Variants in exons and in transcription factors affect gene expression in trans

*Genome Biology* 2013, **14**:R71  doi:10.1186/gb-2013-14-7-r71

Anat Kreimer (anat.kreimer@gmail.com)
Itsik Pe'er (itsik@cs.columbia.edu)

**ISSN** 1465-6906

**Article type** Research

**Submission date** 25 February 2013

**Acceptance date** 31 May 2013

**Publication date** 11 July 2013

**Article URL** [http://genomebiology.com/2013/14/7/R71](http://genomebiology.com/2013/14/7/R71)

This peer-reviewed article can be downloaded, printed and distributed freely for any purposes (see copyright notice below).

Articles in *Genome Biology* are listed in PubMed and archived at PubMed Central.

For information about publishing your research in *Genome Biology* go to [http://genomebiology.com/authors/instructions/](http://genomebiology.com/authors/instructions/)

© 2013 Kreimer and Pe'er

This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Variants in exons and in transcription factors affect gene expression in trans

Anat Kreimer$^{1,2,*}$ & Itsik Pe'er$^3$

$^1$Department of Biomedical Informatics, Columbia University, 622 west 168th St. New York, NY 10032, USA;

$^2$Center of Computational Biology and Bioinformatics, Columbia University, New York, NY 10032, USA;

$^3$Department of Computer Science, Columbia University, 500 west 120th St. New York, NY 10027, USA.

Anat Kreimer (corresponding author) - anat.kreimer@gmail.com

Itsik Pe'er - itsik@cs.columbia.edu
Abstract

**Background:** In recent years many genetic variants (eSNPs) have been reported as associated with expression of transcripts in *trans*. However, the causal variants and regulatory mechanisms through which they act remain mostly unknown. In this paper we follow two kinds of usual suspects: SNPs that alter coding regions or transcription factors, identifiable by sequencing data with transcriptional profiles in the same cohort. We show these interpretable genomic regions are enriched for eSNP association signals, thereby naturally defining source-target gene pairs. We map these pairs onto a protein-protein interaction (PPI) network and study their topological properties.

**Results:** For exonic eSNP sources, we report source-target proximity and high target degree within the PPI network. These pairs are more likely to be co-expressed and the eSNPs tend to have a *cis* effect, modulating the expression of the source gene. In contrast, transcription factor source-target pairs are not observed to have such properties, but instead a transcription factor source tends to assemble into units of defined functional roles along with its gene targets, and to share with them the same functional cluster of the PPI network.

**Conclusions:** Our results suggest two modes of *trans* regulation: transcription factor variation frequently acts via a modular regulation mechanism, with multiple targets that share a function with the transcription factor source. Notwithstanding, exon variation often acts by a local *cis* effect, delineating shorter paths of interacting proteins across functional clusters of the PPI network.

**Key words:** eSNPs, eQTLs, protein-protein-interaction networks, regulatory networks, systems biology, computational biology, systems genetics, regulation, transcriptome
Background

Creating the complete human regulatory map is an active field of study. Many previous studies have used genomic analyses of gene expression, binding motifs, epigenetic marks and other local features to infer regulatory interactions [1-5]. In recent years it has been established that genetic variation can contribute an additional angle to this investigation [6-9]. Formally, transcription level is considered as a quantitative trait that is altered by allelic variation with thousands of SNPs reported as associated with changes in gene expression [10-13]. Such markers, called eSNPs are further found to contribute to variation of disease phenotypes and other clinically-relevant traits [14-16].

Variation in genomic DNA can affect transcription in multiple ways. Most intuitively perhaps, level of transcripts \textit{in cis} of an eSNP may be altered due to allelic variation in regulatory elements [17]. Alternatively, such levels may be auto-regulated by changes in protein structure that reflect variation of the sequence content of local transcripts. Therefore, \textit{cis} eSNPs have been studied extensively. However, \textit{cis} associations are limited in their ability to inform us regarding the network of regulatory interactions between one gene and another. This motivates more focused study of the effects of genetic variants on expression of distal transcripts (\textit{trans} associations). Unfortunately, while \textit{trans} eSNPs can identify downstream effects and previously un-annotated regulatory pathways, they are harder to statistically and biologically justify than \textit{cis} eSNPs. From a statistical perspective, since \textit{trans} eSNPs can be associated with any distal transcript, the multiple testing burden dramatically increases, thus only a small number of results is detected. From a biological perspective, more complex mechanisms are needed to explain \textit{trans} associations. An example of such a mechanism is an eSNP with local \textit{cis} effect on a gene which codes for a transcription factor known to regulate other genes \textit{in trans}. Indeed, across multiple eSNP studies [7, 18], even when statistically significant \textit{trans} or \textit{cis} eSNPs associations are detected aplenty, the regulatory mechanisms by which they alter gene expression remain mostly unknown.

A large fraction of SNPs identified by GWAS [19] have been reported to be associated with disease phenotypes [14] despite being neither coding, nor linked to coding SNPs in \textit{cis}. Furthermore, since large-scale genetic studies had been predominantly based on SNP arrays,
SNP alleles that are reported as associated, in studies of either disease [20] or gene expression [7] are often merely tags for causal variants, whose identity is challenging to track down. More generally, the multitude of phenotypes for eSNPs represents an opportunity for tackling the central question of causation in association.

Protein-Protein Interaction (PPI) networks capture various experimental data, e.g. from yeast two-hybrid systems [21], regarding physical binding of proteins, often in order to perform a specific biological function. Recently improved data on signal transduction, metabolic and molecular networks have contributed to the fidelity and accuracy of the reconstructed PPI networks. Notwithstanding, the data represented by these networks can sometimes be partial and noisy. PPI networks were modeled as theoretical graphs and their topological properties were extensively studied [22-24]. This provided insights pertaining to functional, structural and evolutionary characterization of these networks, primarily in model organisms. Genetic interactions in yeast were studied in the context of a network of protein complexes [25], motivating investigation of genetic variants that alter gene expression as interactions with respect to the human PPI network [26]. Studies of PPI networks in the context of genetic variation had thus far focused on GWAS-detected SNPs that are associated with common traits and disease, reporting genes that harbor such SNPs to frequently code for interacting proteins [26-30]. Yet, such studies only considered the PPI-network nodes that correspond to the associated SNP, without a PPI network node that would correspond to the phenotype.

