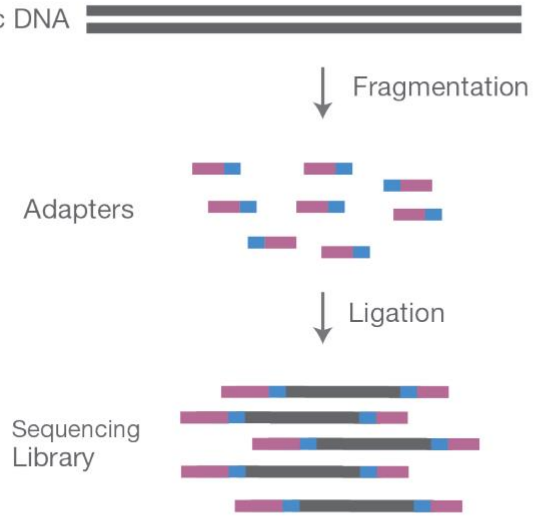


NGS (next generation sequencing) generates short reads, and usually the short reads need to be mapped (aligned) to a reference sequence. For genome sequencing, this is a reference genome.

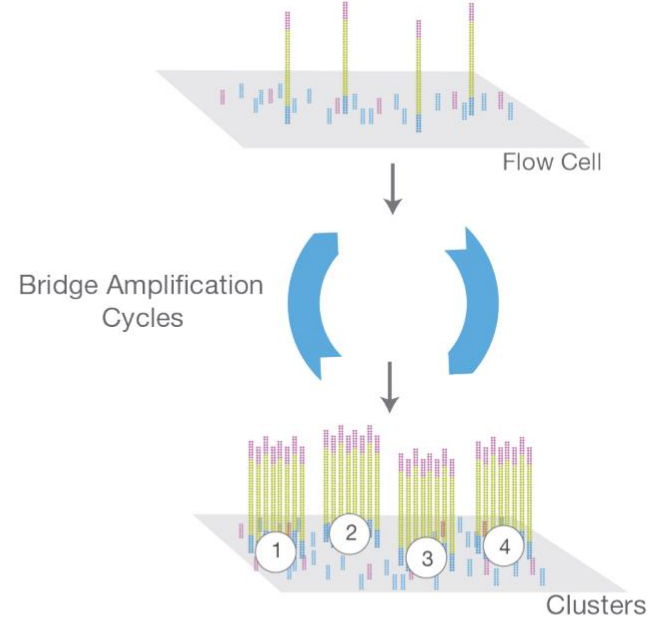
Why does the technology generate short reads?

A. Library Preparation



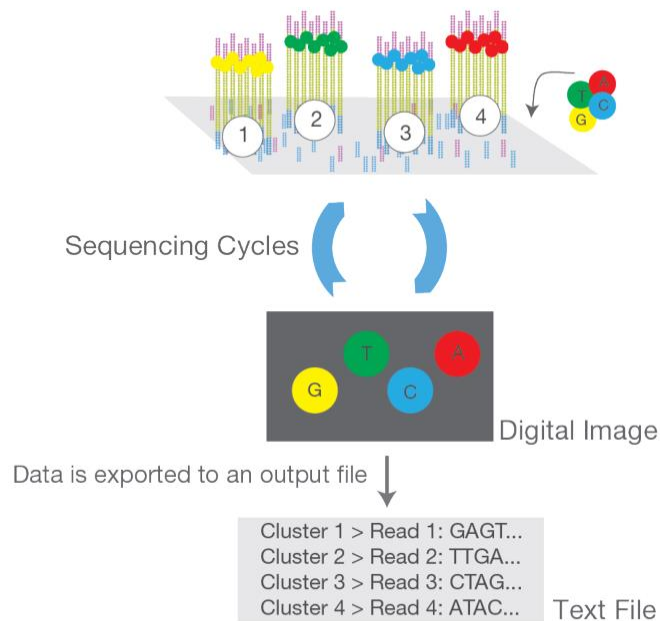
NGS library is prepared by fragmenting a gDNA sample and ligating specialized adapters to both fragment ends.

B. Cluster Amplification



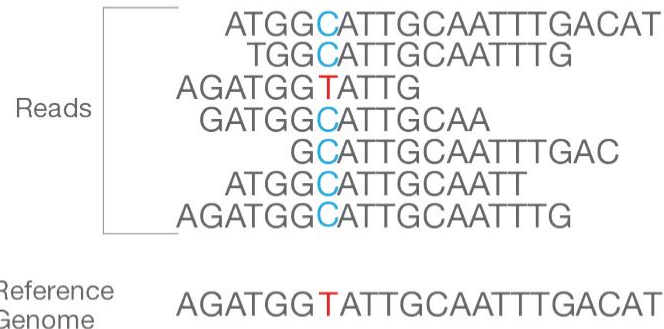
Library is loaded into a flow cell and the fragments are hybridized to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification.

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 Analysis



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.

