Precision Medicine: Cancer and Beyond

2016 Whitehead Institute Spring Lecture Series for High School Students

Using bioinformatics to advance precision medicine

http://jura.wi.mit.edu/bio/education/HS2016/

Exercise 1: Browsing James Watson's genome

We'll use the free IGV Genome Browser (http://www.broadinstitute.org/igv/) to look at his sequenced genome to get a peek at the first steps of precision medicine.

1. On the top left pulldown bar, select the version of the human genome to be "Human (hg38)". If you use another version of the genome, the coordinates won't match up and the genome-mapped DNA sequences won't make sense.
2. Note that the numbers across the top refer to chromosomes, and the blue graph (one set of data called a "track") at the bottom shows where the genes are. Since we start out looking at the whole genome, the "Gene" track doesn't make much sense.
3. Click on 6 to go to chromosome 6. The top now shows a graphical representation of chr6.
4. On the top right, zoom in by clicking on the "+" in the box. The red box on the chr6 figure now indicates the zoomed-in region. Individual genes should start appearing at the bottom. The blue boxes are the gene exons, connected by lines (spanning the introns) with and arrowheads indicating the gene direction ("forward" if to the left or "reverse" if to the right). If you keep zooming in, the actual genome DNA sequence appears. This is the "reference genome" sequence assembled from several different people.
5. To load James Watson's genome in the browser, go to File > "Load from File…" and navigate to the file **Watson_chr6.bwa.bam**. (IGV also needs a file called Watson_chr6.bwa.bam.bai, which should be in the same folder.) If it loaded OK, you should see some gray lines in a new track. These are sequenced DNA fragments of his genome. If you can't see anything, zoom in more.
6. Let's concentrate on how Watson's genome differs from the reference genome. This can be due to
    a. deletions (black horizontal bars indicating missing DNA letters),
    b. insertions (purple vertical bars [with a dot in the middle] indicating extra DNA letters in his genome), and

c. single-nucleotide variants (SNVs, indicated by colored vertical bars, which change to letters when you zoom in)

7. Many of these differences are due to technical reasons. This is typically the case if a position represented by overlapping reads shows only rare differences.
8. Try to find a genome location with a consistent difference between Watson's genome and the reference genome. Here are some examples. You can enter them into the box before the "Go" button.

chr6:41,670,335-41,670,375
chr6:41,674,433-41,674,473
chr6:85,624,634-85,624,674
chr6:41,671,536-41,671,576

Why are some sites only partially different from the reference genome?

9. We'd go crazy is we had to find all of the SNVs manually, so we ran a computer program to identify them automatically. We'll load that set of SNVs as a new track. Go to File > "Load from File…" and load **Watson.DP_5.SNVs.bwa.vcf.sorted.bed**. Zoom out and select some of these automatically found SNVs. Did the computer program do a good job?
10. Are these positions where Watson's genome is different from the reference genome expected or unexpected? To help answer this question we can load sites of known genetic variation (SNPs, defined as SNVs where at least 1% of the population has a different DNA letter). Go to File > "Load from File…" and load **SNP144.sorted.bed** This track also displays the rs ID for each common variant (SNP), and in some cases, these can be used to mine information about predicted/expected effect of the SNP. Note that most of Watson's variants are known variants.
11. Let's zoom in on just one variant (rs1799971) that brings up some interesting issues. It's in the region of chr6:154,039,642-154,039,682. Searching the rs ID on the web should bring you to SNPedia, which describes the expected effect of having different alleles (DNA letters) of the site. Remember that this is only one of 3 million SNVs in Watson's genome. How would you react to a finding like this in your genome? Would you want to publicly share this information? Why or why not? Watson's genome is one of few that are publicly available. Most sequenced genomes – and there lots of them – are available only to doctors and/or specific researchers. People have very different opinions on privacy of genomic information! Another interesting SNP (rs2802292) is in the region chr6:108,587,295-108,587,335.
12. [Just to think about] Is there anything in his genome that Dr. Watson should be worried about? Precision medicine researchers and companies can use computers to analyze all of his variants and compare them to variants that are known to cause disease. Sometimes this is very informative – but we still have a lot to learn!