NGS technologies approaches, applications and challenges!

Jean-François Martin

Centre de Biologie pour la Gestion des Populations

Centre international d'études supérieures en sciences agronomiques

Who am I? Why am I here?

I am an associate professor in genetics and ecology

Interested in adaptation at the genome level (butterfly / fish)

I could probaby define myself as an experienced user / wet lab developper playing with NGS in the biodiversity field

Goals and expectations

The aim of this discussion today : make sure that everyone is on the same page with regards to NGS approaches

It is also to guide the ones begining with NGS through my own experience -> interaction !

What NGS changes for biologists

General improvements and changes

The history of NGS development techniques is young (around 10 years)

It is characterized by general trends

- more and more sequences
- and /or longer sequences
- diminishing prices

General improvements and changes

From a few 1kb sanger sequences to hundreds of millions reads

This shift in data acquisition has direct an undirect consequences on lab's life.

General improvements and changes

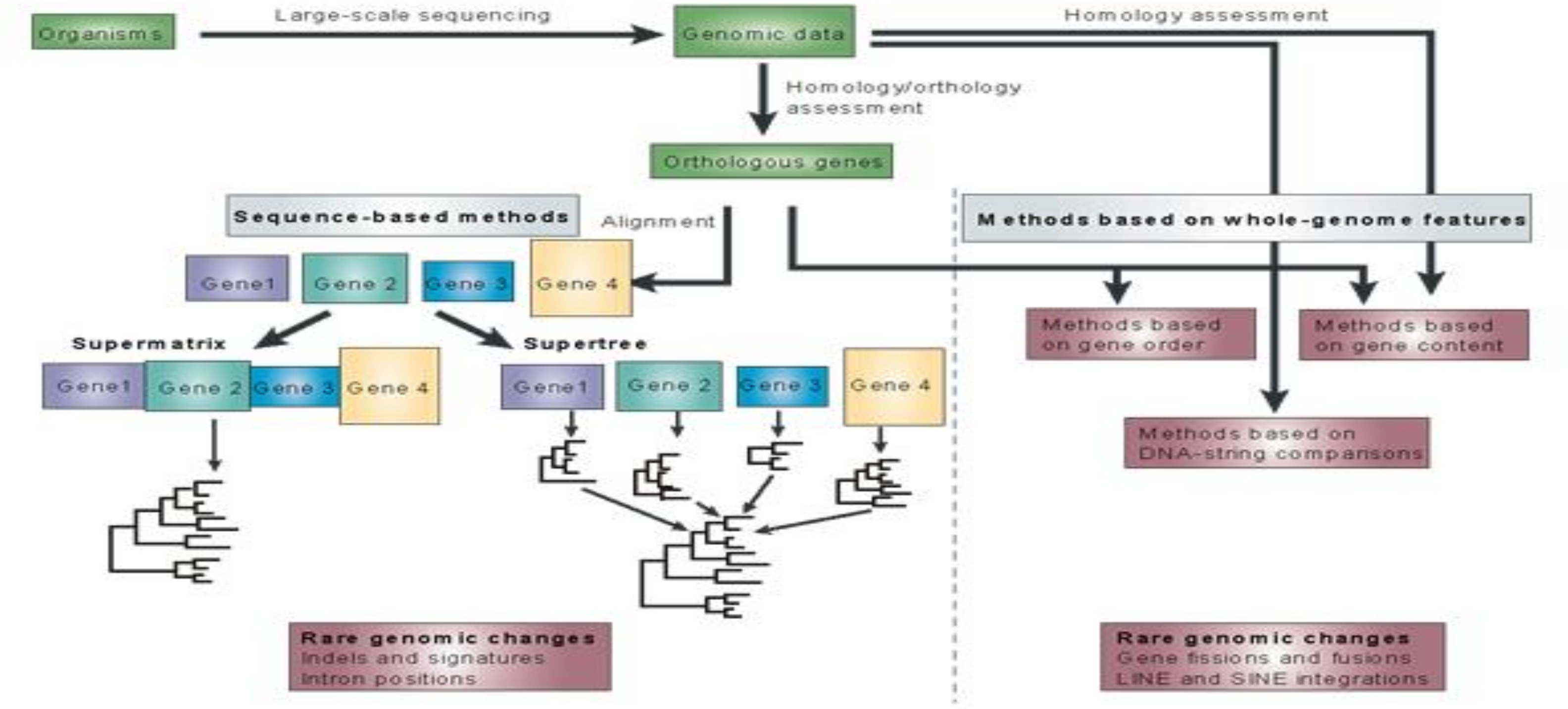
Important parameters for the available technologies:

- Length
- Quantity of reads
- Quality of the reads
- Price ?

What NGS changes for biologists

Long standing scientific questions that can be addressed

Improving phylogenies through multiple markers



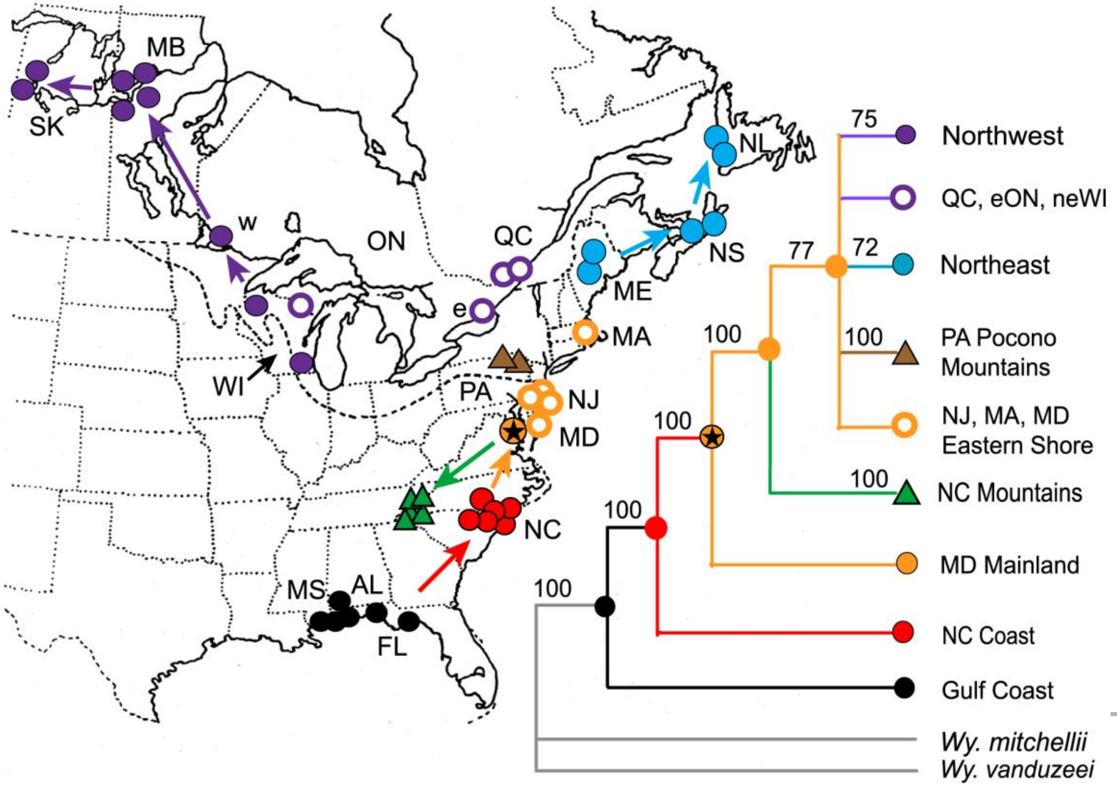
Nature Reviews | Genetics

By Jean-François Martin

What NGS changes for biologists

Long standing scientific questions that can be addressed

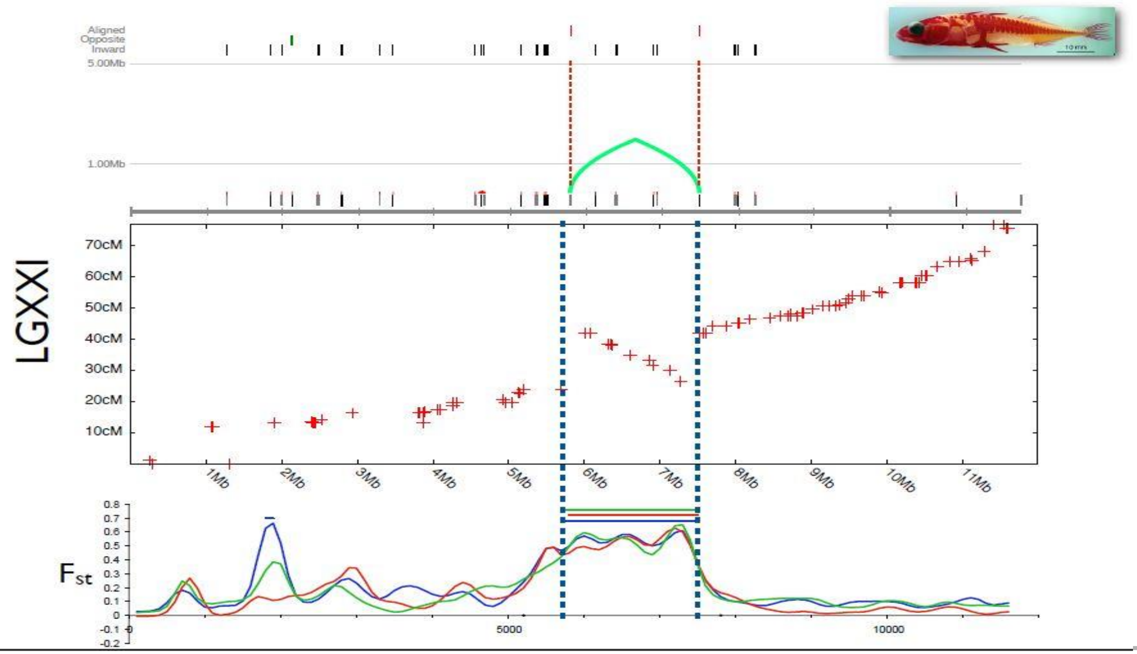
Resolving phylogeography and relationships in species complexes



What NGS changes for biologists

Long standing scientific questions that can be addressed

Testing selection and demography scenarios



What NGS changes for biologists

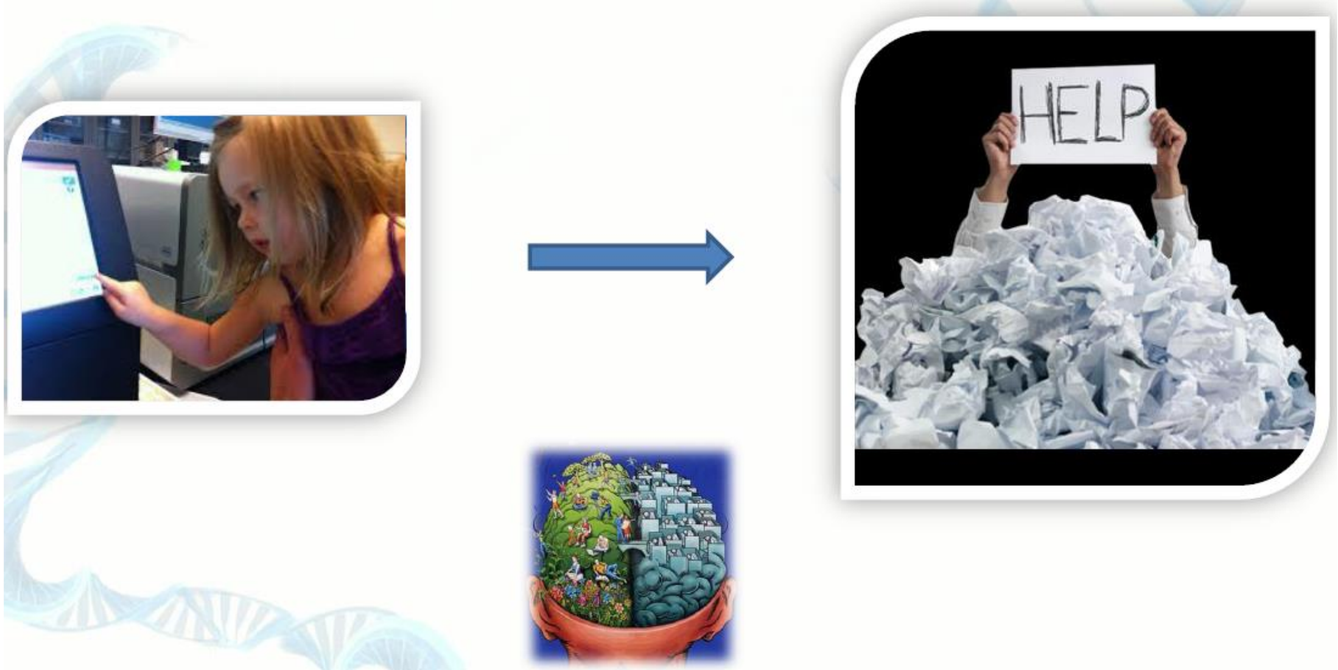
Long standing scientific questions that can be addressed

From population genetics to population genomics in general

Basically, analyzing genomes in interaction with their environment is now feasible and accessible to anyone

What NGS changes for biologists

Basically, analyzing genomes in interaction with their environment is now feasible and accessible to anyone

