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

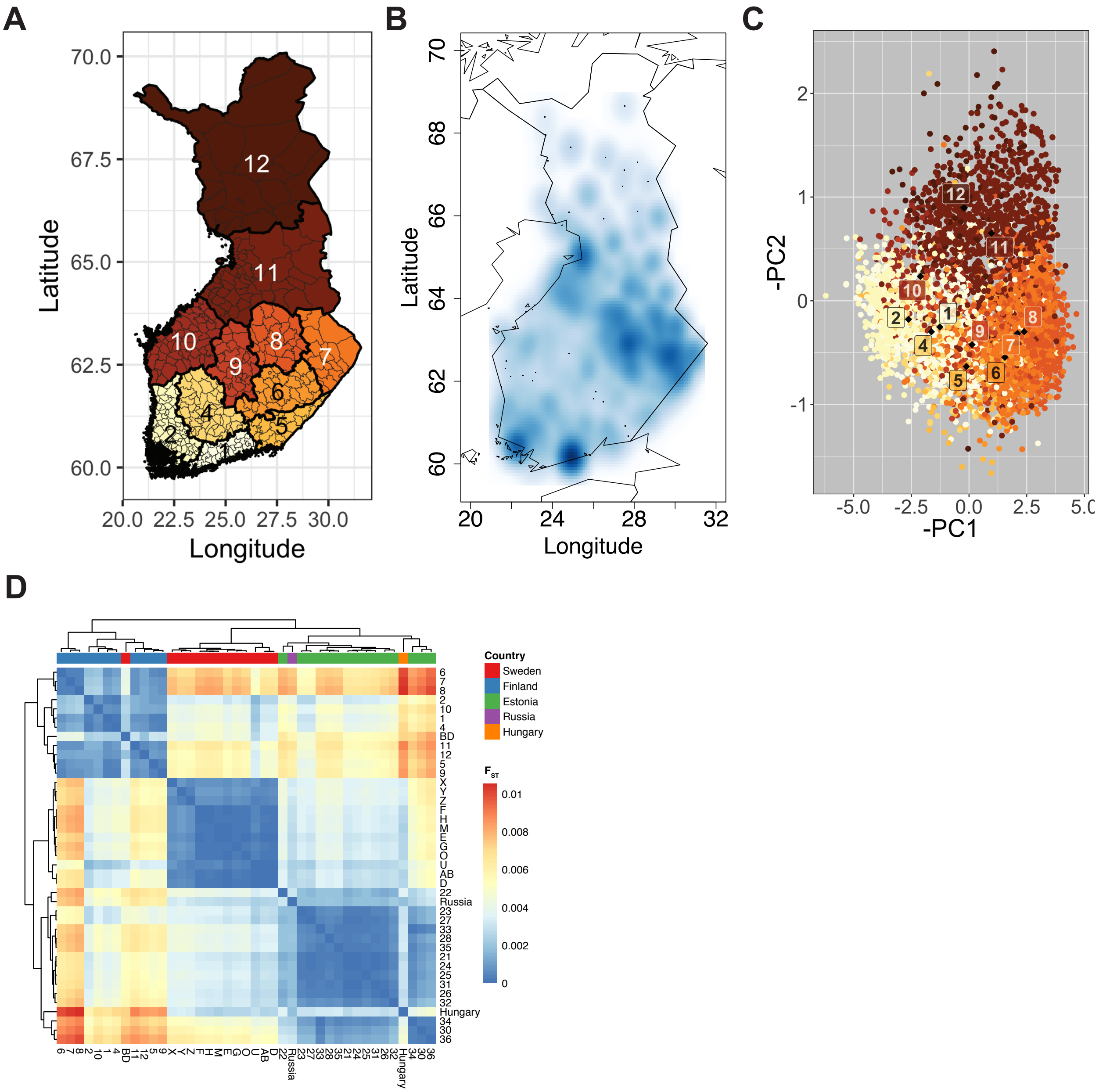

Figure S1 – **Principal components analysis recapitulates geographical birth record data by region**. A) Labeled map of Finland, as in Figure 1A, with colors highlighting regional differences. Notably, forced relocation uprooted many individuals and communities following WWII for example, when Finland 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.

