

KNIME and Next Generation
Sequencing:
From data cleansing to systems
biology

Bernd Jagla
PF2 – Transcriptome et Epigenome
Institut Pasteur

NGS at PF2

Technological approaches for NGS:

RNA seq

- Gene expression profiling (mRNAs, miRNAs, small RNAs...)
- Transcriptome annotation (TSS mapping, isoforms...)

ChiPSeq

- DNA-protein interactions (histone modifications, transcription factor binding sites...)

NGS at PF2

Many different organisms under study:

Viruses: Rift valley fever, Measles

Bacteria: *Listeria*, *Streptococcus*, *Enterococcus*, *Legionella*, *Clostridium*,
Thiomonas, *Helicobacter*

Yeasts: *Candida*, *Yarrowia*, *Aspergillus*, *Saccharomyces*, *Trichoderma*

Protozoans: *Plasmodium*, *Entamoeba*

Insects: *Drosophila*

Mammals: Mouse, Human

NGS at PF2

Biological questions under study:

Developmental biology

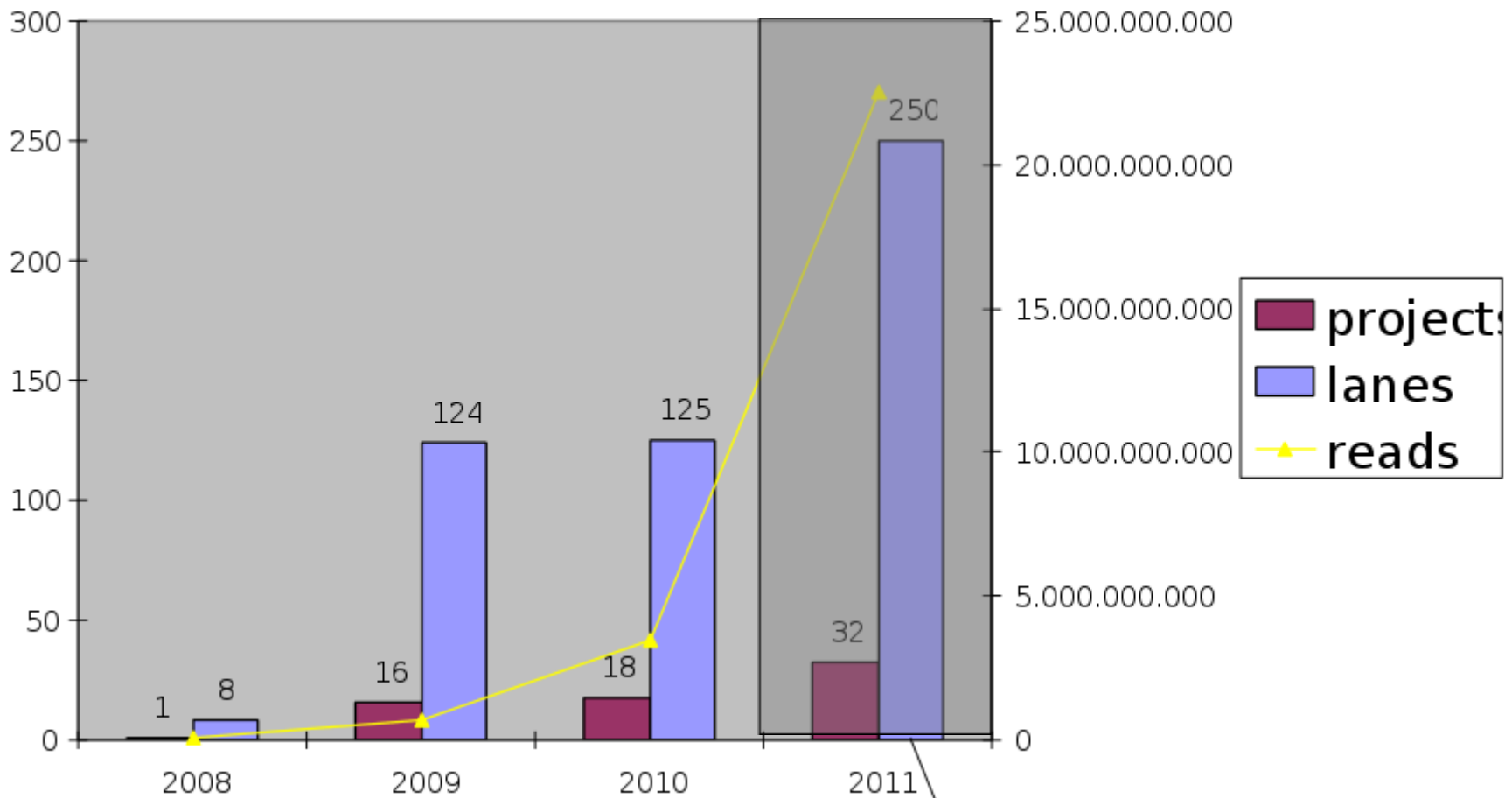
Infectious diseases (host-pathogen interactions, virulence factors...)

