RNA-Seq (2)

BCB 5250 Introduction to Bioinformatics II

Spring 2020

Tae-Hyuk (Ted) Ahn
Department of Computer Science
Program of Bioinformatics and Computational Biology
Saint Louis University

SAINT LOUIS
UNIVERSITY
EST. 1818

Tools:

- Data Quality Control
- Read Mapping
- Differential Gene Expression

Data Quality Control

Quality assessment

- FastQC, MultiQC

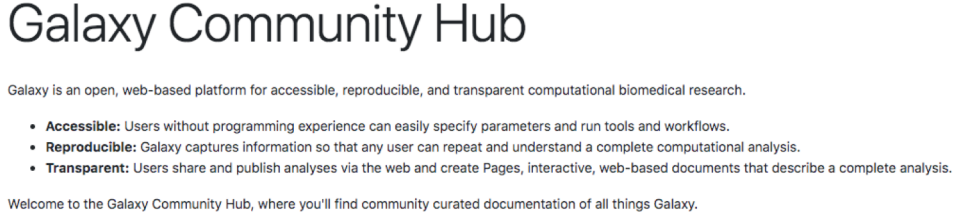
Trim and filtering:

- FASTX Tool Kit: http://hannonlab.cshl.edu/fastx_toolkit/
- Trimmomatic: http://www.usadellab.org/cms/?page=trimmomatic
- Tim Galore: http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
- fastp: https://github.com/OpenGene/fastp
- BBTools – BBDuk: https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/
- Cutadapt: https://cutadapt.readthedocs.io/en/stable/

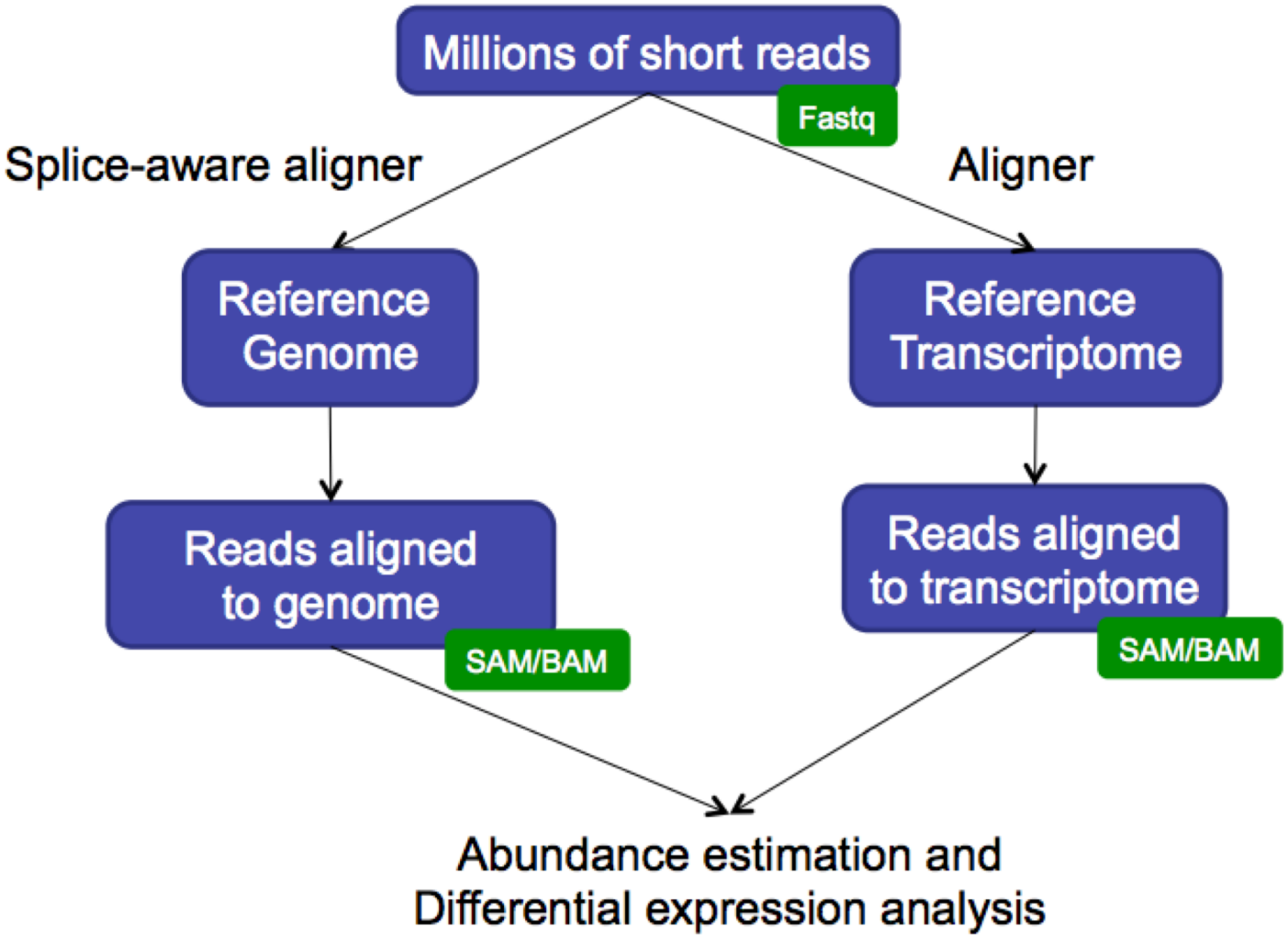
Error correction:

- BBTools – Tadpole: https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/tadpole-guide/
- https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1784-8
- For long sequencing, check the recently published papers and archive such as http://www.pacb.com/asset_tags/error-correction/

https://galaxyproject.org/



Mapping Reads



Aligner

- Short Read Mapper
  - Bowtie2: http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
  - BWA: http://bio-bwa.sourceforge.net/
- Slice Aligner
  - TopHat2: https://ccb.jhu.edu/software/tophat/index.shtml
  - HiSAT2: https://daehwankimlab.github.io/hisat2/
  - STAR: https://github.com/alexdobin/STAR

Mapping

- Input
  - Fastq files
  - Index of genome/transcriptome
  - Annotation file (optional for some, but required for others)
- Output
  - SAM (text) / BAM (binary) alignment files
    - SAMtools – SAM/BAM file manipulation (http://samtools.sourceforge.net/) (http://www.htslib.org/)
    - Picard-tools – SAM/BAM file manipulation (https://broadinstitute.github.io/picard/)
  - Summary statistics (per read library)
    - reads with unique alignment
    - reads with multiple alignments
    - reads with no alignment
    - reads properly paired (for paired-end libraries)

Most Aligner

Most aligners have modules for below two steps:

1. Generating genome indexes files.
2. Mapping reads to the genome.

Why?

Let us test bowtie2 aligner

1. Open manual (http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#getting-started-with-bowtie-2-lambda-phage-example)
2. Let us work "Getting Started"
   - Copy the genome to your lab directory and index a reference genome
     - /public/ahnt/courses/bcb5250/rna_seq_lab/bowtie2_example/reference/lambda_virus.fa
   - Aligning example reads
     - /public/ahnt/courses/bcb5250/rna_seq_lab/bowtie2_example/reads/reads_1.fq
   - Paired-end example
     - reads_1.fq, reads_2.fq
   - From the paired-end aligning output, let us test SAMtools/BCFtools downstream analysis
     - Convert SAM to BAM
     - Sort
     - To generate variant calls in VCF format
     - Then to view the variants

SAM (BAM) Format

- Sequence Alignment/Map format
  - Universal standard
  - Human-readable (SAM) and compact & binary (BAM) forms
- Structure
  - Header
    - version, sort order, reference sequences, read groups, program/processing history
  - Alignment records

Check header

$ samtools view -H -S XXX.sam

[samopen] SAM header is present: 1 sequences.

@HD VN:1.0 SO:unsorted

@SQ SN:genome LN:451226

@PG ID:bowtie2 PN:bowtie2 VN:2.3.4.3 CL:"/usr/local/miniconda/bin/bowtie2-align-s --wrapper basic-0 -x ../GENOME_data/genome -S Sp_ds.sam -p 12 -t -1 Sp_ds.left.fq.gz -2 Sp_ds.right.fq.gz"

