# Computational and Statistical analyses of Molecular Evolution and Demography using Large-scale Sequencing Data

THÈSE N$^{O}$ 8748 (2018)

PRÉSENTÉE LE 14 SEPTEMBRE 2018

À LA FACULTÉ DES SCIENCES DE LA VIE

UNITÉ DU PROF. DEPLANCKE

PROGRAMME DOCTORAL EN BIOTECHNOLOGIE ET GÉNIE BIOLOGIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Adamantia KAPOPOULOU

acceptée sur proposition du jury:

Prof. M. Dal Peraro, président du jury
Prof. B. Deplancke, Prof. J. D. Jensen, directeurs de thèse
Prof. T. Flatt, rapporteur
Dr M. Foll, rapporteur
Prof. J. Fellay, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Suisse
2018

# Acknowledgements

I owe my deepest acknowledgement to my thesis supervisor Prof. Jeffrey Jensen, not only for the opportunity he gave me to test myself on this exercise and the freedom to express and try my ideas and at the same time the immense support to accomplish this task. I could not dream for a better supervision combining those two aspects. Most importantly, I thank him for the motivation and the confidence he gave me to start a PhD, it is not exaggerated to say that without him I would not have thought or dare starting this adventure.

I am also deeply grateful to Dr. Stefan Laurent for the great scientific contribution to not only my thesis but also to the understanding of population genetics field. His continuous and immediate help on every aspect of my work as his infinite good mood and humor have largely contributed to the completion of this work.

I would also like to express my gratitude to my colleagues Matthieu Foll, Sebastian Matuszewski, Anna Ferrer-Admetlla, Severine Vuilleumier, and Kristen Irwin for all the scientific help they have provided me, together with their support and for making a perfect working environment.

I want to thank my co-supervisor Prof. Bart Deplancke for hosting me in his group for the last months of my thesis and my colleagues of the Deplancke lab, especially the “fly team” members Michael Frochaux, Roel Bevers, Masha Litovchenko, Virginie Braman, Sakshi Sharda, and Brian Hollis for all their support over the last months. Huge thanks to Anna-Sapfo Malaspinas and Thanassis Kousathanas for their help and scientific advises over the last months. I also thank the Bucher group members Romain Groux, René Dreos, and Giovanna Ambrosini for “adopting” me when I was alone in the office the last months of my thesis. And of course, a huge gratitude to Sophie Barret, administrative assistant of the Jensen Lab at EPFL, Geneviève Rossier from VDG lab and Sonja Bodmer, administrative assistant of the Doctoral School.

Finalement, je remercie mes parents et je dédie ce travail de thèse à Ermis qui me donne des raisons à me surpasser et à Armand qui m’aide à me surpasser.

# Abstract

Evolution can be described as the change of allele frequencies over time. Four forces - mutation, migration, genetic drift, and selection, drive this change. The aim of my thesis was to accurately estimate and differentiate the parameters governing each of these four mechanisms by utilizing various types of Next-Generation Sequencing datasets.

More specifically, in chapters 1 and 2, I focused on investigating how the past demographic history of African and European *D. melanogaster* affected its genomic polymorphism. Modern genomes of flies carry signatures of past events such as migration to new regions, adaptation to new environments, and population size changes. By studying whole genome sequences of 29 wild strains from West Africa, 14 from Sweden and comparing them with genomes from Zambia (the putative ancestral range of the species), we were able to report for the first time, colonization time of the Western part of the African continent at approximately 72k years ago. Additionally, we demonstrated the importance of gene flow between the two populations, as well as, current and past effective population sizes. Our estimations confirmed already published predictions (Current Zambian and Swedish population size, ancestral African population size). Finally, we demonstrated the importance of inversions when accounting for demographic events of *D. melanogaster*.

In chapter 3 of my thesis, I evaluated the importance of selection acting on the DNA-binding residues of the biggest family of transcription factors in the primates, namely KRAB-ZF genes. We were able to demonstrate the existence of two distinct sub-groups, based on the type of polymorphism (synonymous or not) carried by the DNA-contacting nucleotides. The two groups of genes differ by their expression breadth and intensity, as well as at the number of paralogs and orthologs and their evolutionary age. Additionally, we manually annotated the complete catalog of human KRAB-ZF genes, thereby providing a valuable resource for further investigation of this family of genes.

In conclusion, the work carried during my thesis enabled to refine the evolution and demography of *D. melanogaster* African and Northern European populations, underlying the importance of modelling migration flows between populations for

accurate estimation of split time. The second component of my thesis demonstrated the applicability of transcriptomics and epigenomics datasets to study evolution of the KRAB-ZF family. The proposed methodologies are applicable to other transcription factor gene families and our manually curated dataset is relevant to other scientists deciphering the function of these genes.

**Keywords**: demographic inference, Drosophila melanogaster, inversion polymorphisms, population genomics, colonization history, KRAB-containing zinc-finger genes, regulatory evolution, DNA-contacting residues, transcription factors, endogenous retroelements

# Résumé

L'évolution peut être décrite en tant que changement des fréquences alléliques au fil du temps. Ce changement est le produit de quatre mécanismes: mutation, migration, derive génétique et sélection. Le but de ma thèse est d'estimer les différents paramètres qui gouvernent chacun de ces quatre mécanismes, en utilisant plusieurs types de données issues du Séquençage à Haut Débit.

Les chapitres 1 et 2 traitent de l'histoire démographique de *D. melanogaster* pour les continents Africain et Européen. Ces chapitres étudient la façon dont cette démographie a influencée les différents polymorphismes génétiques. En analysant le génome des mouches contemporaines, nous pouvons distinguer les marques laissées par les événements du passé. Ces marques incluent la migration vers des nouvelles régions et la conséquente adaptation à ses nouveaux environnements, ainsi que l'expansion (croissance) de ces populations. Nous avons analysé le génome de 29 souches natives (« sauvages ») collectées en Afrique de l'Ouest et 14 souches collectées en Suède. Nous avons comparé ces génomes avec ceux de mouches en provenance de Zambie (le lieu supposé d'origine de l'espèce) et nous avons pu identifier pour la première fois la date de colonisation de l'Afrique de l'Ouest (environ 72000 années). Nous avons aussi pu démontrer l'importance des échanges génétiques entre les populations, ainsi que l'évolution des tailles des populations ancestrales et contemporaines. Nos estimations sont en accord avec d'autres études précédentes. Finalement, nous avons aussi démontré l'importance des inversions dans les études démographiques de *D. melanogaster*.

Au 3e chapitre de ma thèse, j'ai évalué l'importance de la sélection agissant sur les résidus en contact avec l'ADN de la plus large famille des facteurs de transcription des primates, contenant les gènes KRAB-ZF. Nous avons pu démontrer l'existence de deux sous-groupes, basé sur la nature de leur polymorphismes (dépendant de conséquences synonymes ou non-synonymes) des nucleotides se liant à l'ADN. Les deux groupes diffèrent par leur expression, par leur différent nombre des gènes paralogues et orthologues et par leur âge. Nous avons aussi annoté manuellement la liste complète des gènes KRAB-ZF présents au génome humain, offrant une ressource importante pour l'étude de leur fonction.

En conclusion, le travail effectué pendant ma thèse a permis d'affiner les connaissances sur l'évolution et la démographie de *D. melanogaster* présente au continent Africain et au Nord de l'Europe, en soulignant l'importance d'inclure dans les modèles démographiques la migration existant entre les populations pour estimer avec plus grande précision le temps de divergence. La deuxième composante de ma thèse démontre comment utiliser les données transcriptomiques et épigénétiques afin d'étudier l'évolution de la famille des gènes KRAB-ZF. Les méthodologies proposées sont applicables à d'autres familles de facteurs de transcription et nos données annotées peuvent être utiles à d'autres projets scientifiques étudiant la fonction de cette famille de gènes.

**Mots-clés**: Inférence démographique, *Drosophila melanogaster*, inversions, génomique des populations, doigts de zinc associés à un domaine KRAB, évolution de la régulation, résidus au contact d'ADN, facteurs de transcription, retroéléments endogènes

# Contents

Acknowledgements ..... 3

Abstract ..... 4

Résumé ..... 6

Contents ..... 8

INTRODUCTION ..... 13

Review of *Drosophila melanogaster*'s demographic history ..... 14

The KRAB-containing Zinc Finger family ..... 18

References ..... 22

CHAPTER 1 ..... 25

The demographic history of African *Drosophila melanogaster* ..... 25

Abstract ..... 25

Introduction ..... 26

Inferring Population History ..... 28

Concluding Thoughts ..... 32

MATERIALS AND METHODS ..... 32

Samples ..... 32

Inferring Population Structure ..... 33

Demographic Inference ..... 33

ACKNOWLEDGEMENTS ..... 34

REFERENCES ..... 34

Supplementary Material ..... 38

CHAPTER 2 ..... 43

Population genomics analyses of a Swedish population of *Drosophila melanogaster* push back the divergence time between tropical and temperate populations. ..... 43

Abstract ..... 44

Introduction ..... 45

MATERIALS AND METHODS ..... 46

Data collection ..... 46

Mapping pipeline ..... 47

Quality control ..... 47

Variant calling ..... 48

Bioinformatic karyotyping ..... 48

Principal Component Analysis ..... 49

Demographic analyses ..... 49

RESULTS ..... 50

Summary statistics of mapping ..... 50

Patterns of genetic variation in the Swedish sample ..... 51

Demographic modeling ..... 52

DISCUSSION ..... 58

ACKNOWLEDGEMENTS ..... 60

REFERENCES ..... 60

Data availability ..... 65

Supplementary Material ..... 66

CHAPTER 3 ..... 73

The evolution of gene expression and binding specificity of the largest transcription factor family in primates ..... 73

Abstract ..... 73

Introduction ..... 74

MATERIALS AND METHODS ..... 76

Manual curation of all human KRAB-containing Zinc-Finger (ZF) genes ..... 76

Polymorphism data ..... 77

Expression data ..... 78

Expression breadth and conservation ..... 78

Histone data ..... 78

Orthologous gene and domain annotation ..... 79

Tests for selection ..... 79

GC content ..... 80

Paralogs ..... 80

RESULTS ..... 81

Expression of orthologous KRAB-ZF genes is species-specific ..... 81

Expression breadth and expression conservation of KRAB-ZF genes ..... 83

Expression of KRAB-ZF genes correlates with polymorphism in their Zinc-Finger Binding amino acids ..... 85

Histone modification H3K9me3 on ZF-coding exon correlates with polymorphism in their Zinc-Finger Binding amino acids ..... 87

Expression breadth and expression conservation of the two groups of KRAB-ZF genes ..... 88

The newest KRAB-ZF genes are enriched for nonsynonymous SNPs in their contacting amino acids relative to older KRAB-ZF genes ..... 89

Evolutionary analysis of orthologous KRAB-ZF genes ..... 91
Discussion ..... 93
Author's contributions ..... 97
ACKNOWLEDGEMENTS ..... 98
REFERENCES ..... 98
Supplementary Material ..... 101
CONCLUSION ..... 113
Demographic History of *Drosophila melanogaster* ..... 113
Evolution of KRAB-containing Zinc Finger family ..... 114
References ..... 115
CURRICULUM VITAE ..... 117

# INTRODUCTION

As early as 2500 years ago, Greek philosophers began thinking and developing theories about the origin of the world and the evolution of species. Anaximander was one of the first philosophers expressing the idea that humans together with other terrestrial animals have evolved from another form of life, coming from the sea and having adapted to life on terrestrial earth. Two thousand years later, Charles Darwin formulated in *The Origin of Species* (1859) the theory of *Natural Selection* - that is, how traits that enhance survival and reproduction increase in frequency in a population. Almost a century later, addressing genotypic evolution rather than phenotypic evolution, Motoo Kimura described in his Neutral Theory (1968) how populations continuously evolve by the influx of new mutations and the loss of that variation via genetic drift.

With environment change, populations may adapt to new conditions by fixing mutations beneficial to their survival. By studying DNA polymorphism data from individuals from modern populations, we can infer past adaptive processes by measuring the amount of genetic variation and the frequency distribution of the alleles in the population. In this way, we can identify genomic loci conferring selective advantages (positive selection) as well as regions highly constrained by purifying selection.

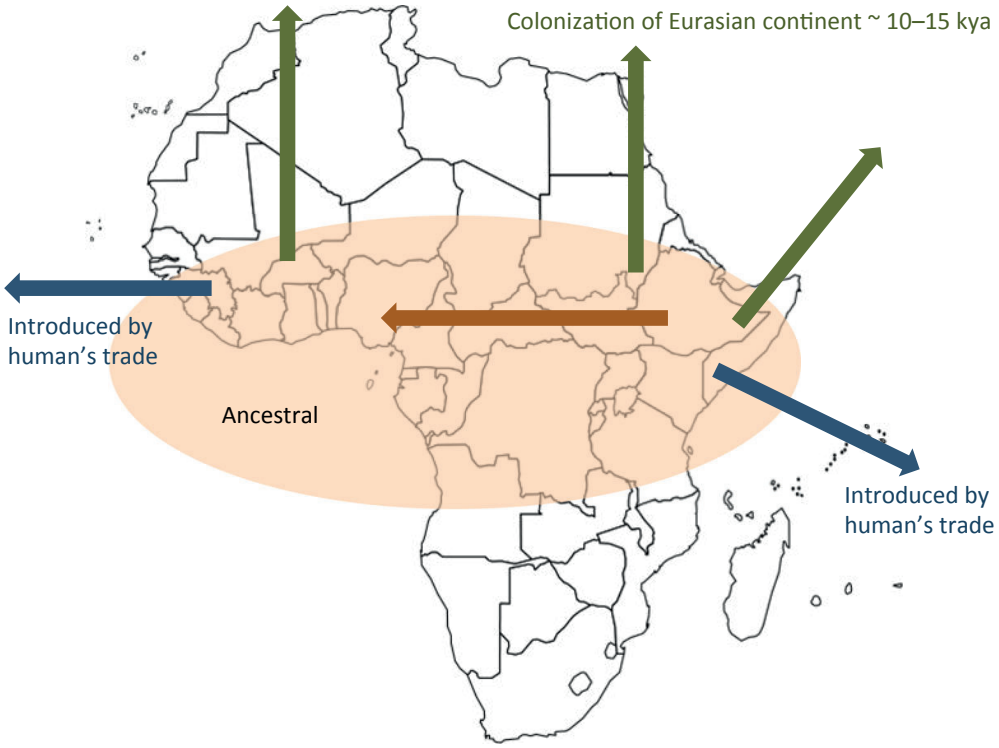
Over the past decades, many methods have been developed to identify genomic regions displaying patterns of variability characteristic of positive selection. The idea behind these tests is to compare expectations under the standard neutral model with observations from samples. When data shows deviation from neutral equilibrium assumptions, the neutral model is rejected in favor of either a demographic change in the populations such as a bottleneck or size expansion (i.e., non-equilibrium) or selective effects (i.e., non-neutral). A caveat of these tests resides in the fact that demographic events may leave similar signatures in the genome as selective events. For example, expanding populations bring an excess of rare alleles, but this can also be a sign of purifying selection. Additionally, recent admixture may result in many alleles of intermediate frequency, but this can also be the result of balancing selection. A severe

reduction in population size (bottleneck) followed by a restoration of size can create an excess of both low and intermediate frequency variants, often confounding signatures of positive selection.

Thus, the need for a correct neutral demographic model is very important as a null distribution for tests of selection. The inference of selection fully relies on an accurate understanding of the species' demographic history. One of the most studied organisms due to its late and documented worldwide colonization is *Drosophila melanogaster*, which coupled with the facility of studying its molecular mechanisms in a laboratory, makes it an ideal candidate for studying molecular evolution and adaptation to diverse environments.

## Review of *Drosophila melanogaster*'s demographic history

Lachaise and Tsacas (1974) first hypothesized the African origin of *D. melanogaster* due to the abundance of the species in the African continent. David and Capy (1988) confirmed this afro-tropical origin with the first lines of evidence provided by genetics: the extant sub-Saharan populations were more polymorphic than the rest of the world with respect to the number of alleles at various loci. Furthermore, variants found around the globe largely exist in these sub-Saharan populations. They also classified *D. melanogaster* population into three groups: Ancestral (possibly originating from the mountains of eastern equatorial Africa and then colonizing to the West sub-Saharan region), Ancient (Eurasian continent colonized after the last glaciation), and New (American continent, Australia, and oceanic islands, introduced by humans).




From these studies, as well as more recent work (Veuille et al. 2004; Baudry et al. 2004), it became clear that population structure also exists within the African continent. This complexity needs to be accounted for in demographic studies. Specifically, studies with samples originating from different mixes of African lines could potentially lead to conflicting results if the underlying structure is ignored. Another potential issue, pointed out by the aforementioned publications (Baudry et al. 2004; Veuille et al. 2004) is that cryptic populations may stem from specific genomic re-arrangements, such as inversions. These two publications agree on the importance of inversions and their impact on population structure. The study by Baudry *et al*. (2004) was the first analyzing multi-loci DNA datasets from a large number of African and non-African populations (including a sample from Madagascar). From such a large panel, they were able to exclude the Madagascar origin hypothesis and suggest an Eastern Africa origin. Regarding the supra-Saharan populations, Dieringer *et al*. (2004) using a Bayesian method, reported the existence of a distinct Northern African population carrying levels of variability similar to European populations.

For the non-African populations, the most probable scenario was formulated as a severe out-of-Africa bottleneck, just after the Neolithic revolution and the development of agriculture (minimum 6400 years ago). This bottleneck is thought to have drastically

reduced sequence variation (Baudry et al. 2004; Haddrill et al. 2005; Li and Stephan 2006; Thornton and Andolfatto 2006). Specifically, for the American continent, Caracristi and Schlötterer (2003) sampled 13 distinct populations and analyzed microsatellites on the second and X chromosomes. This in-depth study was able to confirm the hypothesis of an African admixture of the European-derived American populations. They insist, also, on the importance of using neutral markers for demographic analyses, *a contrario* with previous studies that used markers influenced by natural selection.

In an effort to disentangle natural selection from neutral demographic events, Glinka *et al*. (2003) compared a population from the ancestral range of the species (Zimbabwe) to a derived population from the Netherlands. They found multiple regions potentially under positive selection for the European sample, resulting from local adaptation to the new environment. The African population had an excess of singletons chromosome-wide, indicating a recent size expansion accompanying the transition between a full glacial to an interglacial period and the wild-to-domestic habit shift (Stephan and Li 2007). For the same purpose, Li and Stephan (2006) developed a maximum likelihood method to infer demographic changes and to simultaneously detect selective sweeps. They reported an African expansion at around 60000 years ago [26000-95000] and a split between the African and European populations followed by a bottleneck for this out-of-Africa expansion at around 15800 years ago [12000-19000]. With a small caveat that their estimated split time is in reality older than their estimation because they neglected gene flow between the two populations.

In 2011, Laurent et al. (2011) using an ABC method confirmed the previous predictions concerning the out-of-Africa timing and estimated the settlement of *D. melanogaster* in South-East Asia at approximately 5000 years ago for the third chromosome and 2500 years ago for the X chromosome (owing to the differing effective population sizes, and thus time-scaling, between these chromosomes). They postulated the existence of a European common ancestor for the Southeast Asian flies. Nevertheless, their model lacks to account for the effect of migration on genomic polymorphism.

The advances in DNA sequencing technologies have allowed Pool et al. (2012) to analyze full genome variation for more than 100 wild-derived lines from sub-Saharan Africa, thus refining the species' history. They identified the most diversity in the South-

Central of Africa, where the Zambian samples were isolated, indicating that the geographic origin of all extant populations might not be East Africa as previously believed, but rather South/Central Africa. Three years later, Lack et al. (2015) sequenced full genomes of 197 Zambian strains, completing a catalog of 623 *D. melanogaster* genomes, all analyzed in a similar way, making them comparable. This confirmed the largest pool of genetic diversity in the Zambian samples, and thus the best candidate for the ancestral range of the species.

Thus, to summarize current knowledge on the demographic history of *D. melanogaster*: the origin of the species is thought to be near Zambia (South/Central Africa); the ancestral population underwent a significant size expansion around 60k years ago [26k-95k] (Li and Stephan 2006) corresponding to a climate change and to the potential wild-to-human commensal shift of *D. melanogaster*; at least 16k years ago [12k-19k] an out-of-Africa migration brought flies to the European continent and was accompanied by a severe size reduction of the European population (Li and Stephan 2006); colonization of Southeastern Asia from European strains occurred at [700-11000] years ago (Laurent et al. 2011); introduction of flies to the American continent by human trade took place in two waves: in North America from European strains and in the Caribbean islands from West African strains (Kao et al. 2015) giving rise to a clinal pattern of African ancestry according to the latitude.

Part of my PhD studies (chapters 1 & 2) aim to improve our understanding of the demographic history of *D. melanogaster* by using the latest high quality dataset (full genomes sequenced at high coverage; Lack et al. 2016). By utilizing only neutral markers genome-wide, we can minimize the effect of natural selection in causing demographic mis-inference. Therefore, only neutral introns (Parsch et al. 2010) and four-fold degenerated sites were used for the demographic analyses.

Due to the previously limited geographic sampling from the African continent, the South/Central African origin of the species was misplaced in earlier studies, resulting in mis-inference when evaluating the demographic history of derived populations. To solve this issue, I have used the population having the highest degree of polymorphism amongst all populations (Zambia), as a base to compare against the derived populations. Taking advantage of recent developments in statistical inference, I utilized a widely used diffusion approximation approach (**∂a∂i**, Gutenkunst et al. 2009). Importantly, this method allows for gene flow between populations, a parameter lacking

from previous studies thereby leading to under-estimation for split time. Finally, I have studied the impact from large genomic inversions and provide recommendations to avoid these re-arrangements obscuring population structure in lines sampled from a unique location.

In the two demographic studies above, I have examined how the migration and random genetic drift influence the polymorphism present in the wild *Drosophila melanogaster* present-day genomes. At the last chapter of my thesis, I have foccused on how the selection acts on the nucleotide polymorphism using the fast evolving family of KRAB-ZFs.

## The KRAB-containing Zinc Finger family

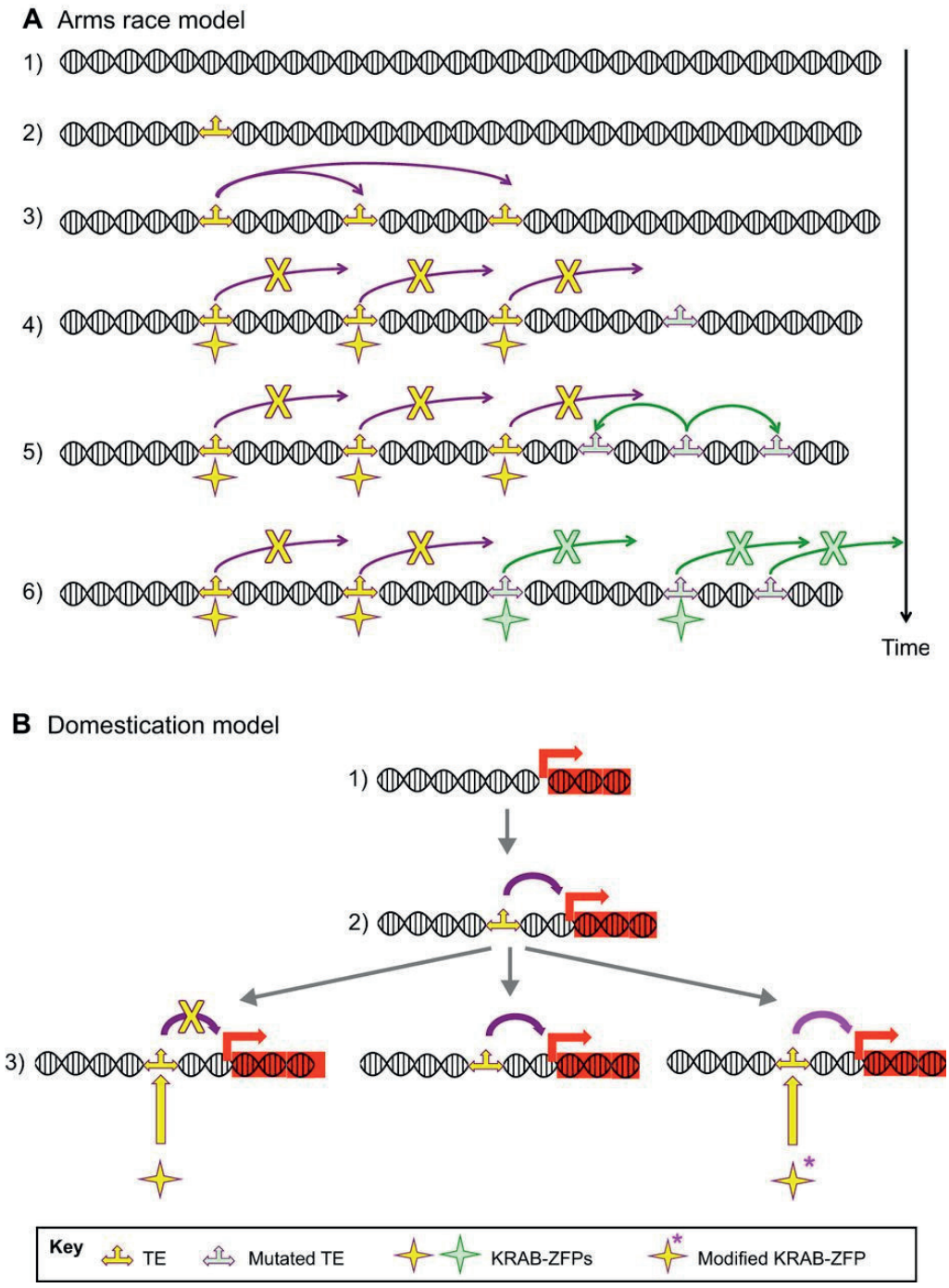
The Krüppel-associated box domain Zinc Finger (KRAB-ZF) family is the largest and the fastest growing family of transcription factors in primates (Vaquerizas et al. 2009). They emerged from an ancestral group of Zinc Fingers through repeated cycles of duplications and expanded independently in several lineages (Emerson and Thomas 2009). This rapid expansion makes them the perfect candidates for facing newly emerging retrotransposons (Thomas and Schneider 2011). Genomes are under constant re-arrangements and mobile elements participate in this evolution by their retrotransposing activity. KRAB-ZFs are in continuous arms race with the Transposable Element (TE) expansion. Their interaction can be described by two complementary mechanisms: an evolutionary arms race between KRAB-ZFs and TEs or a "domestication" of the TEs (Figure below).

The first mechanism (Panel A in following Figure) describing the interaction between the KRAB-ZFs and the TEs is described as an "arms race" and is explained as follows: when a new TE gets incorporated into the host's genome, it is expressed and retrotranscribed. The control of its invading retrotranscription is first ensured by small RNAs such as piRNAs (Russell et al. 2017). Over time, the capacity of KRAB-ZFs to continuously generate new paralogs by segmental duplication creates one of the paralogs to bind to the TE in question. Next, the KAP1 repressive complex is recruited and controls the TE expression. Some transposons accumulate mutations to escape from

this repression and the KRAB-ZF with their continuous evolution adapt to suppress the expression of the escapees. Some other TEs accumulate deleterious mutations and decay. As a consequence, the KRAB-ZF is no longer needed and can evolve toward a non-functional pseudogene or acquire a specialized function (Lupo et al. 2013).

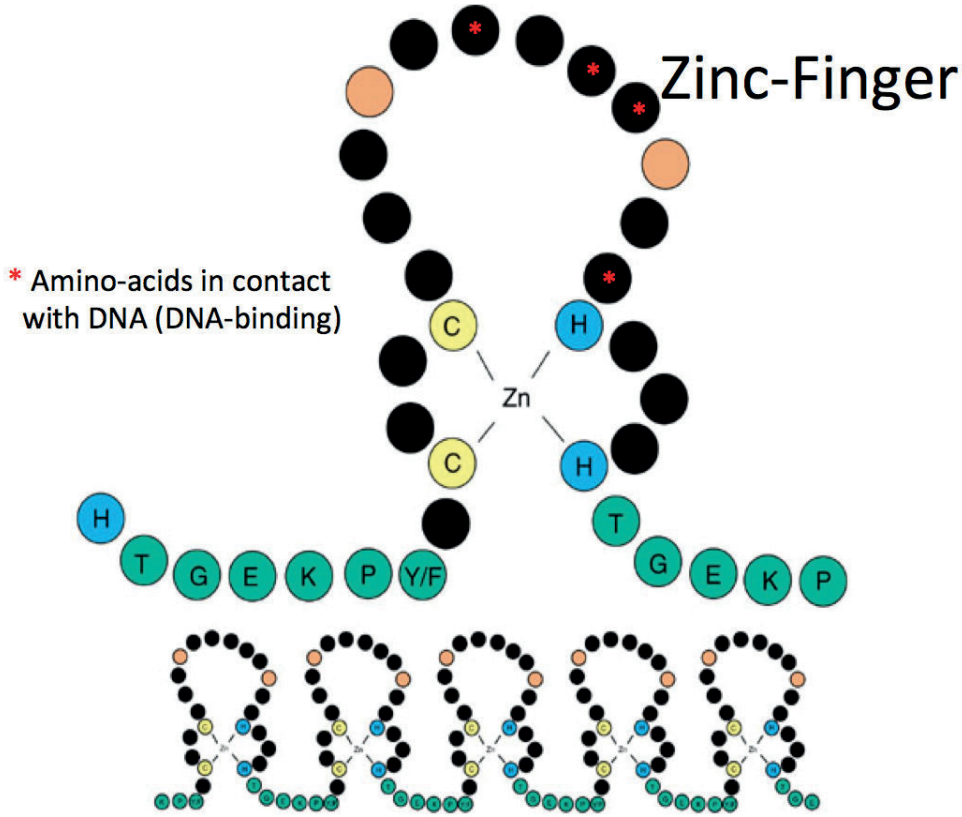
The second mechanism (Figure panel B), the "domestication" of TEs by the KRAB-ZFs can be explained by the following: A new TE is integrated in the host's genome near a functional locus, thus acquiring a function and conferring selective advantage to the host. KRAB-ZFs control its expression and retrotransposition and the pair TE/KRAB-ZF may become fixed in the population.

Figure from Ecco et al. (2017):




A KRAB domain and at least one Zinc Finger (ZF) domain compose the "canonical" KRAB-ZF genes. The number of ZFs per gene is variable and they bind to the DNA with four amino acids of their alpha helix, namely the amino acids at position -1, 2, 3, and 6, also referred to as "fingerprints" (Yang et al. 2017).

Figure adapted from Knight and Shimeld (2001):




Although, the consensus sequence of a Zinc-Finger is well characterized, automatic annotation tools may lead to some errors due to the repetitive nature of ZF proteins.

Emerson and Thomas (2009) explored evolutionary forces acting in this family making use of the paralogous genes inside species as well as orthologous genes between species. They reported that positive selection acts specifically on the DNA-binding residues, creating raw material for adaptive evolution. By contrast, KRAB-ZF genes with conserved orthologs in other species are evolutionary stable, under purifying selection, to maintain their binding specificities.

The KRAB-ZF family contributed at the formation of the human lineage and plays an important role in transcriptional regulation. Yet, little is known about *in vivo* functions of the large majority of human KRAB-ZFs.

In the third chapter of my thesis, I have investigated the evolutionary history of the KRAB-ZF family using both genetic (DNA polymorphism) and epigenetic (expression profiles) datasets, in an effort to elucidate the complex function of the KRAB-ZF transcriptional machinery.

## References


Baudry, E., B. Viginier, and M. Veuille. 2004. Non-African Populations of Drosophila melanogaster Have a Unique Origin. Mol. Biol. Evol. 21:1482–1491.

Caracristi, G., and C. Schlötterer. 2003. Genetic Differentiation Between American and European Drosophila melanogaster Populations Could Be Attributed to Admixture of African Alleles. Mol. Biol. Evol. 20:792–799.

David, J. R., and P. Capy. 1988. Genetic variation of Drosophila melanogaster natural populations. Trends Genet. 4:106–111.

Dieringer Daniel, Nolte Viola, and Schlötterer Christian. 2004. Population structure in African Drosophila melanogaster revealed by microsatellite analysis. Mol. Ecol. 14:563–573.

Ecco, G., M. Imbeault, and D. Trono. 2017. KRAB zinc finger proteins. Development 144:2719–2729.

Emerson, R. O., and J. H. Thomas. 2009. Adaptive Evolution in Zinc Finger Transcription Factors. PLOS Genet. 5:e1000325.

Glinka, S., L. Ometto, S. Mousset, W. Stephan, and D. D. Lorenzo. 2003. Demography and Natural Selection Have Shaped Genetic Variation in Drosophila melanogaster: A Multi-locus Approach. Genetics 165:1269–1278.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante. 2009. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLOS Genet. 5:e1000695.

Haddrill, P. R., K. R. Thornton, B. Charlesworth, and P. Andolfatto. 2005. Multilocus patterns of nucleotide variability and the demographic and selection history of Drosophila melanogaster populations. Genome Res. 15:790–799.

Kao Joyce Y., Zubair Asif, Salomon Matthew P., Nuzhdin Sergey V., and Campo Daniel. 2015. Population genomic analysis uncovers African and European admixture in Drosophila melanogaster populations from the south‐eastern United States and Caribbean Islands. Mol. Ecol. 24:1499–1509.

Knight, R. D., and S. M. Shimeld. 2001. Identification of conserved C2H2 zinc-finger gene families in the Bilateria. Genome Biol. 2:research0016.

Lachaise, D., and L. Tsacas. 1974. Les Drosophilidae des savanes preforestieres de la region tropicale de Lamto (Cote-d'Ivoire). Ann. Univ. Abidj. 7:153–192.

