Brief Guide for NGS Transcriptomics: From gene expression to genetics.

by
Aureliano Bombarely
ab782@cornell.edu

Lectures:

1. **Basics of the Next Generation Sequencing (NGS).**
   1.1. The sequencing revolutions.
   1.2. Strengths and weaknesses of the different technologies.
   1.3. Inputs and outputs.
2. **RNAseq experiment design.**
   2.1. Reference vs Non-reference.
   2.2. High heterozygosity and polyploid polyploid problem.
   2.3. Tissue selection and treatments.
   2.4. Sequencing technology.
3. **RNAseq expression analysis.**
   3.1. Reference preparation and read mapping.
   3.2. Gene expression.
   3.3. Analysis and visualization.
4. **Use of RNAseq reads for phylogeny and genetics.**
   4.1. Recovering full length mRNA: Reference guided assembly.
   4.2. Phylogeny though RNAseq: From gene tree to species tree.
   4.3. From reads to markers: SNP calling.
   4.4. Population genetics and NGS.

Exercises:

1. **Basic Linux commands.**
2. **Sequencing evaluation.**
3. **Simple read mapping.**
4. **Simple de-novo assembly.**
5. **Basic R commands**
6. **Functional annotation.**
7. **Differential gene expression.**
8. **Cluster analysis for gene expression.**
9. **Selecting genes for phylogeny.**
10. **SNP calling and filtering.**
11. **Analysis of the population structure.**

Lectures:

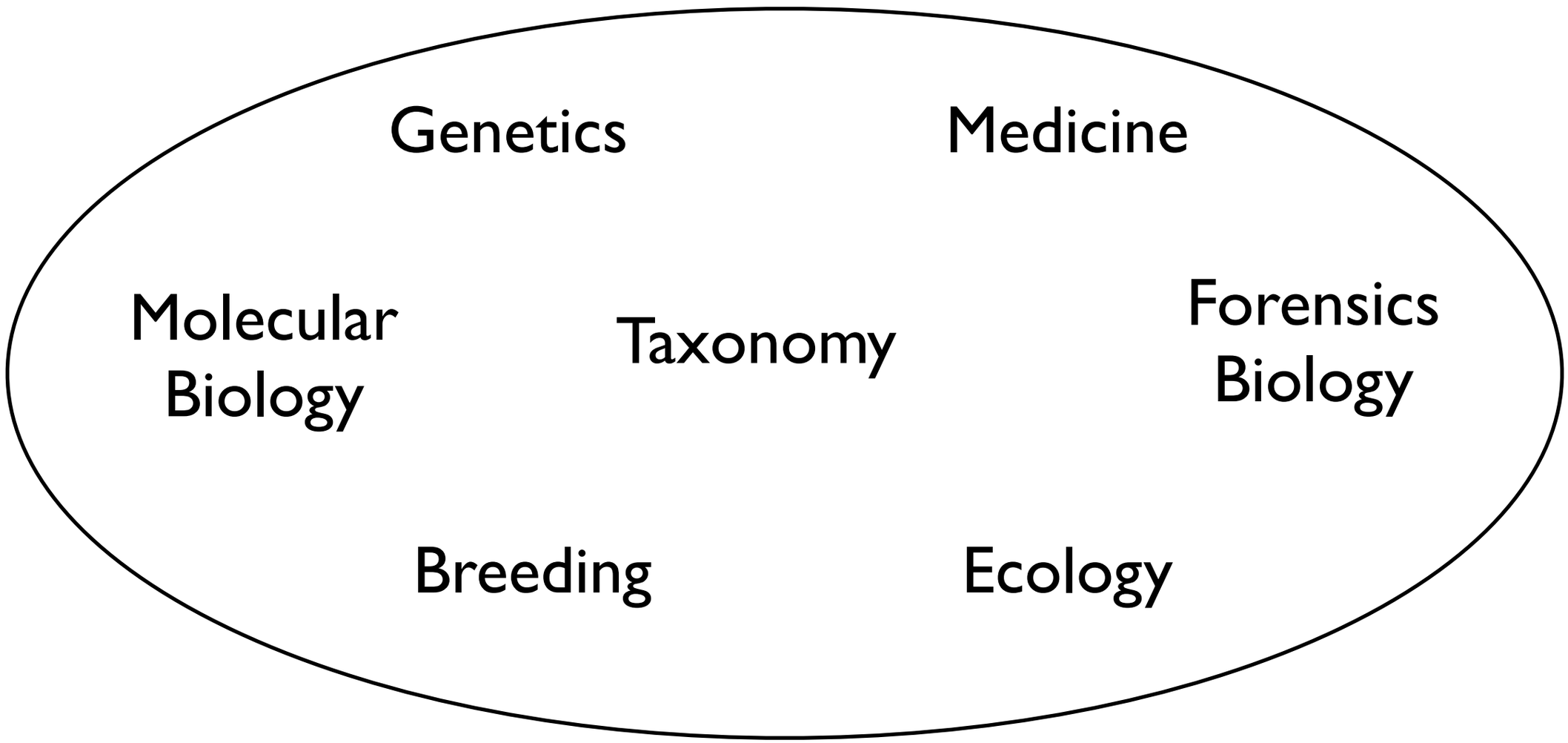
1. **Basics of the Next Generation Sequencing (NGS).**
   - 1.1. The sequencing revolutions.
   - 1.2. Strengths and weaknesses of the different technologies.
   - 1.3. Inputs and outputs.
2. **RNAseq experiment design.**
   - 2.1. Reference vs Non-reference.
   - 2.2. High heterozygosity and polyploid polyploid problem.
   - 2.3. Tissue selection and treatments.
   - 2.4. Sequencing technology.
3. **RNAseq expression analysis.**
   - 3.1. Reference preparation and read mapping.
   - 3.2. Gene expression.
   - 3.3. Analysis and visualization.
4. **Use of RNAseq reads for phylogeny and genetics.**
   - 4.1. Recovering full length mRNA: Reference guided assembly.
   - 4.2. Phylogeny though RNAseq: From gene tree to species tree.
   - 4.3. From reads to markers: SNP calling.
   - 4.4. Population genetics and NGS.

I. **Basics of the Next Generation Sequencing (NGS).**

DNA Sequencing:

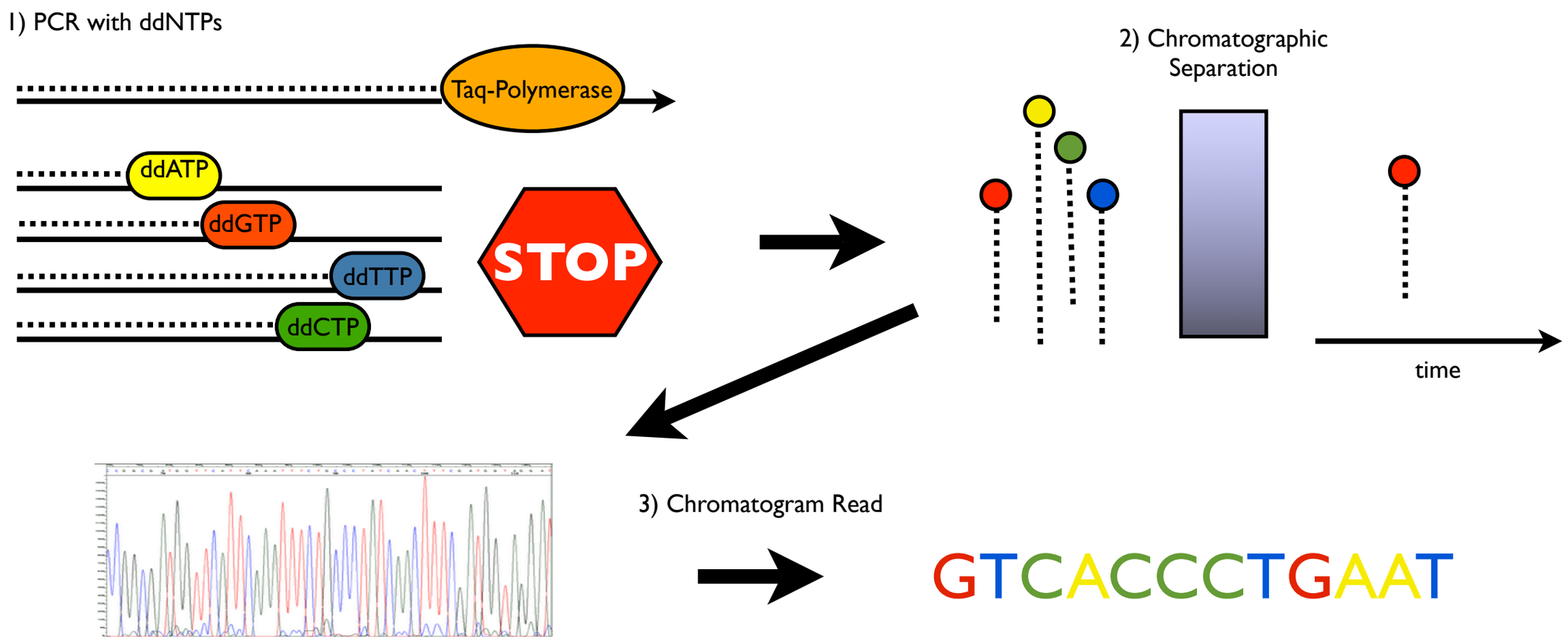
"Process of determining the precise order of nucleotides within a DNA molecule."

-Wikipedia



1. Basics of the Next Generation Sequencing (NGS).

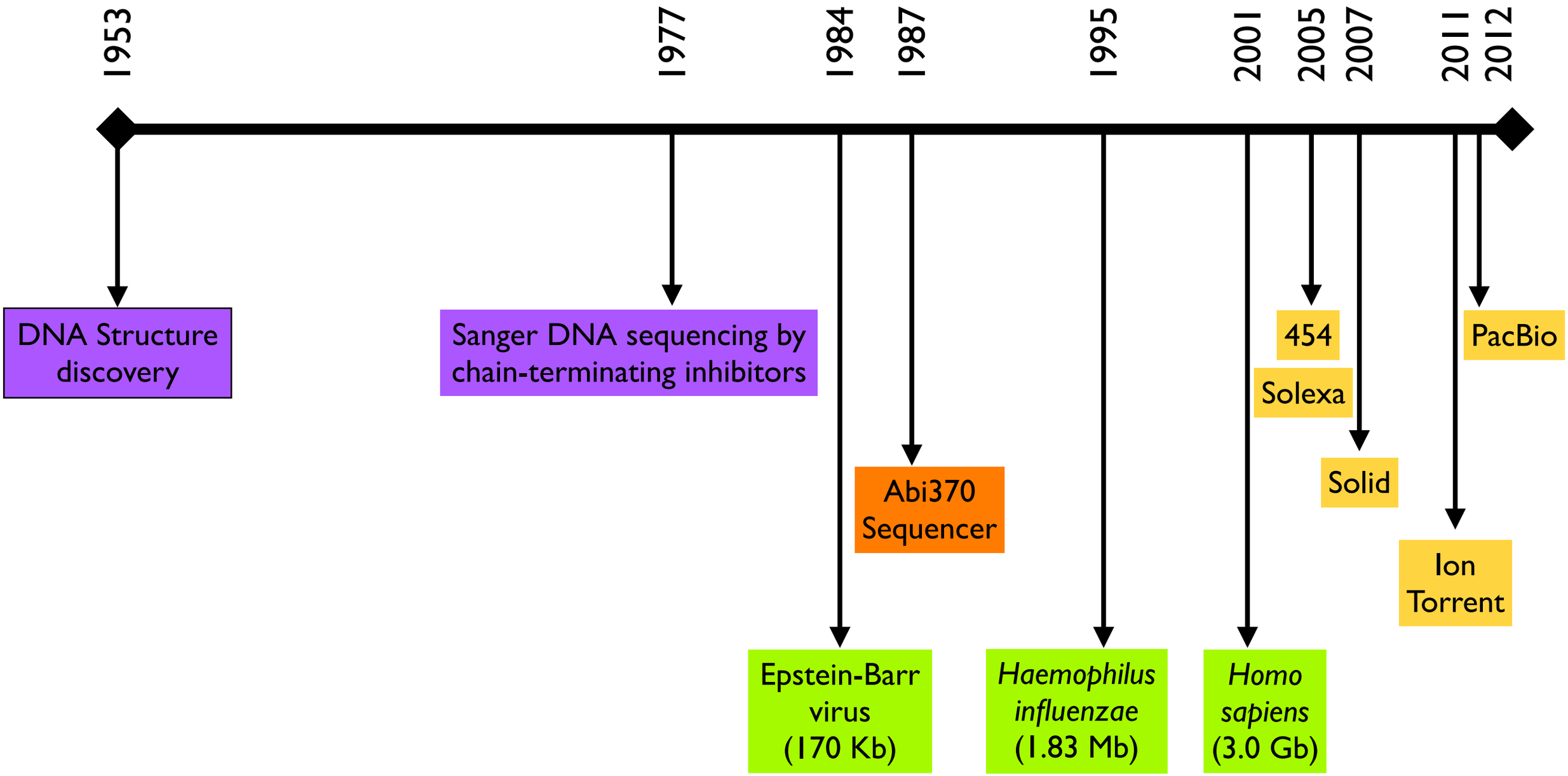
DNA Sanger Sequencing



	Run Time	Sequence Length	Reads/Run	Total nucleotides sequenced
Capillary Sequencing (ABI37000)	~2.5 h	800 bp	386	0.308 Mb

1.1 The sequencing revolutions.

DNA Sequencing NGS



I.I The sequencing revolutions.

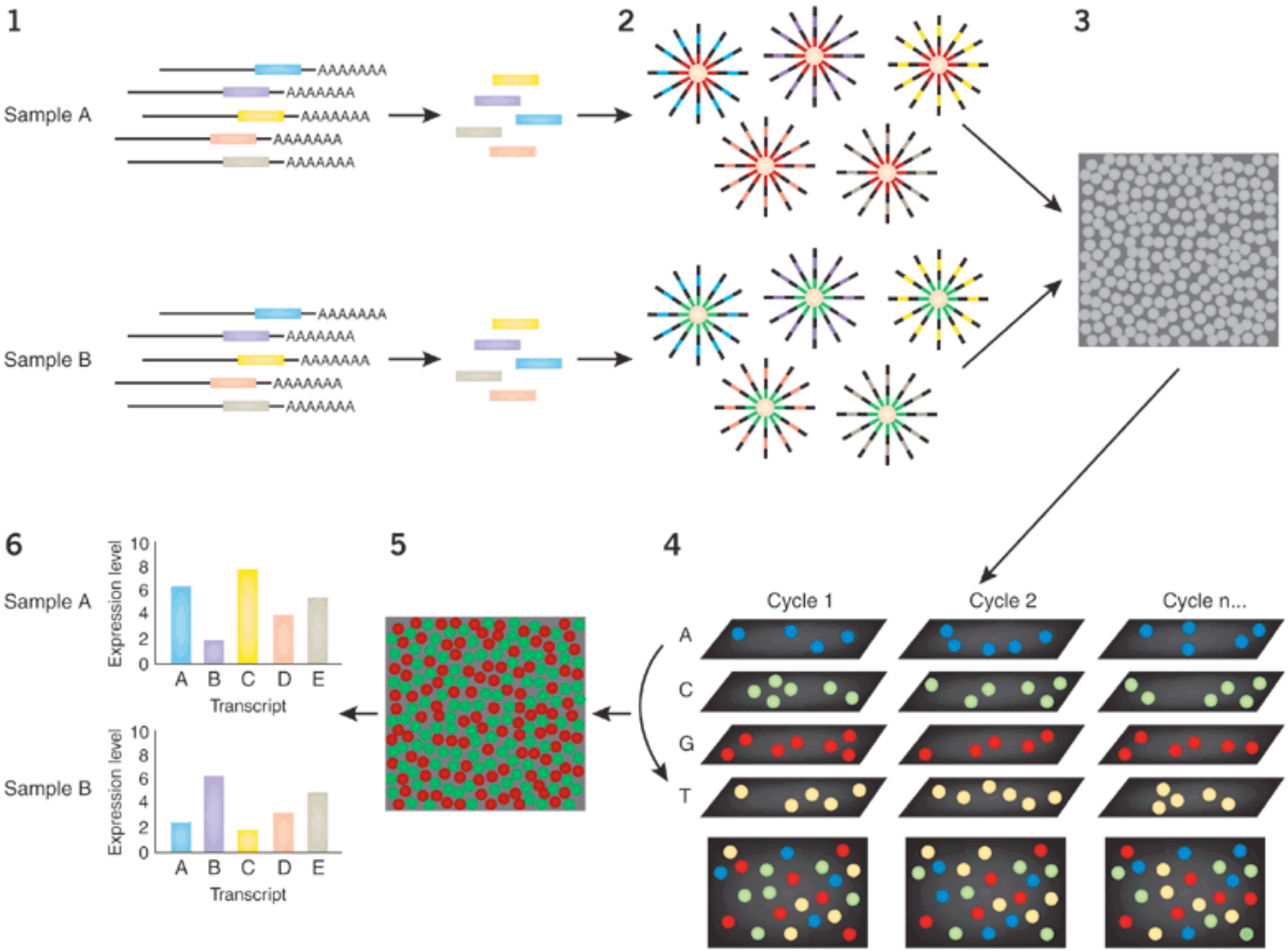
Features of Next Generation Sequencing.

1. **Massive sequence production (from 0.1 to 300 Gb).**
2. **Wide range of sequence lengths (from 50 to 3,000 bp).**
3. **Same or bigger error rate that the traditional sequencing (from 87 to 99.9%).**
4. **Cheap price per base.**

Next Generation Sequencing technologies

- Pyrosequencing (454/Roche).
- Illumina sequencing
- SOLID sequencing
- Ion semiconductor sequencing (IonTorrent)
- Single Molecule SMRT sequencing (PacBio)

1.1 The sequencing revolutions.



Velculescu VE & Kinzler KW *Nature Biotechnology* **25**, 878 - 880 (2007)

1.1 The sequencing revolutions.

Next Generation Sequencing technologies

	Run Time	Sequence Length	Reads/Run	Total nucleotides sequenced per run
Capillary Sequencing (ABI37000)	~2.5 h	800 bp	386	0.308 Mb
454 Pyrosequencing (GS FLX Titanium XL+)	~23 h	700 bp	1,000,000	700 Mb (0.7 Gb)
Illumina (HiSeq 2500)	264 h / 27 h (11 days)	2 x 100 bp 2 x 150 bp	2 x 3,000,000,000 2 x 600,000,000	600,000 / 120,000 Mb (600 / 120 Gb)
Illumina (MiSeq)	39 h	2 x 250 bp	2 x 17,000,000	8,500 Mb (8.5 Gb)
SOLID (5500xl system)	48 h (2 days)	75 bp	400,000,000	30,000 Mb (30 Gb)
Ion Torrent (Ion Proton I)	2 h	100 bp	100,000,000	10,000 Mb (10 Gb)
PacBio (PacBioRS)	1.5 h	~3,000 bp	25,000	100 Mb (0.1 Gb)

1.2 **Strengths and weaknesses of the different technologies.**

Next Generation Sequencing technologies

	Strenghs	Weaknesses
454 Pyrosequencing (GS FLX Titanium XL+)	– Long reads (450/700 bp). – Long insert for mate pair libraries (20Kb). – Low observed raw error rate (0.1%) – Low percentage of PCR duplications for mate pair libraries	– Homopolymer error. – Low sequence yield per run (0.7 Gb). – Preferred assembler (gsAssembler) uses overlapping methodology.
Illumina (HiSeq 2500)	– High sequence yield per run (600 Gb) – Low observed raw error rate (0.26%)	– High percentage of PCR duplications for mate pair libraries. – Long run time (11 days) – High instrument cost (~ $650K)
Illumina (MiSeq)	– Medium read size (250 bp) – Faster run than Illumina HiSeq	– Medium sequence yield per run (8.5 Gb)
SOLID (5500xl system)	– 2-base encoding reduce the observed raw error rate (0.06%)	– 2-base color coding makes difficult the sequence manipulation and assembly. – Short reads (75 bp)
Ion Torrent (Ion Proton I)	– Fast run (2 hours) – Low instrument cost ($80K). – Medium read size (200 bp)	– Medium sequence yield per run (10 Gb) – Medium observed raw error rate (1.7%)
PacBio (PacBioRS)	– Long reads (3000 bp) – Fast run (2 hours)	– Really high observed raw error rate (12.7%) – High instrument cost (~ $700K) – No pair end/mate pair reads

1.3 **Inputs and Outputs.**

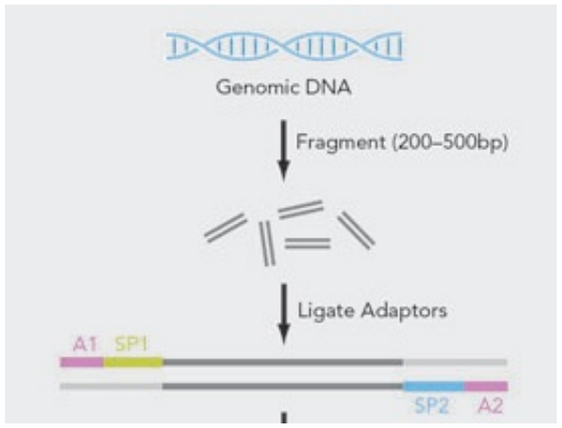
Next Generation Sequencing technologies

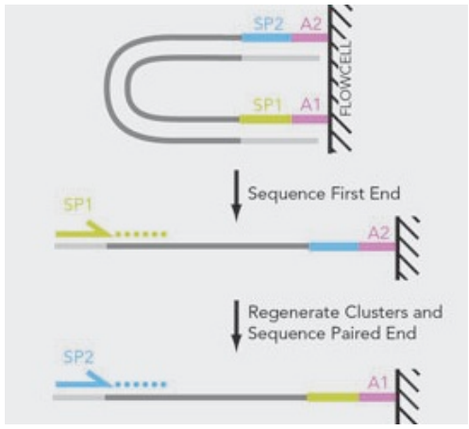
	Inputs	Outputs
454 Pyrosequencing (GS FLX Titanium XL+)	– Single Reads Library. – Pair End Library (3 to 20 Kb insert size). – Multiplexed sample.	– sff files – (fasta and fastq files)
Illumina (HiSeq 2500)	– Single Reads Library. – Pair End Library (170-800 bp insert size). – Mate Pair Library (2 to 10 Kb insert Size) – Multiplexed sample.	– fastq files (Phred+64) – fastq files (Phred+33, Illumina 1.8+)
Illumina (MiSeq)		
SOLID (5500xl system)	– Single Reads Library. – Mate Pairs Library (0.6 to 6 Kb insert size). – Multiplexed sample.	– fastq files (Phred+33)
Ion Torrent (Ion Proton I)	– Single Reads Library. – Pair End Library (0.6 to 6 Kb insert size). – Multiplexed sample.	
PacBio (PacBioRS)	– Single Reads Library.	

Next Generation Sequencing technologies

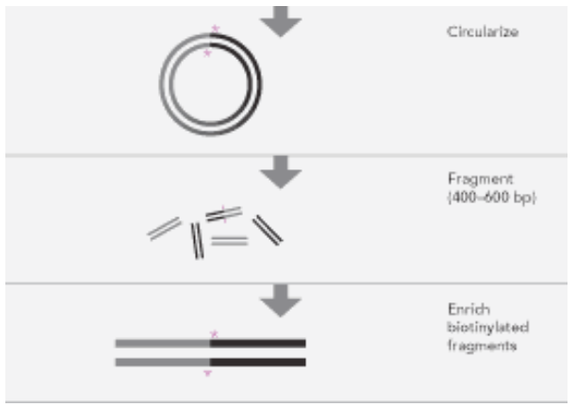
★ Library types:

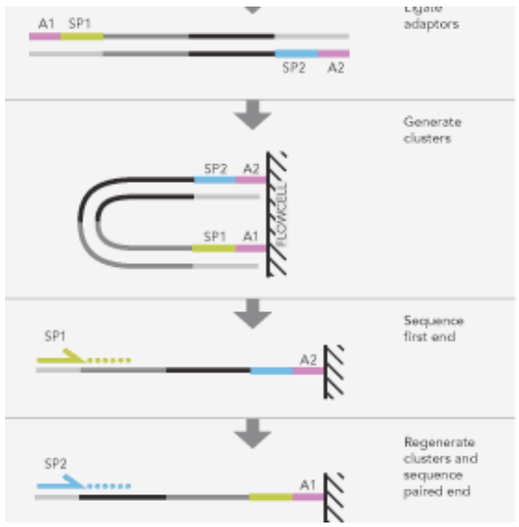
- Single reads
- Pair ends (PE) (from 150-800 bp)





- Mate pairs (MP) (from 2Kb to 20 Kb)





Next Generation Sequencing technologies

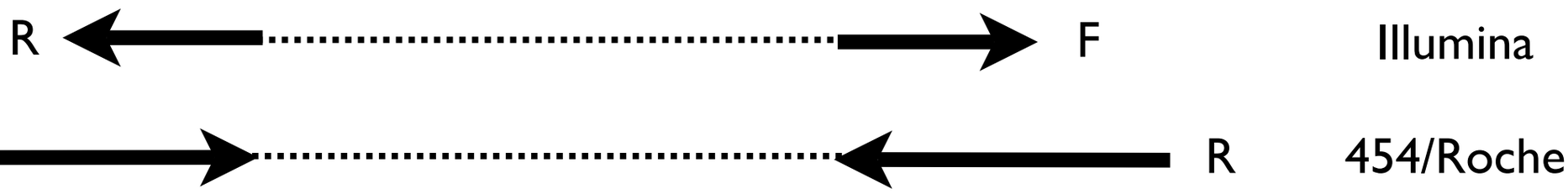
★ Library types (orientations):

- Single reads

F ⟶

- Pair ends (PE) (150-800 bp insert size)

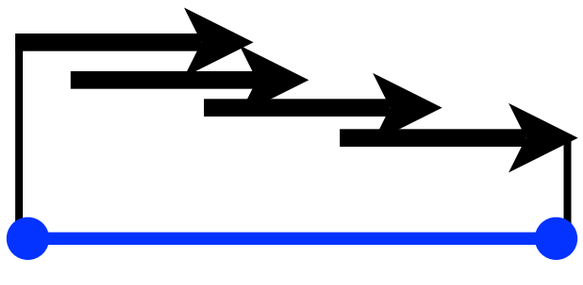
- Mate pairs (MP) (2-20 Kb insert size)



Next Generation Sequencing technologies

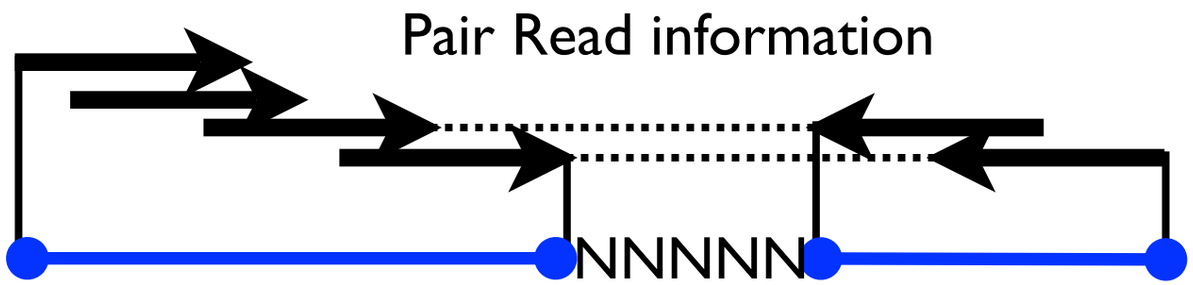
- Why is important the pair information ?

