WORKSHOP: Variant calling in humans, animals and plants with Galaxy

Dr Gareth Price
Head of Computational Biology, QFAB
Galaxy Australia – Service Manager

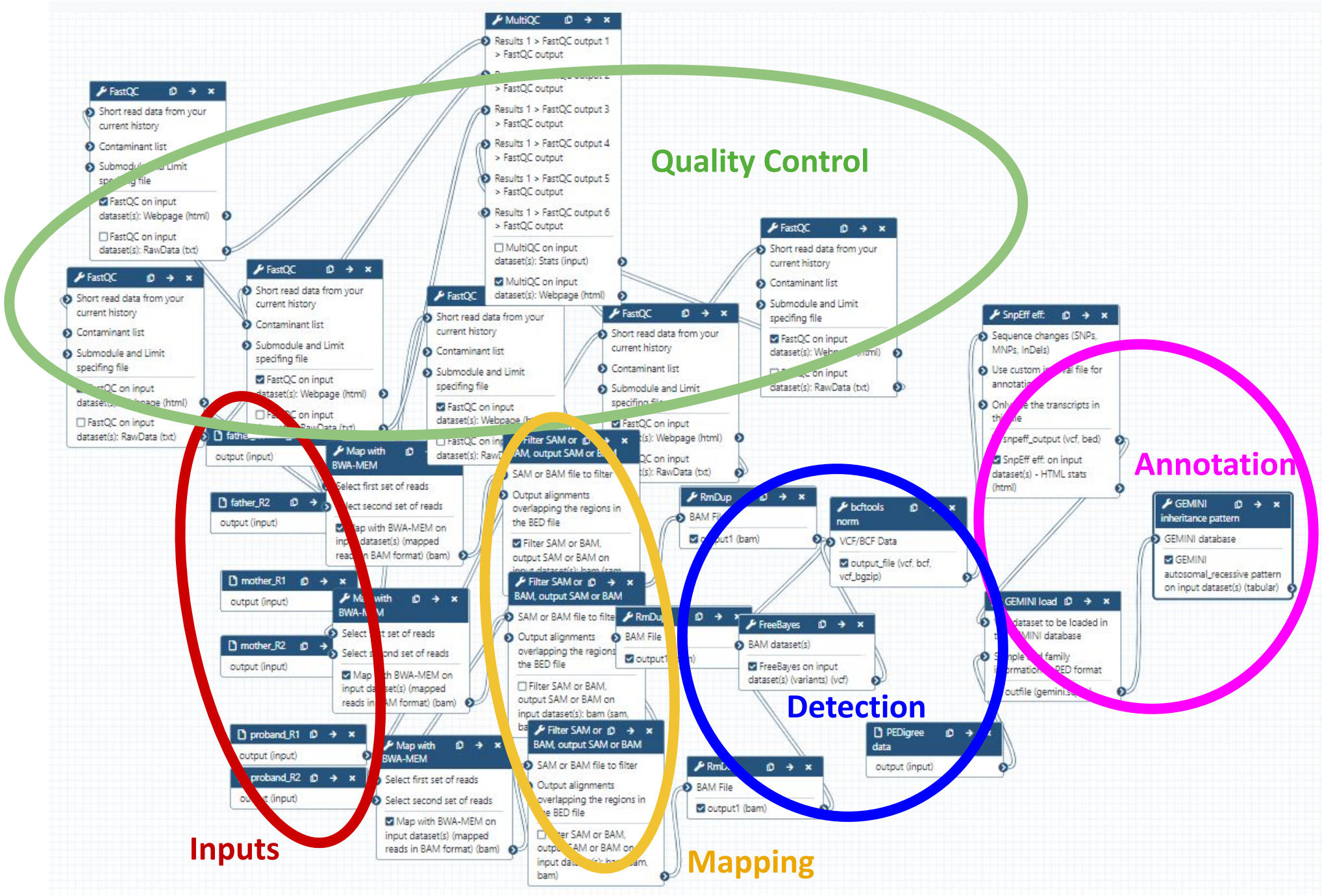
25th May 2021

QCIF BIOINFORMATICS

Galaxy AUSTRALIA

Variant Detection and Annotation in a polyploid organism

- **Questions**
  - How do you identify genetic variants in samples based on exome sequencing data?
  - How do you, among the set of detected variants, identify candidate causative variants for a given phenotype/disease?
- **Objectives**
  - Jointly call variants and genotypes for a family trio from whole-exome sequencing data
  - Use variant annotation and the observed inheritance pattern of a phenotype to identify candidate causative variants and to prioritize them
- **Genome vs Exome**
  - More depth of sequencing concentrated on variants presumed to have a phenotypic influence
  - Cheaper
  - Requires knowledge of genome to build exome capture probes
- **SNP vs SNV**
  - Single Nucleotide Polymorphism: Implied population frequency
  - Single Nucleotide Variant: Observed reference difference


Quality Control
Inputs
Mapping
Detection
Annotation
MultiQC
FastQC
Map with BWA-MEM
Filter SAM or BAM, output SAM or BAM
RmDup
bcftools norm
FreeBayes
SnpEff eff:
GEMINI inheritance pattern
GEMINI load
PEDigree data
father_R2
mother_R1
mother_R2
proband_R1
proband_R2

Assumptions - Galaxy and Biology

- **Working knowledge of Galaxy Australia**
  - GTN: Introduction to Galaxy Analyses
  - https://training.galaxyproject.org/training-material/topics/introduction/

- **Quality Control assessment of Illumina short-read sequence data**
  - GTN: Quality Control
  - https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html

- **Mapping of reads to a reference genome**
  - GTN: Mapping (using Bowtie2)
  - https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/mapping/tutorial.html#map-reads-on-a-reference-genome

Tools for Variant Analysis

- **Freeware**
  - **Genome Analysis Toolkit (GATK)**
  - **Virtual Labs / Machines**
    - Galaxy
    - R Studio (Bioconductor)
    - Command Line
- **Commercial**
  - **Agilent**
    - Cartagenia Bench Lab for Molecular Pathology
  - **Illumina**
    - BaseSpace
  - **Qiagen**
    - CLC-Bio Suite of Analysis Products
    - Ingenuity Variant Analysis
    - ANNOVAR
  - **ThermoFisher**
    - Ion Reporter

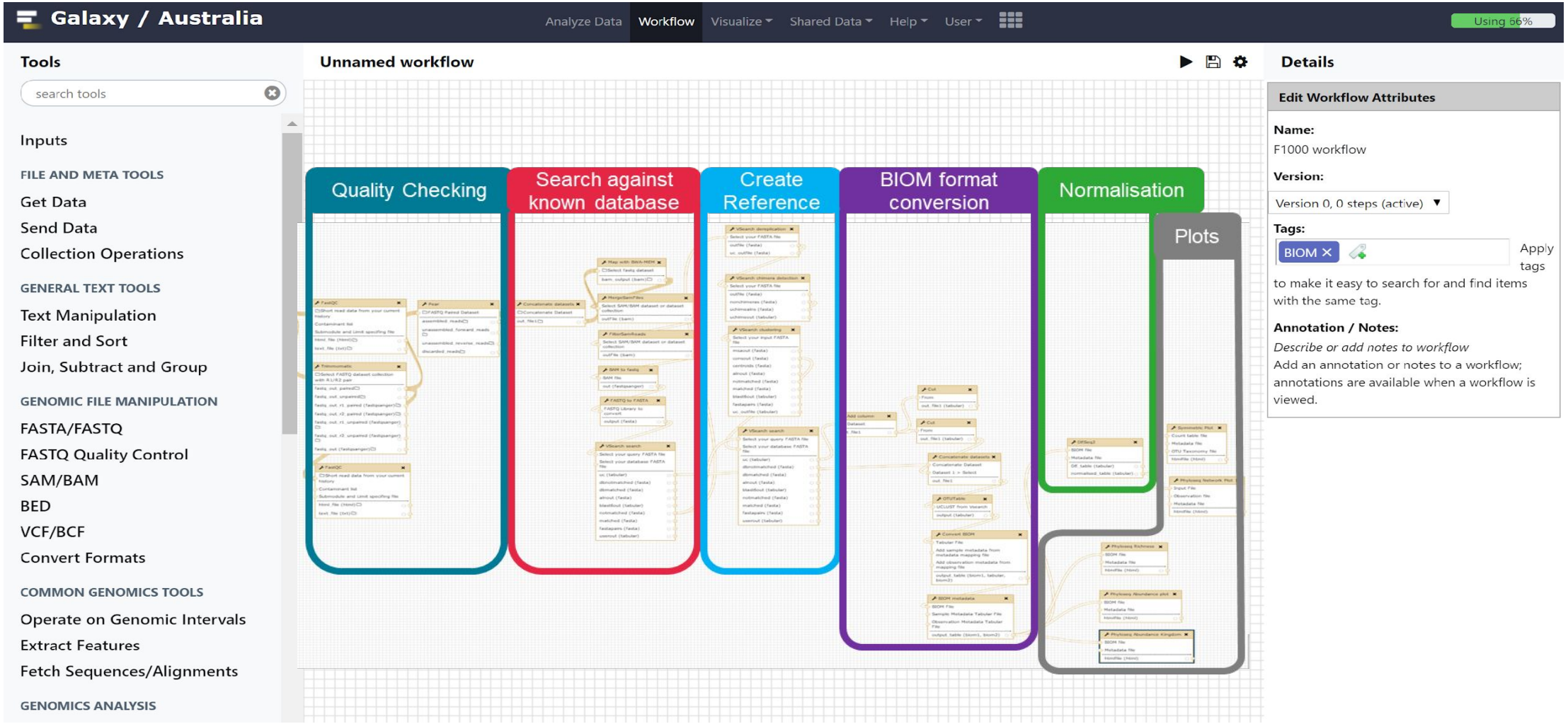
  - ~~**Microsoft**~~
    - ~~Excel~~

Galaxy Australia

QCIF BIOINFORMATICS

Efficiency through Galaxy controlled scheduling

Galaxy AUSTRALIA



Login to Galaxy Australia

Log into Galaxy Australia and join training server

1. Galaxy Australia: https://usegalaxy.org.au

2. Click on this link: https://usegalaxy.org.au/join-training/variants-polyploid/

Congratulations: you are successfully registered in **variants-polyploid**
Return to Galaxy

How It Works

We have deployed a dedicated compute node just for your training session to use. No one outside of your training has access to this machine. Completely transparently to you, all of the jobs that you run in Galaxy, during the period of the training event, will "prefer" to run on this machine. If there is no room on that machine, they will run on any other machine in our cluster with resources.

How can we do so much variant analysis?

- **7,558bn –** DNA bases are output by the Sequencing Centre every day
- **588 –** our Sequencing Centre provided the equivalent of gold-standard (30x) human genomes a week
- ***every 17 mins –*** *we read the equivalent of a single gold-standard (30x) human genome*

https://www.wellcomegenomecampus.org/scienceandinnovation/achievements-uniqueness/

You know you've made it when..

