Next Generation Sequencing data Analysis at Genoscope

Jean-Marc Aury

Paris, 24 mars 2010

Introduction

- ✓ Presentation of Genoscope and NGS activities
- ✓ Overview of sequencing technologies
- ✓ Sequencing and assembly of prokaryotics genomes
- ✓ Annotating genomes using massive-scale RNA-Seq
- ✓ Future projects

Genoscope (National Sequencing center)

http://www.genoscope.cns.fr

✓ Among the largest sequencing center in Europe

✓ Part of the CEA Institut de Génomique since 05/2007

✓ Provide high-throughput sequencing data to the French Academic community, and carry out in-house genomic projects

✓ Involved in large genome projects : human genome project, arabidopsis, rice, …

✓ Coordination of large genome projects : tetraodon, paramecium, vitis, oikopleura, …

✓ and as well fungal genomes (botrytis, tuber) and prokaryotic genomes

Genoscope (National Sequencing center)

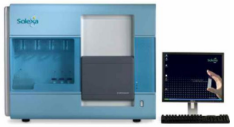
http://www.genoscope.cns.fr

✓NGS activities :

- ✓ **Sequencing of prokaryotic genomes (2007)**
- ✓ **RNA-Seq / Annotation of eukaryotic genomes (2008)**
- ✓ **SNP calling : identification of mutations (2008)**
- ✓ Metagenomic projects (2008)
- ✓ Sequencing of large eukaryotic genomes (2009/2010)
- ✓ Chip-Seq, detection of structural variations, … (2009/2010)

Genoscope (National Sequencing center)

✓ Sequencing capacity :

	19	ABI 3730
	3 Titanium	454/Roche
	2 GA2	Illumina
	1 d v3	Soli

http://www.genoscope.cns.fr

Genoscope (National Sequencing center)

Access to the Genoscope sequencing capacity by call for tender

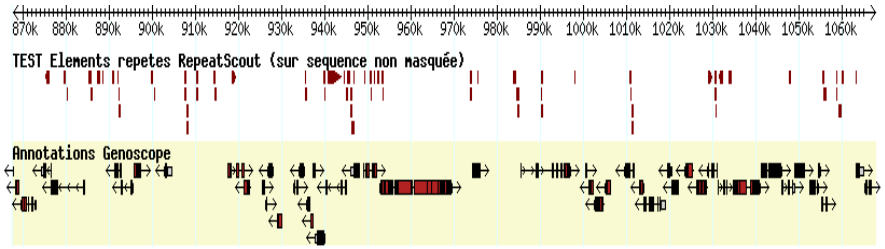
Project

Sequencing

Assembly / Finishing

Prokaryotics genomes annotation (MAGE)

Eukaryotics genomes annotation (GAZE)



Sequencing technologies

Applied Biosystems
ABI 3730XL

Roche / 454
Genome Sequencer FLX

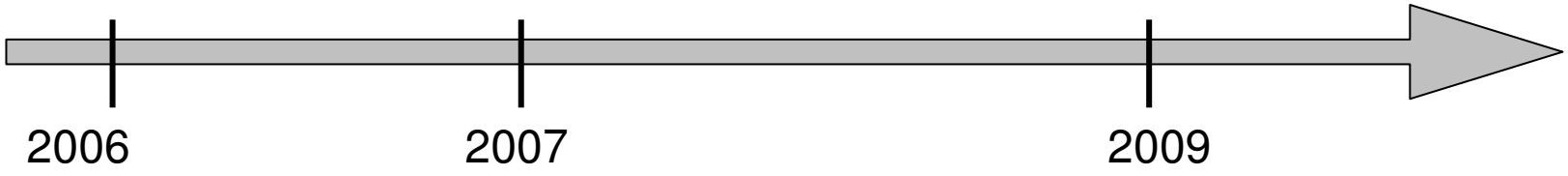
Illumina / Solexa
Genetic Analyzer

Applied Biosystems
SOLiD

What's different :
– Quantity and types of data
– Quality of data

Sequencing technologies

454 / Roche – Genome Sequence FLX



2006	2007	2009
Gs20	GSFLX Standard	GSFLX Titanium
20Mb / run	100Mb / run	500Mb / run
100bp / read	250bp / read	500bp / read

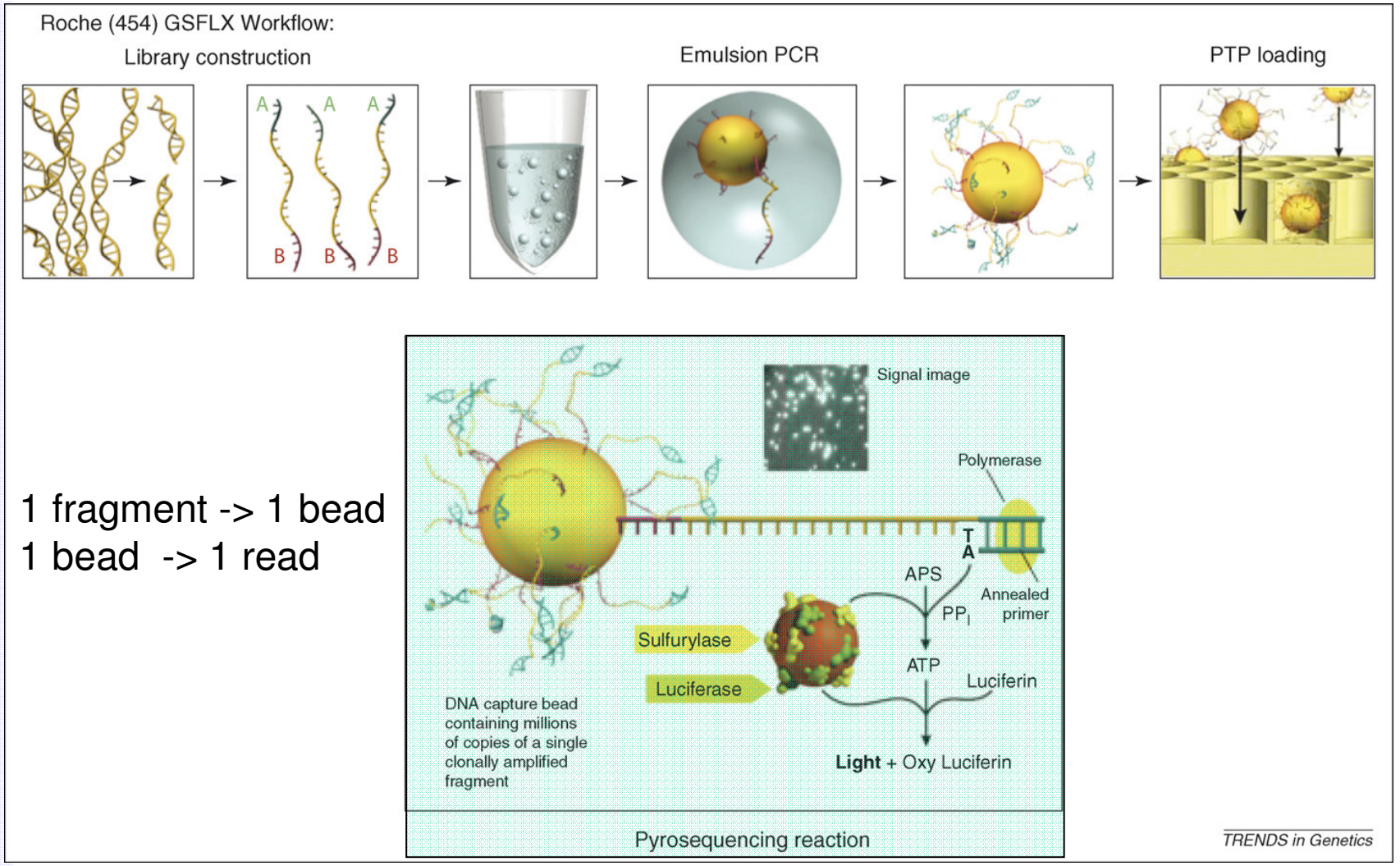
✓ Actual version (GSFLX Titanium) :

- ✓ Majority of 500bp reads
- ✓ Around 1.000.000 reads / run and 500Mbp / run
- ✓ Run duration : 8h

- ✓ High error rate in homopolymer sequences
- ✓ Good assemblies at 20X of coverage
- ✓ No cloning biases

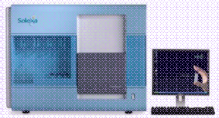
Sequencing technologies

454 / Roche – Genome Sequence FLX



1 fragment -> 1 bead
1 bead -> 1 read

Sequencing technologies



Illumina / Solexa – Genetic Analyzer

2007

GA 1
1Gb / run
32bp / read

2008

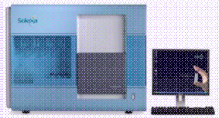
GA II
8Gb / run
50bp / read
Paired reads

2009

GA IIx
30Gb / run
100bp / read
10 days / run
Paired reads and mate pairs

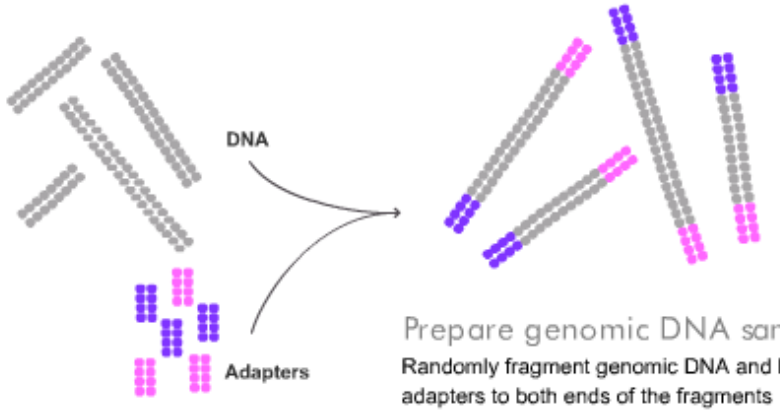
- ✓ Actual version (Genome Analyzer IIx):
  - ✓ Reads of 108bp
  - ✓ Around 240M reads / run and 25Gbp / run
  - ✓ 10 days / run
  - ✓ very low error rate (70% of perfect reads)
  - ✓ No indels => good complementarity with the 454 technology
  - ✓ No coverage gaps
  - ✓ Price

