(ttmap_part1_hda$Dc.Dmat))
```

---

hyperrectangle_deviation_assessment

Calculation of deviation components

Description

hyperrectangle_deviation_assessment function calculates the hyperrectangle deviation assessment (HDA) that calculates the deviation components using normal_hda2 which calculates the normal component of the test sample and deviation_hda2 which calculates the deviation component.

Usage

```
hyperrectangle_deviation_assessment(x,
k = dim(x$Normal.mat)[2], dataname,
normalname,Org.directory = getwd())
```

Arguments

x	output object given back by control_adjustment, list
k	A factor if not all the lines in the control group should be kept
dataname	the name of the project
normalname	A name for the corrected control files
Org.directory	where the outputs should be saved

Details

The function performs the hyperrectangle deviation assessment (HDA)

Value

Outputs

Tdis.pcl	The matrix of the deviation components for each test sample
Tnorm.pcl	The matrix of the normal components for each test sample
NormalModel.pcl	The normal model used

Values

Dc.Dmat	the deviation component matrix composed of the deviation components of all the samples in the test group
m	the values of the filter function per sample in the test group

Author(s)

Rachel Jeitziner

See Also

control_adjustment, hyperrectangle_deviation_assessment, ttmap_sgn_genes

Examples

```
##a full example can be found in ttmap_sgn_genes
##--
library(airway)
data(airway)
airway <- airway[rowSums(assay(airway))>80,]
assay(airway) <- log(assay(airway)+1,2)
ALPHA <- 1
the_experiment <- TTMap::make_matrices(airway,
seq_len(4), seq_len(4) + 4,
rownames(airway), rownames(airway))
TTMAP_part1prime <-TTMap::control_adjustment(
normal.pcl = the_experiment$CTRL,
tumor.pcl = the_experiment$TEST,
normalname = "The_healthy_controls",
dataname = "Effect_of_cancer",
org.directory = tempdir(), e = 0, P = 1.1, B = 0);
Kprime <- 4;
```

```
TTMAP_part1_hda <-
TTMap::hyperrectangle_deviation_assessment(x =
TTMAP_part1prime,
k = Kprime, dataname = "Effect_of_cancer",
normalname = "The_healthy_controls");
```

---

make_matrices *Prepares the matrices for* control_adjustment

Description

make_matrices generates the control and the test matrice in the right format

Usage

```
make_matrices(mat, col_ctrl, col_test, NAME, CLID,
GWEIGHT = rep(1, dim(mat)[1]), EWEIGHT = 0)
```

Arguments

mat	the gene expressions can be **matrix**, **data.frame**, " RangedSummarizedExperiment", " ExpressionSet" format
col_ctrl	the columns in the matrix "mat" of the control samples
col_test	the columns in the matrix "mat" of the test samples
NAME	Name of genes,or annotation, e.g. WNT4
CLID	Identities of genes,e.g. ENSMUSG00000000001
GWEIGHT	the weight for each gene
EWEIGHT	the weight for each experiment

Details

make_matrices generates the test matrix and the control matrix in the format accepted by control_adjustment from a matrix object

Value

junk	A list containing $CTRL and $TEST the matrices to impute in control_adjustment

Author(s)

Rachel Jeitziner

See Also

control_adjustment, hyperrectangle_deviation_assessment, ttmap_sgn_genes, " RangedSummarizedExper

Examples

```
##--
##--
Aa = 6
B1 = 3
B2 = 3
C0 = 100
D0 = 10000
a0 = 4
b0 = 0.1
a1 = 6
b1 = 0.1
a2 = 2
b2 = 0.5
ALPHA = 1
E = 1
Pw = 1.1
Bw = 0
RA <- matrix(rep(0, Aa * D0), nrow = D0)
RB1 <- matrix(rep(0, B1 * D0), nrow = D0)
RB2 <- matrix(rep(0, B2 * D0), nrow = D0)
RA <- lapply(seq_len(D0 - C0), function(i) rnorm(Aa,
mean = a0, sd = sqrt(b0)))
RA<-do.call(rbind, RA)
RB1<- lapply(seq_len(D0 - C0), function(i) rnorm(B1,
mean = a0, sd = sqrt(b0)))
RB1 <- do.call(rbind, RB1)
RB2 <- lapply(seq_len(D0 - C0), function(i) rnorm(B2,
mean = a0, sd = sqrt(b0)))
RB2 <- do.call(rbind, RB2)
RA_c <- lapply(seq_len(C0), function(i) rnorm(Aa,
mean = a0, sd = sqrt(b0)))
RA_c <- do.call(rbind, RA_c)
RB1_c <- lapply(seq_len(C0), function(i) rnorm(B1,
mean = a1, sd = sqrt(b1)))
RB1_c <- do.call(rbind, RB1_c)
RB2_c <- lapply(seq_len(C0), function(i) rnorm(B2,
mean = a2, sd = sqrt(b2)))
RB2_c <- do.call(rbind, RB2_c)
norm1 <- rbind(RA, RA_c)
dis <- cbind(rbind(RB1, RB1_c), rbind(RB2, RB2_c))
colnames(norm1) <- paste("N", seq_len(Aa), sep = "")
rownames(norm1) <- c(paste("norm", seq_len(D0 - C0), sep = ""),
paste("diff", seq_len(C0), sep = ""))
colnames(dis) <- c(paste("B1", seq_len(B1), sep=""),
paste("B2", seq_len(B2), sep  =""))
rownames(dis)<-c(paste("norm",
seq_len(D0 - C0), sep = ""),
paste("diff", seq_len(C0), sep = ""))
the_experiment <- TTMap::make_matrices(cbind(norm1, dis),
col_ctrl = colnames(norm1),
col_test = colnames(dis), NAME = rownames(norm1),
CLID = rownames(norm1))
###other example using SummarizedExperiment
library(airway)
data(airway)
```

```
airway <- airway[rowSums(assay(airway))>80,]
assay(airway) <- log(assay(airway)+1,2)
the_experiment <- TTMap::make_matrices(airway,
seq_len(4), seq_len(4) + 4,
rownames(airway), rownames(airway))
```

`make_matrices-methods` *Prepares the matrices for* `control_adjustment`

Description

make_matrices generates the control (output $CTRL) and the test (output $TEST) matrice in the right format for `control_adjustment`

Methods

`signature(mat = "data.frame")` Method make_matrice for `data.frame` object.

`signature(mat = "matrix")` Method make_matrice for `matrix` object.

`signature(mat = "SummarizedExperiment")` Method make_matrice for `SummarizedExperiment` object.

`signature(mat = "RangedSummarizedExperiment")` Method make_matrice for `RangedSummarizedExperiment` object.

`signature(mat = "ExpressionSet")` Method make_matrice for `ExpressionSet` object.

`ttmap` *Visualisation of the clustering*

Description

Enables a quick view on the groups in the dataset (globally) and how locally they differ.

Usage

```
ttmap(ttmap_part1_hda, m1,
select = row.names(ttmap_part1_hda$Dc.Dmat),
ddd, e, filename = "TEST", n = 3, ad = 0, bd = 0, piq = 1,
dd = generate_mismatch_distance(ttmap_part1_hda = ttmap_part1_hda,
select = select), mean_value_m1 = "N", ni = 2)
```

Arguments

`ttmap_part1_hda`	list output of `hyperrectangle_deviation_assessment`
`m1`	either a user imputed vector whose names are the names of the samples with addition of .Dis. or by default it is the amount of deviation
`select`	Should all the features (default) or only a sublist be considered to calculate the distance

ddd	Annotation matrix with rownames the different sample names with addition of .Dis. There can be as many columns as wanted, but only the column n will be selected to annotated the clusters
e	integer parameter defining under which value two samples are considered to be close
filename	Name for the description file annotating the clusters
n	The column to be considered to annotate the clusters
ad	if ad!=0 then the clusters on the output picture will not be annotated
bd	if different than 0 (default), the output will be without outliers of the test data set (clusters composed of only "piq" element)
piq	parameter used to determine what small clusters are, see bd
dd	the distance matrix to be used
mean_value_m1	if == "N" the average of the values in m1 divided by the number of the samples are put into the legend (by default represents the average of the samples in a cluster of the mean-deviation of the features) otherwise it will show the average value of the values in m1 (is useful for instance if m1 represents the age of the samples)
ni	The column to consider to annotate the samples (is put into parenthesis) for the description file

Details

Is the Two-tiers Mapper function. The output is an interactive image of the clusters in the different layers.

Value

all	the clusters in the overall group
low	the clusters in the lower quartile group
mid1	the clusters in the first middle quartile group
mid2	the clusters in the second middle quartile group
high	the clusters in the higher quartile group

Author(s)

Rachel Jeitziner

See Also

control_adjustment, hyperrectangle_deviation_assessment, ttmap_sgn_genes

Examples

```
##--
library(airway)
data(airway)
airway <- airway[rowSums(assay(airway))>80,]
assay(airway) <- log(assay(airway)+1,2)
ALPHA <- 1
the_experiment <- TTMap::make_matrices(airway,
```

```
seq_len(4), seq_len(4) + 4,
rownames(airway), rownames(airway))
TTMAP_part1prime <-TTMap::control_adjustment(
normal.pcl = the_experiment$CTRL,
tumor.pcl = the_experiment$TEST,
normalname = "The_healthy_controls",
dataname = "Effect_of_cancer",
org.directory = tempdir(), e = 0, P = 1.1, B = 0);
Kprime <- 4;
TTMAP_part1_hda <-
TTMap::hyperrectangle_deviation_assessment(x =
TTMAP_part1prime,
k = Kprime,dataname = "Effect_of_cancer",
normalname = "The_healthy_controls");
annot <- c(paste(colnames(
the_experiment$TEST[,-(seq_len(3))]),"Dis", sep = "."),
paste(colnames(the_experiment$CTRL[,
-seq_len(3)]), "Dis", sep = "."))
annot <- cbind(annot, annot)
rownames(annot)<-annot[, 1]
dd5_sgn_only <-TTMap::generate_mismatch_distance(
TTMAP_part1_hda,
select=rownames(TTMAP_part1_hda$Dc.Dmat), alpha = ALPHA)
TTMAP_part2 <-
TTMap::ttmap(TTMAP_part1_hda, TTMAP_part1_hda$m,
select = rownames(TTMAP_part1_hda$Dc.Dmat), annot,
e = TTMap::calcul_e(dd5_sgn_only, 0.95, TTMAP_part1prime, 1),
filename = "first_comparison", n =  1, dd = dd5_sgn_only)
```

---

ttmap_sgn_genes *Gives a list of associated genes per cluster*

Description

ttmap_sgn_genes function

Usage

```
ttmap_sgn_genes(ttmap_part2_gtlmap, ttmap_part1_hda,
ttmap_part1_ctrl_adj, c, n = 2, a = 0,
filename = "TEST2", annot = ttmap_part1_ctrl_adj$tag.pcl,
col = "NAME", path = getwd(), Relaxed = 1)
ttmap_sgn_genes_inter2(q, ttmap_part1_hda, alpha = 0)
ttmap_sgn_genes_inter(q, ttmap_part1_hda, alpha = 0)
```

Arguments

ttmap_part2_gtlmap
: output of ttmap

ttmap_part1_hda
: output of hyperrectangle_deviation_assessment

ttmap_part1_ctrl_adj
: output of control_adjustment

c	annotation file of the samples
n	column to give the name to the cluster
a	cutoff to be considered different than noise
filename	Name of the files
annot	annotation file
col	which column should be considered to annotate the features
path	where to put the output files
Relaxed	If Relaxed then one allows sample to be as the control and for all the others in one cluster to be going in the same direction (more than alpha) otherwise all the features must be deviating to be considered a significant feature
q	The sample in one cluster
alpha	cutoff to be considered different than noise inherited by a

Details

Is giving per cluster the features that vary in the same direction

Value

generates a file per cluster of significant features with an annotation

Author(s)

Rachel Jeitziner

Examples

```
##--
library(airway)
data(airway)
airway <- airway[rowSums(assay(airway))>80,]
assay(airway) <- log(assay(airway)+1,2)
ALPHA <- 1
the_experiment <- TTMap::make_matrices(airway,
seq_len(4), seq_len(4) + 4,
rownames(airway), rownames(airway))
TTMAP_part1prime <-TTMap::control_adjustment(
normal.pcl = the_experiment$CTRL,
tumor.pcl = the_experiment$TEST,
normalname = "The_healthy_controls",
dataname = "Effect_of_cancer",
org.directory = tempdir(), e = 0, P = 1.1, B = 0);
Kprime <- 4;
TTMAP_part1_hda <-
TTMap::hyperrectangle_deviation_assessment(x =
TTMAP_part1prime,
k = Kprime,dataname = "Effect_of_cancer",
normalname = "The_healthy_controls");
annot <- c(paste(colnames(
the_experiment$TEST[,-(seq_len(3))]),"Dis", sep = "."),
paste(colnames(the_experiment$CTRL[,
-seq_len(3)]), "Dis", sep = "."))
annot <- cbind(annot, annot)
```

```
rownames(annot)<-annot[, 1]
dd5_sgn_only <-TTMap::generate_mismatch_distance(
TTMAP_part1_hda,
select=rownames(TTMAP_part1_hda$Dc.Dmat), alpha = ALPHA)
TTMAP_part2 <-
TTMap::ttmap(TTMAP_part1_hda, TTMAP_part1_hda$m,
select = rownames(TTMAP_part1_hda$Dc.Dmat), annot,
e = TTMap::calcul_e(dd5_sgn_only, 0.95, TTMAP_part1prime, 1),
filename = "first_comparison", n =  1, dd = dd5_sgn_only)
TTMap::ttmap_sgn_genes(TTMAP_part2, TTMAP_part1_hda,
TTMAP_part1prime, annot,
n = 2, a = 1, filename = "first_list_of_genes",
annot = TTMAP_part1prime$tag.pcl, col = "NAME",
path = getwd(), Relaxed = 1)
```

`write_pcl` *Reading, writing and annotation files*

Description

Reading (`read_pcl`), writing (`write_pcl`) files and annotating matrices (mat2pcl)

Usage

```
mat2pcl(mat, tag)
write_pcl(df, dataname, fileaddress = "")
read_pcl(filename, na.type = "", Nrows = -1,
Comment.char = "", ...)
```

Arguments

`df`	PCL object to be saved
`dataname`	Name of the file
`fileaddress`	Where to save the file
`filename`	File name to be loaded on R
`na.type`	feels the parameter na.strings of read.table
`Nrows`	Number of rows to be ignored (nrows of read.table)
`Comment.char`	comment.char of read.table
`...`	other read.table arguments
`mat`	matrix to be changed in annotated
`tag`	annotation

Details

The file (called filename) MUST contain 3 columns before the actual values, which are called CLID, NAME and GWEIGHT, described bellow. The first row must be the header of the columns (starting with CLID,NAME and GWEIGHT) and the second row must be EWEIGHT. Representing how much weight each column has: if some columns are n replicates they can have each a weight of 1/n.

Value

Data frame composed of

CLID	Column called CLID which is the ID of the features, which will then be the rownames of the dataframe
NAME	A possibly longer name, more meaningfull than CLID, text format
GWEIGHT	A weight for each gene or feature. If some genes are less important than others or only a pathway should be selected than the file (called filename) should have this information
Matrix	The matrix with numbers of the different observations

Author(s)

Rachel Jeitziner

See Also

control_adjustment

Examples

```
library(airway)
data(airway)
airway <- airway[rowSums(assay(airway))>80,]
assay(airway) <- log(assay(airway)+1,2)
ALPHA <- 1
to_be_saved <- TTMap::make_matrices(airway,
seq_len(4), seq_len(4) + 4,
rownames(airway), rownames(airway))
TTMap::write_pcl(to_be_saved, "tempfile()", getwd())
```

Index

Two-Tier Mapper: a user-independent clustering method for global gene expression analysis based on topology

Rachel Jeitziner
UPBRI& ISREC
EPFL Lausanne

Contents

1 Introduction

We developed a new user-independent analytical framework, called *Two-Tier Mapper (TTMap)*. This tool is separated into two parts. TTMap consists of two separated and independent parts : 1. Hyperrectangle Deviation assessment (HDA) and 2. Global-to-Local Mapper (GtLMap), where the first step establishes properties of the control group and removes outliers in order to calculate the deviation of each vector in the test group from the corrected control group. The second step uses the traditional Mapper algorithm [1] with a two-tier cover and a special distance. This topological tool detects both global and local differences in the patterns of deviations and thereby captures the structure of the test group. The samples are clustered according to the shape of their deviation (do they both deviate positively, negatively or are they as the control). To still keep on the information about the amount of deviation, one separates the data into 4 clusters according to a function measuring the amount of deviation. These represent then the second tier. Each cluster is colored by the extent of the deviation. A list of the differentially expressed genes is also provided. The

functions and methods presented on this *vignette* provide explanation on how to use TTMap, by default and what can be changed by the user.

2 Prepare the data

Upload the file(s) to compare in R. Log transform and subselect them. Use *make_matrices* to create the needed files for the first function of TTMap since it generates the control and the test matrix in the right format. As an example, we use the airway data set available at Bioconductor.

