IX International Course of Massive Data Analysis FOR GENOMICS

EMBL-EBI

ciberer
CENTRO DE INVESTIGACIÓN BIOMÉDICA EN RED
DE ENFERMEDADES RARAS

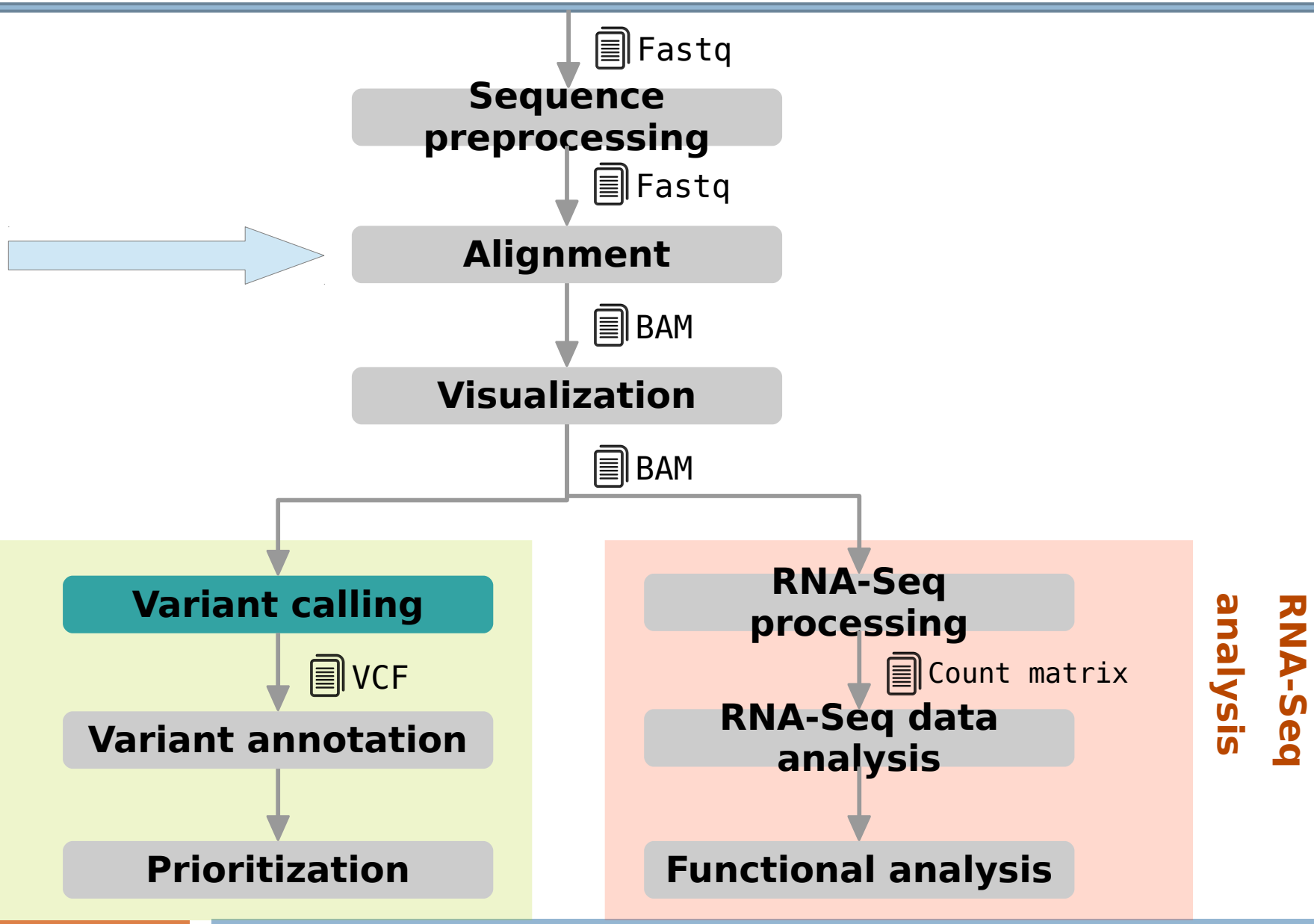
PRINCIPE FELIPE
CENTRO DE INVESTIGACION

INB

Ignacio Medina
imedina@ebi.ac.uk

Mapping NGS reads for genomic studies

Where are we?



