nature genetics

ARTICLES

https://doi.org/10.1038/s41588-018-0160-6

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

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

Key challenges for human genetics, precision medicine and evolutionary biology include deciphering the regulatory code of gene expression and understanding the transcriptional effects of genome variation. However, this is extremely difficult because of the enormous scale of the noncoding mutation space. We developed a deep learning–based framework, ExPecto, that can accurately predict, ab initio from a DNA sequence, the tissue-specific transcriptional effects of mutations, including those that are rare or that have not been observed. We prioritized causal variants within disease- or trait-associated loci from all publicly available genome-wide association studies and experimentally validated predictions for four immune-related diseases. By exploiting the scalability of ExPecto, we characterized the regulatory mutation space for human RNA polymerase II–transcribed genes by in silico saturation mutagenesis and profiled > 140 million promoter-proximal mutations. This enables probing of evolutionary constraints on gene expression and ab initio prediction of mutation disease effects, making ExPecto an end-to-end computational framework for the in silico prediction of expression and disease risk.

Sequence-dependent control of gene transcription contributes to the complexity of multicellular organisms. Expression-altering genomic variation can thus have a 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 relating to the expression effects of common genome variation and their relevance to disease[5]. However, such approaches are generally limited to mutations that are observed frequently and that have matched expression observations, ideally in the relevant tissue or cell type. Moreover, central 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 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. A computational approach is especially important in humans, where only 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 changes to consequent gene expression changes. 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 the direct prediction of fitness impact of genomic changes and the resulting expression alterations using only the sequence and evolutionary information it contains.

Human gene expression profiles show a wide diversity of expression patterns across genes, cell types and cellular states. However, understanding of sequences that activate or repress expression in specific tissues, let alone the ability to quantify the transcriptional modulation strength of a sequence element, remains 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 that are smaller than 2 kb[6–10]. As such, sequence-based prediction of expression in humans is still a critical open challenge.

Here we describe ExPecto (see URLs), a tissue-specific modeling framework for predicting gene expression levels ab initio from sequences 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 40-kb promoter-proximal sequences. A critical feature of this framework is that it does not use any variant information for training, which enables prediction of expression effects for any variant, even those that are rare or that have never previously been observed.

The resulting ExPecto models make highly accurate cell-type-specific predictions of expression from DNA sequences, as evaluated with known expression quantitative trait loci (eQTLs) and validated causal variants from a massively parallel reporter assay. With this capability, we prioritized putative causal variants associated with human traits and diseases from hundreds of publicly available genome-wide associated studies (GWAS). We experimentally validated newly predicted putative causal variants for Crohn's disease, ulcerative colitis, Behçet's disease and hepatitis B virus (HBV) infection, demonstrating that these ExPecto-predicted functional SNPs showed allele-specific regulatory potential, whereas the lead SNPs from the GWAS did 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 comprehensively

[1]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA. [2]Graduate Program in Quantitative and Computational Biology, Princeton University, Princeton, NJ, USA. [3]Flatiron Institute, Simons Foundation, New York, NY, USA. [4]Department of Computer Science, Princeton University, Princeton, NJ, USA. *e-mail: ogt@cs.princeton.edu

probe the tissue-specific impact of human gene 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', are indicative of the phenotypic impact of expression-altering mutations.

By 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 larger scale with lower cost, including those for inferring disease-causal mutations. We demonstrate the 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 tissue-specific expression from human promoter-proximal sequences, we built a modular framework (Fig. 1a and Methods). First, we used deep learning to generate a repertoire of potential regulatory sequence representations capable of predicting the epigenomic effects of any genomic variant from sequence only. This was accomplished with a deep convolutional neural network that was trained to predict 2,002 different histone mark, transcription factor and DNA accessibility profiles for > 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 40-kb regions (Supplementary Fig. 1). Third, tissue-specific regularized linear models used the transformed epigenomic information, which was centered around the transcription start site (TSS), to predict expression of RNA polymerase II (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 sequences, with 0.819 median Spearman correlation between predicted and observed expression log(RPKM) values 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 the 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 ($P = 6.9 \times 10^{-25}$ by two-sided Wilcoxon rank-sum test), likely owing to a lack of causal dependency information, probably resulting from that DNase I–hypersensitive sites are generated by binding proteins of distinct or even opposite effects on expression.

Furthermore, in addition to accurately capturing global expression, ExPecto predictions recapitulated the tissue specificity of expression, with expression predictions being substantially more similar to the experimental measurements from identical or similar cell types than from other cell types on holdout sequences (Fig. 1c). As the 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 of the top five features with positive weights were transcription factors (ER-α and GR) in the breast cancer cell lines T-47D and ECC-1; for the whole blood model, all of the top five features were from the blood-derived cell lines and erythroblast cells (Supplementary Table 1).

The ability of ExPecto to predict tissue-specific gene expression from sequences provided the basis for estimating the transcriptional effects of genomic variation (Fig. 1a,d). These computational predictions of variant effects do not use any variant-specific information for training and thus can be scaled to all human population variants, including billions of potentially 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 in silico mutagenesis to probe the effects of > 140 million variants, including all of the 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 which complements quantitative genetics approaches by avoiding their limitations and predicting the effects of rare and unobserved disease-relevant variants undetected by quantitative genetics.

Effect of genomic variants on tissue-specific expression. To evaluate ExPecto's predictions of the tissue-specific effects of genomic variants on gene expression, we compared these to eQTL data from multiple studies. ExPecto's ab initio sequence-based prediction was especially useful for prioritizing causal eQTL variants, as it was unconfounded by linkage disequilibrium (LD). Thus, even though the majority of eQTL variants are expected to have no effect on expression[12] (of Genotype-Tissue Expression (GTEx) Project lead variants, only 3.5–11.7% are estimated to be causal variants, which is < 1% of all GTEx eQTL variants[13]), we expected 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 variants with the strongest effects and provided accurate predictions for tens of thousands of variants (Supplementary Fig. 3). This suggests that a higher proportion of eQTLs with strong predicted effects are causal as compared to background levels, as the eQTL effect direction of non-causal SNPs should be independent from the predicted directions. Moreover, ExPecto models for the tissue matching eQTL detection provided more accurate predictions than for any other tissue (Supplementary Fig. 3). We also demonstrated accuracy on three other large-scale eQTL studies from brain, primary immune cells and blood, respectively[14–16] (Fig. 2a).

