GMS6850 – Core Concepts in Bioinformatics
Lecture 3
-- NGS short-read mapping and variant calling as a case study in bioinformatics for low-level data analysis
-- Short-read-alignment viewer IGV
-- Genome variation and variant callers
-- Genome organization at the sequence level and genome browsers

2021 01 18 and 20

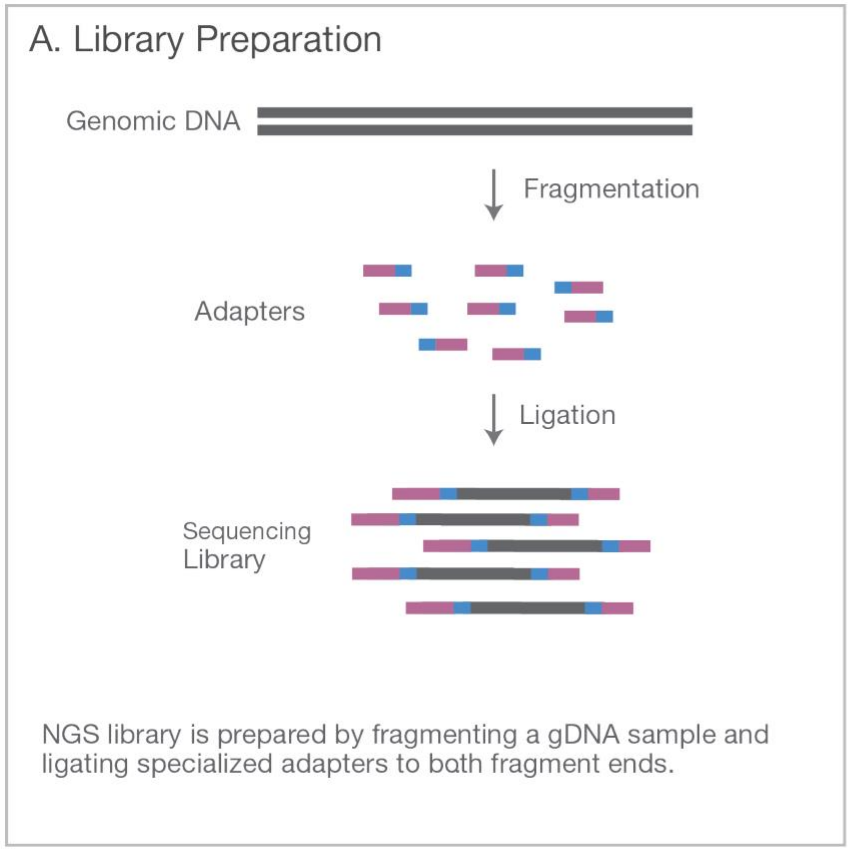
Steve Rozen

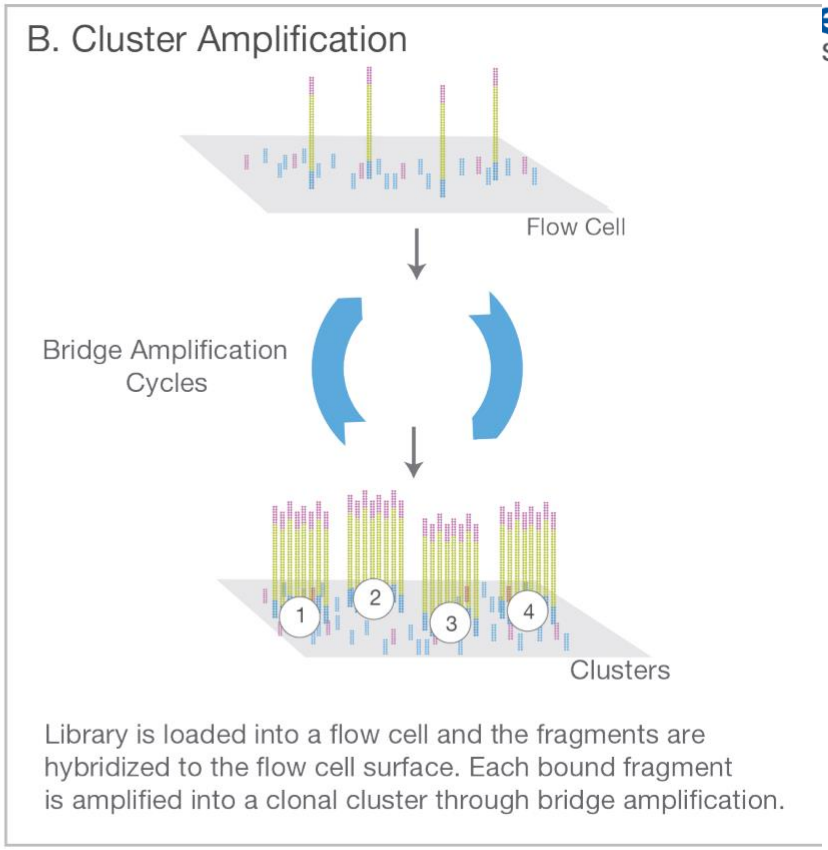
Last Wednesday…

- Translation
- Sequence alignment
  - Global, local, scoring matrices, dot plots, multiple alignment, RNA secondary structure
- Did not really start with short read mapping
- Any follow up questions on last class?

NGS (next generation sequencing) generates short reads, and usually the short reads need to be mapped (aligned) to a reference sequence. For genome sequencing, this is a reference genome.

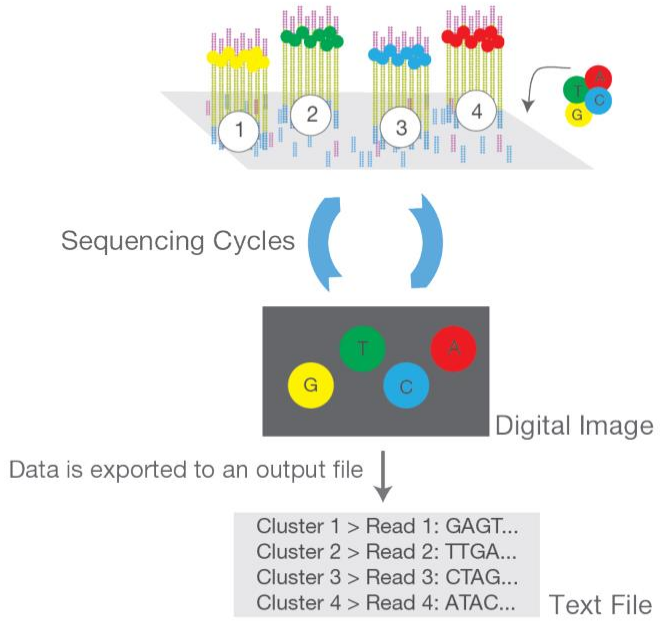
Why does the technology generate short reads?





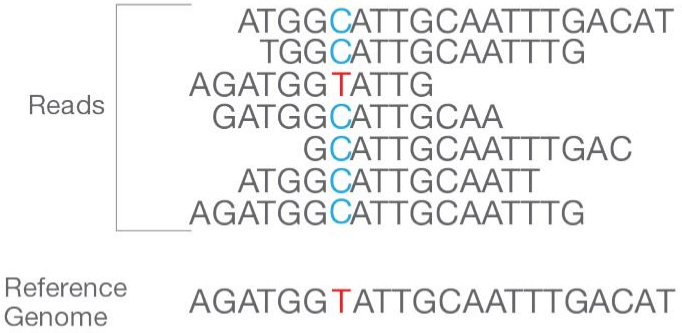
https://sapac.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

C. Sequencing



Sequencing reagents, including fluorescently labeled nucleotides, are added and the first base is incorporated. The flow cell is imaged and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This cycle is repeated "n" times to create a read length of "n" bases.

D. Alignment and Data Anaylsis



Reads are aligned to a reference sequence with bioinformatics software. After alignment, differences between the reference genome and the newly sequenced reads can be identified.

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

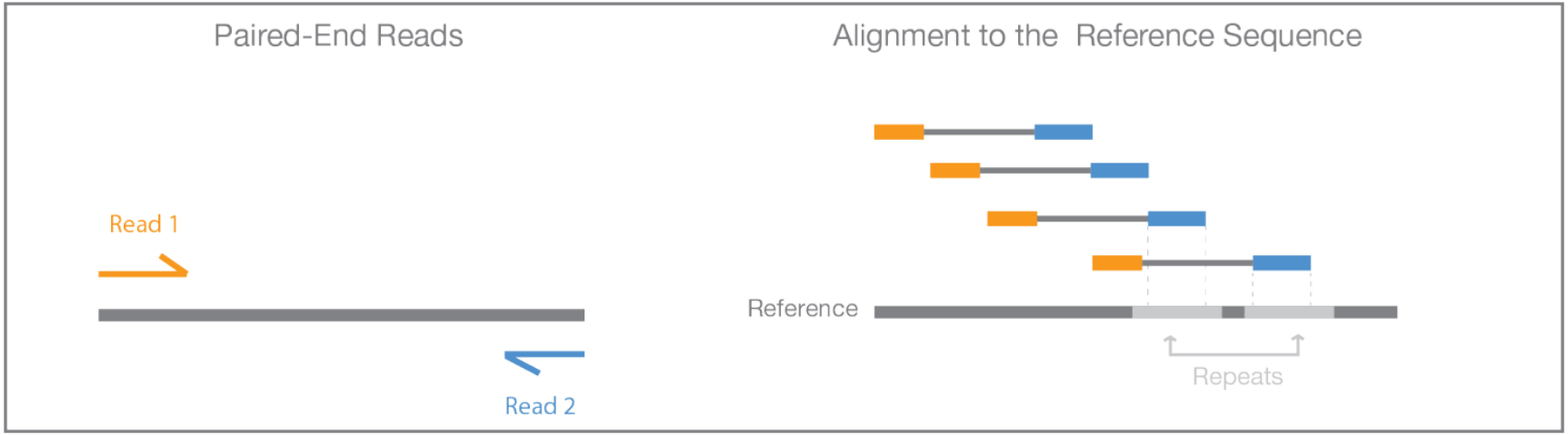


Figure 4: Paired-End Sequencing and Alignment—Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in better alignment of reads, especially across difficult-to-sequence, repetitive regions of the genome.

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

Gritty details for reference, do not need to know: https://seekdeep.brown.edu/illumina_paired_info.html

Flow cell

5 minutes of much more detail from Illumina

- https://youtu.be/fCd6B5HRaZ8
- What you need to know
  - DNA (or cDNA) is fragmented
  - You can optionally select fragments (e.g. just from exons) by hybridization or PCR amplification)
  - Various adapters and index sequences are added on to the ends of the fragments
  - Illumina generates reads from both ends (“paired reads”) or just one end
  - Sometimes, if the fragments are short, paired reads overlap, or reads capture adapter or index sequence at the far end of the fragments
  - Reading is done by incorporating fluorescently labelled nucleotides
  - For most applications you have to map (align) the short reads to a reference; this could be a genome or a transcriptome (= the set of all transcripts in an organism)

Review: Whole-genome Analysis Pipeline

Sequencing Provider

Illumina Hiseq X
30x coverage
150 bp paired-end

FASTQ file

FASTQC

FASTQ quality checker

FASTQ files (2 X 30 Gb / sample)

BWA-MEM

Short read aligner (used often for genomic data)

Sambamba

Miscellaneous post-processing

BAM files (~70 Gb / sample)

Qualimap

HLA typing

Virome profiling

Qualimap
http://qualimap.bioinfo.cipf.es/
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4708105/

Freebayes Joint Calling

Single Nucleotide Variant (SNV) caller

Parliament

Structural (SV) and Copy Number Variant (CNV) caller

Multisample VCF (18 Gb for 100 samples)

SNVs + Indels

SVs + CNVs

Annotation

Annotation

Annotation: is this an known variant?
Does it affect protein sequence?

