Genetic variations

(A) Substitution
Wild-Type: AACGGCC
Mutant: AACGGCC

(B) Insertion
Wild-Type: AACGGCC
Mutant: AACGGCC

(C) Deletion
Wild-Type: AACGGCC
Mutant: AACGGCC

(D) Indel
Wild-Type: AACGGCC
Mutant: AACGGCC

(B) Structural Variations
Deletion
Wild-Type: AACGGCC
Mutant: AACGGCC

Insertion
Wild-Type: AACGGCC
Mutant: AACGGCC

Translocation
Wild-Type: AACGGCC
Mutant: AACGGCC

(C) Single Nucleotide Polymorphism (SNP) Variations
Wild-Type: AACGGCC
Mutant: AACGGCC

Individual 1: AACGGCC
Individual 2: AACGGCC
Individual 3: AACGGCC
Individual 4: AACGGCC
Individual 5: AACGGCC
Individual 6: AACGGCC
Individual 7: AACGGCC
Individual 8: AACGGCC
Individual 9: AACGGCC
Individual 10: AACGGCC
Individual 11: AACGGCC
Individual 12: AACGGCC

DOI: 10.3389/fbioe.2015.00013

Variations at the (A) nucleotide level and (B) structural level. (C) Single nucleotide polymorphism (SNP) across a population.
Terminology

- Polymorphism: stated regarding a population
- Variant: said of anything that is different regarding a reference
- Beware: SNP not equal SNV.
- InDel: Insertion – Deletion. Used by bioinformaticians to designate either an insertion or a deletion

- In practice: people (even us) often use the term SNP when they talk about SNVs.

- Process of extracting information about genomic variants is called “Variant Calling”
Why would variant calling be complicated?

- Identify true genetic variations and discard false positives.

- False positive variations may arise from any step of the previous analysis:
  - PCR artefacts
  - Sequencing artefacts / errors / quality
  - Alignment
  - Realignment, Recalibration
  - Any other step, even the most insignificant might add its layer of dust
Naïve approach

- What are we trying to do with Variant Calling (VC)?

A BRCA1 SNV from an ovary cancer sample viewed on IGV
Naïve approach

● Advantages
  > Convenient for the “trust what I see” consortium
  > Adds some visual context to your variant

● Disadvantages
  > Too much time and data
  > Reproducibility
  > For large genomes, recognized as a risk factor for major depression

● Keep that to check variants found in silico
In silico approach

- Use software called “Variant Callers”
- Advantages
  - Relatively automatic
  - Gives some methodological background to your study
- Disadvantages
  - Can be computationally intensive
  - Potentially complex algorithms leading to incomplete comprehension of results (may give you some headaches)
  - No “one caller fits all” approach
ZOOMING ON A FEW ALGORITHMS
Calling variant bound to reference: Heuristic approach

- Find all variations
- Various filters *
- Test for significance (Fisher’s exact test)
- Discard variant
- Keep variant

* Base quality, mapping quality, depth, variant depth, context, strand bias, other sample presenting the variation, …
Heuristic p-value: the Varscan2 example

- Get the reference reads and the alternate reads
- Compute coverage at position
- Use baseline errors to compute expected alternate erroneous read counts
- Compute t-test to compare expectation with observation

- Though this is a very naïve approach, it is widely used and has the advantage of having easily understood results. Some publication talk about Varscan2 being the best variant caller.
Calling variant bound to reference: the Broad Institute contributions

- **Broad Institute**
  - IGV (Integrative Genomics Viewer)
  - [http://www.tumorportal.org](http://www.tumorportal.org)
  - GATK
  - (hail – Post-calling variant analysis – In progress)

  - MuTecT (Somatic Cancer Variant Caller)
  - GISTIC (MCR)
  - MutSig (Identify driver mutations)
Calling variant bound to reference: let's get back to GATK

- **GATK: The Genome Analysis Toolkit**
  - 2985 citations (24th October 2016)
  - Collection of many tools
  - Developed by the Broad Institute
  - TCGA, 1000 Genome, …

- GATK isn’t a variant caller, GATK-HC and GATK-UG are
- GATK-UG is deprecated and should not be used anymore
- GATK-HC should not be used for somatic variant calling

GATK-HC : GATK-HaplotypeCaller
GATK-UG : GATK-UnifiedGenotyper
The GATK Best Practices

Best Practices for Germline SNPs and Indels in Whole Genomes and Exomes - June 2016
When should you use GATK-HC