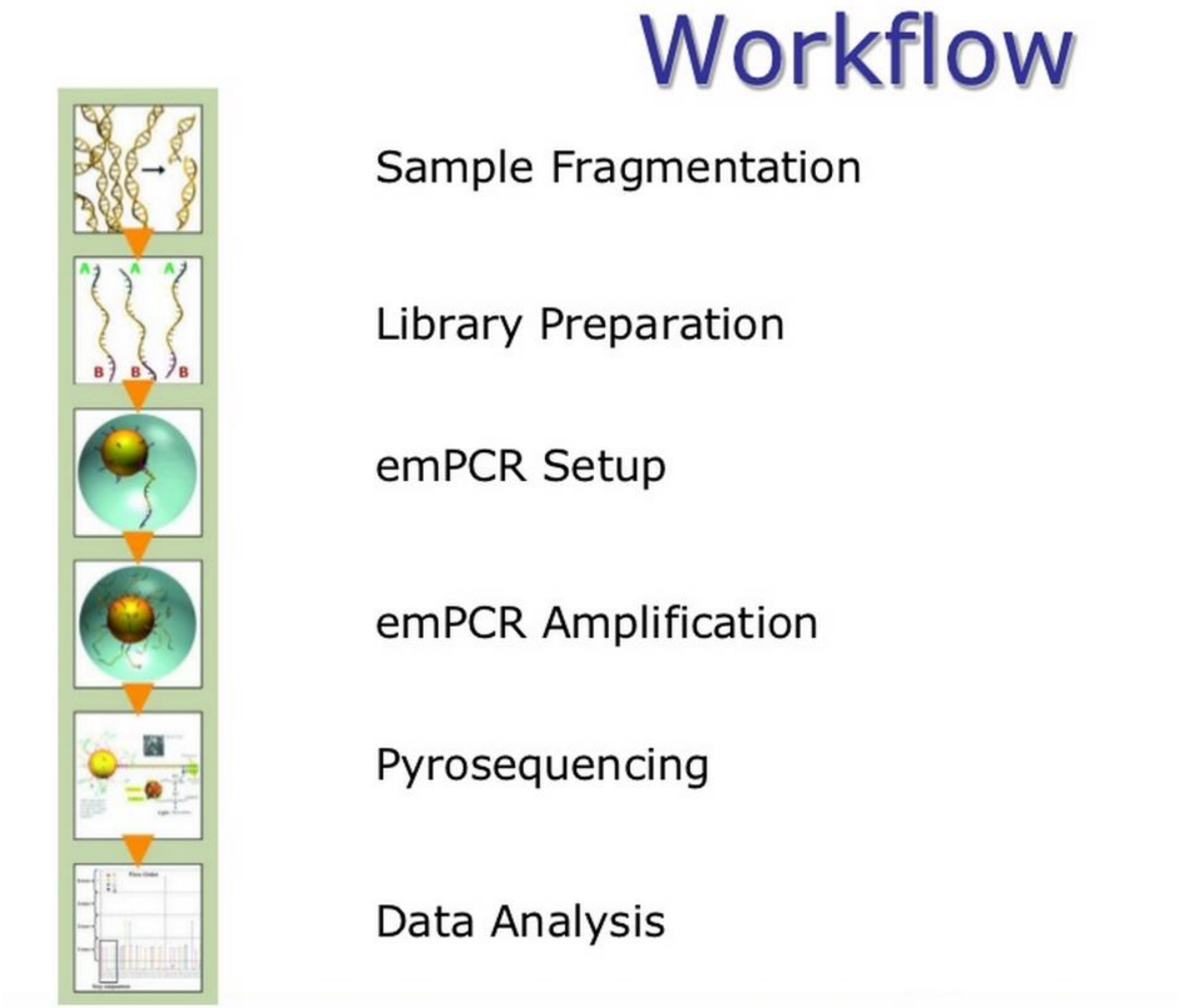


Current technologies & perspectives

Currently available technologies

Roche 454

Current technologies & perspectives



Workflow

Sample Fragmentation

Library Preparation

emPCR Setup

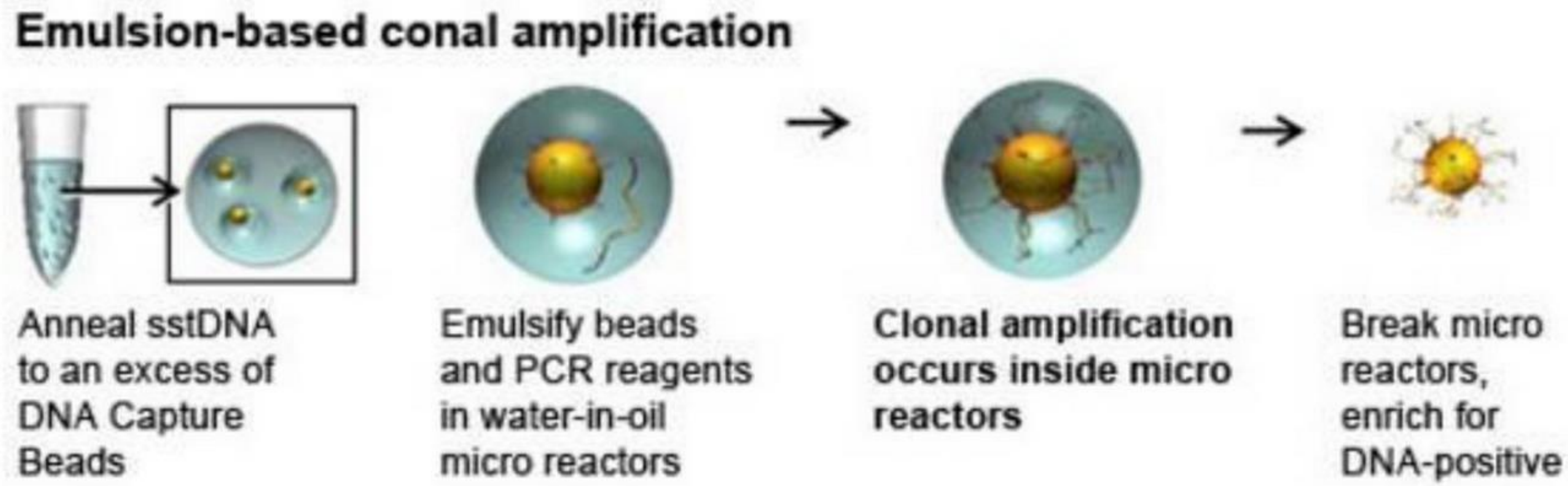
emPCR Amplification

Pyrosequencing

Data Analysis

emPCR

Emulsion PCR is a method of clonal amplification which allows for millions of unique PCRs to be performed at once through the generation of micro-reactors.



Emulsion-based conal amplification

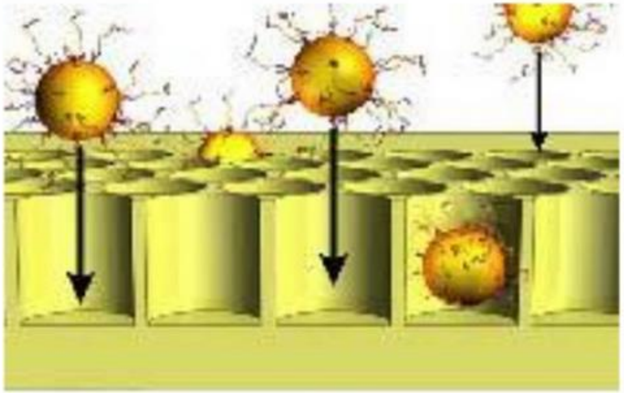
Anneal sstDNA to an excess of DNA Capture Beads

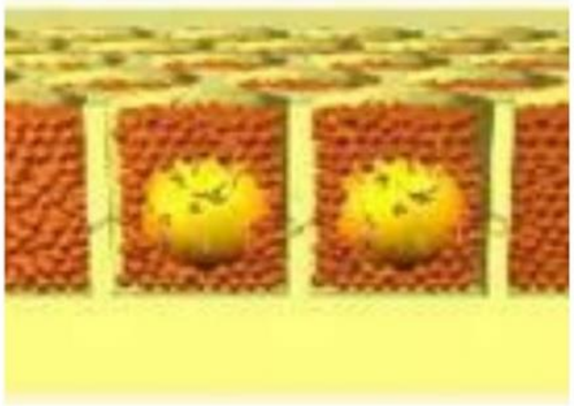
Emulsify beads and PCR reagents in water-in-oil micro reactors

