ARTICLE

Haplotype Sharing Provides Insights into Fine-Scale Population History and Disease in Finland

Alicia R. Martin,[1,2,3,*] Konrad J. Karczewski,[1,2] Sini Kerminen,[4] Mitja I. Kurki,[1,2,3,4,5] Antti-Pekka Sarin,[4,6] Mykyta Artomov,[1,2,3] Johan G. Eriksson,[6,7,8] Tõnu Esko,[2,9] Giulio Genovese,[2,3] Aki S. Havulinna,[4,6] Jaakko Kaprio,[4,10] Alexandra Konradi,[11,12] László Korányi,[13] Anna Kostareva,[11,12] Minna Männikkö,[14] Andres Metspalu,[9] Markus Perola,[4,9,15] Rashmi B. Prasad,[17] Olli Raitakari,[15,16] Oxana Rotar,[11] Veikko Salomaa,[6] Leif Groop,[4,17] Aarno Palotie,[2,3,4,5] Benjamin M. Neale,[1,2,3] Samuli Ripatti,[4,10] Matti Pirinen,[4,10,18] and Mark J. Daly[1,2,3,4,*]

Finland provides unique opportunities to investigate population and medical genomics because of its adoption of unified national electronic health records, detailed historical and birth records, and serial population bottlenecks. We assembled a comprehensive view of recent population history (≤100 generations), the timespan during which most rare-disease-causing alleles arose, by comparing pairwise haplotype sharing from 43,254 Finns to that of 16,060 Swedes, Estonians, Russians, and Hungarians from geographically and linguistically adjacent countries with different population histories. We find much more extensive sharing in Finns, with at least one ≥ 5 cM tract on average between pairs of unrelated individuals. By coupling haplotype sharing with fine-scale birth records from more than 25,000 individuals, we find that although haplotype sharing broadly decays with geographical distance, there are pockets of excess haplotype sharing; individuals from northeast Finland typically share several-fold more of their genome in identity-by-descent segments than individuals from southwest regions. We estimate recent effective population-size changes through time across regions of Finland, and we find that there was more continuous gene flow as Finns migrated from southwest to northeast between the early- and late-settlement regions than was dichotomously described previously. Lastly, we show that haplotype sharing is locally enriched by an order of magnitude among pairs of individuals sharing rare alleles and especially among pairs sharing rare disease-causing variants. Our work provides a general framework for using haplotype sharing to reconstruct an integrative view of recent population history and gain insight into the evolutionary origins of rare variants contributing to disease.

Introduction

Most rare variants that play a critical role in diseases today arose during approximately the last 100 generations and provide signatures of population history.[1] Recent large-scale DNA sequencing consortia efforts have demonstrated that one of the most predictive features of pathogenicity is allele frequency, given that most disease-causing variants are rare and thus relatively young.[2,3] These variants have not yet been fully exposed to the forces of natural selection that common, older variants have survived. However, aside from *de novo* variants in early-onset developmental phenotypes, the role of recently evolved, large-effect variants in common disease is largely uncharacterized. Stronger effects are most likely not confined to *de novo* variants but could persist for several generations; however, it has been difficult to identify this class of variation with single-variant analyses because such analyses have extremely limited power, especially for scenarios involving incomplete penetrance.[4,5] It is imperative that we better understand recent population genetic history because it bounds the ability of natural selection to purge deleterious variants during the most relevant period for producing disease-conferring variants subject to negative selection.[6,7] Furthermore, standard GWAS approaches typically include principal components to correct for population structure, but this is insufficient for rare variants.[8] Haplotype-based methods have two major benefits over single-variant approaches for inferences into demographic history and disease association: (1) as opposed to commonly used site-frequency-based approaches,[9] they are more informative of population history during the last tens to hundreds of generations, and (2) they can expose disease-causing rare variants at the population level without necessitating

[1]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA; [2]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; [3]Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA; [4]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki 00014, Finland; [5]Psychiatric and Neurodevelopmental Genetics Unit, Department of Psychiatry, Massachusetts General Hospital, Boston, MA 02114, USA; [6]National Institute for Health and Welfare of Finland, Helsinki 00271, Finland; [7]Folkhälsan Research Center, Helsinki 00290, Finland; [8]Department of General Practice and Primary Health Care, University of Helsinki and Helsinki University Hospital, Helsinki 00014, Finland; [9]Estonian Genome Center, University of Tartu, Tartu 50090, Estonia; [10]Department of Public Health, University of Helsinki, Helsinki 00014, Finland; [11]Almazov National Medical Research Centre, Saint Petersburg 197341, Russia; [12]National Research University of Information Technologies, Mechanics, and Optics, Saint Petersburg 197101, Russia; [13]Heart Center Foundation, Drug Research Centre, Balatonfured H-8230, Hungary; [14]Center for Life Course Health Research, Faculty of Medicine, University of Oulu, Oulu 90014, Finland; [15]Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, Turku University Hospital, Turku 20520, Finland; [16]Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital, Turku 20520, Finland; [17]Lund University Diabetes Centre, Department of Clinical Sciences, Lund University CRC, Skåne University Hospital Malmö, SE-205 02, Malmö, Sweden; [18]Helsinki Institute for Information Technology and Department of Mathematics and Statistics, University of Helsinki, 00014 Helsinki, Finland
*Correspondence: armartin@broadinstitute.org (A.R.M.), mjdaly@broadinstitute.org (M.J.D.)
https://doi.org/10.1016/j.ajhg.2018.03.003.

deep whole-genome sequencing. Rather, haplotype sharing can take advantage of massive, readily-available GWAS array data. Although these advantages have been theoretically recognized when sample sizes were relatively small,[10,11] they have been underutilized in the modern genomics era.

Finland provides a convenient example from which population history and rare disease associations can be inferred because of its unified electronic health records as well as the founder effect elicited by serial population bottlenecks. In addition to the out-of-Africa bottleneck experienced by all of Europe, Finland underwent multiple additional bottlenecks over the last few thousand years, and the Finnish founder population size is estimated to have included 3,000–24,000 individuals.[12–16] Archaeological evidence indicates that Finland has been continuously but sparsely inhabited since the end of the last ice age ~10.9 kya,[14] when a small population of not more than a few thousand early hunter-gatherers first settled throughout Finland mostly from the south and to a lesser extent from the east and west.[17] An ancient DNA study using samples dated to 6,000–9,500 years old across Sweden, Norway, and the Baltic Islands found evidence of multiple migration events into Scandinavia, where an east-west genetic gradient opposed the geographical gradient of modern populations.[18] A cultural split in approximately 2,300 BC was hypothesized to separate the western and eastern areas of Finland, termed the early- and late-settlement regions (ESRs and LSRs), upon the arrival of the Corded Ware culture, which was primarily restricted to the southwestern and coastal regions of the country; this split has been supported by Y chromosome and mitochondrial DNA as well as historical data.[17,19,20] During the last two millennia, a series of founding, extinction, and re-colonization events took place before continuous habitation coincident with agriculture.[21] The ESR, encompassing the southern and western colonized regions of Finland, was more densely and permanently settled beginning ~4,000 years ago, whereas the LSR, encompassing the northern and eastern regions of Finland, was more permanently inhabited beginning in the 1500s, pushing existing nomadic Sami people farther north into Lapland. Although Finland was a part of the Swedish Kingdom until 1809 and then became a semi-autonomous grand duchy controlled by tsarist Russia until it gained independence in 1917, immigration into western and especially eastern Finland was relatively low until after the collapse of the Soviet Union.[22] Linguistically, the mother tongue of roughly 5% of the population is Swedish, and both Finnish and Swedish are taught at school. Bilingual Finns who speak Swedish as their mother tongue live mostly in the early-settlement region in restricted western and southern coastal regions.

Because of serial bottlenecks in Finland, the site-frequency spectrum is skewed toward more common variants than in other European populations, and deleterious alleles are more likely to be found in a homozygous state.[16] The consequence of this is exemplified in the Finnish Disease Heritage (FinDis) database, which to date contains 36 monogenic diseases that are much more common in Finns than in any other population.[23] Several complex diseases also show strong regional clines within Finland. For example, risk of schizophrenia and familial hypercholesterolemia is greatest in northeastern Finland.[24,25] Current Finnish demographic models are primarily based on single locus markers (i.e., the Y chromosome and mitochondria),[12,19,20] and a few studies have recently expanded to incorporate autosomal data.[22,26,27] Methods based on the site-frequency spectrum consider sites independently and are therefore optimized for inferring old demographic events (>100 generations ago); by contrast, haplotype-based demographic inference is optimized for detailing population history during the period most relevant for negatively selected traits (i.e., the last 100 generations).[28–31] Multiple lines of evidence indicate that recent history is particularly important for disadvantageous traits. For example, long runs of homozygosity (ROH), a special case of recent haplotype sharing, are enriched for deleterious variation,[32] and increased ROH have been associated with decreased educational attainment as well as intellectual disability.[33,34] Furthermore, allele dating techniques indicate that pathogenic variants are on average considerably younger than neutral variants.[3]

Prior studies have used haplotype-based inference of pairwise sharing to query the recent demographic history of other populations, for example those with massive, densely connected genetic networks across the US.[35] This type of inference has provided regional insights into how African American migration routes, which differ markedly from those of European Americans, have changed since the dawn of slavery in the US.[36,37] High levels of haplotype sharing have been observed in a number of other founder populations, including the Druze;[38] Ashkenazi and other Jews;[39–41] Indians;[42] French Canadians;[43] Hispanic/Latino and Native Americans;[44–46] European isolates;[47–51] and other European populations.[30,52–54] However, few studies have investigated recent demographic history in depth with pairwise haplotype sharing in Finns, who are among the best-studied population isolates.[13,22,55]

In this study, we combined biobank-scale genetic and detailed birth-record data to assemble a comprehensive inquiry into recent population history by employing genetic data from 43,254 Finnish individuals (~0.8% of Finland's total population) and 16,060 demographically distinct individuals from geographically or linguistically neighboring countries, including Sweden, Estonia, Russia, and Hungary. Although Finland is a poised example for population insights from haplotype sharing because of its serial population bottlenecks, our approach provides a general framework for using haplotype sharing to reconstruct an integrative view of recent population history within and across countries (e.g., elucidation of changes in migration, divergence, and population size over time). Through these analyses, we also demonstrate that elevated haplotype-sharing patterns

resulting from multiple population bottlenecks provide insights into the origins of certain genetic diseases.

Material and Methods

Genotyping Datasets

Finnish samples were genotyped for various projects, all of which have been published previously and most of which were described in Surakka et al.[56] In brief, study participants were as follows: European Network for Genetic and Genomic Epidemiology (ENGAGE) Consortium, Myocardial Infarction Genetics Exome Array Consortium,[16] FINRISK (1992, 1997, 2002, and 2007) cohorts, Northern Finland Birth Cohort 1966, Corogene controls (which are also from FINRISK), Health 2000 samples from the GenMets study, the Helsinki Birth Cohort Study, the Cardiovascular Risk in Young Finns Study, and the Finnish Twin Cohort. All birth records are from the FINRISK study, which is a superset of several projects. The FINRISK 1997 cohort contains municipality-level birth records (n = 3,942), and the 2007 cohort contains region-level birth records (n = 5,448), which were genotyped across different projects and/or arrays (Table S1). Swedish samples used here were waves 5 and 6 (Sw5 and Sw6) and were genotyped as part of a schizophrenia study.[57] Swedish genotype data are available upon application from the National Institute of Mental Health (NIMH) Genetics Repository (see Web Resources). Estonian samples are from the Estonian Genome Center, University of Tartu.[58] Genotyping for individuals from St. Petersburg, Russia was performed as a part of a starvation study ongoing at the Broad Institute on a cohort previously described in Rotar et al.[59] Hungarian samples included in the study were genotyped as part of the Hungarian Transdanubian Biobank.[60] Genotyping details and sample sizes are shown in Table S2.

Exome Sequencing Data and Quality Control

Exome sequencing data from multiple studies of Finnish individuals were collected and harmonized as part of Sequencing Initiative Suomi (SISU) study (Table S3). The Finnish sequence data processing and variant calling is similar to that described previously.[61] In brief, we filtered these data so that what remained were exomes with overlapping GWAS data from unrelated individuals in this study (n = 9,363), as described below in "Haplotypes Overlapping Exome Variants." These individuals were primarily from the FINRISK study obtained through dbGaP.[62] We performed sample and variant quality control after joint calling to assess the relationship between rare variation and pairwise haplotype sharing. We first removed any individuals with missingness > 10% at sites where allele frequency is greater than 0.001 and missingness is less than 10%. We then filtered to variants meeting genotype quality filters as follows: genotype quality ≥ 20, depth ≥ 10X, and allele balance > 0.2. We then extracted variants present at least twice, with a call rate > 0.8, and excluded variants that failed GATK VQSR quality.

Phasing and Imputation

All Finnish genotypes underwent quality control, phasing, and imputation, as described previously.[56] Imputation was performed with the 1000 Genomes Project data.

Principal-Component Analysis

We combined best-guess genotypes for 43,254 Finnish individuals whose variants had been imputed with info score > 0.99 across all arrays, including the Affymetrix Genome-Wide Human SNP 6.0, Illumina Human 370k, 610k, 670k, Core Exome, and OmniExpress arrays. This resulted in ~3.4 million accurately imputed common SNPs across all individuals. From these sites, we performed linkage-disequilibrium (LD) pruning by using PLINK v1.90b3f[63] and keeping SNPs with minor-allele frequencey (MAF) > 0.05, missingness < 10%, and $R^2 \leq 0.50$ by using a window size of 50 SNPs and a 5 SNP overlap between windows. PCs were computed across 232,332 sites for all Finnish individuals with flashpca.[64] We also generated a multi-population dataset of unrelated individuals with birth records, when available, from Finland; Sweden; Estonia; Hungary; and St. Petersburg, Russia. As before, we extracted best-guess Finnish imputed sites with info score > 0.99. We also filtered to individuals with ≤10% missingness, sites with ≤10% missingness, MAF ≥ 0.05, and LD $R^2 < 0.5$. Because of array heterogeneity, we also filtered to sites on the Illumina Global Screening Array to avoid removing all Russian individuals because of high missingness. We then ran principal-component analysis (PCA) with 65,224 sites across n = 11,287 individuals.

Genetic Divergence

We computed F_{ST} among geographical regions by using PLINK v1.90b3f.[63] For all analyses, we used the weighted Weir-Cockerham F_{ST} estimate.

Genetic Relatedness

We identified the maximal set of unrelated individuals separated by at least two degrees of relatedness by using KING v2.0[65] within each population. We identified a maximal unrelated set of 34,737 Finnish individuals, 7,863 Swedish individuals, 6,328 Estonian individuals, 294 Hungarian individuals, and 210 Russian individuals.

Haplotype Calling

We generated two sets of haplotypes: one by using IBDseq to assess effective population-size changes over time for Finland-only analyses,[66] and another by using GERMLINE for all other analyses.[67] We used IBDSeq rather than GERMLINE for the IBDNe analyses because of previous recommendations[68] stating that switch errors in estimated haplotypes can cause erroneous haplotype breaks, resulting in spuriously recent inferences regarding the time to most recent common ancestor; IBDseq is less susceptible to these errors because it does not rely on phased data as input. We ran IBDseq on the maximal set of unrelated individuals with birth-record data (n = 9,008 individuals with 169,306 SNPs). To perform effective population-size inferences per region, we took the subset of haplotypes where both pairs of individuals were born in the same region.

For all other analyses, we used haplotypes called with GERMLINE. We first phased all genotype data together by using Eagle v2.3.2.[69] We then generated haplotype calls by using GERMLINE (because of its computational tractability at large sample sizes) with the following parameters: -err_hom 0 -err_het 2 -bits 25 -h_extend -haploid. To investigate the decay of identity-by-descent (IBD) tract length, we used a minimum haplotype size of 1 cM (−min_m 1) within each population for unrelated samples with birth-record data and/or exome-sequencing data. When assessing haplotype sharing across the full set of unrelated genotyped Finns without respect to birth records, we set a minimum haplotype size (--min_m) of 3 cM for computational tractability and reasonable storage sizes. We removed haplotypes that

fall partially or fully within centromeres, telomeres, acrocentric short chromosomal arms, heterochromatic regions, clones, and contigs identified in the UCSC hg19 genome "gaps" table.

Because we had used imputed sites to harmonize genotype data across arrays (albeit with very high fidelity with info score > 0.99), we assessed haplotype calling concordance from imputed data versus genotype-array data. As expected, we found high concordance between genotyped and high-quality imputed sites across datasets and geographical regions, especially for haplotypes longer than 2–3 cM (Figure S7).

Haplotype Calling for Effective Population-Size Analyses

Variants imputed with an info score > 0.99 that intersected across all six arrays on which Finnish samples were genotyped (Table S2) were included in the haplotype analyses, resulting in 3.4 million accurately imputed common SNPs across 43,254 individuals. High-imputation-quality best-guess genotypes were subsequently filtered to have MAF > 0.05, no indels, and LD $R^2 < 0.5$. We ran IBDNe across regions of Finland by taking the subset of pairs of individuals who were both born in the same region. Demographic analyses included pairwise haplotypes for individuals from the FINRISK 1997 and 2007 cohorts, which contained the following number of individuals by region: 1,123 in region 1, 1,078 in region 2, 378 in region 4, 224 in region 5, 304 in region 6, 1,581 in region 7; 1,547 in region 8, 225 in region 9, 228 in region 10, 1,697 in region 11, and 184 in region 12 (region names are as in Table S4).

