ETH zürich

High Performance Computing for genomic applications

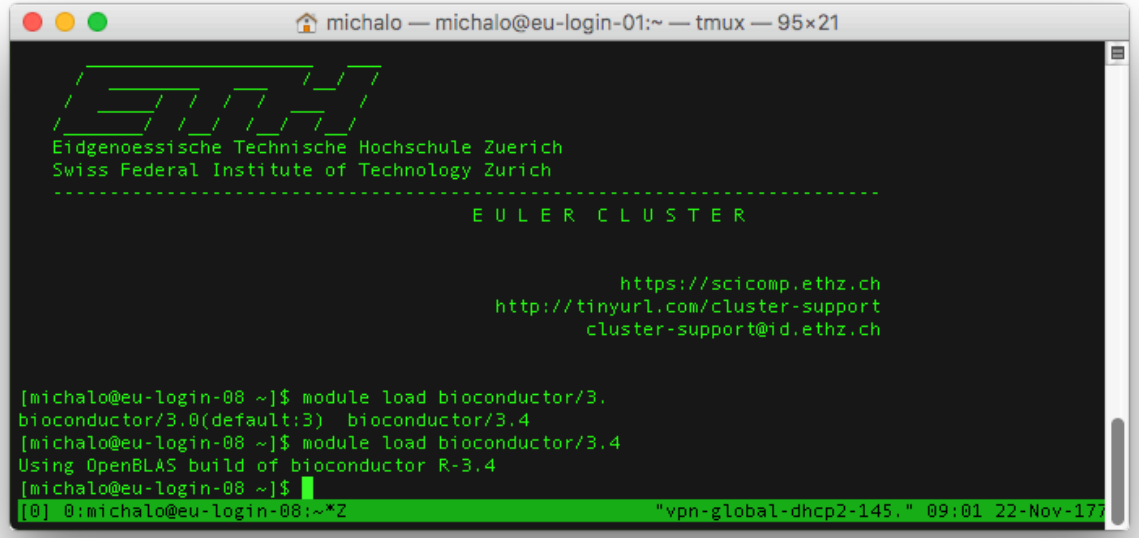
Using genomic software on Euler

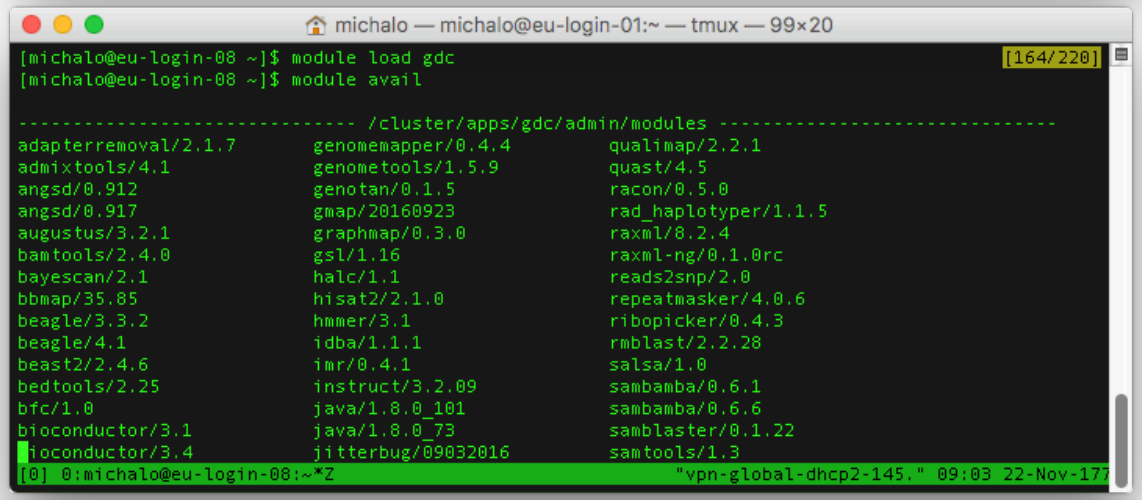
Scientific IT Services
Michal Okoniewski

Bioinformatic modules on EULER

- module load
- module load gdc
- module avail
- module purge
- env2lmod, lmod2env