Lack, J. B., C. M. Cardeno, M. W. Crepeau, W. Taylor, R. B. Corbett-Detig, K. A. Stevens, C. H. Langley, and J. E. Pool. 2015. The Drosophila Genome Nexus: A Population Genomic Resource of 623 Drosophila melanogaster Genomes, Including 197 from a Single Ancestral Range Population. Genetics 199:1229–1241.

Lack, J. B., J. D. Lange, A. D. Tang, R. B. Corbett-Detig, and J. E. Pool. 2016. A Thousand Fly Genomes: An Expanded Drosophila Genome Nexus. Mol. Biol. Evol. msw195.

Laurent, S. J. Y., A. Werzner, L. Excoffier, and W. Stephan. 2011. Approximate Bayesian Analysis of Drosophila melanogaster Polymorphism Data Reveals a Recent Colonization of Southeast Asia. Mol. Biol. Evol. 28:2041–2051.

Li, H., and W. Stephan. 2006. Inferring the Demographic History and Rate of Adaptive Substitution in Drosophila. PLOS Genet. 2:e166.

Lupo, A., E. Cesaro, G. Montano, D. Zurlo, P. Izzo, and P. Costanzo. 2013. KRAB-Zinc Finger Proteins: A Repressor Family Displaying Multiple Biological Functions. Curr. Genomics 14:268–278.

Parsch, J., S. Novozhilov, S. S. Saminadin-Peter, K. M. Wong, and P. Andolfatto. 2010. On the Utility of Short Intron Sequences as a Reference for the Detection of Positive and Negative Selection in Drosophila. Mol. Biol. Evol. 27:1226–1234.

Pool, J. E., R. B. Corbett-Detig, R. P. Sugino, K. A. Stevens, C. M. Cardeno, M. W. Crepeau, P. Duchen, J. J. Emerson, P. Saelao, D. J. Begun, and C. H. Langley. 2012. Population Genomics of Sub-Saharan Drosophila melanogaster: African Diversity and Non-African Admixture. PLOS Genet. 8:e1003080.

Russell, S. J., L. Stalker, and J. LaMarre. 2017. PIWIs, piRNAs and Retrotransposons: Complex battles during reprogramming in gametes and early embryos. Reprod. Domest. Anim. 52:28–38.

Stephan, W., and H. Li. 2007. The recent demographic and adaptive history of *Drosophila melanogaster*. Heredity 98:65–68.

Thomas, J. H., and S. Schneider. 2011. Coevolution of retroelements and tandem zinc finger genes. Genome Res. 21:1800–1812.

Thornton, K., and P. Andolfatto. 2006. Approximate Bayesian Inference Reveals Evidence for a Recent, Severe Bottleneck in a Netherlands Population of Drosophila melanogaster. Genetics 172:1607–1619.

Vaquerizas, J. M., S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe. 2009. A census of human transcription factors: function, expression and evolution. Nat. Rev. Genet. 10:252–263.

Veuille, M., E. Baudry, M. Cobb, N. Derome, and E. Gravot. 2004. Historicity and the population genetics of *Drosophila melanogaster* and *D. simulans*. Pp. 61–70 *in* Drosophila melanogaster, Drosophila simulans: So Similar, So Different. Springer, Dordrecht.

Yang, P., Y. Wang, and T. S. Macfarlan. 2017. The Role of KRAB-ZFPs in Transposable Element Repression and Mammalian Evolution. Trends Genet. 33:871–881.

# CHAPTER 1

# The demographic history of African *Drosophila melanogaster*

*Preprint version of the article published at Genome Biology and Evolution*

Adamandia Kapopoulou[1], Susanne P. Pfeifer[1,2], Jeffrey D. Jensen[1,2], and Stefan Laurent[1,3,*]

[1] School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Switzerland

[2] School of Life Sciences, Center for Evolution & Medicine, Arizona State University, USA

[3] Department of Comparative Development and Genetics, Max Planck Institute for Plant Breeding Research, Cologne, Germany

*Keywords: demographic inference, Drosophila melanogaster, inversion polymorphisms*

## Abstract

As one of the most commonly utilized organisms in the study of local adaptation, an accurate characterization of the demographic history of *Drosophila melanogaster* remains as an important research question. This owes both to the inherent interest in characterizing the population history of this model organism, as well as to the well-established importance of an accurate null demographic model for increasing power and decreasing false positive rates in genomic scans for positive selection. While considerable attention has been afforded to this issue in non-African populations, less is known about the demographic history of African populations, including from the

ancestral range of the species. While qualitative predictions and hypotheses have previously been forwarded, we here present a quantitative model fitting of the population history characterizing both the ancestral Zambian population range as well as the subsequently colonized west African populations, which themselves served as the source of multiple non-African colonization events. These parameter estimates thus represent an important null model for future investigations in to African and non-African *D. melanogaster* populations alike.

## Introduction

Populations of *Drosophila melanogaster* span five continents, making this organism a widely utilized system to study patterns of local adaptation. Yet, this complex underlying demographic history represents unique challenges for disentangling non-neutral from non-equilibrium processes (*e.g*. Jensen et al. 2005; Teshima et al. 2006; Thornton & Jensen 2007; Pavlidis et al. 2010), and thus numerous studies have worked to better illuminate the correct demographic null model. Considerable effort has been made in understanding the species' expansion to Europe (*e.g*. Thornton & Andolfatto 2006; Li & Stephan 2006), Asia (*e.g.* Laurent et al. 2011), and the Americas (*e.g.* Kao Joyce Y. et al. 2015).

However, it is only in the past decade that African demographic history has been similarly scrutinized. In one of the earliest studies, Dieringer Daniel et al. (2004) surveyed X-chromosomal microsatellite variation from thirteen sampling locations across Africa, describing considerable population structure between North, West, and East Africa. Pool & Aquadro (2006) surveyed nucleotide variation at four 1-kb fragments in 240 individuals from sub-Saharan Africa, and described a distinct East-West geographic pattern, suggesting that western Africa may have been recently colonized from the East. Simultaneously, Li & Stephan (2006) examined dozens of non-coding X-chromosome regions from a population sampled in Zimbabwe, suggesting strong evidence of population growth. In a much larger-scale study, Pool et al. (2012) sequenced whole-genomes from 139 wild-derived strains from 22 sampling locations in sub-Saharan Africa. Based on levels of variation and $F_{st}$, they qualitatively described a fit to a model in which Zambia represents the species origin, with subsequent population

expansion, structuring and gene flow across the continent – though they concluded on the need for proper demographic model fitting in order to better elucidate these patterns. In addition, Singh et al. (2013) examined a 2Mb region in 20 individuals sampled from Uganda, also finding support for population expansion, but also suggested an associated population bottleneck out of the initial ancestral range (presumably being Zambia, hundreds of miles to the south).

Following this important work, we here focus our study on Zambia as the likely population of origin, and West Africa as a likely source of multiple widely studied non-African populations (Figure 1). We quantify the demographic history of these regions, including the timing of West African colonization, effective population sizes, and rates of gene flow (Supplementary Figure 1). Furthermore, given known segregating inversions as well as the associated difficulties that may arise if they are left unaccounted for, we have carefully curated a dataset for the purposes of inferring these underlying neutral demographic parameters, which may serve as the basis for future studies.

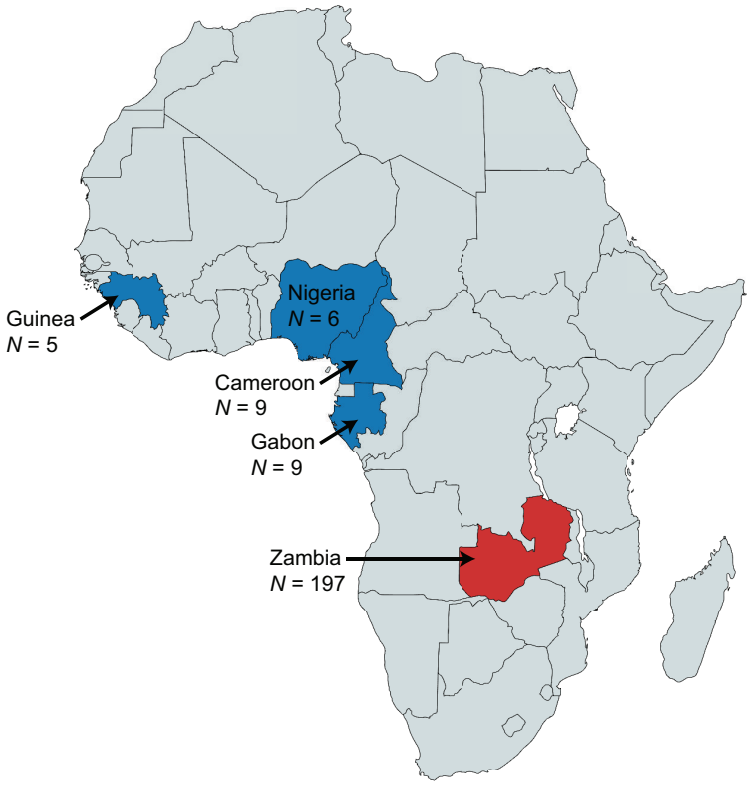



**Figure 1: Geographic distribution of the five *D. melanogaster* populations.** Samples (sample sizes indicated by *N*) were obtained from the Phase 2 (blue) and Phase 3 (red) of the Drosophila Population Genomics Project (Pool et al. 2012; Lack et al. 2015).

## Inferring Population History

The levels of genetic differentiation between individuals were assessed using a principal component analysis. The first principal component, explaining 2.7% of the variation, separates the Zambian individuals from the West African individuals, which cluster according to their sampling location (*i.e.*, Cameroon, Gabon, Guinea, and Nigeria; Supplementary Figure 2). In contrast, Zambian individuals cluster in two distinct groups based on chromosomal inversions carried by the individuals (Supplementary Figure 3). This pattern was well described by Corbett-Detig & Hartl (2012) who noted that polymorphic inversions in *D. melanogaster* affect genomic variation chromosome-wide, with trans-effects beyond the inversions' breakpoints. To avoid the confounding effects of these segregating inversions on subsequent demographic inference, 121 Zambian individuals carrying at least one inversion (*i.e.*, In2RNS, In2Lt, In3R, and In3LOk) were excluded from any further analyses.

Population structure was then assessed using an admixture model to infer individual ancestry proportions using *sNMF* (Frichot & François 2015), a statistical method to evaluate the ideal number of ancestral populations. The best-fit model (*i.e.*, the model with the lowest minimal cross-entropy) had two ancestry components (Figure 2a), strongly supporting the division of individuals from Zambian and West African populations, with evidence of admixture between them (Figure 2b). Principal component analysis confirms the two population clusters inferred by *sNMF*, with no additional sub-genetic stratification of the Zambian individuals (Figure 2c).

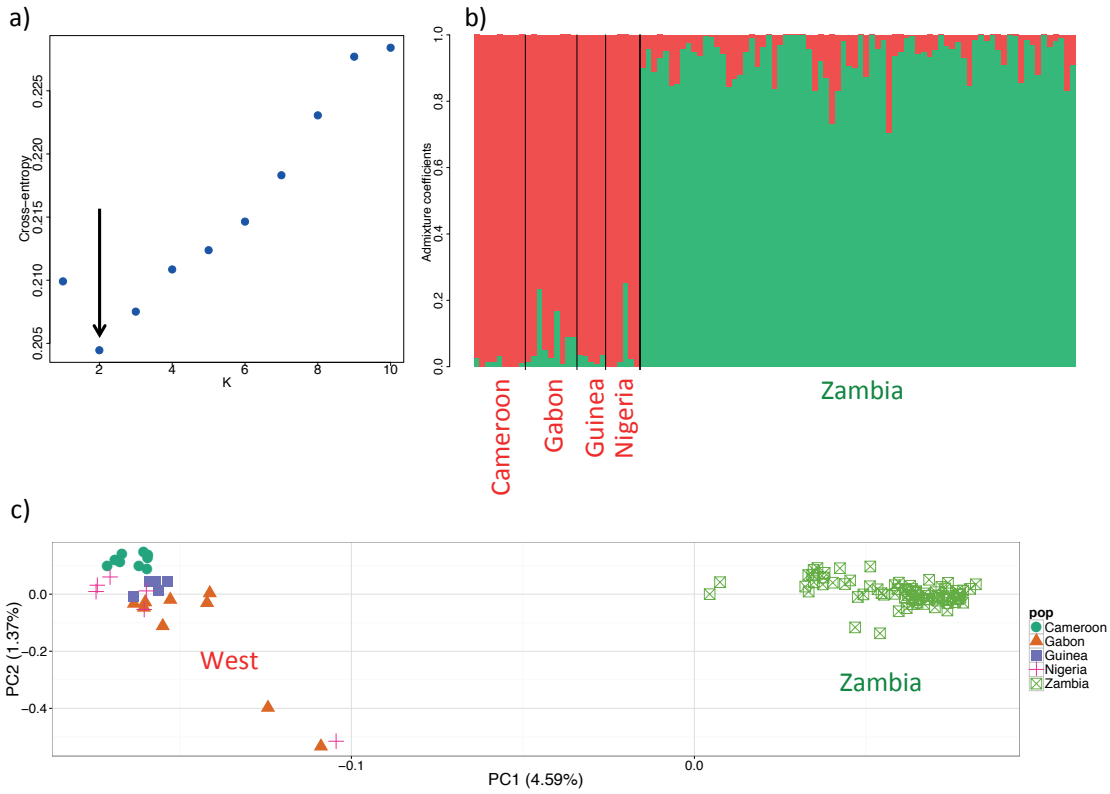



**Figure 2: Genetic structure of African D. melanogaster populations.** a) The number of K ancestry components best explaining the data was assessed by calculating the cross-entropy corresponding to the model. The best-fit model (i.e., the model with the lowest minimal cross-entropy) had two ancestry components (K=2). b) Individual admixture proportions. c) Principal component analysis (symbols correspond to individuals from different populations; green square: Zambia (N=76 individuals which do not carry the chromosome arm's specific inversion); green circle: Cameroon (N=9); orange triangle: Gabon (N=9); purple square: Guinea (N=5); red cross: Nigeria (N=6)). Data was thinned to prune for linkage, excluding SNPs with an r2>0.2 within a 50 SNP window. Percentages indicate the variance explain by each principle component.

Given the observed population structure, the demographic history of Zambian and West African populations was investigated using six different two-population demographic models, allowing for both size changes as well as gene flow among the populations. Three of the six models assumed that populations remained at a constant size with either no gene flow, symmetric migration, or asymmetric migration between them (Supplementary Figure 1). To account for the fact that West African populations exhibit lower nucleotide diversity levels than populations from south-central Africa ($\pi$ = 0.0086 in Zambia, $\pi$ = 0.0077 in West Africa; and see Pool et al. 2012; Lack et al. 2015),

suggesting a potential population bottleneck during their recent colonization from the ancestral range (Haddrill et al. 2005), the remaining three models allowed for population size changes (Supplementary Figure 1). The demographic model best fitting the data (Figure 3; Supplementary Table 1) inferred exponential growth for both the Zambian and West African populations after their split around 70kya, with on-going gene flow. In addition, the parameter estimates obtained for the ancestral and present effective population sizes ($N_e$(anc) = 1,525,061 (95% CI: 1,498,713 - 1,562,754); $N_e$(Zambia) = 3,160,475 (95% CI: 2,933,313 - 3,447,248)) reiterate the higher levels of variation observed in the putative ancestral range of the species.

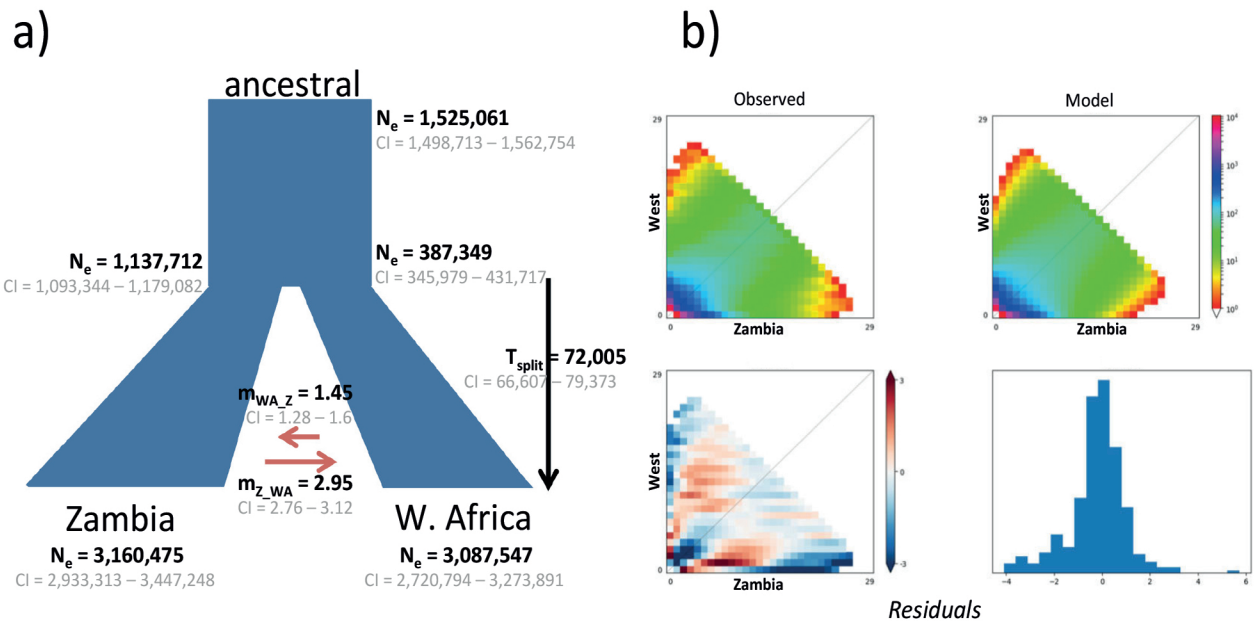



**Figure 3: Parameter estimates inferred by ∂a∂i under the best demographic model.**

a) At time $T_{split}$, the ancestral population splits into two distinct populations, which grow exponentially with asymmetric migration ($m$) between them. The time of the split ($T_{split}$) was estimated in generation times, which were converted to years, assuming ten generations per year (Laurent et al. 2011). Effective population sizes ($N_e$) for the ancestral, West African, and Zambian populations were directly estimated by fixing the mutation rate ($\mu$) to $1.3 \times 10^{-9}$ per base pair per generation (Laurent et al. 2011). 95% confidence intervals (CI) were calculated for each parameter estimate by generating 150 parametric bootstrap replicates of the best model. Note that the mode of the bootstrapped parameter estimates corresponds approximately to the obtained maximum likelihood value estimate. b) Comparison of Joint SFS for the observed data (left) and the best model (right). Below are shown the residuals of the model.

While the specific parameter values inferred are of particular importance for explicitly modelling an appropriate demographic null in future studies, and represent the first estimates of split times between the ancestral range and West Africa, the qualitative patterns are largely consistent with previous supposition. Namely, the

estimated ancestral split times (Li & Stephan 2006), population structure (Pool & Aquadro 2006), and effective population sizes (Laurent et al. 2011), as well as the underlying growth and colonization models themselves (Pool et al. 2012), are all largely in agreement with previous studies.

## Concluding Thoughts

In concordance with Corbett-Detig & Hartl (2012), we have demonstrated the ability of inversions to create significant sub-structure within a single population sampled from a single location, potentially confounding downstream demographic inference. Indeed, we find that even when polymorphisms within the inversion breakpoints were not considered in the analysis, the signature persists and is visible when analyzing other markers located on the same chromosomal arm (Supplementary Figure 3). By removing these individuals from the analysis, and by carefully curating the dataset for neutral sites, we have quantified the demographic histories characterizing these sampling locations. We find evidence for strong growth in populations inhabiting both regions, consistent structure separating West Africa from Zambia, as well as evidence for on-going gene flow particularly in the direction of south/central to west. Thus, this well-fit non-equilibrium demographic model of both the ancestral range of the species as well as the source population of subsequent non-African colonization events, represents a uniquely appropriate null model for future investigations pertaining to the demographic and adaptive histories of both African and non-African populations of *D. melanogaster*.

## MATERIALS AND METHODS

### Samples

Publicly available whole-genome sequence data from haploid *D. melanogaster* embryos originating from Guinea (*N*=5), Nigeria (*N*=6), Cameroon (*N*=9), Gabon (*N*=9), as well as from Zambia (*N*=197) was obtained from the Phase 2 and Phase 3 of the Drosophila Population Genomics Project (DPGP) (Pool et al. 2012; Lack et al. 2015,

2016), respectively (Figure 1). Specifically, genomes previously aligned to a common *D. melanogaster* reference sequence were downloaded from the Drosophila Genome Nexus (DGN) (Lack et al. 2015, 2016) and variants on both arms of chromosome 2 (*i.e.*, chr2L and chr2R) and chromosome 3 (*i.e.*, chr3L and chr3R) were identified using the SNP-sites C program (Page et al. 2016).

As chromosomal inversions may be targeted by natural selection in *D. melanogaster* (Corbett-Detig & Hartl 2012), known inversions were excluded from all demographic analyses (information on inversion breakpoints was obtained from the DGN (Lack et al. 2015; http://www.johnpool.net/Updated_Inversions.xls). To further minimize the confounding effects of linked selection on demographic inference, the dataset was limited to putatively neutral regions of the genome, including four-fold synonymous degenerate sites (Grenier et al. 2015) as well as the 8$^{th}$ to the 30$^{th}$ base of introns smaller than 65bp (Parsch et al. 2010). The resulting dataset contained 82149 variants.

**Inferring Population Structure**

Population structure was investigated using two methods, which cluster individuals based on their genetic similarity using a set of independent SNPs (*i.e.*, SNPs with an $r^2$>0.2 within a 50 SNP window were excluded from the dataset using PLINK v1.07 (Purcell et al. 2007)). Evidence of population structure was assessed using both a principal component analysis (PCA) as well as the *sNMF* function implemented in the R package LEA v2.0.0 (Frichot & François 2015). The latter implements an admixture model (Pritchard et al. 2000; Patterson et al. 2006) which uses sparse non-negative matrix factorization to infer individual ancestry proportions based on *K* potential components. Using a cross-validation technique, *K* values ranging from 1 to 10 were examined, and, following (Frichot et al. 2014), the best *K* was selected to minimize the cross entropy.

**Demographic Inference**

The demographic history of south-western African *D. melanogaster* populations was inferred from the distribution of minor allele frequencies (*i.e.*, the folded joint site frequency spectrum) obtained from the putatively neutral segregating sites using ∂a∂i

1.7.0 (Gutenkunst et al. 2009), a diffusion approximation method. Given the genetic differentiation between populations, six different two-population scenarios (corresponding to samples originating from West Africa - *i.e.*, Guinea, Nigeria, Cameroon, and Gabon, as well as Zambia) were tested, allowing for both population size changes as well as gene flow among the populations (Supplementary Figure 1). Thereby, gene flow was modelled either as symmetric or asymmetric, and considered only between the time of the population split and the present.

For every demographic model, 10 independent runs were performed using different starting points and the parameter estimates for the best run (*i.e.*, the estimation with the highest likelihood) reported. 95% confidence intervals (CI) were calculated for each parameter estimate by generating 150 parametric bootstrap replicates of the best model. Effective population sizes ($N_e$) were directly estimated by fixing the mutation rate ($\mu$) to $1.3 \times 10^{-9}$ per base pair per generation (Laurent et al. 2011). Generation times were converted to years, assuming ten generations per year (Laurent et al. 2011). The best-fitting demographic model was selected based on the Akaike's information criterion (AIC) score (Akaike 1974).

## ACKNOWLEDGEMENTS

We thank Roman Arguello for helpful discussions and for providing the coordinates of short introns and four-fold degenerate coding sites for the neutral set of loci. We also thank Athanasios Kousathanas and Anna-Sapfo Malaspinas for sharing their population genetics and statistical knowledge with us. This work was supported by grants from the Swiss National Science Foundation and the European Research Council to JDJ.

## REFERENCES

Akaike H. 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control. 19:716–723. doi: 10.1109/TAC.1974.1100705.

Corbett-Detig RB, Hartl DL. 2012. Population Genomics of Inversion Polymorphisms in Drosophila melanogaster. PLOS Genet. 8:e1003056. doi: 10.1371/journal.pgen.1003056.

Dieringer Daniel, Nolte Viola, Schlötterer Christian. 2004. Population structure in African Drosophila melanogaster revealed by microsatellite analysis. Mol. Ecol. 14:563–573. doi: 10.1111/j.1365-294X.2004.02422.x.

Frichot E, François O. 2015. LEA : An R package for landscape and ecological association studies O'Meara, B, editor. Methods Ecol. Evol. 6:925–929. doi: 10.1111/2041-210X.12382.

Frichot E, Mathieu F, Trouillon T, Bouchard G, François O. 2014. Fast and Efficient Estimation of Individual Ancestry Coefficients. Genetics. 196:973–983. doi: 10.1534/genetics.113.160572.

Grenier JK et al. 2015. Global Diversity Lines - A Five-Continent Reference Panel of Sequenced Drosophila melanogaster Strains. G3 Genes Genomes Genet. g3.114.015883. doi: 10.1534/g3.114.015883.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLOS Genet. 5:e1000695. doi: 10.1371/journal.pgen.1000695.

Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P. 2005. Multilocus patterns of nucleotide variability and the demographic and selection history of Drosophila melanogaster populations. Genome Res. 15:790–799. doi: 10.1101/gr.3541005.

Jensen JD, Kim Y, DuMont VB, Aquadro CF, Bustamante CD. 2005. Distinguishing Between Selective Sweeps and Demography Using DNA Polymorphism Data. Genetics. 170:1401–1410. doi: 10.1534/genetics.104.038224.

Kao Joyce Y., Zubair Asif, Salomon Matthew P., Nuzhdin Sergey V., Campo Daniel. 2015. Population genomic analysis uncovers African and European admixture in Drosophila melanogaster populations from the south-eastern United States and Caribbean Islands. Mol. Ecol. 24:1499–1509. doi: 10.1111/mec.13137.

Lack JB et al. 2015. The Drosophila Genome Nexus: A Population Genomic Resource of 623 Drosophila melanogaster Genomes, Including 197 from a Single Ancestral Range Population. Genetics. 199:1229–1241. doi: 10.1534/genetics.115.174664.

Lack JB, Lange JD, Tang AD, Corbett-Detig RB, Pool JE. 2016. A Thousand Fly Genomes: An Expanded Drosophila Genome Nexus. Mol. Biol. Evol. msw195. doi: 10.1093/molbev/msw195.

Laurent SJY, Werzner A, Excoffier L, Stephan W. 2011. Approximate Bayesian Analysis of Drosophila melanogaster Polymorphism Data Reveals a Recent Colonization of Southeast Asia. Mol. Biol. Evol. 28:2041–2051. doi: 10.1093/molbev/msr031.

Li H, Stephan W. 2006. Inferring the Demographic History and Rate of Adaptive Substitution in Drosophila. PLOS Genet. 2:e166. doi: 10.1371/journal.pgen.0020166.

Page AJ et al. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb. Genomics. 2. doi: 10.1099/mgen.0.000056.

Parsch J, Novozhilov S, Saminadin-Peter SS, Wong KM, Andolfatto P. 2010. On the Utility of Short Intron Sequences as a Reference for the Detection of Positive and Negative Selection in Drosophila. Mol. Biol. Evol. 27:1226–1234. doi: 10.1093/molbev/msq046.

Patterson N, Price AL, Reich D. 2006. Population Structure and Eigenanalysis. PLoS Genet. 2. doi: 10.1371/journal.pgen.0020190.

Pavlidis P, Jensen JD, Stephan W. 2010. Searching for Footprints of Positive Selection in Whole-Genome SNP Data From Nonequilibrium Populations. Genetics. 185:907–922. doi: 10.1534/genetics.110.116459.

Pool JE et al. 2012. Population Genomics of Sub-Saharan Drosophila melanogaster: African Diversity and Non-African Admixture. PLOS Genet. 8:e1003080. doi: 10.1371/journal.pgen.1003080.

Pool JE, Aquadro CF. 2006. History and Structure of Sub-Saharan Populations of Drosophila melanogaster. Genetics. 174:915–929. doi: 10.1534/genetics.106.058693.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of Population Structure Using Multilocus Genotype Data. Genetics. 155:945–959.

Purcell S et al. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. 81:559–575. doi: 10.1086/519795.

Singh ND, Jensen JD, Clark AG, Aquadro CF. 2013. Inferences of Demography and Selection in an African Population of Drosophila melanogaster. Genetics. 193:215–228. doi: 10.1534/genetics.112.145318.

Teshima KM, Coop G, Przeworski M. 2006. How reliable are empirical genomic scans for selective sweeps? Genome Res. 16:702–712. doi: 10.1101/gr.5105206.

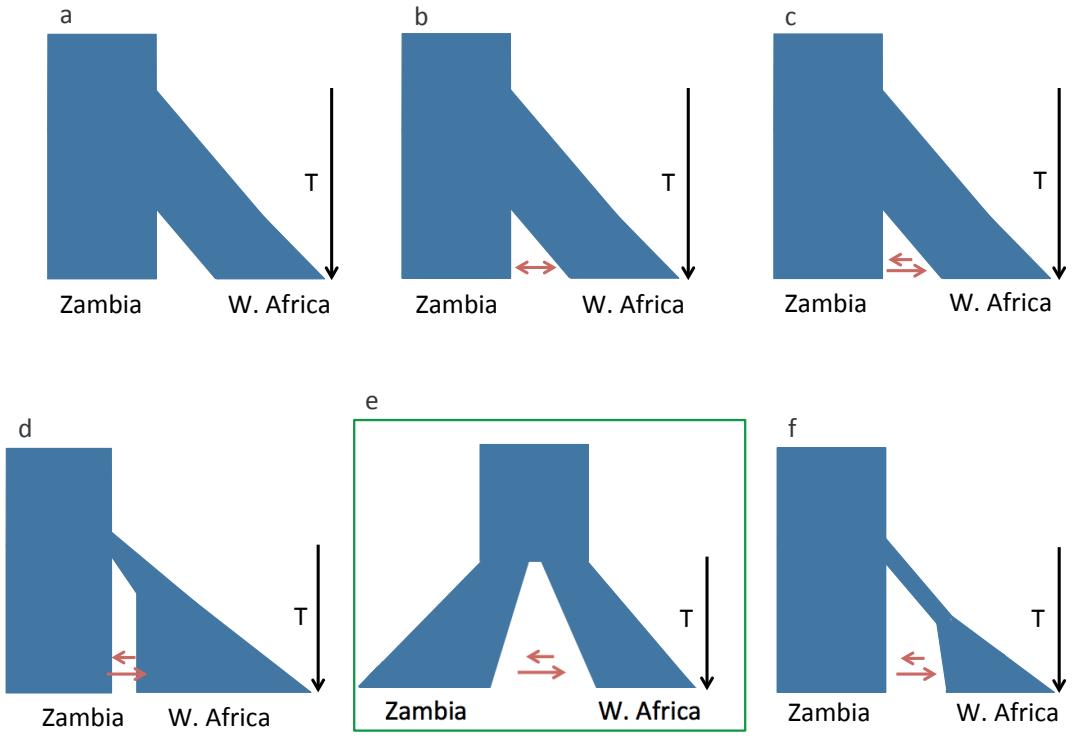
Thornton K, Andolfatto P. 2006. Approximate Bayesian Inference Reveals Evidence for a Recent, Severe Bottleneck in a Netherlands Population of Drosophila melanogaster. Genetics. 172:1607–1619. doi: 10.1534/genetics.105.048223.

Thornton KR, Jensen JD. 2007. Controlling the False-Positive Rate in Multilocus Genome Scans for Selection. Genetics. 175:737–750. doi: 10.1534/genetics.106.064642.

