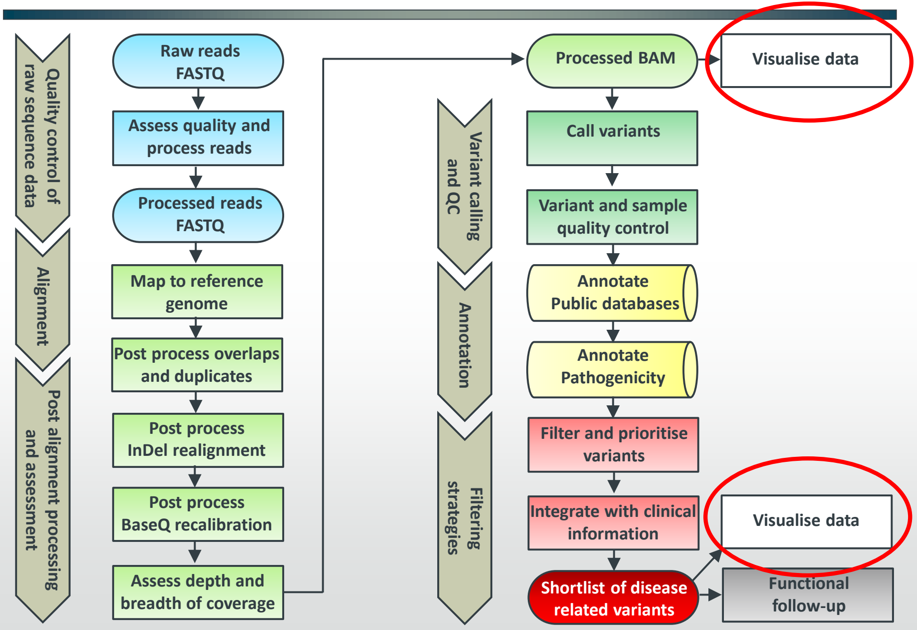
MEDI6215

Visualisation of genomic data

Dr Jane Gibson
J.Gibson@soton.ac.uk
6th February 2017

Analysis workflow



Quality control of raw sequence data

Alignment

Post alignment processing and assessment

Raw reads FASTQ

Assess quality and process reads

Processed reads FASTQ

Map to reference genome

Post process overlaps and duplicates

Post process InDel realignment

Post process BaseQ recalibration

Assess depth and breadth of coverage

Variant calling and QC

Annotation

Filtering strategies

Processed BAM

Visualise data

Call variants

Variant and sample quality control

Annotate Public databases

Annotate Pathogenicity

Filter and prioritise variants

Integrate with clinical information

Visualise data

Shortlist of disease related variants

Functional follow-up

Lecture Outline

- Why visualise?
- What can you visualise?
- Visualisation tools
- Focus on Integrative Genome Viewer (IGV)

Learning Objectives

By the end of this session you should…

- Understand the importance of data visualisation and different types of visualisations
- Be aware of several available tools to view alignment/variant data
- Be familiar with the layout of the IGV software, how to upload your data and navigate to regions of interest
- Gain the information you will need in order to complete the practicals

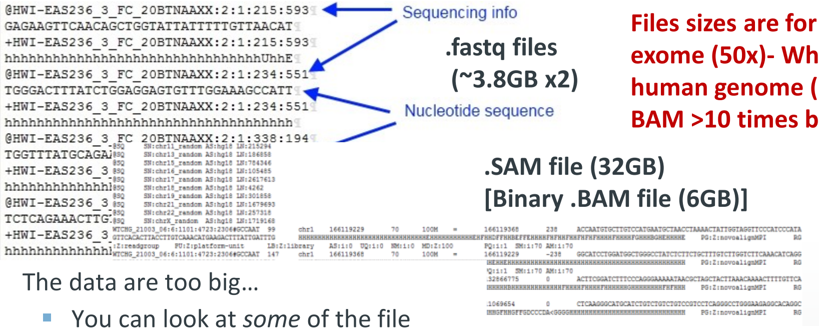
Why visualise?

Why visualise the data?

