TIMM
Trends in Medical Mycology

New Generation Sequencing any use?

Stéphane Bretagne
National Reference Center Invasive Mycoses & Antifungals,
Molecular Mycology Unit,
CNRS URA3012, Institut Pasteur
St Louis Hospital, Paris

Hôpitaux Universitaires SAINT-LOUIS LARIBOISIÈRE FERNAND-WIDAL

cnrs

INSTITUT PASTEUR

Sanger sequencing method

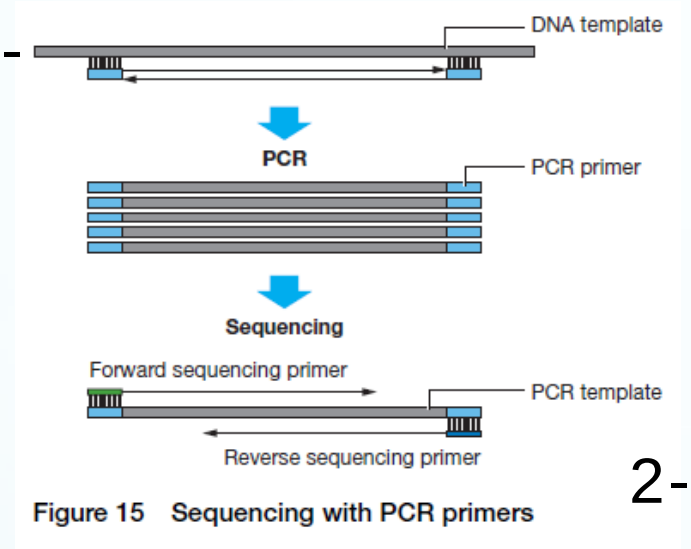
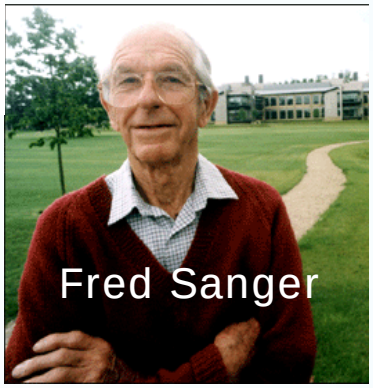
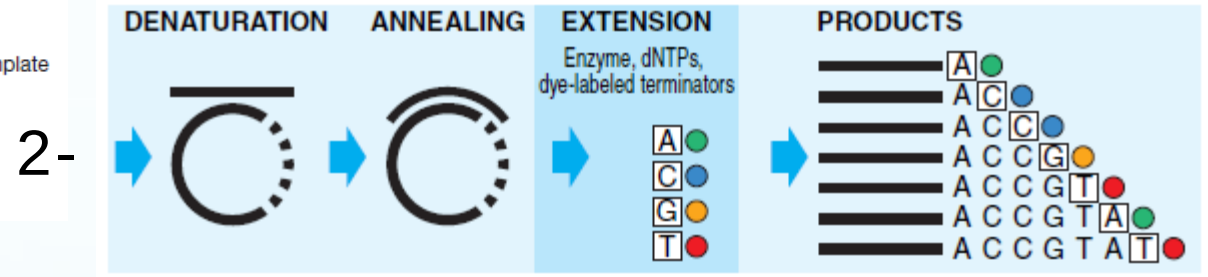


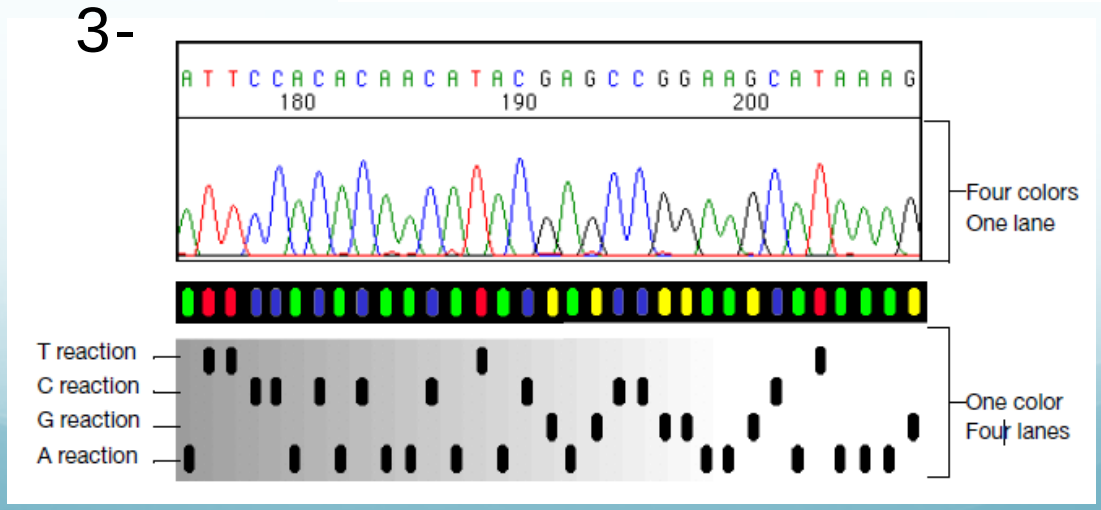
Figure 15 Sequencing with PCR primers

The Nobel Prize in Chemistry 1980

Shared with Walter Gilbert and Paul Berg







Next-Generation sequencing

- DNA-sequencing methods that involve chemical assays other than the traditional Sanger deoxy-chain-termination method.[1]
  - Deep Sequencing
  - Massively Parallel Sequencing
  - Second/Third-generation Sequencing
    - 2nd: Undergo amplification of the template molecules
    - 3rd: Single-molecule sequencing
- Instead of 1 read per bp, multiple sequence reads per bp

1. *Lupski, et al (2010) Whole Genome Sequencing in a Patient with Charcot-Marie Tooth Neuropathy. NEJM. 362(13):1181-1191.*

1. Library Preparation

- **Input DNA fragmented**
  - Shearing by:
    - nebulization
    - sonication
    - enzymatic digestion
- **Fragments have terminal overhangs**
  - Blunt-end repair and phosphorylation
- **Adapter ligation**
  - Platform-specific adapters are ligated to the fragments
- **Final Library**
  - Short DNA fragments with platform-specific adapters

Voelkerding KV (2010) *Next Generation Sequencing for Clinical Diagnostics-Principles and Application to Targeted Resequencing for Hypertrophic Cardiomyopathy.* Journal of Molecular Diagnostics 5(12): 539-551

