HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2019 January 16.

Published in final edited form as:

Nat Genet. 2018 August ; 50(8): 1171–1179. doi:10.1038/s41588-018-0160-6.

Deep learning sequence-based *ab initio* prediction of variant effects on expression and disease risk

Jian Zhou[1,2,4], **Chandra L. Theesfeld**[1], **Kevin Yao**[4], **Kathleen M. Chen**[4], **Aaron K. Wong**[4], and **Olga G. Troyanskaya**[1,3,4]

[1]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey, United States of America

[2]Graduate Program in Quantitative and Computational Biology, Princeton University, Princeton, New Jersey, United States of America

[3]Department of Computer Science, Princeton University, Princeton, New Jersey, United States of America

[4]Flatiron Institute, Simons Foundation, New York, New York, United States of America

Abstract

A key challenge for human genetics, precision medicine, and evolutionary biology is deciphering the regulatory code of gene expression, including understanding the transcriptional effects of genome variation. Yet this is extremely difficult due to the enormous scale of the noncoding mutation space. We developed a deep-learning-based framework, ExPecto, that can accurately predict, *ab initio* from DNA sequence, the tissue-specific transcriptional effects of mutations, including rare or never observed. We prioritized causal variants within disease/trait-associated loci from all publicly-available GWAS studies, and experimentally validated predictions for four immune-related diseases. Exploiting the scalability of ExPecto, we characterized the regulatory mutation space for all human Pol II-transcribed genes by *in silico* saturation mutagenesis, profiling >140 million promoter-proximal mutations. This enables probing of evolutionary constraints on gene expression and *ab initio* prediction of mutation disease effect, making ExPecto an end-to-end computational framework for *in silico* prediction of expression and disease risk.

Correspondence to: Olga G. Troyanskaya, ogt@cs.princeton.edu.

URLs

ExPecto web portal for tissue-specific gene expression effect predictions for human mutations, http://hb.flatironinstitute.org/expecto. GTEx Analysis V6 eQTLs, dbGaP Accession phs000424.v6.p1, https://www.gtexportal.org/home/datasets. The 1000 Genomes project human population genomic variants, ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20130502/. GWAS Catalog, version 06192016, https://www.ebi.ac.uk/gwas/. PhyloP scores from 10 primate species, http://hgdownload.cse.ucsc.edu/goldenpath/hg19/phyloP46way/primates/. The source code for running and training ExPecto models is available at https://github.com/FunctionLab/ExPecto.

Author Contributions

JZ and OGT conceived and designed the study. JZ developed the computational methods and performed the analyses. CLT designed and performed experimental studies. KY, KMC, and AKW developed the ExPecto webserver. JZ, CLT, and OGT wrote the manuscript.

Competing interests

The authors declare no competing interests.

Introduction

Sequence-dependent control of gene transcription is at the foundation of the complexity of multicellular organisms. Expression-altering genomic variation can thus have wide impact on human diseases and traits. Empirical observations of expression-genotype association from population genetics studies[1,2] and predictive models based on matched expression and genotype data[3,4] have provided valuable information for the expression effect of common genome variation and their relevance to disease[5]. However, such approaches are generally limited to mutations that are observed frequently and with matched expression observations in ideally the relevant tissue/cell type. Moreover, core to the understanding of the regulatory potential for both common and rare variants is disentangling causality from association and extracting the dependency between sequence and expression effect, which remains as a major challenge.

A quantitative model that accurately predicts expression level *ab initio* from only sequence information will provide a new perspective on expression effects of genomic sequence variations. The computational approach is especially important in human, where limited experiments can be performed directly. Furthermore, *ab initio* sequence-based prediction is capable of extracting causality because of the unidirectional flow of information from sequence change to consequent gene expression change. Moreover, we envision that the potential of estimating effects for all possible variants, including previously unobserved ones, will enable a new framework for the study of sequence evolution and evolutionary constraints on gene expression. This will allow direct prediction of fitness impact due to genomic changes and the resulting expression alteration using only sequence and evolutionary information it contains.

Human gene expression profiles reveal a wide diversity of expression patterns across genes, cell types, and cellular states. Yet our understanding of sequences that activate or repress expression in specific tissues, let alone our ability to quantify the transcriptional modulation strength of a sequence element, is vastly incomplete. Progress in quantitative expression modeling has focused on model organisms with relatively small noncoding regions such as yeast and fly, and in the context of reporter expression prediction in human cell lines[6–10]. As a result, current sequence-based expression prediction models are limited in accuracy or restricted to small subsets of genes, and utilize narrow regulatory regions smaller than 2kb[6–10]. As such, sequence-based prediction of expression in human is still a critical open challenge, and to our knowledge no prior *in vivo* expression prediction model can predict the effect of sequence alterations, especially in tissue-specific context.

Here we describe ExPecto (see **URLs**), a tissue-specific modeling framework for predicting gene expression levels *ab initio* from sequence for over 200 tissues and cell types. The ExPecto framework integrates a deep-learning method with spatial feature transformation and L2-regularized linear models to predict tissue-specific expression from a wide regulatory region of 40kb promoter-proximal sequences. A critical feature of this framework is that it does not use any variant information for training, enabling prediction of expression effect for any variant, even those that are rare or never previously observed.

The resulting ExPecto models make highly accurate cell-type-specific predictions of expression from DNA sequence, as evaluated with known eQTLs and validated causal variants from a massively parallel reporter assay. With this capability, we prioritize putative causal variants associated with human traits and diseases from hundreds of publicly available GWAS studies. We experimentally validated newly predicted putative causal variants for Crohn's disease, ulcerative colitis, Behcet's disease, and HBV infection, demonstrating that these ExPecto-predicted functional SNPs show allele-specific regulatory potential while the GWAS lead SNPs do not.

The scalability of our computational approach allowed us to systematically characterize the predicted expression effect space of potential mutations for each gene, via profiling over 140 million promoter proximal mutations. This enabled us to systematically probe the tissue-specific impact of gene human transcription dysregulation '*in silico*' at a scale not yet possible experimentally, defining the evolutionary constraints on human gene expression. We show that the effects of potential mutations on each gene, which we call the gene's 'variation potential', is indicative of the phenotypic impact of expression-altering mutations.

Integrating expression effect predictions and inferred evolutionary constraints, we propose an end-to-end computational framework for full *in silico* prediction of disease-associated regulatory variation, from sequence to expression effects and subsequent fitness impacts. This framework is complementary to quantitative genetics and experimental approaches at a substantially larger scale and lower cost, including for inferring disease-causal mutations. We demonstrate the far-reaching potential of this approach for interpreting clinically-relevant mutations (even ones not captured by quantitative genetics) by successfully predicting disease risk.

Results

Sequence-based cell type-specific expression prediction

To predict the tissue-specific expression from human promoter-proximal sequences, we built a modular framework (Fig. 1a, **Methods**). First, we used deep learning to generate a repertoire of potential regulatory sequence representations capable of predicting epigenomic effects of any genomic variant from sequence only. This was accomplished with a deep convolutional neural network trained to predict 2,002 different histone mark, transcription factor and DNA accessibility profiles for over 200 tissues and cell types. This substantially extends the epigenomic effect prediction method we described previously[11] with redesigned architecture, expanded feature space and wider sequence context. Second, through a spatial feature transformation approach, the framework integrated predicted sequence-based epigenomic information across 40kb region (Supplementary Fig. 1). Third, tissue-specific regularized linear models used the transformed epigenomic information, centered around the TSS, to predict expression of Pol II transcribed genes in each of the 218 tissues and cell types (one model per tissue for all genes). The resulting ExPecto framework is capable of predicting cell-type specific gene expression and the effects of genomic variants *ab initio*, having never trained on any variant information (neither matched expression or epigenetic data nor any genomic variant data).

ExPecto makes accurate predictions of gene expression levels from sequence, with 0.819 median Spearman correlation with observed expression log RPKM across 218 tissues and cell types (Fig. 1b). This was evaluated on proximal sequences held out during all training of both the regulatory representations and the expression models. When we examined information 'behind' the predictions, we found that the expression models preferentially exploited sequence representations of transcription factors and histone marks (Supplementary Table 1). DNase I sequence features, in contrast, had consistently lower weights ($p = 6.9\times10^{-25}$ two-sided Wilcoxon rank sum test) likely due to the lack of causal dependency information, probably because DNase I hypersensitive sites can be caused by binding proteins of various functions.

Furthermore, in addition to accurately capturing global expression, ExPecto predictions recapitulated the tissue specificity of expression, with expression predictions being significantly more similar to the experimental measurements from the correct cell type than other cell types on holdout sequences (Fig. 1c). As cell type specificity of gene expression in the human body is determined by differential utilization of regulatory DNA sequences, we examined whether the framework learned such cell type regulatory specificity. Indeed, our expression models could automatically learn to preferentially utilize sequence features from the most relevant cell type, even though no explicit tissue labels for these features were used. For example, the top weighted sequence features specific for the liver model corresponded to binding of seven transcription factors (TFs) in HepG2 cells of liver origin. For the breast-mammary gland model – all top five positive features are TFs (ERα and GR) in the breast cancer cell lines T-47D and ECC-1, and for the whole blood model – all top five features are from the blood-derived cell lines and erythroblast cells (Supplementary Table 1).