Clonal amplification occurs inside micro reactors

Break micro reactors, enrich for DNA-positive

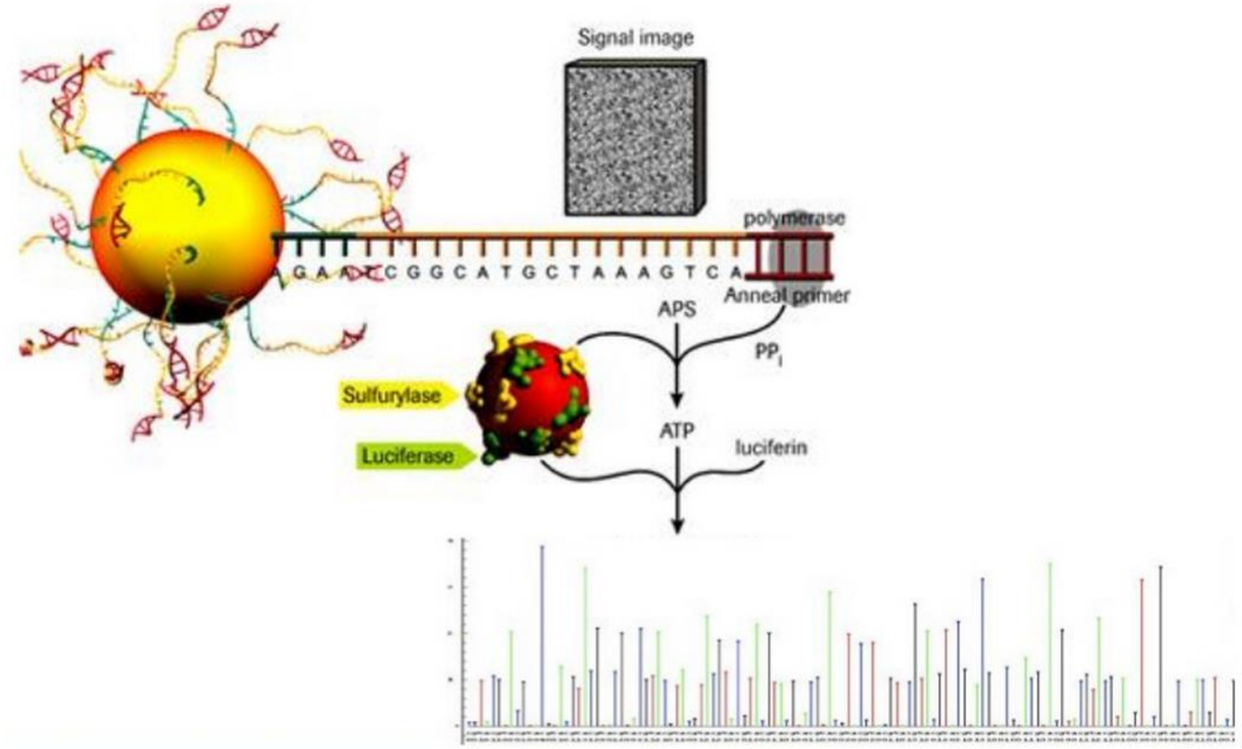
Massively Parallel Sequencing

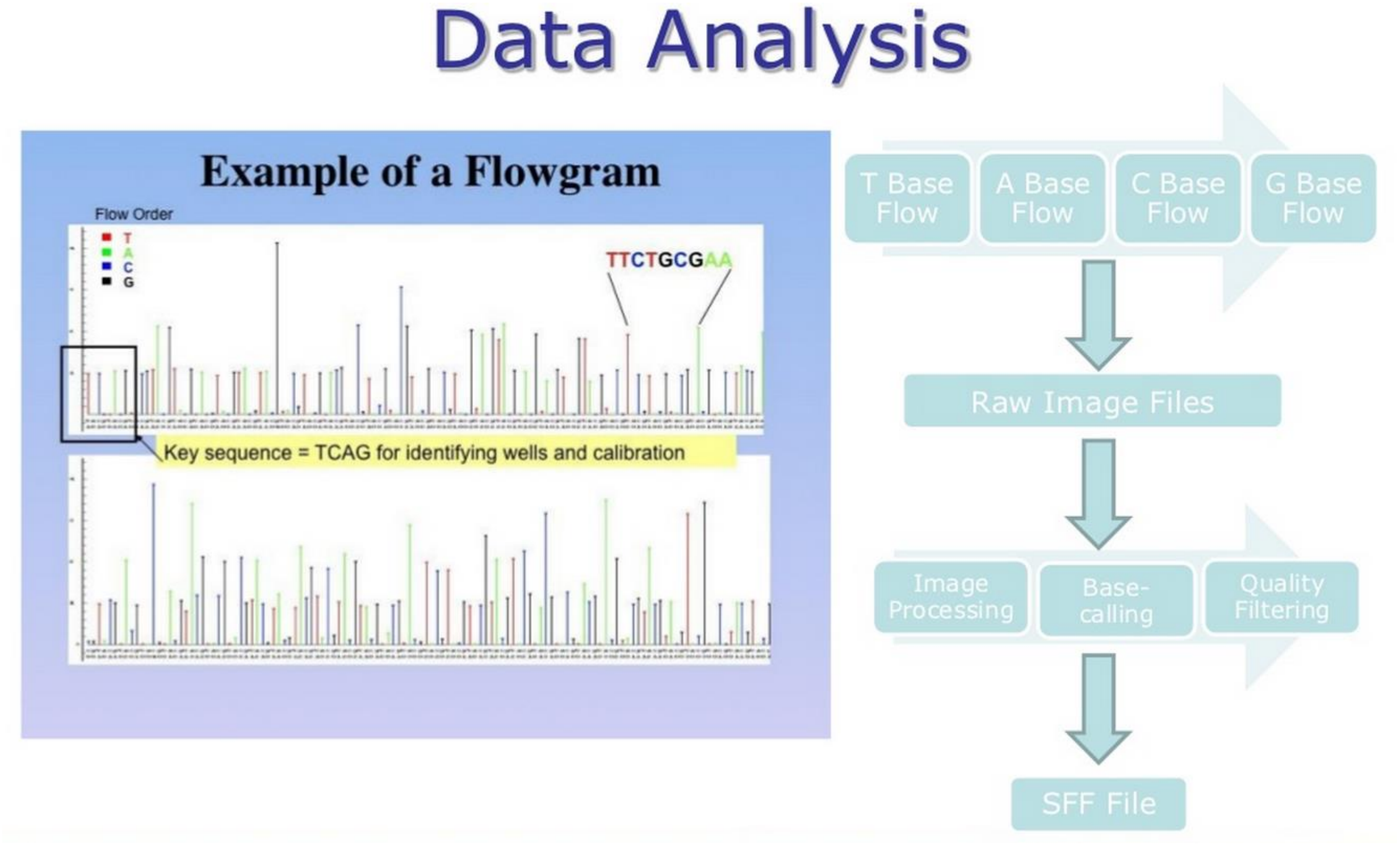




Current technologies & perspectives

Pyrosequencing





Data Analysis

Example of a Flowgram

Key sequence = TCAG for identifying wells and calibration

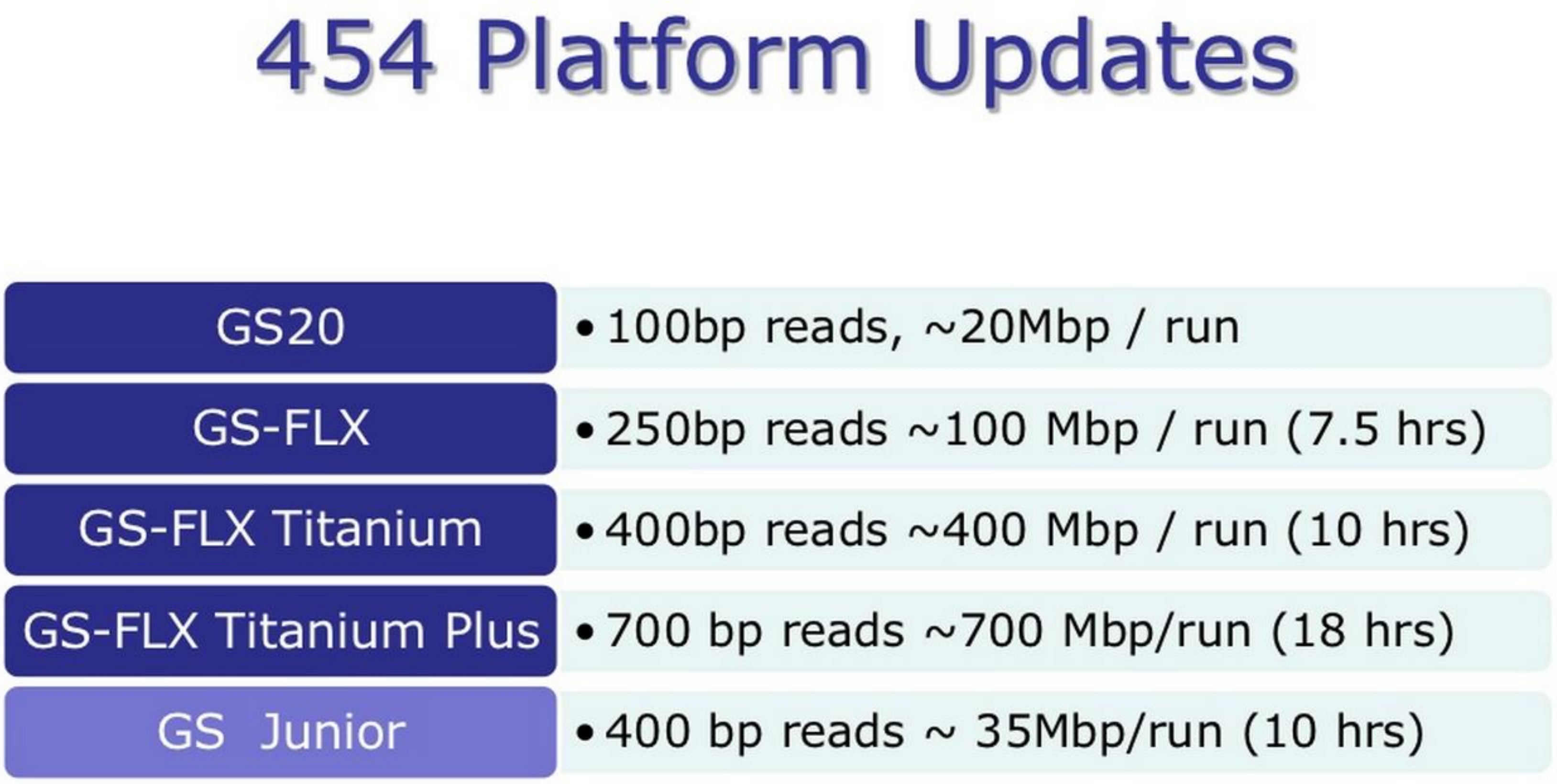
T Base Flow → A Base Flow → C Base Flow → G Base Flow

Raw Image Files

Image Processing → Base-calling → Quality Filtering

SFF File

Current technologies & perspectives



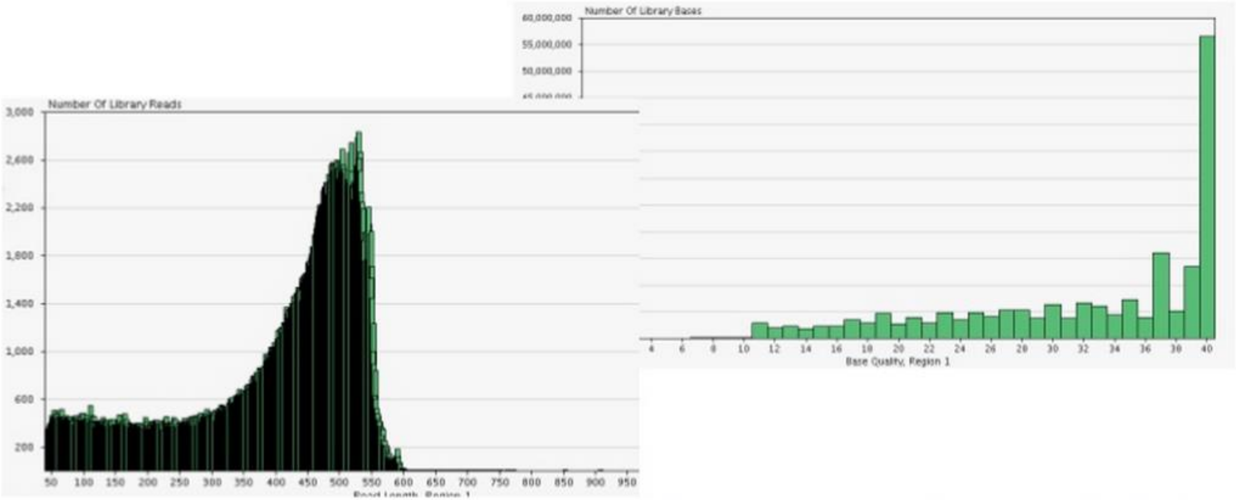
454 Platform Updates

GS20	• 100bp reads, ~20Mbp / run
GS-FLX	• 250bp reads ~100 Mbp / run (7.5 hrs)
GS-FLX Titanium	• 400bp reads ~400 Mbp / run (10 hrs)
GS-FLX Titanium Plus	• 700 bp reads ~700 Mbp/run (18 hrs)
GS Junior	• 400 bp reads ~ 35Mbp/run (10 hrs)

Current technologies & perspectives

454 Sequencing Output

- *.sff *(standard flowgram format)*
- *.fna *(fasta)*
- *.qual *(Phred quality scores)*



Current technologies & perspectives

So 454 is well adapted when long sequences are needed or at least beneficial?

Yes !

But no

Current technologies & perspectives

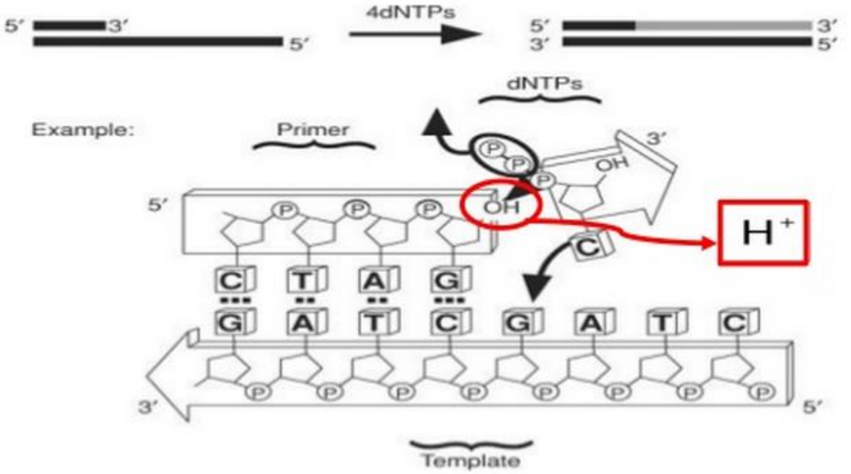
Currently available technologies

Ion torrent

Applied Biosystems:
Ion Torrent PGM

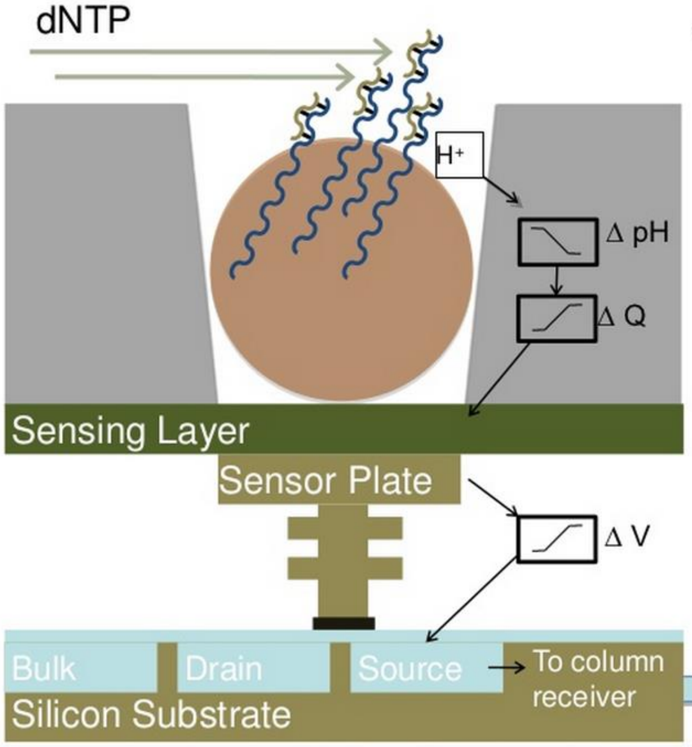
Ion Torrent

- Ion Semiconductor Sequencing
- Detection of hydrogen ions during the polymerization DNA
- Sequencing occurs in microwells with ion sensors
- No modified nucleotides
- No optics

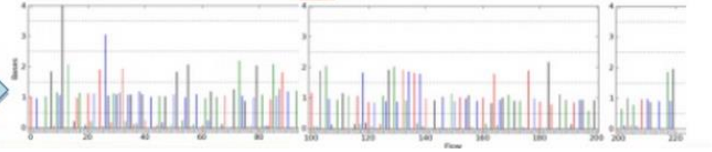


Current technologies & perspectives

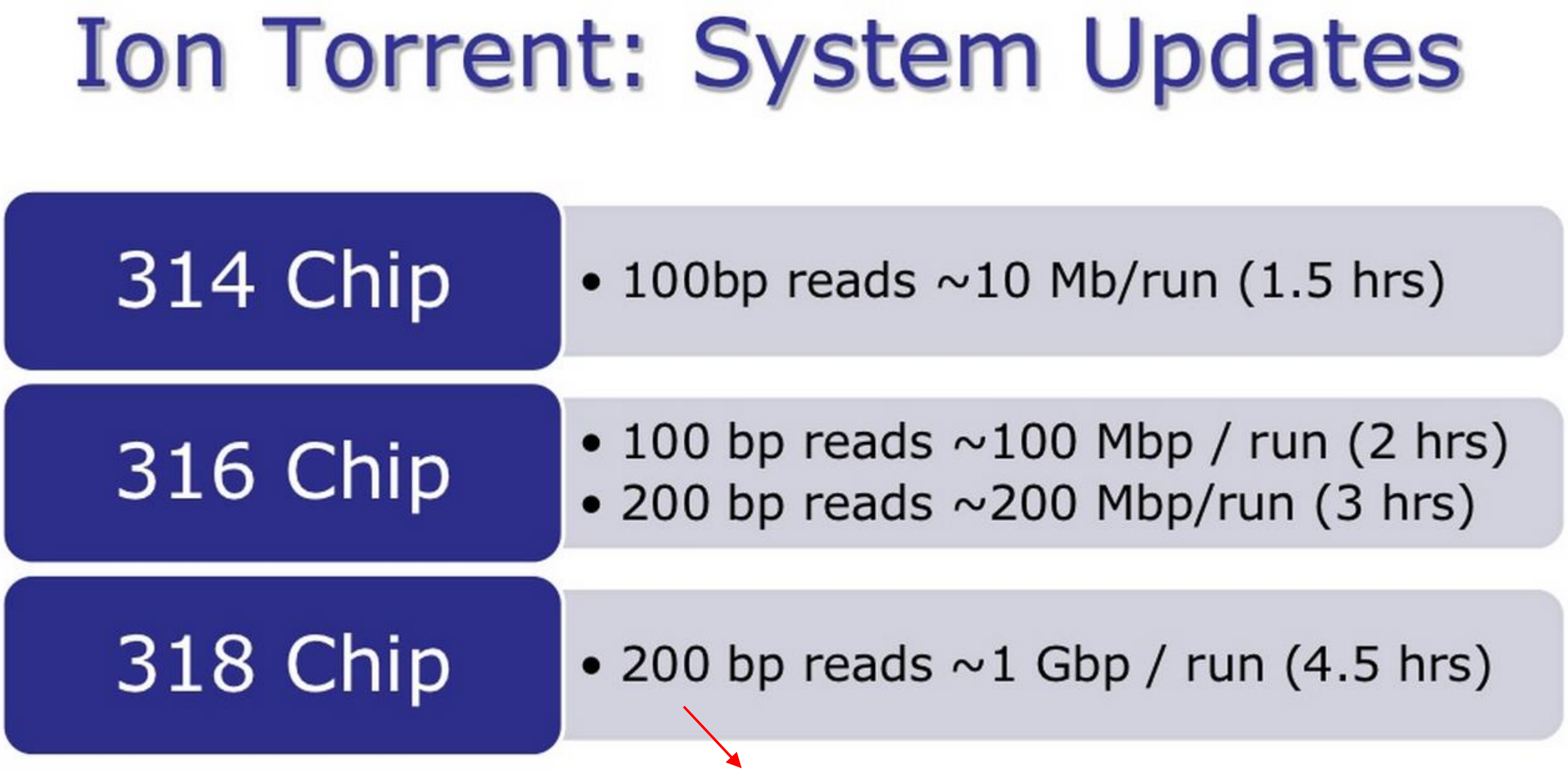
Ion Torrent



- DNA → Ions → Sequence
  - Nucleotides flow sequentially over Ion semiconductor chip
  - One sensor per well per sequencing reaction
  - Direct detection of natural DNA extension
  - Millions of sequencing reactions per chip
  - Fast cycle time, real time detection



