Identification of Functional cis-regulatory Polymorphisms in the Human Genome

Abstract

Polymorphisms in regulatory DNA regions are believed to play an important role in determining phenotype, including disease, and in providing raw material for evolution.

We devised a new pipeline for the systematic identification of functional variation in human regulatory sequences. The algorithm is based on the identification of SNPs leading to significant changes in both the affinity of a regulatory region for transcription factors and the expression *in vivo* of the regulated gene. We tested the algorithm by identifying SNPs leading to altered regulation by STAT3 in human promoters and introns, and experimentally validated the top-scoring ones, showing that most of the SNPs identified by the algorithm indeed correspond to differential binding of STAT3 and differential induction of the target gene upon stimulation with IL6. Using the same computational approach we compiled a database of thousands of predicted functional regulatory SNPs for hundreds of human transcription factors which we provide as a public resource. We discuss possible applications to the interpretation of non-coding SNPs associated to human diseases.

The method we propose and the database of predicted functional cis-regulatory polymorphisms will be useful in future studies of regulatory variation and in particular to interpret the results of past and future genome-wide association studies.

Background

Genetic variation is responsible for phenotypic differences, including susceptibility to disease. Genome-wide association studies (GWAS) can identify statistical correlations between polymorphisms and phenotypes. Once such an association has been determined, one is usually faced with the daunting problem of understanding the functional mechanism by which the polymorphism influences the phenotype. This task is particularly difficult when the polymorphism might not affect any gene product, as in the case of polymorphisms located in intergenic regions or introns.

It is thus reasonable to assume that some polymorphisms exert their effect by affecting the expression level of a neighboring gene, or possibly several genes (see e.g. Lappalainen and Dermitzakis (2010) for a review of regulatory variation in human populations). This can happen if the polymorphism modifies regulatory information such as the binding affinity of a trans-acting regulatory element. Indeed several cases have been documented in which the simplest kind of polymorphism, a single nucleotide polymorphism (SNP), can alter the ability of a transcription factor (TF) to bind the regulatory region and thus correctly regulate the expression of a gene (see e.g. Epstein (2009) for a review).

In this work we present a method for the systematic identification of functional regulatory SNPs based on the integration of SNPs and gene expression data. We define as functional a SNP able to influence the expression level of a neighboring gene by altering the binding *in vivo* of a TF. This should be contrasted with other

approaches in which the functionality of a SNP is defined, and experimentally tested, based solely on its ability to affect the binding of a TF, irrespective to any effect on gene expression, or, *vice versa*, on its effect on gene expression irrespective of the binding of any TF.

The method is based on the availability of genetic variation information and gene expression profiling of the same individuals, and identifies SNPs in regulatory regions that have, on one hand, the potential to alter the affinity of a TF and, on the other, a significant correlation with gene expression.

We then tested the top-scoring SNPs for the STAT3 TF by stimulating human cells with IL6 and determining whether different alleles corresponded to both differential gene expression and differential binding of STAT3. STAT3 is a member of the Signal transducers and activators of transcription (STAT) family of TFs, and is known to be involved in a wide variety of physiological and pathological processes including inflammation, regeneration, proliferation, energy homeostasis and many forms of cancer Poli and Alonzi (2003). This validation was done in two ways: by comparing individuals homozygous for different alleles, and different alleles within heterozygous individuals.

Since the rate of experimental confirmation of our predictions for STAT3 turned out to be satisfactory, we extended the analysis to a large collection of human TFs, thus generating a publicly available database of thousands of predicted functional regulatory SNPs, each associated to one or more TFs and target genes. We believe this resource will be useful in interpreting the results of past and future GWAS. Previous studies such as Andersen et al. (2008); Lapidot et al. (2008); Ameur et al. (2009); Wang et al. (2011) have employed some of the same techniques, but we believe that the integration of TF binding site prediction and gene expression analysis with functional validation *in vivo* makes our approach especially reliable.

Results and Discussion

Genome-wide prediction of functional regulatory SNPs

We considered a SNP to significantly change the affinity of a binding site for STAT3 if (a) only one of the two alleles corresponds to a score above the score cutoff c=9.6, previously determined in Vallania et al. (2009) by optimizing the predictivity on known binding sites; and (b) the difference in score between the two alleles is at least 0.5. A total of 4,166 SNPs in the human genome met these criteria. We retained for further analysis the ones located within 10kbp of an annotated transcription start site (TSS) and outside of annotated exons (250 SNPs), since we expect these regions to be enriched in regulatory elements.

