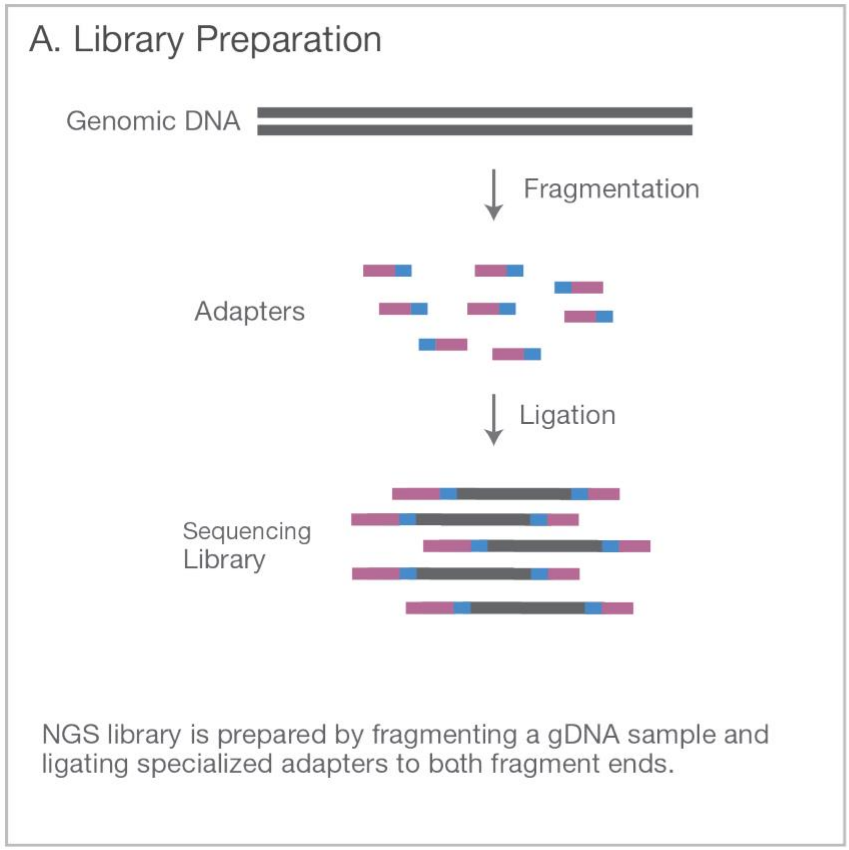
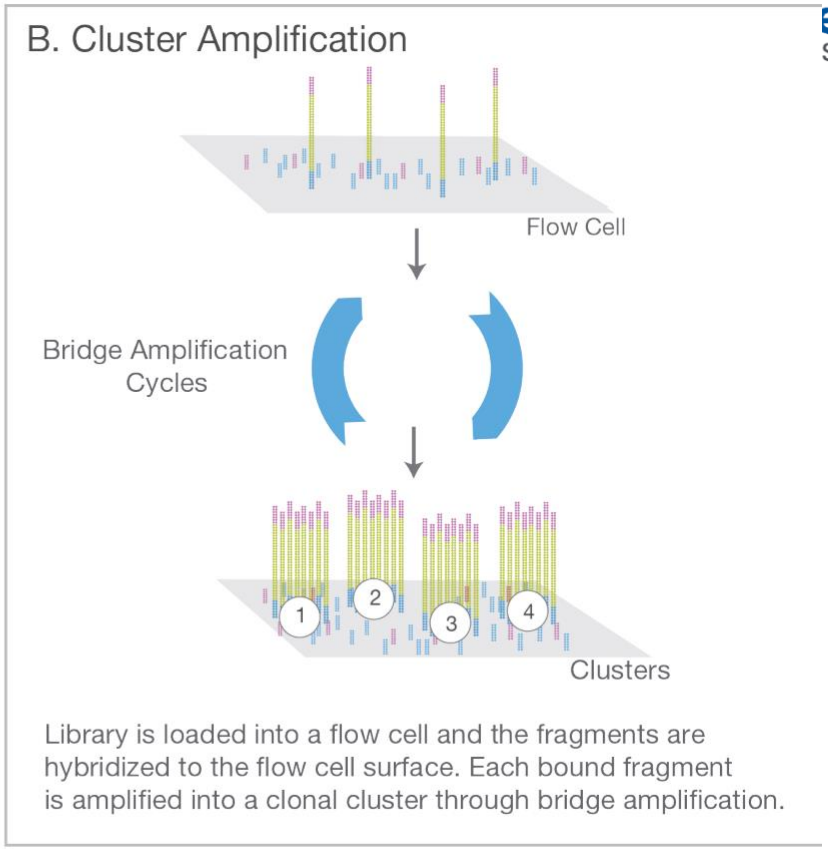
NGS (next generation sequencing) generates short reads, and usually the short reads need to be mapped (aligned) to a reference sequence. For genome sequencing, this is a reference genome.

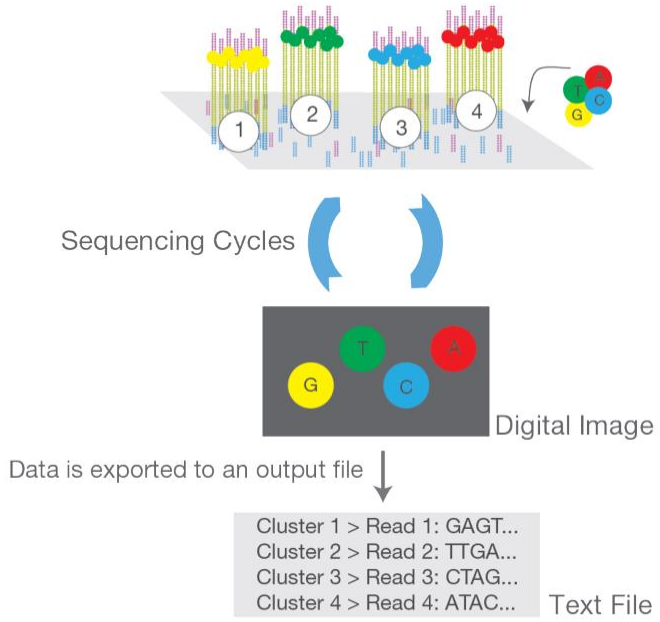
Why does the technology generate short reads?





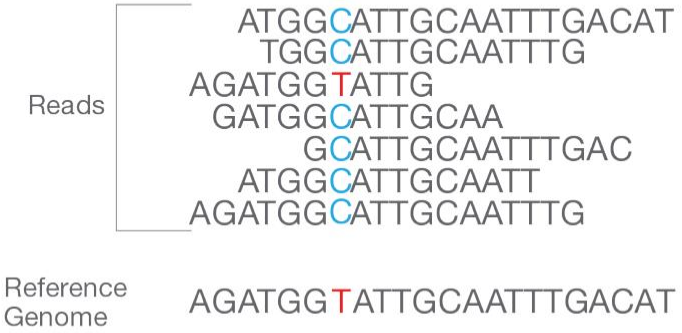
https://sapac.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

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

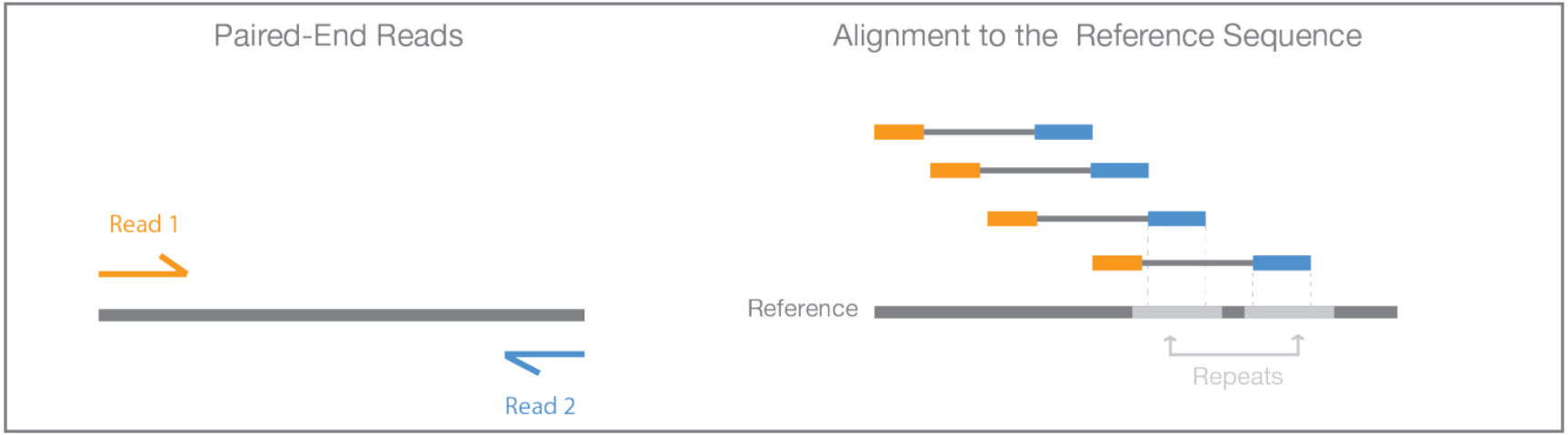


Figure 4: Paired-End Sequencing and Alignment—Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in better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