Ignacio Medina
imedina@ebi.ac.uk

Mapping NGS reads for genomic studies

Index

- Introduction
- Algorithms and Tools
- HPG Aligner
- SAM/BAM specification
- Best practices
- Data repositories
- Hands on
- QC alignment

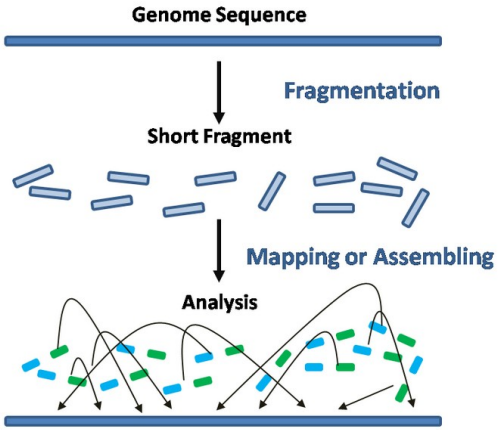
Introduction
The NGS data, some numbers and features

- Current read sizes ranging from 100-800bp, up to 15kb coming soon
- Single-end and paired-end reads
- Sequencing errors, low quality reads, duplicated reads
- Analysis pipelines: Exome vs Genome sequencing, RNA-seq (transcriptomics), germline vs somatic, BS-seq, ChIP-seq, ...
- Illumina **HiSeq 2500** provides high-quality 2x125bp: 176Gb in 40h, 90.2% bases above Q30
  - Human genome 3Gb ~ 60x coverage
  - Each sample produces a *fastq* file ~500GB size containing ~550M reads
- New **Illumina X Ten**: Consists if ten ultra-high-throughput sequencers. First $1000 human genome sequencer. Produces 18.000 genomes per year
- Mapping goes from FASTQ to SAM/BAM files

Ignacio Medina
imedina@ebi.ac.uk

Introduction
Aligning reads, the challenges

- Mapping reads onto a **reference genome**, a simple concept but there are some ***challenges***:
  - *Natural variability*: SNPs, *de novo* mutations, INDELS, copy number, translocations, …
  - *Repetitive regions*
  - *Sequencing errors*
  - *RNA-seq*: gapped alignment
  - *BS-seq*: C → T conversion strategy
  - *High computing resources needed: multicore CPUs and a lot of RAM*
- We must deal with genomic variation an efficient way



Simple idea, but with some challenges

ARTICLE

doi:10.1038/nature11632

An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium*

By characterizing the geographic and functional spectrum of human genetic variation, the 1000 Genomes Project aims to build a resource to help to understand the genetic contribution to disease. Here we describe the genomes of 1,092 individuals from 14 populations, constructed using a combination of low-coverage whole-genome and exome sequencing. By developing methods to integrate information across several algorithms and diverse data sources, we provide a validated haplotype map of 38 million single nucleotide polymorphisms, 1.4 million short insertions and deletions, and more than 14,000 larger deletions. We show that individuals from different populations carry different profiles of rare and common variants, and that low-frequency variants show substantial geographic differentiation, which is further increased by the action of purifying selection. We show that evolutionary conservation and coding consequence are key determinants of the strength of purifying selection, that rare-variant load varies substantially across biological pathways, and that each individual contains hundreds of rare non-coding variants at conserved sites, such as motif-disrupting changes in transcription-factor-binding sites. This resource, which captures up to 98% of accessible single nucleotide polymorphisms at a frequency of 1% in related populations, enables analysis of common and low-frequency variants in individuals from diverse, including admixed, populations.

Ignacio Medina
imedina@ebi.ac.uk

Introduction
Getting a reference genome

- A ***reference genome*** is a consensus sequence built up from high quality sequencing samples from different populations. It is the control reference sequence to compare our samples
- ***Genome Reference Consortium (GRC)*** created to deliver assemblies:
  - http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/
- Current human assembly is ***GRCh38***, released in the summer of 2014. Major projects are considering to use it.
- Reference genomes can be downloaded from:
  - ***GRC***: Human genome available at: ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/Homo_sapiens/GRCh37/Primary_Assembly/assembled_chromosomes/FASTA/
  - ***Ensembl***: many available vertebrates genomes http://www.ensembl.org/info/data/ftp/index.html
  - ***Ensembl Genomes***: http://ensemblgenomes.org/

Introduction
NGS in clinics, proof of concept

Published in final edited form as:
Nat Genet. 2010 January ; 42(1): 30–35. doi:10.1038/ng.499.