- https://scicomp.ethz.ch/wiki/New_SPACK_software_stack_on_Euler

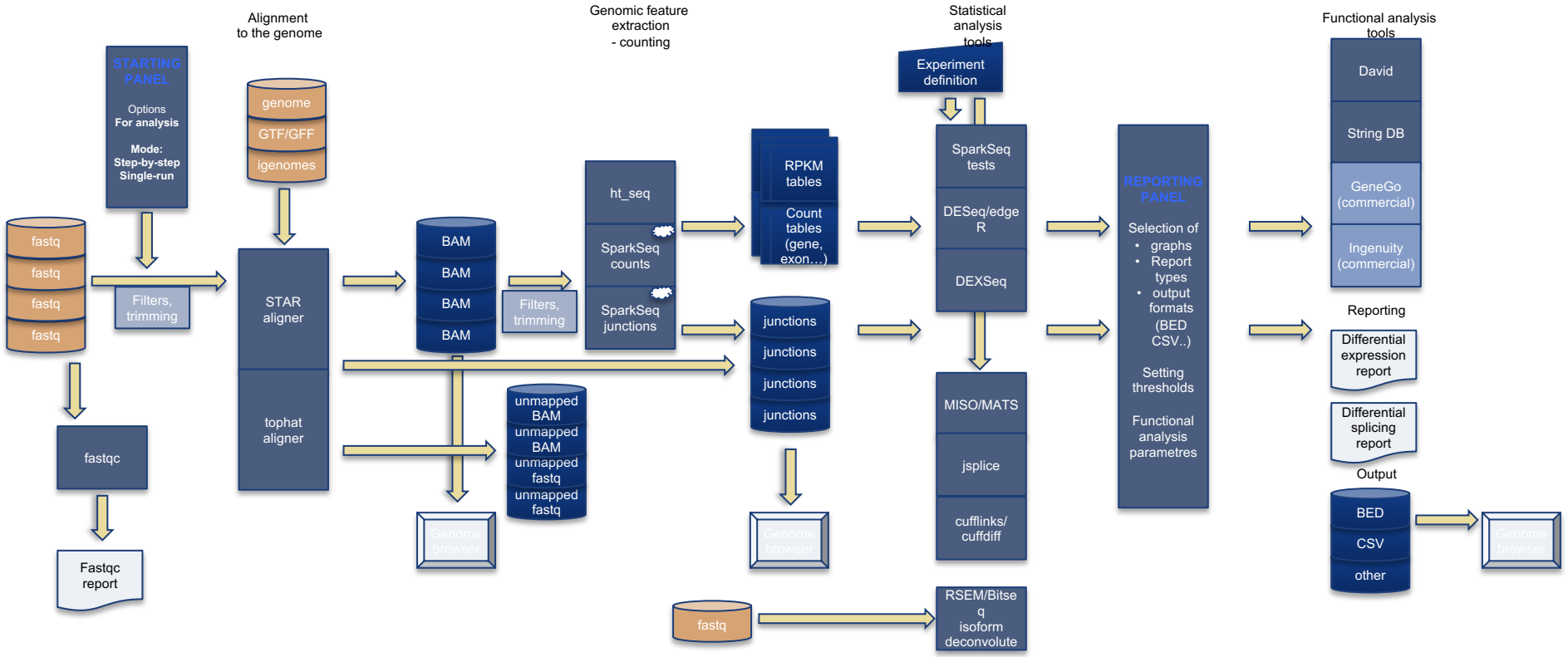




Bioinformatic software jungle - categories

- Trimming tools: trimmomatic, cutadapt
- Aligners:
  - General purpose: bwa, bowtie, SHRiMP
  - RNA aligners: STAR, tophat, subjunc, hisat2
  - Transcriptome aligners: kallistio, salmon, sailfish, RSEM
  - Other aligners: Blast, Blat, VMATCH
- De-novo assemblers: trinity, velvet, spades
- Feature extraction, counting: HTSeq, featureCount
  - Transcript discovery: cufflinks,
- Specialized tools: MISO, blast2go,…
- Utilities and conversion tools: samtools, bcftools, bedtools, picard tools

Bioinformatic software jungle (as seen back in 2015)



Trimming tools

- trimmomatic, cutadapt
- Trimming:
  - Removing adapters
  - By quality: eg sliding window
  - Fixed position, eg several base pairs at each end
- Typicaly: fastq on input, fastq on output
  - itets more complex with paired reads

Trimming tools

```
module load java
module load trimmomatic

trimmomatic SE {input} {output}  ILLUMINACLIP:adapters.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

```
>TruSeqR2_nextera
CTGTCTCTTATACACATCTCCGAGCCCACGAGACTAAGGCGAATCTCGTATGCCGTCTTCTGCTTGAAAA
>nextera_R2_right_side_adapter
CTGTCTCTTATACACATCTGACGCTGCCGACGAGCGATCTAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAA
>I5_Nextera_Transposase_1
CTGTCTCTTATACACATCTGACGCTGCCGACGA
>I7_Nextera_Transposase_1
CTGTCTCTTATACACATCTCCGAGCCCACGAGAC
>I5_Nextera_Transposase_2
CTGTCTCTTATACACATCTCTGATGGCGCGAGGGAGGC
>I7_Nextera_Transposase_2
CTGTCTCTTATACACATCTCTGAGCGGGCTGGCAAGGC
>I5_Primer_Nextera_XT_and_Nextera_Enrichment_[N/S/E]501
GACGCTGCCGACGAGCGATCTAGTGTAGATCTCGGTGGTCGCCGTATCATT
>I5_Primer_Nextera_XT_and_Nextera_Enrichment_[N/S/E]502__and__I5_Primer_Nextera_XT_Index_Kit_v2_S502
GACGCTGCCGACGAATAGAGAGGTGTAGATCTCGGTGGTCGCCGTATCATT
>I5_Primer_Nextera_XT_and_Nextera_Enrichment_[N/S/E]503__and__I5_Primer_Nextera_XT_Index_Kit_v2_S503
GACGCTGCCGACGAAGAGGATAGTGTAGATCTCGGTGGTCGCCGTATCATT
>I5_Primer_Nextera_XT_and_Nextera_Enrichment_[N/S/E]504
GACGCTGCCGACGATCTACTCTGTGTAGATCTCGGTGGTCGCCGTATCATT
>I5_Primer_Nextera_XT_and_Nextera_Enrichment_[N/S/E]505__and__I5_Primer_Nextera_XT_Index_Kit_v2_S505
GACGCTGCCGACGACTCCTTACGTGTAGATCTCGGTGGTCGCCGTATCATT
>I5_Primer_Nextera_XT_and_Nextera_Enrichment_[N/S/E]506__and__I5_Primer_Nextera_XT_Index_Kit_v2_S506
GACGCTGCCGACGATATGCAGTGTGTAGATCTCGGTGGTCGCCGTATCATT
>I5_Primer_Nextera_XT_and_Nextera_Enrichment_[N/S/E]507__and__I5_Primer_Nextera_XT_Index_Kit_v2_S507
GACGCTGCCGACGATACTCCTTGTGTAGATCTCGGTGGTCGCCGTATCATT
```

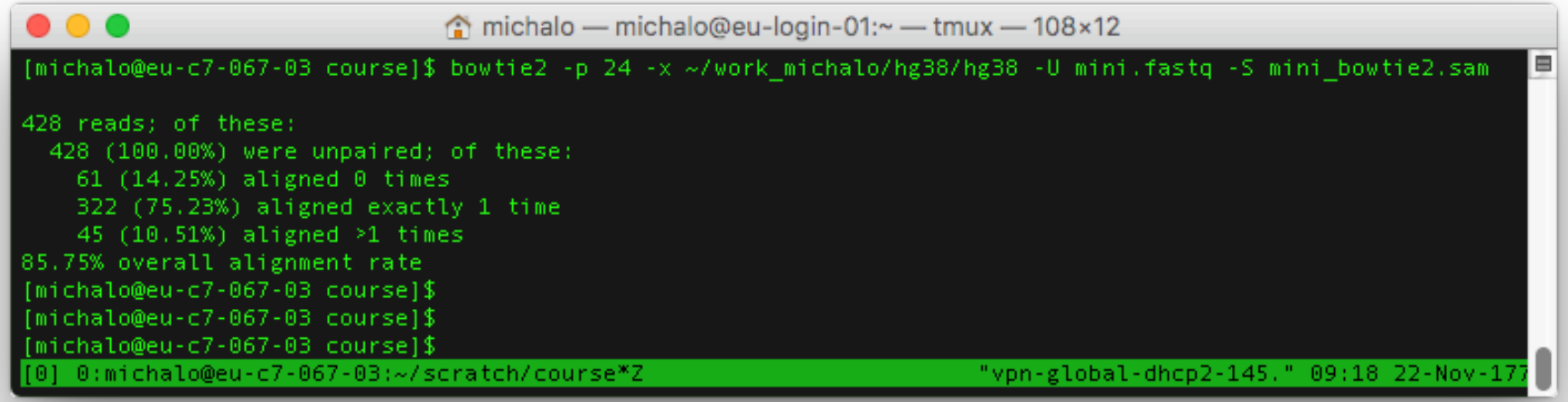
Bowtie, bowtie2

- Building genome index