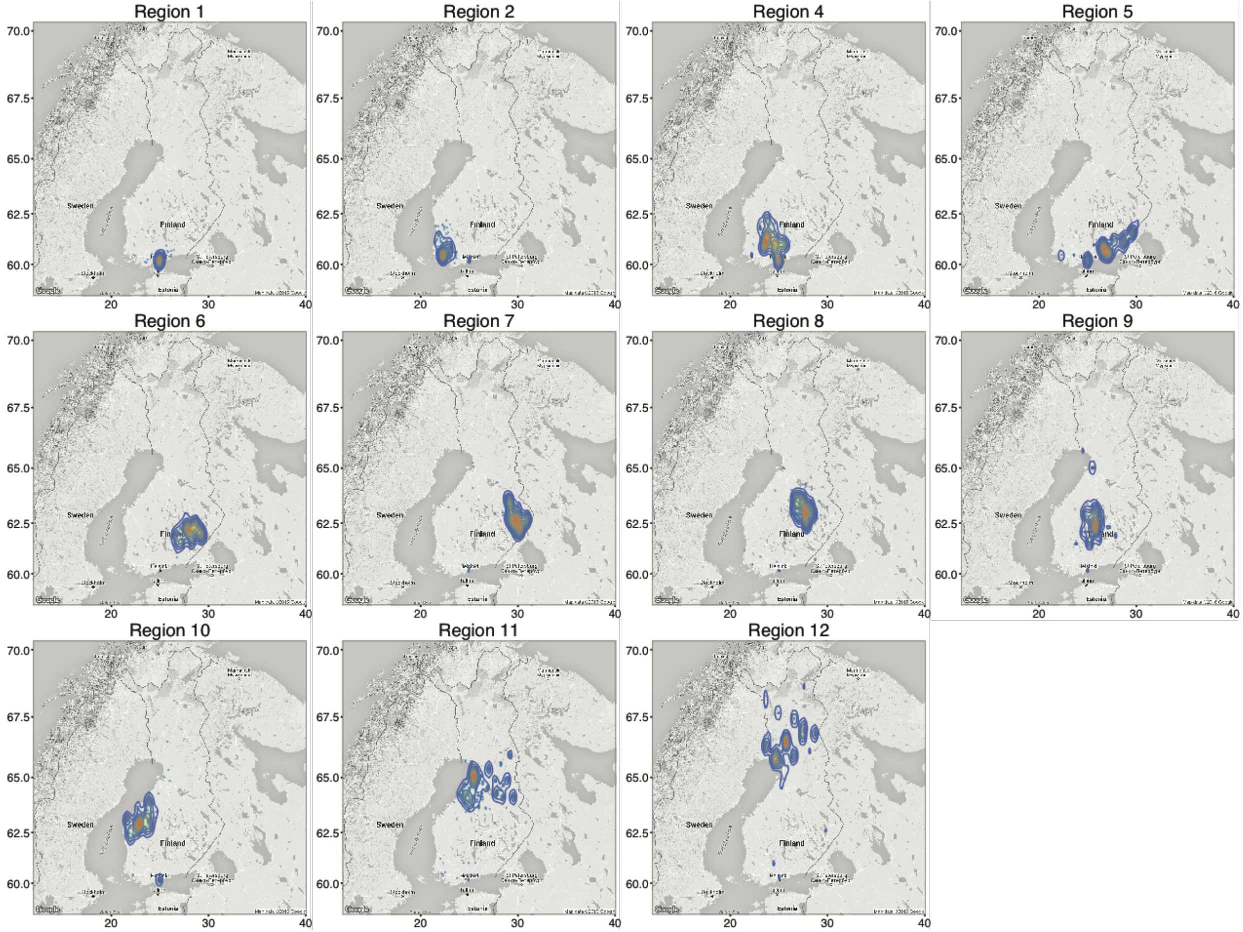

Figure S2 – Birthplace of offspring whose parents are both born in the same region (N=3,132), as indicated by panel titles.