Exome sequencing identifies the cause of a Mendelian disorder

Sarah B. Ng[1,*], **Kati J. Buckingham**[2,*], **Choli Lee**[1], **Abigail W. Bigham**[2], **Holly K. Tabor**[2], **Karin M. Dent**[3], **Chad D. Huff**[4], **Paul T. Shannon**[5], **Ethylin Wang Jabs**[6,7], **Deborah A. Nickerson**[1], **Jay Shendure**[1,†], and **Michael J. Bamshad**[1,2,8,†]

[1]Department of Genome Sciences, University of Washington, Seattle, Washington, USA [2]Department of Pediatrics, University of Washington, Seattle, Washington, USA [3]Department of Pediatrics, University of Utah, Salt Lake City, Utah, USA [4]Department of Human Genetics, University of Utah, Salt Lake City, Utah, USA [5]Institute of Systems Biology, Seattle WA, USA [6]Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, New York, USA [7]Department of Pediatrics, Johns Hopkins University, Baltimore, Maryland [8]Seattle Children's Hospital, Seattle, Washington, USA

Abstract

We demonstrate the first successful application of exome sequencing to discover the gene for a rare, Mendelian disorder of unknown cause, Miller syndrome (OMIM %263750). For four affected individuals in three independent kindreds, we captured and sequenced coding regions to a mean coverage of 40X, and sufficient depth to call variants at ~97% of each targeted exome. Filtering against public SNP databases and a small number of HapMap exomes for genes with two novel variants in each of the four cases identified a single candidate gene, *DHODH*, which encodes a key enzyme in the pyrimidine *de novo* biosynthesis pathway. Sanger sequencing confirmed the presence of *DHODH* mutations in three additional families with Miller syndrome. Exome sequencing of a small number of unrelated, affected individuals is a powerful, efficient strategy for identifying the genes underlying rare Mendelian disorders and will likely transform the genetic analysis of monogenic traits.

Genetic Mapping and Exome Sequencing Identify Variants Associated with Five Novel Diseases

Article | Metrics | Related Content | Comments: 0

Erik G. Puffenberger[1,2*], **Robert N. Jinks**[2], **Carrie Sougnez**[3], **Kristian Cibulskis**[3], **Rebecca A. Willert**[2], **Nathan P. Achilly**[2], **Ryan P. Cassidy**[2], **Christopher J. Fiorentini**[2], **Kory F. Heiken**[2], **Johnny J. Lawrence**[2], **Molly H. Mahoney**[2], **Christopher J. Miller**[2], **Devika T. Nair**[2], **Kristin A. Politi**[2], **Kimberly N. Worcester**[2], **Roni A. Setton**[2], **Rosa DiPiazza**[2], **Eric A. Sherman**[4], **James T. Eastman**[5], **Christopher Francklyn**[6], **Susan Robey-Bond**[6], **Nicholas L. Rider**[1,2,7], **Stacey Gabriel**[3], **D. Holmes Morton**[1,2,7], **Kevin A. Strauss**[1,2,7]

1 Clinic for Special Children, Strasburg, Pennsylvania, United States of America, **2** Department of Biology and Biological Foundations of Behavior Program, Franklin & Marshall College, Lancaster, Pennsylvania, United States of America, **3** The Broad Institute, Boston, Massachusetts, United States of America, **4** Department of Biology, Swarthmore College, Swarthmore, Pennsylvania, United States of America, **5** Department of Pathology and Laboratory Medicine, School of Medicine and Public Health, University of Wisconsin, Madison, Wisconsin, United States of America, **6** College of Medicine, University of Vermont, Burlington, Vermont, United States of America, **7** Lancaster General Hospital, Lancaster, Pennsylvania, United States of America

To **add a note**, highlight some text. Hide notes
Make a general comment

Jump to
Abstract
Introduction
Results
Discussion
Materials and Methods
Acknowledgments
Author Contributions
References

View All Figures

Abstract Top

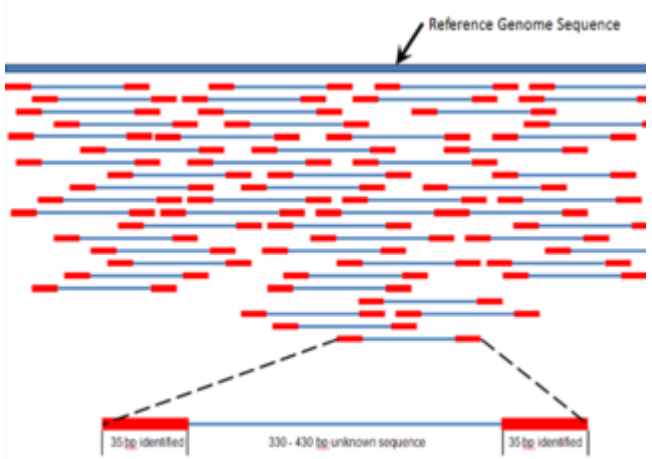
The Clinic for Special Children (CSC) has integrated biochemical and molecular methods into a rural pediatric practice serving Old Order Amish and Mennonite (Plain) children. Among the Plain people, we have used single nucleotide polymorphism (SNP) microarrays to genetically map recessive disorders to large autozygous haplotype blocks (mean = 4.4 Mb) that contain many genes (mean = 79). For some, uninformative mapping or large gene lists preclude disease-gene identification by Sanger sequencing. Seven such conditions were selected for exome sequencing at the Broad Institute; all had been previously mapped at the CSC using low density SNP microarrays coupled with autozygosity and linkage analyses. Using between 1 and 5 patient samples per disorder, we identified sequence variants in the known disease-causing genes *SLC6A3* and *FLVCR1*, and present evidence to strongly support the pathogenicity of variants identified in *TUBGCP6*, *BRAT1*, *SNIP1*, *CRADD*, and *HARS*. Our results reveal the power of coupling new genotyping technologies to population-specific genetic knowledge and robust clinical data.