SPECTRA DB

Database to store variant calls

Review: Whole-genome Analysis Pipeline

Sequencing Provider

30x coverage
150 bp paired-end

FASTQ file

FASTQC

FASTQ quality checker

FASTQ files (2 X 30 Gb / sample)

BWA-MEM

Sh… data)

BWA-MEM is the mapper / aligner (multithreaded, with some random choices, so results change slightly from run to run)

Original program (BWA since upgraded twice) https://academic.oup.com/bioinformatics/article/25/14/1754/225615

https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq/

Miscellaneous post-processing

Sambamba

BAM File

(~70 Gb / sample)

HLA…

Qualimap

Virom…

Qualimap
http://qualimap.bioinfo.cipf.es/
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4708105/

Freebayes Joint Calling

Parliament

Single Nucleotide Variant (SNV) caller

Multisample VCF (18 Gb for 100 samples)

Structural Number Variant (CNV) caller

SNVs + Indels

SVs + CNVs

Annotation

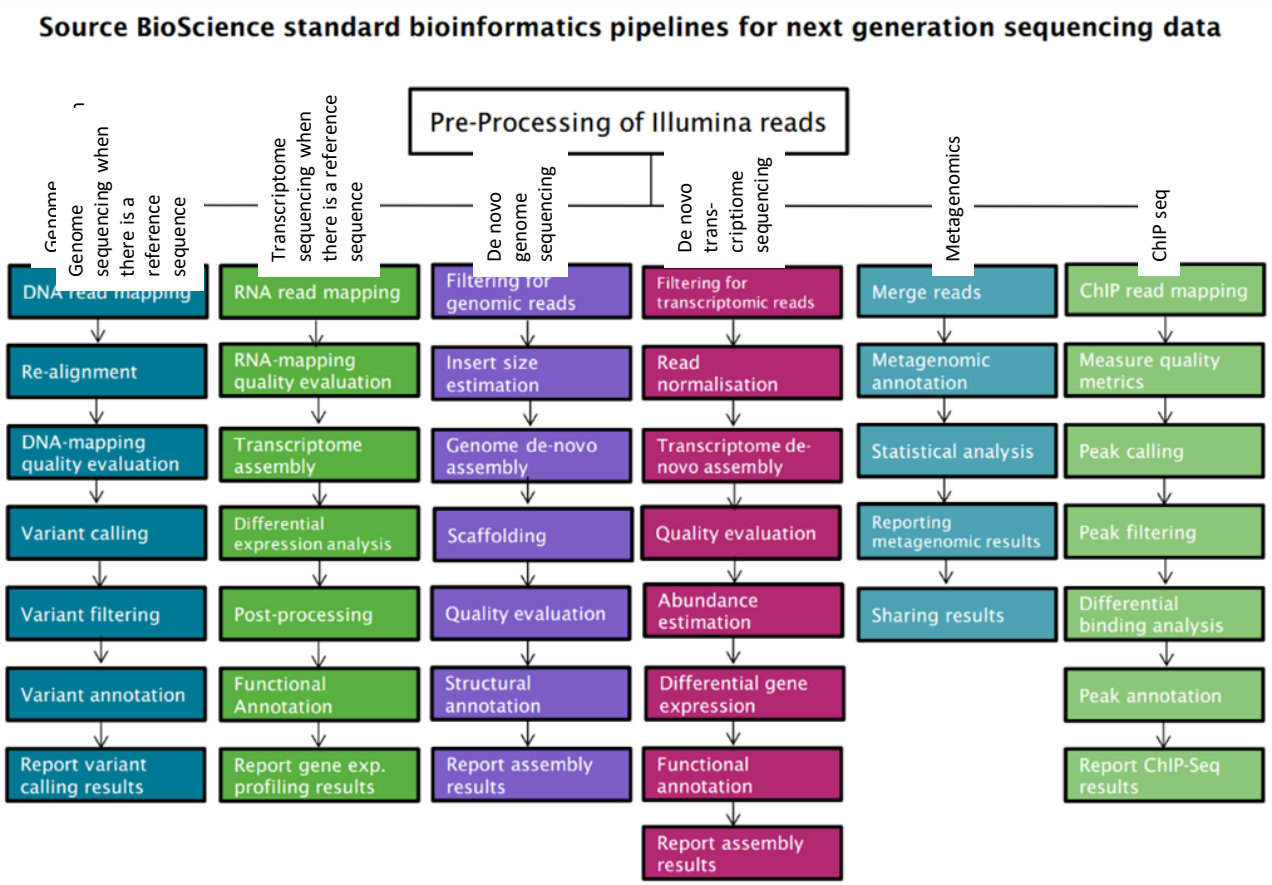
Annotation

Annotation: is this an known variant?
Does it affect protein sequence?

SPECTRA DB

Database to store variant calls

Different pipelines for different objectives



https://www.sourcebioscience.com/media/1714/adv_bioinf_pipeline-overview-5.pdf

Short read mappers

only a few in wide use

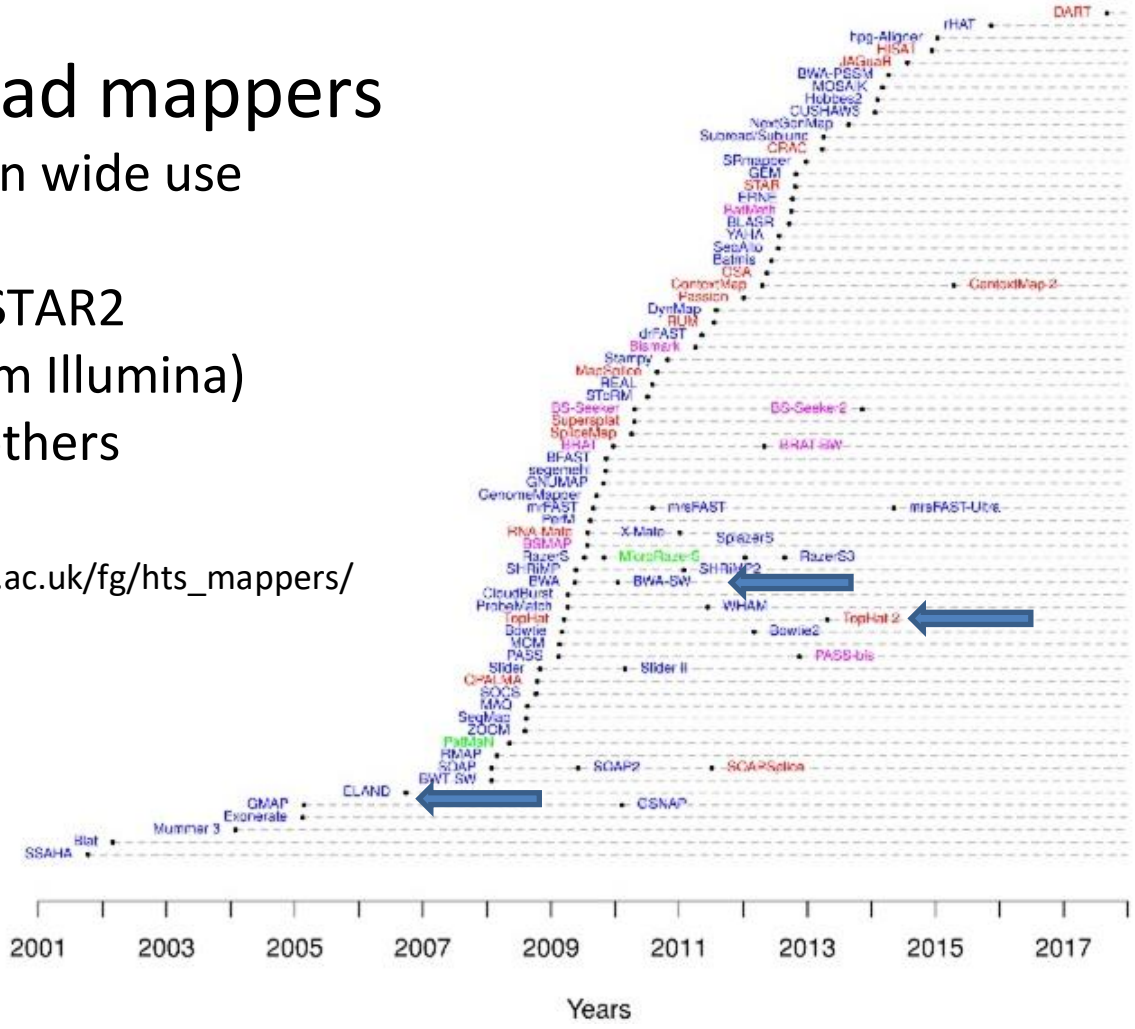
BWA-MEM

TOPHAT2, STAR2

ELAND (from Illumina)

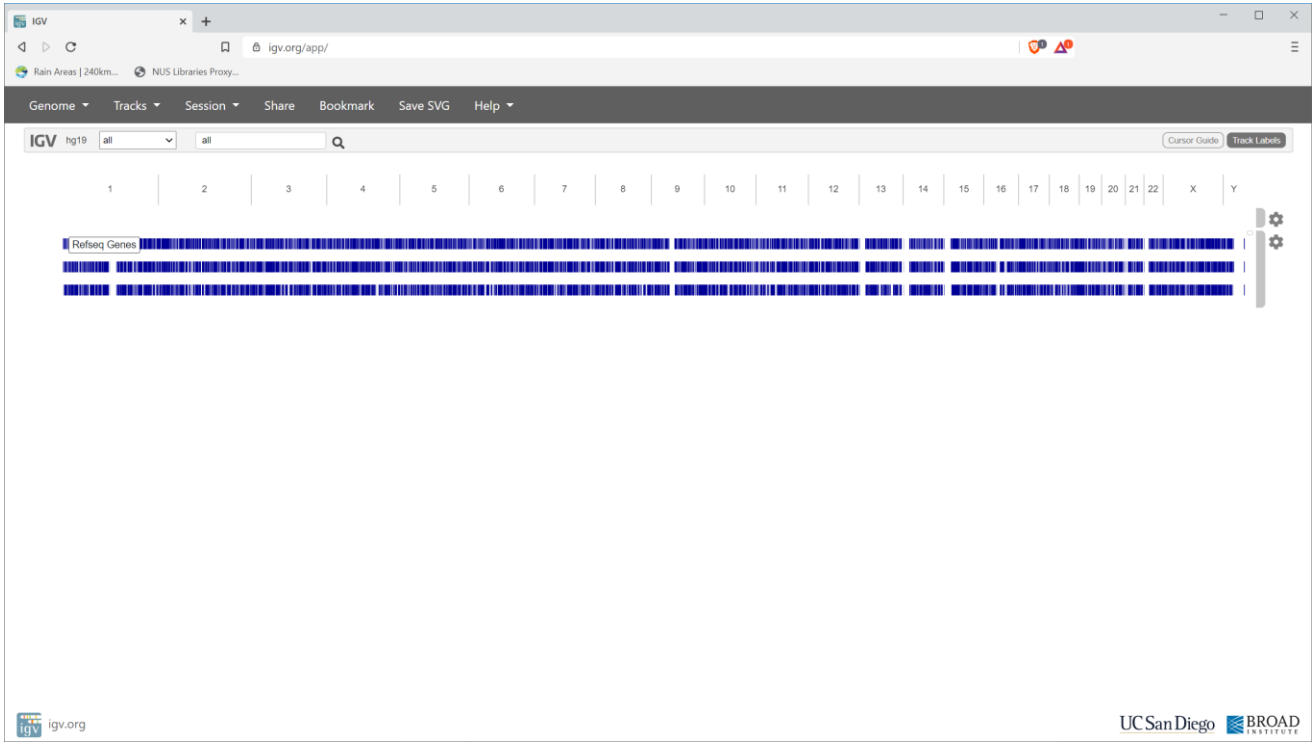
and a few others

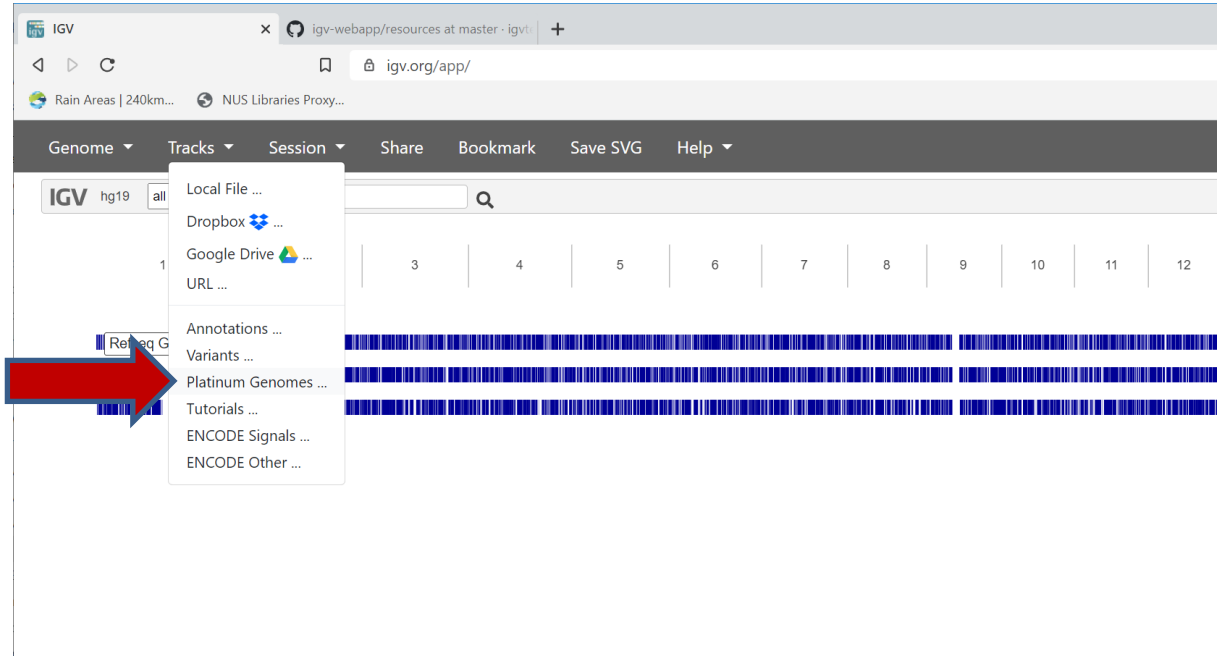
http://wwwdev.ebi.ac.uk/fg/hts_mappers/

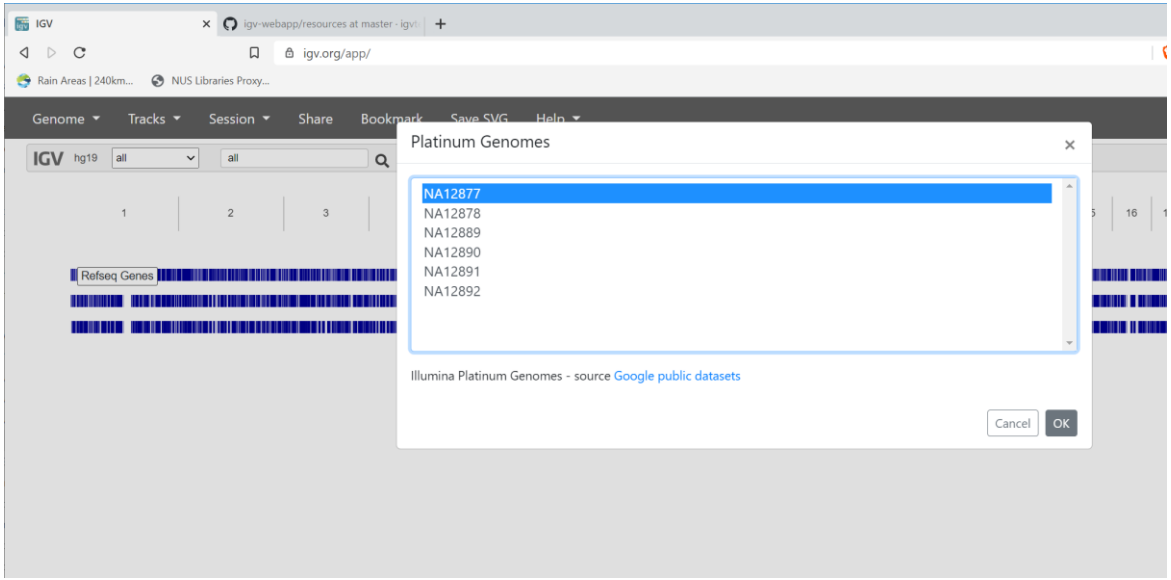


IGV (genome assembly browser)