Here we perform a comprehensive study of trans genetic associations and their large scale properties as manifested on a PPI network. We use SNPs from sequencing data [31] that are candidates to be causal based on their genomic location, and then project their association to gene expression on a PPI network. We hypothesized that genes involved in true eSNP associations have distinct PPI-network properties that differ significantly from spurious genes with candidate association signals. To address this hypothesis, we focus on trans association of eSNPs in exons and transcription factors (TFs), analyzing their properties as reflected on the PPI-network topology and annotations of the genes involved. Our focus on expression quantitative traits allows consideration of paths along the PPI network, whose links with genetic variation had previously only studied w.r.t. SNPs, rather than the transcripts they modulate.
Our results suggest that a significant fraction of eSNPs in exons act *in trans* through mild effects *in cis*, with a regulation mechanism that is mediated by PPI paths that are shorter than expected by chance and tend to traverse across functional clusters of the PPI network. These paths highlight zinc ion binding genes as a possible mechanism of transcript-eSNP feedback across the PPI network. In comparison to such coding eSNPs, we observe that TFs harboring eSNPs and their associated genes create units of genes that are functionally enriched for biological annotations. This suggests a different, modular regulatory mechanism for such TF eSNPs. Altogether, our analysis offers insights concerning a variety of mechanisms by which genetic variation at functional loci shapes the structure of human regulatory networks.

**Results**

Computational Framework for Mapping *Trans* Associations onto the PPI Network

We were interested in pinpointing directly associated variants rather than indirectly imputed ones. We thus used a publicly available dataset of 50 fully sequenced Yoruban samples [31] along with their transcription profiles from RNA-seq data [32], bearing in mind that such available cohorts are limited in size. Due to this small sample size, we have limited power in detecting association. Therefore, most candidate eSNPs can only be designated with various levels of uncertainty.

We were intrigued to examine *trans*-eSNPs interactions with respect to an independent space of interactions, i.e. PPI network. Therefore, we evaluated two categories of candidate eSNPs that reside within regions along the genome with known regulatory potential and can be mapped onto a PPI network, i.e. exons and TFs (Figure S1 and Table S1 in Additional file 1; see Methods). Examining the distribution of p-values across these two categories of candidate *trans*-eSNPs, we observe that candidate eSNPs within exons show evidence of including true positive eSNPs (Figure S2a in Additional file 1), as been previously shown [33]. In contrast, eSNP candidates in TFs show association signal distributions consistent with random expectation (Figure S2b in Additional file 1). We further examine if TF candidate eSNPs exhibit qualities that are different than random. We hypothesized that a single TF will be associated to multiple transcripts via eSNPs. To address this hypothesis, we created 1,000 permuted sets of pairs of TF and transcript (see Methods). We observe that the number of multiple associated transcripts is significantly
higher (Wilcoxon rank sum test \( p < 0.05 \)) in the real dataset (973/1000 permuted sets, empirical \( p \)-value = 0.027). Following these two observations, we focus on eSNPs within exons as the first subject of our investigation, and compare them to eSNPs within the span of transcription factor genes. We set out to characterize and compare these two modes of \textit{trans} regulation.

For each candidate eSNP that is associated with levels of a transcript in \textit{trans}, we denote this transcript as the \textit{target} of the eSNP. When this eSNP is located within the span of a gene (see Methods), we define this gene as \textit{source}. We attempted to characterize eSNPs interactions on the molecular level, by mapping these pairs of source-target genes onto a PPI network (see Methods and Figure 1) and study their functional annotations and topological properties.

Identifying Topological Properties of Exonic eSNP Interactions

We first considered pairs of exon eSNP source and target that demonstrate an association signal which is exome-wide significant for a particular transcript (association \( p \)-value < \( 10^{-7} \)). We observe such pairs to be significantly closer (\( p=0.03 \)) on the PPI network when compared to randomly permuted candidate eSNPs (see Methods). Beyond pairwise properties of sources and targets, we further attempted to characterize each by their single-node features. Specifically, the targets of exon eSNPs have significantly higher (\( p=0.003 \)) degree than expected based on random pairs.

We reasoned that the cutoff of association \( p \)-value we used (\( p<10^{-7} \)) is in many ways arbitrary, as we are interested in statistical properties of the set of results rather than the significance of a particular result amid the testing burden. We therefore considered multiple \( p \)-value thresholds of eSNP association, at each evaluating topological properties of eSNP source and target pairs, while assessing significance vis-à-vis randomly permuted sets of candidate eSNPs in exons (see Methods). We observe that the lower the association \( p \)-values for source-target pairs, the more their topological properties differ compared to random pairs (Table S2 in Additional file 2). For example, for source-target pairs of exon eSNP, the average target degree among the 52 pairs exceeding an association \( p \)-value cutoff of \( 10^{-6.5} \) is 16.42, but it reaches as much as 22.22 among the more focused set of 22 pairs that exceed association \( p \)-value cutoff \( 10^{-6.8} \). These averages are each significant (\( p \)-values 0.02 and 0.006, resp.) when compared to permuted pairs of exon eSNPs, whose target degree is only 9.36 on average. These trends are consistent with properties
of true positives being diluted by false positives at less significant p-value thresholds. We quantify such trends by regressing each topological property on the negative log10 of the association p-value (Figure 2). We confirm that for exonic source-target pairs, network distance decreases and the target degree increases with the –significance of association (Spearman rank correlation coefficients \( r = -0.98 \) and \( 0.97 \), respectively; permutation p-value \( p = 0.001 \) and \( 0.002 \), respectively – see Methods).

These results highlight unique properties of part of the transcripts whose trans regulation is due to coding variation. Specifically, we show that loci implicated by eSNPs, encode for proteins that physically interact in a non-random fashion. Furthermore, target proteins are likely to interact with significantly more nodes of the PPI network than expected by chance.

Characterization of Exon and TF Sources and Targets

Based on these results, for further analysis, we focused on the maximal p-value cutoff of \( 10^{-6.463} \), for which all topological properties show significant difference between true source-target pairs of exon eSNPs and random ones (Wilcoxon rank sum test \( p < 0.05 \), (Figure S3 and Table S3 in Additional file 1, Table S2 in Additional file 2).

There are 343 pairs of source and target, 295 unique pairs, 59 of them on the network. 318 (92.71%) of these pairs are on different chromosomes and 25 (7.29%) are on the same chromosome, at least 1MB apart (See Table S4a in Additional file 1). At this cutoff there are 333 unique eSNPs in exons, 286 unique gene sources and 267 unique gene targets (Table S5 in Additional file 3). When comparing the effect sizes (absolute values of betas in the linear regression) of previously published 929 cis eQTLs in [32] with the distribution of exonic and TF trans eSNPs effect sizes. We find that the trans effect sizes are (mean 1.198) significantly higher than those of corresponding cis effects (mean 0.964) (Wilcoxon rank sum test p-value < 2.25\( \cdot 10^{-49} \) and 3.56\( \cdot 10^{-54} \) for exonic and TF eSNPs, respectively) (See Figure S4 in additional file 1). We binned eSNPs and SNPs in exons by first, middle and last exons (See Figure S5 in additional file 1). We also examined the position of the eSNP along the transcript and compared these results to SNPs in exons (See Figure S6 in additional file 1). We observe that these trans exonic eSNPs tend to be located along middle exons, rather than in first or last exons (Fisher’s exact test p-value < 0.009). We further observe that they tend to lie farther away down the transcript
(Wilcoxon rank sum test p-value = 0.0058). These results are different from what was observed for *cis* eQTLs. Montgomery et al. [34] reported that eQTLs with higher confidence are located in the first and last exons significantly more than in middle exons.