Mapping Cumulative Haplotype Sharing

Municipality-level maps of Finland, Sweden, and Estonia were downloaded in R SpatialPolygonsDataFrame (file format S4) format from the GADM database of Global Administrative Areas (see Web Resources) on 9/14/2015, 4/13/2017, and 7/24/2017, respectively. Pairwise sharing was computed for a maximal unrelated set of individuals ($\geq 2^{nd}$ degree relatives) with municipality- or region-level birth-record data (n = 8,630 individuals total: n = 5,020 with municipality-level data from FR97, and n = 3,610 with region-level data from FR07). From each city, all pairs where at least one individual had parents born within 80 km of each other and whose mean birth location was within 80 km of the city of interest were included. Municipalities are official and were numbered as described in the Web Resources, with three additional codes: 198 = no home in Finland, 199 = unknown, and 200 = abroad. To account for uncertainty when only region-level data were available, even weights were assigned to all municipalities within that region with the sum of the weights equal to 1; in contrast, a single municipality was given a weight of 1 in the municipality-level data.

Estimating Effective Migration Surfaces

We used estimating effective migration surfaces (EEMS) tool[70] to estimate migration and diversity relative to geographic distance. We computed genetic dissimilarities for all unrelated pairwise individuals for whom municipality-level birth-record data were available and whose parents were both within 80 km; we used mean parental latitude and longitude when the parents' data differed. We computed pairwise genetic dissimilarities by using the *bed2diffs* tool provided with EEMS on the intersected Finnish data, which included 232,332 SNPs for 2,706 individuals, as well as on the intersected Finnish, Swedish, Estonian, and Russian data, which included 88,080 genotyped SNPs across 10,993 individuals. We set the number of demes to 300 (but actually observed fewer than this) and adjusted the variances for all proposal distributions of migration, diversity, and degree-of-freedom parameters such that most were accepted 20%–30% of the time and all were accepted 10%–40% of the time, per recommendations in the manual. We increased the number of Markov chain Monte Carlo (MCMC) iterations, burn-in iterations, and thin iterations until the MCMC converged.

Whereas Finland birth records used in this analysis are at the municipality level, Swedish and Estonian birth records are at the region level. Because of differing birth-record densities and boundaries in Finland-only versus multi-country analyses, there are differing densities and numbers of observed demes. When setting nDemes = 300 across Finland, Sweden, Estonia, and St Petersburg, Russia, we observed 110 out of 274 demes. When setting nDemes = 300 across Finland alone, we observed 167 out of 266 demes.

Haplotypes Overlapping Exome Variants

All analyses of haplotypes paired with exome sequencing data were performed with Hail version 0.1. To map IDs between the genotype and exome sequencing data, we filtered genotype and exome data to variants with at least 1% frequency and less than 10% missingness in each dataset and subsequently removed individuals with greater than 10% missingness. We intersected these datasets, repeated the same filtering process, and identified 9,363 individuals with both data types by using the Hail IDB function, which is equivalent to plink --genome (minimum pi_hat = 0.95). Using haplotypes called with GERMLINE, we filtered out regions of the genome shared across all pairs at a rate greater than three times the standard deviation above the mean level of sharing (Figure S6) for improved computational tractability and to remove false positives that are common in regions of the genome with very high levels of sharing. We performed this filtering to make pairing of exome and haplotype pair data computationally tractable and remove false positives that are common in regions of the genome with very high levels of sharing. We overlaid the haplotype data with the exome data by using the annotate_variants_table function and calculated the number of pairs of individuals sharing haplotypes and genotypes for each variant (we excluded singletons and variants that failed VQSR filtering) by using a custom script in the Hail expression language. In brief, we determined the set of individuals carrying each genotype and then iterated over the pairs of individuals who share haplotypes; we counted cases where both members of the pair harbored the same genotype. The number of pairs that did not share a given genotype was simply computed as the number of pairs with the genotype $(n * (n-1)/2)$ minus the number of pairs that shared the genotype. We subsequently annotated variants with VEP version 85 by using transcripts from Gencode v19 and the LOFTEE plugin (Web Resources). We then computed a simple enrichment ratio for haplotype sharing at exome sequencing variants shared among heterozygous individuals (i.e., carriers) versus homozygous reference individuals, as follows: $ratio = ((Heterozygous\ pairs\ that\ share/All\ heterozygous\ \ pairs)/(Homozygous\ reference\ pairs\ that\ share/\ All\ homozygous\ reference\ pairs))$. We stratified haplotype enrichments across allele frequencies and predicted functional variant consequence as well as variants known to cause diseases in FinDis. Starting with a list of 50 FinDis annotated autosomal variants, we found that 40 exome sequencing variants were polymorphic and had overlapping haplotypes. Table S5 contains haplotype enrichments among carrier pairs relative to homozygous reference individuals.

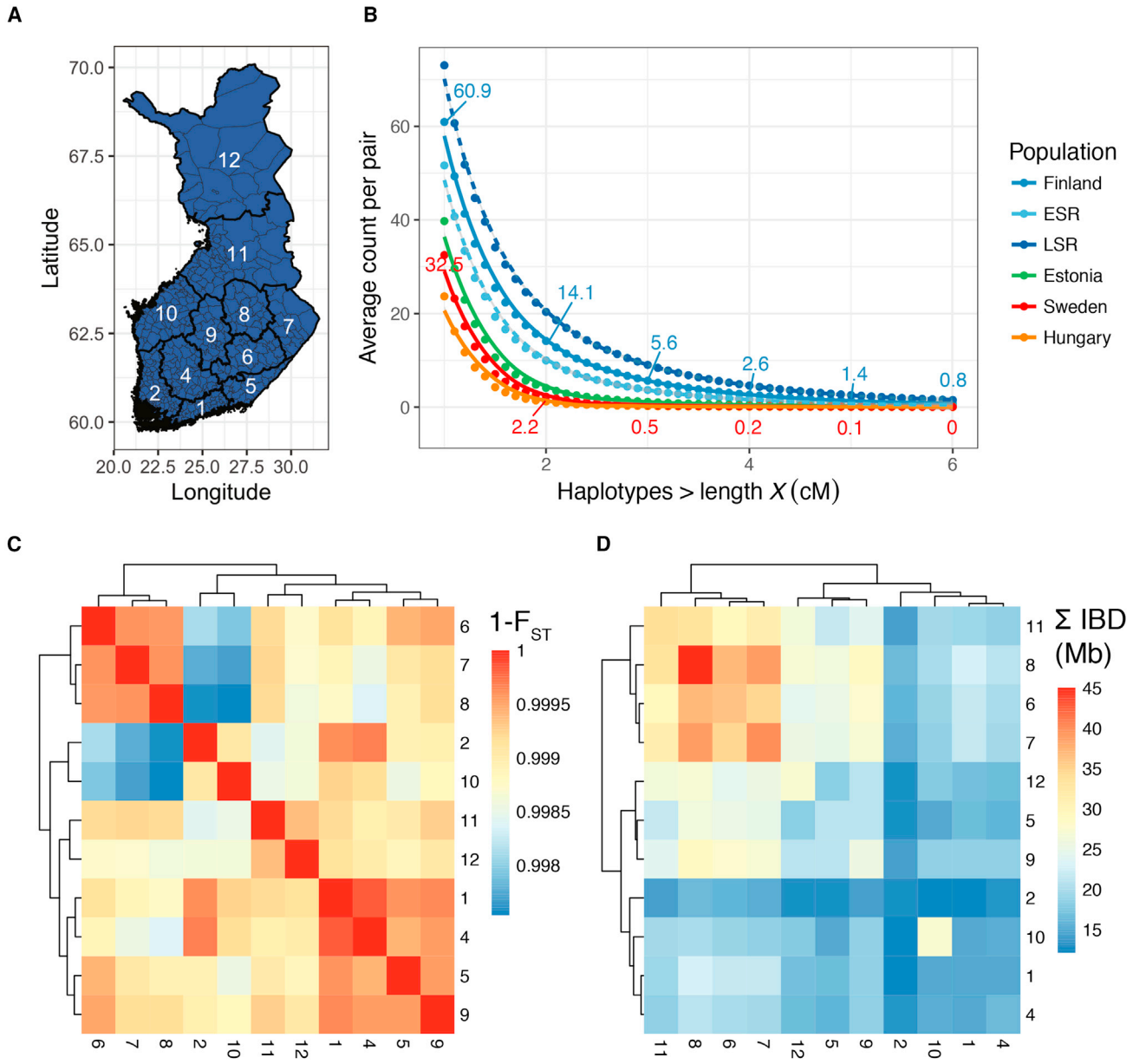

Figure 1. Identity-by-Descent Haplotype Sharing and Genetic Divergence across Regions of Finland
(A) Regional map of Finland. Region names are shown in Table S4. Thin lines within regions represent municipality boundaries. Region 3 corresponds to the Åland Islands (not shown), a small Swedish-speaking archipelago in the Gulf of Bothnia.
(B) Distribution of average pairwise shared IBD segments in Finland (N = 7,669), specifically within two birth regions defined previously as having >95% posterior probability of clustering geographically in the early-settlement region (ESR; n = 428) and late-settlement region (LSR; n = 592),[22] Estonia (n = 6,328), Sweden (n = 7,863), and Hungary (n = 294). All individuals included are unrelated and ancestrally representative of a given region or country. Numbers indicate average pairwise haplotypes shared at 1, 2, 3, 4, and 5 cM in Finland and Sweden.
(C) Hierarchical clustering of genetic similarity, as measured by $1 - F_{ST}$ across regions of Finland. Regions are numbered as in Table S4.
(D) Hierarchical clustering of cumulative IBD (minimum haplotype ≥ 3 cM) sharing across regions of Finland. Regions are numbered as in Table S4.

Results

Population Substructure across Regions of Finland

To investigate fine-scale population structure within Finland, we assembled a panel of 43,254 Finnish individuals (Table S2, Material and Methods). We performed PCA on all individuals and used the subset of individuals with recorded birth record data to show that genetic variation in Finland broadly reflects geographical birthplace (Material and Methods, Figure 1A); we found highly significant correlations between PC1 and longitude ($\rho = -0.72$, $p < 1 \times 10^{-200}$) and between PC2 and latitude ($\rho = -0.55$, $p < 1 \times 10^{-200}$). The PCA and birth-record data also reflect variability in sampling and population density; high

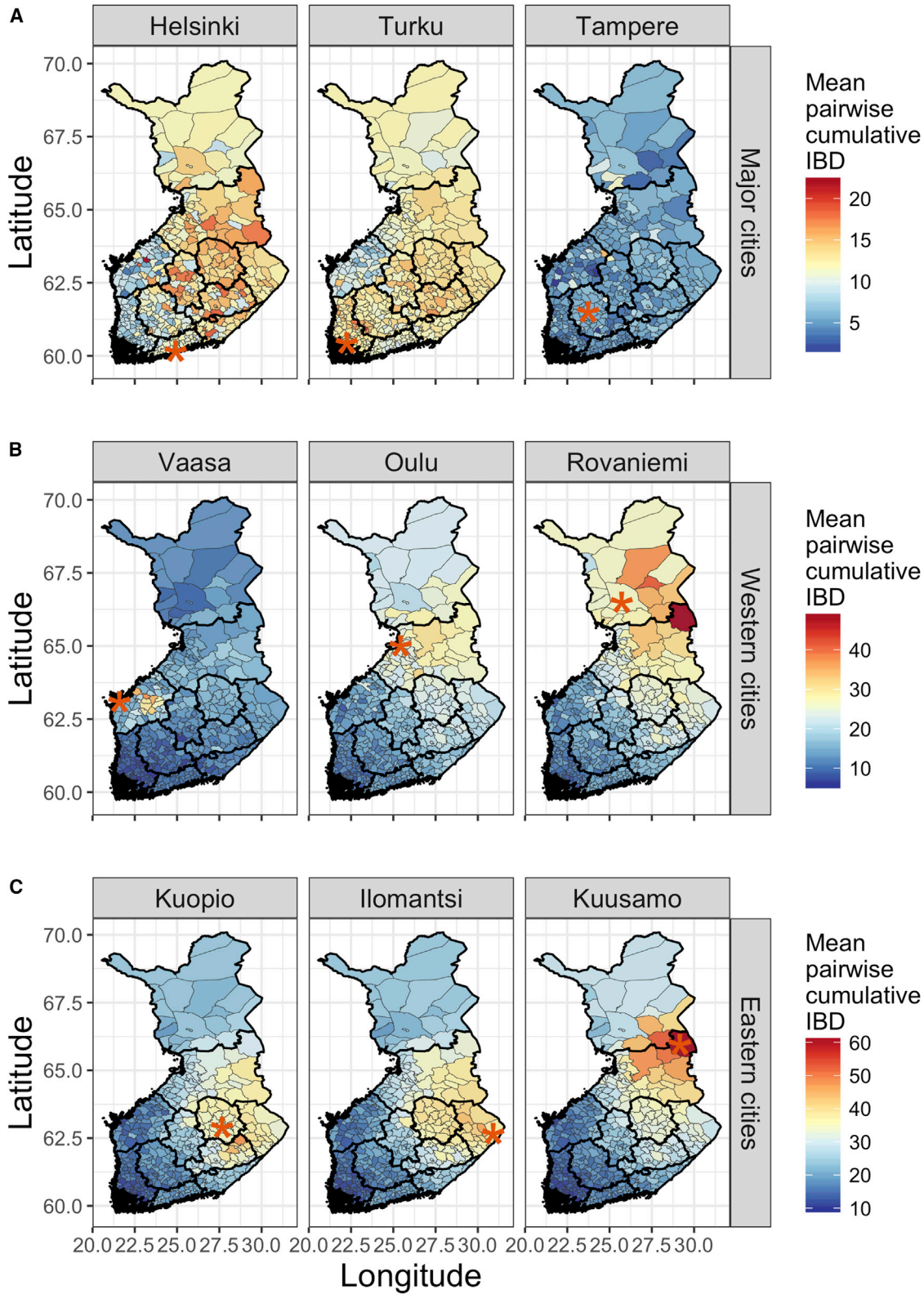

Figure 2. Geographically Structured Haplotype Sharing between Pairs of Individuals across Finland
We used genetic data to take the subset of pairs of individuals whose birth records indicate that both parents were born within 80 km (~50 miles) of each other. For each panel, we further took the subset of haplotypes from pairs of individuals in which at least one of the individual pairs lives within 80 km of cities indicated by red asterisks. Thinner lines outline municipalities, and thicker lines outline regions. The color shaded in each municipality indicates the weighted mean of cumulative IBD sharing for haplotypes $\geq$ 3 cM. Finland-wide city comparisons are grouped into three categories (note different scales for each): (A) major cities, (B) western cities, and (C) eastern cities.

(legend continued on next page)

density in Helsinki and Turku contrasts with low density in the northernmost Lapland region (Figures S1A and S1B). Mean PC1 and PC2 across birth regions closely mirror geographical patterns, with the exception of southern Finland (region 1), which projects closer to central Finland than expected geographically; containing the capital city of Helsinki, southern Finland is the most populous region of Finland and consequently draws from across the country (Figures S1B and S1C). We also assessed genetic divergence across regions in Finland and identified relatively high levels of regional divergence in Finland compared to other European countries, e.g., the UK, Germany, Sweden, and Estonia;[26,71] mean F_{ST} between region pairs was 0.001 (Figure S1C). These results are consistent with an additional Finnish bottleneck with respect to nearby countries.

Regionally across Finland, we identified geographical clusters with high degrees of similarity. For example, Southern Savonia, Northern Karelia, and Northern Savonia (regions 6, 7, and 8, respectively) exhibit high degrees of genetic similarity (Figure 1C). We also identified genetic similarity clusters in the southern central regions of Southern Finland, Tavastia, Southern Karelia, and Central Finland (i.e., regions 1, 4, 5, and 9); western coastal regions of Southwest Finland and Ostrobothnia (2 and 10); and northern regions of Northern Ostrobothnia and Lapland (11 and 12). By comparing parent and offspring birthplaces, we show that within a single generation, offspring across Finland tend to move south, e.g., toward Helsinki (Kolmogorov-Smirnov two-sided test between and child's and mean parents' latitude: $p = 8.7 \times 10^{-3}$ (Figure S2).

Population Bottlenecks in Finland Are Reflected in Identity-by-Descent Sharing

To better understand the recent population history of Finland, we computed pairwise IBD sharing across all unrelated Finnish pairs of individuals (Material and Methods, Figures 1B and 1D). We performed hierarchical clustering of cumulative IBD sharing across pairs of individuals within and between regions of Finland, and we identified excess sharing in eastern Finland (regions 6, 7, and 8) compared with southwestern Finland (regions 1, 2, 4, and 10), where sharing was depleted (Figure 1D). Compared to genetic similarity from common variants (Figure 1C), haplotype-based clustering is more consistent with historical records that have documented the early- versus late-settlement regions in southwest and northeast Finland, respectively. Nonetheless, pairwise regional IBD and F_{ST} are highly correlated (Mantel test $\rho = 0.89$, $p < 1 \times 10^{-4}$ with 1,000 Monte Carlo repetitions). Previous work on serial founder effects showed that global genetic divergence increases with geographical distance,[72] and we recapitulated this finding at the sub-country level within Finland (Figure S3A); we also identified decaying IBD sharing with increasing geographical distance within Finland (Figure S3B).

Because Finland historically has shared trade, language, and migration with neighboring countries and/or regions, including Sweden; Estonia; and St. Petersburg, Russia, we compared the relative level of allelic and haplotypic sharing within each population. We also compared these genetic data with individuals from Hungary because although it is geographically distal, it shares common linguistic roots; Finnish is a Uralic language that forms an outgroup to most European languages but is related to Estonian and Hungarian. Comparing pairwise IBD sharing within each of these countries, we found that cumulative IBD sharing between pairs of individuals is on average significantly greater across pairs of individuals in Finland than in Sweden, Estonia, Russia, and Hungary, which is expected from the Finnish population bottleneck (cumulative total of tracts $\geq$ 1 cM in length: $\mu_{Sweden} = 22.9$ cM versus $\mu_{Finland} = 107.0$ cM, $p < 1 \times 10^{-50}$). Consistent with this observation, the average pair of Finns shares more haplotypes than the average pair in the other countries compared here, and these haplotypes are also longer: for example, 5.6 haplotypes $\geq$ 3 cM shared in Finland versus an order of magnitude fewer (0.5 haplotypes $\geq$ 3 cM) in Sweden (Figure 1B).

