bioRxiv preprint doi: https://doi.org/10.1101/2020.05.20.106831; this version posted May 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

High-resolution population-specific recombination rates and their effect on phasing and genotype imputation

Running Title: Population-specific recombination maps in phasing & imputation

Shabbeer Hassan, *Institute for Molecular Medicine Finland, FIMM, HiLIFE, University of Helsinki, Helsinki, Finland*

Ida Surakka, *Institute for Molecular Medicine Finland, FIMM, HiLIFE, University of Helsinki, Helsinki, Finland*

Marja-Riitta Taskinen, *Clinical and molecular metabolism, Research program unit, University of Helsinki, Helsinki, Finland*

Veikko Salomaa, *Finnish Institute for Health and Welfare, Helsinki, Finland*

Aarno Palotie, *Institute for Molecular Medicine Finland, FIMM, HiLIFE, University of Helsinki, Helsinki, Finland, Psychiatric & Neurodevelopmental Genetics Unit, Department of Psychiatry, Analytic and Translational Genetics Unit, Department of Medicine, Department of Neurology, Massachusetts General Hospital, Boston, MA, USA*

Maija Wessman, *Institute for Molecular Medicine Finland, FIMM, HiLIFE, University of Helsinki, Helsinki, Finland*

Taru Tukiainen, *Institute for Molecular Medicine Finland, FIMM, HiLIFE, University of Helsinki, Helsinki, Finland*

Matti Pirinen, *Institute for Molecular Medicine Finland, FIMM, HiLIFE, University of Helsinki, Helsinki, Finland, Public Health, Clinicum, University of Helsinki, Helsinki, Finland, Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland*

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.20.106831; this version posted May 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Priit Palta, *Institute for Molecular Medicine Finland, FIMM, HiLIFE, University of Helsinki, Helsinki, Finland , Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia*

Samuli Ripatti, *Institute for Molecular Medicine Finland, FIMM, HiLIFE, University of Helsinki, Helsinki, Finland, Broad Institute of the Massachusetts Institute of Technology and Harvard University, Cambridge, MA, USA, Public Health, Clinicum, University of Helsinki, Helsinki, Finland*

CORRESPONDING AUTHOR

Samuli Ripatti, **Email**: samuli.ripatti@helsinki.fi

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.20.106831; this version posted May 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract:

Recombination is an essential part of meiosis as it facilitates novel combinations of homologous chromosomes, following their successive segregation in offspring. Founder population size, demographic changes (eg. population bottlenecks or rapid expansion) can lead to variation in recombination rates across different populations. Previous research has shown that using population-specific reference panels has a significant effect on downstream population genomic analysis like haplotype phasing, genotype imputation and association, especially in the context of population isolates. Here, we developed a high-resolution recombination rate mapping at 10kb and 50kb scale using high-coverage (20-30x) whole-genome sequencing (WGS) data of 55 family trios from Finland and compared it to recombination rates of non-Finnish Europeans (NFE). We then tested the downstream effects of the population-specific recombination rates in statistical phasing and genotype imputation in Finns as compared to the same analyses performed by using the NFE-based recombination rates . Finnish recombination rates have a moderately high correlation (Spearman's ρ =0.67-0.79) with non-Finnish Europeans, although on average (across all autosomal chromosomes), Finnish rates (2.268±0.4209 cM/Mb) are 12-14% lower than NFE (2.641±0.5032 cM/Mb). Population-specific effective population sizes were found to have no significant effect in haplotype phasing accuracy (switch error rates, SER ~ 2%) and average imputation concordance rates (with reference panels in phasing: rates were 97-98% for common, 92-96% for low frequency and 78-90% for rare variants) irrespective of the recombination map used. Similarly, we found no effect of population-specific (Finnish) recombination maps in phasing with comparable switch error rates (SER) across autosomes when compared to HapMap based maps. Our results suggest that

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.20.106831; this version posted May 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

downstream population genomic analyses like haplotype phasing and genotype imputation mostly depend on population-specific contexts like appropriate reference panels and their sample size, but not on population-specific recombination maps or effective population sizes. Currently, available HapMap recombination maps seem robust for population-specific phasing and imputation pipelines, even in the context of relatively isolated populations like Finland.

Keywords: recombination, phasing, imputation, Finland, population genomics

1. Introduction:

Recombination is not uniform across the human genome with large areas having lower recombination rates, so-called 'coldspots', which are then interspersed by shorter regions marked by a high recombinational activity called 'hotspots' [1]. With long chunks of human genome existing in high linkage disequilibrium, LD [2], and organised in the form of 'haplotype blocks', the 'coldspots' tend to coincide with such regions of high LD [3].

Direct estimation methods of recombination are quite time-consuming, and evidence has suggested that they do not easily scale up to genome-wide, fine-scale recombinational variation estimation [4]. A less time-consuming but computationally intensive alternative is to use the LD patterns surrounding the SNPs [5]. Such methods have been used in the past decade or so, to create fine-scale recombination maps [6]. Besides the International HapMap project that focused on capturing common variants and haplotypes in diverse populations, international WGS-based collaborations like 1000 Genomes Project, provided genetic variation data for 20 worldwide populations

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.20.106831; this version posted May 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[7]. This led to further refinement of the recombination maps coupled with methodological advances of using coalescent methods for recombination rate [8, 9]. With the rise of international collaborative projects, it was realised that founder populations can often have very unique LD patterns [10], subsequently also displaying unique increased genetics-driven health risks [11], suggesting that population-specific reference datasets should be used to leverage the LD patterns to better detect disease variants in downstream genetic analysis [12]. Genomic analysis methods like haplotype phasing and imputing genotypes require recombination maps and other population genetic parameters as input to obtain optimal results [13, 14, 15, 16]

In theis study, we set to test this by 1) estimating recombination rates along the genome in Finnish population using ~55 families of whole-genome sequenced (20-30x) Finns, 2) comparing these rates to some other European populations, and 3) comparing the effect of using Finnish recombination rate estimates and cosmopolitan estimates in phasing and imputation errors in Finnish samples.