Current technologies & perspectives



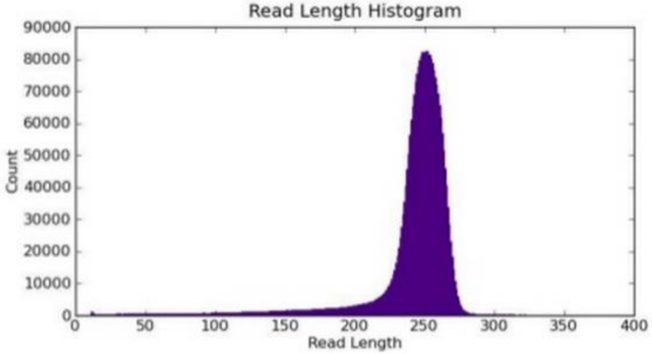
Ion Torrent: System Updates

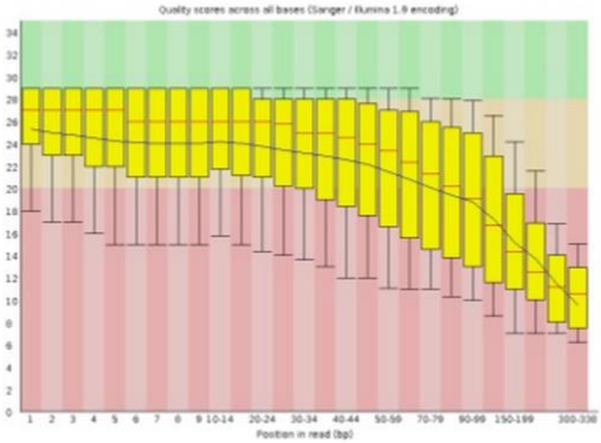
314 Chip	• 100bp reads ~10 Mb/run (1.5 hrs)
316 Chip	• 100 bp reads ~100 Mbp / run (2 hrs) • 200 bp reads ~200 Mbp/run (3 hrs)
318 Chip	• 200 bp reads ~1 Gbp / run (4.5 hrs)

400 bp reads

Ion Torrent Reads

- *.sff *(standard flowgram format)*
- *.fastq (*sequence and corresponding quality score encoded with an ASCII character, phred-like quality score + 33)*





