Thèse n° 10 132

HIV host genomics in the era of effective antiretroviral therapy

Présentée le 20 mars 2020

à la Faculté des sciences de la vie
Groupe Fellay
Programme doctoral en approches moléculaires du vivant

pour l’obtention du grade de Docteur ès Sciences

par

Christian Axel Wandall THORBALL

Acceptée sur proposition du jury

Prof. A. Ablasser, présidente du jury
Prof. J. Fellay, Prof. D. Trono, directeurs de thèse
Prof. A. Telenti, rapporteur
Prof. O. Michielin, rapporteur
Prof. B. Deplancke, rapporteur

École
polytechnique
fédérale
de Lausanne

2020

Acknowledgements

The path to the completion of this thesis has not been ordinary but has brought along many great things that I could never have imagined when I first started. For that, I have a lot of people to be grateful for. First and foremost, I owe big gratitude to Jacques Fellay for not only taking me in, in the hour of need, despite my complete lack of computational skills at the time. But also, for encouraging growth and fostering independence in the daily work, while always being there when things were difficult or when an epic run was needed. It has made the years very enjoyable.

Secondly, I would like to thank my co-supervisor Didier Trono, who also took me into his lab and is responsible for greatly expanding my scientific horizons and thinking while also putting me together with many excellent and fun people. I learned a lot there.

I would like to thank the members of the Fellay Lab; Thomas, Dylan, Konstantin, Olivier, Flavia, Mack, Chris, Alessandro, Nimisha, Sina, Petar, Paul, Istvan, and Samira, for all the good discussions and the excellent atmosphere we have had in the office. One could not have asked for better colleagues. I want to extend a special thanks to Paul McLaren for taking me under his wings in the first year and teaching me all the basics of computational biology.

I would also like to thank the members of the Trono Lab for all the fun and interesting times we have had together. In particular, I would like to thank Pila, Pierre-Yves, Jonas, and Coluccio for all the nice discussions and collaborations. Julien Duc for the many Linux tips and tricks and last but not least Evarist for all the many enjoyable hours we have spent together both working and biking.

I would like to thank all the patients, nurses, and laboratory personnel who contributed and made these studies possible. Furthermore, I would like to thank all the collaborators I have been working with during both these and other projects over the years. It has been very inspiring.

I want to thank HL and all my friends, both in Denmark and Lausanne, for all the fun times together and their many encouragements.

A special word of appreciation goes out to my grandfather Henning and my late grandmother, Agnethe, who have always been extraordinarily supportive and encouraging me to follow my dreams.

Furthermore, I want to thank my Mom, Dad, Betina, and Alexander for always being there when needed and for being who you are and someone whom I can always count on.

I would like to thank my two kids, Noah and Mila, for making me prove that you do not need a single full night's sleep to complete a thesis, but also for always bringing a smile to my face and reminding of the other important aspects of life.

Finally, a special thanks go out to my wife Noemi for all the love, support, and good times during these years.

Abstract

The introduction and widespread use of antiretroviral therapy against Human Immunodeficiency Virus (HIV) has had a remarkable effect on disease progression and the longevity of infected individuals. However, the establishment of a latent viral reservoir and the inability of antiretroviral compounds to purge it has resulted in HIV infection becoming a chronic disease with a high prevalence of comorbidities, adding significant strain to the healthcare systems as well as the affected patients. Human genetic variation has previously been shown to influence HIV pathogenesis as well as the risk of developing multiple common diseases in the general population. However, the influence of genetic variation on HIV positive individuals under suppressive antiretroviral therapy remains largely unknown.

This thesis examines the role of human genetic variation in determining the size and long-term dynamics of the viral reservoir, genetic risk factors for developing HIV-related non-Hodgkin lymphoma (NHL), and how genetic risk scores (GRS) can improve the prediction of chronic kidney disease (CKD) in HIV-infected individuals. Taken together, these studies delineate the role of human genetic variation in phenotypic outcomes that are highly relevant in the era of suppressive antiretroviral treatment, while also suggesting that newly developed genetic risk scores will be capable of enhancing the predictive power of current clinical risk scores.

Keywords

HIV, genetics, GWAS, exome sequencing, genetic risk scores, cancer, chronic kidney disease, latency

Résumé

L'introduction et l'utilisation à grande échelle des traitements antirétroviraux contre le virus de l'immunodéficience humaine (VIH) ont eu un effet remarquable sur la progression de la maladie et la longévité des personnes infectées. Cependant, la création d'un réservoir viral latent et l'incapacité des médicaments antirétroviraux à le purger ont fait de l'infection par le VIH une maladie chronique avec une haute prévalence de comorbidités, ajoutant une contrainte importante aux systèmes de santé ainsi qu'aux patients affectés. Il a déjà été démontré que la variation génétique humaine influence la pathogénèse du VIH ainsi que le risque de développer de nombreuses maladies communes dans la population générale. Toutefois, l'influence de la variation génétique sur les personnes séropositives au VIH sous traitement antirétroviral efficace reste encore largement inconnue.

Cette thèse examine le rôle de la variation génétique humaine dans la détermination de la taille et de la dynamique à long terme du réservoir viral, les facteurs de risque génétiques pour le développement d'un lymphome non hodgkinien (LNH) lié au VIH, et comment les scores de risque génétique (GRS) peuvent améliorer la prédiction de la maladie rénale chronique (MRC) chez les personnes vivant avec le VIH. Ensemble, ces études permettent de mesurer l'influence de la variation génétique humaine sur des phénotypes extrêmement importants à l'ère des thérapies antirétrovirales suppressives ; elles suggèrent d'autre part que de nouvelles approches fondées sur les scores de risque génétique seront capables d'améliorer le pouvoir prédictif des scores de risque clinique actuels.

Mots-clés

VIH, génétique, GWAS, séquençage de l'exome, scores de risque génétique, cancer, maladie rénale chronique, latence

Contents

List of Figures

List of Tables

Chapter 1 Introduction

Since the identification of Human Immunodeficiency Virus (HIV) as the cause of acquired immunodeficiency syndrome (AIDS) in the 1980s, HIV infection has become one of the most significant public health issues of our time. In 2018 an estimated 37.9 million (32.7-44.0 million) individuals were living with HIV globally, causing 770,000 (570,000-1.1 million) AIDS-related deaths in just that year alone (1). The high human toll has prompted an unprecedented effort by the biomedical research community, resulting in a wealth of information on the pathogenesis of HIV disease and the immune system in general, resulting in the rapid development of life-changing drugs that can efficiently fight the infection. Since the first antiretroviral drug was introduced in 1987, more than 25 different compounds have been developed for the treatment of HIV (2), resulting in a remarkable improvement in the quality of life and longevity of people living with HIV. However, while antiretroviral therapy (ART) is capable of efficiently suppressing HIV replication and restoring CD4+ T cell counts, the drugs are not capable of eradicating the virus from its cellular reservoir. This has resulted in the HIV infection shifting from an acute to a life-long (e.g., chronic) illness accompanied by an elevated prevalence of comorbidities usually associated with more advanced ageing.

Currently, the main challenges of the HIV research community remain treatment optimization, vaccine development, developing a cure, as well as the prevention and management of comorbidities. The following chapters will focus on subjects related to the latter two areas from a host genetic perspective: dynamics of the latent HIV reservoir and risk factors associated with the development of comorbidities.

1.1 HIV pathogenesis and the viral reservoir

Upon HIV infection, the disease progression follows three major phases. First, a short period knows as the acute phase occurs where high amounts of viral RNA can be observed in the blood accompanied by a substantial loss of CD4+ T cells and flu-like symptoms (Figure 1.1). The end of the acute phase and transition to the chronic phase is marked by a partial recovery of CD4+ T cell numbers and a decrease in viral RNA. If the infection is left untreated, the CD4+ T cell counts will slowly diminish during the chronic phase at highly variable paces while the viral RNA remains relatively stable (so called "set point viral load") for several years, until the individual develops overt immunodeficiency and AIDS-related illnesses. However, the initiation of ART is capable of restoring the CD4+ T cell counts to some degree and severely delaying or preventing the progression to AIDS. Although current therapies cannot eradicate the virus, they are capable of reducing the viral RNA levels to below the limit of detection of classical assays (< 50 copies of viral RNA per mL), thus preventing further CD4+ T cell decline.

The establishment of a latent viral reservoir during primary infection remains the main barrier preventing a cure for HIV. The viral reservoir is defined as cell types or anatomical sites in which replication competent virus can accumulate and persist stably over time. Following the entry into primarily activated CD4+ T cells, HIV will integrate its reverse-transcribed DNA into the host genome, leading to productive infection with a fast and efficient viral replication cycle (3). While most of the infected CD4+ T cells will be short-lived due to immune surveillance and HIV-induced pathogenicity, some will revert to a long-lived resting memory state following the conventional way in which immunologic memory is established (4, 5).

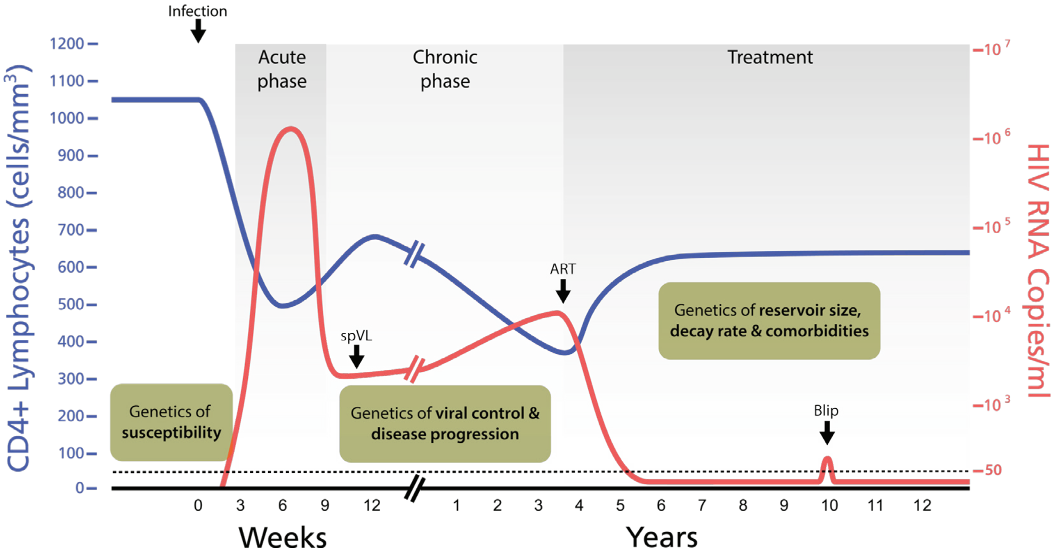


Figure 1.1. HIV phenotypes for host genetic studies over the time course of HIV infection with treatment intervention. Figure is modified from Wikimedia Commons

It is not entirely understood how the reservoir is persistently maintained in individuals under ART, but identifying the sources and factors influencing the HIV reservoir during ART is key for the development of potentially curative therapies (6, 7). Currently, clonal expansion of HIV-infected cells, long-term persistence, or ongoing cycles of active HIV replication are all considered as potential mechanisms through which the reservoir is maintained (8–11). In particular, continuous replication in lymphoid tissues in patients on ART has been reported as an important mechanism for maintaining the HIV reservoir (11). This remains controversial, as demonstrated by recent work by Bozzi *et al.* that found no evidence of ongoing HIV replication in patients on ART (12). Furthermore, the authors did not find any evidence of any local replication in blood or tissues, but only of clonal expansion of already infected cells. The latter is substantiated by the fact that treatment intensification does not seem to affect the decay of the viral reservoir (13).

Studies on the long-term dynamics of the HIV reservoir during therapy have produced a wide range of estimates on the decay rate of the viral reservoir, including large inter-individual differences, ranging from reservoir half-lifes of 2.5 months to reports of increases in reservoir size over time (14–22). A recent study in the Swiss HIV Cohort Study confirmed this variability, with 26.8% of the individuals displaying increases in their reservoir size over time, while the general reservoir decay rate for the remaining individuals decreased over time (23).

Multiple factors have been found to influence both the size of the viral reservoir and its decay rate. In particular viral blips and low-level viremia are both associated with increase in reservoir size and reduced decay rates (24, 23). The other factors that have been convincingly associated with reservoir size are CD4+ T cell counts and viral load prior to ART initiation, time from HIV infection to ART initiation, HIV subtype, ethnicity, and transmission route (23).

1.2 The rise of HIV comorbidities in the ART era

The introduction of ART has drastically increased the longevity and the quality of life of HIV infected individuals (25). Yet, the nature of this now chronic infection means that individuals infected with HIV have a highly increased risk of developing multiple comorbidities, most of which usually associated with ageing. Numerous cohort studies have established that HIV infection is associated with an increased risk of liver disease due to hepatitis B virus (HBV) and hepatitis C virus (HCV) co-infections (26), chronic kidney disease (CKD) (27, 28), acute myocardial infarction (29), ischemic stroke (30), heart failure (31), hypertension (32), osteoporosis and fractures (33, 34), as well as several types of cancer (35–37).

The rate of observed comorbidities increases with age, as also seen in the general population for these diseases, but is more common in HIV patients than in the general population (38). Furthermore, with the increased ageing of the HIV population, the prevalence of these comorbidities is also rising (39, 40). Models of the ageing HIV positive population in the Dutch ATHENA cohort predicts that by 2030, 73% of HIV positive individuals will be above 50 years old, up from 28% in 2010, with 84% of the patients having at least one comorbidity. The latter represents a substantial increase from 29% of the HIV positive individuals in 2010. Additionally, 28% of the HIV positive individuals are in 2030 predicted to suffer from three or more comorbidities (41).

A consequence of the increased prevalence of comorbidities is the potential for drug-drug interactions in the management of these comorbidities together with ART. Frequent drug-drug interactions for patients on ART have previously been reported, especially for cardiovascular drugs (42). Overall, the development of comorbidities in HIV patients is associated with decreased survival rates compared to HIV infection only (43). Thus, healthcare management of the ageing HIV population will increase in complexity over time, putting a serious strain on hospitals and clinics in limited-resource areas.

The high prevalence of comorbidities in the HIV positive population is considered to be caused by increased inflammation levels. Several studies have found elevated levels of multiple biomarkers of inflammation, including CRP, IL-6, CXCL10, soluble CD14, among others, in HIV suppressed individuals compared to HIV negative individuals (44, 45). The increased levels of these biomarkers reflect the chronic inflammation sustained in well-treated HIV positive individuals. In the general population, low-grade inflammation has also been linked to the risk of developing diabetes, cardiovascular diseases (CVDs), CKD, depression, as well as mortality (reviewed in (46)). The cause of the increased inflammation in HIV positive individuals is thought to be multi-factorial (Figure 1.2). Continuous low-level HIV production and replication during ART is considered a key driver of the low-level inflammation seen in HIV patients. Although the majority of proviruses are classified as defective at the chronic infection stage, when most individuals initiate ART (47, 48), these defective proviruses are still capable of transcribing HIV-RNA transcripts stimulating innate immunity pathways and causing low-levels of inflammation (49).

Frequent co-infections with HBV, HCV (50), human herpesvirus-8 (HHV-8) (51), Epstein-Barr virus (EBV) (52) and cytomegalovirus (CMV) (53) also contribute to the elevated inflammatory levels and constitute additional independent risk factors for the development of multiple types of comorbidities.

Traditional risk factors such as smoking, alcohol, and recreational drug use is more prevalent in HIV patients than in the general population (54–56). ART toxicity has also been associated with the development of certain comorbidities, including CKD (57). More prolonged exposure to both HIV itself and ART also increases the probability of developing comorbidities (58).

Damage to the gastrointestinal tract (GI) caused by HIV infection can also cause increased systemic inflammation contributing to the pathogenesis of comorbidities. This is due to the translocation of microbial products from the lumen of the GI tract into the circulation (59). In untreated HIV patients, viral load is correlated with levels of soluble CD14, a marker of inflammation, and levels of intestinal fatty acid-binding protein (I-FABP), a marker of GI tract enterocyte damage (60). The loss of interleukin-17 (IL-17) T_H17 cells has been suggested to contribute to increased microbial translocation (61). Additionally, a lower number of T_H17 cells has been associated with an increase in the fraction of T regulatory (T_{reg}) cells, causing further immunosuppression and contributing to the maintenance of the chronic inflammatory state observed in HIV positive individuals (62).

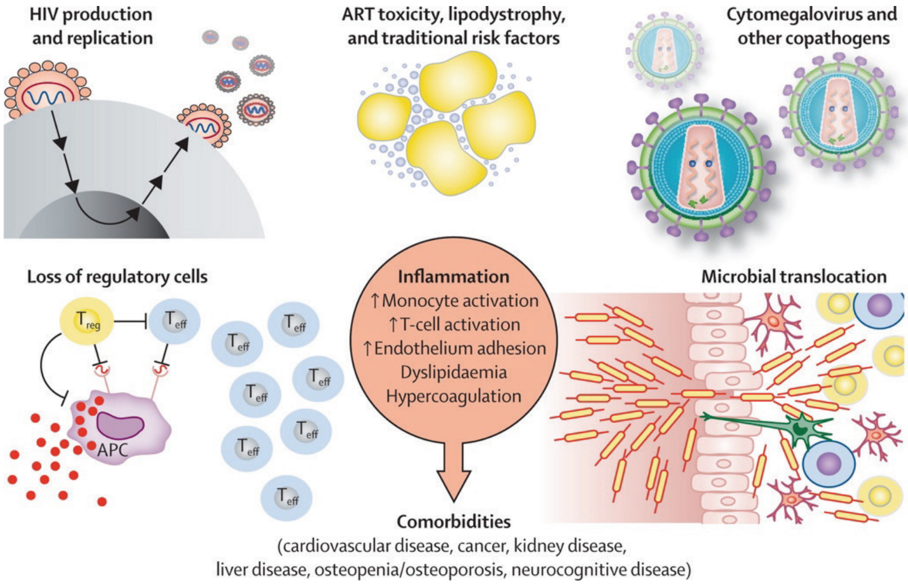

HIV production and replication
ART toxicity, lipodystrophy, and traditional risk factors
Cytomegalovirus and other copathogens
Loss of regulatory cells
T reg
T eff
APC
Inflammation
↑Monocyte activation
↑T-cell activation
↑Endothelium adhesion
Dyslipidaemia
Hypercoagulation
Microbial translocation
Comorbidities
(cardiovascular disease, cancer, kidney disease, liver disease, osteopenia/osteoporosis, neurocognitive disease)

Figure 1.2. Causes, mechanisms and consequences of inflammation in HIV infected individuals. Figure is from Deeks SG, 2013

The notion of an “accelerated ageing” process in HIV patients stems from the accumulation of all the risk factors mentioned above and the associated low-grade inflammation (63, 64). Chronically infected HIV individuals display premature ageing of the immune system believed to be caused by the accumulation of senescent CD8+ T cells producing pro-inflammatory cytokines (65, 66). Furthermore, HIV infected individuals display increased coronary artery ageing in the range of approximately 15 additional years compared to the general population (67). Similarly, the prevalence of having more than one comorbidity among HIV positive individuals equals that of 10-15 years older individuals in the general population (38). However, the exact causes of the accelerated ageing remain unclear. For CVD, a common HIV comorbidity (68), the reported cause(s) of the increased risk has been inconsistent, including inflammation caused by the retrovirus itself, the use of ART causing dyslipidemia, as well as traditional risk factors such as smoking and sedentarity (69).

1.3 Human genetic variation

Genetic variation influences thousands of human traits, with many still to be discovered. Shaped by evolutionary bottlenecks, local adaptation, and selective pressures over thousands of years, genetic variation is observed at millions of sites across the human genome. It is what makes each of us unique. In recent years, remarkable progress has been made in our understanding of human genetic variability and its association with disease susceptibility. This has primarily been driven by declining costs, improved DNA genotyping and sequencing technology and bioinformatics tools, paving the way for the establishment of large catalogs of human genetic variation within large-scale initiatives like The HapMap project (70), The 1000 Genomes Project (71) and more recently gnomAD (72) and The UK Biobank (73). Each individual carries between 4.1 and 5 million single nucleotide polymorphisms (SNPs) (71) and most human traits are influenced to some degree by genetic variation. In particular, this genetic variation has been shown to shape our individual immune responses to pathogens (74) as well as our susceptibility to diseases (75).

1.4 Methods for genetic studies

1.4.1 Genome-wide association studies

The completion of the Human Genome Project (76, 77) and subsequent population studies like The HapMap Project (70) and The 1000 Genomes Project (71), allowed for the development of a novel experimental design, the genome-wide association study (GWAS). In GWAS, the whole genome is systematically scanned for associations between genetic variants and the trait of interest. Beside genome-wide coverage, the main advantage of GWAS is its completely unbiased approach without assumptions as to the location of the associated variant(s). Before the introduction and adaptation of GWAS, genetic research (on complex traits) was primarily carried out in the form of candidate gene studies. This type of study usually focused solely on a couple of cherry-picked genetic variants hypothesized to influence the trait under question. However, the lenient threshold for significance often used in these studies ($P < 0.05$), has meant that it has not been possible to replicate the majority of the reported associations in subsequent studies (78).

The widespread adoption of GWAS in recent years, with some cohorts now including millions of individuals (79), has been facilitated by the development of relatively cheap genotyping arrays. These chips contain between 200,000 – 5 million SNPs spread across the genome. A key aspect underlying GWAS is the reliance and selection of tagging SNPs based on the linkage disequilibrium (LD) between them. LD is a measure of the correlation between genetic variants within the population. Genetic variants with a high LD exhibit similar allele frequencies as they tend to segregate together as part of genomic regions known as haplotypes. By exploiting this knowledge and using fully sequenced haplotype reference panels, it is possible to infer (e.g., impute) the missing SNPs on the genotyping arrays to obtain genome-wide coverage of the majority of common variants in the human population. Common variants are typically designated as having a minor allele frequency (MAF) above 1-5% in the population, while rare variants will have a MAF below 1%.

Similar approaches to imputation using large reference panels has also been shown to be highly effective in determining alleles of genes in highly polymorphic regions like the major histocompatibility complex (MHC) and killer cell immunoglobulin-like receptor (KIR) regions (80, 81).

The number of GWAS has increased significantly in recent years. Many such studies have had a widespread impact on our understanding of complex traits. According to the NHGRI-EBI GWAS Catalog, from 2008 to 2018, the number of published GWAS rose from 139 to 5687 studies and resulted in the identification of

71673 SNP-trait associations (82). Furthermore, the strict threshold of genome-wide significance implemented in GWAS (P < 5e-8) has meant that the replication rate of identified associations has been relatively high, even across studied population groups (83).

The identification of variants associated with a trait or disease of interest is only the first step. From there, the determination of the actual causal variant(s) is not straightforward. While the LD structure of the genome allows for the usage of SNP arrays, it also complicates the determination of the causal variant(s) (known as fine-mapping) and their effect on specific genes or pathways. Multiple bioinformatic approaches for conducting fine-mapping have been developed, and this is an area of continuous development. The use of information on different population LD structures, chromatin marks, transcription factor binding sites, and expression quantitative trait loci (eQTL) co-localization are all variables implemented in various fine-mapping approaches. Nevertheless, imperfect imputation and lack of tagging of rare variants in SNP array-based GWAS mean that this approach cannot identify the true causal variant(s) in cases where these are either rare or located in untagged haplotypes.

While the causal variant(s) are not always identified, a general conclusion from GWAS has been that most traits are polygenic. This means that genetic variants mapping to multiple loci across the genome contribute to the examined phenotypic variation. Thus, at the individual level, the combined contribution of all these variants may differ tremendously. The implication is that the effect size of each associated variant identified through GWAS is usually very small.

1.4.2 Genome sequencing studies

In recent years, further advances in sequencing technologies have enabled the previously cost-prohibited implementation of high-throughput sequencing (NGS), like whole genome sequencing (WGS) and exome sequencing. The main advantage of WGS or exome sequencing over genotyping arrays is that it makes imputation of missing SNPs and HLA alleles redundant, thus allowing for more precise determination of causal variant(s).

As with the genotyping data described above, association studies for common variants identified by means of WGS are also considered GWAS. The main distinction, besides the still-elevated costs, being the density of variants covered and more reliable coverage of rare variants. Thus, when only searching for associations with common variants, using genotyping chips is still the most cost-effective approach. However, exome and WGS approaches have the added benefit of allowing for the testing for the burden of rare variants across genes or pathways, which cannot be reliably performed using genotyping data. Furthermore, these sequencing approaches also allow for the analysis of large structural variants (≥ 50 base pair long) such as copy number variations (CNVs), rearrangements, and insertions of transposable elements.

Exome sequencing represents a cheaper alternative to WGS for studies interested in including rare protein-altering variants. Exome sequencing, as the name suggests, focuses solely on the 1-2% of the genome coding for proteins. By using capture probes targeting the coding exons of each gene, these can be isolated and sequenced as high coverage at much lower costs than WGS. Thus, exome sequencing is a powerful and cost-effective method for identifying rare protein-coding variants. However, it fails to provide any direct information on genetic variation outside the coding regions.

1.5 Host genetics of HIV pathogenesis

To date, the main goal of the HIV genetic research has been on identifying genetic variations influencing susceptibility to infection or viral control and progression to AIDS, to improve therapy options (Figure 1.2) (84). This work has primarily been driven by the early finding that individuals with a homozygous 32 base pair long deletion within the C-C motif chemokine receptor (*CCR5*) gene, known as the CCR5Δ32 deletion, are resistant to HIV acquisition (85, 86). Multiple other genetic variations have also been proposed to confer resistance to HIV acquisition in candidate gene studies. Still, only the homozygous CCR5Δ32 deletion has been replicated in other cohorts or GWAS (87). Additionally, the presence of a single CCR5Δ32 allele has also been associated with a slower progression to AIDS (86, 88). The level of viral RNA following the acute phase is relatively constant, known as the set-point viral load (spVL). Still, large inter-individual differences exist and are often used as a marker of the pace of disease progression due to its correlation with progression to AIDS (89). In fact, the first HIV GWAS examined the genetic determinants of HIV host control using spVL as outcome, discovered a major role of the HLA class I allele, *HLA-B*57*, on the level of viral load and rate of CD4+ T cell decline (90), which has later been replicated in both African and European populations (88, 91). The largest GWAS to date on spVL, involving 6315 individuals of European descent, identified only the known HLA-B*57:01 allele, and the CCR5Δ32 deletion has significant (88). Furthermore, 24.6% of the observed variation in spVL could be attributed to common genetic variants (e.g., minor allele frequency above 5%), with variants outside the MHC or CCR5 regions only explaining 5% of the total variation. GWAS, however, does not include rare variants and, as such, does not assess the impact of all functional variants. The contribution of rare functional variants to spVL was examined by exome sequencing of 1327 individuals of European descent. This study did not uncover any new associations for either rare or common variants outside of the MHC region (92). Notably, no WGS studies have been performed in HIV infected individuals as of this date.

1.5.1 Relationship between host and viral diversity

The HIV genome is characterized by its high mutation rate due to the high error rate of the reverse transcriptase enzyme and replication rate, allowing for the generation of a large degree of variation, which can affect its virulence, drug resistance and ability to evade host immune responses (93). This viral diversity is, in addition to host genetic variation, known to affect disease progression, with subtype diversity of just the polymerase gene explaining around 5.7% of the variance in spVL (94). Later studies in the Swiss HIV Cohort Study using near-full length viral sequences increased the degree of spVL variance attributed to the viral sequence up to 30% (95, 96). In a combined host/viral analysis it was found that the majority of the variance explained by the viral diversity (23.6% in total) could be explained by variation in viral epitopes or other HLA-associated positions (95), indicating that this is the result of viral evolution due to host immune pressure via HLA recognition by cytotoxic T-cells. Thus, while viral diversity is shaped by the evolutionary pressure of the host immune system, it in turn also affects the virulence and outcome of HIV disease. However, to which degree viral sequence diversity affects the risk of developing comorbidities or the viral reservoir remains unknown.

1.5.2 Influence of host genetics following treatment initiation

Most genetic studies on HIV have been conducted on ART-naïve patients, to determine factors associated with viral control and progression, as described above. However, as treatment policies have evolved towards earlier intervention, clinically relevant outcomes have also been modified. Reflecting this change, the focus of the genetic research in the HIV population must be altered to support these new challenges. The goals of human genetic research in the treated HIV population are to identify host genetic variants influencing the risk of developing comorbidities, guiding vaccine response and development, as well as the effort to create a cure (Figure 1.2). To this date, little is known about the genetic influence on these traits, as no genome-wide studies, besides those presented in chapters 3 and 4 of this thesis, have so far been performed looking at genetic variants associated with the known comorbidities in the HIV population. A few candidate gene studies have been published, but the track-record of this type of study suggests that their results are unreliable. The only relevant genome-wide study previously undertaken, although in untreated individuals, looked for genetic determinants of gut damage and microbial translocation as a consequence of HIV pathogenesis in 717 individuals. However, it did not discover any significant genetic variants associated with microbial translocation or chronic inflammation, as determined by plasma levels of soluble CD14 and I-FABP (60).

Despite the lack of previous research in the HIV setting, lessons from studies in the general population on CVD, diabetes, CKD and various cancers highlights the importance of human genetic variation in the determination of individual risk, which will also impact HIV positive individuals. In the general population, the heritability estimates (e.g., the extent to which common genetic variation contribute to a trait) for the most common comorbidities ranged from 16% for the NHL subtype diffuse large B cell lymphoma (DLBCL) (97), to 21-33% for CKD (98) and 40-60% for CAD (99). Traits with a high degree of SNP-based heritability are more likely to also include a high degree of genetic-based variation in the HIV population, making them prime areas for future HIV genetic research.

1.6 Genetic risk scores and predicting future adverse events

A primary goal for the healthcare system is to be able to predict the risk or occurrence of disease accurately. For most common diseases, including HIV comorbidities, clinical risk scores have already been generated and implemented to help guide clinical decision making (28). However, these scores do not consider the genetic factors, since the development of genetic risk scores (GRS) is still in its infancy.

In recent years, data from large GWAS has been shown to be useful for predicting genomic risk through the calculation of GRS (sometimes called polygenic risk scores (PRS)). GRS relies on the combined effect from all variants tested in a GWAS to calculate the genomic risk of an individual, as determined by the presence/absence and effect size of the deleterious allele at each associated SNP. Due to the distribution of SNPs across the genome, the combined GRS are like many other biomarkers normally distributed across the tested cohort. Thus, a GRS value for a single individual is often meaningless by itself but has to be compared in the context of a similar population group to identify outliers of increased or lowered risk across the distribution (Figure 1.3).

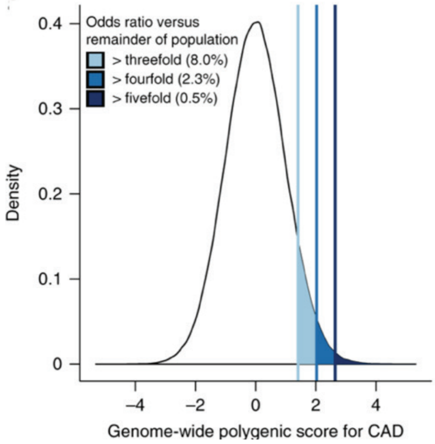


Figure 1.3. Distribution of GRS score for CAD across 288,978 individuals in the UK Biobank from (100). The coloring represents the proportion of the individuals with a three-, four- or fivefold increased risk of CAD versus the rest of the population.

Three main types of genetic risk scores have been proposed. 1) Simple risk score based on a few SNPs, in which the number of risk alleles carried by each individual is combined while ignoring their estimated effect sizes. These types of scores are often based on a few significant SNPs from GWAS or candidate gene studies. 2) GRS using the estimated effect sizes of each included significant SNP from an earlier large and well-powered GWAS. 3) Genome-wide GRS based on up to around a million SNPs. These scores incorporate effect sizes of all tested SNPs while accounting for LD, from a selected GWAS to capture all their small effects. The term GRS or polygenic risk score (PRS) is often used interchangeably between scores type 2 and 3.

The earliest GRS consisted of only the genome-wide significant variants found in the reference GWAS. However, as the polygenic nature of most complex traits became clear and it was shown that variants below the genome-wide significant threshold contributed to a large fraction of the SNP heritability (101), the use of GRS including all variants gained traction. Notably, these GRS explain more of the variance than the previous, smaller versions. This improvement has also been aided by the fact that the accuracy of a GRS is highly dependent on the heritability of the trait examined as well as the power of the GWAS for which the allelic effects sizes are obtained.

The major breakthrough for GRS came when it was shown that a genome-wide GRS had the same predictive power for breast cancer, diabetes type 2, and CAD risk as known single high-risk (e.g., mendelian) variants (100). Additionally, since GRS are based on germline DNA, they can even be calculated at or prior to birth, making early-life screening possible. Such an approach has been verified by a GRS ability to predict, at a significant degree, future-life obesity in new-born babies (102). However, despite the promising results of GRS so far, important questions regarding their usefulness across population groups remain to solved. Differences in LD patterns and allele frequencies between population groups means that GRS based on GWAS in one population do not perform well in other populations (103, 104).

1.6.1 Genetic risk scores in the HIV population

In the HIV population, the development and testing of GRS have been limited so far. The rare ones that have been tested only included a limited set of variants. Despite this, they were able to identify patients with a high risk of developing CAD (105) as well as individuals with a high risk of early treatment discontinuation (106). For diabetes, individuals with the most unfavorable GRS based on 22 SNPs identified in previous GWAS in the general population, had a significantly increased risk of developing diabetes compared to individuals with a favorable GRS. However, the addition of the GRS to a clinical risk model did not improve the predictive power of the model (107). Another GRS consisting of 42 SNPs, was found to explain the same variance in dyslipidemia as ART drugs, a well-established contributor to dyslipidemia (108). No GRS utilizing all genome-wide variants have been generated and used in the HIV population prior to the one presented in chapter 5.

1.7 Aims and overview of the thesis

The aim of this thesis was to examine the host genetic contributions to clinically important HIV phenotypes in patients on therapy. The first aim was to identify genetic factors influencing the HIV reservoir size or its decay rate, as this could potentially point to clinically actionable genes or pathways to accelerate the decay of the reservoir, paving the way for developing a cure. Meanwhile, the current lack of a cure, along with the ageing HIV population, has resulted in an ever-increasing prevalence of comorbidities in HIV infected individuals, affecting their quality of life and straining the healthcare systems in resource-limited countries. Thus, the second aim was to examine the role of common genetic variants explaining the increased prevalence of one of these comorbidities, non-Hodgkin lymphoma, in the HIV population. Last, the third aim was to examine how the development and addition of genetic risk scores can improve current clinical risk scores for HIV-related chronic kidney disease.

Chapter 2 describes a study aimed at identifying host genetic factors influencing the size of the HIV reservoir and its long-term dynamics in treated HIV infected patients.

In **Chapter 3**, the focus moves to genetic risk factors associated with the development of a common HIV comorbidity, non-Hodgkin lymphoma, using data from three international HIV cohorts.

In **Chapter 4**, a new approach combining genetic and clinical data to predict HIV infected individuals' risk of developing a common HIV comorbidity, chronic kidney disorder, is described.

Chapter 5 provides a discussion and perspectives on the future directions of genetic research in the HIV population during therapy.

Finally, **Chapter 6** summarizes the conclusions that can be made from the thesis.

1.8 References

1. UNAIDS, “UNAIDS Data 2019” (2019) (October 3, 2019).

2. A. Tseng, J. Seet, E. J. Phillips, The evolution of three decades of antiretroviral therapy: challenges, triumphs and the promise of the future. *Br. J. Clin. Pharmacol.* **79**, 182–194 (2015).

3. M. Stevenson, HIV-1 pathogenesis. *Nat. Med.* **9**, 853–860 (2003).

4. C. Van Lint, S. Bouchat, A. Marcello, HIV-1 transcription and latency: an update. *Retrovirology* **10**, 67 (2013).

5. J. D. Siliciano, *et al.*, Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4 + T cells. *Nat. Med.* **9**, 727–728 (2003).

6. T.-W. Chun, S. Moir, A. S. Fauci, HIV reservoirs as obstacles and opportunities for an HIV cure. *Nat. Immunol.* **16**, 584–589 (2015).

7. L. Shan, R. F. Siliciano, From reactivation of latent HIV-1 to elimination of the latent reservoir: The presence of multiple barriers to viral eradication. *BioEssays* **35**, 544–552 (2013).

8. M. J. Buzón, *et al.*, HIV-1 replication and immune dynamics are affected by raltegravir intensification of HAART-suppressed subjects. *Nat. Med.* **16**, 460–465 (2010).

9. N. N. Hosmane, *et al.*, Proliferation of latently infected CD4+ T cells carrying replication-competent HIV-1: Potential role in latent reservoir dynamicsProliferation of cells in the HIV reservoir. *J. Exp. Med.* **214**, 959–972 (2017).

10. Z. Wang, *et al.*, Expanded cellular clones carrying replication-competent HIV-1 persist, wax, and wane. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E2575–E2584 (2018).

11. R. Lorenzo-Redondo, *et al.*, Persistent HIV-1 replication maintains the tissue reservoir during therapy. *Nature* **530**, 51–56 (2016).

12. G. Bozzi, *et al.*, No evidence of ongoing HIV replication or compartmentalization in tissues during combination antiretroviral therapy: Implications for HIV eradication. *Sci. Adv.* **5**, eaav2045 (2019).

13. International AIDS Society Scientific Working Group on HIV Cure, *et al.*, Towards an HIV cure: a global scientific strategy. *Nat. Rev. Immunol.* **12**, 607–614 (2012).

14. M. Zanchetta, *et al.*, Long-term decay of the HIV-1 reservoir in HIV-1-infected children treated with highly active antiretroviral therapy. *J. Infect. Dis.* **193**, 1718–1727 (2006).

15. T.-W. Chun, *et al.*, Decay of the HIV reservoir in patients receiving antiretroviral therapy for extended periods: implications for eradication of virus. *J. Infect. Dis.* **195**, 1762–1764 (2007).

16. M. C. Strain, *et al.*, Heterogeneous clearance rates of long-lived lymphocytes infected with HIV: intrinsic stability predicts lifelong persistence. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 4819–4824 (2003).

17. J. Izopet, *et al.*, Decay of HIV-1 DNA in patients receiving suppressive antiretroviral therapy. *J. Acquir. Immune Defic. Syndr. Hum. Retrovirology Off. Publ. Int. Retrovirology Assoc.* **19**, 478–483 (1998).

18. B. Ramratnam, *et al.*, The decay of the latent reservoir of replication-competent HIV-1 is inversely correlated with the extent of residual viral replication during prolonged anti-retroviral therapy. *Nat. Med.* **6**, 82–85 (2000).

19. R. P. van Rij, *et al.*, Persistence of viral HLA-DR- CD4 T-cell reservoir during prolonged treatment of HIV-1 infection with a five-drug regimen. *Antivir. Ther.* **7**, 37–41 (2002).

20. A. Pires, G. Hardy, B. Gazzard, F. Gotch, N. Imami, Initiation of antiretroviral therapy during recent HIV-1 infection results in lower residual viral reservoirs. *J. Acquir. Immune Defic. Syndr. 1999* **36**, 783–790 (2004).

21. M. C. Strain, *et al.*, Effect of treatment, during primary infection, on establishment and clearance of cellular reservoirs of HIV-1. *J. Infect. Dis.* **191**, 1410–1418 (2005).

22. M. Fischer, *et al.*, Biphasic decay kinetics suggest progressive slowing in turnover of latently HIV-1 infected cells during antiretroviral therapy. *Retrovirology* **5**, 107 (2008).

23. N. Bachmann, *et al.*, Determinants of HIV-1 reservoir size and long-term dynamics during suppressive ART. *Nat. Commun.* **10**, 1–11 (2019).

24. B. Ramratnam, *et al.*, The decay of the latent reservoir of replication-competent HIV-1 is inversely correlated with the extent of residual viral replication during prolonged anti-retroviral therapy. *Nat. Med.* **6**, 82–85 (2000).