- From initial exploration of the data…
  - “I just want to have a *look* at it”
- …to publication of the data
  - “I need to *illustrate* a point”
- To check that an algorithm has dealt with the data as expected
- View our own data *in the context* of the information already known
- Just like in the public genome viewers showing the human reference genome eg. ensembl, UCSC
- Integrate annotation of various types, and visualise all relative to a particular genome ‘map’

‘A picture paints a thousand words’

Lets just look at the files?



```
@HWI-EAS236_3_FC_20BTNAAXX:2:1:215:593
GAGAAGTTCAACAGCTGGTATTATTTTTGTTAACAT
+HWI-EAS236_3_FC_20BTNAAXX:2:1:215:593
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhUhhE
@HWI-EAS236_3_FC_20BTNAAXX:2:1:234:551
TGGGACTTTATCTGGAGGAGTGTTTGGAAAGCCATT
+HWI-EAS236_3_FC_20BTNAAXX:2:1:234:551
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
@HWI-EAS236_3_FC_20BTNAAXX:2:1:338:194
TGGTTTATGCAGA
+HWI-EAS236_3
hhhhhhhhhhhhh
@HWI-EAS236_3
TCTCAGAAACTTG
+HWI-EAS236_3
hhhhhhhhhhhhh
```

.fastq files (~3.8GB x2)

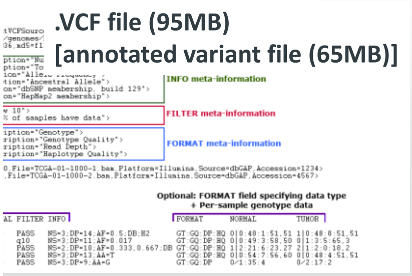
```
@SQ     SN:chr11_random AS:hg18 LN:215294
@SQ     SN:chr13_random AS:hg18 LN:186858
@SQ     SN:chr15_random AS:hg18 LN:784346
@SQ     SN:chr16_random AS:hg18 LN:105485
@SQ     SN:chr17_random AS:hg18 LN:2617613
@SQ     SN:chr18_random AS:hg18 LN:4262
@SQ     SN:chr19_random AS:hg18 LN:301858
@SQ     SN:chr21_random AS:hg18 LN:1679693
@SQ     SN:chr22_random AS:hg18 LN:257318
@SQ     SN:chrX_random  AS:hg18 LN:1719168
WTCHG_21003_06:6:1101:4723:2306#GCCAAT  99   chr1  166119229  70  100M  =  166119368  238   ACCAATGTGCTTGTCCATGAATGCTAACCTAAAACTATTGGTAGGTTCCCATCCCATA
GTTCACACTTACCTTGTCAAACATGAAGACTTTATTGATTTG  HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHEHHHHHHHHHHHHHEHFHHDFFHHBEFFEHHHHFHFHHFHHFHFHFHHHHFHHHHFGHHHBGHEHHHHE   PG:Z:novoalignMPI   RG
:Z:readgroup   PU:Z:platform-unit   LB:Z:library   AS:i:0   UQ:i:0   NM:i:0   MD:Z:100   PQ:i:1   SM:i:70  AM:i:70
WTCHG_21003_06:6:1101:4723:2306#GCCAAT  147  chr1  166119368  70  100M  =  166119229  -238  GGCATCCTGGATGGCTGGGCCTATCTCTTCTGCTTTGTCTTGGTCTTCAAACATCAGG
HEHHEHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH   PG:Z:novoalignMPI   RG
Q:i:1   SM:i:70 AM:i:70
32866775   0   ACTTCGGATCTTTCCCAGGGAAAAATAACGCTAGCTACTTAAACAAAACTTTTGTTCA
HHHHHHBHHHHHHHHHHHHHHHFHHHHFHHHHFHHHHHHGHHHHHHHHHFHFHHH   PG:Z:novoalignMPI   RG

1069654   0   CTCAAGGGCATGCATCTGTCTGTCTGTCCGTCCTCAGGGCCTGGGAAGAGGCACAGGC
HHGFHHGFFGDCCCDA<GGGGHHHHHHHHHHHHHHHHHHHHHHHHHHHHH   PG:Z:novoalignMPI   RG
```

Files sizes are for exome (50x)- Whole human genome (30x) BAM >10 times bigger

.SAM file (32GB)
[Binary .BAM file (6GB)]

- The data are too big…
  - You can look at *some* of the file
  - The files can be very wide and very long
- V. difficult to make sense of lists of numbers, letter, symbols and get the important information out –not intuitive
- Want to be able to go straight to data for a specific region of interest