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

Because expression models can accurately predict the causal gene expression effects of single-nucleotide variants (SNVs) and small insertion–deletions (indels) among eQTLs, we examined the expression effect of variants found in the human population across the full range of allele frequencies (16.5 million variants from the 1000 Genomes project) (Supplementary Data 1). In contrast

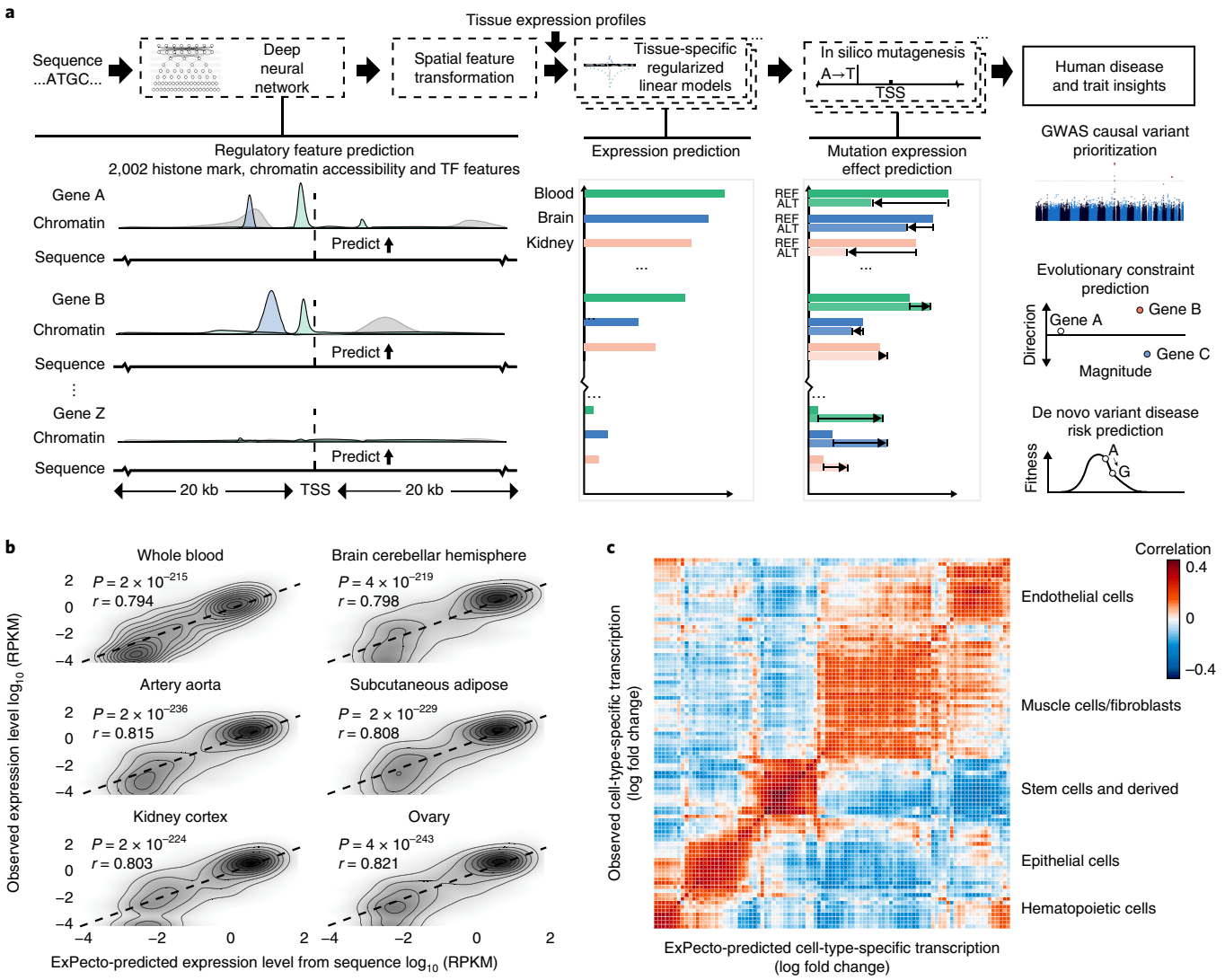

Fig. 1 | Deep learning-based sequence model accurately predicts cell-type-specific gene expression. a, Schematic overview of the ExPecto sequence-based gene expression prediction framework. The predictive model contains three components, a deep convolutional neural network trained on chromatin profiling 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(RPKM) values on 990 genes from the holdout chromosome 8 (chr8; on *x* axis) were compared with experimentally measured log(RPKM) values (*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 heat map shows, on holdout genes, correlations between cell-type-specific expression profiles, as measured by log(fold change) value over cell-type average and the sequence-based predicted log(fold change) values. **d**, Predicted mutation effects from in silico mutagenesis of promoter-proximal regions of 23,779 genes showed substantial variation, as indicated by color. The average of the predicted effects for different variants at the same position was computed. Genes were sorted by gene-wise average predicted mutation effects. Only positions with > 0.5 average absolute log(fold change) values are shown. -1,000 is upstream of the TSS, and + 1,000 is downstream (by base pair). The whole blood model predictions are shown.

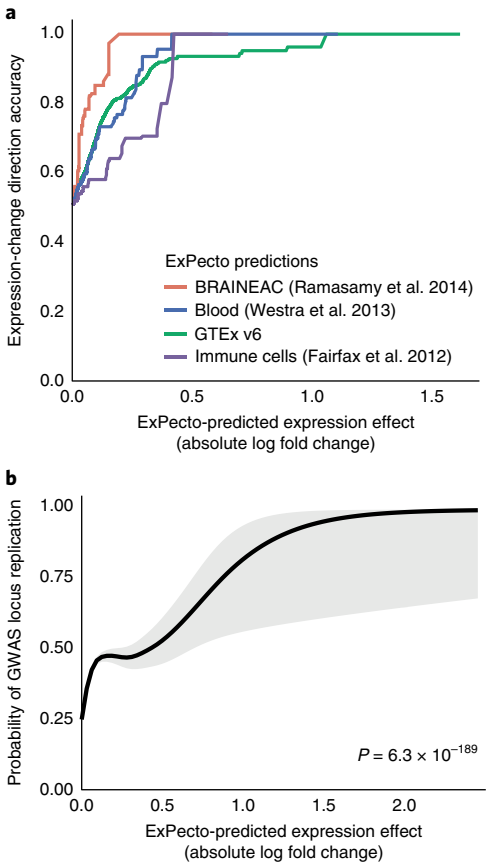

Fig. 2 | Tissue-specific prediction of expression-altering variations. **a**, eQTL direction prediction accuracy increases with the predicted magnitude of variant effect. Each line shows performance for one eQTL study. The *x* axis represents the predicted effect magnitude cutoff, as measured by absolute log(fold change) value. The *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 is shown with the 95% confidence interval. The *x* axis shows the maximum predicted expression absolute log(fold change) value across all non-cancer tissues. A GWAS locus is considered as replicated if it is within 10 kb of the reported SNP of a different study.

