



# Amplicon sequencing by NGS methods

Jean-François Martin

Centre de Biologie pour la Gestion des Populations  
Centre international d'études supérieures en sciences agronomiques

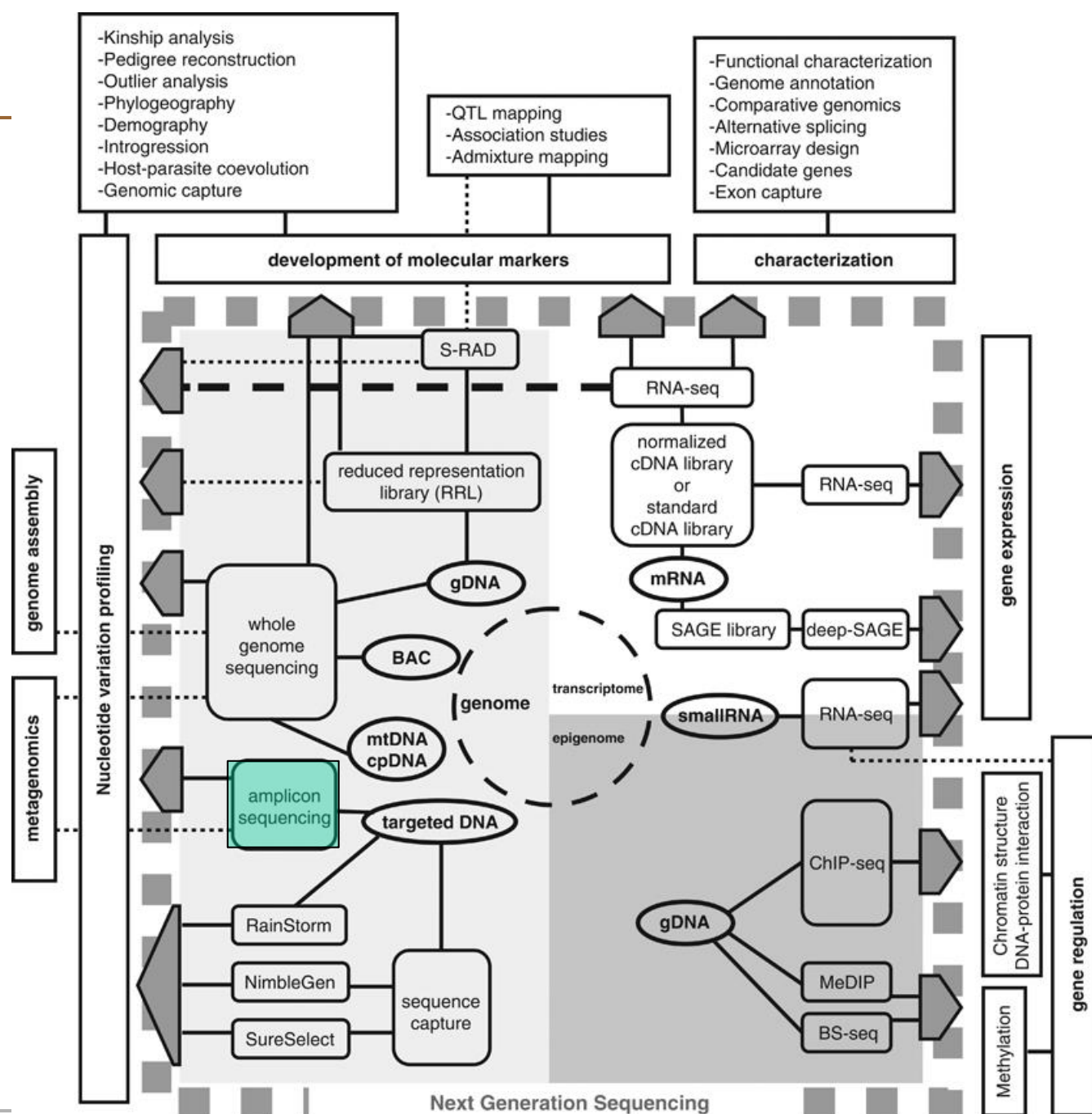
# Goals and expectations

---

■	<b>Part 1</b>
	Part 2
	Part 3
	Part 4
	Part 5

The aim of this discussion today : review the different options to massively sequence amplicons through NGS technology and show some current work on adapting protocols.

## Why sequence amplicons ?



## Most commonly used applications:

- Barcoding
- Metagenomics (including metabarcoding)
- Phylogeny
- Microsatellites genotyping (seriously?!)

# Amplicon sequencing

---

■	<b>Part 1</b>
	Part 2
	Part 3
	Part 4
	Part 5

Most of those applications could also be addressed  
by capture methods

# Amplicon sequencing

---

■	<b>Part 1</b>
	Part 2
	Part 3
	Part 4
	Part 5

The applications will require different analysis methods. This is important in designing the project.

