Master Thesis

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

Anna Cuomo
Delft University of Technology

to be defended publicly on Thursday September 22, 2016 at 12:30
in partial fulfillment of the requirements for the degree of
Master of Science in Computer Science, track Bioinformatics

Supervisors: J.de Ridder UMC Utrecht
S.L. Pulit UMC Utrecht
A. Allahyar TU Delft
Thesis Committee: M.J.T. Reinders TU Delft
W.M. Ruszel TU Delft

Abstract

Motivation: Genome-wide association studies (GWAS) aim to uncover the genetic basis of traits and common diseases. Due to the large number of common variants, most studies use a single-locus approach. However, those fail to explain most of the heritability, especially for complex diseases. Epistatic interactions, where two or more loci have a synergistic effect on the phenotype, are worth investigating to improve our understanding of the genetic architecture of human disease. Most studies that have investigated epistatic association in GWAS focus on gene-gene interactions only. However, regulatory elements such as enhancers have the power of increasing and decreasing the expression level of their target genes, playing a fundamental role in determining their effects, also in relation to diseases. Thus, enhancer-promoter, or enhancer-enhancer interactions should be included in the search for GWAS epistasis. More and more studies show that the control of gene expression can occur over large genomic distances. Enhancers loop over to get in physical contact with their target genes. *In vivo,* enhancers and promoters are therefore found in close 3D spatial proximity. Chromatin loops can be detected using the chromosome conformation capture (3C) technique and its derivatives. Particularly, Hi-C combines 3C with next generation sequencing (NGS), identifying all contacts between all pairs of genomic regions.

Results: In this study, we investigate GWAS epistatic effects of single nucleotide polymorphisms pairs (SNPs) engaging in long-range chromatin interactions. To this end, we overlay GWAS hits, using the T2D (type 2 diabetes) dataset from the WTCCC (Wellcome Trust Case-Control Consortium), with high resolution Hi-C maps. We show that chromatin loops are enriched for common variants, particularly when highly associated with the phenotype. Moreover, looping regions are associated with enhancer activity. We find three sets of SNP pairs engaging in epistatic interactions, on chromosomes 2, 3 and 12. The SNPs are found in either regions with high enhancer activity or in genes involved in metabolic pathways, which supports their potential role in type 2 diabetes (T2D).

Availability: An electronic version of this thesis is available at http://repository.tudelft.nl/

Supplementary information: Supplementary data are attached separately and also available at http://repository.tudelft.nl/

1 Introduction

Genome-wide association studies (GWAS)

Genome-wide association studies (GWAS) have been widely successful in discovering association between genomic variants and the risk of having diseases, or presenting specific traits (McCarthy *et al*. 2008, Pulit 2016). (See Box D.1, in the Supplementary Information, for an illustra-

tion of GWAS). Since the first successful GWAS in 2005, more than 15,000 single nucleotide polymorphisms (SNPs) have been found to be associated with over 1,200 diseases and traits (Li *et al*, 2016). Understanding the genetic basis of disease has the potential to truly revolutionize medicine by uncovering biochemical pathways for drug targets and by enabling personalized risk assessments (Ward *et al*, 2012). However, a full understanding of the molecular mechanisms that link common sequence variation to disease predisposition is still far from being accomplished, especially since the majority (~93%) of found GWAS hits sit on poorly characterized non-coding regions of the genome (Visel *et al*, 2009, Maurano *et al*, 2012, Ward *et al*, 2012, Schierding *et al*, 2014). Moreover, single SNPs often only have a rather small effect on the phenotype (Li *et al*, 2011, Stringer *et al*, 2011) and only a small portion of the heritability of complex diseases can be explained by individual GWAS hits. This phenomenon has been referred to as 'missing heritability' (Maher, 2008). Complex traits and diseases are believed to be caused by multiple genetic loci and their interaction, rather than by one leading variant, as it is for very rare diseases. Thus, studies have looked for the effect of multiple variants at a time (Dinu *et al*, 2012, Grubert *et al*, 2015, Jamshidi *et al*, 2015). Epistatic interactions, where two or more loci have a synergistic influence on the phenotype, could help explain the 'missing heritability'. Epistatic interactions are often called gene-gene interactions, since most studies have only investigated such effects for coding genomic regions (Turner *et al*, 2011, Ritchie *et al*, 2011, Wei *et al*, 2014). However, similar effects could occur between other functional regions. Regulatory elements such as enhancers, for example, can actually increase or decrease the level of expression of a gene. A mutation on such an element would have the power to dysregulate the activity of that gene exerting a similar effect to if the mutation was on the gene itself. Furthermore, regulatory elements occupy a much larger portion of the genome (~40%) than genes (<2%), thus their inclusion can perhaps contribute to the interpretation of non-coding SNPs.

From a technical point of view, it is unfeasible to test all possible pairwise combinations of the several million known common variants. An exhaustive evaluation of all possible scenarios would require too much computational time and power, and it would severely limit statistical power (Brinza *et al*, 2006, Bush *et al*, 2009, So *et al*, 2011, Piriyapongsa *et al*, 2012, Ayati *et al*, 2014, Goudey *et al*, 2015). This is mostly due to our limited sample size (Evans & Purcell, 2012, Hong *et al*, 2012). Clearly, a smart way of selecting testable pairs is necessary.

In this work, we propose to use the three-dimensional organization of the genome and the consequent physical contact between genomic regions in the 3D space to prioritize SNP-SNP interaction pairs.

The genome in 3D

If we were to stretch our genome, we would reach a structure of nearly 2 metres in length. 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. More and more studies have recently shown that the 3D organization of the genome is not random, and it is actually believed to play a role in key cellular functions. Particularly, it would act as an additional layer to the nucleus regulatory mechanisms (de Wit *et al*, 2013,

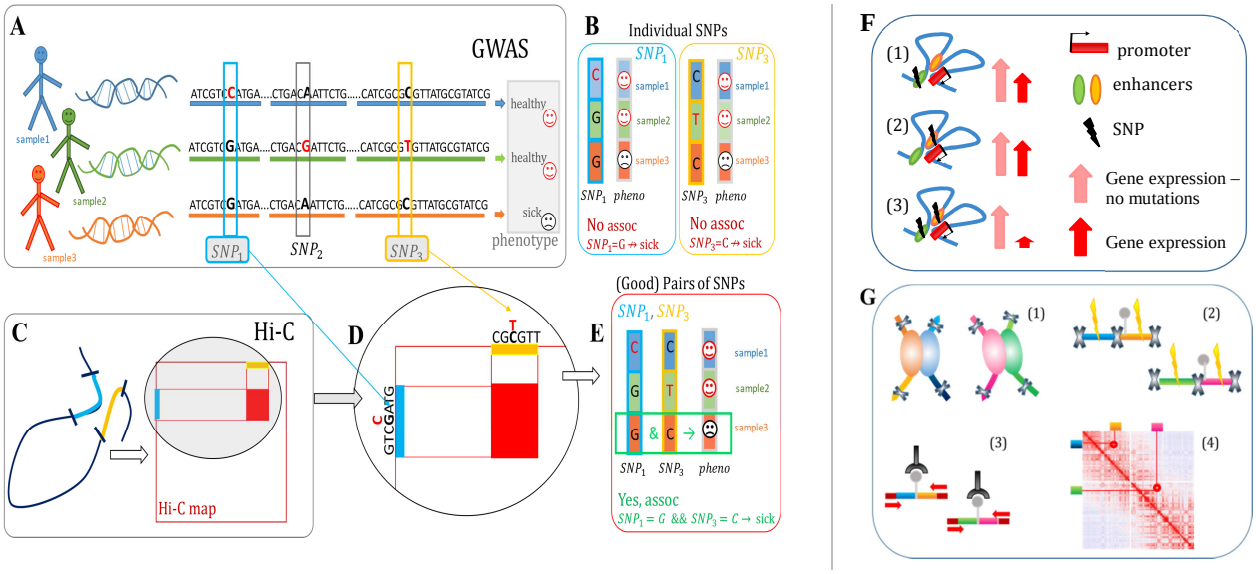

Fig. 1. Overview: integration of GWAS and Hi-C. A, schematic overview of GWAS. In GWAS, many individuals (samples), here indicated by the different colours (blue, green, orange) are genotyped at a number of genomic locations (SNPs, sites that are commonly mutated in the population). Samples are divided into healthy (controls, or unaffected) and diseased (cases, or affected). The aim is to find an association between exhibiting a certain genotype at one specific position, and having the disease. As sketched in **B,** sometimes the association of individual SNPs with the phenotype (sick/healthy) is hard to find. One reason might be that it is actually the synergistic effect of two SNPs, that has an effect on the disease predisposition. As evaluating all possible pairwise combinations of SNPs is unfeasible, potential pairs must be prioritized. **C**, we use the three-dimensional organization of the genome, measured with Hi-C, to prioritize SNP-SNP pairs. Candidate pairs are therefore defined, as in figure, as two sites sitting on two genomic regions (light blue and yellow) engaging in a long-range chromatin interaction, identified as a peak in the Hi-C contact heat-map (sketch). In this way, we hope to identify epistatic interactions between SNPs on two enhancers targeting the same gene, or one on the promoter and one on the enhancer of a gene. In **D**, we show how out of the many GWAS hits, we select only the ones sitting on regions engaging in a loop (here, SNP_1 and SNP_3). In **E**, including both those two SNPs at the same time allows us to find an association with the phenotype (having both a G at position 1 and a C at position 3 has a very high association with having the disease). **F,** example of epistatic interaction detected with Hi-C. 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 (1),(2), 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 (3) shows an epistatic effect of the two SNPs on the two enhancers, on one another. **G,** overview of the Hi-C technique. In one experiment, thousands of pairs of regions get glued together in their original 3D conformation (here, two pairs pf regions are depicted). The chromatin is crosslinked, digested (1) and let re-ligate. The ligation products are marked with biotin, the chromatin is shred into small pieces using sonication (2) and the biotin is pulled down. Finally, the ligation products can be PET-sequenced (3). The result of Hi-C is a contact matrix, where the brighter the colour the more contacts were detected between the corresponding regions, across many cells. This (4) is a real data Hi-C matrix (*ref*: homer.salk,edu), unlike the simplistic sketches in **C** and **D.**

Bouwman *et al*, 2015, Pombo *et al*, 2015). Regulatory elements such as enhancers, for example, loop over to get in close proximity with the promoter of their target gene, to initiate transcription (Botta *et al*, 2010, Ghavi-Helm *et al*, 2014, Matharu *et al*, 2015) (Fig 1C, 1F. Supplementary Box D.4). Enhancers have the ability to help initiate transcription, and thus regulate the expression of the gene they target. A mutated enhancer would not be able to carry out its function anymore, dysregulating the level of the expression of the target gene. However, it is not the gene that is directly mutated. In many cases, the regulation of genes is a complex mechanism that involves multiple enhancers, all looping over, forming so called active chromatin hubs (ACHs)(Bouwman *et al*, 2015, Pombo *et al*, 2015). An important example is given by the human beta-globin locus (Tolhuis *et al*, 2002, de Laat and Grosveld, 2003, Sanyal *et al*, 2012, Jin *et al*, 2013). These observations suggest that there exist regulatory networks of enhancers that become active upon transcription. Thus, we can use these regulatory circuits to detect combined effects of SNPs contained in the elements involved (promoters, enhancers) on the phenotype (Fig 1F). As those networks act by forming chromatin loops between enhancers and promoters, the measure of the 3D genome organization allows us to uncover this type of epistatic interaction.

