Characterizing transcriptomes using ngs data

T. Källman

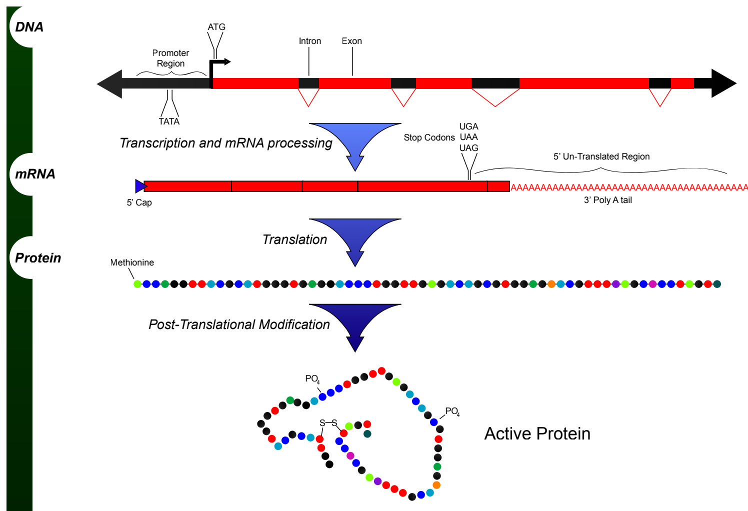
BILS/Scilife Lab/Uppsala University

Feb. 2015

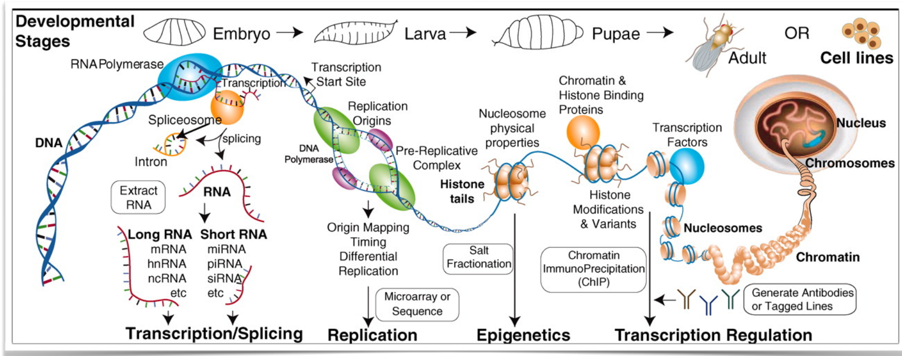
Outline

1. The transcriptome
2. RNA sequence technologies
3. RNA-seq analysis
   - Mapping based approach
   - Tools for working with ngs alignments
   - Gene expression from RNA-seq
   - de-novo assembly

The Central Dogma



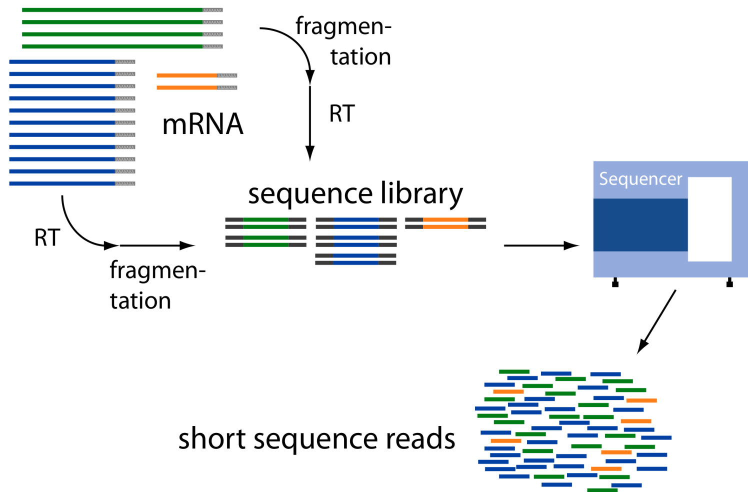
A more complex view



Transcriptomes vs genomes

- Dynamic, not the same over tissues and time points
- Smaller sequence space
- Less repetitive (but large gene families can be found)
- Fairly stable in size? (*eg.* 2-4 fold change among eukaryotes, whereas genome size can vary 1000-fold)
- Genes are often expressed in multiple different splice-variants
- RNA often from only one strand