Think about the output information and format you will use for further analysis

# Amplicon sequencing

Part 1

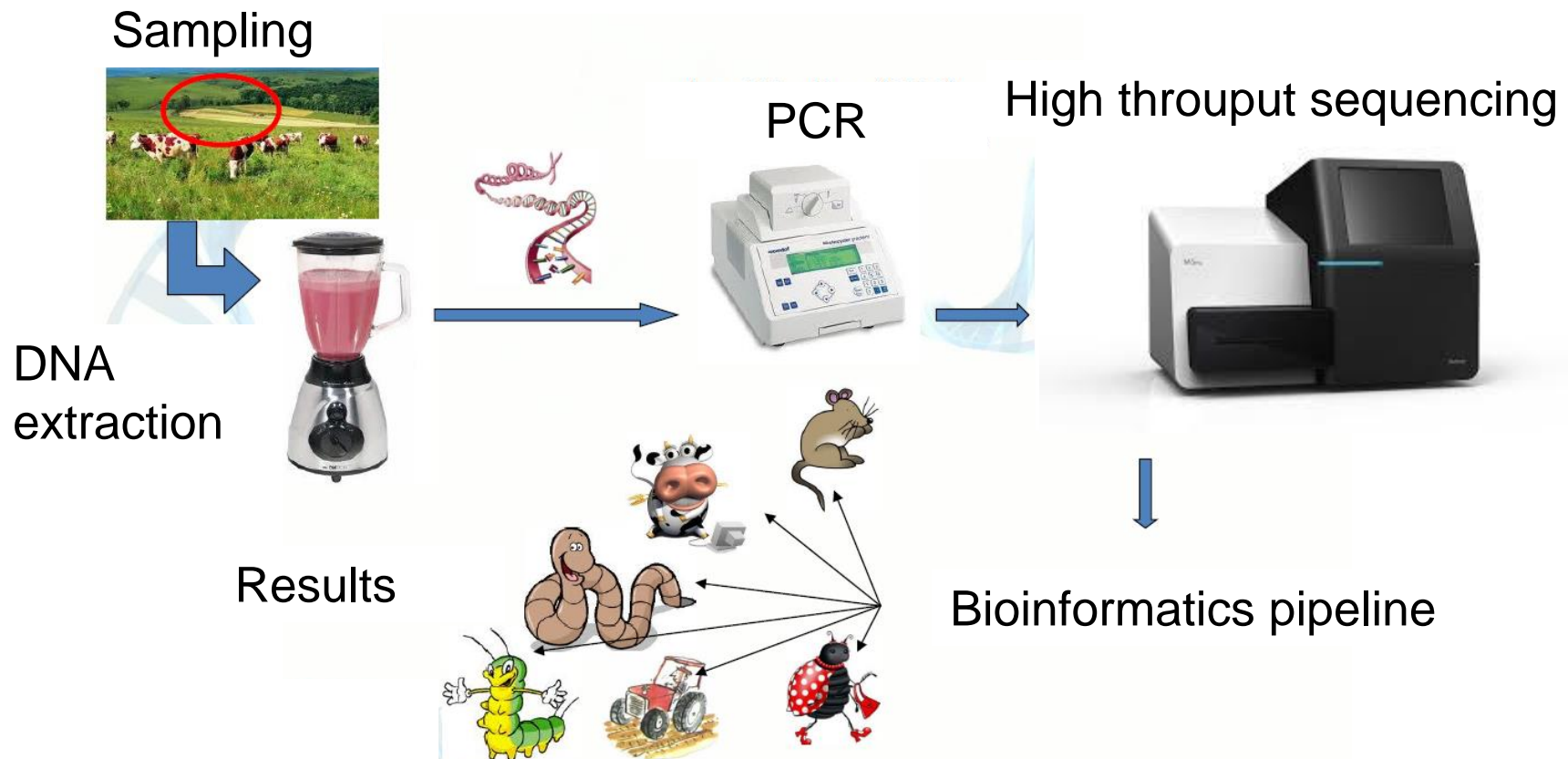
Part 2

Part 3

Part 4

Part 5

Example among others : environmental barcoding





## Goals and applications

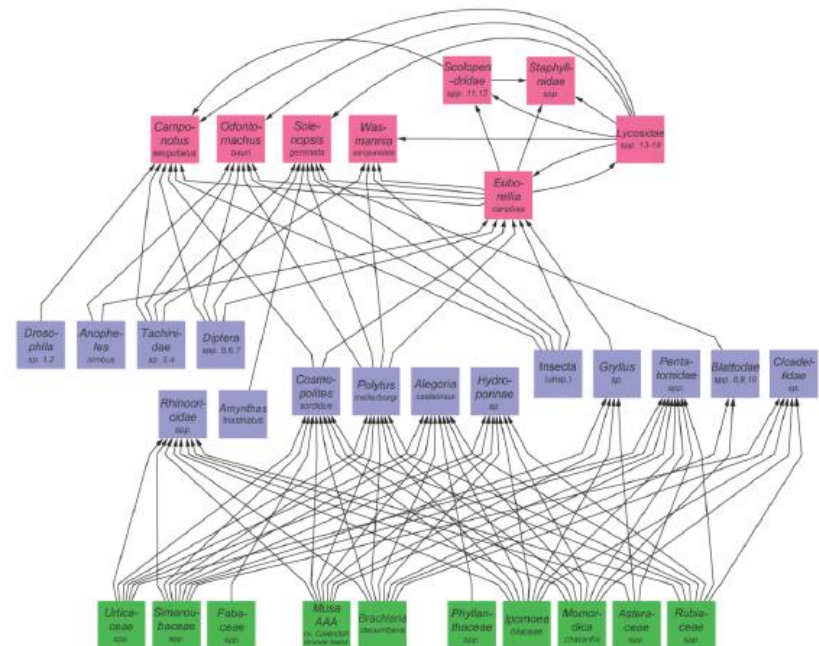
Characterization of the environmental species diversity

Characterization of parasites

Characterization of diet

Trophic network

Microbiome analysis



# Amplicon sequencing

---

- **Part 1**
- Part 2
- Part 3
- Part 4
- Part 5

It is required to setup a simple and efficient methodology to acquire data so it makes amplicon sequencing accessible to anyone

What technology should we choose for sequencing amplicons ?

It depends !

but in any case error rate should be as low as possible

# Amplicon sequencing

---

- **Part 1**
- Part 2
- Part 3
- Part 4
- Part 5

Preliminary testing of pacbio RS on long amplicons

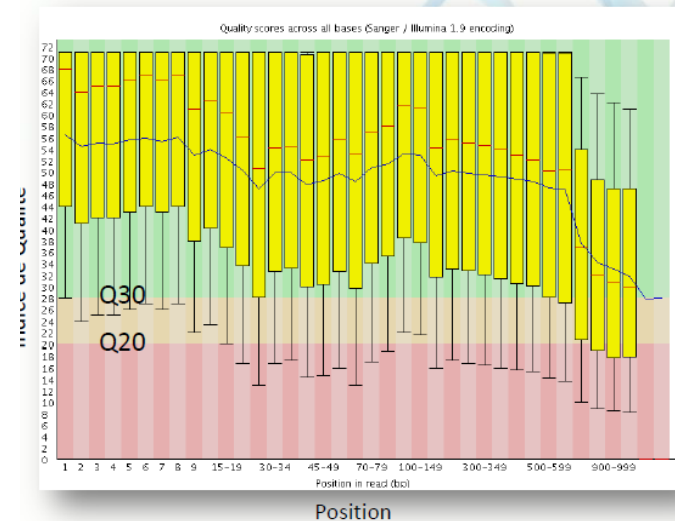


# Amplicon sequencing

Pacbio RS sequencer

## Preliminary testing of pacbio RS on long amplicons

- 100k sequences, including 19-21k ccs
- Up to 17k bases / sequence at the time (march 2014)
- Highly variable quality from one run to the next
- 15% error rate on controls

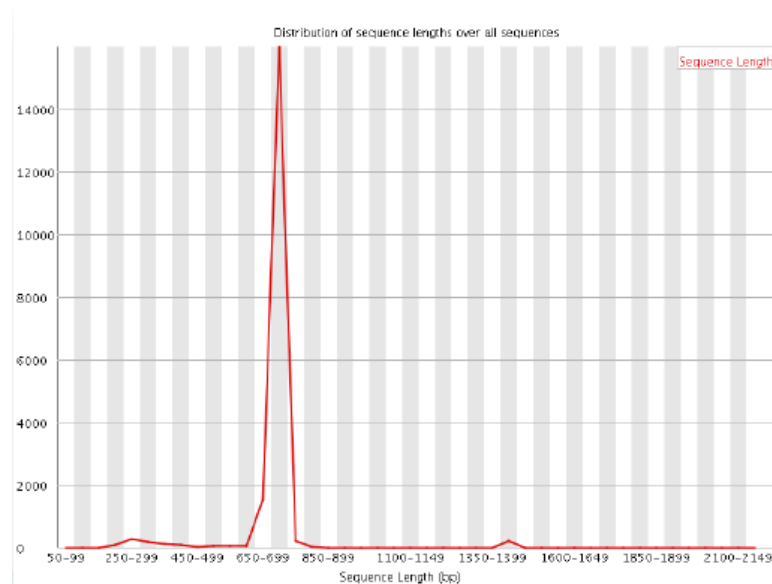


Circular consensus sequence (CCS)

Today on a Pacbio RS II, 15kb median, 40kb max

## Preliminary testing of pacbio RS on long amplicons

98% of the css fragments are of the correct size (658pb)



## Preliminary testing of pacbio RS on long amplicons

- 42% of quantitative variation between technical replicates
- Majoritary sequence (17% in average) always correspond to the expected
- Random error based correction always recovers the correct sequence
- The number of errors/sequence never exceeds 2 (out of 658 bases)
- When mixed samples occur, the reproducibility of ratios is very low

## Preliminary testing of pacbio RS on long amplicons

Conclusion : SMRT is very useful for barcoding long amplicons (658pb) but not usable in environmental applications.

Although compared to sanger it is still expensive



## Designing amplicon sequencing strategy with Miseq sequencing

## Designing amplicon sequencing strategy with Miseq sequencing

Goal : simultaneously sequence up to thousands of amplicons

- How many sequences for each amplicon ?
- How do I multiplex the amplicons ?
- What are the limits of the technology ?

## Designing amplicon sequencing strategy with Miseq sequencing

How many sequences for each amplicon ?

- What is the expected variation for an amplicon ?
- How many sequences to validate a variant ?
- Also means that all amplicons are equitably represented in the sequencing run and all sequences are usable

## Designing amplicon sequencing strategy with Miseq sequencing

How do I multiplex the amplicons?

What multiplexing level do I need ?

What are the built-in multiplex limits for Illumina ?

How do I combine amplification and sequencing ?

## Designing amplicon sequencing strategy with Miseq sequencing

What are the limits of the technology?

Are there artefacts that matter for my application ?

How many amplicons can I sequence in a given period of time ?