2. Materials & Methods:

2.1 Datasets used:

Finnish Migraine Families Collection

Whole-genome sequenced trios (n = 55) consisting of the parent-offspring combination were drawn from a large Finnish migraine families collection consisting of 1,589 families totalling 8,319 individuals [17]. The trios were used for the recombination map construction using LDHAT version 2. The families were collected over 25 years from various headache clinics in Finland (Helsinki, Turku, Jyväskylä, Tampere, Kemi, and Kuopio) and via advertisements in the national migraine patient organisation web page (https://migreeni.org/). The families consist of different pedigree sizes from small to

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.20.106831; this version posted May 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

large (1-5+ individuals). Of the 8319 individuals, 5317 have a confirmed migraine diagnosis based on the third edition of the established International Classification for Headache Disorders (ICHD-3) criteria [18].

EUFAM cohort

To check the phasing accuracy of our Finnish recombination map, we used an independently sourced 49 trios from the European Multicenter Study on Familial Dyslipidemias in Patients with Premature Coronary Heart Disease (EUFAM). Finnish familial combined hyperlipidemia (FCH) families were identified from patients initially admitted to hospitals with premature cardiovascular heart disease (CHD) diagnosis who also had elevated levels of total cholesterol (TC), triglycerides (TG) or both in the ≥ 90th Finnish population percentile. Those families who had at least one additional first-degree relative also affected with hyperlipidemia were also included in the study apart from individuals with elevated levels of TG. [19, 20, 21].

FINRISK cohort

The imputation accuracy of the Finnish and previously published HapMap based recombination maps [8, 9] was subsequently tested on an independent FINRISK CoreExome chip dataset consisting of 10,481 individuals derived from the national-level FINRISK cohort. Primarily, it comprises of respondents of representative, cross-sectional population surveys that are conducted once every 5 years since 1972 to get a national assessment of various risk factors of chronic diseases and other health behaviours among the working-age population drawn from 3 to 4 major cities in Finland [22].

FINNISH reference panel cohort

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.20.106831; this version posted May 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The whole-genome sequenced samples used were obtained from PCR-free methods and PCR-amplified methods, which was followed by sequencing on a Illumina HiSeq X platform with a mean depth of ~30×. The obtained reads were then aligned to the GRCh37 (hg19) human reference genome assembly using BWA-MEM. Best practice guidelines from Genome Analysis Toolkit (GATK) were used to process the BAM files and variant calling. Several criteria were used in this stage for sample exclusion: relatedness (identity-by-descent (IBD) > 0.1), sex mismatches, among several others. Furthermore, samples were filtered based on other criteria such as: non-reference variants, singletons, heterozygous/homozygous variants ratio, insertion/deletion ratio for novel indels, insertion/deletion ratio for indels observed in dbSNP, and transition/transversion ratio.

After this stage, some exclusion criteria were applied to set some variants as missing: GQ < 20, phred-scaled genotype likelihood of reference allele < 20 for heterozygous and homozygous variant calls, and allele balance <0.2 or >0.8 for heterozygous calls. A truth sensitivity percentage threshold of 99.8% for SNVs and of 99.9% for indels was used based on the GATK Variant Quality Score Recalibration (VQSR) to filter variants with, quality by depth (QoD) < 2 for SNVs and < 3 for indels, call rate < 90%, and Hardy-Weinberg equilibrium (HWE) p-value < 1×10-9. Some other variants like monomorphic, multi-allelic and low-complexity regions [23] were further excluded.

The final reference dataset used in this study for imputation consisted of high coverage (20-30x) whole-genome sequence-based reference panel of 2690 individuals from the SISu project (Sequencing Initiative Suomi, http://www.sisuproject.fi/, [24]).

2.2 Recombination map construction:

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.20.106831; this version posted May 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Coalescent-based fine-scale recombination map construction [8] is greatly eased by using trios which provide more accurate haplotype phasing resolution [25]. Hence, we used trio data (n=55, 110 independent parents) from the Finnish Migraine Families Cohort described above. These were filtered primarily using VCFtools [26] and custom R scripts. Firstly, sites were thinned with within 15bp of each other such that only one site remained followed by a filtering step of removing variants with a minor allele frequency of <5% [27]. The resultant data were then phased using family-aware method of SHAPEIT [28] using the standard HapMap recombination map [8, 9], which was then split into segments of ~10000 SNPs with a 1000 SNP overhang on each side of the segments. LDhat version 2 was run for 10^7 iterations with a block penalty of 5, every 5000 iterations of them of which the first 10% observations were discarded [8, 29]. The CEU based maps, used here for comparison, were obtained similarly using LDhat [29].

However, LDHat is computationally intensive, and calculations suggest that the 1000 Genomes OMNI data set [30] would be too much computationally intensive to complete [31], hence limiting the maximum number of haplotypes which could be used.

To overcome this and make the recombination map independent of the underlying methodology, we used a machine learning method implemented in FastEPRR [31, 32]. It supports the use of larger sample sizes, than LDHat and the recombination estimates for sample sizes > 50, yields smaller variance than LDHat based estimates [31]. The method was then applied to each autosome with overlapping sliding windows (*i.e.*, window size, 50 kb and step length, 25 kb) under default settings for diploid organisms.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.20.106831; this version posted May 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

As seen in [31] both methods produce similar estimates, with only variance of the estimate of mean being different.

The output of LDHat and FastEPRR is in terms of population recombination rate (p) and to convert them into per-generational rate (r) used in phasing/imputation algorithms we used optimal effective population size values derived from our testing (as explained in the Supplementary Text). The estimates from LDHat and FastEPRR were then averaged, to obtain a new combined estimate with the lowest variance amongst all the three [31].

2.3 Phasing and imputation accuracy

To test whether the usage of different recombination maps affects the efficiency of haplotype phasing and imputation , we used the aforesaid Finnish genotype data to evaluate: (i) switch error rates across all chromosomes and (ii) imputation concordance rates for chromosome 20.

2.3.1 Phasing Accuracy

The gold standard method to estimate haplotype phasing accuracy is to count the number of switches (or recombination events) needed between the computationally phased dataset and the true haplotypes [33].The number of such switches divided by the number of all possible switches is called switch error rate (SER).

For testing the influence of recombination maps on phasing accuracy, we used three different recombination maps: HapMap, fine-scale Finnish recombination map and a constant background recombination rate (1cM/Mb), to phase the 55 offspring haplotypes without using any reference dataset. To check whether reference panels used during haplotype phasing made any impact on the switch error rates, we used the

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.20.106831; this version posted May 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Finnish SISU based reference (n=2690), to check whether the size of the reference panel made any impact on the results in phasing the offspring's haplotypes (Figure 1). The SER in the offspring's phased haplotypes were then calculated by determining the true offspring haplotypes using data from the parents (98 individuals) with a custom script [34].

