Genomics Research Hub
School of Biosciences

The Biocomputing Hub

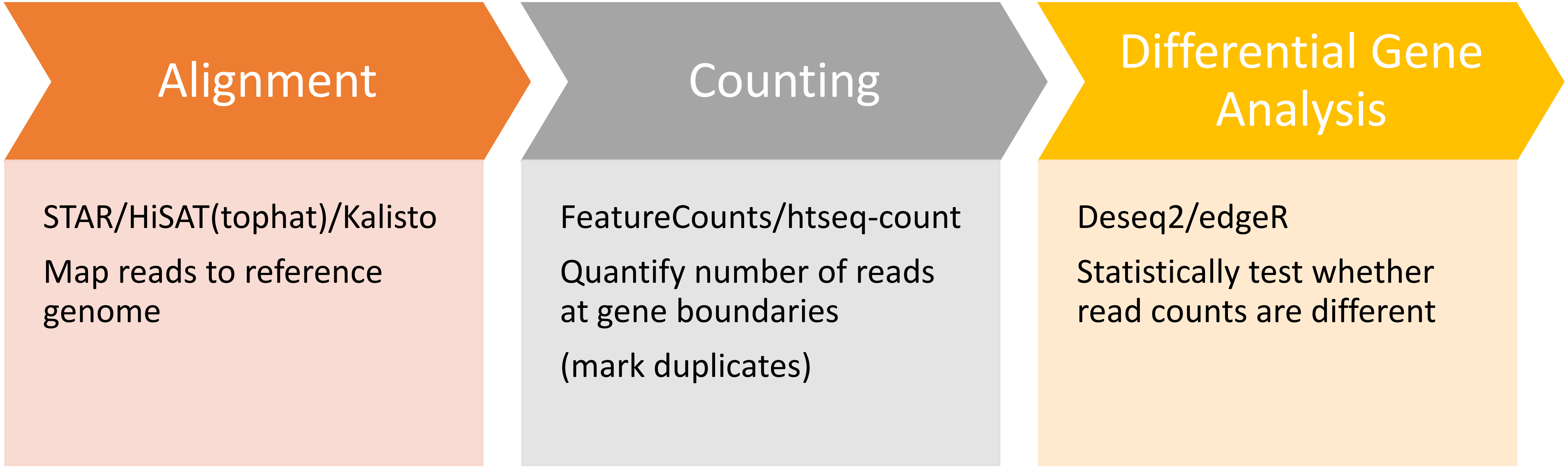
CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Processing RNAseq

- A jungle of file formats

Prof Peter Kille

Process



Alignment

STAR/HiSAT(tophat)/Kalisto

Map reads to reference genome

Counting

FeatureCounts/htseq-count

Quantify number of reads at gene boundaries

(mark duplicates)

Differential Gene Analysis

Deseq2/edgeR

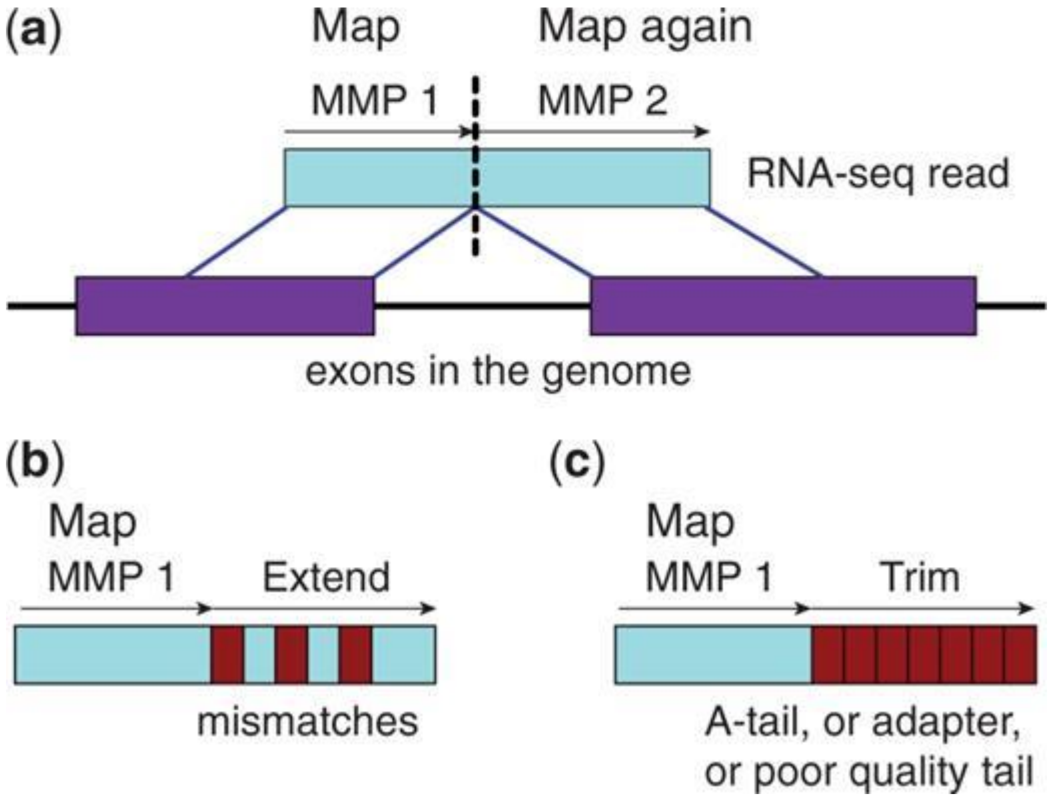
Statistically test whether read counts are different

