Variant Analysis with Galaxy

Galaxy Workshop Glasgow, UK

09/06/2015

Mani Mudaliar

Glasgow Polyomics

Manikhandan.Mudaliar@glasgow.ac.uk

University of Glasgow

Glasgow Polyomics
www.glasgow.ac.uk/polyomics

Outline

- Introduction
- File formats and conventions
- Databases used in variant analysis
- Variant analysis: Options
- Benchmarking and validation
- Variant analysis: A worked example

Objectives

By the end of this session, you will

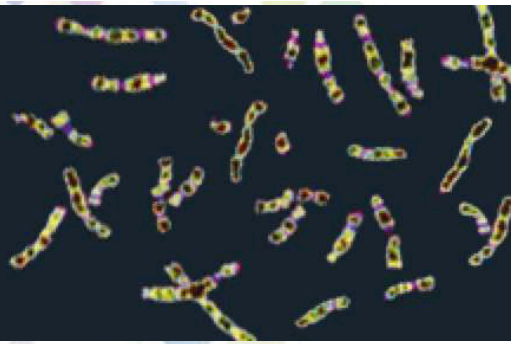
1. Know the tools and workflows in variant calling
2. Understand the file formats
3. Perform variant calling
4. Perform functional annotation of the variants
5. Visualize the variants in a genome browser
6. Understand the importance of benchmarking and validation of workflows

- **Introduction**
- File formats and conventions
- Databases used in variant analysis
- Variant analysis: Options
- Benchmarking and validation
- Variant analysis: A worked example

Major Genome Projects - 1000 Genomes Project

1000 Genomes

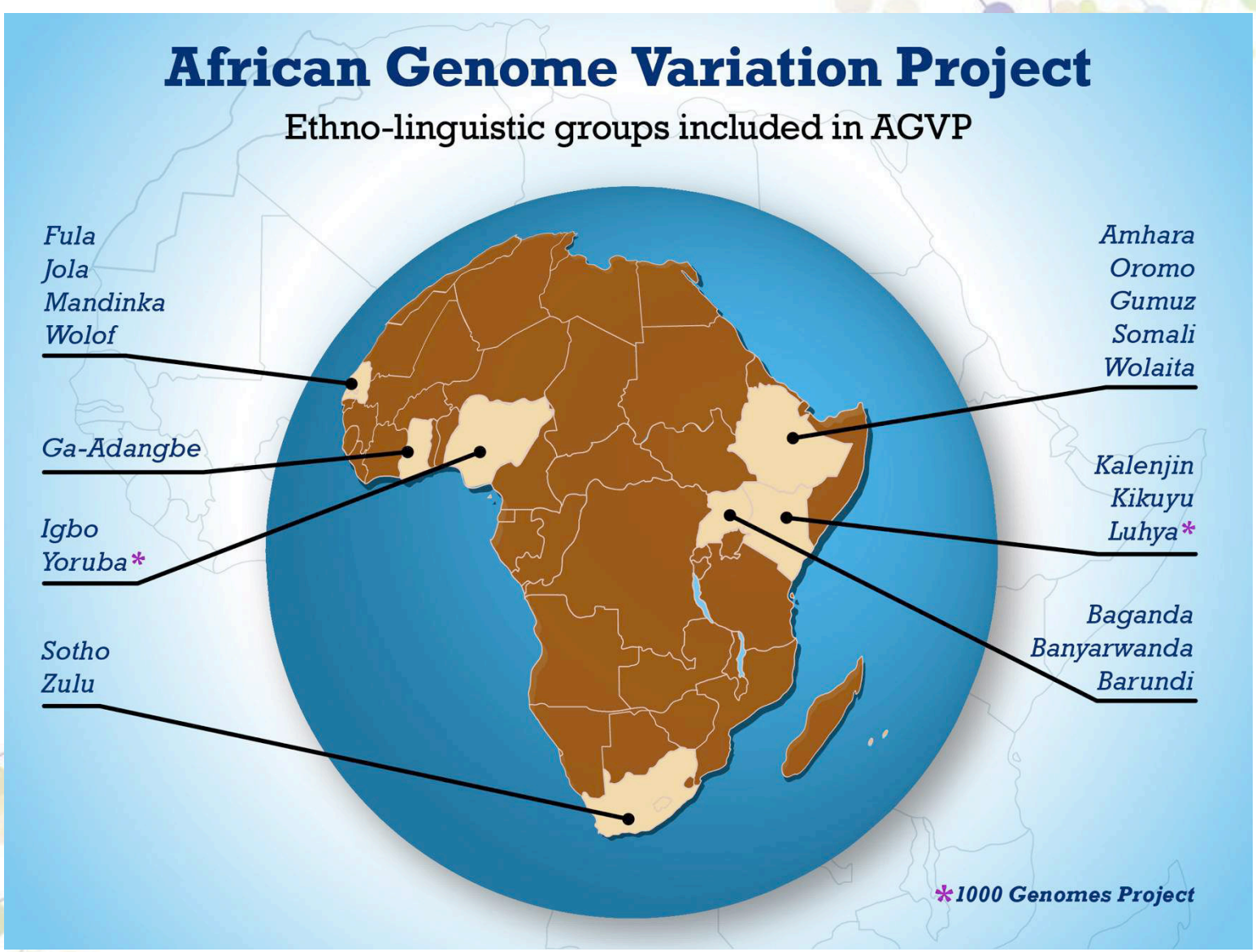
A Deep Catalog of Human Genetic Variation



- The first project to sequence the genomes of a large number of people, to provide a comprehensive resource on human genetic variation
- Launched in **2007**
- Genomes of about 2500 unidentified people from about 25 populations around the world at 4X coverage
- Discovery of SNP, variants at low frequencies (0.1-0.5%), and structural variants.

http://www.1000genomes.org

African Genome Variation Project



doi:10.1038/nature13997

University of Glasgow

Glasgow Polyomics
www.glasgow.ac.uk/polyomics

Major Genome Projects - UK10K

UK10K

Rare Genetic Variants in Health and Disease (2010-2013)

UK10K Study Samples

Follow the links below for more information about the UK10K Study Samples:

- **Whole genome cohorts (4000)**
- **Neurodevelopment Sample Sets (up to 3000 whole exomes)**
- **Obesity Sample Sets (2000 whole exomes)**
- **Rare Diseases Sample Sets (1000 whole exomes)**

➢ To understand the link between low-frequency and rare genetic changes, and human disease.

http://www.uk10k.org

The 100,000 Genomes Project

Genomics England

Home | About us | 100,000 Genomes Project | GeCIP | GENE Consortium | Library & resources | News | Contact us

Genomics England, with the consent of participants and the support of the public, is creating a lasting legacy for patients, the NHS and the UK economy through the sequencing of 100,000 genomes: the 100,000 Genomes Project.

Genomics England was set up by the Department of Health to deliver the 100,000 Genomes Project. Initially the focus will be on rare disease, cancer and infectious disease.

Read more…

http://www.genomicsengland.co.uk

University of Glasgow

Glasgow Polyomics
www.glasgow.ac.uk/polyomics

Saudi Human Genome Project

برنامج الجينوم البشري السعودي
Saudi Human Genome Program

a national program for sequencing the genome ...

مدينة الملك عبد العزيز للعلوم والتقنية KACST

Arabic

- 3 year project to find genes responsible for genetic diseases
- Launched in **2013**
- **Aim:** To eliminate the recessive genetic diseases from the population in 5 years, through a process of discover, screening and pre-marital counselling

http://rc.kfshrc.edu.sa/sgp/Index.asp

University of Glasgow

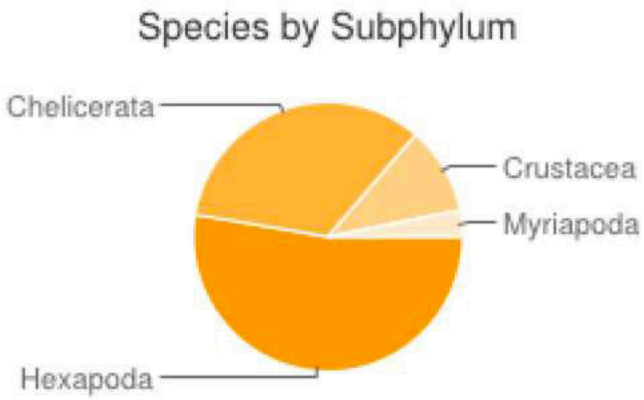
Glasgow Polyomics
www.glasgow.ac.uk/polyomics

Major Genome Projects - Genome 10K Project

- Launched in **April 2009** at the University of California, Santa Cruz.
- To assemble a **genomic zoo**
- Target: 10,000 vertebrate species; Achieved: more than **16,000** vertebrate species

https://genome10k.soe.ucsc.edu

Major Genome Projects - i5k



Total no. of species: 809

Hexapoda	702
Chelicerata	64
Crustacea	20
Myriapoda	6

i5k Genome Sequencing Initiative for Insects and Other Arthropods

The *i5k* initiative is a transformative project that aims to sequence and analyze the genomes of 5,000 arthropod species. Species selection is driven by our common goal to better understand arthropod evolution and phylogeny through studies of species known to be important to worldwide agriculture, food safety, medicine, energy production, models in biology, those species most abundant in world ecosystems, and representatives in every branch of insect phylogeny. Our initiative is broad and inclusive. We intend to involve scientists from around the world to strengthen our combined research and form partnerships to seek funding from academia, governments, industry, and private sources.

- i5k Genome Sequencing Initiative for Insects and Other Arthropods
- Launched in **2011**
- Sequence and analyse the genomes of 5,000 arthropod species

http://www.arthropodgenomes.org/wiki/i5K

Major Transcriptome Projects - Fish-T1K

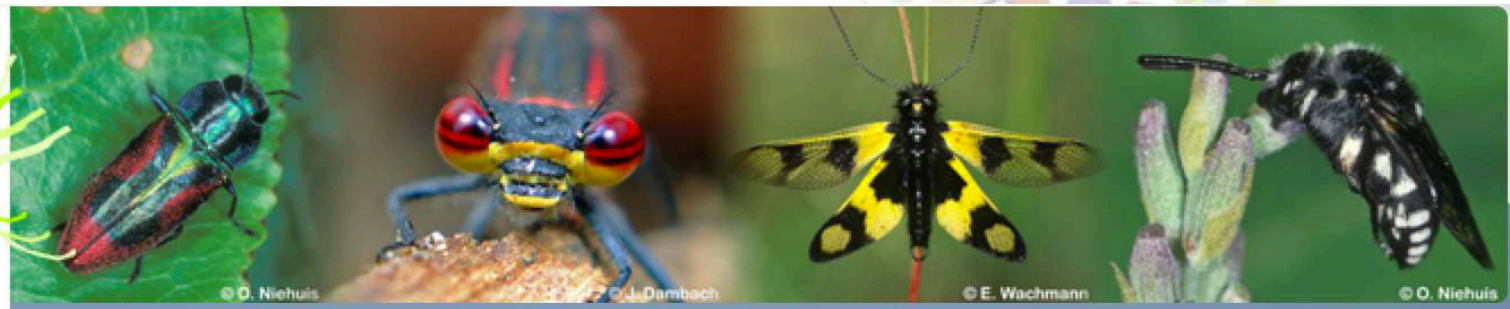
- Fish-T1K: Transcriptomes of 1,000 Fishes
- Launched in **November 2013**
- BGI, Marine Genomics institute (Shenzhen, China)
- Phylogenetic tree of all fishes
- Adaptations
- Evolution of sex-determining systems
- Evolution of the immune system

http://www.fisht1k.org

University of Glasgow

Glasgow Polyomics
www.glasgow.ac.uk/polyomics

Major Transcriptome Projects - 1KITE



- 1K Insect Transcriptome Evolution
- Launched in **2012**
- BGI, Marine Genomics institute (Shenzhen, China)
- Completed for more than 1,200 species

http://www.1kite.org

University of Glasgow

Glasgow Polyomics
www.glasgow.ac.uk/polyomics

Genomics Projects Database

JGI JOINT GENOME INSTITUTE
UNITED STATES DEPARTMENT OF ENERGY

GOLD
Genomes Online Database

JGI/IMG Gold User Log in

Home | Search | Distribution Graphs | Biogeographical Metadata | Statistics | References | Team | Help | News

Studies	20592
Biosamples	61670
Sequencing Projects	61852
Analysis Projects	48582

Download Excel Data file

Welcome to the Genomes OnLine Database

GOLD Release v.5

GOLD:Genomes Online Database, is a World Wide Web resource for comprehensive access to information regarding genome and metagenome sequencing projects, and their associated metadata, around the world.

Studies	Biosamples	Projects	Organisms
• Metagenomic 553	Classification	Complete Projects 6651	• Organisms 58695
• Non-Metagenomic 20025	• Ecosystems	Permanent Drafts 23543	Archaea 1037
	Host-associated 11910	Incomplete Projects 28411	Bacteria 44576
	Engineered 1669	Targeted Projects 1253	Eukarya 8181
	Environmental 6881		

Genomes OnLine Database (GOLD)

Started in 1997; Over 60,000 projects

https://gold.jgi-psf.org

- The African Genome Variation Project
- The ENCODE Project: ENCyclopedia Of DNA Elements
- Genomics of inflammation and immunity – WT

University of Glasgow

Glasgow Polyomics
www.glasgow.ac.uk/polyomics

Genomics Projects

National Human Genome Research Institute
Advancing human health through genomics research

SEARCH GENOME.GOV

Research Funding | Research at NHGRI | Health | Education | Issues in Genetics | Newsroom | Careers & Training | About | Español

Home › Research Funding › Research Funding Divisions › Division of Genome Sciences › NHGRI Genome Sequencing Program (GSP) › Large-Scale Genome Sequencing and Analysis Centers (LSAC) › Approved Sequencing Targets

Approved Sequencing Targets

Please note: To sort the table by column, click on the link in the header. To review a list and database of the previous approved sequencing targets, see: www.genome.gov/10002154.