Short read mappers

only a few in wide use

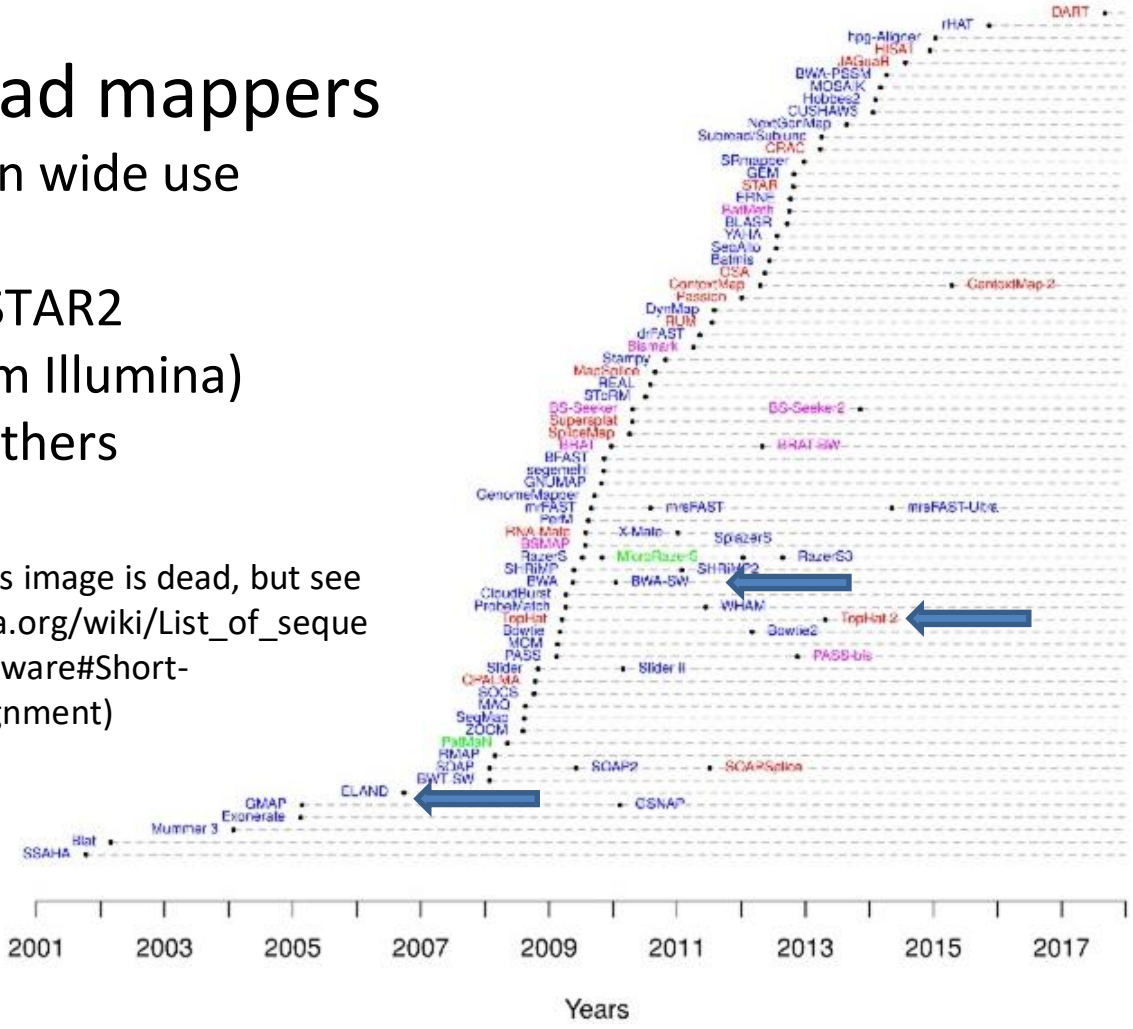
BWA-MEM

TOPHAT2, STAR2

ELAND (from Illumina)

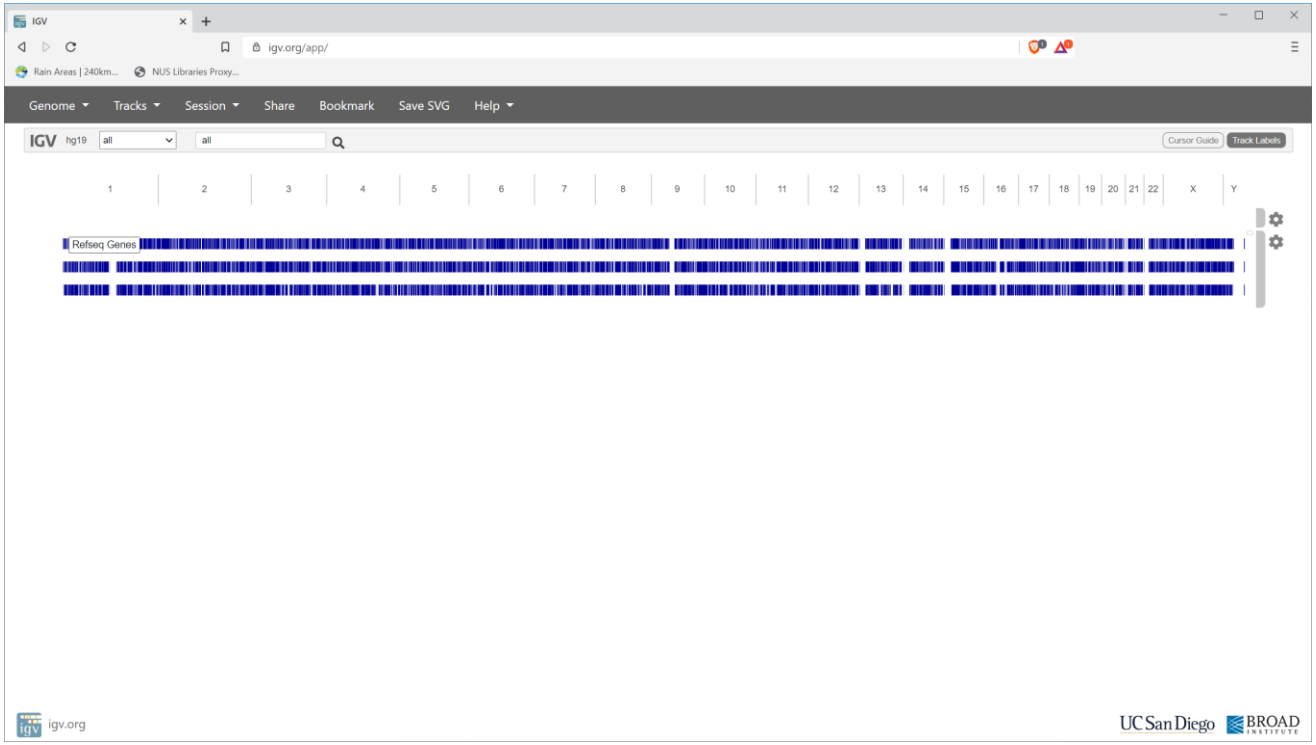
and a few others

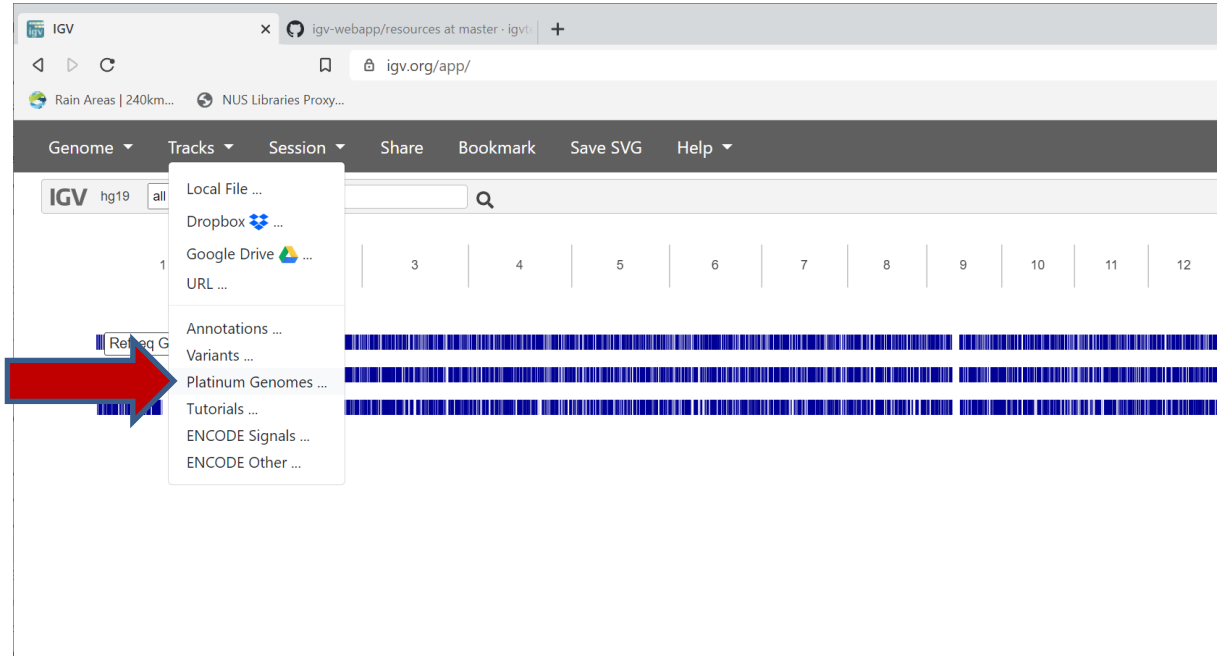
(Original URL for this image is dead, but see https://en.wikipedia.org/wiki/List_of_sequence_alignment_software#Short-read_sequence_alignment)