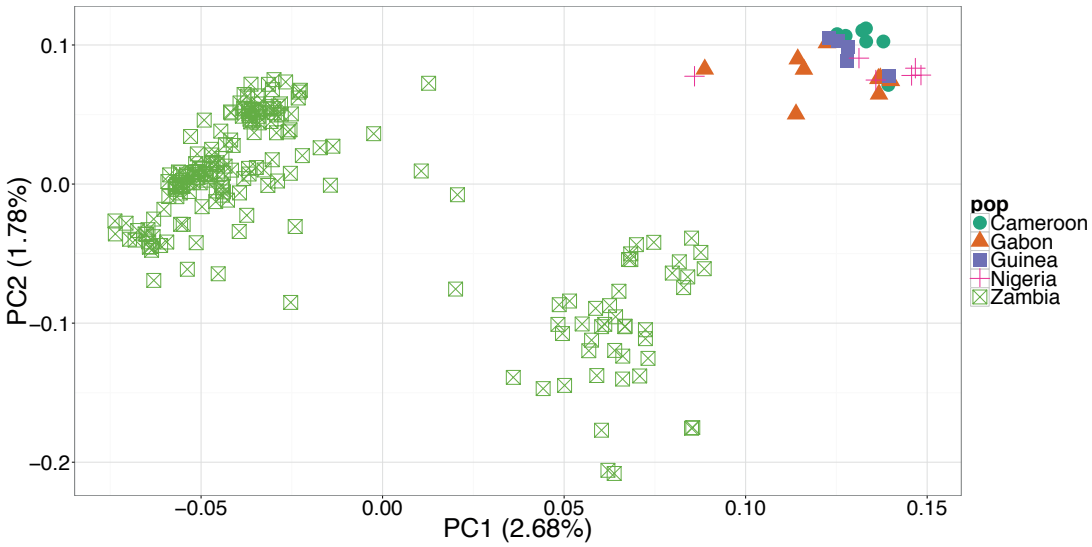
# Supplementary Material

| Model | MCL | AIC | Θ | $N_e$(anc) | $N_e$(WA) | $N_e$(Z) | $T_{split}$ | $m_{WA\_Z}$ | $m_{Z\_WA}$ | $N_e$(WA)split | $N_e$(Z)split |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Exponential growth (Z + WA) Asymmetric migration | -1,457 | 2,928 | 8,012 | 1,525,061 | 3,087,547 | 3,160,475 | 72,005 | 1.45 | 2.95 | 387,349 | 1,137,712 |
| Bottleneck (WA) Constant size (Z) Asymmetric migration | -1,508 | 3,030 | 8,029 | 1,528,297 | 3,233,219 | 2,426,431 | 64,957 | 1.67 | 2.71 | 339,129 | NA |
| Exponential growth (WA) Constant size (Z) Asymmetric migration | -1,802 | 3,618 | 7,438 | 1,415,802 | 3,519,570 | 2,157,781 | 72,558 | 0.85 | 3.09 | 1,899,511 | NA |
| Constant size (Z + WA) Asymmetric migration | -1,855 | 3,722 | 7,562 | 1,439,405 | 1,617,416 | 2,181,836 | 71,524 | 1.04 | 3.08 | NA | NA |
| Constant size (Z + WA) Symmetric migration | -1,930 | 3,870 | 7,612 | 1,448,922 | 1,426,787 | 2,480,221 | 66,022 | 2 | 2 | NA | NA |
| Constant size (Z + WA) No migration | -2,495 | 4,998 | 8,407 | 1,600,248 | 1,823,339 | 3,296,719 | 34,036 | 0 | 0 | NA | NA |

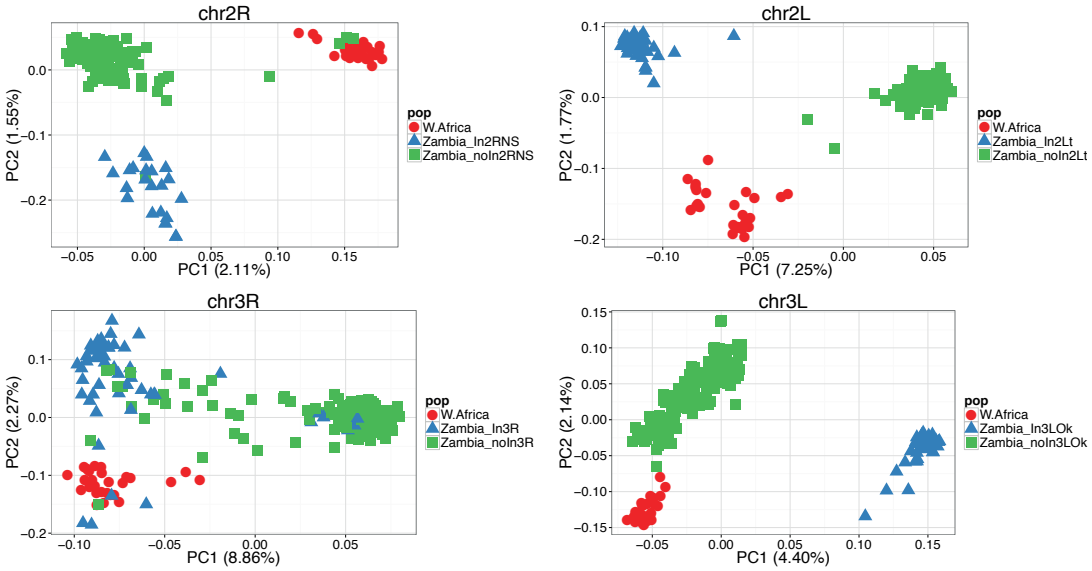
**Supplementary Table 1**: Parameter estimates for the best run (*i.e.*, the estimation with the highest maximum composite likelihood (MCL)) for each of the six two-population demographic models tested (Supplementary Figure 1). Results are ordered based on Akaike's information criterion (AIC) score (Akaike 1974), with the best-fitting demographic model being displayed on the top of the table. Effective population sizes ($N_e$) for the ancestral (anc), West African (WA), and Zambian (Z) populations were directly estimated by fixing the mutation rate ($\mu$) to $1.3 \times 10^{-9}$ per base pair per generation (Laurent et al. 2011). The time of the split ($T_{split}$) was estimated in generation times, which were converted to years, assuming ten generations per year (Laurent et al. 2011). Genetic diversity, described by Watterson's estimate Θ, was estimated together with the other parameters from the software ∂a∂i (Gutenkunst et al. 2009).




**Supplementary Figure 1**: Topologies of the six two-population demographic models tested, with populations corresponding to Zambia ($N$=197) and West Africa (*i.e.*, Guinea ($N$=5), Nigeria ($N$=6), Cameroon ($N$=9), and Gabon ($N$=9)). (top) At time $T$, the ancestral population splits into two distinct populations which remain at a constant size with (a) no gene flow, (b) symmetric migration, and (c) asymmetric migration between them. (bottom) At time $T$, the ancestral population splits into two distinct populations with asymmetric migration between them. (d) The Zambian population remains at a constant size while the West African population grows exponentially. (e) The two populations grow exponentially. (f) The Zambian population remains at a constant size while the West African population experiences a bottleneck before recovering to its current size. The best-fitting demographic model is framed by a green box.




**Supplementary Figure 2**: Principal component analysis (symbols correspond to individuals from different populations; green square: Zambia (*N*=197); green circle: Cameroon (*N*=9); orange triangle: Gabon (*N*=9); purple square: Guinea (*N*=5); red cross: Nigeria (*N*=6)). Data was thinned to prune for linkage, excluding SNPs with an $r^2$>0.2 within a 50 SNP window. Percentages indicate the variance explained by each principle component.




**Supplementary Figure 3**: Principal component analysis of individuals from West Africa (*i.e.*, Cameroon, Gabon, Guinea, and Nigeria; red circle; *N*=29) and Zambia (coloured according to their inversion-carrier status; blue triangle: individual carries the chromosome arm's specific inversion (*N*=121); green square: individual does not carry the chromosome arm's specific inversion (*N*=76)) stratified by chromosomal arms (*i.e.*, chr2L, chr2R, chr3L, and chr3R). Note that SNPs within known inversions were excluded from the analysis (see "Material and Methods" suggesting that polymorphic inversions in *D. melanogaster* affect genomic variation chromosome-wide (as previously noted by Corbett-Detig and Hartl 2012)).

# CHAPTER 2

# Population genomics analyses of a Swedish population of *Drosophila melanogaster* push back the divergence time between tropical and temperate populations.

*Manuscript in preparation. In this chapter, I conducted the totality of the demographic analysis with dadi and participated in the writing of the manuscript.*

Running title: Population genomics of Swedish *Drosophila melanogaster*

Adamandia Kapopoulou[*1], Martin Kapun[*2], Pavlos Pavlidis[3], Bjorn Pieper[4], Ricardo Wilches[5], Wolfgang Stephan[6], Stefan Laurent[§4]

* equal contributions

§ corresponding author

1: School of Life Sciences, École Polytechnique Fédérale de Lausanne, Station 19, CH-1015 Lausanne, Switzerland

2: Université de Fribourg, Département de Biologie, CH-1700 Fribourg Switzerland

3: Institute of Computer Science, Foundation for Research and Technology-Hellas, Crete, Greece

4: Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829 Köln, Germany

5: Section of Evolutionary Biology, Department of Biology II, University of Munich, Munich, Germany

6: Leibniz Institute for Evolution and Biodiversity, Invalidenstr. 43, 10115 Berlin, Germany

## Abstract

Natural populations of the fruit fly *Drosophila melanogaster* have been used extensively as a model system to investigate the effect of neutral and selective processes on genetic variation. The species expanded outside its Afrotropical ancestral range during the last glacial period and numerous studies have focused on identifying molecular adaptations associated with the colonization of northern habitats. The sequencing of many genomes from African and non-African natural populations has facilitated the analysis of the interplay between adaptive and demographic processes. However, most of the non-African sequenced material has been sampled from American and Australian populations that have been introduced within the last hundred years following recent human dispersal and are also affected by recent genetic admixture with African populations. Northern European populations, at the contrary, are expected to be older and less affected by complex admixture patterns and are therefore more appropriate to investigate neutral and adaptive processes. Here we present a new dataset consisting of 14 fully sequenced haploid genomes sampled from a natural population in Umeå, Sweden. We co-analyzed this new data with an African population to compare the likelihood of several competing demographic scenarios for European and African populations. We show that allowing for gene flow between populations in neutral demographic models leads to a significantly better fit to the data and strongly affects estimates of the divergence time and of the size of the bottleneck in the European population.

**Keywords:** *Drosophila melanogaster*, population genomics, colonization history, demography, neutral processes, demographic modeling

## Introduction

*Drosophila melanogaster* originated in sub-Saharan Africa where it diverged from its sister species *Drosophila simulans* approximately 2.3 million years ago (David & Capy, 1988). Accordingly, South and East African populations display genetic diversity patterns closer to mutation-drift expectations compared to western African and non-African populations, providing further evidence that this geographic area represents the ancestral range of the species (David & Capy, 1988; Haddrill et al., 2005; Veuille et al., 2004). Previous genetic analyses of European and Asian samples indicated that non-African populations started expanding beyond their ancestral range around 13,000 years ago, eventually colonizing large areas in Europe and Asia (Laurent et al., 2011; Li & Stephan, 2006). By contrast, the introduction of the species in the Americas and Australia is very recent (couple of hundred years) and has been witnessed and documented by early entomologists (reviewed in Keller, 2007). Interestingly, demographic analyses of a North-American and Australian populations revealed significant African ancestry (between 15 and 40%) in a dominantly European background (Bergland et al., 2016; Caracristi & Schlotterer, 2003; Duchen et al., 2013; Kao et al., 2015).

Natural populations of *D. melanogaster* have also been used extensively to study the effect of positive and negative selection on functional and linked neutral variants (reviewed in Casillas and Barbadilla, 2017; Charlesworth, 2012; Sella et al., 2009), providing estimates for the rate of adaptive events, the magnitude of the fitness effects of beneficial mutations, and identifying genes displaying molecular signatures of hitchhiking events. However, these studies also highlighted the necessity, and difficulty, of considering jointly the effect of positive selection, background selection, and demographic processes (Elyashiv et al., 2016). Studying the joint effect of neutral and selective forces on genetic variation has been facilitated by the recent sequencing of large numbers of complete genomes from natural populations (Grenier et al., 2015; Lack et al., 2015; Lack et al., 2016; Langley et al., 2012; Mackay et al., 2012; Pool et al., 2012).

However, most non-African full-genome datasets have been obtained from new world populations implying that analyses of this material must deal with the additional complexity of recent genome-wide admixture. A small number of European and Asian samples have been sequenced recently (Grenier et al., 2015), but the nature of the sequenced biological material (inbred lines) does not allow obtaining phased data.

Here, we present a new genomic dataset consisting of 14 fully sequenced haploid genomes sampled from a Swedish population. We describe patterns of genetic diversity and compare these to previously available data from a Zambian population located in the ancestral range of the species. We use this new dataset to re-visit different competing hypotheses concerning the demographic history of European populations. We show that accounting for historical gene flow in demographic models of European and African populations significantly improves the fit to the data compared to previously published model and that, as a consequence, the estimate for the divergence time between African and non-African gene pools is older than previously reported.

## MATERIALS AND METHODS

### Data collection

A total of 96 inseminated female *D. melanogaster* were sampled in the locality of Umeå in northern Sweden in August 2012. Then full-sibling mating was performed for 10 generations, which produced 80 inbred lines. Out of these, 20 lines were randomly selected from which haploid embryos were generated following the protocol described by (Langley et al., 2011). Standard genomic libraries were constructed using up to 10 µg (~200 ng/µl) of DNA. Library construction and sequencing of one haploid embryo for each of the 20 haploid-embryo lines were carried out on an Illumina HiSeq 2000 sequencer at GATC Biotech (Konstanz, Germany). In addition to the newly established and sequenced inbred lines from Umeå/Sweden, we randomly chose 10 lines not carrying the chromosomal inversion *In(2L)t* from the DPGP3 dataset. They were collected in Siavonga/Zambia in July 2010 and sequenced as haploid embryos similar to our data. Since four of the Swedish lines carried the chromosomal inversion *In(2L)t*, we additionally chose four lines at random from Zambian strains that also carried *In(2L)t* to

match the number and distribution of inversion karyotypes in our Swedish dataset (see Table S1).

**Mapping pipeline**

Prior to mapping, we tested raw read libraries in FASTQ format for base quality, residual sequencing adapter sequences and other overrepresented sequences with FASTQC (v0.10.1; http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). We trimmed both the 5' and 3' end of each read for a minimum base quality ≥ 18 and only retained reads with a minimum sequence length ≥ 75bp using cutadapt (v 1.8.3 Martin, 2011). We used bbmap (v 35.50 Bushnell, 2017) with default parameters to map intact read pairs, where both reads fulfilled all quality criteria, against a compound reference consisting of the genomes from *D. melanogaster* (v6.12) and genomes from other common pro- and eukaryotic symbionts including *Saccharomyces cerevisiae* (GCF_000146045.2), *Wolbachia pipientis* (NC_002978.6), *Pseudomonas entomophila* (NC_008027.1), *Commensalibacter intestine* (NZ_AGFR00000000.1), *Acetobacter pomorum* (NZ_AEUP00000000.1), *Gluconobacter morbifer* (NZ_AGQV00000000.1), *Providencia burhodogranariea* (NZ_AKKL00000000.1), *Providencia alcalifaciens* (NZ_AKKM01000049.1), *Providencia rettgeri* (NZ_AJSB00000000.1), *Enterococcus faecalis* (NC_004668.1), *Lactobacillus brevis* (NC_008497.1), and *Lactobacillus plantarum* (NC_004567.2) to avoid paralogous mapping of reads belonging to different species. We further filtered for mapped reads with mapping qualities ≥ 20, removed duplicate reads with Picard (v2.17.6; http://picard.sourceforge.net) and re-aligned sequences flanking insertions-deletions (indels) with GATK (v3.4-46 McKenna et al, 2010)

**Quality control**

Since all libraries were constructed from haploid embryos, we assumed that polymorphisms within a library represent either (1) sequencing- or (2) mapping-errors. Accordingly, we expected to find erroneous alleles only at very low frequencies in each dataset. Alternatively, any problem during the construction of haploid embryos would lead to diploid sequences that result in residual heterozygosity characterized by an excess of polymorphisms with frequencies close to 0.5 in the affected library. To test for these hypotheses, we investigated the distribution of minor - putatively erroneous -

allele frequencies for each library separately. In addition, we divided the number of erroneous alleles by the total coverage at variant and invariant positions to calculate library-specific error-rates.

**Variant calling**

We identified single nucleotide polymorphisms (SNPs) based on a combination of stringent heuristic criteria to exclude sequencing and mapping errors in each of the Swedish and Zambian datasets using custom software: For each library, we excluded polymorphic positions with minor frequencies > 0.1. In all other cases, we considered the major allele as the correct allelic state for a given individual. To avoid erroneous SNPs due to inflated sampling error at low-coverage sites or due to paralogous alleles at sites with excessive coverage from mapping errors, we only considered positions with more than 10-fold and less than 200-fold coverage. We further ignored positions where less than 14 of the 28 samples (14 Swedish and 14 Zambian) fulfilled the above-mentioned quality criteria. At last, we refined the SNP dataset by excluding SNPs located either within known transposable elements (TE) based on the *D. melanogaster* reference genome (v.6.12) or within a 5 base-pairs distance to indel polymorphisms supported by 10 reads across all samples. Finally, the same set of filters was applied to each other non-polymorphic chromosomal position in the data. This allowed us to obtain the total number of monomorphic sites in our dataset, which is needed for demographic inference (Laurent et al., 2016).

**Bioinformatic karyotyping**

Following the approach in Kapun et al. (2014), we used a panel of karyotype-specific marker SNPs that are diagnostic for seven chromosomal inversion (*In(2L)t, In(2R)NS, In(3L)P, In(3R)C, In(3R)K, In(3R)Mo* and *In(3R)Payne*) to karyotype all Swedish samples based on presence or absence of alleles which are in tight linkage with the corresponding inversion. We further used the same method to confirm the inversion-status in the previously karyotyped samples from Zambia. We only considered a sample to be positive for an inversion if it carried ≥ 95% of all alleles that are specific to the corresponding inversion.

**Principal Component Analysis**

Principal component analyses were conducted with the "auto_SVD" function from the R package *bigsnpr* (Prive et al., 2017). This algorithm uses clumping instead of pruning to thin SNPs based on linkage disequilibrium, removes SNPs in long-range LD regions, and uses the thinned data to perform dimensionality reduction by singular value decomposition (SVD). Analyses were conducted on the full data and on each chromosomal arm separately.

**Demographic analyses**

For the demographic inference, we used SNPs from all neutral introns (smaller than 65bp, bases from the 8$^{th}$ to the 30$^{th}$, described as the most appropriate sites to be used for such analyses in Parsch et al. (2010) together with 4-fold degenerate sites present in chromosomes 2R, 3L, 3R, and X. The latter SNP lass was obtained following Grenier et al. (2015). Autosomal and X-linked data were treated separately. All genomic regions spanned by common inversions were excluded from the analyses (as defined by coordinates of inversion breakpoints obtained from Corbett-Detig and Hartl (2012)). Additionally, long runs of Identity-By-Descent were masked from the African lines using a perl program available from the DPGP website (http://www.johnpool.net/genomes.html). Genomic regions that were identified as of European ancestry in the DPGP 2 and DPGP 3 project were not masked, because our demographic analyses were intended to evaluate the possibility of gene flow between the two populations. All coordinates were transformed to the dm6 assembly using an in-house python script. In total, 390,852bp (42,306 SNPs) were used for the 3 autosomes together and 183,502bp (27,972 SNPs) for the chromosome X. We used the software *dadi* (Gutenkunst et al., 2009) to test four different demographic scenarios. In all models, the ancestral African population experienced a stepwise expansion at time $T_{exp}$. After the expansion, (forward in time) the European population splits from the African population at time $T_{split}$. Immediately after the split, the size of the new European population is instantaneously reduced to a population size $N_{bot}$, whereas the size of the African population does not change. After the bottleneck, the European population is allowed to recover exponentially until it reaches its current size $N_{eu}$. The four scenarios differ in the modeling of migration following the population split. Model 1 (NOMIG) does not implement gene flow and is therefore similar to previously published models (Duchen et

al., 2013; Laurent et al., 2011; Li & Stephan, 2006). Model 2 (SYMIG) implements symmetrical migration between the populations, starting immediately after the split and lasting until the end of the simulation (present). Model 3 (ASYMIG) is similar to model 2 but allows for asymmetrical migration rates. Finally, Model 4 (RASYMIG) is similar to model 3 except that asymmetrical migration only starts at time $T_{mig}$. These four models have six, seven, eight, and nine parameters, respectively. For every scenario, at least 10 independent runs with different initial parameters values were performed and the run achieving the highest likelihood was kept for parameter estimation and model choice. Model choice was done by comparing the Akaike information criteria (AIC) between models (Akaike, 1974). Confidence Intervals (CI) were calculated using the following procedure: First, 150 datasets were simulated using the best demographic model. These simulations were treated as pseudo-observed data and used to re-estimate demographic parameters under the best model. The set of 150 estimates for each demographic parameter was then used to construct the confidence intervals. Because the re-estimated parameters are not normally distributed, confidence intervals were calculated as the 2.5-97.5% percentiles (see Table 2). Nucleotide diversity, Tajima's D, FST, and the observed 1D and 2D site frequency spectra presented in Figure S3 were calculated with the built-in functions implemented in *dadi*.

Past changes in coalescent rate, and consequently ancestral population size changes, were inferred using the program MSMC (Schiffels & Durbin, 2014). The analysis was performed on 20 pairs of strains drawn at random from the Swedish and Zambian populations respectively, and on 40 pairs consisting of a single strain drawn from each population at random. All available SNPs from chromosomes 2R, 3R, and 3L were used. The software was invoked with the following options: msmc -i 30 -t 8 -p "20*1+30*2". The scaled times and the coalescent rates output by MSMC were converted to generations and $N_e$, respectively using a per base-pair mutation rate of 1.3e-9.

## RESULTS

### Summary statistics of mapping

Our sequencing effort of the Swedish lines yielded homogenous average coverage across all autosomal arms ranging from 53.3x on *2R* to 57.2x on *2L.* In contrast, we

observed a slightly higher coverage on the *X* chromosome (63.7x). These patterns were consistent with the data from the Zambian lines, where we also found slight coverage excess on the *X*. We, however, identified pronounced variation in library-specific read-depth, ranging from 19.4x in SU93n to 87.7x in SU02n for the 14 Swedish and to a lesser extent also in the Zambian lines, which ranged from 26.9x in ZI200 to 38.7x in ZI472 (Figure S1). We found no evidence for residual heterozygosity, which confirms that all sequenced libraries were based on haploid genomes only and are thus fully phased (see Figure S2). Furthermore, we observed that errors occurred at very low frequencies corresponding to an average error rate of 0.365% in the Swedish and 0.348% in the Zambian libraries.

**Patterns of genetic variation in the Swedish sample**

Previous studies based on smaller number of loci showed that European flies derived from an ancestral sub-Saharan population from which they diverged at the end of the last glacial maximum (Stephan & Li, 2007) and that the colonization was associated with a founder event during which European flies were subject to high genetic drift (Li & Stephan, 2006; Thornton & Andolfatto, 2006). This scenario predicts observable genetic differences between Swedish and Zambian flies as well as a lower amount of diversity in the former due to the population size bottleneck associated with the founding event. We used PCA analysis to explore whether these expectations were also observed in our new genome-wide diversity data (Figure 1). This analysis showed that the first principal component always clustered separately European and African samples and that the Swedish lines consistently displayed a smaller dispersion along the second principal component, reflecting lower diversity compared to the Zambian sample (Table 1, McVean, 2009). One important exception to this general pattern was observed on chromosome *2L*. In addition to the population specific clustering on PC1, we identified an equally strong clustering on PC2 that was perfectly consistent with the presence or absence of the known chromosomal inversion *In(2L)t*, whose occurrence in Sweden is here reported for the first time (Table S1). The effect of *In(2L)t* on population genetic structure has already been described in the DPGP3 dataset and, interestingly, has also been shown to extend beyond the chromosomal breakpoints of the inversion, which could reflect the effect of historical positive selection on the inverted arrangement (Corbett-Detig & Hartl, 2012). We show here for the first time that the

genetic differentiation between Swedish and African lines carrying the inverted arrangement of *In(2L)t* is smaller than for lines carrying the standard (non-inverted) arrangement.

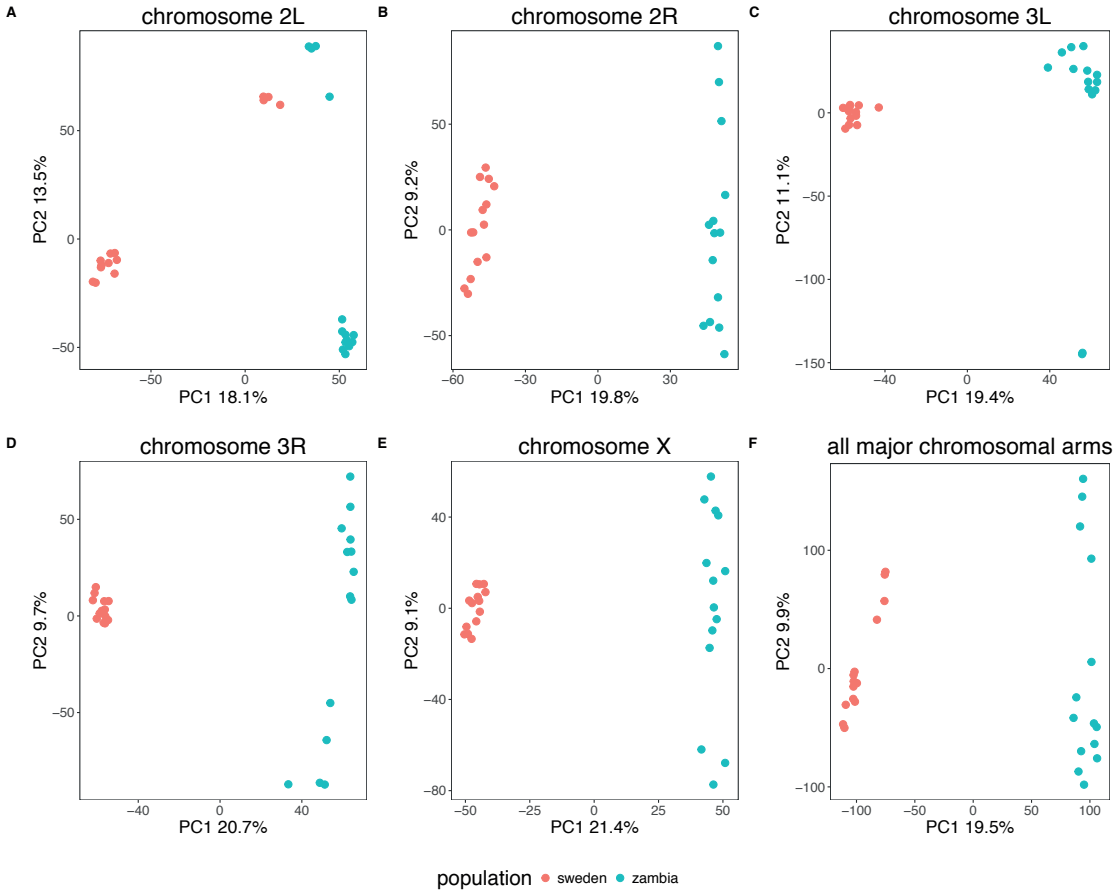



**Figure 1: SVD results.** Results of the principal SVD analyses are presented for each major chromosomal arm separately and for all chromosomes together. Only the first two components are shown. Individuals tend to cluster according to their sampling location except for chromosome 2L, for which flies carrying the inverted variant of the inversion *In(2L)t* cluster together regardless of their geographical origin.

## Demographic modeling

To test whether migration represented an important evolutionary force after the split between the European and African populations, we designed four demographic models recapitulating the main assumptions about the possible role of migration in this system (Figure 2, see Materials and Methods for a description of the models). Model choice and parameter estimation were conducted using the software *dadi* 1.7.0 with a neutral subset of the data (see Materials and Methods). Population genetic statistics of

the observed data (Table 1, Figure S3) were in line with values reported by previous studies based on smaller numbers of loci (Ometto et al., 2005). Our demographic analyses showed that models including migration provided a better fit to the neutral data compared to the model without migration for both the autosomal and the X-linked dataset (Figure 2). For the autosomal data, the best fit was provided by model "ASYMIG" (ongoing asymmetrical migration, Figure 2). Under this model, divergence between the Zambian and Swedish samples for the autosomal data occurred 43,540 years ago (assuming 10 generations per year) and was followed by ongoing asymmetrical migration with the migration rate from Sweden to Zambia ($M_{SZ}=2Nm_{SZ}=2.24$) being larger than from Zambia to Sweden ($M_{ZS}=0.53$). As expected, including gene flow into the models yielded older estimates for the age of the population split (Table 2, Table S2). We accordingly report here older divergence time than previous studies who did not take migration into account (Duchen et al., 2013; Laurent et al., 2011; Li & Stephan, 2006). For the X-chromosomal data, the best model was "SYMIG" (ongoing symmetrical migration). Under this model, divergence time was estimated to be 25,999 years with an ongoing symmetrical migration rate of 1.23 (number of genomes migrating per generation). X-chromosome modeling also confirmed the stronger estimated bottleneck for the X versus autosomes (Hutter et al., 2007; Laurent et al., 2011). The comparison between observed data and predictions of the best models showed that our modeling approach yielded a good absolute fit to the data (Figure S4).

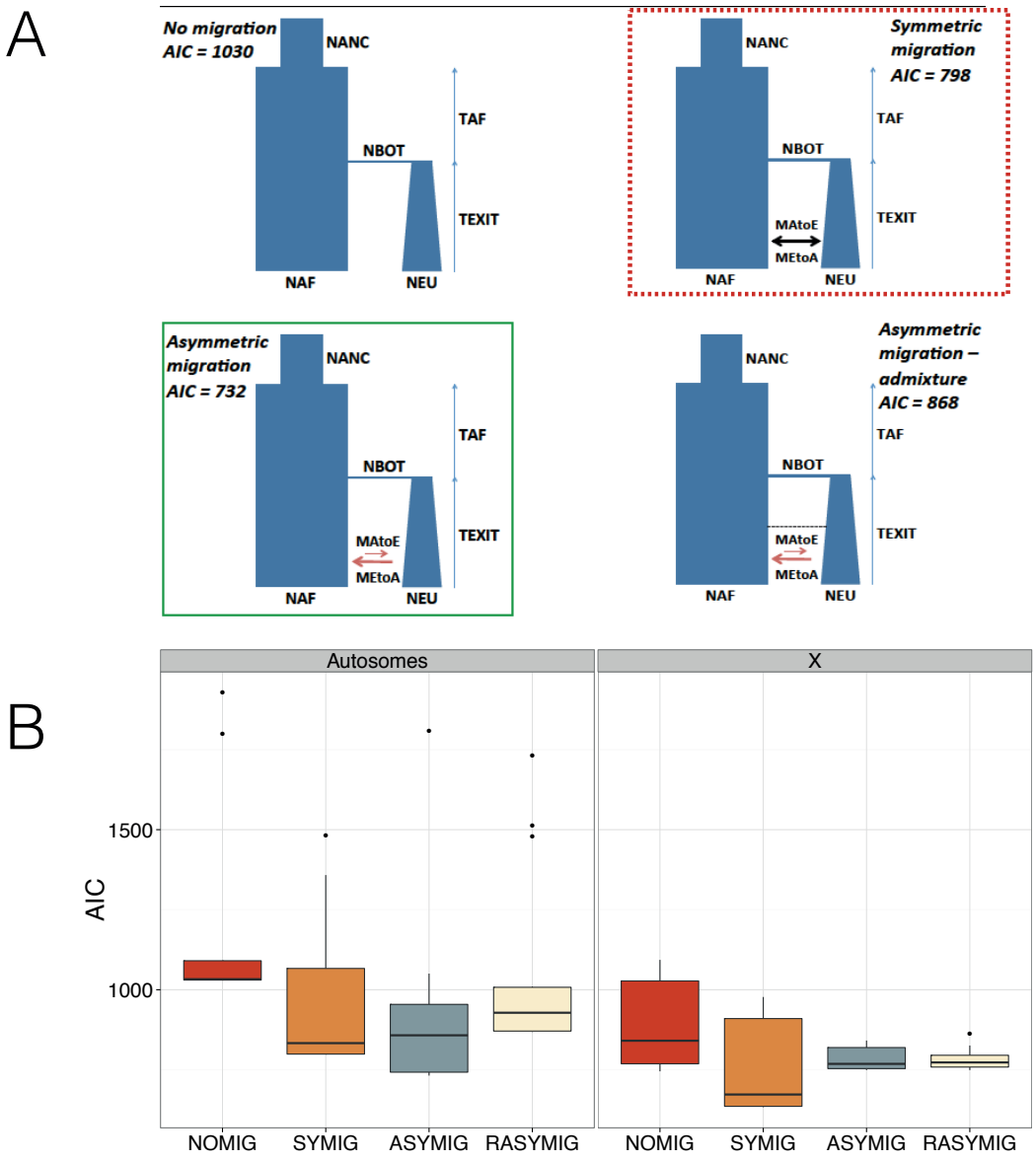



**Figure 2: Results of the model choice analyses.** A) The four demographic models tested in this study. Lowest AIC out of 10 replicates are reported for each model. The green box with a continuous line indicates the best model for the autosomal data. The red box with the dotted line indicates the best model for the X-linked data. B) Distribution of AIC for each for the Autosomal and X-linked datasets across 10 replicates. Lower values of the AIC statistic indicate a better fit between the observed data and the demographic models.

| | Umeå (Sweden) | | Siavonga (Zambia) | |
|---|---|---|---|---|
| | **Autosomes** | **X** | **Autosomes** | **X** |
| **$\theta_w$ (per bp)** | 0.007 | 0.005 | 0.013 | 0.014 |
| **$\theta_\pi$ (per bp)** | 0.008 | 0.005 | 0.012 | 0.013 |
| **Tajima's D** | 0.16 | 0.32 | -0.36 | -0.475 |
| **$F_{ST}$ (Umea - Siavonga)** | 0.2 | 0.26 | | |

**Table 1:** Summary statistics of genetic diversity measured on our neutral dataset (i.e. introns smaller than 65bp, and 4-fold degenerated third codon positions). All known inversions have been removed as well as chromosome 2L. All statistics have been calculated with *dadi* on the same site frequency spectra used for demographic inference.

| Parameters | Autosomes | | X | |
|---|---|---|---|---|
| | This study ML | Laurent et al. (2011) ABC | This study ML | Laurent et al. (2011) ABC |
| Current African population size ($N_{AF}$) | 4,639,014 (4,000,067; 5,217,550) | 3,134,891 (1,371,066; 28,013,950) | 7,537,910 (6,425,986; 9,845,444) | 4,786,360 (2,040,701; 29,208,295) |
| Current European population size ($N_{EU}$) | 957,941 (591,256; 1,917,823) | 878,506 (383,361; 4,775,964) | 529,902 (355,231; 1,146,428) | 1,632,505 (780,907; 4,870,580) |
| European Bottleneck population size ($N_{BOT}$) | 112,191 (71,831; 242,384) | 32,128 (15,968; 95,162) | 41,507 (22,529; 81,830) | 22,066 (14,338; 81,102) |
| African-European divergence time ($T_{SPLIT}$) | 43,540 (33,154; 74,498) | 12,843 (7,095; 31,773) | 25,999 (18,810; 37,809) | 16,849 (9,392; 33,452) |
| African expansion time ($T_{EXP}$) | 61,334 (24,011; 119,173) | 37,323 (3,636; 379,212) | 79,776 (44,008; 117,303) | 25,553 (1,698; 376,730) |
| Ancestral African population size ($N_{ANC}$) | 2,058,317 (1,949,391; 2,145,813) | 1,705,328 (609,393; 2,458,653) | 2,147,406 (2,040,027; 2,253,918) | 1,837,229 (931,637; 2,530,609) |
| Migration rate Europe to Africa ($M_{SZ}$) | 2.24 (1.26; 2.67) | not estimated | 1.23 (0.83; 1.58) | not estimated |
| Migration rate Africa to Europe (MZS) | 0.53 (0.09; 1.31) | not estimated | 1.23 (0.83; 1.58) | not estimated |

**Table 2: Demographic estimates from this study compared to the demographic estimates obtained by Laurent et al. (2011) for the same populations.** For the dadi estimates we report the maximum likelihood estimates and the confidence interval obtained with parametric bootstrapping. The estimates from Laurent et al. (2011) correspond to the mode and the 2.5$^{th}$ and 97.5$^{th}$ quantiles of the posterior distribution.