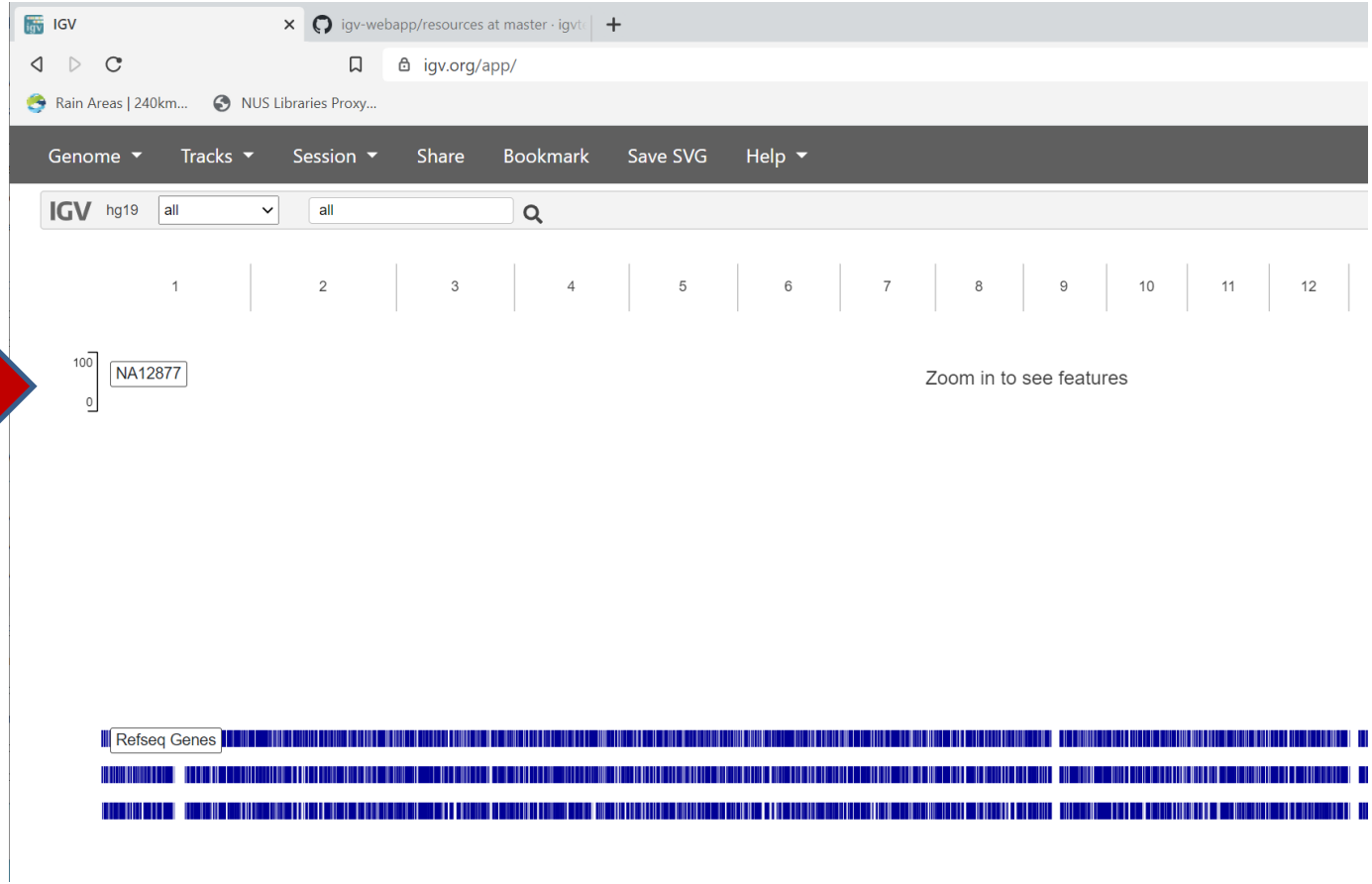
- https://igv.org/app

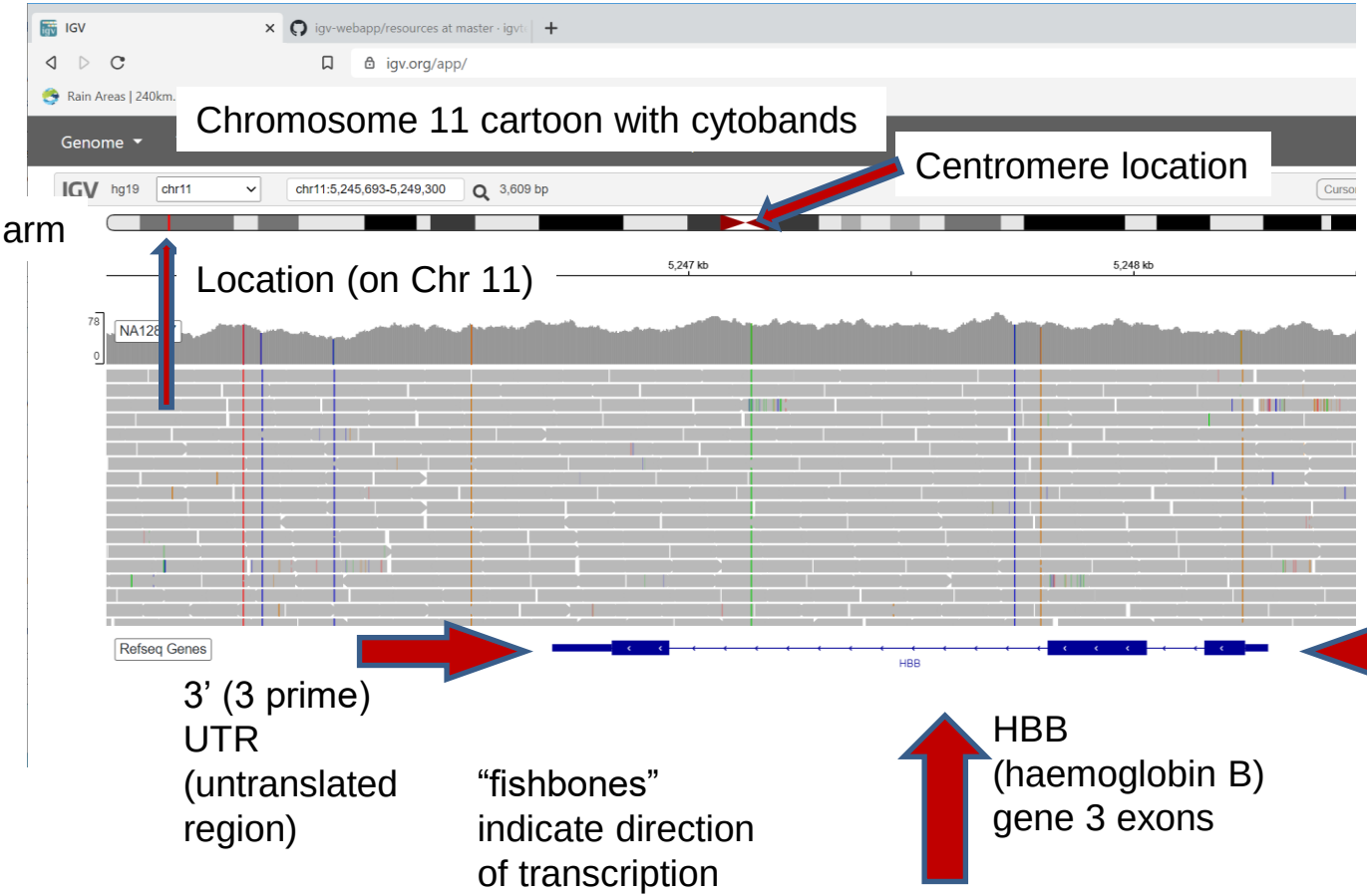


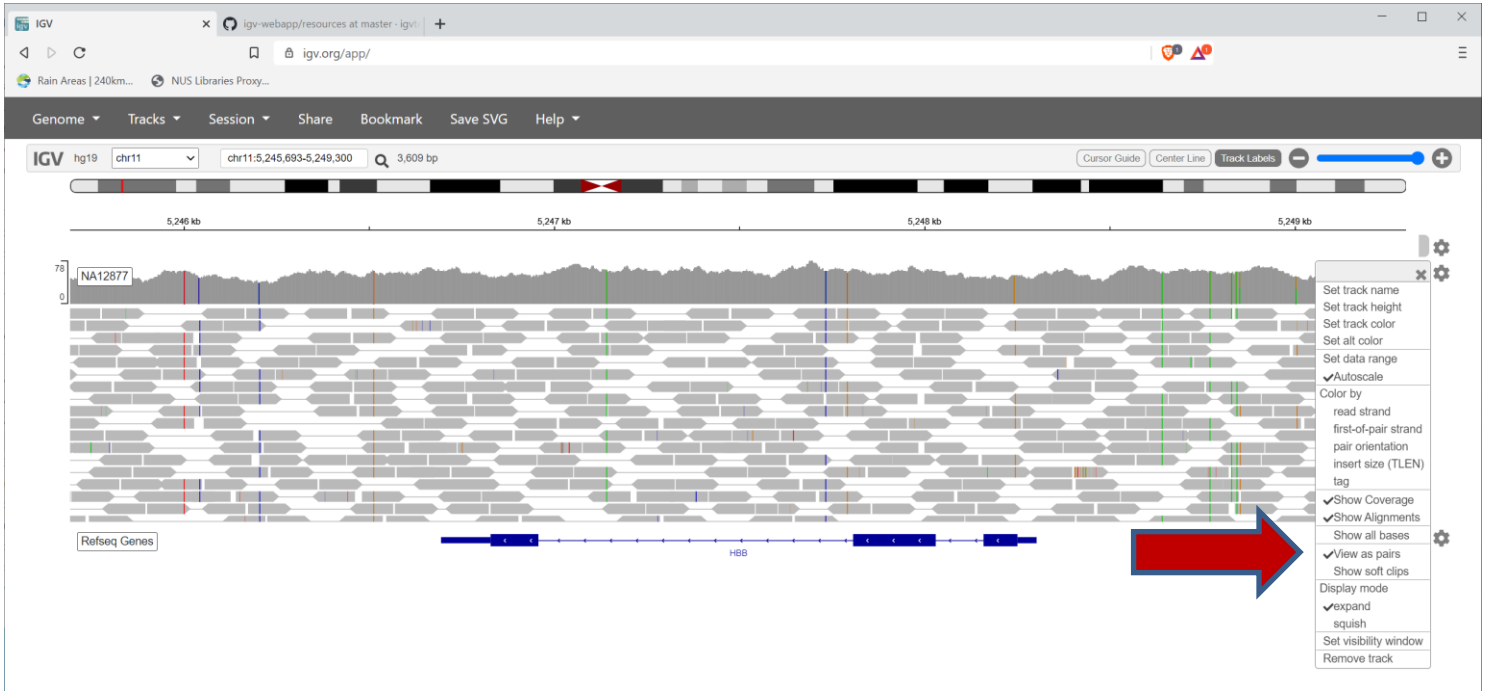

IGV
igv-webapp/resources at master · igv
igv.org/app/
Rain Areas | 240km...
NUS Libraries Proxy...
Genome
Tracks
Session
Share
Bookmark
Save SVG
Help
IGV
hg19
all
Local File ...
Dropbox ...
Google Drive ...
URL ...
Annotations ...
Variants ...
Platinum Genomes ...
Tutorials ...
ENCODE Signals ...
ENCODE Other ...
1
3
4
5
6
7
8
9
10
11
12


IGV
igv-webapp/resources at master - igv
igv.org/app/
Rain Areas | 240km...
NUS Libraries Proxy...
Genome
Tracks
Session
Share
Bookmark
Save SVG
Help
IGV
hg19
all
all
1
2
3
16
Refseq Genes
Platinum Genomes
NA12877
NA12878
NA12889
NA12890
NA12891
NA12892
Illumina Platinum Genomes - source Google public datasets
Cancel
OK

