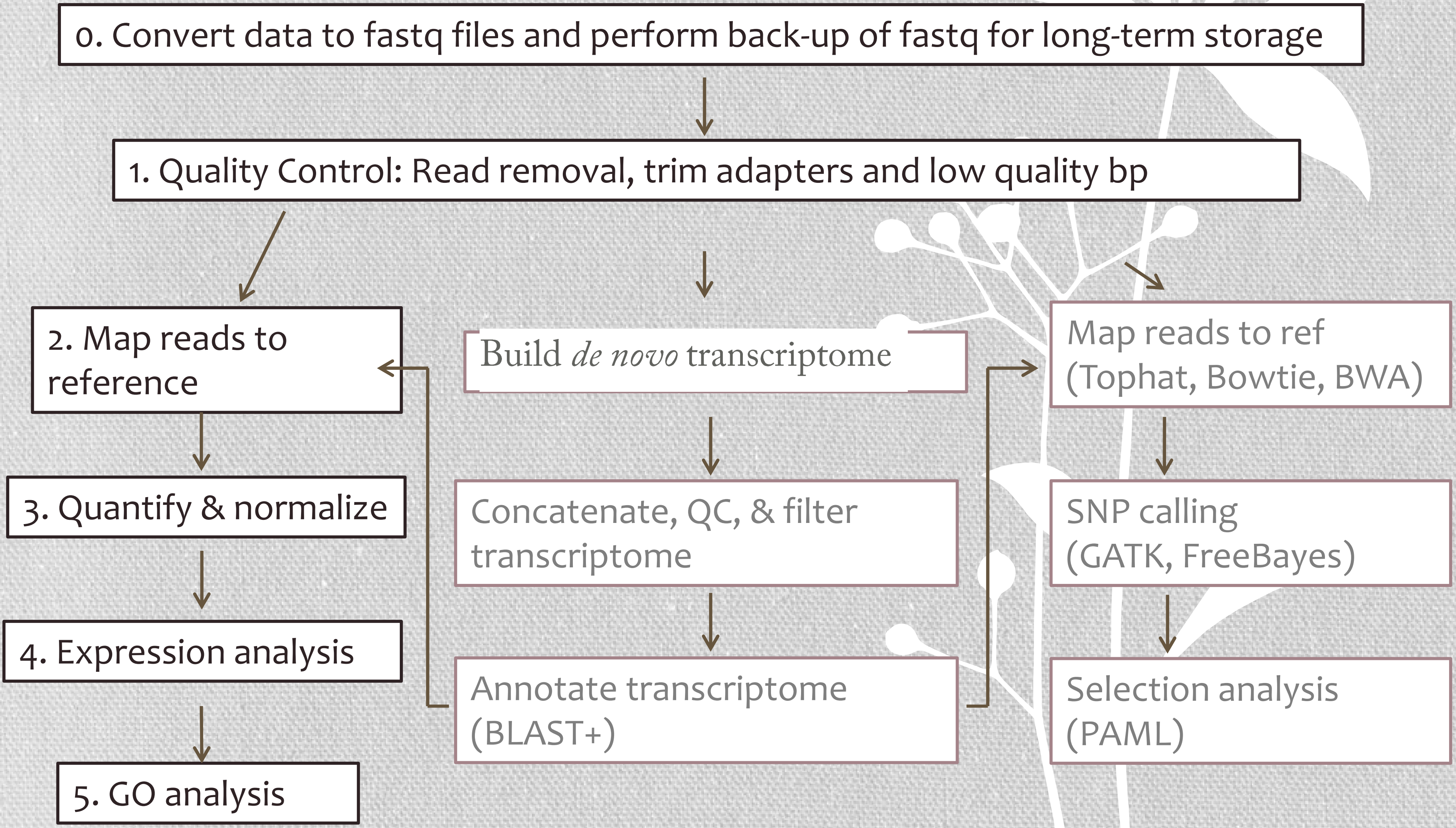
UC CONSERVATION GENOMICS CONSORTIUM

RNA-SEQ MADE EASY: A "HOW TO" MANUAL FROM RAW READS TO COUNTS

Alice Mouton & Wayne lab
Assistant Project Scientist-Postdoc fellow
EEB-UCLA, Wayne lab
Asilomar, September 2016

Bioinformatics Workflow



0. Convert data to fastq files and perform back-up of fastq for long-term storage

1. Quality Control: Read removal, trim adapters and low quality bp

2. Map reads to reference

3. Quantify & normalize

4. Expression analysis

5. GO analysis

Build *de novo* transcriptome

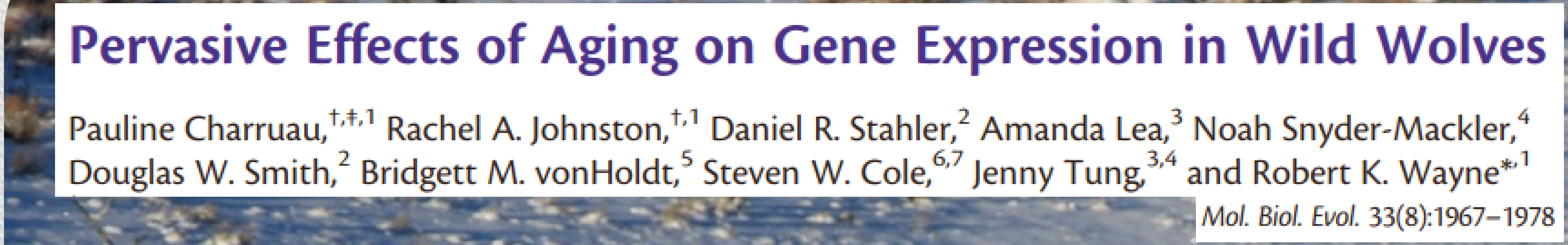
Concatenate, QC, & filter transcriptome

Annotate transcriptome (BLAST+)

Map reads to ref (Tophat, Bowtie, BWA)

SNP calling (GATK, FreeBayes)

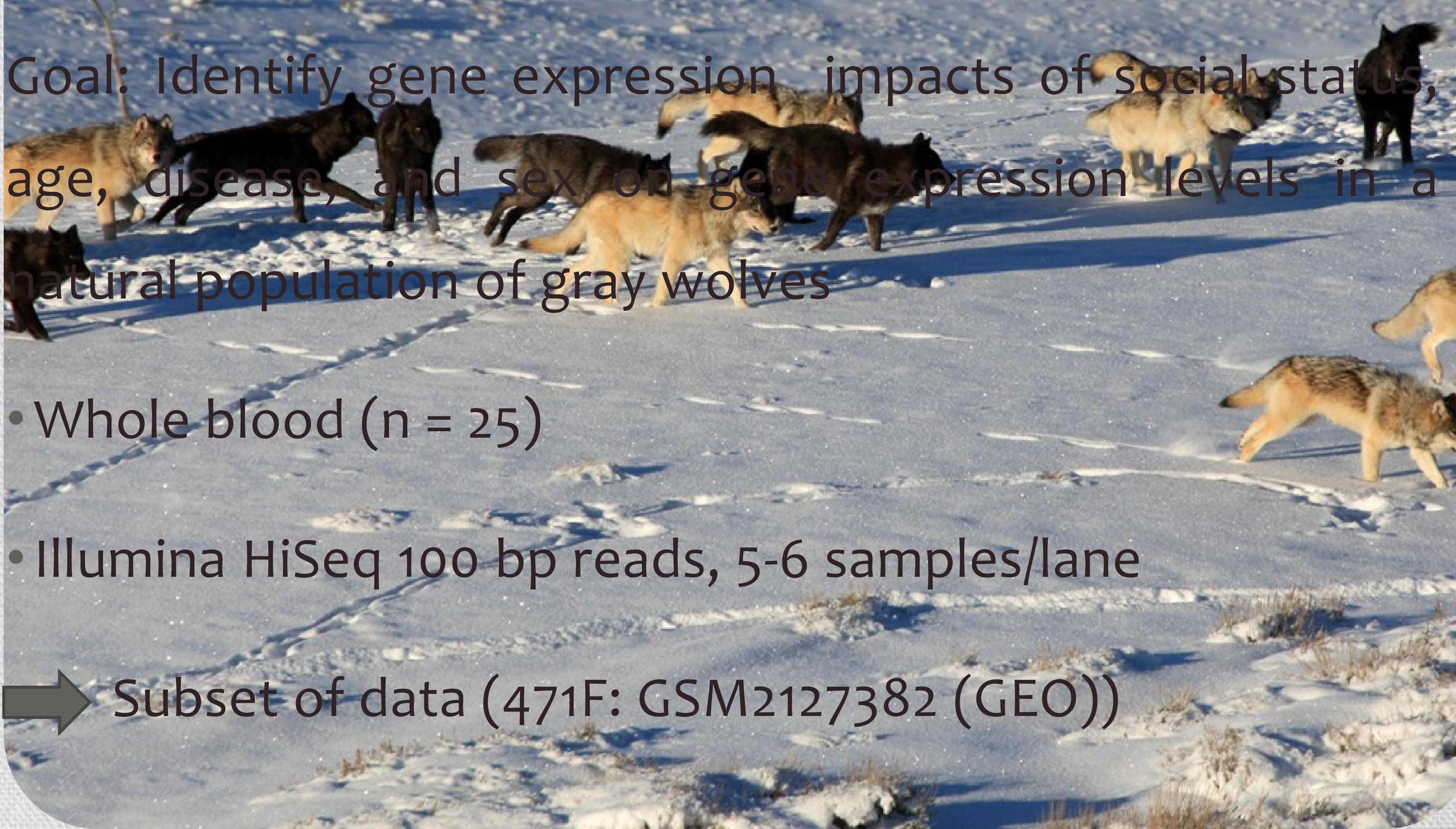
Selection analysis (PAML)



Pervasive Effects of Aging on Gene Expression in Wild Wolves

Pauline Charruau,[†,‡,1] Rachel A. Johnston,[†,1] Daniel R. Stahler,[2] Amanda Lea,[3] Noah Snyder-Mackler,[4] Douglas W. Smith,[2] Bridgett M. vonHoldt,[5] Steven W. Cole,[6,7] Jenny Tung,[3,4] and Robert K. Wayne[*,1]