It has been shown that large contiguous sequence information from a sample of size two contains information about historical changes in coalescent rates (Li & Durbin, 2011). In theory, this approach should complement classical model-based inference procedures like the one presented in Figure 2, because no assumptions are required about how often the coalescent rate can change during the history of the sample. Figure 3 summarizes the results of our estimations of historical coalescent rates in the European and African populations for the autosomal data (excluding chromosome 2L). Estimates younger than 10 thousand years (kyr) display a large variance across replicates consistent with the fact that samples of size two are not expected to contain

much statistical information about recent coalescent rates (Schiffels & Durbin, 2014). Similarly, the dramatic decreases in coalescent rates observed in the oldest time intervals of both the European and African samples are unlikely to reflect neutral demographic processes (see Discussion) and we therefore restrict our interpretation of these results to the time interval spanning from 10kyr to 300kyr. As expected, the coalescence rate in the African sample is lower than the one in the European sample. The African rate also displays a continuous reduction between 100 kyr and 200 kyr that likely corresponds to the ancestral population size expansion inferred by *dadi* in this study, as well as by previous studies (Laurent et al., 2011; Li & Stephan, 2006). To our knowledge, the steady increase of the African coalescent rate in the last 25kyr has not been documented before and it is unclear whether this observation is caused by poorly resolved portions of the data, lack of statistical signal, or true evolutionary processes. Interestingly, the results for the European sample indicate a steady increase of the coalescent rate, which starts approximately at the same time as the African expansion (250kyr). This result indicates that the ancestors of non-African flies could have started diverging from the ancestral population earlier than suggested by our results obtained with *dadi* (44kyr, Table 2). However, the coalescence rates measured between the two populations (grey lines in Figure 3) displays a minimum around 44kyr, which is more consistent with the divergence time estimated by *dadi*.

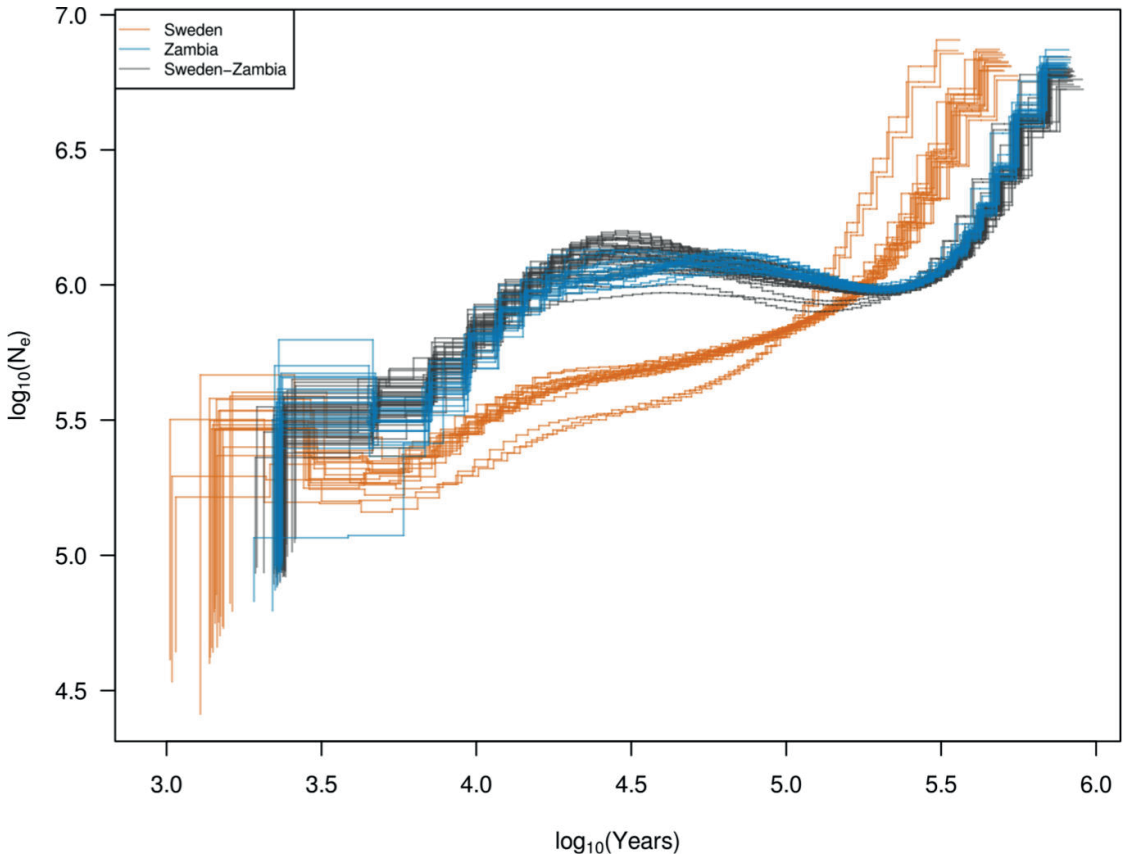



**Figure 3: MSMC results – Historical changes in coalescence rates.** Coalescence rates are inferred with the program MSMC. Each line shows past changes in rates for a single pair of strains drawn at random from the populations. Pairs from the Swedish and Zambian populations are shown in orange and blue, respectively, and pairs consisting of a single strain from each population are shown in grey. The time scale on the x-axis was derived considering 10 generations per year. We decided to rescale estimated coalescence rates into $N_e$ values to facilitate comparisons with similar estimates obtained with the software *dadi*. Coalescence rates equal the inverse of two times $N_e$.

## DISCUSSION

The SVD analysis presented in Figure 1a illustrates the important effect that chromosomal inversions can have on neutral polymorphism data. Importantly, the structure created by *In(2L)t* in the data extends several megabases beyond the inversion's breakpoint (Corbett-Detig & Hartl, 2012; Huang et al., 2014). Therefore, we excluded the totality of chromosome 2L for the demographic analyses in this study and we recommend that future demographic studies of natural populations of *Drosophila*

*melanogaster* address the potential effect of this inversion prior to model fitting. Alternatively, coalescent models that explicitly account for the effect of chromosomal inversions (Guerrero et al., 2012; Peischl et al., 2013) could be used to jointly take into account demographic processes and the specific patterns of recombination caused by the inversion. We note that such models could in principle be used to investigate why the genetic differentiation between African and European populations is lower for *In(2L)t* compared to the standard arrangement (Figure 1a). We speculate that lines carrying *In(2L)t* may have colonized Europe more recently than lines carrying the standard arrangement, leaving less time for drift to increase differentiation.

Our model-based demographic analyses (Figure 2, Table 2, Figure S4) confirmed that European populations do not exhibit patterns of African admixture comparable to the ones that have been measured in American and Australian populations (Bergland et al., 2016; Caracristi & Schlotterer, 2003; Duchen et al., 2013). This indicates that European natural populations of *D. melanogaster*, and ancient populations in general may be better suited for studying local adaptation at the genetic level because neutral models serving as a null hypothesis for selection detection methods do not have to account for the additional complexity caused by genetic admixture (but see Lohmueller et al., 2011). The new Swedish panel presented in this study therefore represents an appropriate sample to address the long lasting issue of the respective contributions of hard versus soft selective sweeps in the adaptation of *D. melanogaster* to northern latitudes (Garud et al., 2015; Jensen, 2014).

Nevertheless, accounting for ongoing gene flow between Africa and Europe improved the fit to the data compared to models that do not allow for migration (Figure 2b). As predicted by Li and Stephan (2006), allowing for gene exchange between Africa and Europe in the demographic model provides an older estimate for the age of the split between the two populations (Table 2, Table S2). The age of the divergence obtained from neutral autosomal under our best model (43,540 years) suggests that the split between African and European ancestral lineages occurred during the last glacial period. Another interesting consequence of including migration is that estimates for $N_{BOT}$ (the size of the European population directly after the split) are roughly two times larger than in models without migration (Table S2). The potential effect of a less severe bottleneck and gene flow on the performance of selection detection in *D. melanogaster* remains to be investigated and is beyond the scope of this study.

By providing a more detailed description of how instantaneous coalescence rates change through time, our MSMC analysis provides new insights into the demographic history of European flies. Our estimate for the time of split between Africa and Europe (43,540 years) is in agreement with the part of the graph where the coalescence rate between populations (grey lines, Figure 3) becomes smaller than the coalescence rate within population. The increase in coalescence rates in the recent history of the African sample is not expected under our best model and could be explained by the action of selection on linked neutral sites, but more work is needed to understand how MSMC results are affected by violations of the assumption of neutrality. Finally, the steep decrease in coalescence rates in both samples for the oldest time intervals likely reflects the presence of short clusters of false-positive heterozygous sites arising in low-complexity regions of the genome.

## ACKNOWLEDGEMENTS

We thank Roman Arguello for helpful discussions and for providing the coordinates of short introns and fourfold degenerate coding sites for the neutral set of loci. Sequencing of the Swedish lines was supported by grant STE325/12-2 of the DFG Research Unit 1078 to W.S.

## REFERENCES

Akaike, H. (1974). New Look at Statistical-Model Identification. *Ieee Transactions on Automatic Control, Ac19*(6), 716-723. doi:Doi 10.1109/Tac.1974.1100705

Bergland, A. O., Tobler, R., Gonzalez, J., Schmidt, P., & Petrov, D. (2016). Secondary contact and local adaptation contribute to genome-wide patterns of clinal variation in Drosophila melanogaster. *Mol Ecol, 25*(5), 1157-1174. doi:10.1111/mec.13455

Bushnell, B. (2017). BBMap URL http://sourceforge. net/projects/bbmap. *BBTools software package*.

Caracristi, G., & Schlotterer, C. (2003). Genetic differentiation between American and European Drosophila melanogaster populations could be attributed to admixture of African alleles. *Mol Biol Evol, 20*(5), 792-799. doi:10.1093/molbev/msg091

Casillas, S., & Barbadilla, A. (2017). Molecular Population Genetics. *Genetics, 205*(3), 1003-1035. doi:10.1534/genetics.116.196493

Charlesworth, B. (2012). The role of background selection in shaping patterns of molecular evolution and variation: evidence from variability on the Drosophila X chromosome. *Genetics, 191*(1), 233-246.

Corbett-Detig, R. B., & Hartl, D. L. (2012). Population genomics of inversion polymorphisms in Drosophila melanogaster. *PLoS Genet, 8*(12), e1003056. doi:10.1371/journal.pgen.1003056

David, J. R., & Capy, P. (1988). Genetic variation of Drosophila melanogaster natural populations. *Trends Genet, 4*(4), 106-111.

Duchen, P., Zivkovic, D., Hutter, S., Stephan, W., & Laurent, S. (2013). Demographic inference reveals African and European admixture in the North American Drosophila melanogaster population. *Genetics, 193*(1), 291-301. doi:10.1534/genetics.112.145912

Elyashiv, E., Sattath, S., Hu, T. T., Strutsovsky, A., McVicker, G., Andolfatto, P., . . . Sella, G. (2016). A Genomic Map of the Effects of Linked Selection in Drosophila. *PLoS Genet, 12*(8), e1006130. doi:10.1371/journal.pgen.1006130

Garud, N. R., Messer, P. W., Buzbas, E. O., & Petrov, D. A. (2015). Recent selective sweeps in North American Drosophila melanogaster show signatures of soft sweeps. *PLoS Genet, 11*(2), e1005004. doi:10.1371/journal.pgen.1005004

Grenier, J. K., Arguello, J. R., Moreira, M. C., Gottipati, S., Mohammed, J., Hackett, S. R., . . . Clark, A. G. (2015). Global diversity lines - a five-continent reference panel of sequenced Drosophila melanogaster strains. *G3 (Bethesda), 5*(4), 593-603. doi:10.1534/g3.114.015883

Guerrero, R. F., Rousset, F., & Kirkpatrick, M. (2012). Coalescent patterns for chromosomal inversions in divergent populations. *Philos Trans R Soc Lond B Biol Sci, 367*(1587), 430-438. doi:10.1098/rstb.2011.0246

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet, 5*(10), e1000695. doi:10.1371/journal.pgen.1000695

Haddrill, P. R., Thornton, K. R., Charlesworth, B., & Andolfatto, P. (2005). Multilocus patterns of nucleotide variability and the demographic and selection history of Drosophila melanogaster populations. *Genome Res, 15*(6), 790-799. doi:10.1101/gr.3541005

Huang, W., Massouras, A., Inoue, Y., Peiffer, J., Ramia, M., Tarone, A. M., . . . Mackay, T. F. (2014). Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. *Genome Res, 24*(7), 1193-1208. doi:10.1101/gr.171546.113

Hutter, S., Li, H., Beisswanger, S., De Lorenzo, D., & Stephan, W. (2007). Distinctly different sex ratios in African and European populations of Drosophila melanogaster inferred from chromosomewide single nucleotide polymorphism data. *Genetics, 177*(1), 469-480. doi:10.1534/genetics.107.074922

Jensen, J. D. (2014). On the unfounded enthusiasm for soft selective sweeps. *Nat Commun, 5*, 5281. doi:10.1038/ncomms6281

Kao, J. Y., Zubair, A., Salomon, M. P., Nuzhdin, S. V., & Campo, D. (2015). Population genomic analysis uncovers African and European admixture in Drosophila melanogaster populations from the south-eastern United States and Caribbean Islands. *Mol Ecol, 24*(7), 1499-1509. doi:10.1111/mec.13137

Kapun, M., van Schalkwyk, H., McAllister, B., Flatt, T., & Schlotterer, C. (2014). Inference of chromosomal inversion dynamics from Pool-Seq data in natural and laboratory populations of Drosophila melanogaster. *Mol Ecol, 23*(7), 1813-1827. doi:10.1111/mec.12594

Keller, A. (2007). Drosophila melanogaster's history as a human commensal. *Curr Biol, 17*(3), R77-81. doi:10.1016/j.cub.2006.12.031

Lack, J. B., Cardeno, C. M., Crepeau, M. W., Taylor, W., Corbett-Detig, R. B., Stevens, K. A., . . . Pool, J. E. (2015). The Drosophila genome nexus: a population genomic resource of 623

Drosophila melanogaster genomes, including 197 from a single ancestral range population. *Genetics, 199*(4), 1229-1241. doi:10.1534/genetics.115.174664

Lack, J. B., Lange, J. D., Tang, A. D., Corbett-Detig, R. B., & Pool, J. E. (2016). A Thousand Fly Genomes: An Expanded Drosophila Genome Nexus. *Mol Biol Evol, 33*(12), 3308-3313. doi:10.1093/molbev/msw195

Langley, C. H., Crepeau, M., Cardeno, C., Corbett-Detig, R., & Stevens, K. (2011). Circumventing heterozygosity: sequencing the amplified genome of a single haploid Drosophila melanogaster embryo. *Genetics, 188*(2), 239-246. doi:10.1534/genetics.111.127530

Langley, C. H., Stevens, K., Cardeno, C., Lee, Y. C., Schrider, D. R., Pool, J. E., . . . Begun, D. J. (2012). Genomic variation in natural populations of Drosophila melanogaster. *Genetics, 192*(2), 533-598. doi:10.1534/genetics.112.142018

Laurent, S., Pfeifer, S. P., Settles, M. L., Hunter, S. S., Hardwick, K. M., Ormond, L., . . . Rosenblum, E. B. (2016). The population genomics of rapid adaptation: disentangling signatures of selection and demography in white sands lizards. *Mol Ecol, 25*(1), 306-323. doi:10.1111/mec.13385

Laurent, S. J., Werzner, A., Excoffier, L., & Stephan, W. (2011). Approximate Bayesian analysis of Drosophila melanogaster polymorphism data reveals a recent colonization of Southeast Asia. *Mol Biol Evol, 28*(7), 2041-2051. doi:10.1093/molbev/msr031

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature, 475*(7357), 493-496. doi:10.1038/nature10231

Li, H., & Stephan, W. (2006). Inferring the demographic history and rate of adaptive substitution in Drosophila. *PLoS Genet, 2*(10), e166. doi:10.1371/journal.pgen.0020166

Lohmueller, K. E., Bustamante, C. D., & Clark, A. G. (2011). Detecting directional selection in the presence of recent admixture in African-Americans. *Genetics, 187*(3), 823-835. doi:10.1534/genetics.110.122739

Mackay, T. F., Richards, S., Stone, E. A., Barbadilla, A., Ayroles, J. F., Zhu, D., . . . Gibbs, R. A. (2012). The Drosophila melanogaster Genetic Reference Panel. *Nature, 482*(7384), 173-178. doi:10.1038/nature10811

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal, 17*(1), 10-12.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., . . . DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res, 20*(9), 1297-1303. doi:10.1101/gr.107524.110

McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS Genet, 5*(10), e1000686. doi:10.1371/journal.pgen.1000686

Ometto, L., Glinka, S., De Lorenzo, D., & Stephan, W. (2005). Inferring the effects of demography and selection on Drosophila melanogaster populations from a chromosome-wide scan of DNA variation. *Mol Biol Evol, 22*(10), 2119-2130. doi:10.1093/molbev/msi207

Parsch, J., Novozhilov, S., Saminadin-Peter, S. S., Wong, K. M., & Andolfatto, P. (2010). On the utility of short intron sequences as a reference for the detection of positive and negative selection in Drosophila. *Mol Biol Evol, 27*(6), 1226-1234. doi:10.1093/molbev/msq046

Peischl, S., Koch, E., Guerrero, R. F., & Kirkpatrick, M. (2013). A sequential coalescent algorithm for chromosomal inversions. *Heredity (Edinb), 111*(3), 200-209. doi:10.1038/hdy.2013.38

Pool, J. E., Corbett-Detig, R. B., Sugino, R. P., Stevens, K. A., Cardeno, C. M., Crepeau, M. W., . . . Langley, C. H. (2012). Population Genomics of sub-saharan Drosophila melanogaster: African diversity and non-African admixture. *PLoS Genet, 8*(12), e1003080. doi:10.1371/journal.pgen.1003080

Prive, F., Aschard, H., Ziyatdinov, A., & Blum, M. G. B. (2017). Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics*. doi:10.1093/bioinformatics/bty185

Schiffels, S., & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nat Genet, 46*(8), 919-925. doi:10.1038/ng.3015

Sella, G., Petrov, D. A., Przeworski, M., & Andolfatto, P. (2009). Pervasive natural selection in the Drosophila genome? *PLoS Genet, 5*(6), e1000495. doi:10.1371/journal.pgen.1000495

Stephan, W., & Li, H. (2007). The recent demographic and adaptive history of Drosophila melanogaster. *Heredity (Edinb), 98*(2), 65-68. doi:10.1038/sj.hdy.6800901

Thornton, K., & Andolfatto, P. (2006). Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of Drosophila melanogaster. *Genetics, 172*(3), 1607-1619. doi:10.1534/genetics.105.048223

Veuille, M., Baudry, E., Cobb, M., Derome, N., & Gravot, E. (2004). Historicity and the population genetics of Drosophila melanogaster and D. simulans. *Genetica, 120*(1-3), 61-70.

## Data availability

Short reads have been made available in Genbank (see Supplementary Table 1).

## Supplementary Material

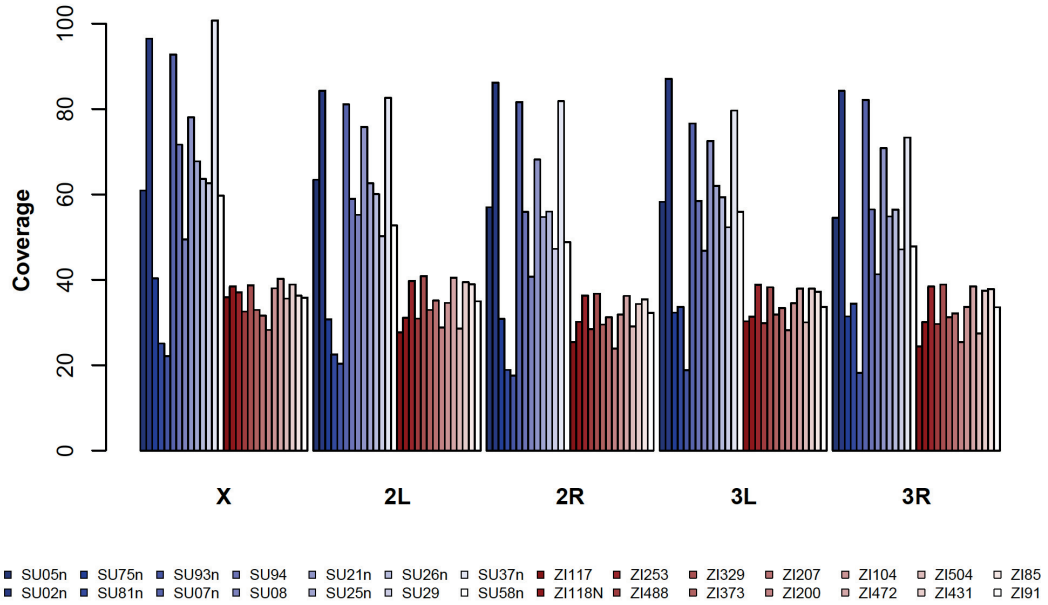



**Figure S1. Sequencing depth per line.** Barplots showing the sequencing depths for all 28 samples and 5 chromosomal arms. Line names with the prefix "SU" (blue) and "ZI" (red) indicate the Swedish and Zambian samples, respectively.

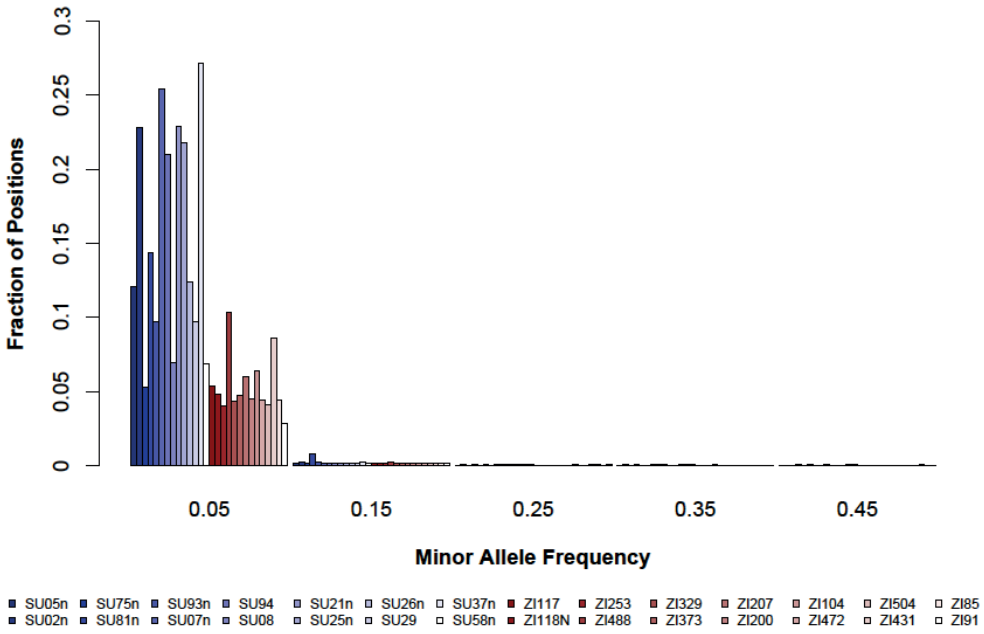



**Figure S2. Sample-specific error rates.** Barplots showing the sample-specific frequencies of false positive alleles due to sequencing or mapping errors. The y-axis shows the proportion of total positions that contain false position alleles of the corresponding frequency class.

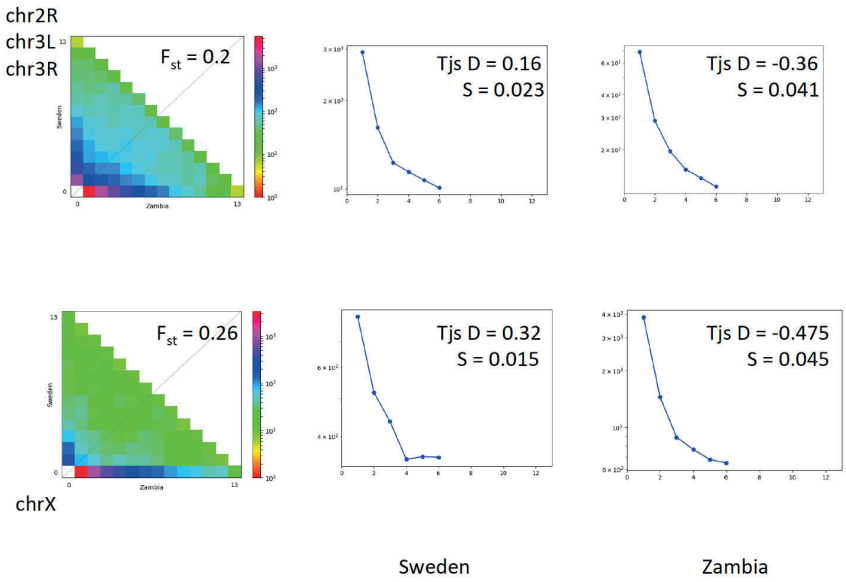
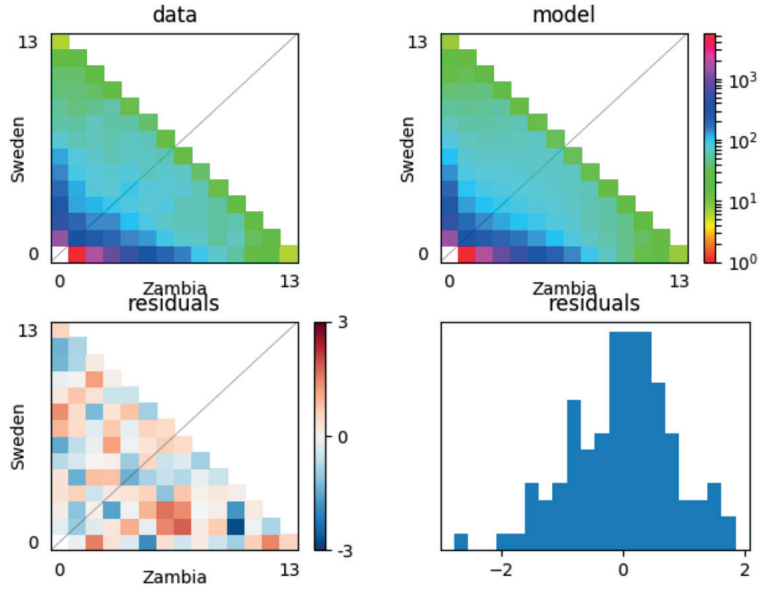



**Figure S3. 2D Allele Frequency Spectra (JAFS)** for Sweden and Zambia (on the left) and for the 3 autosomal arms (on the top) and chromosome X (at the bottom). On the right, the individual folded AFS for Zambia and for Sweden in the middle.




Tajima's D Sweden **data** = 0.16
Tajima's D Zambia **data** = -0.36
$F_{st}$ **data** = 0.2

Tajima's D Sweden **model** = 0.1
Tajima's D Zambia **model** = -0.26
$F_{st}$ **model** = 0.2

**Figure S4. Residuals for the best model.** The Joint Allele Frequency Spectra for the observed dataset (top left) and calculated for the best model (top right). Below, the residuals obtained for the best model.

| Library | Town | Country | Collection Date | SRA Accession | X | 2L | 2R | 3L | 3R | AvCov | Error rate | *In(2L)t* | *In(2R)NS* | *In(3L)P* | *In(3R)K* | *In(3R)Mo* | *In(3R)P* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SU02n | Umeå | Sweden | 07/2010 | SRR2347216 | 96.5 | 84.3 | 86.2 | 87.1 | 84.3 | 87.7 | 0.004176886 | ST | ST | ST | ST | ST | ST |
| SU05n | Umeå | Sweden | 07/2010 | SRR2347265 | 60.9 | 63.5 | 57.0 | 58.2 | 54.6 | 58.8 | 0.003297887 | ST | ST | ST | ST | ST | ST |
| SU07n | Umeå | Sweden | 07/2010 | SRR2347336 | 92.8 | 81.1 | 81.6 | 76.6 | 82.2 | 82.8 | 0.004518216 | ST | ST | ST | ST | ST | ST |
| SU08 | Umeå | Sweden | 07/2010 | SRR2347337 | 49.5 | 55.2 | 40.7 | 46.8 | 41.3 | 46.7 | 0.002996531 | INV | ST | ST | ST | ST | ST |
| SU21n | Umeå | Sweden | 07/2010 | SRR2347338 | 78.1 | 75.8 | 68.2 | 72.5 | 70.8 | 73.1 | 0.004580965 | INV | ST | ST | ST | ST | ST |
| SU25n | Umeå | Sweden | 07/2010 | SRR2347339 | 67.8 | 62.6 | 54.7 | 62.0 | 54.9 | 60.4 | 0.005104961 | ST | ST | ST | ST | ST | ST |
| SU26n | Umeå | Sweden | 07/2010 | SRR2347340 | 63.7 | 60.1 | 56.0 | 59.4 | 56.5 | 59.1 | 0.003000216 | ST | ST | ST | ST | ST | ST |
| SU29 | Umeå | Sweden | 07/2010 | SRR2347341 | 62.7 | 50.2 | 47.2 | 52.3 | 47.1 | 51.9 | 0.002701509 | ST | ST | ST | ST | ST | ST |
| SU37n | Umeå | Sweden | 07/2010 | SRR2347342 | 100.8 | 82.6 | 81.8 | 79.7 | 73.3 | 83.6 | 0.005359326 | ST | ST | ST | ST | ST | ST |
| SU58n | Umeå | Sweden | 07/2010 | SRR2347343 | 59.7 | 52.8 | 48.9 | 56.0 | 47.8 | 53.0 | 0.00252525 | ST | ST | ST | ST | ST | ST |
| SU75n | Umeå | Sweden | 08/2012 | SRR2347308 | 40.3 | 30.7 | 30.8 | 32.3 | 31.4 | 33.1 | 0.002448266 | ST | ST | ST | ST | ST | ST |
| SU81n | Umeå | Sweden | 08/2012 | SRR2347331 | 25.1 | 22.5 | 18.9 | 33.7 | 34.4 | 26.9 | 0.010984268 | INV | ST | ST | ST | ST | ST |
| SU93n | Umeå | Sweden | 08/2012 | SRR2347333 | 22.2 | 20.3 | 17.6 | 18.8 | 18.2 | 19.4 | 0.007010904 | ST | ST | ST | ST | ST | ST |
| SU94 | Umeå | Sweden | 07/2010 | SRR2347334 | 71.7 | 58.9 | 55.9 | 58.5 | 56.5 | 60.3 | 0.0050117 | INV | ST | ST | ST | ST | ST |
| ZI104 | Siavonga | Zambia | 08/2012 | SRR654551 | 38.0 | 34.6 | 31.9 | 34.5 | 33.7 | 34.5 | 0.002898396 | ST | ST | ST | ST | ST | ST |
| ZI117 | Siavonga | Zambia | 08/2012 | SRR248130 | 36.0 | 27.7 | 25.4 | 30.3 | 24.4 | 28.8 | 0.003159745 | INV | ST | ST | ST | ST | ST |
| ZI118N | Siavonga | Zambia | 08/2012 | SRR654664 | 38.5 | 31.1 | 30.2 | 31.4 | 30.1 | 32.2 | 0.002510175 | INV | ST | ST | ST | ST | ST |
| ZI200 | Siavonga | Zambia | 08/2012 | SRR203234 | 28.3 | 28.9 | 23.9 | 28.2 | 25.4 | 26.9 | 0.002822433 | ST | ST | ST | ST | ST | ST |
| ZI207 | Siavonga | Zambia | 08/2012 | SRR202075 | 31.7 | 35.1 | 31.2 | 33.4 | 32.2 | 32.7 | 0.002704208 | ST | ST | ST | ST | ST | ST |
| ZI253 | Siavonga | Zambia | 07/2010 | SRR203350 | 37.1 | 39.7 | 36.3 | 38.9 | 38.4 | 38.1 | 0.001771278 | INV | ST | ST | ST | ST | ST |
| ZI329 | Siavonga | Zambia | 07/2010 | SRR204006 | 38.7 | 40.9 | 36.7 | 38.3 | 38.9 | 38.7 | 0.001852544 | ST | ST | ST | ST | ST | ST |
| ZI373 | Siavonga | Zambia | 08/2012 | SRR210782 | 32.9 | 32.9 | 29.5 | 31.9 | 31.3 | 31.7 | 0.002430329 | ST | ST | ST | ST | ST | ST |
| ZI431 | Siavonga | Zambia | 08/2012 | SRR654556 | 38.9 | 39.4 | 34.3 | 37.9 | 37.5 | 37.6 | 0.003444254 | ST | ST | ST | ST | ST | ST |
| ZI472 | Siavonga | Zambia | 07/2010 | SRR203465 | 40.3 | 40.5 | 36.2 | 38.0 | 38.5 | 38.7 | 0.00187972 | ST | ST | ST | ST | ST | ST |
| ZI488 | Siavonga | Zambia | 08/2012 | SRR326792 | 32.6 | 30.9 | 28.5 | 29.8 | 29.7 | 30.3 | 0.004945238 | INV | ST | ST | ST | ST | ST |
| ZI504 | Siavonga | Zambia | 08/2012 | SRR248124 | 35.6 | 28.6 | 29.1 | 30.0 | 27.4 | 30.1 | 0.002328564 | ST | ST | ST | ST | ST | ST |
| ZI85 | Siavonga | Zambia | 08/2012 | SRR203508 | 36.3 | 39.0 | 35.4 | 37.2 | 37.8 | 37.2 | 0.001879926 | ST | ST | ST | ST | ST | ST |
| ZI91 | Siavonga | Zambia | 08/2012 | SRR189423 | 35.8 | 35.0 | 32.3 | 33.7 | 33.6 | 34.1 | 0.001560109 | ST | ST | ST | ST | ST | ST |