The ability of ExPecto to predict tissue-specific gene expression from sequence provides the basis for estimating transcriptional effects of genomic variation (Fig. 1a,d). These computational predictions of variant effect do not use any variant-specific information for training and thus can scale to all human population variants and even billions of potential small alterations in the human genome. Thus, in contrast to quantitative genetics approaches, which detect mostly high frequency variants, the ExPecto approach is not biased by allele frequencies and works for both common and rare variants (Supplementary Fig. 2). Therefore, we applied exhaustive *in silico* mutagenesis to probe the effects of over 140 million variants, including all variants around 23,779 TSSs (Fig. 1d), all GWAS loci, and eQTL variants. Below, we demonstrate the potential of ExPecto for accurately identifying causal variants for human traits and diseases and complementing quantitative genetics approaches by avoiding their limitations and predicting rare and unobserved disease-relevant variant effects undetected by quantitative genetics.

Effect of genomic variants on tissue-specific expression

To evaluate ExPecto's predictions of tissue-specific effects of genomic variants on gene expression, we compared with eQTL data from multiple studies. ExPecto's *ab initio* sequence-based prediction is especially useful for prioritizing causal eQTL variants as it is unconfounded by linkage disequilibrium. Thus, even though the majority of eQTL variants are expected to cause no expression effect[12] (of GTEx lead variants, only 3.5%- 11.7% are

estimated to be causal variants, which is <1% of all GTEx eQTL variants[13]), we expect the strong ExPecto predicted effect variants to be highly enriched in *bona fide* causal variants. Among the GTEx-identified eQTLs[2], ExPecto correctly predicted the direction of expression change for 92% of the top 500 strongest effect variants (Supplementary Fig. 3) and provides accurate predictions for tens of thousands of variants (Supplementary Fig. 3). This suggests that a high proportion of eQTLs with strong predicted effects are causal in contrast to the background, as the eQTL effect direction of non-causal SNPs should be independent from the predicted directions. Moreover, ExPecto models for the correct tissue provided more accurate predictions than any other tissue (Supplementary Fig. 3). We also demonstrated accuracy on three other large-scale eQTL studies on brain, primary immune cells and blood respectively[14–16] (Fig. 2a).

In addition, ExPecto can accurately predict causal eQTLs when evaluated with data from *in vitro* massively parallel reporter assays (MPRA) in lymphoblastoid cells[17]. The strongest predicted effect variants from the lymphoblastoid expression model differentially activated transcription, and the model was able to predict expression change directionality with nearly perfect accuracy for top prioritized variants (Supplementary Fig. 4). Notably, the lymphoblastoid model outperformed all other tissue models (AUROC=0.815), again demonstrating the importance of tissue-specific expression modeling in causal-effect predictions.

As expression models can accurately predict causal gene expression effects of SNVs and small INDELs among eQTLs, we examined the expression effect of human population variants across the full range of allele frequencies (16.5 million variants from the 1000 Genomes project) (Supplementary Data 1). In contrast to quantitative genetics approaches, which detect mostly high frequency variants (Supplementary Fig. 2), the ExPecto approach is not biased by allele frequencies and can detect both common and rare variants. Indeed, the ExPecto high expression effect variants have similar MAF distribution to all 1000 Genomes variants (Supplementary Fig. 2). As expected, variants with stronger predicted expression effect are enriched for GTEx eQTLs at all allele frequencies (Supplementary Fig. 5). Thus, sequence-based expression models can be powerful tools in the interpretation of rare functional variants.

Prioritizing and experimental study of causal GWAS variants

We next applied ExPecto's variant expression effect predictions to prioritize causal variants from disease/trait loci of 3,000 GWAS studies[18] (Supplementary Table 2). While GWASs reveal the genetic basis of human diseases and traits by identifying a multitude of associated loci, this approach generally lacks the resolution to pinpoint causal genomic variants due mainly to linkage disequilibrium. Assessing overall performance of ExPecto prioritized variants, we found that loci with the stronger predicted effect variants were significantly more likely to be replicated in a different GWAS study ($p = 6.3\times10^{-189}$, two-sided Wald test with logistic regression, Fig. 2b, see also Supplementary Fig. 6 for analysis using only $p<5\times10^{-8}$ variants). Moreover, the stronger predicted effect GWAS LD variants were more likely to be the exact replicated variant ($p = 5.6\times10^{-14}$, two-sided Wald test with logistic regression). For instance, an earlier Venous Thromboembolism GWAS[19] identified

rs3756008 as the lead causal variant, however, ExPecto-prioritized LD variant rs4253399 near the F11 locus was discovered with a later study using a larger cohort[20] (Supplementary Fig. 7a). Similar examples include variants in autoimmune diseases-associated loci (rs7528684, rs2618476)[21–26] (Supplementary Fig. 7b-c). These results support the potential of using predicted expression effect information to improve identification of causal associated loci from GWAS studies.

We then focused on immunity-related diseases and experimentally measured the expression alteration effects of the top three ExPecto-prioritized SNPs and compared their allele-specific regulatory potential to that of the lead SNPs from the corresponding GWASs (Fig. 3a,c,e). We found that these expression effect-prioritized LD SNPs, while having no prior evidence of functionality, showed transcriptional regulatory activity whereas lead GWAS SNPs did not (as measured by reporter expression assays). (Fig. 3b,d,f). The top ExPecto-prioritized SNP, rs1174815, is predicted to decrease the expression of IRGM, an innate immune response gene significantly associated with Crohn's disease, ulcerative colitis and general inflammatory bowel disease, and indeed we observed significantly decreased reporter expression (Fig. 3a-b, $p = 3\times10^{-6}$). The second top SNP rs147398495, associated with Behcet's disease and near CCR1, a chemokine receptor gene, also significantly changed transcriptional regulatory activity (Fig. 3c-d, $p = 7\times10^{-10}$). For a Chronic HBV infection-associated GWAS locus, our third top SNP, rs381218, was predicted by ExPecto to affect the expression of HLA-DOA, a MHC II gene functional in B-cell lysosomes, and indeed, results in 4-fold change in reporter activity (Fig. 3e-f, $p = 1\times10^{-9}$). In all these cases, none of the lead SNPs in the seven GWASs showed significant differences in transcriptional regulatory activity. Importantly, the directionalities of the expression changes for all three top LD variants were also correctly predicted by ExPecto (Supplementary Table 2). This demonstrates the potential of expression-prediction based causal variant prioritization for identifying disease and trait-associated alleles of true functional impact.

Variation potentials and evolutionary constraints of genes

A substantial gap still exists between predicting expression effect and estimating subsequent phenotypic consequences. The complexity of human as an organism poses significant difficulties in predicting phenotypic or disease consequences of expression alteration where perturbations of different genes elicit distinct consequences. As our model enables exploration of tissue-specific expression effects of genomic sequence variation at an unprecedented scale, essentially providing an '*in silico*' assay of every possible mutation's effects, it enables us to analyze the trace of selection on the regulatory sequences from the space of all potential mutations. We propose that the collective effects of potential mutations on each gene, which we call the gene's 'variation potential' (VP) (Supplementary Fig 8.), is indicative of the phenotypic impact of expression-altering mutations. Furthermore, we found that variation potential is indicative of innate expression properties of genes (e.g. tissue specificity of expression and activation/repression status).

We computed a catalog of predicted effects for more than 140 million mutations that include all possible single nucleotide mutations 1kb upstream and downstream of the TSS for each Pol II-transcribed gene. This identifies over 1.1 million mutations with a strong predicted

expression effect (at high confidence). As expected, mutations with predicted negative effect (mutations predicted to decrease expression) were generally positioned at the immediate upstream of the TSS near −50bp (Supplementary Fig. 9), which is the typical position of core promoter elements[27]. Also reassuringly, the bases with stronger predicted mutation effects showed significantly higher evolutionary constraints both in the modern human population (Supplementary Fig. 10a) and in ancestral evolutionary history (Supplementary Fig. 10b).

We observed that tissue-specific variation potential of a gene is highly predictive of expression properties for that gene (Figure 4). Specifically, we can predict both whether a gene is ubiquitous versus tissue- or condition-specific and whether a gene is active or repressed (Figure 4a,b, Supplementary Fig. 11, Supplementary Data 2). Genes with low VP magnitude are characteristically ubiquitously expressed genes involved in essential cellular processes (e.g. splicing, translation, protein folding, and energy metabolism) (Figure 4b). In contrast, genes with high VP magnitude are tissue-specific (e.g. synaptic transmission genes) or condition-specific (e.g. innate immune response genes) (Supplementary Fig. 11). Among non-ubiquitous genes, VP's directionality (positive/negative cumulative mutation effect) in a given tissue predicts that gene's activation status: negative VP predicts actively expressed genes and positive VP indicates repressed expression in the modeled tissue (Supplementary Fig. 11). For example, synaptic transmission genes have negative VP and are thus predicted to be actively expressed in brain tissue. On the contrary, in non-neuronal tissues. they have positive VP and are thus predicted to be repressed (as in Figure 4a) (Supplementary Data 2). In a given tissue, genes with strong positive predicted VP indeed appear to be repressed, as they are expressed significantly higher in other tissues compared to genes with similar expression level but low VP magnitude (analysis based on GTEx, Supplementary Fig. 12). Thus, variation potentials are not simply reflecting the magnitude of gene expression, but rather distilling expression properties from sequence.