25. A. Trickey, *et al.*, Survival of HIV-positive patients starting antiretroviral therapy between 1996 and 2013: a collaborative analysis of cohort studies. *Lancet HIV* **4**, e349–e356 (2017).

26. D. Joshi, J. O'Grady, D. Dieterich, B. Gazzard, K. Agarwal, Increasing burden of liver disease in patients with HIV infection. *The Lancet* **377**, 1198–1209 (2011).

27. F. M. Islam, J. Wu, J. Jansson, D. P. Wilson, Relative risk of renal disease among people living with HIV: a systematic review and meta-analysis. *BMC Public Health* **12**, 234 (2012).

28. A. Mocroft, *et al.*, Development and Validation of a Risk Score for Chronic Kidney Disease in HIV Infection Using Prospective Cohort Data from the D:A:D Study. *PLOS Med.* **12**, e1001809 (2015).

29. M. S. Freiberg, *et al.*, HIV Infection and the Risk of Acute Myocardial Infarction. *JAMA Intern. Med.* **173**, 614–622 (2013).

30. J. L. Marcus, *et al.*, HIV infection and incidence of ischemic stroke. *AIDS* **28**, 1911 (2014).

31. A. A. Butt, *et al.*, Risk of Heart Failure With Human Immunodeficiency Virus in the Absence of Prior Diagnosis of Coronary Heart Disease. *Arch. Intern. Med.* **171**, 737–743 (2011).

32. O. Sitbon, *et al.*, Prevalence of HIV-related Pulmonary Arterial Hypertension in the Current Antiretroviral Therapy Era. *Am. J. Respir. Crit. Care Med.* **177**, 108–113 (2008).

33. A.-B. Hansen, *et al.*, Incidence of low and high-energy fractures in persons with and without HIV infection: a Danish population-based cohort study. *Aids* **26**, 285–293 (2012).

34. V. A. Triant, T. T. Brown, H. Lee, S. K. Grinspoon, Fracture Prevalence among Human Immunodeficiency Virus (HIV)-Infected Versus Non-HIV-Infected Patients in a Large U.S. Healthcare System. *J. Clin. Endocrinol. Metab.* **93**, 3499–3504 (2008).

35. P. Patel, *et al.*, Incidence of Types of Cancer among HIV-Infected Persons Compared with the General Population in the United States, 1992–2003. *Ann. Intern. Med.* **148**, 728–736 (2008).

36. M. Vogel, *et al.*, Cancer risk in HIV-infected individuals on HAART is largely attributed to oncogenic infections and state of immunocompetence. *Eur. J. Med. Res.* **16**, 101 (2011).

37. H. A. Robbins, *et al.*, Excess Cancers Among HIV-Infected People in the United States. *J. Natl. Cancer Inst.* **107** (2015).

38. G. Guaraldi, *et al.*, Premature Age-Related Comorbidities Among HIV-Infected Persons Compared With the General Population. *Clin. Infect. Dis.* **53**, 1120–1126 (2011).

39. B. Hasse, *et al.*, Morbidity and Aging in HIV-Infected Persons: The Swiss HIV Cohort Study. *Clin. Infect. Dis.* **53**, 1130–1139 (2011).

40. J. Gallant, P. Y. Hsue, S. Shreay, N. Meyer, Comorbidities Among US Patients With Prevalent HIV Infection—A Trend Analysis. *J. Infect. Dis.* **216**, 1525–1533 (2017).

41. M. Smit, *et al.*, Future challenges for clinical care of an ageing population infected with HIV: a modelling study. *Lancet Infect. Dis.* **15**, 810–818 (2015).

42. C. Marzolini, *et al.*, Ageing with HIV: medication use and risk for potential drug–drug interactions. *J. Antimicrob. Chemother.* **66**, 2107–2111 (2011).

43. N. Obel, *et al.*, Impact of Non-HIV and HIV Risk Factors on Survival in HIV-Infected Patients on HAART: A Population-Based Nationwide Cohort Study. *PLOS ONE* **6**, e22698 (2011).

44. J. Neuhaus, *et al.*, Markers of inflammation, coagulation, and renal function are elevated in adults with HIV infection. *J. Infect. Dis.* **201**, 1788–1795 (2010).

45. N. I. Wada, *et al.*, The effect of HAART-induced HIV suppression on circulating markers of inflammation and immune activation. *AIDS Lond. Engl.* **29**, 463–471 (2015).

46. D. Furman, *et al.*, Chronic inflammation in the etiology of disease across the life span. *Nat. Med.* **25**, 1822–1832 (2019).

47. Y.-C. Ho, *et al.*, Replication-competent noninduced proviruses in the latent reservoir increase barrier to HIV-1 cure. *Cell* **155**, 540–551 (2013).

48. K. M. Bruner, *et al.*, Defective proviruses rapidly accumulate during acute HIV-1 infection. *Nat. Med.* **22**, 1043–1049 (2016).

49. H. Imamichi, *et al.*, Defective HIV-1 proviruses produce novel protein-coding RNA species in HIV-infected patients on combination antiretroviral therapy. *Proc. Natl. Acad. Sci.* **113**, 8783–8788 (2016).

50. M. J. Alter, Epidemiology of viral hepatitis and HIV co-infection. *J. Hepatol.* **44**, S6–S9 (2006).

51. E. Rohner, *et al.*, HIV and human herpesvirus 8 co-infection across the globe: Systematic review and meta-analysis. *Int. J. Cancer* **138**, 45–54 (2016).

52. A. Telenti, *et al.*, Epstein-Barr virus infection in HIV-positive patients. *Eur. J. Clin. Microbiol. Infect. Dis. Off. Publ. Eur. Soc. Clin. Microbiol.* **12**, 601–609 (1993).

53. D. J. Lang, *et al.*, Seroepidemiologic studies of cytomegalovirus and epstein-barr virus infections in relation to human immunodeficiency virus type 1 infection in selected recipient populations. *J. Acquir. Immune Defic. Syndr.* **2**, 540–549 (1989).

54. R. Mdodo, *et al.*, Cigarette smoking prevalence among adults with HIV compared with the general adult population in the United States: cross-sectional surveys. *Ann. Intern. Med.* **162**, 335–344 (2015).

55. F. H. Galvan, *et al.*, The prevalence of alcohol consumption and heavy drinking among people with HIV in the United States: results from the HIV Cost and Services Utilization Study. *J. Stud. Alcohol* **63**, 179–186 (2002).

56. G. Chander, S. Himelhoch, R. D. Moore, Substance Abuse and Psychiatric Disorders in HIV-Positive Patients. *Drugs* **66**, 769–789 (2006).

57. L. Ryom, *et al.*, Association Between Antiretroviral Exposure and Renal Impairment Among HIV-Positive Persons With Normal Baseline Renal Function: the D:A:D Studya. *J. Infect. Dis.* **207**, 1359–1369 (2013).

58. G. Guaraldi, *et al.*, Aging with HIV vs. HIV Seroconversion at Older Age: A Diverse Population with Distinct Comorbidity Profiles. *PLOS ONE* **10**, e0118531 (2015).

59. G. Marchetti, C. Tincati, G. Silvestri, Microbial Translocation in the Pathogenesis of HIV Infection and AIDS. *Clin. Microbiol. Rev.* **26**, 2–18 (2013).

60. M. R. Perkins, *et al.*, The Interplay Between Host Genetic Variation, Viral Replication, and Microbial Translocation in Untreated HIV-Infected Individuals. *J. Infect. Dis.* **212**, 578–584 (2015).

61. J. M. Brenchley, *et al.*, Differential Th17 CD4 T-cell depletion in pathogenic and nonpathogenic lentiviral infections. *Blood* **112**, 2826–2835 (2008).

62. D. Favre, *et al.*, Tryptophan Catabolism by Indoleamine 2,3-Dioxygenase 1 Alters the Balance of TH17 to Regulatory T Cells in HIV Disease. *Sci. Transl. Med.* **2**, 32ra36-32ra36 (2010).

63. J. Capeau, Premature Aging and Premature Age-Related Comorbidities in HIV-Infected Patients: Facts and Hypotheses. *Clin. Infect. Dis.* **53**, 1127–1129 (2011).

64. S. Pathai, H. Bajillan, A. L. Landay, K. P. High, Is HIV a Model of Accelerated or Accentuated Aging? *J. Gerontol. Ser. A* **69**, 833–842 (2014).

65. J. P. Chou, C. M. Ramirez, J. E. Wu, R. B. Effros, Accelerated Aging in HIV/AIDS: Novel Biomarkers of Senescent Human CD8+ T Cells. *PLOS ONE* **8**, e64702 (2013).

66. R. B. Effros, The silent war of CMV in aging and HIV infection. *Mech. Ageing Dev.* **158**, 46–52 (2016).

67. G. Guaraldi, *et al.*, Coronary Aging in HIV-Infected Patients. *Clin. Infect. Dis.* **49**, 1756–1762 (2009).

68. Shah Anoop S.V., *et al.*, Global Burden of Atherosclerotic Cardiovascular Disease in People Living With HIV. *Circulation* **138**, 1100–1112 (2018).

69. Freiberg Matthew S., So-Armah Kaku, HIV and Cardiovascular Disease: We Need a Mechanism, and We Need a Plan. *J. Am. Heart Assoc.* **5**, e003411 (2016).

70. The International HapMap 3 Consortium, Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).

71. The 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

72. K. J. Karczewski, *et al.*, Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*, 531210 (2019).

73. C. Bycroft, *et al.*, The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

74. P. Scepanovic, *et al.*, Human genetic variants and age are the strongest predictors of humoral immune responses to common pathogens and vaccines. *Genome Med.* **10**, 59 (2018).

75. S. J. Chapman, A. V. S. Hill, Human genetic susceptibility to infectious disease. *Nat. Rev. Genet.* **13**, 175–188 (2012).

76. E. S. Lander, *et al.*, Initial sequencing and analysis of the human genome (2001).

77. J. C. Venter, *et al.*, The sequence of the human genome. *science* **291**, 1304–1351 (2001).

78. K. C. M. Siontis, N. A. Patsopoulos, J. P. A. Ioannidis, Replication of past candidate loci for common diseases and phenotypes in 100 genome-wide association studies. *Eur. J. Hum. Genet.* **18**, 832–837 (2010).

79. B. M. L. Baselmans, *et al.*, Multivariate genome-wide analyses of the well-being spectrum. *Nat. Genet.* **51**, 445–451 (2019).

80. X. Jia, *et al.*, Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens. *PLOS ONE* **8**, e64683 (2013).

81. D. Vukcevic, *et al.*, Imputation of KIR Types from SNP Variation Data. *Am. J. Hum. Genet.* **97**, 593–607 (2015).

82. A. Buniello, *et al.*, The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

83. U. M. Marigorta, J. A. Rodríguez, G. Gibson, A. Navarro, Replicability and Prediction: Lessons and Challenges from GWAS. *Trends Genet.* **34**, 504–517 (2018).

84. P. J. McLaren, M. Carrington, The impact of host genetic variation on infection with HIV-1. *Nat. Immunol.* **16**, 577–583 (2015).

85. M. Samson, *et al.*, Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature* **382**, 722 (1996).

86. M. Dean, *et al.*, Genetic Restriction of HIV-1 Infection and Progression to AIDS by a Deletion Allele of the CKR5 Structural Gene. *Science* **273**, 1856–1862 (1996).

87. P. J. McLaren, *et al.*, Association Study of Common Genetic Variants and HIV-1 Acquisition in 6,300 Infected Cases and 7,200 Controls. *PLoS Pathog* **9**, e1003515 (2013).

88. P. J. McLaren, *et al.*, Polymorphisms of large effect explain the majority of the host genetic contribution to variation of HIV-1 virus load. *Proc. Natl. Acad. Sci.* **112**, 14658–14663 (2015).

89. J. W. Mellors, *et al.*, Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. *Science* **272**, 1167–1170 (1996).

90. J. Fellay, *et al.*, A Whole-Genome Association Study of Major Determinants for Host Control of HIV-1. *Science* **317**, 944–947 (2007).

91. F. Pereyra, *et al.*, The Major Genetic Determinants of HIV-1 Control Affect HLA Class I Peptide Presentation. *Science* **330**, 1551–1557 (2010).

92. P. J. McLaren, *et al.*, Evaluating the Impact of Functional Genetic Variation on HIV-1 Control. *J. Infect. Dis.* **216**, 1063–1069 (2017).

93. J. Coffin, R. Swanstrom, HIV Pathogenesis: Dynamics and Genetics of Viral Populations and Infected Cells. *Cold Spring Harb. Perspect. Med.* **3**, a012526 (2013).

94. E. Hodcroft, *et al.*, The Contribution of Viral Genotype to Plasma Viral Set-Point in HIV Infection. *PLOS Pathog.* **10**, e1004112 (2014).

95. I. Bartha, *et al.*, Estimating the Respective Contributions of Human and Viral Genetic Variation to HIV Control. *PLOS Comput. Biol.* **13**, e1005339 (2017).

96. F. Bertels, *et al.*, Dissecting HIV Virulence: Heritability of Setpoint Viral Load, CD4+ T-Cell Decline, and Per-Parasite Pathogenicity. *Mol. Biol. Evol.* **35**, 27–37 (2018).

97. J. R. Cerhan, *et al.*, Genome-wide association study identifies multiple susceptibility loci for diffuse large B cell lymphoma. *Nat. Genet.* **46**, 1233–1238 (2014).

98. M. Gorski, *et al.*, 1000 Genomes-based meta-analysis identifies 10 novel loci for kidney function. *Sci. Rep.* **7**, 45040 (2017).

99. A. V. Khera, S. Kathiresan, Genetics of coronary artery disease: discovery, biology and clinical translation. *Nat. Rev. Genet.* **18**, 331–344 (2017).

100. A. V. Khera, *et al.*, Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* (2018) https:/doi.org/10.1038/s41588-018-0183-z (August 13, 2018).

101. J. Yang, *et al.*, Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).

102. A. V. Khera, *et al.*, Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. *Cell* **177**, 587-596.e9 (2019).

103. A. R. Martin, *et al.*, Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).

104. A. R. Martin, *et al.*, Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).

105. M. Rotger, *et al.*, Contribution of Genetic Background, Traditional Risk Factors, and HIV-Related Factors to Coronary Artery Disease Events in HIV-Positive Persons. *Clin. Infect. Dis.* **57**, 112–121 (2013).

106. R. Lubomirov, *et al.*, Association of Pharmacogenetic Markers with Premature Discontinuation of First-line Anti-HIV Therapy: An Observational Cohort Study. *J. Infect. Dis.* **203**, 246–257 (2011).

107. M. Rotger, *et al.*, Impact of Single Nucleotide Polymorphisms and of Clinical Risk Factors on New-Onset Diabetes Mellitus in HIV-Infected Individuals. *Clin. Infect. Dis.* **51**, 1090–1098 (2010).

108. Rotger Margalida, *et al.*, Contribution of Genome-Wide Significant Single-Nucleotide Polymorphisms and Antiretroviral Therapy to Dyslipidemia in HIV-Infected Individuals. *Circ. Cardiovasc. Genet.* **2**, 621–628 (2009).

Chapter 2 Host genomics of the HIV-1 reservoir size and its decay rate during suppressive antiretroviral treatment

Christian W. Thorball[1,*], Alessandro Borghesi[1,2,*], Nadine Bachmann[3,4], Chantal von Siebenthal[3.4], Valentina Vongrad[3,4], Teja Turk[3,4], Kathrin Neumann[3,4], Niko Beerenwinkel[5,6], Jasmina Bogojeska[7], Volker Roth[8], Yik Lim Kok[3,4], Sonali Parbhoo[8,9], Mario Wieser[8], Jürg Böni[4], Matthieu Perreau[10], Thomas Klimkait[11], Sabine Yerly[12], Manuel Battegay[13], Andri Rauch[14], Patrick Schmid[15], Enos Bernasconi[16], Matthias Cavassini[17], Roger D. Kouyos[3,4], Huldrych F. Günthard[3,4], Karin J. Metzner[3,4], Jacques Fellay[1,18,§] and the Swiss HIV Cohort Study

** these authors contributed equally to the manuscript*

[1]School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland; [2]Neonatal Intensive Care Unit, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy; [3]Department of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, Zurich, Switzerland; [4]Institute of Medical Virology, University of Zurich, Zurich, Switzerland; [5]Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland; [6]SIB Swiss Institute of Bioinformatics, Basel, Switzerland; [7]IBM Research - Zurich, Rüschlikon, Switzerland; [8]Department of Mathematics and Computer Science, University of Basel, Basel, Switzerland; [9]Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA; [10]Division of Immunology and Allergy, Lausanne University Hospital (CHUV) and University of Lausanne, Lausanne, Switzerland; [11]Division Infection Diagnostics, Department Biomedicine—Petersplatz, University of Basel, Basel, Switzerland; [12]Division of Infectious Diseases and Laboratory of Virology, University Hospital Geneva, University of Geneva, Geneva, Switzerland; [13]Department of Infectious Diseases and Hospital Epidemiology, University Hospital Basel, Basel, Switzerland; [14]Department of Infectious Diseases, University Hospital Bern, Bern, Switzerland; [15]Division of Infectious Diseases, Cantonal Hospital of St. Gallen, St. Gallen, Switzerland; [16]Infectious Diseases Service, Regional Hospital of Lugano, Lugano, Switzerland; [17]Division of Infectious Diseases, Centre Hospitalier Universitaire Vaudois, University of Lausanne, Lausanne, Switzerland; [18]Precision Medicine Unit, Lausanne University Hospital (CHUV) and University of Lausanne, Lausanne, Switzerland.

Contribution to the study: I performed all the exome, HLA and CNV analyses and together with Alessandro Borghesi the GWAS with the genotyping data. I wrote the manuscript together with Alessandro Borghesi.

This work has been submitted and is available as preprint on ***medRxiv*** (09.12.2019)

2.1 Abstract

Introduction. A major hurdle to HIV-1 eradication is the establishment of a latent viral reservoir early after primary infection. Several factors are known to influence the HIV-1 reservoir size and decay rate on suppressive antiretroviral treatment (ART), but little is known about the role of human genetic variation.

Methods. We measured the reservoir size at three time points over a median of 5.4 years, and searched for associations between human genetic variation and two phenotypic readouts: the reservoir size at the first time point and its decay rate over the study period. We assessed the contribution of common genetic variants using genome-wide genotyping data from 797 patients with European ancestry enrolled in the Swiss HIV Cohort Study and searched for a potential impact of rare variants and exonic copy number variants using exome sequencing data generated in a subset of 194 study participants.

Results. Genome- and exome-wide analyses did not reveal any significant association with the size of the HIV-1 reservoir or its decay rate on suppressive ART.

Conclusions. Our results point to a limited influence of human genetics on the size of the HIV-1 reservoir and its long-term dynamics in successfully treated individuals.

2.2 Introduction

Combination antiretroviral treatment (ART) has turned the previously lethal infection by human immunodeficiency virus type 1 (HIV-1) into a chronic disease. Despite this significant achievement, HIV-1 as retrovirus, self-integrating its genome into the host chromosome, persists indefinitely in infected individuals during treatment [1–4], and life-long ART is required to control the infection.

A major hurdle to HIV-1 eradication is the establishment, already during primary infection, of a latent viral reservoir of HIV-1 DNA persisting as provirus in resting memory CD4$^+$ T cells [1,2,5–8]. At the molecular level, chromatin remodeling, epigenetic modifications, transcriptional interference, and availability of transcription factors have been considered as possible mechanisms contributing to HIV-1 latency [9]. The viral reservoir is measurable through different methods, including viral outgrowth assay and intracellular HIV-1 DNA quantification [10,11]. Currently, there is no consensus on the best HIV-1 reservoir biomarker. Total cell-associated HIV-1 DNA, easy to measure in different cell and tissue samples and applicable to large populations, has been shown to be a good proxy for the reservoir size [12]. Indeed, while HIV-1 DNA measurement is able to detect both integrated and nonintegrated viral genomes coding for intact or defective viruses [13], total HIV-1 DNA levels have been shown to correlate with viral outgrowth [14], and to predict the time to viral rebound at treatment interruption [15]. Moreover, the substantial loss of nonintegrated HIV-1 DNA genomes following ART initiation suggests that total HIV-1 DNA after prolonged suppression is largely accounted for by integrated viral genomes [16].

After an initial rapid decay following ART initiation, changes of the viral reservoir size over time display wide inter-individual variability. By limiting dilution culture assay, the half-life of the viral reservoir was first estimated to be 44 months (95% confidence interval 27.4-114.5) in individuals with undetectable viremia [4]. A more recent study showed a slow decline of total HIV-1 DNA with a half-life of 13 years after the first four years of suppressive ART [17]. Generally, different studies show a broad variability of the average decay rate, from 2.5 months to no measurable decay [18–26]. One study even reported an increase in the viral reservoir size in as much as 31% of patients in the 4-7 years following ART initiation [27], and recent data from our group confirm this observation, reporting an increase in the reservoir size in 26.8% of individuals in the 1.5-5.5 years after ART initiation [28].

Several factors are known to influence the decay rate of the viral reservoir: initiation of ART during acute HIV-1 infection substantially accelerates the decay rate, while viral blips and low-level viremia during ART slow it down, as shown in previous studies [22] and in recent data from our cohort [28]. Conversely, treatment intensification, i.e. treating with additional drugs, does not appear to influence the decay rate, suggesting that residual replication is not the main driver of the viral reservoir [29] or that it may happen in sanctuary sites.

Human genetic variants have been shown to influence the outcome of various infections, including HIV. Previous genome-wide association studies (GWAS) addressed the role of common genetic polymorphisms in several HIV-related phenotypes, including plasma viral load (HIV-1 RNA) at set point, exceptional capacity to control viral replication, pace of CD4+ T lymphocyte decline, time to clinical AIDS, rapid progressor status or long-term non-progressor status (LTNP) [30–37], and, in one single study, the amount of intracellular HIV-1 DNA, measured at a single time point during the chronic phase of infection [38]. Rare genetic variants that are detectable through DNA sequencing technologies have been investigated far less. However, a large exome sequencing study did not reveal any convincing association of such variants with the natural history of HIV disease [39].

To date, no studies have addressed the role of human genetic variation in determining the initial viral reservoir size and the reservoir decay rate over time. In the current study, we searched for host genetic factors associated with the HIV-1 reservoir size and its long-term dynamics in a cohort of 797 HIV-1 positive individuals on suppressive ART for at least five years.

2.3 Methods

2.3.1 Ethics statement

The SHCS was approved by the local ethical committees of the participating centres: Kantonale Ethikkommission Zürich (KEK-ZH-NR: EK-793); Ethikkommission beider Basel ("Die Ethikkommission beider Basel hat die Dokumente zur Studie zustimmend zur Kenntnis genommen und genehmigt."); Kantonale Ethikkommission Bern (21/88); Comité departmental d'éthique des specialités médicales es de médecine communautarie et de premier recours, Hôpitaux Universitaires de Genève (01–142); Commission cantonale d'éthique de la recherche sur l'être humain, Canton de Vaud (131/01); Comitato etico cantonale, Repubblica e Cantone Ticino (CE 813); Ethikkommission des Kantons St. Gallen (EKSG 12/003), and written informed consent was obtained from all participants.

2.3.2 Study participants

The SHCS is an ongoing, nation-wide cohort study of HIV-positive individuals, including more than 70% of all persons living with HIV in Switzerland. Clinical and laboratory information has been prospectively recorded at follow-up visits every 3-6 months since 1988 [40]. The general enrolment criteria have been described previously [28]. Additionally, availability of genome-wide genotyping data from previous studies or of a DNA sample for genotyping was required for inclusion in this study (Figure 2.1). DNA samples used in this study were collected as part of the regular follow-up visits between May 2007 and October 2017.

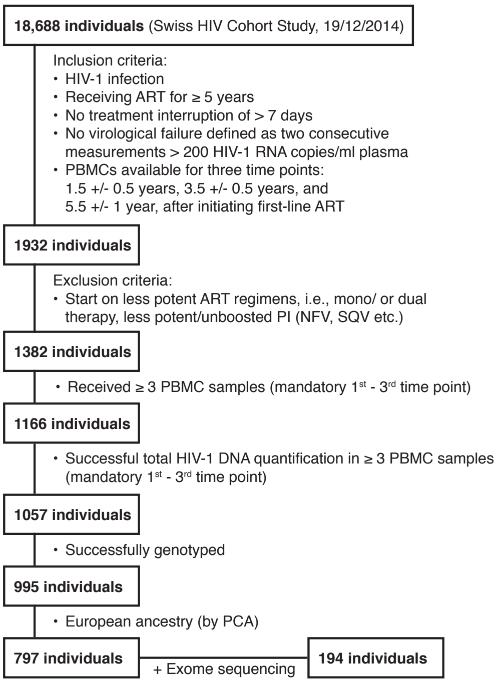

18,688 individuals (Swiss HIV Cohort Study, 19/12/2014)
Inclusion criteria:
• HIV-1 infection
• Receiving ART for ≥ 5 years
• No treatment interruption of > 7 days
• No virological failure defined as two consecutive measurements > 200 HIV-1 RNA copies/ml plasma
• PBMCs available for three time points: 1.5 +/- 0.5 years, 3.5 +/- 0.5 years, and 5.5 +/- 1 year, after initiating first-line ART
1932 individuals
Exclusion criteria:
• Start on less potent ART regimens, i.e., mono/ or dual therapy, less potent/unboosted PI (NFV, SQV etc.)
1382 individuals
• Received ≥ 3 PBMC samples (mandatory 1st - 3rd time point)
1166 individuals
• Successful total HIV-1 DNA quantification in ≥ 3 PBMC samples (mandatory 1st - 3rd time point)
1057 individuals
• Successfully genotyped
995 individuals
• European ancestry (by PCA)
797 individuals
+ Exome sequencing
194 individuals

Figure 2.1. Patient selection flowchart. Specific inclusion and exclusion criteria are listed for each selection step. ART (antiretroviral therapy); PBMCs (peripheral blood mononuclear cells); PI (protease inhibitor); PCA (principal component analysis).

2.3.3 Quantification of total HIV-1 DNA

The collection of longitudinal cryopreserved peripheral blood mononuclear cells (PBMCs) from eligible participants and the quantification of total HIV-1 DNA by droplet digital PCR has been described previously along with the calculation of the reservoir decay rate [28]. Briefly, this study utilized total HIV-1 DNA quantifications from three time points at a median of ~1.5 years, ~3.5 years, and ~5.4 years after initiation of ART.

2.3.4 Genotyping and genome-wide association analyses

Genome-wide genotyping data were obtained from previous GWAS that used various microarrays, including the HumanCore-12, HumanHap550, Human610, Human1M and Infinium CoreExome-24 BeadChips (Illumina Inc., San Diego, CA, USA), or generated from

DNA extracted from peripheral blood mononuclear cells using the HumanOmniExpress-24 BeadChip (Illumina Inc., San Diego, CA, USA).

Genotypes from each genotyping array were filtered and imputed separately, with variants first flipped to the correct strand with BCFTOOLS (v1.8) according to the human GRCh37 reference genome and filtered based on a less than 20% deviation from the 1000 genomes phase 3 EUR reference panel. Genotypes were phased, and missing genotypes were imputed with EAGLE2 [41] and PBWT [42] respectively, using the 1000 Genomes Project Phase 3 reference panel on the Sanger Imputation Service [43]. Study participants were filtered based on European ancestry as determined by principal component analysis (PCA) using EIGENSTRAT (v6.1.4) [44] and the HapMap project [45] as reference populations (Figure S2.1A). Imputed variants were filtered by minor allele frequency (MAF) < 5%, missingness > 10%, deviation from Hardy-Weinberg equilibrium (P_{HWE} < 1e-6) and imputation quality score (INFO < 0.8). The remaining genotypes were then combined using PLINK (v1.90b5) [46] prior to analyses.

To carry out the GWASs, genome-wide genotypes were tested for association with each of the two study phenotypes (reservoir size or reservoir decay rate) in two separate genome-wide association analyses. Statistical significance was set to the standard genome-wide significance threshold of P < 5e-8 to correct for multiple testing. The associations were computed using linear mixed models with genetic relationship matrixes calculated between pairs of individuals according to the leave-one-chromosome-out method as implemented in GCTA mlma-loco (v1.91.4beta) [47,48], only including age and sex as covariates, to avoid masking of true associations by confounders. To further assess the contribution of variables previously shown to be associated with either reservoir size or decay rate, we ran multiple genome-wide association analyses, each including age, sex, and one single covariate, for each of the two study phenotypes. Finally, we conducted a GWAS including all the covariates except those showing mutual correlations. These covariates included time on ART, time to viral suppression, infection stage (acute or chronic), HIV-1 RNA pre-ART, last CD4+ T cell count pre-ART, HIV-1 subtype, transmission group, and occurrence of blips or low-level viremia during treatment.

Classical HLA alleles at the four-digit level and variable amino acids within HLA proteins were imputed using SNP2HLA (v1.03) with the T1DGC reference panel consisting of 5,225 individuals of European ancestry [49]. Association analyses with the imputed HLA alleles and multi-allelic amino acids was performed using linear regressions in PLINK and multivariate omnibus tests, respectively. For all HLA analyses age, sex and the first principal component was included as covariates.

Genotypes at specific loci, i.e. the CCR5Δ32 deletion (rs333) and the HLA-B*57:01 allele, known to influence the setpoint viral load (spVL) [50,51], available from genome-wide genotyping data, were tested for association with the reservoir size and its decay rate in 797 patients. High quality genotyping information on the CCR5Δ32 deletion was available for most individuals (N = 687), while all had available HLA information.

2.3.5 Exome sequencing and analysis

All coding exons were captured using either the Illumina Truseq 65 Mb enrichment kit (Illumina Inc., San Diego, CA, USA) or the IDT xGen Exome Research Panel v1.0 (Integrated DNA Technologies Inc., Coralville, IA, USA) and sequenced on the Illumina HiSeq4000. Sequence reads were aligned to the human reference genome (GRCh37) including decoys with BWA-MEM (v0.7.10) [52]. PCR duplicates were flagged using Picard tools (v2.18.14) and variant calling performed using GATK (v3.7) [53].

To ensure a high-quality variant set across capture kit batches, all samples were merged and variants filtered based on sequencing depth (DP ≥ 20) and genotype quality (GQ ≥ 30) using BCFTOOLS (v1.8). Furthermore, individual genotypes were set as missing in cases of low depth (DP < 10) or low quality (GQ < 20). The effect of the included variants was annotated with SnpEff (v4.3T) [54].

For single variant association analysis, the VCF file was converted to PLINK format using BCFTOOLS and PLINK. Only variants with a MAF above 5%, missingness per variant below 5% and absence of severe deviation from Hardy-Weinberg equilibrium (P_{HWE} > 1e-6) were retained for the subsequent association analyses using PLINK. Sex, age and the first principal component were included as covariates. Only individuals of European descent were retained for the analyses, as determined by PCA (Figure S2.1B).

The combined effect of rare protein-altering variants (MAF < 5%), defined as either missense, stop-gain, frameshift, essential splice variant or an indel by SnpEff, on the reservoir size and decay rate was analyzed using optimal sequence kernel association tests (SKAT-O) [55]. For the decay rate, individuals were split into two groups due to the non-normal distribution; one exhibiting a very high decay over time (< -0.03 -log10(DNA)) and another with a stable reservoir size (≥ -0.03 and ≤ 0.03 -log10(DNA)). For this case-control analysis we used the SKATbinary function with linear weighted variants as implemented in the SKAT R package. In both cases, the analyses were adjusted for age, sex, and the first principal component.

Classical HLA class I and II alleles at the four-digit level were imputed from the exome sequencing data using HLA*LA [56]. All reads mapping to the MHC region or marked as unmapped were extracted using Samtools (v1.8) and used as input into HLA*LA. For association analyses, the 4-digit HLA alleles were extracted and analyzed using PyHLA [57] assuming an additive model, a minimum frequency of 5% and including age, sex and the first principal component as covariates.

2.3.6 Copy number variation

Copy number variations (CNVs) were called from exome sequencing data using CLAMMS [58]. CNVs were called for all samples in batches according to the exome capture kit used. Within batches, samples were normalized based on coverage and potential intra-batch effects adjusted for through the use of recommended mapping metrics extracted with Picard tools (v2.18.14). After CNV calling, samples with the number of CNVs two times above the median were excluded (N = 2). CNV association analyses were performed for duplications and deletions separately for common CNVs (frequency > 5 %) with PLINK adjusting for age and sex. Potential rare CNVs (frequency < 5%) impacting immune related genes were examined by overlapping called CNVs with curated immune-related genes from Immport [59] which were also listed as protein coding in GENCODE (v25).

2.3.7 Statistical analyses

All statistical analyses were performed using the R statistical software (v3.5.2), unless otherwise specified.

2.4 Results

2.4.1 Host genetic determinants of the reservoir size and long-term dynamics

To investigate the effects of host genetic variation on the size of the HIV-1 reservoir 1.5 years after ART initiation and its long-term dynamics under ART over a median duration of 5.4 years, we performed a GWAS, including 797 well-characterized HIV-1 positive individuals. All study participants were enrolled in the SHCS and were of European ancestry with longitudinal total HIV-1 DNA measurements available (Table 2.1). The median HIV-1 reservoir size was 2.76 (IQR: 2.48-3.03) log10 total HIV-1 DNA copies/1 million genomic equivalents measured ~1.5 years after initiation of ART (Figure S2.2A). The median decay rate between 1.5-5.4 years after initiation of ART was -0.06 (IQR: -0.12-0.00) log10 total HIV-1 DNA copies/1 million genomic equivalents per year (Figure S2.2B). With our sample size we had 80% power to detect variants with a MAF of 10% explaining at least 5% of the variance in HIV-1 reservoir size or decay rate [60].

Table 2.1. Patient characteristics

Total number of individuals	
Genotyped	797
Genotyped + exome sequenced	194
Age at first HIV-1 DNA sample in years	
median (IQR)	44 (38, 50)
Sex	
Female	123 (15.4%)
Male	674 (84.6%)
Transmission group	
HET	241 (30.2%)
IDU	77 (9.7%)
MSM	448 (56.2%)
Other	31 (3.9%)
HIV-1 subtype	
B	550 (69.0%)
Non-B	128 (16.1%)
Unknown	119 (14.9%)
Occurrence of blips or low-level viremia	
Blips	200 (25.1%)
Low-level viremia	68 (8.5%)
None	529 (66.4%)
Time on ART	
median (IQR)	1.50 (1.28, 1.69)
Infection stage	
Acute	140 (17.6%)
Chronic	657 (82.4%)
Time to viral suppression	
median (IQR)	0.34 (0.23, 0.51)
Log$_{10}$ HIV-1 plasma RNA pre-ART per mL	
median (IQR)	480 (248, 684)
CD4+ cell count pre-ART cells/µL blood	
median (IQR)	186 (90, 270)
HIV-1 reservoir size	
median (IQR)	2.76 (2.48, 3.03)
HIV-1 reservoir decay rate	
median (IQR)	-0.06 (-0.12, -0.00)

Transmission group indicates the self-reported route of infection (heterosexual (HET), intravenous drug usage (IDU), men who have sex with men (MSM), and other (including transfusions and unknown)). The occurrence of viral blips was defined by measurements of ≥ 50 HIV-1 RNA copies/mL plasma within a 30-day window. Individuals with consecutive measurements of ≥ 50 HIV-1 RNA copies/mL plasma for longer durations were classified as exhibiting low-level viremia. Time to viral suppression was the time from initiation of ART to the first viral load measurement below 50 copies/mL HIV-1 plasma RNA. HIV-1 reservoir size was measured in log$_{10}$ total HIV-1 DNA/1 million genomic equivalents ~1.5 years after initiating ART. The HIV-1 reservoir decay rate was based on the three measurements of total HIV-1 DNA levels taken at the median of 1.5, 3.5 and 5.4 years after initiation of ART.

First, we performed GWAS using age and sex as covariates. We did not observe any genome-wide significant variant (P < 5e-8) associated with either HIV-1 reservoir size or long-term dynamics (Figure 2.2 and S2.3). However, as we have previously determined, multiple factors are associated with the HIV-1 reservoir size and its decay rate [28], some of which are correlated (Figure S2.4). Thus, we performed additional analyses iteratively including these factors to test whether they could mask genetic associations. We ran multiple GWAS each adjusting for age, sex, plus one of the associated covariates, as well as all of the covariates together. The addition of the covariates did not have any significant effect on the results nor the genome-wide inflation factor (lambda) (Table S1).

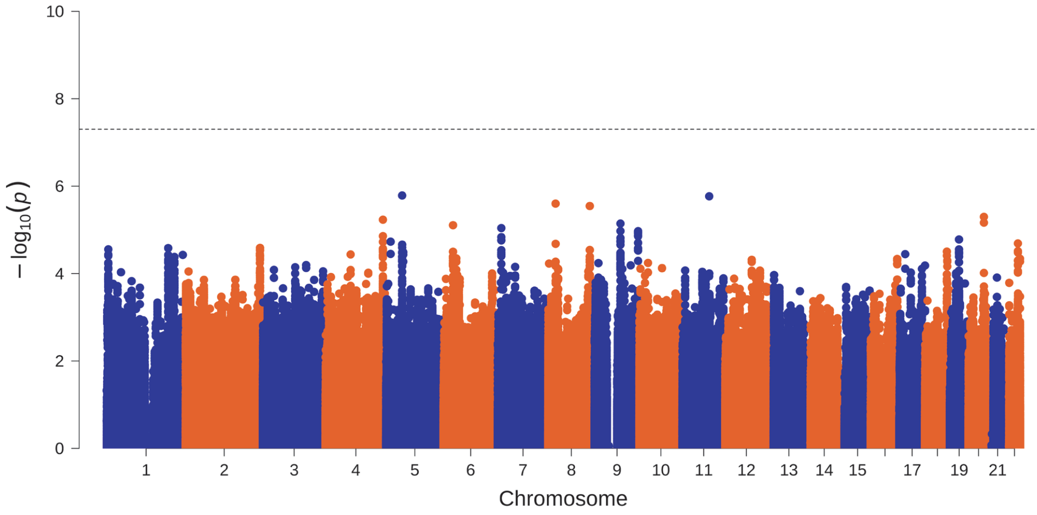


Figure 2.2. Association results with HIV-1 reservoir size. Manhattan plot with association p-values (-log$_{10}$(P)) per genetic variant plotted by genomic position. Dashed line indicates the threshold for genome-wide significance (P = 5e-8). No variants were found to be genome-wide significant.

Genetic variation in the HLA region has previously been associated with multiple HIV-related outcomes, including spVL [51]. To test whether specific HLA variants were associated with reservoir size or long-term dynamics, we imputed the HLA alleles and amino acids for all 797 individuals from the genotyping data. In line with the previous results, we did not observe any genome-wide significant associations with any HLA allele or amino acid.

2.4.2 Impact of protein-coding and rare variants

To assess the impact of rare variants as well as protein-coding variants missed by genotyping arrays on the HIV-1 reservoir size and long-term dynamics, we performed exome sequencing in 194 of the 797 study participants. Patients were selected at the two extremes of the observed reservoir decay rate: either very rapid, or absent (no change in reservoir size over ~5.4 years), while individuals with increasing HIV-1 reservoir sizes were excluded (N=12). Thus, the long-term dynamics phenotype was binarized for subsequent analyses of the decay rate, while the HIV-1 reservoir size phenotype remained normally distributed (Figure S2.5).

To ensure that no common variants, missed by the genotype chips, were associated with the HIV-1 reservoir size or long-term dynamics, we performed a GWAS for common variants using age and sex as covariates. As with the genotyping data, we observed no genome-wide significant variants for either phenotype (Figure S2.6).

We then examined the potential role of rare variants (MAF < 5%) with a functional impact defined as either missense, frameshift, stop gained, splice acceptor or donor. Since HIV-1 primarily infects CD4+ T cells, we only included variants within genes expressed in these cells as determined by Gutierrez-Arcelus *et al.* [61]. The significance threshold after correcting for the number of tests performed was P = 1.21e-5. We did not observe any significant associations for either the HIV-1 reservoir or the decay rate. The *AMBRA1* gene showed the strongest association with HIV-1 reservoir size (P = 4.15e-5, not significant) (Figure S2.7).