NGS data



Machine output

fastq — less — 195×69


```
@HWI-ST0866_0110:5:1101:1264:2090#GATCAG/1
AGGCACTCCCTGCAGGTGTTGGACCACCTGGCTGAGCCACAGCGTCGCTTCCTGCTGCCAGGGCCTCGGAGAGGGTGGCTGTGGAGACACTGTGGGAGCA
+HWI-ST0866_0110:5:1101:1264:2090#GATCAG/1
^_P\`ccceeceeeee[b[beedaae_fdddde_cfhheedfeeh__`aeadd`d]baccc\[TKT\]_\ZQT^a[W[^^aW`^`aX^X^`_Y]^aBBBB
@HWI-ST0866_0110:5:1101:1418:2201#GATCAG/1
TCTTTATTGGCATCAGGCATCACCACACCATGGTTCTTGGCTCCCATGTTGGCCTGGACTCTCTTGCCATTCCGGGATCCTCTCTCATAGATGTACTCGC
+HWI-ST0866_0110:5:1101:1418:2201#GATCAG/1
__P`ccceegge]eghhhhdfhhhhhhhhhfhhefghffffhffhhfheg^eeffgfegf`fghhhffhhggadcX[`bbbbbbbbbcbbbcbR]aabaa
@HWI-ST0866_0110:5:1101:1561:2232#GATCAG/1
CCGAAACCCCGAAAGCACCCCAAAATCCCTGTGGGGAACCCCGAAAATCCCGAAATTACCCCAAAATACCTGTGGGATACCCTGAAAACCCGAAAGCACC
+HWI-ST0866_0110:5:1101:1561:2232#GATCAG/1
__[VJ\``\e[eefdgbafagfffagfd`Rc[cac`a_ef[a_N`]aced]\X]Z^RGYYYXa^_``bb_YYYbBBBBBBBBBBBBBBBBBBBBBBBBBB
@HWI-ST0866_0110:5:1101:1675:2246#GATCAG/1
GCTCAAGTCCCGGAGGAGGTCAGAGCTGGCATCTCTTCCCCAGCTGCTGCTCAGGAGTGTAAGCACCTGCAAACAGCTGCCAGCCAGGGAGCTGTGACTT
+HWI-ST0866_0110:5:1101:1675:2246#GATCAG/1
__J\`accccec[eagag`gggedbffhhffgfhhhhheaaefaghhhfdghhhdfd`ddgbd]_^abbbb__^ababbGXY_[aa^`aOOT[`bbGYYS
@HWI-ST0866_0110:5:1101:1752:2075#GATCAG/1
CAGCNGCTCTGGGCACCCTGTGCCAGGGCNTGNCCACCCTCCCAGCCAAGAATTCCTTCCCNATATCTAACCCAAATTTCTTCCCNGTAGGAGCAGGATG
+HWI-ST0866_0110:5:1101:1752:2075#GATCAG/1
Z_[aBQQQ`ccace_d_Y`a_Xd^ecc]fBPYBOYacedeZeVRbWVW\_\bcS\bdde`VBKKT^accab]GT\Z_YY``_]YGBKKWWO]`_W^[W_R
@HWI-ST0866_0110:5:1101:1888:2141#GATCAG/1
CAGATGAGGACTTTTGCTCCAAATGGGAAAGGAGAAAACCTCAGTCCGTAGAGATGCTCCCAGAGATAGTCCTAAAGATGAAGATGATGAAGACCCTTTG
+HWI-ST0866_0110:5:1101:1888:2141#GATCAG/1
abbeeeeeggggiihiiiiiiiiiiiiiiiiiiiiiiiiiiiiiihiiiihighiihiihichifggdgeebaceebbddbbdcecdcccbccbccc
@HWI-ST0866_0110:5:1101:1930:2172#GATCAG/1
ATCCAAGTTAAAACAGAGGCCTGTGACAGACTCTTGGCCCATCGTGTTGATACTAAAATGAAAGGAAACAAAGTGAATGAAGTACTGAATAGATTACACT
+HWI-ST0866_0110:5:1101:1930:2172#GATCAG/1
_^_a^cccegcgghhgZc`ghhc^egggd^_[d]defcdfd^Z^OXWaQ^adghWWaff_H_cbdbbd\dbddV^_ZRMHHZGUZ_b_YRTGTT]`__b[
@HWI-ST0866_0110:5:1101:1945:2183#GATCAG/1
CTCACGATGGTCCCCAGGCTGTCCACAGTTGCCACACACTGATAATATCCTTCATCAGGTTTATTATGCTTGGAATGCACCACACTGTTAATTAATAAAG
+HWI-ST0866_0110:5:1101:1945:2183#GATCAG/1
^__ccc\ccY`^ae`Z_`bR`b]fa]dec^ceeeffc^fcdceXc]cehehaebefd`eW\\b]bebeeedde`R\_aa_c]b\baaZ`accdc[`]a`a
@HWI-ST0866_0110:5:1101:1920:2205#GATCAG/1
GCCAGTACAGCTGTAGTAGTCTGTCCTTCCCATCCGTGCCCATGTGACACAGCAGGTTCACAGCATGGTGACCAGTTTGAAGCTTCCTACCTCTGTGGTA
