KNIME and Next Generation Sequencing

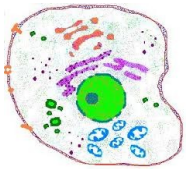
Bernd Jagla

PF2 – Transcriptome et Epigenome

Institut Pasteur

2 min Biology

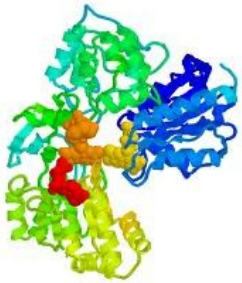
Organism



Cell

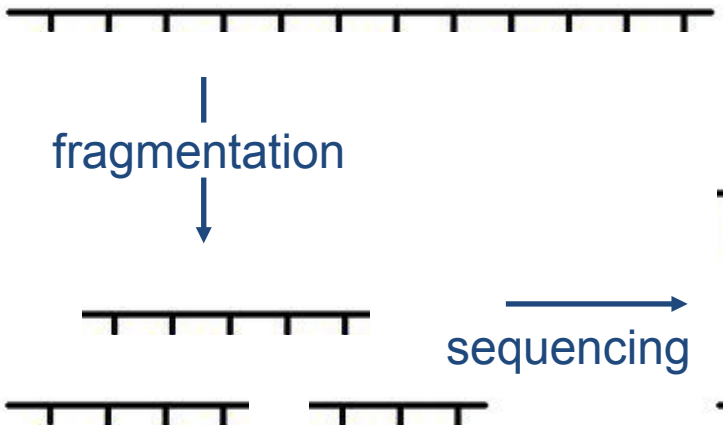
DNA

RNA



Protein

extraction



Alignment/
assembly

statistics

Organism	Estimated size (m bases)	Estimated gene number	Average gene density (bases)	Chromosome number
Homo sapiens (human)	2900	~30,000	1 gene per 100,000	46
Rattus norvegicus (rat)	2,750	~30,000	1 gene per 100,000	42
Mus musculus (mouse)	2,500	~30,000	1 gene per 100,000	40
Drosophila melanogaster (fruit fly)	180	13,600	1 gene per 9,000	8
Arabidopsis thaliana (plant)	125	25,50	1 gene per 4,000	10
Caenorhabditis elegans (roundworm)	97	19,100	1 gene per 5,000	12
Saccharomyces cerevisiae (yeast)	12	6300	1 gene per 2,000	32
Escherichia coli (bacteria)	4.7	3200	1 gene per 1,400	1
H. influenzae (bacteria)	1.8	1700	1 gene per 1,000	1

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

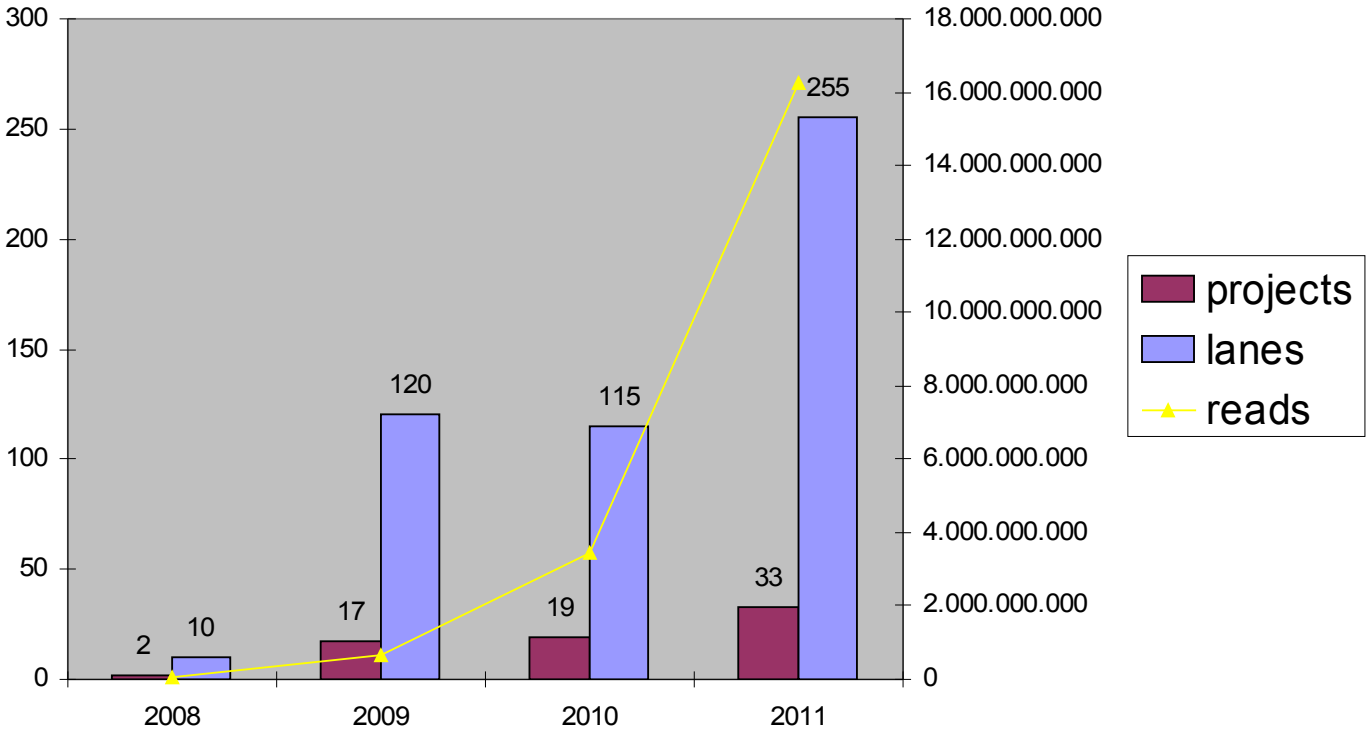
Infectious diseases (host-pathogen interactions, virulence factors...)

