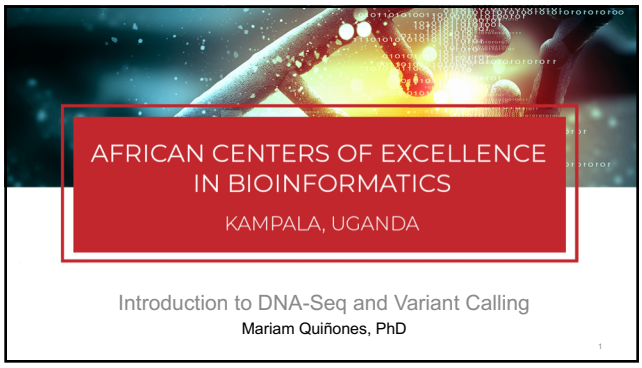


Introduction to DNA-Seq and Variant Calling

Mariam Quiñones, PhD

1

Today's Instructor

Dr. Mariam Quiñones, Ph.D. in Molecular Microbiology

Ongoing Computational Biology projects:
- 16S microbiome
- Ion Torrent Zika virus genome assembly

- Bioinformatics and Computational Biosciences Branch (BCBB), NIAID
- National Institutes of Health, Bethesda, MD USA.
- Contact our team via email:
  - Email: ace@icermali.org
  - Listserv: ACE-GLOBAL-ANNOUNCE@list.nih.gov
  - Instructor: mariam.quinones@nih.gov

2

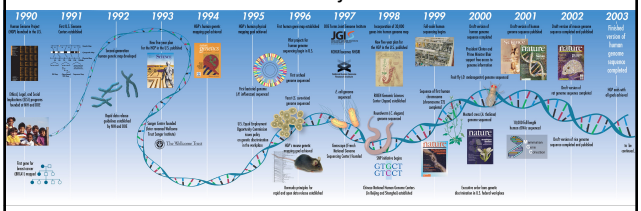
Overview

- Relevant technology concepts
- Introduction to genomic variant studies
- WGS or Exomes
- Step by step workflow for variant calling using GATK

Acknowledgement: Some of the slides included here were taken or adapted from materials created by other BCBB team members (Andrew Oler and Celine Hong). Thank you BCBB!

3

Important developments in studying genomic variations: Timeline to Human Genome Project



https://www.mun.ca/biology/scarr/Human_Genome_Project_timeline.html

Genomic variations have been studied in large populations studies

International HapMap Project

- Project that started on 2002 with the goal of describing patterns of human genetic variation and create a haplotype map using SNPs present in at least 1% of the population, which were deposited in dbSNPs.

1000 Genomes
A Deep Catalog of Human Genetic Variation

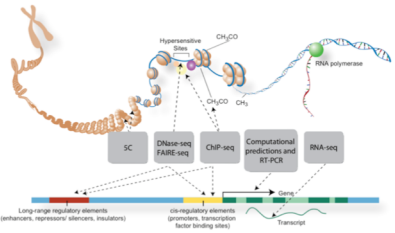
- Started in 2008 with a goal of using at least 1000 individuals (about 2,500 samples at 4X coverage), interrogate 1000 gene regions in 900 samples (exome analysis), find most genetic variants with allele frequencies above 1% and to a 0.1% if in coding regions as well as Indels and structural variants
- Make data available to the public ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/
  or via Amazon Cloud http://s3.amazonaws.com/1000genomes

Variants are not only found in coding sequences

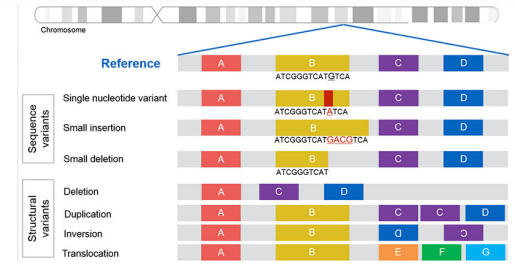
ENCODE

Main Goal:

- Find all functional elements in the genome

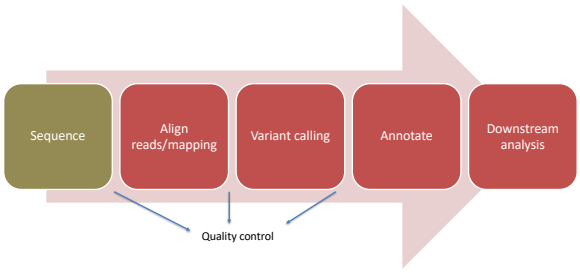


There are many types of variants



http://www.labspaces.net/blog/1627/When_Whole_Genome_Sequencing_Doesn___t_Give_Us_the_Whole_Genome