Alignment with Star

Spliced Transcripts Alignment to a Reference (STAR)

- Novel RNA-seq alignment algorithm that uses sequential maximum mappable seed search.
- "STAR outperforms other aligners by a factor of >50 in mapping speed"
- Note: Multimapping

Combine reference genome (fasta), gene boundaries (gtf) and RNAseq (fastq)

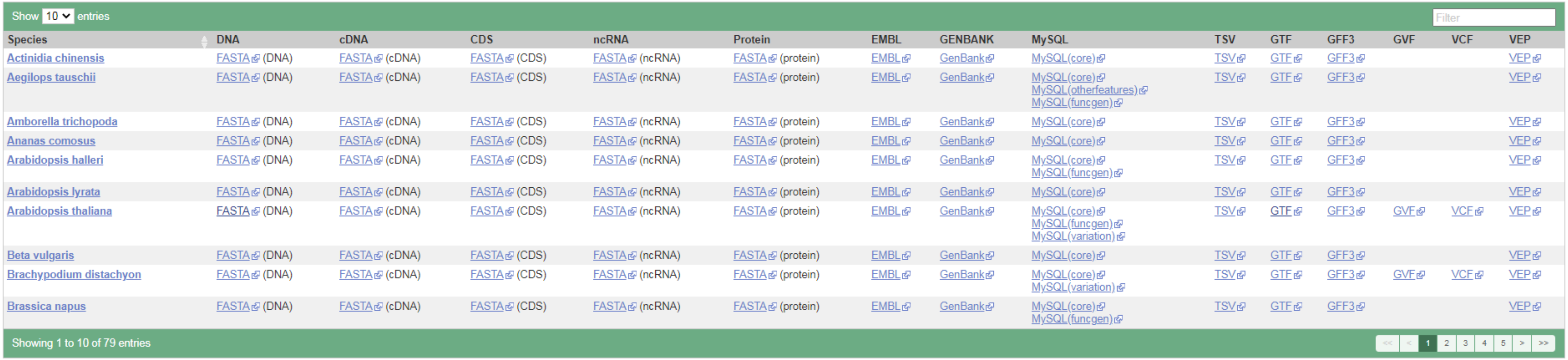


Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, Thomas R. Gingeras, STAR: ultrafast universal RNA-seq aligner, *Bioinformatics*, Volume 29, Issue 1, January 2013, Pages 15–21, https://doi.org/10.1093/bioinformatics/bts635

Ncbi booooooooooo!