**Table S1. Sample origin, mapping coverage, error rates and karyotype status**

| | Texp (years) | Naf | Nbot | Neu | Tsplit (years) | MCL | theta | Nanc | AIC | Msz | Mzs | Tm (part of Tsplit) | 1/(4muL) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *ASYMIG autosomes* | 61334 | 4639014 | 112191 | 957941 | 43540 | -357.98 | 4183 | 2058317 | 732 | 2.2366 | 0.5337 | | 492 |
| *SYMIG autosomes* | 92690 | 4109293 | 136334 | 1222359 | 48144 | -392.55 | 4087 | 2011047 | 798 | 1.1755 | 1.1755 | | 492 |
| *RASYMIG autosomes* | 15273 | 3943794 | 104161 | 681555 | 384129 | -425.02 | 1560 | 767341 | 868 | 1.0988 | 0.3581 | 297426 | 492 |
| *NOMIG autosomes* | 79851 | 4553674 | 55204 | 6594654 | 24950 | -509.19 | 4211 | 2072103 | 1030 | 0 | 0 | | 492 |
| | | | | | | | | | | | | | |
| *SYMIG_chrX* | 79776 | 7537910 | 41507 | 529902 | 25999 | -309.67 | 2049 | 2147406 | 634 | 1.2266 | 1.2266 | | 1048 |
| *NOMIG_chrX* | 71064 | 8075528 | 17864 | 2579288 | 15519 | -366.62 | 2097 | 2197907 | 746 | 0 | 0 | | 1048 |
| *RASYMIG_chrX* | 92470 | 4100305 | 2041499 | 6654 | 502327 | -365.49 | 561 | 587656 | 748 | 0.6667 | 0.5157 | 155858 | 1048 |
| *ASYMIG_chrX* | 488769 | 4158847 | 455780 | 170468 | 121149 | -366.56 | 725 | 759904 | 750 | 0.9483 | 0.5924 | | 1048 |

**Table S2: Estimations for all tested models**

# CHAPTER 3

# The evolution of gene expression and binding specificity of the largest transcription factor family in primates

*Postprint version of the article published in Evolution*

Adamandia Kapopoulou[1, 2]; Lisha Mathew[1, 2]; Alex Wong [3]; Didier Trono[1] and Jeffrey D. Jensen[1, 2]

[1]School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

[2]Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland

[3]Department of Biology, Carleton University, Ottawa, Canada

Running Title: Population genetics of KRAB-ZF genes

## Abstract

The KRAB-containing zinc finger (KRAB-ZF) proteins represent the largest family of transcription factors in humans, yet for the great majority, their function and specific genomic target remain unknown. However, it has been shown that a large fraction of these genes arose from segmental duplications, and that they have expanded in gene and zinc finger number throughout vertebrate evolution. To determine whether this

expansion is linked to selective pressures acting on different domains, we have manually curated all KRAB-ZF genes present in the human genome together with their orthologous genes in three closely related species and assessed the evolutionary forces acting at the sequence level as well as on their expression profiles. We provide evidence that KRAB-ZFs can be separated in to two categories according to the polymorphism present in their DNA-contacting residues. Those carrying a nonsynonymous SNP in their DNA-contacting amino acids exhibit significantly reduced expression in all tissues, have emerged in a recent lineage, and seem to be less strongly constrained evolutionarily than those without such a polymorphism. This work provides evidence for a link between age of the transcription factor, as well as polymorphism in their DNA contacting residues and expression levels – both of which may be jointly affected by selection.

**Keywords**: KRAB-containing zinc-finger genes, regulatory evolution, DNA-contacting residues, transcription factors, endogenous retroelements, population genetics

## Introduction

Gene duplication can play a major role in species evolution: redundancy provides a medium for novelty while maintaining initial function. In the particular case of transcription factor (TF) genes, alterations in their expression profiles or binding properties can affect the expression of many target genes, often with a major functional impact. The KRAB-zinc finger family of transcription factors, the largest family of TFs in the human genome, arose through tandem segmental duplications and contains arrays of C2H2 (also called *Krüppel*-type) zinc fingers (ZFs) combined with a KRAB (*Krüppel*-associated box) domain. Despite being so numerous, the function and specific genomic targets of the great majority of KRAB-ZF proteins remain unknown (Constantinou-Deltas et al. 1992; Huntley et al. 2006; Thomas and Emerson 2009).

KRAB-ZF regulatory specificity is determined by a zinc finger-DNA recognition code, implicating interaction between specific amino acids within the zinc finger motifs and nucleotides at the binding sites (Choo and Klug 1994; Kim and Berg 1996). The

amino acids playing the most critical role in this DNA recognition are those at the -1, 2, 3, and 6 positions relative to the alpha-helical regions in each zinc finger domain (Pavletich and Pabo 1991; Elrod-Erickson et al. 1998). The strong conservation of some DNA-binding domains suggests that some genes have been stably integrated into essential regulatory relationships; however, in spite of this, little functional information from these genes is currently available (Liu et al. 2014).

In primates, KRAB-ZF genes duplicate at a higher rate than any other family. Paralogs diverge from the initial copy by a series of changes in the number and structure of zinc finger motifs, resulting in a dramatic diversity of binding specificities (Shannon et al. 2003; Hamilton et al. 2006). This DNA-binding diversity makes them ideal raw material for responding to newly emerging retrotransposons. Thomas and Schneider (2011) suggested that there is a continuous arms race between newly emerging retrotransposons and KRAB-ZFs acting as retrotransposon-specific repressors. Supporting this hypothesis, Jacobs et al. (2014) identified two KRAB-ZF genes involved in the repression of retrotransposons. They proposed a model where modifications to lineage-specific KRAB-ZFs result in repression of newly emerging families of retrotransposons, which in turn evolve to escape this repression. This evolutionary arms race may drive expansion and diversity of the KRAB-ZF genes and suggests a potential role for positive selection acting on affinity-modifying mutations in KRAB-ZFs. However, the extent to which positive selection has acted to shape this gene family is largely unknown.

One way to identify the relationships between sequence, function, and evolutionary process is to explore intra-species (polymorphic) variation of functional elements – specifically, the relationship between observed polymorphism and measured function (Spivakov et al. 2012). Interestingly, Lockwood et *al*. (2014) assessed polymorphism in the zinc finger DNA-contacting amino acids and reported that the majority of missense SNPs in these DNA-contacting residues did not have any effect on fitness. This example suggests that relaxed selective constraint may potentially explain the diversity of binding amino acids of KRAB-ZFs.

The purpose of this study is to examine the underlying mechanisms behind the large expansion of the KRAB-ZF family in primates. By assessing the expression levels of KRAB-ZF genes in various tissues and taking into account polymorphism in the DNA-contacting amino acids, we link the sequence of the KRAB-ZFs with their underlying

function. By manually curating all human KRAB-ZF genes and orthologous regions in three closely related species, and collecting polymorphism data from the 1000 genomes consortium, we were able to partition all human KRAB-ZF genes into two distinct categories according to the nature of SNPs occurring in the four DNA-contacting amino acids. Those two groups of genes differ significantly in their expression level for all tested tissues, the histone marks they bear in the gene body, and the time of emergence during primate evolution. This work thus represents a novel application of population genetic and transcriptomic data to an evolutionary study of a large family of transcription factors, resulting in insights that will allow future characterization of the regulatory role played by this family of genes.

## MATERIALS AND METHODS

### Manual curation of all human KRAB-containing Zinc-Finger (ZF) genes

All human and mouse KRAB-ZF gene coordinates were obtained as described in Corsinotti et al. (2013). The resulting list was manually checked: from genes containing at least one Zinc-Finger domain and one KRAB domain (based on PFAM annotation, http://pfam.xfam.org), the longest protein-coding transcript was selected (based on Ensembl release 71, http://www.ensembl.org), resulting in 346 human KRAB-ZF genes (Suppl. Table 1). Genomic coordinates were downloaded from Ensembl for all genes as well as for all individual ZF and KRAB domains. The DNA sequences for the ZF domains were then translated into amino acid sequences using EMBOSS Transeq web-server (http://www.ebi.ac.uk/Tools/st/emboss_transeq/). As Ensembl annotation is automated, the start and end coordinates of the ZF domain may periodically be incorrect. We thus performed an extra check to ensure that the start and end of the well-characterized Zinc Finger domains correspond to the consensus sequence of a Zinc Finger (XX-C-XX-C-XXXXX**X**X**XX**XX**X**-H-XXX-H). If the protein sequence did not match the consensus sequence, we corrected the DNA coordinates in such a way that every ZF domain has the correct coordinate. Given that all further analyses depended on the accuracy of these datasets, annotation of the different domains was particularly rigorous.

In the ZF consensus sequence, positions -1,2,3, and 6 (marked in bold) are the putative DNA-binding amino acids and were therefore treated specially within the ZF domains.We only kept complete (containing all 23 amino acids) and perfect (containing at least a C2 or H2 signature) ZFs. All degenerate and atypical ZF domains were removed for downstream analyses. In total, 733 KRAB and 3909 ZF domains were used.

**Polymorphism data**

Human SNP data were obtained from the 1000 Genomes Consortium phase 1, release version 3 (Consortium 2012). Variant Calling Format (.vcf) files aligned to the human reference genome (hg19) were downloaded for all KRAB-ZF genes with tabix-0.2.6. We included 1092 individuals from 14 populations. Only high quality SNPs were kept and indels were removed, resulting in a total of 97,465 SNPs. Filtering was carried out using vcftools version 0.1.7 (Danecek et al. 2011) with the following parameters: minMQ = 10, minGQ = 40, minDP = 5, and minQ = 100. All variants marked as "SysErr" and "lowQual" were removed as well. The resulting SNPs were classified according to their correspondence in the KRAB domain, in the ZF domain, or as ZF Binding amino acids. Because of the repetitive nature of the ZF domains, it is feasible that the amount of polymorphism may have been over- or under-estimated. To check for possible biases, we downloaded the mappability tracks available from the UCSC genome browser (hg19). Because the read lengths are a mixture of 36 to more than 100 base pairs, we downloaded four tracks (of lengths 36, 50, 75, and 100 bp) according to their ability to uniquely align to different parts of the genome. In other words, each position in the genome has a mappability score (ranging from 0 to 1, 1 corresponding to a uniquely aligned read) that depends on the length of the short read (36bp reads map less uniquely in the genome than 100bp reads). We investigated whether there is a bias in read mapping and allele frequency. In Suppl. Table 2, we calculated the Spearman correlation between the Minor Allele Frequency (MAF) of the binding site SNPs and the mappability of the reads (for the four different read lengths used by the 1000 Genomes project for SNP calling). There is no significant correlation between the mappability score and the MAF ($p > 0.15$ in all cases). Furthermore, when comparing the mappability of synonymous versus non-synonymous SNPs, there is no significant difference between them (Wilcoxon test, p-value = 0.7722).

**Expression data**

RNA-Seq expression data for three species (humans, chimpanzees, and rhesus macaques) in six tissues (brain and cerebellum separately, heart, kidney, liver, and testis) were obtained from Brawand et al. (2011), in the form of FPKM values (processing steps described therein). Human Embryonic Stem Cell RNA-Seq data was downloaded from the Gene Expression Omnibus with accession number GSE57989 and processed in a similar way.

**Expression breadth and conservation**

Expression conservation describes the degree of conservation of tissue-specific expression between two homologous genes, and was calculated between human-chimpanzee orthologous genes using the Expression Conservation Index (ECI) according to Yang et al. (2005). More specifically, for a given gene, the ECI is equal to the number of tissues where the gene is expressed in both species ("conserved expression") divided by the mean number of tissues with gene expression in humans and in chimpanzees. ECI values range from 0 and 1, where 1 corresponds to a gene with conserved expression in all tissues for the two species.

Expression breadth corresponds to the number of tissue types in which a given gene is expressed above some threshold value. We used a threshold of FPKM > 1 to define a gene as "expressed" in a given tissue.

**Histone data**

We analyzed the H3K9me3 histone mark, which is marking an inactive chromatin state and therefore a repressed gene. Histone modification data, along with their input control for human adult kidney, liver, and heart tissues, were downloaded (in .wig format) from the Epigenomics Project (http://www.ncbi.nlm.nih.gov/epigenomics) with accessions codes: ESX000002152, ESX000002139, ESX000006561, ESX000006547, ESX000005777, ESX000005738. In order to extract only the significantly enriched regions for H3K9me3, only regions with a minimum two-fold signal over the input control and an input signal greater than the cutoff were used (third quartile + 1.5*IQR).

**Orthologous gene and domain annotation**

The annotation of orthologous genes for humans, chimpanzees, and rhesus macaques was downloaded from the Ensembl Web Browser (http://www.ensembl.org). Only 1-to-1 orthologs were kept. Human-mouse orthologous genes were defined as described in Corsinotti et al. (2013).

All human Zinc-Finger and KRAB domains were separately aligned to the chimpanzee (panTro4), rhesus macaque (rheMac2), and mouse (mm10) genomes using the blat software from the UCSC genome browser (https://genome.ucsc.edu/cgi-bin/hgBlat). From the resulting matches, only those belonging to orthologous genes were kept and in cases of multiple matches, manual inspection was used to confirm the correct corresponding ZF domain. Hence, only the best correspondences between the individual ZF and KRAB domains were used for the 4 species, providing exact 1-to-1 correspondence between all of the amino acids of the ZF domains (including the DNA-binding amino acids).

**Tests for selection**

To evaluate the selection history of KRAB-ZF genes, we performed two types of analyses: McDonald-Kreitman tests (MK, 1991) and tests from the Phylogenetic Analysis by Maximum Likelihood (PAML) package (Yang 2007). We used all alignments of the ZF and KRAB domains for the orthologous genes of the four species, as described in the previous paragraph.

For the MK tests, synonymous and non-synonymous divergence was calculated only for the fixed differences between two species (i.e., all human polymorphic positions as defined from the 1000 genomes dataset were excluded). Statistical significance in each contingency table was determined using a chi-square test and a two-tailed Fisher's exact test.

For the second analysis, the codeml package from the PAML suite (version 4.8, Yang 2007) was used to test different models (as described in Simkin et al., 2013). We used all KRAB-ZF genes having 1:1:1:1 orthologs in the four species: humans, chimpanzees, rhesus macaques, and mice ($n$ = 52). Every ZF domain was used for the analysis by concatenating one after the other per gene (i.e., all Zinc-Finger domains per gene were concatenated by excluding the linker residues existing between them). We

evaluated several models: M0 (a site-model with one omega for all branches) compared to the branch-model (omega varying among lineages); site-model 7 (beta distribution with 0 < omega < 1) versus 8 (model M7 plus another site category assessing omega > 1), 8 versus 8a (an alternate null model for M8, with omega fixed at 1), and 1a (nearly neutral) versus 2a (positive selection). Sites evolving under positive selection were defined as having a posterior probability of > 95% for omega being > 1 using the Bayes empirical Bayes method. Lastly, we compared the branch-site neutral model versus the branch-site model (two or more omega values are accepted for the branches). The lineages are separated in to two groups: one "background" lineage evolving neutrally or under negative selection and a "foreground" lineage that may contain some positively selected sites. In all cases, twice the difference of the two log-likelihood values (null versus alternative model) has been compared to a chi-square distribution to assess significance.

The tree structure used for the analyses differed according to the tested model: for the M0, M1a, M2a, M7, M8, and M8a models a rooted tree was utilized. For the Branch model and Branch-sites models, unrooted trees were used (3 different trees according the lineages tested: human-specific, chimp-specific or human-chimp lineage-specific).

**GC content**

GC content data was downloaded from the UCSC genome table browser for the human genome assembly hg19 (http://genome.ucsc.edu/cgi-bin/hgTables?command=start).

**Paralogs**

Paralogs for the KRAB-ZF genes were obtained from the Ensembl website.

# RESULTS

## Expression of orthologous KRAB-ZF genes is species-specific

We investigated gene expression patterns for orthologous genes in six tissues (brain and cerebellum separately, heart, kidney, liver, and testis). Our analysis used RNA-Seq data from Brawand et al. (2011) and focused on three species (humans, chimpanzees, and rhesus macaques) for which we performed manual curation of all KRAB-ZF genes. Using hierarchical clustering (with Spearman correlation), we observe that expression levels of all orthologous genes from the whole transcriptome cluster in a tissue-specific manner (Figure 1a). In other words, gene expression is conserved across the three species for a given tissue. This is fully in accordance with global patterns of gene expression among mammals demonstrated by Brawand et al. (2011), where data is arranged according to tissue. By contrast, when focusing only on KRAB-ZF gene orthologs ($n$ = 238) the clustering becomes species-specific (Figure 1c). The tissue-specific gene expression is lost, suggesting a rapid change in function for the KRAB-ZF family in primates. As a control, we did the same analysis using all transcription factors-orthologous genes for the three species (except zinc fingers, $n$ = 726, downloaded from Animal Transcription Factor Database: http://www.bioguo.org/AnimalTFDB/index.php). Figure 1b reproduces results from Fig 1a: all orthologous genes, but KRAB-ZF, cluster in a tissue-specific manner, while KRAB-ZF gene expression clusters in a species-specific manner, indicating that this family of TFs has very different expression patterns than other transcription factors. Principal-component analysis (PCA, Suppl. Figure 1) reached the same conclusions.

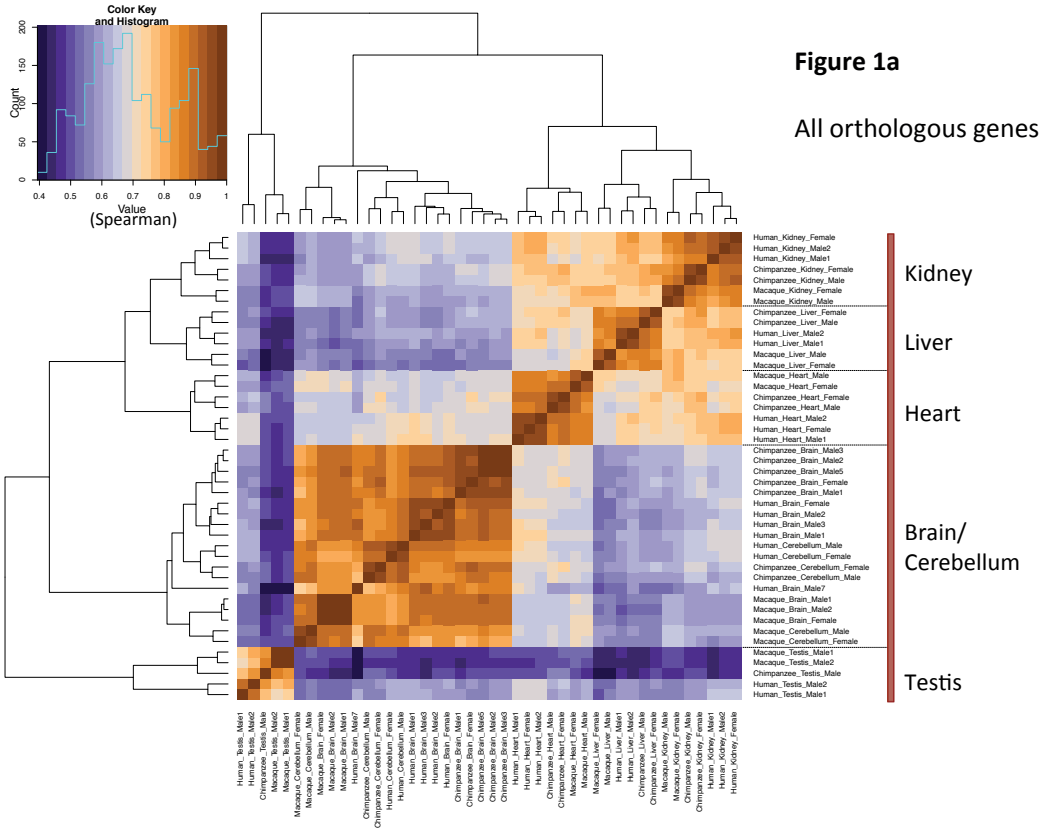



**Figure 1a**

All orthologous genes

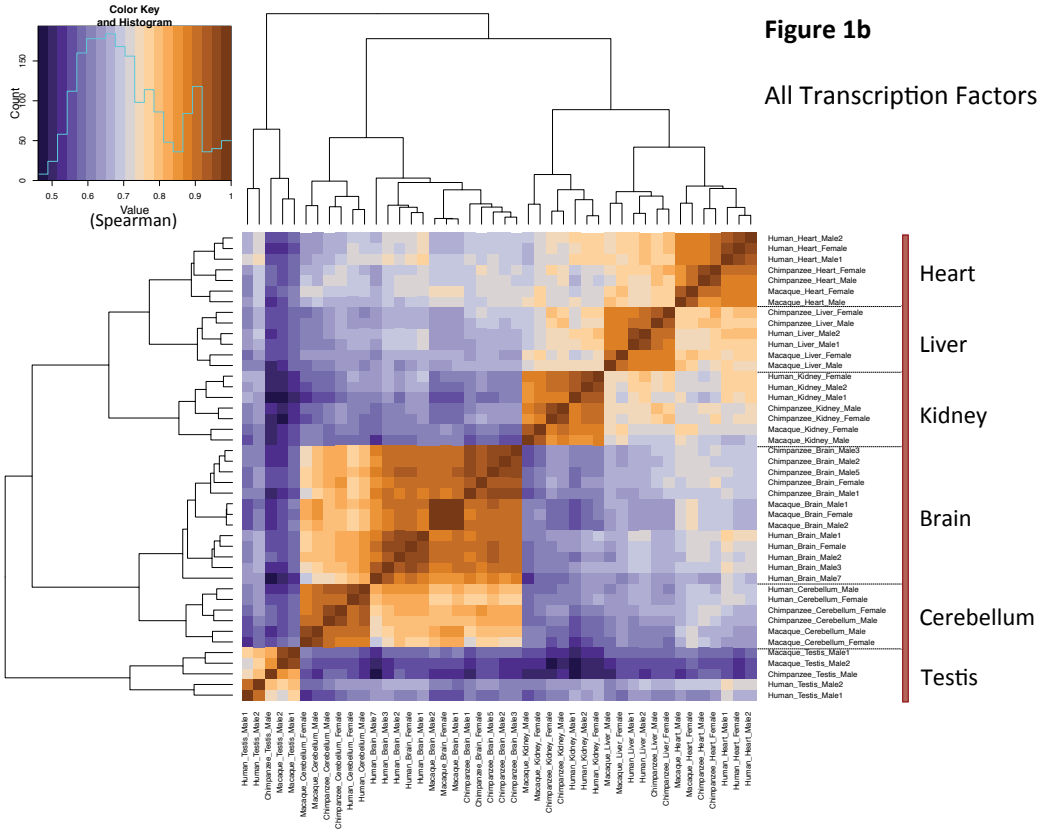



**Figure 1b**

All Transcription Factors

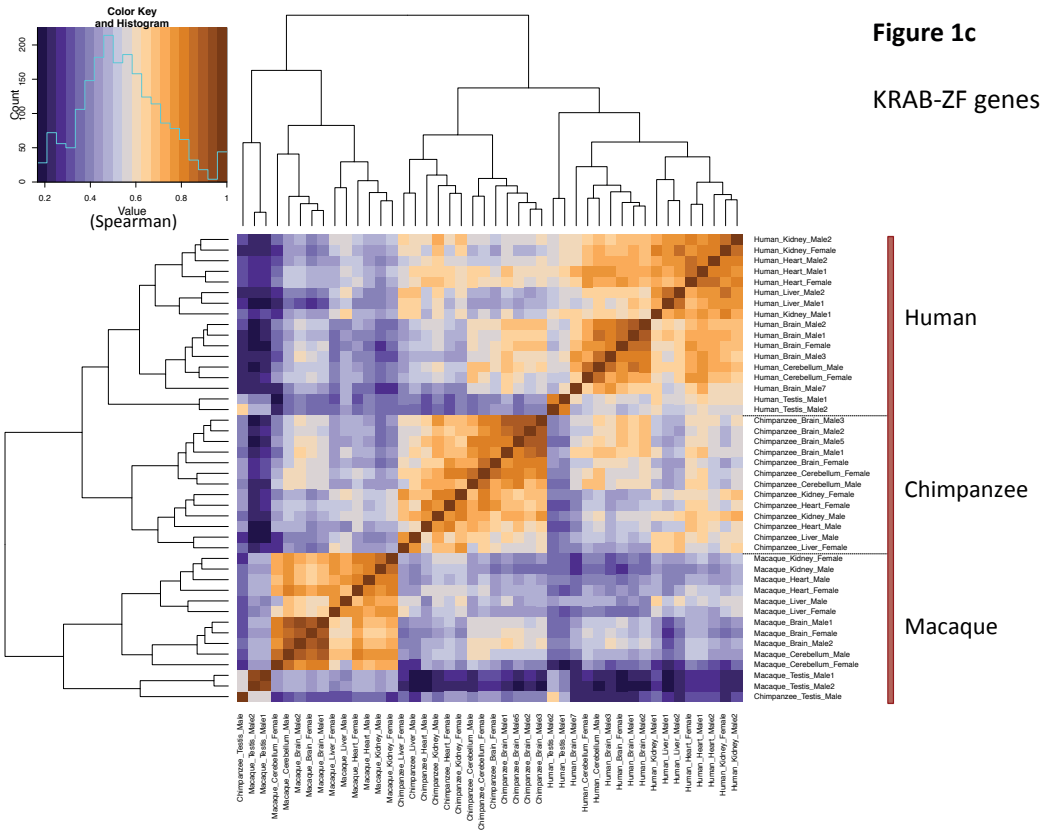



**Figure 1: Correlations of mRNA levels for human, chimpanzee, and rhesus macaque orthologous genes:** Spearman correlation heatmaps and hierarchical clustering for (a) all orthologous genes, (b) all transcription factors orthologous genes (except ZFs) and (c) KRAB-ZF only. The highest Spearman correlation coefficients correspond to brown colors. (a) Expression of all orthologous genes and (b) expression of all human transcription factors cluster according to tissue, with a high Spearman correlation coefficient. (c) Expression of KRAB-ZF genes clusters according to species, with a high Spearman correlation coefficient.

## Expression breadth and expression conservation of KRAB-ZF genes

Many studies highlight the importance of measuring the expression breadth and expression conservation across tissues and organisms when studying evolutionary rates (e.g. Yang et al. 2005; Park and Choi 2010). We calculated the number of genes expressed in all six tissues. Only 29% of KRAB-ZF genes were "expressed" in the six human tissues, whereas 47% of the totality of genes was expressed (with FPKM > 1) in all tissues. As an additional control, we used all human TFs (except the zinc-fingers) to calculate how many are expressed in the six tissues (Table 1). There were significantly fewer KRAB-ZF genes with ubiquitous expression in all tested tissues when compared to

either all transcription factors ($\chi^2$ p < 1.254e-5) or all genes ($\chi^2$ p < 1.948e-8), indicating a narrower pattern of expression for the KRAB-ZFs.

| | Broad expression (in all tissues, FPKM > 1) | Limited expression (in some tissues only) | Percentage (expressed/total) |
|---|---|---|---|
| **KRAB-ZFs** | 68 | 170 | 29% |
| **All TFs (except ZFs)** | 578 | 736 | 44% |
| **All genes** | 7606 | 8548 | 47% |

**Table 1: Expression breadth of KRAB-ZFs, all TFs, and all genes for six human tissues.** The number of genes expressed in all tissues is reported.

We also calculated the ECI (expression conservation index, cf. Methods) for orthologous genes between humans/chimpanzees, and tallied those with an ECI equal to one (i.e., conserved expression in all six tissues for humans and chimpanzees). Results are shown in Table 2. Roughly 16% of KRAB-ZF genes had a conserved expression (i.e., genes expressed in all six tissues in humans and in chimpanzees) whereas 39% of all orthologous genes were conserved ($\chi^2$ p < 8.22e-13). Also, when compared with all TFs, the difference is also significant ($\chi^2$ p < 1.584e-9) and is in accordance with previously reported conservation of tissue-specific gene expression for all orthologous genes (Ramsköld et al. 2009). However, we find that tissue-specific KRAB-ZF gene expression is not as well conserved between the two species. This result indicates that the KRAB-ZF gene family is more narrowly expressed than others and this pattern of expression is not conserved between two closely related species. This can be attributed to the fast evolving expression of KRAB-ZF genes.

| | Expressed in all tissues in humans and chimpanzees (ECI = 1) | Expression not conserved between humans and chimpanzees (ECI < 1) | Percentage (expressed/total) |
|---|---|---|---|
| **KRAB-ZFs** | 38 | 200 | 16% |
| **All TFs (except ZFs)** | 476 | 838 | 36% |
| **All genes** | 6289 | 9865 | 39% |

**Table 2: Expression conservation of KRAB-ZFs, all TFs, and all genes from human and chimpanzee tissues.** The number of genes is reported.

**Expression of KRAB-ZF genes correlates with polymorphism in their Zinc-Finger Binding amino acids**

The Zinc-Finger contacting amino acids correspond to the three positions from the ZF domain contacting the primary strand of the DNA (positions -1, 3, and 6 of the alpha-helix) and one amino acid contacting the secondary strand of the DNA (position 2 of the alpha-helix) (Elrod-Erickson et al. 1998). Those four amino acids are also called the ZF "fingerprint" (Liu et al. 2014). From the 1000 Genomes polymorphism data, we have extracted the SNPs occurring in those four amino acids, and separated the 346 human KRAB-ZF genes into two categories: KRAB-ZF genes with a non-synonymous SNP in at least one of the four contacting amino acids, and KRAB-ZF genes without any non-synonymous SNPs in any of the four contacting amino acids. Fig 2a shows the expression levels between these two categories of KRAB-ZF genes in the six adult tissues and in the human embryonic stem cells (hES).

Human KRAB-ZFs, having non-synonymous polymorphism(s) located in their four binding amino acids, have significantly lower expression levels than those without such polymorphism (Wilcoxon's rank sum test, Benjamini-Hochberg adjusted p-values < 0.05 for all comparisons, Figure 2a). As a control, we also separated the 346 human KRAB-ZFs into two new categories: KRAB-ZF genes with a *synonymous* SNP in at least one of the four contacting amino acids and KRAB-ZF genes without any synonymous SNPs in any of the four contacting amino acids (this category contains both KRAB-ZF genes with non-synonymous SNPs only and those without any SNPs). Figure 2b compares expression levels between these two categories of KRAB-ZF genes in the six adult tissues and the hES cells, observing no difference in expression levels (Wilcoxon's rank sum test). As an additional control, KRAB-ZF genes were separated according to the presence or absence of nonsynonymous polymorphisms in their KRAB domains. Figure 2c illustrates that there is no significant difference in expression levels between the two categories. This re-enforces our conclusion that the presence of a nonsynonymous SNP in a binding site uniquely correlates with the reduced expression of the gene.