```
Read:      GACTGGGCGATCTCGACTTCG
           |||||  ||||||||||| |||
Reference: GACTG--CGATCTCGACATCG
```

```
bowtie2-build  --threads 24
/cluster/home/michalo/work_michalo/hg38/Homo_sapiens.GRCh38.dna.primary_assembly.fa
/cluster/scratch/michalo/hg38/hg38
```

- Alignment:



Tophat

- Classic splice-aware aligner
- Uses bowtie2 as engine, so also bowtie2 index

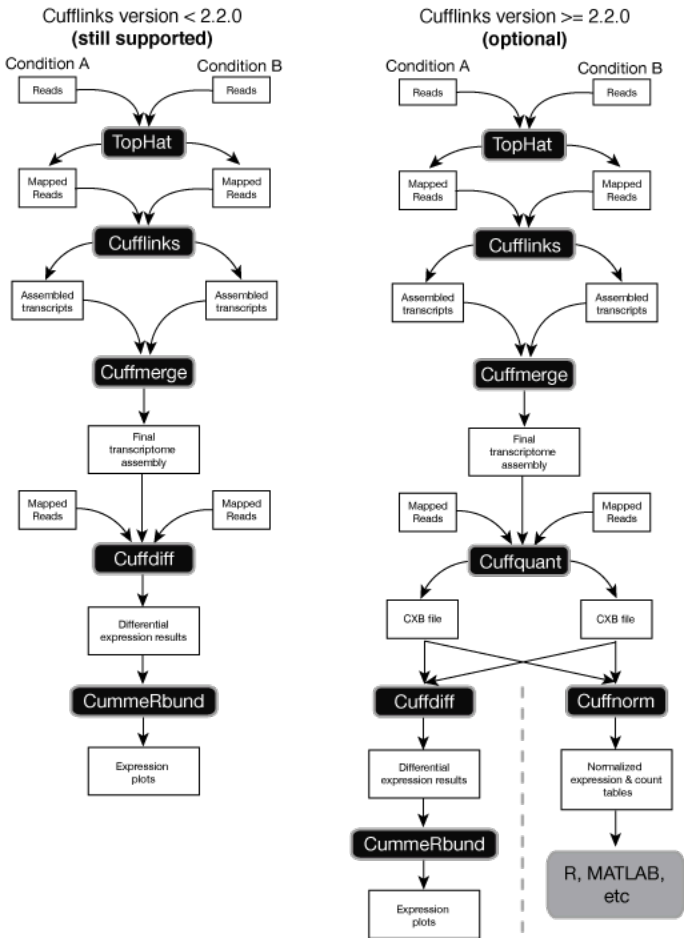
```
tophat -p 24 -o tophat_out --library-type fr-firststrand ~/work_michalo/hg38/hg38 mini.fastq.gz
```

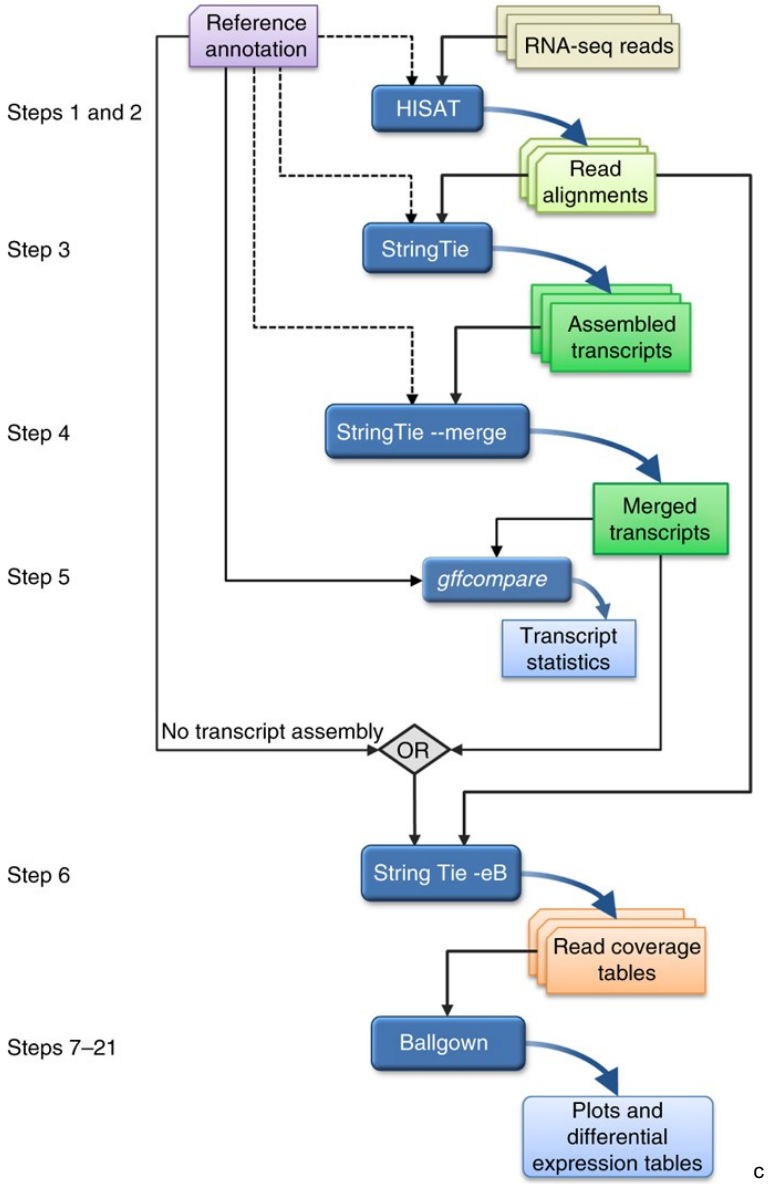
- Manual: http://ccb.jhu.edu/software/tophat/index.shtml

Splice aware aligners for RNA-seq, memory needs

	Index type	Index size=memory needs for hg38	Computing node
Tophat, 2011	bowtie index in a file	in a file	any
STAR, 2014	In memory	ca 40GB	EULER normal (65G)
Subjunc, 2014	In memory	ca 40GB	EULER normal (65G)
Hisat2, 2016	In memory	ca 7GB	good laptop (>=8G RAM)

Tuxedo suite and “new tuxedo”





“New tuxedo suite”

nature protocols

Explore Content | Journal Information | Publish With Us

nature > nature protocols > protocols > article

Published: 11 August 2016

Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown

Mihaela Pertea, Daehwan Kim, Geo M Pertea, Jeffrey T Leek & Steven L Salzberg

Nature Protocols **11**, 1650–1667(2016) | Cite this article

Cufflinks and stringtie

- Transcript discovery tools
- Uses coverage and junctions from a BAM file