The combined set of exon sources is enriched for MHC protein complex genes (FDR < 0.046) with concordance to findings in previous studies, indicating HLA SNPs were 10-fold enriched for *trans*-eSNPs [35]. We further observe that the set of target genes is enriched for multitude functional processes (see Table S6 in Additional file 4 for full list of annotations). The three highest scoring functional annotations of the target set: macromolecule modification, phosphatidylinositol-3,5-bisphosphate binding and protein modification process provide an additional support for the role of exonic eSNP targets as network hubs [36].

For further investigation and comparison, we considered source-target pairs of TF candidate eSNPs, a set with similar order of magnitude, corresponding to association signals passing the p-value cutoff of $10^{-6}$. There are 370 such pairs of TF source-target, 193 of them unique, 58 of which are on the network. 359 (97.03%) of these pairs are on different chromosomes and 11 (2.97%) are on the same chromosome, at least 1MB apart (See Table S4b in Additional file 1). There are 358 unique eSNPs in TFs, 77 unique TF sources and 192 unique targets (Table S5 in Additional file 3). Out of the 358 unique eSNPs in TFs, 15 are in exons, significantly more than expected by chance (hypergeometric p-value < 0.00018). When we examine the combined set of TF targets, we observe that this gene set is enriched for various annotation categories (see Table S6 in Additional file 4 for full list of annotations).

To further establish the association between the source and target genes, we examined the co-expression between eSNP source and target for all candidate pairs of associated genes in this dataset by evaluating Spearman’s rank-correlation coefficient $r$ (Methods). For pairs of exon-source eSNPs and their corresponding targets, the absolute value of $r$ was significantly higher than expected from the entire distribution of co-expression measurements in this dataset (Wilcoxon rank sum test $p < 5.4 \cdot 10^{-5}$). In contrast, for pairs of TF-source eSNPs and their corresponding targets, there was no significant difference in terms of co-expression. We observe the fraction of non-synonymous SNPs to be 0.082 out of exon eSNPs, which is higher than their overall fraction 0.071 among all exonic SNPs [37] (fisher exact $p=0.1$). Furthermore, we
examined *cis* effects that are too mild to be detected at genome-wide significance threshold, of each eSNP by testing for its association to the expression of its source gene. In total, 50 pairs of exonic eSNP and source gene were nominally (P<0.05) *cis* associated, out of 286 such unique sources (p=3.6⋅10^{-15} see Methods). Furthermore, we estimate how many of the SNPs in exons have a *cis*-effect (linear regression p-value < 0.05) on the expression of their host gene. We find that out of 97,135 exonic SNPs, 9,661 show *cis*-effect on their host gene at the nominal significance level, p-value<0.05. Compared to this background distribution, the observed 50 out of 286 *trans* eSNPs having such *cis*-effects are significantly more than expected by chance (Fisher’s exact test p-value < 9.6⋅10^{-5}). This provides additional support for the *cis*-effect phenomena. For comparison, we do not observe a nominally significant *cis* effect between TF eSNP and its source gene more than expected by chance (3 out of the 66 TF sources in this dataset). These results suggest a mechanism where exonic variation often operates in *trans* eSNPs via alteration of gene expression in *cis*, and the source and target genes have correlated expression.

TFs are known to control the transcription of multiple genes; we were therefore interested in examining if we observe the same phenomena in TF variation. We notice that each TF source forms, along with its targets, a set of genes that we call a “unit”. We observe that these units tend to be enriched for functional annotation categories. Specifically, for the 33 TF sources with 2 target genes or more (Tables S7 and S8 in Additional file 1), 26 out of 33 define units that are functionally enriched (2 or more annotated genes, FDR < 0.05) [38] in KEGG [39] and GO [40] categories (Table S9 in Additional file 5). Interestingly, eSNP targets do not tend to share exon sources. Specifically, out of 286 unique sources, 278 have a single target, 7 (AKNA, CDK7, BLK, ATP5G1, RPL8, TRAPPC12, MUC2) of the remaining ones have two and one (HLA-C) has three (Table S5 in Additional file 3). The difference between the number of associated targets in TF and exon variation is statistically significant (Wilcoxon rank sum test p < 3.4⋅10^{-4}). These results support the hypothesis that TF variation frequently acts via a modular regulation mechanism, with multiple targets that share a function with the TF source.

We systematically looked for pairs of TF source-target that were experimentally validated as binding. We found such enrichment, 6/34 TF source-target pairs compared to 551 out of 6,904 random pairs (Fisher’s exact test p < 0.05, see Methods) in a database reporting binding of
transcription factors to DNA, based on ChIP-X experiments [41]. Furthermore, we used the data in [42] to find the closest DNaseI hypersensitive site (DHS) window to the gene target, and examined whether the TF eSNP is associated with the DHS levels in this window. We find that 33 of 370 such pairs of TF eSNP and gene target are significantly associated (P<0.05) indicating significant enrichment (p<5.5·10^{-4}) of this phenomenon. This enrichment is not an artifact of TF eSNP ascertainment: we tested association of 29,212 TF SNPs to DHS levels in a randomly picked DHS window; As expected by chance, 1,400 of these SNPs show such association at the nominal significance level, p-value<0.05. Compared to this background distribution, the observed set of 33 out of 370 trans eSNPs having such association is significantly larger than expected by chance (Fisher’s exact test p-value < 6·10^{-4}). This shows that even in results where the number of true positives is diluted with false positives, due to a small sample size, we still recover true signal.

We were intrigued by potential connections between source-target pairs and cluster properties in the PPI network. Therefore, we partitioned the PPI network into clusters of genes, optimizing the modularity measure [43]. Out of the resulting 249 PPI clusters with two genes or more, 225 (90%) demonstrate functional enrichment for a biological category (Table S10 in Additional file 6). TF source-target pairs are found in the same PPI clusters more than expected by chance: 26/58 TF pairs compared to 26,966 out of 100,000 random pairs (Fisher’s exact test p < 0.0043).