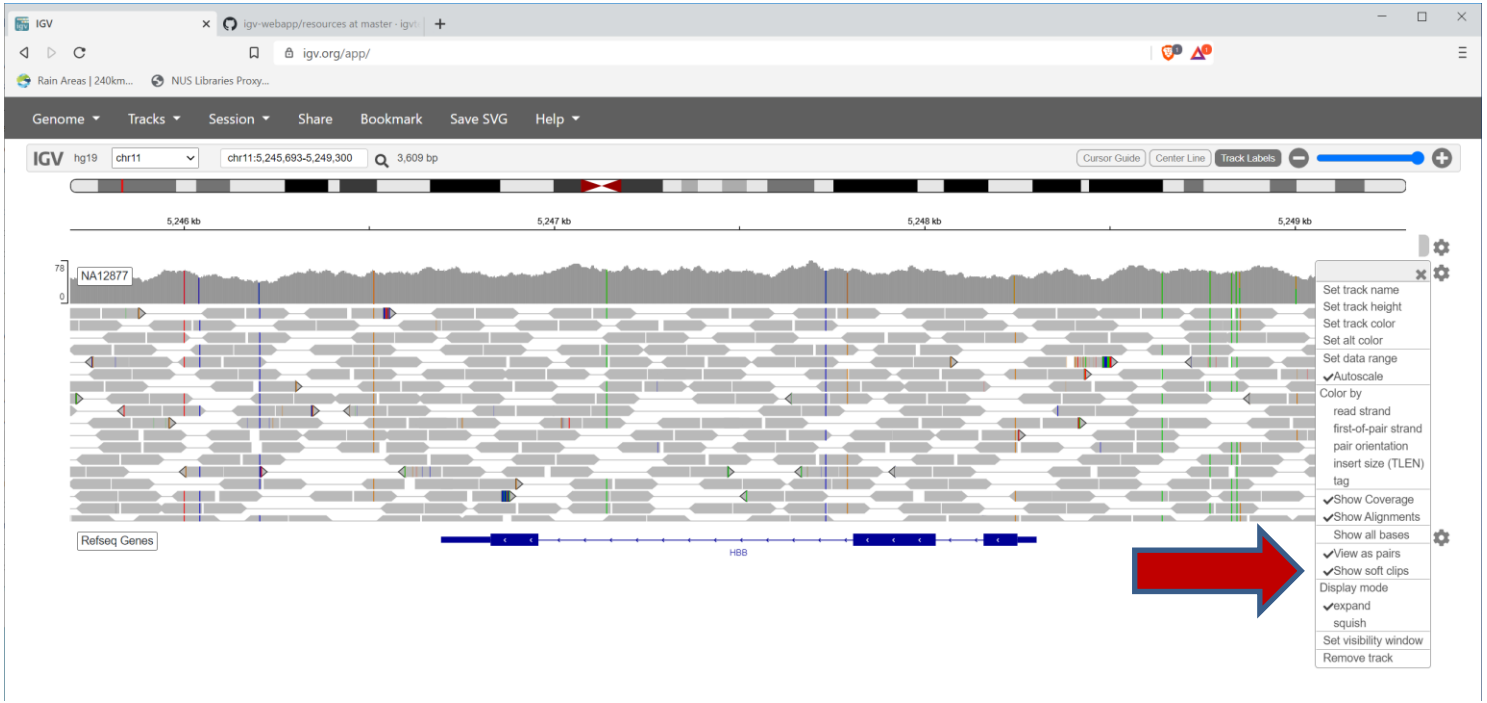
Sample ID
NA12877

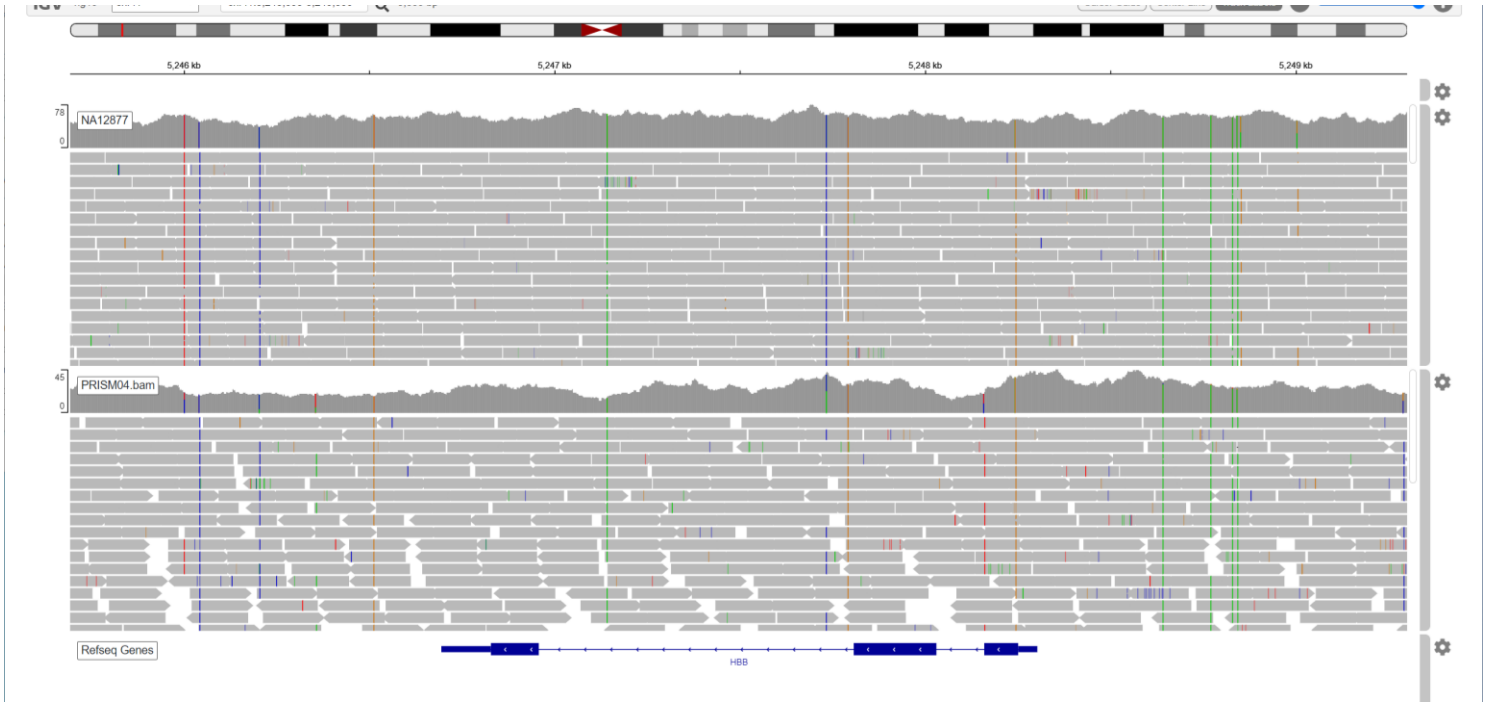


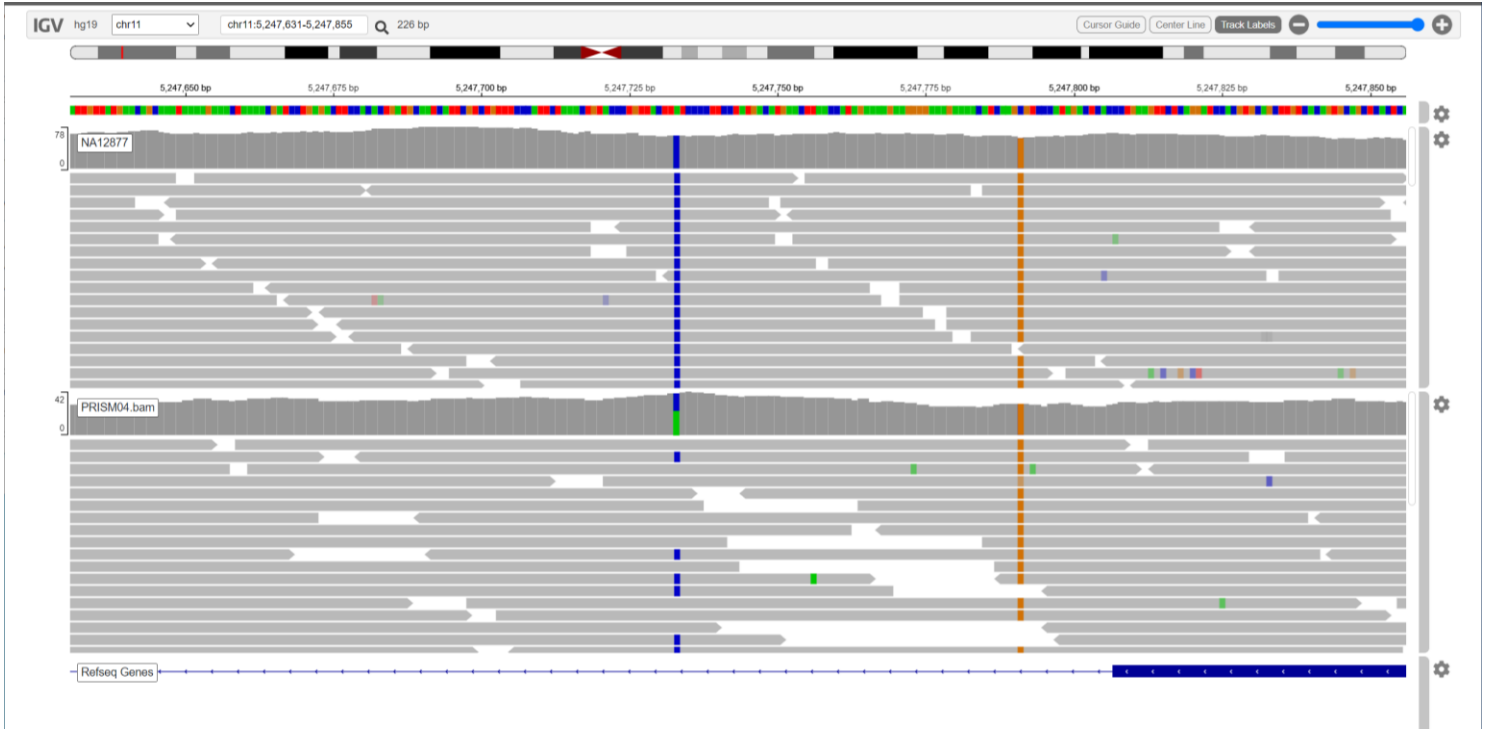

IGV
igv-webapp/resources at master · igv
igv.org/app/
Rain Areas | 240km.
Genome
Chromosome 11 cartoon with cytobands
Centromere location
IGV
hg19
chr11
chr11:5,245,693-5,249,300
3,609 bp
p arm
q arm
Location (on Chr 11)
5,247 kb
5,248 kb
78
0
Read depth plot
Reads
Refseq Genes
HBB
5’ (5 prime) UTR (untranslated region)
3’ (3 prime) UTR (untranslated region)
“fishbones” indicate direction of transcription
HBB (haemoglobin B) gene 3 exons