Mol. Biol. Evol. 33(8):1967–1978



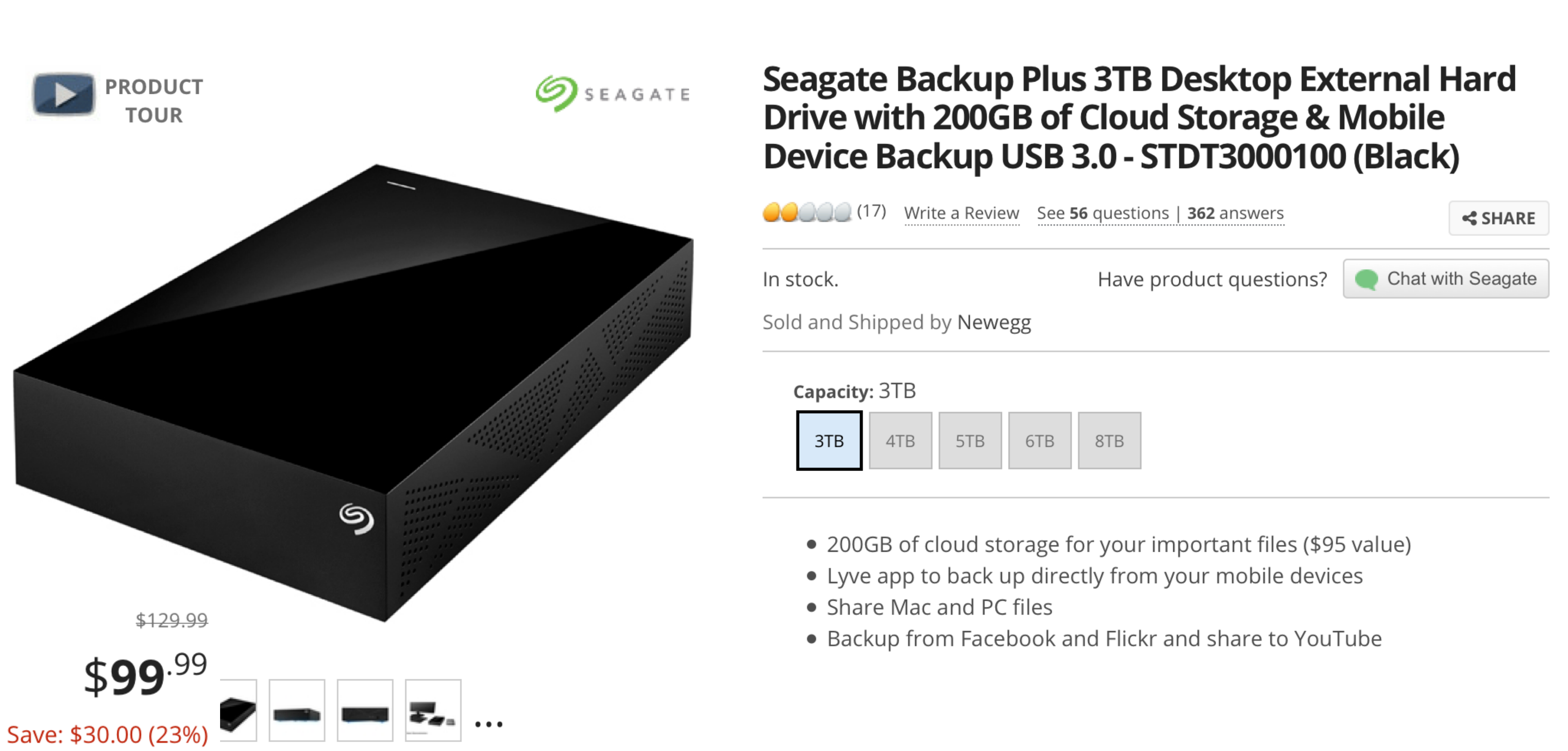
Goal: Identify gene expression impacts of social status, age, disease, and sex on gene expression levels in a natural population of gray wolves

- Whole blood (n = 25)

- Illumina HiSeq 100 bp reads, 5-6 samples/lane

➡ Subset of data (471F: GSM2127382 (GEO))

Part 0. Backup data for long-term storage



Seagate Backup Plus 3TB Desktop External Hard Drive with 200GB of Cloud Storage & Mobile Device Backup USB 3.0 - STDT3000100 (Black)

●●○○○ (17) Write a Review See **56** questions | **362** answers

In stock. Have product questions?

Sold and Shipped by Newegg

Capacity: 3TB

[✓ 3TB] 4TB 5TB 6TB 8TB

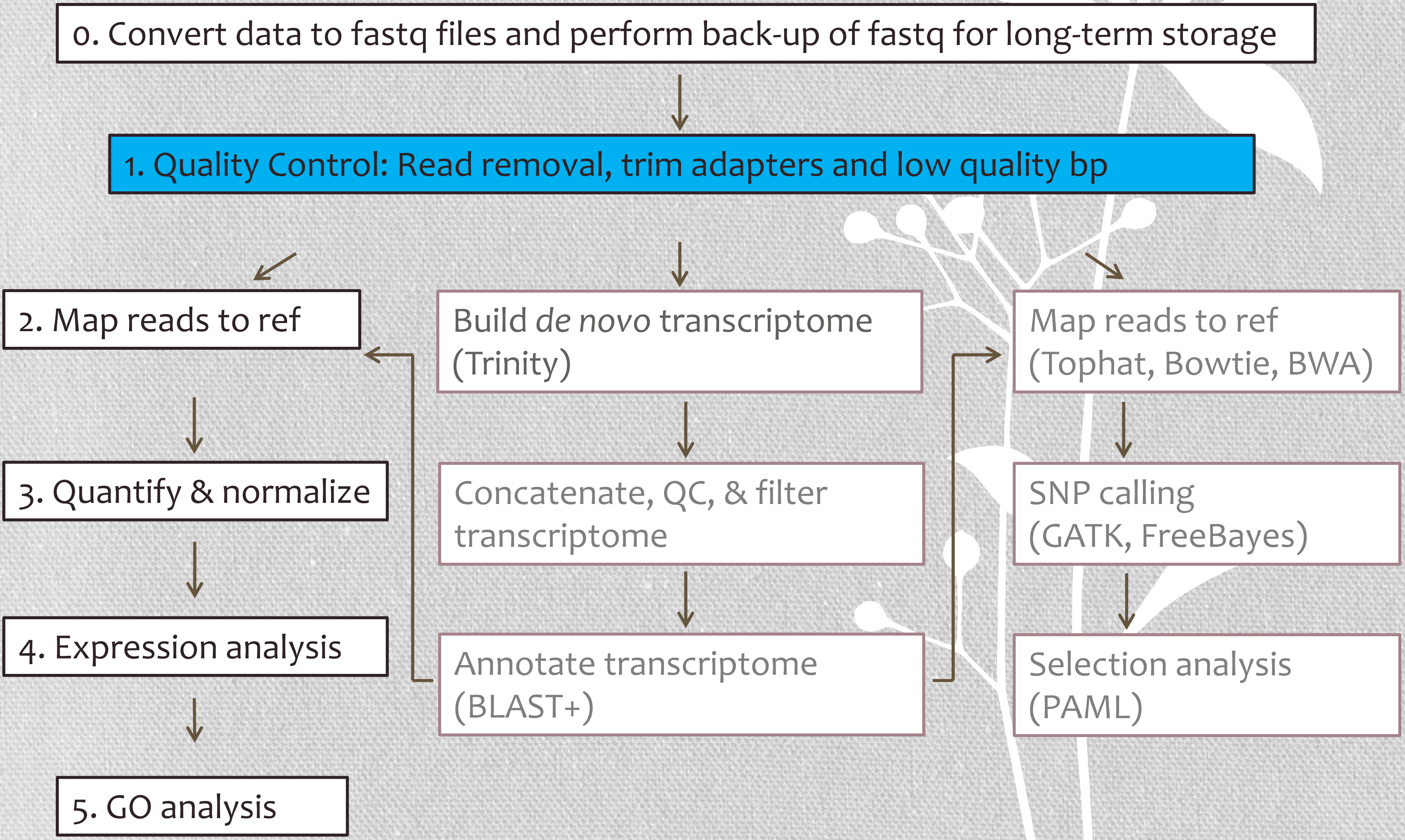
- 200GB of cloud storage for your important files ($95 value)
- Lyve app to back up directly from your mobile devices
- Share Mac and PC files
- Backup from Facebook and Flickr and share to YouTube

~~$129.99~~

$99.99

Save: $30.00 (23%)

Bioinformatics Workflow



0. Convert data to fastq files and perform back-up of fastq for long-term storage

1. Quality Control: Read removal, trim adapters and low quality bp

2. Map reads to ref

3. Quantify & normalize

4. Expression analysis

5. GO analysis

Build *de novo* transcriptome (Trinity)

Concatenate, QC, & filter transcriptome

Annotate transcriptome (BLAST+)

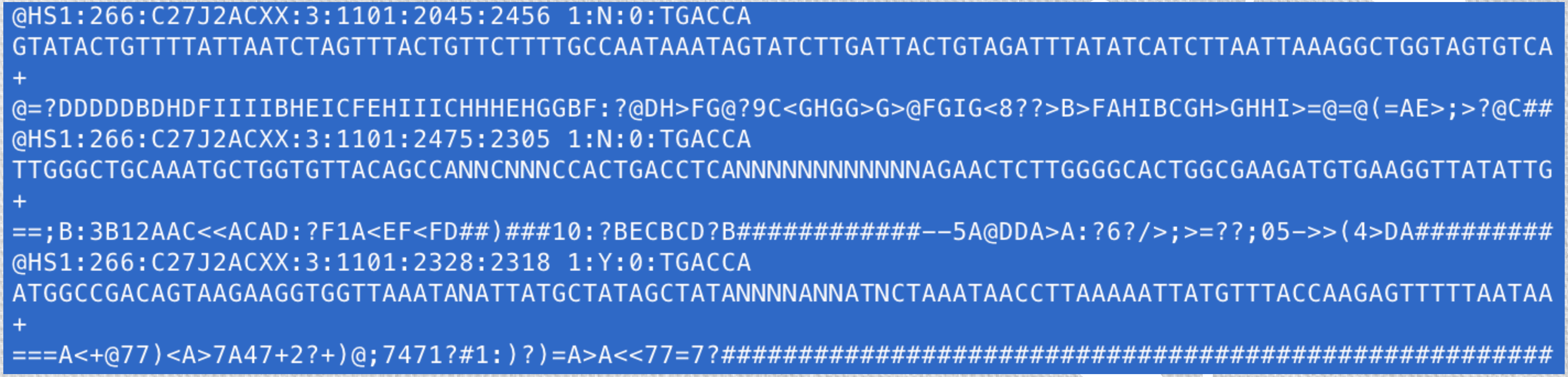
Map reads to ref (Tophat, Bowtie, BWA)