For each of the SNPs selected above we asked whether the expression of nearby transcripts in lymphoblastoid cell lines Stranger et al. (2007) showed significant correlation with the number of copies of the high-affinity allele. For each SNP we analyzed the expression of the gene(s) with a TSS within 10kbp, and we identified 12 SNPs significantly correlated with gene expression at 0.05 significance level (Pearson correlation test with Bonferroni correction for multiple testing).

Experimental validation

The computational analysis described above produced a list of SNPs associated to a significant change in the affinity for STAT3 and correlated to change in expression of the neighboring gene. However we do not expect in all cases that the differential binding of STAT3 to the locus involved will be the causal factor leading to differences in gene expression. Indeed, we cannot exclude the possibility that the effect is due, for example, to other SNPs in linkage disequilibrium with the one analyzed, possibly altering the binding of other TFs or other cis-regulatory information.

Therefore we performed a systematic experimental validation of the top-ranking SNPs identified by our pipeline, on the same cell lines collected by the Hapmap project, stimulated with IL6, a well-established activator of STAT3. For each SNP we tested both the differential binding of STAT3 to the two alleles and the differential regulation of the target gene. We tested differential binding and inducibility of the target gene in individuals homozygous for the two variants. Then we repeated the test on heterozygous individuals, testing the differential binding and inducibility of the two alleles.

Differential induction

We took the inducibility by the IL6 stimulus as a measure of functionality of the STAT3 binding sites in regulating gene expression. With the strategy described in the Materials and Methods section we chose a team of individuals containing at least 2 homozygotes for each allele of 11 candidate SNPs selected above (one SNP had to be excluded since only one individual homozygous for one of the alleles is represented in Hapmap). For each target gene and each individual, we evaluated the fold change in expression (induction) with respect to the unstimulated sample at 1.5, 6 and 24 hours after stimulation. We then compared the induction levels of individuals with high-affinity and low-affinity alleles.

Comparing the induction levels of the high- and low-affinity alleles for the 11 genes tested we found, as expected, that high-affinity alleles corresponded to higher induction. A Wilcoxon signed-rank test revealed a significant overall difference in induction between high- and low-affinity alleles at 6 and 24 hours after induction (respectively P=0.0068 and P=0.0098), while at 1.5 hours the result was not significant (P=0.067). We concluded that, overall, the SNPs identified by our method are indeed functional as they give rise to significantly different levels of induction by IL6.

To analyze individual SNPs, we compared the induction of each homozygote carrying one allele to each homozygote with the other, using an unpaired two–tailed *t*–test on the three replicate expression values obtained by qRT-PCR. The results of these 66 comparisons at 6 hours after treatment with IL6 are shown in Figure 1. In 32 cases the difference in induction was statistically significant, corresponding to a ~10% false discovery rate. In all of these 32 cases the induction was higher for the individual carrying the high-affinity variant. PCR expression data for individual genes are shown in Supplementary Figures 1 and 2.

For the 7 intronic SNPs we also tested the differential induction of the two alleles after induction with IL6, in heterozygous individuals. The results are in general agreement with what found comparing homozygous individuals: for example, at 6 hours after stimulation, in 4 out of 7 cases the induction was significantly higher for

the high-affinity allele (Fig. 2). Results at all timepoints are shown in Supplementary Figure 2.

Differential binding

To confirm that the differential induction of the genes considered above is indeed due to differential binding of STAT3 to the site including the SNP we performed chromatin immunoprecipitation experiments comparing the binding of STAT3 between homozygous individuals carrying different alleles and, for intronic SNPs, between the two alleles of heterozygous individuals, after induction with IL6.

For each of the 11 SNPs tested we performed a ChIP experiment on one homozygous individual carrying the high-affinity allele and one carrying the low-affinity one. The experiment was performed after 1.5 hours or 6 hours after IL6 stimulation, depending on the location of the induction peak for the high-affinity individual.

Evidence for STAT3 binding (defined as STAT3 ChIP signal significantly higher than the IgG negative control signal) was found for 6 of the high-affinity alleles (*ARHGEF1*, *DHX36*, *GTF3C6*, *LCP1*, *NCF4*, *SKA1*) and only one of the low-affinity ones (*NCF4*). For these 6 genes STAT3 binding (defined as STAT3 ChIP signal after subtraction of the IgG signal) was significantly higher for the high-affinity allele than for the low-affinity one.