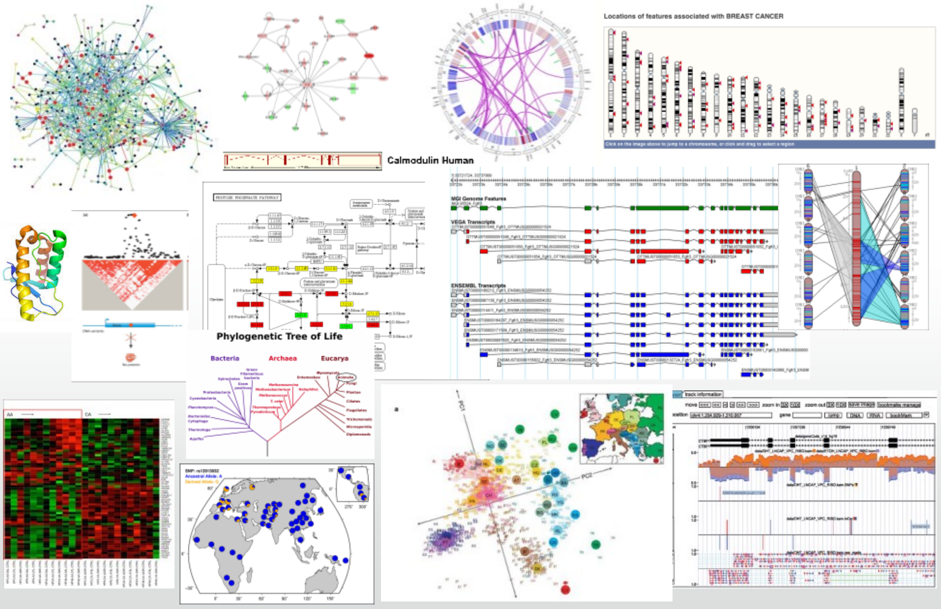
.VCF file (95MB)
[annotated variant file (65MB)]

```
tVCFSourc
/genomes/
36,md5=f1
ption="Nu
ption="To
ion="Alle
tion="Ancestral Allele">
on="dbSNP membership, build 129">
on="HapMap2 membership">
w 10">
% of samples have data">
iption="Genotype">
ription="Genotype Quality">
ription="Read Depth">
ription="Haplotype Quality">
0,File=TCGA-01-1000-1.bam,Platform=Illumina,Source=dbGAP,Accession=1234>
,File=TCGA-01-1000-2.bam,Platform=Illumina,Source=dbGAP,Accession=4567>

AL FILTER INFO                              FORMAT       NORMAL              TUMOR
   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ  0|0:48:1:51,51      1|0:48:8:51,51
   q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ  0|0:49:3:58,50      0|1:3:5:65,3
   PASS   NS=2;DP=10;AF=0.333,0.667;DB      GT:GQ:DP:HQ  1|2:21:6:23,27      2|1:2:0:18,2
   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ  0|0:54:7:56,60      0|0:48:4:51,51
   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP     0/1:35:4            0/2:17:2
```

INFO meta-information
FILTER meta-information
FORMAT meta-information
Optional: FORMAT field specifying data type
+ Per-sample genotype data

Types of biological visualisation



Locations of features associated with BREAST CANCER
Calmodulin Human
Phylogenetic Tree of Life
Bacteria
Archaea
Eucarya
PC2

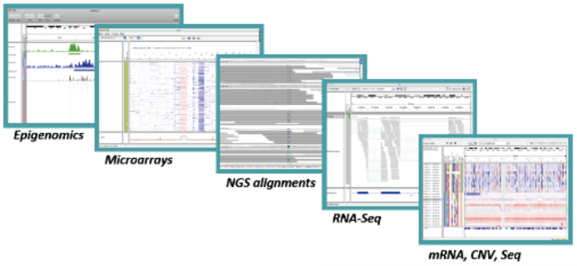
Genome Browsers

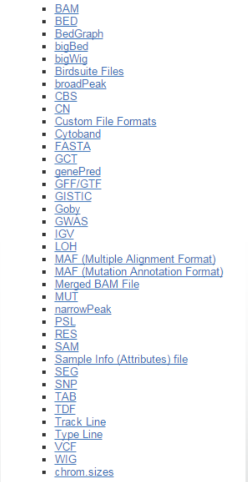
- *"A graphical interface for display of information from a biological database of genomic data"*
- Early days of the Human Genome Project (1999-2001) a need to visualise of genomic data, as a way of making it available to the research community in a more user friendly format than a database
- Public genome viewers showing the human reference genome
  - eg. ensembl, UCSC Genome Browser, (*NCBI mapviewer/Genome data viewer)*
- Many more specific to particular genomes (plants, animals…) or particular projects/datasets 1000 genomes, COSMIC, ENCODE project
- Interactive browsers, move around and investigate data
- Linear representation of genome ~3bn bp, left to right, one strand, lots of tracks below all lined up
- Customisable view; squishable, expandable, add or remove tracks, add your own tracks

Introduced to ensembl browser in Omic Techniques module (uploaded bam and vcf files to ensembl)

What kind of data can we visualise?

