

Exploring overlooked interactions involved in missing heritability

Mio Shibata

McGill-Kyoto International Joint Ph.D. Program in Genomic Medicine

Department of Human Genetics, McGill University, Montreal

Graduate School of Medicine and Faculty of Medicine, Kyoto University, Kyoto

February 2026

A thesis submitted to Kyoto University in partial fulfillment
of the requirements of the degree of

Doctoral Philosophy

© Mio Shibata, 2026

Abstract

Human health status and disease onset are shaped by both genetic and environmental factors through intricate interactions. An accurate estimate of heritability—the proportion of phenotypic variance due to genetic differences—is important for understanding the mechanisms underlying physiological changes and disease. High heritability for many traits has been reported by twin and family studies, leading to the expectation that this heritability would be replicated by genome-wide association studies (GWAS), which identify significant genetic loci at the population level. However, conventional GWAS based on SNPs explain much less heritability than estimates from twin and family studies, a phenomenon known as “missing heritability”, which may obscure the true relationship between phenotype and genetic factors. Some of the “missing factors” that are thought to underlie this discrepancy include: the incomplete capture of relevant genetic contributions in previous GWAS, and the misspecification of gene-environment relationships due to insufficient consideration of population background.

To improve heritability estimates and thereby enhance the validity of genetic association, it is essential to identify the missing factors and incorporate them appropriately. In this thesis, I hypothesize that two key components are required to address the issue of missing heritability: A) the identification of currently unrecognized genetic or environmental factors, and B) an optimized model capable of incorporating these factors.

Focusing on genetic factors, some of the missing factors are thought to involve elements that are difficult to detect using conventional GWAS due to their combinatorial complexity and low frequency—particularly epistatic interactions with synergistic effects. Additionally, since GWAS generally assumes a linear effect model, an alternative modeling is required to capture such synergistic effects. The first study in this thesis focuses on applying GWAS to identify

epistatic interactions in Japanese rheumatoid arthritis (RA) studies. The GWAS identified 90 high-order SNP combinations associated with RA, each involving three or more SNPs in the discovery cohort. In the validation cohort, 82.2% of these combinations (74 out of 90) were successfully replicated, and one combination composed of SNPs located on different chromosomes exhibited a synergistic effect as an optimized genetic model.

Additionally, I explored the relationship between viral infection, as a representative environmental factor, and host genetics since it is known to be associated with health status and diseases onset. The prevalence of viruses in human cells, particularly in the case of latent infections, varies across geographic and host genetic backgrounds, potentially influencing heritability estimation. In the second study of this thesis, I focused on viral quantities in human blood cells detected by whole-genome sequencing (WGS) across cohorts from Europe, Australia, Canada, and Japan. Comparative association analyses based on geographic and genetic differences showed that both backgrounds contribute to the prevalence of DNA viruses. To investigate host genetic factors, GWAS were conducted and identified significant associations in the MHC region. Subsequent *HLA* typing further revealed specific alleles and amino acids residues associated with the viral prevalence.

These findings reveal missing factors spanning relationships from host genetic variants to gene-environment contexts, potentially underlying missing heritability. The complexity and close interrelationship between the detection of missing factors and the construction of appropriate models highlight the need for their simultaneous evaluation in accurate heritability estimation. By systematically accounting for missing factors in intra-genetic and genetic-environment interactions, this thesis addresses some of the limitations of conventional GWAS and can serve as a tool to improve future studies.

Résumé

L'état de santé humain et l'apparition de maladies sont déterminés par des facteurs génétiques et environnementaux via des interactions complexes. Une estimation précise de l'héritabilité — la proportion de la variance phénotypique attribuable aux différences génétiques — est essentielle pour comprendre les mécanismes sous-jacents aux changements physiologiques et aux maladies. Des études de jumeaux et de liaisons ont révélé une forte héritabilité pour de nombreux traits. On s'attendait à ce que cette héritabilité soit répliquée par des études d'association pangénomique (GWAS), identifiant des loci génétiques significatifs au niveau populationnels. Cependant, les GWAS classiques fondées sur les SNPs expliquent moins d'héritabilité que les estimations issues des études de jumeaux et de liaison, un phénomène connu sous le nom d'« héritabilité manquante », qui peut masquer la véritable relation entre le phénotype et les facteurs génétiques. Différents « facteurs manquants » seraient à l'origine de cette discordance, incluant : la prise en compte limitée des facteurs génétiques appropriés dans les GWAS précédentes, et une spécification erronée des relations gènes-environnement en raison du manque de considération du contexte populationnel.

Pour améliorer les estimations de l'héritabilité et donc renforcer la validité des associations génétiques, il est essentiel d'identifier les facteurs manquants et de les intégrer de manière appropriée. Cette thèse, émet l'hypothèse que deux éléments clés sont nécessaires pour aborder la question de l'héritabilité manquante : A) l'identification de facteurs génétiques ou environnementaux actuellement non pris en compte, et B) l'élaboration de modèles optimisés capable d'intégrer ces facteurs.

Concernant les facteurs génétiques, on suppose que certains facteurs manquants concernent des éléments difficiles à détecter par les GWAS conventionnelles, en raison de leur complexité combinatoire et de leur basse fréquence — en particulier les interactions épistatiques à effets synergiques. Par ailleurs, les GWAS reposent généralement sur un modèle d'effet linéaire,

nécessitant une modélisation alternative pour capturer de tels effets. La première étude de cette thèse porte sur l'application des GWAS pour l'identification d'interactions épistatiques dans des études japonaises sur la polyarthrite rhumatoïde (PR). Ces analyses ont identifié, dans la cohorte de découvertes, 90 combinaisons de SNP d'ordre élevé associées à la PR, chacune impliquant trois SNP ou plus. Dans la cohorte de validation, 82,2 % de ces combinaisons (74 sur 90) ont été répliqués avec succès. Une combinaison, composée de SNP localisés sur différents chromosomes, a présenté un effet synergique, constituant un modèle génétique optimisé.

Nous avons également exploré la relation entre l'infection virale — en tant que facteur environnemental représentatif — et la génétique de l'hôte car elle est étroitement liée à l'état de santé et à l'apparition des maladies. La prévalence des virus dans les cellules humaines, notamment dans le cas d'infections latentes, varie selon l'origine géographique et le bagage génétique de l'hôte influençant l'estimation de l'héritabilité. La deuxième étude de cette thèse se consacre aux quantités virales présentes dans les cellules sanguines humaines, détectées par séquençage du génome entier (WGS) dans des cohortes issues d'Europe, d'Australie, du Canada et du Japon. Les analyses comparatives d'association, axées sur les différences géographiques et génétiques, ont révélé que ces deux facteurs contribuent à la prévalence des virus à ADN. Pour étudier les facteurs génétiques de l'hôte, des GWAS ont été menées et ont identifié des associations significatives dans la région du MHC. Le typage HLA effectué ultérieurement a permis d'identifier des allèles spécifiques et des résidus d'acides aminés associés à la prévalence virale.

Ces résultats révèlent des facteurs manquants impliquant des relations allant des variants génétiques de l'hôte aux contextes gène-environnement, qui pourraient être à l'origine de l'héritabilité manquante. La complexité et l'étroite interdépendance entre la détection de ces facteurs et la construction de modèles appropriés soulignent la nécessité de les évaluer

conjointement pour une estimation précise de l'héritabilité. En prenant systématiquement en compte certains facteurs manquants dans les interactions intra-génétiques et gène-environnement, cette thèse aborde les limites des GWAS conventionnelles et pourra contribuer à améliorer leur cadre méthodologique.

Table of Contents

Abstract.....	i
Résumé	iii
Table of Contents	vi
List of Abbreviations.....	ix
List of Figures.....	xi
List of Tables.....	xiii
Acknowledgements	xiv
Contribution of original knowledge	xv
Format of the thesis	xvi
Contributions of Authors	xvii
1. Introduction.....	1
1.1 Overview	1
1.2 Epistatic interaction in complex disease	4
1.3 Identification of interactions across host and virus genomes	8
1.4 New strategy for GWAS and accurate heritability estimation (Objective and hypothesis)	13
2. Identification of epistatic SNP combinations in rheumatoid arthritis using LAMPLINK and Japanese cohorts	15
2.1. Abstract.....	16
2.2. Introduction.....	16

2.3.	Materials and methods	19
2.4.	Results	24
2.5.	Discussion	31
2.6.	Conflict of interest.....	33
2.7.	Acknowledgements	33
2.8.	References	34
2.9.	Supplementary information	37
Connecting Text: Bridge Between Chapter 2 and Chapter 3.....		38
3. Geographic and genetic influences on the human blood DNA virome.....		39
3.1.	Abstract.....	40
3.2.	Introduction.....	41
3.3.	Results	43
3.4.	Discussion	61
3.5.	Methods.....	64
3.6.	Acknowledgements	83
3.7.	Author contributions	83
3.8.	References	83
3.9.	Supplementary Figures	90
4. General discussion		104
4.1.	Addressing the missing heritability	104
4.2.	Components overlooked by traditional heritability	105
4.3.	Modelling strategies for incorporating overlooked components	107
4.4.	Entangled effects of missing factors and modelling assumptions.....	109

4.5. Design of the study incorporating the missing factors and models	110
5. Conclusions and Future Directions	113
6. Master Reference List.....	115

List of Abbreviations

AFR	African
AIC	Akaike Information Criterion
AMR	Admixed American
CDF	Cumulative Distribution Function
CSA	Central/South Asian
DP	Depth
DR	Direct Repeat
DZ	Dizygotic
EAS	Eastern Asian
EBV	Epstein-Barr virus
EEA	Equal Environmental Assumption
EUR	European
FWER	Family-wise Error Rate
G-DOMAINS	Groove Domains
gnomAD	Genome Aggregation Database
GQ	Genotype Quality
GWAS	Genome-wide association studies
HBV	Hepatitis B virus
HGDP	Human Genome Diversity Project
HHV	Human Herpesvirus
HLA	Human Leukocyte Antigen
HTLV-1	Human T-lymphotropic Virus Type 1
HWE	Hardy-Weinberg equilibrium
ici	Inherited Chromosomal Integrations
IFN	Interferon
k-NN	k-Nearest Neighbour
KDE	Kernel Density Estimation
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
MDR	Multifactor-dimensionality Reduction
MID	Middle Eastern
MZ	Monozygotic
NGS	Next-generation Sequencing
NPC	Nasopharyngeal Carcinoma
OR	Odds Ratio
PCA	Principal Component Analysis

PCs	Principal Components
PRS	Polygenetic Risk Score
RA	Rheumatoid Arthritis
SNPs	Single nucleotide polymorphisms
SNVs	Single Nucleotide Variants
VQSLOD	Variant Quality Score Log-odds
WTCCC	Wellcome Trust Case Control Consortium

List of Figures

Chapter 2

Fig. 1	25
Fig. 2	28
Fig. 3	29
Supplementary Fig. 1	37

Chapter 3

Fig. 1 Demographic and genetic characteristics	45
Fig. 2 Workflow of viral read detection and association analyses	47
Fig. 3 Differences in viral quantities across geographic backgrounds or SARS-CoV-2 infection status	52
Fig. 4 Genetic associations of Anellovirus with host factors	57
Fig. 5 Genetic associations of EBV/HHV-4 with host factors	59
Supplementary Fig. 1 WGS results and demographics across diverse cohorts.	90
Supplementary Fig. 2 Average coverage of iciHHV-6A/B viral reads	91
Supplementary Fig. 3 Detected artificial reads and coverage depth of mapped reads	92
Supplementary Fig. 4 Evaluation of the correction method for viral quantification.	93
Supplementary Fig. 5 The correlation between age and abundance of detected viruses	94
Supplementary Fig. 6 Genetic associations of HHV-7 with host factors	95
Supplementary Fig. 7 HLA haplotypes associated with viral prevalence	96
Supplementary Fig. 8 Association analysis of amino acid residues with Anellovirus prevalence	97
Supplementary Fig. 9 Association analysis of amino acid residues with EBV/HHV-4	

prevalence	98
Supplementary Fig. 10 Association analysis of amino acid residues with HHV-7 prevalence	99
Supplementary Fig. 11 Linkage patterns of amino acid residues associated with the prevalence of EBV/HHV-4.	100
Supplementary Fig. 12 The changes in genotype concordance rate based on DP per genotype in each cohort.	101
Supplementary Fig. 13 Parameters for population and sex inference	102
Supplementary Fig. 14 The illustration of an unbiased comparison of viral prevalence and abundance.	103

List of Tables

Chapter 2

Table 1.....	26
--------------	----

Chapter 3

Table 1.....	55
--------------	----

Supplementary Tables are provided in this single Excel workbook. Sheets follow the naming convention "chapterNumber-SupplementaryTable<tableNumber>"

Acknowledgements

I would like to express my deepest gratitude to my supervisors, Dr. Guillaume Bourque, Dr. Ryo Yamada, and Dr. Fumihiko Matsuda, whose guidance, encouragement, and invaluable advice made it possible for my Ph.D. journey to encompass two distinct research projects conducted in Canada and Japan, thereby broadening the international dimension of my doctoral program and making it truly unique. I express my sincere appreciation to my mentors, Dr. Shuji Kawaguchi, Dr. Yoichiro Kamatani, Dr. Mathieu Bourgey, and Dr. Emmanuel Gonzalez. Their dedicated support and invaluable advice helped me to overcome the challenges and move forward with my research from the early stages of my Ph.D. I am also thankful to the supervisors committee members of my thesis, Dr. Simon Gravel and Dr. Sylvia Vidal, who gave consistent feedback, guiding my projects smoothly.

These research projects were accomplished thanks to the wonderful collaborators and laboratory members I had the privilege to work with. I am truly grateful to my friends, whose support, discussions, and encouragement sustained my motivation and helped me pursue my Ph.D. across two countries.

For the financial support, I would like to thank the Japan Student Services Organization and the Tobitate Japan scholarship program, and the Japan Science and Technology Agency SPRING, Grant Number JPMJSP2110.

Lastly, and most importantly, I would like to thank my family for their unwavering encouragement and understanding.

Contribution of original knowledge

The presence of missing heritability limits the accuracy of disease association studies based on genetic factors. To improve the accuracy of heritability estimates, identifying the overlooked components—missing factors—is essential.

Chapter 2 detected epistatic interactions associated with RA using Japanese case-control cohorts and optimized the corresponding genetic model. This study contributes to addressing missing heritability by identifying previously overlooked genetic factors.

Chapter 3 analyzed WGS data from multiple cohorts (Australia, Europe, Canada, and Japan), quantifying viral reads and exploring their associations with geographic and genetic background as well as disease status (SARS-CoV-2 infection). In addition, the study identified amino acid residues associated with the presence or absence of viral reads in WGS data. This work contributed to identifying missing factors arising from the interaction between environmental and host genetic factors.

Format of the thesis

This doctoral thesis is presented in a manuscript-based format, as permitted by the regulations set by McGill-Kyoto International Joint Ph.D. Program in Genomic Medicine.

The studies described in this thesis were conducted under the supervision of Dr. Guillaume Bourque, Dr. Ryo Yamada, and Dr. Fumihiko Matsuda, and focus on uncovering missing contributors to heritability, with particular attention to epistatic and host–environment interactions based on genomic data. Chapter 1 introduces the concept of missing heritability, summarizing prior research and outlining possible missing factors and the challenges in detecting them. Chapter 2 is a published article (*Journal of Human Genetics*, 2024) that investigates epistatic effects associated with rheumatoid arthritis. Chapter 3 is a manuscript in preparation that explores associations between viral sequences detected from whole-genome sequencing and host-related variables such as geographic origin, population structure, and genotype as host–environment interactions. Chapter 4 discusses how findings from Chapters 2 and 3 relate to each other and considers their potential impact. Chapter 5 concludes the thesis and outlines future directions.

Contributions of Authors

Chapter 2 is presented as it appeared in *Journal of Human Genetics* (July 2024), and co-authored with: Aika Terada, Takahisa Kawaguchi, Yoichiro Kamatani, Daigo Okada, Kazuhisa Nagashima, Koichiro Ohmura, Fumihiko Matsuda, Shuji Kawaguchi, Jun Sese, Ryo Yamada, and myself. I performed the analysis and wrote the manuscripts with the help of Shuji Kawaguchi. Aika Terada, Yoichiro Kamatani, and Jun Sese tested the associations for the discovery cohort. Daigo Okada and Kazuhisa Nagashima performed the preliminary data analysis for the validation cohort. Takahisa Kawaguchi contributed to refining the analytical approach. Koichiro Ohmura and Fumihiko Matsuda were responsible for the recruitment of subjects and the collection of genomic data in the validation cohort. Shuji Kawaguchi, Jun Sese, and Ryo Yamada designed and supervised the project.

Chapter 3 is a manuscript currently in preparation, co-authored with: Shuji Kawaguchi, Mathieu Bourgey, Takahisa Kawaguchi, Yuichi Inadomi, Emmanuel Gonzalez, Robert Eveleigh, Pascale Marquis, Hans Grabe, Karen Mather, Helena Schmidt, Stephanie Debette, Ryo Yamada, Yoichiro Kamatani, Fumihiko Matsuda, Guillaume Bourque, and myself. Mathieu Bourgey, Pascale Marquis, and I executed the DNA-seq pipeline. Shuji Kawaguchi and I developed the methodology. I implemented analysis scripts, including those for the developed methods, and conducted procedures across the entire pipeline. Mathieu Bourgey, Takahisa Kawaguchi, Yuichi Inadomi, Emmanuel Gonzalez, Robert Eveleigh, Stephanie Debette, Ryo Yamada, and Yoichiro Kamatani contributed to refining the approach to interpreting the results. Hans Grabe, Karen Mather, Helena Schmidt, Stephanie Debette, and Fumihiko Matsuda collected and managed the samples and data construction. I wrote the manuscript with the help of Shuji Kawaguchi. Fumihiko Matsuda and Guillaume Bourque reviewed and provided feedback.

1. Introduction

1.1 Overview

1.1.1 The development and application of GWAS in genomic medicine

Elucidating the relationship between genetic variants and phenotypes is crucial for unraveling the mechanisms underlying disease onset and progression. Family-based analyses, which rely on detailed pedigrees, have been instrumental in identifying causative genes for diseases with strong familial inheritance, such as Huntington's disease [1]. However, this approach is less effective for diseases that occur in unrelated individuals and are influenced by multiple genetic loci, such as autism [2]. To address such cases, genomic studies have increasingly adopted a genome-wide approach that leverages genetic variants like single nucleotide polymorphisms (SNPs), to scan the genome for associations with disease. In these studies, known as genome-wide association studies (GWAS), statistical significance is determined by comparing the frequency of each SNP between cases (individuals with the disease) and controls (individuals without the disease).

At first, SNPs used in GWAS were measured using PCR-based assays. For example, a study in Japan in 2002 analyzed 92,788 SNPs and identified significant associations between two SNPs in the *LTA* gene and myocardial infarctions in a cohort of 1,133 cases and 1,006 controls [3]. Subsequently, GWAS studies shifted to microarray-based SNP detection methods. In 2005, a case and control study of Age-related Macular Degeneration investigated 116,204 SNPs using microarray and found significant associations with two SNPs in the *CFH* gene based on 96 cases and 50 controls [4]. As the technology advanced to measure more SNPs, the need for multiple-test correction required a more conservative threshold to achieve statistical power to help distinguish between spurious and true associations. Increasing sample sizes has been regarded as an effective strategy to address this challenge. Therefore, later studies expanded

both the sample size and SNP density for statistical testing. For example, The Wellcome Trust Case Control Consortium (WTCCC) measured 500,000 SNPs in 14,000 cases and 3,000 controls and identified significant associations with one SNP in bipolar disorder, one SNP in coronary artery disease, nine SNPs in Crohn's disease, three SNPs in rheumatoid arthritis (RA), seven SNPs in type 1 diabetes and three SNPs in type 2 diabetes [5].

1.1.2 Missing heritability

Observational studies in epidemiology have demonstrated that environmental factors such as diet [6], smoking [7], alcohol consumption [8,9], and infectious diseases [10,11] significantly contribute to disease progression. However, genetic factors have also been considered to influence the progression of the disease. Heritability assesses the extent to which genetic factors explain phenotypic variation and is defined as the proportion of this variation explained by additive genetic effects, which had been established before GWAS developed [12]. In the early stages, heritability was estimated in classical twin studies as additive narrow-sense heritability by comparing correlation coefficients between monozygotic (MZ) and dizygotic (DZ) twin pairs based on the premise that MZ twins share all their genetic material and environmental influences, whereas DZ twins share the same environment but only approximately half of their genetic material [13,14]. This assumption of equal environmental sharing within a family became known as the Equal Environmental Assumption (EEA), based on a study showing minimal differences between twins whose parents misclassified their zygosity [15]. Following this approach, a twin study in Australia estimated that the heritability of height was 0.91 based on an additive genetic model [16]. However, when genetic studies attempt to identify specific genetic factors contributing to phenotypic variation, they often explain only a small fraction of the expected heritability. For instance, the genome-wide significant SNPs associated with height identified in early GWAS accounted for just 0.02 to

0.04 of the additive phenotypic variance in population-scale analyses of unrelated participants [17–19]. This discrepancy between the heritability estimated from twin studies and the proportion explained by genome-wide significant SNPs is referred to as "missing heritability" [20].

The concept of "missing heritability" in genetics is thought to stem from two primary factors [21]: 1) missing factors and 2) modelling limitations. First, heritability estimates from twin studies accounted for all genetic factors, encompassing both common and rare genetic loci with varying effect sizes, which has already been pointed out in the early stage of heritability estimates [14]. Meanwhile, conventional GWAS often fail to capture all genetic variants that may be associated with a phenotype. One approach to addressing missing heritability is to aggregate subthreshold SNPs and assess their combined effect. Even when considering the aggregate contribution of common genotyped SNPs, including loci that are not genome-wide significant, the SNP-based heritability of height only reached 0.45, well below the 0.91 estimated from twin studies [22]. This discrepancy partly reflects limitations of microarray-based studies: rare SNPs are generally not represented on standard genotyping arrays, and structural variants (SVs) are often technically difficult to detect; as a result, such variants are either not represented or are difficult to detect using microarray-based approaches [21]. On the other hand, in twin studies, gene-gene interactions (epistasis), where synergistic effects between genes may be absorbed into twin-based heritability estimates, with the true epistatic contributions remaining missing since twins share genotypes [23]. Since epistatic interactions are not comprehensively known, these epistatic interactions are typically not accounted for in GWAS. Second, from a modelling perspective, an individual effect size can vary substantially depending on the genetic model applied. The heritability estimated by SNPs may misestimate the true genetic contribution because the additive genetic model commonly used in GWAS captures only additive effects and may fail to account for non-additive interactions such as

dominance and epistasis, highlighting the uncertainty in capturing the full genetic architecture. In contrast, heritability estimates derived from twin studies may be inflated by both shared environmental influences that are correlated with genetic similarity, resulting in a violation of the EEA, and by gene–environment interactions [24]. In such cases, environmental factors may amplify or suppress genetic effects in a context-dependent manner, leading to an overestimation or underestimation of genetic contributions. For example, as demonstrated in a previous GWAS [25], genetic factors can promote or suppress smoking behavior, which is typically considered an environmental factor. However, such interactions are not fully incorporated into GWAS analyses due to limited information on the specific genetic and environmental factors involved, as well as the complexity of modelling these interactions.

Heritability estimates can also vary depending on environmental distributions, which are shaped by chronological or geographical differences, even within the same population. For instance, viral infections are significant environmental factors where both genetic susceptibility and regional spread of the virus contribute to phenotypic outcomes. The Human Leukocyte Antigen (HLA), with alleles that exhibit considerable diversity across geographic regions, has been reported to be associated with latent infections, such as Hepatitis B virus (HBV) [26] and Human T-lymphotropic virus type 1 (HTLV-1) [27]. Therefore, it is imperative to incorporate both genetic and environmental factors when trying to identify the association between those factors belonging to individuals, such as case-control studies.

1.2 Epistatic interaction in complex disease

1.2.1. Challenges in the discovery of epistatic interactions

1.2.1.1 Multiple testing correction for unlimited SNP combinations

Given the large number of SNPs analyzed in GWAS, the significance level α must be adjusted

to account for multiple testing. The commonly used Bonferroni correction controls the family-wise error rate (FWER) [28], and the significance level is expressed as $\alpha = \frac{0.05}{M}$, where M is the total number of tests performed. In practice, many GWAS studies employ a significance level of $\alpha = 5.0 \times 10^{-8} = \frac{0.05}{10^6}$. This is based on the estimated number of haplotype blocks (10^6) in the human genome defined by loci with low recombination rates [29].

In the context of epistatic SNP interactions, it is necessary to test combinations of two or more SNPs. As such, the total number of possible combinations with N SNPs is given by $2^N - 1$, which results in an excessively stringent implementation of the Bonferroni correction. To reduce the number of tests and shrink potential candidates, previous studies focused on significantly associated SNPs before exploring epistatic interactions with synergistic effects. However, this approach may overlook SNPs that exhibit significant associations only when interacting with other SNPs. For example, a cardiovascular study used gene function analyses and identified an epistatic interaction involving SNPs, including one that did not have a significant association with cardiovascular risk when considered individually [30].

1.2.1.2 Statistical modelling of the epistatic interactions

In GWAS, it is required to select an appropriate genetic model that aligns with the phenotypic characteristics under investigation. To estimate the genetic effects of independent variants, the linear-additive logistic model is often employed in conditional analyses in conventional GWAS, particularly in case-control studies. However, when assessing the synergistic effects arising from epistatic interactions, it is typically assumed that a non-additive and/or non-linear model is required. Unfortunately, these models are often not rigorously evaluated in conditional analyses [31]. Therefore, prior to exhibiting statistical significance, it is essential to optimize the model to prevent the selection of an arbitrary model. This can be achieved by utilizing

criteria such as the Akaike Information Criterion (AIC) to identify the most optimized model among the potential candidates [32].

1.2.2. Identification and evaluation of epistatic interactions

1.2.2.1. Effective screening of SNP combinations

As the number of SNPs and the order of interactions considered increase, the Bonferroni correction for multiple testing leads to a conservative significance threshold due to the large number of SNP combinations. This leads to a trade-off between discovering novel interactions and maintaining statistical power for detection. Consequently, the effective screening of meaningful SNP combinations is necessary to mitigate the impact of the lowered significance threshold. The process generally consists of two steps: 1) pruning or selecting SNPs, and 2) searching for SNP combinations. The specific order and iterations of these two steps depend on the applications. Since exhaustively searching for all possible combinations is computationally expensive, knowledge-based and statistical-based approaches have been implemented. Drawing on prior reports, the knowledge-based approach succeeded in detecting epistatic interactions by testing combinations of candidate SNPs previously reported in high-density lipoprotein cholesterol levels studies [33], as well as by aggregating autism-associated SNPs into biologically informed networks or pathway sets that were subsequently tested for interaction effects. [34]. However, this approach may introduce bias due to reliance on existing research, called publication bias. In contrast, the data-driven approach identifies interactions based on the frequency of co-occurrence of SNPs in the input data, thus avoiding publication bias. Ultimately, high-order interactions can be simplified to pairwise interactions to facilitate analysis. The BOOST algorithm, which can perform an exhaustive search for all SNPs [35], can reduce the total number of tests ${}_nC_2 = \frac{n \times (n-1)}{2}$. However, the availability of computing resources remains a limiting factor. To resolve this problem, pruning methods have been

developed to focus on the high joint frequency of SNPs in the high-LD regions, such as LINDEN [36], making them suitable for genetic datasets. However, these methods cannot exhaustively search for all interactions. The multifactor-dimensionality reduction (MDR) was developed to identify high-order interactions [37]. By utilizing algorithms developed in informatics, such as item mining algorithms, it becomes possible to avoid redundant testing by excluding combinations that are not worth testing. Among these, LAMPLINK [38] uses the LAMP [39,40] algorithm to efficiently prune the search space and enumerate only a reduced set of combinations, followed by identification of statistically significant ones after Bonferroni correction.

1.2.2.2 Optimizing statistical model for epistatic interactions

Several models can be used to test associations, including Fisher's exact probability test or Pearson's Chi-squared test, as applied in LINDEN [36] and LAMPLINK [38], as well as logistic regression in BOOST [35]. The effect sizes of epistatic interactions are typically expressed through genotype models, which are chosen arbitrarily, depending on the application. In contrast, the application of MDR can identify high-risk genotype models from contingency tables of genotypes for each SNP by comparing the cases and controls. These models are evaluated through cross-validation and permutation tests [37], with the added constraint of limiting the number of SNPs or genes [41]. Although genetic models are dependent on SNPs, current applications do not allow flexible adjustments across all possible genome-wide combinations. After screening candidates for epistatic interactions under a given model, re-evaluating the genetic model can further strengthen the detection process.

1.2.3 Rheumatoid arthritis as a relevant phenotype in the epistatic interactions study

Rheumatoid arthritis (RA) is a complex autoimmune disease causing multiple joint inflammation with a global prevalence ranging from 0.5% to 1% [42]. Twin studies have demonstrated the heritability of RA to be 53% in the United Kingdom data, 65% in the Finnish data from European [43] and 56.1% in the Japanese cohorts [44]. Additionally, it is estimated that 30-50% of the heritability is attributable to the *HLA-DRB1* alleles [45,46]. The association with *HLA* regions has been consistently replicated across various cohorts in GWAS [47–49]. In terms of epistatic interactions, combinations of *HLA* genes and non-*HLA* regions [50,51] or exclusively of non-*HLA* regions, such as *BANK1* and *BLK* [52], have been reported to be associated in different cohorts. These interactions have primarily been identified through knowledge-based approaches. Consequently, RA is expected to involve epistatic interactions that could be uncovered through data-driven approaches.

1.3 Identification of interactions across host and virus genomes

1.3.1 The associations between host and viral factors

In addition to the previous sections highlighting the importance of epistatic interactions within the host genome, it is also necessary to consider environmental factors, particularly the microbiome, which represents external influences that can shape host phenotypes. One study showed the gut microbiome was more similar in MZ twins than in DZ [53]. Another study involving 21 MZ twin pairs found that intestinal virome similarity between twins [54]. These findings suggest that estimates of heritability may be confounded, as MZ twins typically share not only their genomes but also environmental exposures such as viruses, potentially leading to overestimates of genetic contributions.

There are numerous interactions between human cells and various infectious viruses. For example, Anellovirus, typically infecting humans within the first two years of life, is considered a commensal virus [55]. Human herpesvirus (HHV) 6 establishes latent infections by

interacting with T-cells following primary infection and persists in monocytes or macrophages. During severe immunosuppression to deactivate T-cells, such as after hematopoietic stem cells or solid organ transplantation, HHV-6 can be reactivated, potentially triggering encephalitis [56]. Burkitt lymphoma was the first human malignancy reported to be associated with Epstein-Barr virus (EBV) [57]. This association is mechanistically explained by the t(8;14) chromosomal translocation, which leads to constitutive activation of the *c-myc* gene transcription [58]. In Nasopharyngeal carcinoma (NPC) and gastric cancers, EBV infection is associated with the expression of viral genes such as EBNA1 and LMP1, which have been implicated in oncogenic processes [59]. Additionally, a study utilizing genomic DNA from NPC detected integration sites near tumour suppressor genes [60], suggesting that the integration of the EBV genome into somatic cells during latency might influence tumour progression through a different mechanism. The integration of HBV around cancer driver genes is known to trigger hepatocellular carcinoma [61]. Recently, Anellovirus, which has been considered a commensal virus, has been reported to be associated with autoimmune diseases and SARS-CoV-2 infection [62], illustrating how viral infection status and the corresponding immune response can contribute to the development and progression of various diseases. In contrast, integration of HHV-6 into the human germline genome has occurred in the distant past. These HHV-6 sequences are now inherited across generations, with a prevalence of approximately 1% in the UK based on blood donor samples [63] and 0.8% in the Japanese cohort based on cord blood donor samples [64].

On the other hand, viruses exhibit extensive polymorphism, as reflected in their high mutation rates, ranging from 10^{-8} to 10^{-4} mutations per nucleotide per cell infection, including insertions and deletions [65], compared to a rate of 2.66×10^{-9} per cell division in human somatic cells [66]. A vast number of viral sequences have been cataloged in public databases: as of November 2024, Refseq [67] contains representative sequences for 18,688 viral species, while GenBank [68] hosts 13,699,372 virus sequences, including redundant entries for the same

species. The questions to address are which viruses, and to what extent, are these viruses associated with human health status among the vast diversity of viral species.

Additionally, specific *HLA* alleles have been reported to be associated with chronic infections by HBV [26] or HTLV-1 [27], illustrating the interactions between genetic factors (host genome) and environmental factors (viruses). However, twin studies struggle to distinguish these effects when estimating heritability, as MZ twins typically share the same *HLA* alleles and the infection risk that is determined by the environment. To address this limitation, association studies in the general population are necessary to elucidate host genome and virus interactions, taking into account genetic and geographical backgrounds, which is crucial for advancing disease research.

1.3.2. The challenges to distinguish genetic and environmental associations

Meta-analyses enhance the statistical power of GWAS by aggregating summary data across multiple studies of similar or diverse populations, with consistency of effects across studies pointing towards more robust associations. These analyses also help identify genetic factors that may be associated with different populations. However, when significant associations are observed only within a particular population, these findings may result from genetic and/or environmental differences. Distinguishing between these two factors becomes challenging when meta-analyses are conducted on cohorts with vastly different genetic and geographical backgrounds. To address this issue, additional comparisons using data from participants of cohorts from similar populations but collected from different locations are necessary to disentangle genetic associations from environmental influences.

When a cohort includes populations from multiple populations, it is crucial to ensure concordance in genetic architecture across cohorts. For example, a single-cell RNA study of PBMCs conducted in the United States using samples with ancestry inferred from genetic data

reported that infection-responsive genes showed stronger enrichment of type I and II interferon (IFN) pathways in certain cell types following influenza A virus from European ancestry males than in the corresponding cell types from African ancestry males[69]. To accurately compare geographical differences, it is essential to classify individuals into populations with similar genetic backgrounds. Principal component analysis (PCA) [70] can be used to project individuals into a space based on eigenvectors calculated from genome-wide SNP data of reference populations, such as those in the 1000 Genomes Project and the Human Genome Diversity Project (HGDP) [71]. By measuring the distance between an individual's projected point and the cluster of reference populations in this space, it is possible to annotate them to the most probable population. Additionally, incorporating variables, such as age and sex, into cohort comparisons is critical, as immune responses are significantly influenced by these phenotypes. For instance, immunosenescence is known to weaken immune responses in aging [72], and immune responses tend to be more highly activated in females compared to males [73].

