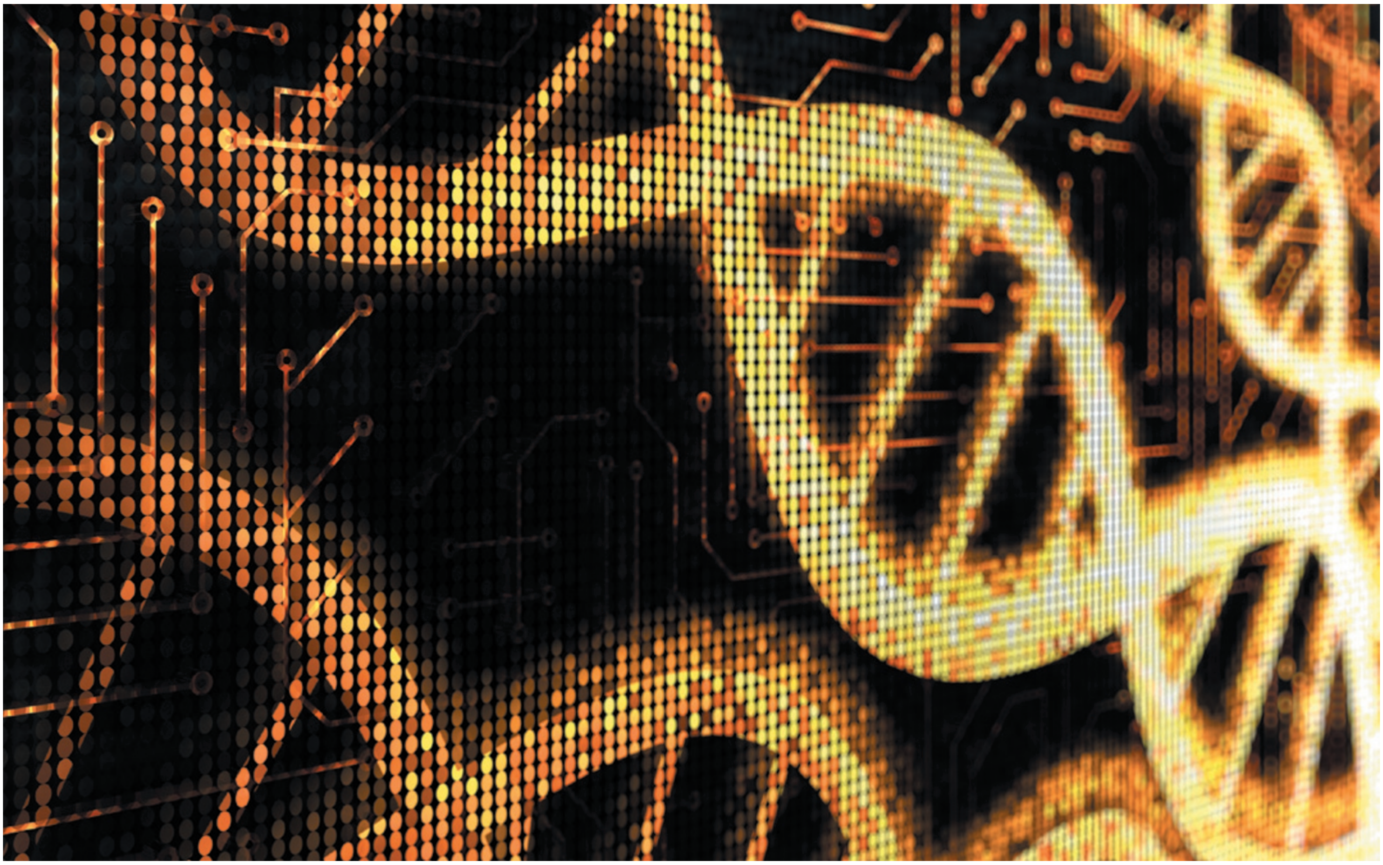
TECHNOLOGY FEATURE

THE CHANGES THAT COUNT

As more mutations are found across the genome, geneticists are focusing on learning which ones are likely to cause human disease, and how.



