Characterizing transcriptomes using ngs data

T. Källman

BILS/Scilife Lab/Uppsala University

Sep. 2015

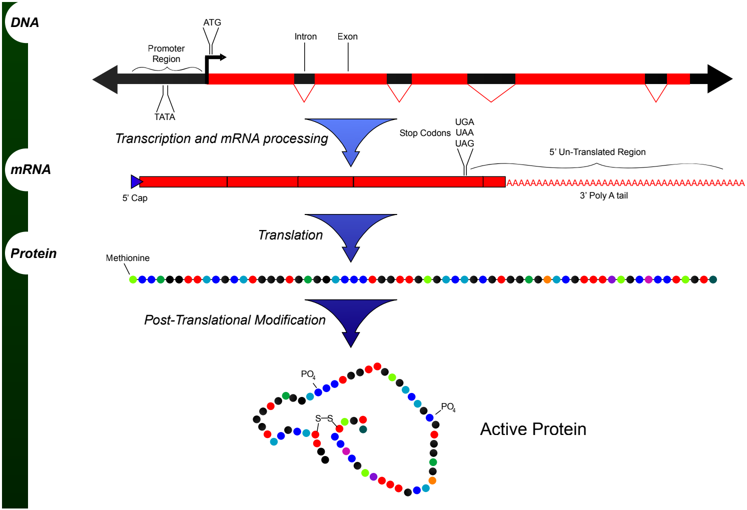
Outline

1. The transcriptome

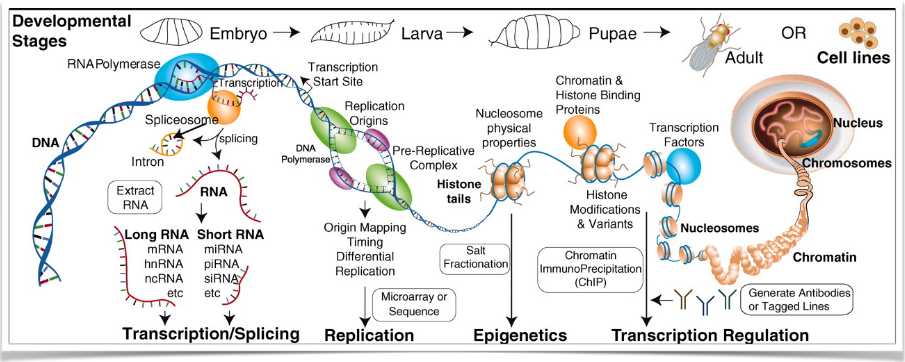
2. RNA sequence technologies

3. RNA-seq analysis
   - Mapping based approach
   - Gene expression from RNA-seq
   - de-novo assembly

The Central Dogma



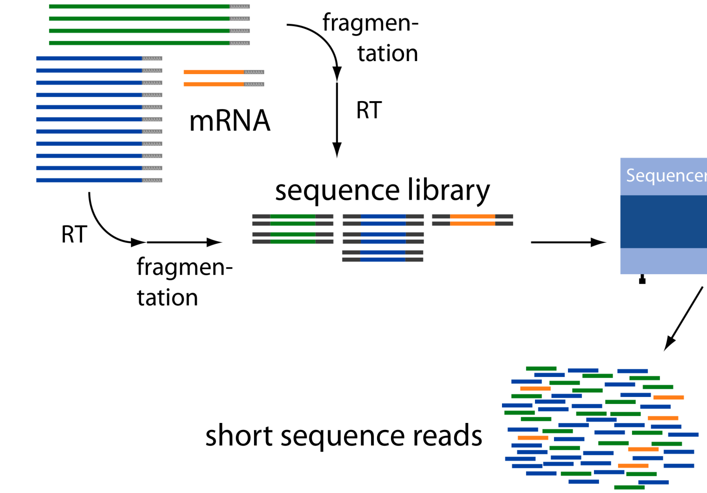
A more complex view



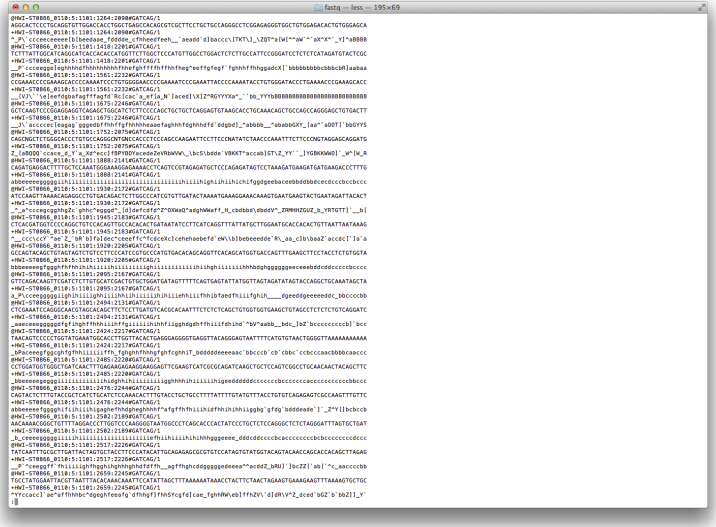
Transcriptomes vs genomes

- 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

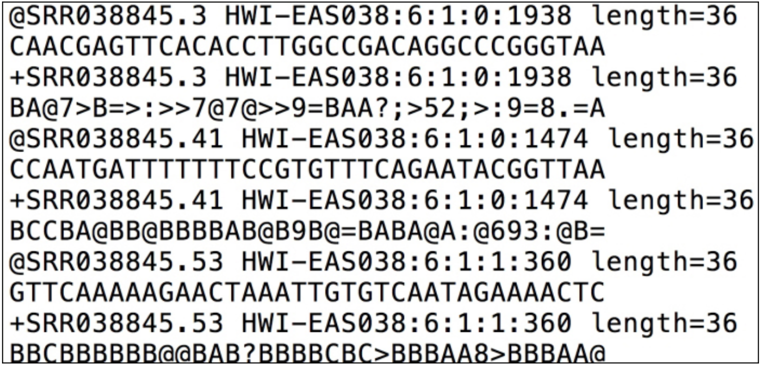

fragmen-
tation
RT
mRNA
sequence library
Sequencer
RT
fragmen-
tation
short sequence reads

Machine output



```
@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[`bbbbbbbbcbbbcbR]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
abbeeeeeggggiihiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiihiiiihighiihiihichifggdgeebaceebbddbbdcecdcccbccbccc
@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
