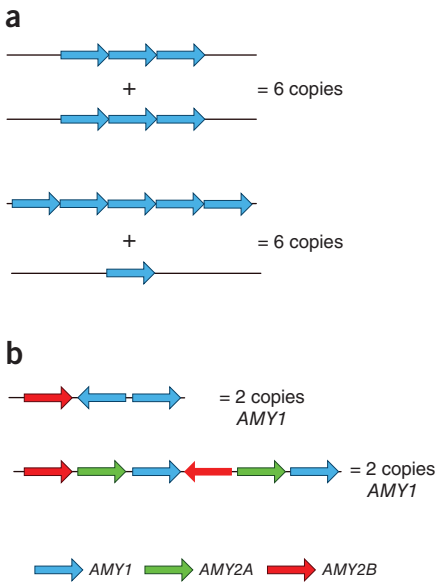

Figure 1 Amylase locus variation. (**a**) The same diploid copy number can be derived from different allele combinations. (**b**) Haplotype structure analysis can show that there is more variation than just copy number: for example, two haplotypes with the same copy number of a gene can have different structures, potentially leading to different levels of gene product.

kilobases in length[11]. The authors constructed models for eight different haplotypes, three of which had not been described previously. The 4 most common haplotypes in a panel of 480 European-ancestry individuals consisted of single copies of *AMY2A* and *AMY2B* and odd numbers (1, 3, 5 or 7) of *AMY1* in tandem arrays. These structures of the locus match those described earlier in 2015 (ref. 12). Less common haplotypes contained differing copy numbers of *AMY2A* and/or *AMY2B*, as well as even number copies of *AMY1* that were not always part of tandem arrays.

AMY1 copy number was then determined in 3 different cohorts (totaling 3,500 individuals) selected for divergent BMI scores. SNPs (unlinked to *AMY1*) previously associated with BMI were analyzed as positive controls[13], and, indeed, these variants showed significant association. Despite having more than sufficient power to detect an effect size equivalent to that previously reported for *AMY1* copy number, no significant association with BMI was observed. The authors considered the most likely explanation for this to be the different assays used in the two studies.

Amylase variation and BMI

The fact that the most common haplotypes defined in both this report and Carpenter *et al.*[12] contained an odd number of *AMY1* copies explains the finding that the majority of individuals possess an even number of *AMY1* copies when the copy numbers on the two chromosomes are combined. That the qPCR-based studies did not observe this supports the conclusion that the copy number estimates made with qPCR were not precise. This does not, however, preclude a role for amylase in BMI.

One feature that commonly used genotyping methodologies have in common is that they provide a single copy number estimate, based on the sum of two alleles. Ideally, the copy number of each allele should be determined separately and the resulting values then combined, as a diploid copy number can be the product of different haplotype combinations (**Fig. 1a**). However, structural variation at the amylase locus is not restricted simply to copy number: different *AMY1* copies are not always part of tandem arrays, and some gene copies can be inverted. It has been shown that the amount of enzyme produced correlates with copy number[5], but it may be the case that positional effects, as well as other genetic variants, also have a role in determining the amount of amylase produced (**Fig. 1b**). As such, it is possible that any association between *AMY1* and obesity is in fact due to specific haplotypes rather than simply the diploid copy number. There is also a potential role for an amylase gene other than *AMY1*, given that both *AMY2A* and *AMY2B* also show CNV. Indeed, Carpenter *et al.*[12] speculated a role for *AMY2*, on the basis of correlation with *AMY1* copy number.

With improved sequencing technology, it has become possible to generate longer read lengths. The Pacific Biosciences RS II Sequencer can produce read lengths that average ~15 kb in length (with individual reads of up to 60 kb), and the power of this technology for resolving regions of the genome intractable to analysis with shorter reads has already been shown[14]. As read lengths increase further, it will eventually be possible to sequence across regions such as the amylase locus (which can be >600 kb in length[12]), allowing a qualitative determination of copy number and removing the uncertainty associated with quantitative analyses.