ALENGO/ISTOCKPHOTO

BY MONYA BAKER

Even before the first draft of the human genome was complete, researchers knew that one genome wouldn't be enough. They needed sequence data from many individuals to reveal the mutations that make people different and sometimes make them ill. Now, tens of thousands of people have had their genomes fully or partially sequenced. Each person's genome contains an average of more than 3 million variants, or differences from the reference genome. A partial sequence, focusing on the 1.5% of the genome that codes for proteins, usually has about 20,000.

For the most part, scientists don't know what those variants do. "The ultimate goal is to sequence a person's genome and make credible predictions just given the list of variants," says Greg Cooper, a genomicist at the Hudson-Alpha Institute for Biotechnology in Huntsville, Alabama. "We're a really long way from that."

Scientists have sorted through the most common variants, using genome-wide association studies to learn which occur more often in people with disease, but these variants tend to have small effects, with the biology behind those effects largely unknown. And as techniques that use sequencing to identify genetic variation become cheaper and more reliable, more rare variants are being uncovered. That is changing the questions that researchers are asking, says David Goldstein, director of the Center for Human Genome Variation at Duke University in Durham, North Carolina. "The field will transition from doing primarily association work to figuring out what implicated variants do biologically."

Disparate strands of research are coming together to do exactly that. A host of increasingly sophisticated algorithms predict whether a mutation is likely to change the function of a protein, or alter its expression. Sequencing data from an increasing number of species and larger human populations are revealing ▶

▶ which variants can be tolerated by evolution and exist in healthy individuals. Huge research projects are assigning putative functions to sequences throughout the genome and allowing researchers to improve their hypotheses about variants. And for regions with known function, new techniques can use yeast and bacteria to assess the effects of hundreds of potential mammalian variants in a single experiment.

ALIGNMENTS AND ALGORITHMS

Many bioinformatics tools rely on evolution to rate how likely a variant is to be harmful. Most focus on identifying the 'non-synonymous' mutations that alter the amino acids that make up the proteins for which genes code. It is expected that the more species have evolved with a certain amino acid in a certain place, the more likely a change is to be harmful. "The idea is that evolution has tested it and that's why you don't see that mutation," says Pauline Ng, a genomicist at the Genome Institute of Singapore. Ng co-wrote an algorithm called SIFT (sorting intolerant from tolerant; http://sift-dna.org), one of the first programs for predicting the effects of protein changes and still one of the most popular. It was originally designed to evaluate one gene at a time, but Ng has updated the protocol to accommodate genomic data files produced by sequencing analyses.

The algorithm first identifies mutations that affect highly conserved amino acids, then predicts whether a particular change is likely to be harmful. To train it for such assessments, Ng used published data that assessed amino-acid changes in a well-studied bacterial protein. That showed how often a change from one particular amino acid to another altered protein function. When researchers run SIFT on their sequencing data, the algorithm uses evolutionary conservation and patterns inferred from that original data set to evaluate whether mutated human proteins are likely to behave in similar ways to their non-mutated counterparts.

"The field will transition from association work to figuring out what variants do biologically."
David Goldstein

Another popular algorithm is PolyPhen (prediction of functional effects of human non-synonymous single-nucleotide polymorphisms; http://genetics.bwh.harvard.edu/pph2), which was co-written by Shamil Sunyaev, a geneticist at Harvard Medical School in Boston, Massachusetts. This algorithm, too, uses evolutionary data in its predictions, but it also incorporates biochemical predictors of stability and spatial structure. Sunyaev trained it using single-gene mutations that are known to cause diseases, reasoning that they did so by disabling proteins.