Remarkably, the genes showing differential binding coincide almost exactly with the ones shown to undergo differential induction (see Tab. Errore. L'origine riferimento non è stata trovata.). These results strongly suggest that the difference in induction that we observed is indeed due to differential STAT3 binding. The results of individual ChIP experiments are shown in Supplementary Figures 1 and 2.

For the 7 intronic SNPs we performed allele-specific ChIP experiments as described in the Methods, finding similar results: in all of them we obtained a significant binding signal (difference between STAT3 and IgG ChIP signal) for the high-affinity allele, and only in two cases for the low-affinity allele. In 6 out of 7 cases the binding of the high-affinity allele was significantly higher than the low-affinity one. Moreover, in all the 4 SNPs in which differential induction of the two alleles was observed also a significant difference in binding was detected (Table Errore. L'origine riferimento non è stata trovata.).

Figure 3 shows, as an example, all the experimental results for the intronic SNPs rs9400435 in gene *GTF3C6*. Results for all the genes tested are shown in Supplementary Figure 2.

A database of putative functional SNPs

Since the experimental validation carried out for STAT3 showed that most of the top-predicted SNPs are indeed functional, we derived similar predictions for all the PWMs included in the JASPAR "Core Vertebrate" collection. For each PWM we used the same definition of regulatory regions used above for STAT3. The score cutoff used for each PWM is a quadratic function of the maximum possible score of the PWM. Such function was determined by an optimization procedure based on publicly available ChIP/seq data and described in the Methods.

We thus obtained 6,682 SNPs located in the regulatory regions and such that (a) only one of the two alleles corresponds to a score above the cutoff; (b) the difference in score between the two alleles is at least 0.5 for one or more PWMs; and (c)

correlation with expression is significant (nominal P-value <0.05). The complete list can be found in Supplementary Table 5.

Association with disease

Genetic variation in regulatory regions is thought to be relevant to genetic predisposition to diseases Epstein (2009). Therefore we asked whether any of the SNPs identified with our approach were significantly associated with the diseases examined in a large scale GWA studyConsortium (2007). About 2,800 of the 6,682 SNPs identified above were included in this study and passed quality controls for each of the seven complex diseases considered. We therefore used a trend test Armitage (1955) on these SNPs for the 7 diseases included in the study to test for significant associations. With a Benjamini-Hochberg FDR of 10%, we found 50 significant associations for 42 unique SNPs and 6 of the 7 diseases (no significant association was found for type-2 diabetes). The complete list is shown as Supplementary Table 6.

Of course association does not imply causation, and we cannot exclude linkage disequilibrium with other SNPs as an alternative explanation for the associations found. However, at least in some cases, both the TF whose binding site is altered by the SNP and the target gene involved were previously associated to the disease or to relevant biological processes, naturally leading to functional hypotheses.

For example we found a significant association between SNP rs9393708 and type 1 diabetes (P=1.54·10\s\up5(–8), trend test). The SNP is located ~2700 bases upstream of *BTN3A2*, a gene whose association with type 1 diabetes was previously established Viken et al. (2009). Expression of *BTN3A2* is very strongly correlated with the allele (correlation P=8.0·10\s\up5(–32)) in the Hapmap gene expression data. This SNP is on the other hand predicted to alter the binding of PDX1, known to be involved in the early development of the pancreas and in the regulation of insulin gene expression Kaneto et al. (2008). Therefore a possible mechanistic explanation of the association of this SNP with type 1 diabetes is that the SNP influences the regulation of *BTN3A2* by PDX1.

Similarly, rs12610384 is significantly associated with bipolar disorder (P=1.91·10\s\up5(–4)), and is located ~ 100 bp upstream of *POLRMT*, a mitochondrial DNA-directed RNA polymerase involved in the expression of mitochondrial genes. The SNP alters a putative GATA2 binding site: given the known role of mitochondrial dysfunction in bipolar and other psychiatric disorders (see e.g. Ref.Clay et al. (2011)), and of GATA2 in the transcriptional network responsible for the differentiation of serotonergic neurons Alenina et al. (2006), also in this case a possible mechanistic explanation of the association between SNP and disorder emerges.

Conclusions

We have shown that the integration of genetic variation and gene expression data can be exploited to predict functional regulatory SNPs with high specificity. The use of correlation with gene expression to select our candidate SNPs helps ensure that they are located in regions of open chromatin, accessible *in vivo* to TFs.