Miller *et al. Genome Medicine* (2015) 7:100
DOI 10.1186/s13073-015-0221-8

Genome Medicine

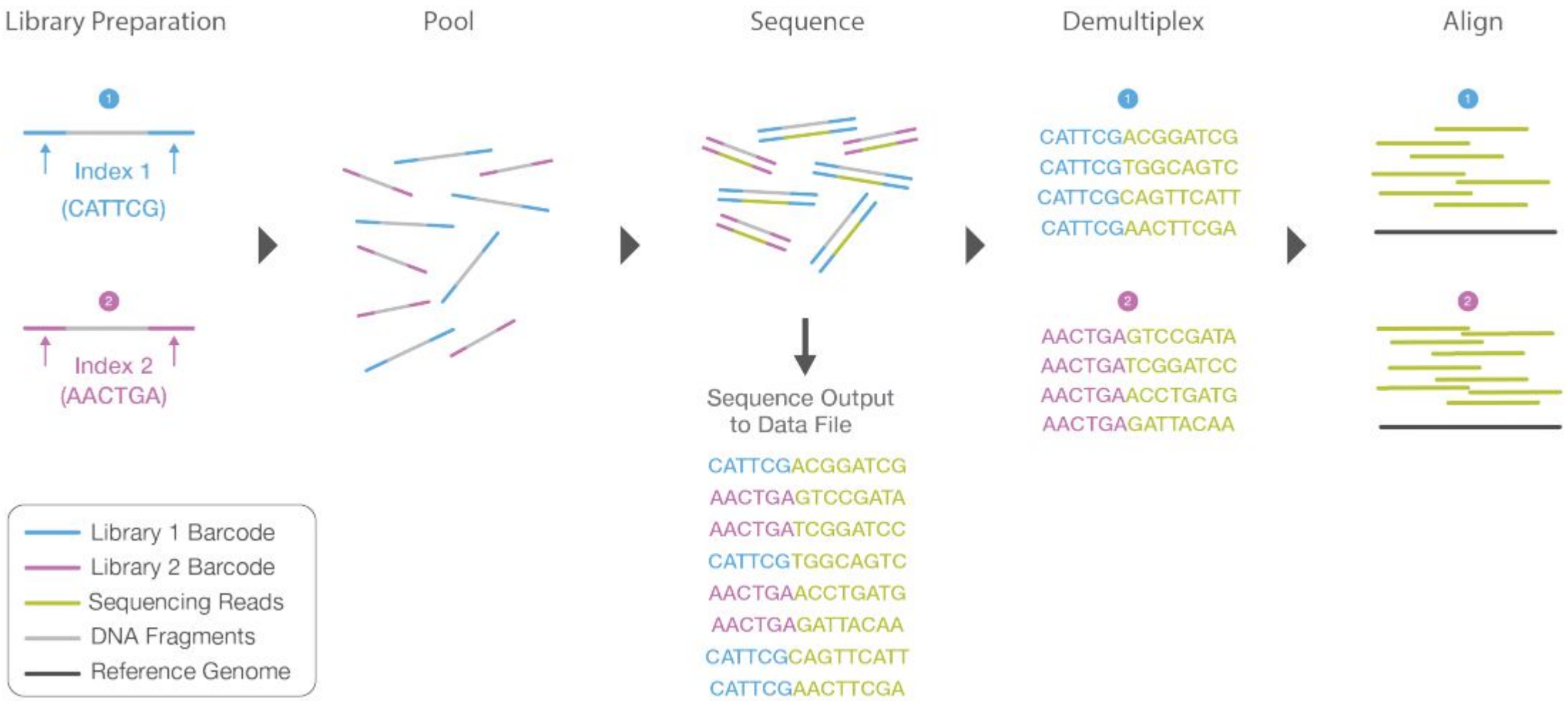
METHOD Open Access

A ~~26~~ 19.5-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases

Neil A. Miller[1†], Emily G. Farrow[1,2,3,4†], Margaret Gibson[1], Laurel K. Willig[1,2,4], Greyson Twist[1], Byunggil Yoo[1], Tyler Marrs[1], Shane Corder[1], Lisa Krivohlavek[1], Adam Walter[1], Josh E. Petrikin[1,2,4], Carol J. Saunders[1,2,3,4], Isabelle Thiffault[1,3], Sarah E. Soden[1,2,4], Laurie D. Smith[1,2,3,4], Darrell L. Dinwiddie[5], Suzanne Herd[1], Julie A. Cakici[1], Severine Catreux[6], Mike Ruehle[6] and Stephen F. Kingsmore[1,2,3,4,7*]

- *Dr. Kingsmore receives the GUINNESS WORLD RECORDS™ certificate for the fastest genetic diagnosis.*
- **San Diego—Feb. 12, 2018**
- https://www.rchsd.org/about-us/newsroom/press-releases/new-guinness-world-records-title-set-for-fastest-genetic-diagnosis/

Generalised NGS workflow



https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

File type definitions

FASTA
typical file extension: .fasta
text file, often gzipped (.fasta.gz)
very simple format for **DNA/RNA** or **protein** sequences

>gi**|12345678|**gb|AA0123567**.**1**|** cytochrome b [Homo sapien]
AGTAGTAGATGATAGAGCTCAGCTACGACT

FASTQ
typical file extension: .fastq, .fq
text file, often gzipped (–> .fastq.gz)
contains raw read information – 4 lines per read:

- read ID
- base calls
- additional information or empty line
- sequencing quality measures - 1 per base call

note that there is no information about where in the genome the read originated from

File type definitions

BAM (Binary Alignment Map)

typical file extension: .bam

contains information about sequenced reads (typically) *after alignment* to a reference genome

each line = 1 mapped read, with information about:

- its mapping quality (how likelihood that the reported alignment is correct)
- its sequencing quality (the probability that each base is correct)
- its sequence
- its location in the genome

highly recommended format for storing data

VCF (Variant Call File)

typical file extension: .vcf

Contains information on the reference the variants are derived from plus observational information to allow for filtering of variants

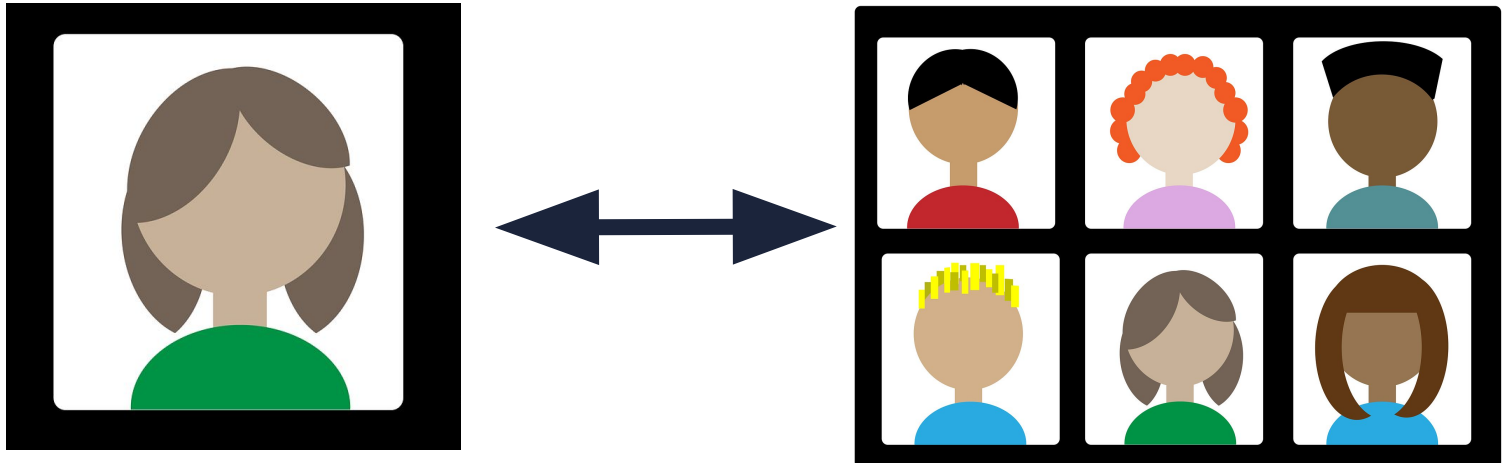
Header

```
##fileformat=VCFv4.1
##fileDate=20110413
##source=VCFtools
##reference=file:///refs/human_NCBI36.fasta
##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens">
##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
```

Body

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	40	PASS	.	GT:DP	1/1:13	2/2:29
1	2	.	C	T,CT	.	PASS	H2;AA=T	GT	0|1	2/2
1	5	rs12	A	G	67	PASS	.	GT:DP	1|0:16	2/2:20
X	100	.	T	<DEL>	.	PASS	SVTYPE=DEL;END=299	GT:GQ:DP	1:12:.	0/0:20:36

Breakout Rooms



Say hello!

Turn on your camera and microphone and introduce yourself

Breakout Rooms

Training materials: **https://tinyurl.com/variant-polyploid-materials**

Galaxy Training! | Variant Analysis | Help | Extras | Search Tutorials

Exome sequencing data analysis for diagnosing a genetic disease

By: Wolfgang Maier, Bérénice Batut, Torsten Houwaart, Anika Erxleben, Björn Grüning

Overview

Questions
- How do you identify genetic variants in samples based on exome sequencing data?
- How do you, among the set of detected variants, identify candidate causative variants for a given phenotype/disease?

Objectives
- Jointly call variants and genotypes for a family trio from whole-exome sequencing data
- Use variant annotation and the observed inheritance pattern of a phenotype to identify candidate causative variants and to prioritize them

Requirements
- Introduction to Galaxy Analyses
- Sequence analysis
  - Quality Control: slides - hands-on
  - Mapping: slides - hands-on

