GENOMICS

In search of rare human variants

The 1000 Genomes Project has completed its pilot phase, sequencing the whole genomes of 179 individuals and characterizing all the protein-coding sequences of many others. Welcome to the third phase of human genomics. SEE ARTICLE P.1061

RASMUS NIELSEN

The goal of the 1000 Genomes Project[1] is to find most of the variants in the human genome that have a frequency of at least 1% in the populations studied. The consortium of researchers participating in the project now reports the results of its pilot phase (page 1061 of this issue[2]).

But first let's take a step back. A decade ago, the reference copy of the human genome was sequenced[3,4]. Although that project is undoubtedly one of the greatest scientific achievements of our time, its potential societal impact will be fully realized only if genomic regions that are responsible for various traits of medical importance, such as response to a drug or susceptibility to a disease, can be identified. After the initial sequencing of the human genome, therefore, a second phase of human genomics emerged, focusing on identifying genomic variations responsible for hereditary diseases and other medically relevant traits. Such genome-wide association studies (GWAS) are based on examining the genomes of thousands of individuals for correlations between the presence of genomic variants and the trait of interest.

Many successes have come out of GWAS[5,6], but there has also been some disappointment that perhaps the pickings from these studies have been too slim[7]. For instance, although certain disorders — including obesity, diabetes and cardiovascular disease — are known to have a strong genetic component, their associated genomic variants detected through GWAS cannot explain most of the experimentally identified genetic effects found in affected families. Human geneticists call this problem the 'missing heritability'[7].

There are many possible explanations for the missing heritability, the most popular being the effect of rare variants. GWAS are based on examining a battery of different variants across the genome. Until recently, however, the cost of including both common and rare variants in such studies was prohibitively high, pushing the focus towards identifying common variants that occur at a relatively high frequency in the population. Consequently, if many rare variants, rather than a few common ones, are responsible for a disease, the rare variants would have been missed in most GWAS.

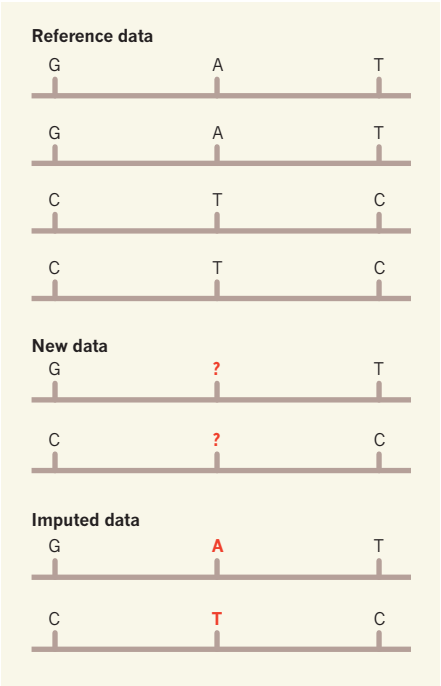

Figure 1 | Gene sequencing by imputation. On the basis of the pattern in a set of reference sequences, the missing nucleotides (indicated by question marks) in a new data set can be imputed. For example, because all sequences in the reference data with a G and a T in the first and third positions, respectively, have an A in the second position, the missing nucleotide in the first sequence of the new data is likely to be an A. Imputation methods are an integral component of the paper[2] reporting the pilot phase of the 1000 Genomes Project.

An obvious solution to this problem is to sequence whole genomes. But this is easier said than done: GWAS require sample sizes of thousands, making whole-genome sequencing extremely expensive. However, computational-biology studies have provided crucial insight that is helping to pave the way for more-comprehensive genomic studies. The idea is that if most of both common and rare variants can be characterized in just a few individuals through whole-genome sequencing, a relatively small battery of variants could then be identified in the remaining individuals in the genome-wide association study, and the pattern of those variants could be inferred computationally on the basis of the few whole-genome sequences.

Sceptics may find this notion — using the data from some individuals to 'invent' data for others — alarming. But if done correctly, this method, called imputation[8], can significantly increase the statistical power of GWAS (Fig. 1). This idea is one of the main motivating forces behind the 1000 Genomes Project.