```
cufflinks mini_star.sorted.bam
```

- Other
  - cuffmerge, cuffdiff, cuffquant, cuffnorm, CummeRbund

Cufflinks and stringtie

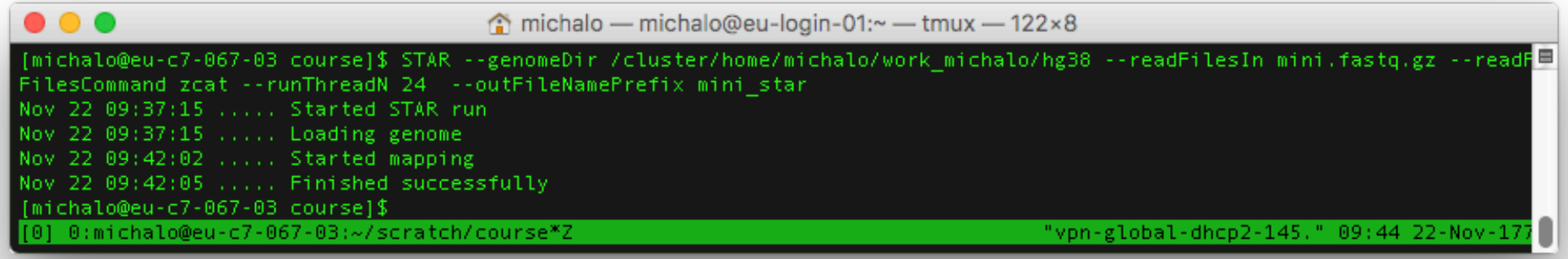
- Produces GTF

michalo — michalo@eu-login-01:~ — tmux — 265×41