```
> library(airway)
> data(airway)
> dw <- rowSums(assay(airway))>80
> dw <- names(dw[dw==TRUE])
> airway <- airway[dw,]
> assay(airway) <- log(assay(airway)+1,2)
> experiment <- TTMap::make_matrices(airway,seq_len(4),seq_len(4)+4,
+ NAME = rownames(airway), CLID =rownames(airway))
```

This function can directly be used on a normalised count table from RNA- seq precising what are the columns of the control group (in col_ctrl) and what are the columns in the test group (in col_test) .

3 TTMap part1: Adjustement of the control group (ctrl_adj)

The first part of the method checks if the control and the test matrices have the same row-names, and if not the method subselects the common rows. It outputs the files with the common rows subselected (with the extension mesh). It then calculates the corrected control matrix, which removes outliers and replaces them by a chosen method (given by a function with input the matrix with NAs where there is an outlier and should return a matrix without NAs), or by the median of the other values by default. The inputs can even be given by the CTRL and TEST variables of the list given by the output of *make_matrices* or by imputed control and test matrices in pcl format (see [2]). The name of the control group and the project name need to be inputed as well as the working directory, in which the output files will be created. A value for what to consider as an outlier (called e) can be imputed or use the data-driven default value given by the method. If there are any batch effects to consider, they can be imputed using the variable B, which is a vector of numbers representing the batches. Last, the parameter P is a value which will remove the genes that have a higher percentage than P of outlier values.

```
> E=1
> Pw=1.1
> Bw=0
> TTMAP_part1prime <-TTMap::control_adjustment(normal.pcl = experiment$CTRL,
+ tumor.pcl = experiment$TEST,
+ normalname = "The_healthy_controls", dataname = "The_effect_of_cancer",
+ org.directory = getwd(), e=E,P=Pw,B=Bw);
```

This outputs:

- A file with the number of outliers per sample (Dataname followed by the number of the batch followed by na_numbers_per_col.txt)
- A file with the number of outliers per row (Dataname followed by the number of the batch followed by na_numbers_per_col.txt)
- A picture of the distribution of the mean against variance for each gene, before (Dataname followed by _mean_vs_variance.pdf) and
- after correction of outliers (Dataname followed by _mean_vs_variance_after_correction.pdf).

The corrected control matrix is output in the next step. A possible output after this first step is shown in figure 2

4 TTMap part1: Hyperrectangle deviation assessement (hda)

This part consists in calculating deviation components from a hyperrectangle. This enables the calculation in the third function (ttmap_part2_gtlmap) of the shape of deviation. One parameter k is given by if all the vectors of the control group should be kept or if only the the top k-dimensional principal component approximation of the control matrix should be kept using the singular value decomposition (as in [2]). The default is to keep all the vectors.

```
> TTMAP_part1_hda <- TTMap::hyperrectangle_deviation_assessment(x =
+ TTMAP_part1prime,k=dim(TTMAP_part1prime$Normal.mat)[2],
+ dataname = "The_effect_of_cancer", normalname = "The_healthy_controls");
> head(TTMAP_part1_hda$Dc.Dmat)
```

	SRR1039516.Dis	SRR1039517.Dis	SRR1039520.Dis	SRR1039521.Dis
ENSG00000000003	2.5977598	3.540946	-5.017074	-2.8395353
ENSG00000000419	5.4810747	5.552852	-3.522795	-0.5227945
ENSG00000000457	2.8373989	5.467910	-3.973458	-2.6042244
ENSG00000000460	2.9880457	4.234311	-2.688056	-0.7705182
ENSG00000000971	0.2050992	3.742591	-3.976432	-1.3737670
ENSG00000001036	1.2333603	-2.783987	-4.568258	-2.3631438

The outputs of this step are the following.

- The corrected control matrix, calculated at the first step is given in *The_healthy_controls.NormalModel.pcl*, with a possible trimming of columns if k is different than the number of columns in the corrected matrix.
- The deviation component of each test sample is written in *The_effect_of_cancer.Tdis.pcl*. An example of the deviation component is found in the previous script by writing *head(TTMAP_part1_hda$Dc.*
- The normal component of each test sample is written in *The_effect_of_cancer.Tnorm.pcl*.

The two values of this function are the deviation component matrix and the overall deviation (calculated by summing in absolute values the deviation components).

5 TTMap part2: Global-to-local Mapper (gtlmap)

The third part corresponds to the Global-to-local Mapper part. One starts with an annotation file of our samples, in order to annotate the obtained clusters. In this example here we just copied several times the column names. This annotation file needs to have as rownames the columns of the test samples followed by ".Dis". We then calculate the distance matrix between the samples using the *generate_mismatch_distance* function, which uses a cutoff parameter α in order to decide what is a considered as noise. Any other distance matrix can be computed here and used for the next step. Then, we calculate and output the clusters using *ttmap_part2_gtlmap*, which needs as inputs the values of *ttmap_part1_ctrl_adj, ttmap_part1_hda*. The default parameter uses all the genes to calculate the overall deviation, but if a subset should be selected (only one pathway for example), it can be imputed here. ttmap_part2_gtlmap then calculates using *calcul_e* a parameter of closeness using the data, in order to know what distance is "close" enough to clusters samples together. The parameter n determines which column of metadata should be chosen for the output files. Two more parameters of convenience, if ad is set to something different than 0 (the default) then the clusters

on the output picture will not be annotated and if bd is different than 0 (default), the output will be without outliers of the test data set. After the picture has been adjusted to what one wants to see one can save it using the *rgl.postscript* function.

```
> library(rgl)
> ALPHA <- 1
> annot <- c(paste(colnames(experiment$TEST[,-seq_len(3)]),"Dis",sep=".")
+ ,paste(colnames(experiment$CTRL[,-seq_len(3)]),"Dis",sep="."))
> annot <- cbind(annot,annot)
> rownames(annot)<-annot[,1]
> dd5_sgn_only <-TTMap::generate_mismatch_distance(TTMAP_part1_hda,
+ select=rownames(TTMAP_part1_hda$Dc.Dmat),alpha = ALPHA)
> TTMAP_part2_gtlmap <-
+ TTMap::ttmap(TTMAP_part1_hda,TTMAP_part1_hda$m,
+ select=rownames(TTMAP_part1_hda$Dc.Dmat),
+ annot,e= TTMap::calcul_e(dd5_sgn_only,0.95,TTMAP_part1prime,1),
+ filename="first_comparison",
+ n=1,dd=dd5_sgn_only)

[1] "e_map = 0.278978511367175"
[1] "e_map = 0.278978511367175"
[1] "e_map = 0.278978511367175"
[1] "e_map = 0.278978511367175"
[1] "e_map = 0.278978511367175"

> rgl.postscript("first_output.pdf","pdf")
```

6 TTMap: Finding the significant genes (sgn_genes)

This last function analyses the different clusters for significant features. It outputs a file per level (one for overall, called all, one for the lower quartile, called low, one for the second quartile, called mid1, the third, mid2, and the higher quartile, called high). In each of them one file per cluster is given, with the list of significant genes linked to the cluster. Relaxed is a parameter permitting to select as a match one sample that would be 0 for the deviation component, while the others deviate in the same shape.