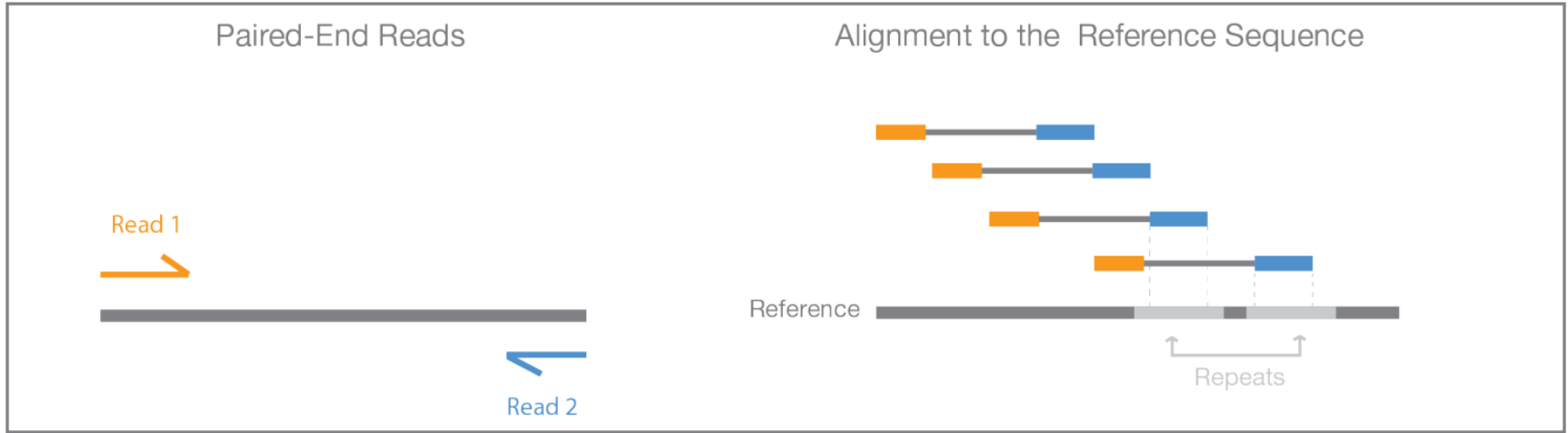


Figure 4: Paired-End Sequencing and Alignment—Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in better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

BWA-MEM

TOPHAT2, STAR2

ELAND (from Illumina)

and a few others

Read mappers

in wide use

STAR2

(from Illumina)