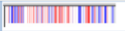
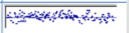
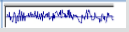
Input files





- BAM
- BED
- BedGraph
- bigBed
- bigWig
- Birdsuite Files
- broadPeak
- CBS
- CN
- Custom File Formats
- Cytoband
- FASTA
- GCT
- genePred
- GFF/GTF
- GISTIC
- Goby
- GWAS
- IGV
- LOH
- MAF (Multiple Alignment Format)
- MAF (Mutation Annotation Format)
- Merged BAM File
- MUT
- narrowPeak
- PSL
- RES
- SAM
- Sample Info (Attributes) file
- SEG
- SNP
- TAB
- TDF
- Track Line
- Type Line
- VCF
- WIG
- chrom.sizes

- NGS data, we want to visualise the alignments (how good is the coverage?) and the variants (how good is the variant?)
- Alignments file = Bam (bai), Variants file = vcf (idx)
- Indexing:
  - Often files need to be indexed (IGVtools can do this) to speed up the navigating around
  - Can jump to specific region without reading in whole file from the beginning

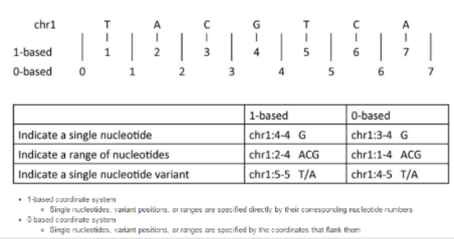
Heatmap	
Bar chart	
Scatter plot	
Line plot	

Reference genomes

- What's in a name?
- Everything is annotated with respect to a particular reference genome

TP53 isoform a Chr17	Start	End	length
hg18	7,512,445	7,531,588	19,144
hg19	7,571,720	7,590,868	19,149
hg38	7,668,402	7,687,550	19,149

- Genome builds
  - UCSC = hg18, hg19
  - NCBI = b36
  - Genome Reference Consortium (GRC) = GRCh37
  - Merged = hg38, GRCh38
- Other variations
  - Patches; GRCh37.p1 (regional fixes do not change coordinates)
  - Sorting; ordered lexicographically/ karyotypically
  - Chr1 Vs 1
  - Zero/1-based format for specifying locations



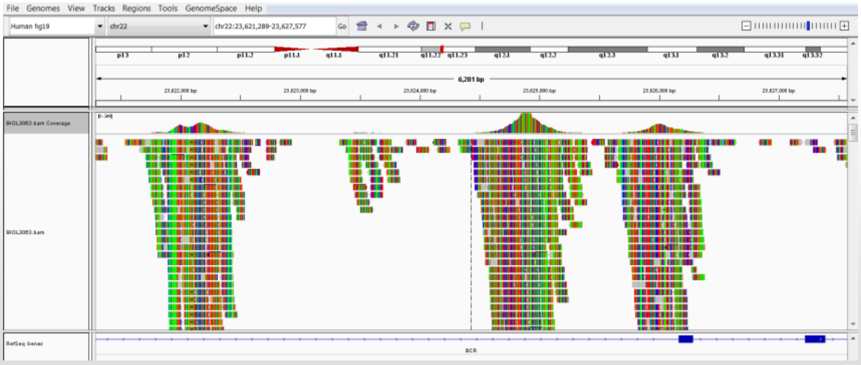
	1-based	0-based
Indicate a single nucleotide	chr1:4-4 G	chr1:3-4 G
Indicate a range of nucleotides	chr1:2-4 ACG	chr1:1-4 ACG
Indicate a single nucleotide variant	chr1:5-5 T/A	chr1:4-5 T/A

- 1-based coordinate system
  - Single nucleotides, variant positions, or ranges are specified directly by their corresponding nucleotide numbers
- 0-based coordinate system
  - Single nucleotides, variant positions, or ranges are specified by the coordinates that flank them

https://www.biostars.org/p/84686/
Tutorial: Cheat Sheet For One-Based Vs Zero-Based Coordinate Systems
Obi Griffith

Which genome to use?

- Generally use the closest major build, hg18, hg19, hg38
- Check the meta-information/header of your files (sam, vcf)
- Currently in transition between hg19 (2009) and hg38 (2013)
- Archive versions of databases/browsers are available with previous builds
- Conversion tools – eg. UCSC liftOver tool / Assembly Converter @ensembl



Human hg19

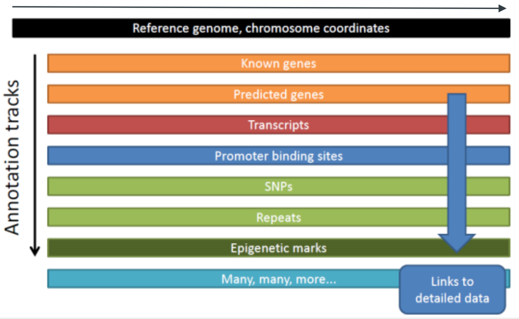
chr22

chr22:23,621,289-23,627,577

When you get it wrong – hg38 bam file viewed on hg19 genome

Tracks (annotation)