Time estimation: 5 hours

Supporting Materials

Topic Overview slides | Datasets | Workflows | Available on these Galaxies

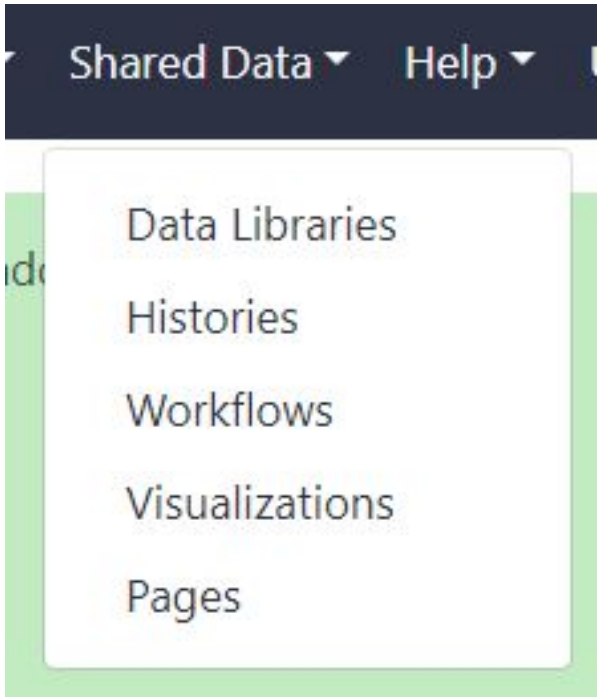
Last modification: Mar 12, 2021

OPEN CHAT

Getting Data into Galaxy

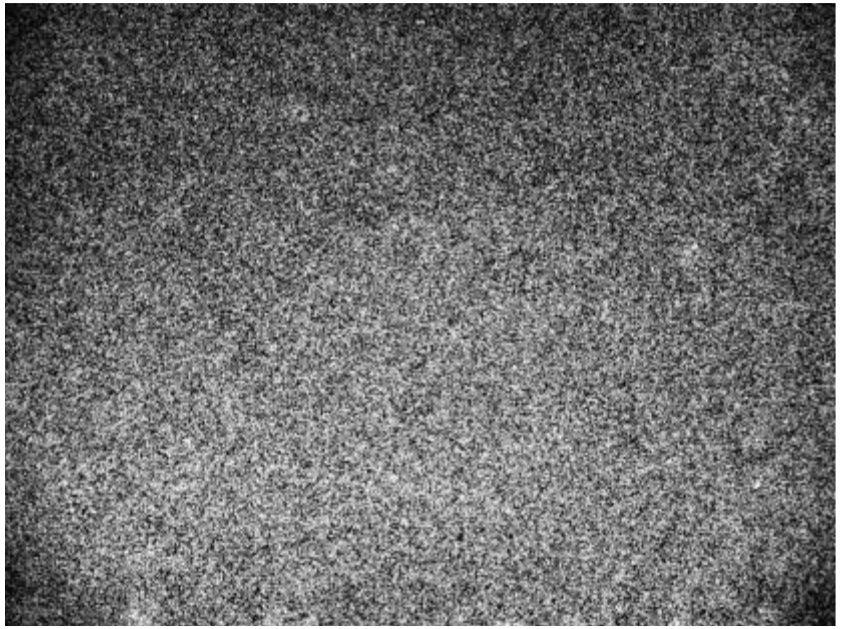
We are doing:

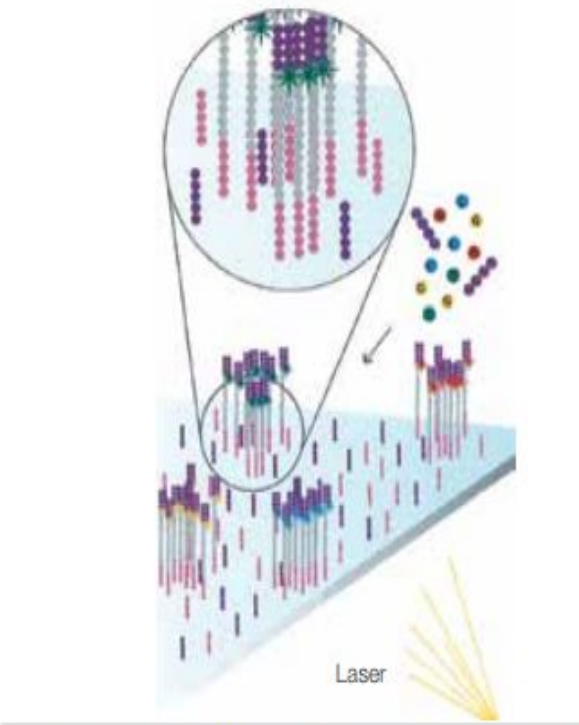
- Data Import (from Libraries)
  - GTN Material, page 2
  - Variant Analysis
  - Exome Sequencing…
  - Import: Pedigree, Father, Mother, Proband
  - Set Database: hg19
  - Add hashtags
- Data Preparation
- ~~Quality Control~~
- ~~Read Mapping~~
- Mapping reads postprocessing
- Variant Calling
- Remainder of tutorial….



Quality Control

- Quality Control assessment of Illumina short-read sequence data
  - GTN: Quality Control
  - https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html





Cycle	Call	Result
1	G G G G	= G
2	A A A A	= A
3	G - G G	= G
4	T G T T	= T
5	C T C C	= C
6	A C - A	= A?
7	G A A G	= G / A?
8	T G G T	= T / G?

FASTQ format

A ***read*** is a sequence with quality score values produced by a sequencing machine

Multiple reads in a single FASTQ file

Each read is described by four lines

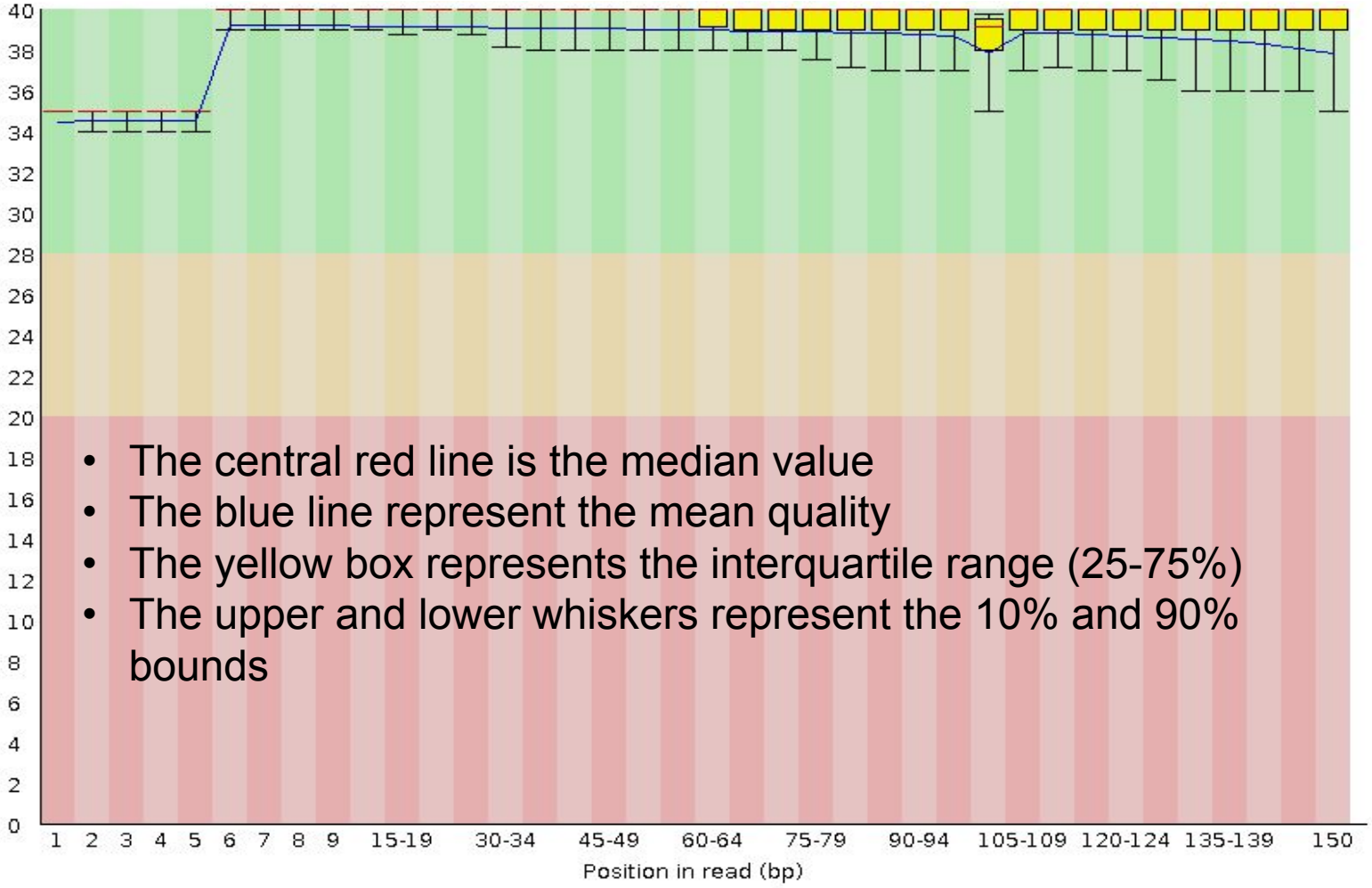
```
@SRR3145.19 ILLUMINA-C32_FC:3:1:80:12/1
TAGCAGCACATCATGGTTTACATCGTATGC
+
IIHIDIIIIIIIIIIIIIIHIHIIIIIDGIB
```

Name always starts with @
Sequence
Always starts with +; may have name
Encoded Phred quality score

FastQC output

very good quality

Phred Score



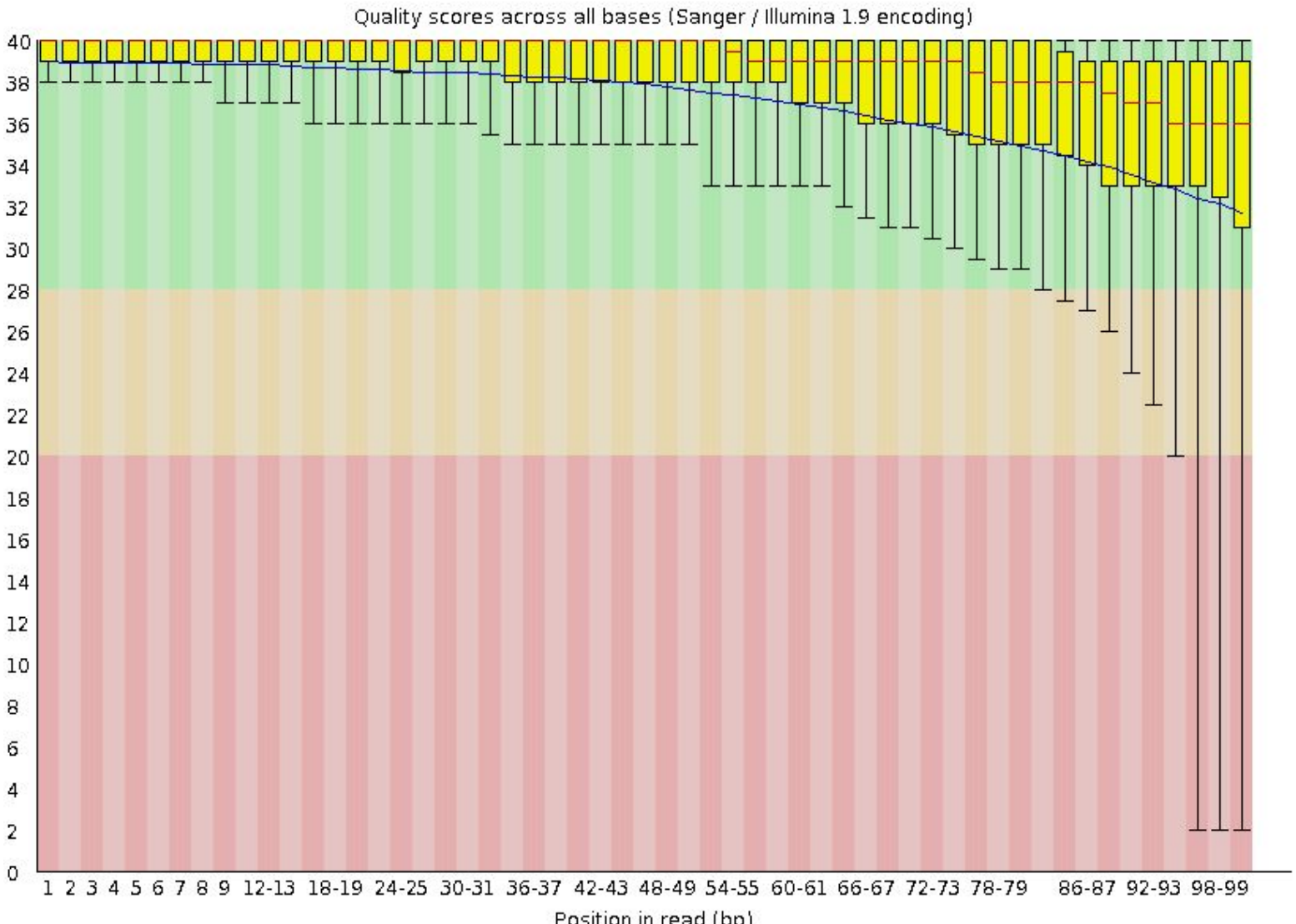
- The central red line is the median value
- The blue line represent the mean quality
- The yellow box represents the interquartile range (25-75%)
- The upper and lower whiskers represent the 10% and 90% bounds