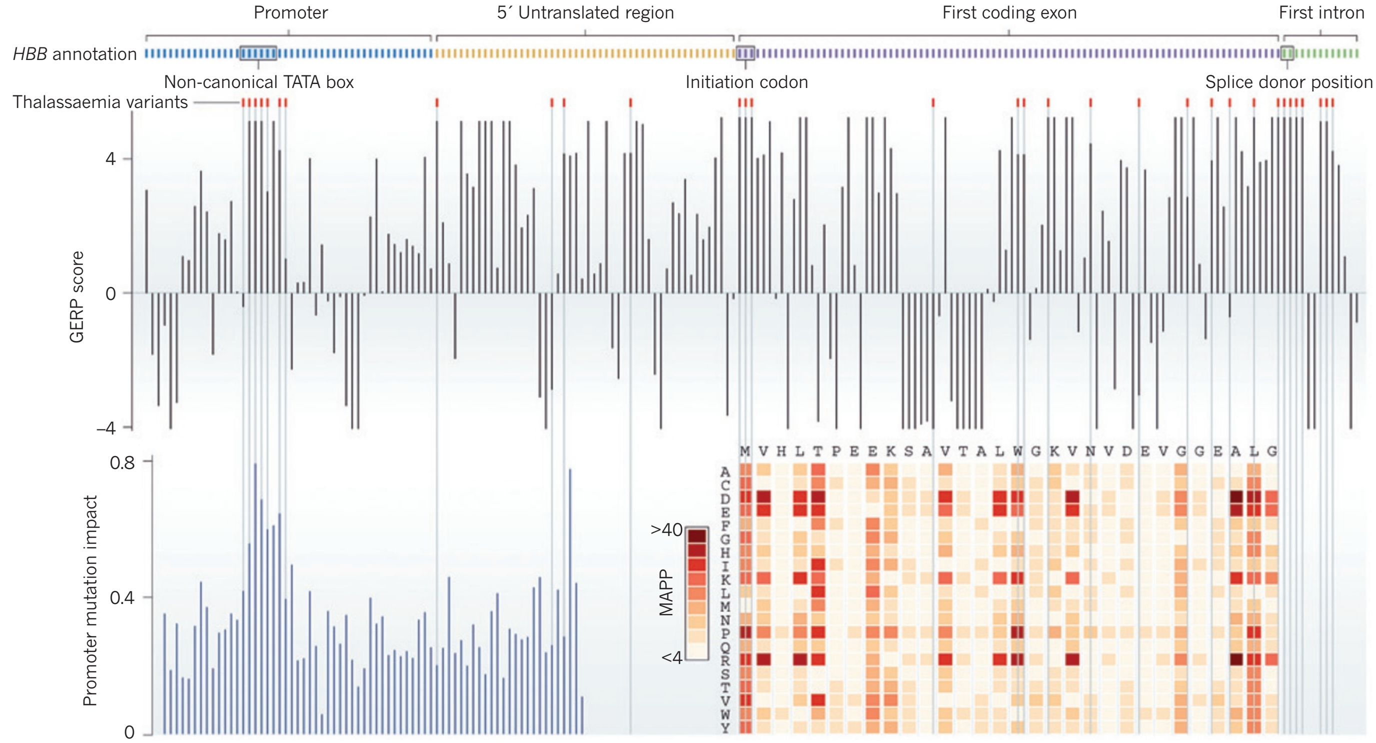
Stephanie Hicks and Marek Kimmel, statisticians at Rice University in Houston, Texas, were part of a team that evaluated[1] the abilities of 4 popular algorithms to predict the effects of 267 well-understood 'missense mutations', which swap one amino acid for another. The algorithms all had accuracies of about 80%. However, even when working from the same 'alignment data' — comparisons of protein sequences — the algorithms made different predictions about the same set of proteins. And Kimmel cautions that algorithms may perform less effectively with mutations that aren't well-known.

Even if algorithms were 100% accurate, knowing that a variant causes a protein to lose function is a very long way from knowing whether it contributes to disease, says Sunyaev. The effects of loss-of-function mutations can be surprisingly minimal, buffered by redundancies in cellular machinery. Algorithms alone are certainly not good enough for clinical diagnostics, he says, and he frets that some clinicians are starting to take an interest in these scores. "This is how I lose sleep at night."

