Working With Sequence Data—Part 1

Christopher G. Wilson, Ph.D.

Loma Linda University
Dept. of Pediatrics and Center for Perinatal Biology

April 23, 2015

Outline

1. Review
2. DNA Background
3. Changes in sequence technology
4. Handling the data
5. First Project

Review from last week

- Signal Theory
- Playing with time-series data
- FFT with sine wave(s)

YouTube video to watch

https://www.youtube.com/watch?v=jV4YMQHZmMk

Genetic coding

- DNA, *deoxyribonucleic acid*, is the primary coding molecule underlying all biological organisms (included hereditary and mutated code).
- DNA strands are known as polynucleotides since they are composed of simpler units called nucleotides.
- Each nucleotide is composed of a nitrogen-containing base—either guanine (G), adenine (A), thymine (T), or cytosine (C)
- DNA $\rightarrow$ mRNA $\rightarrow$ Protein

Limits to sequence information

- Sequence information does NOT tell us about function in the whole organism.
- Disease can be *inferred*.
- The genome is NOT the person.

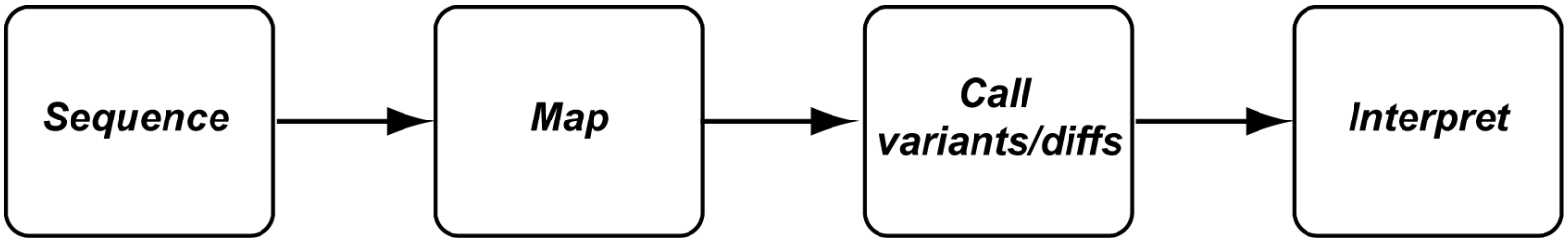
Old Technology

- 3×10^9 characters of DNA in total genome.
- Full genome sequencing cost about $\$2.6 \times 10^9$ in the late 1990s to early 2000s.
- Variant genome sequences were the focus.
- Limited strands to match sequences for efficiency/cost-effectiveness.

New Technology

- Sequence the *whole* thing rather than short stretches of sequence.
- Compare to reference genome (white guys from Buffalo!).
- Re-sequence as needed.
- Now costs about $1000 for a full genome sequence now.

Sequencing Work-flow



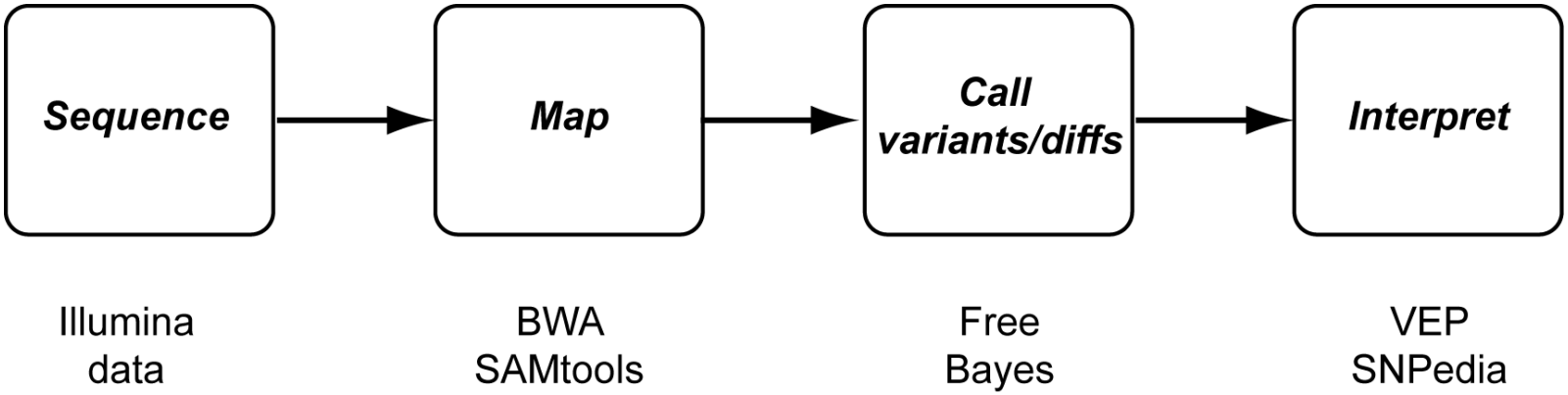
Short-read sequence methodology

- Raw data: `AACCCCTCCATGCTTACAAGCA...`
- Locate sequence in reference and BAM gives you differences.
- Variant detection after mapping gives columns for sequence homology/diffs
- Can take around 1500 CPU hours!

What are we looking for?

- Is variant known to have an effect?
- Is it actually a gene? (signal-to-noise is an issue!)
- Is it a gene with an *obvious* effect?

Tools for each stage of analysis



Data handling

- BAM file can be 108 gigabytes.
- Human genome is around 3 gigabytes or so (raw data).
- Data munging will give files of various sizes.

Data handling (continued)

- *BWA* takes reference genome and set of reads and yields tab-delimited output.
- Must be able to handle noise/nonsense in the data.
- Calling variants with *FreeBayes* (list of places where reads differ from the baseline genome).

Food for thought

- Only 2% of the genome $\rightarrow$ genes.
- Only $\sim$5% is thought to be functional.
- Look at *SNPedia* to find disease/pathologies associated with variants.

First Project

1. Pick a genome and variant (can be whole organism—including bacterium, mammal, whatever—protein, channel, etc.).
2. You can make your own "mutation" by randomizing some base pairs and then chopping the whole genome into random lengths using a *Python* program.
3. Develop a tool-stack based on tools that you learned about in Dr. Brown's talk or based on your own research (suggestions: *BioPython*, *Velvet*, *MIRA*, *MUMmer*, *Mauve*, *Artemis/ACT*, *BLAST+*, *EMBOSS*, *BRIG*, etc.
4. Figure out a visualization and data-storage strategy.
5. Papers in the class repository for background.
6. Write-up is due on May 7^{th}

Some links that may be of use. . .

http://seqanswers.com/wiki/Software
http://software-carpentry.org/v4/shell/
http://quinlanlab.org/tutorials/cshl2013/gemini.html