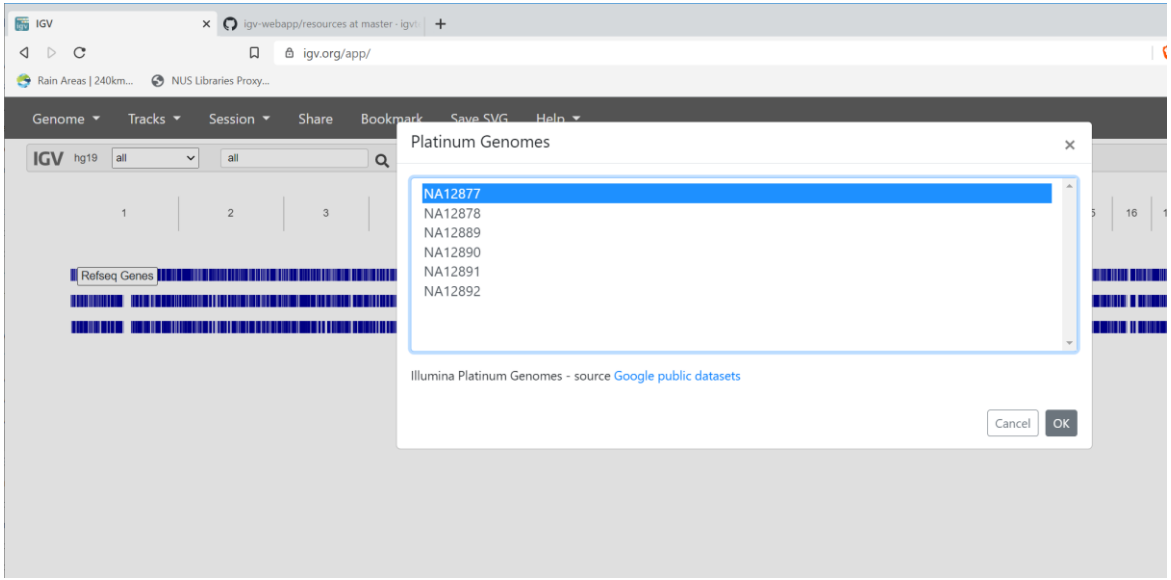


IGV (genome assembly browser)

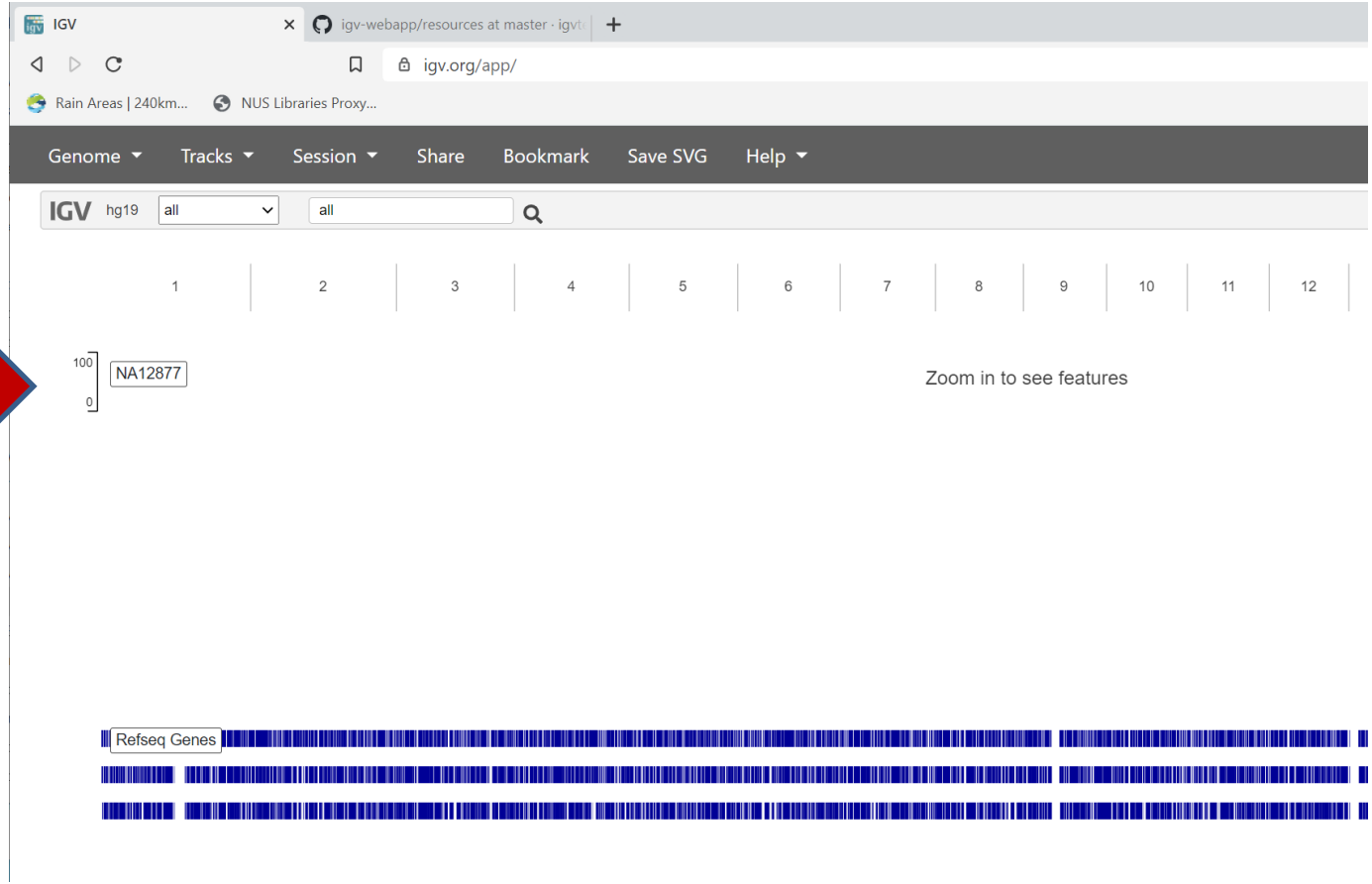
- https://igv.org/app

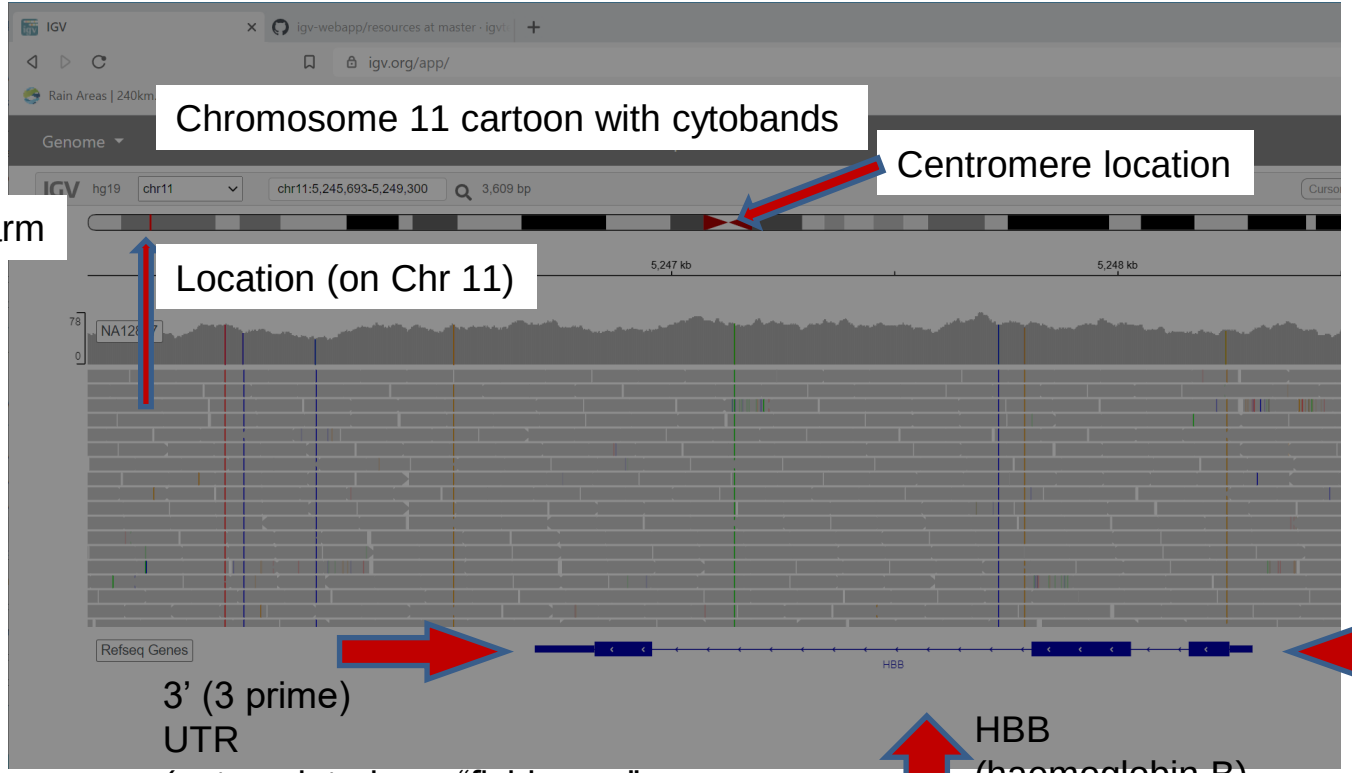



IGV
igv-webapp/resources at master · igv
igv.org/app/
Rain Areas | 240km...
NUS Libraries Proxy...
Genome
Tracks
Session
Share
Bookmark
Save SVG
Help
IGV
hg19
all
Local File ...
Dropbox ...
Google Drive ...
URL ...
Annotations ...
Variants ...
Platinum Genomes ...
Tutorials ...
ENCODE Signals ...
ENCODE 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