1.3.3 Detection of viral reads from whole genome sequence of human cells.

Beyond human genetic factors, environmental contributors have long included viruses, and many studies have pursued association analyses targeting individual viruses. In most studies, viral sequences and their quantities were measured using DNA amplification techniques such as quantitative PCR [74]. Since PCR requires primers designed from known virus genome sequence, it may overlook certain viral sequences present in samples that are not specifically targeted. In contrast, WGS was developed to detect DNA sequences in a sample without requiring predefined targets, thus enabling a data-driven approach that is independent of prior knowledge. In fact, WGS has demonstrated high reproducibility for inherited human genetic variants, identifying 91% of single nucleotide variants (SNVs) across different sample

preparations and bioinformatics pipelines in a study [75]. While WGS is initially used to detect the host genome in cells, the sequencing reads also contain DNA from viruses that have established latent infections or have been integrated into the host genome. These virus-derived reads can be detected from WGS by focusing on reads that do not align to the human genome reference [76]. The quantity of viral DNA in cells can be estimated by comparing the read counts and genome sizes of the virus and the host, using the assumption that each host human cell contains two copies of the genome.

In practical applications of WGS, phages such as PhiX developed by Illumina Inc. are sometimes spiked into the sample at low proportions to improve the base quality of the reads, which may result in artificial reads appearing in the sequencing data [77]. These artificial reads are usually negligible in general WGS data analysis; however, they become a significant issue in viral detection due to their high homology with viral sequences. Thus, artificial reads, which are not present in the cell, can be mistakenly captured as viral sequences, leading to false-positive estimates of viral quantities [78]. The types and amounts of artificial reads vary depending on the library preparation method [79]; therefore, no standardized approach currently exists for their correction. However, a blood microbiome study using WGS data from healthy individuals, including viral reads, identified contaminants based on their high prevalence or strong correlation with other contaminants within the same batch [80]. Nevertheless, the study also concluded that non-human-derived reads that are not classified as contaminants may still persist. On the other hand, several papers have reported that empty wells (i.e., no input DNA) were sequenced alongside experimental samples as negative controls to monitor background contamination [78,81]. Nevertheless, because WGS has typically been performed for purposes other than viral detection, it is difficult to evaluate background sequences under strictly matched conditions. Therefore, to distinguish between contaminant sequences and DNA viruses truly present in human cells, filtering of WGS reads using novel bioinformatics approaches is required.

1.3.4. Unbiased viral quantification observed in different dynamic ranges

The majority of viruses detected in previous WGS studies were DNA viruses, and these are typically much less abundant than host DNA [82,83]. The estimates from real-time PCR studies [84,85] indicate that only 1.7, 0.5, and 1.7 read pairs of EBV, HHV-6B, and HHV-7, respectively, would be expected in WGS at a mean coverage of 30× with 150×2 paired-end reads because latently infected cells constitute only a fraction of total white blood cells. Under such conditions, it is likely that some samples do not contain viral reads in WGS data even when the virus is present in human cells, resulting in false negative estimates of viral abundance. The number of reads derived from viruses is correlated with that of human-derived reads, which constitute the majority of WGS data. Therefore, the probability of detecting viral reads depends on the total number of WGS reads, and viral quantities across different cohorts may be biased, especially when the target coverage of WGS differs between datasets. Such differences have not been a major concern so far, as most studies have been conducted using WGS data within a single cohort. However, investigations of geographic and population genetic differences require diverse WGS datasets from multiple cohorts, often generated by different studies, making cross-cohort comparisons challenging due to experiment-dependent biases. Therefore, it is necessary to address this emerging challenge to enable unbiased comparisons of viral quantities derived from WGS of human cells across different platforms.

1.4 New strategy for GWAS and accurate heritability estimation (Objective and hypothesis)

In this thesis, I will address missing heritability through two studies. In Chapter 1.2, I discussed the potential epistatic interactions within the host genome that were difficult to detect due to

the combinatorial explosions when we apply all combinations. These interactions are typically not accounted for or appropriately modeled in conventional estimates of heritability. Using the item mining approach, implemented in LAMPLINK, the first study will focus on identifying epistatic interactions and optimizing genetic models using data from rheumatoid cohort studies. Environmental factors may be mistakenly treated as genetic factors, potentially inflating heritability estimates. Therefore, it is important to investigate whether viral exposures, as a key environmental factor, have any association across genetic or geographical backgrounds. In the second study, I will investigate interactions between viruses and humans based on the hypothesis that the viral reads, annotated from unmapped reads in WGS data, may reflect infections in the human body and exhibit patterns specific to geographical regions or population genetics. This analysis involves extracting viral sequences from WGS data of human cells and comparing their abundance across populations with diverse geographical and genetic backgrounds. These findings will contribute to improving heritability estimates by integrating epistasis and viral infections into GWAS analyses.

2. Identification of epistatic SNP combinations in rheumatoid arthritis using LAMPLINK and Japanese cohorts

Mio Shibata^{1,2}, Aika Terada³, Takahisa Kawaguchi¹, Yoichiro Kamatani⁴, Daigo Okada¹, Kazuhisa Nagashima¹, Koichiro Ohmura⁵, Fumihiko Matsuda^{1,2}, Shuji Kawaguchi^{1,2}, Jun Sese^{3,6}, Ryo Yamada¹

Published in *Journal of Human Genetics*: 69:541–547 on July 16th, 2024.

Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan

1. Kyoto-McGill International Collaborative School in Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan
2. Humanome Lab., Inc., Tokyo, Japan
3. Laboratory for Statistical Analysis, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan
4. Department of Rheumatology and Clinical Immunology, Graduate School of Medicine, Kyoto University, Kyoto, Japan
5. Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan

2.1. Abstract

Genome-wide association studies have enabled the identification of important genetic factors in many trait studies. However, only a fraction of the heritability can be explained by known genetic factors, even in the most common diseases. Genetic loci combinations, or epistatic contributions expressed by combinations of single nucleotide polymorphisms (SNPs), have been argued to be one of the critical factors explaining some of the missing heritability, especially in oligogenic/polygenic diseases. Rheumatoid arthritis (RA) is a complex disease with more than 100 reported SNP associations, as well as various *HLA* haplotypes and amino acids; however, many associations between RA and inter-chromosomal SNP combinations are unknown. To discover novel associations of epistatic interactions with high odds ratios in RA, we applied the LAMPLINK method, a systematic enumerative procedure for identifying high-order SNP combinations, to a Japanese RA cohort (discovery cohort; 4024 patients with RA and 7731 controls). We validated the identified associations in a different Japanese cohort (validation cohort; 810 RA patients and 6303 controls). In this study, we identified 90 significant genetic associations in the discovery cohort. Among these, 74 (82.2%) associations were replicated in the validation cohort, and eight combinations were inter-chromosomal, all of which comprised rs7765379 or rs35265698 located in the HLA region. These two SNPs exhibited strong correlations with valine at amino acid position 11 in HLA-DRB1 (HLA-DRB1-11-Val). Finally, we discovered that rs9624 showed an association with RA through an epistatic interaction with HLA-DRB1-11-Val. Overall, LAMPLINK showed high reliability for identifying epistatic genetic contributions hidden in complex traits.

2.2. Introduction

Genome-wide association studies (GWAS) identify genetic factors associated with disease; therefore, they can help unravel the etiology of diseases. Most GWAS studies have focused on

single nucleotide polymorphisms (SNPs). For even the resolution of SNP, we must consider the massive number of variants in the human genomes and need a large sample size and high calculation cost [1]. Oligogenic/polygenic diseases are associated with multiple loci, whereas Mendelian diseases are associated with a few genetic loci. There are many studies of oligogenic/polygenic diseases that the associations tested by single SNP showed the insufficient effect sizes, implying the presence of undiscovered SNP combinations that has the association with disease in high odds ratio only when all SNPs in the combination have effect alleles [2]. To identify the genetic factors associated with these oligogenic/polygenic diseases with high odds ratio, an association analysis based on multiple SNPs, the epistatic interactions, are crucial. However, considering all possible combinations calculated from n SNPs, which is the astronomical number of $2^n - 1$ makes the association analysis meaningless due to the correction of multiple testing. Moreover, the allele frequency of the SNP combination is decreased with an increase of the number of SNPs with low linkage disequilibrium (LD). Inter-chromosomal combination involving SNPs of different chromosomes does not have LD. Consequently, using conventional GWAS, SNP combinations tend to have lower detection power than those of the single SNP.

Several approaches have been developed for association analysis of SNP combinations, including exhaustive methods and filtering. The exhaustive methods, BitEpi [3] and BOOST [4], test combinations of multiple SNPs while the number of SNPs included in combinations is limited in these methods, owing to computer resources. The knowledge-based filtering method estimates the effect size of the SNP combinations [5]. However, these approaches may not detect all SNPs associated with diseases or traits only when combined with other SNPs. To overcome these limitations, LAMPLINK was developed; it detects statistically significant epistatic interactions of two or more SNPs [6,7]. This software is based on the LAMP algorithm [8], which selects candidates for testing from all $2^n - 1$ combinations by pruning the

enumerative combinations in which no possibility of statistical significance is confirmed. Based on Bonferroni's method, using the count of selected combinations, LAMPLINK can detect significant associations in high-order combinations using smaller correction factors for multiple testing.

Rheumatoid arthritis (RA) is an autoimmune disease that systemically causes inflammation in multiple joints and is considered an oligogenic/polygenic disease. To date, more than 100 SNPs for RA, including those in the HLA region (located on 6p21.3), particularly *HLA-DRB1*, have been reported to be associated using different ethnicities. The large meta-analysis of RA in a GWAS study, using the European and East Asian populations showed that the odds ratios of non-HLA regions estimated from a single SNP were between 0.93 and 1.81 [9]. Regarding the HLA region, odds ratios showed the wide range according to the alleles or amino acids, an odds ratio of 4.44 for *HLA-DRB1*04:01*, 3.8 for valine at amino acid position 11 in *HLA-DRB1*, whereas 0.38 for serine at amino acid position 11 in *HLA-DRB1* [10]. In addition, RA is associated with epistatic interactions, such as those between *HLA-DRB1* and *PTPN2* [11], *BLK* and *BANK1* [12], and 564 interactions between the HLA and non-HLA regions [13]. These findings are based on single SNP based typical GWAS analyses. Polygenetic risk score (PRS) of RA based on these findings alone, linearly considering multiple SNPs [14], is insufficient; thus, determining SNPs epistatic interactions (associated with combinatorial effects) is crucial. Here, we applied LAMPLINK to a Japanese cohort of 4024 RA patients and 7731 controls recruited from the nationwide in Japan as a discovery cohort to detect high-order epistatic combinations of SNPs. The results were evaluated using another independent Japanese cohort of 810 patients with RA cases and 6303 controls.

2.3. Materials and methods

Genotype data, quality control, and imputation

For the use of two Japanese cohort datasets in this study, ethical approval was obtained from institutional review boards of RIKEN and Kyoto University. The first cohort was the discovery cohort from BioBank Japan (BBJ) [15], which included 4024 patients with RA and 7731 controls; genotyping was performed using the Illumina OmniExpressExome genotyping array. The second cohort was the validation cohort, which included 810 patients with RA (from the Kyoto University [9]) and 6303 controls (from the NAGAHAMA study [16]); genotyping was performed using the Illumina HumanHap610-Quad BeadChip and Infinium Japanese Screening Array, respectively. The genotype data, obtained from for both cohorts, was evaluated by applying the following: 1) SNPs with an overall call-rate of individuals was more than 0.9 were selected; 2) Individuals in which the overall call-rate of SNPs was more than 0.99 were selected; 3) SNPs with an overall call-rate of individuals was more than 0.99 were selected; 4) SNPs with p -values higher than 1.0×10^{-6} (based on the Hardy-Weinberg equilibrium exact test) were extracted; and 5) SNPs with minor allele frequencies (less than 0.01) were removed. Consequently, 497,190 SNPs were used in the discovery cohort. These quality filtering steps were conducted using PLINK [17].

After quality control, the SNP haplotypes were estimated using ShapeIT [18], with the dataset of the 1000 Genomes Project (1000G_Phase3 under the release of 20130502). Imputation was conducted on each chromosome for the Phese1.v3panel of the 1000 Genomes Project using Minimac3 [19]. After extracting the genotype on the array and removing mismatch alleles, 497,190 SNPs were used for the discovery cohort. We imputed the genotype of the validation cohort using the same reference as the discovery cohort, independent of the array.

Detection of the combinatorial effect

LAMPLINK [7] was used to detect SNP combinations that are significantly associated with RA in the discovery cohort. Let n be the number of SNPs used in the analysis. LAMPLINK surveys $2^n - 1$ combinations of SNPs while omitting combinations that would not produce a statistically significant threshold from their marginal counts to reduce the number of tests and increase the significance threshold. Let the diploid genotype be “AA,” “Aa,” and “aa,” where “a” is the effect allele. For each combination, the frequency of “aa” (the recessive genotype) was tested between RA patients and controls using Pearson’s chi-squared test (the recessive exclusive model). Subsequently, the resulting p -value (p_cor) was corrected using Bonferroni correction based on the reduced number of SNP combinations.

Using LAMPLINK, we analyzed the following six SNP sets: 1) coding regions (13,285 SNPs), 2) regulatory regions defined by 5k bp upstream and 0.5k bp downstream regions from transcription starting sites (49,955 SNPs), 3) Union of set 1 and set 2 (61,097 SNPs), 4) wide regulatory regions in which the upstream area were expanded to 10k bp (86,724 SNPs), 5) Gene regions which correspond to areas from the transcription starting site to the transcription end site (214,885 SNPs), and 6) whole SNPs (497,190 SNPs). For each of the SNP sets, SNPs were selected in analyses with a minor allele frequency (MAF) filter. MAF was adjusted in increments of 0.05 from 0.05 to 0.35 (seven thresholds) in the discovery cohort. The SNP annotations and locations were based on NCBI dbSNP for Genome Reference Consortium Human Build 37 (GRCh37) release 147, and the transcription start sites were based on GRCh37 release 75. To avoid unfeasible calculations for large combinations of SNPs, we stopped LAMPLINK when the program did not finish within five days. For all LAMPLINK executions, the statistical significance level was set at 1.0. LAMPLINK defined an effect allele as a SNP combination or a single SNP with the least frequent allele, while the reference allele was defined as the most frequent allele.

We combined SNPs and SNP combinations and their p_cor values detected by LAMPLINK performed in each SNP set and MAF threshold. When the same combination was detected in

multiple runs, the maximum p_{cor} value was considered. Finally, SNPs and SNP combinations with p_{cor} values equal to or less than 0.05 were considered statistically significant (confirming an association with RA). We also calculated the nominal p value (p_{nom}) for a single SNP belonging to significant SNPs and SNP combinations using a two-tailed Pearson's chi-squared test under the recessive exclusive model. Odds ratios and the corresponding 95% confidence intervals (95% CI) for SNPs between RA patients and controls were also calculated.

Validation of detected combinatorial effect

The 90 significant associations of SNPs and SNP combinations, including single SNP associations, were detected in the LAMPLINK runs, and validated in the validation cohort. To prepare the genotype of the significantly associated SNP combinations, identified in the discovery cohort, we extracted the genotypes imputed by Minimac3, confirming that the RSq was greater than 0.3, if the allele was not genotyped in the array. A two-tailed Pearson's chi-squared test was used to calculate the p -values of the SNP combinations and SNPs. The effective alleles were confirmed to be harmonized in the validation cohort. In the LAMPLINK results, many detected SNPs were located in the HLA region and were shared among multiple SNP combinations. Therefore, we used the composite LD (CLD) based method, defined by a total ratio of eigenvalues in descending order (0.995) [20] for the correction of multiple testing (instead of Bonferroni correction of the total number of tests). We used the recessive genotypes of the SNP combinations and SNPs to calculate the CLD. Consequently, a correction factor of 50 was used for multiple testing and let p_{v} was the p -value multiplied of the factor in the discovery cohort. Finally, signals with $p_{\text{v}} < 0.05$ were considered to be replicated.

Characterization of the detected SNP combinations by the combined model

Correlations between p -values and odds ratios of the validated signal were evaluated using the Kendall coefficient without ties, estimating r as a correlation coefficient [21]. The combined

effect sizes of the two cohorts in each combination were estimated using the fixed-effect model with the *metafor* [22] package in R. The product of the odds ratio in each inter-chromosomal combination was calculated by multiplying the odds ratios with the single SNP of the non-HLA region and the SNP combination or single SNP of the HLA region. We used UpSet [23] and UpSetPlot (<https://github.com/jnothman/UpSetPlot>) to illustrate the SNPs in the combination and their combined odds ratios. To evaluate our findings, we searched for the alleles in the combinations of GWAS Catalogs downloaded from 2019-07-19 as RA (Experimental Factor ID as EFO_0000685) [24]. We compared alleles with p -values less than or equal to 5.0×10^{-8} .

HLA allele imputation

We imputed *HLA-B*, *-DRB1*, and *-DPB1* alleles of individuals in the validation cohort with two-field *HLA* allele from SNP arrays used in this study with DEEP*HLA [25]. The Pan-Asia panel [26,27] was used as the reference and lifted over to GRCh37 based on the rsIDs using the TogoVar database [28]. Amino acids in the MHC groove domains (G-DOMAINS) were annotated using the IPD-IMGT/HLA database (release 3.44.0) [29] and an external annotation script [30]. Amino acids were symbolized by the HLA gene name-position-amino acid. Among the significantly associated combinations in the validation cohort, we defined the correlation $r_h(\alpha, s)$ between HLA amino acid α and single SNP s in the HLA region as:

$$r_h(\alpha, s) = \frac{(P_{\alpha,s} - P_{\alpha} \times P_s)^2}{P_{\alpha}(1 - P_{\alpha})P_s(1 - P_s)},$$

where $P_{\alpha,s}$ is the joint frequency of α and s in heterozygous, including compound heterozygous and homozygous individuals, and P_{α} and P_s are frequencies of α and s in heterozygous and homozygous individuals, respectively.

Estimation of combinatorial genetic models

The genetic model for RA of the inter-chromosomal combination by the HLA amino acid,

whose association with RA under the additive effect model was reported [10], and the SNP in the non-HLA region replicated in this study was estimated as follows. Among the eight non-HLA SNPs, we excluded two SNPs whose odds ratio between RA patients and controls in the recessive model by single SNP were below one in either the discovery or validation cohort. Let $y \in \{0, 1\}$ be an indicator variable of the RA patient ($y = 1$) and control ($y = 0$) and $\mathbf{x} = (x_1, x_2, x_3)$, where $x_k \in \{0, 1\}$ be indicator variables for heterozygous of HLA-DRB1-11-Val ($x_1 = 1$), homozygous of HLA-DRB1-11-Val ($x_2 = 1$), and non-HLA SNP ($x_3 = 1$) whose status varies depending on the model. We assume that the posterior probability of $y = 1$ given in dataset $\mathbf{x}$ is expressed by binomial logistic regression of the generalized linear model, defined as:

$$P(y = 1|\mathbf{x}) = \frac{\exp(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \beta x_3)}{1 + \exp(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \beta x_3)}, \quad P(y = 0|\mathbf{x}) = 1 - P(y = 1|\mathbf{x}),$$

where $(\alpha_0, \alpha_1, \alpha_2, \beta) \in \mathbf{R}^4$ are parameters for variables. Either “Aa” or “aa” was defined as a dominant genotype, and the association test used the frequency of this dominant genotype as the “dominant model”. In this study, we estimated the parameters in seven models defined as follows: 1, $\beta = 0$ as a null hypothesis (null model); 2, $x_3 = 1$ for individuals carrying heterozygous or homozygous of non-HLA SNP (dominant model) as the additive effect of the HLA region; 3, $x_3 = 1$ for individuals carrying homozygous of non-HLA SNP (recessive model) as the additive effect of the HLA region; 4, $x_3 = 1$ for individuals carrying both heterozygous or homozygous of HLA-DRB1-11-Val and heterozygous or homozygous of non-HLA SNP; 5, $x_3 = 1$ for individuals carrying both heterozygous or homozygous of HLA-DRB1-11-Val and homozygous of non-HLA SNP; 6, $x_3 = 1$ for individuals carrying both homozygous of HLA-DRB1-11-Val and heterozygous or homozygous of non-HLA SNP, and 7, $x_3 = 1$ for individuals carrying both homozygous of HLA-DRB1-11-Val and homozygous of non-HLA SNP. Models 4–7 are considered epistatic interaction models. The parameters of each model were determined using the maximum likelihood estimation using the glm package

of R. The optimal model among the seven models was selected by the Akaike information criteria (AIC). If the null model was selected, we rejected the additive or epistatic effect of the non-HLA SNP. The significant level of the p -value in the coefficient of β was set to $0.05/3 = 0.0167$ by the Bonferroni correction of multiple testing, considering the three parameters. If the p -value was less than the threshold and the 95% confidential interval of the odds ratio was higher than 1, we concluded that the non-HLA SNP had the effect size apart from that attributed to HLA-DRB1-11-Val.

2.4. Results

Run results of LAMPLINK

In this study, LAMPLINK was used, with seven MAF thresholds, to analyze and detect six SNP sets in the discovery cohort (4024 RA patients and 7731 controls). A total of 23 conditions were completed within five days of runtime (Supplementary Table 1). The number of SNPs used in LAMPLINK linearly increased according to the MAF in any SNP set (Fig. 1A). In contrast, the number of SNP combinations selected by LAMPLINK exponentially increased according to the MAF (Fig. 1B). Specifically, the count of used SNPs and SNP combinations in ALL SNPs with MAF 0.15 are 159,754 and 540,985, respectively, while these in Coding regions with MAF 0.35 are 10,006 and 59,394,760, respectively. This result indicates that the number of SNP combinations strongly depends on the MAF, regardless of the SNP set.

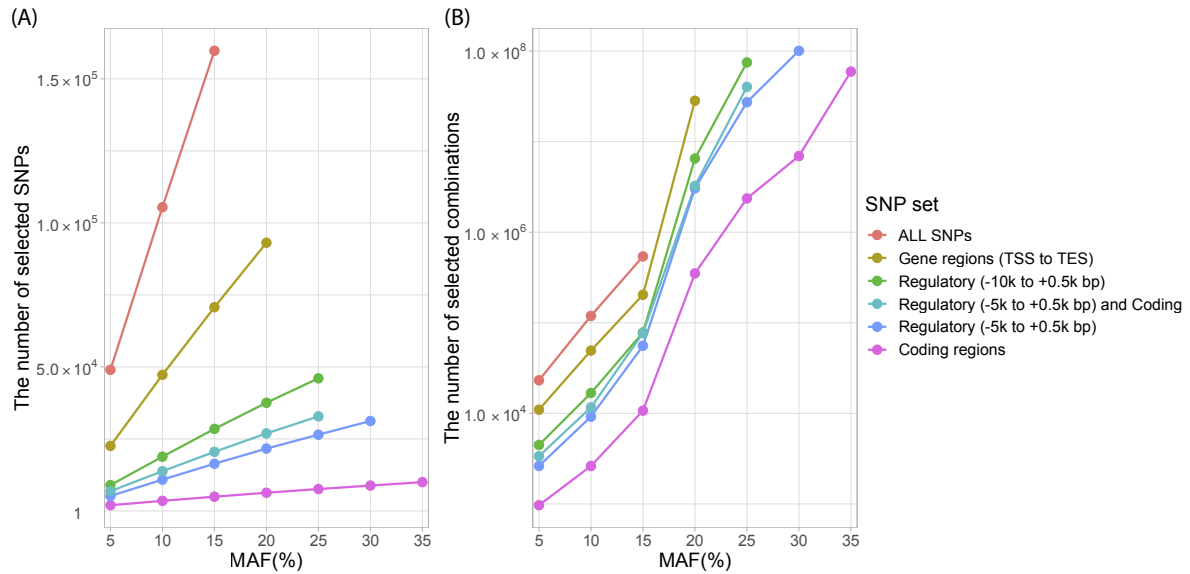


Fig. 1

Counts of selected SNPs (A) and SNP combinations (B) for the six SNP sets, according to a minor allele frequency (MAF).

Detected SNP combinations associated with RA

LAMPLINK resulted in 23 conditions and detected 102 SNPs and SNP combinations that were significantly associated with RA after Bonferroni correction, in each condition, by the number of selected SNP combinations. If the selected SNP combinations overlapped in multiple conditions, the maximum p_{cor} value was considered and 90 (78 SNP combinations and 12 single SNPs) significant associations (Table 1 and Supplementary Table 2) were observed.

Among the 90 associations, 54 distinct SNPs were identified, 37 of which were located in the HLA region. In addition, twenty-two of the 78 SNP combinations were inter-chromosomal, with all of the SNPs located in the HLA region (chr6:28477797-33448354). 12 of the 22 inter-chromosomal combinations were SNP pairs, nine of which were trios, and one was a quartet (Table 1).

The distribution of SNP combinations by minimum MAF showed that 97.8% of 90 associations were detected in the range of 0.15–0.30 in MAF (Supplementary Table 3). The p_{cor} of 12

combinations fell above or below the LAMPLINK condition threshold (Supplementary Fig. 1). Owing to the increase in the number of selected SNP combinations, a higher MAF was selected and a higher p_{cor} was obtained (Fig. 1B). The SNP set with a combination in the regulatory region (100,813,625 SNP combinations) and MAF of 0.30 balanced the false positive and false negative rates.

The number of SNPs in the combination	The number of identified combinations	The number of identified inter-chromosomal combinations	The number of replicated combinations	The number of replicated inter-chromosomal combinations
1	12	-	11	-
2	26	12	18	5
3	25	9	19	3
4	8	1	7	0
5	7	0	7	0
6	6	0	6	0
7	4	0	4	0
8	2	0	2	0
total	90	22	74	8

Table 1

The number of combinations categorized by SNP counts

Replicated SNP combinations within the validation cohort

Associations between RA and 78 SNP combinations and 12 single SNPs were tested using a discovery cohort of 810 patients with RA and 6303 controls. A total of 63 (80.8%) SNP combinations and 11 (91.2%) single SNPs were replicated in this validation cohort ($p_{\text{v}} < 0.05$).

A total of 43 distinct SNPs were detected in the replicated associations, including 35 SNPs located in the HLA region (Supplementary Table 2, 4 and 5). The details of the replicated associations and their SNPs are presented in Table 1. Among the 74 replicated associations, eight SNP combinations were inter-chromosomal combinations (8/22), while the remaining 55 SNP combinations (55/56) and 11 single SNPs (11/12) were composed only of SNPs in the HLA region. All combinations of five or more SNPs were intrachromosomal. The 90 significant associations identified in the discovery cohort were positively correlated with the validation cohort in both p -values ($r = 0.43$, $p = 1.57 \times 10^{-9}$, Fig. 2) and odds ratios ($r = 0.38$, $p = 8.76 \times 10^{-8}$, Fig. 2), leading to the detection of confidently positive combinations of inter- and intra-chromosomal. Stratifying the inter- and intra-chromosomal combinations, the positive correlations in p -values and odds ratios were exhibited in the intra-chromosomal combinations (p -values: $r = 0.42$, $p = 4.64 \times 10^{-7}$, odds ratios: $r = 0.62$, $p = 8.27 \times 10^{-14}$), while those of the intra-chromosomal combinations did not show the significant correlations (p -values: $r = 0.03$, $p = 0.87$, odds ratios: $r = 0.25$, $p = 0.11$).

All eight inter-chromosomal combinations replicated in the validation cohort had both or either of the SNPs in the HLA region, rs7765379 and rs35265698. While non-HLA alleles in SNP combinations did not show significant associations in single SNP only (Supplementary Table 5), seven out of the eight inter-chromosomal SNP combinations had larger odds ratios, calculated by the combinations rather than by each single SNP (Supplementary Table 2, Supplementary Table 5), indicating that effect alleles in non-HLA regions have specifically high odds ratio in the combinatorial effect.

The SNPs in the non-HLA regions were located within eight genes: *PGAMIP13*, *OR10G7*, *ARHGAP42*, *FAM176A*, *FGFR3P2*, *NDRG2*, *TPPP2*, and *LYZLI*. Comparison of the replicated 43 SNPs with previous GWAS reports registered in the GWAS Catalog [24] revealed that only rs7765379, located at the *MTCO3P1* locus in the HLA region [31] was consistent with the results of our study.

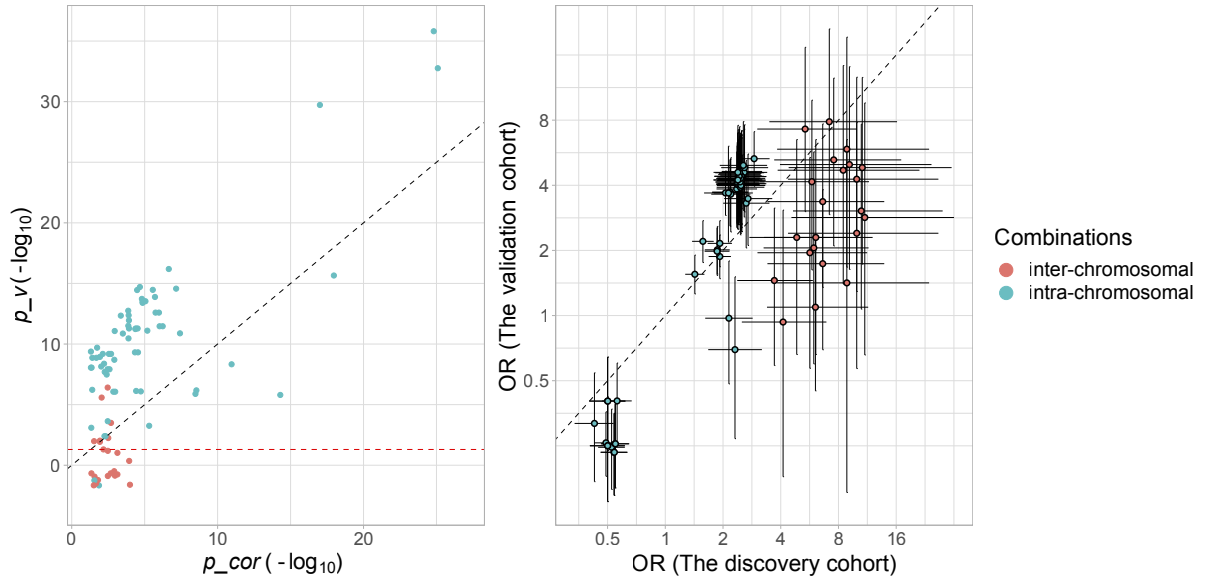


Fig. 2

Comparing p -values (p_{cor} in the discovery cohort and p_v in the validation cohort) and ORs of the validated signals between the discovery and validate cohorts. Each dot indicates one combination. The red dashed line represents the threshold (0.05).

Combinatorial effects of replicated combinations

To estimate a more concrete effect size for replicated associations in the Japanese population, we estimated the odds ratio by combining individuals in the discovery and validation cohorts under the fixed-effect model. All inter-chromosomal combinations ($n=8$) had higher odds ratios than the maximum odds ratio of the individual SNP in the combinations (Fig. 3, Supplementary Table 5). In addition, these odds ratios were higher than the product of the odds ratios of the SNP combinations in the HLA region and the SNP in the non-HLA region (Supplementary Table 2), showing that this high odds ratio was yielded specifically in the SNP combination. In contrast, 38 of the 66 SNP combinations, in which SNPs were located only in the HLA region, had lower odds ratios than the maximum odds ratios of single SNP in the combinations (Fig. 3, Supplementary Table 5).

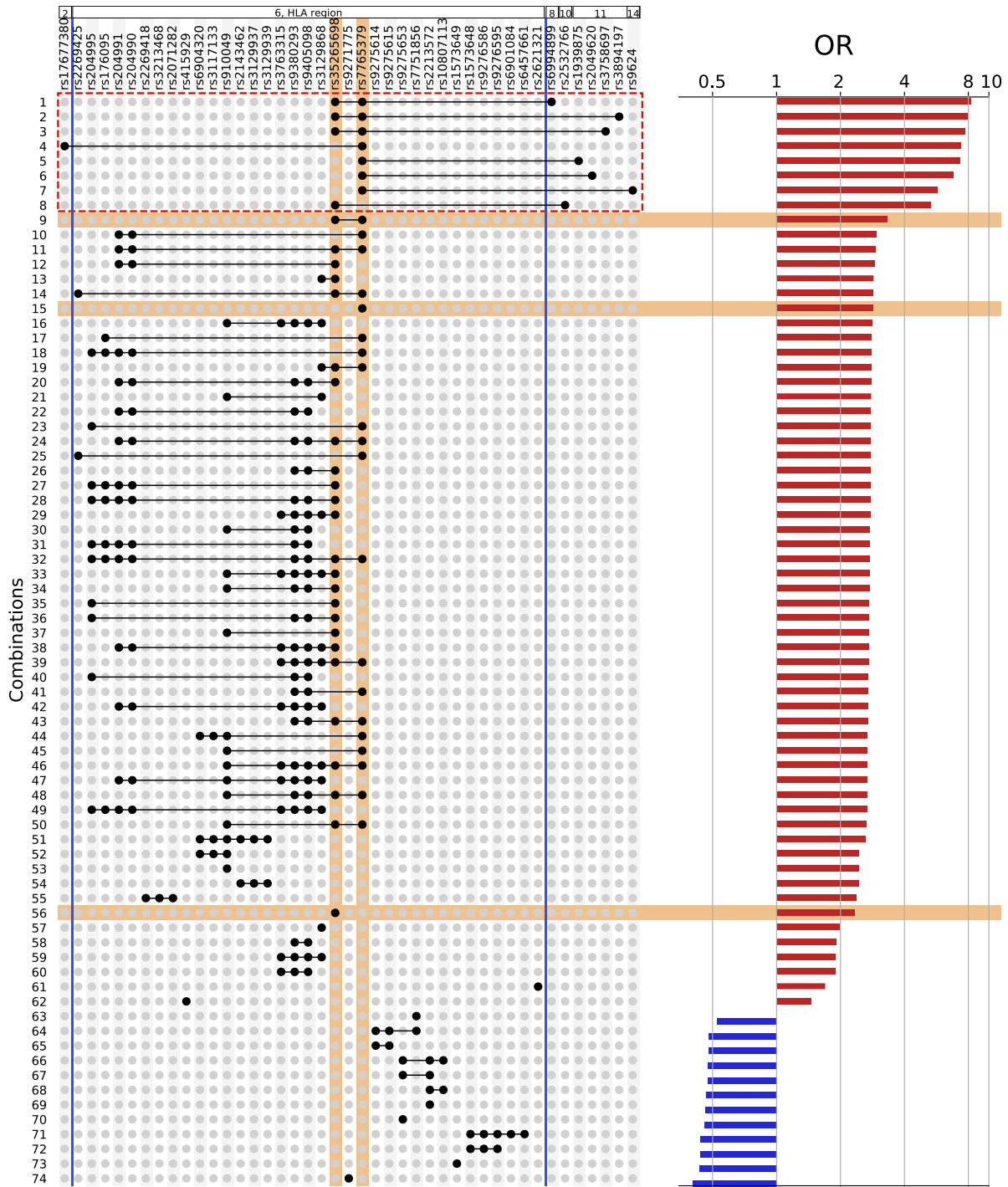


Fig. 3

The left side of the dot and line shows the SNP composition in every signal, and the bars on the left show the odds ratios. The top eight combinations are inter-chromosomal combinations (circled with a dashed red line). The two SNPs highlighted in the columns show the SNPs in the HLA region the inter-chromosomal combinations.

