KNIME & Next Generation Sequencing at the Pasteur Institute

Bernd Jagla, Ph.D.
Institut Pasteur
Génopole – Plate-forme 2 Puces à ADN

Thanks

- KNIME
  - Karol Kozak
  - KNIME community
- PF2:
  - Odile Sismeiro
  - Marie-Agnes Dillies
  - Jean-Yves Coppee
  - Caroline Proux
  - Guillaume Soubigou
- GBrowse:
  - Lincoln Stein and the rest of the GBrowse/BioPerl community
- Mobyle:
  - Hervé Ménager, Bertrand Néron

Content

- Introduction
- NGS data processing
- How KNIME fits in this

Introduction

- ***Plate-forme 2 Puces à ADN – what we do***
  - ***Tiling arrays, Microarrays, NGS***
- ***Putting it into context – Introduction to small RNAs***
- ***State-of-the-art technologies: 2nd generation sequencing***

Plate-forme 2 Puces à ADN

- Tiling arrays
  - Scan a known genome with equally spaced probes
- Micro arrays
  - Scan a known genome for regions of interest (e.g. genes/mRNA)
- Next generation sequencing
  - Unbiased sequencing of unknown RNA or DNA molecules

Small RNA

- size ranges from 20 to 27 nucleotides
- miRNA, siRNA and piRNA
- play essential roles in the all eukaryotes kingdoms (protists, animals, plants, and fungi)
- involved in different phenomena essential for
  - genome stability, development, adaptive responses to biotic and abiotic pressures, control of transposable elements, antiviral defence...
- effects of small RNAs on gene expression are generally inhibitory

Illumina Solexa technology

- Ultra High Throughput sequencing:
  - One machine = 150 M sequences of 36 - ~150 nucleotides per week
  - One human genome in two weeks
  - Sanger has about 40 machines
  - Bejing Genomic Institute has eq. >300 machines with the throughput described here