- You would like to find DNA / RNA variants in a constitutional context with a reference genome.
- You have computational resources (complete GATK workflow duration: approx. 1 days / 8 cores / 16gb RAM / sample)
- You are aware and cool with the fact that sometimes, when you check in alignment file, you see something looking like a variant that was not called by GATK (look at bamOutput)
- You like that the people who wrote the software answer you and mostly keep the math away from you.
How does GATK-HC works: step 1

- Sliding window along the reference
- Count mismatches, indels and soft-clips
- Measure of entropy

Stolen from: https://software.broadinstitute.org/gatk/events/slides/1604/presentations/
How does GATK-HC works: step 2

- Local realignment via graph assembly
- Traverse graph to collect most likely haplotypes
- Align haplotypes to ref using Smith-Waterman

Likely haplotypes + candidate variant sites
How does GATK-HC works: step 2

Multiple caller artifacts that are hard to filter out, since they are well supported by read data

NA12878 original read data

Haplotype Caller (validated)
How does GATK-HC works: step 2

• Mapper can represent two different ways, at random:

• HaplotypeCaller will settle on one representation -> cleaner output call
How does GATK-HC works: step 3

- PairHMM aligns each read to each haplotype
- Considers base qualities

Likelihood of the haplotype given reads
How does GATK-HC works: step 4

- Determine most likely combination of allele(s) for each site
- Based on allele likelihoods (from PairHMM)
- Apply Bayes’ theorem with ploidy assumption

Genotype calls

![Genotype sample table]

<table>
<thead>
<tr>
<th>Genotype sample</th>
<th>0/0</th>
<th>0/1</th>
<th>1/1</th>
</tr>
</thead>
<tbody>
<tr>
<td>A/ C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A/ G</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

GLs + annotations
How does GATK-HC works: step 4

\[ P(G|R) = \frac{P(R|G)P(G)}{\sum_i P(R|G_i)P(G_i)} \]

Awful Bayesian math that is here for reference

Plug in the numbers!

Determines the most likely genotype of the sample at each event in the haplotypes
How does GATK-HC works: step 4

- PL is the normalized Phred-scaled probability of each genotype

<table>
<thead>
<tr>
<th></th>
<th>A/A</th>
<th>A/C</th>
<th>C/C</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(G</td>
<td>R)</td>
<td>0.128</td>
<td>0.305</td>
</tr>
<tr>
<td>Raw PL</td>
<td>8.94</td>
<td>5.15</td>
<td>2.46</td>
</tr>
<tr>
<td>Normalized PL</td>
<td>6</td>
<td><strong>3</strong></td>
<td>0</td>
</tr>
</tbody>
</table>

\((-10) \times \log_{10}\{P(G|R)\}\) subtract smallest PL

- GQ is the genotype quality and is the smaller of the 2nd PL or 99
- PLs are in increasing order of possible genotypes, e.g. 0/0, 0/1 and 1/1.

Only few people know what this “PL” is all about

```
#... REF ALT ... FORMAT SAMPLE
... A  C ... GT...:GQ:PL... 1/1...:3:6,3,0...
```
How does GATK-HC works: summary

Stolen from: https://software.broadinstitute.org/gatk/events/slides/1604/presentations/
Calling variants in RNA-Seq bound to reference: the k-mer approach

- Used by CraC
- Advantage: one method to do everything, so you do not accumulate errors. CraC performs the mapping and SNV, InDels, Fusion and splice junctions for each reads
- Disadvantage: At Gustave Roussy, Fusion analysis underperforms other tools. I have no opinion on SNV analysis

(a) Analysis of the location profile
(b) Analysis of the support profile
(c) Substitution
(d) Deletion
(e) Insertion

Variant Calling – Yannick Boursin
How to discover variants without reference

- **Two methods:**
  - Perform de-novo assembly, then map the reads back on the assembly and call variants accordingly.
  - Use *de Bruijn* graphs to model the reads then look for “bubbles”

**Figure 2.** Toy example of a *Bubble* in the *de Bruijn Graph* \((k=4)\). Bubble generated by a single nucleotide polymorphism. The two polymorphic sequences are …*CTGACCT*… and …*CTGTCCT*…

**DOI:** 10.1093/nar/gku1187
Calling variants from assemblies

- First align assemblies using nucmer from Mummer
- Second use show-snps to get snps from the alignment

- If I may, I strongly discourage you from doing that. If you only got some assembly (e.g. 454 – newbler), ask for the reads …
Calling InDels?

A

Paired-end mapping

B

Read-depth analysis

C

Split-read analysis

The growing importance of CNVs: new insights for detection and clinical interpretation – Frontiers in genetics 2013
How to find the right caller with no reference genome?

