Plateforme Bioinformatique Midi-Pyrénées

Aligning NGS reads and SNP finding

Philippe Bardou & Christophe Klopp

Plateforme Bioinformatique Midi-Pyrénées

Organisation

Morning (9h30 -12h30) :

- Sequence quality
  - Theory + exercises
- Read mapping
  - Theory + exercises

Afternoon (14h-17h) :

- SAM format
  - Theory + exercises
- Visualisation
  - Theory + exercises
- SNP calling
  - Theory + exercises

Plateforme Bioinformatique Midi-Pyrénées

Why are you here?

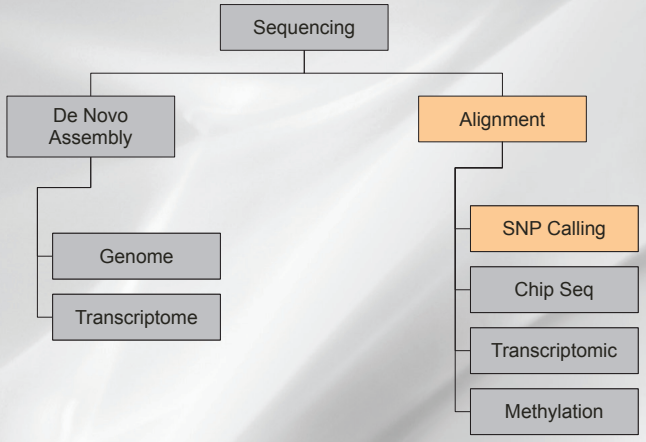
Because

- You have produced or are going to produce a lot of reads for which you want to find variation between conditions or individuals or populations?
- You use :
  - Whole Genome Sequencing,
  - Sequence capture,
  - Reduced Representation Libraries,
  - Long range Polymerase Chain Reaction,
  - Amplification Fragment Length Polymorphisms.
- You use Solexa or Roche 454 platforms !
- You are curious!

Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

Where are we?



Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

What are you going to learn?

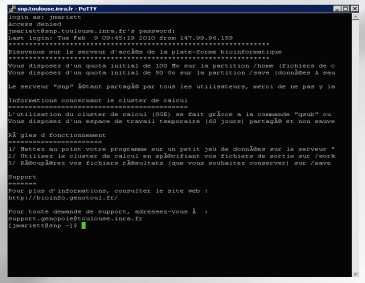
- To extract reads and reference genome from the NCBI
- To verify the read quality
- To format reference sequence
- To align the reads on the reference genome
- To index the reference sequence and the aligned reads
- To visualise the alignments and variations
- To call SNPs

Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

What you should already know?

- How to connect to a remote unix server (putty)?
- What a unix command looks like?
- How to move around the unix environment?
- How to edit a file?



Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

The pieces of software

- Quality : fastqc
- BWA : alignment
- Samtools : formating SNP discovery
- IGV : visualisation

Babraham Bioinformatics FastQC

Manual Reference Pages - bwa (1)

Integrative Genomics Viewer

SAMtools

SOURCEFORGE.NET

Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

The 1000 genomes project

- Joint project NCBI / EBI
- Common data formats :
  - fastq
  - SAM (Sequence Alignment/Map)

1000 Genomes
A Deep Catalog of Human Genetic Variation

LATEST ANNOUNCEMENTS

March 2010 Data Release

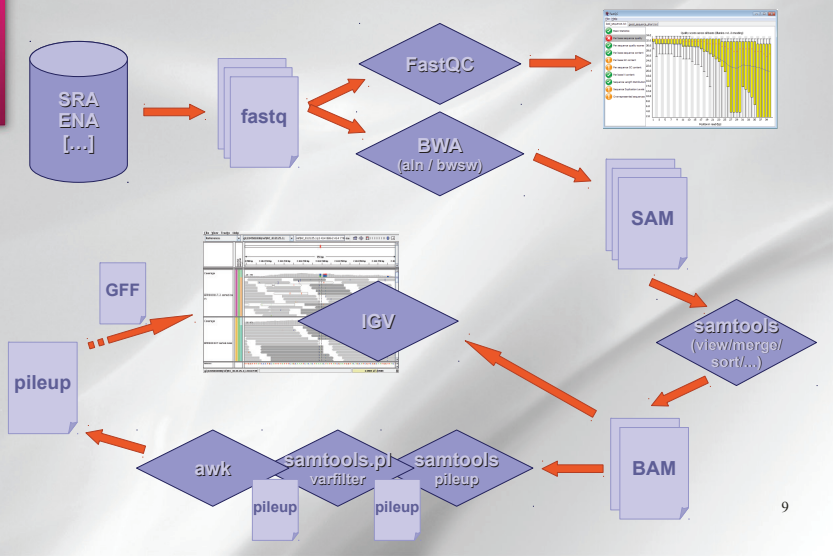
Final release of pilot project SNP calls

Recent project announcements

Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

Overview

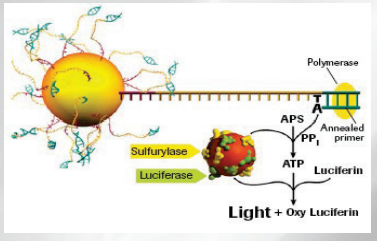


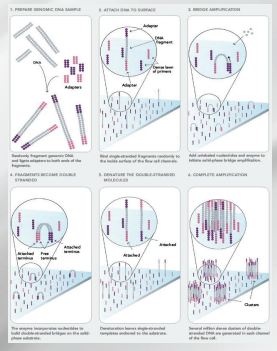
Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

NGS platforms

- Two platforms :
  - Illumina Solexa
  - Roche 454





Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

NGS reads

454 / Roche – Genome Sequence FLX

2006	2007	2009
Gs20 20Mb / run 100pb / lecture	GSFLX Standard 100Mb / run 250pb / lecture	GSFLX Titanium 500Mb / run 500pb / lecture

Illumina / Solexa – Genetic Analyzer

2007	2008	2009	2010
GA 1 1Gb / run 32pb / lecture	GA II 8Gb / run 50pb / lecture Lectures en paires	GA IIx 30Gb / run 100pb / lecture 10 jours / run Lectures en paires	**HiSeq 2000** **200 Gb/run** **100 pb /lecture** **Paired reads**

Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

Sequencing bias bibliography

Research

Open Access

Evaluation of next generation sequencing platforms for population targeted sequencing studies

Olivier Harismendy[¤*], Pauline C Ng[¤†], Robert L Strausberg[†], Xiaoyun Wang[*], Timothy B Stockwell[†], Karen Y Beeson[†], Nicholas J Schork[*], Sarah S Murray[*], Eric J Topol[*], Samuel Levy[†] and Kelly A Frazer[*]

Addresses: [*]Scripps Genomic Medicine - Scripps Translational Science Institute - The Scripps Research Institute, N. Torrey Pines Court, La Jolla, CA 92037, USA. [†]The J Craig Venter Institute, Medical Center Drive, Rockville, MD 20850, USA.

[¤] These authors contributed equally to this work.

Correspondence: Samuel Levy. Email: slevy@jcvi.org. Kelly A Frazer. Email: kfrazer@scripps.edu

Published: 27 March 2009

Genome Biology 2009, **10**:R32 (doi:10.1186/gb-2009-10-3-r32)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2009/10/3/R32

Received: 14 December 2008
Revised: 23 February 2009
Accepted: 27 March 2009

Published online 26 July 2008

Nucleic Acids Research, 2008, Vol. 36, No. 16 **e105**
doi:10.1093/nar/gkn425

Substantial biases in ultra-short read data sets from high-throughput DNA sequencing

Juliane C. Dohm[1], Claudio Lottaz[2], Tatiana Borodina[1] and Heinz Himmelbauer[1,*]

[1]Max Planck Institute for Molecular Genetics, Ihnestr. 63-73, 14195 Berlin and [2]Institute for Functional Genomics, Computational Diagnostics, University of Regensburg, Josef-Engert-Str. 9, 93053 Regensburg, Germany

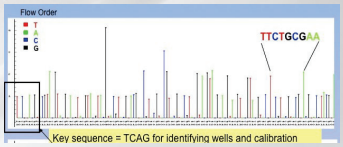
Received December 21, 2007; Revised June 16, 2008; Accepted June 19, 2008

Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

Sequencing bias

- Platform related
- Roche 454 (data from Jean-Marc Aury CNS)
  - 99,9% mapped reads
  - Mean error rate : 0,55%
  - 37% deletions, 53% insertions, 10% substitutions.
  - homopolymers errors
  - emPCR duplications
- Solexa (data from Jean-Marc Aury CNS)
  - 98,5% mapped reads
  - Mean error rate : 0,38%
  - 3% deletions, 2% insertions, 95% substitutions
  - Low A/T rich coverage



Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

What data will we use?