MORE THAN MISSENSE

Even if their predictions were perfect, algorithms that focus on protein sequences would miss many variants that potentially cause disease. Evolutionary analyses indicate that natural selection has conserved five times more base pairs that don't code for proteins than ones that do, which implies that these sequences have some sort of function, even if that is not yet obvious — and mutations in these genomic regions could therefore have a biological effect.

Researchers have now introduced computational tools that use evolution to rank variants in non-coding regions[2]. These include GERP



REF. 2

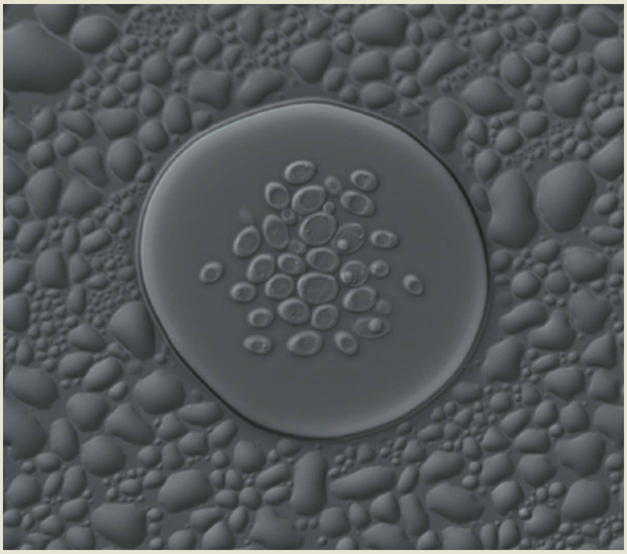
Analysis of variants in coding and non-coding regions of part of a haemoglobin gene (*HBB*). The variants marked in red cause the blood disorder thalassaemia.

Variants in context

The key to human individuality may not be genetic variants so much as the interactions between them. Consider this example: in 2002, a knockout library of more than 5,000 genes from yeast (*Saccharomyces cerevisiae*) found about 1,000 that the microbes literally can't live without[10]. In 2010, Charles Boone, a molecular geneticist at the University of Toronto, Canada, and Gerald Fink and David Gifford, molecular geneticists at the Broad Institute of the Massachusetts Insitute of Technology and Harvard in Cambridge, made a similar library[11] using a second strain of the same species. Startlingly, dozens of 'essential genes' were unique to one strain or the other. And the strains are about as similar to each other genetically as individual humans.

The implications of such studies are frightening, says David Goldstein, director of the Center for Human Genome Variation at Duke University in Durham, North Carolina. "It means that the whole concept of whether variants are pathogenic is not well formulated. When we consider pathogenicity at the level of the individual, we don't always know what we're talking about."

Indeed, researchers have very strong ideas about what contributes to pathogenicity at the population level, but don't necessarily know how to translate them to the individual. Last year, a team at the University of Geneva, Switzerland, characterized[12] an often-overlooked type of interaction using established cell lines for many individuals and data from the international 1000 Genomes Project. "We asked: 'If there was a deleterious coding variant, how likely is it that there is a regulatory variant modifying that effect?'" says Stephen Montgomery, a member of the team and now a geneticist at Stanford University in California. The answer is, pretty likely. For nearly half of the coding variants that the researchers examined, they also found at least one individual who expressed the gene at atypical levels, a situation that could decrease or increase levels of a pathogenic protein and perhaps affect the course of disease.



Yeast cells can be used to demonstrate how the effects of one genetic variant depend on those of other variants.

L. KRUGLYAK LAB, PRINCETON

Researchers at the University of Nottingham, UK, and the Wellcome Trust Sanger Institute in Hinxton, UK, mated a heat-tolerant yeast strain that normally grows on tree bark with a heat-sensitive strain used to make palm wine[13]. They bred the progeny for 12 generations, giving variants on the same chromosome many chances to shuffle and reassort. That helped them to pinpoint loci — defined regions on a chromosome — that contain variants that help yeast to survive. They found around 20 loci containing variants that boost yeast's ability to withstand heat — but surprisingly, one-third of these loci originated in the heat-sensitive strain. "Once you put those mutations in a random background, then you can see their positive effect," says Leopold Parts, first author of the study and now a postdoc at the University of Toronto.