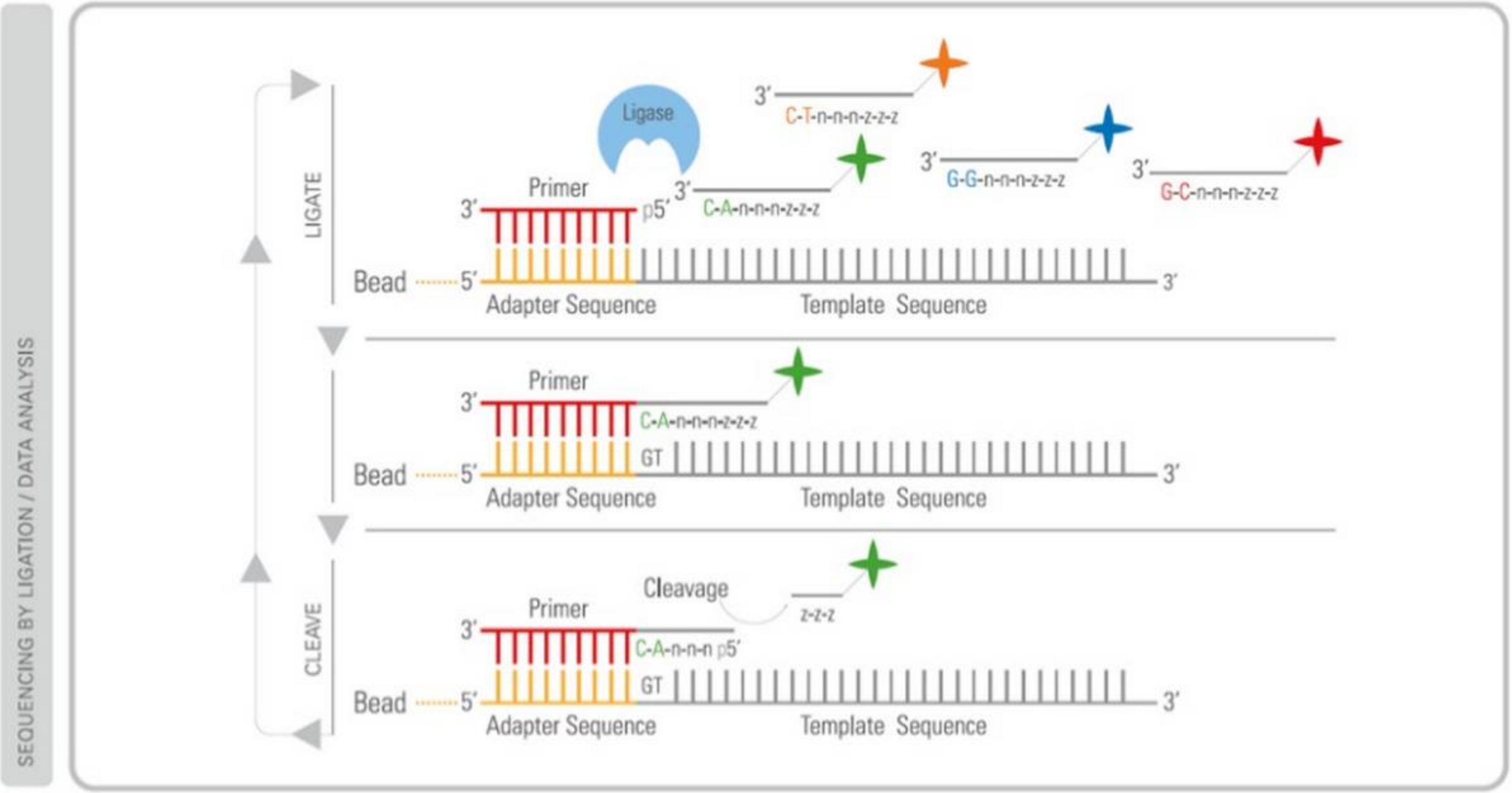
Current technologies & perspectives

Currently available technologies

Applied Biosystems SOLiD

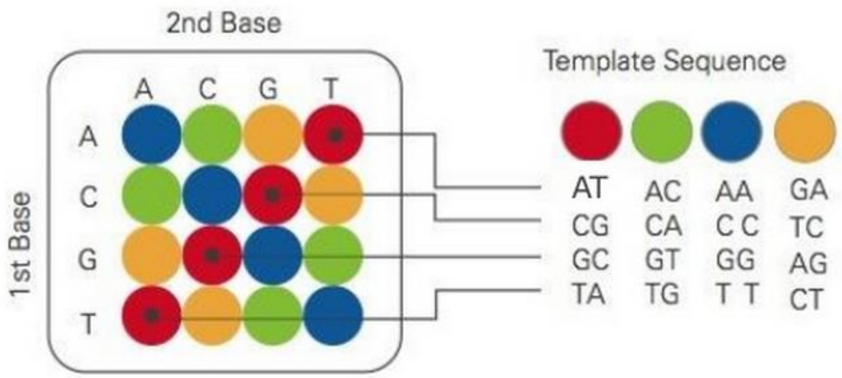
SOLID

Sequencing by Ligation



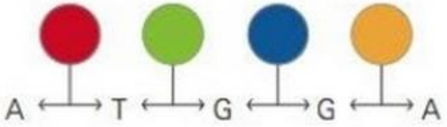
2 Base encoding

Possible Dinucleotides Encoded By Each Color

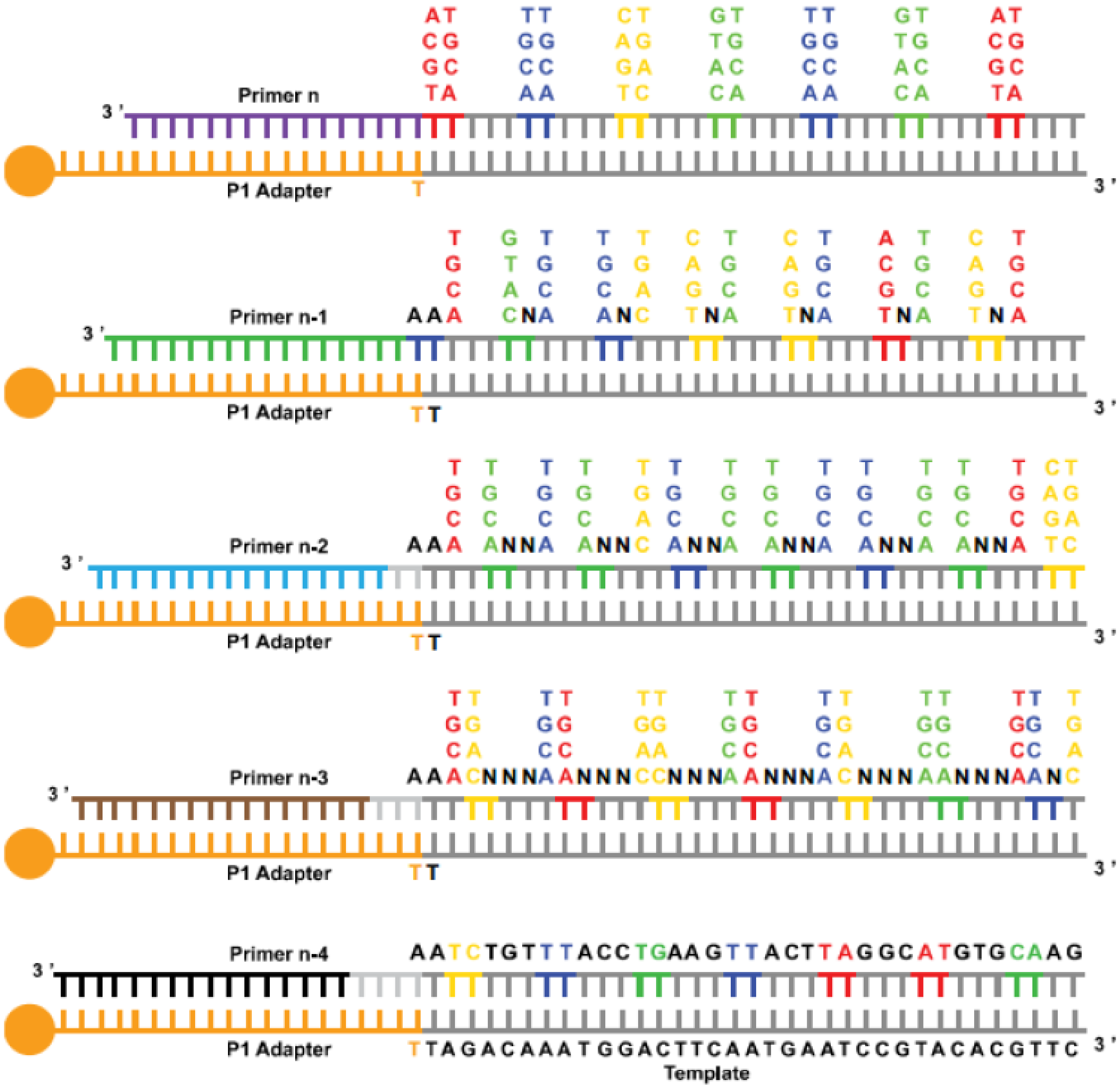


Double Interrogation

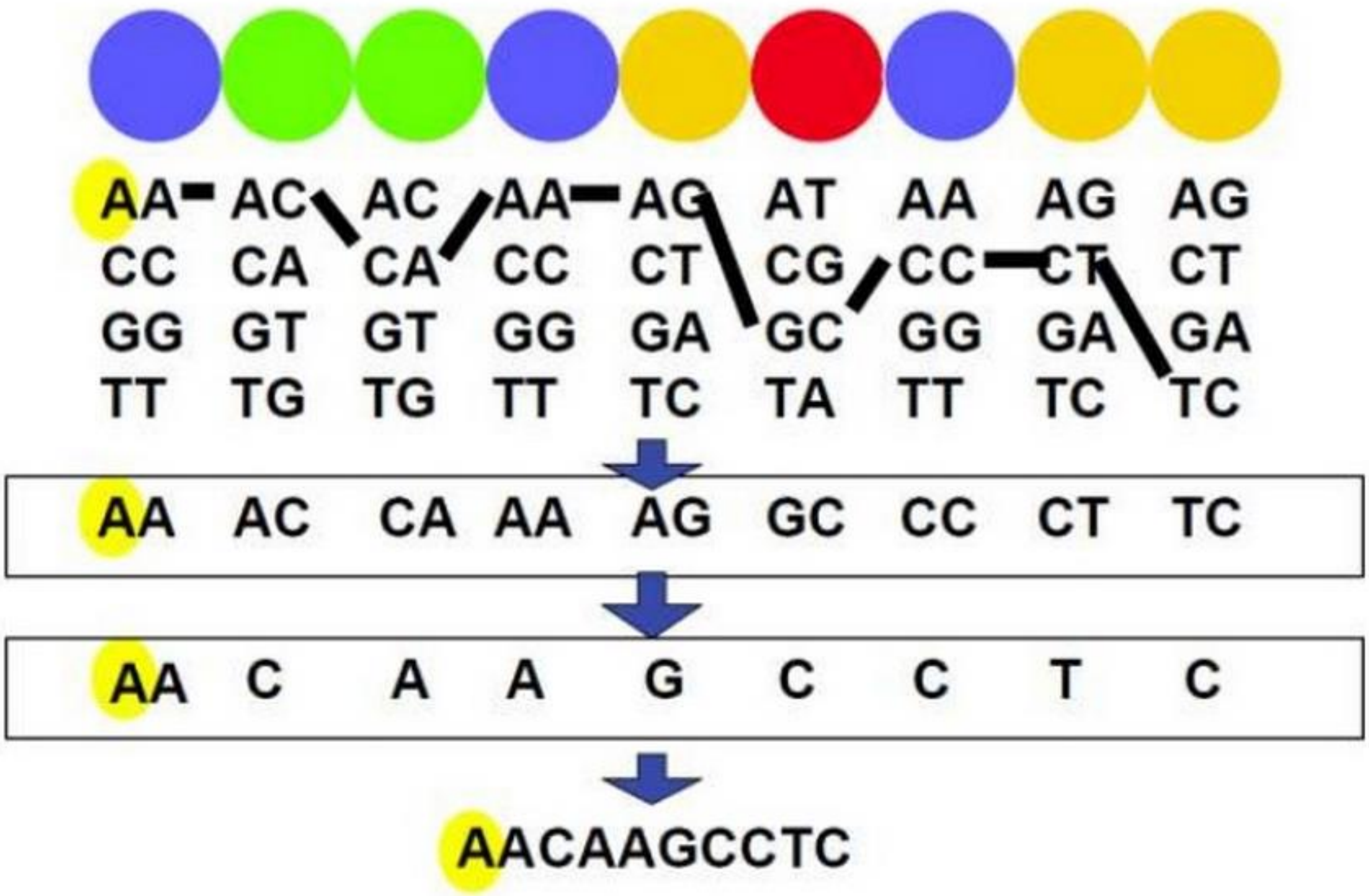
With 2 base encoding each base is defined twice



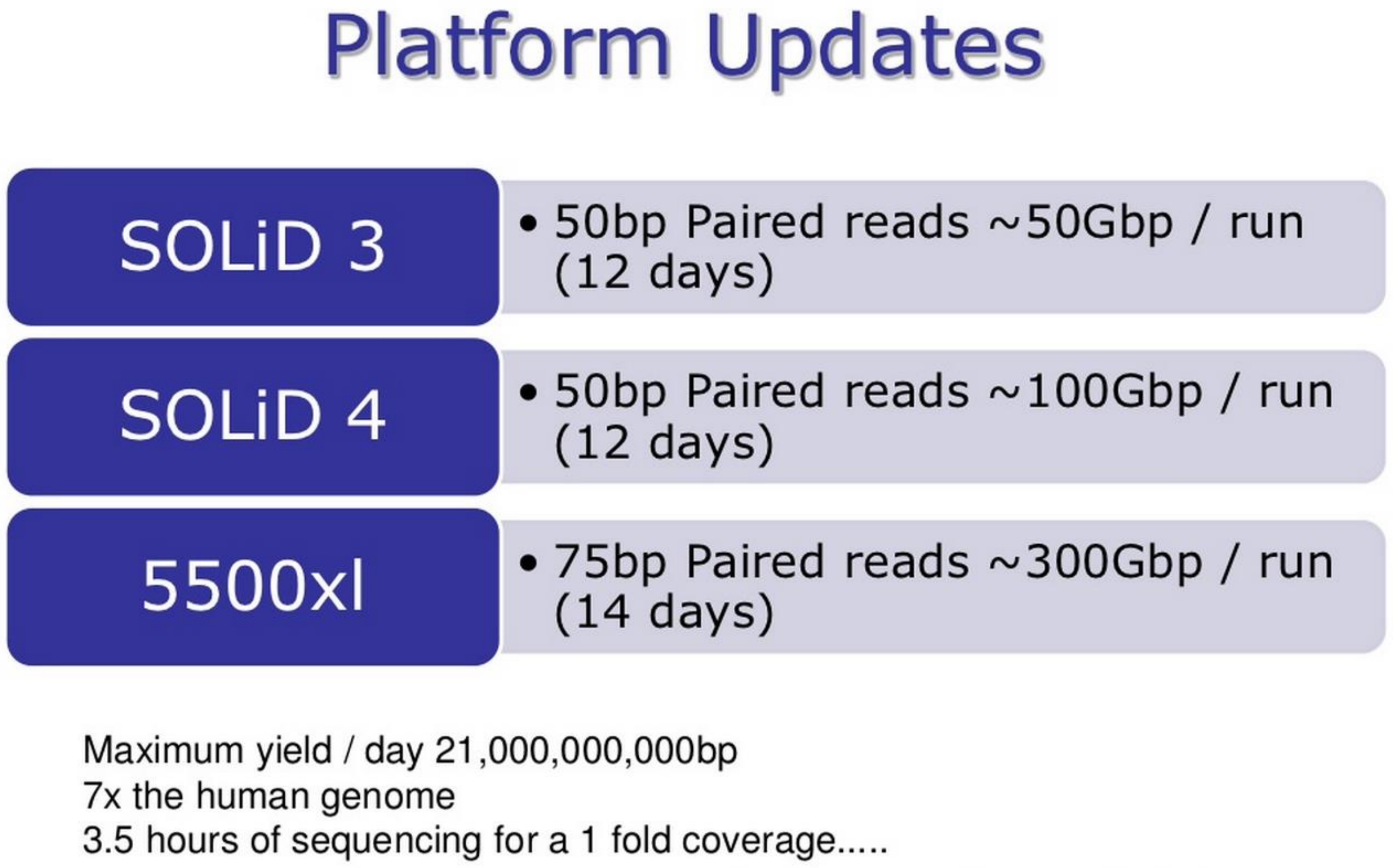
Current technologies & perspectives



Current technologies & perspectives



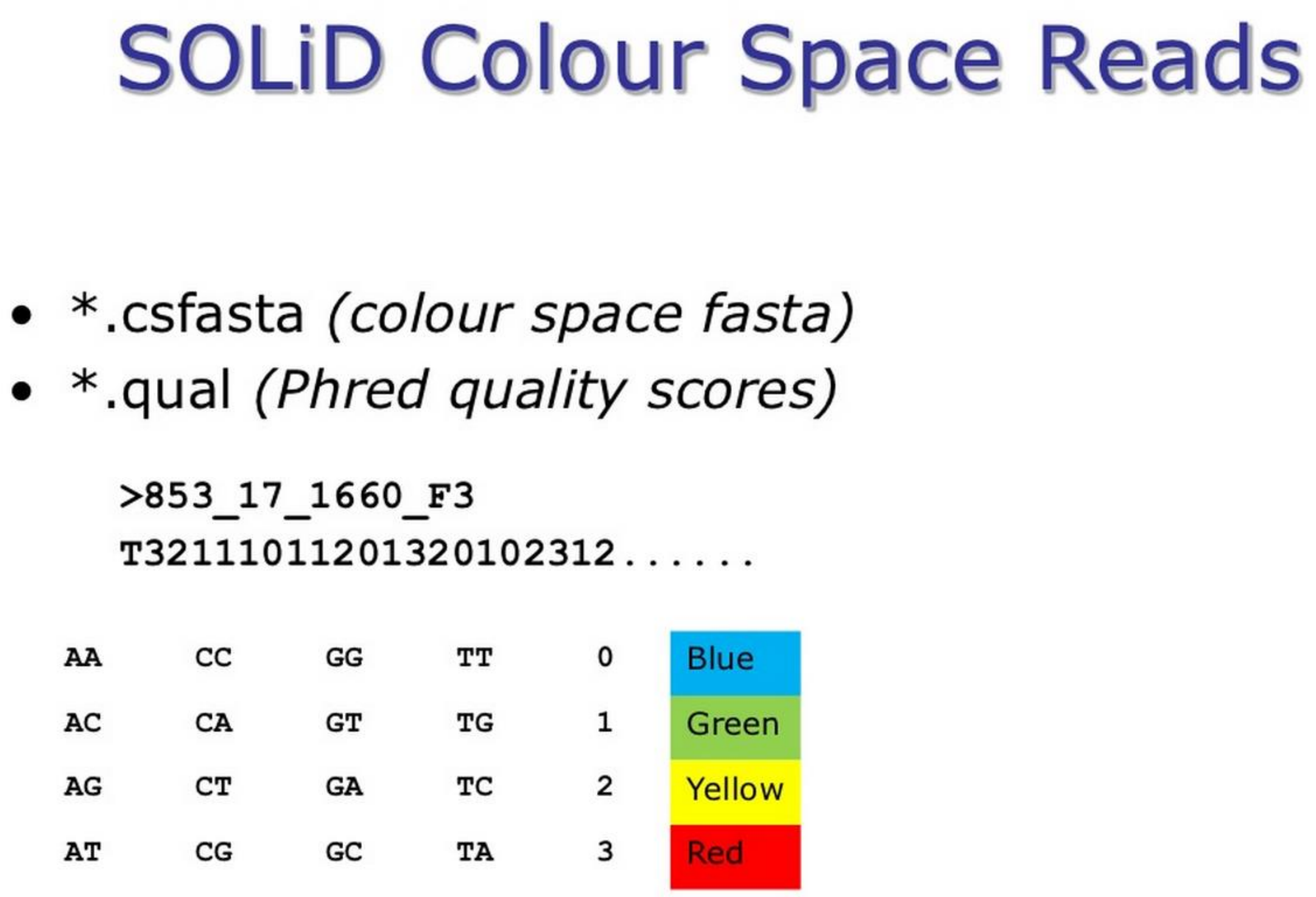
Current technologies & perspectives



Platform Updates

SOLiD 3	• 50bp Paired reads ~50Gbp / run (12 days)
SOLiD 4	• 50bp Paired reads ~100Gbp / run (12 days)
5500xl	• 75bp Paired reads ~300Gbp / run (14 days)

Maximum yield / day 21,000,000,000bp
7x the human genome
3.5 hours of sequencing for a 1 fold coverage.....



SOLiD Colour Space Reads

- *.csfasta *(colour space fasta)*
- *.qual *(Phred quality scores)*

```
>853_17_1660_F3
T32111011201320102312......
```

AA	CC	GG	TT	0	Blue
AC	CA	GT	TG	1	Green
AG	CT	GA	TC	2	Yellow
AT	CG	GC	TA	3	Red

Currently available technologies

SMRT pacific biosciences

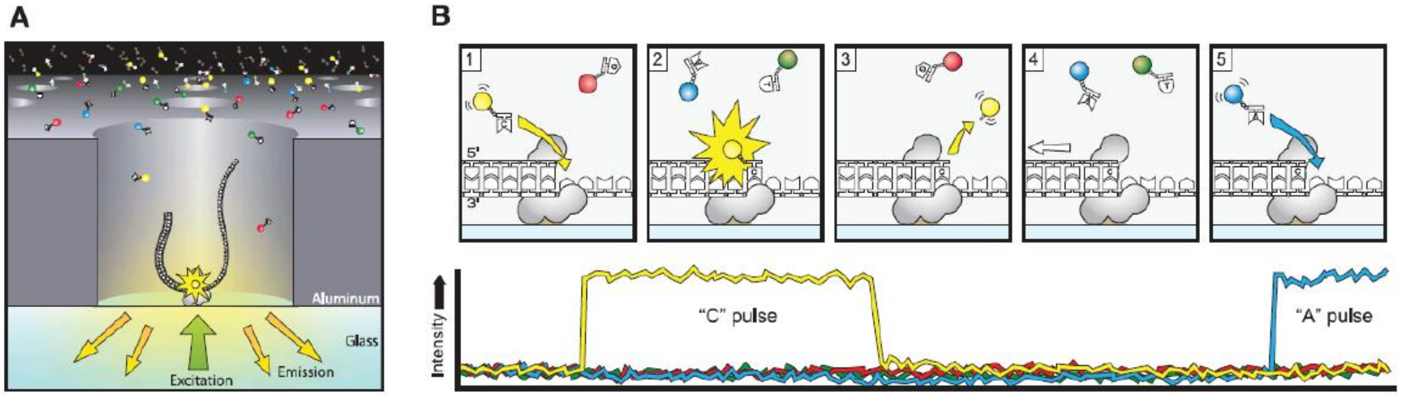
Single Molecule Real Time sequencing – Pacific Bioscience

Specificity : uses a DNA polymerase as real time sequencing engine

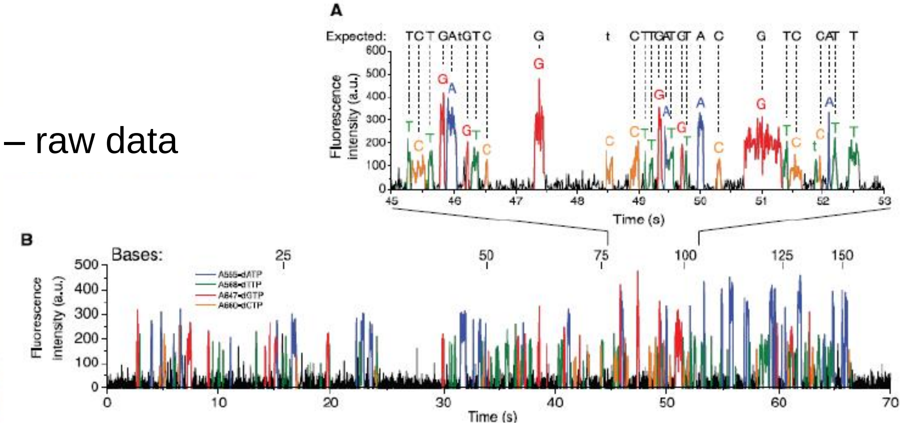
Challenges : accomodate the intrinsic speed and processivity of the enzymes

1. The DNA synthesis speed shows stochastic variations, what implies that the observation has to be ate the molecular level
2. The chemical contact surface should allow for the reaction to inhibate non specific marked dNTPs adsorption
3. The dNTPs carrying the marker should not inhibate the polymerisation
4. The instrument should be reliable at detecting the synthesis and distinguish between each dNTP.