Fortunately, we are now able to measure the three-dimensional DNA, using the Chromosome Conformation Capture ('3C') techniques. In 3C (Dekker *et al*, 2002), the chromatin is fixed in its effective, *in vivo*, organization by crosslinking DNA-DNA contacts with chromatin-associated proteins. Next, this configuration is cut with a restriction enzyme, and allowed to re-ligate. In this way, two genomic regions that are close in the 3D conformation are glued together. Finally the ligation products are de-crosslinked, quantified, sequenced and mapped to a reference genome. Then, it is possible to quantify the amount of contact between the two regions (de Wit and de Laat, 2012) (see Supplementary Box D.5 for a visual representation of the 3C technique). Several other technologies derive from 3C, the most popular being 4C (Simonis *et al*, 2006), 5C (Dostie *et al*, 2006), ChIA-PET (Fullwood *et al*, 2009) and Hi-C (Lieberman-Aiden *et al*, 2009). Using 3C, we can detect the number of contacts between two selected regions: it is called a *one-to-one* technique. Hi-C (high throughput 3C), on the other hand, is the first of the 3C technologies to be genome-wide, and is therefore called an *all-to-all* technique. At the cost of a little lower resolution, Hi-C is able to capture all contacts between all pairwise combinations of genomic regions. The result of a Hi-C experiment is a contact map, a symmetrical matrix M where M(i,j) contains the count of contacts between regions i and j (Fig 1G).

Contribution

In this project, we integrate data from GWAS and Hi-C, to investigate epistatic effects of SNPs pairs that are co-localized in 3D (Fig 1A-E). These two data-types have seldom been combined before. In Grubert *et al.* the authors aim to unravel which genomic variations have an effect on gene expression (eQTLs) or histone modifications (hQTLs), not only in cis, but also in trans. To find putative long-range driving SNPs, they use Hi-C and ChIA-PET (Grubert *et al*, 2015). The 3D genome is therefore used to find SNP-histone modification or SNP-mRNA production pairs, rather than SNP-SNP interaction pairs. On the other hand, Xu *et al.* use Hi-C to improve the interpretation of non-coding GWAS hits, but is limited to single-loci (Xu *et al*, 2016). Similarly, in Dryden *et al.* Capture-C is used to find potential target genes of breast cancer susceptibility loci (Dryden *et al*, 2014). In Martin *et al*. the authors look at promoter-enhancer SNP-SNP interactions but pre-select known autoimmune risk loci as candidate genes, and look for the responsible enhancer mutation (Martin *et al*, 2015). Finally, in Li *et al.* SNP-SNP interactions are searched, involving at least one non-coding SNP. Putative pairs are prioritized based on either mRNA overlap, or shared biological annotations. The count of contacts in 3D (measured with ChIA-PET) is used to validate found pairs (Li *et al*, 2016). In contrast, in our approach, the spatial vicinity in the 3D genome conformation is the prioritization method itself, the way we define potentially interacting SNPs in the first place.

Detecting epistatic interactions among GWAS hits to improve disease association is a challenging task, mainly due to the large number of tests to be performed. To maintain statistical power, a selection of promising candidate pairs is necessary, and only those should be tested. The selection phase is extremely important, as it determines what kinds of pairs we are to find and on the other hand what information we are losing. Most existing approaches pre-select GWAS hits on genes only, and group genes based on shared function or linear vicinity on the genome. In this work, we aim to uncover the relation between SNPs on regions that are close in the 3D space instead. In this way, we are not limited to coding regions. In fact, as we mentioned, gene regulation can occur over large genomic distance, by means of chromatin loops bringing together enhancers and promoters upon transcription. SNPs on enhancers might be just as deleterious as SNPs on genes, when they drastically decrease their transcription levels. We hypothesize that pairs of SNPs in spatial three-dimensional proximity might sit on promoter-enhancer pairs, or on two enhancers targeting the same gene.

We propose a statistical framework to measure the GWAS association improvement of SNP-SNP pairs, prioritized as variants sitting on regions that are in contact in the 3D context. We incorporate SNP-SNP interactions in GWAS using a logistic regression model, containing an interaction term. Logistic regression is a powerful statistical learning technique that can be used to model a binary outcome (e.g. sick or healthy, in a case-control setup) using multiple factors, including SNP genotypes. A likelihood ratio test (LRT) approach is then used to compare two logistic regression models, with and without the interaction term, to detect epistatic effects (see Supplementary section A for a detailed explanation of the logistic regression model and the likelihood ratio test).

Both the SNP-SNP pair prioritization using the 3D genome organization and the genome-wide LRT approach to measure SNP-SNP epistatic interactions are novel to our method, to the best of our knowledge.

2 Approach

2.1 Data

2.1.1 GWAS

We used GWAS data from the WTCCC (Wellcome Trust Case-Control Consortium), as published by Burton *et al*, with their permission (Burton *et al*, 2007). We focus on the type two diabetes (T2D) dataset. Before any further analyses, the raw data are cleaned, following a number of steps, as described in numerous articles in the field (Price *et al*, 2006, Burton *et al*, 2007, McCarthy, 2008). Detailed quality control (QC) steps are illustrated in the Supplementary Information, section B. After quality control, which is performed with Plink (Purcell *et al*, 2007) and informed by the quality control steps taken by the WTCCC, we have data for 3,343 samples and 450,242 sites. SNPs positions are lifted over to build hg19 (from hg18) using LiftOver (Kuhn *et al*, 2013).

In a classical GWAS experiment, only a number of variants (generally between 500,000 and a 2 million) are genotyped, rather than all known

variable loci. This is done using pre-defined SNP-arrays. The selected positions are sampled rather uniformly along the genome, and act as proxies for all other common variants in their close vicinity. This is related to a phenomenon called linkage disequilibrium (LD). (See supplementary Box D.2 for a visual overview of LD). Linkage disequilibrium is the non-random association between alleles at two different loci, as a result of generations of recombination events (Gibbs *et al*, 2003). The resulting highly correlated SNPs are said to be in the same 'LD block'. Thus, it is sufficient to know the genotype at one locus to be reasonably confident in inferring the genotypes at all other positions in the same block. When a GWAS hit is found, in relation to some phenotype, it is not immediately obvious whether that particular SNP is responsible, or another SNP is, in the same block. Thus, it is necessary to impute all other variants. To impute variants means to infer the genotypes at loci that are not directly sampled based on those that are sampled, using a very large population as reference. Imputation helps tremendously in finding the true causing SNPs in GWAS. This is particularly true in our work, since we aim to find causal pairs of SNPs, not single SNPs. Thus, we subsequently imputed a large amount of SNPs (109,126,218 variants in total, for chromosomes 1-22), using an available online imputation server (http://imputationserver.sph.umich.edu). The genotypes resulting from imputation come with a measure of how confident the server was in imputing them. R-squared (r^2) is a measure of the correlation between the imputed genotypes and the true genotypes calculated as they impute (Howie *et al*, 2012, Fuchsberger *et al*, 2015, Loh *et al*, 2016). We filter out SNPs with a very low imputation quality, choosing a not so stringent threshold ($r^2 > 0.3$). After filtering, we are left with 47,073,880 variants in total (~40% of all imputed SNPs). Moreover, the results of imputation are not genotypes, but dosages. A dosage is a scalar between 0 and 2, which is calculated from the probabilities of the different genotypes (encoded as 0/1/2), for one sample, at one locus. For SNP i, sample j:

$$dosage(i,j) = 2P(geno(i,j) = 2) + P(geno(i,j) = 1)$$

For simplicity, we discretize such dosages back to the closest genotypes (0, 1, 2).

$$\widehat{geno}(\mathrm{i,j}) = arg\min_{k=0,1,2}\{dosage(i,j) - k\}$$

Note, with the notation 'hat' ($\widehat{\ }$) we indicate estimated values. As we have already filtered on imputation quality, the difference between dosages and estimated genotypes is never large. To be even more stringent, however, we filter out all SNPs for which such difference is larger than 0.1. Details of all the steps in the GWAS pipeline are described in the Supplementary Information, section B.

2.1.2 Hi-C loops

The Hi-C data is obtained from the up-to-date highest resolution publicly available Hi-C maps database, published by the Lieberman-Aiden lab (Rao *et al*, 2014 & Sanborn *et al*, 2015). Out of the 8 different human cell lines for which Hi-C data is available at high resolution (5 kb), we select the GM18278 cell line. This is a human lymphoblastoid cell line. It is a popular cell line and is one of the original HapMap cell lines (Gibbs *et al*, 2003). We argue that the 3D conformation, although locally quite variable, is rather well conserved across different cell types, as far as long-range chromatin interactions are concerned (Dixon *et al*, 2012, Meuleman 2013, Pope *et al*, 2014). This legitimates overlaying data types obtained from different cells. Together with the high resolution Hi-C maps, the documentation provided by Rao *et al.* includes a list of validated chromatin loops. Those are actual loops that the chromatin fibre forms within the cell nucleus to bring together two genomic regions in the 3D space. We are able to detect them as pairs of regions that form a peak in the Hi-C contact map (Fig 2A). The two regions that are brought together by a loop are also referred to as the anchors of the loop. A Hi-C contact map is the result of a Hi-C experiment. It is a symmetrical matrix that for every pair of genomic regions, at a given resolution, counts how often they are found together, across different cells, as measured with the genome-wide chromosome conformation capture technique (Fig 1G. The 3C technique is also illustrated in the Supplementary Information, box D.5). A peak is defined as a pair of regions for which the contact count is significantly higher than that of neighboring (pairs of) regions, squares in the Hi-C matrix. Peaks are detected with HiCCUPS (Hi-C Computational Unbiased Peak Search), as defined in Rao *et al.* (Rao *et al*, 2014). An illustration of how HiCCUPS works is shown in Fig S5. We use the list of loops found with HiCCUPS for this cell-line, as provided by Rao *et al*. The list includes loops obtained at 5, 10 and 25 kb resolution (size of the two regions forming the loops). There are in total 8,334 loops, all intra-chromosomes (both regions looping to each other are contained in the same chromosome).

2.2 Definitions

To avoid confusion, we shall set some definitions that will be used from now on in this text (Fig 2).

Regions: a region is a genomic portion, of given size. The size is determined by the resolution of the Hi-C map, and one region is one bin of the Hi-C matrix. One region is uniquely defined by the chromosome it belongs to and its starting and ending positions, in base pairs:

$$Region = \{chr,\ x1,\ x2 \mid x1, x2 \in chr\}$$

Loops: one loop is defined as a pair of two regions, of equal size. Loops are defined as peaks in the Hi-C map (Fig.2A). Loops are uniquely identified by two regions *A* and *B*, on the same chromosome *chr*:

$$\begin{aligned} Loop = \{&chr,\ A, B \mid A, B\ are\ regions,\ A, B \in chr,\\ &A = [x_1, x_2],\ B = [y_1, y_2]:\ (y_2 - y_1) = (x_2 - x_1),\\ &A, B\ are\ bins\ of\ Hi\text{-}C\ map, and\ form\ a\ peak\} \end{aligned}$$

Looping Regions: a looping region is a region involved in a loop (region, loop as just defined).

