MEDI6215

UNIVERSITY OF Southampton

Variant Calling

Bioinformatics, Interpretation, and Data Quality Assurance in Genome Analysis

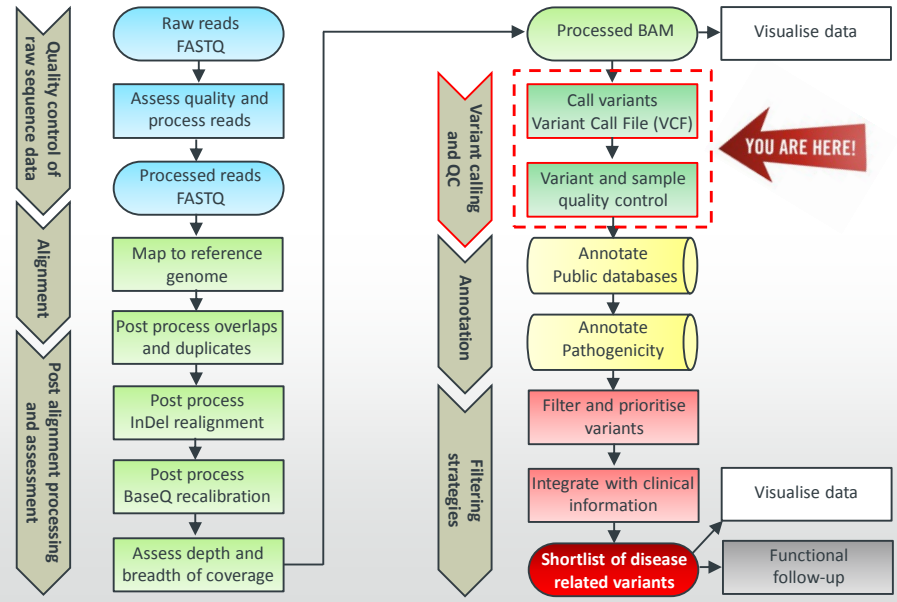
Will Tapper

7th February 2017

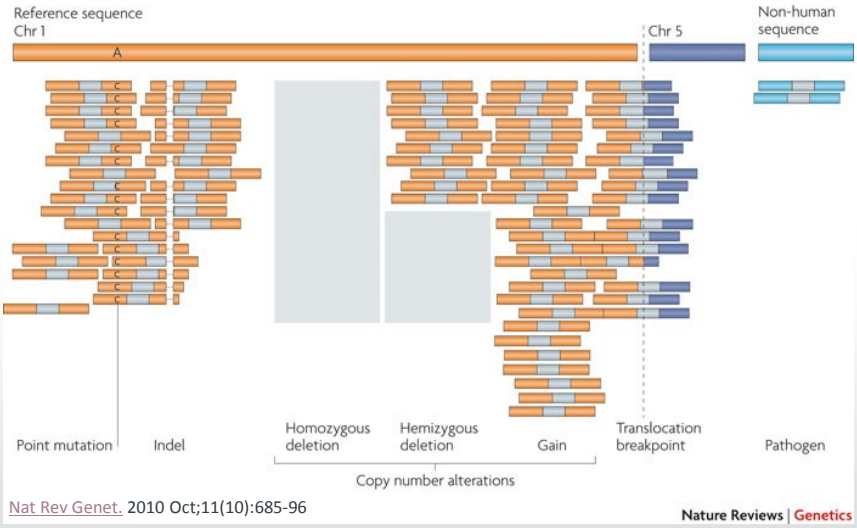
Lecture outline

- Types of genetic variation detected by NGS
- Methodology for variant calling and genotyping
  - Allele counting, heuristic and probabilistic
- Concordance between variant calling software
- Format of Variant Call Files (VCF)
- Evaluation of variant calling as a whole
  - Variant count, overlap with known variation, transition to transversion ratio and heterozygous to homozygous ratio
- Variant quality control
  - Visualisation, hard filtering and variant quality recalibration

