3D chromatin loops measured with Hi-C bring together SNP-SNP pairs engaging in epistatic interactions in GWAS data

Supplementary Information

Table of Contents

A. Statistical Appendix: Logistic Regression and Likelihood Ratio Test 2

A1. Regression models 2

A.2 Regression in GWAS 2

A.3 Likelihood-ratio test (LRT) and Wilks' theorem 3

A.4 In this work 4

B. GWAS preparation: Sample and Site quality control (QC) 5

B.1 Sample QC 5

B.2 Site QC 5

B.3 Association 5

B.4 PCA (Principal Component Analysis) 6

B.5 Imputation and Dosage Convertor 7

C. Supplementary figures 8

GWAS QC: intermediate Steps 8

HiCCUPS (Hi-C Computational Unbiased Peak Search) 9

Loops are enriched for enhancers and common SNPs, especially when highly associated with the phenotype. P-values. 10

SNPs sitting on the same looping region show similar association with the phenotype 10

Epistasis occurs with great diversity from chromosome to chromosome 11

Percentages of pairs passing the different tests 12

KEGG Pathways: we found epistatic pairs between SNPs sitting on or in the vicinity of genes involved in these pathways 13

Future direction: from SNP pairs to SNP groups 14

D. Boxes: main concepts illustrated 15

D.1 Genome-wide association studies (GWAS) 15

D.2 Linkage disequilibrium (LD) 15

D.3 Epistasis 16

D.4 The three-dimensional genome organization 16

D. 5 Chromosome Conformation Capture (3C) 17

References 17

A. Statistical Appendix: Logistic Regression and Likelihood Ratio Test

A1. Regression models

Regressions are statistical models that aim to estimate the relation between variables. Linear regressions are regression that model the relationship between a scalar variable y, which is the dependent variable, and one or more independent variables x_i, which are called predictors. With a linear regression we aim to explain y as a linear combination of a number n of x_i's:

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \varepsilon$$

The coefficients β_i's weigh the predictors, indicating the size of the influence that each predictor has on y. ε is the error, which follows a Gaussian distribution. Thus, the conditional probability $y \mid x$ follows a Gaussian distribution.

Linear regressions are widely used in biological and social sciences, as well as in finance and economics. There exist many extensions of the linear regression models that are for example able to model non-linear relations between the variables. We refer to these models as generalized linear models (GLM). One type of GLM is called logistic regression. If linear regressions model a scalar continuous variable, logistic regressions model a binary dependent variable y. The conditional probability $y \mid x$ therefore follows a Bernoulli distribution. A logistic regression model can be written similarly to a linear one:

$$y^* = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

However, y^* is not directly the outcome y, but rather a transformation (g) of it:

$$y^* = g(y) = \ln(\frac{y}{1-y})$$

Therefore, to explicitly derive the dependent variable y:

$$\ln(\frac{y}{1-y}) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

$$\frac{y}{1-y} = e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}$$

$$y = \frac{e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}}{e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i} + 1} = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)}}$$

The last formula, with y as a function of $(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)$ is the definition of the logistic function, thus the name logistic regression. A property of this function is that the outcome is restricted to be between 0 and 1, and can therefore be interpreted as the probability of the dependent variable being 1.

A.2 Regression in GWAS

Among other applications, linear and logistic regression models are used in genome-wide association studies (GWAS) to model the state of individuals, in relation to a genetic disease or trait, as a function of genetic variables. To be more specific, GWAS aim to find an association between having common genetic variants, typically in the form of single nucleotide polymorphisms (SNPs) and presenting a certain trait or disease. In a regression model, therefore, we model the phenotypic state using the SNPs as predictors. We shall now distinguish between traits and diseases. Traits are often continuous, think for example of somebody's weight, height or body mass index (BMI). Such traits can be modelled using a linear regression, where simply y is the value for a given trait (height in cm, weight in kg and so on). For diseases, on the other hand, we typically distinguish between having a disease or not having it. Thus, a logistic regression is better suited. In a GWAS setting, we definitely refer to y as to the outcome (affected/unaffected). We model the outcome of every individual as a combination of the genotypes of that same individual at the same given location (SNP).

Logistic regression models are not the only association measures used in GWAS experiments. The most traditional approaches are based on a contingency table: those are the χ^2 test and the Fisher's exact test. A standard GWAS experiment is performed in a case-control setup. Individuals with the disease of interest (cases) are compared with healthy individuals (controls) on the basis of the respective genotypes at specific positions. For each locus, the contingency table is:

	AA	Aa	aa
cases			
controls			

Where A and a are the two possible alleles at that position, typically classified as minor and major allele (more or less frequent in the studied population). Recently, studies have preferred logistic regression models over the contingency table methods, for mainly two reasons (Clarke et al, 2011):

1. Regression models are naturally well-suited to incorporate many variables. By adding as covariates potentially confounding factors (age, sex, ethnicity..) it is possible to correct for intrinsic variation in the population we are studying.

2. The interpretation of the contribution of each SNP to the outcome prediction is straight forward, it is sufficient to look at the estimated coefficients

There are several ways of measuring the goodness of fit of a logistic regression model. Those include the Wald statistic, which assesses the contribution of each individual predictor and the likelihood ratio test, which assesses the goodness of fit of a given model, when compared to the null model (model with no predictors). However, LRTs go beyond that, and are able to compare any two regression models, when one is a special case of the other. LRT is discussed in further detail in the next paragraph.

A.3 Likelihood-ratio test (LRT) and Wilks' theorem

The likelihood ratio test is a statistical test meant to compare the goodness of fit of two models, one being a special case of the other. For a linear model (or a generalized linear model, e.g. our logistic regression), the common practice is to compare the model of interest with either the null model (only intercept, no predictors) or on the other extreme to the full or saturated model (all predictors available).

For example, in formulas, if I have 10 predictors in total, and I want to measure how well the first three predictors alone perform, I want to compare the model:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

With the null model: $y_i = \beta_0$

Or the full model: $y_i = \beta_0 + \sum_{i=1}^{10} \beta_i x_i$

In the first case, I am measuring how much those three characteristics are able to predict, compared to not having any information on the samples. In the second, I measure how much those three factors weigh in determining the outcome, out of all the available information.

However, one does not necessarily need to compare a given model to the full model, or to the null one. As long as they are nested, any two models can be compared using LRT. One might be interested to see how much adding a fourth factor improves the prediction: an LRT could answer that, when comparing the models with 3 and 4 predictors.

For every two models having as set of parameters θ_0 and θ, with $\theta_0 \subset$ θ, the likelihood ratio statistic is:

$$\Lambda = \frac{\sup\{L(\theta|x), \theta \in \theta_0\}}{\sup\{L(\theta|x), \theta \in \theta\}}$$

To obtain a significance measure of the improvement, basically a p-value out of the likelihood ratio, we can use the Wilk's theorem. For a very large number of samples, $n \to \infty$

$$-2 \ln \Lambda \sim \chi^2(k) \qquad (1)$$

Where the number of degrees of freedom k is $|\theta| - |\theta_0|$

(#) can also be written: $-2ln\frac{L_{reduced}}{L_{complete}} = -2[\ln(L_{reduced}) - \ln(L_{complete})] = 2[LL_{complete} - LL_{reduced}]$

The only values to fill in in the formula are therefore the log-likelihoods of the two model. For one logistic regression model

$$logit(p) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

Its log-likelihood is computed as follows:

$$LL = \sum_{i=1}^{n} y_i \ln(p_i) + (1 - y_i)\ln(1 - p_i) \qquad (2)$$

Where $y_i \in \{0,1\}$ are the true binary labels (sick/healthy)

And $p_i \in [0,1]$ are the predicted values, which can be interpreted as a probability of positive outcome (1=sick)

$$p_i = \frac{e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i}} = \frac{1}{1 + e^{-\beta_0 - \sum_{i=1}^{n} \beta_i x_i}}$$

It is therefore sufficient to compute the log-likelihood (2) of both models, insert them in (1), and we obtain a χ^2 value and therefore a p-value indicating how confident we are in refusing the (null) hypothesis that the two models provide the same level of information, in other words are an equally good fit of the data.

A.4 In this work

In this work, we use a GWAS case-control dataset, for type 2 diabetes (T2D). We use logistic regression models to model the outcome of the individuals tested based on a number of SNPs as predictors. To correct for population structure, we include in the models the first ten principal components (PCs). We also add the information about the sex of all individuals. Moreover, we are interested in the performance of pairs of SNPs, rather than individuals. The logistic regression setup allows to include those in an additive manner:

$$y^{*} \sim covar + SNP_1 + SNP_2$$

It is equally straight forward to include a term that measures the interaction between variants:

$$y^{*} \sim covar + SNP_1 + SNP_2 + SNP_1 SNP_2$$

To assess the goodness of fit of these models we use a likelihood-ratio test.