Leonid Kruglyak, a geneticist at Princeton University in New Jersey, has found a way to combine high-throughput genotyping with yeast mating to work out how many spots on a genome contribute to a trait[14]. He says that attributing disease heritability to multiple common variants that each have small effects just doesn't add up. "If you project from the numbers that are being reported," he says, "you end up with preposterous numbers, multiple variants for every single gene in the genome." It will take empirical work to learn the relative importance of common variants, rare variants and the interactions between them, he says.

The problem is that biological experiments are set up to get information about averages, not individuals, says Ben Lehner, a systems biologist at the Center for Genome Regulation in Barcelona, Spain, who is studying how yeast-sequencing data can be used to predict phenotypes. "We talk about the typical effect of an allele in the population, but that is not useful if you want to find out what that means for an individual," he says. M.B.

(genomic evolutionary rate profiling; http://mendel.stanford.edu/SidowLab/downloads/gerp) and phastCons (phylogenetic analysis with space/time models, conservationl; http://compgen.bscb.cornell.edu/phast). Like algorithms that assess protein-coding genes, they evaluate variants on the basis of how often the sequence changes between species. However, because non-coding regions evolve very quickly, sequences can be compared only among mammals. "Even if you go to chickens, nearly all the non-coding stuff won't align," says Cooper, who co-wrote GERP.

And it is not always clear what the rankings mean. Because non-coding regions do not have a corresponding protein, rules regarding amino-acid changes are irrelevant, and there are no data sets appropriate for training such algorithms. "The evolutionary data we do have are informative, but it's early days, so you have to take them with a grain of salt," says Arend Sidow, a genomicist at Stanford University in California, who co-wrote GERP and other predictive algorithms. But algorithms for non-coding sequences can provide evidence that a mutation has an impact by looking at conservation, says Sidow. For example, if a child with a rare disease has an unknown mutation not shared by his or her healthy parents, a score indicating that the mutation is in an evolutionarily conserved region would encourage researchers to examine it more carefully in follow-up experiments.

Alternatively, researchers can consider the results of human-sequencing experiments. One algorithm, VAAST (variant annotation, analysis and search tool; www.yandell-lab.org/software/vaast.html), received a lot of attention last year when researchers used it[3] on just two newly sequenced genomes to pinpoint the mutation that causes Ogden syndrome, a fatal condition linked to the X chromosome in males. The algorithm was also able to re-identify single genes already known to cause some conditions and implicated in more complex diseases[4].

VAAST was developed by Mark Yandell, a geneticist at the University of Utah in Salt Lake City, and Martin Reese, chief executive of genetic-analysis company Omicia in Emeryville, California. It is different from other predictive algorithms that focus on protein-coding and non-coding regions, says Yandell. "Instead of saying, is this conserved?' The algorithm asks, 'How often do we see humans with these variants?'" Unlike many other algorithms, which score each variant as 'probably harmful' or 'probably benign', VAAST provides a ranked list of which variants are most likely to contribute to disease.

"The idea is that evolution has tested it and that's why you don't see that mutation."
Pauline Ng

The algorithm integrates many sources of information: whether a variant has been observed in healthy individuals; whether it occurs in a known functional region; and, for protein-coding variants, what its functional impact is expected to be. When working out whether a single gene is likely to contribute to a condition, it also looks at all the variants that occur in that gene throughout the surveyed population. "You dump all the variants for each gene into a bucket and then see which bucket has the most likely damaging variants. That goes to the top of the list," says Yandell. Future iterations of the algorithm, he says, will consider variants in genes that are associated with common biological pathways.

VAAST is just one in a wave of algorithms to incorporate human-sequencing data. Another is ANNOVAR (http://www.openbioinformatics.org/annovar), which was developed at the Children's Hospital of Philadelphia in Pennsylvania. Knome, a genetic-analysis company in Cambridge, Massachusetts, provides informatics and services for interpreting genomes, and Softgenetics in State College, Pennsylvania, and GenomeQuest, in Westborough, Massachusetts, pluck out variants that might affect patients' health.