This procedure, however, makes our predictions context-dependent, since, strictly speaking, they apply only to the cellular context in which the gene expression data were obtained. Indeed our experimental validation was carried out in the same cells.

However the concept itself of functional regulatory SNP as we defined is intrinsically context-dependent, and it is in principle impossible to develop a context-independent prediction unless one is willing to settle for a less biologically relevant definition of "functional" (e.g. by requiring only differential binding *in vitro*).

On the other hand, we do expect our predicted SNPs to be functional also in other contexts, since chromatin states in proximal regulatory regions are much more conserved across cell types than those in distal enhancers, as shown for example in a recent genome-wide mapping of chromatin state in human cellsErnst et al. (2011). Therefore we believe our database can be profitably used independently of the cellular context we used to derive it.

An important limit of the present work is that we only analyzed single nucleotide polymorphisms, when it is natural to expect that an important portion of regulatory variation is due to insertion, deletions and copy number variation (CNV). For example CNV was shown Stranger et al. (2007) to explain ~20% of genetic variation in gene expression in individuals analyzed within the HapMap project. Massive resequencing projects such as the 1000 Genomes Project 1000 Genomes Project Consortium (2010) will enable genome-wide investigations of altered TF binding patterns due to this type of polymorphisms.

Methods

Identification of affinity-changing SNPs

We considered all SNPs in the Affymetrix GenomeWideSNP 6 platform (annotation version na30). Each SNP has two alleles: A and B. We reconstructed two genomic sequences for each SNP, S\s\do5(A) and S\s\do5(B), that differ only in the SNP position, while the sequences flanking the SNP nucleotide were reconstructed from the reference genome (UCSC hg18) based on chromosomal coordinates of the SNPs as reported in the Affymetrix annotation file.

For STAT3 we used the position weight matrix (PWM) determined in Vallania et al. (2009). For other TFs we considered the 130 PWMs included in the JASPAR Core Vertebrate collection (2009 release). For each PWM we computed the log-likelihood ratios (LLR) of the PWM for the two sequences: L(S\s\do5(A)) and L(S\s\do5(B)). The LLR is the log\s\do5(2) of the ratio of the probability of obtaining the sequence from the PWM and the probability of obtaining it from the background model. The latter is given by the nucleotide frequencies in the whole intergenic part of the human genome.

A SNP is called *affinity-changing* for a given PWM if one of the two LLRs is greater than a cutoff c, the other one is not and $|L(S\text{\s\do5}(A))-L(S\text{\s\do5}(B))|>0.5$. The threshold score c was chosen equal to 9.6. This value was obtained in Vallania et al. (2009) by requiring the highest statistical significance of the enrichment of positive hits in a set of confirmed STAT3 promoters Vallania et al. (2009). For other PWMs in the Jaspar Core database we determined the cutoff c with an optimization procedure described below.

Correlation with expression

We downloaded the normalized gene expression data of 210 unrelated HapMap individuals Stranger et al. (2007) from GEO (series identifier: GSE6536). For each

gene g and each affinity-changing SNP s located within ±10Kbp from the transcription start site of g, we computed the correlation r\s\do5(gs) as follows: Given an affinity-changing SNP s, let s\s\do5(⊕) be the allele with greater affinity and s\s\do5(⊖) the other allele, so that $L(S$\s\do5(⊕))>$L(S$\s\do5(⊖)). r\s\do5(gs) is the Pearson's product moment correlation of the number of s\s\do5(⊕) alleles in an individual (0, 1 or 2) and the expression level of g. Positive (negative) correlation suggest that s changes the affinity of the sequence for an activator (repressor) of g. The significance of the correlation was assessed with a standard correlation test.

Experimental validation: choice of individuals

Given a set of affinity-changing SNPs, to check if they are functional we needed to choose a team of individuals with different genotypes for the given SNPs. The ideal team is the one that allows the functional characterization of the greatest number of SNPs using the smallest possible set of individuals, to minimize experimental costs.

We selected 12 SNPs that change the affinity for STAT3 and significantly correlated with gene expression (Bonferroni corrected P<0.05). We then selected 88 individuals for which cell lines were already available in our lab from the 210 present in the gene expression dataset used.

We define "satisfactory" a team of individuals containing at least n individuals for each SNP and each genotype (we used n=2 in the present study). SNPs with less than n individuals per each genotype in the set of 88 individuals are discarded a priori.