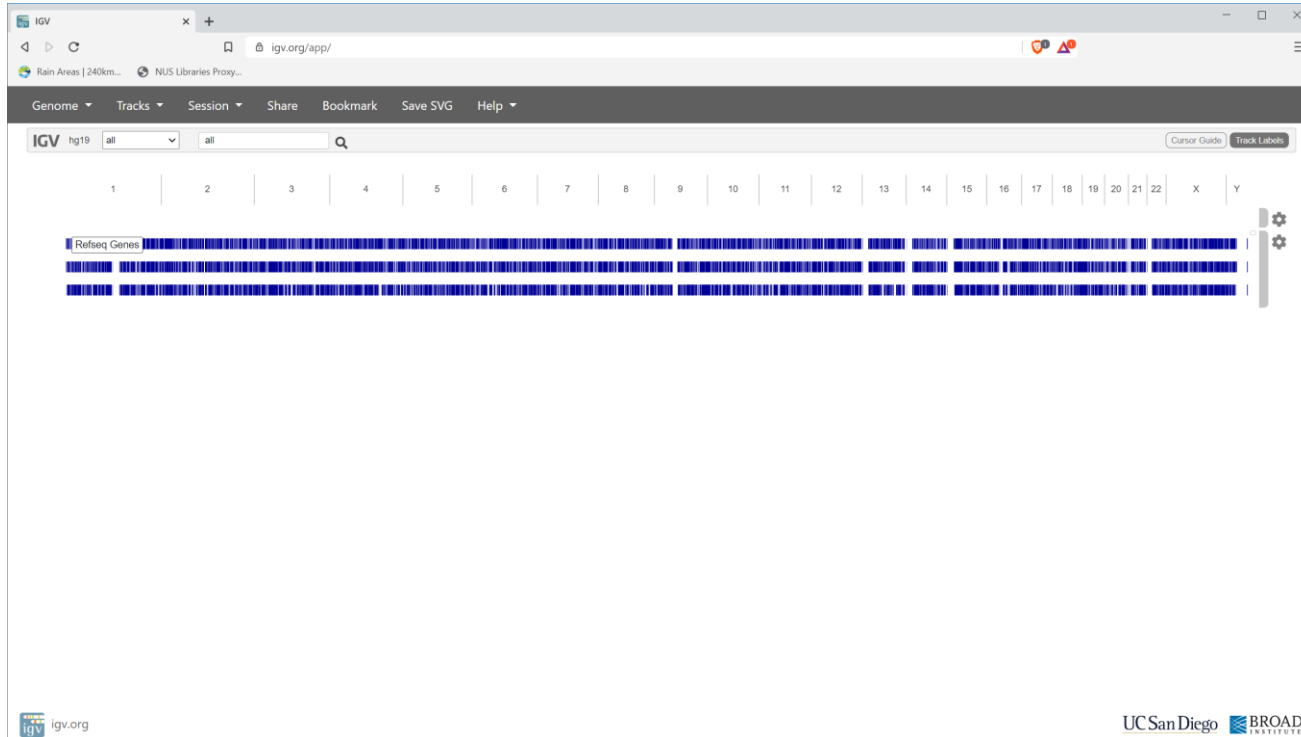
others

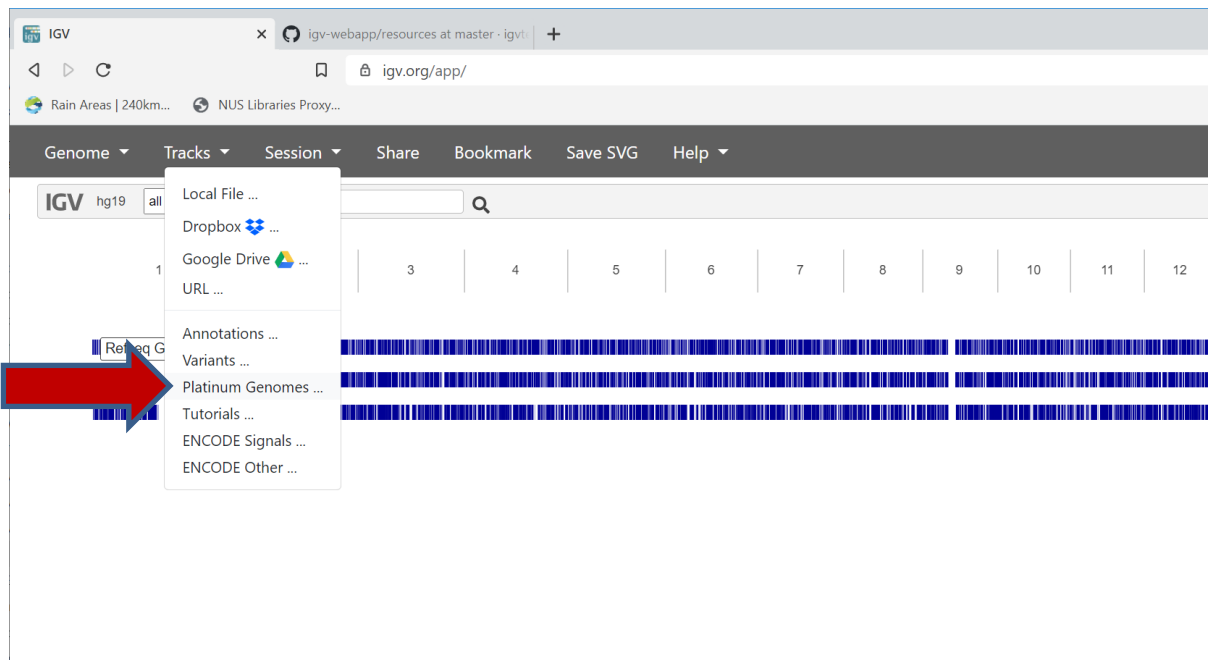
This image is dead, but see www.biorxiv.org/wiki/List_of_sequence_mapping_software#Short-read_mapping

Years

IGV (genome assembly browser)

- <https://igv.org/app>





IGV hg19 all

Local File ...

Dropbox ...

Google Drive ...

URL ...

Annotations ...

Variants ...

Platinum Genomes ...

Tutorials ...

ENCODE Signals ...

ENCODE Other ...