But the results of such algorithms can't be trusted without further verification. Predictive algorithms can tell researchers which variants should be flagged up for follow-up studies, but not which ones cause disease, says Cooper. "The best we can do computationally is to prioritize things. It's still going to be a lot of work to nail it."

And there are few ways to assess predictive algorithms, particularly those that go beyond evaluating missense mutations, says John Moult, a bioinformatician at the University of Maryland in Rockville. Moult is one of the co-organizers of the Critical Assessment of Genome Interpretation, a contest in which bioinformatics teams compete to predict a phenotype — an organism's characteristics — from genetic data. Of 13 teams that competed last year, only 2 tried to predict how nucleotide sequences might affect gene expression and splicing.

But the field is still young, says Moult. For algorithms to improve, researchers will need more data — and the data are coming, he says. Not only are more genomes being sequenced, but researchers are working out protocols to share data without compromising patient privacy. Last year, the contest could provide data for only ten whole genomes. This year, Moult expects data for 500.

EXPERIMENTS REQUIRED

Laboratory experiments are essential for verifying the effects of variants, but with so many new variants cropping up, there is currently no way to test them all. "What we need are functional approaches that have a bit of the feel of genomics," says Goldstein. "They need to be scalable; they need to be applied if not to every variant, at least to an awful lot of variants." In particular, Goldstein wants to know whether a variant associated with a gene affects RNA splicing or transcription rates. To find out, he is collecting genome-wide gene-expression data alongside sequencing data. That allows him to find out whether genetic variants correlate with changes in messenger RNA. "It's an affordable additional expense," he says.

Other researchers are developing high-throughput techniques for testing protein variants. Just changing one amino acid at a time, a protein containing 1,000 amino acids would have 19,000 variants. In the past, variants had to be tested individually or in small batches, limiting assays to a few hundred. New methods allow the testing of hundreds of thousands at a time.

Stan Fields, a molecular geneticist at the University of Washington in Seattle, is designing assays that exploit the basic principle of natural selection. He places many variants of a protein-coding gene into viruses or cells that depend on the protein variants that they produce to grow and reproduce, allowing him to interrogate characteristics such as the protein's stability, structure, enzymatic activity and interaction with other proteins. Sequencing can log which variants become more common and which become less so over several generations. "You can come up with all sorts of assays," says Fields, "and the answer comes down to a simple sequence run."

With his postdoc Doug Fowler, Fields has demonstrated[5] that this approach, called deep mutational scanning, can be used to assess the binding activity of hundreds of thousands of variants of the WW domain, a stretch of 40 amino acids that is found in many human proteins and is often important in protein–protein interactions. Fields and Fowler are working out ways to analyse the residues that contribute to protein function, and so learn about general principles of protein design. Fowler is also using the technique to assess which mutations confer drug resistance on Src-kinase, an enzyme implicated in cancer.

It should be possible eventually to assess all the single-amino-acid mutations that could occur in important genes, says Fields. "Then if someone shows up with any mutation, you can say: 'Looking at that particular protein activity, we know what the mutation means.'"

Last year, Dan Bolon, a protein biochemist at the University of Massachusetts Medical School in Worcester, described[6] a similar approach, which he calls EMPIRIC (extremely methodical and parallel investigation of randomized individual codons). He and his colleagues used this technique to test every possible point mutation in a short stretch of Hsp90, a protein that is necessary for yeast growth. The team examined some 500 genetic changes that collectively encoded 180 protein variants. After growing yeast for several generations, Bolon could see which variants enabled the fastest growth, by measuring which showed up the most often in sequencing data. Previous approaches would have required one-by-one testing, but Bolon's method evaluated all the variants at once. An experiment that would normally have taken years was completed in days.