COMPETING FINANCIAL INTERESTS
The author declares no competing financial interests.

1. Pronk, J.C. *et al. Hum. Genet.* **60**, 32–35 (1982).
2. Groot, P.C. *et al. Genomics* **5**, 29–42 (1989).
3. Perry, G.H. *et al. Nat. Genet.* **39**, 1256–1260 (2007).
4. Falchi, M. *et al. Nat. Genet.* **46**, 492–497 (2014).
5. Cantsilieris, S. & White, S.J. *Hum. Mutat.* **34**, 1–13 (2013).
6. Cantsilieris, S. *et al. BMC Genomics* **15**, 329 (2014).
7. Aldhous, M.C. *et al. Hum. Mol. Genet.* **19**, 4930–4938 (2010).
8. Usher, C.L. *et al. Nat. Genet.* **47**, 921–925 (2015).
9. Handsaker, R.E. *et al. Nat. Genet.* **43**, 269–276 (2011).
10. Hindson, B.J. *et al. Anal. Chem.* **83**, 8604–8610 (2011).
11. Hastie, A.R. *et al. PLoS ONE* **8**, e55864 (2013).
12. Carpenter, D. *et al. Hum. Mol. Genet.* **24**, 3472–3480 (2015).
13. Speliotes, E.K. *et al. Nat. Genet.* **42**, 937–948 (2010).
14. Chaisson, M.J. *et al. Nature* **517**, 608–611 (2015).

Running spell-check to identify regulatory variants

Martin Kircher & Jay Shendure

A major challenge in human genetics is pinpointing which non-coding genetic variants affect gene expression and disease risk. A new study in this issue describes a broadly applicable approach for this task that explicitly models cell type–specific regulatory motifs and generates variant effect predictions that are more accurate and interpretable than those of alternative tools.

Our capacity to sequence human genomes has vastly outpaced our ability to interpret the resulting catalogs of genetic variants[1]. Thousands of haplotypes have been found to contribute to complex disease risk through genome-wide association studies (GWAS), but we rarely know which variants or genes underlie these associations. The path forward is being partly illuminated by large-scale efforts such as the ENCODE, BLUEPRINT, and Roadmap Epigenomics Projects, which perform diverse biochemical assays in hundreds of cell types or tissues at a genome-wide scale. Genetic variants associated with complex diseases tend to occur in regions that bear the marks of active regulatory

Martin Kircher and Jay Shendure are in the Department of Genome Sciences, University of Washington, Seattle, Washington, USA.
e-mail: shendure@uw.edu