IGV hg19 all all

Genome Tracks Session Share Bookmark Save SVG Help

1 2 3

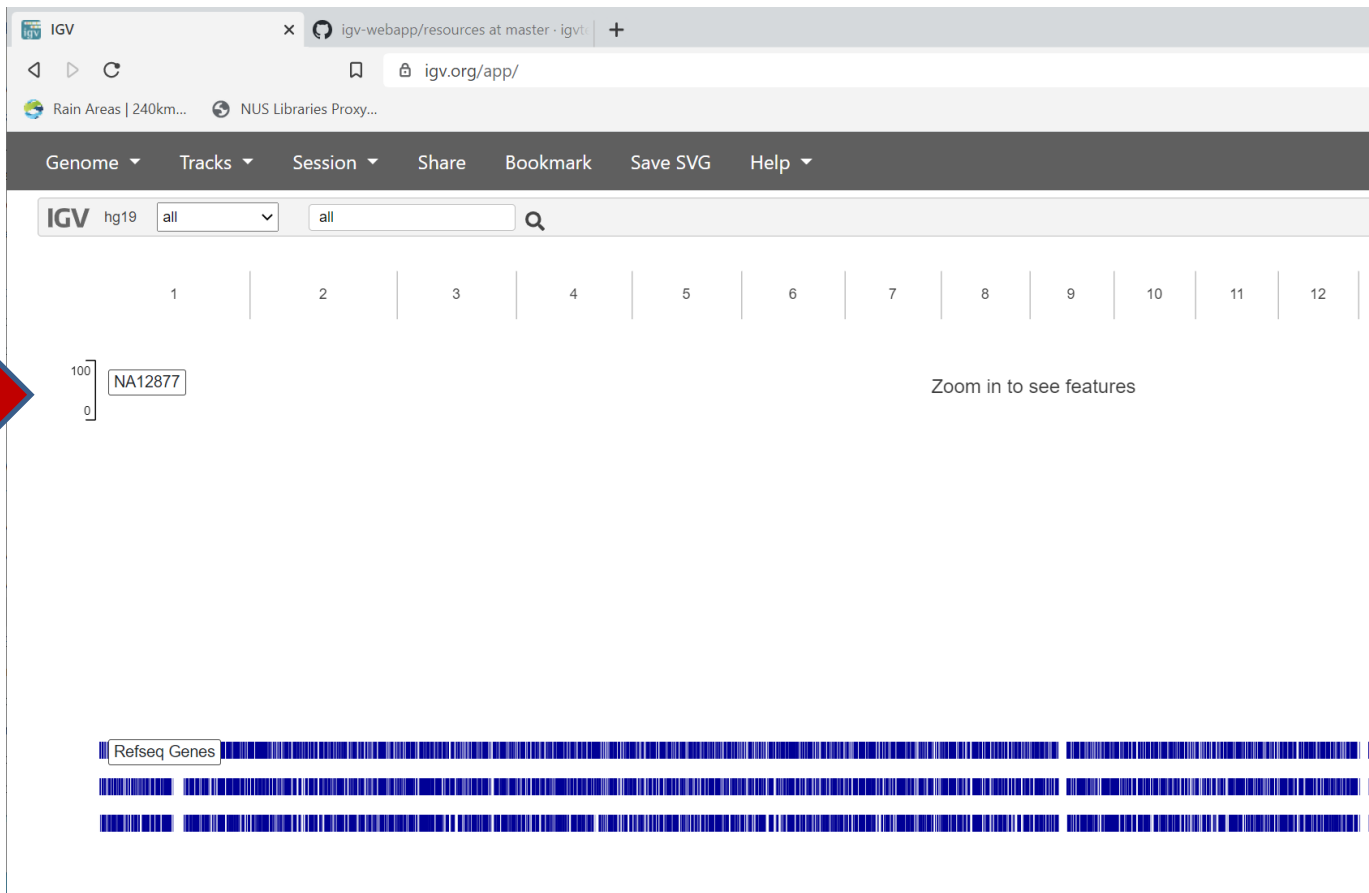
Refseq Genes

Platinum Genomes

- NA12877
- NA12878
- NA12889
- NA12890
- NA12891
- NA12892

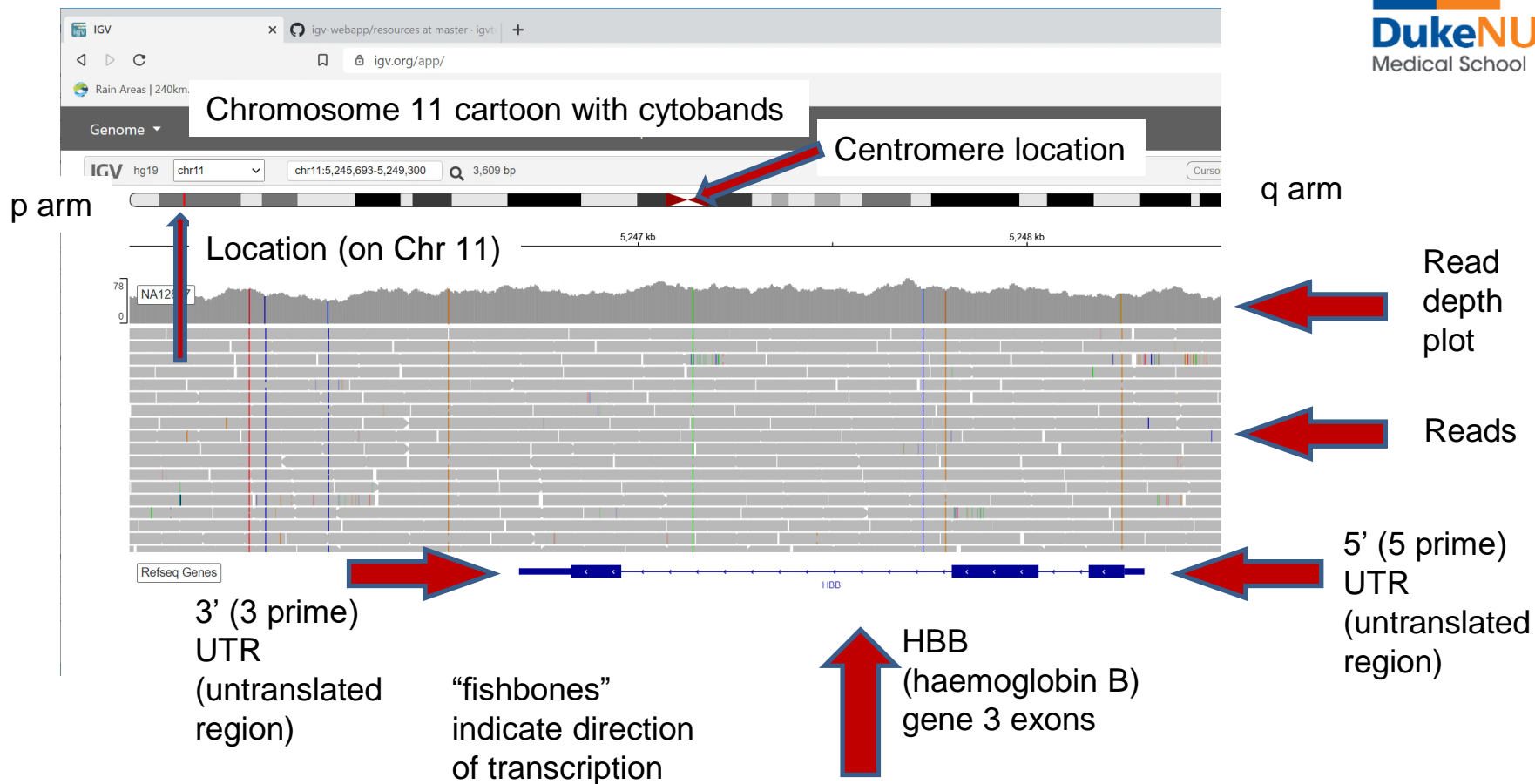
Illumina Platinum Genomes - source [Google public datasets](#)