_aaeceeeeggggdfgfihghffhhhiiihffgiiiiiihihhfiigghdgdhffhiiifdhihd`^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
_bbeeeeegegggiiiiiiiiiiiiihidghhihiiiiiiiiiigghhhhihiiiiiihigeedddddddcccccccbcccccccacccccccccccbbccc
@HWI-ST0866_0110:5:1101:2476:2244#GATCAG/1
CAGTACTCTTTGTACCGCTCATCTGCATCTCCAAACACTTTGTACCTGCTGCCTTTTATTTTGTATGTTTACCTGTGTCAGAGAGTCGCCAAGTTTGTTC
+HWI-ST0866_0110:5:1101:2476:2244#GATCAG/1
abbeeeeefggggghifiihiiihigaghefhhdgheghhhhf^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_ceeeeggggiiiiihiiiiiiiiiiiiiiiiiiiiiefhiihiiiihihihhhgggeeee_dddcddccccbcacccccccccbcbccccccccdccc
@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`
:
```

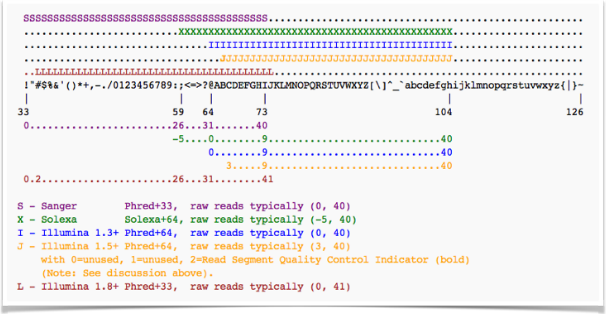
Machine output



```
@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

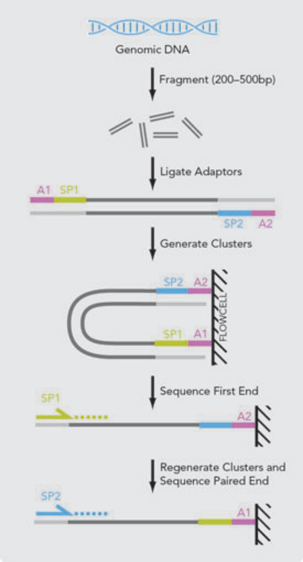
- Phred quality scores: Q = -10 x log P (High Q = high probability of the base being correct
- A Phred quality score of 20 to a base, means that the base is called incorrectly in 1 out of 100 times.