```
> TTMap::ttmap_sgn_genes(TTMAP_part2_gtlmap,
+ TTMAP_part1_hda, TTMAP_part1prime,
+ annot, n = 2, a = ALPHA,
+ filename = "first_trial", annot = TTMAP_part1prime$tag.pcl, col = "NAME",
+ path = getwd(), Relaxed = 0)
```

7 Conclusion

Two-Tier Mapper (TTMap) is a topology-based clustering tool, which is user- friendly and reliable. The algorithm first provides an overall clustering, in an unbiased manner, since all the parameters are defined in a data-driven manner or by reliable default parameters. his method enables a refined view on the composition of the clusters by delineating how clusters differ locally and how the local clusters relate to the global structure of the dataset. The output is a visual interpretation of the data given by a colored graph that is easy to interpret, which describes the shape of the data according to the chosen distance.

References

[1] P. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Vejdemo-Johansson, M. Alagappan, J. Carlsson, and G. Carlsson. Extracting insights from the shape of complex data using topology. *Scientific Reports*, 3, 2013.

[2] Monica Nicolau, Arnold Levine, and Gunnar Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Science*, 108(17):7265–7270, 2011.

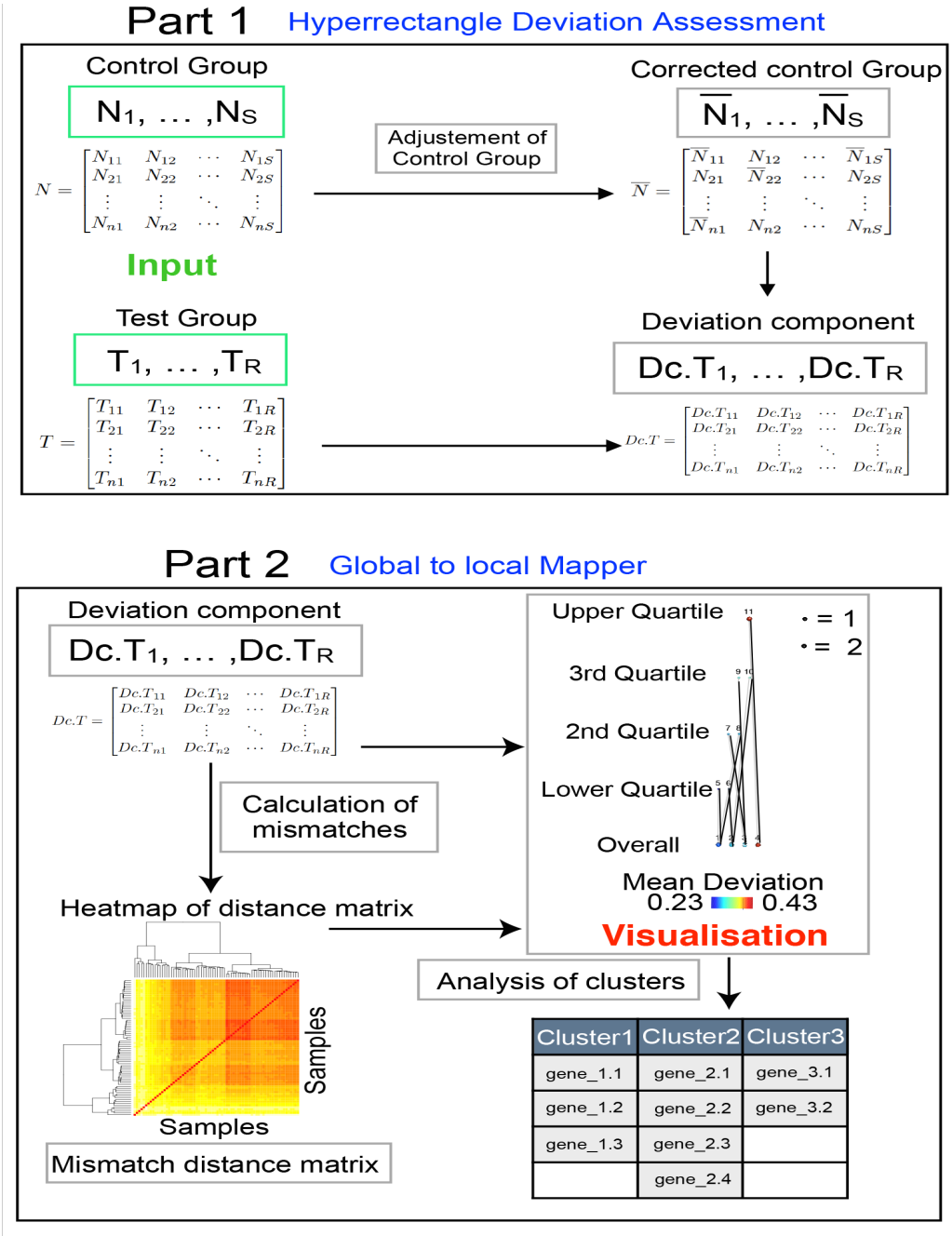


Figure 1: Schematic overview of TTMap. The inputs (green) are given by two gene expression matrices, the control (N) and the test group (T), rows represent genes and columns samples. In Part 1, TTMap adjusts the control group for outlier values ($\bar{N}_*$), feature by feature. It calculates deviation from this corrected control group for individual samples in the test group ($Dc.T_*$). In Part 2, TTMap computes a similarity measure, the mismatch distance (represented as a heatmap) using the deviation components. The Mapper [**?**] algorithm is used with a two-tier cover to generate a visual representation of the clustering creating a network of global clusters (Overall) and local clusters (1st, 2nd, 3rd, 4th quartile of a filter function). It takes as inputs the mismatch distance and the deviation components.

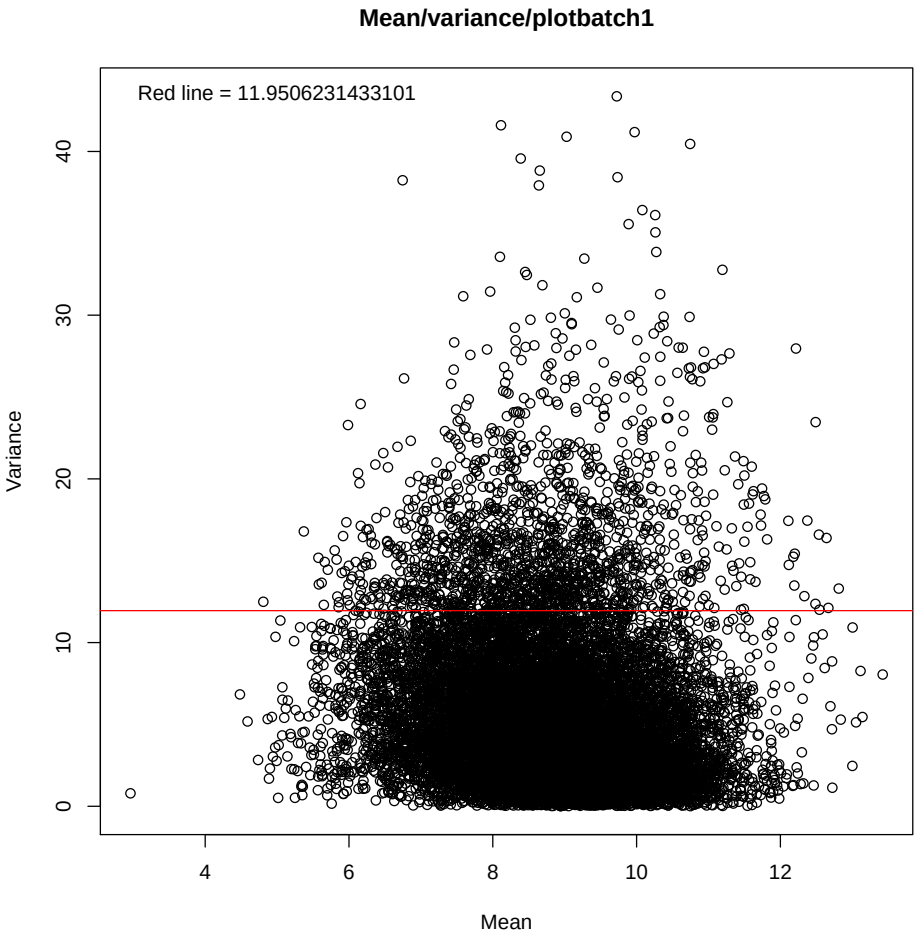


Figure 2: `barplotSignifSignatures`: Plot of mean against variance per gene.

B Appendix: Additional Theory

B.0.1 Multidimensional size functions for shape comparison

This section is an introduction to multidimensional size functions and how to compare them. We are starting to use these concepts 6.1.

Goal. *The major goal is comparing several descriptions of a space $\mathscr{M}$ expressed in terms of certain functions in order to understand which functions provide similar information on the space. Therefore, given two functions f, g from a space $\mathscr{M}$ to $\mathbb{R}^n$, a possibility would be to calculate*

$$\|f - g\|_\infty .$$

However, for some application and in particular once $\mathscr{M}$ is the space S^1, calculating the infinity norm is not sufficient to compare the shape descriptors. Indeed, two of them could be the same up to a certain delay, i.e., $f(e^{i2\pi t}) = g(e^{i2\pi(t+m)})$, where $m \in \mathbb{R}$ is the shift. In that case, the goal becomes to compare the functions f and g by calculating

$$\|f - g \circ h\|_\infty ,$$

where $h : S^1 \to S^1$ is a diffeomorphism (a differentiable bijection, which in the case of study of the menstrual cycle takes into account delays, see section 6.1). The solution is however not computable. Indeed, in order to compare two functions, f and g, infinitely many diffeomorphisms need to be computed and each time the ∞-norm calculated, which is not achievable in a finite time.

The complexity of the problem should be reduced. The idea is to calculate a distance, for which the following statement is true : for every $\varepsilon > 0$, there exists a $\delta > 0$ such that if two functions are at a computable distance δ from each other, there exists a diffeomorphism $h : S^1 \to S^1$ as described earlier, which would verify that

$$\|f - g \circ h\|_\infty < \varepsilon.$$

Therefore, understanding the problem would be reduced to calculating a distance, which should lower the complexity of the computations.

This section is a compilation of articles [248], [249] and [31]. All the proofs of the theorems can be found in [248]. In order to understand the shape of a space, a theory emerged in the early 90s that describes the space using a size function. These functions are used in order to

gain information about the space. The final result, theorem B.0.38, enables the classification of size functions by calculating only a certain type of homology groups. This lowers the complexity of the problem since it is easier to calculate the rank of those groups rather than directly comparing the curves.

B.0.2 1-dimensional Size Theory

Size theory is used to understand shapes, by using functions from the space to a reference space, which is easier to understand, where a specific example is $\mathbb{R}^k$, and $k \in \mathbb{N}^*$ is called the dimension. The 1-dimensional size theory refers to the study of space descriptors that are continuous function from the space to $\mathbb{R}$, whereas k-dimensional size theory will analyse continuous function from the space to $\mathbb{R}^k$.

For the whole section $\mathscr{M}$ will be a compact locally connected Hausdorff space (see [1] for definitions).

Intuition B.0.1. Space descriptors are functions on the space and have the same role as filter functions (see section 1.2.9).

Definitions

Definition B.0.2. A continuous function $\varphi : \mathscr{M} \to \mathbb{R}^k$ describing a feature of interest is called a **k-dimensional measuring function**.

Definition B.0.3. Let $\varphi : \mathscr{M} \to \mathbb{R}^k$ and $\psi : \mathscr{N} \to \mathbb{R}^k$ be two k-dimensional measuring functions. Let H be the set of all homeomorphisms between $\mathscr{M}$ and $\mathscr{N}$. The **natural pseudo-distance** between φ and ψ, is defined as

$$d((\mathscr{M},\varphi),(\mathscr{N},\psi)) = \inf_{f\in HP\in\mathscr{M}} \max \|\varphi(P) - \psi(f(P))\|_\infty .$$

Intuition B.0.4. If $\mathscr{M} = \mathscr{N}$, then the problem becomes finding out how close the two descriptors are. Hence, the user will find out which descriptions of a space give different types of information (those will be discovered when the distance is non zero). For instance, since the weight and the diameter of a tumor are gradually increasing together, it is expected that the natural pseudo-distance between the two functions is small.

Definition B.0.5. Let $\varphi : \mathscr{M} \to \mathbb{R}$ be a measuring function. The **lower level sets** are defined for $x \in \mathbb{R}$ by

$$\mathscr{M}\langle\varphi \leq x\rangle = \{P \in \mathscr{M} \mid \varphi(P) \leq x\}.$$

Definition B.0.6. Let $y \in \mathbb{R}$. Two points $P, Q \in \mathscr{M}$ are said to be $\langle\varphi \leq y\rangle$**-connected** if and only if a connected subset of $\mathscr{M}\langle\varphi \leq y\rangle$ exists containing P and Q.

It is easy to verify that this defines an equivalence relation.

Example B.0.7. In the following example (Supplementary Fig. S1) $\varphi : \mathscr{M} \to \mathbb{R}$ is defined as the distance to the point O. The points P, Q are not $\langle\varphi \leq a + \varepsilon\rangle$-connected as long as ε is smaller than $(b - a)$ (at b the two green components merge and at that moment P and Q are in the same component).

Definition B.0.8. The pair $(\mathscr{M},\varphi)$ gives rise to a 1**-dimensional size function** defined as

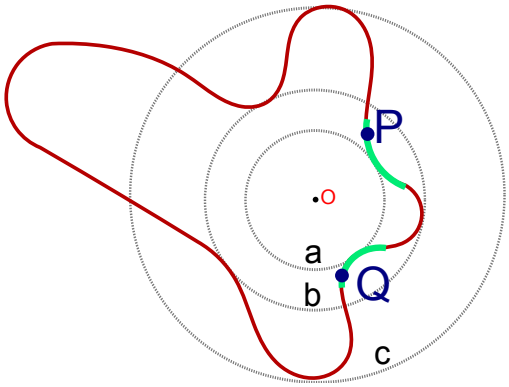


Figure S1: Illustration of a measuring function given by the distance to O and the equivalence relation $\langle \varphi \leq a + \varepsilon \rangle$-connectedness.

$$l_{(\mathscr{M},\varphi)} : \{(x, y) \mid x < y\} \to \mathbb{N},$$

where $l_{(\mathscr{M},\varphi)}(x, y)$ is equal to the number of equivalence classes in which the set $\mathscr{M}\langle \varphi \leq x \rangle$ is divided by the $\langle \varphi \leq y \rangle$-connectedness relation.

Example B.0.9. Continuing the example B.0.7,if $a \leq x < y < b$, then $l_{(\mathscr{M},\varphi)}(x, y) = 2$. Indeed, the two component (shown in green) cannot be joined under y. At the point b however, they join, hence $l_{(\mathscr{M},\varphi)}(x, z) = 1$, for $z \geq b$. The size function is represented in Supplementary Fig. S2. Another way to see $l_{(\mathscr{M},\varphi)}(x, y)$ is by observing that it corresponds to the number of connected component of $\mathscr{M}\langle \varphi \leq y \rangle$ that have at least one point in $\mathscr{M}\langle \varphi \leq x \rangle$.

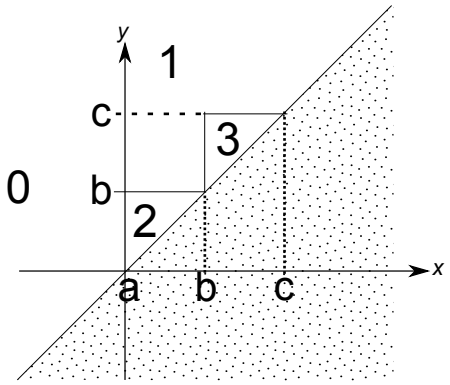


Figure S2: Illustration of the 1-dimensional size function corresponding to the pair $(\mathscr{M}, \varphi)$ defined in Supplementary Fig. S1.

Describing Size functions with cornerpoints and cornerlines

The information contained in the graph in Supplementary Fig. S2 can be summarised using only certain points and lines in this plot and is described in this paragraph.

Definition B.0.10. For every vertical line r, with equation $x = k$, we define the **multiplicity** of r, denoted $\mu(r)$ as the minimum, over all the positive real numbers ε verifying $k + \varepsilon < 1/\varepsilon$, of

$$l_{(\mathscr{M},\varphi)}(k + \varepsilon, 1/\varepsilon) - l_{(\mathscr{M},\varphi)}(k - \varepsilon, 1/\varepsilon).$$

When the multiplicity of r is a strictly positive number, the line r is called a **cornerline** for the size function.

Example B.0.11. Continuing the example B.0.7, for the vertical line $x = 0$ and any $\varepsilon > 0$ such that $k + \varepsilon < 1/\varepsilon$, we obtain that

$$l_{(\mathscr{M},\varphi)}(k+\varepsilon, 1/\varepsilon) - l_{(\mathscr{M},\varphi)}(k-\varepsilon, 1/\varepsilon),$$

is equal to either 1 or 2, which is bigger than 0 (Supplementary Fig. S3). Hence, the line $x = 0$ is a cornerline.

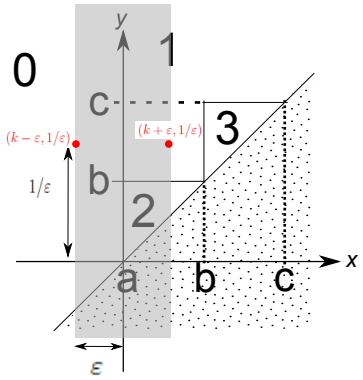


Figure S3: Illustration of the cornerline $x = 0$ and the calculation for a given ε of the multiplicity of the line $x = 0$ to prove that it is a cornerline.

Definition B.0.12. For every point $p = (x, y)$, with $x < y$, we define the **multiplicity** of p, denoted $\mu(p)$ as the minimum, over all the positive real numbers ε verifying $x + \varepsilon < y - \varepsilon$, of

$$l_{(\mathscr{M},\varphi)}(x+\varepsilon, y-\varepsilon) - l_{(\mathscr{M},\varphi)}(x-\varepsilon, y-\varepsilon) - l_{(\mathscr{M},\varphi)}(x+\varepsilon, y+\varepsilon) + l_{(\mathscr{M},\varphi)}(x-\varepsilon, y+\varepsilon).$$

When the multiplicity of p is a strictly positive number, the point p is called a **cornerpoint** for the size function.

Remark B.0.13. There is another similar theory, called **persistent homology** (section 1.3) that describes when a new component is "born" and when this component dies (finally merged with another component). Persistent diagrams correspond to studying the cornerpoints of the 1-dimensional size function, which are given by (a, b) where a denotes when a component was born, and b when it died.

Example B.0.14. Still continuing example B.0.7, considering the point $p = (b, c)$, for any $\varepsilon > 0$ such that $a + \varepsilon < b - \varepsilon$, we obtain that

$$l_{(\mathscr{M},\varphi)}(b+\varepsilon, c-\varepsilon) - l_{(\mathscr{M},\varphi)}(b-\varepsilon, c-\varepsilon) - l_{(\mathscr{M},\varphi)}(b+\varepsilon, c+\varepsilon) + l_{(\mathscr{M},\varphi)}(b-\varepsilon, c+\varepsilon)$$

is equal to 2, which is bigger than 0 (Supplementary Fig. S4). Hence, the point $p = (b, c)$ is a cornerpoint.

Representation theorem

We are now able to link the multiplicity of the points and the size function.

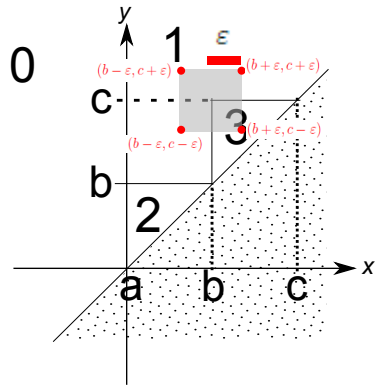


Figure S4: Illustration of the cornerpoint $p = (b, c)$.

Theorem B.0.15 (Representation theorem)**.** *Let $l_{(M,\varphi)}$ be a size function. Then for every $(\bar{x}, \bar{y}) \in \mathbb{R}^2$, with $\bar{x} < \bar{y} < \infty$, we have that*

$$l_{(M,\varphi)}(\bar{x}, \bar{y}) = \sum_{\substack{(x,y):x<y\leq\infty \\ x\leq\bar{x}, y\geq\bar{y}}} \mu(x, y).$$

Example B.0.16. Back to example B.0.7 (Supplementary Fig. S5) reflects that the representation theorem is working since the green node, which has size function equal to 3 (as can be read from the graph) and the sum of the multiplicities that lie in the shaded grey area corresponds to the multiplicity of the cornerpoint (b, c) $\mu(b, c) = 2$ and the multiplicity of the cornerline $x = a$ which is $\mu(a) = 1$ coincide.

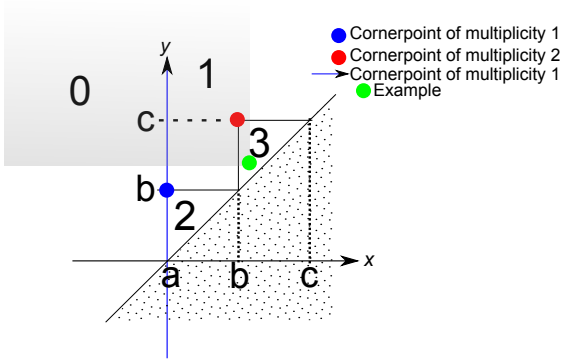


Figure S5: Illustration of theorem B.0.15.

Remark B.0.17. The representation theorem B.0.15 states that the size function at a specific point is equal to the multiplicity of the cornerpoints and cornerline on its upper-left quadrant. In other words two size function can be compared through their cornerpoints and lines, an intuition that will be justified in the next paragraph.

Comparing Size functions

In the previous section, the study of size function was reduced to the comparison of their cornerpoints and cornerlines. Let l_1, l_2 be two size functions. We define the associated sets C_1 (respectively C_2) to be the multiset of all cornerpoints taken with their multiplicity and

cornerlines of l_1 (respectively l_2) where infinitely many points are added from the diagonal $\{(x, y) \in \mathbb{R}^2 \mid x = y\}$ and where cornerlines $x = k$ are represented by points (k, ∞).

Definition B.0.18. Let l_1, l_2 be two size functions and C_1 and C_2 their associated multisets as defined above. Let B be the set of all bijections between C_1 and C_2. The **matching distance** between size functions arising from 1-dimensional measuring functions is given by

$$d_{match}(l_1, l_2) = \min_{\sigma \in B} \max_{p \in C_1} \delta(p, \sigma(p)),$$

where

$$\delta((x, y), (x', y')) = \min \left\{ \max\{| x - x' |, | y - y' |\}, \max \left\{ \frac{y - x}{2}, \frac{y' - x'}{2} \right\} \right\}.$$

This is calculated using the convention that $\infty - \infty = 0$, $\infty - x = x - \infty = \infty$ for all $x \neq \infty$, $\infty / 2 = \infty$, $| \infty | = \infty$, $\min(c, \infty) = c$, $\max(c, \infty) = \infty$.

Example B.0.19. In Supplementary Fig. S6, we illustrate two size functions represented by their cornerpoints and cornerlines. The multiplicities different from one are written above the points. In the lower panel, an explanation of how the distance between the two size functions are calculated is illustrated.

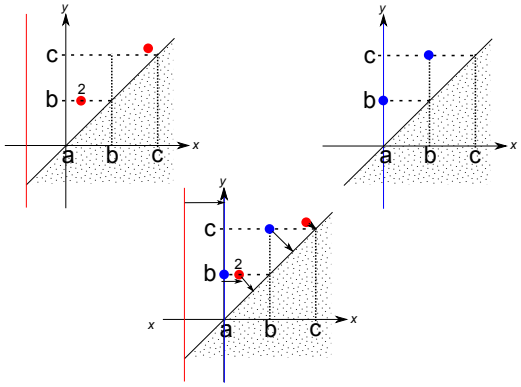


Figure S6: Illustration of the calculation of the matching distance.

Theorem B.0.20. *Let $(\mathscr{M}, \varphi)$ and $(\mathscr{N}, \psi)$ be size pairs. Then,*

$$d_{match}(l_{(\mathscr{M},\varphi)}, l_{(\mathscr{N},\psi)}) \leq d((\mathscr{M}, \varphi), (\mathscr{N}, \psi)).$$