The number of selected SNP and individuals in our case allow modern workstation hardware to solve the combinatorial problem of finding the smallest satisfactory team trough exhaustive search for small team size. So we adopted an iterative approach:

1. we obtained the best team for size 1, 2, 3, 4 and 5. If at a certain size a satisfactory team is obtained the procedure ends and the individuals in the current team are marked,
2. if there is no satisfactory team of size ≤5 we mark the individuals found in the best team of size 5 and discard the SNP that have at least n individuals for genotype in the team of marked individuals,
3. repeat from 1 using the new (smaller) set of SNP until the goal is reached.

The marked individuals at the end of the procedure are those selected for the experimental validation.

Cell lines and treatments

Human lymphoblastoid cell lines were obtained from the Coriell Cell Repository (Coriell Institute, New Jersey, USA) and grown in RPMI 1640 (Gibco-BRL) supplemented with 15% (v/v) heat-inactivated FCS (foetal calf serum; Gibco-BRL) at 37\s\up5(o)C under 5% carbon dioxide. All genotypes were confirmed by direct sequencing after PCR amplification of the region containing the SNP of interest. Primer sequences used for PCR amplification are reported in Supplementary Table 1. Forward primers were used as sequencing primers. Cells were treated with recombinant IL6 (500 ng/ml) plus soluble receptor (250 ng/ml) for 1.5, 3 or 6 hours.

Total RNA extraction, retro-transcription and qRT-PCR (quantitative real-time PCR)

Total RNA extraction and retro-transcription were performed as previously described Vallania et al. (2009). qRT-PCRs were performed using the Universal Probe Library system (UPL; Roche). Primer sequences and probe numbers are reported in Supplementary Table 2. Results were analyzed with the 2\s\up5(−ΔΔ*C*\s\do4(*t*)) method using the 18S rRNA pre-developed TaqMan assay (Applied Biosystems) as internal control.

Nuclear RNA extraction

10\s\up5(7) cells were precipitated by centrifugation (5 minutes at 1200 rpm) and resuspended in 200 µl of LB Buffer (10 mM NaCl, 2 mM MgCl\s\do5(2), 10 mM Tris-HCl pH 7.8, 5 mM DTT, 0.5% Igepal CA630). After 10 minutes of incubation on ice, nuclei were collected by centrifugation (5 minutes at 8000 rpm at 4°C), washed twice with cold PBS and resuspended in 200µl of LB Buffer. Nuclei were then treated with 200µl of 2x ProtK Buffer (0.2 M Tris-HCl pH 7.5, 25 mM EDTA, 0.3 M NaCl, 2% SDS), 20µl of Proteinase K (Sigma-Aldrich, 10 mg/ml) and incubated at 37\s\up5(o)C for 20 minutes. After incubation, nuclear RNA was purified with a classic Trizol (Invitrogen) – Chloroform (Merck Chemicals) protocol. Samples were treated with DNase I Amp Grade (Invitrogen) in order to remove genomic DNA contaminations. RNA integrity was checked on 1% denaturing agarose gel.

Allele-specific gene expression

A Real-Time Amplification Refractory Mutation System quantitative PCR (ARMS-qPCR) approach was used to evaluate allele-specific gene expression. A common reverse primer was used to amplify both alleles, while were used two different 3\s\up5(')-mismatched forward primers to discriminate the two different SNP-containing alleles Newton et al. (1989). To increase the specificity of the ARMS reactions, two additional mismatches were introduced at the two nucleotides immediately 5\s\up5(') to the SNP as previously described by Bai and co-workers Bai and Wong (2004). The specificity of ARMS-qPCR reactions was tested using homozygous individuals as a control and only 100% allele-specific primers were chosen for subsequent analysis. Primers used in ARMS-qPCR reactions are reported in Supplementary Table 3, mismatched nucleotides are underlined. The RT ARMS-qPCRs, were performed with the ABI Prism 7300 real-time PCR system (Applied Biosystems) using Platinum SYBR Green qPCR SuperMix-UDG (Invitrogen) and 500 nM of each primer in a total volume of 20µl. Real-time conditions were 2 min at 50\s\up5(o)C and 10 min at 95\s\up5(o)C, followed by 45 cycles of denaturation for 15 sec at 95\s\up5(o)C and annealing/extension for 1 min at 60\s\up5(o)C. Dissociation curves for the amplicons were generated after each run to confirm the specificity of the signal. Calibration curves were included in each run. The allelic expression ratio was calculated with the following formula:

$$\F(A,B)=2\s\up5(C\s\do4(t)(B)-C\s\do4(t)(A))$$

Experiments were performed in triplicate with at least two independent samples.