```
 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

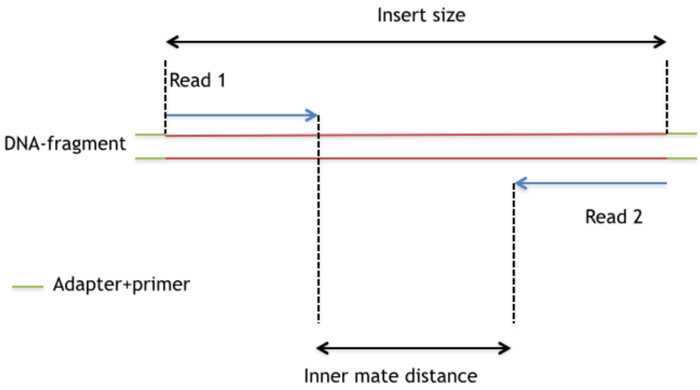
File format

- 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

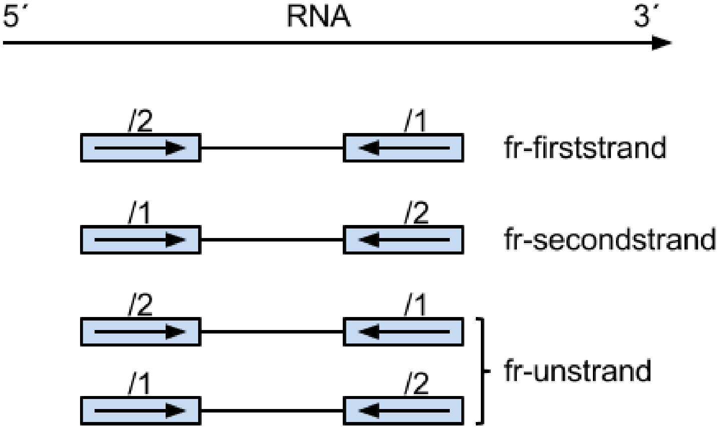
```
@61DFRAAXX100204:1:100:10494:3070/1
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
+
ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCBC?CCCCCCCCC@@CACCCCCA
```

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

Pair-end data

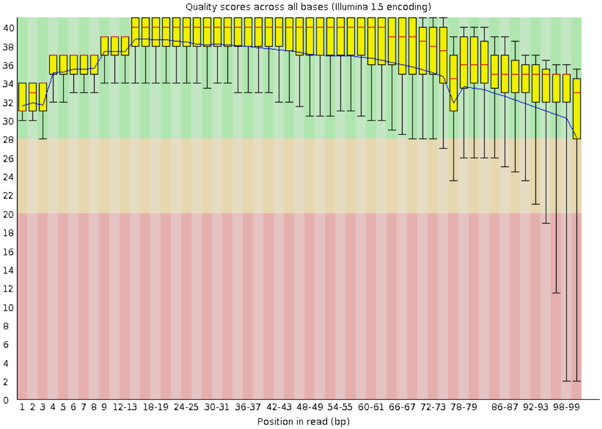


Stranded or not



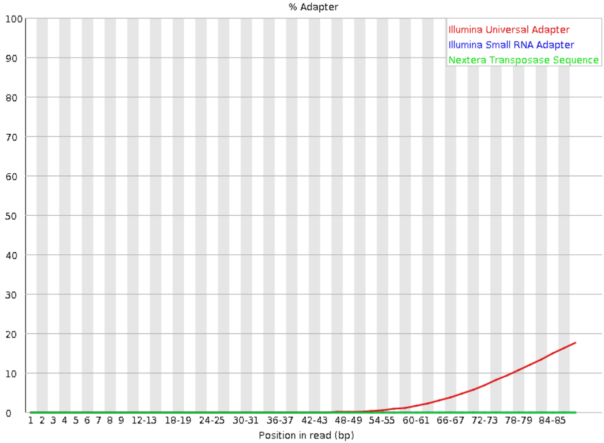
Basic quality control of raw reads

- FastQC



Basic quality control of raw reads

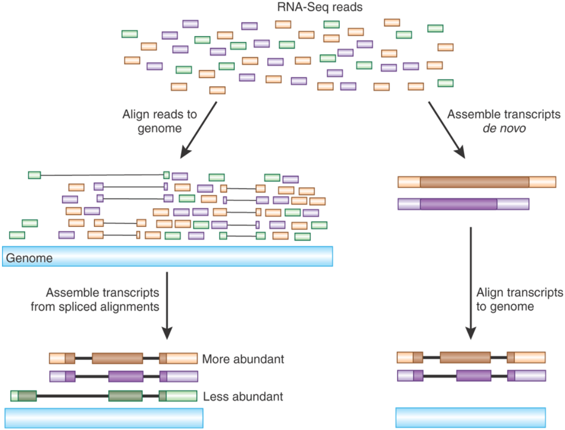
- FastQC



Basic quality control of raw reads

- RNA-seq is not random sample from the genome eg. GC content might be different
- Highly expressed genes can be frequent and create warnings in quality controls that assumes whole genome data
- Random hexamer in cDNA synthesis might create 'biases' in base frequencies in the beginning of reads