  - *novo* assembly:

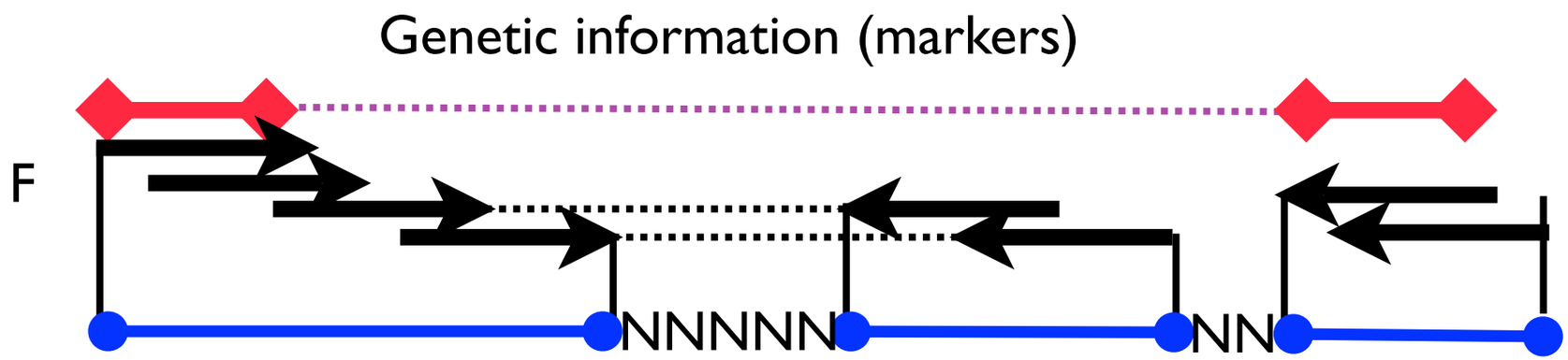


Reads

Consensus sequence
(Contig)



Scaffold
(or Supercontig)

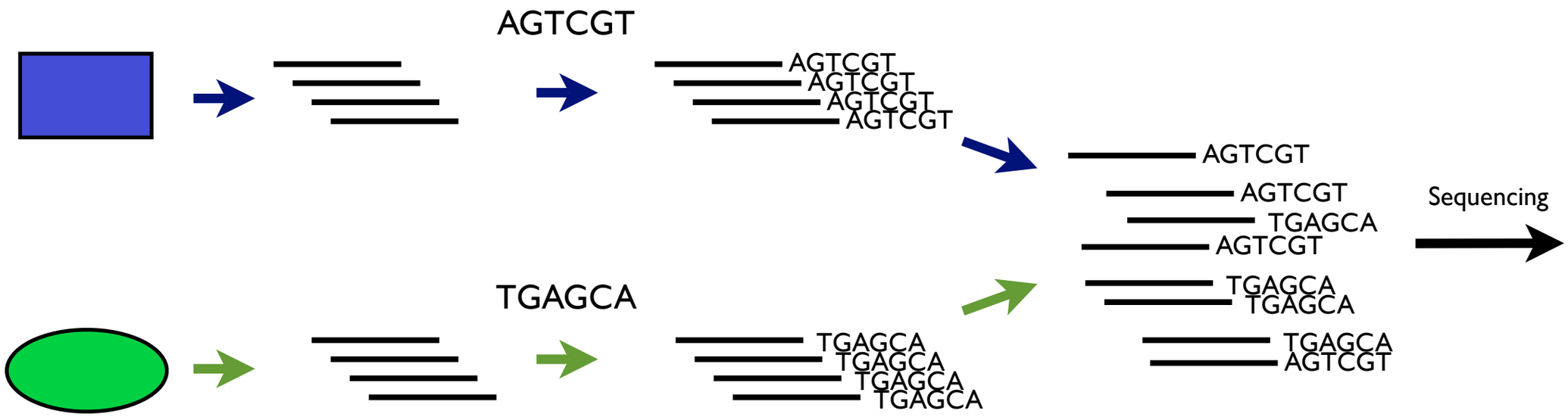


Pseudomolecule
(or ultracontig)

Next Generation Sequencing technologies

★ Multiplexing:

Use of different tags (4-6 nucleotides) to identify different samples in the same lane/ sector.



1.3 **Inputs and Outputs.**

Sff files:

Standard flowgram format (SFF) is a **binary file** format used to encode results of pyrosequencing from the **454** Life Sciences platform for high-throughput sequencing. SFF files can be viewed, edited and converted with DNA Baser SFF Workbench (graphic tool), or converted to FASTQ format with **sff2fastq** or **sff_extract**.

-Wikipedia

sff2fastq, (program written in C)
https://github.com/indraniel/sff2fastq

sff_extract, (program written in Python)
http://bioinf.comav.upv.es/sff_extract/download.html

1.3 **Inputs and Outputs.**

Fasta files:

It is a text-based format for representing either nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes.

-Wikipedia

```
>SEQUENCE_ID1 DESCRIPTION
ATGCGCGCGCGCGCGCGCGGGTAGCAGATGACGACACAGAGCGAGGATGCGCTGAGAGTA
GTGTGACGACGATGACGGAAAATCAGATGGACCCGATGACAGCATGACGATGGGACGGGA
AAGATTGGACCAGGACAGGACCAGGACCAGGACCAGGGATTAGA
>SEQUENCE_ID2 DESCRIPTION
ATGGGGGGGACGACGATGGACACAGAGACAGAGACGACGACAGCAGACAGATTTACCTTA
GACGAGATAGGAGAGACGACAGATATATATATATATAGCAGACAGACAGACATTTAGACGAG
ACGACGATAGACGAT
```

1.3 **Inputs and Outputs.**

Fastq files:

FASTQ format is a **text-based format** for storing both a biological **sequence** (usually nucleotide sequence) and its corresponding **quality scores.**

-Wikipedia

```
@SEQUENCE_ID1
ATGCGCGCGCGCGCGCGGGTAGCAGATGACGACACAGAGCGAGGATGCGCTGAGAGTA
GTGTGACGACGATGACGGAAAATCAGA
+
BBBBBPPPPPXXXXX^^^^^^^^^^^^^^^^^^^^^^^^^__^^^^^^^^^^^^^_eeeeeee
[[[[[^^^^]]]]XXXXXPPPPPBBBB
```

1. **Single line ID** with at symbol ("@") in the first column.
2. There should be not space between "@" symbol and the first letter of the identifier.
3. Sequences are in multiple lines after the ID line
4. Single line with plus symbol ("+") in the first column to represent the quality line.
5. Quality ID line can have or have not ID
6. Quality values are in multiple lines after the + line

1.3 **Inputs and Outputs.**

Fastq files:

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................
.........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
.............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
.................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                                     |                               |
33                        59   64       73                                   104                             126