Overview of next-generation sequencing processing for variant analysis



Sequencers

Machine	Cost	methods	Throughput per run	Read length	Error rate
Illumina MiSeq	$128K	Small genomes, targeted gene	1.5-2Gb	2X300	0.8%
Ion Torrent	$80K	Small genomes, targeted gene	1Gb	400	1.71%
Illumina NextSeq	$250K	Exomes/transcriptome	120Gb	2x150	0.8%
Illumina HiSeq	$654K	Genomes/exomes/transcriptomics	600Gb	2X150	0.76%
Illumina X Ten	$10 Mil.	Genomes	1.6Tb	2x150	0.5%
PacBio	$695K	Genomes	100Mb	15K	12.86%

Next-generation sequencing overview

	Exomes (ES/WES)	Genomes (GS/WGS)
Cost	~$1000	~$2000 (~$1000 with HiSeq X Ten)
Size of bam files	~10 Gb	~200 Gb
DNA	Targeted and captured	Sheared DNA
What you can get	Most coding regions (+UTR)	coding and non-coding
Variants that can be examined	SNVs, indels (CNVs)	SNVs, indels, CNVs, structural variations

Processing reads for Variant Analysis

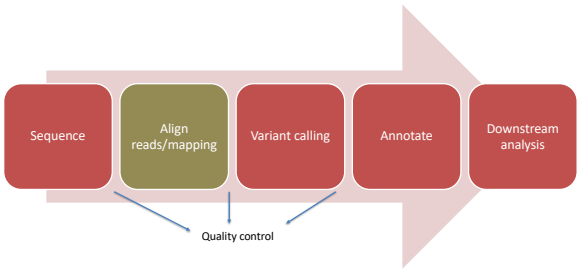
- Trimming:
  - Unless quality of read data is poor, low quality end trimming of short reads (e.g. Illumina) prior to mapping is usually not required.
  - If adapters are present in a high number of reads, these could be trimmed to improve mapping.
    - *Recommended tools: Prinseq, Btrim64, FastQC*

- Marking duplicates:
  - It is not required to remove duplicate reads prior to mapping but instead it is recommended to mark duplicates after the alignment.
    - *Recommended tool: Picard.*
  - A library that is composed mainly of PCR duplicates could produce inaccurate variant calling.

http://www.biomedcentral.com/1471-2164/13/S8/S8

Clip: http://bestclipartblog.com

Overview of next-generation sequencing processing for variant analysis



Variant analysis can be done using the entire genome (WGS) or only coding sequences (exome)

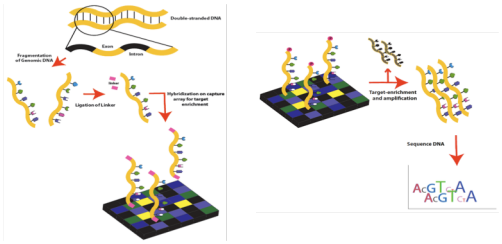
Whole Genome

- Useful for identifying SNPs but also structural variants in any genomic region
- It is routinely done when studying genomes from bacteria or viruses

Exome-Seq

- Interrogates variants in specific genomic regions, usually coding sequences.
- Cheaper and provides high coverage

Exome sequencing overview



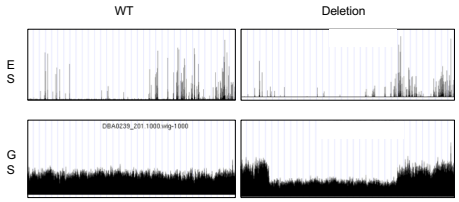
Ann https://en.wikipedia.org/wiki/Exome_sequencing#/media/File:Exome_Sequencing_workflow_1b.png. 2012 Jan; 71(1):5-14

Partial list of capture enrichment kits

Manufacturer	Kits	Regions targeted	Bases covered
Illumina	Nextera Rapid capture	Exons + UTRs+ miRNA	62 Mb
Nimblegen	SeqCap EZ Exome	Exons + UTR	96 Mb
	SeqCap EZMedExome	Disease-associated regions	47 Mb
Agilent	SureSelect Human All Exon V6	Exons+UTRs	60Mb
	Clinical Research Exome	Disease-relevant targets	51Mb

How exomes and genomes look

- Genome has even coverage
- Even a deletion is observed by eye



All about exomes

- What is exome?
  - Sequencing targeted exonic regions
  - ~2% of genome
- Important to know
  - You will NOT get a whole exome!
  - Not all exons in all genes are captured!
  - Important to know the negative results vs no data
- Coverage will vary in targets

What to know about your data

- What is the sequence depth?
  - The depth your sequence
  - 10x, 30x, 50x, 100x
- Read length
  - How long is your read length?
- What software was used to align?
- Which reference was used?
- Which capture kit was used? What is covered?
- What variant calling was used to call variants?