Good region

Acceptable region

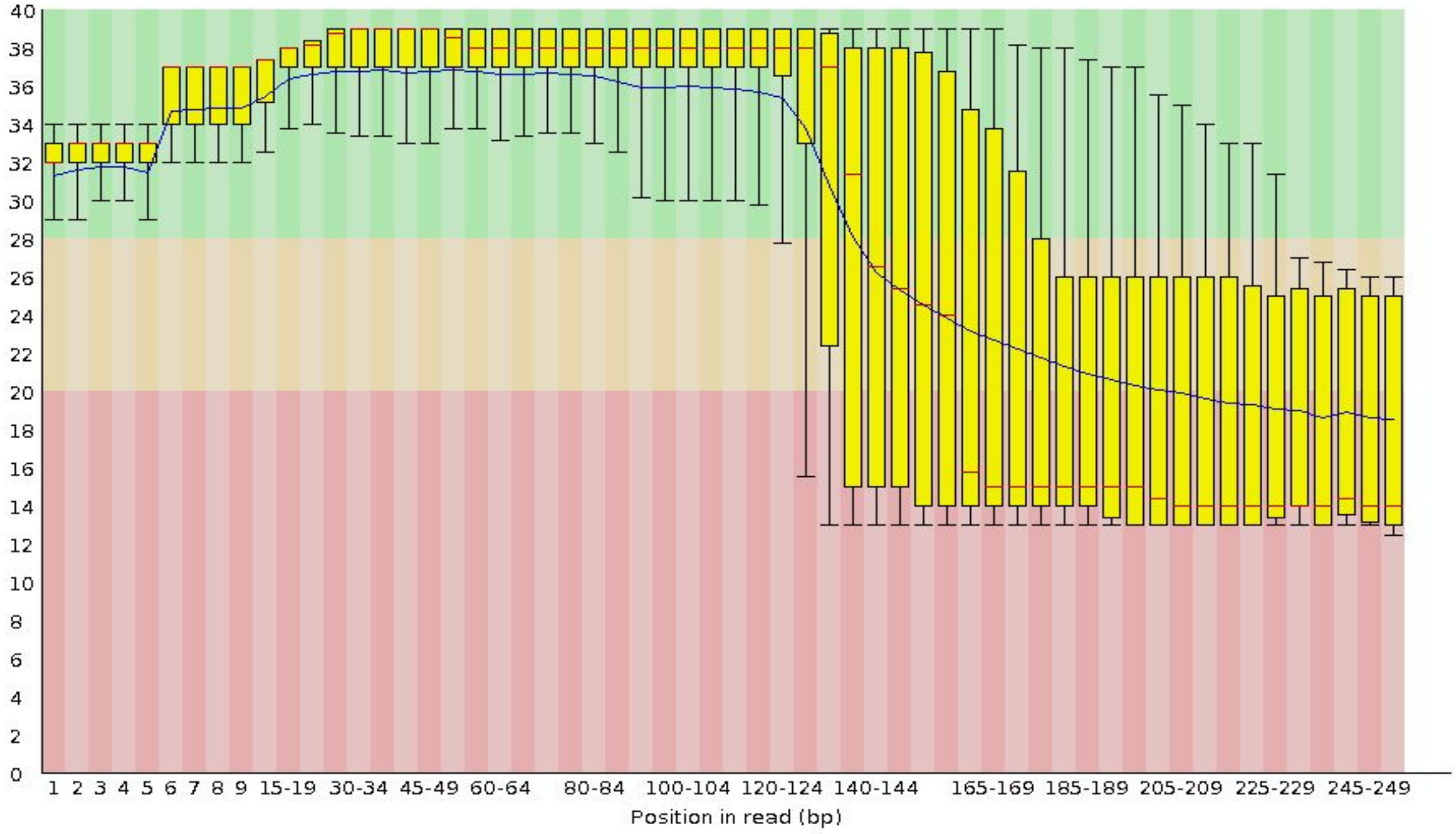
Bad region

FastQC output
typical quality



FastQC output

bad quality



FASTQ Emoji (FASTQE)

FASTQ + Emoji = FASTQE 🤔

Compute quality stats for FASTQ files and print those stats as emoji... for some reason.

Scores can also be binned:

Bin	Emoji
N	🚫
2-9	💀
10–19	💩
20–24	⚠️
25–29	😄
30–34	😆
35–39	😎
≥ 40	😍

FASTQE Report 🤔

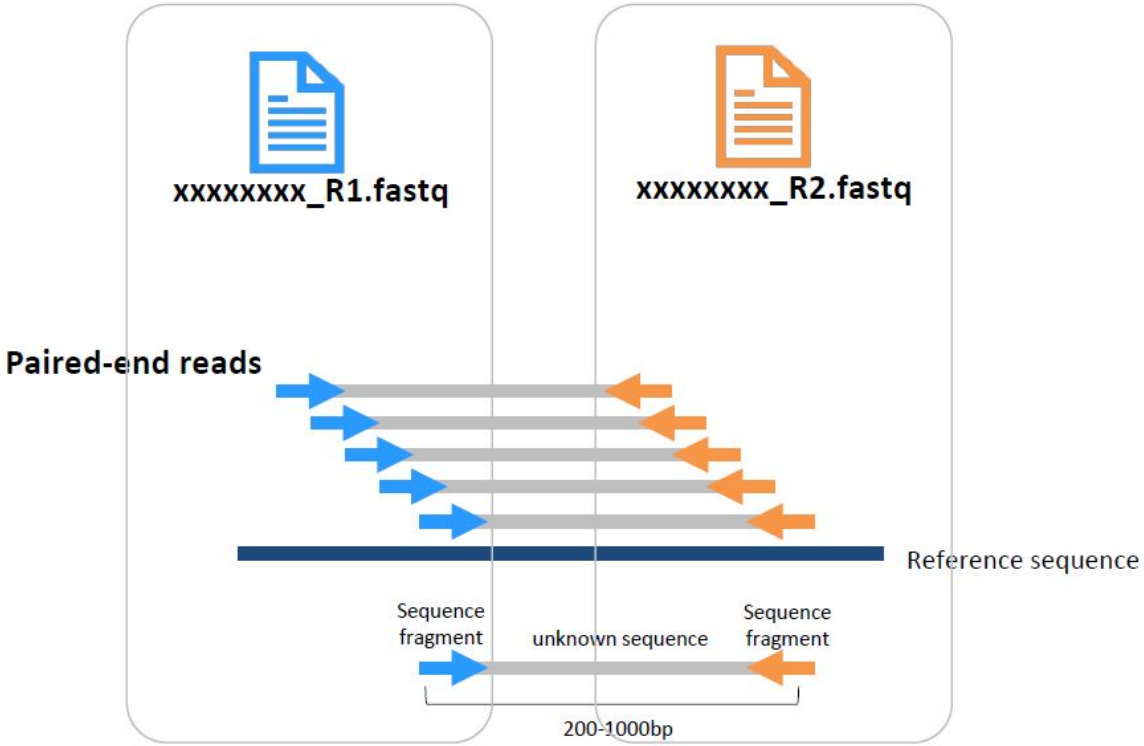
https___zenodo.org_record_3567224_files_sweet-potato-chloroplast-illumina-reduced.fastq: max

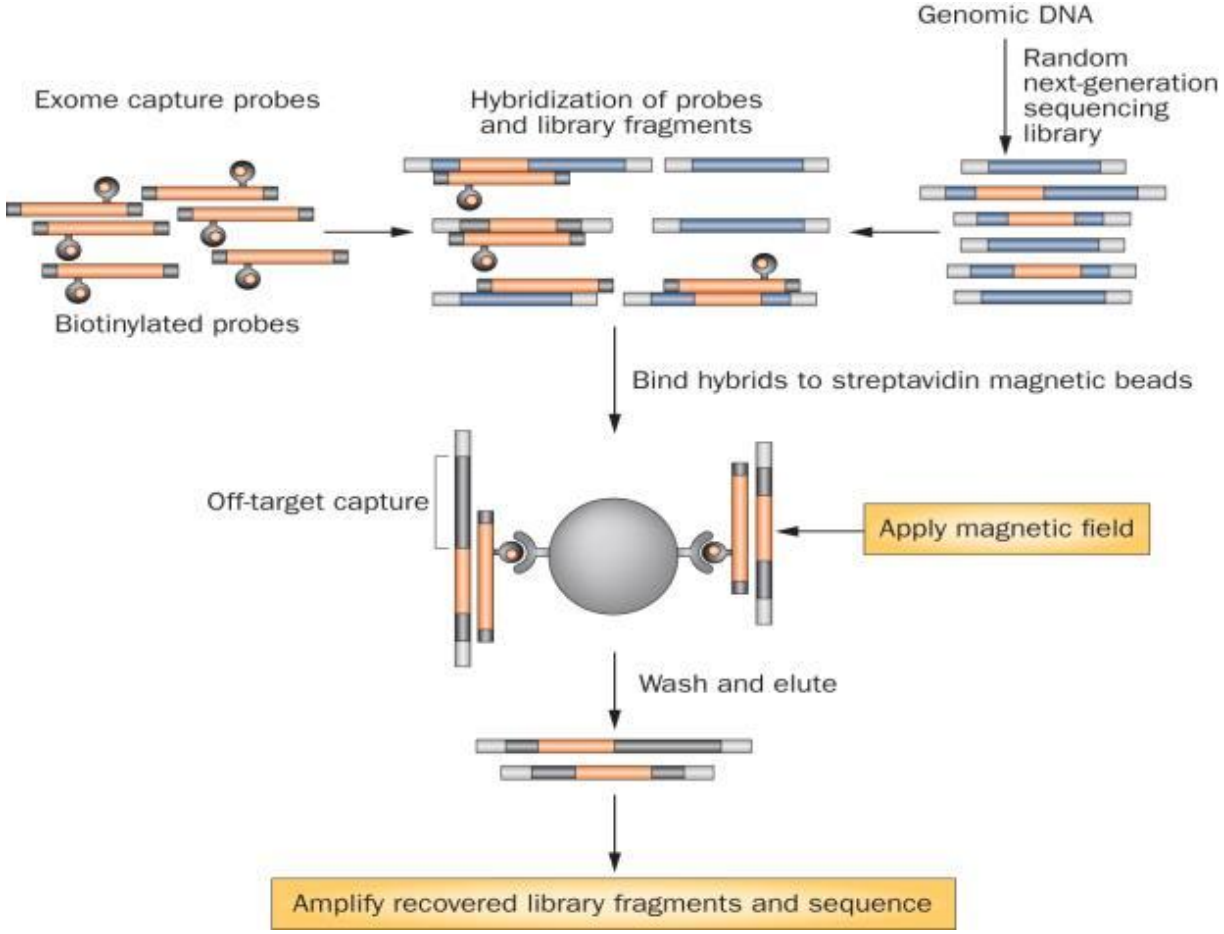
https___zenodo.org_record_3567224_files_sweet-potato-chloroplast-illumina-reduced.fastq: mean

https___zenodo.org_record_3567224_files_sweet-potato-chloroplast-illumina-reduced.fastq: min

Mapping - to a reference genome

- Mapping of reads to a reference genome
  - https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/mapping/tutorial.html#map-reads-on-a-reference-genome





Mardis, E. R. (2012) Applying next-generation sequencing to pancreatic cancer treatment *Nat. Rev. Gastroenterol. Hepatol.* doi:10.1038/nrgastro.2012.126

Reference genome

Genome Reference Consortium: … a consensus representation of the genome.