Equipment

Flow cell

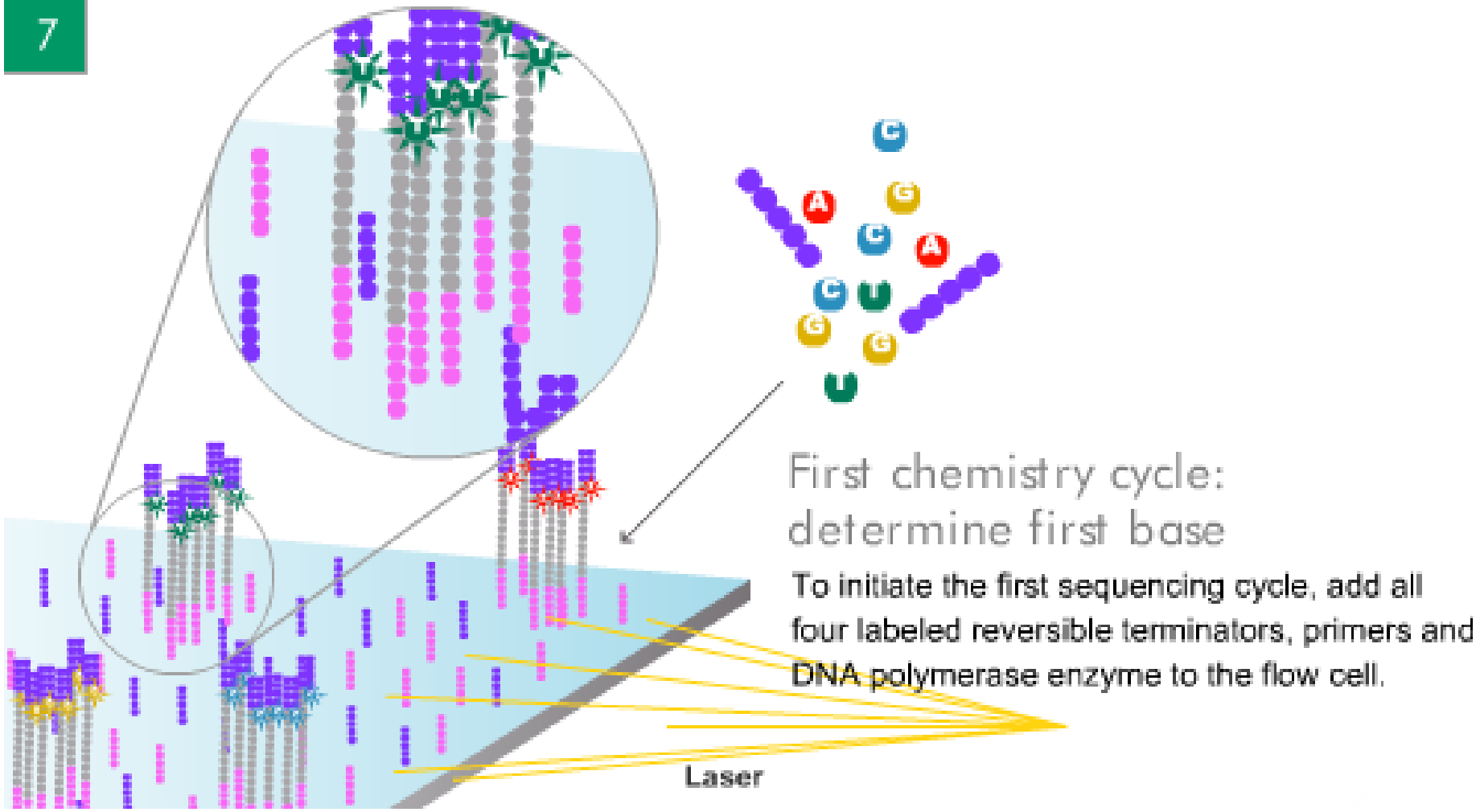
Cluster Station

Genome Analyser

Paired-end module

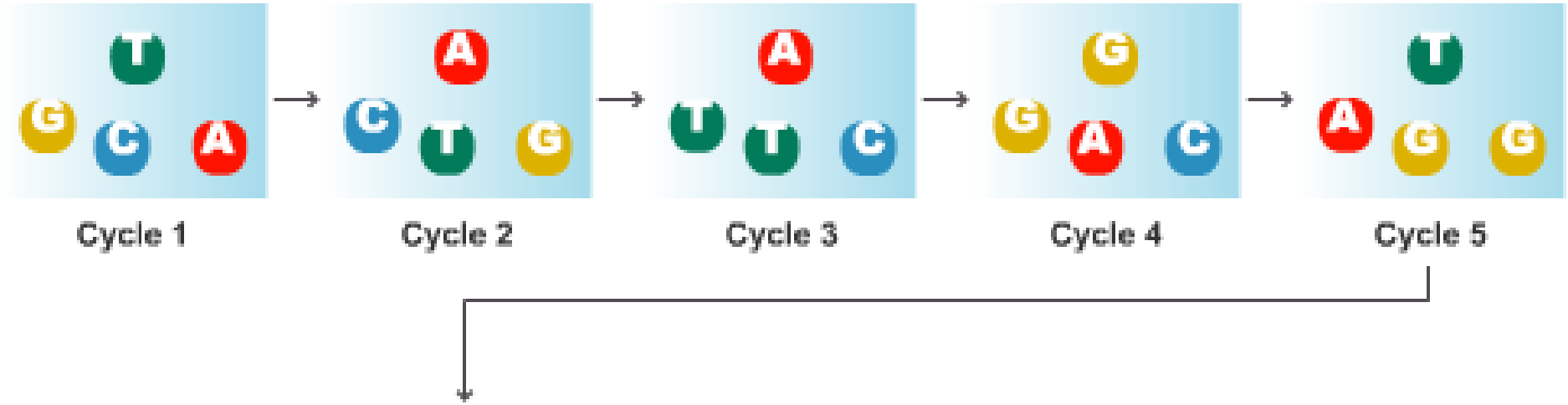
Principle: sequencing by synthesis

7



Sequence by synthesis

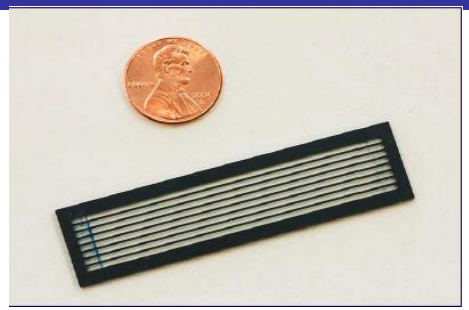
11



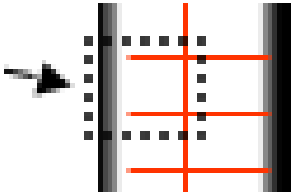
Sequence read over multiple chemistry cycles

Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at a time.