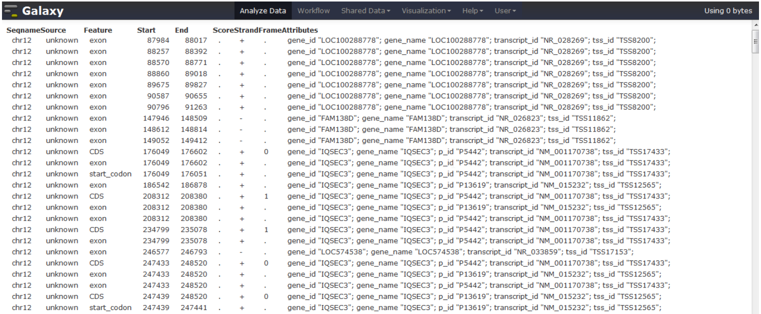
Two main routes for analysis


RNA-Seq reads
Align reads to genome
Assemble transcripts *de novo*
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

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

Aligning short reads from RNA to genomes

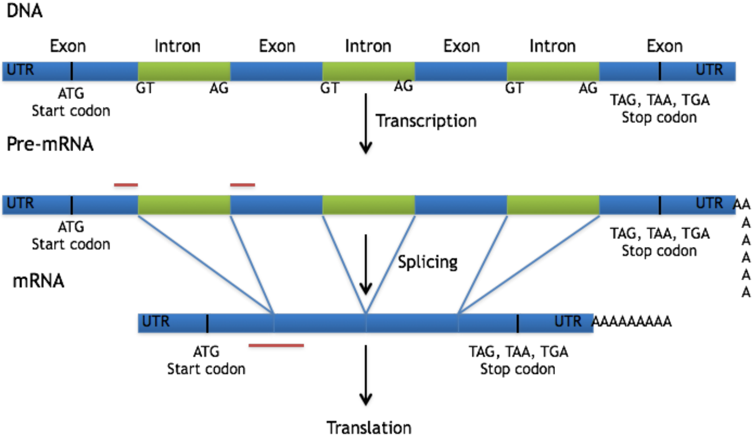
- If available map to the genome sequence
- If no genome sequence one can also map to transcriptome reference
- Make use of available genome annotation (GTF, GFF, BED files)



Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attributes
chr12	unknown	exon	87984	88017	.	+	.	gene_id "LOC100288778"; gene_name "LOC100288778"; transcript_id "NR_028269"; tss_id "TSS8200";
chr12	unknown	exon	88257	88392	.	+	.	gene_id "LOC100288778"; gene_name "LOC100288778"; transcript_id "NR_028269"; tss_id "TSS8200";
chr12	unknown	exon	88570	88771	.	+	.	gene_id "LOC100288778"; gene_name "LOC100288778"; transcript_id "NR_028269"; tss_id "TSS8200";
chr12	unknown	exon	88860	89018	.	+	.	gene_id "LOC100288778"; gene_name "LOC100288778"; transcript_id "NR_028269"; tss_id "TSS8200";
chr12	unknown	exon	89675	89827	.	+	.	gene_id "LOC100288778"; gene_name "LOC100288778"; transcript_id "NR_028269"; tss_id "TSS8200";
chr12	unknown	exon	90587	90655	.	+	.	gene_id "LOC100288778"; gene_name "LOC100288778"; transcript_id "NR_028269"; tss_id "TSS8200";
chr12	unknown	exon	90796	91263	.	+	.	gene_id "LOC100288778"; gene_name "LOC100288778"; transcript_id "NR_028269"; tss_id "TSS8200";
chr12	unknown	exon	147946	148509	.	-	.	gene_id "FAM138D"; gene_name "FAM138D"; transcript_id "NR_026823"; tss_id "TSS11862";
chr12	unknown	exon	148612	148814	.	-	.	gene_id "FAM138D"; gene_name "FAM138D"; transcript_id "NR_026823"; tss_id "TSS11862";
chr12	unknown	exon	149052	149412	.	-	.	gene_id "FAM138D"; gene_name "FAM138D"; transcript_id "NR_026823"; tss_id "TSS11862";
chr12	unknown	CDS	176049	176602	.	+	0	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	exon	176049	176602	.	+	.	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	start_codon	176049	176051	.	+	.	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	exon	186542	186878	.	+	.	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P13619"; transcript_id "NM_015232"; tss_id "TSS12565";
chr12	unknown	CDS	208312	208380	.	+	1	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	exon	208312	208380	.	+	.	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P13619"; transcript_id "NM_015232"; tss_id "TSS12565";
chr12	unknown	exon	208312	208380	.	+	.	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	CDS	234799	235078	.	+	1	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	exon	234799	235078	.	+	.	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	exon	246577	246793	.	-	.	gene_id "LOC574538"; gene_name "LOC574538"; transcript_id "NR_033859"; tss_id "TSS17153";
chr12	unknown	CDS	247433	248520	.	+	0	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	exon	247433	248520	.	+	.	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P13619"; transcript_id "NM_015232"; tss_id "TSS12565";
chr12	unknown	exon	247433	248520	.	+	.	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	CDS	247439	248520	.	+	0	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P13619"; transcript_id "NM_015232"; tss_id "TSS12565";
chr12	unknown	start_codon	247439	247441	.	+	.	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P13619"; transcript_id "NM_015232"; tss_id "TSS12565";