FASTA format

The human reference sequence GRCh37 (hg19) contains the mitochondrial genome, 22 autosomes, chrX, chrY, 9 haplotype chromosomes, 39 unplaced contigs, and 20 unlocalized contigs.

Genome assemblies can be big. GRCh38.p10 has 3,080,585,178 non-N bases.

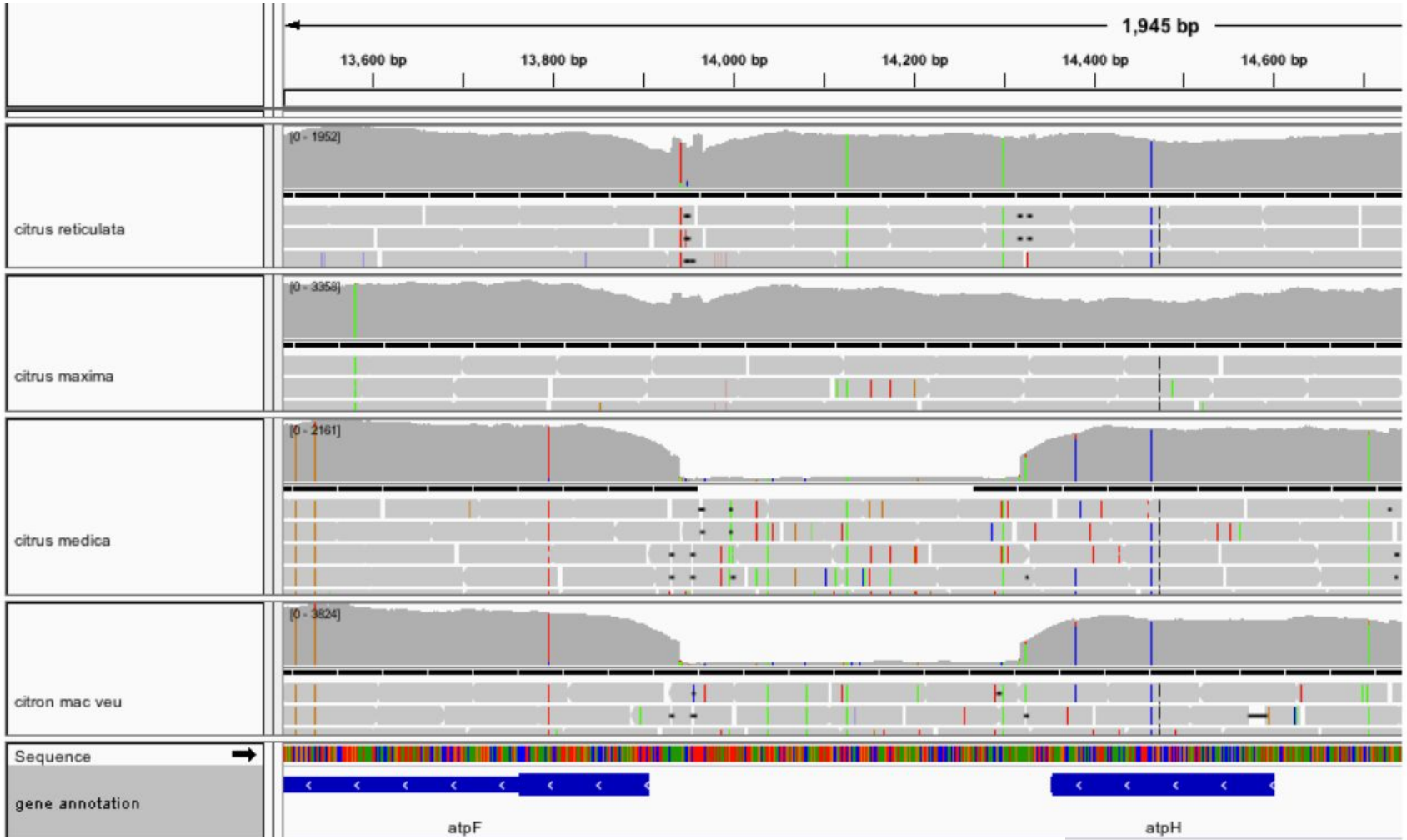
Genomes may have many assembly versions (releases, build): mm9, mm10.

Use the same assembly version for the reference sequence and gene annotations.

Order of sequences / contigs might be important for some tools.

“chr1” and “1” are not identical for some tools.

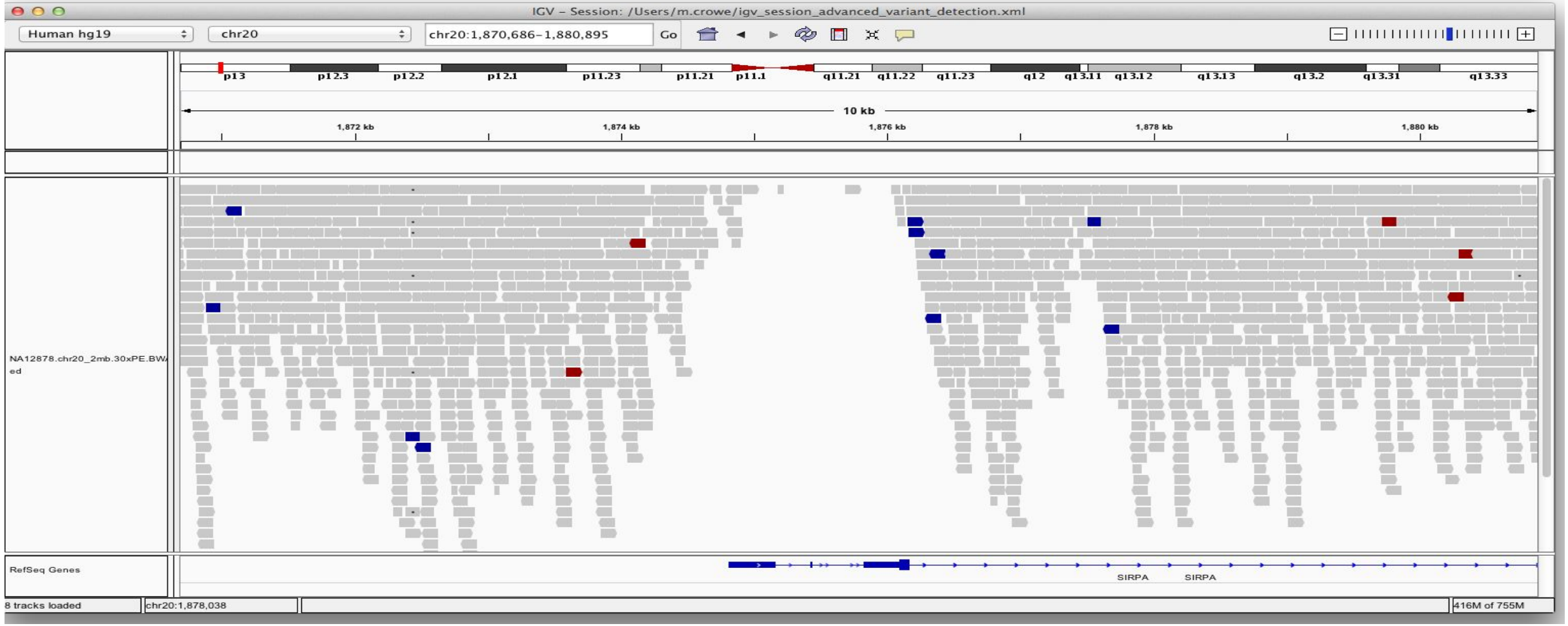
Coverage



A Phylogenetic Analysis of 34 Chloroplast Genomes Elucidates the Relationships between Wild and Domestic Species within the Genus Citrus

DOI: 10.1093/molbev/msv082

Low Coverage

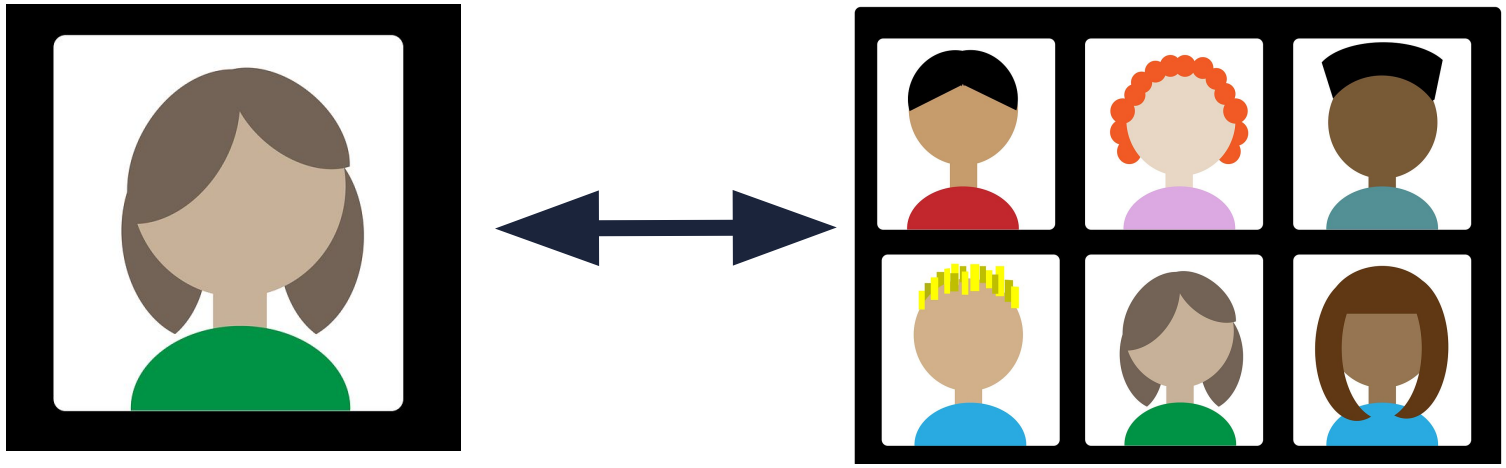


Depth of coverage

The challenges of variant calling

- Poor choice of source material
- Repetitive sequence
- Sequencing errors
- Uneven coverage
- Heterozygosity
- All these and/or combinations of these result in gaps

Doing - Post Processing



Say hello!