2. Clonal Amplification

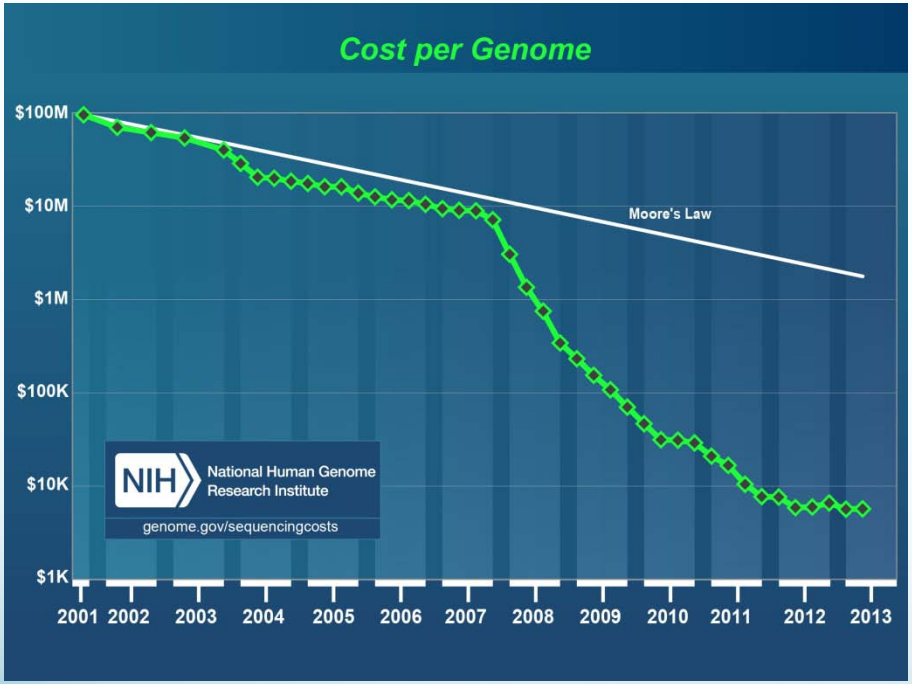
3. Sequencing

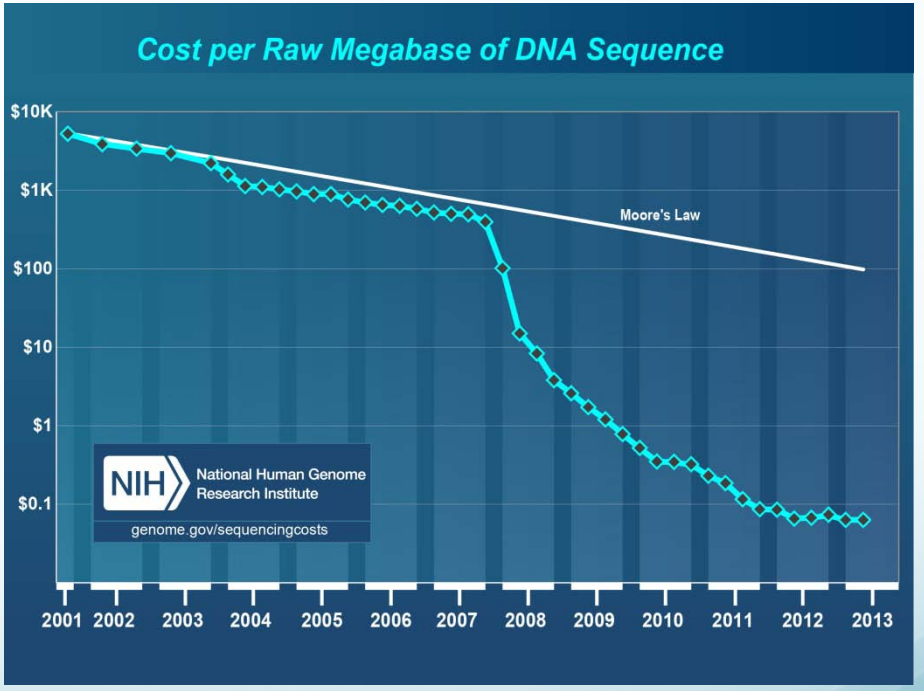
- **Pyrosequencing**
- **Sequencing by ligation**
- **Reversible dye terminators**
- **Emission of photons**

Voelkerding KV (2010) *Next Generation Sequencing for Clinical Diagnostics-Principles and Application to Targeted Resequencing for Hypertrophic Cardiomyopathy.* Journal of Molecular Diagnostics 5(12): 539-551

Cost of Sequencing Over Time

- Human Genome Project: $3 billion and 13 years
- ***NOW:*** Sequencing centers and laboratories: ~$15K and ~15 days



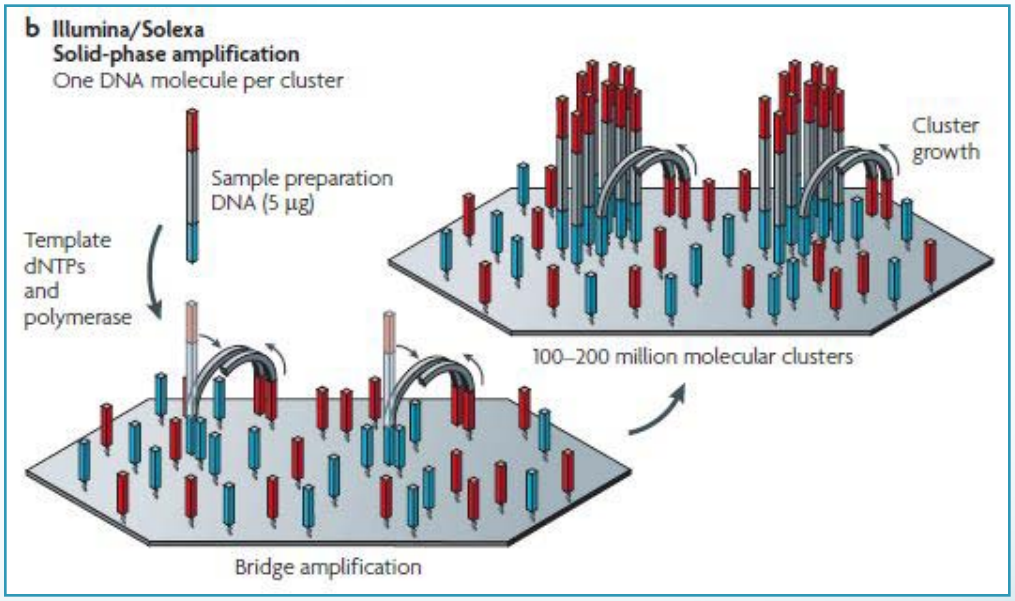


Data from the National Human Genome Research Institute (NHGRI)

WHICH TECHNOLOGY?

Next-Gen Technologies on the Market

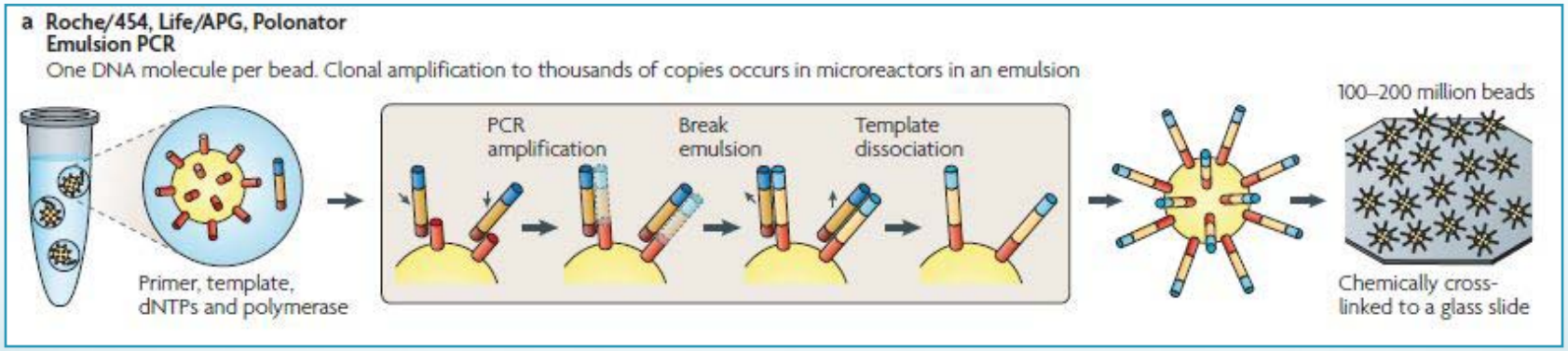
- **Illumina**
  - HiSeq *(HiSeq2000, HiSeq 1000)*
  - Genome Analyzer *(IIx, IIe)*
  - MiSeq
  - iScanSQ



