Variant calling and annotation

Vincent Plagnol

Inivata- Head of Computational Biology
UCL- Reader in Statistical Genetics

Outline

1. Variant calling algorithms and strategies
2. Variant calling format
3. Annotating your variants

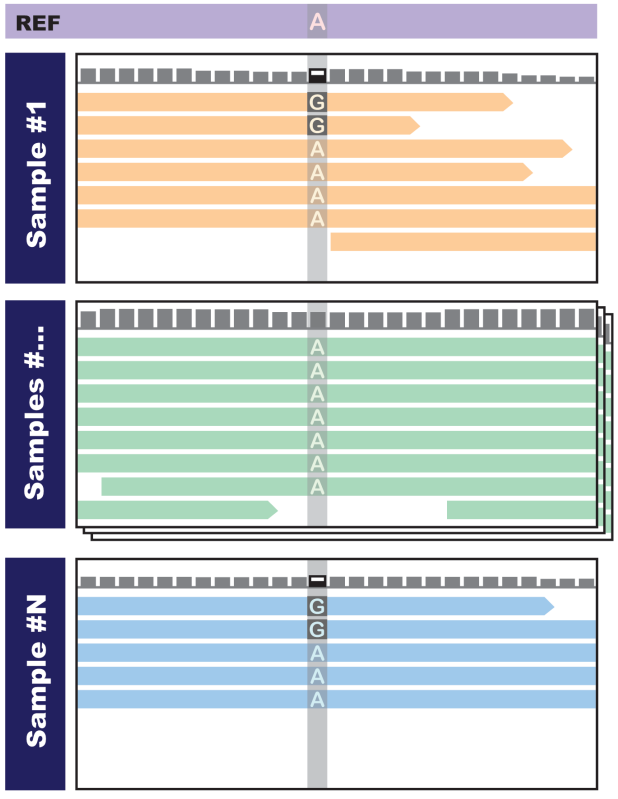
There are several calling algorithms available to you

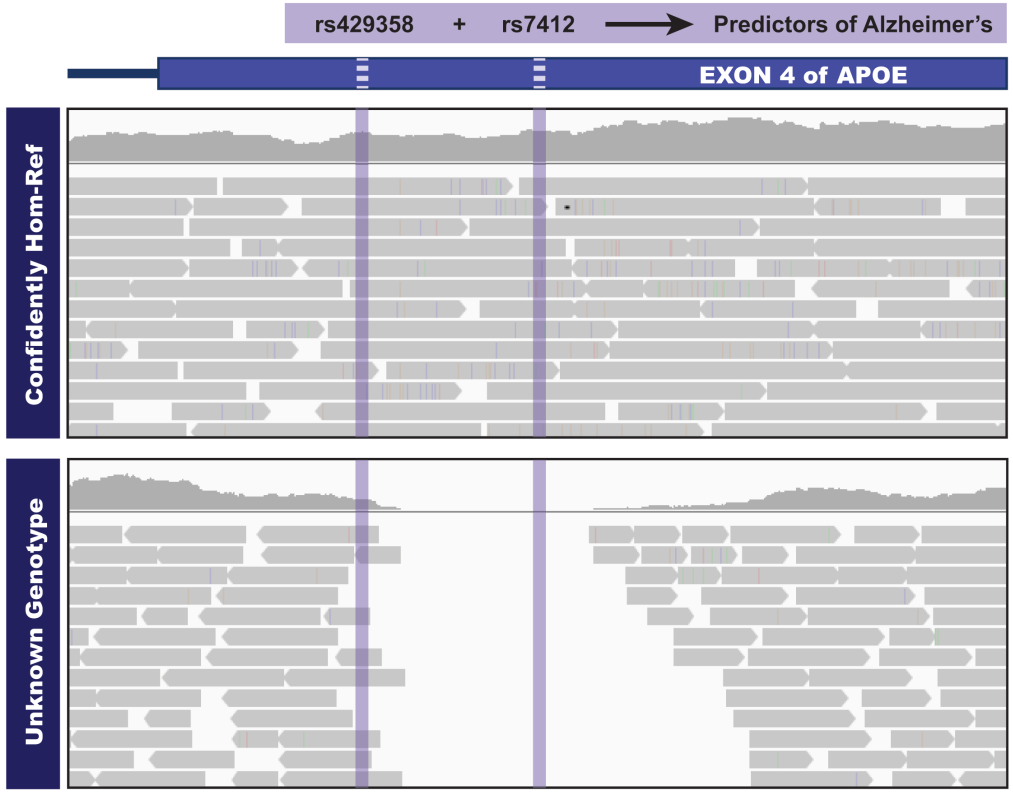
- `samtools` has a perfectly valid variant calling algorithm.
- `GATK` is usually considered the gold standard but there are several ways to use it, which we will discuss today.
- Other options include the recently release `platypus`.