2.3.2 Imputation Accuracy

Imputation concordance was used as the metric for calculating the imputation accuracy. For this, we randomly masked FINRISK CoreExome chip data consisting of 10,480 individuals [22] from chromosome 20. To test the role of reference panel size in influencing the imputation accuracy in conjunction with varying the population genetics parameters, we imputed the masked dataset with BEAGLE (Browning *et al*., 2016) using the Finnish reference panel (n = 2690). The concordance was then calculated between the imputed genotypes and the original masked variants. Masking was done by randomly removing ~10% of variants from the chip dataset.

The influence of recombination maps on imputation accuracy was checked by calculating the concordance values between imputed and original variants, using the Finnish reference panel in various combinations of recombination maps (constant rate, HapMap, Finnish map) during the imputation (Figure 1).

3. Results:

3.1 Finnish recombination map and its comparison to the HapMap recombination map:

The primary aim of our study was to derive a high-resolution genetic recombination map for Finland and use it for comparative tests in commonly used analyses like haplotype phasing and imputation. To derive a population-specific Finnish

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.20.106831; this version posted May 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

recombination map, we used the high-coverage WGS data and an average of different estimation methods (LDHat and FastEPRR). We used the Ne value of 10,000 derived from our extensive testing of different Ne values (See supplementary text) to get the per-generation recombination rates. The average recombination rates of Finnish population isolate depicted 12-14% lower values (autosome-wide average 2.268±0.4209 cM/Mb) for all chromosomes compared to CEU based maps (2.641±0.5032 cM/Mb) (Figure 2).

These differences in average recombination rates are reflected in the correlation values across all chromosomes (Spearman's $\rho \sim 0.67 - 0.79$) between the developed Finnish map and HapMap based one (Figure 2). We also present a direct comparison between the two maps, of the recombination rates at 5Mb scales, which presents a similar visual pattern of rates across the genome (Supplementary Figure 1).

3.2 Effects of the population-specific recombinations map on haplotype phasing

Variation in population-specific recombination maps (and effective population sizes) can affect the downstream genomic analyses like haplotype phasing and imputation.

We tested the Finnish map, HapMap map and a constant recombination rate map (1cM/Mb) to understand the effects of population-specific maps on downstream genomic analyses. The phasing accuracy was tested under two different conditions: using no additional reference panel and using an population-specific . SISu v2 reference panel (n= 2690) in phasing. We observed that, on average, SER ranged between 1.8-3.7% (Supplementary Figure 2) across the different chromosomes and recombination maps. We found statistically significant differences within both no-reference panel and the Finnish reference panel results (Kruskal Wallis, p-value = 5.3e-10 and 4.7e-10, respectively; Figure 3). The constant recombination map (1cM/Mb) had significantly

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.20.106831; this version posted May 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

higher SER values when compared to the Finnish map or the HapMap map (Figure 3) both when no reference panels were used (p-value = 2.9e-11 and 2.6e-09, respectively) and when the Finnish reference panel was used (p-value = 2.9e-11 and 9.5e-13, respectively). The choice of recombination maps mattered more when no reference panel was used (p-value = 0.0046), however when using the Finnish reference panel, the difference in SER was statistically insignificant (p-value = 0.25).

3.3 Effects of the population-specific recombinations map on genotype imputation

Imputation accuracy was similarly tested using the reference panel under three different recombination map settings. We observed that when the imputation target dataset was phased and imputed using the Finnish reference panel (n=2690) irrespective of the population-specific recombination maps, it had a high imputation accuracy (overall concordance rate ~98%, Figure 4) across MAF bins (>0.1%). Though some differences in concordance rates are seen in for rare variants (MAF <0.1%). The concordance rate was lower when the test dataset was phased without reference panels (concordance rate 72~77%, Figure 5).

4. Discussion:

Population isolates like Finland, have had a divergent demographic history as compared to the outbred European populations, with a less historic migration, more fluctuating population sizes and higher incidences of bottleneck events and founder effects [35, 36] This unique demographic history then affects different population genetic parameters, like recombination rates [37]. It has been shown previously that using population-specific genomic reference panels augmented the accuracy of imputation accuracy leading to better mapping of diseases specific variants in GWAS [12]. Since recombination rates (in the form of recombination maps), features in much of the

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.20.106831; this version posted May 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

downstream genomic analyses' methods like imputation and haplotype phasing [15, 34], we wanted to study their effect on downstream analyses.

Firstly, we characterised the Finnish recombination map using high-coverage (~30x) whole-genome sequencing (WGS) samples from large SISu v2 reference panel (n=2690). Previously used recombination maps hail from the HapMap and 1000Genomes projects which used sparse genotypic datasets or low-depth sequencing samples. This is a first attempt in creating a recombination map for Finland using population-specificWGS samples. We used two different methods in estimating the recombination rates, to achieve accurate estimates with lower variance [29,31]. In addition, we estimated effective population sizes using identity-by-descent (IBD) based methods [15] for both Finnish and CEU based datasets. The obtained recombination map was then used to test their role and importance in two selected downstream genomic analyses – haplotype phasing and imputation concordance. Since the recombination rate determination requires effective population size estimates, we also tested the role of varying effective population size on these two analyses (See Supplementary Text). The extensive testing of Ne yielded the estimate of 10,000 originally derived theoretically [38] and most used commonly for humans fits quite rightly for the recombination map.

The Finnish recombinational landscape when compared to the HapMap based map, showed, on average, a high degree of correlation across scales (10, 50kb and 5Mb), however, on average, Finnish recombination rates across chromosomes were found to be lower. Such moderate to high correlations (Figure 2) and similar recombinational landscape (Supplementary Figure 1) could be due to high sharing of recombinations in individuals from closely-related populations. The degree of dissimilarity in the

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.20.106831; this version posted May 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

population-level differences between Finnish and mainland Europeans in terms of recombination rates could be due to population-specific demographic processes like founder effects, bottleneck events and migration [39], or chromatin structure PRDM9 binding locations, for example [40]. And the broad similarity in terms of correlational structure seen here, reflects a shared ancestral origin of Finns and other mainland Europeans [41]. Other studies on population isolates like Iceland [9] have previously found a high degree of correlation with CEU based maps, albeit with substantial differences as seen here. Previous studies [42] have additionally explored the relationship between recombination rate differences between populations and allele frequency differences, with evidence suggesting that the differences between rates show the selection impact in the past 100,000 years since the out-of-Africa movement of humans.