Analysis workflow

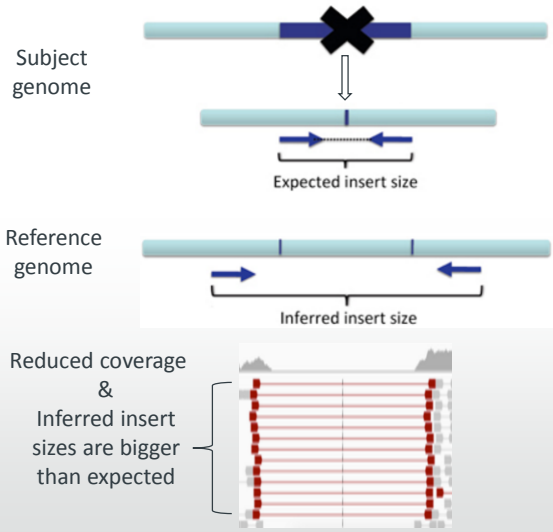


Types of variants detected by NGS



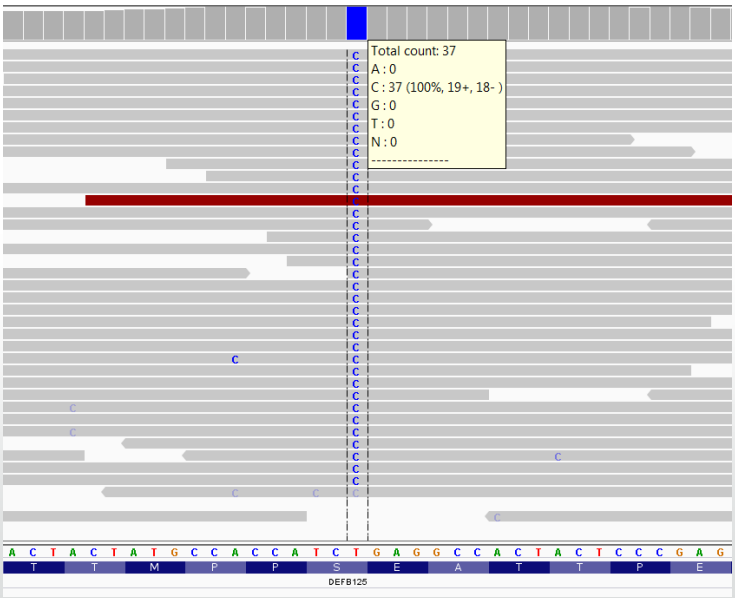
The aim is to detect clinically relevant variants from sequence data

Deletions and insert size

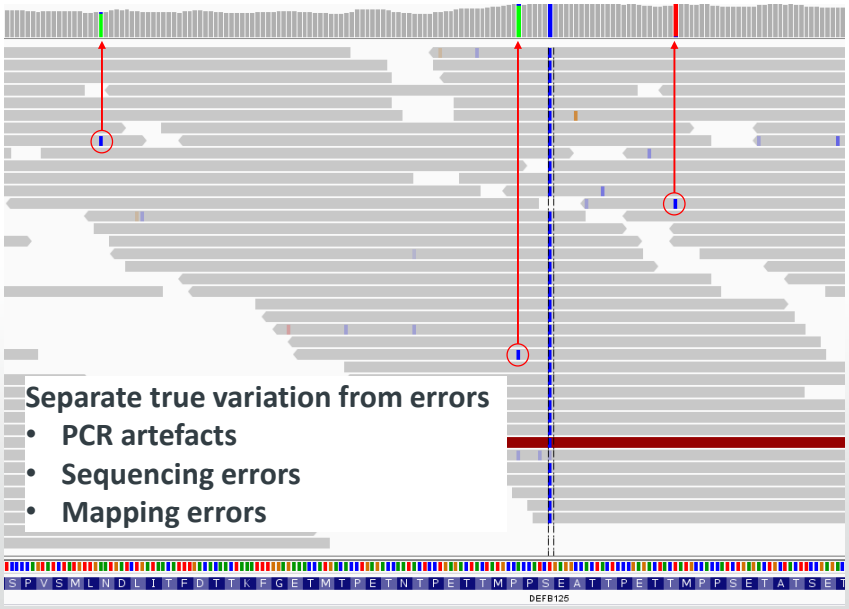


http://software.broadinstitute.org/software/igv/interpreting_insert_size

Variant calling: In principal it is simple



In reality it is more complex



Variant calling software

- Over 15 different variant calling programmes
- 3 main categories: Allele counting, heuristic and probabilistic

Software	Calling method	Metric	Reference
Bambino	Allele counting	SNP Score	Edmonson MN et al (2011)
VarScan	Heuristic	Phred	Koboldt D et al (2012)
GNUMAP	Probabilistic	Phred	Clement NL et al (2009)
SOAPsnp	Probabilistic	Phred	Li R et al (2009)
SAMtools	Probabilistic	Phred	Li H et al (2009)
SNVer	Probabilistic	Phred	Wei Z et al (2011)
GATK: UnifiedGenotyper	Probabilistic	Phred	DePristo MA et al (2011)
GATK: HaplotypeCaller	Probabilistic	Phred	DePristo MA et al (2011)

Variant calling methods

Allele counting

- Filter (≥Q20), count alleles, heterozygous if 20-80% otherwise homozygous
- Requires ≥20 reads for heterozygotes to have 20-80% alternate allele freq.
- Under-calls heterozygous variants when depth is moderate to low
- Does not consider base quality and no measure of confidence (equally likely)

Heuristic approach

- Based on thresholds for read depth, base quality, variant allele frequency
- Provides a measure of statistical significance
- Robust to outlying data that violate the assumptions of other models

Probabilistic methods (eg Bayesian model)

- Calculate a posterior probability for each genotype based on:
  1. Prior probability of genotypes (how probable irrespective of the data)
  2. Likelihood of alleles given the observed read data and base qualities
  3. Probability of the data under all hypotheses
- Posterior genotype probabilities used to measure of genotype confidence

Which variant caller to use?

Which variant-caller to use?