- Sequence – individual nucleotides
- Gene datasets (refGene, Ensembl)
- Databases of known variants
- Sequence conservation
- Repetitive sequence, CpGs
- Phenotype (Omim, COSMIC)
- Many many others…
- Compare samples (trio)
- In context of other peoples data
- In context of other types of data (omics)



Reference genome, chromosome coordinates

Annotation tracks

Known genes

Predicted genes

Transcripts

Promoter binding sites

SNPs

Repeats

Epigenetic marks

Many, many, more…

Links to detailed data

Jon K. Lærdahl, Structural Bioinformatics, Department of Informatics, University of Oslo

UNIVERSITY OF Southampton

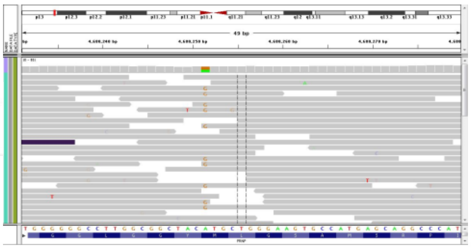
Data visualisation tools

Tools for visualisation

UNIVERSITY OF Southampton

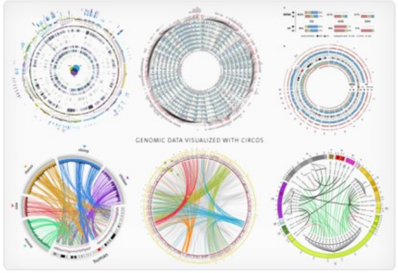
- **Integrative Genomics Viewer (IGV)**
  - Free
  - Download to PC
  - Windows, Java

Integrative Genomics Viewer



Robinson, et al. Integrative Genomics Viewer. Nature Biotechnology 29, 24–26 (2011)
Thorvaldsdóttir,et al Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics 14, 178-192 (2013).

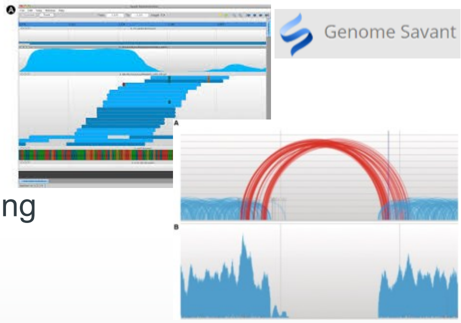
- **Circos**
  - (http://circos.ca/)
  - Free
  - Download to PC
  - Perl, command-line but possible on windows



Krzywinski, M. *et al.* Circos: an Information Aesthetic for Comparative Genomics. Genome Res (2009) 19:1639-1645

Tools for visualisation

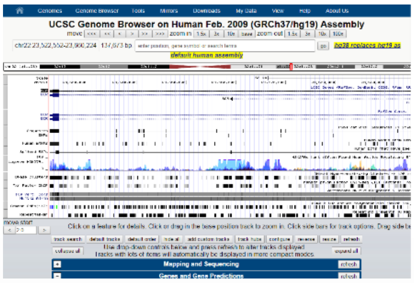
- **Genome Savant**
  - Free
  - Download to PC
  - Windows, Java
  - Arc view good for visualising structural variants



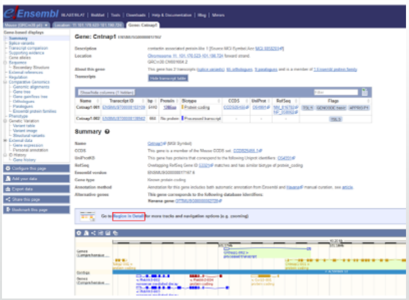
- **Trackster (Circster) in Galaxy**
  - Free
  - Available via the galaxy server
  - Also available UCSC, ensembl, IGV, IGB

Tools for visualisation

- **UCSC genome browser**
  - Free
  - Web-based
  - Upload your own data track
  - Vast amount of data available through UCSC


UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
Mapping and Sequencing
Genes and Gene Predictions

- **Ensembl**
  - Free
  - Web-based
  - Upload your own data track
  - Vast amounts of data available through ensembl



Which tool to use?

- Pros and Cons for each
    - Download loads of annotation from UCSC to view in IGV

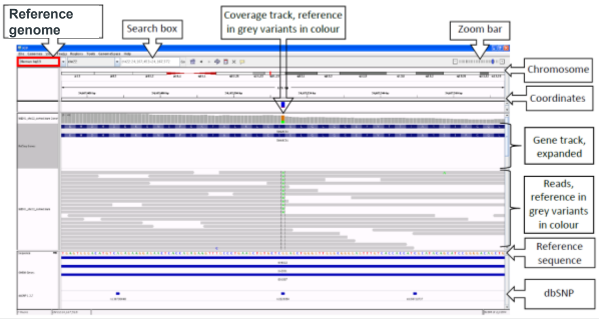
  Or
    - Upload loads of your own files to UCSC?
- Share with colleagues? Available online? Private data?
- How powerful is your PC? Got lots of RAM?
- Easiest to integrate with the external information you want?
- Which do you find easiest?