+HWI-ST0866_0110:5:1101:1920:2205#GATCAG/1
bbbeeeeegfggghfhfhhihihiiiiihiiiiiiiiighiiiiiiiiiiiihiihghiiiiiiihhhbdghggggggeeceeebddcddcccccbcccc
@HWI-ST0866_0110:5:1101:2095:2167#GATCAG/1
GTTCAGACAAGTTCGATCTCTTGTGCATCGACTGTGCTGGATGATAGTTTTTCAGTGAGTATTATGGTTAGTAGATATAGTACCAGGCTGCAAATAGCTA
+HWI-ST0866_0110:5:1101:2095:2167#GATCAG/1
a_P\cceeggggiighihiiiighhiiiihhiihiiiiiihihiiiehhiiifhhibfaedfhiiifghih____dgeeddgeeeeeddc_bbccccbb
@HWI-ST0866_0110:5:1101:2494:2131#GATCAG/1
CTCGAAATCCAGGGCAACGTAGCACAGCTTCTCCTTGATGTCACGCACAATTTCTCTCTCAGCTGTGGTGGTGAAGCTGTAGCCTCTCTCTGTCAGGATC
+HWI-ST0866_0110:5:1101:2494:2131#GATCAG/1
_aaeceeegggggdfgfihghffhhhiiihffgiiiiiihihhfiigghdgdhffhiiifdhihd`^bV^aabb__bdc_]bZ`bccccccccb]`bcc
@HWI-ST0866_0110:5:1101:2424:2217#GATCAG/1
TAACAGTCCCCCTGGTATGAAATGGCACCTTGGTTACACTGAGGGAGGGGTGAGGTTACAGGGAGTAATTTTCATGTGTAACTGGGGTTAAAAAAAAAAA
+HWI-ST0866_0110:5:1101:2424:2217#GATCAG/1
_bPaceeegfggcghfgfhhiiiiiiffh_fghghhfhhhgfghfcghhiT_bdddddeeeeaac`bbcccb`cb`cbbc`ccbcccaacbbbbcaaccc
@HWI-ST0866_0110:5:1101:2485:2220#GATCAG/1
CCTGGATGGTGGGCTGATCAACTTTGAGAAGAGAAGGAAGGAGTTCGAAGTCATCGCGCAGATCAAGCTGCTCCAGTCGGCCTGCAACAACTACAGCTTC
+HWI-ST0866_0110:5:1101:2485:2220#GATCAG/1
_bbeeeeegegggiiiiiiiiiiiiihidghhihiiiiiiiiiigghhhhihiiiiiihigeeddddddccccccbcccccccaccccccccccccbbccc
@HWI-ST0866_0110:5:1101:2476:2244#GATCAG/1
CAGTACTCTTTGTACCGCTCATCTGCATCTCCAAACACTTTGTACCTGCTGCCTTTTATTTTGTATGTTTACCTGTGTCAGAGAGTCGCCAAGTTTGTTC
+HWI-ST0866_0110:5:1101:2476:2244#GATCAG/1
abbeeeeefggghifiihiiihigaghefhhdgheghhhhf^afgffhfhiiihidfhhihihhiiggbg`gfdg`bdddeade`]`_Z^Y[]bcbccb
@HWI-ST0866_0110:5:1101:2502:2189#GATCAG/1
AACAAAACGGGCTGTTTTAGGACCCTTGGTCCCAAGGGGTAATGGCCCTCAGCACCCACTATCCCTGCTCTCCAGGGCTCTCTAGGGATTTAGTGCTGAT
+HWI-ST0866_0110:5:1101:2502:2189#GATCAG/1
_b_ceeeeggggiiiiihiiiiiiiiiiiiiiiiiiiiefhiihiiiihihihhhgggeeee_dddcddccccbcaccccccccbcbcccccccccdccc
@HWI-ST0866_0110:5:1101:2517:2226#GATCAG/1
TATCAATTTGCGCTTGATTACTAGTGCTACCTTCCCATACATTGCAGAGAGCGCGTGTCCATAGTGTATGGTACAGTACAACCAGCACCACAGCTTAGAG
+HWI-ST0866_0110:5:1101:2517:2226#GATCAG/1
__P`^ceeggff`fhiiiiighfhgghihghhhghhdfdffh__agffhghcddgggggedeeea^^acddZ_bRU]`]bcZZ[`ab[`^c_aaccccbb
@HWI-ST0866_0110:5:1101:2659:2245#GATCAG/1
TGCCTATGGAATTACGTTAATTTACACAAACAAATTCCATATTAGCTTTAAAAAATAAACCTACTTCTAACTAGAAGTGAAAGAAGTTTAAAAGTGCTGC
+HWI-ST0866_0110:5:1101:2659:2245#GATCAG/1
^YYccacc]`ae^affhhhbc^dgeghfeeafg`dfhhgf]fhhSYcgfd]cae_fghhRW\eb]ffhZV\`d]dR\V^Z_dced`bGZ`b`bbZ]]_Y`
:
```