Recent Gene Flow and Migration Inference from IBD Sharing

We coupled haplotype sharing between pairs of individuals with municipality- and region-level birth-record data to determine relative rates of sharing among fine-scale locations in Finland. We used pairwise IBD to take the subset of individuals in which both parents were born within 80 km (~50 miles) of each other. For each analysis, we further took the subset of pairs of individuals in which at least one individual had municipality-level birth records from within 80 km of a given city and assessed average pairwise IBD with other individuals across municipalities and regions of Finland. By comparing pairwise sharing from different Finnish cities, we found that IBD sharing is very uneven throughout the country, varying by several-fold, and that different geographical regions exhibit considerable substructure with differential IBD sharing patterns (Figure 2). This fine-scale structure is most likely driven by multiple bottlenecks, recent migration patterns, and variable population density (e.g.,

For each city, the number of unique individuals whose parents are both from within an 80 km radius and total pairwise comparisons across Finland are as follows: n = 152 in Helsinki, 677,844 total pairwise comparisons; n = 227 in Turku, 1,003,794 total pairwise comparisons; n = 102 in Tampere, 457,419 total pairwise comparisons; n = 50 in Vaasa, 225,525 total pairwise comparisons; n = 185 in Oulu, 821,955 total pairwise comparisons; n = 13 in Rovaniemi, 58,877 total pairwise comparisons; n = 566 in Kuopio, 2,406,915 total pairwise comparisons; n = 363 in Ilomantsi, 1,580,502 total pairwise comparisons; and n = 25 in Kuusamo, 113,075 total pairwise comparisons.

genetic diversity is higher, and thus IBD sharing is lower, in densely populated Helsinki than many rural areas because Helsinki ancestors have more diverse origins).

Haplotype sharing is on average lowest when at least one individual lives in a major southern Finnish city (Figure 2A). Specifically, pairwise haplotype sharing is relatively low across Finland when at least one individual lives in Helsinki, Turku, and Tampere, which exhibit the lowest structure of cities compared here. Among individuals born in Helsinki, a relatively young capital (it became so in 1812), there is a subtle structure indicated by greater haplotype sharing with eastern Finland than western Finland on average; in contrast, individuals from the historical capital of Turku have more elevated haplotype sharing with nearby southwestern Finland (Figure 2). IBD sharing among western coastal cities (e.g., Vaasa, Oulu, and Rovaniemi) are intermediate and show varying patterns of regional haplotype sharing (Figure 2B). For example, Vaasa, a bilingual city with mostly Finnish and Swedish speakers surrounded by municipalities where people mostly speak Swedish, shows restricted patterns of elevated sharing specifically in Ostrobothnia (region 10). In contrast, Oulu and Rovaniemi in Northern Ostrobothnia and Lapland (regions 11 and 12) show broadly elevated patterns of sharing in the late-settlement region and depleted sharing in the early-settlement area. IBD sharing is generally highest among individuals living in northeastern cities in the late-settlement region (e.g., Kuopio, Ilomantsi, or Kuusamo); more structure is evident in the cosmopolitan cities, and greater sharing is evident in the late-settlement region (Figure 2C). Of all cities investigated, Kuusamo shows the most elevated IBD sharing: in haplotypes $>$ 3 cM, ~60 Mb on average is shared with nearby individuals, whereas ~5–15 Mb is shared near Helsinki, Turku, and Tampere.

Fine-Scale Population Differentiation and Migration Rate Inference between Finland and Nearby Countries

We assessed how much sharing occurs within and between regions of Finland and neighboring countries and/or regions, including Sweden; Estonia; St. Petersburg, Russia; and Hungary (Figure 3A). PCA recapitulates geographic boundaries and Finnish bottlenecks: PC1 separates Finland from non-Finnish Europeans, and PC2 separates non-Finnish European populations along a cline (Figure 3B).[58,73] Birth regions also recapitulate expected trends; for example, southern Finns project closer in PCA space with northern Estonians than with individuals from other regions of either country (Figure 3B). Hierarchical clustering of genetic divergence (F_{ST}) within and between regions and countries demonstrates that divergence is typically smallest within countries, with the exception of Finland and the northernmost Swedish region, Norrbotten, which neighbors Finnish Lapland. These two areas cluster together, albeit with the greatest divergence within Finland plus Norrbotten (Figure S1D). Together with the migration-rate analysis, our results suggest that although Norrbotten is most genetically similar to Finnish Lapland, a migration barrier still separates these two counties. Individuals from the southwest coastal regions of Finland (regions 1, 2, 10, and 4; i.e., Southern Finland, Southwestern Finland, Ostrobothnia, and Tavastia) are more genetically similar to cosmopolitan Swedes than other Finns are (Figure S1D, Figure 3A). The divergence is greatest (F_{ST} ~0.01) between eastern Finland (regions 6, 7, 8; i.e., Southern Savonia, North Karelia, and Northern Savonia) and the regions located within Hungary and southern Estonia (regions 30, 34, and 36) (Figure S1D). The elevated IBD sharing in Finland and the elevated divergence in relation to neighboring countries supports the utility of haplotypes for investigating recent population history as well as IBD mapping for identifying rare associations.[74]

We also utilized the granular birth records to investigate geospatial migration rates (m) in Finland and among neighboring countries. We used a spatially explicit statistical model to estimate effective migration surfaces (via EEMS) by measuring effective migration rates from genetic differentiation (i.e., resistance distance) across neighboring demes.[70] By measuring the genetic distance between evenly spaced demes relative to other pairs of demes across Finland and/or neighboring countries, we inferred locations where migration was uncommon (such locations are referred to as migration barriers and are depicted in dark orange) and where migration excesses occurred (these locations are depicted in blue) (Figures 3C and 3D). Across Finland, we found variable migration rates, many of which are consistent with known historical events (Figure 3C). For example, we identify migration barriers generally separating the early and late settlement area (i.e., between Tampere and Kuopio) as well as the northernmost Lapland region from the rest of Finland. In contrast, within Finland, there is increased migration in and directly surrounding several coastal cities, including Helsinki, Turku, Vaasa, and Oulu.

When one considers migration rates among individuals with birth records from Finland; Sweden; Estonia; and St. Petersburg, Russia (Figure 3D), the major migration routes within Finland remain broadly consistent. For example, a barrier to migration between the early and late settlement regions between Tampere and Kuopio remain, along with a barrier of migration into Lapland. The starkest difference is a barrier to migration along nearly the entire Finnish border (Figure 3D), most likely due to the absence of some neighboring comparison demes in Figure 3C (see also Figure S5), indicating little significant migration into Finland in the last 100 generations, consistent with the described patterns of low frequency variation presenting as a bottleneck or isolate. Apart from migration rate inferences along the border, subtle changes within Finland are most likely due to additional smoothing because of a larger area over which demes are spread (Material and Methods, Figure S5). Migration rates within Sweden are most elevated in southern regions near the

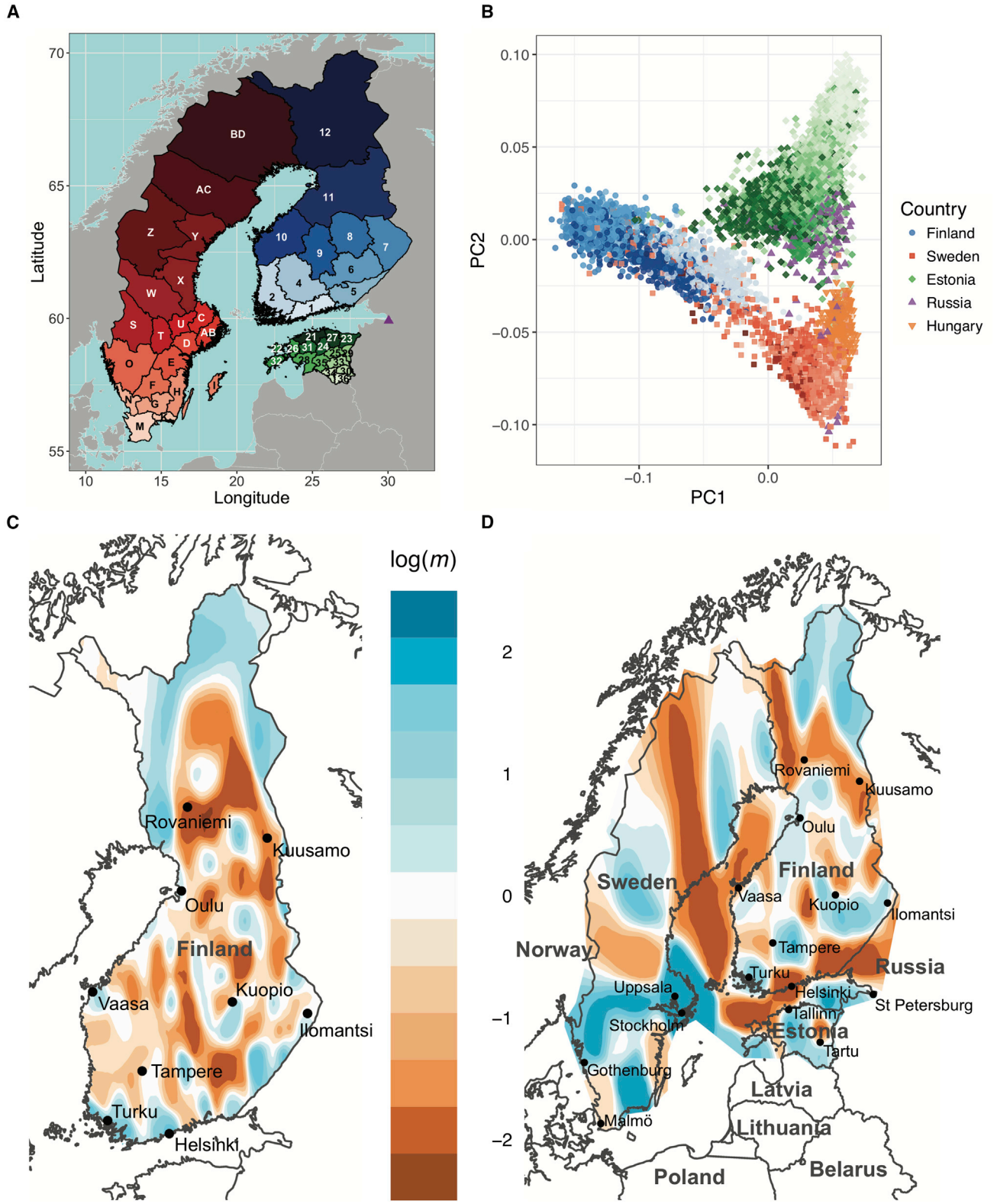

Figure 3. Migration Rates and Haplotype Sharing within Finland and Between Neighboring Countries
(A) Map of regional Finnish, Swedish, and Estonian birthplaces. A purple triangle indicates St. Petersburg, Russia. Hungary is not shown. Finnish, Swedish, and Estonian region labels are shown in Table S4.
(B) PCA of unrelated individuals, colored by birth region (if available; otherwise, by country) as shown in (A).
(C and D) Migration rates inferred with EEMS. Values and colors indicate inferred rates, *m*: shades of blue indicate logarithmically higher migration at a given point on average (i.e., log(m) = 1 corresponds to effective migration that is 10-fold faster than the average), and shades of orange indicate migration barriers. (C) Migration rates among municipalities in Finland. (D) Migration rates within and between Finland; Sweden; Estonia; and St. Petersburg, Russia.

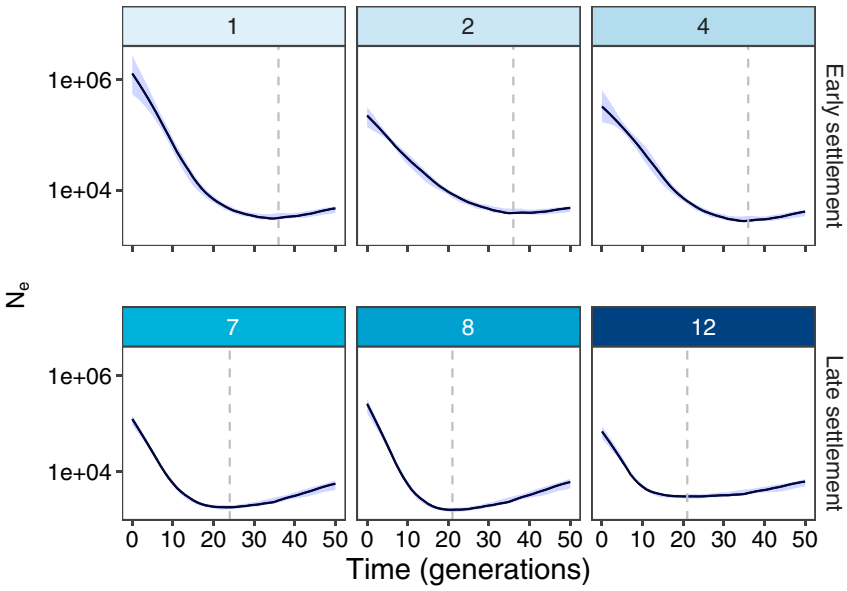

Figure 4. Effective Population Size over Time by Birth Region in Finland
Representative regions within the early- and late-settlement regions are numbered as shown in Table S4. Dashed lines indicate the time at which the minimum N_e over the last 50 generations occurred in each region. The number of individuals in each region is shown in Figure S4. Error bars indicate 95% bootstrap confidence interval.

largest cities, including Stockholm and Uppsala. As speculated previously,[58] migration rates are generally elevated within Estonia but depleted along the west coast and between Tallinn and Tartu; rates are also depleted between the Estonia mainland and both Finland and Sweden. The strongest barriers to migration in and near Sweden are in the northwest as well as along the northwestern Finnish border separating Finnish Lapland and Sweden, although there are notably few individuals either sampled or living there, resulting in increased noise.

Regional Recent Effective Population Size Changes over Time

Haplotype sharing also enables an assessment of fluctuations in effective population size over time and across geographical regions. We inferred changes in effective population size over recent time across birth regions in Finland by using the haplotype-based IBDNe method.[68] Across all birth regions, we identified a population expansion in the last 50 generations from around 10^3 to 10^5 and 10^6 (Figure 4 and Figure S4). The region with the largest current effective population size is Southern Finland (region 1, current $N_e = 1.3 \times 10^6$, 95% CI = [5.5×10^5, 2.8×10^6]), which contains the capital city of Helsinki; these findings closely approximate current census data (current census population = $\sim 1.6 \times 10^6$). We inferred that Lapland (region 12), the northernmost and least populated region, had the least growth: current $N_e = 6.9 \times 10^4$, 95% CI = [4.9×10^4, 8.8×10^4] (current census population = $\sim 1.8 \times 10^5$). The inferred effective population size is expected to be smaller than the census size because the census size includes multiple generations, variance in reproductive rates, and other factors.[68]

When comparing the early- and late-settlement regions, we found consistently earlier onset of population expansions in the early-settlement region. In the early-settlement region, for example, the population began expanding around 30–40 generations ago (circa 760–060 AD, if we assume a generation time of 30 years[75]). In contrast, the late-settlement region began expanding between approximately 15 and 25 generations ago (circa 1210–1510 AD) and had lower minimum effective population sizes (Figure 4). We also found significant evidence of a geographical cline, wherein populations began expanding earlier in regions farther south ($\rho = 0.79$, $p = 4.2 \times 10^{-3}$). For example, whereas Southern Finland and Southwestern Finland (regions 1 and 2) began growing ~36 generations ago, the northernmost region of Lapland (region 12) only began growing ~21 generations ago. We also infer larger current effective population sizes in the early- rather than late-settlement region, consistent with a higher population density for Finland in the early-settlement region. Together, the estimation of the regional expansion of the population in conjunction with IBD sharing within and between municipalities provides a clear picture of the population history as calculated entirely from genetic analysis of the modern Finnish population.

