Inter-Chromosomal Linkage Reveals Errors in the 1000 Genomes Dataset

Fabrizio Mafessoni[1,*], Rashmi B Prasad[2], Leif Groop[2,3], Ola Hansson[2], Kay Prüfer[1,*]

[1] Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany
[2] Department of Clinical Sciences, Diabetes and Endocrinology, Lund University Diabetes Center, Malmö 20502, Sweden
[3] Finnish Institute for Molecular Medicine (FIMM), Helsinki University, 00014 Helsinki, Finland
[*] Correspondence should be addressed to: fabrizio_mafessoni@eva.mpg.de, pruefer@eva.mpg.de

Abstract

It is often unavoidable to combine data from different sequencing centers or sequencing platforms when compiling datasets with a large number of individuals. To measure quality differences between individual genomes and to identify erroneous variants, we study pairs of variants that co-occur in individuals. Applying our method to the 1000 Genomes dataset, we find that exonic regions are enriched for errors, where about 1% of the higher-frequency variants are predicted to be erroneous.

Main text

Next generation sequencing technologies allowed for the generation of datasets that include genetic data of a large number of individuals. To produce these datasets, sequencing data of different coverage and from different platforms or different batches of sequencing chemistry may need to be combined. This can result in differences in the types of errors and different amounts of errors across samples[1-5].

Here, we introduce a method to identify individual genomes with a higher error rate in large datasets and to predict which variants are likely due to error. The method first tests pairs of variants that reside on different chromosomes for signals of linkage disequilibrium. Linkage between separate chromosomes is not expected by population genetics theory for a randomly mating population, unless strong epistatic interactions are present. However, such signals can occur if errors affect individual genomes differently, leading to co-occurring erroneous variants on different chromosomes in the same individuals (Fig. 1). This first step of testing pairs is computationally intensive and we therefore limited the computation of linkage to pairs of variants in a subset of the genome. In the second step, we compare the contribution of individual genomes to the total linkage

signal to identify outlier individuals that carry more potentially erroneous variants. As a last step, we use the differences in the number of linked pairs between individuals to test which variants are present primarily in those individuals that carry more predicted errors (Fig. 1). This last step can be applied to all variants, not only those that have been tested for linkage, resulting in a list of predicted erroneous variants for the complete dataset. Removing these errors, we repeat the procedure starting from the second step until no significant differences in the burden of predicted errors is observed between individuals. No knowledge of differences in sequencing technologies or other factors are required by this approach.

We applied our method to the widely used 1000 Genomes dataset[6]. The 1000 Genomes data have been acquired over 7 years, involving 10 sequencing centers, 5 sequencing technologies, and several platform versions[6,7]. Individuals also differ in sequencing coverage and whether additional exome sequencing data are included. We limited our analysis to populations with at least 95 unrelated individuals, resulting in a total of 12 populations that we were able to test. Since many individuals from the 12 populations contained data generated via exome capture, we considered for our analysis all rare and common variants in coding regions (>1% minor allele frequency (MAF); 107087 sites over all 12 populations) and, as a separate dataset, an equal number of rare and common intergenic variants. To minimize the influence of population substructure on our measure, we calculate inter-chromosomal linkage for each population independently.

For both the intergenic and exome datasets and for all populations, we observe an excess of linked pairs over the expected number at a false discovery rate of 5% or when comparing to an expectation derived from randomly assigning chromosomes to individuals (Fig. 2a, Supplementary Figs. 2-3). Linked pairs are often shared between populations, but this sharing does not reflect population relationships (Fig. 2b). However, much more significant links are discovered in the exome dataset compared to the intergenic dataset. In the exome we find that variants are often linked to other variants on several different chromosomes, leading to large clusters of paired-variants (Supplementary Fig. 12). Maintaining such large clusters would require implausible selection pressures that favor the co-inheritance of minor alleles (Supplementary Fig. 16-19). This contrasts with the concept of synergistic epistatic interaction among deleterious variants, that would lead to a repulsion between rare variants[8].