Pairs: the term pairs will refer from now on to SNP-SNP pairs. One pair is defined as:

$$\begin{aligned} Pair = \{&chr, i, j \mid i, j\ are\ SNPs,\\ &i, j \in chr\} \end{aligned}$$

True pairs: these are pairs (as just defined) selected such that one SNP sits on one genomic region, the other on another genomic region, and the two regions form a loop (as just defined)(Fig 2A).

$$\begin{aligned} True\ pair = \{&chr, i, j \mid chr, i, j\ \ is\ a\ pair,\ \ i \in A, j \in B,\\ &chr, A, B\ \ is\ a\ loop\} \end{aligned}$$

We have defined a loop to be made up of two genomic regions. The size of such regions ranges from 5kb to 25 kb. The several million known common SNPs span the entire genome, with a relatively high frequency (~3-4% of the human genome, on average). As we show in 3.1, moreover, common variants are particularly abundant in looping regions.

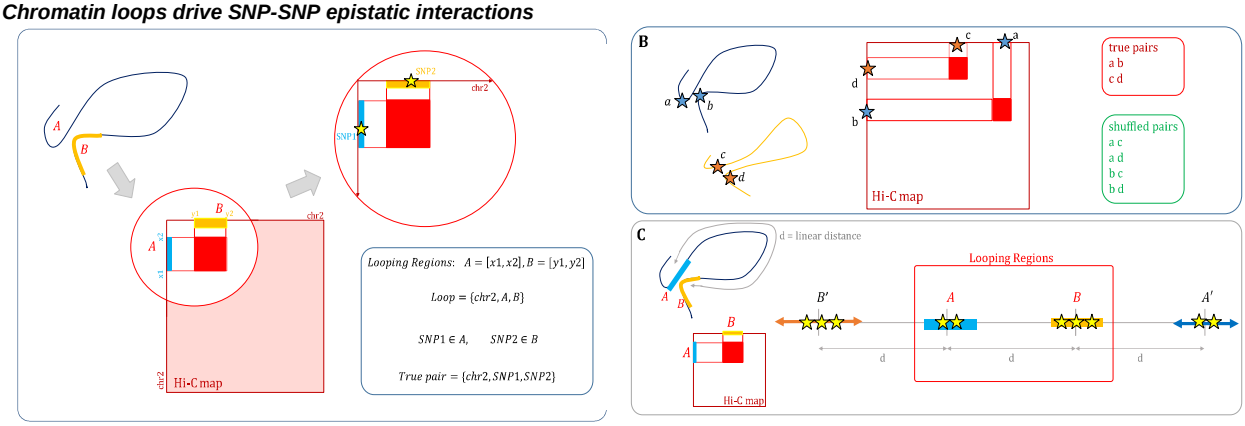

Fig. 2. Definitions: loops, true and artificial pairs. A: How a loop is defined, and how true pairs are defined after it. Regions A and B form a loop in the 3D chromatin conformation, which is detected as a peak (in brighter red) in the Hi-C contact map (sketch). Those two regions are on the same chromosome (chr2) and have the same length. They are characterized by a starting position (in bp, x1 and y1 respectively for A and B) and an ending position (x2, y2). To define pairs, we can look at SNPs found on the two genomic regions. SNP1 happens to sit on region A, SNP2 on region B. The pair SNP1, SNP2 (on chromosome 2 in this example) is therefore a true pair. Figures **B, C:** Two approaches to generate artificial pairs. In **B** only SNPs on looping regions are taken into consideration. Here we have two loops, involving four regions, and one SNP per region for simplicity. Note that since there is only one SNP per region, we extend the notation A, B, C and D to the SNPs. Artificial pairs are randomly 'shuffled' matches of SNPs in loops. In **C,** for every region a region loops to, there is another region, at the same linear distance on the other side, which is not involved in a loop. A and A' (similar to B and B') are at same linear distance from B (A) and contain the same number of SNPs. The only reason why pairs of SNPs sampled from AB would perform differently than pairs of SNPs sampled from A'B (and AB') is the fact the A and B engage in one long-range chromatin interaction that puts them in physical contact in the 3D space.

Therefore, we expect each of the two regions to contain several SNPs. This means that from one loop, we can build multiple pairs of SNPs. Those pairs, furthermore, are not independent: if we were to take all possible combinations, multiple pairs would include the same SNP. Moreover, all SNPs in the same region are so close to each other, that very likely they will be in strong linkage disequilibrium (LD). It can be therefore necessary to choose one pair of SNPs 'representing' the loop. We define:

'Representative' pairs (per loop): the chosen pairs to represent one loop's behaviour. In different situations those are the pairs for which the association with the phenotype is highest. For one loop:

$$\underset{i\in A, j\in B}{\operatorname{argmax}}\{assoc(geno_i, geno_j \leftrightarrow phenotype)\}$$

Where $geno_i, geno_j$ are the genotypes of SNPs *i, j* and *A,B* are the two regions forming the loop.

It can also be one of the possible combinations, randomly sampled:

$$i, j \quad | \quad i \in A, j \in B$$

The second approach allows us not to compute the association for every single pair. We believe this choice is valid, since SNPs belonging to the same region are expected to have similar association with the phenotype. We have indeed verified this is the case, as we show in Fig S7.

Artificial pairs: throughout the rest of this work, in order to verify that true pairs actually improve GWAS association by engaging in epistatic interaction, we will need to compare their association with the association that random pairs would obtain, as reference. Artificial pairs are pairs, as defined earlier, but the two SNPs are not on looping regions. The construction of artificial pairs is shown in Figures 2B and 2C and described in Detailed Methods.

2.3 Modeling epistatic interaction with logistic regression models

Using logistic regression models to measure GWAS association is common practice (Wason *et al*, 2010, Bush *et al*, 2012). The once very popular 'contingency table' methods, such as the Fisher's exact test and the χ^2test, measure the association between genotype and phenotype based on a pure count: if one SNP is observed (significantly) more often in affected samples than in unaffected ones (for one disease of interest), then that SNP is associated with an increased risk for that disease. However, many confounding factors could arise and provoke false positives (or hide interesting discoveries). Sex, age and ancestry of the analyzed samples can have a major impact on the results, and need to be corrected for. In a logistic regression model, a binary outcome (e.g. sick or healthy, in a case-control setup) is modelled using multiple factors. Logistic regression is designed to include multiple predictors, other covariates as well as the SNPs genotypes. Their ability to account for confounders, particularly population structure, makes the regression model the preferred association model in recent GWAS (Pulit 2016). Due to the large number of testable SNPs, however, the majority of such studies perform single-locus analyses only. Being able to include more predictors does not come for free: logistic regression models, compared to χ^2 tests for example, are much more computationally expensive (Shete *et al*, 2009, Chen *et al*, 2011, Chahal *et al*, 2016). Moreover, the complexity grows as we add more predictors, since more parameters need to be estimated.
Here, we prioritize the candidate SNP-SNP pairs, reducing drastically the number of models to be tested. This allows us not only to use logistic regression models (including covariates accounting for confounders), but more importantly to include multiple SNPs at a time, and their interaction. For one pair SNP_1, SNP_2 , the interaction model is:

$$logit\big(P(diseased)\big) \sim covar + SNP_1 + SNP_2 + SNP_1 SNP_2 \qquad (0)$$

To measure whether this model (0) improves the prediction compared to models that include the two SNPs taken individually (1)(2), we use an LRT approach, described in the next paragraph (2.4).

$$logit\big(P(diseased)\big) \sim covar + SNP_1 \qquad (1)$$

$$logit\big(P(diseased)\big) \sim covar + SNP_2 \qquad (2)$$

The logistic regression here acts as a classifier, which predicts class 0 (unaffected) or class 1 (affected) based on some features, or predictors (here the SNP genotypes), which can be better or worse at discriminating between the two classes. If one SNP (say SNP_1) was already very highly

associated with the phenotype, then the model that includes SNP_2, too, would also result in a good prediction. We are not interested in these kinds of pairs. Instead, we aim to identify pairs of SNPs whose individual association with the phenotype is rather low. From a classification perspective the individual SNPs genotypes have low discriminative power. Their interaction, however, (the term $SNP_1 * SNP_2$ in the model) is a good predictor. If this situation happens, we have detected an epistatic effect.

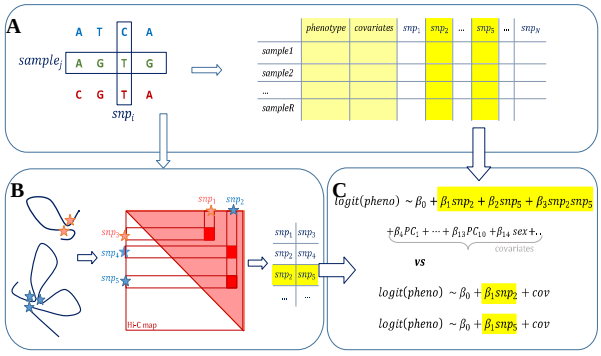

Fig3. Overview of our method. Integration of GWAS (A) and Hi-C loops (B) in a logistic regression framework (C) to detect SNP-SNP epistatic interactions. In **A**, after cleaning the GWAS data, we build a 'sample matrix': for every sample (row), we know the phenotype – this is encoded as 0 (controls, healthy, unaffected) and 1 (cases, diseased, affected). Furthermore, we have collected some covariates: the sex of the individuals and the first 10 principal components, which account for population structure. Finally, we have genotypic information for all sites (SNPs). We build one sample matrix per chromosome. In **B**, based on the peaks called from the Hi-C map (schematic representation), which identify the 3D genome organization into loops, we define true SNP-SNP pairs. Different colours indicate SNPs on different loops. Only the upper triangular portion of the Hi-C matrix is coloured, to indicate that since the matrix is symmetric looking at one half is sufficient. To explain the rest of the procedure, we select one of the pairs, snp_2snp_5 (highlighted in yellow). **C:** every time one SNP-SNP pair is selected, the corresponding columns in the sample matrix are selected and used in building the logistic regression model (s), together with the phenotype and the covariates. Every model that includes both SNPs in a pair, and their interaction, is then compared to the two individual models of the two SNPs taken individually to assess whether there is in fact an epistatic effect.

2.4 Modeling GWAS association improvement with LRT

To compare two logistic regression models, and have a statistical measure of the improvement of one model over the other, we use a likelihood ratio test (LRT)(Neyman & Pearson, 1933). An LRT is a statistical test designed to compare the goodness of fit of two models, where one is a special case of the other. The two models must be nested: the set of parameters of one must be a subset of parameters of the other. If that is the case, then the likelihood ratio statistic is:

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

According to the Wilks' theorem (Wilks, 1938), moreover, for a sufficiently large number of samples,

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

where the number of degrees of freedom k is given as $|\theta| - |\theta_0|$.
Likelihood ratio tests are often used in Statistics, Probability and Economics (Yang, 1998, Bai, 1999, Moreira, 2003). However, they are seldom used in Bioinformatics (Marioni *et al*, 2008, Basu *et al*, 2011, Petersen *et al*, 2013). LRTs have been used in the context of GWAS, for example in Cortes *et al*, to analyze in detail the MHC (major histocompatibility complex) on chromosome 6, and detect whether considering all variants in one region improves association over single variants (Cortes *et al*, 2015), but never truly genome-wide, to the best of our knowledge.
For every pair of SNPs - SNP_1SNP_2 - we perform the test twice, comparing the full model (0), first with (1), then with (2). This provides us with two χ^2 statistics, and two derived p-values. When both p-values are significant, we have identified a synergistic effect.