O'Rawe et al 2013, compared 5 variants callers (15 exomes, 120x mean coverage)

Concordance with SNP genotyping arrays

Software	Sensitivity	Specificity
GATK v1.5	95.3	99.7
SOAPsnp	94.7	99.8
SAMtools	94.5	99.6
SNVer	92.3	99.8
GNUMAP	86.6	99.6

- High sensitivity and specificity
- Common SNPs, in regions with little repeat DNA and without extreme GC contents
- Not a true measure of performance because arrays do not represent a personal genome

Number of *de novo* non-synonymous SNVs detected by all 5 callers

Family 1	2 generation
Child A	241
Child B	211
Child C	102
Child D	242

- ~1-2 per ind. expected (Roach et al 2010)
- 102-242 identified using parents
- 0-6 identified using parents & grandparents
- False -ves in parents

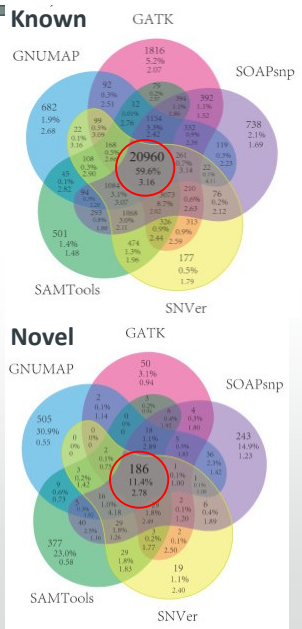
Which variant-caller to use?

SNV concordance

- 60% of **known** SNVs (dbSNP) called by all 5 methods
- 90% were called by GATK and SOAPsnp
- 11% of **novel** SNVs called by all 5 methods
- 36% were called by GATK and SOAPsnp

Validation of SNVs from GATK and SOAPsnp (n=919)

- 99% of overlapping SNVs were real (312/315)
- 97% of SNV unique to GATK were real (306/315)
- 60% of SNV unique to SOAPsnp were real (174/289)



Genome Med. 2013 Mar 27;5(3):28. doi: 10.1186/gm432.

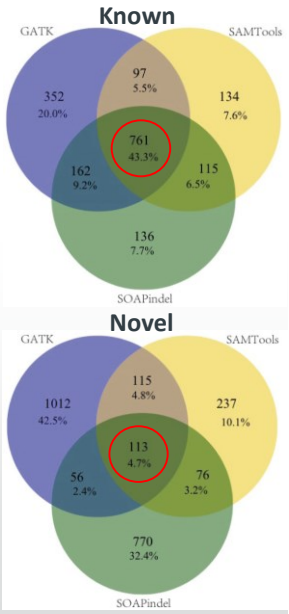
Which variant-caller to use?

Indel concordance

- 43% of **known** indels were called by all 3 methods
- 67% were called by GATK and SOAPsnp
- 5% of **novel** indels were called by all 3 methods
- 13% were called by GATK and SOAPsnp

Validation of indels from GATK and SOAPsnp (n=837)

- 78% of overlapping indels were validated (132/169)
- 54% of indels unique to GATK validated (180/336)
- 45% of indels unique to uSOAPsnp validated (148/332)



Genome Med. 2013 Mar 27;5(3):28. doi: 10.1186/gm432.

Which variant-caller to use?

- SNVs are more reliably called than indels
- Improve overall calling accuracy by using a combination of callers
- But… this increases cost and turnaround time
- Consensus approach used to create benchmarks (eg Illumina Platinum Genomes)
- Popitsch et al 2016: Overlap between variant calling software depends mainly on genomic context rather than the sequencing data
- Suggests that the genome can be split into regions that can and cannot be reliably genotyped by a single method
- Hard to call regions have low sequence complexity
- False +ve variant calls are mainly due to PCR and alignment errors
- Exclude or flag unreliable regions or use a combination of variant callers in hard to call areas

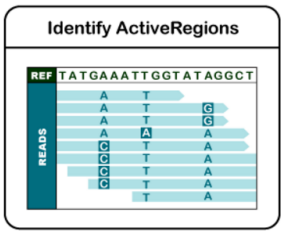
In practice

- Setup pipelines using SAMtools, GATK UnifiedGenotyper and GATK HaplotypeCaller for germline variation
- VarScan to identify somatic mutations from paired tumour and normal samples

Software	Calling method	Metric	Reference
VarScan	Heuristic	Phred	Koboldt D et al (2012)
SAMtools	Probabilistic	Phred	Li H et al (2009)
GATK: UnifiedGenotyper	Probabilistic	Phred	DePristo MA et al (2011)
GATK: HaplotypeCaller	Probabilistic	Phred	DePristo MA et al (2011)

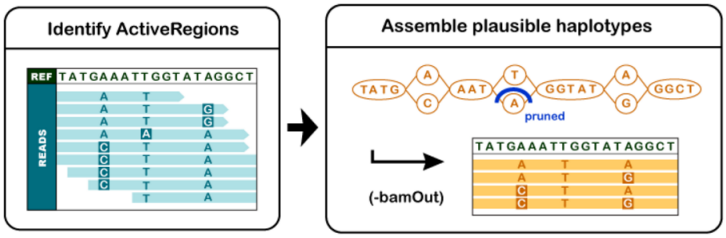
GATK HaplotypeCaller

Step 1: Use a sliding window to identify regions with significant evidence for variation relative to the reference genome