To assess whether a pair of SNPs performs better than the two individual SNPs in predicting the outcome, we measure (with LRT) both:

$$y^{*} \sim covar + SNP_1 + SNP_2 + SNP_1 SNP_2 \qquad vs \qquad y^{*} \sim covar + SNP_1$$

$$y^{*} \sim covar + SNP_1 + SNP_2 + SNP_1 SNP_2 \qquad vs \qquad y^{*} \sim covar + SNP_2$$

To measure if the interaction model is good overall we compare (with LRT):

$$y^{*} \sim covar + SNP_1 + SNP_2 + SNP_1 SNP_2 \qquad vs \qquad y^{*} \sim covar$$

Finally, to assess whether the improvement is due to the interaction of the variants, rather than just their sum, we calculate (with LRT):

$$y^{*} \sim covar + SNP_1 + SNP_2 + SNP_1 SNP_2 \qquad vs \qquad y^{*} \sim covar + SNP_1 + SNP_2$$

B. GWAS preparation: Sample and Site quality control (QC)

Raw genotypes were data generated for Burton et al. Nature 2007. Only one set of controls (NBS) was available (1500 samples). For the disease of choice, T2D, genotypes were available for 2000 samples. Several steps were taken to clean the raw data. Samples are tested for 500k SNPs, sitting on chromosomes 1-22 and X. All analyses described in the following paragraphs are performed using plink version 1.9 (Purcell et al 2006).

B.1 Sample QC

For sample QC (quality control), a ‘white list’ of sites is selected. Those are variants with extremely low missingness (missing<0.01), high minor allele frequency (MAF>0.5), not in LD (for every 50-SNP window, sliding by 5, all pairwise r^2 are computed, and if $r^2 < 0.5$ the SNP with higher missingness is filtered out). SNPs should not be of the form AT or CG. Such SNPs are furthermore not on chromosome X, and not on known variable regions (the lactase gene, LCT, 2:129883530-140283530; the major histocompatibility complex, MHC, 6:24092021-38892022, inversions on chromosomes 8 and 17, 8:6612592-13455629 and 17:40546474-44644684). After such filtering, the white list consists of 57,852 SNPs. These SNPs are ‘well-behaved’ SNPs, for which we have enough data (low missingness), which are not rare (the allele that appears less often, is present in at least 5% of the samples), which are rather independent from each other (not in LD) and are not on regions that are extremely variable in the population. Moreover, SNPs on chromosome X and Y are not included here, since they might introduce gender-related biases. This is not the set of SNPs we are going to analyse later on, but only an intermediate step for the sample quality control (QC). It is a precaution, so that the sample QC can truly be only about the samples, as we are only looking at extremely stable sites.

Based on this selection of variants, samples are filtered out when they have missingness higher than 5% and mismatched sex (genetic sex does not correspond to reported sex). As for relatedness, we remove all samples that are related with more than 100 other samples (coefficient of relationship or identical by descent IBD>0.125). After sample QC 3,343 samples are left (95.5%).

B.2 Site QC

Site QC is then performed only for the remaining samples, but reintroducing all variants from the start. The steps of site QC are the following: first, Hardy-Weinberg Equilibrium (HWE), with P<1e-13, either in controls only or overall. The Hardy-Weinberg Equilibrium is the notion for which, given a minor allele frequency of p, the probabilities of the three possible unordered genotypes (*a/a*, *A/a*, *A/A*) at a bi-allelic locus with minor allele *A* and major allele a, are $(1-p)^2$, $2p\left(1-p\right)$ and p^2, respectively. In a large, randomly mating, homogenous population, these probabilities should be stable from generation to generation (Clarke et al, 2011). Variants for which this is not the case are filtered out. Secondly, we filter on missingness rate (if < 0.05), both overall and differentiate between cases and controls. Finally, we only keep non-rare variants (MAF <0.05). All other sites are kept. After site QC 450,242 sites are left (90%).

B.3 Association

The association with the phenotype is measured with a logistic regression, using as covariates the first 10 principal components, and sex. No information about the age of the samples was available, nor batches information (or any other details on how the data was collected).

The results of a GWAS analysis are typically displayed as a Manhattan plot, along with a QQ plot. The first represents the SNPs as the p-values of their association with the phenotype, per chromosome. Every dot is the –log10(P) of a SNP. The higher up the dots, the higher the association of those variants with the disease of interest. Here, the only SNPs that are over the significance threshold sit on chromosome 10, which agrees with the results obtained in the original paper. A QQ plot compares, on the two axes, the observed (-log10) p-values, with the expected ones. In the figure we can see that the two mostly agree (most dots are on the diagonal) apart from a few ones that deviate from it, which are the highly associated ones.

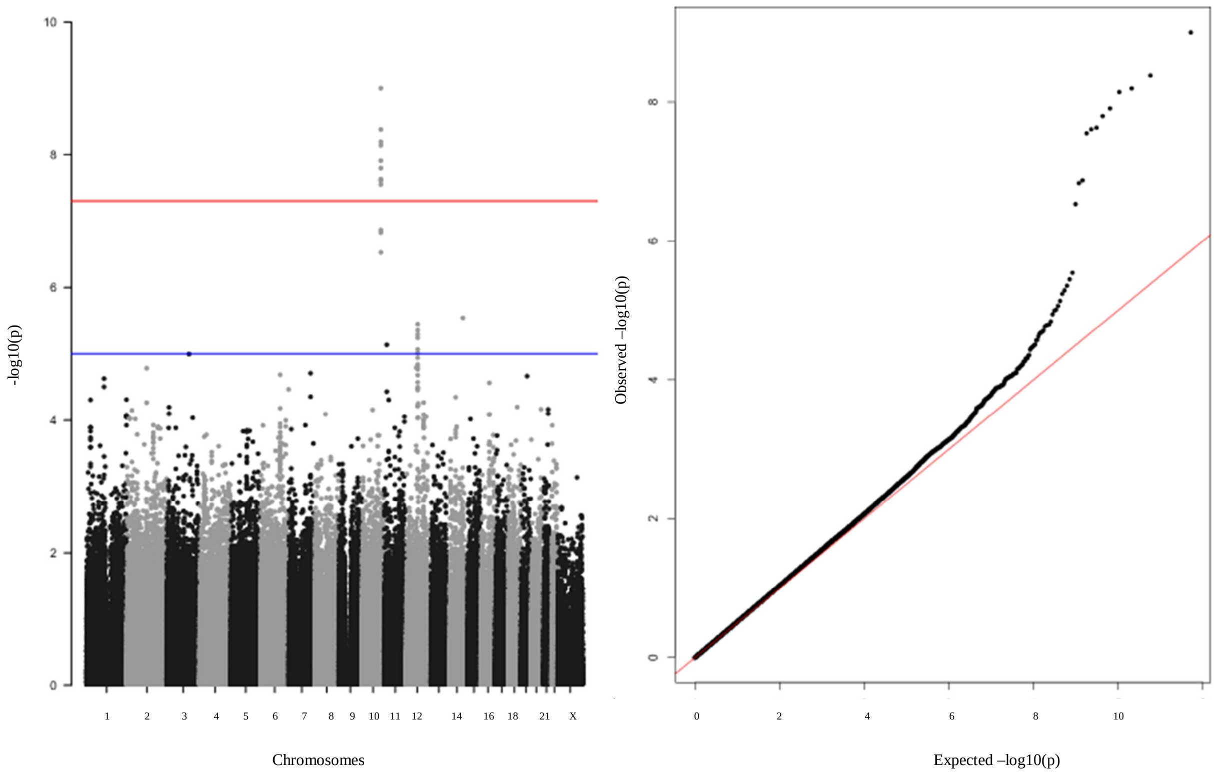


Fig S1. Results of the GWAS analyses on the WTCCC T2D dataset. The dataset is built with data for type two diabetes (T2D) compared to the controls (NBS). The **left** plot is a Manhattan plot. On the x axis are the chromosomes, in alternate colours. 23 indicates chromosome X. On the y axis, the negative log 10 of the p-values. Every dot is a SNP. In blue and red, significance thresholds. The only SNPs over the threshold are on chromosome 10. These results agree with those of the original paper. The **righ**t plot is a qq-plot, where the same negative log10 of the p-values for the observed data is compared to the same as it would be expected, if there was no association. Most sites (dots) are on the diagonal, with a few deviating from it showing some association which was not expected by chance.