What are the costs (direct and indirect) for the different strategies ?

# Amplicon sequencing

## Part 1

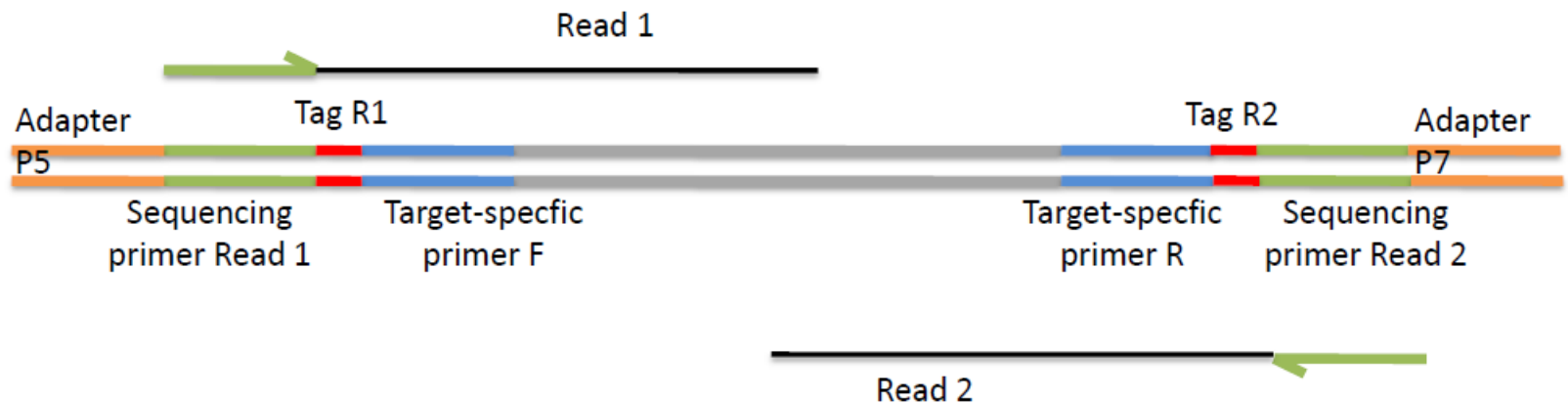
Part 2

Part 3

Part 4

Part 5

## The 1-step tagged PCR + library strategy



### Step 1: PCR + Tag



### Step 2: Pooling + Ligation (kit TrueSeq)



## The 1-step tagged PCR + library strategy

- Amplify the target with tagged primers
- Pool the amplicons by marker / size ...
- Prepare a library for each pool
- Sequence (Miseq v3)

- Part 1
- Part 2
- Part 3
- Part 4
- Part 5

## « Tag »

		#tag																																									
Forward	1	T	C	G	A	T	C	A	C	G	A	T	G	T	T	C	C	A	C	T	A	A	T	C	A	C	A	A	R	G	A	T	A	T	T	G	G	T	A	C			
	2	C	G	A	T	C	G	T	C	A	T	C	A	C	G		T	C	C	A	C	T	A	A	T	C	A	C	A	A	R	G	A	T	A	T	T	G	G	T	A	C	
	3	G	A	T	C	G	A	C	A	G	A	T	C	T	C	C	A	C	T	A	A	T	C	A	C	A	C	A	A	R	G	A	T	A	T	T	G	G	T	A	C		
	4	A	C	G	A	T	C	C	A	C	A	G	T	G	T	C	C	A	C	T	A	A	T	C	A	C	A	A	R	G	A	T	A	T	T	G	G	T	A	C			
	5	T	G	A	T	C	G	A	T	G	A	T	C	A	G		T	C	C	A	C	T	A	A	T	C	A	C	A	A	R	G	A	T	A	T	T	G	G	T	A	C	
	6	C	A	T	C	G	A	G	T	A	G	A	G	T	C	C	A	C	T	A	A	T	C	A	C	A	A	R	G	A	T	A	T	T	G	G	T	A	C				
	7	G	T	C	G	A	T	C	A	T	G	T	C	A	T	C	C	A	C	T	A	A	T	C	A	C	A	A	R	G	A	T	A	T	T	G	G	T	A	C			
	8	A	G	A	T	C	G	T	A	C	T	A	G	C	T	T	C	C	A	C	T	A	A	T	C	A	C	A	A	R	G	A	T	A	T	T	G	G	T	A	C		
	9	T	A	T	C	G	A	C	G	A	T	G	T	G	A	A	A	A	T	C	A	T	A	A	T	G	A	A	G	G	C	A	T	G	A	G	C						
	10	C	T	C	G	A	T	G	A	T	C	A	C	G	G	A	A	A	T	C	A	T	A	A	T	G	A	A	G	G	C	A	T	G	A	G	C						
	11	G	C	G	A	T	C	A	G	C	A	G	A	T	C	G	A	A	A	A	T	C	A	T	A	A	T	G	A	A	G	G	C	A	T	G	A	G	C				
	12	A	T	A	T	C	G	A	C	A	G	T	G	G	A	A	A	A	T	C	A	T	A	A	T	G	A	A	G	G	C	A	T	G	A	G	C						
	Reverse	13	T	A	T	C	G	A	C	G	A	T	G	T	G	A	A	A	A	T	C	A	T	A	A	T	G	A	A	G	G	C	A	T	G	A	G	C					
14		C	T	C	G	A	T	G	A	T	C	A	C	G	G	A	A	A	T	C	A	T	A	A	T	G	A	A	G	G	C	A	T	G	A	G	C						
15		G	C	G	A	T	C	A	G	C	A	G	A	T	C	G	A	A	A	A	T	C	A	T	A	A	T	G	A	A	G	G	C	A	T	G	A	G	C				
16		A	T	A	T	C	G	A	C	A	G	T	G	G	A	A	A	A	T	C	A	T	A	A	T	G	A	A	G	G	C	A	T	G	A	G	C						
17		T	C	T	C	G	A	T	G	A	T	C	A	G	G	A	A	A	A	T	C	A	T	A	A	T	G	A	A	G	G	C	A	T	G	A	G	C					
18		C	G	C	G	A	T	C	T	G	T	A	G	A	G	G	A	A	A	T	C	A	T	A	A	T	G	A	A	G	G	C	A	T	G	A	G	C					
19		G	A	G	A	T	C	A	T	G	T	C	A	G	A	A	A	A	T	C	A	T	A	A	T	G	A	A	G	G	C	A	T	G	A	G	C						
20		A	C	A	T	C	G	A	C	G	T	A	C	G	G	A	A																										

## Percentage for each base



# Amplicon sequencing

## Part 1

Part 2

Part 3

Part 4

Part 5

## The 1-step tagged PCR + library strategy

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12
F1	F1R1	F1R2	...									
F2												
F3												
F4												
F5												
F6												
F7												
F8												