As seen in previous studies, much of the downstream genomic analyses like getting more refined GWAS hits or, accurate copy number variants (CNV) imputation, can be highly improved with the addition/use of population-specific datasets [12]. To test this in the context of population-specific recombination maps, we used them to test the haplotype phasing and imputation accuracy and observed that despite large differences in the effective population sizes between populations, it did not affect the tested metrics. One possible explanation for the insignificant effect seen here is that the role of parameters like effective population size and recombination maps is to scale over the haplotypes for efficient coverage of the whole genome. However, when sufficiently large, population-specific genomic reference panels are available with tens of thousands of haplotypic combinations, such scaling over for specific populations, does not yield in substantial improvements. As we showed here, reference panel size could play an

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.20.106831; this version posted May 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

important role in the downstream genomic analyses and in most cases, the current practice of using the standard HapMap recombination map can be reasonably used. Another point of interest here is that the use of different Ne parameters during phasing/imputation might be redundant as we observed no change in the accuracy of our estimates on varying the Ne parameters. Similarly, when using population-specific recombination maps, we did not find any tangible benefits in using them over the current standard maps based on the HapMap data.

Our study suggests a couple of important points for future studies: (a) varying effective population size for downstream genomic analyses, such as phasing and imputation, might have a relatively small impact, and it might be better to use the default option of the particular software; (b) when available, it is beneficial to use a population-specific genomic reference panel as they increase the accuracy; (c) HapMap can be used for current downstream genomic analyses like haplotype phasing or genotype imputation in European-based populations. And, if need be, can be substituted for using population-specific maps, as the accuracy rates are quite similar to the population-based maps.

Though the sample used here is from a disease cohort but is nevertheless representative of Finland's population and hence provides a reasonable recombination rate estimates. On the other hand, our reliance on disease cohorts could lead to minor variation in the resultant recombination. Though as we share a similar out-of-Africa origin, much of our history is shared and though biological differences in the recombinational landscape do exist between different populations, much of the downstream genomic analyses (haplotyping, imputation or, GWAS), might not be affected by recombination map or values of effective population size.

Funding

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.20.106831; this version posted May 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

This work was financially supported by the Academy of Finland (251217 and 255847 to S.R.). S.R. was further supported by the Academy of Finland Center of Excellence for Complex Disease Genetics, the Finnish Foundation for Cardiovascular Research, Biocentrum Helsinki, and the Sigrid Jusélius Foundation. S.H. was supported by FIMM-EMBL PhD program doctoral funding and I.S. by Academy of Finland Postdoctoral Fellowship (298149). V.S. was supported by the Finnish Foundation for Cardiovascular Research. T.T. was supported by Academy of Finland grant number 315589.

Acknowledgements

We would like to thank Sari Kivikko and Huei-Yi Shen for management assistance. The FINRISK analyses were conducted using the THL biobank permission for project BB2015_55.1. We thank all study participants for their generous participation in the FINRISK study.

Conflict of Interest: VS has received honoraria from Novo Nordisk and Sanofi for consulting and has ongoing research collaboration with Bayer ltd (all unrelated to the present study).

References

1. Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M et al. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science 2009; 327: 836–840
2. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. High-resolution haplotype structure in the human genome. Nature Genetics 2001; 29: 229–232

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.20.106831; this version posted May 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3. Hudson RR, Kaplan NL. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 1985; 111: 147-164
4. Chan AH, Jenkins PA, Song YS. Genome-wide fine-scale recombination rate variation in Drosophila melanogaster. PLoS Genet 2012; 8: e1003090
5. McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. The fine-scale structure of recombination rate variation in the human genome. Science 2004; 304: 581-584
6. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. A fine-scale map of recombination rates and hotspots across the human genome. Science 2005; 310: 321-324.
7. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO et al. A global reference for human genetic variation. Nature 2015; 526: 68-74
8. Auton A, McVean G. Recombination rate estimation in the presence of hotspots. Genome Res 2007; 17: 1219-1227.
9. Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A et al. Fine-scale recombination rate differences between sexes, populations and individuals. Nature 2010; 467: 1099-1103.
10. Service S, DeYoung J, Karayiorgou M, Roos JL, Pretorious H, Bedoya G et al. Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. Nat Genet 2006; 38: 556-560.
11. Peltonen L, Jalanko A, Varilo T. Molecular genetics of the Finnish disease heritage. Hum Mol Genet 1999; 8: 1913-1923.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.20.106831; this version posted May 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

12. Surakka I, Kristiansson K, Anttila V, Inouye M, Barnes C, Moutsianas L et al. Founder population-specific HapMap panel increases power in GWA studies through improved imputation accuracy and CNV tagging. Genome Res 2010; 20: 1344-1351.
13. Tewhey R, Bansal V, Torkamani A, Topol EJ, Schork NJ. The importance of phase information for human genomics. Nat Rev Genet 2011; 12: 215-223.
14. Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. Nat Rev Genet 2011; 12: 703-714.
15. Browning BL, Browning SR. Genotype Imputation with Millions of Reference Samples. Am J Hum Genet 2016; 98: 116-126.
16. Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods 2013; 10: 5-6.
17. Gormley P, Kurki MI, Hiekkala ME, Veerapen K, Häppölä P, Mitchell AA et al. Common Variant Burden Contributes to the Familial Aggregation of Migraine in 1,589 Families. Neuron 2018; 98: 743-753.e4.
18. The International Classification of Headache Disorders, 3rd edition (beta version). Cephalalgia 2013; 33: 629-808.
19. Borodulin K, Vartiainen E, Peltonen M, Jousilahti P, Juolevi A, Laatikainen T et al. Forty-year trends in cardiovascular risk factors in Finland. Eur J Public Health 2015; 25: 539-546.
20. Porkka KV, Nuotio I, Pajukanta P, Ehnholm C, Suurinkeroinen L, Syvänne M et al. Phenotype expression in familial combined hyperlipidemia. Atherosclerosis 1997; 133: 245-253.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.20.106831; this version posted May 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

21. Ripatti P, Rämö JT, Söderlund S, Surakka I, Matikainen N, Pirinen M et al. The Contribution of GWAS Loci in Familial Dyslipidemias. PLOS Genetics 2016; 12: e1006078.