SNP calling (GATK, FreeBayes)

Selection analysis (PAML)

Part 1: Quality control

RNA-Seq fastq files looks like any other fastq

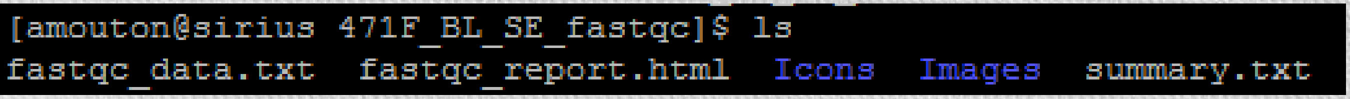


```
@HS1:266:C27J2ACXX:3:1101:2045:2456 1:N:0:TGACCA
GTATACTGTTTTATTAATCTAGTTTACTGTTCTTTTGCCAATAAATAGTATCTTGATTACTGTAGATTTATATCATCTTAATTAAAGGCTGGTAGTGTCA
+
@=?DDDDDBDHDFIIIIBHEICFEHIIICHHHEHGGBF:?@DH>FG@?9C<GHGG>G>@FGIG<8??>B>FAHIBCGH>GHHI>=@=@(=AE>;>?@C##
@HS1:266:C27J2ACXX:3:1101:2475:2305 1:N:0:TGACCA
TTGGGCTGCAAATGCTGGTGTTACAGCCANNCNNNCCACTGACCTCANNNNNNNNNNNNAGAACTCTTGGGGCACTGGCGAAGATGTGAAGGTTATATTG
+
==;B:3B12AAC<<ACAD:?F1A<EF<FD##)###10:?BECBCD?B############--5A@DDA>A:?6?/>;>=??;05->>(4>DA#########
@HS1:266:C27J2ACXX:3:1101:2328:2318 1:Y:0:TGACCA
ATGGCCGACAGTAAGAAGGTGGTTAAATANATTATGCTATAGCTATANNNNANNATNCTAAATAACCTTAAAAATTATGTTTACCAAGAGTTTTTAATAA
+
===A<+@77)<A>7A47+2?+)@;7471?#1:)?)=A>A<<77=7?######################################################
```

1. (starting with an @) is a read identifier
2. the second is the DNA sequence
3. the third another identifier (same as line 1,but starting with a +(or sometimes only consisting of a +))
4. the fourth is a Phred quality score symbol for each base in the read.

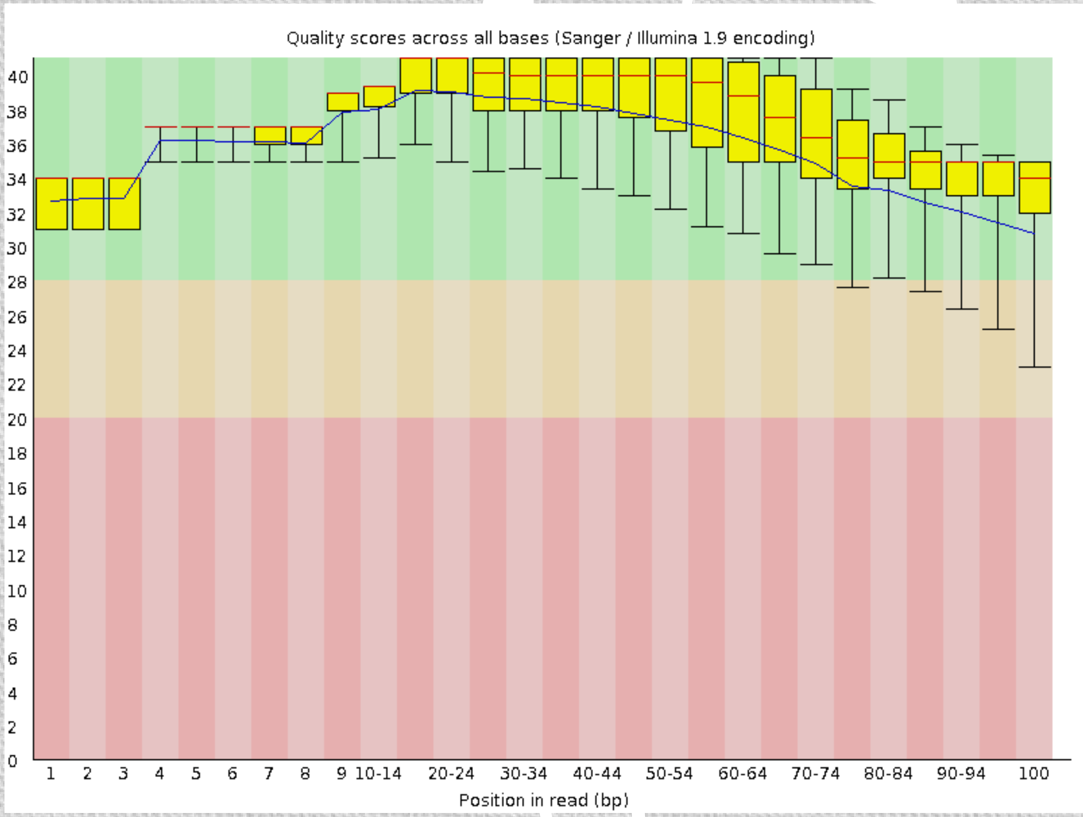
Part 1: Quality control

- Step 1.1. Look at quality of the sequence data : FASTQC



```
[amouton@sirius 471F_BL_SE_fastqc]$ ls
fastqc_data.txt  fastqc_report.html  Icons  Images  summary.txt
```

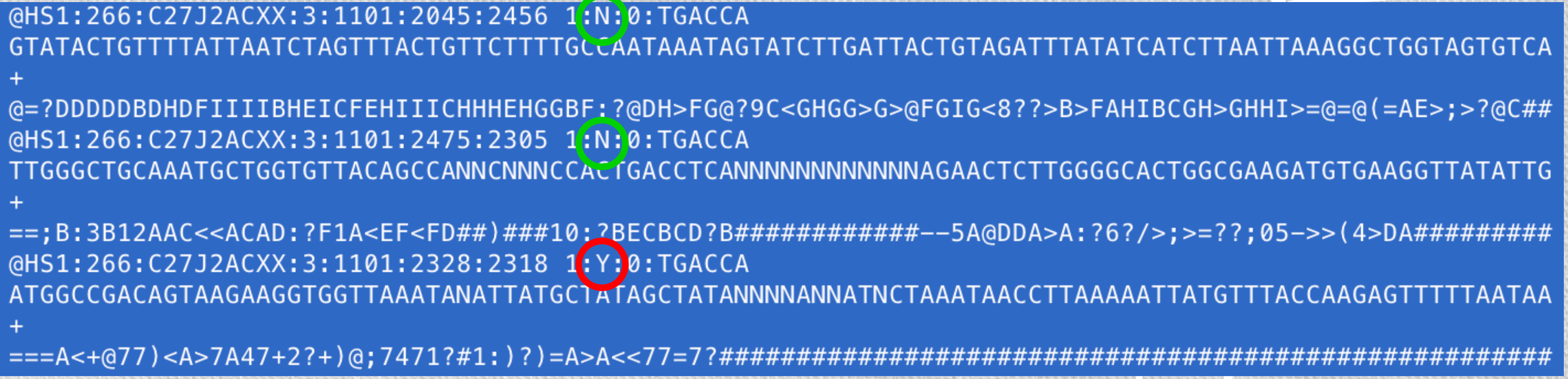
Before



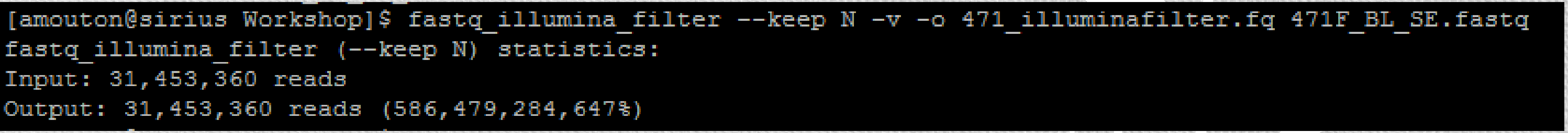
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
https://www.youtube.com/watch?v=bz93ReOv87Y

Part 1: Quality control filter provided

- Step 1.2 Remove reads that did not pass Y/N : Illumina filter



```
@HS1:266:C27J2ACXX:3:1101:2045:2456 1:N:0:TGACCA
GTATACTGTTTTATTAATCTAGTTTACTGTTCTTTTGCCAATAAATAGTATCTTGATTACTGTAGATTTATATCATCTTAATTAAAGGCTGGTAGTGTCA
+
@=?DDDDDBDHDFIIIIBHEICFEHIIICHHHEHGGBF:?@DH>FG@?9C<GHGG>G>@FGIG<8??>B>FAHIBCGH>GHHI>=@=@(=AE>;>?@C##
@HS1:266:C27J2ACXX:3:1101:2475:2305 1:N:0:TGACCA
TTGGGCTGCAAATGCTGGTGTTACAGCCANNCNNNCCACTGACCTCANNNNNNNNNNNNAGAACTCTTGGGGCACTGGCGAAGATGTGAAGGTTATATTG
+
==;B:3B12AAC<<ACAD:?F1A<EF<FD##)###10:?BECBCD?B############--5A@DDA>A:?6?/>;>=??;05->>(4>DA#########
@HS1:266:C27J2ACXX:3:1101:2328:2318 1:Y:0:TGACCA
ATGGCCGACAGTAAGAAGGTGGTTAAATANATTATGCTATAGCTATANNNNANNATNCTAAATAACCTTAAAAATTATGTTTACCAAGAGTTTTTAATAA
+
===A<+@77)<A>7A47+2?+)@;7471?#1:)?)=A>A<<77=7?######################################################
```