2.4.1 Selecting epistatic pairs

To find SNP-SNP pairs that engage in epistatic interactions, we select all pairs for which the interaction model (0) is better than the single SNP models (1) and (2), as described previously. Those pairs already show a synergistic effect, as their interaction reaches higher association with the phenotype than the two SNPs taken individually. Out of those, we only keep SNP-SNP pairs that show association in absolute terms as well. This does not happen automatically even if (0) is better than (1) and (2). An LRT measures whether one model is significantly better than the other. As including more features naturally improves the prediction, larger models are punished, since they require more parameters to be estimated and therefore increase their complexity. It can therefore happen, if the two SNPs are very poorly associated with the phenotype, that their interaction has higher association than them, but still not high enough to pass the significance threshold. We measure this again with an LRT. Now (0) is compared with the null model:

$$logit\big(P(diseased)\big) \sim covar \qquad (3)$$

The pairs that have passed this selection perform better than the individual variants, and well overall. Furthermore, we want to verify that we are truly capturing a synergistic effect, on top of the additive one. For every pair, we compare (0) with the additive model:

$$logit\big(P(diseased)\big) \sim covar + SNP_1 + SNP_2 \qquad (4)$$

with an LRT. To be even more stringent, we compare each candidate pair SNP_1, SNP_2 with two artificial pairs. Those pairs are built such that SNP_1 is paired up with SNP_2', which is at similar distance from SNP_1 as SNP_2 is, but on its other side. Similarly, the other artificial pair is SNP_1', SNP_2 (See Detailed Methods and Figure 2C). We only keep the true pair if the corresponding artificial pairs do not satisfy the other two conditions. To sum up, we define a true pair SNP_1, SNP_2 to be a candidate epistatic pair if $(y = logit\big(P(diseased)\big))$:

1. $y \sim covar + SNP_1 + SNP_2 + SNP_1SNP_2 > y \sim covar + SNP_1$
   and
   $y \sim covar + SNP_1 + SNP_2 + SNP_1SNP_2 > y \sim covar + SNP_2$
2. $y \sim covar + SNP_1 + SNP_2 + SNP_1SNP_2 > y \sim covar$
3. $y \sim covar + SNP_1 + SNP_2 + SNP_1SNP_2 > y \sim covar + SNP_1 + SNP_2$
4. 1. 2. and 3. do not hold for SNP_1, SNP_2' and SNP_1', SNP_2, artificial pairs.

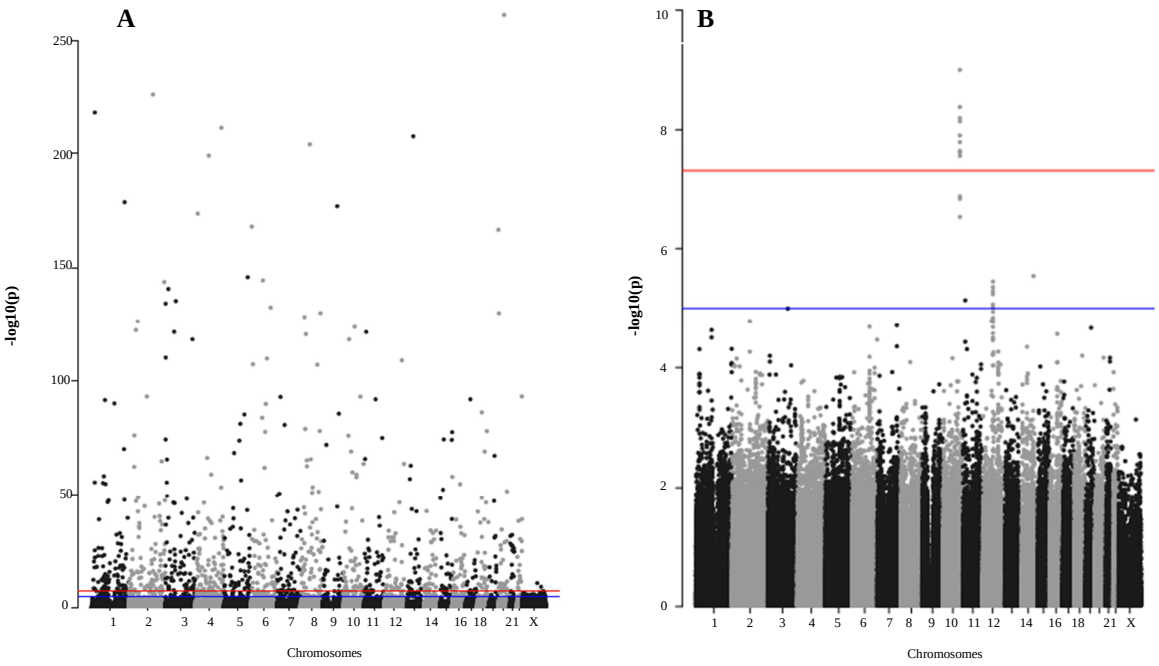

Fig 4. Manhattan plots, before and after QC. A. Before: here we show the results of the first GWAS experiment we performed on the WTCCC T2D dataset. We included all the variants and all the samples available, 495,472 and 3,499 respectively. We performed the association with a Fisher's exact test. The results are not simply noisy, they are completely misleading. Hundreds of variants shows an association with the phenotype, above the threshold, some even with p-values < e-200. The characteristic little columns of SNPs, representing LD blocks, where close variants show the same association, so clear in the right plot, are hardly noticeable on the left. There is no simple reason for these results. The human genetic heritage is extremely complex, and many factors play a role. To capture one single aspect, i.e. the predisposition to a disease of interest, we must take several measures to correct for other factors. For example, individuals from different geographic areas have several different SNPs. Individuals that are related, on the other hand, have extremely similar genomes. All these aspects introduce biases, thus those individuals must be removed from the study. As for SNPs, SNPs that are extremely rare might just by chance only be detected in one of the two classes, and be mistakenly found to have an extremely high association. For technical reasons, we might have information about one variant for cases, but not for controls. Missingness is also an element to correct for, and there are many more examples like those. **B. After:** after sample and site QC, we perform a better informed experiment, with 450,242 variants and 3,433 samples. Alongside eliminating noisy samples and sites, we calculate the association with a logistic regression, including the sex of the individuals and the first 10 principal components (PCs) which account for population structure as covariates. The results look much better, and reproduce the results found by the original paper. The LD columns are easily spotted, and only very few sites make the significance threshold (plots generated using the R package qqman, Turner 2014).

3 Results

Data preparation

The WTCCC dataset is perhaps the most used GWAS dataset in subsequent GWAS analysis studies, for many different venues, including epistatic analysis (Wan *et al*, 2009, Hu *et al*, 2010, Wan *et al*, 2010, Yung *et al*, 2011, Lippert *et al*, 2013). To utilize the data, which come as raw genotypes, a great cleaning effort is required. To produce a manageable dataset the raw genotypes must undergo an extensive pipeline. The QC steps (described in detail in the Supplementary Information, section B) are extensive, non-trivial and truly needed. With QC, we filtered out a rather small but not negligible portion of both SNPs and samples (~10% in both cases). Moreover, it is very important to include population structure confounders in the association model. Figure S4 shows the slight improvements as we performed the QC steps, and the biggest jump in quality occurred when we added the first ten principal components (PCs), as covariates to the logistic regression model. The first two principal components alone are able to identify the largest ethnicity clusters (Fig S2). To emphasize the importance of QC, in Fig 4 we show the difference between the results of the GWAS experiments that we performed on the T2D dataset, before and after QC. We represent the GWAS results with a Manhattan plot, as is customary. In a Manhattan plot every SNP is represented by a dot measuring its –log10(p-value), in association with the phenotype. Variants on all chromosomes are shown, with the different chromosomes indicated by the alternating colours. Fig 4 clearly shows that without performing a thorough quality control the results of GWAS are biased, and can be misleading. As every dataset is different, the steps should be flexible and data-driven. The SNPs that appear to be highly associated in Fig 4A are either extremely rare, or are contained in very variable regions, and should not be included in the experiment. Importantly, if we do not correct for biases introduced by the various ancestries, we are likely to capture ethnic differences across the samples, rather than the case-control discrimination we are interested in. Such correction can be done by including the first principal components (PCs) as covariates in the logistic regression. We described all the QC steps in detail in the Supplementary Information, and hence hope to contribute an easily accessible GWAS QC guide for non-geneticists.

3.1 Genomic regions that engage in long-range chromatin interactions are enriched for common variants

The existence and formation of chromatin loops demonstrated that the chromatin fibre can be extremely flexible, at small scales, as described in Sanborn *et al.* (Sanborn *et al*, 2015). Observations have shown that in order to form loops, the chromatin is often found in an open state. Chromatin loops are believed to form for regulatory purposes. Thus, we explored whether the loops we have collected do in fact frequently host enhancers.

3.1.1 Chromatin loops are associated with enhancer activity

To investigate the relation between the chromatin conformation and its regulatory function we overlaid our chromatin loops with validated enhancers from the Epigenome Roadmap, derived from the same cell as the Hi-C loops, GM17828 (Roadmap *et al*, 2015). We found that a large portion of enhancers are contained in loops, compared to how little coverage the same loops have (see Detailed Methods). The difference in percentage is striking (Fig 5A), and the enrichment is significant for all chromosomes (Binomial test, worst p-value is 5e-14). Although perhaps not surprising, this finding is encouraging, and justifies our quest for epistatic interactions driven by enhancer-promoter chromatin loops.
Chromatin loops are enriched for enhancers, as we have shown. Although to a lesser extent than for genes, we expect enhancers to be particularly prone to common sequence variation, even more so since they are regions of open, and therefore accessible, chromatin.

3.1.2 Loops are enriched for SNPs

To study the relation between the SNPs distribution and the chromatin loops, we collected an extensive list of positions of all known common variants from the UCSC repository (University of California, Santa Cruz) and overlaid them with our loops, collected from Sanborn *et al.*
As we show in Fig 5B, SNPs are more often in loops compared to non-variable base pairs. The portion of SNPs in loops (light blue bars, out of the total number of SNPs) is consistently larger than the loop coverage (light red bars, base pairs in loops) as calculated from the Hi-C map, for all chromosomes but one. The enrichment is significant for most chromosomes (Fig 5B). Specifically, 18/23 (78.3%) using Bonferroni multiple testing correction, and 19/23 (82.6%) using Benjamini-Hochberg. The significance is calculated using a Binomial test (see Detailed Methods).
We have shown that there is a bias toward regions engaging in long-range interactions, which are more prone to containing common variants. First of all, this confirms the validity of overlaying GWAS and Hi-C data-types, even though they are obtained from different cell-types. Secondly, this can have an impact on genomic analyses involving SNPs. For example, this finding implies that common variants are not uniformly distributed along the genome. Genomic methods that include random sampling of SNPs, among others, should take this information into account.