B.4 PCA (Principal Component Analysis)

Principal component analysis is performed using EIGENSTRAT (eigensoft package eig4.2, converf, smartpca, eigenstrat), where the principal components (first 10) are computed on the reference downloaded froms HapMap, and then true samples are overlayed on it. No actual cleaning is performed based on the principal component analyses. However, the first 10 PCs are included as covariates in the logistic regression, which is the association measured used to calculate the pvalues in the Manhattan plot. Furthermore, by performing sample QC we eliminate inbred, related and other 'unreliable' samples, and this is visible in the PCA plot, where only the samples that fall in the European cluster are kept. This is to be expected as all collected samples are (supposably) of UK origins.

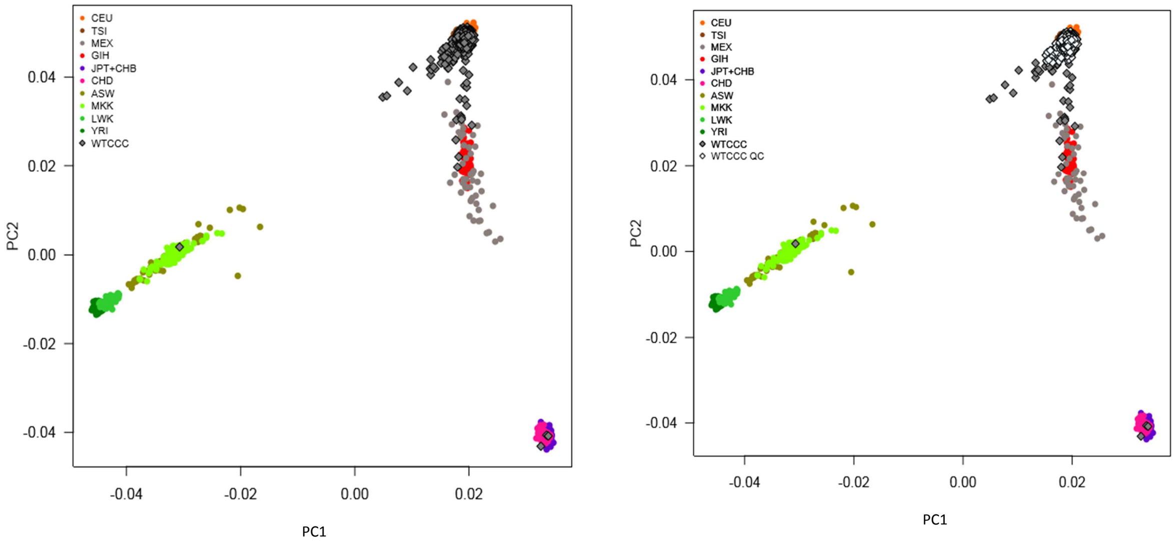


Fig S2: PCA plots. On the axis, the first principal component PC1, on the y axis, the second principal component PC2. In colours, the results for reference samples, from HapMap. Clearly, the two principal components capture the population structure of the samples. Four clusters are clearly noticeable: the elongated green one on the left represents the African samples, the circular purple one in bottom right represents the south Asian individuals, the orange top right the Europeans, the red elongated one on the right mostly mixed races. On top of the reference we overlay our WTCCC samples, in dark grey (**left** plot). Although officially all collected samples are of British origins, we observe same samples away from the Caucasian cluster. On the **right**, we highlight the samples that passed sample QC, in lighter grey. This clearly shows the importance of

this step, as the remaining samples nicely cluster to the top right, and the other ones have been filtered out.

B.5 Imputation and Dosage Convertor

Imputation is done uploading the (clean) data to the Michigan Imputation Server (http://imputationserver.sph.umich.edu). We imputed 109,126,218 new sites. We filter out on imputation quality ($r^2 > 0.3$). This filtering is done per chromosome and left us with 47,073,880 imputed sites (~40%).

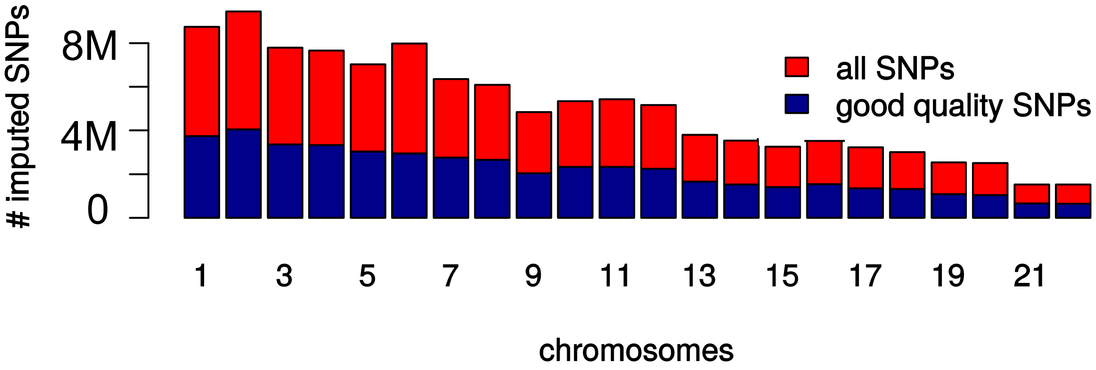


Fig S3. Percentage of imputed SNPs after quality filter ($r^2 > 0.3$). On the x axis, the chromosomes' numbers. On the y axis, the count of SNPs. In red is depicted the total number of imputed variants, per chromosome. For chromosome 1, for example, 8,740,001 variants were imputed. In blue, the number of SNPs we keep for further analyses, after filtering on imputation quality. The imputation quality is measured as a correlation measure (r^2) between imputed and true genotypes. We chose the rather lenient threshold of 0.3. For example, for chromosome 1, we kept 3,338,284 sites.

Genotypes are provided as dosages. Dosages are continuous numbers between 0 and 2, calculated as $2p(AA) + p(Aa)$, where A is the alternate allele, a is the reference allele. This indicates somehow the quality of the imputation. If the imputed genotype was certainly AA, p(AA) would be 1, p(Aa) and p(aa) would be 0, and the dosage would be exactly 2. Similarly, whenever one probability is 1 and therefore the other two are 0, the dosage and the imputed genotype (0,1,2) coincide. However, in some cases the imputing server is not as sure, and the dosages are not integers.

However, since we already filtered on imputation quality, those dosages are all quite close to the imputed results. Therefore, for simplicity, we convert dosages into discrete genotypes 0,1,2 for further analyses. First however we make sure that the difference is never larger than 0.1, and filter out SNPs for which this is not the case.

C. Supplementary figures

GWAS QC: intermediate Steps

Figure S4 shows some intermediate GWAS results (Manhattan and qq plots) at different stages of quality control (QC). Essentially, we show some of the steps between Fig 4A and Fig 4B (and Fig S1).