Machine output

```
@SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
CAACGAGTTCACACCTTGGCCGACAGGCCCGGGTAA
+SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
BA@7>B=>:>>7@7@>>9=BAA?;>52;>:9=8.=A
@SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
CCAATGATTTTTTTCCGTGTTTCAGAATACGGTTAA
+SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
BCCBA@BB@BBBBAB@B9B@=BABA@A:@693:@B=
@SRR038845.53 HWI-EAS038:6:1:1:360 length=36
GTTCAAAAAGAACTAAATTGTGTCAATAGAAAACTC
+SRR038845.53 HWI-EAS038:6:1:1:360 length=36
BBCBBBBBB@@BAB?BBBBCBC>BBBAA8>BBBAA@
```

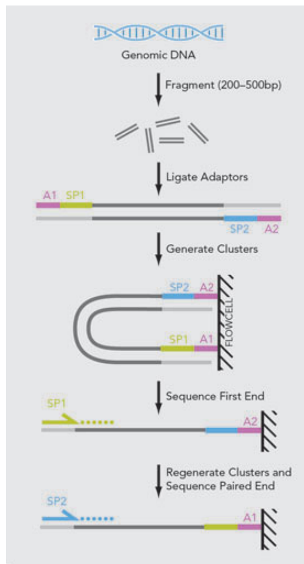
Sequence quality

- Phred quality scores: Q = -10 x log P (High Q = high probability of the base being correct
- A Phred quality score of 20 to a base, means that the base is called incorrectly in 1 out of 100 times.

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
.................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                              |                     |
33                        59   64       73                            104                  126
0.........................26...31.......40
                         -5....0........9..............................40
                               0........9..............................40
                                  3.....9..............................40
0.2.......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

Pair-end (PE) sequencing



Pair-end reads

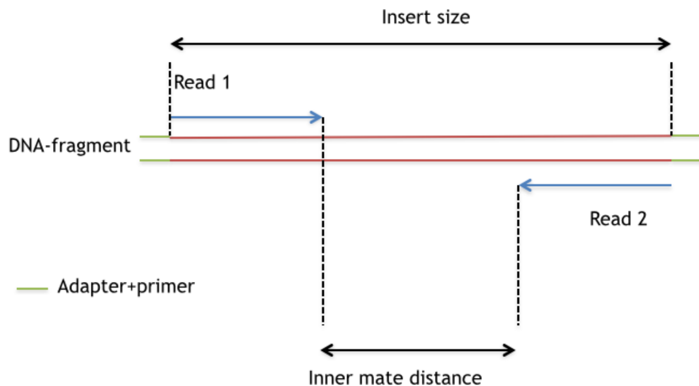
File format

- Two files are created
- The order in files identical and naming of reads are the same with the exception of the end
- The way of naming reads are changing over time so the read names depend on software version