As an illustration for our results, we show an example (Figure 3a) of a specific source and its gene target, examining TCF7L2 (transcription factor 7-like 2; T-cell specific, HMG-box) and its transcript target TLE4 (transducin-like enhancer of split 4). There is a significant cis effect (p < 0.012) of the associated intronic eSNP rs7087006 with the expression of TCF7L2, but the co-expression correlation of the source and target, is not statistically significant in this dataset. TCF7L2 and its 5 targets (unit number 28, Table S8 in Additional file 1) comprise a unit that is enriched (2 out of 6) for cell proliferation (FDR<0.03) (Table S9 in Additional file 5). This TF plays a key role in the Wnt signaling pathway, activating MYC expression in the presence of CTNNB1. The gene target TLE4 within the PPI network is a transcriptional co-repressor that represses transactivation mediated by TCF7L2 and CTNNB1. These annotations implicate that TCF7L2, TLE4 and MYC act as the network motif I1-Feed Forward Loop (a pulse generator and response accelerator) [44] where the two arms of the FFL act in opposition: TCF7L2 activates
MYC (in presence of CTNNB1) but also represses MYC by activating the repressor TLE4 (via an eSNP). We note that TCF7L2 harbors the common allele most strongly associated with increased risk of type 2 diabetes. Correspondingly, TLE4 was recently discovered as a T2D locus [45]. Specifically, TLE4 encodes a protein that forms complexes with TCF proteins, including TCF7L2, to modulate transcription at target sites [46]. The source and target are part of the same PPI network cluster, which is enriched (1257 out of 4627) for regulation of transcription (FDR < 2.4·10⁻⁸⁸, Table S10 in Additional file 6) (Figure 3a). This demonstrates a case of shared function between a source TF and its target.

In contrast, only 19 (32%) of exon eSNP sources are found in the same PPI network cluster as their respective single targets, consistent with chance expectation (Methods). Yet, as such pairs are linked by relatively shorter paths (Figure 2a) it follows that coding variants affect transcription in trans not in a modular way but rather in a linear fashion that defines shorter paths than expected by chance. We recorded the proteins along such paths (Table S11 in Additional file 1) and evaluated the enrichment of functional annotation for each path (Table S12 in Additional file 7).

We show an example (Figure 3b) of exon source and its gene target, examining the path between gene source PIDD and gene target PLK3 (path number 18, Tables S11 in Additional file 1 and S12 in Additional file 7). This path is enriched for p53 signaling pathway (FDR < 0.01, Table S12 in Additional file 7). The eSNP exon source is PIDD (p53-induced death domain protein). This gene promotes apoptosis downstream of the tumor suppressor as component of the DNA damage/stress response pathway that connects p53 to apoptosis. The gene target PLK3 polo-like kinase 3 is a serine/threonine kinase that plays a role in regulation of cell cycle progression and potentially in tumor-genesis. EGF containing fibulin-like extracellular matrix protein 2 (EFEMP2) and tumor protein p53 (TP53) reside along the shortest path between PIDD and PLK3 (Figure 3b). There is evidence from ChIP-ChIP and ChIP-seq experiments that TP53 has binding sites in the promoter of PLK3 [41] and it is annotated as a zinc ion binding protein. Furthermore, the combination of a pair of genes with TF-DNA and PPI edge between them is a known network motif (mixed-feedback loop) [47], suggesting a mechanism by which the expression of the target gene is altered. In support, the co-expression correlation of the source and target genes is significant (Spearman rank-correlation test r= 0.3223, p-val<0.02). The exon
gene source and target reside in different PPI network clusters: PIDD resides in a cluster that is
enriched for regulation of cell death (FDR < 4.5 \cdot 10^{-6}, Table S10 in Additional file 6) and
PLK3 resides in a cluster which is enriched for regulation of transcription (FDR < 2.4 \cdot 10^{-88},
Table S10 in Additional file 6).

These results beg a mechanistic explanation that would clarify how is the network interaction at
the protein level leading to the observed changes in transcript levels. Fortunately, examination of
the genes along the reported paths provides a plausible answer, as they are strongly enriched for
zinc ion binding proteins. Specifically, when we examine the enrichment for annotations of
genes along shortest paths in the real dataset we observe 410 enriched categories (minimum of
10 genes from a category, FDR < 0.05, Table S13 in Additional file 8). For comparison, across
1,000 permuted datasets we observed a total of 1,870 categories satisfying the same enrichment
criteria. We focus on the 6 categories that are enriched in real data and not in permutations: ion
binding, metal ion binding, cation binding, intracellular, zinc ion binding and transition metal ion
binding (Table S13 in Additional file 8). We compare two properties in real vs. permuted
datasets (Methods): (1) the number of genes from each category (empirical p-values 0.005 and
0.014 for zinc ion binding and transition metal ion binding respectively) (2) number of paths
where we observe at least one gene from each category (empirical p-values 0.016 and 0.038 for
zinc ion binding and transition metal ion binding respectively). These results replicate in a
second permuted dataset. For comparison, only 7 and 10 out of the 404 joint categories achieve
an empirical p-value smaller than 0.05 for these two properties respectively. These results
indicate that the genes in real paths are enriched for zinc ion binding which is associated with
regulation of transcription, suggesting a possible mechanism by which the expression level of the
target transcript is modified.

**Discussion**

We presented a computational approach to study characteristics of *trans* regulation. We observed
that candidate eSNPs within exons exhibit overabundance of significant association signals. We
consequently focused on eSNPs that reside within an exon of a source gene, and are associated
with the expression level of a different gene target. Furthermore, we observed that candidate
eSNPs within TFs are associated with a higher number of transcripts than expected by chance.
We subsequently examined eSNPs that reside within the span of source TFs. We mapped these pairs of source and target onto a PPI network, and analyzed their topological properties.

We applied our approach to publicly available genetics and genomics [32] data from the same samples. We demonstrate that by combining association data with information on PPI it is possible to unravel topological properties for the two trans association types. We find that for an eSNP exon source and its gene target, the stronger the association, the closer the source-target distance and the higher the target degree in the PPI network. Expression analysis shows these source–target pairs to be frequently co-expressed, and that these exon eSNPs often have significant cis effects on the expression of the source genes. The observed phenomenon of exonic variation leaving a signature on PPI paths raises speculations regarding the mechanisms of transcription regulation. These speculations regarding the connection between eSNP regulation and the PPI space have been indirectly tackled in previous studies. Specifically, Rossin et al. find that PPI connections between loci defined in GWAS of a specific disease are more densely connected than chance expectation [26] and Nicolae et al. [14] observe that GWAS SNPs are more likely to be eSNPs. The comprehensiveness of this work relies on combining eQTL data with the PPI network and not merely GWAS data, as described in previous studies [29]. This allowed us to examine source-target connections across the network, rather than be limited to studying the source nodes as in GWAS-PPI analyses. The novel observation is that genetic variation that modifies PPI network properties is associated with normal expression landscape and not only with extreme cases of disease.

We attempted to go beyond topological results and shed light on the regulatory mechanism by which gene expression of the target gene is altered in these shorter paths. We thus systematically compared genes along real and permuted shortest paths and found enrichment for ion zinc binding proteins, suggesting a plausible mechanism by which the expression level of the target transcript is modified. More generally, the paths of interacting protein pairs, from a source protein to the target protein, are consistent with concatenation of two pathways (Figure S7 in Additional file 1). The prefix of the path is consistent with a regulatory pathway, leading to some regulatory protein (transcription factor or other), that affects expression of the target. The suffix of the path may match a self feedback loop in reverse: from the target protein back to the same regulatory protein [48].
Furthermore, we demonstrate it is possible to characterize regulatory variation in TFs. We observe that eSNP TF sources and their gene targets create units of genes that are enriched for functional annotations. When decomposing the PPI network to clusters, we observe that these source–target pairs tend to reside within the same cluster.

