



Galaxy KNIME integration
(NGS problematic and ?solutions?)
Bernd Jagla
Pasteur Institute, Paris

KOS, Konstanz, 7.9.2011

Thanks

PF2

- Jean-Yves
- Marie-Agnès
- Odile
- Caroline
- Guillaume

KNIME

- Bernd Wiswedel
- Michael Berthold
- Thorsten Meinl
- Thomas Gabriel
- And the rest of the KNIME team
- KNIME users

Collaborators

- Karol Kozak
- 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 Sobacki)
- NGS users

advertisement

Extending KNIME for next-generation sequencing data analysis

Bernd Jagla, Bernd Wiswedel, and Jean-Yves Coppée

Bioinformatics (2011) 27(20): 2907-2909 first published online
August 27, 2011 doi:[10.1093/bioinformatics/btr478](https://doi.org/10.1093/bioinformatics/btr478)

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

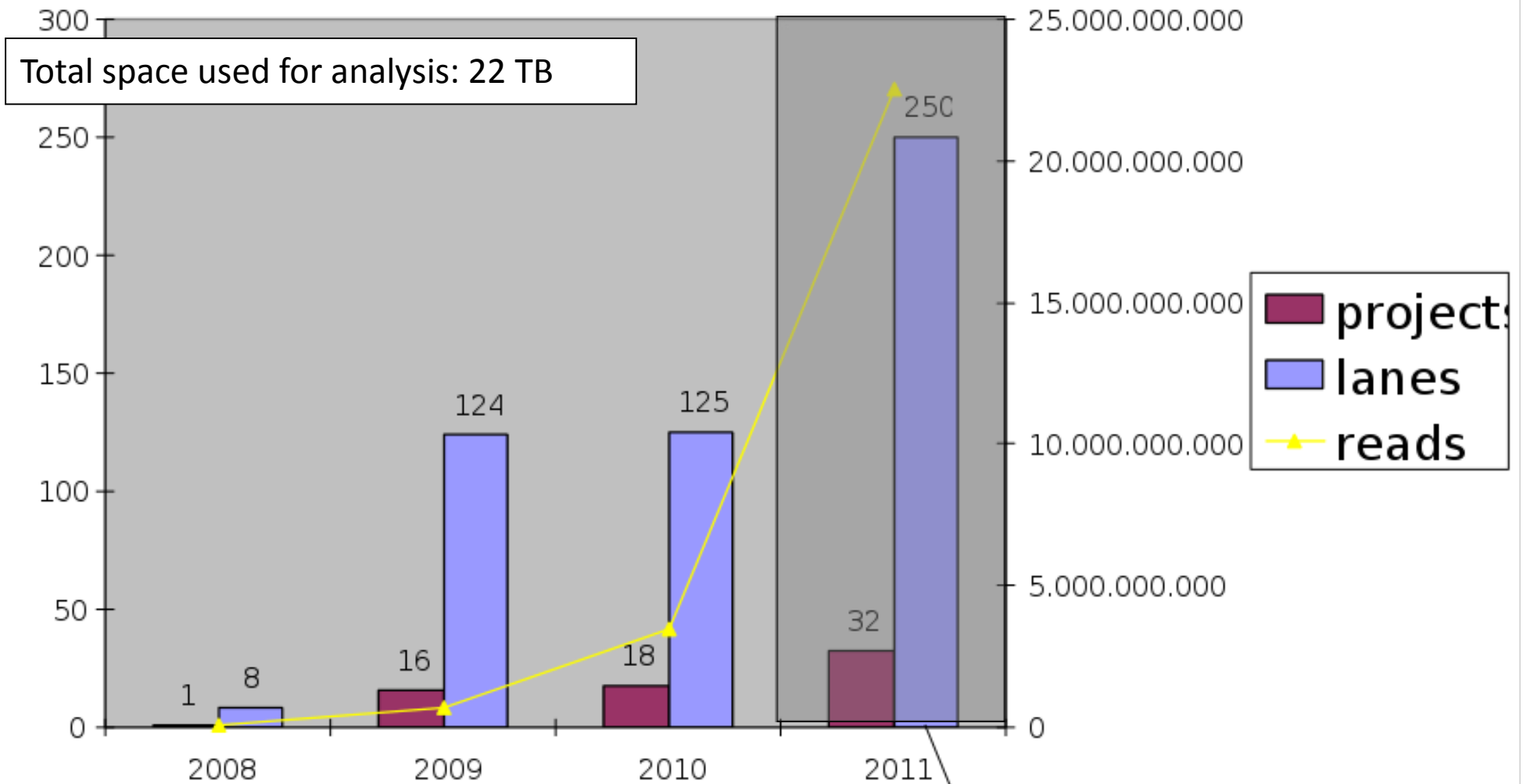
Infectious diseases (host-pathogen interactions, virulence factors...)