Files and tools used for SNPs and Small INDELs

File type	Origin
FASTQ	Raw reads from sequencer
SAM BAM	Sequence Alignment/Map Binary version of SAM
gVCF/VCF	Variant call format

Tool	Purpose
BWA mem	Read alignment to reference
Picard	Mark duplicates
GATK (Haplotype caller)	QC/variant calling
Samtools	Sort sam/bam, convert sam<->bam

How to identify structural variants?

Try using tools which leverage information on read pairs or depth.

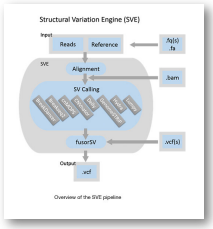
1- Read Pair – use pair ends.
GASVPro (Sindi 2012), BreakDancer (Chen 2009)

2- Read depth – CNVnator (Abyzov, 2010)

3- Split read – PinDEL (Ye 2009)

4- Denovo assembly

FUSORSV



20

VCF format (version 4.3)

- Format used to report information about a position in the genome
- Use by 1000 genomes to report all variants

VCF FORMAT

The Variant Call Format (VCF) is a TAB-delimited format with each data line consists of the following fields:

Col	Field	Description
1	CHROM	CHROMosome name
2	POS	the left-most POSition of the variant
3	ID	unique variant IDentifier
4	REF	the REFerence allele
5	ALT	the ALTernate allele(s), separated by comma
6	QUAL	variant/reference QUALity
7	FILTER	FILTers applied
8	INFO	INFOrmation related to the variant, separated by semi-colon
9	FORMAT	FORMAT of the genotype fields, separated by colon (optional)
10+	SAMPLE	SAMPLE genotypes and per-sample information (optional)

21

VCF format

See http://samtools.github.io/hts-specs/VCFv4.3.pdf

```
[HEADER LINES]
#CHROM  POS     ID          REF     ALT     QUAL    FILTER  INFO            FORMAT          NA12878
chr1    873762          T       G       5231.78 PASS    [ANNOTATIONS]   GT:AD:DP:GQ:PL  0/1:173,141:282:99:255,0,255
chr1    877664  rs3828047       A       G       3931.66 PASS    [ANNOTATIONS]   GT:AD:DP:GQ:PL  1/1:0,105:94:99:255,255,0
chr1    899282  rs28548431      C       T       71.77   PASS    [ANNOTATIONS]   GT:AD:DP:GQ:PL  0/1:1,3:4:25.92:103,0,26
chr1    974165  rs9442391       T       C       29.84   LowQual [ANNOTATIONS]   GT:AD:DP:GQ:PL  0/1:14,4:14:60.91:61,0,255
```

How variation is represented in a VCF

Each line represents one variant (here everything is a SNP, but some could be indels or CNVs) as well as the genotype of our sample, NA12878, at that variant. I've chosen these four variants because they each represent an important aspect in interpreting a VCF file:

- chr1:873762 is a novel T/G polymorphism, found with very high confidence (QUAL = 5231.78).
- chr1:877664 is a known A/G SNP (rs3828047), found with very high confidence (QUAL = 3931.66)
- chr1:899282 is a known C/T SNP (rs28548431), but has a relative low confidence (QUAL = 71.77)
- chr1:974165 is a known T/C SNP but we have so little evidence for this variant in our data that although we write out a record for it (book keeping, really) our statistical evidence is so low that we filter the record out as a bad site "LowQual".

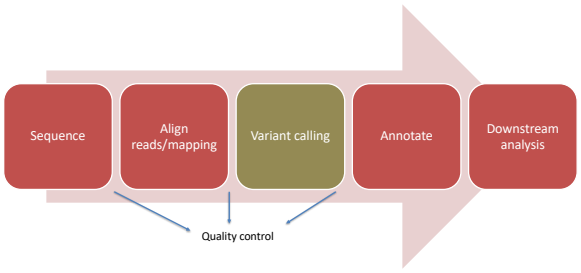
Popular tools for mapping, variant calling and annotation

Task	Popular tools
Align reads/mapping	BWA-mem Novoalign Isaac
Variant calling	GATK (Broad) Platypus (Wellcome Trust) Starling (Illumina)
Annotating variants	AnnoVar VEP SnpEff Gemini

Commonly used sources for reference genomes and annotations

Task	Popular source
Control population frequency	ExAC 1000GP ESP
Annotation	RefSeq Ensemble UCSC genes GENCODE
Visualization	UCSC genome browser IGV
Clinical relevance	HGMD OMIM CGD ClinVar
Variants and effect on gene expression	GTEX

Overview of next-generation sequencing processing for variant analysis



Example of variant calling with GATK

Variation in coagulation and fibrinolysis genes evaluated for their contribution to cerebrovascular complications in adults with bacterial meningitis in the Netherlands