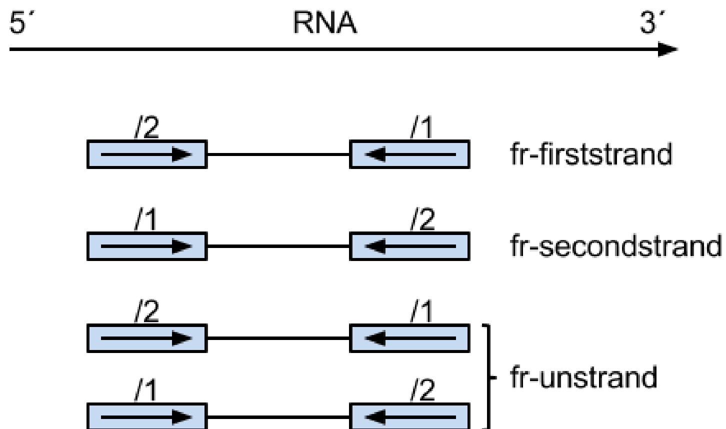
```
@61DFRAAXX100204:1:100:10494:3070/1
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
+
ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCBC?CCCCCCCCC@@CACCCCCA
```

```
@61DFRAAXX100204:1:100:10494:3070/2
ATCCAAGTTAAAACAGAGGCCTGTGACAGACTCTTGGCCCATCGTGTTGATA
+
_^_a^cccegcgghhgZc`ghhc^egggd^_[d]defcdfd^Z^OXWaQ^ad
```

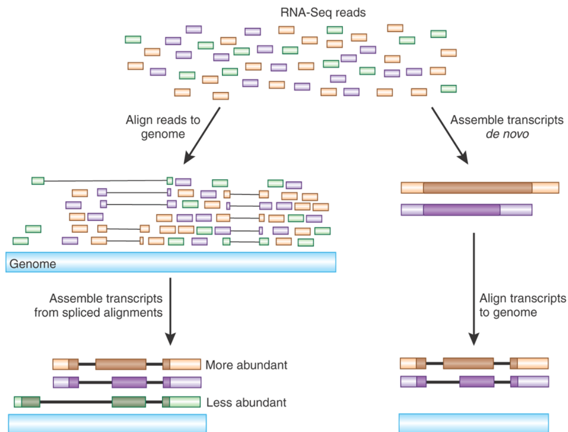
Pair-end data



Stranded or not



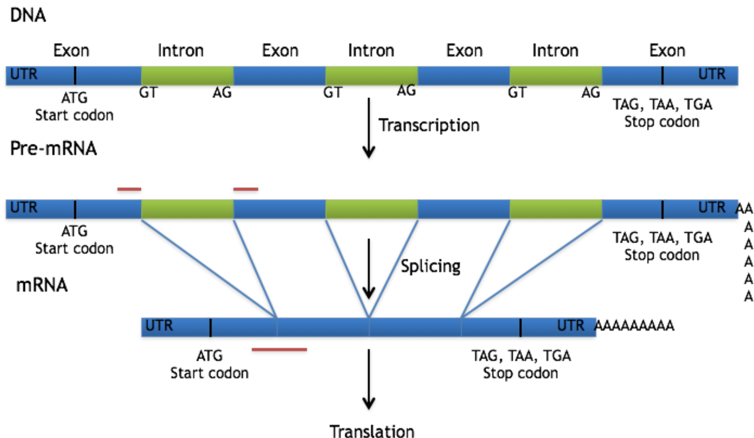
Two main routes for analysis



Haas & Zody (2010), Nature Biotechnology 28, 421–423

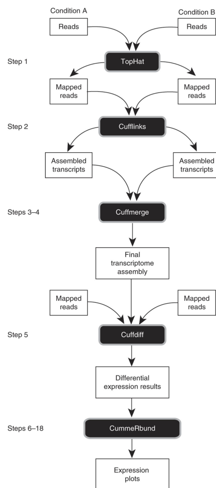
Aligning short reads from RNA to genomes