To confirm the lack of HLA association seen with the genotyping data, we imputed the HLA haplotypes from the exome data using HLA*LA. Again, we did not observe any significant HLA association with the study outcomes.

2.4.3 Copy number variations

To examine the role of large exonic CNVs not captured by standard genotyping and exome pipelines, we called CNVs from the mapped sequencing reads of the exome samples using the software CLAMMS. The contribution of common CNVs to HIV-1 reservoir size and long-term dynamics was analyzed by association analyses including age, sex and the first principal component as covariates. No significant association was observed after Bonferroni correction (Figure S2.8). We also searched for rare CNVs in curated immune-related genes from Immport [59] but did not discover any suggestive immune-related CNVs .

2.4.4 Influence of HLA-B*57:01 and the CCR5Δ32 deletion on reservoir size and long-term dynamics

We have previously shown that pre-ART RNA viral load levels are associated with the HIV-1 reservoir size and the occurrence of blips [28]. The HLA-B*57:01 allele and the CCR5Δ32 deletion are well known genetic variants influencing HIV-1 spVL [50,51], and could thus also be associated with the with the HIV-1 reservoir size or its decay rate. However, we did not observe any nominal association (all P > 0.05) with either reservoir size or its long-term dynamics for HLA-B*57:01 and CCR5Δ32 (Figure S2.9).

2.5 Discussion

We used a combination of genomic technologies to assess the potential role of human genetic factors in determining both the HIV-1 reservoir size and its long-term dynamics in a well-characterized, population-based cohort. We studied 797 HIV-1-positive individuals of European origin under suppressive ART over a median of 5.4 years, for whom extensive clinical data are available, allowing detailed characterization and correction for potential confounders [28]. We measured the HIV-1 reservoir size at three time points and selected two phenotypes for our genomic study: the reservoir size at ~1.5 years after ART initiation and the slope of the reservoir decay rate calculated over the three time points. Previous HIV host genetic studies mostly focused on phenotypes reflecting the natural history of HIV-1 infection, prior to ART initiation, including spontaneous viral control and disease progression [30–37]. A single study specifically tested for associations between common genetic variants and the amount of intracellular HIV-1 DNA, measured at a single time point during the chronic phase of infection prior to initiation of any antiretroviral therapy [38]. Here, in contrast, we longitudinally assessed samples collected from patients under suppressive ART to search for human genetic determinants of the long-term dynamics of the HIV-1 reservoir during treatment.

We first conducted a GWAS on 797 individuals to test for association between common genetic variants and the phenotypes. Given the small proportion of non-European subjects in the initial study cohort, we only included patients of European ancestry to avoid any false positive associations or masking of true positive associations due to different allele frequencies in small proportions of individuals belonging to different subpopulations (Figure 2.1) [62]. Regardless of including or not independent covariates other than the standard ones (i.e., sex and age), no genetic variant reached the genome-wide significance threshold for association with any of the two phenotypes. This may reflect a small effect size of genetic variants on the HIV-1 reservoir size and decay rate. We acknowledge that a larger sample size and thus increased statistical power may allow detecting genetic variants with a smaller effect size associated with the phenotypes. However, it should be noted that this study is by far the largest today that has investigated the size and decay of the HIV-reservoir in well characterized and well suppressed HIV-positive individuals over a longer time period. Alternatively, the control of the HIV-1 reservoir size and its long-term dynamics may be under the control of viral or host factors other than the individual germline genetic background. A previous report from our group had shown a correlation between viral blips during the first 1.5 years of suppressive ART and the HIV-1 reservoir size 1.5 years after ART initiation, and between viral blips after 1.5-5.4 years of suppressive ART or low-level viremia and a slower decay rate [28]. Importantly, viral blips are generally thought to reflect transient increases in viral replication, and probably occur under multifactorial influence from viral and host factors [63–70], with these latter possibly including, but not being limited to, germline genetic variation. The biological relations between viral reservoir, decay rate, viral blips, and the contribution of the individual genetic background still need full elucidation.

Standard GWAS is designed to detect associations with common genetic variants (i.e., with a MAF of at least 0.05), with little power to investigate the role of rare variants. Thus, to further assess the contribution of rare variants in individuals at the extreme of the decay rate distribution, we used exome sequencing in a selected subset of 194 study participants with very high decay rate, or conversely, a stable reservoir size over time (Figure S2.5). Here again, our analyses did not detect significant associations with the phenotypes. Although not reaching statistical significance, a rare genetic variant with potential functional impact in *AMBRA1* had a p-value for association just below the corrected threshold. The expression of *AMBRA1*, a core component of the autophagy machinery, has previously been associated with long-term viral control in HIV-1 non-progressors [71]. Future studies may further elucidate whether genetic variation in *AMBRA1* may account for inter-individual differences in the long-term dynamics of the HIV-1 reservoir.

Large deletions or duplications of genomic material may be implicated in human phenotypes, with CNVs impacting the exonic regions being more likely to have a functional role. Thus, we further investigated whether any common or rare CNV spanning exonic regions was associated with the phenotype. Again, no CNV was statistically associated with the phenotypes both in the exome-wide analyses and in analyses focused on immune-related genes.

An inherent limitation of our exome-based association analyses was their inability to detect rare variants outside the coding or splice-site regions. The exonic regions account for approximately 1-2% of the whole human genome. Because many regulatory sequences are located in extra-genic sites, our analysis did not fully investigate the role of highly conserved, non-coding genetic regions in influencing the phenotypes linked to HIV-1 latency.

Additionally, we focused on specific genetic variants, i.e., the HLA haplotypes and the CCR5Δ32 deletion, previously demonstrated to have a role in HIV-1 related phenotypes [30,51]. Indeed, previous studies unraveled a robust association between variation in the HLA region and the HIV-1 spVL [30]. Likewise, heterozygosity for the CCR5Δ32 deletion has been shown to influence spontaneous HIV-1 control [51]. Thus, we imputed HLA genotypes from genotyping and exome data, and studied the CCR5Δ32 deletion, without, however, detecting any significant associations with the phenotypes or the covariates (Figure S2.9). Specifically, we found no correlation between HLA genotypes and HIV-1 RNA plasma levels prior to ART initiation, apparently contrasting with the previous findings of an association between HLA-B*57:01 haplotype and spVL. This probably reflects historical changes in the therapeutic approach following a diagnosis of HIV-1 infection, given that ART is currently initiated soon after clinical diagnosis, before most patients reach a stable plateau of plasma viral load.

In our study, the quantification of the reservoir size at different time points may have been influenced by factors as, for example, blips and low-level viremia, which may have reduced our ability to detect significant genetic effects. It is also possible that, in the future, novel methods to assess the viral reservoir will allow the detection of significant contributions of genetic factors [72]. So far, it remains unanswered whether the initial response to acute infection, the containment of ongoing replication, and the control of latently infected cells are under the influence of the same or different molecular networks. It needs to be noted that in previous work we have shown that host genetic factors as defined by GWAS did not explain the severity of symptoms during acute HIV-infection, although severity of symptoms correlated well with viral load and CD4 cell counts [73].

2.6 Conclusion

In conclusion, our study suggests that human individual germline genetic variation has little, if any, influence on the control of the HIV-1 viral reservoir size and its long-term dynamics. Complex, likely multifactorial biological processes govern HIV-1 viral persistence. Larger studies will possibly clarify the role of common or rare genetic variants explaining small proportions of the variability of the phenotypes related to viral latency.

2.7 Declarations

2.7.1 Data availability

The datasets generated during and/or analyzed during the current study are not publicly available due to privacy reasons, the sensitivities associated with HIV infections, and the representativeness of the dataset, but is available on request.

2.7.2 Conflicts of interest statement

H.F.G. has received unrestricted research grants from Gilead Sciences and Roche; fees for data and safety monitoring board membership from Merck; consulting/advisory board membership fees from Gilead Sciences, ViiV, Merck, Sandoz and Mepha.

T.K. has received consulting/advisory board membership fees from Gilead Sciences and from ViiV Healthcare for work that has no connection to the work presented here.

K.J.M. has received travel grants and honoraria from Gilead Sciences, Roche Diagnostics, GlaxoSmithKline, Merck Sharp & Dohme, Bristol-Myers Squibb, ViiV and Abbott; and the University of Zurich received research grants from Gilead Science, Roche, and Merck Sharp & Dohme for studies that Dr. Metzner serves as principal investigator, and advisory board honoraria from Gilead Sciences.

A.R. reports support to his institution for advisory boards and/or travel grants from MSD, Gilead Sciences, Pfizer and Abbvie, and an investigator initiated trial (IIT) grant from Gilead Sciences. All remuneration went to his home institution and not to A.R. personally, and all remuneration was provided outside the submitted work.

All other authors declare no competing financial interests.

2.7.3 Author contributions

N.B., J.B., V.R., R.D.K., H.F.G., K.J.M., and J.F. contributed to the conception and design of the study. C.v.S., V.V., K.N., Y.I.K., and K.J.M. contributed to the acquisition of data. A.B., C.W.T., N.B., T.T., S.P., M.W., R.D.K., and J.F. contributed to the analysis and interpretation of data. J.B., M.P., T.K., S.Y., M.B., A.R., P.S., E.B., M.C., H.F.G., and the members of the Swiss HIV Cohort Study (SHCS) conceived and managed the cohort, collected and contributed patient samples and clinical data. A.B., C.W.T., and J.F. contributed to the drafting the article. All authors read and approved the final manuscript.

2.7.4 Acknowledgements

We thank the patients for participating in the SHCS, the study nurses and physicians for excellent patient care, A. Scherrer, A. Traytel, and S. Wild for excellent data management and D. Perraudin and M. Amstutz for administrative assistance.

This work was funded within the framework of the Swiss HIV Cohort Study (SNF grant# 33CS30_177499 to H.F.G.). The data were gathered by the Five Swiss University Hospitals, two Cantonal Hospitals, 15 affiliated hospitals and 36 private physicians (listed in http://www.shcs.ch/180-health-care-providers). The work was furthermore supported by the Systems.X grant # 51MRP0_158328 (to N.B., J.B., J.F.,V.R., R.D.K., M.W., H.F.G. and K.J.M.), by SNF grant 324730B_179571 (to H.F.G.), SNF grant SNF 310030_141067/1 (to H.F.G. and K.J.M.), SNF grants no. PZ00P3-142411 and BSSGI0_155851 to R.D.K., the Yvonne-Jacob Foundation (to H.F.G.), the University of Zurich's Clinical Research Priority Program viral infectious disease, ZPHI (to H.F.G) and the Vontobel Foundation (to H.F.G. and K.J.M.). M.W. is partially supported by the NCCR MARVEL, funded by the Swiss National Science Foundation.

Members of the Swiss HIV Cohort Study:

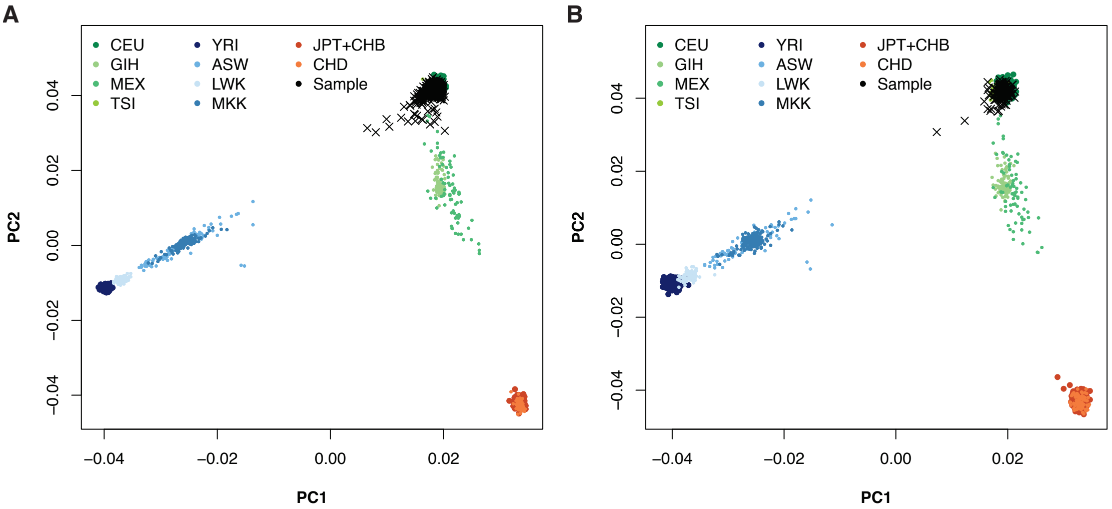
Anagnostopoulos A, Battegay M, Bernasconi E, Böni J, Braun DL, Bucher HC, Calmy A, Cavassini M, Ciuffi A, Dollenmaier G, Egger M, Elzi L, Fehr J, Fellay J, Furrer H, Fux CA, Günthard HF (President of the SHCS), Haerry D (deputy of "Positive Council"), Hasse B, Hirsch HH, Hoffmann M, Hösli I, Huber M, Kahlert CR (Chairman of the Mother & Child Substudy), Kaiser L, Keiser O, Klimkait T, Kouyos RD, Kovari H, Ledergerber B, Martinetti G, Martinez de Tejada B, Marzolini C, Metzner KJ, Müller N, Nicca D, Paioni P, Pantaleo G, Perreau M, Rauch A (Chairman of the Scientific Board), Rudin C, Scherrer AU (Head of Data Centre), Schmid P, Speck R, Stöckle M (Chairman of the Clinical and Laboratory Committee), Tarr P, Trkola A, Vernazza P, Wandeler G, Weber R, Yerly S.

2.8 Supplementary tables and figures

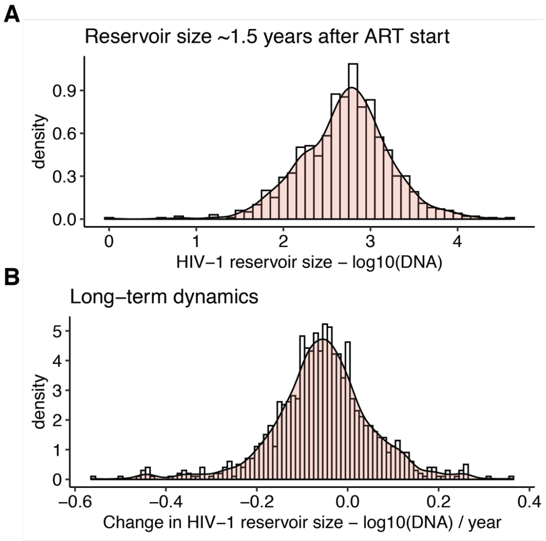
Supplementary Table 2.1. GWAS sensitivity analysis

Phenotype	Covariates	min P	Lambda
HIV-1 reservoir size	Basic (age + sex)	1.64E-06	1.02
	RNA pre-ART	2.78E-06	1.01
	CD4 pre-ART	1.54E-06	1.02
	Time on ART	1.96E-06	1.02
	Time to suppression	1.13E-06	1.02
	Stage	1.10E-06	1.02
	T_group.HET	1.71E-06	1.02
	T_group.IDU	1.82E-06	1.02
	T_group.MSM	1.68E-06	1.02
	T_group.OTHER	1.53E-06	1.02
	Subtype - B	1.21E-06	1.02
	Subtype - Non B	1.69E-06	1.02
	Subtype - Unknown	1.52E-06	1.02
	Blips	9.99E-07	1.02
	Low-level viremia	2.76E-06	1.01
	No blips or viremia	1.43E-06	1.02
	All	2.39E-06	1.01
HIV-1 reservoir decay rate	Basic (age + sex)	2.05E-06	1.00
	RNA pre-ART	2.12E-06	1.00
	CD4 pre-ART	1.98E-06	1.00
	Time on ART	1.97E-06	1.00
	Time to suppression	1.91E-06	1.00
	Stage	2.03E-06	1.00
	T_group.HET	2.02E-06	1.00
	T_group.IDU	1.94E-06	1.00
	T_group.MSM	1.94E-06	1.00
	T_group.OTHER	2.21E-06	1.00
	Subtype - B	2.08E-06	1.00
	Subtype - Non B	2.14E-06	1.00
	Subtype - Unknown	2.11E-06	1.00
	Blips	2.21E-06	1.00
	Low-level viremia	2.24E-06	1.00
	No blips or viremia	2.07E-06	1.00
	All	2.26E-06	0.99

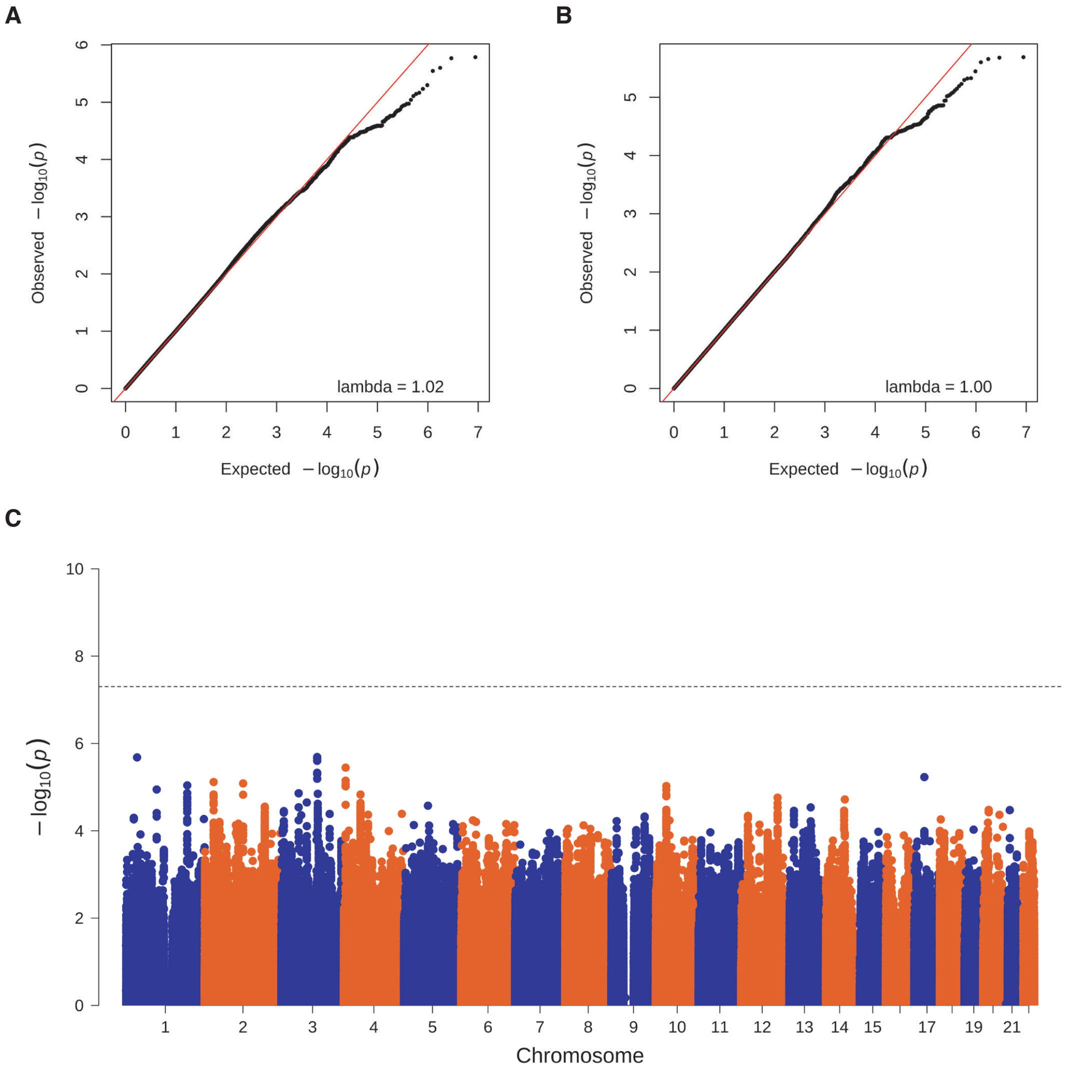
Covariates indicates the covariates added in the linear mixed model together with age and sex. Transmission group indicates the self-reported route of infection (heterosexual (HET), intravenous drug usage (IDU), men who have sex with men (MSM), and other (including transfusions and unknown)). The occurrence of viral blips was defined by measurements of ≥ 50 HIV-1 RNA copies/mL plasma within a 30-day window. Individuals with consecutive measurements of ≥ 50 HIV-1 RNA copies/mL plasma for longer durations were classified as exhibiting low-level viremia. Time to viral suppression was the time from initiation of ART to the first viral load measurement below 50 copies/mL HIV-1 plasma RNA. All, indicates that all independent covariates listed in the table was added to the model. Min P is the minimal observed p-value in the corresponding GWAS. Lambda indicates the genomic inflation factor. All values were ~1.00, indicating an absence of genomic inflation of the test statistics due to confounding factors.



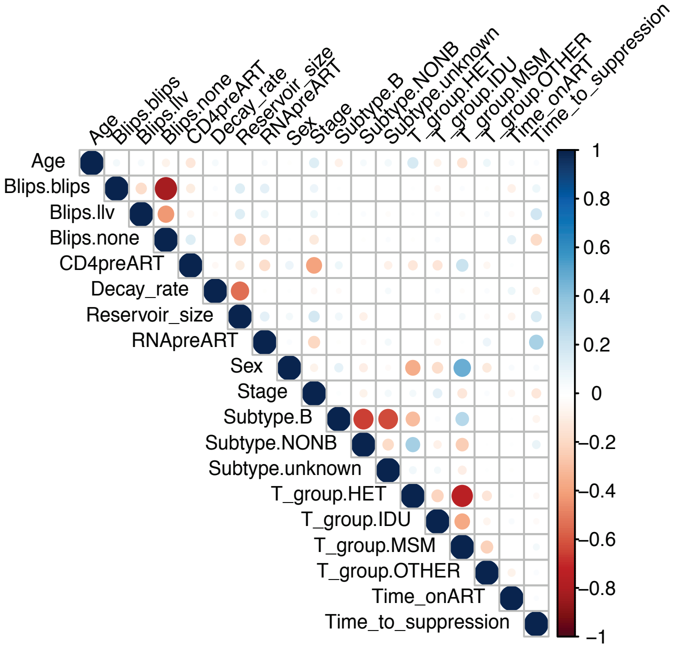
Supplementary Figure 2.1. Principal component analyses (PCA) with population references from the HapMap project. (A) The 797 genotyped individuals (black crosses) colocalizing with the HapMap reference samples from CEU (Northern Europeans from Utah) and TSI (Tuscans from Italy). (B) The 194 individuals also exome sequenced (black crosses) cluster as expected still with CEU and TSI.



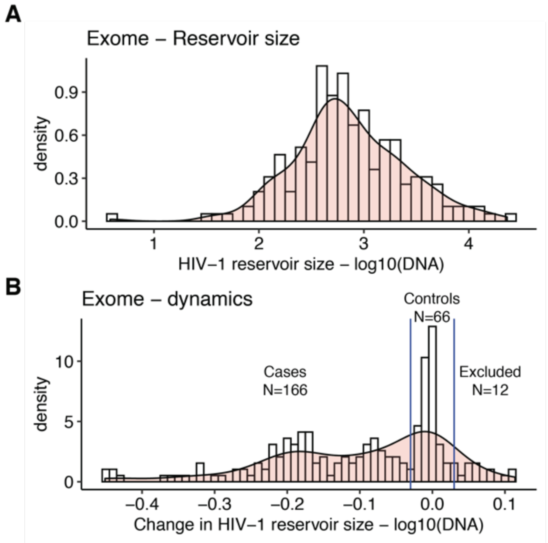
Supplementary Figure 2.2. The HIV-1 reservoir size and decay rates for 797 genotyped individuals on ART for a median of 5.4 years. (A) Histograms of the HIV-1 reservoir size in $\log_{10}$ total HIV-1 DNA/1 million genomic equivalents measured ~1.5 after initiating ART. (B) Histogram of the HIV-1 reservoir decay rate in $\log_{10}$ total HIV-1 DNA/1 million genomic equivalents per year based on the three measurements of total HIV-1 DNA levels taken at the median of 1.5, 3.5 and 5.4 years after initiation of ART.



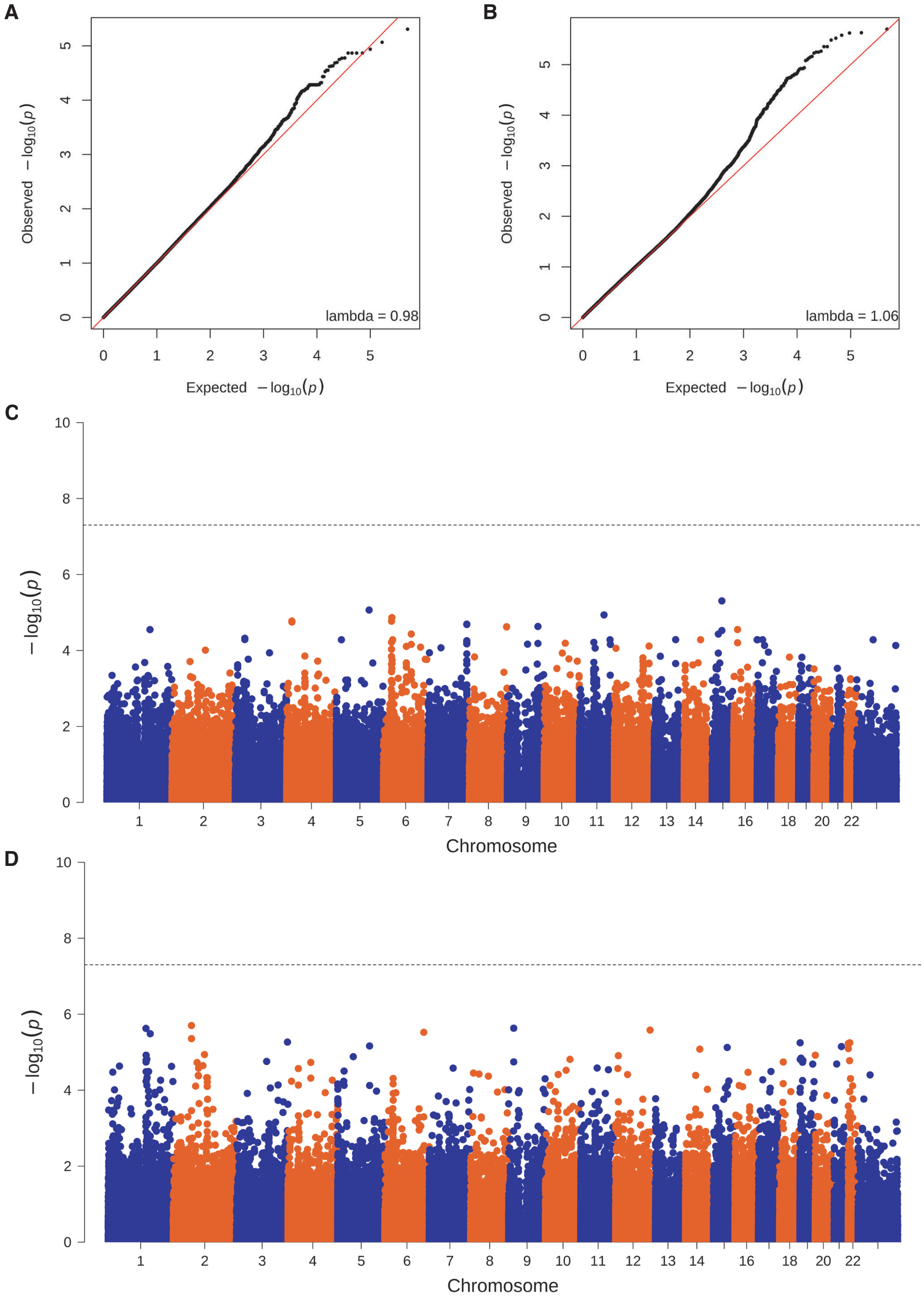
Supplementary Figure 2.3. GWAS results for the HIV-1 reservoir size and decay rate. (A) Quantile-quantile plot for the GWAS of HIV-1 reservoir size showing the observed -log10(P) (black dots, y-axis) versus expected -log10(P) under the null hypothesis (red line). Lambda indicates the genomic inflation factor. Values ~1 indicates the lack of genomic inflation due to confounding factors (B) Quantile-quantile plot for the GWAS of the HIV-1 decay rate between 1.5 – 5.5 years after ART initiation. (C) Manhattan plot for the GWAS of the HIV-1 decay rate with association p-values per genetic variant plotted by genomic position. Dashed line indicates the threshold for genome-wide significance (P = 5e-8). No variants were found to be genome-wide significant.



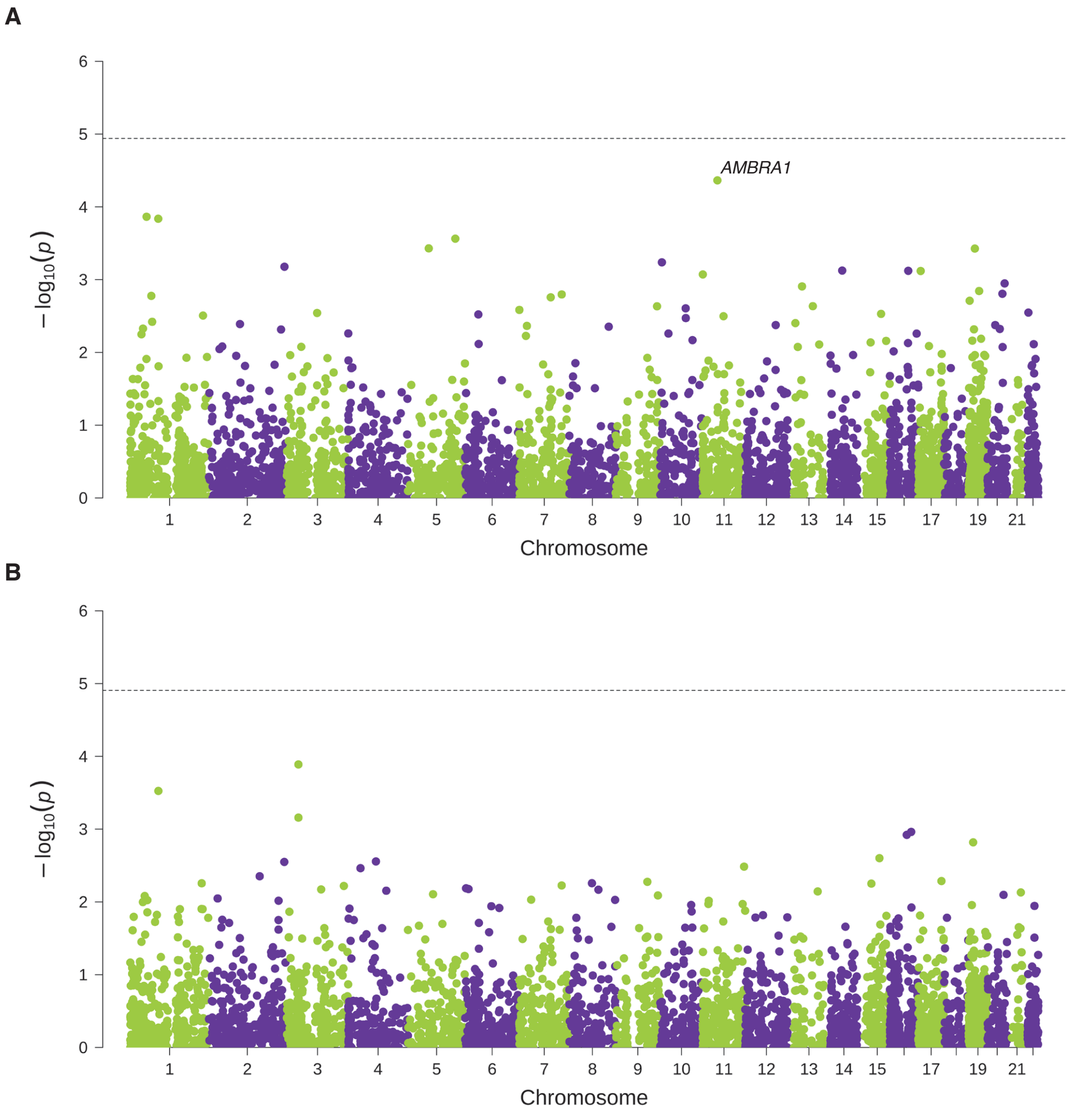
Supplementary Figure 2.4. Correlations between known determinants of the HIV-1 reservoir size and decay rate and other variables. The size and color intensity of the circles indicates the level of correlation between variables. Blips.blips refers to the occurrence of viral blips defined by measurements of ≥ 50 HIV-1 RNA copies/mL plasma within a 30-day window. Blips.llv refers to individuals exhibiting low-level viremia defined as consecutive measurements of ≥ 50 HIV-1 RNA copies/mL plasma for more than 30 days. Transmission group (T_group) indicates the self-reported route of infection (heterosexual (HET), intravenous drug usage (IDU), men who have sex with men (MSM), and other (including transfusions and unknown)). Time to viral suppression was the time from initiation of ART to the first viral load measurement below 50 copies/mL HIV-1 plasma RNA. HIV-1 reservoir size was measured in log10 total HIV-1 DNA/1 million genomic equivalents ~1.5 years after initiating ART.



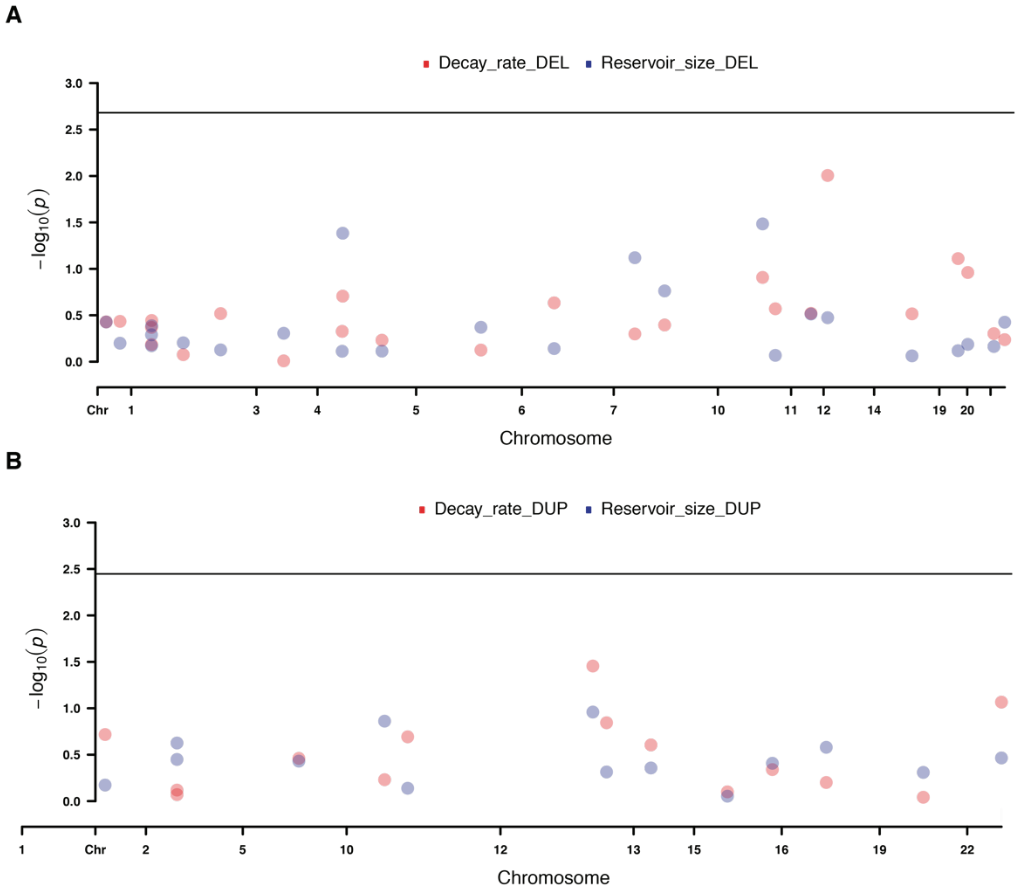
Supplementary Figure 2.5. The HIV-1 reservoir size and decay rates for the 194 exome sequenced individuals. (A) Histogram of the HIV-1 reservoir size in $\log_{10}$ total HIV-1 DNA/1 million genomic equivalents measured ~1.5 after initiating ART. (B) The non-normal distribution of the decay rate (long-term dynamics) meant that the individuals were split into cases with a decreasing HIV-1 reservoir and controls with no change in their HIV-1 reservoir size over 5.4 years. The blue lines indicate the cutoff points. Individuals with increasing HIV-1 reservoir sizes were excluded (N=12).



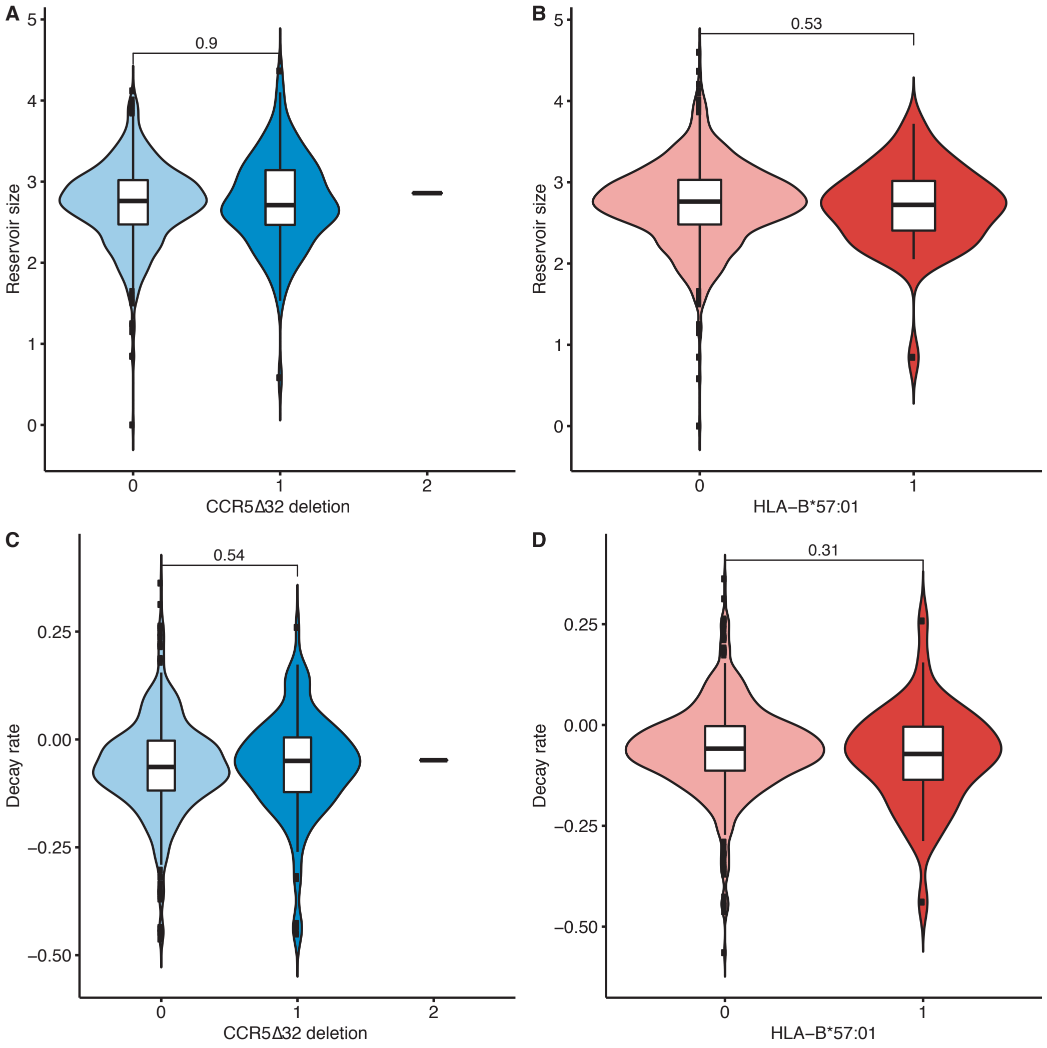
Supplementary Figure 2.6. Associations with common variants in 194 exome sequenced individuals. (A) Quantile-quantile plot for GWAS of HIV-1 reservoir size showing the observed -log10(P) (black dots, y-axis) versus expected -log10(P) under the null hypothesis (red line). (B) Quantile-quantile plot for GWAS of the HIV-1 decay rate between 1.5 – 5.5 years after ART initiation. (C) Manhattan plot for GWAS of the HIV-1 reservoir size with association p-values per genetic variant plotted by genomic position. Dashed line indicates the threshold for genome-wide significance (P = 5e-8). No variants were found to be genome-wide significant. (D) Manhattan plot for the GWAS of the HIV-1 decay rate. No variants were found to be genome-wide significant.