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)
```

Phred score of a base is: **Qphred=-10 log10 (e)**

Q=15	e=0.03	(min. used Sanger)
Q=20	e=0.01	(min. used 454 and Illumina)
Q=30	e=0.001	(standard used 454)

Exercises:

1. **Basic Linux commands.**
2. **Sequencing evaluation.**
3. **Simple read mapping.**
4. **Simple de-novo assembly.**
5. **Basic R commands**
6. **Functional annotation.**
7. **Differential gene expression.**
8. **Cluster analysis for gene expression.**
9. **Selecting genes for phylogeny.**
10. **SNP calling and filtering.**
11. **Analysis of the population structure.**

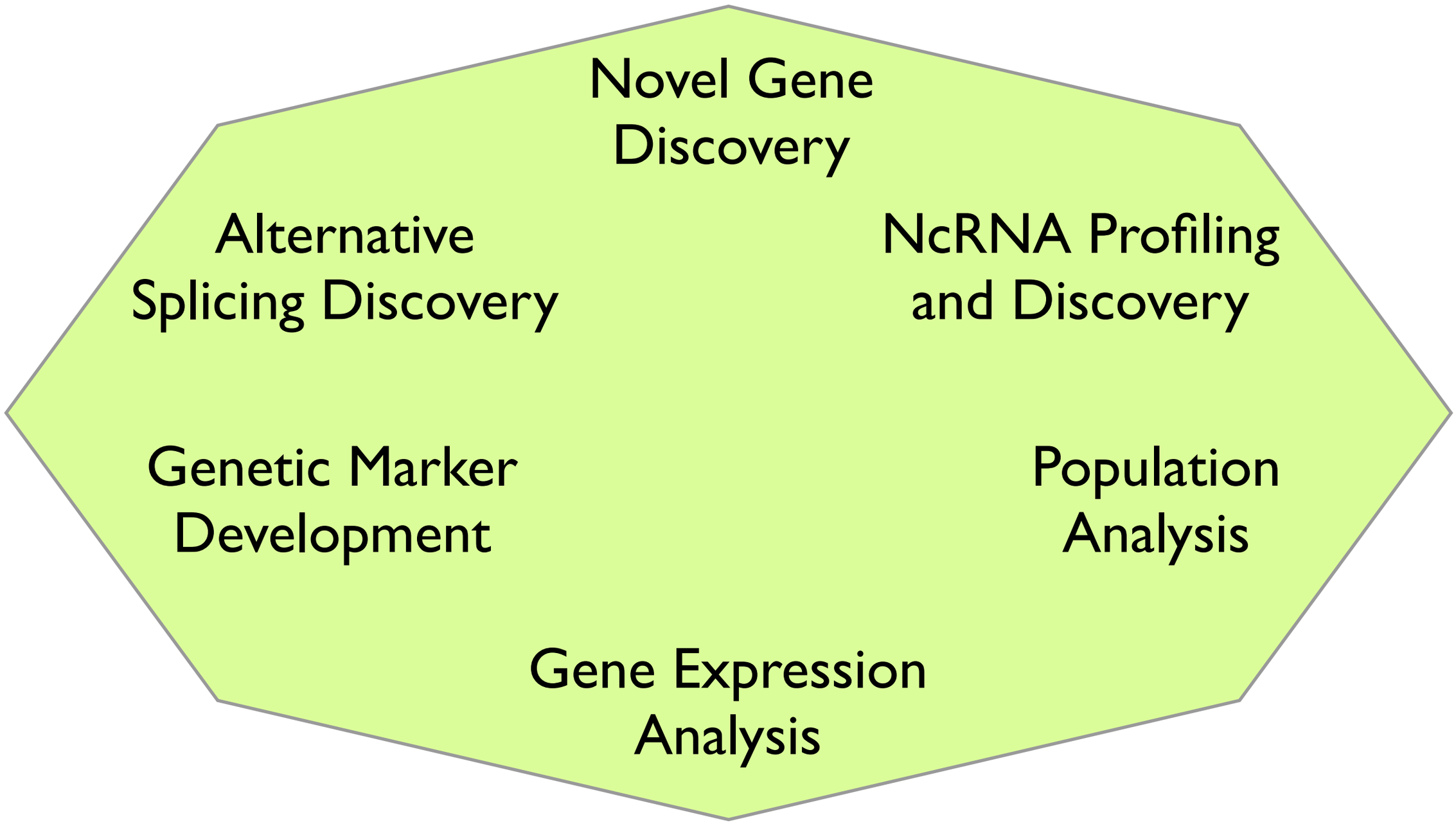
Exercises:

1. **Basic Linux commands.**
2. **Sequencing evaluation.**
3. **Simple read mapping.**
4. **Simple de-novo assembly.**
5. **Basic R commands**
6. **Functional annotation.**
7. **Differential gene expression.**
8. **Cluster analysis for gene expression.**
9. **Selecting genes for phylogeny.**
10. **SNP calling and filtering.**
11. **Analysis of the population structure.**

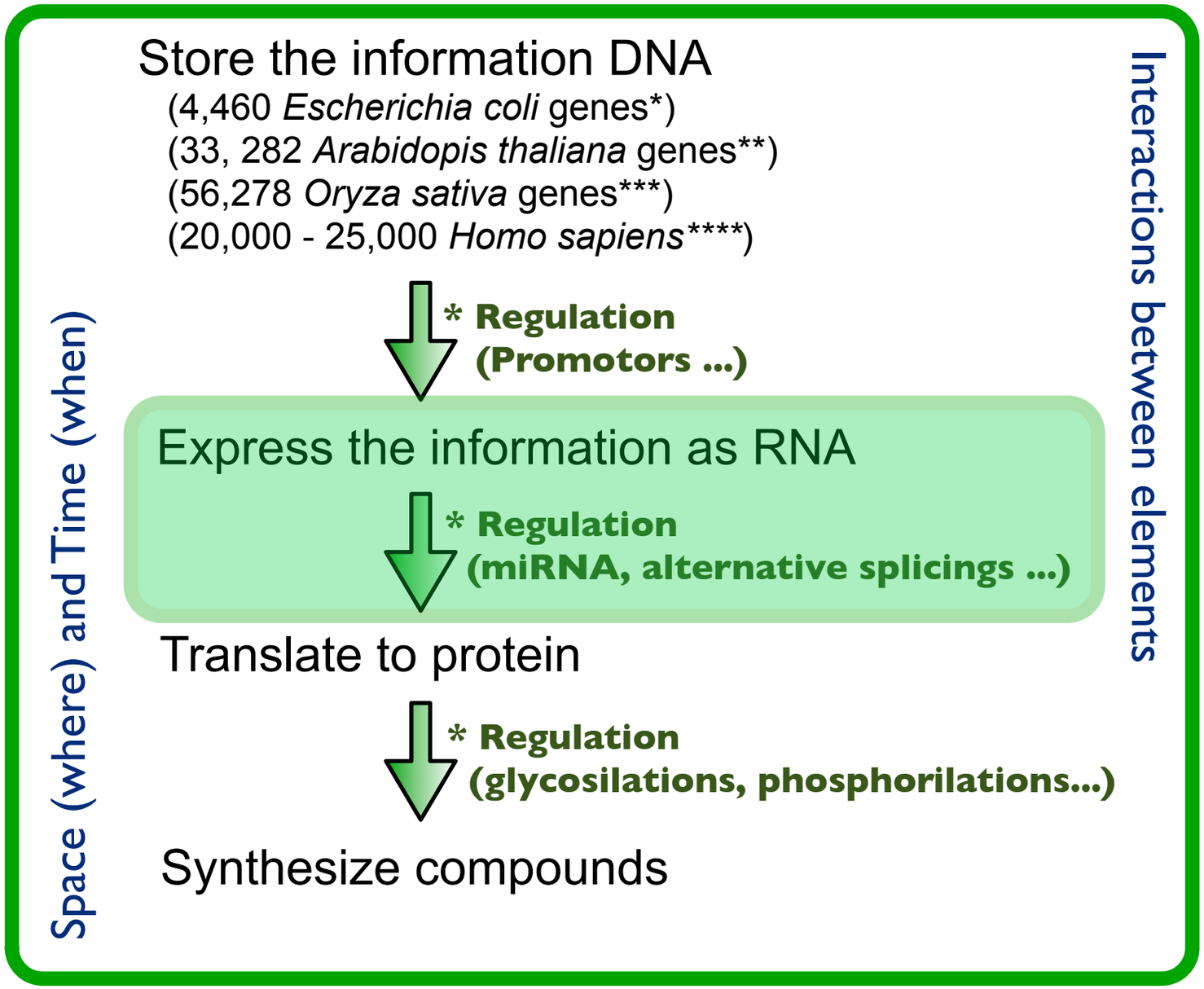
Lectures:

1. **Basics of the Next Generation Sequencing (NGS).**
   - 1.1. The sequencing revolutions.
   - 1.2. Strengths and weaknesses of the different technologies.
   - 1.3. Inputs and outputs.
2. **RNAseq experiment design.**
   - 2.1. Reference vs Non-reference.
   - 2.2. High heterozygosity and polyploid polyploid problem.
   - 2.3. Tissue selection and treatments.
   - 2.4. Sequencing technology.
3. **RNAseq expression analysis.**
   - 3.1. Reference preparation and read mapping.
   - 3.2. Gene expression.
   - 3.3. Analysis and visualization.
4. **Use of RNAseq reads for phylogeny and genetics.**
   - 4.1. Recovering full length mRNA: Reference guided assembly.
   - 4.2. Phylogeny though RNAseq: From gene tree to species tree.
   - 4.3. From reads to markers: SNP calling.
   - 4.4. Population genetics and NGS.

Applications for NGS Transcriptomics:



Complex systems may create complex transcriptomes

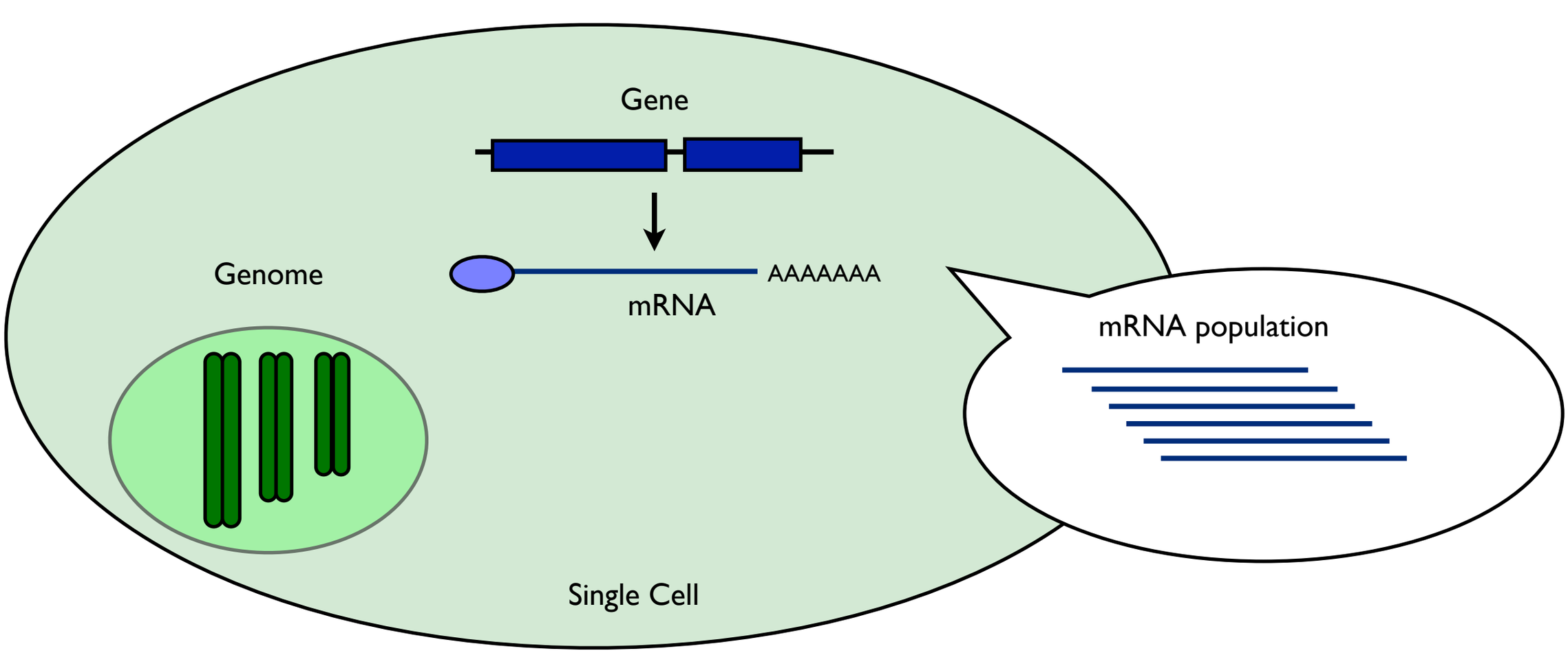


* Karp et al. *Multidimensional annotation of the Escheriichia coli K-12 genome.* Nucleid Acid Research. 2007:35:7577-7590
** http://www.arabidopsis.org/portals/genAnnotation/genome_snapshot.jsp
*** http://rice.plantbiology.msu.edu/riceInfo/info.shtml
**** http://www.sanger.ac.uk/Info/Press/2004/041020.shtml

Transcriptome Complexity:

Simple System:

One Genome => Gene 1 copy => Single mRNA

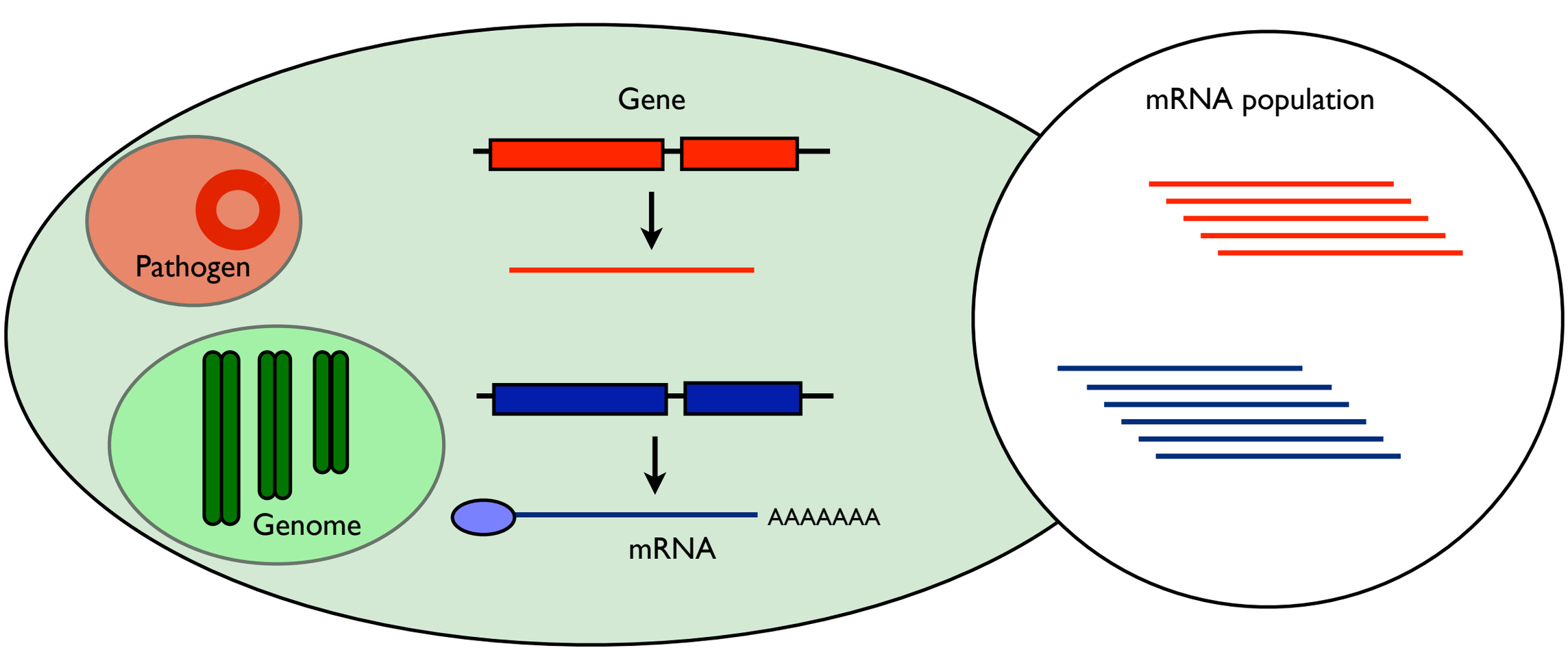


Transcriptome Complexity:

How many species we are analyzing ?

1) Problems to isolate a single species (rhizosphere)

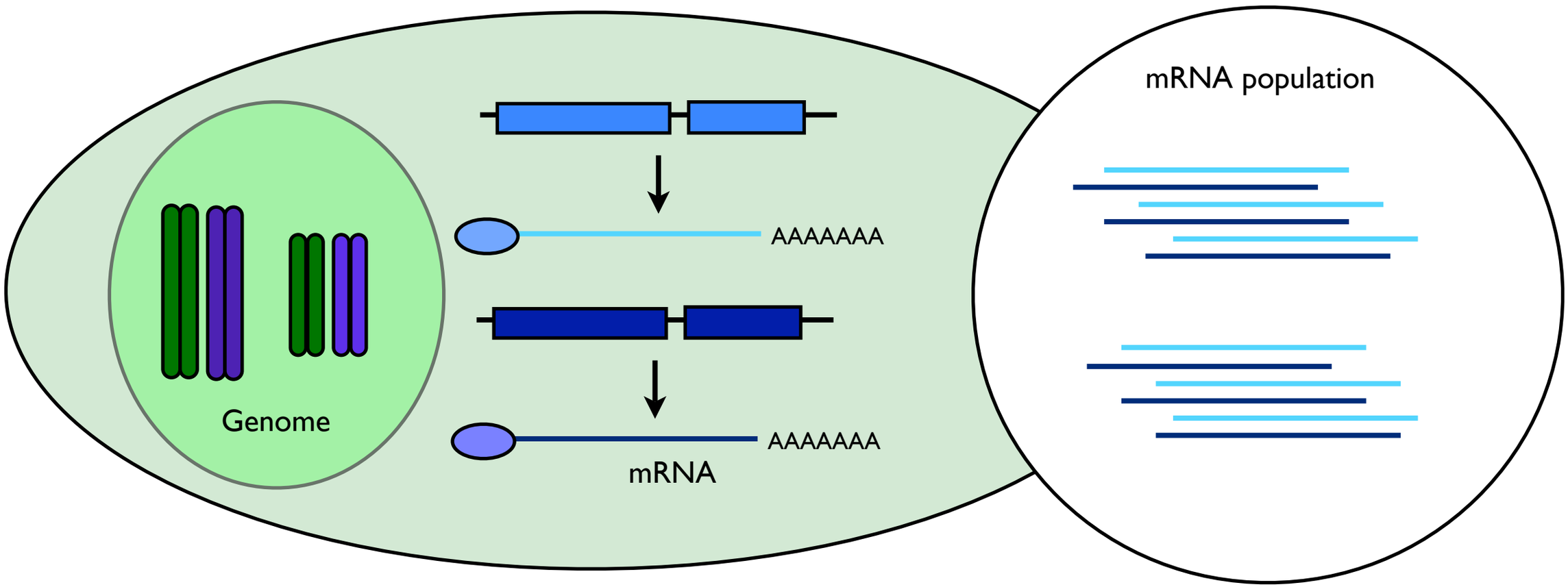
2) Species interaction study (plant-pathogen)



Transcriptome Complexity:

How many possible alleles we expect per gene ?

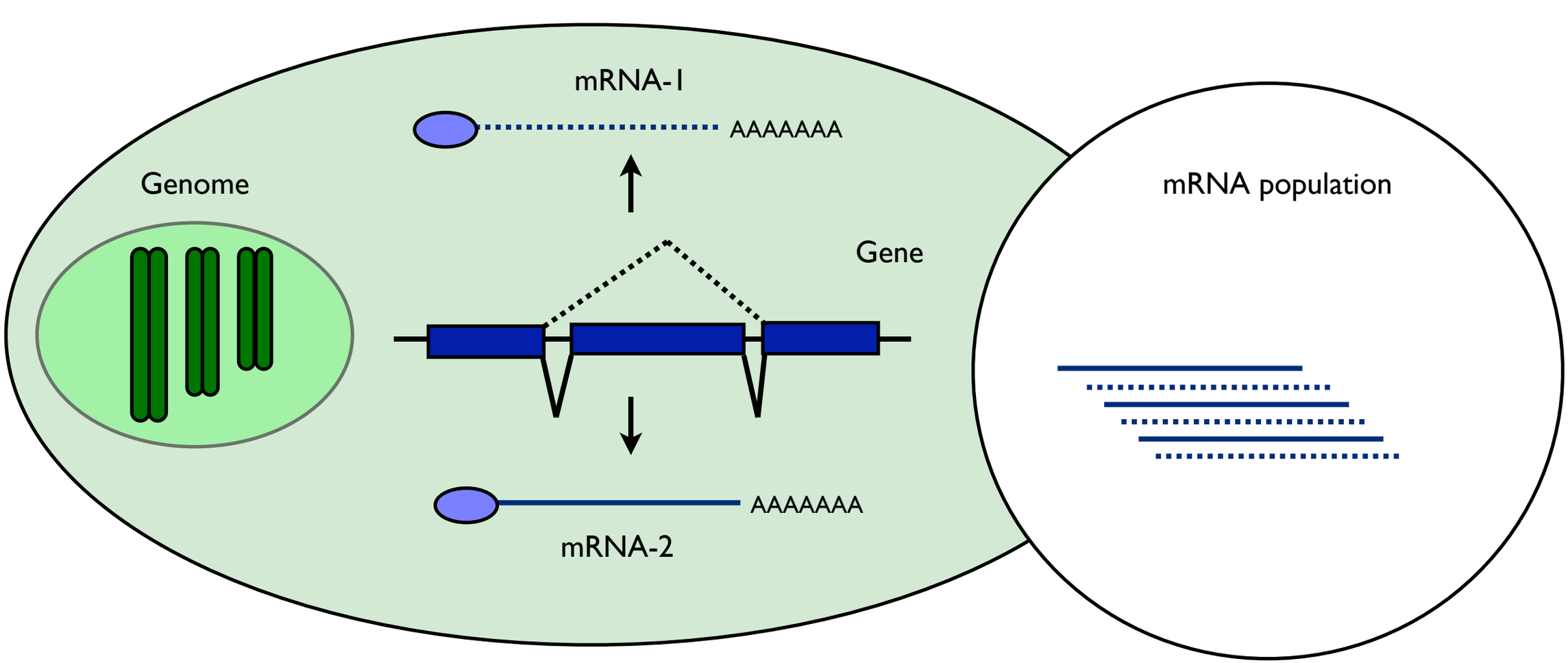
1) Polyploids (autopolyploids, allopolyploids).
2) Heterozygosity
3) Complex Gene Families (tandem duplications)



Transcriptome Complexity:

How many isoforms we expect for each allele ?

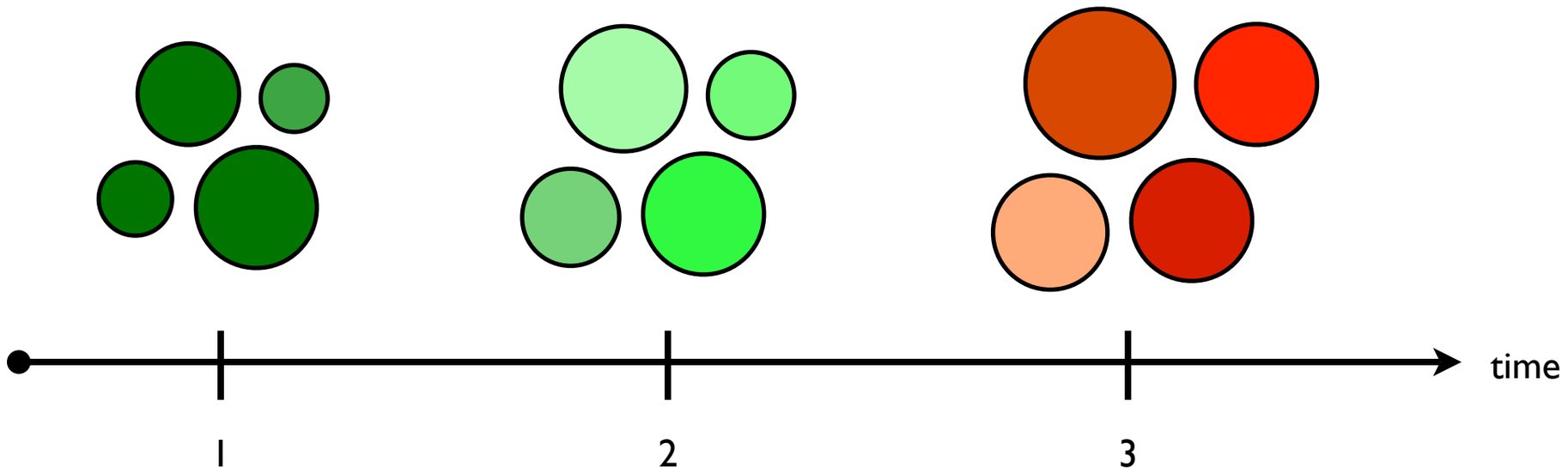
1) Alternative splicings



Transcriptome Complexity:

Is the study performed at different time points?

1) Developmental stages (difficult to select the same)
2) Response to a treatment

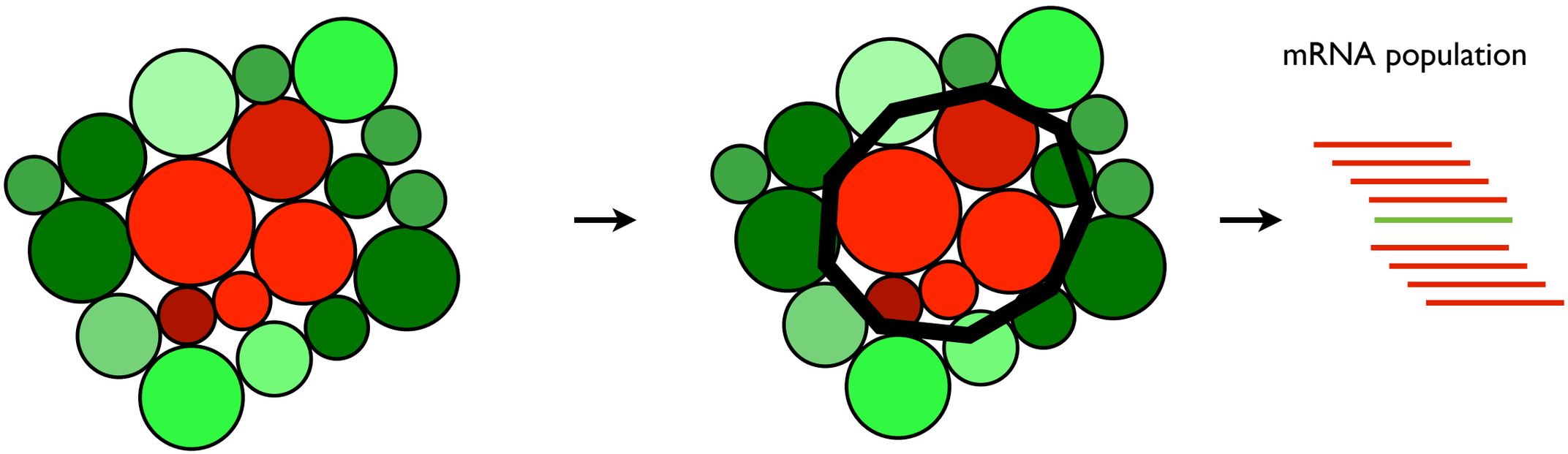


Transcriptome Complexity:

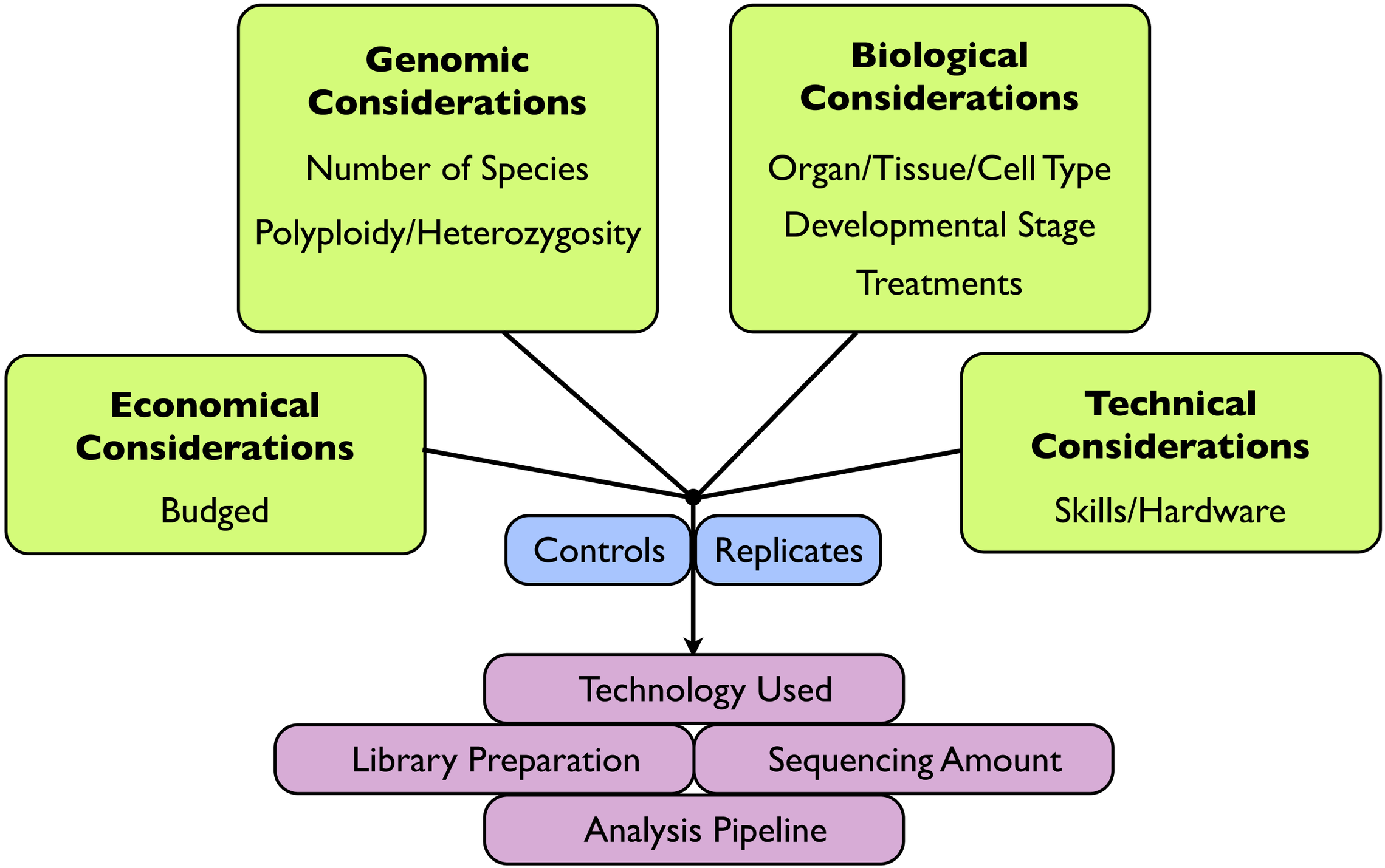
Is the study performed with different parts?

1) Organ specific
2) Tissue/Cell type specific

   (Laser Capture Microdissection, LCM)



Experimental Design:



Reference:

Generally a genomic sequence with gene models used to align the reads, but a reference can be a *de-novo* assembled transcriptome.

The most important advantage of the use of a reference is that the analysis is **computationally less intense** because it only needs to align reads.

Methodology	Technology	Program	Minimum RAM*	Time*
Mapping	454	gsMapper	1 Gb	several hours
	Illumina	Bowtie2	2 Gb**	< 1 hour
De-novo	454	gsAssembler	8 Gb	> 1 day
	Illumina	Trinity	16 Gb	> 1 day
		SOAPdenovo-trans	16 Gb	several hours

* Rough approach
** Human genome size (~3 Gb)

2.1 Reference vs. Non-reference

Plant Genomes:

Plant genomes in this database (26 genomes)					
Species name	Common name	Release version	Gene number	Access	Reference
Arabidopsis lyrata	Lyrate rockcress	Version 1.0 (Apr 2011)	32,670	JGI	*Nature Genetics*
Arabidopsis thaliana	Arabidopsis	TAIR 9.0 (Jun 2009)	27,379	TAIR	*Nature*
Brachypodium distachyon	Purple false brome	Phytozome v6.0	32,255	JGI	*Nature*
Brassica rapa	Chinese cabbage	Version 1.1	22,285	BRAD	*Nature Genetics*
Cajanus cajan	Pigeonpea	Nov 2011	48,680	IIPG	*Nature Biotechnology*
Carica papaya	Papaya	Dec 2007	25,536	Hawaii	*Nature*
Chlamydomonas reinhardtii	Green algae	Version 4.2	16,036	JGI	*Science*
Cucumis sativus	Cucumber	Phytozome v6.0	21,491	JGI	*Nature Genetics*
Fragaria vesca	Strawberry	Dec 2010	34,809	PFR	*Nature Genetics*
Glycine max	Soybean	Release 1 (Dec 2008)	66,153	JGI	*Nature*
Lotus japonicus	Lotus	Release 2.5	42,399	Kazusa	*DNA research*
Musa acuminata	Banana	Jul 2012	36,542	CIRAD	*Nature*
Malus x domestica	Apple	Aug 2010	57,386	IASMA	*Nature Genetics*
Medicago truncatula	Barrel medic	Mt3.5 v3 (Jun 2011)	45,108	JCVI	*Nature*
Oryza sativa	Rice	RAP 2.0 (Nov 2007)	30,192	RAP	*Nature*
Physcomitrella patens	Moss	Version 1.6 (Jan 2008)	32,272	JGI	*Science*
*Prunus persica**	Peach	Version 1.0	27,864	JGI	-
Populus trichocarpa	Western poplar	JGI 2.0 (Feb 2010)	45,778	JGI	*Science*
Ricinus communis	Castor bean	Release 0.1 (May 2008)	38,613	JCVI	*Nature Biotechnology*
Sorghum bicolor	Sorghum	Sbi 1.4 (Dec 2007)	34,496	JGI	*Nature*
Solanum lycopersicum	Tomato	Version 2.3	34,727	SGN	*Nature*
Selaginella moellendorffii	Selaginella	Version 1.0 (Dec 2007)	22,273	JGI	*Science*
Solanum tuberosum	Potato	Version 3.4	39,031	PGSC	*Nature*
Theobroma cacao	Cacao	Release 0.9 (Sep 2010)	28,798	CIRAD	*Nature Genetics*
Vitis vinifera	Grape vine	Genoscope (Aug 2007)	26,346	Genoscope	*Nature*
Zea mays	Maize	Release 5a (Nov 2010)	32,540	AGI	*Science*

http://chibba.agtec.uga.edu/duplication/index/home

Reference and phylogenetic relations:

Can I use as a reference a different accession ?

Yes

Can I use as a reference a different species ?

Same genus

For most of them, but some losses are expected for the most polymorphic genes

Same family

Probably not. Still some reads will map with the most conserved genes.

Reference and phylogenetic relations:

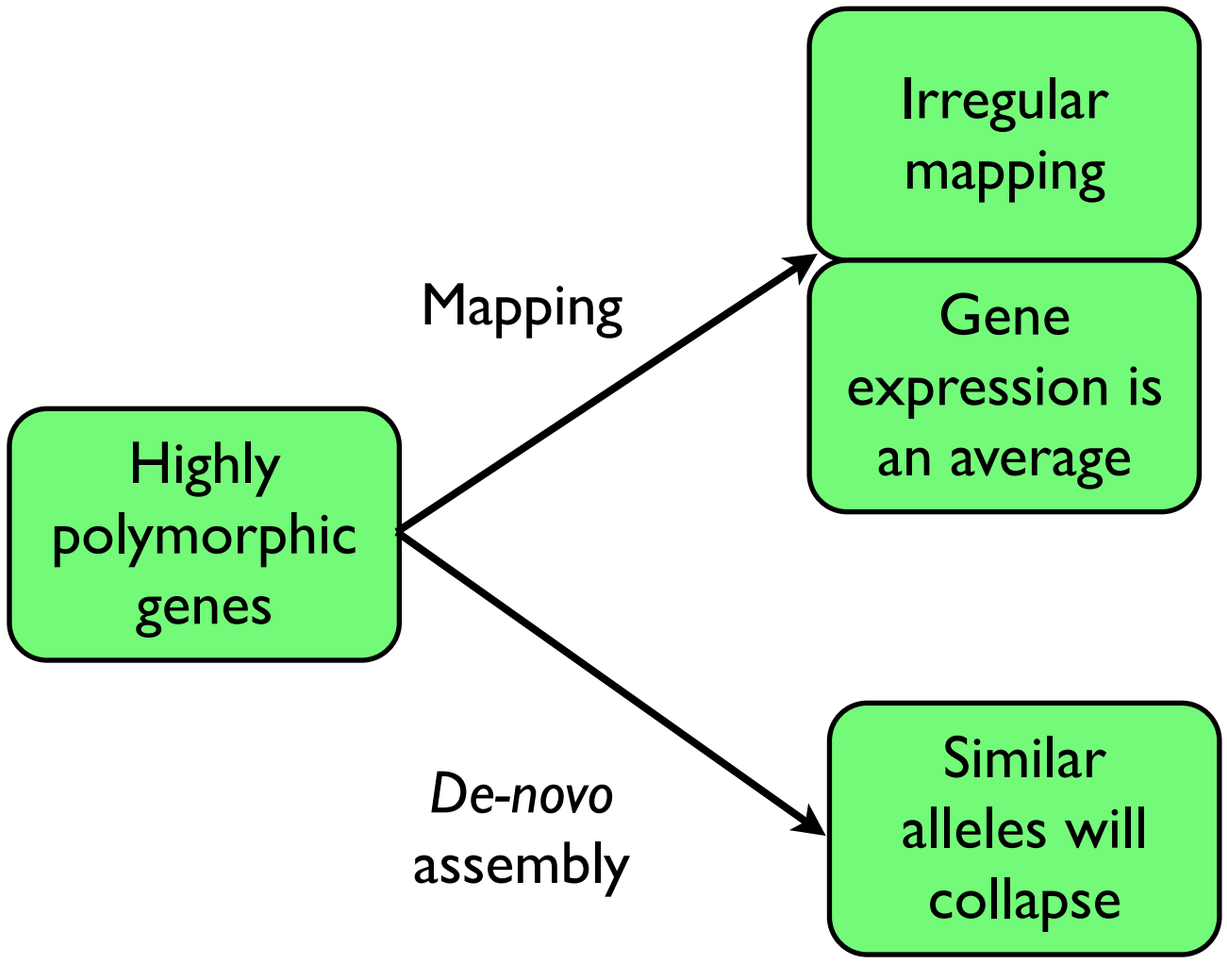
Percentage of mapped reads using *Arabidopsis thaliana* col. as reference

Species	Accession	SRA	Reads	% Mapped Read	Time
Arabidopsis thaliana	Col	SRR513732	12672866	75%	00:11:45
	Ler	SRR392121	9752382	71%	00:07:05
	C24	SRR392124	6186734	72%	00:04:29
Arabidopsis lyrata	-	SRR072809	9214967	69%	00:10:11
Brassica rapa	-	ERR037339	29230003	20%	00:23:15

This test was performed using 1 core. The memory peak was 155 Mb. Reads were preprocessed with Q20 L30. Mapping tool: Bowtie2

High heterozygosity/Polyploid problem:

mRNA from species with a high heterozygosity or a polyploid genome can produce highly polymorphic reads for the same gene.



Reference Gene I

ATGCGCGCTAGACGACATGACGACA

CACTTGACGACATGACG **Gene I A**

CTTGACGACATGACGAC
CCCTTGACGACATGACG **Gene I B**
CGCCCTTGACGACATGA

Expression Gene I = A + B

CACTTGACGACATGACG **Gene I A**

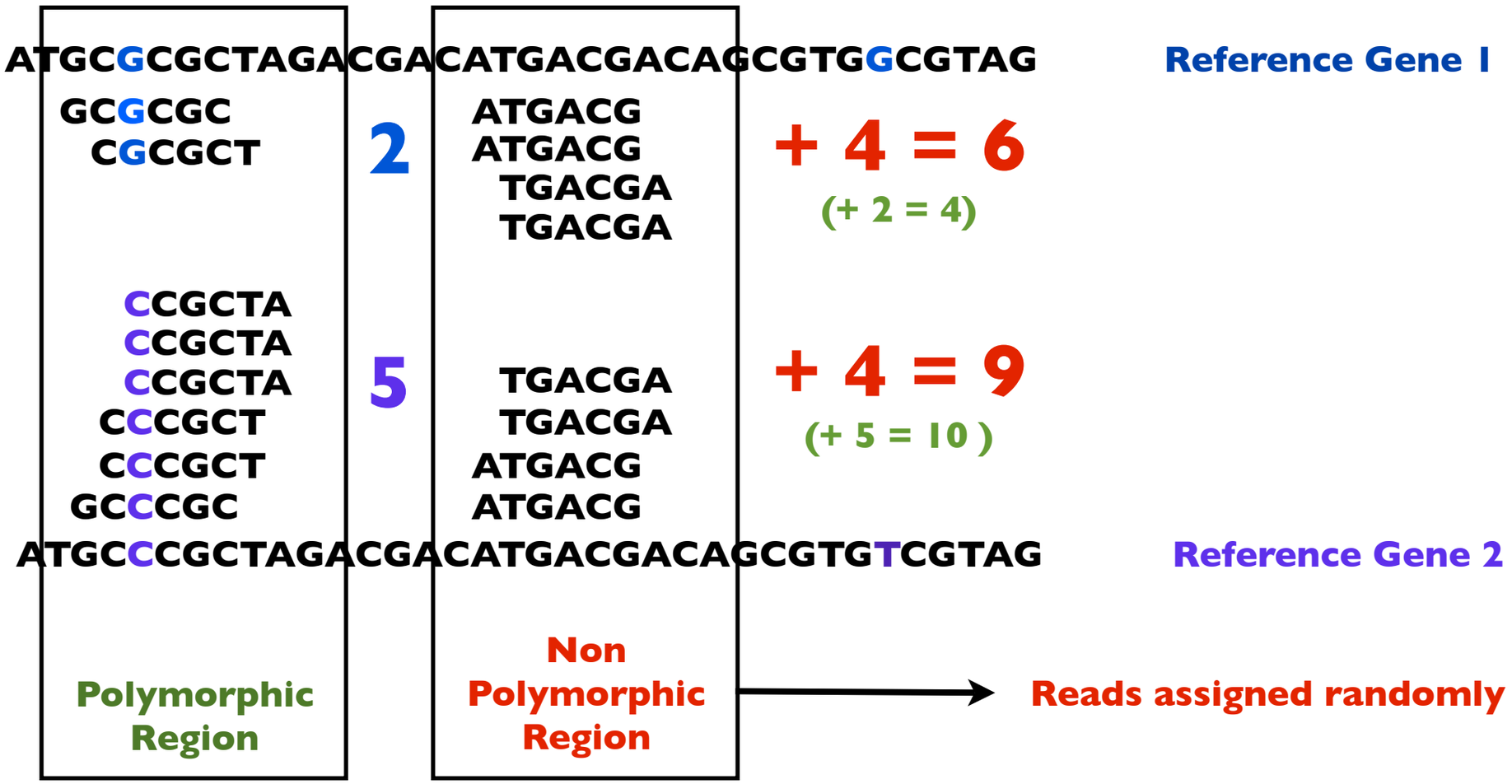
CTTGACGACATGACGAC
CCCTTGACGACATGACG **Gene I B**
CGCCCTTGACGACATGA

CGCCCTTGACGACATGACGACA

Collapsed consensus Gene A + Gene B

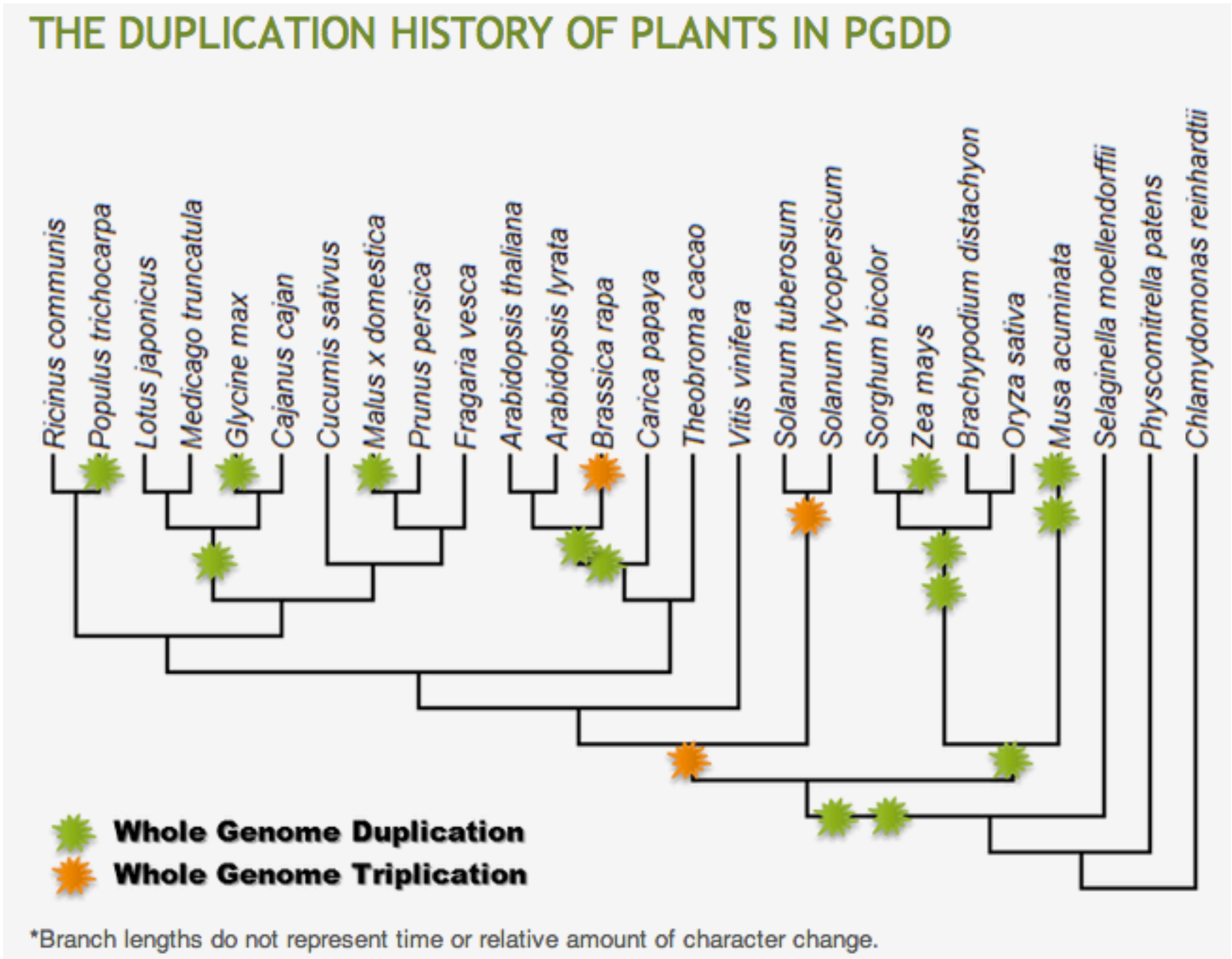
High heterozygosity/Polyploid problem:

If the reference is a paleoploid and it had recent WGD event, the mapping can be irregular and produce an important bias.



High heterozygosity/Polyploid problem:

If the reference is a paleoploid and it had recent WGD event, the mapping can be irregular.



http://chibba.agtec.uga.edu/duplication/index/home

Tissue selection and treatments:

Different organs, tissues or cell types can produce different mRNA extraction yields.

For samples where a low yield is expected a common practice is 1 to 3 rounds of **cDNA amplification**, specially using techniques such as LCM.

↓

Amplifications produce severe **bias** for between low/high represented transcripts

↓

Best Practices:

1) Compare samples with same number of amplification rounds

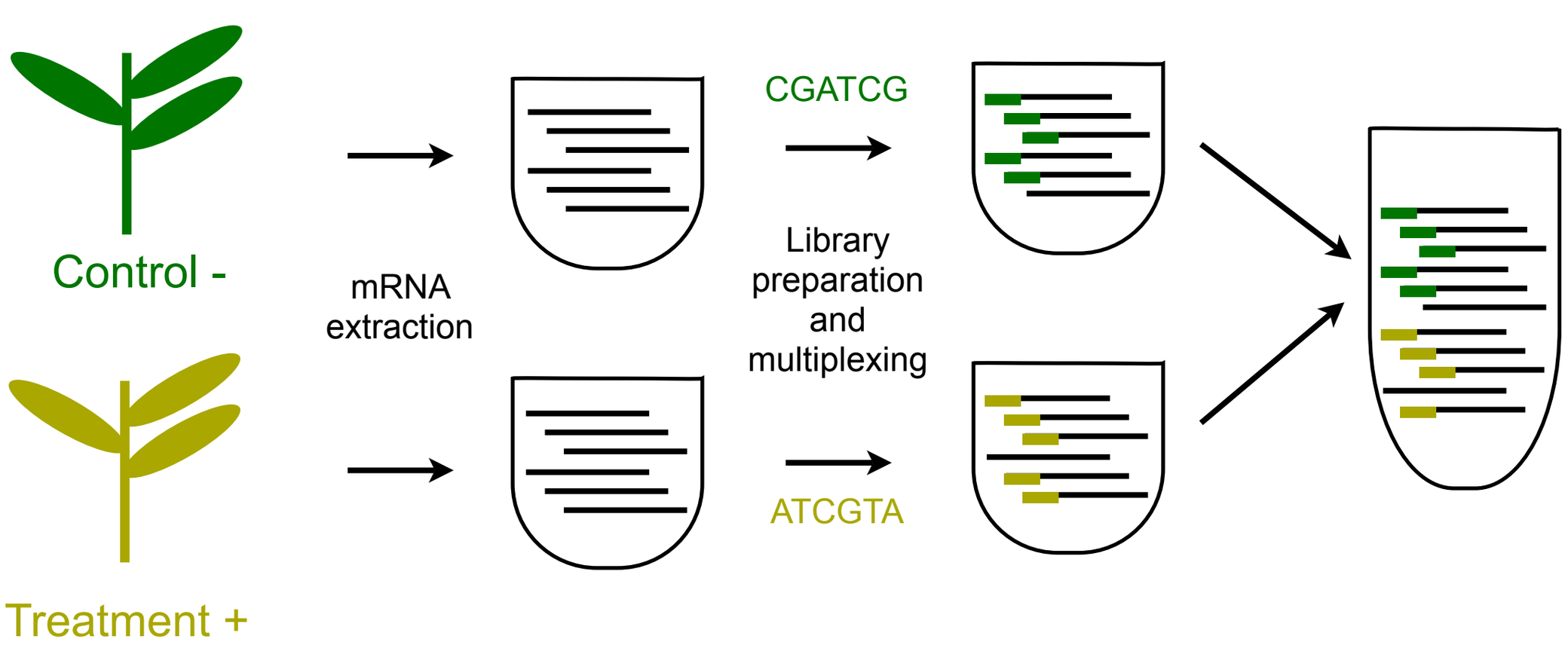
2) Use software to measure and correct the bias

(example: *seqbias* from R/Bioconductor, Jones DC *et al.* 2012)

Tissue selection and treatments:

Sequencing of multiple samples can be performed using **multiplexing.**

The multiplexing add a tag/**barcode** of 4-6 nucleotides during the library preparation to identify the sample. Common kits can add up to 96 different tags.



Selecting the right sequencing technology.

1) How many reads I need per sample ?

Enough to represent the mRNA population.

ENCODE consortium's *Standards, Guidelines and Best Practices for RNA-Seq*

"Experiments whose purpose is to evaluate the **similarity between the transcriptional profiles** of two polyA+ samples may require only modest depths of sequencing (e.g. 30M pair-end reads of length > 30NT, of which **20-25M are mappable to the genome** or known transcriptome."

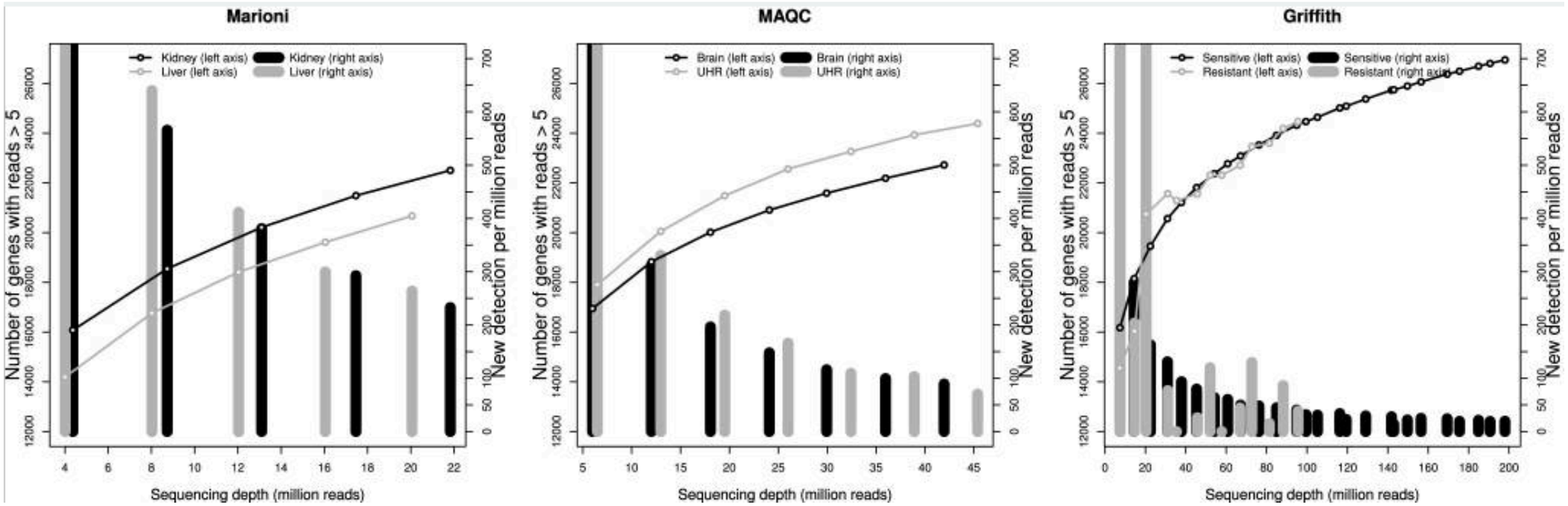
"Experiments whose purpose is discovery of **novel transcribed elements and strong quantification** of known transcript isoforms… a minimum depth of **100-200 M** 2 x 76 bp or longer reads is currently recommended."

http://www.rna-seqblog.com/information/how-many-reads-are-enough/

Selecting the right sequencing technology.

1) How many reads I need per sample ?

Enough to represent the mRNA population.



Tarazona S. *et al.* (2012) **Differential expression in RNA-seq: a matter of depth.** *Genome Res.*21:2213-23

Selecting the right sequencing technology.

1) How many reads I need per sample ?

	Run Time	Sequence Length	Reads/Run	Total nucleotides sequenced per run
Capillary Sequencing (ABI37000)	~2.5 h	800 bp	386	0.308 Mb
454 Pyrosequencing (GS FLX Titanium XL+)	~23 h	700 bp	1,000,000	700 Mb (0.7 Gb)
Illumina (HiSeq 2500)	264 h / 27 h (11 days)	2 x 100 bp 2 x 150 bp	2 x 3,000,000,000 2 x 600,000,000	600,000 / 120,000 Mb (600 / 120 Gb)
Illumina (MiSeq)	39 h	2 x 250 bp	2 x 17,000,000	8,500 Mb (8.5 Gb)
SOLID (5500xl system)	48 h (2 days)	75 bp	400,000,000	30,000 Mb (30 Gb)
Ion Torrent (Ion Proton I)	2 h	100 bp	100,000,000	10,000 Mb (10 Gb)
PacBio (PacBioRS)	1.5 h	~3,000 bp	25,000	100 Mb (0.1 Gb)

Selecting the right sequencing technology.

2) How long should be these reads?

Depending if you need to do a *de-novo* assembly, mapping with a reference with recent WGD or mapping with a reference without recent WGD.

- *de-novo* assembly
- Reference with recent WGD

} ▸ Longer is better (at least 100 bp)
▸ Pair ends recommended

- Reference without recent WGD ▸ Any size beyond 35 bp

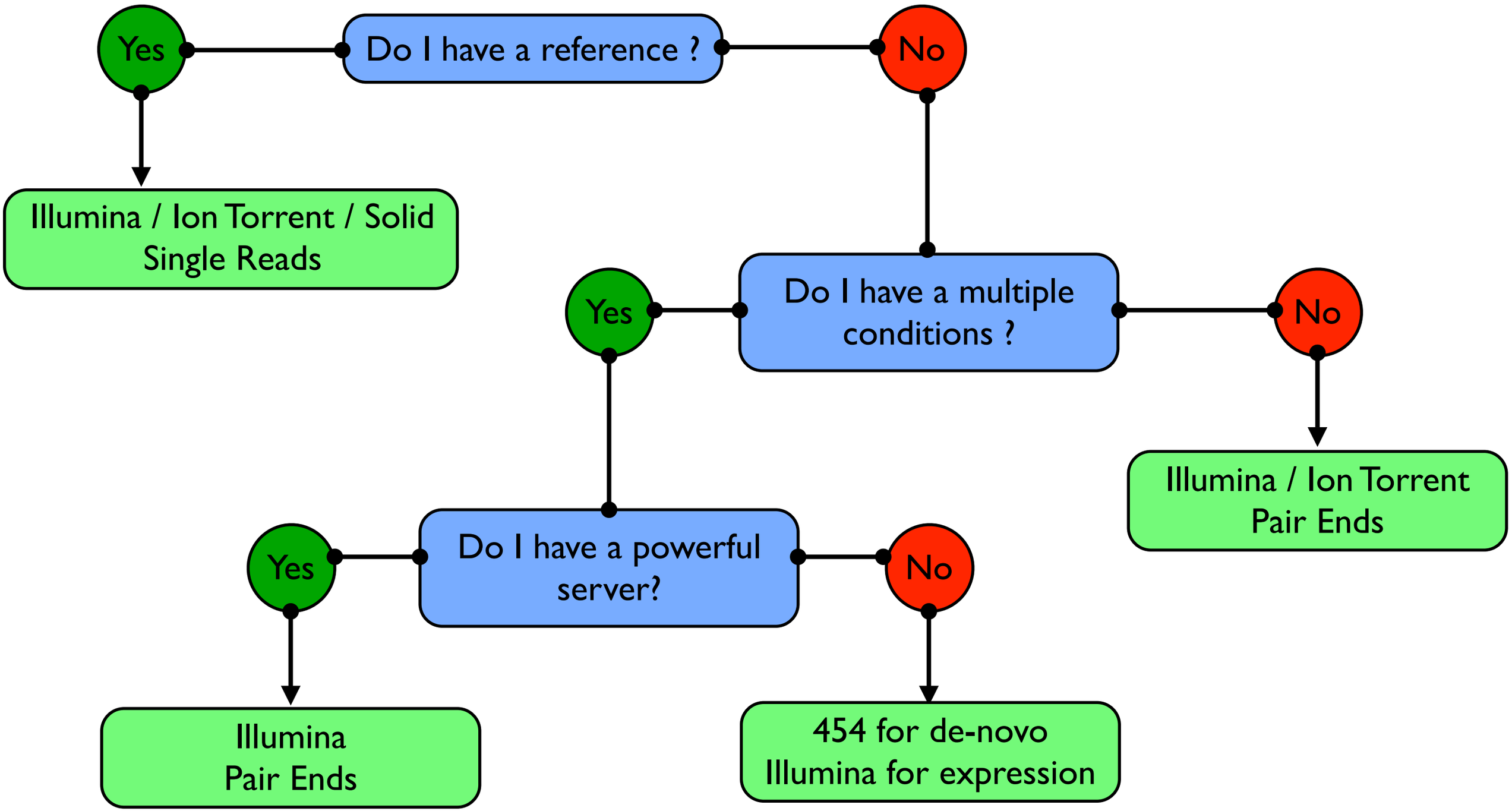
Selecting the right sequencing technology.

3) Do I have software/hardware limitations?

Some tools have some limitations to work with long reads.

Other tools doesnt work with color space reads produced by Solid.

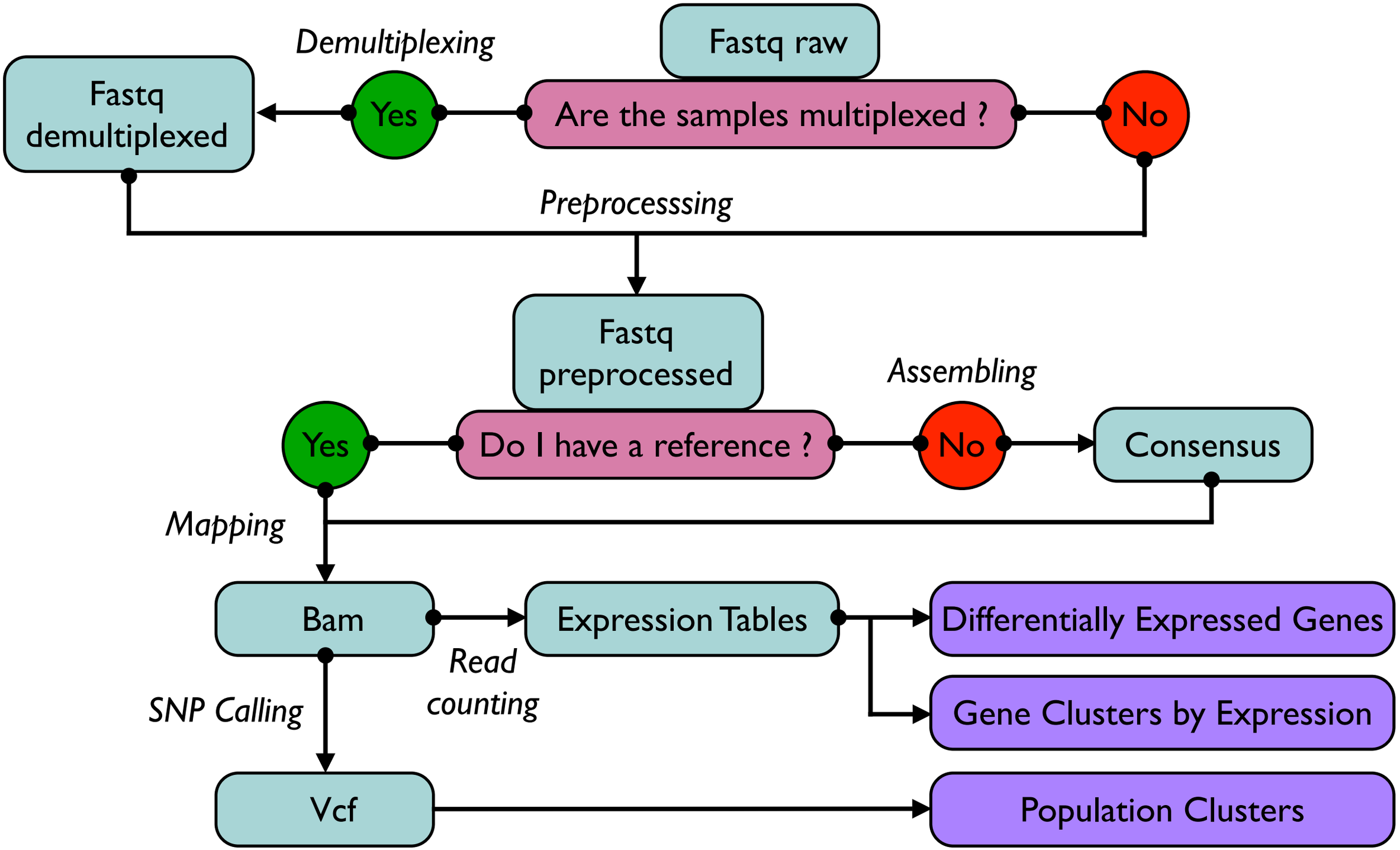
Selecting the right sequencing technology.



Lectures:

1. **Basics of the Next Generation Sequencing (NGS).**
   - 1.1. The sequencing revolutions.
   - 1.2. Strengths and weaknesses of the different technologies.
   - 1.3. Inputs and outputs.
2. **RNAseq experiment design.**
   - 2.1. Reference vs Non-reference.
   - 2.2. High heterozygosity and polyploid polyploid problem.
   - 2.3. Tissue selection and treatments.
   - 2.4. Sequencing technology.
3. **RNAseq expression analysis.**
   - 3.1. Reference preparation and read mapping.
   - 3.2. Gene expression.
   - 3.3. Analysis and visualization.
4. **Use of RNAseq reads for phylogeny and genetics.**
   - 4.1. Recovering full length mRNA: Reference guided assembly.
   - 4.2. Phylogeny though RNAseq: From gene tree to species tree.
   - 4.3. From reads to markers: SNP calling.
   - 4.4. Population genetics and NGS.

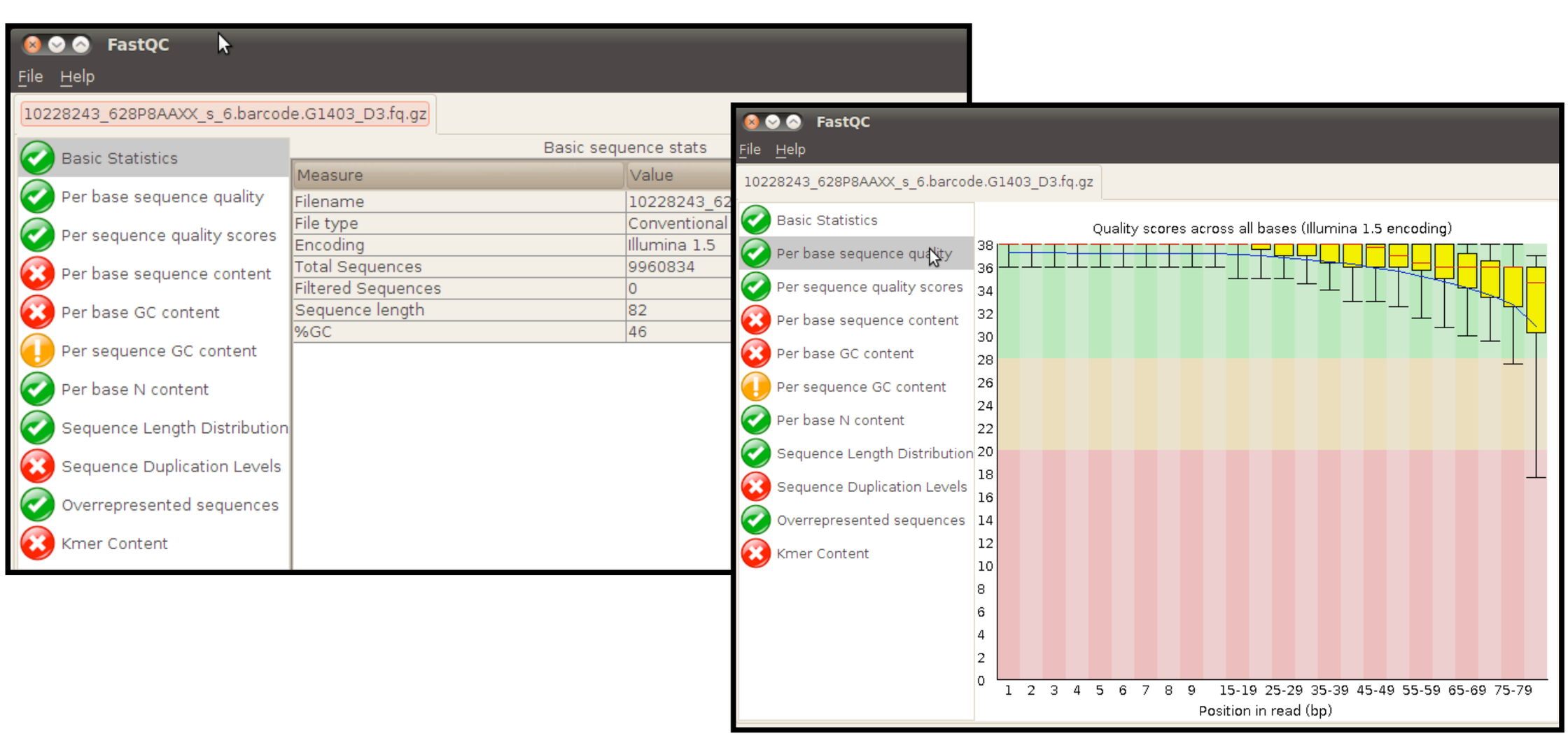
RNAseq Data Analysis Steps:



0. Read quality evaluation.

a) Length of the read.
b) Bases with qscore > 20 or 30.

- FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)



1. Read demultiplexing, filtering and trimming

- Separation of multiplexed samples.
- Adaptors removing.
- Low quality extreme trimming.
    - Minimum Q20.
    - Suggested Q30.
- Short sequence removing.
    - Suggested L50

1. Read demultiplexing, filtering and trimming

- Fastx-Toolkit (http://hannonlab.cshl.edu)
- Ea-Utils (http://code.google.com/p/ea-utils/)
- PrinSeq (http://edwards.sdsu.edu/cgi-bin/prinseq/prinseq.cgi)

Software	Multiplexing	Trimming/Filtering
Fastx-Toolkit	fastx_barcode_splitter	fastq_quality_filter
Ea-Utils	fastq-multx	fastq-mcf
PrinSeq	PrinSeq	PrinSeq

2.1 Read mapping

- Reference is a genome with gene model annotations.

↓

Tophat (Bowtie2)

- Reference is not a genome with gene model annotations.

↓

Bowtie2

BWA

GsMapper (454) ...

2.1 Read mapping

Software	Sequencing technology	Features	URL
bwa	Illumina	Mapping	http://bio-bwa.sourceforge.net/bwa.shtml
bowtie	Illumina, SOLID	Mapping	http://bowtie-bio.sourceforge.net/index.shtml
bowtie2	Illumina, 454 (fastq)	Mapping	http://bowtie-bio.sourceforge.net/bowtie2
novoalign	Illumina, SOLID	Mapping	http://www.novocraft.com/main/index.php
gsMapper	454 (sff)	Mapping, annotation	http://454.com/products/analysis-software/index.asp#reference-tabbing
SOAPaligner	Illumina	Mapping	http://soap.genomics.org.cn/soapaligner.html
TopHat (bowtie)	Illumina	Mapping, splicing	http://tophat.cbcb.umd.edu/index.html

2.1 Read mapping

- Reference is a genome with gene model annotations.

↓

1. **Fasta** file with genome sequence

2. **Gff** file with gene model annotations