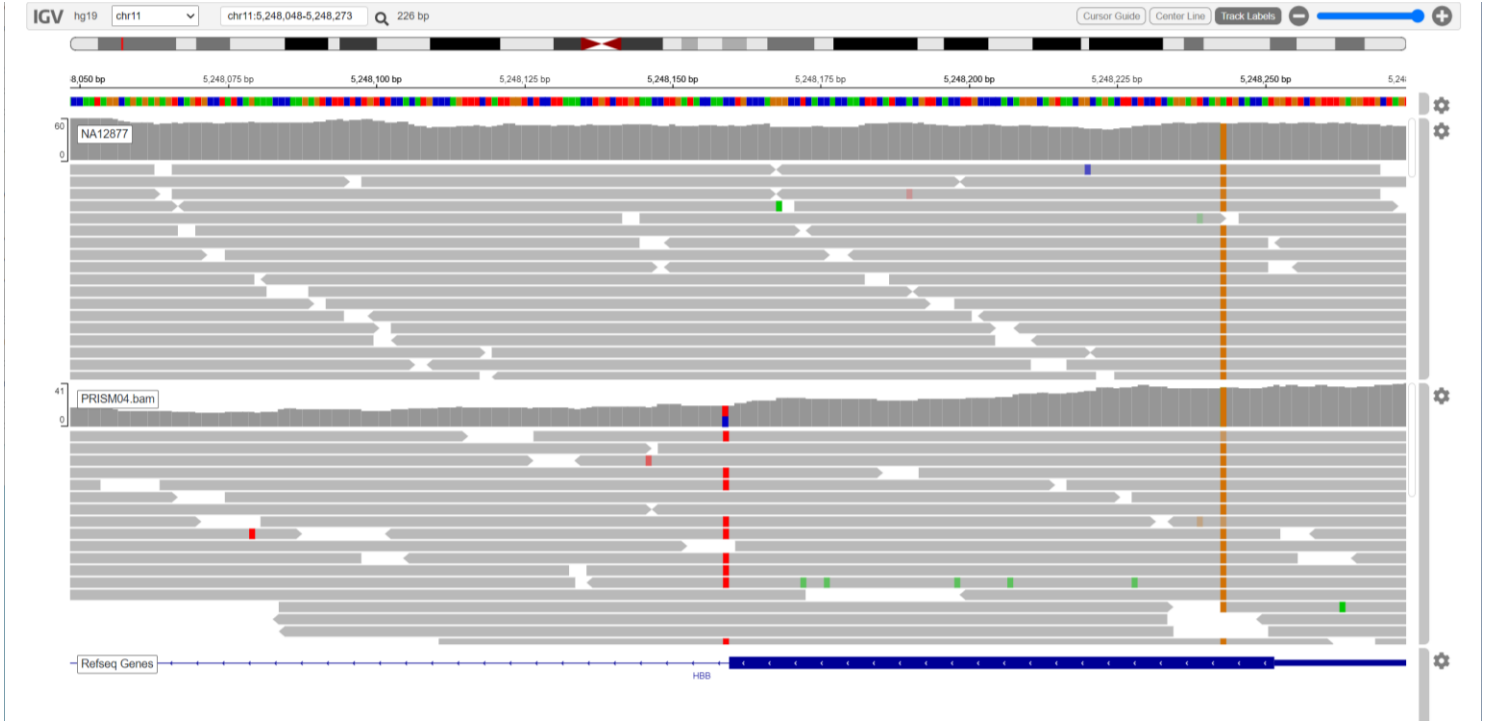


DukeNUS Medical School

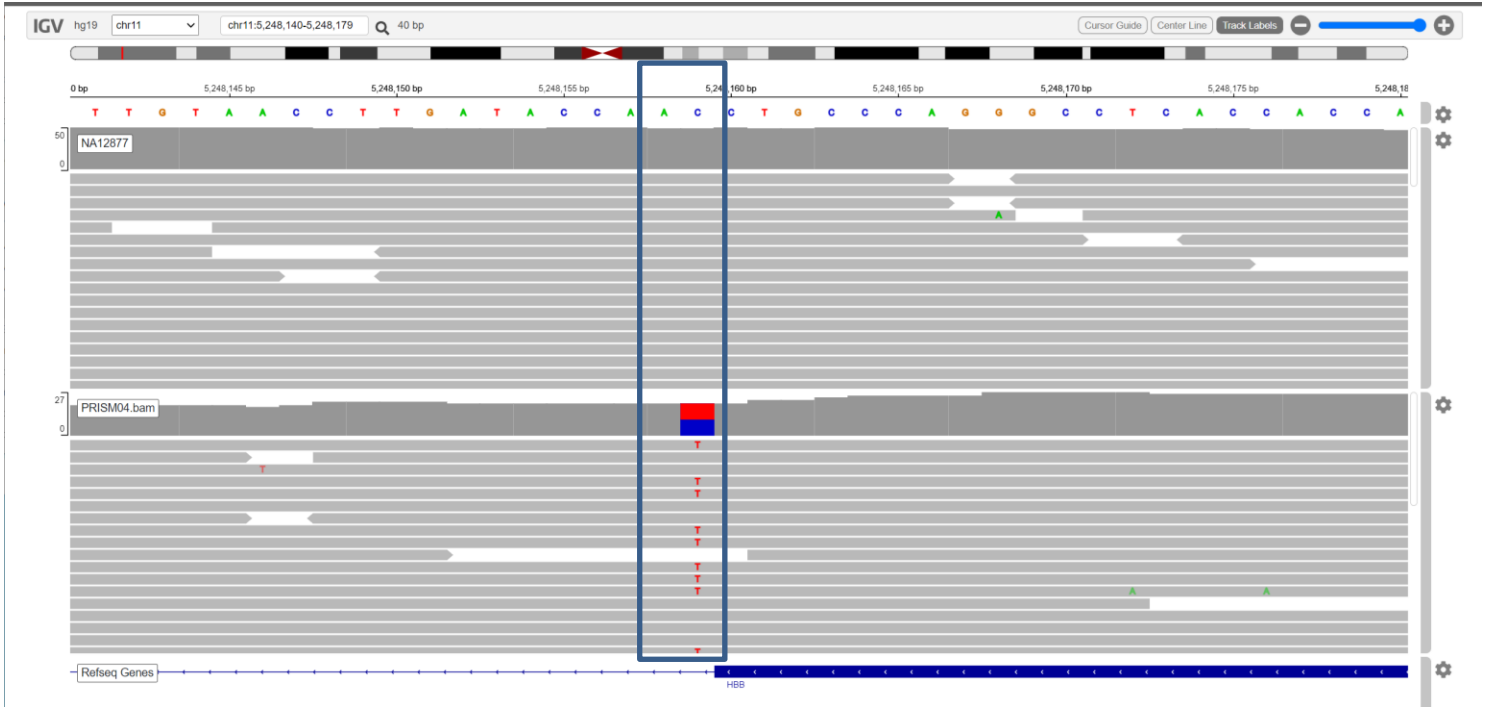

IGV
igv-webapp/resources at master · igv
igv.org/app/
Rain Areas | 240km...
NUS Libraries Proxy...
Genome
Tracks
Session
Share
Bookmark
Save SVG
Help
IGV
hg19
chr11
chr11:5,245,693-5,249,300
3,609 bp
Cursor Guide
Center Line
Track Labels
5,246 kb
5,247 kb
5,248 kb
5,249 kb
NA12877
Refseq Genes
HBB
Set track name
Set track height
Set track color
Set alt color
Set data range
Autoscale
Color by
read strand
first-of-pair strand
pair orientation
insert size (TLEN)
tag
Show Coverage
Show Alignments
Show all bases
View as pairs
Show soft clips
Display mode
expand
squish
Set visibility window
Remove track


5,246 kb
5,247 kb
5,248 kb
5,249 kb
78
0
NA12877
45
0
PRISM04.bam
Refseq Genes
HBB


IGV
hg19
chr11
chr11:5,247,631-5,247,855
226 bp
Cursor Guide
Center Line
Track Labels
5,247,650 bp
5,247,675 bp
5,247,700 bp
5,247,725 bp
5,247,750 bp
5,247,775 bp
5,247,800 bp
5,247,825 bp
5,247,850 bp
NA12877
PRISM04.bam
Refseq Genes


IGV
hg19
chr11
chr11:5,248,048-5,248,273
226 bp
Cursor Guide
Center Line
Track Labels
8,050 bp
5,248,075 bp
5,248,100 bp
5,248,125 bp
5,248,150 bp
5,248,175 bp
5,248,200 bp
5,248,225 bp
5,248,250 bp
NA12877
PRISM04.bam
Refseq Genes
HBB

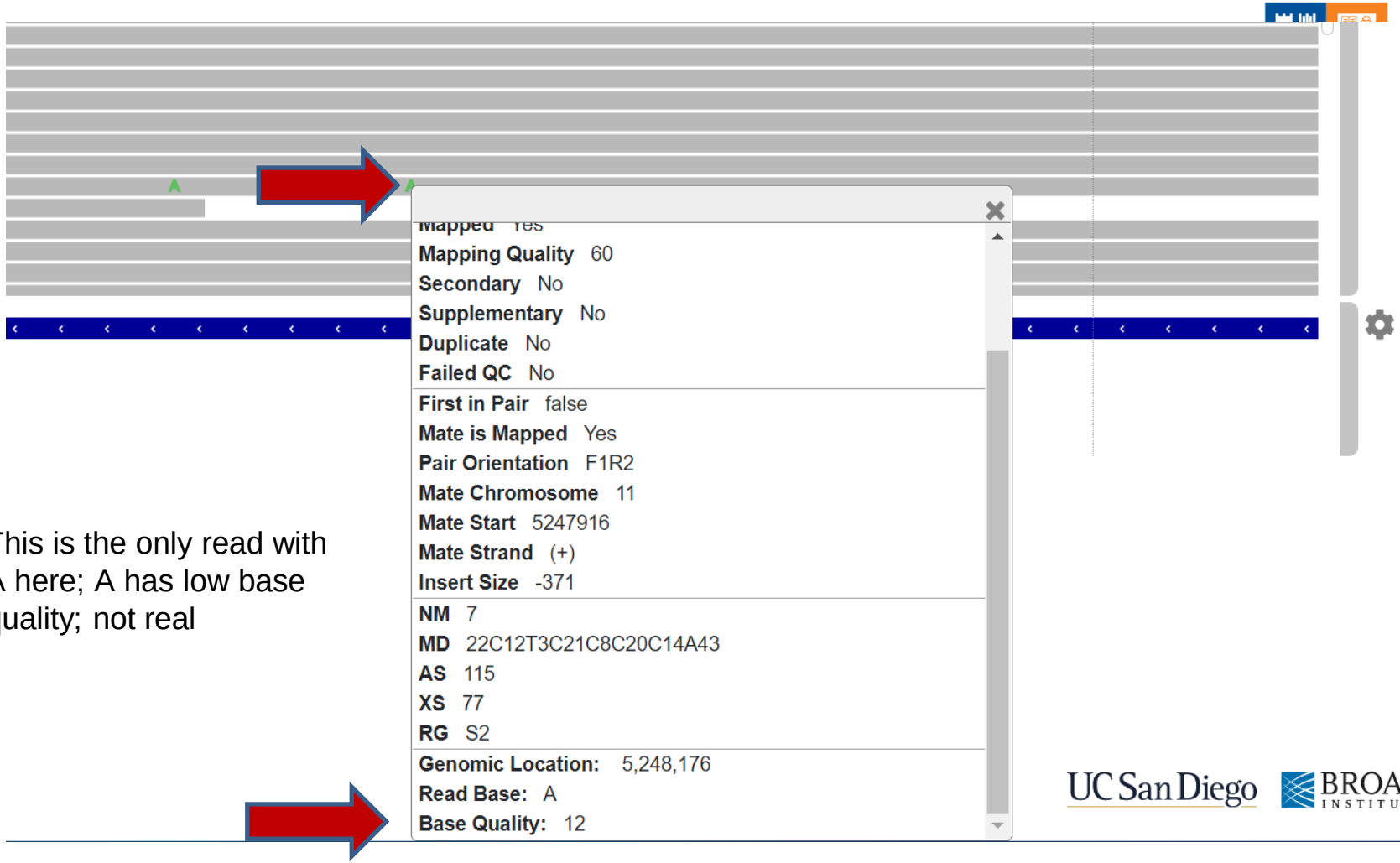
Heterozygous mutation at 3’ splice site of intron 2 of HBB (beta haemoglobin), “beta-thalassemia trait”



Sequence at splice sites very conserved: 5’ end of intron usually 5’-GT-3’, 3’ end is 5’-AG-3’; this variant takes