Next, we calculated the contribution of each individual to the overall signal of linkage in a population by summing over the estimated number of linked pairs of minor alleles this individual carries (called nAB)[9]. We then compared the distribution of nAB over the individuals in a given population to the distribution calculated after randomly assigning chromosomes to individuals. We found that the variance in nAB is over 80-fold higher in intergenic regions and over 100-fold higher in exomes compared to the expectation from randomization (Wilcoxon-rank test: intergenic p-

value<7.4×10^{-7}, exome p-value<10^{-20}), showing that the signal of linkage is driven by individuals that carry an excess of linked pairs. Interestingly, in the exome most populations show groups of individuals that have similar nAB values but differ from the values observed for individuals of other groups (Supplementary Figs. 5-6). Consistent with this observation, the nAB distribution in all populations fit a model of a mixture of several Gaussian distributions significantly better than a model with just one Gaussian distribution, and we use the fitted Gaussians to assign individuals to groups (Fig. 2c; Supplementary Fig. 4).

To explain the differences in nAB between individuals for the exome dataset, we correlated the group assignment of individuals from all populations with technical features of individuals annotated as part of the 1000 Genomes dataset. We found that both coverage and sequencing center are significantly associated with differences in nAB (coverage: p-value<10^{-18}; sequencing center: p-value<10^{-20}). While both factors are also found significant in most populations, the sequencing center explains more of the variation in nAB (likelihood ratio test p-value<10^{-12}) (Fig. 2d). We notice that alternative explanations are incompatible with the observed signal: for example, the possibility that polymorphic genetic rearrangements contribute to a large extent to the linkage signal is incompatible with the clustering of individuals according to their nAB, while population substructure would not generate the same linked variants across different populations. We found no significant correlation with technical features for the intergenic dataset (Supplementary Fig. 7). Since the intergenic dataset contains an equal number of variants as the exome but much fewer linked variants, this negative result is likely due to low power.

We next searched for variants where a minor allele is preferentially encountered in individuals with a high nAB value. Genome-wide we identify 4445 variants (>1% MAF) in the 1000 Genomes data that are significantly associated with nAB and constitute our set of error candidates. Interestingly, these candidates are not distributed randomly over the genome, but are enriched in the exome, where around 1143 variants (~1%) are predicted to be errors (Supplementary Table 1). In comparison, in intergenic regions only a small fraction (<0.001%) were labeled as candidates, even when more variants are sampled to increase power in the prediction (see Supplementary Material). To further test the enrichment in coding regions, we used the software *admixture*[10], which labels individuals by components of ancestry, on variants in coding regions and on all variants genome-wide. Coding regions showed components that corresponded to the grouping of individuals by nAB, while such an effect was not observed for intergenic variants, suggesting that variants in coding regions are enriched for error (Supplementary Fig. 8).

We would expect that our predicted errors are shared less often than real variants with datasets produced at high coverage, by different technologies. To test this prediction, we compare how often our candidate variants are found in the genomes of 69 individuals generated by Complete

Genomics and compare this number to the sharing of other frequency matched variants from the 1000 Genomes dataset. While on average 86% of the matched variants in exomes are found in the Complete Genomics datasets, only 18% of our candidates are shared (chi-square test p-value$<10^{-15}$). In intergenic regions 85% of variants match on average while only 29% of candidates are shared (p-value$<10^{-15}$). This suggests that around 80% of our predicted variants in exons and 66% of intergenic variants are indeed due to error, assuming conservatively that Complete Genomics is devoid of errors that are shared with the 1000 Genomes dataset[3]. We also tested how often our exome candidates show evidence for allelic imbalance in high-coverage genomes from the 1000 Genomes dataset and find strong enrichment (chi-square test p-value $<10^{-6}$).

Our method predicts errors that likely originate from the combination of technologies, but should not find errors when the dataset is generated with just one technology and does not contain other batch effects. To test how many false positive errors we predict, we applied our method to such a uniform dataset, containing 654 unrelated individuals from the Botnia region[11]. The dataset was produced as part of a diabetes genome-wide association study using OmniChip. Our method found no excess of linkage in this cohort and the distribution of nAB across individuals is comparable to that observed after randomly permutating chromosomes across individuals.