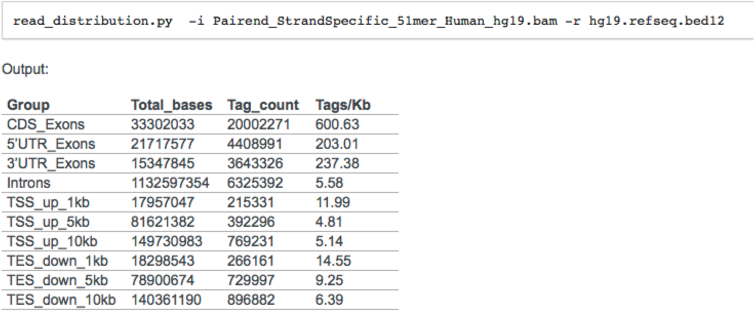
Aligning short reads from RNA to genomes

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



Aligning short reads from RNA to genomes

- After mapping perform QC of the output



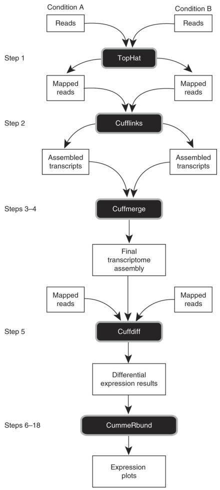
```
read_distribution.py  -i Pairend_StrandSpecific_51mer_Human_hg19.bam -r hg19.refseq.bed12
```

Output:

Group	Total_bases	Tag_count	Tags/Kb
CDS_Exons	33302033	20002271	600.63
5'UTR_Exons	21717577	4408991	203.01
3'UTR_Exons	15347845	3643326	237.38
Introns	1132597354	6325392	5.58
TSS_up_1kb	17957047	215331	11.99
TSS_up_5kb	81621382	392296	4.81
TSS_up_10kb	149730983	769231	5.14
TES_down_1kb	18298543	266161	14.55
TES_down_5kb	78900674	729997	9.25
TES_down_10kb	140361190	896882	6.39

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
- Cuffdiff: Detect differential gene expression



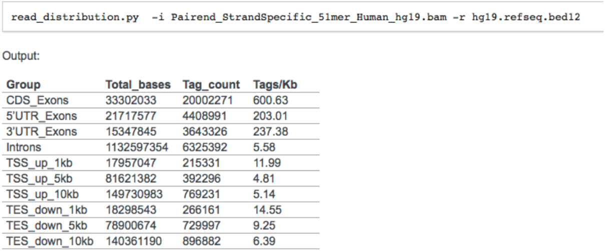
Trapnell *et al.* (2012), Nature Protocols 7, 562–578

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 using the splice information

QC of mapped reads

Reads should mostly map to known (annotated genes)