```
17	Cufflinks	transcript	41801952	41802294	1000	.	.	gene_id "CUFF.1"; transcript_id "CUFF.1.1"; FPKM "385640.0529891439"; frac "1.000000"; conf_lo "224014.123428"; conf_hi "547265.982551"; cov "8.014439";
17	Cufflinks	exon	41801952	41802294	1000	.	.	gene_id "CUFF.1"; transcript_id "CUFF.1.1"; exon_number "1"; FPKM "385640.0529891439"; frac "1.000000"; conf_lo "224014.123428"; conf_hi "547265.982551"; cov "8.014439";
17	Cufflinks	transcript	41802365	41803426	337	.	.	gene_id "CUFF.2"; transcript_id "CUFF.2.1"; FPKM "66188.3599707013"; frac "0.252522"; conf_lo "20863.162216"; conf_hi "111513.557726"; cov "1.382272";
17	Cufflinks	exon	41802365	41802981	337	.	.	gene_id "CUFF.2"; transcript_id "CUFF.2.1"; exon_number "1"; FPKM "66188.3599707013"; frac "0.252522"; conf_lo "20863.162216"; conf_hi "111513.557726"; cov "1.382272";
17	Cufflinks	exon	41803289	41803426	337	.	.	gene_id "CUFF.2"; transcript_id "CUFF.2.1"; exon_number "2"; FPKM "66188.3599707013"; frac "0.252522"; conf_lo "20863.162216"; conf_hi "111513.557726"; cov "1.382272";
17	Cufflinks	transcript	41802365	41803426	1000	.	.	gene_id "CUFF.2"; transcript_id "CUFF.2.2"; FPKM "195920.7516271057"; frac "0.747478"; conf_lo "129358.547713"; conf_hi "262482.955541"; cov "4.091593";
17	Cufflinks	exon	41802365	41802979	1000	.	.	gene_id "CUFF.2"; transcript_id "CUFF.2.2"; exon_number "1"; FPKM "195920.7516271057"; frac "0.747478"; conf_lo "129358.547713"; conf_hi "262482.955541"; cov "4.091593";
17	Cufflinks	exon	41803287	41803426	1000	.	.	gene_id "CUFF.2"; transcript_id "CUFF.2.2"; exon_number "2"; FPKM "195920.7516271057"; frac "0.747478"; conf_lo "129358.547713"; conf_hi "262482.955541"; cov "4.091593";
17	Cufflinks	transcript	41806822	41809842	1000	.	.	gene_id "CUFF.3"; transcript_id "CUFF.3.1"; FPKM "390983.8092033406"; frac "1.000000"; conf_lo "222379.790173"; conf_hi "559587.828234"; cov "8.190210";
17	Cufflinks	exon	41806822	41806879	1000	.	.	gene_id "CUFF.3"; transcript_id "CUFF.3.1"; exon_number "1"; FPKM "390983.8092033406"; frac "1.000000"; conf_lo "222379.790173"; conf_hi "559587.828234"; cov "8.190210";
17	Cufflinks	exon	41807859	41808004	1000	.	.	gene_id "CUFF.3"; transcript_id "CUFF.3.1"; exon_number "2"; FPKM "390983.8092033406"; frac "1.000000"; conf_lo "222379.790173"; conf_hi "559587.828234"; cov "8.190210";
17	Cufflinks	exon	41809706	41809842	1000	.	.	gene_id "CUFF.3"; transcript_id "CUFF.3.1"; exon_number "3"; FPKM "390983.8092033406"; frac "1.000000"; conf_lo "222379.790173"; conf_hi "559587.828234"; cov "8.190210";
17	Cufflinks	transcript	41810873	41811019	1000	.	.	gene_id "CUFF.4"; transcript_id "CUFF.4.1"; FPKM "2721172.9130021906"; frac "1.000000"; conf_lo "1080243.905346"; conf_hi "4362101.920658"; cov "56.977279";
17	Cufflinks	exon	41810873	41811019	1000	.	.	gene_id "CUFF.4"; transcript_id "CUFF.4.1"; exon_number "1"; FPKM "2721172.9130021906"; frac "1.000000"; conf_lo "1080243.905346"; conf_hi "4362101.920658"; cov "56.977279";
17	Cufflinks	transcript	41811134	41811385	1000	.	.	gene_id "CUFF.5"; transcript_id "CUFF.5.1"; FPKM "386460.4071371864"; frac "1.000000"; conf_lo "136927.788094"; conf_hi "635993.026180"; cov "8.059207";
17	Cufflinks	exon	41811134	41811385	1000	.	.	gene_id "CUFF.5"; transcript_id "CUFF.5.1"; exon_number "1"; FPKM "386460.4071371864"; frac "1.000000"; conf_lo "136927.788094"; conf_hi "635993.026180"; cov "8.059207";
17	Cufflinks	transcript	41811684	41811948	1000	.	.	gene_id "CUFF.6"; transcript_id "CUFF.6.1"; FPKM "365523.9830166674"; frac "1.000000"; conf_lo "145104.727827"; conf_hi "585943.238206"; cov "7.632440";
17	Cufflinks	exon	41811684	41811948	1000	.	.	gene_id "CUFF.6"; transcript_id "CUFF.6.1"; exon_number "1"; FPKM "365523.9830166674"; frac "1.000000"; conf_lo "145104.727827"; conf_hi "585943.238206"; cov "7.632440";
17	Cufflinks	transcript	41825205	41835234	1000	.	.	gene_id "CUFF.7"; transcript_id "CUFF.7.1"; FPKM "386421.1203223147"; frac "0.686190"; conf_lo "311028.834987"; conf_hi "461813.405657"; cov "8.087935";
17	Cufflinks	exon	41825205	41825657	1000	.	.	gene_id "CUFF.7"; transcript_id "CUFF.7.1"; exon_number "1"; FPKM "386421.1203223147"; frac "0.686190"; conf_lo "311028.834987"; conf_hi "461813.405657"; cov "8.087935";
17	Cufflinks	exon	41827426	41827626	1000	.	.	gene_id "CUFF.7"; transcript_id "CUFF.7.1"; exon_number "2"; FPKM "386421.1203223147"; frac "0.686190"; conf_lo "311028.834987"; conf_hi "461813.405657"; cov "8.087935";
17	Cufflinks	exon	41828790	41828952	1000	.	.	gene_id "CUFF.7"; transcript_id "CUFF.7.1"; exon_number "3"; FPKM "386421.1203223147"; frac "0.686190"; conf_lo "311028.834987"; conf_hi "461813.405657"; cov "8.087935";
17	Cufflinks	exon	41830801	41830890	1000	.	.	gene_id "CUFF.7"; transcript_id "CUFF.7.1"; exon_number "4"; FPKM "386421.1203223147"; frac "0.686190"; conf_lo "311028.834987"; conf_hi "461813.405657"; cov "8.087935";
17	Cufflinks	exon	41832392	41832477	1000	.	.	gene_id "CUFF.7"; transcript_id "CUFF.7.1"; exon_number "5"; FPKM "386421.1203223147"; frac "0.686190"; conf_lo "311028.834987"; conf_hi "461813.405657"; cov "8.087935";
17	Cufflinks	exon	41835070	41835116	1000	.	.	gene_id "CUFF.7"; transcript_id "CUFF.7.1"; exon_number "6"; FPKM "386421.1203223147"; frac "0.686190"; conf_lo "311028.834987"; conf_hi "461813.405657"; cov "8.087935";
17	Cufflinks	exon	41835203	41835234	1000	.	.	gene_id "CUFF.7"; transcript_id "CUFF.7.1"; exon_number "7"; FPKM "386421.1203223147"; frac "0.686190"; conf_lo "311028.834987"; conf_hi "461813.405657"; cov "8.087935";
17	Cufflinks	transcript	41825205	41836162	306	.	.	gene_id "CUFF.7"; transcript_id "CUFF.7.2"; FPKM "118402.6643432135"; frac "0.313810"; conf_lo "81225.981370"; conf_hi "155579.347316"; cov "2.478211";
17	Cufflinks	exon	41825205	41825657	306	.	.	gene_id "CUFF.7"; transcript_id "CUFF.7.2"; exon_number "1"; FPKM "118402.6643432135"; frac "0.313810"; conf_lo "81225.981370"; conf_hi "155579.347316"; cov "2.478211";
17	Cufflinks	exon	41827426	41827626	306	.	.	gene_id "CUFF.7"; transcript_id "CUFF.7.2"; exon_number "2"; FPKM "118402.6643432135"; frac "0.313810"; conf_lo "81225.981370"; conf_hi "155579.347316"; cov "2.478211";
17	Cufflinks	exon	41828790	41828952	306	.	.	gene_id "CUFF.7"; transcript_id "CUFF.7.2"; exon_number "3"; FPKM "118402.6643432135"; frac "0.313810"; conf_lo "81225.981370"; conf_hi "155579.347316"; cov "2.478211";
17	Cufflinks	exon	41830801	41830890	306	.	.	gene_id "CUFF.7"; transcript_id "CUFF.7.2"; exon_number "4"; FPKM "118402.6643432135"; frac "0.313810"; conf_lo "81225.981370"; conf_hi "155579.347316"; cov "2.478211";
17	Cufflinks	exon	41832392	41832477	306	.	.	gene_id "CUFF.7"; transcript_id "CUFF.7.2"; exon_number "5"; FPKM "118402.6643432135"; frac "0.313810"; conf_lo "81225.981370"; conf_hi "155579.347316"; cov "2.478211";
17	Cufflinks	exon	41835070	41835272	306	.	.	gene_id "CUFF.7"; transcript_id "CUFF.7.2"; exon_number "6"; FPKM "118402.6643432135"; frac "0.313810"; conf_lo "81225.981370"; conf_hi "155579.347316"; cov "2.478211";
17	Cufflinks	exon	41835859	41836162	306	.	.	gene_id "CUFF.7"; transcript_id "CUFF.7.2"; exon_number "7"; FPKM "118402.6643432135"; frac "0.313810"; conf_lo "81225.981370"; conf_hi "155579.347316"; cov "2.478211";
~
~
~
~
~
(END)
[0] 0:michalo@eu-c7-067-03:~/scratch/course*Z                "vpn-global-dhcp2-145." 10:06 22-Nov-17
```

STAR

- Splice aware aligner, loading index into memory
- Results similar to tophat, but faster
- --genomeLoad LoadAndKeep
- With specific options, can produce BAM and do the counting too



- https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf

STAR statistics

michalo — michalo@eu-login-01:~ — tmux — 82×33