The matching distance defined above is the best possible approximation to the natural pseudo-distance, using size functions.

Theorem B.0.21. *If d' is another distance on size functions verifying $d'(l_{(\mathscr{M},\varphi)}, l_{(\mathscr{N},\psi)}) \leq d((\mathscr{M}, \varphi), (\mathscr{N}, \psi))$, then*

$$d'(l_{(\mathscr{M},\varphi)}, l_{(\mathscr{N},\psi)}) \leq d_{match}(l_{(\mathscr{M},\varphi)}, l_{(\mathscr{N},\psi)}).$$

Concluding remarks B.0.22. This concludes the 1-dimensional case. Indeed, d_{match} is a computable distance that is the best lower bound using size functions.

B.0.3 k-dimensional Size Theory

The question of how to compare k-dimensional size functions is of greater difficulty. Indeed, cornerpoints and cornerlines as defined in the previous section do not have a direct analogue in the k-dimensional case. The task here will be to reduce the comparison of k-dimensional size functions to the comparison of corresponding 1-dimensional functions.

Definition B.0.23. A pair $(l, b) \in \mathbb{R}^k \times \mathbb{R}^k$ is **linearly admissible** if $l = (l_1, \dots, l_k)$ verifies $\sum_{i=1}^k l_i = 1$ where $l_i > 0$ for $i = 1, \dots k$, and $b = (b_1, \dots, b_k)$ in $\mathbb{R}^k$ verifies $\sum_{i=1}^k b_i = 0$.

Notation. We denote the set of admissible pairs in $\mathbb{R}^k \times \mathbb{R}^k$ by Adm_k.

Example B.0.24. If $(l, b) \in \mathbb{R}^2 \times \mathbb{R}^2$ is an admissible pair, then, by definition B.0.23, $l \in \mathbb{R}^2$ and verifies $l_1 + l_2 = 1$. In addition, l verifies $l_i > 0$. Hence, $l = (l_1, 1 - l_1) \in (0, 1) \times (0, 1)$. And $b = (b_1, b_2) \in \mathbb{R}^2$ is given by $b_1 + b_2 = 0$, which then corresponds to all the $b = (b_1, -b_1) \in \mathbb{R}^2$.

Definition B.0.25. For every linearly admissible pair (l, b), the **associated half-plane** to this linearly admissible pair, written $\pi_{(l,b)} \in \mathbb{R}^2 \times \mathbb{R}^2$, is determined by the equations $u = \sigma l + b$ and $v = \tau l + b$ with $\sigma, \tau \in \mathbb{R}$ and $\sigma < \tau$.

The following theorem explains the usefulness of admissible pairs.

Theorem B.0.26. *For every (u, v) such that $u < v$, there exists one and only one linearly admissible pair (l, b) such that $(u, v) \in \pi_{(l,b)}$.*

The following statement explains the construction of a 1-dimensional size function for every admissible pair which is consistent with the k-dimension size function.

Theorem B.0.27. *Let $(l, b) \in Adm_k$, and $\varphi = (\varphi_1, \dots, \varphi_k)$ a k-dimensional size function and let $F^{\varphi}_{(l,b)} : \mathscr{M} \to \mathbb{R}$ be the function given by*

$$F^{\varphi}_{(l,b)}(x) = \max_{i=1,\dots,k} \left\{ \frac{\varphi_i(x) - b_i}{l_i} \right\}.$$

Then, for every $(u, v) = (\sigma l + b, \tau l + b) \in \pi_{(l,b)}$, we get $l_{(\mathscr{M},\varphi)}(u, v) = l_{\left(\mathscr{M},F^{\varphi}_{(l,b)}\right)}(\sigma, \tau)$.

Hence the 1-dimensional size functions $l_{(\mathscr{M},F^{\varphi}_{(l,b)})}$ for (l, b) varying through Adm_k is characterize completely the k-dimensional size function $l_{(\mathscr{M},\varphi)}$. In addition, the following theorems clarify the relation between the original natural pseudo-distance between size pairs $(\mathscr{M}, \varphi)$ and $(\mathscr{N}, \psi)$ and the matching distance on $l_{(\mathscr{M},F^{\varphi}_{(l,b)})}$ and $l_{(\mathscr{N},F^{\psi}_{(l,b)})}$.

Theorem B.0.28. *Let $(\mathscr{M}, \varphi)$ and $(\mathscr{M}, \psi)$ be two size pairs, on the same space. Then*

$$d_{match}(l_{(\mathscr{M},F^{\varphi}_{(l,b)})}, l_{(\mathscr{M},F^{\psi}_{(l,b)})} \leq \max_{P \in \mathscr{M}} \left\{ \frac{|| \varphi(P) - \psi(P) ||_\infty}{\min_{i=1,\dots,k} l_i} \right\}.$$

Theorem B.0.29. *Let $(\mathscr{M}, \varphi)$ be a size pair. Moreover, let $(l, b) \in Adm_k$ be an admissible pair and $\varepsilon > 0$ a real number smaller than $\min_{i=1,\dots,k} l_i$. Then, for every admissible pair (l', b') that verifies $|| (l, b) - (l'b') ||_\infty \leq \varepsilon$, it holds that*

$$d_{match}(l_{(\mathscr{M},F^{\varphi}_{(l,b)})}, l_{(\mathscr{M},F^{\varphi}_{(l',b')})} \leq \varepsilon \cdot \frac{\max_{x \in \mathscr{M}} \left\{ \frac{||\varphi(P)||_\infty + ||l||_\infty + ||b||_\infty}{\min_{i=1,\dots,k} l_i} \right\}}{\min_{i=1,\dots,k} \{l_i(l_i - \varepsilon)\}}.$$

Definition B.0.30. Let $(\mathscr{M}, \varphi)$ and $(\mathscr{N}, \psi)$ be two size pairs with φ and ψ taking values in $\mathbb{R}^k$. Then we define the **k-dimensional matching distance** by

$$D_{match}(l_{(\mathscr{M},\varphi)}, l_{(\mathscr{N},\psi)}) = \sup_{(l,b)\in Adm_k} \min_{i=1,\dots,k} l_i d_{match}\left(l_{\left(\mathscr{M},F^{\varphi}_{(l,b)}\right)}, l_{\left(\mathscr{N},F^{\psi}_{(l,b)}\right)}\right).$$

Remark B.0.31.

- We therefore replaced a not computable distance with another distance that is again not computable.

- As shown in the next pages, this is not problematic, since there is a good approximation of this distance in the 2-dimensional case. Indeed, for every $\varepsilon > 0$ an algorithm constructing a *finite* set $A \subseteq Adm_k$ such that

$$\tilde{D}_{match}(l_{(\mathscr{M},\varphi)}, l_{(\mathscr{N},\psi)}) = \max_{(l,b)\in Ai=1,\dots,k} \min l_i d_{match}\left(l_{\left(\mathscr{M},F^{\varphi}_{(l,b)}\right)}, l_{\left(\mathscr{N},F^{\psi}_{(l,b)}\right)}\right)$$

verifies

$$|D_{match}(l_{(\mathscr{M},\varphi)}, l_{(\mathscr{N},\psi)}) - \tilde{D}_{match}(l_{(\mathscr{M},\varphi)}, l_{(\mathscr{N},\psi)})| < \varepsilon,$$

can be found.

B.0.4 Algorithm to approximate the 2-dimensional matching distance

This algorithm was described in [249] and implemented in [250]. In order to explain how the algorithm works, some useful constants are needed. First, for every couple of 2-dimensional size pairs $(\mathscr{M}, \varphi)$ and $(\mathscr{N}, \psi)$, where $\varphi = (\varphi_1, \varphi_2)$ et $\psi = (\psi_1, \psi_2)$, let C be a constant defined by

$$C = \max\left\{\max_{x\in\mathscr{M}} \max\{|\ \varphi_1(x)\ |, |\ \varphi_2(x)\ |\}, \max_{y\in\mathscr{N}} \max\{|\ \psi_1(y)\ |, |\ \psi_2(y)\ |\}\right\}.$$

Then, for every linearly admissible pair $(l, b) = (l_1, 1 - l_1, b_1, -b_1)$ let $m(l)$ be the constant defined by $m(l) = \min\{l_1, 1 - l_1\}$.

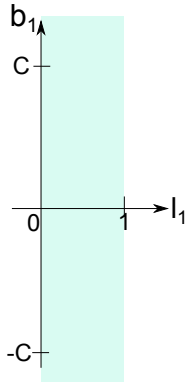


Figure S7: Possible region for b_1 and l_1 such that $(l, b) = (l_1, 1 - l_1, b_1, -b_1)$ is linearly admissible.

Theorem B.0.32. *Let $(l, b) = (l_1, 1 - l_1, b_1, -b_1)$ be an admissible pair. Let $(\mathscr{M}, \varphi)$ and $(\mathscr{N}, \psi)$ be two 2-dimensional size pairs, where $\varphi = (\varphi_1, \varphi_2)$ et $\psi = (\psi_1, \psi_2)$. If $|b_1| \geq C$, then*

it follows that

$$m(l)d_{match}\Big(l_{\left(\mathscr{M},F^{\varphi}_{(l,b)}\right)},l_{\left(\mathscr{N},F^{\psi}_{(l,b)}\right)}\Big)=\begin{cases}\frac{m(l)}{l_1}d_{match}(l_{(\mathscr{M},\varphi_1)},l_{(\mathscr{N},\psi_1)}) & \text{if } b_1\leq -C\\ \frac{m(l)}{1-l_1}d_{match}(l_{(\mathscr{M},\varphi_2)},l_{(\mathscr{N},\psi_2)}) & \text{if } b_1\geq C.\end{cases}$$

Remark B.0.33. Since the 2-dimensional matching distance is searching for a supremum, only the highest possible value that $m(l)d_{match}\Big(l_{\left(\mathscr{M},F^{\varphi}_{(l,b)}\right)},l_{\left(\mathscr{N},F^{\psi}_{(l,b)}\right)}\Big)$ could reach is of interest. Therefore, looking at theorem B.0.32, the supremum restricted to when b_1 is smaller than $-C$ and when b_1 is greater than C is easily computable. In those cases, the function does not depend any more on b_1. Therefore, it makes sense to compute out the supremum of this function on all (l,b) where $b_1\leq -C$ and where $b_1\geq C$.

In the former case, we need to determine when $\frac{m(l)}{l_1}$ is maximal. The following calculation

$$\max_{l_1\in[0,1]}\frac{m(l)}{l_1}=\max\{\max_{l_1\leq\frac{1}{2}}\frac{l_1}{l_1},\max_{l_1\geq\frac{1}{2}}\frac{1-l_1}{l_1}\}=\max\{1,\max_{l_1\geq\frac{1}{2}}\frac{1-l_1}{l_1}\}=1,$$

shows that the supremum in this case is reached when l_1 is smaller or equal to $\frac{1}{2}$ (to understand this formula, observe that when $l_1\leq\frac{1}{2}$, then $m(l)=l_1$ and when $l_1\geq\frac{1}{2}$, $m(l)=1-l_1$). This helps us conclude that

$$\sup_{\{(l,b)|b_1\leq -C\}}m(l)d_{match}\Big(l_{\left(\mathscr{M},F^{\varphi}_{(l,b)}\right)},l_{\left(\mathscr{N},F^{\psi}_{(l,b)}\right)}\Big)=d_{match}\Big(l_{(\mathscr{M},\varphi_1)},l_{(\mathscr{N},\psi_1)}\Big).$$

This supremum is attained, for example, for the pair $(l,b)=(\frac{1}{2},\frac{1}{2},-(C+1),(C+1))$.

In the latter case, with the same procedure, we discover that $\frac{m(l)}{1-l_1}$ is maximal when l_1 is greater than or equal to $\frac{1}{2}$. This helps us conclude that

$$\sup_{\{(l,b)|b_1\geq C\}}m(l)d_{match}\Big(l_{\left(\mathscr{M},F^{\varphi}_{(l,b)}\right)},l_{\left(\mathscr{N},F^{\psi}_{(l,b)}\right)}\Big)=d_{match}\Big(l_{(\mathscr{M},\varphi_2)},l_{(\mathscr{N},\psi_2)}\Big).$$

This supremum is attained, for example, for the pair $(l,b)=(\frac{1}{2},\frac{1}{2},(C+1),-(C+1))$.

Hence, the problem is reduced to calculating

$$\begin{aligned}D_{match}(l_{(\mathscr{M},\varphi)},l_{(\mathscr{N},\psi)})=\max\ \Big\{&d_{match}\Big(l_{(\mathscr{M},\varphi_1)},l_{(\mathscr{N},\psi_1)}\Big),d_{match}\Big(l_{(\mathscr{M},\varphi_2)},l_{(\mathscr{N},\psi_2)}\Big),\\ &\max_{(l,b)\in Adm*}m(l)d_{match}\Big(l_{\left(\mathscr{M},F^{\varphi}_{(l,b)}\right)},l_{\left(\mathscr{N},F^{\psi}_{(l,b)}\right)}\Big)\Big\},\end{aligned}$$

where $Adm*=\{(l,b)=(l_1,1-l_1,b_1,-b_1)\in Adm_2\mid b_1\in[-C,C]\}$.

Error Bound Theorem

Lemma B.0.34. *Let $\delta>0$ a positive number. For every linearly admissible pair $(l,b)\in Adm_2^*$ and every linearly admissible pair $(l',b')\in Adm_2$ such that $||\ (l,b)-(l',b')\ ||_\infty\leq\delta$, it follows*

that

$$\left| m(l) d_{match}\left(l_{\left(\mathscr{M}, F^{\varphi}_{(l,b)}\right)}, l_{\left(\mathscr{N}, F^{\psi}_{(l,b)}\right)} \right) - m(l') d_{match}\left(l_{\left(\mathscr{M}, F^{\varphi}_{(l',b')}\right)}, l_{\left(\mathscr{N}, F^{\psi}_{(l',b')}\right)} \right) \right| \leq \delta * (16C + 2).$$

We can state the Error Bound Theorem:

Theorem B.0.35 (Error Bound Theorem)**.** *Let $\delta > 0$ a positive number. For every couple of linearly admissible pairs $(l,b), (l',b') \in Adm_2^*$ such that $|| (l,b) - (l',b') ||_\infty \leq \delta$, it follows that*

$$\left| m(l) d_{match}\left(l_{\left(\mathscr{M}, F^{\varphi}_{(l,b)}\right)}, l_{\left(\mathscr{N}, F^{\psi}_{(l,b)}\right)} \right) - m(l') d_{match}\left(l_{\left(\mathscr{M}, F^{\varphi}_{(l',b')}\right)}, l_{\left(\mathscr{N}, F^{\psi}_{(l',b')}\right)} \right) \right| \leq \delta * (16C + 2).$$

Algorithm

In this paragraph, we enumerate the different steps of the algorithm illustrating the steps of the computation

1. Let $\varepsilon > 0$ be a fixed error threshold.
2. The algorithm starts with a value $\delta = 1/16$. Afterwards, in every iteration, this value changes.
3. Calculate the matching distance for $(l,b) = (\frac{1}{2}, \frac{1}{2}, -(C+1), C+1)$ et $(l,b) = (\frac{1}{2}, \frac{1}{2}, C+1, -(C+1))$ (this gives the maximum for the dotted area in Supplementary Fig. S8). The maximum between these two values is written M_1.

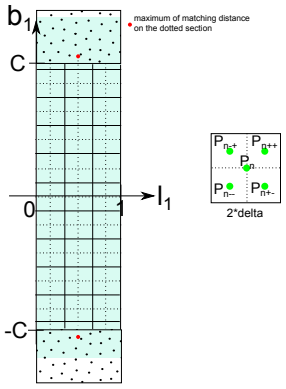


Figure S8: Algorithm illustration to approximate the matching distance.

4. Then, find a set of points in $\mathbb{R}^2$, $\mathscr{P} = \{P_n = (\alpha_n, \beta_n)\}$ such that P_n is the center of the square of side 2δ. The union of all those squares should cover the space $]0,1[\times]-C, C[$. In that way any admissible pair (l,b) verifies that there exists an $n \in \mathbb{N}$ such that $|| (l,b) - (\alpha_n, \beta_n) ||_\infty \leq \delta$.
5. We can calculate the distance for the associated pairs $(l_{P_n}, b_{P_n}) = (\alpha_n, 1-\alpha_n, \beta_n, -\beta_n)$. The maximum of those values is written M_δ.
6. Let $\bar{D}$ be the maximum of the values obtained until now, i.e. $\bar{D} = \max\{M_1, M_\delta\}$.

- If $\delta \cdot (16C + 2) \leq \varepsilon$ holds then the theorem B.0.35 implies that $|D_{match} - \bar{D}| \leq \varepsilon$. The algorithm finishes here.
- Otherwise, the algorithm deletes from $\mathscr{P}$ every point P_n such that

$$\bar{D} - D_{(l_{P_n}, b_{P_n})}(l_{(\mathscr{M},\varphi)}, l_{(\mathscr{N},\psi)}) > \delta \cdot (16C + 2)$$

  and remove also the corresponding square.
- For the remaining points P_n divide the square associated to that point into four sets, and each point P_n is replaced with the four points $P_{n++}, P_{n-+}, P_{n+-}, P_{n--}$ (Supplementary Fig. S9), so δ is replaced by $\delta/2$ and the algorithm can start again.

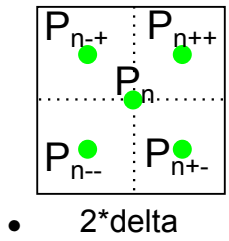


Figure S9: Algorithm illustration to approximate the matching distance.

B.0.5 Comparing generic curves

We summarise the major findings and most useful theorems for our applications, for complete literature see [31], [251] and [249].

A dense subset of C^1 [31] functions is needed for shape comparisons.

Definition B.0.36. A function $f : S^1 \to \mathbb{R}^2$ is called **generic** if it verifies the following properties.

- $f \in C^1$.
- f is an immersion, i.e. $d_\theta f$ has rank equal to one for every $\theta \in S^1$.
- $f(S^1)$ has at most a finite number of multiple points, all of them are double points and $f(\theta_1) = f(\theta_2)$ and $\mathrm{im} d_{\theta_1} f = \mathrm{im} d_{\theta_2} f$ together imply that $\theta_1 = \theta_2$, for every $\theta_1, \theta_2 \in S^1$.

This category of functions must be further subselected in order to prove theorem B.0.38.

Definition B.0.37. Fix $k > 0$. We define the set F_k for functions f that match the following criteria.

1. Let $f : S^1 \to \mathbb{R}^2$ be a generic function (see previous definition).
2. f is in $C^2(S^1, \mathbb{R}^2)$.
3. $f(S^1)$ is contained in a disk of $\mathbb{R}^2$ centred at $(0, 0)$ with radius k.
4. f is a curve of length l_f with $l_f \leq k$.
5. the curvature of the curve f is everywhere not greater than k.

6. every C^1 function $f_1 : S^1 \to \mathbb{R}^2$ such that f_1 has a distance less than $1/k$ to f with respect to the C^1-norm is also generic.

The following theorem will be used to discover hormone-responsive genes in 6.1.

Theorem B.0.38 (Frosini and Landi, 2011, [31]). *Let $k > 0$. For every $\varepsilon > 0$, a $\delta > 0$ exists such that if $f, g \in F_k$ and the matching distance between the functions $rank H_0^{\cdot\cdot}(s \circ f)$ and $rank H_0^{\cdot\cdot}(s \circ g)$ is not greater than δ for every $s \in \Sigma_2$, then there exists a C^1-diffeomorphism $h : S^1 \to S^1$ such that $|| f - g \circ h ||_\infty \leq \varepsilon$.*

Furthermore, if f and g are generic function, the following result holds as well :