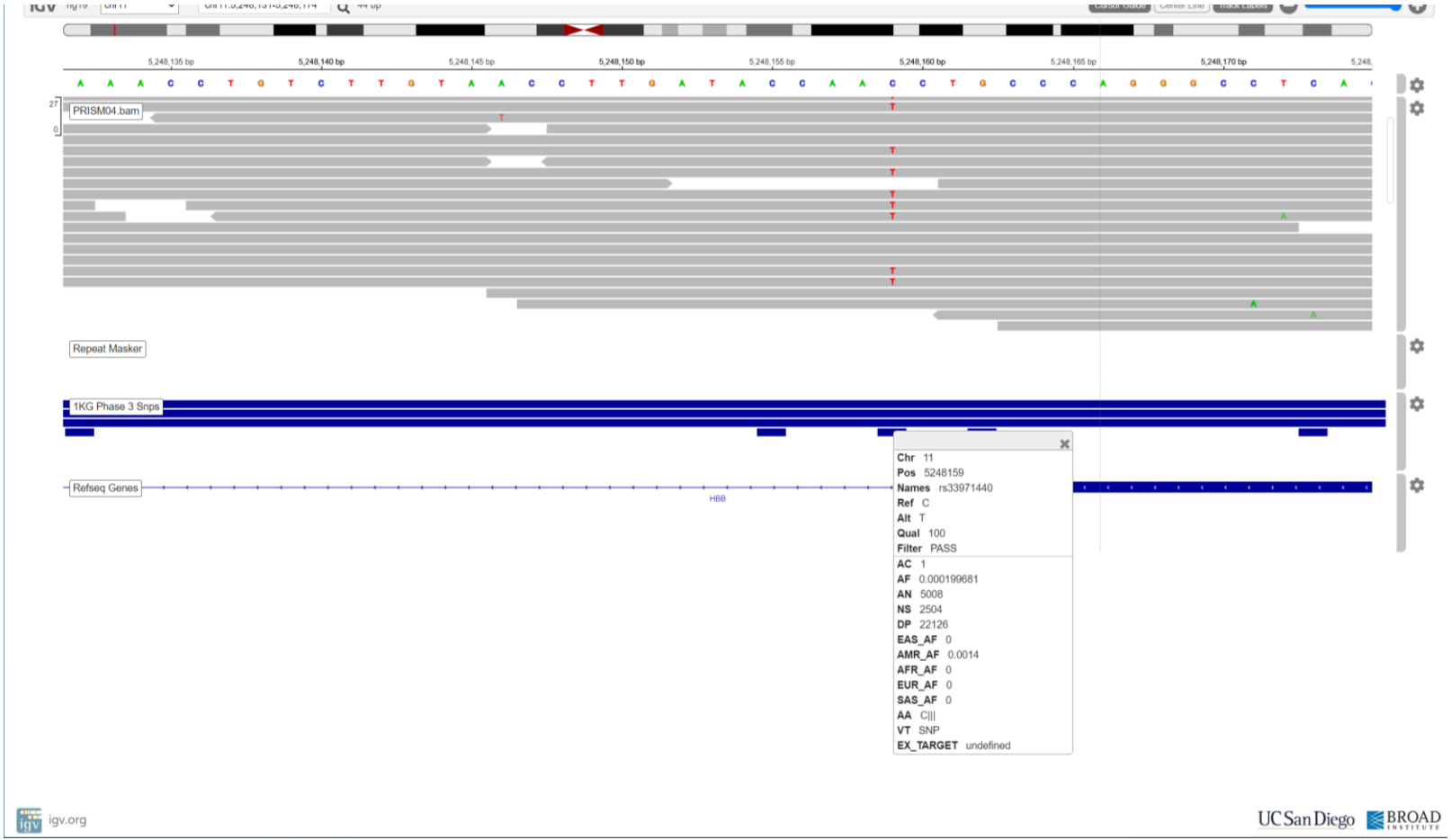
3′-TG-5′ → 3′-TA-5′
5′-AC-3′ → 5′-AT-3′

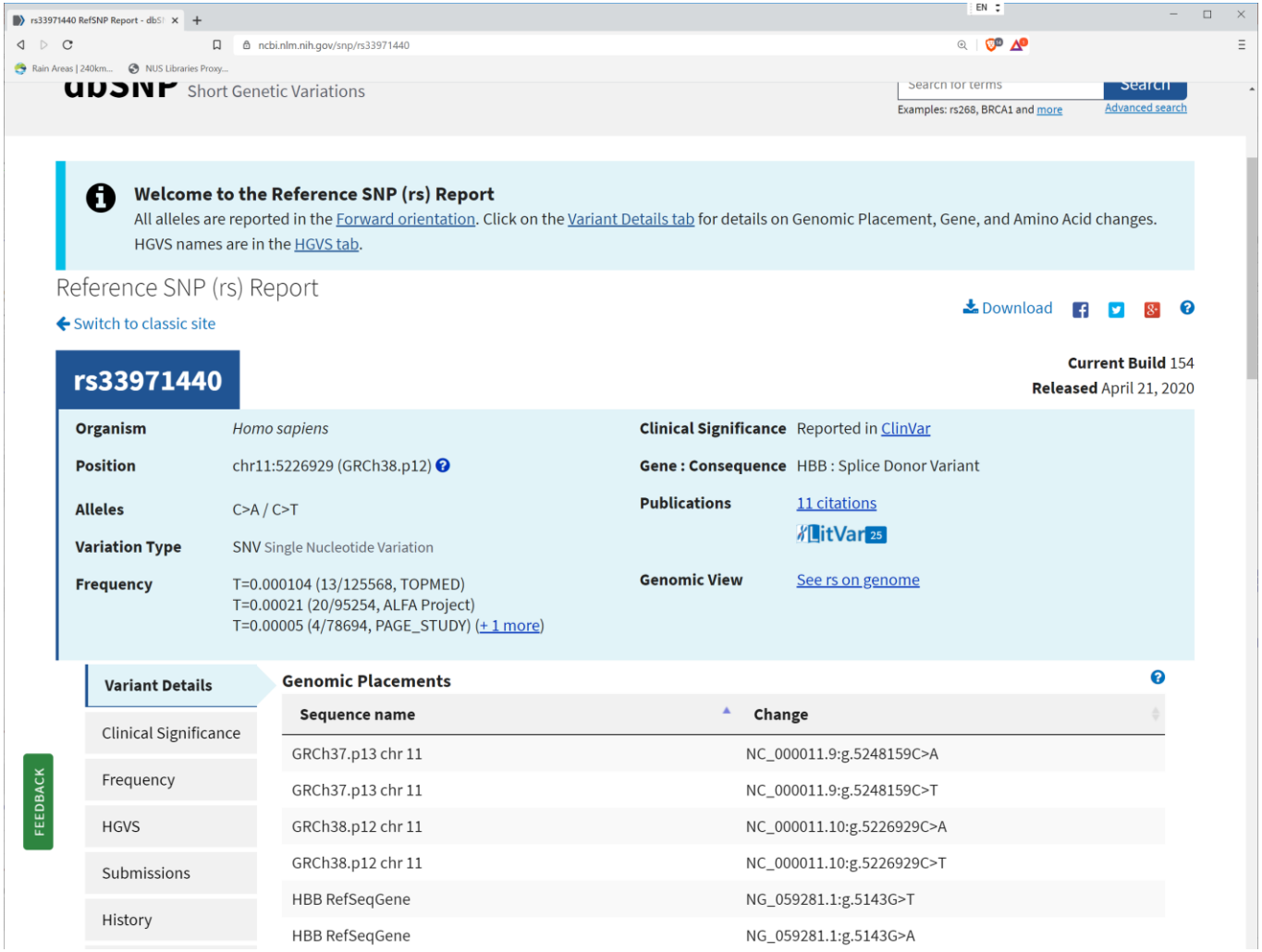
Info on one read and one base in that read
High base quality

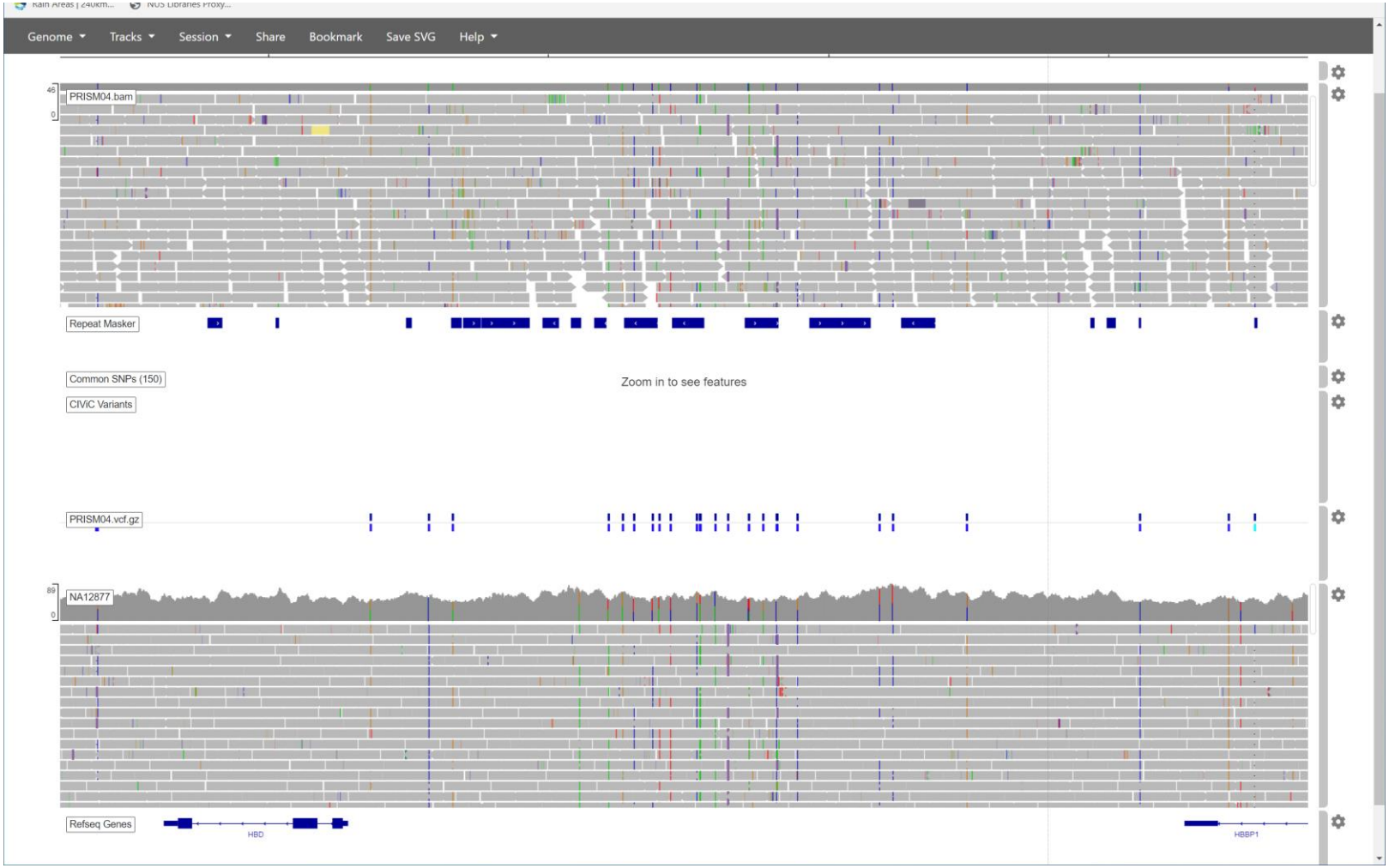


This is the only read with A here; A has low base quality; not real

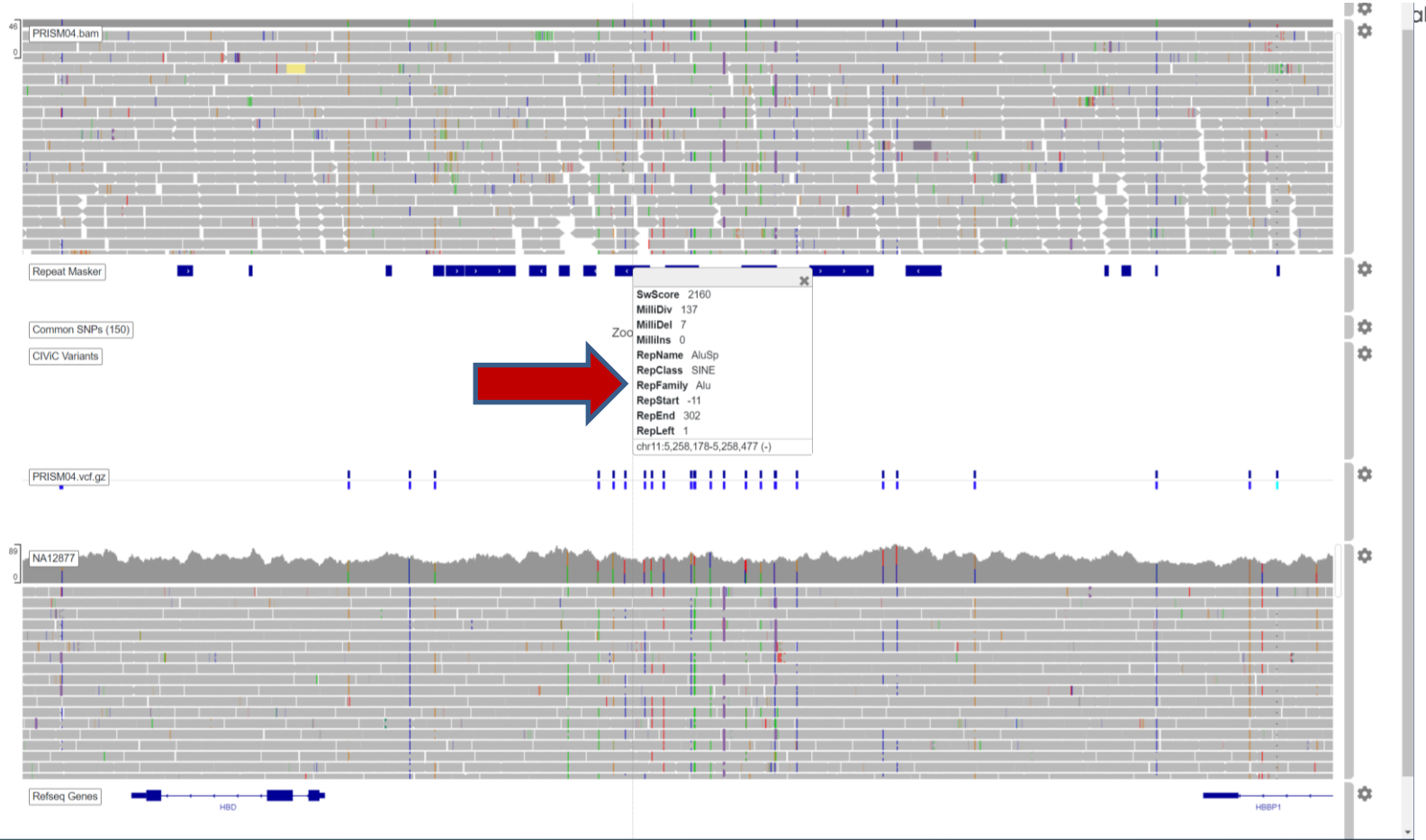
UC San Diego BROAD INSTITUTE

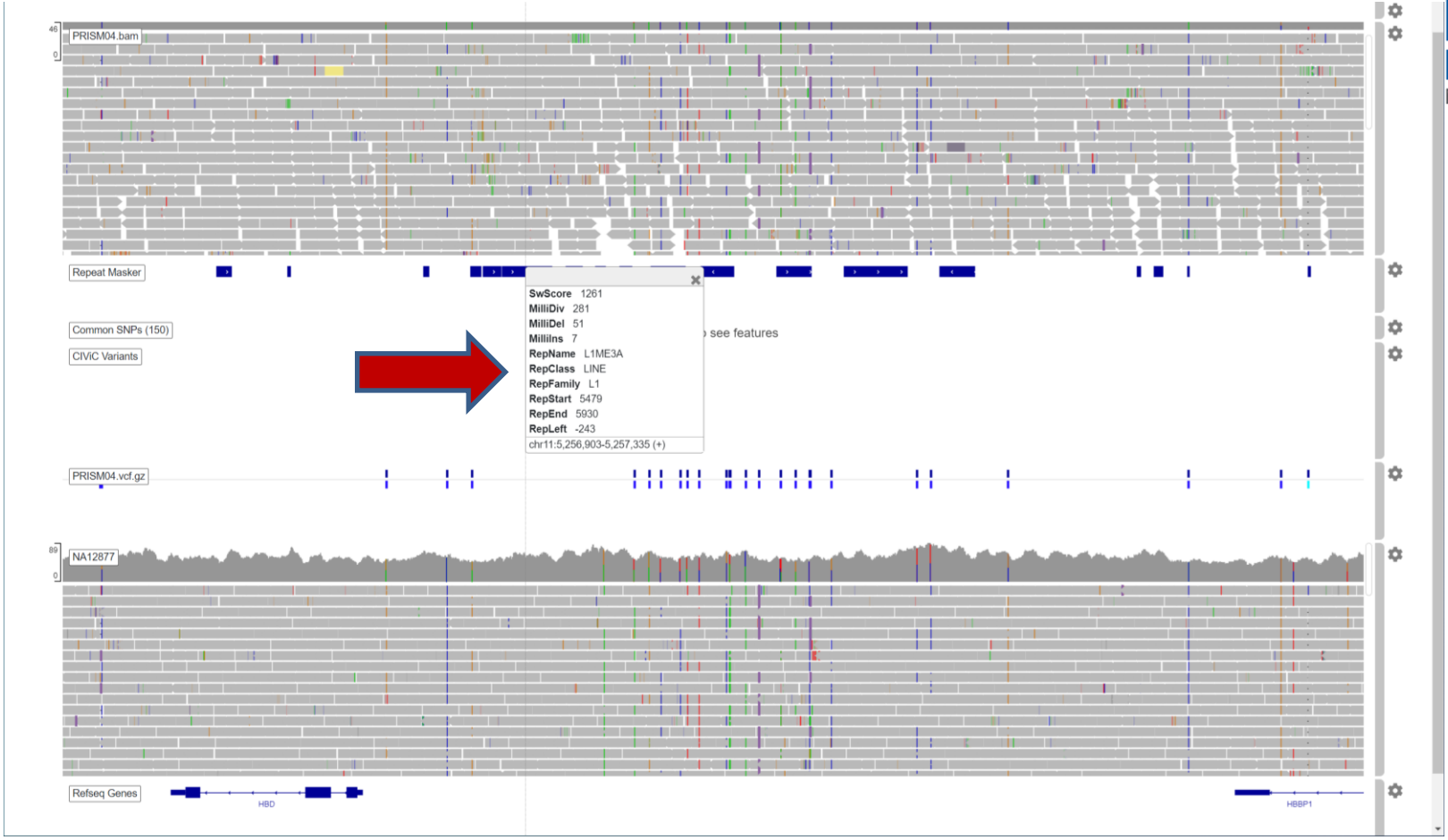

5,248,135 bp
5,248,140 bp
5,248,145 bp
5,248,150 bp
5,248,155 bp
5,248,160 bp
5,248,165 bp
5,248,170 bp
A A A C C T G T C T T G T A A C C T T G A T A C C A A C C T G C C C A G G G C C T C A
PRISM04.bam
Repeat Masker
1KG Phase 3 Snps
Refseq Genes
HBB
Chr 11
Pos 5248159
Names rs33971440
Ref C
Alt T
Qual 100
Filter PASS
AC 1
AF 0.000199681
AN 5008
NS 2504
DP 22126
EAS_AF 0
AMR_AF 0.0014
AFR_AF 0
EUR_AF 0
SAS_AF 0
AA C|||
VT SNP
EX_TARGET undefined