22. Vartiainen E, Laatikainen T, Peltonen M, Juolevi A, Mannisto S, Sundvall J et al. Thirty-five-year trends in cardiovascular risk factors in Finland. International Journal of Epidemiology 2009; 39: 504–518.

23. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics 2014; 30: 2843–2851.

24. Mart Kals, Tiit Nikopensius, Kristi Läll, Kalle Pärn, Timo Tõnis Sikka, Jaana Suvisaari, Veikko Salomaa, Samuli Ripatti, Aarno Palotie, Andres Metspalu, Tõnu Esko, Priit Palta, Reedik Mägi Advantages of genotype imputation with ethnically matched reference panel for rare variant association analyses bioRxiv 579201; doi: https://doi.org/10.1101/579201

25. Roach JC, Glusman G, Hubley R, Montsaroff SZ, Holloway AK, Mauldin DE et al. Chromosomal haplotypes by genetic phasing of human families. Am J Hum Genet 2011; 89: 382-397.

26. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA et al. The variant call format and VCFtools. Bioinformatics 2011; 27: 2156-2158.

27. Stevison LS, Woerner AE, Kidd JM, Kelley JL, Veeramah KR, McManus KF et al. The Time Scale of Recombination Rate Evolution in Great Apes. Mol Biol Evol 2016; 33: 928-945.

28. O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M et al. A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genet 2014; 10: e1004234.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.20.106831; this version posted May 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

29. Auton A, Fledel-Alon A, Pfeifer S, Venn O, Ségurel L, Street T et al. A fine-scale chimpanzee genetic map from population sequencing. Science 2012; 336: 193-198.

30. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE et al. An integrated map of genetic variation from 1,092 human genomes. Nature 2012; 491: 56-65.

31. Gao F, Ming C, Hu W, Li H. New Software for the Fast Estimation of Population Recombination Rates (FastEPRR) in the Genomic Era. G3 (Bethesda) 2016; 6: 1563-1571.

32. Lin K, Futschik A, Li H. A fast estimate for the population recombination rate based on regression. Genetics 2013; 194: 473-484.

33. Bansal V. Integrating read-based and population-based phasing for dense and accurate haplotyping of individual genomes. Bioinformatics 2019; 35: i242-i248.

34. Loh PR, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet 2016; 48: 1443-1448.

35. Martin AR, Karczewski KJ, Kerminen S, Kurki MI, Sarin AP, Artomov M et al. Haplotype Sharing Provides Insights into Fine-Scale Population History and Disease in Finland. Am J Hum Genet 2018; 102: 760-775.

36. Kerminen S, Havulinna AS, Hellenthal G, Martin AR, Sarin AP, Perola M et al. Fine-Scale Genetic Structure in Finland. G3 (Bethesda) 2017; 7: 3459-3468.

37. Wang J, Santiago E, Caballero A. Prediction and estimation of effective population size. Heredity (Edinb) 2016; 117: 193-206.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.20.106831; this version posted May 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

38. Takahata N, Satta Y, Klein J. Divergence time and population size in the lineage leading to modern humans. Theor Popul Biol 1995; 48: 198-221.

39. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A et al. Genes mirror geography within Europe. Nature 2008; 456: 98-101.

40. Ségurel L. The complex binding of PRDM9. Genome Biol 2013; 14: 112.

41. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature 2016; 538: 201-206.

Keinan A, Reich D. Human population differentiation is strongly correlated with local recombination rate. PLoS Genet 2010; 6: e1000886.

Figure 1: Flowchart overview of the analyses and comparisons performed.

Figure 2: Average (± standard deviation) recombination rates of Finnish v/s CEU per autosome measured in cM/Mb and Correlation between Finnish and CEU recombination rates across all chromosomes. The comparisons are made for similar physical positions.

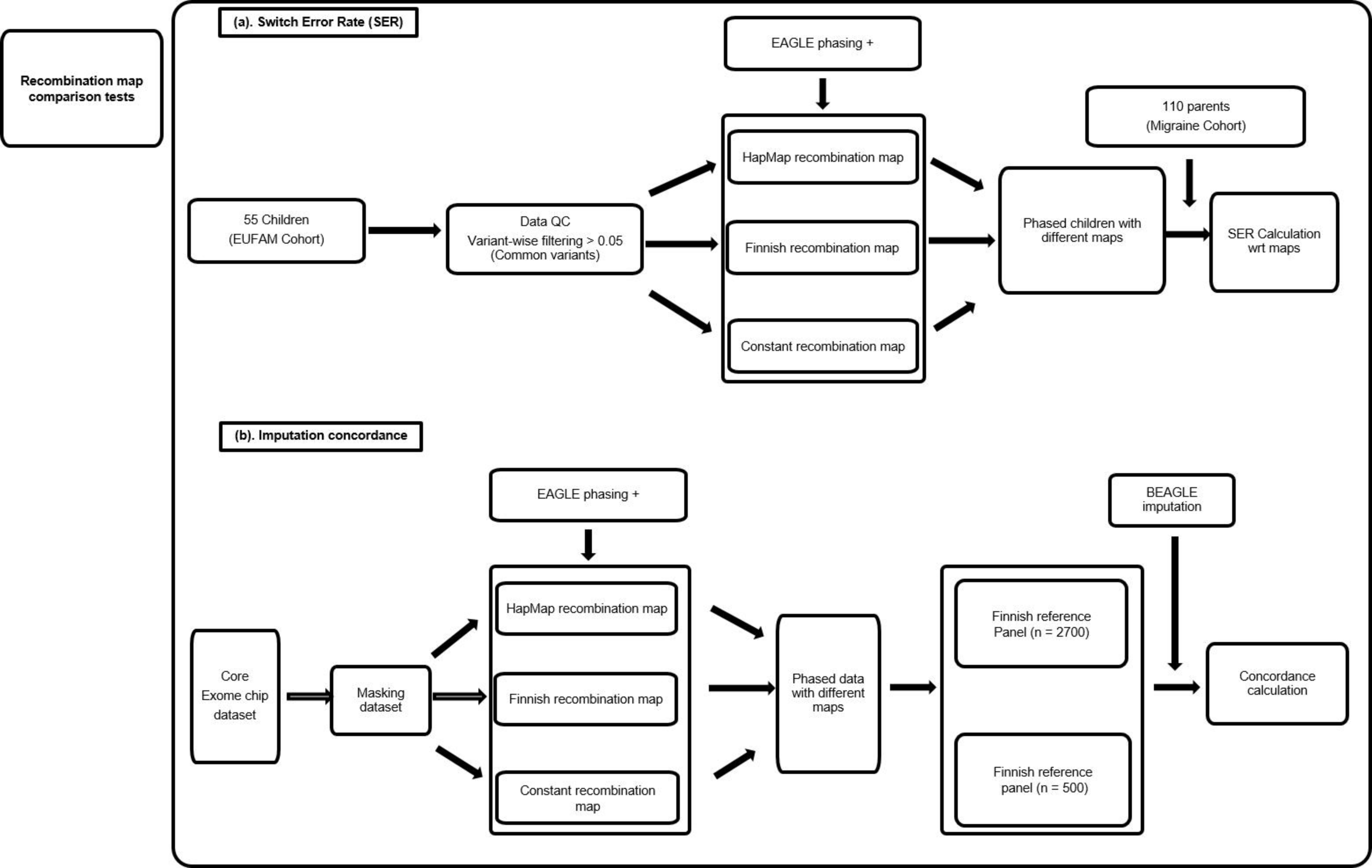
Figure 3: Statistical comparison of Switch Error Rates across all autosomes calculated for all children in the trios using different recombination maps with respect to different