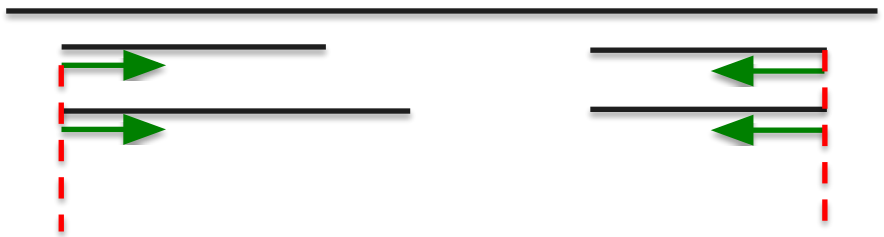
Turn on your camera and microphone and introduce yourself

De-duplication with Picard MarkDuplicates

Removal of (PCR) duplicates

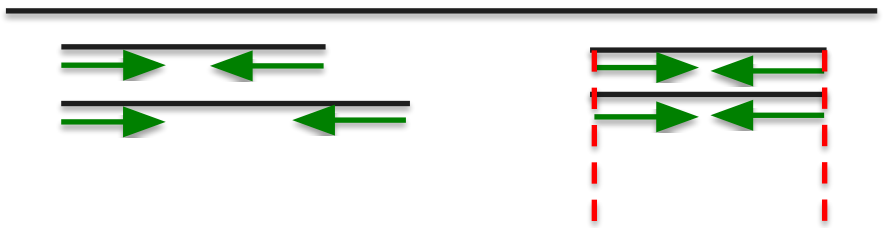
Reference sequence

Single end sequencing



Both reads pairs are duplicates

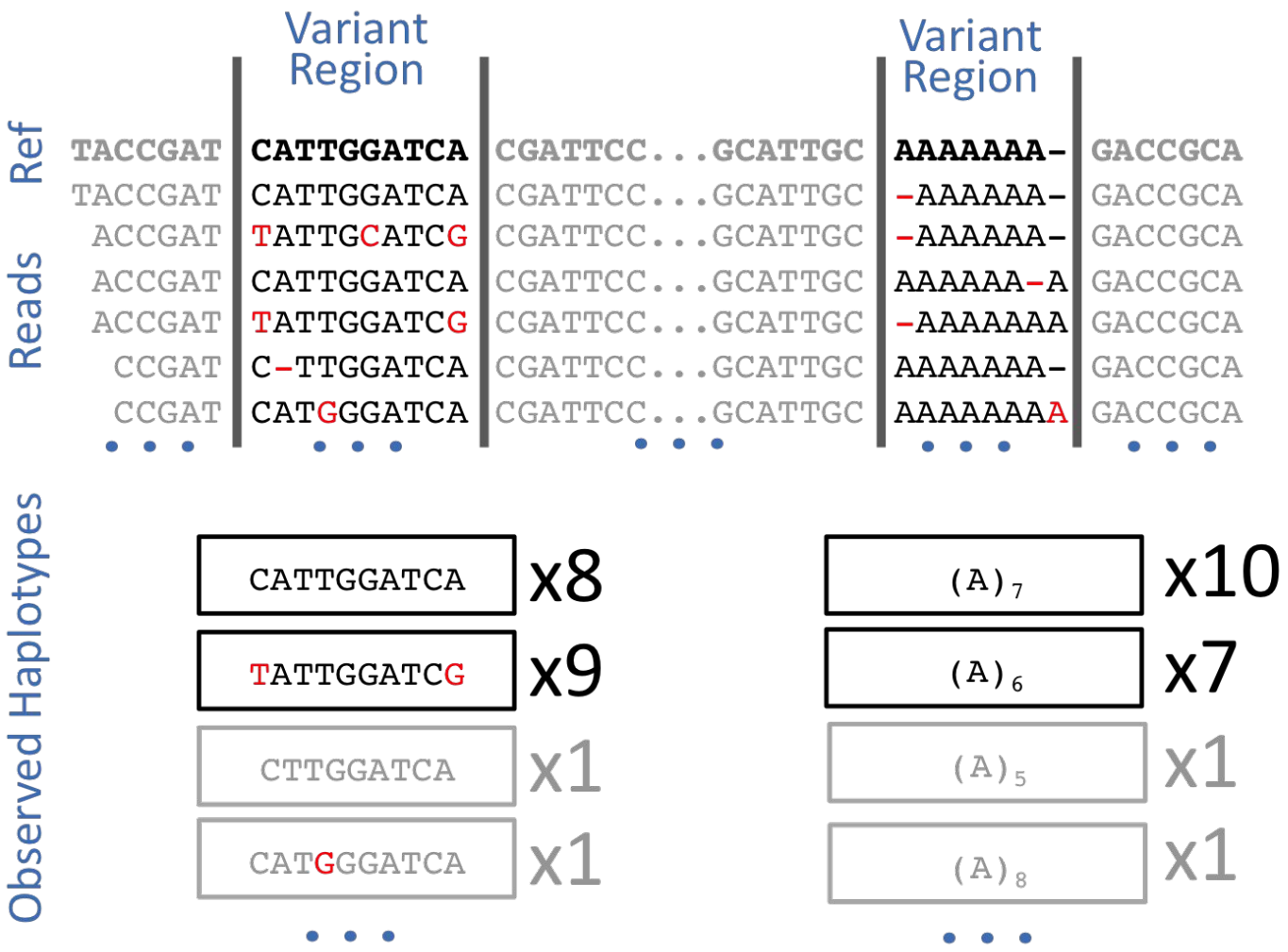
Paired-end sequencing



Duplicates

Variant Calling - FreeBayes

- **Simple diploid calling**:
  - BAM input only - no other parameters set.
- **Simple diploid calling with filtering and coverage**:
  - --min-mapping-quality 30 --min-base-quality 20 --min-supporting-allele-qsum 0 --genotype-variant-threshold 0
  - --min-coverage
- **Frequency-based pooled calling**:
  - --haplotype-length 0 --min-alternate-count 1 --min-alternate-fraction 0 --pooled-continuous --report-monomorphic
  - Best for calling variants in mixtures such as viral, bacterial, or organellar genomes
- **Frequency-based pooled calling with filtering and coverage**:
  - --min-mapping-quality 30 --min-base-quality 20 --min-supporting-allele-qsum 0 --genotype-variant-threshold 0
  - --min-coverage



https://github.com/freebayes/freebayes/blob/master/README.md

Variant Calling - Indel normalisation

Reference and alternative alleles of a CA short tandem repeat (STR)

```
REF    GGGCACACACAGGG
ALT    GGGCACACAGGG
```

CA deletion from the reference

Genome Reference

```
        GGGCACACACAGGG
REF             CA
ALT             .
REF          CAC
ALT          C
REF       GCACA
ALT       GCA
REF     GGCA
ALT     GG
REF       GCA
ALT       G
```

Alleles represented against the human genome reference. Allele pairs are colored the same, all are representations of the same variant.

Variant Call Format

POS	REF	ALT	
8	CA	.	Not left aligned and alternate allele is empty
6	CAC	C	Not left aligned but parsimonious
3	GCACA	GCA	Not right trimmed
2	GGCA	GG	Not left trimmed
3	GCA	G	Normalized (left aligned & parsimonious)

Alleles represented in Variant Call Format, all are representations of the same variant.

Variant Calling - Preparation: SNPeff

SnpEff download: download a pre-built database (Galaxy Version 4.3+T.galaxy2)

Favorite | Versions | Options

Select the annotation database you want to download (e.g. GRCh38.86, mm10 etc.)

The list of available databases can be obtained with 'SnpEff databases' tool

Email notification

No

Send an email notification when the job completes.

Execute

hg19

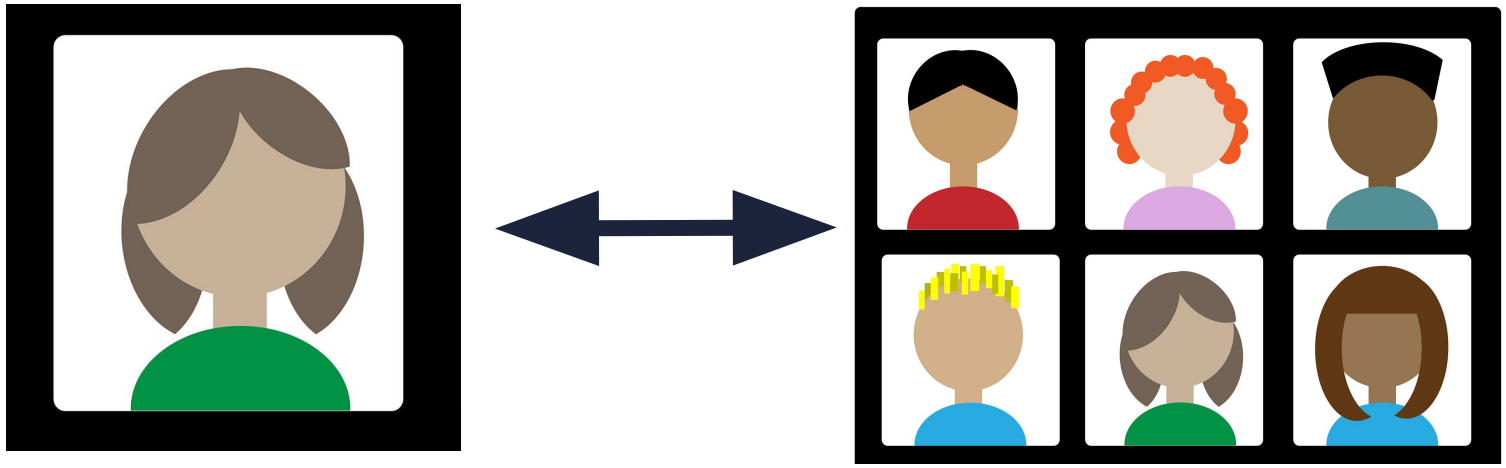
Variant Calling - Preparation: Pedigree

#family_id	name	paternal_id	maternal_id	sex	phenotype
FAM	father	0	0	1	1
FAM	mother	0	0	2	1
FAM	proband	father	mother	1	2

- **family_id** is an alphanumeric identifier of a family
- **name** is the identifier of the sample described by the line
- **paternal_id** is the identifier of the sample's father
- **maternal_id** is the identifier of the sample's mother
- **sex** is a numeric code for the sample's sex (1=male, 2=female, any other number=unknown sex)
- **phenotype** is a numeric code for the sample's phenotypic affection status (1=unaffected, 2=affected)

If the sample's status is unknown, a placeholder of 0 or -9 can be used to indicate this fact.