Status Approved Sequencing Targets

Center	Active / Historical	Proposal or Project Name	Sub-Project Name	Common Name (Species Name), Tumor Type, Phenotype, or Disease	Data Type	Name registered in dbGap/ BioProject	Capacity needed for project (Gb)	% sequencing complete	Project Finished?	Human / Non-Human
WASHU	ACTIVE	1000 Genomes	Full Scale		Whole exome	PRJNA28889	7850	100%	Yes	Human
BAYLOR	ACTIVE	1000 Genomes	Full Scale		Whole exome	PRJNA59773	3193	100%	Yes	Human
WASHU	ACTIVE	1000 Genomes	Full Scale		Whole genome	PRJNA28889	5650	100%	Yes	Human
BAYLOR	ACTIVE	1000 Genomes	Full Scale		Whole genome	PRJNA59771	4050	100%	Yes	Human
BROAD	ACTIVE	1000 Genomes	Phase 2		Whole exome			100%	Yes	Human
BROAD	ACTIVE	1000 Genomes	Phase 2		Whole genome			100%	Yes	Human
BROAD	ACTIVE	1000 Genomes	Phase 3		Whole exome			100%	Yes	Human
BROAD	ACTIVE	1000 Genomes	Phase 3		Whole genome			100%	Yes	Human
BROAD	ACTIVE	1000 Genomes	Phase 3 Validation		Whole exome			100%	Yes	Human
BROAD	ACTIVE	1000 Genomes	Phase 3 Validation		Whole genome			100%	Yes	Human
BROAD	ACTIVE	1000 Genomes	Pilot + Phase 1		Whole exome			100%	Yes	Human
BROAD	ACTIVE	1000 Genomes	Pilot + Phase 1		Whole genome			100%	Yes	Human

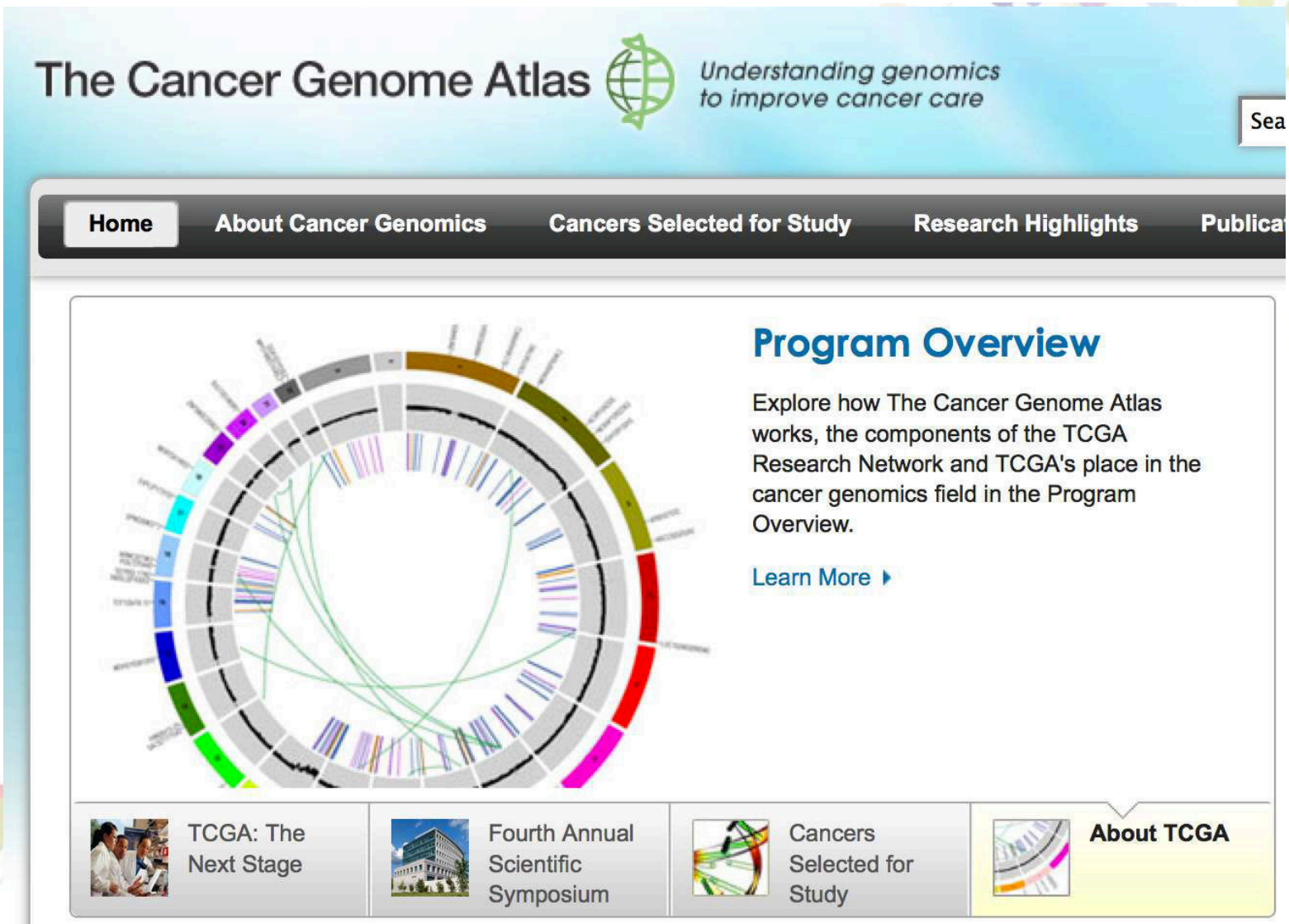
Status of Approved Sequencing Targets

https://www.genome.gov/27557963

University of Glasgow

Glasgow Polyomics
www.glasgow.ac.uk/polyomics

Cancer Genome Projects – TCGA



The Cancer Genome Atlas (TCGA)

http://cancergenome.nih.gov

University of Glasgow

Glasgow Polyomics
www.glasgow.ac.uk/polyomics

Cancer Genome Projects - ICGC

International Cancer Genome Consortium

ICGC Cancer Genome Projects

Committed projects to date: 77

Sort by: Project

ICGC Goal: To obtain a **comprehensive** description of **genomic, transcriptomic and epigenomic changes** in **50 different tumor types and/or subtypes** which are of clinical and societal importance across the globe.

Read more »

https://icgc.org

Cancer genome project

wellcome trust sanger institute

Search for Enter search here...

Home | Research | Scientific resources | Work & study | About us

Areas of research | Projects | Academic faculty | Scientific publications

Cancer genome project

The Wellcome Trust Sanger Institute's Cancer Genome Project is led jointly by Professor Mike Stratton and Dr Peter Campbell. All cancers occur due to abnormalities in DNA sequence. Cancer affects people at all ages with the risk for most types increasing with age.

One in three people in the Western world develop cancer and one in five die of the disease. Cancer is therefore the most common genetic disease.

[Anne Weston, Wellcome Images]

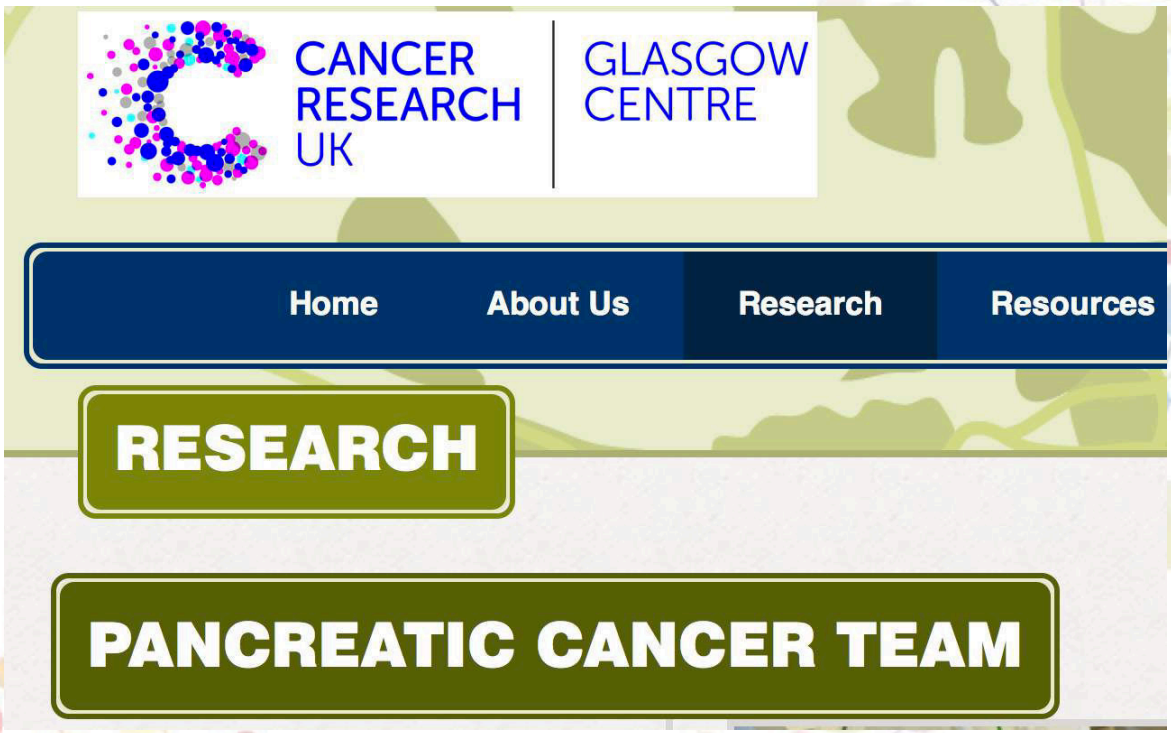
- **Wellcome Trust** initiative
- Part of the **International Cancer Genome Consortium**

https://www.sanger.ac.uk/research/projects/cancergenome/

University of Glasgow

Glasgow Polyomics
www.glasgow.ac.uk/polyomics

Large NGS projects in Glasgow



- **Leading Pancreatic Cancer Centre in the world**

http://www.glasgowcancer.org/Research/Pancreatic-Cancer-Team.html

Large NGS projects in Glasgow

Stratified Medicine Scotland

SMS Making a Reality of Stratified Medicine

Contact Us | Our Partners | Scottish SME Partners | News

Innovation Centres

Supported by The Scottish Funding Council, Highlands and Islands Enterprise and Scottish Enterprise.

Ovarian Cancer | Oesophageal Cancer | Rheumatoid Arthritis | IBD/COPD

Rheumatoid Arthritis (RA) is the most common of the chronic inflammatory arthritic conditions. There are approximately 400,000 new cases in Europe and the US each year. RA has major impacts at three levels - the well being of patients, health care cost and societal costs. Many patients experience significant pain, disability and premature mortality. Within Europe the direct cost in terms of managing RA is in the region of £11.6Bn pa whilst the indirect cost of managing long term social security costs are an additional £14.1Bn pa . Current RA drug therapy is

http://www.stratmed.co.uk/

http://www.race-gbn.org/

http://www.glazgodiscoverycentre.co.uk/

I am working in two stratified medicine biomarker discovery projects

University of Glasgow

Glasgow Polyomics
www.glasgow.ac.uk/polyomics

Think!

Our journey from 2001

One human genome to 100,000 genomes

What led to this fast pace improvement?

NGS (Technological improvements)

Bioinformatics (Efficient tools)

Where are we heading to ??????

Polyomics, Clinical and omics data integration, Stratified medicine, Systems biology, Virtual cell/organ/human .....

Genetic Variation

variant
alteration
polymorphism
sequence-variant
mutation
allelic-variant

"a change / changes in the genomic sequence compared with the reference genomic sequence"

E.g., Substitution, Indel, Copy number variation, Translocation, Polyploidy or Aneuploidy

Genetic Variation

Nucleus of a human somatic cell contains 46 chromosomes (23 pairs)

- 22 autosomal pairs + 1 pair of sex chromosomes XX or XY
- One set of chromosomes inherited from each parent
- Mitochondrial circular DNA in cytoplasm from mother

Germline mutation

- Mutation inherited from the parents
- Fertilization (syngamy): Unique mixture resulting from four genetically unique haploid strands of the maternal and paternal chromosomes
- Independent assortment, genetic linkage and linkage disequilibrium (Meiosis)

Somatic mutation

- Not inherited from parents
- Acquired from spontaneous mutations during DNA replication (Mitosis)
- Frequent in tissues with high cell turnover (e.g., intestinal villi)
- Results in cancer

"Variation" or "Polymorphism" – Nomenclature for the description of sequence variants

- **Polymorphism**
  - A change found at a frequency of 1% or higher in the population
  - Generally a non disease-causing change
  - Single Nucleotide Polymorphism (SNP) and Copy Number Polymorphism (CNP)
  - Pathogenic variant, affects function, variants of unknown significance (VUS)
- **Human Genome Variation Society (HGVS)** http://www.hgvs.org/mutnomen/

General recommendations of the HGVS

- All variants should be described at the most basic level, i.e. the DNA level
- Descriptions should always be in relation to a reference sequence
- Describing genes / proteins, only official HGNC gene symbols should be used
- Should be preceded by a letter indicating the type of reference sequence used:
  - ‘c.’ for a coding DNA sequence (e.g., c.76A>T)
  - ‘g.’ for a genomic sequence (e.g., g.476A>T)
  - ‘m.’ for a mitochondrial sequence (e.g., m.8993T>C)
  - ‘n.’ for a non-coding RNA sequence

http://www.hgvs.org/mutnomen/recs.html

Types of Genome Sequence Variants

- Single Nucleotide Variant (**SNV**) or Single Nucleotide Polymorphism (**SNP**)

Individual 1

Chr 2 ...CGATATTCC**T**ATCGAATGTC...
copy1 ...GCTATAAGG**A**TAGCTTACAG...