3.1.3 Looping regions are more enriched for sites that are associated with diseases

To determine whether looping regions are particularly enriched for SNPs that are associated with the disease (T2D), we then looked at the association of the SNPs with the phenotype, alongside their positions. To do so, we performed a one-sided rank-sum test on the p-values of the SNPs in loops, versus those sitting on non-looping regions. We found that when we consider all variants, this is not the case (Mann-Whitney U test, p-value = 0.48). However, it is not all variants we are interested in. We are investigating whether loops are enriched for highly associated variants, thus it makes sense to only look at the 'lower tail' of the p-values distribution, at those SNPs that do have an effect, although minor, on the disease of interest. Indeed, when we perform the test again, this time only for SNPs with an association measure of P<0.5, looping regions are enriched for SNPs (Mann-Whitney U test, p-value = 0.00593), and they are even more so when we reduce the threshold further (P<0.2) (Mann-Whitney U test, p-value = 0.00386) (Fig 6). The enrichment is true genome-wide, and at the single-chromosome level too (largest p-value = 0.0099, Fig S6C). Hence, chromatin loops are not only enriched for common variants, but particularly so for those that are disease-associated.

We conclude that disease-associated SNPs preferentially occur in looping chromatin, and that loops are associated with higher enhancer activity, as we have shown, and as it has been observed before (Dixon *et al*, 2012, Babaei *et al*, 2015). These two findings combined represent a promising starting point for the rest of this work. Moreover, the results found in the last paragraph (3.1.3) suggest the possibility to leave out variants that show no association at all, as they seem not to carry any signal and no potential for epistatic interactions. In fact, by definition, an epistatic interaction implies that the effect of one variant influences the effect that another variant has on the phenotype, either by silencing it or enhancing it, for instance. If there is no initial effect on the phenotype in the first place, there cannot be an epistatic effect either.

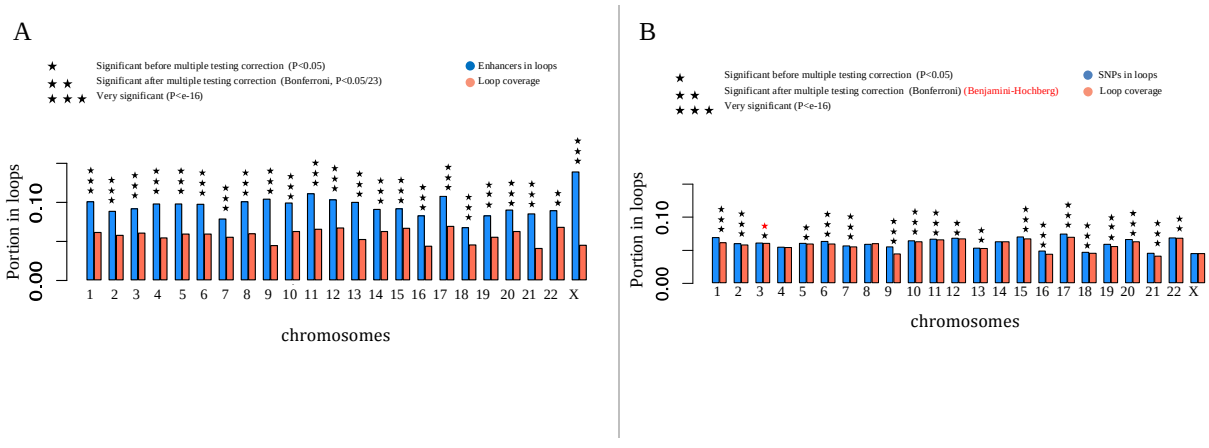

Fig. 5. Loops are enriched for common variants, especially when associated with the phenotype, and for enhancer activity. In **A** compared bar-plots: the portion of enhancers in looping regions (blue bars) is much higher than the coverage that those loop have (red bars). We tested with a Binomial test how significant the difference is. Stars indicate the level of significance, as indicated. No star means the test was not significant. The enhancers' positions are downloaded from Roadmap *et al.* X axis, chromosomes, y axis portion in loops. **B** similar to A, but for SNPs. The portion of SNPs in loops is systematically (with the exception of chromosome 8) larger than the loop coverage of the chromosomes, in base pairs and based on the Hi-C maps. The stars show how significant the difference is, per chromosome, as obtained with a Binomial test. With the Binomial test, we test if the number of common variants found in the loops is larger than expected by chance, and it appears to be so in the majority of the experiments. Two stars indicate significance after multiple testing correction, using a Bonferroni correction. We also used Benjamini-Hochberg. The red star for chromosome 3 means that the test is only significant for Benjamini-Hochberg, and not for Bonferroni.

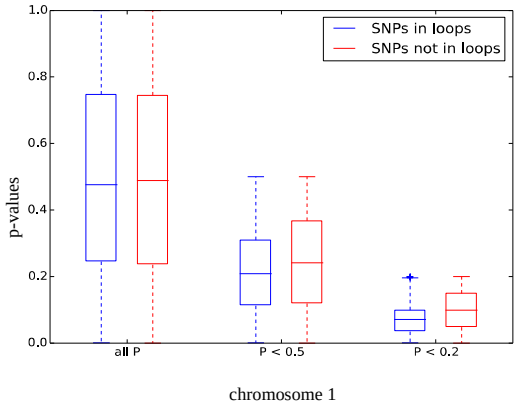

Fig. 6. Loops are enriched for variants that are associated with the phenotype. Loops are enriched for highly associated SNPs, when we select only SNPs that show a little association with the phenotype on their own. Results are illustrated for chromosome 1. The distributions of p-values of SNPs in loops (blue) and not in loops (red) are compared using boxplots. We furthermore compare such distributions when we reduce the initial set of p-values, by considering all SNPs with P<0.5 first, and then P<0.2. On the y axis are the p-values, on the x axis the three experiments, for different P thresholds. We can see that the difference between the two distributions (SNPs in loops, SNPs not in loops) grows larger, the more we reduce the p-value threshold.

3.2 Detecting epistatic interactions

By only selecting pairs in loops, we dramatically reduce the number of SNP-SNP pairs to be tested. If we were to take all pairwise combinations of our (450,242) variants, we would have to test over 100,000,000,000 hypotheses. Even if we were to start with only SNPs with some association with the phenotype (P<0.5), the number of tests to be performed would remain huge (>20,000,000,000). By select pairs in loops, instead, we only need to test 35,425 pairs, and have therefore much more statistical power. The number of pairs varies from chromosome to chromosome, but not in a surprising fashion, rather according to the chromosome's length (Pearson's correlation = 0.8152).

3.2.1 SNP-SNP pairs engage in epistatic interactions more often than expected by chance

We found that, genome-wide, epistatic effects occur more often in true pairs than we would expect by chance. As reference, we built artificial pairs. The artificial pairs are built as a re-shuffling of the true ones, in such a way that only SNPs that are on some looping region are used, but the true matches are disrupted (Fig 2B). When modelling the artificial pairs, we want to keep the structure similar to that of the true pairs as much as possible, to capture the role played by the three-dimensional chromatin conformation only. As we have shown, genomic regions that engage in long-range chromatin interactions are associated with enhancer activity, and are hotspots for common single nucleotide polymorphisms. By shuffling the pairs, therefore, we take SNPs that all belong to one of those hotspots, but we disrupt the loop structure, by combining SNPs sitting on two regions that do not loop over each other (Fig.2B). In this way, we aim to detect the effect that real chromatin loops have on the interaction between variants (see Detailed Methods).
We define an epistatic interaction between two SNPs as described in the approach section (2.4, 2.5). Thus, we counted for how many pairs the full interaction model (0) significantly improves the association with the phenotype compared to the models with the two individual SNPs only (1) and (2), as measured with an LRT. We then shuffled the SNPs into the same total number of artificial pairs, and counted the same number

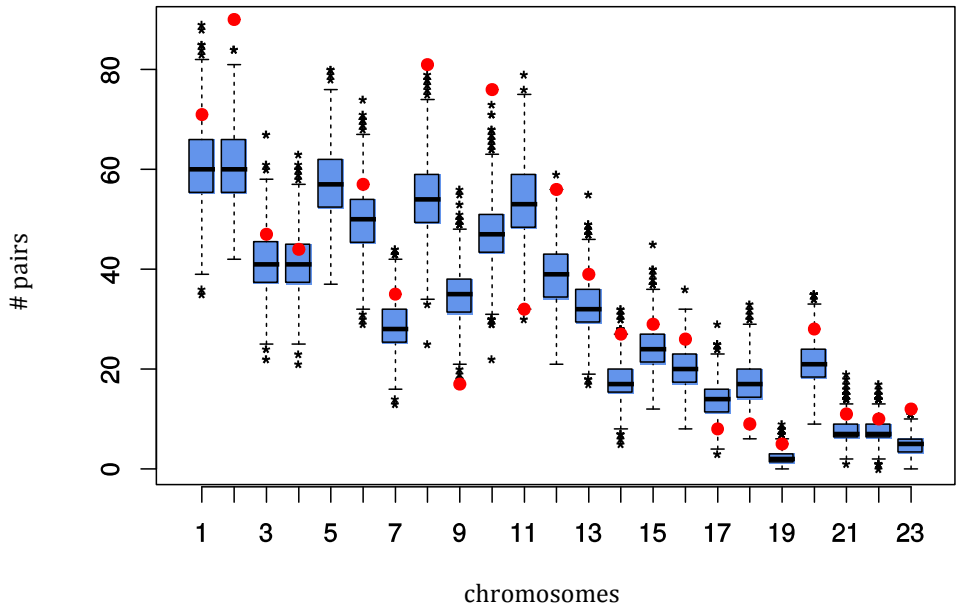

Fig. 7. True pairs engage in epistatic interactions more often than expected by chance. Boxplots for all chromosomes. For every chromosome we counted how many true pairs engage in epistatic interactions (red dot). Then, we re-shuffled the pairs 1000 times and at each permutation we counted the number of epistatic pairs. The resulting distributions, one per chromosome, are shown as boxplots. We observe that overall, true pairs engage in epistatic interactions more often than random. The difference is obvious for some chromosomes, such as 2, 5, 8, 10, 19 and 23. For others, it is less evident. In general, we observe great variation across the chromosomes. Extremely puzzling, furthermore, are chromosomes 9, 11, 17 and 18, which show an exactly opposite trend.

for the resulting artificial pairs. For every chromosome, we used 1000 permutations and built the null distribution onto which we compared the number for the true pairs (Fig 7 and Fig S8). Across all chromosomes, the number of pairs that show GWAS association improvement is rather low compared to the total number of pairs (~3%). However, overall, it is larger than the same number for the artificial pairs. As the null distribution obtained with the shuffled pairs was not Gaussian (Shapiro-Wilk test, largest p-value 0.0024), we calculated an estimated p-value, for all chromosomes, as:

$$\hat{P} = \frac{a+1}{b+1}$$

Where a is the number of values more extreme than the count for true pairs (red dot in the plots) and b is the number of permutations (10000). By definition, this p-value cannot be smaller than 0.001 (1/1000). Over all chromosomes, P = 0.01 (average p-value, see Supplementary Figure S8).

3.2.2 Epistasis occurs with great diversity from chromosome to chromosome

Although genome-wide epistasis occurs overall more frequently for true pairs than for shuffled pairs, which supports our hypothesis that chromatin loops could drive SNP-SNP epistatic interactions, the results are extremely variable when we look at individual chromosomes (Fig 7, Fig S8). While for some chromosomes artificial pairs never perform better than the true ones (a = 0) and for some other chromosomes the difference is not as striking. Finally, four chromosomes (specifically 9, 11, 17 and 18) show a radically opposite behaviour. Unfortunately, we are not able to explain this extremely unexpected result, for now.