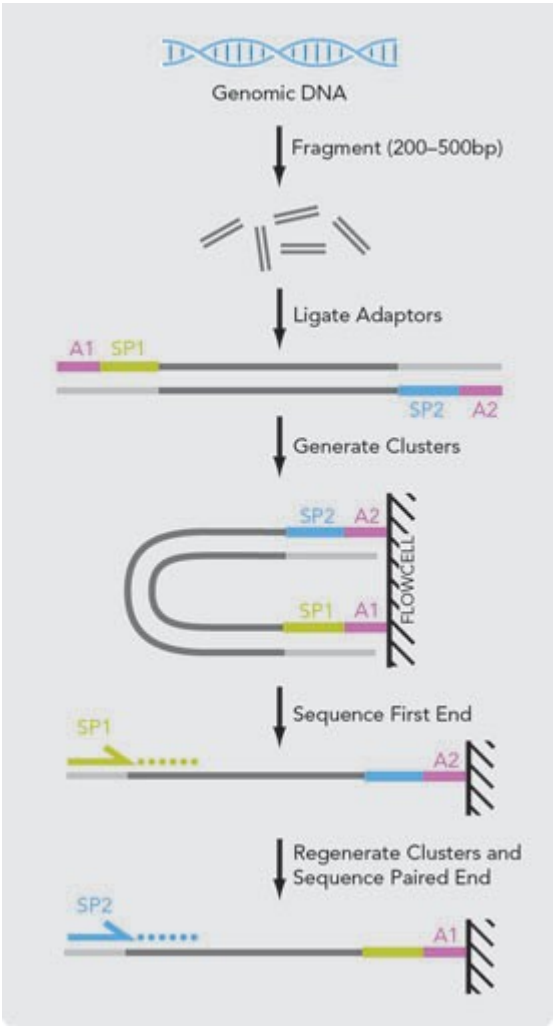
Ignacio Medina
imedina@ebi.ac.uk

Mapping NGS reads for genomic studies

Introduction
The mapping process considerations

- Considerations:
  - Which tool to use? What am I looking for? SNVs? INDELS? Long reads?
  - Is it DNA or RNA?
  - Single-end or paired-end? Paired-end when:
    - For very short reads, reduce the number of false positives alignments
    - Re-sequencing projects, Rna-seq?
    - Am I interested in Structural variation or gene fusions?
    - Reduce number of false positive variants
  - Should I allow multiple hits?
  - Should I remove low quality reads?
- In general for *genomic variant analysis* we need high quality reads, paired-end datasets work better, and **no** multiple hits must be allowed





Taken from Illumina

Ignacio Medina
imedina@ebi.ac.uk

Algorithms and tools
Desirable features of a aligner

- Goals
  - **Sensitivity**, we are looking for genomic variants, reads with mismatches and INDELS must be properly aligned
  - **Specificity**, no wrong alignments should be provided
  - Being able to perform gapped alignments (RNA), exons must be correctly located
  - Good performance, efficiency matters
  - Easy to use
  - Open-source and maintained
  - Capable of align different data types: DNA, RNA-seq, BS-seq, …
- Unfortunately... most tools or algorithms only work well in a specific scenario
- New project called *High-Performance Genomics* (HPG) that is part of the OpenCB initiative tries to solve this

Algorithms and tools
Smith-Waterman (SW) algorithm

SW finds the optimal local alignment between:

Sequence 1 = ACACACTA

Sequence 2 = AGCACACA

Given gap-scoring penalties:

w(match) = +2

w(a,-) = w(-,b) = w(mismatch) = -1

$$H = \begin{pmatrix} & - & A & C & A & C & A & C & T & A \\ - & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & 0 & 2 & 1 & 2 & 1 & 2 & 1 & 0 & 2 \\ G & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ C & 0 & 0 & 3 & 2 & 3 & 2 & 3 & 2 & 1 \\ A & 0 & 2 & 2 & 5 & 4 & 5 & 4 & 3 & 4 \\ C & 0 & 1 & 4 & 4 & 7 & 6 & 7 & 6 & 5 \\ A & 0 & 2 & 3 & 6 & 6 & 9 & 8 & 7 & 8 \\ C & 0 & 1 & 4 & 5 & 8 & 8 & 11 & 10 & 9 \\ A & 0 & 2 & 3 & 6 & 7 & 10 & 10 & 10 & 12 \end{pmatrix}$$