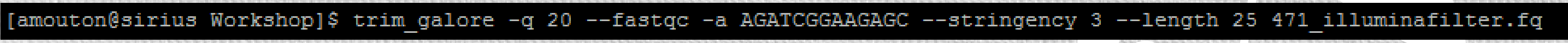
```
[amouton@sirius Workshop]$ fastq_illumina_filter --keep N -v -o 471_illuminafilter.fq 471F_BL_SE.fastq
fastq_illumina_filter (--keep N) statistics:
Input: 31,453,360 reads
Output: 31,453,360 reads (586,479,284,647%)
```

Y = Low quality reads
N = High quality reads

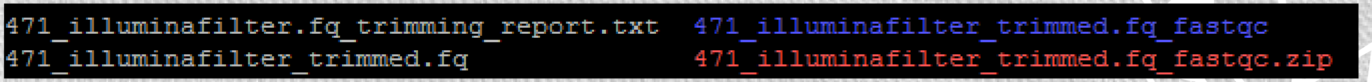
http://cancan.cshl.edu/labmembers/gordon/fastq_illumina_filter/

Part 1: Quality control

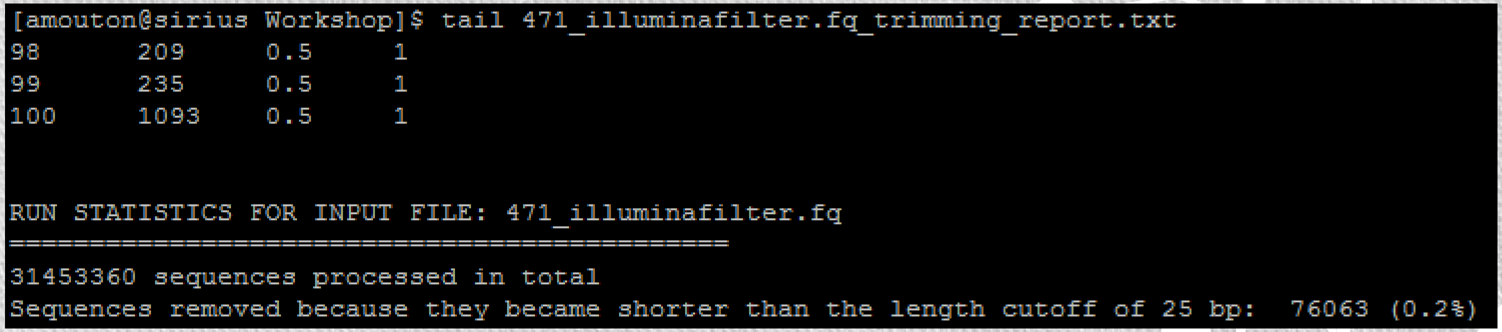
- Step 1.3: Remove the low quality base calls as well as adaptor contamination : Trim Galore



```
[amouton@sirius Workshop]$ trim_galore -q 20 --fastqc -a AGATCGGAAGAGC --stringency 3 --length 25 471_illuminafilter.fq
```



```
471_illuminafilter.fq_trimming_report.txt  471_illuminafilter_trimmed.fq_fastqc
471_illuminafilter_trimmed.fq              471_illuminafilter_trimmed.fq_fastqc.zip
```



```
[amouton@sirius Workshop]$ tail 471_illuminafilter.fq_trimming_report.txt
98      209     0.5     1
99      235     0.5     1
100     1093    0.5     1


RUN STATISTICS FOR INPUT FILE: 471_illuminafilter.fq
=============================================
31453360 sequences processed in total
Sequences removed because they became shorter than the length cutoff of 25 bp:  76063 (0.2%)
```

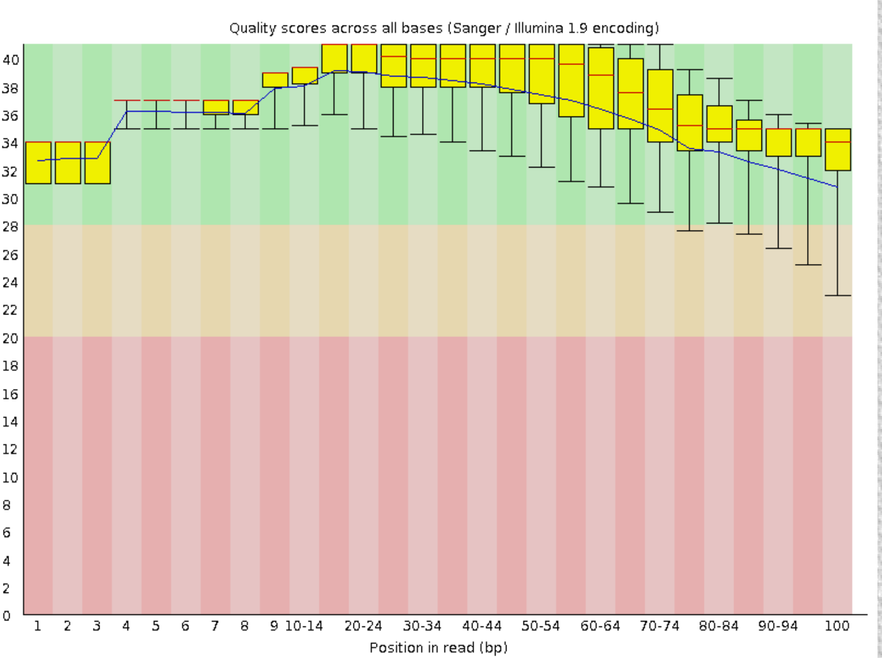
Note: A functional version of Cutadapt and optionally FastQC are required

http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/trim_galore_User_Guide_v0.3.7.pdf

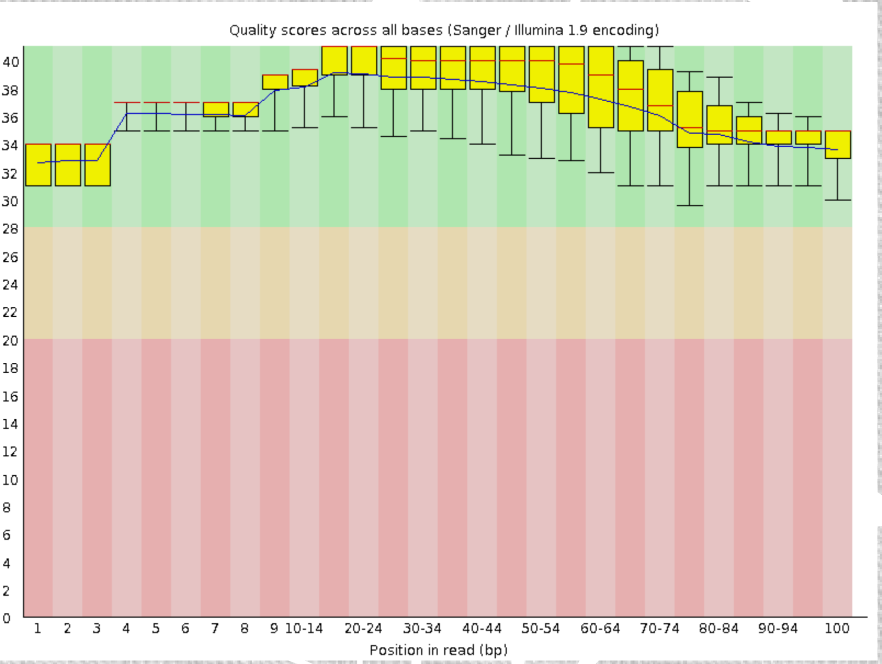
Part 1: Quality control

- Step 1.3: Remove the low quality base calls as well as adaptor contamination : Trim Galore

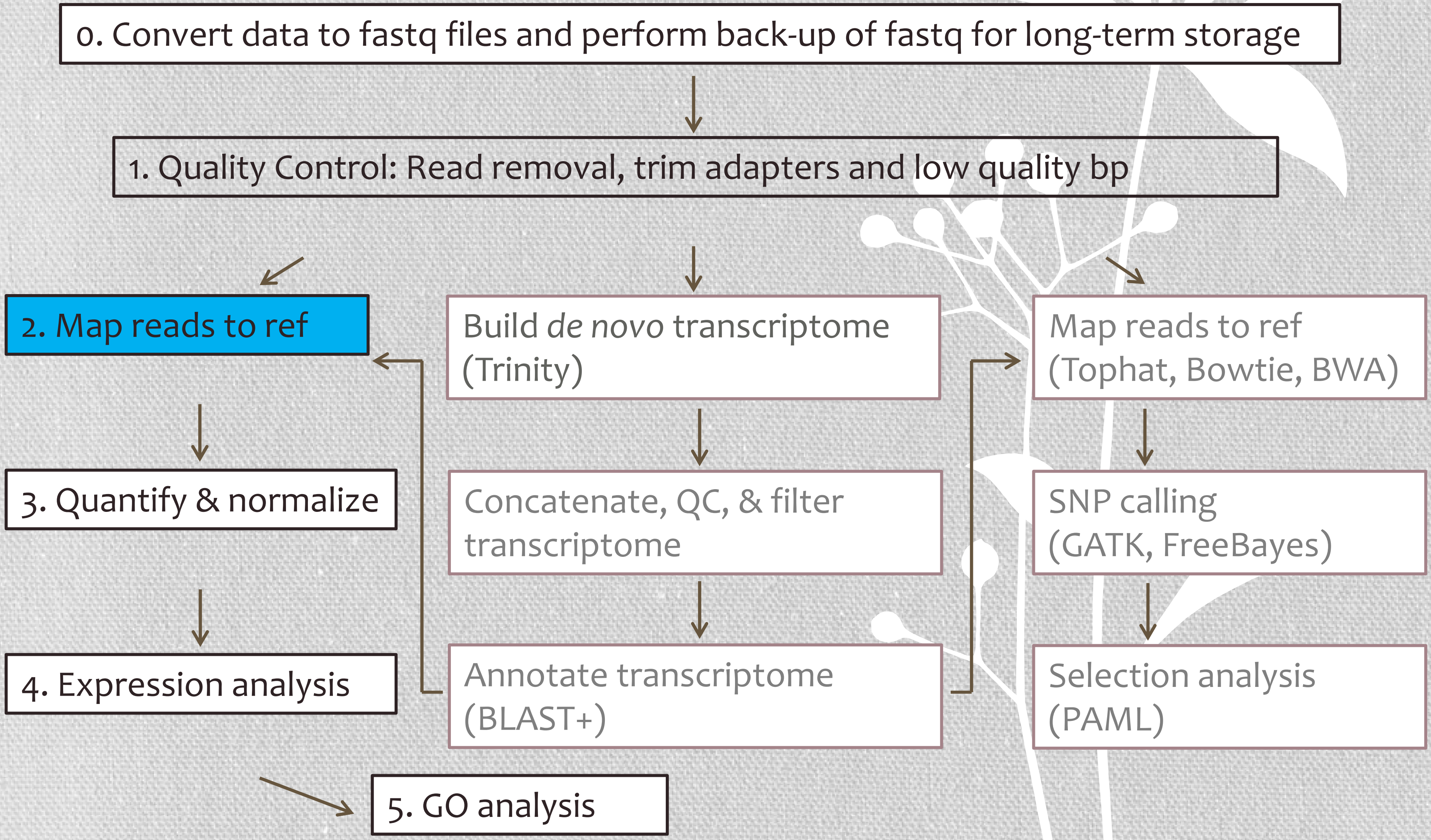
Before



After



Bioinformatics Workflow



0. Convert data to fastq files and perform back-up of fastq for long-term storage

1. Quality Control: Read removal, trim adapters and low quality bp

2. Map reads to ref

3. Quantify & normalize

4. Expression analysis

5. GO analysis

Build *de novo* transcriptome (Trinity)