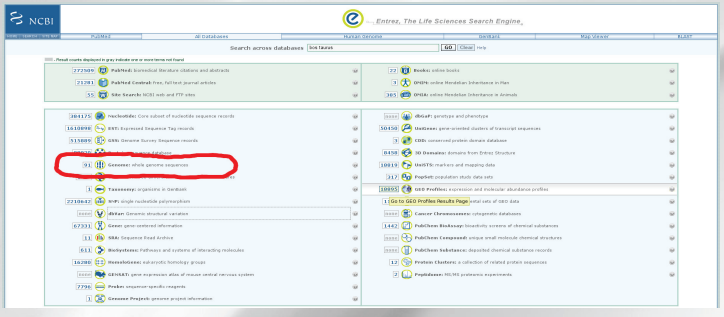
- The needed data :
  - A reference sequence :
    - Genome
    - Parts of the genome
    - Transcriptome
  - Short reads

Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

Where to get a reference genome?

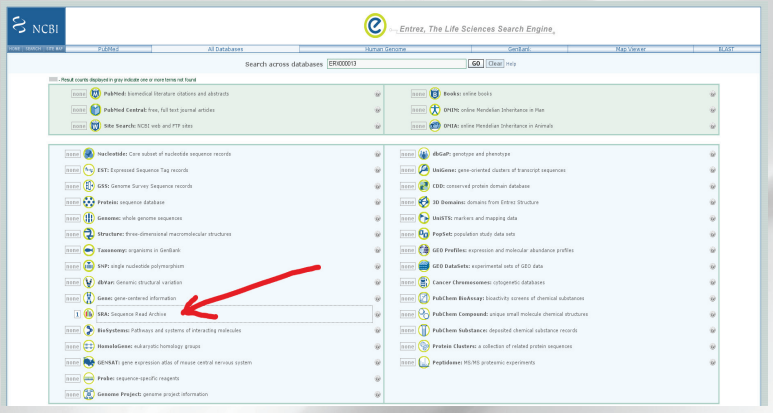
- Assemble your own
- Use a public assembly :
  - NCBI : Genbank
  - EMBL



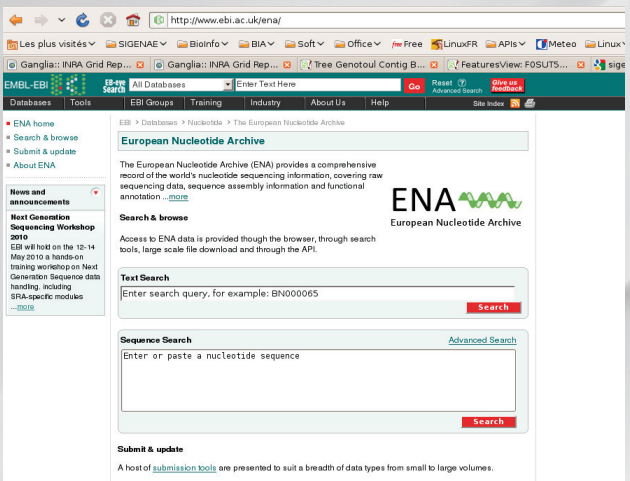
Where to get short reads?

- Produce your own sequences :
  - CNS
  - Local platform
  - Private company
- Use public data :
  - SRA : NCBI Sequence Read Archive
  - ENA : EMBL/EBI European Nucleotide Archive

NCBI SRA?



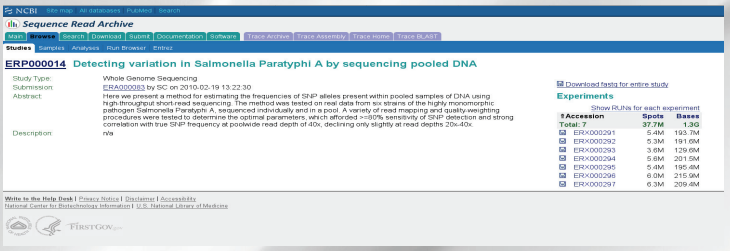
EBI ENA



Plateforme Bioinformatique Midi-Pyrénées

Meta data

- Meta data structure :
  - Experiment
  - Sample
  - Study
  - Run
  - Data file



Plateforme Bioinformatique Midi-Pyrénées

What is a fastq file

FASTQ format stores sequences and Phred qualities in a single file. It is concise and compact. FASTQ is first widely used in the Sanger Institute and therefore we usually take the Sanger specification and the standard FASTQ format, or simply FASTQ format. Although Solexa/Illumina read file looks pretty much like FASTQ, they are different in that the qualities are scaled differently. In the quality string, if you can see a character with its ASCII code higher than 90, probably your file is in the Solexa/Illumina format.

Example

```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
;;;;;;;;;;;7;;;;;-;;;3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;;9;7;;.7;393333
```


Plateforme Bioinformatique Midi-Pyrénées

fastq file formats

Published online 16 December 2009 *Nucleic Acids Research, 2010, Vol. 38, No. 6* **1767–1771**
doi:10.1093/nar/gkp1137

SURVEY AND SUMMARY

The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants

Peter J. A. Cock[1,*], Christopher J. Fields[2], Naohisa Goto[3], Michael L. Heuer[4] and Peter M. Rice[5]

Table 1. The three described FASTQ variants, with columns giving the description, format name used in OBF projects, range of ASCII characters permitted in the quality string (in decimal notation), ASCII encoding offset, type of quality score encoded and the possible range of scores

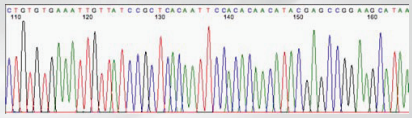
Description, OBF name	ASCII characters		Quality score	
	Range	Offset	Type	Range
Sanger standard `fastq-sanger`	33–126	33	PHRED	0 to 93
Solexa/early Illumina `fastq-solexa`	59–126	64	Solexa	−5 to 62
Illumina 1.3+ `fastq-illumina`	64–126	64	PHRED	0 to 62

$$Q_{\text{PHRED}} = -10 \times \log_{10}(P_e)$$

$$Q_{\text{Solexa}} = -10 \times \log_{10}\left(\frac{P_e}{1 - P_e}\right)$$

Sequence quality

- Phred : base calling



What is Phred Quality?

Traditionally, Phred quality is defined on base calls. Each base call is an estimate of the true nucleotide. It is a random variable and can be wrong. The probability that a base call is wrong is called error probability.

Explanation about the quality values :

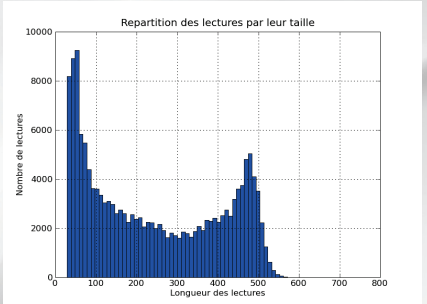
source http://maq.sourceforge.net/qual.shtml

Which reads should I keep?

- All
- Some : what criteria and threshold should I use
  - Composition (number of Ns, complexity,...),
  - Quality,
  - Alignment based criteria,
- Should I trim the reads using :
  - Composition
  - Quality

Basic reads statistics

- Number of reads
- Length histogram
- Number of Ns in the reads
- Reads quality
- Reads redundancy
- Reads complexity

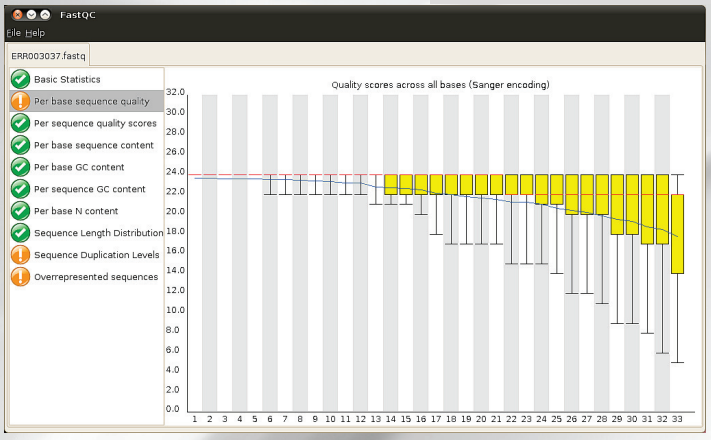


Plateforme Bioinformatique Midi-Pyrénées

Sequence quality analysis

- FastQC :

Babraham Bioinformatics
FastQC

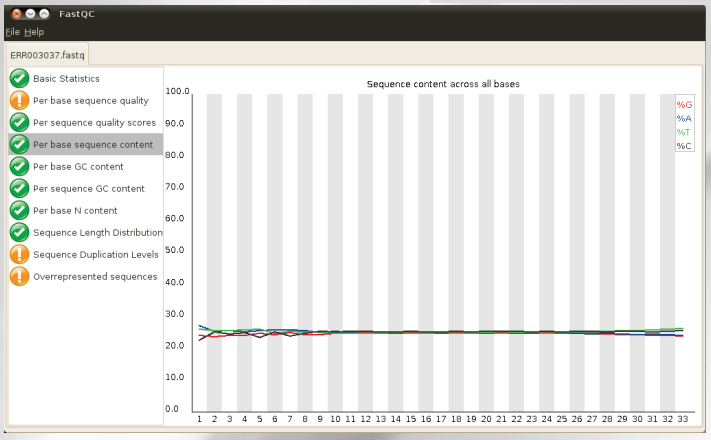


Plateforme Bioinformatique Midi-Pyrénées

Sequence quality analysis

- FastQC :