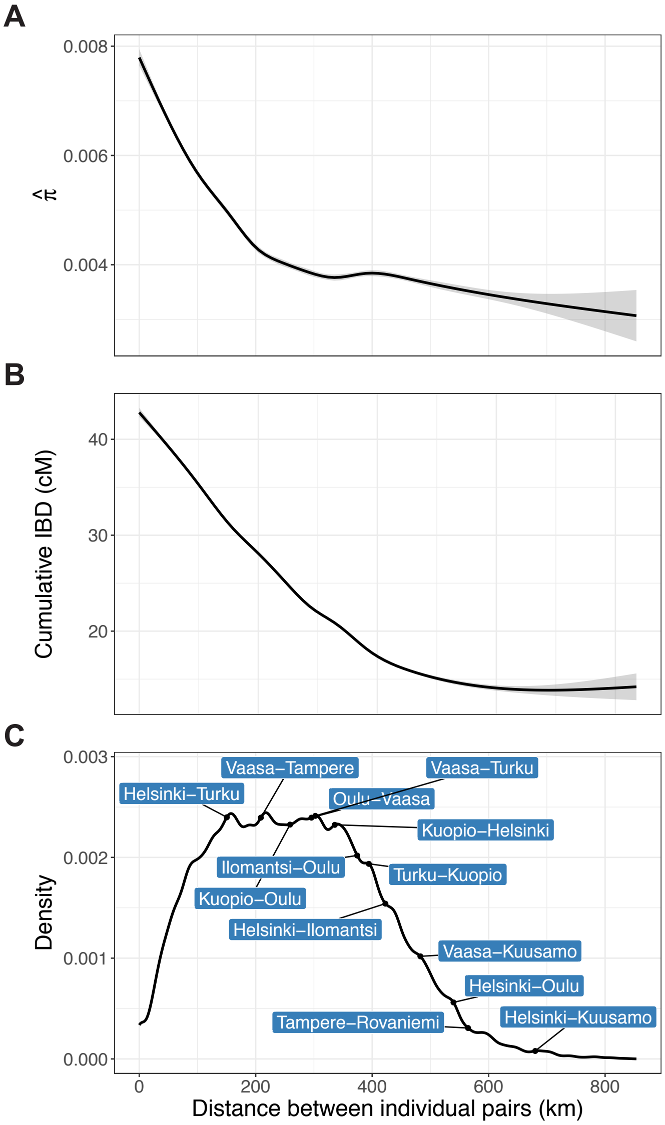

Figure S3 – Geographical distance between pairs of Finnish individuals and genetic sharing. A) Pairwise genetic sharing among unrelated individuals by geographical distance. B) Cumulative IBD sharing (minimum haplotype length ≥ 3 cM) across the genome among unrelated individuals by geographical distance.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