Supplementary Figure 2.7. Association results for rare functional variants with optimal sequence kernel association tests (SKAT-O). (A) Manhattan plot for rare variant association analysis using SKAT-O of the HIV-1 reservoir size with association p-values per genetic variant plotted by genomic position. Dashed line indicates the threshold for genome-wide significance following Bonferroni correction (P = 1.21e-5). No variants were found to be genome-wide significant. (B) Manhattan plot for the rare variant association analysis of the HIV-1 decay rate. No variants were found to be genome-wide significant.



Supplementary Figure 2.8. Association results for common exonic copy number variations (CNVs). (A) Manhattan plot for CNV deletions with association p-values per CNV plotted by genomic position. The association of each CNV with either reservoir size (yellow dots) or decay rate (blue dots) is indicated. (B) Manhattan plot for CNV duplications with association p-values per CNV plotted by genomic position. The association of each CNV with either reservoir size (yellow dots) or decay rate (blue dots) is marked.



Supplementary Figure 2.9. Effect of CCR5Δ32 deletion and the HLA-B*57:01 allele on HIV-1 reservoir size and decay rate in 797 individuals. Statistical differences between groups were determined by Wilcoxon signed-rank tests. (A) Violin plots showing the HIV-1 reservoir size in log10 total HIV-1 DNA/1 million genomic equivalents of individuals carrying either none, one or two alleles of the CCR5Δ32 deletion as tagged by the rs333 SNP. (B) Violin plot with the HIV-1 reservoir size levels grouped by presence of the HLA-B*57:01 allele. (C) Violin plot showing HIV-1 reservoir decay rates of individuals with and without the CCR5Δ32 deletion. (D) Violin plot with the decay rates of individuals with and without the HLA-B*57:01 allele.

2.9 References

1. Finzi D, Hermankova M, Pierson T, Carruth LM, Buck C, Chaisson RE, et al. Identification of a Reservoir for HIV-1 in Patients on Highly Active Antiretroviral Therapy. Science. 1997 Nov 14;278(5341):1295–300.

2. Wong JK, Hezareh M, Günthard HF, Havlir DV, Ignacio CC, Spina CA, et al. Recovery of Replication-Competent HIV Despite Prolonged Suppression of Plasma Viremia. Science. 1997 Nov 14;278(5341):1291–5.

3. Chun T-W, Stuyver L, Mizell SB, Ehler LA, Mican JAM, Baseler M, et al. Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. Proc Natl Acad Sci. 1997 Nov 25;94(24):13193–7.

4. Siliciano JD, Kajdas J, Finzi D, Quinn TC, Chadwick K, Margolick JB, et al. Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4$^+$ T cells. Nat Med. 2003 Jun;9(6):727–8.

5. Chun TW, Carruth L, Finzi D, Shen X, DiGiuseppe JA, Taylor H, et al. Quantification of latent tissue reservoirs and total body viral load in HIV-1 infection. Nature. 1997 May 8;387(6629):183–8.

6. Chun TW, Engel D, Berrey MM, Shea T, Corey L, Fauci AS. Early establishment of a pool of latently infected, resting CD4(+) T cells during primary HIV-1 infection. Proc Natl Acad Sci U S A. 1998 Jul 21;95(15):8869–73.

7. Smith MZ, Wightman F, Lewin SR. HIV reservoirs and strategies for eradication. Curr HIV/AIDS Rep. 2012 Mar;9(1):5–15.

8. Siliciano RF, Greene WC. HIV latency. Cold Spring Harb Perspect Med. 2011 Sep;1(1):a007096.

9. Ruelas DS, Greene WC. An integrated overview of HIV-1 latency. Cell. 2013 Oct 24;155(3):519–29.

10. Han Y, Wind-Rotolo M, Yang H-C, Siliciano JD, Siliciano RF. Experimental approaches to the study of HIV-1 latency. Nat Rev Microbiol. 2007 Feb;5(2):95–106.

11. Hodel F, Patxot M, Snäkä T, Ciuffi A. HIV-1 latent reservoir: size matters. Future Virol. 2016 Dec;11(12):785–94.

12. Avettand-Fènoël V, Hocqueloux L, Ghosn J, Cheret A, Frange P, Melard A, et al. Total HIV-1 DNA, a Marker of Viral Reservoir Dynamics with Clinical Implications. Clin Microbiol Rev. 2016 Oct;29(4):859–80.

13. Bruner KM, Wang Z, Simonetti FR, Bender AM, Kwon KJ, Sengupta S, et al. A quantitative approach for measuring the reservoir of latent HIV-1 proviruses. Nature. 2019;566(7742):120–5.

14. Kiselinova M, De Spiegelaere W, Buzon MJ, Malatinkova E, Lichterfeld M, Vandekerckhove L. Integrated and Total HIV-1 DNA Predict Ex Vivo Viral Outgrowth. PLoS Pathog. 2016 Mar;12(3):e1005472.

15. Williams JP, Hurst J, Stöhr W, Robinson N, Brown H, Fisher M, et al. HIV-1 DNA predicts disease progression and post-treatment virological control. eLife. 2014 Sep 12;3:e03821.

16. Koelsch KK, Liu L, Haubrich R, May S, Havlir D, Günthard HF, et al. Dynamics of total, linear nonintegrated, and integrated HIV-1 DNA in vivo and in vitro. J Infect Dis. 2008 Feb 1;197(3):411–9.

17. Gandhi RT, McMahon DK, Bosch RJ, Lalama CM, Cyktor JC, Macatangay BJ, et al. Levels of HIV-1 persistence on antiretroviral therapy are not associated with markers of inflammation or activation. PLOS Pathog. 2017 Apr 20;13(4):e1006285.

18. Zanchetta M, Walker S, Burighel N, Bellanova D, Rampon O, Giaquinto C, et al. Long-term decay of the HIV-1 reservoir in HIV-1-infected children treated with highly active antiretroviral therapy. J Infect Dis. 2006 Jun 15;193(12):1718–27.

19. Chun T-W, Justement JS, Moir S, Hallahan CW, Maenza J, Mullins JI, et al. Decay of the HIV reservoir in patients receiving antiretroviral therapy for extended periods: implications for eradication of virus. J Infect Dis. 2007 Jun 15;195(12):1762–4.

20. Strain MC, Günthard HF, Havlir DV, Ignacio CC, Smith DM, Leigh-Brown AJ, et al. Heterogeneous clearance rates of long-lived lymphocytes infected with HIV: intrinsic stability predicts lifelong persistence. Proc Natl Acad Sci U S A. 2003 Apr 15;100(8):4819–24.

21. Izopet J, Salama G, Pasquier C, Sandres K, Marchou B, Massip P, et al. Decay of HIV-1 DNA in patients receiving suppressive antiretroviral therapy. J Acquir Immune Defic Syndr Hum Retrovirology Off Publ Int Retrovirology Assoc. 1998 Dec 15;19(5):478–83.

22. Ramratnam B, Mittler JE, Zhang L, Boden D, Hurley A, Fang F, et al. The decay of the latent reservoir of replication-competent HIV-1 is inversely correlated with the extent of residual viral replication during prolonged anti-retroviral therapy. Nat Med. 2000 Jan;6(1):82–5.

23. van Rij RP, van Praag RME, Prins JM, Rientsma R, Jurriaans S, Lange JMA, et al. Persistence of viral HLA-DR- CD4 T-cell reservoir during prolonged treatment of HIV-1 infection with a five-drug regimen. Antivir Ther. 2002 Mar;7(1):37–41.

24. Pires A, Hardy G, Gazzard B, Gotch F, Imami N. Initiation of antiretroviral therapy during recent HIV-1 infection results in lower residual viral reservoirs. J Acquir Immune Defic Syndr 1999. 2004 Jul 1;36(3):783–90.

25. Strain MC, Little SJ, Daar ES, Havlir DV, Gunthard HF, Lam RY, et al. Effect of treatment, during primary infection, on establishment and clearance of cellular reservoirs of HIV-1. J Infect Dis. 2005 May 1;191(9):1410–8.

26. Fischer M, Joos B, Niederöst B, Kaiser P, Hafner R, von Wyl V, et al. Biphasic decay kinetics suggest progressive slowing in turnover of latently HIV-1 infected cells during antiretroviral therapy. Retrovirology. 2008 Nov 26;5:107.

27. Besson GJ, Lalama CM, Bosch RJ, Gandhi RT, Bedison MA, Aga E, et al. HIV-1 DNA decay dynamics in blood during more than a decade of suppressive antiretroviral therapy. Clin Infect Dis Off Publ Infect Dis Soc Am. 2014 Nov;59(9):1312–21.

28. Bachmann N, Siebenthal C von, Vongrad V, Turk T, Neumann K, Beerenwinkel N, et al. Determinants of HIV-1 reservoir size and long-term dynamics during suppressive ART. Nat Commun. 2019 Jul 19;10(1):1–11.

29. International AIDS Society Scientific Working Group on HIV Cure, Deeks SG, Autran B, Berkhout B, Benkirane M, Cairns S, et al. Towards an HIV cure: a global scientific strategy. Nat Rev Immunol. 2012 20;12(8):607–14.

30. Fellay J, Shianna KV, Ge D, Colombo S, Ledergerber B, Weale M, et al. A whole-genome association study of major determinants for host control of HIV-1. Science. 2007 Aug 17;317(5840):944–7.

31. Fellay J, Ge D, Shianna KV, Colombo S, Ledergerber B, Cirulli ET, et al. Common genetic variation and the control of HIV-1 in humans. PLoS Genet. 2009 Dec;5(12):e1000791.

32. Pelak K, Goldstein DB, Walley NM, Fellay J, Ge D, Shianna KV, et al. Host determinants of HIV-1 control in African Americans. J Infect Dis. 2010 Apr 15;201(8):1141–9.

33. International HIV Controllers Study, Pereyra F, Jia X, McLaren PJ, Telenti A, de Bakker PIW, et al. The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. Science. 2010 Dec 10;330(6010):1551–7.

34. Herbeck JT, Gottlieb GS, Winkler CA, Nelson GW, An P, Maust BS, et al. Multistage genomewide association study identifies a locus at 1q41 associated with rate of HIV-1 disease progression to clinical AIDS. J Infect Dis. 2010 Feb 15;201(4):618–26.

35. Le Clerc S, Limou S, Coulonges C, Carpentier W, Dina C, Taing L, et al. Genomewide association study of a rapid progression cohort identifies new susceptibility alleles for AIDS (ANRS Genomewide Association Study 03). J Infect Dis. 2009 Oct 15;200(8):1194–201.

36. Limou S, Le Clerc S, Coulonges C, Carpentier W, Dina C, Delaneau O, et al. Genomewide association study of an AIDS-nonprogression cohort emphasizes the role played by HLA genes (ANRS Genomewide Association Study 02). J Infect Dis. 2009 Feb 1;199(3):419–26.

37. Limou S, Coulonges C, Herbeck JT, van Manen D, An P, Le Clerc S, et al. Multiple-cohort genetic association study reveals CXCR6 as a new chemokine receptor involved in long-term nonprogression to AIDS. J Infect Dis. 2010 Sep 15;202(6):908–15.

38. Dalmasso C, Carpentier W, Meyer L, Rouzioux C, Goujard C, Chaix M-L, et al. Distinct genetic loci control plasma HIV-RNA and cellular HIV-DNA levels in HIV-1 infection: the ANRS Genome Wide Association 01 study. PloS One. 2008;3(12):e3907.

39. McLaren PJ, Pulit SL, Gurdasani D, Bartha I, Shea PR, Pomilla C, et al. Evaluating the Impact of Functional Genetic Variation on HIV-1 Control. J Infect Dis. 2017 27;216(9):1063–9.

40. Schoeni-Affolter F, Ledergerber B, Rickenbach M, Rudin C, Günthard HF, Telenti A, et al. Cohort Profile: The Swiss HIV Cohort Study. Int J Epidemiol. 2010 Oct 1;39(5):1179–89.

41. Loh P-R, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, Finucane HK, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016 Nov;48(11):1443–8.

42. Durbin R. Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). Bioinformatics. 2014 May 1;30(9):1266–72.

43. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet [Internet]. 2016 Aug 22 [cited 2016 Aug 31];advance online publication. Available from: http://www.nature.com/ng/journal/vaop/ncurrent/full/ng.3643.html

44. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006 Aug;38(8):904–9.

45. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. Nature. 2010 Sep;467(7311):52–8.

46. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015 Dec 1;4(1):1–16.

47. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A Tool for Genome-wide Complex Trait Analysis. Am J Hum Genet. 2011 Jan 7;88(1):76–82.

48. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. Nat Genet. 2014 Feb;46(2):100–6.

49. Jia X, Han B, Onengut-Gumuscu S, Chen W-M, Concannon PJ, Rich SS, et al. Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens. PLOS ONE. 2013 Jun 6;8(6):e64683.

50. Fellay J, Shianna KV, Ge D, Colombo S, Ledergerber B, Weale M, et al. A Whole-Genome Association Study of Major Determinants for Host Control of HIV-1. Science. 2007 Aug 17;317(5840):944–7.

51. McLaren PJ, Coulonges C, Bartha I, Lenz TL, Deutsch AJ, Bashirova A, et al. Polymorphisms of large effect explain the majority of the host genetic contribution to variation of HIV-1 virus load. Proc Natl Acad Sci. 2015 Nov 24;112(47):14658–63.

52. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinforma Oxf Engl. 2009 Jul 15;25(14):1754–60.

53. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep 1;20(9):1297–303.

54. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012 Jun;6(2):80–92.

55. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet. 2012 Aug 10;91(2):224–37.

56. Dilthey AT, Gourraud P-A, Mentzer AJ, Cereb N, Iqbal Z, McVean G. High-Accuracy HLA Type Inference from Whole-Genome Sequencing Data Using Population Reference Graphs. PLOS Comput Biol. 2016 Oct 28;12(10):e1005151.

57. Fan Y, Song Y-Q. PyHLA: tests for the association between HLA alleles and diseases. BMC Bioinformatics. 2017 Feb 6;18(1):90.

58. Packer JS, Maxwell EK, O'Dushlaine C, Lopez AE, Dewey FE, Chernomorsky R, et al. CLAMMS: a scalable algorithm for calling common and rare copy number variants from exome sequencing data. Bioinformatics. 2016 Jan 1;32(1):133–5.

59. Bhattacharya S, Dunn P, Thomas CG, Smith B, Schaefer H, Chen J, et al. ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. Sci Data. 2018 27;5:180015.

60. Purcell S, Cherny SS, Sham PC. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics. 2003 Jan 1;19(1):149–50.

61. Gutierrez-Arcelus M, Teslovich N, Mola AR, Polidoro RB, Nathan A, Kim H, et al. Lymphocyte innateness defined by transcriptional states reflects a balance between proliferation and effector functions. Nat Commun. 2019 Feb 8;10(1):1–15.

62. Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. Int J Methods Psychiatr Res. 2018;27(2):e1608.

63. Jones LE, Perelson AS. Transient viremia, plasma viral load, and reservoir replenishment in HIV-infected patients on antiretroviral therapy. J Acquir Immune Defic Syndr 1999. 2007 Aug 15;45(5):483–93.

64. Fletcher CV, Staskus K, Wietgrefe SW, Rothenberger M, Reilly C, Chipman JG, et al. Persistent HIV-1 replication is associated with lower antiretroviral drug concentrations in lymphatic tissues. Proc Natl Acad Sci U S A. 2014 Feb 11;111(6):2307–12.

65. Lorenzo-Redondo R, Fryer HR, Bedford T, Kim E-Y, Archer J, Pond SLK, et al. Persistent HIV-1 replication maintains the tissue reservoir during therapy. Nature. 2016 Feb 4;530(7588):51–6.

66. Podsadecki TJ, Vrijens BC, Tousset EP, Rode RA, Hanna GJ. Decreased adherence to antiretroviral therapy observed prior to transient human immunodeficiency virus type 1 viremia. J Infect Dis. 2007 Dec 15;196(12):1773–8.

67. Young J, Rickenbach M, Calmy A, Bernasconi E, Staehelin C, Schmid P, et al. Transient detectable viremia and the risk of viral rebound in patients from the Swiss HIV Cohort Study. BMC Infect Dis. 2015 Sep 21;15:382.

68. Simonetti FR, Sobolewski MD, Fyne E, Shao W, Spindler J, Hattori J, et al. Clonally expanded CD4+ T cells can produce infectious HIV-1 in vivo. Proc Natl Acad Sci U S A. 2016 Feb 16;113(7):1883–8.

69. Wang Z, Gurule EE, Brennan TP, Gerold JM, Kwon KJ, Hosmane NN, et al. Expanded cellular clones carrying replication-competent HIV-1 persist, wax, and wane. Proc Natl Acad Sci U S A. 2018 13;115(11):E2575–84.

70. Lee GQ, Orlova-Fink N, Einkauf K, Chowdhury FZ, Sun X, Harrington S, et al. Clonal expansion of genome-intact HIV-1 in functionally polarized Th1 CD4+ T cells. J Clin Invest. 2017 Jun 30;127(7):2689–96.

71. Nardacci R, Amendola A, Ciccosanti F, Corazzari M, Esposito V, Vlassi C, et al. Autophagy plays an important role in the containment of HIV-1 in nonprogressor-infected patients. Autophagy. 2014 Jul;10(7):1167–78.

72. Gaebler C, Lorenzi JCC, Oliveira TY, Nogueira L, Ramos V, Lu C-L, et al. Combination of quadruplex qPCR and next-generation sequencing for qualitative and quantitative analysis of the HIV-1 latent reservoir. J Exp Med. 2019 Oct 7;216(10):2253–64.

73. Braun DL, Kouyos R, Oberle C, Grube C, Joos B, Fellay J, et al. A novel Acute Retroviral Syndrome Severity Score predicts the key surrogate markers for HIV-1 disease progression. PloS One. 2014;9(12):e114111.

Chapter 3 Genetic variation near *CXCL12* is associated with susceptibility to HIV-related non-Hodgkin lymphoma

Christian W. Thorball[1,2], Tiphaine Oudot-Mellakh[3], Christian Hammer[4,5], Federico A. Santoni[6], Jonathan Niay[3], Dominique Costagliola[7], Cécile Goujard[8,9], Laurence Meyer[10], Sophia S. Wang[11], Shehnaz K. Hussain[12], Ioannis Theodorou[3], Matthias Cavassini[13], Andri Rauch[14], Manuel Battegay[15], Matthias Hoffmann[16], Patrick Schmid[17], Enos Bernasconi[18], Huldrych F. Günthard[19,20], Paul J. McLaren[21,22], Charles S. Rabkin[23], Caroline Besson[23-25,27], Jacques Fellay[1,2,26,27]

[1]School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland; [2]Swiss Institute of Bioinformatics, Lausanne, Switzerland; [3]Centre de génétique moléculaire et chromosomique, GH La Pitié Salpêtrière, Paris, France; [4]Department of Cancer Immunology, Genentech, South San Francisco, CA, USA; [5]Department of Human Genetics, Genentech, South San Francisco, CA, USA; [6]Service of Endocrinology, Diabetology and Metabolism, Lausanne University Hospital, Lausanne, Switzerland; [7]Sorbonne Universités, INSERM, UPMC Université Paris 06, Institut Pierre Louis d'épidémiologie et de Santé Publique (IPLESP UMRS 1136), Paris, France; [8]Inserm, CESP, U1018, Paris-Sud University, Le Kremlin-Bicêtre, France; [9]Department of Internal Medicine, Bicêtre Hospital, AP-HP, Le Kremlin-Bicêtre, France; [10]INSERM U1018, Centre de recherche en Épidémiologie et Santé des Population, Paris-Sud University, Paris-Saclay University, Le Kremlin-Bicêtre, France; [11]Division of Health Analytics, City of Hope Beckman Research Institute and City of Hope Comprehensive Cancer Center, Duarte, CA, USA ; [12]Department of Medicine, Cedars-Sinai Medical Center, Los Angeles, CA, USA; [13]Service of Infectious Diseases, Lausanne University Hospital and University of Lausanne, 1015, Lausanne, Switzerland; [14]Department of Infectious Diseases, Bern University Hospital, University of Bern, Switzerland; [15]Department of Infectious Diseases and Hospital Epidemiology, University Hospital Basel, University of Basel, 4031, Basel, Switzerland; [16]Division of Infectious Diseases and Hospital Epidemiology, Kantonsspital Olten, Switzerland; [17]Division of Infectious Diseases, Cantonal Hospital of St. Gallen, 9007, St. Gallen, Switzerland; [18]Division of Infectious Diseases, Regional Hospital of Lugano, 6900, Lugano, Switzerland; [19]Department of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, 8091, Zurich, Switzerland; [20]Institute of Medical Virology, University of Zurich, 8057, Zurich, Switzerland; [21]JC Wilt Infectious Diseases Research Centre, National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg, Canada; [22]Department of Medical Microbiology and Infectious Diseases, University of Manitoba, Winnipeg, Canada; [23]Infections and Immunoepidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA; [24]CESP, UVSQ, INSERM, Université Paris-Saclay, 94805, Villejuif, France; [25]Department of Hematology and Oncology, Hospital of Versailles, 78150 Le Chesnay, France; [26]Precision Medicine Unit, Lausanne University Hospital (CHUV) and University of Lausanne, Lausanne, Switzerland.

[27]J.F. and C.B. jointly directed this work.

Contribution to the study: I came up with the idea for the study and wrote the initial grant that permitted the genotyping of samples in the Swiss HIV Cohort. I performed all the analyses beginning from raw genotyping data to final results for all samples across all three cohorts and I wrote the paper.

This work has been submitted and is available as preprint on ***medRxiv*** (15.11.2019)

3.1 Abstract

Human immunodeficiency virus (HIV) infection is associated with an increased risk of non-Hodgkin lymphoma (NHL). Even in the era of suppressive antiretroviral treatment, HIV-infected individuals remain at higher risk of developing NHL compared to the general population. To identify potential genetic risk loci, we performed case-control genome-wide association studies and a meta-analysis across three cohorts of HIV+ patients of European ancestry, including a total of 278 cases and 1924 matched controls. We observed a significant association with NHL susceptibility in the C-X-C motif chemokine ligand 12 (CXCL12) region on chromosome 10. A fine mapping analysis identified rs7919208 as the most likely causal variant (P = 4.77e-11), with the G>A polymorphism creating a new transcription factor binding site for BATF and JUND. These results suggest a modulatory role of CXCL12 regulation in the increased susceptibility to NHL observed in the HIV-infected population.

3.2 Introduction

Human immunodeficiency virus (HIV) infection is associated with a markedly increased risk of several types of cancer compared to the general population.[1–3] This elevated cancer risk can be attributed partly to viral-induced immunodeficiency, frequent co-infections with oncogenic viruses (e.g., Epstein-Barr virus (EBV), hepatitis B and hepatitis C viruses, human herpesvirus 8 (HHV-8) and papillomavirus), and increased prevalence of traditional risk factors such as smoking.[4,5] However, all of these risk factors may not entirely explain the excess cancer burden seen in the HIV+ population.[6]

A previous study performed in the Swiss HIV Cohort Study (SHCS) identified two AIDS-defining cancers, Kaposi sarcoma and non-Hodgkin lymphoma (NHL) as the main types of cancer found among HIV positive patients (NHL representing 34% of all identified cancers).[4] The relative risk of developing NHL in HIV patients was highly elevated compared to the general population (period-standardized incidence ratio (SIR) = 76.4).[4] High HIV plasma viral load, absence of antiretroviral therapy (ART) as well as low CD4+ T cell counts are known predictive factors for NHL.[7,8] The introduction of ART into clinical practice has led to improved overall survival and restoration of immunity by decreasing viral load and increasing CD4+ T cell counts, and has led to a decreased risk of developing NHL. However, the risk remains substantially elevated compared to the general population (SIR = 9.1 (8.3–10.1))[9] and NHL still represents 20% of all cancers in people living with HIV in the ART era.[10] NHLs associated with HIV are predominantly aggressive B-cell lymphomas. Although they are heterogeneous, they share several pathogenic mechanisms involving chronic antigen stimulation, impaired immune response, cytokine deregulation and reactivation of the oncogenic viruses EBV and HHV-8.[11]

The emergence of genome-wide approaches in human genomics has led to the discovery of many associations between common genetic polymorphisms and susceptibility to several diseases including HIV infection and multiple types of cancer.[12,13] Recent genome-wide association studies (GWAS) of NHL have identified multiple susceptibility loci in the European population.[14–22] These variants are located in the genes *LPXN*[21]*, BTNL2*[23]*, EXOC2*, *NCOA1*[14]*, PVT1*[14,22], *CXCR5, ETS1, LPP,* and *BCL2*[22] for various subtypes of NHL, as well as *BCL6* in the Chinese population.[24] Strong associations with variation in human leukocyte antigen (HLA) genes have also been reported.[15,18,22] However, in the setting of HIV infection, no genome-wide analysis has been reported concerning the occurrence of NHL and the specific mechanisms driving their development remain largely unknown.

Here we report the results of the first genome-wide analysis of NHL susceptibility in individuals chronically infected with HIV. We combined three HIV cohort studies from France, Switzerland and the USA and searched for associations between >6 million single nucleotide polymorphisms (SNPs) and a diagnosis of NHL. We identified a novel genetic locus near *CXCL12* as associated with the development of NHL among HIV+ individuals.

3.3 Results

3.3.1 Study participants and association testing

To identify human genetic determinants of HIV-associated NHL, we performed case-control GWAS in three groups of HIV+ patients of European ancestry (SHCS, ANRS and MACS). The characteristics of the study participants are presented in Table 1. In total, genotyping data were obtained for 278 cases (NHL+/HIV+) and 1924 matched controls (NHL-/HIV+). With this sample size, we had 80% power to detect a common genetic variant (10% minor allele frequency) with a relative risk of 2.5, assuming an additive genetic model and using Bonferroni correction for multiple testing ($P_{threshold}$ = 5e-8).[25]

Table 3.1. Summary of included samples and studies

Cohort	Cases	Controls	Lambda	Genotyping chips	Years of NHL diagnosis	Control inclusion criteria
SHCS *Age (median)* *Sex (%Male)*	145 *61* *91%*	1090 *58* *80%*	1.00	Illumina HumanOmniExpress-24, Human1M, Human610, HumanHap550, HumanCore-12	2000 - 2017	HIV < 2005, no cancer diagnosis as of 2017 & matched with age
ANRS *Age (median)* *Sex (%Male)*	61 *50* *89%*	562 *34* *87%*	1.00	Illumina Human Omni5 Exome 4v 1-2, Illumina 300	2008 - 2015	No cancer diagnosis
MACS *Age (median)* *Sex (%Male)*	72 *69* *100%*	272 *68* *100%*	1.01	Illumina 1MV1, Human1M-Duo, HumanHap550	1985 - 2013	Matched to cases in terms of age, treatment & time of infection

After genome-wide imputation and quality control, 6.2 million common variants were tested for association with the development of NHL using linear mixed models including sex as a covariate. Results were combined across cohorts using a weighted Z-score-based meta-analysis (Figure 3.1A). The genomic inflation factor (lambda) was in all cases very close to 1 [1.00–1.01], indicating an absence of systematic inflation of the association results (Figure 3.1B; Supplementary Figure 3.2).

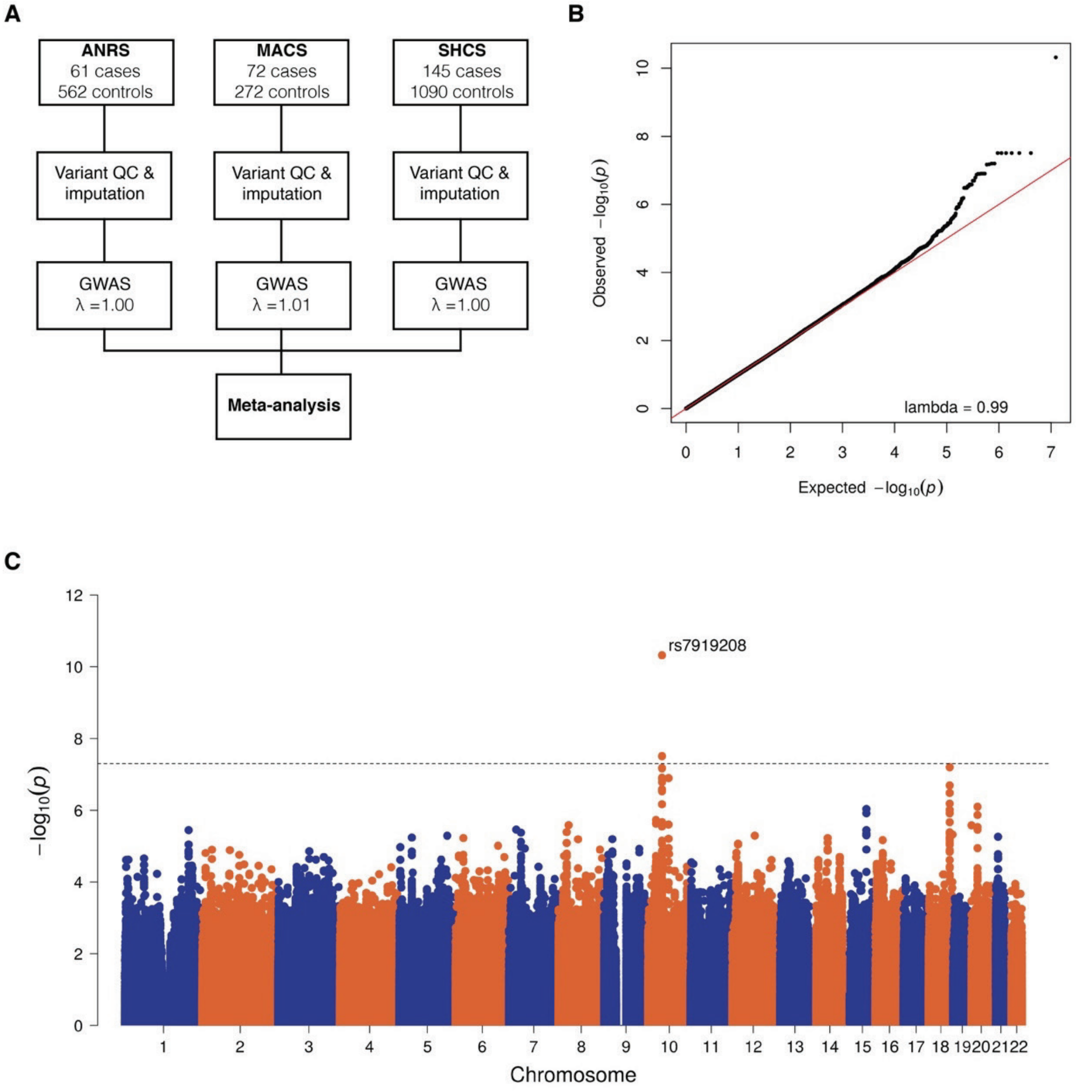


Figure 3.1. Genome-wide association analysis. (A) Schematic of analysis pipeline. (B) Quantile-quantile plot of the observed -log10(p-value) (black dots, y-axis) versus expected -log10(p-values) under the null hypothesis (red line) to check for any genomic inflation of the observed p-values. No genomic inflation is observed, with the genomic inflation factor lambda = 0.99. (C) Manhattan plot of all obtained p-values for each variant included in the meta-analysis. The genome-wide threshold (P = 5e-8) for significance is marked by a dotted line. Only variants at the CXCL12 locus were found to be significant.

3.3.2 Association results

We observed significant associations with the development of HIV-related NHL at a single locus on chromosome 10, downstream of *CXCL12* (Figure 3.1C). A total of 7 SNPs in this locus had p-values lower than the genome-wide significance threshold (P < 5e-8), with rs7919208 displaying the strongest association (Table 3.2). This association was only detected in the SHCS and ANRS cohorts and not among MACS study participants (Supplementary Table 3.1).

Table 3.2. Significant association with HIV-related NHL

Chr	**Pos**	**SNP**	**Ref**	**Alt**	**P**	**OR**
10	44673557	rs7919208	A	G	4.77e-11	1.23
10	44677967	rs149399290	T	C	3.09e-08	1.20
10	44678218	rs17155463	T	A	3.09e-08	1.20
10	44678262	rs17155474	C	T	3.09e-08	1.20
10	44678454	rs17155478	T	C	3.09e-08	1.20
10	44678898	rs12249837	G	A	3.09e-08	1.20
10	44680902	rs10608969	T	TAAAGA	3.09e-08	1.20

Variants significantly associated with HIV-related NHL in a weighted Z-score-based meta-analysis of all individuals included in the SHCS, ANRS and MACS cohorts. Odds ratios (OR) were transformed from betas using the formula OR = exp(beta).

3.3.3 Fine mapping of the CXCL12 locus

To identify the causal variant(s) among associated SNPs and determine their potential functional effects, we used a multi-level fine mapping approach, combining the statistical fine mapping tool PAINTOR to obtain a 99% credible set and the deep learning framework DeepSEA to predict any effects on chromatin marks and transcription factor binding these variants may have.

Using PAINTOR, we identified a single variant, rs7919208, having a high posterior probability (= 100%) of being causal among the 99% credible set based on the integration of the association results, LD structure and enrichment of genomic features in this locus (Figure 3.2).

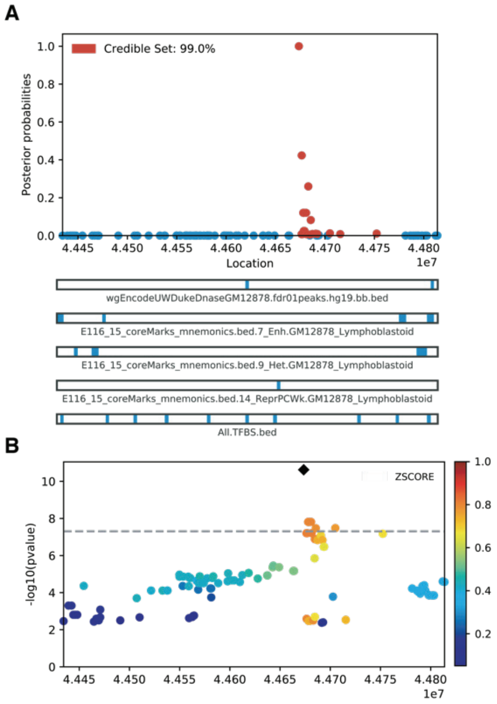


Figure 3.2. Fine mapping of genome-wide significant hits with PAINTOR. (A) The 99% credible set and posterior probabilities of being the causal variant. The genomic positions are listed on the x-axis. Bottom tracks represent DNAase and chromatin marks obtained from GM12878 cells as well as TFBS from the Roadmap Epigenomics Project and ENCODE in the region. (B) Locus plot of the associated variants, highlighting the LD relationship, based on the SHCS cohort. The top variant rs7919208 is marked by a black diamond.

Consistent with the PAINTOR result, DeepSEA also identified rs7919208 as the sole variant, among the 99% credible set, predicted to have a functional impact by significantly increasing the probability of binding by the B cell transcription factors BATF (log2 fold-change = 3.27) and JUND (log2 fold-change = 2.91) (Supplementary Table 3.2). Further analysis of the genomic sequence surrounding rs7919208 and the JASPAR transcription factor binding site (TFBS) motifs for BATF and JUND revealed that rs7919208 G->A polymorphism creates the TFBS motif required for the binding of these transcription factors (Figure 3.3A).

3.3.4 Long-range chromatin interactions

To assess the potential functional links between the TFBS created in the presence of the minor allele of rs7919208 and the nearby genes, we performed an analysis of promoter capture Hi-C data and topologically associating domains (TADs). We used the well-characterized GM12878 lymphoblastoid cell line produced by EBV transformation of B lymphocytes collected from a female European donor as model.

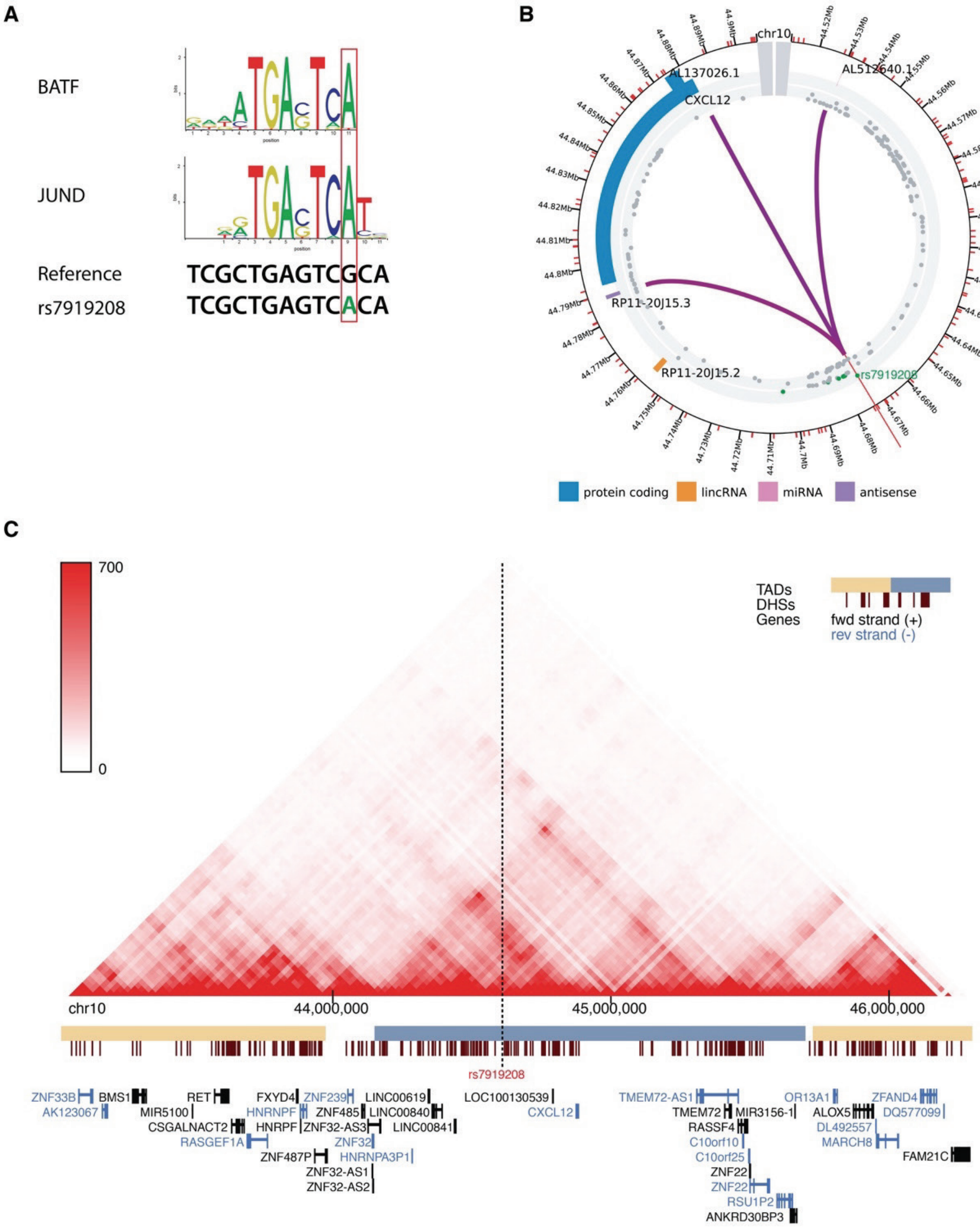


Figure 3.3. Novel transcription factor binding site and long-range interactions. (A) Canonical motifs of BATF and JUND with the underlying genomic reference sequence and the nucleotide change caused by rs7919208. (B) Promoter capture Hi-C analysis in the GM12878 cell line of the region with the predicted causal variant and CXCL12. Variants and their level of association in the meta-analysis are marked in the inner grey circle. Genome-wide significant variants are colored green. Purple lines indicate significant interactions between promoter and other genomic regions. (C) TADs in the GM12878 cell line in the region of CXCL12. The yellow and blue boxes indicate the called TADs from the Hi-C contact map above. The plot is centered on rs7919208.