# Amplicon sequencing

Part 1

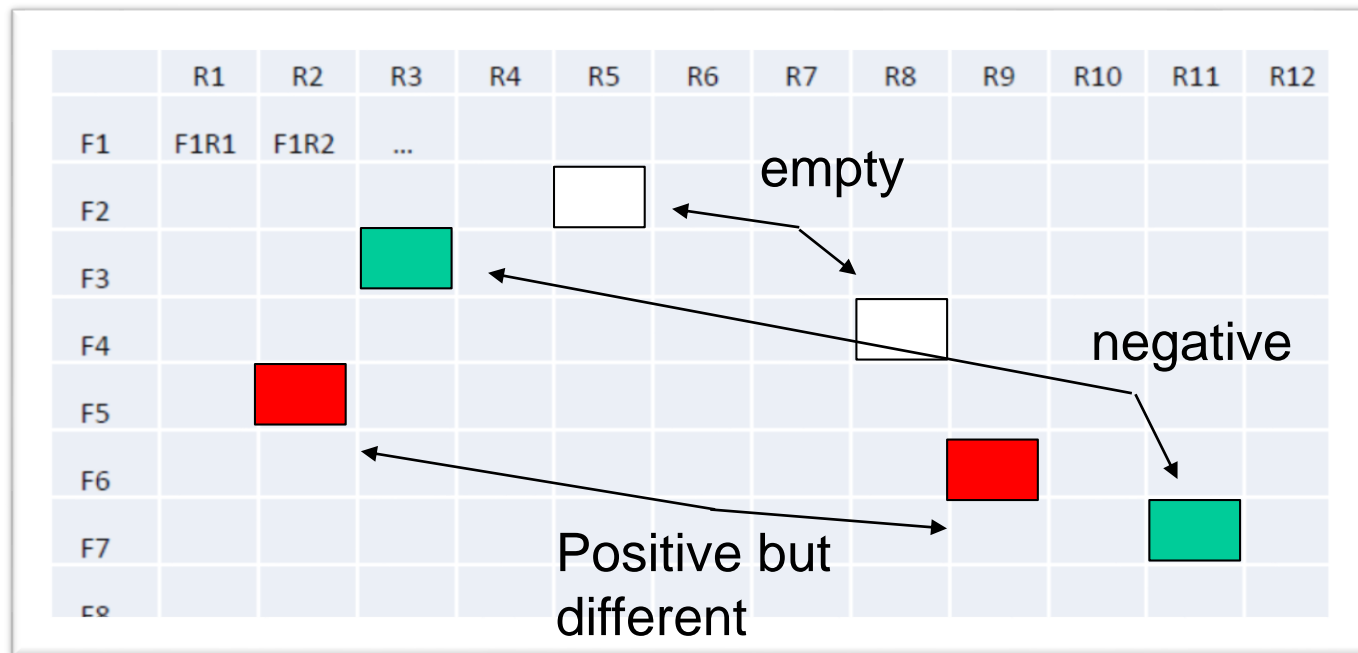
Part 2

Part 3

Part 4

Part 5

## The 1-step tagged PCR + library strategy



**Plate design**

## The 1-step tagged PCR + library strategy

How many tag combinations and how many indexes ?

In other words should I maximise the number of PCR ?

For reference, 1 index / library -> 90€ + handling effort

## Step 2 : combining amplicons as pools

How do I normalize the amplicons before pooling ?

old-school method : they all look the same on my gel,  
I pool them as is.

refined old-school : I create a couple of categories  
according to intensities on gel and pool different  
volumes accordingly

# Amplicon sequencing

## Part 1

Part 2

Part 3

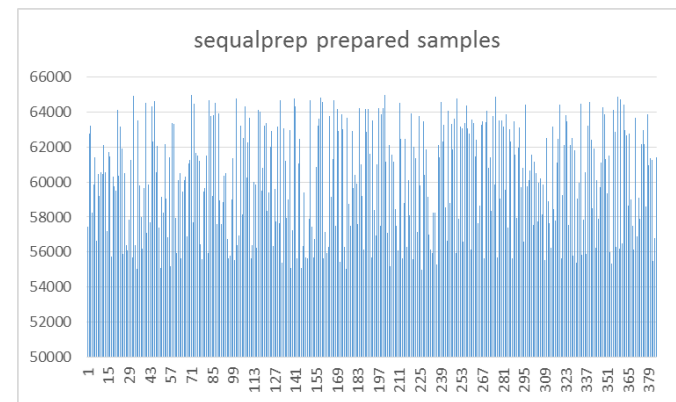
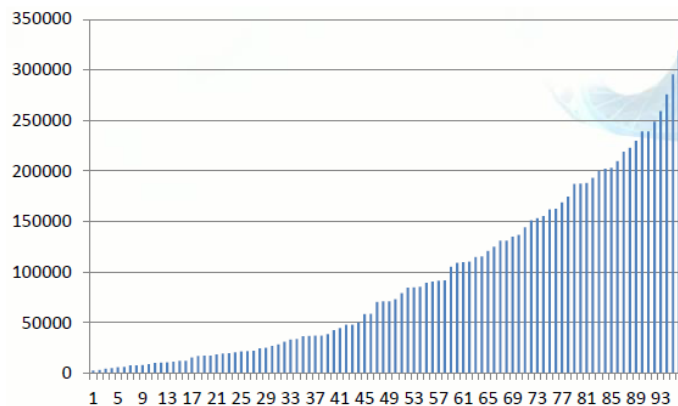
Part 4

Part 5

## Step 2 : combining amplicons as pools

How do I normalize the amplicons before pooling ?

the « what I can I do best for my precious ? » strategy, uses a sequalprep kit to retain known quantity of amplicon (25ng) then pool them



## Step 2 : combining amplicons as pools

It is advised not to mix different size amplicons as it would favor the shortest ones in the process (PCR if any during the library and sequencing).

1 pool = 1 library

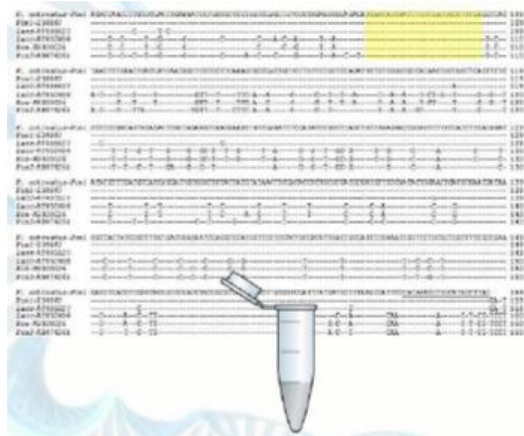
Should I do replicates ?

## Step 2 : combining amplicons as pools

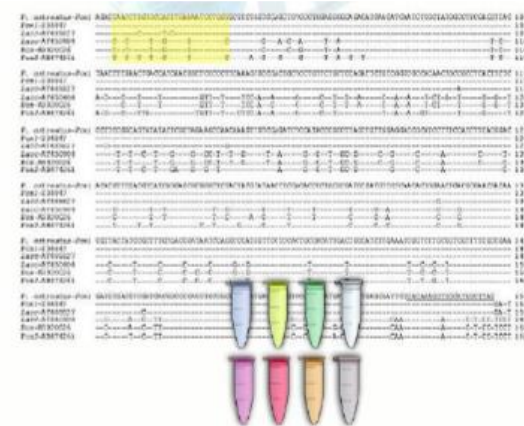
Common problems are linked to successfully amplify a wide range of species with an equal efficiency

Improve  
efficiency

Universal primers



Primers cocktail



# Amplicon sequencing

---

■	<b>Part 1</b>
	Part 2
	Part 3
	Part 4
	Part 5

## Step 2 : combining amplicons as pools

Gel purification of the pools

Qualifying the pools on bioanalyzer

Quantifying the pools with what is convenient (nanodrop)

Normalize pools to 0.4ng/μL (20 ng in 50μL)



## Step 3 : building libraries for each pool

This protocol uses part of the Truseq nano LT kit (4h)

1. End-repair to make the amplicons blunt-ended.
2. A-tailing to allow further ligation
3. Ligate the illumina adapters (holding indexes)
4. Enrich in P5-P7 libraries through amplification