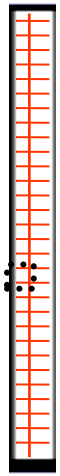
1 Cycle



« tile »

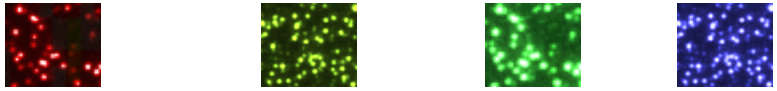


x 100



x 8

genome analyser



x 100 x 8

7 Mb	+	7 Mb	+	7 Mb	+	7 Mb	=	28 Mb	→	2,8 Gb	→	**22,4 Gb**
1	+	1	+	1	+	1	=	4 files	→	400 files	→	3200 files

* Each file contains 50 000 - 150 000 lines (clusters)

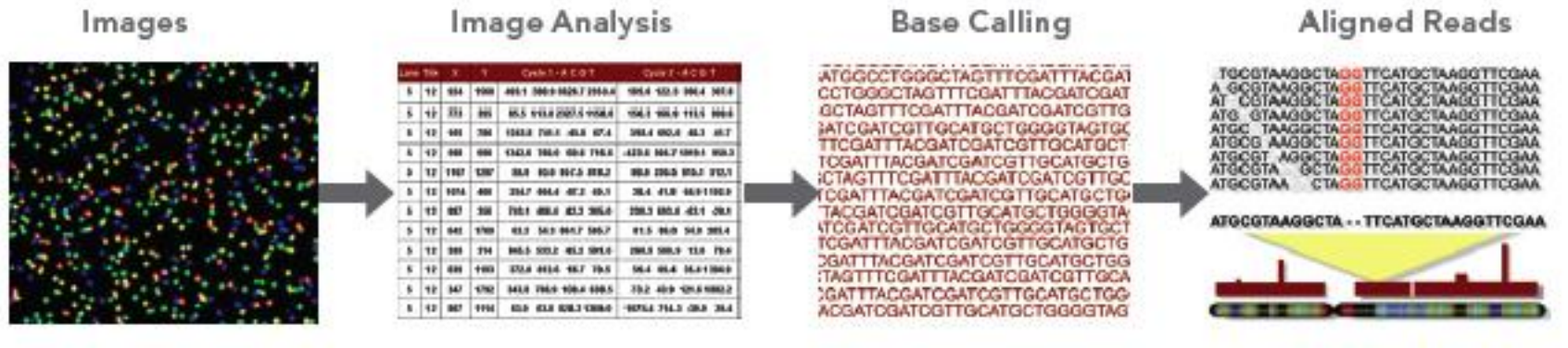
1 run

1 cycle	22.4 Gb – 3200 files
36 cycles	806 Gb – 115.200 files
72 cycles	1.6 Tb – 230.400 files
72 cycles Paired end	3.2 Tb – 460.800 files

NGS data processing

- ***NGS Data primary data analysis***
  - ***Image analysis***
  - ***Base calling***
  - ***Alignment***
  - ***Quality control***
- ***NGS Data secondary data analysis***
  - ***Functional annotation***
  - ***Visualization***
  - ***Classification***
  - ***Quantitative/Qualitative analysis***

NGS – primary data analysis



	Images	Image Analysis	Base Calling	Aligned Reads
	• Transfert of image files • Identification of clusters	• Cluster position • Clusters intensities • Signal to noise analysis => One file per cycle	• Intensity correction • Quality assesment • Calculate sequence for each cluster	• Align sequences to reference genome • Sequence files • Statistics • Graphics
Number of files	360.000+	5.000	10.000	36.000
Disc space used	TBs	~100 GB	~ 80 GB	~ 25 GB

NGS – secondary data analysis

- Making sense of the alignments
  - Visualization (Genome Browser)
  - Statistical interpretation

KNIME - Content

- ***Why KNIME***
- ***Mobyle***
- ***GBrowse***
- ***Implemented workflows***
- ***Under development:***
  - ***Upload to GBrowse***
  - ***Mobyle***
  - ***SSH***
  - ***Cytoscape***
- ***Availablity***
- ***Missing Features***