rs33971440 RefSNP Report - dbSNP
ncbi.nlm.nih.gov/snp/rs33971440
dbSNP Short Genetic Variations
Search for terms
Search
Examples: rs268, BRCA1 and more
Advanced search
Welcome to the Reference SNP (rs) Report
All alleles are reported in the Forward orientation. Click on the Variant Details tab for details on Genomic Placement, Gene, and Amino Acid changes.
HGVS names are in the HGVS tab.
Reference SNP (rs) Report
Download
Switch to classic site
rs33971440
Current Build 154
Released April 21, 2020
Organism Homo sapiens
Position chr11:5226929 (GRCh38.p12)
Alleles C>A / C>T
Variation Type SNV Single Nucleotide Variation
Frequency T=0.000104 (13/125568, TOPMED)
T=0.00021 (20/95254, ALFA Project)
T=0.00005 (4/78694, PAGE_STUDY) (+ 1 more)
Clinical Significance Reported in ClinVar
Gene : Consequence HBB : Splice Donor Variant
Publications 11 citations
LitVar 25
Genomic View See rs on genome
Variant Details
Clinical Significance
Frequency
HGVS
Submissions
History
Genomic Placements
Sequence name
Change
GRCh37.p13 chr 11 NC_000011.9:g.5248159C>A
GRCh37.p13 chr 11 NC_000011.9:g.5248159C>T
GRCh38.p12 chr 11 NC_000011.10:g.5226929C>A
GRCh38.p12 chr 11 NC_000011.10:g.5226929C>T
HBB RefSeqGene NG_059281.1:g.5143G>T
HBB RefSeqGene NG_059281.1:g.5143G>A
FEEDBACK



DukeNUS Medical School


PRISM04.bam
Repeat Masker
SwScore 2160
MilliDiv 137
MilliDel 7
MilliIns 0
RepName AluSp
RepClass SINE
RepFamily Alu
RepStart -11
RepEnd 302
RepLeft 1
chr11:5,258,178-5,258,477 (-)
Common SNPs (150)
CIViC Variants
PRISM04.vcf.gz
NA12877
Refseq Genes
HBD
HBBP1


PRISM04.bam
Repeat Masker
SwScore 1261
MilliDiv 281
MilliDel 51
MilliIns 7
RepName L1ME3A
RepClass LINE
RepFamily L1
RepStart 5479
RepEnd 5930
RepLeft -243
chr11:5,256,903-5,257,335 (+)
see features
Common SNPs (150)
CIViC Variants
PRISM04.vcf.gz
NA12877
Refseq Genes
HBD
HBBP1

Genome variation

- Germline or cancer context is important
- Single nucleotide variants (can be mutations or SNPs; SNP == “single nucleotide polymorphism”; terminology is often used loosely).
- Indels (small insertion deletions)
- Copy number alterations (larger insertions and deletions)
- Structural variants (inversions, inter-chromosomal events, includes larger insertions and deletions, but focus is on breakpoints)

Germ-line variants versus somatic mutations in cancer

- Cancer is (partly) a disease caused by gene mutations that occur in cancer cells
- These are the somatic mutations
- Method for finding somatic mutations is conceptually simple
  1. Sequence DNA from the tumor and find genetic variants
  2. Sequence DNA from non-tumor tissue from the same patient and find variants
  3. Subtract

Reference genome versions

- Human
  - hg19 (CRCh37)
  - hg38 (CRCh38)
- Other species: Mouse mm10, mm9, dog, C. elegans, Drosophila, ….
- Remember, a BAM file has reads that are aligned to one reference genome, so hard to interpret e.g. an hg19-aligned BAM file in the context of the hg38 reference genome
- The “liftover” program can move annotations from one reference genome to another, but the mapping is lossy
- If you want to change the reference genome of a BAM file you need to do the alignment over against the new reference genome

“Variant callers” (For SNVs and small indels)

- Freebayes https://github.com/freebayes/freebayes (germline)
- HaplotypeCaller (germline, part of the GATK pipeline https://gatk.broadinstitute.org/hc/en-us) https://gatk.broadinstitute.org/hc/en-us/articles/360050814612-HaplotypeCaller
- Mutect2 (part of the GATK pipeline / tool suite) Somatic variants (“tumor minus normal) https://gatk.broadinstitute.org/hc/en-us/articles/360051306691-Mutect2
- Strelka (germline and somatic) https://www.biorxiv.org/content/10.1101/192872v2 https://github.com/Illumina/strelka

More on SNV variant callers

- Now staring to use convolutional neural nets to filter variants: “CNNScoreVariants” https://gatk.broadinstitute.org/hc/en-us/articles/360037226672-CNNScoreVariants
- Google DeepVariant https://github.com/google/deepvariant -- trained on many sequencing technologies
- LIM Weng Khong

For copy number variants (CNVs) and structural variants (SVs)

- Smoove https://github.com/brentp/smoove
- SG10K is using
  - CNVpytor for CNVs https://github.com/abyzovlab/CNVpytor
  - Manta for SVs https://github.com/Illumina/manta
  - SurVindel (for a special kind of SV – tandem duplication)

VCF (variant call file) output of variant callers

##fileformat=VCFv4.2
##FILTER=<ID=alignment_artifact,Description="Alignment artifact">
##FILTER=<ID=bad_haplotype,Description="Variant near filtered variant on same haplotype.">
##FILTER=<ID=base_quality,Description="alt median base quality">
….
##FILTER=<ID=clustered_events,Description="Clustered events observed in the tumor">
##FILTER=<ID=contamination,Description="contamination">
##FILTER=<ID=duplicate_evidence,Description="evidence for alt allele is overrepresented by apparent duplicates">
##FILTER=<ID=fragment_length,Description="abs(ref - alt) median fragment length">
…
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=PID,Number=1,Type=String,Description="Physical phasing ID information, where each unique ID within a given sample (but not across samples) con
records within ##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
…
##GATKCommandLine=<ID=FilterAlignmentArtifacts,CommandLine="FilterAlignmentArtifacts --output
/home/gmsywss/temp_folder/…vcf_folder/mutect2_##GATKCommandLine=<ID=FilterMutectCalls,CommandLine="FilterMutectCalls –output ….. …
##INFO=<ID=TLOD,Number=A,Type=Float,Description="Log odds ratio score for variant">
##MutectVersion=2.1
##contig=<ID=1,length=249250621>
##contig=<ID=2,length=243199373>
…
##contig=<ID=GL000247.1,length=36422>
##contig=<ID=GL000203.1,length=37498>
…
##filtering_status=These calls have been filtered by FilterMutectCalls to label false positives with a list of failed filters and true positives with PASS.
##normal_sample=MCF10A_DHG11565_HH2GKALXX
##source=FilterAlignmentArtifacts

VCF (v

Starts with lots of “metadata” – data about how the file was generated and how to interpret the file

There are different formats, but they are not very standardized

```
##fileformat=VCFv4.2
##FILTER=<ID=alignment_artifact,Description="Alignment artifact">
##FILTER=<ID=bad_haplotype,Description="Variant near filtered variant on same haplotype.">
##FILTER=<ID=base_quality,Description="alt median base quality">
....
##FILTER=<ID=clustered_events,Description="Clustered events observed in the tumor">
##FILTER=<ID=contamination,Description="contamination">
##FILTER=<ID=duplicate_evidence,Description="evidence for alt allele is overrepresented by apparent duplicates">
##FILTER=<ID=fragment_length,Description="abs(ref - alt) median fragment length">
...
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=PID,Number=1,Type=String,Description
...
##GATKCommandLine=<ID=FilterAlignmentArtifacts,CommandLine="FilterAlignmentArtifacts --output
/home/gmsywss/temp_folder/...vcf_folder/mutect2_##GATKCommandLine=<ID=FilterMutectCalls,CommandLine="FilterMutectCalls –output ..... ...
##INFO=<ID=TLOD,Number=A,Type=Float,Description="Log odds ratio score for variant">
##MutectVersion=2.1
##contig=<ID=1,length=249250621>
##contig=<ID=2,length=243199373>
...
##contig=<ID=GL000247.1,length=36422>
##contig=<ID=GL000203.1,length=37498>
...
##filtering_status=These calls have been filtered by FilterMutectCalls to label false positives with a list of failed filters and true positives with PASS.
##normal_sample=MCF10A_DHG11565_HH2GKALXX
##source=FilterAlignmentArtifacts
```

Information content is extensible with this kind of metadata

ss samples) ....

For humans, chromosomes are called “contigs”, a historical term for a contiguous stretch of sequence

These contigs are little bits of the human genome that are “lost” – we don’t know where they came from

DukeNUS Medical School

VCF, the actual variants and associated information

DukeNUS Medical School

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT
1	813109	.	A	C	.	PASS	DP=57;ECNT=2;NLOD=8.13;N_ART_LOD=-1.447e+00;POP_AF=0.020;P_CONTAM=5.695e-09;P_GERMLINE=-5.618e+00;RCNTS=1,4;TLOD=19.95	GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:PGT:
1	813112	.	T	C	.	PASS	DP=55;ECNT=2;NLOD=8.13;N_ART_LOD=-1.447e+00;POP_AF=0.020;P_CONTAM=5.695e-09;P_GERMLINE=-5.618e+00;RCNTS=1,4;TLOD=19.96	GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:PGT:
1	815757	.	G	A	.	PASS	DP=66;ECNT=2;NLOD=9.03;N_ART_LOD=-1.503e+00;POP_AF=0.182;P_CONTAM=2.348e-04;P_GERMLINE=-7.431e+00;RCNTS=2,0;TLOD=10.61	GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:SA_M
1	906904	.	C	A	.	PASS	DP=70;ECNT=1;NLOD=11.09;N_ART_LOD=-1.579e+00;POP_AF=2.500e-06;P_CONTAM=0.00;P_GERMLINE=-1.071e+01;RCNTS=2,0;TLOD=38.81	GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:SA_M
1	1584047	.	C	T	.	PASS	DP=80;ECNT=1;NLOD=5.85;N_ART_LOD=-3.220e-01;POP_AF=0.017;P_CONTAM=1.328e-14;P_GERMLINE=-5.902e+00;RCNTS=0,4;TLOD=22.18	GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:SA_M
1	1821952	.	C	T	.	PASS	DP=47;ECNT=1;NLOD=5.42;N_ART_LOD=-1.279e+00;POP_AF=2.500e-06;P_CONTAM=0.00;P_GERMLINE=-4.727e+00;RCNTS=2,0;TLOD=42.72	GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:SA_M
...								
1	4785169	.	C	A	.	PASS	DP=74;ECNT=1;NLOD=11.99;N_ART_LOD=-1.621e+00;POP_AF=2.500e-06;P_CONTAM=0.00;P_GERMLINE=-1.173e+01;RCNTS=2,0;TLOD=32.32	GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:SA_M
1	5071298	.	A	G	.	PASS	DP=61;ECNT=1;NLOD=6.32;N_ART_LOD=-1.351e+00;POP_AF=2.500e-06;P_CONTAM=0.00;P_GERMLINE=-7.059e+00;RCNTS=2,0;TLOD=28.53	GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:SA_M
1	5288645	.	TG	T	.	PASS	DP=82;ECNT=1;NLOD=10.53;N_ART_LOD=-1.556e+00;POP_AF=2.500e-06;P_CONTAM=0.00;P_GERMLINE=-1.022e+01;RCNTS=2,0;RPA=2,1;RU=G;STR;	GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:SA_M
1	5416531	.	A	C	.	PASS	DP=96;ECNT=1;NLOD=10.53;N_ART_LOD=-1.561e+00;POP_AF=2.500e-06;P_CONTAM=0.00;P_GERMLINE=-1.013e+01;RCNTS=0,4;TLOD=64.03	GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:SA_M
1	5534473	.	TC	AT	.	PASS	DP=61;ECNT=1;NLOD=9.63;N_ART_LOD=-1.524e+00;POP_AF=2.500e-06;P_CONTAM=0.00;P_GERMLINE=-8.931e+00;RCNTS=2,0;TLOD=48.80	GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:SA_M
1	5815799	.	G	T	.	PASS	DP=59;ECNT=1;NLOD=8.38;N_ART_LOD=-1.474e+00;POP_AF=2.500e-06;P_CONTAM=0.00;P_GERMLINE=-7.961e+00;RCNTS=2,0;TLOD=56.31	GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:SA_M
1	5851107	.	G	A	.	PASS	DP=66;ECNT=1;NLOD=6.62;N_ART_LOD=-1.385e+00;POP_AF=2.500e-06;P_CONTAM=0.00;P_GERMLINE=-6.060e+00;RCNTS=0,4;TLOD=70.66	GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:SA_M
1	5906421	.	G	A	.	PASS	DP=67;ECNT=1;NLOD=10.79;N_ART_LOD=-1.577e+00;POP_AF=2.500e-06;P_CONTAM=0.00;P_GERMLINE=-1.012e+01;RCNTS=2,0;TLOD=37.78	GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:SA_M
1	6248930	.	C	T	.	PASS	DP=50;ECNT=1;NLOD=6.02;N_ART_LOD=-1.331e+00;POP_AF=2.500e-06;P_CONTAM=0.00;P_GERMLINE=-6.110e+00;RCNTS=0,4;TLOD=26.73	GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:SA_M
1	6361849	.	G	T	.	PASS	DP=80;ECNT=1;NLOD=9.03;N_ART_LOD=-1.509e+00;POP_AF=2.500e-06;P_CONTAM=0.00;P_GERMLINE=-8.336e+00;RCNTS=2,0;TLOD=69.58	GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:SA_M
...								