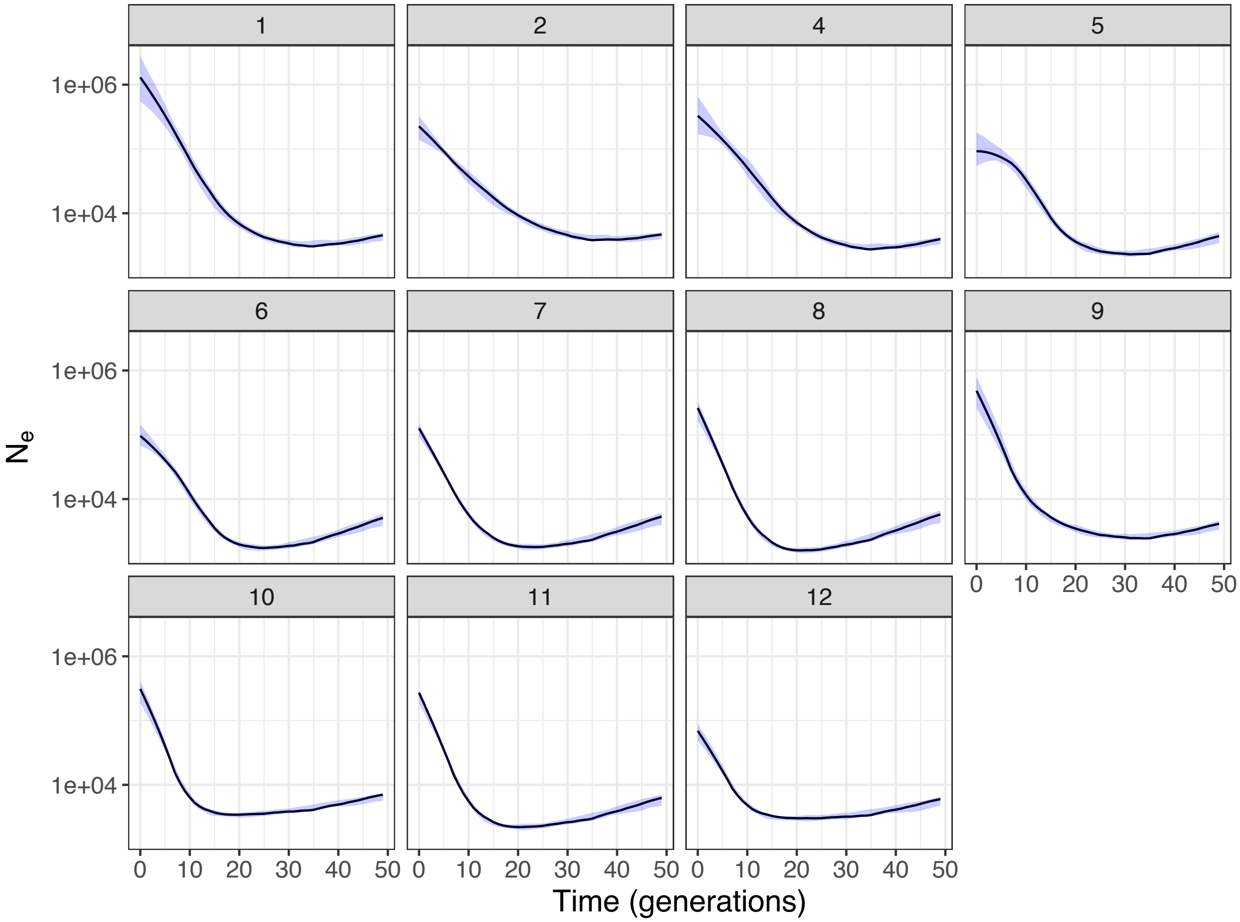

Figure S4 – Effective population size change over time by region of Finland. Number of individuals in each region are: 1: 1,123, 2: 1,078, 4: 378, 5: 224, 6: 304, 7: 1,581, 8: 1,547, 9: 225, 10: 288, 11: 1,697, 12: 184**.**

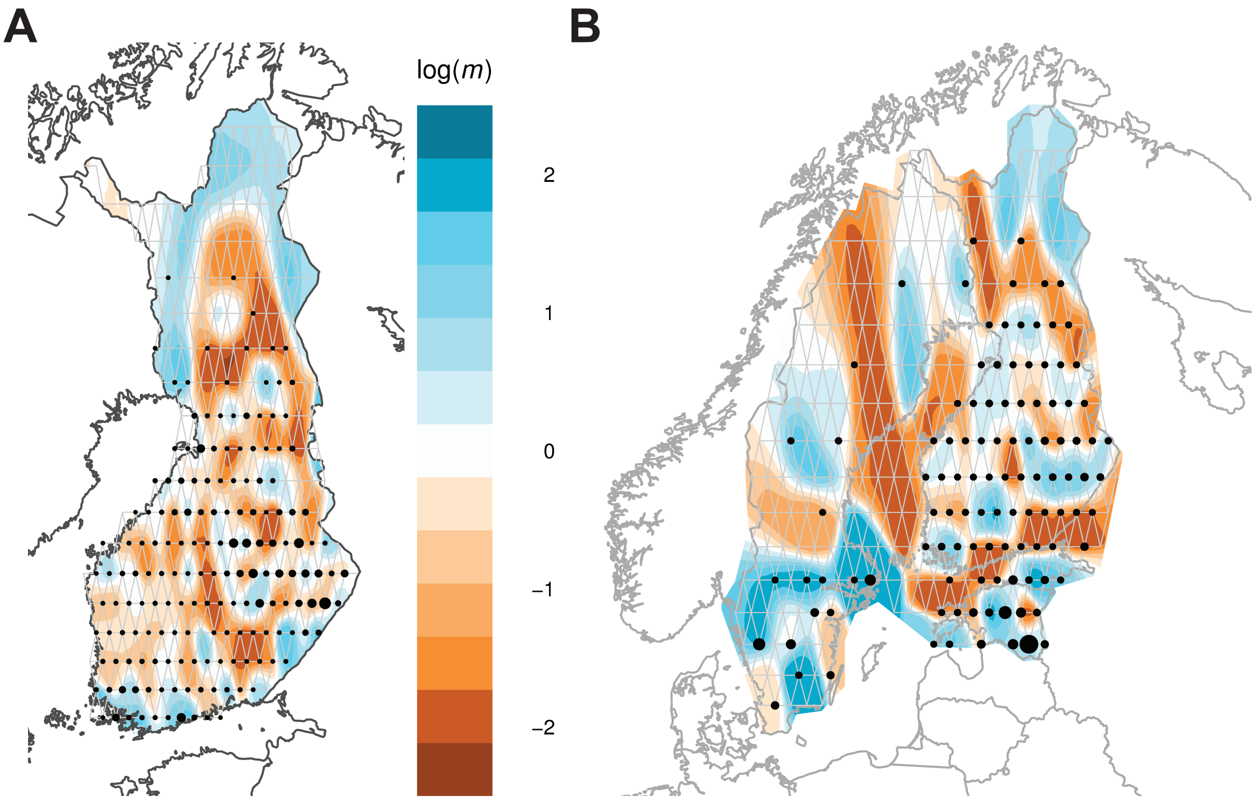

Figure S5 – Deme assignment for EEMS analyses in/near Finland. Black dots at center of demes are proportional to sample size. A) Finland deme assignment from municipality-level birth records. B) Deme assignment in Finland with municipality-level birth records and for region-level birth records in neighboring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