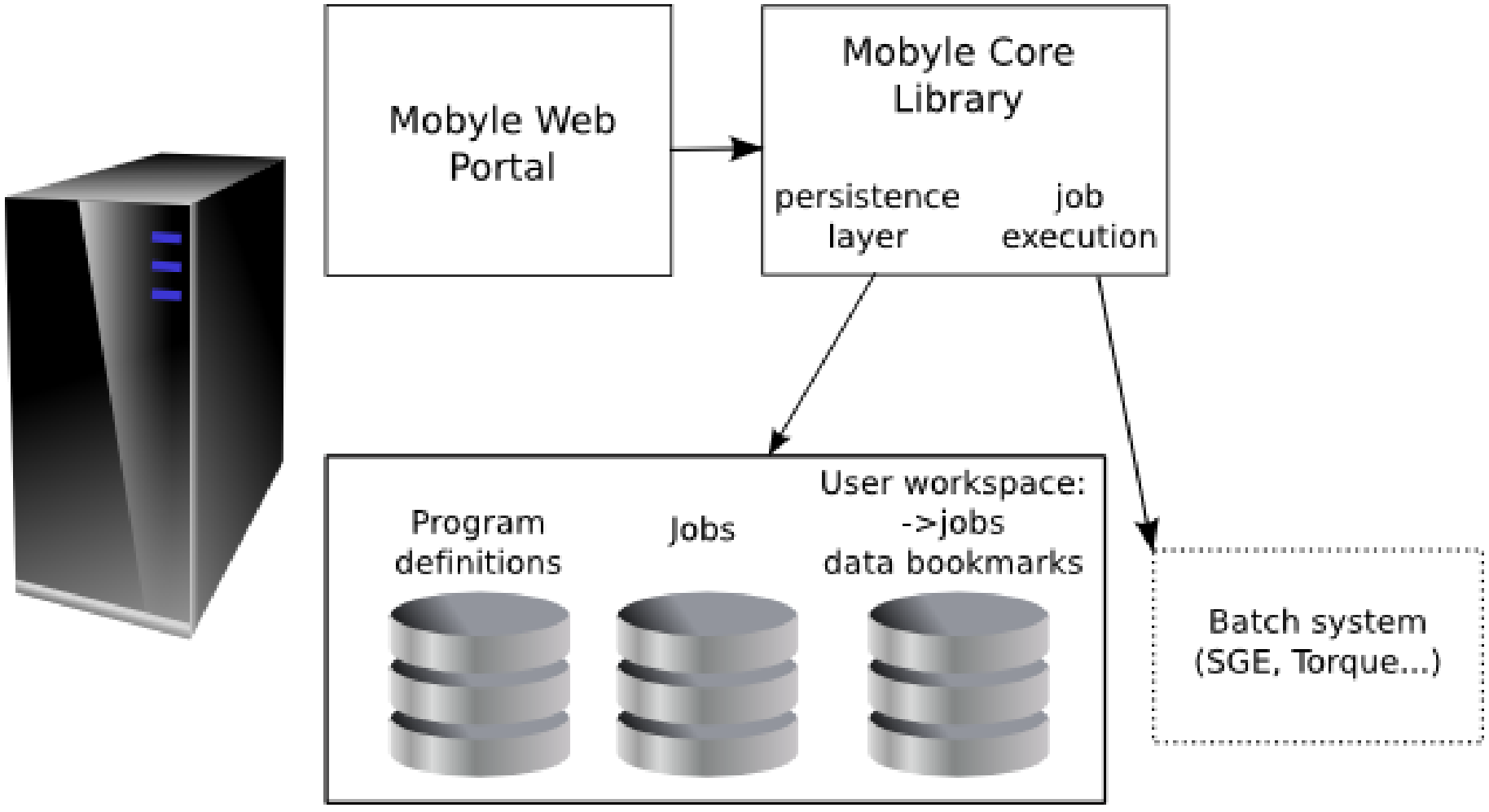
Why KNIME

- Replace command line scripts
- Graphical
- Tool integration
- Can deal with large number of rows
- Open-source
- Flexible
- Workflows are executable from command line
- Interact with data (Hilight)
- **Very good support**