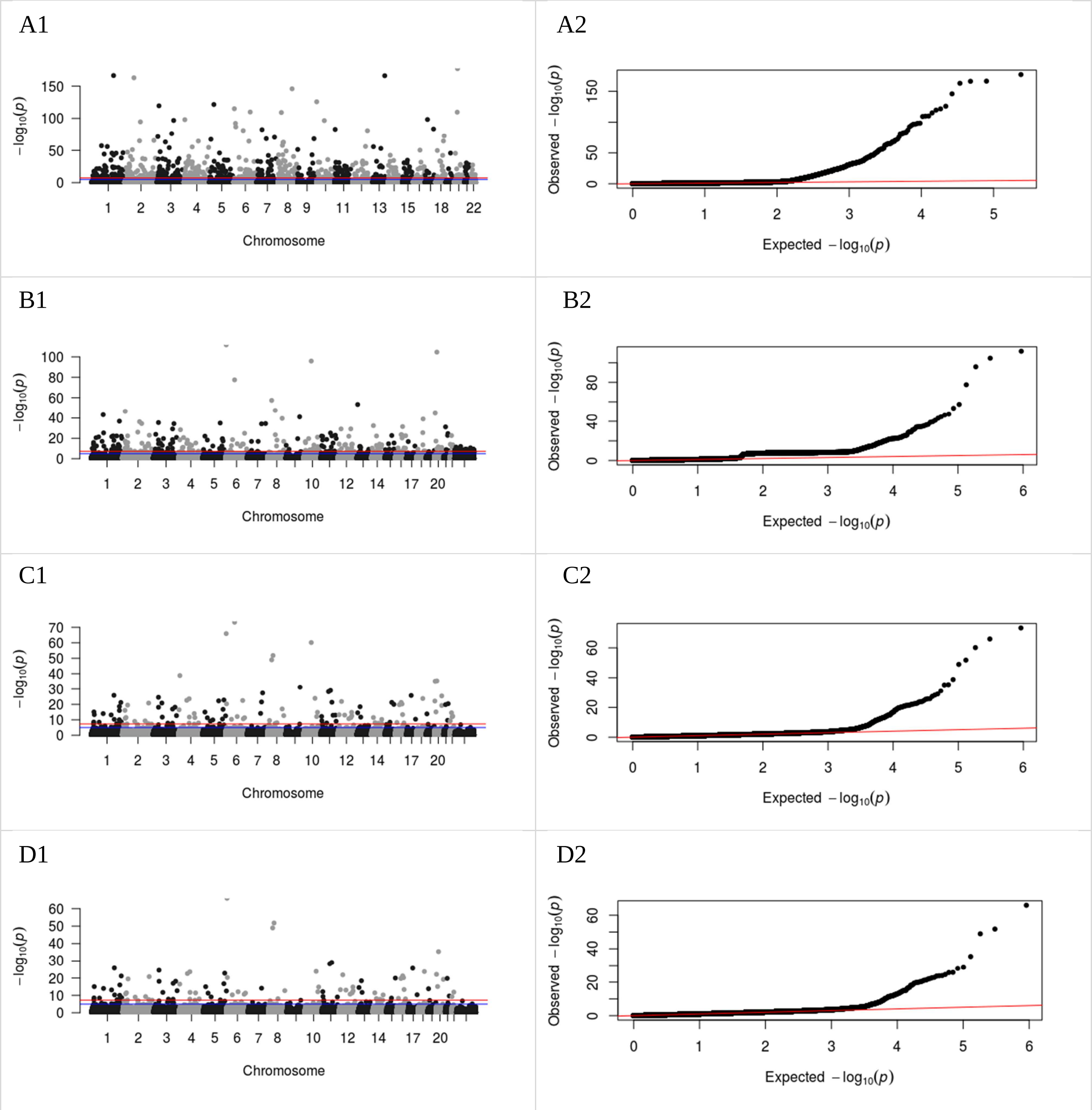


Fig S4. GWAS QC steps: Manhattan and qq plots. Figures are paired up. At every step of qc we generated both a Manhattan plot and qq plot. Manhattan plots: x axis, chromosomes, in alternating colours, y axis, negative log 10 of the p-value of every SNP in association with the phenotype. QQ plot, expected –log10 of the same p-values versus observed –log10 (p-values). When QC is performed carefully, the Manhattan plot should show only a limited number of SNPs passing the p-value threshold, and the QQ plot should be almost a diagonal line, with only a few dots deviating from it (the same significant SNPs that pass the threshold in the Manhattan plot). See Fig S1 for reference.

A: dataset after sample QC. At this stage, we have eliminated samples that do not behave very well: essentially individuals that are related and duplicates. The results are still pretty bad, not much improvement is observed compared to the first experiment. **B: dataset after sample QC and preliminary site QC.** The results look slightly better when we start filtering out bad SNPs. Here we eliminated missing SNPs, and rare ones. **C: dataset after sample QC and more advanced site QC, specifically after HWE step.** Another important step is eliminating variants that do not satisfy the Hardy-Weinberg Equilibrium (HWE) assumptions describe in this document, section B2. **D: dataset at one step from the end. Results of sex check. Only inclusion of PCs in the logistic regression model is missing.** Finally, we found that eliminating sex mismatches improved our results a great deal, especially since we were including sex as a covariate in the logistic regression. However, the biggest jump in quality (from S4D to S1) occurred with the last step, where we included the first 10 principal components as covariates in the model, which corrected for the population structure bias.

HiCCUPS (Hi-C Computational Unbiased Peak Search)

We have defined loops to be pairs of genomic regions that form a peak in the Hi-C map. A peak in the Hi-C map indicates that the two indicated regions are found co-localized in the 3D genome context more often than expected. We detect peaks using HiCCUPS (Hi-C Computational Unbiased Peak Search), as implemented by Rao *et al.*(Rao et al, 2014). The contacts count of very square in the matrix, at a given resolution, is compared to that of neighbouring regions, namely the horizontal neighbours (blue), the vertical closest rectangles (green), the surrounding 'doughnut' (black) and finally the bottom left corner (yellow). If the count in the centre is at least 50% more than that of each of the surrounding regions, then it is called a peak. The bottom left corner deserves perhaps a special explanation. As the matrix is symmetric, the peak search is only performed in half of the matrix, namely the upper triangle (row index ≤ column index). The search is furthermore performed more than once, at different resolutions, starting from the highest resolution (smallest squares). Thus, if the centre of the currently investigated square has more counts than all surrounding regions but the bottom left one, it is most likely only part of a peak of larger size, which will consequently be detected at lower resolution. Before performing the peak search, the Hi-C map is normalized, to correct for the 'linear genome bias'. Linearly close genomic regions are naturally also close in 3D. The expected distribution of contacts decreases exponentially as a function of the linear distance. To observe counts that are not expected, we must correct for this expected background distribution. In Rao *et al.* the maps are normalized according to a matrix balancing algorithm described in Knight and Ruiz (Knight and Ruiz, 2012).

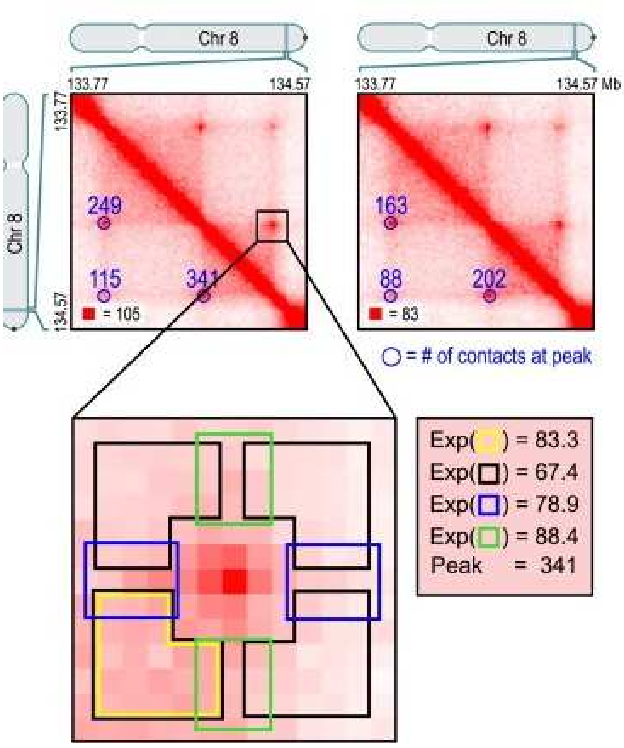


Fig S5. HiCCUPs. Taken from Rao *et al.* Figure 3A. '*We identify peaks by detecting pixels that are enriched with respect to four local neighborhoods (blowout): horizontal (blue), vertical (green), lower-left (yellow), and donut (black). These "peak" pixels indicate the presence of a loop and are marked with blue circles (radius = 20 kb) in the lower-left of each heatmap. The number of raw contacts at each peak is indicated. Left: primary GM12878 map; Right: replicate; annotations are completely independent.*' (Rao et al, 2014).

Loops are enriched for enhancers and common SNPs, especially when highly associated with the phenotype. P-values.

