

Our Experience with Genome Analyzer Applications:

From Sample Preparation to Data Analysis

Laurent FARINELLI

Illumina Seminars

Berlin and Paris

19 & 21 February 2008

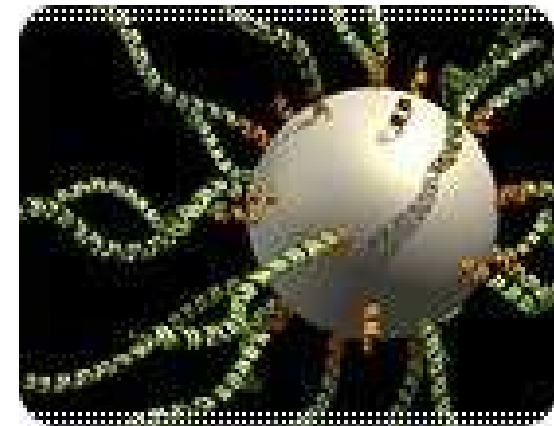


1996: Alternative Sequencing Project

- Whole bacterial genome project at Glaxo's *Geneva Biomedical Research Institute*
=> Need for higher throughput sequencing
- Pharmacogenomics
"The Right Drug to the Right Patient"
- Sydney Brenner's Massively Parallel approach

Bead Arrays

- ✈ **Lynx Therapeutics'**
MegaClone
- ✈ **Massively Parallel
Signature Sequencing
(MPSS)**



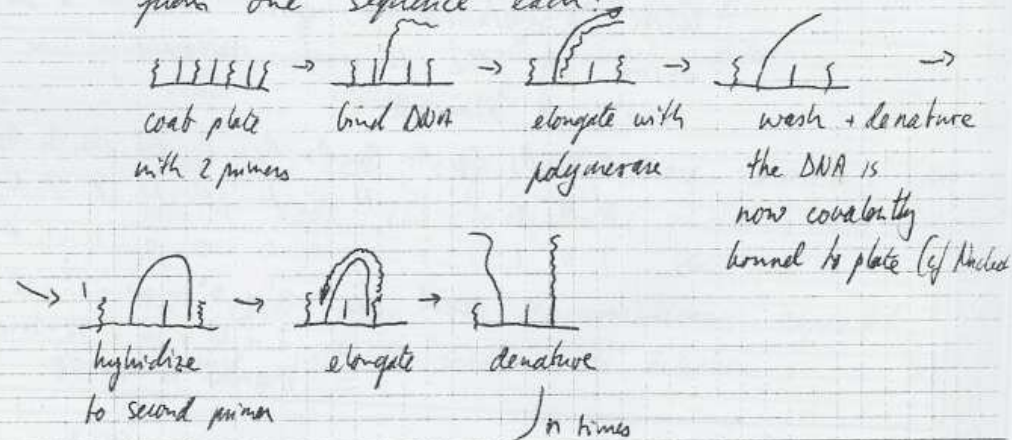
LYNX

- **Gene Expression Profiling**

Brenner, S. et al., Nature Biotechnology 18:630-634 (2000).

① PCR colonies (Pascal's idea)

- Coat a NucleoLink-like surface with 2 primers
- Apply diluted template DNA so that each molecule is $\approx 5 \mu m$ apart.
- Do a PCR reaction without primers in solution. The result should be spots of DNA amplified from one sequence each:



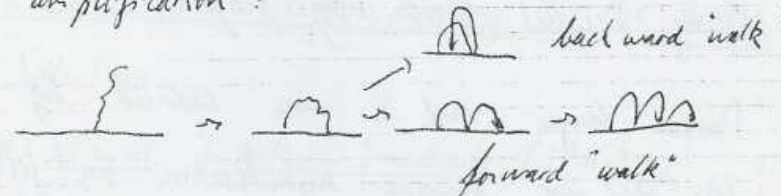
Signature: *Amelk*

Read and understood: *SV*

Date: 13.11.96

Date: 15.11.96

"Thus each spot of DNA will have been initiated with one molecule only, which "walked" during PCR amplification:



The resulting "PCR colonies" would be the equivalent to the coated beads.

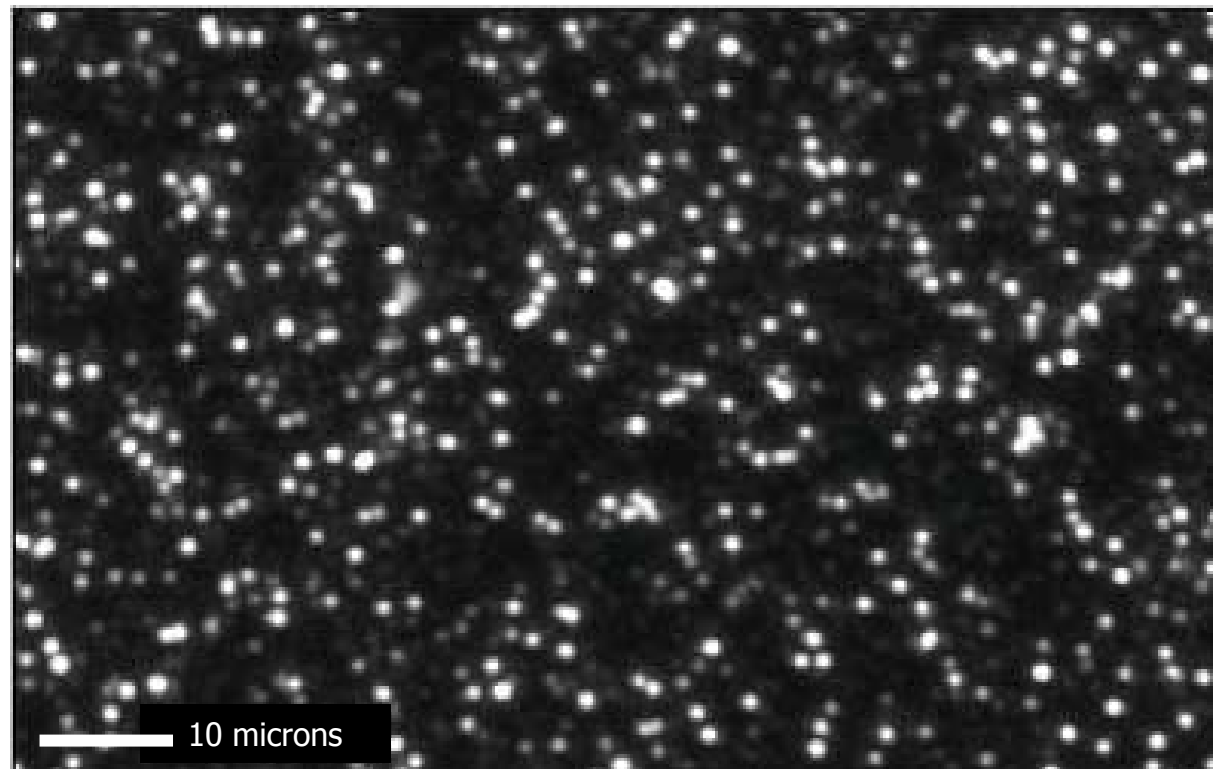


1996-1997: GlaxoWellcome's
Geneva Biomedical Research Institute

Mayer P., Farinelli L. and Kawashima, E., 1997, Patent application WO 98/44151

DNA Colonies: Random Arraying

"DNA Colonies"



Up to 10'000'000 DNA colonies / cm²

1996-1997: GlaxoWellcome's
Geneva Biomedical Research Institute

1998-2000: Serono

2000-2003: Manteia Predictive Medicine

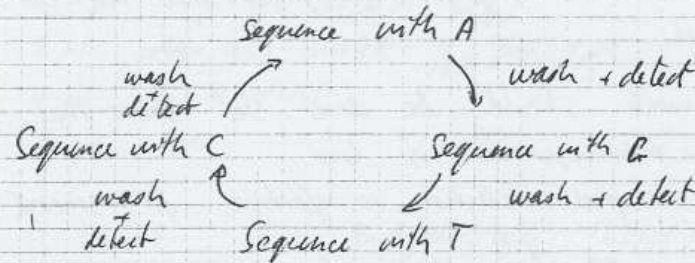
2004- : Solexa illumina



② Step-by-step sequencing (Sanger's idea)

Using a single stranded molecule as template, with a specific oligo as a primer, like for a normal sequencing reaction:

- do a polymerization with only 1 type of nucleotide present in solution, wash
- Detect presence of incorporated nucleotide(s) (by fluorescence for example)
- Repeat polymerization with next nucleotide type
- Repeat washing and detection.
- Continue the cycle with the 4 bases:



Thus on each step, only those colonies / beads which have the incorporated nucleotide will light on.