Why calling each sample individually is not always a great idea

- When one works genome-wide, there are always artifacts and technical issues to deal with.
- Sometimes you see the same “rare variant” coming up across multiple smaples.
  - This is obviously extremely unlikely to happen.
- Looking at multiple samples can point to these errors and help you better understand the data.

One can also gain power

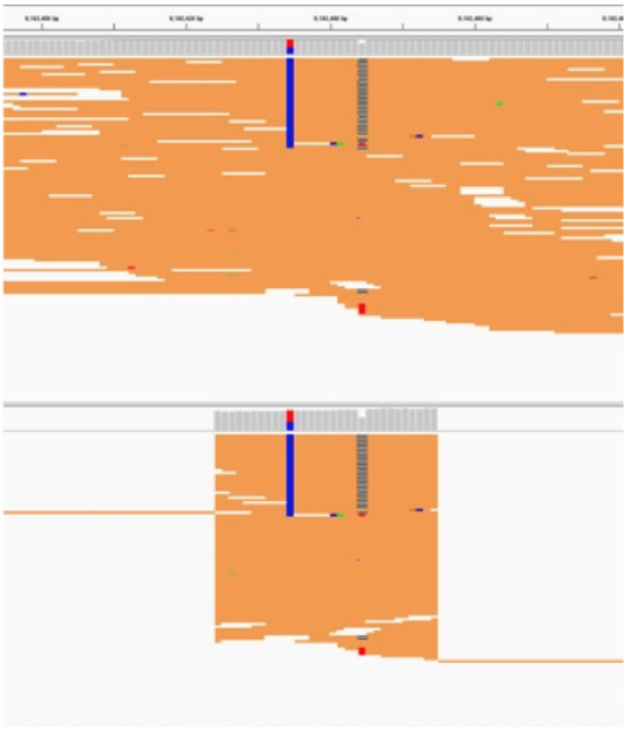




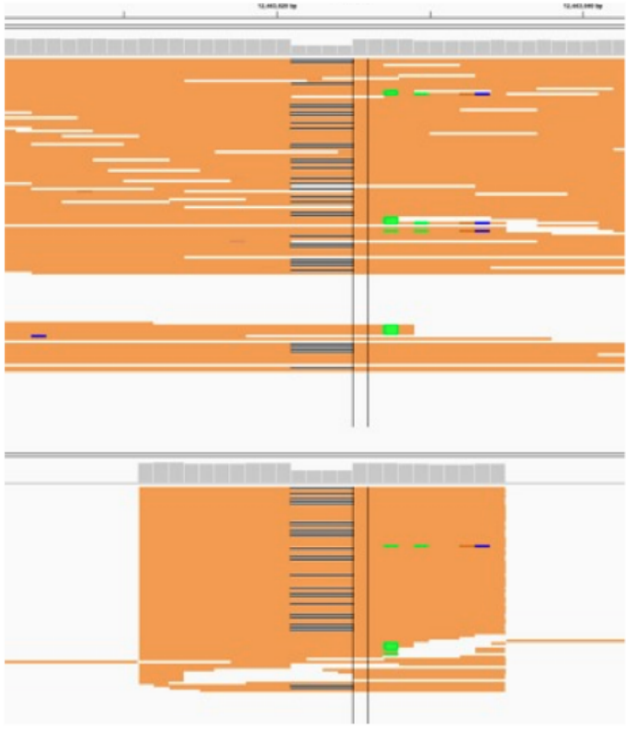
A bit of history: the reduceReads format