M. L. Metzker Nature Reviews Genetics 11 January 2010; 31-46

Next-Gen Technologies on the Market

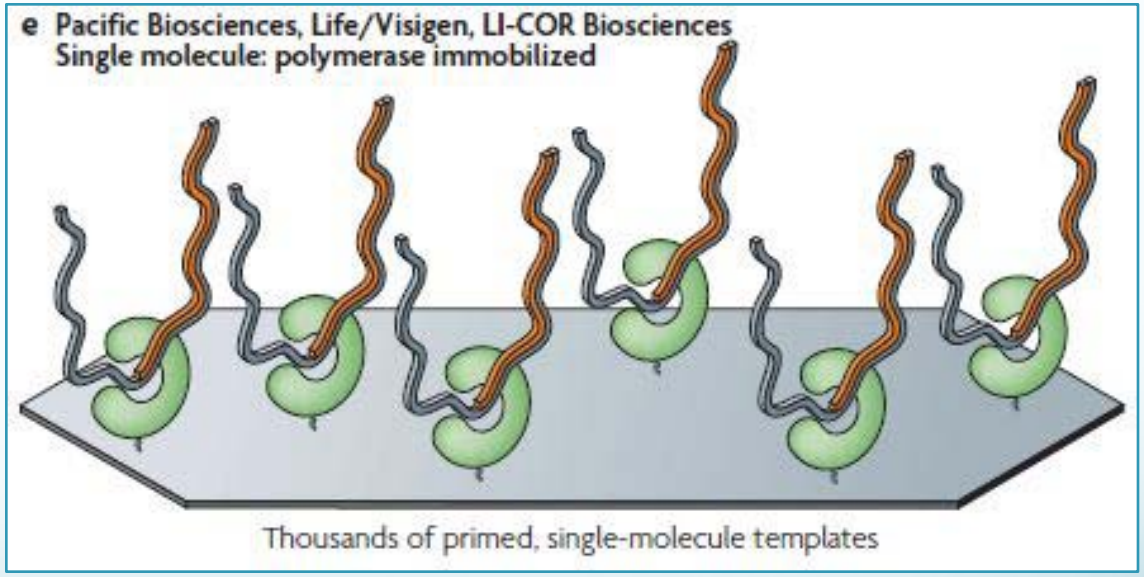
- **Roche 454**
  - *GS FLX Titanium, GS Junior)*



M. L. Metzker Nature Reviews Genetics 11 January 2010; 31-46

Next-Gen Technologies on the Market

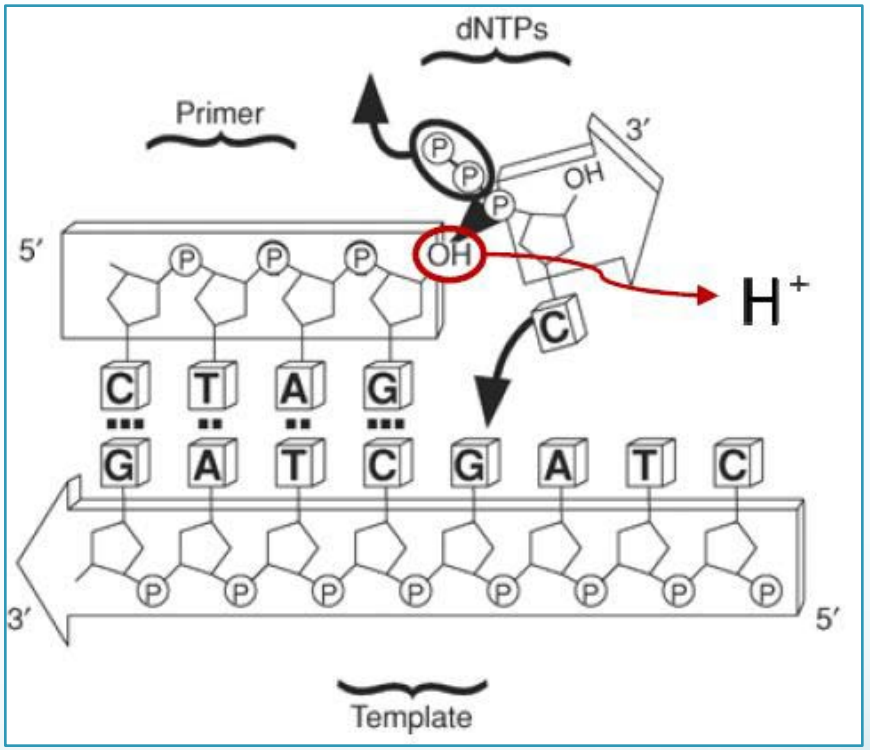
- **Pacific Biosciences**
  - PacBio RS



M. L. Metzker Nature Reviews Genetics 11 January 2010; 31-46

Next-Gen Technologies on the Market

- **Ion Torrent PGM**
  - 314, 316 318 ™ chips
- **Ion Proton Sequencer**



Performance comparison of benchtop high-throughput sequencing platforms

Table 2 Run and alignment metrics for benchtop sequencers

Platform (run)	Number of reads	Total bases	Modal read length in bases	Mean read length in bases (s.d.)	Alignment coverage		
					Chromosome	Large plasmids	Reads aligned (%)
454 GS Junior (1)	135,992	70,999,968	518	522 (46)	11.50	5.66	99
454 GS Junior (2)	137,528	71,710,564	516	521 (47)	11.54	5.39	99
Ion Torrent PGM (1)	2,483,868	303,579,279	123	122 (11)	46.60	53.33	90
Ion Torrent (2)	2,154,577	260,017,346	123	120 (16)	39.33	43.80	89
MiSeq (1)	11,708,156	1,652,529,000	150	141 (22)	–	–	–
MiSeq (1) demulti-plexed strain 280	1,766,516	250,356,566	150	141 (21)	22.11	625.46	99

Metrics for each sequencing run are shown as well as results of alignment against the reference sequence. Depth of coverage for the chromosome and two large plasmids (pESBL and pAA) are shown with the percentage of reads that align. For the MiSeq run, the sequence metrics are shown for the entire run as well as the results of de-multiplexing *E. coli* O104:H4 strain 280. Alignment statistics for the entire run are not shown as two strains sequenced were of *E. coli* isolates unrelated to the outbreak strain.

Table 1 Price comparison of benchtop instruments and sequencing runs

Platform	List price	Approximate cost per run	Minimum throughput (read length)	Run time	Cost/Mb	Mb/h
454 GS Junior	$108,000	$1,100	35 Mb (400 bases)	8 h	$31	4.4
Ion Torrent PGM						
(314 chip)	$80,490[a,b]	$225[c]	10 Mb (100 bases)	3 h	$22.5	3.3
(316 chip)		$425	100 Mb[d] (100 bases)	3 h	$4.25	33.3
(318 chip)		$625	1,000 Mb (100 bases)	3 h	$0.63	333.3
MiSeq	$125,000	$750	1,500 Mb (2 × 150 bases)	27 h	$0.5	55.5

Note pricing may vary between countries and/or sales territories. Instrument prices do not include service contracts. Sample prices do not include the cost of generating the initial fragmented genomic DNA library with adaptors (an additional cost of between $75–200 depending on method used). Cost per megabase assumes one sample and one sample sequencing kit per run. Unless stated, pricing information is from the online supplement of ref. 3.
[a]Ion Torrent PGM pricing from Invitrogen US territory website (http://www.invitrogen.com/, accessed 21 February 2012). [b]Price includes Ion Torrent PGM, server, OneTouch and OneTouch ES sample automation systems. [c]Ion Torrent PGM prices include chip and sample preparation kit. [d]Configuration used in this study.