3.2.3 Synergistic pairs often show phenotype association

We next established which epistatic pairs are also highly associated with the phenotype, as described in 2.4.1. To this end, we collected all pairs for which the interaction model (0) performs better than the two single-loci models (1) and (2), and checked whether it also performs significantly better than the null model (3). We found that this was true for more than 70% of the pairs, averaged over the chromosomes (Likelihood ratio test, P<0.05, see Supplementary Figure S9B). In order to demonstrate that the improvement in association is truly due to the epistatic interaction between the variants, rather than simply caused by their sum, we compared the interaction model (0) with the additive model (4), as described in the approach section (2.4.1). We found that for 96% of the pairs the interaction model explained the phenotype significantly better than the additive (Fig S9C). Moreover, for 71% of the rest, the association of the interaction model with the phenotype was higher than that of the additive, although not significantly. Among the remaining 9 pairs, no additive model passed the significance threshold (best p-value = 0.0013). These findings confirmed our hypothesis that epistatic effects between SNPs, rather than just additive effects can actually yield large effects on the phenotype.

Next, in order to demonstrate that the association of the selected synergistic pairs is higher than what we would expect by chance, we performed the same tests for artificial pairs. To this end, we first grouped all pairs based on the loop they were built from. Every loop connects two regions, at a given linear distance. For each of these two regions we selected another portion of the genome, on the other side, but at the same linear distance (Fig 2C). We can now sample SNPs from these two new regions, and build artificial pairs (see Detailed Methods). We filtered out loops for which more than one artificial pair passed both tests (LRT between interaction model and both individual models and LRT between interaction model and null model), and the deriving pairs, which only made up for 7% of the total (Fig S9D). In Figure 8 we show the Manhattan plot of the remaining 607 pairs, built from 301 chromatin loops, over all chromosomes. Every dot in the figure represents a pair of SNPs, rather than one single SNP. The p-values are obtained as the result of the likelihood ratio test between the interaction model (0) and the null model (3). This is a measure of how well the epistatic interaction between those SNPs can predict the phenotype.

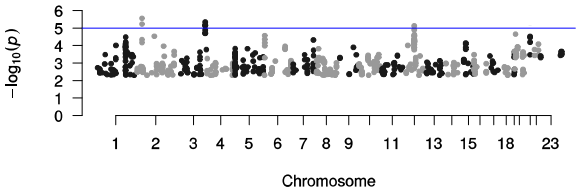

Fig 8. Manhattan plot of candidate epistatic pairs. Manhattan plot showing the association with the phenotype of SNP-SNP pairs, rather than individual SNPs. In this plot every dot is one pair. Other than that, the rest is similar to other Manhattan plots. As usual, the y axis indicates the –log10(p-values), the x axis the alternating chromosomes. As all pairs are within the same chromosome, we kept this distinction. As for genomic position in base pairs, we use the mean position of the two SNPs, as a default. The blue line represents the significance threshold, after multiple testing correction. We here show only the 607 candidate pairs, thus the Manhattan plot is much more sparse than usual, and all dots are above y=2, as we previously selected 'good' pairs. Note that the little columns here do not indicate LD, but rather pairs from the same loops. Three sets of pairs, from chromosome 2, 3 and 12 pass the threshold.

Fig 8 shows a few pairs for which the interaction model (0) passed the significance threshold. These 7 pairs derive from 2 loops on chromosomes 2, 3 and 12. All those pairs show a striking epistatic effect on the phenotype (Fig 9).

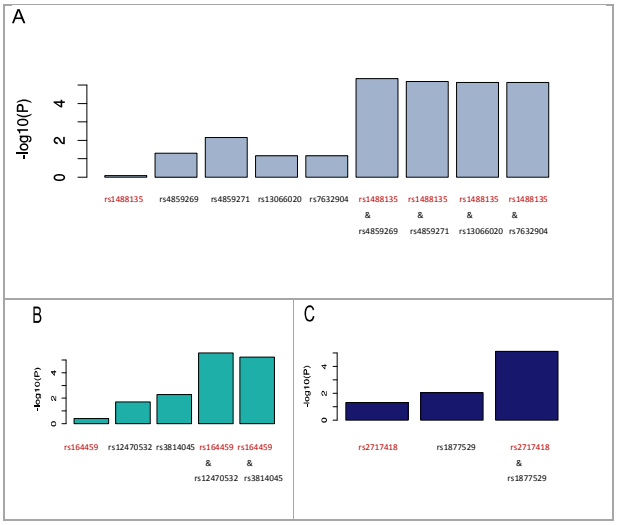

Fig 9. Synergistic pairs, compared association between single-variant models and interaction model. In **A**, 4 out of the best 7 pairs derive from the same loop, on chromosome 3. To show the increased performance of the pairs compared to the single-SNP models, we compare the –log10(P) of the models' association with the phenotype. The epistatic improvement is evident (around 4X improvement). Moreover, we observe that different pairs deriving from the same loop perform extremely similarly. **B** and **C** indicate similar behaviours to the two pairs on chromosome 2, and the pair on chromosome 12. SNP IDs and chromatin loops coordinates are indicated. X axis: SNPs or pairs of SNPs of the model tests. Y axis: –log10(p-values), p-values obtained as an LRT of the corresponding model vs the null model.

3.2.4 The identified top seven synergistic pairs involve SNPs on enhancer active regions or genes involved in metabolic pathways

The top seven identified epistatic pairs (Figures 8 and 9) are built from chromatin loops on chromosomes 2, 3 and 12. We briefly investigated known annotation of the involved variants and the surrounding genomic regions. First of all, we found no overlap between our selected SNPs and a list of known T2D-associated variants from the GWAS catalogue (see Detailed Methods). Moreover, none of the selected have a CADD score higher than 8. CADD scores give an indication of how deleterious one variant is, based on a number of features (Kircher *et al*, 2014). These two observations confirm our hypothesis that SNPs engaging in epistatic interactions are not known or predicted disease-associated variants, when taken individually.

The two SNP pairs on chromosome 2 both involve SNP rs11694459, contained in one of the looping regions, or anchors of a loop (chr2:46,400,000-46,425,000). This SNP shows synergistic effects on the phenotype when interacting with rs12470532 and rs3814045 which are found very close to each other, at the other anchor (chr2:46,725,000-46,750,000). The first SNP (rs11694459) sits a few base pairs downstream of gene PRKCE, which is found in the type II diabetes mellitus pathway (Fig S10). The other two SNPs (rs12470532 and rs3814045) are just upstream of gene ATP6V1E2. ATP6V1E2 is …, and is involved in multiple pathways, including the rheumatoid arthritis (RA) pathway (Fig S10). It is known that complex diseases often share risk genomic loci, so this could be an example of this phenomenon. However, we make another observation. Only a few base pairs downstream of ATP6V1E2 is another gene, RHOQ. This gene belongs to the insulin signaling pathway (Fig S10), which is one pathway that is disrupted in diabetes. Although the SNPs do not sit directly on either of the two genes, the entire area is highly enriched with H3K4me1, which is a known mark for enhancers (Fig 10).

The four epistatic pairs we detected on chromosome 3 all involve SNP rs1488135, interacting with rs4859269, rs4859271, rs13066020 and rs7632904. The first is found upstream of a long non-coding RNA (FLJ46066), highly enriched for H3K4me1. The others sit on the MCCC1 gene, which hosts numerous other common variants. This gene is involved in multiple metabolic pathways and its malfunction causes a number of congenital disorders of metabolism (Fig 10).

Finally, the SNP pair we discovered on chromosome 12 is composed of SNPs rs2717418 and rs1877529. The first sits on the PTPRB gene, whose function is in blood vessel remodeling and angiogenesis, and is enriched for H3K4me1. The second is downstream of the PTPRR gene, which is part of the MAPQ signaling pathway, and shows high levels of H3K4me1 and H3K27ac, both marks for enhancers (Fig 10). Interestingly, both genes also play a role in cancer.

We observed that both pairs on chromosome 2 and all four pairs on chromosome 3 derive in fact from two loops only. In order to investigate whether those loops considered as a whole would have a higher association score than their deriving pairs, we measured the performance of a logistic regression model including all involved variants, with an LRT. For both loops, however, the models per pair outperformed the model per loop. Moreover, pairs built out of the same loop have extremely similar association with the phenotype. This result confirms and extends an earlier observation: SNPs in close vicinity show similar association measures not only when considered individually, but also when paired with other SNPs (Fig 9 and Fig S7). Thus, including more than one SNP from one region is redundant, and does not improve the model.

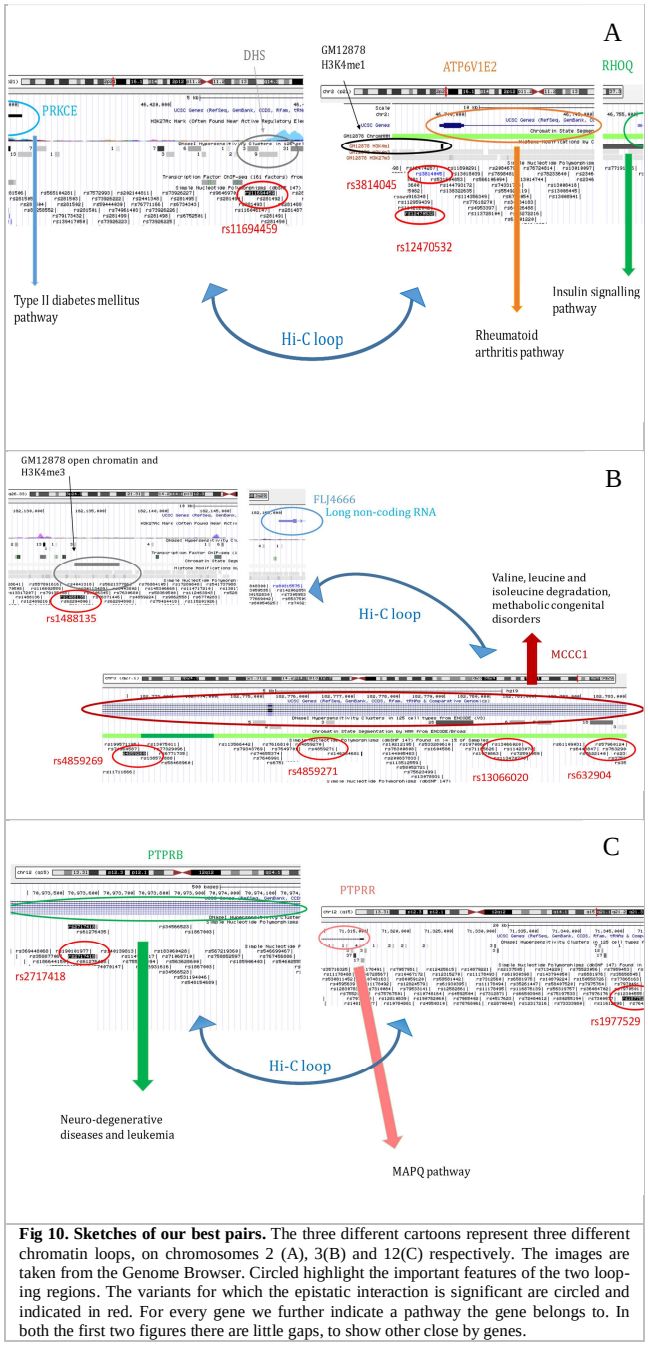

Fig 10. Sketches of our best pairs. The three different cartoons represent three different chromatin loops, on chromosomes 2 (A), 3(B) and 12(C) respectively. The images are taken from the Genome Browser. Circled highlight the important features of the two looping regions. The variants for which the epistatic interaction is significant are circled and indicated in red. For every gene we further indicate a pathway the gene belongs to. In both the first two figures there are little gaps, to show other close by genes.