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 variants with high expression effects had a similar minor allele frequency (MAF) distribution as all of the 1000 Genomes variants (Supplementary Fig. 2). As expected, variants with stronger predicted expression effects were 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 the disease or trait loci of 3,000 GWAS[18] (Supplementary Table 2). Although GWAS inform on 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, largely owing to LD. By assessing the overall performance of ExPecto-prioritized variants, we found that loci having variants with stronger predicted effects were significantly more likely to be replicated in a different GWAS ($P = 6.3 \times 10^{-189}$ by two-sided Wald test with logistic regression; Fig. 2b) (see also Supplementary Fig. 6 for analysis using only variants with $P < 5 \times 10^{-8}$). Moreover, GWAS LD variants with stronger predicted effects were more likely to be the exact replicated variant ($P = 5.6 \times 10^{-14}$ by two-sided Wald test with logistic regression). For instance, an earlier GWAS for venous thromboembolism[19] identified rs3756008 as the lead causal variant; however, ExPecto-prioritized LD variant rs4253399 near the *F11* locus was discovered in a later study by using a larger cohort[20] (Supplementary Fig. 7a). Similar examples include variants in autoimmune disease–associated loci (rs7528684 and 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.

We then focused on immunity-related diseases. We 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 GWAS (Fig. 3a,c,e). We found that these LD SNPs prioritized by expression effect, although 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, was predicted to decrease the expression of *IRGM*, a gene involved in the innate immune response that is significantly associated with Crohn's disease, ulcerative colitis and general inflammatory bowel disease; indeed, we observed significantly decreased reporter expression ($P = 3 \times 10^{-6}$; Fig. 3a,b). The second top SNP, rs147398495, which is associated with Behçet's disease and is near *CCR1* (a chemokine receptor–encoding gene), also significantly changed transcriptional regulatory activity ($P = 7 \times 10^{-10}$; Fig. 3c,d). For a GWAS locus associated with chronic HBV infection, our third top SNP, rs381218, was predicted by ExPecto to affect the expression of *HLA-DOA* (a gene encoding a major histocompatibility complex (MHC)-II molecule functional in B cell lysosomes) and, indeed, resulted in a fourfold change in reporter activity ($P = 1 \times 10^{-9}$; Fig. 3e,f). In all of these cases, none of the lead SNPs in the seven GWAS showed significant differences in transcriptional regulatory activity. Notably, 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 organisms poses significant difficulties in predicting phenotypic or disease consequences of expression alteration where perturbations of different genes elicit distinct consequences. Because our model enables exploration of the tissue-specific expression effects of genomic sequence variation at large scale, which essentially provides an 'in silico' assay of a mutation's effects, it enabled us to analyze the trace of natural 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' (Supplementary Fig. 8), is indicative of the phenotypic impact of expression-altering mutations. Furthermore, we found that variation potential was indicative of the innate expression properties of genes (for example, tissue specificity of expression and activation or repression status).

We computed a catalog of predicted effects for more than 140 million mutations that included all possible single-nucleotide mutations 1 kb upstream and downstream of the TSS for each Pol II–transcribed gene. This identified > 1.1 million mutations