First, to examine the interaction potential of the rs7919208 region with nearby promoters, we analyzed available promoter capture Hi-C data obtained from the GM12878 cell line. This analysis revealed a significant interaction between the rs7919208 region and the *CXCL12* promoter, suggesting a possible modulating impact of rs7919208 on the transcription of that gene (Figure 3.3B). Second, to further validate this observed genomic interaction, we analyzed available TAD calls from GM12878 cells[26], using the 3D Genome Browser for visualization[27] (Figure 3.3C). We observed that rs7919208 is located within a large TAD together with *CXCL12*, signifying the interaction potential of the new TFBS at rs7919208 and *CXCL12*.

3.3.5 Transcriptomic effects of rs7919208

We did not observe any association between rs7919208 and mRNA expression levels of *CXCL12* in peripheral blood or PBMCs from multiple publicly available datasets, including GTEx (v7)[28], GEUVADIS[29] and the Milieu Intérieur Consortium[30] (Supplementary Figure 3.3). Of note, *CXCL12* expression levels were very low in all datasets.

HIV infection causes many profound transcriptomic changes.[31] Thus, in order to examine the effect of rs7919208 on *CXCL12* in the context of HIV infection, we extracted RNA from PBMCs of 452 individuals in the SHCS with available genotyping data and sequenced them using the Bulk RNA Barcoding and sequencing (BRB-seq) approach.[32] However, the expression levels of *CXCL12* were below the limit of detection for most individuals, preventing an eQTL analysis.

3.3.6 No replication of susceptibility loci found in the general population

To assess whether the genetic contribution to the risk of developing NHL is similar or distinct in the HIV+ population compared to the general population, we extracted the p-values of all variants found to be genome-wide significant in previous GWAS performed in the general population[14,21–24,33] and compared them to our results. We did not replicate any of the previously published genome-wide associated variants, even at nominal significance level ($P < 0.05$), despite sufficient statistical power for many of the variants, thus indicating that the genetic susceptibility of NHL is distinct between the HIV+ and the general population (Supplementary Table 3.3). To further examine this possibility, we tested whether the NHL/HIV+ associated variant rs7919208 is associated with an increased risk of NHL in the general population. We performed a series of case/control GWAS of four NHL subtypes (CLL, DLBCL, FL and MZL) as well as a combined GWAS with all NHL subtypes (Supplementary Table 3.4; Supplementary Figure 3.4) and assessed the association evidence at rs7919208. We found no association between rs7919208 and any of the subtypes in the general population, even at nominal significance.

3.4 Discussion

In this genome-wide analysis, including a total of 278 NHL HIV+ cases and 1924 HIV+ controls from three independent cohorts, we identified a novel NHL susceptibility locus on chromosome 10 near the *CXCL12* gene. The strong signal observed in the meta-analysis was driven by the associations detected in the SHCS and ANRS cohorts and there was no evidence of association in the MACS cohort. Notably, most NHL cases in the MACS cohort date back to the pre-ART era, while only NHL cases diagnosed after the year 2000 were included in the SHCS and ANRS analyses. Conceivably, NHL occurring in the early years of the HIV pandemic may have been primarily driven by severe immunosuppression, which could have obscured any influence of human genetic variation among the cases in the MACS sample. Precise phenotype definition is crucial in designing large-scale genetic studies since any environmental noise tends to decrease the likelihood of identifying potential genetic influences.

NHL is a relatively rare cancer even among HIV infected individuals, making it difficult to collect the large numbers of cases that would typically be included in contemporary genome-wide genetic studies. Indeed, a recent study from the Data Collection on Adverse events of Anti-HIV Drugs (D:A:D) group showed an NHL incidence rate of 1.17/1000 person-years of follow-up over the past 15 years (392 new cases in >40,000 HIV-infected individuals).[8] Still, we were able to obtain clinical and genetic data from a total of 278 patients with confirmed NHL diagnosis. By matching them with a larger number of controls from the same cohorts, we had enough power to identify associated variants of relatively large effects in the *CXCL12* region.

Several groups have already suggested a potential role for *CXCL12* variation in HIV-related NHLs. A prospective study correlated increased *CXCL12* expression with subsequent NHL development in HIV-infected children but not in uninfected children.[34] The number of A alleles at the CXCL12-3' variant (rs1801157) has also previously been associated with an increased risk of developing HIV-related NHL during an 11.7 year follow-up period.[35] Thus, our data further support the role of CXCL12 as a critical modulator of the individual risk of developing NHL in the HIV population.

The role of CXCL12 and its receptor chemokine receptor 4 (CXCR4) in cancer in the general population is well established, with the levels of *CXCL12* and *CXCR4* found to be increased in multiple types of cancer and to be associated with tumor progression.[36,37] Furthermore, *in vivo* inhibition of either CXCR4 or CXCL12 signaling is capable of disrupting early lymphoma development in severe combined immunodeficient (SCID) mice transfused with EBV+ PBMCs.[38] These results and others have already led to the development and testing of several small molecules targeting either CXCL12 or CXCR4 to inhibit tumor progression.[36]

We could not identify any significant relationship between rs7919208 and the expression levels of *CXCL12* in PBMCs or EBV transformed lymphocytes. This can be due to multiple factors such as the low expression levels of *CXCL12* in most tissues, aside from stromal cells, or that rs7919208 through creation of the BATF and JUND binding site represent an induced or dynamic eQTL. These types of eQTLs are often found in regions deprived of regulatory annotations, since these have been examined in static cell types.[39] HIV-induced overexpression of *BATF*[40] could also explain why rs7919208 is only a risk factor in the HIV population and not in the general population. Allele specific expression (ASE) analyses constitutes a novel method, with more power, which can uncover the effect of heterozygous variants on a given gene. In fact, recent ASE data from Pejman Mohammadi at the Scripps Research Institute within GTEx (v8) showed a significant positive effect of rs7919208 on the expression of *CXCL12* in fibroblasts (P = 0.0006), supporting our findings that this variant increases the expression of *CXCL12*. Furthermore, the fact that this signal was only observed in fibroblasts, the GTEx tissue most closely assembling stromal cells, underscores the clinical importance of these cells in the development of HIV-related NHL.

Previous analyses in the general population have discovered both shared and distinct associations for NHL subtypes.[14,21–24,33] However, similar analyses were not possible in our sample since NHL subtype information was not available for many of our cases. Furthermore, information on serostatus for relevant co-infections with EBV or other oncogenic viruses was not available and could therefore not be assessed. In particular, EBV has been largely associated with the development of NHL and other lymphomas and is considered a driver of a subset of NHLs in the general population.[41] Variants in the HLA region have consistently been associated with all NHL subtypes in HIV uninfected populations regardless of EBV serostatus. We did not find any evidence of HLA associations in our analyses of HIV-related NHL. This lack of replication of HLA variants and of all other previously identified risk variants from the general population suggests that distinct genes or pathways influence susceptibility to NHL in the HIV+ population compared to the general population.[42]

In summary, we have identified variants significantly associated with the development of NHL in the HIV population. Fine mapping of the associated locus and subsequent analyses of TADs, promoter capture Hi-C data as well as deep-learning models of mutational effects on transcription factor binding, points to a causative model involving the gain of a BATF and JUND transcription binding site downstream of *CXCL12* capable of physically interacting with the *CXCL12* promoter. These results suggest an important modulating role of CXCL12 in the development of HIV-related NHL.

3.5 Methods

3.5.1 Ethics statement

The Swiss HIV Cohort Study (SHCS), the Primo ANRS and ANRS CO16 Lymphovir cohorts (ANRS) and the Multicenter AIDS Cohort Study (MACS) cohorts have been approved by the competent ethics committees / institutional review boards of all participating institutions. A written informed consent, including consent for human genetic testing, was obtained from all study participants.

3.5.2 Study participants and contributing centers

Swiss HIV Cohort Study (SHCS)

The SHCS is a large, ongoing, multicenter cohort study of HIV-positive individuals that includes >70% of adult living with HIV in Switzerland. At follow-up visits every 6 months, demographic, clinical, laboratory, and ART information has been prospectively recorded since 1988.[43] Cancer diagnoses are verified thoroughly using checking charts including information on biopsies and imaging. To minimize potential treatment bias and population stratification, we only considered as cases patients diagnosed with NHL between 2000 and 2017 and of European ancestry, as determined by principal component analysis (PCA) (Supplemental Figure 3.1A). Controls were matched based on age, ancestry, CD4+ T cell counts and viral load results. To be eligible as controls, they also had to be diagnosed with HIV prior to 2005 and have no registered cancer diagnosis of any type as of 2017. Patients were genotyped using Illumina HumanOmniExpress-24 Beadchips, or genotypes were obtained in the context of a previous GWAS in the SHCS on various platforms including Illumina HumanCore-12, HumanHap550, Human610 and Human1M Beadchips.

French Primo ANRS and ANRS CO16 Lymphovir cohorts (ANRS)

The French ANRS CO16 lymphovir cohort of HIV related lymphomas enrolled adult patients at diagnosis of lymphoma in 32 centers between 2008 and 2015.[44] Pathological materials were centralized, and diagnoses of NHL were based on World Health Organization criteria. Patients were genotyped using Illumina Human Omni5 Exome 4v beadchips. Additional cases and controls were included from the ANRS PRIMO Cohort, which has been enrolling patients during primary HIV-1 infection in 95 French Hospitals since 1996.[45] Patients were genotyped using Illumina Sentrix Human Hap300 Beadchips. Only patients of European ancestry, as determined by PCA, were included in the study (Supplemental Figure 3.1B).

The Multicenter AIDS Cohort Study (MACS)

The MACS has enrolled gay and bisexual HIV infected men in 4 US cities since 1984. The NHL cases were predominately diagnosed prior to the year 2000. Data collected include demographic variables (age, race, ethnicity and HIV transmission category), CD4+ T cell count, HIV viral load and tumor histology. Eligible cases had a diagnosis of HIV-related NHL, available genotyping data and at least one CD4+ T cell count obtained within 2 years of the NHL diagnosis. Controls were matched on MACS study site, age at NHL diagnosis (+/- 2 years) and CD4+ T cell count at NHL diagnosis (within the following groups 0-99 / 100 -199 / 200-499 / >499 cells/µL). Patients were genotyped using Illumina HumanHap550 and Human1M Beadchips.[46] As in the other cohorts, only individuals of European ancestry were included, as determined by PCA (Supplementary Figure 3.1C).

3.5.3 Quality control and imputation of genotyping data

The genotyping data from each cohort was filtered and imputed in a similar way, with each genotyping array processed separately to minimize potential batch effects. All variants were first flipped to the correct strand orientation with BCFTOOLS (v1.8) using the human genome build GRCh37 as reference. Variants were removed if they had a larger than 20% minor allele frequency (MAF) deviation from the 1000 genomes phase 3 EUR reference panel or if they showed a larger than 10% MAF deviation between genotyping chips in the same cohort.

The QC filtered genotypes were phased with EAGLE2[47] and missing genotypes were imputed using PBWT[48] with the Sanger Imputation Service[49], taking the 1000 Genomes Project Phase 3 panel as reference. Only high-quality variants with an imputation score (INFO > 0.8) were retained for further analyses.

Genome-wide association testing and meta-analysis

To search for associations between human genomic variation and the development of HIV-related NHL, we first performed separate GWAS within each cohort (SHCS, ANRS and MACS) prior to combining the results in a meta-analysis.

For each cohort separately, the imputed variants were filtered out using PLINK (v2.00a2LM)[50] based on missingness (> 0.1), minor allele frequency (< 0.02) and deviation from Hardy-Weinberg Equilibrium (P_{HWE} < 1e-6). Determination of population structure and calculation of principal components was done using EIGENSTRAT (v6.1.4)[51] and the HapMap3 reference panel[52]. All individuals not clustering with the European HapMap3 samples were excluded from further analyses. The samples were screened using KING (v2.1.3)[53] to ensure no duplicate or cryptic related samples were included. Single-marker case-control association analyses were performed using linear mixed models, with genetic relationship matrices calculated between pairs of individuals according to the leave-one-chromosome-out principle, as implemented in GCTA mlmaloco (v1.91.4beta).[54,55] Sex was included as a covariate, except in the MACS cohort, which only includes men.

The results of the three GWAS were combined across cohorts using a weighted Z-score-based meta-analysis in PLINK (v1.90b5.4), after exclusion of the variants that were not present in all three cohorts.

3.5.4 Fine mapping of associated regions

Fine mapping of the *CXCL12* locus was performed using PAINTOR (v3.1)[56] to identify the most likely causal variant(s). All variants within 200kb of the top associated SNP and with a p-value below 0.005 were included in the model. The linkage disequilibrium (LD) matrix was created using PLINK and genotype data from the SHCS cohort. PAINTOR was first run against all genomic annotation databases provided with the software, including the FANTOM5, ENCODE and the Roadmap Epigenomics Project. For the final model, the top 5 annotations based on improvement to model fit and cell type relevance were selected to obtain the posterior probabilities and the 99% credible set of the variants most likely to be causal based on the association from Bayes' factors.

3.5.5 Predictive effect of potentially causal variants

The potential functional impact of the predicted causal variants was assessed using DeepSEA[57], a deep learning-based sequence model trained on available chromatin and transcription factor data from ENCODE and Roadmap Epigenomics. DeepSEA provides a functional significance score for each variant, which is a measure of the evolutionary conservation and the significance of the magnitude of the predicted chromatin effects. For the variants with a functional significance score of less than 0.01, we analyzed the predicted changes in specific chromatin modifications or transcription factor (TF) binding probabilities. Chromatin or TF binding changes with E-values below 0.001 and normalized probabilities of observing a binding event above 0.2 were considered relevant. The TF position weight matrices (PWMs) for TFs with a high probability of binding (normalized probability ≥ 50%) were obtained from the JASPAR CORE 5.0 database.[58]

3.5.6 Long-range chromatin interactions

Predicted topological associating domains (TADs) near the genome-wide significant locus in GM12878 lymphoblastoid cells were obtained from publicly available data[26] and visualized using the 3D Genome Browser.[27]

Potential interactions between the genome-wide significant locus and promoters of nearby genes were analyzed using publicly available promoter capture Hi-C data in GM12878 lymphoblastoid cells. The Hi-C data was processed through the CHiCAGO pipeline and visualized with CHiCP.[59,60] Interaction scores ≥ 5 were considered significant, as described previously.[61]

3.5.7 Expression quantitative trait loci (eQTL) analyses

The role of rs7919208 as an eQTL was examined in GEUVADIS[29] and in response to various pathogens, although not including HIV, in the Milieu Intérieur Consortium cohort.[30] Furthermore, eQTL information was also obtained from the GTEx (v7)[28] Portal on 03/22/2019.

Bulk RNA Barcoding and sequencing (BRB-seq)[32] was performed on RNA from peripheral blood mononuclear cells (PBMCs) of 452 individuals from the SHCS with available genotyping data.

3.5.8 Comparison to GWAS hits in the general population

An attempt at replicating variants previously associated with NHL in the general population was performed by extraction of the p-values of the SNPs reported to be associated in previous NHL GWAS. A variant was considered replicated if it had a nominally significant association p-value ($P < 0.05$) plus similar effect direction in the meta-analysis.

The effect of rs7919208 in the general population cohorts was assessed directly using the NIH database for Genotypes and Phenotypes (dbGaP) accession # phs000801 cohorts for chronic lymphocytic leukemia (CLL), DLBCL (Diffuse large B-cell lymphoma), FL (Follicular lymphoma) and MZL (Marginal zone lymphoma) and corresponding controls.[14,22,23,62] The genotype data was imputed, processed and analyzed using the same pipeline and methods as described above for the HIV cohorts, with duplicate samples identified and removed using KING and including age and sex as covariates.

3.5.9 Statistical analyses

All statistical analyses were performed using the R statistical software (v3.3.3), unless otherwise specified.

3.5.10 Data sharing statement

Full summary statistics will be made available in the GWAS catalog (https://www.ebi.ac.uk/gwas) upon publication. The raw genotype data can be obtained through the respective cohorts.

3.6 Declarations

3.6.1 Acknowledgments

This study has been financed within the framework of the Swiss HIV Cohort Study, supported by the Swiss National Science Foundation (grant #177499), by SHCS project #789 and by the SHCS research foundation. The data are gathered by the Five Swiss University Hospitals, two Cantonal Hospitals, 15 affiliated hospitals and 36 private physicians (listed in http://www.shcs.ch/180-health-care-providers).

This work further benefited from the ANRS funding of both the Primo and Lymphovir cohorts.

Foundation Monahan and Fulbright funded the stay of CB at the National Cancer Institute (NCI).

The Genome-Wide Association Study (GWAS) of Non-Hodgkin Lymphoma (NHL) project was supported by the intramural program of the Division of Cancer Epidemiology and Genetics (DCEG), National Cancer Institute (NCI), National Institutes of Health (NIH). The datasets have been accessed through the NIH database for Genotypes and Phenotypes (dbGaP) under accession # phs000801. A full list of acknowledgements can be found in the supplementary note (Berndt SI et al., Nature Genet., 2013, PMID: 23770605).

3.6.2 Author contributions

C.W.T., J.F., P.J.M., C.S.R., C.B., C.H. and T.O.M. contributed to the conception and design of the study. C.W.T., J.F., P.J.M., F.A.S., D.C., L.M., C.G., I.T., S.K.H., M.C., A.R., M.B., M.H., P.S., E.B., H.F.G., C.S.R. and C.B. contributed to the acquisition of data. C.W.T., T.O.M., C.H., F.A.S., C.B., C.S.R. and J.F. contributed to the analysis and interpretation of data. C.W.T., J.F., C.S.R., C.B. and S.W. contributed to the drafting the article and revising it critically for important intellectual content.

All authors critically reviewed and approved the final manuscript.

3.6.3 Competing interests

Conflict of Interest Disclosure: Christian Hammer is a full-time employee of F. Hoffmann–La Roche/Genentech. The remaining authors declare no competing financial interests.

3.7 Supplementary tables and figures

Supplementary Table 3.1. Cohort level association statistics for genome-wide significant variants in the meta-analysis

SNP	**ANRS BETA**	**ANRS P**	**SHCS BETA**	**SHCS P**	**MACS BETA**	**MACS P**
rs7919208	0.32	2.78E-10	0.12	2.66E-03	-0.06	0.57
rs149399290	0.28	7.93E-07	0.13	2.30E-03	-0.03	0.79
rs17155463	0.28	7.93E-07	0.13	2.30E-03	-0.03	0.79
rs17155474	0.28	7.93E-07	0.13	2.30E-03	-0.03	0.79
rs17155478	0.28	7.93E-07	0.13	2.30E-03	-0.03	0.79
rs12249837	0.28	7.93E-07	0.13	2.30E-03	-0.03	0.79
rs10608969	0.28	7.93E-07	0.13	2.30E-03	-0.03	0.79

The analyses were performed using linear mixed models with GCTA within each cohort. No associations were seen in the MACS cohort. P-values and beta values are presented to show the level and direction of the association.

Supplementary Table 3.2. Top predicted changes associated with rs7919208 allelic variation in GM12878

Transcription factor	Effect (Log2fold change)	E-value	Normalized Prob. (Reference)	Normalized Prob. (Alternative)
BATF	3.27	0.00004	0.13	0.60
JUND	2.91	0.00009	0.12	0.50
MEF2A	2.06	0.00026	0.09	0.28
MEF2C	1.98	0.00022	0.08	0.26
BCL11A	2.19	0.00040	0.07	0.25
P300	1.74	0.00055	0.08	0.22
IRF4	2.12	0.00033	0.06	0.21

Significant changes induced by rs7919208 as predicted by DeepSEA for transcription factors with a normalized probability (Prob.) above 0.20 in the GM12878 lymphoblastoid cell line. The E-value is the expected proportion of variants with a larger predicted effect between the reference and alternative allele for a certain chromatin feature based on predicted effects calculated for variants in The 1000 Genomes Project.

Supplementary Table 3.3. Comparisons with genome-wide significant variants identified in GWAS of NHL in the general population

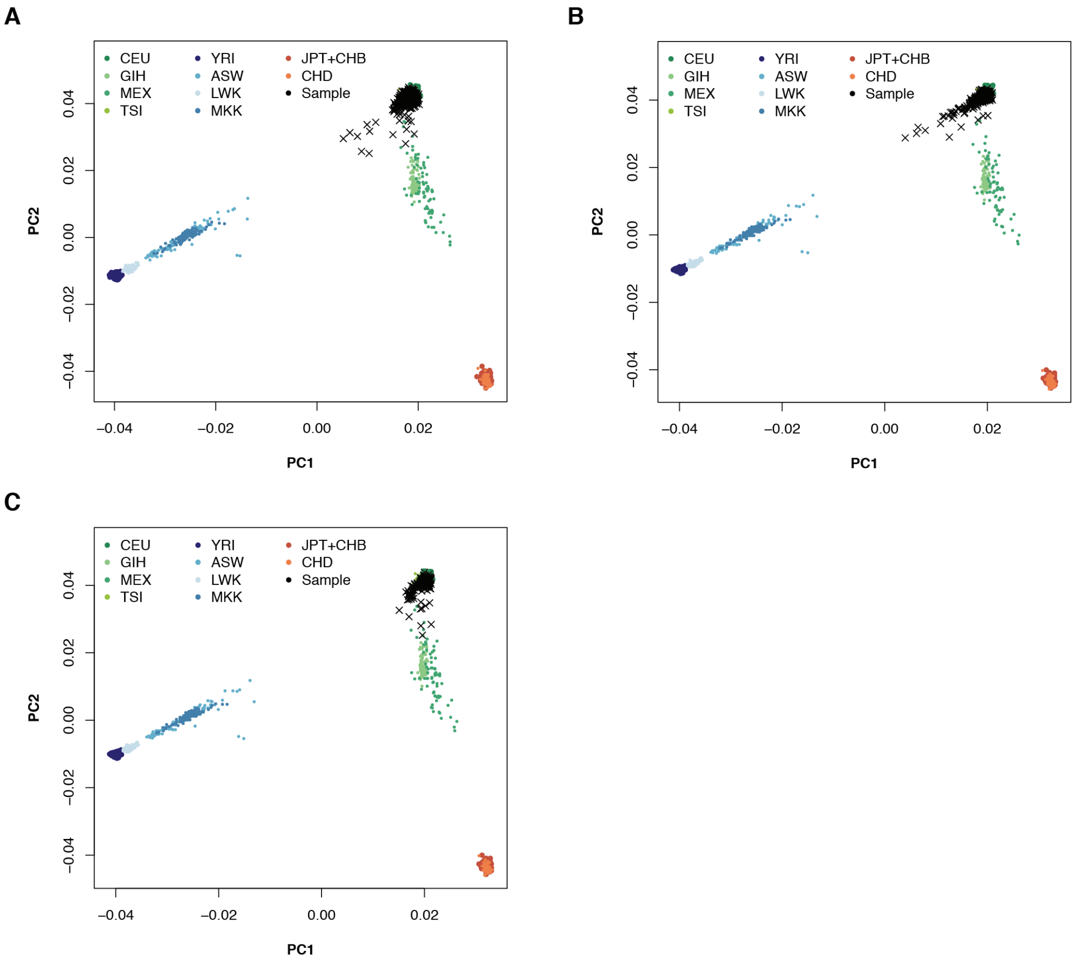
SNP	Gene	Publication	Subtype	META P	SHCS P	ANRS P	MACS P	OR	Mean MAF	POWER (P<0.05)	POP
rs116446171	EXOC2	Cerhan et al.[14]	DLBCL	NA	NA	NA	0.08	2.20	0.01	58%	EUR
rs12195582	HLA region	Skibola et al.[22]	FL	0.33	0.47	0.51	NA	1.78	0.41	100%	EUR
rs12289961	LPXN	Vijai et al. (2013)[21]	DLBCL+FL	0.88	0.40	0.14	0.46	1.29	0.22	74%	EUR
rs13254990	PVT1	Skibola et al.[22]	FL	0.75	NA	0.58	0.82	1.18	0.32	48%	EUR
rs13255292	PVT1	Cerhan et al.[14]	DLBCL	0.83	NA	0.72	0.90	1.22	0.32	62%	EUR
rs17203612	HLA class II	Skibola et al.[22]	FL	NA	0.82	NA	NA	1.44	0.38	98%	EUR
rs17749561	BCL2	Skibola et al.[22]	FL	0.74	NA	0.72	0.98	1.34	0.10	61%	EUR
rs2523607	HLA-B	Cerhan et al.[14]	DLBCL	0.50	0.30	0.77	NA	1.32	0.08	49%	EUR
rs2922994	HLA-B	Vijai et al. (2015)[23]	MZL	0.55	0.30	0.66	NA	1.64	0.08	93%	EUR
rs3130437	HLA class I	Skibola et al.[22]	FL	0.33	0.47	NA	0.46	1.23	0.38	68%	EUR
rs4733601	PVT1	Cerhan et al.[14]	DLBCL	0.29	NA	0.38	0.57	1.18	0.49	51%	EUR
rs4937362	ETS1	Skibola et al.[22]	FL	0.62	0.61	0.90	NA	1.17	0.45	47%	EUR
rs4938573	CXCR5	Skibola et al.[22]	FL	0.60	0.25	0.72	0.18	1.34	0.19	81%	EUR
rs6444305	LPP	Skibola et al.[22]	FL	NA	NA	NA	NA	1.21	NA	NA	EUR
rs6457327	HLA	Lim et al.[53]	DLBCL+FL	0.66	0.80	0.85	0.18	1.30	0.35	85%	EUR
rs6773854	BCL6	Tan et al.[22]	DLBCL	0.30	0.51	0.73	0.07	1.44	0.21	95%	CHN
rs79480871	NCOA1	Cerhan et al.[14]	DLBCL	NA	NA	NA	NA	1.34	0.08	54%	EUR
rs9461741	BTNL2	Vijai et al. (2015)[23]	MZL	0.97	0.72	0.67	NA	2.66	0.03	99%	EUR

Comparisons with genome-wide significant variants identified in published GWAS of NHL in the general population. The NHL subtypes includes DLBCL (Diffuse large B-cell lymphoma), FL (Follicular lymphoma) and MZL (Marginal zone lymphoma). P-values for the HIV meta-analysis and the individual cohort GWAS are shown per variant. The statistical power to replicate the published variants under an additive model at $P < 0.05$, given their published odds ratios (OR) and the mean observed minor allele frequencies (MAF) in the HIV cohorts is also listed. Most of the published GWAS was on European (EUR) patients and Chinese (CHN).

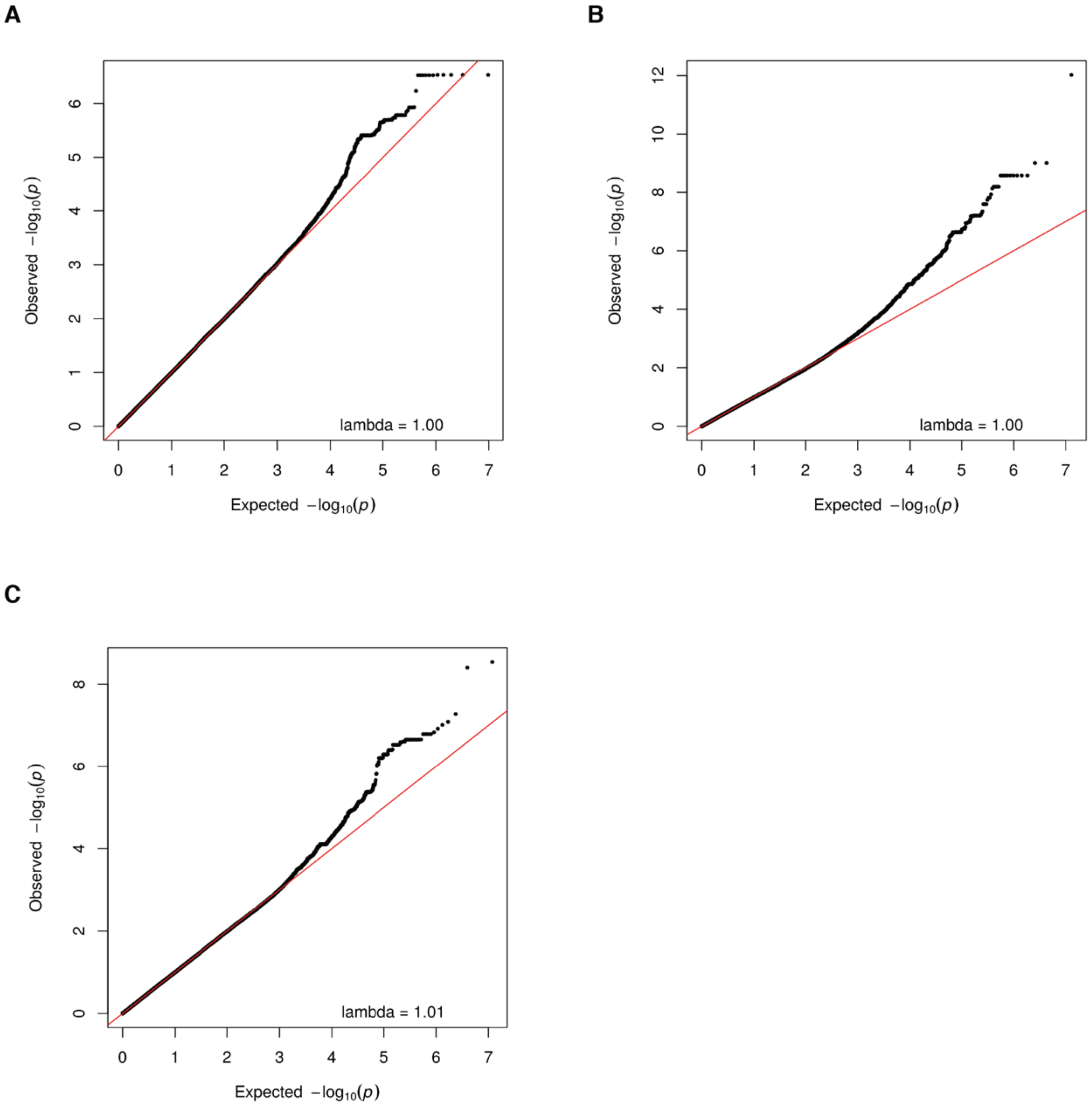
Supplementary Table 3.4. The effect of rs7919208 in GWAS in the general population

Subtype	Cases	Controls	Lambda	P (rs7919208)	OR (rs7919208)
CLL	1033	2635	1.04	0.29	0.97
DLBCL	2173	2635	1.04	0.65	1.01
FL	1753	2635	1.03	0.36	0.97
MZL	617	2635	1.01	0.56	0.98
Combined	5556	2635	1.04	0.60	0.99

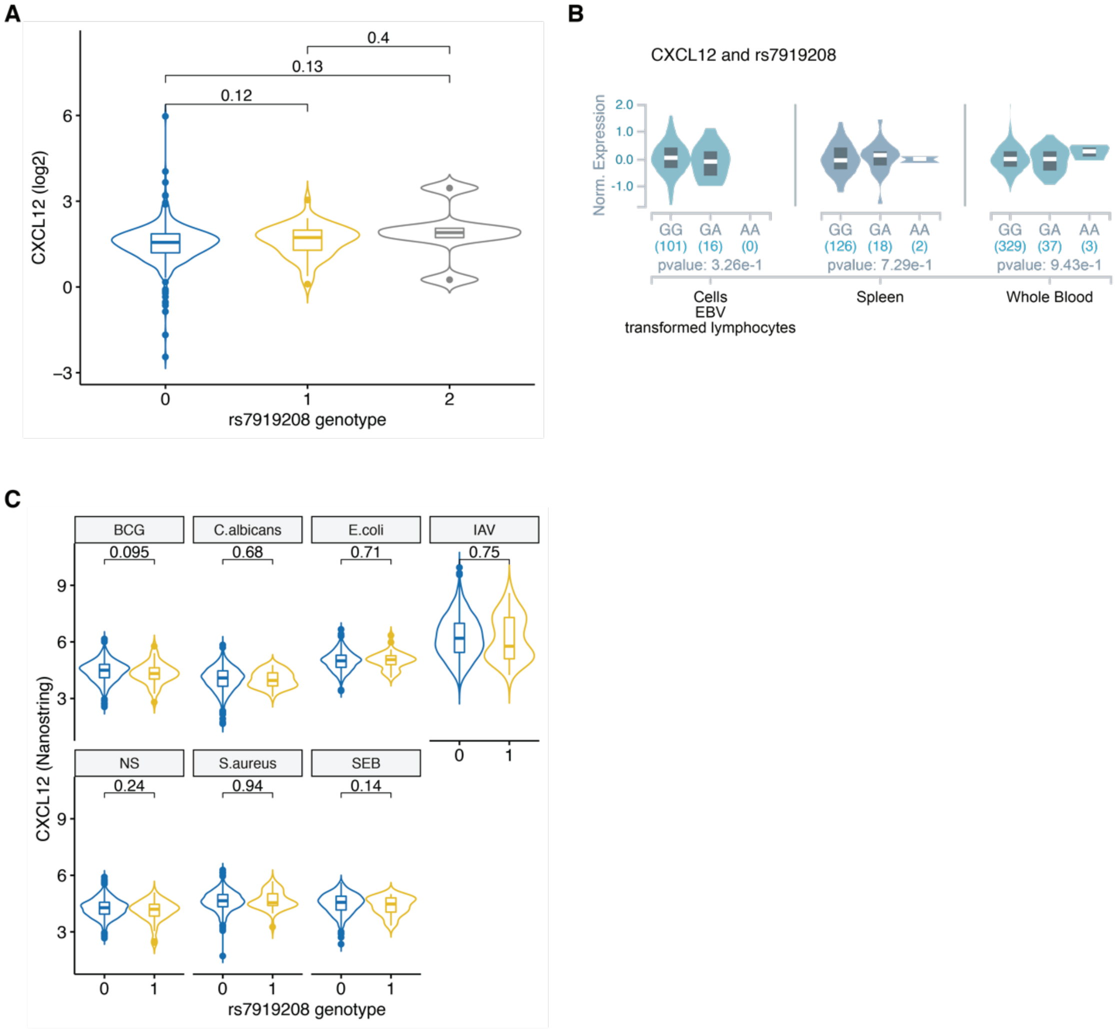
The association of rs7919208 in the general (non-HIV) population across NHL subtypes. Lambda indicates the genome-wide inflation factor for the GWAS performed for each subtype to ensure the test-statistics observed are valid. The calculated p-values and odds ratios for rs7919208 are listed for each GWAS.



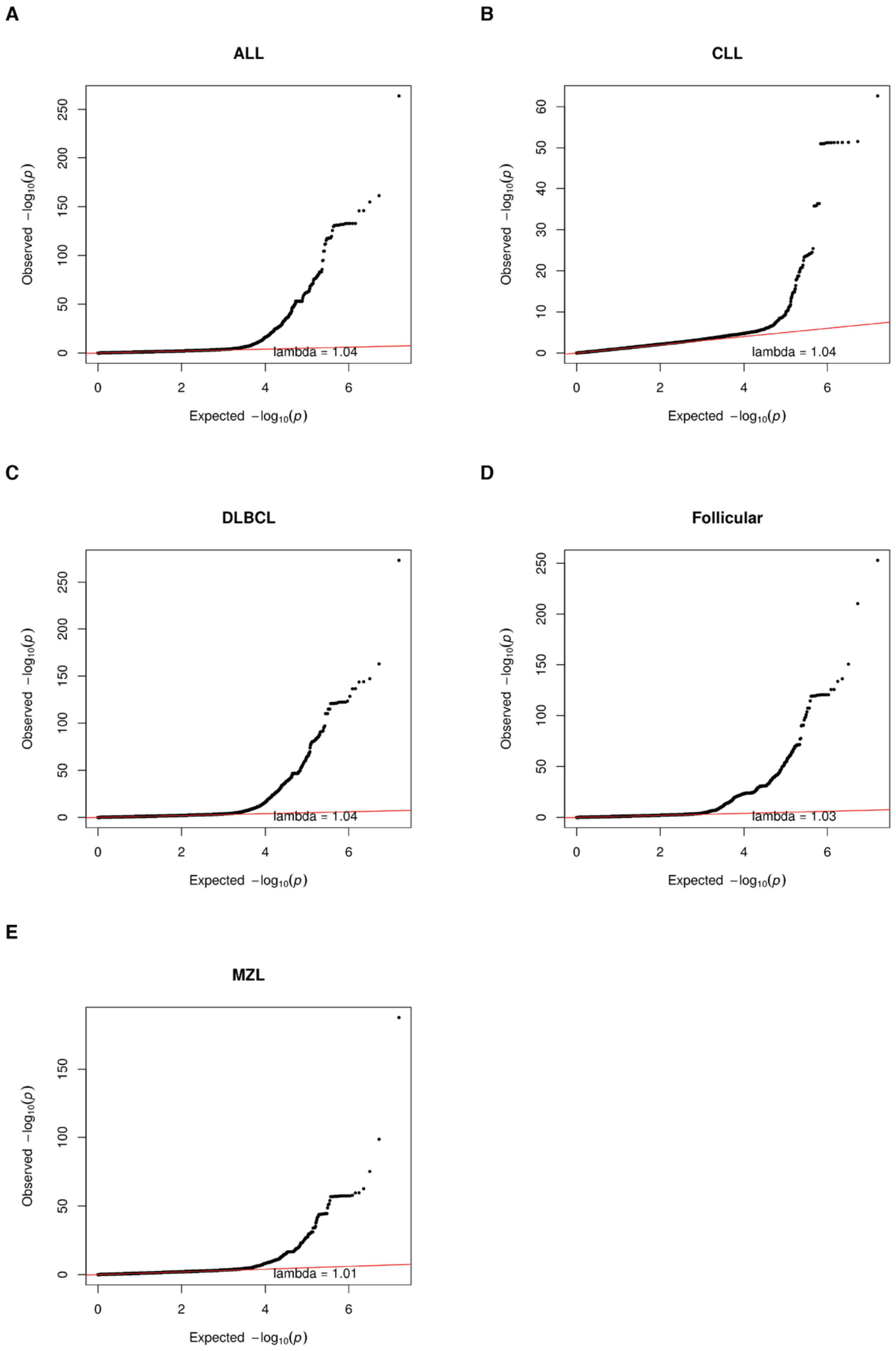
Supplementary Figure 3.1. Principal component analyses (PCA) with the HapMap project. The black crosses represent individuals genotyped and included in this study. Individuals of European ancestry colocalizes with the HapMap reference samples from CEU (Northern Europeans from Utah) and TSI (Tuscans from Italy). (A) The SHCS cohort. (B) The ANRS cohort. (C) The MACS cohort.