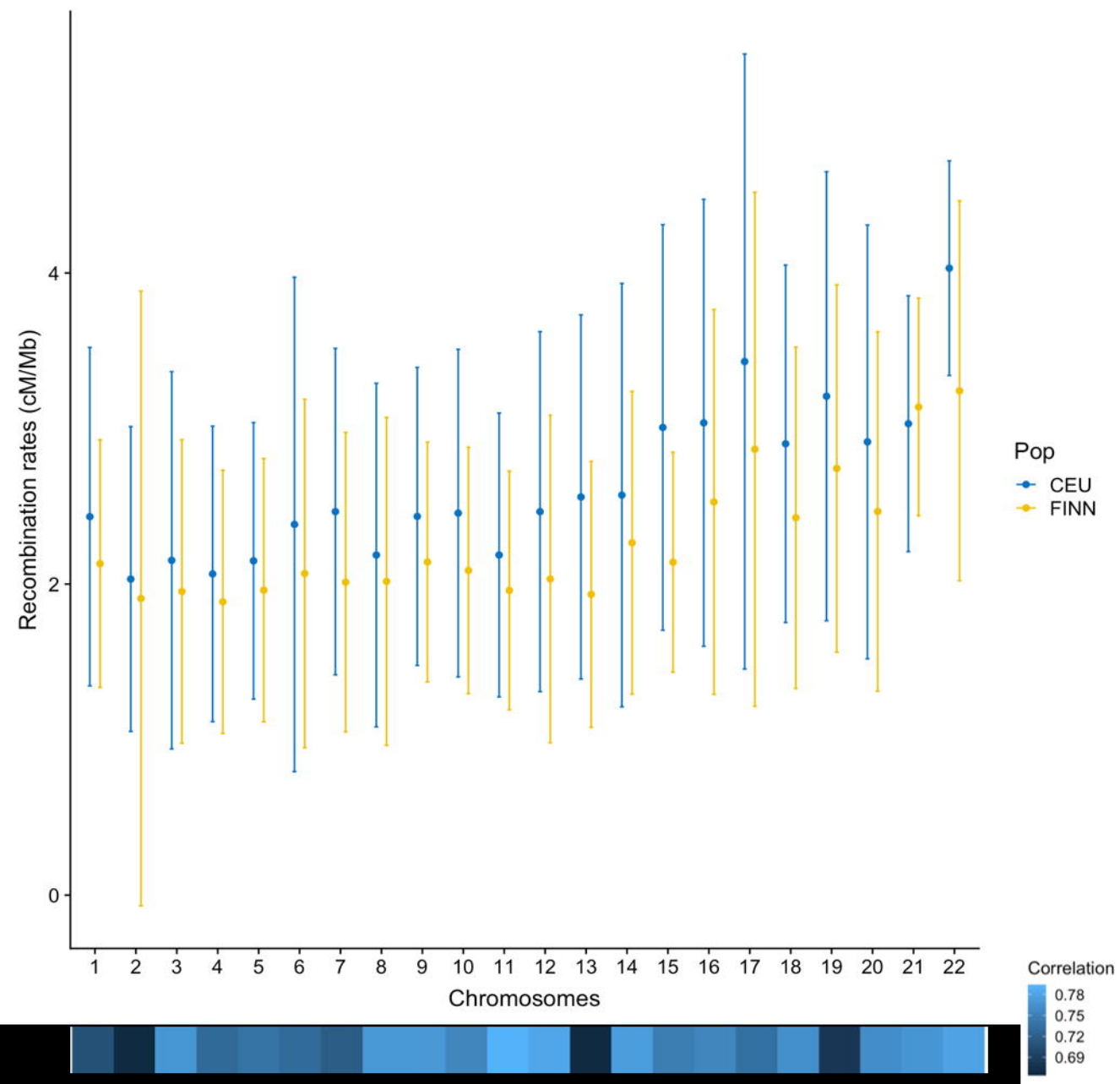
bioRxiv preprint doi: https://doi.org/10.1101/2020.05.20.106831; this version posted May 22, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