Correlation with reported HLA amino acids

To elucidate the relationship between the replicated 35 SNPs in the HLA region and nine HLA amino acids (HLA-DRB1-11-Val, HLA-DRB1-11-Pro, HLA-DRB1-11-Gly, HLA-DRB1-11-Ser, HLA-DRB1-71-Lys, HLA-DRB1-71-Ala, HLA-DRB1-74-Ala, HLA-B-9-Asp, and HLA-DPB1-9-Phe) whose associations with RA were reported [10], we imputed *HLA-B*, *-DRB1*, and *-DPB1* alleles in the validation cohort using DEEP*HLA. Two SNPs, rs35265698 and rs7765379, which were components of all replicated inter-chromosomal combinations exhibited strong correlations with known risk amino acid, HLA-DRB1-11-Val ($r_h = 0.9412$ for rs35265698 and 0.4742 for rs7765379, Supplementary Table 6), whose position corresponds to 7th amino acid residue in the G-BETA domain of HLA-DRB1 [32]. In contrast, rs9271775 showed the lowest combined odds ratio (OR = 0.40 [0.33, 0.50], Supplementary Table 2), had a strong correlation with HLA-DRB1-71-Ala ($r_h = 0.7196$) reported as the protective amino acid [10]. The strong correlations in both risk and protective HLA amino acids indicated the high reproducibility of p -values and odds ratios in the intra-chromosomal combinations (Fig. 2).

Redefinition of genetic models for inter-chromosomal combinations

Finally, we evaluated the effects of non-HLA SNPs within the replicated eight inter-chromosomal combinations by conditioning the effect of HLA-DRB1-11-Val using two additive models and four interactive models via logistic regression of the generalized linear model. SNPs rs3758697 and rs3894197 were excluded from the analysis as their odds ratios in the recessive model were lower than 1.0 in the discovery cohort (Supplementary Table 5). Consequently, four SNPs exhibited lower AIC in either combinatorial model than in the model using only the HLA-DRB1-11-Val effect (null model). However, among them, rs6994899 and rs1939875 did not demonstrate significant effect sizes ($p \geq 0.0167$, Supplementary Table 7). SNP rs2049620 exhibited an additive effect under the dominant model independent of HLA-

DRB1-11-Val (model 2, $p = 0.00180$, OR = 1.27 [1.09, 1.47]). In addition, rs9624 had a significant effect for RA under the dominant model within HLA-DRB1-11-Val (model 6, $p = 0.0109$, OR = 1.75 [1.14, 2.71]), suggesting an epistatic interaction between rs9624 and HLA-DRB1-11-Val with RA risk. The odds ratio of RAs individuals carrying HLA-DRB1-11-Val homozygote was 5.06 in the null hypothesis model. However, the odds ratio of those with the rs9624 dominant genotype was estimated to be 6.58, whereas those without the rs9624 dominant allele, it was estimated at 3.76.

2.5. Discussion

LAMPLINK detects SNP combinations that are potentially difficult to detect using conventional GWAS, owing to the explosion of considerable combinations. Therefore, LAMPLINK is an effective method for identifying SNP combination associations in oligogenic/polygenic diseases. In this study, we applied LAMPLINK to a Japanese cohort of 4,024 patients with RA and 7,731 controls (the discovery cohort) to detect novel SNP combinations significantly associated with RA. LAMPLINK identified 78 SNP combinations, of which 12 single SNPs were significantly associated with RA (Table 1). Among the identified associations, 63 (82.2%) SNP combinations and 11 (91.2%) single SNPs were replicated in the other independent Japanese cohort of 810 RA patients and 6303 controls (the validation cohort). In addition, the eight SNP combinations were inter-chromosomal and composed of SNPs located in the HLA region, where specific amino acids in *HLA-DRB1*, *HLA-B* and *HLA-DPBI* have been reported as risk factors for RA [10], and the other regions of eight genes (Supplementary Table 2). In particular, the inter-chromosomal combinations increased the odds ratios compared to those independently calculated using a single SNP belonging to the combinations in both the discovery and validation cohorts.

In the present study, SNPs rs35265698 and rs7765379, located in the HLA region and

comprising all inter-chromosomal combinations, exhibited strong correlations with HLA-DRB1-11-Val, indicating that these SNPs might be proxy effects of known risk factor [10]. The redefinition of the genetic model for inter-chromosomal combinations concluded that rs2049620 indicated an additive effect under the dominant model. However, this result raises some concern because the additive effect should be derived from the conditional analysis in the conventional GWAS approach. In contrast, rs9624 showed an association with RA under the dominant model within HLA-DRB1-11-Val homozygote, indicating that this effect is considered an epistatic interaction and is difficult to identify through the conventional GWAS approach owing to the low detection power in complex genetic models. The discovery of novel epistatic interactions using LAMPLINK could improve polygenic risk scores for RA.

HLA-DRB1-11-Val is mainly translated from *HLA-DRB1*04* alleles; however, they exhibited a large difference in odds ratios under the additive effect model, such as *HLA-DRB1*04:01* (OR = 4.44), **04:04*, **04:05* or **04:08* (OR = 4.22), **04:03* or **04:07* (OR = 1.65), and **04:02* (OR = 1.43) in a previous report [10]. The haplotype of amino acids positions 71 and 74 is considered to contribute to these differences. In addition, transgenic mice expressing *HLA-DRB1*04:01* or **04:04* have been reported to increase osteoclasts in primary bone marrow cells, whereas those expressing *HLA-DRB1*04:02* or **04:03* did not increase osteoclasts [33]. Moreover, alterations in bone marker levels have been observed in the donated blood of individuals who developed RA after the donation, suggesting changes in osteoclast activity before the onset of RA [34]. SNP rs9624 is located in the intron or upstream region of *NDRG2*, whose overexpression in monocytes suppresses osteoclast differentiation [35]. These evidence indicate a strong association with RA resulting from the interaction between rs9624 mutation and *HLA-DRB1* risk alleles of RA. However, it remains unclear why the results differed between homozygous and heterozygous of HLA-DRB1-11-Val. In addition, while we analyzed biallelic information (SNP or amino acid) in this study, the use of multiallelic in LAMPLINK could elucidate the detailed mechanisms in high polymorphic regions.

We have shown that LAMPLINK has the potential to identify associations by SNP combinations from genome-wide SNP array data. Although LAMPLINK can test the high-order SNP combinations with high detection power by pruning the ineligible SNP combinations, we had to set MAF thresholds and perform LAMPLINK under the recessive exclusive model to reduce SNP combinations. All identified SNP combinations included SNPs in the HLA region where many factors strongly associated with RA are known. In addition, SNPs in the HLA region have high LD each other. Under these conditions, the efficiency of the pruning is low because LAMPLINK endlessly yields high-order combinations among the SNPs in high LD block that decreases the significance threshold by Bonferroni correction. It might be better to prune combinations that include SNPs in high LD forcibly to test novel SNP combinations in a high significant threshold. As another solution, an implementation of dummy variables based on factors whose associations with the phenotype are known, as input of LAMPLINK instead of SNPs piled up in the LD block increases detection power. Therefore, developing of an effective LAMP algorithm could handle complex genetic models to detect novel epistatic interactions.

2.6. Conflict of interest

There are no conflicts of interest to declare.

2.7. Acknowledgements

This work was supported by JST, CREST Grant Number JPMJCR1502 and JST, AIP Acceleration Research JPMJCR21U2, Japan. We express our gratitude to all the participants in the BioBank Japan Project and Nagahama Study. M.S. was supported by the Kyoto-McGill International Collaborative Program in Genomic Medicine.

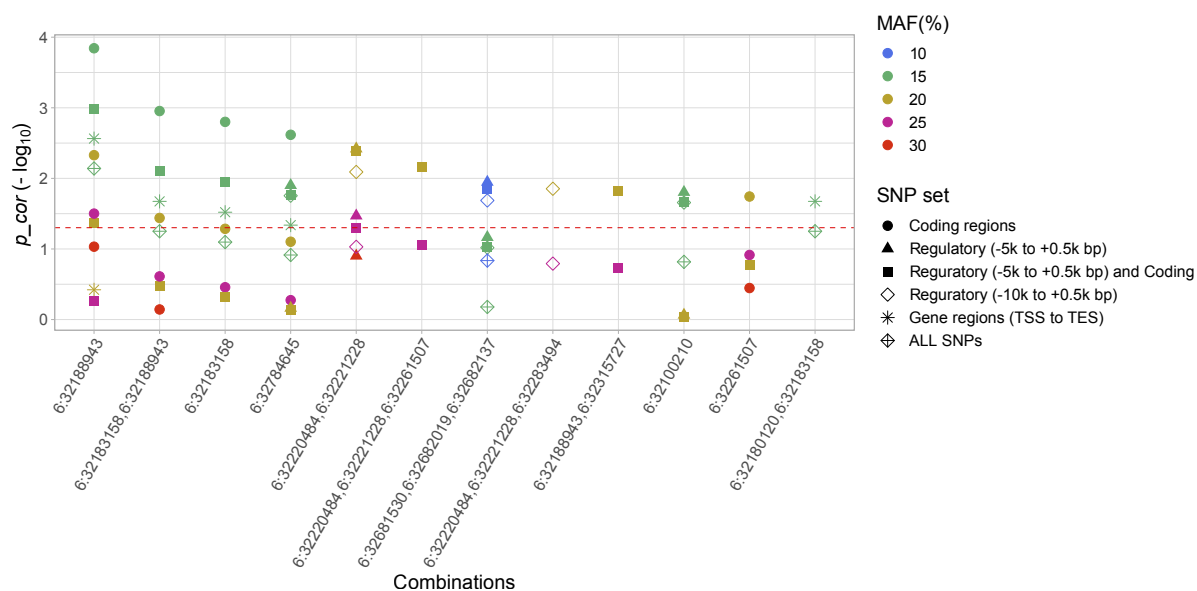
2.8. References

1. Tam V, Patel N, Turcotte M, Bosse Y, Pare G, Meyre D. Benefits and limitations of genome-wide association studies. *Nat Rev Genet.* 2019 Aug;20(8):467–84.
2. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell.* 2017 Jun 15;169(7):1177–86.
3. Bayat A, Hosking B, Jain Y, Hosking C, Kodikara M, Reti D, et al. Fast and accurate exhaustive higher-order epistasis search with BitEpi. *Sci Rep.* 2021 Aug 5;11(1):15923.
4. Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NLS, et al. BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-wide Case-Control Studies. *Am J Hum Genet.* 2010 Sep 10;87(3):325–40.
5. Ma L, Brautbar A, Boerwinkle E, Sing CF, Clark AG, Keinan A. Knowledge-Driven Analysis Identifies a Gene–Gene Interaction Affecting High-Density Lipoprotein Cholesterol Levels in Multi-Ethnic Populations. *PLOS Genet.* 2012 May 24;8(5):e1002714.
6. Terada A, Okada-Hatakeyama M, Tsuda K, Sese J. Statistical significance of combinatorial regulations. *Proc Natl Acad Sci U A.* 2013 Aug;110(32):12996–3001.
7. Terada A, Yamada R, Tsuda K, Sese J. LAMPLINK: detection of statistically significant SNP combinations from GWAS data. *Bioinformatics.* 2016 Nov;32(22):3513.
8. Minato S, Uno T, Tsuda K, Terada A, Sese J. A fast method of statistical assessment for combinatorial hypotheses based on frequent itemset enumeration. In Springer; 2014. p. 422.
9. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature.* 2014 Feb;506(7488):376–81.
10. Raychaudhuri S, Sandor C, Stahl EA, Freudenberg J, Lee HS, Jia X, et al. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat Genet.* 2012 Jan;44(3):291–6.
11. Qiao B, Huang CH, Cong L, Xie J, Lo SH, Zheng T. Genome-wide gene-based analysis of rheumatoid arthritis-associated interaction with PTPN22 and HLA-DRB1. *BMC Proc.* 2009 Dec;3 Suppl 7:S132.
12. Genin E, Coustet B, Allanore Y, Ito I, Teruel M, Constantin A, et al. Epistatic interaction between BANK1 and BLK in rheumatoid arthritis: results from a large trans-ethnic meta-analysis. *PLoS One.* 2013;8(4):e61044.

13. Diaz-Gallo LM, Ramskold D, Shchetynsky K, Folkersen L, Chemin K, Brynedal B, et al. Systematic approach demonstrates enrichment of multiple interactions between non-HLA risk variants and HLA-DRB1 risk alleles in rheumatoid arthritis. *Ann Rheum Dis*. 2018 Oct;77(10):1454–62.
14. Stahl EA, Wegmann D, Trynka G, Gutierrez-Achury J, Do R, Voight BF, et al. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat Genet*. 2012 Mar;44(5):483–9.
15. Nagai A, Hirata M, Kamatani Y, Muto K, Matsuda K, Kiyohara Y, et al. Overview of the BioBank Japan Project: Study design and profile. *J Epidemiol*. 2017 Mar;27(3S):S2–8.
16. Setoh K, Matsuda F. Cohort Profile: The Nagahama Prospective Genome Cohort for Comprehensive Human Bioscience (The Nagahama Study). In: Yano M, Matsuda F, Sakuntabhai A, Hirota S, editors. *Socio-Life Science and the COVID-19 Outbreak: Public Health and Public Policy* [Internet]. Singapore: Springer Singapore; 2022. p. 127–43. Available from: https://doi.org/10.1007/978-981-16-5727-6_7
17. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007 Sep;81(3):559–75.
18. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nat Methods*. 2011 Dec;9(2):179–81.
19. Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. *Nat Genet*. 2016 Oct;48(10):1284–7.
20. Gao X, Starmer J, Martin ER. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet Epidemiol*. 2008;32(4):361–9.
21. KENDALL MG. A NEW MEASURE OF RANK CORRELATION. *Biometrika*. 1938;30(1–2):81–93.
22. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw*. 2010;36(3):1–48.
23. Lex A, Gehlenborg N, Strobel H, Vuilleumot R, Pfister H. UpSet: Visualization of Intersecting Sets. *IEEE Trans Vis Comput Graph*. 2014 Dec;20(12):1983–92.
24. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. 2019 Jan;47(D1):D1005–12.

25. Naito T, Suzuki K, Hirata J, Kamatani Y, Matsuda K, Toda T, et al. A deep learning method for HLA imputation and trans-ethnic MHC fine-mapping of type 1 diabetes. *Nat Commun*. 2021 Mar 12;12(1):1639.
26. Okada Y, Kim K, Han B, Pillai NE, Ong RT, Saw WY, et al. Risk for ACPA-positive rheumatoid arthritis is driven by shared HLA amino acid polymorphisms in Asian and European populations. *Hum Mol Genet*. 2014 Dec;23(25):6916–26.
27. Pillai NE, Okada Y, Saw WY, Ong RTH, Wang X, Tantoso E, et al. Predicting HLA alleles from high-resolution SNP data in three Southeast Asian populations. *Hum Mol Genet*. 2014 Aug 15;23(16):4443–51.
28. Mitsuhashi N, Toyo-oka L, Katayama T, Kawashima M, Kawashima S, Miyazaki K, et al. TogoVar: A comprehensive Japanese genetic variation database. *Hum Genome Var*. 2022 Dec 12;9(1):1–9.
29. Robinson J, Soormally AR, Hayhurst JD, Marsh SGE. The IPD-IMGT/HLA Database – New developments in reporting HLA variation. *Hum Immunol*. 2016 Mar 1;77(3):233–7.
30. Kawaguchi S, Matsuda F. High-Definition Genomic Analysis of HLA Genes Via Comprehensive HLA Allele Genotyping. *Methods Mol Biol*. 2020;2131:31–8.
31. Freudenberg J, Lee HS, Han BG, Shin HD, Kang YM, Sung YK, et al. Genome-wide association study of rheumatoid arthritis in Koreans: population-specific loci as well as overlap with European susceptibility loci. *Arthritis Rheum*. 2011 Apr;63(4):884–93.
32. Lefranc MP, Duprat E, Kaas Q, Tranne M, Thiriot A, Lefranc G. IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. *Dev Comp Immunol*. 2005 Jan 1;29(11):917–38.
33. Holoshitz J, Liu Y, Fu J, Joseph J, Ling S, Colletta A, et al. An HLA-DRB1–Coded Signal Transduction Ligand Facilitates Inflammatory Arthritis: A New Mechanism of Autoimmunity. *J Immunol*. 2013 Jan 1;190(1):48–57.
34. Schaardenburg D van, Nielen MMJ, Lems WF, Twisk JWR, Reesink HW, Stadt RJ van de, et al. Bone metabolism is altered in preclinical rheumatoid arthritis. *Ann Rheum Dis*. 2011 Jun 1;70(6):1173–4.
35. Kang K, Nam S, Kim B, Lim JH, Yang Y, Lee MS, et al. Inhibition of osteoclast differentiation by overexpression of NDRG2 in monocytes. *Biochem Biophys Res Commun*. 2015 Dec 25;468(4):611–6.

2.9. Supplementary information



Supplementary Fig. 1

The 12 combinations in which the p_cor values were below or above the LAMPLINK condition threshold. The red dotted line represents the threshold.

Connecting Text: Bridge Between Chapter 2 and Chapter 3

Chapter 2 focuses on the identification of missing factors related to epistatic interactions, alongside the development of optimized models for each, to help address missing heritability. Although these factors are detectable in genetic data, they were often overlooked by conventional approaches. This highlights an underexplored level of resolution that may contribute to the explanation of missing heritability. In Chapter 3, beyond the genetic data in the human cells, the environmental information that may also contribute to addressing the missing heritability was examined. Environmental factors encompass all external elements surrounding an individual, many of which are not typically measured or incorporated into conventional GWAS. This presents a challenge, as such unmeasured components may act as missing factors contributing to missing heritability. In this context, next-generation sequencing (NGS) data, such as WGS, yield vast numbers of reads derived from DNA present in the target cells. Some studies have suggested that, because NGS captures DNA regardless of its origin, it may also detect viral sequences from non-human genomes, which could represent environmental contributors to missing heritability. By focusing on non-human genomic elements captured within human WGS data, Chapter 3 aimed to detect and quantify environmentally derived missing factors that are usually excluded from GWAS.

3. Geographic and genetic influences on the human blood DNA virome

Mio Shibata^{1,2}, Shuji Kawaguchi¹, Mathieu Bourgey³, Takahisa Kawaguchi¹, Yuichi Inadomi¹, Emmanuel Gonzalez³, Robert Eveleigh³, Pascale Marquis³, Hans Grabe⁴, Karen Mather⁵, Helena Schmidt⁶, Stephanie Debette⁷, Ryo Yamada¹, Yoichiro Kamatani¹, Fumihiko Matsuda¹, Guillaume Bourque^{2,3}

1.Centre for Genomic Medicine, Kyoto University Graduate School of Medicine, Kyoto, Japan

2.Department of Human Genetics, McGill University, Montreal, Canada

3.Canadian Centre for Computational Genomics, McGill University, Montreal, Canada

4.Department of Psychiatry and Psychotherapy, University Medicine Greifswald, Greifswald, Germany

5. Centre for Healthy Brain Ageing, Discipline of Psychiatry and Mental Health, School of Clinical Medicine, University of New South Wales, Sydney, Australia

6.Institute of Molecular Biology and Biochemistry, Medical University of Graz, Graz, Austria

7.Bordeaux Population Health research center, University of Bordeaux, Inserm, UMR 1219, Bordeaux, France

3.1. Abstract

Viruses present in human cells can impact health to varying degrees, depending on their specific characteristics and their interactions with the host. Additionally, the presence and abundance of these viruses may be shaped by environmental conditions and the host's genetic makeup, though these relationships remain poorly understood. In this study, we developed a pipeline to assess the presence and quantities of viruses from unmapped reads from whole genome sequencing and applied it to 7,059 individuals obtained from several cohorts from Europe, Australia, Canada and Japan. In total, we detected 52 distinct types of viruses in these cohorts, including EBV/HHV-4, HHV-6B, HHV-7, and Anellovirus, which were found in over 1% of the individuals across all cohorts. Among these four viruses, we identified significant differences in prevalence and abundance between geographically or genetically different groups. Notably, the prevalence of Anellovirus in Japan (58.29%) was higher than in the other regions ($\leq 33.63\%$, $p < 1.66 \times 10^{-40}$). GWAS signals in the MHC region prompted *HLA* typing and amino acid-level association analysis, identifying three residues for Anellovirus, eight for EBV/HHV-4, and four for HHV-7. These results shed light on some of the environmental and genetic factors that contribute to the interactions between human cells and viruses.

3.2. Introduction

Viruses present in the human body are known to influence disease, including cancers and autoimmune diseases, some of them through latent infections that extend beyond the acute infectious phase [1,2]. Several viruses, such as HIV and HTLV-1, also have the ability to integrate into the human genome [3]. Latent infection and subsequent reactivation, are known to be influenced by host genetic factors, such as human leukocyte antigen (HLA), and are associated with disease development [4,5] as exemplified by Hodgkin lymphoma [6] and in HTLV-1-associated myelopathy [7]. These associations may be influenced by geographic and genetic backgrounds, yet this remains poorly understood.

Human herpesviruses and Anellovirus are considered to persist in blood cells as latent infections [2]. Real-time PCR has been commonly utilized to measure viral load in cells [8]; however, this approach targets only a single species and has amplification bias due to high polymorphism of viral sequences. In contrast, whole-genome sequencing (WGS) in human cells has the potential to identify circulating DNA viruses together with chromosomally integrated proviruses without predefined targets. Indeed, most viral sequences do not have homology with human genome and consequently could appear as unmapped reads when WGS data is mapped to the human genome reference [9].

Such an approach was applied to a cohort not selected for health status with WGS from 8,000 individuals in the United States, which detected viral reads from 94 distinct viral species [10]. The highest prevalence, measured as the proportion of individuals with at least one read from a given viral species, was observed for human herpesvirus 7 (HHV-7) at 20.4% and significant associations. This study also reports that depending on the virus, either prevalence, abundance or both are associated with phenotypic traits such as sex, age, and population, with abundance defined as the quantity of viral reads in an individual. An independent Chinese cohort of 10,585 individuals, reported a high prevalence of Anellovirus (76.7%), along with higher prevalence

rates of HHV-7 and Epstein-Barr virus (EBV) in individuals aged ≤ 40 years (24% and 22%, respectively) than in individuals aged > 60 years (9% and 38%, respectively) [11].

Previous studies have shown that the number of detected viral reads associated with latent infections was extremely low in most individuals [10,11]. Indeed, real-time PCR analysis revealed 197 ± 433 standard deviation EBV genome copies per 1.0×10^6 PBMCs obtained from 15 people who were seropositive for EBV [12]. This finding indicates that the expectation of read pairs in human (3.10Gbp) WGS with $30\times$ coverage for EBV (1.73Kbp) is estimated to only be ~ 1.7 pairs. Under such conditions, even if EBV is present in blood cells, its reads from EBV may not be included in some individuals, resulting in false negatives. This indicates that the number of detected viral reads will be proportional to the total read count of WGS. Additionally, unmapped reads, often contain a large number of artificial reads, notably phage-derived reads added to improve base quality [13] and synthetic sequence reads to trace cross-contaminating samples within the sequencing run. Independent of technical issues, these cohort studies [10,11] are restricted to a single-country study, limiting the ability to determine whether viral quantities are influenced by population or geographic location. However, inter-cohort comparisons are hindered by the protocol-related discrepancies, including variation in total read counts and protocol-specific artificial reads, as previously described.

In addition to viruses that establish latent infections within the host, transient infections may also influence the host as environmental factors. Among these, pandemics exert a rapid impact on the host, and several reports have described changes in the latent infection status (viral load) in SARS-CoV-2–infected individuals. For example, the frequency of EBV reactivation is higher in SARS-CoV-2 positive individuals compared to SARS-CoV-2 negative groups [14]. Conversely, studies on Anellovirus present inconsistent findings: some report increased viral loads in critically ill COVID-19 patients [15], while others observe decreases among healthcare workers [16]. Unlike previous events, the COVID-19 pandemic occurred after the advent of

WGS, enabling direct comparisons of genomic data collected before and during its spread.

In the present study, we used WGS blood cells data from three independent cohorts with diverse genetic background and conducted in different locations (Europe, Australia, Japan, and Canada). This includes a case-control cohort from Canada put together during the COVID-19 pandemic, with known infection status. Our goal was to develop a robust viral read detection pipeline and identify geographic, environmental and host genetic factors associated with latent infections.

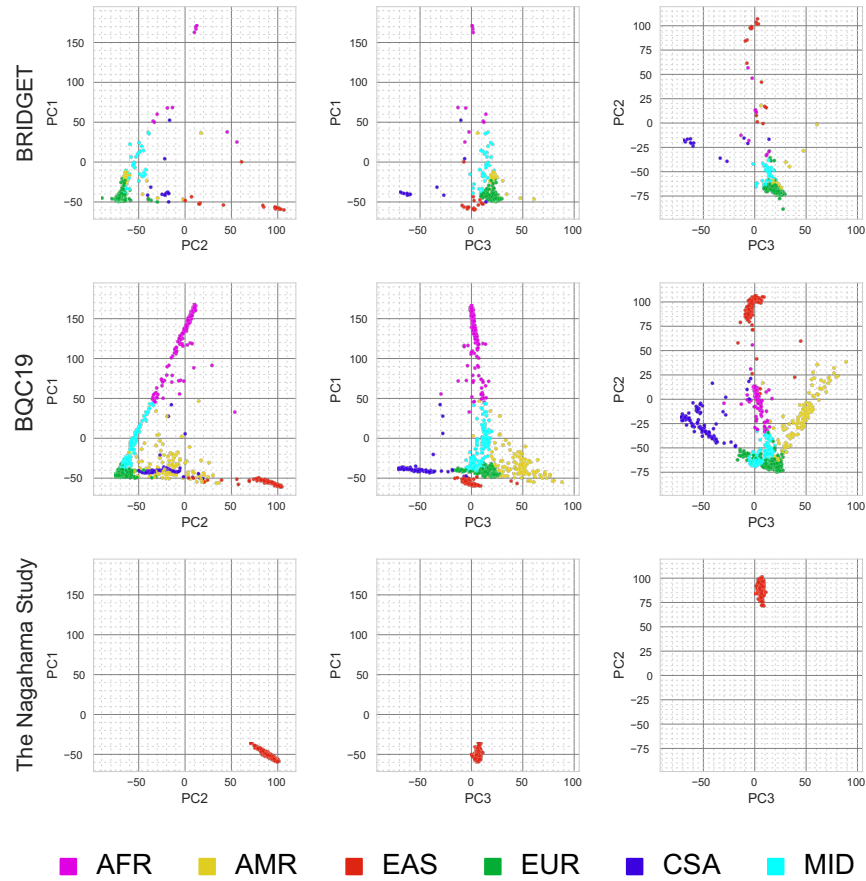
3.3. Results

Genetic and phenotypic diversity in the cohorts analyzed

In this study, we used WGS data from: 1) 2,560 individuals of the BRain Imaging, cognition, Dementia and next-generation GENomics (BRIDGET) cohort, collected in Germany, France, Austria and Australia; 2) 3,066 individuals of the Biobanque québécoise de la COVID-19 (BQC19), a case-control cohort of SARS-CoV-2 infection collected in Quebec, Canada [17] (729 negative and 2,337 positive for SARS-CoV-2) and; 3) 1,583 individuals of the Nagahama Study conducted in Nagahama city in Japan [18] (Supplementary Table 1). The WGS datasets were generated from blood samples and variants were called using GenPipes [19] and GATK best-practices [20] for quality filtering (see Methods). Some of the parameters had to be adjusted to take into account the variable sequencing coverage of the cohorts (Supplementary Table 2, Supplementary Fig. 1a). Next, we assigned to each individual (Fig. 1a) to one of the six populations as defined by Human Genome Diversity Project (HGDP) and 1000 Genomes Project [21] using their joint-call data (Supplementary Fig. 1b). Individuals in BQC19 were genetically diverse, whereas the majority of BRIDGET (95.1%=2,329/2,449) were assigned to the European category (EUR), and all individuals of the Nagahama Study (100%=1,583/1,583)

were assigned to the Eastern Asian category (EAS) (Supplementary Table 1). Sex was determined based on genotype-derived binary sex categories, revealing that 64.0%, 49.5% and 77.4%, were female in BRIDGET, BQC19, and the Nagahama Study, showing a concordance rate exceeding 99% with the reported sex (Supplementary Table 3). The age range was broad in BQC19, from 0 to 103 years (mean 57.4 years, s.d.=20.0, Fig. 1b). In BRIDGET, age exhibited a bimodal distribution between 20 to 30 years and 40 to 93 years (mean 60.5 years, s.d.=22.5, Supplementary Table 1, subgroup distribution in each cohort shown in Supplementary Fig. 1c). In contrast, all individuals in the Nagahama Study were older than 30 years of age. (mean 58.0 years, s.d.=12.8).

a



b

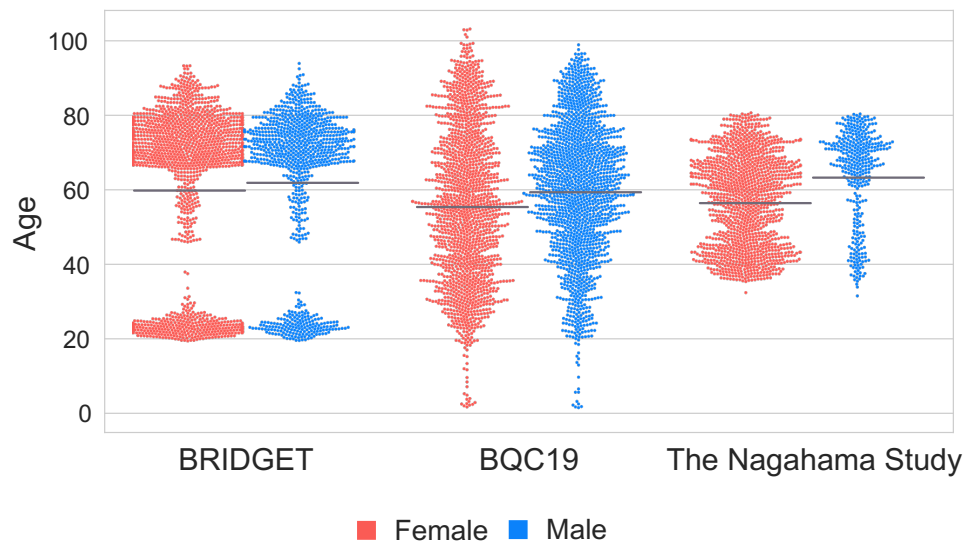


Fig. 1 Demographic and genetic characteristics

a, Top three principal components (PCs) for BRIDGET, BQC19 and the Nagahama Study cohorts. Each dot represents an individual, coloured according to one of six assigned populations. b, Age distribution by sex in the three cohorts.

Construction of the viral read detection pipeline

We developed a pipeline to detect viral reads from human WGS data while minimizing biases and errors due to sequence protocols (Fig. 2a, Methods). In short, paired-end reads in WGS, in which both forward and reverse template strands were unmapped to the human reference genome, were classified as unmapped read pairs (Step 1). After filtering out low-quality base reads and any spiked-in or tagmentation reads introduced during DNA preparation (Step 2), the unmapped reads were then mapped to 10,532 viral genomes (Step 3). To evaluate viral quantities, the prevalence was defined as the number of individuals with at least one viral read, while abundance referred to the quantity of viral reads in an individual. Viruses in individuals whose abundances were disproportionately high based on a generalized gamma distribution were classified as outliers while the ones near one copy number were classified as likely inherited chromosomal integrations (ici) (Step 4). An additional filtering step looking at the uniformity of coverage across the full length of viruses was implemented to avoid spurious hits (Step 5). Finally, to ensure unbiased comparisons between cohorts with varying levels of total reads, the prevalence and abundance were adjusted (Step 6). These adjusted viral quantities were subsequently used in association tests that accounted for geographic, phenotypic, and genetic factors.