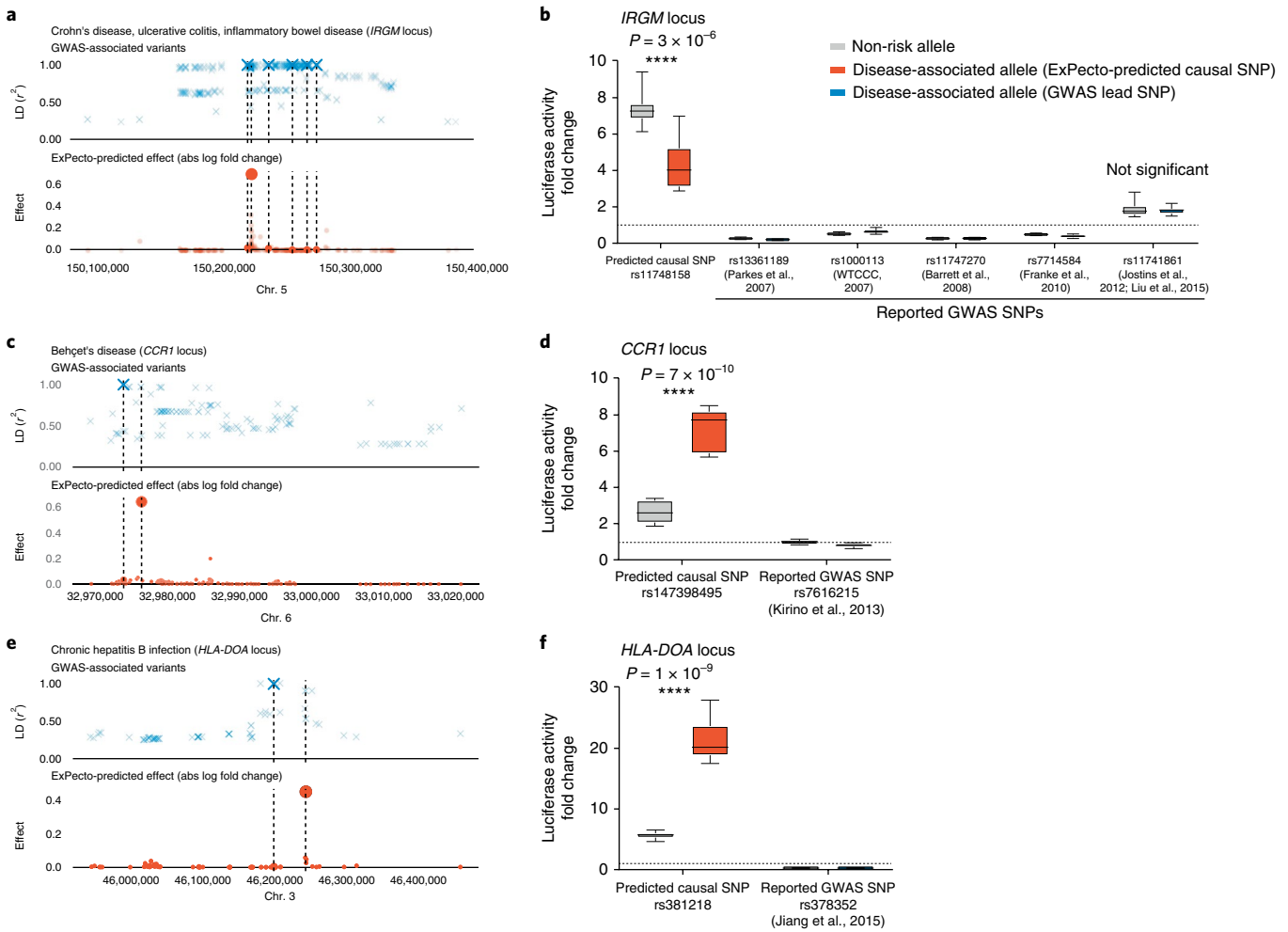

Fig. 3 | Prioritized putative causal variants from GWAS loci with expression effect prediction. a–f, ExPecto expression effect prediction prioritizes putative causal SNPs in inflammatory bowel disease (**a**), Behçet's disease (**c**) and chronic HBV infection (**e**) GWAS loci and luciferase reporter assays for the predicted causal SNPs and the reported GWAS SNPs (**b**,**d**,**f**). In **a**, **c** and **e**, LD r^2 scores between the reported variant and LD variants in the study population (variants are indicated by the '×' symbols) (top) and the predicted expression effects (maximum across tissues; variants are indicated by the dot symbols) (bottom) are shown. The upper panels (GWAS-associated variants) show the reported SNP(s) from the GWAS, and all variants in LD with this variant ($r^2 > 0.25$). The lower panels (ExPecto-predicted effect) show the predicted effects of all LD variants. The reported GWAS SNPs and the ExPecto-predicted causal variant are indicated by the dash lines. In **b**, **d** and **f**, luciferase reporter assays 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**) show differential transcriptional regulatory activity in the predicted direction, whereas the reported GWAS SNPs show either no transcriptional activation activity or no detectable activity alteration. Luciferase activity was normalized to that for the empty vector, which is indicated by the dashed line. Statistical significance was based on a two-sided *t* test. Each allele was tested with at least 11 total replicates from three independent experiments ($n = 11$ for the rs7616215 non-risk allele, $n = 12$ for all other alleles). Central values of the box plot represent the median; the box extends from the 25th to the 75th percentile; and whiskers extend to the maximum and minimum values.

with a strong predicted expression effect (at high confidence). As expected, mutations with a predicted negative effect (i.e., mutations predicted to decrease expression) were generally positioned immediately upstream of the TSS, near the position at −50 bp (Supplementary Fig. 9), which is the typical position of core promoter elements[27]. Notably, the bases with stronger predicted mutation effects showed significantly higher evolutionary constraints both in the modern human population ($P = 2.4 \times 10^{-36}$; Supplementary Fig. 10a) and in ancestral evolutionary history ($P = 2.2 \times 10^{-16}$; Supplementary Fig. 10b).

We observed that the tissue-specific variation potential of a gene was highly predictive of expression properties for that gene (Fig. 4). Specifically, we could predict both whether a gene's expression was ubiquitous versus whether it was tissue or condition specific and whether a gene was active or repressed (Fig. 4a,b, Supplementary Fig. 11 and Supplementary Data 2). Genes with low magnitudes of variation potential were characteristically ubiquitously expressed genes involved in essential cellular processes (for example, splicing, translation, protein folding and energy metabolism) (Fig. 4b). In contrast, genes with high magnitudes of variation potential were tissue specific (for example, those associated with synaptic transmission) or condition specific (for example, those associated with the innate immune response) (Supplementary Fig. 11). Among the set of non-ubiquitous genes, the directionality of variation potential (positive or negative cumulative mutation effect) in a given tissue predicted the gene's activation status: negative variation potential was predictive of actively expressed genes, and positive variation potential was indicative of repressed expression in the modeled