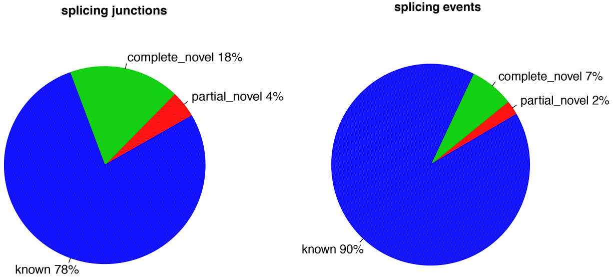
```
read_distribution.py  -i Pairend_StrandSpecific_51mer_Human_hg19.bam -r hg19.refseq.bed12
```

Output:

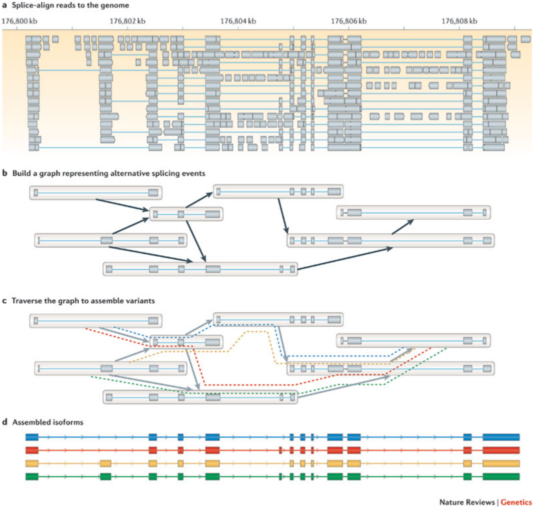
Group	Total_bases	Tag_count	Tags/Kb
CDS_Exons	33302033	20002271	600.63
5'UTR_Exons	21717577	4408991	203.01
3'UTR_Exons	15347845	3643326	237.38
Introns	1132597354	6325392	5.58
TSS_up_1kb	17957047	215331	11.99
TSS_up_5kb	81621382	392296	4.81
TSS_up_10kb	149730983	769231	5.14
TES_down_1kb	18298543	266161	14.55
TES_down_5kb	78900674	729997	9.25
TES_down_10kb	140361190	896882	6.39

QC of mapped reads

Most splice event should be known and canonical (GU-AG)



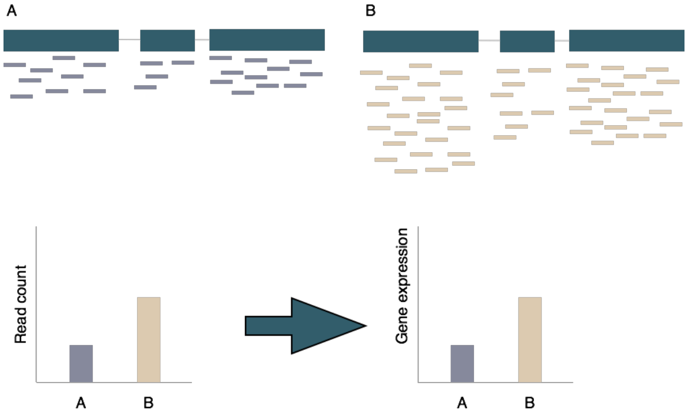
Cufflinks


a Splice-align reads to the genome
176,800 kb
176,802 kb
176,804 kb
176,806 kb
176,808 kb
b Build a graph representing alternative splicing events
c Traverse the graph to assemble variants
d Assembled isoforms
Nature Reviews | Genetics

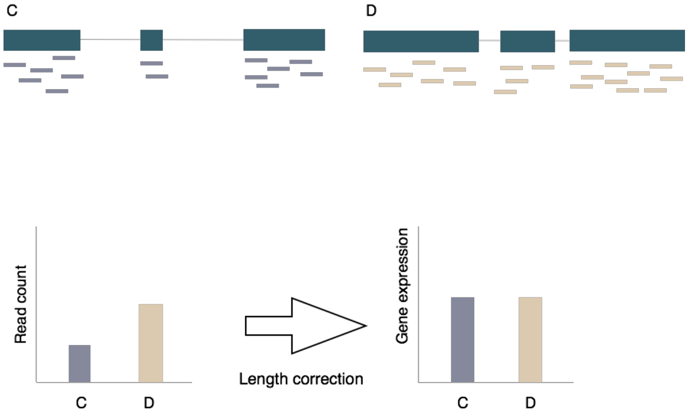
Cuffdiff

- Program that estimate expression levels and identify differentially expressed genes from ngs alignments
- Basically uses the read data to estimate dispersion parameters (the amount of deviation from a Poisson distr.)
- Genes that show patterns deviating from the above expectations are differentially expressed between treatments
- Will work also for detection of isoform differential expression