- Large number of programs available: Star, Tophat, Subread etc
- Important feature: Allow for spliced mapping



Example workflow

- Tophat: Aligns reads to genome (allows for spliced read mapping)
- Cufflinks: Extract transcripts from spliced read alignments
- Cuffmerge: Merge results from multiple Cufflinks results

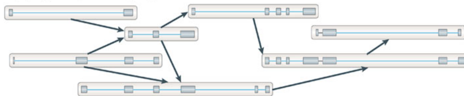


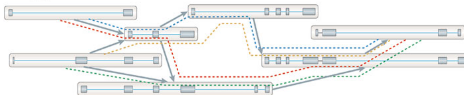
Trapnell *et al.* (2012), Nature Protocols 7, 562–578

Tophat

1. Efficient and fast alignment to the genome using bowtie2
2. Create a data base of putative splice junctions from the reads mapping in step 1
3. Map reads that did not map in step 1 run using the splice information

Cufflinks





Nature Reviews | Genetics

Samtools

- Program to work with ngs alignment files (SAM, BAM, CRAM)
- Can be used to view data, calculate basic info, extract subsets of alignments and convert between file formats
- http://www.htslib.org

Picard

- A set of Java command line tools with the same (or similar functionality as samtools)
- Note that even though they largely aim at doing similar functions Picard and Samtools is not always generating compatible file formats
- http://broadinstitute.github.io/picard/

Samtools tview, a text-based alignment viewer

$ samtools view alignment.bam target.fasta

IGV: Integrative Genomics Viewer

Viewer

- Home
- Downloads
- Documents
  - Hosted Genomes
  - FAQ
  - IGV User Guide
  - File Formats
  - Release Notes
  - IGV for iPad
  - Credits
- Contact

Search website

search

Broad Home
Cancer Program

BROAD INSTITUTE
© 2013 Broad Institute

What's New

NEWS

September 2014. The IGV iPad app can now be installed from the Apple App Store. ***IGV for iPad*** is a lightweight genomic data viewer that provides some of the functionality available in our regular desktop IGV. See the IGV for iPad documentation for details.

June 2014. We're hiring! See the job description on the Broad Institute careers website.

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.

Downloads

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

Citing IGV

To cite your use of IGV in your publication:

Helga Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov. **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.** Briefings in Bioinformatics 2012.

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)

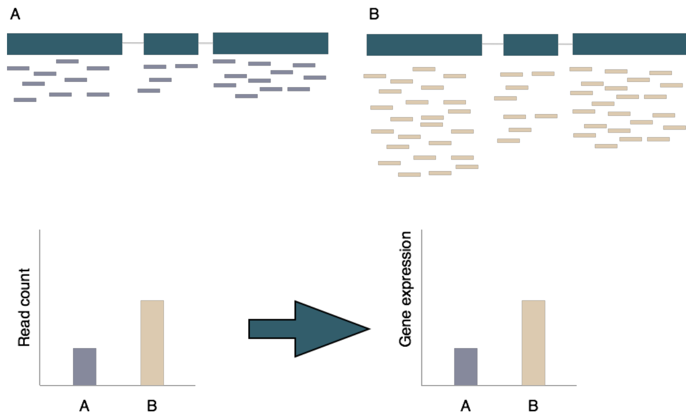
Funding

Development of IGV is made possible by funding from the National Cancer Institute, the National Institute of General Medical Sciences of the National Institutes of Health, and the Starr Cancer Consortium.