4 Discussion

In this article, we present an analysis to explore the contribution of the three-dimensional genome organization to GWAS epistasis. To this end, we overlay the WTCCC type 2 diabetes GWAS dataset with long-range chromatin loops measured with Hi-C. We find that chromatin loops show a significant enrichment for common SNPs, especially for SNPs that are associated with the disease, and that they are extremely enriched for enhancer activity. These results indicate that long-range chromatin interactions form for regulatory purposes and generate hotspots of dis-

ease-associated common variants, in the 3D context. This suggests that the 3D genome organization plays a role in driving GWAS epistasis, by yielding the co-occurrence of SNPs on promoter-enhancer pairs, or pairs of enhancers targeting the same gene.
We propose a statistical framework to detect epistatic interactions, where we model GWAS association with a logistic regression and measure the improvement of the synergistic model over the single-loci models with a likelihood ratio test. We identify a handful of pairs showing epistatic interaction, which are built from three different loops on chromosomes 2, 3 and 12. All variants involved in these pairs were not previously known to be related to T2D, nor known as deleterious variants according to the CADD score. This supports our hypothesis that epistatic effects occur as synergistic interactions of variants that are not highly enough associated with the phenotype on their own. Moreover, these SNPs are either sitting on regions enriched for histone modification marks associated with enhancers (H3K4me1 and H3K27ac), or are found on or in close vicinity with genes involved in metabolic pathways, including the type 2 diabetes mellitus pathway and the insulin signaling pathway. First of all, these findings confirm that analyzing chromatin loops helps identify enhancer-promoter pairs, and that those indeed are able to drive GWAS epistatic interactions. Moreover, our proposed statistical approach is able to successfully to capture true epistatic interactions, which involve genes that appear to be related to the disease of interest. As a preliminary biological validation those findings are very promising, and we expect interesting future results after further and thorough genetic analyses.

5 Future work

This work can be improved and extended in a number of ways. First of all, we only considered chromatin loops called from the Hi-C map with very high confidence. With a slightly less stringent threshold we could have examined more loops and potentially found more information, maintaining still a reasonably small number of tests to be performed. Moreover, the three-dimensional genome organization could have been further investigated at other scales as well, for example looking at whether we find epistatic interactions within TADs (topological association domains), more often than between TADs. Secondly, we were very strict in the detection of our final epistatic pairs, which had to pass many tests and therefore needed to be very highly associated to pass the significance threshold after multiple testing correction. Again, more lenient thresholds might have led to further discoveries, although we could not be as confident about them. Thirdly, further biological interpretation of our findings would help validate and interpret our results. A thorough integration of the Hi-C chromatin loops with epigenetic states from Roadmap *et al*, for example, could yield interesting results, while in this work, we only looked at enhancers.

Extension to more GWAS phenotypes and Hi-C cell lines

In this study, we focused on one single phenotype, type 2 diabetes. There are many other GWAS datasets, starting with the other six diseases of the WTCCC study (bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis and type 1 diabetes). The investigation of similarities and differences among different diseases could shed new light on the specific diseases etiology and common mechanisms. Promising findings in other datasets, moreover, would provide a nice validation for our method. Furthermore, including multiple diseases could be combined with collecting Hi-C maps from different cell-lines, and investigating the association of different cell types with different diseases. However, those analyses would still incur the same limitation that we encountered, due to the difference in the cell-lines that two data types (GWAS and Hi-C) are obtained from.
While we showed that the integration of the two data types allowed us to find interesting results, we cannot imply a direct connection unless the data was collected from the same cells. A substantial improvement in the interpretation of the results and their quality would occur if we could integrate GWAS data and Hi-C maps from the same source.

Correlation between SNPs that are linearly distal but spatially co-localized

Another possible direction for investigation is the exploration of a potential correlation between the genotypes of variants co-localizing in the 3D context. Similar to how linearly close variants are highly correlated due to LD, we wonder whether there is something analogous for variants in close 3D proximity, a sort of a map of conserved three-dimensional blocks. Calculating a correlation measure between bi-allelic loci with unknown phase is hardly a trivial task (Rogers and Huff, 2009). We attempted a simple approach, using a χ^2 test to compare expected and observed co-occurrence of different alleles at two loci. Although we did find some mild correlations, the results were not entirely convincing. Perhaps more informed and integrated methods, that would take into account nucleotide frequencies and genotype phasing, for example, could provide better outcomes.

Extension from SNP pairs to SNP groups

Finally, larger sets of SNPs should be included. As we briefly mentioned in the introduction, the regulatory mechanisms of human genes are not always understood, and are often extremely complex. As the example of the human beta-globin locus illustrates, they can involve multiple players. If we are to capture regulatory epistatic interactions between SNPs on these elements, we should consider all of them at a time, not just two. Thus, we must go beyond pairs of SNPs, toward larger SNP groups. As it happens, one of the major advantages of logistic regressions compared to other association measures is that they are perfectly well suited to include multiple predictors. Nevertheless, Hi-C is only able to measure contacts between two regions at a time. We can assume, if one region *A* forms a loop with region *B* as found across some cells, and a loop with region *C* in other cells, that *A*, *B* and *C* are actually all in the same place and it is the technology that can only see two at a time. On the other hand, the fact that there is no obvious loop between *B* and *C* might suggest that on the contrary, *A* either interacts with one or the other, in a mutually exclusive fashion. Unfortunately, we are not able to answer these questions with the existing experimental techniques. This makes the task of extending the search for epistatic interactions from pairs to larger groups a non-trivial one. An interesting way to visualize and interpret Hi-C results was proposed by Sanborn *et al*: the network of loops. One could imagine genomic regions as nodes, connected to each other by an edge, if engaging in a chromatin loop. Graphs come with nice mathematical properties and one could for example find all clusters of looping regions as connected components within the graph. Clusters of variants engaging in epistatic interactions might be found by assigning one or more variants to a node. Then starting from a node, it could be possible to progressively add other nodes as long as the model including the new variants keeps improving association, as measured with an LRT (a representation of this proposed approach can be found in Fig S11).

Detailed Methods

Enhancers' positions

A list of enhancers is collected from the Epigenome Roadmap (Roadmap *et al*, 2015). The data is derived from the same cell as the Hi-C loops (GM17828, ENCODE ID: E116). In Roadmap *et al*, genomic bins of 200 bp are assigned to one of 15 states, based on an HMM (Hidden Markov Model) that takes into account various features including histone modification markers, conservation rates and chromatin state, among others. We select those bins that are confidently assigned to state 7 (state: Enhancer), and see how many of those are contained in chromatin loops.

Common SNPs' positions

Common variants are downloaded from the UCSC repository (University of California, Santa Cruz, http://hgdownload.soe.ucsc.edu). A list of loops is collected from Sanborn *et al* (accession number = GSE74072 on http://www.ncbi.nlm.nih.gov). The portion of SNPs in loops is calculated as the ratio between the number of SNPs in loops, and the total number of SNPs, per chromosome. The loop coverage (base pairs in loops) is calculated from the Hi-C map. It is the ratio between the overall size of chromosome bins forming a loop, over the sum of the span of all chromosome bins. If for some portions of the chromosome we have no Hi-C information, then we do not include those.

Binomial test

In 3.1.1 and 3.1.2, the significance of the enrichment is calculated using a Binomial test:

$$H_0: \; x \sim Bi(n,p)$$
$$H_1: \; x \nsim Bi(n,p)$$

Where x is the number of enhancers/SNPs in looping regions, n is the total number of enhancers/SNPs, and p is the ratio of base pairs in loops out of the total number of base pairs we have Hi-C data from.

Artificial Pairs: SNPs in loops

The first set of artificial pairs used in 3.2.1 and 3.2.2 are built as follows. Only SNPs that are on a region that engages in a loop are considered, but the pairs are shuffled around. If we have two loops, bringing together two independent genomic regions each: the first loop is composed of regions *A* and *B*, the second loop of regions *C* and *D*. For simplicity, each region only hosts one SNP. We call the SNPs *a*, *b*, *c* and *d*. The true pairs in this scenario are *ab*, *cd*. The artificial pairs we build, instead, are *ac*, *ad*, *bc*, *bd*. In 3.2.1 we want to keep the number of pairs fixed. Thus, we would select only two out of the four (say *ad* and *bc*). This approach is depicted in Fig 2B.

Artificial Pairs: similar linear distance

For the second strategy to build artificial pairs, we start from a loop. Every loop is made up of two regions, that we can call A and B. A comes 'before' B on the chromosome, if we consider it as a straight line. A and B are at a certain linear distance *d*, measured in bp, midpoint to midpoint. We consider two other regions, at the two opposite sides: B', at distance *d* from region A, but on its left hand side (upstream), and A', at distance *d* from B, to its right (downstream). As we expect A and B to be rich in SNPs, we take A' and B' twice as large (as A, B), symmetrically from the midpoint, and only later shrink them until they contain the same number of SNPs as their counterpart (Fig 2C):

$$\#\{SNPs\ in\ A\} = \#\{SNPs\ in\ A'\}$$

$$\#\{SNPs\ in\ B\} = \#\{SNPs\ in\ B'\}$$

T2D-associated SNPs

The list of SNPs known to be associated with type 2 diabetes are collected from the NHGRI (National Human Genome Research Institute) GWAS Catalog (Welter, MacArthur *et al*.2014). Diabetes mellitus type II (or simply type 2 diabetes) is a metabolic disorder characterized by high blood sugar and insulin resistance. Type 2 diabetes is only partly genetic. Similar to most common complex diseases, many factors play a role. For type 2 diabetes, the main cause is obesity. As for genetic causes, most known common SNPs related to T2D sit on genes involved in beta cells function (insulin storage and release).

References

Ayati, M. and M. Koyutürk (2014). Prioritization of genomic locus pairs for testing epistasis. *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, ACM.

Babaei, S., *et al*. (2015). "3D hotspots of recurrent retroviral insertions reveal long-range interactions with cancer genes." *Nature communications* 6.

Bai, J. (1999). "Likelihood ratio tests for multiple structural changes." *Journal of Econometrics* 91(2): 299-323.

Bannister, A. J. and T. Kouzarides (2011). "Regulation of chromatin by histone modifications." *Cell research* 21(3): 381-395.

Basu, S. and W. Pan (2011). "Comparison of statistical tests for disease association with rare variants." *Genetic epidemiology* 35(7): 606-619.

Botta, M., *et al*. (2010). "Intra-and inter-chromosomal interactions correlate with CTCF binding genome wide." *Molecular systems biology* 6(1): 426.

Bouwman, B. A. and W. de Laat (2015). "Getting the genome in shape: the formation of loops, domains and compartments." Genome biology 16(1): 1-9.