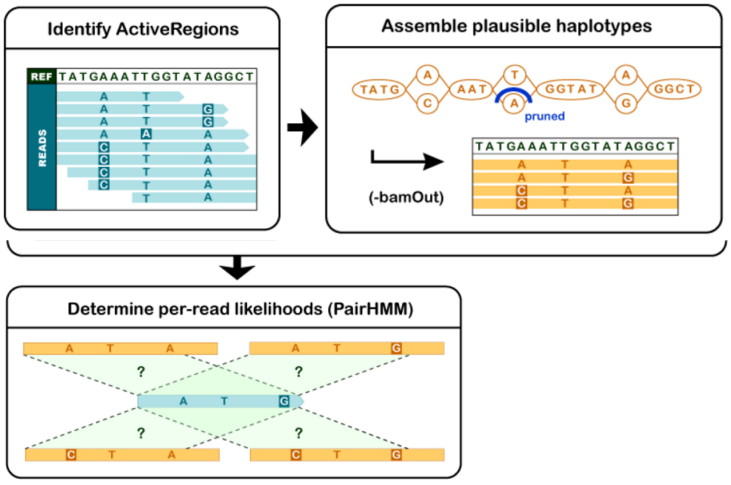
GATK HaplotypeCaller

Step 2: Make all plausible haplotypes in active region using a De Bruijn-like graph
Identify variant sites by realigning each haplotype against reference



GATK HaplotypeCaller

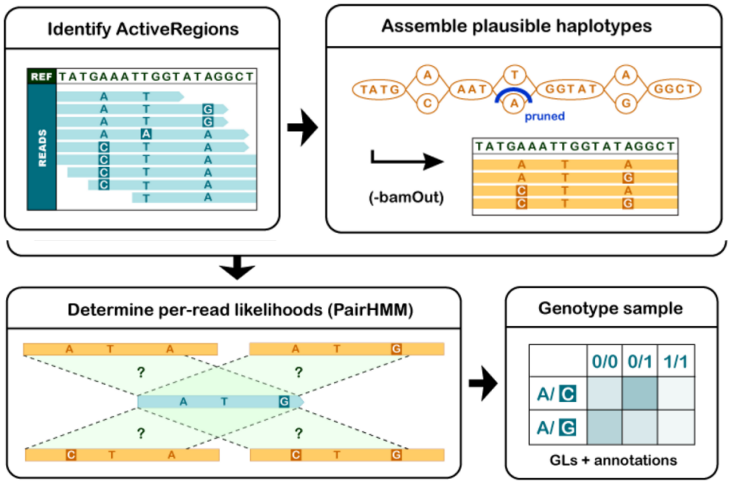
Step 3: Pairwise alignment of each read against each haplotype using PairHMM
This produces a matrix of likelihoods of haplotypes given the read data
Marginalise likelihoods to give likelihood of alleles per read for each variant



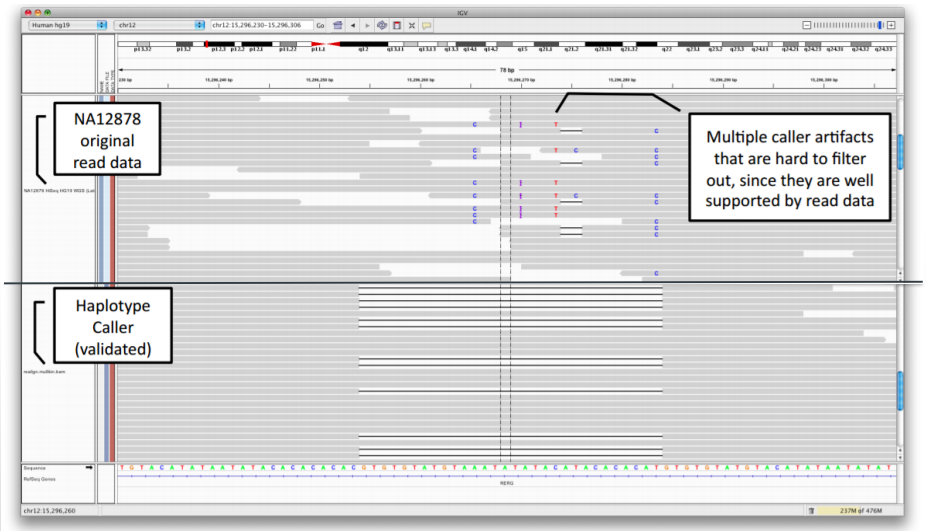
GATK HaplotypeCaller

Step 4: For each variant, calculate the posterior likelihood of each genotype given the read data

The most likely genotype is used as the genotype call



GATK HaplotypeCaller



Variant Call Files (VCF):