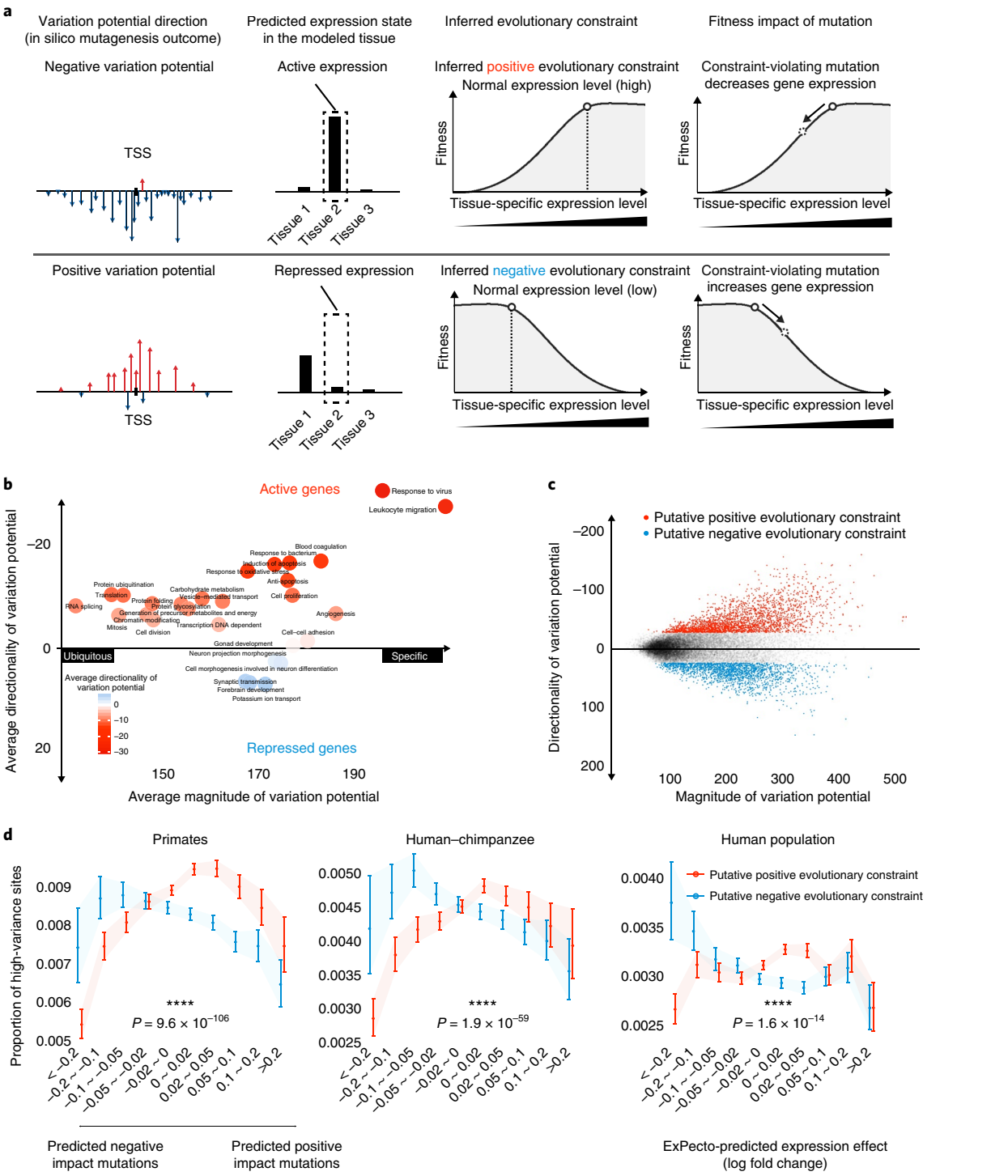

Fig. 4 | Variation potential is predictive of gene regulatory specificity, activation status and evolutionary constraints. a, Schematic overview of association between variation potential, gene expression and evolutionary constraints. **b**, Gene expression specificity and activation status can be predicted from the magnitude and directionality of gene variation potential. The position of each gene set was computed as the average cumulative mutation effects (directionality) and average 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 the relationship between variation potential 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. The *x* and *y* axes show the cumulative predicted mutation effects and the log(fold change) values of mutations with positive and negative impact within 1 kb of the TSS, respectively. See Supplementary Fig. 13 and the Methods for details on determining the threshold for calling putative constrained genes. Predictions from the subcutaneous adipose tissue model are shown. **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) were estimated based on the proportion of high-variance sites among primate species (phyloP < −2.3, which corresponds to $P < 0.005$ for acceleration) (left), divergent sites between human and the inferred human–chimpanzee common ancestor (middle) and common variant sites (minor allele frequency > 0.001) in human populations (right). The error bars show 90% confidence intervals.

tissue (Supplementary Fig. 11). For example, in brain tissue, genes involved in synaptic transmission had a negative variation potential and were thus predicted to be actively expressed. On the contrary, in non-neuronal tissues, they had positive variation potential and were thus predicted to be repressed (as in Fig. 4a) (Supplementary Data 2). In a given tissue, genes with a strong positive predicted variation potential indeed appeared to be repressed, as they were expressed at significantly higher levels in other tissues as compared to genes with similar expression levels but a low magnitude of variation potential (analysis based on GTEx; Supplementary Fig. 12). Thus, variation potentials are not simply reflecting the magnitude of gene expression, but rather they are distilling expression properties from sequence.

We hypothesized that this connection between variation potentials and the expression properties of genes was imposed by evolutionary constraint. Specifically, we proposed that genes strongly enriched with mutations of predicted negative effect were under positive evolutionary constraint (i.e., decreasing expression of that gene would be deleterious) and vice versa (Fig. 4a (third and fourth panels),c,d, Methods and Supplementary Fig. 13). In support of this differential evolutionary constraints hypothesis, both variant allele frequencies in human populations and evolutionary conservation evidence supported divergent selection signatures for genes with putative positive and negative constraints ($P < 1.6 \times 10^{-14}$ 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 had a lower allele frequency and higher evolutionary conservation, as compared to negative constraint genes, and vice versa (Fig. 4d). These divergent selection signatures were not simply driven by gene expression levels, as randomly selecting genes with matching expression levels did not show differential selection (Supplementary Fig. 14). Moreover, genes with stronger inferred evolutionary constraints were significantly more enriched in GWAS disease-associated genes ($P = 6.1 \times 10^{-53}$ by 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 an analysis framework for pathogenic regulatory mutations that addresses both the impact of a variant on gene expression and the fitness impact of expression alterations (Fig. 4a). Recognizing pathogenic regulatory mutations is very challenging because both of these problems are difficult to address experimentally[5]. Our proposed computational sequence-based approach addressed these problems across the genome, even for mutations that had not previously been observed experimentally.

We thus assessed the ability of our approach to predict disease risk ab initio from sequences. At the individual-variant level, we predicted whether a specific sequence alteration was 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 caused a positively constrained, highly expressed gene to have substantially decreased expression, then it was likely to be deleterious (Fig. 4a, fourth panel). We then evaluated our predictions on the curated pathogenic regulatory mutations in the Human Gene Mutation Database (HGMD) and prioritized putative causal LD SNPs derived from the GWAS catalog[18,28].

Most ExPecto-predicted mutations of strong effect from HGMD were predicted to decrease expression (Fig. 5a), and correspondingly, all of them were near genes with putative positive evolutionary constraints, as would be expected from our findings above. Positive constraints are also consistent with the current understanding of these diseases being caused by a deficiency in

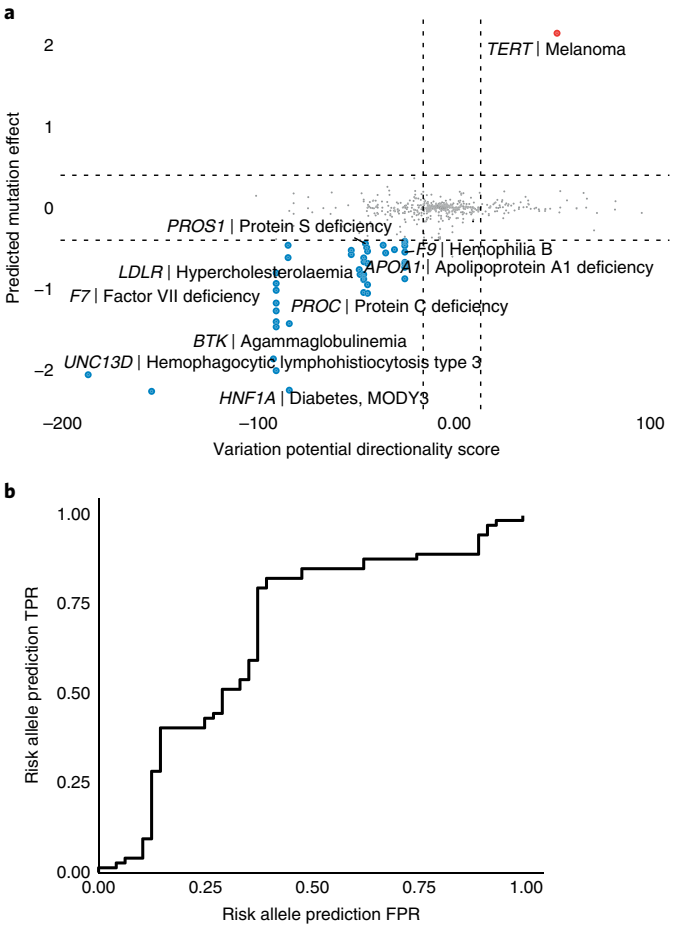

Fig. 5 | Ab initio prediction of allele-specific disease risk integrating predicted expression effects and inferred evolutionary constraints. **a**, HGMD regulatory disease-associated mutations with strong predicted effects are violators of the putative evolutionary constraints. The *y* axis shows the ExPecto-predicted effects of annotated deleterious mutations (maximum across tissues). The *x* axis shows the inferred evolutionary constraints, as measured by the variation potential directionality score (sum of gene-wise predicted mutation effects within 1 kb of the TSS) for the tissue with the maximum predicted effect. Mutations of negative effect with the nearest gene being putatively constrained to be highly expressed are shown in blue, and those of positive effect with the nearest gene being putatively constrained to have low expression are shown in red. **b**, Prioritized GWAS LD variant constraint violation score is predictive of whether the reference allele or the alternative allele is the risk allele. The *y* and *x* axes show the true-positive rate (TPR) and false-positive rate (FPR), respectively, of the receiver operating characteristic, which shows prediction performance of the constraint violation score for the GWAS disease risk allele. The constraint violation score is the product of the predicted variant effect and the variation potential directionality score. The median constraint violation scores across all non-cancer tissue or cell types for each variant were used.