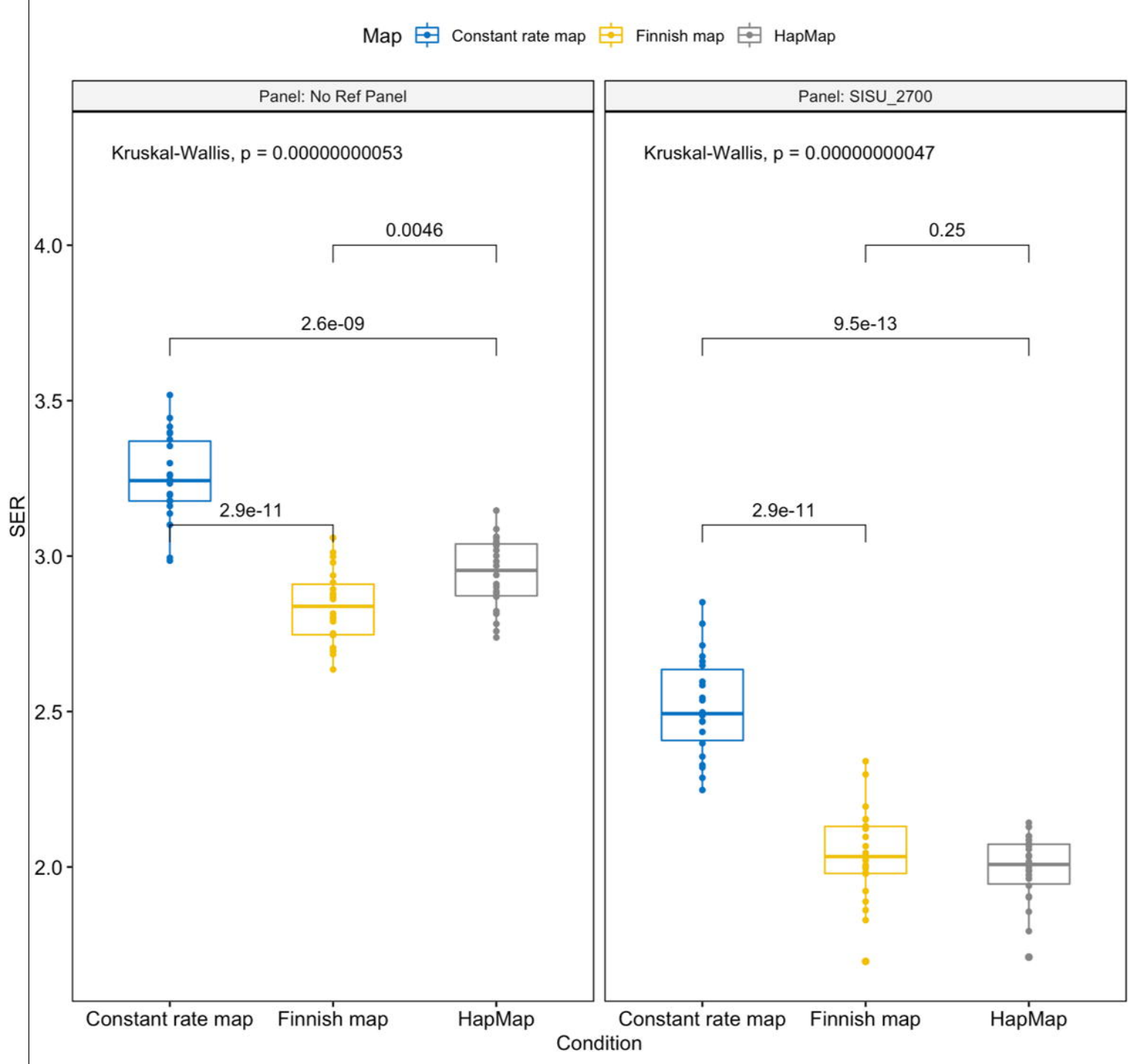
reference panel conditions (absent or present). The p-values are shown at the top of each panel from Kruskal Wallis ANOVA testing between panel groups and ones between boxplots for within-group comparisons.

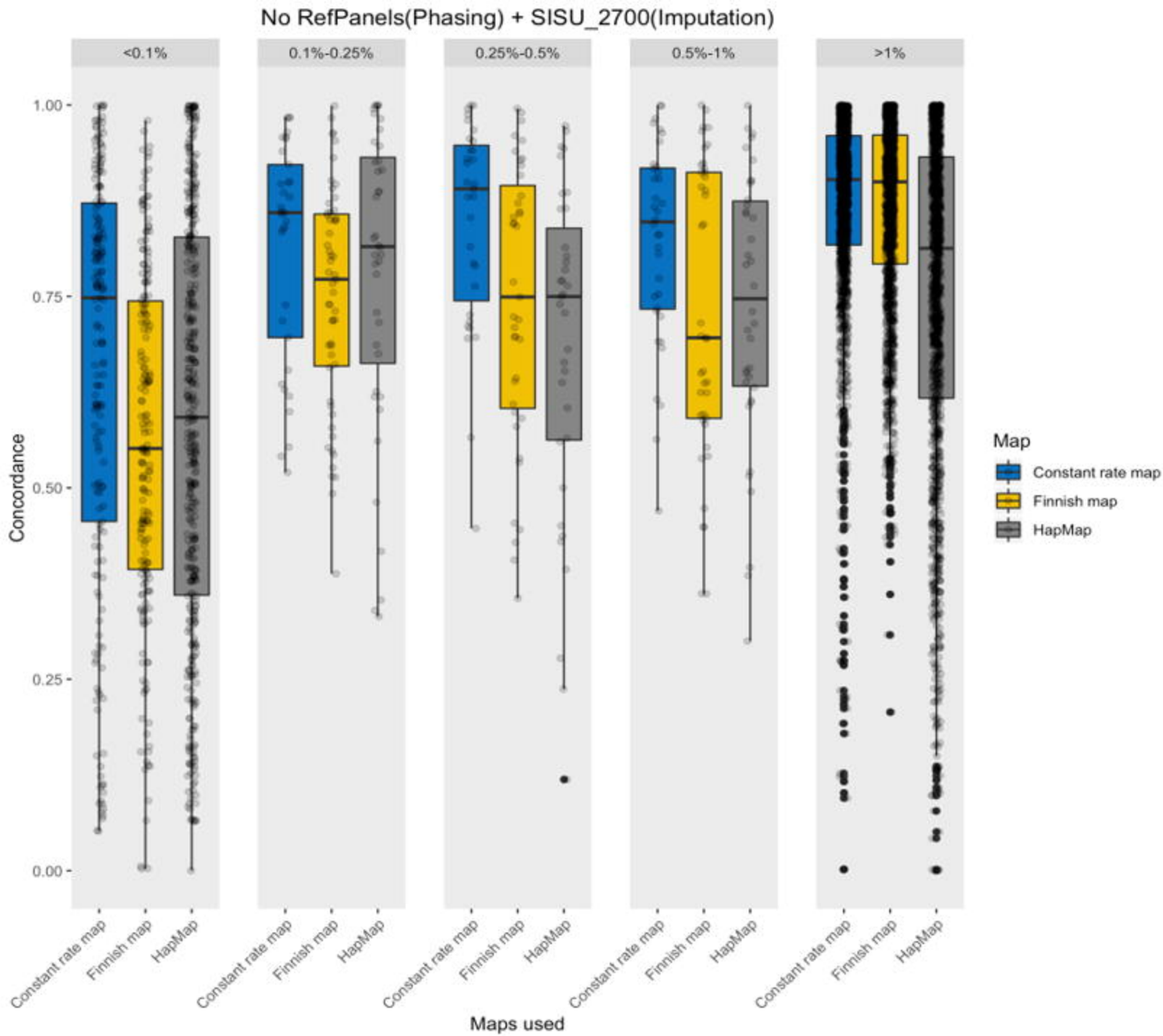
Figure 4: Comparison of Imputation Concordance across different Minor Allele Frequency (MAF) groups for a range of different recombination map combinations phased with NO reference panels