Current technologies & perspectives



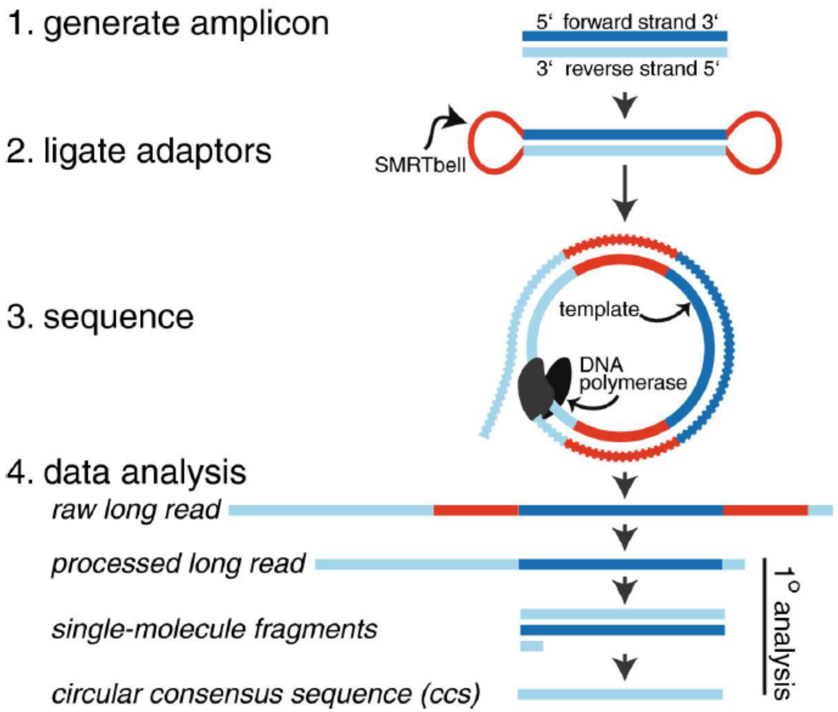
Pacbio RS – raw data



Technical specifications (v3.0)

- Speed : 4.7 bases / s, no spatial correlation
- Signal noise ratio above 24
- 37% ZMWs produce unique and full length sequences
- Error rate is around 14% (D:7,4%; I:4,5%; S:2,1%)

Circular consensus sequencing

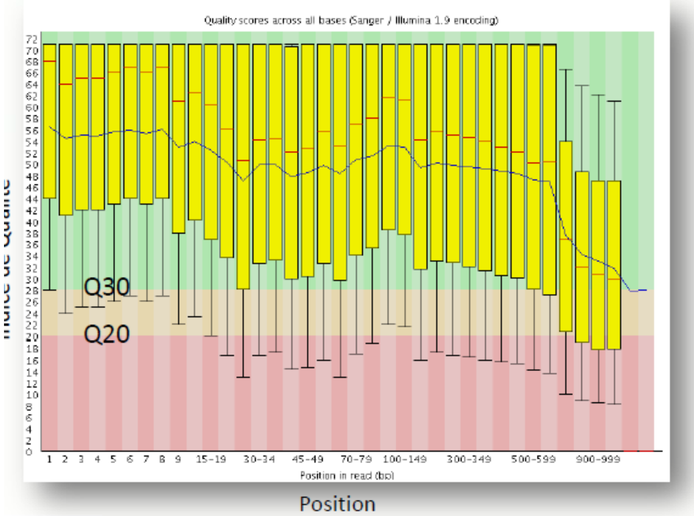


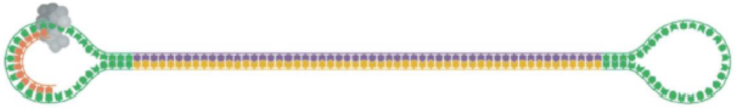
Fichot, E. B., & Norman, R. S. (2013). Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform. *Microbiome*, *1*(1), 10. doi:10.1186/2049-2618-1-10

Pacbio RS sequencer

Main results from the field

- 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% eror rate on controls





Circular consensus sequence (CCS)

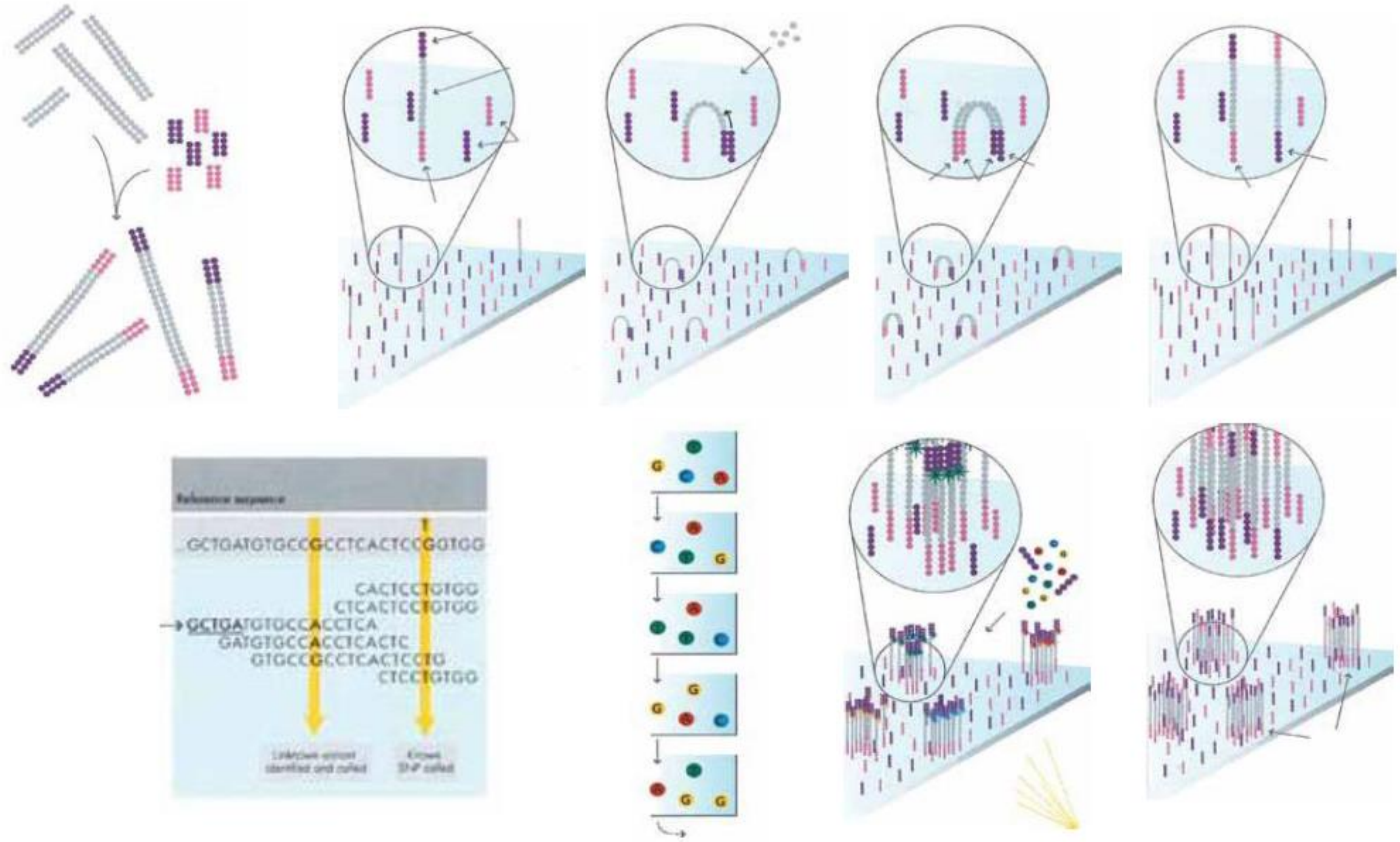
Today on a Pacbio RS II, 15kb median, 40kb max