Babraham Bioinformatics
FastQC

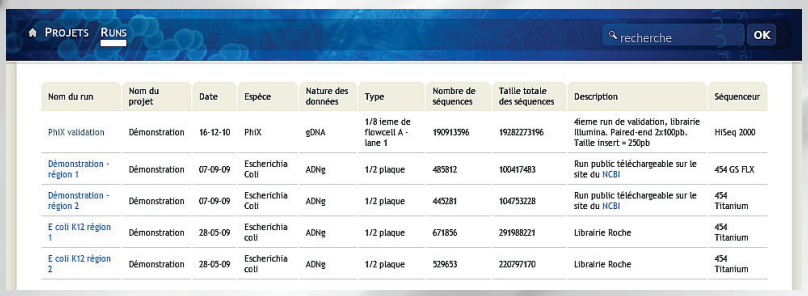


Plateforme Bioinformatique Midi-Pyrénées

NG6 - Home

Plateforme Bioinformatique Midi-Pyrénées

NG6 - Runs

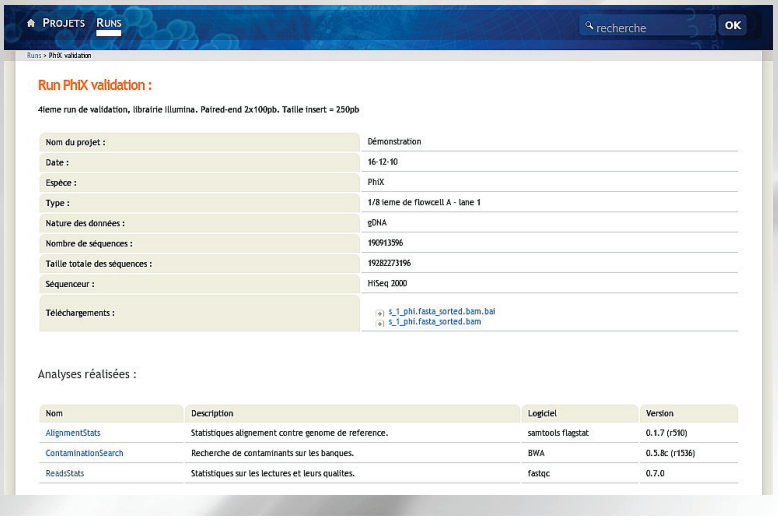


PROJETS RUNS recherche OK

Nom du run	Nom du projet	Date	Espèce	Nature des données	Type	Nombre de séquences	Taille totale des séquences	Description	Séquenceur
PhiX validation	Démonstration	16-12-10	PhiX	gDNA	1/8 lieme de flowcell A - lane 1	190913596	19282273196	4ieme run de validation, librairie Illumina. Paired-end 2x100pb. Taille insert = 250pb	HiSeq 2000
Démonstration - région 1	Démonstration	07-09-09	Escherichia Coli	ADNg	1/2 plaque	485812	100417483	Run public téléchargeable sur le site du NCBI	454 GS FLX
Démonstration - région 2	Démonstration	07-09-09	Escherichia Coli	ADNg	1/2 plaque	445281	104753228	Run public téléchargeable sur le site du NCBI	454 Titanium
E coli K12 région 1	Démonstration	28-05-09	Escherichia coli	ADNg	1/2 plaque	671856	291988221	Librairie Roche	454 Titanium
E coli K12 région 2	Démonstration	28-05-09	Escherichia coli	ADNg	1/2 plaque	529653	220797170	Librairie Roche	454 Titanium

Plateforme Bioinformatique Midi-Pyrénées

NG6 - Run



PROJETS RUNS recherche OK

Runs > PhiX validation

Run PhiX validation :

4ieme run de validation, librairie Illumina. Paired-end 2x100pb. Taille insert = 250pb

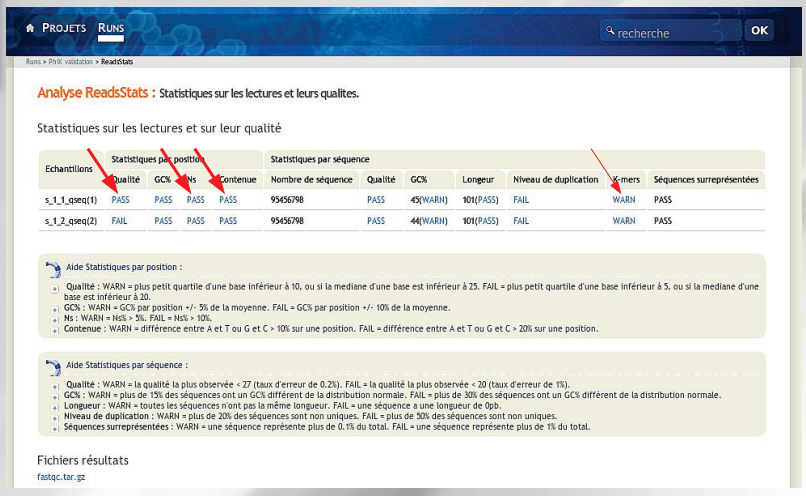
Nom du projet :	Démonstration
Date :	16-12-10
Espèce :	PhiX
Type :	1/8 lieme de flowcell A - lane 1
Nature des données :	gDNA
Nombre de séquences :	190913596
Taille totale des séquences :	19282273196
Séquenceur :	HiSeq 2000
Téléchargements :	s_1_phi.fasta_sorted.bam.bai s_1_phi.fasta_sorted.bam

Analyses réalisées :

Nom	Description	Logiciel	Version
AlignmentStats	Statistiques alignement contre genome de reference.	samtools flagstat	0.1.7 (r510)
ContaminationSearch	Recherche de contaminants sur les banques.	BWA	0.5.8c (r1536)
ReadsStats	Statistiques sur les lectures et leurs qualites.	fastqc	0.7.0

Plateforme Bioinformatique Midi-Pyrénées

NG6 - Stats



PROJETS RUNS recherche OK

Runs > PhiX validation > ReadsStats

Analyse ReadsStats : Statistiques sur les lectures et leurs qualites.

Statistiques sur les lectures et sur leur qualité

Echantillons	Statistiques par position				Statistiques par séquence						
	Qualité	GC%	Ns	Contenue	Nombre de séquence	Qualité	GC%	Longeur	Niveau de duplication	K-mers	Séquences surreprésentées
s_1_1_qseq(1)	PASS	PASS	PASS	PASS	95456798	PASS	45(WARN)	101(PASS)	FAIL	WARN	PASS
s_1_2_qseq(2)	FAIL	PASS	PASS	PASS	95456798	PASS	44(WARN)	101(PASS)	FAIL	WARN	PASS

Aide Statistiques par position :

- Qualité : WARN = plus petit quartile d'une base inférieur à 10, ou si la mediane d'une base est inférieur à 25. FAIL = plus petit quartile d'une base inférieur à 5, ou si la mediane d'une base est inférieur à 20.
- GC% : WARN = GC% par position +/- 5% de la moyenne. FAIL = GC% par position +/- 10% de la moyenne.
- Ns : WARN = Ns% > 5%. FAIL = Ns% > 10%.
- Contenue : WARN = différence entre A et T ou G et C > 10% sur une position. FAIL = différence entre A et T ou G et C > 20% sur une position.

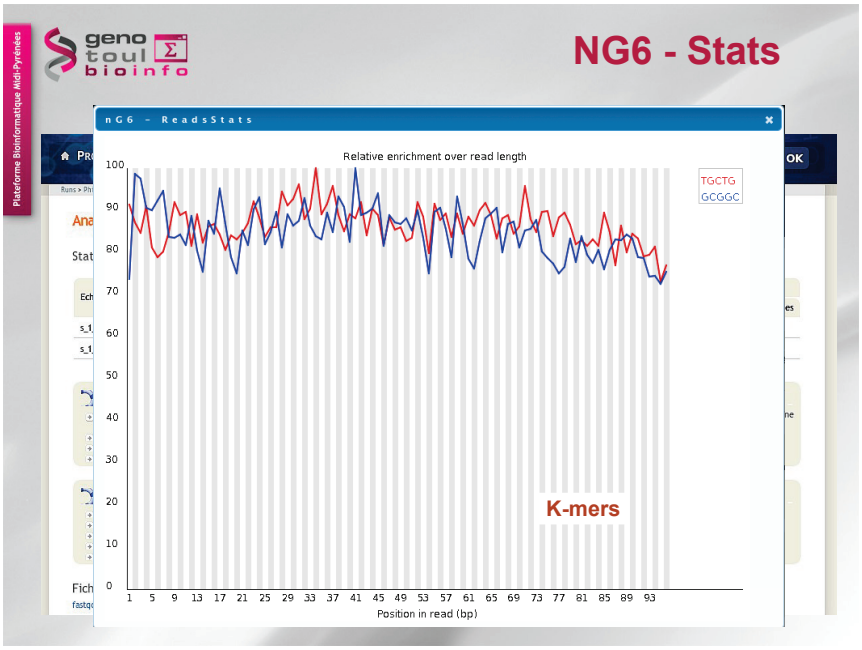
Aide Statistiques par séquence :

- Qualité : WARN = la qualité la plus observée < 27 (taux d'erreur de 0.2%). FAIL = la qualité la plus observée < 20 (taux d'erreur de 1%).
- GC% : WARN = plus de 15% des séquences ont un GC% différent de la distribution normale. FAIL = plus de 30% des séquences ont un GC% différent de la distribution normale.
- Longueur : WARN = toutes les séquences n'ont pas la même longueur. FAIL = une séquence a une longueur de 0pb.
- Niveau de duplication : WARN = plus de 20% des séquences sont non uniques. FAIL = plus de 50% des séquences sont non uniques.
- Séquences surreprésentées : WARN = une séquence représente plus de 0.1% du total. FAIL = une séquence représente plus de 1% du total.