Experimental Design

- Small reads
- Preferably long reads
- Metagenomics
- Phenotype difference in population

- Stacks
- DiscoSnp++
- MaryGold
- Cortex

TP

RADSeq (Yvan Le Bras - Mercredi)

Reference-free SNP detection: dealing with the data deluge - BMC Genomics 2014
How to find the right caller for your aligned data?

Experimental Design
- Constitutive study
  - GATK HC*, GATK UG, VarScan2, samtools, varDict, Platypus, Freebayes …
- Somatic study
  - MuTecT2*, VarScan2, Strelka, SomaticSniper …
- Trios
  - GATK HC* (mirTrios)*, VarScan2, Scalpel …

Type of algorithm for SNV calling: Heuristic, Statistical Model, Assembly-based

* Note: GATK HC engine uses de novo assembly and a hidden Markov model

Decision help for reference-based genome callers (own made) 1/2
How to find the right caller for your aligned data?

- **Experimental Design**
  - Include specific InDels algorithms
    - pindel, Scalpel, softsearch, ...
  - Large cohorts (+1000)
    - Reveel
  - RNA-Seq
    - GATK HC*, samtools, CraC

Type of algorithm for SNV calling: **Heuristic, Statistical Model, Assembly-based, k-mer**
Type of algorithm for InDel calling: **Read depth, Pairs, Split reads, Assembly-based**

*Note: GATK HC engine uses de-novo assembly and a hidden Markov model*

*Decision help for reference-based genome callers (own made) 2/2*
How to find the right caller for your aligned data?

- The provided list is just a snapshot: there are many other variant callers.

- Advise n°1: check bibliography. Has anyone conducted the same experiment? If someone did, what tool did they use?

- Advise n°2: check if the tool is available to you. Will you get some support would you need it? Is it on Galaxy?

  - E.g.: ABSOLUTE (estimation of ploidy and subclonality). This tool is a mess, though everyone wants to use it.
VARIANT CALLER OUTPUTS
What variant callers outputs: VCF & gVCF
and why researchers are almost never happy about it

What variant callers outputs: VCF & gVCF

and why researchers are almost never happy about it

<table>
<thead>
<tr>
<th>#CHROM</th>
<th>POS</th>
<th>ID</th>
<th>REF</th>
<th>ALT</th>
<th>QUAL</th>
<th>FILTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>877664</td>
<td>rs3828047</td>
<td>A</td>
<td>G</td>
<td>3931.66</td>
<td>PASS</td>
</tr>
<tr>
<td>1</td>
<td>899282</td>
<td>rs28548431</td>
<td>C</td>
<td>T</td>
<td>71.77</td>
<td>PASS</td>
</tr>
<tr>
<td>1</td>
<td>974165</td>
<td>rs9442391</td>
<td>T</td>
<td>C</td>
<td>29.84</td>
<td>LowQual</td>
</tr>
</tbody>
</table>

INFO

<table>
<thead>
<tr>
<th>FORMAT</th>
<th>NA12878</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT:AD:DP:GQ:PL</td>
<td>0/1:1,3:4:26:103,0,26</td>
</tr>
<tr>
<td>GT:AD:DP:GQ:PL</td>
<td>0/1:14,4:14:61:61,0,255</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Field</th>
<th>What it is</th>
<th>Field</th>
<th>What it is</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT</td>
<td>Genotype</td>
<td>GQ</td>
<td>Genotype Quality</td>
</tr>
<tr>
<td>AD</td>
<td>Allelic Depth</td>
<td>PL</td>
<td>Phred-scaled Likelihood</td>
</tr>
<tr>
<td>DP</td>
<td>Global Depth</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What variant callers outputs: VCF & gVCF
and why researchers are almost never happy about it

<table>
<thead>
<tr>
<th>#CHROM</th>
<th>POS</th>
<th>ID</th>
<th>REF</th>
<th>ALT</th>
<th>QUAL</th>
<th>FILTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>877664</td>
<td>rs3828047</td>
<td>A</td>
<td>G</td>
<td>3931.66</td>
<td>PASS</td>
</tr>
<tr>
<td>1</td>
<td>899282</td>
<td>rs28548431</td>
<td>C</td>
<td>T</td>
<td>71.77</td>
<td>PASS</td>
</tr>
<tr>
<td>1</td>
<td>974165</td>
<td>rs9442391</td>
<td>T</td>
<td>C</td>
<td>29.84</td>
<td>LowQual</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Column</th>
<th>What the hell it is</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHROM</td>
<td>Chromosome number</td>
</tr>
<tr>
<td>POS</td>
<td>Genomic position</td>
</tr>
<tr>
<td>ID</td>
<td>Database ID if any (default: ‘.’)</td>
</tr>
<tr>
<td>REF</td>
<td>Reference allele (namely, the 0 allele, as in “patient 0”)</td>
</tr>
<tr>
<td>ALT</td>
<td>Alternate allele (if multiple, comma-separated, each gets a number in order).</td>
</tr>
<tr>
<td>QUAL</td>
<td>Phred-scaled quality score (complicated in HaplotypeCaller)</td>
</tr>
<tr>
<td>FILTER</td>
<td>PASS (may be any value there, it is supposed to help you filtering)</td>
</tr>
</tbody>
</table>
What variant callers outputs: VCF & gVCF
and why researchers are almost never happy about it