Obviously, one should acknowledge that the design choices for a study of this kind convey a few methodological limitations. First, since we were interested in detecting putatively causal variants based on their exact genomic location, we used a dataset of fully sequenced individuals along with their transcription profiles. Such cohort sizes are limited in size, reducing the power to detect association and enabling us to see only the strongest effects. Second, we were interested in understanding the mechanisms underlying eSNPs interactions. This required the use of a well-established interaction network. We examine our results on a PPI network, rather than a TF-DNA interaction network or co-expression network derived from this dataset, to establish a broad and independent network of interactions. Overall, both the raw datasets [31, 32] and supporting databases [39-41, 49-51] in this work are noisy and limited. The fact that we observe statistically significantly plausible results in such a small dataset combined with noisy databases is encouraging. Potential increase in sample size may enable detection of eSNP associations at more significant p-values for even milder effects.

**Conclusions**

Over the last decade, causal interpretation of genetic association signals for common variants and common traits had been impeded by two hurdles. First, many of the signals had been obtained as indirect association to proxy genetic markers, without access to the directly and causally associated variant. Second, often the trait under investigation is not understood at the molecular mechanistic level well enough to decipher the connection between variant and phenotype. This work bridges the gap between association and causality by considering both direct association to sequencing-ascertained variants, as well as expression quantitative traits. The ability to tie together these loose ends of genetic association using an interaction map constitutes a notable stride towards understanding the thousands of such connections that recent genetics have discovered.
Our main findings suggest two modes of trans regulation via genetic variation in exons and TFs. Exonic variation possibly acts through mild cis effects that alter the expression of the source gene and delineates shorter paths between functional clusters (Figure 4a) and exonic eSNP targets might play an important role in the PPI network as hubs. Whereas, TF variation frequently acts via a modular regulation mechanism, with multiple targets that share a function with the TF source (Figure 4b).

Future studies could extend the approach presented here to investigate how genetic variation in different meaningful genomic locations (e.g. enhancers, insulators, miRNAs) correlates with gene targets. Datasets that combine sequenced variants coupled with gene expression and phenotypic traits are limited in human, but available for other model organisms [52, 53]. It would be insightful to combine this type of study with phenotypic data, to see how trans association tracks with phenotypes. Specifically, applying our approach to samples under various conditions (e.g. disease), could improve understanding of condition-specific regulatory processes [26]. Moreover, considering genetics-genomics data across different tissues along with a tissue-specific PPI network [54] could be telling regarding the underlying regulatory mechanisms characterizing these tissues.

**Materials and methods**

**Data Details and Processing**

We analyze a cohort of 50 Yoruban samples, for which genotypes of SNVs that are fully ascertained from sequencing data [31] along with RNA-seq data[32] are publicly available. Briefly, the raw dataset consists of 10,553,953 genotyped SNVs and expression measurements (quantile-quantile normalized values) of 18,147 genes with Ensembl gene ID across these 50 samples. Standard filters have been applied to the genetic data: Minor allele frequency > 0.05, SNP missingness rate < 0.1 and individual missingness rate < 0.1 [55]. After filtering, data for analysis consists of 50 samples with 7,206,056 SNPs.

**Association Testing**

For association analysis, we consider only SNPs that reside within candidate regulatory regions along the genome. For trans association, we test for association between a SNP and every gene;
we consider SNPs within the span of known exons and TFs (including introns) [56]. We test for association (Figure S2 in Additional file 1) using linear regression.

**Identifying Pairs of Source and Target for Exonic and TF Variation**

For each eSNP that is putatively associated with levels of a transcript in *trans*, we define this transcript as *target*. When this eSNP that is located within an exon or in the span of a TF, we define this gene as *source*.

**Obtaining a Random Distribution of Association Test-Statistics**

Examining the random distribution of association tests is helpful in evaluating the empirical significance of results. This is achieved by generating 100,000 random pairs of sources and targets for exonic and TF variation separately. We use a strict randomization process of edges switching. We pick a source gene from all sources in the real data; we then pick a target gene from all targets in the real data with a p-value cutoff of $10^{-6}$. When evaluating the number of targets per TF source, we created 1,000 sets of random TF source and gene target pairs; each set contains 370 such pairs corresponding to 370 TF source-target pairs at a p-value cutoff of $10^{-6}$ in the real data.

**Identifying Topological Properties of Source-Target Pairs Projected on the PPI Network**

We use the PPI network provided by the Human Protein Reference Database (HPRD) [51]. The undirected network contains 9,671 nodes and 37,041 edges. For each node, we calculate its degree: the number of edges incident on the node. We define a distance between every two nodes as the number of edges on the shortest path between them. All pair-wise shortest paths are determined, using the Floyd–Warshall algorithm [57]. In cases where the network has more than one connected component, nodes from two different components are defined to have a distance of twice the maximal distance obtained within the components.

**Identifying Topological Trends across Association P-values**

We observe for exons, the emergence of true positive associations between p-values $10^{-6}$ and $10^{-7}$ (Figure S2a in Additional file 1). Therefore, we focus of p-values $< 10^{-6}$ and sort all source-target pairs according to the significance of their association signal. We consider each prefix of this list,
i.e. each subset of source-target pairs exceeding a particular threshold for significance of
association signal. For each such subset we report, each one of the topological properties defined
above, averaged over the subset. We calculate Spearman’s correlation coefficient between
significance thresholds and each of these cumulative averages. In a similar process we randomly
choose an equal number of arbitrary source-target pairs on the PPI network. Adding these pairs
one by one creates a distribution of analogous cumulative averages for permuted pairs. We
record the Spearman correlation coefficient for these 100,000 permuted distributions. We
calculate the empirical p-value for the significance of the observed correlation coefficients by
counting the number of times when |permuted r| > real r and divide this number by the number of
permutations.

Expression Analysis
We calculated all pairwise co-expression correlations for all gene pairs in the dataset using
Spearman rank-correlation test, and therefore obtained the distribution of the correlation
coefficient \( r \). To determine whether the distribution of \( r \) between source-target pairs differs from
its background distribution, we employed the Wilcoxon ranked-sum test.

Enrichment of eSNPs for cis-Effects
We examine whether eSNPs that are associated with a target’s expression level also affect
expression levels of the corresponding source. We test this by considering, for each source-target
pair, the one eSNP most associated to the expression for the target. We tally the source-target
pairs for which this eSNP is also significantly (p<0.05) associated with the expression level of
the source. Under the null, the number of such pairs is a random variable that is binomially-
distributed Bin(n = #unique source genes, P=0.05).