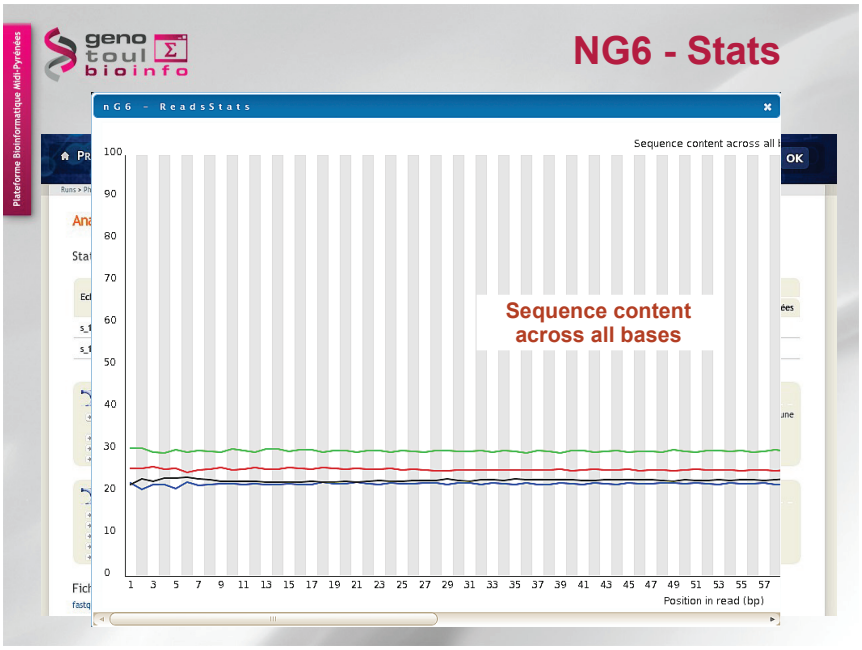
Fichiers résultats

fastqc.tar.gz

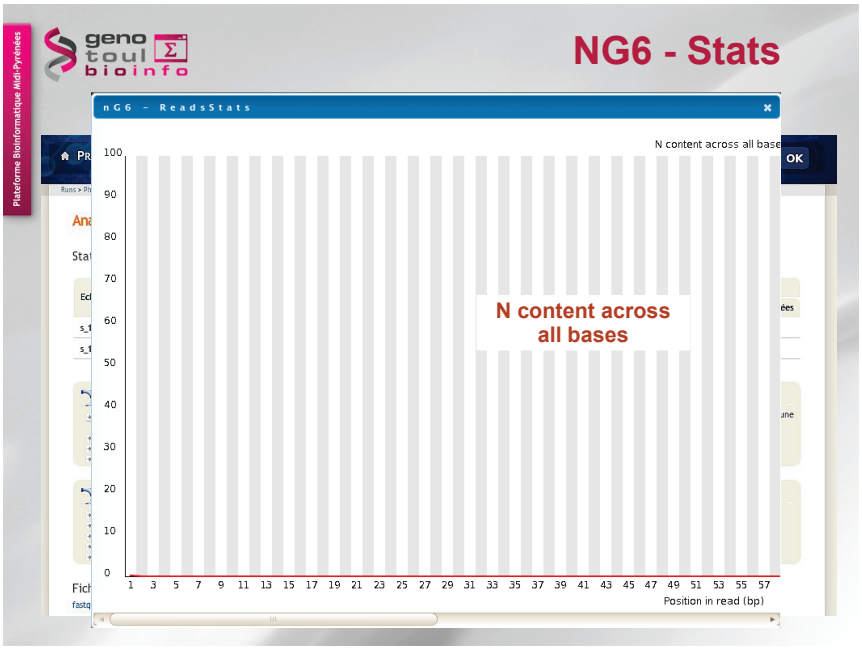
NG6 - Stats



NG6 - Stats



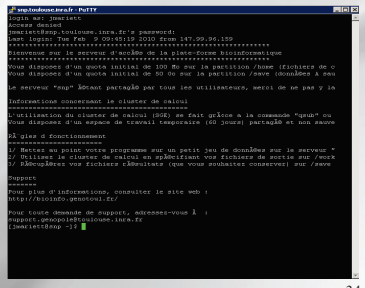
NG6 - Stats



Exercises / set 1

- snp server connexion
- Short read retrieval (wget)
- Read statistics (fastqc)

- Data sets :
  - SRX002048
  - ERR003037
  - ERR000017



Read alignment

- What are the ideas?
- The different software generations :
  - Smith-Waterman / Needleman-Wunch (1970)
  - BLAST (1990)
  - MAQ (2008)
  - BWA (2009)

BWA

- Much faster than MAQ
- Exact match
- Limited number of errors (2 for 32bp, 4 for 100 bp)

http://bio-bwa.sourceforge.net/

BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 14 2009, pages 1754–1760
doi:10.1093/bioinformatics/btp324

Sequence analysis

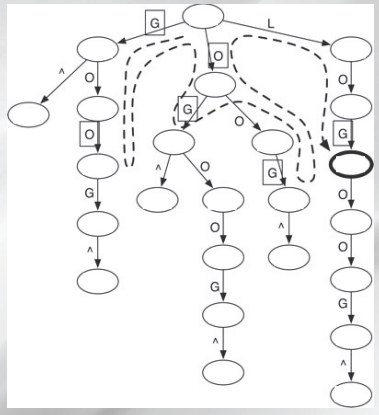
Fast and accurate short read alignment with Burrows–Wheeler transform

Heng Li and Richard Durbin*
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK

Received on February 20, 2009; revised on May 6, 2009; accepted on May 12, 2009
Advance Access publication May 18, 2009
Associate Editor: John Quackenbush

BWA prefix trie

- Word 'googol'
- ^ = start character
- --- Search of 'lol' with one error



The prefix trie is compressed to fit in memory in most cases (1Go for the human genome).

http://bio-bwa.sourceforge.net/bwa.shtml

Manual Reference Pages - bwa (1)

NAME

bwa - Burrows-Wheeler Alignment Tool

CONTENTS

SYNOPSIS

```
bwa index -a bwtsw database.fasta

bwa aln database.fasta short_read.fastq > aln_sa.sai

bwa samse database.fasta aln_sa.sai short_read.fastq > aln.sam

bwa sampe database.fasta aln_sa1.sai aln_sa2.sai read1.fq read2.fq > aln.sam

bwa bwasw database.fasta long_read.fastq > aln.sam
```


Commands

Reference sequence indexing :

```
bwa index -a bwtsw db.fasta
```

Read Alignment :

```
bwa aln db.fasta short_read.fastq > aln_sa.sai

bwa bwasw database.fasta long_read.fastq > aln.sam
```

Formatting unpaired reads :

```
bwa samse db.fasta aln_sa.sai short_read.fastq > aln.sam
```

Formatting pair ends :

```
bwa sampe database.fasta aln_sa1.sai aln_sa2.sai read1.fq
read2.fq > aln.sam
```

Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

Index

index bwa index [-p prefix] [-a algoType] [-c] <in.db.fasta>

Index database sequences in the FASTA format.

OPTIONS:

-c Build color-space index. The input fast should be in nucleotide space.

-p STR Prefix of the output database [same as db filename]

-a STR Algorithm for constructing BWT index. Available options are:

is IS linear-time algorithm for constructing suffix array. It requires 5.37N memory where N is the size of the database. IS is moderately fast, but does not work with database larger than 2GB. IS is the default algorithm due to its simplicity. The current codes for IS algorithm are reimplemented by Yuta Mori.

bwtsw Algorithm implemented in BWT-SW. This method works with the whole human genome, but it does not work with database smaller than 10MB and it is usually slower than IS.

Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

aln

aln bwa aln [-n maxDiff] [-o maxGapO] [-e maxGapE] [-d nDelTail] [-i nIndelEnd] [-k maxSeedDiff] [-l seedLen] [-t nThrds] [-cRN] [-M misMsc] [-O gapOsc] [-E gapEsc] [-q trimQual] <in.db.fasta> <in.query.fq> > <out.sai>

Find the SA coordinates of the input reads. Maximum *maxSeedDiff* differences are allowed in the first *seedLen* subsequence and maximum *maxDiff* differences are allowed in the whole sequence.