Previous studies used local patterns of linkage disequilibrium (LD) in order to improve the quality of haplotype and SNP-calls in large-scale studies[12,13]. An example is the fastPhase method, that allowed for the identification of over 1,500 low frequency SNPs with high error rates in the HapMap datasets[12]. Our method uses a different source of information and can be combined with these approaches to predict errors.

Here we have shown that long-distance linkage between pairs of sites that reside on different chromosomes can be used to predict individuals that show an excess of error and to label variants that are likely errors. Our results show that these errors affect estimates of population demography (Supplementary Fig. 8) and estimates of mutational load (Supplementary Fig. 11), and we anticipate that the identification of strong batch effects will be crucial for the study of epistatic interactions.

Methods

Data handling

We downloaded the 1000 genomes phase 3 dataset (version 6/25/2014). One representative individual was sampled for each set of related individuals using the 1000 genomes annotation. Only populations with at least 95 unrelated individuals were analysed further, retaining 12 populations and 1117 individuals (Table S1). Variants not marked as PASS in the FILTER field were excluded. The remaining variants were classified according to frequency using bcftools (common variants:

>5% frequency, rare variants: 1% < frequency <= 5%)[14] . Variants were annotated as coding when they fell within the coding exons of the UCSC known gene annotation[15], and as intergenic when they did not overlap UCSC known genes and were not annotated as a potentially functional variant by the Variance Effective Predictor[16]. The Botnia data include 327 trios from the Botnia population, in Finland[11]. We excluded all offspring and related individuals.

Outline of pipeline

We implemented our analyses in a pipeline to detect inter-chromosomal linkage disequilibrium and detect variants affected by batch effects or inhomogeneity in the treatment of samples. This pipeline is outlined in Fig. S1, and the different steps are described in the following sections.

Step-1: Linkage Disequilibrium

When the phase is unknown, as for two physically unlinked loci A and B with possible alleles A-a and B-b, respectively, a composite genotypic linkage disequilibrium can be calculated, by relying on a maximum likelihood estimate of the amount of AB-gametes that are present in samples. Following Weir et al, 1996[21] , we can arrange the counts of the nine possible observed genotypes for the two loci in a matrix:

A \ B	B/B	B/b	b/b
A/A	n1	n2	n3
A/a	n4	n5	n6
a/a	n7	n8	n9

so that

$$n_{AB}=2n_1+n_2+n_4+n_5/2. \quad (1)$$

The composite genetic disequilibrium equals $D_{AB} = n_{AB}/n+2p_Ap_B$ where n is the number of samples. The sign of the composite linkage disequilibrium D_{AB} indicates whether alleles A and B (or a and b) occur preferentially in combination (D_{AB}>0) or whether the alleles occuring most often together are A and b (or a and B) (D_{AB}<0). We can test statistical association between two variants by either considering a two-tailed test (i.e. Fisher's exact test, adopting normalization proposed by Kuliskaya et al., 2009[18]) or by performing a 1-tailed Fisher's exact test for the positive and negative associations between minor alleles, thus denoted as A and B.

In order to speed up calculations approximate p-values were first determined with the chi-square based T2 method[9,19,20], and exact p-values were calculated only for those pairs with an approximate p-value < $100/n_{SNP}^2$, where n_{SNP} is the total number of variants examined. While

negative association between variants might occurs because of synergistic interaction between deleterious variants[8], batch effects are expected to result in the positive association between rare variants (Fig. 1). Thus, we restricted our analyses to significantly linked variants with a positive association.

We combine p-values across populations, using Fisher's method to obtain a single combined 1-tailed p-value for each pair of variants. These combined p-values are then compared to those obtained in a dataset generated by randomly redistributing chromosomes across individuals. The False Discovery Rate was calculated as the fraction of allele-pairs that have an equal or lower p-value in the randomized dataset, versus the original data. In order to test the excess of inter-chromosomal linkage disequilibrium we restricted further analyses to instances in which at least one pair of variant is significantly associated (Fig. S2-S3).