Alignment result:

Sequence 1 = A - C A C A C T A

Sequence 2 = A G C A C A C - A

- Very popular algorithm developed in 1981
- Provides a very **high sensitivity,** allowing alignments with any number of mismatches, insertions and deletions
- Gives an *optimal alignment* between two sequences given a penalties, **it is not a mapper but an sequence aligner**
- No suitable for whole genome alignment: for a 100bp read and the human genome 3Gb, the matrix dimension: 100×3.10^9, using 4 Bytes for integers: **1.2TB of RAM !!**
- Although *dynamic programming* techniques are applied to make SW more efficient, the CPU requirements are still too high, **SW is too slow for NGS**

Ignacio Medina
imedina@ebi.ac.uk

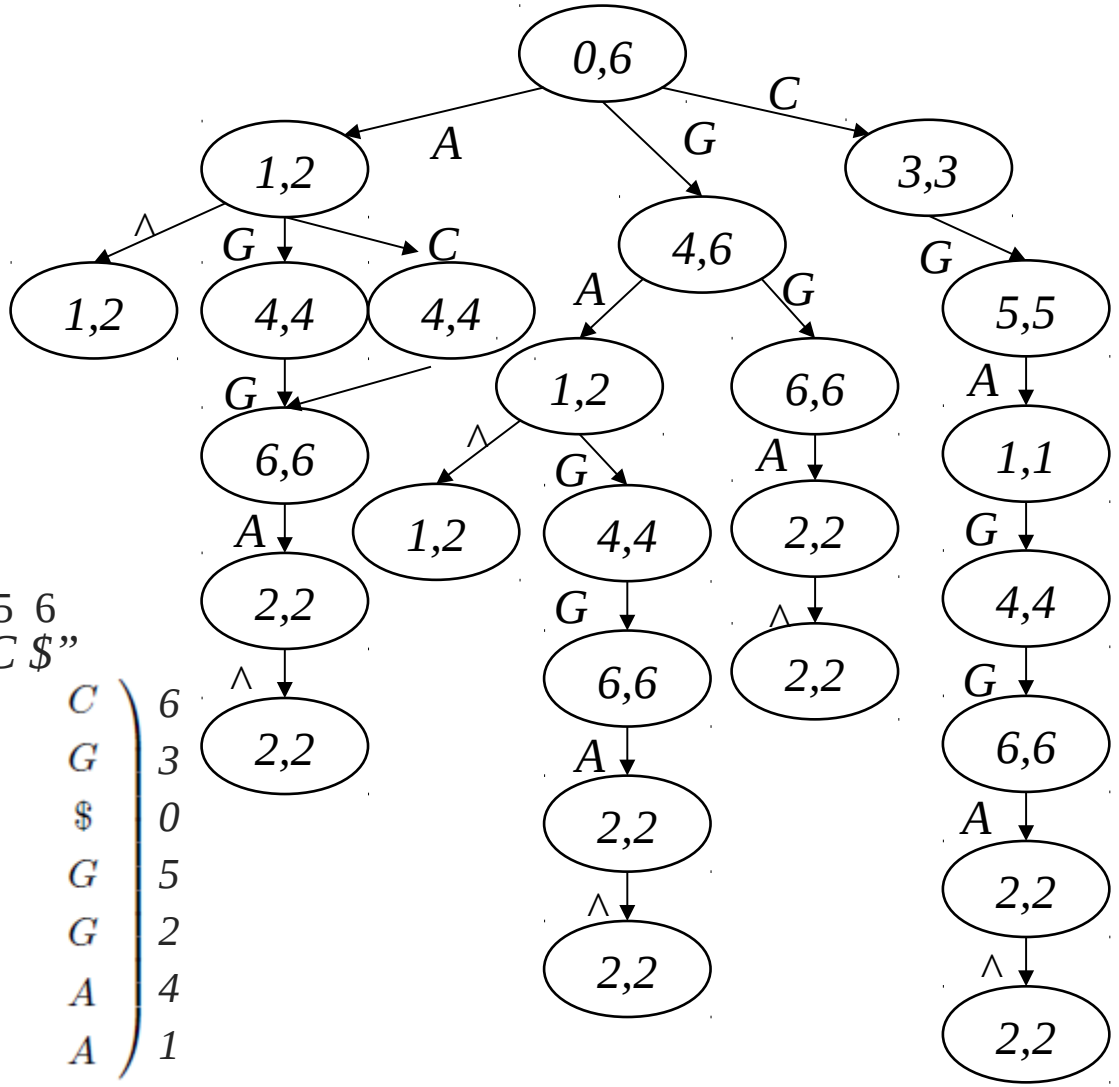
Algorithms and tools
BLAST, Basic Local Alignment Search Tool