Sequencing technologies



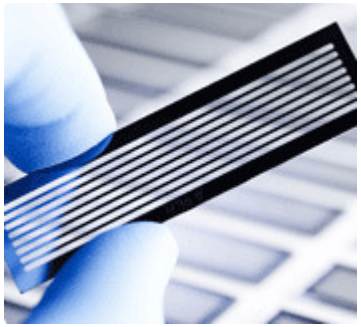
Illumina / Solexa – Genetic Analyzer

1

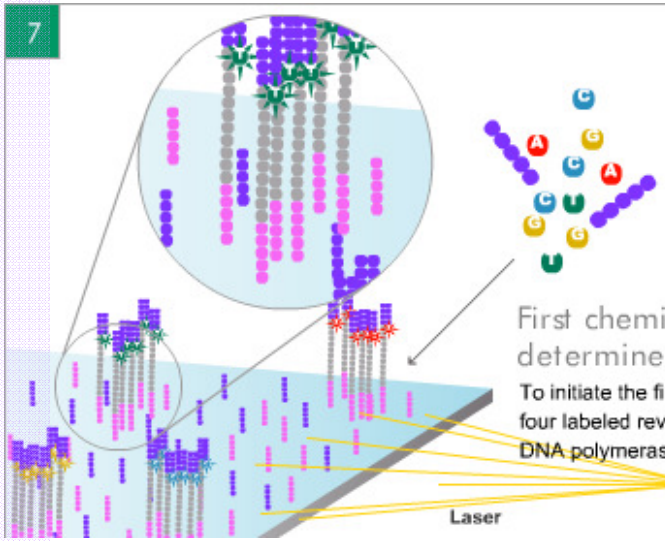


Prepare genomic DNA sample

Randomly fragment genomic DNA and ligate adapters to both ends of the fragments



7



First chemistry cycle: determine first base

To initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

8

Image of first chemistry cycle

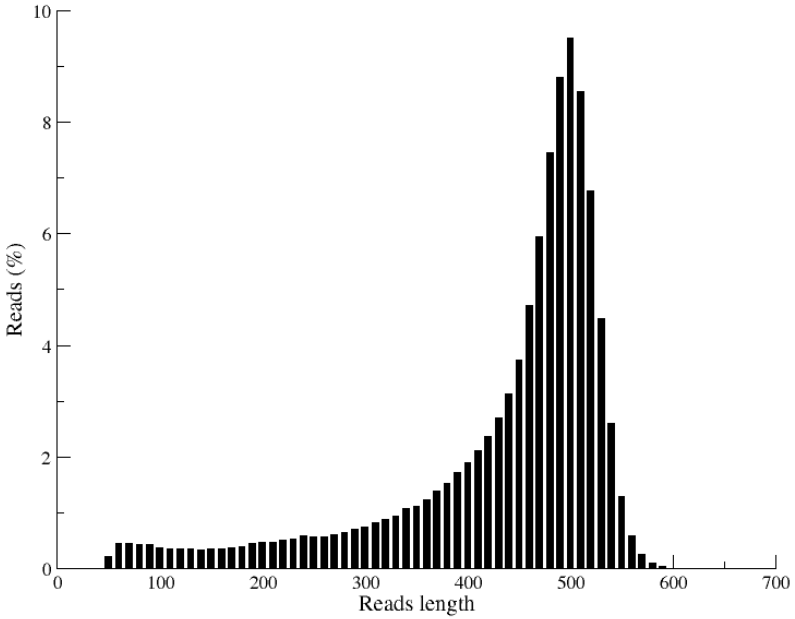
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

Before initiating the next chemistry cycle

The blocked 3' terminus and the fluorophore from each incorporated base are removed.

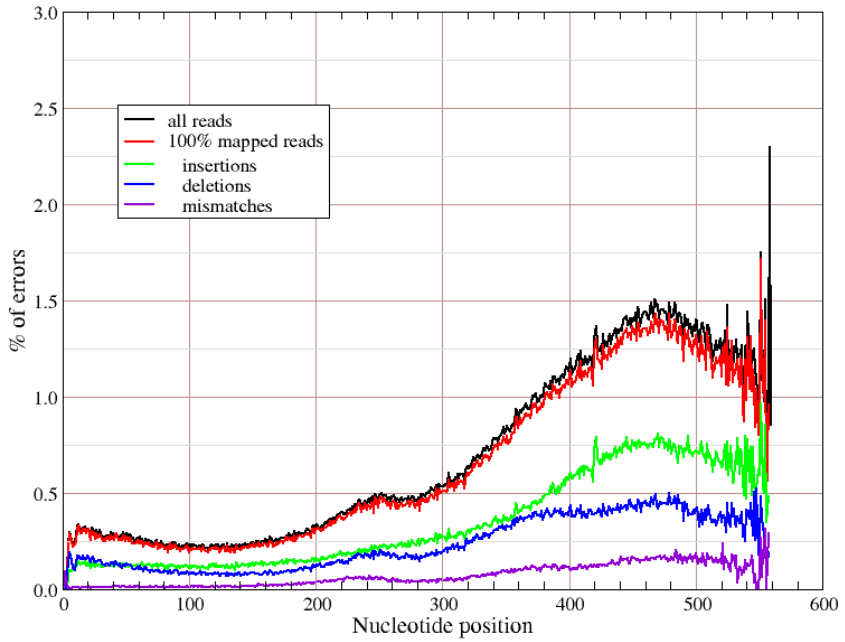
Sequencing technologies

454 / Roche – Genome Sequence FLX

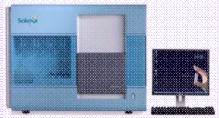


- ✓ ¼ run on *Acinetobacter baylyi* (3,5Mb)
- ✓ ~300.000 reads
- ✓ Cumulative size of 130Mb

- ✓ 99,9% of aligned reads
- ✓ Average error rate : 0,55%
- ✓ 37% deletions, 53% insertions, 10% substitutions.
- ✓ Errors accumulated around homopolymers => error rate is not constant



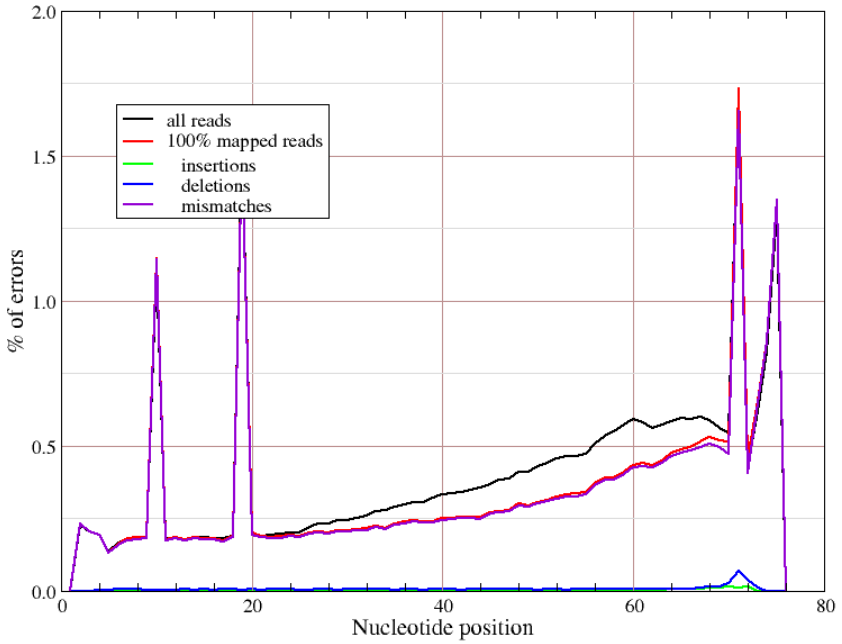
Sequencing technologies



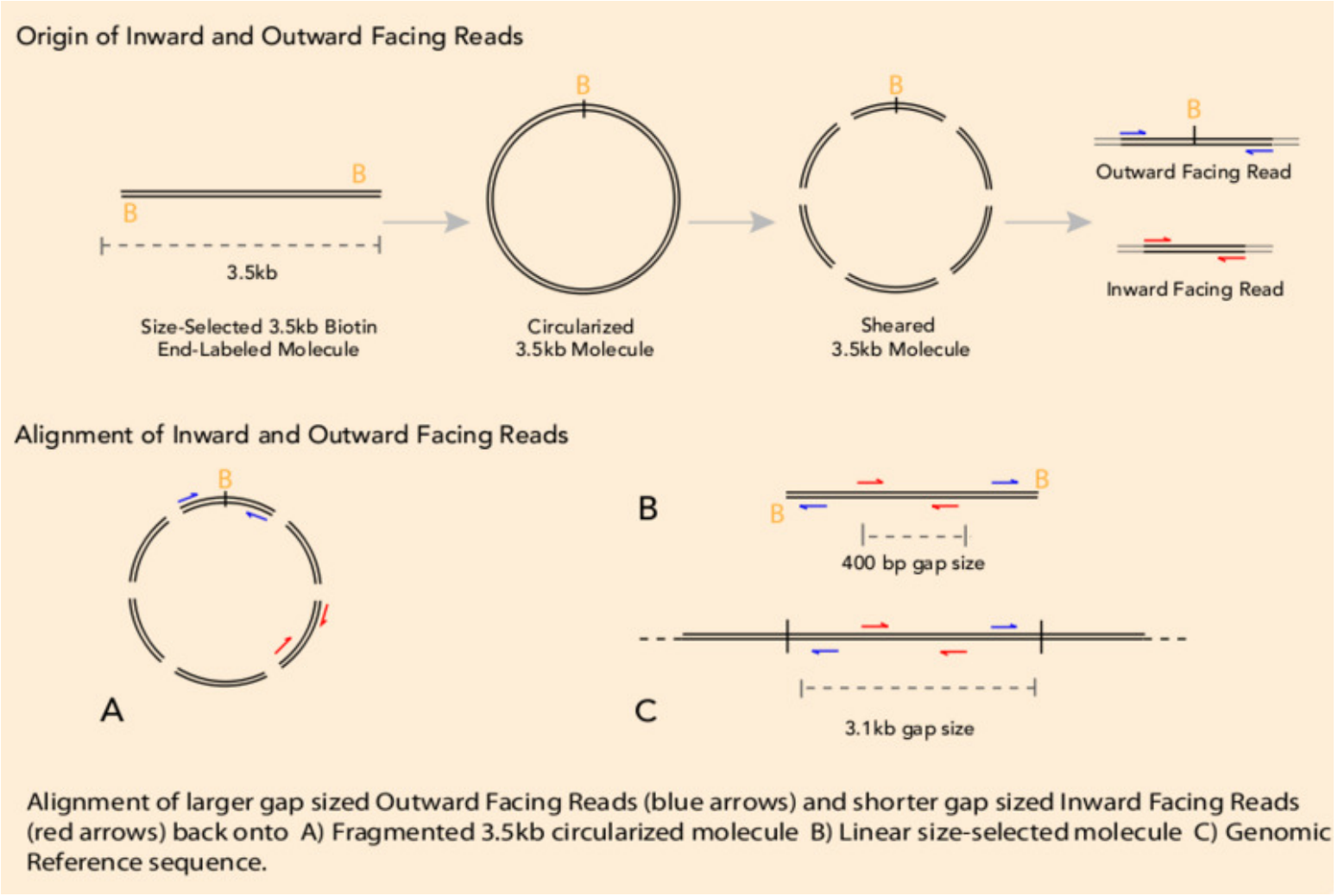
Illumina / Solexa – Genetic Analyzer

- ✓ 1 lane on *Acinetobacter baylyi* (3,5Mb)
- ✓ 11,4M reads
- ✓ cumulative size of 900Mb

- ✓ 98,5% aligned reads
- ✓ Average error rate : 0,38%
- ✓ 3% deletions, 2% insertions, 95% substitutions.