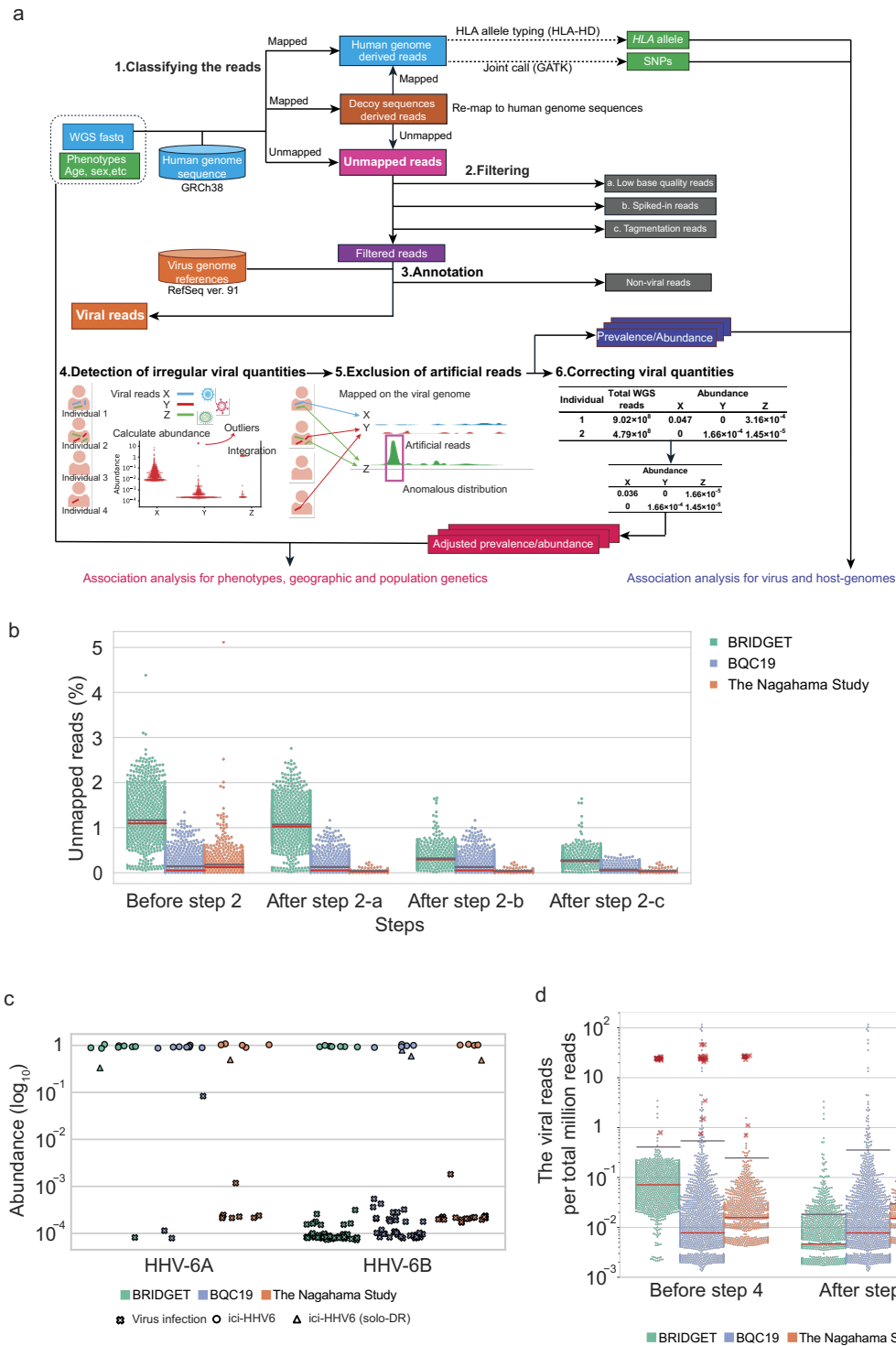


Fig. 2 Workflow of viral read detection and association analyses

a, Workflow for viral read detection from WGS data, followed by cohort-based artificial read exclusion and adjusted estimation of prevalence and abundance to ensure unbiased inter-cohort

comparisons. The adjusted viral quantities are used to compare groups across different recruitment locations. Generated with BioRender.com. b, Distribution of unmapped read rates per individual across three cohorts after filtering low-quality (Step 2-a), spiked-in (Step 2-b), and tagmentation-derived reads (Step 2-c). Gray and red bars represent the mean and median values, respectively. c, Classification results of HHV-6A- and HHV-6B-derived reads into virus infection or inherited chromosomal integration. d, Distributions of viral reads per million total reads for each individual before Step 4 (left) and after Step 5 (right). Gray and red bars represent the mean and the median values, respectively. The individual marked with a red cross in the left plot represents an outlier or a case of iciHHV-6 detected in Step 4.

Common detection of four viruses across all cohorts

We applied this pipeline to the three cohorts in our study and observed that the mean frequency of raw unmapped read pairs was 1.16% (s.d.=0.41%) in BRIDGET, 0.14% (s.d.=0.16%) in BQC19, and 0.18% (s.d.=0.24%) in the Nagahama Study (Fig. 2b, Step 1). After filtering low-quality reads and PhiX and Kapa tags (Step 2), these values were reduced to 0.28% (s.d.=0.12%) in BRIDGET, 0.07% (s.d.=0.05%) in BQC19, and 0.04% (s.d.=0.02%) in the Nagahama Study, depending on the DNA preparation protocol used by each cohort (Supplementary Table 2). After annotating these unmapped reads to viral genomes (Step 3), we found that the prevalences of any virus in each individual were 98.86% (2,421/2,449) in BRIDGET, 59.43% (1,799/3,027) in BQC19, and 69.80% (1,105/1,583) in the Nagahama Study, (Supplementary Table 4). Next, in Step 4, putative iciHHV-6A were found in nine individuals from BRIDGET, eight from BQC19, and six from Nagahama; iciHHV-6B was predicted in seven, seven, and five individuals from the same cohorts, respectively, among whom five individuals were solo-DR, defined as the remaining direct repeat (DR) region of HHV-6 integrated in the human genome [22] (Fig. 2c, Supplementary Fig. 2). Among them,

one BQC19 individual had both iciHHV-6A or iciHHV-6B. Additionally, one BRIDGET individual showed outlier abundance for EBV/HHV-4 (abundance=15.53), and one BQC19 individual showed outlier abundance for HHV-7 (abundance=0.12).

After excluding viral reads from outliers, we further removed artificial reads when the coverage of the viral genomes was not uniform (Step 5). As a result of this step, the prevalences were reduced to 59.53% (1,458/2,449), 59.13% (1,790/3,027), and 68.295% (1,081/1,583), in BRIDGET, BQC19 and Nagahama Study respectively. The large reduction in BRIDGET was attributed to artificial reads mapped to HHV-5 and *Autographa californica* nucleopolyhedrovirus (Supplementary Fig. 3a, Supplementary Table 4). Regarding the average viral per million total reads per individual excluding outliers, it decreased significantly from 0.0843 to 0.0108 in BRIDGET, whereas it remained nearly unchanged in BQC19 (from 0.2095 to 0.2095) and the Nagahama Study (from 0.0215 to 0.0201), respectively, Fig. 2d, Supplementary Table 4). After all of these steps, viral reads corresponding to 52 species viruses were detected, although most of these species exhibited low prevalence (Supplementary Table 5). Among them, the prevalence of EBV/HHV-4, HHV-6B, HHV-7 and Anellovirus, exceeded 1% in all three cohorts with no artificial reads included (Supplementary Fig. 3b, Supplementary Table 6), and these four viruses were selected for further analyses.

Finally, we validated the quantification correction step (Step 6) using simulated data. The WGS dataset in BRIDGET was first randomly split into two subsets (Set 1 and Set 2). For each WGS data in subsets, half of the reads were randomly sampled (Set 1h and Set 2h). Viral quantities were then calculated for both the full and half subsets, with and without adjustment (Supplementary Fig. 4a). Fisher's exact test showed that the prevalence of Anellovirus, EBV/HHV-4, and HHV-7 was significantly higher in Set 1 than in Set 1h, and in Set 2 than in Set 2h (Supplementary Fig. 4b and Supplementary Table 7). In contrast, Student's *t*-test demonstrated statistically significant decreases in the abundance of EBV/HHV-4, HHV-6B and

HHV-7 (Supplementary Fig. 4c). After applying the adjustment, no differences were observed in either prevalence or abundance, confirming the robustness of the approach.

Geographical diversity of viral prevalence and abundance

Next, we wanted to estimate the effects of age, sex and of the geographic location on the viral quantities observed. For this comparison, individuals were categorized into one of four groups, Australia (n=336), Europe (n=2,113), Canada (n=723) and Japan (n=1,583), based on recruited location without stratifying by genetic data (Supplementary Table 1). To avoid the impact of SARS-CoV-2 infection, the BQC19 cohort analysis was limited to SARS-CoV-2–negative individuals (n=723). We first tested viral prevalence and abundance along with sex and age within each of the four groups using a generalized linear regression model. As a result, the prevalence of Anellovirus showed a significant association ($p<0.01$) with age in Europe ($p=1.48\times10^{-23}$, effect size=1.03, 95%CI=[1.02, 1.03]), Canada ($p=1.93\times10^{-8}$, effect size=1.02, 95%CI=[1.01, 1.03]) and Japan ($p=5.12\times10^{-19}$, effect size=1.04, 95%CI=[1.03, 1.05], Supplementary Table 8). For sex, the prevalence of Anellovirus in females was significantly lower than in males for all groups, with odds ratios (ORs) ranging from 0.41 in Australia to 0.55 in Canada. Strong positive correlations were also observed between abundance of Anellovirus and age in Canada ($r=0.286$, $p=2.72\times10^{-6}$) and Japan ($r=0.205$, $p=2.49\times10^{-10}$). Conversely, HHV-7 in Europe exhibited the negative correlation ($r=-0.185$, $p=2.23\times10^{-3}$) between the abundance and age (Supplementary Fig. 5).

Next, prevalence and abundance were compared across geographically distinct locations, while age and sex were included as covariates. The prevalence of Anellovirus showed statistically significant differences ($p<8.33\times10^{-3}$) among all locations, increasing in the order of Australia (18.55%, s.d.=1.10%), Europe (23.98%, s.d.=0.51%), Canada (33.64%, s.d.=0.80%), and

Japan (58.28%, s.d.=0.25%) (Fig. 3a, Supplementary Table 9). For EBV/HHV-4, Japan exhibited a higher prevalence (14.27%, s.d.=0.17%) than both Australia (11.38%, s.d.=1.05%) and Europe (11.60%, s.d.=0.45%). Additionally, the prevalence of HHV-7 was lower in both Canada (6.18%, s.d.=0.59%) and Japan (4.69%, s.d.=0.12%) compared to Europe (10.00%, s.d.=0.41%). Moreover, despite similar prevalence between Australia (5.91%, s.d.=0.88%) and Canada, a statistically significant difference was observed. Regarding abundance, both Canada and Japan showed significantly higher levels of Anellovirus compared to Europe ($p=1.35\times 10^{-4}$, effect size=1.33, 95%CI=[1.24, 1.43] for Canada, and $p=5.42\times 10^{-6}$, $\beta=1.28$, 95%CI=[1.23, 1.34] for Japan, Fig. 3b, Supplementary Table 9).

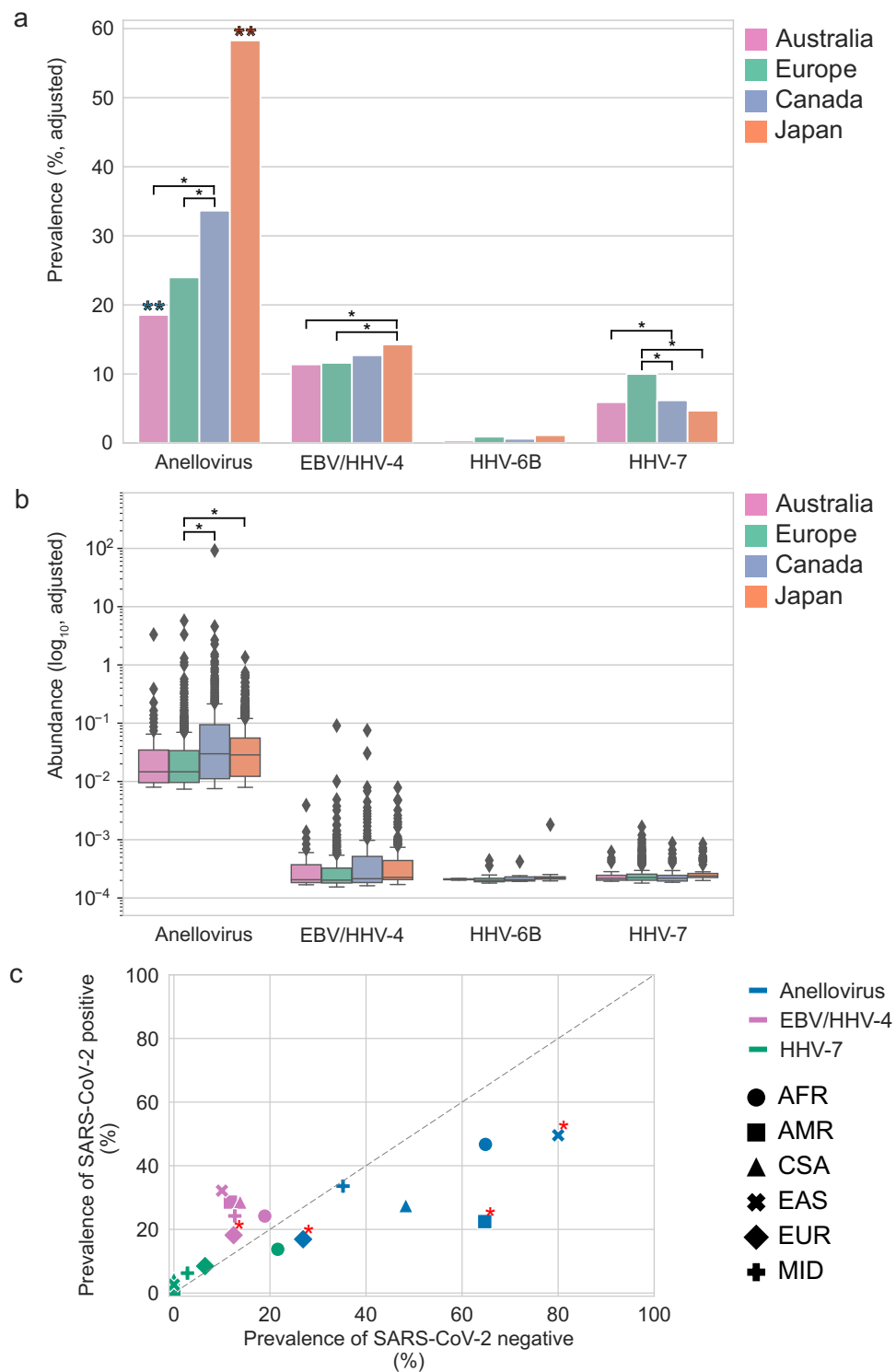


Fig. 3 Differences in viral quantities across geographic backgrounds or SARS-CoV-2 infection status

a,b, Adjusted prevalences (a) and abundances (b) of four DNA viruses across four groups categorized by recruitment locations. Statistically significant pairwise comparisons are marked

with asterisks. Double asterisks indicate significantly lower (blue) or higher (red) prevalence compared to all other groups. c, Differences in prevalence between SARS-CoV-2 infection statuses in BQC19, stratified by populations. Statistically significant pairwise comparisons are indicated by red asterisks on their right shoulder.

Divergent prevalence within and between populations

Next, we wanted to look at the genetic effects and we stratified samples based on genetic background across the different recruited locations, focusing on EUR in Australia, Europe and Canada and EAS in Japan and Canada. In this analysis, the prevalence of Anellovirus was found to be significantly different for EUR individuals in Australia (18.45%, s.d.=1.10 %) compared to Europe (24.03%, s.d.=0.53%, $p=3.42\times10^{-6}$), as well as between Australia and Canada (27.31 %, s.d.= 0.85%, $p=2.83\times10^{-5}$) (Table 1). Notably, for EAS, the rate in Canada was shown to be significantly higher than in Japan (76.63%, s.d.=4.33% vs 58.28%, s.d.=0.25%, $p=9.01\times10^{-3}$). No significant differences were detected for the other viruses and HHV-6B had to be excluded from this analysis because of its low prevalence.

Finally, to evaluate differences based on genetic background within the same location, we examined viral prevalence between AFR (n=37), AMR (n=17), CSA (n=29), EAS (n=30) and EUR (n=539) within Canada. The prevalence of Anellovirus in EUR (27.31%, s.d.= 0.85%) was found to be significantly lower ($p<3.33\times10^{-3}$) than in other four groups (CSA:47.57%, s.d.=3.72%, AMR:53.41%, s.d.=7.74%, AFR:64.22%, s.d.= 4.46% and EAS:76.63%, s.d.=4.33%, Supplementary Table 10). The prevalence of EBV/HHV-4 did not show a significant difference in any of the group comparisons. In the case of HHV-7 prevalence, viral reads were detected only in individuals from AFR (20.01%, s.d.=4.29 %) and EUR (5.73%, s.d.= 0.65%) under the adjusted data.

Changes in viral quantities following SARS-CoV-2 infection

Using BQC19 cohort, we compared the prevalence of three viruses (Anellovirus, EBV/HHV-4 and HHV-7) based on the infection status of SARS-CoV-2 within the stratified populations. We observed that SARS-CoV-2 infection was associated with a significant decreased of the prevalence of Anellovirus in AMR ($p=1.03\times 10^{-3}$, OR=0.15, 95%CI=[0.05, 0.47]), EAS ($p=1.30\times 10^{-3}$, OR=0.22, 95%CI=[0.08, 0.57]) and EUR ($p=5.84\times 10^{-5}$, OR=0.60, 95%CI=[0.48, 0.77]) individuals. In contrast, it was associated with a significant increase in the prevalence of EBV/HHV-4 in EUR ($p=5.00\times 10^{-5}$, OR=1.79, 95%CI=[1.35, 2.42], Fig. 3c and Supplementary Tables 11). An opposite trend in the OR between Anellovirus and EBV/HHV-4 was observed in the other populations, although comparisons of them did not reach the significance level. The comparison of the difference was difficult for HHV-7, which had a naturally low prevalence in non-infected individuals.

Table 1

Differences of prevalence depend on geographic differences in the same population.

^aOdds ratio of region 2 to region 1 are shown. ^bCanada (BQC19 SARS-CoV-2 negative)

P values and odds ratios that show the significant differences ($p < 1.25 \times 10^{-2}$) are bolded.

Virus	Population	Recruited location 1	Recruited location 2	Recruited location 1		Recruited location 2		<i>p</i> value	OR [95%CI] ^a
				Total	Adjusted	Total	Adjusted		
					prevalence (%) (s.d.)		prevalence (%) (s.d.)		
Anellovirus	EUR	Australia	Europe	331	18.45 (1.10)	1,998	24.03 (0.53)	3.42×10^{-6}	2.00 [1.48, 2.73]
		Australia	Canada ^b	331	18.45 (1.10)	539	27.31 (0.85)	2.83×10^{-5}	2.12 [1.49, 3.05]
		Europe	Canada ^b	1,998	24.03 (0.53)	539	27.31 (0.85)	0.844	1.02 [0.81, 1.28]
	EAS	Japan	Canada ^b	1,583	58.28 (0.25)	30	76.63 (4.33)	9.01×10^{-3}	3.24 [1.32, 9.22]
EBVHHV-4	EUR	Australia	Europe	331	11.35 (1.06)	1,998	11.74 (0.46)	0.375	1.18 [0.82, 1.75]
		Australia	Canada ^b	331	11.35 (1.06)	539	12.12 (0.74)	0.197	1.34 [0.86, 2.10]
		Europe	Canada ^b	1,998	11.74 (0.46)	539	12.12 (0.74)	0.835	1.03 [0.76, 1.38]
	EAS	Japan	Canada ^b	1,583	14.27 (0.17)	30	13.96 (3.01)	0.452	0.64 [0.15, 1.88]
HHV-7	EUR	Australia	Europe	331	6.00 (0.89)	1,998	9.56 (0.42)	0.867	0.96 [0.59, 1.62]
		Australia	Canada ^b	331	6.00 (0.89)	539	5.73 (0.65)	0.012	0.39 [0.18, 0.82]
		Europe	Canada ^b	1,998	9.56 (0.42)	539	5.73 (0.65)	0.048	0.69 [0.46, 1.00]
	EAS	Japan	Canada ^b	1,583	4.69 (0.12)	30	1.41 (1.65)	0.062	NA

Host genome variants associated with viral prevalence

We evaluated associations between host variants and prevalences of four viruses (Anellovirus, EBV/HHV-4, HHV-6B, HHV-7) in three groups whose sample size was greater than 1,000: EUR of BRIDGET (hereafter BRIDGET-EUR) and BQC19 SARS-CoV-2 positive (hereafter BQC19-pos-EUR) and the Nagahama Study. This analysis was conducted within each cohort; therefore, prevalence was not adjusted. The genome wide association study (GWAS) showed statistically significant associations ($p < 5.00 \times 10^{-8}$) only in the major histocompatibility complex (MHC) region (chr6:28,510,120-33,480,577) (Fig. 4a-c, Fig. 5a-c and Supplementary Fig. 6a-c), including *HLA* genes associated in BRIDGET-EUR with *HLA-DQA1* for Anellovirus, and *HLA-DQA1* and *-DQB1* for EBV/HHV-4; and in the Nagahama Study with *HLA-DQA1* and *-DQB1* for Anellovirus (Fig. 4d-f, Fig. 5d-f, Supplementary Fig. 6d-f and Supplementary Table 12). Subsequently, suggestive associations ($p < 5.00 \times 10^{-6}$) were identified between *HLA* genes and Anellovirus (*HLA-DQB1*), EBV/HHV-4 (*HLA-C* and *-B*) and HHV-7 (*HLA-B* and *-DQA1*) in BRIDGET-EUR, HHV-7 (*HLA-B*) in BQC19-pos-EUR, and Anellovirus (*HLA-B*) and EBV/HHV-4 (*HLA-C* and *-B*) in the Nagahama Study. No significant association was detected between *HLA* genes and HHV-6B in any cohorts (Supplementary Table 12).

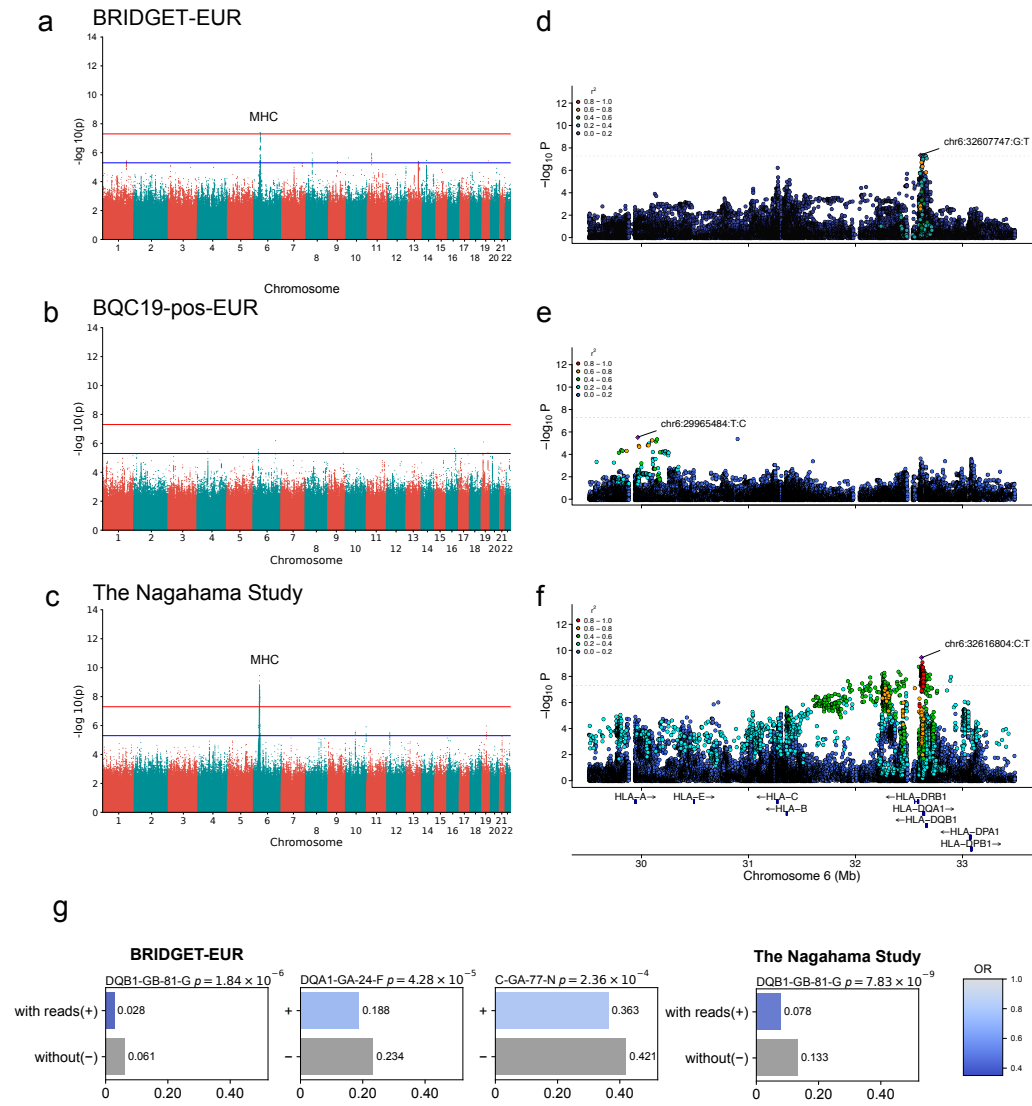


Fig. 4 Genetic associations of Anellovirus with host factors

a-c, GWAS results for Anellovirus prevalence in BRIDGET-EUR (a), BQC19-pos-EUR (b) and the Nagahama Study (c). d-f, Associations between prevalence of Anellovirus and variants within the MHC region in BRIDGET-EUR (d), BQC19-pos-EUR (e), and the Nagahama Study (f). The variant shown in each panel represents the most significant association with viral prevalence. Positions are based on hg38. g, Frequencies (x-axis) of amino acid residues significantly associated with Anellovirus prevalence.

Owing to the low estimation accuracy of genome-wide SNV genotypes for high polymorphic

HLA genes, the associations were further tested in *HLA* alleles. As a result, Anellovirus, EBV/HHV-4 and HHV-7 had significant risk ($OR \geq 1$) associations ($p < 2.84 \times 10^{-4}$) with two and five *HLA* alleles, respectively, while protective ($OR < 1$) associations were observed in eight, 12 and four alleles ($p < 2.68 \times 10^{-4}$), respectively, in at least one cohort (Supplementary Tables 13). For Anellovirus, the strongest associations were observed with *DRB1*13:02* in both BRIDGET-EUR ($p = 1.69 \times 10^{-5}$, $OR = 0.45$, $95\%CI = [0.31, 0.64]$) and the Nagahama Study ($p = 4.98 \times 10^{-9}$, $OR = 0.48$, $95\%CI = [0.37, 0.61]$), while no association was found in BQC19-pos-EUR. Haplotype analysis showed the link ($D' = 0.97$) between *DRB1*13:02-DQA1*01:02-DQB1*06:04* in the Nagahama Study, whereas even no *DQA1* allele showed the association in BRIDGET-EUR, the link ($D' = 0.98$) between *DRB1*13:02* and *DQB1*06:04* were observed (Supplementary Fig. 7 and Supplementary Table 13). EBV/HHV-4 had the strongest associations with *B*07:02*, *B*14:02* and *DQA1*03:01* in BRIDGET-EUR ($p = 5.03 \times 10^{-8}$, $OR = 1.89$, $95\%CI = [1.50, 2.37]$), BQC19-pos-EUR ($p = 1.55 \times 10^{-5}$, $OR = 0.28$, $95\%CI = [0.15, 0.48]$) and the Nagahama Study ($p = 1.50 \times 10^{-8}$, $OR = 2.22$, $95\%CI = [1.68, 2.93]$), respectively. For these three alleles, haplotypes *C*-07:02-B*07:02*, *C*08:02-B*14:02*, and *DRB1*04:03* or *DRB1*04:06-DQA1*03:01-DQB1*03:02* were in high linkage disequilibrium (LD) for BRIDGET-EUR ($D' = 0.94$), BQC19-pos-EUR ($D' = 0.96$) and the Nagahama Study ($D' = 0.96$ for *DRB1*04:03* and $D' = 0.97$ for *DRB1*04:06*), respectively. Both BRIDGET-EUR and BQC19-pos-EUR groups showed a significant association between *B*08:01* and the prevalence of HHV-7. The strongest significant associations with HHV-7, *DQA1*02:01* ($p = 7.20 \times 10^{-6}$, $OR = 0.52$, $95\%CI = [0.39, 0.69]$) and *C*07:01* ($p = 5.20 \times 10^{-5}$, $OR = 1.81$, $95\%CI = [1.35, 2.41]$), and *DRB1*08:03* ($p = 4.69 \times 10^{-5}$, $OR = 2.43$, $95\%CI = [1.56, 3.69]$) were observed in BRIDGET-EUR, BQC19-pos-EUR and the Nagahama Study, respectively. The haplotypes with these alleles were shown in *DRB1*07:01-DQA1*02:01-DQB1*02:02* ($D' = 0.98$) in BRIDGET-EUR and *C*07:01* and *B*08:01* ($D' = 0.97$) in BQC19-pos.

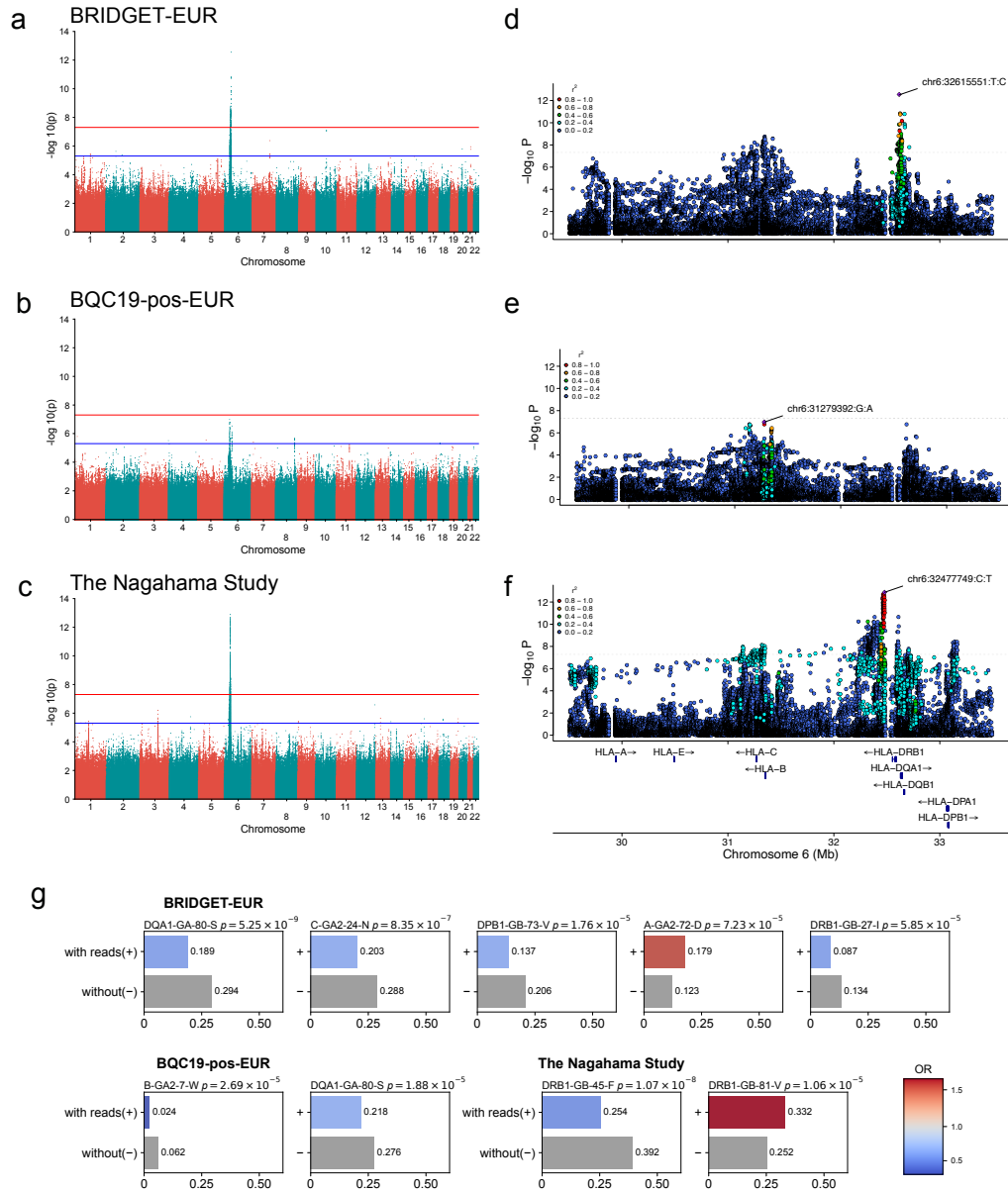


Fig. 5 Genetic associations of EBV/HHV-4 with host factors

a-c, GWAS results for EBV/HHV-4 prevalence in BRIDGET-EUR (a), BQC19-pos-EUR (b) and the Nagahama Study (c). d-f, Associations between prevalence of EBV/HHV-4 and variants within the MHC region in BRIDGET-EUR (d), BQC19-pos-EUR (e), and the Nagahama Study (f). The variant shown in each panel represents the most significant association with viral prevalence. Positions are based on hg38. g, Frequencies (x-axis) of amino acid residues significantly associated with EBV/HHV-4 prevalence.