Step2: Individual contribution to LD

We calculated the contribution of each sample/diploid individual to the linkage signal by summing up its contribution to the nAB value of significant linkage pairs. For positive associations (D>0), the contribution of each sample is directly the weight in equation (1) corresponding to a specific genotype configuration, e.g. 2 if a sample has genotype n1 (A/A, B/B) since both gametes necessarily had alleles A and B on both chromosomes, and 0 if it has genotype n3 (A/A, b/b), since no gametes had alleles A and B on both chromosomes.

In order to test whether samples contribute uniformly to the linkage signal or not we compared the variance of the observed data and random reshuffling of the chromosomes across samples and within populations. To identify samples with similar features, we perform a finite mixture analysis (R package Mixtools), by assuming a normal distribution for the underlying model of the contribution to the linkage signal of each group of samples, within a population. We calculated the corrected Akaike Information Criterion (AICc) weights for each model, with 1 to 9 possible underlying Gaussians for each population.

The models with the highest weights are shown in Fig. S4. Since a Gaussian might deviate from the real underlying distribution, we tested whether a finite mixture analysis on the null datasets in which chromosomes are redistributed across individuals would provide less support to the presence of groups of samples with different nAB. We first calculated the relative support for the best supported model against the null model with only one Gaussian to explain the data, and compared it with the same statistics for 100 null datasets. Higher support for multiple clusters was present in the observed data compared to the null distribution (Wilcoxon rank test, p-value$<10^{-16}$ for coding regions, p-value=0.00103 for intergenic). Note that although generating permutations of the data is computationally expensive, the high number of potential links give a very narrow

distribution of all statistics related to this empirical null distribution. For the 1000 genomes dataset 3 randomizations are sufficient to provide highly significant p-values when adopting a one-tailed t-test and comparing to the real data. When a dataset showed a significantly higher variance and higher clustering than its empirical null distribution in terms of nAB, we proceed into identifying the variants responsible for the signal and the sources of the bias.

In order to assess whether the identified clusters correspond to specific features of the samples we tested the role of several technical predictors extracted from the sample spreadsheet of the 1000 genomes dataset (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130606_sample_info/20130606_sample_info.xlsx). For exonic regions, the clusters identified with Mixtools for the coding regions of the 1000 genomes are significantly associated with Sequencing Center (combined chi-square p-value$<10^{-15}$) (Fig.S5-S6). We also compared the variance in the mean coverage of the exome per cluster with 10,000 permutation of the samples across the clusters, and not a single permutation displayed higher variance than the original data.

To directly assess the association between nAB and technical features of the samples, we applied a Generalized Linear Mixed Model using the R package lme4. We considered as response variable the observed nAB value for each sample and the features of the individual samples as predictors. Populations were considered as random predictor.

In order to identify variants whose presence is explained by the occurrence in specific clusters of samples, we used a Generalized Linear Mixed Model (R package lme4) iteratively on each variant, considering the contribution of each sample to nAB as a predictor. The underlying assumption is that samples that show a consistent excess of linked variants are more likely affected by technical artifacts. Hence, variants that are present only in these samples would be more likely spurious. To assess whether the presence of an allele is predicted by the contribution of each individual to the linkage signal, we built two models: a full model, including the nAB value for each individual as a predictor, and a null model, in which nAB is not included. We then compared the two models with a likelihood ratio test, so that for each variant we assess the significance of the relationship between nAB and the presence of the minor variant (Fig. S10).

In more detail, the response variable of the linear models is the presence or absence of the minor allele per sample. A sample can have three states for this minor allele (absent homozygous a/a, heterozygous a/A or present homozygous A/A). To model non-independence of the two chromosomes of each individual we consider each allele separately and introduce a predictive variable (factor) “Sample” that groups the two alleles for each chromosome of each individual (Fig. S9). The other predictor that is present in both the full and the null model is the population to which each individual belongs. In the full model, the contribution to the linkage signal per sample (nAB),

is also present as a continuous covariate.

We can thus write the two models as:

- full model: presence allele A at site i ~ contribution nAB + (1| Population) + (0 + contribution nAB | Population) + (1|Sample).
- null model: presence allele A at site i ~ (1| Population) + (1|Sample).