VCF, the actual variants and associated information

Chromosome (contig) in which the variant occurred

Location of the variant in the chromosome

The reference “allele” (base or short sequence)

The “alternative” allele (the variant)

Post-processing assessment of the variants – is it believable?

Complicated stuff with interpretation that depends on the metadata in the header

Most are “SNVs” (or “SBS”), e.g. C to T here

Deletion of a “G” (a type of indel)

TC mutated to AT

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT
1	813109	.	A	C	.	PASS	DP=57;ECNT=2;NLOD=8.13;N_ART_LOD=-1.447e+00;POP_AF=0.020;P_CONTAM=5.695e-09;P_GERMLINE=-5.618e+00;RCNTS=1,4;TLOD=19.95	GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:PGT
1	813112	.	T	C	.	PASS	DP=55;ECNT=2;NLOD=8.13;N_ART_LOD=-1.447e+00;POP_AF=0.020;P_CONTAM=5.695e-09;P_GERMLINE=-5.618e+00;RCNTS=1,4;TLOD=19.96	GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:PGT
1	815757	.	G	A	.	PASS	DP=66;ECNT=2;NLOD=9.0	GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:SA_
1	906904	.	C	A	.	PASS	DP=70;ECNT=1;NLOD=11	GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:SA_
1	1584047	.	C	T	.		D=5.8	GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:SA_
1	1821952	.	C	T	.	PASS	DP=47;ECNT=1;NLOD=5.42	GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:SA_
...								
1	4785169	.	C	A	.	PASS	DP=74;ECNT=1;NLOD=11.9	NTS=2,0;TLOD=32.32 GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:SA_
1	5071298	.	A	G	.	PASS	DP=61;ECNT=1;NLOD=6.32	NTS=2,0;TLOD=28.53 GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:SA_
1	5288645	.	TG	T	.		=10.5	NTS=2,0;RPA=2,1;RU=G;STR; GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:SA_
1	5416531	.	A	C	.	PASS	DP=96;ECNT=1;NLOD=10.5	NTS=0,4;TLOD=64.03 GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:SA_
1	5534473	.	TC	AT	.		=9.63	M=0.00;P_GERMLINE=-8.931e+00;RCNTS=2,0;TLOD=48.80 GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:SA_
1	5815799	.	G	T	.	PASS	DP=59;ECNT=1;NLOD=8.38	M=0.00;P_GERMLINE=-7.961e+00;RCNTS=2,0;TLOD=56.31 GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:SA_
1	5851107	.	G	A	.	PASS	DP=66;ECNT=1;NLOD=6.62;N_ART_LOD=-1.385e+00;POP_AF=2.500e-06;P_CONTAM=0.00;P_GERMLINE=-6.060e+00;RCNTS=0,4;TLOD=70.66	GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:SA_
1	5906421	.	G	A	.	PASS	DP=67;ECNT=1;NLOD=10.79;N_ART_LOD=-1.577e+00;POP_AF=2.500e-06;P_CONTAM=0.00;P_GERMLINE=-1.012e+01;RCNTS=2,0;TLOD=37.78	GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:SA_
1	6248930	.	C	T	.	PASS	DP=50;ECNT=1;NLOD=6.02;N_ART_LOD=-1.331e+00;POP_AF=2.500e-06;P_CONTAM=0.00;P_GERMLINE=-6.110e+00;RCNTS=0,4;TLOD=26.73	GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:SA_
1	6361849	.	G	T	.	PASS	DP=80;ECNT=1;NLOD=9.03;N_ART_LOD=-1.509e+00;POP_AF=2.500e-06;P_CONTAM=0.00;P_GERMLINE=-8.336e+00;RCNTS=2,0;TLOD=69.58	GT:AD:AF:DP:F1R2:F2R1:MBQ:MFRL:MMQ:MPOS:ORIGINAL_CONTIG_MISMATCH:SA_
...								

Somatic variant VCF, more info to the left

Information about the “tumor”

Information about the “normal”

MCF10A_Carb_Low_cl2	MCF10A_DHG11565_HH2GKALXX
0/1:22,6:0.233:28:10,0:12,6:33,33:263,267:37:43:0:0|1:813109_A_C:0.00,0.212,0.214:0.509,3.349e-03,0.488	0/0:27,0:2.860e-06:27:10,0:17,0:33,0:286,0:0:0:0:0|1:813109_A_C
0/1:22,6:0.233:28:10,0:12,6:33,32:263,267:37:40:0:0|1:813109_A_C:0.00,0.212,0.214:0.509,3.349e-03,0.488	0/0:27,0:2.860e-06:27:10,0:17,0:33,0:286,0:0:0:0:0|1:813109_A_C
0/1:29,5:0.166:34:17,2:12,3:33,34:258,261:37:24:0:0.152,0.00,0.147:4.228e-03,0.300,0.696	0/0:30,0:0.032:30:10,0:20,0:33,0:296,0:0:0:0
0/1:20,13:0.400:33:5,4:15,9:34,34:258,205:60:30:0:0.364,0.364,0.394:0.043,0.014,0.943	0/0:37,0:2.907e-03:37:19,0:18,0:33,0:285,0:0:0:0
0/1:40,9:0.196:49:18,4:22,5:34,34:246,262:57:33:0:0.182,0.00,0.184:6.355e-03,0.148,0.845	0/0:27,1:0.052:28:15,0:12,1:33,34:294,262:21:35:0
0/1:15,14:0.490:29:4,3:11,11:34,34:253,254:60:20:0:0.434,0.465,0.483:0.034,0.017,0.949	0/0:18,0:1.803e-04:18:9,0:9,0:33,0:308,0:0:0:0
0/1:20,12:0.382:32:10,8:10,4:34,33:253,264:60:31:0:0.364,0.323,0.375:0.014,0.040,0.947	0/0:40,0:0.026:40:23,0:17,0:33,0:309,0:0:0:0
0/1:25,10:0.297:35:9,2:16,8:34,34:243,252:60:29:0:0.253,0.263,0.286:0.026,0.015,0.959	0/0:21,0:0.023:21:11,0:10,0:33,0:290,0:0:0:0
0/1:18,27:0.596:45:6,6:12,21:34,34:242,272:60:21:0:0.576,0.576,0.600:0.021,0.025,0.954	0/0:35,0:2.312e-04:35:16,0:19,0:34,0:288,0:0:0:0
0/1:34,25:0.424:59:18,11:16,14:34,29:249,274:60:26:0:0.354,0.414,0.424:0.079,9.270e-03,0.912	0/0:35,0:0.014:35:14,0:21,0:35,0:288,0:0:0:0
0/1:13,13:0.500:26:4,11:9,2:33,34:263,236:60:33:0:0.495,0.444,0.500:0.018,0.035,0.947	0/0:32,0:0.016:32:14,0:18,0:34,0:291,0:0:0:0
0/1:11,17:0.600:28:6,9:5,8:33,34:260,251:60:23:0:0.586,0.586,0.607:0.029,0.021,0.951	0/0:28,0:0.037:28:18,0:10,0:33,0:295,0:0:0:0
0/1:17,22:0.559:39:11,9:6,13:34,34:293,256:60:28:0:0.556,0.515,0.564:0.013,0.055,0.932	0/0:22,0:0.060:22:12,0:10,0:32,0:293,0:0:0:0
0/1:14,12:0.464:26:9,6:5,6:35,34:239,229:60:27:0:0.414,0.444,0.462:0.025,0.022,0.953	0/0:36,0:0.029:36:14,0:22,0:33,0:288,0:0:0:0
0/1:19,9:0.333:28:5,3:14,6:34,34:252,251:60:29:0:0.283,0.303,0.321:0.022,0.019,0.958	0/0:20,0:0.024:20:9,0:11,0:33,0:270,0:0:0:0
0/1:23,22:0.489:45:13,13:10,9:34,33:272,250:60:19:0:0.475,0.455,0.489:0.016,0.031,0.953	0/0:30,0:0.046:30:10,0:20,0:33,0:294,0:0:0:0

Header line, continued

Genome Browsers

Genome browsers; some of the same info as alignment viewers, but don't show reads and have a lot more "genome annotation"

UCSC genome browser

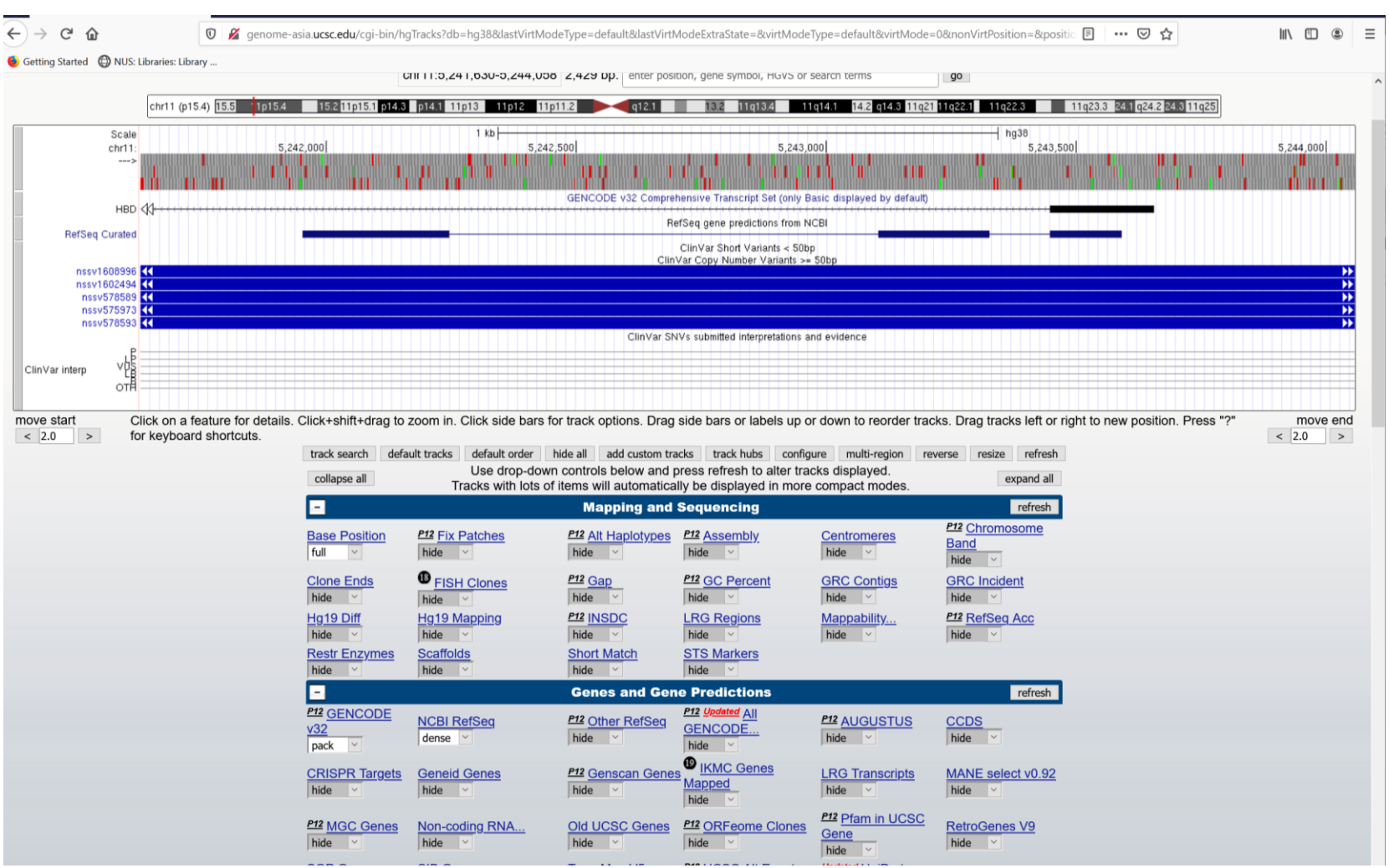
http://genome-asia.ucsc.edu/cgi-bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr11%3A5241442%2D5243870&hgsid=741515590_lwevGa81AEJZBs3CgBH3PRmsIFKE

Ensembl genome browser

http://asia.ensembl.org/Homo_sapiens/Location/View?h=G3HBB6.1%20(1-32);r=8:6730731-6730826

The genome browser gather lots of information and try to organize it by genome location.

It is possible to add new tracks to the UCSC genome browser



DukeNUS Medical School

Challenges with WGS compared to whole exome (extra material)

SingHealth DukeNUS ACADEMIC MEDICAL CENTRE
Institute of Precision Medicine (PRISM)

1. Disk space: 70 GB per genome vs 4 GB per exome
   a) Takes much longer to transfer
   b) > 10X more expensive to store
2. Processing time: 1 day per genome vs 30 minutes per exome
   a) Requires either very long time to process
   b) Or massive investment in computing resources
3. Analysis
   a) 100% of genome to annotate/interpret vs ~2% in exomes
   b) Huge diversity in variation:
      - Coding variants
      - Noncoding variants
      - Structural variation
      - Copy number variation
      - Virome
      - HLA-typing