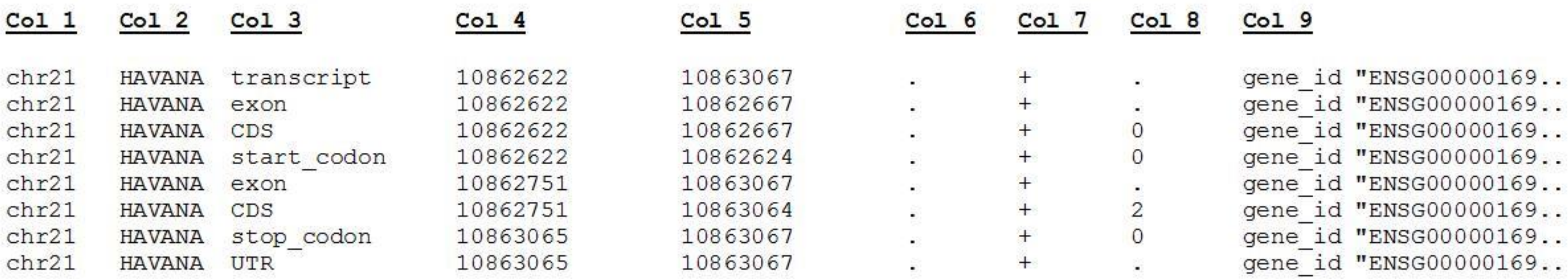
Ensmbl reference databases

https://plants.ensembl.org/info/website/ftp/index.html



Species	DNA	cDNA	CDS	ncRNA	Protein	EMBL	GENBANK	MySQL	TSV	GTF	GFF3	GVF	VCF	VEP
Actinidia chinensis	FASTA (DNA)	FASTA (cDNA)	FASTA (CDS)	FASTA (ncRNA)	FASTA (protein)	EMBL	GenBank	MySQL(core)	TSV	GTF	GFF3			VEP
Aegilops tauschii	FASTA (DNA)	FASTA (cDNA)	FASTA (CDS)	FASTA (ncRNA)	FASTA (protein)	EMBL	GenBank	MySQL(core) MySQL(otherfeatures) MySQL(funcgen)	TSV	GTF	GFF3			VEP
Amborella trichopoda	FASTA (DNA)	FASTA (cDNA)	FASTA (CDS)	FASTA (ncRNA)	FASTA (protein)	EMBL	GenBank	MySQL(core)	TSV	GTF	GFF3			VEP
Ananas comosus	FASTA (DNA)	FASTA (cDNA)	FASTA (CDS)	FASTA (ncRNA)	FASTA (protein)	EMBL	GenBank	MySQL(core)	TSV	GTF	GFF3			VEP
Arabidopsis halleri	FASTA (DNA)	FASTA (cDNA)	FASTA (CDS)	FASTA (ncRNA)	FASTA (protein)	EMBL	GenBank	MySQL(core) MySQL(funcgen)	TSV	GTF	GFF3			VEP
Arabidopsis lyrata	FASTA (DNA)	FASTA (cDNA)	FASTA (CDS)	FASTA (ncRNA)	FASTA (protein)	EMBL	GenBank	MySQL(core)	TSV	GTF	GFF3			VEP
Arabidopsis thaliana	FASTA (DNA)	FASTA (cDNA)	FASTA (CDS)	FASTA (ncRNA)	FASTA (protein)	EMBL	GenBank	MySQL(core) MySQL(funcgen) MySQL(variation)	TSV	GTF	GFF3	GVF	VCF	VEP
Beta vulgaris	FASTA (DNA)	FASTA (cDNA)	FASTA (CDS)	FASTA (ncRNA)	FASTA (protein)	EMBL	GenBank	MySQL(core)	TSV	GTF	GFF3			VEP
Brachypodium distachyon	FASTA (DNA)	FASTA (cDNA)	FASTA (CDS)	FASTA (ncRNA)	FASTA (protein)	EMBL	GenBank	MySQL(core) MySQL(variation)	TSV	GTF	GFF3	GVF	VCF	VEP
Brassica napus	FASTA (DNA)	FASTA (cDNA)	FASTA (CDS)	FASTA (ncRNA)	FASTA (protein)	EMBL	GenBank	MySQL(core) MySQL(funcgen)	TSV	GTF	GFF3			VEP