Theorem B.0.39 (Landi, 2011, [31]). *If f, g are generic functions and $H_0^{(u,v)}(s \circ f) = H_0^{(u,v)}(s \circ g)$, for every $(u, v) \in \Delta_+ = \{(u, v) \in \mathbb{R}^2 \mid u \leq v\}$ and for every $s \in \Sigma_2$, then there exists a C^1-diffeomorphism $h : S^1 \to S^1$ such that $f = g \circ h$.*

Therefore, we conclude that if f, $g \in F_k$ than we need only to calculate the matching distance (where an algorithm approximating this distance is given in section B.0.4) to establish the existence of a diffeomorphism $h : S^1 \to S^1$ such that $|| f - g \circ h ||_\infty$ is small.

Bibliography

[1] James R Munkres. *Topology*. Prentice Hall, Incorporated, 2000.

[2] B Alberts, A Johnson, J Lewis, D Morgan, M Raff, K Roberts, and P Walter. *Molecular Biology of the Cell, Sixth Edition:*. Taylor & Francis Group, 2014.

[3] B Di Fabio and C Landi. Stable shape comparison of surfaces via Reeb graphs. 2014.

[4] *Range size functions*, volume 2356, 1994.

[5] Patrizio Frosini and Claudia Landi. Size functions and morphological transformations. *Acta Applicandae Mathematica*, 49(1):85–104, Oct 1997.

[6] Patrizio Frosini and Claudia Landi. Size functions and formal series. *Applicable Algebra in Engineering, Communication and Computing*, 12(4):327–349, Aug 2001.

[7] P Frosini and C Landi. Size Theory as a Topological Tool for Computer Vision. Technical report, Pattern Recognition and Image Analysis.

[8] G Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46:255–308, 2009.

[9] M Nicolau, A Levine, and G Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Science*, 108(17):7265–7270, 2011.

[10] G Singh, F Mémoli, and Gr Carlsson. Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. In *Symposium on Point Based Graphics*, 2007.

[11] P Lum, G Singh, A Lehman, T Ishkanov, M Vejdemo-Johansson, M Alagappan, J Carlsson, and G Carlsson. Extracting insights from the shape of complex data using topology. *Scientific Reports*, 3, 2013.

[12] F Chazal and Michel B. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *arXiv*, 2017.

[13] PG Cámara. Topological methods for genomics: Present and future directions. *Current Opinion in Systems Biology*, 2016.

[14] G Reeb. Sur les points singuliers d'une forme de pfaff complètement intégrable ou d'une fonction numérique. *Compte Rendu de l'Académie des Science de Paris*, 222:847–849, 1946.

[15] Brittany Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, and Aarti Singh. Confidence Sets for Persistence Diagrams. *The Annals of Statistics*, 42(6):2301–2339, 2014.

[16] C Maria. Gudhi, simplicial complexes and persistent homology packages, 2014.

[17] Artificial Intelligence for the Enterprise | Ayasdi.

[18] Yongjin Lee, Senja D Barthel, Paweł Dłotko, S Mohamad Moosavi, Kathryn Hess, and Berend Smit. Quantifying similarity of pore-geometry in nanoporous materials. *Nature Communications*, 8, May 2017.

[19] AH Rizvi, PG Camara, EK Kandror, TJ Roberts, I Schieren, T Maniatis, and R Rabadan. Single-cell topological rna-seq analysis reveals insights into cellular differentiation and development. *Nature Biotechnology*, 35.

[20] L De Cecco, M Nicolau, M Giannoccaro, MG Daidone, P Bossi, L Locati, L Licitra, and S Canevari. Head and neck cancer subtypes with biological and clinical relevance: Meta-analysis of gene-expression data. *Oncotarget*, 6:9627–9642, 2015.

[21] J Chang, M Nicolau, TR Cox, D Wetterskog, JWM Martens, HE Barker, and JT Erler. Loxl2 induces aberrant acinar morphogenesis via erbb2 signaling. *Breast Cancer Research*, 15.

[22] Mathieu Carrière and Steve Oudot. Structure and Stability of the 1-Dimensional Mapper. In *Proceedings of the 32nd Symposium on Computational Geometry*, volume 51, pages 25:1–25:16, 2016.

[23] Mathieu Carrière and Steve Oudot. Structure and Stability of the 1-Dimensional Mapper. *Foundations of Computational Mathematics*, 2017.

[24] Tamal K Dey, Facundo Memoli, and Yusu Wang. Mutiscale Mapper: A Framework for Topological Summarization of Data and Maps. *arXiv:1504 03763 [cs, math]*, April 2015. arXiv: 1504 03763.

[25] Tamal K Dey, Facundo Memoli, and Yusu Wang. Topological Analysis of Nerves, Reeb Spaces, Mappers, and Multiscale Mappers. *arXiv:1703 07387 [cs, math]*, March 2017. arXiv: 1703 07387.

[26] C Hennig, M Meila, F Murtagh, and R Rocci. *Handbook of cluster analysis*. CRC Press, 2015.

[27] Artificial Intelligence for the Enterprise | Ayasdi.

[28] Herbert Edelsbrunner, John Harer, Patrizio Frosini, Massimo Ferri, and Vanessa Robins. Persistent homology — a survey.

[29] C Uras and A Verri. On the recognition of the alphabet of the sign language through size functions. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol 3 - Conference C: Signal Processing (Cat No 94CH3440-5)*, volume 2, pages 334–338 vol 2, Oct 1994.

[30] S Banerjee. Size functions in the study of the evolution of cyclones. 36:39–46, 03 2011.

[31] Patrizio Frosini and Claudia Landi. Uniqueness of models in persistent homology: The case of curves. 27, 12 2010.

[32] Vanessa Robins. Towards computing homology from approximations. 24, 01 1999.

[33] S T Hyde, S J Ramsden, and V Robins. Unification and classification of two-dimensional crystalline patterns using orbifolds. *Acta Crystallographica Section A*, 70(4):319–337, May 2014.

[34] Vanessa Robins, Mohammad Saadatfar, Olaf Delgado-Friedrichs, and Adrian P Sheppard. Percolating length scales from topological persistence analysis of micro-ct images of porous materials. *Water Resources Research*, 52:315–329, 1 2016.

[35] H Edelsbrunner and J Harer. *Computational Topology: An Introduction*. Applied Mathematics. American Mathematical Society, 2010.

[36] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. *Discrete and Computational Geometry*, 28:511–533, 2002.

[37] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete and Computational Geometry*, 33(2):249–274, 2005.

[38] Gunnar Carlsson, Tigran Ishkhanov, Vin de Silva, and Afra Zomorodian. On the local behavior of spaces of natural images. *International Journal of Computer Vision*, 76(1):1–12, Jan 2008.

[39] S Y Oudot. *Persistence Theory: From Quiver Representations to Data Analysis:*. Mathematical Surveys and Monographs. American Mathematical Society, 2015.

[40] Vin de Silva and Gunnar Carlsson. Topological estimation using witness complexes. In Markus Gross, Hanspeter Pfister, Marc Alexa, and Szymon Rusinkiewicz, editors, *SPBG'04 Symposium on Point - Based Graphics 2004*. The Eurographics Association, 2004.

[41] Milka Doktorova and Afra Zomorodian. Constructing simplicial complexes over topological spaces, May 2014.

[42] Herbert Edelsbrunner and Dmitriy Morozov. Persistent homology: Theory and practice.

[43] Peter Bubenik and Jonathan A Scott. Categorification of persistent homology. *Discrete & Computational Geometry*, 51(3):600–627, Apr 2014.

[44] Peter Bubenik. Statistical topological data analysis using persistence landscapes. *J Mach Learn Res*, 16(1):77–102, January 2015.

[45] Andrew Robinson and Katharine Turner. Hypothesis testing for topological data analysis. *Journal of Applied and Computational Topology*, 1(2):241–261, Dec 2017.

[46] Yanjie Li, Dingkang Wang, Giorgio A Ascoli, Partha Mitra, and Yusu Wang. Metrics for comparing neuronal tree shapes based on persistent homology. *PLoS ONE*, 12(8):e0182184, 2017.

[47] Lida Kanari, Pawel Dlotko, Martina Scolamiero, Ran Levi, Julian C Shillcock, Kathryn Hess, and Henry Markram. A topological representation of branching neuronal morphologies. *Neuroinformatics*, 16(1):3–13, 2018.

[48] Akihiro Takiyama, Takashi Teramoto, Hiroaki Suzuki, Katsushige Yamashiro, and Shinya Tanaka. Persistent homology index as a robust quantitative measure of immunohistochemical scoring. *Scientific Reports*, 7:14002, 2017.

[49] Noah Giansiracusa, Robert Giansiracusa, and Chul Moon. Persistent homology machine learning for fingerprint classification. *arXiv:1711 09158 [math, stat]*, November 2017. arXiv: 1711 09158.

[50] Allen Hatcher. *Algebraic Topology*. 2001.

[51] Peter J May. *A Concise Course in Algebraic Topology*. 1999.

[52] Peter Bubenik and Jonathan A Scott. Categorification of persistent homology. *Discrete & Computational Geometry*, 51(3):600–627, April 2014. arXiv: 1205 3669.

[53] F Chazal, V de Silva, M Glisse, and S Oudot. *The Structure and Stability of Persistence Modules*. Springer, 2016.

[54] Frédéric Chazal, David Cohen-Steiner, Marc Glisse, Leonidas Guibas, and Steve Oudot. Proximity of Persistence Modules and their Diagrams. In *Proceedings of the 25th Symposium on Computational Geometry*, pages 237–246, 2009.

[55] Morse Theory for Cell Complexes. *Advances in Mathematics*, 134(1):90–145, March 1998.

[56] V de Silva, E Munch, and A Patel. Categorified Reeb Graphs. *Discrete and Computational Geometry*, 55:854–906, 2016.

[57] D Cohen-Steiner, H Edelsbrunner, and J Harer. Extending persistence using Poincaré and Lefschetz duality. *Foundation of Computational Mathematics*, 9(1):79–103, 2009.

[58] U Bauer, X Ge, and Y Wang. Measuring Distance Between Reeb Graphs. In *Proceedings of the 30th Symposium on Computational Geometry*, pages 464–473, 2014.

[59] Jay Shendure, Shankar Balasubramanian, George M Church, Walter Gilbert, Jane Rogers, Jeffery A Schloss, and Robert H Waterston. DNA sequencing at 40: past, present and future. *Nature*, 550(7676):345–353, October 2017.

[60] Frederick Sanger - Nobel Lecture: The Chemistry of Insulin.

[61] Robert W Holley, Jean Apgar, George A Everett, James T Madison, Mark Marquisee, Susan H Merrill, John Robert Penswick, and Ada Zamir. Structure of a ribonucleic acid. *Science*, 147(3664):1462–1465, 1965.

[62] Support Center for Microsystems Education. How Does a DNA Microarray Work?

[63] L Klebanov and A Yakovlev. How high is the level of technical noise in microarray data. *Biology Direct*, 2.

[64] Alvis Brazma, Pascal Hingamp, John Quackenbush, Gavin Sherlock, Paul Spellman, Chris Stoeckert, John Aach, Wilhelm Ansorge, Catherine A Ball, Helen C Causton, Terry Gaasterland, Patrick Glenisson, Frank C P Holstege, Irene F Kim, Victor Markowitz, John C Matese, Helen Parkinson, Alan Robinson, Ugis Sarkans, Steffen Schulze-Kremer, Jason Stewart, Ronald Taylor, Jaak Vilo, and Martin Vingron. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data, December 2001.

[65] StatQuest with Josh Starmer. StatQuest: A gentle introduction to RNA-seq.

[66] Maurizio Callari, Ankita Sati Batra, Rajbir Nath Batra, Stephen-John Sammut, Wendy Greenwood, Harry Clifford, Colin Hercus, Suet-Feung Chin, Alejandra Bruna, Oscar M Rueda, and Carlos Caldas. Computational approach to discriminate human and mouse sequences in patient-derived tumour xenografts. *BMC Genomics*, 19(1):19, December 2018.

[67] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14:R36, April 2013.

[68] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, April 2012.

[69] Daehwan Kim, Ben Langmead, and Steven L Salzberg. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, 12(4):357–360, April 2015.

[70] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–2079, August 2009.

[71] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven Salzberg, John Rinn, and Lior Pachter. Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. 7:562–78, 03 2012.

[72] Picard Tools - By Broad Institute.

[73] Malachi Griffith, Jason R Walker, Nicholas C Spies, Benjamin J Ainscough, and Obi L Griffith. Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud. *PLOS Computational Biology*, 11(8):e1004393, 2015.

[74] Yang Liao, Gordon Smyth, and Wei Shi. Featurecounts: An efficient general purpose program for assigning sequence reads to genomic features. 30, 11 2013.

[75] Vincent Gardeux, Fabrice P A David, Adrian Shajkofci, Petra C Schwalie, and Bart Deplancke. Asap: a web-based platform for the analysis and interactive visualization of single-cell rna-seq data. *Bioinformatics*, 33(19):3123–3125, 2017.

[76] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15:R29, February 2014.

[77] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[78] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15:550, December 2014.

[79] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015.

[80] Martin Morgan, Seth Falcon, and Robert Gentleman. *GSEABase: Gene set enrichment data structures and methods*. R package version 1 32 0.

[81] Adrian Alexa, Jörg Rahnenführer, and Thomas Lengauer. Improved scoring of functional groups from gene expression data by decorrelating go graph structure. 22:1600–7, 08 2006.

[82] Andrea Komljenovic, Julien Roux, Marc Robinson-Rechavi, and Frederic B Bastian. BgeeDB, an R package for retrieval of curated expression datasets and for gene list expression localization enrichment tests. *F1000Research*, 5:2748, November 2016.

[83] Gene Ontology Consortium | Gene Ontology Consortium.

[84] PANTHER - Gene List Analysis.

[85] STRING: functional protein association networks.

[86] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang, and Ali Mortazavi. A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17, 2016.

[87] Peipei Li, Yongjun Piao, Ho Sun Shon, and Keun Ho Ryu. Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. *BMC Bioinformatics*, 16:347, October 2015.

[88] Charlotte Soneson and Mauro Delorenzi. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, 14:91, March 2013.

[89] Hannah Dueck, Mugdha Khaladkar, Tae Kyung Kim, Jennifer M Spaethling, Chantal Francis, Sangita Suresh, Stephen A Fisher, Patrick Seale, Sheryl G Beck, Tamas Bartfai, Bernhard Kuhn, James Eberwine, and Junhyong Kim. Deep sequencing reveals cell-type-specific patterns of single-cell transcriptome variation. *Genome Biology*, 16(1):122, Jun 2015.

[90] Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11(7):740–742, July 2014.

[91] Guo-Cheng Yuan, Long Cai, Michael Elowitz, Tariq Enver, Guoping Fan, Guoji Guo, Rafael Irizarry, Peter Kharchenko, Junhyong Kim, Stuart Orkin, John Quackenbush, Assieh Saadatpour, Timm Schroeder, Ramesh Shivdasani, and Itay Tirosh. Challenges and emerging directions in single-cell analysis. *Genome Biology*, 18:84, May 2017.

[92] M Ester, HP Kriegel, J Sander, and X Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226–231, 1996.

[93] Saket J Swarndeep and Pandya Sharnil. An overview of partitioning algorithms in clustering techniques. *International Journal of Advanced Research in Computer Engineering & Technology*, 5(6):1943–1946, 2016.

[94] J A García, J Fdez-Valdivia, F J Cortijo, and R Molina. A dynamic approach for clustering data. *Signal Processing*, 44(2):181–196, June 1995.

[95] JA Hartigan and MA Wong. A k-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.

[96] C Fraley and AE Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

[97] KMeans function | R Documentation.

[98] C Fraley and A E Raftery. Mclust: Software for model-based cluster analysis. *Journal of Classification*, pages 297–306, 1999.