Loman NJ et al, Nature Biotechnology 2012 30 (5) 434-9

Performance comparison of benchtop high-throughput sequencing platforms (*E. coli* genome)

Table 2 Run and alignment metrics for benchtop sequencers

Platform (run)	Number of reads	Total bases	Modal read length in bases	Mean read length in bases (s.d.)	Alignment coverage		
					Chromosome	Large plasmids	Reads aligned (%)
454 GS Junior (1)	135,992	70,999,968	518	522 (46)	11.50	5.66	99
454 GS Junior (2)	137,528	71,710,564	516	521 (47)	11.54	5.39	99
Ion Torrent PGM (1)	2,483,868	303,579,279	123	122 (11)	46.60	53.33	90
Ion Torrent (2)	2,154,577	260,017,346	123	120 (16)	39.33	43.80	89
MiSeq (1)	11,708,156	1,652,529,000	150	141 (22)	–	–	–
MiSeq (1) demulti-plexed strain 280	1,766,516	250,356,566	150	141 (21)	22.11	625.46	99

The MiSeq had the highest throughput per run (1.6 Gb/run, 60 Mb/h) and lowest error rates. The 454 GS Junior generated the longest reads (up to 600 bases) and most contiguous assemblies but had the lowest throughput (70 Mb/run, 9 Mb/h). Run in 100-bp mode, the Ion Torrent PGM had the highest throughput (80–100 Mb/h). Unlike the MiSeq, the Ion Torrent PGM and 454 GS Junior both produced homopolymer-associated indel errors (1.5 and 0.38 errors per 100 bases, respectively).

Note pricing may vary between countries and/or sales territories. Instrument prices do not include service contracts. Sample prices do not include the cost of generating the initial fragmented genomic DNA library with adaptors (an additional cost of between $75–200 depending on method used). Cost per megabase assumes one sample and one sample sequencing kit per run. Unless stated, pricing information is from the online supplement of ref. 3.
[a]Ion Torrent PGM pricing from Invitrogen US territory website (http://www.invitrogen.com/, accessed 21 February 2012). [b]Price includes Ion Torrent PGM, server, OneTouch and OneTouch ES sample automation systems. [c]Ion Torrent PGM prices include chip and sample preparation kit. [d]Configuration used in this study.

Loman NJ et al, Nature Biotechnology 2012 30 (5) 434-9

A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers

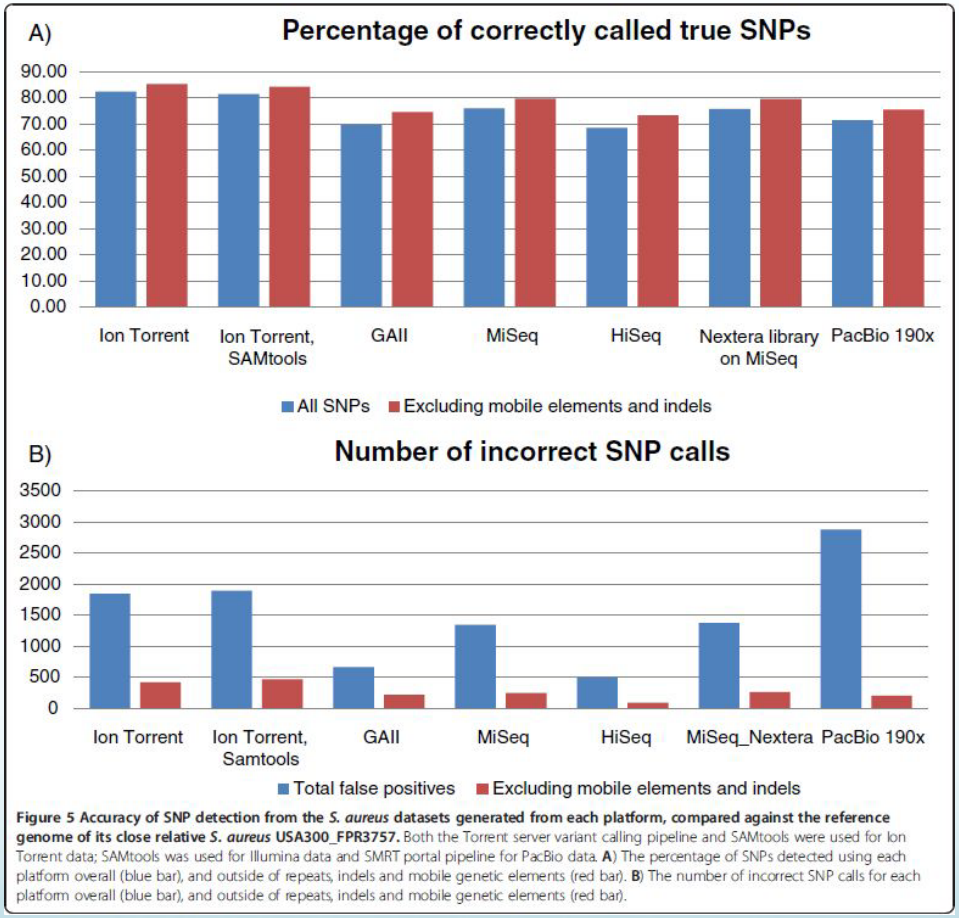
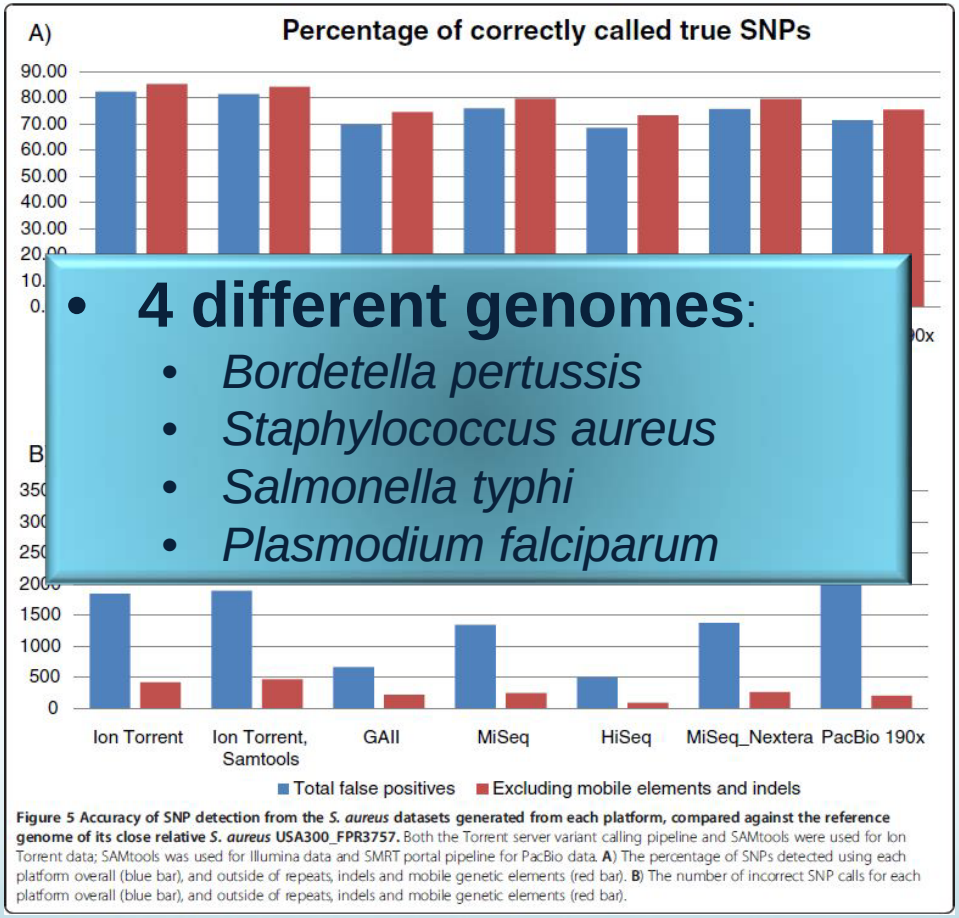


Figure 5 Accuracy of SNP detection from the *S. aureus* datasets generated from each platform, compared against the reference genome of its close relative *S. aureus* USA300_FPR3757. Both the Torrent server variant calling pipeline and SAMtools were used for Ion Torrent data; SAMtools was used for Illumina data and SMRT portal pipeline for PacBio data. **A**) The percentage of SNPs detected using each platform overall (blue bar), and outside of repeats, indels and mobile genetic elements (red bar). **B**) The number of incorrect SNP calls for each platform overall (blue bar), and outside of repeats, indels and mobile genetic elements (red bar).