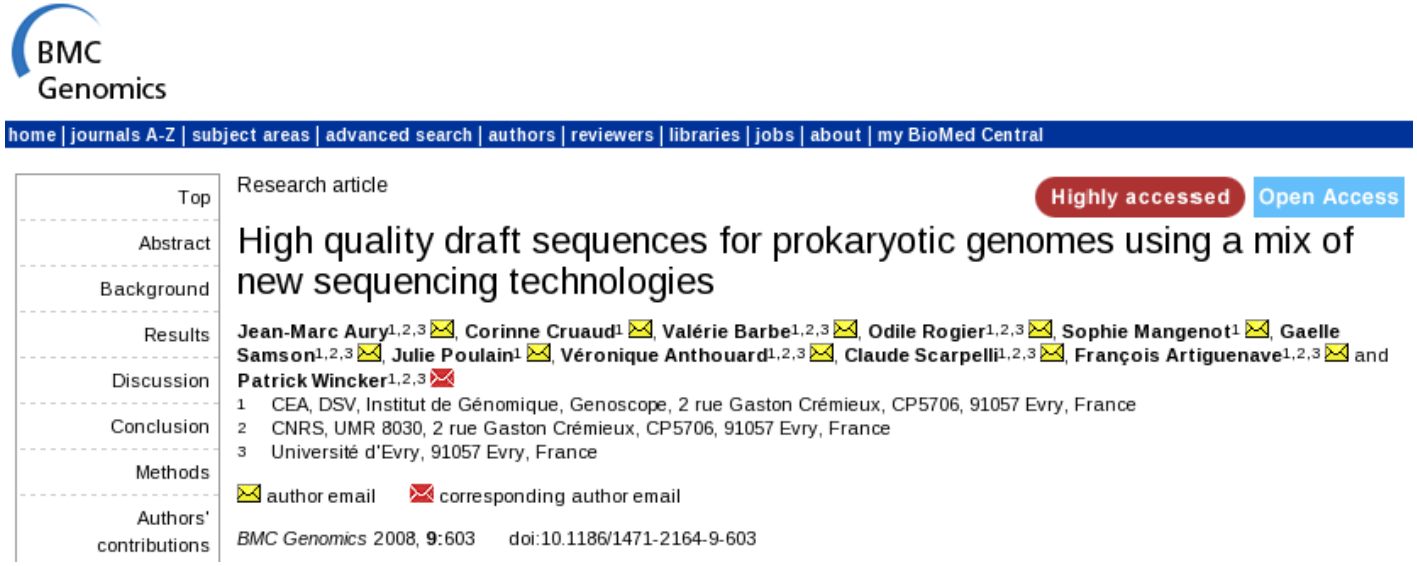


Sequencing technologies



Alignment of larger gap sized Outward Facing Reads (blue arrows) and shorter gap sized Inward Facing Reads (red arrows) back onto A) Fragmented 3.5kb circularized molecule B) Linear size-selected molecule C) Genomic Reference sequence.

Sequencing of prokaryotic genomes



BMC Genomics

home | journals A-Z | subject areas | advanced search | authors | reviewers | libraries | jobs | about | my BioMed Central

Top
Abstract
Background
Results
Discussion
Conclusion
Methods
Authors' contributions

Research article

Highly accessed | Open Access

High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies

Jean-Marc Aury[1,2,3], **Corinne Cruaud**[1], **Valérie Barbe**[1,2,3], **Odile Rogier**[1,2,3], **Sophie Mangenot**[1], **Gaelle Samson**[1,2,3], **Julie Poulain**[1], **Véronique Anthouard**[1,2,3], **Claude Scarpelli**[1,2,3], **François Artiguenave**[1,2,3] and **Patrick Wincker**[1,2,3]

1 CEA, DSV, Institut de Génomique, Genoscope, 2 rue Gaston Crémieux, CP5706, 91057 Evry, France
2 CNRS, UMR 8030, 2 rue Gaston Crémieux, CP5706, 91057 Evry, France
3 Université d'Evry, 91057 Evry, France

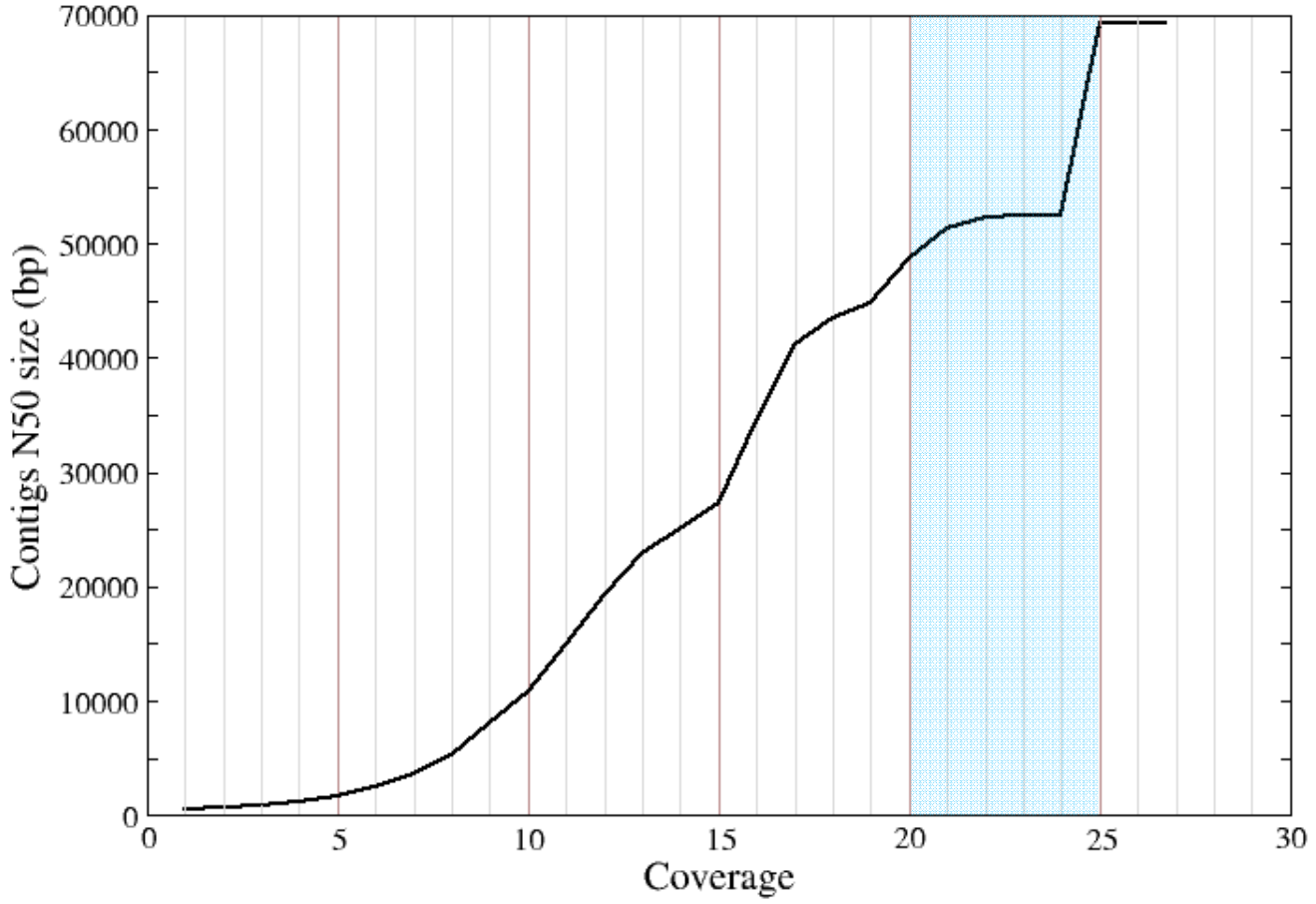
author email | corresponding author email

BMC Genomics 2008, **9**:603 doi:10.1186/1471-2164-9-603

Prokaryotic genomes sequencing

454 / Roche – Genome Sequence FLX

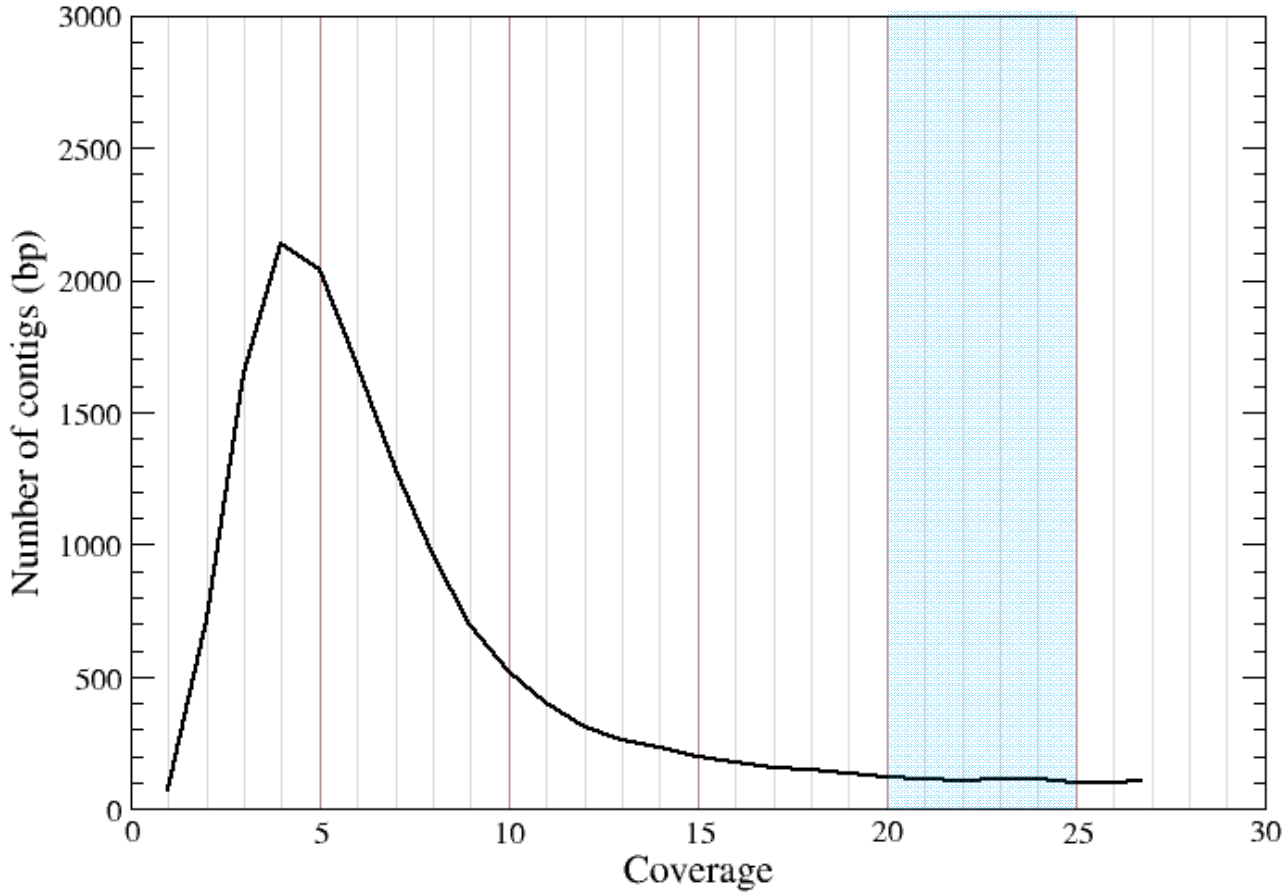
✓ Required genome coverage :