For cases when the same base is repeated, sequence will be read with detection

Signature:

Farinelli

Read and understood:

SW

Date:

13.11.96

Date:

15.11.96



1996-1997: GlaxoWellcome's
Geneva Biomedical Research Institute

Farinelli L., Mayer P. and Kawashima E., 1997, Patent application WO 98/44152

in situ Sequencing

Primer : 5'-gactagcgtcat-3'

Template : 3'-ggatgctgatcgcactattgatgggcacgaactca-5'

Cycle of stepwise base extension

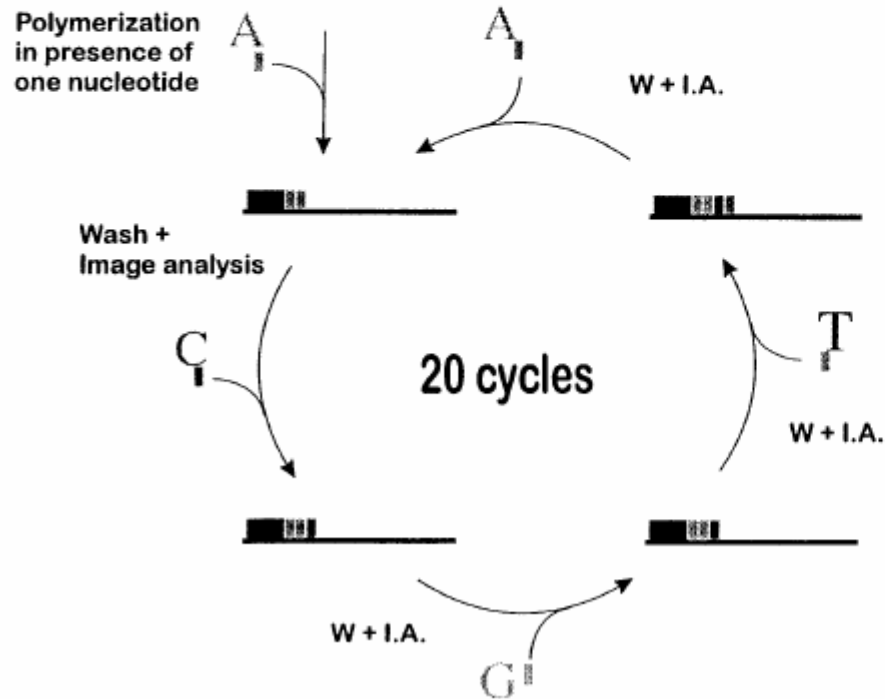
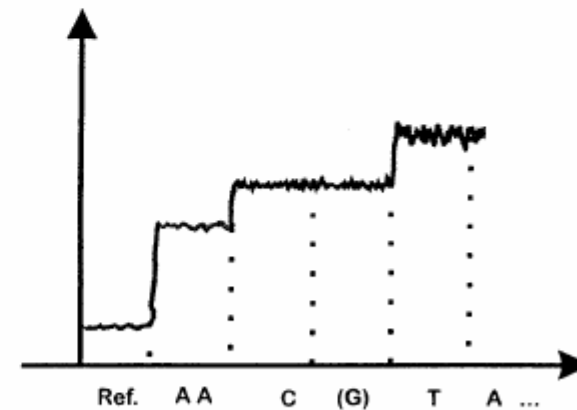


FIG. 2.

Fluorescence increase on individual spots



1996-1997: GlaxoWellcome's
Geneva Biomedical Research Institute

**Farinelli L., Mayer P. and
Kawashima E.,** 1997, Patent
application WO 98/44152

DNA Colonies Project

- ✈ 1996-1997 *Geneva Biomedical Research Institute, Switzerland*
- ✈ 1998-2001 *Serono Pharmaceutical Research Institute*
- ✈ 2001-2003 *GenInEx / Manteia Predictive Medicine*
- ✈ 2004 Technology sold to *Solexa* UK and *Lynx* USA
- ✈ 2004 Merge of *Solexa* and *Lynx*
- ✈ 2007 Acquisition by *illumina*, USA

Fasteris SA

- ✈ Founded 2003
 - Laurent FARINELLI and Magne OSTERAS
- ✈ No external funding
- ✈ Sustainable business

Core Business ***High-quality Capillary DNA sequencing***

- ✈ Importance of **SERVICE**
- ✈ Next day results
- ✈ Competitive cost
- ✈ Personalized protocols
- ✈ Analyses are not charged when no sequence can be delivered



➡ ***Enabled leasing a
NEXT generation instrument***

Additional services

- Large scale plasmid preparation



- Custom projects in molecular biology and microbiology

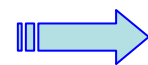


- Low density Micro-arrays



2006: Tough Decision

- ✈ NEXT-generation systems came on the market earlier than we had anticipated
- ✈ Too big for a small company?
- ✈ Can we afford to wait?
- ✈ Which system to choose?
- ✈ Will clients adopt the new technology?
- ✈ Can they pay for such services?



We ordered the Genome Analyzer in December 2006



One Year Ago

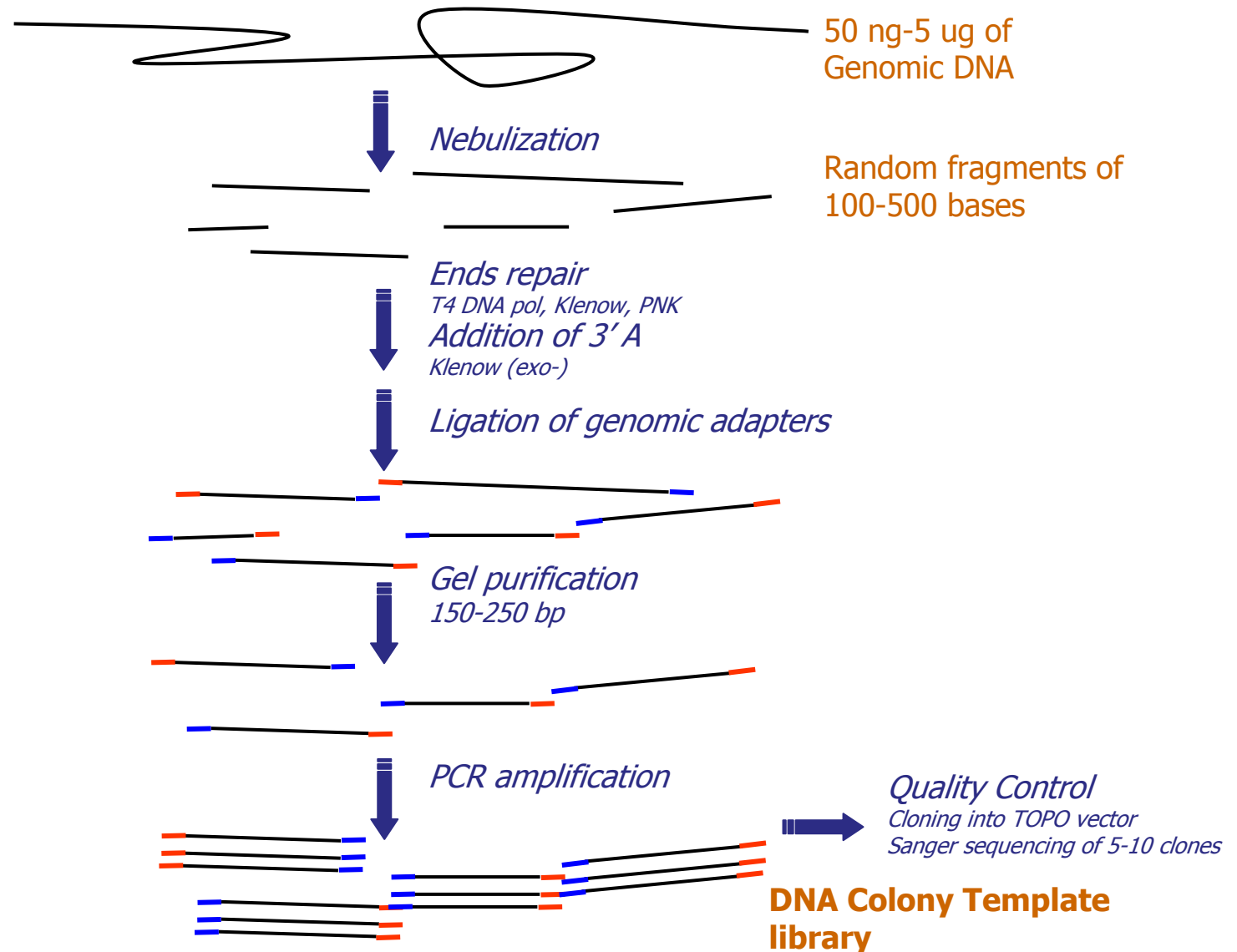
- ✈ Project to close gaps of a bacterial genome
- ✈ Approx. 100 gaps:
 - 200 PCR primers US\$ ~2000\$
 - 100 PCR reactions US\$ ~500\$
 - 100 sequencing reactions US\$ ~1000\$

=> over \$4000 just for consumables
- ✈ Issue: many difficult PCR reactions
- ✈ Optimization => high costs \$\$\$

One Year Ago *Sample prep*

- ✈ We had just received Solexa genomic sample prep kit and protocol
- ✈ First attempt with bacterial gDNA was successful

Genomic Sample Preparation



One Year Ago Installation Run

- ✈ 2 channels with BAC control library
- ✈ 6 channels with bacterial library
 - 20.6 mio reads of 26 bases
 - > 500 mio bases
 - > 185x coverage



Spring 2007: Analysis needs

- ✈ Re-sequencing
 - Mapping reads of reference genome
 - Repeats
 - Generate consensus
 - Identify SNPs (+ filtration)
 - Coverage information
 - Graphical views
- ✈ *De novo* assembly
 - Contigs from unsequenced organisms
 - Combinations between millions of reads
 - Very time-consuming
 - Short reads

Spring 2007: Analysis needs

- ✈ First re-sequencing softwares available (ELAND, AMOS)
- ✈ Need to develop *de novo* assembly
 - EDENA, written by David Hernandez as part of a collaboration with the group of Prof. Schrenzel, Geneva University Hospital
 - Reads of identical lengths
 - Exact matches

Bioinformatics analysis tools

✈ Re-sequencing

- ELAND + BeadStudio
- M.A.Q.