Quail et al. BMC Genomics 2012, 13:341

A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers



- **4 different genomes**:
  - *Bordetella pertussis*
  - *Staphylococcus aureus*
  - *Salmonella typhi*
  - *Plasmodium falciparum*

Figure 5 Accuracy of SNP detection from the *S. aureus* datasets generated from each platform, compared against the reference genome of its close relative *S. aureus* USA300_FPR3757. Both the Torrent server variant calling pipeline and SAMtools were used for Ion Torrent data; SAMtools was used for Illumina data and SMRT portal pipeline for PacBio data. **A**) The percentage of SNPs detected using each platform overall (blue bar), and outside of repeats, indels and mobile genetic elements (red bar). **B**) The number of incorrect SNP calls for each platform overall (blue bar), and outside of repeats, indels and mobile genetic elements (red bar).

Quail et al. BMC Genomics 2012, 13:341

A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers

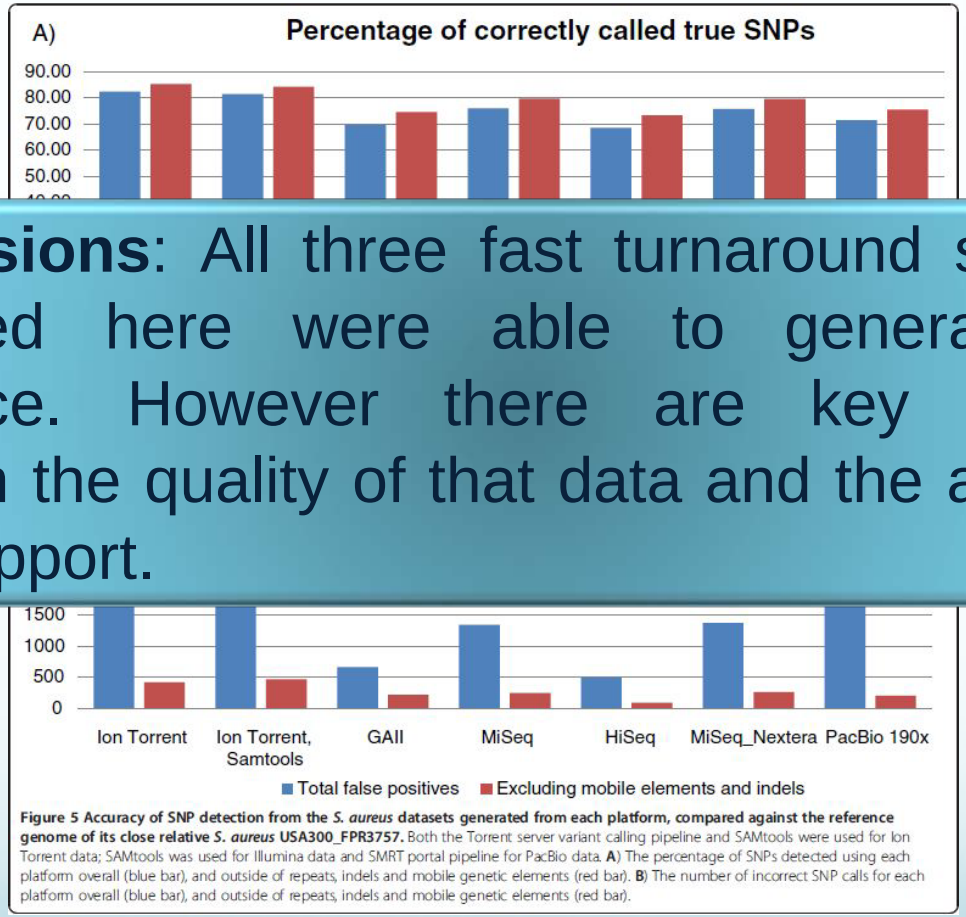


Figure 5 Accuracy of SNP detection from the *S. aureus* datasets generated from each platform, compared against the reference genome of its close relative *S. aureus* USA300_FPR3757. Both the Torrent server variant calling pipeline and SAMtools were used for Ion Torrent data; SAMtools was used for Illumina data and SMRT portal pipeline for PacBio data. **A**) The percentage of SNPs detected using each platform overall (blue bar), and outside of repeats, indels and mobile genetic elements (red bar). **B**) The number of incorrect SNP calls for each platform overall (blue bar), and outside of repeats, indels and mobile genetic elements (red bar).

Conclusions: All three fast turnaround sequencers evaluated here were able to generate usable sequence. However there are key differences between the quality of that data and the applications it will support.

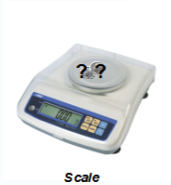
Quail et al. BMC Genomics 2012, 13:341

Which question?

- **Real-Time PCR**
  - Identify and quantify a microorganism
- **Broad identification of multiple micoorganisms**
  - Microarrays
  - PCR ESI Mass Spectromatry (PlexID)
- **Targeting sequencing**
  - Identification and quantitation of sequence variants in an amplicons; haplotype
- **Deep sequencing**
  - ***de novo* sequencing**
    - Sequencing of a genome that has not been sequenced before or does not have a reference
  - **Re-sequencing**
    - Sequencing a genome for which a reference sequence exists
    - Generally done for the purpose of mutation detection
  - **Others**

PCR ESI Mass Spectrometry (PlexID)

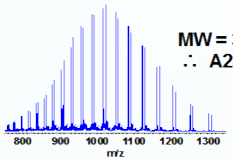
Penny = 2.500 g
Nickel = 3.950 g
Dime = 2.268 g
Quarter = 5.670 g



Scale

Weight = 4.6 grams
∴ 2 dimes

A = 313.0576 amu
G = 329.0526 amu
C = 289.0464 amu
T = 304.0461 amu



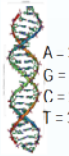
Mass spectrum

MW = 32,588.90 amu
∴ A28 G29 C25 T24

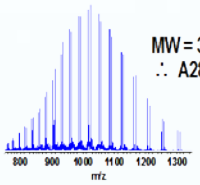
Penny = 2.500 g
Nickel = 3.950 g
Dime = 2.268 g
Quarter = 5.670 g