Conventional and allele-specific chromatin immunoprecipitation (ChIP and HaploChIP)

Lymphoblastoid cell were treated or not with IL6 plus soluble receptor as described above. ChIP assays were performed as previously described Vallania et al. (2009). Immunoprecipitations were performed by incubating overnight at 4\s\up5(o)C 1 ml of sheared chromatin with anti-Stat3 antibodies (Cell Signaling Technology; 1:50), or with negative control IgG (Sigma-Aldrich, 2μg). Primer sequences used in qRT-PCRs for conventional ChIPs are reported in Supplementary Table 4. Allele-specific qRT-PCRs were performed using the ARMS approach described above. Primer sequences are reported in Supplementary Table 3.

Statistical significance of differential induction, STAT3 binding and differential binding was assessed with a two-sided *t*-test. For differential induction the *t*-test compared the induction (expression ratio treated/untreated) of the two alleles. Significance of STAT3 binding was assessed by a *t*-test comparing the ChIP signal of STAT3 to the IgG signal. For differential binding we compared the STAT3 binding signal of the high- and low-affinity alleles after subtracting the corresponding IgG signal.

Extension to other transcription factors

Putative regulatory SNPs altering the binding of other transcription factors were identified with the same method used for STAT3. We considered the PWMs included in the JASPAR core vertebrate collection, and performed a general LLR threshold optimization. We downloaded from ENCODE Myers et al. (2011) (UCSC genome browser version hg19) chip-seq peak data for 55 TF associated to known PWMs in JASPAR.

Then we obtained computationally predicted binding sites (TFBS) for each PWM by applying different LLR thresholds. For each PWM and each threshold S (ranging from 5 to 20 at step 1) we counted true positive binding sites (TP: predicted TFBS with LLR≥*S* that overlap with a chip-seq peak of the TF), false positives (FP: predicted TFBS with *LLR*≥*S* that do not overlap with a chip-seq peak of the TF), and false negative (FN: chip-seq peak that have no TFBS with a score ≥*S*.

From TP, FP an FN we computed precision, recall and the harmonic mean of precision and recall, i.e. the balanced F-score. F is a function of the LLR threshold S, and we denoted as *S*\s\do5(*best*) of a given PWM the score that maximizes *F*(*S*) for the appropriate TF. This procedure gave us an optimal value of the LLR cutoff for each PWM that is associated to a TF for which ChIP/seq data are available. To estimate the optimal cutoff for all PWMs, we fitted the values of *S*\s\do5(*best*) to a quadratic model where the idependent variable is a measure of information content of the PWM. We considered three such measures: the entropyof the PWM, the IC as defined by JASPAR and the maximum possible LLR score *S*\s\do5(*max*) that the PWM can produce.

The best fit was obtained using *S*\s\do5(*max*) as the indpendent variable (P=0.0001). This P-value was better than those obtained from linear models, while using polynomial model of higher order did not further improve the significance of the fit. The best fit was given by *S*\s\do5(*best*)=−0.021*S*\s\up(2,*max*)+1.020*S*\s\do5(*max*)+0.032. This function was used

to determine the cutoff in LLR to define putative binding sites for each PWM (including the ones for which ChIP/seq data are available).

Association with disease

We evaluated disease associations of our predicted regulatory SNPs against the genome-wide datasets from the Wellcome Trust Case Control Consortium (WTCCC) Consortium (2007). The original study included 7 complex disease datasets (~2000 samples each) and a dataset of shared controls (~3000 samples): bipolar disorder (BD), coronary artery disease (CAD), Crohn disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D) and type 2 diabetes (T2D).

The original data were downloaded from the WTCCC repository and then filtered by removing samples and SNPs excluded by the WTCCC quality control, and SNPs with genotype quality score (CHIAMO) lower than 0.9. Minimum minor allele frequency and maximum genotype missingness among samples were both set at 0.05. Hardy-Weinberg equilibrium threshold was tested at a cutoff of 0.001. Different cutoff choices (10\s\up5(−5) and 10\s\up5(−7)) resulted in marginal changes on the final set of SNPs.