Sample and Population were introduced as two random factors (categorical random predictors), in order to control for the non-independence between chromosomes belonging to the same individual and individuals belonging to the same population. The effects of random factors are denoted as (1| Factor). The effects of a covariate, when dependent on a random factor, are denoted as (0 + covariate | Factor). In particular, we allowed for different effects of nAB in different populations (0 + contribution nAB | Population), due to potential differences in population composition and treatment.

The contribution of each sample to the nAB signal has been z-transformed and the p-values of the likelihood ratio test are corrected for multiple testing with the Benjamini-Hochberg criterion. In order to speed up calculations we preliminarily scanned each sample with an analogous simpler logistic model, in which random factors are neglected, and populations are considered independently. The p-values of each population are combined with Fisher's method, and the full model including all random components was performed only for variants for which the combined p-value was below 0.01.

We applied our method to all variants present in the 1000 genomes datasets. For the exonic regions, where linkage pairs are abundant, we directly use the nAB values estimated exclusively on significantly linked variants (Table S1). This procedure is under-powered for intergenic variants, where the amount of linked pairs is much smaller and the distribution of nAB has a low resolution. To increase the amount of bona-fide linked variants in the intergenic dataset, we first increased by 10-fold the number of pairwise inter-chromosomal comparisons by subsampling a larger amount of intergenic variants. In addition, we relax the FDR cutoff to define linked pairs to FDR<20%. Notice that while this reduced cutoff may increase the noise in the nAB profile due to additional randomly linked pairs, we expect no systematic bias that would increase the number of predicted errors. Consistent with an increase in power, we observe approximately 4-times more predicted variants with these more permissive parameters for intergenic linked pairs (Table S2). In contrast, increasing the FDR cutoff has only a minimal effect on the number of predicted errors in exons, suggesting that the estimation of the nAB profile for the exome is not under-powered.

For both datasets, we estimated the false discovery rate for each variant with the Benjamini-Hochberg method (Table S1 and Table S2). We report also an empirical false discovery measure, based on the overlap of the candidate variants from the 1000 genomes dataset to variants present in

Complete Genomics (Table S1-S2 and Supplementary Figure S13-S14). Significant variants for both datasets show a reduced overlap with the high quality Complete Genomics dataset, indicating an enrichment in error among our candidates.

For the exome dataset, we calculated linkage between all pairs of SNPs. When we apply our model only to those exonic variants that already showed significant linkage, we notice that we are able to increase the number of predicted errors (1429 while retaining a similar false positive rate based on the overlap with variants present in Complete Genomics (21.1%) (Supplementary Figure S15). This indicates that the linear model alone is underpowered and that looser cutoffs could be potentially used.

In order to further disentangle the effect of the two main technical features of the data (coverage and sequencing center) we repeated the same approach by substituting the factor "contribution nAB" with the continuous variable coverage and the random predictor sequencing center, respectively, as shown in the main text.

When computing linkage disequilibrium in all datasets after candidate variants are removed, the association between nAB and batches (sequencing center or coverage) is not detectable anymore (p-value > 0.05) (Fig. S10).

Effects of selection

In order to illustrate the possible selection scenarios that could lead to inter-chromosomal linked variants we calculated the dynamics in time of the average linkage-disequilibrium coefficient D, in presence of epistatic interactions between two different genomic variants leading to a difference in survival rates of the different gametes. We consider two biallelic sites, with alleles A-a, and B-b, respectively. We denote the number and selection coefficient of gametes AB,Ab, Ab and aB with n_{AB},n_{Ab},n_{aB} and n_{ab}, and s_{AB},s_{Ab},s_{aB} and s_{ab}, respectively. We performed for each selection scenario 10000 simulations. In each generation, we first simulated recombination, then selection.