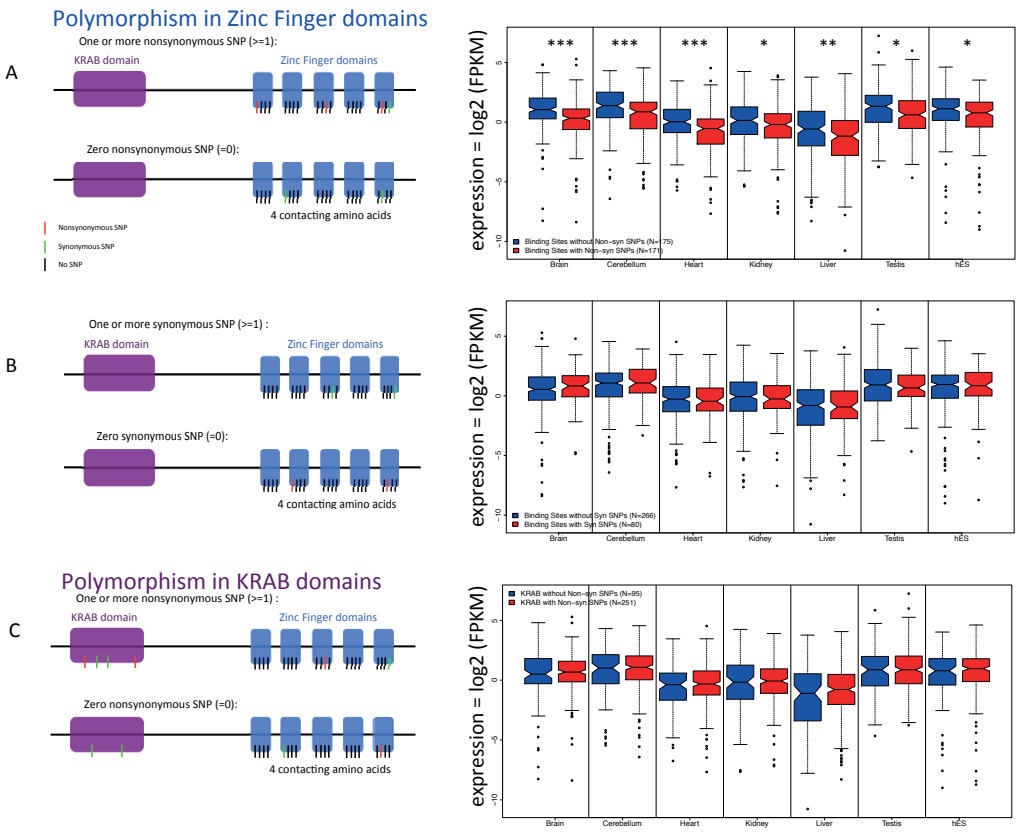



**Figure 2: Comparison of human mRNA levels for two categories of KRAB-ZF genes (with and without nonsynonymous SNP in their DNA-contacting residues):** Expression values of all KRAB-ZF genes with (red boxes) and without (blue boxes) *non-synonymous* polymorphism(s) in at least one of the four binding amino acids (panel a). As a control, in panel b, expression values of all KRAB-ZF genes with (red boxes) and without (blue boxes) *synonymous* polymorphisms in at least one of the four binding amino acids are given. In panel c, expression values of all KRAB-ZF genes with (red boxes) and without (blue boxes) *non-synonymous* polymorphism(s) in the KRAB domain. Accompanying cartoons illustrate examples of the corresponding two categories of KRAB-ZF genes compared. (a) Genes with nonsynonymous SNP(s) in their contacting residues are significantly less expressed in all tested tissues than genes without nonsynonymous SNP in their contacting residues. FDR: < 0.05 (*), < 0.01 (**), < 0.001 (***). (b) There is no significant difference in expression level between genes with synonymous SNP(s) in their contacting residues when compared with genes without synonymous SNP(s) in their contacting residues. (c) There is no significant difference in expression level between genes with nonsynonymous SNP(s) in the KRAB domain when compared with genes without nonsynonymous SNP(s) in the KRAB domain.

To test whether the observed difference in expression may be due to the number of nonsynonymous SNPs present in the genes, we separated the genes in two categories: only/mostly non-synonymous SNPs and only/mostly synonymous SNPs. There is no

significant difference between the two categories regarding their expression levels (Wilcoxon test p-value = 0.06), thus indicating that it is not the number of nonsynonymous SNPs per gene (i.e. nonsynonymous SNP density at the gene-level) but the presence of a nonsynonymous SNP in the binding site only that correlates with the reduced expression.

Finally, we controlled for a possible relationship between the number of Zinc-Fingers per gene and our observed expression differences. We did not found any significant correlation between the number of Zinc-Finger domains per gene and their expression for the 6 tissues and human embryonic stem cells (hESC, Table 3).

| Tissue | Spearman's rho | p-value |
|---|---|---|
| Brain | -0.0548 | 0.3093 |
| Cerebellum | -0.05 | 0.3538 |
| Heart | -0.0757 | 0.16 |
| Kidney | -0.0568 | 0.2923 |
| Liver | -0.082 | 0.1273 |
| Testis | -0.0876 | 0.1038 |
| hES | -0.038 | 0.4819 |

**Table 3: Spearman's correlation coefficient (rho)** and p-values between number of ZF per gene and gene expression for six tissues and hES cells.

## Histone modification H3K9me3 on ZF-coding exon correlates with polymorphism in their Zinc-Finger Binding amino acids

Chromatin immunoprecipitation of histones followed by sequencing (ChIP-Seq) is used to identify chromatin states at very high resolution. The modification of histones changes the DNA compaction, resulting in differences in the accessibility of DNA fragments for transcription factors, and thus influences transcriptional regulation (Tollefsbol 2011). Using publicly available ChIP-Seq data, we analyzed one type of histone modification (H3K9me3, a marker of transcriptionally inactive chromatin) for presence or absence on the ZF-coding exon of all KRAB-ZF genes for the human kidney, liver, heart, and spleen. The 346 human KRAB-ZF genes were separated in the two categories described earlier (KRAB-ZF genes with/without a non-synonymous SNP in at least one of the four contacting amino acids). Figure 3 compares the enrichment of H3K9me3 for the two groups of genes. Results indicate that KRAB-ZF genes bearing

nonsynonymous SNP(s) in one of their four binding amino acids are significantly enriched for repressive histone marks (H3K9me3) than those without such polymorphism. Though this analysis is based on a different dataset (see Methods), it corresponds to the same three tissues used from the RNA-Seq expression results (Figure 2).

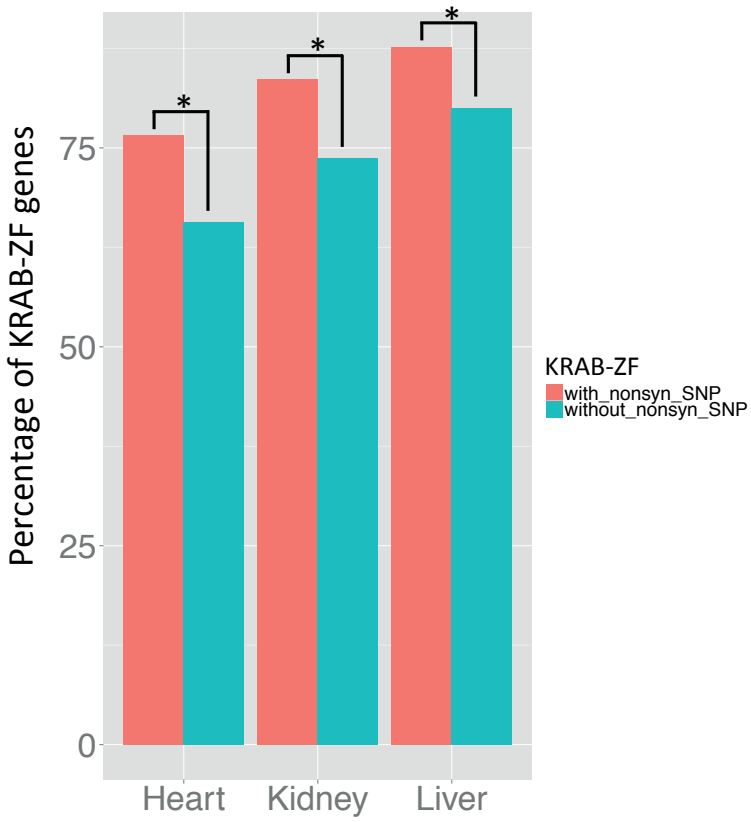



**Figure 3: H3K9me3 on ZF-coding exon.** Comparison of repressive (H3K9me3) histone mark for KRAB-ZF genes with (in red) and without (in green) nonsynonymous SNPs in their four DNA-contacting amino acids. There is a significant enrichment (Fisher's exact test two-tailed p-values < 0.05) of H3K9me3 occupancy in the ZF-coding exon of KRAB-ZF genes carrying a nonsynonymous SNP in their contacting residues, indicating a repressed gene.

## Expression breadth and expression conservation of the two groups of KRAB-ZF genes

We investigated the expression breadth and conservation separately for the two groups of KRAB-ZF genes described above. Only 9/171 KRAB-ZF genes carrying a nonsynonymous SNP in their DNA-recognizing amino acids have conserved expression in all tissues for the two species (i.e., ECI=1), whereas 28/175 genes without

nonsynonymous SNPs meet this criterion (Fisher's exact test, two-tailed, p-value = 0.0015). Similarly, there is a significant difference in the proportion of expression breadth between the two groups of KRAB-ZF genes, with those carrying nonsynonymous SNP(s) in their DNA-recognizing amino acids being less broadly expressed than the others (Fisher's exact test, two-tailed p-value = 0.00038).

**The newest KRAB-ZF genes are enriched for nonsynonymous SNPs in their contacting amino acids relative to older KRAB-ZF genes**

Jacobs et al. (2014) presented a phylogenetic tree with all KRAB-ZF genes and the lineages on which they emerge. We used these data to infer the number of genes emerging in the Primate, Simian/Catarrhine, and Hominoid/Hominid lineages having nonsynonymous polymorphism in their binding amino acids (Figure 4a). 70% of the total genes that emerged in the Hominoid/Hominid lineage have nonsynonymous SNPs in the binding amino acids, whereas genes that emerged during the primate lineage are more constrained (47% contain a nonsynonymous SNP). This indicates that older KRAB-ZF genes may be experiencing stronger purifying selection to maintain their four-contacting amino acids. Another indicator of such constraint is their allele frequency; in Figure 4b, the minor allele frequencies (MAFs) of the nonsynonymous SNPs (only in the four contacting residues) for the three categories of KRAB-ZF genes are plotted according to the lineage on which they appear. Interestingly, nonsynonymous SNPs from KRAB-ZF genes emerging in the Hominoid/Hominid lineage have a significantly higher MAF than SNPs from genes emerging in older lineages (Wilcoxon Mann-Whitney p-values < 0.01). This result is consistent with stronger selective constraints acting on the oldest members of the KRAB-ZF family.

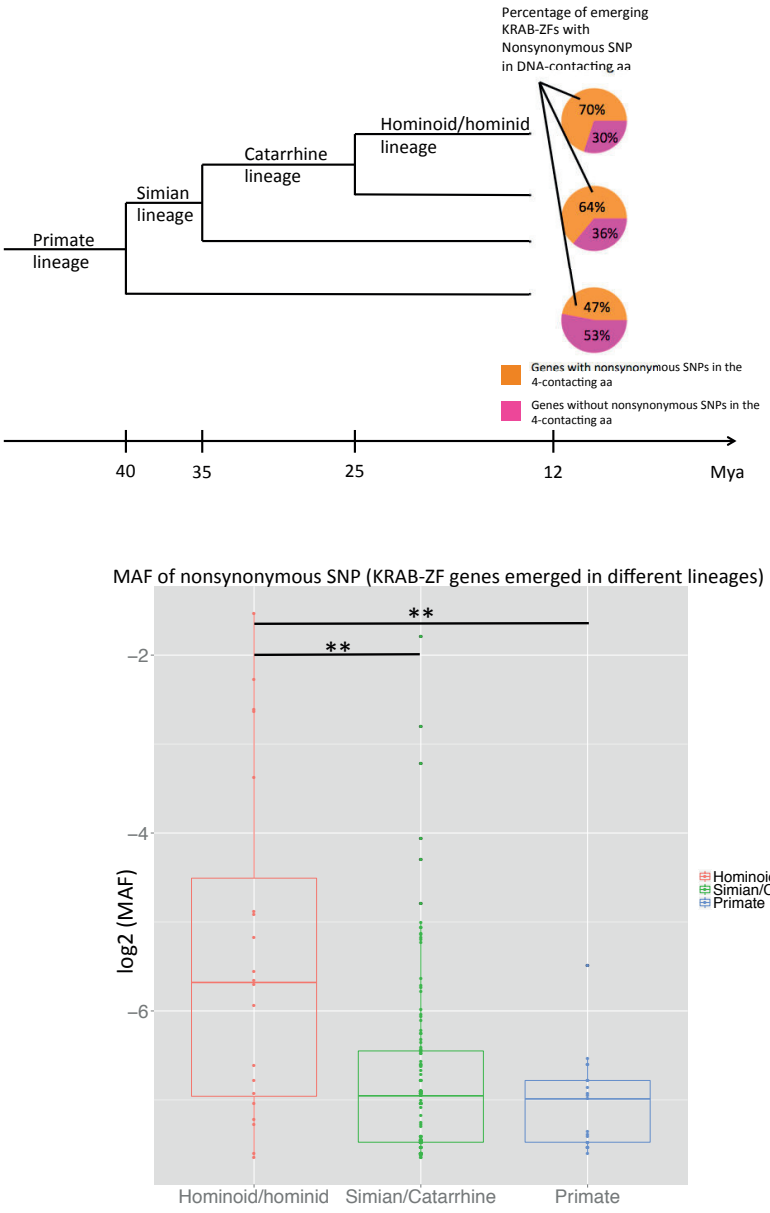



Figure 4

---

**Figure 4: Minor allele frequency (MAF) and number of KRAB-ZF genes emerging in different lineages.** a) Proportion of KRAB-ZF genes with/without a nonsynonymous SNP in their contacting residues emerging in recent lineages. In total 70% of the genes emerging in the Hominoid/Hominid

lineage carry a nonsynonymous SNP in their binding residues, 64% in the Simian/Catarrhine lineage, and only 47% in the primate lineage, indicating a potential relaxation of selective constraint for genes emerging in the most recent lineages. b) Minor Allele Frequencies (MAF) of nonsynonymous SNPs (in the four contacting residues). SNPs from genes emerging in the Hominoid/Hominid lineage have a significantly higher MAF than SNPs from genes emerging in older lineages. Both a) and b) demonstrate the strong selective constraint acting on older genes to maintain their binding residues, thus indicating strong functional relevance. Conversely, contacting residues from younger genes seem to be under weaker purifying selection, potentially because of the lack of a specific target.

## Evolutionary analysis of orthologous KRAB-ZF genes

To investigate the selective pressures acting on the KRAB-ZF genes, we performed two different analyses. All amino acids present in the Zinc-Finger domains were tested for positive selection using the codeml program implemented in the PAML suite. Three different approaches were implemented (see Materials and Methods).

First, we investigated the possibility that the ratio dN/dS (ratio of nonsynonymous changes to synonymous changes, or omega) of a single branch was different from the rest of the phylogenetic tree (composed of four organisms: humans, chimpanzees, rhesus macaques, and mice). For this, we compared the null site-model (one omega for all lineages) with the branch-model (estimates of omega are produced for each lineage). No significant difference was found between the likelihood values of the two models; therefore, we assumed that the selective pressure for the Zinc-Finger domains does not vary across the phylogeny.

Next, we used three different sites-model comparisons to estimate selective constraints on individual amino acids across the length of the Zinc-Fingers. The comparison of model 7 versus model 8 identified only three genes rejecting neutrality in favor of positive selection (ZNF212, ZNF263, ZNF473). ZNF212 had three individual amino acids with high probability of positive selection according to the Bayes Empirical Bayes method, ZNF263 had four amino acids identified and ZNF473 had no site localized. No sites from the four contacting amino acids were found to be experiencing positive selection. The other two site-model comparisons (M8 versus M8a and M1a versus M2a) did not identify specific sites undergoing positive selection.

Lastly, we tested the hypothesis that positively selected individual sites are present only in specific lineages. We used the comparison of the branch-site model

against the branch-site neutral model. This test did not identify any positively selected site in any lineage.

To estimate levels of between-species divergence, we compared humans with closely related species (chimpanzees and rhesus macaques), as well as with mice. We separated the surveyed fragments into three categories that are likely to differ in the intensity and mode of selection acting on them, namely, the Zinc Finger domains, the KRAB domains, and the four DNA-contacting amino acids. The MK-test is designed to distinguish neutrality in protein-coding genes from negative or positive selection by comparing levels of polymorphism within-species (humans) and divergence between-species (human-chimpanzee, human-macaque, and human-mouse). If the sites evolve neutrally, the ratio of polymorphism to divergence for the *nonsynonymous* sites (dN/dS) should be similar to that for *synonymous* sites (pN/pS). Detailed results of each MK-test are shown in supplementary Table 3. Using all genes pooled together for the ZF domains, there are fewer *nonsynonymous* substitutions between species than *synonymous* substitutions ($dN/dS < pN/pS$, $\chi^2$, p-value < 0.0001), indicating purifying selection, or the purging of deleterious mutations. However, as the zinc-finger domain is highly conserved, comparing the average rate of *synonymous* and *nonsynonymous* substitutions for the whole ZF domain may mask specific positively selected sites. For this reason, we performed a separate MK-test for the four DNA-contacting amino acids, pooling all genes together to gain statistical power. The results remain the same as for the ZF domain ($dN/dS < pN/pS$, Fisher's exact test, two-tailed, p-value < 0.0001) for all three comparisons (human-chimpanzee, human-rhesus macaque, and human-mouse). For the KRAB domain, all MK-tests indicate neutrality for the two comparisons (human/chimpanzee and human/rhesus macaque, Fisher's exact test, two-tailed, p-value > 0.05, $dN/dS \sim pN/pS$). The pattern is different for the comparison with mice, where significant evidence of purifying selection is present ($dN/dS < pN/pS$, Fisher's exact test, two-tailed, p-value < 0.05). Using only genes presenting a *nonsynonymous* SNP in the four contacting amino acids, the test is no longer significant, indicating that the KRAB domain is evolving neutrally for those genes. This result points towards weaker purifying selection acting on this group of genes.

## Discussion

The expression of many orthologous genes appears to be tissue-specific. This has been previously demonstrated in a study of global patterns of gene expression differences among mammals (Brawand et al. 2011). From the same dataset, we focused on the cross-species, cross-tissue expression of KRAB-ZF genes. We found that the expression of orthologous KRAB-ZF genes follows a species-specific pattern rather than a tissue-specific pattern. This finding is in line with previous studies suggesting that KRAB-ZF genes have different tissue preferences in different species (Nowick et al. 2010) and supports the independent expansion and functional diversification of KRAB-ZFs in different vertebrate lineages (Liu et al. 2014). This loss of tissue-specific expression implies a rapid change in function for the KRAB-ZF family in primates, providing additional support for the hypothesis that this family of transcription factors plays a role in speciation by regulating evolutionarily divergent traits (see also Nowick et al., 2013).

Next, we analyzed the breadth and the conservation of expression for the KRAB-ZF genes. We confirmed that the KRAB-ZF genes do not have tissue-conserved expression among species, and are narrowly expressed in only a few tissues. Yang et al. (2005) and Park and Choi (2010) showed that gene expression evolves rapidly for genes expressed in only a limited number of tissues. They also demonstrated that, in many cases, tissue-specific gene expression may be transient and not evolutionarily stable. Our results support the hypothesis that the expression of KRAB-ZF genes is fast evolving in primates and this alteration in gene regulatory networks is playing a major role in primate evolution. New endogenous retroelements (EREs) are continuously emerging during evolution and their expression needs to be constrained in a tissue-specific manner. Thus, it is important for the organism to have a fast-evolving modular system capable of regulating retroelement expression at precise developmental stages and in a tissue-specific manner. Thus, KRAB-ZFs are good candidates to control aberrant expression of EREs.

Given that the expression of KRAB-ZF genes is rapidly evolving, we next evaluated models of selection at the nucleotide level. Both the MK test and PAML found that the KRAB and zinc-finger domains are evolving under purifying selection. This

conclusion aligns with previous results, which have demonstrated that orthologs of each KRAB-ZF are subject to negative constraint across the entire set of DNA-binding domains to retain its DNA-binding specificity (Thomas and Schneider 2011), with the nucleotide contacting residues being amongst the slowest evolving (Thomas and Schneider 2011). Also, there is evidence of selection against common SNPs at DNA-contacting amino acids given that substitutions in the DNA-contacting positions could alter the DNA-binding specificity of the KRAB-ZF protein and disrupt the transcription factor function (Lockwood et al. 2014). However, studies on KRAB-ZF paralogous genes show evidence for a very short period of positive selection occurring just after duplication, followed by a long period of strong purifying selection (Thomas and Schneider 2011). Thus, signals of positive selection driving the acquisition of new DNA-binding specificities may be obscured by subsequent purifying selection to maintain those specificities (Emerson and Thomas 2009).

Since the expression divergence of KRAB-ZF genes seems to be an important parameter in their evolutionary process (Nowick et al. 2010), and because the drive for novelty in their function may be based on alterations of their DNA-contacting amino acids, we studied the expression of KRAB-ZF genes in the light of polymorphism in their four binding residues. We divided the 346 human KRAB-ZF genes in to two categories: the ones bearing a nonsynonymous polymorphism in at least one of their DNA-contacting amino acids (171 genes in total) and the ones without nonsynonymous polymorphism(s) in any of their DNA-contacting amino acids (175 genes in total). We found that the average expression of the 171 genes having at least one non-synonymous SNP was significantly lower. We extend this result using another dataset of histone ChIP-Seq that showed enrichment of repressive histone marks in the ZF region of the 171 KRAB-ZFs compared with genes without nonsynonymous SNPs. Comparison of global GC content also supports this result, where genes with lower expression have a smaller percentage of GCs. These findings shed light on the relationship between KRAB-ZF gene expression and the presence of polymorphisms in their zinc finger binding amino acids.

By searching for more elements differentiating the two groups of KRAB-ZF genes (cf. Table 4), we discovered that the KRAB-ZFs with nonsynonymous SNP(s) in their binding site(s) have significantly fewer mouse orthologs than those without, which could be a consequence of their younger age. At the same time, they have more paralogs and ZF domains per gene on average, indicating formation by recent gene duplication

(Emerson and Thomas 2009). Further investigation confirmed that KRAB-ZF genes emerging in the Simian, Catarrhine, and Hominoid/hominid lineages were enriched for genes presenting a nonsynonymous SNP in their contacting residues (Fisher's exact test two-tailed p-value = 6.4e-5). Those SNPs have a significantly higher minor allele frequency (MAF), indicating a relaxation of strong purifying selection for the younger KRAB-ZF genes - as also observed by the nonsynonymous SNPs in their binding residues. In contrast, only 47% of genes emerging in the primate lineage bear a nonsynonymous SNP in their contacting amino acids and have a significantly lower MAF, strongly suggesting the action of purifying selection.

| **Comparison** | **KRAB-ZFs with *nonsynonymous* SNPs in their DNA-contacting amino acids** | **KRAB-ZFs without *nonsynonymous* SNPs in their DNA-contacting amino acids** | **P-value** |
|---|---|---|---|
| Expression level (FPKM) | Less expressed | More expressed | < 0.05 |
| H3K9me3 on the ZF-coding exon | More present | Less present | < 0.05 |
| ECI and expression breadth | Narrowly expressed (i.e. tissue expression evolves rapidly) | Broadly expressed (i.e. tissue expression more conserved) | 0.0015 |
| GC content | Lower GC content (average = 42%, i.e. less expressed) | Higher GC content (average = 43%, i.e. more expressed) | 0.03 |
| Number of orthologous genes human/mouse | Fewer mouse orthologs (i.e. younger) | More mouse orthologs (i.e. older) | 0.00047 |
| Number of paralogs per gene | More paralogs (average = 25/gene) | Fewer paralogs (average = 21/gene) | 0.01 |
| Number of zinc-finger domains per gene | More ZF domains/gene (average = 12 ZFs/gene, i.e. more newly formed ZF domains) | Fewer ZF domains/gene (average = 10 ZFs/gene, i.e. older ZF domains) | $6.5 * 10^{-5}$ |
| Emergence in lineage | Simian, Catarrhine or Hominoid/Hominid lineage | Primate lineage | $6.4 * 10^{-5}$ |

**Table 4: Differences between the two groups of KRAB-ZF genes (with or without nonsynonymous SNP(s) in the four DNA-contacting amino acids).** The group having nonsynonymous SNP(s) is globally

less expressed, with repressive histone marks occupying their gene body, and less GC content. In addition, they appear to be younger, generally emerging in the Simian, Catarrhine or Hominoid/Hominid lineage, thus having fewer mouse orthologs and more paralogs and zinc finger domains per gene.

In summary, through analyses combining transcriptomic data, histone-modification marks, and population genetics, we conclude that human KRAB-ZF genes can be separated in to two categories according to the type of polymorphisms located within their four DNA-contacting residues. Genes without nonsynonymous polymorphism(s) seem to be the oldest members of this family and are significantly more expressed in humans, indicating that members of this sub-group are essential for the organism and therefore are highly conserved. The second category contains newer KRAB-ZFs, with significantly lower expression in all tested tissues and, in human populations, frequent polymorphisms present in their binding sites. Because EREs mutate in order to escape the KRAB-ZF control, slight changes in the four DNA-contacting residues provide the opportunity for the KRAB-ZF genes to re-create a new DNA-binding fingerprint able to control this newly generated binding site. Genetic diversity is generated very quickly from existing contacting residues, providing ground for fine-tuning of their DNA-binding specificity, without having a deleterious effect on the fitness of the organism. This reduced expression enables them to make slight modifications of their DNA-contacting residues and eventually establish high affinity between zinc finger residues and binding site. Since little is known about where these proteins bind, which zinc fingers they use or which genes they regulate, future results on their targets will reveal more about this family and its members' putative function.

## Author's contributions

Conceived and designed the experiments: AK LM AW DT JDJ. Analyzed the data: AK LM. Wrote the paper: AK JDJ.

## ACKNOWLEDGEMENTS

We are grateful to Stefan Laurent, Kristen Irwin, Simon Quenneville, and Anamaria Necsulea for discussions and manuscript comments. This study was supported by grants from the Swiss National Science Foundation and a European Research Council (ERC) Starting Grant to JDJ.

## REFERENCES

Brawand, D., M. Soumillon, A. Necsulea, P. Julien, G. Csárdi, P. Harrigan, M. Weier, A. Liechti, A. Aximu-Petri, M. Kircher, F. W. Albert, U. Zeller, P. Khaitovich, F. Grützner, S. Bergmann, R. Nielsen, S. Pääbo, and H. Kaessmann. 2011. The evolution of gene expression levels in mammalian organs. Nature 478:343–348.

Choo, Y., and A. Klug. 1994. Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage. Proc. Natl. Acad. Sci. 91:11163–11167.

Consortium, T. 1000 G. P. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56–65.

Constantinou-Deltas, C. D., J. Gilbert, R. J. Bartlett, M. Herbstreith, A. D. Roses, and J. E. Lee. 1992. The identification and characterization of KRAB-domain-containing zinc finger proteins. Genomics 12:581–589.

Corsinotti, A., A. Kapopoulou, C. Gubelmann, M. Imbeault, F. R. Santoni de Sio, H. M. Rowe, Y. Mouscaz, B. Deplancke, and D. Trono. 2013. Global and Stage Specific Patterns of Krüppel-Associated-Box Zinc Finger Protein Gene Expression in Murine Early Embryonic Cells. PLoS ONE 8:e56721.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, and R. Durbin. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

Elrod-Erickson, M., T. E. Benson, and C. O. Pabo. 1998. High-resolution structures of variant Zif268–DNA complexes: implications for understanding zinc finger–DNA recognition. Structure 6:451–464.

Emerson, R. O., and J. H. Thomas. 2009. Adaptive Evolution in Zinc Finger Transcription Factors. PLoS Genet 5:e1000325.

Hamilton, A. T., S. Huntley, M. Tran-Gyamfi, D. M. Baggott, L. Gordon, and L. Stubbs. 2006. Evolutionary expansion and divergence in the ZNF91 subfamily of primate-specific zinc finger genes. Genome Res. 16:584–594.

Huntley, S., D. M. Baggott, A. T. Hamilton, M. Tran-Gyamfi, S. Yang, J. Kim, L. Gordon, E. Branscomb, and L. Stubbs. 2006. A comprehensive catalog of human KRAB-associated zinc finger genes: Insights into the evolutionary history of a large family of transcriptional repressors. Genome Res. 16:669–677.

Jacobs, F. M. J., D. Greenberg, N. Nguyen, M. Haeussler, A. D. Ewing, S. Katzman, B. Paten, S. R. Salama, and D. Haussler. 2014. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. Nature 516:242–245.

Kim, C. A., and J. M. Berg. 1996. A 2.2 A resolution crystal structure of a designed zinc finger protein bound to DNA. Nat. Struct. Biol. 3:940–945.

Liu, H., L.-H. Chang, Y. Sun, X. Lu, and L. Stubbs. 2014. Deep vertebrate roots for mammalian zinc finger transcription factor subfamilies. Genome Biol. Evol. evu030.

Lockwood, S. H., A. Guan, A. S. Yu, C. Zhang, A. Zykovich, I. Korf, B. Rannala, and D. J. Segal. 2014. The Functional Significance of Common Polymorphisms in Zinc Finger Transcription Factors. G3 GenesGenomesGenetics 4:1647–1655.

McDonald, J. H., and M. Kreitman. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652–654.

Nowick, K., M. Carneiro, and R. Faria. 2013. A prominent role of KRAB-ZNF transcription factors in mammalian speciation? Trends Genet. 29:130–139.

Nowick, K., A. T. Hamilton, H. Zhang, and L. Stubbs. 2010. Rapid Sequence and Expression Divergence Suggest Selection for Novel Function in Primate-Specific KRAB-ZNF Genes. Mol. Biol. Evol. 27:2606–2617.

Park, S. G., and S. S. Choi. 2010. Expression breadth and expression abundance behave differently in correlations with evolutionary rates. BMC Evol. Biol. 10:241.

Pavletich, N. P., and C. O. Pabo. 1991. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. Science 252:809–817.

Ramsköld, D., E. T. Wang, C. B. Burge, and R. Sandberg. 2009. An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data. PLoS Comput Biol 5:e1000598.

Shannon, M., A. T. Hamilton, L. Gordon, E. Branscomb, and L. Stubbs. 2003. Differential Expansion of Zinc-Finger Transcription Factor Loci in Homologous Human and Mouse Gene Clusters. Genome Res. 13:1097–1110.

Simkin, A., A. Wong, Y.-P. Poh, W. E. Theurkauf, and J. D. Jensen. 2013. Recurrent and recent selective sweeps in the piRNA pathway. Evol. Int. J. Org. Evol. 67:1081–1090.

Spivakov, M., J. Akhtar, P. Kheradpour, K. Beal, C. Girardot, G. Koscielny, J. Herrero, M. Kellis, E. E. Furlong, and E. Birney. 2012. Analysis of variation at transcription factor binding sites in Drosophila and humans. Genome Biol. 13:R49.

Thomas, J. H., and R. O. Emerson. 2009. Evolution of C2H2-zinc finger genes revisited. BMC Evol. Biol. 9:51.

Thomas, J. H., and S. Schneider. 2011. Coevolution of retroelements and tandem zinc finger genes. Genome Res. 21:1800–1812.

Tollefsbol, T. O. (ed). 2011. Using ChIP-Seq Technology to Generate High-Resolution Profiles of Histone Modifications - Springer. Humana Press.

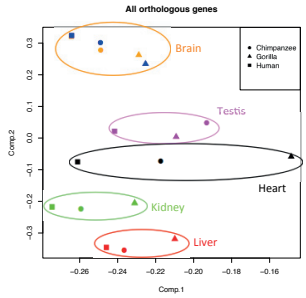
Yang, J., A. I. Su, and W.-H. Li. 2005. Gene Expression Evolves Faster in Narrowly Than in Broadly Expressed Mammalian Genes. Mol. Biol. Evol. 22:2113–2118.

Yang, Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. 24:1586–1591.

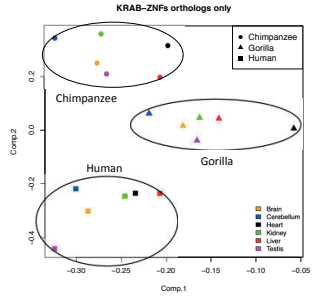
# Supplementary Material

Supplementary Figure 1: Gene expression patterns for human, chimpanzee, and gorilla orthologous genes

a)




b)




Legend: Principal Component Analysis (PCA) on standardized expression values for (a) all orthologous genes and (b) KRAB-ZF only. (a) Expression data from all orthologous genes separates according to tissue. PC1 explains 68% of the variance while PC2 explains 10%. PC2 shows a clear tissue-specific segregation, while PC1 shows partial separation. (b) Data from only the KRAB-ZF orthologous genes (N = 238) separates according to species. PC1 explains 68% of the variance while PC2 explains 9%.