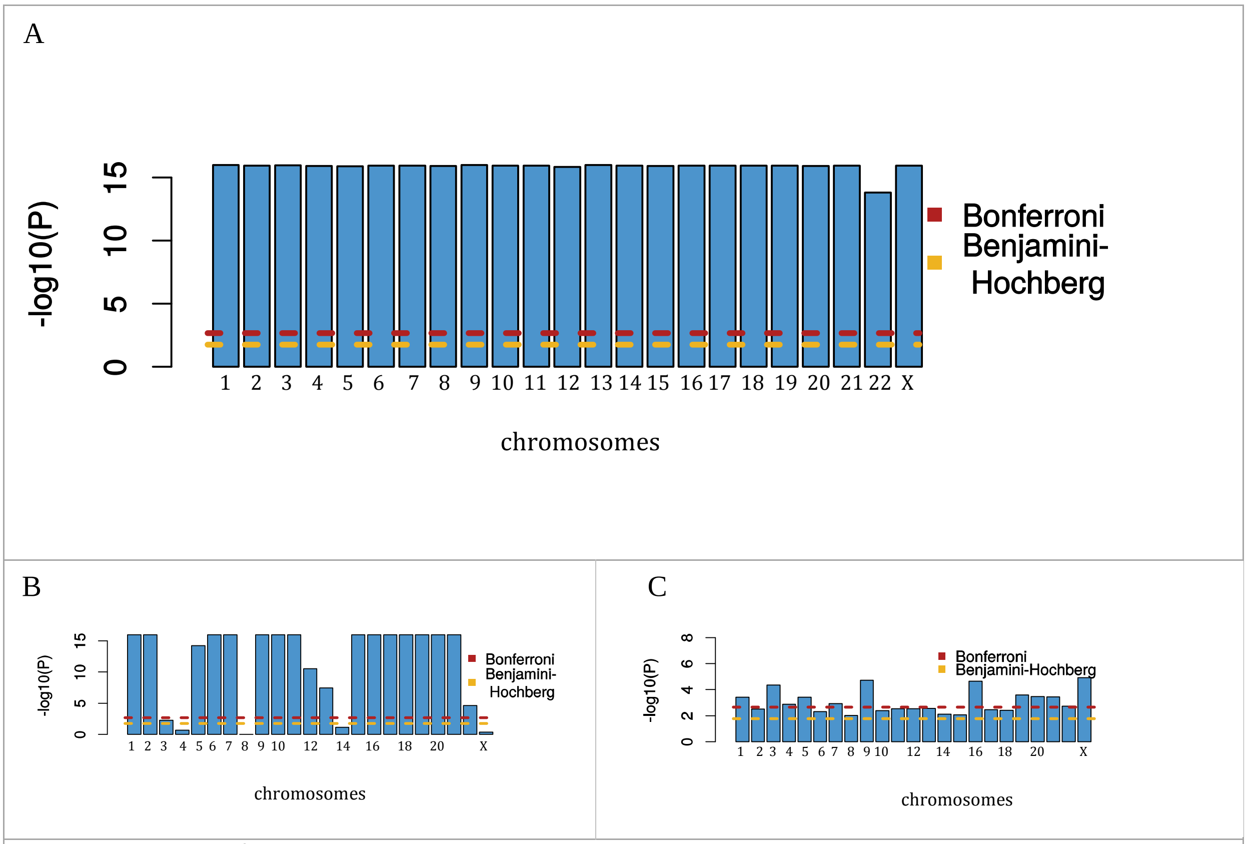


Fig S6. Loops are enriched for enhancer activity (A) and common variants (B), particularly when highly associated with the phenotype (C). P-values per chromosome. A: loops are enriched for enhancers. P-values obtained from a Binomial test, as described in 3.1.1. X axis: chromosomes, y axis: -log10(P). As indicated, the two lines represent the significance threshold after multiple testing correction. Red line indicates threshold using Bonferroni, yellow using Benjamini-Hochberg. The enrichment is highly significant for all chromosomes. The stars in Fig 5A were determined based on this result. **B:** loops are enriched for common SNPs. Similar to A, significance is determined with a Binomial test. The enrichment is not significant only for chromosomes 4,8, 14 and X. The stars in Fig 5B were determined based on this result. **C:** loops are especially enriched for SNPs that are highly associated with the phenotype, when we only look at SNPs that do show some association with the phenotype. We select only SNPs whose association measure is P<0.5. For those variants, we performed a one-sided Wilcoxon rank-sum test between the p-values of SNPs in loops and the p-values of SNPs not in loops. The result are indicated in the bar-plot. The axes are the same as in A and B, with the only exception that the p-values are obtained with a different test: not Binomial, but rank-sum, as described.

SNPs sitting on the same looping region show similar association with the phenotype

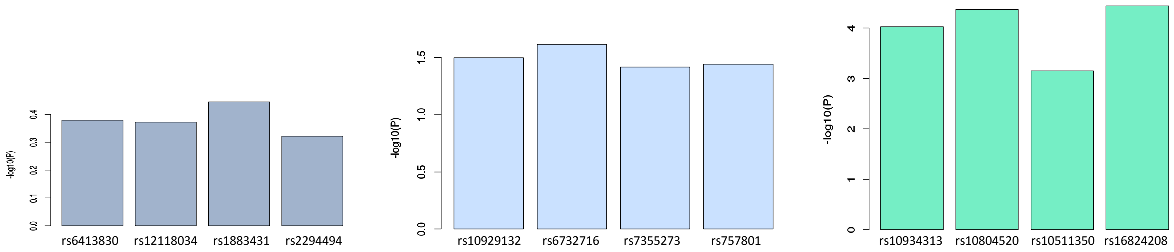


Fig S7. SNPs sitting on the same looping region have very similar association with the phenotype, even when not in strong LD. In order to show that SNPs sitting on the same looping region show similar association with the phenotype, we selected three regions on chromosome 1, 2 and 3 respectively that had exactly four SNPs sitting on them (from our original dataset, before imputation). Moreover, we chose three sets that have different levels of association, from very bad to quite good, to show that this factor does not influence the result. On the y axis, is the –log10 of the p-values calculated as an LRT of the logistic regression models containing the different variants compared to the null model. As you can notice, the scale varies in the three plots. On the x axis, are the IDs of the variants. Clearly, most of these behaviours are explained by linkage. Those SNPs sit on the same region, thus quite close to each other, thus in the same LD block. Naturally, then, they have very similar association measure. However, looping regions can be up to 25kb long, which is larger than the typical LD block. Rs109291322 and rs757801 in the same plot, for example, are highly correlated ($r^2 = 0.4$). We have performed these checks extensively if not exhaustively, and believe it is a good approximation to assume SNPs on the same looping region have similar association with the phenotype.

Epistasis occurs with great diversity from chromosome to chromosome

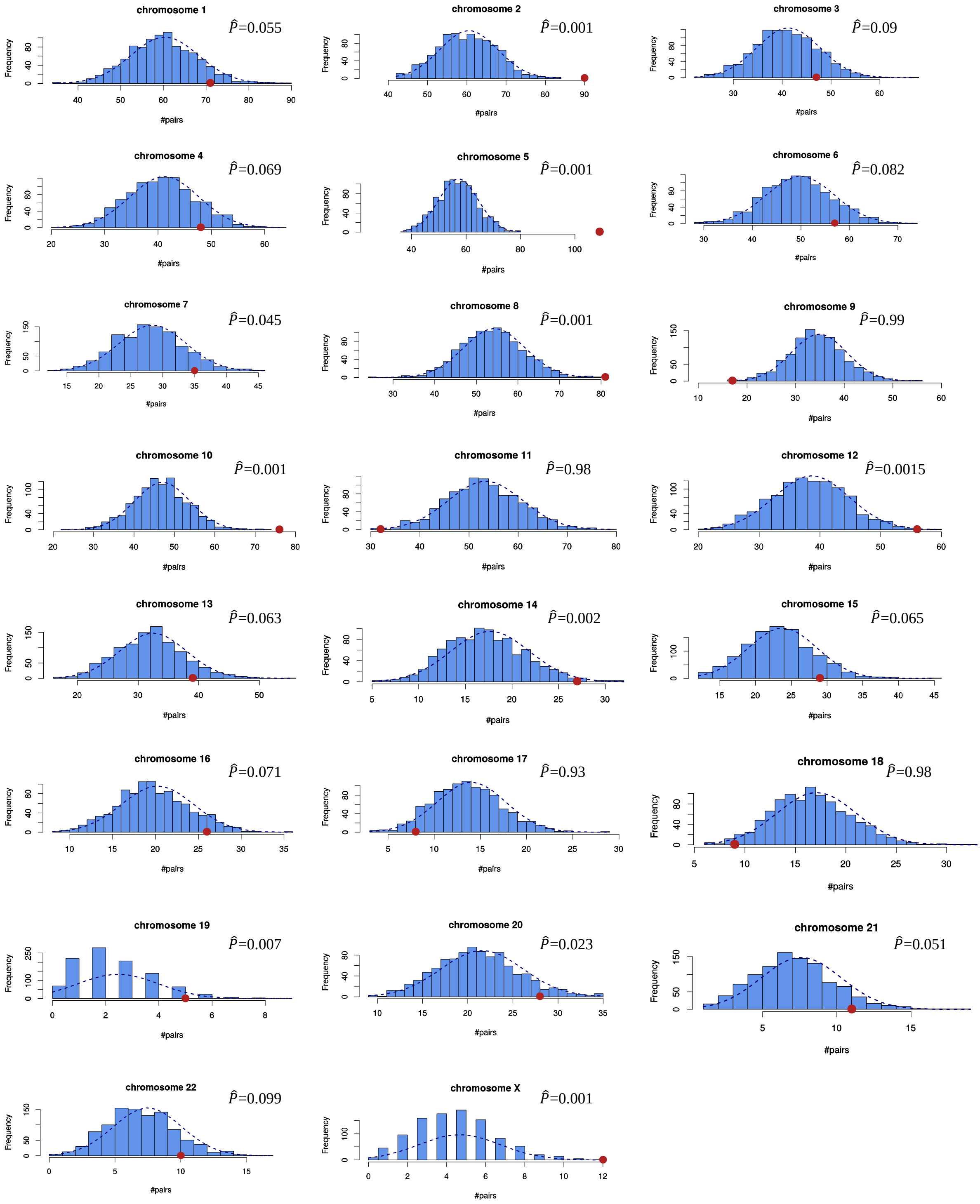


Fig S8. Histogram: distribution of the number of shuffled pairs showing epistatic effects, compared to the same number for 'true' pairs, for all chromosomes. Extension of Fig 6 from the main text. X axis, count of pairs showing significant epistatic interaction (LRT, $p<0.05$), for 1000 permutations. The red dot represents the same number for the 'true' pairs. On the y axis the frequency with which the different counts occur. In the majority of the cases the red dot is far to the right compared to the distribution of the shuffled pairs. Over all chromosomes, the number of significant epistatic interactions for the 'true' pairs is significantly larger than that of 'artificial' pairs (empirical estimated p-value = 0.01). However, we observe a puzzling opposite behaviour for four chromosomes. Chromosomes 9, 11, 17 and 18 show a completely opposite trend than the rest of the chromosomes. Not only the red dot is not far to the right compared to the distribution of the shuffled pairs: instead it is far to the left, as if the 'true' pairs were engaging in epistatic interactions particularly rarely. Unfortunately, we do not have an explanation for this now, although it is definitely worth investigating it in the future.