```
                          Started job on |       Nov 22 09:37:15
                      Started mapping on |       Nov 22 09:42:02
                             Finished on |       Nov 22 09:42:05
    Mapping speed, Million of reads per hour |       0.51

                       Number of input reads |       428
                   Average input read length |       49
                                UNIQUE READS:
              Uniquely mapped reads number |       419
                   Uniquely mapped reads % |       97.90%
                     Average mapped length |       48.74
                  Number of splices: Total |       47
       Number of splices: Annotated (sjdb) |       0
                  Number of splices: GT/AG |       47
                  Number of splices: GC/AG |       0
                  Number of splices: AT/AC |       0
          Number of splices: Non-canonical |       0
                 Mismatch rate per base, % |       0.41%
                    Deletion rate per base |       0.00%
                   Deletion average length |       0.00
                   Insertion rate per base |       0.00%
                  Insertion average length |       0.00
                       MULTI-MAPPING READS:
    Number of reads mapped to multiple loci |       9
         % of reads mapped to multiple loci |       2.10%
    Number of reads mapped to too many loci |       0
         % of reads mapped to too many loci |       0.00%
                             UNMAPPED READS:
    % of reads unmapped: too many mismatches |       0.00%
              % of reads unmapped: too short |       0.00%
                  % of reads unmapped: other |       0.00%
(END)
[0] <chalo@eu-c7-067-03:~/scratch/course*Z "vpn-global-dhcp2-145." 09:46 22-Nov-17
```

subread

- Includes subjunc similar to STAR and featureCounts
- Building index

```
subread-buildindex -o /cluster/home/michalo/work_michalo/hg38/subread_index/hg38
/cluster/home/michalo/work_michalo/hg38/Homo_sapiens.GRCh38.dna.primary_assembly.fa
```

- Alignment

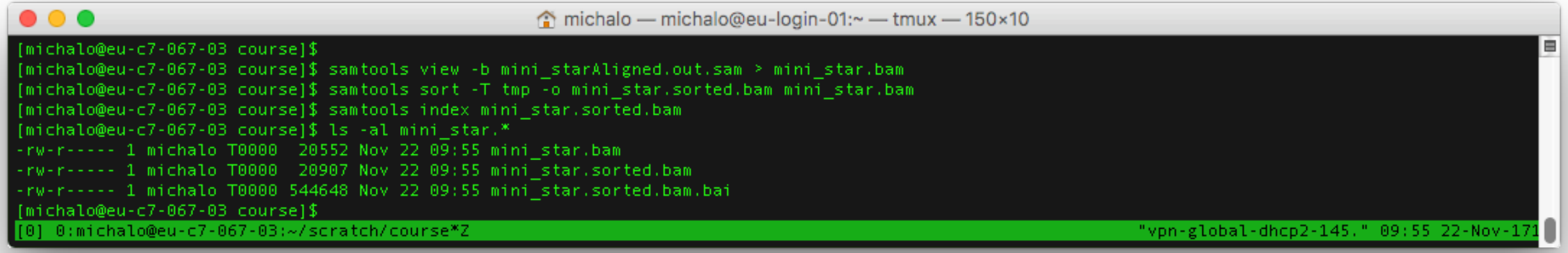
```
subread -T 24 -i /cluster/home/michalo/work_michalo/hg38/subread_index/hg38 -r mini.fastq -o
mapped_reads_subjunc/mini.bam
```