Variant call files (VCF)

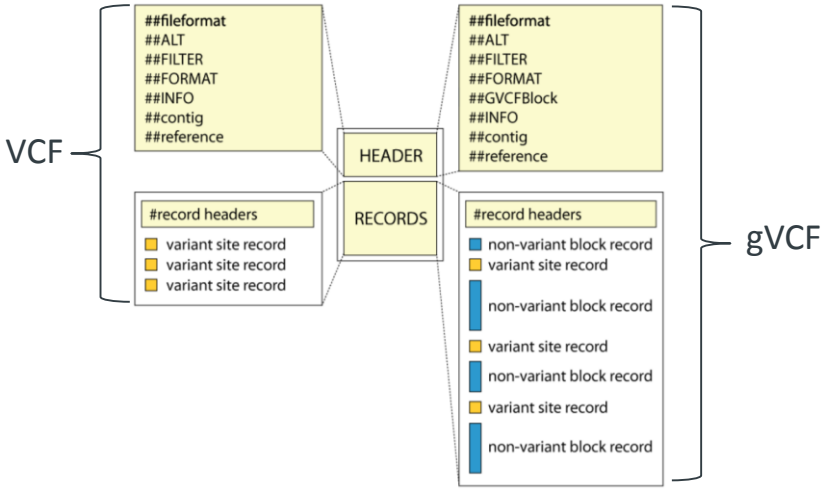
- **QUAL:** Quality score for the ALT assertion (variant or not?) from Bayesian model
- **GT**: Genotype, variant sites only (0/1, 1/1) no reference homozygotes (0/0)
- **AD**: Allelic depth, number of reads with reference and alternate allele
- **DP**: Depth, total number of reads
- **GQ**: Genotype quality. Phred-scaled confidence that genotype is correct, derived from the genotype PL, maximum of 99
- **PL:** Phred scaled genotype likelihoods for ref. hom (0/0), het (0/1), alt. hom (1/1) Most likely genotype has Q score = 0 and the other are scaled relative to this

##fileformat=VCFv4.1

CHROM	POS	REF	ALT	QUAL	INFO	FORMAT	SAMPLE-A
20	66370	G	A	691.96	AC=1;AF=0.5;AN=2;BaseQRankSum=1.1;DP=43;Dels=0; FS=0;HRun=0;HaplotypeScore=1;MQ=60;MQ0=0; MQRankSum=-0.4;QD=16.09;ReadPosRankSum=-1.7	GT:AD:DP: GQ:PL	0/1 20,23 43 99 722,0,612
20	68749	T	C	1009.13	AC=2;AF=1;AN=2;DP=30;Dels=0;FS=0;HRun=0; HaplotypeScore=0;MQ=60.00;MQ0=0;QD=33.64	GT:AD:DP: GQ:PL	1/1 0,30 30: 87.25 1042,87,0

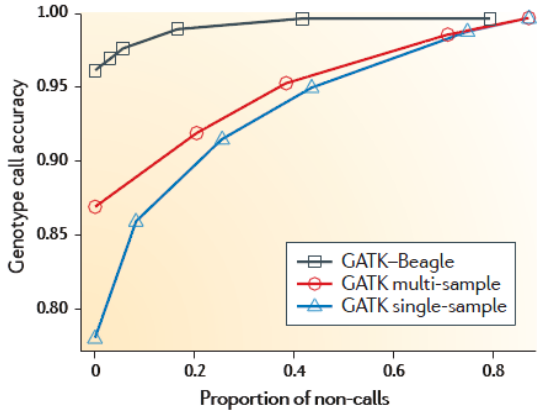
Genomic variant call files (gVCF)

- Records all variant and non-variant sites (block) and gives call confidence
- Information on all sites is used for subsequent multi-sample analysis



Single or multi-sample variant calling?

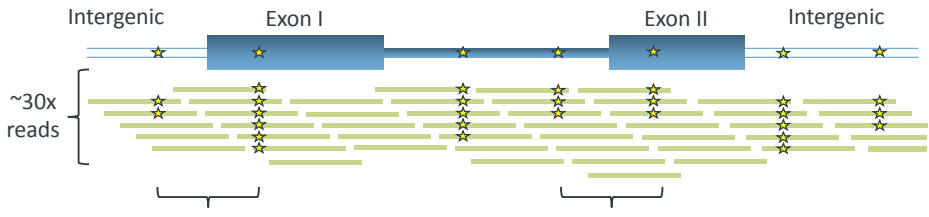
- In multi-sample calling, read info from all samples is combined
- Better discrimination between real and false +ve variants
- Multi-sample calling has higher genotype call accuracy



Nat Rev Genet. 2011 Jun;12(6):443-51

Evaluation of variant calling:

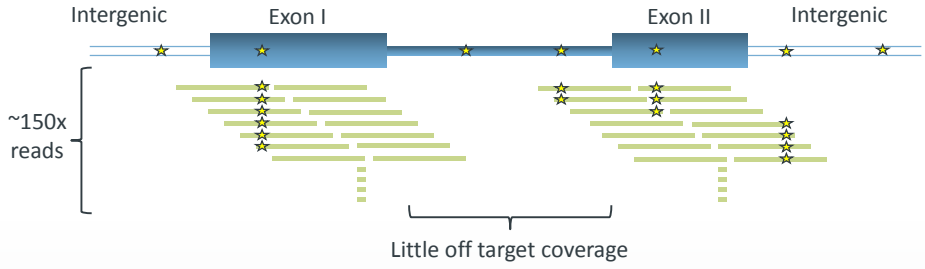
Whole Genome Sequencing



Data requirements per sample	
Targeted bases	~3 bn bases
Depth	Avg. 30x
#Sequenced bases	100Gb
#Lanes of HiSeq	~8 lanes

Expected variant detection	
Variants per sample	~4.6M (1/650bp)
Amount of variation in dbSNP 129	~83%
Transition to transversion ratio	>2
Heterozygous to homozygous ratio	1.6
Estimated contamination level	<3%
Heterozygosity of X chromosome	♂ 10% ♀ 60%

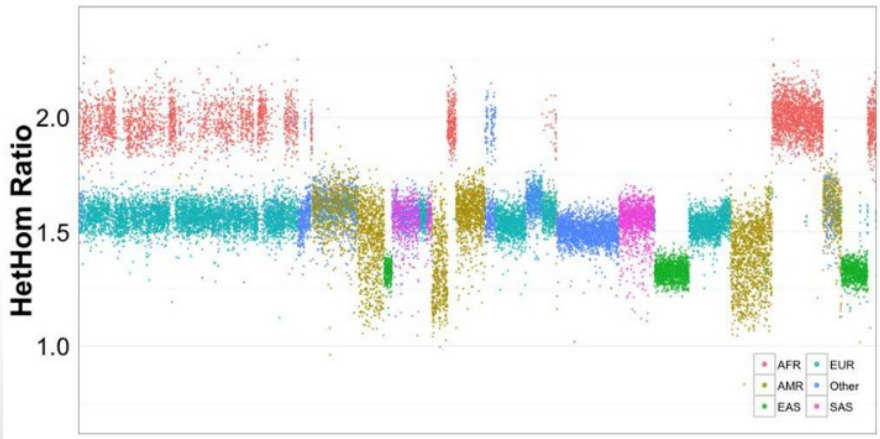
Whole Exome Sequencing



Data requirements per sample	
Targeted bases	~32Mb
Coverage	>80% >20x
#Sequenced bases	5Gb
#Lanes of HiSeq	~0.33 lanes

Variant detection	
Variants per sample	~25K (1/1280bp)
Amount of variation in dbSNP 129	~83%
Transition to transversion ratio	>3
Heterozygous to homozygous ratio	1.6
Estimated contamination level	<3%
Heterozygosity of X chromosome	♂ 10% ♀ 60%

Adjust expectations according to ethnicity



Monkol Lek, 2014

Deviation in variant evaluation metrics

- After accounting for ethnicity, deviation from expected values may indicate problems with data or analysis BUT they could also reflect the underlying biology
- Too few variants may suggest low coverage
- Low TiTv ratio: suggests callset has more false positives
- Excess heterozygosity could be due to sample contamination or recent admixture
- Low heterozygosity could occur due to parental relatedness, large deletions, chromosomal loss or acquired uniparental disomy (both chromosomes from one parent)

UNIVERSITY OF Southampton

Quality control of variants:

Variant quality control

- Use QC metrics in VCF file to remove or flag artifactual variants

1) **Hard filtering**
   - Apply lenient thresholds to quality metrics
2) **Visualize aligned data (BAM)**
   - Assess QC metrics and sequence context
3) **Variant quality recalibration**
   - Generate a model from quality metrics that is used to recalculate QUAL scores
   - Requires a list of known/real variants and data from 30+ exomes

GATK Unified Genotyper

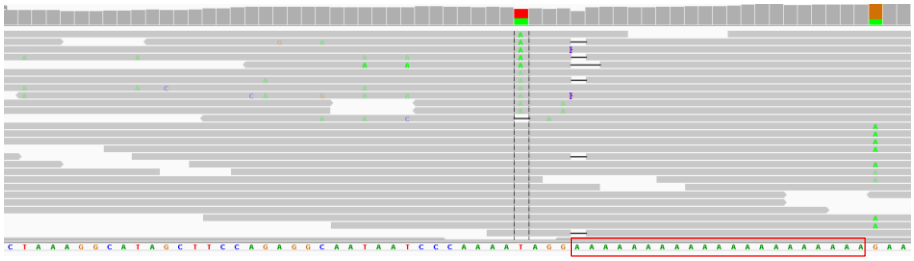
CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
20	61098	.	C	T	409	.	AC=1;AF=0.50;AN=2;BaseQRankSum=2.404;DP=28;Dels=0.00;FS=4.154;HRun=1;HaplotypeScore=0.0000; MQ=60.00;MQ0=0;MQRankSum=-0.601;QD=14.60;ReadPosRankSum=0.134
20	80655	.	A	G	778	.	AC=2;AF=1.00;AN=2;DP=21;Dels=0.00;FS=0.000;HRun=0;HaplotypeScore=0.0000; MQ=60.00;MQ0=0;QD=37.05