Chr 2 ...CGATATTCC**C**ATCGAATGTC...
copy2 ...GCTATAAGG**G**TAGCTTACAG...

Individual 2

Chr 2 ...CGATATTCC**C**ATCGAATGTC...
copy1 ...GCTATAAGG**G**TAGCTTACAG...

Chr 2 ...CGATATTCC**C**ATCGAATGTC...
copy2 ...GCTATAAGG**G**TAGCTTACAG...

Individual 3

Chr 2 ...CGATATTCC**T**ATCGAATGTC...
copy1 ...GCTATAAGG**A**TAGCTTACAG...

Chr 2 ...CGATATTCC**T**ATCGAATGTC...
copy2 ...GCTATAAGG**A**TAGCTTACAG...

Individual 4

Chr 2 ...CGATATTCC**T**ATCGAATGTC...
copy1 ...GCTATAAGG**A**TAGCTTACAG...

Chr 2 ...CGATATTCC**C**ATCGAATGTC...
copy2 ...GCTATAAGG**G**TAGCTTACAG...

Individual 5

Chr 2 ...CGATATTCC**C**ATCGAATGTC...
copy1 ...GCTATAAGG**G**TAGCTTACAG...

Chr 2 ...CGATATTCC**T**ATCGAATGTC...
copy2 ...GCTATAAGG**A**TAGCTTACAG...

Individual 6

Chr 2 ...CGATATTCC**C**ATCGAATGTC...
copy1 ...GCTATAAGG**G**TAGCTTACAG...

Chr 2 ...CGATATTCC**T**ATCGAATGTC...
copy2 ...GCTATAAGG**A**TAGCTTACAG...

Types of Genome Sequence Variants

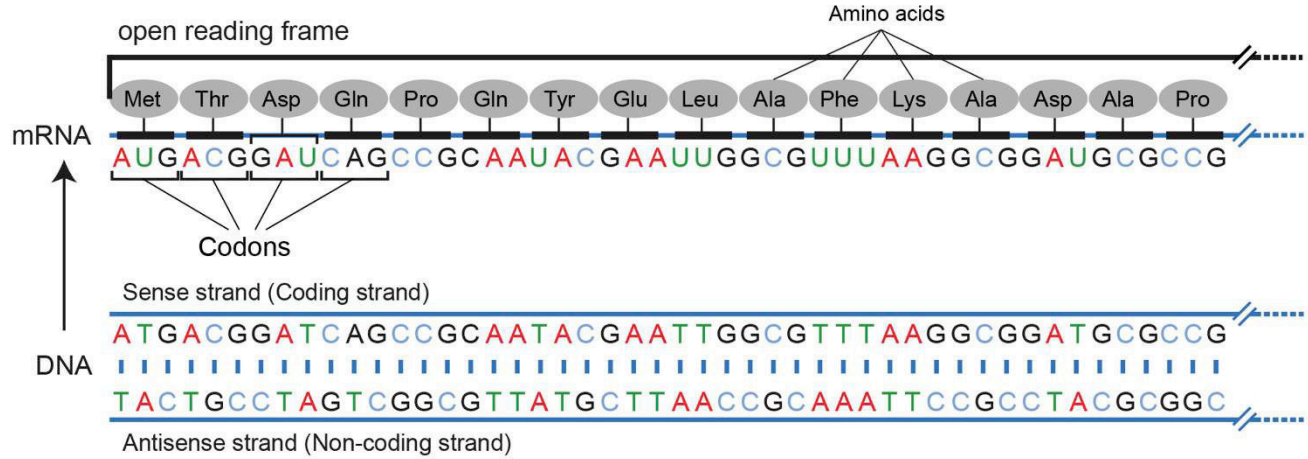
- Single Nucleotide Variant (**SNV**) or Single Nucleotide Polymorphism (**SNP**)
  - A single nucleotide — A, T, C or G — in the genome differs between members of a population
  - Bi-allelic or Multi-allelic
  - Can be in the coding sequences of genes, non-coding regions of genes or in the intergenic region
  - SNPs occur one in every 300 nucleotides, roughly 10 million SNPs in the human genome (minor allele frequency >1%)

Types of Genome Sequence Variants

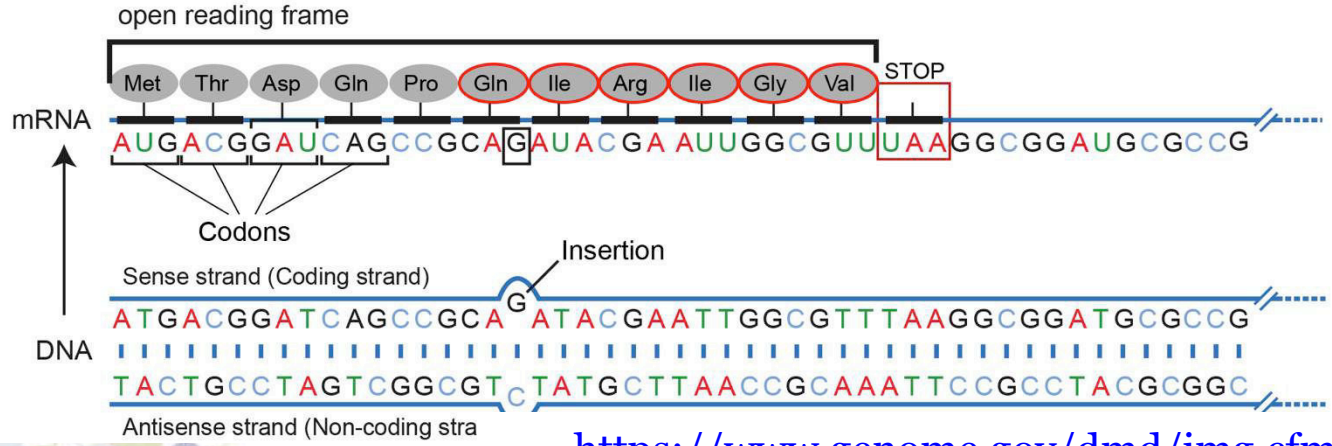
- Small insertions and deletions (**Indel**)
  - Insertion / deletion of bases
  - Ranges from 1 to 10,000 bp in length
  - Can be in the coding sequences of genes, non-coding regions of genes or in the intergenic region
- Structural variation (**SV**)
  - Approximately 1 kb and larger in length
  - Inversions and translocations or copy number variants (**CNV**s)

Sequence Variants – Frameshift Mutation

Normal



Frameshift mutation - single nucleotide insertion

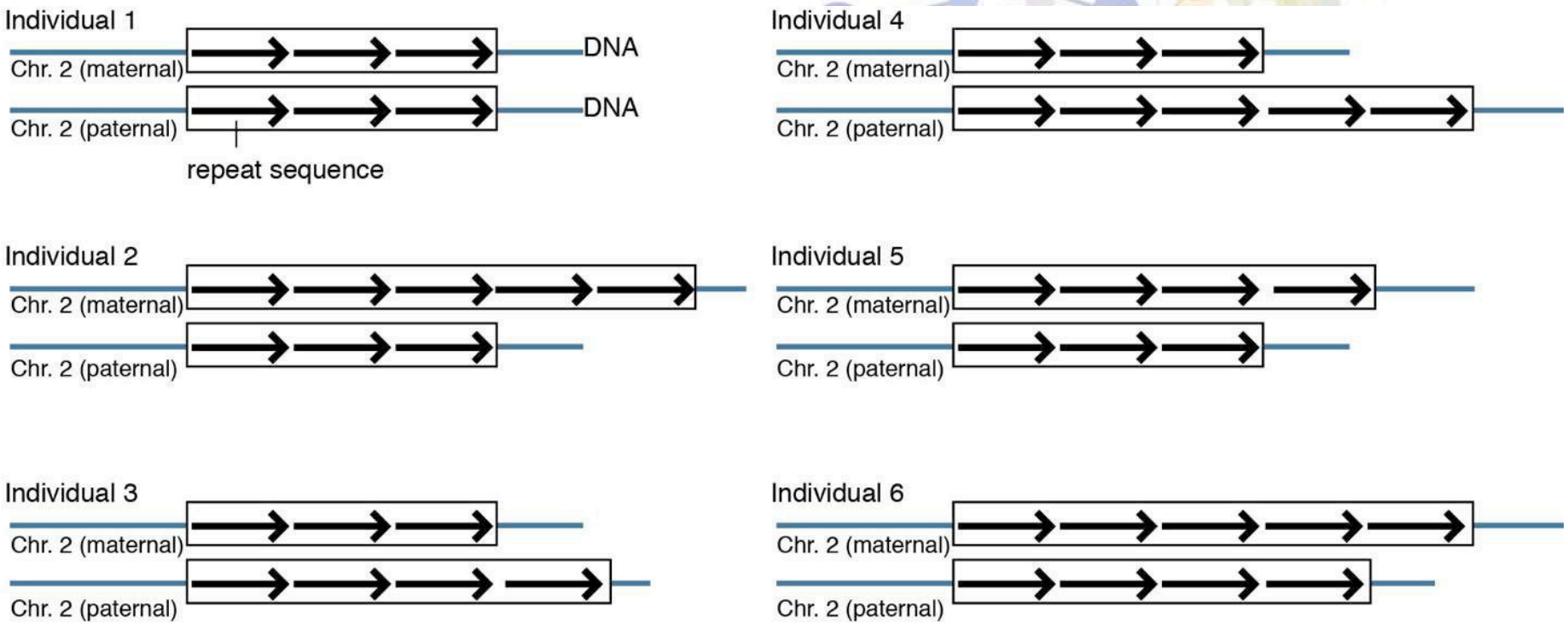


https://www.genome.gov/dmd/img.cfm?node=Photos/Graphics&id=85168

University of Glasgow

Glasgow Polyomics
www.glasgow.ac.uk/polyomics

Copy number variation (CNV)



https://www.genome.gov/dmd/img.cfm?node=Photos/Graphics&id=85286

Single nucleotide polymorphism (SNP)

Individual 1

```
Chr 2   ..CGATATTCCTATCGAATGTC..
copy1   ..GCTATAAGGAUAGCTTACAG..

Chr 2   ..CGATATTCCCATCGAATGTC..
copy2   ..GCTATAAGGGTAGCTTACAG..
```

Individual 2

```
Chr 2   ..CGATATTCCCATCGAATGTC..
copy1   ..GCTATAAGGGTAGCTTACAG..

Chr 2   ..CGATATTCCCATCGAATGTC..
copy2   ..GCTATAAGGGTAGCTTACAG..
```

Short tandem repeat polymorphism (STRP)

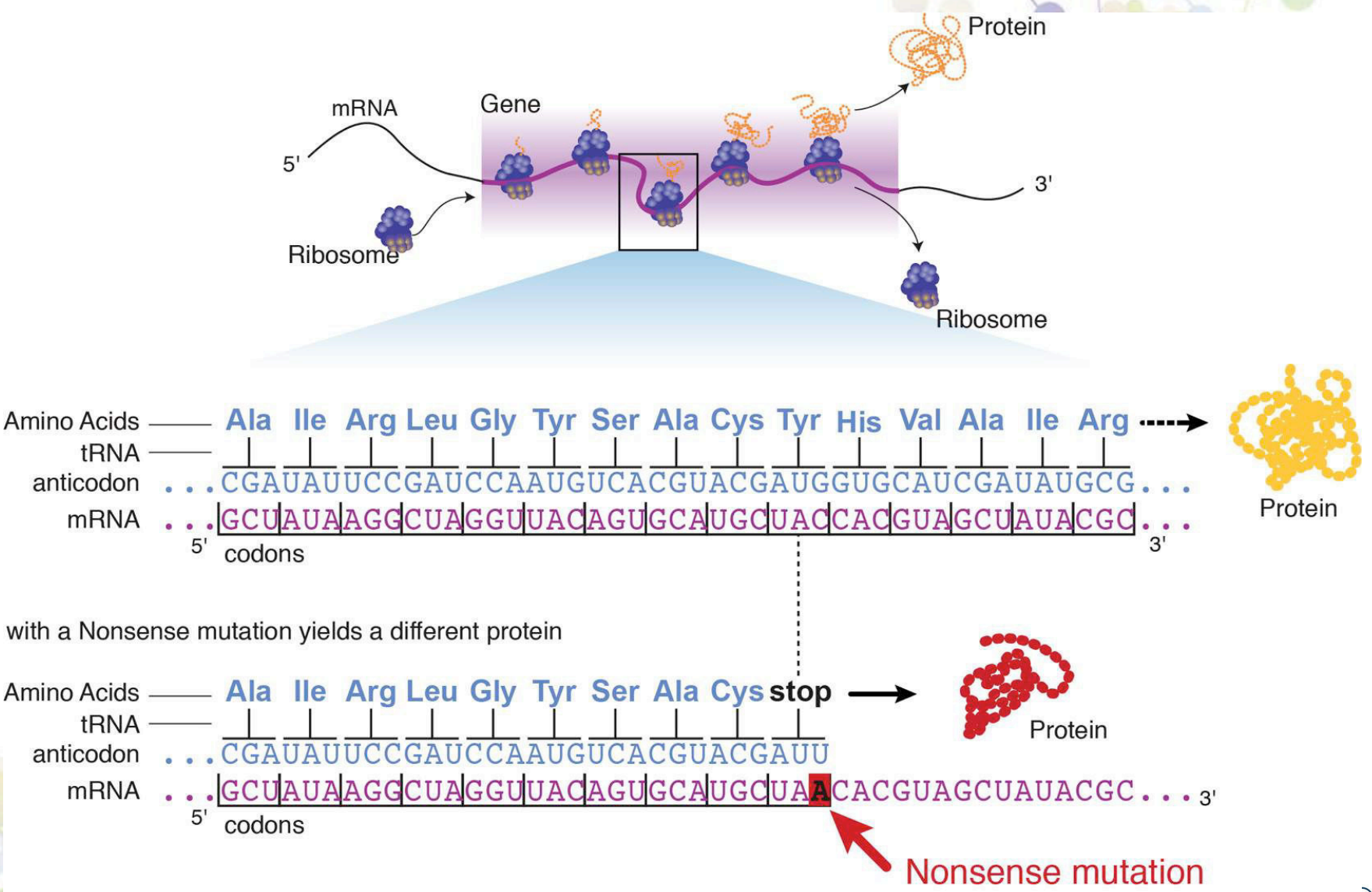
Individual 3 Repeat unit

```
Chr 2   ..CGATATTCCCAGCAGCAGATCGAATGTC..
copy1   ..GCTATAAGGCAGCAGCAGTAGCTTACAG..

Chr 2   ..CGATATTCCCAGCAGCAGCAGCAGATCGAATGTC..
copy2   ..GCTATAAGGCAGCAGCAGCAGCAGTAGCTTACAG..
```