Viral prevalence–associated HLA amino acid residues

Next, the associations between virus prevalences and HLA amino acid residues corresponding to the genotyped allele name were tested. First, virus prevalence and amino acid residues was tested per aligned position of antigen presentation groove domains (G-DOMAINS) [23]. The association analysis revealed the significant ($p < 2.16 \times 10^{-4}$) and strongest association between the prevalence of Anellovirus and the 81st position in G-BETA domain of HLA-DQB1 (hereafter DQB1-GB-81) in both BRIDGET-EUR ($p = 9.99 \times 10^{-7}$) and the Nagahama Study ($p = 2.92 \times 10^{-8}$) groups (Supplementary Fig. 8). The subsequent analysis with conditioning DQB1-GB-81 identified a significant association at the 24th position in G-ALPHA domain of HLA-DQA1 (DQA1-GA-24, $p = 3.30 \times 10^{-5}$) and the 77th position in G-ALPHA1 domain of HLA-C (C-GA1-7, $p = 2.15 \times 10^{-4}$) only in BRIDGET-EUR. For the EBV/HHV-4 prevalence, DQA1-GA-80 ($p = 1.12 \times 10^{-9}$), C-GA2-24 ($p = 3.81 \times 10^{-7}$), DPB1-GB-73 ($p = 6.08 \times 10^{-6}$), A-GA2-72 ($p = 9.32 \times 10^{-5}$) and DRB1-GB-27 ($p = 1.38 \times 10^{-5}$) in BRIDGET-EUR had significant associations with the virus in the sequence of identified positions. B-GA2-7 ($p = 1.14 \times 10^{-5}$) and DQA1-GA-80 ($p = 1.17 \times 10^{-5}$) were shown to be associated with the prevalence in BQC19-pos-EUR by conditional analysis (Supplementary Fig. 9). In the Nagahama Study group, DRB1-GB-45 ($p = 2.20 \times 10^{-9}$) and DRB1-GB-81 ($p = 1.34 \times 10^{-5}$) were associated with the EBV/HHV-4 prevalence. Lastly, HHV-7 was associated with DRB1-GB-73 ($p = 4.17 \times 10^{-7}$) followed by B-GA2-23 ($p = 5.79 \times 10^{-7}$) and B-GA1-74 ($p = 3.66 \times 10^{-5}$) in BRIDGET-EUR, and B-GA1-9 ($p = 1.16 \times 10^{-4}$) in BQC19-pos-EUR, while no amino acid position was associated with HHV-7 in the Nagahama Study (Supplementary Fig. 10).

The analysis of amino acid residues on the identified positions exhibited the strongest significant association between Anellovirus and glycine at DQB1-GB-81 (DQB1-GB-81-G) in BRIDGET-EUR ($p = 1.84 \times 10^{-6}$, OR=0.40, 95%CI=[0.27, 0.58]) and the Nagahama Study ($p = 7.83 \times 10^{-9}$, OR=0.47, 95%CI=[0.36, 0.61]; Fig. 4g, Supplementary Table 14). With respect

to EBV/HHV-4 prevalence, the strongest significant associations were identified in serin at DQA1-GA-80 (DQA1-GA-80-S, $p=5.25\times10^{-9}$, OR=0.57, 95%CI=[0.48, 0.69]; Fig. 5g), in BRIDGET-EUR, tryptophan at B-GA2-7 (B-GB-7-W, $p=2.69\times10^{-5}$, OR=0.34, 95%CI=[0.20, 0.55]) in BQC19-pos-EUR, phenylalanine of DRB1-GB-45 (DRB1-GB-45-F, $p=1.07\times10^{-8}$, OR=0.52, 95%CI=[0.41, 0.65]) in the Nagahama Study. In BQC19-pos-EUR with the EBV/HHV-4, the condition analysis including amino acid residues of B-GA2-7 demonstrated serine at DQA1-GA-80 (DQA1-GA-80-S, $p=1.88\times10^{-5}$, OR=0.63, 95%CI=[0.51, 0.78]), which were shown as the strongest amino acids residues in BRIDGET-EUR. The strongest significant association was observed between the prevalence of HHV-7 and DRB1-GB-73-V ($p=2.17\times10^{-6}$, OR=0.51, 95%CI=[0.38, 0.67]; Supplementary Fig. 6g) in BRIDGET-EUR, as well as, B-GA1-9-D ($p=1.63\times10^{-5}$, OR=2.09, 95%CI=[1.49, 2.92]) in BQC19-pos-EUR.

Among amino acid positions whose associations with virus were not identified in conditional analysis, DQA1-GA-80 in BQC19-pos-EUR, and C-GA2-24 and DPB1-GB-73 in the Nagahama Study showed significant associations with the prevalence of EBV/HHV-4 ($p<2.16\times10^{-4}$, Supplementary Table 14). The positional arrangement of amino acids corresponding to the haplotypes showed that the links were identified between C-GA2-24-N and B-GA2-7-W ($D'=0.96$) in BQC19-pos-EUR, as well as DRB1-GB-45-F and DPB1-GB-73-V ($D'=0.52$) in the Nagahama Study (Supplementary Fig. 11).

3.4. Discussion

We have established a framework for a DNA viral genome detection from WGS dataset. When applied to three independent cohorts, the inclusion artificial reads showed cohort-specific biases. As inferred from the previous PCR study [12], the WGS data contained only a small number of viral read pairs in most individuals (Supplementary Table 6), highlighting the importance of adjusting viral quantities prior to comparisons.

We applied the framework to WGS data from 7,059 individuals recruited in Europe, Australia, Canada and Japan to evaluate the contributions of geographic and host genetic background to latent infections. For Anellovirus, while the prevalence of EAS was higher than in EUR populations, differences were also observed within EUR (Australia vs Europe or Canada), suggesting that latent infection of Anellovirus may be influenced by both geographical and host genetical factors. A significant difference in the prevalence of EBV/HHV-4 was observed between Europe and Japan, whereas no such difference was found between EUR and EAS within Canada. Therefore, we could not conclude whether geographic or genetic factors contributed to the infection. A previous study reported a higher prevalence of HHV-7 in Japan than in European countries [24], however, it did not distinguish between genetic and geographic contributions. Our study showed that the prevalence in Europe was higher than in both Canada and Japan, while no significant difference was observed between EUR in Europe and Canada, suggesting that geographic factors may have a stronger effect on the prevalence of HHV-7. Different blood cell types were used across cohorts (Supplementary Table 2), as germline WGS does not need consideration of the differences. It is necessary to experimentally evaluate how variations in blood cell types affect viral quantities.

A previous qPCR study showed the decrease of serum viral load of Anellovirus by the onset of COVID-19 [16]. In our study, this result is considered to reflect a decrease of abundance in SARS-CoV-2 infected group leading to falling below the detection limit of viral reads, which in turn resulted in a lower prevalence. A previous Japanese WGS study also showed the correlation between high Anellovirus load (=high abundance) and COVID-19 [25]; however, our study could not evaluate the result due to the low prevalence in EUR and low sample size of EAS in BQC19 cohort. In contrast to Anellovirus, the prevalence of EBV was higher in infected group in EUR, which seems to reflect that the abundance exceeded the detection limit in some individuals. The higher frequency of EBV reactivation during acute COVID-19 [14]

suggests we detected DNA viruses not in a latent state, but in a reactivated state. Further investigation may involve comparing WGS data from individuals recruited before and after the pandemic.

The presence of an association between Anellovirus and DQB1-GB-81-G in both BRIDGET-EUR and the Nagahama Study suggests that the relationship outweighs the influence of geographical and genetic backgrounds. The lack of the association in BQC19-pos-EUR may be attributed by the association between COVID-19 infection and *HLA* alleles [26]. Conversely, the reproducibility of four *HLA* alleles and DQA1-GA-80-S that showed associations with EBV/HHV-4 in BRIDGET-EUR and BQC19-pos-EUR, but not in the Nagahama Study, representing the EAS group, indicates a strong influence of genetic background. Among the four alleles, a protective association between *DQB1*03:01* and antibodies against Epstein-Barr nuclear antigen was reported in a French study [27]. Most individuals carrying the DQA1-GA-80-S also carry the corresponding *DQA1*05:05*, which is in strong LD with *DQB*03:01* ($D' > 0.91$). Therefore, this amino acid may be identified as a confounding factor of *DQB*03:01*. Unlike the two viruses, associations between *HLA* alleles and HHV-7 were not shared across the three groups, except for *B*08:01*. Although the effect of SARS-CoV-2 infection should be taken into account for BQC19-pos-EUR, this result may imply a geographical influence on the association between HHV-7 and HLA. The underlying factors related to genetic and geographical backgrounds may involve either host or viral factors, or both, as well as differences in past environmental exposures. A careful elucidation of the causal mechanisms is therefore warranted.

The viral genome reference based on representative sequences, used in the pipeline except for Anellovirus, may reduce detection power; thus, the use of a redundant GenBank reference [28] could be considered, despite the increased computational cost. While WGS data may contain novel DNA viruses that could be constructed by *de novo* assembly of unmapped reads, their

low prevalence and read counts would require far more samples than were available in this study. As an additional limitation, due to its limited dynamic range reaching the detection limit makes it unlikely to be suitable for individual-level diagnostics.

The framework established in this study can be applied to WGS datasets from past and future studies. Aggregating association findings derived from viral DNA detected in WGS may aid in identifying novel factors in case-control studies, potentially reflecting host immune responses.

3.5. Methods

Whole genome sequencing (WGS) and mapping to the human genome reference

DNA samples from human blood were obtained from BRIDGET (buffy coat and whole blood) and the Nagahama Study (buffy coat) before the end of 2019, and from BQC19 (whole blood) from 2020 to 2022, with written informed consent provided by all participants. During DNA preparation, PhiX (Illumina, PhiX Sequencing ControlV3) was spiked in BRIDGET samples, and Kapa-tag (Kapa Biosystems) were added to BRIDGET samples and 1,250 BQC19 samples (Supplementary Table 2). Sequencing was performed at an average coverage of 30× for BRIDGET and BQC19 samples, whereas samples from the Nagahama Study were sequenced at an average coverage of 15 (Supplementary Table 2).

The human genome assembly GRCh38 (GCA_000001405.15) and 10,532 viral genome sequences from the NCBI Reference Sequence Database (RefSeq release 91) were used as reference sequences for human and viral genomes, respectively. The homology between the two references was evaluated using blastn with default parameters. A sequence was considered homologous if the alignment satisfied either of the following criteria: (1) The aligned region covered more than 70% of any viral reference sequence, or (2) The aligned region covered more than 70% of any contig in the human genome assembly. Sequences fulfilling criterion (1)

were classified as viral integrations incorporated into the human genome, whereas those meeting criterion (2) were classified as decoy sequences.

The GenPipes DNaseq pipeline (version 3.1.2) [19] with bwa mem was employed to map the sequenced reads. Mapping was performed against the human genome reference GRCh38. Haplotypes were called using GATK (version 3.7) [20]. Paired-end reads were defined as “unmapped read pairs” if both forward and reverse reads were not mapped to the human genome reference. Next, paired-end reads mapped to decoy sequences were extracted and re-mapped to the GRCh38 sequence, excluding the decoy sequences. Paired-end reads that remained unmapped during this step were reclassified and included in the “unmapped read-pairs” category.

Variants from BRIDGET and the Nagahama Study were aggregated using GenomicsDB Import in GATK (version 4.2) from chromosomes 1 through 22, as well as the X and Y chromosomes. Variants for BQC19 were extracted from a previously published study [29].

Quality Control of the variant calls

Quality control of the variant calls was performed using GATK (version 4.2) with parameters following GATK best practice [20]. Initially, bi-allelic variants were extracted based on a Phred-scaled quality score from the exact test of excess heterozygosity greater than 54.69 and an inbreeding coefficient “inbreeding coeff” less than -0.8. Subsequently, model-based variant recalibration was performed in the SNP mode. Finally, variants with a variant quality score log-odds (VQSLOD) sensitivity below 99.7% were excluded.

To filter out false-positive calls due to extremely low or high depth, we compared the genotyping results of WGS with those obtained using genome-wide SNP arrays. These comparisons included 579 samples from BRIDGET, 2,233 samples from BQC19 (Affymetrix

Axiom Precision Medicine research array, Affymetrix, USA), and 1,564 the Nagahama Study samples (Illumina Asian Screening Array, Illumina, USA, Supplementary Table 15). For each cohort, all individuals had a call rate > 0.9 and SNVs with a call rate ≥ 0.95 were extracted. Let l and h be the thresholds for the total read depth (DP) calculated from the WGS mapping results. We then computed the concordance rates c_d of the genotypes (0/0, 0/1, 1/1, or ./.) between WGS and arrays for all extracted SNVs and samples at each $DP = d$. The thresholds l and h were optimized as:

$$\hat{l}, \hat{h} = \underset{l, h}{\operatorname{argmax}} h - l, c_d \geq 0.99 \text{ for } l \leq d \leq h.$$

Supplementary Fig. 12 illustrates the changes in concordance rates as a function of DP, as well as the estimated l and h . We assigned d_i^j and q_i^j to represent the DP and the phred-scaled probability of Genotype Quality (GQ) of the i -th variant in the j -th sample, respectively. In this study, we assumed that the genotype of the i -th variant of the j -th sample was not called (./.) if $d_i^j < \hat{l}, \hat{h} < d_i^j$ or $q_i^j \leq 20$.

SNV and Individual selection

From the variants and genotypes that passed quality control, SNVs on the host genome and individuals were selected using Plink (version 1.9) [30], following these steps, whereby the variants from BRIDGET and the Nagahama Study were individually allocated to distinct datasets. First, variants in autosomal chromosomes were extracted, and the missing call rate for each individual m denoted as r_m was calculated. The missing call rate of k -th variant ($k = 1, \dots, K$) in all individuals in the dataset was denoted as v_k . The probability density function P of v was estimated using kernel density estimation (KDE) with Gaussian kernels, implemented in the Python (version 3.10) module “statsmodels” (version 0.14.0) [31], based on a random sample of 1,000,000 variants. The cumulative distribution function (CDF) of P ,

denoted as $F_v(x)$, was derived and variants with v_k values greater than $F_v(0.95)$, were excluded. Next, SNVs that passed the Hardy-Weinberg equilibrium (HWE) exact test ($p < 1.0 \times 10^{-6}$). These steps were applied independently in each cohort, resulting in 73,345,839, 125,639,463 and 73,420,545 SNVs for the GWAS in BRIDGET, BQC19 and the Nagahama Study cohorts, respectively. Related individual pairs with $PI_HAT > 0.9$ were excluded, as well as individuals without age information.

Populations assignment

Individuals in this study were assigned to one of six superpopulations (AFR: African, AMR: admixed American, CSA: Central/South Asian, EAS: East Asian, EUR: European and MID: Middle Eastern) as defined in the 1000 Genomes project and Human Genome Diversity Project (HGDP) [21]. We obtained vcf files from the Genome Aggregation Database (gnomAD), which included jointed calls based on 948 samples from the 1000 Genomes Project and 3,202 samples from HGDP. From these datasets, we extracted 165,348 bi-allelic SNVs on autosomal chromosomes that met the thresholds and the 3,407 unrelated samples defined in the study [21]. Quality checks of the remaining SNVs and individuals were performed using the same procedure as in the SNV and individual selection step, with the exception of additional filtering of SNVs annotated to high-LD SNV pairs ($r^2 > 0.1$) that were within 500Kbp of each other and located in high-LD regions as defined in a previous study [32]. Additionally, SNVs with missing call rates greater than 0 were extracted. These remaining SNVs and individuals were subjected to the same procedure described above, with an additional filtration step for SNVs located in the high-LD region. Finally, we selected SNVs that passed the HWE test ($p < 1.0 \times 10^{-6}$) in each cohort (BRIDGET, BQC19 and the Nagahama Study) and retained those present in both the 1000 Genomes project and HGDP datasets, after applying the filtering steps described above.

PCA was conducted on the genotyped dataset of selected SNVs from the 1000 Genomes Project and HGDP, as provided by gnomAD [21], and the resulting factor loadings were extracted. SNV datasets of individuals from the three cohorts were then projected onto the top three PCs using these loadings. For the projection, “not-called” genotypes were replaced with the mean of called genotypes for the same SNV within each cohort. Individuals from the three cohorts were assigned to one of six populations using the k-nearest neighbour (k-NN) algorithm based on their scores in the PC space. The parameter k of k-NN, representing the number of nearest items, was determined using samples with assigned population from either the 1000 Genomes project or HGDP as the training dataset. For each value of k from 1 to 20, training samples were assigned to population using k-NN with leave-one-out cross-validation based on the Euclidian distance of the top three PCs. The F1 score for population assignment accuracy was calculated for each of the six populations, and the minimum F1 score among them was selected. The optimal k was chosen as the value that maximized the minimum F1 score. In this study, $k = 5$ was selected, and individuals from the three cohorts were assigned to populations based on their five nearest neighbors among the training samples (Supplementary Fig. 13a).

To identify DNA sample contamination, we estimated the CHARR value [33] for all individuals using the SNVs included in the PCA, with allele frequencies retrieved from TogoVar [34]. Individuals with $\text{CHARR} > 0.05$ were excluded from the analysis as contaminated samples.

Sex calling from WGS data

Sex calling was based on the average coverage and heterogeneity of autosomal and sex chromosomes, as estimated by Somalier [35]. The average coverages of the sex chromosomes were normalized by the average coverage of all autosomal chromosomes, defining norm-X and

norm-Y values. Individuals with $\text{norm-Y} > 0$ were called male. Subsequently, individuals with $\text{norm-X} > 0.75$ were called females (Supplementary Fig. 13b). For the remaining 14 individuals (three from BRIDGET, ten from BQC19 and, two from the Nagahama Study), who were sequenced in different WGS run batches, the average coverages of 21q21.2 and, Yq11.223 were calculated using GATK (version 4.2) DepthOfCoverage. The average coverage of Yq11.223 was normalized by that of 21q21.2, defining norm-Yb. 11 individuals with a $\text{norm-Yb} > 0.01$ individuals were called male, while the remaining three individuals were called female (Supplementary Fig. 13c).

Detection of viral reads from unmapped reads

Unmapped read pairs extracted from the WGS data were trimmed using Trimmomatic (version 0.35) in paired-end mode with the parameters, “TruSeq2-PE. fa:2:30:15 TRAILING:30 MINLEN:50” [36]. Trimmed reads were subsequently aligned to the PhiX sequence (NC_001422.1, 5,386 bp) using blastn (version 2.10.0+) [37] with options -qcov_hsp_perc 100 -perc_identity 99, and any matching reads, along with their paired read, were excluded. Similarly, reads derived from Kapa-tags (reference of Kapa-tag, 58-60 bp per a tag) were excluded using blastn with adjusted options -qcov_hsp_perc 100 -perc_identity 98 -word size 50. The remaining read pairs were annotated against 10,532 virus references from the NCBI Reference Sequence Database (RefSeq release 91) by blastn with default parameter. For Anellovirus reference, 4,746 redundant sequences were collected from the NCBI viral database with the taxid as “687329” and “Nucleotide Completeness” as ‘complete’ (December 8, 2022)[38]. Phage sequences included in Refseq were identified based on the accession IDs from hosts recognized as bacteria in the taxonomy table “taxid10239.nbr” downloaded from <https://www.ncbi.nlm.nih.gov/genome/viruses/>) on May 23, 2020 and through keyword search in the Nucleotide database (<https://www.ncbi.nlm.nih.gov/nucleotide/>) for

“phage”, “bacteriophage”, “virophage,” and “prophage” as of the release date of Refseq version 91.

Assume that the base positions from $s_{k_\alpha}^v$ to $e_{k_\alpha}^v$ on the read k_α are aligned to the v -th virus ($v \in \{1, \dots, V\}$). The index α is f and r , representing the forward and reverse directions of the read, respectively. Let L_{k_f} and L_{k_r} be the lengths of k_f and k_r , respectively. Additionally, we denote $b_{k_\alpha}^v, p_{k_\alpha}^v, i_{k_\alpha}^v, d_{k_\alpha}^v$ and $m_{k_d}^v$ as the bitscore, E-value, and counts of insertions, deletions, and match nucleotides, respectively, from the blastn result between k_d and the v -th virus. A read pair k_f and k_r was assigned to the v -th virus if it satisfied the following criteria.

1. $b_{k_f}^v \geq b_{k_f}^{v'} \cup b_{k_r}^v \geq b_{k_r}^{v'}$ for $\forall v' \in \{1, \dots, V\}$.
2. $p_{k_f}^v < 1.0 \times 10^{-10} \cap p_{k_r}^v < 1.0 \times 10^{-10}$.
3. $\frac{e_{k_f} - s_{k_f} + 1}{L_{k_f}} > 0.9 \cap \frac{e_{k_r} - s_{k_r} + 1}{L_{k_r}} > 0.9$.

Aggregation of homological viruses

The viral sequences, to which unmapped read pairs were assigned in at least one individual, were aligned using blastn. Let L_{v_k} and L_{v_l} denote the sequence lengths of viruses v_k and v_l , respectively. Additionally, $i_{k,l}$, $d_{k,l}$ and $m_{k,l}$ represent the counts of insertions, deletions, and match nucleotides, respectively, obtained from the blastn alignment with v_k as the query and v_l as the subject. The sequence similarity between v_k and v_l , denote as $r_{k,l}$ is then defined as:

$$r_{k,l} = 0.5 \left(\frac{m_{k,l}}{L_{v_k} + i_{k,l} - d_{k,l}} + \frac{m_{l,k}}{L_{v_l} + i_{l,k} - d_{l,k}} \right).$$

Viruses v_k and v_l were merged into a single homogenous virus if $r_{k,l} > 0.9$. The RefSeq IDs of the viruses merged into each group are provided in Supplementary Table 5.

Calculation of viral quantities

For quantitative comparisons between geographical or genetic backgrounds, the prevalence of each virus was calculated. Prevalence was defined as the proportion of individuals within a cohort or population who had at least one read pair assigned to the virus.

Viral quantity in the blood cells of each individual was evaluated in terms of abundance. Let n_v^i be the number of reads (twice the number of read pairs) assigned to viral genome v in the WGS reads of individual i , and l_v represent the genome size of v . Similarly, n_h^i corresponds to the number of reads mapped to the human genome reference in the WGS reads of individual i and l_h refers to its diploid genome size (3,099,750,718×2). The abundance of virus v for individual i , a_v^i is defined as follows:

$$a_v^i = \frac{\frac{n_v^i}{l_v}}{\frac{n_h^i}{l_h}}.$$

In addition, we evaluated the total viral quantity of an individual as the number of viral read pairs assigned per million WGS read pairs.

Detection of inherited integration and outliers

Inherited chromosomal integration (ici) is defined as the insertion of partial or full viral sequences into the human germline genome [39]. First, viruses observed in at least ten individuals in each cohort were selected. We assumed that the abundance distribution for integrated proviruses exhibits a distinct mode around 1.0, independent of the distribution for

unintegrated viruses. Using the assigned viral reads, the probability density function of abundances was estimated under the assumption of a bimodal distribution. This estimation was performed using the `locmodes` function from the “multimode” package (version 1.5) [40] in R (version 4.4.0). Let $d(x)$ be the estimated probability density function of the bimodal distribution with modes $m_i, i = 1, 2, m_1 < m_2$. We then identified two positions $q_i^l \leq q_i^r$ that satisfied the following criteria:

$$h(q_i^l) = h(q_i^r) = 0.95 \times h(m_i).$$

We concluded that a virus included individuals with ici if the distribution satisfied the following criteria.

$$0.9 < h_i < 1.1 \ (i = 1 \text{ or } 2).$$

$$h_i > 0.05 \times h_{3-i}.$$

Next, the detected ici cases were classified as either HHV-6A or HHV-6B. Unmapped reads were mapped to HHV-6A and HHV-6B sequences for each individual using `bwa mem` with the parameters `-Y -K 1000000000`. Assign N_A and N_B to the sets of read pairs mapped to HHV-6A and HHV-6B, respectively. From the total set of mapped read pairs for each individual, denoted as $N_T = N_A \cup N_B$, the proportions of HHV-6A and HHV-6B were calculated as:

$$p_A = \frac{|N_A|}{|N_T|}, p_B = \frac{|N_B|}{|N_T|},$$

where the operator $|| \cdot ||$ means the number of elements in the set. We considered $N_B = \emptyset$ if $p_B - p_A > 0.5$ and $p_B > 0.9$, indicating that no reads derived from HHV-6B were obtained for the individual. Similarly, $N_A = \emptyset$ if $p_A - p_B > 0.5$ and $p_A > 0.9$. Otherwise, both N_A and N_B were retained and assigned to HHV-6A and HHV-6B, respectively. From the sets N_A or N_B , the abundance of HHV-6A or HHV-6B was calculated, adjusting for viral genome length by considering the number of positions where at least one read was mapped.

Finally, an individual was considered to have iciHHV-6A or HHV-6B if the corresponding abundance ranged from 0.3 to 1.5, consistent with solo-direct repeat (DR) HHV-6 integration [22].

After excluding abundance values corresponding to iciHHV-6, outliers were identified based on the distribution of abundances within each cohort and virus. In this study, we assumed that, after excluding outliers, abundances followed a generalized gamma distribution [41]. Let $A = (a_1, \dots, a_{n_s})$ be the set of abundances, where n_s is sample size of the cohort. Then, set A was fitted to a generalized gamma distribution using `scipy.stats.goodness_of_fit` function with “SciPy” (version 1.11.2) [42] module in Python. Using the estimated distribution parameters, right-tailed percentiles of the abundances were calculated, and abundances with percentiles exceeding $1 - 0.05/n_s$ were classified as outliers. These outliers were excluded from the association of viral quantities and geographical location or with populations.

Exclusion of artificial viral reads

For each cohort and virus, we re-mapped the read pairs assigned to the virus from all individuals in the cohort to the virus's reference genome using `bwa mem` with the options `-Y -K 100000000`. For viral species represented by multiple reference genomes, the mapping was conducted separately for each genome. In Anelloviruses, which include 2294 subtypes, genomic positions were standardized based on open reading frame (ORF) region. ORF coordinates for 2,144 annotated subtypes were retrieved using Biopython Entrez.efetch [43]. For the remaining 150 subtypes lacking annotated ORF information, genomic positions were aligned using `blastn` against the best-matching viral sequence with known ORF annotations. The ORF start and end nucleotides were defined as the 500th and 2500th positions of each sequence, respectively, based on the derived alignment results. The mapped region of

each read on a sequence was adjusted for offset and read length relative to the defined start and end positions, and coverage at each sequence position was calculated accordingly, including the upstream and downstream 500 bp regions from the ORF. The orientation of sequences within the multi-FASTA file was detected using MAFFT (version 7.526) [44] with the --adjustdirection option. If the strand orientation of a sequence was determined to be reversed, the mapped read positions were flipped accordingly.

We then calculated the proportion of positions in the reference genome covered by at least one read. If this rate was below 0.85, the distribution of coverage across the genome was tested to identify and exclude potential artificial reads. Let $c_p^v, p = 1, \dots, L_v$ be the coverage at position p , where L_v is the genome size of virus v . Denote $r_k^v, k = 1, \dots, K_v$ as the reads assigned to virus v . We calculated the distribution of coverage depth $d_i^v, i \in N^0$ defined as:

$$d_i^v = \frac{1}{R_a} \sum_{p=1}^{L_v} I(c_p^v = i), I(c_p^v = i) = \begin{cases} 1, & \text{if } c_p^v = i, \\ 0, & \text{otherwise,} \end{cases}$$

where R_a is the average length of the paired-end reads in K_v . We assumed that the distribution $D_v = (d_0^v, d_1^v, \dots)$ followed a Poisson distribution, provided that no artificial reads were included in the mapped reads. The distribution D_v was then evaluated using a generalized goodness-of-fit test under the non-central chi-square distribution with non-centrality parameter δ [45]. Assign e_c^v to the expected probability of coverage $c \in \{1, \dots, C\}$, where $C = \max \{i = 0, 1, \dots | d_i^v > 0\}$, under a Poisson distribution with parameter $\lambda = \frac{R_a K_v}{L_v}$. The statistic E was then calculated by:

$$E = \sum_{c \leq C} \frac{(d_c^v - \delta e_c^v)^2}{e_c^v}.$$

The percentile of E under the non-central chi-square distribution with $\delta = 1.5$, and $df = C - 1$, was then calculated using `scipy.stats.ncx2`. If the percentile exceeded 0.95, we

concluded that D_v did not follow a Poisson distribution. When the null hypothesis was rejected, the position with the maximum coverage on the reference genome

$$p_m = \underset{p}{\operatorname{argmax}} c_p^v,$$

was estimated. Paired-end reads, in which either or both reads mapped to points P_m defined by $P_m = \{p \in \{1, \dots, L_v\} \mid c_i^v > 0 \text{ for } p \leq i \leq p_m \text{ or } p_m \leq i \leq p\}$, were removed as artificial reads. After removal, c_p^v and d_i^v were calculated, and D_v was re-tested. This testing and the removal process was iterated until the test was accepted. For species represented by multiple reference genomes, unfiltered reads from all genomes were combined and treated as a union set, provided that at least one genome passed the exclusion filtering based on its individual mapping result.

Unbiased comparison of viral prevalences and abundances

Let N be the total read count of the WGS and $d_t^v > 0$ and d_o^v be the true and observed abundance of virus v , respectively. Suppose N_h and N_v are the number of reads mapped to the human and viral genome sequences, respectively, and L_h and L_v as the sizes of the human and virus genomes, respectively. Considering that $N \approx N_h$, $N_v \propto N$, $L_v \ll L_h$ and $d_o^v \doteq \frac{N_v L_h}{N L_v}$.

It follows that when $d_t^v L_v$ is extremely low compared to $a L_h$, where a is the average coverage of the human genome sequencing, $d_o^v = 0$ in certain cases due to $N_v = 0$. Therefore, the prevalence is underestimated in instances where $d_o^v = 0 \cap d_t^v > 0$. This also truncates the left tail of the abundance distribution for $d_t < \frac{L_h}{N L_v}$. By definition, the frequency of $d_o^v = 0 \cap d_t^v > 0$ is inversely proportional to the total read count N .

In this study, we developed a correction method for prevalence and abundance based on random sampling, to enable unbiased comparisons between individuals with widely varying total read

counts. Assume that WGS was performed on M individuals and the total read count T_m for each individual was obtained from the sequencing results. The set of reads mapped to virus v , denoted as $R_m^v \subset T_m$ was also obtained for individual $m = 1, \dots, M$. For an individual m , define random variable X that follows a normal distribution with mean μ and variance σ^2 , where μ and σ^2 are estimated from the total read counts of a control group. A realization x_m is generated from X , and a sampling ratio $\alpha_m = \min\left(\frac{x_m}{N_m}, 1\right)$ is calculated, where N_m is the total number of reads in T_m . Subsequently, a random variable U following a uniform distribution on the support $[0,1]$ is defined. For each element $r_{i,m}^v \in R_m^v, i = 1, \dots, n_m^v$, a realization $u_{i,m}^v \sim U$ is generated. The adjusted read count $\hat{n}_m^v$ is then given by:

$$\hat{n}_m^v = \sum_{i=1}^{n_m^v} I(u_{i,m}^v < \alpha_m).$$

The adjusted abundance for the virus v in m -th individual is calculated from $\hat{n}_m^v$ and $\alpha_m N_m$.

To reduce random fluctuations, sampling from X and U is iterated 1,000 times for each virus and individual (Supplementary Fig. 14). Let $\hat{d}_{k,m}^v, k = 1, \dots, 1000, m = 1, \dots, M$ be the abundance of virus v in the k -th iteration for the m -th individual calculated from $\hat{n}_m^v$. For each iteration k , the abundances of the M individuals are sorted in descending order as:

$$\hat{d}_{k,o_1}^v \geq \hat{d}_{k,o_2}^v \geq \dots \geq \hat{d}_{k,o_M}^v.$$

The adjusted abundances of the M individuals are obtained by:

$$\bar{d}_i^v = \text{median}_{j=1, \dots, 1000} \hat{d}_{j,o_i}^v, i = 1, \dots, M.$$

In addition, the prevalence in the k -th iteration, $\hat{p}_k^v$ is given by:

$$\hat{p}_k^v = \frac{1}{M} \sum_{i=1}^M I(\hat{d}_{k,i}^v > 0), k = 1, \dots, 1000.$$

and the adjusted prevalence is expressed as their mean $\bar{p}^v = \frac{1}{1000} \sum_{k=1}^{1000} \hat{p}_k^v$.

Downsampling simulation using cohort WGS data

We evaluated the correction method using BRIDGET WGS dataset. The dataset was split evenly into two subsets, Set 1 and Set 2. For each WGS in Set 1 and Set 2, half of the reads were randomly extracted and designated as Set 1h and Set 2h, respectively. For each set, prevalence and abundance were calculated, with values for the half sets averaged over 1,000 iterations of read extraction using different random seeds. Next, the adjusted prevalence and abundance were calculated for each set using $X \sim N(465658879.13, 44673404.23)$, which was estimated from the distribution of half the total read counts in the WGS dataset of BRIDGET. Prevalences and abundance were calculated for Anellovirus, EBV/HHV-4, HHV-6B and HHV-7, and comparisons were conducted between before and after the adjustment. Let $\mathbf{d}_1 = \{\check{d}_{i,1}^v | \check{d}_{i,1}^v > 0, i = 1, \dots, M_1\}$ and $\mathbf{d}_2 = \{\check{d}_{j,2}^v | \check{d}_{j,2}^v > 0, j = 1, \dots, M_2\}$ be the vectors of adjusted or unadjusted abundance values greater than zero for M_1 and M_2 individuals from two different sets, respectively. These abundance values were transformed to a logarithmic scale of 10 and then compared using Student's t -test between $\mathbf{d}_1$ and $\mathbf{d}_2$. Next, let $\check{p}_1^v$ and $\check{p}_2^v$ represent the adjusted or unadjusted prevalences for N_1 and N_2 individuals from the two sets, respectively. The prevalence between the two sets was tested using the Fisher's exact test based on a 2×2 table with values of $\check{p}_1^v N_1$, $N_1 - \check{p}_1^v N_1$, $\check{p}_2^v N_2$ and $N_2 - \check{p}_2^v N_2$. For both statistical tests, Bonferroni corrected significance threshold of $0.025 = 0.05/2$ was applied.

Association analysis of viral reads between geographical or phenotypic groups

EBV/HHV-4, HHV-6B, HHV-7 and Anellovirus, all of which had a prevalence greater than 1% in all BRIDGET, BQC19 and the Nagahama Study were tested for associations with geographical or phenotypic differences. In the correction of viral quantities for each cohort,

N (361922901.62,35197511.96), which was estimated from the distribution of the total reads in the Nagahama Study, was used.