OPTIONS:

-n NUM Maximum edit distance if the value is INT, or the fraction of missing alignments given 2% uniform base error rate if FLOAT. In the latter case, the maximum edit distance is automatically chosen for different read lengths. [0.04]

-o INT Maximum number of gap opens [1]

-e INT Maximum number of gap extensions, -1 for k-difference mode (disallowing long gaps) [-1]

-d INT Disallow a long deletion within INT bp towards the 3'-end [16]

-i INT Disallow an indel within INT bp towards the ends [5]

-l INT Take the first INT subsequence as seed. If INT is larger than the query sequence, seeding will be disabled. For long reads, this option is typically ranged from 25 to 35 for '-k 2'. [inf]

-k INT Maximum edit distance in the seed [2]

-t INT Number of threads (multi-threading mode) [1]

-M INT Mismatch penalty. BWA will not search for suboptimal hits with a score lower than (bestScore-misMsc). [3]

-O INT Gap open penalty [11]

-E INT Gap extension penalty [4]

-R INT Proceed with suboptimal alignments if there are no more than INT equally best hits. This option only affects paired-end mapping. Increasing this threshold helps to improve the pairing accuracy at the cost of speed, especially for short reads (~32bp).

-c Reverse query but not complement it, which is required for alignment in the color space.

-N Disable iterative search. All hits with no more than *maxDiff* differences will be found. This mode is much slower than the default.

-q INT Parameter for read trimming. BWA trims a read down to argmax_x{\sum_{i=x+1}^l(INT-q_i)} if q_l<INT where l is the original read length. [0]

Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

samse & sampe

samse bwa samse [-n maxOcc] <in.db.fasta> <in.sai> <in.fq> > <out.sam>

Generate alignments in the SAM format given single-end reads. Repetitive hits will be randomly chosen.

OPTIONS:

-n INT Maximum number of alignments to output in the XA tag for reads paired properly. If a read has more than INT hits, the XA tag will not be written. [3]

sampe bwa sampe [-a maxInsSize] [-o maxOcc] [-n maxHitPaired] [-N maxHitDis] [-P] <in.db.fasta> <in1.sai> <in2.sai> <in1.fq> <in2.fq> > <out.sam>

Generate alignments in the SAM format given paired-end reads. Repetitive read pairs will be placed randomly.

OPTIONS:

-a INT Maximum insert size for a read pair to be considered being mapped properly. Since 0.4.5, this option is only used when there are not enough good alignment to infer the distribution of insert sizes. [500]

-o INT Maximum occurrences of a read for pairing. A read with more occurrneces will be treated as a single-end read. Reducing this parameter helps faster pairing. [100000]

-P Load the entire FM-index into memory to reduce disk operations (base-space reads only). With this option, at least 1.25N bytes of memory are required, where N is the length of the genome.

-n INT Maximum number of alignments to output in the XA tag for reads paired properly. If a read has more than INT hits, the XA tag will not be written. [3]

-N INT Maximum number of alignments to output in the XA tag for disconcordant read pairs (excluding singletons). If a read has more than INT hits, the XA tag will not be written. [10]

bwasw

```
bwasw    bwa bwasw [-a matchScore] [-b mmPen] [-q gapOpenPen] [-r gapExtPen] [-t nThreads]
         [-w bandWidth] [-T thres] [-s hspIntv] [-z zBest] [-N nHspRev] [-c thresCoef]
         <in.db.fasta> <in.fq>

         Align query sequences in the <in.fq> file.

         OPTIONS:

          -a INT   Score of a match [1]
          -b INT   Mismatch penalty [3]
          -q INT   Gap open penalty [5]
          -r INT   Gap extension penalty. The penalty for a contiguous gap of size k is
                   q+k*r. [2]
          -t INT   Number of threads in the multi-threading mode [1]
          -w INT   Band width in the banded alignment [33]
          -T INT   Minimum score threshold divided by a [37]
          -c FLOAT Coefficient for threshold adjustment according to query length. Given an
                   l-long query, the threshold for a hit to be retained is
                   a*max(T,c*log(l)). [5.5]
          -z INT   Z-best heuristics. Higher -z increases accuracy at the cost of speed.
                   [1]
          -s INT   Maximum SA interval size for initiating a seed. Higher -s increases
                   accuracy at the cost of speed. [3]
          -N INT   Minimum number of seeds supporting the resultant alignment to skip
                   reverse alignment. [5]
```


Exercises / set 2

- Retrieving the reference sequence in fasta format :
  - gi|224581838|ref|**NC_012125.1**| Salmonella enterica subsp. enterica serovar Paratyphi C strain RKS4594, complete genome
- Indexing the reference sequence
- Aligning the reads (fastq format)
- Formatting the alignment in SAM

Sequence Alignment/Map (SAM) format

- Data sharing was a major issue with the 1000 genomes
- Capture all of the critical information about NGS data in a single indexed and compressed file
- Sharing : data across and tools
- Generic alignment format
- Supports short and long reads (454 – Solexa – Solid)
- Flexible in style, compact in size, efficient in random access

Website :

http://samtools.sourceforge.net

Paper :

Li H.*, Handsaker B.*, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9. [PMID: 19505943]

Sequence Alignment/Map (SAM) format

Aligners natively generating SAM

- BFAST, `Blat-like Fast Accurate Search Tool' for Illumina and SOLiD reads.
- Bowtie. Highly efficient short read aligner. Natively support SAM output in recent version. A convertor is also available in samtools-C.
- BWA, Burrows-Wheeler Aligner for short and long reads.
- GEM library. Short read aligner. Convertor provided by the developers.
- Karma, the K-tuple Alignment with Rapid Matching Algorithm.
- Mosaik. The latest version support SAM output.
- Novoalign. An accurate aligner capable of gapped alignment for Illumina short reads. Academic free binary. Convertor is also available in samtools.
- SNP-o-matic, short read aligner and SNP caller.
- SOLiD BaseQV Tool. Developed by Applied Biosystems for converting SOLiD output files.
- SSAHA2 (since v2.4). Classical aligner for both short and long reads.
- Stampy, by Gerton Lunter. An accurate read aligner capable of gapped alignment for Illumina short reads. Used for indel discovery on the 1000 genomes data. Not released.
- TopHat for mapping short RNA-seq reads bridging exon junctions.

SAM format Header section

- Header lines start with @ followed by a two-letter TAG
- Header fields are TYPE:VALUE pairs

Type	Tag	Description
HD – header	VN*	File format version.
	SO	Sort order. Valid values are: *unsorted*, *queryname* or *coordinate*.
	GO	Group order (full sorting is not imposed in a group). Valid values are: *none*, *query* or *reference*.
SQ – Sequence dictionary	SN*	Sequence name. Unique among all sequence records in the file. The value of this field is used in alignment records.
	LN*	Sequence length.
	AS	Genome assembly identifier. Refers to the reference genome assembly in an unambiguous form. Example: HG18.
	M5	MD5 checksum of the sequence in the uppercase (gaps and space are removed)
	UR	URI of the sequence
	SP	Species.
RG – read group	ID*	Unique read group identifier. The value of the ID field is used in the RG tags of alignment records.
	SM*	Sample (use pool name where a pool is being sequenced)
	LB	Library
	DS	Description
	PU	Platform unit (e.g. lane for Illumina or slide for SOLiD); should be a full, unambiguous identifier
	PI	Predicted median insert size (maybe different from the actual median insert size)
	CN	Name of sequencing center producing the read.
	DT	Date the run was produced (ISO 8601 date or date/time).
	PL	Platform/technology used to produce the read.
PG – Program	ID*	Program name
	VN	Program version
	CL	Command line
CO – comment		One-line text comments

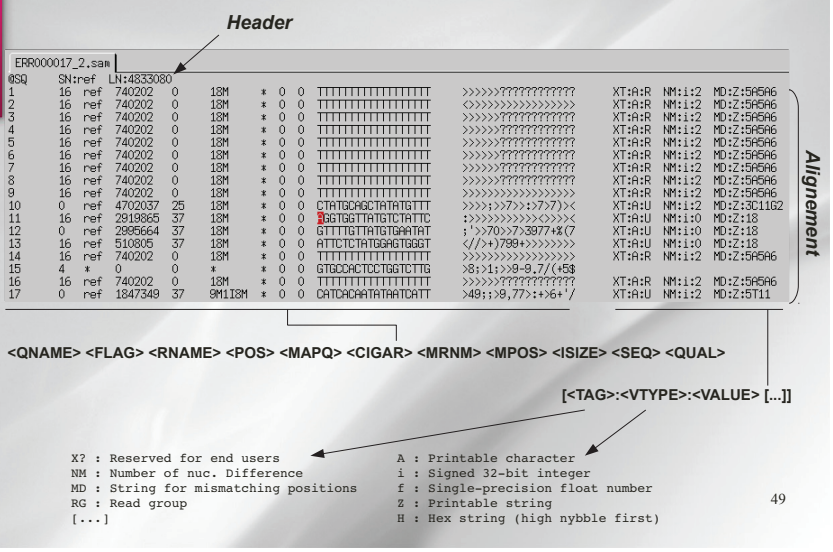
```
@HD     VN:1.0
@SQ     SN:chr20 LN:62435964
@RG     ID:L1 PU:SC_1_10 LB:SC_1 SM:NA12891
@RG     ID:L2 PU:SC_2_12 LB:SC_2 SM:NA12891
```