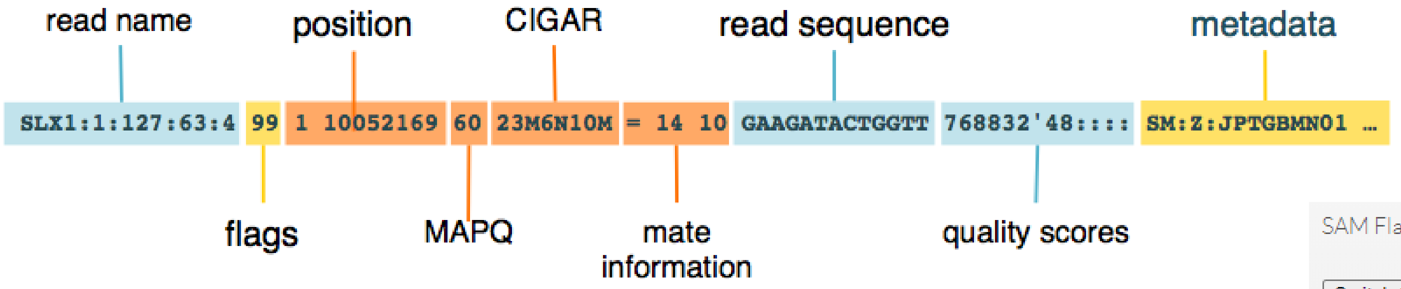
Showing 1 to 10 of 79 entries



Col 1	Col 2	Col 3	Col 4	Col 5	Col 6	Col 7	Col 8	Col 9
chr21	HAVANA	transcript	10862622	10863067	.	+	.	gene_id "ENSG00000169..
chr21	HAVANA	exon	10862622	10862667	.	+	.	gene_id "ENSG00000169..
chr21	HAVANA	CDS	10862622	10862667	.	+	0	gene_id "ENSG00000169..
chr21	HAVANA	start_codon	10862622	10862624	.	+	0	gene_id "ENSG00000169..
chr21	HAVANA	exon	10862751	10863067	.	+	.	gene_id "ENSG00000169..
chr21	HAVANA	CDS	10862751	10863064	.	+	2	gene_id "ENSG00000169..
chr21	HAVANA	stop_codon	10863065	10863067	.	+	0	gene_id "ENSG00000169..
chr21	HAVANA	UTR	10863065	10863067	.	+	.	gene_id "ENSG00000169..

Output from alignment: bam/sam files

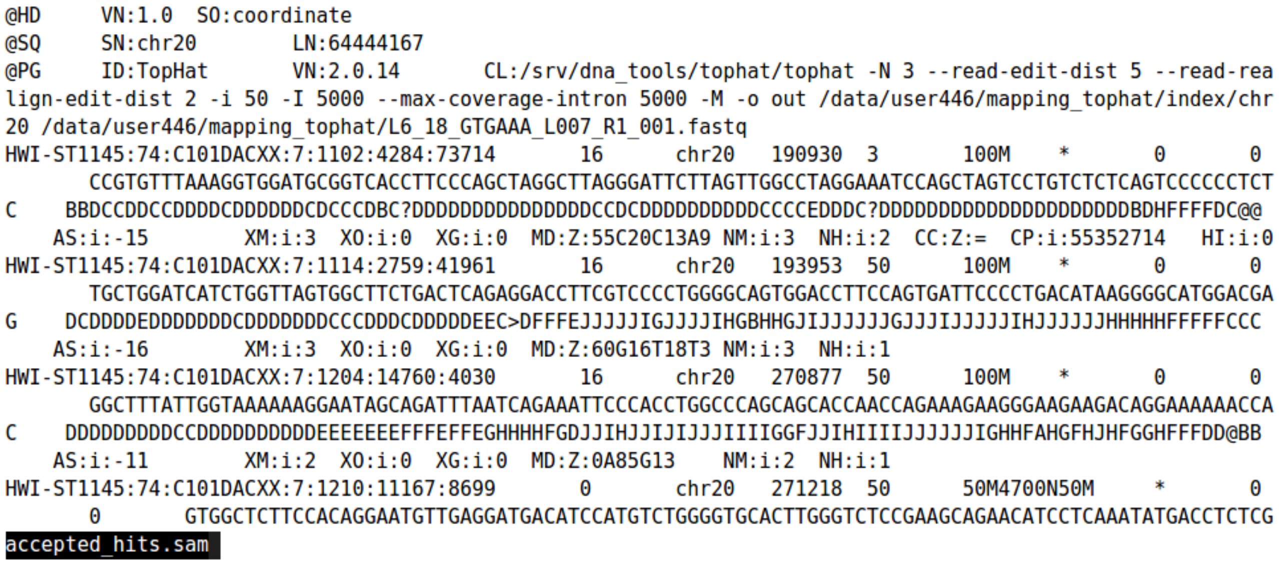
HEADER containing metadata (sequence dictionary, read group definitions etc)
RECORDS containing structured read information (1 line per read record)