<table>
<thead>
<tr>
<th>INFO</th>
<th>FORMAT</th>
<th>NA12878</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ANNOTATIONS]</td>
<td>GT:AD:DP:GQ:PL</td>
<td>0/1:1,3:4:26:103,0,26</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Column</th>
<th>What the hell it is</th>
</tr>
</thead>
<tbody>
<tr>
<td>INFO</td>
<td>Place for any position wide information (like GD quality info, variant genomic effect, cosmic count …)</td>
</tr>
<tr>
<td>FORMAT</td>
<td>Works with the next column (sample, here NA12878 is the sample id) and defines the order in which the data is going to be displayed</td>
</tr>
<tr>
<td>NA12878</td>
<td>Sample-dependent data.</td>
</tr>
</tbody>
</table>
What variant callers outputs: VCF & gVCF
and why researchers are almost never happy about it

**gVCF**: Genomic VCF

- Contains info about every regions.
- Useful when calling variants on cohorts
- How do you know you can compare a specific genomic region in a cohort? Maybe you say there is no variation whereas the region has just not been sequenced.
- Get rid of some computational costs and improve quality.
What variant callers outputs: VCF & gVCF

and why researchers are almost never happy about it

- Each variant caller has its own output format, though it is generally VCF compliant.
- VCF is very flexible
- VCF is complicated to read
- Any kind of variation can be stored in VCF

- Documentation: [https://samtools.github.io/hts-specs/VCFv4.2.pdf](https://samtools.github.io/hts-specs/VCFv4.2.pdf)
- Getting out of it: use specific tools to get tables or views out of VCFs
Looking at VCF files?

References

Variant Density
(click bottom chart to select a reference)

TSTV Ratio

Base Changes

http://vcf.iobio.io
Looking at VCF files?

- With some learning: Gemini
  - [http://dx.doi.org/10.1371/journal.pcbi.1003153](http://dx.doi.org/10.1371/journal.pcbi.1003153)
  - [https://gemini.readthedocs.io](https://gemini.readthedocs.io)

- Learning way more: R with VcfR
  - [http://dx.doi.org/10.1101/041277](http://dx.doi.org/10.1101/041277)

- You don’t need this course anymore (but some of those tools can be found on galaxy)
  - vcftools, bcftools, tabix, vcfli
  - pyVCF, pySam
Haplotype?

**A** SNPs


**B** Haplotypes

Haplotype 1: C T C A A A G T A C G G T T C A G G C A 
Haplotype 2: T T G A T T G C G C A A C A G T A A T A 
Haplotype 3: C C C G A T C T G T G A T T A C T G G T G 
Haplotype 4: T C G A T T C C G C G T T C A G A C A 

**C** Tag SNPs

A / G
T / C
C / G

How does GATK-HC works: step 4

\[
\mathcal{L}(G|R) = \prod_j \left( \frac{\mathcal{L}(H_1|R_j)}{2} + \frac{\mathcal{L}(H_2|R_j)}{2} \right), \quad G = H_1H_2 \text{ for diploids}
\]

Genotype likelihoods for $G_{C/C}$, $G_{C/A}$ and $G_{A/A}$ given reads $R_{1-3}$:

- $L(G_{C/C}|R_{1-3}) = \frac{([0.10+0.10]/2)[(0.10+0.10)/2][0.12+0.12]/2} = 0.00120$
- $L(G_{C/A}|R_{1-3}) = \frac{([0.10+0.06]/2)[(0.10+0.09)/2][0.12+0.05]/2} = 0.00065$
- $L(G_{A/A}|R_{1-3}) = \frac{([0.06+0.06]/2)[(0.09+0.09)/2][0.05+0.05]/2} = 0.00027$

Genotype probability:

- $P(G_{C/C}|R_{1-3}) = 0.567$
- $P(G_{C/A}|R_{1-3}) = 0.305$
- $P(G_{A/A}|R_{1-3}) = 0.128$

• Assigns highest probability genotype $C/C$