GENOMICS

SNPing away at anonymity

New findings challenge the assumption that aggregate genotype data, in which the single-nucleotide polymorphism (SNP) profiles of many people are pooled, conceal the identity of the individuals within that pool.

In the last few years there has been a proliferation of genome-wide association studies, in which relationships are mapped between genomic sequence variants and predisposition to a disease or a trait of interest. These studies depend upon the participation of thousands of individuals in the research process. It has been assumed that it is not possible to determine, based on aggregate single-nucleotide polymorphism (SNP) data, whether or not a particular individual is present in a pool.

However, using a statistical approach, David Craig and colleagues at the Translational Genomics Research Institute now show that this assumption is incorrect.

Statistical analysis can identify whether an individual SNP profile is present in pooled genotype data.

"One way you can understand what we're doing is if you think of a roulette table," says Craig. "The colors it can have are either red or black, and let's say I want to know if the table is slightly biased towards black. If I spin it once, I wouldn't really get a good idea of bias because there isn't much information in a single measurement. But if I spin it half a million times, you can bet I could find some pretty subtle biases." In other words, Craig and colleagues take advantage of the fact that it is possible to monitor hundreds of thousands of SNPs to determine whether or not the SNP profile of a particular individual is present in pooled profile data.

"What we do is essentially a *t*-test," says Craig. They compare the allele frequencies for the person in question to the mean allele frequencies in a reference population and in the pooled test population. When this is done across hundreds of thousands of SNPs, it is possible to assess statistically whether or not the pooled data are shifted significantly in the direction of the person in question. It may even be possible, the researchers report, to use a relative of the person for this purpose. Notably, for the method to work, high-density SNP data for the person must already be available.

PROTEIN BIOCHEMISTRY

EVOLVING A BETTER-EXPRESSING GPCR

Researchers describe a method for evolving G protein–coupled receptors (GPCRs) with greater stability and enhanced expression.

Sixty percent of all drugs target the class of membrane proteins known as G protein–coupled receptors (GPCRs), but the disconnect between their biomedical importance and the number of atomic resolution structures is very large, with only a handful of GPCR structures solved. This certainly does not reflect a lack of interest or attempts, but the tremendous challenges involved in GPCR expression, purification and crystallization. Even when overexpressed in heterologous systems, GPCRs typically express at very low levels in the cell membrane. They are also not very stable in detergents and exhibit conformational flexibility, making them difficult to crystallize. All of these bottlenecks add up, making the structure determination of GPCRs a Herculean task.

Andreas Plückthun of the University of Zurich and his colleagues hope to change this with a new directed evolution method intended to address these bottlenecks. According to Plückthun, the GPCR structure field has so far relied on "finding the lucky break," or basically using brute-force methods to obtain a structure that can in turn be used to model the rest of the family. However, "I always thought that all GPCRs are interesting and we have to find a method that eventually will make all of them amenable for study," he says.

Using the GPCR rat neurotensin receptor-1 (NTR1) as an example, the researchers tested whether they could modify its sequence via directed evolution to make the protein more expressible while still maintaining its function. Such an approach has not been tried before for GPCRs. They constructed a *Ntsr1* library via error-prone PCR and expressed the constructs with N-terminal maltose binding protein and C-terminal thioredoxin fusion partners in *Escherichia coli*. Because GPCRs bind specific ligands, the researchers took advantage of high-throughput fluorescence-activated cell sorting (FACS) to identify cells expressing high levels of NTR1 variants that bound a fluorescently labeled ligand. After several rounds of directed evolution and FACS, they sequenced and analyzed the enriched clones.

They identified a mutant, named D03, which in the *E. coli* membrane exhibited a tenfold increase in expression compared to the wild-type NTR1. This mutant had just 14 nucleotide substitutions throughout the sequences encoding helices and loops, five of which were silent. Agonist binding to D03 was just as strong as to the wild-type NTR1, and the mutant maintained signaling properties when expressed in mammalian cells. Notes Plückthun: "The interesting finding, which to me at least was somewhat unexpected, is that the mutations showed improvement [in expression levels] in every expression system" that they tested, which also included the yeast *Pichia*

Using both simulations and experimental analysis with high-density SNP microarrays, Craig and his team show that it is possible to identify an individual in a mixture of hundreds to thousands of genomic samples, even when the DNA of the person in question is present only in trace amounts (as low as 0.1% of the total). Craig estimates that this may not be the limit of sensitivity. "My guess is it could go down to about one in ten thousand," he says.