**Suppl. Figure 1: PCA of orthologous gene expression across tissues and across species.**

| Gene symbol | Ensembl Gene ID | Ensembl Protein ID | Ensembl Transcript ID | Chromosome | Gene start | Gene end | # of zinc fingers |
|---|---|---|---|---|---|---|---|
| ZNF436 | ENSG00000125945 | ENSP00000313582 | ENST00000314011 | 1 | 23685941 | 23695935 | 12 |
| ZNF69B | ENSG00000187801 | ENSP00000399664 | ENST00000411995 | 1 | 40915774 | 40929390 | 9 |
| ZNF642 | ENSG00000187815 | ENSP00000361790 | ENST00000372705 | 1 | 40942887 | 40962015 | 9 |
| ZNF684 | ENSG00000117010 | ENSP00000361784 | ENST00000372699 | 1 | 40997233 | 41013841 | 8 |
| ZNF678 | ENSG00000181450 | ENSP00000440403 | ENST00000397097 | 1 | 227751236 | 227847594 | 14 |
| ZNF695 | ENSG00000197472 | ENSP00000341236 | ENST00000339986 | 1 | 247108849 | 247171395 | 9 |
| ZNF670 | ENSG00000135747 | ENSP00000355459 | ENST00000366503 | 1 | 247108849 | 247242113 | 8 |
| ZNF669 | ENSG00000188295 | ENSP00000342818 | ENST00000343381 | 1 | 247261406 | 247267674 | 9 |
| ZNF124 | ENSG00000196418 | ENSP00000440365 | ENST00000543802 | 1 | 247285277 | 247335318 | 7 |
| ZNF496 | ENSG00000162714 | ENSP00000355454 | ENST00000366498 | 1 | 247460714 | 247495148 | 4 |
| ZNF514 | ENSG00000144026 | ENSP00000295208 | ENST00000295208 | 2 | 95813075 | 95831158 | 7 |
| ZNF2 | ENSG00000163067 | ENSP00000411051 | ENST00000453539 | 2 | 95831177 | 95850065 | 8 |
| ZNF860 | ENSG00000197385 | ENSP00000373274 | ENST00000360311 | 3 | 32023263 | 32033120 | 12 |
| ZNF619 | ENSG00000177873 | ENSP00000411132 | ENST00000447116 | 3 | 40518604 | 40531727 | 10 |
| ZNF620 | ENSG00000177842 | ENSP00000322265 | ENST00000314529 | 3 | 40547483 | 40560227 | 8 |
| ZNF621 | ENSG00000172888 | ENSP00000340841 | ENST00000339296 | 3 | 40566369 | 40616176 | 7 |
| ZNF662 | ENSG00000182983 | ENSP00000329264 | ENST00000328199 | 3 | 42947223 | 42960825 | 8 |
| ZNF445 | ENSG00000185219 | ENSP00000379387 | ENST00000396077 | 3 | 44481262 | 44519162 | 14 |
| ZNF852 | ENSG00000178917 | ENSP00000389841 | ENST00000436261 | 3 | 44540462 | 44552128 | 13 |
| ZNF167 | ENSG00000196345 | ENSP00000273320 | ENST00000273320 | 3 | 44596685 | 44635665 | 13 |
| ZNF197 | ENSG00000186448 | ENSP00000345809 | ENST00000396058 | 3 | 44626380 | 44689963 | 22 |
| ZNF589 | ENSG00000164048 | ENSP00000346729 | ENST00000354698 | 3 | 48282590 | 48340743 | 4 |
| ZNF717 | ENSG00000227124 | ENSP00000409514 | ENST00000422325 | 3 | 75758794 | 75834734 | 18 |
| ZNF732 | ENSG00000186777 | ENSP00000415774 | ENST00000419098 | 4 | 264464 | 299110 | 9 |
| ZNF141 | ENSG00000131127 | ENSP00000240499 | ENST00000240499 | 4 | 331603 | 378653 | 10 |
| RP11-1396O13.13.1 | ENSG00000219492 | ENSP00000421652 | ENST00000508324 | 4 | 9385743 | 9390709 | 2 |
| PRDM9 | ENSG00000164256 | ENSP00000296682 | ENST00000296682 | 5 | 23507264 | 23528706 | 13 |
| ZNF300 | ENSG00000145908 | ENSP00000397178 | ENST00000446148 | 5 | 150273954 | 150284545 | 12 |
| ZNF354A | ENSG00000169131 | ENSP00000337122 | ENST00000335815 | 5 | 178138593 | 178157703 | 13 |
| ZNF354B | ENSG00000178338 | ENSP00000327143 | ENST00000322434 | 5 | 178286954 | 178315123 | 13 |
| ZNF454 | ENSG00000178187 | ENSP00000326249 | ENST00000320129 | 5 | 178368192 | 178393434 | 12 |
| ZNF879 | ENSG00000234284 | ENSP00000414887 | ENST00000444149 | 5 | 178450753 | 178462065 | 13 |
| ZNF354C | ENSG00000177932 | ENSP00000324064 | ENST00000315475 | 5 | 178487416 | 178510538 | 11 |
| ZNF184 | ENSG00000096654 | ENSP00000211936 | ENST00000211936 | 6 | 27418522 | 27440897 | 19 |
| ZNF192 | ENSG00000198315 | ENSP00000332750 | ENST00000330236 | 6 | 28109716 | 28124089 | 9 |
| ZKSCAN4 | ENSG00000187626 | ENSP00000366509 | ENST00000377294 | 6 | 28212401 | 28227011 | 7 |
| ZKSCAN3 | ENSG00000189298 | ENSP00000252211 | ENST00000252211 | 6 | 28317691 | 28335336 | 7 |
| ZNF311 | ENSG00000197935 | ENSP00000366384 | ENST00000377179 | 6 | 28962562 | 28973387 | 14 |
| ZFP57 | ENSG00000204644 | ENSP00000418259 | ENST00000488757 | 6 | 29640169 | 29648887 | 6 |
| RBAK | ENSG00000146587 | ENSP00000275423 | ENST00000353796 | 7 | 5023349 | 5112854 | 14 |
| ZNF12 | ENSG00000164631 | ENSP00000385939 | ENST00000405858 | 7 | 6728064 | 6746554 | 15 |
| ZNF713 | ENSG00000178665 | ENSP00000416662 | ENST00000429591 | 7 | 55955169 | 56009918 | 5 |
| ZNF479 | ENSG00000185177 | ENSP00000333776 | ENST00000331162 | 7 | 57187321 | 57207571 | 11 |
| ZNF716 | ENSG00000182111 | ENSP00000394248 | ENST00000420713 | 7 | 57509883 | 57533265 | 9 |

| ZNF727 | ENSG00000257482 | ENSP00000447987 | ENST00000550760 | 7 | 63505821 | 63538927 | 10 |
|---|---|---|---|---|---|---|---|
| ZNF679 | ENSG00000197123 | ENSP00000255746 | ENST00000255746 | 7 | 63688852 | 63727309 | 7 |
| ZNF736 | ENSG00000234444 | ENSP00000347210 | ENST00000355095 | 7 | 63767837 | 63810017 | 9 |
| ZNF680 | ENSG00000173041 | ENSP00000309330 | ENST00000309683 | 7 | 63980262 | 64023484 | 12 |
| ZNF138 | ENSG00000197008 | ENSP00000303533 | ENST00000307355 | 7 | 64254766 | 64294054 | 2 |
| ZNF273 | ENSG00000198039 | ENSP00000418719 | ENST00000476120 | 7 | 64330550 | 64391344 | 10 |
| ZNF92 | ENSG00000146757 | ENSP00000332595 | ENST00000328747 | 7 | 64838712 | 64866038 | 11 |
| ZNF394 | ENSG00000160908 | ENSP00000337363 | ENST00000337673 | 7 | 99084142 | 99097947 | 6 |
| ZKSCAN5 | ENSG00000196652 | ENSP00000322872 | ENST00000326775 | 7 | 99101607 | 99132323 | 12 |
| ZKSCAN1 | ENSG00000106261 | ENSP00000323148 | ENST00000324306 | 7 | 99613204 | 99639312 | 6 |
| ZNF3 | ENSG00000166526 | ENSP00000306372 | ENST00000303915 | 7 | 99661656 | 99680171 | 8 |
| ZNF786 | ENSG00000197362 | ENSP00000417470 | ENST00000491431 | 7 | 148766735 | 148787874 | 13 |
| ZNF425 | ENSG00000204947 | ENSP00000367300 | ENST00000378061 | 7 | 148799876 | 148823438 | 19 |
| ZNF398 | ENSG00000197024 | ENSP00000439340 | ENST00000540950 | 7 | 148823508 | 148880116 | 7 |
| ZNF282 | ENSG00000170265 | ENSP00000262085 | ENST00000262085 | 7 | 148892554 | 148923339 | 5 |
| ZNF212 | ENSG00000170260 | ENSP00000338572 | ENST00000335870 | 7 | 148936742 | 148952700 | 4 |
| ZNF783 | ENSG00000204946 | ENSP00000410890 | ENST00000434415 | 7 | 148959262 | 148994393 | 4 |
| ZNF777 | ENSG00000196453 | ENSP00000247930 | ENST00000247930 | 7 | 149128454 | 149158214 | 9 |
| ZNF746 | ENSG00000181220 | ENSP00000395007 | ENST00000458143 | 7 | 149169885 | 149194908 | 3 |
| ZNF596 | ENSG00000172748 | ENSP00000310033 | ENST00000308811 | 8 | 182137 | 197342 | 11 |
| ZNF705G | ENSG00000215372 | ENSP00000445477 | ENST00000400078 | 8 | 7213039 | 7243080 | 2 |
| ZNF705B | ENSG00000215356 | ENSP00000382987 | ENST00000400120 | 8 | 7783859 | 7812271 | 3 |
| ZNF705D | ENSG00000215343 | ENSP00000382957 | ENST00000400085 | 8 | 11961898 | 11973025 | 3 |
| ZNF707 | ENSG00000181135 | ENSP00000351482 | ENST00000358656 | 8 | 144766622 | 144796068 | 7 |
| ZNF251 | ENSG00000198169 | ENSP00000292562 | ENST00000292562 | 8 | 145946298 | 145981802 | 12 |
| ZNF34 | ENSG00000196378 | ENSP00000341528 | ENST00000343459 | 8 | 145997611 | 146012730 | 12 |
| ZNF517 | ENSG00000197363 | ENSP00000353058 | ENST00000359971 | 8 | 146024261 | 146036554 | 9 |
| ZNF7 | ENSG00000147789 | ENSP00000393260 | ENST00000446747 | 8 | 146052849 | 146072894 | 14 |
| ZNF250 | ENSG00000196150 | ENSP00000292579 | ENST00000292579 | 8 | 146076632 | 146127553 | 13 |
| ZNF658 | ENSG00000196409 | ENSP00000366853 | ENST00000377626 | 9 | 40760700 | 40836415 | 19 |
| ZNF484 | ENSG00000127081 | ENSP00000378882 | ENST00000395506 | 9 | 95607874 | 95640304 | 15 |
| ZNF169 | ENSG00000175787 | ENSP00000378792 | ENST00000395395 | 9 | 97021593 | 97063736 | 11 |
| ZNF510 | ENSG00000081386 | ENSP00000223428 | ENST00000223428 | 9 | 99518147 | 99540411 | 9 |
| ZNF782 | ENSG00000196597 | ENSP00000419397 | ENST00000481138 | 9 | 99578754 | 99637905 | 11 |
| ZNF189 | ENSG00000136870 | ENSP00000342019 | ENST00000339664 | 9 | 104161155 | 104172942 | 16 |
| ZNF483 | ENSG00000173258 | ENSP00000311679 | ENST00000309235 | 9 | 114287439 | 114340124 | 11 |
| ZFP37 | ENSG00000136866 | ENSP00000452552 | ENST00000553380 | 9 | 115800660 | 115819039 | 12 |
| ZNF79 | ENSG00000196152 | ENSP00000362446 | ENST00000342483 | 9 | 130186661 | 130207651 | 11 |
| ZNF248 | ENSG00000198105 | ENSP00000349882 | ENST00000357328 | 10 | 38091751 | 38147034 | 7 |
| ZNF25 | ENSG00000175395 | ENSP00000302222 | ENST00000302609 | 10 | 38238500 | 38265561 | 12 |
| ZNF33A | ENSG00000189180 | ENSP00000402467 | ENST00000432900 | 10 | 38299578 | 38356282 | 16 |
| ZNF37A | ENSG00000075407 | ENSP00000329141 | ENST00000351773 | 10 | 38383264 | 38412276 | 10 |
| ZNF33B | ENSG00000196693 | ENSP00000352444 | ENST00000359467 | 10 | 43069633 | 43134018 | 16 |
| ZNF487P | ENSG00000243660 | ENSP00000388421 | ENST00000431662 | 10 | 43932282 | 43991517 | 3 |
| ZNF485 | ENSG00000198298 | ENSP00000354694 | ENST00000361807 | 10 | 44101855 | 44113351 | 11 |
| ZNF195 | ENSG00000005801 | ENSP00000382511 | ENST00000399602 | 11 | 3360491 | 3400448 | 9 |
| ZNF215 | ENSG00000149054 | ENSP00000278319 | ENST00000278319 | 11 | 6947635 | 7005863 | 4 |
| ZNF214 | ENSG00000149050 | ENSP00000278314 | ENST00000278314 | 11 | 7020549 | 7041599 | 10 |
| ZNF202 | ENSG00000166261 | ENSP00000337724 | ENST00000336139 | 11 | 123594885 | 123612383 | 8 |

| ZNF705A | ENSG00000196946 | ENSP00000352233 | ENST00000359286 | 12 | 8290733 | 8332642 | 3 |
|---|---|---|---|---|---|---|---|
| ZNF641 | ENSG00000167528 | ENSP00000301042 | ENST00000301042 | 12 | 48733791 | 48745197 | 5 |
| ZNF605 | ENSG00000196458 | ENSP00000376135 | ENST00000392321 | 12 | 133498047 | 133532892 | 17 |
| ZNF26 | ENSG00000198393 | ENSP00000333725 | ENST00000328654 | 12 | 133562951 | 133589154 | 13 |
| ZNF84 | ENSG00000198040 | ENSP00000331465 | ENST00000327668 | 12 | 133613878 | 133639885 | 19 |
| ZNF140 | ENSG00000196387 | ENSP00000347755 | ENST00000355557 | 12 | 133656424 | 133684130 | 10 |
| ZNF891 | ENSG00000214029 | ENSP00000380480 | ENST00000397313 | 12 | 133694740 | 133707059 | 7 |
| ZNF10 | ENSG00000256223 | ENSP00000248211 | ENST00000248211 | 12 | 133707161 | 133736051 | 10 |
| ZNF268 | ENSG00000090612 | ENSP00000444412 | ENST00000536435 | 12 | 133707570 | 133783698 | 24 |
| ZNF205 | ENSG00000122386 | ENSP00000219091 | ENST00000219091 | 16 | 3162561 | 3170518 | 8 |
| ZNF213 | ENSG00000085644 | ENSP00000380087 | ENST00000396878 | 16 | 3185057 | 3192804 | 5 |
| ZNF263 | ENSG00000006194 | ENSP00000219069 | ENST00000219069 | 16 | 3313800 | 3341460 | 9 |
| ZNF75A | ENSG00000162086 | ENSP00000459566 | ENST00000574298 | 16 | 3355406 | 3368852 | 5 |
| ZNF597 | ENSG00000167981 | ENSP00000301744 | ENST00000301744 | 16 | 3486104 | 3493537 | 7 |
| ZNF500 | ENSG00000103199 | ENSP00000219478 | ENST00000219478 | 16 | 4798240 | 4817219 | 5 |
| ZKSCAN2 | ENSG00000155592 | ENSP00000331626 | ENST00000328086 | 16 | 25247322 | 25269252 | 6 |
| ZNF747 | ENSG00000169955 | ENSP00000441702 | ENST00000535210 | 16 | 30537244 | 30546668 | 4 |
| ZNF764 | ENSG00000169951 | ENSP00000252797 | ENST00000252797 | 16 | 30565085 | 30569819 | 7 |
| ZNF688 | ENSG00000229809 | ENSP00000223459 | ENST00000223459 | 16 | 30580667 | 30584055 | 2 |
| ZNF785 | ENSG00000197162 | ENSP00000378642 | ENST00000395216 | 16 | 30585061 | 30597092 | 7 |
| ZNF689 | ENSG00000156853 | ENSP00000287461 | ENST00000287461 | 16 | 30613879 | 30635333 | 10 |
| ZNF267 | ENSG00000185947 | ENSP00000300870 | ENST00000300870 | 16 | 31885079 | 31929914 | 14 |
| ZFP90 | ENSG00000184939 | ENSP00000381304 | ENST00000398253 | 16 | 68563993 | 68601039 | 13 |
| ZNF19 | ENSG00000157429 | ENSP00000288177 | ENST00000288177 | 16 | 71507493 | 71598992 | 9 |
| ZFP1 | ENSG00000184517 | ENSP00000377080 | ENST00000393430 | 16 | 75182390 | 75206134 | 8 |
| ZNF778 | ENSG00000170100 | ENSP00000405289 | ENST00000433976 | 16 | 89284118 | 89295363 | 14 |
| ZNF18 | ENSG00000154957 | ENSP00000315664 | ENST00000322748 | 17 | 11880762 | 11900785 | 5 |
| ZNF286A | ENSG00000187607 | ENSP00000464218 | ENST00000464847 | 17 | 15602891 | 15640874 | 10 |
| ZNF287 | ENSG00000141040 | ENSP00000379168 | ENST00000395824 | 17 | 16454701 | 16472520 | 14 |
| ZNF624 | ENSG00000197566 | ENSP00000310472 | ENST00000311331 | 17 | 16524051 | 16557158 | 20 |
| ZNF519 | ENSG00000175322 | ENSP00000464872 | ENST00000590202 | 18 | 14057456 | 14132489 | 9 |
| ZNF554 | ENSG00000172006 | ENSP00000321132 | ENST00000317243 | 19 | 2819872 | 2836733 | 7 |
| ZNF555 | ENSG00000186300 | ENSP00000334853 | ENST00000334241 | 19 | 2841433 | 2860472 | 15 |
| ZNF556 | ENSG00000172000 | ENSP00000302603 | ENST00000307635 | 19 | 2867333 | 2878501 | 9 |
| ZNF57 | ENSG00000171970 | ENSP00000303696 | ENST00000306908 | 19 | 2900896 | 2918474 | 13 |
| ZNF77 | ENSG00000175691 | ENSP00000319053 | ENST00000314531 | 19 | 2933216 | 2944969 | 12 |
| ZNF557 | ENSG00000130544 | ENSP00000252840 | ENST00000252840 | 19 | 7069471 | 7087979 | 10 |
| ZNF558 | ENSG00000167785 | ENSP00000301475 | ENST00000301475 | 19 | 8920382 | 8933565 | 9 |
| ZNF317 | ENSG00000130803 | ENSP00000247956 | ENST00000247956 | 19 | 9251056 | 9274090 | 13 |
| ZNF699 | ENSG00000196110 | ENSP00000311596 | ENST00000308650 | 19 | 9404951 | 9415795 | 14 |
| ZNF559 | ENSG00000188321 | ENSP00000377461 | ENST00000393883 | 19 | 9434448 | 9461838 | 9 |
| ZNF177 | ENSG00000188629 | ENSP00000473346 | ENST00000602738 | 19 | 9435021 | 9493293 | 7 |
| ZNF560 | ENSG00000198028 | ENSP00000301480 | ENST00000301480 | 19 | 9577031 | 9609279 | 13 |
| ZNF426 | ENSG00000130818 | ENSP00000253115 | ENST00000253115 | 19 | 9638683 | 9649303 | 11 |
| ZNF561 | ENSG00000171469 | ENSP00000303915 | ENST00000302851 | 19 | 9715356 | 9732075 | 8 |
| ZNF562 | ENSG00000171466 | ENSP00000411784 | ENST00000448622 | 19 | 9759330 | 9785776 | 6 |
| ZNF812 | ENSG00000224689 | ENSP00000395629 | ENST00000457674 | 19 | 9800600 | 9811452 | 7 |
| ZNF846 | ENSG00000196605 | ENSP00000380999 | ENST00000397902 | 19 | 9868151 | 9879410 | 11 |
| ZNF627 | ENSG00000198551 | ENSP00000354414 | ENST00000361113 | 19 | 11708235 | 11729974 | 11 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| ZNF823 | ENSG00000197933 | ENSP00000340683 | ENST00000341191 | 19 | 11832081 | 11849824 | 15 |
| ZNF441 | ENSG00000197044 | ENSP00000350576 | ENST00000357901 | 19 | 11877815 | 11894893 | 14 |
| ZNF440 | ENSG00000171295 | ENSP00000305373 | ENST00000304060 | 19 | 11925099 | 11946016 | 10 |
| ZNF439 | ENSG00000171291 | ENSP00000305077 | ENST00000304030 | 19 | 11959576 | 11980306 | 11 |
| ZNF69 | ENSG00000198429 | ENSP00000402985 | ENST00000429654 | 19 | 11998599 | 12025144 | 13 |
| ZNF763 | ENSG00000197054 | ENSP00000369774 | ENST00000343949 | 19 | 12035890 | 12090390 | 6 |
| ZNF433 | ENSG00000197647 | ENSP00000339767 | ENST00000344980 | 19 | 12125547 | 12146556 | 18 |
| ZNF878 | ENSG00000257446 | ENSP00000472036 | ENST00000602107 | 19 | 12154620 | 12163754 | 14 |
| ZNF844 | ENSG00000223547 | ENSP00000392024 | ENST00000439326 | 19 | 12175514 | 12192380 | 6 |
| ZNF20 | ENSG00000132010 | ENSP00000335437 | ENST00000334213 | 19 | 12203658 | 12251222 | 11 |
| ZNF625 | ENSG00000257591 | ENSP00000394380 | ENST00000439556 | 19 | 12251032 | 12267546 | 8 |
| ZNF136 | ENSG00000196646 | ENSP00000344162 | ENST00000343979 | 19 | 12273879 | 12300064 | 13 |
| ZNF44 | ENSG00000197857 | ENSP00000348419 | ENST00000356109 | 19 | 12358092 | 12405702 | 14 |
| ZNF563 | ENSG00000188868 | ENSP00000293725 | ENST00000293725 | 19 | 12428291 | 12444534 | 10 |
| ZNF442 | ENSG00000198342 | ENSP00000242804 | ENST00000242804 | 19 | 12460185 | 12476719 | 14 |
| ZNF799 | ENSG00000196466 | ENSP00000411084 | ENST00000430385 | 19 | 12490003 | 12512088 | 15 |
| ZNF443 | ENSG00000180855 | ENSP00000301547 | ENST00000301547 | 19 | 12540521 | 12551926 | 16 |
| ZNF709 | ENSG00000242852 | ENSP00000380840 | ENST00000397732 | 19 | 12571998 | 12624668 | 19 |
| ZNF564 | ENSG00000249709 | ENSP00000340004 | ENST00000339282 | 19 | 12636185 | 12662327 | 15 |
| ZNF490 | ENSG00000188033 | ENSP00000311521 | ENST00000311437 | 19 | 12688775 | 12750912 | 13 |
| ZNF791 | ENSG00000173875 | ENSP00000342974 | ENST00000343325 | 19 | 12721732 | 12742735 | 17 |
| ZNF333 | ENSG00000160961 | ENSP00000292530 | ENST00000292530 | 19 | 14800613 | 14844557 | 10 |
| ZNF101 | ENSG00000181896 | ENSP00000319716 | ENST00000318110 | 19 | 19779605 | 19791761 | 9 |
| ZNF14 | ENSG00000105708 | ENSP00000340514 | ENST00000344099 | 19 | 19821282 | 19843921 | 17 |
| ZNF506 | ENSG00000081665 | ENSP00000393835 | ENST00000443905 | 19 | 19902620 | 19932560 | 8 |
| ZNF253 | ENSG00000256771 | ENSP00000468720 | ENST00000589717 | 19 | 19976695 | 20005483 | 11 |
| ZNF93 | ENSG00000184635 | ENSP00000342002 | ENST00000343769 | 19 | 20011722 | 20046860 | 16 |
| ZNF682 | ENSG00000197124 | ENSP00000380351 | ENST00000397165 | 19 | 20115227 | 20150277 | 10 |
| ZNF90 | ENSG00000213988 | ENSP00000410466 | ENST00000418063 | 19 | 20188803 | 20237885 | 15 |
| ZNF486 | ENSG00000256229 | ENSP00000335042 | ENST00000335117 | 19 | 20278037 | 20311299 | 9 |
| ZNF737 | ENSG00000237440 | ENSP00000395733 | ENST00000427401 | 19 | 20720799 | 20748626 | 13 |
| ZNF626 | ENSG00000188171 | ENSP00000469958 | ENST00000601440 | 19 | 20802867 | 20844402 | 12 |
| ZNF85 | ENSG00000105750 | ENSP00000329793 | ENST00000328178 | 19 | 21106028 | 21133503 | 15 |
| ZNF430 | ENSG00000118620 | ENSP00000261560 | ENST00000261560 | 19 | 21203426 | 21242852 | 11 |
| ZNF714 | ENSG00000160352 | ENSP00000472368 | ENST00000596143 | 19 | 21264965 | 21308073 | 12 |
| ZNF431 | ENSG00000196705 | ENSP00000308578 | ENST00000311048 | 19 | 21324840 | 21368805 | 12 |
| ZNF708 | ENSG00000182141 | ENSP00000349401 | ENST00000356929 | 19 | 21473963 | 21512212 | 14 |
| ZNF493 | ENSG00000196268 | ENSP00000376110 | ENST00000392288 | 19 | 21579931 | 21610375 | 18 |
| ZNF429 | ENSG00000197013 | ENSP00000351280 | ENST00000358491 | 19 | 21688437 | 21721079 | 15 |
| ZNF100 | ENSG00000197020 | ENSP00000351042 | ENST00000358296 | 19 | 21905568 | 21950330 | 11 |
| ZNF43 | ENSG00000198521 | ENSP00000347045 | ENST00000354959 | 19 | 21990085 | 22034830 | 19 |
| ZNF208 | ENSG00000160321 | ENSP00000380315 | ENST00000397126 | 19 | 22148897 | 22193745 | 36 |
| ZNF257 | ENSG00000197134 | ENSP00000470209 | ENST00000594947 | 19 | 22235254 | 22274282 | 11 |
| ZNF676 | ENSG00000196109 | ENSP00000380310 | ENST00000397121 | 19 | 22361903 | 22379753 | 14 |
| ZNF729 | ENSG00000196350 | ENSP00000350085 | ENST00000357491 | 19 | 22469252 | 22499951 | 32 |
| ZNF98 | ENSG00000197360 | ENSP00000350418 | ENST00000357774 | 19 | 22573899 | 22605148 | 13 |
| ZNF492 | ENSG00000229676 | ENSP00000413660 | ENST00000456783 | 19 | 22817126 | 22850472 | 12 |
| ZNF99 | ENSG00000213973 | ENSP00000380293 | ENST00000397104 | 19 | 22939007 | 22952784 | 26 |
| ZNF730 | ENSG00000183850 | ENSP00000472959 | ENST00000597761 | 19 | 23258012 | 23330021 | 10 |