Check alignment

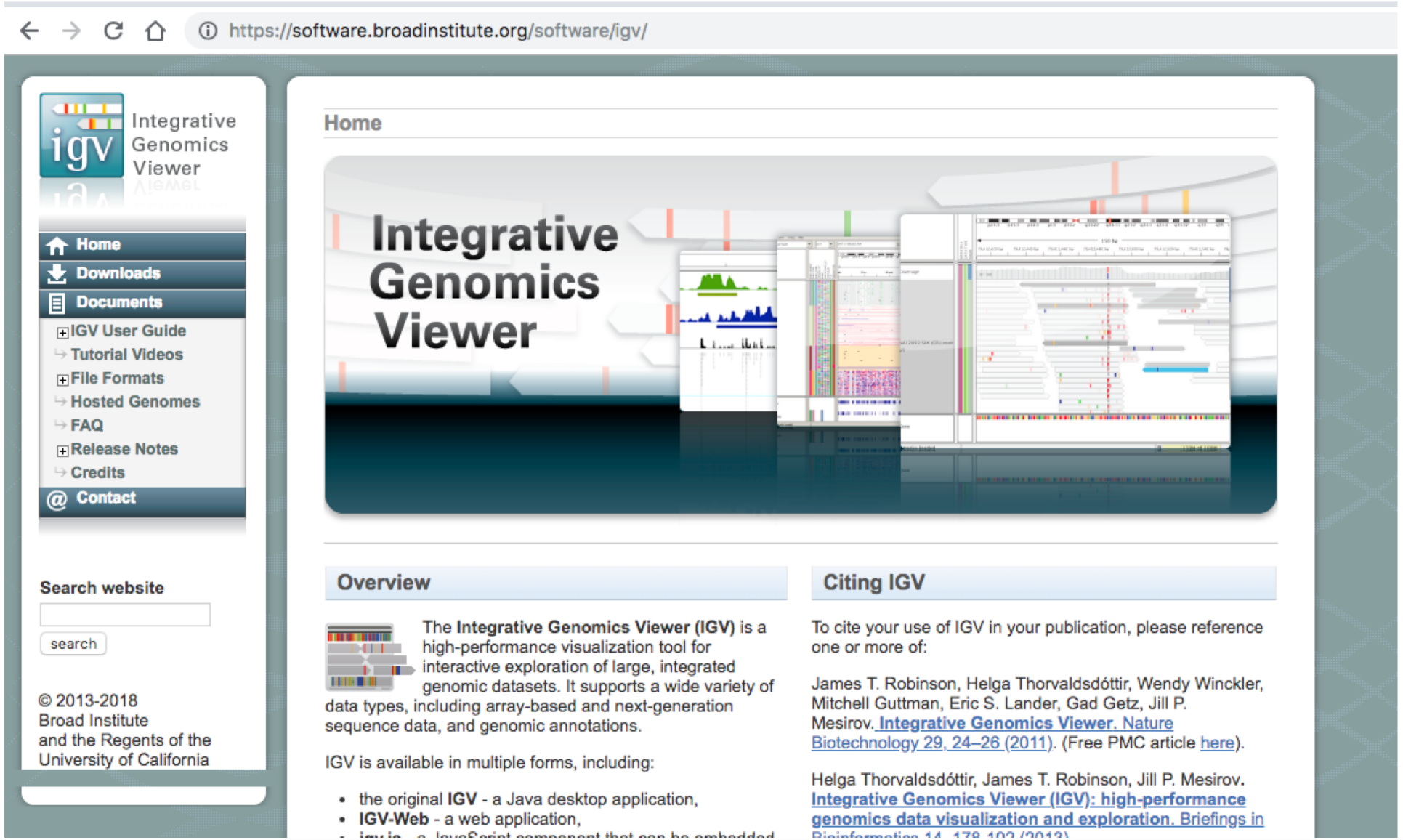
$ samtools view -S XXX.sam | more

```
61G9EAAXX100520:5:100:10001:20086       83      genome  59809   42      68M     =       59622   -255    GGTTTTAAAATTGTGCATGTTTATAGTTACCTTCATTTATGCTGGAAGAAATGAAATGAATTTT
TGTT    :CBE?EGFDFAEFEEEE?EFEGGGGGGGDFGGGGGFDDGGGGGGGGGGGGGGEGGGGGGGGEGGGGGGG    AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:68 YS:i:0  YT:Z:CP
61G9EAAXX100520:5:100:10001:20086       163     genome  59622   42      68M     =       59809   255     GTGTTAAATACTTGGAGGAGAACTTTGGCGCTCTCAAGGTTAAATTATCCGATGCCACTGTTAA
GGAA    GFGGGGGGGGGGGGGDGGFGGGGGGGGGGGGGGGGGGGFDGGGGGGGGDGGBFGFFEDGGBEEEBEDG    AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:68 YS:i:0  YT:Z:CP
61G9EAAXX100520:5:100:10002:14745       99      genome  400042  42      68M     =       400185  211     GGTAGTAGTAGCAGCAGCATTTGTGGAAGCTGAAGTAGGCAAGGTGCTGGGAGTAGATTCAACA
GCCT    EEEEEEFFCFFGGGFGFFEFGFFDEECEDCDCC?C=A;C@BAAA9CAACCC?C6A=36?8<?:?5<=@    AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:68 YS:i:0  YT:Z:CP
61G9EAAXX100520:5:100:10002:14745       147     genome  400185  42      68M     =       400042  -211    CAGTCTCAGTGGTGGCAGGATAGTGATAGGCGGCGGACTTGCTCTTAGTACTAACATCAGTAGG
TTCA    GDEFBGBEGFFGF?GFFGFDGGFGGGGFGGEGGGGGFGGGGGEFGGGGEFGGGFGGGGGGGGGGEGGG    AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:68 YS:i:0  YT:Z:CP
61G9EAAXX100520:5:100:10007:9987        83      genome  130215  42      68M     =       130091  -192    GAGTCTTGTGGTTGCAAGAGCAGCGCTGTTGATAACTGCAAGTGCTCTTCTTGCAAATGTGCTT