A.T. Kloek [a], H.N. Khan [a, b], M. Valls Seron [a], A. Jongejan [b], A.H. Zwinderman [b], F. Baas [c], A. van der Ende [d], D. van de Beek [a, 1], R. B. Ferwerda [a, 1], M.C. Brouwer [a, 1]

Journal of Infection
Volume 77, Issue 1, July 2018, Pages 54-59

- Sequencing: Solid 5500xl sequencer (Life technologies)
- Mapping: Alignment to **GRch37/HG19** reference genome using the Lifescope aligner (version 2.5.1, Applied Biosystems).
- Preprocessing and QC: Reads were realigned, recalibrated and called using **GATK HaplotypeCaller** (version 3.3-0). Variants were filtered on minimal mapping quality score ≥30, minimal read depth of 20, allele balance between 0.2 and 0.8 for heterozygotes, >90 for homozygotes and genotype quality score of 99.
- Additional QC and Filtering: All samples were combined and converted to PLINK format.[18] Final QC was performed to exclude related individuals and samples with non-European ancestry based on a previous multidimensional scaling test.[18] Markers with a missingness of 5% and Hardy-Weinberg equilibrium (HWE) *p*-value of $< 1.0^{-5}$ were excluded from the analysis. For the regression analysis only markers with minor allele frequency **(MAF) ≥ 0.01** were included.

Example of variant calling and filtering with FreeBayes / samtools

frontiers in Genetics
Front Genet. 2018; 9: 147.
Published online 2018 Apr 24. doi: 10.3389/fgene.2018.00147
PMCID: PMC5928233
PMID: 29740479

A New Single Nucleotide Polymorphism Database for Rainbow Trout Generated Through Whole Genome Resequencing

Guangtu Gao,[1] Torfinn Nome,[2] Devon E. Pearse,[3] Thomas Moen,[4] Kerry A. Naish,[5] Gary H. Thorgaard,[6] Sigbjørn Lien,[2] and Yniv Palti[3]*

- **SNP quality filter (QUAL):** SNPs not located within 4 bases distance to an indel, and have the phred-scaled variant quality score, QUAL, larger than 30.
- **Low-complexity filter (LC):** We removed the SNPs that are in the low-complexity regions as using *mdust* (https://github.com/lh3/mdust
- **Maximum depth filter (DP):** maximum depth value of 1,500, corresponding to a coverage of 24.6 reads per sample
- **Minimum sample filter (NS):** Only SNP sites with sequence reads coverage from at least 58 of the 61 samples passed this filter,
- **Double haploid filter (DH):** Briefly, we filtered out SNP sites with heterozygous genotypes in at least two of the 11 DH lines that we re-sequenced in this study.

Additional consideration: germline vs somatic variants

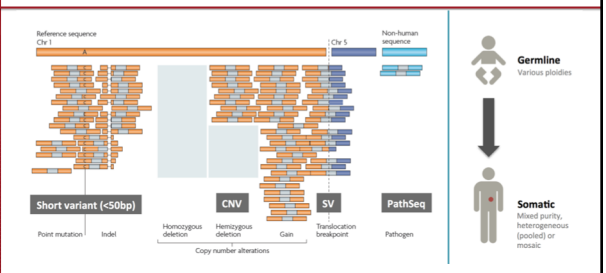


Image source: GATK Broad Institu

IGV browser: a must-have for variant inspection

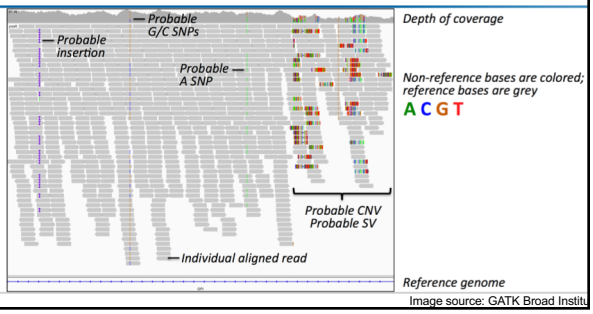


Image source: GATK Broad Institu

Reference genome

- There are different versions of human reference genome

Reference Name	Chr notation	Mitochondrial sequence	Additional sequences included
GRCh38 (Genome Reference Consortium)	1, 2… X,Y,MT	Yes	• Unlocalized • Unplaced • Alternate loci • Centromere sequences
hg19 (UCSC genome browser)	Chr1, Chr2…ChrX, ChrY, ChrM	Copied from previous release	• Unlocalized • Unplaced • Alternate loci
b37/b37+decoy/hs37d5 (1000GP)	1, 2… X,Y,MT	Yes	• Unlocalized • Unplaced • "decoy" sequence • Human herpesvirus 4 type 1