Microbiology of model organisms

 **This represents 35 different projects in the last 2 years**

Most of them require specific bioinformatics developments

Data throughput

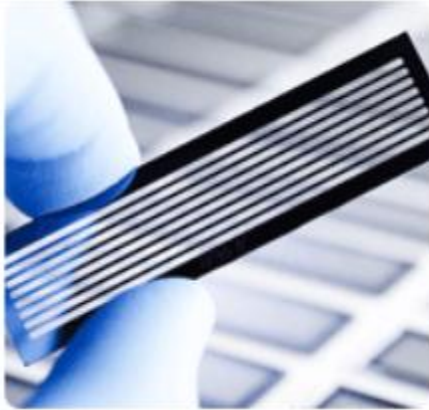


GAll: average number of reads: 25M / lane
HiSeq: first run: 85M / lane

Projected throughput
with HiSeq

Equipment

Flow cell



Cluster Station



Genome Analyser

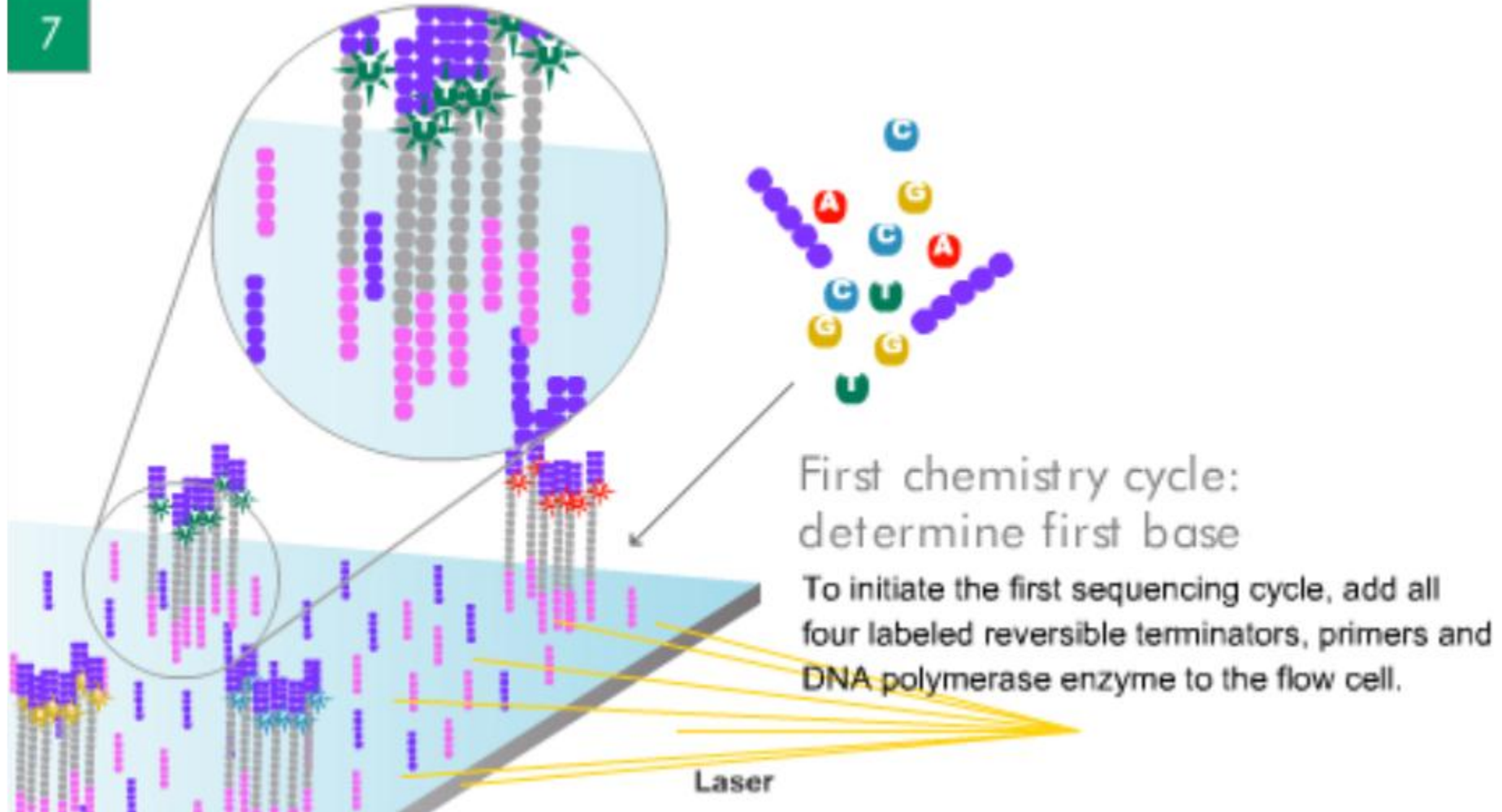


Paired-end module



Principle: sequencing by synthesis

7



Sequence by synthesis

11

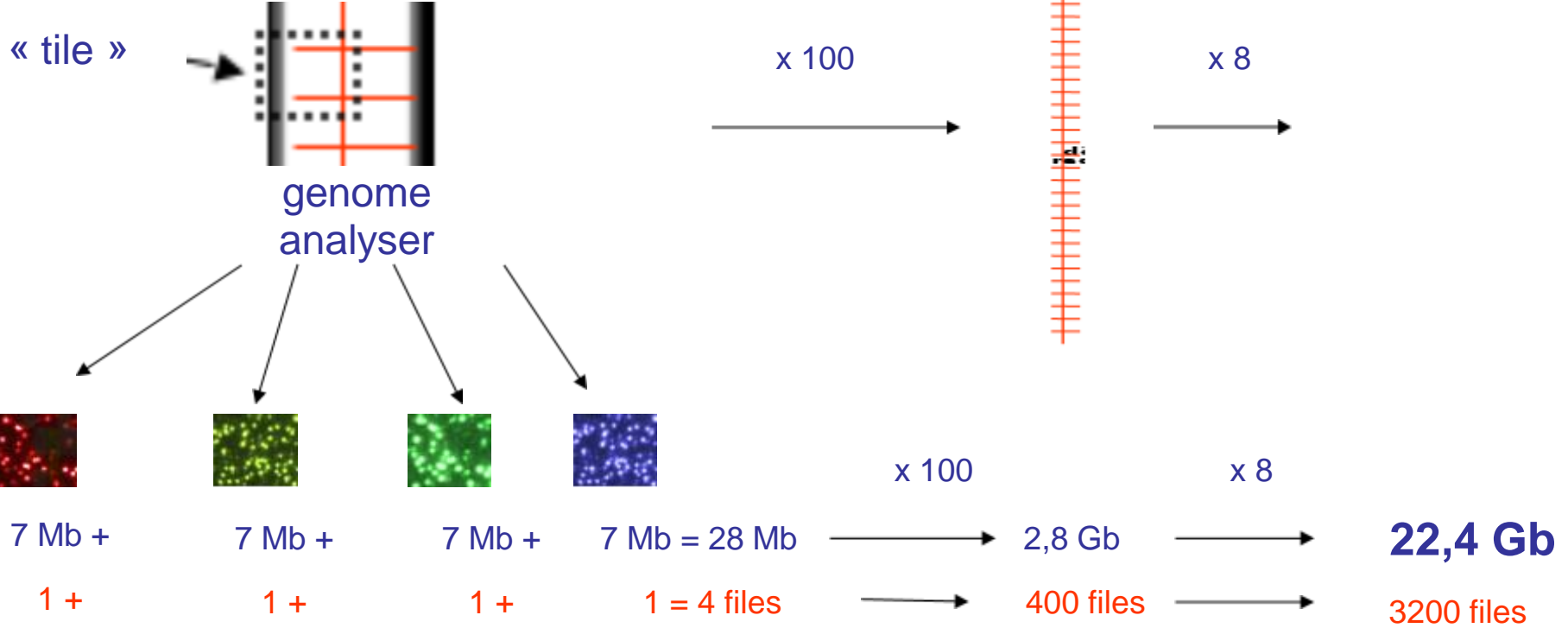
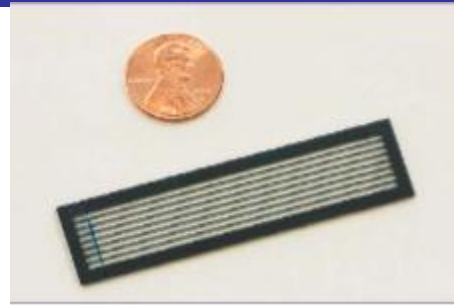