Individual 4

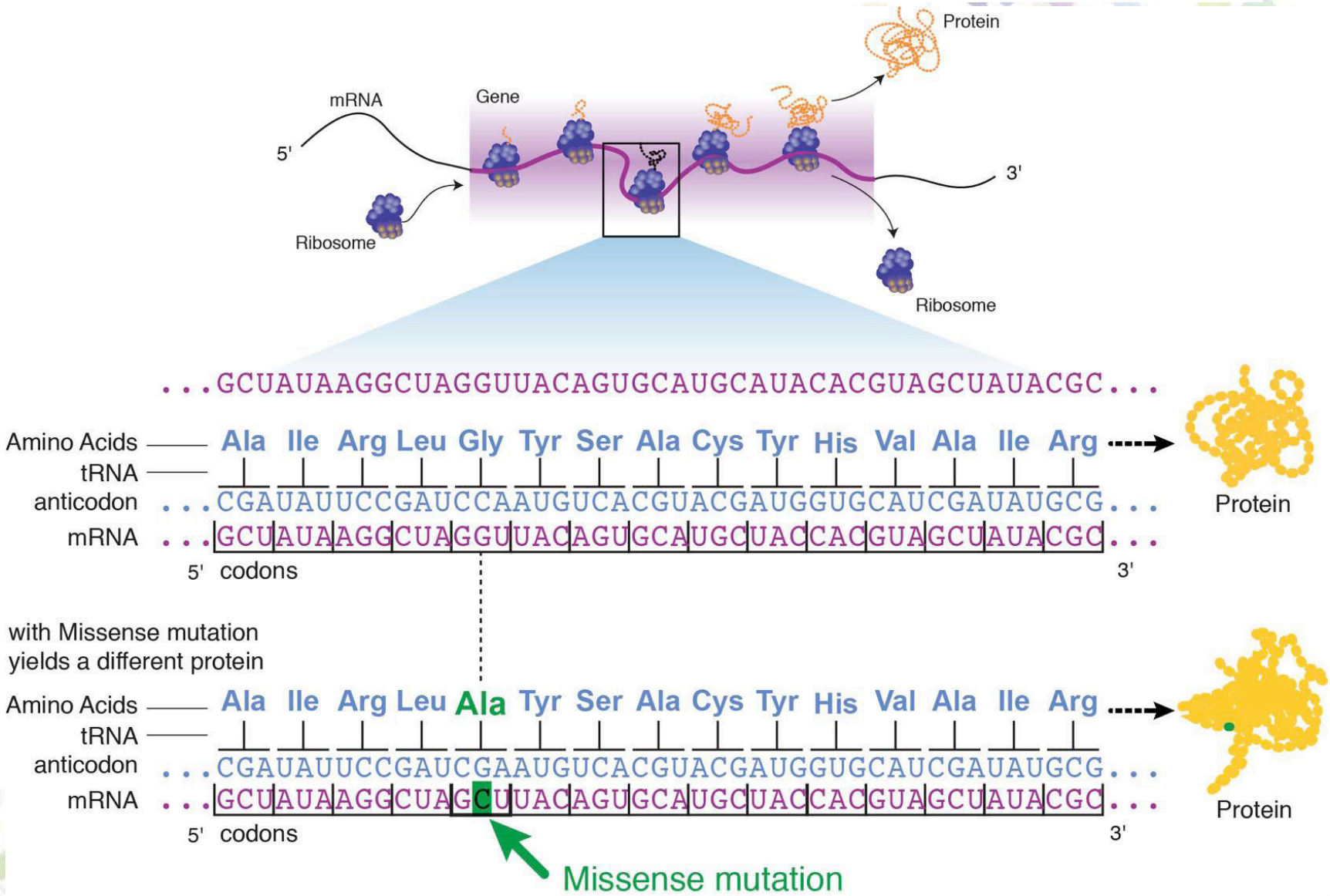
```
Chr 2   ..CGATATTCCCAGCAGCAGCAGCAGATCGAATGTC..
copy1   ..GCTATAAGGCAGCAGCAGCAGCAGTAGCTTACAG..

Chr 2   ..CGATATTCCCAGCAGCAGCAGCAGCAGCAGATCGAATGTC..
copy2   ..GCTATAAGGCAGCAGCAGCAGCAGCAGCAGTAGCTTACAG..
```

Sequence Variants – Nonsense Mutation



https://www.genome.gov/dmd/img.cfm?node=Photos/Graphics&id=85207

Sequence Variants – Missense Mutation



https://www.genome.gov/dmd/img.cfm?node=Photos/Graphics&id=85201

- **Introduction**
- **File formats and conventions**
- Databases used in variant analysis
- Variant analysis: Options
- Benchmarking and validation
- Variant analysis: A worked example

File formats and conventions - FASTA

- First used by Bill Pearson
- A single-line description (defline), followed by lines of sequence data
- The defline has a greater-than (">") symbol at the beginning
- Traditionally the sequence lines are limited to a width of 60 characters

```
>MT:647-1601
AATAGGTTTGGTCCTAGCCTTTCTATTAGCTCTTAGTAAGATTACACATGCAAGCATCCC
CGTTCCAGTGAGTTCACCCTCTAAATCACCACGATCAAAAGGAACAAGCATCAAGCACGC
AGCAATGCAGCTCAAAACGCTTAGCCTAGCCACACCCCCACGGGAAACAGCAGTGATTAA
CCTTTAGCAATAAACGAAAGTTTAACTAAGCTATACTAACCCCAGGGTTGGTCAATTTCG
```

File formats and conventions - FASTQ

- FASTQ files have sequence and quality data (PHRED quality score), and the quality values are single-byte encoded.
- *Reference: DOI: 10.1093/nar/gkp1137*

```
@NS500205:27:H15V6BGXX:1:11101:11986:1033 1:N:0:8
GCCCTNAGCGACCTGCACGCGCACAAGCTTCGGGTNGACCCGGTCAACTTCAAGCTCCTAAGCCACNGCCTGC
+
AAAAA#AFFF<FFFFFFFFFFFFFF<FFFFFFFFF#F<FFFFFFFF<FFFAFFFFFFFFFFFFFAF#FFFFFF
```

```
@NS500205:27:H15V6BGXX:1:11101:8152:1033 1:N:0:8
GCAAANCTGAAACTTAAAGGAATTGACGGAAGGGCNCCACCAGGAGAGGAGACTGCGGCTTAAAAANACACA
+
)A<A<#)FF<AFFFF<<.FAFAFF))FA<A)F..<#)F.A)FFF<F<..F..F.AF.)F.FFF.FA#)7)F.
```

File formats and conventions - GFF / GTF

- Generic Feature Format (GFF) / Gene Transfer Format (GTF)
- GFF2 = GTF

- One line per feature, each containing 9 tab-separated columns of data, plus optional track definition lines

- ID, Source, Feature type name, Start, End, Score, Strand, Frame, Attribute and track definition

- http://www.ensembl.org/info/website/upload/gff.html
- http://www.sequenceontology.org/gff3.shtml

File formats and conventions - SAM / BAM

- SAM stands for Sequence Alignment/Map format
- https://samtools.github.io/hts-specs/SAMv1.pdf
- TAB-delimited text format with a header section, and an alignment section
- The header lines begin with the character '@'
- The alignment lines have 11 mandatory fields and optional aligner specific fields
- BAM Format – 64Kb BGZF block compression on top of the standard gzip format

File formats and conventions - SAM / BAM

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

- SAM - 1-based coordinate system
- BAM - 0-based coordinate system
  - Supports random access through indexing
  - BAI index file

File formats and conventions - VCF

- Variant Call Format (VCF)
- http://www.1000genomes.org/wiki/Analysis/vcf4.0
- VCF contains meta-information lines, a header line, and data lines.
- Meta-information begins with ## string, often as key=value pairs
- The data lines each containing information about a position in the genome

File formats and conventions - VCF

8 mandatory columns:

1. #CHROM
2. POS
3. ID
4. REF
5. ALT
6. QUAL
7. FILTER
8. INFO

FORMAT (parameters) and values for each Sample

e.g. GT:GQ:DP:RO:QR:AO:QA:GL 1/1:17.0545:1:0:0:1:40:-4,-0.30103,0

File formats and conventions - BED

- The Browser Extensible Data (BED) format was developed by UCSC Genome Bioinformatics team to display data lines for genome browser annotation tracks
- The BED format consists of one line per feature, each containing 3-12 columns of data, plus optional track definition lines
- Required fields: chrom, chromStart and chromEnd
- Optional fields: name, score, strand, thickStart, thickEnd and itemRgb
- Track lines: space-separated key=value pairs
- 0-based coordinate system

https://genome.ucsc.edu/FAQ/FAQformat.html#format1

http://www.ensembl.org/info/website/upload/bed.html

Coordinate Systems

- **1-based coordinate system:** The first base of a sequence is one.
  - Region is specified by a closed interval. Eg. The region between the 3rd and the 7th bases inclusive is [3,7].
  - SAM, VCF and GFF formats, and Ensembl

- **0-based coordinate system:** the first base of a sequence is zero.
  - a region is specified by a half-closed-half-open interval. Eg. The region between the 3rd and the 7th bases inclusive is [2,7].
  - BAM, BED formats, and RefSeq and UCSC

- **Introduction**
- **File formats and conventions**
- **Databases used in variant analysis**
- Variant analysis: Options
- Benchmarking and validation
- Variant analysis: A worked example

Reference databases

- NCBI – RefSeq http://www.ncbi.nlm.nih.gov/refseq
- UCSC http://genome.ucsc.edu
- Ensembl http://www.ensembl.org/index.html
- dbSNP: Database for Short Genetic Variations
  http://www.ncbi.nlm.nih.gov/SNP/index.html
- dbVar: Database of genomic structural variations
  http://www.ncbi.nlm.nih.gov/dbvar
  http://www.ncbi.nlm.nih.gov/dbvar/content/overview
- ClinVar: Genomic variations and their relationship to human health and disease
- dbGaP: Database of Genotypes and Phenotypes (interactions of genotypes and phenotypes)
  http://www.ncbi.nlm.nih.gov/gap

Reference databases

- NCBI – RefSeq http://www.ncbi.nlm.nih.gov/refseq
- UCSC http://genome.ucsc.edu
- Ensembl http://www.ensembl.org/index.html

- UCSC/RefSeq and Genome Reference Consortium (GRCh)
  - hg18, hg19, hg38 = GRCh36, GRCh37, GRCh38
  - Latest version - hg38 (GRCh38)
  - Differences in naming chromosomes and sorting order

Ensembl Human Assembly and Annotation

- **Database version:** 79.38 (Jan 2015)
- **Base Pairs:** 3,384,269,757

- **Gene counts (Primary assembly)**
  - Coding genes: 20,300
  - Non coding genes: 24,885
  - Small non-coding genes: 7,715
  - Long non-coding genes: 14,863
  - Pseudogenes: 14,424
  - Gene transcripts: 198,622
  - Short Variants: 65,897,584
  - Structural variants: 4,168,103

Reference databases

- **dbSNP:**
  - Central repository for SNPs and Indels; Established in September 1998
  - Information for variants: Population, Sample Size, allele frequency, genotype frequency, heterozygosity, etc
  - High False Positive rate; About 40% not validated SNPs

BUILD STATISTICS:

Organism	dbSNP Build	Genome Build	Number of Submissions (ss#'s)	Number of RefSNP Clusters (rs#'s) (# validated)	Number of (rs#'s) in gene	Number of (ss#'s) with genotype	Number of (ss#'s) with frequency
Homo sapiens	144	38.2	505,875,709	149,735,377 (97,535,033)	85,591,044	73,909,260	45,812,686

- **ClinVar:** ClinVar aggregates information about genomic variation and its relationship to human health.

  http://www.ncbi.nlm.nih.gov/clinvar

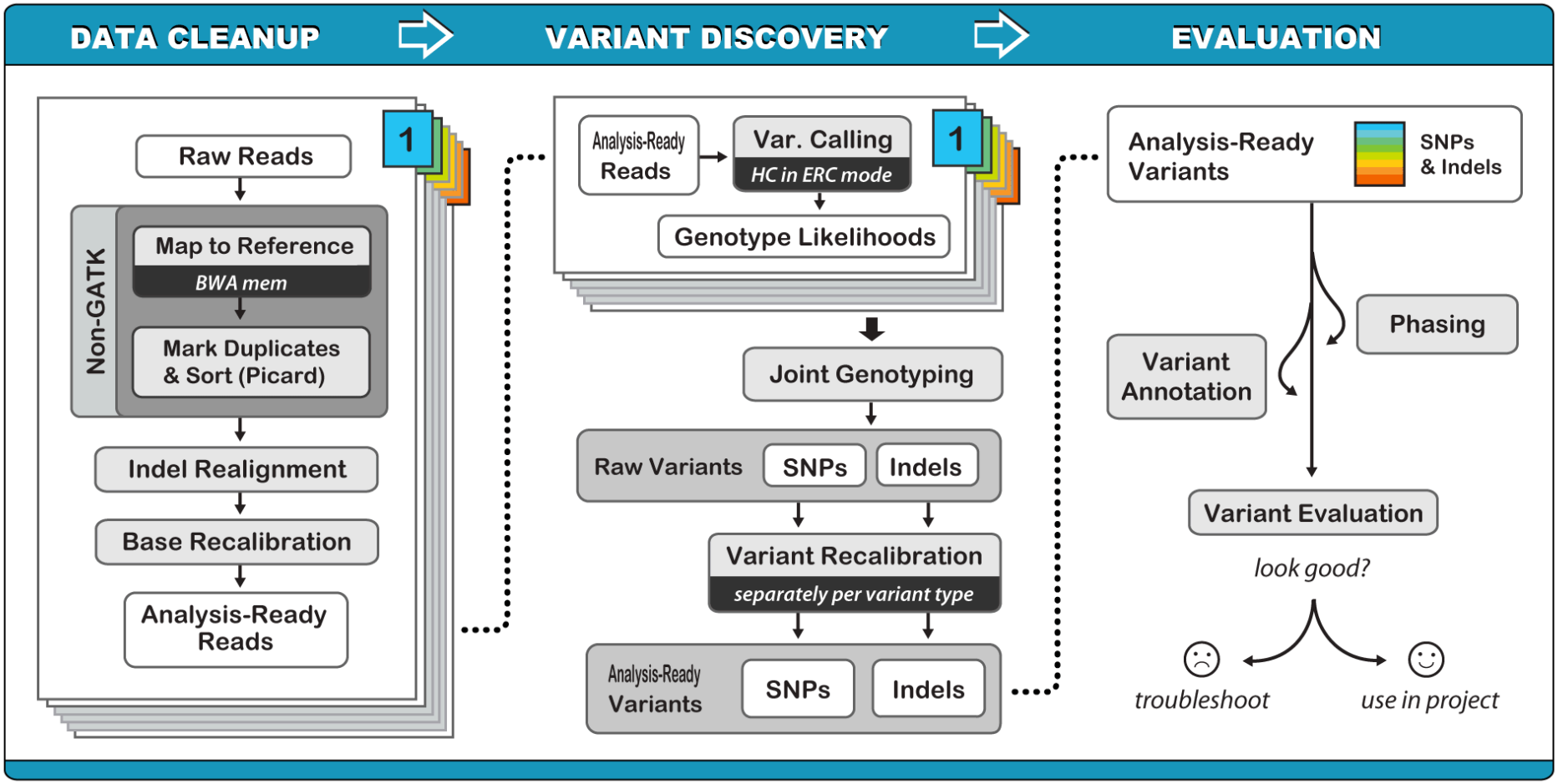
HGVS databases

- Locus Specific Mutation Databases
- Disease Centered Central Mutation Databases
- National & Ethnic Mutation Databases
- Mitochondrial Mutation Databases
- Chromosomal Variation Databases
- Other Mutation Databases
- Clinical & Patient Aspects Databases
- Non Human Mutation Databases
- Artificial Mutations Only

http://www.hgvs.org/content/databases-tools

- **Introduction**
- **File formats and conventions**
- **Databases used in variant analysis**
- **Variant analysis: Options**
- Benchmarking and validation
- Variant analysis: A worked example

Variant analysis workflow



University of Glasgow

Glasgow Polyomics
www.glasgow.ac.uk/polyomics

Open source tools for QC before mapping

- Remove non genomic sequences (barcodes, adapter, ...)
- Remove contaminations (PRINSEQ, DeconSeq)
- Quality Trimming – Remove bad quality reads
- FASTQC - http://www.bioinformatics.babraham.ac.uk/projects/fastqc
- Cutadapt - https://code.google.com/p/cutadapt
- Sickle - https://github.com/ucdavis-bioinformatics/sickle
- Scythe - https://github.com/vsbuffalo/scythe
- Fastx toolkit - http://hannonlab.cshl.edu/fastx_toolkit
- DeconSeq - http://deconseq.sourceforge.net
- PRINSEQ - http://prinseq.sourceforge.net

Open source tools for alignment

Mapping with reference genome

- Burrows-Wheeler Aligner (BWA)
  http://bio-bwa.sourceforge.net
- Bowtie 2
  http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

Quality check of alignment

- Qualimap - http://qualimap.bioinfo.cipf.es
- Picard - https://broadinstitute.github.io/picard
- Samtools - http://samtools.sourceforge.net
  http://www.htslib.org
- Bamstats - http://bamstats.sourceforge.net

BAM pre-processing

- **Sorting**
  - **Coordinate sorting –** Picard / Samtools
- **Remove / mark PCR duplicates:** Picard / Samtools
- **Local Realignment Around Indels -** GATK
- **Base Quality Score Recalibration (BQSR) -** GATK

Variant Discovery

- **SAMTools** http://samtools.sourceforge.net
  http://www.htslib.org
- **GATK** https://www.broadinstitute.org/gatk
- **Platypus** http://www.well.ox.ac.uk/platypus
- **Freebayes** https://github.com/ekg/freebayes
- **BreakDancer** http://breakdancer.sourceforge.net
- **Pindel (**doi: 10.1093/bioinformatics/btp394)
- **Dindel** https://www.sanger.ac.uk/resources/software/dindel

Variant Annotation

- **GATK-VariantAnnotator**
- **SnpEff** http://snpeff.sourceforge.net
- **ANNOVAR** http://annovar.openbioinformatics.org
- **Ensembl** - Variant Effect Predictor
  http://www.ensembl.org/info/docs/tools/vep/index.html
- **PheGenI** - Phenotype-Genotype Integrator
  http://www.ncbi.nlm.nih.gov/gap/phegeni
- **Variation Reporter** - accessing the content of NCBI human variation resources
  http://www.ncbi.nlm.nih.gov/variation/tools/reporter

Variant Interpretation

We are interested in identifying the consequences of every variation

- **Genomic location –** coding, non-coding region,….
- **Co-located known variants**
- **SIFT:** (http://sift.jcvi.org) predict if an amino acid substitution affects protein function
- **PolyPhen:** (http://genetics.bwh.harvard.edu/pph2) predict possible impact of an amino acid substitution on the structure and function of a protein
- **SuSPect:** (http://www.sbg.bio.ic.ac.uk/~suspect) sequence-, structure- and systems biology-based features to predict the phenotypic effects of missense mutations
- **MutationTaster** (http://www.mutationtaster.org)

Variant Filtering

- **Variant filtering with Variant Quality Score Recalibration (VQSR) - GATK**
- **Filter by minor allele frequency (MAF)**
- **Results for variants in coding regions only**
- **Show selected consequence only**
- **Transition Transversion Ratio (Ti/Tv)**

- **https://github.com/ekg/vcflib**
- **http://vcftools.sourceforge.net/index.html**

Variant Visualization

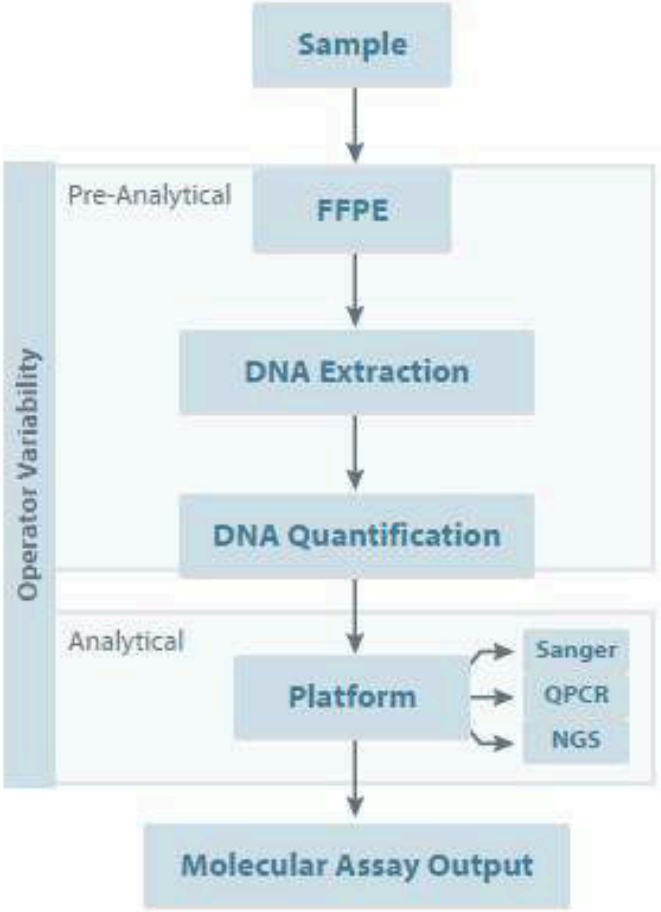
- Examine the results using a genome browser
  - IGV https://www.broadinstitute.org/igv/home
  - UCSC Genome Browser
  - Tablet http://ics.hutton.ac.uk/tablet

University of Glasgow

Glasgow Polyomics
www.glasgow.ac.uk/polyomics

- **Introduction**
- **File formats and conventions**
- **Databases used in variant analysis**
- **Variant analysis: Options**
- **Benchmarking and validation**
- Variant analysis: A worked example

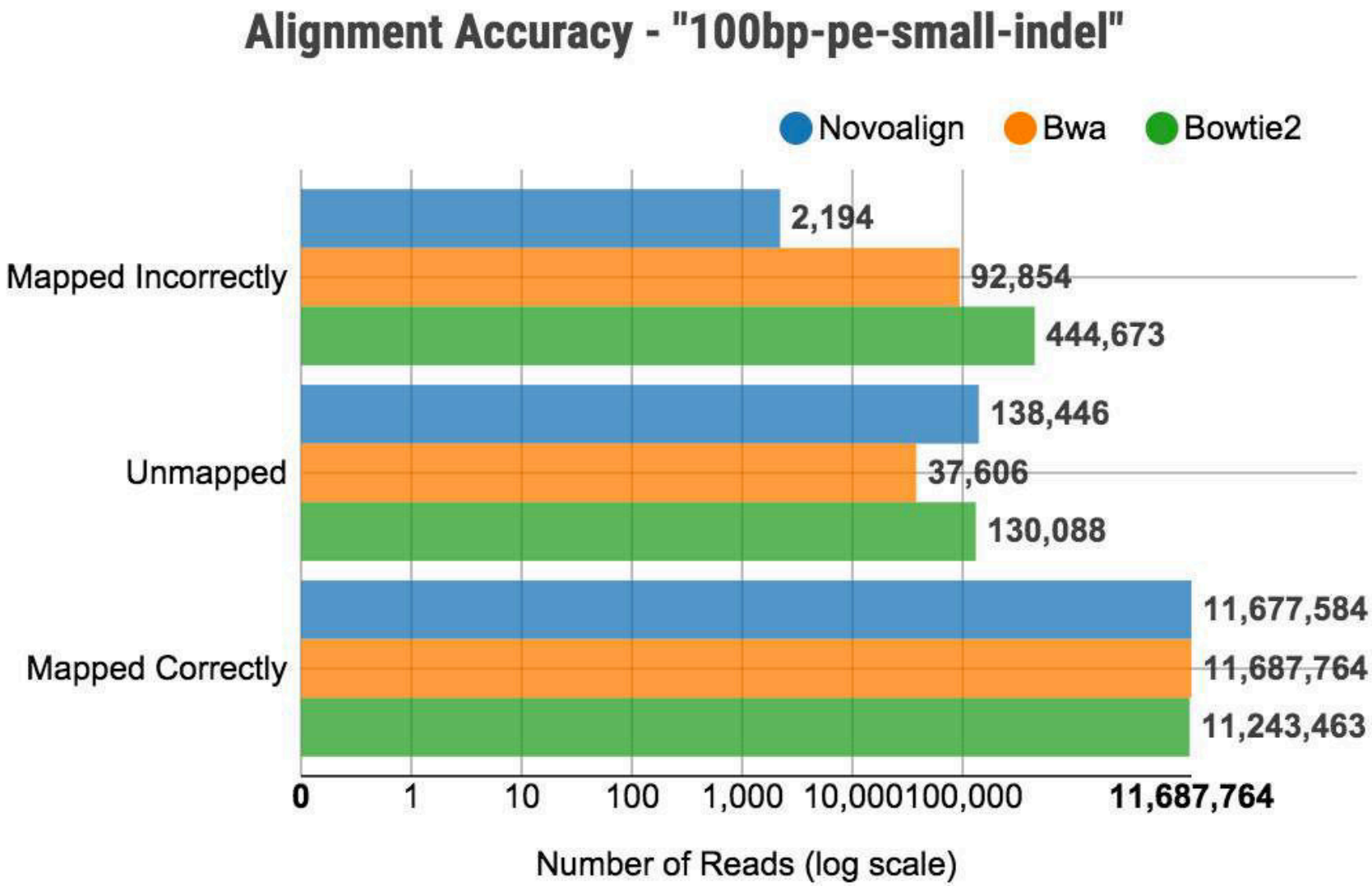
Importance of benchmarking and validation