We hypothesize that this connection between variation potentials and expression properties of genes is imposed by evolutionary constraint. Specifically, we propose that genes strongly enriched with mutations of predicted negative effects are under positive evolutionary constraint (i.e. decreasing expression of that gene is deleterious) and vice versa (Figure 4a third panel, c,d, Supplementary Fig. 13, Methods). Supporting this differential evolutionary constraints hypothesis, both variant allele frequencies in human populations and evolutionary conservation evidence supports divergent selection signatures for genes with putative positive and negative constraints ($p < 1.6\times10^{-14}$ in all cases by two-sided Wald test with logistic regression on coefficient of interaction term). Specifically, for putative positive constraint genes, variant sites predicted to decrease expression have lower allele frequency and higher evolutionary conservation, as compared to negative constraint genes, and vice versa (Figure 4d). These divergent selection signatures are not simply driven by gene expression levels, because randomly selecting genes with matching expression levels do not show differential selection (Supplementary Fig. 14). Moreover, genes with stronger inferred evolutionary constraints are significantly more enriched in GWAS disease genes ($p=6.1\times10^{-53}$ two-sided Wald test with logistic regression), indicating that changes in expression of these genes are more likely to lead to adverse consequences (Supplementary Fig. 15).

Ab initio inference of disease risk alleles

With the inference of putative evolutionary constraints, we have collected key components for a regulatory disease mutation analysis framework that addresses both the impact of a variant on gene expression and the fitness impact of expression alterations (Figure 4a). Recognizing regulatory disease mutations is very challenging because both of these problems are difficult to address experimentally[5]. Our proposed computational sequence-based approach addresses these problems genome-wide, even for mutations that have not been previously observed experimentally.

We thus assess the ability of our approach to predict disease risk *ab initio* from sequence. At the individual variant level, we predict whether a specific sequence alteration is likely to be deleterious or protective via integrating the expression effect and variation-potential-based constraint directionality through the constraint violation score (Methods). For example, if a variant causes a positively constrained highly expressed gene to substantially decrease expression, it is likely to be deleterious (Figure 4a fourth panel). We then evaluated our predictions on the curated pathogenic regulatory mutations in human gene mutation database (HGMD) and the prioritized putative causal LD SNPs derived from GWAS catalog[18,28].

Most strong ExPecto predicted effect mutations from HGMD are predicted to decrease expression (Fig. 5a), and correspondingly, all of them are near genes with putative positive evolutionary constraints as would be expected from our findings above. Positive constraints are also consistent with the current understandings of these diseases as caused by deficiency of certain proteins, such as coagulation factor genes *F7* and *F9* for factor VII deficiency[29] and Haemophilia B respectively; *PROC* and *PROS1* genes for blood clotting disorders caused by protein C deficiency and protein S deficiency; *APOA1* and *LDLR* for hypolipoproteinemia and hypercholesterolemia caused by apolipoprotein deficiency and decreased receptor-mediated endocytosis of LDL cholesterol deficiency respectively. *UNC13D*, essential for intracellular trafficking and exocytosis of lytic granules is associated with hemophagocytic lymphohistiocytosis type 3[30]. Decreased expression of the *BTK* gene, an essential gene for B cell development and maturation, causes agammaglobulinemia[31]. *HNF1A* mutation is well known as a major cause for maturity onset of diabetes (MODY)[32] and considered important in beta-cell differentiation[33].

Only one HGMD disease mutation was predicted to strongly increase transcriptional activity and it is near a gene with putative negative constraints in all tissues, *TERT* (Fig. 5a). *TERT* encodes telomerase reverse transcriptase, and overexpression of *TERT* thus supports unconstrained proliferation. Indeed, mutations in the *TERT* promoter were found to be highly recurrent in 71% of melanoma samples[34], as well as in many other cancer types including bladder and central nervous system cancers[35], and many of these mutations generate new ETS binding sites and increase transcriptional activity in reporter assays, consistent with our predictions. Note that even though HGMD disease mutations are known to be deleterious, these results demonstrate that ExPecto can correctly predict the disease allele versus the non-disease allele without any prior knowledge of disease association.

To assess the potential for ExPecto to predict disease risk for relatively common variants in the population, we evaluated whether constraint violation scores were predictive for GWAS

risk loci. Positive violation scores suggest the alternative allele is likely more deleterious while a negative violation score suggests the reference allele is likely more deleterious. This GWAS evaluation standard directly includes both deleterious and protective variants (risk alleles are reference alleles for 37 loci, alternative alleles for 63 loci). Our approach is significantly predictive (p=0.002, Wilcoxon rank sum test, AUC = 0.67, Fig. 5b, Supplementary Data 3) of the known risk alleles detected from GWAS studies. This evaluation thus demonstrates that predicted effects that violate inferred constraints are predictive of risk alleles in GWAS, indicating that ExPecto can predict which allele is deleterious or protective without any prior variant-disease association information. Together our results suggest this approach as a promising direction for large-scale prediction of disease risks, which will be especially useful for interpreting the enormous amount of potential disease mutations for which there is with little or no prior knowledge.

Discussion

ExPecto thus provides robust and scalable *ab initio*, sequence-based prediction of variant effects, enabling genome-wide studies of human genomic variation and disease. We demonstrated that computational prediction of causal variants in trait-associated loci, including eQTLs and GWAS disease-associated loci, is capable of identifying causal variants, and this can be routinely performed at whole-genome level. Our approach can potentially be further combined with statistical models (reviewed in Pasaniuc & Price[36]) for further improvement on causal variant identification. Moreover, as the method is equally applicable to rare or common variants, it allows wider application to mutation space outside the power of traditional quantitative genetics.

The ExPecto expression models also make possible the probing of variation potentials and evolutionary constraints through *in silico* mutagenesis analysis. We expect these predictions of evolutionary constraints on gene expression to be especially valuable for understanding human disease by identifying the fitness consequence of expression alteration that are otherwise very difficult to study in human. We propose using variation potentials as a proxy for evolutionary constraints, and we show that with only sequence information, it is possible to predict the disease risk allele of HGMD regulatory mutations with very high accuracy and to identify GWAS risk alleles.

The prediction of expression effects and evolutionary constraints thus provides an end-to-end computational framework for regulatory disease mutation analysis. While the ExPecto models are accurate, scalable, and robust, there is still potential for future improvement in both accuracy and coverage of predictable variants. More comprehensive chromatin profiling, especially of chromatin marks and transcription factor binding, additional data capturing ultra-distal regulatory sequences, especially those mediated by long-range interactions, and epigenetic inheritance mechanisms that affect expression independently from sequence, such as imprinting via DNA-methylation could be incorporated into the ExPecto framework and are likely important for improvement of sequence-based expression models.

In the long run, we expect that sequence-based expression analysis will become an important part of research and clinical studies of whole genome sequences, especially for identifying clinically relevant non-coding variants and expression perturbations. Such analyses could, in the future, be used for grouping patients in drug and other treatment trials, disease subtyping, and eventually personalized treatment. At the same time, we expect that tapping into human genome evolution information, allowed by sequence-based expression modeling, will provide valuable insights required for the comprehensive understanding of the healthy and disease modes of human gene expression.

Online Methods

ExPecto framework architecture

The ExPecto sequence-based expression prediction framework includes three components that act sequentially (Figure 1a). First, a deep neural network epigenomic effects model scans the long input sequence with a moving window and outputs predicted probabilities for histone marks transcription factors, and DNase hypersensitivity profiles at each spatial position. Then a series of spatial transformation functions summarize each predicted spatial pattern of chromatin profiles to generate a reduced set of spatially transformed features (Supplementary Figure 1). Last, the spatially-transformed features are used to make tissue-specific predictions of expression for every gene by regularized linear models.

The first component of ExPecto uses a deep convolutional neural network to transform genomic sequences to epigenomic features. Our approach generates a cell-type specific model for 2,002 genome-wide histone marks, transcription factor binding and chromatin accessibility profiles (based on training data from ENCODE and Roadmap Epigenomics projects[45,46], Supplementary Data 4), substantially extending the deep learning-based method that we described previously[11] with redesigned architecture and more features. Specifically, the model architecture was extended to double the number of convolution layers for increased model depth, broader genomic context was incorporated with increased window size (2000bp), and the new model was trained to predict twice as many regulatory features for over 200 cell types (Supplementary Note). Critically, this deep learning model does not use any mutation data for training. The deep convolutional neural network model predicts epigenomic features of a 200bp region, while also using the 1800bp surrounding context sequence. For each Pol II-transcribed gene, surrounding its representative transcriptional start site (TSS, see the ‘Identification of representative transcription start sites’ section below), the deep convolutional neural network model scans the genomic sequence between +20kb upstream and −20kb downstream to predict spatial chromatin organization patterns using a moving window with 200bp step size, yielding 200 spatial bins with a total number of 400400 features.

