“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 XY or XX)
- 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 → homozygotes.
• If the alleles are different → 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) → SNV (single nucleotide variant) if observed very rarely
• SNP, SNV → 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\[1\] (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

• The Archon Genomics X PRIZE will 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

• How to do define the accuracy?

Archon Genomics X PRIZE
Validation Protocol
Version 2-24-2011
Granger Sutton, Edison Liu, Victor Jongsma², 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 (AGxP) competition. The purpose of this article in *Nature Previews* is to enlist constructive criticism from the genomic and genetic community on the outlined approaches. The intent for the final version of this Validation Protocol (VP) is to become a useful standard by which to gauge the capability of whole genome sequencing technologies that emerge even after 2012.
Blog: Blue Collar Bioinformatics

- This is where I keep up with the status of the competition :-)
Simulate a dataset and call SNPs

• Generate a simulated dataset with no errors, no insertions and deletions – only SNPs

```bash
albert@porthos:~/work/leci6
$ ~/bin/wgsim -e 0.0 -R 0.0 refs/ec.fa r1.fq r2.fq > mutation.txt

albert@porthos:~/work/leci6
$ head -5 mutation.txt
chrI 1608 A T -
chrI 1635 T A -
chrI 3058 G C -
chrI 3862 T A -
chrI 5308 G S +

albert@porthos:~/work/leci6
$ 
```

Note! The sequences in the fasta file must have the same lengths on each size

samtools index of the fasta file

A second option to extract sequenced from a fasta file

Pileup output from a BAM file

• Generate a BAM file
• Use samtools mpileup to generate a report for each base

```bash
albert@porthos:~/work/leci6
$ ~/bin/samtools mpileup -f refs/mini.fa bwa.bam | head

Usage: samtools mpileup [options] [input] [output]

Input options:
-a       assume the quality is in the Illumina-33 encoding
-b       disable BQ computation
-c       CID parameter for adjusting mapq, 0 to disable [0]
-d       max per-base depth to avoid excessive memory usage [300]
-e       extended SAQ for higher sensitivity but lower specificity
-f       force indexed reference sequence file [null]
-g       skip unmapped reads
-i       list of positions (chr pos) or regions [BND] [null]
-l       SAM line separator [\t] [\n]
-m       max mapping quality at INT (60)
```

Note! The sequences in the fasta file must have the same lengths on each size

MPileup output

```
albert@porthos:~/work/leci6
$ ~/bin/samtools mpileup -f refs/mini.fa bwa.bam | head

Usage: samtools mpileup [options] [input] [output]

Input options:
-a       assume the quality is in the Illumina-33 encoding
-b       disable BQ computation
-c       CID parameter for adjusting mapq, 0 to disable [0]
-d       max per-base depth to avoid excessive memory usage [300]
-e       extended SAQ for higher sensitivity but lower specificity
-f       force indexed reference sequence file [null]
-g       skip unmapped reads
-i       list of positions (chr pos) or regions [BND] [null]
-l       SAM line separator [\t] [\n]
-m       max mapping quality at INT (60)
```
Random region retrieval

Samtools manual has detailed information on the formatting

Actual SNP calling

- It used to be that everyone made up their own formats (that even persists to some extent today)
- New data formats have been adopted VCF (Variant Call Format) and BCF (Binary Variant Call Format) analogous to SAM and BAM
- Understanding `mpileup` is still very useful → you may be able to query your data in ways other data formats cannot

Advanced topic: build more complex queries

Find all the lines where a T is replaced by a C and the measurements are supported by at least five bases of C measured on the forward strand.

All the work is done via the awk regular expression:

```
$5 ~ /C\{5\}/
```

Homework 16

- Generate mutated reads with `wgsim` and the default error models.
- Create an aligned, sorted and indexed BAM file from these reads.
- How many of your mutations are homozygous T to C mutations?
- Generate a `mpileup` output and find out how many of bases have at least one read indicating a T to C mutation.