GCTGA....

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



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

HiSeq 2000 - throughput

HiSeq 2000 (Paired-end run)

Gb per run 150-200

Gb per day 20-25

Cluster density in KClusters/mm²** 260-350

Read length 2 x100

Available surface area (mm²)* 2880



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 3.2 Tb – 460.800 files

Paired end

Illumina Centers

BGI (Beijing Genomics Inst., China)

128 HiSeq 2000

Broad (Cambridge, MA, USA)

51 HiSeq 2000

Sanger Center (UK)

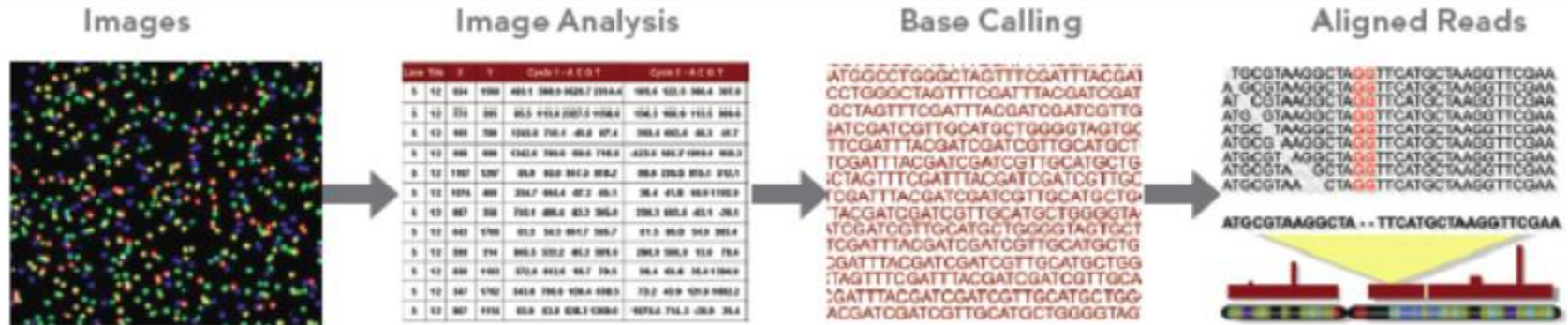
~ 37 Solexa machines

=> just these 3 centers can produce almost
1 PB per month

NGS data processing

- ***NGS Data primary data analysis***
 - *types and size; workflow overview; Image analysis; Base calling; Alignment; Quality control*
- ***NGS Data secondary data analysis***
 - *Functional annotation; Visualization; Classification; Quantitative/Qualitative analysis*