Sources of variability within a standard molecular assay workflow

http://www.horizondx.com/scientific-support/sanger-qpcr-sequencing/ffpe-sections.html

Benchmarking and Validation

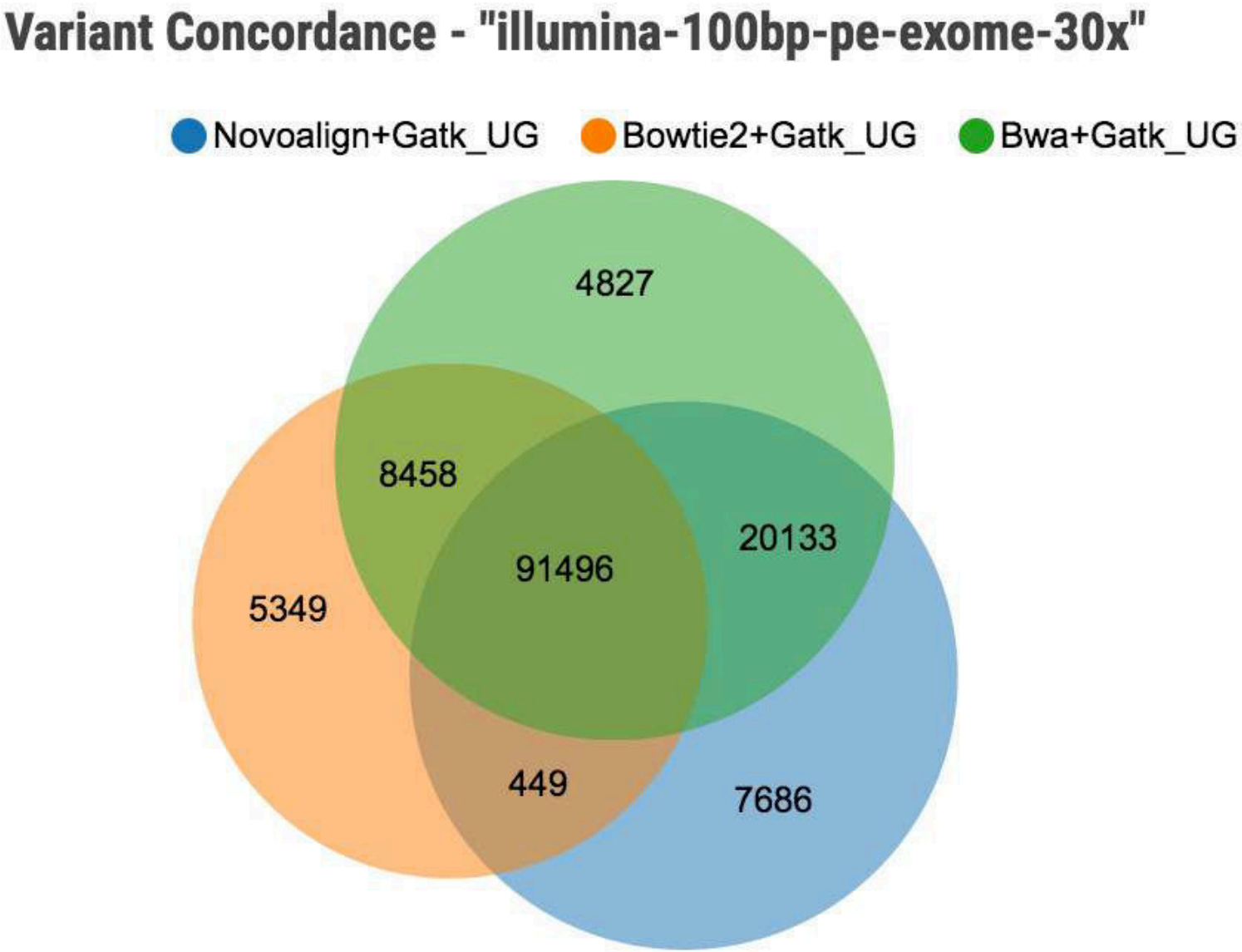


http://www.bioplanet.com/gcat

University of Glasgow

Glasgow Polyomics
www.glasgow.ac.uk/polyomics

Benchmarking and Validation

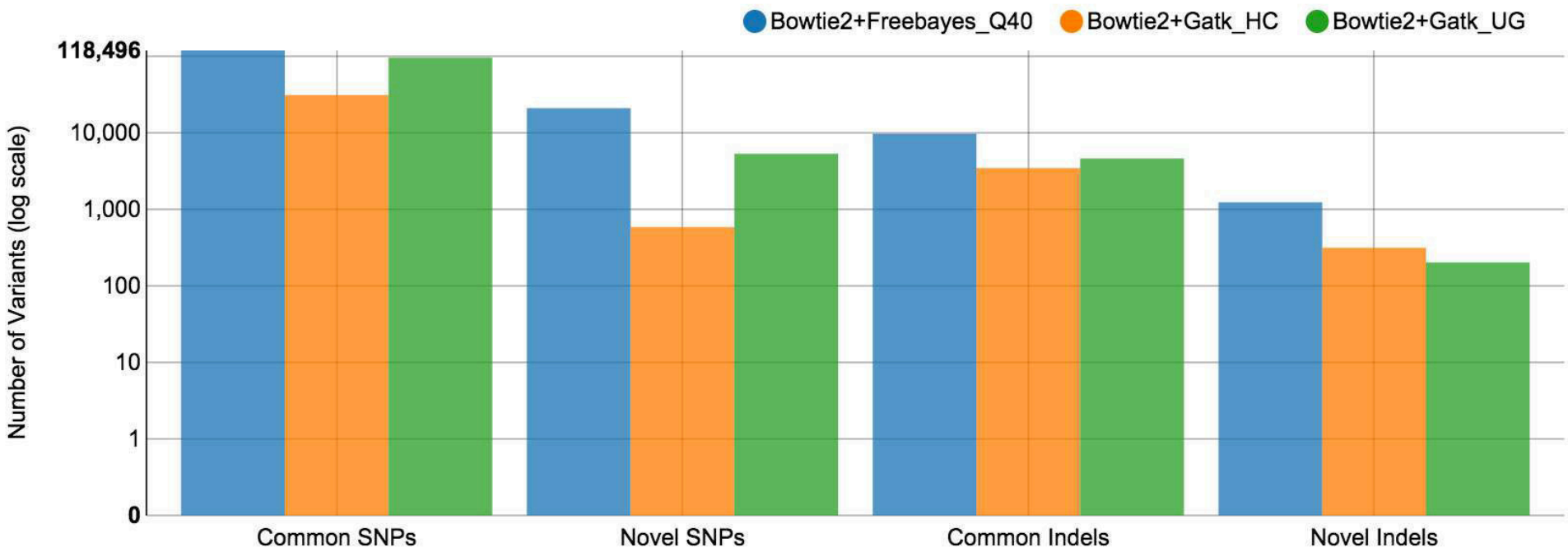


http://www.bioplanet.com/gcat

University of Glasgow

Glasgow Polyomics
www.glasgow.ac.uk/polyomics

Benchmarking and Validation

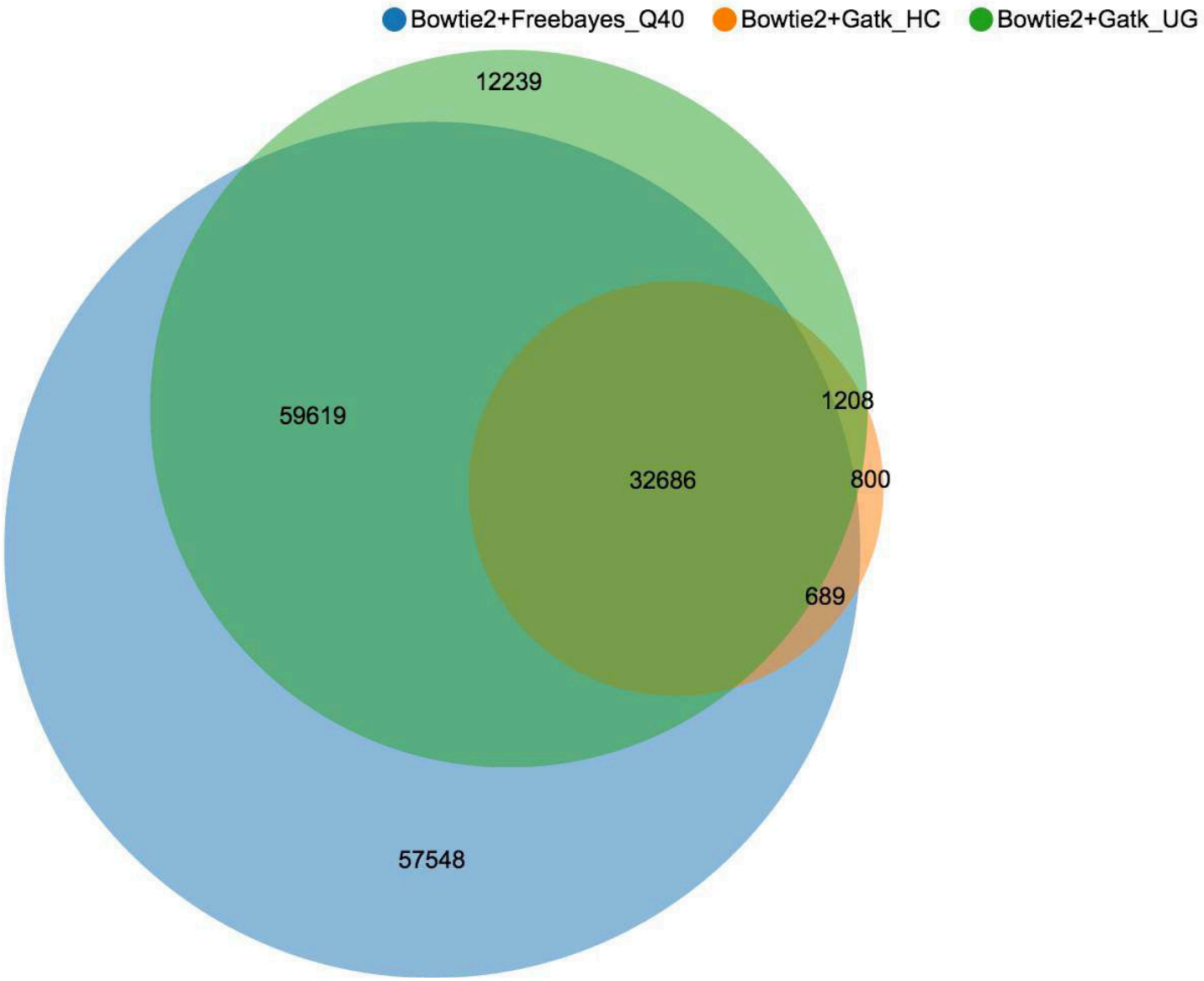


http://www.bioplanet.com/gcat

University of Glasgow

Glasgow Polyomics
www.glasgow.ac.uk/polyomics

Benchmarking and Validation



http://www.bioplanet.com/gcat

University of Glasgow

Glasgow Polyomics
www.glasgow.ac.uk/polyomics

Benchmarking and Validation

Advances in Biological and Medical Measurement Science

Log in with WebAuth

Home | Program Components | People | ERCC 2.0 Workshop | **Genome in a Bottle Consortium** | Membership

Local User Login

Genome in a Bottle Consortium

Home » Genome in a Bottle Consortium

The Genome in a Bottle Consortium is a public-private-academic consortium initiated by NIST to develop the technical infrastructure (reference standards, reference methods, and reference data) to enable translation of whole human genome sequencing to clinical practice. The consortium was initiated in a set of meetings in 2011 and 2012, and the consortium now holds biannual public workshops in January at Stanford University and in August at NIST in Gaithersburg, MD. The next workshop will be held on August 27-28, 2015 in Maryland. Agenda and registration information will be available soon.

Below you will find the most recent blog posts, and in the bar to the left you can find links to the Reference Materials and Data we are producing, as well as a link to sign up for our email list on Google Groups.

Members

life technologies™

NOVARTIS

- NIST - Genome in a Bottle Consortium
  https://sites.stanford.edu/abms/giab
- A public-private-academic consortium initiated by NIST to develop the technical infrastructure (reference standards, reference methods, and reference data) to enable translation of whole human genome sequencing to clinical practice.