Currently available technologies

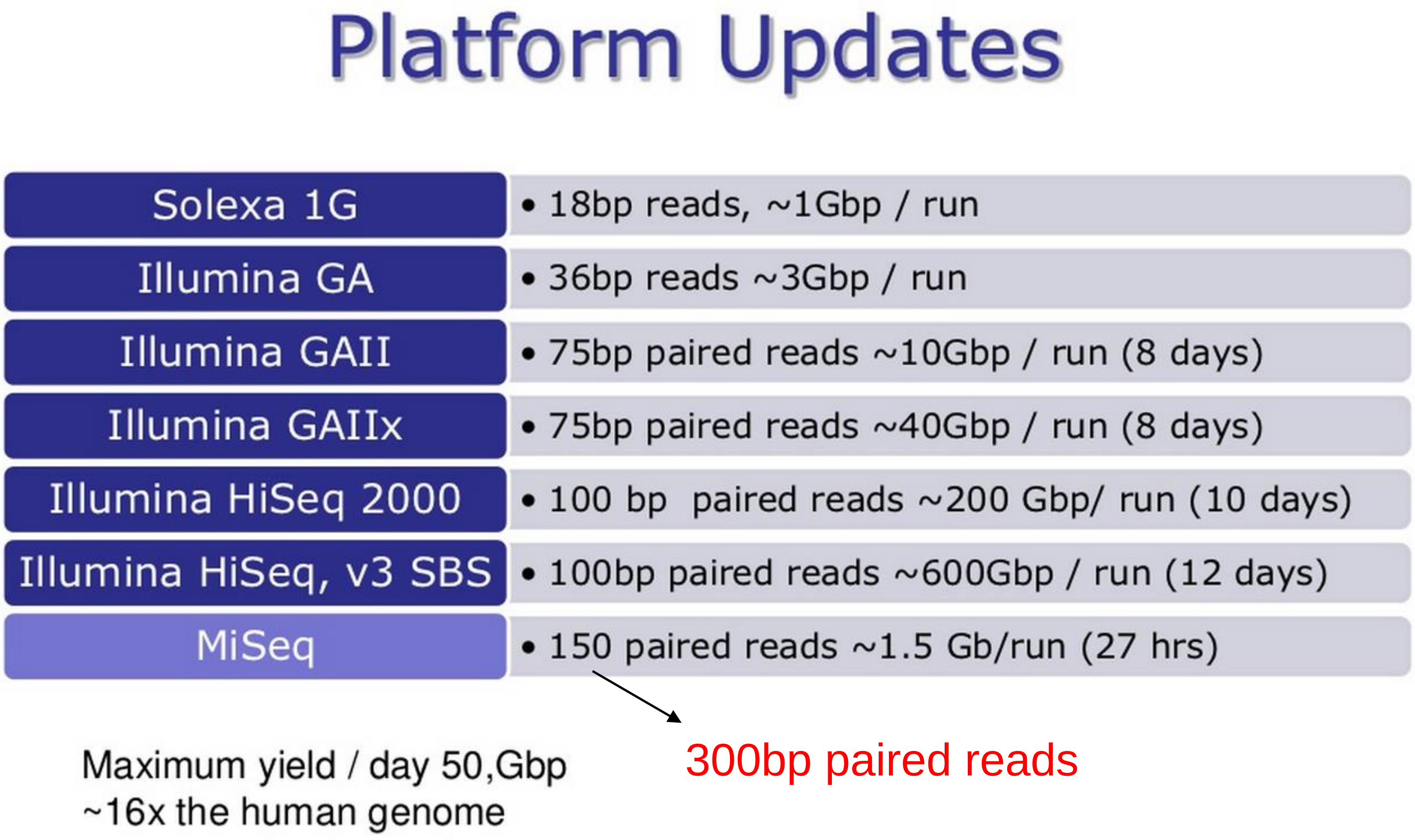
Illumina

Illumina HiSeq

Current technologies & perspectives



Current technologies & perspectives

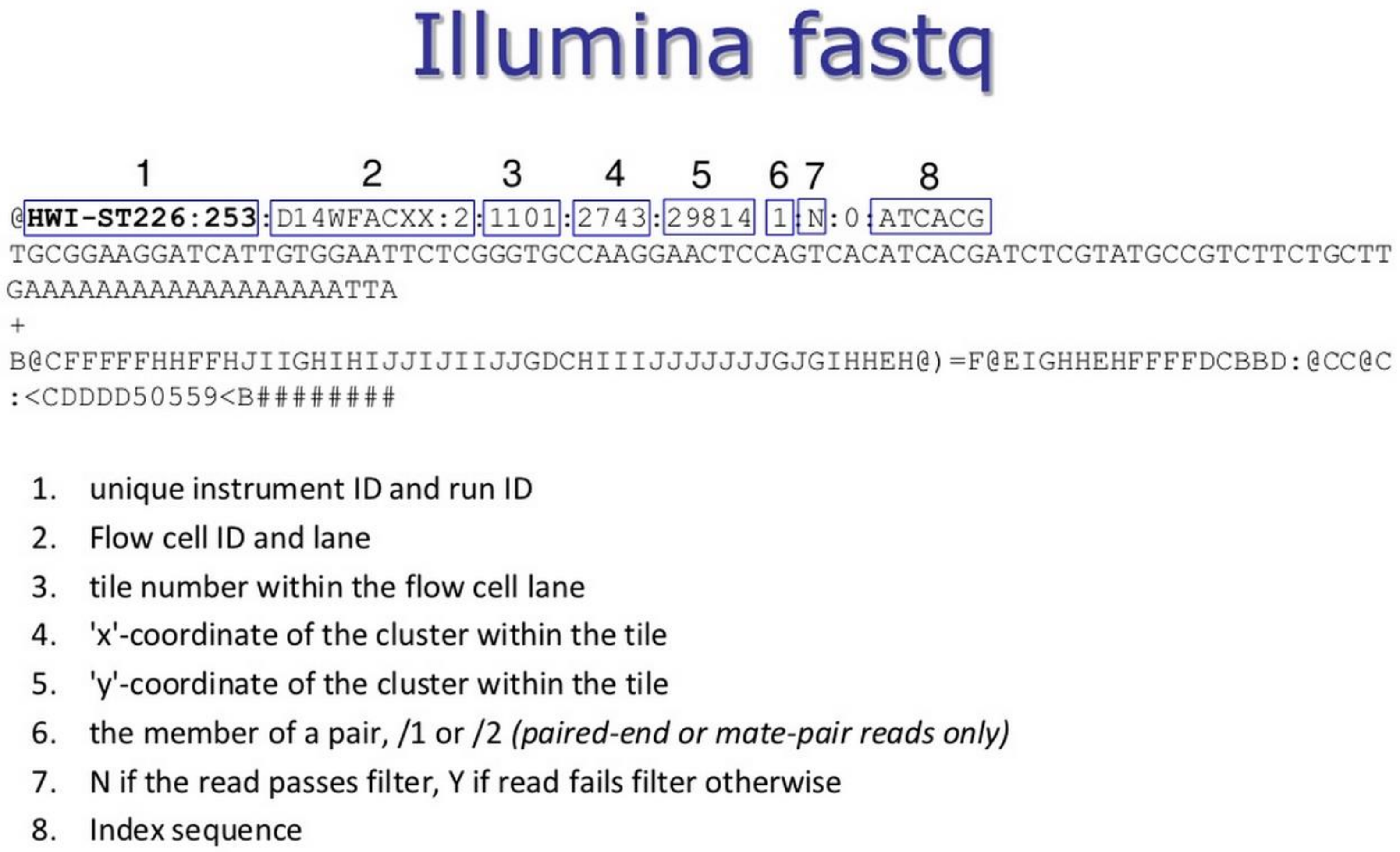


Platform Updates

Solexa 1G	• 18bp reads, ~1Gbp / run
Illumina GA	• 36bp reads ~3Gbp / run
Illumina GAII	• 75bp paired reads ~10Gbp / run (8 days)
Illumina GAIIx	• 75bp paired reads ~40Gbp / run (8 days)
Illumina HiSeq 2000	• 100 bp paired reads ~200 Gbp/ run (10 days)
Illumina HiSeq, v3 SBS	• 100bp paired reads ~600Gbp / run (12 days)
MiSeq	• 150 paired reads ~1.5 Gb/run (27 hrs)

Maximum yield / day 50,Gbp
~16x the human genome

300bp paired reads



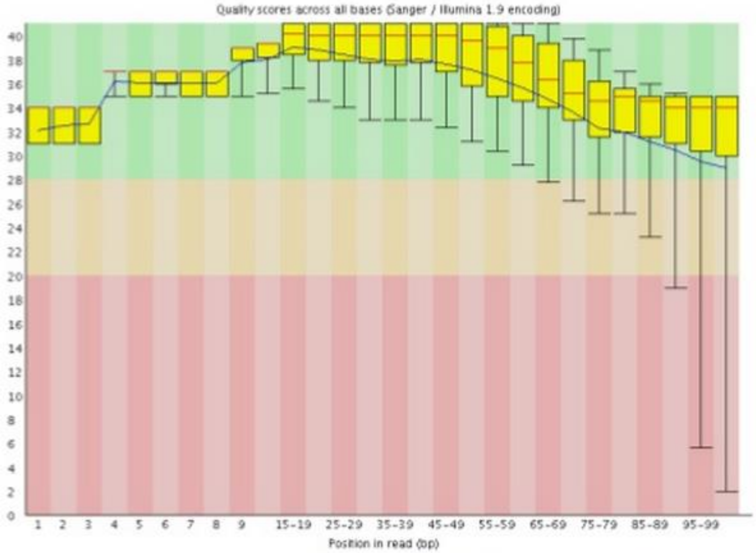
Illumina fastq

```
@HWI-ST226:253:D14WFACXX:2:1101:2743:29814 1:N:0:ATCACG
TGCGGAAGGATCATTGTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTT
GAAAAAAAAAAAAAAAAAATTA
+
B@CFFFFFHHFFHJIIGHIHIJJIJIIJJGDCHIIIJJJJJJGJGIHHEH@)=F@EIGHHEHFFFFDCBBD:@CC@C
:<CDDDD50559<B#########
```

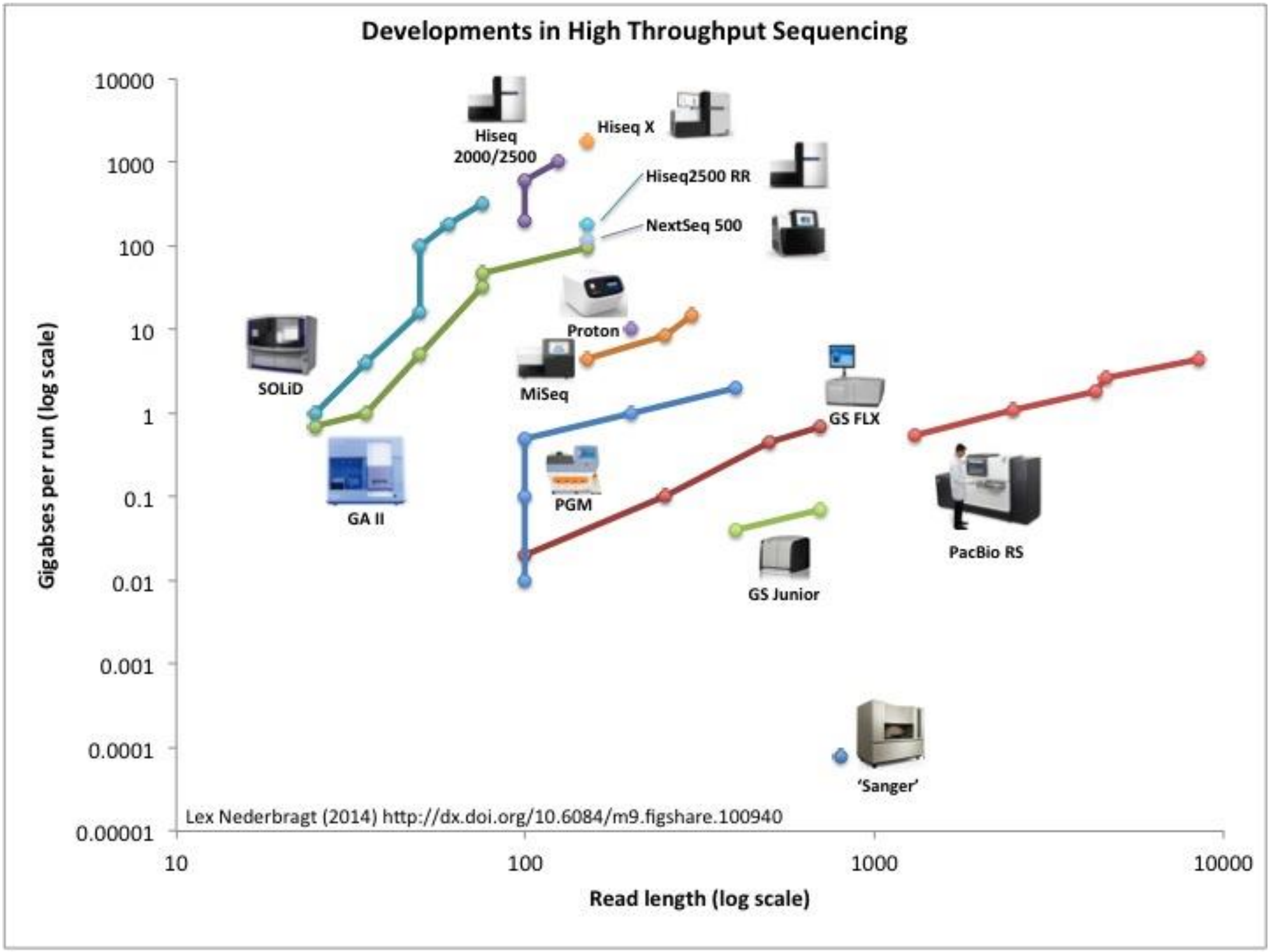
1. unique instrument ID and run ID
2. Flow cell ID and lane
3. tile number within the flow cell lane
4. 'x'-coordinate of the cluster within the tile
5. 'y'-coordinate of the cluster within the tile
6. the member of a pair, /1 or /2 *(paired-end or mate-pair reads only)*
7. N if the read passes filter, Y if read fails filter otherwise
8. Index sequence

Illumina Sequencing Output

- *.fastq *(sequence and corresponding quality score encoded with an ASCII character, phred-like quality score + 33)*



Current technologies & perspectives



Lex Nederbragt (2014) http://dx.doi.org/10.6084/m9.figshare.100940

Perspectives
What to expect for tomorrow?

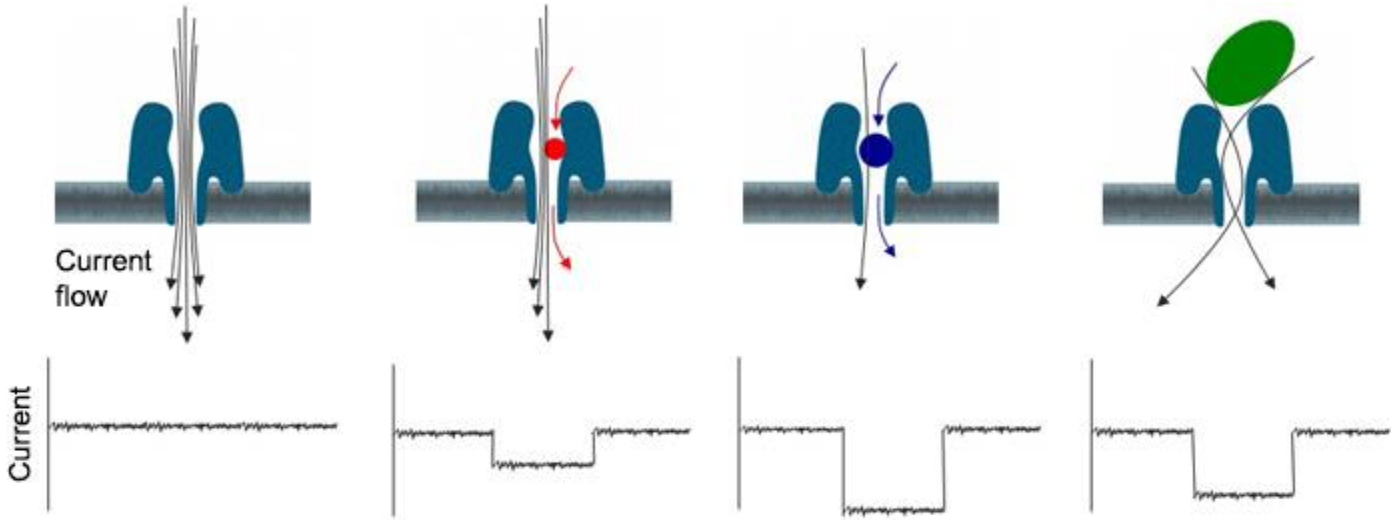
- Longer and even more cheaper sequences
- Faster and easier libraries preparation

-> The wait and sample strategy

Current technologies & perspectives

Oxford Nanopore

https://nanoporetech.com/technology/analytes-and-applications-dna-rna-proteins/dna-an-introduction-to-nanopore-sequencing