Doing - Variant Calling

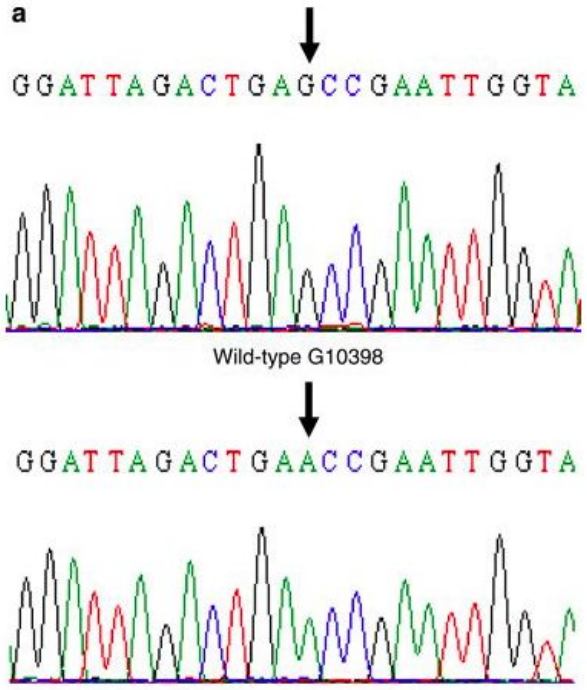


Say hello!

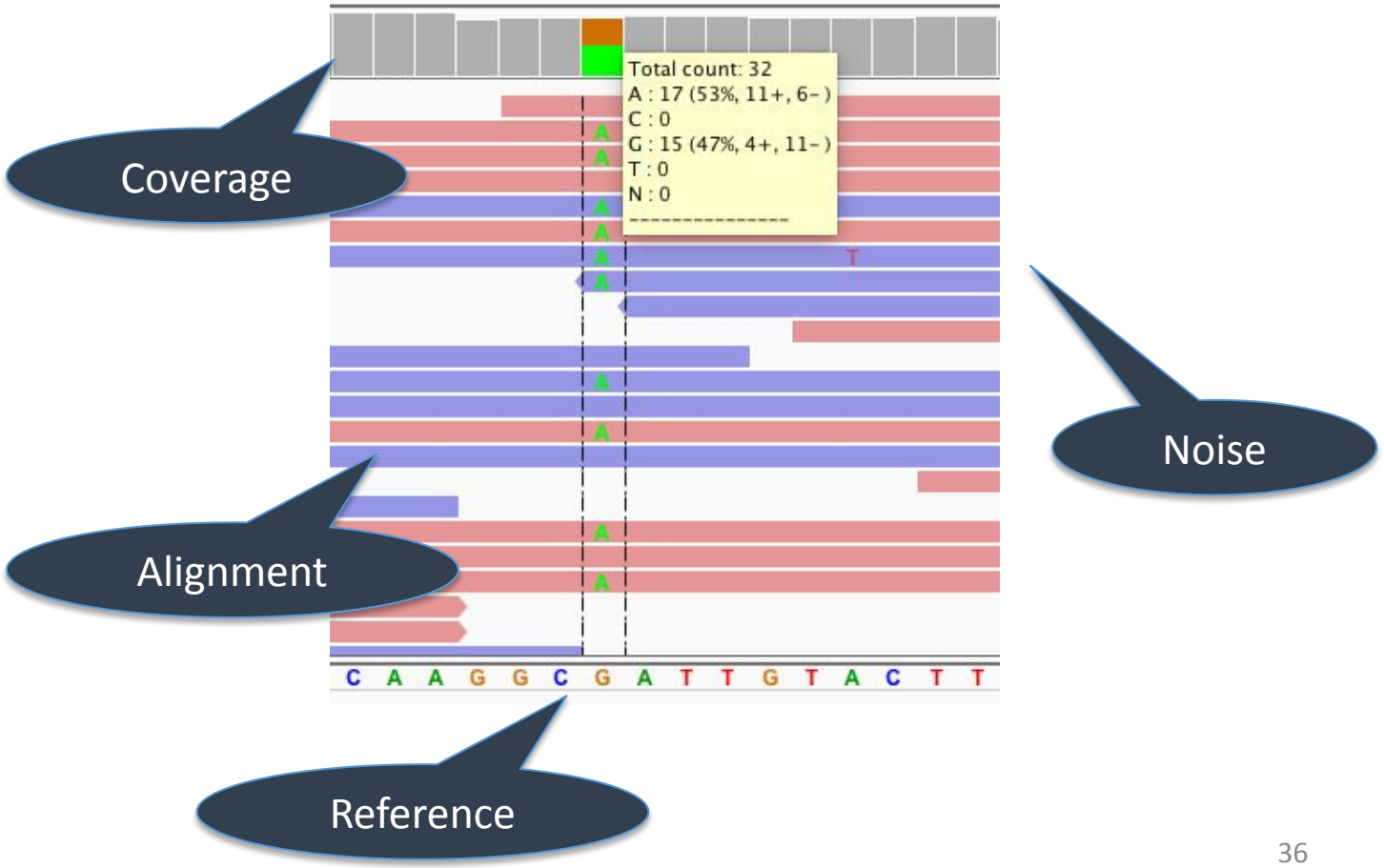
Turn on your camera and microphone and introduce yourself

Variant Detection

Sanger sequencing
(capillary sequencing)



High throughput sequencing



The chromatograms are from:
Kulawiec et al. Journal of Human Genetics 54, 647–654 (2009)
https://www.nature.com/articles/jhg200989