Supplementary Figure 3.2. Quantile-quantile plots for the initial cohort level GWAS. Lambda indicates the genome-wide inflation factor. Values ~1 denotes the lack of genomic inflation due to confounding factors. (A) Plot for the GWAS in the SHCS cohort. (B) Plot or the GWAS in the ANRS cohort. (C) Plot for the GWAS in the ANRS cohort.



Supplementary Figure 3.3. eQTL information on rs7919208. (A) Expression levels of CXCL12 in EBV transformed lymphocytes from the GEUVADIS consortium according to the rs7919208 genotype. Differences between genotype groups were tested using Wilcoxon rank-sum tests with the obtained p-values shown on the figures. (B) Relationship between CXCL12 and the rs7919208 genotype in GTEx across EBV transformed lymphocytes, Spleen and Whole Blood. (C) Relationship between CXCL12 expression and rs1919208 using Nanostring in the Milieu Interieur Consortium for stimulated and non-stimulated (NS) PBMCs. Stimulants used were Mycobacterium bovis (BCG), Candida albicans, Escherichia coli, Influenza A virus (IAV), Staphylococcus aureus and Staphylococcal enterotoxin B (SEB).



Supplementary Figure 3.4. Quantile-quantile plots for the general population NHL GWAS. Lambda indicates the genome-wide inflation factor. Values ~1 denotes the lack of genomic inflation due to confounding factors. (A) Plot for the GWAS of all NHL subtypes combined. (B) Plot for GWAS of chronic lymphocytic leukemia (CLL). (C) Plot for Diffuse large B-cell lymphoma (DLBCL). (D) Plot for Follicular lymphoma. (E) Plot for Marginal zone lymphoma (MZL).

3.8 References

1. Patel, P. *et al.* Incidence of Types of Cancer among HIV-Infected Persons Compared with the General Population in the United States, 1992–2003. *Ann Intern Med* **148**, 728–736 (2008).

2. Vogel, M. *et al.* Cancer risk in HIV-infected individuals on HAART is largely attributed to oncogenic infections and state of immunocompetence. *European Journal of Medical Research* **16**, 101 (2011).

3. Robbins, H. A. *et al.* Excess Cancers Among HIV-Infected People in the United States. *J. Natl. Cancer Inst.* **107**, (2015).

4. Clifford, G. M. *et al.* Cancer Risk in the Swiss HIV Cohort Study: Associations With Immunodeficiency, Smoking, and Highly Active Antiretroviral Therapy. *JNCI Journal of the National Cancer Institute* **97**, 425–432 (2005).

5. Engels, E. A. Non-AIDS-defining malignancies in HIV-infected persons: etiologic puzzles, epidemiologic perils, prevention opportunities. *AIDS* **23**, 875–885 (2009).

6. Borges, Á. H., Dubrow, R. & Silverberg, M. J. Factors contributing to risk for cancer among HIV-infected individuals, and evidence that earlier combination antiretroviral therapy will alter this risk: *Current Opinion in HIV and AIDS* **9**, 34–40 (2014).

7. Guiguet, M. *et al.* Effect of immunodeficiency, HIV viral load, and antiretroviral therapy on the risk of individual malignancies (FHDH-ANRS CO4): a prospective cohort study. *The Lancet Oncology* **10**, 1152–1159 (2009).

8. Shepherd, L. *et al.* Differences in Virological and Immunological Risk Factors for Non-Hodgkin and Hodgkin Lymphoma. *J Natl Cancer Inst* **110**, 598–607 (2018).

9. Hleyhel, M. *et al.* Risk of AIDS-Defining Cancers Among HIV-1–Infected Patients in France Between 1992 and 2009: Results From the FHDH-ANRS CO4 Cohort. *Clinical Infectious Diseases* **57**, 1638–1647 (2013).

10. Robbins, H. A. *et al.* Excess Cancers Among HIV-Infected People in the United States. *JNCI: Journal of the National Cancer Institute* **107**, (2015).

11. Swerdlow, S. H. *WHO classification of tumours of haematopoietic and lymphoid tissues*. (International Agency for Research on Cancer, 2017).

12. McLaren, P. J. & Carrington, M. The impact of host genetic variation on infection with HIV-1. *Nat Immunol* **16**, 577–583 (2015).

13. Sud, A., Kinnersley, B. & Houlston, R. S. Genome-wide association studies of cancer: current insights and future perspectives. *Nature Reviews Cancer* **17**, 692–704 (2017).

14. Cerhan, J. R. *et al.* Genome-wide association study identifies multiple susceptibility loci for diffuse large B cell lymphoma. *Nature Genetics* **46**, 1233–1238 (2014).

15. Conde, L. *et al.* Genome-wide association study of follicular lymphoma identifies a risk locus at 6p21.32. *Nat Genet* **42**, 661–664 (2010).

16. Frampton, M. *et al.* Variation at 3p24.1 and 6q23.3 influences the risk of Hodgkin's lymphoma. *Nature Communications* **4**, (2013).

17. Kumar, V. *et al.* Common variants on 14q32 and 13q12 are associated with DLBCL susceptibility. *J Hum Genet* **56**, 436–439 (2011).

18. Moutsianas, L. *et al.* Multiple Hodgkin lymphoma–associated loci within the HLA region at chromosome 6p21.3. *Blood* **118**, 670–674 (2011).

19. Skibola, C. F. *et al.* Genetic variants at 6p21.33 are associated with susceptibility to follicular lymphoma. *Nat Genet* **41**, 873–875 (2009).

20. Urayama, K. Y. *et al.* Genome-Wide Association Study of Classical Hodgkin Lymphoma and Epstein–Barr Virus Status–Defined Subgroups. *JNCI J Natl Cancer Inst* **104**, 240–253 (2012).

21. Vijai, J. *et al.* Susceptibility Loci Associated with Specific and Shared Subtypes of Lymphoid Malignancies. *PLoS Genetics* **9**, e1003220 (2013).

22. Skibola, C. F. *et al.* Genome-wide Association Study Identifies Five Susceptibility Loci for Follicular Lymphoma outside the HLA Region. *The American Journal of Human Genetics* **95**, 462–471 (2014).

23. Vijai, J. *et al.* A genome-wide association study of marginal zone lymphoma shows association to the HLA region. *Nat Commun* **6**, (2015).

24. Tan, D. E. K. *et al.* Genome-wide association study of B cell non-Hodgkin lymphoma identifies 3q27 as a susceptibility locus in the Chinese population. *Nature Genetics* **45**, 804–807 (2013).

25. Johnson, J. L. & Abecasis, G. R. GAS Power Calculator: web-based power calculator for genetic association studies. *bioRxiv* 164343 (2017) doi:10.1101/164343.

26. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665–1680 (2014).

27. Wang, Y. *et al.* The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *Genome Biology* **19**, 151 (2018).

28. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

29. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).

30. Piasecka, B. *et al.* Distinctive roles of age, sex, and genetics in shaping transcriptional variation of human immune responses to microbial challenges. *PNAS* **115**, E488–E497 (2018).

31. Mohammadi, P. *et al.* 24 Hours in the Life of HIV-1 in a T Cell Line. *PLOS Pathogens* **9**, e1003161 (2013).

32. Alpern, D. *et al.* BRB-seq: ultra-affordable high-throughput transcriptomics enabled by bulk RNA barcoding and sequencing. *Genome Biology* **20**, 71 (2019).

33. Lim, U. *et al.* Pleiotropy of Cancer Susceptibility Variants on the Risk of Non-Hodgkin Lymphoma: The PAGE Consortium. *PLoS ONE* **9**, e89791 (2014).

34. Sei, S. *et al.* Increased Level of Stromal Cell-Derived Factor-1 mRNA in Peripheral Blood Mononuclear Cells from Children with AIDS-related Lymphoma. *Cancer Res* **61**, 5028–5037 (2001).

35. Rabkin, C. S. *et al.* Chemokine and Chemokine Receptor Gene Variants and Risk of Non-Hodgkin's Lymphoma in Human Immunodeficiency Virus-1–Infected Individuals. *Blood* **93**, 1838–1842 (1999).

36. Meng, W., Xue, S. & Chen, Y. The role of CXCL12 in tumor microenvironment. *Gene* **641**, 105–110 (2018).

37. Peled, A., Klein, S., Beider, K., Burger, J. A. & Abraham, M. Role of CXCL12 and CXCR4 in the pathogenesis of hematological malignancies. *Cytokine* **109**, 11–16 (2018).

38. Piovan, E. *et al.* Chemokine receptor expression in EBV-associated lymphoproliferation in hu/SCID mice: implications for CXCL12/CXCR4 axis in lymphoma generation. *Blood* **105**, 931–939 (2005).

39. Strober, B. J. *et al.* Dynamic genetic regulation of gene expression during cellular differentiation. *Science* **364**, 1287–1290 (2019).

40. Quigley, M. *et al.* Transcriptional analysis of HIV-specific CD8+ T cells shows that PD-1 inhibits T cell function by upregulating BATF. *Nat. Med.* **16**, 1147–1151 (2010).

41. Gasser, O. *et al.* HIV Patients Developing Primary CNS Lymphoma Lack EBV-Specific CD4þ T Cell Function Irrespective of Absolute CD4þ T Cell Counts. *PLoS Medicine* **4**, 6 (2007).

42. SH, S. *et al. WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues*.

43. Schoeni-Affolter, F. *et al.* Cohort Profile: The Swiss HIV Cohort Study. *Int J Epidemiol* **39**, 1179–1189 (2010).

44. Besson, C. *et al.* Outcomes for HIV-associated diffuse large B-cell lymphoma in the modern combined antiretroviral therapy era. *AIDS* **31**, 2493 (2017).

45. Dalmasso, C. *et al.* Distinct Genetic Loci Control Plasma HIV-RNA and Cellular HIV-DNA Levels in HIV-1 Infection: The ANRS Genome Wide Association 01 Study. *PLoS ONE* **3**, e3907 (2008).

46. Fellay, J. *et al.* Common Genetic Variation and the Control of HIV-1 in Humans. *PLOS Genetics* **5**, e1000791 (2009).

47. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics* **48**, 1443–1448 (2016).

48. Durbin, R. Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). *Bioinformatics* **30**, 1266–1272 (2014).

49. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **advance online publication**, (2016).

50. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 1–16 (2015).

51. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

52. The International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).

53. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).

54. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A Tool for Genome-wide Complex Trait Analysis. *The American Journal of Human Genetics* **88**, 76–82 (2011).

55. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet* **46**, 100–106 (2014).

56. Kichaev, G. *et al.* Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics* **33**, 248–255 (2017).

57. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods* **12**, 931–934 (2015).

58. Khan, A. *et al.* JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res* **46**, D260–D266 (2018).

59. Schofield, E. C. *et al.* CHiCP: a web-based tool for the integrative and interactive visualization of promoter capture Hi-C datasets. *Bioinformatics* **32**, 2511–2513 (2016).

60. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics* **47**, 598–606 (2015).

61. Cairns, J. *et al.* CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol.* **17**, 127 (2016).

62. Berndt, S. I. *et al.* Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia. *Nature Genetics* **45**, 868–876 (2013).

Chapter 4 Contribution of genetic background and clinical D:A:D risk score to chronic kidney disease in Swiss HIV-positive persons with normal baseline estimated glomerular filtration rate

Léna G., Dietrich[1*], Catalina, Barceló[2*], **Christian W. Thorball[3,13*]**, Lene, Ryom[4], Felix, Burkhalter[5], Barbara, Hasse[6], Hansjakob, Furrer[7], Maja, Weisser[8], Ana, Steffen[9], Enos, Bernasconi[10], Matthias, Cavassini[11], Sophie, de Seigneux[12], Chantal, Csajka[2], Jacques, Fellay[3,13], Bruno Ledergerber[6], Philip E., Tarr[1] and the Swiss HIV Cohort Study

** these authors contributed equally to the manuscript*

[1]University Department of Medicine and Infectious Diseases Service, Kantonsspital Baselland, University of Basel, Bruderholz, Switzerland; [2] Division of Clinical Pharmacology, Centre Hospitalier Universitaire Vaudois, University of Lausanne, Switzerland; [3] Swiss Institute of Bioinformatics, Lausanne, Switzerland; [4]CHIP, Center of excellence for Health, immunity and infections, Dept. of Infectious Diseases, Rigshospitalet, University of Copenhagen, Denmark; [5] University Department of Medicine and Nephrology Service, Kantonsspital Baselland, University of Basel, Bruderholz, Switzerland; [6] Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, University of Zurich; [7] Department of Infectious Diseases, Bern University Hospital, University of Bern, Switzerland; [8] Division of Infectious Diseases and Hospital Epidemiology, University Hospital Basel, Basel, Switzerland; [9] Division of Infectious Diseases, Kantonsspital St. Gallen, Switzerland; [10] Division of Infectious Diseases, Ospedale Regionale, Lugano, Switzerland; [11] Division of Infectious Diseases, Lausanne University Hospital, Lausanne, Switzerland; [12] Division of Nephrology, Geneva University Hospitals and Faculty of Medicine, Geneva, Switzerland; [13] School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Switzerland

Contribution to the study: I came up with the idea for the type of genetic risk score to use, performed all genetic related analyses (genetic risk score and GWAS) including quality control of the data. I wrote the paper together with Lêna G. Dietrich, Catalina Barcelo and Philip Tarr.

This work has been published in ***Clinical Infectious Diseases*** (06.04.2019)

4.1 Abstract

Background: In HIV, the relative contribution of genetic background, clinical risk factors, and antiretrovirals to chronic kidney disease (CKD) is unknown.

Methods: We applied a case-control design and performed genome-wide genotyping in white Swiss HIV Cohort participants with normal baseline estimated glomerular filtration rate (eGFR >90 mL/min/1.73 m2). Uni- and multivariable CKD odds ratios (OR) were calculated based on the D:A:D score that summarizes clinical CKD risk factors and a polygenic risk score that summarizes genetic information from 86613 single nucleotide polymorphisms..

Results: We included 743 cases (79% male; median age, 42 years; baseline eGFR 106 mL/min/1.73 m2) with confirmed eGFR drop to <60 mL/min/1.73 m2 (n=144) or >25% eGFR drop to <90 mL/min/1.73 m2 (n=599), and 322 controls (eGFR drop <15%; 81% male; median age, 39 years, baseline eGFR 107 mL/min/1.73 m2). Polygenic risk score and D:A:D score contributed to CKD. In multivariable analysis, CKD ORs were 2.13 (95% confidence interval, 1.55-2.97) in participants in the 4th (most unfavorable) vs. 1st (most favorable) genetic score quartile; 1.94 (1.37-2.65) in the 4th vs. 1st D:A:D score quartile; and 2.98 (2.02-4.66), 1.70 (1.29-2.29), and 1.83 (1.45-2.40), per 5-years exposure to atazanavir/ritonavir, lopinavir/ritonavir, and tenofovir disoproxil fumarate, respectively. Participants in the 1st genetic score quartile had no increased CKD risk, even if they were in the 4th D:A:D score quartile.

Conclusions: Genetic score increased CKD risk similar to clinical D:A:D score and potentially nephrotoxic antiretrovirals. Irrespective of D:A:D score, individuals with the most favorable genetic background may be protected against CKD.

4.2 Introduction

Chronic kidney disease (CKD) is a major long-term concern in HIV-positive persons.[1-4] The D:A:D study, the largest consortium of observational HIV studies with rigorous endpoint ascertainment and validation, has documented clinical and HIV-related risk factors for CKD, which can be summarized in a 9-item risk score.[5] HIV-positive persons with low, medium and high risk D:A:D score had a 1:393, 1:47, and 1:6 chance of developing CKD over 5 years.[5] In addition, the D:A:D study described atazanavir/ritonavir (ATV/r), lopinavir/ritonavir (LPV/r), and tenofovir disoproxil fumarate (TDF) as being associated with an increased CKD incidence rate in HIV-positive persons with normal kidney function at baseline.[6]

CKD has a strong hereditary component.[7-9] Genetic studies of CKD in HIV have focused on HIV-associated nephropathy (HIVAN) which develops predominantly in persons of African ancestry with untreated HIV infection, and there is a strong association with *APOL1* gene variants.[10,11] Candidate gene studies have suggested an association of e.g. *ABCC2* polymorphisms and TDF-associated kidney dysfunction in HIV, but were limited by the assessment of single or few gene variants only, by their cross-sectional design, and small study populations.[12,13]

Genome-wide association studies (GWAS) have now identified >50 common genetic variants that reproducibly contribute to CKD in the general population.[7-9] The aim of the present study was therefore to quantitate the contribution of genome-wide genetic variation to CKD in HIV-positive participants. Analyzed in the context of clinical risk factors (summarized in the D:A:D score) and potentially nephrotoxic antiretroviral drugs, we hypothesized that genetic background may partially explain CKD risk in HIV. Our study represents the most comprehensive genetics–CKD evaluation undertaken to date in HIV-positive persons.

4.3 Methods

4.3.1 Study population

Eligible participants included HIV-positive persons enrolled in the Swiss HIV Cohort Study (www.shcs.ch), with ≥3 months follow-up after 1.1.2004. The study was approved by the respective local ethics committees. Participants provided written informed consent for genetic testing. Baseline was defined as first estimated glomerular filtration rate (eGFR) measured after 1.1.2004. CKD cases included participants with normal baseline eGFR (>90 mL/min/1.73 m^2; using the CKD-EPI formula) who developed a CKD event during follow-up, as defined in the D:A:D study[6] and in the renal subproject of the START trial, i.e. eGFR drop to <60 mL/min/1.73 m^2, confirmed over a ≥90 day period. Because only 1% of D:A:D study participants with normal baseline eGFR later experienced an eGFR drop to <60 mL/min/1.73 m^2 [6]), we also included participants who developed mild CKD, defined as >25% eGFR drop to <90 mL/min/1.73 m^2, confirmed over a ≥90 day period. To better separate the phenotypes of cases and controls, and thereby to increase power to detect genetic effects, [14,15] only participants with ≤15% eGFR drop at last SHCS follow-up were eligible as controls. Only controls with GWAS genotyping data already available were included. Because previous CKD GWAS in the general population were conducted in populations of predominantly European descent,[7-9] the study was restricted to participants of European descent.

4.3.2 Case-control matching

We performed 1:1 matching. The last available eGFR measurement of controls had to be after the CKD event date of the corresponding case. Matching was done using incidence density sampling, ^{11} i.e. controls were required to have the first available eGFR measurement +/- 1 year of the corresponding case. In other words, controls were matched on similar follow-up *duration*, and their observation *period* was at similar calendar times, in an effort to correct for differences in potentially nephrotoxic ART compounds in use at different times and other differences during the study period. More specifically, cases and controls were put in random order and cases were sequentially matched to a control, which was then removed from the list of potential controls. While baseline characteristics of cases at the time of their first available eGFR determination remained unchanged, controls could be matched to different cases in the 2000 re-sampled datasets. As a consequence, their baseline characteristics (i.e. eGFR measurement within +/- 1 year of the first available eGFR of the respective case, and last eGFR measurement after the CKD event date of the corresponding case) could vary. Therefore, we present averaged values of the time-varying baseline characteristics of controls.

Since we had more cases than controls, only a subset of cases was successfully matched, and we therefore repeated the matching process 2000 times with random re-sampling from cases and controls.[16] This bootstrap resampling method yielded effect estimates (CKD odds ratio) for both D:A:D score and genetic score with appropriately narrow confidence intervals (**Supplementary Figure 4.1**).

4.3.3 Genotyping and Quality Control

DNA samples obtained from peripheral blood mononuclear cells were genotyped with the Infinium CoreExome-24 BeadChip (Illumina, San Diego, CA), or in the context of previous GWAS in the SHCS.

SHCS control participants were previously genome-wide genotyped using various platforms including the HumanCore-12, HumanHap550, Human610, Human1M and HumanOmniExpress-24 BeadChips (Illumina, San Diego, CA). Each cohort was filtered and imputed separately, with variants first flipped to the correct strand with BCFTOOLS (v1.8) according to the human GRCh37 reference genome and filtered based on a <20% deviation from the 1000 genomes phase 3 EUR reference panel. Genotypes were phased and missing genotypes were imputed with EAGLE2 and PBWT respectively,[28,29] using the 1000 Genomes Project Phase 3 reference panel on the Sanger Imputation Service.[30] Study participants were filtered based on European ancestry, while imputed variants were filtered by minor allele frequency (>1%), missingness (>10%), deviation from Hardy-Weinberg equilibrium (P < 1e-6) and an imputation quality score (INFO>0.8). The filtered genotypes were then combined using PLINK (v1.90b5) prior to analyses.[31]

4.3.4 Non-genetic CKD risk factors

Only variables included in the D:A:D score[5] were used, i.e. mode of HIV transmission, hepatitis C co-infection, age, baseline eGFR, gender, CD4 nadir, hypertension, prior cardiovascular disease, and diabetes mellitus. Each antiretroviral agent is recorded with start and stop dates in the SHCS database. We adjusted only for those ART exposures that contributed to CKD in patients with normal baseline eGFR in the D:A:D study,[6] i.e. cumulative exposure to ATV/r, LPV/r, and TDF. Hypertension was defined as blood pressure ≥140/90 mmHg or use of antihypertensive medication. Diabetes mellitus was diagnosed with confirmed plasma glucose >7.0 mmol/L (fasting) or >11.1 (non-fasting), or use of antidiabetic medication.

4.3.5 Genome-wide Polygenic Risk Score

The effect estimate for each SNP included in the polygenic risk score ("genetic score") was obtained from the summary statistics in a recent genetic meta-analysis reference paper of eGFR.[7] The genetic score was calculated with PRSice (v1.1.3b),[17] using p-value thresholding to identify the best model, because including common variants of smaller effect sizes in addition to only the genome-wide significant variants has been shown to increase the predictive power of genetic risk scores. [18-20] The final genetic score model included 86'813 independent SNPs after clumping.[7]

Prior to matching of the imputed SNPs with the SNPs included in the reference paper, the imputed SNPs were filtered by missingness (>10%), minor allele frequency (>1%), Hardy-Weinberg equilibrium (P <1e-6) and also clumped (r^2=0.8), to control for SNPs in linkage disequilibrium.

We excluded future case patients from potentially serving as controls until the CKD event date,[32] because it makes no sense that individual genetic background could both be permissive for and protective against CKD. Also, extending on previous non-genetic studies,[28,29] we avoided re-using control patients for multiple cases because, according to Robins,[33] this would cause biased results if cases themselves do not serve as controls until CKD diagnosis.

4.3.6 Statistical analyses

Univariable and multivariable conditional logistic regression analyses were used to estimate associations of the different quartiles of the genetic and D:A:D scores with CKD events for each of the 2000 case-control sets. The indicators of quartiles and not the scores themselves were included in the models. In multivariable analyses, we included the cumulative exposure to ATV/r, LPV/r, and TDF per 5 years use until the event date among cases, or, for controls, up to the CKD event date of the corresponding case. To assess any potential effect modification of the D:A:D score by the genetic score, we added a model with an interaction term between genetic and D:A:D scores. The average odds ratio was then calculated as the antilog of the mean of the 2000 log-transformed odds ratios, and the 95% confidence interval was based on the 2.5 and 97.5 percentiles. We used Stata/SE 15.1 (StataCorp, College Station, TX, USA).

4.3.7 Sensitivity analyses

To capture the genetic effect in subgroups of participants who develop different degrees of kidney impairment, we performed sensitivity analyses, defining CKD as either (i) eGFR drop to <60 mL/min/1.73 m^2; (ii) eGFR drop >25% to <70 mL/min/1.73 m^2 ; (iii) or as eGFR drop to <60 mL/min/1.73 m^2 OR of $\geq$40%. In further sensitivity analyses, we excluded participants treated with; (i) dolutegravir, (ii) any integrase inhibitor, (iii) cobicistat, and (iv) rilpivirine, because these ART agents can increase serum creatinine (eGFR) without changing the actual GFR [21][22]. To quantify the potential bias introduced by the imbalance of matching frequencies we added a sensitivity analysis in which cases and controls were weighted with the inverse probability of being sampled, i.e. participants who were sampled less often were attributed more weight.

4.3.8 Exploratory genome-wide association analysis and analysis of previously published candidate SNP

In an exploratory GWAS, we separately tested all genotyped or imputed SNPs on the genetic arrays for association with CKD. We also attempted to replicate previously published associations between candidate SNPs (**Supplementary Table 4.1**) and CKD by extraction of the p-values from the exploratory GWAS. A SNP was considered replicated if found nominally significant (P<0.05).

The SNPs included in exploratory GWAS were filtered as for the genetic score by missingness (>10%), minor allele frequency (>1%), Hardy-Weinberg equilibrium (P <1e-6). The associations were computed using linear mixed models with genetic relationship matrixes calculated between pairs of individuals after the leave-one-chromosome-out principal as implemented in GCTA mlma-loco (v1.91.4beta) with age as covariable.[34,35] The genomic inflation factor, lambda, was 1.02.

4.4 Results

4.4.1 Participants, CKD events

We included 743 cases with confirmed eGFR drop to <60 mL/min/1.73 m^2 (n=144) or eGFR drop >25% to <90 mL/min/1.73 m^2 (n=599). We included 335 controls with eGFR drop of <15% during the observation period, of whom 322 were successfully matched to a case. All cases were matched 377-2000 (out of 2000) times, with a median (IQR) of 660 (565-916) times. Only 6 cases were matched <500 times. All analyses are therefore based on 1065 participants whose baseline characteristics are shown in **Table 4.1.** There were 20% women and the median age at CKD event date was 41 years. Cases and controls had similar baseline eGFR (106 mL/min/1.73 m^2); cases were slightly older, less likely to be injection drug users or to be hepatitis C co-infected, had lower CD4 nadir, were more likely to have diabetes, and exposure to ATV/r, LPV/r, and TDF was longer.

Table 4.1. Characteristics of cases and controls

		Entire Case Population (n=743)	**Cases with eGFR drop to <60 mL/min/1.73 m^2 (n=144)**	**Controls** (n=322)
Male gender, n (%)		587 (79)	109 (76)	261 (81)
Age (years), median (interquartile range)		42 (36-47)	45 (41-54)	39 (34-44)
Baseline eGFR (mL/min/1.73 m^2), median (interquartile range)		106 (99-113)	100 (95-107)	107 (98-115)
Median (IQR) time from baseline to CKD date (years), median (IQR)		7.74 (4.98 - 10.81)	9.72 (7.37-12.18)	n.a.
Presumed mode of HIV transmission, n (%)	heterosexual	201 (27)	39 (27)	87 (27)
	MSM	380 (51)	64 (44)	126 (39)
	IDU	137 (18)	34 (24)	101 (31)
	other	25 (3)	7 (5)	8 (2)
Current* smoking, n (%)		410 (55)	78 (54)	241 (76)
Hepatitis C co-infection, n (%)		198 (27)	42 (29)	141 (44)
Duration of atazanavir-ritonavir treatment (years), median (IQR)	All participants**	0 (0-0.97)	0 (0-2.77)	0 (0-0.0)
	Ever exposed***	2.49 (0.78-4.94)	3.61 (1.42-6.81)	1.63 (0.18-3.48)
Duration of lopinavir-ritonavir treatment (years), median (IQR)	All participants**	0 (0-1.17)	0 (0-1.55)	0 (0-0.10)
	Ever exposed***	2.11 (0.70-4.92)	2.76 (1.11-5.2)	1.65 (0.62-4.00)
Duration of tenofovir disoproxil fumarate treatment (years), median (IQR)	All participants**	4.52 (1.76-7.03)	6.68 (2.76-9.18)	1.75 (0-5.21)
	Ever exposed***	5.11 (2.67-7.38)	7.19 (3.85-9.60)	4.19 (1.47-6.05)
CD4+ T-cell count nadir (IQR), (cells/µL)		209 (64-370)	152 (43-295)	280 (150-405)
Hypertension		89 (12)	22 (15)	40 (12)
Prior cardiovascular disease		12 (1.6)	3 (2.1)	5 (1.6)
Diabetes mellitus		23 (3.1)	6 (4)	5 (1.6)

Notes. Data are no. (%) of participants, unless otherwise indicated. *at baseline +/- 1 year. **all CKD cases and controls, irrespective of whether ever treated with the respective ART drug or not. ***only those CKD cases and controls who were ever treated with the respective ART drug. CI, confidence interval; eGFR, estimated glomerular filtration rate; IDU, injection drug use; MSM, men who have sex with men; n.a., not applicable

4.4.2 CKD risks according to clinical D:A:D score, genetic score, and ART, univariable analyses

CKD odds ratio was associated with D:A:D score, genetic score, and cumulative ATV/r, LPV/r, and TDF exposure in univariable analyses (**Figure 4.1A**). Compared to the first (most favorable) D:A:D score quartile, participants in the 2nd, 3rd, and 4th (most unfavorable) quartiles had CKD odds ratios (OR) of 1.51 (95% confidence interval, 1.11-2.03), 1.77 (1.36-2.35), and 2.32 (1.70-3.06), respectively. Compared to the 1st (most favorable) genetic score quartile, participants in the 2nd, 3rd, and 4th (most unfavorable) quartiles had CKD OR of 1.12 (0.86-1.46), 1.46 (1.16-1.84), and 1.88 (1.47-2.45), respectively. Cumulative 5-year exposure to ATV/r, LPV/r, and TDF was associated with CKD OR of 2.93 (2.05-4.45), 1.64 (1.32-2.06), and 1.96 (1.59-2.52), respectively.

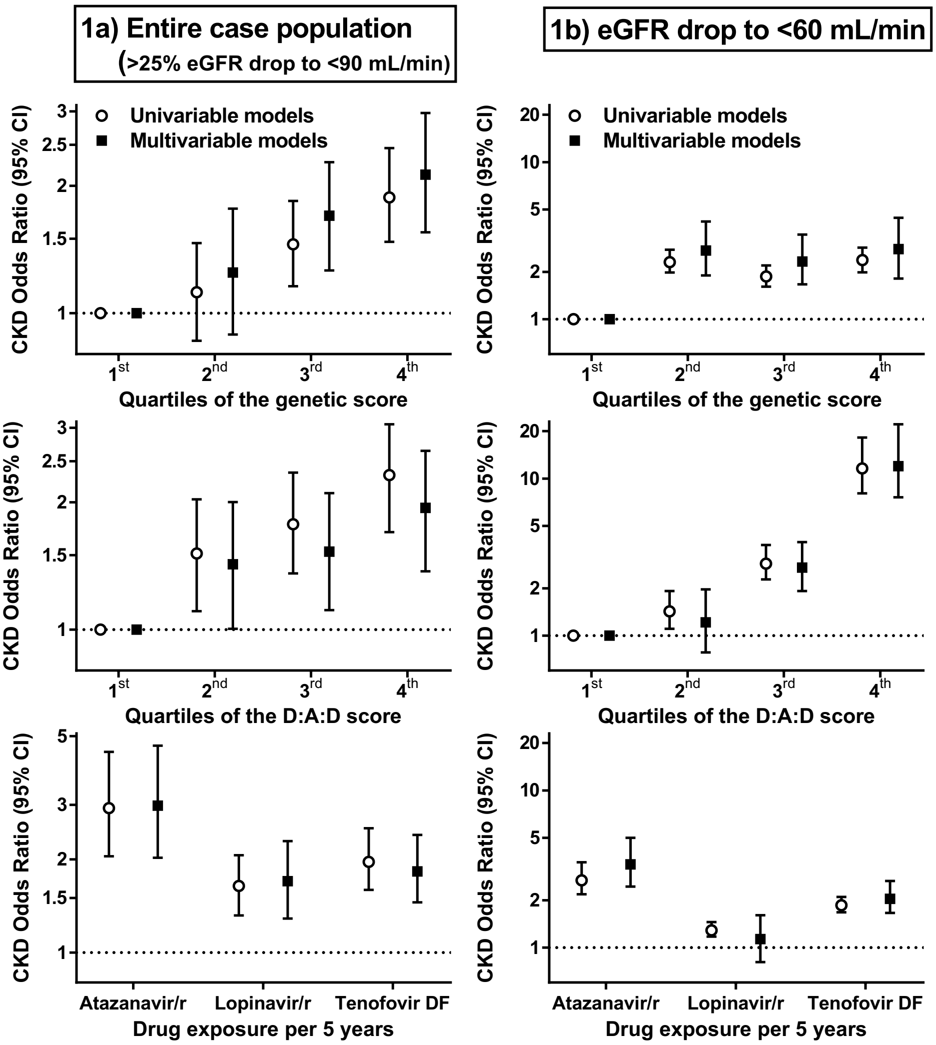


Figure 4.1. CKD odds ratio according to quartiles of genetic score, quartiles of D:A:D score, and per 5-year antiretroviral exposures. Uni- and multivariable conditional logistic regression of associations with CKD. Results are pooled estimates from 2000 re-sampled 1:1 case-control pairs involving 743 cases and 322 controls. Multivariable models are adjusted for all variables displayed, i.e. for genetic score, D:A:D score, and drug exposures, respectively.

4.4.3 CKD risks according to clinical D:A:D Score, genetic score, and ART, multivariable analyses

CKD odds ratio remained associated with D:A:D score, genetic score, and cumulative ATV/r, LPV/r, and TDF exposure in multivariable analyses (**Figure 4.1A**). Compared to the first D:A:D score quartile, participants in the 2nd, 3rd, and 4th quartiles had CKD odds ratios (OR) of 1.43 (1.00-2.00), 1.53 (1.11-2.10), and 1.94 (1.37-2.65), respectively. Compared to the 1st genetic score quartile, participants in the 2nd, 3rd, and 4th quartiles had CKD OR of 1.25 (0.89-1.77), 1.70 (1.26-2.27), and 2.13 (1.55-2.97), respectively. Cumulative 5-year exposure to ATV/r, LPV/r, and TDF was associated with CKD OR of 2.98 (2.02-4.66), 1.70 (1.29-2.29), and 1.83 (1.45-2.40), respectively.

4.4.4 Interaction of clinical and genetic risk score, adjusted for ART exposure

To evaluate whether genetic background modifies the clinical CKD odds ratio captured in the D:A:D score, we introduced an D:A:D score - genetic score interaction term to the multivariable model. The low CKD risk in the most favorable 1st D:A:D score quartile was not significantly modified by the participant's genetic score quartile (**Figure 4.2, Supplementary Table 4.2**). Participants in the 2nd D:A:D score quartile only had a significantly increased CKD odds ratio when they were in the most unfavorable (4th) genetic score quartile, when compared to the most favorable profile (D:A:D quartile 1, genetic score quartile 1). For participants in the highest (4th) CKD risk D:A:D score quartile there was no evidence for an increased CKD odds ratio when they had the most favorable genetic score (1st quartile).

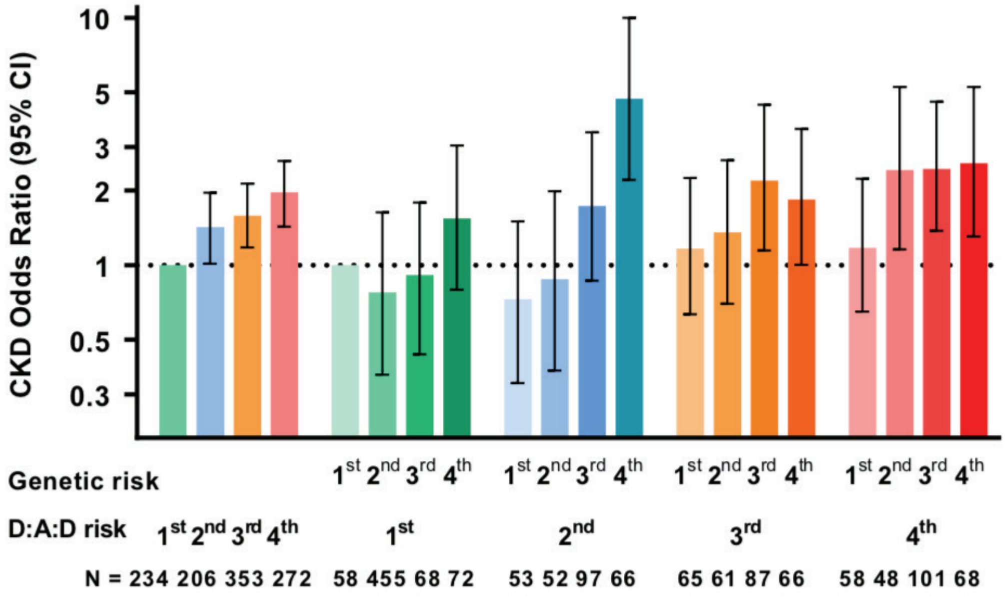


Figure 4.2. CKD odds ratios according to quartiles of genetic score and quartiles of D:A:D score, adjusted for antiretroviral exposures. The first of these 16 groups, i.e., participants who are in D:A:D score quartile 1 and in genetic score quartile 1, is the reference (odds ratio = 1, without confidence interval). The adjusted odds ratios and 95% confidence intervals displayed here in Figure 4.2 are tabulated in Supplementary Table 4.2. Results from two conditional logistic regression analyses of associations with CKD. Results are pooled estimates from 2000 re-sampled 1:1 case-control pairs involving 743 cases and 322 controls. The leftmost four bars show estimates for quartiles of the D:A:D risk score adjusted for drug exposure to ATV/r, LPV/r, and TDF, without consideration of genetic score. Participants are then stratified into 16 groups by genetic score quartile (quartile 1, 2, 3, and 4) and by D:A:D score quartile (quartile 1, 2, 3, and 4), and these odds ratios are also adjusted for ATV/r, LPV/r, and TDF exposure.

4.4.5 Sensitivity analyses – additional CKD case definitions

When restricting the analyses to CKD cases with eGFR drop to <60 mL/min/1.73 m^2 (n=144), this case population was older, had lower baseline eGFR, and time to CKD event was longer compared to the entire case population (**Table 4.1**). CKD odds ratio remained associated with D:A:D score, genetic score, and cumulative ATV/r, LPV/r, and TDF exposure in uni- and multivariable analyses (**Figure 4.1B**). In multivariable analysis, compared to the first D:A:D score quartile, participants in the 2nd, 3rd, and 4th quartiles had CKD OR of 1.22 (.78-1.97), 2.71 (1.93-3.94), and 11.97 (7.61-22.17), respectively. Compared to the 1st genetic score quartile, participants in the 2nd, 3rd, and 4th quartiles had CKD OR of 2.74 (1-90-4.18), 2.33 (1.67-3.46), and 2.79 (1.81-4.43), respectively (see also **Supplementary Table 4.3**).

Results were similar when applying the intermediate CKD case definitions (i.e. >25% eGFR drop to <70 mL/min/1.73 m^2 [n=449]; or as eGFR drop to <60 mL/min/1.73 m^2 OR of $\geq$40%; n=204) (**Supplementary Table 4.3).**

4.4.6 Sensitivity analyses – exclusion of certain ART agents

When participants treated with dolutegravir (n=146) were excluded, genetic score remained significantly associated with CKD but the effect size was slightly attenuated (**Supplementary Table 4.4)**. For example, in the 4th vs. 1st genetic score quartile, CKD OR was 1.80 (1.34-2.47) and 1.96 (1.33-2.95) in univariable and multivariable models, respectively. When all participants treated with any integrase inhibitor (n=244) were excluded, genetic score remained significantly associated with CKD but the effect size was attenuated (**Supplementary Table 4.5**). For example, in the 4th vs. 1st genetic score quartile, CKD OR was 1.58 (1.15-2.20) and 1.68 (1.10-2.61) in univariable and multivariable models, respectively. When participants treated with rilpivirine and cobicistat were excluded, results remained essentially unchanged (**Supplementary Tables 4.6 and 4.7**).

4.4.7 Sensitivity analysis – weighting of cases and controls with the inverse probability of being sampled

Results remained very similar when patients who were sampled less often get more weight (**Supplementary Table 4.8; Supplementary Figure 4.3 and 4.4**).

4.4.8 Exploratory GWAS, candidate SNP replication analysis

In exploratory GWAS, no SNPs were found to be genome-wide significant (P<5e-8, **Supplementary Figure 4.2**). Of 59 previously published candidate SNPs, 2 SNPs replicated as nominally significant, with P-values of 0.03 and 0.05 in the GWAS (**Supplementary Table 4.1)**.