IGV

Getting started with IGV

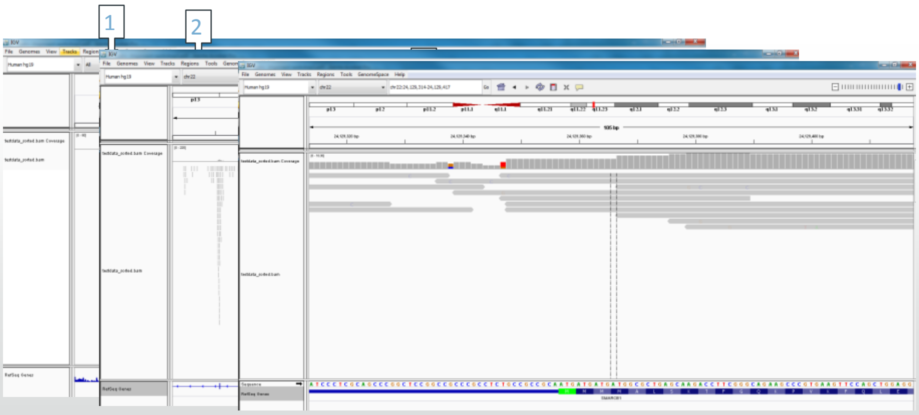
- Launching IGV
- Selecting relevant genome
  - Choose ‘more…’ if not present in drop down menu
- Loading your data (alignments)
  - Load from file…
- The index file must have the same file name and must reside in the same directory as the bam file
  - mysample.bam (choose this one)
  - mysample.bam.bai

Figure 6. The Integrative Genome Viewer (IGV)



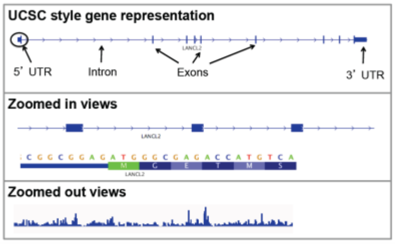
Navigating

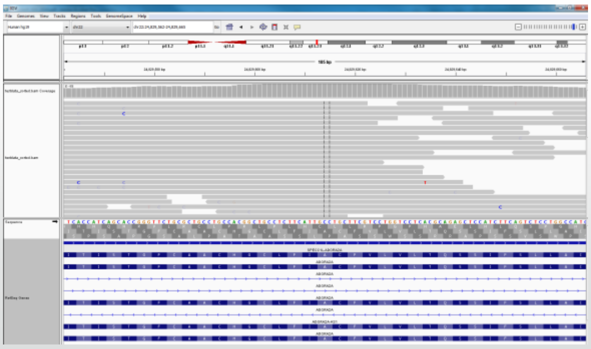
- You will not see your reads when you first load up your data
- Zoom in to your region of interest
  1. Select Chr from drop down menu
  2. Type in location (chr1 or Chr1:100,000-200,000) or locus (gene symbol)
  3. Highlight regions on genomic ruler
  4. Scroll left and right by dragging with the mouse (Home/End, left/right arrows)
- Different views at different resolutions (zooms)

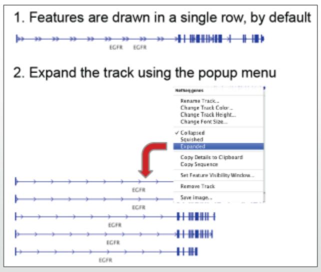


Genes and sequence

- Sequence
  - Click to get 3 frame translation
  - Click arrow to reverse the strand
- Genes
  - Multiple transcripts – click to expand
  - Positive/negative strand



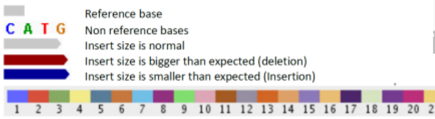



1. Features are drawn in a single row, by default

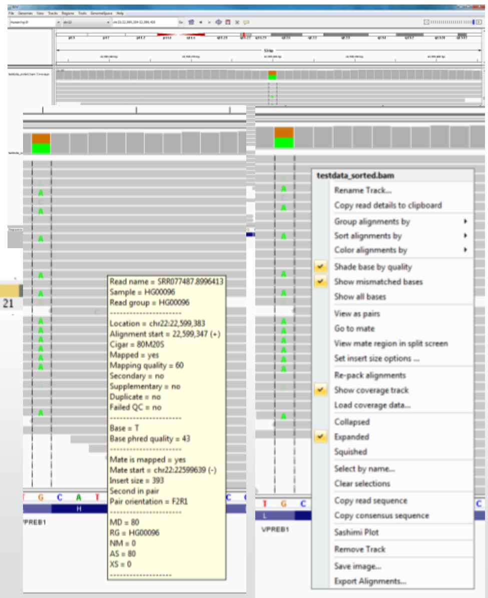
2. Expand the track using the popup menu