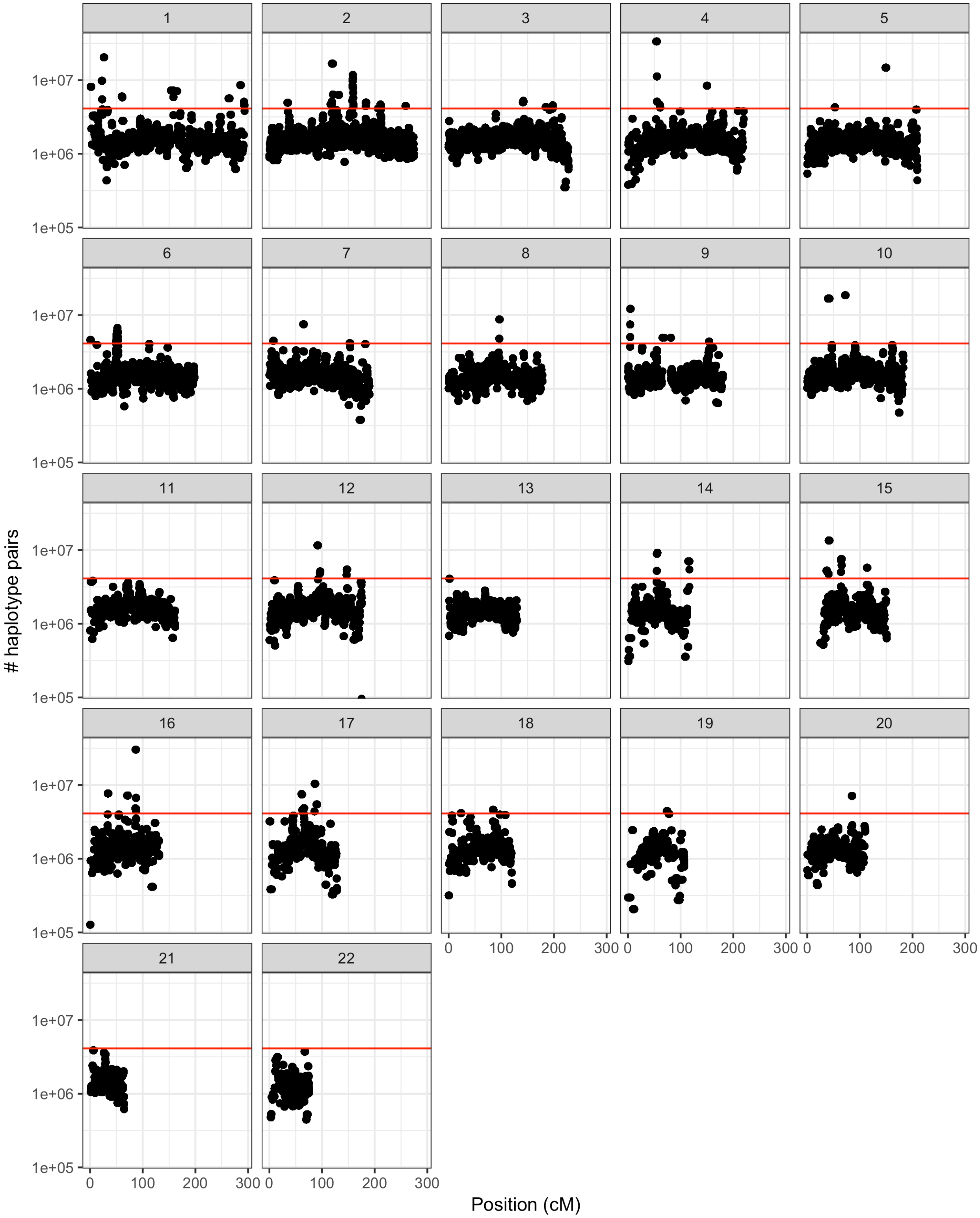

Figure S6 – Haplotype sharing rate genome-wide by chromosome. At each best-guess genotype used to call haplotypes, we quantified the number of pairs of individuals who shared haplotypes. Included individuals are unrelated and have corresponding

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.