Haplotype Insights into Disease

To better understand the utility of IBD sharing for rare-variant interpretation, we coupled haplotype tracts with exome sequencing data (Material and Methods). A key consideration here is the quality and consistency of haplotype calls. To call the primary haplotypes used throughout this study, we used phased best-guess genotypes from very high-quality imputed sites (info score > 0.99) as input in GERMLINE to ensure overlap across several genotyping arrays. To look at an explicitly genotyped set, we compared haplotypes from the imputed set to haplotypes for overlapping samples genotyped as part of the ENGAGE consortium on the IlluminaCoreExome array because these had the largest number of individuals for whom birth records were available and who were also genotyped on a single array (Table S1 and Table S2). By phasing these overlapping samples and calling haplotypes separately, we could assess the fraction of haplotypes that overlapped, that were genotype specific, and that were imputation specific as a function of haplotype length and geography (Figure S7). As expected, shorter haplotypes (i.e., < 1.5 cM) were less concordant across call sets. In contrast, longer haplotypes were more

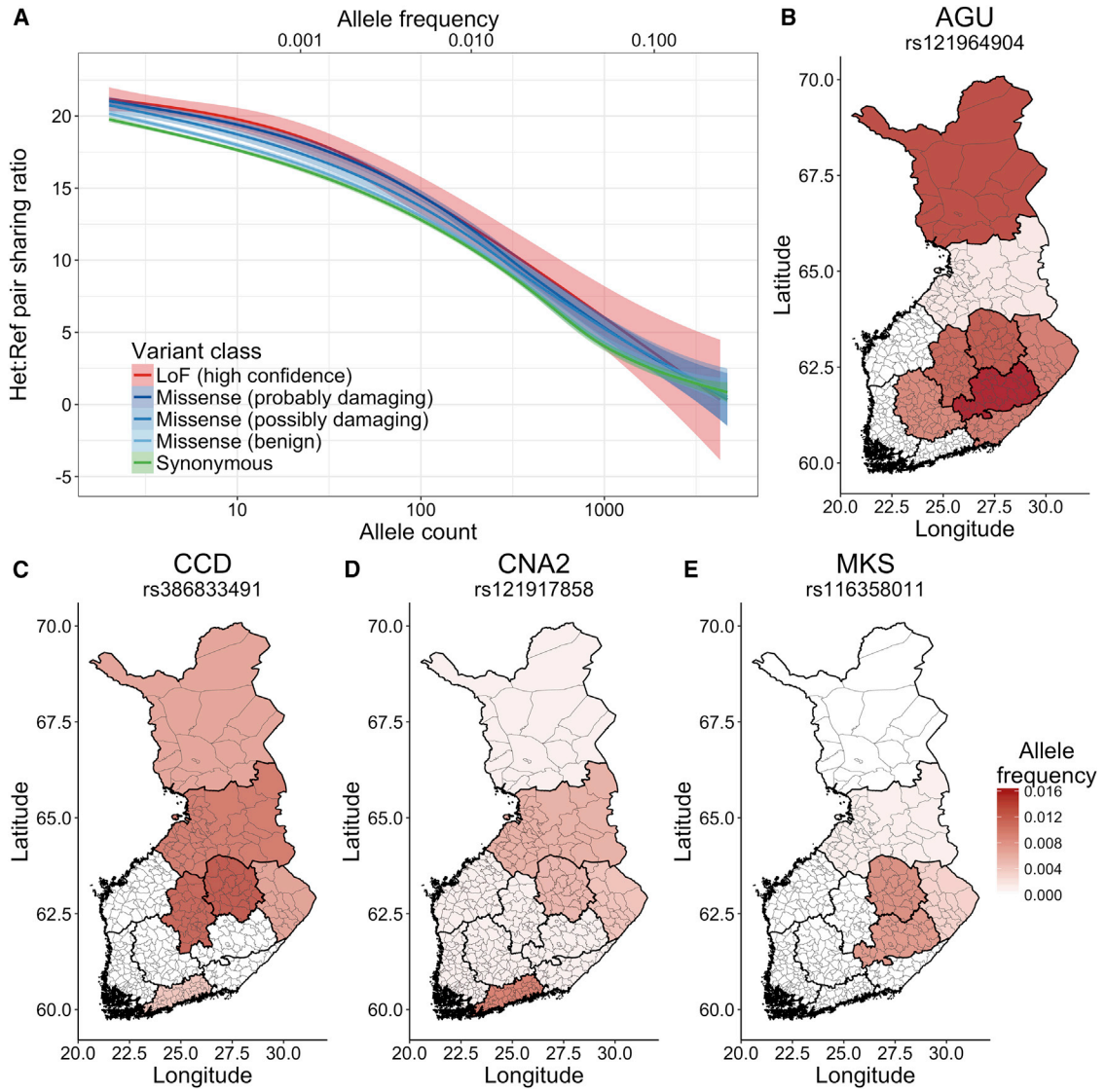

Figure 5. Haplotype Sharing Enrichment across Variant Classes and in Finnish Heritage Diseases
(A) Haplotype sharing enrichment among pairs of individuals who are heterozygous versus homozygous reference (Material and Methods). Variant class curves differ most at the very low to low frequency ranges. The normal confidence intervals show the standard errors for each variant class.
(B)–(E) Allele frequency maps for known Finnish heritage disease variants. The same allele frequency scale is included for each of these plots and is shown on the bottom right. (B) AGU, Aspartylglucosaminuria. (C) CCD, congenital chloride diarrhea. (D) CNA2, cornea plana 2. (E) MKS, Meckel syndrome.
Additional haplotype summaries of these variants are shown in Table 1.

concordant across call sets; a plateau began between 2 and 3 cM. We compared haplotype concordance across geographical regions, including the early- and late-settlement regions, and found highly consistent concordance, indicating that haplotype calls from high-quality imputed sites do not induce haplotype calling error rates that vary by geography specifically within Finland (Figure S7). Because haplotype sharing rates differ substantially across populations (Figure 1), this empirical relationship will most likely vary with population history (e.g., longer haplotype thresholds will most likely be useful in Sweden because the true rate of haplotype sharing is an order of magnitude lower for haplotypes $>$ 3 cM). This empirical framework provides a valuable metric for assessing the appropriate threshold for reliably discovering high-fidelity IBD segments in different populations.

Because previous work in population genetics has suggested that haplotype lengths provide insight into the age of alleles[76] and that younger alleles are more likely to be deleterious,[3] we quantified the extent of haplotype sharing across predicted functional classes of variants and across genotype states. We found, as expected, that there is generally more haplotype sharing at the rare end of the allele-frequency spectrum (Figure 5A). Additionally, we identified greater haplotype sharing at the rarest allele frequencies in the predicted relative order of deleteriousness among

Table 1. Enrichment of Haplotype Sharing that Overlaps FinDis Variants

Disease Code	Gene	rsID	Chr	Pos	Ref	Alt	Freq	Reference Pair Ratio	Carrier Pair Ratio	Haplotype Enrichment
AGU	*AGA*	rs121964904	4	178359918	C	G	0.79%	0.02	0.25	10.4
CNA2	*KERA*	rs121917858	12	91449319	T	C	0.52%	0.03	0.82	23.5
CCD	*SLC26A3*	rs386833491	7	107427289	AACC	A	0.60%	0.02	0.90	36.1
MKS	*CC2D2A*	rs116358011	4	15538697	C	T	0.30%	0.03	0.38	14.4

Haplotype enrichment is computed as in Figure 5 and the Material and Methods (in brief, the rate of haplotype sharing among pairs of heterozygous individuals per total number of heterozygous pairs relative to homozygous reference pairs). AGU, aspartylglucosaminuria; CNA2, cornea plana 2; CCD, congenital chloride diarrhea; MKS, Meckel syndrome.
See Table S5 for all reviewed FinDis variants.

missense variants (i.e., probably damaging > possibly damaging > benign), all of which exhibit greater haplotype sharing than synonymous variants. CpGs modestly disrupt haplotype patterns at the rarest allele frequencies (Figure 5A and Figure S8), which is most likely a product of mutational recurrence. Haplotype sharing rates are similar both in loss-of-function and missense-constrained genes (Figure S9), and these rates show similar signatures of mutational recurrence modestly disrupting haplotype sharing at CpG sites.

We also assessed the overlap of haplotypes for several known disease variants from the FinDis database (Material and Methods, Figures 5B and 5C). Across the genome, there is a 3% chance that two unselected Finns share a $\geq$1cM haplotype at any position. Considering a set of disease variants with 0.25%–1% frequency, we first confirmed that indeed homozygous reference individuals (non-carriers) share a haplotype spanning the mutation site at this same background rate. For pairs of individuals who are both carriers of a FinDis variant, however, the likelihood of sharing a haplotype $\geq$1 cM is an order of magnitude higher (~30% or higher, Table 1). This enrichment of sharing among carriers belies the conceptual framework of IBD mapping, highlighting the power to detect rare, disease-associated loci. We focused on several validated causal variants known to confer disease (Table 1 and Table S5); such variants included alleles of the in-frame deletion of rs386833491 (ref = AACC and alt = A in Table 1), which is known to confer congenital chloride diarrhea. This allele is not imputable with the standard 1000 Genomes Project reference panel. We find a significant enrichment of haplotype lengths among pairs of individuals who are both heterozygous (mean length = 6.7 cM) versus both homozygous reference (mean length = 5.6 cM) for the rs386833491 deletion (t test $p = 5.3 \times 10^{-71}$, Figure S10). This deletion is most likely slightly more enriched for haplotype sharing beyond the other FinDis variants because of the regional specificity and origins in the late-settlement region (Figure 5C).

Discussion

Bringing together genetic and birth-record data, we have constructed one of the most comprehensive genetic studies of population history to date. By coupling Finnish population history with spatial genetic analyses, we have inferred the timing of bottlenecks and expansions, deduced movement across Finland and neighboring countries, and assessed divergence and similarity during recent epochs. Our results demonstrate that prior dichotomous descriptions of the early- versus late-settlement regions correlate with our findings but are insufficient to explain the multi-generational continuous southwest-to-northeast migration trends that correspond with additional bottleneck signatures. We have recapitulated the observation that genes mirror geography at a broad scale,[77] but we have shown that at a more granular level within this founder population, demographic fluctuations have been spatially structured by the local environment and other movement-inhibiting or movement-accelerating factors such as linguistic or cultural differences and forced migration events, e.g., by Swedish kings. This comprehensive study of Finnish population history was especially powerful for investigations of recent history over the last 100 generations through statistical analyses of pairwise genomic sharing via haplotypes among fine-scale regions.

The concept that haplotype tracts assessed from common-variant GWAS arrays can provide insight into both population history and rare disease without sequencing data harkens back to the International HapMap Project and earlier.[10] Although these ideas have been around for decades, their implementation in biobank-scale data is now feasible and shows promise in isolated populations.[78] Using data from Finland, we demonstrate that haplotypes provide insight into the evolutionary timeline and class of variants of greatest interest for this study: recent population history over the past 100 generations and rare, deleterious variants. Coupled with birth-record data, haplotype tracts allow deeper insight into fine-scale substructure, including differential sharing within and across coastal and inland municipalities in the early- and late-settlement regions of Finland, than common allele approaches alone.

Finland is particularly amenable for an investigation of recent population history because it has gone through multiple well-documented bottlenecks, has considerable population substructure compared to those of many other

countries,[22,26,27] and has a universal health care system with integrated registry information. The relatively high genetic divergence between the early- and late-settlement regions has been well documented in prior genetic analyses; we demonstrate much more granular resolution into differential rates of haplotypes across Finland at the level of municipality: for example, the several-fold differences in cumulative sharing across Finland between major urban southwest cities (e.g., Turku and Helsinki) and isolated late-settlement regions (e.g., Kuusamo).

The founder effects in Finland have resulted in a massive enrichment of longer haplotypes in Finns relative to non-Finnish European neighbors. Additionally, these effects have depleted genetic diversity overall and increased relatively common deleterious variants with respect to non-Finnish Europeans.[16] A consequence of these bottleneck signatures is the utility of population-based linkage analysis for discovering deleterious variants at the rare end of the frequency spectrum. Many of the founder mutations contributing to the FinDis database were originally discovered through family-based linkage analysis.[23] The emergence of biobank-scale genetic and clinical data allows researchers to use population-based linkage analysis to discover rare-variant associations with previously undiscovered diseases or in populations where risk was previously unrealized, such as in the case of a rare orthopedic collagen disorder that conferred extreme short stature and dysmorphic features in Puerto Ricans.[78] Our work and previous studies suggest that coupling population-based linkage analysis with electronic health records provides a powerful tool for gaining rare-disease insights, particularly in populations that have gone through a historical bottleneck.[29,78,79] Furthermore, researchers can query the role of these rare variants in complex disease by using GWAS arrays to construct kinship matrices from pairwise haplotype sharing to understand a more complete spectrum of allele frequencies in overall heritability.[80,81] This study demonstrates the utility of haplotype sharing for historical demographic inference and for identifying rare variants that confer risk of rare disorders in isolated populations, such as Finland, that have data from unified health care registries.

Supplemental Data

Supplemental Data include 10 figures, five tables, and one Supplementary Note and can be found with this article online at https://doi.org/10.1016/j.ajhg.2018.03.003.

Acknowledgments

Thanks to the participants in the Finnish cohort studies. Thanks to the sequencing centers at Washington University, the Broad Institute, and the UK10K project for generation and deposition of exome sequencing data from FINRISK and other Finnish cohorts. We thank Eimear Kenny and Gillian Belbin for helpful discussions and Cotton Seed and Tim Poterba for helping to scale computational analyses. The Sweden Schizophrenia Study was supported by NIMH grant R01 MH077139. The Estonian Biobank was funded by an Estonian Research Council personal research grant (PUT1660). This research was supported by the Russian Science Foundation (17-15-01177); the National Human Genome Research Institute (5U54HG003079); the European Foundation for the Study of Diabetes (EFSD) New Horizons Programme (L.G. and L.K.); an EFSD/Novo Nordisk grant to R.B.P.; the Academy of Finland Center of Excellence in Complex Disease Genetics (grant 312063 to L.G., 312074 to A.P., and 312062 to S.R.); the Academy of Finland (grants 263401 and 267 882 to L.G.; 286500 to A.P.; 265240 and 263278 to J.K.; and 285380 to S.R.); the Sigrid Juselius Foundation (A.P., L.G., and S.R.); the Finnish Foundation for Cardiovascular Research (A.P. and S.R.); the Nordic Information for Action eScience Center (62721 to A.P.); the 7th Research and Innovation Framework Programme (602633 to A.P.) (EUROHEADPAIN); and the Horizon 2020 Research and Innovation Programme (667301 to A.P. [COSYN] and 692145 to S.R. [ePerMed]); Biocentrum Helsinki (S.R.); University of Helsinki HiLIFE Fellow grant (S.R.); and the National Institutes of Health (R01HL113315-01 to A.P. and S.R.). M.J.D. is on the scientific advisory board of Ancestry DNA.

Received: November 10, 2017
Accepted: February 28, 2018
Published: April 26, 2018

Web Resources

Database of Global Administrative Areas, http://www.gadm.org/
Hail, https://hail.is/
LOFTEE, https://github.com/konradjk/loftee
National Institute of Mental Health Genetics Repository, https://www.nimhgenetics.org/
GitHub (scripts from study), https://github.com/armartin/haplotypes
SISu Project, http://www.new.sisuproject.fi
Wikipedia.org (Finnish municipality codes), https://fi.wikipedia.org/wiki/Luettelo_Suomen_kuntanumeroista

References

1. Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Rieder, M.J., Altshuler, D., Shendure, J., et al.; NHLBI Exome Sequencing Project (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature *493*, 216–220.
2. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
3. Rasmussen, M.D., Hubisz, M.J., Gronau, I., and Siepel, A. (2014). Genome-wide inference of ancestral recombination graphs. PLoS Genet. *10*, e1004342.
4. Kiezun, A., Pulit, S.L., Francioli, L.C., van Dijk, F., Swertz, M., Boomsma, D.I., van Duijn, C.M., Slagboom, P.E., van Ommen, G.J.B., Wijmenga, C., et al.; Genome of the Netherlands Consortium (2013). Deleterious alleles in the human genome are on average younger than neutral alleles of the same frequency. PLoS Genet. *9*, e1003301.
5. Zuk, O., Schaffner, S.F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M.J., Neale, B.M., Sunyaev, S.R., and Lander,