| ZNF724P | ENSG00000196081 | ENSP00000413411 | ENST00000418100 | 19 | 23404401 | 23433162 | 14 |
|---|---|---|---|---|---|---|---|
| ZNF91 | ENSG00000167232 | ENSP00000300619 | ENST00000300619 | 19 | 23540501 | 23578269 | 31 |
| ZNF675 | ENSG00000197372 | ENSP00000352836 | ENST00000359788 | 19 | 23835708 | 23870017 | 11 |
| ZNF681 | ENSG00000196172 | ENSP00000384000 | ENST00000402377 | 19 | 23921997 | 23941693 | 11 |
| ZNF726 | ENSG00000213967 | ENSP00000317125 | ENST00000322487 | 19 | 24097678 | 24127961 | 16 |
| ZNF254 | ENSG00000213096 | ENSP00000349494 | ENST00000357002 | 19 | 24216276 | 24312643 | 13 |
| ZNF302 | ENSG00000089335 | ENSP00000396379 | ENST00000446502 | 19 | 35168544 | 35177302 | 7 |
| ZNF181 | ENSG00000197841 | ENSP00000376065 | ENST00000392232 | 19 | 35225061 | 35233777 | 11 |
| ZNF599 | ENSG00000153896 | ENSP00000333802 | ENST00000329285 | 19 | 35248979 | 35270385 | 14 |
| ZNF30 | ENSG00000168661 | ENSP00000403441 | ENST00000439785 | 19 | 35417807 | 35436074 | 16 |
| ZNF792 | ENSG00000180884 | ENSP00000385099 | ENST00000404801 | 19 | 35447258 | 35454953 | 12 |
| ZNF565 | ENSG00000196357 | ENSP00000347234 | ENST00000355114 | 19 | 36673188 | 36737159 | 12 |
| ZFP14 | ENSG00000142065 | ENSP00000270001 | ENST00000270001 | 19 | 36827162 | 36870078 | 13 |
| ZFP82 | ENSG00000181007 | ENSP00000446080 | ENST00000392171 | 19 | 36874593 | 36909558 | 12 |
| ZNF566 | ENSG00000186017 | ENSP00000376010 | ENST00000392170 | 19 | 36936021 | 36980804 | 7 |
| ZNF529 | ENSG00000186020 | ENSP00000465578 | ENST00000591340 | 19 | 37025676 | 37096178 | 9 |
| ZNF382 | ENSG00000161298 | ENSP00000292928 | ENST00000292928 | 19 | 37095719 | 37119499 | 9 |
| ZNF461 | ENSG00000197808 | ENSP00000467931 | ENST00000588268 | 19 | 37128094 | 37157755 | 10 |
| ZNF567 | ENSG00000189042 | ENSP00000441838 | ENST00000536254 | 19 | 37178514 | 37218603 | 14 |
| ZNF790 | ENSG00000197863 | ENSP00000349161 | ENST00000356725 | 19 | 37309224 | 37341215 | 12 |
| ZNF829 | ENSG00000185869 | ENSP00000429266 | ENST00000391711 | 19 | 37379026 | 37407193 | 9 |
| ZNF568 | ENSG00000198453 | ENSP00000334685 | ENST00000333987 | 19 | 37407231 | 37488834 | 15 |
| ZNF420 | ENSG00000197050 | ENSP00000338770 | ENST00000337995 | 19 | 37569337 | 37621212 | 19 |
| ZNF585A | ENSG00000196967 | ENSP00000349440 | ENST00000356958 | 19 | 37597636 | 37663643 | 21 |
| ZNF585B | ENSG00000245680 | ENSP00000433773 | ENST00000532828 | 19 | 37675722 | 37709055 | 21 |
| ZNF383 | ENSG00000188283 | ENSP00000340132 | ENST00000352998 | 19 | 37717366 | 37734566 | 11 |
| HKR1 | ENSG00000181666 | ENSP00000315505 | ENST00000324411 | 19 | 37808813 | 37855355 | 13 |
| ZNF527 | ENSG00000189164 | ENSP00000390179 | ENST00000436120 | 19 | 37862059 | 37883968 | 11 |
| ZNF569 | ENSG00000196437 | ENSP00000325018 | ENST00000316950 | 19 | 37902062 | 37958339 | 18 |
| ZNF570 | ENSG00000171827 | ENSP00000331540 | ENST00000330173 | 19 | 37959982 | 37976260 | 11 |
| ZNF793 | ENSG00000188227 | ENSP00000396402 | ENST00000445217 | 19 | 37997841 | 38034237 | 6 |
| ZNF571 | ENSG00000180479 | ENSP00000333660 | ENST00000328550 | 19 | 38053552 | 38085673 | 16 |
| ZNF540 | ENSG00000171817 | ENSP00000324598 | ENST00000316433 | 19 | 38085731 | 38105000 | 17 |
| ZFP30 | ENSG00000120784 | ENSP00000343581 | ENST00000351218 | 19 | 38123389 | 38147162 | 12 |
| ZNF607 | ENSG00000198182 | ENSP00000347338 | ENST00000355202 | 19 | 38187264 | 38210691 | 18 |
| ZNF573 | ENSG00000189144 | ENSP00000465020 | ENST00000590414 | 19 | 38226734 | 38307940 | 19 |
| ZNF546 | ENSG00000187187 | ENSP00000339823 | ENST00000347077 | 19 | 40490041 | 40523514 | 22 |
| ZNF780B | ENSG00000128000 | ENSP00000391641 | ENST00000434248 | 19 | 40534167 | 40562116 | 21 |
| ZNF780A | ENSG00000197782 | ENSP00000400997 | ENST00000455521 | 19 | 40570428 | 40596845 | 17 |
| ZNF283 | ENSG00000167637 | ENSP00000327314 | ENST00000324461 | 19 | 44331444 | 44353307 | 15 |
| ZNF404 | ENSG00000176222 | ENSP00000319479 | ENST00000324394 | 19 | 44376519 | 44384291 | 14 |
| ZNF45 | ENSG00000124459 | ENSP00000269973 | ENST00000269973 | 19 | 44416776 | 44439411 | 15 |
| ZNF221 | ENSG00000159905 | ENSP00000251269 | ENST00000251269 | 19 | 44455380 | 44471752 | 15 |
| ZNF155 | ENSG00000204920 | ENSP00000385163 | ENST00000407951 | 19 | 44488346 | 44502477 | 11 |
| ZNF230 | ENSG00000159882 | ENSP00000409318 | ENST00000429154 | 19 | 44507077 | 44518072 | 7 |
| ZNF222 | ENSG00000159885 | ENSP00000375822 | ENST00000391960 | 19 | 44529494 | 44537260 | 8 |
| ZNF223 | ENSG00000178386 | ENSP00000401947 | ENST00000434772 | 19 | 44556164 | 44572142 | 8 |
| ZNF284 | ENSG00000186026 | ENSP00000411032 | ENST00000421176 | 19 | 44576297 | 44591623 | 11 |
| ZNF224 | ENSG00000186019 | ENSP00000337368 | ENST00000336976 | 19 | 44598503 | 44612919 | 18 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| ZNF225 | ENSG00000256294 | ENSP00000262894 | ENST00000262894 | 19 | 44617548 | 44637255 | 17 |
| ZNF226 | ENSG00000167380 | ENSP00000400878 | ENST00000426739 | 19 | 44645710 | 44681836 | 18 |
| ZNF227 | ENSG00000131115 | ENSP00000321049 | ENST00000313040 | 19 | 44716691 | 44741420 | 18 |
| ZNF235 | ENSG00000159917 | ENSP00000291182 | ENST00000291182 | 19 | 44732882 | 44809199 | 15 |
| ZNF233 | ENSG00000159915 | ENSP00000375820 | ENST00000391958 | 19 | 44754318 | 44779470 | 7 |
| ZNF112 | ENSG00000062370 | ENSP00000346305 | ENST00000354340 | 19 | 44830708 | 44871377 | 13 |
| ZNF285 | ENSG00000267508 | ENSP00000333595 | ENST00000330997 | 19 | 44886459 | 44905774 | |
| ZNF229 | ENSG00000167383 | ENSP00000291187 | ENST00000291187 | 19 | 44930426 | 44952665 | 16 |
| ZNF180 | ENSG00000167384 | ENSP00000221327 | ENST00000221327 | 19 | 44979861 | 45004574 | 12 |
| ZNF114 | ENSG00000178150 | ENSP00000318898 | ENST00000315849 | 19 | 48774654 | 48790863 | 3 |
| ZNF473 | ENSG00000142528 | ENSP00000270617 | ENST00000270617 | 19 | 50529212 | 50552029 | 18 |
| ZNF175 | ENSG00000105497 | ENSP00000262259 | ENST00000262259 | 19 | 52074551 | 52092991 | 13 |
| ZNF577 | ENSG00000161551 | ENSP00000301399 | ENST00000301399 | 19 | 52359055 | 52394203 | 7 |
| ZNF649 | ENSG00000198093 | ENSP00000347043 | ENST00000354957 | 19 | 52392477 | 52408293 | 10 |
| ZNF613 | ENSG00000176024 | ENSP00000293471 | ENST00000293471 | 19 | 52430400 | 52452012 | 12 |
| ZNF350 | ENSG00000256683 | ENSP00000243644 | ENST00000243644 | 19 | 52467594 | 52490079 | 8 |
| ZNF615 | ENSG00000197619 | ENSP00000473089 | ENST00000602063 | 19 | 52494585 | 52511483 | 19 |
| ZNF614 | ENSG00000142556 | ENSP00000270649 | ENST00000270649 | 19 | 52516021 | 52531680 | 10 |
| ZNF432 | ENSG00000256087 | ENSP00000221315 | ENST00000221315 | 19 | 52536361 | 52552106 | 16 |
| ZNF841 | ENSG00000197608 | ENSP00000374185 | ENST00000389534 | 19 | 52567719 | 52599018 | 20 |
| ZNF616 | ENSG00000204611 | ENSP00000471000 | ENST00000600228 | 19 | 52616344 | 52643170 | 21 |
| ZNF836 | ENSG00000196267 | ENSP00000325038 | ENST00000322146 | 19 | 52658125 | 52674896 | 24 |
| ZNF766 | ENSG00000196214 | ENSP00000409652 | ENST00000439461 | 19 | 52772824 | 52795977 | 8 |
| ZNF480 | ENSG00000198464 | ENSP00000471754 | ENST00000595962 | 19 | 52800430 | 52829175 | 10 |
| ZNF610 | ENSG00000167554 | ENSP00000327597 | ENST00000327920 | 19 | 52839498 | 52870375 | 8 |
| ZNF880 | ENSG00000221923 | ENSP00000406318 | ENST00000422689 | 19 | 52873170 | 52889048 | 13 |
| ZNF528 | ENSG00000167555 | ENSP00000353652 | ENST00000360465 | 19 | 52901102 | 52921657 | 15 |
| ZNF534 | ENSG00000198633 | ENSP00000327538 | ENST00000332323 | 19 | 52932440 | 52955568 | 14 |
| ZNF578 | ENSG00000258405 | ENSP00000459216 | ENST00000421239 | 19 | 52956829 | 53020131 | 10 |
| ZNF808 | ENSG00000198482 | ENSP00000352846 | ENST00000359798 | 19 | 53030905 | 53067717 | 23 |
| ZNF701 | ENSG00000167562 | ENSP00000444339 | ENST00000540331 | 19 | 53059075 | 53090427 | 7 |
| ZNF611 | ENSG00000213020 | ENSP00000322427 | ENST00000319783 | 19 | 53206066 | 53238307 | 13 |
| ZNF28 | ENSG00000198538 | ENSP00000397693 | ENST00000457749 | 19 | 53300662 | 53360853 | 15 |
| ZNF468 | ENSG00000204604 | ENSP0000047038 | ENST00000595646 | 19 | 53341261 | 53360902 | 10 |
| ZNF320 | ENSG00000182986 | ENSP00000375660 | ENST00000391781 | 19 | 53379425 | 53393592 | 11 |
| ZNF816 | ENSG00000180257 | ENSP00000350295 | ENST00000357666 | 19 | 53430388 | 53466164 | 15 |
| ZNF160 | ENSG00000170949 | ENSP00000409597 | ENST00000418871 | 19 | 53569867 | 53606687 | 20 |
| ZNF415 | ENSG00000170954 | ENSP00000388787 | ENST00000455735 | 19 | 53611132 | 53636330 | 11 |
| ZNF347 | ENSG00000197937 | ENSP00000405218 | ENST00000452676 | 19 | 53641958 | 53662322 | 17 |
| ZNF665 | ENSG00000197497 | ENSP00000379702 | ENST00000396424 | 19 | 53666552 | 53696619 | 18 |
| ZNF677 | ENSG00000197928 | ENSP00000334394 | ENST00000333952 | 19 | 53727087 | 53758126 | 10 |
| ZNF845 | ENSG00000213799 | ENSP00000388311 | ENST00000458035 | 19 | 53837002 | 53858122 | 26 |
| ZNF525 | ENSG00000203326 | ENSP00000417696 | ENST00000474037 | 19 | 53868946 | 53889846 | 8 |
| ZNF765 | ENSG00000196417 | ENSP00000379689 | ENST00000396408 | 19 | 53893046 | 53930574 | 8 |
| ZNF813 | ENSG00000198346 | ENSP00000379684 | ENST00000396403 | 19 | 53970989 | 54006950 | 13 |
| ZNF331 | ENSG00000130844 | ENSP00000253144 | ENST00000253144 | 19 | 54024235 | 54083523 | 12 |
| ZNF582 | ENSG00000018869 | ENSP00000301310 | ENST00000301310 | 19 | 56894648 | 56904889 | 9 |
| ZNF583 | ENSG00000198440 | ENSP00000291598 | ENST00000291598 | 19 | 56915383 | 56938733 | 12 |
| ZNF667 | ENSG00000198046 | ENSP00000344699 | ENST00000342634 | 19 | 56950696 | 56988770 | 14 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| ZNF471 | ENSG00000196263 | ENSP00000309161 | ENST00000308031 | 19 | 57019212 | 57040270 | 15 |
| ZFP28 | ENSG00000196867 | ENSP00000301318 | ENST00000301318 | 19 | 57050317 | 57068169 | 14 |
| ZNF470 | ENSG00000197016 | ENSP00000333223 | ENST00000330619 | 19 | 57078890 | 57094261 | 17 |
| ZIM2.1 | ENSG00000259486 | ENSP00000221722 | ENST00000221722 | 19 | 57285920 | 57352097 | 5 |
| ZIM3 | ENSG00000141946 | ENSP00000269834 | ENST00000269834 | 19 | 57645464 | 57656570 | 11 |
| ZNF264 | ENSG00000083844 | ENSP00000263095 | ENST00000263095 | 19 | 57702868 | 57734212 | 13 |
| ZNF805 | ENSG00000204524 | ENSP00000412999 | ENST00000414468 | 19 | 57751973 | 57766503 | 13 |
| ZNF460 | ENSG00000197714 | ENSP00000353491 | ENST00000360338 | 19 | 57791419 | 57805436 | 11 |
| ZNF543 | ENSG00000178229 | ENSP00000322545 | ENST00000321545 | 19 | 57831865 | 57842144 | 13 |
| ZNF304 | ENSG00000131845 | ENSP00000401642 | ENST00000443917 | 19 | 57862645 | 57871266 | 16 |
| ZNF547 | ENSG00000152433 | ENSP00000282282 | ENST00000282282 | 19 | 57874891 | 57890923 | 10 |
| ZNF548 | ENSG00000188785 | ENSP00000337555 | ENST00000336128 | 19 | 57901218 | 57913917 | 11 |
| ZNF17 | ENSG00000186272 | ENSP00000302455 | ENST00000307658 | 19 | 57922529 | 57933307 | 18 |
| ZNF749 | ENSG00000186230 | ENSP00000333980 | ENST00000334181 | 19 | 57946697 | 57956853 | 13 |
| ZNF772 | ENSG00000197128 | ENSP00000341165 | ENST00000343280 | 19 | 57978031 | 57988938 | 10 |
| ZNF419 | ENSG00000105136 | ENSP00000388864 | ENST00000424930 | 19 | 57999079 | 58006048 | 11 |
| ZNF773 | ENSG00000152439 | ENSP00000282292 | ENST00000282292 | 19 | 58011309 | 58024436 | 9 |
| ZNF549 | ENSG00000121406 | ENSP00000365407 | ENST00000376233 | 19 | 58038693 | 58068910 | 13 |
| ZNF550 | ENSG00000251369 | ENSP00000446224 | ENST00000325134 | 19 | 58046625 | 58071231 | 8 |
| ZNF416 | ENSG00000083817 | ENSP00000196489 | ENST00000196489 | 19 | 58082935 | 58090243 | 11 |
| ZIK1 | ENSG00000171649 | ENSP00000472867 | ENST00000597850 | 19 | 58095510 | 58105145 | 9 |
| ZNF530 | ENSG00000183647 | ENSP00000332861 | ENST00000332854 | 19 | 58111253 | 58119637 | 13 |
| ZNF211 | ENSG00000121417 | ENSP00000299871 | ENST00000299871 | 19 | 58141761 | 58154147 | 11 |
| ZNF551 | ENSG00000204519 | ENSP00000282296 | ENST00000282296 | 19 | 58193357 | 58202022 | 14 |
| ZNF154 | ENSG00000179909 | ENSP00000442370 | ENST00000426889 | 19 | 58208735 | 58220579 | 10 |
| ZNF671 | ENSG00000083814 | ENSP00000321848 | ENST00000317398 | 19 | 58231120 | 58238995 | 9 |
| ZNF776 | ENSG00000152443 | ENSP00000321812 | ENST00000317178 | 19 | 58258164 | 58269527 | 9 |
| ZNF586 | ENSG00000083828 | ENSP00000379458 | ENST00000396154 | 19 | 58281023 | 58331307 | 8 |
| ZNF552 | ENSG00000178935 | ENSP00000375582 | ENST00000391701 | 19 | 58315209 | 58326281 | 6 |
| ZNF587 | ENSG00000198466 | ENSP00000345479 | ENST00000339656 | 19 | 58331094 | 58376485 | 13 |
| ZNF814 | ENSG00000204514 | ENSP00000410545 | ENST00000435989 | 19 | 58380747 | 58400442 | 23 |
| ZNF417 | ENSG00000173480 | ENSP00000311319 | ENST00000312026 | 19 | 58417142 | 58427978 | 12 |
| ZNF418 | ENSG00000196724 | ENSP00000407039 | ENST00000425570 | 19 | 58433252 | 58446755 | 16 |
| ZNF256 | ENSG00000152454 | ENSP00000282308 | ENST00000282308 | 19 | 58452206 | 58459077 | 15 |
| ZNF606 | ENSG00000166704 | ENSP00000343617 | ENST00000341164 | 19 | 58488421 | 58514717 | 13 |
| ZNF135 | ENSG00000176293 | ENSP00000441410 | ENST00000401053 | 19 | 58570607 | 58597677 | 16 |
| ZNF274 | ENSG00000171606 | ENSP00000321209 | ENST00000326804 | 19 | 58694396 | 58724927 | 5 |
| ZNF544 | ENSG00000198131 | ENSP00000269829 | ENST00000269829 | 19 | 58740070 | 58775010 | 12 |
| ZNF8 | ENSG00000083842 | ENSP00000196548 | ENST00000196548 | 19 | 58790318 | 58807254 | 7 |
| ZNF584 | ENSG00000171574 | ENSP00000306756 | ENST00000306910 | 19 | 58912871 | 58929694 | 8 |
| ZNF132 | ENSG00000131849 | ENSP00000254166 | ENST00000254166 | 19 | 58944181 | 58951589 | 17 |
| ZNF324B | ENSG00000249471 | ENSP00000337473 | ENST00000336614 | 19 | 58962971 | 58969199 | 9 |
| ZNF324 | ENSG00000083812 | ENSP00000196482 | ENST00000196482 | 19 | 58978459 | 58984781 | 9 |
| ZNF446 | ENSG00000083838 | ENSP00000472802 | ENST00000594369 | 19 | 58985384 | 58992597 | 3 |
| ZNF343 | ENSG00000088876 | ENSP00000278772 | ENST00000278772 | 20 | 2462463 | 2505348 | 12 |
| ZNF133 | ENSG00000125846 | ENSP00000400897 | ENST00000396026 | 20 | 18269121 | 18297640 | 14 |
| ZNF337 | ENSG00000130684 | ENSP00000252979 | ENST00000252979 | 20 | 25654851 | 25677477 | 19 |
| ZNF334 | ENSG00000198185 | ENSP00000255129 | ENST00000347606 | 20 | 45129709 | 45142198 | 14 |
| ZNF74 | ENSG00000185252 | ENSP00000349098 | ENST00000356671 | 22 | 20748405 | 20762745 | 12 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| ZNF674 | ENSG00000251192 | ENSP00000429148 | ENST00000523374 | X | 46357162 | 46404892 | 11 |
| ZNF157 | ENSG00000147117 | ENSP00000366273 | ENST00000377073 | X | 47229982 | 47273704 | 12 |
| ZNF41 | ENSG00000147124 | ENSP00000380243 | ENST00000397050 | X | 47305278 | 47342345 | 17 |
| ZNF81 | ENSG00000197779 | ENSP00000366153 | ENST00000376954 | X | 47696301 | 47861960 | 12 |
| ZNF182 | ENSG00000147118 | ENSP00000380165 | ENST00000396965 | X | 47834250 | 47863377 | 14 |
| ZNF630 | ENSG00000221994 | ENSP00000393163 | ENST00000442455 | X | 47842756 | 47931025 | 11 |
| ZNF75D | ENSG00000186376 | ENSP00000359802 | ENST00000370766 | X | 134382867 | 134478012 | 5 |
| ZNF275 | ENSG00000063587 | ENSP00000411097 | ENST00000440091 | X | 152599613 | 152618384 | 11 |

**Suppl. Table 1: Manually curated list of KRAB-ZF genes.**

| Read length | Spearman's rho | p-value |
|---|---|---|
| 36bp | -0.06261553 | 0.1955 |
| 50bp | -0.08199701 | 0.08984 |
| 75bp | -0.05795235 | 0.231 |
| 100bp | -0.0655479 | 0.1754 |

**Suppl. Table 2: Correlation coefficiants between MAF and mappability.**

## Zinc-Finger domains (all genes pooled together)

**247 manually annotated orthologous genes:**

| | Non-synonymous | Synonymous |
|---|---|---|
| Divergence with Chimpanzee | 302 | 460 |
| Polymorphism in Humans | 1430 | 794 |

chi-square test p-value < 2.2e-16***

**108 manually annotated orthologous genes:**

| | Non-synonymous | Synonymous |
|---|---|---|
| Divergence with Macaque | 406 | 890 |
| Polymorphism in Humans | 542 | 344 |

chi-square test p-value < 2.2e-16***

**61 manually annotated orthologous genes:**

| | Non-synonymous | Synonymous |
|---|---|---|
| Divergence with Mouse | 552 | 1556 |
| Polymorphism in Humans | 246 | 189 |

chi-square test p-value < 2.2e-16***

## DNA-contacting amino acids only (all genes pooled together)

| | Non-synonymous | Synonymous |
|---|---|---|
| Divergence with Chimpanzee | 42 | 86 |
| Polymorphism in Humans | 177 | 105 |

Fisher's exact test two-tailed p-value = 2.069e-08***

| | Non-synonymous | Synonymous |
|---|---|---|
| Divergence with Macaque | 82 | 202 |
| Polymorphism in Humans | 46 | 47 |

Fisher's exact test two-tailed p-value = 0.0003901***

| | Non-synonymous | Synonymous |
|---|---|---|

| | | |
|---|---|---|
| Divergence with Mouse | 68 | 301 |
| Polymorphism in Humans | 21 | 17 |

Fisher's exact test two-tailed p-value = 2.115e-06***

## KRAB domains (all genes pooled together)

**90 manually annotated orthologous genes:**

| | Non-synonymous | Synonymous |
|---|---|---|
| Divergence with Chimpanzee | 78 | 46 |
| Polymorphism in Humans | 156 | 84 |

Fisher's exact test two-tailed p-value = 0.7297

**29 manually annotated orthologous genes:**

| | Non-synonymous | Synonymous |
|---|---|---|
| Divergence with Macaque | 84 | 75 |
| Polymorphism in Humans | 54 | 28 |

Fisher's exact test two-tailed p-value = 0.05587

**24 manually annotated orthologous genes:**

| | Non-synonymous | Synonymous |
|---|---|---|
| Divergence with Mouse | 264 | 254 |
| Polymorphism in Humans | 51 | 21 |

Fisher's exact test two-tailed p-value = 0.001561**

**Suppl. Table 3: Details about all MK-tests.**

# CONCLUSION

## Demographic History of *Drosophila melanogaster*

In this study, we used whole genome sequences from wild strains of *D. melanogaster* from Zambia, West Africa (Lack et al. 2016) and Sweden. Our results confirmed and extended previous reports of significant structure present in sub-Saharan Africa between West and South/Central populations (Veuille et al. 2004; Pool et al. 2012). We estimated the division time between those two populations at approximately 72k years ago [66.5k-79.5k]. We demonstrated their consequent increase in population size, as well as the importance of migration in shaping the variation in their genomes. In agreement with our estimations, previous studies (Lachaise and Silvain 2004; Li and Stephan 2006; Stephan and Li 2007; Laurent et al. 2011) reported an African expansion at the same time, corresponding to the transition period from full glacial to interglacial and the wild-to-domestic habit shift. *D. melanogaster* became a human commensal and it has been reported that humans were present in West Africa at the same time (Nielsen et al. 2017).

Similarly, for the European sample, we estimated the out-of-Africa exodus at approximately 43.5k years ago for the autosomes and 26k years ago for the X chromosome. In comparison with previous results (David and Capy 1988; Baudry et al. 2004; Haddrill et al. 2005; Li and Stephan 2006; Laurent et al. 2011), our results refine the estimated exit out-of-Africa to much earlier dates. These refinements were possible through the use of more sophisticated models taking into account the existence of constant gene flow. Previous studies underestimated the exit time by not accounting for migration between the two populations. Therefore, the divergence between populations appears to be more recent, because populations differentiate faster than in presence of constant gene flow which reduces differentiation. The bigger the migration rate, the more populations will tend to look similar in terms of allele frequencies. Additional comparison with the human out-of-Africa estimates (55-65kya, Nielsen et al. 2017) lends additional support to our conclusions and model assumptions.

We believe that our results are crucial for further studies aiming to identify alleles responsible for local adaptation. The Swedish flies demographic history is important in genomic scans aiming to identify regions responsible for adaptation to the Northern European cold climate. The colonization of the western part of Africa might also have undergone adaptive processes. Our model estimates that after the split, a size reduction occurred but not a severe bottleneck. This better explains the genomic variation present in the *D. melanogaster* genome and represents an improved null model for future genomic scans to detect selection. The West African demography is also necessary when studying American colonization patterns. Specifically, the West African population gave rise to the American strains, admixed with the European strains (Caracristi and Schlötterer 2003; Kao *et al.* 2015). Therefore our results from West African demography have profound repercussion on the study of evolution and demographic models for other *D. melanogaster* populations present in distant continents.

## Evolution of KRAB-containing Zinc Finger family

In this chapter, we characterized the evolution of gene expression together with the binding specificity of the largest family of transcription factors in humans, namely the KRAB-ZFs. Such study required the development of a carefully annotated and reliable dataset of the KRAB-ZF genes. Due to their repetitive sequence, automated predictions fail to correctly identify them. We have manually inspected and annotated all KRAB-ZF genes present in the human genome, which enabled us to refine the automated predictions and to perform our subsequent analyses with high-quality data.

Specifically, I focused on gene expression analyses, supplemented with large-scale epigenomic data, for all KRAB-ZFs identified in primates. From their gene expression patterns, we found support for their rapid evolution, suggesting their important role in primate evolution and subsequent KRAB-ZF lineage expansion. Additionally, we analyzed their binding specificities and were able to characterize KRAB-ZF genes into two distinct groups according to the presence or absence of nonsynonymous polymorphisms located within at least one of the four DNA-contacting

amino acids. Subsequent analyses showed that those two groups had different age and expression patterns across the tissues.

Globally, this study was able to link gene expression patterns, regulatory gene expression networks, evolutionary history and DNA-binding polymorphism. Our approaches can serve other evolutionary studies focusing on gene expression data (RNA-seq or microarray), possibly with access to epigenetic data (*e.g.*, chip-seq). Both our results on KRAB-ZNF and our manually annotated dataset can constitute a valuable resource to other scientists studying the KRAB-ZNF family or more broadly to scientists interested in the evolution of gene expression regulation. Demonstrating the utility of this work, two independent studies have used our manually curated dataset of KRAB-ZF genes (Ward et al. 2017) and our conclusions (Ecco et al. 2017) to guide and support their investigations.

## References

Baudry, E., B. Viginier, and M. Veuille. 2004. Non-African Populations of Drosophila melanogaster Have a Unique Origin. Mol. Biol. Evol. 21:1482–1491.

Caracristi, G., and C. Schlötterer. 2003. Genetic Differentiation Between American and European Drosophila melanogaster Populations Could Be Attributed to Admixture of African Alleles. Mol. Biol. Evol. 20:792–799.

David, J. R., and P. Capy. 1988. Genetic variation of Drosophila melanogaster natural populations. Trends Genet. 4:106–111.

Ecco, G., M. Imbeault, and D. Trono. 2017. KRAB zinc finger proteins. Development 144:2719–2729.

Haddrill, P. R., K. R. Thornton, B. Charlesworth, and P. Andolfatto. 2005. Multilocus patterns of nucleotide variability and the demographic and selection history of Drosophila melanogaster populations. Genome Res. 15:790–799.

Kao Joyce Y., Zubair Asif, Salomon Matthew P., Nuzhdin Sergey V., and Campo Daniel. 2015. Population genomic analysis uncovers African and European admixture in

Drosophila melanogaster populations from the south‐eastern United States and Caribbean Islands. Mol. Ecol. 24:1499–1509.

Lachaise, D., and J.-F. Silvain. 2004. How two Afrotropical endemics made two cosmopolitan human commensals: the Drosophila melanogaster-D. simulans palaeogeographic riddle. Pp. 17–39 *in* Drosophila melanogaster, Drosophila simulans: So Similar, So Different. Springer.

Lack, J. B., J. D. Lange, A. D. Tang, R. B. Corbett-Detig, and J. E. Pool. 2016. A Thousand Fly Genomes: An Expanded Drosophila Genome Nexus. Mol. Biol. Evol. msw195.

Laurent, S. J. Y., A. Werzner, L. Excoffier, and W. Stephan. 2011. Approximate Bayesian Analysis of Drosophila melanogaster Polymorphism Data Reveals a Recent Colonization of Southeast Asia. Mol. Biol. Evol. 28:2041–2051.

Li, H., and W. Stephan. 2006. Inferring the Demographic History and Rate of Adaptive Substitution in Drosophila. PLOS Genet. 2:e166.

Nielsen, R., J. M. Akey, M. Jakobsson, J. K. Pritchard, S. Tishkoff, and E. Willerslev. 2017. Tracing the peopling of the world through genomics. Nature 541:302–310.

Pool, J. E., R. B. Corbett-Detig, R. P. Sugino, K. A. Stevens, C. M. Cardeno, M. W. Crepeau, P. Duchen, J. J. Emerson, P. Saelao, D. J. Begun, and C. H. Langley. 2012. Population Genomics of Sub-Saharan Drosophila melanogaster: African Diversity and Non-African Admixture. PLOS Genet. 8:e1003080.

Stephan, W., and H. Li. 2007. The recent demographic and adaptive history of *Drosophila melanogaster*. Heredity 98:65–68.

Veuille, M., E. Baudry, M. Cobb, N. Derome, and E. Gravot. 2004. Historicity and the population genetics of *Drosophila melanogaster* and *D. simulans*. Pp. 61–70 *in* Drosophila melanogaster, Drosophila simulans: So Similar, So Different. Springer, Dordrecht.

Ward, M. C., S. Zhao, K. Luo, B. J. Pavlovic, M. M. Karimi, M. Stephens, and Y. Gilad. 2017. Silencing Of Transposable Elements May Not Be A Major Driver Of Regulatory Evolution In Primate Induced Pluripotent Stem Cells. bioRxiv 142455.

# CURRICULUM VITAE


## Adamandia KAPOPOULOU

Bioinformatician / Population Geneticist


Ten years of experience in bioinformatics, with a broad expertise ranging from protein 3D to genomics.

Over the past 7 years, strong focus on the analysis of next-generation sequencing data (Gene expression using RNA-Seq, NanoString, and Microarrays, Chromatin modifications using ChIP-Seq, Transcription Factor binding using ChIP-Seq, Methylation profiles using Bisulfite Sequencing).


Greek, Swiss C permit

076 230 4721

adamandia.kapopoulou@epfl.ch

www.linkedin.com/in/kapopoulou

Avenue de la Gare 73

1022 – Chavannes-Près-Renens


### WORK EXPERIENCE

**SWISS FEDERAL SCHOOL OF TECHNOLOGY (EPFL)**

**Bioinformatician / Population Geneticist**

*2014 – PRESENT*

Computational and Statistical Analysis of research projects in the field of Population Genetics and Evolutionary Biology

**Main Tasks:**

- NGS Data Analysis (Genomics, Genetics, Epigenomics)
- Development and Optimization of NGS pipelines
- Data Integration from multiple data sources (Private and Public)
- Statistical Analysis using R
- Develop innovative approaches to process and analyse data
- Statistical Population Genetics Teaching to B.Sc. and Master's level at EPFL (300 hours)

**Relevant Skills:**

- Programming (Perl/Shell scripting)
- Ability to Interpret Results and Present to experts/non-experts

**SWISS FEDERAL SCHOOL OF TECHNOLOGY (EPFL)**

Lead Scientist in NGS Data Analysis, Data Management, and Visualization

**Embedded Bioinformatician**

*2010 - 2014*

**Main Tasks:**

- NGS Data Analysis (Genomics, Genetics, Epigenomics)
- Development and Implementation of NGS pipelines
- Data Integration from multiple data sources (Private and Public)
- Statistical Analysis using R
- Develop innovative approaches to process and analyse data
- Two student supervision (6 months each)

**Relevant Skills:**

- Programming (Perl/Shell scripting)
- Ability to Interpret Results
- Effective communication to non-specialists

**SWISS FEDERAL SCHOOL OF TECHNOLOGY (EPFL)**

Project Manager of Tuberculosis / Leprosy Database and Website

**Project Manager / Bioinformatics Scientist**

*2007 – 2010*

**Main Tasks:**

- Development of a Web Portal (PHP/Javascript) for Mycobacterial Genomes (http://tuberculist.epfl.ch, https://mycobrowser.epfl.ch/)
- Database Management (PostgreSQL)
- Project Management (establish international collaborations with Stanford, Broad Institute, and Institut Pasteur Paris)
- Implementation of an annotation pipeline (Perl)

**Relevant Skills:**

- Programming (Perl/Shell scripting)
- Web, Database programming

**EUROPEAN BIOINFORMATICS INSTITUTE (EBI) – CAMBRIDGE, UK**

Validation and Curation of 3D protein structures submitted to Protein Data Bank (PDB)

**Biologist / Scientific Database Curator**

*2004 – 2007*

**Main Tasks:**

- Optimization of PDB annotation protocols
- PDB weekly releases and correspondences
- Representation of PDB in international conferences
- Implementation of annotation pipeline (Perl)
- Develop innovative approaches to process and analyse data
- Two student supervision (6 months each)

**Relevant Skills:**

- Programming (Perl/Shell scripting)
- Data Scientist (Protein Structures)
- Database specifications, Annotation specifications

## EDUCATION

**SWISS FEDERAL SCHOOL OF TECHNOLOGY (EPFL)**
**PhD**
*2014 – EXPECTED SUMMER 2018*

PhD Population Genetics and Evolution

**UNIVERSITÉ BORDEAUX I, FRANCE**
**Master's Degree**
*2003 – 2004*

Master's Degree in Bioinformatics

**UNIVERSITÉ BORDEAUX II**
**Master's Degree**
*2001 – 2003*

Master's Degree in Genetics

**UNIVERSITÉ BORDEAUX II, FRANCE**
**Bachelor's Degree**
*1999 – 2001*

Bachelor's Degree in Physiology and Cell Biology

**UNIVERSITÉ BORDEAUX II**
**First Year Medical Studies**

First Year Medical Studies

*1997 – 1999*

## LANGUAGES

- **ENGLISH**: Fluent
- **FRENCH**: Fluent
- **GREEK**: Native speaker

## SKILLS SUMMARY

- **Programming** (Perl, Shell, AWK)
- **Statistical Computing** (R)
- **Databases** (SQL)
- **Web Programming** (PHP, Javascript)
- **Bioinformatics**
- **Genetics**
- **Genomics** (Next Generation Sequencing)
- **Protein 3D Structure** (Quality Control, Rasmol, PDB format)

## EXTRA-CURRICULAR ACTIVITIES

- **PHD REPRESENTATIVE** at EPFL Doctoral Commission
- Active member of Bioscience Network Lausanne **(BSNL)**

## PUBLICATIONS

- ***20 PUBLICATIONS IN PEER-REVIEWED JOURNALS***
- ***H-INDEX: 18***

**Population Genetics (PhD)**

**The evolution of gene expression and binding specificity of the largest transcription factor family in primates**

A Kapopoulou, L Mathew, A Wong, D Trono, JD Jensen

*Evolution, 2016*

**Genomics (Lead Bioinformatician)**

**Release of human cytomegalovirus from latency by a KAP1/TRIM28 phosphorylation switch**

B Rauwel, SM Jang, M Cassano, A Kapopoulou, I Barde, D Trono

*eLife, 2015*

**TRIM28 Represses Transcription of Endogenous Retroviruses in Neural Progenitor Cells**

L Fasching, A Kapopoulou, et al.

*Cell reports, 2015*

**Loss of transcriptional control over endogenous retroelements during reprogramming to pluripotency**

M Friedli, P Turelli, A Kapopoulou, et al.

*Genome research, 2014*

**Interplay of TRIM28 and DNA methylation in controlling human endogenous retroelements**

P Turelli, N Castro-Diaz, F Marzetta, A Kapopoulou, et al.

*Genome research, 2014*

**Evolutionally dynamic L1 regulation in embryonic stem cells**

N Castro-Diaz, G Ecco, A Coluccio, A Kapopoulou, et al.

*Genes & development, 2014*

**Contrôle de la mitophagie par les microARN-Une étape clé de l'érythropoïèse**

I Barde, B Rauwel, RM Marin-Florez, A Corsinotti, E Laurenti, S Verp, Sandra Offner, Julien Marquis, A Kapopoulou, et al.

*médecine/sciences, 2014*

**A KRAB/KAP1-miRNA cascade regulates erythropoiesis through stage-specific control of mitophagy**

I Barde, B Rauwel, RM Marin-Florez, A Corsinotti, E Laurenti, S Verp, Sandra Offner, Julien Marquis, A Kapopoulou, et al.

*Science, 2013*

**TRIM28 repression of retrotransposon-based enhancers is necessary to preserve transcriptional dynamics in embryonic stem cells**

HM Rowe, A Kapopoulou, et al.

*Genome research, 2013*

**Global and stage specific patterns of Krüppel-associated-box zinc finger protein gene expression in murine early embryonic cells**

A Corsinotti, A Kapopoulou, et al.

*PLoS One, 2013*

**KAP1 regulates gene networks controlling T-cell development and responsiveness**

FRS de Sio, I Barde, S Offner, A Kapopoulou, et al.

*The FASEB Journal, 2012*

**The KRAB-ZFP/KAP1 system contributes to the early embryonic establishment of site-specific DNA methylation patterns maintained during development**

S Quenneville, P Turelli, K Bojkowska, C Raclot, S Offner, A Kapopoulou, Didier Trono

*Cell reports, 2012*

**Liver-specific ablation of Krüppel-associated box–associated protein 1 in mice leads to male-predominant hepatosteatosis and development of liver adenoma**

K Bojkowska, F Aloisio, M Cassano, A Kapopoulou, et al.

*Hepatology, 2012*

**KAP1 regulates gene networks controlling mouse B-lymphoid cell differentiation and function**

FRS de Sio, J Massacand, I Barde, S Offner, A Corsinotti, A Kapopoulou, et al.

*Blood, 2012*

**In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions**

S Quenneville, G Verde, A Corsinotti, A Kapopoulou, et al.

*Molecular cell, 2011*

**A gene-rich, transcriptionally active environment and the pre-deposition of repressive marks are predictive of susceptibility to KRAB/KAP1-mediated silencing**

S Meylan, AC Groner, G Ambrosini, N Malani, S Quenneville, N Zangger, A Kapopoulou, et al.

*BMC genomics, 2011*

**Tuberculosis and Leprosy (Lead Bioinformatician)**

**Probable zoonotic leprosy in the southern United States**

RW Truman, P Singh, R Sharma, P Busso, J Rougemont, Alberto Paniz-Mondolfi, A Kapopoulou, et al.

*New England Journal of Medicine, 2011*

**The MycoBrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes**

A Kapopoulou, JM Lew, ST Cole

*Tuberculosis, 2011*

**TubercuList–10 years after**

JM Lew, A Kapopoulou, LM Jones, ST Cole

*Tuberculosis, 2011*

**Protein 3D Structures (Database Curator)**

**E-MSD: improving data deposition and structure quality**

M Tagari, J Tate, GJ Swaminathan, R Newman, A Naim, W Vranken, A Kapopoulou, et al.

*Nucleic acids research, 2006*