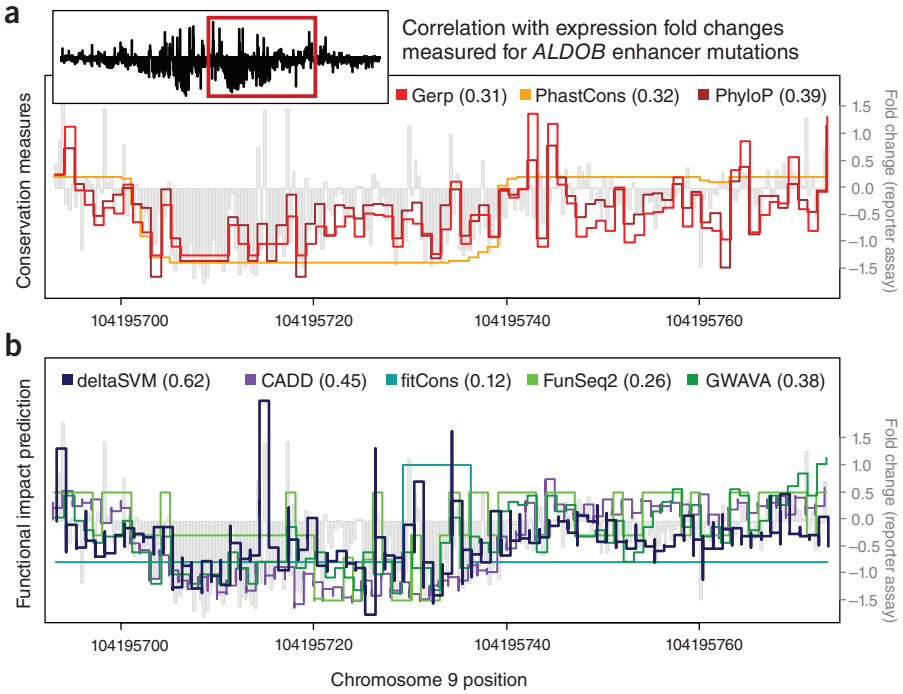

Figure 1 Correlation of conservation measures and functional effects with gene expression. (**a**,**b**) Conservation measures (Gerp[13], PhyloP[14], PhastCons[15]) (**a**) and functional impact scores (CADD[6], deltaSVM[3], fitCons[8], FunSeq2[5], GWAVA[7]) (**b**) overlaid with expression fold changes (gray bars) for an *ALDOB* (aldolase B, fructose-bisphosphate) enhancer as determined with a massively parallel *in vivo* reporter assay[10]. Pearson correlation values for the whole region are provided in parentheses for each method.

elements in cell types related to that disease[2], but new methods are needed to precisely identify which variants causally underlie genome-wide associations, as well as the mechanisms by which they do so. It also remains possible that regulatory variants underlie a substantial proportion of Mendelian disorders that are not resolved by exome sequencing, although this is challenging to explore given that there are 100-fold more variants in the genome than in the exome. To tackle these issues, Michael Beer and colleagues[3] developed a method that predicts the functional effects of regulatory mutations from only DNA sequence and open chromatin information.

Many computational tools have been developed for predicting the effects of coding variants. These predictors exploit evolutionary conservation and the physicochemical properties of amino acids, and some also score the effects of variants on well-defined sequence signals (for example, splice sites). However, as we broaden our scope to the effects of regulatory variants, evolutionary constraint is a poorer guide, because of both the rapid evolution of *cis*-regulatory elements[4], and the fact that the variants that underlie known genome-wide associations are at allele frequencies inconsistent with large deleterious fitness effects. Nonetheless, we and others have developed approaches that integrate evolutionary constraint, functional annotations, and/or knowledge of the sequence preferences of DNA binding proteins to score the relative impact of non-coding variants (for example, FunSeq[5], CADD[6], GWAVA[7] and fitCons[8]). However, these predictions are far from perfect, and it remains unclear how to best predict regulatory effects that are only present in specific cell types or tissues, rather than at the organismal level.

What's different?

Instead of using evolutionary information or the coordinates of functional annotations, Lee *et al.*[3] analyzed the sequence composition of regulatory regions in the relevant cell or tissue type. Specifically, their method contrasts the abundance of gapped *k*-mers in a training data set of regions of open chromatin, for example, as defined by chromatin immunoprecipitation–sequencing (ChIP-seq) or DNase-seq data sets, with a set of random regions matched for their base composition and repeat content. The gapped *k*-mers are sequence strings of length 10, with up to four uninformative positions and at least six informative bases. They learn weights—that is, relative representations in the training data—using a support vector machine (SVM), which serves as a predictor of cell type–specific regulatory elements[9]. Then, by accumulating the weight changes incurred by a particular sequence variant, Lee *et al.*[3] determine a score, deltaSVM, which effectively captures how much the variant alters the surrounding sequence's regulatory potential.

To validate the deltaSVM method, the authors use data from quantitative trait loci as well as from *in vitro* assays of enhancer variants. For example, as shown in **Figure 1**, deltaSVM outperforms all alternative predictors that we evaluated on the task of predicting the results of *in vivo* saturation mutagenesis of an enhancer[10]. The authors show that performance is highly dependent on matching the cell type used for model training with the cell type in which the functional readout is obtained.