E.S. (2014). Searching for missing heritability: designing rare variant association studies. Proc. Natl. Acad. Sci. USA *111*, E455–E464.
6. Henn, B.M., Botigué, L.R., Peischl, S., Dupanloup, I., Lipatov, M., Maples, B.K., Martin, A.R., Musharoff, S., Cann, H., Snyder, M.P., et al. (2016). Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. Proc. Natl. Acad. Sci. USA *113*, E440–E449.
7. Lohmueller, K.E. (2014). The impact of population demography and selection on the genetic architecture of complex traits. PLoS Genet. *10*, e1004379.
8. Mathieson, I., and McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. Nat. Genet. *44*, 243–246.
9. Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H., and Bustamante, C.D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. *5*, e1000695.
10. International HapMap Consortium (2003). The International HapMap Project. Nature *426*, 789–796.
11. Bonnen, P.E., Pe'er, I., Plenge, R.M., Salit, J., Lowe, J.K., Shapero, M.H., Lifton, R.P., Breslow, J.L., Daly, M.J., Reich, D.E., et al. (2006). Evaluating potential for whole-genome studies in Kosrae, an isolated population in Micronesia. Nat. Genet. *38*, 214–217.
12. Sajantila, A., Salem, A.H., Savolainen, P., Bauer, K., Gierig, C., and Pääbo, S. (1996). Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population. Proc. Natl. Acad. Sci. USA *93*, 12035–12039.
13. Peltonen, L., Palotie, A., and Lange, K. (2000). Use of population isolates for mapping complex traits. Nat. Rev. Genet. *1*, 182–190.
14. Palo, J.U., Ulmanen, I., Lukka, M., Ellonen, P., and Sajantila, A. (2009). Genetic markers and population history: Finland revisited. Eur. J. Hum. Genet. *17*, 1336–1346.
15. Wang, S.R., Agarwala, V., Flannick, J., Chiang, C.W.K., Altshuler, D., Hirschhorn, J.N.; and GoT2D Consortium (2014). Simulation of Finnish population history, guided by empirical genetic data, to assess power of rare-variant tests in Finland. Am. J. Hum. Genet. *94*, 710–720.
16. Lim, E.T., Würtz, P., Havulinna, A.S., Palta, P., Tukiainen, T., Rehnström, K., Esko, T., Mägi, R., Inouye, M., Lappalainen, T., et al.; Sequencing Initiative Suomi (SISu) Project (2014). Distribution and medical impact of loss-of-function variants in the Finnish founder population. PLoS Genet. *10*, e1004494.
17. Salmela, E. (2012). Genetic structure in Finland and Sweden: aspects of population history and gene mapping. PhD thesis (University of Helsinki).
18. Günther, T., Malmström, H., Svensson, E.M., Omrak, A., Sánchez-Quinto, F., Kılınç, G.M., Krzewińska, M., Eriksson, G., Fraser, M., Edlund, H., et al. (2018). Population genomics of Mesolithic Scandinavia: Investigating early postglacial migration routes and high-latitude adaptation. PLoS Biol. *16*, e2003703.
19. Poznik, G.D., Xue, Y., Mendez, F.L., Willems, T.F., Massaia, A., Wilson Sayres, M.A., Ayub, Q., McCarthy, S.A., Narechania, A., Kashin, S., et al.; 1000 Genomes Project Consortium (2016). Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. Nat. Genet. *48*, 593–599.
20. Kittles, R.A., Perola, M., Peltonen, L., Bergen, A.W., Aragon, R.A., Virkkunen, M., Linnoila, M., Goldman, D., and Long, J.C. (1998). Dual origins of Finns revealed by Y chromosome haplotype variation. Am. J. Hum. Genet. *62*, 1171–1179.
21. Tallavaara, M., Pesonen, P., and Oinonen, M. (2010). Prehistoric population history in eastern Fennoscandia. J. Archaeol. Sci. *37*, 251–260.
22. Kerminen, S., Havulinna, A.S., Hellenthal, G., Martin, A.R., Sarin, A.-P., Perola, M., Palotie, A., Salomaa, V., Daly, M.J., Ripatti, S., and Pirinen, M. (2017). Fine-scale genetic structure in Finland. G3 (Bethesda) *7*, 3459–3468.
23. Peltonen, L., Jalanko, A., and Varilo, T. (1999). Molecular genetics of the Finnish disease heritage. Hum. Mol. Genet. *8*, 1913–1923.
24. Stoll, G., Pietiläinen, O.P.H., Linder, B., Suvisaari, J., Brosi, C., Hennah, W., Leppä, V., Torniainen, M., Ripatti, S., Ala-Mello, S., et al. (2013). Deletion of TOP3β, a component of FMRP-containing mRNPs, contributes to neurodevelopmental disorders. Nat. Neurosci. *16*, 1228–1237.
25. Lahtinen, A.M., Havulinna, A.S., Jula, A., Salomaa, V., and Kontula, K. (2015). Prevalence and clinical correlates of familial hypercholesterolemia founder mutations in the general population. Atherosclerosis *238*, 64–69.
26. Salmela, E., Lappalainen, T., Fransson, I., Andersen, P.M., Dahlman-Wright, K., Fiebig, A., Sistonen, P., Savontaus, M.-L., Schreiber, S., Kere, J., and Lahermo, P. (2008). Genome-wide analysis of single nucleotide polymorphisms uncovers population structure in Northern Europe. PLoS ONE *3*, e3519.
27. Jakkula, E., Rehnström, K., Varilo, T., Pietiläinen, O.P.H., Paunio, T., Pedersen, N.L., deFaire, U., Järvelin, M.-R., Saharinen, J., Freimer, N., et al. (2008). The genome-wide patterns of variation expose significant substructure in a founder population. Am. J. Hum. Genet. *83*, 787–794.
28. Palamara, P.F., Lencz, T., Darvasi, A., and Pe'er, I. (2012). Length distributions of identity by descent reveal fine-scale demographic history. Am. J. Hum. Genet. *91*, 809–822.
29. Browning, S.R., and Thompson, E.A. (2012). Detecting rare variant associations by identity-by-descent mapping in case-control studies. Genetics *190*, 1521–1531.
30. Ralph, P., and Coop, G. (2013). The geography of recent genetic ancestry across Europe. PLoS Biol. *11*, e1001555.
31. Lawson, D.J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. PLoS Genet. *8*, e1002453.
32. Szpiech, Z.A., Xu, J., Pemberton, T.J., Peng, W., Zöllner, S., Rosenberg, N.A., and Li, J.Z. (2013). Long runs of homozygosity are enriched for deleterious variation. Am. J. Hum. Genet. *93*, 90–102.
33. Joshi, P.K., Esko, T., Mattsson, H., Eklund, N., Gandin, I., Nutile, T., Jackson, A.U., Schurmann, C., Smith, A.V., Zhang, W., et al. (2015). Directional dominance on stature and cognition in diverse human populations. Nature *523*, 459–462.
34. Gamsiz, E.D., Viscidi, E.W., Frederick, A.M., Nagpal, S., Sanders, S.J., Murtha, M.T., Schmidt, M., Triche, E.W., Geschwind, D.H., State, M.W., et al.; Simons Simplex Collection Genetics Consortium (2013). Intellectual disability is associated with increased runs of homozygosity in simplex autism. Am. J. Hum. Genet. *93*, 103–109.
35. Han, E., Carbonetto, P., Curtis, R.E., Wang, Y., Granka, J.M., Byrnes, J., Noto, K., Kermany, A.R., Myres, N.M., Barber, M.J., et al. (2017). Clustering of 770,000 genomes reveals post-colonial population structure of North America. Nat. Commun. *8*, 14238.

36. Baharian, S., Barakatt, M., Gignoux, C.R., Shringarpure, S., Errington, J., Blot, W.J., Bustamante, C.D., Kenny, E.E., Williams, S.M., Aldrich, M.C., and Gravel, S. (2016). The Great Migration and African-American genomic diversity. PLoS Genet. *12*, e1006059.
37. Mathias, R.A., Taub, M.A., Gignoux, C.R., Fu, W., Musharoff, S., O'Connor, T.D., Vergara, C., Torgerson, D.G., Pino-Yanes, M., Shringarpure, S.S., et al.; CAAPA (2016). A continuum of admixture in the Western Hemisphere revealed by the African Diaspora genome. Nat. Commun. *7*, 12522.
38. Zidan, J., Ben-Avraham, D., Carmi, S., Maray, T., Friedman, E., and Atzmon, G. (2015). Genotyping of geographically diverse Druze trios reveals substructure and a recent bottleneck. Eur. J. Hum. Genet. *23*, 1093–1099.
39. Atzmon, G., Hao, L., Pe'er, I., Velez, C., Pearlman, A., Palamara, P.F., Morrow, B., Friedman, E., Oddoux, C., Burns, E., and Ostrer, H. (2010). Abraham's children in the genome era: major Jewish diaspora populations comprise distinct genetic clusters with shared Middle Eastern Ancestry. Am. J. Hum. Genet. *86*, 850–859.
40. Behar, D.M., Metspalu, M., Baran, Y., Kopelman, N.M., Yunusbayev, B., Gladstein, A., Tzur, S., Sahakyan, H., Bahmanimehr, A., Yepiskoposyan, L., et al. (2013). No evidence from genome-wide data of a Khazar origin for the Ashkenazi Jews. Hum. Biol. *85*, 859–900.
41. Campbell, C.L., Palamara, P.F., Dubrovsky, M., Botigué, L.R., Fellous, M., Atzmon, G., Oddoux, C., Pearlman, A., Hao, L., Henn, B.M., et al. (2012). North African Jewish and non-Jewish populations form distinctive, orthogonal clusters. Proc. Natl. Acad. Sci. USA *109*, 13865–13870.
42. Nakatsuka, N., Moorjani, P., Rai, N., Sarkar, B., Tandon, A., Patterson, N., Bhavani, G.S., Girisha, K.M., Mustak, M.S., Srinivasan, S., et al. (2017). The promise of discovering population-specific disease-associated genes in South Asia. Nat. Genet. *49*, 1403–1407.
43. Gauvin, H., Moreau, C., Lefebvre, J.-F., Laprise, C., Vézina, H., Labuda, D., and Roy-Gagnon, M.-H. (2014). Genome-wide patterns of identity-by-descent sharing in the French Canadian founder population. Eur. J. Hum. Genet. *22*, 814–821.
44. Moreno-Estrada, A., Gravel, S., Zakharia, F., McCauley, J.L., Byrnes, J.K., Gignoux, C.R., Ortiz-Tello, P.A., Martínez, R.J., Hedges, D.J., Morris, R.W., et al. (2013). Reconstructing the population genetic history of the Caribbean. PLoS Genet. *9*, e1003925.
45. Gravel, S., Zakharia, F., Moreno-Estrada, A., Byrnes, J.K., Muzzio, M., Rodriguez-Flores, J.L., Kenny, E.E., Gignoux, C.R., Maples, B.K., Guiblet, W., et al.; 1000 Genomes Project (2013). Reconstructing Native American migrations from whole-genome and whole-exome data. PLoS Genet. *9*, e1004023.
46. Homburger, J.R., Moreno-Estrada, A., Gignoux, C.R., Nelson, D., Sanchez, E., Ortiz-Tello, P., Pons-Estel, B.A., Acevedo-Vasquez, E., Miranda, P., Langefeld, C.D., et al. (2015). Genomic insights into the ancestry and demographic history of South America. PLoS Genet. *11*, e1005602.
47. Kong, A., Masson, G., Frigge, M.L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P.I., Ingason, A., Steinberg, S., Rafnar, T., et al. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. Nat. Genet. *40*, 1068–1075.
48. Moorjani, P., Patterson, N., Loh, P.-R., Lipson, M., Kisfali, P., Melegh, B.I., Bonin, M., Kádaši, L., Rieß, O., Berger, B., et al. (2013). Reconstructing Roma history from genome-wide data. PLoS ONE *8*, e58633.
49. Panoutsopoulou, K., Hatzikotoulas, K., Xifara, D.K., Colonna, V., Farmaki, A.-E., Ritchie, G.R.S., Southam, L., Gilly, A., Tachmazidou, I., Fatumo, S., et al. (2014). Genetic characterization of Greek population isolates reveals strong genetic drift at missense and trait-associated variants. Nat. Commun. *5*, 5345.
50. Glodzik, D., Navarro, P., Vitart, V., Hayward, C., McQuillan, R., Wild, S.H., Dunlop, M.G., Rudan, I., Campbell, H., Haley, C., et al. (2013). Inference of identity by descent in population isolates and optimal sequencing studies. Eur. J. Hum. Genet. *21*, 1140–1145.
51. Gilbert, E., Carmi, S., Ennis, S., Wilson, J.F., and Cavalleri, G.L. (2017). Genomic insights into the population structure and history of the Irish Travellers. Sci. Rep. *7*, 42187.
52. Botigué, L.R., Henn, B.M., Gravel, S., Maples, B.K., Gignoux, C.R., Corona, E., Atzmon, G., Burns, E., Ostrer, H., Flores, C., et al. (2013). Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. Proc. Natl. Acad. Sci. USA *110*, 11791–11796.
53. Genome of the Netherlands Consortium (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat. Genet. *46*, 818–825.
54. Fiorito, G., Di Gaetano, C., Guarrera, S., Rosa, F., Feldman, M.W., Piazza, A., and Matullo, G. (2016). The Italian genome reflects the history of Europe and the Mediterranean basin. Eur. J. Hum. Genet. *24*, 1056–1062.
55. Browning, S.R., and Browning, B.L. (2013). Identity-by-descent-based heritability analysis in the Northern Finland Birth Cohort. Hum. Genet. *132*, 129–138.
56. Surakka, I., Horikoshi, M., Mägi, R., Sarin, A.-P., Mahajan, A., Lagou, V., Marullo, L., Ferreira, T., Miraglio, B., Timonen, S., et al.; ENGAGE Consortium (2015). The impact of low-frequency and rare variants on lipid levels. Nat. Genet. *47*, 589–597.
57. Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J.L., Kähler, A.K., Akterin, S., Bergen, S.E., Collins, A.L., Crowley, J.J., Fromer, M., et al.; Multicenter Genetic Studies of Schizophrenia Consortium; Psychosis Endophenotypes International Consortium; and Wellcome Trust Case Control Consortium 2 (2013). Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat. Genet. *45*, 1150–1159.
58. Haller, T., Leitsalu, L., Fischer, K., Nuotio, M.-L., Esko, T., Boomsma, D.I., Kyvik, K.O., Spector, T.D., Perola, M., and Metspalu, A. (2017). MixFit: Methodology for computing ancestry-related genetic scores at the individual level and its application to the Estonian and Finnish population studies. PLoS ONE *12*, e0170325.
59. Rotar, O., Moguchaia, E., Boyarinova, M., Kolesova, E., Khromova, N., Freylikhman, O., Smolina, N., Solntsev, V., Kostareva, A., Konradi, A., and Shlyakhto, E. (2015). Seventy years after the siege of Leningrad: does early life famine still affect cardiovascular risk and aging? J. Hypertens. *33*, 1772–1779.
60. Prasad, R.B., Lessmark, A., Almgren, P., Kovacs, G., Hansson, O., Oskolkov, N., Vitai, M., Ladenvall, C., Kovacs, P., Fadista, J., et al. (2016). Excess maternal transmission of variants in the THADA gene to offspring with type 2 diabetes. Diabetologia *59*, 1702–1713.
61. Rivas, M.A., Graham, D., Sulem, P., Stevens, C., Desch, A.N., Goyette, P., Gudbjartsson, D., Jonsdottir, I., Thorsteinsdottir,

U., Degenhardt, F., et al.; UK IBD Genetics Consortium; and NIDDK IBD Genetics Consortium (2016). A protein-truncating R179X variant in RNF186 confers protection against ulcerative colitis. Nat. Commun. *7*, 12342.

62. Borodulin, K., Vartiainen, E., Peltonen, M., Jousilahti, P., Juolevi, A., Laatikainen, T., Männistö, S., Salomaa, V., Sundvall, J., and Puska, P. (2015). Forty-year trends in cardiovascular risk factors in Finland. Eur. J. Public Health *25*, 539–546.
63. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.
64. Abraham, G., and Inouye, M. (2014). Fast principal component analysis of large-scale genome-wide data. PLoS ONE *9*, e93766.
65. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867–2873.
66. Browning, B.L., and Browning, S.R. (2013). Detecting identity by descent and estimating genotype error rates in sequence data. Am. J. Hum. Genet. *93*, 840–851.
67. Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman, J.M., and Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. Genome Res. *19*, 318–326.
68. Browning, S.R., and Browning, B.L. (2015). Accurate nonparametric estimation of recent effective population size from segments of identity by descent. Am. J. Hum. Genet. *97*, 404–418.
69. Loh, P.-R., Palamara, P.F., and Price, A.L. (2016). Fast and accurate long-range phasing in a UK Biobank cohort. Nat. Genet. *48*, 811–816.
70. Petkova, D., Novembre, J., and Stephens, M. (2016). Visualizing spatial population structure with estimated effective migration surfaces. Nat. Genet. *48*, 94–100.
71. Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., Day, T., Hutnik, K., Royrvik, E.C., Cunliffe, B., Lawson, D.J., et al.; Wellcome Trust Case Control Consortium 2; and International Multiple Sclerosis Genetics Consortium (2015). The fine-scale genetic structure of the British population. Nature *519*, 309–314.
72. Ramachandran, S., Deshpande, O., Roseman, C.C., Rosenberg, N.A., Feldman, M.W., and Cavalli-Sforza, L.L. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc. Natl. Acad. Sci. USA *102*, 15942–15947.
73. Nelis, M., Esko, T., Mägi, R., Zimprich, F., Zimprich, A., Toncheva, D., Karachanak, S., Piskácková, T., Balascák, I., Peltonen, L., et al. (2009). Genetic structure of Europeans: a view from the North-East. PLoS ONE *4*, e5472.
74. Zhang, Q.S., Browning, B.L., and Browning, S.R. (2014). Genome-wide haplotypic testing in a Finnish cohort identifies a novel association with low-density lipoprotein cholesterol. Eur. J. Hum. Genet. *23*, 672–677.
75. Tremblay, M., and Vézina, H. (2000). New estimates of intergenerational time intervals for the calculation of age and origins of mutations. Am. J. Hum. Genet. *66*, 651–658.
76. Sousa, V., and Hey, J. (2013). Understanding the origin of species with genome-scale data: modelling gene flow. Nat. Rev. Genet. *14*, 404–414.
77. Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., et al. (2008). Genes mirror geography within Europe. Nature *456*, 98–101.
78. Belbin, G.M., Odgis, J., Sorokin, E.P., Yee, M.-C., Kohli, S., Glicksberg, B.S., Gignoux, C.R., Wojcik, G.L., Van Vleck, T., Jeff, J.M., et al. (2017). Genetic identification of a common collagen disease in puerto ricans via identity-by-descent mapping in a health system. eLife *6*, 68.
79. Vacic, V., Ozelius, L.J., Clark, L.N., Bar-Shira, A., Gana-Weisz, M., Gurevich, T., Gusev, A., Kedmi, M., Kenny, E.E., Liu, X., et al. (2014). Genome-wide mapping of IBD segments in an Ashkenazi PD cohort identifies associated haplotypes. Hum. Mol. Genet. *23*, 4693–4702.
80. Zaitlen, N., Kraft, P., Patterson, N., Pasaniuc, B., Bhatia, G., Pollack, S., and Price, A.L. (2013). Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. PLoS Genet. *9*, e1003520.
81. Evans, L., Tahmasbi, R., Jones, M., Vrieze, S., Abecasis, G., Das, S., Bjelland, D., deCandia, T., Yang, J., Goddard, M., et al. (2017). Narrow-sense heritability estimation of complex traits using identity-by-descent information. bioRxiv. https://doi.org/10.1101/164848.