NGS – primary data analysis



- 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

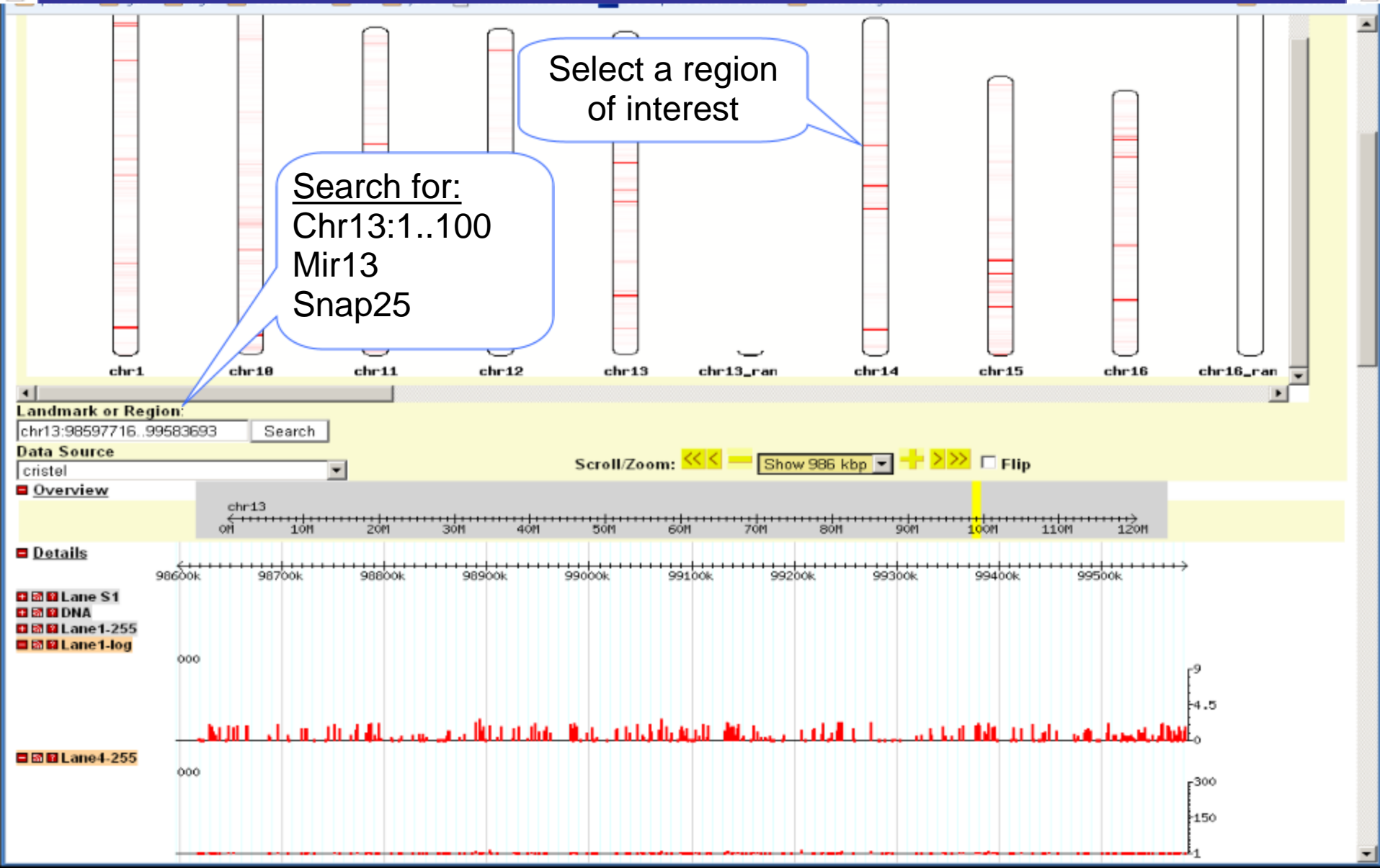
Bioinformatics tasks

- Quality control
- Descriptive statistics
- Mapping to reference genomes
- De-novo alignment
- Intersection with annotation
- Preparation for statistical analysis

NGS – *secondary data analysis*

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

Visualization: GBrowse



Nodes implemented

- **FastQReader** Reads in FastQ file into table. One FASTQ entry (i.e. 4 lines) are translated into one row. This node is using BioJava
- **FastQWriter** Writes out FastQ file into a file. This node is using BioJava.
- **BEDGraphWriter** Writes out BED files.
- **SAMReader** Reads Sam or Bam files.
- **AdapterRemoval** Node to remove adapter sequences.

Nodes implemented

- **Bash** Executes commands in bash or cmd.exe
- **CmdwInput** Similar to the bash node only that it takes the input table and executes strings within that table.
- **JoinSorted** Creates a full outer join of two sorted tables.
- **CountSorted** Counts occurrences within a sorted column. It is faster than the ValueCounter and useful for counting reads from a FASTQ file as they are already sorted. It also uses minimum amount of memory.