Out of the 6,682 predicted regulatory SNPs about 1600 were retained for each disease after such filtering, and were analysed with the Cochran-Armitage test for trend Armitage (1955). This test is more conservative than allele-count test and does not rely on an assumption of Hardy-Weinberg equilibrium. It checks the hypothesis of zero slope for the straight line that best fits the risk estimates for the three genotypes. This choice is coherent with the linear-dependence hypothesis used in determining correlation with expression, and agrees with the widely held belief Balding (2006) that the contribution of single alleles to complex diseases is roughly additive.

Population structure, due to the small genomic coverage of the resulting SNP set, was assessed examining results of the original GWA study (PCA and over-dispersion estimates). The authors concluded that, after excluding the non-European ancestry samples, the overall effect of population structure on association results seemed to be small. Therefore our analysis, as the one in the original reference Consortium (2007) does not correct for structure.

Authors contributions

IM, DS, GM, VP and PP conceived, designed and coordinated the study. DS carried out the experiments. IM and FR wrote the analysis software and carried the computational analysis. All authors contributed to the writing of the manuscript, and read and approved its final form.

Acknowledgements

We are grateful to Gabriele Sales and Francesco Vallania for stimulating discussions and suggestions. This work is supported by the following funding agencies: Italian Association for Cancer Research (AIRC) thorugh grants IG-9272 to VP, IG-9408 to PP, CR-4016 to GM; Italian Ministry of University and Scientific Research through PRIN and FIRB grants to VP; Human Genetic Foundation grant to GM; European Union Sixth Framework Program, under the project “Environmental Cancer Risk Nutrition and Individual Susceptibility” (GM). This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the

investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113

References

1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73.

Alenina N, Bashammakh S, and Bader M. 2006. Specification and differentiation of serotonergic neurons. *Stem Cell Reviews*, 2(1):5–10.

Ameur A, Rada-Iglesias A, Komorowski J, and Wadelius C. 2009. Identification of candidate regulatory SNPs by combination of transcription-factor-binding site prediction, SNP genotyping and haploChIP. *Nucleic acids research*, 37(12):e85.

Andersen MC, Engström PG, Lithwick S, Arenillas D, Eriksson P, Lenhard B, Wasserman WW, and Odeberg J. 2008. In silico detection of sequence variations modifying transcriptional regulation. *PLoS computational biology*, 4(1):e5.

Armitage P. 1955. Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3):375–386.

Bai RK and Wong LJC. 2004. Detection and quantification of heteroplasmic mutant mitochondrial DNA by real-time amplification refractory mutation system quantitative PCR analysis: a single-step approach. *Clinical chemistry*, 50(6):996–1001.

Balding DJ. 2006. A tutorial on statistical methods for population association studies. *Nature reviews. Genetics*, 7(10):781–91.

Clay HB, Sillivan S, and Konradi C. 2011. Mitochondrial dysfunction and pathology in bipolar disorder and schizophrenia. *International journal of developmental neuroscience : the official journal of the International Society for Developmental Neuroscience*, 29(3):311–24.

Consortium WTCC. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–78.

Epstein DJ. 2009. Cis-regulatory mutations in human disease. *Briefings in functional genomics & proteomics*, 8 (4):310–6.

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, and Bernstein BE. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–9.

Kaneto H, Matsuoka Ta, Miyatsuka T, Kawamori D, Katakami N, Yamasaki Y, and Matsuhisa M. 2008. PDX-1 functions as a master factor in the pancreas. *Frontiers in bioscience : a journal and virtual library*, 13:6406–20.

Lapidot M, Michal L, Mizrahi-Man O, and Pilpel Y. 2008. Functional characterization of variations on regulatory motifs. *PLoS genetics*, 4(3):e1000018.

Lappalainen T and Dermitzakis ET. 2010. Evolutionary history of regulatory variation in human populations. *Human molecular genetics*, 19(R2): R197–203.

Myers RM, Stamatoyannopoulos J, Snyder M, Dunham I, Hardison RC, Bernstein BE, Gingeras TR, Kent WJ, Birney E, Wold B, and Crawford GE. 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS biology*, 9(4):e1001046.

Newton CR, Graham A, Heptinstall LE, Powell SJ, Summers C, Kalsheker N, Smith JC, and Markham AF. 1989. Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). *Nucleic acids research*, 17(7):2503–16.