Some recommendations for hard filtering

Quality metric	SNV	Indels
QD: Variant quality / depth (QUAL/DP)	< 2	< 2
FS: Strand bias	FS > 60	FS > 200
ReadPosRankSum: Tail bias	< -8	< -20
BaseQRankSum: Base quality bias	< -2	?
MQRankSum: Mapping bias	< -12.5	NA
HaplotypeScore: Consistency with ≤2 haplotypes	> 13	NA

- **QD:** Confidence should increase with increasing depth
- **FS:** Is the distribution reads mapping to +ve and –ve strand similar for ref and alt alleles?
- **ReadPosRankSum:** Are alt bases located evenly throughout reads?
- **BaseQRankSum:** Artefact if reads with alt. allele have lower base quality than ref allele?
- **MQRankSum:** Is average mapping quality similar for ref and alt alleles?
- **HaplotypeScore:** Probability that the reads in a window around the variant can be explained by at most two haplotypes

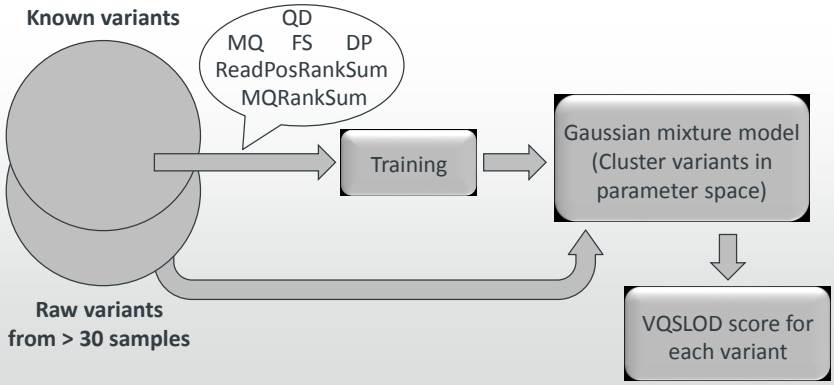
Assess variant quality and context



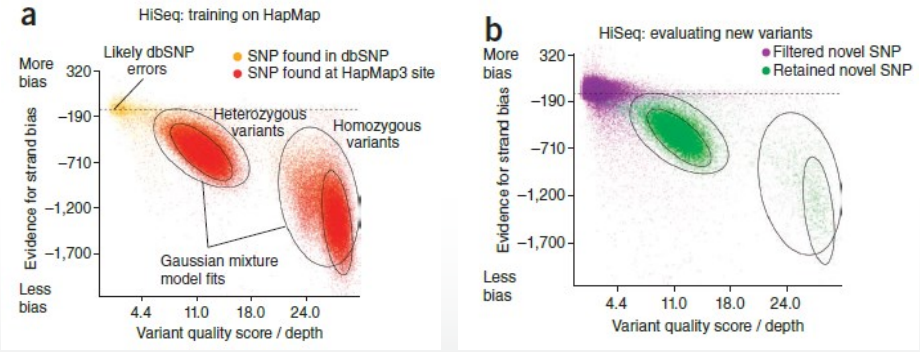
Quality metric	SNV	A	B
QD	< 2	1.38	1.73
FS	FS > 60	13.6	23.3
ReadPosRankSum	< -8	0.9	0.4
HaplotypeScore	> 13	9.7	2.4
BaseQRankSum	< -2	-3.1	-2.1
HRun	> 4	4	21

Variant quality recalibration (GATK)

- Recalculate variant probability and use it to generate a highly accurate call set
- Takes overlap between known variants (HapMap and dbSNP) and raw variants and models their distribution relative to QC parameters to create clusters
- Clusters used to assign VQSLOD score (log odds ratio of being a true variant versus false under Gaussian model) requires data from >30 samples (GATK)



Variant quality recalibration (GATK)

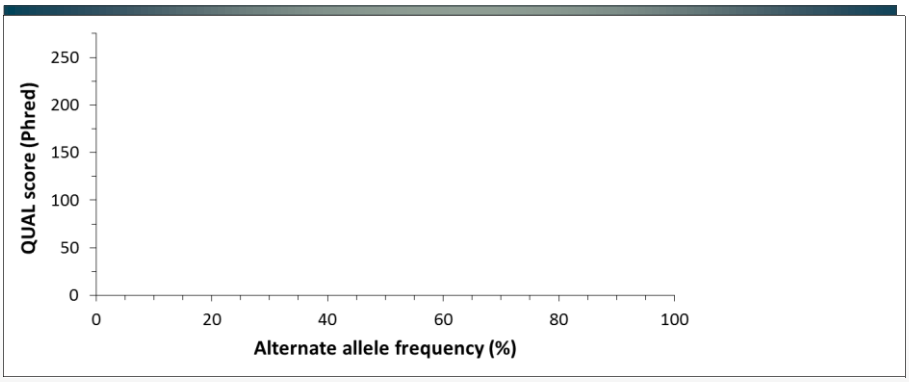


DePristo et al. (2011)