Prokaryotic genomes sequencing

454 / Roche – Genome Sequence FLX

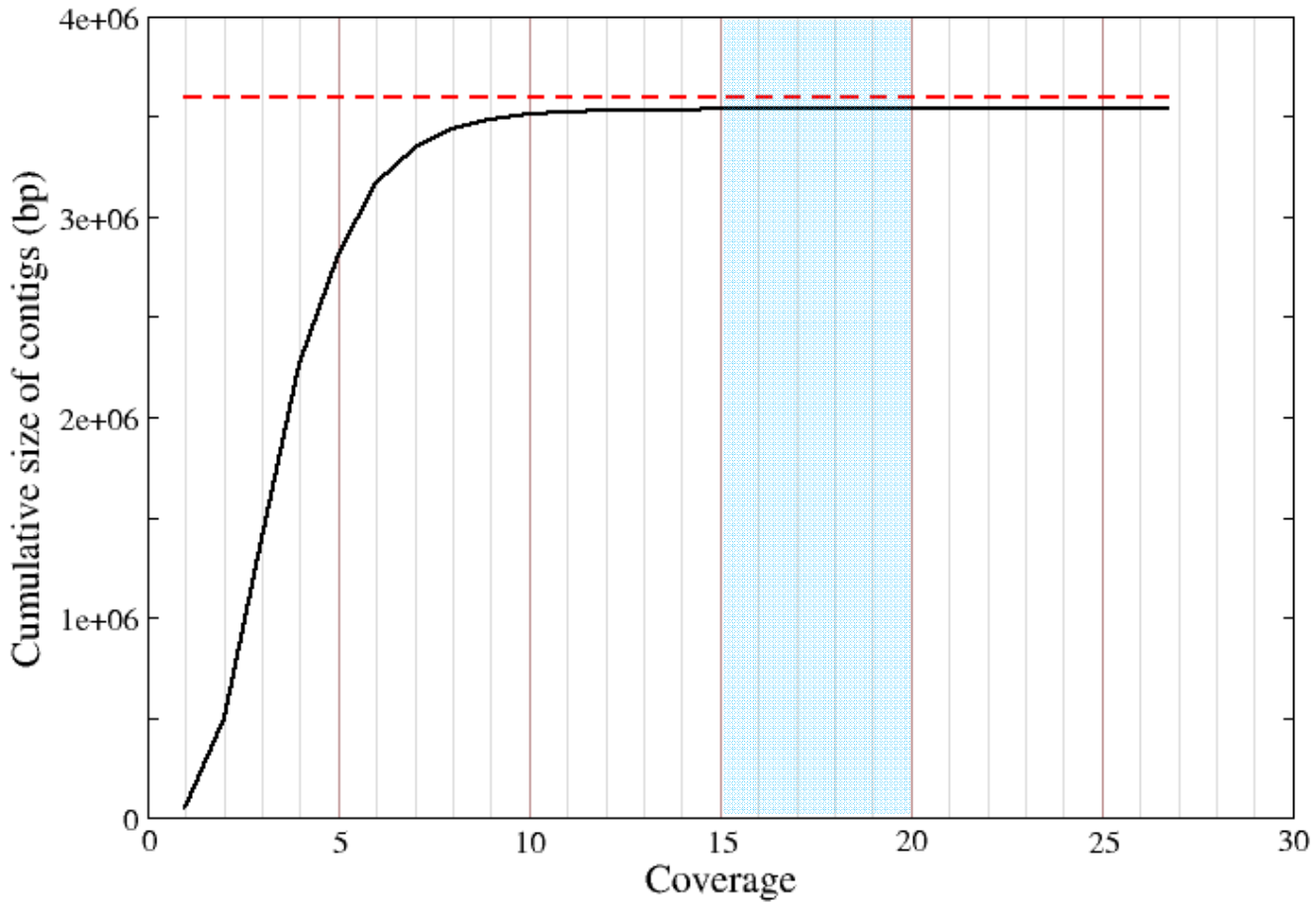
✓ Required genome coverage :



Prokaryotic genomes sequencing

454 / Roche – Genome Sequence FLX

✓ Required genome coverage :



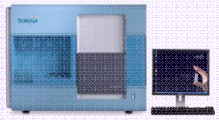
Procaryotic genomes sequencing

	Sanger	Unpaired 454	Unpaired + PE 454
Coverage	7.4X	20X	25X
Assembler	Arachne (Broad Institute)	Newbler (454/Roche)	Newbler (454/Roche)
# of contigs	173	**119**	**119**
Contigs N50 (Kb)	39.0	48.7	**58.2**
# of scaffolds	**2**	119	10
Scaffolds N50 (Kb)	**2,200**	48.7	1,000
Assembly size (% of reference)	3.417Mb (95%)	3.542 Mb (98%)	**3.544 Mb (98%)**
Mis-assemblies	0	0	0
# of errors	3,442	420	431
Substitutions	2,494	67	75
Insertions / Deletions	948	353	356

Prokaryotic genomes sequencing

- Good assembly structure (more scaffolds => library of 3 and 10Kb for the Sanger assembly against 3Kb for the 454 assembly)
- Good representativeness of the genome (homogeneous coverage)
- Error rate is still too high for a high quality draft : ~ 1 error / 8,5Kb. The vast majority are indels (introducing frameshifts in coding regions)
- Rational : polish the consensus of the 454 assembly with a complementary technology.

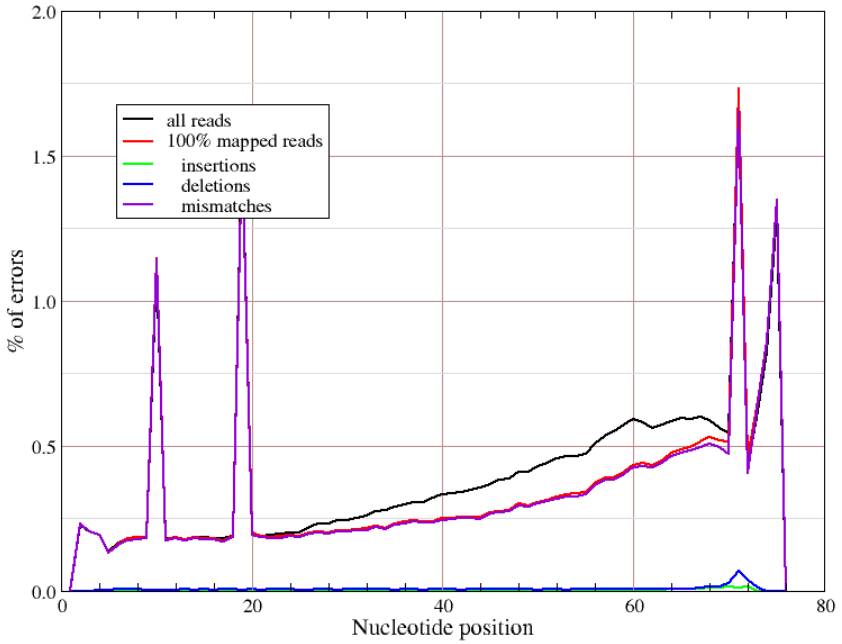
Prokaryotic genomes sequencing



Illumina / Solexa – Genetic Analyzer

- ✓ 1 lane on *Acinetobacter baylyi* (3,5Mb)
- ✓ 11,4M reads
- ✓ cumulative size of 900Mb

- ✓ 98,5% aligned reads
- ✓ Average error rate : 0,38%
- ✓ 3% deletions, 2% insertions, 95% substitutions.

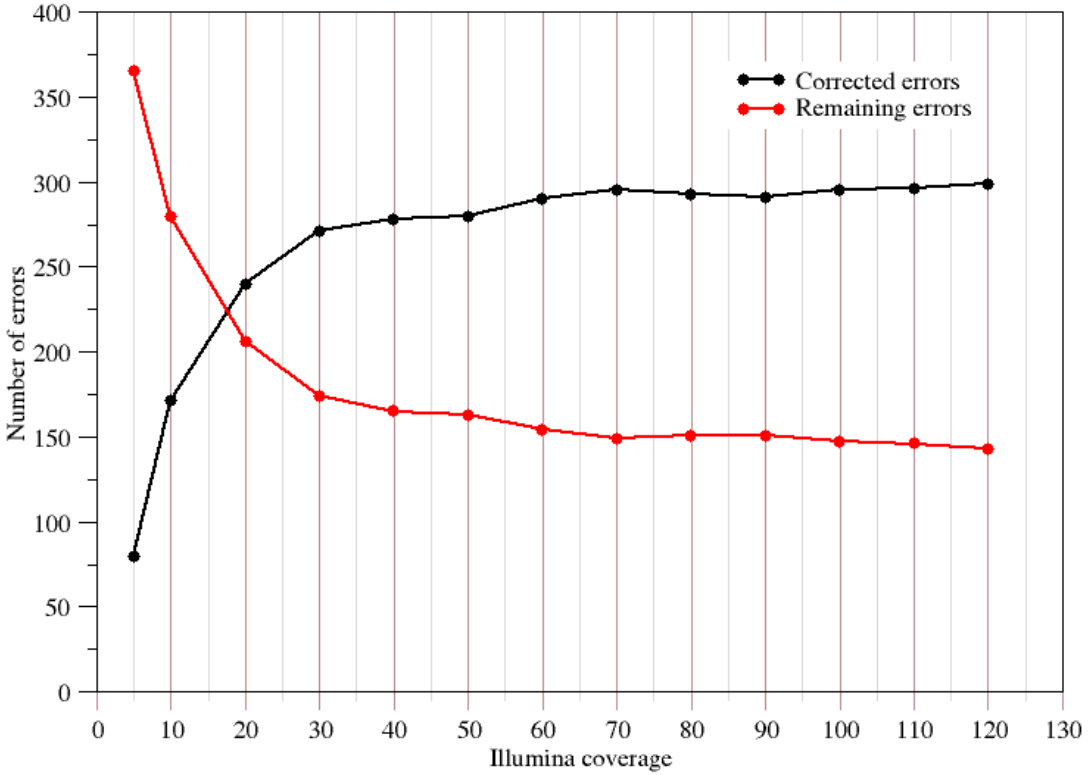


Prokaryotic genomes sequencing

- Alignment of illumina reads on the 454 assembly using Soap (gapped alignments) : 2 mismatches and 3 gaps
- Only uniquely mapped reads were retained
- Each difference was kept only if it met the following three criteria :
  - Error is not located in the first 5bps or the last 5bps
  - Quality of the considered base, the previous and the next one are above 20
  - Remaining sequences (around the error) are not homopolymers
- Each detected difference is considered as a sequencing error if :
  - At least three reads detected the given error
  - 70% of the reads located at that position agree

Prokaryotic genomes sequencing

- Illumina sequencing coverage :



- At 50X, still remains 163 errors :
  - 51 were attributed to errors in the original consensus sequence or to the presence of variations occurring during cultivation
  - 112 are found in repetitive regions or low coverage (with illumina reads) regions (contigs extremity).