# Amplicon sequencing

## Part 1

Part 2

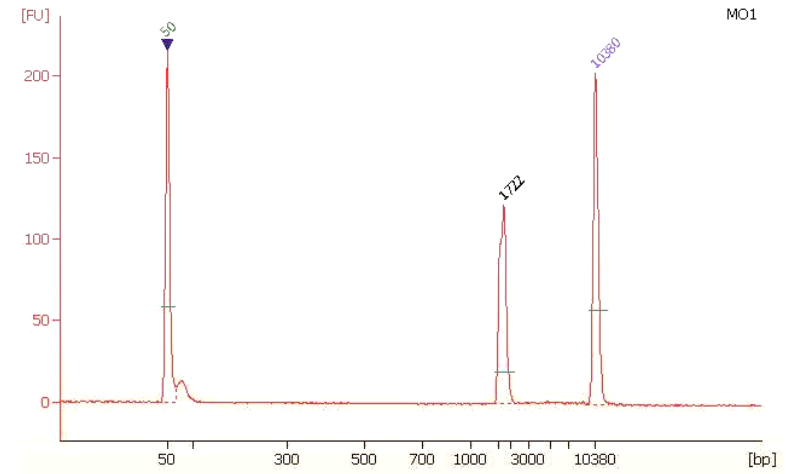
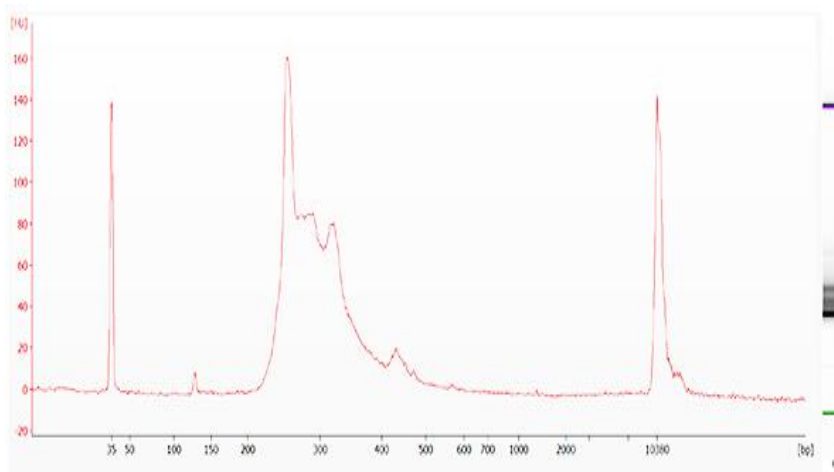
Part 3

Part 4

Part 5

## Step 3 : building libraries for each pool

When the libraries are built, I need to qualify them using the bioanalyzer



# Amplicon sequencing

## Part 1

Part 2

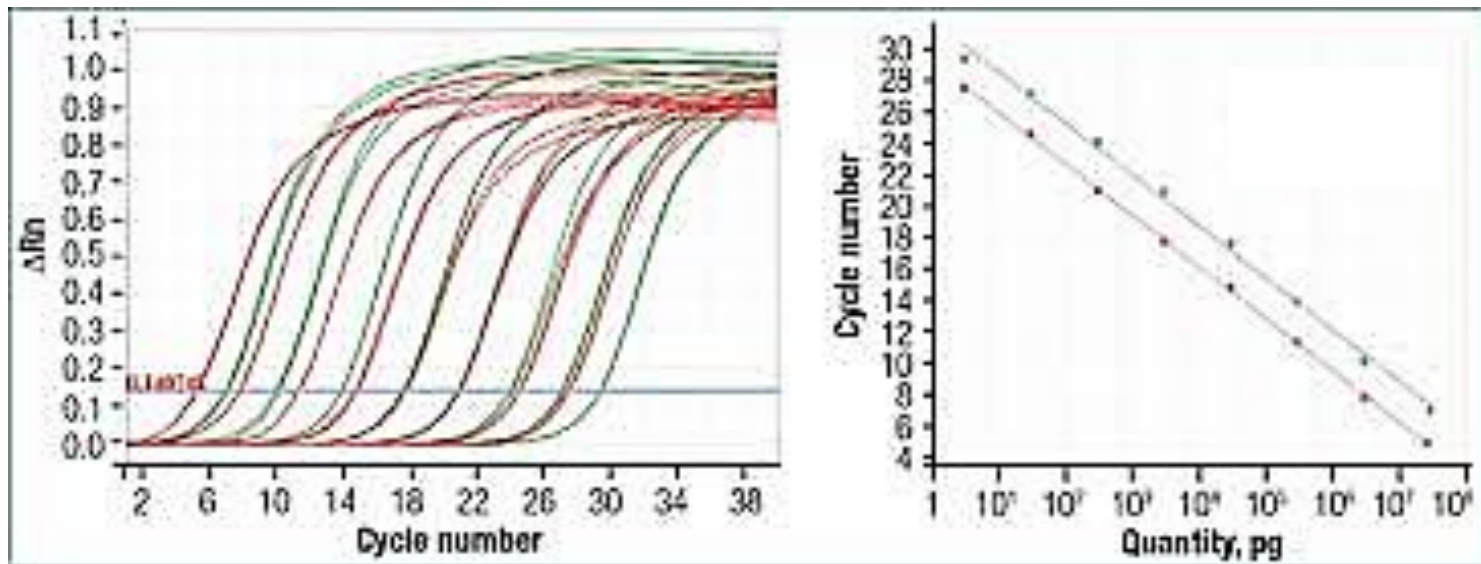
Part 3

Part 4

Part 5

## Step 3 : building libraries for each pool

The last step is to quantify the libraries with qPCR using adapter sequences as primers.



## Step 4 : Miseq sequencing

Preparing the sample (pooled libraries) involves dilution and denaturation in NaOH.

Setup of the sequencer takes around an hour (warming)

The run is 46 hours long

# Amplicon sequencing

## Part 1

Part 2

Part 3

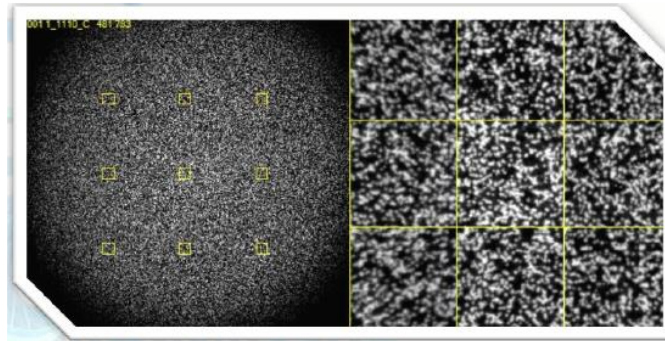
Part 4

Part 5

## Step 4 : Miseq sequencing

What can go wrong at the sequencing step?

- Overclusterization
- Bad quality
- Door opened..
- Road construction.....