Using KNIME

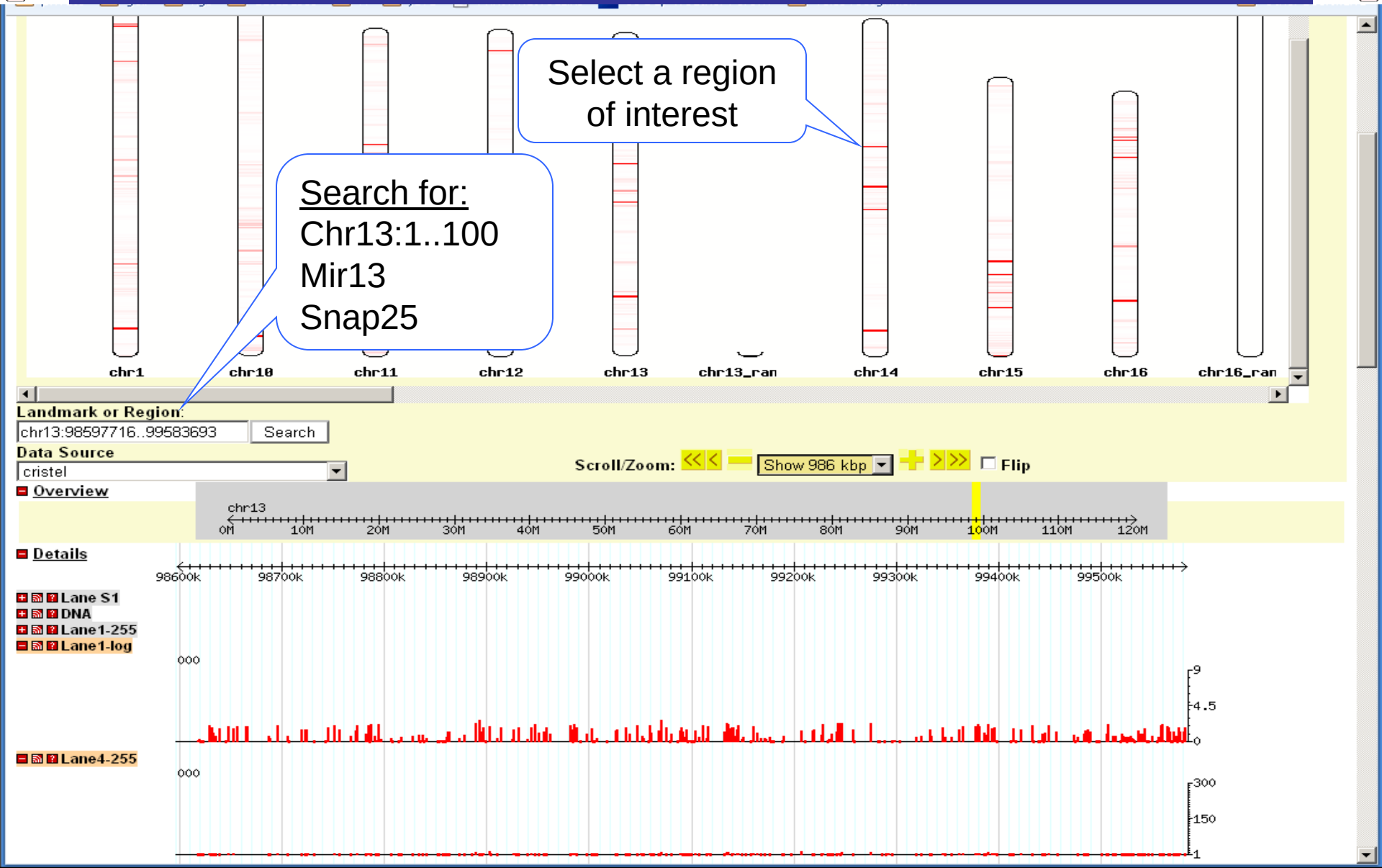
1. Develop workflow using sample data
2. Execute through command line via
   - Mobyle, Web-interface, directly…
3. Examine results using KNIME
4. Goto 1.