Chromosome 11 cartoon with cytobands

Centromere location

p arm

q arm

Location (on Chr 11)

Read depth plot

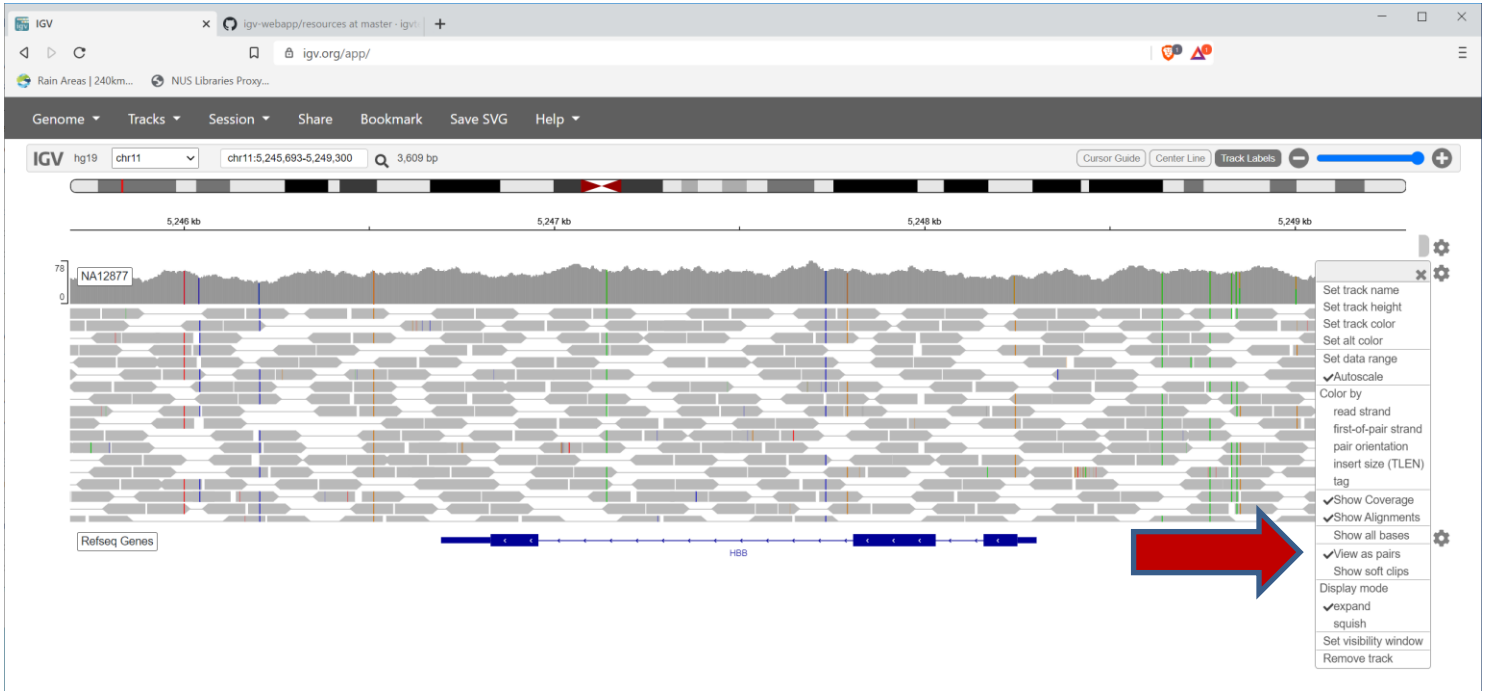
Reads

5’ (5 prime) UTR (untranslated region)

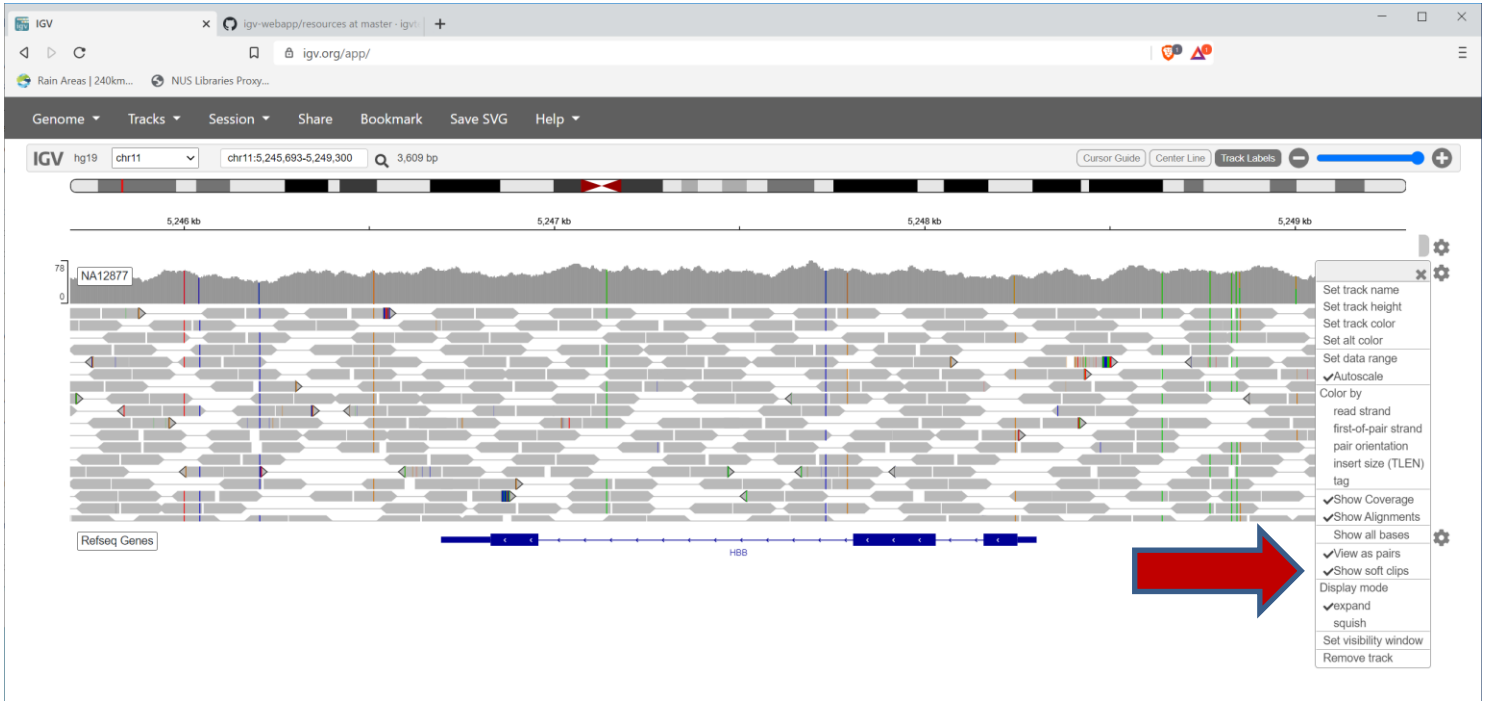
3’ (3 prime) UTR (untranslated region)