⇒ *Only reads with 0, 1 or 2 SNPs*

⇒ *Does not map reads with indels
(without paired-ends)*

✈ de novo assembly

- EDENA
- Velvet

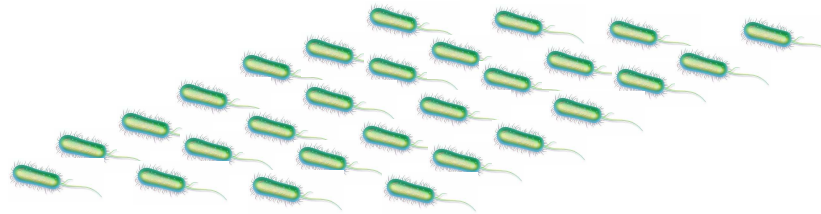
⇒ **Contigs extension blocked**

- Low coverage
- Repeated regions

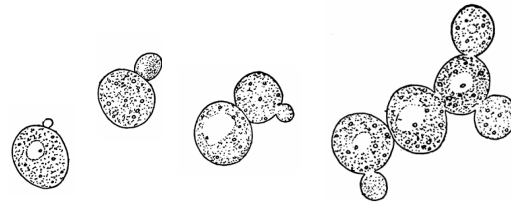
One Channel

(one Genome Analyzer Flow-cell/run is 8 channels)

- 2-5 mio reads of 30-35 bases
- ~140 mio bases
- 140 books of 1000 pages



- 28x coverage of a 5 Mb bacterium



- 10x coverage of a 12 Mb Yeast

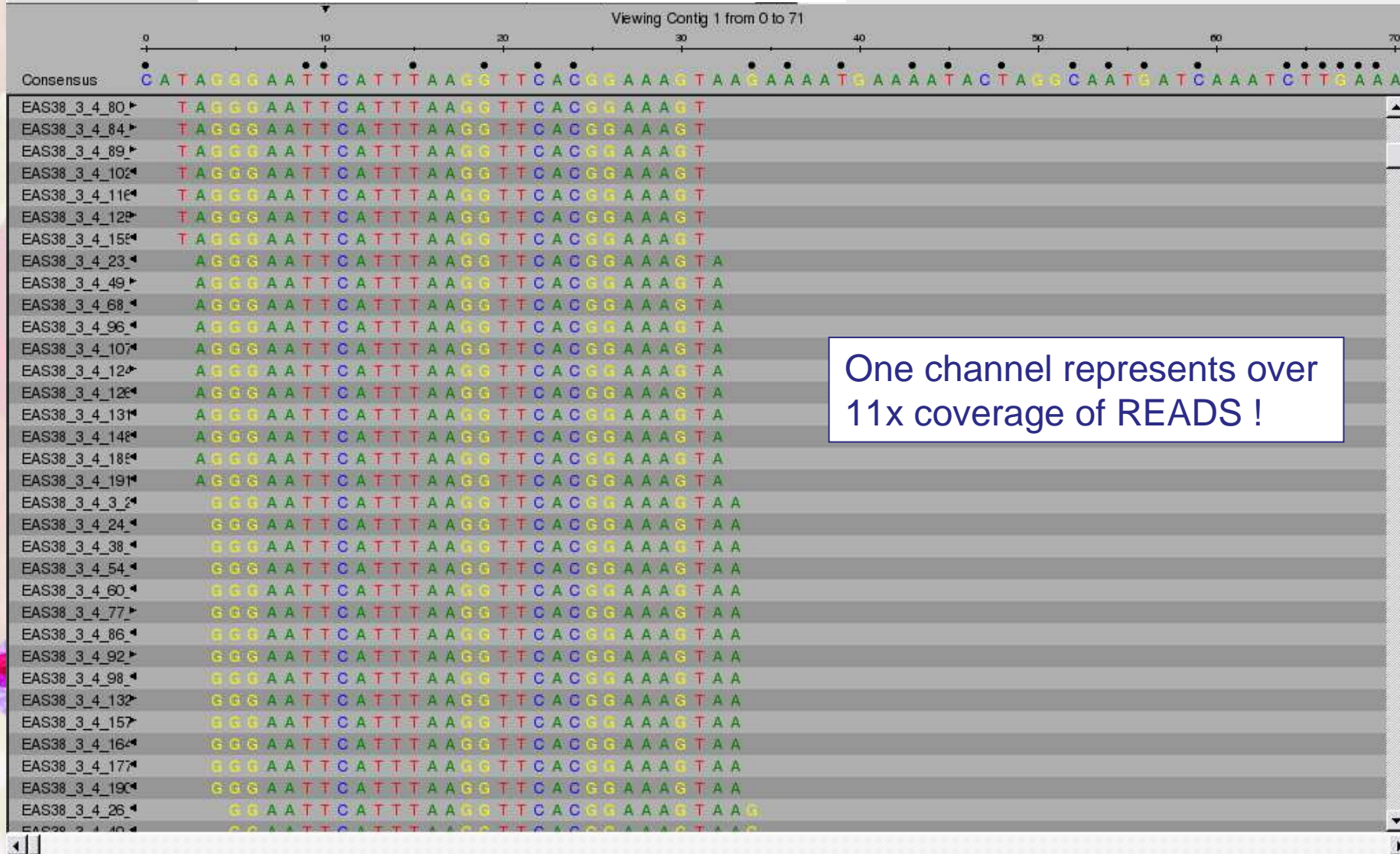
- 1x coverage of 128 Mb Arabidopsis



BAC Control

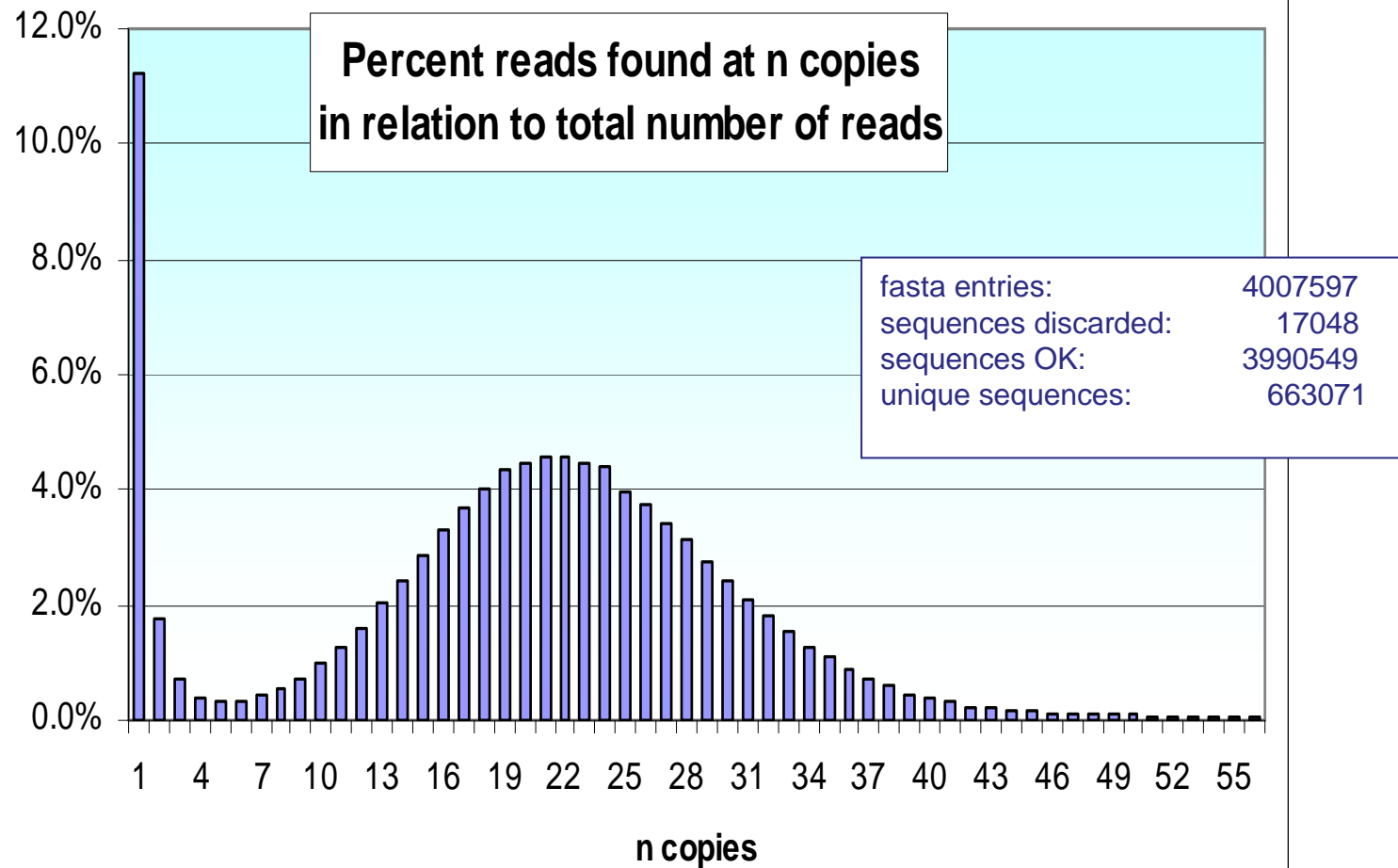
- BAC of 175 kb
- Can generate 175'000 unique fragments of 35 bases
- If one channel produces 2 mio reads, each unique fragment should be present at 11 copies

BAC Resequencing



One channel represents over 11x coverage of READS !