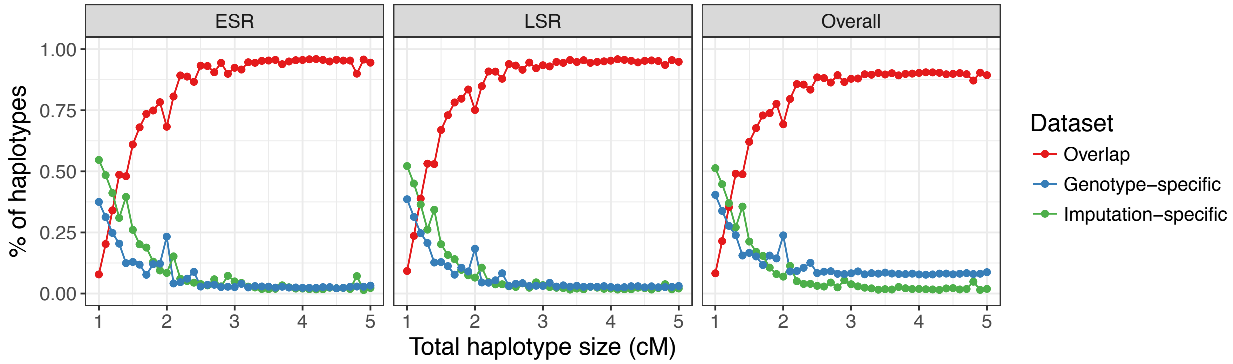

Figure S7 – Comparison of pairwise haplotype sharing for the same samples when using high quality imputed vs genotyped sites. Best guess hard call genotypes passed a filter of INFO > 0.99. Haplotypes were designated as overlapping (maximum start of either haplotype was less than the minimum end of either 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.