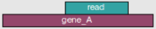
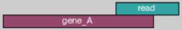
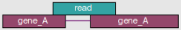
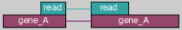
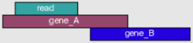
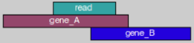
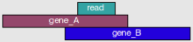
From counts to gene expression



From counts to gene expression



Not all reads are the same

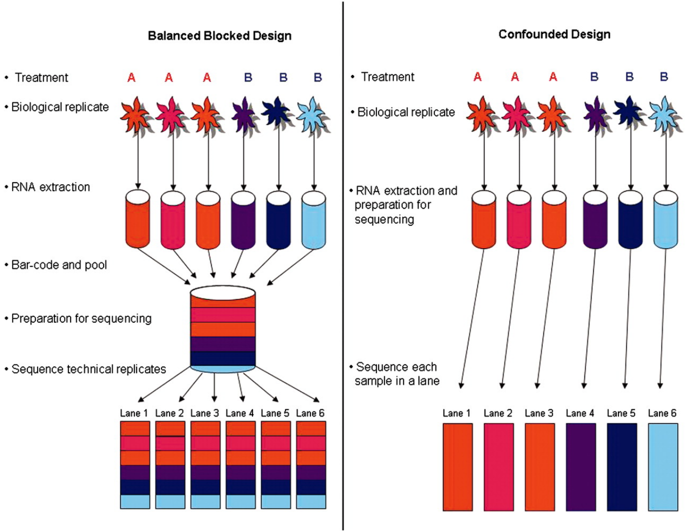
	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

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

Normalized expression Values

- Mapped read counts are normalized for both length of the transcript they map to 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

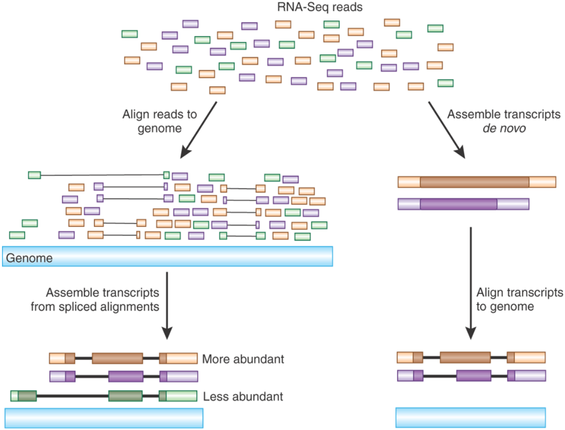


Experimental design

- 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

Two main routes for analysis


RNA-Seq reads
Align reads to genome
Assemble transcripts *de novo*
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

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

Major challenges in relation to genome assembly

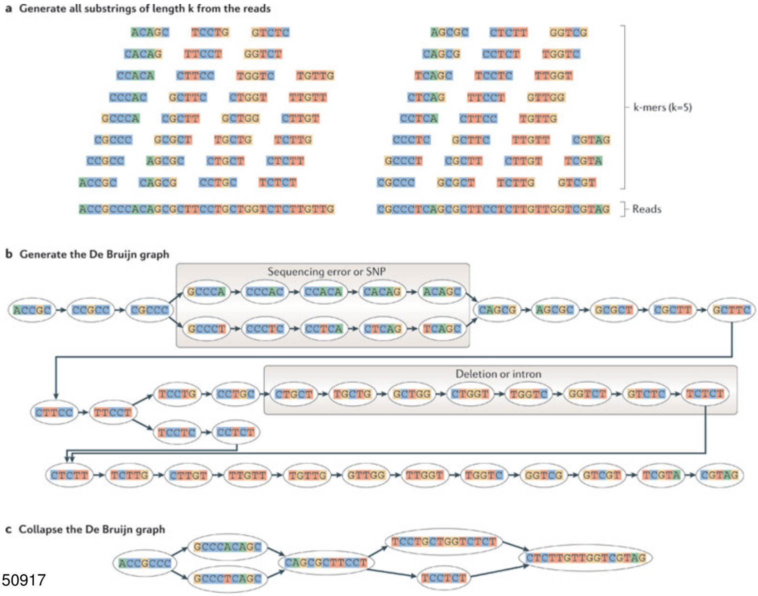
- Genes show different levels of gene expression, hence uneven coverage among genes
- Many genes are expressed in different isoforms
- As sequence depth increase detected number of loci increase. (What is actually expressed?)
- Sequence error from highly expressed genes might be seen more often than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