The adjusted abundance $\bar{d}^v$ defined in the previous section was assigned to each individual using the following procedure. Let $o_{k,m}^v$ be the descending rank of $\hat{d}_{k,m}^v$ for individual m in the k -th iteration. In cases where multiple individuals had tied abundances, the minimum rank among them was assigned to all tied individuals. We then calculated the mean rank $\bar{o}_{k,m}^v$ across 1,000 iterations as:

$$\bar{o}_m^v = \sum_{k=1}^{1000} o_{k,m}^v.$$

Finally, the adjusted abundance for individual m was determined by:

$$\tilde{d}_m^v = \bar{d}_{r_m}^v, r_m = \sum_{i=1}^M I(\bar{o}_m^v \geq \bar{o}_i^v).$$

First, the association between prevalence and age or sex was tested within each group. We assigned $y \in \{0, 1\}$ to an incidence of an individual with positive abundance ($y = 1$ for $\tilde{d}_m^v > 0$) or zero abundance ($y = 0$ for $\tilde{d}_m^v = 0$) as the response variable. However, the incidence y was exceptionally set to 0 in the following case.

$$y = 0, \text{ if } r_m > \lfloor \bar{p}^v M + 0.5 \rfloor,$$

where $\lfloor \cdot \rfloor$ means the floor function. Additionally, we defined $\mathbf{x} = (x_1, x_2) \in \{0, 1\} \times \mathbb{R}^+$, where x_1 and x_2 denote sex (male=0, female=1) and age (years), respectively, as explanatory variables. We assumed that the posterior probability of a positive abundance given $\mathbf{x}$ is calculated by:

$$P(y = 1|\mathbf{x}) = \frac{\exp(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2)}{1 + \exp(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2)}, P(y = 0|\mathbf{x}) = 1 - P(y = 1|\mathbf{x}),$$

where $(\alpha_0, \alpha_1, \alpha_2) \in \mathbb{R}^3$ are parameters for the explanatory variables. The parameters α_1 and α_2 were estimated using the maximum likelihood estimation (MLE) via the glm package in R. To test the association with the prevalence of a given virus, a likelihood ratio test was

applied under the null hypothesis, either $\alpha_1 = 0$ for sex or $\alpha_2 = 0$ for age. To evaluate the association between abundance and age, the Spearman rank-order correlation coefficient and the corresponding p value were used with `scipy.stats.spearmanr`. Considering both age and sex, we modeled a random variable Z representing the abundance greater than zero and transformed to a logarithmic scale of 10, followed by a Gaussian distribution with mean $(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2)$ and variance σ^2 . Let z be a realization of Z ; the parameters were estimated using MLE under a generalized linear model defined as:

$$z = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2.$$

The abundance model was evaluated using the likelihood ratio test with $\alpha_1 = 0$ or $\alpha_2 = 0$ as the null hypothesis.

Next, prevalence was compared between two groups. A variable $x_3 \in \{0, 1\}$ was defined as an indicator distinguishing the two groups, and the model was extended to:

$$P(y = 1|\mathbf{x}) = \frac{\exp(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \beta x_3)}{1 + \exp(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \beta x_3)}, \beta \in R, P(y = 0|\mathbf{x}) = 1 - P(y = 1|\mathbf{x}).$$

The prevalence between the two groups was compared using a likelihood ratio test with $\beta = 0$ as the null hypothesis.

For abundance, a Gaussian distribution with mean $(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \gamma x_3)$ and variance σ^2 was assumed for Z . The parameters were then estimated using MLE within the generalized linear model of

$$z = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \gamma x_3,$$

where z is a realization of Z . Finally, the abundances between the two groups were tested using a likelihood ratio test with $\gamma = 0$ as the null hypothesis.

The significance level was set to $0.0125=0.05/4$ for the tests of α_1 and α_2 , $8.33 \times 10^{-3}=0.05/4C_2$ for geographic differences, $0.0125 = 0.05/4$ for geographic difference within the

same population and $3.33 \times 10^{-3} = 0.05/6C_2$ for the difference in infection status of SARS-CoV-2, respectively.

Genome wide association study of viral quantities

In the GWAS, the SNVs used for individual selection were further filtered to include only those whose minor allele frequency is equal or greater than 0.01 in each set. Among individuals in the sets, we estimated 15, 85 and 24 kinship pairs in EUR of BRIDGET, BQC19 SARS-CoV-2 positive and the Nagahama Study, respectively, using the filtered SNVs and genome-wide complex trait analysis, based on genetic relatedness coefficients greater than 0.2. Additionally, PCA was performed using PLINK (version 1.9) in each set. In both the kinship and PCA analyses, SNVs with high-LD ($r^2 > 0.1$) within 500Kbp of each other or located in high-LD regions as defined in a previous study[32] were excluded.

The association between viral prevalence and the dosage of a given SNV allele was tested in the GWAS. Let y_i be the presence of viral reads for a given virus (0: absence, 1: presence) in the WGS of individual i . Suppose that $z_{k,i} \in \{0,1,2\}$ is a allele dosage of k -th SNV for individual i . We assume that the posterior probability of y given z_k is expressed by

$$P(y = 1|z_k, x_1, x_2, \mathbf{t}_L) = \frac{\exp(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha'_L \mathbf{t}_L + \beta z_k)}{1 + \exp(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha'_L \mathbf{t}_L + \beta z_k)}, \quad \boldsymbol{\alpha}_L = (\alpha_{L,1}, \dots, \alpha_{L,L})',$$

where x_1 and x_2 denote sex (male=0, female=1) and age, respectively. The vector $\mathbf{t}_p$ represents the factor scores of the top L principal components, where L was estimated for each GWAS dataset using the elbow method, as implemented in kneed [46] with default parameters. The parameters $(\alpha_0, \alpha_1, \alpha_2, \alpha'_L, \beta)$ were estimated using a binomial logistic regression. To eliminate kinship bias, if individuals in a kinship pair had the same y value, the individual with the lower total read count was excluded from the regression analysis. Finally,

the association between viral prevalence and the k -th SNV was evaluated using PLINK with `-logistic` option, applying genome-wide significance and suggestive thresholds [47] of 5.00×10^{-8} and 5.00×10^{-6} , respectively.

Identified SNVs associated with viruses were annotated using TogoVar [34]. The `locuszoomr` (version 0.3.6) [48] R package was used to visualize the LD between the top associated SNV and neighboring variants.

HLA genotyping and association study

HLA alleles of eight classical genes (*HLA-A*, *-B*, *-C*, *-DRB1*, *-DQA1*, *-DQB1*, *-DPA1*, *-DPB1*) and *HLA-E* were genotyped from WGS data using HLA-HD (version 1.3.0) [49]. Raw FASTQ files were used for the genotyping in BRIDGET and the Nagahama Study, while read pairs in which at least one read was mapped to the full-length sequences of *HLA* alleles recorded in IPD-IMGT/HLA database (release 3.48.0) [50] using Bowtie2 (version 2.3.5) [51] with the `--very-fast` parameter, were used for BQC19. The three-field *HLA* allele names genotyped by HLA-HD were translated into two-field names. The association between viral prevalence and *HLA* allele dosage was evaluated using a generalized linear model consistent with the model applied in the GWAS. A total of 164 *HLA* alleles with allele frequency > 0.1 were tested, with a Bonferroni-corrected significance threshold of $3.05 \times 10^{-4} = 0.05/164$. Haplotypes comprising the nine *HLA* genes were estimated using PHASE [52] in each set, with alleles having a frequency of less than 0.01 merged into a single cluster ("Others"). The relative LD coefficient D' was calculated between associated *HLA* alleles [53], and pairs with $D' > 0.5$ were considered to exhibit strong LD. The estimated haplotypes were visualized using `ggsankey` R package (version 0.0.99999), retaining significantly associated *HLA* alleles and those with a population frequency $\geq 7\%$, while grouping the rest as "Others."

Next, associations between viral prevalence and amino acid residues in MHC groove domains (G-DOMAINS) were tested. Initially, the association was tested at each amino acid position as defined by the IMGT unique numbering [23]. Let the dosage vector of amino acid residues at position j be denoted by $\mathbf{z}_j = (z_{j,1}, \dots, z_{j,m_j})'$, $z_{j,i} \in \{0,1,2\}$, where m_j is the number of amino acid residues at j -th position with frequencies > 0.01 , excluding the most frequent amino acid residue in the control group. We assigned dosages of the amino acid residues from $z_{j,1}$ to z_{j,m_j} , and estimated the corresponding parameters $\boldsymbol{\beta} = (\beta_1, \dots, \beta_{m_j})'$ using binomial logistic regression with replacing βz_k in the GWAS model to $\boldsymbol{\beta}'\mathbf{z}_j$. For the estimated model, the log likelihood ratio test was performed with the $\boldsymbol{\beta} = \mathbf{0}$ as the null hypothesis, and then the most significantly associated amino acid position was identified among the nine *HLA* genes. This analysis was performed for 231 amino acid positions where $m_j \geq 1$, applying the significance threshold of $2.16 \times 10^{-4} = 0.05/231$. Subsequently, the binomial logistic regressions and log likelihood ratio tests were applied with conditioning the most significantly associated amino acid position. This step was continued until no further significant amino acid position was obtained. Finally, amino acid residues on the identified significant positions were tested using the Wald test, with the significance threshold set at $0.05/n$, where n is the total number of amino acid residues across all identified positions. The linkage between associated HLA amino acids and those in neighboring *HLA* genes was evaluated using $D' > 0.5$, calculated from PHASE results.

3.6. Acknowledgements

We express our gratitude to all the participants in BRIDGET, BQC19 and the Nagahama Study. BRIDGET was supported by the EU-JPND Funding (grant no. FKZ:01ED1615). BQC19 was made possible through open sharing of data and samples from the Biobanque québécoise de la COVID-19, funded by the Fonds de recherche du Québec - Santé, Génome Québec, the Public Health Agency of Canada and, as of March 2022, the ministère de la Santé et des Services sociaux. The Nagahama Study was funded by the Ministry of Education, Culture, Sports, Science & Technology (MEXT) in Japan, the Center of Innovation Program, and the Global University Project from Japan Science and Technology Agency, Japan Agency for Medical Research and Development (AMED)(grant no. JP16ek0109070), and a research grant from the Takeda Science Foundation. M.S. was supported by the Kyoto-McGill International Collaborative Program in Genomic Medicine and JST SPRING, Grant Number JPMJSP2110.

3.7. Author contributions

M.S., S.K., F.M., and G.B. designed the study. M.S., M.B. and P.M. executed the DNA-seq pipeline using GenPipes. M.S. and S.K. developed the methodology. M.S. implemented analysis scripts, including those for the developed methods, and conducted procedures across the entire pipeline. M.S., S.K. and G.B. wrote the paper. M.B., T.K., Y.I., E.G., R.E., S.D., R.Y., and Y.K. contributed to refining the analytical approach and interpreting the results. H.G., K.M., H.S., S.D. and F.M. collected and managed the samples and data construction.

3.8. References

1. Damania B, Kenney SC, Raab-Traub N. Epstein-Barr virus: Biology and clinical disease. *Cell*. 2022 Sep;185(20):3652–70.

2. Virgin HW, Wherry EJ, Ahmed R. Redefining Chronic Viral Infection. *Cell*. 2009 Jul 10;138(1):30–50.
3. Desfarges S, Ciuffi A. Viral Integration and Consequences on Host Gene Expression. *Viruses Essent Agents Life*. 2012 Sep 25;147–75.
4. Houldcroft CJ, Kellam P. Host genetics of Epstein–Barr virus infection, latency and disease. *Rev Med Virol*. 2015 Mar;25(2):71–84.
5. Kwok AJ, Mentzer A, Knight JC. Host genetics and infectious disease: new tools, insights and translational opportunities. *Nat Rev Genet*. 2021 Mar;22(3):137–53.
6. Urayama KY, Jarrett RF, Hjalgrim H, Diepstra A, Kamatani Y, Chabrier A, et al. Genome-Wide Association Study of Classical Hodgkin Lymphoma and Epstein–Barr Virus Status–Defined Subgroups. *JNCI J Natl Cancer Inst*. 2012 Feb 8;104(3):240–53.
7. Penova M, Kawaguchi S, Yasunaga JI, Kawaguchi T, Sato T, Takahashi M, et al. Genome wide association study of HTLV-1-associated myelopathy/tropical spastic paraparesis in the Japanese population. *Proc Natl Acad Sci U S A*. 2021 Mar 16;118(11):e2004199118.
8. Watzinger F, Ebner K, Lion T. Detection and monitoring of virus infections by real-time PCR. *Real-Time Polym Chain React*. 2006 Apr 1;27(2):254–98.
9. Tae H, Karunasena E, Bavarva JH, McIver LJ, Garner HR. Large scale comparison of non-human sequences in human sequencing data. *Genomics*. 2014 Dec;104(6 0 0):453–8.
10. Moustafa A, Xie C, Kirkness E, Biggs W, Wong E, Turpaz Y, et al. The blood DNA virome in 8,000 humans. *PLoS Pathog*. 2017 Mar;13(3):e1006292.
11. Guo J, Huang X, Zhang C, Huang P, Li Y, Wen F, et al. The blood virome of 10,585 individuals from the ChinaMAP. *Cell Discov*. 2022 Oct;8(1):113.
12. Miceli-Richard C, Gestermann N, Amiel C, Sellam J, Ittah M, Pavy S, et al. Effect of

methotrexate and anti-TNF on Epstein-Barr virus T-cell response and viral load in patients with rheumatoid arthritis or spondylarthropathies. *Arthritis Res Ther*. 2009 May 26;11(3):R77.

13. Chrisman B, He C, Jung JY, Stockham N, Paskov K, Washington P, et al. The human “contaminome”: bacterial, viral, and computational contamination in whole genome sequences from 1000 families. *Sci Rep*. 2022 Jun;12(1):9863.

14. Bernal KDE, Whitehurst CB. Incidence of Epstein-Barr virus reactivation is elevated in COVID-19 patients. *Virus Res*. 2023 Jun 26;334:199157.

15. Forque L, Albert E, Gimenez E, Torres I, Carbonell N, Ferreres J, et al. Monitoring of Torque Teno virus DNAemia in critically ill COVID-19 patients: May it help to predict clinical outcomes? *J Clin Virol*. 2022 Mar;148:105082.

16. Timmerman AL, Commandeur L, Deijns M, Burggraaff MGJM, Lavell AHA, van der Straten K, et al. The Impact of First-Time SARS-CoV-2 Infection on Human Anelloviruses. *Viruses*. 2024 Jan;16(1):99.

17. Tremblay K, Rousseau S, Zawati MH, Auld D, Chassé M, Coderre D, et al. The Biobanque québécoise de la COVID-19 (BQC19)—A cohort to prospectively study the clinical and biological determinants of COVID-19 clinical trajectories. *PLOS ONE*. 2021 May 19;16(5):e0245031.

18. Setoh K, Matsuda F. Cohort Profile: The Nagahama Prospective Genome Cohort for Comprehensive Human Bioscience (The Nagahama Study). In: Yano M, Matsuda F, Sakuntabhai A, Hirota S, editors. *Socio-Life Science and the COVID-19 Outbreak: Public Health and Public Policy*. Singapore: Springer Singapore; 2022. p. 127–43.

19. Bourgey M, Dali R, Eveleigh R, Chen KC, Letourneau L, Fillon J, et al. GenPipes: an open-source framework for distributed and scalable genomic analyses. *GigaScience*. 2019 Jun 1;8(6):giz037.

20. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinforma*. 2013;43:11 10 1-11 10 33.
21. Koenig Z, Yohannes MT, Nkambule LL, Zhao X, Goodrich JK, Kim HA, et al. A harmonized public resource of deeply sequenced diverse human genomes. *Genome Res*. 2024 May 15;gr.278378.123.
22. Liu X, Kosugi S, Koide R, Kawamura Y, Ito J, Miura H, et al. Endogenization and excision of human herpesvirus 6 in human genomes. *PLoS Genet*. 2020 Aug;16(8):e1008915.
23. Lefranc MP, Duprat E, Kaas Q, Tranne M, Thiriout A, Lefranc G. IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. *Dev Comp Immunol*. 2005 Jan 1;29(11):917–38.
24. Krueger GRF, Koch B, Leyssens N, Berneman Z, Rojo J, Horwitz C, et al. Comparison of Seroprevalences of Human Herpesvirus-6 and -7 in Healthy Blood Donors from Nine Countries. *Vox Sang*. 1998;75(3):193–7.
25. Sasa N, Kojima S, Koide R, Hasegawa T, Namkoong H, Hirota T, et al. Blood DNA virome associates with autoimmune diseases and COVID-19. *Nat Genet*. 2025 Jan;57(1):65–79.
26. Letovsky SI, Cao X, Hollenbach JA, Mack SJ, Maier M. Association between HLA genetics and SARS-CoV-2 infection in a large real-world cohort. *Genes Immun*. 2025 Jun 1;26(3):213–21.
27. Scepanovic P, Alanio C, Hammer C, Hodel F, Bergstedt J, Patin E, et al. Human genetic variants and age are the strongest predictors of humoral immune responses to common pathogens and vaccines. *Genome Med*. 2018 Jul 27;10(1):59.

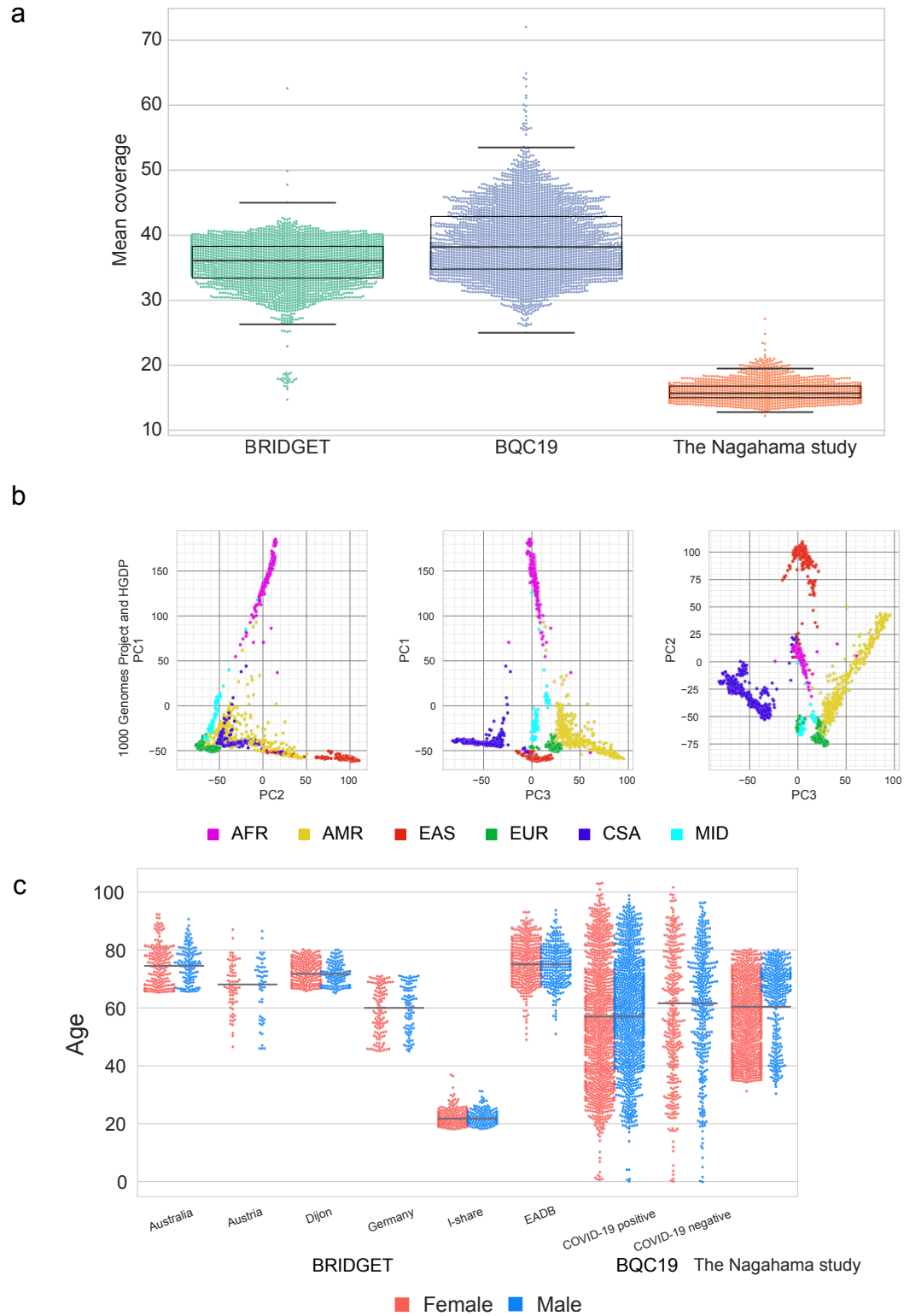
28. Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Schoch CL, Sherry ST, et al. GenBank. *Nucleic Acids Res.* 2020 Nov;49(D1):D92.
29. Yoo S, Garg E, Elliott L, Hung R, Halevy A, Brooks J, et al. HostSeq: a Canadian whole genome sequencing and clinical data resource. *BMC Genomic Data.* 2023 May 2;24(1):26.
30. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007 Sep;81(3):559–75.
31. Seabold S, Perktold J. Statsmodels: Econometric and Statistical Modeling with Python. In: Walt S van der, Millman J, editors. *Proceedings of the 9th Python in Science Conference.* 2010. p. 92–6.
32. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. *Nat Protoc.* 2010 Sep;5(9):1564–73.
33. Lu W, Gauthier LD, Poterba T, Giacomuzzi E, Goodrich JK, Stevens CR, et al. CHARR efficiently estimates contamination from DNA sequencing data. *Am J Hum Genet.* 2023 Dec 7;110(12):2068–76.
34. Mitsuhashi N, Toyo-oka L, Katayama T, Kawashima M, Kawashima S, Miyazaki K, et al. TogoVar: A comprehensive Japanese genetic variation database. *Hum Genome Var.* 2022 Dec 12;9(1):1–9.
35. Pedersen BS, Bhetariya PJ, Brown J, Kravitz SN, Marth G, Jensen RL, et al. Somalier: rapid relatedness estimation for cancer and germline studies using efficient genome sketches. *Genome Med.* 2020 Jul;12(1):62.

36. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014 Aug 1;30(15):2114–20.
37. Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schäffer AA. Database indexing for production MegaBLAST searches. *Bioinformatics*. 2008 Aug 15;24(16):1757–64.
38. Brister JR, Ako-Adjei D, Bao Y, Blinkova O. NCBI viral genomes resource. *Nucleic Acids Res*. 2015 Jan;43(Database issue):D571-577.
39. Luppi M, Marasca R, Barozzi P, Ferrari S, Ceccherini-Nelli L, Batoni G, et al. Three cases of human herpesvirus-6 latent infection: integration of viral genome in peripheral blood mononuclear cell DNA. *J Med Virol*. 1993 May;40(1):44–52.
40. Ameijeiras-Alonso J, Crujeiras RM, Rodriguez-Casal A. multimode: An R Package for Mode Assessment. *J Stat Softw*. 2021 Mar 22;97:1–32.
41. Stacy EW. A Generalization of the Gamma Distribution. *Ann Math Stat*. 1962 Sep;33(3):1187–92.
42. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020 Mar;17(3):261–72.
43. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009 Jun 1;25(11):1422–3.
44. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 2013 Apr;30(4):772–80.
45. McLaren CE, Legler JM, Brittenham GM. The Generalized χ^2 Goodness-of-fit

Test. *J R Stat Soc Ser Stat.* 1994;43(2):247–58.

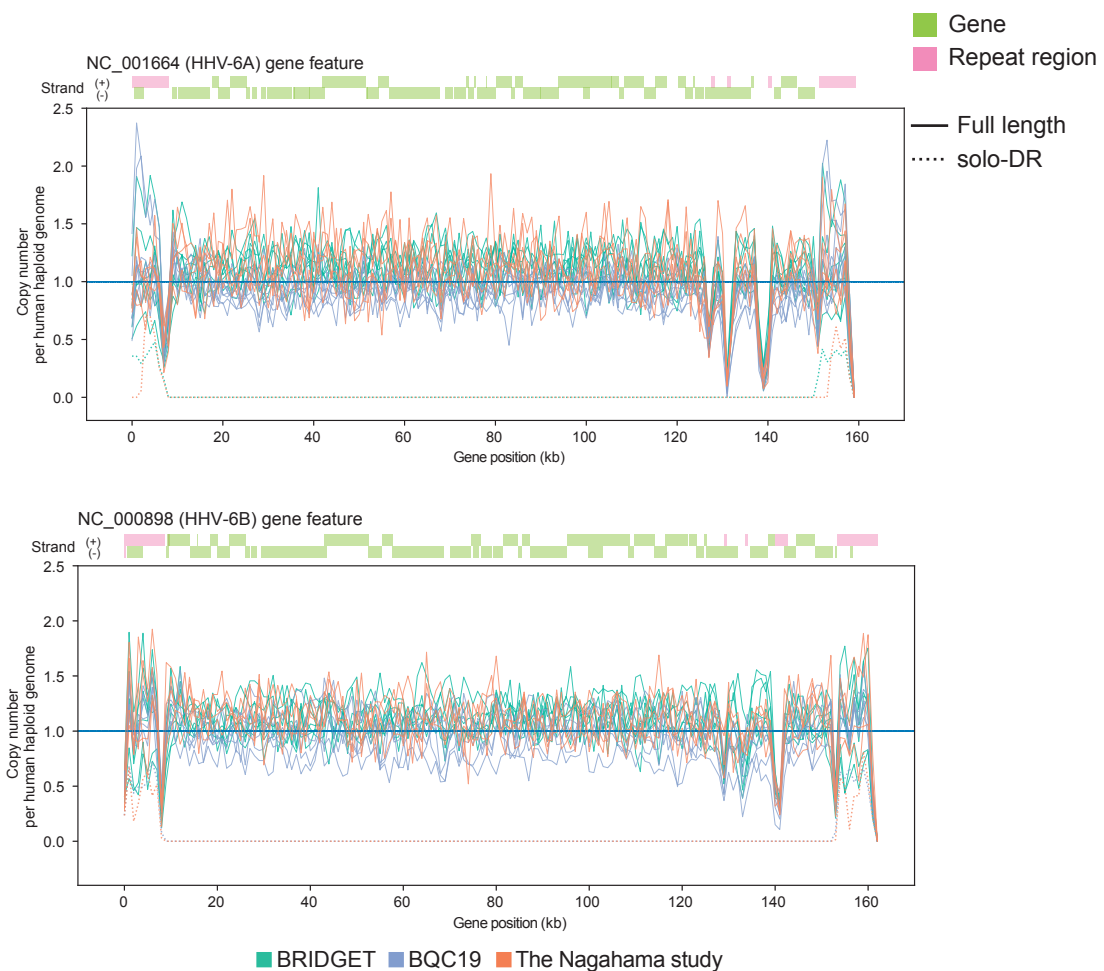
46. V. Satopaa, J. Albrecht, D. Irwin, B. Raghavan. Finding a “Kneedle” in a Haystack: Detecting Knee Points in System Behavior. In: 2011 31st International Conference on Distributed Computing Systems Workshops. 2011. p. 166–71.
47. Hammond RK, Pahl MC, Su C, Cousminer DL, Leonard ME, Lu S, et al. Biological constraints on GWAS SNPs at suggestive significance thresholds reveal additional BMI loci. Zaidi M, Meyre D, editors. *eLife*. 2021 Jan 18;10:e62206.
48. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*. 2010 Sep;26(18):2336–7.
49. Kawaguchi S, Higasa K, Shimizu M, Yamada R, Matsuda F. HLA-HD: An accurate HLA typing algorithm for next-generation sequencing data. *Hum Mutat*. 2017 Jul 1;38(7):788–97.
50. Robinson J, Barker DJ, Georgiou X, Cooper MA, Flicek P, Marsh SGE. IPD-IMGT/HLA Database. *Nucleic Acids Res*. 2020 Jan;48(D1):D948–55.
51. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012 Apr;9(4):357–9.
52. Stephens M, Donnelly P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet*. 2003 Nov;73(5):1162–9.
53. Thomson G, Baur MP. Third order linkage disequilibrium. *Tissue Antigens*. 1984 Oct 1;24(4):250–5.

3.9. Supplementary Figures



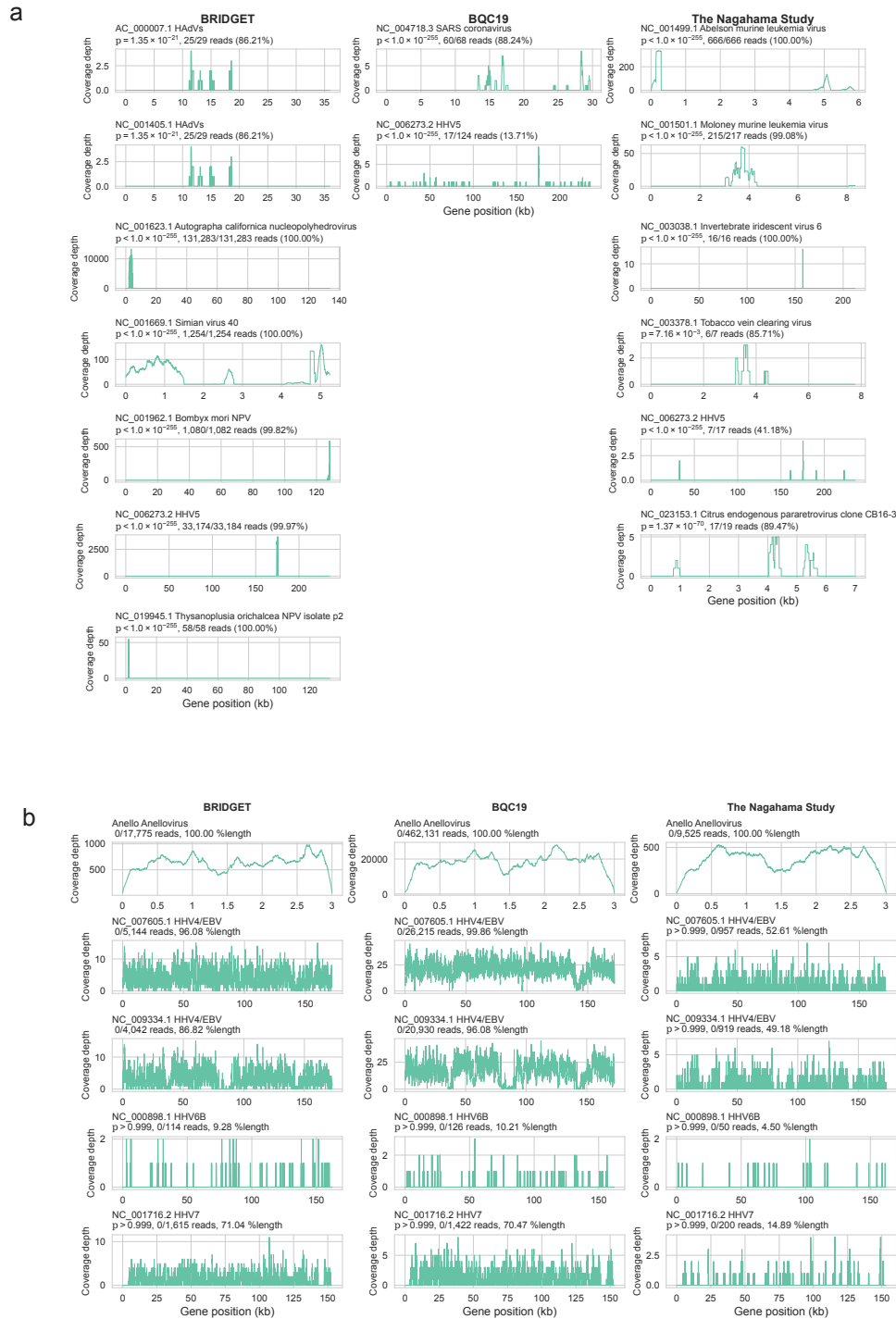
Supplementary Fig. 1 WGS results and demographics across diverse cohorts.

a, Distribution of the mean coverage of the human genome across the three cohorts. b, Principal Components Analysis (PCA) plots of the training data. The dots in the scatter plots of PC1-2, PC1-3, and PC2-3 show the individuals in the training data, coloured by the six populations: AFR (African), AMR (admixed American), CSA (Central/South Asian), EAS (East Asian), EUR (European) and MID (Middle Eastern). c, Age distribution across subgroups of the three cohorts. Each dot represents an individual coloured according to sex.