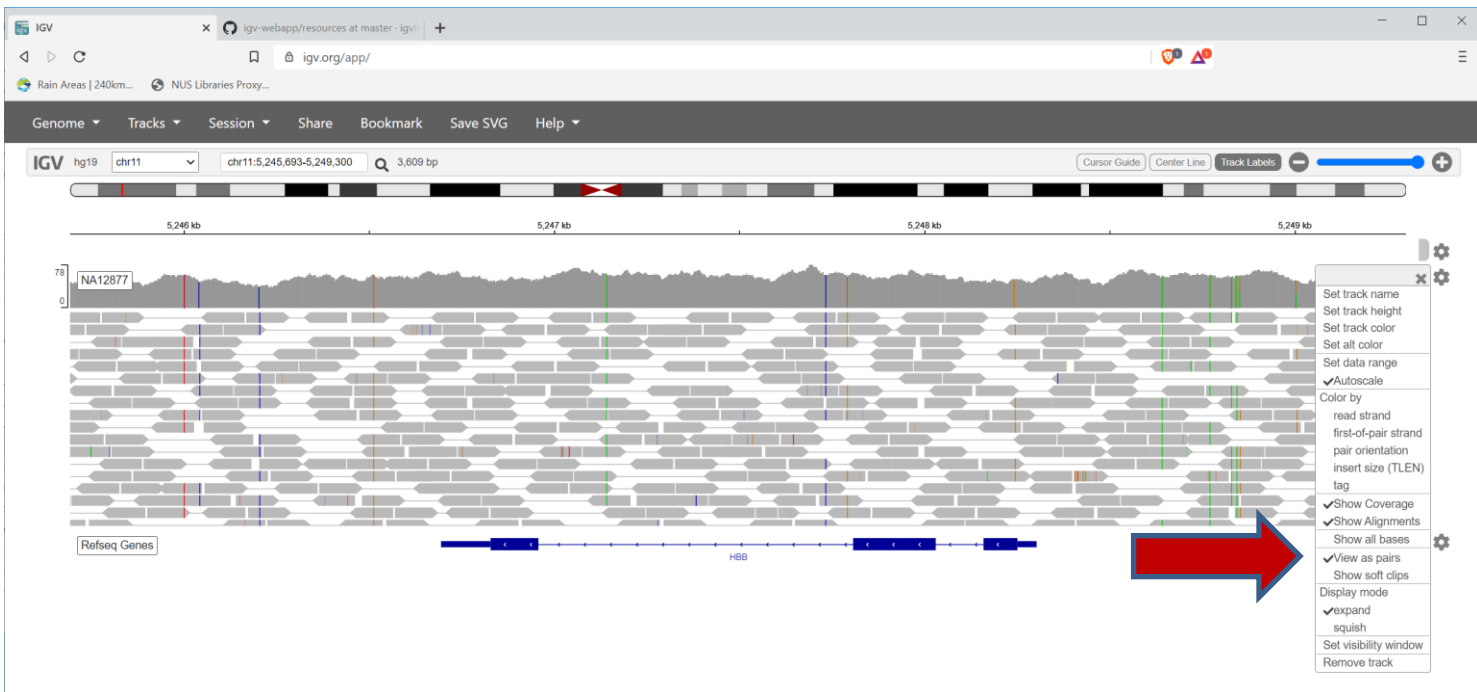
Cancel OK

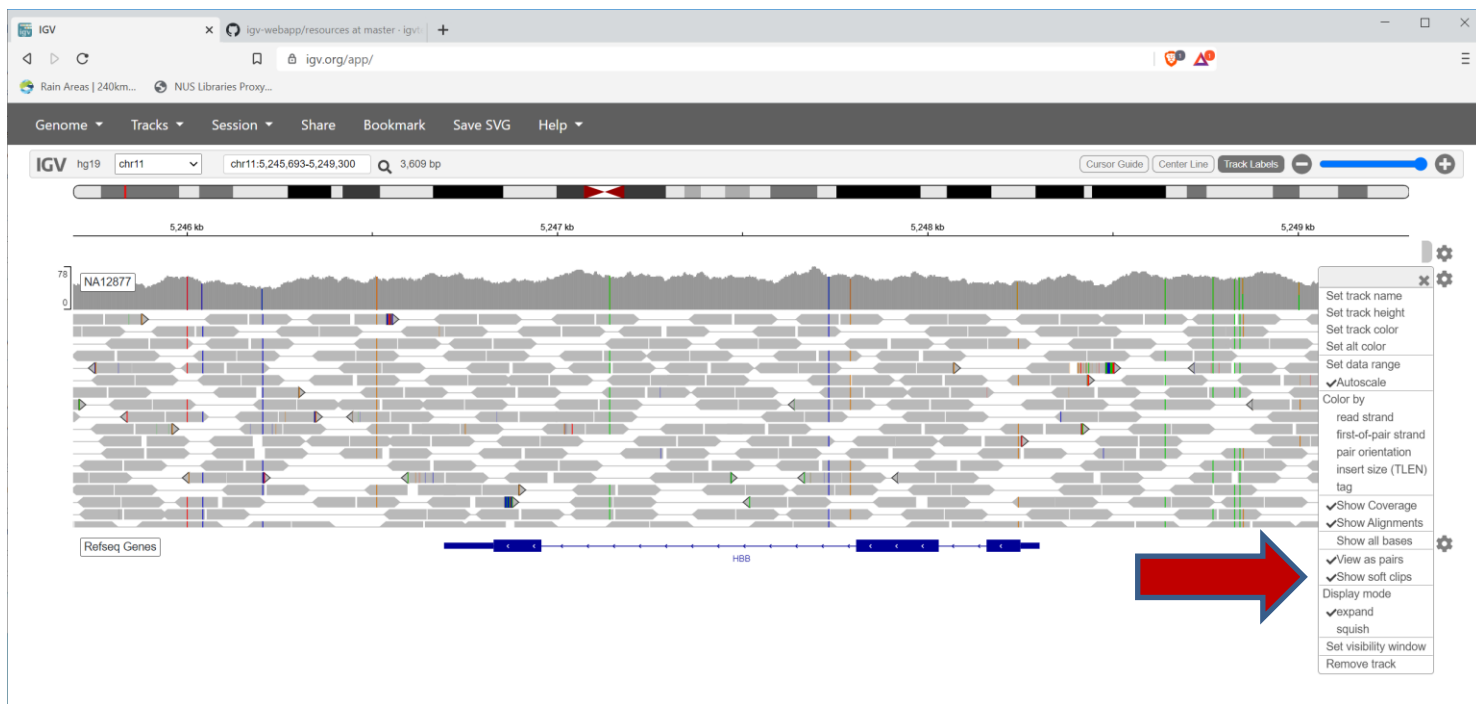


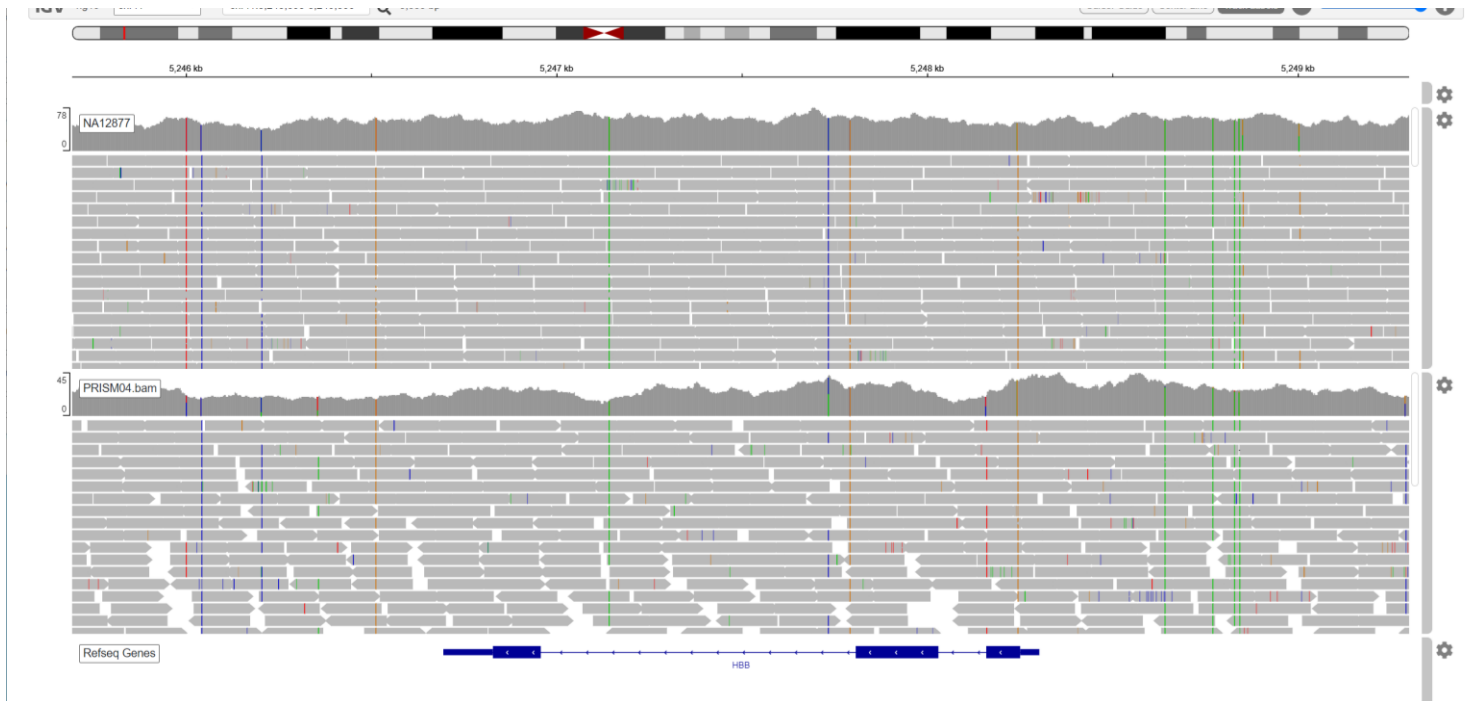
Sample ID
NA12877

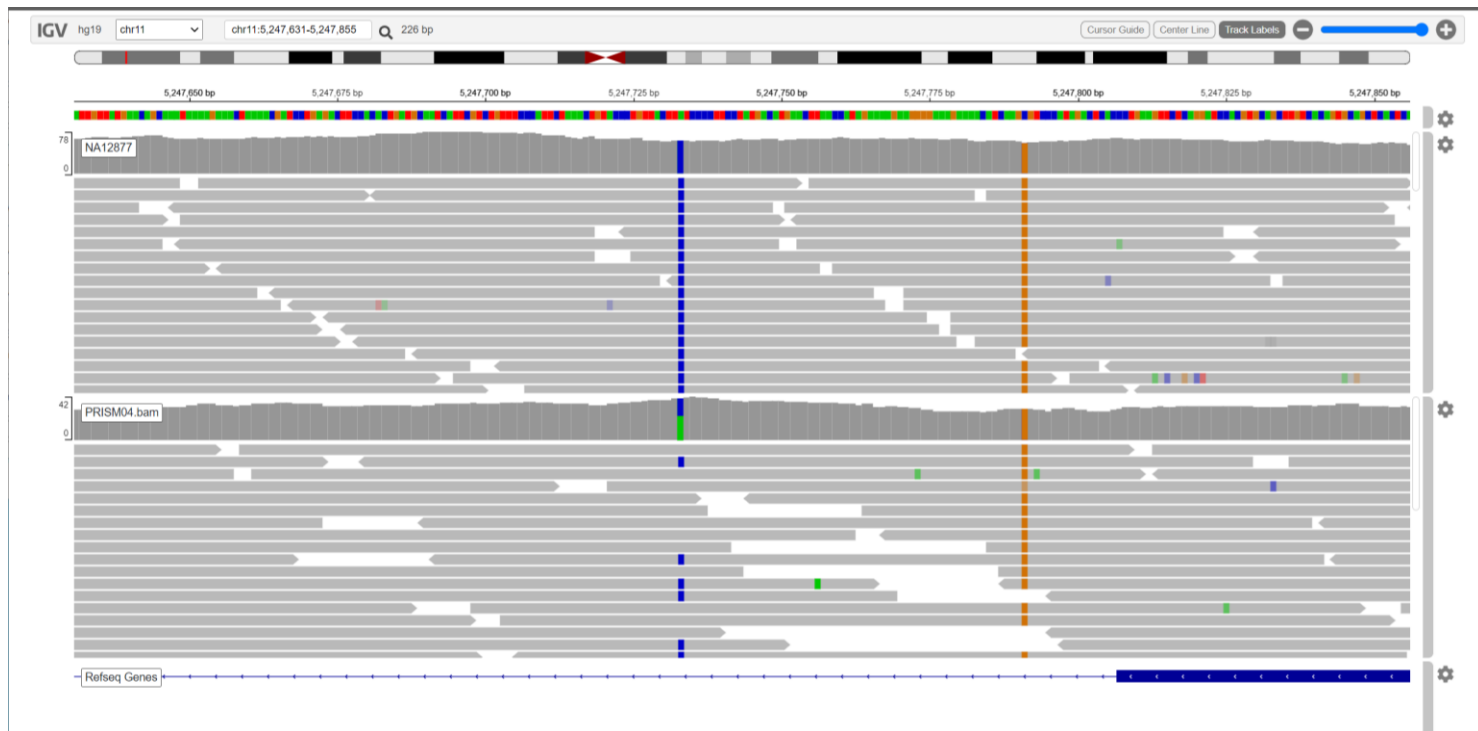


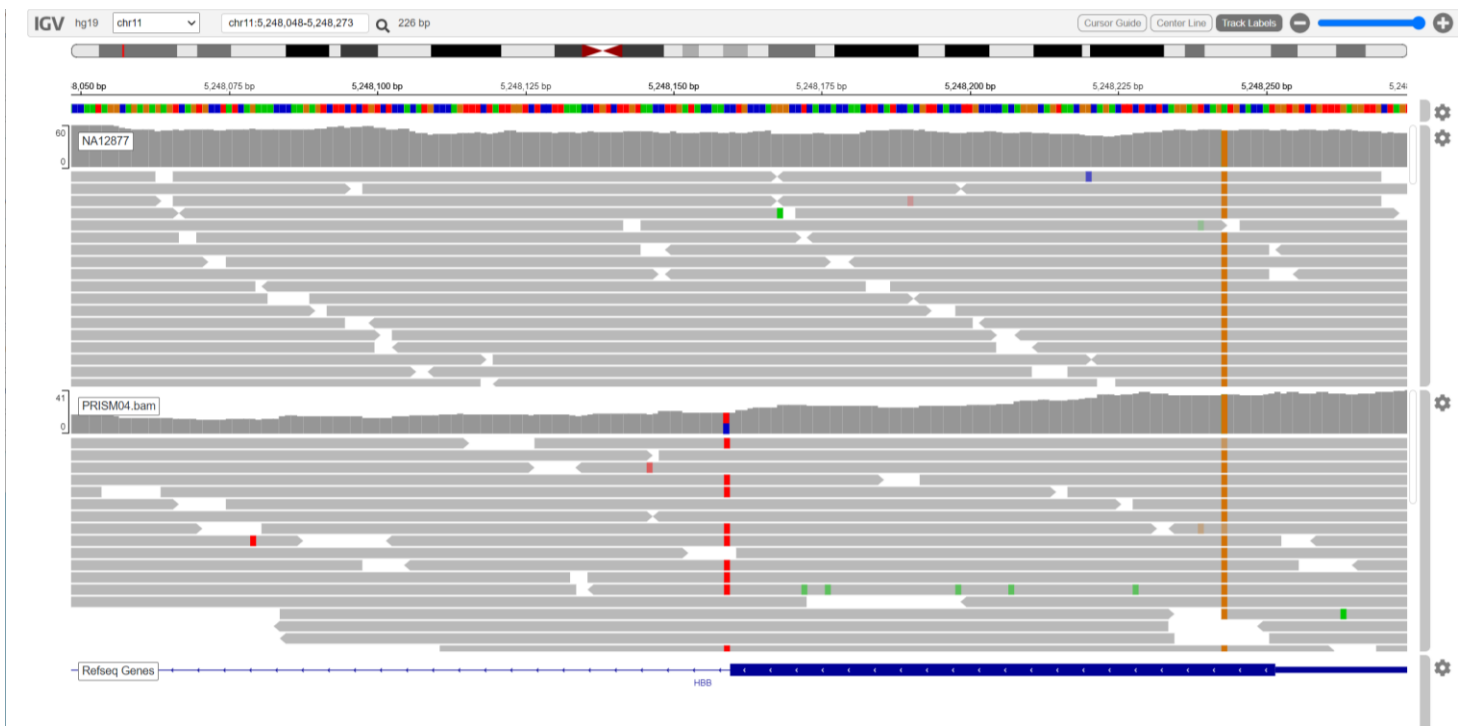












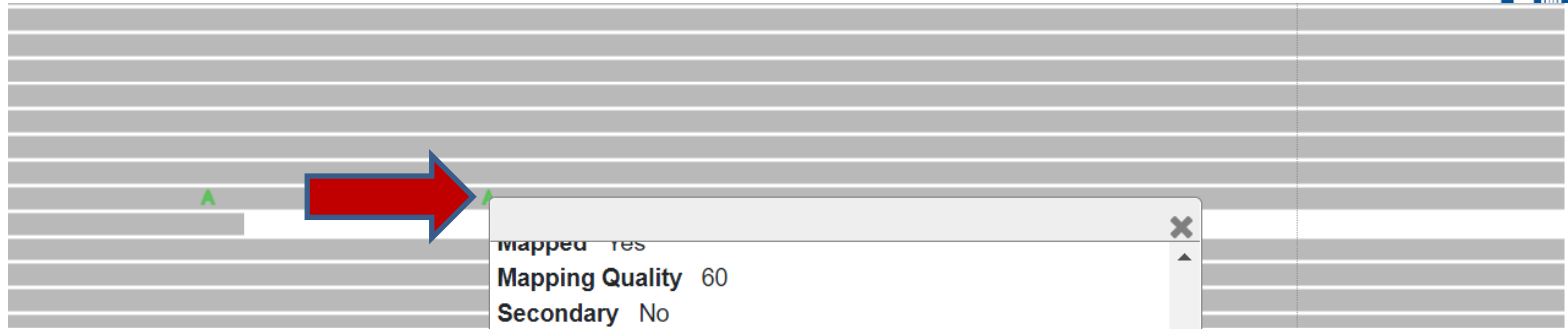
Heterozygous mutation at 3' splice site of intron 2 of HBB (beta haemoglobin), “beta-thalassemia trait”