An appealing aspect of the deltaSVM approach is that its training only requires open chromatin and control regions for the cell type of interest, which can be generated by a single assay, DNase-seq. No *a priori* knowledge of transcription factor activities or binding specificities are required, nor are multiple biochemical assays. Furthermore, in contrast with evolutionary information, the weights used in the deltaSVM method are directly interpretable with respect to mechanism, as they plausibly reflect the binding specificities of cell type–specific transcription factors. In some cases, these may be immediately relatable to our current knowledge of transcription factor binding specificities, while highly weighted but unassignable *k*-mers can be prioritized for experimental follow-up.

Applications

Whereas causal variants for Mendelian disorders primarily affect coding sequence, most of the signal underlying complex trait heritability partitions to regulatory elements, as, for example, defined by DNase I hypersensitivity[11]. As such, we anticipate that deltaSVM may be broadly useful for developing both fine-mapping and mechanistic knowledge of complex trait associations. As it is blind to allele frequency, it will also stratify the regulatory consequences of rare or *de novo* variants. A limitation of this approach is that it is dependent on the availability of training data corresponding to each cell type of interest. However, chromatin accessibility data sets are increasingly straightforward to generate. As we progress toward comprehensive atlases of regulatory DNA, *in vivo* and for all cell types, deltaSVM may help clarify both the cell types and the regulatory grammar, that is, the functional elements involved and the ways in which they combine to orchestrate a regulatory effect, that are most relevant to specific complex diseases. Finally, even as the cell

type or tissue specificity of deltaSVM predictions is an advantage in some contexts, we anticipate that it may be possible to combine the deltaSVM approach with other information (for example, conservation, allele frequencies and functional annotations) to improve organism-level predictors of deleteriousness or pathogenicity.

The strategy of using sequence-based models to predict variant effects in a cell type–specific manner is powerful, and it may be applicable to other types of regulatory sequence (for example, chromatin marks, splicing). However, a limitation is that deltaSVM only captures local sequence effects. Regional, domain-level and large-scale organization probably influences why a sequence is functional in one context but not in another. On a related note, even as deltaSVM and other tools aim to predict cell type–specific or organismal consequences of regulatory variants, they do not tell us which genes are affected by these consequences. We speculate that in addition to direct measurement of enhancer-promoter interactions[12], improved modeling of the larger-scale organization and interdependencies of regulatory sequences in the human genome will improve the utility and quality of 'model-first' approaches such as deltaSVM for regulatory variant effect prediction.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Cooper, G.M. & Shendure, J. *Nat. Rev. Genet.* **12**, 628–640 (2011).
2. Trynka, G. *et al. Nat. Genet.* **45**, 124–130 (2013).
3. Lee, D. *et al. Nat. Genet.* **47**, 955–961 (2015).
4. Villar, D. *et al. Cell* **160**, 554–566 (2015).
5. Fu, Y. *et al. Genome Biol.* **15**, 480 (2014).
6. Kircher, M. *et al. Nat. Genet.* **46**, 310–315 (2014).
7. Ritchie, G.R., Dunham, I., Zeggini, E. & Flicek, P. *Nat. Methods* **11**, 294–296 (2014).
8. Gulko, B., Hubisz, M.J., Gronau, I. & Siepel, A. *Nat. Genet.* **47**, 276–283 (2015).
9. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M.A. *PLoS Comput. Biol.* **10**, e1003711 (2014).
10. Patwardhan, R.P. *et al. Nat. Biotechnol.* **30**, 265–270 (2012).
11. Gusev, A. *et al. Am. J. Hum. Genet.* **95**, 535–552 (2014).
12. Ma, W. *et al. Nat. Methods* **12**, 71–78 (2015).
13. Davydov, E.V. *et al. PLoS Comput. Biol.* **6**, e1001025 (2010).
14. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. *Genome Res.* **20**, 110–121 (2010).
15. Siepel, A. *et al. Genome Res.* **15**, 1034–1050 (2005).