```
##gff-version   3
##sequence-region   ctg123 1 1497228
ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN
ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
```

http://www.sequenceontology.org/resources/gff3.html

2.1 Read mapping

Gff file with gene model annotations:

Tabular format file with 9 columns:

- Column 1: "seqid"
- Column 2: "source"
- Column 3: "type"
- Column 4: "start coordinate" (1 based coordinate)
- Column 5: "end coordinate"
- Column 7: "strand"
- Column 8: "phase"
- Column 9: "attributes" composed by tags such as:
  ID; Name; Alias; Parent; Target; Gaps; Derives_from; Note;Dbxref

2.1 Read mapping

Before run the software some considerations:

- *How similar are the reads with the reference ?*

  By default only 1 SNP is allowed by the mapper.

- *Can one read map equally in different genes (gene duplications) ?*

  By default read mappers assign randomly sequences that map equally.

- *Are you mapping more than one species ?*

  Use combine sets to avoid multiple mappings of the same read.

2.1 Read mapping

Before run the software some considerations:

Example: Mapping reads from infected *N. benthamiana* leaves with *P. syringae*.

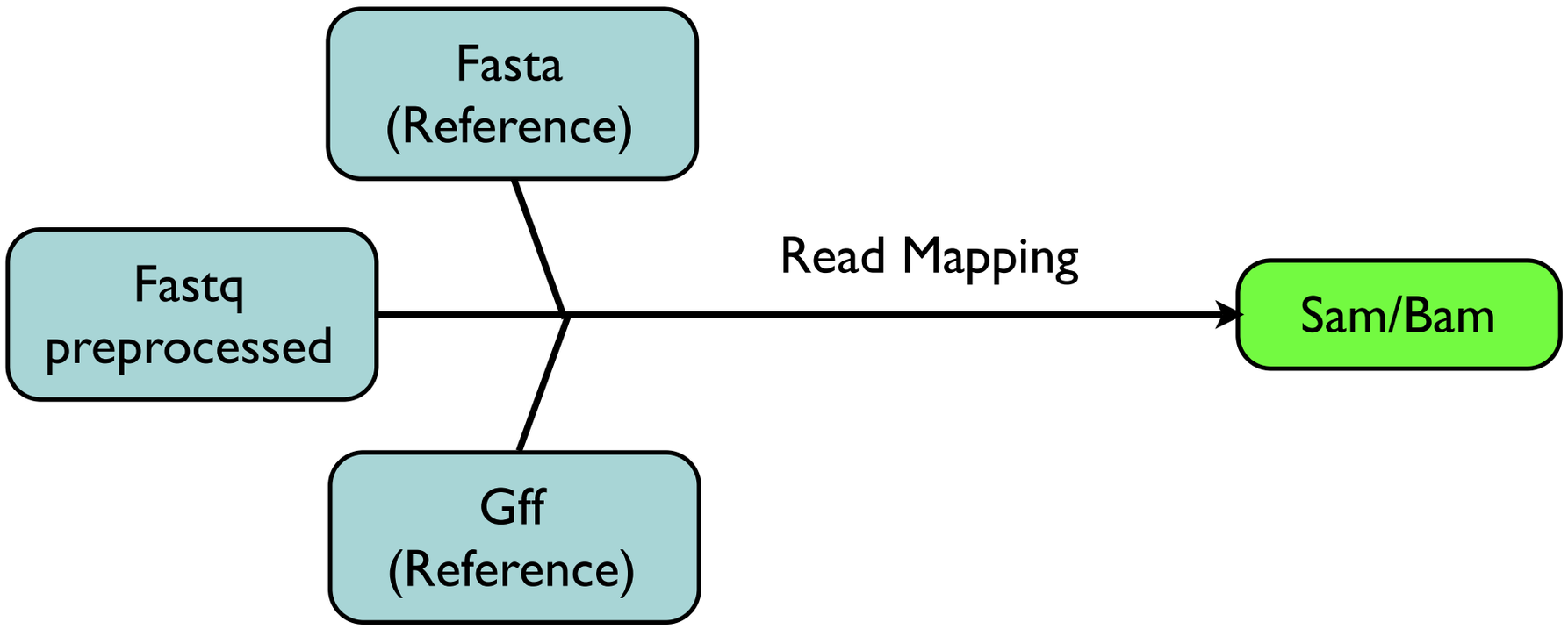
1. Join both datasets (fasta and gff3).
2. Map the reads using Bowtie2:

```
bowtie2-build -f merged_reference.fasta

bowtie2 -N 1 -M 0 -x merged_reference.fasta
seq1.fastq -S results.sam
```

- Exclude reads with more than M+1 matches
- Allow 1 mismatch in the seed

2.1 Read mapping



Note about the hardware and mapping software:

+ Bigger is the reference, more memory the programs needs
  (example: Bowtie2 ~2.1 Gb for human genome with 3 Gb)
+ Longer are the reads, more time the program needs for the mapping.

2.1 Read mapping

The standard output for read alignments is a **sam/bam** format. Sam format is a tabular delimited format with a **header lines starting with the character '@'** and **one line per alignment with 11 mandatory fields**.

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	`[0,2`^{16}`-1]`	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	`[0,2`^{29}`-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 segment
8	PNEXT	Int	`[0,2`^{29}`-1]`	Position of the mate/next segment
9	TLEN	Int	`[-2`^{29}`+1,2`^{29}`-1]`	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

http://samtools.sourceforge.net/SAM1.pdf

2.1 Read mapping

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002        aaaAGATAA*GGATA
+r003      gcctaAGCTAA
+r004                  ATAGCT..............TCAGC
-r003                        ttagctTAGGC
-r001/2                                   CAGCGCCAT
```

Alignment

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001  163  ref   7  30  8M2I4M1D3M  =  37   39  TTAGATAAAGGATACTG  *
r002    0  ref   9  30  3S6M1P1I4M  *   0    0  AAAAGATAAGGATA     *
r003    0  ref   9  30  5H6M        *   0    0  AGCTAA     *       NM:i:1
r004    0  ref  16  30  6M14N5M     *   0    0  ATAGCTTCAGC        *
r003   16  ref  29  30  6H5M        *   0    0  TAGGC      *       NM:i:0
r001   83  ref  37  30  9M          =   7  -39  CAGCGCCAT           *
```

Col	Field
1	QNAME
2	FLAG
3	RNAME
4	POS
5	MAPQ
6	CIGAR
7	RNEXT
8	PNEXT
9	TLEN
10	SEQ
11	QUAL

Sam file

http://samtools.sourceforge.net/SAM1.pdf

2.1 Read mapping

Cigar String