Benchmarking and Validation

Q-Seq HDx™ Reference Standards

Independent external controls designed to routinely validate workflows and assays to ensure consistency and accuracy across laboratories, manufacturers, assays and platforms. Analyze and evaluate variant calling sensitivity, utilize the breadth of variants to understand assay specificity and accurately quantify the limit of detection for each variant.

20 Item(s) Sort By Name Show 25 View as

ASHKENAZIM PGP FATHER - FFPE REFERENCE STANDARD

Format: FFPE (Genome In A Bottle)
Product Code: GM24149

Unit Size: FFPE Section

Please be advised this product is currently out of stock.
The expected date of availability is May 2015.

If you would like to pre-order this product, please contact us here

Learn More

£65.00

ASHKENAZIM PGP MOTHER - FFPE REFERENCE STANDARD

Format: FFPE (Genome In A Bottle)
Product Code: GM24143

Unit Size: FFPE Section

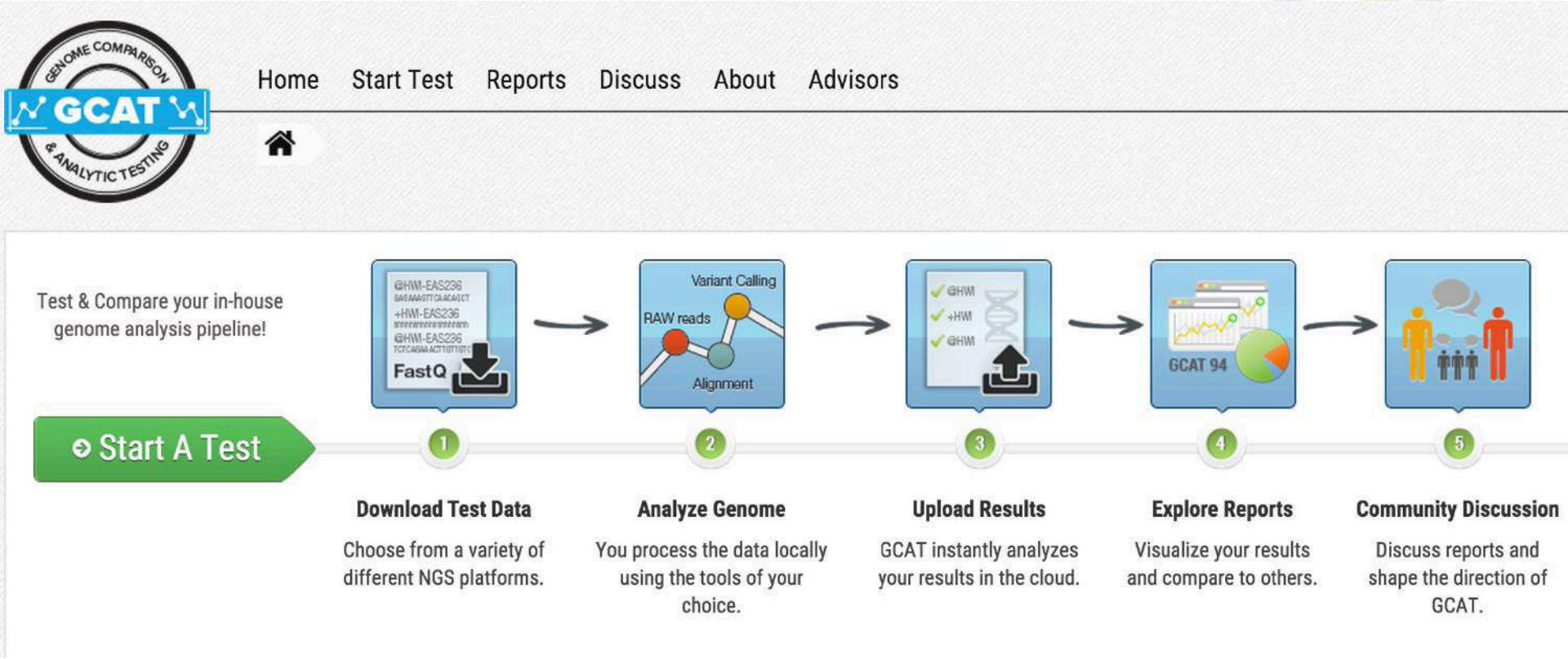
Learn More

£65.00

ASHKENAZIM PGP SON - FFPE REFERENCE STANDARD

- Horizon Diagnostics Q-Seq HDx™ Reference Standards http://www.horizondx.com/products/q-seq-ngs.html

Benchmarking and Validation



- **GCAT** – Genome Comparison & Analytic Testing
  http://www.bioplanet.com/gcat/
- **GCAT** is a collaborative platform for comparing multiple genome analysis tools across a standard set of metrics.

Benchmarking and Validation

bcbio

A python toolkit providing best-practice pipelines for fully automated high throughput sequencing analysis. You write a high level configuration file specifying your inputs and analysis parameters. This input drives a parallel pipeline that handles distributed execution, idempotent processing restarts and safe transactional steps. The goal is to provide a shared community resource that handles the data processing component of sequencing analysis, providing researchers with more time to focus on the downstream biology.

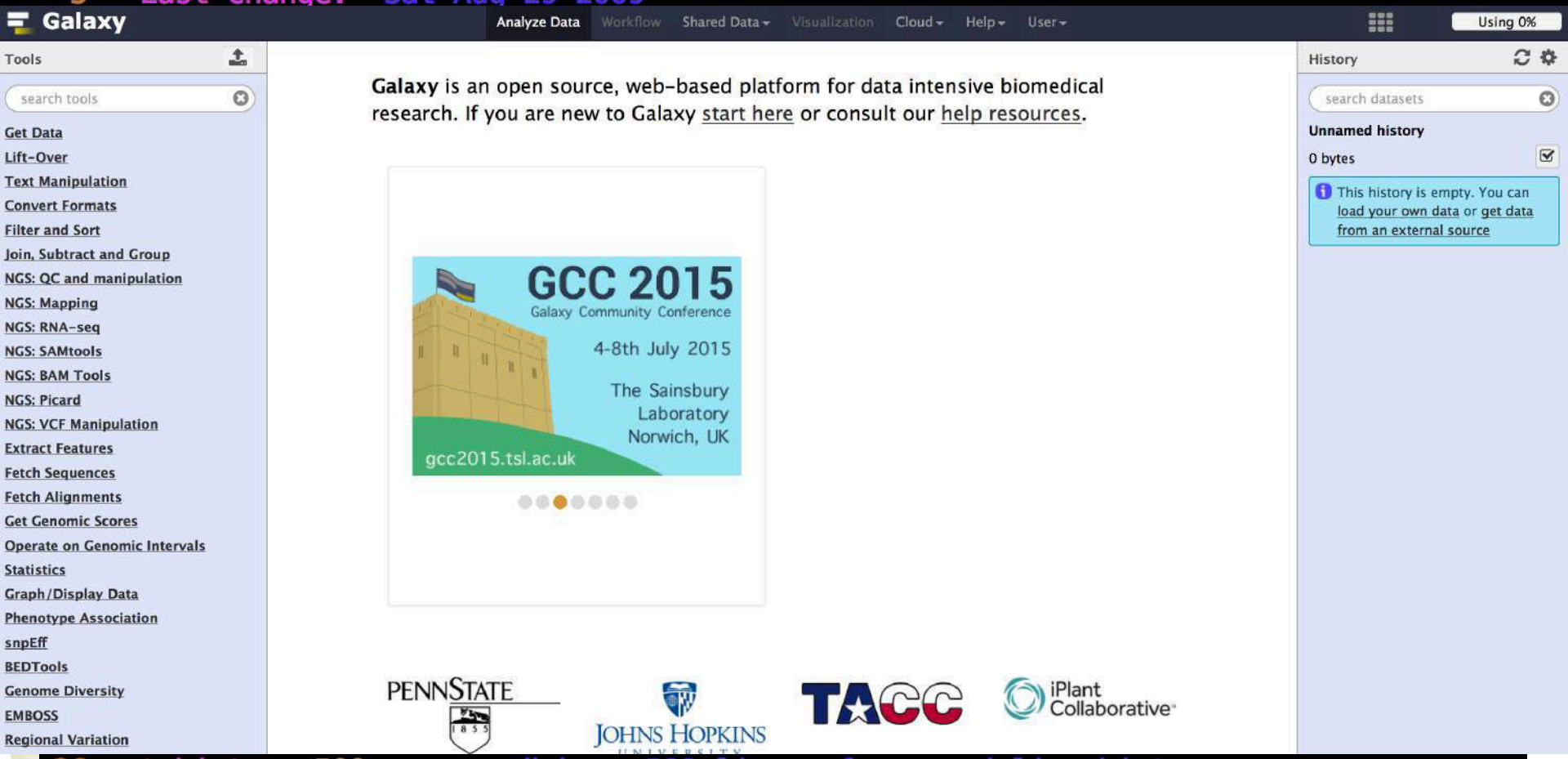
Validated, scalable and community developed analysis pipelines
https://github.com/chapmanb/bcbio-nextgen

- Introduction
- File formats and conventions
- Databases used in variant analysis
- Variant analysis: Options
- Benchmarking and validation
- Variant analysis: A worked example

Hail Galaxy!

- **Galaxy is available online, for free, to every one!**
  - **Empowering biologists!**
  - **Democratizing computational resources and skills!**
  - **Free access to high-performance computers and free data storage (250 GB or more)**
  - **Free tools and workflows**
  - **Training – workshops and online**
  - **Dedicated person/team to answer your queries**
  - **Active community**
  - **Sharing best practices, workflows, histories and data**

```
" This file was copied by Mani Mudaliar on 01-01-2015 from the 'System vimrc file'
" Last change by "Mani Mudaliar" on "02-01-2015"
"
" Maintainer:   Bjorn Winckler <bjorn.winckler@gmail.com>
" Last Change:  Sat Aug 29 2009
```




```
set history=500          " keep 500 lines of command line history
set ruler                " show the cursor position all the time
set showcmd              " display incomplete commands
set incsearch            " do incremental searching
```

Mitochondrial Heteroplasmy Analysis

- This example is based on the Galaxy NGS 101 tutorial available at https://wiki.galaxyproject.org/Learn/GalaxyNGS101

- I have borrowed slides from Dave Clements https://wiki.galaxyproject.org/Documents/Presentations?action=AttachFile&do=view&target=ESHG_2015_Variant.pdf

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Cloud | Help | User

Published Pages | galaxyproject | Galaxy 101 NGS: Introduction to Polymorphism Detection via Variant Analysis

Galaxy 101 NGS: Introduction to Polymorphism Detection via Variant Analysis

Heteroplasmy: Mother-Child mtDNA Variant Polymorphism

- heteroplasmy
- ismb2010-demo

This tutor

- Import Seque
- Interpret FAST
- Execute a seri
- Execute one c
- Filter key resu

What is H

The heteroplast
mtDNA heterop
research in the

Being heteropla

- Go ahead, sta

Experimer

Import child an
(*where Q20 in
Convert the res
and FreeBayes.
Variant Annotat

- Explore the re
- Questions:
  - Can you id
  - Can you id
    How could
  - Are any SN
    Do any po
  - How would

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Cloud | Help | User

Published Pages | aun1 | heteroplasmy

Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study

Hiroki Goto[1], B[illegible]

Published in Geno

Correspondence s

1. How to

This document is a
with them by re-ru
hassle-free proced

access our datasets
re-use workflows li
view and import his
In addition, we crea

Watch the analysis
Watch how the com
If you experience a

2. Accessi

All datasets discuss

A Galaxy Library ca
An S3 bucket on th
From there these d
family is "F4", "F7",

Goto *et al. Genome Biology* 2011, **12**:R59
http://genomebiology.com/2011/12/6/R59

Genome Biology

RESEARCH | Open Access

Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study

Hiroki Goto[1†], Benjamin Dickins[2†], Enis Afgan[3], Ian M Paul[4], James Taylor[3*], Kateryna D Makova[1*] and

PNAS

Maternal age effect and severe germ-line bottleneck in the inheritance of human mitochondrial DNA

Boris Rebolledo-Jaramillo[a,1], Marcia Shu-Wei Su[b,1], Nicholas Stoler[a], Jennifer A. McElhoe[c], Benjamin Dickins[d], Daniel Blankenberg[a], Thorfinn S. Korneliussen[e,f], Francesca Chiaromonte[g], Rasmus Nielsen[e], Mitchell M. Holland[c], Ian M. Paul[h], Anton Nekrutenko[a,2], and Kateryna D. Makova[b,2]

Departments of [a]Biochemistry and Molecular Biology, [b]Biology, and [g]Statistics, [c]Forensic Science Program, Pennsylvania State University, University Park, PA 16802; [d]School of Science and Technology, Nottingham Trent University, Nottingham NG1 4BU, United Kingdom; [e]Department of Integrative Biology, University of California, Berkeley, CA 94720; [f]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, DK-1350 Copenhagen, Denmark; and [h]Department of Pediatrics, College of Medicine, Pennsylvania State University, Hershey, PA 17033

Edited by Michael Lynch, Indiana University, Bloomington, IN, and approved September 8, 2014 (received for review May 20, 2014)

Mitochondrial Heteroplasmy

Mitochondrial heteroplasmy - the existence of different mtDNA sequences within an individual due to somatic or inherited mutations

Dataset: Publicly available
http://www.ncbi.nlm.nih.gov/sra/SRP047378

SRA

SRA | SRP047378

Save search Advanced

Access
Public (156)