# Amplicon sequencing

Part 1

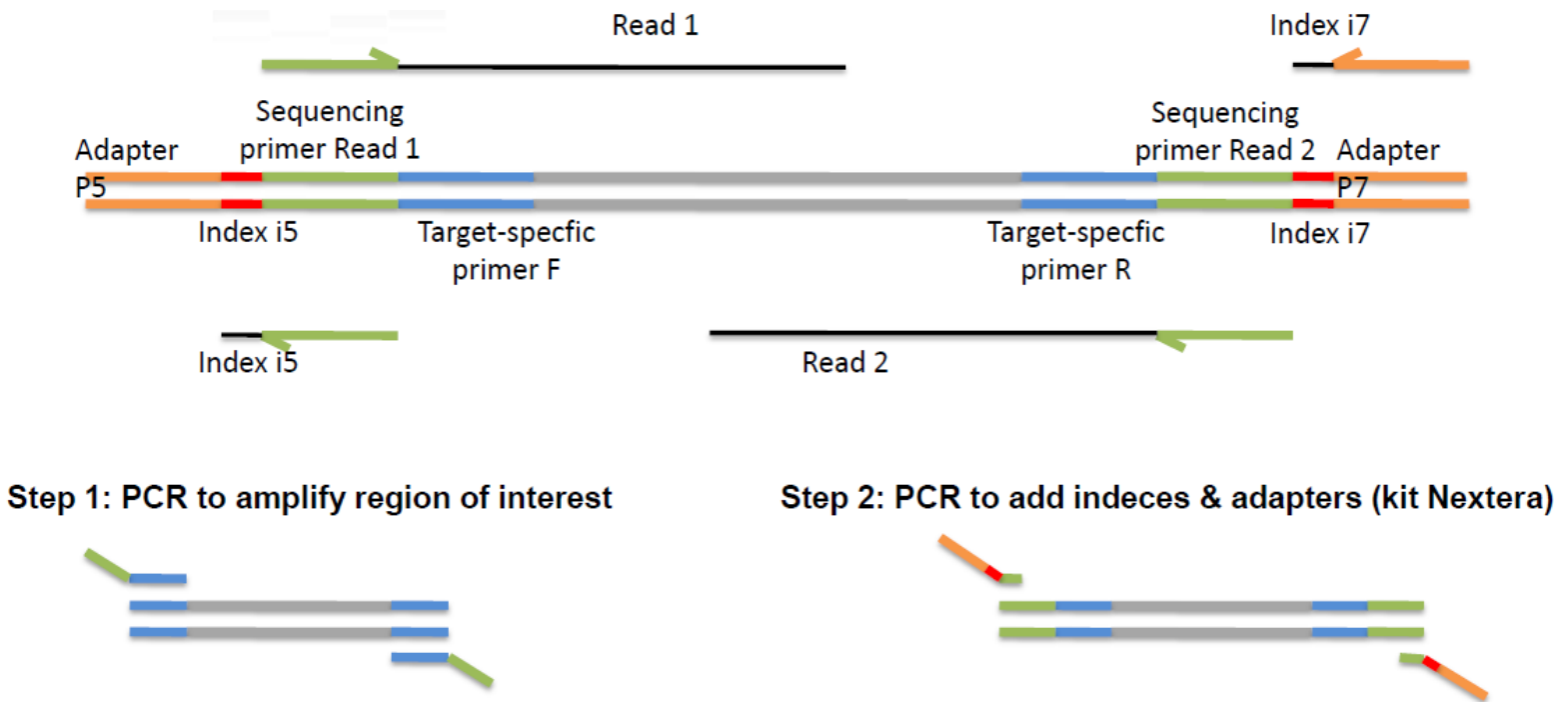
Part 2

Part 3

Part 4

Part 5

## Two other wet lab strategies 2-steps PCR



Advantages : multiple markers, universal and easy

Drawbacks : 2 steps PCR costs, biases and contamination issue

# Amplicon sequencing

Part 1

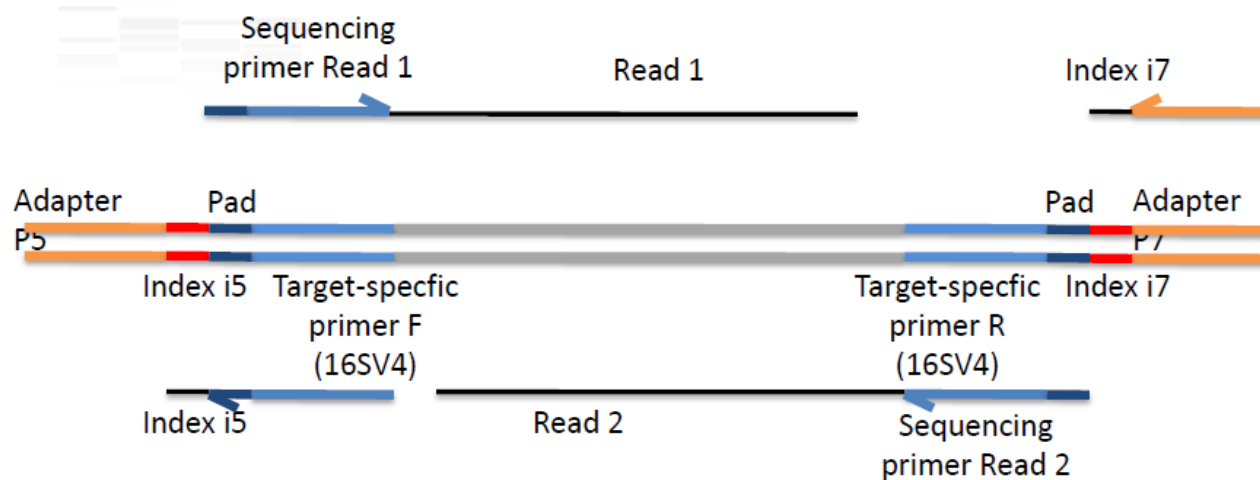
Part 2

Part 3

Part 4

Part 5

## Two other wet lab strategies : 1-step PCR



### Step 1: PCR to amplify region of interest



Advantages : 1 step PCR, no primer sequencing

Drawbacks : add custom primers into the sequencer cartridge, only 16S for now

# Amplicon sequencing

## Part 1

Part 2

Part 3

Part 4

Part 5

## comparing wet lab strategies

	Protocole 2-step PCR	Protocole 1-step PCR	Protocole tag PCR - Truseq
Few markers and < 1728 samples	+++	---	
Metabarcoding 16S (bacteria)	+	+++	
Multiple projects, > 1728 samples		---	+++



# Amplicon sequencing

## Part 1

Part 2

Part 3

Part 4

Part 5



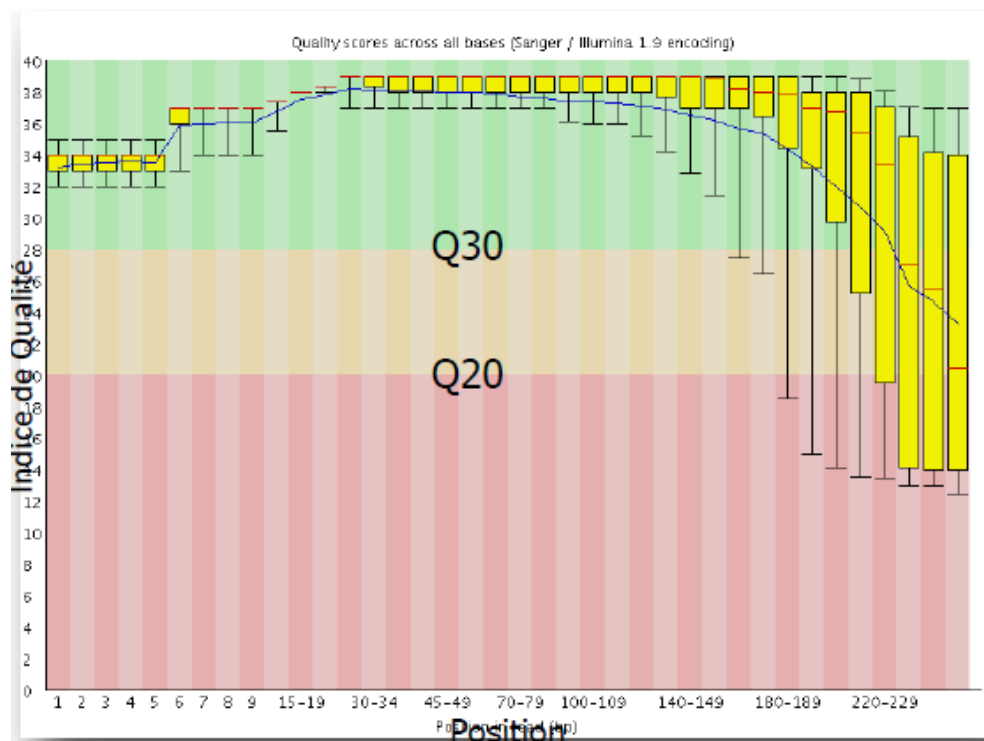
Data analysis, from raw data to usable data

General steps:

1. quality assessment
2. filtering using quality
3. contig of paired reads
4. demultiplexing the markers if relevant
5. demultiplexing the amplicons
6. characterizing variants
7. Aligning variants if necessary
8. assigning to a reference library

## Data analysis, from raw data to usable data

### 1. quality assessment (Galaxy)



Data analysis, from raw data to usable data

2. Filtering the reads (Galaxy)
3. Contig the paired reads (Mothur) and prepare a fasta file

From now, everything is done with

## MOLECULAR ECOLOGY RESOURCES

Molecular Ecology Resources (2012) 12, 1151–1157

doi: 10.1111/j.1755-0998.2012.03171.x

### |SE|S|AM|E| Barcode: NGS-oriented software for amplicon characterization – application to species and environmental barcoding