The second component of ExPecto is the spatial transformation module that reduces the dimensionality of the learning problem by generating spatially-transformed features (Supplementary Fig. 1). The spatial transformation module reduces the input dimensionality with ten exponential functions weighting upstream and downstream regions separately, with weights based on relative distance to TSS (transformed features with higher decay rate are more concentrated near TSSs). This effectively reduces the number of features 20 fold to

20020. The exponential functions were prespecified (based on empirical selection) and not learned during training. This spatial feature transformation followed by learning linear model retains the flexibility of learning spatial patterns, which is equivalent to learning a smooth nonlinear spatial pattern function f constrained to the space of linear combinations of basis functions corresponding to the feature transformations.

Finally, to make tissue-specific expression predictions, spatially-transformed features are used to predict gene expression levels for each tissue (quantified by log RPKM) with L2-regularized linear regression models fitted by gradient boosting algorithm[47,48]. Specifically, the full models including both spatial transformation and linear models are specified as below.

$$\text{expression} = \sum_{d \in D} \sum_{i} p_{id} \left| \sum_{k} 1(t_d < 0)\, \beta^{\text{up}}{}_{ik} e^{-a_k \left\lfloor \frac{|t_d|}{200\text{bp}} \right\rfloor} + \sum_{k} 1(t_d > 0) \beta^{\text{down}}{}_{ik} e^{-a_k \left\lfloor \frac{|t_d|}{200\text{bp}} \right\rfloor} \right|$$

where p_{id} is the predicted probabilities for chromatin feature i at region d relative to the TSS, and D represents the set of 200 × 200bp spatial bins within 20kb of the TSS. 1 represents the indicator function which equals one when the specified condition is satisfied and zero otherwise. t_d represents the mean distance of region d to the TSS. For example, the −200bp to 0bp bin has a distance of −100 bp and the −400bp to −200bp bin has a distance of −300bp. $\beta^{\text{up}}{}_{ik}$ and $\beta^{\text{down}}{}_{ik}$ are the learned expression model coefficients of chromatin feature i and exponential function index k for upstream and downstream regions respectively. The decay constant for exponential function k is indicated by a_k, where $a = \{0.01, \quad 0.02, \quad 0.05, \quad 0.1, \quad 0.2\}$. Note that model coefficients $\beta^{\text{up}}{}_{ik}$ and $\beta^{\text{down}}{}_{ik}$ are shared across spatial bins indexed by d due to spatial transformation, thus significantly decreasing the number of fitted parameters (by 20 fold) and reducing ovefitting. All hyperparameters of ExPecto are chosen by empirical evaluations, including the number and values of exponential terms, model design, and window sizes, while all the neural network model weights and linear model coefficients are learned from data. The +/- 20kbp (40Kbp) window size around the TSS maximizes ExPecto accuracy. While smaller windows decrease prediction performance, increasing the window size to 50kb, 100kb or even 200kb gives negligible performance gain (Supplementary Fig. 16).

Application of ExPecto for sequence-based gene expression level prediction across tissues

While ExPecto models can be trained on any expression profile, here we used 218 tissue expression profiles from GTEx, Roadmap epigenomics and ENCODE projects. A pseudocount was added before log transformation (0.01, except for 0.0001 for GTEx tissues (which were averaged across individuals) due to high coverage from pooling multiple samples). The linear expression models were trained with L2 regularization parameter

lambda=100, shrinkage parameter eta=0.01 and basescore=2 for 100 rounds. The training and prediction time of ExPecto is detailed in Supplementary Table 3.

The gene-wise expression prediction performance was evaluated on whole chromosome holdout of chr8 (990 genes), which was withheld at all stages of the ExPecto training (sequences were not used for training either the linear expression models or the neural network regulatory effects model). We chose a whole chromosome holdout to provide a more conservative evaluation and minimize overlap of regulatory regions. To further minimize the possibility of overfitting through homology, we removed all chr8 genes with paralogs on other chromosomes, and this does not negatively affect performance (Spearman correlation 0.819 for all 990 chr8 genes, 0.821 for after removal of 184 paralogous genes).

For interpretation of tissue-specific signals captured by the models, the most informative cell type-specific sequence features from expression models were extracted as follows:

$$\frac{1}{n_k}\sum_k \beta_{ik}^{c} - \frac{1}{n_k n_{\text{cells}}}\sum_{c'}\sum_k \beta_{ik}^{c'}$$

β_{ik}^{c} represent in the coefficient for chromatin feature i, exponential function k in cell type/tissue c. n_k represents the number of exponential functions ($n_k = 10$ in this case, considering both upstream and downstream coefficients), and n_{cells} represents the number of all cell type/tissue models. c' is index for cell type/tissues. To enable comparison across features from different datasets, we used models retrained with a uniform pseudocount of 0.0001 for all tissue or cell types. The top features with higher than tissue-average coefficients were then selected.

Variant expression effect prediction

Gene expression effect is naturally estimated by the difference of predicted expression levels for reference and alternative allele, which is measured by the predicted log fold change. As the expression effect models are linear combinations of regulatory feature predictions, expression effect prediction computation can be simplified to a function of the variant chromatin effects p and distance to TSS t

$$\text{effect}(p,t) = \sum_{\delta \in \Delta}\sum_i \left(p_{i\delta}^{\text{alt}} - p_{i\delta}^{\text{ref}}\right)\left[\sum_k 1(t<0)\ \beta^{\text{up}}{}_{ik} e^{-a_k\left\lfloor\frac{|t+\delta|}{200\text{bp}}\right\rfloor} + \sum_k 1(t>0)\beta^{\text{down}}{}_{ik} e^{-a_k\left\lfloor\frac{|t+\delta|}{200\text{bp}}\right\rfloor}\right]$$

where $p_{i\delta}^{\text{ref}}$ and $p_{i\delta}^{\text{alt}}$ are the predicted probabilities for chromatin feature i with reference allele or alternative allele at position δ relative to the variant position, $\beta^{\text{up}}{}_{ik}$ and $\beta^{\text{down}}{}_{ik}$ are the expression model coefficients of chromatin feature i and exponential function index k for upstream and downstream variants, respectively. The decay constant for exponential function k is indicated by a_k and the distance to TSS is indicated by t. Notably the predicted variant regulatory effect includes both effects at the variant site and at adjacent positions (as

long as the variant is within range of 2000bp context sequence window for that region), thus the variant expression effect considers regulatory effects in 9 positions specified by $\Delta = \{0\text{bp}, -200\text{bp}, -400\text{bp}, -600\text{bp}, -800\text{bp}, +200\text{bp}, +400\text{bp}, +600\text{bp}, +800\text{bp}\}$.

For small INDELs, we compensate or truncate the alternative allele sequence equally on both sides to total 2000bp.

Evaluation of ExPecto tissue-specific expression effect predictions

The GTEx v6 eQTLs, the 1000 Genomes phase 3 variants, and GWAS Catalog data were downloaded from the websites (see **URLs**). HGMD regulatory mutations were from HGMD professional version 2014.4 and filtered to category DM, which represents “disease-causing/ pathological” mutations reported to be disease causing in the original literature.

The *in vitro* reporter assay eQTL effects were predicted with modifications for adapting to the difference between *in vitro* reporter assay and *in vivo* expression, as only a short element is cloned to a fixed position upstream of a reporter gene in reporter assay. Specifically, we used regulatory effect models trained on 230bp input window instead of 2000bp, and only the in-place chromatin effect but not effect on adjacent regions were computed, as these sequences were not cloned to the reporter vector. The position relative to TSS is fixed at −100bp.

We evaluated ExPecto prioritization of GWAS loci by examining their replication of prioritized loci across studies. In Supplementary Fig. 17 we compare ExPecto to DeepSEA[11] (which predicts just the epigenomic component of the variant effect) in this task. ExPecto predicts variant effects on gene expression while DeepSEA can identify variant effects that do not lead to significant expression change.

Computation of GWAS linkage disequilibrium SNPs

To systematically screen for SNPs in linkage disequilibrium with the reported GWAS lead SNPs from GWAS catalog, we first computed linkage disequilibrium for all 88 million variants in 1000 Genomes phase 3 genotype data (see **URLs**), which includes >99% of SNP variants with a frequency of >1% for a variety of ancestries[49]. Linkage disequilibrium between SNPs in five populations EAS, SAS, AMN, AFR and EUR were computed with PLINK v1.90b. In total, we found 390,085 variants in LD r^2 >0.75 with 15571 distinct GWAS catalog reported variants. We then used ExPecto to systematically predict expression effects for all LD variants to their nearest TSS.

Experimental validation of prioritized candidate GWAS causal SNPs

We experimentally validated the top three ExPecto-prioritized variants that had no prior evidence for functionality, and which were associated with four immune-related diseases in seven GWAS studies. Specifically, we used a luciferase assay to compare the ability of risk versus non-risk alleles to drive expression for the above ExPecto prioritized variants and the seven lead SNPs reported by the corresponding GWAS studies.