read name
position
CIGAR
read sequence
metadata

SLX1:1:127:63:4 99 1 10052169 60 23M6N10M = 14 10 GAAGATACTGGTT 768832'48:::: SM:Z:JPTGBMN01 …

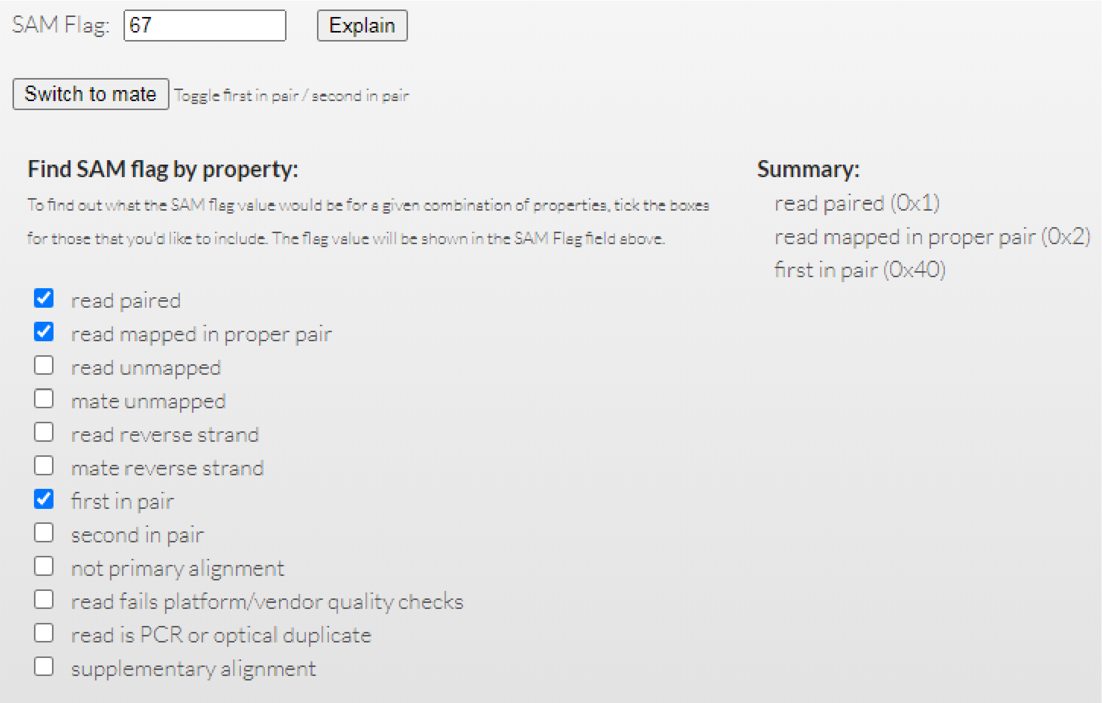
flags
MAPQ
mate information
quality scores



```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714       16      chr20   190930  3       100M    *       0       0
        CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C     BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961       16      chr20   193953  50      100M    *       0       0
        TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G     DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030       16      chr20   270877  50      100M    *       0       0
        GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C     DDDDDDDDDCCDDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699       0       chr20   271218  50      50M4700N50M     *       0
        0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```

Optional step: Mark Duplicates

- Picard



SAM Flag: 67

Find SAM flag by property:

To find out what the SAM flag value would be for a given combination of properties, tick the boxes for those that you'd like to include. The flag value will be shown in the SAM Flag field above.

☑ read paired
☑ read mapped in proper pair
☐ read unmapped
☐ mate unmapped
☐ read reverse strand
☐ mate reverse strand
☑ first in pair
☐ second in pair
☐ not primary alignment
☐ read fails platform/vendor quality checks
☐ read is PCR or optical duplicate
☐ supplementary alignment

Summary:

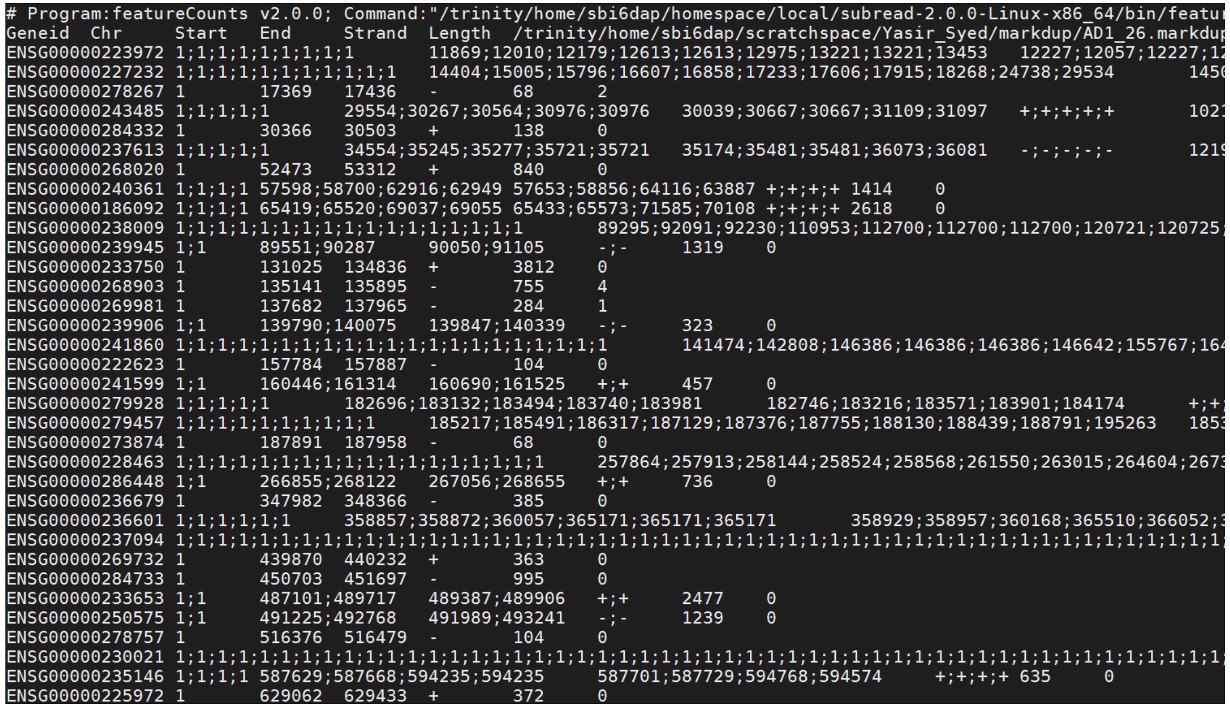
read paired (0x1)
read mapped in proper pair (0x2)
first in pair (0x40)

https://broadinstitute.github.io/picard/explain-flags.html

Counting

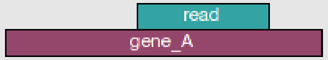
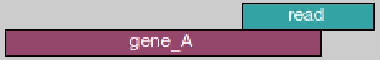
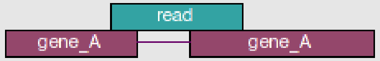
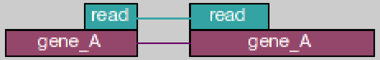
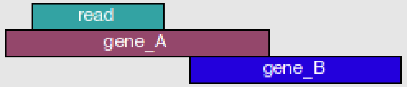
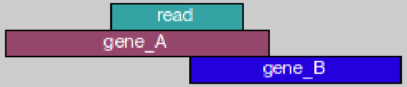
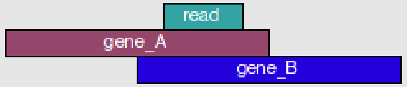
FeatureCounts part of subread

- a software program developed for counting reads to genomic features such as genes, exons, promoters and genomic bins.
  - ~20x faster
  - More definitive on ambiguities
  - Better with paired end data
  - http://subread.sourceforge.net/