In addition to the consequences it will have for forensic analyses, this demonstration has implications for how pooled genotype data will be shared in the future. To protect individual privacy, the US National Institutes of Health and other organizations have already removed aggregate genomic data from public access, instating approval processes for accessing these data, similar to those already in place for accessing individual-level data.

Craig suggests, however, that there is another side to this story. "I hope this will open up the conversation about data sharing," he says. "In my opinion, you really need to share individual-level data, since you lose a lot of power when you just share the aggregate information, and our work now shows that, even in aggregate data, the identity of participants is not completely masked. And I think it's better to work out how to do this responsibly now, when the amount of data is manageable, than in five or ten years."

Natalie de Souza

RESEARCH PAPERS

Homer, N. *et al.* Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* **4**, e1000167 (2008).

pastoris. This suggests that the mutations introduced into the D03 mutant likely confer an overall stability to the protein, resulting in more robust expression. They also purified sixfold more D03 than the wild-type NTR1 from *E. coli*, and the mutant was more thermally stable in detergent-solubilized form, which may promote its crystallization.

Although the researchers have so far only reported results for NTR1, they are currently working on additional mutagenesis of NTR1 as well as testing the generality of the method for other GPCRs. Plückthun notes that the FACS-based selection method is likely to be applicable for evolving better-expressing variants of any membrane receptor that can bind a fluorescent ligand.

They also have yet to test whether their method can streamline the bottlenecks in GPCR crystallization. If such an approach does turn out to be general for evolving more crystallizable variants, it could be extremely powerful and have a major impact on our understanding of GPCR biology. "The interesting part is to understand how the ligand binds, the exact atomic details, and what the differences between an agonist and an antagonist are," says Plückthun. "I just don't think it can really be extrapolated from one model. We have to have an experimental access to basically the whole family."

Allison Doerr

RESEARCH PAPERS

Sarkar, C.A. *et al.* Directed evolution of a G protein–coupled receptor for expression, stability, and binding selectivity. *Proc. Natl. Acad. Sci. USA* **105**, 14808–14813 (2008).

NEWS IN BRIEF

MOLECULAR LIBRARIES

Histone mutant libraries

It is well appreciated that the functions of core histones are largely controlled by combinatorial post-translational modifications, but individual amino acid residues are also important in regulating DNA-damage response, transcriptional activation and heterochromatin formation. Dai *et al.* describe a systematic yeast-based library of histone H3 and H4 mutants, which they used to explore the contribution of each individual residue to nucleosome function.

Dai, J. *et al. Cell* **134**, 1066–1078 (2008).

GENOMICS

Genomic analyses of tumors

To really understand cancer biology it is important to understand all of its genetic and genomic alterations. Several groups have launched large-scale, multidimensional efforts to analyze copy-number variations and gene expression in human glioblastomas and pancreatic cancer. All data of these global genomic analyses are freely accessible.

Jones, S. *et al. Science* **321**, 1801–1806 (2008).
Parsons, D.W. *et al. Science* **321**, 1807–1812 (2008).
The Cancer Genome Atlas Research Network. *Nature*, published online 4 September 2008.

CHEMICAL BIOLOGY

Chemical control of proteins in mice

Banaszynski *et al.* expanded a previously developed method to control protein function in cells. They express a protein of interest as a fusion to an unstable domain. The unstable fusion protein is targeted for degradation, but the presence of a stabilizing ligand protects the fusion protein from degradation, in a dose-dependent manner. By using a viral vector to deliver the fusion protein, they now show they can control protein function in living mice.

Banaszynski, L.A. *et al. Nat. Med.* **14**, 1123–1127 (2008).

STEM CELLS

iPS cells without viral integration

Reprogramming of somatic cells to yield induced pluripotent stem (iPS) cells has only been achieved so far using technology that requires viral integration into the host cell genome. This poses problems for the safety of the approach, particularly in a clinical setting. Stadtfeld *et al.* now show that transient expression of Oct4, Sox2, Klf4 and c-Myc from non-integrating adenoviral vectors can reprogram mouse somatic cells to pluripotency.

Stadtfeld, M. *et al. Science*, published online 25 September 2008.

PROTEIN BIOCHEMISTRY

Evolving streptavidin

The extremely strong interaction between streptavidin and biotin has been exploited for many applications. Levy and Ellington used *in vitro* compartmentalization–based directed evolution methods to generate streptavidin mutants that bind the biotin analog desthiobiotin with the same affinity as the wild-type enzyme but with a 50-fold slower off rate, which may facilitate new applications. The method should also be applicable for evolving other very high affinity protein-ligand interactions.

Levy, M. & Ellington, A.D. *Chem. Biol.* **15**, 979–989 (2008).