*Supplementary Fig. 2 Average coverage of *iciHHV-6A/B* viral reads*

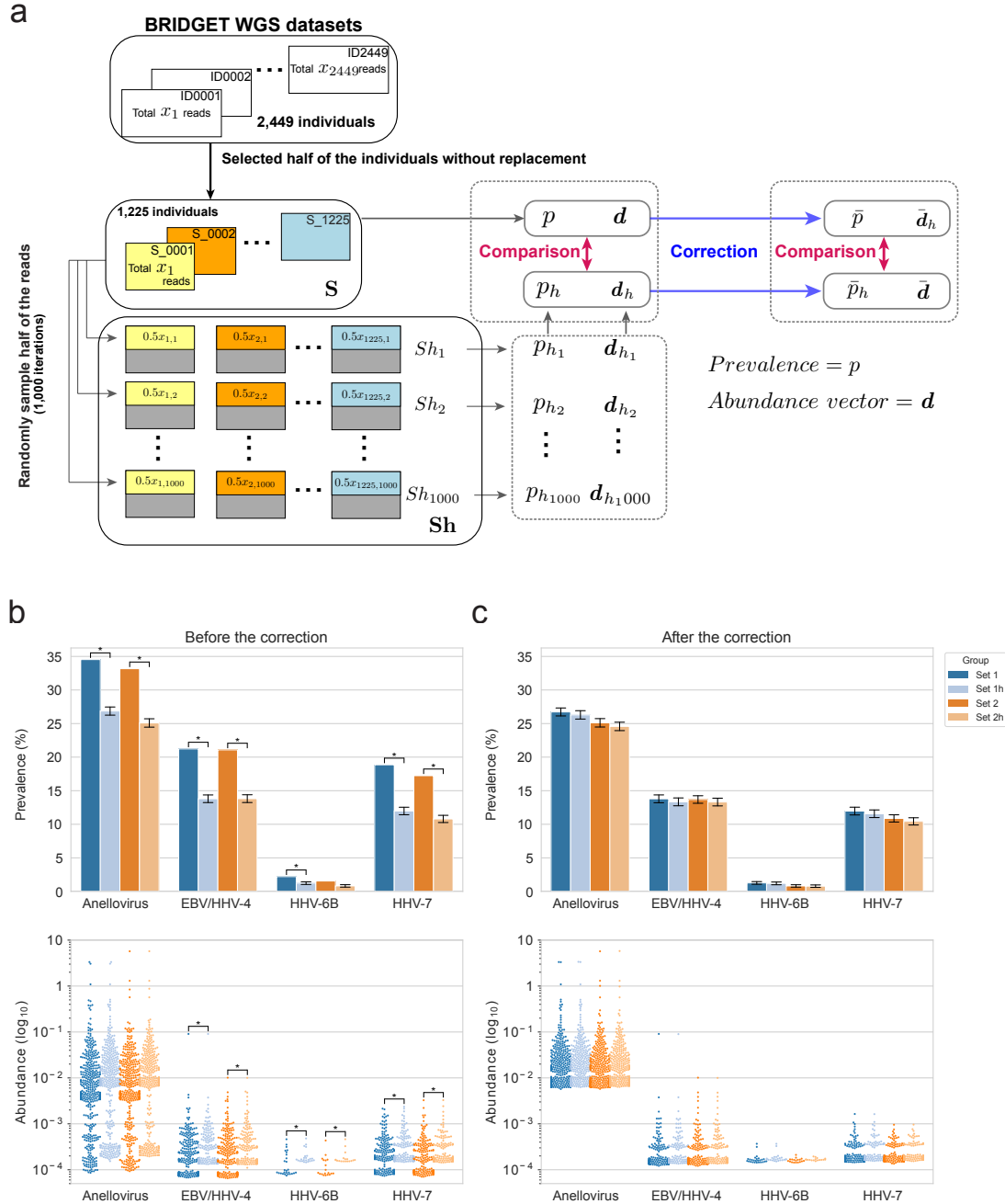
The coverage plot normalized by the mean coverage of human haploid genome are shown. Each curve represents individual coverage. The line style illustrates the full and solo-DR integrations, respectively. The top rectangle is coloured by the genes and repeat regions defined in Genbank.



Supplementary Fig. 3 Detected artificial reads and coverage depth of mapped reads

a, Coverage depth across all individuals in the cohort for viruses in which artificial reads were detected. Percentages represent the proportion of artificial reads among all reads mapped to each viral sequence. b, Coverage depth across all individuals for viruses in which reads were

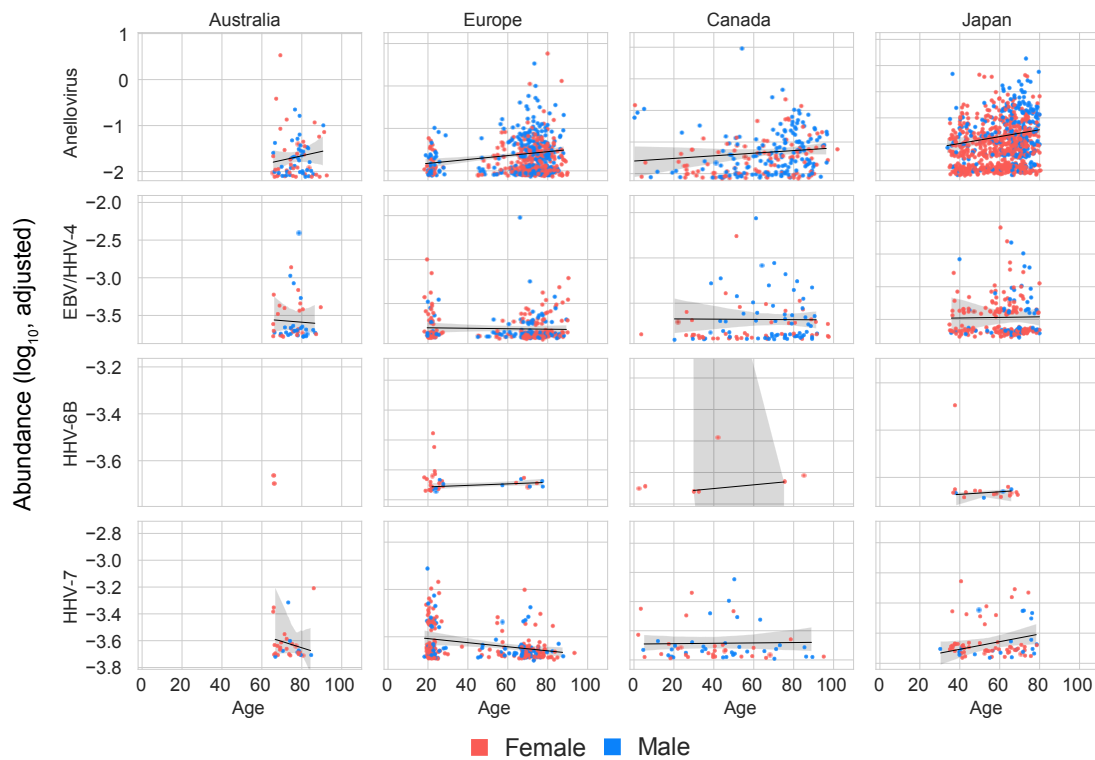
detected at a frequency of $\geq 1\%$ individuals in all cohorts. Percentages indicate the proportion of bases covered by at least one read, relative to the total length of each viral sequence.



Supplementary Fig. 4 Evaluation of the correction method for viral quantification.

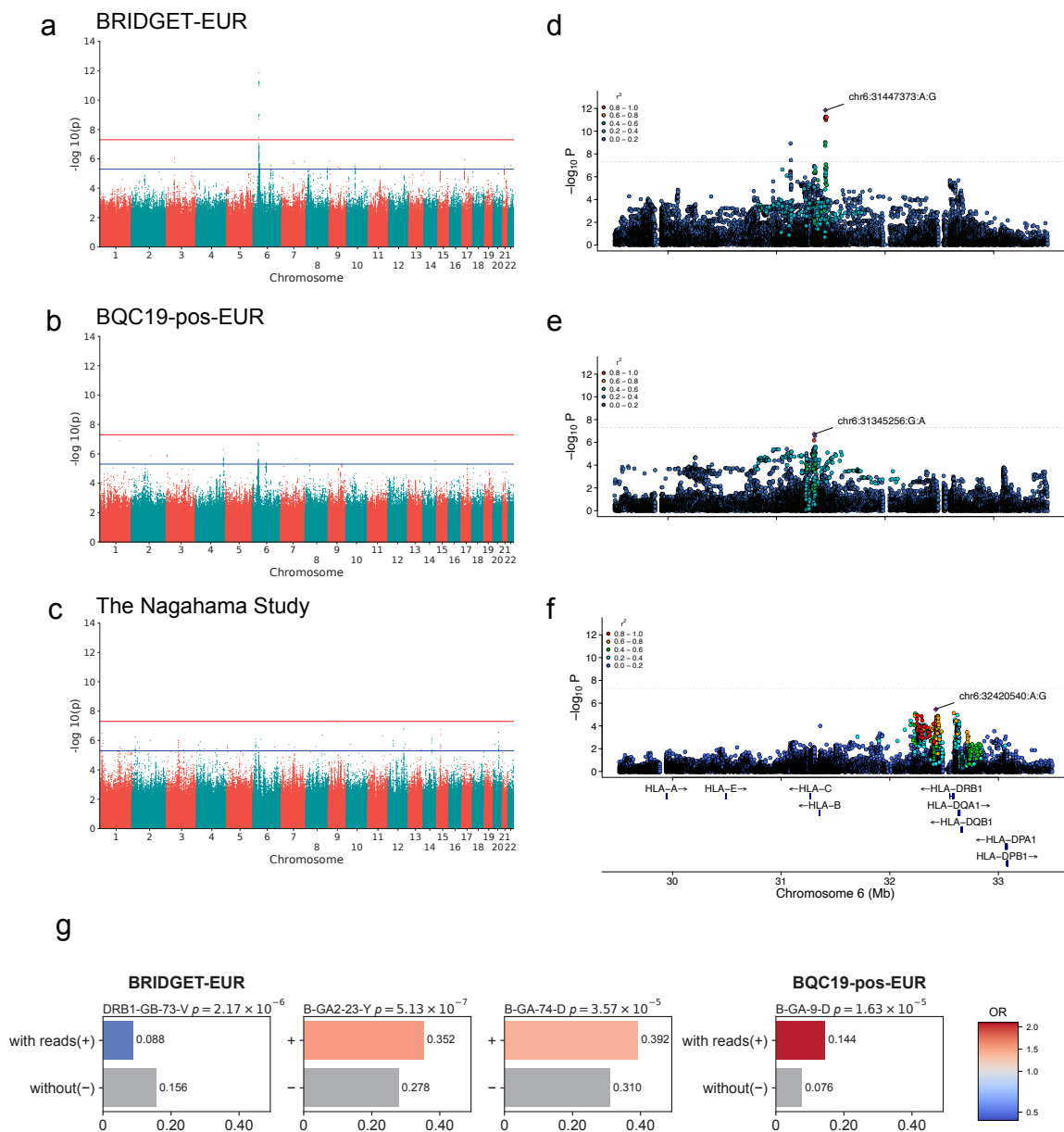
a, Workflow for simulation data generation using BRIDGET cohort and comparison of viral quantities before and after correction. b, c Comparisons of the prevalence and abundance of

detected viral reads between full and half datasets for before (b) and after (c) correction. Error bars indicate standard deviation. Statistically significant comparisons are marked with asterisks.



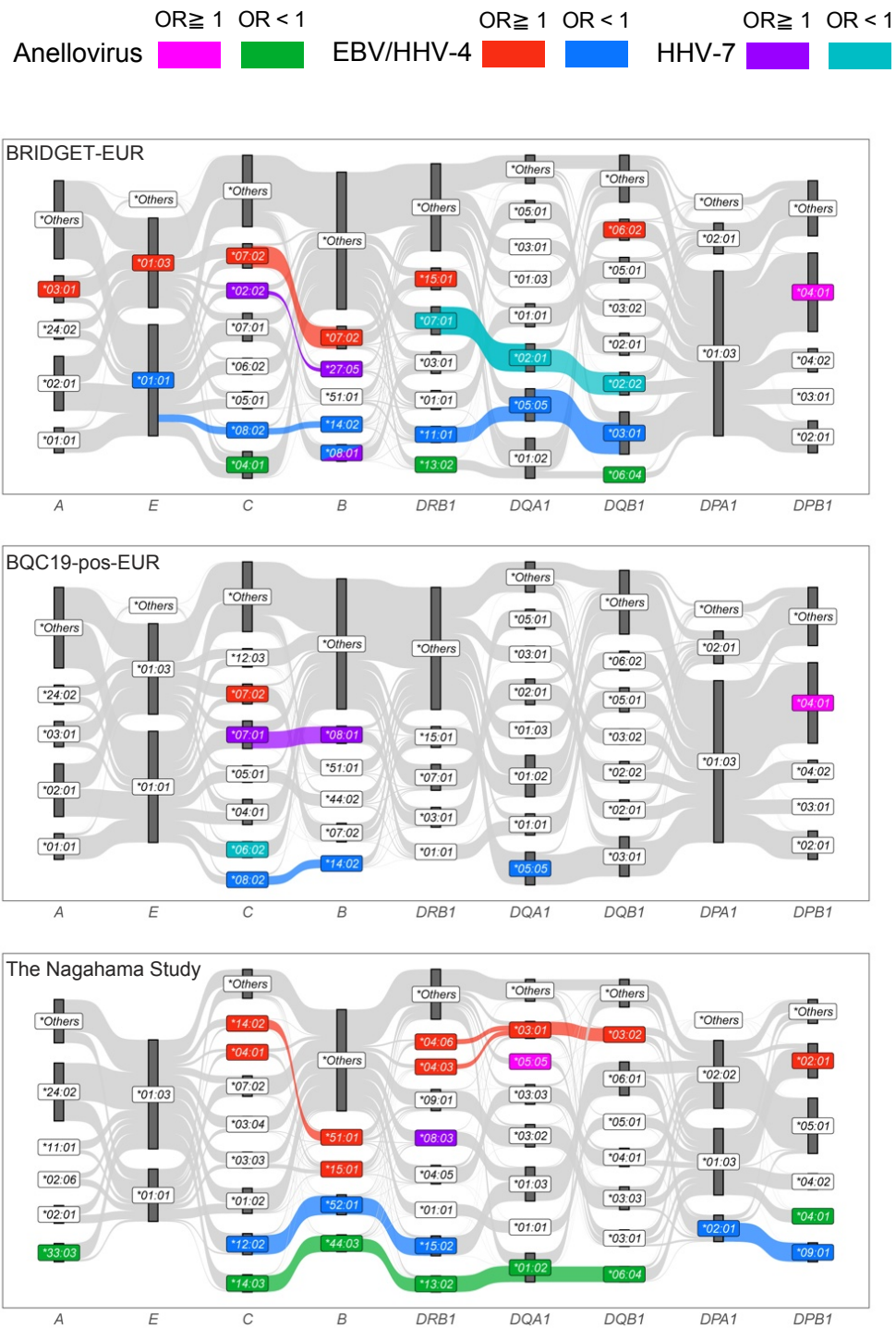
Supplementary Fig. 5 The correlation between age and abundance of detected viruses

The columns and rows are organized by virus and group, and the colour represents sex. The black line represents the regression line estimated from the combined data of both females and males.



Supplementary Fig. 6 Genetic associations of HHV-7 with host factors

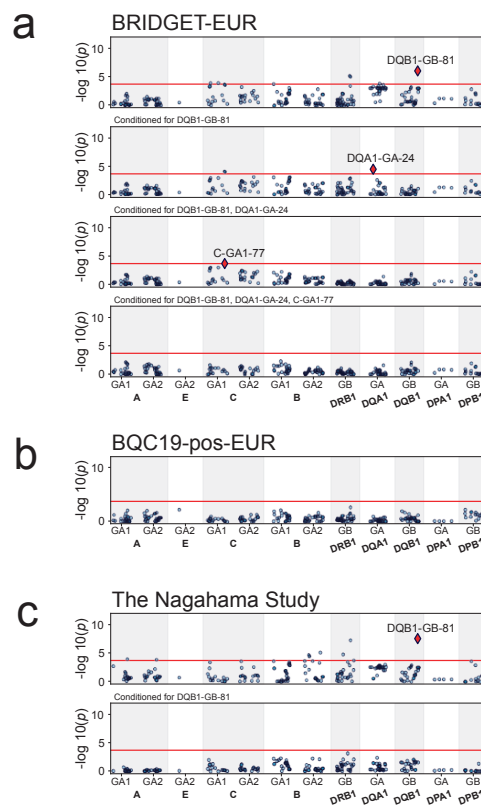
a-c, GWAS results for HHV-7 prevalence in BRIDGET-EUR (a), BQC19-pos-EUR (b) and the Nagahama Study (c). d-f, Associations between prevalence of EBV/HHV-4 and variants within the MHC region in BRIDGET-EUR (d), BQC19-pos-EUR (e), and the Nagahama Study (f). The variant shown in each panel represents the most significant association with viral prevalence. Positions are based on hg38. g, Frequencies (x-axis) of amino acid residues significantly associated with HHV-7 prevalence.



Supplementary Fig. 7 HLA haplotypes associated with viral prevalence

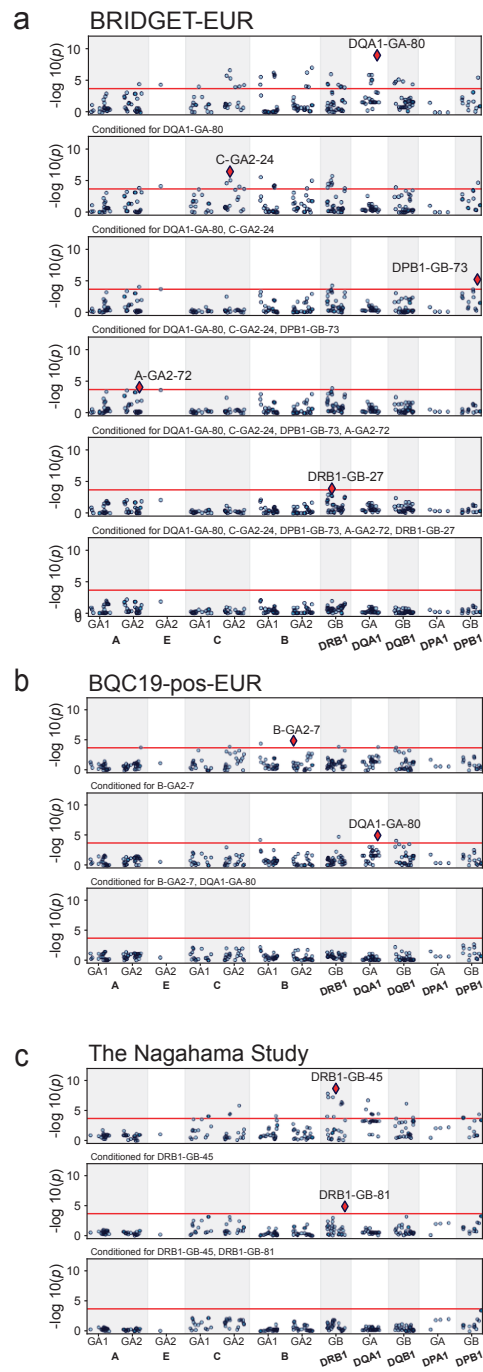
Linkage disequilibrium plot between *HLA* alleles (allele frequency > 0.07) for the individuals in BRIDGET-EUR, BQC19-pos-EUR and the Nagahama Study. Alleles are colour-coded

according to each virus, categorized by whether their odds ratio is ≥ 1 or < 1 . The strong linkage disequilibrium ($D' > 0.5$) within the risk or protective significant alleles in each virus is shown as the flow between the alleles.



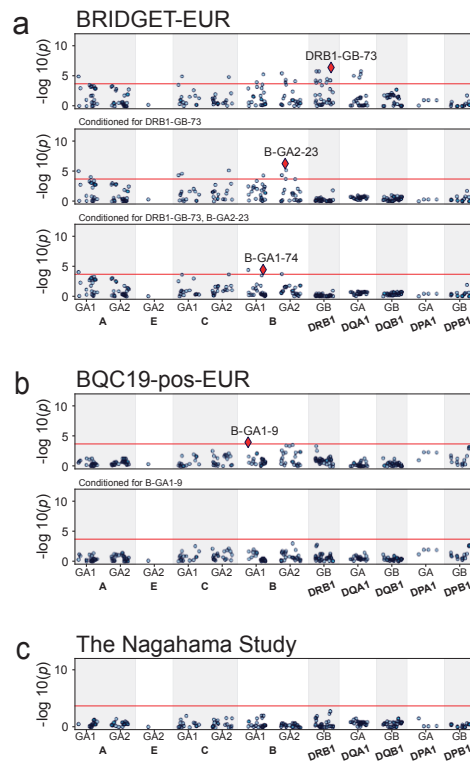
Supplementary Fig. 8 Association analysis of amino acid residues with Anellovirus prevalence

Association between each amino acid position and prevalence of Anellovirus in (a) BRIDGET-EUR, (b) BQC19-pos-EUR and (c) the Nagahama Study. The red dot illustrates the most significantly associated amino acid position.



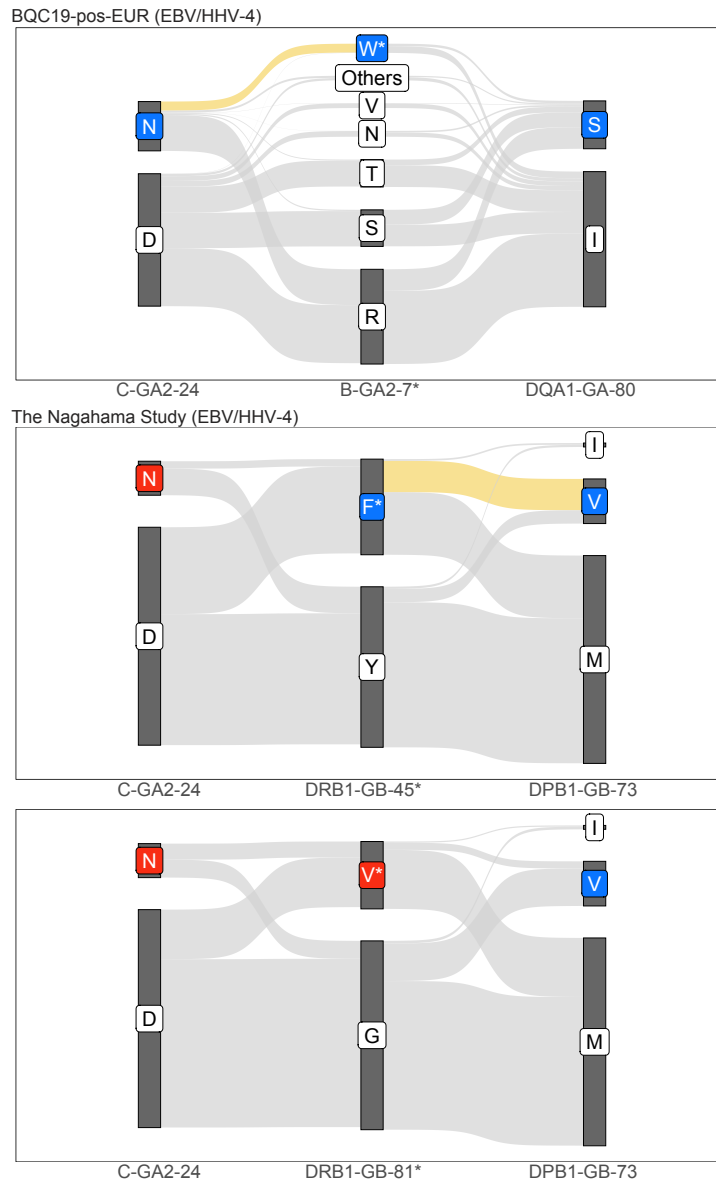
Supplementary Fig. 9 Association analysis of amino acid residues with EBV/HHV-4 prevalence

Association between each amino acid position and prevalence of EBV/HHV-4 in (a) BRIDGET-EUR, (b) BQC19-pos-EUR and (c) the Nagahama Study. The red dot illustrates the most significantly associated amino acid position.



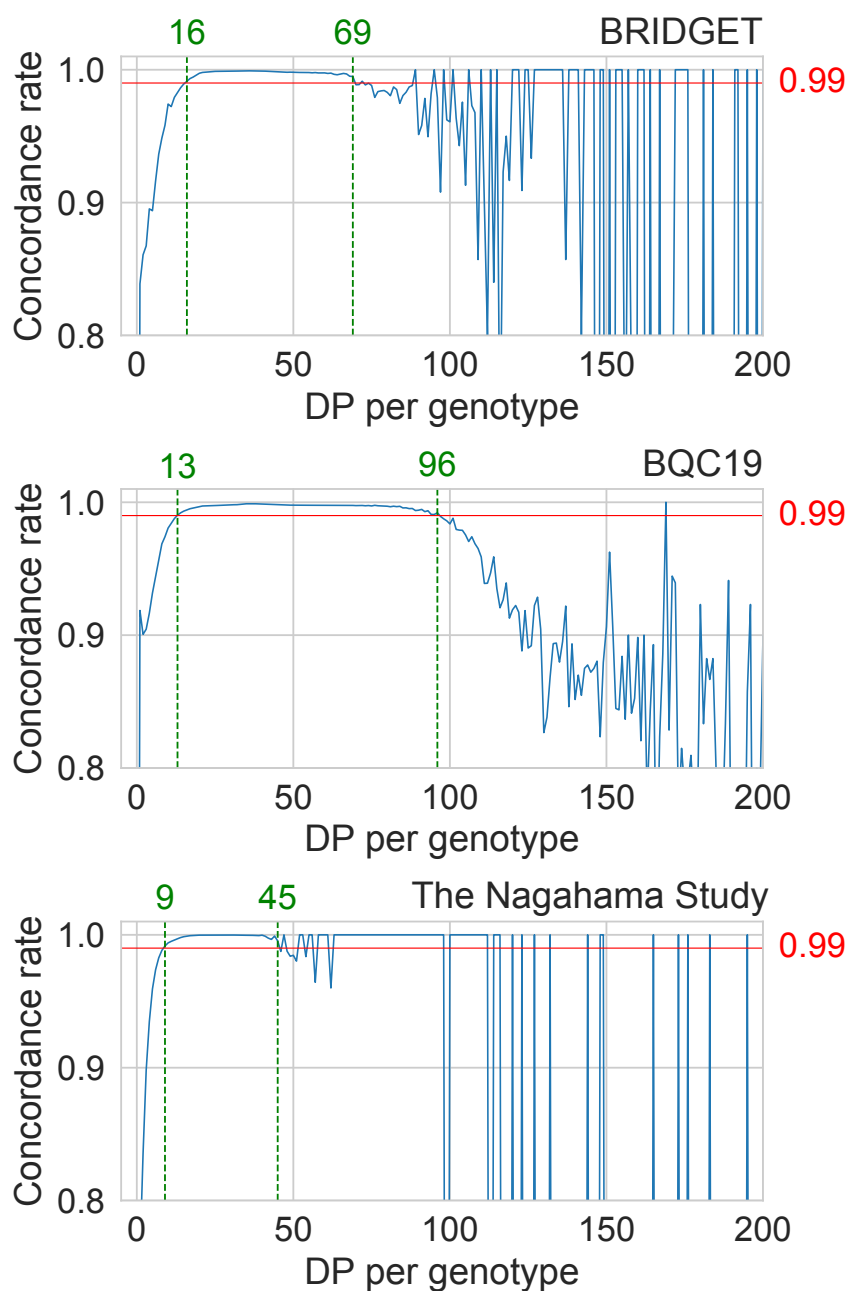
Supplementary Fig. 10 Association analysis of amino acid residues with HHV-7 prevalence

Association between each amino acid position and prevalence of HHV-7 in (a) BRIDGET-EUR, (b) BQC19-pos-EUR and (c) the Nagahama Study. The red dot illustrates the most significantly associated amino acid position.



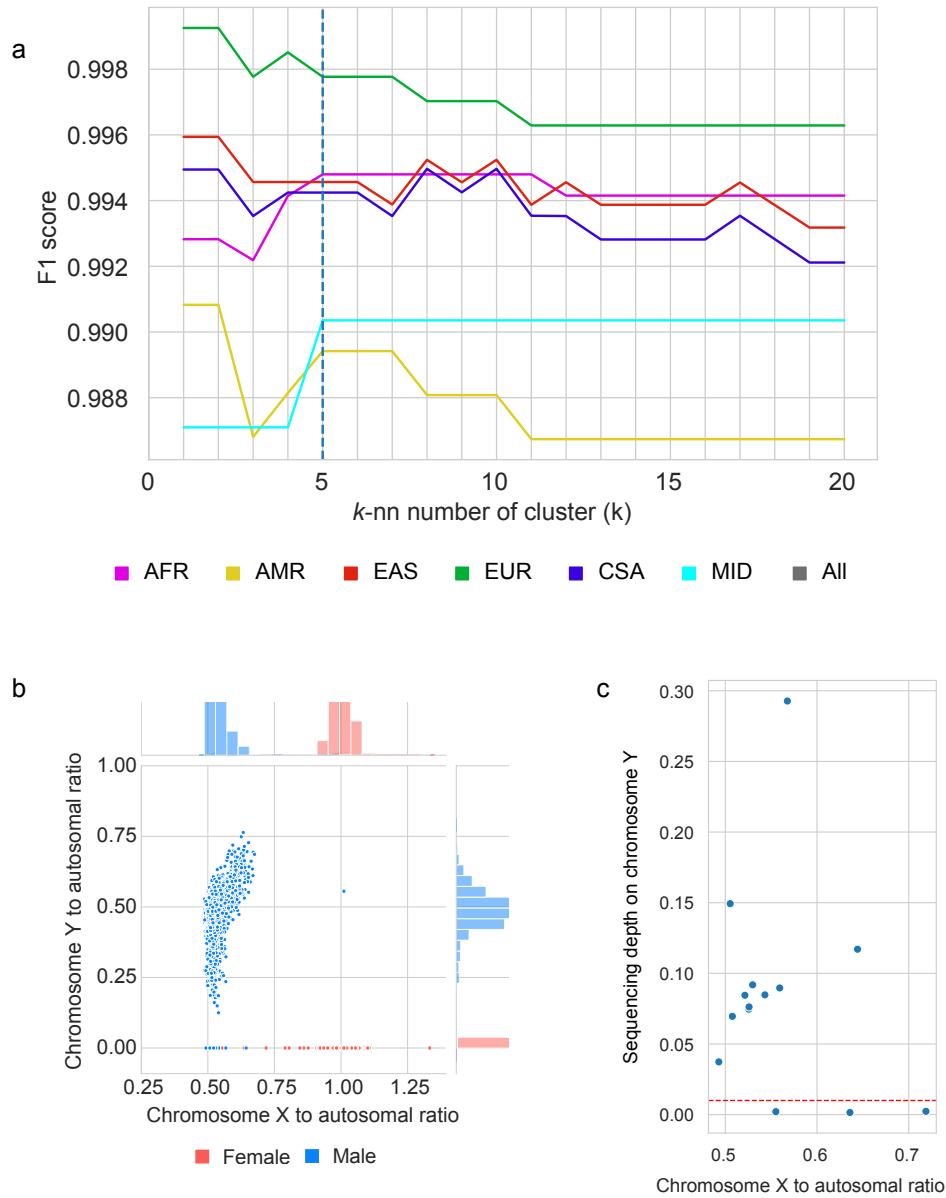
Supplementary Fig. 11 Linkage patterns of amino acid residues associated with the prevalence of EBV/HHV-4.

Strong links ($D' > 0.5$) among significant risk or protective amino acid residues, identified with or without conditional analysis, are visualized as flows highlighted between residues for each virus. Asterisks indicate the amino acids that were identified as significantly associated through conditional analysis.



Supplementary Fig. 12 The changes in genotype concordance rate based on DP per genotype in each cohort.

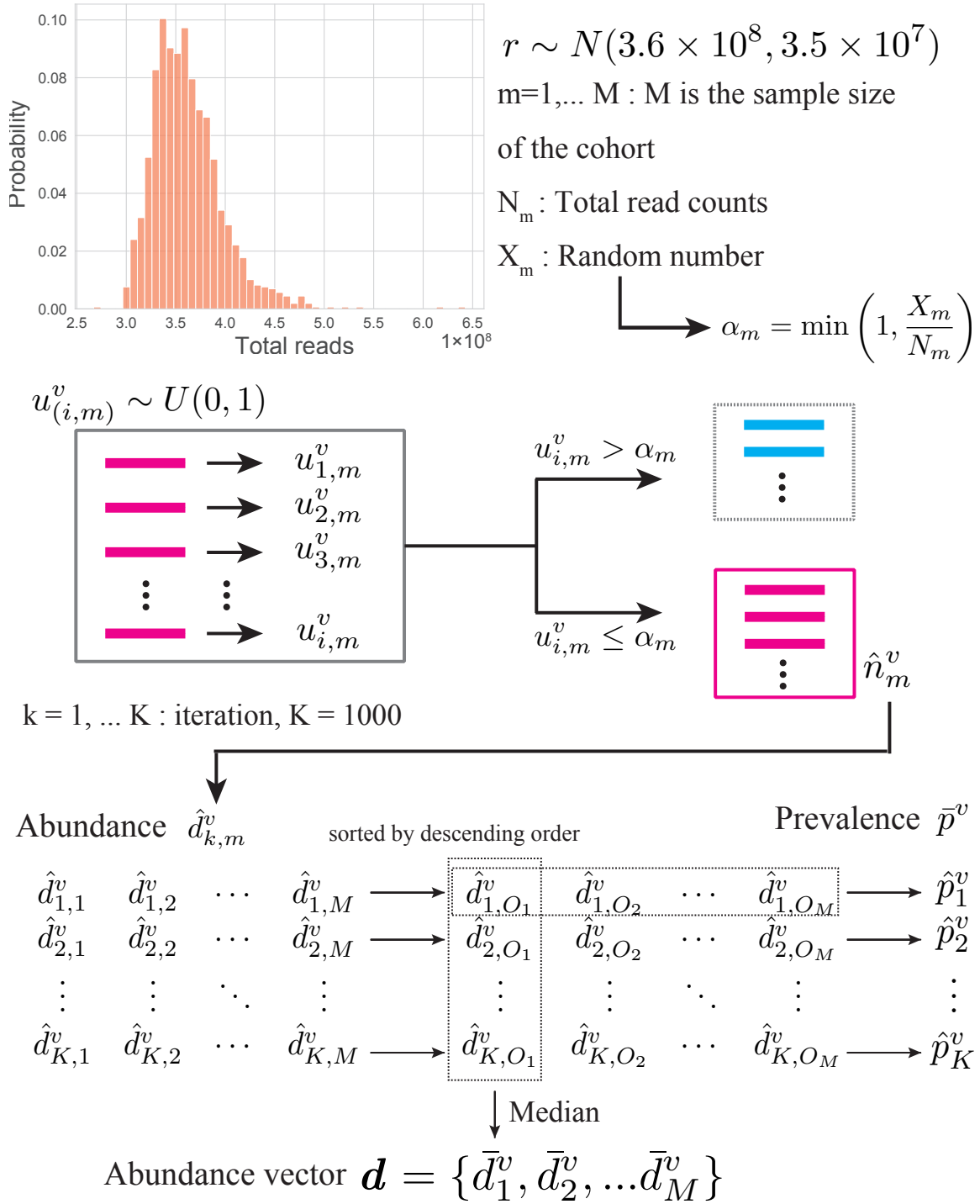
The blue line represents the genotype concordance rate at each DP for genotype. The horizontal red line marks a genotype concordance rate of 0.99, while the vertical green dashed lines indicate the thresholds defined in Supplementary Table 15.