The American Journal of Human Genetics, Volume *102*

Supplemental Data

Haplotype Sharing Provides Insights into Fine-Scale Population History and Disease in Finland

Alicia R. Martin, Konrad J. Karczewski, Sini Kerminen, Mitja I. Kurki, Antti-Pekka Sarin, Mykyta Artomov, Johan G. Eriksson, Tõnu Esko, Giulio Genovese, Aki S. Havulinna, Jaakko Kaprio, Alexandra Konradi, László Korányi, Anna Kostareva, Minna Männikkö, Andres Metspalu, Markus Perola, Rashmi B. Prasad, Olli Raitakari, Oxana Rotar, Veikko Salomaa, Leif Groop, Aarno Palotie, Benjamin M. Neale, Samuli Ripatti, Matti Pirinen, and Mark J. Daly

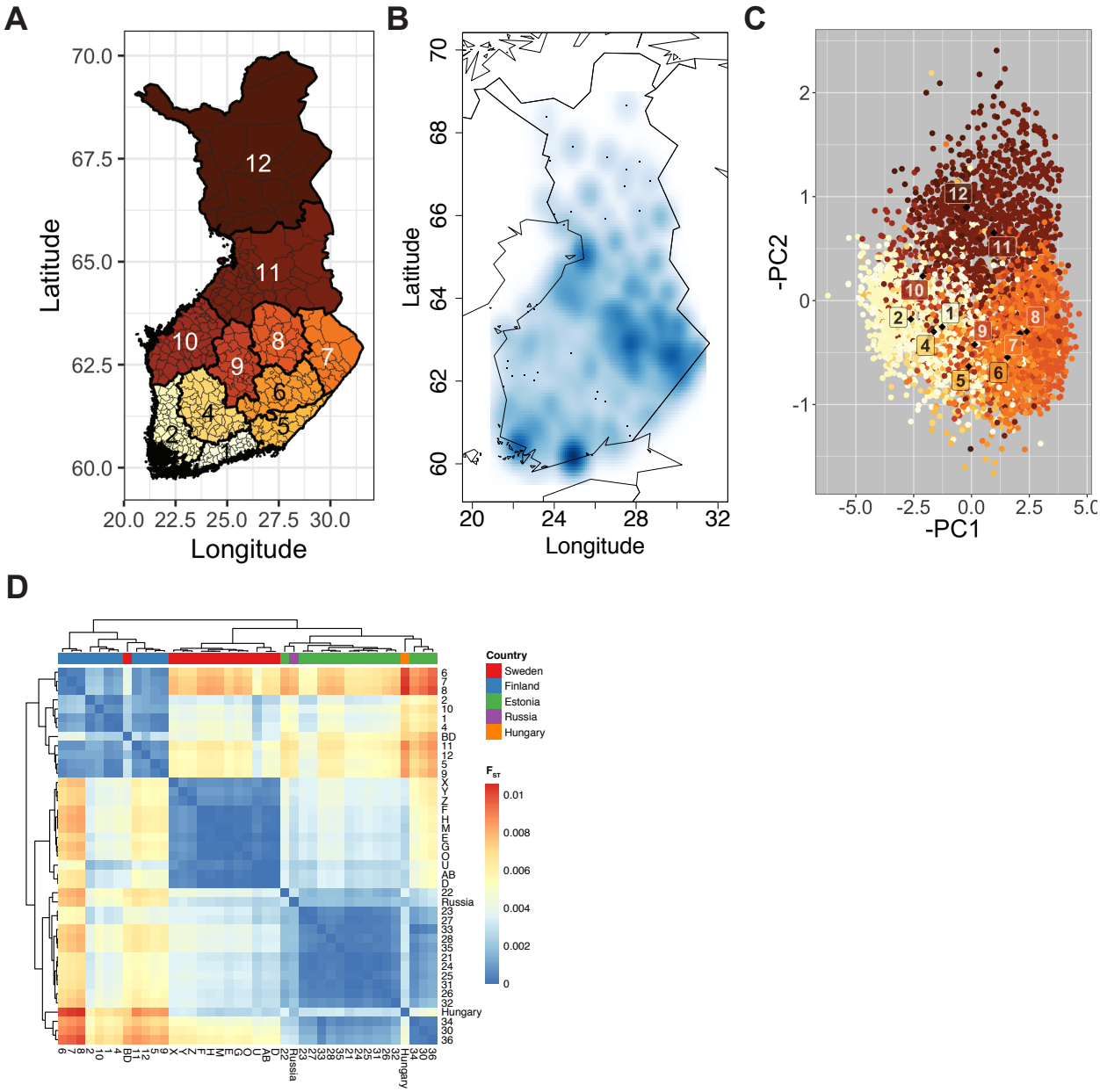

Figure S1 – **Principal components analysis recapitulates geographical birth record data by region**. A) Labeled map of Finland, as in Figure 1A, with colors highlighting regional differences. Notably, forced relocation uprooted many individuals and communities following WWII for example, when Finland ceded its eastern parts (e.g. Karelia) to the Soviet Union and resettled everyone living in the lost areas into the remaining parts of the country[1]. B) Smoothed geographical density map of all Finrisk97 samples with birth record data at the centroids of municipalities (N=5,448). Regional-level birth records not shown. C) PCA positions for all Finrisk97 samples with birth record data. Numbers label the average PC coordinates for all individuals born in a region. Colors are as in A). D) Clustered F_{ST} heat map between individuals born in different regions of Finland, Sweden, Estonia, St Petersburg, Russia, and Hungary. Regions with fewer than 10 individuals were not included. Region labels and names are as in Table S3.

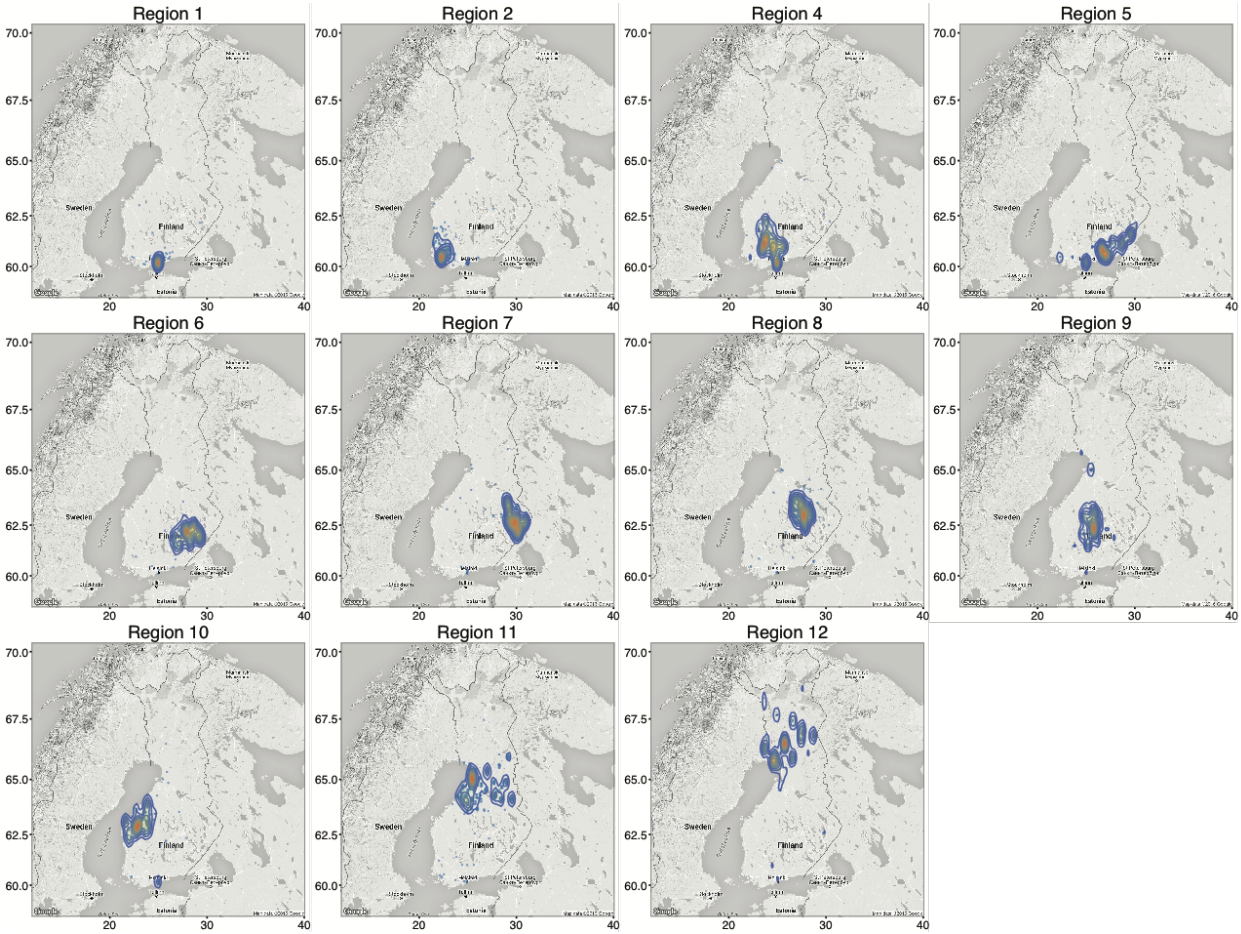

Figure S2 – Birthplace of offspring whose parents are both born in the same region (N=3,132), as indicated by panel titles.

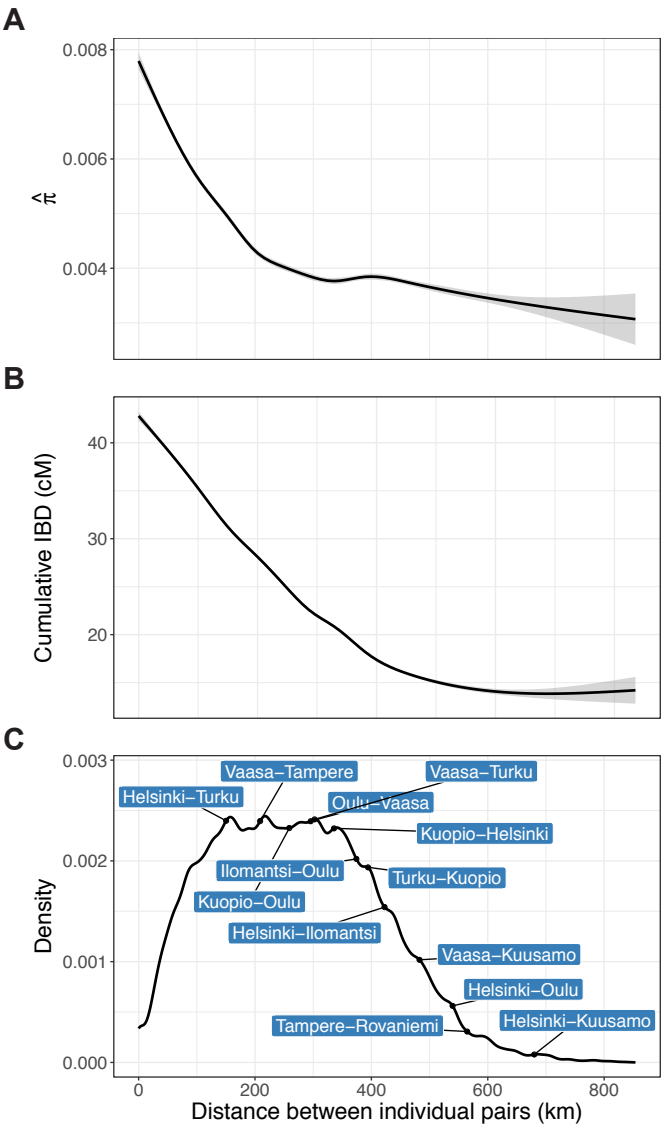

Figure S3 – Geographical distance between pairs of Finnish individuals and genetic sharing. A) Pairwise genetic sharing among unrelated individuals by geographical distance. B) Cumulative IBD sharing (minimum haplotype length ≥ 3 cM) across the genome among unrelated individuals by geographical distance. C) Density of genetic distance between pairs of individuals by geographical distance. The distance between representative city pairs are shown in blue.

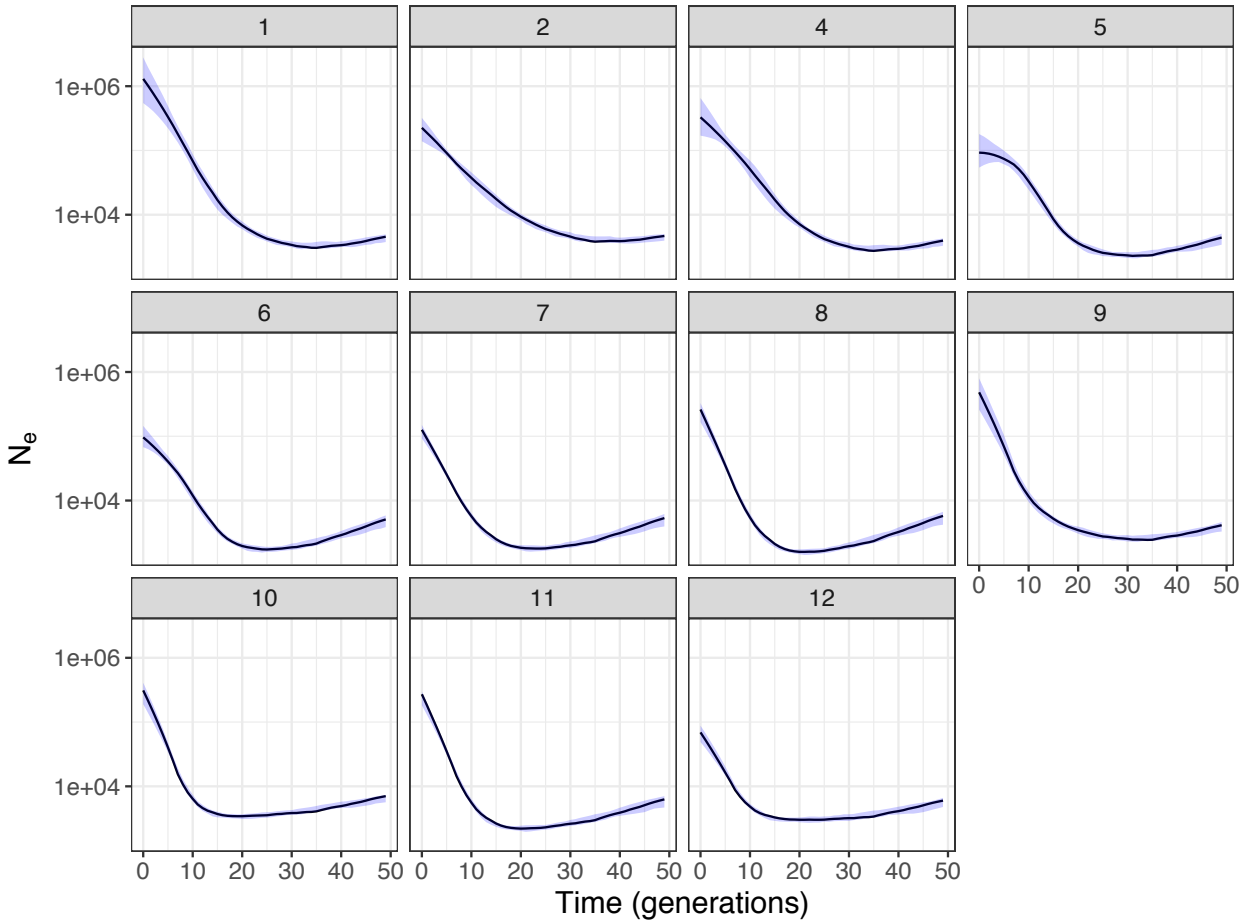

Figure S4 – Effective population size change over time by region of Finland. Number of individuals in each region are: 1: 1,123, 2: 1,078, 4: 378, 5: 224, 6: 304, 7: 1,581, 8: 1,547, 9: 225, 10: 288, 11: 1,697, 12: 184**.**

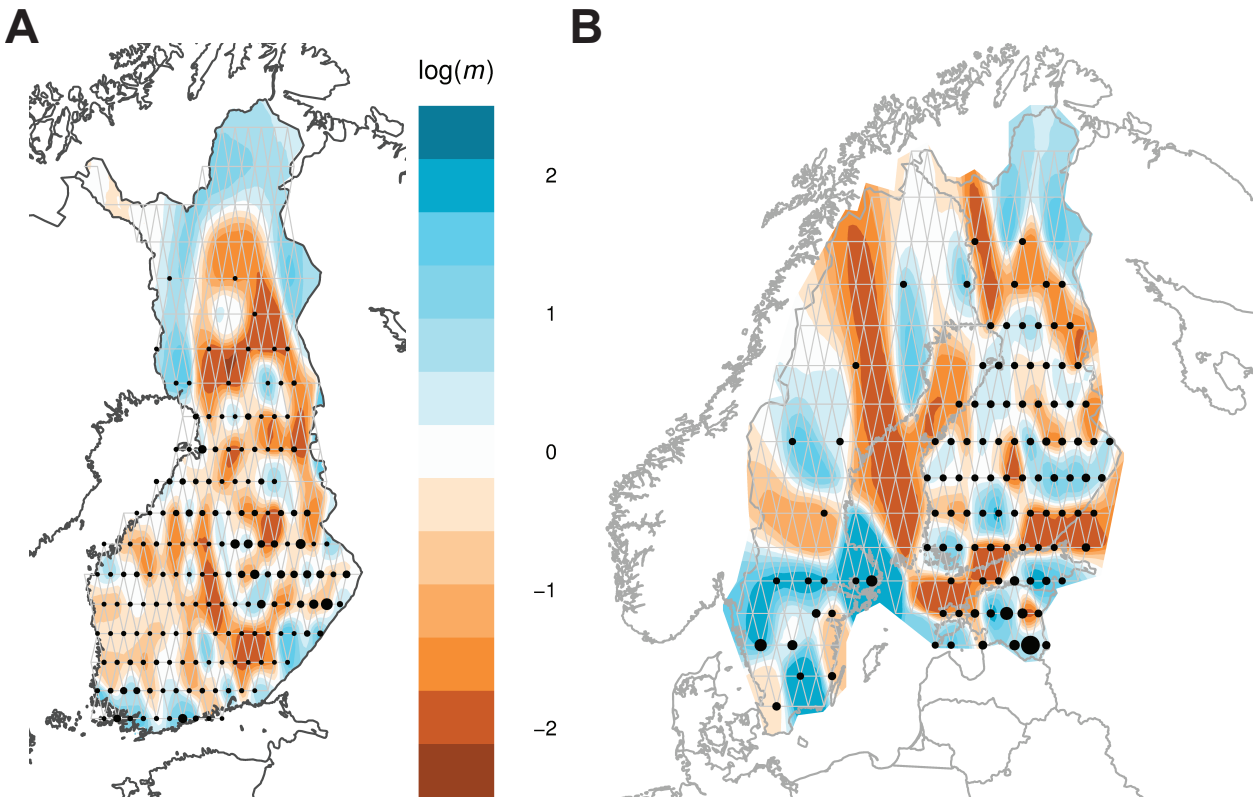

Figure S5 – Deme assignment for EEMS analyses in/near Finland. Black dots at center of demes are proportional to sample size. A) Finland deme assignment from municipality-level birth records. B) Deme assignment in Finland with municipality-level birth records and for region-level birth records in neighboring countries/regions of Sweden, Estonia, and St. Petersburg, Russia.