In the recombination step, we sampled the number of gametes that would change state (i.e. gametes AB recombining with ab, and Ab recombining with aB) after random pairing of the gametes and recombination occurring with probability r. The selection step follows a Wright-Fisher model, with each gamete having fitness $1+s_{AB}$,$1+s_{Ab}$,$1+s_{aB}$ and $1+s_{ab}$, respectively. We consider two possible scenarios: advantageous combinations of minor alleles, with $s_{AB}>0$ and $s_{Ab}=s_{Ab}= s_{ab} =0$, and antagonistic combinations of minor and major alleles (with $s_{AB}=s_{ab}=0$ and $s_{Ab}=s_{aB}<0$). Selection coefficients were either fixed to 1% (strong selection) or sampled from the distribution of selection coefficients estimated for non-synonymous variants in Racimo et al.,2014[22] .

Code and Data Availability

We implemented the genome-wide scan in C, and the logistic models adopted for the identification of biased SNPs in R. The code and tables with the predicted erroneous variants in the 1000 Genomes are available at http://bioinf.eva.mpg.de/LDLD/.

Figures

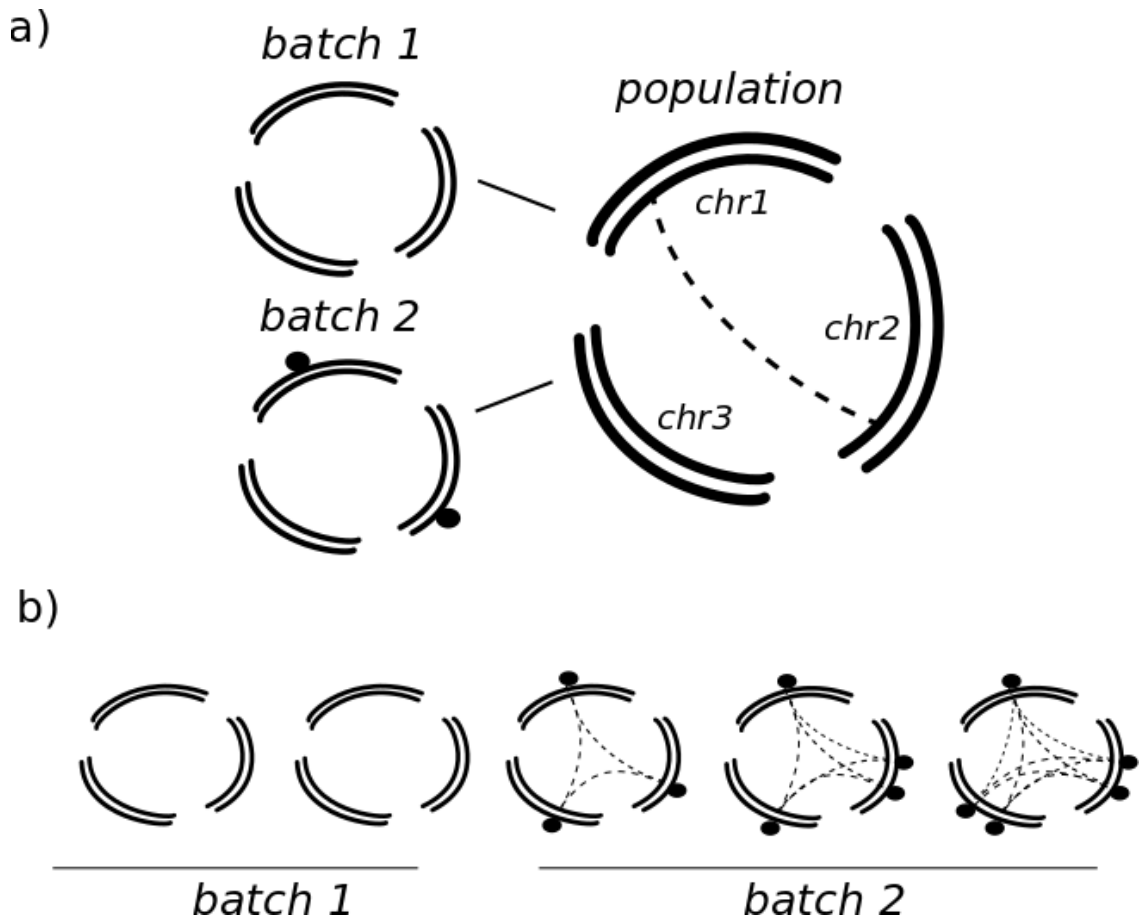

Figure1: Outline of the method. a) Sequencing data generated from samples with different sequencing quality or processing might introduce different errors (black dots). Since these errors will be present in samples coming from the same platform, they will give a signal of linkage between different chromosomes (dashed lines). b) The contribution to the linkage signal can be computed for each sample (dashed lines), and used to identify samples coming from the same batch and with similar error profiles, as well as the errors.