SAM format Alignment section

- 11 mandatory fields
- Variable number of optional fields
- Fields are tab delimited

1. **QNAME:** Query name of the read or the read pair
2. **FLAG:** Bitwise flag (pairing, strand, mate strand, etc.)
3. **RNAME:** Reference sequence name
4. **POS:** 1-Based leftmost position of clipped alignment
5. **MAPQ:** Mapping quality (Phred-scaled)
6. **CIGAR:** Extended CIGAR string (operations: MIDNSHP)
7. **MRNM:** Mate reference name ('=' if same as RNAME)
8. **MPOS:** 1-based leftmost mate position
9. **ISIZE:** Inferred insert size
10. **SEQQuery:** Sequence on the same strand as the reference
11. **QUAL:** Query quality (ASCII-33=Phred base quality)

SAM format
Full example



SAM format
Flag field

Flag	Description
0x0001	the read is paired in sequencing, no matter whether it is mapped in a pair
0x0002	the read is mapped in a proper pair (depends on the protocol, normally inferred during alignment) [1]
0x0004	the query sequence itself is unmapped
0x0008	the mate is unmapped [1]
0x0010	strand of the query (0 for forward; 1 for reverse strand)
0x0020	strand of the mate [1]
0x0040	the read is the first read in a pair [1,2]
0x0080	the read is the second read in a pair [1,2]
0x0100	the alignment is not primary (a read having split hits may have multiple primary alignment records)
0x0200	the read fails platform/vendor quality checks
0x0400	the read is either a PCR duplicate or an optical duplicate

http://picard.sourceforge.net/explain-flags.html

SAM format
Extended CIGAR format

```
Ref: GCATTCAGATGCAGTACGC
Read:  ccTCAG--GCATTAgtg
POS  CIGAR
5    2S4M2D6M3S
```

op	Description
M	Alignment match (can be a sequence match or mismatch)
I	Insertion to the reference
D	Deletion from the reference
N	Skipped region from the reference
S	Soft clip on the read (clipped sequence present in <seq>)
H	Hard clip on the read (clipped sequence NOT present in <seq>)
P	Padding (silent deletion from the padded reference sequence)

SAM format
Extended CIGAR format

P	Padding (silent deletion from the padded reference sequence)

```
  REF: CACGATCA**GACCGATACGTCCGA          REF: CACGATCA**GACCGATACGTCCGA
READ1:    CGATCAGAGACCGATA             READ1:    CGATCAGAGACCGATA
READ2:      ATCA*AGACCGATAC            READ2:      ATCAA*GACCGATAC
READ3:     GATCA**GACCG                READ3:     GATCA**GACCG
```

```
READ1: 6M2I8M                          READ1: 6M2I8M
READ2: 4M1P1I9M                        READ2: 4M1I1P9M
READ3: 5M2P5M                          READ3: 5M2P5M
```

N	Skipped region from the reference

```
  REF: AGCTAGCATCGTGTCGCCCGTCTAGCATACGCATGATCGACTGTCAGCTAGTCAGACTAGTCGATCGATGTG
 READ:            GTGTAACCC..............................TCAGAATA
```

where '...' on the read sequence indicates the intron. The CIGAR for this alignment is: `9M32N8M`.

BAM format

- Binary representation of SAM
- Compressed by BGZF library
- Greatly reduces storage space requirements to about 27% of original SAM

SAMtools

- Library and software package
- Creating sorted and indexed BAM files from SAM files
- Removing PCR duplicates
- Merging alignments
- Visualization of alignments from BAM files
- SNP calling
- Short indel detection

http://samtools.sourceforge.net/samtools.shtml

SAMtools
Example usage

```
inf212:~/Bureau/FormationNGS/JeuxDonnées> ../samtools-0.1.7a/samtools

Program: samtools (Tools for alignments in the SAM format)
Version: 0.1.7 (r510)

Usage:   samtools <command> [options]

Command: view        SAM<->BAM conversion
         sort        sort alignment file
         pileup      generate pileup output
         faidx       index/extract FASTA
         tview       text alignment viewer
         index       index alignment
         fixmate     fix mate information
         glfview     print GLFv3 file
         flagstat    simple stats
         calmd       recalculate MD/NM tags and
         merge       merge sorted alignments
         rmdup       remove PCR duplicates
```

```
inf212:~/Bureau/FormationNGS/JeuxDonnées> ../samtools-0.1.7a/samtools view

Usage:   samtools view [options] <in.bam>|<in.sam> [region1 [...]]

Options: -b       output BAM
         -h       print header for the SAM output
         -H       print header only (no alignments)
         -S       input is SAM
         -u       uncompressed BAM output (force -b)
         -x       output FLAG in HEX (samtools-C specific)
         -X       output FLAG in string (samtools-C specific)
         -t FILE  list of reference names and lengths (force -S) [null]
         -T FILE  reference sequence file (force -S) [null]
         -o FILE  output file name [stdout]
         -f INT   required flag, 0 for unset [0]
         -F INT   filtering flag, 0 for unset [0]
         -q INT   minimum mapping quality [0]
         -l STR   only output reads in library STR [null]
         -r STR   only output reads in read group STR [null]
         -?       longer help
```


SAMtools
Example usage

- Create BAM from SAM

```
samtools view -bS aln.sam -o aln.bam
```

- Sort BAM file

```
samtools sort example.bam sortedExample
```

- Merge sorted BAM files

```
samtools merge sortedMerge.bam sorted1.bam sorted2.bam
```

- Index BAM file

```
samtools index sortedExample.bam
```

- Visualize BAM file

```
samtools tview sortedExample.bam reference.fa
```


Picard

- A SAMtools complementary package
- More format conversion than SAMtools
- Visualization of alignments not available
- SNP calling & short indel detection not available

http://picard.sourceforge.net/

Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

Picard
Example usage

- ValidateSamFile
- SortSam
- MarkDuplicates
- EstimateLibraryComplexity
- MergeSamFiles
- ViewSam
- ReplaceSamHeader
- SamToFastq
- FastqToSam
- SamFormatConverter
- CreateSequenceDictionary
- CleanSam
- CompareSAMs

```
java -Xmx2g -jar PicardCmd.jar OPTION1=value1 OPTION2=value2...
```


Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

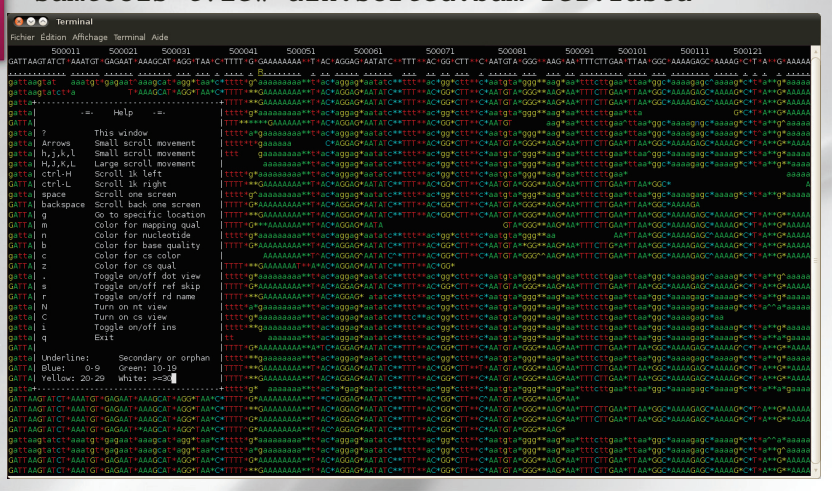
Exercise / set 3

Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

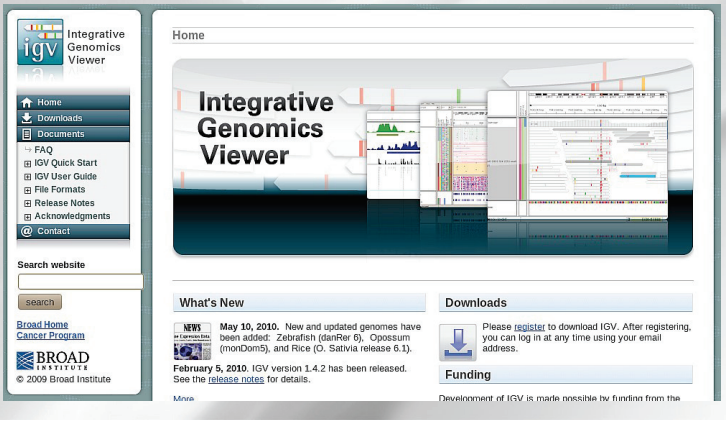
Visualizing the alignment
SAMtools : tview

```
samtools tview aln.sorted.bam ref.fasta
```



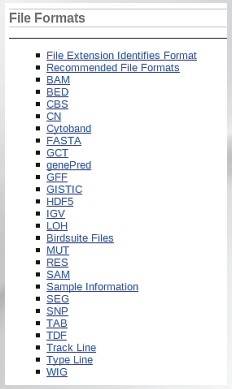
Visualizing the alignment IGV

- IGV : Integrative Genomics Viewer
- Website : http://www.broadinstitute.org/igv