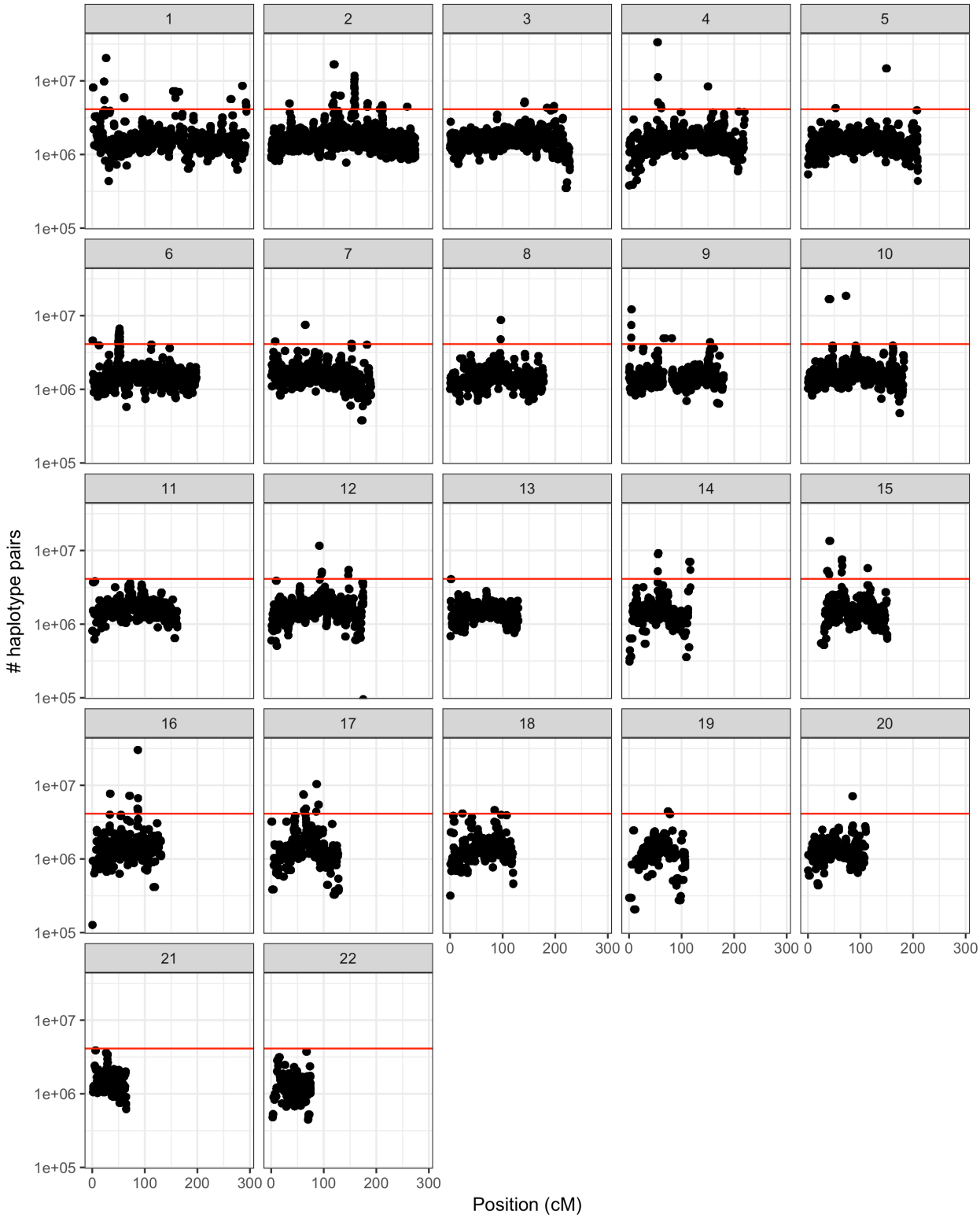

Figure S6 – Haplotype sharing rate genome-wide by chromosome. At each best-guess genotype used to call haplotypes, we quantified the number of pairs of individuals who shared haplotypes. Included individuals are unrelated and have corresponding

exome sequencing data (N=9,363). Red line indicates the mean sharing plus 3 * standard deviation. Total possible number of pairs is $\binom{N}{2}$ = 43,828,203.

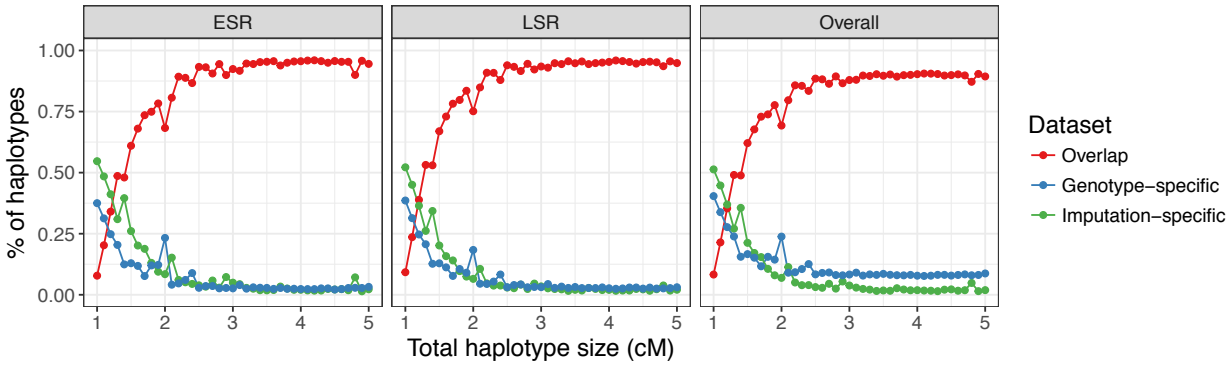

Figure S7 – Comparison of pairwise haplotype sharing for the same samples when using high quality imputed vs genotyped sites. Best guess hard call genotypes passed a filter of INFO > 0.99. Haplotypes were designated as overlapping (maximum start of either haplotype was less than the minimum end of either haplotype) if the overlapping region was at least 50% of the union length across both haplotypes (i.e. minimum end – maximum start ≥ 0.5 * (maximum end – minimum start)). The first two panels indicate individuals genetically and geographically corresponding to the Early Settlement Region (ESR) and Late Settlement Region (LSR), respectively, as in **Figure 1**. The last panel (Overall) includes all individuals regardless of birthplace, and likely has lower overlap rates because of heterogeneity in pairwise geography.

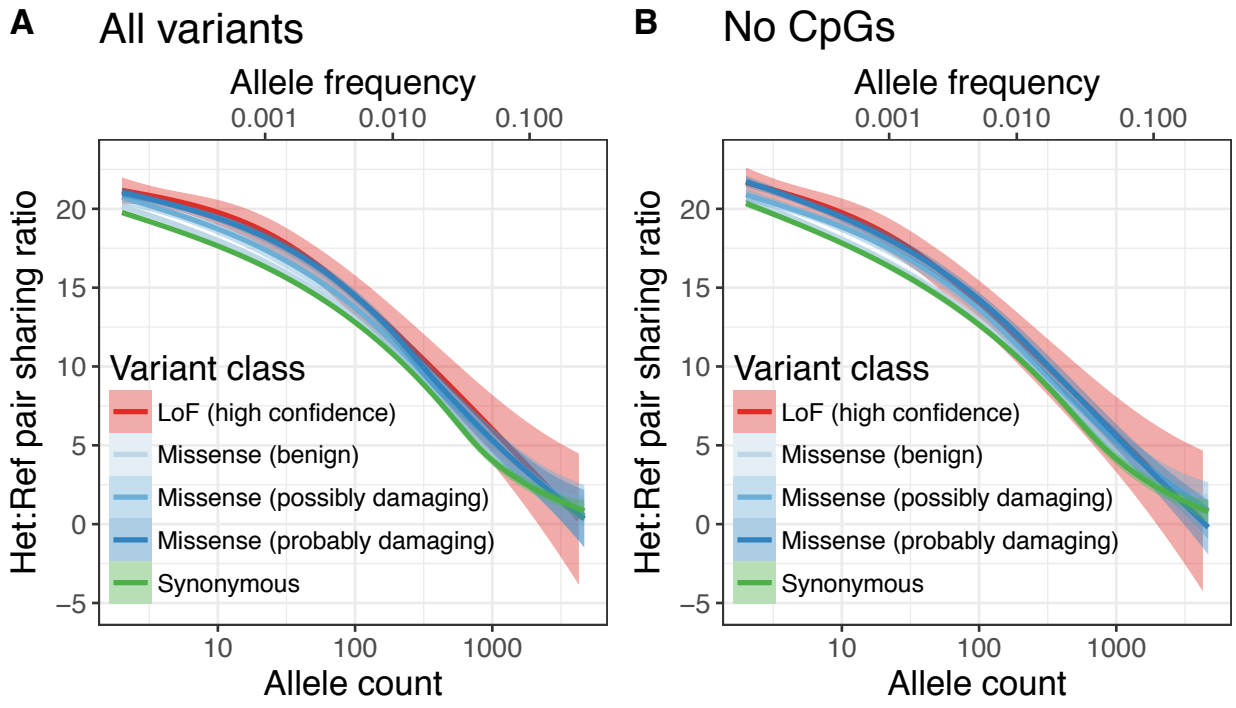

Figure S8 – Slight excess of haplotype sharing at non-CpG sites supports a reduction of haplotype sharing at highly mutable sites. A) Enrichment of haplotype sharing across all variants. B) Enrichment of haplotype sharing across non-CpG variants.

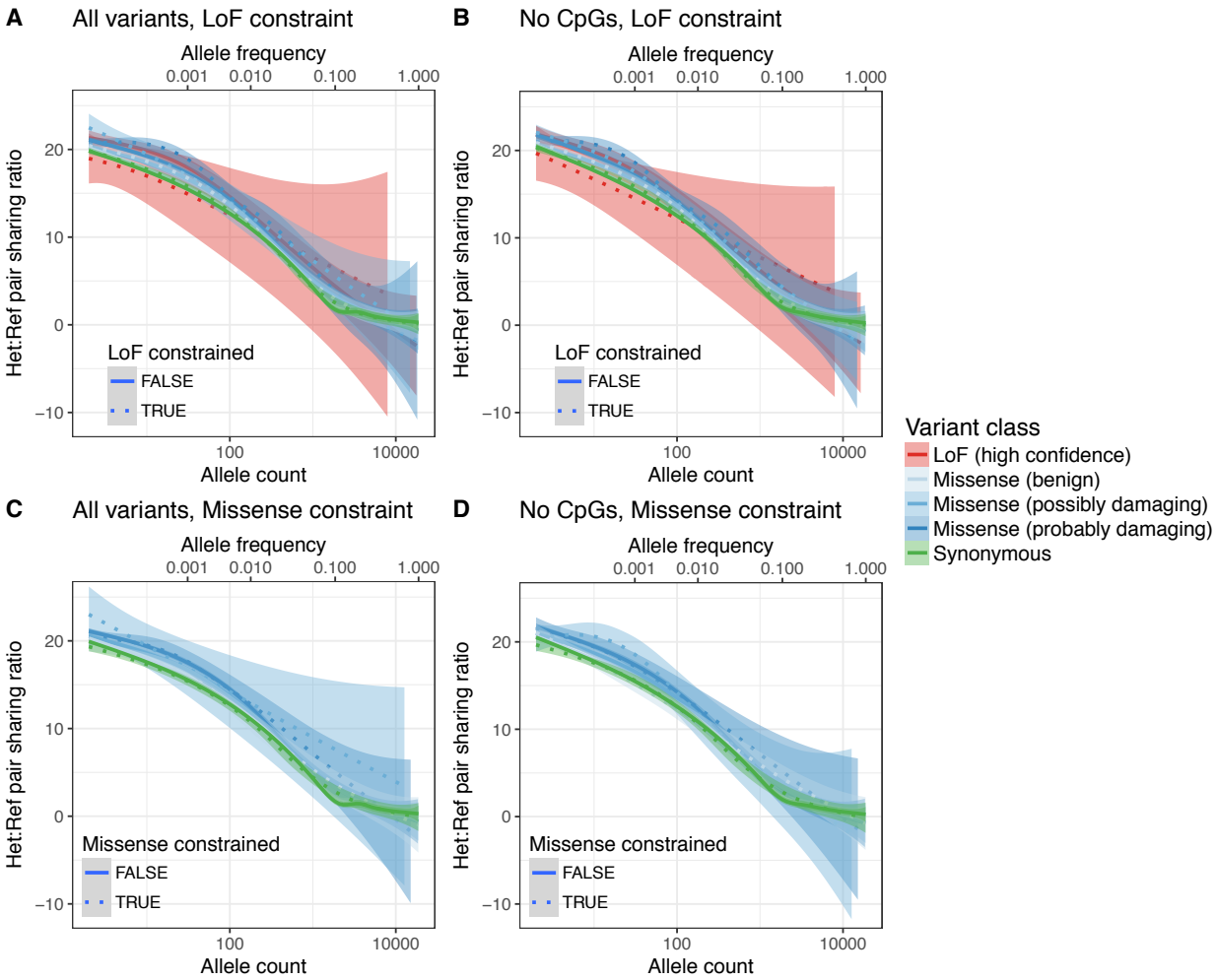

Figure S9 – Haplotype sharing rates are similar across variant classes in missense and loss-of-function (LoF) constrained regions. As calculated in Lek et al, missense constraint indicates regions depleted of missense variation, and LoF constrained regions indicate regions depleted of LoF variation[2,3]. Haplotype sharing rates are similar across different classes of variants when comparing: A) all variants inside and outside LoF constrained regions, B) non-CpG sites inside and outside LoF constrained regions, C) all variants inside and outside missense constrained regions, and D) non-CpG sites inside and outside missense constrained regions.

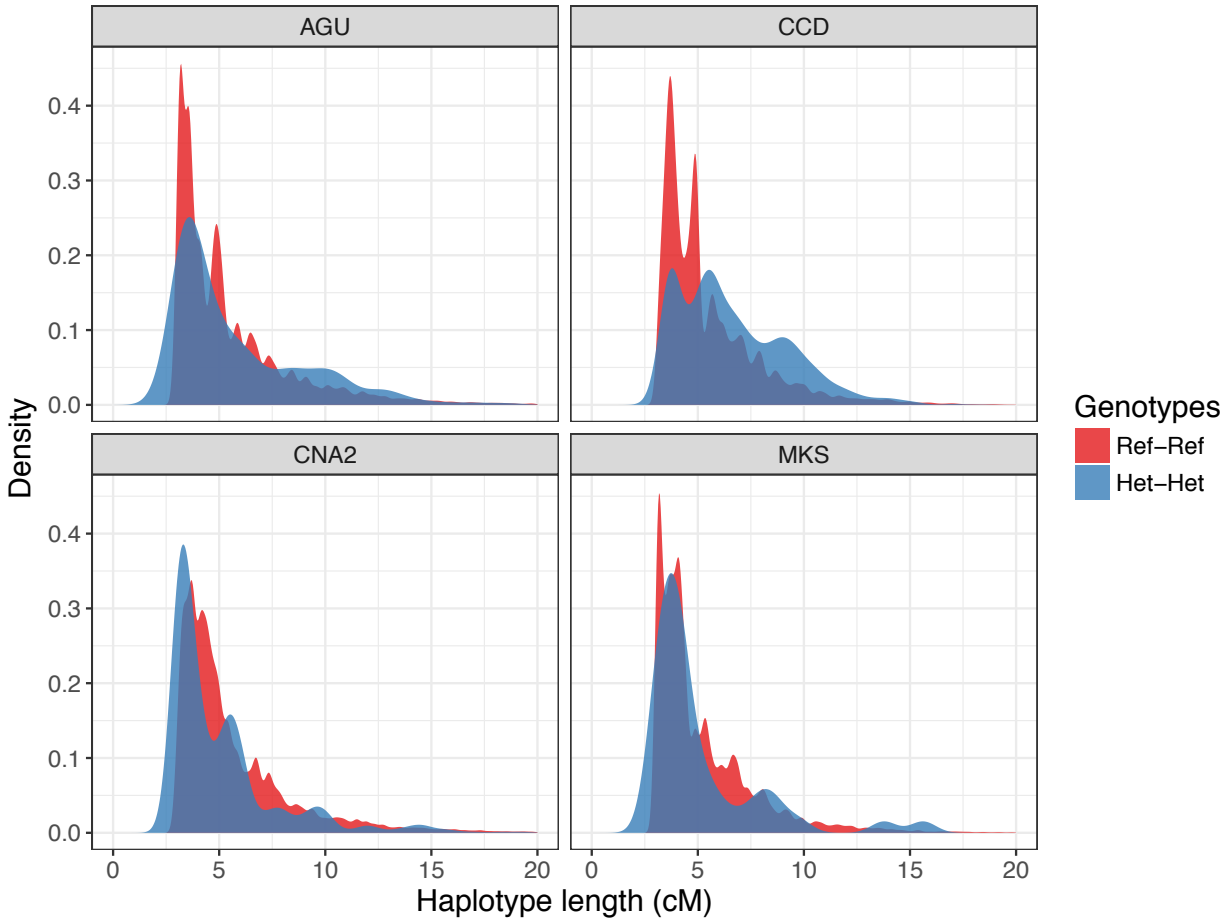

Figure S10 – Haplotype lengths among carrier versus homozygous reference pairs for variants contributing to four FinDis diseases. Variants and diseases correspond to those shown in Figure 5 and Table 1, with diseases as follows: AGU = Aspartylglucosaminuria, CNA2 = Cornea plana 2, CCD = Congenital chloride diarrhea, and MKS = Meckel syndrome.