CGAA    CDBCDCCBDDDCEEEDEEEEFEGCFFGFGFFGGGGGGGGGGGEEGEGGGGFFGGGGGGGGGGGGGGGG    AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:68 YS:i:0  YT:Z:CP
61G9EAAXX100520:5:100:10007:9987        163     genome  130091  42      68M     =       130215  192     ATCGACTTTCAACAACTCATATCACCCTAACAAACTTTCAACGATGGAACACACTACCCAATGT
AAGA    GGGGGGGGGGGGGGGGFGGGGGGGGGGGFDGGFGGGGGGGFGFGGGEGGEGFGGGGGFGEGDFFFGFE    AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:68 YS:i:0  YT:Z:CP
61G9EAAXX100520:5:100:10008:6997        99      genome  400469  42      68M     =       400602  201     GTGGCAGAGCTATAGTCGAGGTTGGCGGGTTGAGGAGGATGAGAGCGTTTAGCGTTAAGGTCTT
GAGT    GGGFGGGGGGGGGDGFEGEGFCFFFEEEE:B@@AA=AA=???=:?ACACBABAC?B<=AA:????###    AS:i:-5 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:0A67       YS:i:0  YT:Z:CP
61G9EAAXX100520:5:100:10008:6997        147     genome  400602  42      68M     =       400469  -201    TGGTTTCATTGCGCTTCTCAGCGGCCTTGACGGCATCATAGGCTTGAGTGGCACAACGAGTGAT
TTCC    EECFEEDCGEEEGCGFDDEEEEGGCDEFFEFGGDFGGGGGGGGGGGGGGGGGFGGGGGGGGGGGFGGG    AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:68 YS:i:-5 YT:Z:CP
61G9EAAXX100520:5:100:10022:7766        153     genome  160514  42      68M     =       160514  0       ATGGTACTGAGTCTTTGCGGAGTAATTTCATTTCGTAATACTGATGCTCTAAGTTTCGGGTTAT
TTTT    ?FFDECFEDBEEFBFEGBDGGEFGGFFEDFGDDGGGGGGBGGGGEGGGGGGDGGEGGGGGGGGGGGE    AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:68 YT:Z:UP