Prokaryotic genomes sequencing

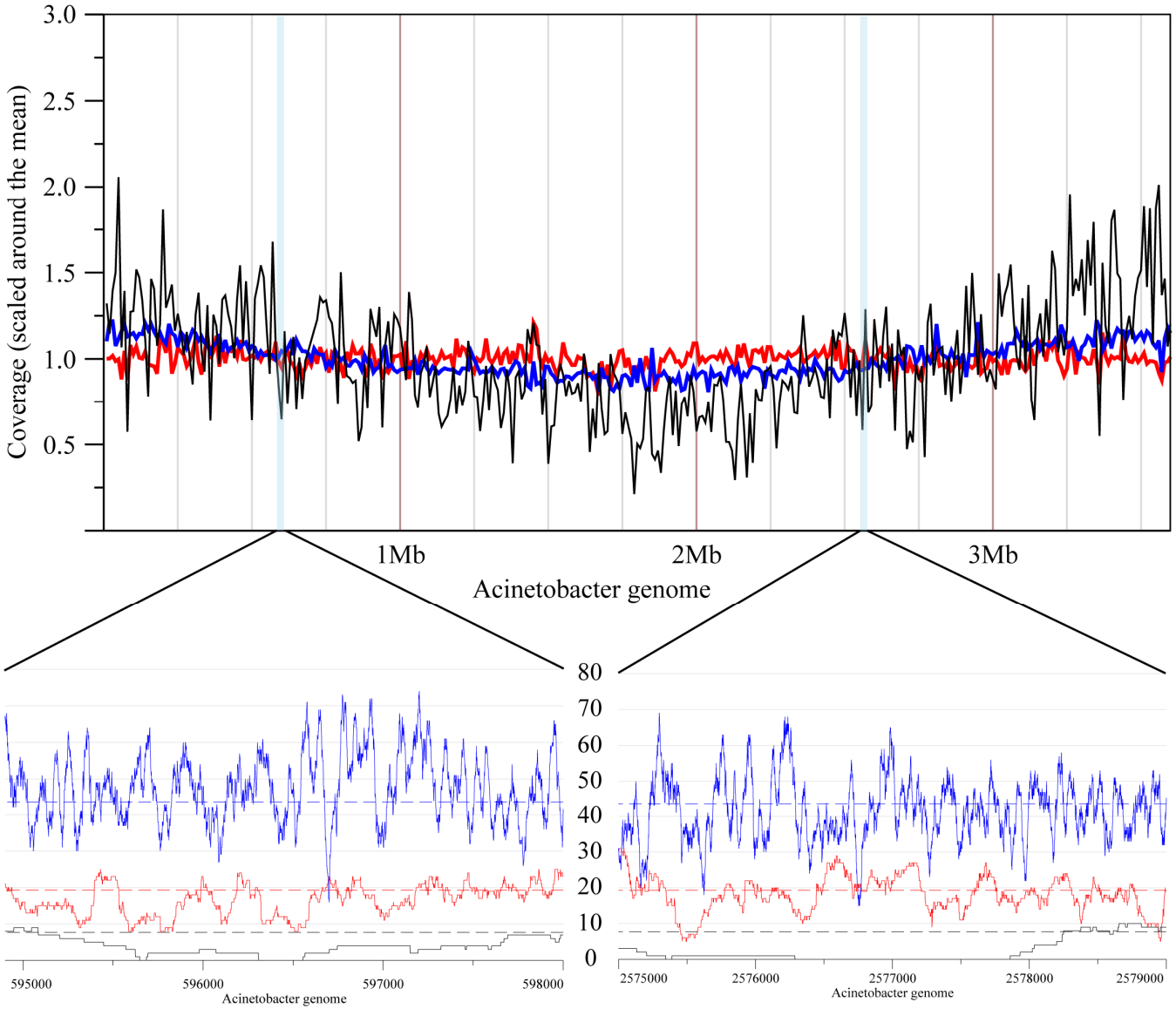
Step	Sequenced reads	Uniquely mapped reads	Filtered reads
Number of reads	5.000.000	4.543.370	3.497.539
Number of bases	180.000.000	163.561.320	60.680.570
Genome coverage	50,0X	45,5X	16,9X

- Alignment of illumina reads on the 454 assembly using Soap (gapped alignments) : 2 mismatches and 3 gaps
- Only uniquely mapped reads were retained
- Each difference was kept only if it met the following three criteria :
  - Error is not located in the first 5bps or the last 5bps
  - Quality of the considered base, the previous and the next one are above 20
  - Remaining sequences (around the error) are not homopolymers

Prokaryotic genomes sequencing

	Sanger	Unpaired + PE 454	unpaired + paired 454 with Illumina / Solexa GA1
Coverage	7.4X	25X	25X and 50X
Assembler	Arachne (Broad Institute)	Newbler (454/Roche)	Newbler (454 / Roche)
# of contigs	173	119	119
Contigs N50 (Kb)	39.0	58.2	58.2
# of scaffolds	2	10	10
Scaffolds N50 (Kb)	2,200	1,000	1,000
Assembly size (% of reference)	3.417Mb (95%)	3.544 Mb (98%)	3.544 Mb (98%)
Mis-assemblies	0	0	0
# of errors	3,442	431 (1 error / 8Kb)	**163 (1 error / 22Kb)**
Substitutions	2,494	75	71
Insertions / Deletions	948	356	92

Prokaryotic genomes sequencing



Prokaryotic genomes sequencing

Genomic DNA

↓ Roche/454 sequenced paired-end library to a ~7x fragment size coverage (for 3Kb fragments)

↓ Add 454 unpaired data to a final 25x coverage

Newbler assembly

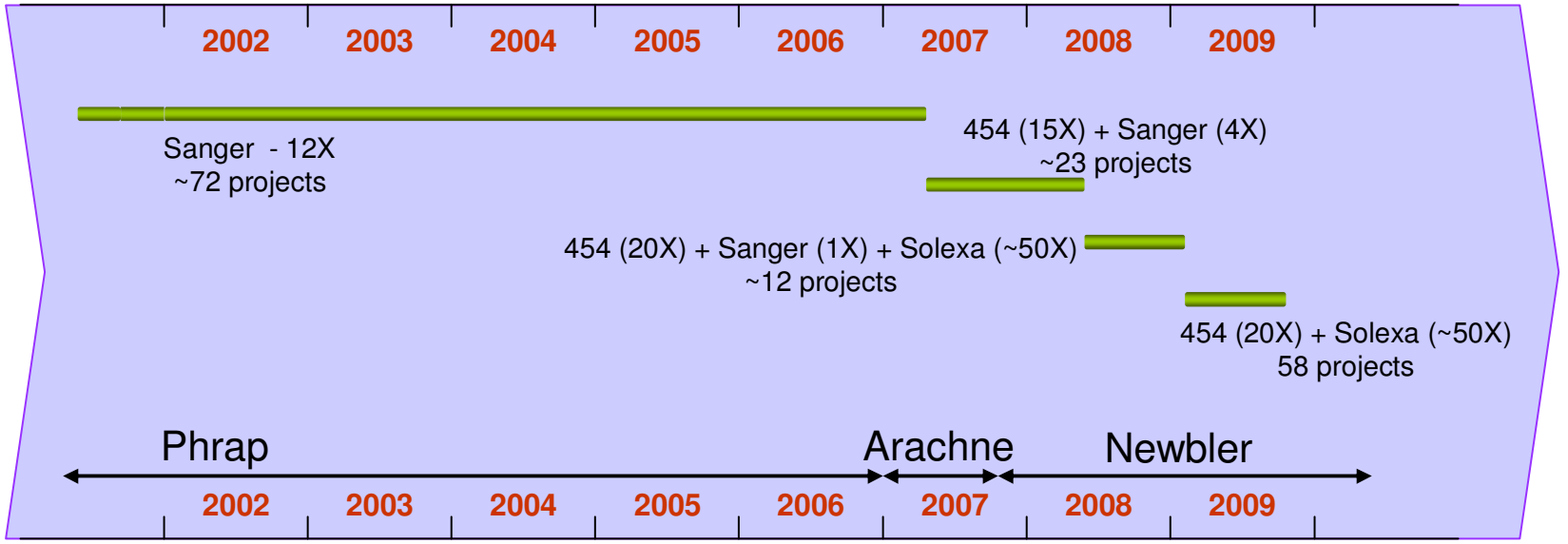
↓ Correct errors with ~50x Solexa/illumina short reads data

High quality draft (< 10^{-4} error rate)

Microbial Genome Sequencing Evolution

✓ Propose a strategy to sequence prokaryotic genomes, accounting for assembly quality and costs

✓ Mixing 454 and illumina technologies to obtain high quality drafts (454 provide read length and illumina low error rate)



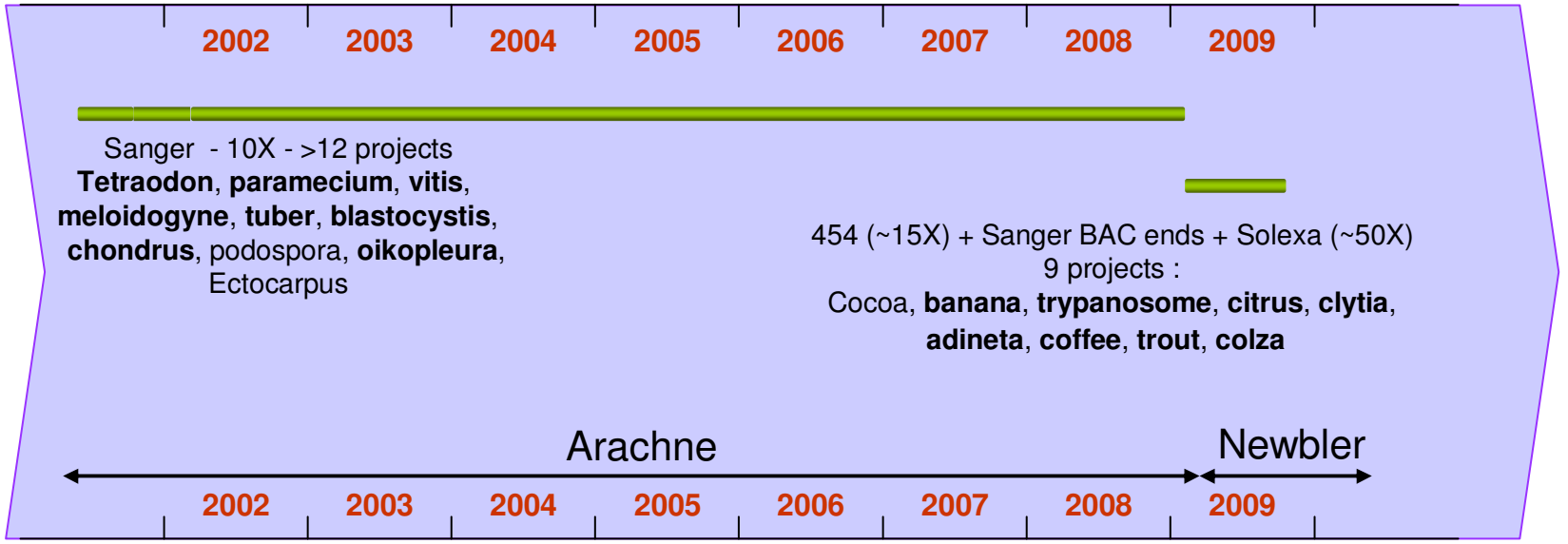
Prokaryotic genomes sequencing

	Sanger	unpaired + paired 454 + illumina GAI 36bp	Unpaired 454 + illumina GAIIx MP 4,5Kb 76bp	Illumina GAIIx MP 4,5Kb 76bp
Coverage	7.4X	25X and 50X	25X and 40X	22X
Assembler	Arachne (Broad Institute)	Newbler (454 / Roche)	Newbler (454 / Roche)	Soap (BGI)
# of contigs	173	119	44	495
Contigs N50 (Kb)	39.0	58.2	197	12.2
# of scaffolds	2	10	9	84
Scaffolds N50 (Kb)	2,200	1,000	1,009	818
Assembly size (% of reference)	3.417Mb (95%)	3.544 Mb (98%)	3.567 (99%)	3.674 Mb (102%)
Mis-assemblies	0	0	0	0
# of errors	3,442	163 1 error / 22Kb	63 1 error / 55Kb	35 1 error / 100Kb
Substitutions	2,494	71	60	33
Insertions / Deletions	948	92	4	2