- BLAST is one of the most widely used programs in Bioinformatics developed in 1990 at NIH. Allows comparing and searching amino-acid and DNA sequences in a database of sequences
- BLAST uses a heuristic algorithm to speed-up searches, it is **much faster** than calculating an optimal alignment with Smith-Waterman, **but it cannot guarantee the optimal alignment** of the query sequence in the database. It searches the most relevant *seeds* from query sequence in exact way and then SW is applied
- It presents a **high sensitivity,** allowing alignments with any number of mismatches, insertions and deletions, it can be used to align sequence between species
- However, it is **still too slow** for NGS mapping, blast can align few thousands sequences per hour

Ignacio Medina
imedina@ebi.ac.uk

Algorithms and tools
Burrows-Wheeler Transform (BWT) algorithm

- BWT is an algorithm used in data compression techniques such as *bzip2*
- It **efficiently** align short sequencing reads against a large reference sequence such as the human genome, a **prefix tree index** is created using reference genome
- In the transformation all permutations are sorted and all suffixes are grouped
- It is **much faster** than BLAST, it can align hundred of thousands sequences per second!
- However, it presents a **lower sensitivity**, it can allow a few mismatches, and in some implementation one INDEL



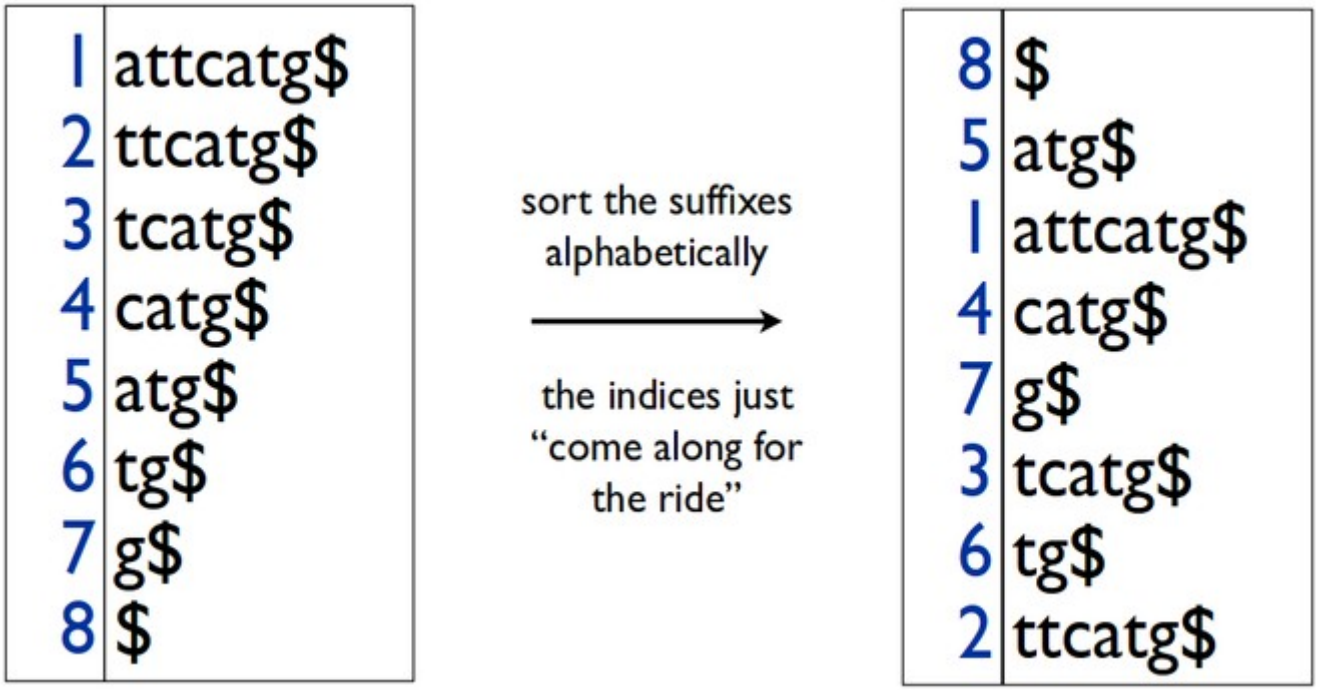
R = "A G G A G C $"