Source
DNA (156)

Clear all

Show additional filters

Display Settings: Summary, 200 per page

Results: 156

1. full length mtDNA sequencing of child SC8C1: whole blood
   1 ILLUMINA (Illumina MiSeq) run: 1.4M spots, 597.8M bases, 382.3Mb downloads
   Accession: SRX707999

2. full length mtDNA sequencing of mother SC8: whole blood
   1 ILLUMINA (Illumina MiSeq) run: 1.3M spots, 570.9M bases, 352.3Mb downloads
   Accession: SRX707998

Mitochondrial Heteroplasmy Dataset

- 39 healthy mother–child pairs
- Two tissues: Blood and Buccal mucosa
- 156 samples (39 mothers × 2 tissues + 39 children × 2 tissues)
- Amplicons: mtDNA from two overlapping 9-kb fragments
- Nextera XT libraries
- MiSeq 250bp paired-end reads
- ~1 million reads per sample (~60x coverage)
- For the workshop purpose – two samples
- Mother and child blood samples

Mitochondrial Heteroplasmy Analysis

- I have shared and published the history of this analysis performed live on this workshop (https://wiki.galaxyproject.org/Events/Glasgow2015)

- The analysis history is available at: https://test.galaxyproject.org/u/mmudaliar/h/variantcallingglasgowworkshop20150609manimudaliar

- Updated slides used in this section are available at: http://www.slideshare.net/drmani_vet

Galaxy Analysis – QC and Manipulation

- **Import Data into current history**
  - **Shared Data → Data Libraries → Training → Heteroplasmy → M512 and import**
  - **M512-bl_1 – Mother, Blood, Forward**
  - **M512-bl_2 – Mother, Blood, Reverse**
  - **M512C2-bl_1 – Child, Blood, Forward**
  - **M512C2-bl_2 – Child, Blood, Reverse**
- **FastQC**
- Adapter trimming - Paired-end mode Cutadapt or Trimmomatic
- Quality trimming?
- Trim off the first 12bp? See FastQC report

Galaxy Analysis - Mapping

- **Reference genome → Homo sapiens b38/ hg38**

- **Samples - M512-bl_1, M512-bl_2, M512C2-bl_1 & M512C2-bl_2**

- **Map with BWA-MEM**
  - **Set Read group ID, Read Group Sample Name , Library Name, Platform unit**

Galaxy Analysis – BAM manipulations

- **Remove PCR duplicates – Picard MarkDuplicates**
  - **Remove Duplicates → Yes**

- **Filter Bam - NGS BAM Tools → Filter**
  - **Mapping Quality → >=20**
  - **Insert Filter → isProperPair: Yes**
  - **Insert Filter → reference: chrM**

Galaxy Analysis – Variant analysis

- **Variant Calling – NGS: Variant Analysis → FreeBayes - bayesian genetic variant detector**
  - Sample BAM file → Mother.bam
  - Sample BAM file → Child.bam
  - Using reference genome → hg38
  - Limit to Region → chrM, Start 1, End 16,500
  - Choose parameter selection level → Complete list of all option
  - Set population model? → Yes
  - Set ploidy for the analysis → 1
  - Output all alleles which pass input filters, regardless of genotyping outcome or model → Yes (--pooled-continuous)

Galaxy Analysis – Variant analysis

- Ploidy for the analysis → 1 (Remember we are analyzing Mitochondrial variation!)
- --pooled-continuous (Remember the buzzword heteroplasmy!)
- FreeBayes can act as a frequency-based pooled caller and describe variants and haplotypes in terms of observation frequency rather than called genotypes. To do so, use --pooled-continuous and set input filters to a suitable level. Allele observation counts will be described by AO and RO fields in the VCF output.

Galaxy Analysis – Variant analysis

- **VCF filtering – NGS: VCF Manipulation → VCFfilter: filter VCF data in a variety of attributes**
  - Specify filtering expression → **-f "DP > 10" –f "QUAL >30"**

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
chrM	73	.	A	G	35390.5	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=1170;CIGAR=1X;DP=1174
chrM	263	.	A	G	14398.1	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=497;CIGAR=1X;DP=497;D
chrM	309	.	CT	CCTC,CC	2524.71	.	AB=0,0;ABP=0,0;AC=2,0;AF=1,0;AN=2;AO=185,86;CIGAR=1M
chrM	513	.	GCACACACACAC	GCACACACACACAC	5134.87	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=262;CIGAR=1M2I11M;DP=
chrM	750	.	A	G	63299.9	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=1977;CIGAR=1X;DP=1979
chrM	1438	.	A	G	82467.2	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=2491;CIGAR=1X;DP=2493
chrM	2706	.	A	G	48619.3	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=1730;CIGAR=1X;DP=1740
chrM	3197	.	T	C	134897	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=4055;CIGAR=1X;DP=4055
chrM	3243	.	A	G	24163.9	.	AB=0;ABP=0;AC=1;AF=0.5;AN=2;AO=1440;CIGAR=1X;DP=27
chrM	4769	.	A	G	57315.2	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=1777;CIGAR=1X;DP=1780
chrM	5539	.	A	G	6718.25	.	AB=0;ABP=0;AC=1;AF=0.5;AN=2;AO=472;CIGAR=1X;DP=780
chrM	7028	.	C	T	78996.6	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=2433;CIGAR=1X;DP=2443

Galaxy Analysis – Variant analysis

Try yourself

- **Compare variants between Mother and Child**
  - Allele observation counts - AO and RO fields in the VCF file
  - NGS: VCF Manipulation → VCFselectsamples: Select samples from a VCF dataset
  - NGS: VCF Manipulation → VCF-VCFintersect: Intersect two VCF datasets
  - NGS: VCF Manipulation → VCFcommonSamples: Output records belonging to samples commong between two datasets

Galaxy Analysis – Variant annotation

http://www.ensembl.org/Homo_sapiens/Tools/VEP/

e!Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Human (GRCh38.p2) ▼ VEP ▼

Web Tools

- Web Tools
  - BLAST/BLAT
  - **Variant Effect Predictor**
  - Assembly Converter

Configure this page

Add your data

Export data

Share this page

Bookmark this page

Variant Effect Predictor

New VEP job:

VEP for Human GRCh37

If you are looking for VEP for Human GRCh37, please go to GRCh37 website.

Input

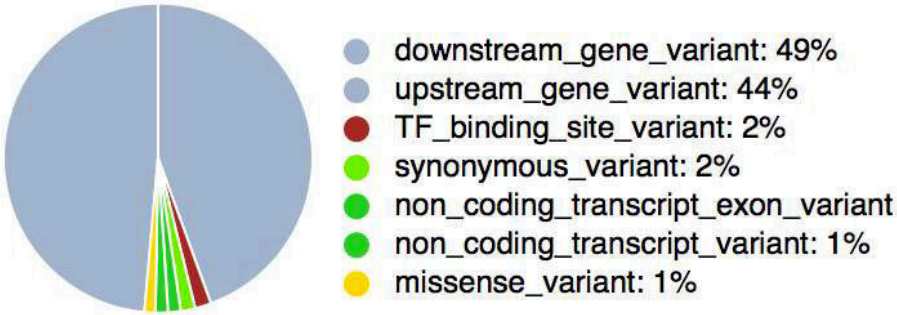
Species:	Human (Homo sapiens) Assembly: GRCh38.p2
Name for this data (optional):	
Either paste data:	**Examples:** Ensembl default, VCF, Variant identifiers, HGVS notations, Pileup
Or upload file:	Browse... No file selected.

Galaxy Analysis – Variant annotation

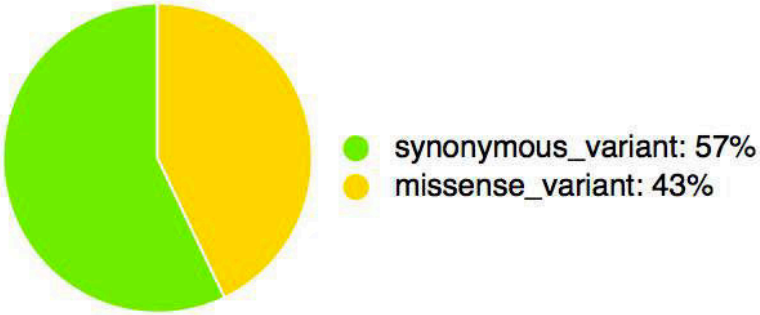
Summary statistics:

Category	Count
Variants processed	28
Variants remaining after filtering	28
Novel / existing variants	4 (14.3%) / 24 (85.7%)
Overlapped genes	74
Overlapped transcripts	74
Overlapped regulatory features	-

Consequences (all)



Coding consequences



FreeBayes

GitHub | This repository | Search | Explore | Features | Enterprise | Blog

ekg / freebayes

https://github.com/ekg/freebayes

Watch 30

Bayesian haplotype-based polymorphism discovery and genotyping. http://arxiv.org/abs/1207.3907

764 commits | 2 branches | 21 releases | 8 contributors

branch: master | freebayes / +

Update Makefile

AlistairNWard authored 4 days ago | latest commit c003c1e602

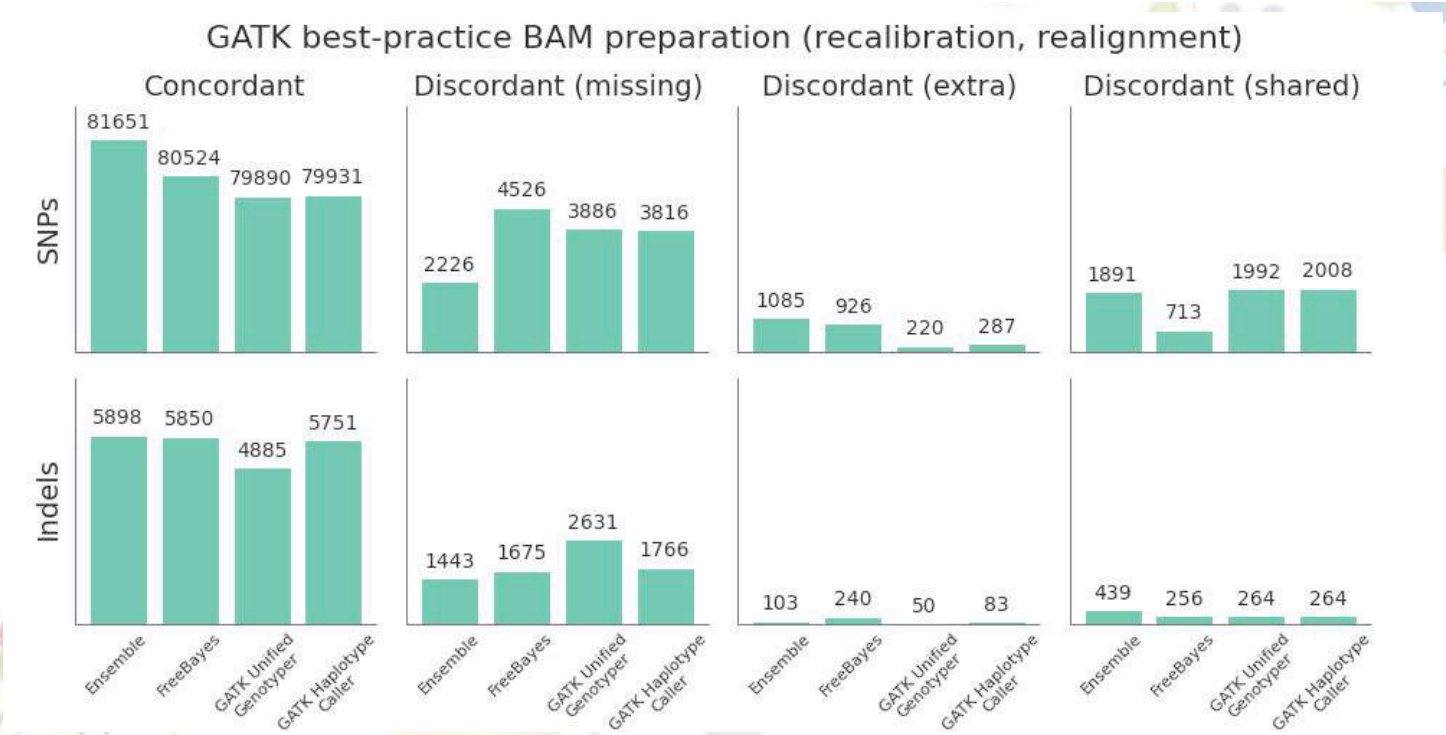
Best practices philosophy .....

- Indel realignment is accomplished internally
- Base quality recalibration is avoided
- Variant quality recalibration is avoided

Benchmarking and Validation

Blue Collar Bioinformatics

Updated comparison of variant detection methods: Ensemble, FreeBayes and minimal BAM preparation pipelines



http://bcb.io/2013/10/21/updated-comparison-of-variant-detection-methods-ensemble-freebayes-and-minimal-bam-preparation-pipelines/

For Sequencing and Bioinformatics Data Analysis Collaborations

Please contact: Allison Jackson

Allison.Jackson@glasgow.ac.uk

Glasgow Polyomics
College of Medical, Veterinary and Life Sciences
University of Glasgow

http://www.polyomics.gla.ac.uk/enquiry.php

Reach me:
Manikhandan.Mudaliar@glasgow.ac.uk
https://twitter.com/ManiMudaliar
https://uk.linkedin.com/in/mudaliar

University of Glasgow

Glasgow Polyomics
www.glasgow.ac.uk/polyomics