61G9EAAXX100520:5:100:10022:7766        101     genome  160514  0       *       =       160514  0       CCATGATTTAAGAAAATTTCCAGCGATTTCCGCAAACAGTACATTCACAAAACGTTGTCATAGG
CTCA    GFFGGGEGGGGGGEFDGGGFGGGGGDGGGDEEGFAGDFGDGEDGFBFEFEEFEF?E=EEC?ECADFF:    YT:Z:UP
61G9EAAXX100520:5:100:10038:8120        99      genome  166132  42      68M     =       166425  361     GGGGCATGTAACATTGGTAGCATAATCGCAACCAGAGCTAGAGCAGTTGAAACGTCCAATTTCG
GCAA    GGGGGEGGGGGGGGGDGEGGFGGGGGGGGFDEGFFBDFFEE@FE@ECEEDD?DD@DDC@CCCD=CCBB    AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:68 YS:i:-4 YT:Z:CP
61G9EAAXX100520:5:100:10038:8120        147     genome  166425  42      68M     =       166132  -361    TACCGGAGTTAGCTTCAGGCTTAGTCGAATCAATTAAACGAATGCCATGGGGGAACGAAGGATA
TACA    ECDDFEEBDECGEBFEDEGEFGGFGGEGFFEFFGGFEDGGGGFFGFGEGGGGGGEG:GFGGGGGGFGG    AS:i:-4 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:56A11      YS:i:0  YT:Z:CP
61G9EAAXX100520:5:100:10040:13843       99      genome  166901  42      68M     =       166913  80      GAAGAGATGGGTTGAGGCAAAACAGCCAAAACAGCAATACCTACACAACTAGGGTTACCGTTTC
CAAC    GFGGGGGGGGGFGGGFGGDGFGGGGFDFFFGFFGFEEFEFEGGFDFEEEBEEEDBFFDDFCCDCCCDC    AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:68 YS:i:0  YT:Z:CP
61G9EAAXX100520:5:100:10040:13843       147     genome  166913  42      68M     =       166901  -80     TGAGGCAAAACAGCCAAAACAGCAATACCTACACAACTAGGGTTACCGTTTCCAACTCCAAAGA
ATAC    DDEEDEEEEDGEEEGEEEGGEEEGERGEEDGEGGEGEGGGGDGEEGGGDGGGGEGEDGGGGGGGGGG    AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:68 YS:i:0  YT:Z:CP
```