Brinza, D. and A. Zelikovsky (2006). Combinatorial methods for disease association search and susceptibility prediction. *International Workshop on Algorithms in Bioinformatics, Springer*.

Burton, P. R., *et al*. (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." *Nature* 447(7145): 661-678.

Bush, W. S., *et al*. (2009). Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pacific Symposium on Biocomputing*. *Pacific Symposium on Biocomputing*, NIH Public Access.

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

Cantor, R. M., *et al*. (2010). "Prioritizing GWAS results: a review of statistical methods and recommendations for their application." The American Journal of Human Genetics 86(1): 6-22.

Chahal, H. S., *et al*. (2016). "Genome-wide association study identifies 14 novel risk alleles associated with basal cell carcinoma." Nature communications 7.

Chen, Z.-J., *et al*. (2011). "Genome-wide association study identifies susceptibility loci for polycystic ovary syndrome on chromosome 2p16. 3, 2p21 and 9q33. 3." Nature genetics 43(1): 55-59.

Cortes, A., *et al*. (2015). "Major histocompatibility complex associations of ankylosing spondylitis are complex and involve further epistasis with ERAP1." Nature communications 6.

de Laat, W. and F. Grosveld (2003). "Spatial organization of gene expression: the active chromatin hub." Chromosome Research 11(5): 447-459.

de Wit, E. and W. de Laat (2012). "A decade of 3C technologies: insights into nuclear organization." Genes & development 26(1): 11-24.

de Wit, E., *et al*. (2013). "The pluripotent genome in three dimensions is shaped around pluripotency factors." Nature 501(7466): 227-231.

Dekker, J., *et al*. (2002). "Capturing Chromosome Conformation." Science 295(5558): 1306-1311.

Dinu, I., *et al*. (2012). "SNP-SNP interactions discovered by logic regression explain Crohn's disease genetics." PLoS One 7(10): e43035.

Dixon, J. R., *et al*. (2012). "Topological domains in mammalian genomes identified by analysis of chromatin interactions." Nature 485(7398): 376-380.

Dostie, J., *et al*. (2006). "Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements." Genome research 16(10): 1299-1309.

Dryden, N. H., *et al*. (2014). "Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C." Genome research 24(11): 1854-1868.

Evans, D. M. and S. Purcell (2012). "Power calculations in genetic studies." Cold Spring Harbor Protocols 2012(6): pdb. top069559.

Fuchsberger, C., *et al.* (2015). "minimac2: faster genotype imputation." *Bioinformatics* **31**(5): 782-784.

Fullwood, M. J., *et al*. (2009). "An oestrogen-receptor-&agr;-bound human chromatin interactome." Nature 462(7269): 58-64.

Ghavi-Helm, Y., *et al*. (2014). "Enhancer loops appear stable during development and are associated with paused polymerase." Nature 512(7512): 96-100.

Gibbs, R. A., *et al*. (2003). "The international HapMap project." Nature 426(6968): 789-796.

Goudey, B., *et al*. (2015). "High performance computing enabling exhaustive analysis of higher order single nucleotide polymorphism interaction in Genome Wide Association Studies." Health Information Science and Systems 3(1): 1.

Grange, L., *et al.* (2015). "Filter-free exhaustive odds ratio-based genome-wide interaction approach pinpoints evidence for interaction in the HLA region in psoriasis." *BMC genetics* 16(1): 1.

Grubert, F., *et al*. (2015). "Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions." Cell 162(5): 1051-1065.

Hong, E. P. and J. W. Park (2012). "Sample size and statistical power calculation in genetic association studies." *Genomics & informatics* 10(2): 117-122.

Howie, B., *et al.* (2012). "Fast and accurate genotype imputation in genome-wide association studies through pre-phasing." *Nature genetics* **44**(8): 955-959

Hu, X., *et al*. (2010). "SHEsisEpi, a GPU-enhanced genome-wide SNP-SNP interaction scanning algorithm, efficiently reveals the risk genetic epistasis in bipolar disorder." Cell research 20(7): 854-857.

Jamshidi, M., *et al.* (2015). "SNP-SNP interaction analysis of NF-κB signaling pathway on breast cancer survival." Oncotarget 6(35): 37979.

Jin, F., *et al*. (2013). "A high-resolution map of the three-dimensional chromatin interactome in human cells." Nature 503(7475): 290-294.

Kircher, M., *et al.* (2014). "A general framework for estimating the relative pathogenicity of human genetic variants." Nature genetics 46(3): 310.

Kuhn, R. M., *et al*. (2012). "The UCSC genome browser and associated tools." Briefings in bioinformatics: bbs038.

Li, J., *et al*. (2011). "Detecting epistatic effects in association studies at a genomic level based on an ensemble approach." Bioinformatics 27(13): i222-i229.

Li, H., *et al.* (2016). "Integrative genomics analyses unveil downstream biological effectors of disease-specific polymorphisms buried in intergenic regions." NPJ Genomic Medicine 1: 16006.

Lieberman-Aiden, E., *et al*. (2009). "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." Science 326(5950): 289-293.

Lippert, C., *et al*. (2013). "An exhaustive epistatic SNP association analysis on expanded Wellcome Trust data." Scientific reports 3: 1099.

Loh, P.-R., *et al.* (2016). "Fast and accurate long-range phasing in a UK Biobank cohort." *Nature genetics.*

Maher, B. (2008). "The case of the missing heritability." Nature 456(7218): 18-21.

Marioni, J. C., *et al.* (2008). "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays." Genome research 18(9): 1509-1517.

Martin, P., *et al*. (2015). "Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci." Nature communications 6.

Matharu, N. and N. Ahituv (2015). "Minor Loops in Major Folds: Enhancer–Promoter Looping, Chromatin Restructuring, and Their Association with Transcriptional Regulation and Disease." PLoS Genet 11(12): e1005640.

Maurano, M. T., *et al*. (2012). "Systematic localization of common disease-associated variation in regulatory DNA." Science 337(6099): 1190-1195.

McCarthy, M. I., *et al*. (2008). "Genome-wide association studies for complex traits: consensus, uncertainty and challenges." Nature reviews genetics 9(5): 356-369.

Meuleman, W., *et al*. (2013). "Constitutive nuclear lamina–genome interactions are highly conserved and associated with A/T-rich sequence." Genome research 23(2): 270-280.

Moreira, M. J. (2003). "A conditional likelihood ratio test for structural models." *Econometrica* 71(4): 1027-1048.

Neyman, J. and E. S. Pearson (1933). The testing of statistical hypotheses in relation to probabilities a priori. *Mathematical Proceedings of the Cambridge Philosophical Society, Cambridge Univ Press.*

Petersen, A., *et al*. (2013). "Assessing methods for assigning SNPs to genes in gene-based tests of association using common variants." *PLoS One* 8(5): e62161.

Piriyapongsa, J., *et al.* (2012). "iLOCi: a SNP interaction prioritization technique for detecting epistasis in genome-wide association studies." *BMC genomics* 13(7): 1.

Pombo, A. and N. Dillon (2015). "Three-dimensional genome architecture: players and mechanisms." *Nature Reviews Molecular Cell Biology.*

Pope, B. D., *et al*. (2014). "Topologically associating domains are stable units of replication-timing regulation." *Nature* 515(7527): 402-405.

Price, A. L., *et al*. (2006). "Principal components analysis corrects for stratification in genome-wide association studies." Nature genetics 38(8): 904-909.

Pulit, S. (2016). "Deciphering the four-letter code: The genetic basis of complex traits and common disease."

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.

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.

Roadmap, E., *et al*. (2015). "Integrative analysis of 111 reference human epigenomes." *Nature* 518: 317-330.

Ritchie, M. D. (2011). "Using biological knowledge to uncover the mystery in the search for epistasis in genome-wide association studies." *Annals of human genetics* 75(1): 172-182.

Rogers, A. R. and C. Huff (2009). "Linkage disequilibrium between loci with unknown phase." Genetics 182(3): 839-844.

Sanborn, A. L., *et al*. (2015). "Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes." Proceedings of the National Academy of Sciences 112(47): E6456-E6465.

Sanyal, A., *et al*. (2012). "The long-range interaction landscape of gene promoters." *Nature* 489(7414): 109-113.

Schierding, W. S., *et al*. (2014). "The missing story behind Genome Wide Association Studies: single nucleotide polymorphisms in gene deserts have a story to tell." *Frontiers in genetics* 5: 39.

Shete, S., *et al.* (2009). "Genome-wide association study identifies five susceptibility loci for glioma." *Nature genetics* 41(8): 899-904.

Simonis, M., *et al*. (2009). "High-resolution identification of balanced and complex chromosomal rearrangements by 4C technology." *Nature methods* 6(11): 837-842.

So, H. C., *et al*. (2011). "Uncovering the total heritability explained by all true susceptibility variants in a genome-wide association study." *Genetic epidemiology* 35(6): 447-456.

Stringer, S., *et al*. (2011). "Underestimated effect sizes in GWAS: fundamental limitations of single SNP analysis for dichotomous phenotypes." *PLoS One* 6(11): e27964.

Tolhuis, B., *et al*. (2002). "Looping and interaction between hypersensitive sites in the active β-globin locus." Molecular cell 10(6): 1453-1465.

Turner, S. D., *et al.* (2011). "Knowledge-driven multi-locus analysis reveals gene-gene interactions influencing HDL cholesterol level in two independent EMR-linked biobanks." *PLoS One* 6(5): e19586.

Visel, A., *et al*. (2009). "Genomic views of distant-acting enhancers." *Nature* 461(7261): 199-205.

Wan, X., *et al*. (2010). "BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies." *The American Journal of Human Genetics* 87(3): 325-340.

Wan, X., *et al*. (2009). "MegaSNPHunter: a learning approach to detect disease predisposition SNPs and high level interactions in genome wide association study." *BMC bioinformatics* 10(1): 1.

Wan, X., *et al*. (2010). "Predictive rule inference for epistatic interaction detection in genome-wide association studies." Bioinformatics 26(1): 30-37.

Wason, J. M. and F. Dudbridge (2010). "Comparison of multimarker logistic regression models, with application to a genomewide scan of schizophrenia." *BMC genetics* 11(1): 1.

Ward, L. D. and M. Kellis (2012). "Interpreting noncoding genetic variation in complex traits and human disease." Nature biotechnology 30(11): 1095-1106.

Wei, W.-H., *et al.* (2014). "Detecting epistasis in human complex traits." *Nature reviews genetics* 15(11): 722-733.

Welter, D., *et al.* (2014). "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations." *Nucleic acids research* **42(**D1): D1001-D1006.

Wilks, S. S. (1938). "The large-sample distribution of the likelihood ratio for testing composite hypotheses." The Annals of Mathematical Statistics 9(1): 60-62.

Xu, Z., *et al*. (2016). "HiView: an integrative genome browser to leverage Hi-C results for the interpretation of GWAS variants." *BMC research notes* 9(1): 1.

Yang, Z. (1998). "Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution." *Molecular biology and evolution* 15(5): 568-573.

Yung, L. S., *et al*. (2011). "GBOOST: a GPU-based tool for detecting gene–gene interactions in genome–wide case control studies." *Bioinformatics* 27(9): 1309-1310.

Zhao, Z., *et al*. (2006). "Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions." *Nat Genet* 38(11): 1341-1347.