

Whole Genome Sequencing May 2011

There has been a lot of information recently in the lay press and scientific and specifically genetics literature about the practice and promise of whole genome sequencing....lots of speculation about how it will change medicine. In addition to the individuals participating in the 100 genomes project, a number of private citizens have opted and paid to have themselves sequenced (for example singer Ozzy Osborne and actress Glenn Close), and other individuals or individual's tumors are being sequenced as part of research protocols. However, as the price of DNA sequencing falls, how and when will this really enter the mainstream for clinical practice? That is anyone's guess, but what has become evident in the past year is that the price of the sequencing itself is only one important consideration in the process of making DNA sequencing technically accurate, clinically valid, and cost effective...or at least affordable to include in the myriad possible tests that clinicians can order and use in their clinical decision making.

Once a genome is "sequenced", which is performed by actually sequencing millions of short sequences that are generated from an individual's DNA specimen, the shorter sequences must be assembled into a whole, aligning the redundant sections so as to locate the presence of sequencing errors and correct them, so first of all, the technology is not yet 100% accurate. In addition, each short section needs to be sequenced multiple times, again to detect errors in the data. Newer sequencing methods, which are cheaper in terms of a 'per base pair' cost, also sequence shorter DNA strands, so each section of DNA must be done more times.....called X-fold coverage (ie. 6-fold, 30-fold, etc). There are approximately 3 billion base pairs in the human haploid genome and each person has twice this amount of DNA that is potentially going to be sequenced.

There was a very interesting article published in *Genetics in Medicine* recently¹, showing the scale of the interpretative challenge: the average genome contains 2.8 to 4.2 MILLION single nucleotide variants or SNPs. That's per individual. Of these, 38% lie within genes. Of these, 36% lie within introns (the genetic code within a gene that does not actually code for protein, but which may have regulatory or chromosome structure functions) and 2% lie within exons; so of the total, 0.64% are within protein coding regions. What this results in is between 19,000 and 26,000 variants per person that could cause or impact a person's risk for disease....if one only looks at the coding portion of the genome....what is being termed the "exome". This does not include the possible mutations that lie outside of coding regions; sequences that may have important regulatory functions that we don't yet know about.

The price of DNA sequencing is falling, but where the cost is going to potentially lie and what may therefore most greatly impact test availability is in two main areas: interpretation of data and data storage. Interpretation entails someone with medical and genetic knowledge going through the variants (hopefully with the help of a computer application specific to the task), looking for one or more variants that could have known clinical impact. What can be important to remember, however, is that as clinical knowledge and our knowledge of the genome

expands, what is thought to not be a variant of clinical significance today may be one next month or next year, or one thought to be associated with disease may turn out to be benign. Another potentially significant cost will be in the realm of data storage. Some experts have said that patient sequencing data should not be stored, but that if a patient needs to be re-sequenced in order to answer a different clinical question months or years later than when he/she was originally sequenced, that re-sequencing would be cheaper than storing this huge amount of data for extended periods of time.

In addition, what is clear from this paper and other recently published articles² is that there are a number of errors in current genetic/genomic databases; since these are used to compare a patient's sequence with 'normal'; errors in these public databases will potentially result in errors in interpretation and possibly diagnosis. The clinical genetics and genetics research community has a lot of work to do before whole genome sequencing will be clinically useful and common; stay tuned for changes in this rapidly evolving field.

1. "Global analysis of disease-related DNA sequence variation in 10 healthy individuals; Implications for whole genome-based clinical diagnostics". *Genetics in Medicine*, 13(3):210-217, 2011.
2. "Effect of Direct-to-Consumer Genomewide Profiling to Assess Disease Risk" *NEJM*, 364:524-534, 2011.