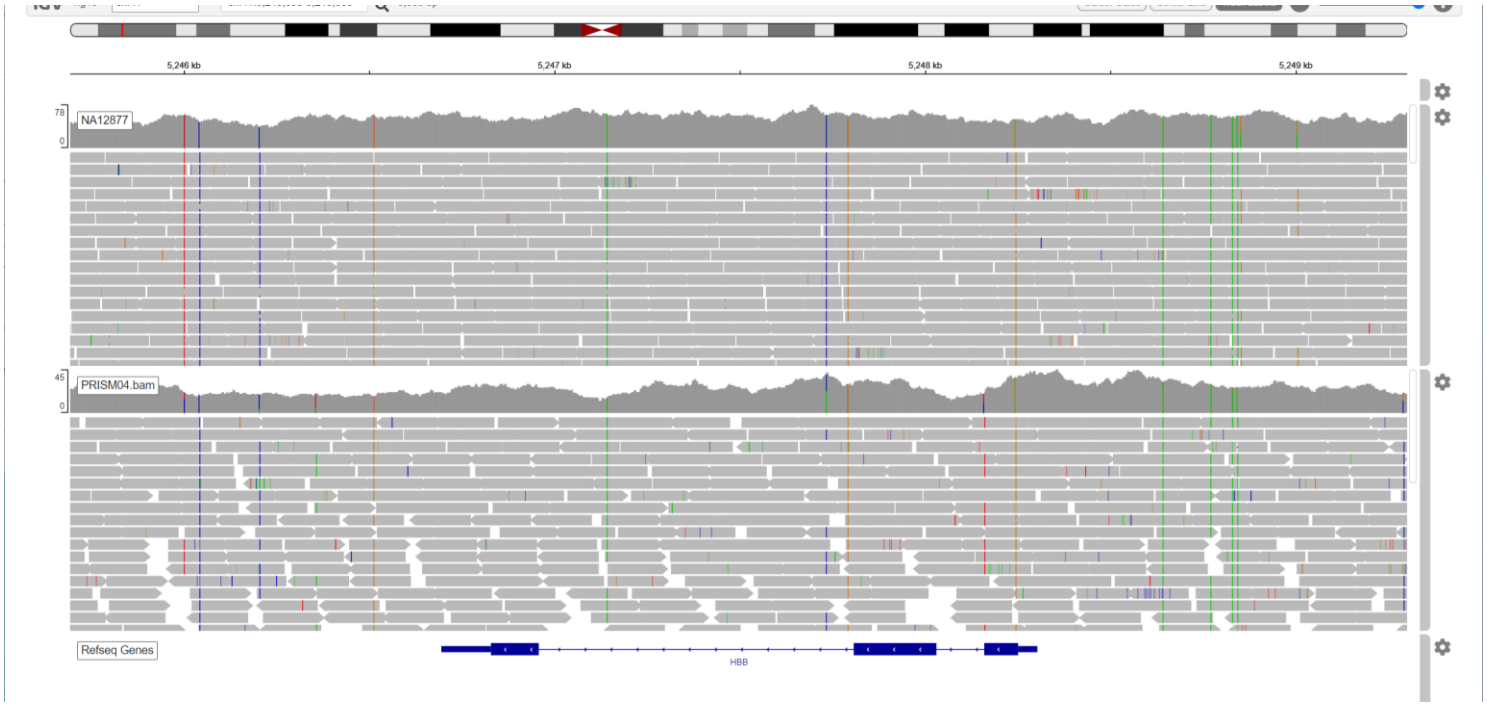
“fishbones” indicate direction of transcription