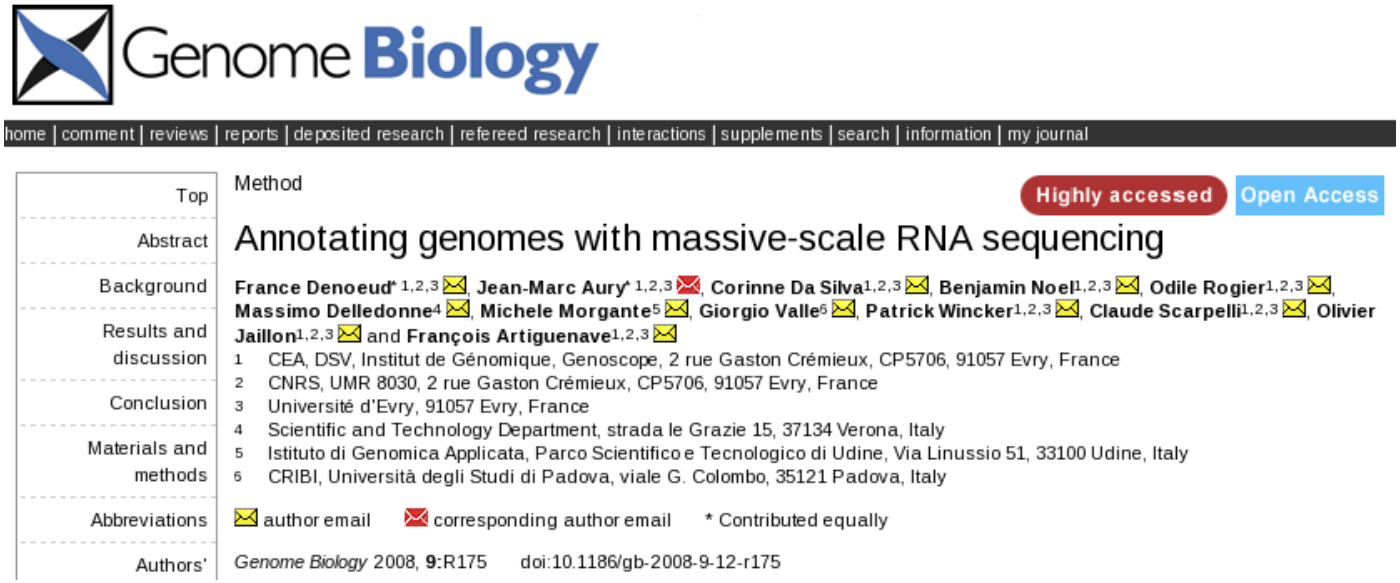
Eukaryotic Genome Sequencing Evolution

✓ Extend prokaryotic strategy to eukaryotic genomes

✓ Sanger sequencing is still used to sequence long DNA fragments : >20Kb, BAC ends, …



Annotating genomes using RNA-Seq



Genome Biology

home | comment | reviews | reports | deposited research | refereed research | interactions | supplements | search | information | my journal

Top
Abstract
Background
Results and discussion
Conclusion
Materials and methods
Abbreviations
Authors'

Method

Highly accessed Open Access

Annotating genomes with massive-scale RNA sequencing

France Denoeud* [1,2,3], **Jean-Marc Aury*** [1,2,3], **Corinne Da Silva**[1,2,3], **Benjamin Noel**[1,2,3], **Odile Rogier**[1,2,3], **Massimo Delledonne**[4], **Michele Morgante**[5], **Giorgio Valle**[6], **Patrick Wincker**[1,2,3], **Claude Scarpelli**[1,2,3], **Olivier Jaillon**[1,2,3] and **François Artiguenave**[1,2,3]

1 CEA, DSV, Institut de Génomique, Genoscope, 2 rue Gaston Crémieux, CP5706, 91057 Evry, France
2 CNRS, UMR 8030, 2 rue Gaston Crémieux, CP5706, 91057 Evry, France
3 Université d'Evry, 91057 Evry, France
4 Scientific and Technology Department, strada le Grazie 15, 37134 Verona, Italy
5 Istituto di Genomica Applicata, Parco Scientifico e Tecnologico di Udine, Via Linussio 51, 33100 Udine, Italy
6 CRIBI, Università degli Studi di Padova, viale G. Colombo, 35121 Padova, Italy

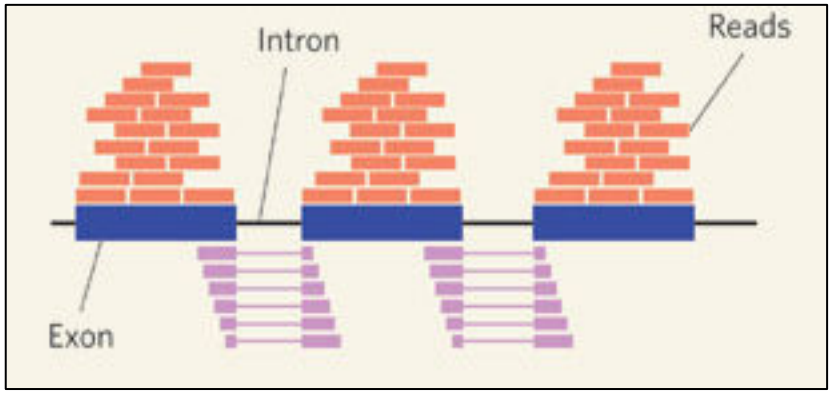
author email corresponding author email * Contributed equally

Genome Biology 2008, **9**:R175 doi:10.1186/gb-2008-9-12-r175

Annotating genomes using RNA-Seq

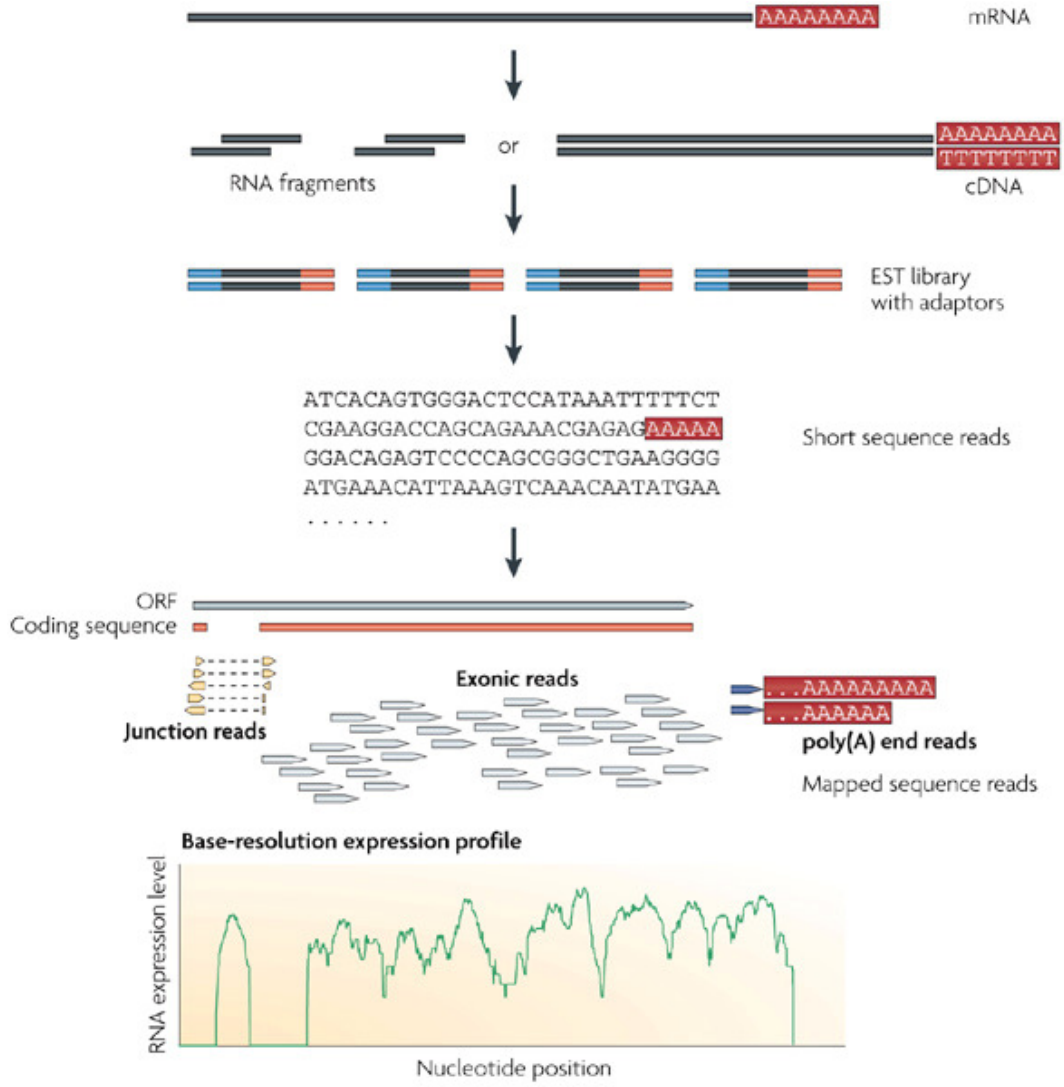
- Goal : annotate eukaryotic genomes using transcriptomic data from ultra-high throughput sequencers : Illumina and Solid

- Difficulties :
  - Predict complete gene structures with 40 bp reads
  - Align short reads to exon/exon junctions (mapping algorithms allow a limited number of gaps during alignments).



Molecular biology: Power sequencing. Brenton R. Graveley. Nature 453, 1197-1198(26 June 2008)

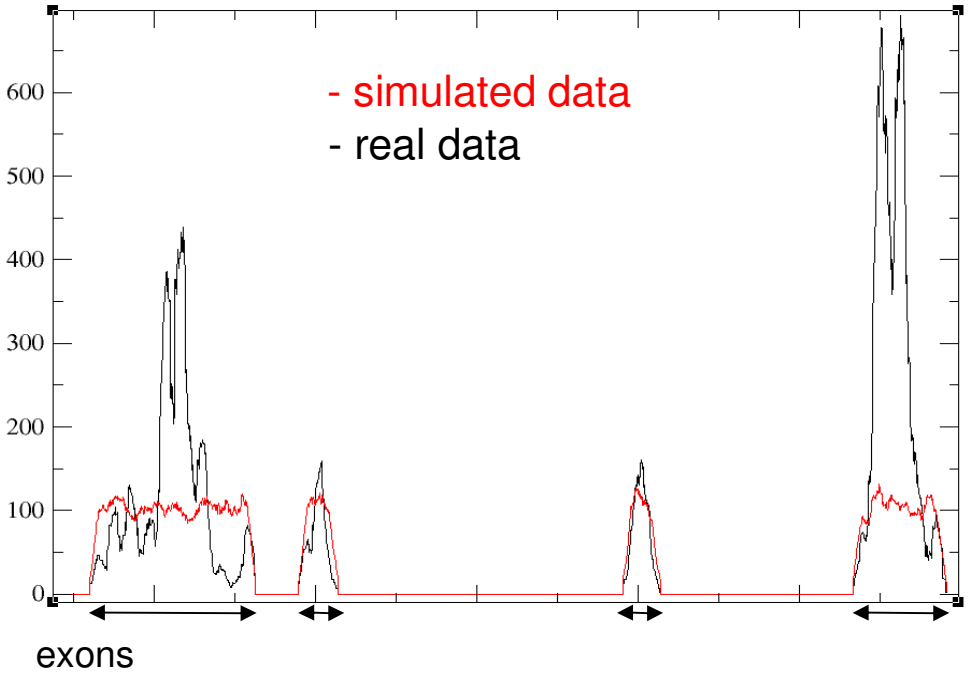
Typical RNA-Seq experiments



Nature Reviews | Genetics

Annotating genomes using RNA-Seq

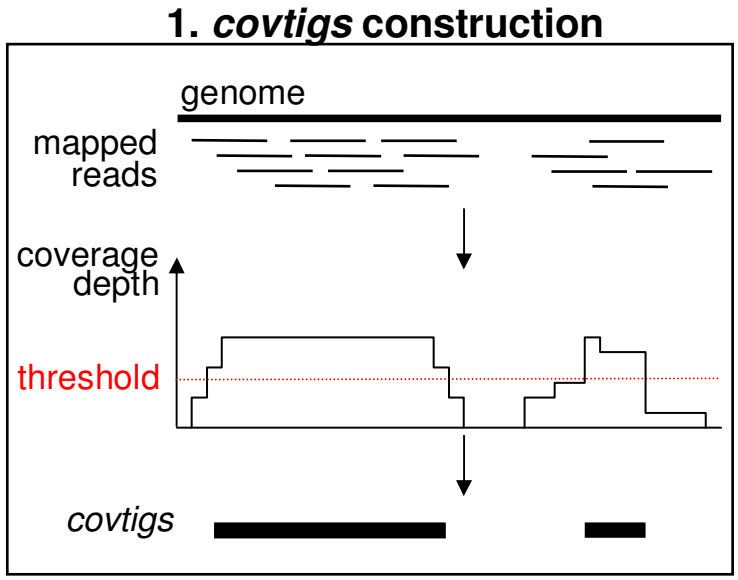
Coverage heterogeneity at the exon and gene level