Viewing alignments

- Zoom in to see grey bars (arrow end shows direction)


Reference base
C A T G Non reference bases
Insert size is normal
Insert size is bigger than expected (deletion)
Insert size is smaller than expected (Insertion)

- Low quality bases are 'faded'
- Many of the thresholds can be altered in preferences
- Hover for more information, right click to get more viewing options


Read name = SRR077487.8996413
Sample = HG00096
Read group = HG00096

Location = chr22:22,599,383
Alignment start = 22,599,347 (+)
Cigar = 80M20S
Mapped = yes
Mapping quality = 60
Secondary = no
Supplementary = no
Duplicate = no
Failed QC = no

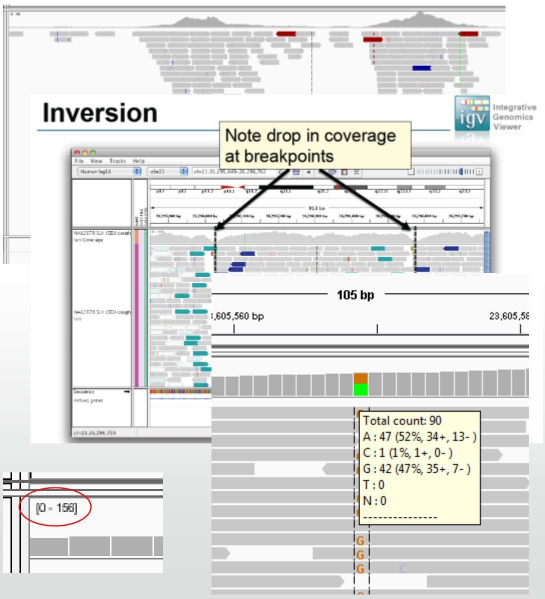
Base = T
Base phred quality = 43

Mate is mapped = yes
Mate start = chr22:22599639 (-)
Insert size = 393
Second in pair
Pair orientation = F2R1

MD = 80
RG = HG00096
NM = 0
AS = 80
XS = 0

Coverage track

- Histogram across the top
- Coverage profile different for Exome/genome
- Does my gene have good coverage/depth?
- Evidence for depth changes associated with structural changes?
- Coloured to reflect alleles – hover for more information
- **Note:** scale changes to maximise view



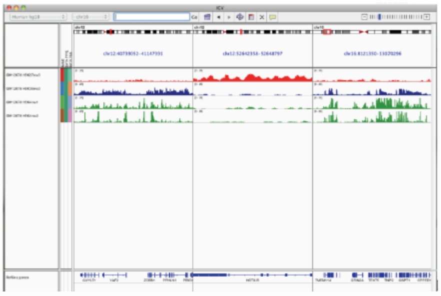
Inversion

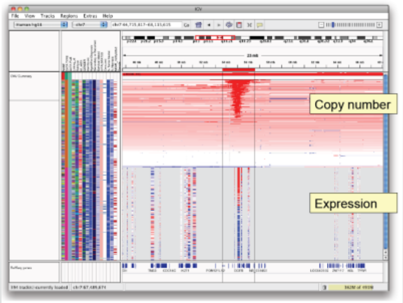
Note drop in coverage at breakpoints

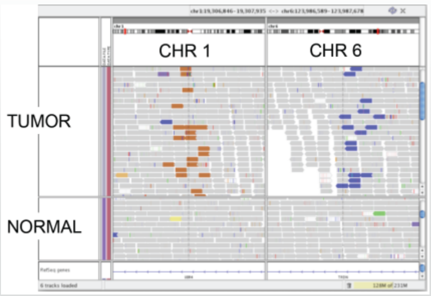
```
Total count: 90
A : 47 (52%, 34+, 13- )
C : 1 (1%, 1+, 0- )
G : 42 (47%, 35+, 7- )
T : 0
N : 0
---------------
```