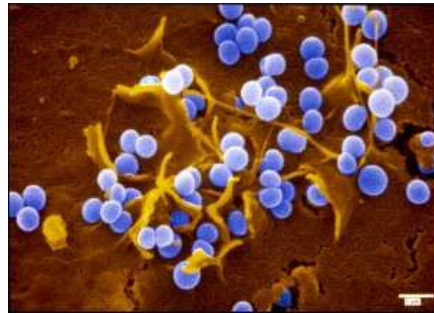
BAC Resequencing



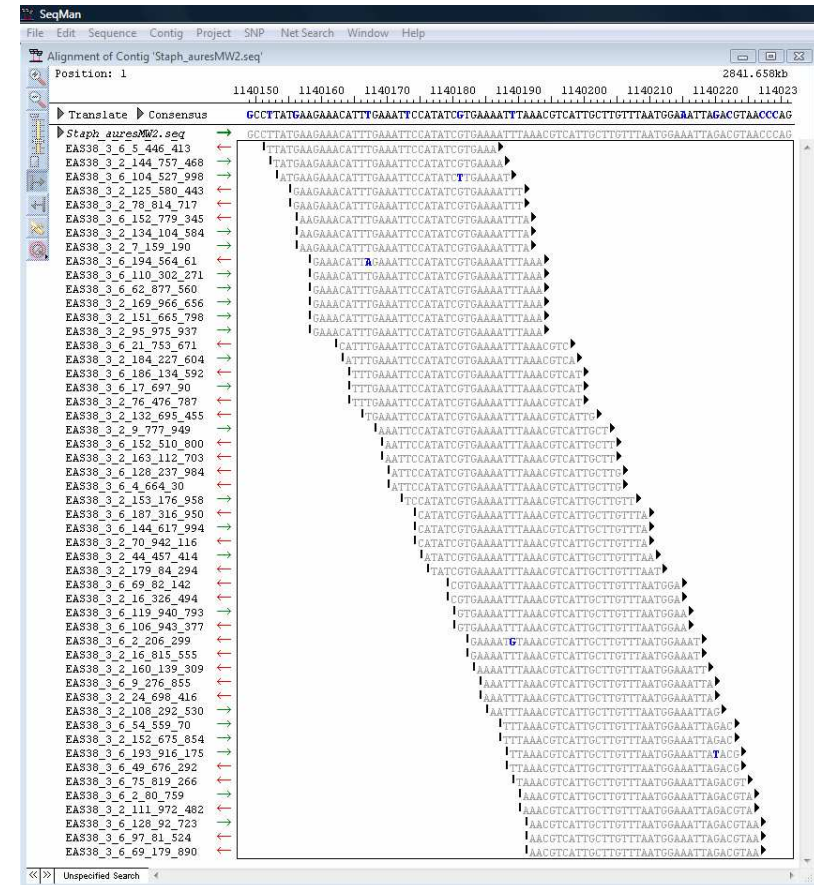
BAC Resequencing

The screenshot shows a software window titled "Applications" with a menu bar containing "Options". Below the menu bar, there are controls for "Contig 1", "Offset 0", and a search field. The main display area shows "Viewing Contig 1 from 0 to 71" with a scale from 0 to 70. A consensus sequence is shown at the top, followed by a list of reads with their corresponding IDs (e.g., 268381-13, 268382-1, etc.). The reads are color-coded to show mismatches from the consensus. A callout box on the right contains the text: "Only the unique reads => exaggerates errors".

Accuracy of Genome Analyzer data *Staphylococcus aureus* MW2



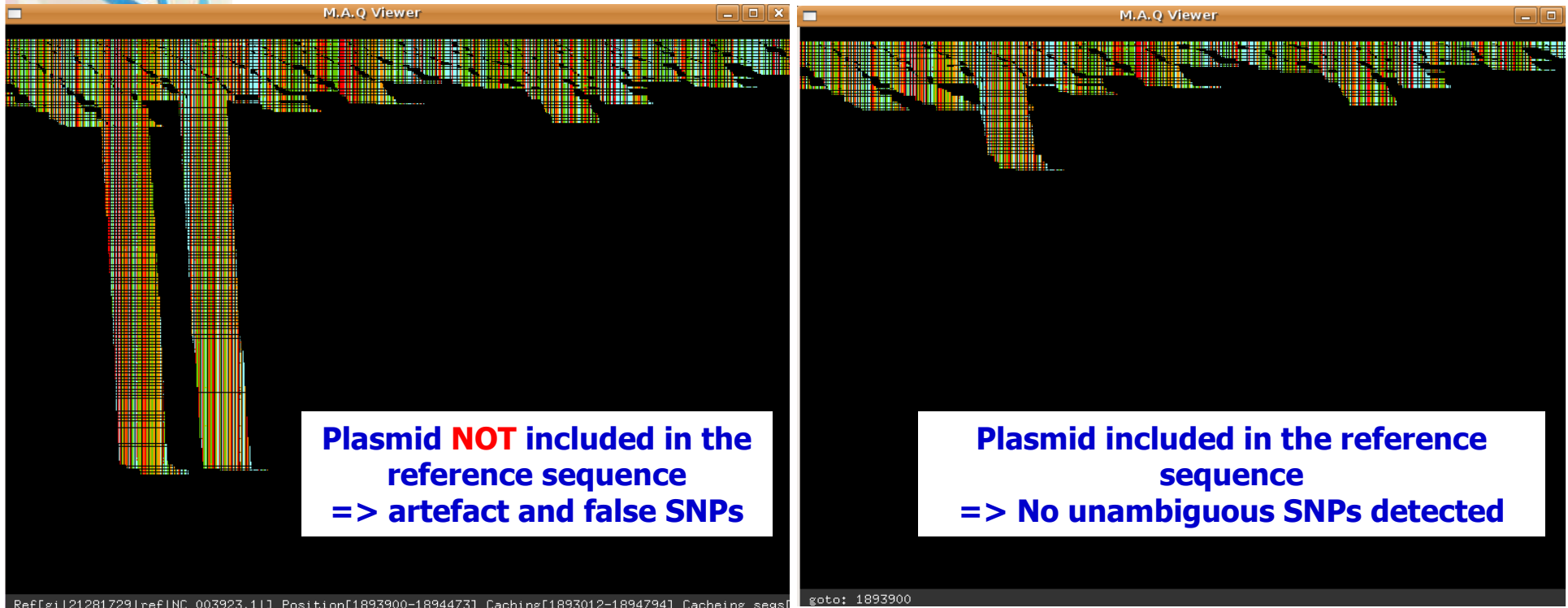
MAQ	Experiment 1	Experiment 2
Genomic DNA prep	Jan 2007	Jan 2005
Ref. / size	NC_003923 / 2.82 Mb AP004832 pMW2 / 20.6 kb	
Reads	3'857'879 (2 channels)	2'684'877 (1 channel)
Reads mapped (%)	3'606'310 (93.5%)	2'530'239 (94.2%)
Coverage gDNA	44.8	30.7
Coverage pMW2	245	92.6
Bases not covered	1	0
SNPs	2	2