	0	1	2	3	4	5	6	
0	$	A	G	G	A	G	C	6
1	A	G	C	$	A	G	G	3
2	A	G	G	A	G	C	$	0
3	C	$	A	G	G	A	G	5
4	G	A	G	C	$	A	G	2
5	G	C	$	A	G	G	A	4
6	G	G	A	G	C	$	A	1

Ignacio Medina
imedina@ebi.ac.uk

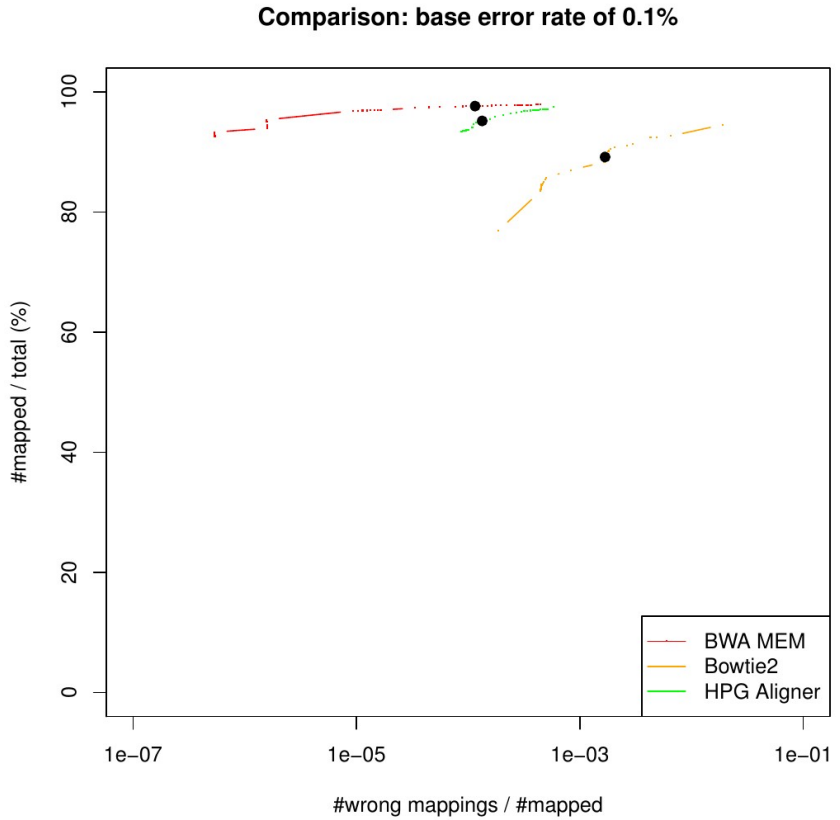
Algorithms and tools
Suffix Arrays (SA) algorithm

- SA uses more memory than other related compressed data structures like BWT
- It **very efficiently** aligns sequencing reads against a large reference sequence such as the human genome, a **suffix array index** is created using reference genome
- Since it is not compressed the **performance** achieved is higher than with BWT, it can align million sequences per second in modern CPUs!
- Depending on the alignment strategy, a **high sensitivity** can be achieved, it can allow a relative high number of mismatches, and in some implementations several INDELs



Ignacio Medina
imedina@ebi.ac.uk

Algorithms and tools
Many aligners available, which to use?



- Many aligners available, more than 70!!
  - http://wwwdev.ebi.ac.uk/fg/hts_mappers/
- Can be difficult to select one, some criteria
  - Type of analysis: dna, rna, meth
  - Number of cites
  - ...
- **Selecting an aligner**: simulate datasets to choose the best:
  - Which one is more sensitive to INDELS?
  - Which produce less false positives alignments
  - Which RNA aligner works better with low coverage?
  - …
- All of them work similarly
  - **Reference genome index**: this index can be a Burrows-Wheeler Transform (BWT), Suffix array (SA), …
  - The reads are **aligned to that index or are split in seeds an then aligned**, seeds aligned are clustered together
  - In general poor performance when high number of mismatches or INDELS are present

Ignacio Medina
imedina@ebi.ac.uk

Algorithms and tools
DNA: BWA, BWA-SW and BWA-MEM

- BWA stands from Burrows-Wheeler Aligner, developed by R. Durbin at Sanger Institute
  - http://bio-bwa.sourceforge.net/
- It was one of the first NGS mappers and is widely used, provides very good results in common scenarios
- It implements BWT and Suffix Arrays (SA) with support for few errors:
  - *BWA-SW and BWA-MEM both tolerate more errors given longer alignment. Simulation suggests that they may work well given 2% error for an 100bp alignment, 3% error for a 200bp, 5% for 500bp and 10% for 1000bp or longer alignment*
- Implementation is in C and it is multi-thread, but lacks some features such as support for RNA-seq or big INDELS
- Not designed to take advantage of new technologies and clusters, not specially fast