More complex visualisation

- View sample attributes
- View multiple samples/regions/tissues

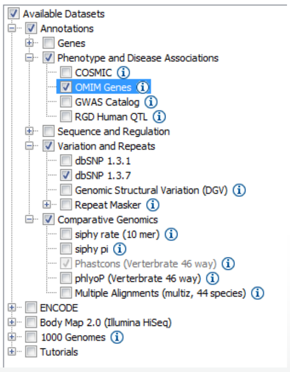




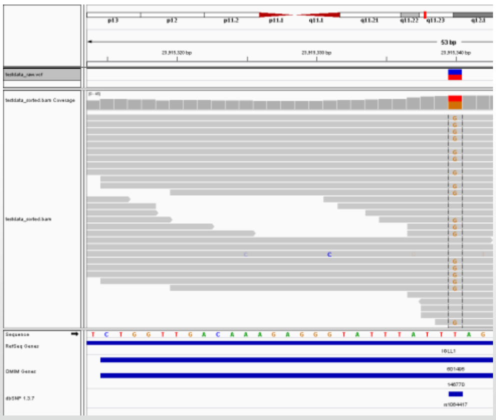


Adding and moving tracks

- Add your own called variants (.vcf)
- Several datasets available (Load from server…)
- Or download your own



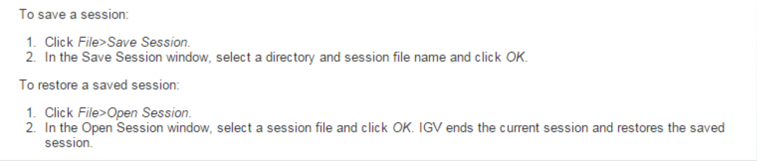
- ☑ Available Datasets
  - ☑ Annotations
    - □ Genes
    - ☑ Phenotype and Disease Associations
      - □ COSMIC ⓘ
      - ☑ OMIM Genes ⓘ
      - □ GWAS Catalog ⓘ
      - □ RGD Human QTL ⓘ
    - □ Sequence and Regulation
    - ☑ Variation and Repeats
      - □ dbSNP 1.3.1
      - ☑ dbSNP 1.3.7
      - □ Genomic Structural Variation (DGV) ⓘ
      - □ Repeat Masker ⓘ
    - ☑ Comparative Genomics
      - □ siphy rate (10 mer) ⓘ
      - □ siphy pi ⓘ
      - ☑ Phastcons (Verterbrate 46 way) ⓘ
      - □ phlyoP (Verterbrate 46 way) ⓘ
      - □ Multiple Alignments (multiz, 44 species) ⓘ
  - □ ENCODE
  - □ Body Map 2.0 (Illumina HiSeq)
  - □ 1000 Genomes ⓘ
  - □ Tutorials



- Eg. OMIM genes, or dbSNP 1.3.7
- Can move tracks around to make comparison easier eg. Move dbsnp track to next to vcf track

Saving your work

- You can save the entire session (.xml), when you restore it, it will be exactly as you left it
  - Saves all your viewing preferences, tracks, location, level of zoom etc.
- Can also save image (.png) of your current view, export visible alignments (.sam), or copy the consensus sequence
- chr22:23915314-23915365
  TCTGGTTGACAAAGAGGGTATTTATTKAGGGTTTACTGGGTACAGGGAGAAG



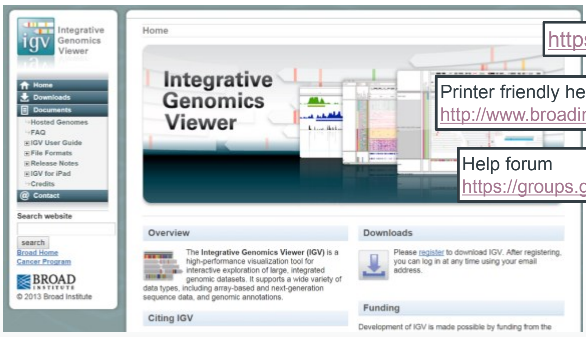
To save a session:

1. Click *File>Save Session*.
2. In the Save Session window, select a directory and session file name and click *OK*.

To restore a saved session:

1. Click *File>Open Session*.
2. In the Open Session window, select a session file and click *OK*. IGV ends the current session and restores the saved session.

More info/help



The **Integrative Genomics Viewer (IGV)** is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations.

Please register to download IGV. After registering, you can log in at any time using your email address.

Development of IGV is made possible by funding from the

https://www.broadinstitute.org/igv/

Printer friendly help guide
http://www.broadinstitute.org/igv/book/export/html/6

Help forum
https://groups.google.com/forum/#!forum/igv-help

Video tutorials; eg.
https://www.youtube.com/watch?v=IILfC3Uc6Vo

Introduction to the Integrative Genomics Viewer (IGV)
Jim Robinson & Helga Thorvaldsdóttir,Broad Institute of MIT and Harvard
BroadE Workshop, June 2014
https://drive.google.com/file/d/0B2dK2q40HDWeMkw4Q2RoQzd4TkU/edit?usp=sharing

Next...

Practical 1: Introduction to Galaxy