Unit and Path Annotation
We define units of genes by considering a TF source and its gene targets. We examine shortest
paths within the PPI network between eSNP exon source and its gene target. The enrichment of
units and paths with gene subsets from the Gene Ontology [40], and KEGG [39] databases was
calculated by Genatomy [38]. We report only units or paths with annotations that have a
significant FDR of 0.05 or better. The description of genes in units or paths is cited from NCBI
Gene database and genecards [58].
Finding TF Source-Target Pairs in Experimental Data Base

ChEA database [41] represents a collection of interactions describing the binding of transcription factors to DNA, collected from ChIP-X (ChIP-chip, ChIP-seq, ChIP-PET and DamID) experiments. For each TF source and target we examine if they are present in ChEA. We repeat the same procedure for 100,000 permuted pairs of a random TF source and a random gene target. We then compare, using Fisher’s exact test, the number of pairs in ChEA between real and permutation pairs, out of all pairs where the TF source is included in the database.

Finding PPI Network Decomposition to Clusters

The decomposition of the PPI network to clusters is computed by using the Louvain algorithm presented in [59]. This is a heuristic method that is based on modularity optimization. The method consists of two phases and partitions the network into clusters such that the number of edges between clusters is significantly less than expected by chance. The method provides a mathematical measure for modularity with network- size normalized values, ranging from 0 (low modularity) to 1 (maximum modularity). This method has been previously applied to various biological networks [60] and specifically to a PPI network [61].

Significance of Source and Target Residing in the Same PPI Cluster

For each exon/TF source-target pairs we record whether both source and target reside in the same PPI cluster. We repeat the same procedure with 100,000 permuted unique source-target pairs from nodes on the PPI network. We then compare the number of cluster co-occurrences between real data and permutations using the Fisher exact test.

Comparing Shortest Paths Annotation Content

We recorded all genes along the shortest paths between exonic sources and targets, both in real and permuted data. We then look for enrichment in this set of genes (at least 10 genes per category, FDR <0.05). We create sets of 1,000 permuted 55 shortest paths (from the 17,564 shortest paths in permutations) that follow the exact length distribution of the real 55 paths. For each one of the six categories that is not enriched in permutations, we preform two analyses: (1) we count how many genes from each category appear in the real paths (with repetitions, i.e. if
gene X from category Y appears in two shortest paths we count it twice), (2) we count how many
of the 55 paths have at least one gene from this category. We repeat the same procedures for the
1,000 permuted sets. For each category, we then count how many of the 1,000 permutations
achieved equal or greater number than the one we see for real data (empirical p-value).

**Abbreviations**
eQTL – expression quantitative trait loci ; eSNP – expression single nucleotide polymorphism;
PPI – Protein Protein Interaction.

**Authors’ contributions**
AK designed and performed research, analyzed the data and drafted the manuscript. IP
conceived and designed research and wrote the paper. All authors read and approved the final
manuscript.

**Acknowledgments**
We want to thank Dr. Joseph Pickrell for enabling access and providing clarifications regarding
the data published in [32]. This work was supported by the following funding: NIH U54,
CA121852-07, NSF08929882 and NSF0845677.

**References**

1. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK: Accurate
inference of transcription factor binding from DNA sequence and chromatin
accessibility data. *Genome Res* 2011, 21:447-455.

2. Segal E, Friedman N, Koller D, Regev A: A module map showing conditional activity
of expression modules in cancer. *Nat Genet* 2004, 36:1090-1098.

3. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R,
Califano A: ARACNE: an algorithm for the reconstruction of gene regulatory
networks in a mammalian cellular context. *BMC Bioinformatics* 2006, 7 Suppl 1:S7.

4. Yosef N, Ungar L, Zalckvar E, Kimchi A, Kupiec M, Ruppin E, Sharan R: Toward
accurate reconstruction of functional protein networks. *Mol Syst Biol* 2009, 5:248.

5. Ihmels J, Bergmann S, Berman J, Barkai N: Comparative gene expression analysis by
differential clustering approach: application to the Candida albicans transcription
program. *PLoS Genet* 2005, 1:e39.

6. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Ghahakura D, Sieberts SK, Monks S,
Reitman M, Zhang C, Lum PY, Leonardson A, Thieringer R, Metzger JM, Yang L,
Castle J, Zhu H, Kash SF, Drake TA, Sachs A, Lusis AJ: An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet* 2005, 37:710-717.

7. Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, Kasarskis A, Zhang B, Wang S, Suver C, Zhu J, Millstein J, Sieberts S, Lamb J, GuhaThakurta D, Derry J, Storey JD, Avila-Campillo I, Kruger MJ, Johnson JM, Rohl CA, van Nas A, Mehrabian M, Drake TA, Lusis AJ, Smith RC, Guengerich FP, Strom SC, Schuetz E, Rushmore TH, et al: Mapping the genetic architecture of gene expression in human liver. *PLoS Biol* 2008, 6:e107.

8. Brem RB, Yvert G, Clinton R, Kruglyak L: Genetic dissection of transcriptional regulation in budding yeast. *Science* 2002, 296:752-755.

9. Cheung VG, Spielman RS: Genetics of human gene expression: mapping DNA variants that influence gene expression. *Nat Rev Genet* 2009, 10:595-604.

10. Rockman MV, Kruglyak L: Genetics of global gene expression. *Nat Rev Genet* 2006, 7:862-872.

11. Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R, Kruglyak L: Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nat Genet* 2003, 35:57-64.

12. Gilad Y, Rifkin SA, Pritchard JK: Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet* 2008, 24:408-415.

13. Kreimer A, Litvin O, Hao K, Molony C, Pe'er D, Pe'er I: Inference of modules associated to eQTLs. *Nucleic Acids Res* 2012, 40:e98.

14. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ: Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* 2010, 6:e1000888.

15. Moffatt MF, Kubesch M, Lian L, Dixon AL, Strachan D, Heath S, Depner M, von Berg A, Bufe A, Rietschel E, Heinzmann A, Simba B, Frischer T, Willis-Owen SA, Wong KC, Illig T, Vogelberg C, Weiland SK, von Mutius E, Abecasis GR, Farrall M, Gut IG, Lathrop GM, Cookson WO: Genetic variants regulating *ORMDL3* expression contribute to the risk of childhood asthma. *Nature* 2007, 448:470-473.

16. Kathiresan S, Melander O, Guiducci C, Surti A, Burtt NP, Rieder MJ, Cooper GM, Roos C, Voight BF, Havulinna AS, Wahlstrand B, Hedner T, Corella D, Tai ES, Ordovas JM, Berglund G, Vartiainen E, Joussilhati P, Hedblad B, Taskinen MR, Newton-Cheh C, Salomaa V, Peltonen L, Groop L, Altshuler DM, Orho-Melander M: Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat Genet* 2008, 40:189-197.

17. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grasso A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavare S, Deloukas P, Hurles ME, Dermitzakis ET: Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 2007, 315:848-853.

18. Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, Ingle C, Beazley C, Gutierrez Arcelus M, Sekowska M, Gagnebin M, Nisbett J, Deloukas P, Dermitzakis ET, Antonarakis SE: Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* 2009, 325:1246-1250.
19. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* 2009, 106:9362-9367.

20. GWAS catalog

21. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM: A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 2000, 403:623-627.

22. Wuchty S, OltvaiZN, Barabasi AL: Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat Genet* 2003, 35:176-179.

23. Yook SH, OltvaiZN, Barabasi AL: Functional and topological characterization of protein interaction networks. *Proteomics* 2004, 4:928-942.

24. Jeong H, Mason SP, Barabasi AL: Lethality and centrality in protein networks. *Nature* 2001, 411:41-42.

25. Kelly R, Iden T: Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol* 2005, 23:561-566.

26. Rossin EJ, Lage K, Raychaudhuri S, Xavier RJ, Tatar D, Benita Y, Cotsapas C, Daly MJ: Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet* 2011, 7:e1001273.

27. Feldman I, Rzhetsky A, Vitkup D: Network properties of genes harboring inherited disease mutations. *Proc Natl Acad Sci U S A* 2008, 105:4323-4328.

28. Akula N, Baranova A, Seto D, Solka J, Nalls MA, Singleton A, Ferrucci L, Tanaka T, Bandinelli S, Cho YS, Kim YJ, Lee JY, Han BG, McMahon FJ: A network-based approach to prioritize results from genome-wide association studies. *PLoS One* 2011, 6:e24220.

29. Jia P, Wang L, Fanous AH, Pato CN, Edwards TL, Zhao Z: Network-assisted investigation of combined causal signals from genome-wide association studies in schizophrenia. *PLoS Comput Biol* 2012, 8:e1002587.

30. Jia P, Zheng S, Long J, Zheng W, Zhao Z: *dmGWAS*: dense module searching for genome-wide association studies in protein-protein interaction networks. *Bioinformatics* 2011, 27:95-102.

31. A map of human genome variation from population-scale sequencing. *Nature* 2010, 467:1061-1073.

32. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK: Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 2010, 464:768-772.

33. Veyrieras JB, Gaffney DJ, Pickrell JK, Gilad Y, Stephens M, Pritchard JK: Exon-specific QTLs skew the inferred distribution of expression QTLs detected using gene expression array data. *PLoS One* 2012, 7:e30629.

34. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET: Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* 2010, 464:773-777.

35. Fehrmann RS, Jansen RC, Veldink JH, Westra HJ, Arends D, Bonder MJ, Fu J, Deelen P, Groen HJ, Smolonska A, Weersma RK, Hofstra RM, Buurman WA, Rensen S, Wolfs MG, Platteel M, Zhernakova A, Elbers CC, Festen EM, Trynka G, Hofker MH, Saris CG,
Ophoff RA, van den Berg LH, van Heel DA, Wijmenga C, Te Meerman GJ, Franke L: Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS Genet* 2011, 7:e1002197.

36. Chang X, Xu T, Li Y, Wang K: Dynamic modular architecture of protein-protein interaction networks beyond the dichotomy of 'date' and 'party' hubs. *Sci Rep* 2013, 3:1691.

37. Gamazon ER, Zhang W, Konkashbaev A, Duan S, Kistner EO, Nicolae DL, Dolan ME, Cox NJ: SCAN: SNP and copy number annotation. *Bioinformatics* 2010, 26:259-262.

38. Litvin O, Causton HC, Chen BJ, Pe'er D: Modularity and interactions in the genetics of gene expression. *Proc Natl Acad Sci U S A* 2009, 106:6441-6446.

39. Kanehisa M, Goto S: KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000, 28:27-30.

40. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: Gene ontology: tool for the unification of biology. *The Gene Ontology Consortium. Nat Genet* 2000, 25:25-29.

41. Lachmann A, Xu H, Krishnan J, Berger SI, Mazloom AR, Ma'ayan A: ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* 2010, 26:2438-2444.

42. Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE, Stephens M, Gilad Y, Pritchard JK: DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 2012, 482:390-394.

43. Newman ME: Modularity and community structure in networks. *Proc Natl Acad Sci U S A* 2006, 103:8577-8582.

44. Alon U: Network motifs: theory and experimental approaches. *Nat Rev Genet* 2007, 8:450-461.

45. Voight BF, Scott LJ, Steinthorsdottir V, Morris AP, Dina C, Welch RP, Zeggini E, Huth C, Aulchenko YS, Thorleifsson G, McCulloch LJ, Ferreira T, Grallert H, Amin N, Wu G, Willer CJ, Raychaudhuri S, McCarthy SA, Langenberg C, Hofmann OM, Dupuis J, Qi L, Segre AV, van Hoek M, Navarro P, Ardlie K, Balkau B, Bennett AJ, Blagieva R, et al: Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet* 2010, 42:579-589.

46. Brantjes H, Roose J, van De Wetering M, Clevers H: All Tcf HMG box transcription factors interact with Groucho-related co-repressors. *Nucleic Acids Res* 2001, 29:1410-1419.

47. Yeger-Lotem E, Sattath S, Kashtan N, Itzkovitz S, Milo R, Pinter RY, Alon U, Margalit H: Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proc Natl Acad Sci U S A* 2004, 101:5934-5939.

48. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U: Network motifs: simple building blocks of complex networks. *Science* 2002, 298:824-827.

49. Birney E, Andrews TD, Bevan P, Caccamo M, Chen Y, Clarke L, Coates G, Cuff J, Curwen V, Cutts T, Down T, Eyras E, Fernandez-Suarez XM, Gane P, Gibbins B, Gilbert J, Hammond M, Hotz HR, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Keenan
S, Lehvaslaiho H, McVicker G, Melsopp C, Meidl P, Mongin E, Pettett R, et al: An overview of Ensembl. *Genome Res* 2004, 14:925-928.

50. Hoffmann R: A wiki for the life sciences where authorship matters. *Nat Genet* 2008, 40:1047-1051.

51. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A: Human Protein Reference Database--2009 update. *Nucleic Acids Res* 2009, 37:D767-772.

52. Bennett BJ, Farber CR, Orozco L, Kang HM, Ghalalpour A, Siemers N, Neubauer M, Neuhaus I, Yordanova R, Guan B, Truong A, Yang WP, He A, Kayne P, Gargalovic P, Kirchgessner T, Pan C, Castellani LW, Kostem E, Furlotte N, Drake TA, Eskin E, Lusis AJ: A high-resolution association mapping panel for the dissection of complex traits in mice. *Genome Res* 2010, 20:281-290.

53. Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M, Furlotte NA, Eskin E, Nellaker C, Whitley H, Cleak J, Janowitz D, Hernandez-Pliego P, Edwards A, Belgard TG, Oliver PL, McIntyre RE, Bhamra A, Nicod J, Gan X, Yuan W, van der Weyden L, Steward CA, Baia S, Stalker J, Mott R, et al: Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* 2011, 477:289-294.