Mapping using the DNASTAR SeqMan
Genome Assembler software

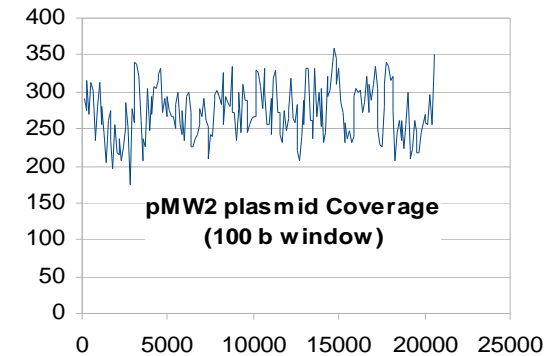
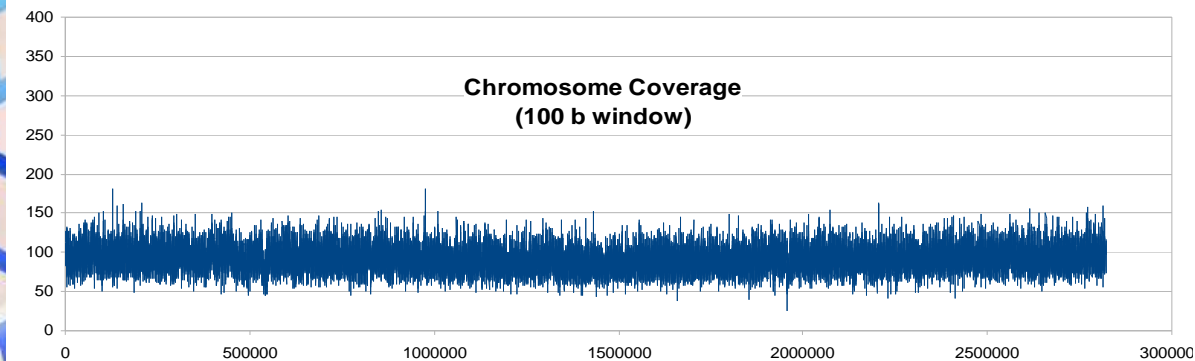
Collaboration with the group of Prof. Jacques Schrenzel, Geneva University Hospitals
www.genomic.ch

Presence of plasmid pMW2 in Staphylococcus aureus MW2



Collaboration with the group of Prof. Jacques Schrenzel, Geneva University Hospitals
www.genomic.ch

Re-sequencing of Staphylococcus aureus MW2



- Complete coverage
- Uniformity:
no bias due to sample preparation,
amplification or Genome Analyzer sequencing
- Higher coverage for pMW2 plasmid

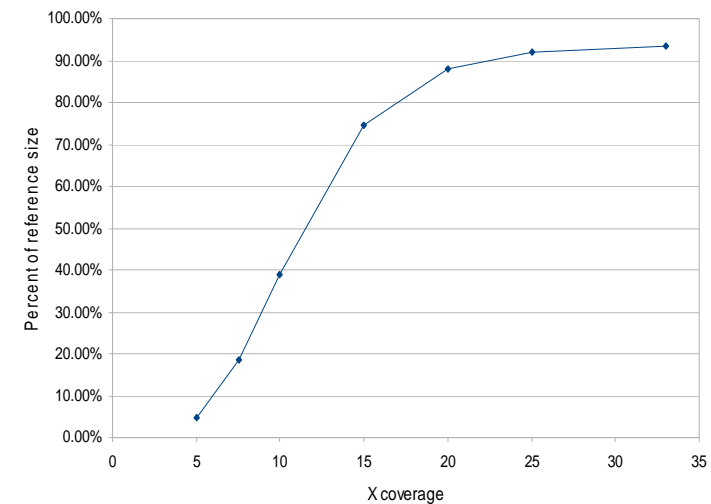
Collaboration with the group of Prof. Jacques Schrenzel, Geneva University Hospitals
www.genomic.ch

SNP detection

- ✧ ELAND and MAQ:
 - Scripts to filter MAQ SNPs
 - Do not map reads with insertions/deletions
(without paired-ends)
- ✧ Using *de novo* assembler EDENA or Velvet
 - Generate contigs
 - Compare contigs with reference sequence
 - Find insertions or deletions

EDENA *de novo* assembly *Staphylococcus aureus* MW2

EDENA 2b	Experiment 1	Experiment 2	Experiment 1+2
Genomic DNA isolation	Jan 2007	Jan 2005	
Ref. / size	NC_003923 / 2.82 Mb AP004832 pMW2 / 20.6 kb		
Reads	3'857'879 (2 channels)	2'684'877 (1 channel)	6'5422'756 (3 channel)
Contigs	1124	2273	570
Total length	2'763'965	2'724'486	2'782'588
Ctgs mapped	1124	2260	561
Coverage	98%	96%	98%

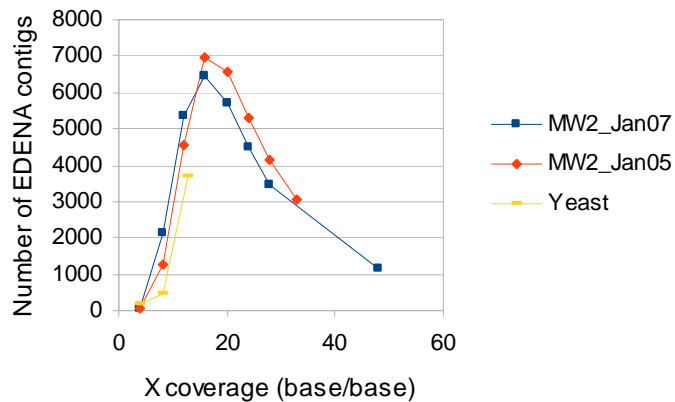


Coverage of the reference sequence by EDENA contigs starting from an increasing number of reads
(plotted as coverage, using data from Experiment 2 with minimum contigs length of 100)

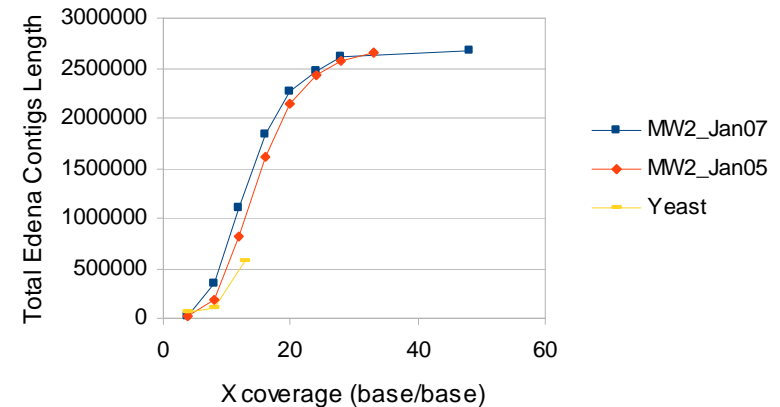
Collaboration with the group of Prof. Jacques Schrenzel, Geneva University Hospitals
www.genomic.ch - EDENA: Hernandez *et al*, Genome Res. In Press 2008

EDENA *de novo* assembly

Contigs Versus coverage



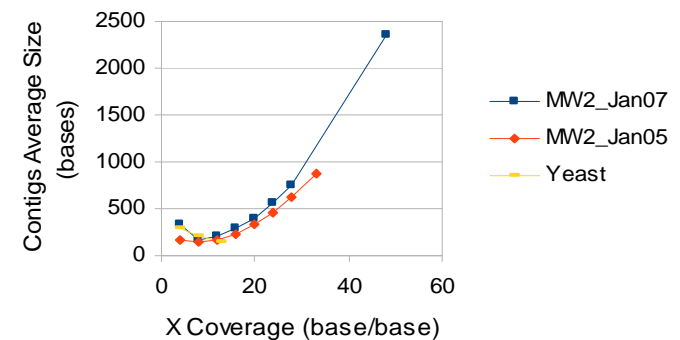
Total Length Versus Coverage



Above ~20x coverage:

- ✈ Number of contigs goes down
- ✈ 80% of genome covered

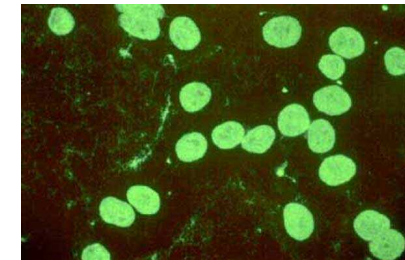
Contigs Average Size Versus Coverage



Collaboration with the group of Prof. Jacques Schrenzel, Geneva University Hospitals
www.genomic.ch
 EDENA: Hernandez *et al*, Genome Res. In Press 2008