Supplementary Fig. 13 Parameters for population and sex inference

a, The change in the F1 score with respect to the parameter k . The lines are coloured according to the six populations, and black represents the mean across all populations. The dashed vertical blue line indicates $k=5$. b, The scatter plot of the normalized coverage of the sex chromosomes (norm-X and norm-Y). Each dot represents an individual from the three cohorts. The x-axis corresponds to norm-X, while the y-axis corresponds to norm-Y. Each dot represents an individual coloured according to sex. c, The scatter plots of the normalized coverages of

chromosome Y(norm-Yb) from the unclassified individuals in the three cohorts. The dashed red line indicates 0.01.



Supplementary Fig. 14 The illustration of an unbiased comparison of viral prevalence and abundance.

4. General discussion

4.1. Addressing the missing heritability

In this thesis, I claim that there are two important issues unresolved in the traditional heritability: missing factors and the use of inappropriate models. Identifying missing factors directly improves the heritability and model optimization enhances the accuracy of heritability estimation by addressing not only missing but also overestimated factors. The human body is maintained through biological reactions driven by vast and complex interactions, which may underlie missing or misestimated heritability. To better understand these, Chapter 2 focused on epistatic interaction, a type of non-additive genetic effect involving two or more genes, as investigated in the rheumatoid arthritis study. Heritability also varies among populations due to the influence of both environmental and genetic factors. Even with improved models that account for population structure using SNP data, discrepancies in heritability estimates between different populations still remain [\[86\]](#), suggesting the presence of unresolved genetic and environmental factors. In contrast, environmental backgrounds vary widely and often reflect regional differences, with geographic location approximating lifetime exposures. However, its correlation with genetic background complicates efforts to disentangle its respective contributions to heritability and identify which environmental factors influence it and how they exert their effects. Chapter 3 examined quantification of DNA viruses in human blood cells—biological measures influenced by both geographical and genetic backgrounds—as indicators of individual health based on interactions between viral and host genetic factors.

The general discussion considers the challenge of missing heritability from two perspectives: (1) which factors to target, and (2) how to incorporate them into models. Building on these, I further discuss a study design for implementing these two issues and a comprehensive evaluation of the resulting findings.

4.2. Components overlooked by traditional heritability

Numerous genetic variants directly associated with diseases or health status have been successfully identified through association analyses; however, identifying interactions, such as epistatic interactions, remains a challenge. Some of the variants related to epistatic interactions may have been identified independently by the analysis focusing only on a single interaction based on the previous findings. However, they are often overlooked due to the low detection power and excluded from heritability estimates if their joint probability is low. Chapter 2 demonstrated the detection of RA-associated SNP combinations that showed epistatic interactions using LAMPLINK [38]. The replicated combinations in the validation cohort were classified as either intra-chromosomal or inter-chromosomal, depending on whether the SNPs were located on the same or different chromosomes, respectively. All intra-chromosomal combinations consisted of SNPs within the *HLA* region, most of these SNPs showed genome-wide significant associations ($p < 5 \times 10^{-8}$) with RA in the analysis of that SNP alone, indicating the strong linkage disequilibrium of *HLA* haplotypes. For inter-chromosomal combinations, at least one SNP was located in the *HLA* region, and notably, none of the non-*HLA* SNPs showed genome-wide significant associations with RA in the analysis of that SNP alone. The conditional analysis of inter-chromosomal SNP combinations poses a substantial challenge for GWAS, highlighting the importance of the use of detection methods specific to epistatic interactions.

Among individuals who share similar genetic backgrounds, environmental changes lead to corresponding changes in biological responses, thereby influencing the heritability linked to environmental factors. Among the environmental factors, viruses interact with host-derived molecules through immune responses, and these interactions can influence health outcomes [87]. The effects of interaction may therefore vary greatly depending on the specific genetic combinations of both the virus and the host. In Chapter 3, we identified viral sequences from

WGS reads that were considered to lack homology to human genetic sequences, thereby enabling the differentiation between human and non-human sequences and facilitating the assessment of environmental factors. We observed different patterns of virus–host associations within the same population, potentially due to differences in geographic location, even when targeting common viruses that were prevalent across all cohorts.

Although the detection results in the two studies represent an improvement over missing heritability issues, there is still room for further enhancement. Each study highlighted different considerations in detecting the missing factor. One factor was repeatedly detected during the exploration process, raising the question of how strictly the candidate list should be filtered. The other study limited the search to a narrower region of the viral genome by using a non-redundant database, highlighting the question of how broadly the search range should be expanded. This contrast raises the question of how to efficiently detect missing factors: in other words, what should be omitted and what should be retained. In the case of RA in Chapter 2, the association with the *HLA* region is already well-established [47,49,88]. Due to the high-LD in this region, a significant association at any locus often leads to related *HLA* alleles appearing repeatedly as candidate combinations during combinatorial exploration. Furthermore, SNPs in the *HLA* region were annotated using biallelic markers (e.g., rsSNPs), which cannot fully capture the complexity of *HLA* alleles, such as two-field resolution. Incorporating more detailed *HLA* allele information into the input data—via whole-genome sequencing or imputation-based approaches—would allow biologically plausible allele combinations to be considered. This may reduce unnecessary detection redundancy and avoid overly conservative statistical corrections, thereby enabling more accurate and interpretable associations.

On the other hand, the viral genome annotation database used in Chapter 3 was non-redundant and may not have accounted for mutation sequences reflecting geographical differences among viruses. Although each virus was analyzed as a single representative genome, a search of the NCBI Virus Genome Database [89] (as of May 12, 2025) revealed that thousands of sequences

have been registered for 12,566 for EBV/HHV-4, 610 for HHV-6B, and 266 for HHV-7. This indicates substantial sequence variation (polymorphism) associated with each virus. Even if the scope is expanded to include polymorphism, the computational cost is unlikely to be significant, as the detected viral read counts were uniformly low across all individuals and cohorts.

4.3. Modelling strategies for incorporating overlooked components

Modelling is used to capture underlying associations; however, incorrect assumptions about the effect size estimated from associations between the phenotype and genetic model, or not accounting for biases caused by technical issues and confounding factors, can further obscure heritability. Regarding conditional analysis in GWAS using a linear regression model, it can only estimate at most the total effect size as the product of the single effect sizes. However, synergistic effects, shown as the epistatic interaction, are defined as interactions in which the combined effect size exceeds the product of the individual component effect sizes, which cannot be identified by standard linear regression. Consequently, model optimization is required for epistatic interactions, as each allele may follow a distinct genetic model. This was demonstrated in Chapter 2, where re-evaluating the interactions uncovered results that diverged from those obtained under the recessive exclusive genetic model assumed by LAMPLINK [\[38\]](#). One combination was optimized for epistatic interaction models, combining the recessive model of valine at position 11 of HLA-DRB1 (hereafter: HLA-DRB1-11-V) with the dominant model of rs9624, located on chromosome 14 at 2,150,012, annotated to *TPPP2* and *NDRG2*. These findings suggest that predefined models may be useful for initial screening, but model optimization should follow based on combinations that show significant associations. On the other hand, Chapter 3 demonstrated that the heritability of quantifying DNA viruses was influenced by detection limitation according to the total number of reads in WGS and

confounders, including age, sex, and population. First, correction is needed for data with very low read counts due to detection limits. This includes “sampling zeros,” where true values fall below detection limits, and “structural zeros,” where the virus is truly absent. Under these conditions, comparing the frequency of individuals with “non-zero” and “zero” viral reads, as well as the viral quantities, can be informative. This approach, known as a two-part model, is valid only when the data are unbiased and the sample size is sufficient [90]. The simulation datasets of half of the individuals in a cohort with a twofold difference in sequence coverage demonstrated that our correction effectively mitigated bias, enabling valid comparisons of viral quantification. Second, identifying relevant covariates, such as age, sex, and population, is essential before applying statistical tests, as it helps control confounding effects and ensures valid inference. If the association analysis is conducted without accounting for these confounders, false positives may be detected, which are not truly associated with the virus or the disease. The results of our study indicated a higher frequency of Anellovirus in males. For instance, if males are more frequently detected in cases than in controls, the detection of Anellovirus could be higher, leading to a significant difference, even if it is not actually related to the presence or absence of the disease of interest.

On the other hand, some aspects of the modelling remain unsatisfactory. The association between HLA-DRB1-11-V and rs9624 in Chapter 2 appears plausible based on previous findings. However, not all *HLA* alleles encoding HLA-DRB1-11-V were shown to be contributing. In Chapter 3, based on previous reports suggesting that viral infections have shaped present-day population genetics by favouring genetic backgrounds better adapted to specific regions [91], we categorized individuals according to established populations. Nevertheless, some findings failed to replicate between Canada and Europe, even within the same European (EUR) population group. This discrepancy may reflect differences between the European populations that migrated to Canada, which are not fully represented in public reference panels such as the 1000 Genomes Project and HGDP [71]. Furthermore, even within

the same EUR population, sampling bias or unrecognized population substructure could have influenced the results. This issue also applies when compared to the Japanese population. While the Japanese cohort was geographically restricted to the small region of Nagahama City, the Canadian EAS population likely includes individuals with diverse backgrounds distributed across a broad area, making direct comparisons potentially inequivalent. To improve the modelling of associations, one approach would be to define subpopulations in other cohorts by matching the variance of individuals in PCA space that characterizes the genetic population structure observed in the Nagahama Study.

4.4. Entangled effects of missing factors and modelling assumptions

The previous chapters have highlighted how missing factors and model optimization contribute to the problem of missing heritability independently. However, as the set of factors under investigation changes, the analytical model must be adapted accordingly. The repeated emergence of specific factors may distort results due to their overrepresentation, and confounders may contribute to false positives. To mitigate these risks, model optimization requires efficient selection of relevant factors, either by referencing prior evidence or by filtering the dataset during preprocessing to retain only essential information. When the set of factors under investigation includes components that are highly interrelated—such as those in strong LD—or that are already known to be associated with the trait of interest, their joint probabilities tend to increase. This can result in a combinatorial expansion of plausible factor sets, increasing computational cost. Improved algorithms and high-performance computing can help; however, even if exhaustive testing of all possible combinations were feasible in principle, that would not resolve the multiple-testing correction problem. Moreover, staged pruning approaches (e.g., LAMPLINK) do not necessarily gain from parallelization and will need further methodological work. Consequently, this necessitates the application of more stringent

significance thresholds, which may, in turn, influence model specification and interpretation. Potential confounding effects, such as between viral infection and host genetic factors, can be evaluated using the own dataset. Based on these findings, such confounding should be incorporated into comparative analyses. Assuming redundant associations are resolved, newly emerging relevant factors necessitate an update to the model or analytical strategy. The identification of factors and the modelling process are not independent tasks; rather, they should be iteratively refined based on the outcomes derived from each other.

4.5. Design of the study incorporating the missing factors and models

In this section, I describe how to design studies that integrate two key aspects: identifying missing factors and constructing models capable of incorporating them into association analyses, along with strategies for their validation. As noted in the Introduction, overcoming missing heritability is the central theme of this thesis, and the present results indicate the existence of missing factors that should be detected, characterized, and integrated into heritability models and validated in independent, well-phenotyped cohorts. Study design should account for population structure to avoid bias, since findings may reflect either common biological pathways shared across humans or variation due to population- or environment-specific factors. Chapter 2 used two Japanese cohorts, in which subgroups may exist; Chapter 3 included participants from multiple populations. How finely samples should be stratified—within or across populations—remains undecided. Although Chapter 2 did not address environmental contributors in detail, shared exposures within a single population may nonetheless have been missed. Moreover, because Chapter 3 considered only environmental factors inferable from genetic data, additional environmental contributors to heritability may remain unobserved. To identify such contributors and improve integrative models, it will be valuable to analyze cohorts that capture a wide range of exposures; cohorts with limited

regional variation or those that are population-based rather than disease-selected may be particularly informative. Large resources such as the Nagahama Study and the UK Biobank are especially well-suited for these purposes.

During this process, known associations at single genetic loci, such as those in the *HLA* region or influenced by LD structure, can be incorporated into the input data prior to the main analysis. For example, representative factors, for instance, a known *HLA* allele associated with a specific trait, can be extracted, while other *HLA* alleles may be excluded from the input data to reduce redundancy and enhance interpretability. A core strategy should involve an integrative and iterative process, where the search for missing factors and the refinement of the model mutually inform each other in a dynamic, data-driven manner.

In the validation step, alongside genetic background, environmental factors, such as viral infections, are influenced by geographical variation and are further intertwined with host immune responses. By comparing cohorts that share similar genetic backgrounds but differ in geographic locations, we can assess whether differences in viral infection prevalence arise from environmental factors, host genetics, or their interaction. In the context of HLA association analysis for quantification of viruses, the failure of certain *HLA* alleles to replicate associations within genetically similar populations suggests that immune responses, potentially shaped by viral exposure, may have shifted due to historical human migration across regions. Conversely, replicating associations with particular *HLA* alleles despite geographic differences in viral prevalence suggests that these viruses' distributions may reflect underlying host genetic structures, including regional allele frequency patterns. These bidirectional relationships between viral prevalence and host genetics underscore the importance of using cohorts recruited from diverse geographical regions, as was previously discussed [\[92\]](#). The availability of comparisons across multiple geographic regions enabled the evaluation of such hypotheses. Based on this refined set of candidate factors, follow-up wet-lab experiments could be conducted to evaluate the associations. For instance, viruses potentially linked to specific *HLA*

alleles could be quantified using PCR. These quantitative measurements would enable us to assess whether the observed genetic associations are driven not only by the presence or absence of viral DNA, but also by the levels at which these DNA viruses can be detected by the measurement, such as qPCR.

5. Conclusions and Future Directions

In this thesis, two studies were conducted to address the issue of missing heritability by focusing on two key perspectives: missing factors and modelling. These issues arise when relying solely on information obtained from WGS and genotype data. In Chapter 2, significant variant combinations associated with RA were identified while avoiding a combinatorial explosion. These combinations were subsequently validated in an independent cohort, and an optimized model was established. These findings captured epistatic effects contributing to heritability, specifically reflecting interactions between genetic variants that are typically difficult to detect using conventional approaches. Chapter 3 extended the analysis to environmental components of heritability, using viral read counts derived from WGS data to identify contributions from viral prevalence and associated genetic factors. Moreover, the observed interactions between environmental exposures and host genetic backgrounds suggested an additional, previously underrepresented source of heritability. Together, these findings highlight the value of considering both genetic interactions and gene–environment interplay to explain heritability more comprehensively. Along with the findings of each study, I discussed the importance of exhaustively pruning a wide range of factors in the context of missing factors, as well as the potential for obtaining results through model re-evaluation in the context of modelling. These efforts are not limited to discovering novel disease-associated factors but may also enhance the predictive accuracy of genetic contributions. For example, polygenic risk scores (PRS) often show reduced accuracy when applied across different populations [\[93\]](#), partly due to differences in heritability. Previous studies have suggested that focusing on traits with higher heritability estimates can improve PRS performance [\[94\]](#). Therefore, refining heritability estimates has the potential to broadly benefit genetic association studies.

Notably, the analyses presented in this thesis were conducted using existing data from array-based genotyping and standard WGS cohort studies, without requiring additional measurements. This highlights the practicality and scalability of the proposed approaches. However, it remains critical to ensure that the approach does not overfit and that the findings reflect robust and generalizable biological insights rather than artifacts specific to the dataset used in our studies. To establish broader relevance, missing factors and their modelling should be validated across multiple independent studies. Given that many large-scale cohorts are already available or ongoing, this presents a timely opportunity to incorporate epistatic interactions and other overlooked components as covariates in association analyses. Such integration may uncover novel risk factors without requiring additional data collection, thereby enhancing the utility of existing resources.

6. Master Reference List

1. Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, et al. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature*. 1983 Nov;306(5940):234–8.
2. Risch N, Spiker D, Lotspeich L, Nouri N, Hinds D, Hallmayer J, et al. A genomic screen of autism: evidence for a multilocus etiology. *Am J Hum Genet*. 1999 Aug;65(2):493.
3. Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, et al. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat Genet*. 2002 Dec;32(4):650–4.
4. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, et al. Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science*. 2005 Apr 15;308(5720):385–9.
5. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007 June;447(7145):661–78.
6. Osman AA. The Importance of Sugar in the Diet of the School Child. *Proc R Soc Med*. 1931 June;24(8):1103.
7. Cutler SJ, Loveland DB. The Risk of Developing Lung Cancer and Its Relationship to Smoking. *JNCI J Natl Cancer Inst*. 1954 Aug 1;15(1):201–11.
8. Dyer AR, Stamler J, Paul O, Berkson DM, Lepper MH, McKean H, et al. Alcohol consumption, cardiovascular risk factors, and mortality in two Chicago epidemiologic studies. *Circulation*. 1977 Dec 1;56(6):1067–74.
9. Sharpe CR. A case-control study of alcohol consumption and drinking behaviour in

patients with acute gout. *Can Med Assoc J.* 1984 Sept;131(6):563.

10. Korbjuhn P, Anagnostopoulos I, Hummel M, Tiemann M, Dallenbach F, Parwaresch MR, et al. Frequent Latent Epstein-Barr Virus Infection of Neoplastic T Cells and Bystander B Cells in Human Immunodeficiency Virus-Negative European Peripheral Pleomorphic T-Cell Lymphomas. *Blood.* 1993 July 1;82(1):217–23.

11. Mosca JD, Bednarik DP, Raj NBK, Rosen CA, Sodroski JG, Haseltine WA, et al. Herpes simplex virus type-1 can reactivate transcription of latent human immunodeficiency virus. *Nature.* 1987 Jan;325(6099):67–70.

12. Fisher RA. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Earth Environ Sci Trans R Soc Edinb.* 1919 Jan;52(2):399–433.

13. Jensen AR. Estimation of the limits of heritability of traits by comparison of monozygotic and dizygotic twins. *Proc Natl Acad Sci U S A.* 1967 July;58(1):149–56.

14. FALCONER DS. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann Hum Genet.* 1965 Aug 1;29(1):51–76.

15. Scarr S. Environmental bias in twin studies. *Eugen Q.* 1968 Mar 1;15(1):34–40.

16. Macgregor S, Cornes BK, Martin NG, Visscher PM. Bias, precision and heritability of self-reported and clinically measured height in Australian twins. *Hum Genet.* 2006 Nov 1;120(4):571–80.

17. Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, Mangino M, et al. Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet.* 2008 May;40(5):575–83.

18. Gudbjartsson DF, Walters GB, Thorleifsson G, Stefansson H, Halldorsson BV, Zusmanovich P, et al. Many sequence variants affecting diversity of adult human height. *Nat*

Genet. 2008 May;40(5):609–15.

19. Lettre G, Jackson AU, Gieger C, Schumacher FR, Berndt SI, Sanna S, et al. Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat Genet.* 2008 May;40(5):584–91.

20. Maher B. Personal genomes: The case of the missing heritability. *Nature.* 2008 Nov;456(7218):18–21.

21. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. *Nature.* 2009 Oct;461(7265):747–53.

22. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet.* 2010 July;42(7):565–9.

23. Kaprio J. Twins and the mystery of missing heritability: the contribution of gene-environment interactions. *J Intern Med.* 2012 Nov;272(5):440–8.

24. Tenesa A, Haley CS. The heritability of human disease: estimation, uses and abuses. *Nat Rev Genet.* 2013 Feb;14(2):139–49.

25. Xu K, Li B, McGinnis KA, Vickers-Smith R, Dao C, Sun N, et al. Genome-wide association study of smoking trajectory and meta-analysis of smoking status in 842,000 individuals. *Nat Commun.* 2020 Oct 20;11(1):5302.

26. Nishida N, Ohashi J, Khor SS, Sugiyama M, Tsuchiura T, Sawai H, et al. Understanding of HLA-conferred susceptibility to chronic hepatitis B infection requires HLA genotyping-based association analysis. *Sci Rep.* 2016 Apr;6:24767.

27. Penova M, Kawaguchi S, Yasunaga JI, Kawaguchi T, Sato T, Takahashi M, et al. Genome wide association study of HTLV-1-associated myelopathy/tropical spastic paraparesis

in the Japanese population. *Proc Natl Acad Sci U S A*. 2021 Mar 16;118(11):e2004199118.

28. Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. *BMJ*. 1995 Jan 21;310(6973):170.

29. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al. The Structure of Haplotype Blocks in the Human Genome. *Science*. 2002 June 21;296(5576):2225–9.

30. Li Y, Cho H, Wang F, Canela-Xandri O, Luo C, Rawlik K, et al. Statistical and Functional Studies Identify Epistasis of Cardiovascular Risk Genomic Variants From Genome-Wide Association Studies. *J Am Heart Assoc*. 2020 Apr 9;9(7):e014146.

31. Balvert M, Cooper-Knock J, Stamp J, Byrne RP, Mourragui S, van Gils J, et al. Considerations in the search for epistasis. *Genome Biol*. 2024 Nov 19;25:296.

32. Akaike H. Factor Analysis and AIC. In: Parzen E, Tanabe K, Kitagawa G, editors. *Selected Papers of Hirotugu Akaike* [Internet]. New York, NY: Springer; 1998. p. 371–86.

33. Ma L, Brautbar A, Boerwinkle E, Sing CF, Clark AG, Keinan A. Knowledge-Driven Analysis Identifies a Gene–Gene Interaction Affecting High-Density Lipoprotein Cholesterol Levels in Multi-Ethnic Populations. *PLOS Genet*. 2012 May 24;8(5):e1002714.

34. Mitra I, Lavillaureix A, Yeh E, Traglia M, Tsang K, Bearden CE, et al. Reverse Pathway Genetic Approach Identifies Epistasis in Autism Spectrum Disorders. *PLoS Genet*. 2017 Jan;13(1):e1006516.

35. Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NLS, et al. BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-wide Case-Control Studies. *Am J Hum Genet*. 2010 Sept 10;87(3):325–40.

36. Cowman T, Koyutürk M. Prioritizing tests of epistasis through hierarchical

representation of genomic redundancies. *Nucleic Acids Res.* 2017 June;45(14):e131.

37. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, et al. Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer. *Am J Hum Genet.* 2001 July 1;69(1):138–47.

38. Terada A, Yamada R, Tsuda K, Sese J. LAMPLINK: detection of statistically significant SNP combinations from GWAS data. *Bioinformatics.* 2016 Nov;32(22):3513.

39. Terada A, Okada-Hatakeyama M, Tsuda K, Sese J. Statistical significance of combinatorial regulations. *Proc Natl Acad Sci U A.* 2013 Aug;110(32):12996–3001.

40. Terada A, Tsuda K. Multiple Testing Tool to Detect Combinatorial Effects in Biology. *Methods Mol Biol Clifton NJ.* 2018;1807:83.

41. Oh S, Lee J, Kwon MS, Weir B, Ha K, Park T. A novel method to identify high order gene-gene interactions in genome-wide association studies: Gene-based MDR. *BMC Bioinformatics.* 2012 June 11;13(9):S5.

42. Almutairi K, Johannes Nossent, David Preen, Helen Keen, Charles Inderjeeth. 749The global prevalence of Rheumatoid Arthritis: A meta-analysis. *Int J Epidemiol.* 2021 Sept 1;50(Supplement_1):dyab168.034.

43. MacGregor AJ, Snieder H, Rigby AS, Koskenvuo M, Kaprio J, Aho K, et al. Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis Rheum.* 2000 Jan;43(1):30–7.

44. Terao C, Ikari K, Nakayamada S, Takahashi Y, Yamada R, Ohmura K, et al. A twin study of rheumatoid arthritis in the Japanese population. *Mod Rheumatol.* 2016 Sept;26(5):685–9.

45. van der Woude D, Houwing-Duistermaat JJ, Toes RE, Huizinga TW, Thomson W,

Worthington J, et al. Quantitative heritability of anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis. *Arthritis Rheum.* 2009 Apr;60(4):916–23.

46. Deighton CM, Walker DJ, Griffiths ID, Roberts DF. The contribution of H LA to rheumatoid arthritis. *Clin Genet.* 1989;36(3):178–82.

47. Raychaudhuri S, Sandor C, Stahl EA, Freudenberg J, Lee HS, Jia X, et al. Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat Genet.* 2012 Jan;44(3):291–6.

48. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature.* 2014 Feb;506(7488):376–81.

49. Okada Y, Eyre S, Suzuki A, Kochi Y, Yamamoto K. Genetics of rheumatoid arthritis: 2018 status. *Ann Rheum Dis.* 2019 Apr;78(4):446–53.

50. Diaz-Gallo LM, Ramskold D, Shchetynsky K, Folkersen L, Chemin K, Brynedal B, et al. Systematic approach demonstrates enrichment of multiple interactions between non-HLA risk variants and HLA-DRB1 risk alleles in rheumatoid arthritis. *Ann Rheum Dis.* 2018 Oct;77(10):1454–62.

51. Kallberg H, Padyukov L, Plenge RM, Ronnelid J, Gregersen PK, van der Helm-van Mil AH, et al. Gene-gene and gene-environment interactions involving HLA-DRB1, PTPN22, and smoking in two subsets of rheumatoid arthritis. *Am J Hum Genet.* 2007 May;80(5):867–75.

52. Genin E, Coustet B, Allanore Y, Ito I, Teruel M, Constantin A, et al. Epistatic interaction between BANK1 and BLK in rheumatoid arthritis: results from a large trans-ethnic meta-analysis. *PLoS One.* 2013;8(4):e61044.

53. Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, et al. Human genetics shape the gut microbiome. *Cell*. 2014 Nov 6;159(4):789–99.
54. Moreno-Gallego JL, Chou SP, Di Rienzi SC, Goodrich JK, Spector TD, Bell JT, et al. Virome Diversity Correlates with Intestinal Microbiome Diversity in Adult Monozygotic Twins. *Cell Host Microbe*. 2019 Feb 13;25(2):261-272.e5.
55. Spandole S, Cimponeriu D, Berca LM, Mihăescu G. Human anelloviruses: an update of molecular, epidemiological and clinical aspects. *Arch Virol*. 2015 Apr;160(4):893–908.
56. Seeley WW, Marty FM, Holmes TM, Upchurch K, Soiffer RJ, Antin JH, et al. Post-transplant acute limbic encephalitis. *Neurology*. 2007 July 10;69(2):156–65.
57. Epstein MA, Achong BG, Barr YM. VIRUS PARTICLES IN CULTURED LYMPHOBLASTS FROM BURKITT'S LYMPHOMA. *Lancet Lond Engl*. 1964 Mar;1(7335):702–3.
58. Erikson J, Ar-Rushdi A, Drwinga HL, Nowell PC, Croce CM. Transcriptional activation of the translocated c-myc oncogene in Burkitt lymphoma. *Proc Natl Acad Sci U S A*. 1983 Feb;80(3):820–4.
59. Minarovits J, Niller HH. Current Trends and Alternative Scenarios in EBV Research. *Methods Mol Biol Clifton NJ*. 2017;1532:1–32.
60. Xu M, Zhang WL, Zhu Q, Zhang S, Yao Y yuan, Xiang T, et al. Genome-wide profiling of Epstein-Barr virus integration by targeted sequencing in Epstein-Barr virus associated malignancies. *Theranostics*. 2019 Jan 30;9(4):1115–24.
61. Shafritz DA, Shouval D, Sherman HI, Hadziyannis SJ, Kew MC. Integration of hepatitis B virus DNA into the genome of liver cells in chronic liver disease and hepatocellular carcinoma. Studies in percutaneous liver biopsies and post-mortem tissue specimens. *N Engl J*

Med. 1981 Oct;305(18):1067–73.

62. Sasa N, Kojima S, Koide R, Hasegawa T, Namkoong H, Hirota T, et al. Blood DNA virome associates with autoimmune diseases and COVID-19. *Nat Genet.* 2025 Jan;57(1):65–79.

63. Leong HN, Tuke PW, Tedder RS, Khanom AB, Eglin RP, Atkinson CE, et al. The prevalence of chromosomally integrated human herpesvirus 6 genomes in the blood of UK blood donors. *J Med Virol.* 2007 Jan;79(1):45–51.

64. Miura H, Kawamura Y, Hattori F, Kozawa K, Ihira M, Ohye T, et al. Chromosomally integrated human herpesvirus 6 in the Japanese population. *J Med Virol.* 2018;90(10):1636–42.

65. Sanjuán Rafael, Nebot Miguel R., Chirico Nicola, Mansky Louis M., Belshaw Robert. Viral Mutation Rates. *J Virol.* 2010 Oct 1;84(19):9733–48.

66. Milholland B, Dong X, Zhang L, Hao X, Suh Y, Vijg J. Differences between germline and somatic mutation rates in humans and mice. *Nat Commun.* 2017 May 9;8(1):15183.

67. Li W, O'Neill KR, Haft DH, DiCuccio M, Chetvernin V, Badretdin A, et al. RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res.* 2020 Dec;49(D1):D1020.

68. Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Schoch CL, Sherry ST, et al. GenBank. *Nucleic Acids Res.* 2020 Nov;49(D1):D92.

69. Randolph HE, Fiege JK, Thielen BK, Mickelson CK, Shiratori M, Barroso-Batista J, et al. Genetic ancestry effects on the response to viral infection are pervasive but cell type specific. *Science.* 2021 Nov 26;374(6571):1127–33.

70. 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.

71. Koenig Z, Yohannes MT, Nkambule LL, Zhao X, Goodrich JK, Kim HA, et al. A harmonized public resource of deeply sequenced diverse human genomes. *Genome Res.* 2024 May 15;gr.278378.123.

72. Nikolich-Žugich J. The twilight of immunity: emerging concepts in aging of the immune system. *Nat Immunol.* 2018 Jan;19(1):10–9.

73. Klein SL, Flanagan KL. Sex differences in immune responses. *Nat Rev Immunol.* 2016 Oct;16(10):626–38.

74. Houldcroft CJ, Beale MA, Breuer J. Clinical and biological insights from viral genome sequencing. *Nat Rev Microbiol.* 2017 Mar;15(3):183–92.

75. Pan B, Ren L, Onuchic V, Guan M, Kusko R, Bruinsma S, et al. Assessing reproducibility of inherited variants detected with short-read whole genome sequencing. *Genome Biol.* 2022 Jan 3;23(1):2.

76. Tae H, Karunasena E, Bavarva JH, McIver LJ, Garner HR. Large scale comparison of non-human sequences in human sequencing data. *Genomics.* 2014 Dec;104(6 0 0):453–8.

77. Mukherjee S, Huntemann M, Ivanova N, Kyrpides NC, Pati A. Large-scale contamination of microbial isolate genomes by Illumina PhiX control. *Stand Genomic Sci.* 2015;10:18.

78. Chrisman B, He C, Jung JY, Stockham N, Paskov K, Washington P, et al. The human “contaminome”: bacterial, viral, and computational contamination in whole genome sequences from 1000 families. *Sci Rep.* 2022 June;12(1):9863.

79. Asplund M, Kjartansdóttir KR, Møllerup S, Vinner L, Fridholm H, Herrera JAR, et al.

Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries. *Clin Microbiol Infect.* 2019 Oct 1;25(10):1277–85.

80. Tan CCS, Ko KKK, Chen H, Liu J, Loh M, Chia M, et al. No evidence for a common blood microbiome based on a population study of 9,770 healthy humans. *Nat Microbiol.* 2023 May;8(5):973–85.

81. de Goffau MC, Lager S, Salter SJ, Wagner J, Kronbichler A, Charnock-Jones DS, et al. Recognizing the reagent microbiome. *Nat Microbiol.* 2018 Aug 1;3(8):851–3.

82. Guo J, Huang X, Zhang C, Huang P, Li Y, Wen F, et al. The blood virome of 10,585 individuals from the ChinaMAP. *Cell Discov.* 2022 Oct;8(1):113.

83. Moustafa A, Xie C, Kirkness E, Biggs W, Wong E, Turpaz Y, et al. The blood DNA virome in 8,000 humans. *PLoS Pathog.* 2017 Mar;13(3):e1006292.

84. Géraudie B, Charrier M, Bonnafe P, Heurté D, Desmonet M, Bartoletti MA, et al. Quantitation of human herpesvirus-6A, -6B and -7 DNAs in whole blood, mononuclear and polymorphonuclear cell fractions from healthy blood donors. *J Clin Virol.* 2012 Feb 1;53(2):151–5.

85. Miceli-Richard C, Gestermann N, Amiel C, Sellam J, Ittah M, Pavy S, et al. Effect of methotrexate and anti-TNF on Epstein-Barr virus T-cell response and viral load in patients with rheumatoid arthritis or spondylarthropathies. *Arthritis Res Ther.* 2009 May 26;11(3):R77.

86. Momin MdM, Shin J, Lee S, Truong B, Benyamin B, Lee SH. A method for an unbiased estimate of cross-ancestry genetic correlation using individual-level data. *Nat Commun.* 2023 Feb 9;14(1):722.

87. Virgin HW, Wherry EJ, Ahmed R. Redefining Chronic Viral Infection. *Cell.* 2009 July 10;138(1):30–50.

88. Kochi Y, Suzuki A, Yamamoto K. Genetic basis of rheumatoid arthritis: a current review. *Biochem Biophys Res Commun*. 2014 Sept;452(2):254–62.
89. Brister JR, Ako-Adjei D, Bao Y, Blinkova O. NCBI viral genomes resource. *Nucleic Acids Res*. 2015 Jan;43(Database issue):D571-577.
90. Lachenbruch PA. Comparisons of two-part models with competitors. *Stat Med*. 2001;20(8):1215–34.
91. Barreiro LB, Quintana-Murci L. From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nat Rev Genet*. 2010 Jan;11(1):17–30.
92. Quintana-Murci L. Human Immunology through the Lens of Evolutionary Genetics. *Cell*. 2019 Mar 21;177(1):184–99.
93. Duncan L, Shen H, Gelaye B, Meijsen J, Ressler K, Feldman M, et al. Analysis of polygenic risk score usage and performance in diverse human populations. *Nat Commun*. 2019 July 25;10:3328.
94. Ruan Y, Lin YF, Feng YCA, Chen CY, Lam M, Guo Z, et al. Improving polygenic prediction in ancestrally diverse populations. *Nat Genet*. 2022 May 1;54(5):573–80.