```
8M2I4M1D3M
3S6M1P1I4M
5H6M
6M14N5M
6H5M
9M
```

6 CIGAR

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

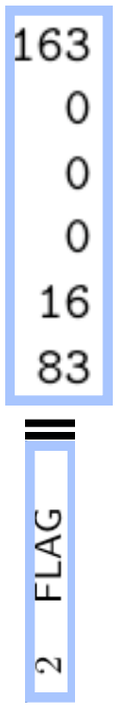
8M2I4M1D3M ⟶

8M	8 aligned nucleotides (match or mismatch)
2I	2 insertions to the reference
4M	4 aligned nucleotides (match or mismatch)
1D	1 deletion from the reference
3M	3 aligned nucleotides (match or mismatch)

http://samtools.sourceforge.net/SAM1.pdf

2.1 Read mapping

Flags String



Bit	Description
0x1	template having multiple segments in sequencing
0x2	each segment properly aligned according to the aligner
0x4	segment unmapped
0x8	next segment in the template unmapped
0x10	SEQ being reverse complemented
0x20	SEQ of the next segment in the template being reversed
0x40	the first segment in the template
0x80	the last segment in the template
0x100	secondary alignment
0x200	not passing quality controls
0x400	PCR or optical duplicate

- Flag = 4 means 0x4 read unmapped
- Flag = 16 means 0x10 read reverse strand
- Flag = 83 means 0x1 read paired, 0x2 read mapped proper pair, 0x10 read reverse strand and 0x40 first in pair

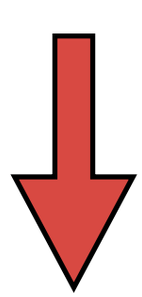
http://samtools.sourceforge.net/SAM1.pdf

2.1 Read mapping

Evaluation of the mapping results:

Software:

Samtools (http://samtools.sourceforge.net/samtools.shtml)



Count all the reads: `samtools view -c file.sam/file.bam`

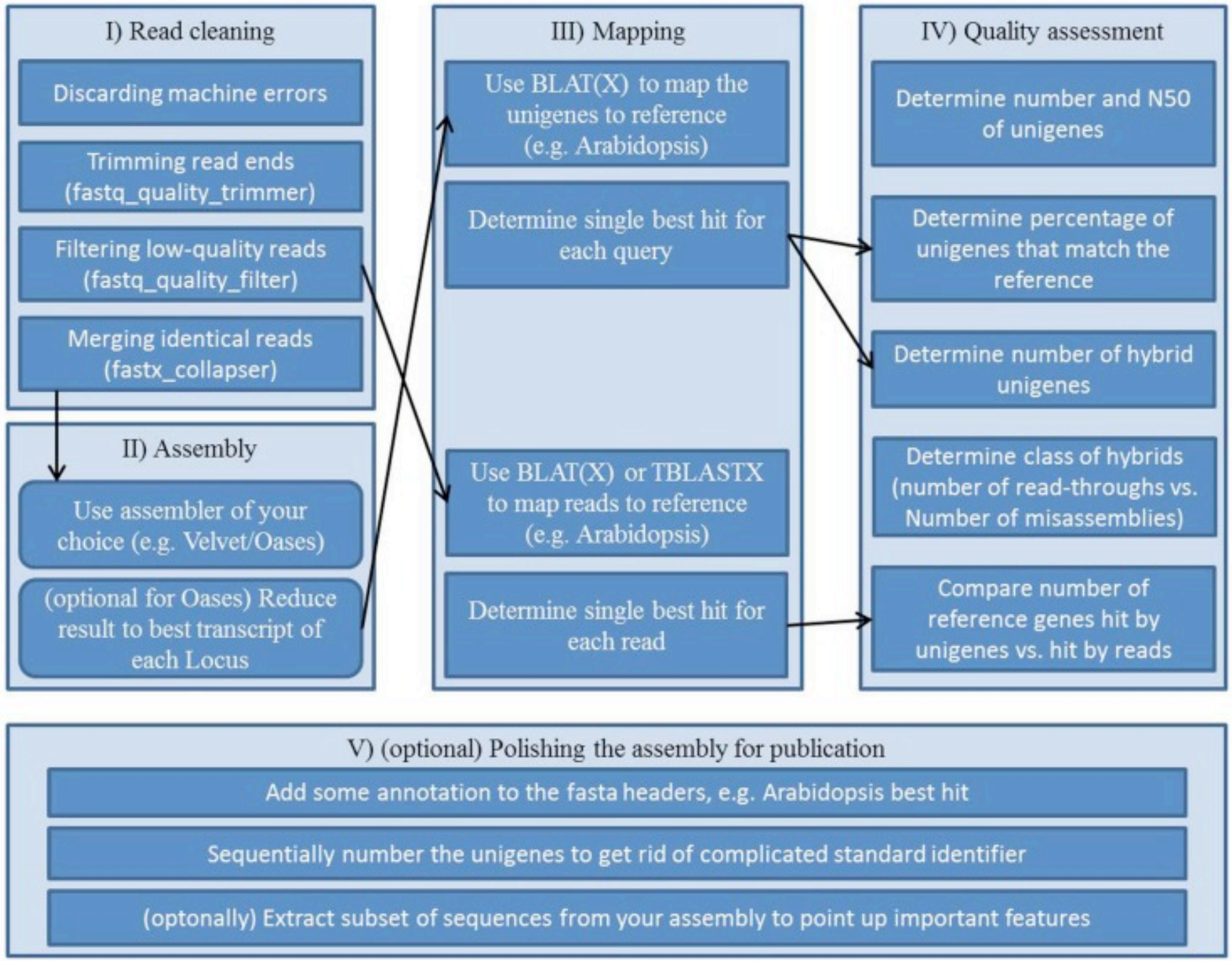
Count mapped reads: `samtools view -c -F 4 file.sam/file.bam`

Sam to Bam: `samtools view -Scb -o file.bam file.sam`

Sam to Bam (mapped): `samtools view -Scb -F 4 -o file.bam file.sam`

2.2 Transcriptome de-novo assembly

Workflow scheme for a transcriptome assembly



Schliesky S, Gowik U, Weber AP, Bräutigam A. (2012) **RNA-Seq Assembly – Are We There Yet?** *Front Plant Sci doi: 10.3389/fpls.2012.00220*

2.2 Transcriptome de-novo assembly

Software	Sequencing technology	Type	Features	URL
MIRA	Sanger, 454	Overlap-layout-consensus	Highly configurable	http://sourceforge.net/apps/mediawiki/mira-assembler
gsAssembler	Sanger, 454	Overlap-layout-consensus	Splicings	http://454.com/products/analysis-software/index.asp
iAssembler	Sanger, 454	Overlap-layout-consensus	Improves MIRA	http://bioinfo.bti.cornell.edu/tool/iAssembler
Trans-ABySS*	454 or Illumina	Bruijn graph	Splicings, Gene fusions	http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss
SOAPdenovo-trans*	454 or Illumina	Bruijn graph	Fastest	http://soap.genomics.org.cn/SOAPdenovo-Trans.html
Velvet/Oases	454 or Illumina or SOLiD	Bruijn graph	SOLiD	http://www.ebi.ac.uk/~zerbino/oases/
Trinity*	454 or Illumina	Bruijn graph	Downstream expression	http://trinityrnaseq.sourceforge.net/

* Comparisons in the Article: Vijay N. *et al* (2012) *Molecular Ecology* DOI: 10.1111/mec.12014

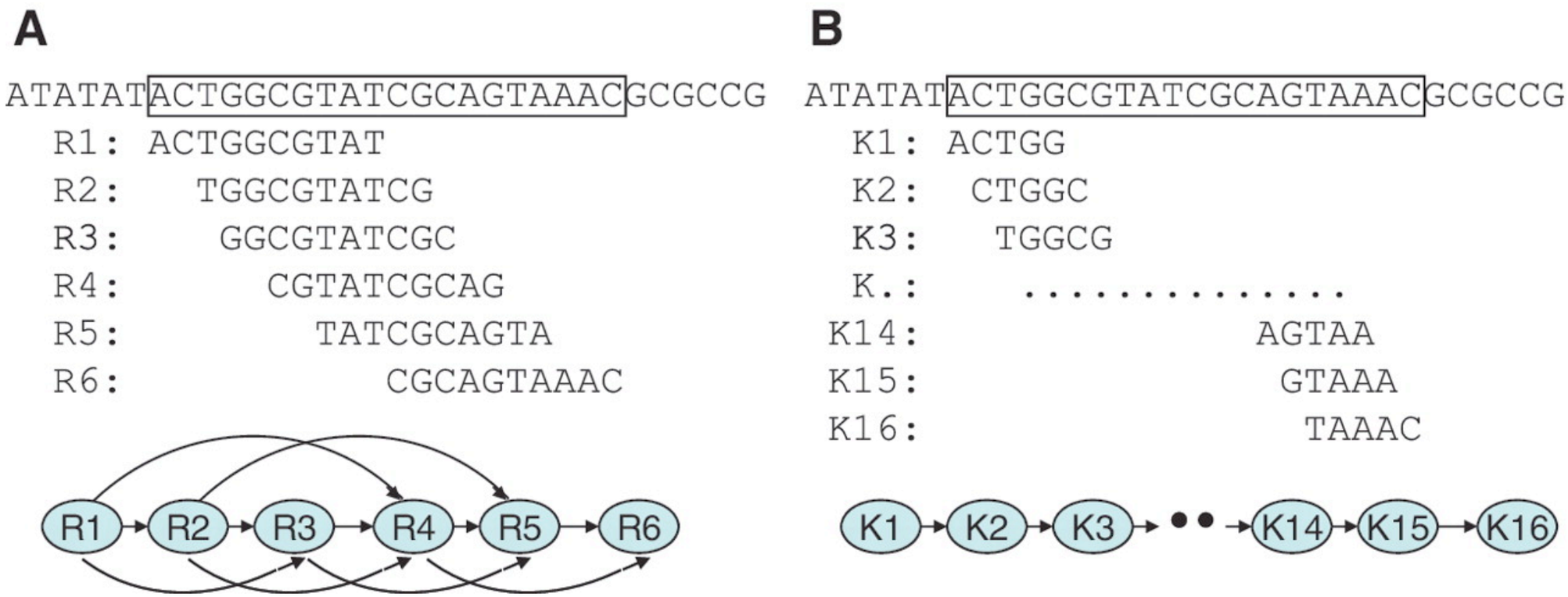
2.2 Transcriptome de-novo assembly

Overlap-layout-consensus

More memory,
percentage of identity configurable

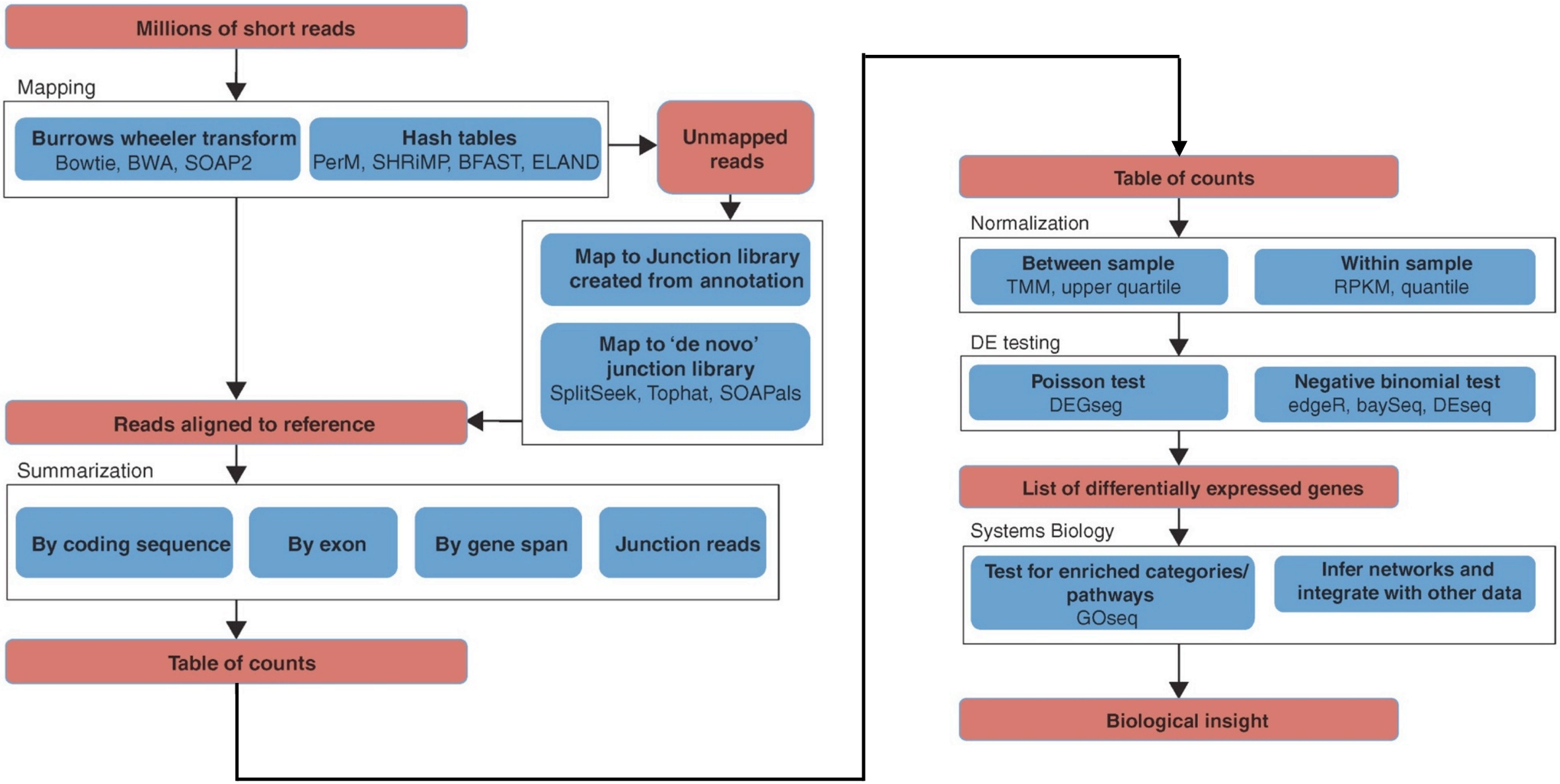
Bruijn graph

Faster,
less memory intensive



Li Z. *et al.* (2011) **Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph**
Brief. Funct. Genomics 11: 25-37. doi: 10.1093/bfgp/elr035

3.1 Gene Expression for RNAseq



Alicia Oshlack A, Robinson MD, and YoungMD, Genome Biology 2010, 11:220

3.1 Gene Expression for RNAseq

Gene expression for RNAseq analysis is based in how many reads map to an specific gene. For comparison purposes the counts needs to be normalized. There are different methodologies.

- **RPKM** (Mortazavi et al. 2008): Reads per Kilobase of Exon perMillion of Mapped reads.
- **Upper-quartile** (Bullard et al. 2010): Counts are divided per upper quartile of counts with at least one read.
- **TMM** (Robinson and Oshlack, 2010): Trimmed Means of M values (EdgeR).
- **FPKM** (Trapnell et al. 2010): Fragment per Kilobase of exon per Million of Mapped fragments (Cufflinks).

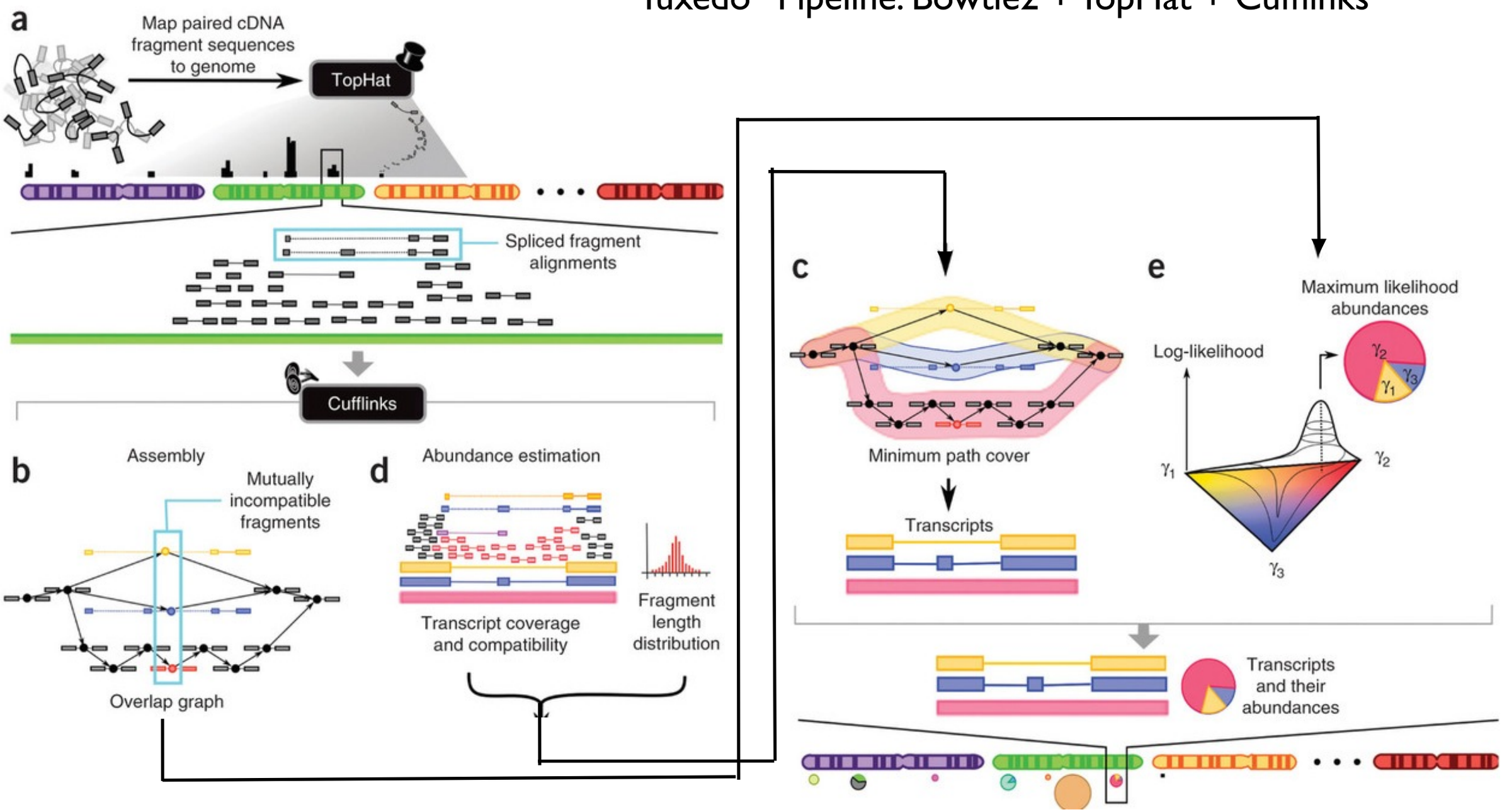
3.1 Gene Expression for RNAseq

Software	Normalization	Notes	URL
ERANGE	RPKM	Python	http://woldlab.caltech.edu/wiki/RNASeq
Scripture	RPKM	Java	http://www.broadinstitute.org/software/scripture
BitSeq*	RPKM	R/Bioconductor, Calculate DE	http://www.bioconductor.org/packages/2.12/bioc/html/BitSeq.html
EdgeR	TMM	R/Bioconductor, Calculate DE	http://www.bioconductor.org/packages/2.11/bioc/html/edgeR.html
Cufflinks*	FPKM	Isoforms, Calculate DE	http://cufflinks.cbcb.umd.edu/
MMSEQ*	FPKM	Isoforms, Haplotypes	http://bgx.org.uk/software/mmseq.html
RSEM*	FPKM	Calculate DE (EBSeq)	http://deweylab.biostat.wisc.edu/rsem/README.html

* Comparisons in the Article: Glaus P. *et al* (2012) *Bioinformatics* 28:1721-1728 doi:10.1093/bioinformatics/bts260

3.1 Gene Expression for RNAseq

"Tuxedo" Pipeline: Bowtie2 + TopHat + Cufflinks



Trapnell C, et al. 2012 Nature Biotechnology doi:10.1038/nbt.1621

3.2 Differential Gene Expression

Statistical test to evaluate if one gene has an differential expression between two or more conditions. These test can be based in different methodologies.

- **Negative binomial distribution** (DESeq, CuffLinks).
- **Bayesian methods for the negative binomial distribution** (EdgeR, BaySeq, BitSeq).
- **Non-parametric:** models the noise distribution of count changes by contrasting fold-change differences (M) and absolute expression differences (D) (NOISeq).

3.2 Differential Gene Expression

Software	Normalization	Need Replicas	Input	URL
EdgeR	Library Size / TMM	Yes	Raw Counts	http://www.bioconductor.org/packages/2.11/bioc/html/edgeR.html
DESeq	Library Size	No	Raw Counts	http://bioconductor.org/packages/release/bioc/html/DESeq.html
baySeq	Library Size	Yes	Raw Counts	http://www.bioconductor.org/packages/2.11/bioc/html/baySeq.html
NOISeq	Library Size / RPKM / UpperQ	No	Raw or Normalized Counts	http://bioinfo.cipf.es/noiseq/doku.php?id=start

Tarazona S. *et al.* (2012) **Differential expression in RNA-seq: a matter of depth.** *Genome Res.*21:2213-23

Exercises:

1. **Basic Linux commands.**
2. **Sequencing evaluation.**
3. **Simple read mapping.**
4. **Simple de-novo assembly.**
5. **Basic R commands**
6. **Functional annotation.**
7. **Differential gene expression.**
8. **Cluster analysis for gene expression.**
9. **Selecting genes for phylogeny.**
10. **SNP calling and filtering.**
11. **Analysis of the population structure.**

Exercises:

1. **Basic Linux commands.**
2. **Sequencing evaluation.**
3. **Simple read mapping.**
4. **Simple de-novo assembly.**
5. **Basic R commands**
6. **Functional annotation.**
7. **Differential gene expression.**
8. **Cluster analysis for gene expression.**
9. **Selecting genes for phylogeny.**
10. **SNP calling and filtering.**
11. **Analysis of the population structure.**

Exercises:

1. **Basic Linux commands.**
2. **Sequencing evaluation.**
3. **Simple read mapping.**
4. **Simple de-novo assembly.**
5. **Basic R commands**
6. **Functional annotation.**
7. **Differential gene expression.**
8. **Cluster analysis for gene expression.**
9. **Selecting genes for phylogeny.**
10. **SNP calling and filtering.**
11. **Analysis of the population structure.**

Exercises:

1. **Basic Linux commands.**
2. **Sequencing evaluation.**
3. **Simple read mapping.**
4. **Simple de-novo assembly.**
5. **Basic R commands**
6. **Functional annotation.**
7. **Differential gene expression.**
8. **Cluster analysis for gene expression.**
9. **Selecting genes for phylogeny.**
10. **SNP calling and filtering.**
11. **Analysis of the population structure.**

Exercises:

1. **Basic Linux commands.**
2. **Sequencing evaluation.**
3. **Simple read mapping.**
4. **Simple de-novo assembly.**
5. **Basic R commands**
6. **Functional annotation.**
7. **Differential gene expression.**
8. **Cluster analysis for gene expression.**
9. **Selecting genes for phylogeny.**
10. **SNP calling and filtering.**
11. **Analysis of the population structure.**

3.3 Explorative Data Mining Methods

Data mining is the process that attempts to discover **patterns in large data sets**. Data mining involves six common classes of tasks:

- Anomaly detection (Outlier/change/deviation detection) - Search of unusual data records
- Association rule learning (Dependency modeling) - Search of relationships between variables.
- Clustering - Discover groups and structures by similarity.
- Classification - Apply known structure to the new data
- Regression - Modeling to find the least error
- Summarization – Including visualization and report generation.

http://en.wikipedia.org/wiki/Data_mining

3.3 Explorative Data Mining Methods

For **gene expression** there are some common tasks and associated methods for the **data mining**:

- Clustering of the expression values and principal component analysis to reduce the variables.
- Classification using Gene Ontology terms and metabolic annotations
- Summarization visualizing the expression data through heat maps.

3.3 Cluster Analysis and Visualization

Cluster analysis or clustering is the task of **assigning a set of objects into groups** (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. Clustering is a main task of **explorative data mining**. The most used **clustering algorithm** for gene expression are:

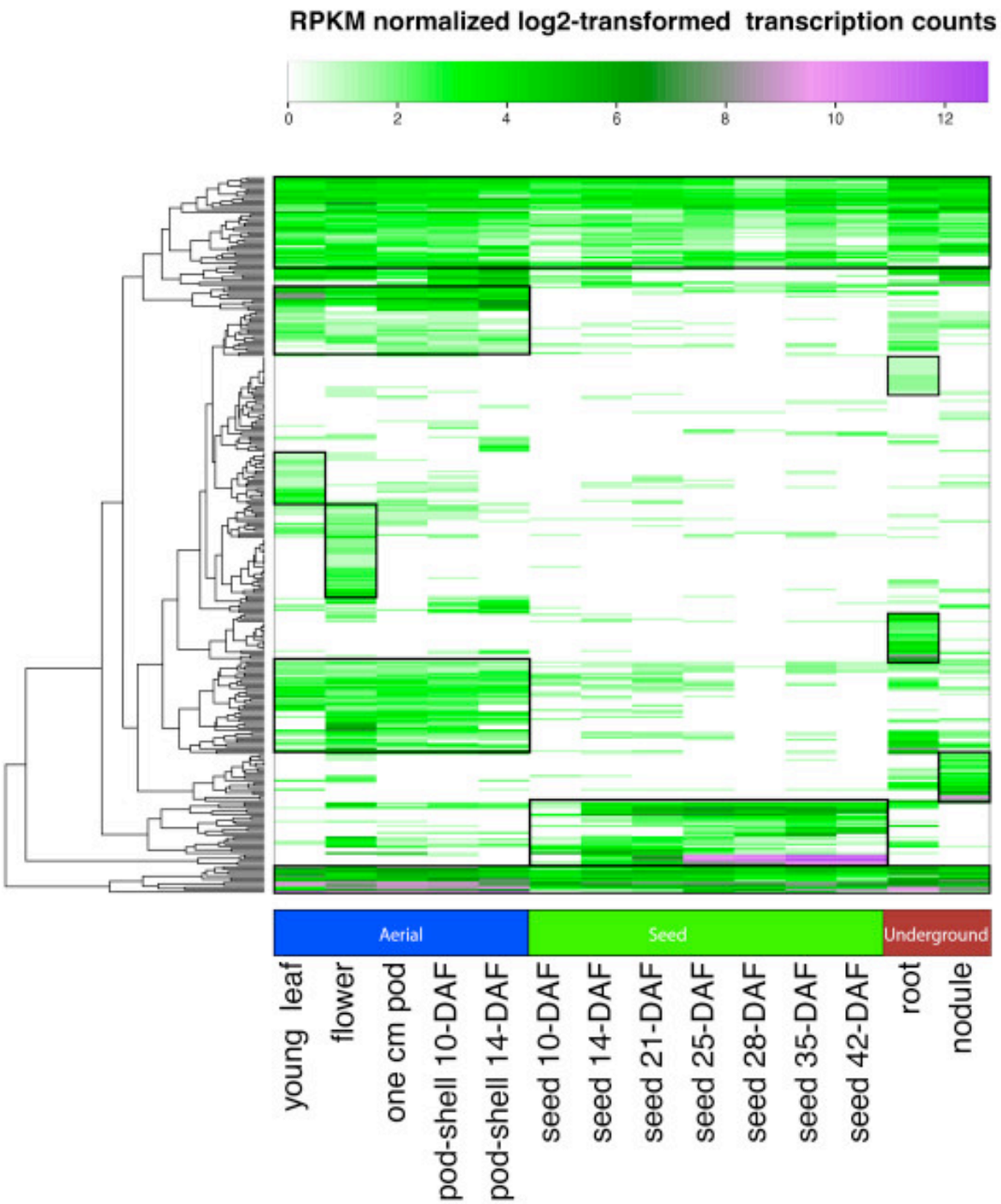
- **Hierarchical clustering (HCL)**, where the distance between elements is used to build the clusters.
- **K-means clustering (KMC)**, where clusters are represented by a vector. The number of clusters is fixed and the elements are assigned based in its distance to the vector.

http://en.wikipedia.org/wiki/Cluster_analysis

3.3 Cluster Analysis and Visualization

Software	Clustering Algorithm	URL
MeV	HC, KMC, visualization	http://www.tm4.org/mev/about
Stats (R package)	HC (hclust() function) KMC (kmeans() function) Visualization (gplots() function)	http://stat.ethz.ch/R-manual/R-patched/library/stats/html/stats-package.html
GENE-E	HC, visualization	http://www.broadinstitute.org/cancer/software/GENE-E/

3.3 Cluster Analysis and Visualization



Severin AJ et al., 2010 BMC Plant Biology, 10:160

3.3 Classification Analysis and Visualization

One of the most common classification data mining method is the use of gene annotations such as GO terms or metabolic annotations. These methodologies compare two groups between them to find if there are term more represented in one group than in other. Some examples are:

- **Gene Set Enrichment Analysis (GSEA)**, computational method that determines whether an a priori defined set of genes shows statistically significant.
- **Profile comparisons**, each group defines a profile based in the annotation groups (generally GO terms). Profiles are compared to find if they are significantly different.

3.3 Classification Analysis and Visualization

One of the most common classification data mining method is the use of gene annotations such as GO terms or metabolic annotations. These methodologies compare two groups between them to find if there are term more represented in one group than in other. Some examples are:

- **Gene Set Enrichment Analysis (GSEA)**, computational method that determines whether an a priori defined set of genes shows statistically significant.
- **Profile comparisons**, each group defines a profile based in the annotation groups (generally GO terms). Profiles are compared to find if they are significantly different.

3.3 Classification Analysis and Visualization

Gene ontologies:

Structured controlled vocabularies (ontologies) that describe **gene products** in terms of their associated

biological processes,

cellular components and

molecular functions

in a species-independent manner

http://www.geneontology.org/GO.doc.shtml

3.3 Classification Analysis and Visualization

Biological processes,
Recognized series of events or molecular functions. A process is a collection of molecular events with a defined beginning and end.

Cellular components,
Describes locations, at the levels of subcellular structures and macromolecular complexes.

Molecular functions
Describes activities, such as catalytic or binding activities, that occur at the molecular level.

http://www.geneontology.org/GO.doc.shtml

3.3 Classification Analysis and Visualization

Bioconductor Packages for GO Terms:

GO.db A set of annotation maps describing the entire Gene Ontology

Gostats Tools for manipulating GO and microarrays

GOSim functional similarities between GO terms and gene products

GOProfiles Statistical analysis of functional profiles

TopGO Enrichment analysis for Gene Ontology

http://www.geneontology.org/GO.doc.shtml

Exercises:

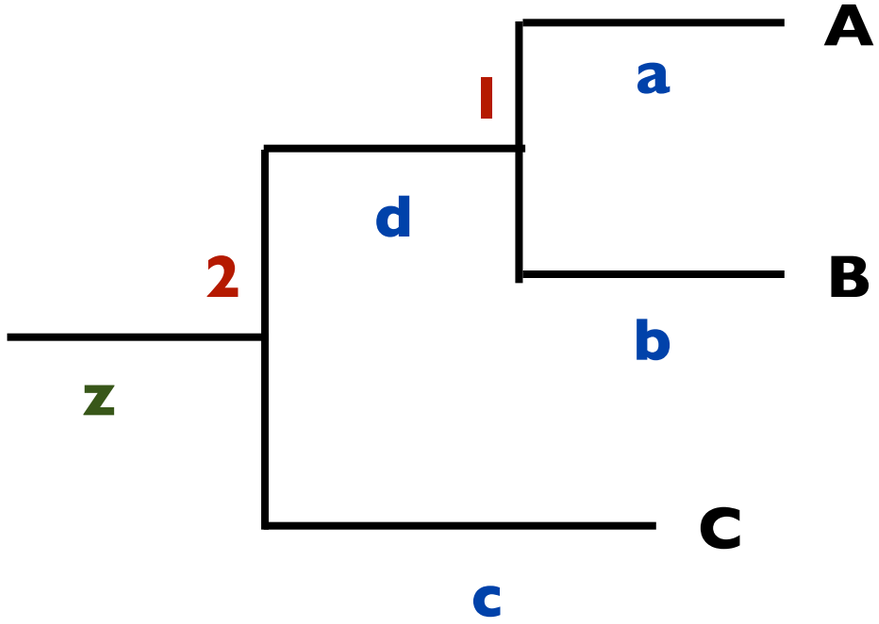
1. **Basic Linux commands.**
2. **Sequencing evaluation.**
3. **Simple read mapping.**
4. **Simple de-novo assembly.**
5. **Basic R commands**
6. **Functional annotation.**
7. **Differential gene expression.**
8. **Cluster analysis for gene expression.**
9. **Selecting genes for phylogeny.**
10. **SNP calling and filtering.**
11. **Analysis of the population structure.**

Lectures:

1. **Basics of the Next Generation Sequencing (NGS).**
   - 1.1. The sequencing revolutions.
   - 1.2. Strengths and weaknesses of the different technologies.
   - 1.3. Inputs and outputs.
2. **RNAseq experiment design.**
   - 2.1. Reference vs Non-reference.
   - 2.2. High heterozygosity and polyploid polyploid problem.
   - 2.3. Tissue selection and treatments.
   - 2.4. Sequencing technology.
3. **RNAseq expression analysis.**
   - 3.1. Reference preparation and read mapping.
   - 3.2. Gene expression.
   - 3.3. Analysis and visualization.
4. **Use of RNAseq reads for phylogeny and genetics.**
   - 4.1. Recovering full length mRNA: Reference guided assembly.
   - 4.2. Phylogeny though RNAseq: From gene tree to species tree.
   - 4.3. From reads to markers: SNP calling.
   - 4.4. Population genetics and NGS.

4.0 What is a phylogenetic tree ?

A phylogenetic tree is a **diagram** that shows the **evolutionary relationships** among **genes** and **organism**. A phylogenetic tree has different parts:



A and **B** are ***leaves***

C is an **external node**

1 and **2** are internal ***nodes***

a, **b**, **c** and **d** are ***branches***

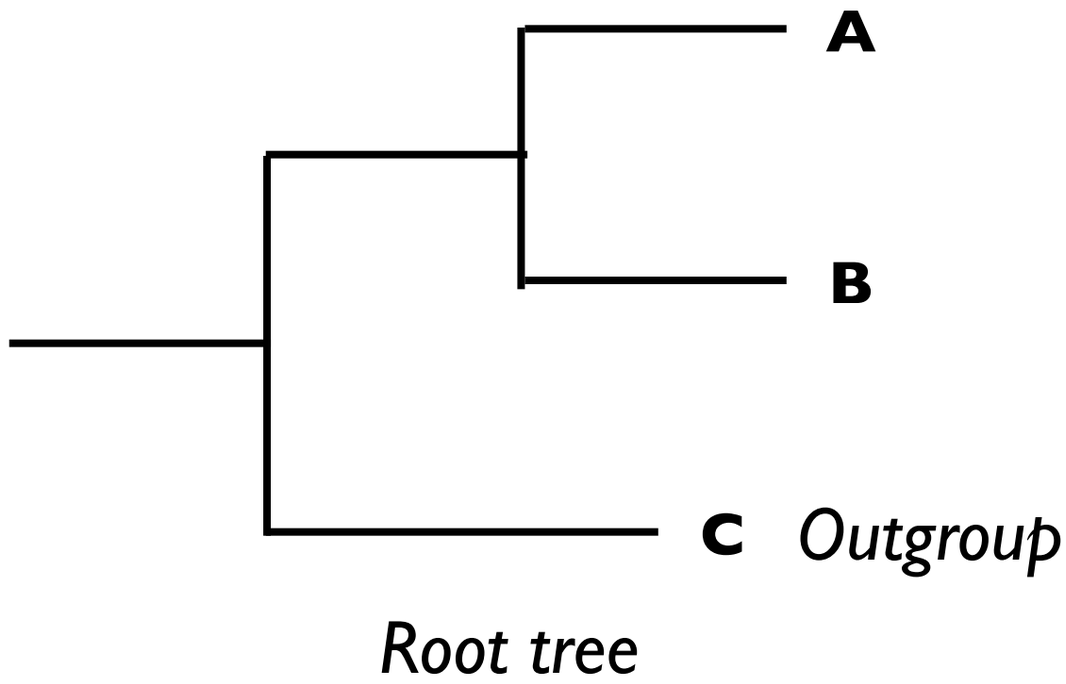
z is the ***root***

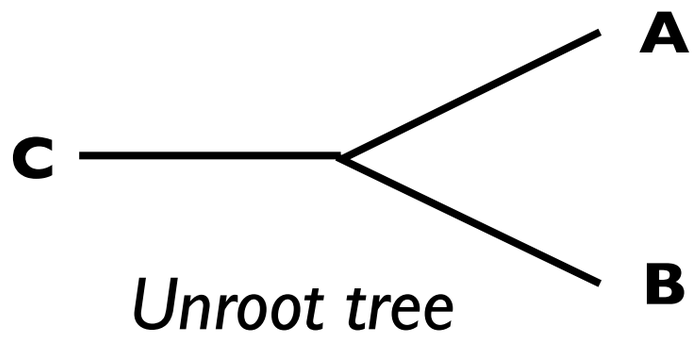
External nodes and **leaves** represents **extant** and **existing taxa** (operational taxonomic units, OTU).

Internal nodes may be called hypothetical taxonomic units (HTU)

The Phylogenetic Handbook, 2nd ed. Cambridge

4.0 What is a phylogenetic tree ?





Unroot tree: Position of each taxa **relative** to each other.

Root tree: Position of the taxa to a **common ancestor.**

A tree can be rooted if at least one of the OTU is an **outgroup**

The Phylogenetic Handbook, 2nd ed. Cambridge

4.0 What is a phylogenetic tree ?

There are two different methods to construct phylogenetic trees:

- **Character state**, uses discrete characters such as morphologic data or sequence data.
- **Distance matrix**, uses a measure of the dissimilarity of two OTUs to produce a *pairwise distance matrix*.

↓

They use a **evolutionary model** to **correct** multiple hits.

↓

Tree evaluation by **optimality search criterion.**

The Phylogenetic Handbook, 2nd ed. Cambridge

4.0 What is a phylogenetic tree ?

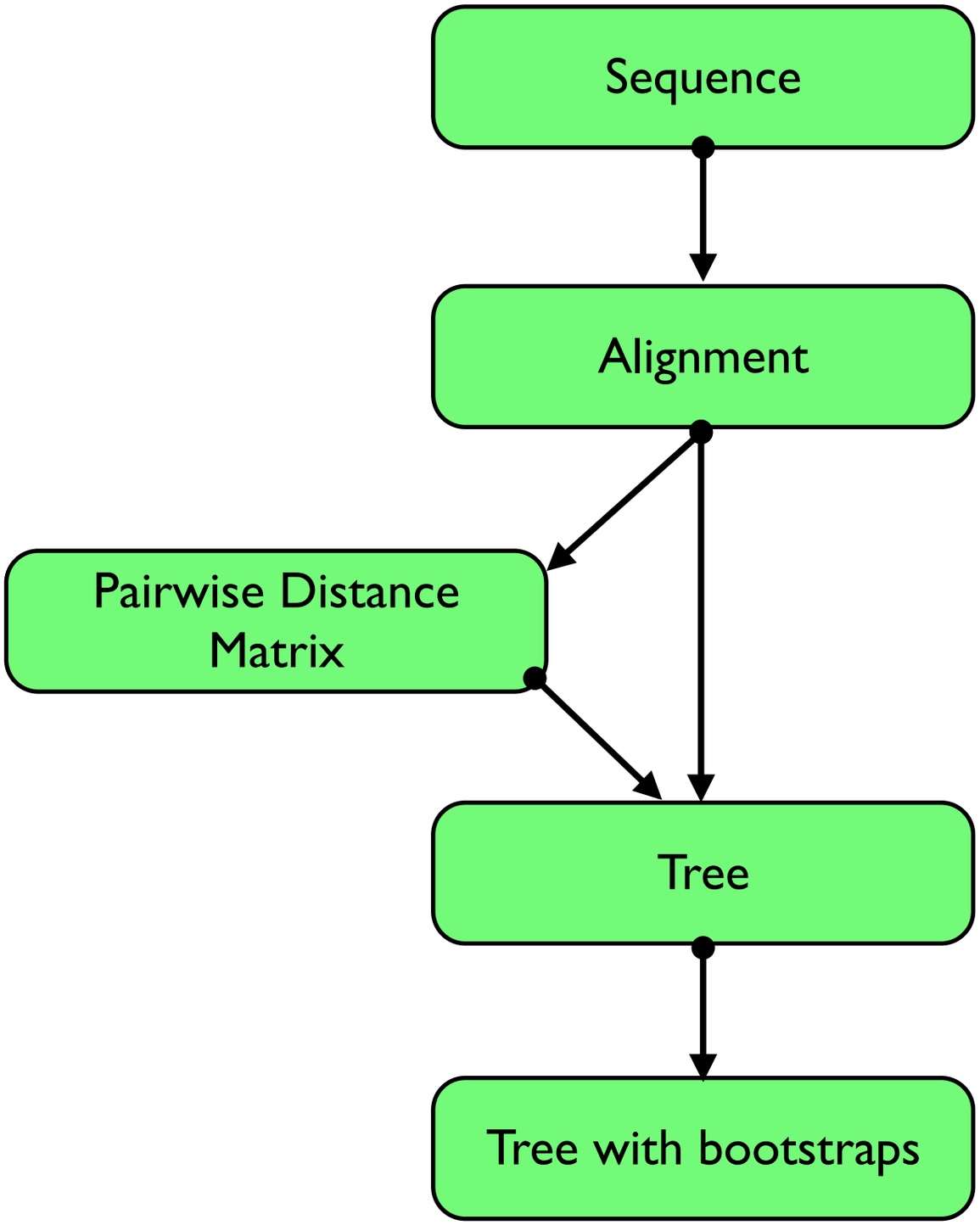
Classification of Phylogenetic Analysis		
	Optimality Search Criterion	Clustering
Character State	Maximum Parsimony (MP)	
	Maximum Likelihood (ML)	
	Bayesian Inference (BI)	
Distance Matrix	Fitch-Margoliash (FM)	UPGMA
	Minimum Evolution (ME)	Neighbor-joining (NJ)

The Phylogenetic Handbook, 2nd ed. Cambridge

4.0 Phylogenetic Software

Software	Methods	URL
Phylip	UPGMA, NJ, Fitch, ML and MP	http://evolution.genetics.washington.edu/phylip/general.html
MEGA4	NJ, ME, ML and MP	http://www.megasoftware.net
PAUP	ML and MP	http://paup.csit.fsu.edu/about.html
FastTree	NJ, ME and ML	http://www.microbesonline.org/fasttree/
PhyML	ML	http://www.atgc-montpellier.fr/phyml/
RAxML	ML	http://sco.h-its.org/exelixis/software.html
MrBayes	BI	http://mrbayes.sourceforge.net/index.php

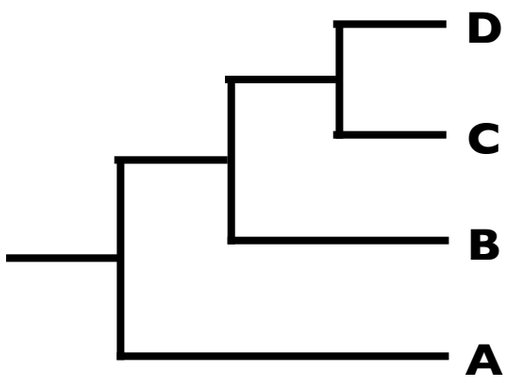
4.0 Steps to Infer a Phylogenetic Tree



4.0 Why a phylogenetic tree needs a bootstrap analysis ?

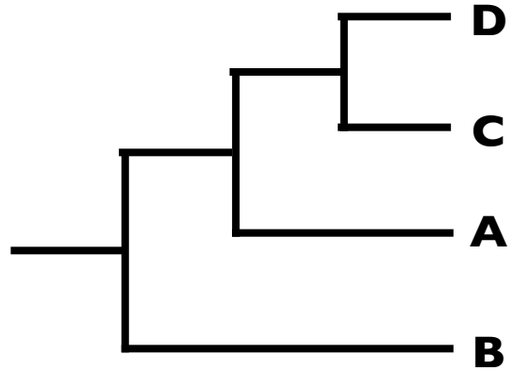
Bootstrap analysis and ***Jackknifing*** are the methodologies used to evaluate the **reliability of the inferred tree**. They can be applied to all tree construction. Under normal circumstances, branches supported by less than **70% of the bootstrap** should be treated with caution.

ATGCGTCGTTAG - **A**
ATGTGTCGTTAG - **B**
ATGTGACGTTAG - **C**
ATGTGACTTTAG - **D**



Original Tree

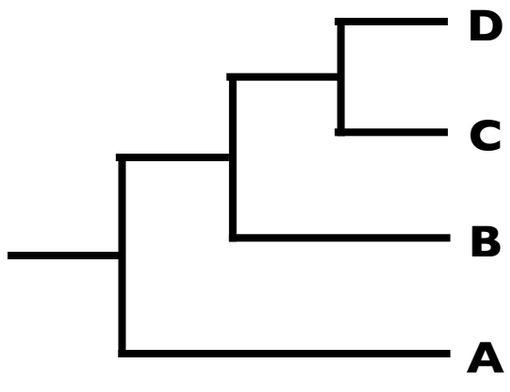
ATGCGTCGTTAG - **A**
AGGTGTCGTTAG - **B**
ATGTGACGTTAG - **C**
ATGTGACTTTAG - **D**



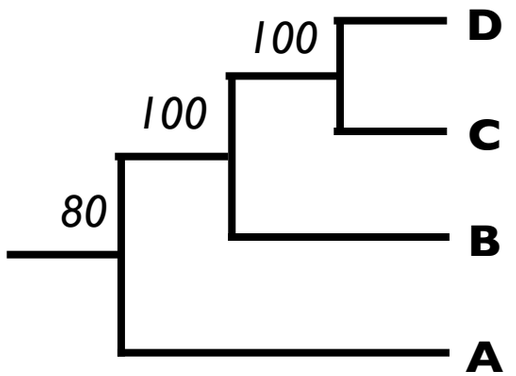
Bootstrap 1

....

ATGCGTCGTGAG - **A**
ATGTGTCGTTAG - **B**
ATGTGACGTTAG - **C**
ATGTGACTTTAG - **D**

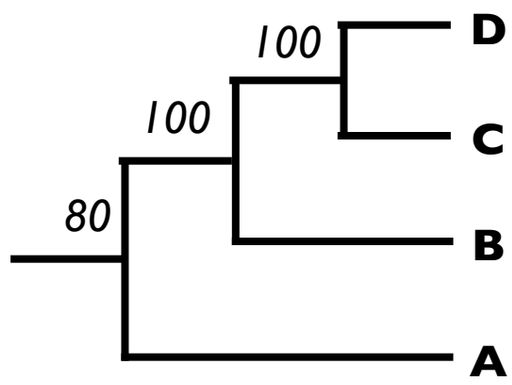


Bootstrap 100



4.0 Why a phylogenetic tree needs a bootstrap analysis ?

So **bootstrap values** are like the error bars for a phylogenetic tree. A tree without bootstrapping values has an incomplete information about how reliable are each of the branches.



4.0 Use of RNAseq for phylogenetic analysis

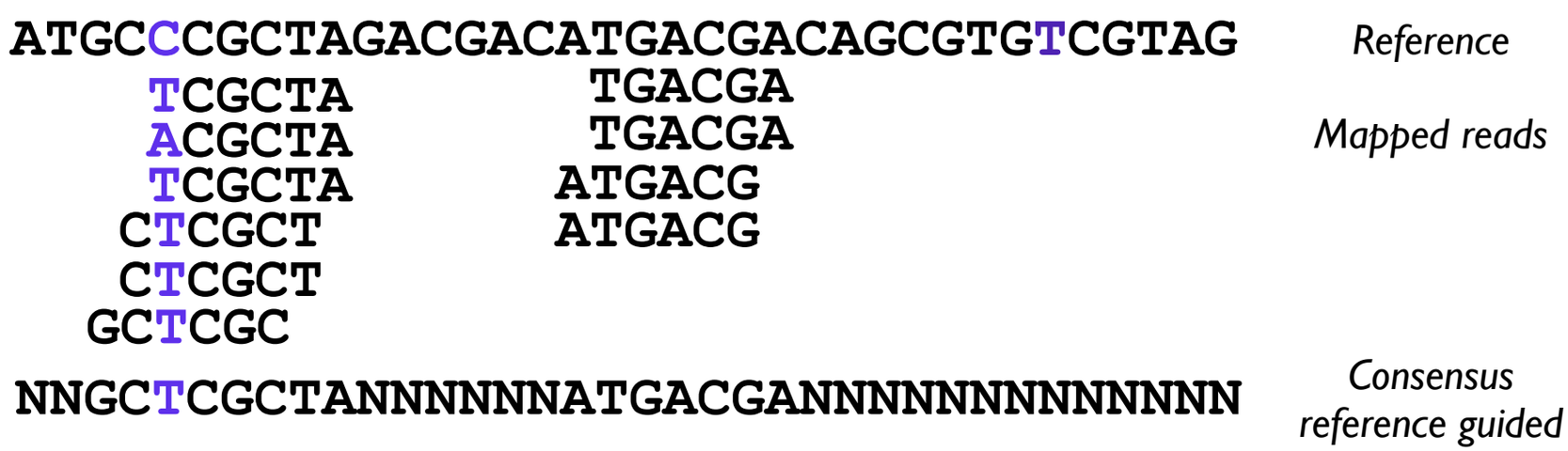
One of the advantages of the RNAseq data to the microarrays is that RNAseq produces thousand of mRNA sequences. These sequences, like any other sequence can be used to perform a phylogenetic analysis:

1. Use CDS sequence, from start codon to the codon before the stop codon. Use full length if they are available.
2. The consensus sequence is supported by enough reads to avoid sequencing errors.

4.1 Consensus sequences

There are two ways to retrieve a consensus sequence of a set of reads from the same gene (or gene family).

1. De-novo assembly
2. Reference guided assembly



4.1 Consensus sequences

The most common tool used to generate a sequence consensus from a read dataset alignment is samtools/bcftools using the SNP information generated by samtools mpileup.