Poli V and Alonzi T. STAT3 function in vivo. In Sehgal P, Levy D, and Hirano T, editors, *Signal Transducers and Activators of Transcription (STATs): Activation*, pages 493–512. Kluwer, Dordrecht, 2003.

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavaré S, Deloukas P, Hurles ME, and Dermitzakis ET. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science (New York, N.Y.)*, 315(5813): 848–53.

Vallania F, Schiavone D, Dewilde S, Pupo E, Garbay S, Calogero R, Pontoglio M, Provero P, and Poli V. 2009. Genome-wide discovery of functional transcription factor binding sites by comparative genomics: the case of Stat3. *Proceedings of the National Academy of Sciences of the United States of America*, 106(13):5117–22.

Viken MK, Blomhoff A, Olsson M, Akselsen HE, Pociot F, Nerup J, Kockum I, Cambon-Thomsen A, Thorsby E, Undlien DE, and Lie BA. 2009. Reproducible association with type 1 diabetes in the extended class I region of the major histocompatibility complex. *Genes and immunity*, 10(4):323–33.

Wang Y, Joseph SJ, Liu X, Kelly M, and Rekaya R. 2011. SNPxGE2: a database for human SNP-coexpression associations. *Bioinformatics (Oxford, England)*.

Figures

Figure 1 - Differential induction in homozygous individuals

The induction at 6 hours after treatment with IL6 of individuals homozygous for the high-affinity variant (rows) compared to low-affinity ones (columns). For each comparison we report the P-value of the two-tailed t-test. Red: higher induction for the high-affinity variant, statistically significant (P<0.05); Pink: higher induction for the high-affinity variant, not statistically significant; Green: lower induction for the high-affinity variant, not statistically significant

Figure 2 - Differential induction in heterozygous individuals

Relative induction of the high vs low-affinity allele in heterozygous individuals 6 hours after stimulation with IL6. Error bars represent the SEM computed on two independent samples

Figure 3 - Experimental validation for an intronic SNP in GTF3C6

Experimental validation for the SNP rs9400435, located in the fifth intron of GTF3C6. (A) Induction of GTF3C6 upon stimulation with IL6 at various timepoints for homozygous individuals with the low-affinity allele (green) and high-affinity allele (red). (B) ChIP signal for STAT3, and IgG (negative control) at 1.5 hours after stimulation. Green and red bars correspond to two individuals with respectively low and high affinity alleles. The ChIP signal is measured in unit of the total input. (C)

Relative expression of the high- to low-affinity allele of a heterozygous individual at various time points. (D) ChIP signal for the two alleles of the same heterozygous individual at 6 hrs after stimulation. Green (red) bars correspond to the low (high) affinity allele.

Additional Files

Additional file 1 — Supplementary Table 1

Primers used for SNP genotyping

Additional file 2 — Supplementary Table 2

Primer sequences and probe numbers (Universal Probe Library; UPL) used for quantitative Real-Time PCRs.

Additional file 3 — Supplementary Table 3

Primer sequences used for ARMS-qPCRs

Additional file 4 — Supplementary Table 4

Primer sequences used for conventional ChIP qPCR

Additional file 5 — Supplementary Table 5

Non-exonic SNPs located within 10Kbp form a TSS and significantly correlated with the expression of a neighboring gene.

Additional file 6 — Supplementary Figure 1

PCR and ChIP results for the non-intronic SNPs. For each gene: (A) Induction upon stimulation with IL6 at various timepoints for homozygous individuals with the low-affinity allele (green) and high-affinity allele (red). (B) ChIP signal for STAT3, and IgG (negative control) at 1.5 hours after stimulation. Green and red bars correspond to two individuals with respectively low and high affinity alleles. The ChIP signal is measured in units of the total input.

Additional file 7 — Supplementary Figure 2

PCR and ChIP results for the intronic SNPs. For each gene: (A) Induction upon stimulation with IL6 at various timepoints for homozygous individuals with the low-affinity allele (green) and high-affinity allele (red). (B) ChIP signal for STAT3, and IgG (negative control) at 1.5 hours after stimulation. Green and red bars correspond to two individuals with respectively low and high affinity alleles. The ChIP signal is measured in unit of the total input. (C) Relative expression of the high- to low-affinity allele of a heterozygous individual at various time points. (D) ChIP signal for the two

alleles of the same heterozygous individual at 6 hrs after stimulation. Green (red) bars correspond to the low (high) affinity allele.