Several programs available

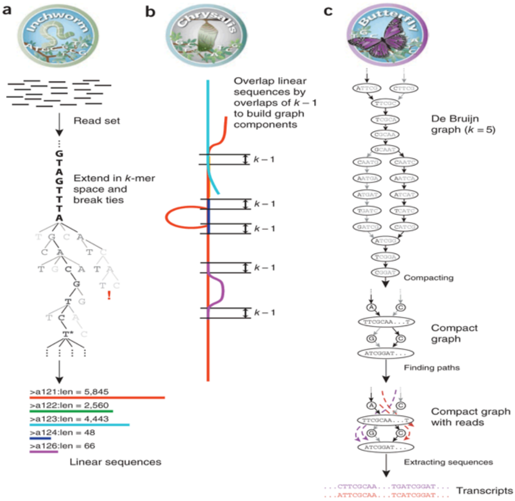
- SOAP-denovo TRANS
- Oases
- Trans-ABYSS
- Trinity

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


a Generate all substrings of length k from the reads
ACAGC TCCTG GTCTC
CACAG TTCCT GGTCT
CCACA CTTCC TGGTC TGTTG
CCCAC GCTTC CTGGT TTGTT
GCCCA CGCTT GCTGG CTTGT
CGCCC GCGCT TGCTG TCTTG
CCGCC AGCGC CTGCT CTCTT
ACCGC CAGCG CCTGC TCTCT
AGCGC CTCTT GGTCG
CAGCG CCTCT TGGTC
TCAGC TCCTC TTGGT
CTCAG TTCCT GTTGG
CCTCA CTTCC TGTTG
CCCTC GCTTC TTGTT CGTAG
GCCCT CGCTT CTTGT TCGTA
CGCCC GCGCT TCTTG GTCGT
k-mers (k=5)
ACCGCCCACAGCGCTTCCTGCTGGTCTCTTGTTG
CGCCCTCAGCGCTTCCTCTTGTTGGTCGTAG
Reads
b Generate the De Bruijn graph
ACCGC CCGCC CGCCC
Sequencing error or SNP
GCCCA CCCAC CCACA CACAG ACAGC
GCCCT CCCTC CCTCA CTCAG TCAGC
CAGCG AGCGC GCGCT CGCTT GCTTC
CTTCC TTCCT
TCCTG CCTGC
Deletion or intron
CTGCT TGCTG GCTGG CTGGT TGGTC GGTCT GTCTC TCTCT
TCCTC CCTCT
CTCTT TCTTG CTTGT TTGTT TGTTG GTTGG TTGGT TGGTC GGTCG GTCGT TCGTA CGTAG
c Collapse the De Bruijn graph
ACCGCCC
GCCCACAGC
GCCCTCAGC
CAGCGCTTCCT
TCCTGCTGGTCTCT
TCCTCT
CTCTTGTTGGTCGTAG

Trinity


a
Inchworm
Read set
G
T
A
G
T
T
T
A
Extend in *k*-mer space and break ties
!
T*
>a121:len = 5,845
>a122:len = 2,560
>a123:len = 4,443
>a124:len = 48
>a126:len = 66
Linear sequences
b
Chrysalis
Overlap linear sequences by overlaps of *k* – 1 to build graph components
k – 1
k – 1
k – 1
k – 1
k – 1
c
Butterfly
ATTCG
CTTCG
TTCGC
TCGCA
CGCAA
GCAAT
CAATG
CAATC
AATGA
AATCA
ATGAT
ATCAT
TGATC
TCATC
GATCG
CATCG
ATCGG
TCGGA
CGGAT
De Bruijn graph (*k* = 5)
Compacting
A
C
TTCGCAA. . .T
G
C
ATCGGAT. . .
Compact graph
Finding paths
Compact graph with reads
Extracting sequences
. . .CTTCGCAA. . .TGATCGGAT. . .
. . .ATTCGCAA. . .TCATCGGAT. . .
Transcripts

Summary - with ref.

- Map to genome allow for spliced alignment
- If novel transcripts of interest: use method that can re-create transcripts from mapped reads (Cufflinks, Scripture or Bayesembler)
  NB! In well annotated genomes most reads should map to known genes
- If interest is expression of known genes/exons: Use available annotation for analysis
- Spend time on experimental design and more replicates gives more power in gene expression analysis

Summary - without ref.

- Assemble using your favourite assembler
- Spend lots of time in assessing the results (compare to related species, look for ORFs etc)
- Often large number of partial transcripts (hence often large number of contigs).
- Merge with other data from transcripts?