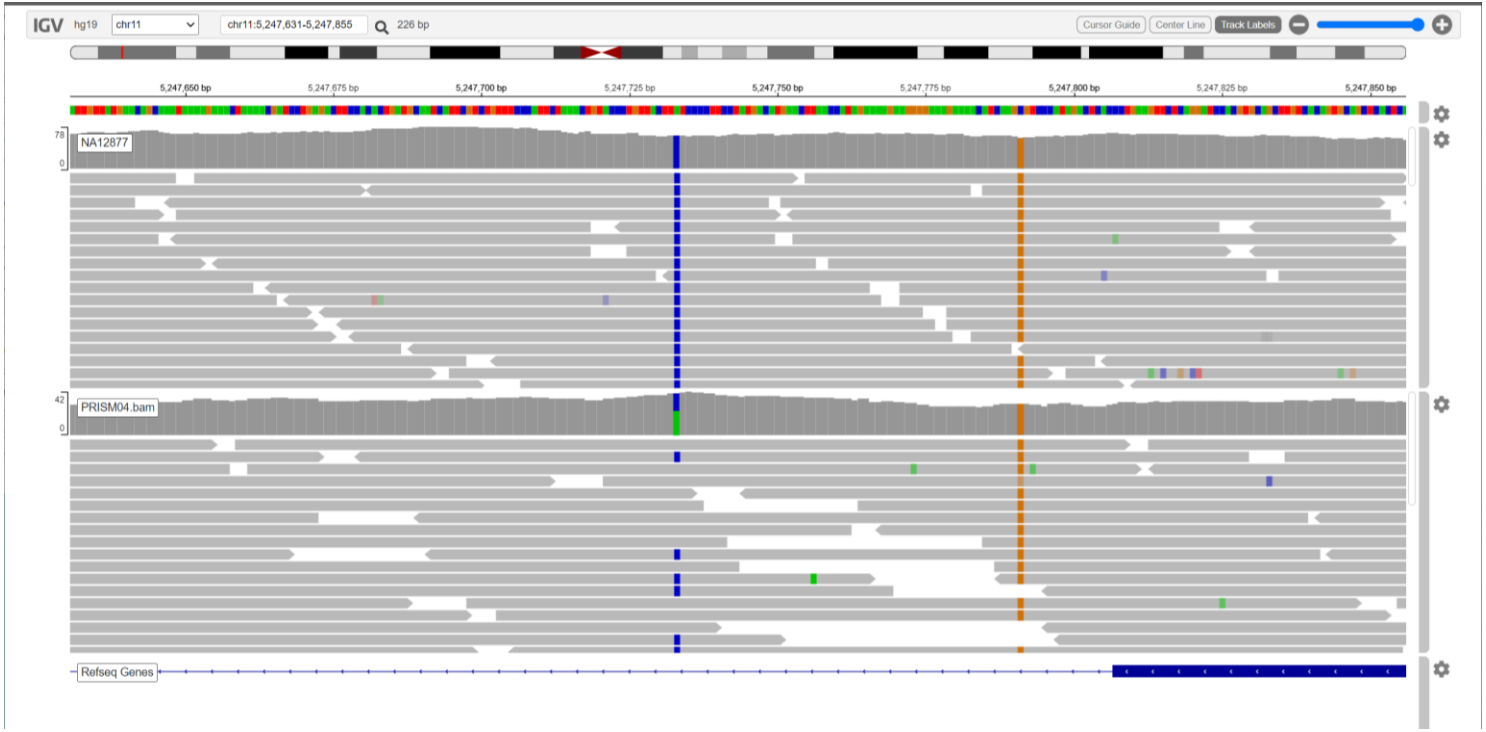
HBB (haemoglobin B) gene 3 exons

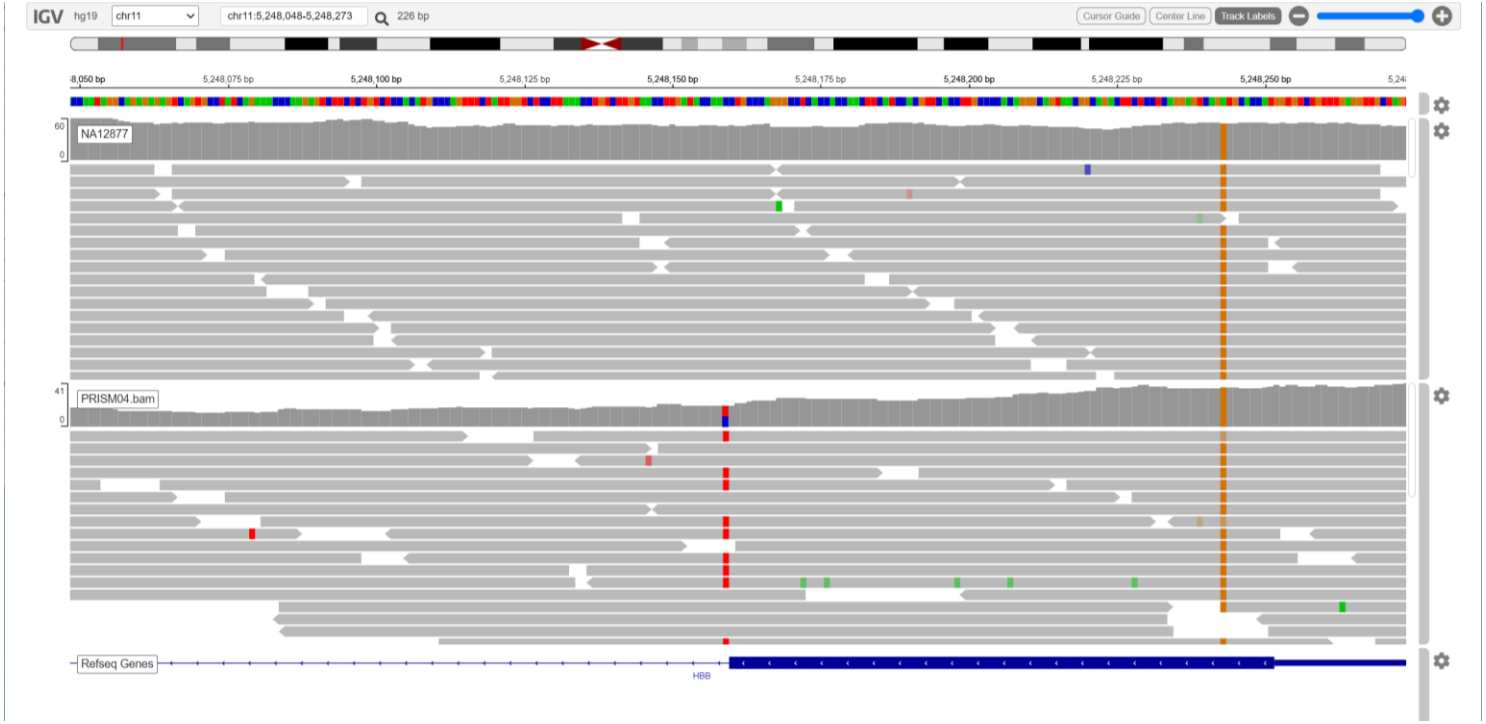


DukeNUS Medical School

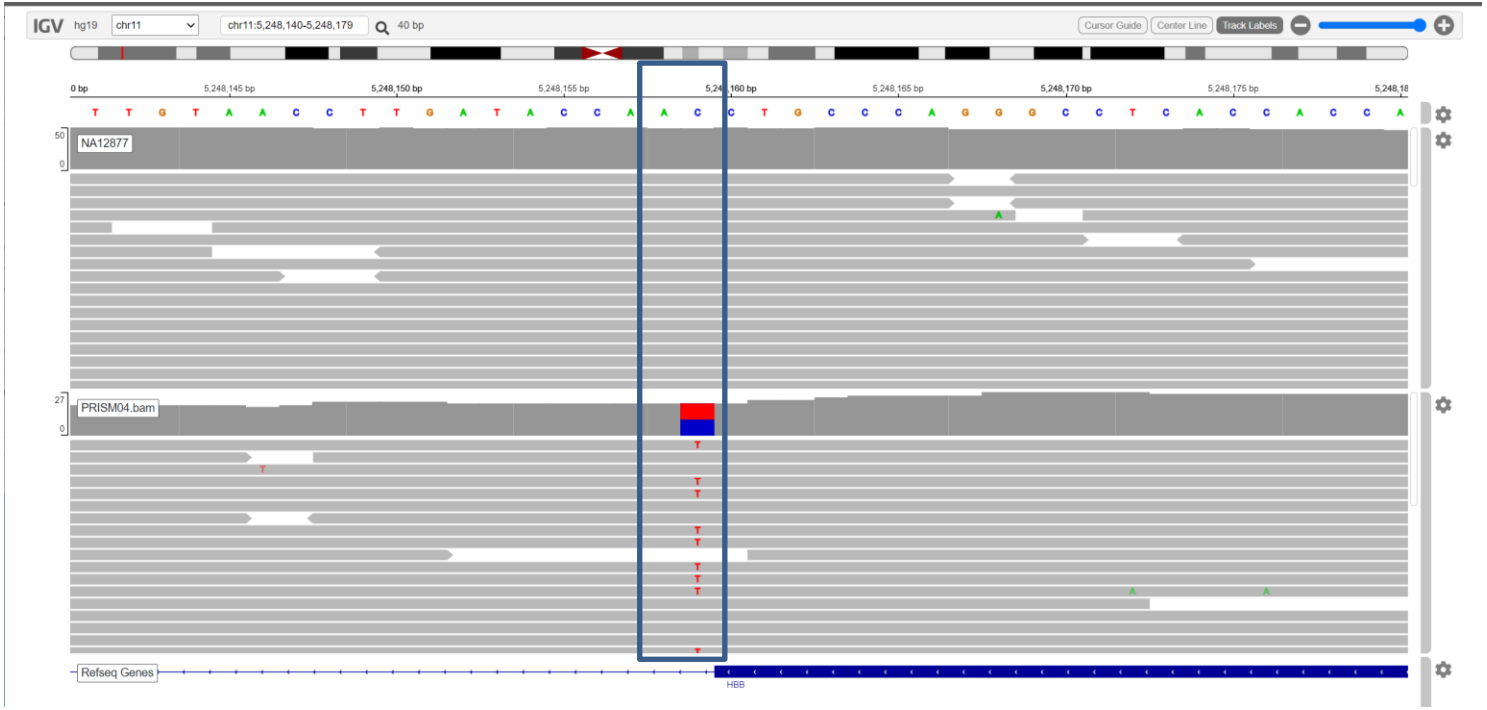

IGV
igv-webapp/resources at master · igv
igv.org/app/
Rain Areas | 240km...
NUS Libraries Proxy...
Genome
Tracks
Session
Share
Bookmark
Save SVG
Help
IGV
hg19
chr11
chr11:5,245,693-5,249,300
3,609 bp
Cursor Guide
Center Line
Track Labels
5,246 kb
5,247 kb
5,248 kb
5,249 kb
NA12877
Refseq Genes
HBB
Set 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

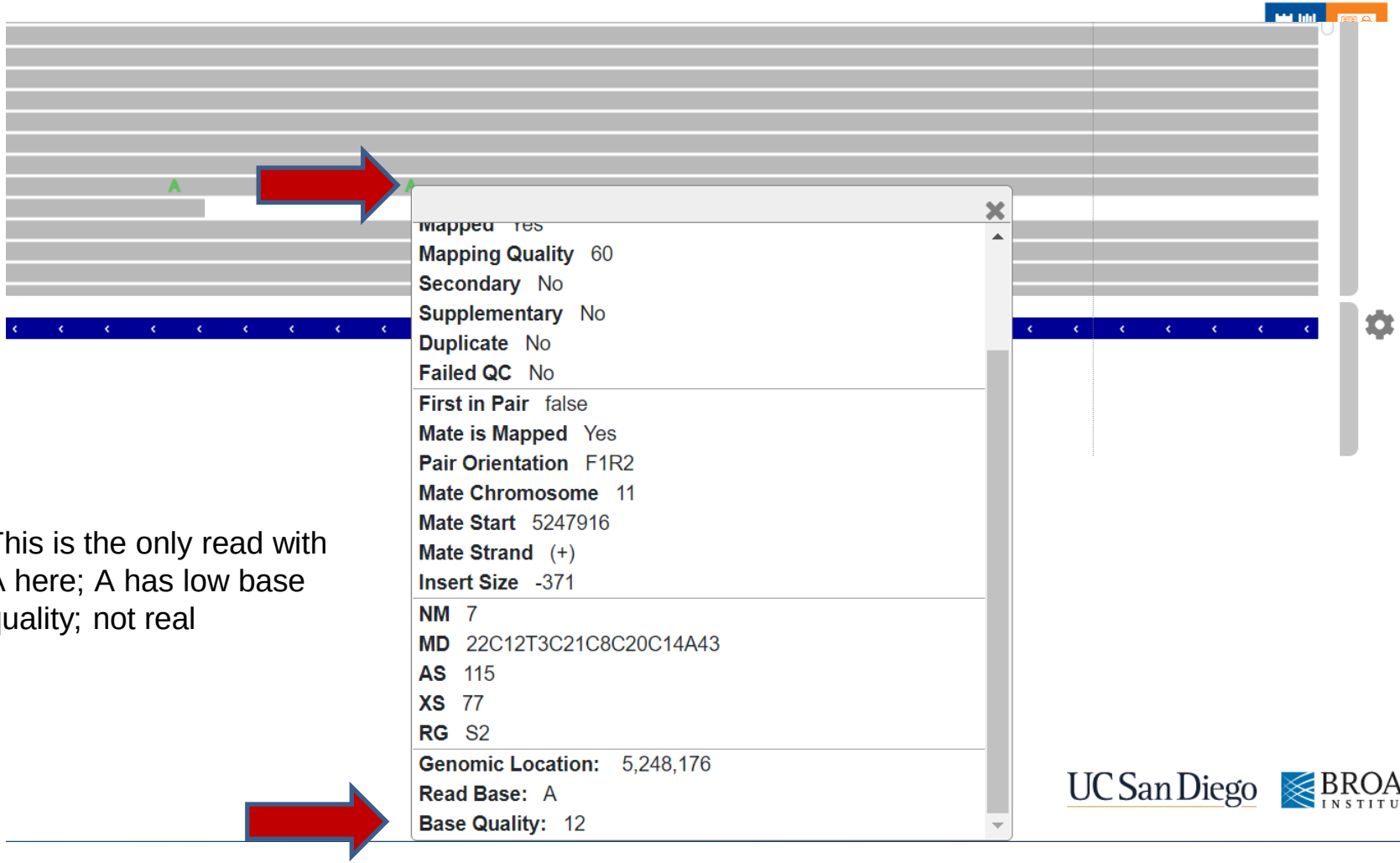
Heterozygous mutation at 3’ splice site of intron 2 of HBB (beta haemoglobin), “beta-thalassemia trait”