- Unlocalized: chromosome known, exact location unknown
- Unplaced: known to originate from human genome, chromosome unknown
- Alternate loci: alternate representation of specific human regions

Index files

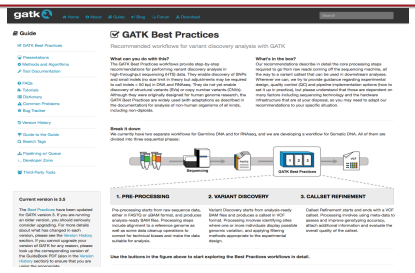
- Index files are needed for files in next-generation analysis, as file sizes are big!
- Enables program to efficiently access the data, rather than having to read the whole file

File type	Index
FASTA	*fai
BAM	*bai
VCF	*vcf.idx

Reference FASTA index file

```
1       249250621       52      60      61
2       243199373       253404903       60      61
3       198022430       500657651       60      61
4       191154276       701980507       60      61
5       180915260       896320740       60      61
6       171115067       1080251307      60      61
7       159138663       1254218344      60      61
8       146364022       1416009371      60      61
9       141213431       1564812846      60      61
10      135534747       1708379889      60      61
11      135006516       1846173603      60      61
12      133851895       1983430202      60      61
13      115169878       2119513896      60      61
14      107349540       2236602526      60      61
15      102531392       2345741279      60      61
16      90354753        2449981581      60      61
17      81195210        2541842300      60      61
18      78077248        2624390817      60      61
19      59128983        2703769406      60      61
20      63025520        2763803926      60      61
21      48129895        2827959925      60      61
22      51304566        2876892038      60      61
X       155270560       2929051733      60      61
Y       59373566        3086910193      60      61
MT      16569   3147273397      70      71
GL000207.1      4262    3147290265      60      61
GL000226.1      15008   3147294661      60      61
GL000229.1      19913   3147309982      60      61
GL000231.1      27386   3147330289      60      61
GL000210.1      27682   3147358194      60      61
GL000239.1      33824   3147386400      60      61
GL000235.1      34474   3147420850      60      61
```

GATK Best Practices Website

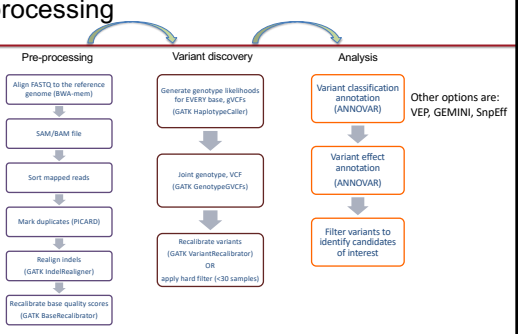


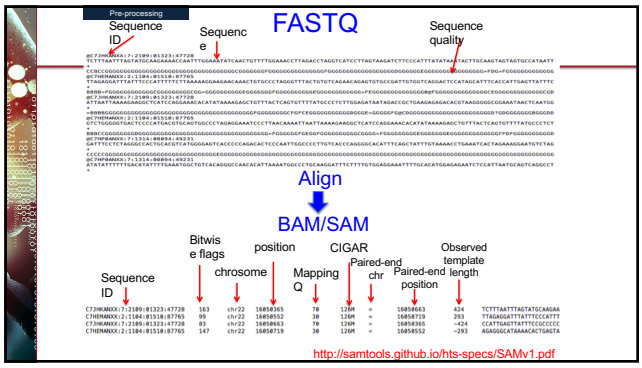
https://www.broadinstitute.org/gatk/guide/best-practices

Why BWA+ GATK Haplotype caller?

- Widely accepted as the “conventional” way of processing next-gen data
- Well assessed
- Well documented
- Software is supported
- Community support for troubleshooting or information

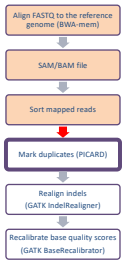
Data processing



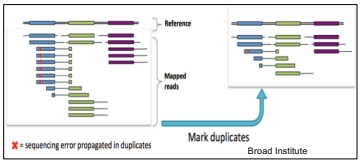


Pre-processing QC

Marking duplicates and why is it necessary?