certain genes, such as the coagulation factor–encoding genes *F7* and *F9* for factor VII deficiency[29] and hemophilia B, respectively; *PROC* and *PROS1* for blood clotting disorders caused by protein C deficiency and protein S deficiency, respectively; and *APOA1* and *LDLR* for hypolipoproteinemia and hypercholesterolemia caused by apolipoprotein deficiency and decreased receptor-mediated endocytosis of LDL cholesterol, respectively. *UNC13D*, which is essential for intracellular trafficking and exocytosis of lytic granules, was associated with hemophagocytic lymphohistiocytosis

type 3[30]. Decreased expression of *BTK*, an essential gene for B cell development and maturation, causes agammaglobulinemia[31]. A mutation in *HNF1A* is well known as a major cause for maturity onset of diabetes (MODY)[32] and is considered to be important in beta cell differentiation[33].

Only one HGMD disease-associated mutation was predicted to strongly increase transcriptional activity and was 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-associated mutations are known to be deleterious, these results demonstrate that ExPecto could 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 suggested that the alternative allele was likely more deleterious, whereas a negative violation score suggested that the reference allele was likely to be more deleterious. This GWAS evaluation standard directly included both deleterious and protective variants (risk alleles are reference alleles for 37 loci and alternative alleles for 63 loci). Our approach was significantly predictive ($P=0.002$ by Wilcoxon rank-sum test, AUROC $=0.67$; Fig. 5b and Supplementary Data 3) of the known risk alleles detected from GWAS. This evaluation thus demonstrated that predicted effects that violated inferred constraints were predictive of risk alleles in GWAS, indicating that ExPecto could predict which allele was deleterious or protective without any prior variant–disease association information. Taken together, our results suggest this approach as a promising direction for large-scale prediction of disease risk, which will be especially useful for interpreting the enormous amount of potential disease-associated mutations for which there is little to no prior knowledge.

Discussion

ExPecto 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 that this can be routinely performed at a whole-genome level. Our approach can potentially be further combined with statistical models (reviewed in Pasaniuc and Price[36]) for further improvement on causal variant identification. Moreover, because the method is equally applicable to rare or common variants, it allows wider application to a 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 consequences of expression alteration that are otherwise very difficult to study in humans. 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 analysis of regulatory disease-associated mutations. Although 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 TF binding; additional data capturing ultra-distal regulatory sequences, especially those mediated by long-range interactions; and epigenetic inheritance mechanisms that affect expression independently of sequence, such as imprinting via DNA methylation, could be incorporated into the ExPecto framework and will likely be important for the 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 noncoding variants and expression perturbations. Such analyses could, in the future, be used for grouping patients in drug and other treatment trials, in disease subtyping and eventually in personalized treatment. At the same time, we expect that tapping into human genome evolution information, as allowed by sequence-based expression modeling, will provide valuable insights required for comprehensive understanding of healthy and disease modes of human gene expression.

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

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0160-6.

Received: 9 December 2017; Accepted: 3 May 2018;
Published online: 16 July 2018

References

1. Pickrell, J. K. et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
2. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
3. Gamazon, E. R. et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
4. Li, X. et al. The impact of rare variation on gene expression across tissues. *Nature* **550**, 239–243 (2017).
5. Edwards, S. L., Beesley, J., French, J. D. & Dunning, A. M. Beyond GWASs: illuminating the dark road from association to function. *Am. J. Hum. Genet.* **93**, 779–797 (2013).
6. Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U. & Gaul, U. Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature* **451**, 535–540 (2008).
7. Beer, M. A. & Tavazoie, S. Predicting gene expression from sequence. *Cell* **117**, 185–198 (2004).
8. Yuan, Y., Guo, L., Shen, L. & Liu, J. S. Predicting gene expression from sequence: a reexamination. *PLoS Comput. Biol.* **3**, e243 (2007).
9. Bussemaker, H. J., Li, H. & Siggia, E. D. Regulatory element detection using correlation with expression. *Nat. Genet.* **27**, 167–171 (2001).
10. Kreimer, A. et al. Predicting gene expression in massively parallel reporter assays: a comparative study. *Hum. Mutat.* **38**, 1240–1250 (2017).
11. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep-learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
12. Aguet, F. et al. Local genetic effects on gene expression across 44 human tissues. *Nature* **550**, 204–213 (2017).

13. Battle, A., Brown, C. D., Engelhardt, B. E. & Montgomery, S. B. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
14. Westra, H.-J. et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
15. Ramasamy, A. et al. Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nat. Neurosci.* **17**, 1418–1428 (2014).
16. Fairfax, B. P. et al. Genetics of gene expression in primary immune cells identifies cell-type-specific master regulators and roles of HLA alleles. *Nat. Genet.* **44**, 502–510 (2012).
17. Tewhey, R. et al. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **165**, 1519–1529 (2016).
18. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
19. Germain, M. et al. Genetics of venous thrombosis: insights from a new genome-wide association study. *PLoS One* **6**, e25581 (2011).
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.* **37**, 512–521 (2013).
21. Plagnol, V. et al. Genome-wide association analysis of autoantibody positivity in type 1 diabetes cases. *PLoS Genet.* **7**, e1002216 (2011).
22. Chu, X. et al. A genome-wide association study identifies two new risk loci for Graves' disease. *Nat. Genet.* **43**, 897–901 (2011).
23. Sawcer, S. et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214–219 (2011).
24. Graham, R. R. et al. Genetic variants near *TNFAIP3* on 6q23 are associated with systemic lupus erythematosus. *Nat. Genet.* **40**, 1059–1061 (2008).
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.* **47**, 1457–1464 (2015).
26. Lee, Y.-C. et al. Two new susceptibility loci for Kawasaki disease identified through genome-wide association analysis. *Nat. Genet.* **44**, 522–525 (2012).
27. Xi, H. et al. Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1. *Genome Res.* **17**, 798–806 (2007).
28. Stenson, P. D. et al. The Human Gene Mutation Database: 2008 update. *Genome Med.* **1**, 13 (2009).
29. Nagaizumi, K. et al. Two double-heterozygous mutations in the *F7* gene show different manifestations. *Br. J. Haematol.* **119**, 1052–1058 (2002).
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* **115**, 461–473 (2003).
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.* **200**, 927–934 (2004).
32. Yamagata, K. et al. Mutations in the hepatocyte nuclear factor-4α gene in maturity-onset diabetes of the young (MODY1). *Nature* **384**, 458–460 (1996).
33. Servitja, J.-M. et al. Hnf-1α (MODY3) controls tissue-specific transcriptional programs and exerts opposed effects on cell growth in pancreatic islets and liver. *Mol. Cell. Biol.* **29**, 2945–2959 (2009).
34. Huang, F. W. et al. Highly recurrent *TERT* promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).
35. Vinagre, J. et al. Frequency of *TERT* promoter mutations in human cancers. *Nat. Commun.* **4**, 2185 (2013).
36. Pasaniuc, B. & Price, A. L. Dissecting the genetics of complex traits using summary-association statistics. *Nat. Rev. Genet.* **18**, 117–127 (2017).
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.* **39**, 830–832 (2007).
38. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
39. Barrett, J. C. et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* **40**, 955–962 (2008).
40. Franke, A. et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* **42**, 1118–1125 (2010).
41. Jostins, L. et al. Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
42. Liu, J. Z. et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).
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.* **45**, 202–207 (2013).
44. Jiang, D. K. et al. Genetic variants in five novel loci including *CFB* and *CD40* predispose to chronic hepatitis B. *Hepatology* **62**, 118–128 (2015).