Scale

Weight = 4.6 grams
∴ 2 dimes



A = 313.0576 amu
G = 329.0526 amu
C = 289.0464 amu
T = 304.0461 amu

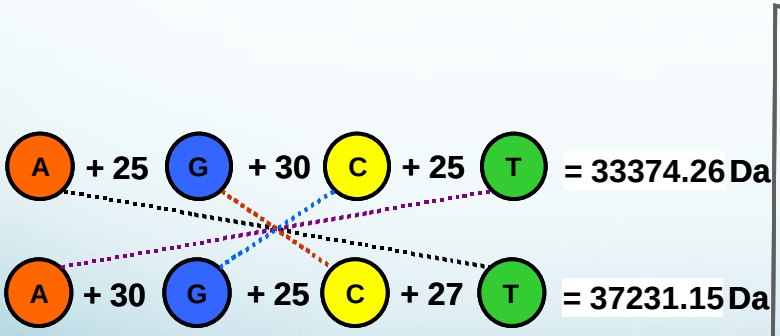


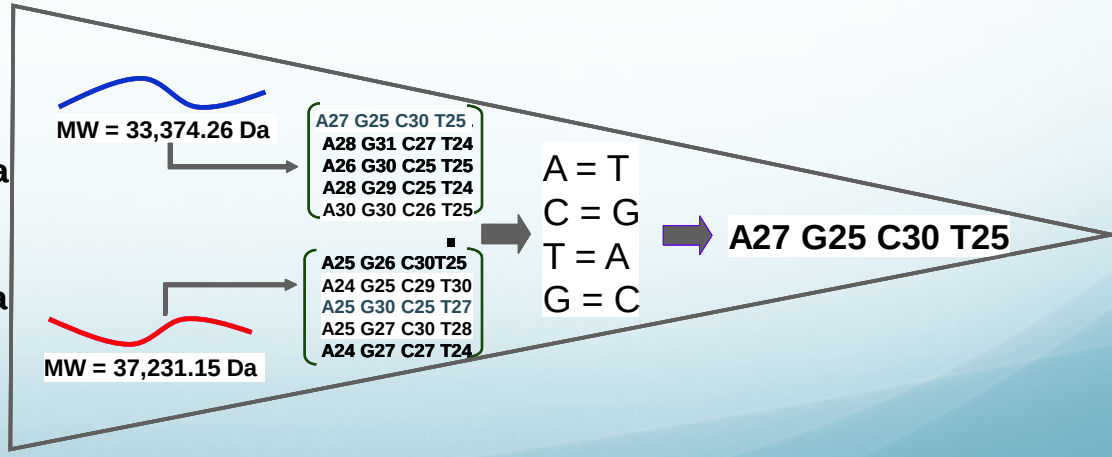
Mass spectrum

MW = 32,588.90 amu
∴ A28 G29 C25 T24

Weight: 36,2 g

1 x 1 euro
5 x 20 cents

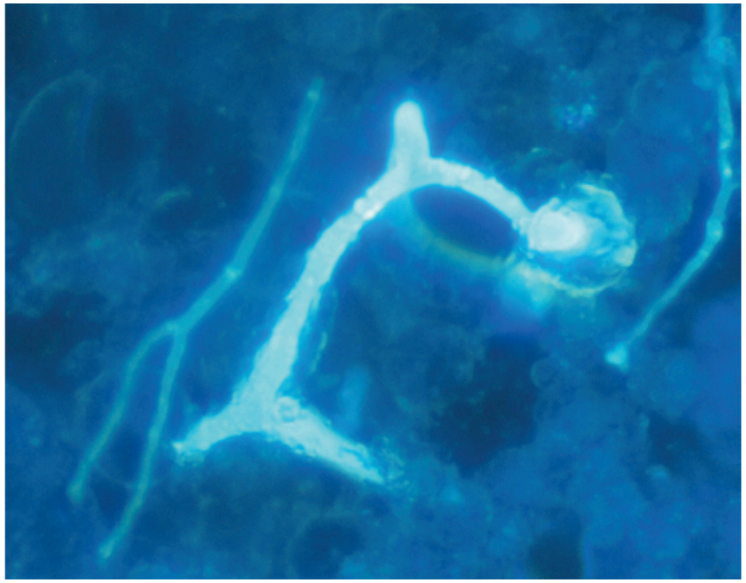




Alanio A, et al, O 1.5 Saturday 12th 6th TIMM

Detection of mixed infection

Deep skin biopy (burn)



Septate filament and non-septate filament

Aspergillus flavus
Purpureocillium lilacinum
Rhizopus microsporus

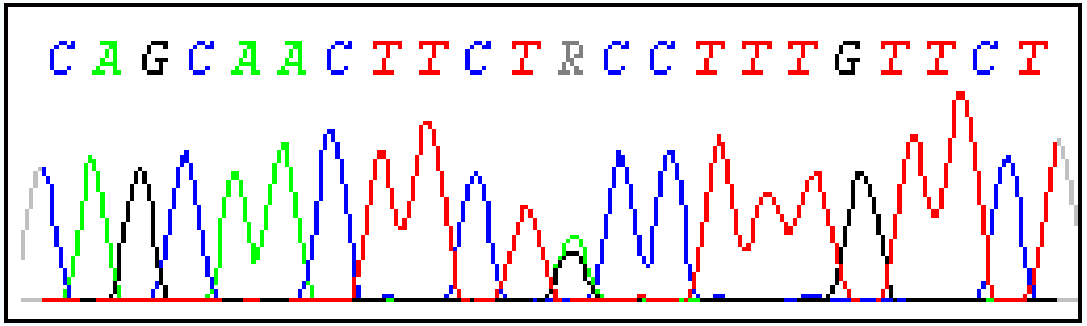
All these species have been recovered from other cutaneous specimens and identified at the species level based on different DNA targets

Alanio A, et al, O 1.5 Saturday 12th 6th TIMM

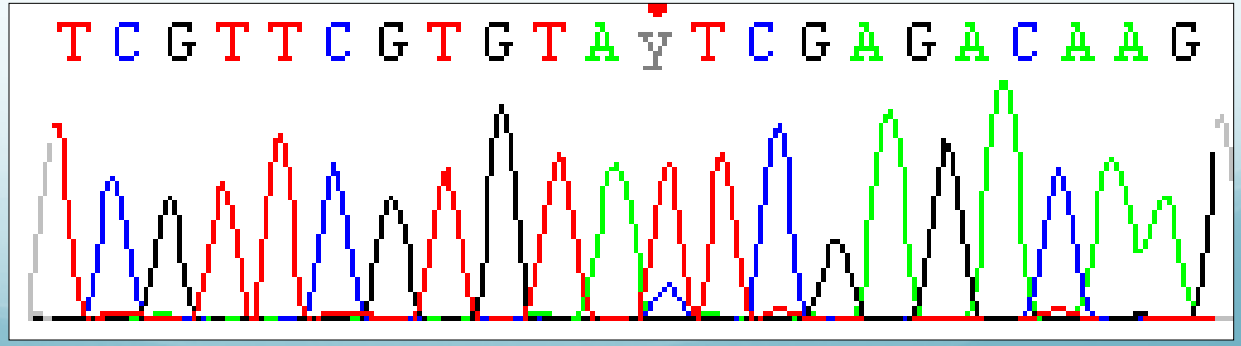
Which question?

- **Real-Time PCR**
  - Identify and quantify a microorganism
- **Broad identification of multiple micoorganisms**
  - Microarrays
  - PCR ESI Mass Spectromatry (PlexID)
- **Targeting sequencing**
  - Identification and quantitation of sequence variants in an amplicons; haplotype
- **Deep sequencing**
  - ***de novo* sequencing**
    - Sequencing of a genome that has not been sequenced before or does not have a reference
  - **Re-sequencing**
    - Sequencing a genome for which a reference sequence exists
    - Generally done for the purpose of mutation detection
  - **Others**