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

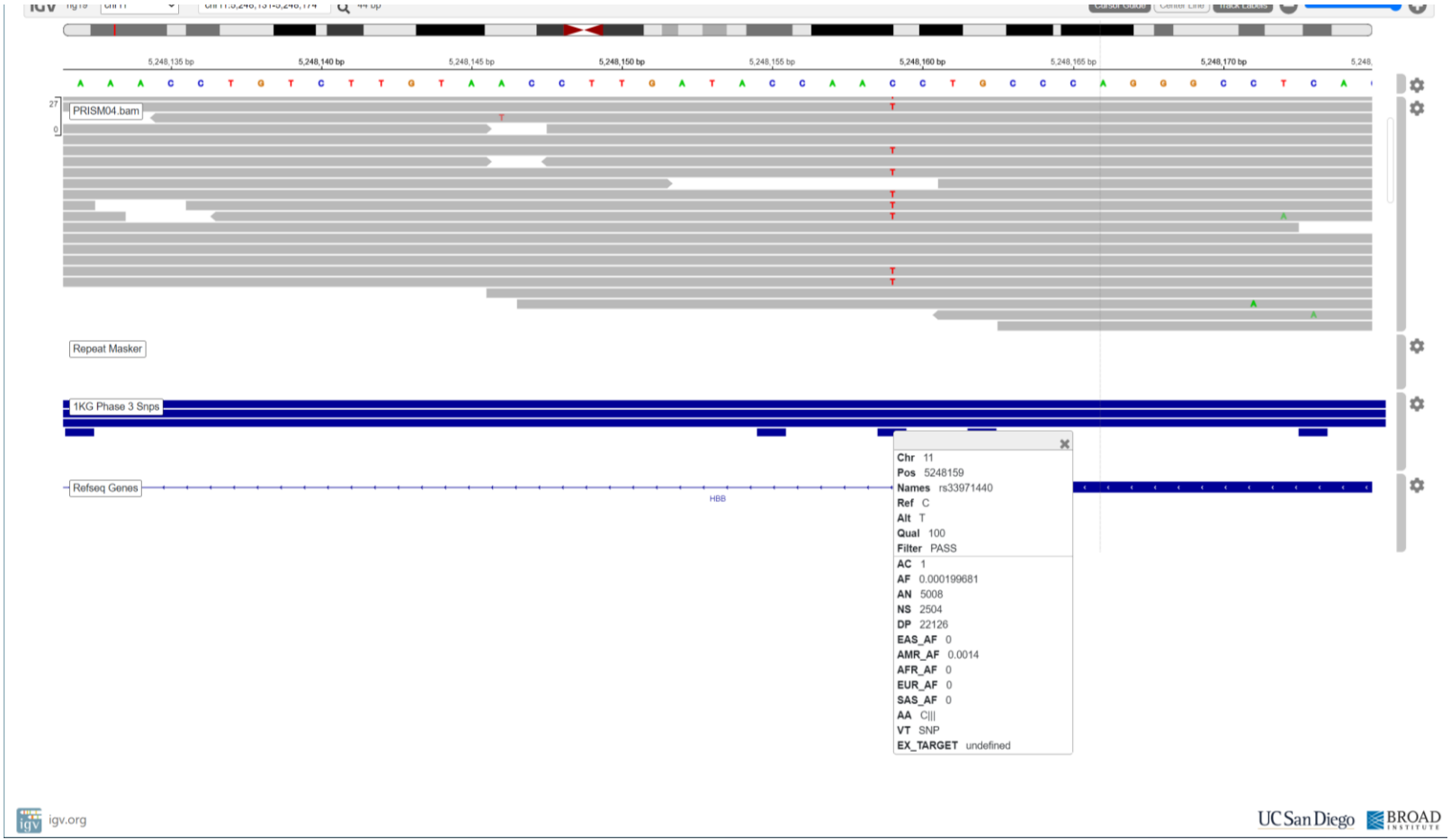
3′-TG-5′ → 3′-TA-5′
5′-AC-3′ → 5′-AT-3′

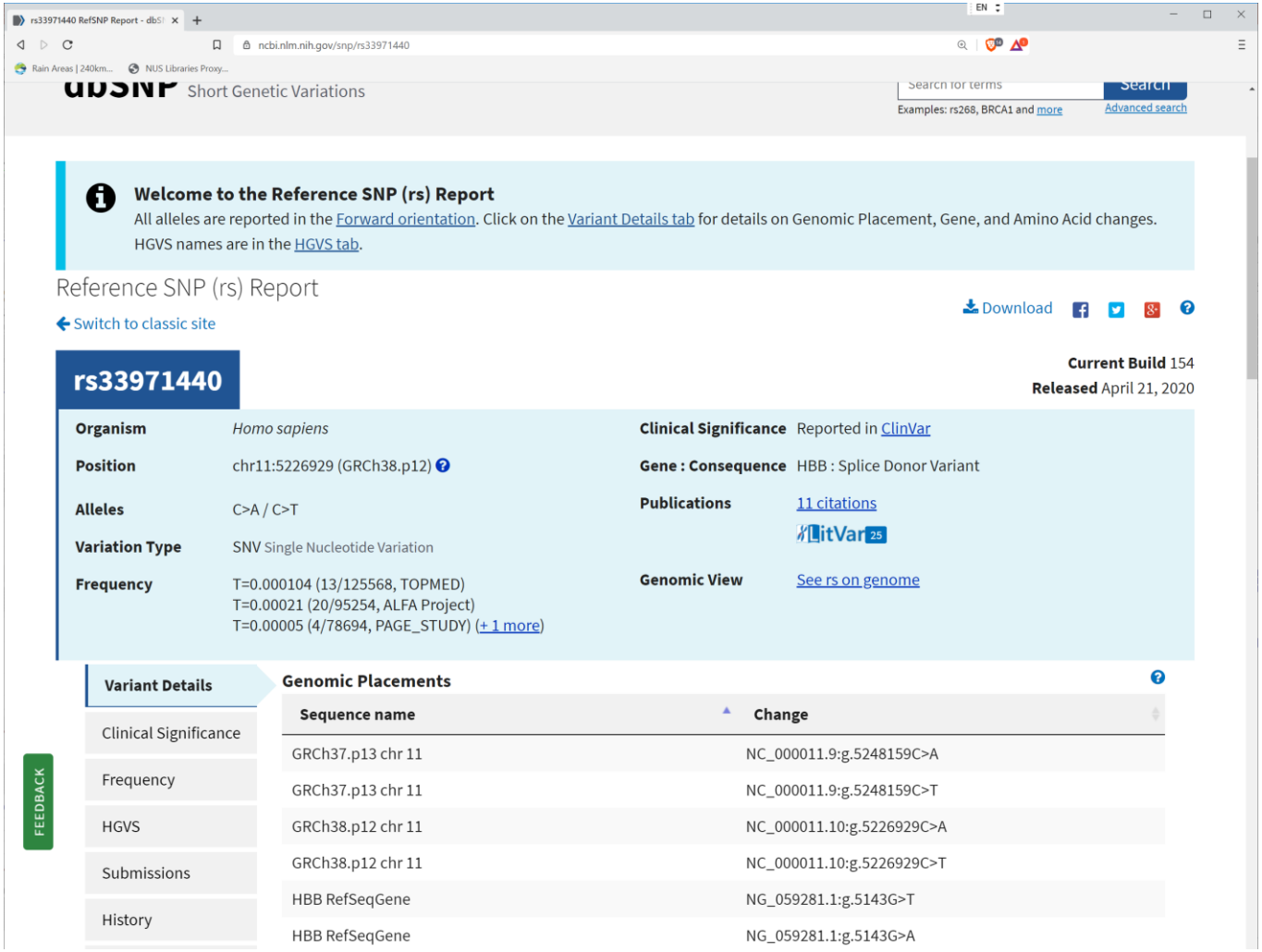
Info on one read and one base in that read
High base quality