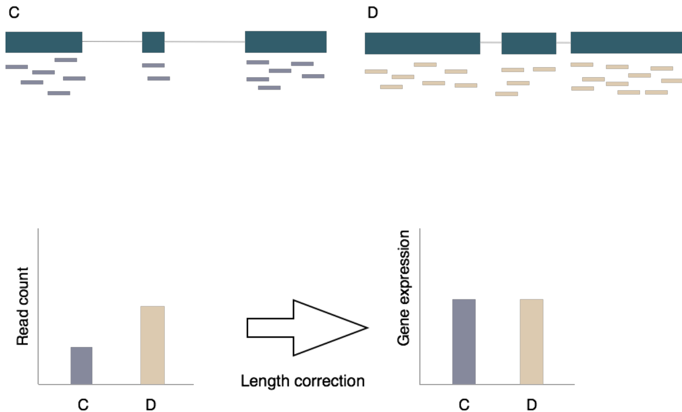
IGV participates in the GenomeSpace initiative, which is funded by the National Human Genome Research Institute.

IGV: Integrative Genomics Viewer

From counts to gene expression



From counts to gene expression



Not all reads are the same

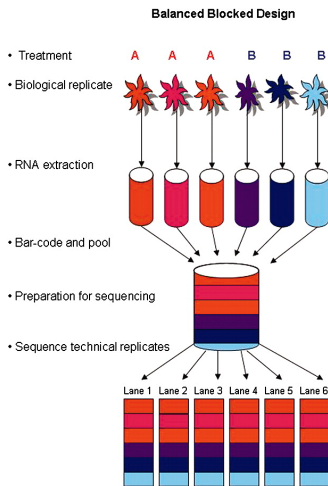
	union	intersection _strict	intersection _nonempty
read / gene_A	gene_A	gene_A	gene_A
read / gene_A	gene_A	no_feature	gene_A
read / gene_A, gene_A	gene_A	no_feature	gene_A
read, read / gene_A, gene_A	gene_A	gene_A	gene_A
read / gene_A / gene_B	gene_A	gene_A	gene_A
read / gene_A / gene_B	ambiguous	gene_A	gene_A
read / gene_A / gene_B	ambiguous	ambiguous	ambiguous

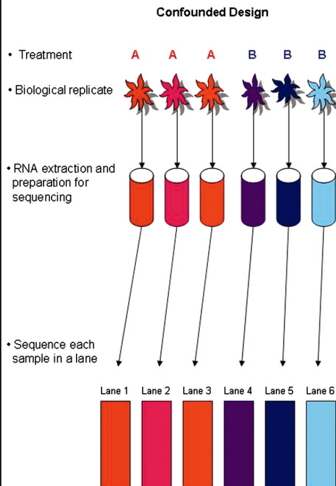
from: http://www-huber.embl.de/users/anders/HTSeq/doc/count.html

Normalized expression Values

- Transcript-mapped read counts are normalized for both length of the transcript and total depth of sequencing.
- Count data is hence converted to: Reads/Fragments per kb of transcript length and million mapped reads (RPKM or FPKM)

Experimental design





Experimental design

- Count reads (convert to RPKM/FPKM?)
- Small number of reads (= low RPKM/FPKM values) often non-significant
- Remember that Fold change is not the same as significance

	Condition 1	Condition 2	Fold_Change	Significant?
Gene A	1	2	2-fold	No
Gene B	100	200	2-fold	Yes

Major challenges in relation to genome assembly

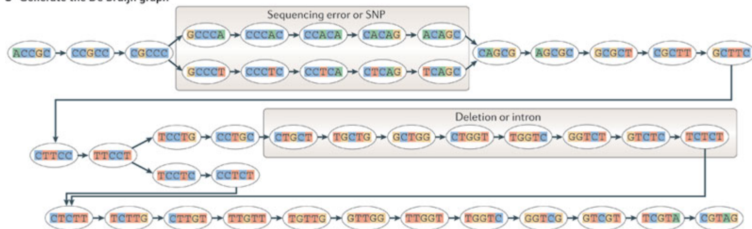
- Genes show different levels of gene expression, hence uneven coverage among genes
- Many genes are expressed in different isoforms
- As sequence depth increase detected number of loci increase. (What is actually expressed?)
- Sequence error from highly expressed genes might be seen more often than "true" sequences from lowly expressed genes

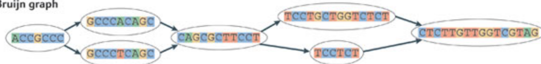
Several programs available

- SOAP-denovo TRANS
- Oases
- Trans-ABYSS
- Trinity

All of them uses de Bruijn graphs to cope with the data and many of them have been developed from a genome assembly program

Trinity





Trinity