Microbiology of model organisms

➡ This represents 35 different projects in the last 2 years

Most of them require specific bioinformatics developments

Data throughput



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

Projected throughput
with HiSeq

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

Preprocessed data: >20GB

General problem

Biological question



Sequences

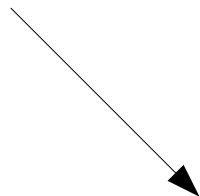
50 – 100 nucleotides

DNA/RNA

Many different biological questions



Statistical analysis



Paper / patent / fame

(i.e. no money)



script

Simple solution

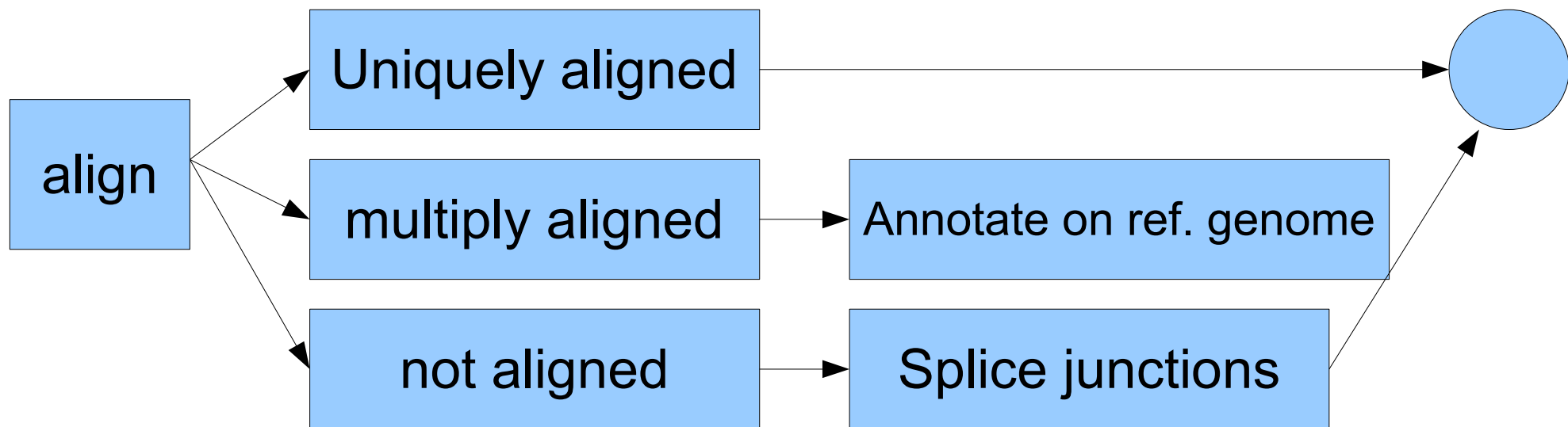
- Alignment / mapping to reference genome
- => only uniquely mapping sequences
- => counts per position/gene
- => statistics

Problem:

Biology is not that simple: solution dependant on e.g. reference genome, biological question, experimental details.

Desired solution

- Flexible workflows that are able to subdivide/join data based on individual requirements



QC, data cleansing, strand specific analysis,
mutations analyzed separately, user specific analysis

Different operation modi

- KNIME GUI on server, running ext. programs on cluster (using qsub)
- KNIME GUI on remote system running ext. programs on cluster (using ssh qsub), same filesystem mounted, maybe different mount points
- KNIME batch running on submit node (low memory consumption, mainly for submitting jobs using qsub)
- KNIME batch running on compute node (large memory single machine, predefine # CPUs), started using qsub

Implemented nodes

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.

Implemented nodes

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.

Burning nodes

DAS client retrieve information from a DAS server

IGVview/Gbrowse open view at genomic position

Mobyle Webservice client launch program through Mobyle

Upload2GBrowse upload features to GBrowse

FastAReader Reads fastA sequence files (and associated functionality)

GroupByLoopStart splits a table by chunks of row with the same value in specified column

EvenBetterSparkLine line graphs (SVG) in cells of table