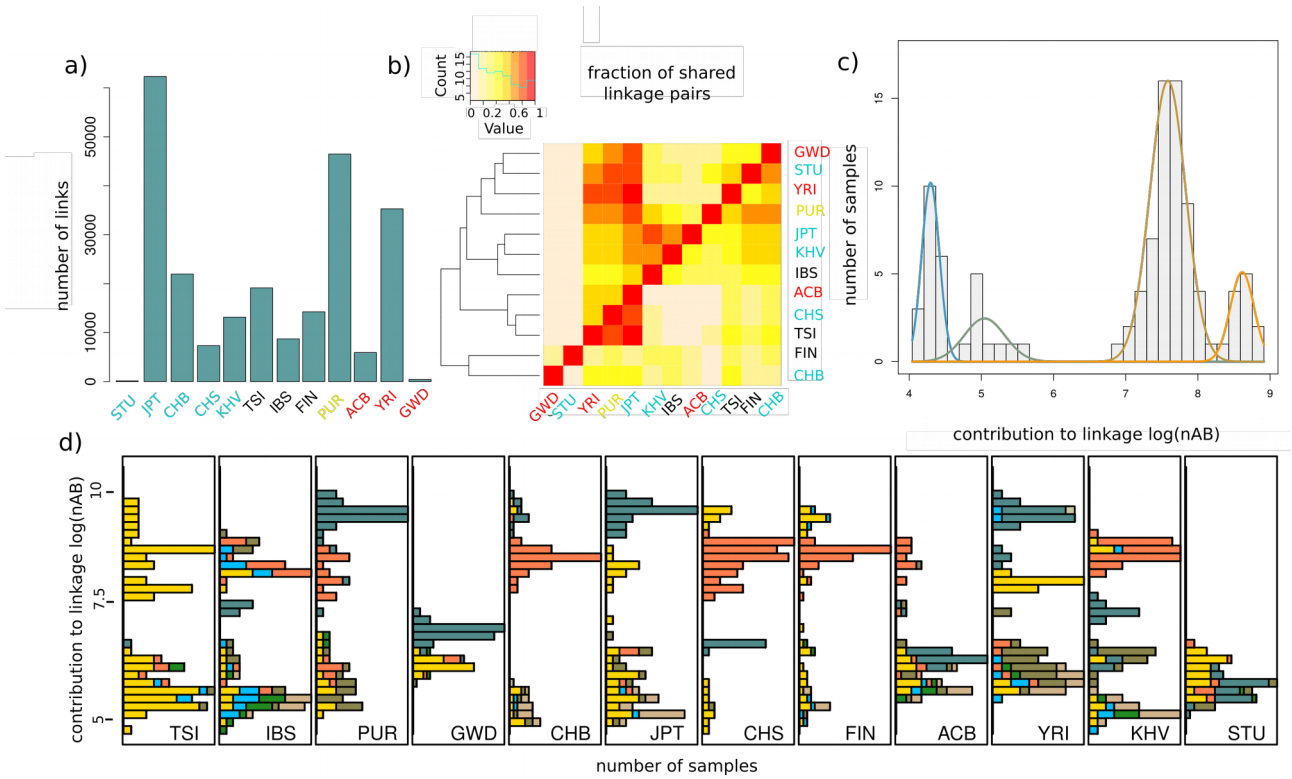

Figure 2. Characteristics of inter-chromosomal linkage. a) Number of inter-chromosomal linkage pairs with a false discovery rate (FDR) < 5% in exomes of the 1000 Genomes populations. The FDR was calculated by comparing the p-value of each linked pair to the distribution of p-values after permuting chromosomes across individuals. Populations labels are colored according to the continent: blue for Asia, red for Africa, black for Europe and yellow for others. b) Fraction of exome inter-chromosomal linked pairs in one population (row) that are also linked in another population (column). c) Contribution of Chinese from Bejing individuals to the linkage signal (bars) in exomes given by the number of linked minor alleles (nAB). Individuals with similar nAB values were grouped by a Gaussian mixture model, whose fitted distributions are shown as colored lines. d) Distribution of nAB in exomes for individuals from different 1000 Genomes populations. Colors indicate the sequencing center per individual. Individuals sequenced in multiple centers were marked with a separate colors.

References

1 Kircher, M., Heyn, P. & Kelso, J. Addressing challenges in the production and analysis of illumina sequencing data. *BMC Genomics* **12**, 382, doi:10.1186/1471-2164-12-382 (2011).

2 Wolpin, B. M. *et al.* Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer. *Nat Genet* **46**, 994-1000, doi:10.1038/ng.3052 (2014).

3 Schirmer, M. *et al.* Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res* **43**, e37, doi:10.1093/nar/gku1341 (2015).
4 Torkamaneh, D., Laroche, J. & Belzile, F. Genome-Wide SNP Calling from Genotyping by Sequencing (GBS) Data: A Comparison of Seven Pipelines and Two Sequencing Technologies. *PLOS ONE* **11**, e0161333, doi:10.1371/journal.pone.0161333 (2016).
5 Wall, J. D. *et al.* Estimating genotype error rates from high-coverage next-generation sequence data. *Genome Res* **24**, 1734-1739, doi:10.1101/gr.168393.113 (2014).
6 The Genomes Project, C. A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).