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'





×▲

mapped	yes
Mapping Quality	60
Secondary	No
Supplementary	No
Duplicate	No
Failed QC	No
First in Pair	false
Mate is Mapped	Yes
Pair Orientation	F1R2
Mate Chromosome	11
Mate Start	5247916
Mate Strand	(+)
Insert Size	-371
NM	7
MD	22C12T3C21C8C20C14A43
AS	115
XS	77
RG	S2
Genomic Location:	5,248,176
Read Base:	A
Base Quality:	12

▼

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



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

i

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

[Switch to classic site](#)

[Download](#)
[f](#)
[t](#)
[g+](#)
[?](#)

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

Genomic Placements ?

Clinical Significance

Frequency

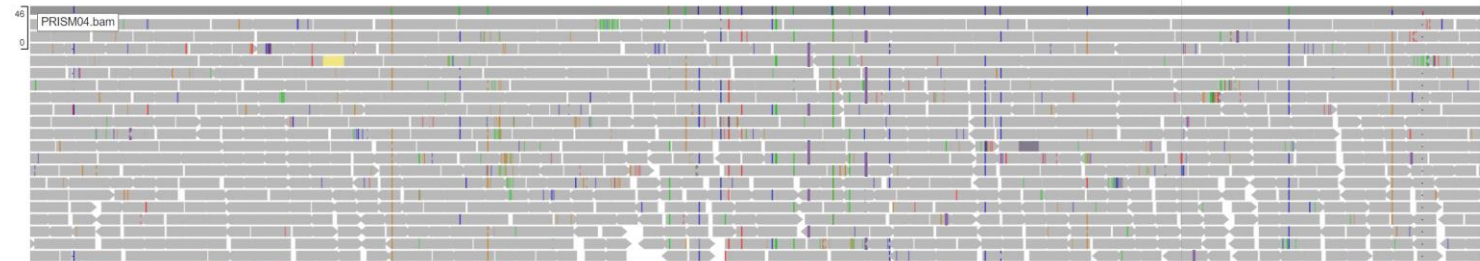
HGVS

Submissions

History

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



Repeat Masker

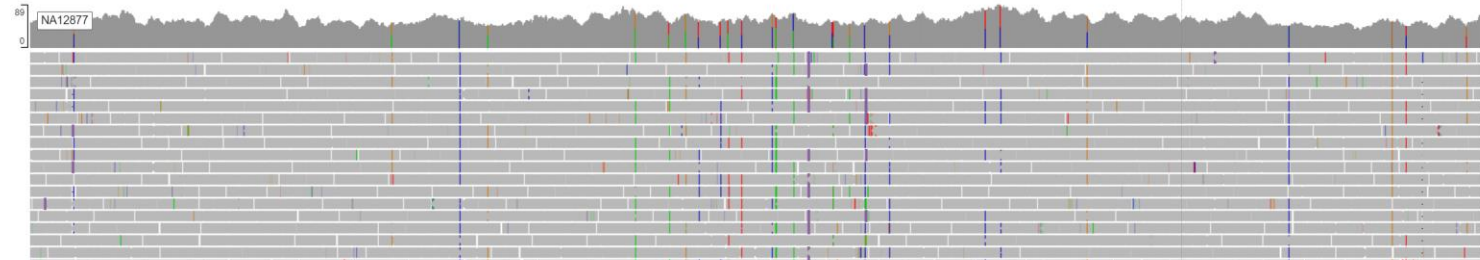
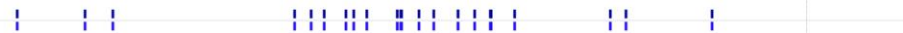


Common SNPs (150)

Zoom in to see features

CIVIC Variants

PRISM04.vcf.gz



Refseq Genes