All genomic sequences were retrieved from hg19 human genome assembly. For each risk allele (reference or alternative), Genewiz synthesized a 260 nucleotide fragment: 230 was

human genomic sequence and 15 nucleotides matched each flank of the plasmid cloning sites (Supplementary Table 4). Each fragment was cut with KpnI and BglII and cloned into pGL4.23 (minP firefly luciferase vector) (Promega) cut with the same enzymes. For luciferase assay, 2×10^4 BE(2)-C cells were plated in 96-well plates, and 24 hours later transfected with Lipofectamine 3000 (L3000-015, Thermofisher Scientific) and 75ng of variant-containing pGL4.23 plasmid (Supplementary Table 4), and 4ng of pNL3.1 NanoLuc plasmid, for normalization of transfection conditions. 42 hours after transfection, luminescence was detected with the Promega NanoGlo Dual Luciferase assay system (N1630) and BioTek Synergy plate reader. Four to six replicates per variant were tested in each experiment. The experiment was performed 2-5 times for the variants. For each sequence tested, the ratio of firefly (variant) luminescence to NanoLuc (transfection control) luminescence was calculated and then normalized to empty vector. Statistics were calculated by combining fold over EV values from each biological replicate.

Systematic profiling of variation potentials and evolutionary constraints by *in silico* mutagenesis

We systematically predicted all (over 140 million) possible single nucleotide substitution variations across all human promoters within 1kb of the representative TSS on both sides. Gene-wise variation potentials were summarized by two measures: directionality, which is computed as the sum of predicted log fold-changes for all mutations per gene, and magnitude, which is computed as the sum of all absolute predicted log fold-changes. We find that genes with negative variation potential directionality (i.e. mutations tend to cause a decrease in tissue-specific expression) are actively expressed in the modeled tissue (see Figure 4b, Supplementary Fig. 11). We infer that these genes are under positive evolutionary constraint, and thus are vulnerable to inactivating mutations. On the other hand, we find that expression of genes with positive variation potential (i.e. mutations cause an increase in tissue-specific expression) is repressed in the modeled tissue (see Figure 4b, Supplementary Fig. 11). We infer that these genes are under negative evolutionary constraint, and thus are vulnerable to activating mutations. Note that evolutionary constraints cannot simply be inferred from gene expression levels (Supplementary Fig. 14).

We use the directionality score to measure the tendency of the potential mutation effect to be biased toward positive or negative, which we propose indicates negative and positive evolutionary constraints, respectively (Supplementary Figure 13). The distribution of mean predicted mutation effects across genes was modeled as a mixture of a Gaussian null distribution, a positive constraint component and a negative constraint component. While the true null distribution is unknown, a conservative estimate of empirical null distribution can be obtained by assuming the other components are only observed at the two tails and estimating a Gaussian distribution using central quantiles of the data, similar to the idea used for measuring local FDR[50]. We fit the empirical null distribution with the truncated Gaussian MLE method implemented in the locfdr R package[50]. With empirical null distribution estimation and density estimation of overall distribution of gene-wise average predicted effects, we can then compute probabilities of genes belonging to the positive or negative constraint components. Probability > 0.5 for each component is used for assigning genes to putative positive or negative evolutionarily constrained genes.

Analysis of conservation and allele frequency for variants

For estimating recent divergence in the modern human population, we used allele frequencies among the 1000 Genomes project phase 3 individuals. For estimating divergence from human-chimpanzee common ancestor, the proportion of divergent sites was computed from the high confidence divergence sites from[51]. For estimating divergence among 10 primate species (including humans), we computed proportion of accelerated evolution sites based on primates phylop scores (see **URLs**). Accelerated evolution sites were decided with the threshold of phyloP < −2.3 which corresponds to p-value<0.005 for accelerated evolution.

Ab initio inference of disease risk alleles

We used the ExPecto-prioritized GWAS LD variants (as described above) for risk allele prediction. We included GWAS LD variants with $r^2 > 0.75$ in a matched 1000 Genomes population, and variants for which the risk allele is ambiguous (different GWAS studies pointing to conflicting risk alleles) were excluded. Only GWAS studies for disease or disease related traits were included. The constraint violation score was computed as the product of the predicted variant effect of the prioritized LD variant and the variation potential directionality score of the nearest TSS. The median constraint violation score across all non-cancer tissue or cell types for each variant was used.

Identification of representative transcription start sites

Most expression profiling datasets were quantified to gene level, as it is often challenging to achieve accurate quantification of TSS expression level from short read sequencing. Even though training expression model should ideally utilize TSS-specific expression quantification, gene level expression measured by RNA-seq or microarray are usually a good approximation of transcription level from the representative TSS of each gene[52,53], and are usually measured with higher sequencing depth. We determined representative TSS for each Pol II transcribed gene based on quantification of aggregated cap analysis of gene expression(CAGE) reads in the FANTOM5 project[54]. Specifically, a CAGE peak is associated to a GENCODE gene if it is within 1000bp from a GENCODE v24 annotated transcription start site (lifted to GRCh37 coordinates). Peaks within 1000bp to rRNA, snRNA, snoRNA or tRNA genes were removed to avoid confusion. Next, we selected the most abundant CAGE peak for each gene, and took the TSS position reported for the CAGE peak as the selected representative TSS for the gene. For genes with no CAGE peaks assigned, we kept the annotated gene start position as the representative TSS. The selected TSSs showed significantly higher conservation level compared to the annotated gene start positions ($p = 5.7\times10^{-8}$, Supplementary Fig. 18).

Statistical analysis

All details of the statistical tests are specified in the associated text or figure legend. Association between two variables is tested via linear regression (or logistic regression if one the variable is categorical) with the null hypothesis that the slope coefficient is zero. For comparing evolution and population genetics signatures between putative positive and putative negative constraint genes, we test the null hypotheses that the coefficient of the

interaction term is zero in a logistic regression model specified by the formula $y \sim e + t + e \cdot t$,. where y is a binary variable representing evolutionary or population genetic information about a site, e represents the ExPecto predicted expression effect, t represents the inferred putative constraint type, and $e \cdot t$ represents the interaction term.

Life Sciences Reporting Summary

Further information on experimental design is available in the Life Sciences Reporting Summary.

Data and code availability

The data supporting the findings of the study are available within the paper and its supplementary information files. The source code is available (see **URLs**).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors acknowledge all Troyanskaya lab members for helpful discussions. This work is supported by NIH grants R01HG005998, U54HL117798 and R01GM071966, HHS grant HHSN272201000054C, and the Simons Foundation grant 395506. The authors are pleased to acknowledge that the work reported on in this paper was substantially performed at the TIGRESS high performance computer center at Princeton University which is jointly supported by the Princeton Institute for Computational Science and Engineering and the Princeton University Office of Information Technology's Research Computing department. OGT is a CIFAR fellow.

References

1. Pickrell JK, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010; 464:768–772. [PubMed: 20220758]
2. GTEx Consortium, T. Gte. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013; 45:580–5. [PubMed: 23715323]
3. Gamazon ER, et al. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. 2015; 47:1091–1098. [PubMed: 26258848]
4. Li X, et al. The impact of rare variation on gene expression across tissues. bioRxiv. 2016; doi: 10.1101/074443
5. Edwards SL, Beesley J, French JD, Dunning M. Beyond GWASs: Illuminating the dark road from association to function. American Journal of Human Genetics. 2013; 93:779–797. [PubMed: 24210251]
6. Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U. Predicting expression patterns from regulatory sequence in Drosophila segmentation. Nature. 2008; 451:535–540. [PubMed: 18172436]
7. Beer MA, Tavazoie S. Predicting gene expression from sequence. TL - 117. Cell. 2004; 117VN: 185–198.
8. Yuan Y, Guo L, Shen L, Liu JS. Predicting gene expression from sequence: A reexamination. PLoS Comput Biol. 2007; 3:2391–2397.
9. Bussemaker HJ, Li H, Siggia ED. Regulatory element detection using correlation with expression. Nat Genet. 2001; 27:167–71. [PubMed: 11175784]
10. Kreimer A, et al. Predicting gene expression in massively parallel reporter assays: A comparative study. Human Mutation. 2017; doi: 10.1002/humu.23197

11. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015; 12:931–4. [PubMed: 26301843]
12. Aguet F, et al. Local genetic effects on gene expression across 44 human tissues. bioRxiv. 2016; doi: 10.1101/074450
13. Aguet F, et al. Genetic effects on gene expression across human tissues. Nature. 2017; 550:204–213. [PubMed: 29022597]
14. Westra H-J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013; 45:1238–43. [PubMed: 24013639]
15. Ramasamy A, et al. Genetic variability in the regulation of gene expression in ten regions of the human brain. Nat Neurosci. 2014; 17:1418–1428. [PubMed: 25174004]
16. Fairfax BP, et al. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. Nat Genet. 2012; 44:502–10. [PubMed: 22446964]
17. Tewhey R, et al. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. Cell. 2016; 165:1519–1529. [PubMed: 27259153]
18. MacArthur J, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017; 45:D896–D901. [PubMed: 27899670]
19. Germain M, et al. Genetics of Venous thrombosis: Insights from a new genome wide association study. PLoS One. 2011; 6
20. Tang W, et al. A Genome-Wide Association Study for Venous Thromboembolism: The Extended Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium. Genet Epidemiol. 2013; 37:512–521. [PubMed: 23650146]
21. Plagnol V, et al. Genome-wide association analysis of autoantibody positivity in type 1 diabetes cases. PLoS Genet. 2011; 7
22. Chu X, et al. A genome-wide association study identifies two new risk loci for Graves' disease. Nat Genet. 2011; 43:897–901. [PubMed: 21841780]
23. Sawcer S, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature. 2012; 476:214–219.
24. Graham RR, et al. Genetic variants near TNFAIP3 on 6q23 are associated with systemic lupus erythematosus. Nat Genet. 2008; 40:1059–61. [PubMed: 19165918]
25. Bentham J, et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. Nat Genet. 2015; 47:1457–1464. [PubMed: 26502338]
26. Lee Y-C, et al. Two new susceptibility loci for Kawasaki disease identified through genome-wide association analysis. Nat Genet. 2012; 44:522–5. [PubMed: 22446961]
27. Xi H, et al. Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1. Genome Res. 2007; 17:798–806. [PubMed: 17567998]
28. Stenson PD, et al. The Human Gene Mutation Database: 2008 update. Genome Med. 2009; 1:13. [PubMed: 19348700]
29. Nagaizumi K, et al. Two double heterozygous mutations in the F7 gene show different manifestations. Br J Haematol. 2002; 119:1052–1058. [PubMed: 12472587]
30. Feldmann J, et al. Munc13-4 Is Essential for Cytolytic Granules Fusion and Is Mutated in a Form of Familial Hemophagocytic Lymphohistiocytosis (FHL3). Cell. 2003; 115:461–473. [PubMed: 14622600]
31. Ng Y-S, Wardemann H, Chelnis J, Cunningham-Rundles C, Meffre E. Bruton's tyrosine kinase is essential for human B cell tolerance. J Exp Med. 2004; 200:927–34. [PubMed: 15466623]
32. Yamagata K, et al. Mutations in the hepatocyte nuclear factor-4alpha gene in maturity-onset diabetes of the young (MODY1). Nature. 1996; 384:458–460. [PubMed: 8945471]
33. Servitja J-M, et al. Hnf1alpha (MODY3) controls tissue-specific transcriptional programs and exerts opposed effects on cell growth in pancreatic islets and liver. Mol Cell Biol. 2009; 29:2945–59. [PubMed: 19289501]
34. Huang FW, et al. Highly recurrent TERT promoter mutations in human melanoma. Science. 2013; 339:957–9. [PubMed: 23348506]
35. Vinagre J, et al. Frequency of TERT promoter mutations in human cancers. Nat Commun. 2013; 4

36. Pasaniuc B, Price AL. Dissecting the genetics of complex traits using summary association statistics. Nature Reviews Genetics. 2017; 18:117–127.
37. Parkes M, et al. Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nat Genet. 2007; 39:830–2. [PubMed: 17554261]
38. Wellcome T, Case T, Consortium, C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447:661–78. [PubMed: 17554300]
39. Barrett JC, et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet. 2008; 40:955–962. [PubMed: 18587394]
40. Franke A, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet. 2010; 42:1118–25. [PubMed: 21102463]
41. Jostins L, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012; 491:119–24. [PubMed: 23128233]
42. Liu JZ, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet. 2015; 47:979–989. [PubMed: 26192919]
43. Kirino Y, et al. Genome-wide association analysis identifies new susceptibility loci for Behçet's disease and epistasis between HLA-B*51 and ERAP1. Nat Genet. 2013; 45:202–7. [PubMed: 23291587]
44. Jiang DK, et al. Genetic Variants in Five Novel Loci Including CFB and CD40 Predispose to Chronic Hepatitis B. Hepatology. 2015; 62:118–128. [PubMed: 25802187]
45. de Souza N. The ENCODE project. Nat Methods. 2012; 9:1046–1046. [PubMed: 23281567]
46. Bernstein BE, et al. The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol. 2010; 28:1045–8. [PubMed: 20944595]
47. Chen T, Guestrin C. XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16. 2016; :785–794. DOI: 10.1145/2939672.2939785
48. Bühlmann P. Boosting for high-dimensional linear models. Ann Stat. 2006; 34:559–583.
49. 1000 Genomes Project Consortium. et al. A global reference for human genetic variation. Nature. 2015; 526:68–74. [PubMed: 26432245]
50. Efron B. Size, power and false discovery rates. Ann Stat. 2007; 35:1351–1377.
51. Kircher M, et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014; 46:310–315. [PubMed: 24487276]
52. Gonzalez-Porta M, Frankish A, Rung J, Harrow J, Brazma A. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome Biol. 2013; 14:R70. [PubMed: 23815980]
53. Uhlen M, et al. Tissue-based map of the human proteome. Science (80-). 2015; 347:1260419–1260419.
54. Forrest ARR, et al. A promoter-level mammalian expression atlas. Nature. 2014; 507:462–470. [PubMed: 24670764]

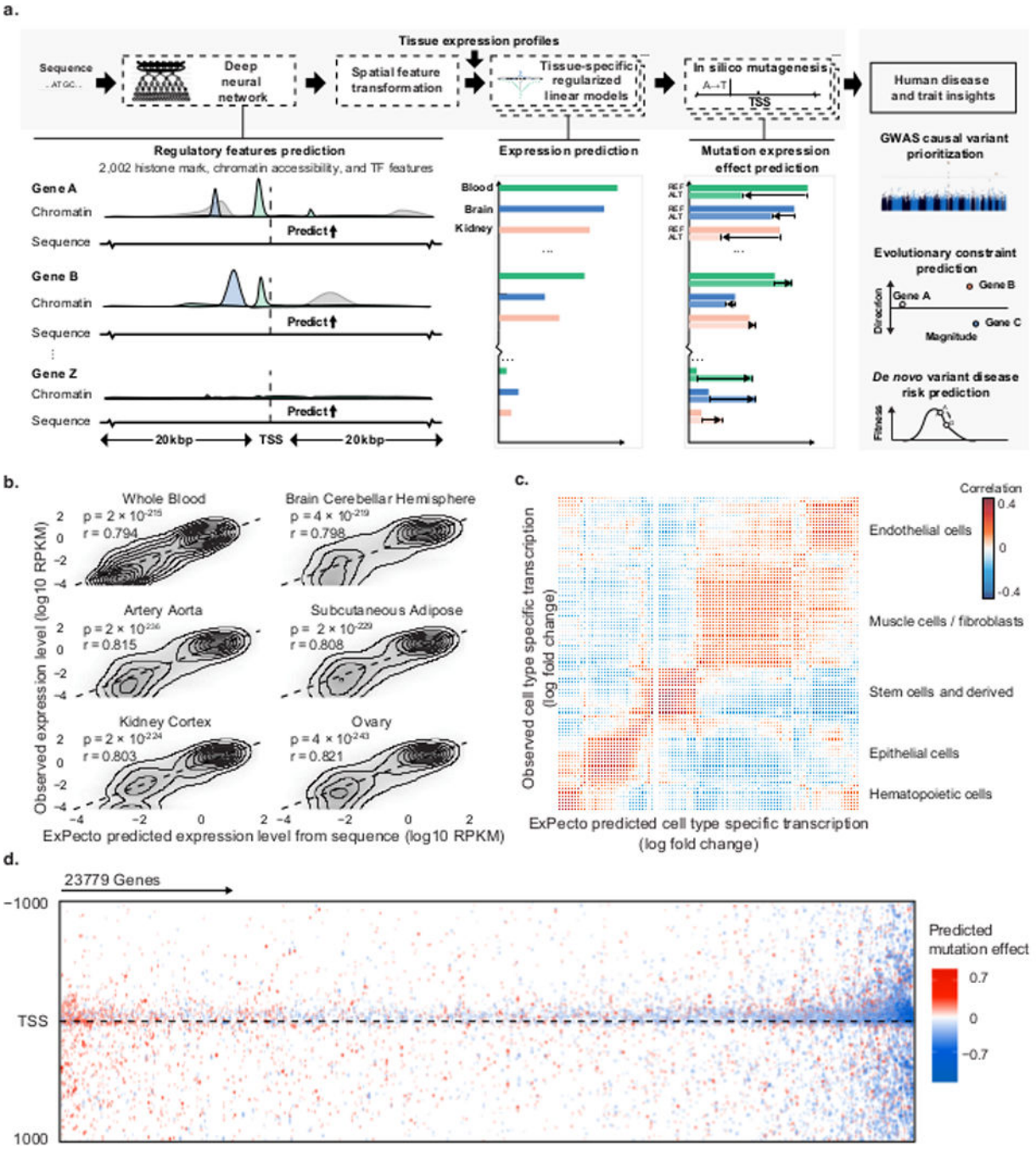

Figure 1. Deep learning-based sequence model accurately predicts cell type-specific gene expression

a) Schematic overview of the ExPecto sequence-based gene expression prediction framework. The predictive model contains three components, a deep convolutional neural network trained on chromatin profiling data that converts sequence to regulatory features, a spatial feature transformation module, and a linear model that predicts gene expression from transformed nonlinear regulatory representations.

b) Sequence-based gene expression predictions on holdout genes are highly correlated with RNA-seq observations. Predicted log RPKMs on 990 genes from the holdout chromosome

chr8 (x-axis) were compared with experimentally measured log RPKMs (y-axis) in each of the six example tissues. Spearman correlations between predicted and observed values are shown.

c) Cell type-specific expression models capture transcription tissue-specificity. The heatmap shows, on holdout genes, correlations between cell type specific expression profiles measured by log fold change over cell-type-average and the sequence-based predicted log fold changes.

d) Predicted mutation effects from *in silico* mutagenesis of promoter-proximal regions of 23,779 genes showed substantial variation, as indicated by color. The predicted effects for different variants at the same position were averaged. Genes were sorted by gene-wise average predicted mutation effects. Only positions with larger than 0.5 average absolute log fold change were shown. −1000 is upstream of the TSS and +1000 is downstream (by base pair). The whole blood model predictions are shown.