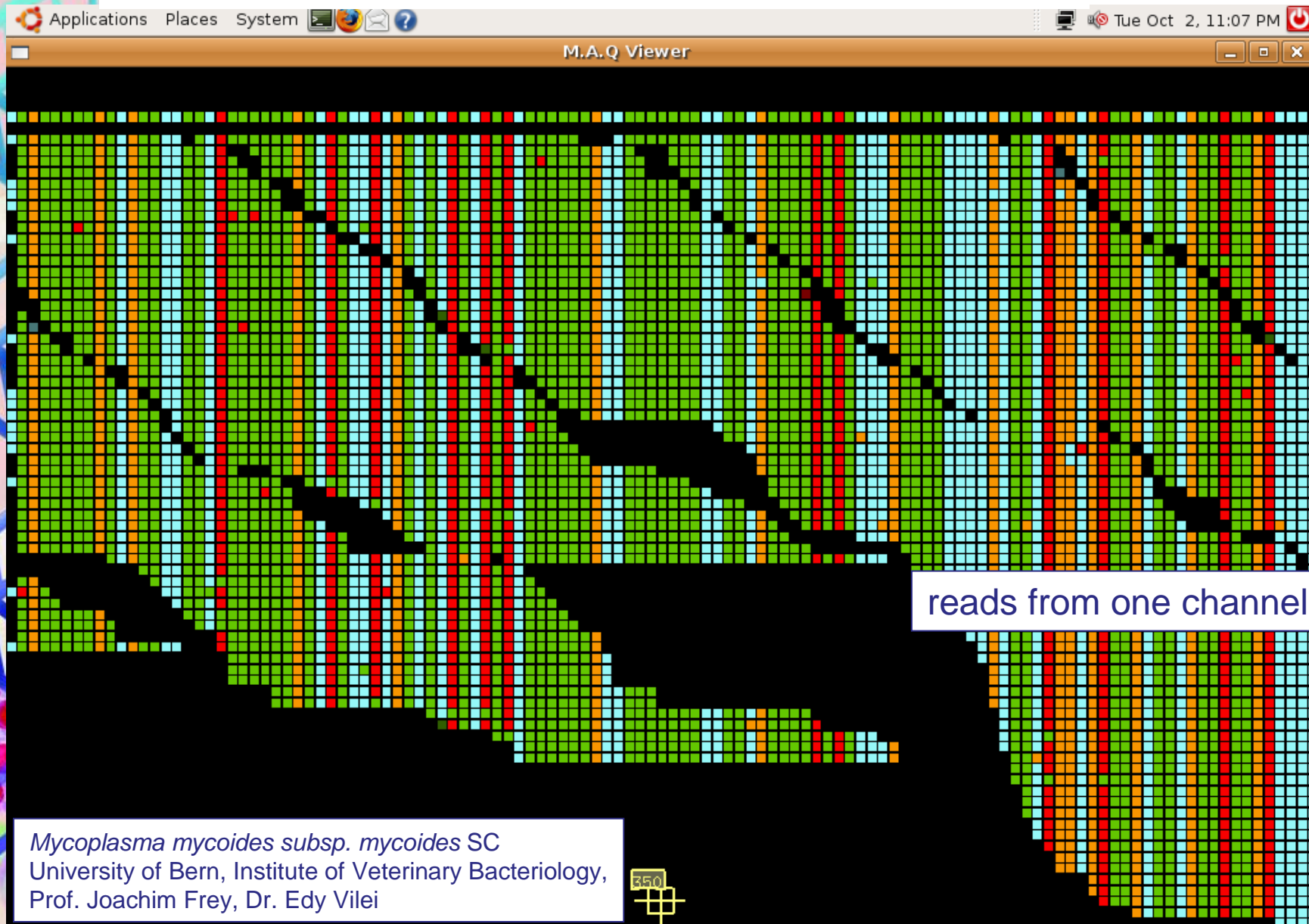
August '07: Comparison of Difficult-to-Sequence Bacteria

- In collaboration with the team of Prof. Joachim Frey of the University of Berne, we re-sequenced the genome of a pathogenic strain and a vaccine strain of the bacterium *Mycoplasma mycoides subs. mycoides SC*. This bacterial species is the agent of contagious bovine pleuropneumonia, a severe, highly contagious respiratory disease of cattle and buffalo



- Genomic DNA is difficult to obtain as the bacteria are expensive to grow
- For each strain, we obtained enough reads on one channel to achieve coverage of 30-60x and align the reads on a reference sequence
- We identified several genetic variants that are being analyzed for confirmation of their role in the phenotypic differences observed between the strains

Whole Bacterium Resequencing

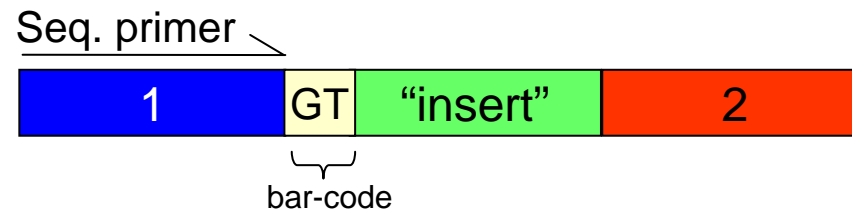


June '07: Bar-coded small RNAs

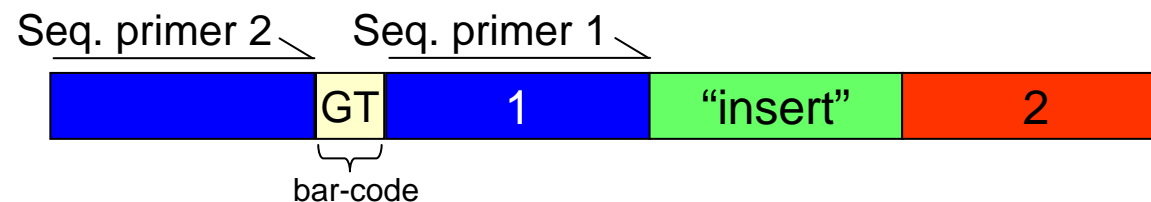
- ✦ We adapted the sample preparation protocols to permit analysis of 2 small RNA samples in the same channel
 - Added on 5' the P5 and P7 sequences by PCR amplification
=> permits amplification in the flow-cell
 - Quality control of the library
 - Designed specific sequencing primer for the GA
- ✦ Data analysis:
 - Trim adapter sequences
 - Count unique small RNAs
 - Group small RNAs with lengths of $n+1$, $n+2$, etc..
 - Compare profiles between samples
 - Map small RNAs on reference sequence
 - Identify phenotype-related small RNA

Bar-Code / Indexing

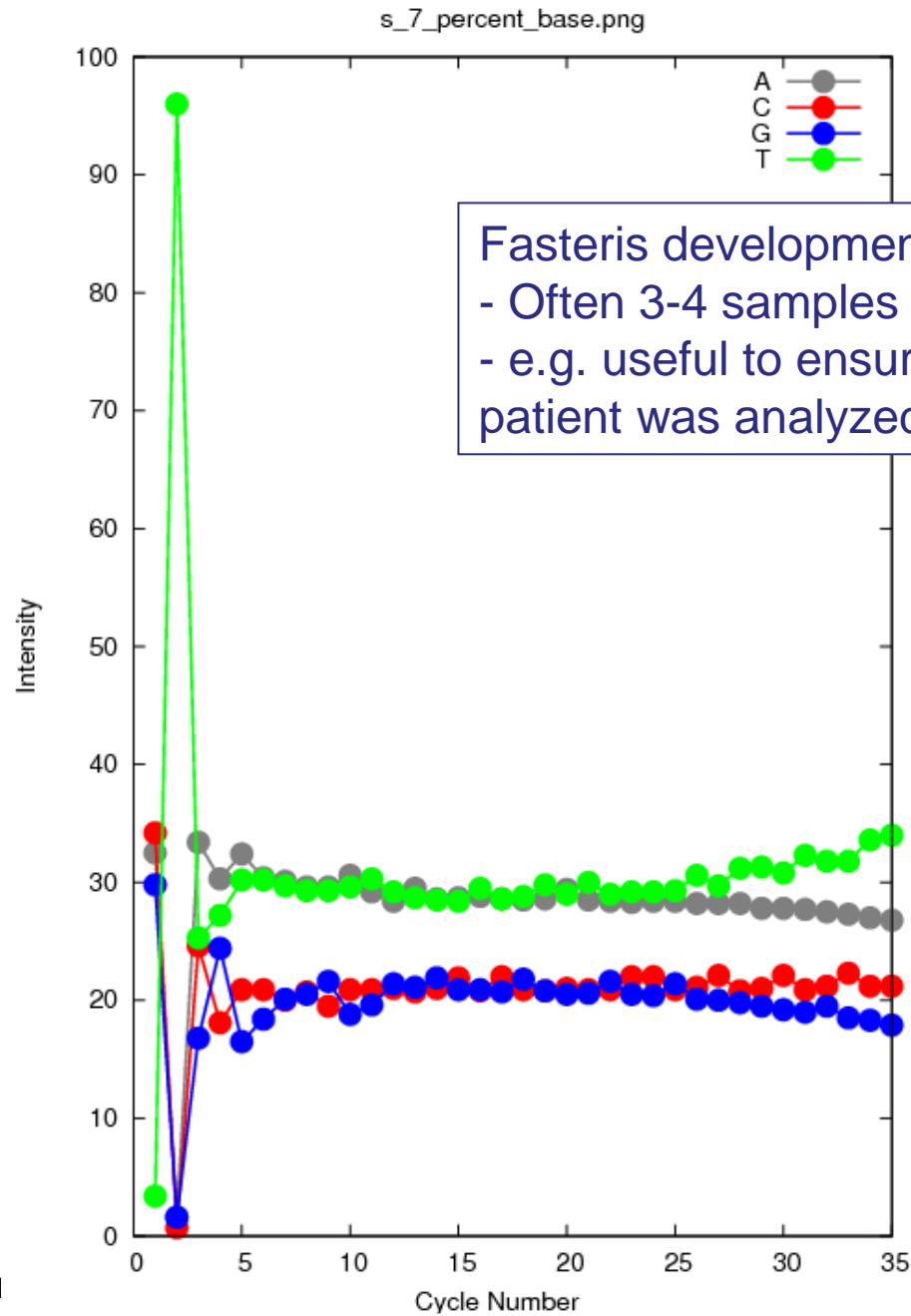
- ✈ Can analyze more than one sample per channel
- ✈ Two options:
 - 2-4 bases in the adapter 1