```
# Program:featureCounts v2.0.0; Command:"/trinity/home/sbi6dap/homespace/local/subread-2.0.0-Linux-x86_64/bin/featu
Geneid  Chr     Start   End     Strand  Length  /trinity/home/sbi6dap/scratchspace/Yasir_Syed/markdup/AD1_26.markdup
ENSG00000223972 1;1;1;1;1;1;1;1;1       11869;12010;12179;12613;12613;12975;13221;13221;13453   12227;12057;12227;12
ENSG00000227232 1;1;1;1;1;1;1;1;1;1;1   14404;15005;15796;16607;16858;17233;17606;17915;18268;24738;29534       1450
ENSG00000278267 1       17369   17436   -       68      2
ENSG00000243485 1;1;1;1;1       29554;30267;30564;30976;30976   30039;30667;30667;31109;31097   +;+;+;+;+       102
ENSG00000284332 1       30366   30503   +       138     0
ENSG00000237613 1;1;1;1;1       34554;35245;35277;35721;35721   35174;35481;35481;36073;36081   -;-;-;-;-       1219
ENSG00000268020 1       52473   53312   +       840     0
ENSG00000240361 1;1;1;1 57598;58700;62916;62949 57653;58856;64116;63887 +;+;+;+ 1414    0
ENSG00000186092 1;1;1;1 65419;65520;69037;69055 65433;65573;71585;70108 +;+;+;+ 2618    0
ENSG00000238009 1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1       89295;92091;92230;110953;112700;112700;112700;120721;120725;
ENSG00000239945 1;1     89551;90287     90050;91105     -;-     1319    0
ENSG00000233750 1       131025  134836  +       3812    0
ENSG00000268903 1       135141  135895  -       755     4
ENSG00000269981 1       137682  137965  -       284     1
ENSG00000239906 1;1     139790;140075   139847;140339   -;-     323     0
ENSG00000241860 1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1       141474;142808;146386;146386;146386;146642;155767;164
ENSG00000222623 1       157784  157887  -       104     0
ENSG00000241599 1;1     160446;161314   160690;161525   +;+     457     0
ENSG00000279928 1;1;1;1;1       182696;183132;183494;183740;183981      182746;183216;183571;183901;184174      +;+;
ENSG00000279457 1;1;1;1;1;1;1;1;1;1     185217;185491;186317;187129;187376;187755;188130;188439;188791;195263   1853
ENSG00000273874 1       187891  187958  -       68      0