Samtools

- http://www.htslib.org/
- https://samtools.github.io/hts-specs/SAMv1.pdf

IGV

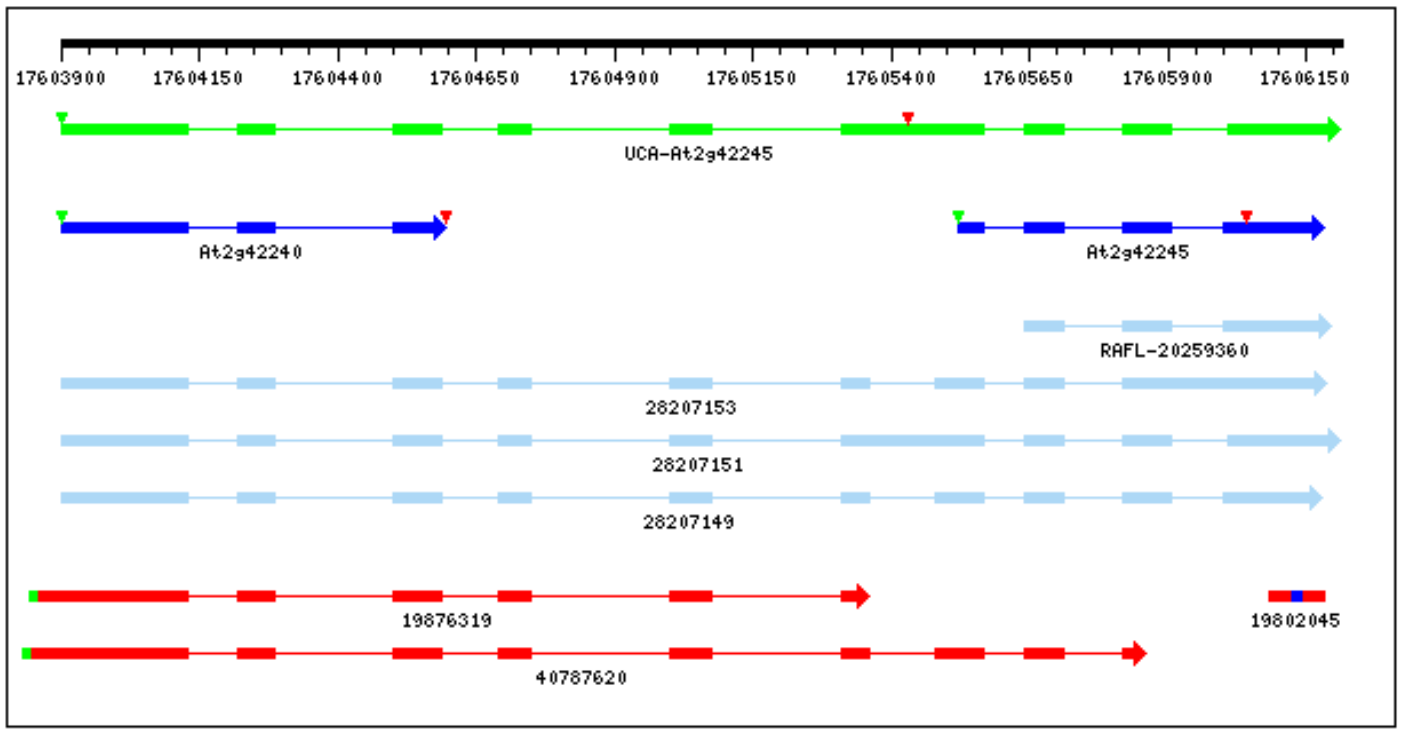

https://software.broadinstitute.org/software/igv/
Integrative Genomics Viewer
Home
Downloads
Documents
IGV User Guide
Tutorial Videos
File Formats
Hosted Genomes
FAQ
Release Notes
Credits
Contact
Search website
search
© 2013-2018 Broad Institute and the Regents of the University of California
Home
Integrative Genomics Viewer
Overview
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.
IGV is available in multiple forms, including:
the original IGV - a Java desktop application,
IGV-Web - a web application,
Citing IGV
To cite your use of IGV in your publication, please reference one or more of:
James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. Integrative Genomics Viewer. Nature Biotechnology 29, 24–26 (2011). (Free PMC article here).
Helga Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in

Genome sequence file: where to get?

- Illumina iGenomes (recommended)
  - http://support.illumina.com/sequencing/sequencing_software/igenome.html
- Ensembl
  - http://ensemblgenomes.org/info/access/ftp

- NCBI genome
  - http://www.ncbi.nlm.nih.gov/genome/
- Organism specific databases/websites.

Genome annotation file



GFF/GTF File Format - Definition

- The GFF (General Feature Format) format
  - one line per feature
  - each containing 9 columns of data
  - plus optional track definition lines.
- GFF has many versions (GFF, GFF2, GFF3)
- GTF (General Transfer Format) identical to GFF2.
- Most spliced aligner supports both GTF and GFF3 (mostly GTF)

GTF/GTF2 format

9 columns:

```
<seqname> <source> <feature> <start> <end> <score> <strand> <frame> [attributes]
```

- seqname - name of the chromosome or scaffold
- source – program or database that generated this feature.
- feature – Examples: "CDS", “gene", “transcript”, and "exon".
- start - The starting position of the feature in the sequence.
- end - The ending position of the feature (inclusive).
- score - A score between 0 and 1000.
- strand – ‘+ ‘(forward) or ‘-’ (reverse) or '.' (don't know/don't care).
- Frame – reading frame '0', '1' or '2'
- attribute – A semicolon-separated list of tag-value pairs, providing additional information about each feature.

Example of GTF2 format

```
AB000381 Twinscan  CDS          380   401   .   +   0  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  CDS          501   650   .   +   2  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  CDS          700   707   .   +   2  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  start_codon  380   382   .   +   0  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  stop_codon   708   710   .   +   0  gene_id "001"; transcript_id "001.1";
```

A simple example with 3 translated exons. Order of rows is not important.

Some annotation sources (e.g. Ensembl) add the gene_name attribute

```
gene_id "ENSBTAG00000020601"; transcript_id "ENSBTAT00000027448"; gene_name "ZNF366";
```

http://mblab.wustl.edu/GTF2.html

Generic Feature Format Version 3 (GFF3)