At the exon level
- experimental biases
- alternative splicing

Biases between 3' and 5' at the gene level

Annotating genomes using RNA-Seq

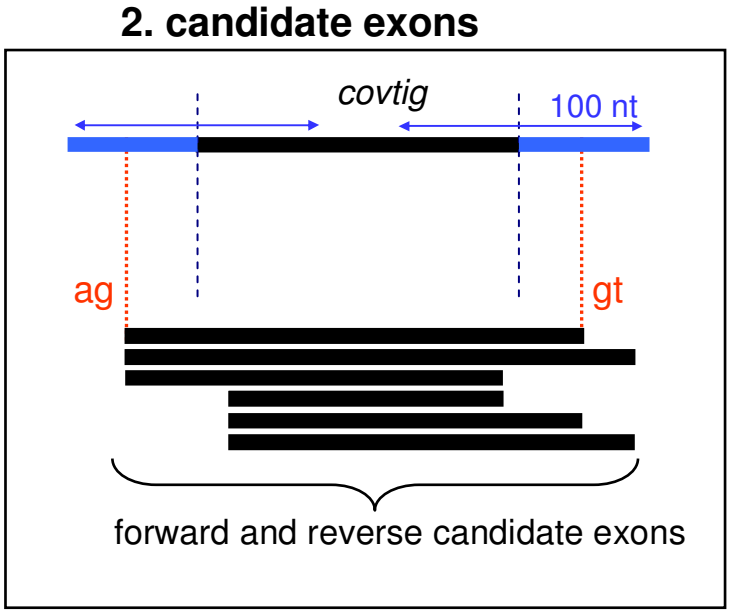


Step 1. covtigs construction

Annotating genomes using RNA-Seq

Inaccurate detection of splicing junctions

Improvement :
Extension of covtigs using unmapped reads

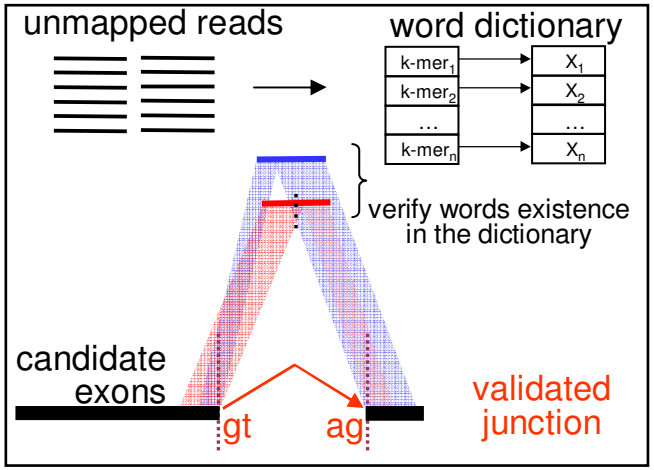


Step 2. extraction of candidate exons

Annotating genomes using RNA-Seq

Step 3: Validation of exon/exon junctions

Validation of junctions between candidate exons using a word dictionary built from the unmapped reads.



covtig1 ....**GGTGTTCACTACTTACCCATGT**.....AG**ATCTACACACTTTTAGAAGCCTGAAAG**.... **covtig2**

Kmers

TTACCCAT
CTTACCCAT
ACTTACCCAT
TACTTACCCAT
CTACTTACCCAT
ACTACTTACCCAT
CACTACTTACCCAT
TCACTACTTACCCAT
TTCACTACTTACCCAT
GTTCACTACTTACCCAT

ATCTACACACTTTTAGA
ATCTACACACTTTTAG
ATCTACACACTTTTA
ATCTACACACTTTT
ATCTACACACTTT
ATCTACACACTT
ATCTACACACT
ATCTACACAC
ATCTACACA
ATCTACAC

Junction validation

Creation of the dictionnary

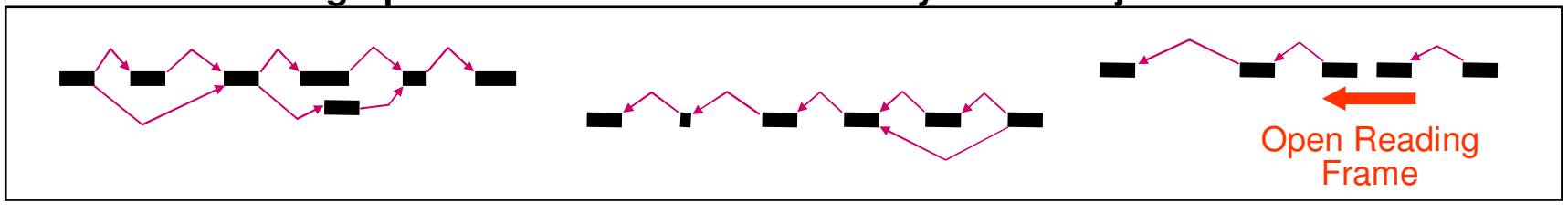
Unmapped reads

TGTTCACTACTTACCCATATCTACACACTTTTAGAA
TCACTACTTACCCATATCTACACACTTTTAGAAGCC
GTTCACTACTTACCCATATCTACACACTTTTAGAAG
TTCACTACTTACCCATATCTACACACTTTTAGAAGC
TGTTCACTACTTACCCATATCTACACACTTTTAGAA
GTTCACTACTTACCCATATCTACACACTTTTAGAAG
GTGTTCACTACTTACCCATATCTACACACTTTTAGA

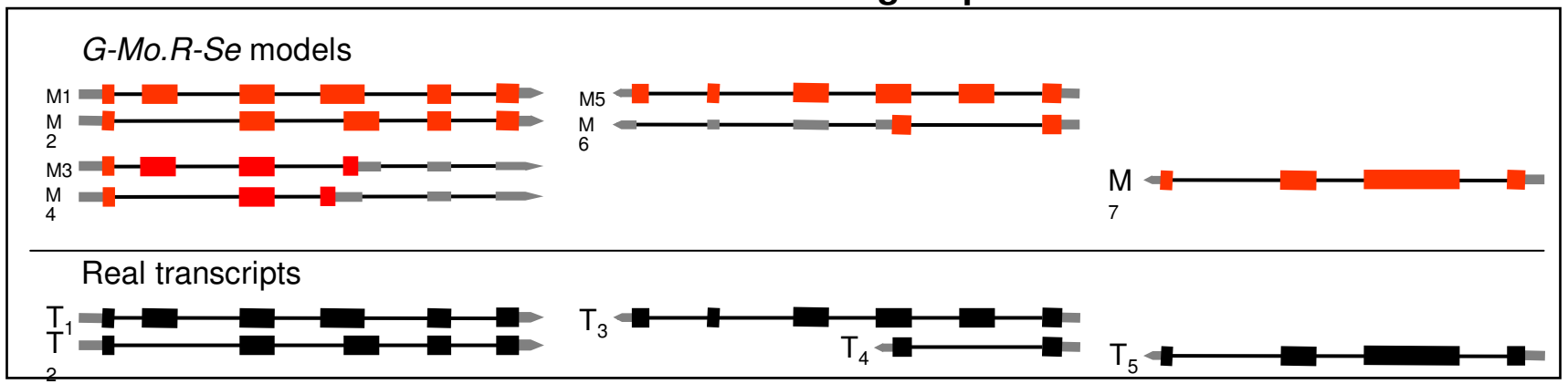
TGTTCACTACTTACCCATATCTACA
GTTCACTACTTACCCATATCTACAC
TTCACTACTTACCCATATCTACACA
TCACTACTTACCCATATCTACACAC
CACTACTTACCCATATCTACACACT
.....

Annotating genomes using RNA-Seq

4. graph of candidate exons linked by validated junctions



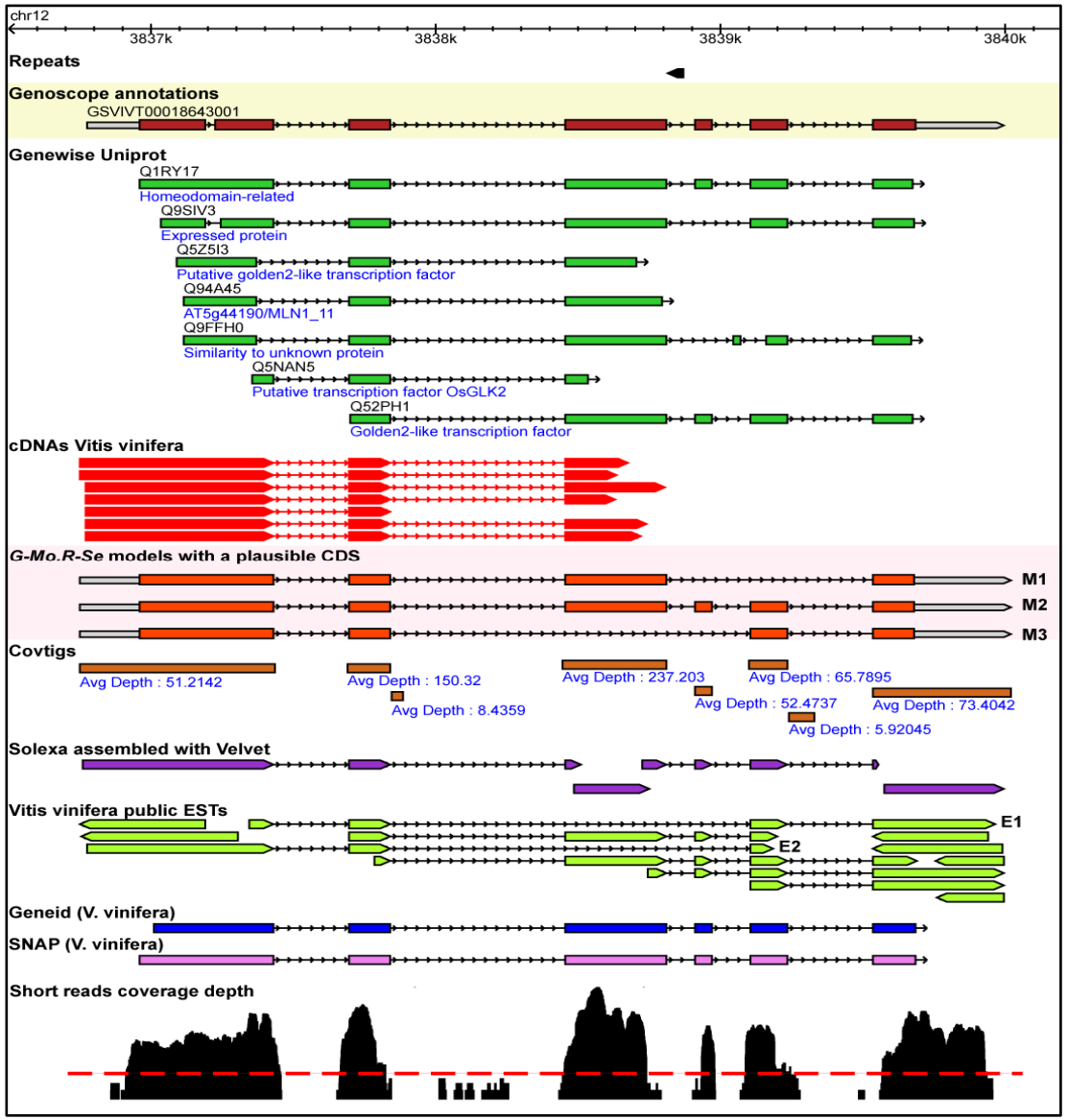
5. model construction and coding sequence detection