Concatenate, QC, & filter transcriptome

Annotate transcriptome (BLAST+)

Map reads to ref (Tophat, Bowtie, BWA)

SNP calling (GATK, FreeBayes)

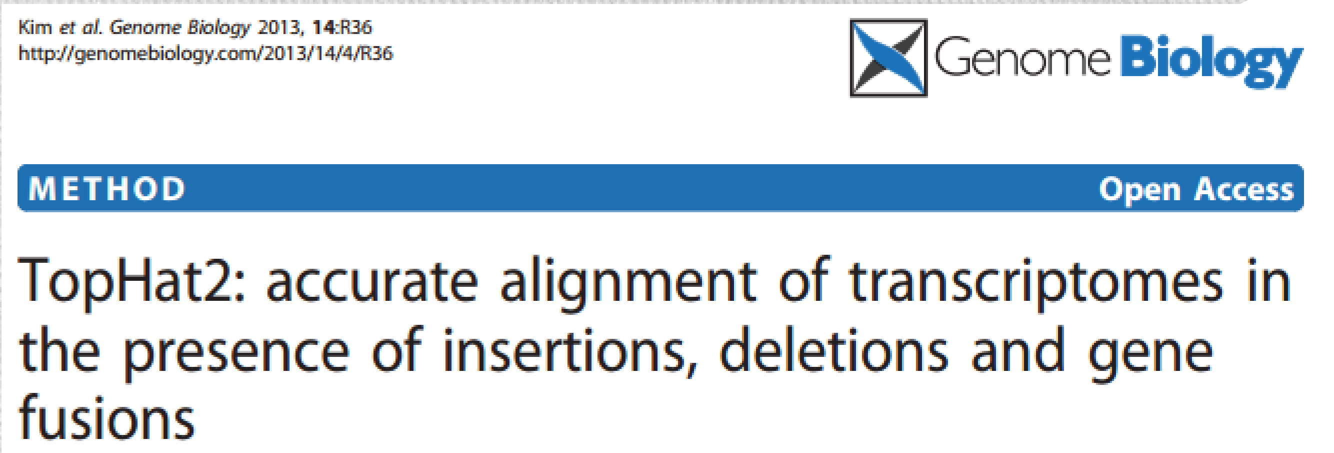
Selection analysis (PAML)

Part 2: Mapping

Available genome vs. *de novo* transcriptome

Reference	Pros/Cons	When to use
Available genome (GTF/GFF required)	• You don't have to spend weeks/months trying to assemble and annotate a transcriptome • Can use more advanced mapping programs	• When files are available for related spp
De novo transcriptome	• Transcriptome assemblies will be incomplete and have redundancy • You still rely on a reference genome for annotation	• Usually never • If no related spp reference is available (100's of million years)

Part 2: Mapping against a genome



Kim *et al. Genome Biology* 2013, **14**:R36
http://genomebiology.com/2013/14/4/R36

Genome **Biology**

METHOD **Open Access**

TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions

- Specifically made for mapping RNA-Seq reads to reference genome
- Fast splice junction mapper for RNA-Seq reads
- Needs genome (fasta) and annotation file (GTF)

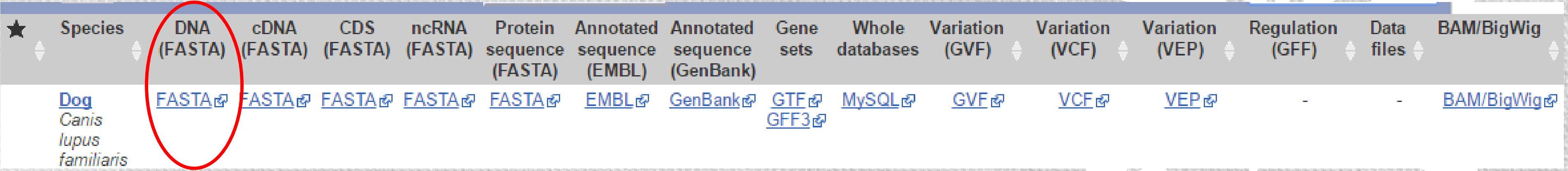
To use TopHat2, you will need the following programs in your PATH:

* bowtie2 and bowtie2-align (or bowtie)
* bowtie2-inspect (or bowtie-inspect)
* bowtie2-build (or bowtie-build)
* samtools
* Python version 2.6 or higher

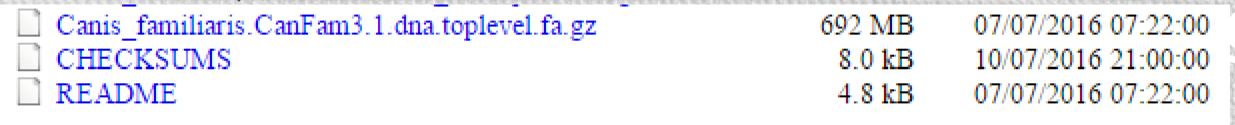
Part 2: Mapping against a genome

Step 0: download my genome and GTF (in the same new directory)

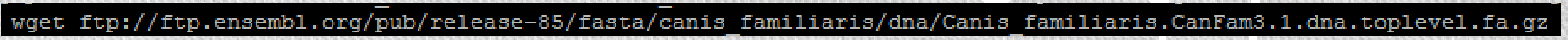
http://www.ensembl.org/info/data/ftp/index.html



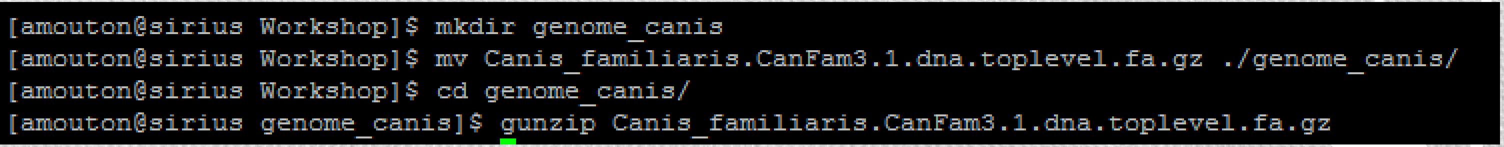
★	Species	DNA (FASTA)	cDNA (FASTA)	CDS (FASTA)	ncRNA (FASTA)	Protein sequence (FASTA)	Annotated sequence (EMBL)	Annotated sequence (GenBank)	Gene sets	Whole databases	Variation (GVF)	Variation (VCF)	Variation (VEP)	Regulation (GFF)	Data files	BAM/BigWig
	Dog *Canis lupus familiaris*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	MySQL	GVF	VCF	VEP	-	-	BAM/BigWig



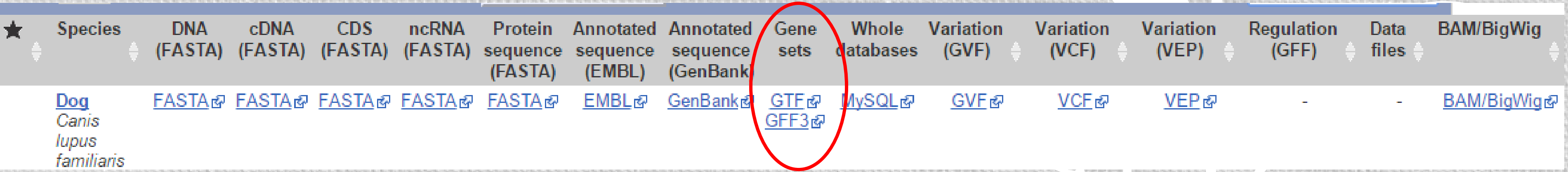
Canis_familiaris.CanFam3.1.dna.toplevel.fa.gz	692 MB	07/07/2016 07:22:00
CHECKSUMS	8.0 kB	10/07/2016 21:00:00
README	4.8 kB	07/07/2016 07:22:00



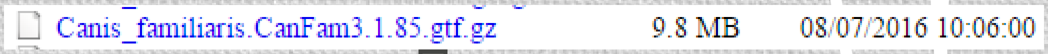
```
wget ftp://ftp.ensembl.org/pub/release-85/fasta/canis_familiaris/dna/Canis_familiaris.CanFam3.1.dna.toplevel.fa.gz
```



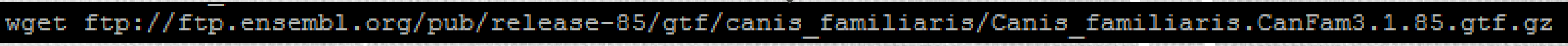
```
[amouton@sirius Workshop]$ mkdir genome_canis
[amouton@sirius Workshop]$ mv Canis_familiaris.CanFam3.1.dna.toplevel.fa.gz ./genome_canis/
[amouton@sirius Workshop]$ cd genome_canis/
[amouton@sirius genome_canis]$ gunzip Canis_familiaris.CanFam3.1.dna.toplevel.fa.gz
```