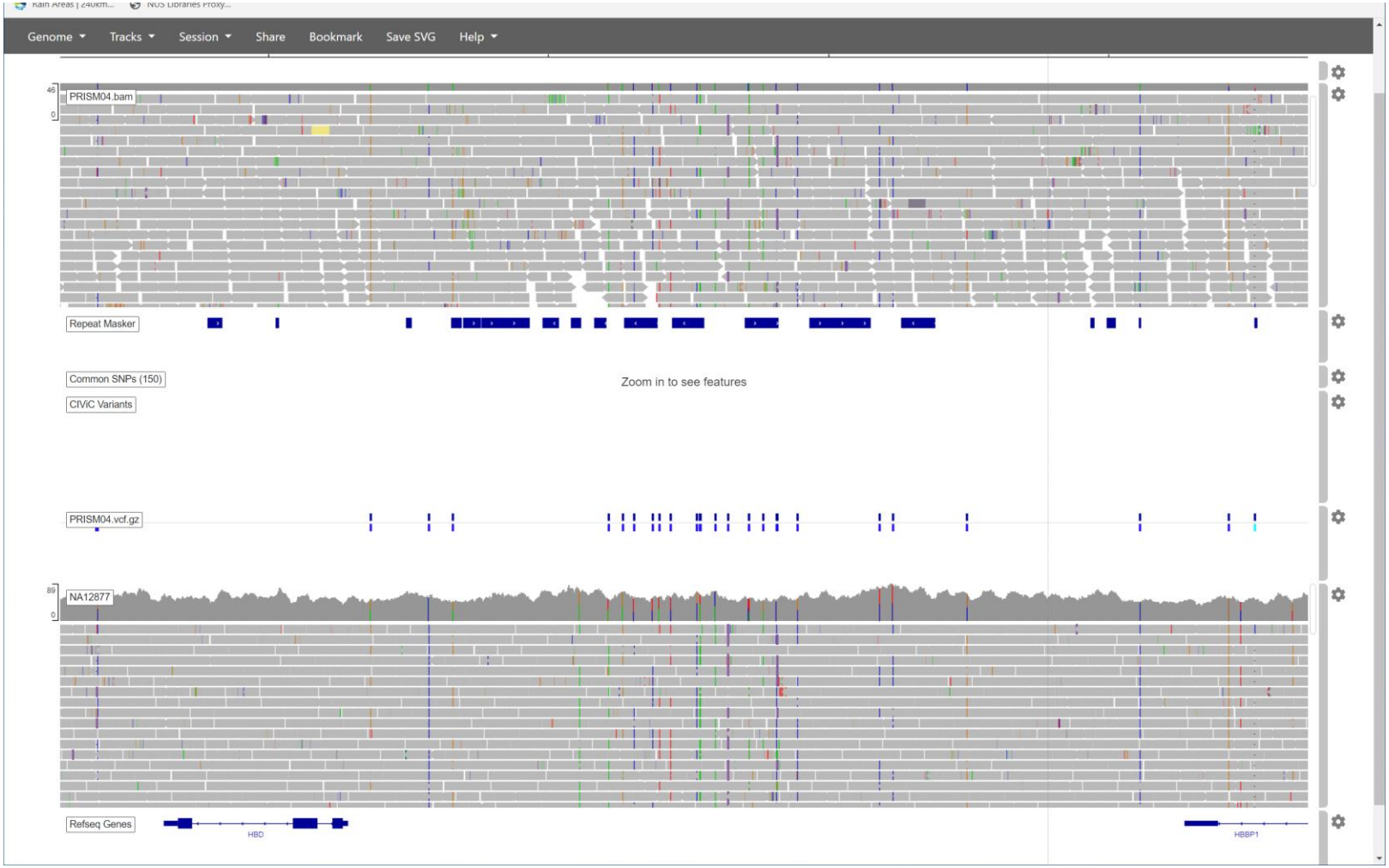


This is the only read with A here; A has low base quality; not real

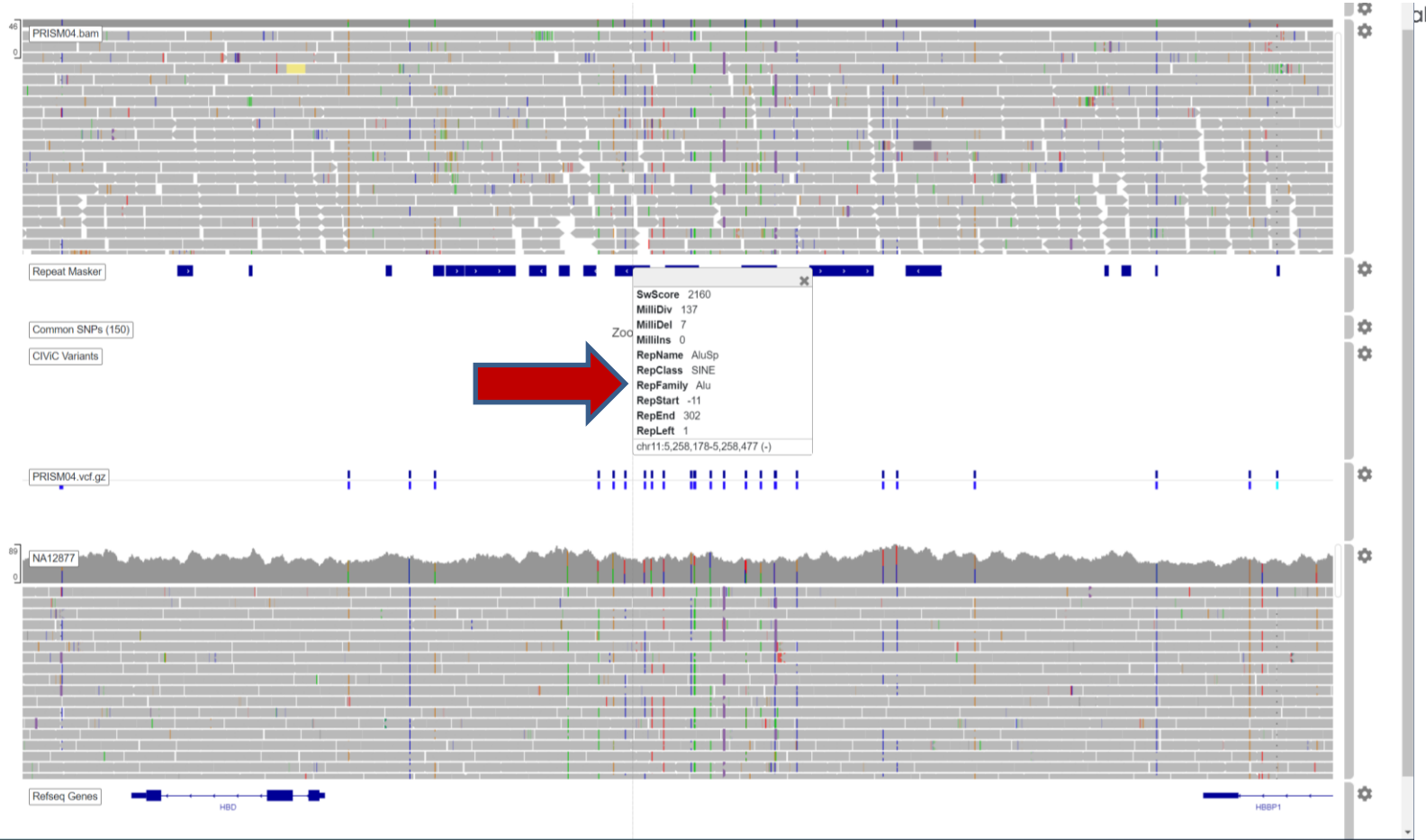
UC San Diego BROAD INSTITUTE

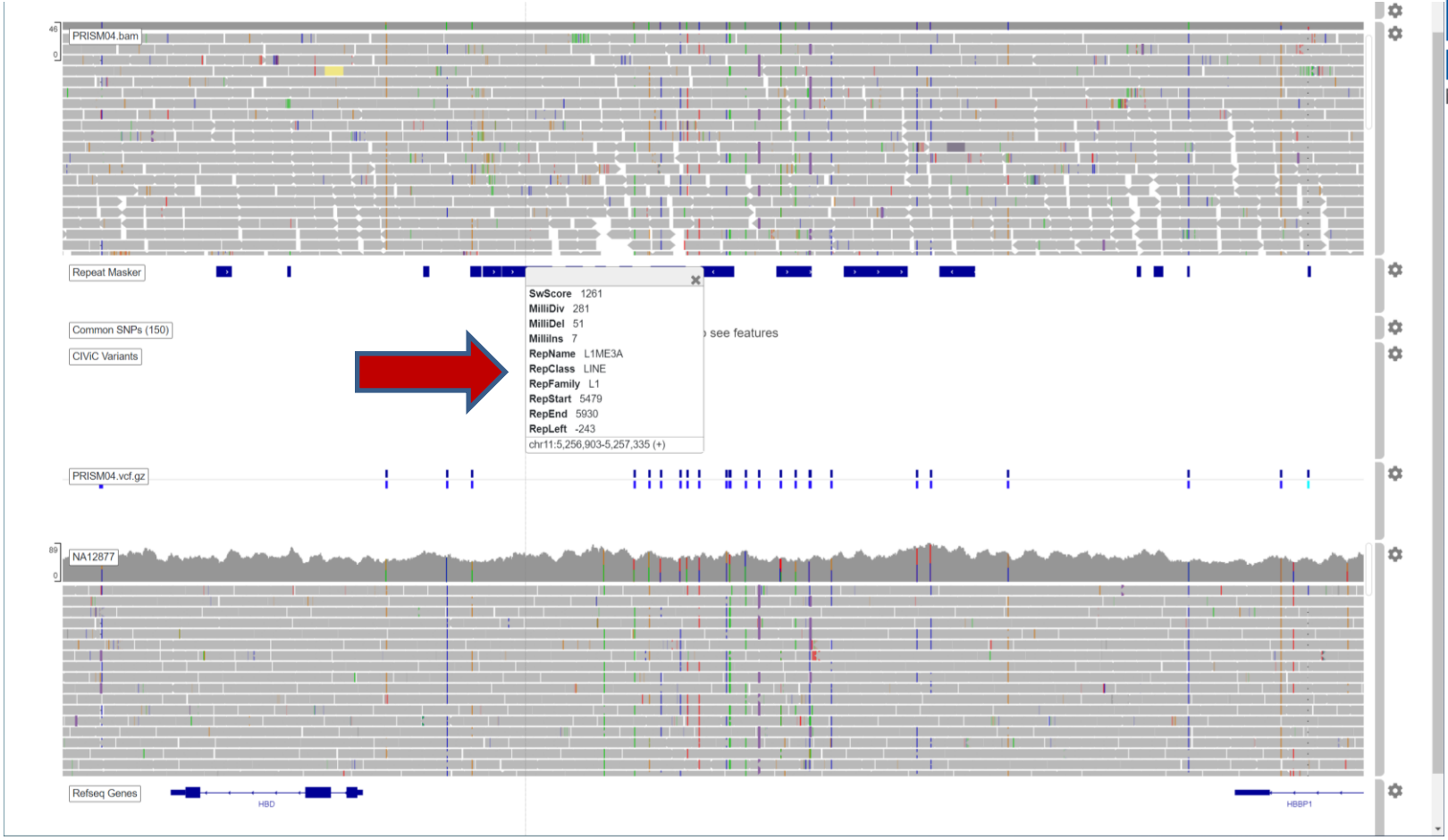

5,248,135 bp
5,248,140 bp
5,248,145 bp
5,248,150 bp
5,248,155 bp
5,248,160 bp
5,248,165 bp
5,248,170 bp
A A A C C T G T C T T G T A A C C T T G A T A C C A A C C T G C C C A G G G C C T C A
PRISM04.bam
Repeat Masker
1KG Phase 3 Snps
Refseq Genes
HBB
Chr 11
Pos 5248159
Names rs33971440
Ref C
Alt T
Qual 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