Percentages of pairs passing the different tests

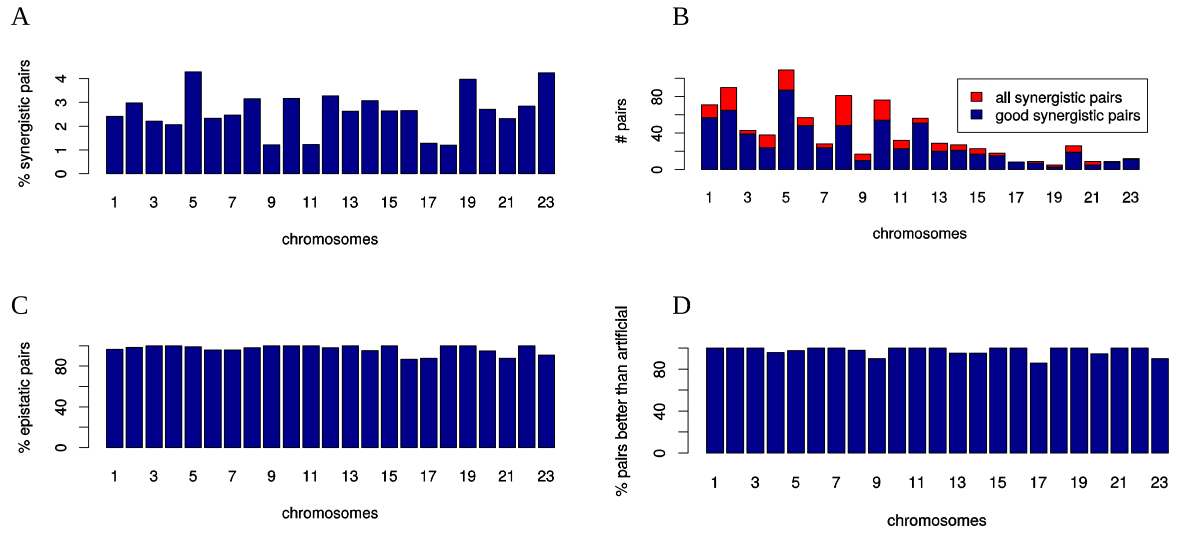


Fig S9. Steps of pairs selection. After selecting all pairs for which the interaction model performs better (LRT) than the two single-loci models (A), we proceed to find how many of those are also significant (better than null model, LRT) (B). Of those, to capture interaction only, we find for how many pairs the interaction model is also better (LRT) than the additive one (C). Finally, we calculate how many of these pairs also perform better than their counterparts ‘artificial’ pairs (D). **A.** X axis, chromosomes. Y axis, percentage of synergistic pairs, out of total pairs. For each chromosome, we counted how many pairs show a synergistic interaction, meaning that the SNP interaction improves association over the single SNPs. Generally, the percentage of success is between 1 and 4%. **B.** X axis, chromosomes. Y axis, count of total pairs (red) for which there is a synergistic effect. In blue, out of the total, number of pairs that are also significantly improving the association with the phenotype, when compared to the null model. On average, those are >70%. **C.** x axis, chromosomes. Y axis, percentage of synergistic and not additive pairs, out of ‘good’ pairs so far. We selected pairs for which the interaction is better than the individuals, and out of those the pairs that are also better than the null model. We here also check that the interaction model 1+2+1*2 performs better than the additive only 1+2. It does, in more than 90% of the cases, over all chromosomes. **D.** Similarly to A and C, we check those pairs that on top of all previous requirements also are better than the artificial counterpart pairs, at similar linear distance. X axis, chromosomes, Y axis, percentage of ‘true’ pairs performing better than both ‘artificial’ counterparts. Again, this is the case for nearly all pairs in all chromosomes.

KEGG Pathways: we found epistatic pairs between SNPs sitting on or in the vicinity of genes involved in these pathways

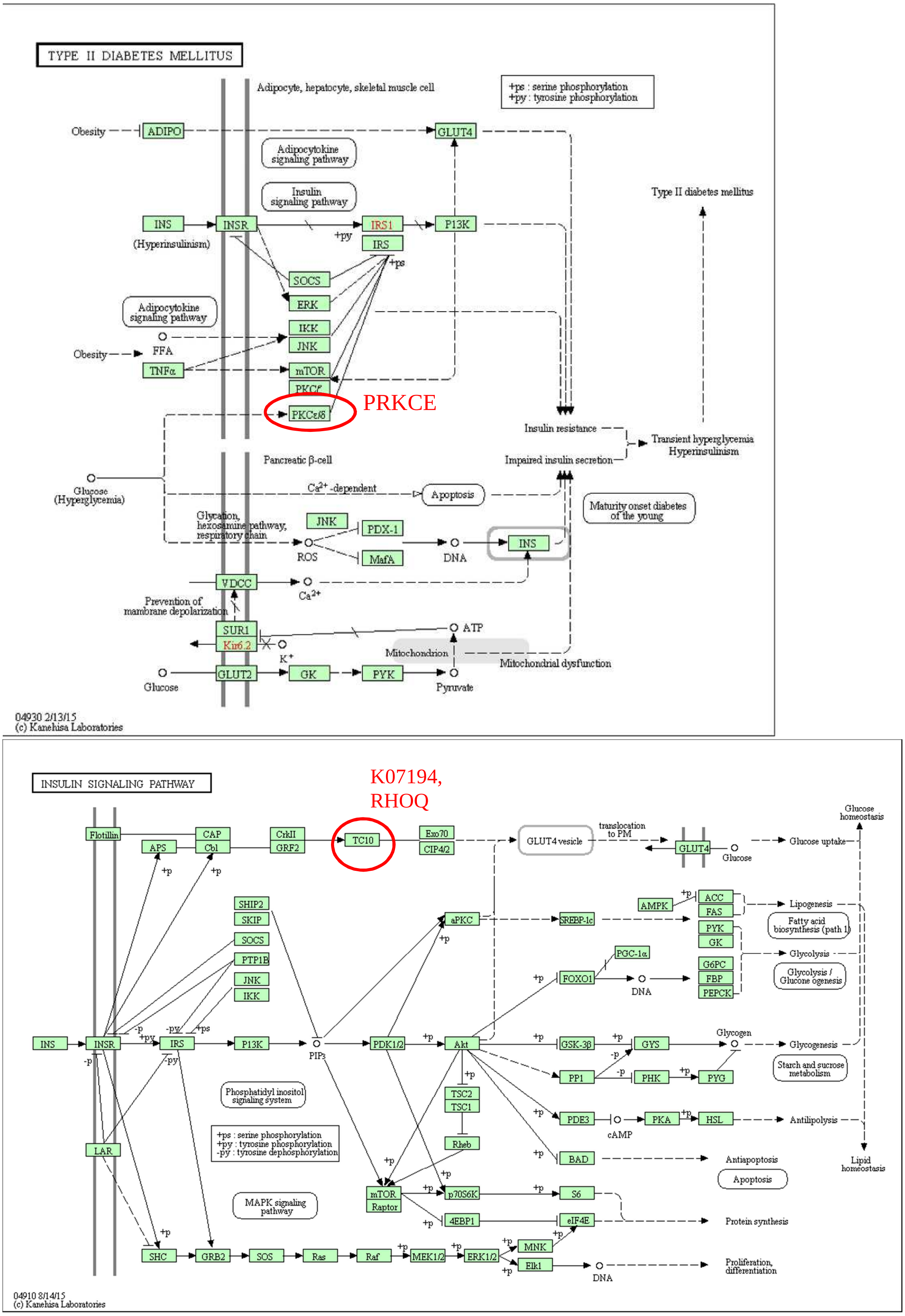
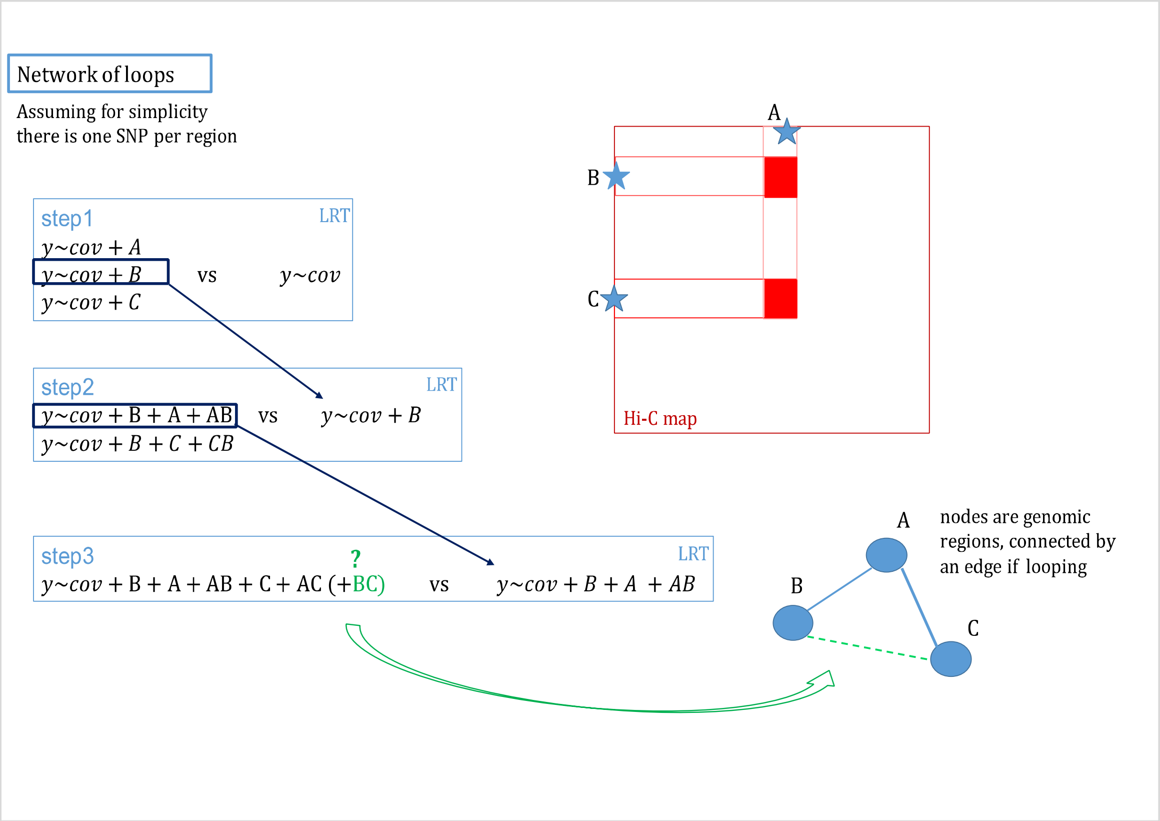