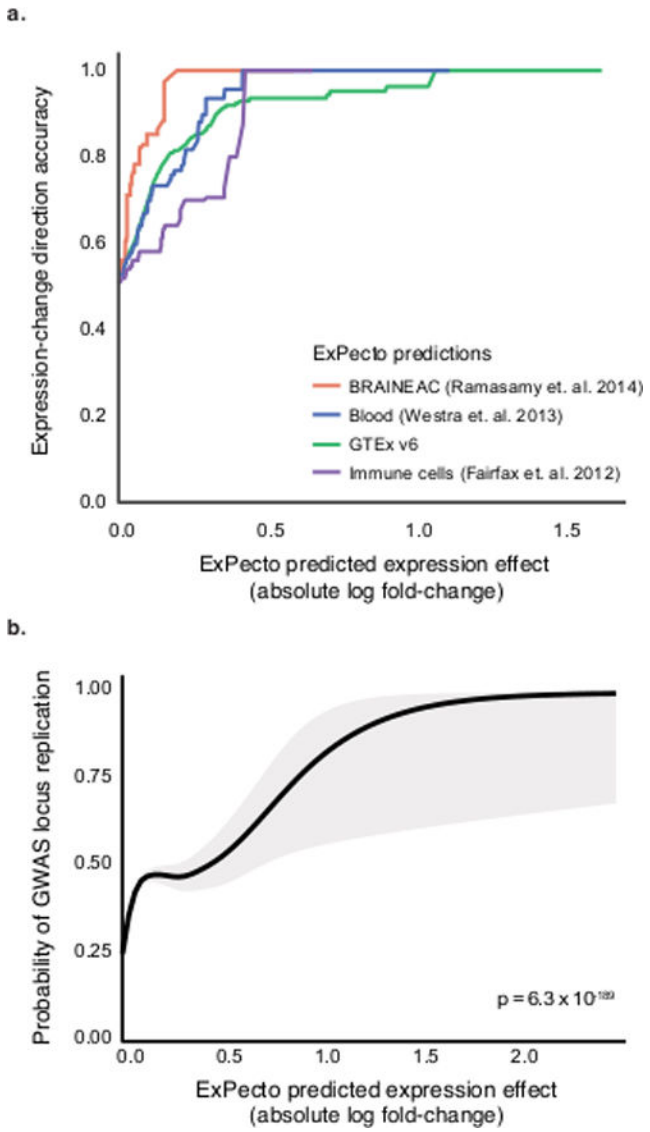

Figure 2. Tissue-specific prediction of expression-altering variations

a). eQTL direction prediction accuracy increases with predicted magnitude of variant effect. Each line shows performance for one eQTL study. x-axis represents the predicted effect magnitude cutoff, as measured by absolute log fold-change. y-axis represents the accuracy of predicting the expression change directionality for the variants above the corresponding effect magnitude.

b) GWAS loci with stronger predicted effect variants are more likely to be replicated by separate studies. The generalized additive model fitted curve of replication probability was shown with 95% confidence interval. x-axis shows the max predicted expression absolute log fold-change across all non-cancer tissues. A GWAS locus is considered as replicated if it is within 10kb to the reported SNP of a different study.

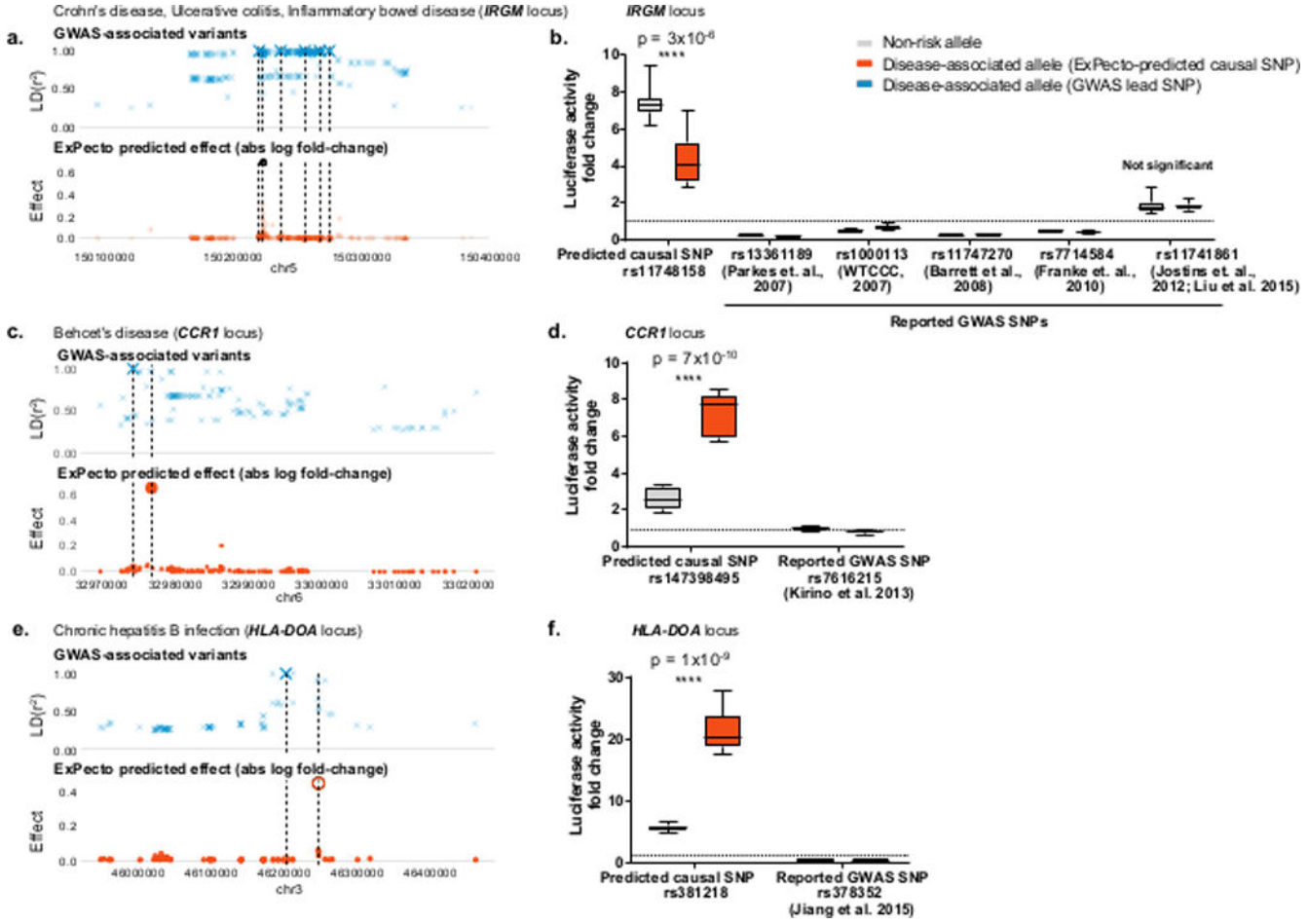

Figure 3. Prioritize putative causal variants from GWAS loci with expression effect prediction

(a, c, e). ExPecto expression effect prediction prioritizes putative causal SNPs in inflammatory bowel disease (a), Behcet's disease(c), and chronic hepatitis B infection (e) GWAS loci. Linkage disequilibrium r^2 scores between the reported variant and LD variants in the study population were shown in the top panel (variants are indicated by the × symbols) and the predicted expression effects (maximum across tissues) were shown in the bottom panel (variants are indicated by the dot symbols). The upper panels (GWAS-associated variants) showed the reported SNP(s) from the GWAS studies, indicated by the dashed lines, and all variants in LD with this variant ($r^2 > 0.25$). The lower panels (ExPecto predicted effect) showed the predicted effects of all LD variants and the ExPecto-predicted causal variant is indicated by the dashed line.