★	Species	DNA (FASTA)	cDNA (FASTA)	CDS (FASTA)	ncRNA (FASTA)	Protein sequence (FASTA)	Annotated sequence (EMBL)	Annotated sequence (GenBank)	Gene sets	Whole databases	Variation (GVF)	Variation (VCF)	Variation (VEP)	Regulation (GFF)	Data files	BAM/BigWig
	Dog *Canis lupus familiaris*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	MySQL	GVF	VCF	VEP	-	-	BAM/BigWig



Canis_familiaris.CanFam3.1.85.gtf.gz	9.8 MB	08/07/2016 10:06:00

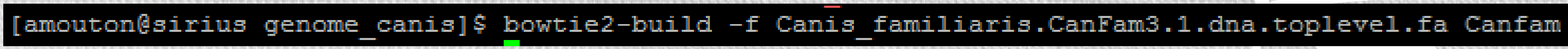


```
wget ftp://ftp.ensembl.org/pub/release-85/gtf/canis_familiaris/Canis_familiaris.CanFam3.1.85.gtf.gz
```

Part 2: Mapping against a genome

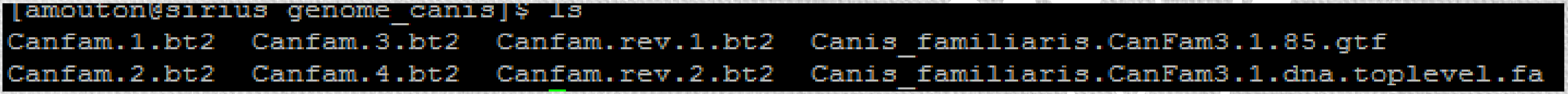
Step 1: Build indexes (take a while but you have to do it only once)

From the directory containing the genome.fa file, run the "bowtie2-build" command.



```
[amouton@sirius genome_canis]$ bowtie2-build -f Canis_familiaris.CanFam3.1.dna.toplevel.fa Canfam
```

This command will create 6 files with a *.bt2 file extension.



```
[amouton@sirius genome_canis]$ ls
Canfam.1.bt2   Canfam.3.bt2   Canfam.rev.1.bt2   Canis_familiaris.CanFam3.1.85.gtf
Canfam.2.bt2   Canfam.4.bt2   Canfam.rev.2.bt2   Canis_familiaris.CanFam3.1.dna.toplevel.fa
```

Important considerations to make when you plan to map your reads

* Find the most closely related species!!!!

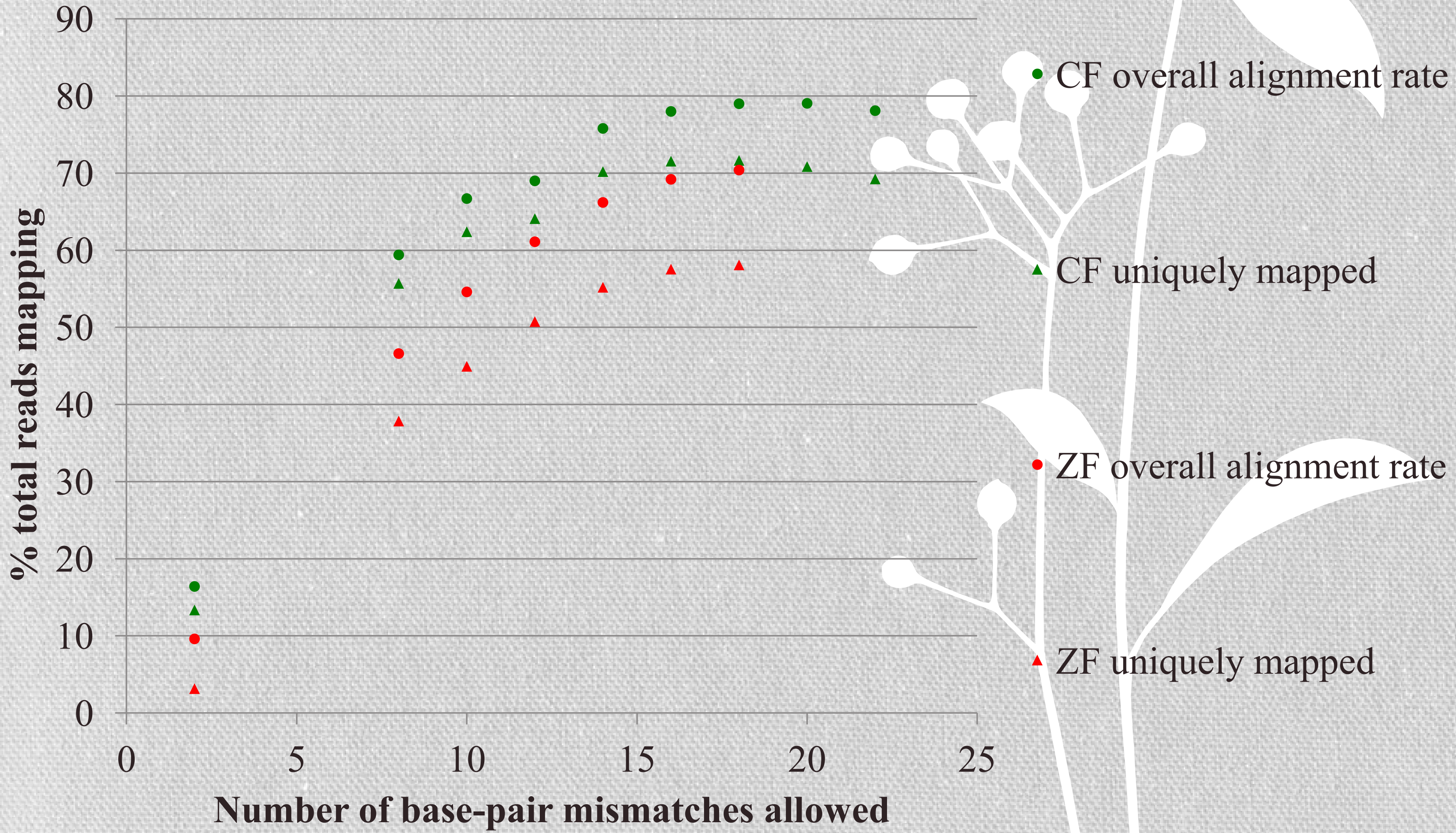
* Optimize mapping parameters for species divergence

Part 2: Mapping against a genome

Swainson’s Thrush to **Collared Flycatcher:** 25 million years

Swainson’s Thrush to **Zebra Finch:** 75 million years

“Better” genome vs more closely related genome: Closer genome wins

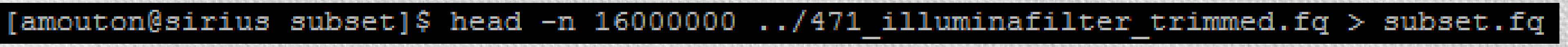


Part 2: Mapping against a genome

Step 2: Alignment with Tophat2

2.1 Work with a subset of samples

head -n 16000000 yoursamples > subset.fq # don't forget that a fastq file has 4 lines
(for instance 4 000 000 reads to play with)



```
[amouton@sirius subset]$ head -n 16000000 ../471_illuminafilter_trimmed.fq > subset.fq
```

2.2 Optimize your parameters (tophat2 –h) !!

=> the high number of reads with high % of unique reads

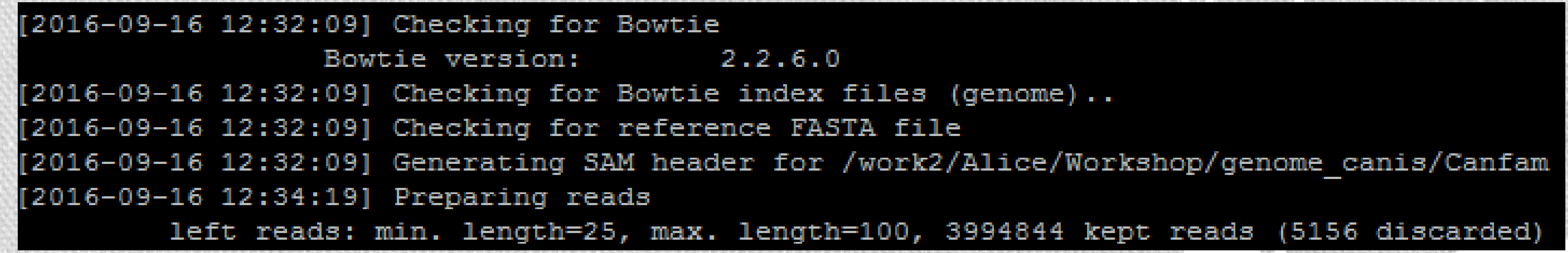
You can play with several options such as
* --read-mismatches
* --read-gap-length
* --read-edit-dist

Part 2: Mapping against a genome

2.2 Optimize your parameters (tophat2 –h) : exemple

tophat2 -p 2 --output-dir ./test1 --library-type fr-secondstrand --b2-very-sensitive -N9 --read-edit-dist 22 --read-gap-length 3 /work2/Alice/Workshop/genome_canis/Canfam subset.fq

tophat2 -p 2 --output-dir ./test2 --library-type fr-secondstrand --b2-very-sensitive -N3 --read-edit-dist 3 --read-gap-length 3 /work2/Alice/Workshop/genome_canis/Canfam subset.fq



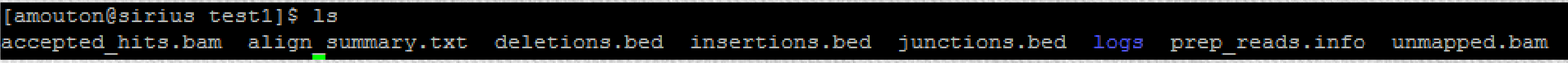
```
[2016-09-16 12:32:09] Checking for Bowtie
                  Bowtie version:        2.2.6.0
[2016-09-16 12:32:09] Checking for Bowtie index files (genome)..
[2016-09-16 12:32:09] Checking for reference FASTA file
[2016-09-16 12:32:09] Generating SAM header for /work2/Alice/Workshop/genome_canis/Canfam
[2016-09-16 12:34:19] Preparing reads
         left reads: min. length=25, max. length=100, 3994844 kept reads (5156 discarded)
```

low complexity of reads and number of N (poly-A and poly-T and so one..)