ENSG00000228463 1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1     257864;257913;258144;258524;258568;261550;263015;264604;2673
ENSG00000286448 1;1     266855;268122   267056;268655   +;+     736     0
ENSG00000236679 1       347982  348366  -       385     0
ENSG00000236601 1;1;1;1;1;1     358857;358872;360057;365171;365171;365171       358929;358957;360168;365510;366052;3
ENSG00000237094 1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;
ENSG00000269732 1       439870  440232  +       363     0
ENSG00000284733 1       450703  451697  -       995     0
ENSG00000233653 1;1     487101;489717   489387;489906   +;+     2477    0
ENSG00000250575 1;1     491225;492768   491989;493241   -;-     1239    0
ENSG00000278757 1       516376  516479  -       104     0
ENSG00000230021 1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;
ENSG00000235146 1;1;1;1 587629;587668;594235;594235     587701;587729;594768;594574     +;+;+;+ 635     0
ENSG00000225972 1       629062  629433  +       372     0
```

Htseq

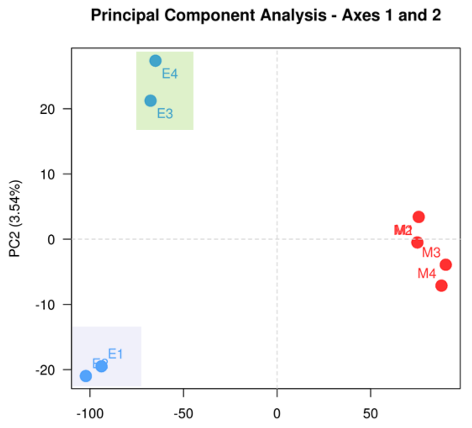
	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

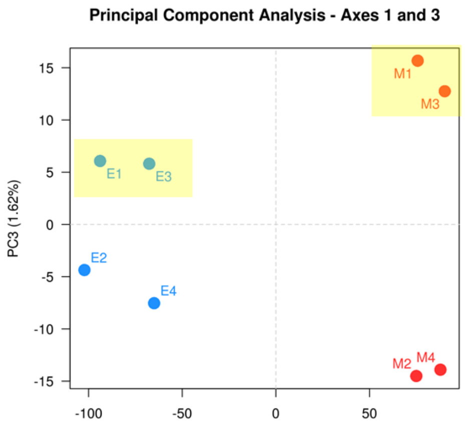
Differential gene Analysis

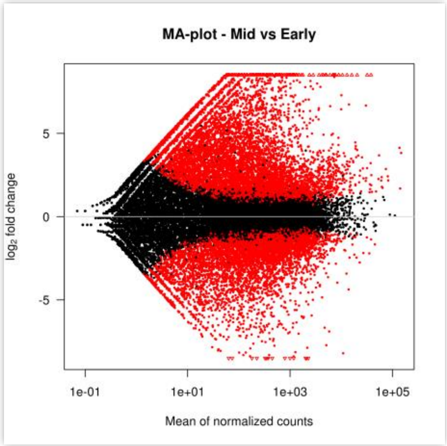
Deseq2 vs edgeR

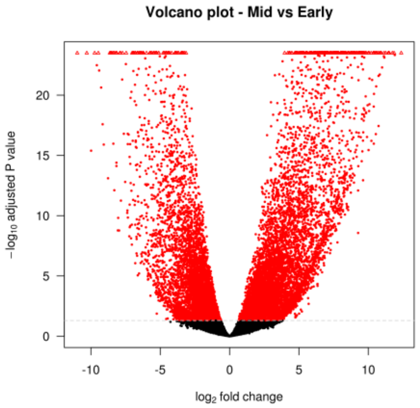
- Both do the same:
  - T-tests / Volcano plots
  - Log2(fold change) vs mean count (MA plots /FC plots)
  - Dendrograms
  - Principal Component Analysis (PCA)
  - ANOVA
  - Multiple Sample correction
  - Hierarchical Cluster Analysis

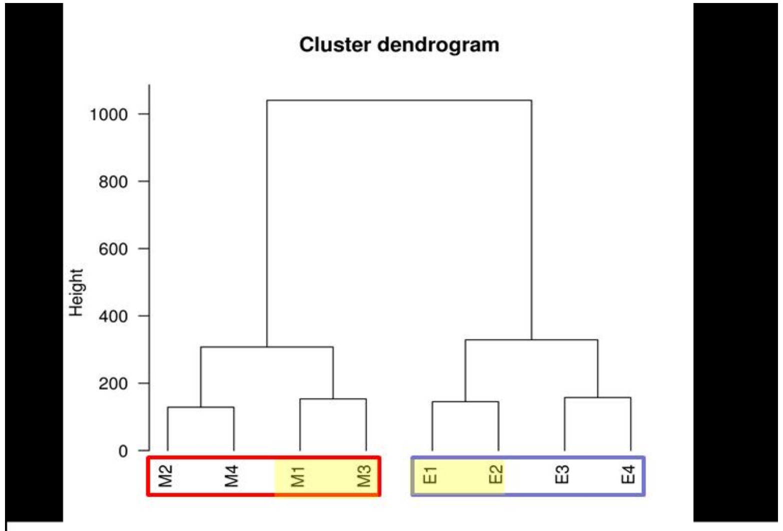
More next time!

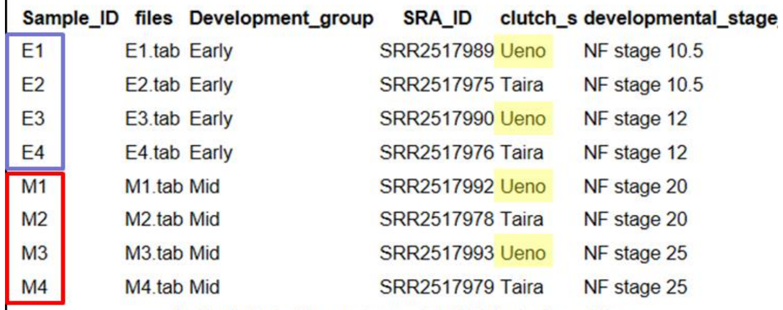












Sample_ID	files	Development_group	SRA_ID	clutch_s	developmental_stage
E1	E1.tab	Early	SRR2517989	Ueno	NF stage 10.5
E2	E2.tab	Early	SRR2517975	Taira	NF stage 10.5
E3	E3.tab	Early	SRR2517990	Ueno	NF stage 12
E4	E4.tab	Early	SRR2517976	Taira	NF stage 12
M1	M1.tab	Mid	SRR2517992	Ueno	NF stage 20
M2	M2.tab	Mid	SRR2517978	Taira	NF stage 20
M3	M3.tab	Mid	SRR2517993	Ueno	NF stage 25
M4	M4.tab	Mid	SRR2517979	Taira	NF stage 25

Table 1: Data files and associated biological conditions.