S. PIRY,\* E. GUIVIER,\* A. REALINI† and J.-F. MARTIN†

\*INRA (UMR CBGP Centre de Biologie Pour la Gestion des Populations), Campus international de Baillarguet, CS 30016, F 34988  
Montferrier sur Lez Cedex, France, †Montpellier SupAgro (UMR CBGP Centre de Biologie Pour la Gestion des Populations),  
Campus international de Baillarguet, CS 30016, F 34988 Montferrier sur Lez Cedex, France

Could also be done with



# Amplicon sequencing

Part 1

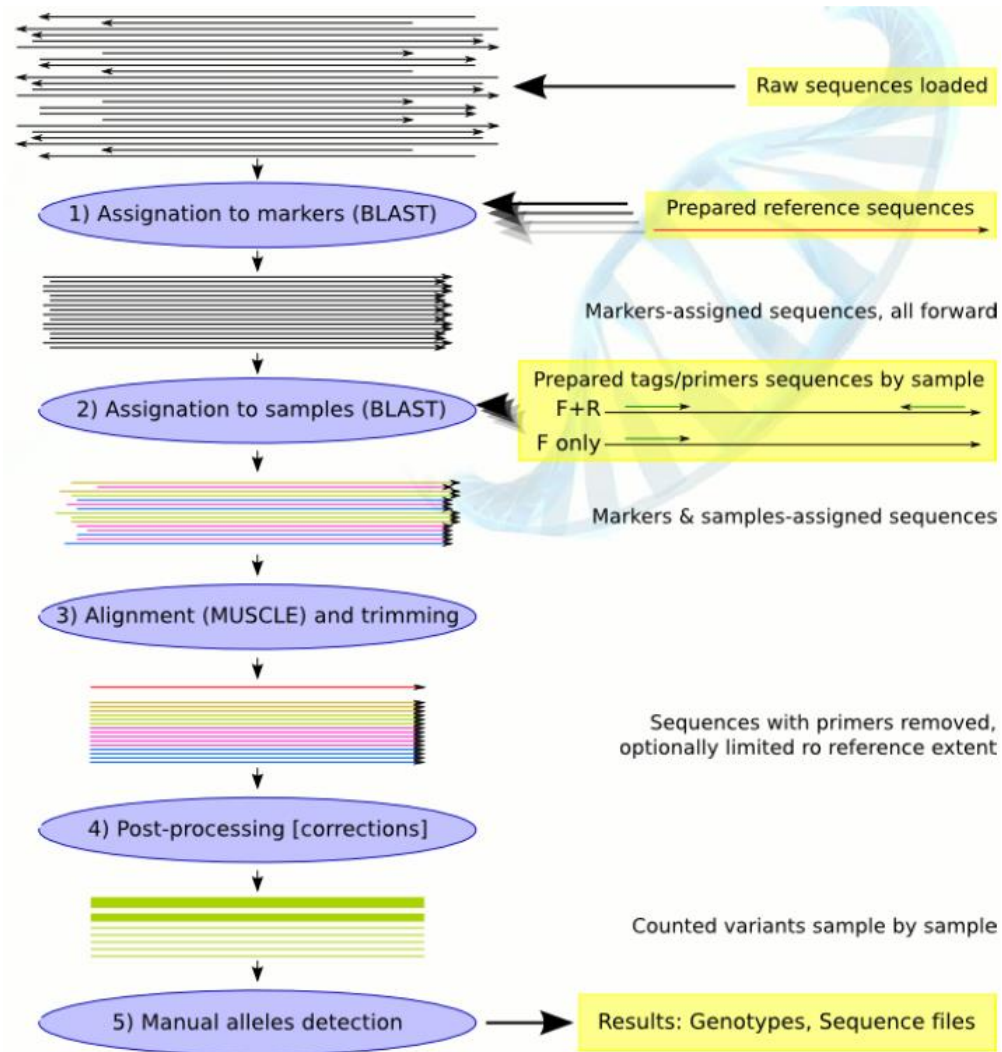
**Part 2**

Part 3

Part 4

Part 5

## The pipeline



# Amplicon sequencing

Part 1

**Part 2**

Part 3

Part 4

Part 5

## The sequences alignment interface (if needed)

INRA Montpellier SupAgro © 2010-2011 Sylvain Piry, CBGP-INRA & Emese Meglecz, Université de Provence ; Jean-François Martin, CBGP-SupAgro

**SESAME** SEquence Sorter & Amplicon Explorer

Logout Login piry

Refresh projects Select a project **Manu** Global filters : Run Max Démo Marker

Markers Runs Samples **Sequences** Alleles Results Administration Help

Refresh Export to CSV View sample

Samples filter (run & marker above):

Name	Species	Population	Standard
	Mus		

Sample name	Std	Marker name	ploidy	Run name	# alleles	# valid sequen	# variants	Sp
L0209		DRB	2	Max Démo		282	55	M
L0210		DRB	2	Max Démo		273	50	M
L0211		DRB	2	Max Démo		119	17	M
L0212		DRB	2	Max Démo		189	39	M
R4362c		DRB	2	Max Démo		189	41	M
R4952c		DRB	2	Max Démo		77	20	M
R4955c		DRB	2	Max Démo		14	5	M
R4984		DRB	2	Max Démo		984	174	M
R4986		DRB	2	Max Démo		124	15	M

Use reference ☒ Reference ☒ Allele ☐ Trash ☐ Min. number of seq. 1 Seq. width

Sequences of this sample

Is allele	Allele ID	# seq	Frequency	# samp	# seq in run	Is trash	Length	10	20	30
		106	0.39	32	2785		172	GCAGCGGTCGG-TTTCTGGAAGATTTCATCTACAAC		
		101	0.37	19	1902		172	.....C.....A.T.....T.....		
		4	0.01	16	1178		172	.....A.....A.....TC.....GA.....		
		4	0.01	8	346		172	.....A.....A.....TC.....T.....		
		4	0.01	6	13		172	.....		
		2	0.01	16	27		173	.....		
		2	0.01	10	16		171	.....		
		2	0.01	9	13		172	.....		
		2	0.01	9	12		173	.....		
		2	0.01	7	10		173	.....C.....A.T.....T.....		
		2	0.01	5	10		172	.....C.....A.T.....T.....		
		2	0.01	3	8		172	.....C.....A.T.....T.....		
		2	0.01	2	3		172	.....		
		2	0.01	1	2		172	.....		