- Downstream a second seq. primer (paired reads)

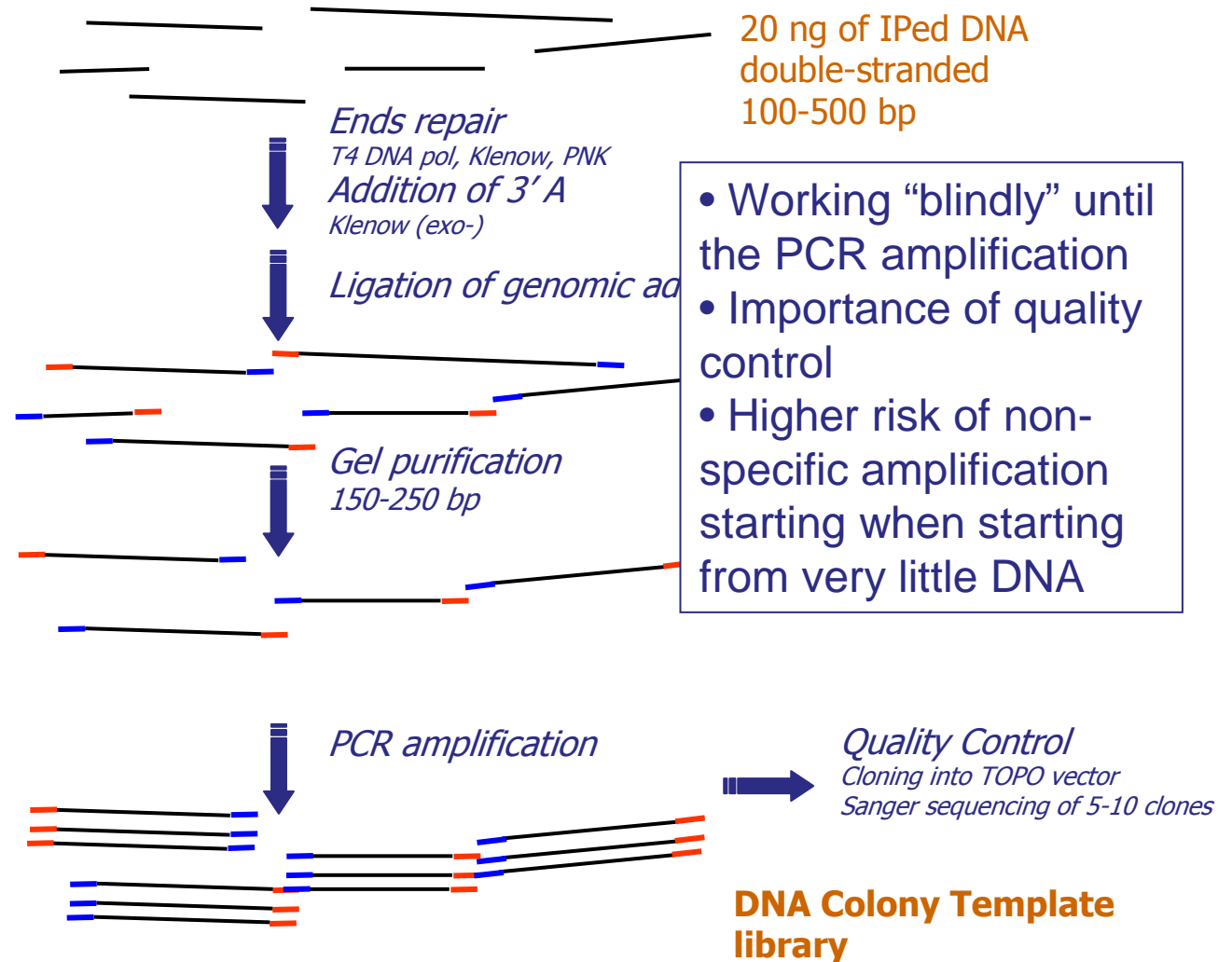


Bar-coding / indexing

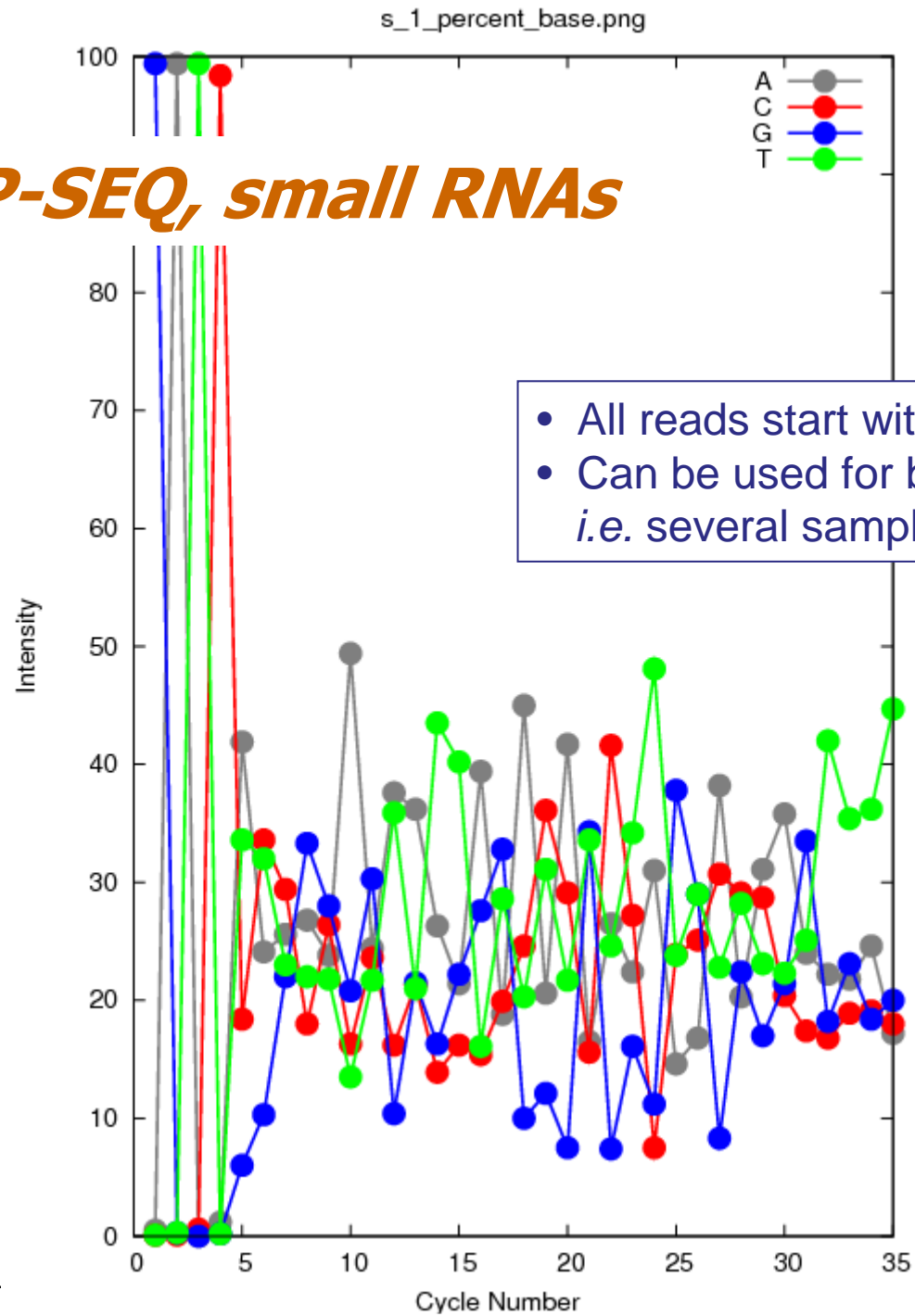


Fasteris development
- Often 3-4 samples in one channel
- e.g. useful to ensure DNA from each patient was analyzed

ChIP-SEQ Sample Preparation



ChIP-SEQ, small RNAs



ChIP-SEQ data analysis

- Mapping reads on reference sequence
- Depending on genome/chromosome size, counts in windows of 100-1000 bases
- Importance of enrichment and control sample

July '07: de novo Assembly of Whole Transcriptomes

- ✈ Dr Michel Schalk from Firmenich is interested in identifying genes involved in specific metabolic pathways
- ✈ Difficulties: no reference sequence available; genes homologous to the target genes are available only for distantly related species
- ✈ Classical approaches of library hybridization or PCR amplification are time consuming
- ✈ Sample prep and data analysis:
 - Purified cDNA (important!)
 - Shotgun approach using nebulization and genomic protocol
 - *de novo* assembly using EDENA
 - Homogeneous coverage



de novo Assembly of Whole Transcriptomes



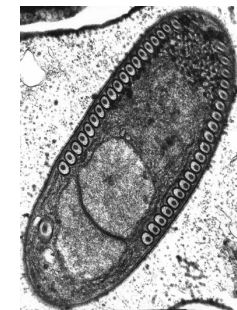
- Large contigs shown to match with plants genes known to be highly expressed, such as RUBISCO
- Candidate sequences for the target genes found among the smaller contigs
- These sequences were used to design specific PCR primers and permitted to recover the full-length transcripts by 5' and 3' RACE