[99] Chris Fraley and Adrian E Raftery. Enhanced Model-Based Clustering, Density Estimation, and Discriminant Analysis Software: MCLUST. *Journal of Classification*, 20(2):263–286, September 2003.

[100] Chris Fraley, Adrian Raftery, Thomas Murphy, and Luca Scrucca. Mclust version 4 for r: Normal mixture modeling for model-based clustering, classification, and density estimation. 01 2012.

[101] Luca Scrucca, Michael Fop, T Brendan Murphy, and Adrian E Raftery. mclust 5: Clustering, classification and density estimation using gaussian finite mixture models. *The R journal*, 8(1):289–317, 08 2016.

[102] A P Dempster, N M Laird, and D B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[103] G J McLachlan and D Peel. *Finite Mixture Models*. John Wiley & Sons, 2000.

[104] Jeffrey D Banfield and Adrian E Raftery. Model-based gaussian and non-gaussian clustering. 49, 09 1993.

[105] Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781–793, May 1995.

[106] Cathrin Brisken, Kathryn Hess, and Rachel Jeitziner. Progesterone and Overlooked Endocrine Pathways in Breast Cancer Pathogenesis. *Endocrinology*, 156(10):3442–3450, October 2015.

[107] Cathrin Brisken and Bert O'Malley. Hormone action in the mammary gland. *Cold Spring Harbor Perspectives in Biology*, 2(12):a003178, December 2010.

[108] Sonia Mallepell, Andrée Krust, Pierre Chambon, and Cathrin Brisken. Paracrine signaling through the epithelial estrogen receptor alpha is required for proliferation and morphogenesis in the mammary gland. *Proceedings of the National Academy of Sciences of the United States of America*, 103(7):2196–2201, February 2006.

[109] C Brisken, S Park, T Vass, J P Lydon, B W O'Malley, and R A Weinberg. A paracrine role for the epithelial progesterone receptor in mammary gland development. *Proceedings of the National Academy of Sciences of the United States of America*, 95(9):5076–5081, April 1998.

[110] C Brisken, S Kaur, T E Chavarria, N Binart, R L Sutherland, R A Weinberg, P A Kelly, and C J Ormandy. Prolactin controls mammary gland development via direct and indirect mechanisms. *Developmental Biology*, 210(1):96–106, June 1999.

[111] C W Daniel, G B Silberstein, and P Strickland. Direct action of 17 beta-estradiol on mouse mammary ducts analyzed by sustained release implants and steroid autoradiography. *Cancer Research*, 47(22):6052–6057, November 1987.

[112] S Z Haslam and G Shyamala. Progesterone receptors in normal mammary glands of mice: characterization and relationship to development. *Endocrinology*, 105(3):786–795, September 1979.

[113] Sandra L Grimm, Tiffany N Seagroves, Elena B Kabotyanski, Russell C Hovey, Barbara K Vonderhaar, John P Lydon, Keiko Miyoshi, Lothar Hennighausen, Christopher J Ormandy, Adrian V Lee, Malinda A Stull, Teresa L Wood, and Jeffrey M Rosen. Disruption of steroid and prolactin receptor patterning in the mammary gland correlates with a block in lobuloalveolar development. *Molecular Endocrinology (Baltimore, Md)*, 16(12):2675–2691, December 2002.

[114] Breast - Wikipedia.

[115] J R Masters, J O Drife, and J J Scarisbrick. Cyclic Variation of DNA synthesis in human breast epithelium. *Journal of the National Cancer Institute*, 58(5):1263–1265, May 1977.

[116] T A Longacre and S A Bartow. A correlative morphologic study of human breast and endometrium in the menstrual cycle. *The American Journal of Surgical Pathology*, 10(6):382–393, June 1986.

[117] Proliferation of breast epithelial cells in healthy women during the menstrual cycle. *American Journal of Obstetrics and Gynecology*, 176(1):123–128, January 1997.

[118] Tamara Tanos, George Sflomos, Pablo C Echeverria, Ayyakkannu Ayyanan, Maria Gutierrez, Jean-Francois Delaloye, Wassim Raffoul, Maryse Fiche, William Dougall, Pascal Schneider, Ozden Yalcin-Ozuysal, and Cathrin Brisken. Progesterone/RANKL Is a Major Regulatory Axis in the Human Breast. *Science Translational Medicine*, 5(182):182ra55–182ra55, April 2013.

[119] Jun Wang, Akash Gupta, Hong Hu, Robert T Chatterton, Charles V Clevenger, and Seema A Khan. Comment on "Progesterone/RANKL is a major regulatory axis in the human breast". *Science Translational Medicine*, 5(215):215le4, December 2013.

[120] N Sleeckx, H de Rooster, E J B Veldhuis Kroeze, C Van Ginneken, and L Van Brantegem. Canine mammary tumours, an overview. *Reproduction in Domestic Animals = Zuchthygiene*, 46(6):1112–1131, December 2011.

[121] R B Clarke, A Howell, C S Potten, and E Anderson. Dissociation between steroid receptor expression and cell proliferation in the human breast. *Cancer Research*, 57(22):4987–4991, November 1997.

[122] T N Seagroves, J P Lydon, R C Hovey, B K Vonderhaar, and J M Rosen. C/EBPbeta (CCAAT/enhancer binding protein) controls cell fate determination during mammary gland development. *Molecular Endocrinology (Baltimore, Md)*, 14(3):359–368, March 2000.

[123] Heidi N Hilton, Tram B Doan, J Dinny Graham, Samantha R Oakes, Audrey Silvestri, Nicole Santucci, Silke Kantimm, Lily I Huschtscha, Christopher J Ormandy, John W Funder, Evan R Simpson, Elizabeth S Kuczek, Peter J Leedman, Wayne D Tilley, Peter J Fuller, George E O Muscat, and Christine L Clarke. Acquired convergence of hormone signaling in breast cancer: ER and PR transition from functionally distinct in normal breast to predictors of metastatic disease. *Oncotarget*, 5(18):8651–8664, September 2014.

[124] Heidi N Hilton, N Santucci, A Silvestri, S Kantimm, L I Huschtscha, J D Graham, and C L Clarke. Progesterone stimulates progenitor cells in normal human breast and breast cancer cells. *Breast Cancer Research and Treatment*, 143(3):423–433, February 2014.

[125] Cathrin Brisken and Stephan Duss. Stem cells and the stem cell niche in the breast: an integrated hormonal and developmental perspective. *Stem Cell Reviews*, 3(2):147–156, June 2007.

[126] Manfred Beleut, Renuga Devi Rajaram, Marian Caikovski, Ayyakkannu Ayyanan, Davide Germano, Yongwon Choi, Pascal Schneider, and Cathrin Brisken. Two distinct mechanisms underlie progesterone-induced proliferation in the mammary gland. *Proceedings of the National Academy of Sciences of the United States of America*, 107(7):2989–2994, February 2010.

[127] Gwen E Dressing, Todd P Knutson, Matthew J Schiewer, Andrea R Daniel, Christy R Hagan, Caroline H Diep, Karen E Knudsen, and Carol A Lange. Progesterone receptor-cyclin D1 complexes induce cell cycle-dependent transcriptional programs in breast cancer cells. *Molecular Endocrinology (Baltimore, Md)*, 28(4):442–457, April 2014.

[128] Biserka Mulac-Jericevic, John P Lydon, Francesco J DeMayo, and Orla M Conneely. Defective mammary gland morphogenesis in mice lacking the progesterone receptor B isoform. *Proceedings of the National Academy of Sciences of the United States of America*, 100(17):9744–9749, August 2003.

[129] Eva Gonzalez-Suarez, Allison P Jacob, Jon Jones, Robert Miller, Martine P Roudier-Meyer, Ryan Erwert, Jan Pinkas, Dan Branstetter, and William C Dougall. RANK ligand mediates progestin-induced mammary epithelial proliferation and carcinogenesis. *Nature*, 468(7320):103–107, November 2010.

[130] John Stingl, Peter Eirew, Ian Ricketson, Mark Shackleton, François Vaillant, David Choi, Haiyan I Li, and Connie J Eaves. Purification and unique properties of mammary epithelial stem cells. *Nature*, 439(7079):993–997, February 2006.

[131] Mark Shackleton, François Vaillant, Kaylene J Simpson, John Stingl, Gordon K Smyth, Marie-Liesse Asselin-Labat, Li Wu, Geoffrey J Lindeman, and Jane E Visvader. Generation of a functional mammary gland from a single stem cell. *Nature*, 439(7072):84–88, January 2006.

[132] Purna A Joshi, Hartland W Jackson, Alexander G Beristain, Marco A Di Grappa, Patricia A Mote, Christine L Clarke, John Stingl, Paul D Waterhouse, and Rama Khokha. Progesterone induces adult mammary stem cell expansion. *Nature*, 465(7299):803–807, June 2010.

[133] Marie-Liesse Asselin-Labat, François Vaillant, Julie M Sheridan, Bhupinder Pal, Di Wu, Evan R Simpson, Hisataka Yasuda, Gordon K Smyth, T John Martin, Geoffrey J Lindeman, and Jane E Visvader. Control of mammary stem cell function by steroid hormone signalling. *Nature*, 465(7299):798–802, June 2010.

[134] Ilaria Taddei, Marie-Ange Deugnier, Marisa M Faraldo, Valérie Petit, Daniel Bouvard, Daniel Medina, Reinhard Fässler, Jean Paul Thiery, and Marina A Glukhova. Beta1 integrin deletion from the basal compartment of the mammary epithelium affects stem cells. *Nature Cell Biology*, 10(6):716–722, June 2008.

[135] Alexandra Van Keymeulen, Ana Sofia Rocha, Marielle Ousset, Benjamin Beck, Gaëlle Bouvencourt, Jason Rock, Neha Sharma, Sophie Dekoninck, and Cédric Blanpain. Distinct stem cells contribute to mammary gland development and maintenance. *Nature*, 479(7372):189–193, October 2011.

[136] L J Young, D Medina, K B DeOme, and C W Daniel. The influence of host and tissue age on life span and growth rate of serially transplanted mouse mammary gland. *Experimental Gerontology*, 6(1):49–56, February 1971.

[137] Renuga Devi Rajaram, Duje Buric, Marian Caikovski, Ayyakkannu Ayyanan, Jacques Rougemont, Jingdong Shan, Seppo J Vainio, Ozden Yalcin-Ozuysal, and Cathrin Brisken. Progesterone and Wnt4 control mammary stem cells via myoepithelial crosstalk. *The EMBO journal*, 34(5):641–652, March 2015.

[138] Kevin Roarty and Jeffrey M Rosen. Wnt and mammary stem cells: hormones cannot fly wingless. *Current Opinion in Pharmacology*, 10(6):643–649, December 2010.

[139] Daisong Wang, Cheguo Cai, Xiaobing Dong, Qing Cissy Yu, Xiao-Ou Zhang, Li Yang, and Yi Arial Zeng. Identification of multipotent mammary stem cells by protein C receptor expression. *Nature*, 517(7532):81–84, January 2015.

[140] Cheguo Cai, Qing Cissy Yu, Weimin Jiang, Wei Liu, Wenqian Song, Hua Yu, Lei Zhang, Ying Yang, and Yi Arial Zeng. R-spondin1 is a novel hormone mediator for mammary stem cell self-renewal. *Genes & Development*, 28(20):2205–2218, October 2014.

[141] Sara Lombardi, Gabriella Honeth, Christophe Ginestier, Ireneusz Shinomiya, Rebecca Marlow, Bharath Buchupalli, Patrycja Gazinska, John Brown, Steven Catchpole, Suling Liu, Ariel Barkan, Max Wicha, Anand Purushotham, Joy Burchell, Sarah Pinder, and Gabriela Dontu. Growth hormone is secreted by normal breast epithelium upon progesterone stimulation and increases proliferation of stem/progenitor cells. *Stem Cell Reports*, 2(6):780–793, June 2014.

[142] I Pardo, HA Lillemoe, RJ Blosser, M Choi, CA Sauder, DK Doxey, T Mathieson, BA Hancock, D Baptiste, R Atale, M Hickenbotham, J Zhu, J Glasscock, AM Storniolo, F Zheng, RW Doerge, Y Liu, S Badve, M Radovich, SE Clare, and Susan G Komen for the Cure Tissue Bank at the IU Simon Cancer Center. Next-generation transcriptome sequencing of the premenopausal breast epithelium using specimens from a normal human breast tissue bank. *Breast Cancer Research*, 16.

[143] Jennifer K Richer, Britta M Jacobsen, Nicole G Manning, M Greg Abel, Douglas M Wolf, and Kathryn B Horwitz. Differential gene regulation by the two progesterone receptor isoforms in human breast cancer cells. *The Journal of Biological Chemistry*, 277(7):5209–5218, February 2002.

[144] Britta M Jacobsen and Kathryn B Horwitz. Progesterone receptors, their isoforms and progesterone regulated transcription. *Molecular and Cellular Endocrinology*, 357(1-2):18–29, June 2012.

[145] Cancer Statistics Review, 1975-2011 - Previous Version - SEER Cancer Statistics Review.

[146] SR Lakhani, IO Ellis, SJ Schnitt, PH Tan, and MJ van de Vijver. Invasive breast carcinoma, 2012.

[147] Brian D Lehmann, Joshua A Bauer, Xi Chen, Melinda E Sanders, A Bapsi Chakravarthy, Yu Shyr, and Jennifer A Pietenpol. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *The Journal of Clinical Investigation*, 121(7):2750–2767, July 2011.

[148] Amir Sonnenblick, Debora Fumagalli, Christos Sotiriou, and Martine Piccart. Is the differentiation into molecular subtypes of breast cancer important for staging, local and systemic therapy, and follow up? *Cancer Treatment Reviews*, 40(9):1089–1095, October 2014.

[149] Early Breast Cancer Trialists' Collaborative Group (EBCTCG). Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *Lancet (London, England)*, 365(9472):1687–1717, May 2005.

[150] M Elizabeth H Hammond, Daniel F Hayes, Mitch Dowsett, D Craig Allred, Karen L Hagerty, Sunil Badve, Patrick L Fitzgibbons, Glenn Francis, Neil S Goldstein, Malcolm Hayes, David G Hicks, Susan Lester, Richard Love, Pamela B Mangu, Lisa McShane, Keith Miller, C Kent Osborne, Soonmyung Paik, Jane Perlmutter, Anthony Rhodes, Hironobu Sasano, Jared N Schwartz, Fred C G Sweep, Sheila Taube, Emina Emilia Torlakovic, Paul Valenstein, Giuseppe Viale, Daniel Visscher, Thomas Wheeler, R Bruce Williams, James L Wittliff, and Antonio C Wolff. American Society of Clinical Oncology/-College Of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology*, 28(16):2784–2795, June 2010.

[151] Cathrin Brisken. Progesterone signalling in breast cancer: a neglected hormone coming into the limelight. *Nature Reviews Cancer*, 13(6):385–396, June 2013.

[152] Keely M McNamara, Nicole L Moore, Theresa E Hickey, Hironobu Sasano, and Wayne D Tilley. Complexities of androgen receptor signalling in breast cancer. *Endocrine-Related Cancer*, 21(4):T161–181, August 2014.

[153] Carmen J Narvaez, Donald Matthews, Erika LaPorta, Katrina M Simmons, Sarah Beaudin, and JoEllen Welsh. The impact of vitamin D in breast cancer: genomics, pathways, metabolism. *Frontiers in Physiology*, 5:213, 2014.

[154] George T Beatson. On the Treatment of Inoperable Cases of Carcinoma of the Mamma: Suggestions for a New Method of Treatment, with Illustrative Cases. *Transactions Medico-Chirurgical Society of Edinburgh*, 15:153–179, 1896.

[155] M C Pike, M D Krailo, B E Henderson, J T Casagrande, and D G Hoel. 'Hormonal' risk factors, 'breast tissue age' and the age-incidence of breast cancer. *Nature*, 303(5920):767–770, June 1983.

[156] D Trichopoulos, B MacMahon, and P Cole. Menopause and breast cancer risk. *Journal of the National Cancer Institute*, 48(3):605–613, March 1972.

[157] John L Young, Susan S Devesa, and Sidney J Cutler. Incidence of Cancer in United States Blacks. *Cancer Research*, 35(11 Part 2):3523–3536, November 1975.

[158] Graham A Colditz, Bernard A Rosner, Wendy Y Chen, Michelle D Holmes, and Susan E Hankinson. Risk factors for breast cancer according to estrogen and progesterone receptor status. *Journal of the National Cancer Institute*, 96(3):218–228, February 2004.

[159] B MacMahon, P Cole, T M Lin, C R Lowe, A P Mirra, B Ravnihar, E J Salber, V G Valaoras, and S Yuasa. Age at first birth and breast cancer risk. *Bulletin of the World Health Organization*, 43(2):209–221, 1970.