# Amplicon sequencing

Part 1

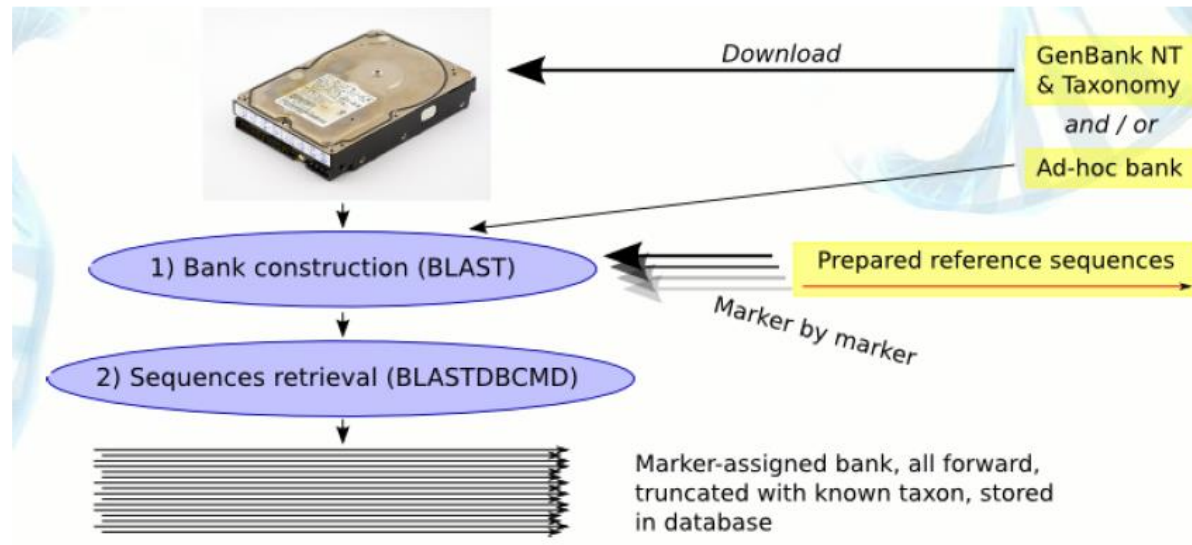
**Part 2**

Part 3

Part 4

Part 5

## Preparing reference library





# Amplicon sequencing

Part 1

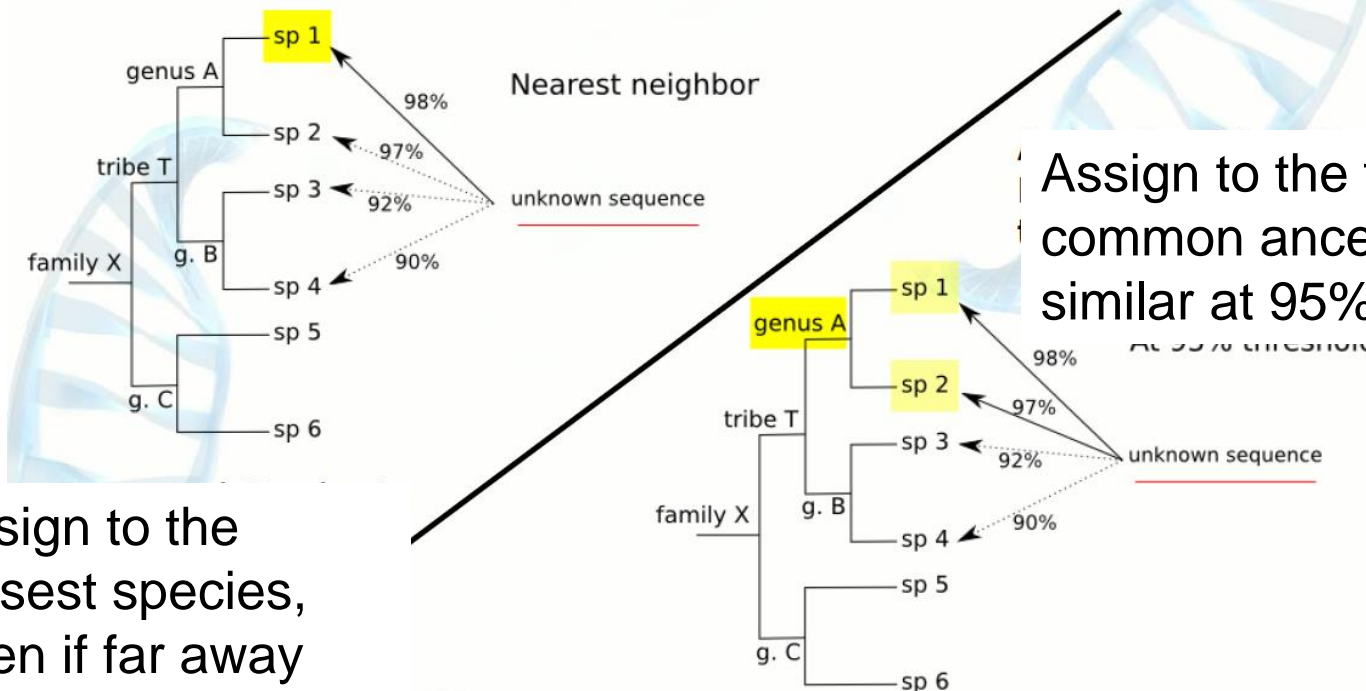
**Part 2**

Part 3

Part 4

Part 5

## Assignment to the library



Assign to the closest species, even if far away

# Amplicon sequencing

Part 1  
**Part 2**  
 Part 3  
 Part 4  
 Part 5

## Exploring results

**ISE|S|AM|E| Sequence Sorter & Amplicon Explorer (Barcode)**  
 ver. 1.18 build 1394

© 2010-2011 Sylvain Piry, CBGP-INRA & Emese Meglécz, Université de Provence; Jean-François Martin, CBGP-SupAgro

Project: **aphids**

Global filters: Run: MIX noised 0.2 Marker: [dropdown]

Similarity threshold: [dropdown] Min. % similarity: 90

Nearest neighbor: [dropdown]

Results grid (double-click for details)

Sample name	Marker name	Run name	# variants	# barcodes
1008.1	col	MIX noised 0.2	4	4
1008.2	col	MIX noised 0.2	3	3
1008.3	col	MIX noised 0.2	2	2
1008.4	col	MIX noised 0.2	8	8
1008.5	col	MIX noised 0.2	3	3
1008.6	col	MIX noised 0.2	2	2
1008.7	col	MIX noised 0.2	5	5
1008.8	col	MIX noised 0.2	3	3
1008.9	col	MIX noised 0.2	4	4
1008.10	col	MIX noised 0.2	3	3
1008.11	col	MIX noised 0.2	4	4
1008.12	col	MIX noised 0.2	6	6
1008.13	col	MIX noised 0.2	2	2
1008.14	col	MIX noised 0.2	4	4
1008.15	col	MIX noised 0.2	3	3
MHB.1	col	MIX noised 0.2	8	8

Results image

Project: aphids - Run: MIX noised

Sequences: 1 - Values: number of

Detected taxa, samples and sequences

Graphical results

List of samples (multiple sel.)

## Concluding remark about future directions

Improving PCR of  
alternative approaches



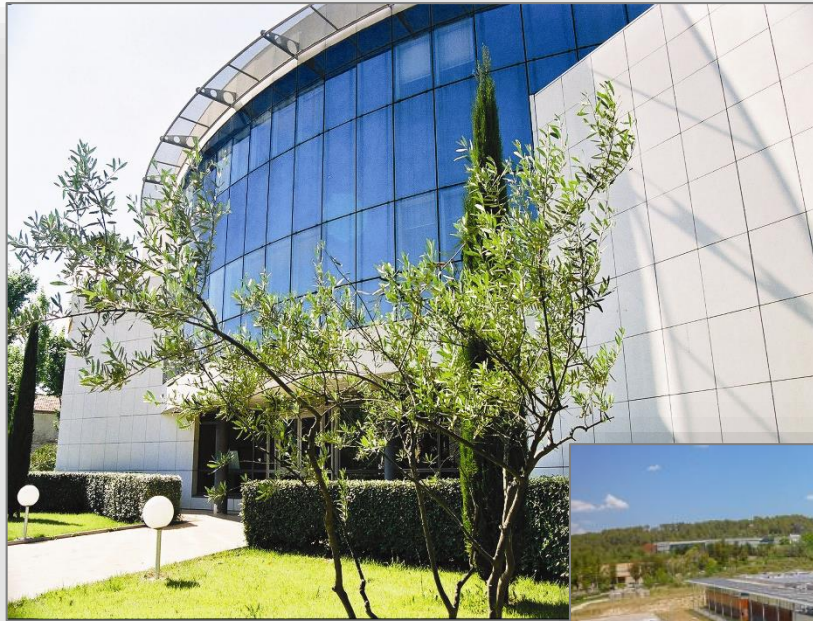
Be aware of the  
continuous evolution  
of technologies

Work on the scalability of  
bioinformatics solutions

# Acknowledgements

- Morgane Ardisson
- Anne-Laure Clamens
- Armelle Cœur d'Acier
- Emmanuel Corse
- Vincent Dubut
- Philippe Gauthier
- André Gilles
- Emmanuel Guivier
- Emese Meglecz
- Grégory Mollot
- Sylvain Piry
- Audrey Réalini





Centre de Biologie pour la Gestion des Populations