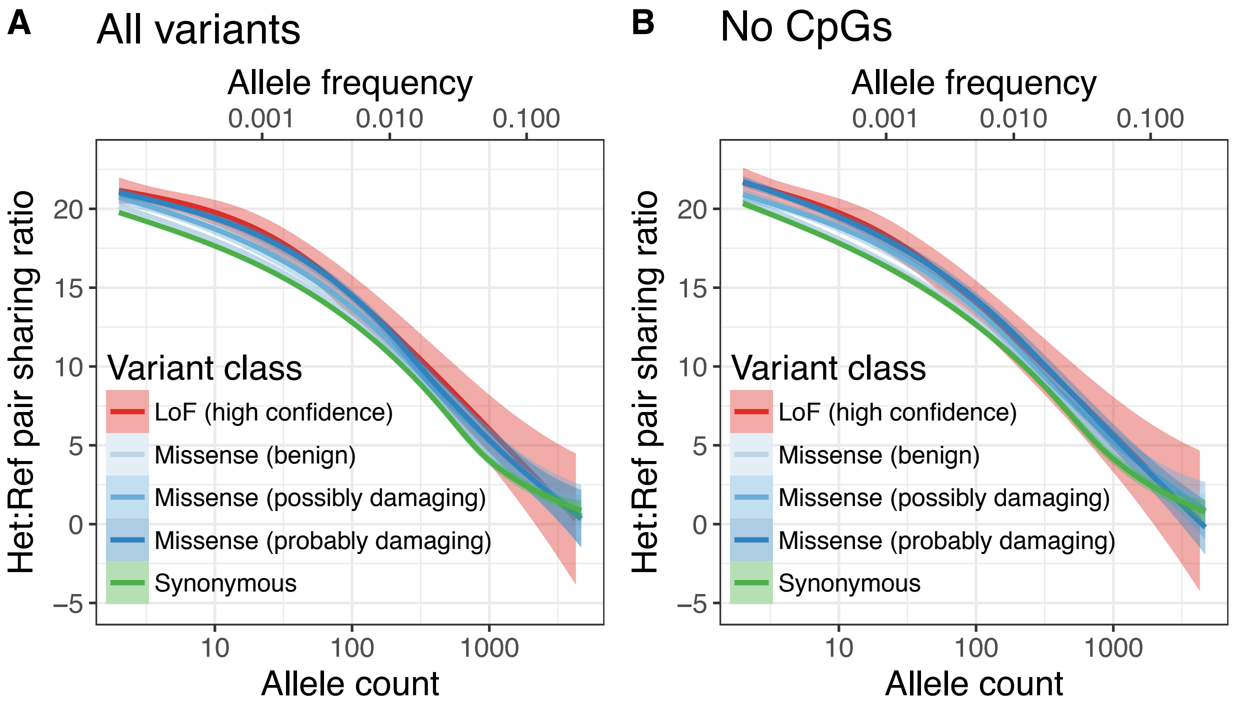

Figure S8 – Slight excess of haplotype sharing at non-CpG sites supports a reduction of haplotype sharing at highly mutable sites. A) Enrichment of haplotype sharing across all variants. B) Enrichment of haplotype sharing across non-CpG variants.

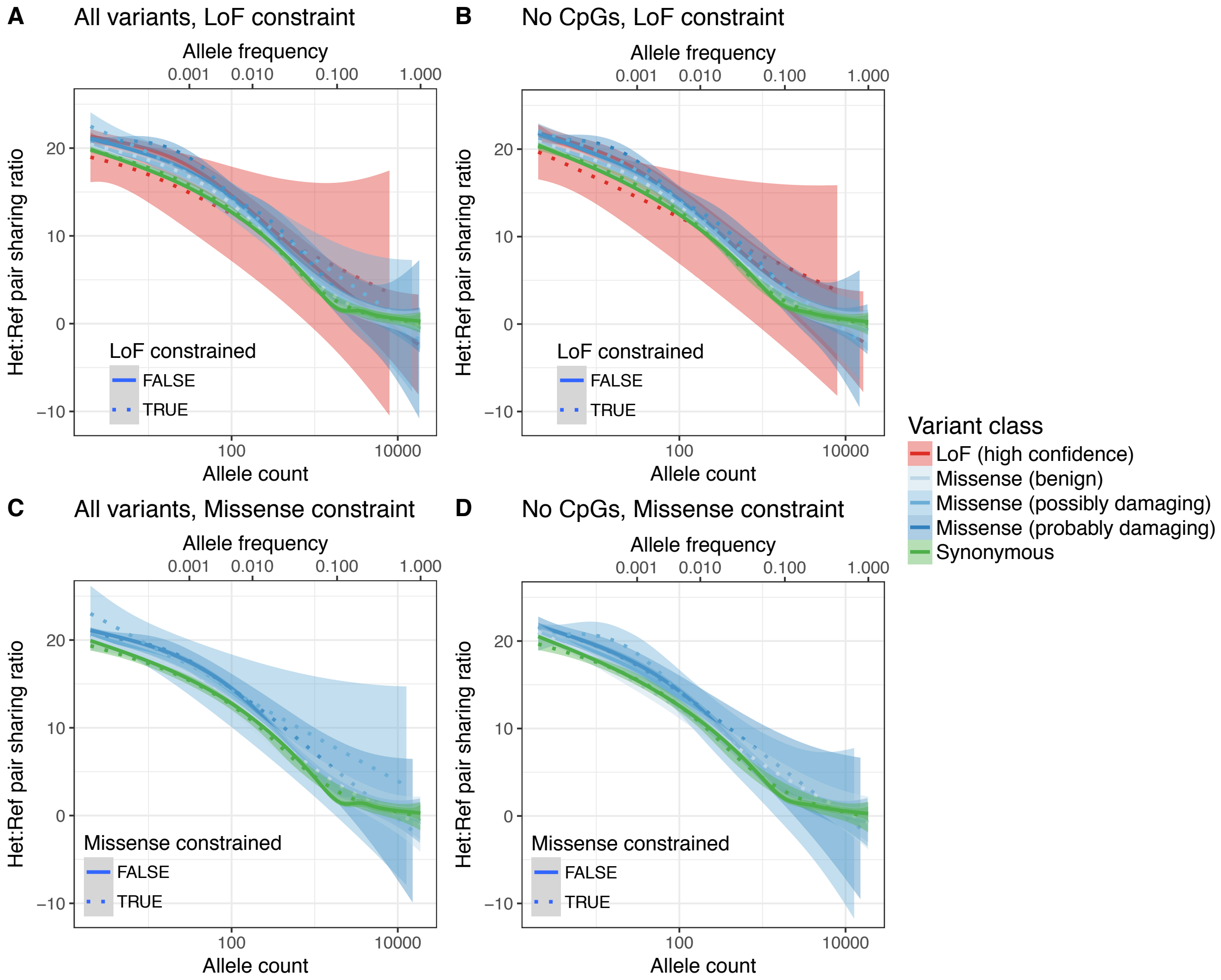

Figure S9 – Haplotype sharing rates are similar across variant classes in missense and loss-of-function (LoF) constrained regions. As calculated in Lek et al, missense constraint indicates regions depleted of missense variation, and LoF constrained regions indicate regions depleted of LoF variation[2,3]. 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.