"The best we can do computationally is to prioritize things. It's still going to be a lot of work to nail it."
Greg Cooper

Bolon found that about 15% of amino-acid substitutions that never occurred in evolution grew just as well in his experiments as the wild type, perhaps because effects of those substitutions were too small to matter over the tested time frame, or were irrelevant under the test conditions. Evolution eventually removes both lethal and slightly deleterious variants, but a variant that has an effect only over many generations might make little difference to an individual.

As well as providing direct information on particular proteins, such subtle analyses could be used to train algorithms and improve their accuracy, says Peter Good, programme director for genome informatics at the US National Human Genome Research Institute (NHGRI) in Bethesda, Maryland.

Both Bolon and Fields expect rapid increases in the number and complexity of variants that can be assessed. Bolon is able to vary 100 amino acids at once, the entire length of some small proteins. Already, he can imagine testing all protein variants within small viral genomes. "The ability to look at systematic libraries across an entire genome is just very exciting in terms of understanding the raw evolutionary basis for an entire organism," he says.

Such sequencing approaches can also be applied to regulatory elements. Knowing that a mutation changes a transcription-factor binding site says nothing about how it will affect the binding of the gene-activating protein, says Gary Stormo, a molecular biologist at Washington University School of Medicine in St Louis, Missouri. The protein may bind just as well as without the mutation; hardly at all; slightly worse; or even slightly better. So Stormo has created experimental systems that link transcription-factor binding to cell proliferation. The cells that grow best are those that contain the best-binding DNA, and next-generation sequencing is allowing a more systematic exploration of more variants than ever before. Only two or three years ago, scientists would manually pick 20–50 of the fastest-growing colonies to examine, says Stormo. "We now just scrape the whole plate. You can get millions of examples in a single experiment." Even better, with that many samples, researchers can derive quantitative data, and so show how much better the best-binding sites are.

However, *in vitro* results are far from perfect in predicting *in vivo* binding, says Stormo. "Some of the best sites won't be bound, and there will be binding to other places that you wouldn't expect." The good thing is that differences observed between test tubes and living cells indicate interesting biology. "That tells you we're missing a lot of information, and that's what we want to figure out," says Stormo (see 'Variants in context').

DECODING REGULATORY ELEMENTS

Before they can work out what a variant might do, researchers need to learn whether it occurs in an active part of the genome. Several genome-wide studies are providing crucial clues. The NHGRI's ENCODE project (Encyclopedia of DNA Elements) hopes to map and annotate all functional elements in the genome, and the International Cancer Genome Consortium is mapping genomic changes in cancer. The International Human Epigenome Consortium and the US National Institutes of Health's Roadmap Epigenomics Mapping Consortium are studying features such as DNA methylation and other modifications across the genome in many types of cell, and so are showing which regions of the genome might be functional in particular tissues. Annotation alone will not demonstrate that a variant is pathogenic, but the information can help researchers to design the right experiment, says Good. "The question is knowing why it's pathogenic, that's where the annotation helps you. It's a big difference to say, 'this variant affects a protein-coding region or a promoter active in particular cell types.'"

In work[7] funded by these consortia, Manolis Kellis, a computational biologist at the Massachusetts Institute of Technology (MIT), along with Bradley Bernstein, a pathologist at Harvard Medical School, and their colleagues, mapped 'chromatin states' — sets of chemical modifications to DNA and DNA-binding proteins that distinguish genomic regions. The location of these states varies across different cell types and is correlated with gene expression. By comparing chromatin states on gene promoters, enhancers and other regulatory regions with data on gene expression, the researchers linked regulatory elements to target genes.

"When we are talking about synonymous changes, we can no longer think of them as neutral."
Manolis Kellis

The team then cross-referenced chromatin states with variants that had been associated with specific diseases. This revealed patterns that made sense: for example, variants that had been statistically associated with leukaemia occurred in what chromatin states revealed to be enhancer regions active in leukaemia cells. Similarly, variants thought to affect lipid and triglyceride levels in blood were found in regulatory elements active in liver cells.