original BAM

reduced BAM



multiple variants merging variant region



long deletion

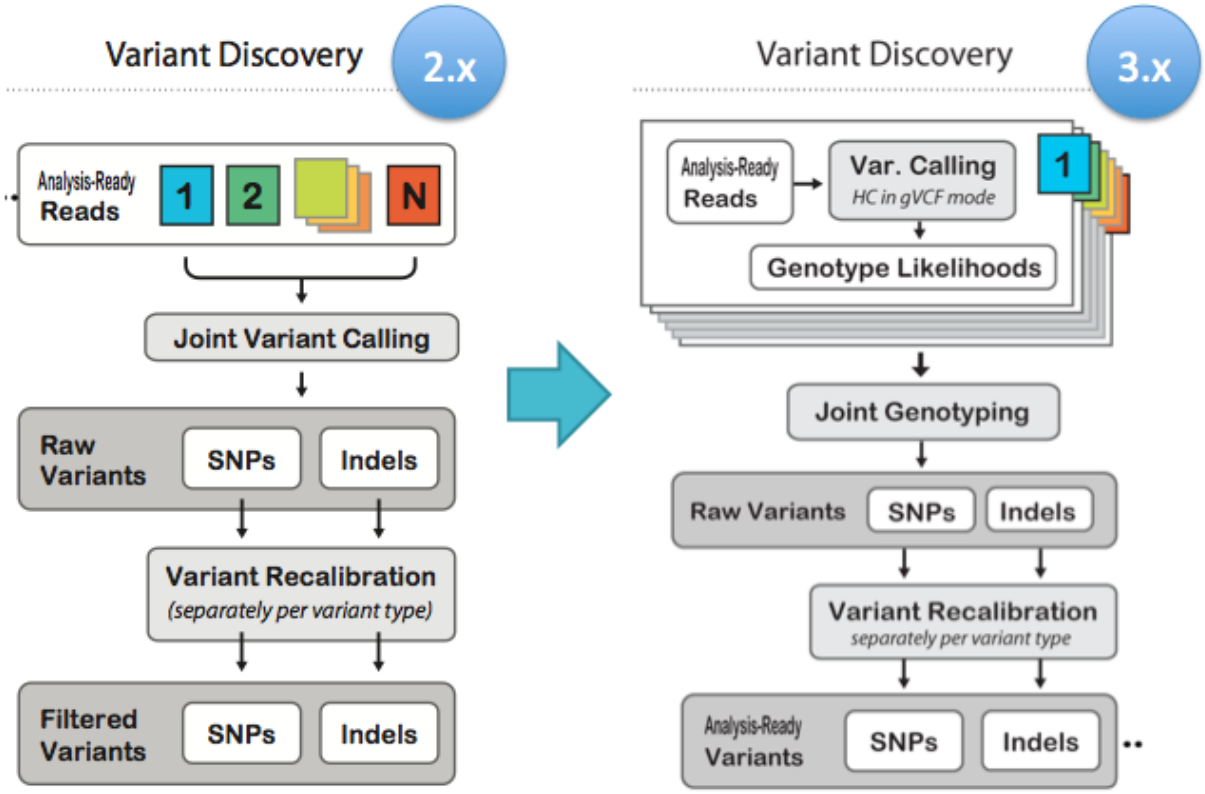
Why is this a hard problem?

- We all agree that we do not want to process 20 Tb of exomes each time we add one extra sample.
- But then, what is the intermediate step?
- We must know not only the variant called, but also the regions where there are NO variant call.
- And when there is a hint of a call, we must also remember that.
    - This is really a compression problem.

The drawback of multi-sample calling

- The computational challenges can be daunting: can you process together 50,000 samples?
- There is also a (N+1) problem: this one extra sample you forgot to process may cost you 1,000 hours of computing.
- To mitigate these issues, the GATK team has put together a hybrid concept, the gVCF workflow that we will look at today.

Joint vs single sample calling



A consequence: massive centralized variant calling efforts

- Given the strong gain in power it really makes sense to have centralized teams handling the calling.
- The most obvious example is the Broad theme (Boston) and the release of the ExAC calls.
  - This website is a “must visit” for anyone working on human genetics.
  - You will find about 80,000 exomes jointly called.
  - It gives an extraordinary perspective on coding regions variability.

New kids in town

- SpeedSeq is supposed to work well.
- All these tools try to find a balance between accuracy and speed.
- Projets like Genomics England really struggle with computational burden and are interested in solutions that are faster than GATK.
- `Platypus` is another new tool from the Oxford group.

My bit of advice

- Think very carefully about your headers, and how you name samples.
- If you get this wrong, it is very hard to fix.
- Also think about your reference genome. Once a choice is made, there is hardly a way back.

The variant calling format (VCF)

- The VCF format is now supported by a team from the “Global Alliance for Genomics and Health”.
- A link to file formats can be found at this location.
- I am not yet familiar with this website but I am hoping that much will happen there, we need this to be successful.

The variant calling format

- This is an incredibly loosely defined format.
- A colleague argued that at the core, it's really just a bunch of tab delimited columns.
- And in particular `samtools` and `GATK` will output slightly different things.
- Nevertheless, the VCF format is ubiquitous and it is important to understand what it stores.

An exercise to go through together

- Compare the flavours of VCF format between `samtools` and `GATK`.
- Note the genotype likelihood, stored in Phred scaled format.
- See what GATK shows that `samtools` does not show.

This is not easy!

- Different tools can give you very different interpretations.
  - The primary issue is differences in the underlying database.
  - A second issue is the handling of the multiple transcripts for each gene.
  - And the third issue is the actual data analysis, with potential minor bugs.
- I have used a lot ANNOVAR in the past...
- ... but really the variant effect predictor (VEP) is a better tool.

More on annotations

- Think carefully about the annotation database that you use.
- Mare careful choices about the way you handle multiple transcripts.