Microbiology of model organisms

This represents 35 different projects in the last 2 years

Most of them require specific bioinformatics developments

Data throughput



GAII: average number of reads: 25M / lane
HiSeq: 80-150M / lane

Stats

- Sequences produced in ~ 10 days
  - 2,287,716,764 (billion) strings of length 100
    (Paired end 100 run) 0.2 trillion nucleotides
- Storage used
  - ~ 2TB (continuously removing “old” data)
- Time

  Min. 2 weeks for data analysis
  (up to several years)

Nodes implemented (NGS specific)

- **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 (sequence region specific)

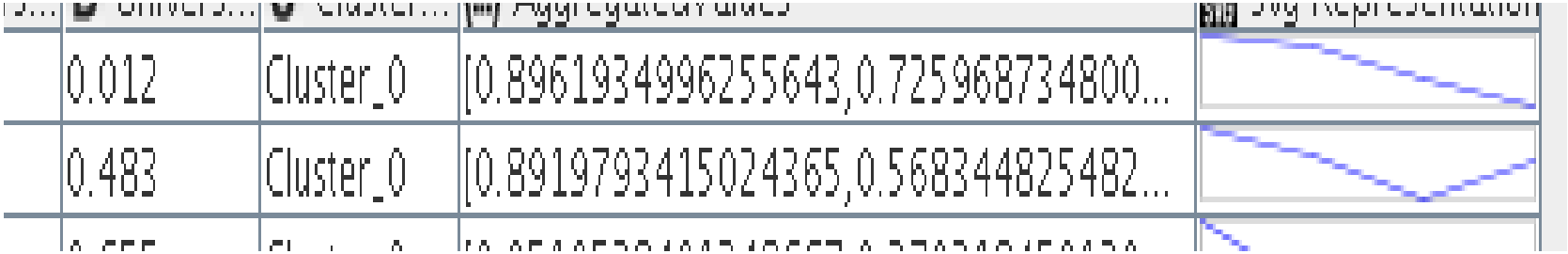
- **GetRegions** Identifies regions of interest (ROI). A ROI is defined as a chromosomal region that has no gaps. This node also produces a string of concatenated values (i.e. counts)
- **PositionStr2Position** Takes a string (chr1_123) and translates it into two columns ("chr1", 123)
- **RegionOverlap** 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 (general purpose)

- **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.
- **NGSconcat** concatenate tables with identical table specs.

Nodes implemented (general purpose)

- **GroupByLoopStart** Loop start node that iterates over parts of the input table that have constant values
- **CollectionLinePlot** Line Plot for numerical collections



- **TableSpecs** Retrieves simple stats for table and columns(n) included are column type, index, lower and upper bound (table 1) number of rows and columns (table2)

Publication

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

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

SequenceReader Reads sequence files from various fromats (fasta, genbank, embl,...)

AnnotationReader Reads annotation from various formats (genbank, embl)

Thanks

PF2

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

KNIME team

- Bernd Wiswedel
- Thorsten Meinl
- Michael Berthold
- Thomas Gabriel

Karol Kozak (ETH Zürich)

Collaborators

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

Please contact me

Bernd.jagla@pasteur.fr

If you have questions/suggestions

Are interested in importing and working sequence data (very large strings) into KNIME