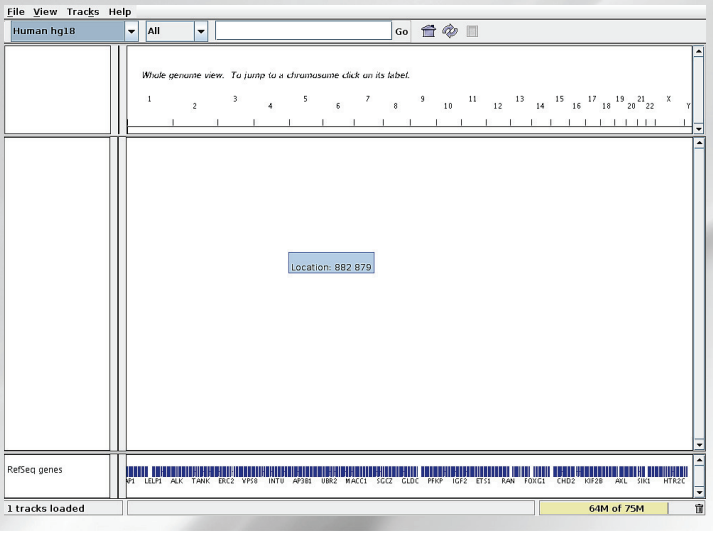


Visualizing the alignment IGV

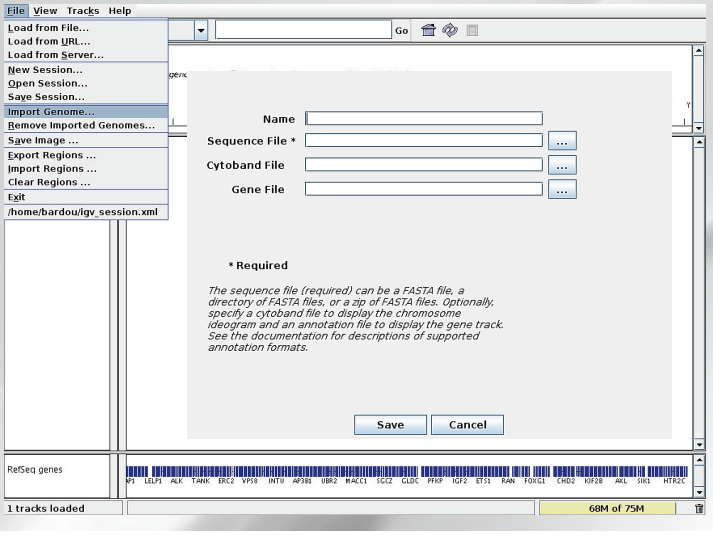
- High-performance visualization tool
- Interactive exploration of large, integrated datasets
- Supports a wide variety of data types
- Documentations
- Developed at the Broad Institute of MIT and Harvard



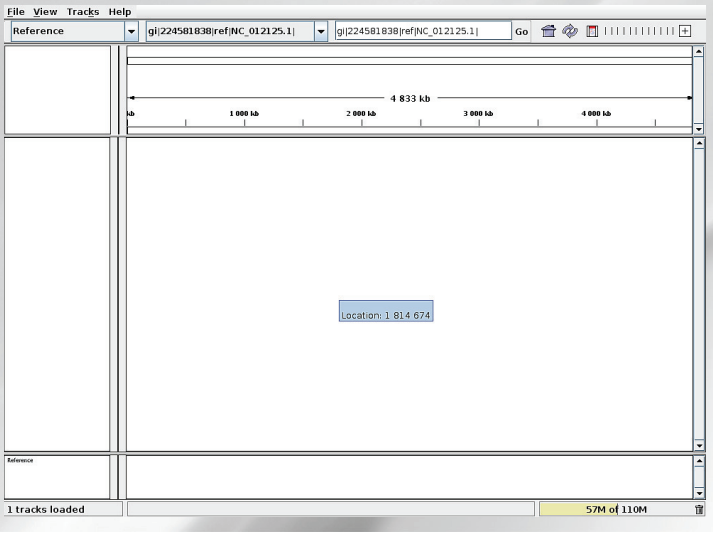
Visualizing the alignment IGV



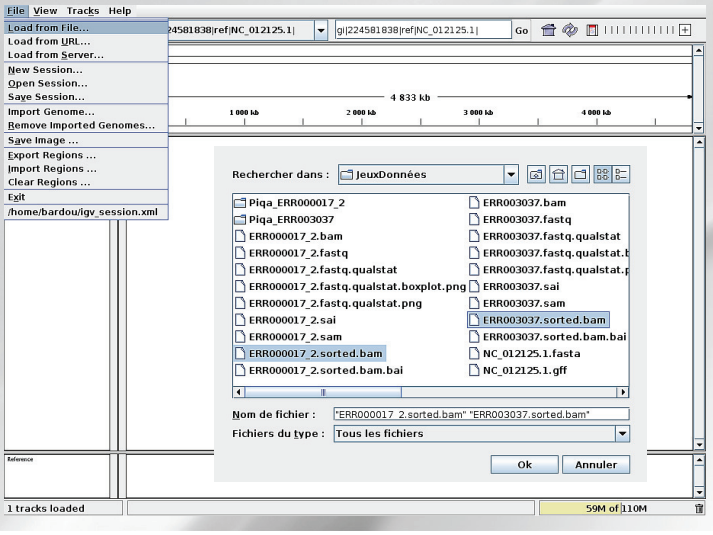
Visualizing the alignment
IGV - Loading the reference



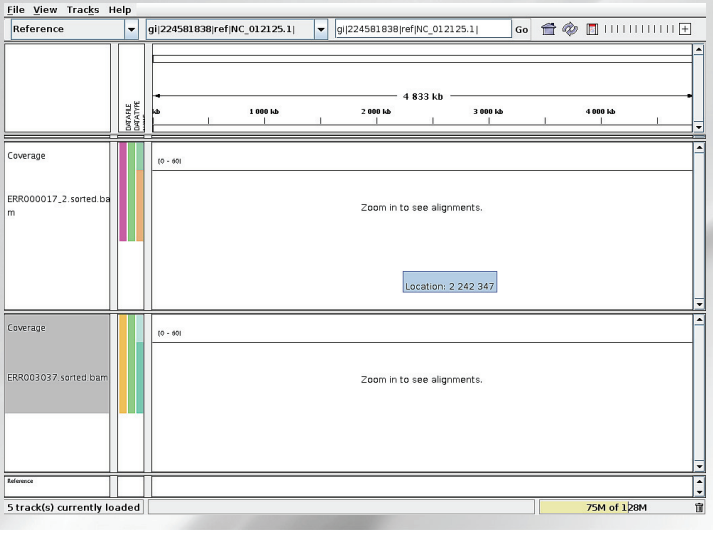
Visualizing the alignment
IGV - Loading the reference



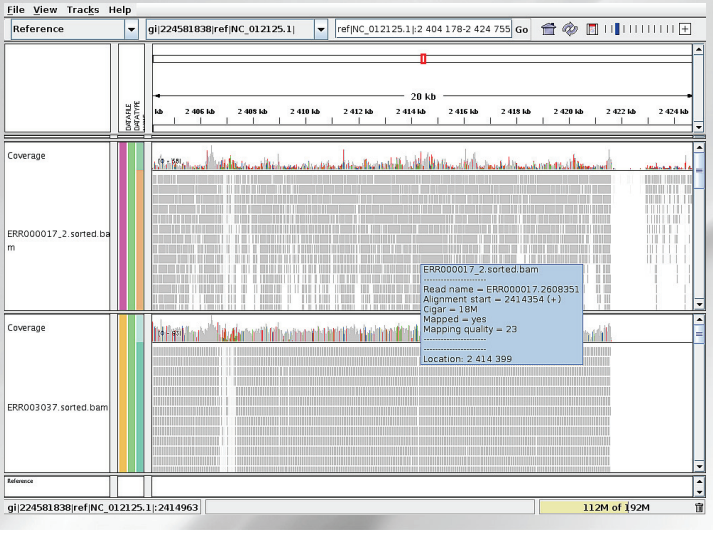
Visualizing the alignment
IGV - Loading the bam file