Table S1 – **Birth record data by cohort**. Municipality-level birth records were available for FR97, regional-level birth records were available for FR07 for this study.

Project/Array	**FR07**	**FR97**
ENGAGE	0	3969
FIN610K	634	458
MIGen	0	110
FINRISK, CoreEX	3065	0
PredictCVD, SUMMIT	243	911
TOTAL (N=9,390)	3942	5448

Table S2 – **Finnish sample genotyping summaries**. Note that some FINRISK samples with birth records have been included as controls for multiple different projects.

Population	**Array**	**Project name**	**SNPs genotyped**	**Sample size**
Finland	Affymetrix 6.0	MIGen	666,979	339
Finland	Illumina 370k	NFBC	324,674	5,363
Finland	Illumina 610k	Corogene, GenMets	535,787	6,240
Finland	Illumina 670k	HBCS, YFS, FTC	521,500	6,492
Finland	Illumina CoreExome	FINRISK, CoreEX	322,929	10,641
Finland	Illumina CoreExome	ENGAGE	342,869	11,639
Finland	Illumina OmniExpress	PredictCVD, SUMMIT	606,310	2,542
Sweden	Illumina OmniExpress	Sw5	733,202	4,465
Sweden	Illumina OmniExpress	Sw6	733,202	3,873
Hungary	Illumina OmniExpressExome	HTB	943,987	506
Estonia	Illumina OmniExpress	EGCUT	710,831	6,946
Russia	Illumina GlobalScreeningArray	RussiaSiege	633,183	262
TOTAL				59,309

Table S3 – Exome sequencing data included in haplotype analyses. Cohorts are ordered by number of individuals contributing to this study. Full descriptions of each cohort are in supplementary note.

Cohort name	**Number of individuals included**
FINRISK_population_cohort	7014
IBD_FINRISK	845
NFBC	525
Health 2000	271
FINRISK_AD	238
Fusion	214
UK10K	68
Migraine	57
METSIM	45
Eufam	42
NFID	30
Twins_AD	25
ADGEN	9

IBD	4
EPILEPSY_EPI25	1
Botnia_T2D	1

Table S4 – Region names by country in Finland, Sweden, and Estonia.

Country	Code	Name
Finland	1	Southern Finland
Finland	2	Southwestern Finland
Finland	3	Åland
Finland	4	Tavastia
Finland	5	Southern Karelia
Finland	6	Southern Savonia
Finland	7	North Karelia
Finland	8	Northern Savonia
Finland	9	Central Finland
Finland	10	Ostrobothnia
Finland	11	Northern Ostrobothnia
Finland	12	Lapland
Sweden	AB	Stockholm
Sweden	AC	Västerbotten
Sweden	BD	Norrbotten
Sweden	C	Uppsala
Sweden	D	Södermanland
Sweden	E	Östergötland
Sweden	F	Jönköping
Sweden	G	Kronoberg
Sweden	H	Kalmar
Sweden	I	Gotland
Sweden	K	Blekinge
Sweden	M	Skåne
Sweden	N	Halland
Sweden	O	Västra Götaland
Sweden	S	Värmland
Sweden	T	Orebro
Sweden	U	Västmanland
Sweden	W	Dalarna
Sweden	X	Gävleborg
Sweden	Y	Västernorrland
Sweden	Z	Jämtland
Estonia	21	Harju
Estonia	22	Hiiu
Estonia	23	Ida-Viru
Estonia	24	Järva
Estonia	25	Jõgeva
Estonia	26	Lääne

Estonia	27	Lääne-Viru
Estonia	28	Pärnu
Estonia	29	Peipsi
Estonia	30	Põlva
Estonia	31	Rapla
Estonia	32	Saare
Estonia	33	Tartu
Estonia	34	Valga
Estonia	35	Viljandi
Estonia	36	Võru

Table S5 – Haplotype sharing rates at Finnish heritage disease (FinDis) variants. FinDis consists of 36 monogenic diseases that are enriched in the Finnish bottleneck. Starting with a list of 50 autosomal variants that are known to be major or minor causes of these diseases, 40 of these variants were polymorphic and in region with high quality haplotype calls. The reference pair ratio is $\frac{\text{\# hom ref pairs sharing a haplotype}}{\text{total \# hom ref pairs}}$, and the carrier pair ratio is $\frac{\text{\# het pairs sharing a haplotype}}{\text{total \# het pairs}}$. The haplotype enrichment is the carrier pair ratio / reference pair ratio.

References

1. Haukka, J., Suvisaari, J., Sarvimäki, M., and Martikainen, P. (2017). The Impact of Forced Migration on Mortality. Epidemiology *28*, 587–593.
2. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
3. Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnström, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of de novo mutation in human disease. Nat Genet *46*, 944–950.

Supplementary Note – Sequencing Initiative Suomi (SISU) cohort descriptions of exome sequencing data included in this study

FINRISK_population_cohort

The FINRISK cohorts comprise the respondents of representative, cross-sectional population surveys that are carried out every 5 years since 1972, to assess the risk factors of chronic diseases (e.g. CVD, diabetes, obesity, cancer) and health behavior in the working age population, in 3-5 large study areas of Finland (Borodulin et al. 2015). DNA samples have been collected in the following survey years: 1987, 1992, 1997, 2002, 2007, and 2012. The cohort sizes are 6000-8800 per survey.

The cohorts have been followed up for disease end-points using annual record linkage with the Finnish National Hospital Discharge Register, the National Causes-of-Death Register and the National Drug Reimbursement Register. The samples sequenced for the current study were enriched for individuals with Northern and Eastern Finnish ancestry.

Borodulin, K. et al., 2015. Forty-year trends in cardiovascular risk factors in Finland. European Journal of Public Health, 25(3), pp.539–546.

Read more at www.nationalbiobanks.fi/index.php/studies2/7-finrisk.

IBD_FINRISK

These FINRISK cohorts have been followed up for IBD and other disease end-points using annual record linkage with the Finnish National Hospital Discharge Register, the National Causes-of-Death Register and the National Drug Reimbursement Register. Controls were chosen to have high polygenic risk score for IBD without IBD diagnosis.

Borodulin, K. et al., 2015. Forty-year trends in cardiovascular risk factors in Finland. European Journal of Public Health, 25(3), pp.539–546.

NFBC

NFBC1966 is a birth cohort from two northern provinces of Oulu and Lapland.
Mothers expected to give birth in the Oulu and Lapland in 1966 were invited to participate in the study, which was originally focused on factors affecting pre-term birth, low birth weight, and subsequent morbidity. The DNA was extracted from a blood sample drawn in 31-year clinical examination.

We thank the late professor Paula Rantakallio (launch of NFBC1966), the participants in the 31yrs study and the NFBC project center.

Järvelin, M.-R. et al., 2004. Early life factors and blood pressure at age 31 years in the 1966 northern Finland birth cohort. Hypertension (Dallas, TX: 1979), 44(6), pp.838–46.

Sabatti, C. et al., 2009. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nature genetics, 41(1), pp.35–46.

Health 2000

Health 2000 Survey, a comprehensive combination of health interview and health examination survey, was carried out in 2000-2001. The study was based on a nationally representative sample of 8028 persons aged 30 and over living in the mainland Finland. In addition a sample of 1894 persons aged 18-29 and a sample of 1260 survivors from the Mini-Finland Health Examination Survey, were included in the data. The Mini-Finland Health Examination Survey, which also was representative of the Finnish population, was carried out in 1978-1980 by The Social Insurance Institution. The main aim of the Health 2000 Survey was to obtain information on the most important public health problems in working-aged and the aged population, their causes and treatment as well as on the population's functional capacity and working capacity.

Read more at: www.nationalbiobanks.fi/index.php/studies2/8-health2000.

FINRISK_AD

The FINRISK cohorts comprise the respondents of representative, cross-sectional population surveys that are carried out every 5 years since 1972, to assess the risk factors of chronic diseases (e.g. CVD, diabetes, obesity, cancer) and health behavior in the working age population, in 3-5 large study areas of Finland (Borodulin et al. 2015). DNA samples have been collected in the following survey years: 1987, 1992, 1997, 2002, 2007, and 2012. The cohort sizes are 6000-8800 per survey.

The cohorts have been followed up for Alzheimer using annual record linkage with the Finnish National Hospital Discharge Register, the National Causes-of-Death Register and the National Drug Reimbursement Register and cases were selected as described in (Tynkkynen et al. 2017)

Borodulin, K. et al., 2015. Forty-year trends in cardiovascular risk factors in Finland. European Journal of Public Health, 25(3), pp.539–546.

Tynkkynen, J. et al., 2017. High-sensitivity cardiac troponin I and NT-proBNP as predictors of incident dementia and Alzheimer's disease: the FINRISK Study. Journal of neurology, 264(3), pp.503–511.

Fusion

The Finland-United States Investigation of NIDDM Genetics (FUSION) dataset is collected for localizing and identifying genetic variants that predispose to type 2 diabetes mellitus (T2D) or are responsible for variability in diabetes-related quantitative traits. The FUSION study sample includes approximately 800 families ascertained for sibling pairs affected with type 2 diabetes, including also parents, unaffected siblings, spouses and children in some cases; ~200 unrelated individuals with normal glucose tolerance at ages 65 and 70 years, with their spouses and children in some cases; and ~8400 mostly unrelated individuals including ~1700 type 2 diabetics selected from the D2D 2004, Finrisk 1987, Finrisk 2002, Health 2000, Action LADA, and Savitaipale Diabetes studies.

Read more at www.nationalbiobanks.fi/index.php/studies2/18-fusion.

UK10K

These Finnish samples have been collected from a population cohort using national registers. Three Finnish cohorts were included in the study, including Kuusamo schizophrenia cohort, a

non-Kuusamo schizophrenia cohort, and autism spectrum disorder (ASD) samples. The entire schizophrenia sample collection consists of 2756 individuals from 458 families of whom 931 are diagnosed with schizophrenia spectrum disorder, each family having at least two affected siblings. 170 families originate from an internal isolate (Kuusamo) with a three-fold life time risk for the trait. The genealogy of the internal isolate is well documented and the individuals form a "megapedigree" reaching to the 17th Century. Families outside Kuusamo (n=288) all had at least two affected siblings. All schizophrenia diagnoses are based on DSM-IV.

The Finnish ASD samples are a nationwide collection. These samples have been collected from Central Hospitals across Finland in collaboration with the University of Helsinki. The samples consist of individuals with a diagnosis of autistic disorder or Asperger syndrome from 36 families with at least two affected individuals. Of these individuals, 16 can be genealogically connected to form two large pedigrees originating from Central Finland, suggesting possible genetic risk factors shared identical by descent within the pedigrees. All diagnoses are based on ICD-10 and DSM-IV diagnostic criteria for ASDs.

https://www.uk10k.org/

Migraine

The Finnish Migraine Family Study sample consists of migraine patients visiting headache clinics, from which extensive questionnaire data for headache and co-morbid disorders has been collected.
Read more at www.nationalbiobanks.fi/index.php/studies2/20-migraine-family-study.

Freilinger, T. et al., 2012. Genome-wide association analysis identifies susceptibility loci for migraine without aura. Nature genetics, 44(7), pp.777–82.

METSIM

The cross-sectional METSIM (METabolic Syndrome In Men) Study includes 10,197 men, aged from 45 to 73 years, randomly selected from the population register of the Kuopio town, Eastern Finland, and examined in 2005-2010. The aim of the study was to investigate genetic and non-genetic factors associated with the risk of type 2 diabetes (T2D), cardiovascular disease (CVD), and insulin resistance –related traits in a cross-sectional and longitudinal setting.

Read more at www.nationalbiobanks.fi/index.php/studies2/10-metsim

Eufam

EUFAM (European Study of Familial Dyslipidemias) study is a project aiming to reveal the molecular and genetic basis of familial combined hyperlipidemia (FCHL) and of familial low high-density cholesterol (HDL-C). The study cohort comprises of over 1500 family members from 140 Finnish families with premature coronary heart disease and with either FCHL or familial low HDL-C.

NFID

From January 2013 subjects for the NFID (Northern Finland Intellectual Disability) Project have been recruited from the Northern Ostrobothnia Hospital District Center for Intellectual Disability Care and from the Department of Clinical Genetics of Oulu University Hospital. In January 2016 the recruitment was expanded to include all the pediatric neurology units and the centers for intellectual disability care in the special responsibility area of Oulu University Hospital. Subjects

of all ages with either intellectual disability or pervasive and specific developmental disorders (ICD-10 codes F70-79 and F80-89, respectively) of unknown etiology were included. Individuals with copy number variations of unknown clinical significance or highly variable phenotype were also included in order to uncover possible other etiologic factors of genetic etiology. Subjects were identified through hospital records and invited by a letter to take part in the study. In addition, they were recruited during routine visits to any of the study centers. All research subjects and or their legal guardians provided a written informed consent to participate in the study. DNA samples of the participants were extracted primarily from peripheral blood. In few sporadic cases where a blood sample could not be obtained, DNA was extracted from saliva. The ethical committees of the Northern Ostrobothnia Hospital District and the Hospital District of Helsinki and Uusimaa reviewed and approved the study.

Kurki et al. Genetic architecture of intellectual disability in high-risk population sub-isolate of Northern Finland. Manuscript to be submitted soon.

Twins_AD

The Finnish Twin Cohort was first established in 1974 to investigate genetic and environmental risk factors for chronic disorders. Twins and their families have been ascertained in three stages from the Central Population Register in 1974 (older like-sexed pairs), 1987 (multiple births 1968-1987) and 1995 (opposite-sex pairs 1938-1957). There are a total of 12,966 MZ and DZ twin pairs (25,932 individuals) with both members currently alive and excluding individuals who refused to participate in studies. Over 15000 DNA samples have been collected in this study, and serum and other biological samples are available from several sub-studies as well. AD cases were identified from Finnish Twin Cohort (Kaprio 2013) study by using combination of Finnish cause of death registries and TELE/TICS interviews (Vuoksimaa et al. 2016).

Read more at www.nationalbiobanks.fi/index.php/studies2/30-finnish-twin-cohort.

Kaprio, J., 2013. The Finnish Twin Cohort Study: An Update. Twin Research and Human Genetics, 16(1), pp.157–162.

Vuoksimaa, E. et al., 2016. Middle age self-report risk score predicts cognitive functioning and dementia in 20-40 years. Alzheimer's & dementia (Amsterdam, Netherlands), 4, pp.118–125.

ADGEN

The ADGEN cohort has been collected for a study focusing on the identification of novel Alzheimer's disease (AD)-associated genes and pathways using existing clinical cohorts from Eastern and Northern Finland. ADGEN is a clinic based collection of AD patients examined in the Department of Neurology in Kuopio University Hospital and Department of Neurology in Oulu University Hospita. All patients were diagnosed with probable AD according to the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) criteria.

Read more at www.nationalbiobanks.fi/index.php/studies2/34-adgen-study.

IBD

Finnish inflammatory bowel disease (IBD) patients were recruited from Helsinki University Hospital and described in more detail in the references below.

Halme, L. et al., 2002. Familial and Sporadic Inflammatory Bowel Disease: Comparison of

Clinical Features and Serological Markers in a Genetically Homogeneous Population. Scandinavian Journal of Gastroenterology, 37(6), pp.692–698.

Heliö, T. et al., 2003. CARD15/NOD2 gene variants are associated with familially occurring and complicated forms of Crohn's disease. Gut, 52(4), pp.558–62.

Rivas, M.A. et al., 2016. A protein-truncating R179X variant in RNF186 confers protection against ulcerative colitis. Nature Communications, 7.

EPILEPSY_EPI25

Finnish epilepsy patients sequenced as part of NHGRI funded EPI25-project.

Read more at: http://epilepsygenetics.net/2017/01/10/year-1-of-the-epi25-collaborative-the-first-6000-epilepsy-exomes/

Botnia_T2D

The aims of Botnia cohort has been collected from the western coast of Finland in the Gulf of Bothnia for four different studies studying type 2 diabetes. The Botnia Study, started in 1990, is one of the largest diabetes family studies in the world. The initial family based Botnia study comprised of 11000 individuals as well as a prospective 10-year follow-up of 2800 individuals. The Botnia study also includes a population based study of 5200 individuals aged 18-75 with an ongoing 6-year follow-up study. A project aiming to cover all diabetic patients in the region has also been launched and includes at the moment more than 4000 individuals. The study includes individuals from about 4000 families (about 1000 independent trios) and extensive phenotype information is available for all study participants.

Read more at www.nationalbiobanks.fi/index.php/studies2/13-the-bosnia-study.