Acknowledgements

The authors acknowledge all members of the Troyanskaya lab for helpful discussions. This work is supported by NIH grants R01HG005998, U54HL117798 and R01GM071966, HHS grant HHSN272201000054C and Simons Foundation grant 395506. The authors are pleased to acknowledge that a substantial portion of the work in this paper was 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. O.G.T. is a CIFAR fellow.

Author contributions

J.Z. and O.G.T. conceived and designed the study; J.Z. developed the computational methods and performed the analyses; C.L.T. designed and performed experimental studies; K.Y., K.M.C. and A.K.W. developed the ExPecto web server; J.Z., C.L.T. and O.G.T. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0160-6.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to O.G.T.

Methods

ExPecto framework architecture. The ExPecto sequence-based expression prediction framework includes three components that act sequentially (Fig. 1a). First, a deep neural network epigenomic effects model scans the long input sequence with a moving window and provides an output of predicted probabilities for histone marks, TFs 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 Fig. 1). Lastly, 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 used a deep convolutional neural network to transform genomic sequences to epigenomic features. Our approach generated a cell-type-specific model for 2,002 genome-wide histone marks, TF-binding and chromatin accessibility profiles (based on training data from the 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 (2,000 bp), and the new model was trained to predict twice as many regulatory features for > 200 cell types (Supplementary Note). Critically, this deep learning model did not use any mutation data for training. The deep convolutional neural network model predicted the epigenomic features of a 200-bp region, while also using the 1,800-bp surrounding context sequence. For each Pol II–transcribed gene, surrounding its representative TSS (see "Identification of representative transcription start sites"), the deep convolutional neural network model scanned the genomic sequence between + 20 kb upstream and −20 kb downstream to predict spatial chromatin organization patterns by using a moving window with a 200-bp step size, which yielded 200 spatial bins with a total number of 400,400 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 reduced the input dimensionality with ten exponential functions by weighting upstream and downstream regions separately, with weights based on relative distance to the TSS (transformed features with a higher decay rate were more concentrated near TSSs). This effectively reduced the number of features 20-fold to 20,020. The exponential functions were prespecified (based on empirical selection) and not learned during training. This spatial feature transformation followed by a learning linear model retained the flexibility of learning spatial patterns, which was equivalent to learning a smooth nonlinear spatial pattern function f constrained to the space of linear combinations of basis functions that corresponded to the feature transformations.

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

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

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 bp × 200 bp spatial bins within 20 kb of the TSS. l represents the indicator function, which equals 1 when the specified condition is satisfied and 0 otherwise. t_d represents the mean distance of region d to the TSS. For example, the bin from −200 bp to 0 bp has a distance of −100 bp and the bin from −400 bp to −200 bp has a distance of −300 bp. $\beta^{\text{up}}{}_{ik}$ 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 $a = \{0.01, 0.02, 0.05, 0.10, 0.20\}$. 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 were chosen by empirical evaluations, including the number and values of exponential terms, model design and window sizes, whereas all of the neural network model weights and linear model coefficients were learned from data. The ± 20-kb (40-kb) window around the TSS maximized ExPecto accuracy. Although smaller windows decreased prediction performance, increasing the window size to 50 kb, 100 kb or even 200 kb gave a negligible performance gain (Supplementary Fig. 16).

Application of ExPecto for sequence-based gene expression level prediction across tissues. Although ExPecto models could be trained on any expression profile, here we used 218 tissue expression profiles from the 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 base score = 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 a whole-chromosome holdout of chromosome 8 (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 chromosome 8 genes with paralogs on other chromosomes, and this did not negatively affect performance (Spearman correlation 0.819 for all 990 chromosome 8 genes, 0.821 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

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

where β_{ik}^{c} represents the coefficient for chromatin feature i, exponential function k in a cell type or in 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 or tissue models. c' is the index for cell types and 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. A gene expression effect is naturally estimated by the difference in the predicted expression levels for the reference and alternative alleles, and it is measured by the predicted log(fold change) value. Because 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

$$\begin{aligned}\text{effect}(p,t) \;=\; \sum_{\delta\in\Delta}\sum_{i}\left(p_{i\delta}^{\text{alt}} - p_{i\delta}^{\text{ref}}\right)\Bigg[&\sum_{k} 1(t<0)\beta_{ik}^{\text{up}} e^{-a_k\frac{|t+\delta|}{200\text{ bp}}} \\ &+ \sum_{k} 1(t>0)\beta_{ik}^{\text{down}} e^{-a_k\frac{|t+\delta|}{200\text{ bp}}}\Bigg]\end{aligned}$$

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, and $\beta^{\text{up}}{}_{ik}$ 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 the TSS is indicated by t. Notably, the predicted variant regulatory effect included both effects at the variant site and at adjacent positions (as long as the variant was within range of a 2,000-bp context sequence window for that region); thus, the variant expression effect considered regulatory effects in nine 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 compensated or truncated the alternative allele sequence equally on both sides to a total of 2,000 bp.

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 or 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 expression in the in vitro reporter assay and expression in vivo, as only a short element was cloned to a fixed position upstream of a reporter gene in the reporter assay. Specifically, we used regulatory effect models trained on a 230-bp input window instead of 2,000 bp, and only the in-place chromatin effect but not the effect on adjacent regions was computed, as these sequences were not cloned into the reporter-expressing vector. The position relative to the TSS was fixed at −100 bp.