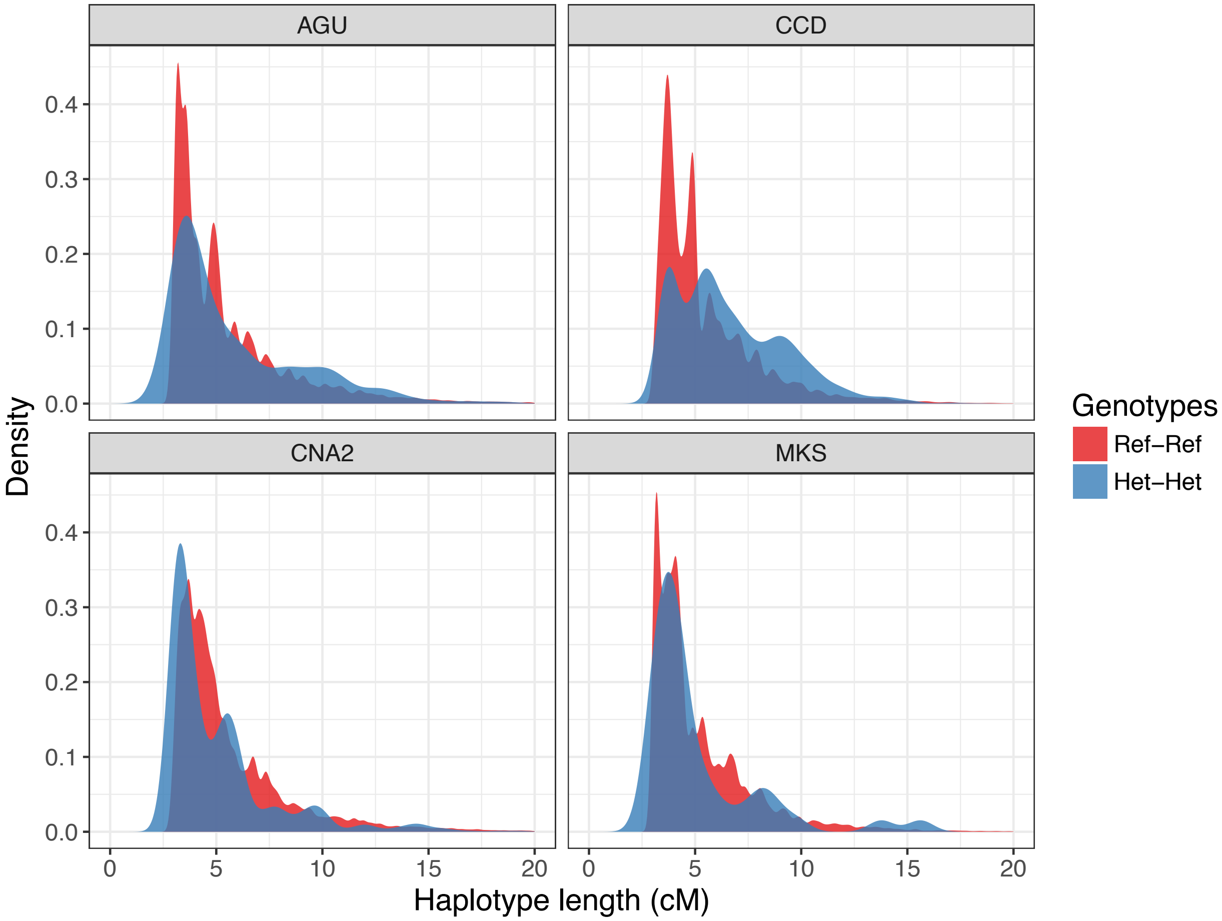

Figure S10 – Haplotype lengths among carrier versus homozygous reference pairs for variants contributing to four FinDis diseases. Variants and diseases correspond to those shown in Figure 5 and Table 1, with diseases as follows: AGU = Aspartylglucosaminuria, CNA2 = Cornea plana 2, CCD = Congenital chloride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