(b, d, f) Luciferase reporter assay test verified predicted differential transcriptional regulatory activities of sequence elements with the risk allele and with the non-risk allele of prioritized variants, while showing no difference for the GWAS lead variants. Three top prioritized variants near *IRGM*[37–42] (b), *CCR1*[43](d), and *HLA-DOA*[44] (f) showed differential transcriptional regulatory activity in the predicted direction while the reported GWAS SNPs show either no transcriptional activation activity or no detectable activity alteration. Luciferase activity is normalized by the empty vector, which is indicated by the dotted line. Statistical significance was based on two-sided t-test. Each allele was tested with

at least 11 total replicates from 3 independent experiments (n=11 for the rs7616215 non-risk allele, n=12 for all other alleles). Central values of the boxplot represent the median, box extends from 25th to 75th percentiles, and whiskers extend to the maximum and minimum values.

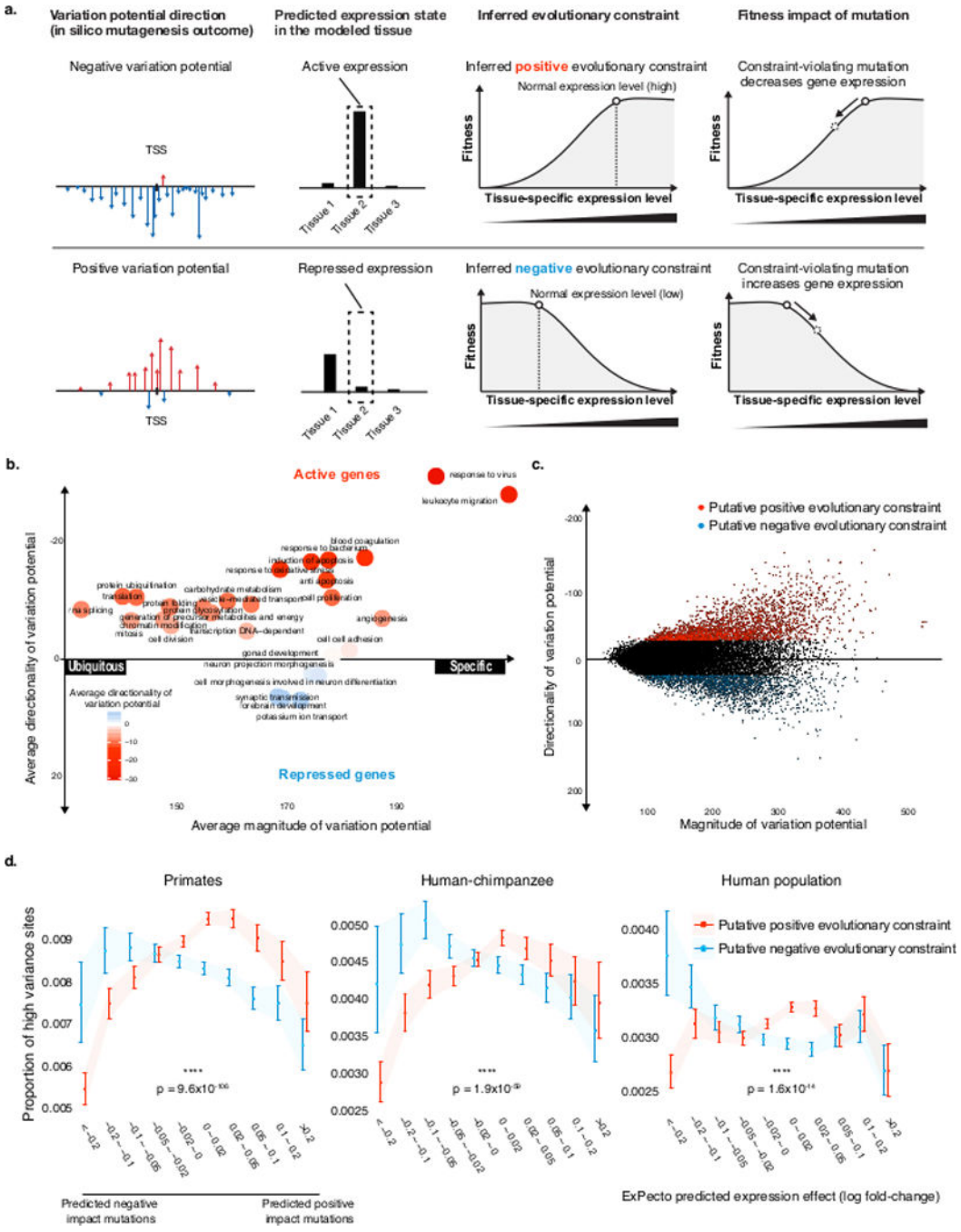

Figure 4. Variation potential is predictive of gene regulatory specificity, activation status, and evolutionary constraints

a). Schematic overview of association between variation potential, gene expression, and evolutionary constraints.

b). Gene expression specificity and activation status can be predicted from the magnitude and directionality of gene variation potential. The position of each gene set is computed as the cumulative mutation effects (directionality) and cumulative absolute mutation effects (magnitude) across all genes in the set. Each gene set is colored by the directionality of

variation potential. See Supplementary Fig. 11 for relationship between VP and gene-wise expression properties. Whole blood model predictions are shown as examples here.
c). Inference of genes with putative directional evolutionary constraints from variation potentials. Each dot represents a gene. x- and y- axis shows the cumulative predicted mutation effects (log fold-change) of positive and negative impact mutations within 1kb off TSS, respectively. See Methods and Supplementary Fig. 13 for details in determining threshold for calling putative constrained genes. This example shows predictions from the subcutaneous adipose tissue model.
d). Evolution and population genetics signatures show differential selective pressure for mutations in putative positive and negative constraint genes across evolutionary time scales. Selection pressures across mutations with different predicted effects (x-axes) are estimated based on proportion of high variance sites among primate species (phyloP < -2.3 which corresponds to $p < 0.005$ for acceleration; left panel y-axis; primates), divergent sites between human and the inferred human-chimpanzee common ancestor (mid panel y-axis; human-chimpanzee), and common variant sites (minor allele frequency > 0.001) in human populations (right panel y-axis; human population). The error bars show 90% confidence intervals.

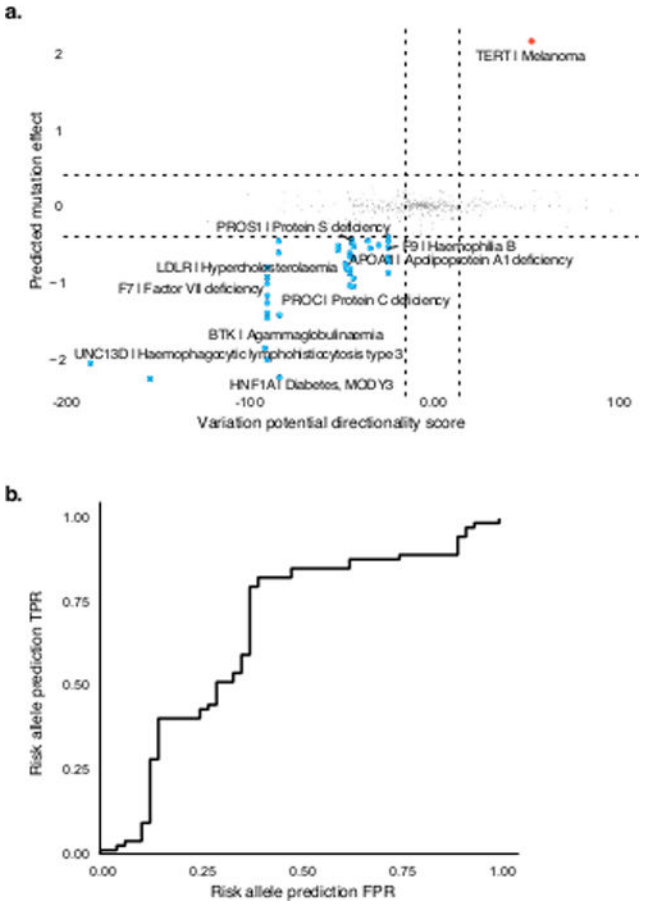

Figure 5. ***Ab initio*** **prediction of allele-specific disease risk integrating predicted expression effects and inferred evolutionary constraints**

a). HGMD regulatory disease mutations with strong predicted effects are violators of the putative evolutionary constraints. y-axis shows the ExPecto predicted effects of annotated deleterious mutations (maximum across tissues). x-axis shows the inferred evolutionary constraints measured by variation potential directionality score (sum of gene-wise predicted mutation effects within 1kb to TSS) of the maximum predicted effect tissue. Negative effect mutations with nearest gene being putatively constrained to be high expressing are shown in blue and positive effect mutations with nearest gene being putatively constrained to be low expressing are shown in red.

b). Prioritized GWAS LD variant constraint violation score is predictive of whether the reference allele or the alternative-risk allele is the risk allele. The y-axis and x-axis shows the true positive rate and false positive rate of the receiver-operating characteristic, which shows prediction performance of constraint violation score for the GWAS disease risk allele. The constraint violation score is the product of predicted variant effect and the variation potential directionality score. The median constraint violation score across all non-cancer tissue or cell types for each variant were used.