“Holistic” Data Analysis

- Put together EVERY STEP of the analysis BEFORE optimizing any of the intermediate steps
- Try to imagine what the end result needs to look like and work towards that goal
- Think of an artist drawing portrait → it is a successive refinement of the entire image

Origins of genetic variation 1

- A regular diploid human cell contains 46 chromosomes
- 23 pairs of homologous chromosomes = 46 (22 pairs + sex chromosomes XX(female) XY(male))
- One set of chromosomes inherited from each parent

Note that the reference genome is a “consensus” across all chromosomes of DNA pooled from multiple individuals

Origins of genetic variation 2

Meiosis → four genetically unique haploid gametes that each contain a unique mixture of the genetic code of the maternal and paternal chromosomes of the cell

Genetic diversity → phenotype → natural selection → adaptation → evolution
Origins of human genetic variation 3

- No two humans are genetically identical (not even monozygous twins that start out as such)
- About 30 new variations per generation.
- An allele is one of two or more forms of a gene or a genetic locus
- Both alleles are the same $$\rightarrow$$ homozygotes.
- If the alleles are different $$\rightarrow$$ heterozygotes.

Single nucleotide polymorphisms: SNP

- A single nucleotide — A, T, C or G — in the genome differs between members of a population or chromosome pairs
- Originally defined as occurring at least in 1% of the population (these definitions may shift in time) $$\rightarrow$$ SNV (single nucleotide variant) if observed very rarely
- SNP, SNV $$\rightarrow$$ may fall within coding sequences of genes, non-coding regions of genes, or in the intergenic regions
- DIP: deletion/insertion polymorphism,
- Single Nucleotide Polymorphism Database (dbSNP)
- As of 26 June 2012, dbSNP listed 187,852,828 SNPs in humans.

SNP Calling

- Not nearly as well standardized as one might think

In 2006 the Archon Genomics X PRIZE was to award $10 million to the first team to rapidly, accurately and economically sequence 100 whole human genomes to a level of accuracy never before achieved.

First X-Prize challenge (cancelled)

Archon Genomics X PRIZE
Validation Protocol
Version 2-24-2011
Granger Sutton¹, Edison Liu¹, Victor Jongeneel², and Larry Kedes³

Preamble
The following document is a collective assembly of techniques designed to test the quality and accuracy of 100 whole human genome sequences resulting from the $10 Million Archon Genomics X PRIZE (AGXPRIZE) competition. The purpose of this article is

Organizers of the Archon Genomics X-Prize called off their $10-million competition Thursday, just two weeks before teams were set to begin work on 100 high-quality human genome sequences, in 30 days, at a cost of less than $10,000 per genome.
**Blog: Blue Collar Bioinformatics**

**SNP calling checklist**

- Unique sample or pooled samples?
  - unique samples → the expectation for each allele will be 50%

- External information → SNPs tend to occur in clusters

- Coverage and quality filtering are very important

---

A large number of SNP callers have been published

- Each is good at some aspects (well publicized) – and not so good at others (less publicized)

- SNP calling seems deceivingly simple – why can’t we just enumerate all the bases at a position?

- Greatest challenge: misalignments → incorrect SNP calls

---

Updated comparison of variant detection methods: Ensemble, FreeBayes and minimal BAM preparation pipelines

<table>
<thead>
<tr>
<th>Variant</th>
<th>concordant: total</th>
<th>GATK discordant: total</th>
<th>samtools discordant: total</th>
</tr>
</thead>
<tbody>
<tr>
<td>total</td>
<td>160851</td>
<td>13925</td>
<td>25368</td>
</tr>
<tr>
<td>SNPs</td>
<td>136146</td>
<td>1315</td>
<td>17247</td>
</tr>
<tr>
<td>indels</td>
<td>24705</td>
<td>12610</td>
<td>8121</td>
</tr>
</tbody>
</table>

8% of GATK and 14% of samtools SNP calls are discordant!

---

10,000 exonic sites where the RNA does not match the DNA RNA-DNA difference (RDD)

All 12 possible categories of discordance have been observed
A few statements from the paper

In total, we generated ~1.1 billion reads of 50 base pairs (bp) (~41 million reads and 2 Gb of

Next, we validated our findings experimentally by Sanger sequencing of both DNA and RNA

Proteomic evidence for RDD.

and gene density among chromosomes. RDD sites are significantly ($P < 10^{-10}$) enriched in genes

Claim: at least 90% of the sites in the paper are false positives

12 possible mismatch types. Before accepting such a fundamental claim, a deeper analysis of the sequencing data is required to discern true differences between RNA and DNA from potential artifacts.
Let’s compare aligners: bwa vs bowtie2

Homework 18

1. Install bowtie2

2. Create alignments with bowtie2 and bwa.

3. Compare their outputs in IGV and look for similarities and dissimilarities in the alignments at the simulated error rate of 10%

Extra credit (10 points):
Are there parameter settings for bowtie2 that would make it more competitive with bwa when mapping reads with high error rates?