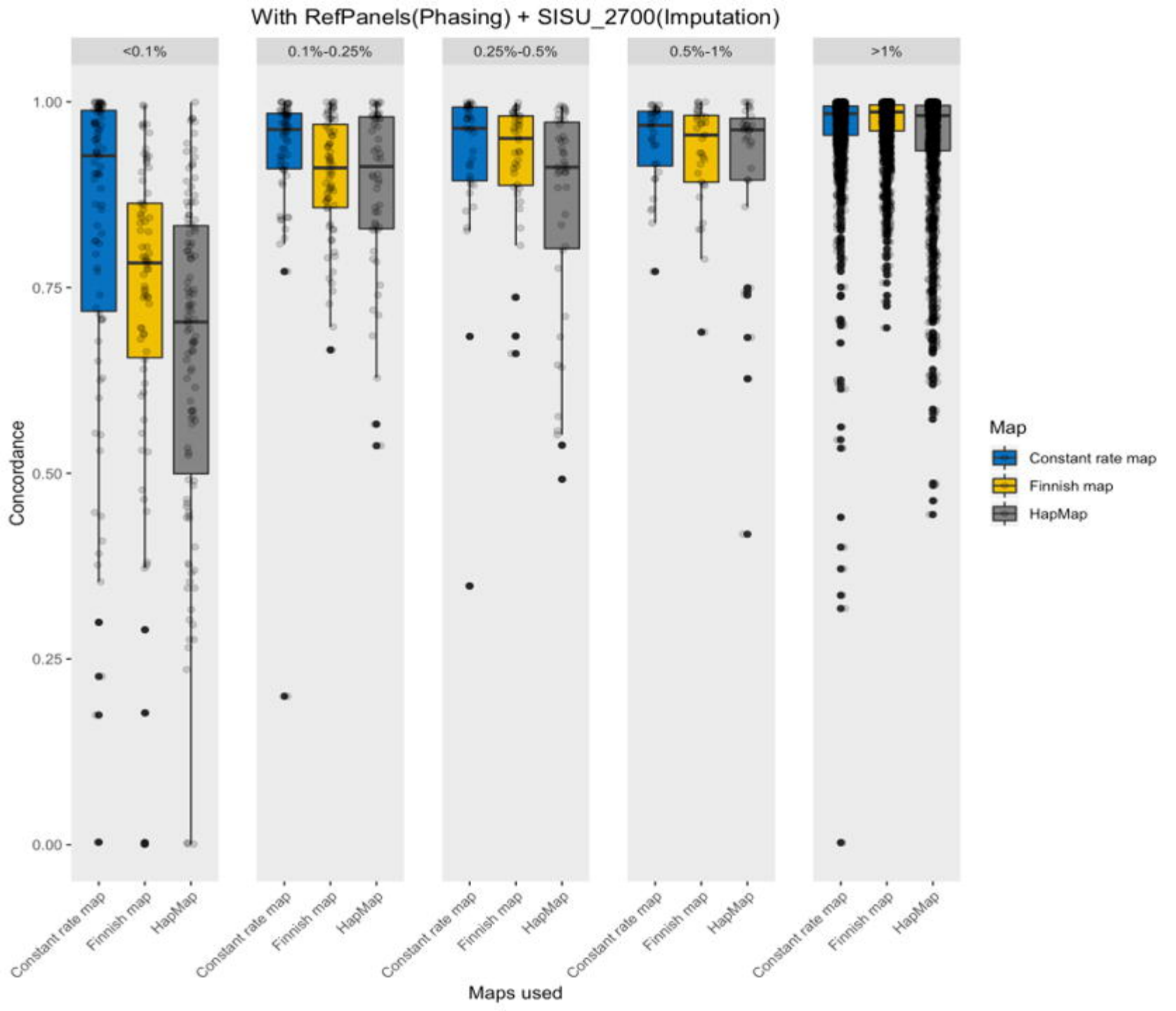
Figure 5: Comparison of Imputation Concordance across different Minor Allele Frequency (MAF) groups for a range of different recombination map combinations phased with reference panels.


Recombination map comparison tests
(a). Switch Error Rate (SER)
EAGLE phasing +
110 parents (Migraine Cohort)
HapMap recombination map
55 Children (EUFAM Cohort)
Data QC Variant-wise filtering > 0.05 (Common variants)
Finnish recombination map
Phased children with different maps
SER Calculation wrt maps
Constant recombination map
(b). Imputation concordance
EAGLE phasing +
BEAGLE imputation
HapMap recombination map
Core Exome chip dataset
Masking dataset
Finnish recombination map
Phased data with different maps
Finnish reference Panel (n = 2700)
Concordance calculation
Constant recombination map
Finnish reference panel (n = 500)


Recombination rates (cM/Mb)
0
2
4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Chromosomes
Pop
CEU
FINN
Correlation
0.78
0.75
0.72
0.69


Map
Constant rate map
Finnish map
HapMap
Panel: No Ref Panel
Panel: SISU_2700
Kruskal-Wallis, p = 0.00000000053
Kruskal-Wallis, p = 0.00000000047
0.0046
2.6e-09
2.9e-11
0.25
9.5e-13
2.9e-11
SER
4.0
3.5
3.0
2.5
2.0
Constant rate map
Finnish map
HapMap
Constant rate map
Finnish map
HapMap
Condition


No RefPanels(Phasing) + SISU_2700(Imputation)
<0.1%
0.1%-0.25%
0.25%-0.5%
0.5%-1%
>1%
Concordance
1.00
0.75
0.50
0.25
0.00
Constant rate map
Finnish map
HapMap
Maps used
Map
Constant rate map
Finnish map
HapMap


With RefPanels(Phasing) + SISU_2700(Imputation)
<0.1%
0.1%-0.25%
0.25%-0.5%
0.5%-1%
>1%
Concordance
1.00
0.75
0.50
0.25
0.00
Constant rate map
Finnish map
HapMap
Maps used
Map
Constant rate map
Finnish map
HapMap