[160] G Albrektsen, I Heuch, S Hansen, and G Kvåle. Breast cancer risk by age at birth, time since birth and time intervals between births: exploring interaction effects. *British Journal of Cancer*, 92(1):167–175, January 2005.

[161] G Thordarson, E Jin, R C Guzman, S M Swanson, S Nandi, and F Talamantes. Refractoriness to mammary tumorigenesis in parous rats: is it caused by persistent changes in the hormonal environment or permanent biochemical alterations in the mammary epithelia? *Carcinogenesis*, 16(11):2847–2853, November 1995.

[162] A Heather Eliassen, Shelley S Tworoger, and Susan E Hankinson. Reproductive factors and family history of breast cancer in relation to plasma prolactin levels in premenopausal and postmenopausal women. *International Journal of Cancer*, 120(7):1536–1541, April 2007.

[163] C Kuperwasser, J Pinkas, G D Hurlbut, S P Naber, and D J Jerry. Cytoplasmic sequestration and functional repression of p53 in the mammary epithelium is reversed by hormonal treatment. *Cancer Research*, 60(10):2723–2729, May 2000.

[164] L Sivaraman, L C Stephens, B M Markaverich, J A Clark, S Krnacik, O M Conneely, B W O'Malley, and D Medina. Hormone-induced refractoriness to mammary carcinogenesis in Wistar-Furth rats. *Carcinogenesis*, 19(9):1573–1581, September 1998.

[165] Valerie Beral and Million Women Study Collaborators. Breast cancer and hormone-replacement therapy in the Million Women Study. *Lancet (London, England)*, 362(9382):419–427, August 2003.

[166] Rowan T Chlebowski, Garnet L Anderson, Margery Gass, Dorothy S Lane, Aaron K Aragaki, Lewis H Kuller, JoAnn E Manson, Marcia L Stefanick, Judith Ockene, Gloria E Sarto, Karen C Johnson, Jean Wactawski-Wende, Peter M Ravdin, Robert Schenken, Susan L Hendrix, Aleksandar Rajkovic, Thomas E Rohan, Shagufta Yasmeen, Ross L Prentice, and WHI Investigators. Estrogen plus progestin and breast cancer incidence and mortality in postmenopausal women. *JAMA*, 304(15):1684–1692, October 2010.

[167] Jacques E Rossouw, Garnet L Anderson, Ross L Prentice, Andrea Z LaCroix, Charles Kooperberg, Marcia L Stefanick, Rebecca D Jackson, Shirley A A Beresford, Barbara V Howard, Karen C Johnson, Jane Morley Kotchen, Judith Ockene, and Writing Group for the Women's Health Initiative Investigators. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results From the Women's Health Initiative randomized controlled trial. *JAMA*, 288(3):321–333, July 2002.

[168] Garnet L Anderson, Rowan T Chlebowski, Aaron K Aragaki, Lewis H Kuller, JoAnn E Manson, Margery Gass, Elizabeth Bluhm, Stephanie Connelly, F Allan Hubbell, Dorothy Lane, Lisa Martin, Judith Ockene, Thomas Rohan, Robert Schenken, and Jean Wactawski-Wende. Conjugated equine oestrogen and breast cancer incidence and mortality in postmenopausal women with hysterectomy: extended follow-up of the

Women's Health Initiative randomised placebo-controlled trial. *The Lancet Oncology*, 13(5):476–486, May 2012.

[169] Peter M Ravdin, Kathleen A Cronin, Nadia Howlader, Christine D Berg, Rowan T Chlebowski, Eric J Feuer, Brenda K Edwards, and Donald A Berry. The decrease in breast-cancer incidence in 2003 in the United States. *The New England Journal of Medicine*, 356(16):1670–1674, April 2007.

[170] Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and hormonal contraceptives: collaborative reanalysis of individual data on 53 297 women with breast cancer and 100 239 women without breast cancer from 54 epidemiological studies. *Lancet (London, England)*, 347(9017):1713–1727, June 1996.

[171] I Sinha-Hikim, S Arver, G Beall, R Shen, M Guerrero, F Sattler, C Shikuma, J C Nelson, B M Landgren, N A Mazer, and S Bhasin. The use of a sensitive equilibrium dialysis method for the measurement of free testosterone levels in healthy, cycling women and in human immunodeficiency virus-infected women. *The Journal of Clinical Endocrinology and Metabolism*, 83(4):1312–1318, April 1998.

[172] H L Judd and S S Yen. Serum androstenedione and testosterone levels during the menstrual cycle. *The Journal of Clinical Endocrinology and Metabolism*, 36(3):475–481, March 1973.

[173] Eva S Schernhammer, Francesca Sperati, Pedram Razavi, Claudia Agnoli, Sabina Sieri, Franco Berrino, Vittorio Krogh, Carlo Abbagnato, Sara Grioni, Giovanni Blandino, Holger J Schunemann, and Paola Muti. Endogenous sex steroids in premenopausal women and risk of breast cancer: the ORDET cohort. *Breast cancer research: BCR*, 15(3):R46, June 2013.

[174] T E Hickey, J L L Robinson, J S Carroll, and W D Tilley. Minireview: The androgen receptor in breast tissues: growth inhibitor, tumor suppressor, oncogene? *Molecular Endocrinology (Baltimore, Md)*, 26(8):1252–1267, August 2012.

[175] Cathrin Brisken and Dalya Ataca. Endocrine hormones and local signals during the development of the mouse mammary gland. *Wiley Interdisciplinary Reviews Developmental Biology*, 4(3):181–195, June 2015.

[176] Sandro Santagata, Ankita Thakkar, Ayse Ergonul, Bin Wang, Terri Woo, Rong Hu, J Chuck Harrell, George McNamara, Matthew Schwede, Aedin C Culhane, David Kindelberger, Scott Rodig, Andrea Richardson, Stuart J Schnitt, Rulla M Tamimi, and Tan A Ince. Taxonomy of breast cancer based on normal cell phenotype predicts outcome. *The Journal of Clinical Investigation*, 124(2):859–870, February 2014.

[177] C M Aldaz, Q Y Liao, M LaBate, and D A Johnston. Medroxyprogesterone acetate accelerates the development and increases the incidence of mouse mammary tumors induced by dimethylbenzanthracene. *Carcinogenesis*, 17(9):2069–2072, September 1996.

[178] A G Jabara, G N Marks, J E Summers, and P S Anderson. Effects of progesterone on mammary carcinogenesis by DMBA applied directly to rat mammae. *British Journal of Cancer*, 40(2):268–273, August 1979.

[179] Daniel Schramek, Andreas Leibbrandt, Verena Sigl, Lukas Kenner, John A Pospisilik, Heather J Lee, Reiko Hanada, Purna A Joshi, Antonios Aliprantis, Laurie Glimcher, Manolis Pasparakis, Rama Khokha, Christopher J Ormandy, Martin Widschwendter, Georg Schett, and Josef M Penninger. Osteoclast differentiation factor RANKL controls development of progestin-driven mammary cancer. *Nature*, 468(7320):98–102, November 2010.

[180] R Nusse and H E Varmus. Many tumors induced by the mouse mammary tumor virus contain a provirus integrated in the same region of the host genome. *Cell*, 31(1):99–109, November 1982.

[181] A S Tsukamoto, R Grosschedl, R C Guzman, T Parslow, and H E Varmus. Expression of the int-1 gene in transgenic mice is associated with mammary gland hyperplasia and adenocarcinomas in male and female mice. *Cell*, 55(4):619–625, November 1988.

[182] Emily J Faivre and Carol A Lange. Progesterone receptors upregulate Wnt-1 to induce epidermal growth factor receptor transactivation and c-Src-dependent sustained activation of Erk1/2 mitogen-activated protein kinase in breast cancer cells. *Molecular and Cellular Biology*, 27(2):466–480, January 2007.

[183] D M Anderson, E Maraskovsky, W L Billingsley, W C Dougall, M E Tometsko, E R Roux, M C Teepe, R F DuBose, D Cosman, and L Galibert. A homologue of the TNF receptor and its ligand enhance T-cell growth and dendritic-cell function. *Nature*, 390(6656):175–179, November 1997.

[184] W S Simonet, D L Lacey, C R Dunstan, M Kelley, M S Chang, R Lüthy, H Q Nguyen, S Wooden, L Bennett, T Boone, G Shimamoto, M DeRose, R Elliott, A Colombero, H L Tan, G Trail, J Sullivan, E Davy, N Bucay, L Renshaw-Gegg, T M Hughes, D Hill, W Pattison, P Campbell, S Sander, G Van, J Tarpley, P Derby, R Lee, and W J Boyle. Osteoprotegerin: a novel secreted protein involved in the regulation of bone density. *Cell*, 89(2):309–319, April 1997.

[185] H Yasuda, N Shima, N Nakagawa, S I Mochizuki, K Yano, N Fujise, Y Sato, M Goto, K Yamaguchi, M Kuriyama, T Kanno, A Murakami, E Tsuda, T Morinaga, and K Higashio. Identity of osteoclastogenesis inhibitory factor (OCIF) and osteoprotegerin (OPG): a mechanism by which OPG/OCIF inhibits osteoclastogenesis in vitro. *Endocrinology*, 139(3):1329–1337, March 1998.

[186] H Yasuda, N Shima, N Nakagawa, K Yamaguchi, M Kinosaki, S Mochizuki, A Tomoyasu, K Yano, M Goto, A Murakami, E Tsuda, T Morinaga, K Higashio, N Udagawa, N Takahashi, and T Suda. Osteoclast differentiation factor is a ligand for osteoprotegerin/osteoclastogenesis-inhibitory factor and is identical to TRANCE/RANKL. *Proceedings of the National Academy of Sciences of the United States of America*, 95(7):3597–3602, March 1998.

[187] D L Lacey, E Timms, H L Tan, M J Kelley, C R Dunstan, T Burgess, R Elliott, A Colombero, G Elliott, S Scully, H Hsu, J Sullivan, N Hawkins, E Davy, C Capparelli, A Eli, Y X Qian, S Kaufman, I Sarosi, V Shalhoub, G Senaldi, J Guo, J Delaney, and

W J Boyle. Osteoprotegerin ligand is a cytokine that regulates osteoclast differentiation and activation. *Cell*, 93(2):165–176, April 1998.

[188] Brendan F Boyce and Lianping Xing. Biology of rank, rankl, and osteoprotegerin. *Arthritis Research & Therapy*, 9(Suppl 1):S1–S1, 2007.

[189] Jimmie E Fata, Young-Yun Kong, Ji Li, Takehiko Sasaki, Junko Irie-Sasaki, Roger A Moorehead, Robin Elliott, Sheila Scully, Evelyn B Voura, David L Lacey, William J Boyle, Rama Khokha, and Josef M Penninger. The Osteoclast Differentiation Factor Osteoprotegerin-Ligand Is Essential for Mammary Gland Development. *Cell*, 103(1):41–50, September 2000.

[190] Biserka Mulac-Jericevic, John P Lydon, Francesco J DeMayo, and Orla M Conneely. Defective mammary gland morphogenesis in mice lacking the progesterone receptor B isoform. *Proceedings of the National Academy of Sciences*, 100(17):9744–9749, August 2003.

[191] Eva Gonzalez-Suarez, Allison P Jacob, Jon Jones, Robert Miller, Martine P Roudier-Meyer, Ryan Erwert, Jan Pinkas, Dan Branstetter, and William C Dougall. RANK ligand mediates progestin-induced mammary epithelial proliferation and carcinogenesis. *Nature*, 468(7320):103–107, November 2010.

[192] Daniel Schramek, Andreas Leibbrandt, Verena Sigl, Lukas Kenner, John A Pospisilik, Heather J Lee, Reiko Hanada, Purna A Joshi, Antonios Aliprantis, Laurie Glimcher, Manolis Pasparakis, Rama Khokha, Christopher J Ormandy, Martin Widschwendter, Georg Schett, and Josef M Penninger. Osteoclast differentiation factor RANKL controls development of progestin-driven mammary cancer. *Nature*, 468(7320):98–102, November 2010.

[193] Pellegrini Pasquale, Cordero Alex, Gallego Marta Ines, Dougall William C , Purificacion Munoz, Pujana Miguel Angel, and Gonzalez-Suarez Eva. Constitutive activation of RANK disrupts mammary cell fate leading to tumorigenesis. *STEM CELLS*, 31(9):1954–1965, October 2013.

[194] R Jeitziner, M Carrière, J Rougemont, S Oudot, K Hess, and C Brisken. Two-Tier Mapper: a user-independent clustering method for global gene expression analysis based on topology. *ArXiv e-prints*, December 2017.

[195] Sara Goodwin, John D McPherson, and W Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351, June 2016.

[196] Michael L Metzker. Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11(1):31–46, January 2010.

[197] Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1842(10):1932–1941, October 2014.

[198] MA Dillies, A Rau, J Aubert, C Hennequet-Antier, M Jeanmougin, N Servant, C Keime, G Marot, D Castel, J Estelle, G Guernec, B Jagla, L Jouneau, Laloë, D, C Le Gall, B Schaëffer, S Le Crom, M Guedj, and F Jaffrézic. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in Bioinformatics*, 14(6):671–683, 2013.

[199] JW Osborne and A Overbay. The power of outliers (and why researchers should always check for them) practical assessment. *Research and Evaluation*, 9.

[200] RL Somorjai, B Dolenko, and R Baumgartner. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics*, 19 12:1484–1491, 2003.

[201] U Von Luxburg. Clustering stability: an overview. *Found Trends Mach Learn*, 2.

[202] H Edelsbrunner and J Harer. *Computational Topology: an introduction*. AMS Bookstore, 2010.

[203] G Carlsson, V de Silva, and D Morozov. Zigzag Persistent Homology and Real-valued Functions. In *Proceedings of the 25th Symposium on Computational Geometry*, pages 247–256, 2009.

[204] M Carrière and S Oudot. Structure and Stability of the 1-Dimensional Mapper. *CoRR*, 2015.

[205] D Cohen-Steiner, H Edelsbrunner, and J Harer. Stability of Persistence Diagrams. *Discrete and Computational Geometry*, 37(1):103–120, 2007.

[206] D Burago, Y Burago, and S Ivanov. *A Course in Metric Geometry*. American Mathematical Society, 2001.

[207] M Carrière, S Oudot, and M Ovsjanikov. Local Signatures using Persistence Diagrams. *HAL preprint*, 2015.

[208] H Edelsbrunner, J Harer, and A Patel. Reeb Spaces of Piecewise Linear Mappings. In *Proceedings of the 24th Symposium on Computational Geometry*, pages 242–250, 2008.

[209] Dey Partha. Stein-chen method for poissson approximation. *University of Warwick*, 2013-2014.

[210] William M Rand. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850, December 1971.

[211] Kazutaka Akagi, Moustafa Sarhan, Abdel-Rahman S Sultan, Haruka Nishida, Azusa Koie, Takumi Nakayama, and Hitoshi Ueda. A biological timer in the fat body comprising Blimp-1, β-Ftz-f1 and Shade regulates pupation timing in Drosophila melanogaster. *Development*, 143(13):2410–2416, July 2016.

[212] E Tian, Ten Hagen, and Kelly G. Expression of the UDP-GalNAc : polypeptide N-acetylgalactosaminyltransferase family is spatially and temporally regulated during Drosophila development. *Glycobiology*, 16(2):83–95, February 2006.

[213] AM Snijders, S Langley, JH Mao, S Bhatnagar, KA Bjornstad, CJ Rosen, A Lo, Y Hang, EA Blakely, GH Karpen, MJ Bissell, and AJ Wyrobek. An interferon signature identified by rna-sequencing of mammary tissues varies across the estrous cycle and is predictive of metastasis-free survival. *Oncotarget*, 5:4011–4025, 2014.

[214] J Amory, R Lawler, and J Hitti. Increased tumor necrosis factor-alpha in whole blood during the luteal phase of ovulatory cycles. 49:678–82, 09 2004.

[215] H HuJun, WA Gupta, A Shidfar, D Branstetter, O Lee, D Ivancic, M Sullivan, RT Chatterton, WC Dougall, and S Khan. Rankl expression in normal and malignant breast tissue responds to progesterone and is up-regulated during the luteal phase. *Cancer Research Treatment*, 146:515–523, 2014.

[216] F Vignon, S Bardon, D Chalbos, and H Rochefort. Antiestrogenic effect of R5020, a synthetic progestin in human breast cancer cells in culture. *The Journal of Clinical Endocrinology and Metabolism*, 56(6):1124–1130, June 1983.

[217] George Sflomos, Marie Shamseddin, and Cathrin Brisken. An Ex vivo Model to Study Hormone Action in the Human Breast. *JoVE (Journal of Visualized Experiments)*, (95):e52436–e52436, January 2015.

[218] Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11:R25, March 2010.