In the pilot phase of the project[2], the authors used several techniques to sequence the whole genomes of 179 individuals. They thereby generated a catalogue of 8 million previously unknown variants affecting single nucleotides — the building blocks of genes — and around 1 million structural variants due to small insertions or deletions of DNA. The study also presents several new methods for analysing genomic data. For example, it convincingly shows that imputation methods can significantly increase the power of GWAS.

New technologies also allow the protein-coding sequences (exons) within genes to be sequenced specifically. The vast majority of genomic DNA falls outside genes, but many of the most important variants are thought to be located within exons. Exon sequencing therefore provides a cost-effective method for identifying most of the functional variants. The consortium[2] reports exon sequences of 697 individuals from different ethnic groups.

Apart from exon sequencing, another way to contain the cost of sequencing based on GWAS is to sequence genomes at only low coverage. This means that, for each individual, only a limited amount of randomly distributed DNA is sequenced. Although, on average, a genome is sequenced several times using this technique, there may be missing data in any particular genomic region. In fact, low coverage was the approach taken for whole-genome sequencing of the 179 individuals[2].

A disadvantage of low-coverage sequencing is a higher error rate; but this can be reduced, again using imputation methods. Indeed, the consortium's low-coverage data produced an overall error rate of only 1–3% thanks to supplementation with such methods. Imputation-based methods may therefore also be the key to maximizing the utility of low-coverage sequencing data. Characterizing variants in heterozygous sites, which contain two versions

of the DNA, is more difficult, and for them the error rate in the present study varied between 5% and 30% depending on the frequency of the variant.

Given the declining cost of DNA sequencing, future discoveries in human genomics are more likely to be based on a combination of exon sequencing and low-coverage, whole-genome sequencing, rather than on the more traditional techniques. Such DNA sequencing gives access to rare and novel variants, as well as being more suitable for identifying DNA insertions and deletions and, in general, for detecting less-common variants that affect only a single nucleotide.

The remaining question is how to accommodate errors in low-coverage sequencing, because an error rate of even a few per cent can lead to drastically reduced power if not accounted for appropriately[9]. Statistical methods that incorporate high error rates will be an essential component of future genomic-sequencing efforts. But no matter which protocol is used, the focus of the third phase of human genomics will clearly be on whole-genome sequencing. ■

Rasmus Nielsen *is in the Departments of Integrative Biology and of Statistics, University of California, Berkeley, Berkeley, California 94720, USA.*
e-mail: rasmus_nielsen@berkeley.edu

1. www.1000genomes.org/page.php?page=about
2. The 1000 Genomes Project Consortium *Nature* **467,** 1061–1073 (2010).
3. International Human Genome Sequencing Consortium *Nature* **409,** 860–921 (2001).
4. Venter, J. C. *et al. Science* **291,** 1304–1351 (2001).
5. Hindorff, L. A. *et al. Proc. Natl Acad. Sci. USA* **106,** 9362–9367 (2009).
6. Hardy, J. & Singleton, A. *N. Engl. J. Med.* **360,** 1759–1768 (2009).
7. Manolio, T. A. *et al. Nature* **461,** 747–753 (2009).
8. Servin, B. & Stephens, M. *PLoS Genet.* **3,** e114 (2007).
9. Huang, L. *et al. Am. J. Hum. Genet.* **85,** 692–698 (2009).

DRUG DEVELOPMENT

Longer-lived proteins

Short residence times in the bloodstream reduce the effectiveness of protein drugs. Application of an approach that combines protein and polymer engineering prolongs circulation time and increases drug uptake by tumours.

JEFFREY A. HUBBELL

The past 25 years have seen an explosion in the number of approved protein drugs produced by genetic engineering, for treating hormonal, metabolic, immunological, haematological and reproductive disorders, as well as cancer[1]. Scientists initially sought to perfectly copy nature's structural expression of these proteins, leading to many first-generation drugs. Subsequently, protein engineers began to adapt nature's structures, either subtly (for example, by changing a few amino-acid residues to make interactions with a target molecule stronger or more specific) or more profoundly (for instance, by attaching two unrelated proteins to create a protein possessing a combined function that nature never considered). Several second-generation drugs have resulted from such efforts.