Visualizing the alignment
IGV - Loading the bam file



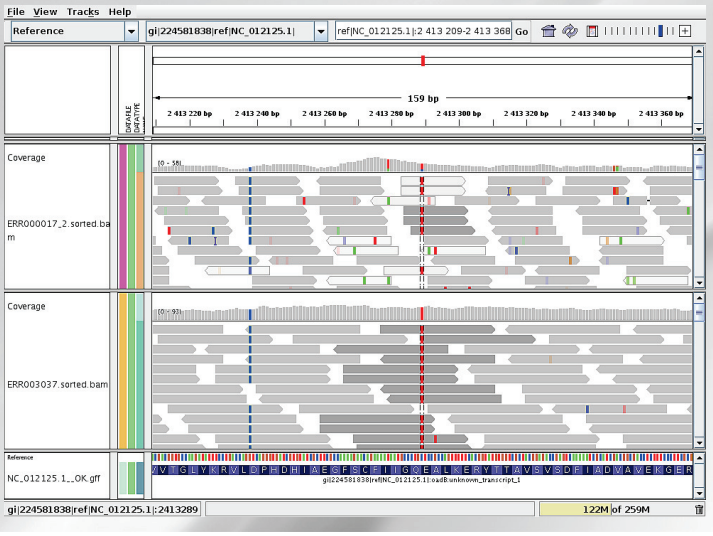
Visualizing the alignment
IGV - Zoom



Visualizing the alignment
IGV - Zoom

Visualizing the alignment
IGV - Loading a gff file

Visualizing the alignment
IGV - Loading a gff file



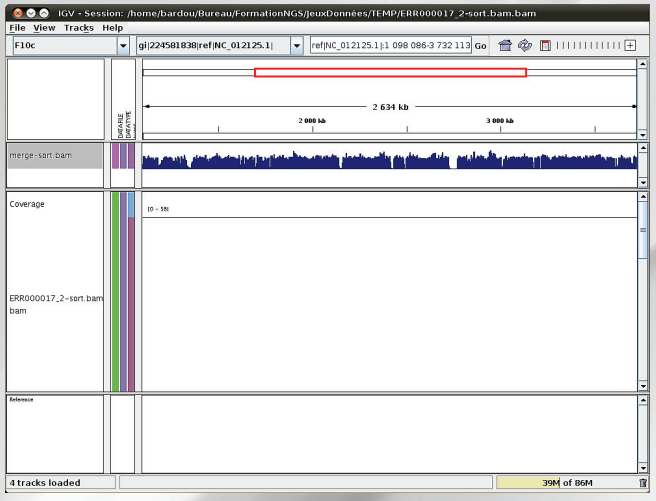
Visualizing the alignment
IGV - Coverage

- Generate the coverage information to be displayed in IGV.

```
java -jar igvtools.jar count aln.bam aln.depth.tdf ref.genome
```

- Remark : ref.genome was generated when we imported the genome sequence
- This step is optional, but it is essential if you want to see the read depth information in large scale.

Visualizing the alignment
IGV - Coverage



Exercise / set 4

The pileup format

Chr. - Coord. - Base('*' for indel) - Number of reads covering the site - **Read bases*** - Base qualities

```
seq1 272 T 24  ,.$.....,,,,,,,,,,,,,,,^+. <<<+;<<<<<<<<<<=<;<;7<&
seq1 273 T 23  ,.....,,,,,,,,,,,,,,,,A <<<;<<<<<<<<3<=<<<;<<+
seq1 274 T 23  ,.$.....,,,,,,,,,,,,,,,,     7<7;<;<<<<<<<<<=<;<;<<6
seq1 275 A 23  ,$.....,,,,,,,,,,,,,,,,^l.  <+;9*<<<<<<<<<=<<;;<<<<
seq1 276 G 22  ...T,,,,,,,,,,,,,,,,  33;+<<7=7<<7<&<<1;<<6<
seq1 277 T 22  ..........C.,,,,,,,G.  +7<;<<<<<<<&<=<<:;<<&<
seq1 278 G 23  .....,,,,,,,,,,,,,,,,^k.   %38*<<;<7<<7<=<<<;<<<<<
seq1 279 C 23  A..T,,,,,,,,,,,,,,,,  ;75&<<<<<<<<<=<<<9<<:<<

seq2 156 A 11  .$......+2AG.+2AG.+2AGGG     <975;:<<<<<
```

Read bases :

- **'.'** and **','** : match to the reference base on the forward/reverse strand
- **'ACTGN'** and **'actgn'** : for a mismatch on the forward/reverse strand
- **'^'** and **'$'** : start/end of a read segment
- **'+[0-9]+[ACGTNacgtn]+'** and **'-[0-9]+[ACGTNacgtn]+'** : insertion/deletion

http://samtools.sourceforge.net/pileup.shtml

Variant Calling with SAMtools

- Get the raw variant :

```
samtools pileup -vcf ref.fa aln.bam > raw.txt
samtools view -u aln.bam X | samtools pileup -vcf ref.fa - > raw-X.txt
```

ref.fa	Fasta formatted file of the reference genome
aln.bam	Sorted BAM formatted file, from the alignments
raw[-X].txt	Output pileup formatted, with consensus calls
-c	Calls the consensus base at each position
-v	Show positions that do not agree with ref.fa
-f	Reference sequence, ref.fa (in FastA format)

The pileup format

Consensus base - Consensus quality - Probability of difference from ref. base - Max. mapping quality

```
seq1  60  T  T  66  0  99  13  ..........^~.^~.   9<<55<;<<<<<<
seq1  61  G  G  72  0  99  15  ............^~.^y. (;975&;<<<<<<<<
seq1  62  T  T  72  0  99  15  .$.............     <;;,55;<<<<<<<<
seq1  63  G  G  72  0  99  15  .$............^~.  4;2;<7:+<<<<<<<
seq1  64  G  G  69  0  99  14  ............  9+5<;;;<<<<<<<
seq1  65  A  A  69  0  99  14  .$.............  <5-2<;;<<<<<<;
seq1  66  C  C  66  0  99  13  ............  &*<;;<<<<<<8<
seq1  67  C  C  69  0  99  14  ............^~.    ,75<.4<<<<<-<<
seq1  68  C  C  69  0  99  14  ............  576<;7<<<<<8<<

seq2  156 A  A  10  0  99  11  .$.....+2AG.+2AG.+2AGGG <975;:<<<<<
seq2  156 *  +AG/+AG  71  252  99  11  +AG  *  3  8  0
```

1st indel allele - 2nd indel allele - Reads supporting 1st - Reads supporting 2nd - Reads supporting 3rdc

~~Reads bases~~ ~~Reads qualities~~

Filter - samtools.pl

- Filter the raw variant calls :

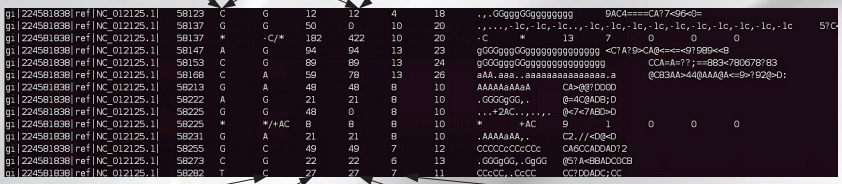
```
samtools.pl varFilter raw.txt > raw_ok.txt

(samtools.pl varFilter -p raw.txt > raw_ok.txt) >& raw_filtered.txt
```

```
Usage:   samtools.pl varFilter [options] <in.cns-pileup>

Options: -Q INT    minimum RMS mapping quality for SNPs [25]
         -q INT    minimum RMS mapping quality for gaps [10]
         -d INT    minimum read depth [3]
         -D INT    maximum read depth [100]

         -G INT    min indel score for nearby SNP filtering [25]
         -w INT    SNP within INT bp around a gap to be filtered [10]

         -W INT    window size for filtering dense SNPs [10]
         -N INT    max number of SNPs in a window [2]

         -l INT    window size for filtering adjacent gaps [30]

         -p        print filtered variants
```

Filter - awk

Acquire final variant calls by setting a quality threshold (50 for indels and 20 for substitutions) :

```
awk '($3=="*"&&$6>=50)||($3!="*"&&$6>=20)' raw.flt.txt > raw.final.txt
```



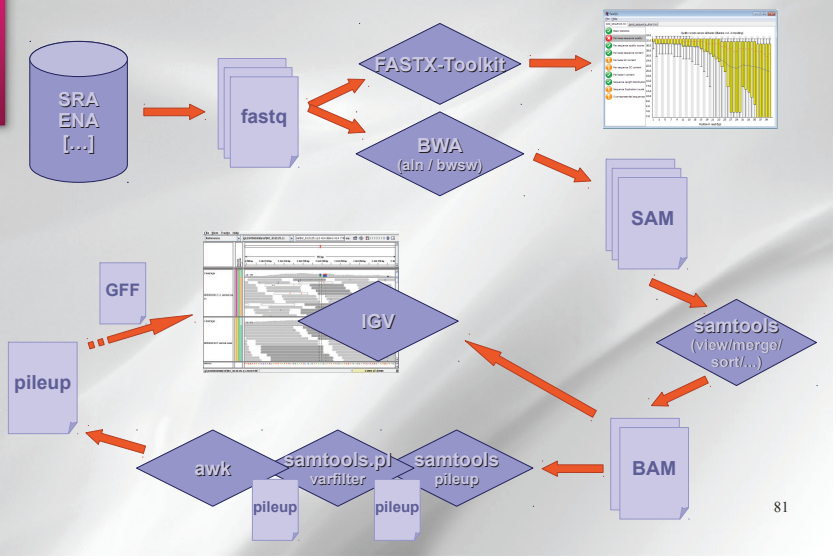
Consensus base - Consensus quality - Probability of difference from ref. base - Max. mapping quality

Consensus - samtools.pl

Consensus : A way of representing the results of a multiple sequence alignment (which residues are most abundant in the alignment at each position).

```
samtools.pl pileup2fq raw.txt > raw.fq
```


Synthesis



Exercise / set 5

The END

Satisfaction form :

http://bioinfo.genotoul.fr/index.php?id=60

Exam :

http://bioinfo.genotoul.fr/index.php?id=122