Work in parallel to save time!!

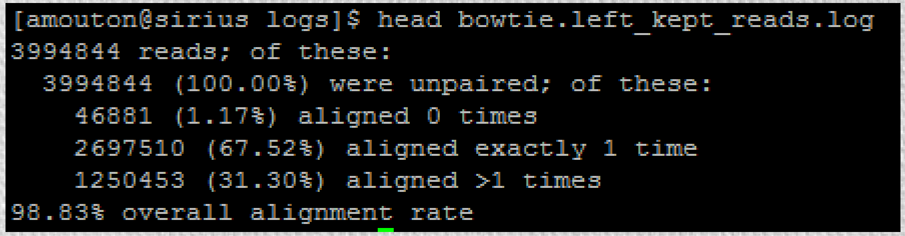
Part 2: Mapping against a genome

2.2 Optimize your parameters (tophat2 –h) : exemple



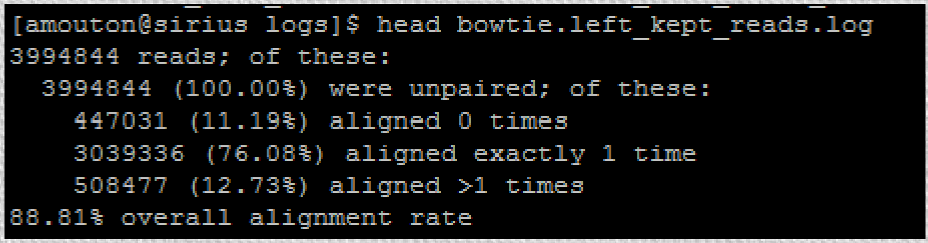
```
[amouton@sirius test1]$ ls
accepted_hits.bam  align_summary.txt  deletions.bed  insertions.bed  junctions.bed  logs  prep_reads.info  unmapped.bam
```

Test 1



```
[amouton@sirius logs]$ head bowtie.left_kept_reads.log
3994844 reads; of these:
  3994844 (100.00%) were unpaired; of these:
    46881 (1.17%) aligned 0 times
    2697510 (67.52%) aligned exactly 1 time
    1250453 (31.30%) aligned >1 times
98.83% overall alignment rate
```

Test 2



```
[amouton@sirius logs]$ head bowtie.left_kept_reads.log
3994844 reads; of these:
  3994844 (100.00%) were unpaired; of these:
    447031 (11.19%) aligned 0 times
    3039336 (76.08%) aligned exactly 1 time
    508477 (12.73%) aligned >1 times
88.81% overall alignment rate
```

	N (mismatches)	% overall alignment rate	% uniq mapped
test1	9	98.83	67.52
test2	3	88.81	76.08

Mapping of all your samples with the parameters of your choice

The file that we’re interested in for now is accepted_hits.bam, which is the reads that were mapped successfully.

https://samtools.github.io/hts-specs/SAMv1.pdf

Part 2: Mapping against a genome

2.3 Quality of the mapping (on sorted bam)

* IGV (Resources:https://www.broadinstitute.org/igv/) => alignment (SAM or BAM) has to be sorted and indexed by coordinates (sorts by chromosome and start position not by read ID)

```
samtools sort accepted_hits.bam accepted_hits_sorted
samtools index accepted_hits_sorted.bam #generate a .bam.bai that can be used for the IGV view
```

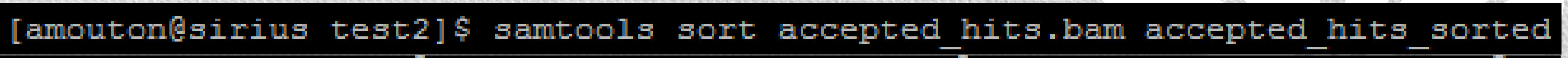
* Qualimap (http://qualimap.bioinfo.cipf.es/)

```
qualimap bamqc -bam accepted_hits_sorted.bam #(html file)
```

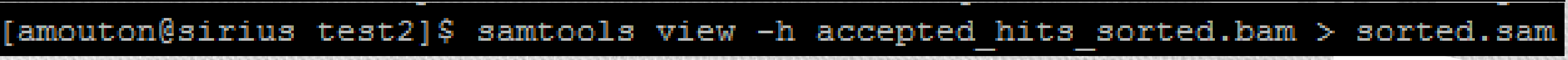
* 'samtools flagstat' to get a basic summary of an alignment

```
samtools flagstat accepted_hits_sorted.bam
```

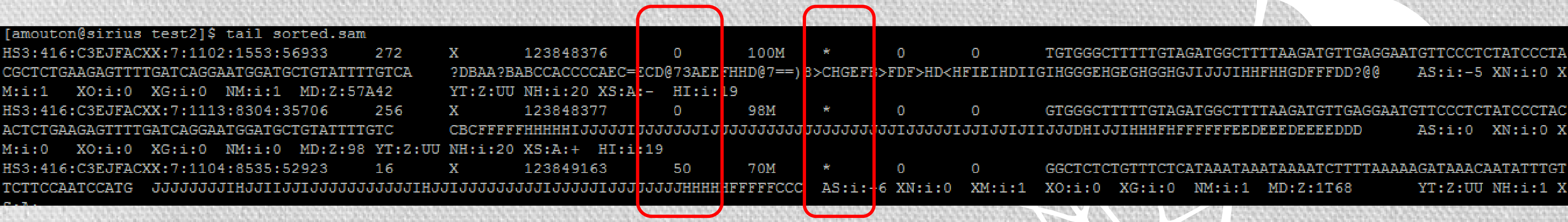
Part 3: SORT and keep UNIQ reads!



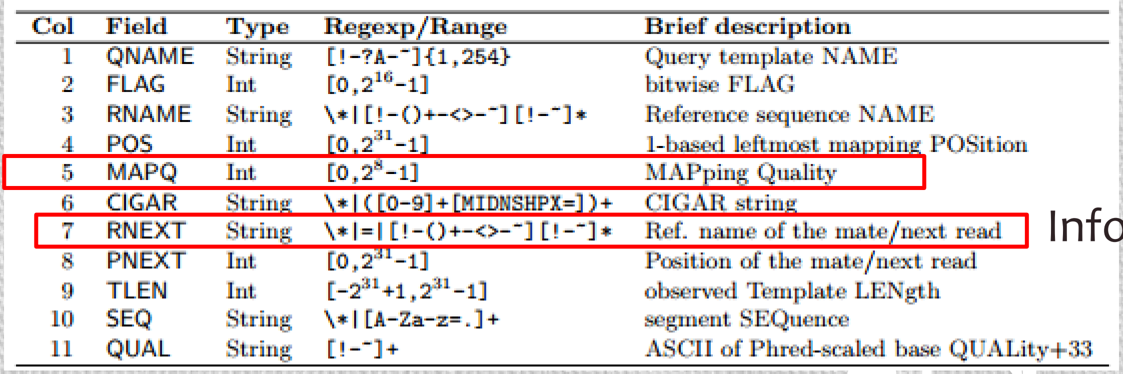
```
[amouton@sirius test2]$ samtools sort accepted_hits.bam accepted_hits_sorted
```



```
[amouton@sirius test2]$ samtools view -h accepted_hits_sorted.bam > sorted.sam
```

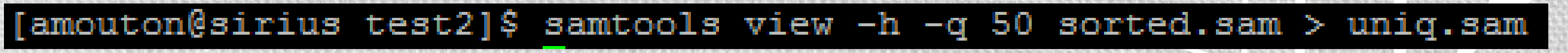


```
[amouton@sirius test2]$ tail sorted.sam
HS3:416:C3EJFACXX:7:1102:1553:56933     272     X       123848376       0       100M    *       0       0       TGTGGGCTTTTTGTAGATGGCTTTTAAGATGTTGAGGAATGTTCCCTCTATCCCTA
CGCTCTGAAGAGTTTTGATCAGGAATGGATGCTGTATTTTGTCA    ?DBAA?BABCCACCCCAEC=ECD@73AEEFHHD@7==)8>CHGEFF>FDF>HD<HFIEIHDIIGIHGGGEHGEGHGGHGJIJJJIHHFHHGDFFFDD?@@    AS:i:-5 XN:i:0 X
M:i:1   XO:i:0  XG:i:0  NM:i:1  MD:Z:57A42      YT:Z:UU NH:i:20 XS:A:-  HI:i:19
HS3:416:C3EJFACXX:7:1113:8304:35706     256     X       123848377       0       98M     *       0       0       GTGGGCTTTTTGTAGATGGCTTTTAAGATGTTGAGGAATGTTCCCTCTATCCCTAC
ACTCTGAAGAGTTTTGATCAGGAATGGATGCTGTATTTTGTC      CBCFFFFFHHHHHIJJJJJIJJJJJJJIJJJJJJJJJJJJJJJJJJIJJJJJIJJIJJIJIIJJJDHIJJIHHHFHFFFFFFFEEDEEEDEEEEDDD      AS:i:0  XN:i:0 X
M:i:0   XO:i:0  XG:i:0  NM:i:0  MD:Z:98 YT:Z:UU NH:i:20 XS:A:+  HI:i:19
HS3:416:C3EJFACXX:7:1104:8535:52923     16      X       123849163       50      70M     *       0       0       GGCTCTCTGTTTCTCATAAATAAATAAAATCTTTTAAAAAGATAAACAATATTTGT
TCTTCCAATCCATG  JJJJJJJJIHJJIIJJIJJJJJJJJJJJIHJJIJJJJJJJJJIJJJJJIJJJJJJJHHHHHFFFFFCCC  AS:i:-6 XN:i:0  XM:i:1  XO:i:0  XG:i:0  NM:i:1  MD:Z:1T68       YT:Z:UU NH:i:1 X
```



Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,254}	Query template NAME
2	FLAG	Int	[0,2^{16}-1]	bitwise FLAG
3	RNAME	String	*|[!-()+-<>-~][!-~]*	Reference sequence NAME
4	POS	Int	[0,2^{31}-1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,2^{8}-1]	MAPping Quality
6	CIGAR	String	*|([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	*|=|[!-()+-<>-~][!-~]*	Ref. name of the mate/next read
8	PNEXT	Int	[0,2^{31}-1]	Position of the mate/next read
9	TLEN	Int	[-2^{31}+1,2^{31}-1]	observed Template LENgth
10	SEQ	String	*|[A-Za-z=.]+	segment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