7 Consortium, T. G. P. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65, doi:10.1038/nature11632 (2012).
8 Sohail, M. *et al.* Negative selection in humans and fruit flies involves synergistic epistasis. *Science* **356**, 539-542, doi:10.1126/science.aah5238 (2017).
9 Schaid, D. J. Linkage disequilibrium testing when linkage phase is unknown. *Genetics* **166**, 505-512 (2004).
10 Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*, doi:10.1101/gr.094052.109 (2009).
11 Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* **536**, 41-47, doi:10.1038/nature18642 (2016).
12 Scheet, P. & Stephens, M. A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. *Am J Hum Genet* **78**, 629-644 (2006).
13 Leek, J. T. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res* **42**, e161, doi:10.1093/nar/gku864 (2014).
14 Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987-2993, doi:10.1093/bioinformatics/btr509 (2011).
15 Rosenbloom, K. R. *et al.* The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res* **43**, D670-D681, doi:10.1093/nar/gku1177 (2015).
16 McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122, doi:10.1186/s13059-016-0974-4 (2016).
17 Reich, D. *et al.* Denisova Admixture and the First Modern Human Dispersals into Southeast Asia and Oceania. *The American Journal of Human Genetics* **89**, 516-528, doi:10.1016/j.ajhg.2011.09.005 (2011).
18 Kulinskaya, E. & Lewin, A. Testing for linkage and Hardy-Weinberg disequilibrium. *Annals of Human Genetics* **73**, 253-262, doi:10.1111/j.1469-1809.2008.00501.x (2009).
19 Zaykin, D. V., Pudovkin, A. & Weir, B. S. Correlation-Based Inference for Linkage Disequilibrium With Multiple Alleles. *Genetics* **180**, 533-545, doi:10.1534/genetics.108.089409 (2008).
20 Wu, X., Jin, L. & Xiong, M. Composite measure of linkage disequilibrium for testing interaction between unlinked loci. *Eur J Hum Genet* **16**, 644-651, doi:10.1038/sj.ejhg.5202004 (2008).
21. Weir, B.S., *Genetic Data Analysis II*. 1996.
22. Racimo, F. and J.G. Schraiber, *Approximation to the Distribution of Fitness Effects across Functional Categories in Human Segregating Polymorphisms.* PLOS Genetics, 2014. **10**(11): p. e1004697.