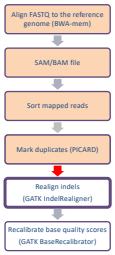


- It is assumed that each read corresponds to an independent DNA fragment from *randomly* sheared DNA
- However, PCR amplification can cause duplicates
  - Identify based on start + stop of reads
  - Choose the best and ignore the rest



Pre-processing QC

Indel realignment

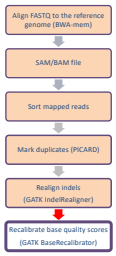


- Misalignment around indels cause high number of SNPs
- These regions are identified and locally realigned to minimize mismatches

Broad Institute

Pre-processing QC

Recalibrate Base Quality Scores



- Base scores are produced by sequencers
- Quality scores are inaccurate and biased
  - Prone to various technical errors
  - QS are often over- or under estimated
- To identify and correct non-random technical error
  - Physics or the chemistry of sequencing reactions
  - Manufacturing flaws in the equipment
- Error covariates e.g.
  - Reported quality score
  - Position within the read (machine cycle)
  - Preceding and current nucleotide (sequencing chemistry)

Broad Institute

Pre-processing QC

Recalibrate Base Quality Scores

Align FASTQ to the reference genome (BWA-mem)

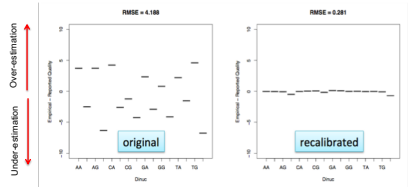
SAM/BAM file

Sort mapped reads

Mark duplicates (PICARD)

Realign indels (GATK IndelRealigner)

Recalibrate base quality scores (GATK BaseRecalibrator)

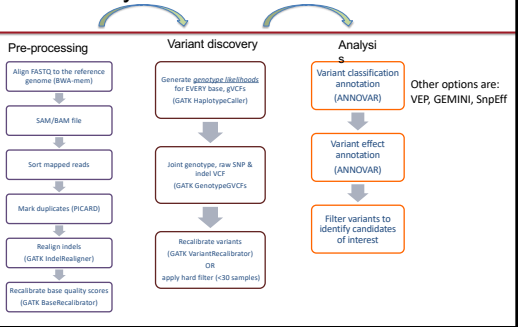


- GATK BSQR builds model based on the known variants set
- Adjusts the base quality scores in the data based on the model

Broad Institute

Now we are ready to call variants!

Variant Discovery



GATK uses a bayesian model

$$\text{Bayesian model}\begin{cases}\Pr\{G|D\} = \dfrac{\overbrace{\Pr\{G\}}^{\text{Prior of the genotype}}\overbrace{\Pr\{D|G\}}^{\text{Likelihood of the genotype}}}{\Sigma_i \Pr\{G_i\}\Pr\{D|G_i\}}, \text{ [Bayes' rule]} \\ \Pr\{D|G\} = \prod_j \left(\dfrac{\Pr\{D_j|H_1\}}{2} + \dfrac{\Pr\{D_j|H_2\}}{2}\right) \text{ where } \overbrace{G = H_1H_2}^{\text{Diploid assumption}} \\ \Pr\{D|H\} \text{ is the haploid likelihood function}\end{cases}$$

- It determines possible SNP and indel alleles
- Computes, for each sample, for each genotype, likelihoods of data given genotypes
- Computes the allele frequency distribution to determine most likely allele count, and emit a variant call if determined
  - When it reports a variant, it assign a genotype to each sample

http://www.broadinstitute.org/gatk//events/2038/GATKwh0-BP-5-Variant_calling.pdf

Variant Discovery

Generate genotype likelihoods in each sample (gVCF)

Generate *genotype likelihoods* for EVERY base, gVCFs
(GATK HaplotypeCaller)

↓

Joint genotype, raw SNP & indel VCF
(GATK GenotypeGVCFs)

↓

Recalibrate variants
(GATK VariantRecalibrator)
OR
apply hard filter (<30 samples)

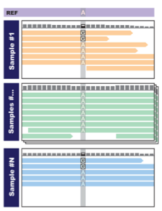
- For a single sample, calculates normalized Phred-scaled likelihoods (PL) for genotypes:
  - "likelihood of the genotype"= "the probability that the genotype is not correct"
  - Normalized so that the most likely genotype is 0

↓

```
16  105459  .  G  <NON_REF>  .  .  .  GT:AD:DP:GQ:PL  0/0:60,0:60:99:0,120,1800
16  105460  .  T  <NON_REF>  .  .  .  GT:AD:DP:GQ:PL  0/0:60,0:60:99:0,120,1800
16  105461  .  T  <NON_REF>  .  .  .  GT:AD:DP:GQ:PL  0/0:61,0:61:99:0,120,1800
16  105462  .  A  <NON_REF>  .  .  .  GT:AD:DP:GQ:PL  0/0:61,1:62:99:0,120,1800
16  105463  .  C  <NON_REF>  .  .  .  GT:AD:DP:GQ:PL  0/0:64,0:64:99:0,120,1800
16  105464  .  T  <NON_REF>  .  .  .  GT:AD:DP:GQ:PL  0/0:64,0:64:99:0,120,1800
16  105465  .  C  <NON_REF>  .  .  .  GT:AD:DP:GQ:PL  0/0:66,0:66:99:0,120,1800
```


Variant Discovery

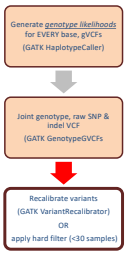
What is a joint genotyping?