Mixed genotypes, SNPs



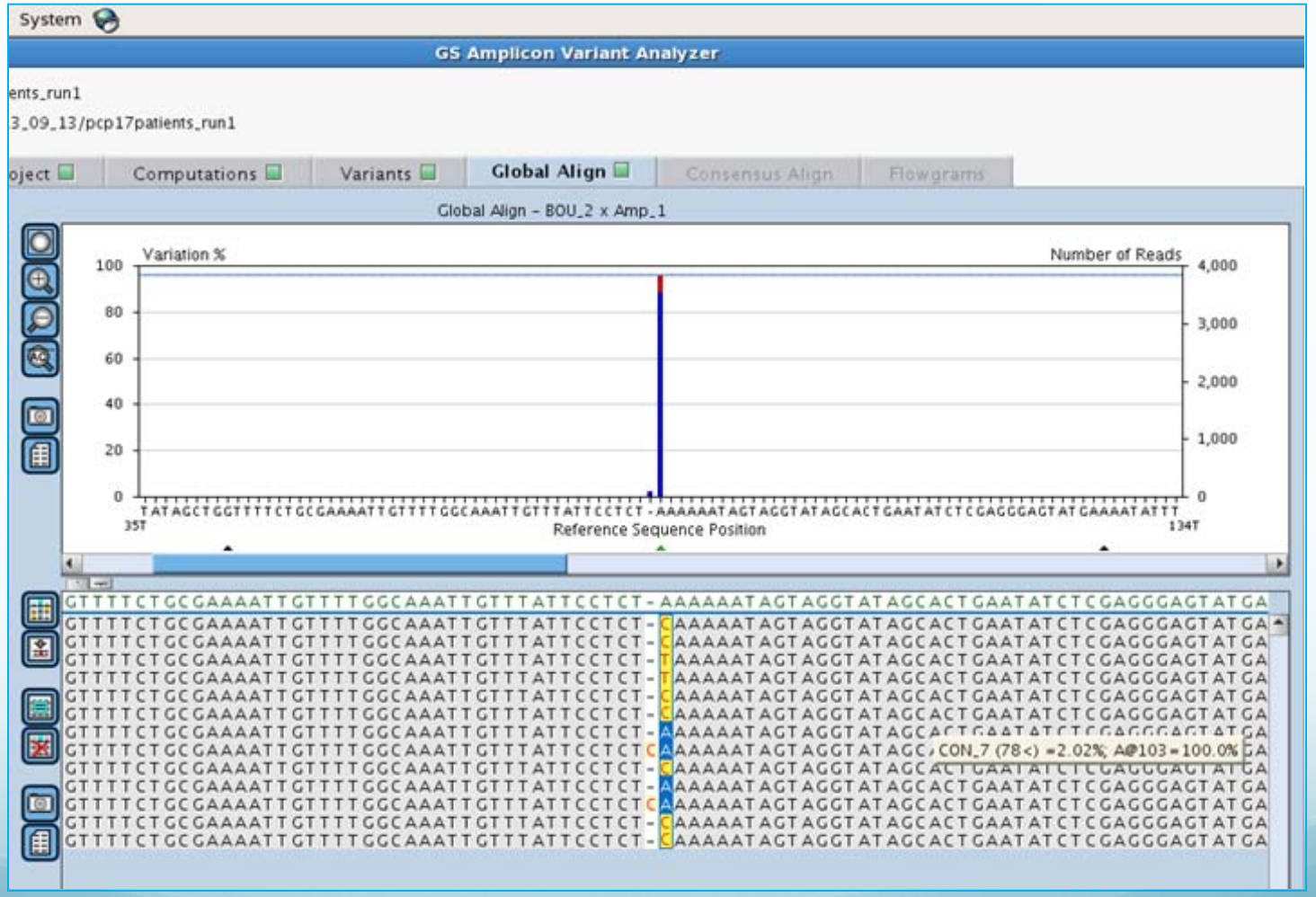
Mixed Sanger sequences: ≈ 50/50 ratio



GS Junior Titanium Chemistry Bi-directional Sequencing

- Improved variant calling vs. uni-directional, especially at longer read lengths
- Ideal when the entire amplicon needs to be sequenced (full length)
- Similar to Sanger-based approaches; required for some applications

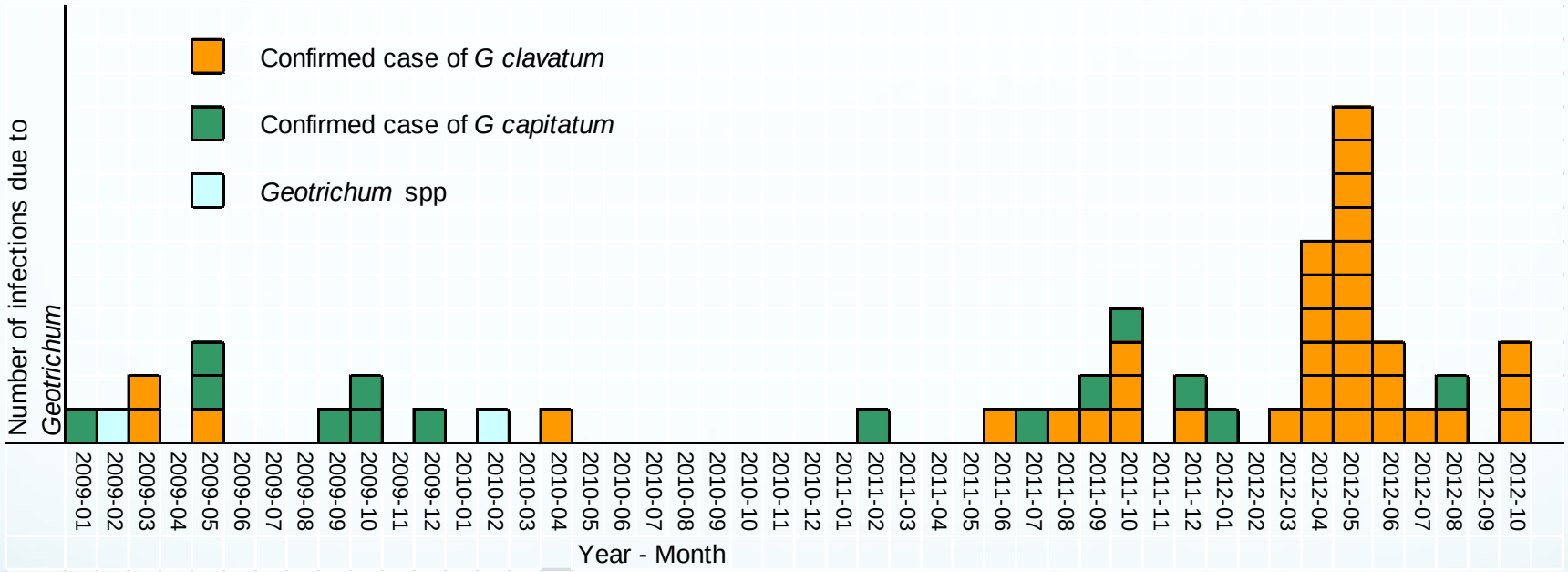
Insertion-deletion



Which question?

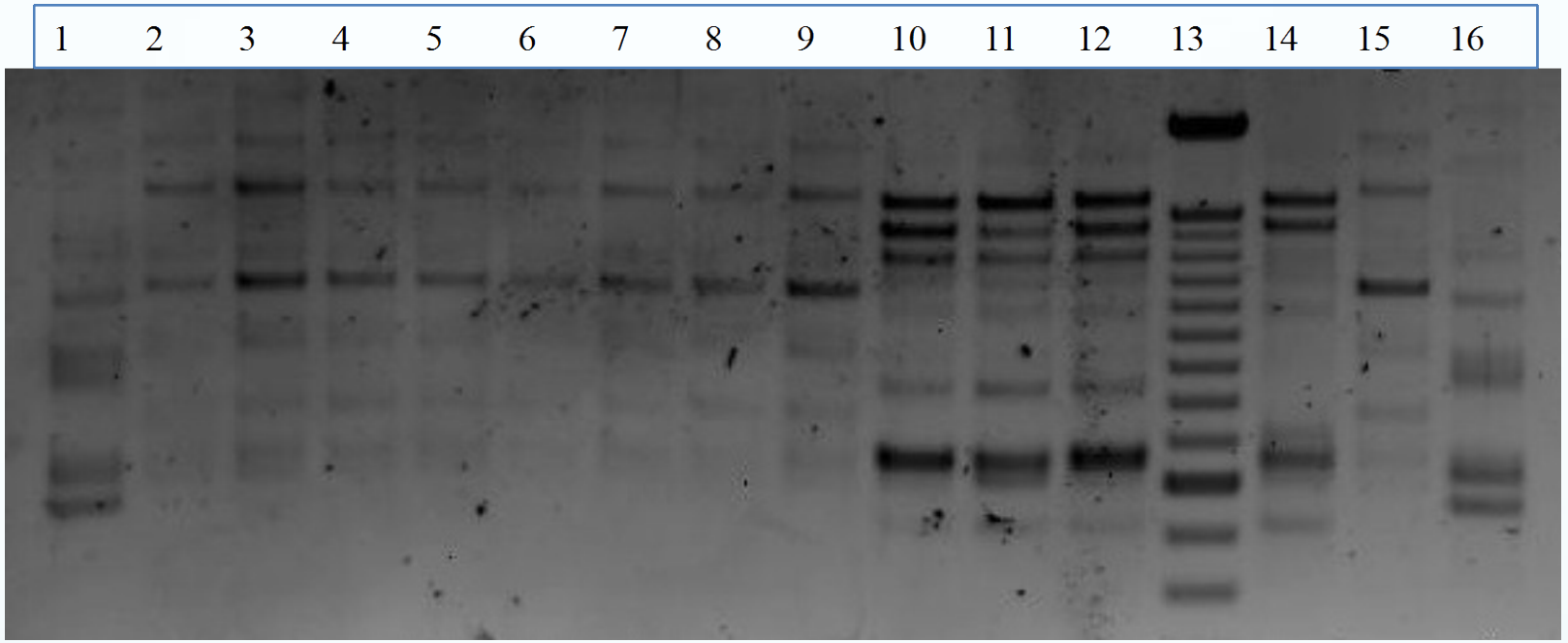
- **Real-Time PCR**
  - Identify and quantify a microorganism
- **Broad identification of multiple micoorganisms**
  - Microarrays
  - PCR ESI Mass Spectromatry (PlexID)
- **Targeting sequencing**
  - Identification and quantitation of sequence variants in an amplicons; haplotype
- **Deep sequencing**
  - ***de novo* sequencing**
    - Sequencing of a genome that has not been sequenced before or does not have a reference
  - **Re-sequencing**
    - Sequencing a genome for which a reference sequence exists
    - Generally done for the purpose of mutation detection
  - **Others**