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, whereas DeepSEA can identify variant effects that do not lead to significant changes in expression.

Computation of GWAS linkage disequilibrium SNPs. To systematically screen for SNPs in LD with the reported GWAS lead SNPs from the GWAS Catalog,

we first computed LD for all 88 million variants in the 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]. LD values between SNPs in five populations (AFR, African; AMR, Ad Mixed American; EAS, East Asian; EUR, European; SAS, South Asian) were computed with PLINK v1.90b. In total, we found 390,085 variants of LD $r^2 > 0.75$ with 15,571 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. Specifically, we used a luciferase assay to compare the ability of risk versus non-risk alleles to drive expression for the above-mentioned ExPecto-prioritized variants and the seven lead SNPs reported by the corresponding GWAS.

All of the genomic sequences were retrieved from the hg19 human genome assembly. For each risk allele (reference or alternative), Genewiz synthesized a 260-nt fragment: a 230-nt human genomic sequence and a 15-nt sequence matching 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) that was digested with the same enzymes. For the luciferase assay, 2×10^4 BE(2)-C cells were plated in 96-well plates, and 24 h later the cells were transfected with Lipofectamine 3000 (L3000-015, Thermo Fisher Scientific), 75 ng of variant-containing pGL4.23 plasmid (Supplementary Table 4) and 4 ng of pNL3.1 NanoLuc plasmid for normalization of transfection conditions. 42 h after transfection, luminescence was detected with the Promega NanoGlo Dual-Luciferase assay system (N1630) and a 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 this was then normalized to the values obtained from cells expressing the empty vector (EV). 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 (> 140 million) possible single-nucleotide substitution variations across all human promoters within 1 kb of the representative TSS on both sides. Gene-wise variation potentials were summarized by two measures: directionality, which was computed as the sum of predicted log(fold change) values for all mutations per gene, and magnitude, which was computed as the sum of all absolute predicted log(fold change) values. We found that genes with negative variation potential directionality (i.e., mutations that tended to cause a decrease in tissue-specific expression) were actively expressed in the modeled tissue (Fig. 4b and Supplementary Fig. 11). We inferred that these genes were under positive evolutionary constraint and thus were vulnerable to inactivating mutations. We found that expression of genes with positive variation potential (i.e., mutations that caused an increase in tissue-specific expression) was repressed in the modeled tissue (Fig. 4b and Supplementary Fig. 11). We inferred that these genes were under negative evolutionary constraint and thus were vulnerable to activating mutations. Note that evolutionary constraints could not simply be inferred from gene expression levels (Supplementary Fig. 14).

We used 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 Fig. 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. Although the true null distribution was unknown, a conservative estimate of the empirical null distribution could be obtained by assuming the other components were observed only 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 the overall distribution of gene-wise average predicted effects, we could then compute the probabilities of genes belonging to the positive or negative constraint components. Probability > 0.5 for each component was 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 the human–chimpanzee common ancestor, the proportion of divergent sites was computed from the high-confidence divergence sites from ref.51. For estimating divergence among ten primate species (including humans), we computed the proportion of accelerated evolution sites based on primate phyloP scores (see URLs). Accelerated evolution sites were decided with the threshold of phyloP < −2.3, which corresponds to $P < 0.005$ 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 was ambiguous (different GWAS studies pointing to conflicting risk alleles) were excluded. Only GWAS 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 the gene level, as it is often challenging to achieve accurate quantification of TSS expression levels from short-read sequencing. Even though training an expression model should ideally utilize TSS-specific expression quantification, gene-level expression as measured by RNA-seq or microarrays is usually a good approximation of transcription level from the representative TSS of each gene[52,53], and this is usually measured with higher sequencing depth. We determined the representative TSS for each Pol II–transcribed gene based on the quantification of aggregated cap analysis of gene expression (CAGE) reads in the FANTOM5 project[54]. Specifically, a CAGE peak is associated with a GENCODE gene if it is within 1,000 bp of a GENCODE v24 annotated TSS (lifted to GRCh37 coordinates). Peaks within 1,000 bp of rRNA-, snRNA-, snoRNA- or tRNA-encoding 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 levels as compared to the annotated gene start positions ($P = 5.7 \times 10^{-8}$; Supplementary Fig. 18).

Statistical analysis. All details of the statistical tests are specified in the associated text or figure legends. Association between two variables was tested via linear regression (or logistic regression if one of the variables was categorical) with the null hypothesis that the slope coefficient was 0. For comparing evolution and population genetics signatures between putative positive and putative negative constraint genes, we tested the null hypotheses that the coefficient of the interaction term was 0 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.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Data and code availability. The data supporting the findings of the study are available within the paper and its supplementary information files. The ExPecto web portal for tissue-specific gene expression effect predictions for human mutations is at http://hb.flatironinstitute.org/expecto. The source code for running and training the ExPecto models is available at https://github.com/FunctionLab/ExPecto.

References

45. de Souza, N. The ENCODE project. *Nat. Methods* **9**, 1046 (2012).
46. Bernstein, B. E. et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28**, 1045–1048 (2010).
47. Chen, T. & Guestrin, C. XGBoost. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 (ACM, San Francisco, 2016).
48. Bühlmann, P. Boosting for high-dimensional linear models. *Ann. Stat.* **34**, 559–583 (2006).
49. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
50. Efron, B. Size, power and false discovery rates. *Ann. Stat.* **35**, 1351–1377 (2007).
51. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
52. Gonzàlez-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.* **14**, R70 (2013).
53. Uhlen, M. et al. Tissue-based map of the human proteome. *Science* **347**, 1260419–1260419 (2015).
54. Forrest, A. R. R. et al. A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).

nature research

Corresponding author(s): Olga G. Troyanskaya

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Experimental design

1. Sample size

   Describe how sample size was determined. — We did not determine sample size with statistical methods.

2. Data exclusions

   Describe any data exclusions. — No data were excluded from the analysis.

3. Replication

   Describe the measures taken to verify the reproducibility of the experimental findings. — All experimental results reported are successfully replicated for at least three biological replicates.

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups. — A grouped design was not used

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis. — A grouped design was not used

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

   For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	Test values indicating whether an effect is present *Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.*
☒	☐	A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	The source code is now available at http://hb.flatironinstitute.org/expecto or https://github.com/FunctionLab/ExPecto

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.	No

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	N/A

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	Low passage BE(2)-C directly from ATCC (CRL-2268)
b. Describe the method of cell line authentication used.	No method used
c. Report whether the cell lines were tested for mycoplasma contamination.	Cell line was tested for mycoplasma and was negative
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	*Provide a rationale for the use of commonly misidentified cell lines OR state that no commonly misidentified cell lines were used.*

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.	N/A

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	N/A