Current technologies & perspectives

Oxford Nanopore

Highly scalable system

MinION

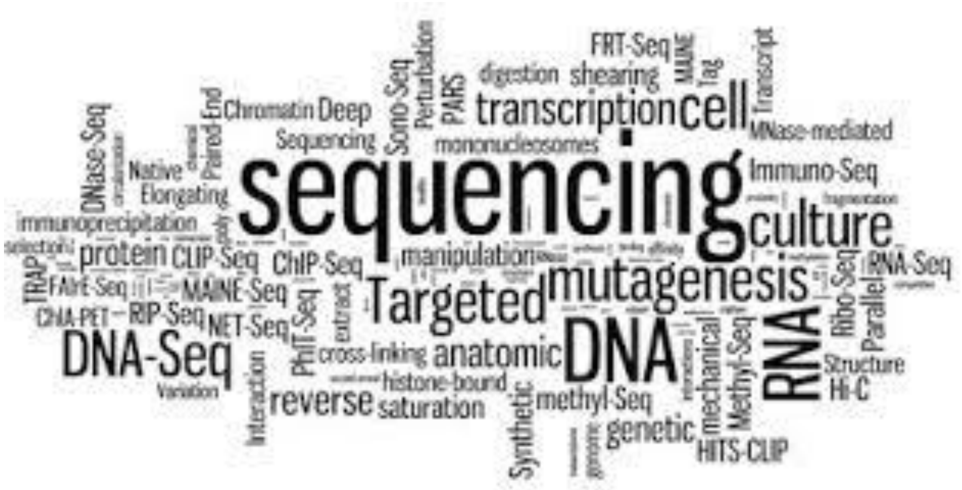
512 pores

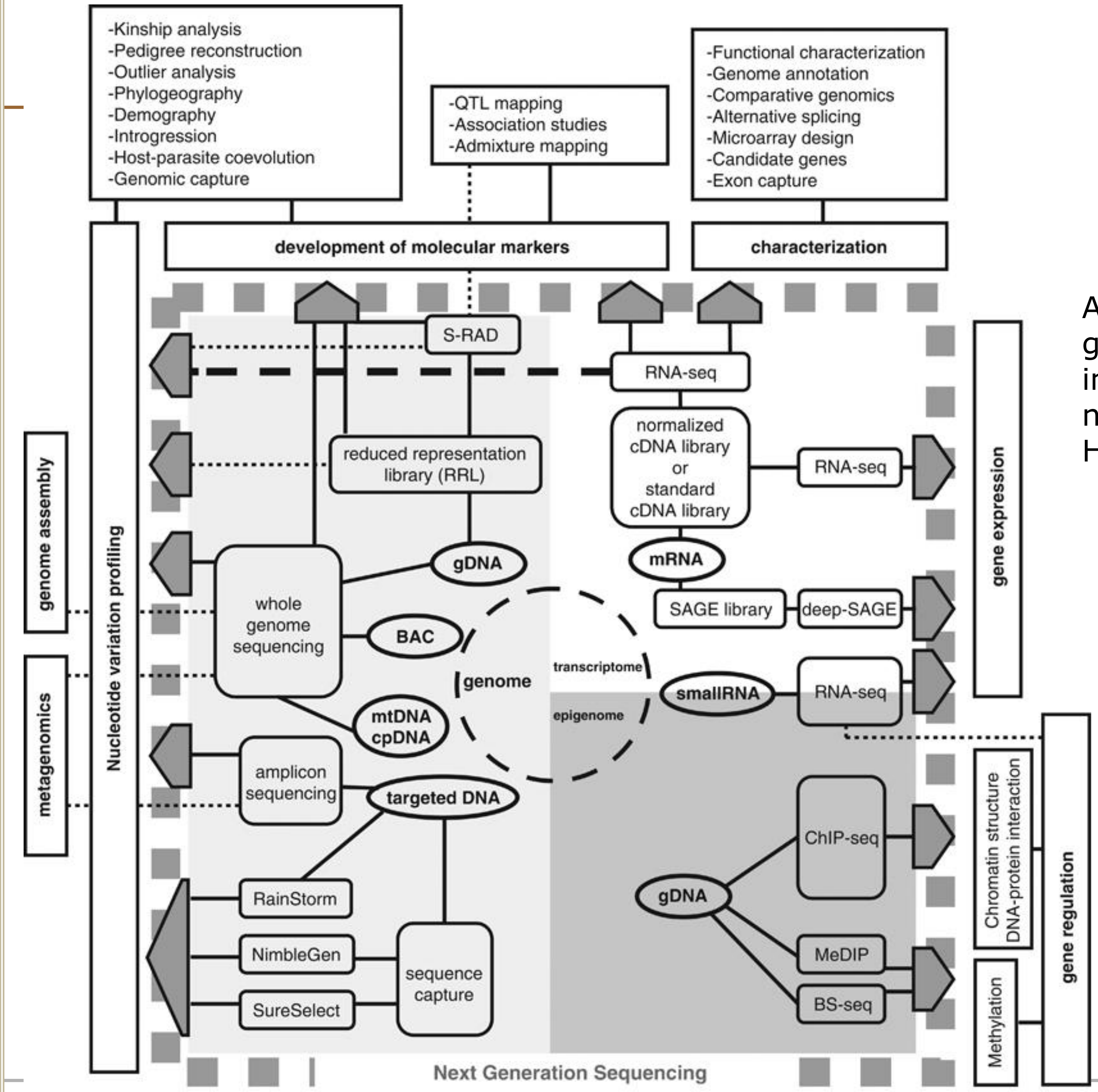
GridION

5 000 pores

Applications overview

Applications overview





Applications of next generation sequencing in molecular ecology of non-model organisms, Heredity, 2011

Challenges

Experimental design

Before starting – thinking ahead

1. Scientific question first
2. What kind of data?
3. How much data?

Challenges

Experimental design

Data acquisition

Commercial kits or not?

Being a geek has a cost

Challenges

Experimental design

Data acquisition

- Number of samples
- Type of read
- Type of library
- Number of reads
- Read length
- Complexity of library
- Which sequencing machine to use

Challenges

Experimental design

Data acquisition

- Steps of library construcCon and sequencing

- Making Fragment libraries (to generate fragment or paired end reads)

- Making Jumping libraries (to generate mate pair reads)

- Pooling with or without barcoding

- Possible artefacts of library construction

Challenges

Experimental design

Data analysis

Huge references list, difficult to sort out

Specialized workshops, bring your own data

Inhouse development and outsourcing

Inhouse development

VS

Outsourcing projects

amazon web services™

GATC BIOTECH

GenoScreen

ADN id
Labeling & AIDC Solutions Provider

BECKMAN COULTER®

eurofins

Inhouse development or outsourcing

Comparing strategies :

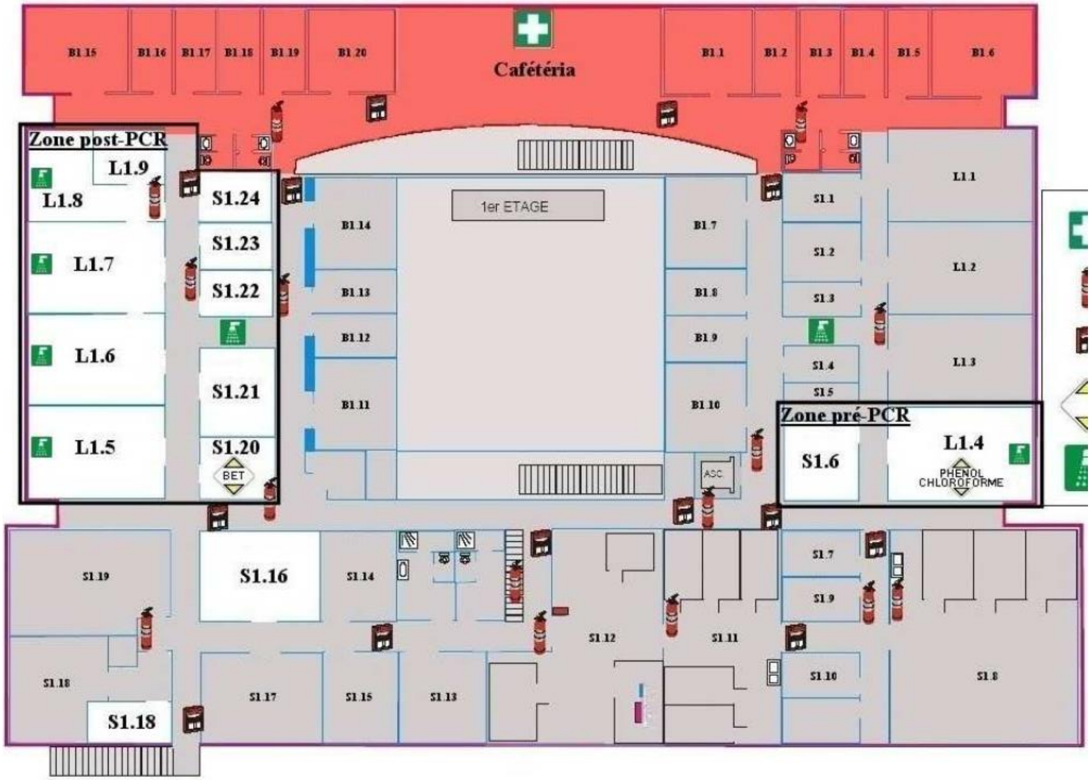
Data acquisition & computing capabilities

	Inhouse	Outsourcing
Cost		
Time		
Quality		
Other		

Inhouse development or outsourcing

Do it yourself : acquire data

1- lab setup



Inhouse development or outsourcing

Do it yourself : acquire data

1- lab setup

70 k€ (x25 in cz crown)

Inhouse development or outsourcing

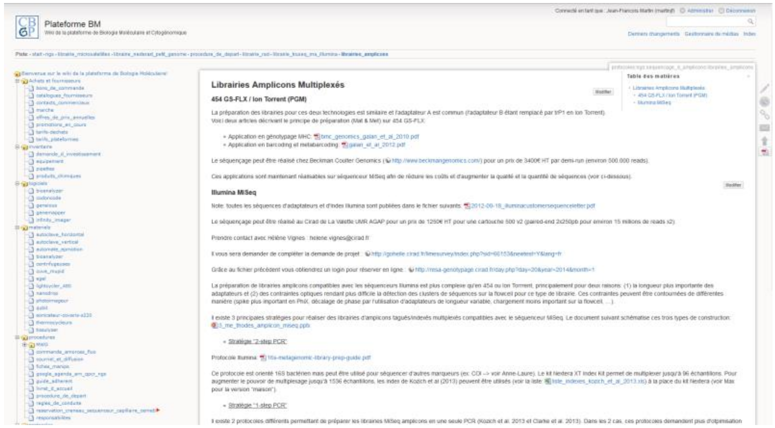
Do it yourself : acquire data

2 – staff training

Communication inside the lab

Team dynamics

Quality management



Librairies Amplicons Multiplexés

454 GS-FLX / Ion Torrent (PGM)

La préparation des librairies pour ces deux technologies est similaire et l'adaptateur A est commun (l'adaptateur B étant remplacé par trP1 en Ion Torrent).
Voici deux articles décrivant le principe de préparation (Mat & Met) sur 454 GS-FLX:

- Application en génotypage MHC: bmc_genomics_galan_et_al_2010.pdf
- Application en barcoding et metabarcoding: galan_et_al_2012.pdf

Le séquençage peut être réalisé chez Beckman Coulter Genomics (http://www.beckmangenomics.com/) pour un prix de 3400€ HT par demi-run (environ 500.000 reads).

Ces applications sont maintenant réalisables sur séquenceur MiSeq afin de réduire les coûts et d'augmenter la qualité et la quantité de séquences (voir ci-dessous).

Illumina MiSeq

Note: toutes les séquences d'adaptateurs et d'index Illumina sont publiées dans le fichier suivants: 2012-09-18_illuminacustomersequenceletter.pdf

Le séquençage peut être réalisé au Cirad de La Valette UMR AGAP pour un prix de 1250€ HT pour une cartouche 500 v2 (paired-end 2x250pb pour environ 15 millions de reads x2)

Prendre contact avec Hélène Vignes : helene.vignes@cirad.fr

Il vous sera demander de compléter la demande de projet : http://gohelle.cirad.fr/limesurvey/index.php?sid=66153&newtest=Y&lang=fr

Grâce au fichier précédent vous obtiendrez un login pour réserver en ligne : http://resa-genotypage.cirad.fr/day.php?day=20&year=2014&month=1

La préparation de librairies amplicons compatibles avec les séquenceurs Illumina est plus complexe qu'en 454 ou Ion Torrent, principalement pour deux raisons: (1) la longueur plus importante des adaptateurs et (2) des contraintes optiques rendant plus difficile la détection des clusters de séquences sur la flowcell pour ce type de librairie. Ces contraintes peuvent être contournées de différentes manière (spike plus important en PhiX, décalage de phase par l'utilisation d'adaptateurs de longueur variable, chargement moins important sur la flowcell, …).

Il existe 3 principales stratégies pour réaliser des librairies d'amplicons tagués/indexés multiplexés compatibles avec le séquenceur MiSeq. Le document suivant schématise ces trois types de construction: 3_methodes_amplicon_miseq.pptx

- Stratégie "2-step PCR"

Protocole Illumina: 16s-metagenomic-library-prep-guide.pdf

Ce protocole est orienté 16S bactérien mais peut être utilisé pour séquencer d'autres marqueurs (ex: COI –> voir Anne-Laure). Le kit Nextera XT Index Kit permet de multiplexer jusqu'à 96 échantillons. Pour augmenter le pouvoir de multiplexage jusqu'à 1536 échantillons, les index de Kozich et al (2013) peuvent être utilisés (voir la liste: liste_indexes_kozich_et_al_2013.xls) à la place du kit Nextera (voir Max pour la version "maison").

- Stratégie "1-step PCR"

Il existe 2 protocoles différents permettant de préparer les librairies MiSeq amplicons en une seule PCR (Kozich et al. 2013 et Clarke et al. 2013). Dans les 2 cas, ces protocoles demandent plus d'optimisation

Labs network and joint meetings at the regional scale

Inhouse development or outsourcing

Do it yourself : store and analyse data

90 k€
cluster

7 k€
storage

A good system and infrastructure administrator : priceless!

Comparing strategies : conclusion

	Inhouse	**Outsourcing**
Cost		
Time		
Quality		
Other		

IT DEPENDS !

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

Montpellier SupAgro

www.supagro.fr

Centre de Biologie pour la Gestion des Populations

Centre international d'études supérieures en sciences agronomiques