Alert: cases of infections due to « *Geotrichum* »



- 52 cases of infections due to *Magnusiomyces capitatus* (*Geotrichum capitatum*) reported
- 36 cases of infections due to *Saprochaete clavata* (*Geotrichum clavatum*)

RAPD studies comparing *S. clavata, M. capitatus* and *Geotrichum candidum*



PCR results with OPE-4 primer of clinical isolate of *G. candidum* (line 1), *S. clavata* (lines 2 to 9) and *M. capitatus* (lines 10-12) clinical isolates. Reference strains: *M. capitatus* (line 14, CBS178.71), *S. clavata* (line 15, CBS162.80), and *G. candidum* (line 16, CBS425.71) - (line 13, size marker).

Sequencing of *S. clavata*

- **Illumina sequencing**
  - 1 reference strain
  - 18 clinical strains
  - 1 strain re-sequenced
- **Data analyses**
  - **Filtering Illumina reads**
  - **Inferring single nucleotide polymorphisms**
    - strains closely related
  - **Phylogenetic analysis**
  - ***de novo* genome assembly of strain CNRMA12.647**

Sequencing of *S. clavata*

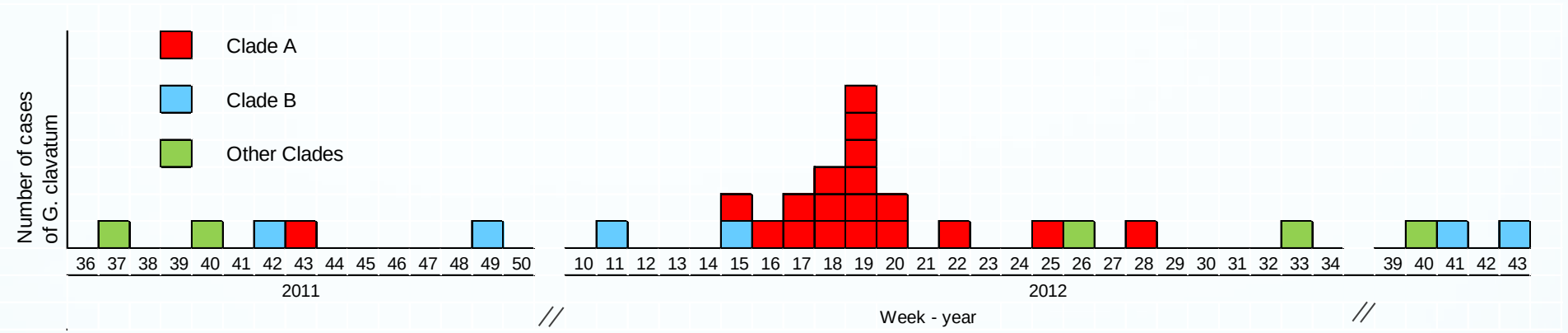
- **Illumina sequencing**
  - 1 reference strain
  - 18 clinical strains
  - 1 strain re-sequenced
- **Data analyses**
  - **Filtering**
  - **In**
  - **netic analysis**
  - ***novo* genome assembly of strain CNRMA12.647**

>10 different softwares (most home-made)

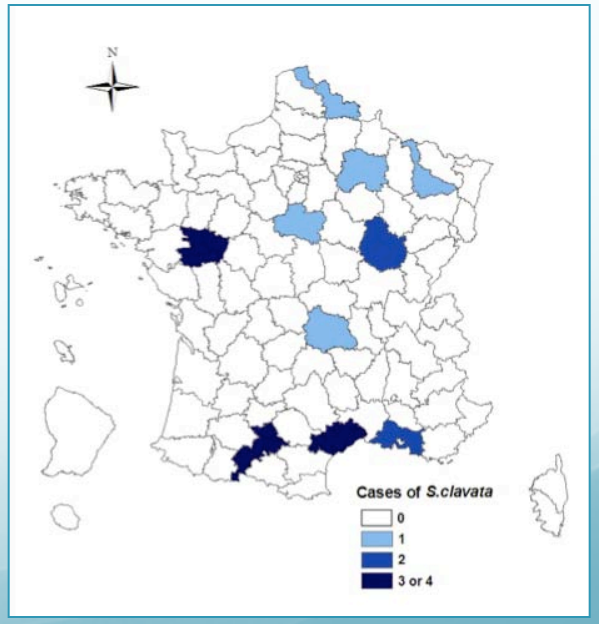
tide polymorphisms

elated

Cases of *S. clavata* by clade, Sept 2011 – Oct 2012



- 19 cases with infections due to *S. clavata* and belonging to the same clade (red)



Which question?

- **Real-Time PCR**
  - Identify and quantify a microorganism
- **Broad identification of multiple micoorganisms**
  - Microarrays
  - PCR ESI Mass Spectromatry (PlexID)
- **Targeting sequencing**
  - Identification and quantitation of sequence variants in an amplicons; haplotype
- **Deep sequencing**
  - ***de novo* sequencing**
    - Sequencing of a genome that has not been sequenced before or does not have a reference
  - **Re-sequencing**
    - Sequencing a genome for which a reference sequence exists
    - Generally done for the purpose of mutation detection
  - **Others**

NGS Applications

- **mRNA sequencing** (transcriptome sequencing)
  - Aligment with references
  - Statistics of gene expression level
- **miRNA sequencing** (small RNA sequencing)
- **ChIP sequencing** (protein-DNA relations)
- **Bisulfite sequencing** (pattern of methylation)

New Generation Sequencing, any use?

- **MANY** use**S**
  - Can answer a specific question not solved with previous methods
  - Necessity of accurate pre-analytical steps (validation of RNAs and DNAs)
  - Use the adapted technology (access to expensive equipment and chemistry)
  - Stay alert since extremely moving
- **Invest in bioinformatics**
  - Analysis and storage of computerized data

Acknowledgments

- Mycology unit
  - Pr F. Dromer
  - Marie Desnos-Olliver
  - Staff members
- Sequencing platform unit
  - V. Caro, S. Brisse
  - Alexis Criscuolo and collaborators

Hôpitaux Universitaires SAINT-LOUIS LARIBOISIÈRE FERNAND-WIDAL

St Louis hospital, founded in 1607

- Severine Mercier-Delarue
- Dr Alexandre Alanio
- Dr Jean Menotti
- Staff members of the hospital laboratory


Thank you for your attention