- If we analyze Sample 1 or Sample N alone, we are not confident that the variant is real
- If we see both samples, we are more confident that there is real variation at this site in this cohort

Variant Discovery-QC

Generate new variant quality score using VQSR

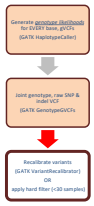


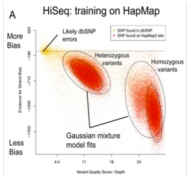
- What is Variant Quality Score Recalibration
  - NOT adjusting scores!!
  - Generate new score VQSLOD (variant quality score log-odds)
- Approach
  - Machine learning to profile good variants vs bad variants
  - Using multiple dimensions (5-8, typically)
  - Uses INFO annotations for each variant (eg. Allele count, allele frequency, etc)

Broad Institute

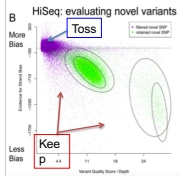
Variant Discovery-QC

Generate new variant quality score using VQSR





Train a model using "truth" set of known variants

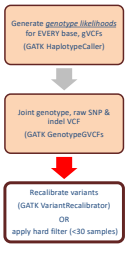


Apply the model to your samples

http://gatkforums.broadinstitute.org/gatk/discussion/39/variant-quality-score-recalibration-

Variant Discovery-QC

Alternatively! IF you have few samples use multiple filters



- Apply hard filter!
- Define how to filter your variant or use default filter parameters
  - QualByDepth (QD) 2.0
  - FisherStrand (FS) 60.0
  - RMSMappingQuality (MQ) 40.0
  - MappingQualityRankSumTest (MQRankSum) -12.5
  - ReadPosRankSumTest (ReadPosRankSum) -8.0 (only het calls)

http://gatkforums.broadinstitute.org/gatk/discussion/39/variant-quality-score-recalibration-

Variant Discovery-Final VCF

Final VCF

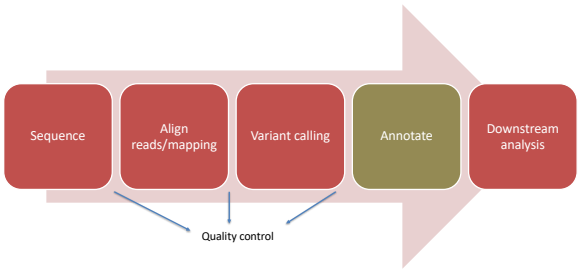
Header

```
##fileformat=VCFv4.1
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
##GATKCommandLine.HaplotypeCaller=<ID=HaplotypeCaller,Version=3.4-3-gd1ac142,Date="Mon May 18 17:36:4
.
.
.
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##contig=<ID=chr1,length=249250621,assembly=b37>
##reference=file:human_genome_b37.fasta
```

Body

```
#CHROM  POS ID      REF ALT QUAL    FILTER  INFO        FORMAT      NA12878
1   873762  .       T   G   5231.78 PASS    [ANNOTATIONS] GT:AD:DP:GQ:PL  0/1:173,141:282:99:255,0,255
1   877664  rs3828047   A   G   3931.66 PASS    [ANNOTATIONS] GT:AD:DP:GQ:PL  1/1:0,105:94:99:255,255,0
1   899282  rs28548431  C   T   71.77   PASS    [ANNOTATIONS] GT:AD:DP:GQ:PL  0/1:1,3:4:26:103,0,26
1   974165  rs9442391   T   C   29.84   LowQual [ANNOTATIONS] GT:AD:DP:GQ:PL  0/1:14,4:14:61:61,0,255
```

Overview of next-generation sequencing processing for variant analysis



Variant Annotation

Annotating Variants

Variant classification annotation (e.g. ANNOVAR, SnpEff or GEMINI)

Variant effect annotation

Filter variants to identify candidates of interest

Gene Annotation	refSeq Gene, mitochondrial variants, USC/EnSembl, GENCODE/CCDS
Region-based Annotation	Conserved genomic elements, transcription factor binding site, cytogenetic band, segmental duplications, GWAS..
Filter-based annotation	1000 GP, dbSNP, ESP, EXAC, non-synonymous variants annotation (SIFT/Polyphen2/MutationTaster/LRT/FATHMM/CADD..), ClinVar...