```
subjunc -T 24 -i /cluster/home/michalo/work_michalo/hg38/subread_index/hg38 -r mini.fastq -o
mapped_reads_subjunc/mini.bam
```

- http://bioinf.wehi.edu.au/subread-package/SubreadUsersGuide.pdf

samtools

- General purpose tool for conversion of BAM ⇔ SAM
- Many other operations: pileup…
- See: http://www.htslib.org/doc/samtools.html

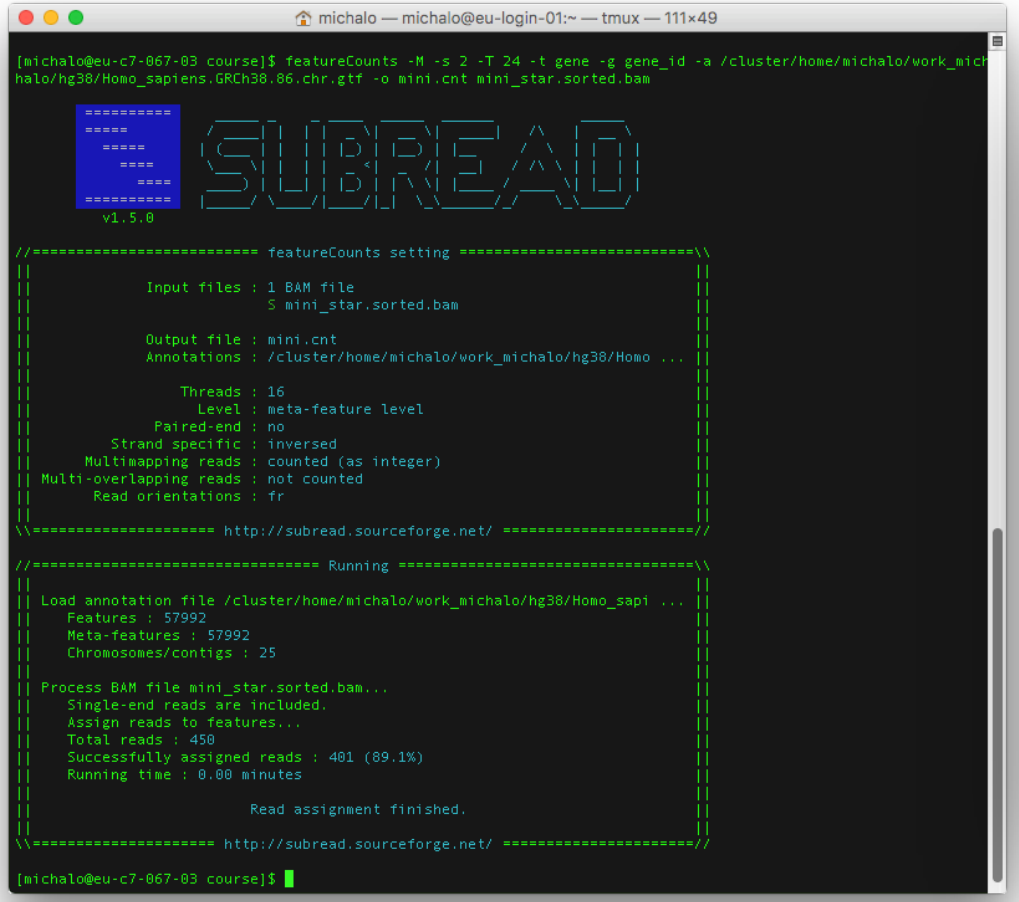


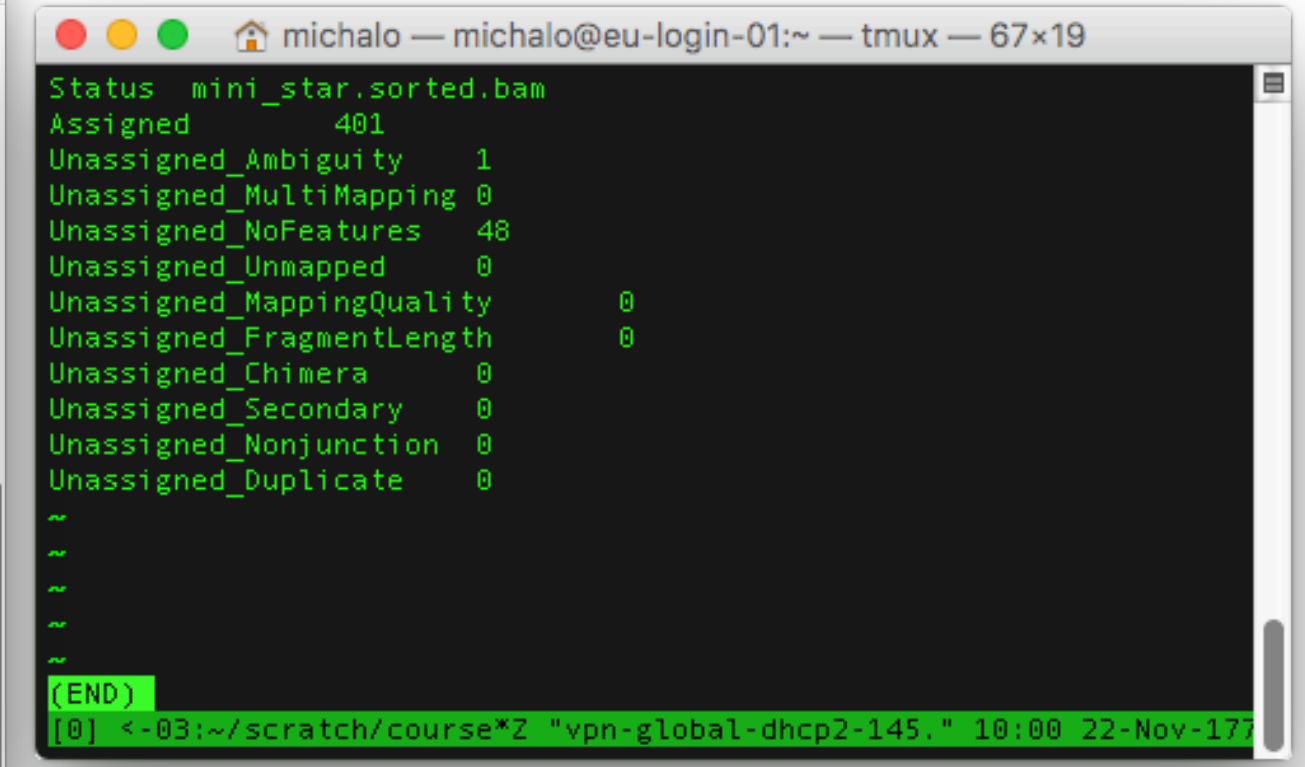
featureCounts

- Fast and flexible counting in genomic features

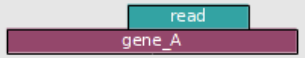
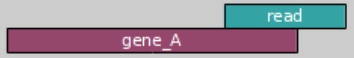
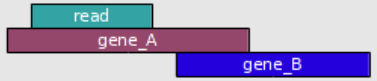
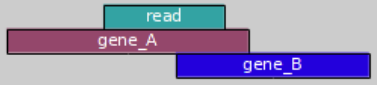
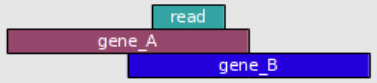
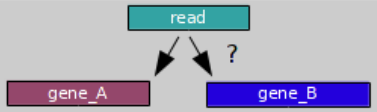
```
featureCounts -M -s 2 -T 24 -t gene -g gene_id -a
/cluster/home/michalo/work_michalo/hg38/Homo_sapiens.GRCh38.86.chr.gtf -o mini.cnt
mini_star.sorted.bam
```

featureCounts





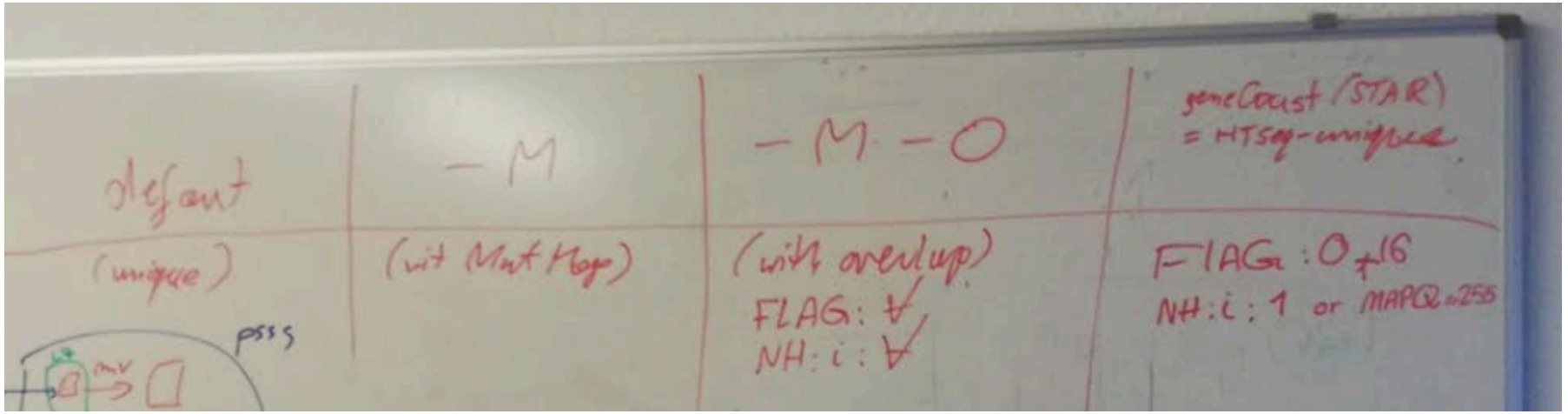
Counting reads in genes is non-trivial!

	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 (both genes with --nonunique all)	gene_A	gene_A
	ambiguous (both genes with --nonunique all)		
	alignment_not_unique (both genes with --nonunique all)		

htSeq counting modes
by Simon Anders

Tuning featureCounts

- Defaults counting features can be fuound in current STAR
- Still, for many cases a more careful counting is needed
- Gene or exon level counting
- Options



SRA tools

The command line download of public datasets from Short Read Archive

http://ncbi.github.io/sra-tools/

fastq-dump

fastq-dump --split-files ERR2811092

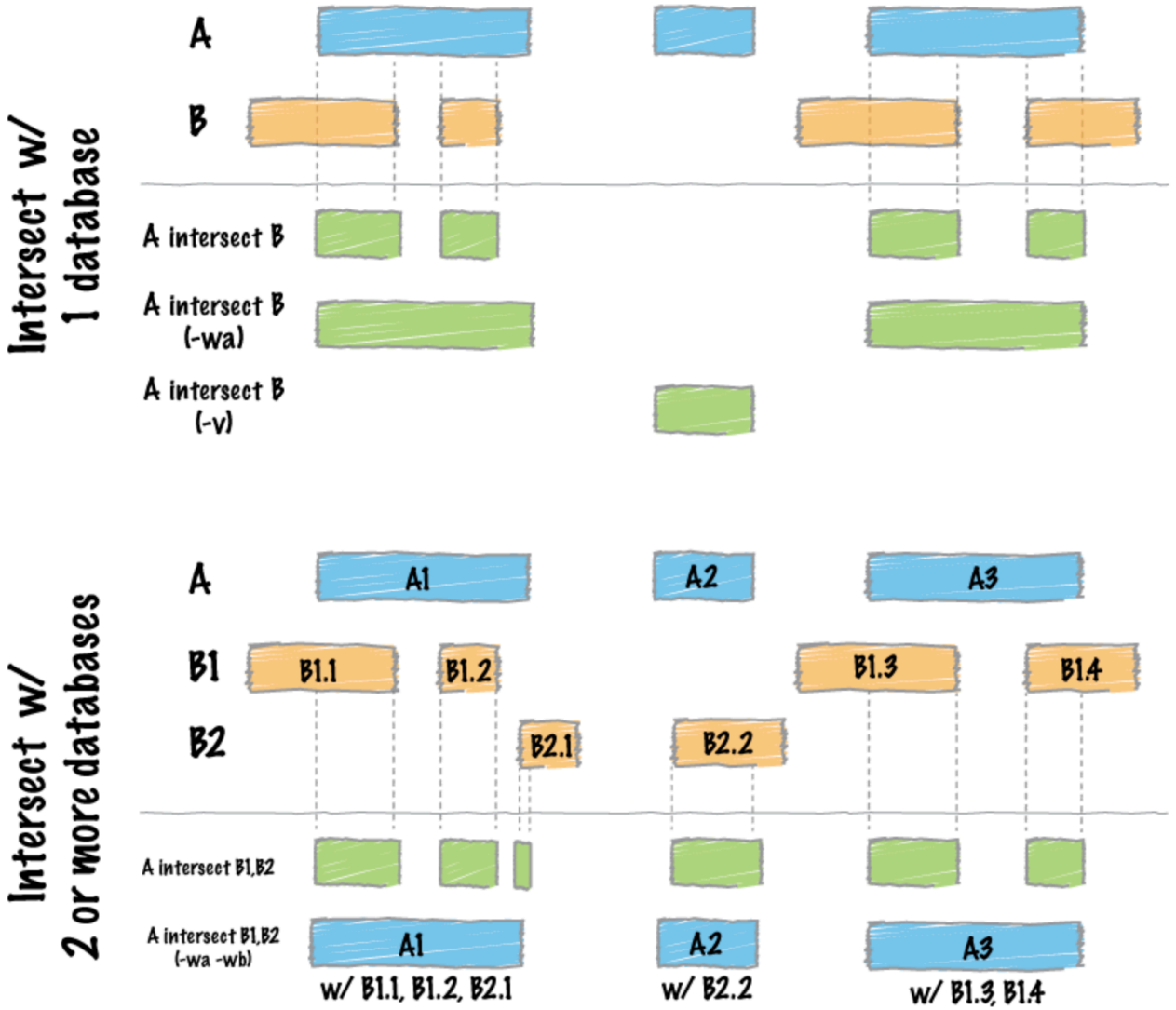
(from v2.9) fasterq-dump --split-files ERR2811092 –e 4

Bedtools

- “swiss army knife” for BED, GFF, SAM/BAM, VCF

- Conversion
- Annotation
- Getting sequences
- Genomic ranges
- ….

Bedtools - intersect



GATK

- GATK is a genomic toolbox for various operations related mainly to genomic variants calling
- Operations include producing a variant file *.vcf from an alignment file *.bam