4.5 Discussion

Our findings suggest that in white HIV-positive individuals an unfavorable genetic background increases the incidence of CKD approximately 2-fold. This genetic effect size was similar to the well validated D:A:D score [5,6], and similar to the CKD effect of 5 years treatment with LPV/r or TDF, but smaller than the CKD effect of 5 years ATV/r treatment. The genetic score appears robust, because in multivariable analyses and in sensitivity analyses, it remained independently associated with CKD after adjusting for D:A:D score and for potentially nephrotoxic ART. To our knowledge, this is the first application of a genome-wide polygenic risk score and its integration with clinical risk factors and ART exposure to better explain individual CKD risk in HIV-positive persons.

Our results further suggest that the individual CKD risk captured in the D:A:D score can additionally be stratified by knowledge of genetic background, based on our identification of a clinically relevant interaction between genetic score and D:A:D score. Most importantly, even individuals in the highest clinical risk category (4th D:A:D score quartile) were protected against CKD if they had the most favorable genetic background (1st genetic score quartile). Therefore, a favorable genetic background might explain why certain HIV-positive persons with high clinical CKD risk may not develop CKD, even in the presence of multiple clinical risk factors. Conversely, the most unfavorable genetic background was associated with CKD even with a relatively low D:A:D score (2nd quartile), but was not associated with CKD with the lowest risk D:A:D quartile, highlighting the interaction of genetic and clinical CKD risk factors.

The polygenic risk score may predict more severe CKD better than milder degrees of CKD. The effect size of unfavorable genetic background increased from an approximately 2-fold to an almost 3-fold increased CKD odds ratio, when restricting the analyses to those with eGFR drop to <60 mL/min/1.73m^2. In these participants, D:A:D score was the strongest predictor of CKD, with the effect size increasing from approx. 2-fold increased CKD odds ratio, as in the entire case population, to an approx. 12-fold increase. This was not unexpected, because the variable with by far the largest effect size in the D:A:D score is age, [5] and those with eGFR drop to <60 mL/min/1.73m^2 were older (median age 45 vs. 41 years in the entire case population). In addition, the D:A:D score was developed in a population with eGFR drop to <60 mL/min/1.73m^2), [5] and not in the much larger segment of individuals with eGFR 60-89 mL/min/1.73m^2.

We exploited clinical, laboratory, and HIV-related data from >1000 HIV-positive participants prospectively followed at regular intervals in the well-established Swiss HIV Cohort Study. This allowed the consideration of all relevant CKD-related risk factors and co-morbidities,[5] and of potentially nephrotoxic ART.[6] The polygenic CKD risk score we used summarizes the genome-wide risk captured by >86’000 SNPs.[7-9] We applied rigorous quality control of the genotyping data, excluded population outliers and corrected for residual population stratification. As in our previous genetic studies of dyslipidemia,[23] diabetes mellitus,[24] coronary artery disease events,[25] and osteoporotic fractures,[26] we based SNP selection on large previous GWAS meta-analyses in the general population.[7-9] As expected, we were unable to confirm most previous candidate-gene kidney association studies in HIV.[12,13]

CKD definitions rely on ultimately arbitrary degrees of eGFR drop, therefore we used a CKD case definition (normal baseline eGFR with subsequent drop to <60 mL/min/1.73 m^2) extensively validated in the D:A:D study[6] and in the renal substudy of the START trial.[27] Because this degree of CKD is uncommon (1% of D:A:D participants [6]), we also included participants who developed less severely decreased kidney function. The polygenic risk score was robust, i.e. it predicted CKD independent of the definition used. As expected, applying a rigorous control definition (longitudinal eGFR drop of <15%) limited the number of controls available, but this allowed us to achieve clear phenotypic separation of cases and controls and to thereby better capture the genetic effects. The issue of fewer controls than cases was successfully addressed by applying a well validated procedure, bootstrap resampling from cases and controls,[16] which yielded effect estimates for D:A:D score and genetic score with appropriately narrow confidence intervals.

Our results apply to individuals of European descent. Because of the relatively small number of women and persons >65 years of age included in our study, the results should be cautiously extrapolated to these populations. Additional studies are needed to confirm preliminary findings from trans-ethnic GWAS meta-analyses which suggest that genetic results may potentially be generalized from persons of European descent to persons of African descent.[8]

In conclusion, genetic background may provide CKD risk information complementary to that afforded by traditional CKD risk factors and antiretroviral regimen. Knowledge of an adverse genetic CKD predisposition might further emphasize the rationale to avoid potentially nephrotoxic antiretroviral and other drugs, and to optimize management of other factors contributing to CKD risk, including hypertension and diabetes. The clinical value of genetic testing will rely on demonstration of improved CKD risk stratification in prospective studies. This was beyond the scope of our study. Finally, CKD odds ratios of the genetic score were attenuated when patients treated with integrase inhibitors were excluded, highlighting the interest in future studies that quantitate the genetic effect in patients using different modern ART combinations.

4.6 Declarations

4.6.1 Conflicts of Interest

MC's institution has received research grants from Gilead and Viiv. MC has received travel grants from Abbvie and Gilead. PET's institution has received research grants and advisory fees from Gilead and Viiv.

4.6.2 Funding

This work was supported by the Swiss HIV Cohort Study, supported by the Swiss National Science Foundation (grant number 148522); by SHCS project 775; and by the SHCS research foundation. SHCS data are gathered by the Five Swiss University Hospitals, two Cantonal Hospitals, 15 affiliated hospitals and 36 private physicians (listed in http://www.shcs.ch/180-health-care-providers).

Members of the Swiss HIV Cohort Study: Anagnostopoulos A, Battegay M, Bernasconi E, Böni J, Braun DL, Bucher HC, Calmy A, Cavassini M, Ciuffi A, Dollenmaier G, Egger M, Elzi L, Fehr J, Fellay J, Furrer H (Chairman of the Clinical and Laboratory Committee), Fux CA, Günthard HF (President of the SHCS), Haerry D (deputy of "Positive Council"), Hasse B, Hirsch HH, Hoffmann M, Hösli I, Huber M, Kahlert CR, Kaiser L, Keiser O, Klimkait T, Kouyos RD, Kovari H, Ledergerber B, Martinetti G, Martinez de Tejada B, Marzolini C, Metzner KJ, Müller N, Nicca D, Paioni P, Pantaleo G, Perreau M, Rauch A (Chairman of the Scientific Board), Rudin C (Chairman of the Mother & Child Substudy), Scherrer AU (Head of Data Centre), Schmid P, Speck R, Stöckle M, Tarr P, Trkola A, Vernazza P, Wandeler G, Weber R, Yerly S.

4.6.3 Acknowledgments

The authors acknowledge the effort and commitment of investigators, study nurses, laboratory personnel, and participants.

4.7 Supplementary tables and figures

Supplementary Table 4.1. Candidate SNPs associated with CKD in the general population that we analysed

SNP	CHR	BP	Allele1	Allele2	MAF[1]	BETA	P-value	PUBMED ID	FIRST AUTHOR
rs2467853	15	45698793	G	T	0.389	-0.11	0.144	19430482	Kottgen A
rs12917707	16	20367690	T	G	0.176	-0.111	0.831	19430482	Kottgen A
rs17319721	4	77368847	A	G	0.417	-0.020	0.961	19430482	Kottgen A
rs10518733	15	53940307	C	A	0.181	-0.028	0.958	20139978	Kamatani Y
rs4821469	22	36616445	C	T	0.099	1.336	0.066	20532800	Bostrom MA
rs9310709	3	23093574	C	T	0.485	0.135	0.740	20686651	Gudbjartsson DF
rs13070584	3	99266337	T	C	0.043	0.273	0.781	20686651	Gudbjartsson DF
rs10941694	5	45197779	A	G	0.170	-0.012	0.983	20686651	Gudbjartsson DF
rs4293393	16	20364588	G	A	0.183	0.006	0.990	20686651	Gudbjartsson DF
rs6569474	6	127493611	A	T	0.480	0.818	**0.045**	21909109	Kim YJ
rs6499166	16	68326917	G	A	0.294	0.358	0.429	22962313	Chasman DI
rs1153831	15	45772448	A	G	0.135	0.957	0.107	23254893	Park H
rs10032549	4	77398015	A	G	0.480	0.216	0.593	23535967	Tin A
rs4859682	4	77410318	A	C	0.434	-0.193	0.641	23535967	Tin A
rs17126268	1	102726399	C	T	0.075	1.450	0.054	24351856	Nanayakkara S
rs10099338	8	9257739	G	A	0.107	-0.288	0.653	24351856	Nanayakkara S
rs2980098	4	4252956	G	A	0.335	0.009	0.982	24351856	Nanayakkara S
rs1965907	4	88438176	C	T	0.447	-0.871	**0.027**	24358131	Thameem F
rs878953	5	163926492	A	G	0.433	-	0.206	24358131	Thameem F

						0.500			
rs1686430	2	10947801	A	G	0.432	-0.512	0.206	24358131	Thameem F
rs1734449	2	10950284	G	C	0.432	-0.512	0.206	24358131	Thameem F
rs762063	14	53098902	G	A	0.469	-0.483	0.217	24358131	Thameem F
rs6879805	5	173541828	C	T	0.376	0.496	0.228	24358131	Thameem F
rs1019603	8	113747816	T	C	0.245	-0.513	0.265	24358131	Thameem F
rs925470	4	26015986	T	A	0.467	0.430	0.272	24358131	Thameem F
rs2180419	6	21728317	A	G	0.359	-0.445	0.288	24358131	Thameem F
rs1420725	12	2751583	G	C	0.302	-0.419	0.334	24358131	Thameem F
rs856830	6	68283357	T	A	0.402	0.282	0.483	24358131	Thameem F
rs1516822	4	153330301	C	T	0.278	0.307	0.514	24358131	Thameem F
rs1703711	10	132641714	G	T	0.438	-0.241	0.542	24358131	Thameem F
rs7037744	9	92903715	T	C	0.428	0.207	0.607	24358131	Thameem F
rs11457	15	63886379	G	C	0.437	0.209	0.617	24358131	Thameem F
rs6901750	6	17019426	G	T	0.447	0.197	0.630	24358131	Thameem F
rs767707	1	167560542	G	A	0.491	-0.169	0.675	24358131	Thameem F
rs2928927	18	49423185	C	T	0.361	-0.160	0.691	24358131	Thameem F
rs1904899	8	21456116	G	A	0.453	0.121	0.768	24358131	Thameem F
rs580839	15	34998829	A	G	0.437	0.117	0.768	24358131	Thameem F
rs74111	6	15786602	G	T	0.464	-0.035	0.933	24358131	Thameem F
rs9481410	6	97677118	G	A	0.222	0.424	0.391	24385048	Cooke Bailey JN
rs3775067	4	2888622	A	G	0.400	-0.426	0.302	24658007	Montasser ME

rs2488815	4	2988636	T	C	0.422	0.159	0.692	24658007	Montasser ME
rs4353	17	61570422	G	A	0.455	-0.093	0.820	24658007	Montasser ME
rs4316	17	61562309	T	C	0.449	-0.080	0.844	24658007	Montasser ME
rs4343	17	61566031	A	G	0.458	-0.054	0.893	24658007	Montasser ME
rs4331	17	61564052	G	A	0.449	-0.051	0.901	24658007	Montasser ME
rs4762	1	230845977	A	G	0.119	-0.069	0.912	24658007	Montasser ME
rs4580098	15	45596909	A	T	0.223	-0.637	0.198	25082825	Sveinbjornsson G
rs77924615	16	20392332	A	G	0.205	-0.318	0.519	25082825	Sveinbjornsson G
rs1044261	10	1065710	T	C	0.084	0.157	0.832	25082825	Sveinbjornsson G
rs10794720	10	1156165	T	C	0.083	0.069	0.925	25082825	Sveinbjornsson G
rs60136849	16	20353815	C	T	0.180	0.023	0.965	25082825	Sveinbjornsson G
rs35772020	10	863482	A	G	0.073	-0.018	0.981	25082825	Sveinbjornsson G
rs17069906	18	60048394	G	A	0.025	2.253	0.072	25478860	Leiherer A
rs4845625	1	154422067	T	C	0.413	-0.097	0.811	25524550	Horibe H
rs699	1	230845794	G	A	0.428	0.350	0.388	25660845	Sarkar S
rs2074379	4	113352899	G	A	0.379	0.816	0.052	25813695	Yamada Y
rs2074388	4	113352397	G	A	0.379	0.816	0.052	25813695	Yamada Y
rs6929846	6	26458265	T	C	0.201	-0.16	0.403	25813695	Yamada Y

Notes. [1]MAF: Minor allelic frequency. Candidate SNPs were excluded if; (i) the study was restricted to a specific condition (e.g. end-stage renal disease in type 1 diabetics); and/or (ii) the study design exclusively focused on linkage analysis, copy number variation, large deletions, burden of rare variants, gene–environment interaction and epistasis.

Supplementary Table 4.2. Interaction of D:A:D score with Genetic score adjusted for ART, multivariable analysis

Variable, Stratum	**Entire case population (n=743)** NB: The results shown in this column are the same adjusted odds ratios and 95% confidence intervals that are illustrated in **Figure 2**.
D:A:D 1st quartile and Genetic score 1st quartile	1.00 (Reference)
D:A:D 1st quartile and Genetic score 2nd quartile	0.78 (0.36-1.63)
D:A:D 1st quartile and Genetic score 3rd quartile	0.91 (0.44-1.79)
D:A:D 1st quartile and Genetic score 4th quartile	1.55 (0.79-3.04)
D:A:D 2nd quartile and Genetic score 1st quartile	0.73 (0.33-1.50)
D:A:D 2nd quartile and Genetic score 2nd quartile	0.88 (0.37-1.99)
D:A:D 2nd quartile and Genetic score 3rd quartile	1.74 (0.86-3.45)
D:A:D 2nd quartile and Genetic score 4th quartile	4.71 (2.21-10.00)
D:A:D 3rd quartile and Genetic score 1st quartile	1.17 (0.63-2.25)
D:A:D 3rd quartile and Genetic score 2nd quartile	1.36 (0.70-2.66)
D:A:D 3rd quartile and Genetic score 3rd quartile	2.20 (1.15-4.44)
D:A:D 3rd quartile and Genetic score 4th quartile	1.84 (1.00-3.56)
D:A:D 4th quartile and Genetic score 1st quartile	1.17 (0.65-2.23)
D:A:D 4th quartile and Genetic score 2nd quartile	2.42 (1.16-5.25)
D:A:D 4th quartile and Genetic score 3rd quartile	2.44 (1.38-4.57)
D:A:D 4th quartile and Genetic score 4th quartile	2.58 (1.30-5.24)

Supplementary Table 4.3. Main results tabulated for the 4 different CKD case definitions: CKD odds ratio (95% confidence interval) according to D:A:D score, genetic score, and ART

	Entire Population of CKD Cases (n=743) *(NB: these are the numbers corresponding to results illustrated in **Fig. 4.1A**)*		**Cases with eGFR drop to <60 mL/min/1.73 m^2 (n=144)** *(NB: these are the numbers corresponding to results illustrated in **Fig. 4.1B**)*		**Cases with eGFR drop ≥40% OR to <60 mL/min/1.73 m^2 (n=204)**		**Cases with eGFR drop >25% AND to <70 mL/min/1.73 m^2 (n=449)**	
Variable	**Univariable analysis**	**Multivariable analysis**	**Univariable analysis**	**Multivariable analysis**	**Univariable analysis**	**Multivariable analysis**	**Univariable analysis**	**Multivariable analysis**
D:A:D 2nd quartile vs. 1st quartile	**1.51** **(1.11-2.03)**	**1.42** **(1.00-2.00)**	1.43 (1.11-1.92)	1.22 (0.78-1.97)	1.34 (1.06-1.75)	1.06 (0.71-1.56)	2.59 (1.87-3.72)	2.46 (1.67-3.82)
D:A:D 3rd quartile vs. 1st quartile	**1.77** **(1.36-2.45)**	**1.53** **(1.11-2.10)**	2.88 (2.28-3.77))	2.71 (1.93-3.94)	2.49 (2.01-3.23)	2.20 (1.64-3.08)	2.99 (2.21-4.25)	2.62 (1.82-4.06)
D:A:D 4th quartile vs. 1st quartile	**2.32** **(1.70-3.06)**	**1.94** **(1.37-2.65)**	11.59 (8.06-18.22)	11.97 (7.61-22.17)	5.87 (4.58-8.10)	5.22 (3.73-7.72)	5.58 (4.13-8.03)	5.05 (3.48-7.80)
Genetic score 2nd quartile vs. 1st quartile	**1.12** **(0.86-1.46)**	**1.24** **(0.89-1.77)**	2.31 (1.98-2.77))	2.74 (1.90-4.18)	1.63 (1.37-1.98)	1.66 (1.19-2.35)	1.73 (1.35-2.28)	2.38 (1.62-3.49)
Genetic score 3rd quartile vs. 1st quartile	**1.46** **(1.16-1.84)**	**1.70** **(1.26-2.27)**	1.87 (1.61-2.20)	2.33 (1.67-3.46)	1.42 (1.21-1.67)	1.62 (1.20-2.25)	1.74 (1.42-2.13)	2.19 (1.61-3.05)
Genetic score 4th quartile vs. 1st quartile	**1.88** **(1.47-2.45)**	**2.13** **(1.55-2.97)**	2.38 (1.99-2.85)	2.79 (1.81-4.43)	1.76 (1.52-2.08)	1.66 (1.19-2.35)	2.31 (1.83-3.00)	3.01 (2.05-4.56)
Cumulative ATV/r exposure, per 5 years	**2.93** **(2.05-4.45)**	**2.98** **(2.02-4.66)**	2.68 (2.19-3.49)	3.38 (2.45-4.99)	3.21 (2.56-4.26)	3.60 (2.62-5.32)	2.83 (2.13-3.96)	3.14 (2.14-4.86)

Cumulative LPV/r, per 5 years	**1.64** **(1.32-2.06)**	**1.70** **(1.29-2.29)**	1.29 (1.18-1.46)	1.13 (0.81-1.61)	1.42 (1.25-1.64)	1.43 (1.07-1.97)	1.59 (1.31-1.96)	1.63 (1.21-2.25)
Cumulative TDF exposure, per 5 years	**1.96** **(1.59-2.52)**	**1.83** **(1.45-2.40)**	1.86 (1.68-2.10)	2.04 (1.66-2.65)	2.21 (1.93-2.63)	2.32 (1.87-2.95)	1.99 (1.64-2.47)	1.94 (1.51-2.54)

Supplementary Table 4.4. Sensitivity Analysis (entire case population) after exclusion of participants treated with dolutegravir (n=147): CKD odds ratio (95% confidence interval) according to D:A:D Score, genetic score, and ART

Variable	**Univariable analysis**	**Multivariable analysis**
D:A:D 2nd quartile vs. 1st quartile	1.43 (1.00-2.03)	1.29 (0.83-1.96)
D:A:D 3rd quartile vs. 1st quartile	1.83 (1.34-2.54)	1.48 (1.01-2.19)
D:A:D 4th quartile vs. 1st quartile	2.50 (1.77-3.37)	1.93 (1.30-2.80)
Genetic score 2nd quartile vs. 1st quartile	1.03 (0.75-1.42)	1.09 (0.71-1.63)
Genetic score 3rd quartile vs. 1st quartile	1.36 (1.03-1.83)	1.49 (1.03-2.14)
Genetic score 4th quartile vs. 1st quartile	1.80 (1.34-2.47)	1.96 (1.33-2.95)
Cumulative ATV/r exposure, per 5 years	3.62 (2.17-7.16)	3.68 (2.07-7.65)
Cumulative LPV/r, per 5 years	2.11 (1.49-3.18)	2.20 (1.45-3.60)
Cumulative TDF exposure, per 5 years	2.92 (2.14-4.23)	2.62 (1.86-3.91)

Note: This sensitivity analysis is based on 597 CKD cases and 321 controls without any dolutegravir exposure

Supplementary Table 4.5. Sensitivity Analysis (entire case population) after exclusion of participants treated with any integrase inhibitor (n=244): CKD odds ratio (95% confidence interval) according to D:A:D Score, genetic score, and ART

Variable	**Univariable analysis**	**Multivariable analysis**
D:A:D 2nd quartile vs. 1st quartile	1.54 (1.05-2.27)	1.36 (0.82-2.19)
D:A:D 3rd quartile vs. 1st quartile	1.84 (1.32-2.64)	1.41 (0.92-2.18)
D:A:D 4nd quartile vs. 1st quartile	2.52 (1.81-3.55)	1.89 (1.24-2.85)
Genetic score 2nd quartile vs. 1st quartile	0.87 (0.61-1.23)	0.91 (0.57-1.44)
Genetic score 3rd quartile vs. 1st quartile	1.31 (0.95-1.84)	1.38 (0.93-2.10)
Genetic score 4th quartile vs. 1st quartile	1.58 (1.15-2.20)	1.68 (1.10-2.61)
Cumulative ATV/r exposure, per 5 years	4.80 (2.46-13.08)	4.94 (2.38-14.21)
Cumulative LPV/r exposure, per 5 years	2.41 (1.59-3.98)	2.51 (1.46-4.61)
Cumulative TDF exposure, per 5 years	3.15 (2.17-4.99)	2.71 (1.77-4.47)

Note: This sensitivity analysis is based on 499 CKD cases and 318 controls without any integrase inhibitor exposure

Supplementary Table 4.6. Sensitivity Analysis (entire case population) after exclusion of participants treated with Rilpivirine (n=61): CKD odds ratio (95% confidence interval) according to D:A:D Score, genetic score, and ART

Variable	**Univariable analysis**	**Multivariable analysis**
D:A:D 2nd quartile vs. 1st quartile	1.49 (1.06-2.05)	1.41 (0.95-2.06)
D:A:D 3rd quartile vs. 1st quartile	1.75 (1.31-2.39)	1.50 (1.06-2.12)
D:A:D 4nd quartile vs. 1st quartile	2.30 (1.67-3.07)	1.89 (1.33-2.62)
Genetic score 2nd quartile vs. 1st quartile	1.08 (0.80-1.48)	1.22 (0.84-1.83)
Genetic score 3rd quartile vs. 1st quartile	1.47 (1.13-1.94)	1.73 (1.26-2.39)
Genetic score 4th quartile vs. 1st quartile	1.85 (1.40-2.54)	2.16 (1.52-3.16)
Cumulative ATV/r exposure, per 5 years	3.05 (2.02-4.94)	3.18 (2.00-5.54)
Cumulative LPV/r exposure, per 5 years	1.87 (1.43-2.55)	1.97 (1.43-2.92)
Cumulative TDF exposure, per 5 years	2.18 (1.70-2.95)	1.94 (1.45-2.68)

Note: This sensitivity analysis is based on 682 CKD cases and 320 controls without any rilpivirine exposure

Supplementary Table 4.7. Sensitivity Analysis (entire case population) after exclusion of participants treated with Cobicistat (n=37): CKD odds ratio (95% confidence interval) according to D:A:D Score, genetic score, and ART

Variable	**Univariable analysis**	**Multivariable analysis**
D:A:D 2nd quartile vs. 1st quartile	1.52 (1.11-2.05)	1.43 (0.98-2.04)
D:A:D 3rd quartile vs. 1st quartile	1.89 (1.43-2.57)	1.63 (1.16-2.28)
D:A:D 4nd quartile vs. 1st quartile	2.43 (1.80-3.27)	2.02 (1.42-2.78)
Genetic score 2nd quartile vs. 1st quartile	1.11 (0.84-1.47)	1.24 (0.87-1.78)
Genetic score 3rd quartile vs. 1st quartile	1.41 (1.11-1.81)	1.64 (1.20-2.23)
Genetic score 4th quartile vs. 1st quartile	1.85 (1.43-2.46)	2.10 (1.48-2.98)
Cumulative ATV/r exposure, per 5 years	3.13 (2.11-5.11)	3.25 (2.10-5.40)
Cumulative LPV/r exposure, per 5 years	1.74 (1.37-2.30)	1.83 (1.34-2.58)
Cumulative TDF exposure, per 5 years	1.97 (1.58-2.58)	1.79 (1.40-2.40)

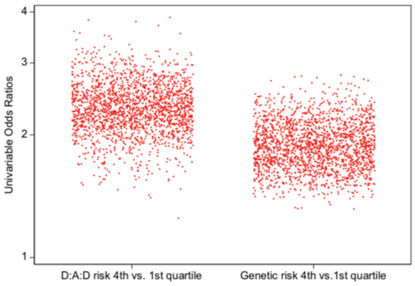
Note: This sensitivity analysis is based on 706 CKD cases and 322 controls without any cobicistat exposure

Supplementary Table 4.8. Sensitivity Analysis (entire case population) after Inverse Probability of Sampling Weighting: CKD odds ratio (95% con-fidence interval) according to D:A:D Score, genetic score, and ART

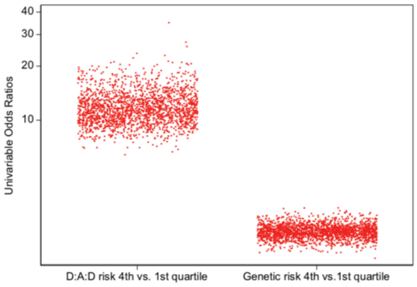
Variable	**Univariable analysis**	**Multivariable analysis**
D:A:D 2nd quartile vs. 1st quartile	1.48 (1.01-2.15)	1.33 (0.92-1.88)
D:A:D 3rd quartile vs. 1st quartile	1.66 (1.21-2.36)	1.41 (1.04-1.94)
D:A:D 4nd quartile vs. 1st quartile	1.68 (1.19-2.39)	1.68 (1.19-2.33)
Genetic score 2nd quartile vs. 1st quartile	1.27 (0.91-1.82)	1.34 (0.95-1.90)
Genetic score 3rd quartile vs. 1st quartile	1.63 (1.21-2.25)	1.87 (1.37-2.56)
Genetic score 4th quartile vs. 1st quartile	1.98 (1.46-2.81)	2.15 (1.55-2.99)
Cumulative ATV/r exposure, per 5 years	3.35 (2.41-4.70)	2.92 (2.02-4.15)
Cumulative LPV/r exposure, per 5 years	1.72 (1.37-2.18)	1.60 (1.22-2.03)
Cumulative TDF exposure, per 5 years	2.28 (1.95-2.75)	2.09 (1.77-2.51)

Note: This sensitivity analysis is based on 743 CKD cases and 322 controls

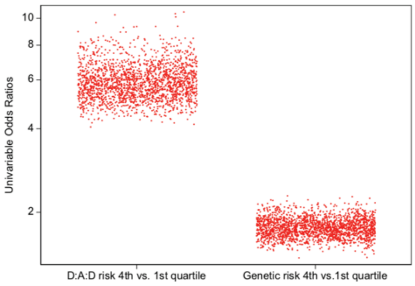
Entire case population (n=743 cases) and 322 controls:



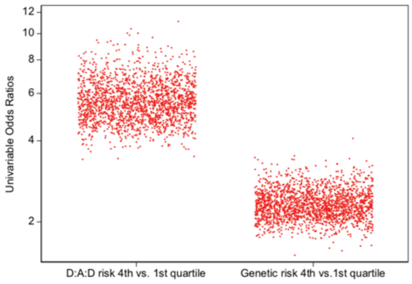
Cases with eGFR drop to <60 mL/min/1.73 m^2 (n=144) and 180 controls



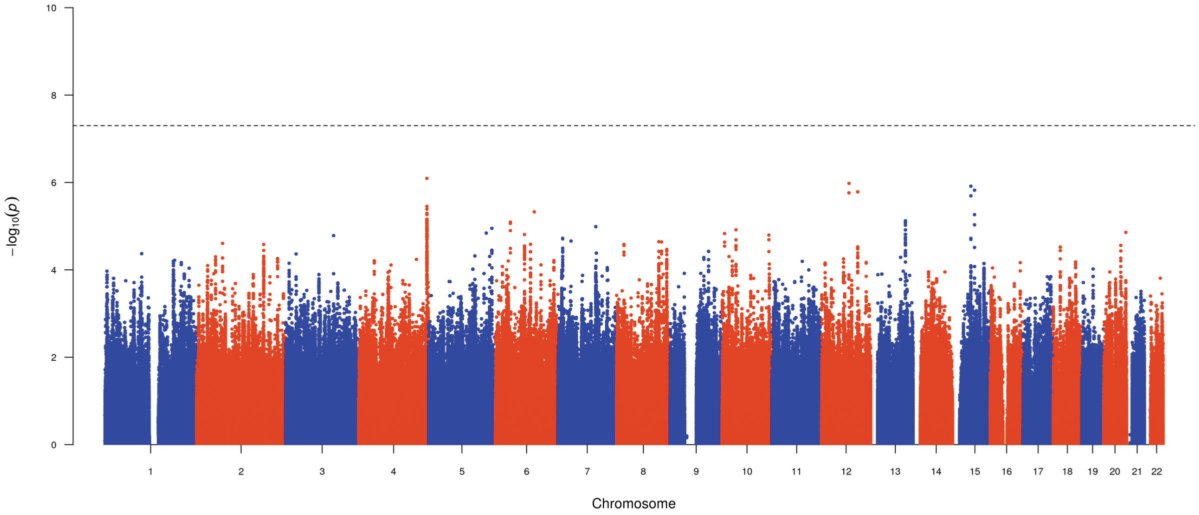
Cases with eGFR drop ≥40% OR to <60 mL/min/1.73 m^2 (n=204) and 230 controls



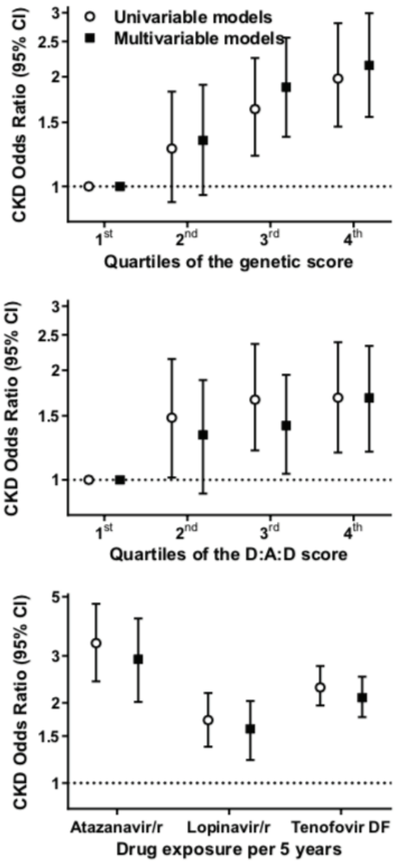
Cases with eGFR drop >25% AND to <70 mL/min/1.73 m^2 (n=449) ad 302 controls



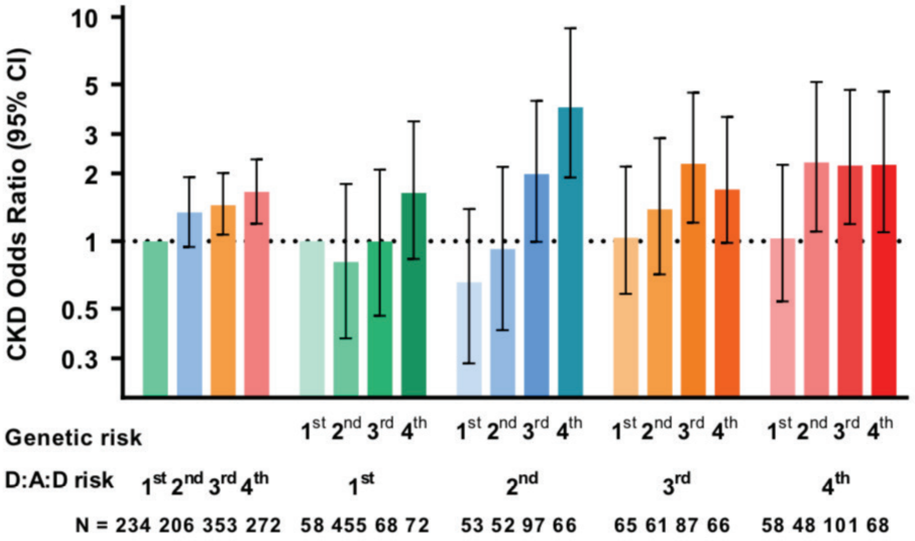
Supplementary Figure 4.1. Limited variability of 2000 univariable estimates of CKD odds ratio in participants with unfavorable D:A:D scores and unfavorable genetic scores, after repeating the matching process 2000 times, with random re-sampling from cases and controls [16]



Supplementary Figure 4.2. Manhattan plot from exploratory GWAS



Supplementary Figure 4.3. CKD odds ratio according to quartiles of genetic score, quartiles of D:A:D score, and per 5-year antiretroviral expo-sures, entire case population (n=743) with inverse probability of sampling weighting. Uni- and multivariable conditional logistic regression of associations with CKD. Results are pooled estimates from 2000 re-sampled 1:1 case-control pairs involving 743 cases and 322 controls. Multivariable models are adjusted for all varia-bles displayed, i.e. for genetic score, D:A:D score, and drug exposures, respectively.



Supplementary Figure 4.4. Interaction of D:A:D with Genetic score adjusted for ART, multivariable analysis with inverse probability of sampling weighting. Results from two conditional logistic regression analyses of associations with CKD. Results are pooled estimates from 2000 re-sampled 1:1 case-control pairs involving 743 cases and 322 controls. The leftmost four bars show estimates for quartiles of the D:A:D risk score adjusted for drug exposure to ATV/r, LPV/r, and TDF, without consideration of genetic score. Participants are then stratified into 16 groups by genetic score quartile (quartile 1, 2, 3, and 4) and by D:A:D score quartile (quartile 1, 2, 3, and 4), and these odds ratios are also adjusted for ATV/r, LPV/r, and TDF exposure. The first of these 16 groups, i.e., participants who are in D:A:D score quartile 1 and in genetic score quartile 1, is the reference (odds ratio = 1, without confidence interval).

4.8 References

1. Mocroft A, Ryom L, Begovac J, et al. Deteriorating renal function and clinical outcomes in HIV-positive persons. AIDS **2014**; 28:727–737.

2. Lucas GM, Ross MJ, Stock PG, et al. Clinical practice guideline for the management of chronic kidney disease in patients infected with HIV: 2014 update by the HIV Medicine Association of the Infectious Diseases Society of America. Clin. Infect. Dis. 2014; 59:e96–138.

3. Abraham AG, Althoff KN, Jing Y, et al. End-stage renal disease among HIV-infected adults in North America. Clin. Infect. Dis. **2015**; 60:941–949.

4. Pelchen-Matthews A, Ryom L, Borges ÁH, et al. Aging and the evolution of comorbidities among HIV-positive individuals in a European cohort. AIDS **2018**; 32:2405–2416.

5. Mocroft A, Lundgren JD, Ross M, et al. Development and validation of a risk score for chronic kidney disease in HIV infection using prospective cohort data from the D:A:D study. PLoS Med. **2015**; 12:e1001809.

6. Mocroft A, Lundgren JD, Ross M, et al. Cumulative and current exposure to potentially nephrotoxic antiretrovirals and development of chronic kidney disease in HIV-positive individuals with a normal baseline estimated glomerular filtration rate: a prospective international cohort study. Lancet HIV **2016**; 3:e23–32.

7. Gorski M, van der Most PJ, Teumer A, et al. 1000 Genomes-based meta-analysis identifies 10 novel loci for kidney function. Sci Rep **2017**; 7:45040.

8. Pattaro C, Teumer A, Gorski M, et al. Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function. Nat Commun **2016**; 7:10023.

9. Ma J, Yang Q, Hwang S-J, Fox CS, Chu AY. Genetic risk score and risk of stage 3 chronic kidney disease. BMC Nephrol **2017**; 18:32.

10. Kopp JB, Nelson GW, Sampath K, et al. APOL1 genetic variants in focal segmental glomerulosclerosis and HIV-associated nephropathy. J. Am. Soc. Nephrol. **2011**; 22:2129–2137.

11. Estrella MM, Li M, Tin A, et al. The association between APOL1 risk alleles and longitudinal kidney function differs by HIV viral suppression status. Clin. Infect. Dis. **2015**; 60:646–652.

12. Izzedine H, Hulot J-S, Villard E, et al. Association between ABCC2 gene haplotypes and tenofovir-induced proximal tubulopathy. Journal of Infectious Diseases **2006**; 194:1481–1491.

13. Nishijima T, Komatsu H, Higasa K, et al. Single nucleotide polymorphisms in ABCC2 associate with tenofovir-induced kidney tubular dysfunction in Japanese patients with HIV-1 infection: a pharmacogenetic study. Clin. Infect. Dis. **2012**; 55:1558–1567.

14. Duncan EL, Danoy P, Kemp JP, et al. Genome-wide association study using extreme truncate selection identifies novel genes affecting bone mineral density and fracture risk. PLoS Genet. **2011**; 7:e1001372.

15. Wheeler E, Huang N, Bochukova EG, et al. Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity. Nat. Genet. **2013**; 45:513–517.

16. Bland JM, Altman DG. Statistics Notes: Bootstrap resampling methods. BMJ **2015**; 350:h2622–h2622.

17. Euesden J, Lewis CM, O'Reilly PF. PRSice: Polygenic Risk Score software. Bioinformatics **2015**; 31:1466–1468.

18. International Schizophrenia Consortium, Purcell SM, Wray NR, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature **2009**; 460:748–752.

19. Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock SJ, Park J-H. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nat. Genet. **2013**; 45:400–5– 405e1–3.

20. Inouye M, Abraham G, Nelson CP, et al. Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention. J. Am. Coll. Cardiol. **2018**; 72:1883–1893.

21. Raffi F, Rachlis A, Stellbrink H-J, et al. Once-daily dolutegravir versus raltegravir in antiretroviral-naive adults with HIV-1 infection: 48 week results from the randomised, double-blind, non-inferiority SPRING-2 study. Lancet **2013**; 381:735–743.

22. Cohen SD, Kopp JB, Kimmel PL. Kidney Diseases Associated with Human Immunodeficiency Virus Infection. N Engl J Med **2018**; 378:1655–1656.

23. Rotger M, Bayard C, Taffé P, et al. Contribution of genome-wide significant single-nucleotide polymorphisms and antiretroviral therapy to dyslipidemia in HIV-infected individuals: a longitudinal study. Circ Cardiovasc Genet **2009**; 2:621–628.

24. Rotger M, Gsponer T, Martinez R, et al. Impact of single nucleotide polymorphisms and of clinical risk factors on new-onset diabetes mellitus in HIV-infected individuals. Clin. Infect. Dis. **2010**; 51:1090–1098.

25. Rotger M, Glass TR, Junier T, et al. Contribution of genetic background, traditional risk factors, and HIV-related factors to coronary artery disease events in HIV-positive persons. Clin. Infect. Dis. **2013**; 57:112–121.

26. Junier T, Rotger M, Biver E, et al. Contribution of Genetic Background and Clinical Risk Factors to Low-Trauma Fractures in Human Immunodeficiency Virus (HIV)-Positive Persons: The Swiss HIV Cohort Study. Open Forum Infect Dis **2016**; 3:ofw101.

27. Mocroft A, Neuhaus J, Peters L, et al. Hepatitis B and C co-infection are independent predictors of progressive kidney disease in HIV-positive, antiretroviral-treated adults. PLoS ONE **2012**; 7:e40245.

28. Loh P-R, Danecek P, Palamara PF, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat. Genet. **2016**; 48:1443–1448.

29. Durbin R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). Bioinformatics **2014**; 30:1266–1272.

30. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet. **2016**; 48:1279–1283.

31. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience **2015**; 4:7.

32. Essebag V, Genest J, Suissa S, Pilote L. The nested case-control study in cardiology. Am. Heart J. **2003**; 146:581–590.

33. Robins JM, Gail MH, Lubin JH. More on "Biased selection of controls for case-control analyses of cohort studies". Biometrics **1986**; 42:293–299.

34. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. **2011**; 88:76–82.

35. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. Nat. Genet. **2014**; 46:100–106.

Chapter 5 Discussion & perspectives

The results described in the previous chapters highlight some of the most recent advances in host genetic research for HIV infected individuals in the era of suppressive ART. The direct consequences of the findings are discussed in their respective chapters, whereas more general issues and opportunities for the HIV research community are discussed here.

5.1 Consequences of the HIV reservoir size and decay rate

The size and long-term dynamics of the HIV reservoir in successfully treated patients represent an essential aspect of HIV pathogenesis after therapy initiation, given that the persistence of the latent reservoir is the main barrier preventing a functional cure. Future treatment or cure strategies may thus benefit from the knowledge of biological genes or pathways influencing the reservoir size as well as its decay rate, which could be uncovered through genetic analyses. While in itself, the lowering of the viral reservoir will likely not represent a cure, unless it successfully purges the complete reservoir, it should at least lower the barrier for obtaining a functional cure. Many approaches have been proposed and tested in order to purge the viral reservoir - most famously, the “shock and kill” theory involving the reactivation of the latent reservoir through cytokine or histone deacetylase inhibitors (1) - but none have yet been successful.

The crucial importance of discovering a cure should not be dismissed, even if people living with HIV generally are in decent health due to the current treatment regiments. The removal of the burden of stigmatization, daily medications, and worries about infecting others would, in addition to the physical benefits, substantially improve their quality of life (2). Just how important it is for individuals infected with HIV to find a cure is underscored by the fact that half of the people infected with HIV are willing to take serious and potentially lethal risks for obtaining a cure (3). This also underscores the importance of proper ethical oversight for experimental HIV cure research.

While we did not discover any genetic association with the HIV reservoir size ~1.5 years after ART initiation nor with the decay rate in chapter 2, it is unlikely that no genetic factors influence these traits. The lack of an observed association is possibly due to the limited sensitivity and specificity of the measurements of the latent HIV reservoir size in our study. Due to the very small amount of integrated HIV DNA, minor fluctuations in reported values will add noise to the phenotype, requiring a larger sample size to observe true associations. Furthermore, the numerous occurrences of viral blips have the potential to mask any genetic effect on the decay rate as they keep reseeding the reservoir at every instance. Thus, the continuous development of improved assays for quantifying the reservoir size along with ever-larger cohorts will be required in order to discover genetic variants affecting the HIV reservoir size and decay rate on treatment. Nevertheless, our study adds important information for the field and will also serve to hopefully prevent future low-powered genetic studies on the HIV reservoir size and dynamics.

An important point for the study described in chapter 2 is that the viral RNA pre-ART variable used as co-variate in regression analyses does not accurately reflect the spVL values used in previous genetic analyses. The pre-ART measurement represents the last viral RNA measurement prior to treatment initiation, a time-point in which CD4+ T cell levels have already decreased significantly (at least in historical samples collected in the early decades of the pandemic). Thus, while the HLA-B*57:01 haplotype has been repeteadly shown to be significantly associated with lower spVL, this effect was no longer seen at the time of reservoir measurement.

The major non-genetic factors influencing the size of the viral reservoir has been well described by now (4). However, an open question remains as to the potential consequences of the size of the viral reservoir on the development of future comorbidities. With the potential of defective provirus to cause inflammation (5), a larger viral reservoir could in theory contribute to increased inflammation levels. Of note, this was not observed by Rajesh *et al.* (6). It is, however, important to note that several different assays for measuring the size of the viral reservoir exists, with relatively poor correlation between them (7, 8). Thus, in order for both experimental studies or clinical trials on the viral reservoir size to be comparable, a gold-standard assay has to be agreed upon first.

5.2 Risk factors of HIV-related comorbidities

The contribution of genetic variation to various common diseases, like CVD, CKD, and multiple types of cancers, have been well studied in the general population, with evidence of a significant genetic contribution to disease risk (9–12). However, to which degree these genetic risk factors contribute to the development of HIV-related comorbidities remains mostly untested. In chapter 3, we sought to answer this question for HIV-related NHL. While we discovered a novel associated locus near the chemokine *CXCL12*, we also did not find any evidence of association with previously discovered genetic variants associated with NHL in the general population. Thus, in the case of HIV-related NHL, and potentially other comorbidities, distinct genes and pathways influencing the risk of developing these comorbidities might exist. Two reasons could explain this distinct risk. 1) In the case of NHLs, this malignancy may constitute a unique subtype in HIV-infected individuals due to specific pathogenic mechanisms involving cytokine deregulation, impaired immune response, chronic antigen stimulation, and reactivation of EBV and HHV-8 (13). 2) HIV infection causes widespread transcriptional changes within the cell (14), which may be further affected by induced eQTLs (e.g., eQTLs that only appear upon stimulation), as seen for both viral- and bacterial infections as well as cellular differentiation stages (15, 16). As of this date, no analysis of HIV infection-induced eQTLs exists, but lessons from studies with other pathogens point to the enrichment of pathogen-induced eQTLs within disease-associated variants (16). Thus, instead of relying on eQTL information from normal cells, the mapping of potential HIV induced eQTLs would be beneficial to HIV genetic studies to better understand the underlying biological mechanisms behind observed disease-associated variants.

Somatic variants constitute an often-overlooked entity in non-cancer genetic studies, thus their impact on health and disease is relatively unknown. While genotyping arrays do not capture somatic variants, both exome sequencing and WGS are able to capture these down to relatively low frequencies (depending on the sequencing depth). Multiple sequencing studies have recently demonstrated that the occurrence and frequency of somatic variants in hematopoietic stem cell populations are increasing with age (17) and are furthermore associated with hematological cancers (18) and CVD (19) in the general population. This phenomenon is known as clonal hematopoiesis of indeterminate potential (CHIP), due to the location of the somatic variants within cancer-associated genes without causing full-blown cancer in most carriers. Interestingly, our preliminary data from exome sequencing of HIV positive individuals in the SHCS indicates that these HIV patients accumulate CHIP at a higher rate than HIV negative individuals, corresponding to approximately ten years in advance (unpublished). In line with this, HIV patients are generally considered to suffer from premature ageing, estimated to an additional 10-15 years compared to HIV negative individuals (20). Chronic low-grade inflammation, toxic drugs, prevalent smoking are all considered risk factors contributing to this observation. However, a more direct biological explanation remains to be found. Thus, the accumulation of CHIP constitutes an area warranting further research, as the premature presence of CHIP may contribute to the increased rate of CVD seen in the HIV population and could point to an increased mutational burden within hematopoietic cells.

In general, many risk factors have been associated with the development of common diseases in the HIV population, and many of these factors are also associated with each other, making it challenging to determine the main causal factors. Thus, determining these causal risk factors, whether genetic or not, will be of enormous value in order to guide treatment and prevention strategies. So far, comprehensive studies measuring the contribution of risk factors to inflammation and the development of comorbidities have focused solely on single or small subset of risk factors at a time (e.g., smoking or microbial translocation). However, few have measured all the main risk factors simultaneously in the same cohort to detangle the contribution of each factor. This was, to some extent, performed for CKD with the D:A:D studies (21)(chapter 4). However, even this study did not include markers of residual HIV production, levels of T_{reg} and T_{17} cells, microbial translocations, or markers of inflammation. Furthermore, the inclusion of novel biomarkers capable of measuring chronic inflammation instead of classical markers of acute inflammation (e.g., CRP) in clinical risk scores might also further enhance their predictive performance (22–24).

5.3 Adaptation and potential use of genetic risk scores

Human genetic discoveries have had a limited impact so far in the clinical care of HIV patients, with only the presence of the HLA-B*57:01 allele routinely checked due to its importance in predicting abacavir hypersensitivity. However, the recent developments in methodology and GWAS cohort sizes have greatly facilitated the path towards clinical translation of GRS. This progress has, in particular, been spearheaded by research into CAD in the general population. The main finding has been that an unfavorable GRS carries the same risk for CAD as monogenic risk variants (25). Recent work further established that polygenic risk captured by GRS is able to modulate the risk conferred by monogenic variants for CAD, breast cancer, and colorectal cancer (26). Thus, accounting for both polygenic- and monogenic risk might improve the overall risk prediction. The clinical relevance of GRSs was further underlined by the finding that adding a GRS for CAD to current clinical risk scores independently improved the predictive accuracy (27), while individuals with the most unfavorable GRS also benefited the most from statin treatment in terms of relative and absolute LDL-cholesterol reduction (28, 29).

In chapter 5, we described the first genome-wide GRS in the HIV population with CKD, demonstrating how adding a GRS to the current clinical risk algorithm improved not only the risk prediction, but also identified a subset of patients at high risk of developing CKD despite having a low clinical risk score. Furthermore, a beneficial GRS was strongly protective against CKD despite the presence of clinical risk factors. Thus, the GRS for CKD in the HIV population represents a risk marker that is independent of other currently used predictors, in line with the previous findings for CAD in the general population.

The independent nature of the risk captured by GRS, as demonstrated for CAD in the general population and here for CKD in the HIV population, suggests that the inclusion of a GRS together with established clinical risk scores can enhance the possibilities of detecting individuals with a high risk of developing certain comorbidities at an early stage. The early identification of high-risk individuals allows for preventive interventions, thus reducing the occurrence and hospital burden of overt clinical diseases.

The implementation of GRS in clinical routine will require a substantial change in the way healthcare is generally performed by moving towards a more preventive approach rather than the current one of dealing with already manifested diseases. Furthermore, as many preventive interventions constitute not only therapeutics but also behavioral changes in lifestyle or drug compliance, innovative support frameworks will also be needed. The practical implementation of GRS faces other challenges: currently, the use of different calculation methods, as well as a lack of best practices to standardize the calculation and validation of GRS, have resulted in the generation of widely variable GRSs for the same disease. The possibility of modifying many parameters and lack of consistency for the GRS calculation and validation also makes it challenging to compare the accuracy of different GRS. Thus, method standardization of GRS calculations and their validation across cohorts and hospitals will be essential for their implementation into clinical use, both in the general- and HIV population.

Importantly, the accuracy of GRS will keep improving as the size of GWAS continues to increase, and the estimated effect sizes thus become more precise. Furthermore, the shift from genotyping arrays to WGS as the primary technique should further improve the information included and, thus, the accuracy of GRS. The use of WGS would also allow for the simultaneous identification of monogenic variants contributing to disease risk. However, currently, only a few studies have been performed using GRS from WGS data due to their limited availability (30). The recently announced initiative in the UK Biobank for performing WGS on 0.5 million individuals will constitute an important opportunity to benchmark GRS based on genotyping arrays versus WGS in large population cohorts, the outcome of which will likely shape the future direction of GRS research and implementation initiatives.

5.3.1 The need for increasing diversity of research participants

Human genome diversity is strongly shaped by demographic history, with the migration out of Africa some 60,000 years ago constituting a major bottleneck event, resulting in the genomes of non-Africans becoming more homogenous than the more diverse and sub-structured African genomes (31, 32). As a result, Africans exhibit in general more genetic diversity compared to non-Africans, which have more extended regions of LD (33). These diverging patterns of LD across population groups represent both challenges and opportunities, as they affect how well genotyping chips capture causal variants. Thus, variants in LD with a causal variant in Europeans might not be in LD with the same variant in Africans, which will affect the ability to replicate GWAS associations. However, differences in LD patterns may also make transethnic fine mapping especially powerful to discover the true causal variants. Genetic drift or local adaptation may also affect allele frequencies and thus the ability to replicate GWAS in other populations. That genetic drift and local adaption can influence the occurrence of variants affecting HIV traits is exemplified by the CCR5Δ32 deletion conferring resistance against HIV infection. The CCR5Δ32 deletion is only found at high frequencies in the European population but exhibits a north to south cline in allele frequency (34).

It is well documented that differences in immune responses exist between African and European individuals. Individuals of African descent display stronger inflammatory responses (35) and several non-genetic studies on HIV-related traits having found significant differences between African and European individuals; African individuals are more prone to develop broadly neutralizing antibodies (bNAbs) than European individuals (36) and tend to have smaller HIV reservoirs (37). The benefit of including more diverse populations into genetic studies is best exemplified by the case of LDL-cholesterol lowering loss-of-function variants found in *PCSK9* in Africans only (38), which has subsequently been developed into a drug benefiting all populations worldwide (39). Thus, focusing on African individuals in future studies may yield new insight into the pathogenesis of HIV and potential new drug targets.

The majority of GWAS performed worldwide so far have primarily focused on European (52%) or Asian (21%) populations (40). This trend is also seen in the field of HIV genetics, as also exemplified in this thesis, with few small GWAS including individuals of African descent (41, 42), despite HIV being much more prevalent on the African continent (43). There is no reason to believe that the burden of comorbidities will be lower in non-European countries with lower incomes. Rather, the lack of primary preventions and proper healthcare systems constitute a substantial emerging problem as the HIV population ages in these countries as well (44). The vast disparity between the number of sequenced or genotyped populations, disease prevalence, and genetic architecture may impede our understanding of these traits and the implementation of GRS for individuals of non-European ancestry.

Like the majority of GWAS, GRS have primarily been developed using genetic information from European populations. As a consequence, GRS have already been shown to perform worse in other population groups (45, 46), severely limiting their applicability in these groups. As the differences in LD cannot currently be solved computationally, the ideal solution would be to establish large GWAS projects in non-European populations, which would serve as a reference for population-specific GRS. Another approach would be to recalibrate either the scores or variant weights according to the population group (47). While the latter approach is still to be validated and requires some population-specific genetic information, it may, in theory, constitute the most practical path forward to obtain the desired accuracy and inclusion for all population groups. Another important and often overlooked issue regarding population groups is the presence of many admixed individuals. In our increasingly connected world, more and more people will be born from parents originating from different population groups. Without the use of recalibrated GRS, how would a clinician or researcher know which GRS is best suited to the child of these parents? As the median of the GRS has been shown to differ significantly between populations (46), the use of either may alter the perceived genetic risk for an admixed individual tremendously.

Finally, despite the obvious benefit to everyone for including more diverse populations in genetic studies, several challenges remain. Recruitment of participants might be difficult in some regions, notably due to lack of finances, infrastructure, and trained personnel. In particular, proper infrastructure and training of personnel is needed to obtain reliable phenotype information, crucial for genetic studies. Without such investments, our understanding of the genetic architecture affecting HIV pathogenesis and the development of comorbidities will never be complete.

5.4 Future opportunities for the HIV genetic research community

Current insight from HIV research on the natural history of HIV infection has primarily been enabled by the availability of large historical cohorts with samples collected before the initiation of ART. However, since treatment guidelines have moved towards immediate treatment initiation upon diagnosis, these samples are becoming rarer, and future research relying on samples from ART naïve patients will become ever more difficult. However, the unique setup of several HIV cohorts, with close monitoring and regular follow-up of patients, means that they also constitute a unique resource for performing longitudinal studies of the development of comorbidities, expansion of somatic variants and response to interventions, above and beyond what is usually possible in most population cohorts. Furthermore, it could be possible to evaluate methods that better capture environmental differences such as transcriptome, proteome, metabolome, or microbiome analyses to decipher the genetic and environmental risk factors during HIV infection better together with GRS due to the ample availability of patient plasma and cell samples. However, the many types of comorbidities and their potential relation to prescribed drugs, means that collecting sufficient numbers of patients for GWAS to discover novel HIV-specific associations for comorbidities will be challenging and will require international collaborations across HIV cohorts. This requires the standardization of patient data collection, but also offers the great opportunity to increase the diversity of study participants.

5.5 References

1. S. G. Deeks, Shock and kill. *Nature* **487**, 439–440 (2012).

2. L. Sylla, *et al.*, If We Build It, Will They Come? Perceptions of HIV Cure-Related Research by People Living with HIV in Four U.S. Cities: A Qualitative Focus Group Study. *AIDS Res. Hum. Retroviruses* **34**, 56–66 (2017).

3. A. Kratka, *et al.*, HIV Cure Research: Risks Patients Expressed Willingness to Accept. *Ethics Hum. Res.* **41**, 23–34 (2019).

4. N. Bachmann, *et al.*, Determinants of HIV-1 reservoir size and long-term dynamics during suppressive ART. *Nat. Commun.* **10**, 3193 (2019).

5. H. Imamichi, *et al.*, Defective HIV-1 proviruses produce novel protein-coding RNA species in HIV-infected patients on combination antiretroviral therapy. *Proc. Natl. Acad. Sci.* **113**, 8783–8788 (2016).

6. R. T. Gandhi, *et al.*, Levels of HIV-1 persistence on antiretroviral therapy are not associated with markers of inflammation or activation. *PLOS Pathog.* **13**, e1006285 (2017).

7. K. M. Bruner, N. N. Hosmane, R. F. Siliciano, Towards an HIV-1 cure: measuring the latent reservoir. *Trends Microbiol.* **23**, 192–203 (2015).

8. F. Hodel, M. Patxot, T. Snäkä, A. Ciuffi, HIV-1 latent reservoir: size matters. *Future Virol.* **11**, 785–794 (2016).

9. A. V. Khera, S. Kathiresan, Genetics of coronary artery disease: discovery, biology and clinical translation. *Nat. Rev. Genet.* **18**, 331–344 (2017).

10. M. Gorski, *et al.*, 1000 Genomes-based meta-analysis identifies 10 novel loci for kidney function. *Sci. Rep.* **7**, 45040 (2017).

11. J. R. Cerhan, *et al.*, Genome-wide association study identifies multiple susceptibility loci for diffuse large B cell lymphoma. *Nat. Genet.* **46**, 1233–1238 (2014).

12. A. Sud, B. Kinnersley, R. S. Houlston, Genome-wide association studies of cancer: current insights and future perspectives. *Nat. Rev. Cancer* **17**, 692–704 (2017).

13. S. SH, *et al.*, *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues* (August 29, 2019).

14. P. Mohammadi, *et al.*, 24 Hours in the Life of HIV-1 in a T Cell Line. *PLOS Pathog.* **9**, e1003161 (2013).

15. B. J. Strober, *et al.*, Dynamic genetic regulation of gene expression during cellular differentiation. *Science* **364**, 1287–1290 (2019).

16. B. Piasecka, *et al.*, Distinctive roles of age, sex, and genetics in shaping transcriptional variation of human immune responses to microbial challenges. *Proc. Natl. Acad. Sci.* **115**, E488–E497 (2018).

17. A. G. Bick, *et al.*, Inherited Causes of Clonal Hematopoiesis of Indeterminate Potential in TOPMed Whole Genomes. *bioRxiv*, 782748 (2019).

18. S. Jaiswal, *et al.*, Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).

19. S. Jaiswal, *et al.*, Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *N. Engl. J. Med.* **377**, 111–121 (2017).

20. G. Guaraldi, *et al.*, Premature Age-Related Comorbidities Among HIV-Infected Persons Compared With the General Population. *Clin. Infect. Dis.* **53**, 1120–1126 (2011).

21. A. Mocroft, *et al.*, Development and Validation of a Risk Score for Chronic Kidney Disease in HIV Infection Using Prospective Cohort Data from the D:A:D Study. *PLOS Med.* **12**, e1001809 (2015).

22. D. Furman, *et al.*, Chronic inflammation in the etiology of disease across the life span. *Nat. Med.* **25**, 1822–1832 (2019).

23. J. Eugen-Olsen, *et al.*, Circulating soluble urokinase plasminogen activator receptor predicts cancer, cardiovascular disease, diabetes and mortality in the general population. *J. Intern. Med.* **268**, 296–308 (2010).

24. D. J. Eapen, *et al.*, Soluble Urokinase Plasminogen Activator Receptor Level Is an Independent Predictor of the Presence and Severity of Coronary Artery Disease and of Future Adverse Events. *J. Am. Heart Assoc.* **3**, e001118 (2014).

25. A. V. Khera, *et al.*, Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* (2018) https:/doi.org/10.1038/s41588-018-0183-z (August 13, 2018).

26. A. C. Fahed, *et al.*, Polygenic background modifies penetrance of monogenic variants conferring risk for coronary artery disease, breast cancer, or colorectal cancer. *medRxiv*, 19013086 (2019).

27. G. Abraham, *et al.*, Genomic prediction of coronary heart disease. *Eur. Heart J.* **37**, 3267–3278 (2016).

28. Natarajan Pradeep, *et al.*, Polygenic Risk Score Identifies Subgroup With Higher Burden of Atherosclerosis and Greater Relative Benefit From Statin Therapy in the Primary Prevention Setting. *Circulation* **135**, 2091–2101 (2017).

29. J. L. Mega, *et al.*, Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials. *The Lancet* **385**, 2264–2271 (2015).

30. Khera Amit V., *et al.*, Whole-Genome Sequencing to Characterize Monogenic and Polygenic Contributions in Patients Hospitalized With Early-Onset Myocardial Infarction. *Circulation* **139**, 1593–1602 (2019).

31. D. Gurdasani, *et al.*, The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517**, 327–332 (2015).

32. A. Prohaska, *et al.*, Human Disease Variation in the Light of Population Genomics. *Cell* **177**, 115–131 (2019).

33. S. A. Tishkoff, *et al.*, The Genetic Structure and History of Africans and African Americans. *Science* **324**, 1035–1044 (2009).

34. F. Libert, *et al.*, The Δccr5 Mutation Conferring Protection Against HIV-1 in Caucasian Populations Has a Single and Recent Origin in Northeastern Europe. *Hum. Mol. Genet.* **7**, 399–406 (1998).

35. J. Sanz, H. E. Randolph, L. B. Barreiro, Genetic and evolutionary determinants of human population variation in immune responses. *Curr. Opin. Genet. Dev.* **53**, 28–35 (2018).

36. P. Rusert, *et al.*, Determinants of HIV-1 broadly neutralizing antibody induction. *Nat. Med.* **22**, 1260–1267 (2016).

37. N. Bachmann, *et al.*, Determinants of HIV-1 reservoir size and long-term dynamics during suppressive ART. *Nat. Commun.* **10**, 1–11 (2019).

38. J. C. Cohen, E. Boerwinkle, T. H. Mosley, H. H. Hobbs, Sequence Variations in PCSK9, Low LDL, and Protection against Coronary Heart Disease. *N. Engl. J. Med.* **354**, 1264–1272 (2006).

39. R. T. Dadu, C. M. Ballantyne, Lipid lowering with PCSK9 inhibitors. *Nat. Rev. Cardiol.* **11**, 563–575 (2014).

40. G. Sirugo, S. M. Williams, S. A. Tishkoff, The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).

41. K. Pelak, *et al.*, Host Determinants of HIV-1 Control in African Americans. *J. Infect. Dis.* **201**, 1141–1149 (2010).

42. P. J. McLaren, *et al.*, Fine-mapping classical HLA variation associated with durable host control of HIV-1 infection in African Americans. *Hum. Mol. Genet.* **21**, 4334–4347 (2012).

43. UNAIDS, “UNAIDS Data 2019” (2019) (October 3, 2019).

44. S. G. Deeks, S. R. Lewin, D. V. Havlir, The end of AIDS: HIV infection as a chronic disease. *The Lancet* **382**, 1525–1533 (2013).

45. L. Duncan, *et al.*, Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10**, 1–9 (2019).

46. A. R. Martin, *et al.*, Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).

47. C. Márquez-Luna, P.-R. Loh, A. L. Price, Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genet. Epidemiol.* **41**, 811–823 (2017).

Chapter 6 Conclusions

Human genetic variation is well known to influence our susceptibility to infections as well as the pathogenesis of HIV during natural infection. The studies presented in this thesis represent some of the first genome-wide studies examining the genetic effects influencing HIV positive individuals in the era of widely adopted ART and provides direct evidence for the influence of genetic variation on the risk of developing HIV-related comorbidities.

In chapter 2, we examined the contribution of host genetic variations to the size of the HIV reservoir and its long-term dynamics during ART, since the establishment and slow decay rate of the HIV reservoir is the main barrier for obtaining a functional cure for HIV. Using both genotyping arrays and exome sequencing, we performed a comprehensive examination of the association of both common- and rare variants as well as CNVs with the size of the HIV reservoir and its decay rate. However, we did not find any genome-wide significant associations, indicating that human genetic variation has a limited influence on the size of the HIV reservoir and its decay rate during therapy.

In chapter 3, we conducted the first genome-wide association study of HIV-related NHL using data from three major HIV cohorts in France, Switzerland, and the USA. We identified significantly associated genetic variants linked to the chemokine CXCL12. Furthermore, our data indicate that the genetic risk of HIV-related NHL is distinct from that of the general population. These findings suggest a unique role of CXCL12 in the development of HIV-related NHL.

In chapter 4, we developed a GRS for HIV-related CKD and evaluated its potential as a marker of disease occurrence. We found that the GRS was capable of independently predicting CKD in HIV patients, with the effect size of an unfavorable GRS found to be similar to that of an unfavorable clinical D:A:D score or use of certain potentially nephrotoxic antiretroviral compounds. Furthermore, the results indicate that individuals with the most favorable GRS are protected against CKD, irrespective of the presence of other clinical risk factors as measured by the D:A:D score.

Altogether, the knowledge obtained here on the role of genetic variation and its influence on HIV positive patients during ART has the potential to improve patient care by identifying patients with an increased risk of developing specific comorbidities early, thus providing the possibility for early treatment interventions to further improve the longevity and quality of life of HIV infected individuals.

Curriculum Vitae

Christian Axel Wandall Thorball

Contact data

Name:	Christian Axel Wandall Thorball
Address:	Route de Berne 61, CH-1010 Lausanne
Mobile:	+41 789454587
E-mail:	christian@thorball.com
LinkedIn:	https://www.linkedin.com/in/cthorball
Born:	October 11, 1987, Copenhagen

Education

1993-03	Lyngby Private Skole, Lyngby
2002	Emerson Junior High School, Davis, CA, USA
2003-06	Virum Gymnasium, Virum, Denmark. Mathematical Track
2007	University of New England, Australia. Biomedical Science
2007-11	University of Copenhagen, BSc Biochemistry Specialization: Immunology and Molecular Genetics Bachelor thesis: Influence of Chronic Hepatitis C Virus on the Innate- and Adaptive Immune Response. Included approved application to the Committee on Research Ethics & development of the best IL-1beta ELISA assay available.
2011-13	University of Copenhagen, MSc Biochemistry, with Qualification Profile in Immunology
2012	ETH Zürich, MSc Microbiology and Immunology (Exchange)
2012-13	Master thesis, Institute of Microbiology, University Hospital Center Lausanne (CHUV), Switzerland, with Professor Amalio Telenti. Project: HIV-1 Latency: Biomarker Discovery & Validation

Independent Courses

2011 Presentation skills training course, Impact Factory, London

2013 New England Biolabs Molecular Biology Summer Workshop, MA, USA

2014 Prince2 Project Management, KnowledgeTrain, London, UK

Certifications

2014 Prince2 Foundation in project management, AXELOS Global Best Practice

2014 Prince2 Practioner, AXELOS Global Best Practice

Experience

2003-05 Webmaster, Alsensa ApS (Allergy Medtech Company)
Webpage development and maintenance etc.

2006-07 Research Assistant, Copenhagen University Hospital, Hvidovre Hospital, Clinical Research Centre

2007-19 Consultant, ViroGates A/S
In charge of analysing suPAR in blood samples (e.g. HIV, HCV, TB & sepsis), and a research project with a general practitioner. Ad-hoc production, QC and QA of the suPARnostic© ELISA. Training of several bio and MD students. Carrying through clinical research studies in Guinea-Bissau and Atlanta, USA.

2009-10 Consultant, Your Global Eye ApS
Recruitment and instruction of photographers throughout Europe. Development of business model and webpage.

2011-13 Student Assistant, Marketing & Communication, Exiqon A/S
Location of new sales leads, updating of scientific material, webpage maintenance and ad-hoc tasks.

2014 PhD student, Institute of Microbiology, University Hospital Center Lausanne (CHUV) and University of Lausanne, Switzerland.
With Professor Amalio Telenti.

2014-15 Research Assistant, EPFL, Lausanne, Switzerland
With Professors Jacques Fellay & Didier Trono

2015-2020 PhD student, EPFL, Lausanne, Switzerland
With Professors Jacques Fellay & Didier Trono
Performing full analyses of genotyping data, whole genome and exome sequencing data. Imputation and fine mapping of HLA alleles as well as the development of polygenic risk scores. Genetic characterization of the KRAB-ZFP transcription factor family

Awards

2017	Keystone Symposia (HIV Vaccines) Scholarship

Language Skills

Danish & English	Fluent
German & French	Basic level

Computer Skills

Microsoft office	Advanced
R	Advanced
Python	Basic level

Sport

Cycling (road amateur races and endurance events)

Alpine- and cross-country skiing

Tennis (previous Danish national competition level)

Publications

An, P., Penugonda, S., **Thorball, C.W**., Bartha, I., Goedert, J.J., Donfield, S., Buchbinder, S., Binns-Roemer, E., Kirk, G.D., Zhang, W., Fellay, J., Yu, X.-F., Winkler, C.A., 2016. Role of APOBEC3F Gene Variation in HIV-1 Disease Progression and Pneumocystis Pneumonia. PLOS Genet. 12, e1005921. https://doi.org/10.1371/journal.pgen.1005921

Bachmann, N., Siebenthal, C. von, Vongrad, V., Turk, T., Neumann, K., Beerenwinkel, N., Bogojeska, J., Fellay, J., Roth, V., Kok, Y.L., **Thorball, C.W.,** Borghesi, A., Parbhoo, S., Wieser, M., Böni, J., Perreau, M., Klimkait, T., Yerly, S., Battegay, M., Rauch, A., Hoffmann, M., Bernasconi, E., Cavassini, M., Kouyos, R.D., Günthard, H.F., Metzner, K.J., 2019. Determinants of HIV-1 reservoir size and long-term dynamics during suppressive ART. Nat. Commun. 10, 3193. https://doi.org/10.1038/s41467-019-10884-9

Dietrich, L.G., Barceló, C., **Thorball, C.W**., Ryom, L., Burkhalter, F., Hasse, B., Furrer, H., Weisser, M., Steffen, A., Bernasconi, E., Cavassini, M., de Seigneux, S., Csajka, C., Fellay, J., Ledergerber, B., Tarr, P.E., 2019. Contribution of genetic background and clinical D:A:D risk score to chronic kidney disease in Swiss HIV-positive persons with normal baseline estimated glomerular filtration rate. Clin. Infect. Dis. https://doi.org/10.1093/cid/ciz280

Eapen, D.J., Manocha, P., Ghasemzadeh, N., Patel, R.S., Kassem, H.A., Hammadah, M., Veledar, E., Le, N.-A., Pielak, T., **Thorball, C.W**., Velegraki, A., Kremastinos, D.T., Lerakis, S., Sperling, L., Quyyumi, A.A., 2014. Soluble Urokinase Plasminogen Activator Receptor Level Is an Independent Predictor of the Presence and Severity of Coronary Artery Disease and of Future Adverse Events. J. Am. Heart Assoc. 3, e001118. https://doi.org/10.1161/JAHA.114.001118

Geraghty, D.E., **Thorball, C.W.**, Fellay, J., Thomas, R., 2019. Effect of Fc Receptor Genetic Diversity on HIV-1 Disease Pathogenesis. Front. Immunol. 10. https://doi.org/10.3389/fimmu.2019.00970

Ghasemzedah, N., Hayek, S.S., Ko, Y.-A., Eapen, D.J., Patel, R.S., Manocha, P., Al Kassem, H., Khayata, M., Veledar, E., Kremastinos, D., **Thorball, C.W**., Pielak, T., Sikora, S., Zafari, A.M., Lerakis, S., Sperling, L., Vaccarino, V., Epstein, S.E., Quyyumi, A.A., 2017. Pathway-Specific Aggregate Biomarker Risk Score Is Associated With Burden of Coronary Artery Disease and Predicts Near-Term Risk of Myocardial Infarction and Death. Circ. Cardiovasc. Qual. Outcomes 10. https://doi.org/10.1161/CIRCOUTCOMES.115.001493

Haastrup, E., Grau, K., Eugen-Olsen, J., **Thorball, C**., Kessing, L.V., Ullum, H., 2014. Soluble Urokinase Plasminogen Activator Receptor as a Marker for Use of Antidepressants. PLOS ONE 9, e110555. https://doi.org/10.1371/journal.pone.0110555

Haedersdal, S., Salvig, J.D., Aabye, M., **Thorball, C.W**., Ruhwald, M., Ladelund, S., Eugen-Olsen, J., Secher, N.J., 2013. Inflammatory Markers in the Second Trimester Prior to Clinical Onset of Preeclampsia, Intrauterine Growth Restriction, and Spontaneous Preterm Birth. Inflammation 36, 907–913. https://doi.org/10.1007/s10753-013-9619-x

Haupt, T.H., Kallemose, T., Ladelund, S., Rasmussen, L.J.H., **Thorball, C.W**., Andersen, O., Pisinger, C., Eugen-Olsen, J., 2014. Risk Factors Associated with Serum Levels of the Inflammatory Biomarker Soluble Urokinase Plasminogen Activator Receptor in a General Population. Biomark. Insights 2014, 91–100. https://doi.org/10.4137/BMI.S19876

Haupt, T.H., Petersen, J., Ellekilde, G., Klausen, H.H., **Thorball, C.W.**, Eugen-Olsen, J., Andersen, O., 2012. Plasma suPAR levels are associated with mortality, admission time, and Charlson Comorbidity Index in the acutely admitted medical patient: a prospective observational study. Crit. Care 16, R130. https://doi.org/10.1186/cc11434

Helleboid, P.-Y., Heusel, M., Duc, J., Piot, C., **Thorball, C.W**., Coluccio, A., Pontis, J., Imbeault, M., Turelli, P., Aebersold, R., Trono, D., 2019. The interactome of KRAB zinc finger proteins reveals the evolutionary history of their functional diversification. EMBO J. 0, e101220. https://doi.org/10.15252/embj.2018101220

Mekonnen, G., Corban, M.T., Hung, O.Y., Eshtehardi, P., Eapen, D.J., Al-Kassem, H., Rasoul-Arzrumly, E., Gogas, B.D., McDaniel, M.C., Pielak, T., **Thorball, C.W**., Sperling, L., Quyyumi, A.A., Samady, H., 2015. Plasma soluble urokinase-type plasminogen activator receptor level is independently associated with coronary microvascular function in patients with non-obstructive coronary artery disease. Atherosclerosis 239, 55–60. https://doi.org/10.1016/j.atherosclerosis.2014.12.025

Mohammadi, P., Iulio, J. di, Muñoz, M., Martinez, R., Bartha, I., Cavassini, M., **Thorball, C.**, Fellay, J., Beerenwinkel, N., Ciuffi, A., Telenti, A., 2014. Dynamics of HIV Latency and Reactivation in a Primary CD4+ T Cell Model. PLOS Pathog. 10, e1004156. https://doi.org/10.1371/journal.ppat.1004156

Mölkänen, T., Ruotsalainen, E., **Thorball, C.W**., Järvinen, A., 2011. Elevated soluble urokinase plasminogen activator receptor (suPAR) predicts mortality in Staphylococcus aureus bacteremia. Eur. J. Clin. Microbiol. Infect. Dis. 30, 1417–1424. https://doi.org/10.1007/s10096-011-1236-8

Nguyen, H., **Thorball, C.W.**, Fellay, J., Böni, J., Yerly, S., Perreau, M., Klimkait, T., Kusejko, K., Bachmann, N., Chaudron, S.E., Paioni, P., Thurnheer, M.C., Battegay, M., Cavassini, M., Vernazza, P., Bernasconi, E., Günthard, H.F., Kouyos, R., Study, the S.H.C., 2019. HIV Transmission Chains Exhibit Greater HLA-B Homogeneity Than Randomly Expected. JAIDS J. Acquir. Immune Defic. Syndr. 81, 508. https://doi.org/10.1097/QAI.0000000000002077

Power, R.A., **Thorball, C.W**., Bartha, I., Perry, J.R.B., McLaren, P.J., Oliveira, T. de, Fellay, J., 2017. A genome-wide polygenic approach to HIV uncovers link to inflammatory bowel disease and identifies potential novel genetic variants. bioRxiv 145383. https://doi.org/10.1101/145383

Rusert, P., Kouyos, R.D., Kadelka, C., Ebner, H., Schanz, M., Huber, M., Braun, D.L., Hozé, N., Scherrer, A., Magnus, C., Weber, J., Uhr, T., Cippa, V., **Thorball, C.W**., Kuster, H., Cavassini, M., Bernasconi, E., Hoffmann, M., Calmy, A., Battegay, M., Rauch, A., Yerly, S., Aubert, V., Klimkait, T., Böni, J., Fellay, J., Regoes, R.R., Günthard, H.F., Trkola, A., The Swiss HIV Cohort Study, Bucher, H.C., Ciuffi, A., Dollenmaier, G., Egger, M., Elzi, L., Fehr, J., Furrer, H., Fux, C.A., Haerry, D., Hasse, B., Hirsch, H.H., Hösli, I., Kahlert, C., Kaiser, L., Keiser, O., Kovari, H., Ledergerber, B., Martinetti, G., Tejada, B.M. de, Marzolini, C., Metzner, K.J., Müller, N., Nicca, D., Pantaleo, G., Paioni, P., Rudin, C., Schmid, P., Speck, R., Stöckle, M., Tarr, P., Vernazza, P., Wandeler, G., Weber, R., 2016. Determinants of HIV-1 broadly neutralizing antibody induction. Nat. Med. 22, 1260–1267. https://doi.org/10.1038/nm.4187

Scepanovic, P., Alanio, C., Hammer, C., Hodel, F., Bergstedt, J., Patin, E., **Thorball, C.W**., Chaturvedi, N., Charbit, B., Abel, L., Quintana-Murci, L., Duffy, D., Albert, M.L., Fellay, J., 2018. Human genetic variants and age are the strongest predictors of humoral immune responses to common pathogens and vaccines. Genome Med. 10, 59. https://doi.org/10.1186/s13073-018-0568-8

Takahashi, N., Coluccio, A., **Thorball, C.W.**, Planet, E., Shi, H., Offner, S., Turelli, P., Imbeault, M., Ferguson-Smith, A.C., Trono, D., 2019. ZNF445 is a primary regulator of genomic imprinting. Genes Dev. 33, 49–54. https://doi.org/10.1101/gad.320069.118

Thorball, C.W., Borghesi, A., Bachmann, N., Siebenthal, C. von, Vongrad, V., Turk, T., Neumann, K., Beerenwinkel, N., Bogojeska, J., Roth, V., Kok, Y.L., Parbhoo, S., Wieser, M., Böni, J., Perreau, M., Klimkait, T., Yerly, S., Battegay, M., Rauch, A., Schmid, P., Bernasconi, E., Cavassini, M., Kouyos, R., Günthard, H.F., Metzner, K., Fellay, J., Study, the S.H.C., 2019a. Host genomics of the HIV-1 reservoir size and its decay rate during suppressive antiretroviral treatment. medRxiv 19013763. https://doi.org/10.1101/19013763

Thorball, C.W., Oudot-Mellakh, T., Hammer, C., Santoni, F.A., Niay, J., Costagliola, D., Goujard, C., Meyer, L., Wang, S.S., Hussain, S.K., Theodorou, I., Cavassini, M., Rauch, A., Battegay, M., Hoffmann, M., Schmid, P., Bernasconi, E., Günthard, H.F., McLaren, P.J., Rabkin, C.S., Besson, C., Fellay, J., 2019b. Genetic variation near CXCL12 is associated with susceptibility to HIV-related non-Hodgkin lymphoma. medRxiv 19011999. https://doi.org/10.1101/19011999

Turelli, P., Playfoot, C., Grun, D., Raclot, C., Pontis, J., Coudray, A., **Thorball, C**., Duc, J., Pankevich, E.V., Deplancke, B., Busskamp, V., Trono, Didier, 2019. Primate-restricted KRAB zinc finger proteins and target retrotransposons control gene expression in human neurons. bioRxiv 856005. https://doi.org/10.1101/856005