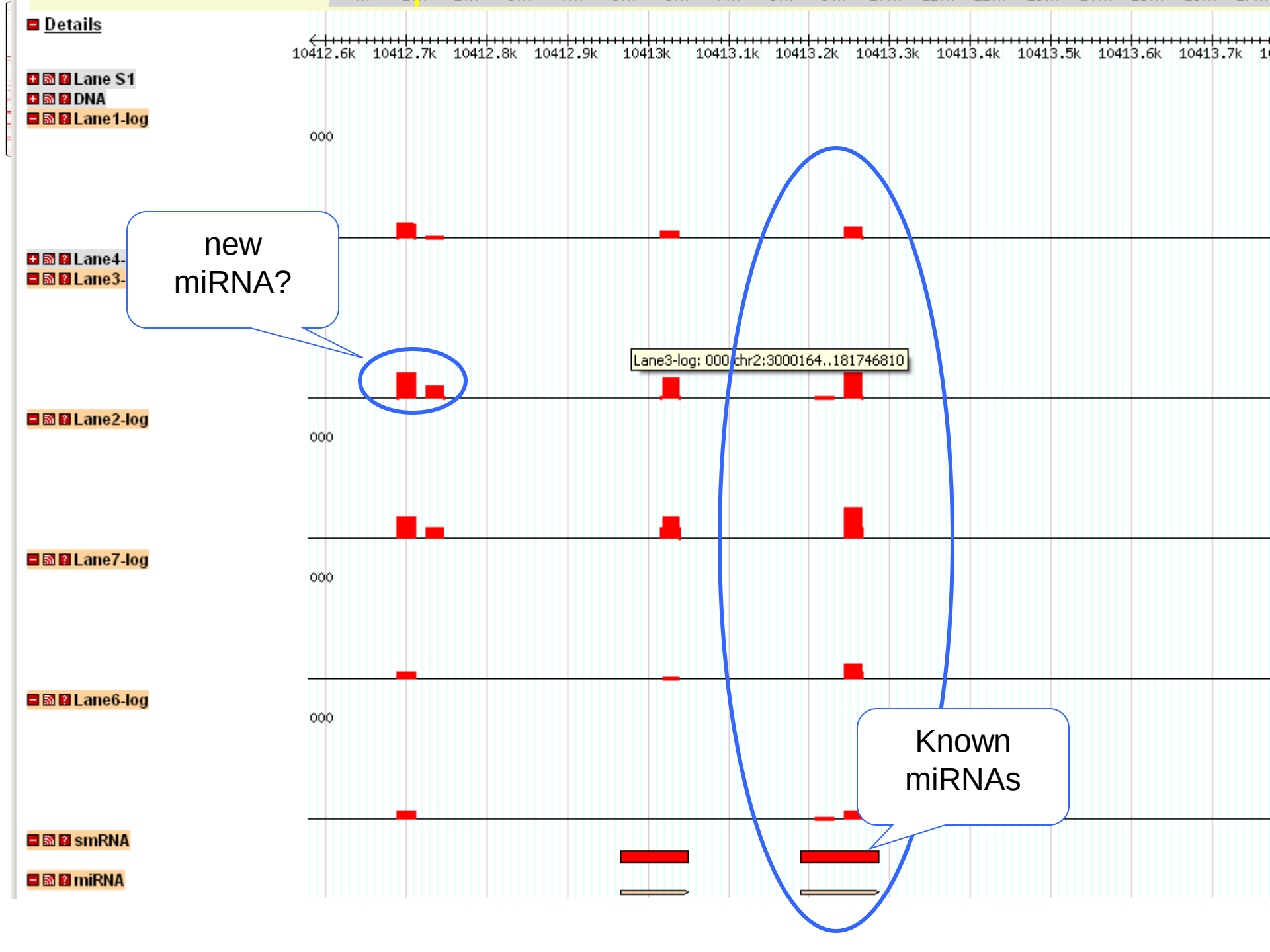
Mobyle



> 250 programs integrating, mostly dealing with sequence data

Visualization: GBrowse




Details
10412.6k
10412.7k
10412.8k
10412.9k
10413k
10413.1k
10413.2k
10413.3k
10413.4k
10413.5k
10413.6k
10413.7k
Lane S1
DNA
Lane1-log
000
Lane4-
Lane3-
new miRNA?
Lane3-log: 000 chr2:3000164..181746810
Lane2-log
000
Lane7-log
000
Lane6-log
000
Known miRNAs
smRNA
miRNA

Landmark or Region:

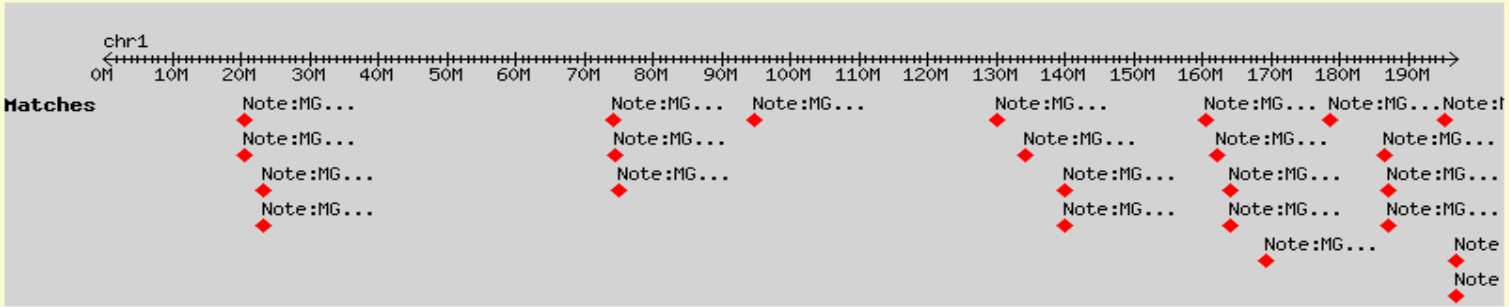
mir Search

Data Source

cristal

The following 555 regions match your request.

Matches on chr1



Note:MGI:2676881 Mir206	**chr1**:20.67..20.67 Mbp (73 bp)	score=10
Note:MGI:3618720 Mir133b	**chr1**:20.67..20.67 Mbp (119 bp)	score=10
Note:MGI:2676907 Mir30a	**chr1**:23.28..23.28 Mbp (71 bp)	score=10
Note:MGI:3619048 Mir30c-2	**chr1**:23.3..23.3 Mbp (84 bp)	score=10
Note:MGI:3836960 Mir1928	**chr1**:74.25..74.25 Mbp (67 bp)	score=10
Note:MGI:2676901 Mir26b	**chr1**:74.44..74.44 Mbp (85 bp)	score=10
Note:MGI:3619376 Mir375	**chr1**:74.95..74.95 Mbp (64 bp)	score=10
Note:MGI:2676834 Mir149	**chr1**:94.75..94.75 Mbp (66 bp)	score=10
Note:MGI:2676813 Mir128-1	**chr1**:130.1..130.1 Mbp (70 bp)	score=10
Note:MGI:3618732 Mir135b	**chr1**:134.1..134.1 Mbp (97 bp)	score=10
Note:MGI:3629589 Mir181a-1	**chr1**:139.9..139.9 Mbp (87 bp)	score=10
Note:MGI:3618735 Mir181b-1	**chr1**:139.9..139.9 Mbp (80 bp)	score=10
Note:MGI:3629597 Mir488	**chr1**:160.4..160.4 Mbp (109 bp)	score=10
Note:MGI:3836959 Mir1927	**chr1**:162.2..162.2 Mbp (109 bp)	score=10
Note:MGI:3618742 Mir199a-2	**chr1**:164.1..164.1 Mbp (110 bp)	score=10
Note:MGI:2676890 Mir214	**chr1**:164.2..164.2 Mbp (110 bp)	score=10
Note:MGI:3629598 Mir689-1	**chr1**:169.3..169.3 Mbp (109 bp)	score=10
Note:MGI:3619366 Mir350	**chr1**:178.7..178.7 Mbp (99 bp)	score=10
Note:MGI:3837225 Mir1981	**chr1**:186.6..186.6 Mbp (82 bp)	score=10
Note:MGI:2676858 Mir194-1	**chr1**:187.1..187.1 Mbp (67 bp)	score=10
Note:MGI:2676891 Mir215	**chr1**:187.1..187.1 Mbp (112 bp)	score=10
Note:MGI:2676880 Mir205	**chr1**:195.3..195.3 Mbp (68 bp)	score=10
Note:MGI:3619047 Mir29b-2	**chr1**:196.9..196.9 Mbp (81 bp)	score=10
Note:MGI:2676906 Mir29c	**chr1**:196.9..196.9 Mbp (88 bp)	score=10

Availability

- Currently through SVN for registered users of Pasteur project web-site (projets.pasteur.fr)
- Planed: open source of workflows and nodes (sourceforge???)

DEMO

- Tiling array analysis
- NGS data analysis