Nodes implemented

- **GetRegions** Identifies regions of interest (ROI). A ROI is defined as a chromosomal region that has no gaps. This node also produces a string of concatenated values (i.e. counts)
- **PositionStr2Position** Takes a string (chr1_123) and translates it into two columns (“chr1”, 123)
- **RegionOverlap** Identifies regions that overlap. This node is usually used within a sub-workflow that divides the data set per chromosome. The first input node is being retained.
- **Seq2PosIncidents** This node splits a sequence into one nucleotide per row.

Nodes implemented

- **OneString** This creates a single cell of type String, Integer, or Double. It is use full when executing workflows from the command line.
- **Wait** Does nothing other than synchronizing executions. This can also be done using the Variable Ports of existing nodes

KNIME workflows implemented

- RNA-Seq
 - Gene expression profiling
 - TSS mapping
 - Small RNA characterization
- ChiP-Seq
 - Histone modifications
 - DNA/protein interactions
- Implemented workflows are routinely being executed by Caroline Proux and Odile Sismeiro

KNIME workflows implemented

- Adapter removal
 - Multiple copies of adapter with variations
 - Length selection
 - Quality score selection
 - Complexity filter

KNIME workflows implemented

- Mapping to reference genome
 - Using bowtie
 - Creating SAM/BAM files
 - Creates Pileup (counts per position)

Node in Alpha

- **FASTAReader** reads fasta, genbank, uniprot, embl, INSDseq files and creates a new sequence object including annotations
- **GetSequenceName** extracts the sequence name from a sequence object
- **SubSequence** extracts a sequence from a sequence object
- **IGVView** open IGV at a particular chromosomal location
- **GBrowseViewer** opens GBrowse at a particular chromosomal position
- **Upload2Gbrowse** uploads data to a GBrowse instance
- **DASClient** retrieve annotation from a DAS server
- **MobyleWebServiceClient** launches a program through Mobyle
- **AdapterRemovalAdv** more advance read cleansing.

KNIME workflows in development

- Region of interest (ROI) identification
 - Select mapped reads
 - Distinguish between forward/reverse strand
 - Distinguish between unique/non-unique/all aligned reads
 - Generate ROIs of ungapped genomic regions
 - Identify possible Mutations/SNPs
 - Generate BED files and table for further analysis
- KNIME integration of Statistical analysis tools for normalization / differential analysis
- Integration of Cytoscape / GOSeq,
- Specific developments according to user's needs

Conclusions

- 2nd stage data analysis for NGS is
complex
time-consuming
individual
rewarding
- Completed (KNIME) workflows available for
 - Data cleansing
 - 1st stage data analysis (mapping)
 - Intersection with annotation
 - Most of the individual tasks are available
(TSS, strand specificity, intersection with annotation, Pivoting,
prototypes for integration with: DAS , UCSC & GBrowse, IGV,
BioJava, Mobyly, SAMtools, ROI tools)

Whish list

- Random access tables (see samtools, bzip)
- Visualization of arbitrary large data sets (pre-rendered graphs?)
- Vector/matrix operations for LARGE data
- Parallelization
- Faster chunk-loops (no rereading of the input table)
- Partial results should be available after cancel or failure...
- Workflows with the same name that are located in different wkfl groups cannot be distinguished
- Templates for java/python/perl snippets (see R Plot from community nodes)
- node to save workflow

Thanks

PF2

- Jean-Yves Coppee
- Marie-Agnès Dilles
- Odile Sismeiro
- Caroline Proux
- Guillaume Soubigoux

KNIME

- Bernd Wiswedel
- Michael Berthold
- Torsten Meinl
- Martin Horn
-

Collaborators

- Anastassia Komarova (Unité de Génomique Virale et Vaccination)
- PF1 (Christiane Bouchier, Sophie Creno)
- PF8 (Ghislaine Guigon)
- ENS (Laurent Jourdren, S Le Crom)
- Mobyly (Hervé Ménager, Bertrand Néron)
- LBD (Nicolas Joly, Bernard Caudron, Louis Jones)
- SR (Youssef Ghorbal, Jerome Sobecki)
- NGS users