One drawback of protein drugs is their rapid clearance from the systemic circulation. Writing in *Proceedings of the National Academy of Sciences,* Chilkoti and colleagues[2] now describe a combined protein- and polymer-engineering approach to prolong protein circulation and enhance drug accumulation in tumours.

The concept of polymer attachment to proteins first arose in the late 1970s, with the demonstration[3] that conjugation of multiple copies of a relatively low-molecular-weight, water-soluble, nonionic polymer, poly(ethylene glycol) (PEG), could prolong the circulation of a therapeutic enzyme. This observation led to a flurry of activity in the 'PEGylation' of protein drugs, several of which have now entered the marketplace[4–6].

An example that illustrates both the benefits and the complexities of PEGylation is interferon-α2a. The drug has been grafted at amine groups on lysine amino-acid residues to a branched, 40-kilodalton PEG chain. Although the protein is grafted with only one polymer chain, the chain can be attached to any one of four sites; as such, the drug is a mixture of four isomers. Grafting increases the hydrodynamic radius, making the drug bulkier to promote longer retention in the circulation[7]. PEGylated interferon-α2a is a very successful drug for treating chronic hepatitis C.

Polymer grafting to protein drugs is associated with many complexities, however, which Chilkoti and colleagues target in their work[2]. One such complexity, as mentioned above, is the possibility of multiple sites of polymer conjugation, leading to a heterogeneous product. A second results from limitations in the size of the polymer chain that can be grafted. Just as it is difficult to find the end of a long rope piled up in a heap, it is difficult to graft the terminus of a long polymer to the surface of a protein. An alternative approach, which is being developed by Chilkoti and colleagues[2,8], is to grow the polymer chain on the protein drug by polymerization; this would, in principle, allow any length of polymer chain to be grafted.

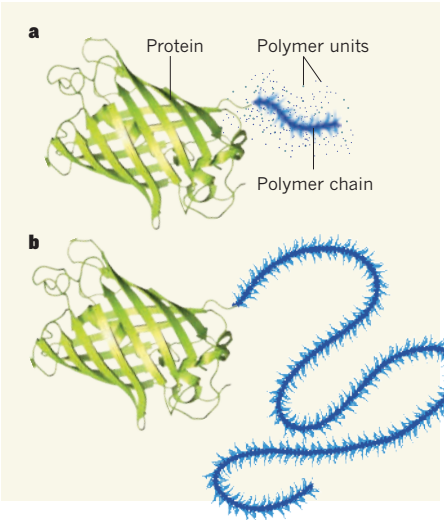

Figure 1 | Drug modification with a feather boa. Chilkoti and colleagues' approach[2] involved growing a long polymer chain from the carboxy terminus of a model protein (green fluorescent protein). The polymer unit used by the authors was oligo(ethylene glycol) methyl ether methacrylate. It is difficult to attach a long polymer chain to the end of a protein, because the two reactive sites only rarely find each other. But a chain much larger than the protein itself can be readily grown by polymerization.

Chilkoti and colleagues' strategy[2] has many advantages. To solve the problem of multiple sites of polymer grafting, they used a protein-engineering trick to place a single chemical group at the carboxy terminus of the protein. They used this group to attach an initiator molecule for a polymerization reaction, selecting a strategically advantageous initiator for a polymerization reaction that gives precise control of polymer length under mild conditions, consistent with the delicate nature of proteins. This allowed the growth of a very long polymer chain, one that looks like a bottlebrush, consisting of a long main chain covered by short PEG chains along its length, with the polymer attached to the protein's carboxy-terminal site.

The result of this convergence of protein and polymer engineering was anything but subtle: the bottlebrush polymer on the carboxy terminus of the model protein increased its hydrodynamic radius almost sevenfold, from 3 to 20 nanometres — an increase in size corresponding to an almost 300-fold increase in hydrodynamic volume. One can imagine the result as being rather like an elfin dancer adorned with an outrageously long and fluffy feather boa (Fig. 1), rather than a few peacock feathers, as would be the effect using previous approaches. The grafted polymer chain resulted in a substantial prolongation in circulation time, which the authors showed to be beneficial in targeting tumours.