```
samtools mpileup -uf reference.fa align.bam |

bcftools view -cg - |

vcfutils.pl varFilter -d 5 |

vcfutils.pl vcf2fq > consensus.fq
```

1) BAM => BCF

2) BCF => VCF

3) Filter SNP depth < 5

4) Generate consensus

BCF (Binary variant Call Format) stores the variant call for the mapped reads at each reference position.

4.1 Consensus sequences

VCF (Variant Call Format) is a common file format to store sequence polymorphism (SNPs and INDELs) based in a reference position. It has three parts:

- Meta-information lines (starting with '#')
- Header line (starting with '#CHROM').
- Data lines, 8 columns separated by tabs.
  - CHROM, chromosome
  - POS, position (1-based coordinate)
  - ID, identifier for the polymorphism
  - REF, reference base
  - ALT, alternative base (for no alternative base, '.')
  - QUAL, phred based quality score for the alternative base
  - FILTER, if the variant call has passed thr filter
  - INFO, additional information, such as read depth (DP), or allele frequency (AF1)

http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41

4.1 Consensus sequences

Once the consensus is generated, the mRNA or the cds can be retrieved using a gene model annotation gff3 file.

```
bp_sreformat.pl -if fastq -of fasta
                -i consensus.fq
                -o consensus.fa