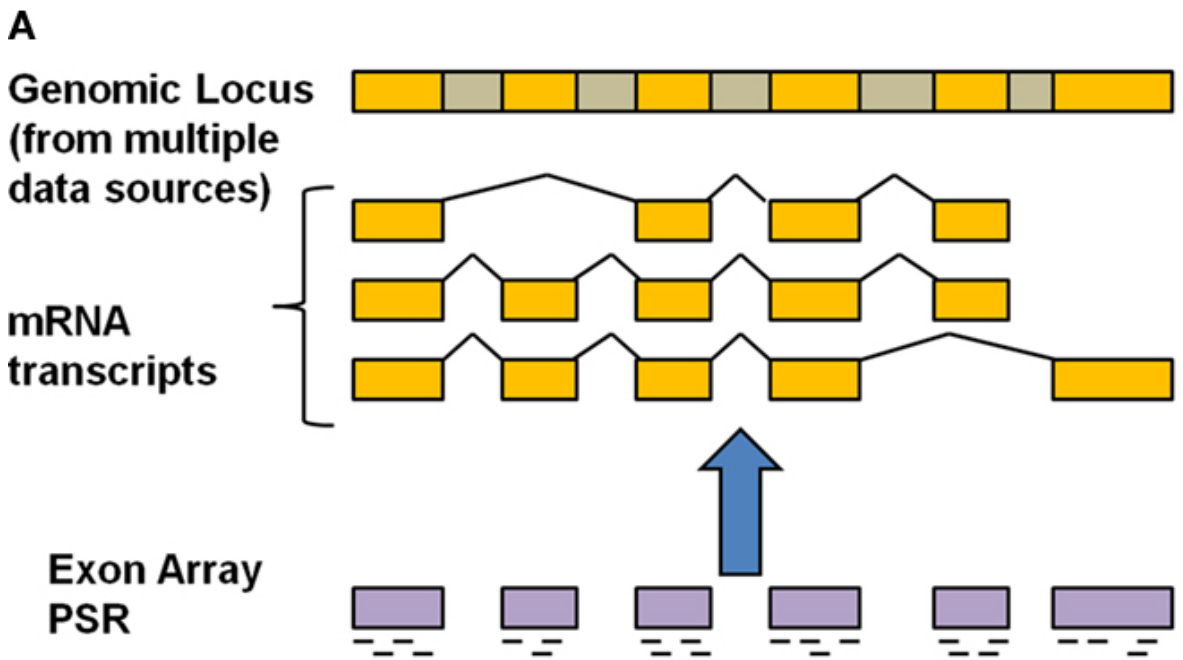
GFF3 adds parent feature

```
##gff-version 3.2.1
##sequence-region   ctg123 1 1497228
ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN

ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001

ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=E
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=E
ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001;Name=E

ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mRNA00001,mRNA00
ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mRNA00001,mRNA00
ctg123 . exon            5000  5500  .  +  .  ID=exon00004;Parent=mRNA00001,mRNA00
ctg123 . exon            7000  9000  .  +  .  ID=exon00005;Parent=mRNA00001,mRNA00

ctg123 . CDS             1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=ed
ctg123 . CDS             3000  3902  .  +  0  ID=cds00001;Parent=mRNA00001;Name=ed
ctg123 . CDS             5000  5500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=ed
ctg123 . CDS             7000  7600  .  +  0  ID=cds00001;Parent=mRNA00001;Name=ed

ctg123 . CDS             1201  1500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=ed
ctg123 . CDS             5000  5500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=ed
ctg123 . CDS             7000  7600  .  +  0  ID=cds00002;Parent=mRNA00002;Name=ed

ctg123 . CDS             3301  3902  .  +  0  ID=cds00003;Parent=mRNA00003;Name=ed
ctg123 . CDS             5000  5500  .  +  1  ID=cds00003;Parent=mRNA00003;Name=ed
ctg123 . CDS             7000  7600  .  +  1  ID=cds00003;Parent=mRNA00003;Name=ed

ctg123 . CDS             3391  3902  .  +  0  ID=cds00004;Parent=mRNA00003;Name=ed
ctg123 . CDS             5000  5500  .  +  1  ID=cds00004;Parent=mRNA00003;Name=ed
ctg123 . CDS             7000  7600  .  +  1  ID=cds00004;Parent=mRNA00003;Name=ed
```

http://www.sequenceontology.org/gff3.shtml

GFF2 vs GFF3

- GFF2
  - two-level hierarchies *transcript* → *exon*
- GFF3
  - three-level hierarchy of *gene* → *transcript* → *exon*



http://gmod.org/wiki/GFF2

Conversion GFF3 To GTF

- Optional
- Use `gffread` (comes with the Cufflinks software suite)

```
$ gffread my.gff3 -T -o my.gtf
```

- See `gffread -h` for more information

Download Prepare GTF/GFF file

- Download it from Illumina's iGenomes project (for model species)
  - http://support.illumina.com/sequencing/sequencing_software/igenome.html

Or

- Download gff3 files from genome database,
  - Ensembl gnome
  - http://ensemblgenomes.org/