Variant detection pipeline - generic

BWA

bowtie2

GATK

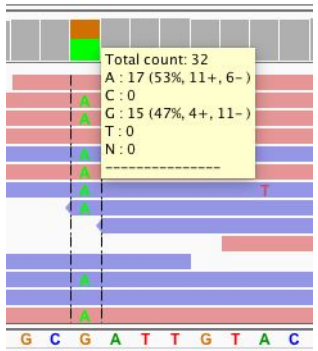
MPileup

FreeBayes

ANNOVAR

SnpEff

IGV



reads

Align reads to a reference genome

alignment

Variant calling

variants

Annotation, interpretation

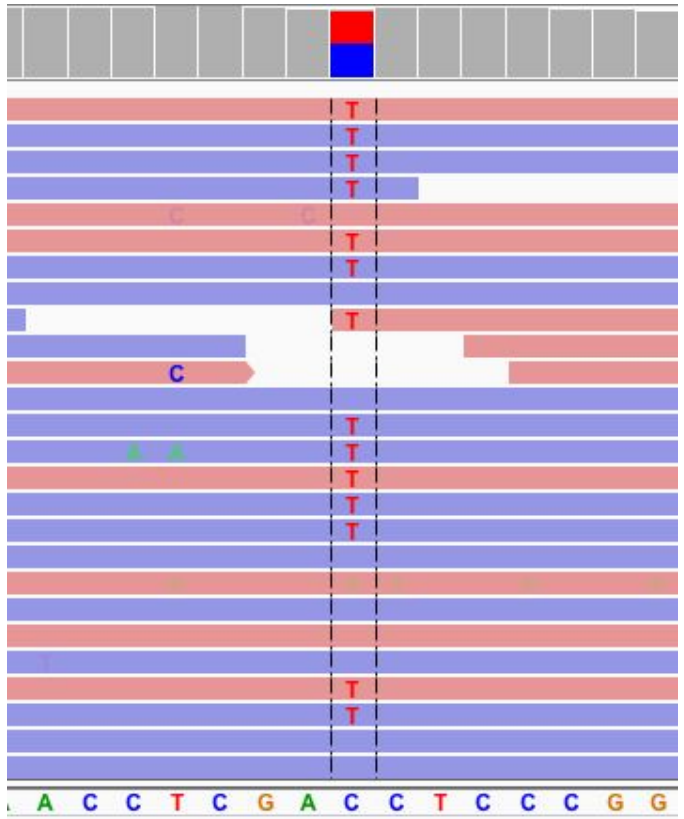
QC, read trimming

Indel realignment, quality score recalibration

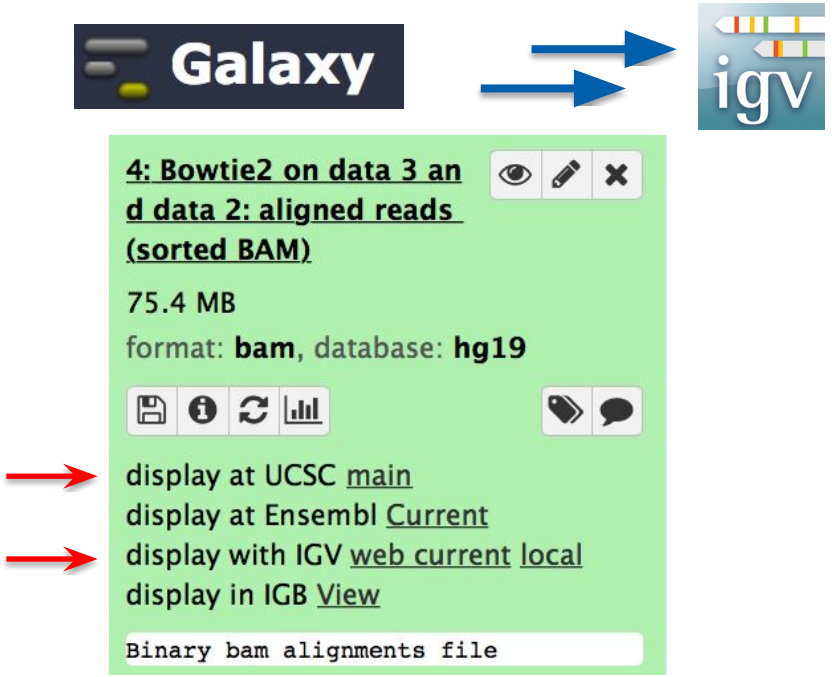
Variant filtering and validation

Visualisation of alignments

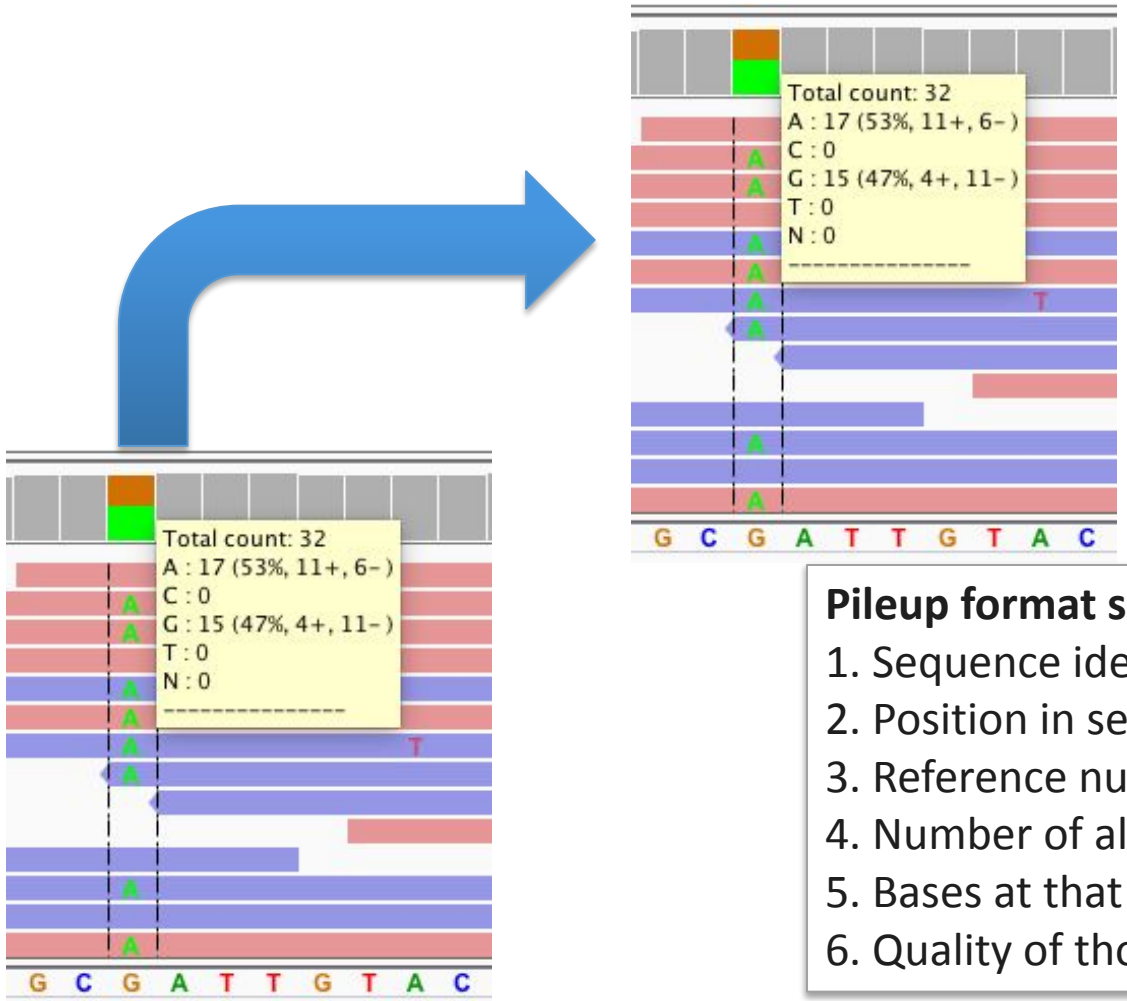
BAM files can be visualised on genome browsers, such as *IGV*.
Galaxy can visualise alignments with build-in browser, *Trackster*
Visualisation of multiple tracks: BAMs, gene annotations, variants…



Galaxy can act as a track hub



Transform alignment into consensus



Write nucleotide composition for every position in alignment, keeping as much information as possible

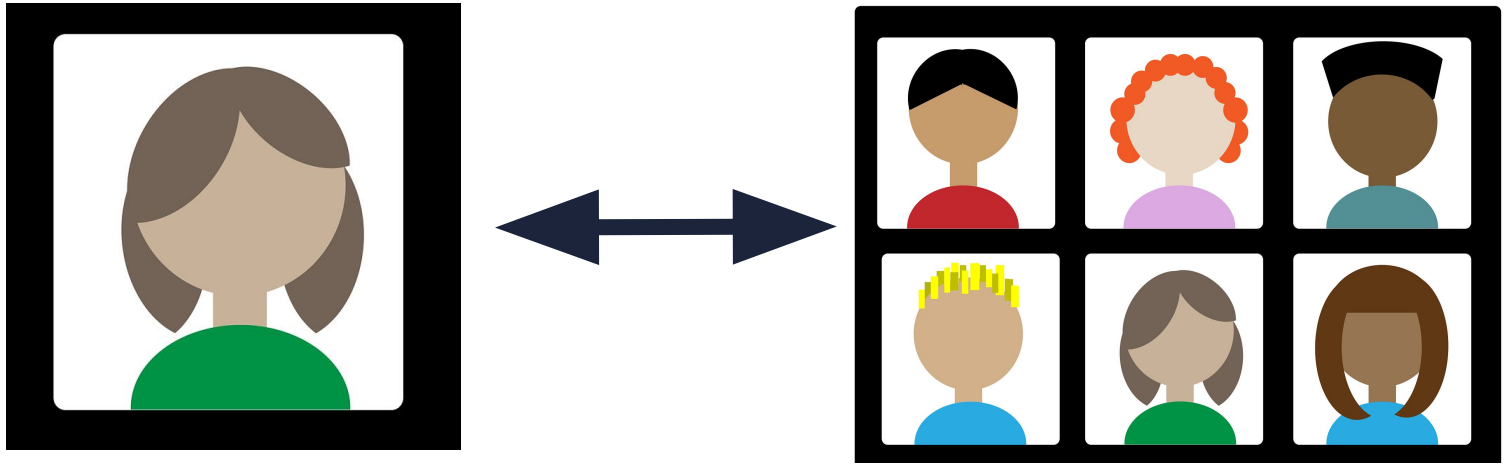
Pileup format specs:

1. Sequence identifier
2. Position in sequence (starting from 1)
3. Reference nucleotide at that position
4. Number of aligned reads covering that position
5. Bases at that position from aligned reads
6. Quality of those bases (OPTIONAL)

```
seq1 272 T 24  ,.$.....,,,,,,,,,,,,,,,.^+. <<<+;<<<<<<<<<<<=<;<;7<&
seq1 273 T 23  ,.....,,,,,,,,,,,,,,,.A <<<;<<<<<<<<<3<=<<<;<<+
```

https://en.wikipedia.org/wiki/Pileup_format

Doing - Variant Annotation



Say hello!

Turn on your camera and microphone and introduce yourself

Gemini scoring

- **GERP scores - Genomic Evolutionary Rate Profiling**
  a. http://mendel.stanford.edu/SidowLab/downloads/gerp/
  b. GERP identifies constrained elements in multiple alignments by quantifying substitution deficits. These deficits represent substitutions that would have occurred if the element were neutral DNA, but did not occur because the element has been under functional constraint. Rejected substitutions are a natural measure of constraint that reflects the strength of past purifying selection on the element.
  c. Positive scores represent highly-conserved positions while negative scores represent highly-variable positions.

- **CADD scores - Combined Annotation Dependent Depletion**
  a. https://cadd.gs.washington.edu/
  b. CADD tool scores the predicted deleteriousness of single nucleotide variants and insertion/deletions variants in the human genome by integrating multiple annotations including conservation and functional information into one metric.
  c. CADD provides a ranking rather than a prediction or default cut-off, with higher scores more likely to be deleterious.
  d. Scores >30 as 'likely deleterious', <30 as 'likely benign'. Variants with scores over 30 are predicted to be the 0.1% most deleterious possible substitutions in the human genome.

Gemini scoring

- **ClinVar**
  a. https://www.ncbi.nlm.nih.gov/clinvar/
  b. ClinVar is a freely accessible, public archive of reports of the relationships among human variations and phenotypes, with supporting evidence.
  c. ClinVar thus facilitates access to and communication about the relationships asserted between human variation and observed health status, and the history of that interpretation.
  d. The level of confidence in the accuracy of variation calls and assertions of clinical significance depends in large part on the supporting evidence, so this information, when available, is collected and visible to users.
  e. Domain experts are encouraged to apply for recognition as an expert panel.

NM_000314.7(PTEN):c.1A>G (p.Met1Val)

Interpretation:	**Pathogenic**
Review status:	★★★☆ reviewed by expert panel FDA RECOGNIZED DATABASE
Submissions:	3 (Most recent: Jan 7, 2021)
Last evaluated:	Nov 22, 2019
Accession:	VCV000484600.6
Variation ID:	484600
Description:	single nucleotide variant

Gemini scoring

- Autosomal recessive
- Autosomal dominant
- X-linked recessive
- X-linked dominant
- Autosomal de-novo
- X-linked de-novo
- Compound heterozygous
- Loss of heterozygosity (LOH) events

Track sequence from bottom up			1		2	3		4	5
Father	Mother	Child	iUPD-P	hUPD-P	BPD	hUPD-M	iUPD-M	MI-S	MS-D
AA	AA	AA	X	X	X	X	X		
		AB						X	
		BB							X
	AB	AA	X	X	X		X		
		AB			X	X			
		BB					X		
	BB	AA	X	X					
		AB			X				
		BB				X	X		
AB	AA	AA	X		X	X	X		
		AB		X	X				
		BB	X						
	AB	AA	X		X		X		
		AB		X	X	X			
		BB	X		X		X		
	BB	AA	X						
		AB		X	X				
		BB	X		X	X	X		
BB	AA	AA				X	X		
		AB			X				
		BB	X	X					
	AB	AA					X		
		AB			X	X			
		BB	X	X	X		X		
	BB	AA							X
		AB						X	
		BB	X	X	X	X	X		

http://pevsnerlab.kennedykrieger.org/php/?q=node/85

Systematic Variant Curation

Genet Med. 2015 May ; 17(5): 405–424. doi:10.1038/gim.2015.30.

Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology

	Benign		Pathogenic			
	Strong	Supporting	Supporting	Moderate	Strong	Very Strong
Population Data	MAF is too high for disorder *BA1/BS1* **OR** observation in controls inconsistent with disease penetrance *BS2*			Absent in population databases *PM2*	Prevalence in affecteds statistically increased over controls *PS4*	
Computational And Predictive Data		Multiple lines of computational evidence suggest no impact on gene /gene product *BP4* Missense in gene where only truncating cause disease *BP1* Silent variant with non predicted splice impact *BP7*	Multiple lines of computational evidence support a deleterious effect on the gene /gene product *PP3*	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before *PM5* Protein length changing variant *PM4*	Same amino acid change as an established pathogenic variant *PS1*	Predicted null variant in a gene where LOF is a known mechanism of disease *PVS1*
Functional Data	Well-established functional studies show no deleterious effect *BS3*		Missense in gene with low rate of benign missense variants and path. missenses common *PP2*	Mutational hot spot or well-studied functional domain without benign variation *PM1*	Well-established functional studies show a deleterious effect *PS3*	
Segregation Data	Non-segregation with disease *BS4*		Co-segregation with disease in multiple affected family members *PP1*	Increased segregation data →		
De novo Data				*De novo* (without paternity & maternity confirmed) *PM6*	*De novo* (paternity & maternity confirmed) *PS2*	
Allelic Data		Observed in *trans* with a dominant variant *BP2* Observed in *cis* with a pathogenic variant *BP2*		For recessive disorders, detected in *trans* with a pathogenic variant *PM3*		
Other Database		Reputable source w/out shared data = benign *BP6*	Reputable source = pathogenic *PP5*			
Other Data		Found in case with an alternate cause *BP5*	Patient's phenotype or FH highly specific for gene *PP4*			

Summary

- Exome sequencing is an efficient way to identify disease-relevant genetic variants.
- Freebayes is a good variant and genotype caller for the joint analysis of multiple samples. It is straightforward to use and requires only minimal processing of mapped reads.
- Variant annotation and being able to exploit genotype information across related individuals is key to identifying candidate disease variants. SnpEff and GEMINI, in particular, are powerful tools offered by Galaxy for that purpose.

- Key to confident variant calls
  - High quality reads
  - Appropriate minimum depth threshold
  - And not looking below that threshold!

Acknowledgements

Simon Gladman, Igor Makunin, Nicholas Rhodes, Catherine Bromhead, Nuwan Goonasekera, Michael Thang, Gareth Price