Info for mate reads

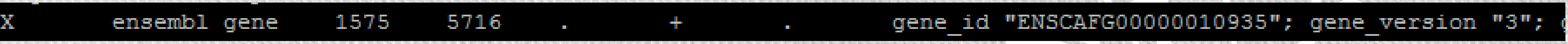


```
[amouton@sirius test2]$ samtools view -h -q 50 sorted.sam > uniq.sam
```

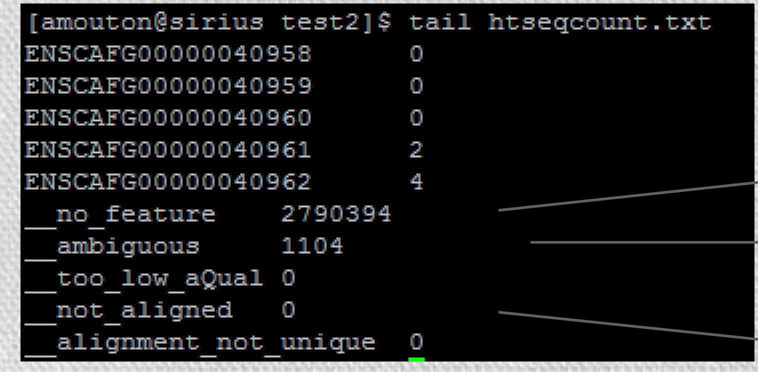
Part 4: Count the reads

samtools view -bS uniq.sam > uniq.bam # you want to convert into a bam again to gain space

htseq-count -f bam -r pos -s yes -i gene_id -m union -q uniq.bam
/work2/Alice/Workshop/genome_canis/ Canis_familiaris.CanFam3.1.85.gtf> htseqcount.txt



```
X       ensembl gene    1575    5716    .       +       .       gene_id "ENSCAFG00000010935"; gene_version "3";
```



```
[amouton@sirius test2]$ tail htseqcount.txt
ENSCAFG00000040958      0
ENSCAFG00000040959      0
ENSCAFG00000040960      0
ENSCAFG00000040961      2
ENSCAFG00000040962      4
__no_feature    2790394
__ambiguous     1104
__too_low_aQual 0
__not_aligned   0
__alignment_not_unique  0
```

reads (or read pairs) which could not be assigned to any feature

reads (or read pairs) which could have been assigned to more than one feature and hence were not counted for any of these

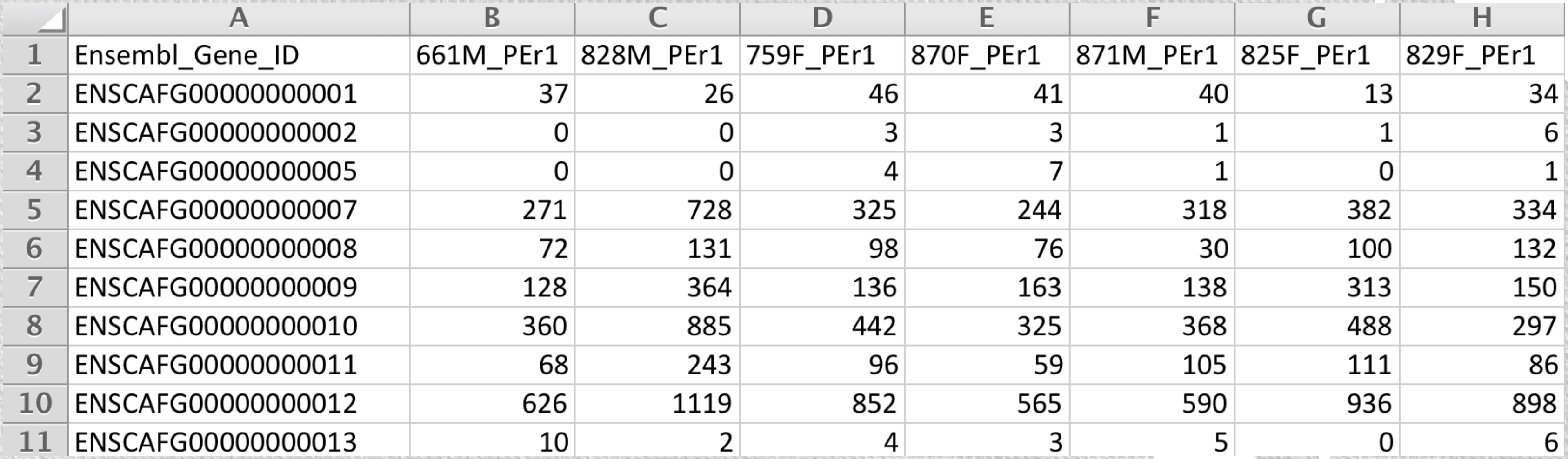
reads (or read pairs) in the SAM file without alignment

Note: Check your kit to know if you have to use stranded or no!*

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

Part 4: Count the reads

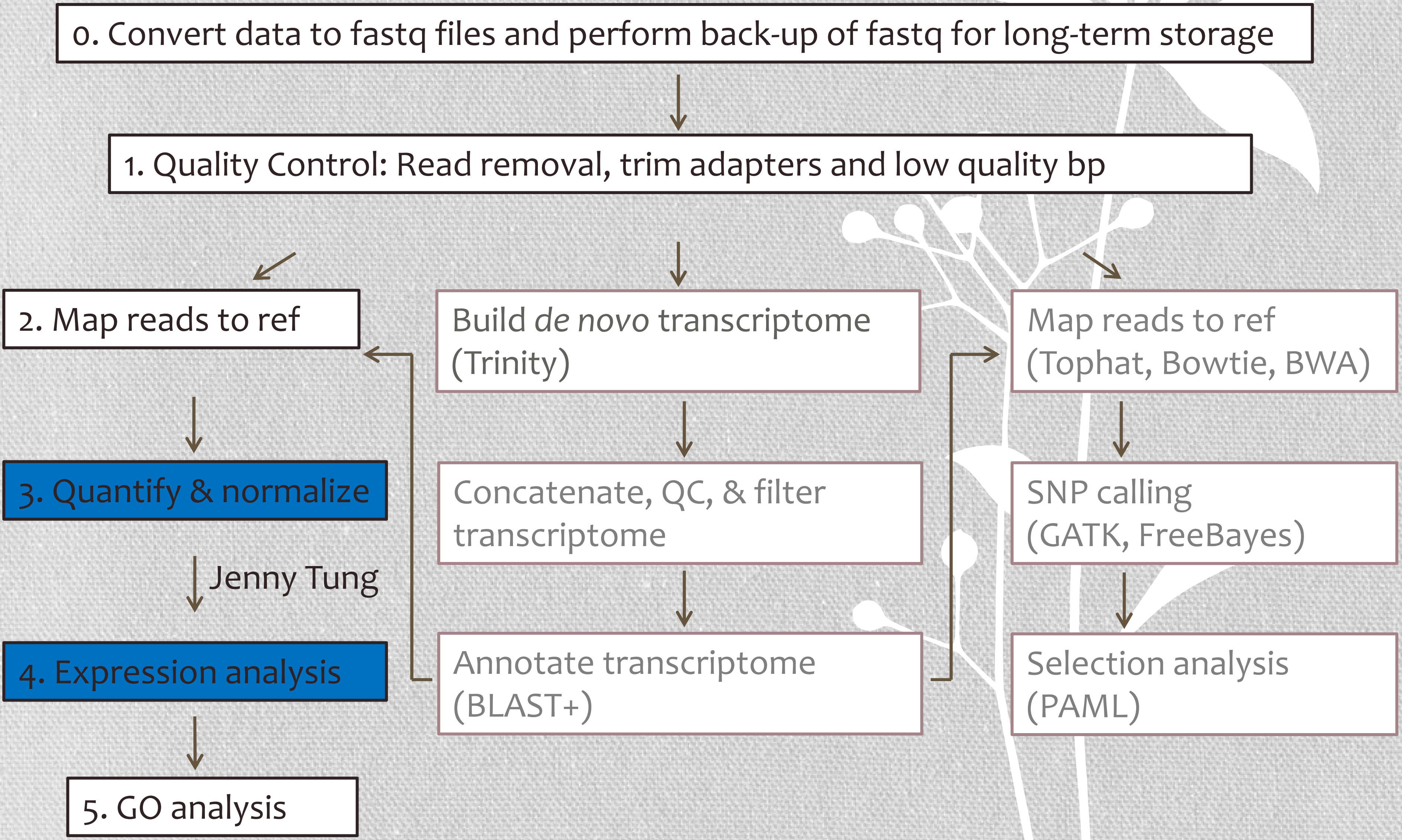
# copy the htseq counts in the same folder and copy on your computer



	A	B	C	D	E	F	G	H
1	Ensembl_Gene_ID	661M_PEr1	828M_PEr1	759F_PEr1	870F_PEr1	871M_PEr1	825F_PEr1	829F_PEr1
2	ENSCAFG00000000001	37	26	46	41	40	13	34
3	ENSCAFG00000000002	0	0	3	3	1	1	6
4	ENSCAFG00000000005	0	0	4	7	1	0	1
5	ENSCAFG00000000007	271	728	325	244	318	382	334
6	ENSCAFG00000000008	72	131	98	76	30	100	132
7	ENSCAFG00000000009	128	364	136	163	138	313	150
8	ENSCAFG00000000010	360	885	442	325	368	488	297
9	ENSCAFG00000000011	68	243	96	59	105	111	86
10	ENSCAFG00000000012	626	1119	852	565	590	936	898
11	ENSCAFG00000000013	10	2	4	3	5	0	6

Now you are ready to analyze your data..

Bioinformatics Workflow



0. Convert data to fastq files and perform back-up of fastq for long-term storage

1. Quality Control: Read removal, trim adapters and low quality bp

2. Map reads to ref

3. Quantify & normalize

Jenny Tung

4. Expression analysis

5. GO analysis

Build *de novo* transcriptome (Trinity)

Concatenate, QC, & filter transcriptome

Annotate transcriptome (BLAST+)

Map reads to ref (Tophat, Bowtie, BWA)

SNP calling (GATK, FreeBayes)

Selection analysis (PAML)