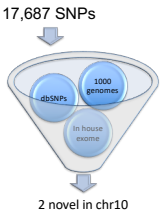
For comprehensive list, see http://annovar.openbioinformatics.org/en/latest/

Variant Analysis… like finding a needle in a 'deep' haystack

Look for evidence of variants of interest

Further filters needed

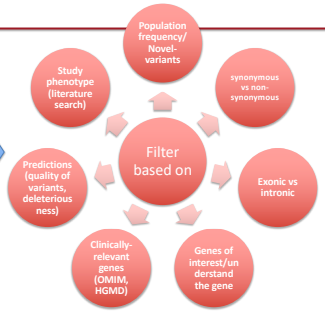
- High number of variants
- The goal is to narrow down your list of variants
- Eliminate variants that are not interesting



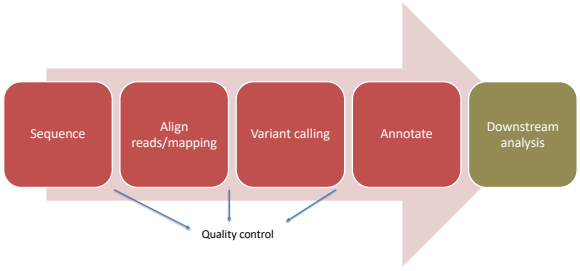
PLoS One. 2012;7(1):e29708

Things to consider

Filter variants to identify candidates of interest



Overview of next-generation sequencing processing for variant analysis

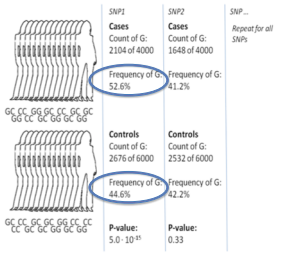


Additional Analysis

- Trio/Family Based Analysis
  - PhaseByTransmission (GATK)
  - GEMINI https://gemini.readthedocs.org/en/latest/
  - pVAAST
- Somatic Variation
  - Tumor/normal
  - MuTech http://archive.broadinstitute.org/cancer/cga/mutect
- Association Studies
- Copy number variation analysis

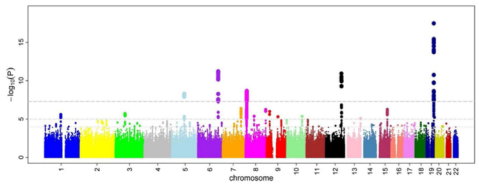
GWAS Association studies

- A typical analysis:
  - Identify SNPs where one allele is significantly more common in cases than controls
  - Hardy-Weinberg Chi Square



http://en.wikipedia.org/wiki/Genome-wide_association_studyhttp://www.ebi.ac.uk/gwas/

GWAS manhattan plot



- Only applicable for large cohort

CNVs from exome

- High variability of read-depth in exomes
- CNV prediction is very challenging problem
  - High false positives
  - Break-points limited due to targets
  - Long range CNVs have higher positive predictive value
- Useful when other alternatives (SNPs or aCGH) are not available
- Lots of prediction tools!! (some examples below)

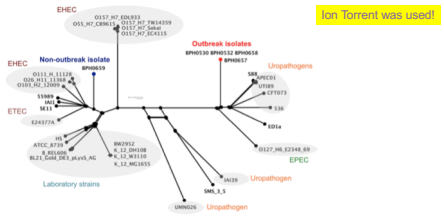
Population Caller	Somatic Caller
XHMM	ADTEx
CoNIFER	ExomeCNV
EXCAVATOR	Varscan2
ExomeDepth	Control-FREEC
CONTRA	

51

Deep Sequence: Fast genotyping

Outbreak investigation using high-throughput genome sequencing within a diagnostic microbiology laboratory. J Clin Microbiol. 2013 Feb 13

Ion Torrent was used!



52

A more integrative approach

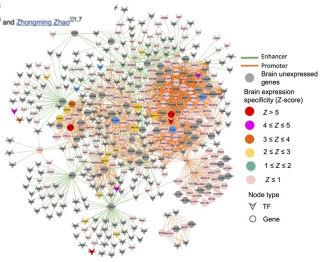
Genome Med. 2018; 10: 7.
Published online 2018 Jan 29. doi: 10.1186/s13073-018-0513-x
PMCID: PMC5789733
PMID: 29378629

An integrative functional genomics framework for effective identification of novel regulatory variants in genome–phenome studies

Junfei Zhao,#1 Feixiong Cheng,#2,3 Peilin Jia,1 Nancy Cox,4,5 Joshua C. Denny,5,6 and Zhongming Zhao[1,7]

A brain-specific TF-target gene regulatory network for AD

The TF-target gene network was generated by mapping the significant SNPs with AD from the PheWAS Catalog into the enhancer or promoter regions via three components: (1) genome-wide mapping of promoters and enhancers; (2) linking TFs to promoters and enhancers; and (3) linking enhancers and promoters to target genes



Let's practice