Annotating genomes using RNA-Seq

- ✓ Method set-up to annotate the vitis genome (500Mb)
  - ✓ Around 175 million of illumina reads
  - ✓ 4 tissues : leaf, root, stem and callus
  - ✓ 140 million of uniquely aligned reads (73,5Mb)
  - ✓ around 380 000 covtigs (38,5Mb)
  - ✓46 062 transcript models (19 486 loci), and 28 399 with a plausible CDS (12 341 loci)
  - ✓Around one week of computation with a desktop computer

Annotating genomes using RNA-Seq



Annotating genomes using RNA-Seq

Characteristics of known and novel G-Mo.R-Se models (all, and with a plausible CDS)

	All models	Models with a plausible CDS (65%)	cDNAs
Number of loci	18,811	12,236	7,895
Number of models	45,290	28,283	9,827*
Number of models per locus	2.4	2.3	1.25

* ~ 90 000 ESTs assembled, 95% of assembled transcripts detected by Gmorse

Alternative splicing events detected in cDNAs and G-Mo.R-Se models

	cDNAs 7,895 loci		Models (all) 19,486 loci		Models (CDS) 12,341 loci		Events common to cDNAs and models (% of cDNA events)
alternative acceptor/donor	690	73.1%	7405	62.5%	2988	58.0%	156 (22.6%)
skipped	250	26.5%	3656	30.9%	1677	32.5%	18 (7.2%)
mutually exclusive	4	0.4%	781	6.6%	487	9.5%	1 (25.0%)
intron retention (IR)	1227	-	-	-	-	-	-
Total	2171 (944 without IR)		11,84		5152		175 (18.5%)
Total number of loci with alternative splicing (% of all identified loci)	783 (9.9%) (598 without IR)		1602 (8.2%)		1029 (8.3%)		-

⇨ G-Mo.R-Se is not optimised to detect splicing events, but it detects more alternative transcripts than classical cDNA sequencing.

Annotating genomes using RNA-Seq

✓G-Mo.R-Se (Gene MOdeling using Rna-Seq), is downloadable from Genoscope website : http://www.genoscope.cns.fr/gmorse

✓ Used with illumina data, but it can be easily adapt to manage Solid data (in colorspace)

✓ Method used to annotate the whole vitis genome

-.-.-.-.-.- G-Mo.R-Se -.-.-.-.-.-
Gene MOdeling using RNA-Seq

Introduction || Download | Example | Contact

Introduction

G-Mo.R-Se is a method aimed at using RNA-Seq short reads to build *de novo* gene models.
First, candidate exons are built directly from the positions of the reads mapped on the genome (without any *ab initio* assembly of the reads), and all the possible splice junctions between those exons are tested against unmapped reads : the testing of junctions is directed by the information available in the RNA-Seq dataset rather than *a priori* knowledge about the genome. Exons can thus be chained into stranded gene models.

Top ↑

Download

At the moment, G-Mo.R-Se is still in development, but the current unstable version can be obtained from here.

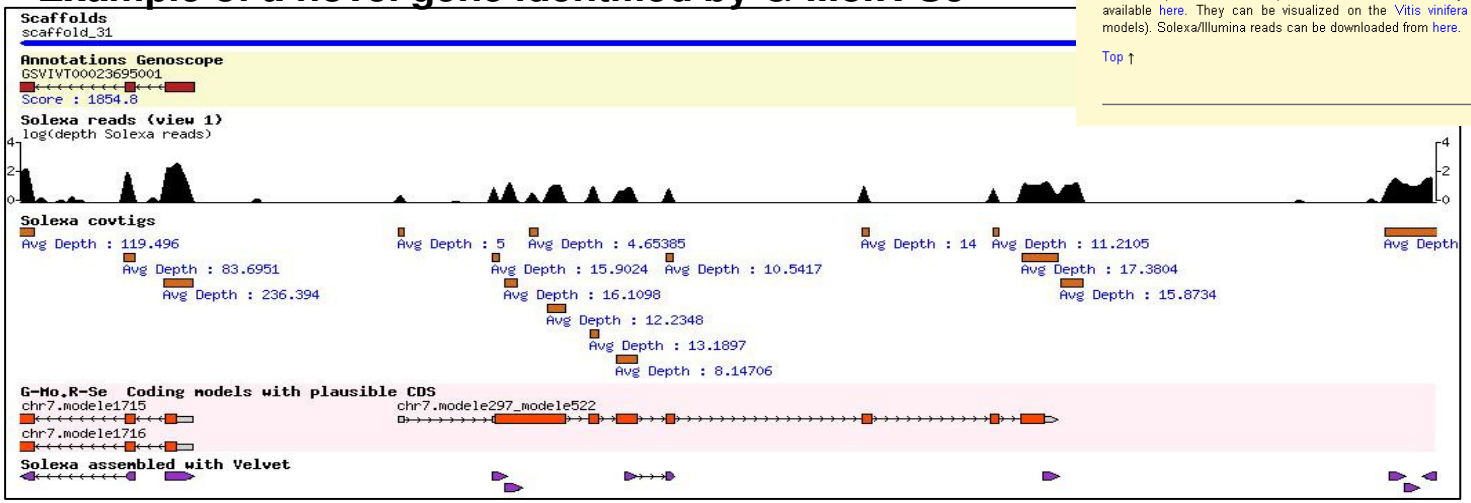
Top ↑

Grapevine genome example

We demonstrate the feasibility of this method on the grapevine genome using ~175 million Solexa/Illumina RNA-Seq reads from four tissues. This allowed the identification of new exons (in known loci) and alternative splice forms, as well as entirely new loci. The G-Mo.R-Se models are available here. They can be visualized on the Vitis vinifera genome browser (tracks G-Mo.R-Se models). Solexa/Illumina reads can be downloaded from here.

Top ↑

Example of a novel gene identified by G-Mo.R-Se



Current and future projects

- Diversification of applications and projects : de novo sequencing, genome annotation, re-sequencing, metagenomic, functional genomic, identification of mutations and structure variations…

- Increase of number and size of projects

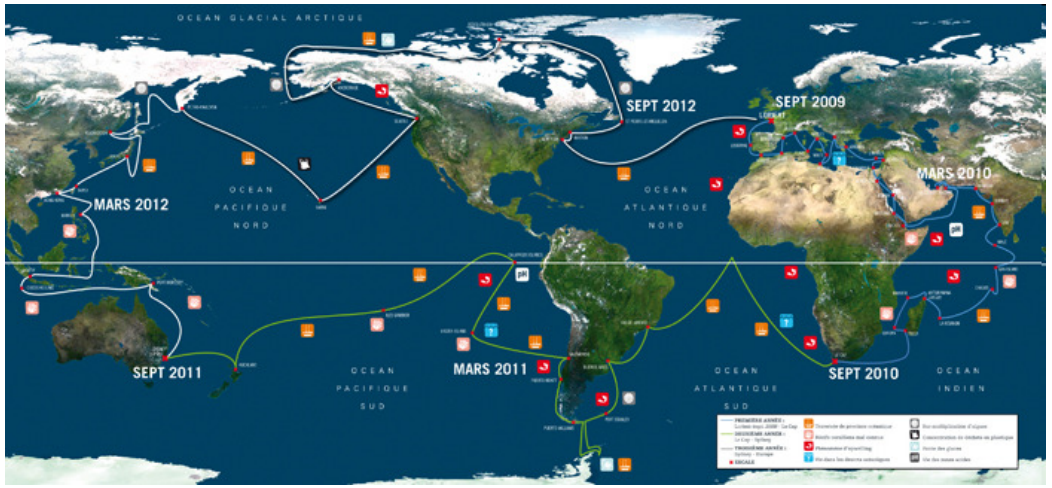
Current and future projects

- De novo sequencing :
  - Assembly of prokaryotic genomes with Illumina sequencing only
  - Sequencing of large eukaryotic genomes with 454 and Sanger : banana (~500Mb ; WGS with 20X 454 + 4X Sanger), cocoa (~400Mb ; WGS with 20X 454 + Sanger BAC ends), trout (~2Gb ; WGS with 454 and Sanger BAC ends), wheat chromosome 3b (~1Gb ; 454)
  - Benchmarking assembly of illumina data for large eukaryotic genomes : banana, cocoa, …

- Re-sequencing :
  - 100 Arabidopsis genomes : transposons mobility and methylation

Current and future projects

- Tara Oceans Project:
  - Eukaryotic meta-genomic and meta-transcriptomic
  - 3 years expedition with regular sampling at different depth and different cell size fractions.
  - Pilot project of 6 months in progress
  - 1 samplig station, 3 different depths, 3 cell size fractions, 2 or 3 sequencing technologies (454, illumina and Solid)
  - Sequencing of DNA, total RNA and messenger RNA
  - Etablish a collection of reference genome sequences and gene catalogue : 454



Conclusion

- Price-cutting => burst of novel applications

- Still at the beginning, NGS change every month :
  - Illumina announce 200Gbases per run (8days) with the HiSeq2000
  - 454/Roche will provide Kb reads in 2010
  - Next-next generation is coming : single molecule sequencing, longer reads, runtime decrease, ...

- Necessity to provide adequate IT infrastructure : production, storage and analysis.

Thank you

Laboratoire d'Analyse Bioinformatique des Séquences

François Artiguenave
Jean-Marc Aury
Christophe Battail
France Denoeud
Olivier Jaillon
Vincent Meyer
Arthur Moisdon
Benjamin Noel
Gaelle Samson
Corinne da Silva
Marc Wessner

Laboratoire de séquençage

Patrick Wincker
Corinne Cruaud
Julie Poulain
Adriana Alberti
Karine Labadie

Laboratoire finishing

Valérie Barbe
Sophie Mangenot

Laboratoire d'informatique

Claude Scarpelli
Veronique Anthouard
Arnaud Couloux
Frederic Gavory
Eric Pelletier

Paris, 24 mars 2010