Why predicting the phenotypic effect of mutations is hard

By now, we’re probably all familiar with Niels Bohr’s famous quote that “prediction is very difficult, especially about the future”. Although Bohr’s experience was largely in quantum physics, the same problem is true in human genetics. Despite a plethora of genetic variants associated with disease – with frequencies ranging from ultra-rare to commonplace, and effects ranging from protective to catastrophic – variants where we can accurately predict the severity, onset and clinical implications are still few and far between. Phenotypic heterogeneity is the norm even for many rare Mendelian variants, and despite the heritable nature of many common diseases, genomic prediction is rarely good enough to be clinically useful.

The breadth of genomic complexity was really brought home to me a few weeks ago while listening to a range of fascinating talks at the Genomic Disorders 2013 conference. Set against a policy backdrop that includes the recent ACMG guidelines recommending opportunistic screening of 57 genes, and ongoing rumblings in the UK about the 100,000 NHS genomes, the lack of predictability in genomic medicine is rather sobering. For certain genes and diseases, we can or will be able to make accurate and clinically useful predictions; but for many, we can’t and won’t. So what’s the problem? In short, context matters – genomic, environmental and phenotypic. Here are six reasons why genomic prediction is hard, all of which were covered by one or more speakers at Genomic Disorders (I recommend reading to the end – the last one on the list is rather surprising!):
(1) The association between a genotype and disease phenotype may be weaker than we think. Apart from obvious stumbling points like small study size (which has largely been mitigated by increasingly large sample sizes and meta-consortia), ascertainment bias remains a problem particularly in rare disease research. Most of our knowledge about the impact and penetrance of rare disease variants comes from individuals and families with a history of the disease in question. We are still relatively ignorant about the effect of actionable mutations, such as those in the BRCA1/2 genes, in the general population. Many putative disease-causing variants in well-respected databases of genomic variation are actually artefacts, or errors in the original publications, and cannot be trusted. In addition, since historically genetic tests have been ordered only for patients meeting a set of diagnostic criteria, the phenotypic spectrum associated with a gene or variant is may be much wider than we currently know, so diagnosis of even Mendelian conditions may be missed.

(2) Composite phenotypes caused by multiple genetic variants may be commonplace. There are a small but increasing number of publications describing a ‘two-hit’ hypothesis in developmental delay (Girirajan NEJM 2012, for example) where two independent copy number variants each account for part of the complete phenotype. Multigene effects are likely to be a common phenomenon, but need large-scale studies of extremely well phenotyped individuals to explore. Composite phenotypes can introduce inherent biases into the literature, where a phenotype becomes associated with one particular variant – the most likely candidate at the time – but is actually caused by another.

(3) We know almost nothing about the dependence of mutations on genetic background. Epistasis (where the expression of one gene depends on another) is the elephant in the genetics lab. We all know it exists, and is important, and yet it’s often explicitly ignored partly due to the inherent difficulties of finding robust and unbiased statistical associations between multiple genes. But both common and rare variation in modifier genes is likely to account for large differences between individuals with a particular mutation, from altering the phenotypic severity to preventing the disease completely. Uncovering and understanding modifier genes is crucial for prediction.

(4) Non-coding DNA may be important for regulating gene activity. In the wake of ENCODE, it’s clear that a large proportion of the genome is transcribed (though not translated) and may be functional. Through a number of complementary mechanisms, the 98.5% of the genome that
doesn’t code for genes plays a crucial role in regulating gene expression – how, where and when individual genes are turned on or off in specific cells. Since many of the GWAs hits for common diseases lie in the non-coding regions of the genome, it is likely that variation here affects the amount of a gene product produced, rather than altering its actual chemical composition. However, most studies currently focus on gene-targeted sequencing for very valid practical reasons – it is currently much cheaper (and will remain so for the foreseeable future) and our ability to interpret variation in coding regions vastly outstrips our ability to understand the effect of variation in non-coding regions.

(5) Gene-environment interactions and epigenetic factors are poorly understood and hard to study systematically in human populations. Again, we know that the genome interacts with the environment in a number of ways, sometimes resulting in semi-permanent chemical modifications (somatic mutations, cytosine methylation, chromatin remodelling, etc.), but robust associations are hard to come by. Some archetypal genetic diseases have a major environmental component – PKU being the most obvious example, where both the mutation and a dietary source of phenylalanine are required for individuals to manifest the disease. The environment can affect which genes are expressed, how they interact, and whether mutations have any phenotypic consequences. However, it is currently much easier to assay someone’s genome than it is to systematically and longitudinally measure their environmental exposures!

(6) There is intrinsic underlying variability in gene expression. Even after accounting for all of that, the same genotype in the same genome exposed to the same environment can still produce a different phenotype! Ben Lehner’s fascinating talk at Genomic Disorders surprised most of the audience by showing that genetically identical worms in the same controlled environment exhibit enormous variation in gene expression levels – to the extent that a single loss-of-function mutation killed some worms but had no effect on others. This is caused by natural variation in the expression of another gene, which either mitigates the effect of the mutation at high levels or makes the worm particularly susceptible to its effect at low levels. What causes this natural variation is just speculation at this point, but could be a maternal factor that was present in the egg (RNA or peptide, for example). Even tiny differences in the levels of these small molecules can have profound knock-on effects on gene expression. Although worms are obviously substantially less complex than humans, there is no reason to believe that such inherent differences don’t also exist

3
and play a major role in human development and susceptibility to disease.

4 Responses to “Why predicting the phenotypic effect of mutations is hard”

• Gholson Lyon (@GholsonLyon) 26/04/2013 at 20:51 This is so incredibly obvious to me that I find it surprising that some people are only somehow just realizing this, and it speaks volumes regarding the paradigm of genetic determinism that is forced down people’s throats from textbooks and other simple-minded venues. There are ~6 billion nucleotides of DNA in every cell of the human body, and there are 25–100 trillion cells in each human body. Given somatic mosaicism, epigenetic changes, and environmental differences, no two human beings are the same, and therefore the expressivity of ANY mutation will be different in each person. Therefore, I would like to get to a world of millions of whole genomes shared and analyzed for numerous additive, epistatic interactions and gene X environment interactions, so that we can only then begin to make any reliable predictions for any one human being. We need to sequence and collate the raw data from thousands and then millions of exomes and genomes, so that we can actually begin to really understand the expressivity patterns of any mutation in the human genome in any one person. I talked about this over at the Nature blog regarding the recent ACMG guidelines as well: Also, read Keller’s book on this topic as well: Mirage Space Between Nature and Nature

• Ruslan Dorfman 26/04/2013 at 21:21 Very interesting post, and I agree with all the points, however I need to add one more. Even for classical genetic diseases like Cystic Fibrosis there is no clear cut–off between “disease” and “norm”. Some mutations have stronger impact than others and are more penetrant leading to classical pancreatic insufficient CF, while other mutations cause milder CF pancreatic sufficiency, yet some variations cause mild diseases that are now coined as “CFTR associated disorders” and include infertility, lild ling disease or pancreatitits. So it important to emphasize that there is a gradient of clinical manifestations for many disorders, and one needs to take this in consideration when running large scale genotype–phenotype analyses. In CF only after 20 years and scores of clinical data we can begin to analyze such correlations in more reliable manner. So check out CFTR2 database and our publications on in silico prediction models for CF mutations.

• David Poster 26/04/2013 at 23:36 You are joking right about “98.5% of the genome that doesn’t code for genes plays a crucial role in regulating gene expression “? Apart from that very thoughtful post. very thoughtful post