Other mapping projects rely on comparative genomics. Last year, researchers based at the Broad Institute of MIT and Harvard in Cambridge, Massachusetts, completed whole-genome sequencing of 20 mammalian species, then analysed[8] these sequences along with those of 9 other mammals that had already been sequenced. This revealed more than 3.5 million evolutionarily constrained elements in the human genome, up from a few hundred thousand that had been previously identified. Still, only about 60% of these could be assigned any putative function. Most of the new elements were located either between genes or in non-coding parts of genes.

Furthermore, even nucleotides in protein-coding genes that would not alter amino acids were under evolutionary constraint, and further analysis suggests that these sites affect RNA-transcript processing, microRNA binding and how chromatin states are established[9]. "When we are talking about synonymous changes, we can no longer think of them as neutral," says Kellis, who was part of the study.

And more regulatory elements are being revealed. Scores of researchers have noticed that non-conserved areas of the genome have activities associated with function. Many such regions are transcribed; others host various DNA-binding proteins. One-half to one-third of 'biochemically active' elements are unique to humans, says Ewan Birney, a bioinformatician at the European Bioinformatics Institute in Hinxton, UK. When the number of these active, non-coding elements was first discovered, their activity was dismissed as an experimental artefact, then discounted as irrelevant noise. But unpublished work shows that many of these regions are in fact evolutionarily conserved in the human population, presumably because they have a function that helps individuals to survive and reproduce.

Of course, changes to evolutionarily conserved sequences do not necessarily contribute to disease, says Birney. But researchers should start thinking about what variation in regulatory regions might do. Six months ago, the Variant Effect Predictor (VEP) tool went live on the Ensembl Genome Browser (www.ensembl.org), which brings together information from several databases, including human-sequencing projects and chromatin signatures across cell types. The tool shows, for example, whether a mutation affects a site that binds known transcription factors.

Other tools are also coming online. Michael Snyder, a geneticist at Stanford, is developing RegulomeDB (www.regulomedb.org), which identifies binding sites and other elements in non-coding DNA. This January, Kellis introduced HaploReg (www.broadinstitute.org/mammals/haploreg/haploreg.php), which brings together data from chromatin-mapping and comparative-genomics studies. Researchers can enter common variants and see whether they fall in a highly conserved region, disrupt a regulatory motif or are associated with a regulatory element in a particular cell type. It provides the same information for common variants that tend to be inherited along with the ones entered.

This is just the beginning of efforts to assign functions to the millions of DNA variants. In time, says Kellis, it will help researchers to pin down the mechanisms that cause disease. "The marriage of human genetics and functional genomics can deliver what the original plan of the human genome promised to medicine." ■

Monya Baker *is technology editor for* Nature *and* Nature Methods.

1. Hicks, S., Wheeler, D. A., Plon, S. E. & Kimmel, M. *Hum. Mutat.* **32,** 661–668 (2011).
2. Cooper, G. M. & Shendure, J. *Nature Rev. Genet.* **12,** 628–640 (2011).
3. Rope, A. F. *et al. Am. J. Hum. Genet.* **89,** 28–43 (2011).
4. Yandell, M. *et al. Genome Res.* **21,** 1529–1542 (2011).
5. Fowler, D. M. *et al. Nature Meth.* **7,** 741–746 (2010).
6. Hietpas, R. T., Jensen, J. D. & Bolon, D. N. A. *Proc. Natl Acad. Sci. USA* **108,** 7896–7901 (2011).
7. Ernst, J. *et al. Nature* **473,** 43–49 (2011).
8. Lindblad-Toh, K. *et al. Nature* **478,** 476–482 (2011).
9. Lin, M. F. *et al. Genome Res.* **21,** 1916–1928 (2011).
10. Giaver, G. *Nature* **418,** 387–391 (2002).
11. Dowell, R. D. *et al. Science* **328,** 469 (2010).
12. Lappalainen, T., Montgomery, S. B., Nica, A. C. & Dermitzakis, E. T. *Am. J. Hum. Genet.* **89,** 459–463. (2011).
13. Parts, L. *et al. Genome Res.* **21,** 1131–1138 (2011).
14. Ehrenreich, I. M. *et al. Nature* **464,** 1039–1042 (2010).