gffread annot.gff -g consensus.fa -w mrna.fa

gffread annot.gff -g consensus.fa -x cds.fa
```

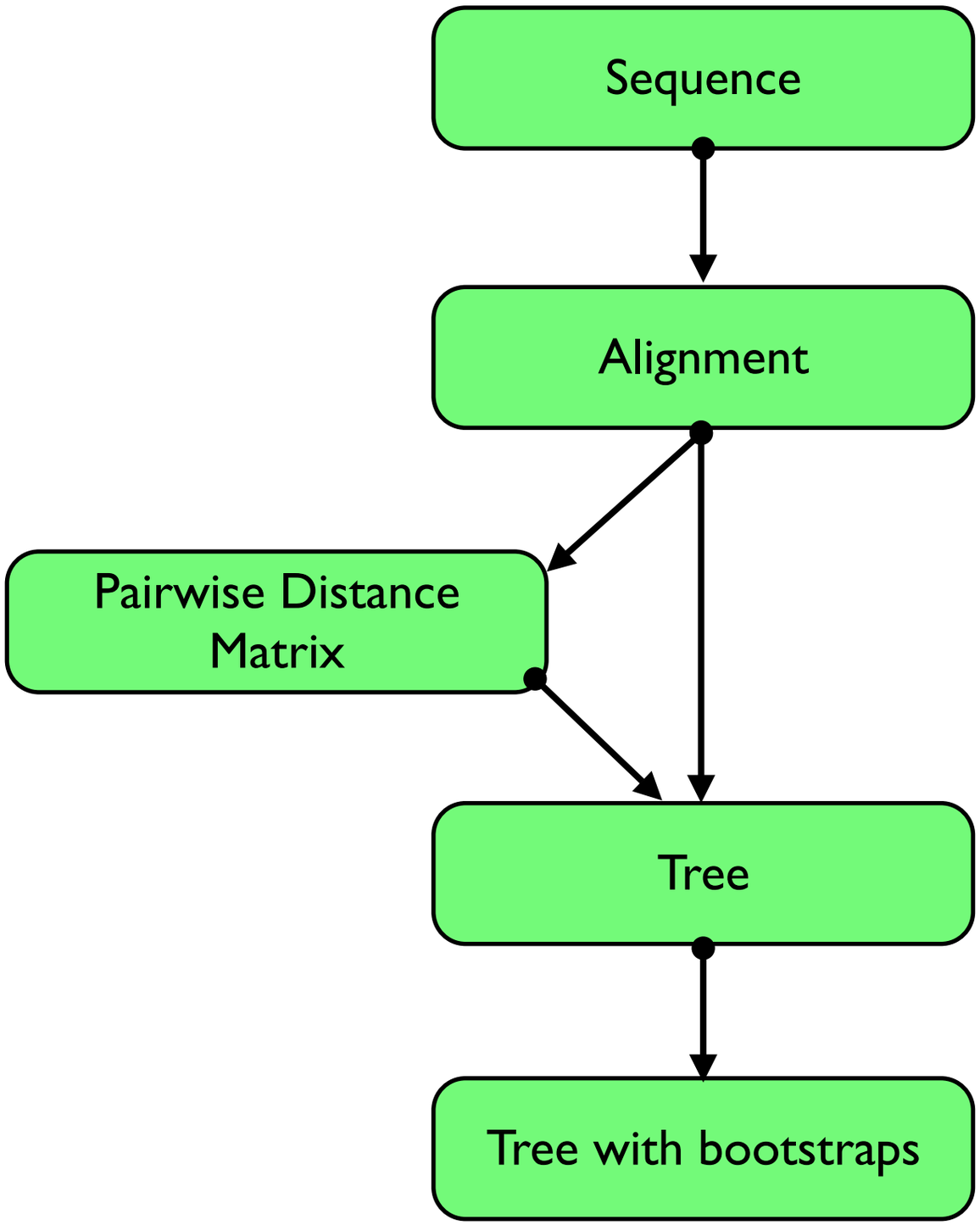
1) Fastq => Fasta

2.1) Get mRNA

2.2) Get CDS

Consensus sequences will have the same ID than the reference so is easy to retrieve the same gene in different samples.

4.2 Inferring the gene tree.



1) Select the gene of interest. If it is possible search for an outgroup at GenBank and add to the fasta file with all the sequences.

2) Select the alignment tool, run the alignment and check the results to get an optimal alignment.

3) Decide the type of phylogenetic analysis to perform and choose the right software based in the method, speed, memory consumption and usability.

4.2 Inferring the gene tree.

There are dozens of multiple sequence alignment tools. The most representatives are:

Software	Feature	Alignment Type	URL
ClustalW	Progressive alignment	Global and Local	http://www.ebi.ac.uk/clustalw/
DiAlign	Segment-based method	Global and Local	http://bibiserv.techfak.uni-bielefeld.de/dialign
Kalign	Progressive alignment	Global	http://msa.cgb.ki.se/
MAFFT	Progressive and iterative alignment	Global and Local	http://align.bmr.kyushu-u.ac.jp/mafft/software/
MUSCLE	Progressive and iterative alignment	Global and Local	http://phylogenomics.berkeley.edu/cgi-bin/muscle/input_muscle.py
TCoffee	Sensitive progressive alignment	Global and Local	http://www.tcoffee.org

http://www.kuleuven.be/aidslab/phylogenybook/Table3.1.html

4.2 Exploring multiple gene phylogenetic trees

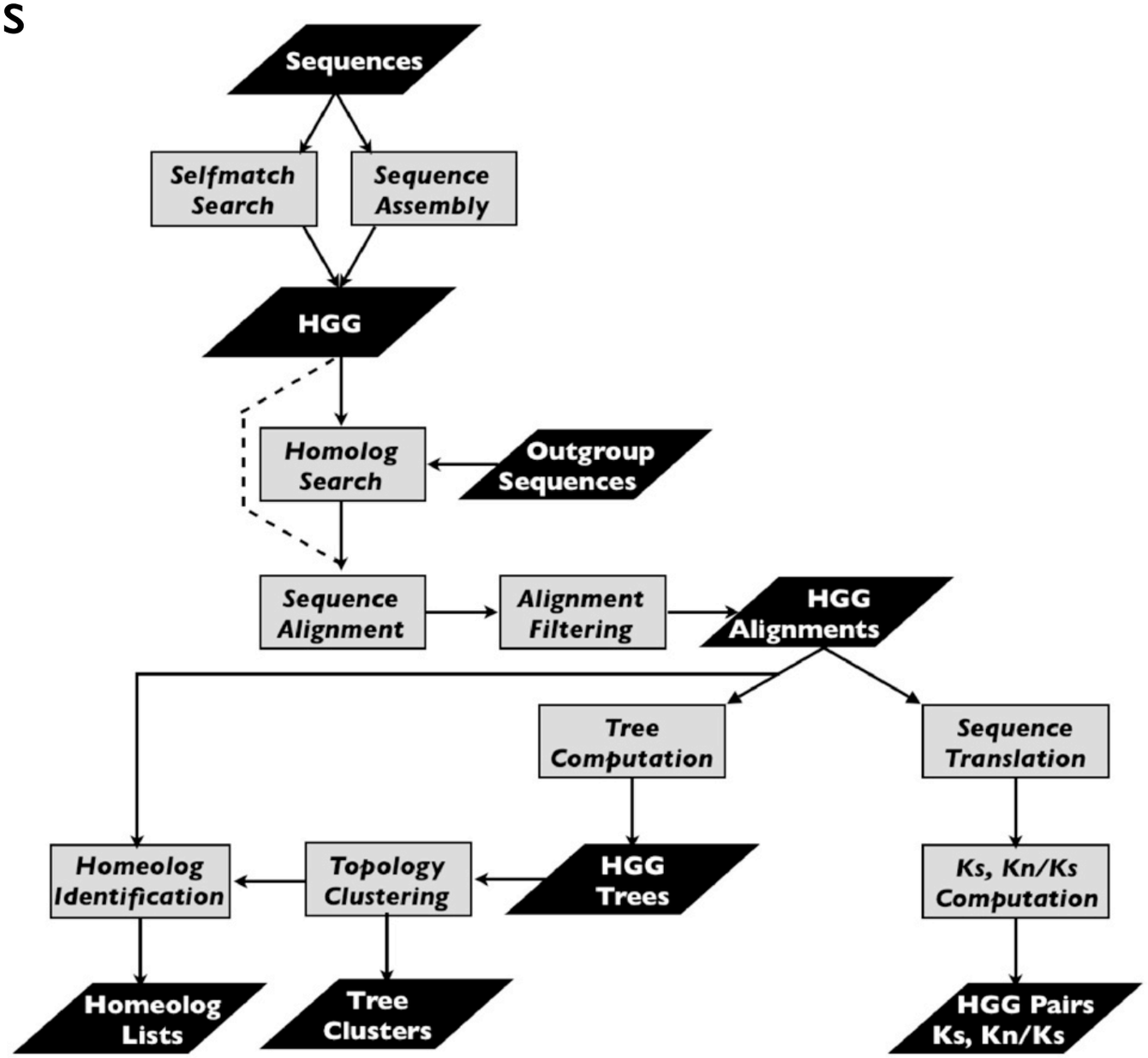
Multiple phylogenetic analysis can be performed for different gene groups.

- Hal (Robbertse B. *et al.* 2011).
- PhygOmics (Bombarely A.)

1. Search of unusual tree topologies.
2. Look into the most represented tree topologies to infer an species tree.

4.2 Exploring multiple gene phylogenetic trees

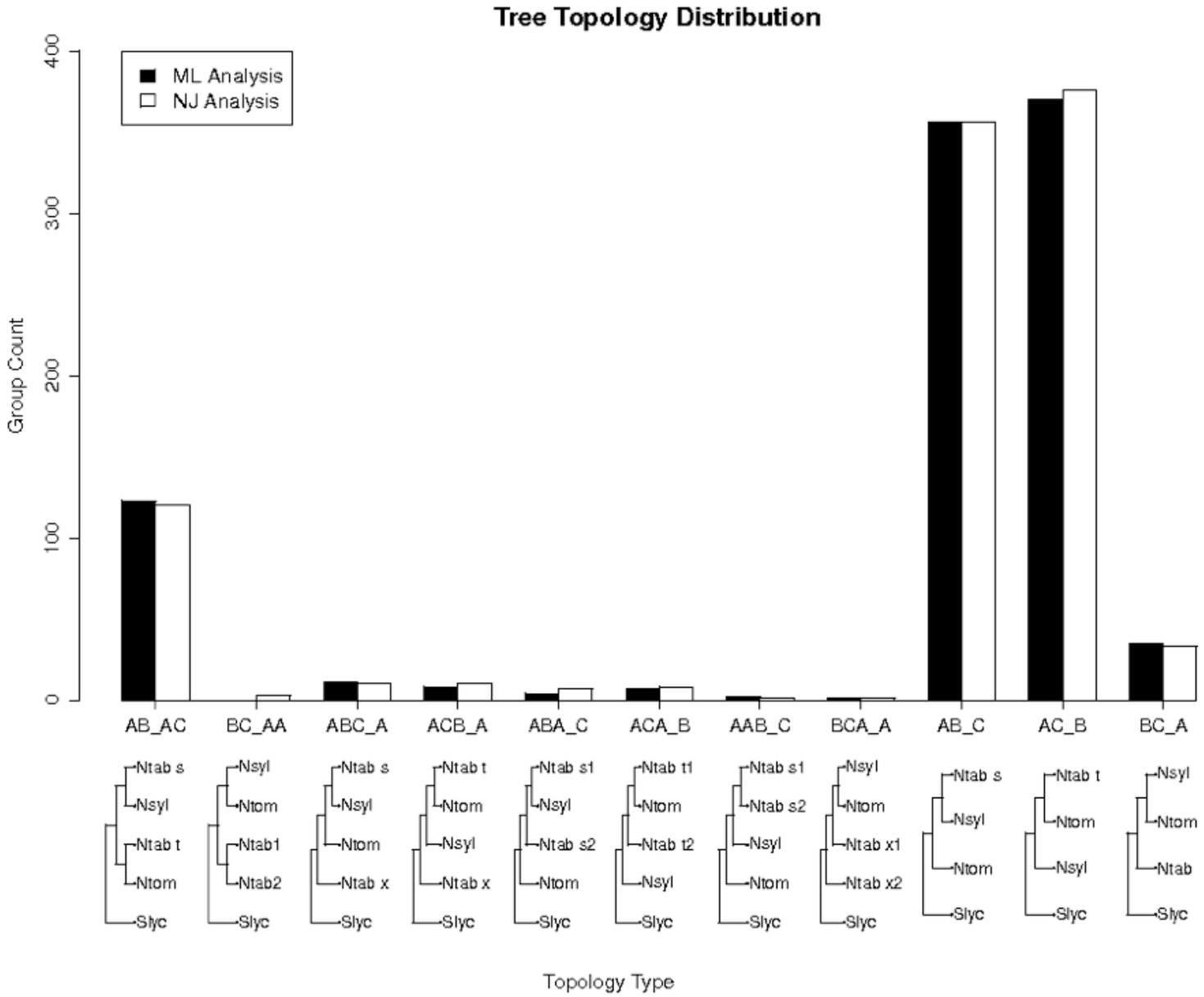
- PhygOmics



https://github.com/solgenomics/PhygOmics

4.2 Exploring multiple gene phylogenetic trees

- PhygOmics



~ 1,000 Gene trees for the allotetraploid *Nicotiana tabacum* and its diploids progenitors, *N. sylvestris* and *N. tomentosiformis* were analyzed to identify the origin of each homoeolog.

Bombarely A et al., 2012 *BMC Genomics*

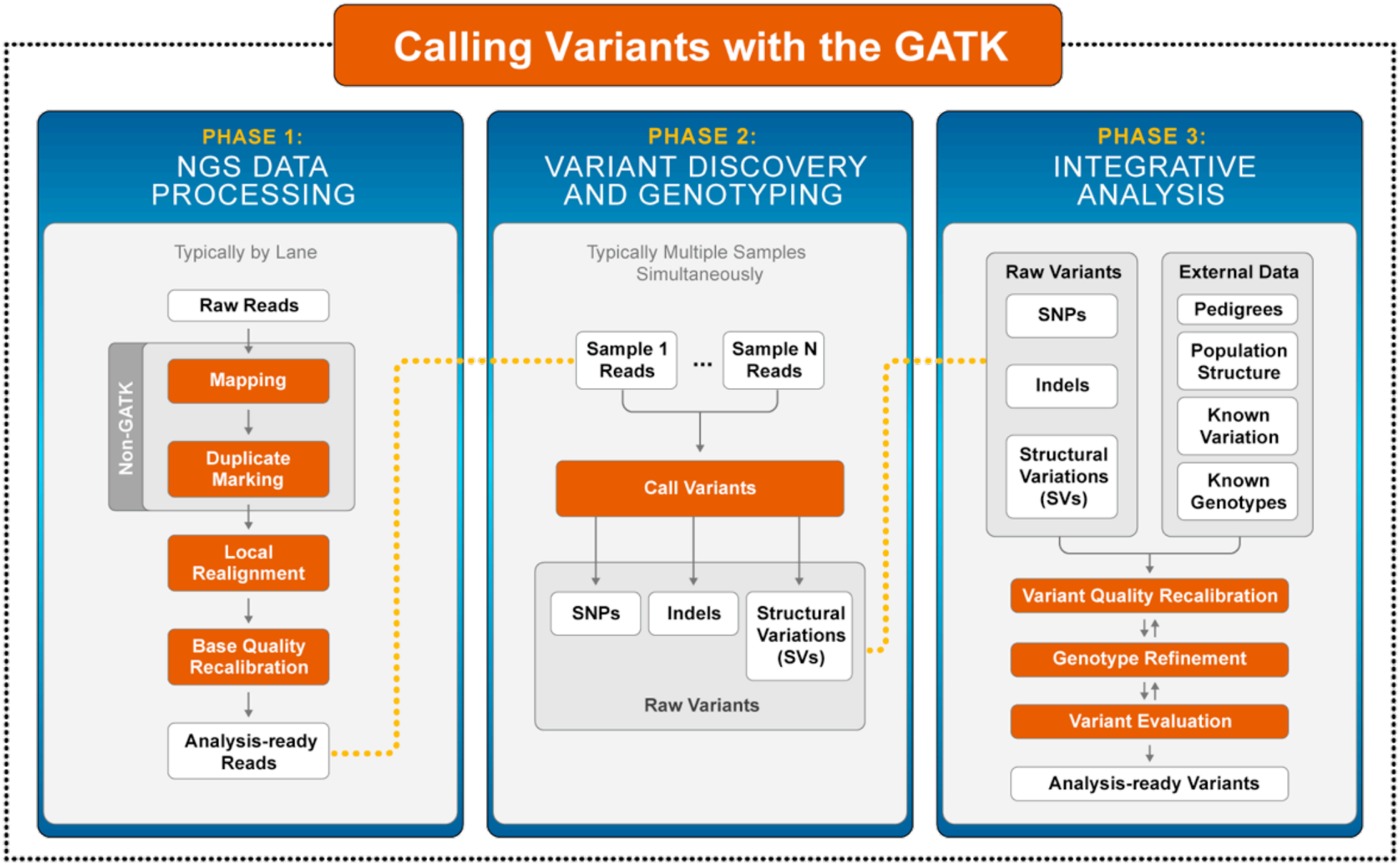
4.3 SNPs from RNAseq

Polymorphism analysis can be performed over the RNAseq alignments. There are several programs that can be used for this purpose:

Software	Features	URL
Samtools/ Bcftools	SNPs, INDELS	http://samtools.sourceforge.net/mpileup.shtml
GATK	SNPs, INDELs and SV (Structural Variations). Polyploids.	http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit
FreeBayes	SNPs, INDELs, MNPs (Multiple Nucleotide Polymorphism), Complex Events. Polyploids	http://bioinformatics.bc.edu/marthlab/FreeBayes
Cortex_var	Pipeline for genome assembly and SNP calling for population	http://cortexassembler.sourceforge.net/index_cortex_var.html
Ngs_backbone	Pipeline to process and align reads and call SNP and SSR	http://bioinf.comav.upv.es/ngs_backbone/

4.3 SNPs from RNAseq

- Example of GATK workflow



http://www.broadinstitute.org/gatk/about

4.3 SNPs from RNAseq

After the SNP calling and before use the SNP data for other analysis is recommended to perform a SNP filtering.

Common SNP calling errors:

- Missing calls for SNPs with overlapping genotype clusters (Anney et al., 2008)
- Homozygote–heterozygote miscalls (Teo et al.,2007),
- False homozygote calls in heterozygous individuals due to allelic dropout (Pompanon et al.,2005)
- Erroneous assessment of monomorphic SNPs as polymorphic (Pettersson et al., 2008).

Pongpanich M et al., 2010 *Bioinfomatics*

4.3 SNPs from RNAseq

Common SNP filtering criteria:

1. Read alignment and SNP calling based.
    - 1.1. By a minimum read depth (DP).
    - 1.2. By a minimum quality variant call (QUAL).
    - 1.3. By a maximum/minimum allele frequency (AF1)
    - 1.4. Biallelic polymorphism.
    - 1.5. By minimum physical distance between SNPs
    - 1.6. By a minimum distance from a genomic/genetic element.

2. Genetic data incongruences.
    - 2.1. Hardy–Weinberg equilibrium (HWE)
    - 2.2. Missing proportion (MSP)
    - 2.3. Minor allele frequency (MAF)

Chagne D. et al. 2012 *Plos One*
Pongpanich M et al., 2010 *Bioinfomatics*

4.4 Uses for SNPs

A whole genome SNP dataset can be an inestimable source of markers with a wide use spectrum such as:

- Marker discovery (example: SNPs and SSRs for Pepper, Ashrafi H. *et al.* 2012;CbCC methods in Chickpea, Azam S. *et al.* 2012)
- Genetic map development (example: Genetic map in *Miscanthus sinensis*, Swaminathan K. *et al.* 2012).
- Gene mapping (example: Gene Mapping via Bulked Segregant RNA-Seq (BSR-Seq), Liu S. *et al.* 2012, Trick M. *et al.* 2012).
- Population genetic analysis (example: Population genetic of sunflowers, Renaut S. *et al.* 2012).
- eQTL (phosphorous supply intake in *Brassica rapa*, Hammond JP. *et al.* 2012)
- Homoeologus regions identification in polyploids (unpublished).

4.4 Uses for SNPs

- Homoeologus regions identification in polyploids

Glycine syndetika
D4
(2n)

X

Glycine tomentella
D3
(2n)

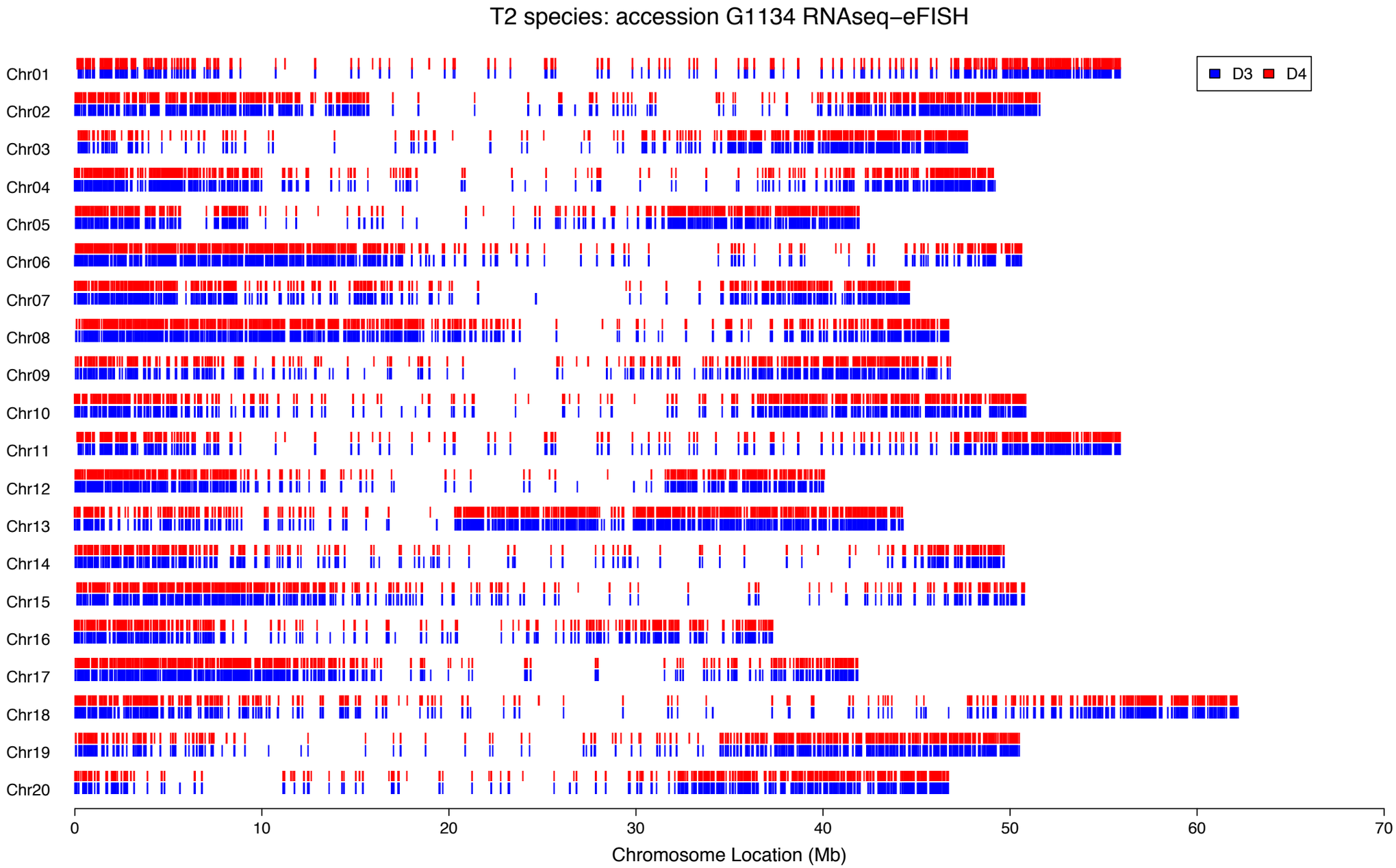
↓

Glycine dolichocarpa
T2
(4n)

1. RNAseq from leaves samples of D3, D4 and T2 (Coates J. *et al.* 2012)
2. Generation of the consensus sequence for D3 and D4 using *G. max* as reference.
3. Selective mapping of T2 reads to D3 or D4 consensus.
4. SNP representation.

4.4 Uses for SNPs

- Homoeologus regions identification in polyploids



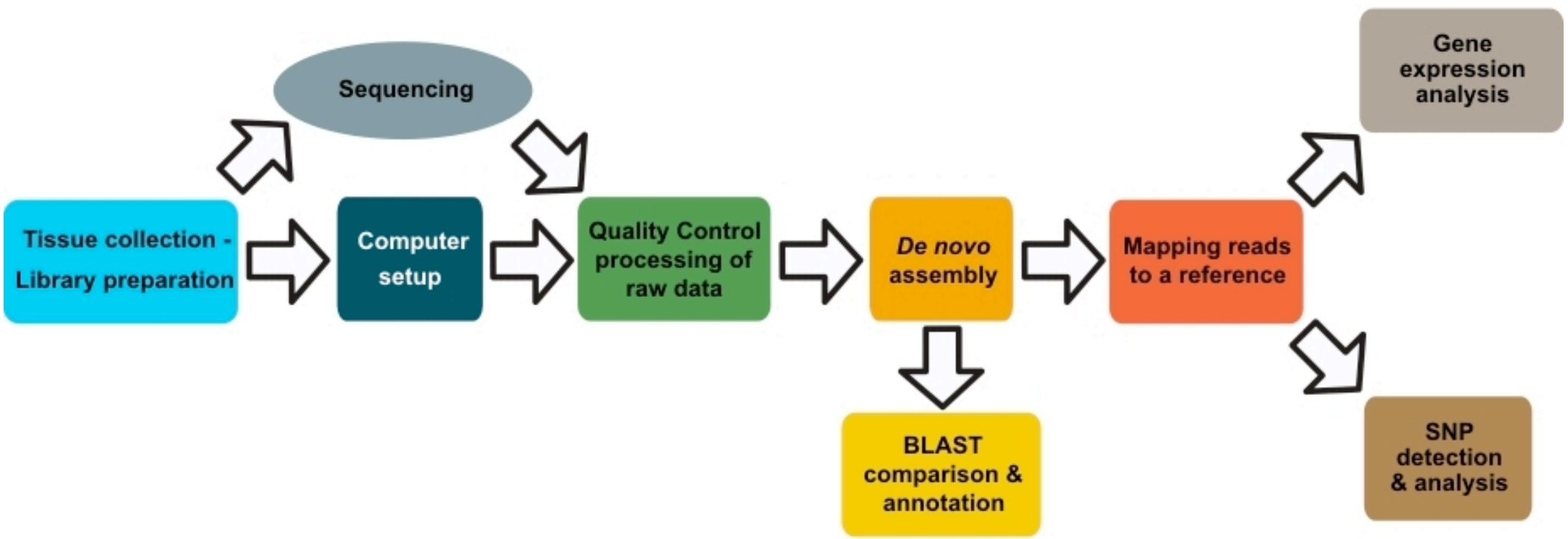
4.4 Uses for SNPs

- Population genetic analysis

  Recommended resource:

  The Simple Fool's Guide to Population Genomics via RNA-Seq

  http://sfg.stanford.edu/



4.4 Uses for SNPs

- Population genetic analysis

Software	Features	URL
TASSEL	Association analysis, PCA for populations	http://www.maizegenetics.net/tassel
Structure	Population structure analysis	http://pritch.bsd.uchicago.edu/structure.html
FineStructure	Population structure analysis for High Throughput Data. PCA	http://paintmychromosomes.com/
Phase	Genetic phasing of alleles	http://stephenslab.uchicago.edu/software.html#phase

4.4 Uses for SNPs

- Population genetic analysis

Population analysis tools frequently use a specific format different from the VCF format produced by SNP calling tools. Options:

1. Write your own script.
2. Use a script that someone wrote

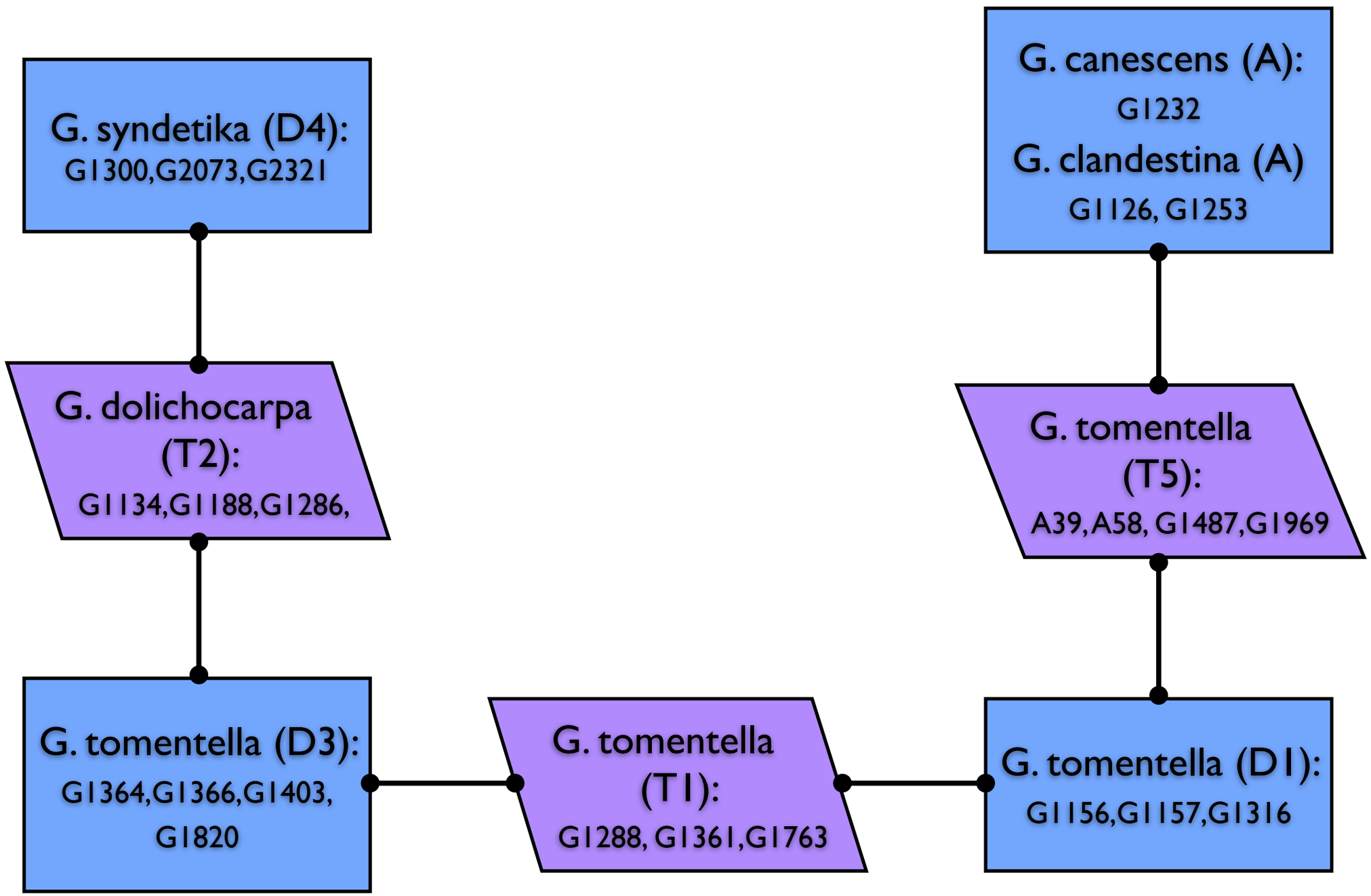
GenoToolBox

MultiVcfTool Hapmap2Structure

https://github.com/aubombarely/GenoToolBox

4.4 Uses for SNPs

- Population genetic analysis: Example Glycine perennials analysis



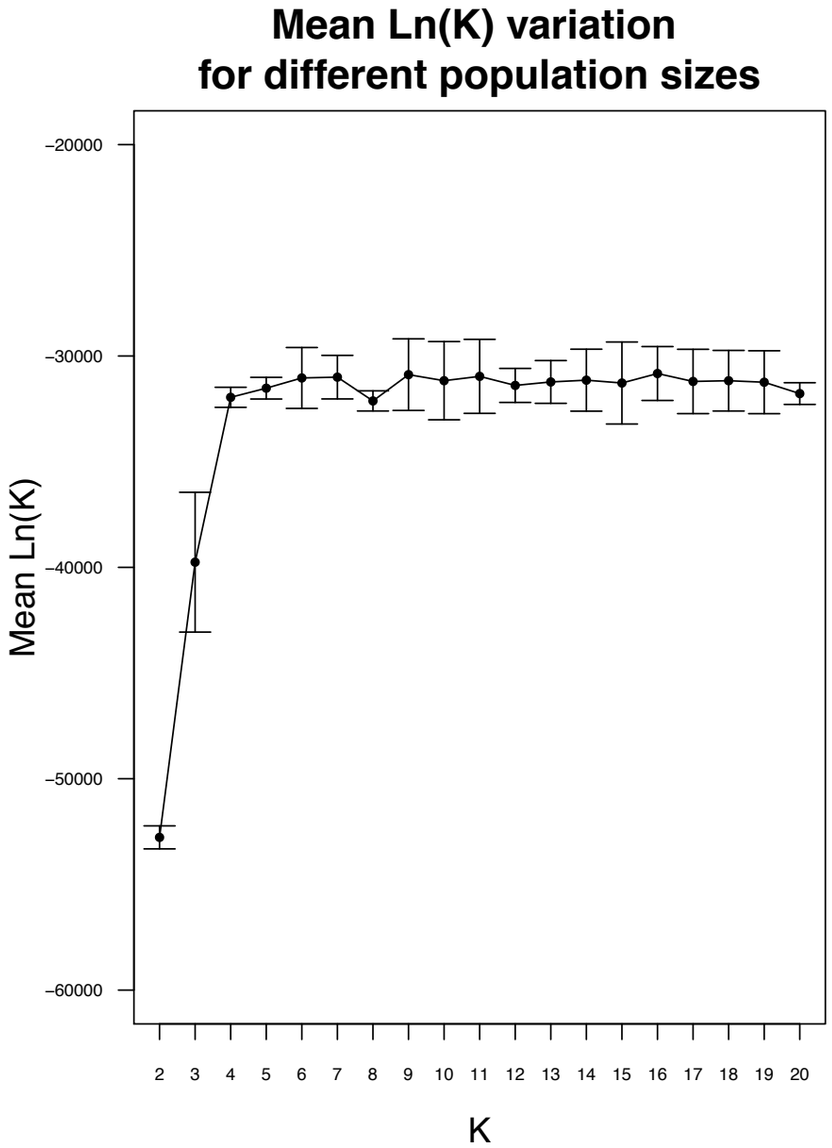
4.4 Uses for SNPs

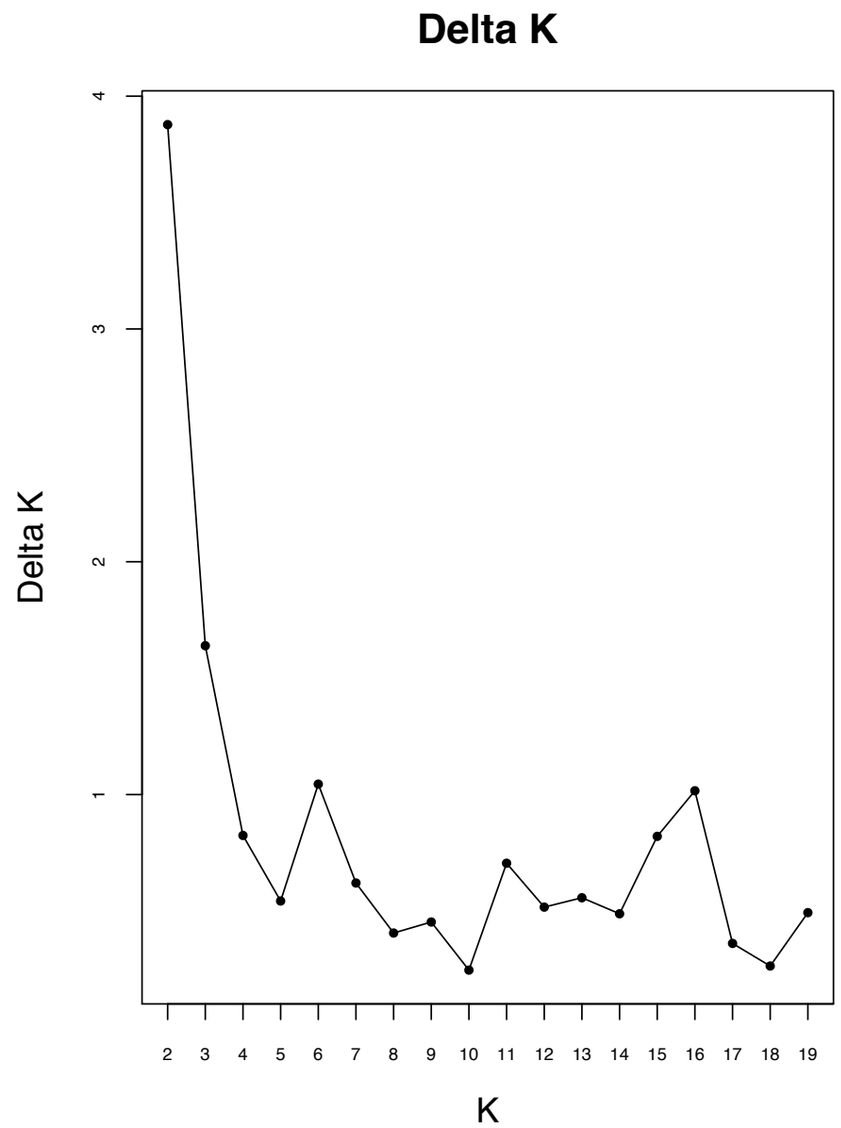
- Population genetic analysis: Example Glycine perennials analysis

1. RNAseq of 8 species (5 diploids (D1, D3, D4, A canescens, A clandestina), 3 allotetraploids (T1, T2, T5)), 25 accessions.
2. Generation of the consensus sequence for A, D1, D3 and D4 using *G. max* as reference.
3. Selective mapping of:
    1. T1 reads to D1 or D3 consensus
    2. T2 reads to D3 or D4 consensus.
    3. T5 reads to A or D1
4. SNP calling.
5. Change format from:
    1. VCF to HapMap for TASSEL
    2. VCF to Structure for Structure
    3. VCF to phase for fineStructure

4.4 Uses for SNPs

- Population genetic analysis: Structure





Structure analysis:

1. Number of clusters optimization (Evanno G. et al. 2005)

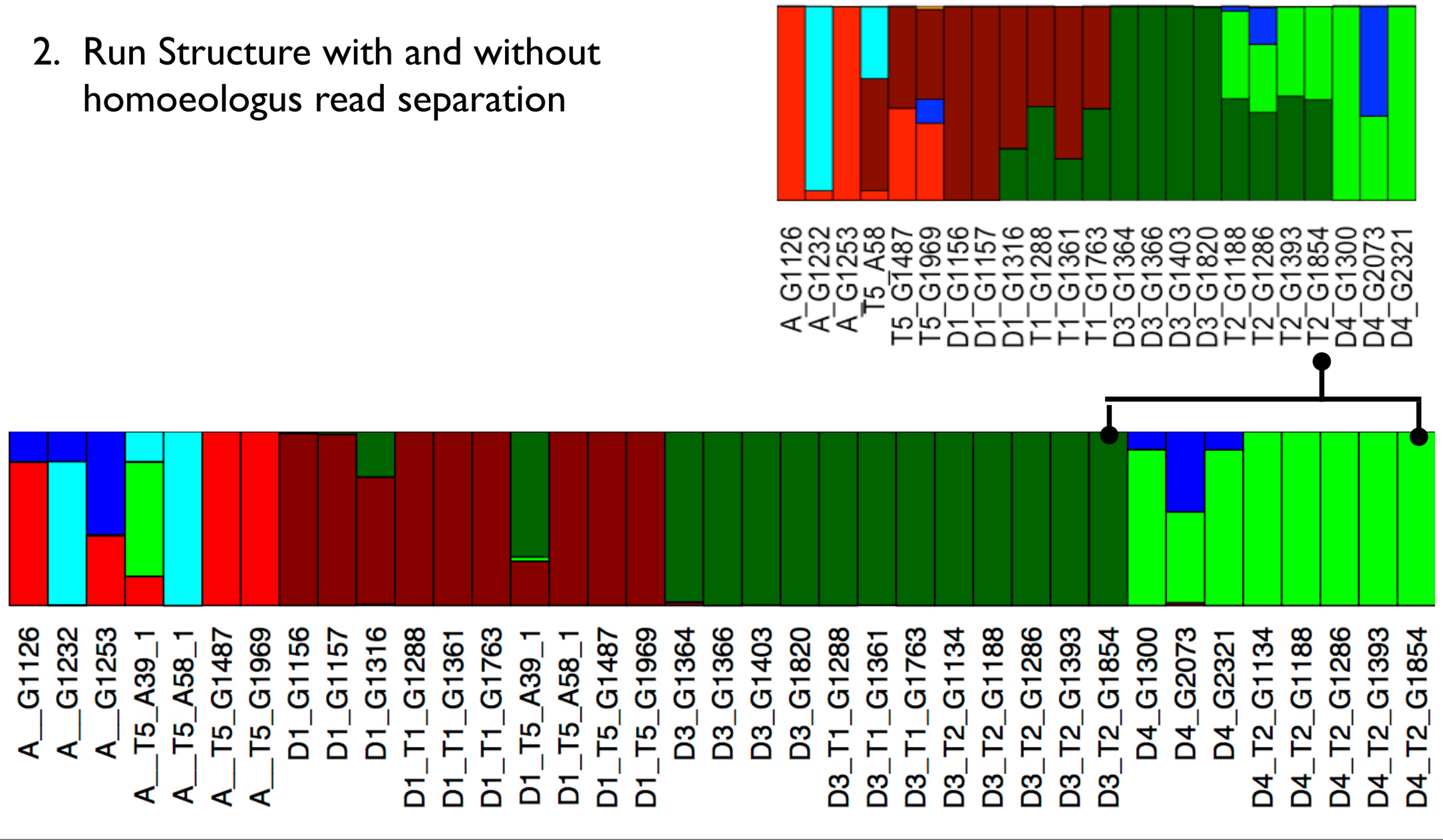
K = 6

(K = 16)

4.4 Uses for SNPs

- Population genetic analysis: Structure

2. Run Structure with and without homoeologus read separation

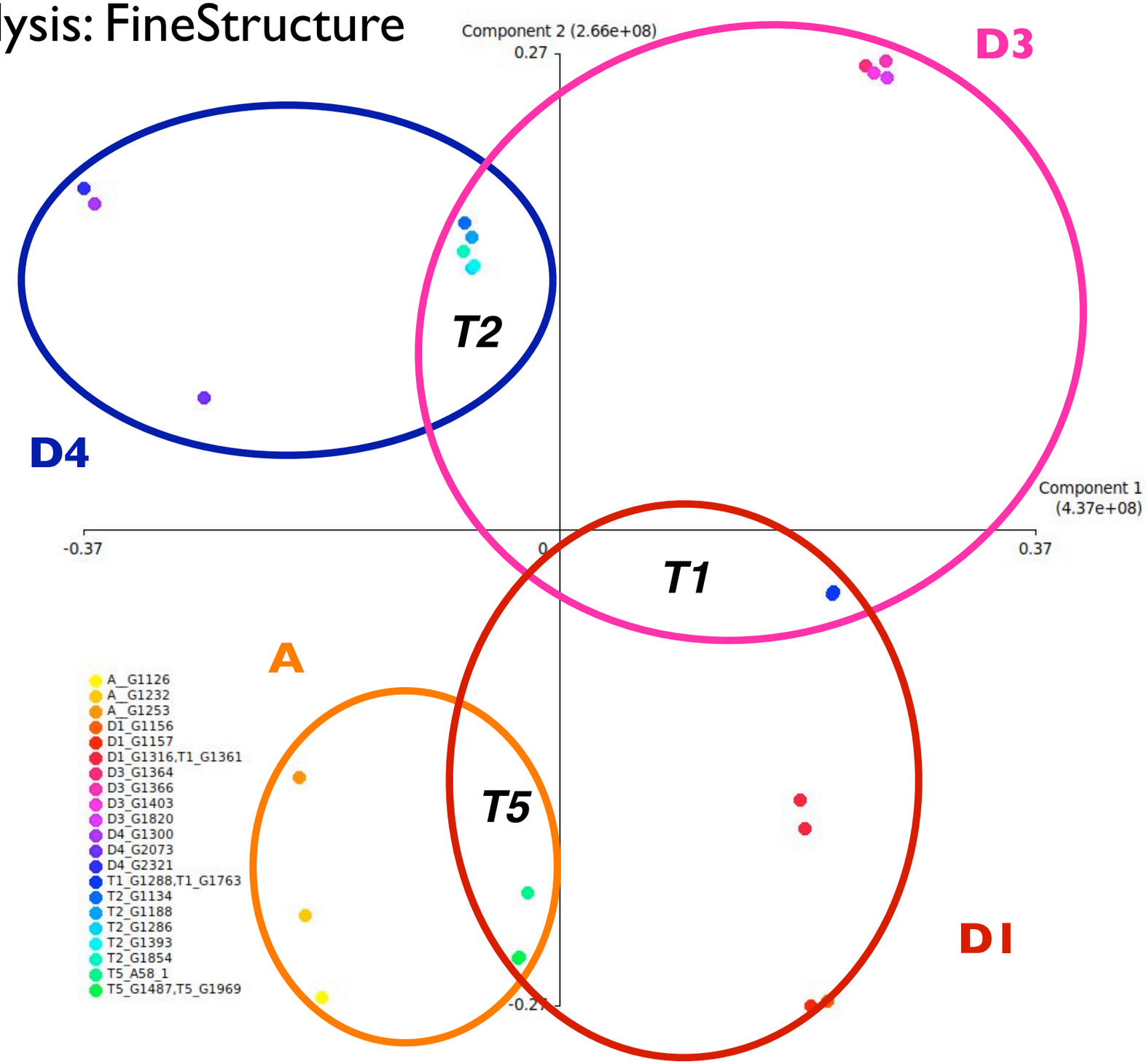


4.4 Uses for SNPs

- Population genetic analysis: FineStructure

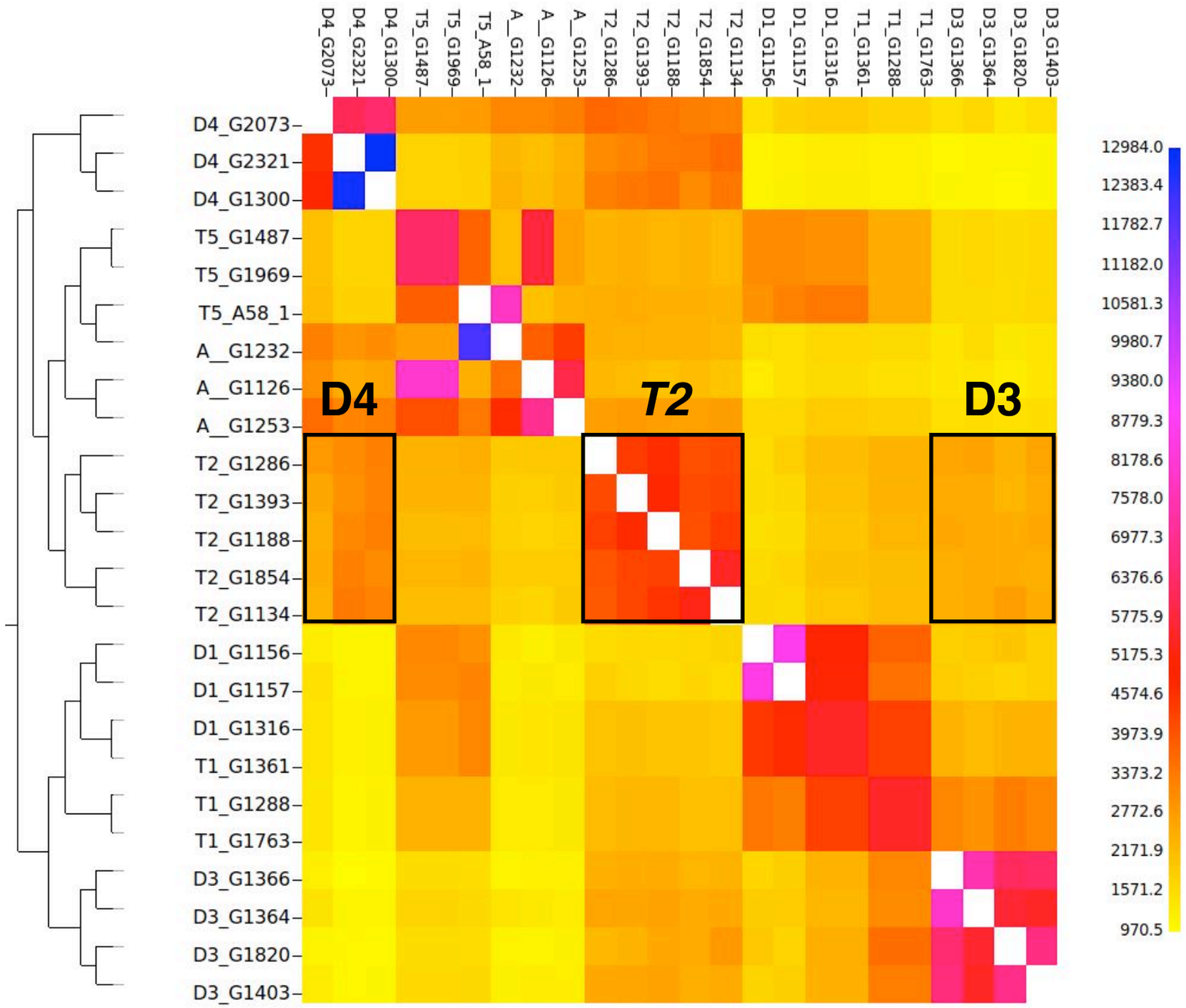
3. Run FineStructure without homoeologus read separation

Admixture model

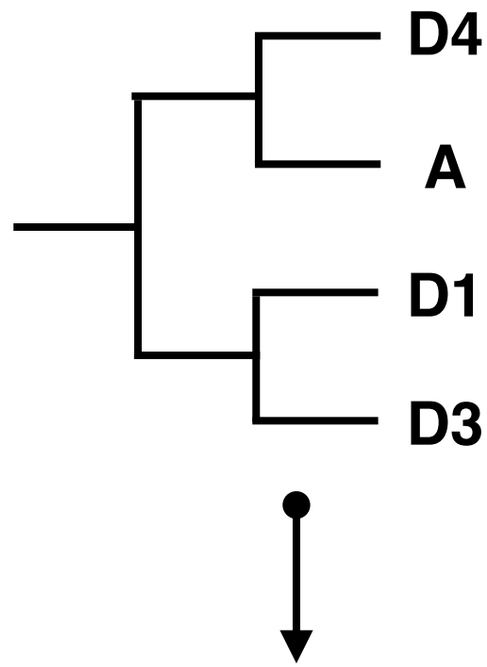


4.4 Uses for SNPs

- Population genetic analysis: FineStructure



Phylogenetic information:



It agrees with previous data from nuclear genes

Exercises:

1. **Basic Linux commands.**
2. **Sequencing evaluation.**
3. **Simple read mapping.**
4. **Simple de-novo assembly.**
5. **Basic R commands**
6. **Functional annotation.**
7. **Differential gene expression.**
8. **Cluster analysis for gene expression.**
9. **Selecting genes for phylogeny.**
10. **SNP calling and filtering.**
11. **Analysis of the population structure.**

Exercises:

1. **Basic Linux commands.**
2. **Sequencing evaluation.**
3. **Simple read mapping.**
4. **Simple de-novo assembly.**
5. **Basic R commands**
6. **Functional annotation.**
7. **Differential gene expression.**
8. **Cluster analysis for gene expression.**
9. **Selecting genes for phylogeny.**
10. **SNP calling and filtering.**
11. **Analysis of the population structure.**

Exercises:

1. **Basic Linux commands.**
2. **Sequencing evaluation.**
3. **Simple read mapping.**
4. **Simple de-novo assembly.**
5. **Basic R commands**
6. **Functional annotation.**
7. **Differential gene expression.**
8. **Cluster analysis for gene expression.**
9. **Selecting genes for phylogeny.**
10. **SNP calling and filtering.**
11. **Analysis of the population structure.**