Fig S10. KEGG pathways: we found epistatic pairs between SNPs sitting on or in the vicinity of genes involved in these pathways: PRKCE in the Type 2 diabetes mellitus pathway and RHOQ in the insulin signalling pathway.

Future direction: from SNP pairs to SNP groups



Network of loops

Assuming for simplicity
there is one SNP per region

step1 LRT

$y \sim cov + A$

$\boxed{y \sim cov + B}$ vs $y \sim cov$

$y \sim cov + C$

step2 LRT

$\boxed{y \sim cov + B + A + AB}$ vs $y \sim cov + B$

$y \sim cov + B + C + CB$

step3 LRT

$y \sim cov + B + A + AB + C + AC\ (+BC)$? vs $y \sim cov + B + A + AB$

nodes are genomic regions, connected by an edge if looping

Fig S11. Extension of method from SNP pairs to SNP groups, using network of loops: an idea. The network of loops is an idea introduced by Sanborn *et al.* It is an innovative manner of looking at Hi-C maps, to try and go beyond its intrinsic pairwise limitation. The technique Hi-C (high-throughput 3C) extends 3C (chromosome conformation capture) from counting how many contacts there are between two selected genomic regions (*one-to-one)* to counting how many contacts there are between all pairs of genomic regions (*all-to-all).* Yet, it is only able to measure the co-occurrence of two regions at a time. However, a more realistic scenario is one where many regions are found in contact together at the same time, in sort of 'hairballs' (*de Laat,* not yet published*).* Those structures are not found with Hi-C, thus it is impossible to tell whether region A and B looping, and A and C looping, implies that A, B and C are all found together. It is possible however to hypothesize that it is so, and one way of detecting and visualizing clusters of genomic portions that are found in physical 3D contact, directly or indirectly, is the network of loops (sketch, bottom right). In a network of loops every node represents a genomic region, and two nodes are connected by an edge if those regions are found to be looping, as measured with Hi-C (sketch, top right). In our data, we identified clusters containing up to 11 nodes (genomic regions) searching connected components in the graph. One possible extension of our method starting from the network of loops is shown on the left. First, we should choose one 'representative' SNP per region. For every cluster in the network, we could start with identifying the best performing SNP (thus, node). We can do that, as usual, by comparing the model including the SNP with the null model with an LRT. Subsequently, we could progressively add to the model new SNPs if they are on connected nodes, and if they improve the overall association measure. This would terminate when all the nodes in the clusters have been examined.

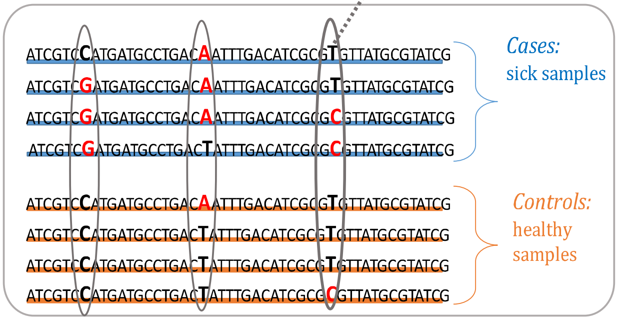
D. Boxes: main concepts illustrated

This section is meant as an illustration of the main concepts, a schematic overview of the key models and notions in this work.

D.1 Genome-wide association studies (GWAS)

GWAS

A single-nucleotide polymorphism (SNP), is defined as a single-nucleotide variation at one specific position on the genome. SNPs are 'common variants', occurring in at least 1% of the population. In the highlighted example, some individuals have a 'T', and some others a 'C', at the same locus.



ATCGTCCATGATGCCTGACAATTTGACATCGCGTGTTATGCGTATCG

ATCGTCGATGATGCCTGACAATTTGACATCGCGTGTTATGCGTATCG

ATCGTCGATGATGCCTGACAATTTGACATCGCGCGTTATGCGTATCG

ATCGTCGATGATGCCTGACTATTTGACATCGCGCGTTATGCGTATCG

Cases:
sick samples

ATCGTCCATGATGCCTGACAATTTGACATCGCGTGTTATGCGTATCG

ATCGTCCATGATGCCTGACTATTTGACATCGCGTGTTATGCGTATCG

ATCGTCCATGATGCCTGACTATTTGACATCGCGTGTTATGCGTATCG

ATCGTCCATGATGCCTGACTATTTGACATCGCGCGTTATGCGTATCG

Controls:
healthy
samples

Genotype-phenotype association

Genome-wide association studies (GWAS) aim to identify an association between a genotype (typically in the form of a SNP) and a phenotype, which can be a disease, or a trait.