54. Lefebvre C, Rieckhoff G, Califano A: Reverse-engineering human regulatory networks. *Wiley Interdiscip Rev Syst Biol Med* 2012.

55. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007, 81:559-575.

56. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, Diekhans M, Dreszer TR, Giardine BM, Harte RA, Hillman-Jackson J, Hsu F, Kirkup V, Kuhn RM, Learned K, Li CH, Meyer LR, Pohl A, Raney BJ, Rosenbloom KR, Smith KE, Haussler D, Kent WJ: The UCSC Genome Browser database: update 2011. *Nucleic Acids Res* 2011, 39:D876-882.

57. T.H. Cormen CEL, R.L. Rivest, C. Stein,: *Introduction to Algorithms*. Cambridge: MIT University Press; 2001.

58. Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, Nativ N, Bahir I, Doniger T, Krug H, Sirota-Madi A, Olender T, Golan Y, Stelzer G, Harel A, Lancet D: GeneCards Version 3: the human gene integrator. *Database (Oxford)* 2010, 2010:baq020.

59. Vincent D Blondel J-LG, Renaud Lambiotte and Etienne Lefebvre: Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008, 2008.

60. Bassett DS, Greenfield DL, Meyer-Lindenberg A, Weinberger DR, Moore SW, Bullmore ET: Efficient physical embedding of topologically complex information processing networks in brains and computer circuits. *PLoS Comput Biol* 2010, 6:e1000748.

61. Stokes A, Drozdov I, Guerra E, Ouzounis CA, Warnakulasuriya S, Gleeson MJ, McGurk M, Tavassoli M, Odell EW: Copy number and loss of heterozygosity detected by SNP
array of formalin-fixed tissues using whole-genome amplification. *PLoS One* 2011, 6:e24503.

**Figure legends**

Figure 1: *trans* associations (solid and dashed red straight arrows) on a PPI network: an eSNP that resides within a known exon (left) or TF (middle) maps to the PPI network (right). It defines a source gene (blue s) is associated in *trans* with the levels of a target transcript (green t). Edges of the PPI network edges are denoted in black, and define the shortest path between the exon source and its target (solid red curved arrow). The association between an eSNP within a TF source and its gene target is denoted by a dashed red curved arrow.

Figure 2: **Topological properties on a PPI network vs. exonic source-target association significance:** Averages for (a) distance between source and target (b) source degree and (c) target degree are evaluated across source-target pairs of candidate exon eSNPs at varying association p-value thresholds (+). The average of randomly permuted pairs (dashed horizontal line) is shown for permuted pairs and Spearman’s rank correlation coefficient (denoted r) is listed when significant at p-value < 0.05 (denoted p).

Figure 3: **Examples of TF and exon source-target pairs:** An eSNP (red tick mark) along a source gene (blue circle), either in exon or TF (blue rectangle), is associated (solid/dashed red lines for exon/TF) with levels of transcription of the target gene (green circle). The source and target genes interact via nodes (black circles) and edges (black solid lines) in the PPI network. Each node belongs to a PPI cluster (purple cloud) with a functional annotation. (a) Network motif I1-FFL [44]: TCF7L2 activates MYC (in presence of CTNNB1) but also represses MYC by activating the repressor TLE4 (via an eSNP). (b) The shortest path on the PPI network between PIDD source and its gene target PLK3. Binding sites of TP53 were found in the promoter of PLK3. TP53 is annotated as a zinc ion binding protein. There is a significant correlation between the expression of the source and target genes.

Figure 4: **Summary illustration - two suggested modes of *trans* regulation:** (a) Exon variation often acts by a local *cis* effect, delineating shorter paths of interacting proteins across functional clusters of the PPI network. (b) TF variation frequently acts via a modular regulation mechanism,
with multiple targets that share a function with the TF source. (See Figure 3 legend for further details).

**Additional files**

The following additional data are available with the online version of this paper:

Additional data file 1: contains supplementary text and figures.

Additional data file 2: Table S2: exon-source with their corresponding eSNP targets, for each p-value smaller than $10^{-6}$, where a source-target pair on the PPI network is added, we record the differences between topological properties of random and real pairs using Wilcoxon rank sum test. The table includes for each p-value, the number of unique pairs on the PPI network, the rank sum test p-values and the mean value for each one of the topological properties (distance and source and target degrees) for real and random pairs.

Additional data file 3: Table S5: for all TF and exonic source-target pairs: eSNP rs number, eSNP chromosome, eSNP location, source gene ID, target gene ID, target chromosome and association p-value. For eSNPs in TF we indicate whether they are within an exon.

Additional data file 4: Table S6: functional enrichment analysis of combined sets of: exon sources, exon targets and TF targets (gene sets include only genes that map to an Entrez ID).

Additional data file 5: Table S9: TF units’ functional enrichment (gene sets include only genes that map to an Entrez ID).

Additional data file 6: Table S10: functional enrichment analysis of clusters in the PPI network (gene sets include only genes that map to an Entrez ID).

Additional data file 7: Table S12: functional enrichment of exon paths, between source and target (gene sets include only genes that map to an Entrez ID).

Additional data file 8: Table S13: enriched annotations (minimum 10 genes, FDR<0.05) of genes along real and permuted data shortest paths, and gene names for the six categories that are enriched in real shortest paths.
Figure 1

Target transcript

Coding eSNP

Exon

TF eSNP

trans association

PPI network

$s$

$\text{Exon}$

$\text{TF}$

$t$
Figure 2
Figure 3
Figure 4
Additional files provided with this submission:

Additional file 1: Final_Revision_Supp_GB.docx, 425K
http://genomebiology.com/imedia/5889091809886466/supp1.docx
Additional file 2: Table_S2_Exons_trans_different_pvals.xlsx, 31K
http://genomebiology.com/imedia/1347190089886466/supp2.xlsx
Additional file 3: Table_S5_Exons_TF_source_target_pairs_details.xlsx, 54K
http://genomebiology.com/imedia/7247118459886481/supp3.xlsx
Additional file 4: Table_S6_Enrichment_exon_sources_target_TF_targets.xlsx, 57K
http://genomebiology.com/imedia/1118705439886481/supp4.xlsx
Additional file 5: Table_S9_TF_units.xlsx, 123K
http://genomebiology.com/imedia/9646883479886481/supp5.xlsx
Additional file 6: Table_S10_PPI_enriched_clusters.xlsx, 1884K
http://genomebiology.com/imedia/1037320563988648/supp6.xlsx
Additional file 7: Table_S12_Exons_paths.xlsx, 523K
http://genomebiology.com/imedia/1481492409988648/supp7.xlsx
Additional file 8: Table_S13_annotations_genes_shortest_paths.xlsx, 155K
http://genomebiology.com/imedia/4513508698864822/supp8.xlsx