```
module load gcc/4.8.2 gdc java/1.8.0_73
gatk/3.5java -jar GenomeAnalysisTK.jar -T UnifiedGenotyper -R
ref/human_g1k_b37_20.fasta -I bams/exp_design/NA12878_wgs_20.bam -o
sandbox/NA12878_wgs_20_UG_calls.vcf -glm BOTH -L 20:10,000,000-10,200,000
```

https://software.broadinstitute.org/gatk/documentation/tooldocs/current/

https://software.broadinstitute.org/gatk/documentation/topic?name=tutorials

http://gatkforums.broadinstitute.org/gatk/discussion/7869/howto-discover-variants-with-gatk-a-gatk-workshop-tutorial

Bioinformatics software stack on EULER

- https://scicomp.ethz.ch/wiki/GDC_software_stack
- Commands to call modules for a specific tool
- List compiled in collaboration with Genomic Diversity Center from D-USYS

Resources needed by bioinformatic tools on EULER

		cores	memory node	time/queue
Trimming	Trimmomatic, cutadapt	1		24h
Aligners - old	bowtie, bwa, SHRiMP, tophat	many: 24		4h
Aligners - RNAseq	STAR, subread	many: 24	65G	4h
Aligners - optimized	hisat2	many: 24		4h
Conversion tools	Samtools, bcftools	1		4h
Counting	featureCount	1		4h
De-novo assembly	Spades, velvet	many	256GB-1T	24h or more
Variant calling	GATK, pileup	many	256GB	24h
…				

ETH zürich

Thank you!

Using genomic software on Euler