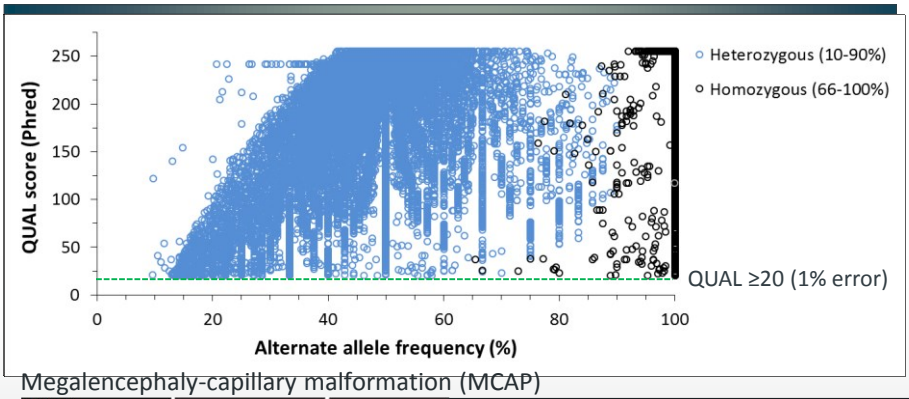
Summary

- Three categories of variant caller (counting, heuristic, probabilistic), probability methods are most commonly used, heuristic can be good for outliers
- The output is a list of *in-silico* variants (VCF file) which have a range of quality scores and a number of false positives
- The overlap between variant calling software is modest (<90% for SNVs and <67% for indels). Real variants are more likely to be concordant but many unique-to-caller variants have also been validated
- The best solution is to use multiple variant callers and multi-sample calling
- The calling process should be evaluated using: variant density, overlap with known variants, het:Hom ratio, transition:transversion ratio and these tests need to consider ethnicity
- Potential false positives can be flagged or excluded by hard filtering or variant quality recalibration
- Evidence for variant calls should be checked, positive and negative calls should be carefully interpreted especially for indels which are more error prone

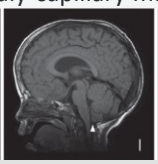
Variant quality scores

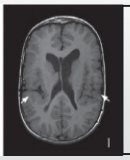


Variant quality scores



Megalencephaly-capillary malformation (MCAP)

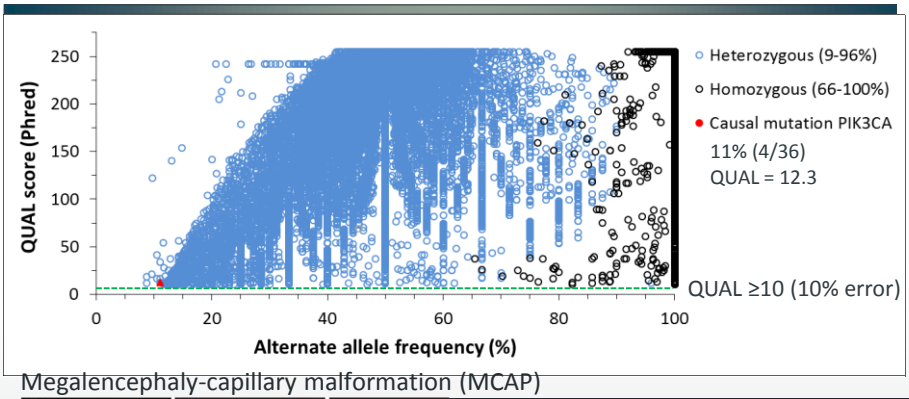




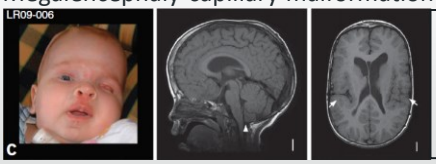
- Macrocephaly, prominent forehead
- Midline facial capillary malformation
- MRI abnormalities
- Rivière et al Nature Genetics 2012

- Postzygotic mutation in PIK3CA: - only present in a proportion of cells (mosaic)
  - low alternate allele frequency

Variant quality scores



Megalencephaly-capillary malformation (MCAP)



- Macrocephaly, prominent forehead
- Midline facial capillary malformation
- MRI abnormalities
- Rivière et al Nature Genetics 2012

- Postzygotic mutation in PIK3CA: - only present in a proportion of cells (mosaic)
  - low alternate allele frequency

Lessons learned

- The Megalencephaly-capillary malformation (MCAP) example is unusual case but it demonstrates some important points:

1) Know the QUAL threshold used to filter VCFs
2) Set the QUAL threshold according to mutation type
3) If causal variants are not detected the QUAL threshold can be reduced but this will introduce more false positives
4) Use different programs to call germline and somatic mutations