[219] Marta Melé, Pedro G Ferreira, Ferran Reverter, David S DeLuca, Jean Monlong, Michael Sammeth, Taylor R Young, Jakob M Goldmann, Dmitri D Pervouchine, Timothy J Sullivan, Rory Johnson, Ayellet V Segrè, Sarah Djebali, Anastasia Niarchou, Fred A Wright, Tuuli Lappalainen, Miquel Calvo, Gad Getz, Emmanouil T Dermitzakis, Kristin G Ardlie, and Roderic Guigó. The human transcriptome across tissues and individuals. *Science (New York, N Y)*, 348(6235):660–665, May 2015.

[220] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)*, 57(1):289–300, 1995.

[221] Hong Hu, Jun Wang, Akash Gupta, Ali Shidfar, Daniel Branstetter, Oukseub Lee, David Ivancic, Megan Sullivan, Robert T Chatterton, William C Dougall, and Seema A Khan. RANKL expression in normal and malignant breast tissue responds to progesterone and is up-regulated during the luteal phase. *Breast Cancer Research and Treatment*, 146(3):515–523, August 2014.

[222] L A Okumu, N Forde, A G Fahey, E Fitzpatrick, J F Roche, M A Crowe, and P Lonergan. The effect of elevated progesterone and pregnancy status on mRNA expression and localisation of progesterone and oestrogen receptors in the bovine uterus. *Reproduction*, 140(1):143–153, July 2010.

[223] Alison J Camden, Maria M Szwarc, Sangappa B Chadchan, Francesco J DeMayo, Bert W O'Malley, John P Lydon, and Ramakrishna Kommagani. Growth regulation by estrogen in breast cancer 1 (GREB1) is a novel progesterone-responsive gene required for human

endometrial stromal decidualization. *Molecular Human Reproduction*, 23(9):646–653, September 2017.

[224] Marie Shamseddin. Progestins used for hormonal contraception in switzerland: study of their effects on the breast epithelium. 2018.

[225] Mikael Häggström. Reference ranges for estradiol, progesterone, luteinizing hormone and follicle-stimulating hormone during the menstrual cycle. 1, 03 2014.

[226] Reto Stricker, Raphael Eberhart, Marie-Christine Chevailler, Frank A Quinn, Paul Bischof, and René Stricker. Establishment of detailed reference values for luteinizing hormone, follicle stimulating hormone, estradiol, and progesterone during different phases of the menstrual cycle on the Abbott ARCHITECT® analyzer. *Clinical Chemistry and Laboratory Medicine (CCLM)*, 44(7):883–887, 2011.

[227] P Franchimont, C Dourcy, J J Legros, A Reuter, Y Vrindts-Gevaert, J R Van Cauwenberge, and U Gaspard. PROLACTIN LEVELS DURING THE MENSTRUAL CYCLE. *Clinical Endocrinology*, 5(6):643–650, November 1976.

[228] George Sflomos, Valerian Dormoy, Tauno Metsalu, Rachel Jeitziner, Laura Battista, Valentina Scabia, Wassim Raffoul, Jean-Francois Delaloye, Assya Treboux, Maryse Fiche, Jaak Vilo, Ayyakkannu Ayyanan, and Cathrin Brisken. A Preclinical Model for ERα-Positive Breast Cancer Points to the Epithelial Microenvironment as Determinant of Luminal Phenotype and Hormone Response. *Cancer Cell*, 29(3):407–422, March 2016.

[229] Chin-Yo Lin, Anders Ström, Vinsensius Berlian Vega, Say Li Kong, Ai Li Yeo, Jane S Thomsen, Wan Ching Chan, Balraj Doray, Dhinoth K Bangarusamy, Adaikalavan Ramasamy, Liza A Vergara, Suisheng Tang, Allen Chong, Vladimir B Bajic, Lance D Miller, Jan-Åke Gustafsson, and Edison T Liu. Discovery of estrogen receptor α target genes and response elements in breast tumor cells. *Genome Biology*, 5(9):R66, 2004.

[230] Chan HS al, et. Serine protease PRSS23 is upregulated by estrogen receptor α and associated with proliferation of breast cancer cells - PubMed - NCBI.

[231] Alison J Camden, Maria M Szwarc, Sangappa B Chadchan, Francesco J DeMayo, Bert W O'Malley, John P Lydon, and Ramakrishna Kommagani. Growth regulation by estrogen in breast cancer 1 (GREB1) is a novel progesterone-responsive gene required for human endometrial stromal decidualization. *Molecular Human Reproduction*, 23(9):646–653, September 2017.

[232] Rodrigo Fernandez-Valdivia, Atish Mukherjee, Chad J Creighton, Adam C Buser, Francesco J DeMayo, Dean P Edwards, and John P Lydon. Transcriptional Response of the Murine Mammary Gland to Acute Progesterone Exposure. *Endocrinology*, 149(12):6236–6250, December 2008.

[233] Mark E Sherman, Jonine D Figueroa, Jill E Henry, Susan E Clare, Connie Rufenbarger, and Anna Maria Storniolo. The Susan G Komen for the Cure Tissue Bank at the IU Simon Cancer Center: a unique resource for defining the "molecular histology" of the breast. *Cancer Prevention Research (Philadelphia, Pa)*, 5(4):528–535, April 2012.

[234] National Cancer Institute (U S) . *Charting the course: priorities for breast cancer research: report of the Breast Cancer Progress Review Group*. Priorities for breast cancer research, report. Breast Cancer Progress Review Group, United States, 1998.

[235] John L Young, Susan S Devesa, and Sidney J Cutler. Incidence of Cancer in United States Blacks. *Cancer Research*, 35(11 Part 2):3523–3536, November 1975.

[236] Traci N Bethea, Lynn Rosenberg, Chi-Chen Hong, Melissa A Troester, Kathryn L Lunetta, Elisa V Bandera, Pepper Schedin, Laurence N Kolonel, Andrew F Olshan, Christine B Ambrosone, and Julie R Palmer. A case-control analysis of oral contraceptive use and breast cancer subtypes in the African American Breast Cancer Epidemiology and Risk Consortium. *Breast cancer research: BCR*, 17:22, February 2015.

[237] Carol E DeSantis, Rebecca L Siegel, Ann Goding Sauer, Kimberly D Miller, Stacey A Fedewa, Kassandra I Alcaraz, and Ahmedin Jemal. Cancer statistics for African Americans, 2016: Progress and opportunities in reducing racial disparities. *CA: a cancer journal for clinicians*, 66(4):290–308, July 2016.

[238] Jennifer Keller, Alan F Schatzberg, and Mario Maj. Current Issues in the Classification of Psychotic Major Depression. *Schizophrenia Bulletin*, 33(4):877–885, July 2007.

[239] JM Chan, G Carlsson, and R Rabadana. Topology of viral evolution. *Proceedings of the National Academy of Science*, 110.

[240] JL Nielson, J Paquette, AW Liu, CF Guandique, AC Tovar, T Inoue, KA Irvine, JC Gensel, J Kloke, TC Petrossian, PY Lum, GE Carlsson, GT Manley, W Young, MS Beattie, MC Bresnahan, and AR Fergusona. Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury. *Nature Communication*, 6:8581, 2015.

[241] M Carrière, B Michel, and S Oudot. Statistical Analysis and Parameter Selection for Mapper. Manuscript, 2017.

[242] R W Hamming. Error detecting and error correcting codes. *The Bell System Technical Journal*, 29(2):147–160, April 1950.

[243] Brittany Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, and Aarti Singh. Confidence Sets for Persistence Diagrams. *The Annals of Statistics*, 42(6):2301–2339, 2014.

[244] Javier Arsuaga, Nils A Baas, Daniel DeWoskin, Hideaki Mizuno, Aleksandr Pankov, and Catherine Park. Topological analysis of gene expression arrays identifies high risk molecular subtypes in breast cancer. *Applicable Algebra in Engineering, Communication and Computing*, 23(1-2):3–15, April 2012.

[245] D DeWoskin, J Climent, I Cruz-White, M Vazquez, C Park, and J Arsuaga. Applications of computational homology to the analysis of treatment response in breast cancer patients. *Topology and its Applications*, 157(1):157–164, January 2010.

[246] Rachel Jeitziner. Package 'ttmap'. 2018.

[247] Rachel Jeitziner. User manual : Two-tier mapper: a user-independent clustering method for global gene expres- sion analysis based on topology. 2018.

[248] S Biasotti, A Cerri, P Frosini, D Giorgi, and C Landi. Multidimensional Size Functions for Shape Comparison. *Journal of Mathematical Imaging and Vision*, 32(2):161, October 2008.

[249] Silvia Biasotti, Andrea Cerri, Patrizio Frosini, and Daniela Giorgi. A new algorithm for computing the 2-dimensional matching distance between size functions. *Pattern Recognition Letters*, 32(14):1735–1746, October 2011.

[250] Silvia Biasotti, Andrea Cerri, Patrizio Frosini, and Daniela Giorgi. Approximating the 2-dimensional matching distance. 06 2018.

[251] Federico Iuricich, Sara Scaramuccia, Claudia Landi, and Leila De Floriani. Computing shape descriptors based on vector-valued functions. September 2016.

Nomenclature

$\mathrm{Dg}(\tilde{f})$ the extended persistence diagram of the Reeb graph using $\tilde{f}$, page 69

$\mathrm{DM}(X, f, \mathscr{I})$ Descriptor of $\mathrm{M}(X, f, \mathscr{I})$, page 69

$\mathrm{M}(X, f, \mathscr{I})$ Mapper of the space X with function f and cover $\mathscr{I}$, page 69

$\mathfrak{I}$ Cover in TTMap, page 63

τ total absolute deviation, page 63

$d(\mathrm{DM}(X, f, \mathscr{I}), \mathrm{DM}(Y, g, \mathscr{I}))$ Descriptor Distance, page 73

$d_b^\infty(D, D')$ Bottleneck Distance, page 31

$d_M(X, Y)$ Mismatch distance between X and Y, page 62

$Dc.T$ Disease Component of T, page 60

$Nc.T$ Normal Component of T, page 60

R the software **R**, page 3

A Adenosine, page 33

AA African American, page 118

Amyg Amygdala cell, page 122

AR androgen receptor, page 50

C Chronic Depression, page 123

C Cytosine, page 33

cDNA complementary DNA, page 34

CGH Comparative Genomic Hybridization, page 143

CHUV Centre Hospitalier Universitaire Vaudois, page 98

$\mathrm{cost}(\Gamma)$ cost of a matching, page 31

D Diestrous, page 95

DA Dopaminergic cell, page 122

DBSCAN Density-Based Spatial Clustering of Application with Noise, page 39

DNA Deoxyribonucleic Acid, page 33

DSGA Disease-Specific Genomic Analysis, page 4

E Estrous, page 95

EC Estrous cycle, page 45

EM expectation-maximisation, page 43

EP Early Psychosis, page 123

EpCAM Epithelial Cell Adhesion Molecule, page 98

ER Estrogen Receptor, page 44

FS fast-spiking (FS) neuron, page 122

G Guanosine, page 33

GLMap Global-to-Local Mapper, page 56

GMM Gaussian mixture models, page 42

HDA Hyperrectangle Deviation Assessment, page 56

HER2 erb-b2 receptor tyrosine kinase 2, page 49

HRT hormone replacement therapy, page 50

IHC immunohistochemistry, page 49

IN stem cell, page 122

KEGG Kyoto Encyclopedia of Genes and Genomes, page 97

LH Luteinizing hormone, page 50

LR Linear Regression, page 128

Mclust Model-based clustering, page 40

MDS Multidimensional Scaling, page 15

MIAME Minimum Information About a Microarray Experiment, page 34

mRNA messenger RNA, page 33

NA Not a Number, page 58

OMap Original use of Mapper, page 128

OPG osteoprotegerin, page 52

P Proestrous, page 95

PAD Progression Analysis of Disease, page 4

PC Pyramidal cell, page 122

PCA Principal Component Analysis, page 15

PCR polymerase chain reaction, page 35

PD Persistent Diagrams, page 20

PH Persistent Homology, page 2

PR Progesterone Receptor, page 44

R5020 Promegestone, page 53

RANKL Receptor Activator of Nuclear Factor κB Ligand, page 48

RI Rand Index, page 90

RNA Ribonucleic Acid, page 33

ssDNA Single stranded DNA, page 33

T Thymine, page 33

TC Termination Condition, page 39

TDA Topological Data Analysis, page 2

TMM trimmed-mean of M values, page 99

TNF Tumour Necrosis Factor, page 96

TTMap Two-Tier Mapper, page 54

UNIX Operation System, page 35

VDR vitamin D receptor, page 50

Wnt4 Wnt family member 4, page 48

Rachel Jeitziner

Chemin de la Tuilière 41 rachel.jeitziner@epfl.ch

1805 Jongny, Vaud 0041 79 715 47 03

Birth date: 23.06.1991 in Lausanne

Civil status: Single

Nationality: Swiss

Training

2014-present	PhD in Life Sciences in molecular biology
2012-2014	Master at the École Polytechnique Fédérale de Lausanne (EPFL) in mathematics
2009-2012	Bachelor at the EPFL in mathematics
2006-2009	Swiss *Maturité fédérale* in Burier (VD) with biology-chemistry as main option and physics as complementary option
2006	Swiss *Baccalauréat* in Corsier (VD) with mathematics-physics as option. Price in mathematics and in literature

Accepted Publications

Brisken C, Hess K, Jeitziner R. *Progesterone and overlooked endocrine pathways in breast cancer pathogenesis.* Endocrinology. 2015;156(10):3442–3450

George Sflomos, Valerian Dormoy, Tauno Metsalu, Rachel Jeitziner, Laura Battista, Valentina Scabia, Wassim Raffoul, Jean-Francois Delaloye, Assya Treboux, Maryse Fiche, Jaak Vilo, Ayyakkannu Ayyanan, Cathrin Brisken, *A Robust Preclinical Model for ERa Positive Breast Cancer Points to the Mammary Epithelial Microenvironment as a Critical Determinant of Luminal Phenotype and Hormone Response,* Cancer Cell, 2016, Mar 14; 29(3):407-22.

Forseen Publications

R. Jeitziner, M. Carrière, J. Rougemont, S. Oudot, K. Hess, C. Brisken, *Two-Tiers Mapper: a user-independent clustering method for global gene expression based on topology,* arXiv: 1801.01841, submitted to *Bioinformatics*.

Program developed

TTMap is an open-source R package deposited at the Bioconductor.

Conferences with oral presentations or posters

- Workshop on Applied and Computational Topology, Southampton, (30.04.2018), a talk
- Applied and Algebraic Topology 2017, Sapporo, Japan, (8-12.8.2017), a talk.
- Fifth Faculty & Staff Retreat of the Swiss Cancer Center Lausanne (9-10.11.2016), a short talk.
- Applications and Statistics of Multidimensional Persistence, EPFL, (22-26.8.2016), a poster.
- Computational algebraic topology meeting, University of Oxford, (20.06.2015), a talk.
- Computational and algebraic topology, Copenhagen, (10-14.11.2014), a poster.
- EPFL, "Journée des gymnasiens 2014", (3.2014), a talk in French for the SV faculty and for the mathematical faculty in German, and "Journée des gymnasiens 2015" (3.2015) in the same way, and

"Journée des gymnasiens 2016" (1-2.12.2016), on my developed method TTMapper (a video can be found in http://sv.epfl.ch/gymnasiens/mars2014).

Organised conference

- Young Topologists Meeting 2015 (I was one of four organizers, with 200 participants) from the 6-10 July 2015.

Other conferences

- Alpine Algebraic and Applied Topology Conference, Saas-Almagell from 18 and 19th August 2016.
- Swiss Bioinformatics Days, Biel/Bienne, Switzerland, 7-8 June 2016.
- Fourth Faculty & Staff Retreat of the Swiss Cancer Center Lausanne, from the 10th to the 11th of November 2015.
- Life Science Symposium 2015 at the EPFL, from 2nd to the 4th of September 2015.

Professional experience

2014-2016	Assistant in linear algebra and Senior Assistant in Topology I and II for Mathematician at EPFL
Summers 2006-2013	Banque Cantonale de Fribourg, summer trainee in banking: cash service, corporate actions and back office activities
2011-2013	Assistanceship and tutoring in different classes at the EPFL including Analysis I and II in German, for Civil engineering and for Mathematicians
2007-2009	Private class support in German

Informatical skills (Programming and scientific writing)

R, LaTeX, C++, Matlab, Back office

Interests

Music, flute, music theory (diplomas of end of studies at the « Association Vaudoise des Conservatoires et Ecoles de Musique ») and violin, hiking, swimming, cooking, reading, knitting.

Languages

German	Mother tongue, perfect written and oral level
French	Second mother tongue, perfect written and oral level
English	Oral level (C2) and written (C2), Advanced certificate of Cambridge
Italian	Level A1, basic knowledge