Case-control setup

A GWAS experiment typically has a case-control setup: both diseased and healthy people are genotyped for the same sites, and the results are compared

*Association measures: **Contingency Table methods***

The Fisher exact test and the χ^2 test are both based on a contingency table, basically counting, per individual site the number of each allele in the cases and in the controls

	Allele1	Allele2
Cases		
Controls		

Contingency table. The distinction between alleles can be major/minor or reference/alternate

*Association measures: **Regression models***

Logistic regression models use genotypes as predictors for binary outcomes (diseased/healthy). Linear regressions are used for continuous traits.

$$y_i = \beta_0 + {}_{j=1}^{N} \beta_j x_{ij}$$

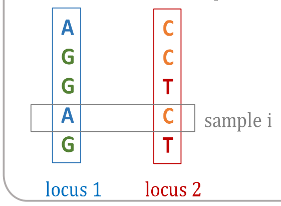
$$y_i = \ln \frac{p_i}{1-p_i}, \qquad p_i = outcome\ of\ sample\ i$$

$$x_{ij} = genotype\ at\ site\ j\ for\ sample\ i$$

D.2 Linkage disequilibrium (LD)

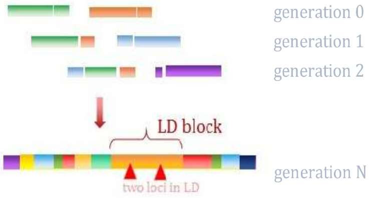
Linkage Disequilibrium (LD)

Linkage Disequilibrium is the non-random association between alleles at two different loci. This basically means that the presence of one specific allele at one position is not independent from the simultaneous presence of another allele at another position. If at one locus one can have an A or a G, and at another one can have a C or a T, then if the two loci are in LD one is more likely to have an A at locus 1 when they have a C at locus 2, for example. The two sites are in a way correlated, and their genotypes not independent from one another.



Recombination

There are many factors contributing to this non-random association, but mostly, LD is due to recombination.



LD formal definition and calculation

In formulas: $D = p(AB) - p(A)p(B)$

Where A is one allele (typically the minor allele) at the first locus, and B is one allele at the second. If the two sites were completely independent, then $p(AB) = p(A)p(B)$, and $D = 0$

Two derived measures, more frequently used are D' and r^2:

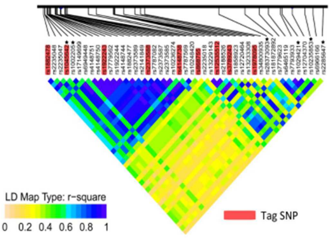
$$D' = D/D_{max}, \text{ where } D_{max} = \begin{matrix} \min(p_A p_B, (1-p_A)(1-p_B)) & if\ D < 0 \\ \min(p_A(1-p_B), p_B(1-p_A)) & if\ D > 0 \end{matrix}$$

$$r^2 = \frac{D^2}{p_A(1-p_A)p_B(1-p_B)}$$

LD plot and TAG SNPs

An LD plot represents the LD blocks, based on the r^2 measure. Triangles in blue represnts block where all SNPs are highly correlated (in LD).

For most genomic analyses is enough to look at one proxy, or TAG SNP, per block, as the genotypes of all other SNPs are easily imputed



Wikipedia, the free encyclopedia

D.3 Epistasis

Epistatic Interactions

Genetic interactions

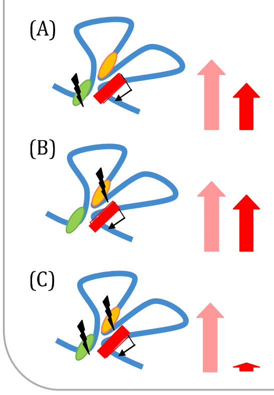
Perhaps the most well-known example of epistasis is the interaction that goes on between two genes responsible for human hair. One gene decides on the hair colour, the other determines whether you will or not get bald. Naturally, the effect on the first disappears, if the second is in action. The same happens for the 'albino gene' in animals. If albino, the dog's fur cannot be brown nor black, no matter what the 'fur-colour gene' says.

This is what we call an epistatic interaction, when the effect of one (gene, in these examples) has an influence on the effect (on the phenotype) of another, for example silencing or enhancing it.

	Blond hair	Red hair
Not Bald		
Bald		

SNP-SNP epistatic interactions

Similar phenomena can also occur among mutations or variants. Many experiments have been performed on yeast showing different effects resulting from different combinations of presence/absence of mutations.



In this scenario two enhancers (yellow and green) loop over to help the transcription of their target gene (red). When only one of the two enhancers is mutated (A),(B), the other is enough to maintain gene expression at an almost normal level (in pink, gene expression when no enhancer is mutated). However, when both enhancers are disabled, almost no transcription happens. This situation (C) shows an epistatic effect of the two SNPs on the two enhancers, on one another.

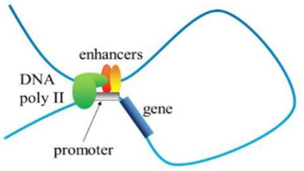
D.4 The three-dimensional genome organization

The 3D genome

The DNA is traditionally regarded as a straight line, a sequence of over 3 billion A's, C's, G's and T's. If we were to strecht it, we would reach a length of almost 2 metres. To fit in the ~10μm-diameter nucleus of a cell, the DNA is wrapped around proteins called histones to form the chromatin fibre and then even further compacted. This generates extensive contact between genomic regions that are very far apart in the linearized unfolded sequence. The mechanisms leading to the final 3D structure of our genome are not random, and play a role in several cellular functions.

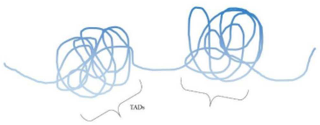
Enhancer-promoter loops

The three-dimensional organization of the genome plays a role in gene regulation. One example of that are enhancer-promoter loops. Enhancers are regulatory elements that aid the transcription of the genes they target. They sit several bp away from their target, but upon transcription loop over, and directly contact their promoter



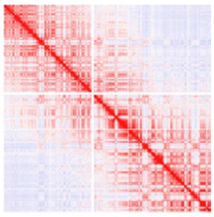
TADs and LADs

The 3D genome organization is hierarchichal, non-random and well conserved at different scales.
Topological associated domains(TADs) are defined as regions with many interactions within and hardly any interaction between.
Equally well conserved across cell-types are LADs (lamina associated domains).



Measuring the 3D genome

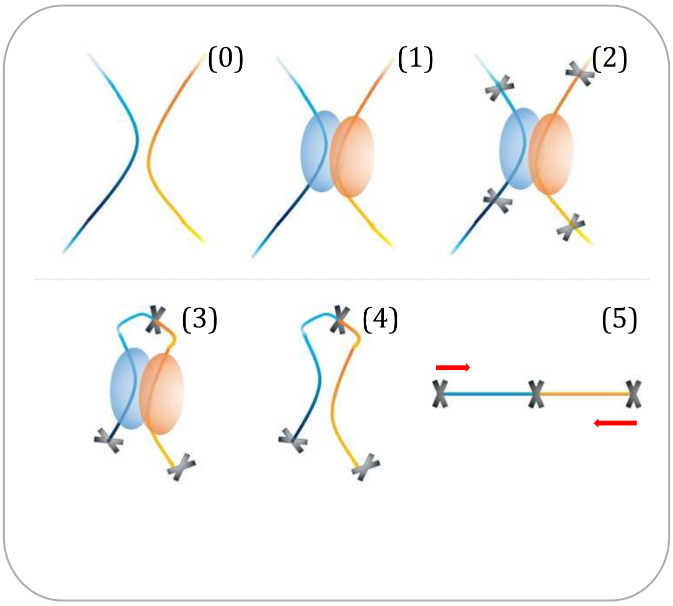
The 3D genome is now measurable thanks to the Chromosome Conformation Capture techniques. The technology used here is Hi-C, which measures genome-wide, pairwise 3D contact. The result of a Hi-C experiment is a contact (heat)map, a symmetric matrix where at position (i,j) the brightness of the colour indicates the number of contacts between genomic regions i and j.



normalized Hi-C map showing TADs

D. 5 Chromosome Conformation Capture (3C)

Chromosome Conformation Capture



In this technique, the chromatin is fixated in its effective, *in vivo*, organization by crosslinking DNA-DNA contacts with chromatin-associated proteins (1). In the next step, this configuration is cut with a restriction enzyme (2), and allowed to re-ligate (3). In this way, two genomic regions that are close in the 3D conformation are glued together. Finally the ligation products are de-crosslinked (4), and with PCR it is possible to quantify the amount of contacts between those two specific regions.

There exist many techniques deriving from the original 3C (2002). The most common and used are 4C (2006), 5C (2006), ChIA-PET (2009), and Hi-C (2009).

References

Bush, W. S. and J. H. Moore (2012). "Genome-wide association studies." *PLoS Comput Biol* 8(12): e1002822.

Clarke, G. M., *et al.* (2011). "Basic statistical analysis in genetic case-control studies." *Nature protocols* 6(2): 121-133.

Purcell, S., *et al.* (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." *The American Journal of Human Genetics* 81(3): 559-575.

Knight, P. A. and D. Ruiz (2012). "A fast algorithm for matrix balancing." *IMA Journal of Numerical Analysis*: drs019.

Rao, S. S., *et al*. (2014). "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping." *Cell* 159(7): 1665-1680.