August '07: *de novo* Assembly from 100 ng of genomic DNA

- ✈ *Microsporidia* are intracellular parasites
- ✈ It is very difficult to obtain DNA, therefore very little sequence knowledge is available
- ✈ Used EDENA for *de novo* assembly of contigs
- ✈ Prof. Dieter Ebert and his group from the University of Basel are using this data to mine for candidate genes for host parasite interactions and for genetic markers (variable number tandem repeats VNTRs)
- ✈ Estimation of the genome size



Daphnia



Microsporidia

Gene Expression Profiling

- ✈ Request to compare profiles from unsequenced plants
- ✈ GEX kits were not available yet
- ✈ Need to develop Fasteris protocol
 - Longer signatures to find genes
 - Possibility to compare different profiles from the same sample

Overview of new Fasteris mRNA Sample Prep Protocol

AAAAA mRNA / total RNA

↓ 1st and 2nd Strand cDNA Synthesis

AAAAA
TTTTT-bio

↓ Restriction Enzyme Digests

CGN
N AAAAA
TTTTT-bio

↓ Adaptor 1 Ligation

EcoP15II
1 CGN AAAAA
GCN TTTTT_bio

↓ EcoP15I digestion

EcoP15II
1 CGN 27 bases* NN
GCN

↓ Adaptor 2 Ligation

EcoP15II
1 CGN NN 2
GCN NN

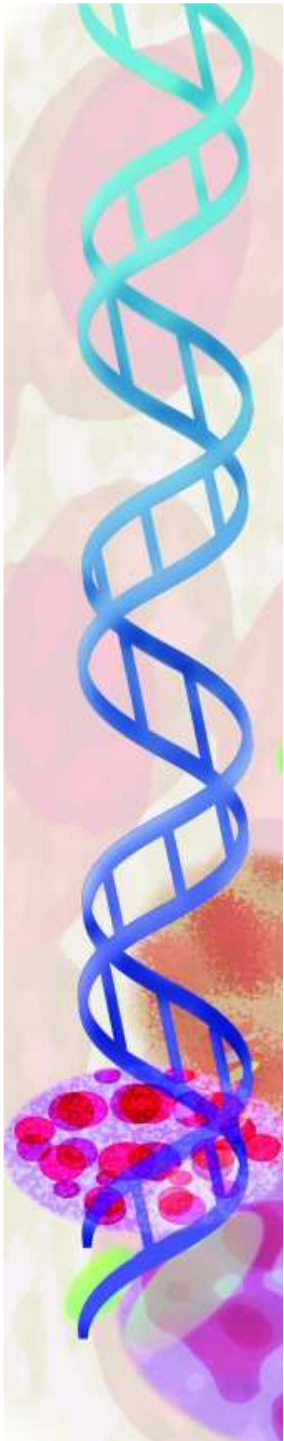
↓ PCR amplification to generate DNA Colonies Templates

← PCR Primer 2 → Signature PCR Primer 1
Sequencing Primer 27 or more bases

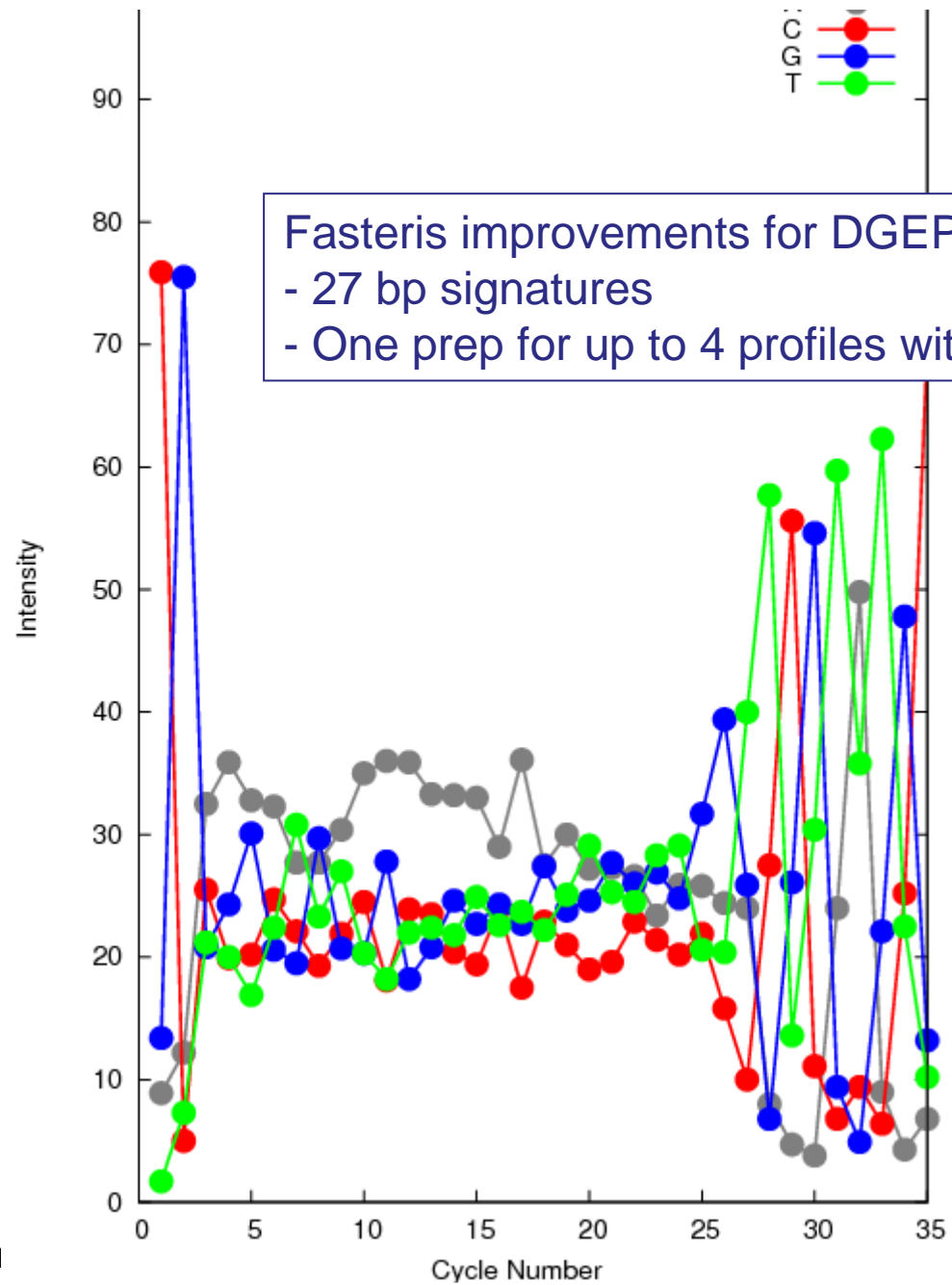
Up to 4 enzymes can be used in separate reactions: HpyCH4IV, HpaII, HinP1I or TaqI
The N is used as a bar-code CGN to identify the enzyme used

*) Due to the nature of EcoP15I, the cutting site has ~50% chances of being located at more than 27 bases. A small proportion of cuts can occur at less than 27 bases

5 June 2007



Digital Gene Expression Profiling **FASTERIS** LIFE SCIENCES



Fasteris improvements for DGEP

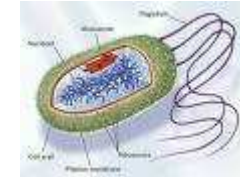
- 27 bp signatures
- One prep for up to 4 profiles with built-in bar-code

Digital Gene Expression Profiling Data analysis

- ✈ Trim Vector
(variable signature lengths even with *MmeI*)
- ✈ Count unique signatures
- ✈ Map signatures on reference sequence
- ✈ Compare profiles
- ✈ Identify genes of interest

Applications performed at FASTERIS

- ✈ Whole Genome re-sequencing
 - Bacteria, yeast, plant, animals, viruses
 - SNP detection
 - PCR products, pooled samples, e.g. 5% threshold
- ✈ *De novo* sequencing
 - Bacteria, whole transcriptome
- ✈ Digital Gene Expression Profiling
 - Signatures, whole transcriptome
- ✈ Binding sites
 - ChIP-SEQ, 4C, etc..
- ✈ Small RNAs
- ✈ Bar-coding / indexing
- ✈ Capture on micro-arrays
- ✈ Other applications



NEXT Generation Applications



Our first cars
still look like
horse-carriages

Explosion
engine has just
been invented!

We know
NEXT-Generation
is a revolution



That's where we
stand with
NEXT-Generation

We can't dream
of Formula 1

NEXT-Gen has great
potential to be discovered



With You To Guarantee Success



Clients

