Variant calling: a world unto its own

Why is it so difficult to call variants?

Genome == Information

Information will vary as much as needed to allow the organism to survive

there is reason for it to be easy to understand

How to tell when you know something?

Bioinformatics in 2015

• We live in an era of massive data generation
• It is also driven by technology
• Current technologies are extremely inefficient → everything needs to be measures hundreds of times
• Most of the data is redundant and most of the information is lost
• Information that is not lost → that’s what Bioinformatics of 2015 is about
Best practices evolve very fast
some things stay the same

Conceptually always the same

Step 1 \rightarrow Step 2 \rightarrow Step 3 \rightarrow Step 4

Can be automatized at the command line

What you actually may change (a lot)

What never stay the same

“How do we call SNPS?”

Always depends on the complexity of the information.

Easy problems \rightarrow can be solved by any tool
Complex problems \rightarrow only some results are correct
Difficult problems \rightarrow needs human intervention

Even comparing SNP calls is hard
(we can’t even tell which SNP callers are correct)

Personalized Medicine

All about the variation

One of the fastest growing fields of Bioinformatics

One with the most promise and hype
GATK is well documented but not easy to use, lots of gotchas – but it works!

Better Ingredients - Better SNP calls

- Base recalibration
  (makes the read quality score better reflect the predictive values)

- Mark (remove) duplicates
  (identical reads are likely artificial, hence will magnify errors)

- Realign around indels
  (Perform local realignment around indels to correct mapping-related artifacts.)

- Refine genotypes
  (compare to known variants and establish stricter discovery to de-novo calls)
Homework 23

• Install the picard tool

• Use alignments that you have previously generated

• Use the picard to perform the following:
  
  – Mark the duplicates in your data. How many of your alignment are duplicates?
  
  – Add read group information to your data. How does your alignment file show the read groups?