Algorithms and tools
DNA: Bowtie and Bowtie2

- Bowtie allowed a few mismatches (<3) and no gaps, claimed to be the fastest, but it missed many reads
- Bowtie2 improved sensitivity when compared to Bowtie:
  - http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
- Widely used, however it is a little bit less sensitivity than BWA, fail to correctly map many mismatches and INDELS
- Implementation is in C and it is multi-thread, but lacks some biological features such as support for RNA or big INDELS
- Not designed to take advantage of new technologies and clusters

Algorithms and tools
RNA-seq: TopHat, the standard RNA-seq aligner

- TopHat is the standard for RNA-seq mapping
  - http://tophat.cbcb.umd.edu/
- It uses Bowtie2 to align reads, so it is not very sensitive, usually maps 75% of reads
- Not ready for long reads (>150bp), mapping decrease to below 50%
- Poor performance, can take several hours to map
- Big memory footprint and a lot of disk used
- Mapping fall down with mismatches, INDELS and longer reads
- Written in Python and C. Not designed to take advantage of new technologies and clusters

Algorithms and tools
RNA-seq: STAR and MapSplice

- STAR developed for ENCODE project
  - https://code.google.com/p/rna-star/
  - High-performance, not very high sensitivity
- MapSplice
  - http://www.netlab.uky.edu/p/bioinfo/MapSplice2
  - Not bad sensitivity but very slow

Algorithms and tools
Meth: Bismark, a BS-seq mapper

- Bismark can map BS-seq data:
  - http://www.bioinformatics.babraham.ac.uk/projects/bismark/
- It uses Bowtie2 for mapping
- Sensitivity and performance very poor
- Written in Perl and Python. Not designed to take advantage of new technologies and clusters

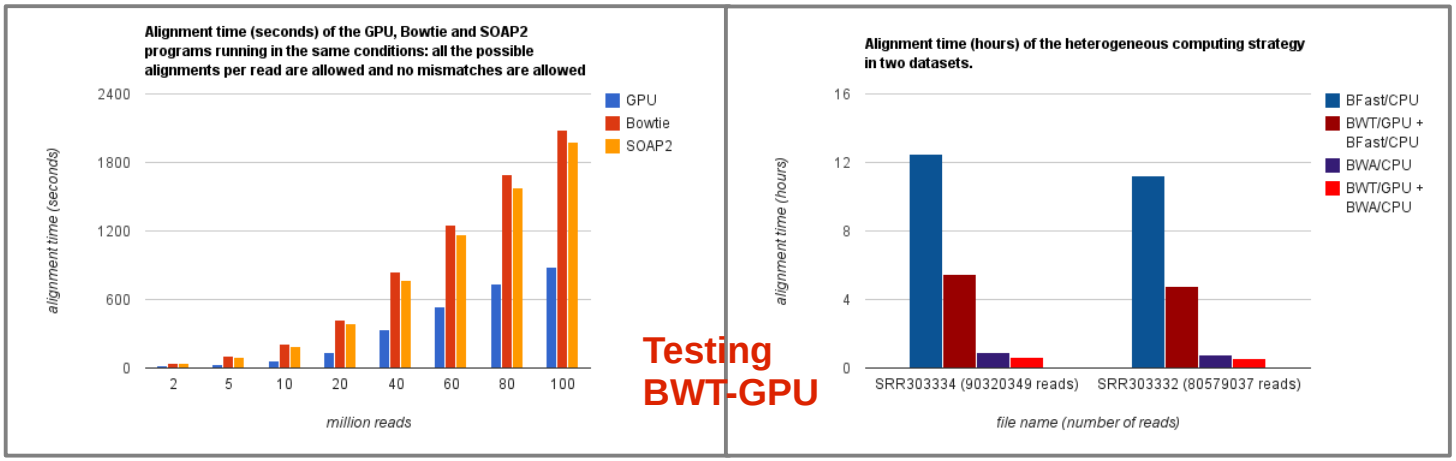
HPG Aligner
Why another NGS read mapper, motivation

- Bioinformatics needs
  - Reads sizes are increasing, first mappers were designed to 50bp
  - More sensitivity is needed: more variability and indels than expected
  - Genomic rearrangements: copy number, translocations, …
  - INDELS realignment, mapping recalibration
  - One tool to rule them all: DNA, RNA-seq, BS-seq, BAM QC, …
  - Only one execution from FASTQ to BAM file, easy pipelines
- Computational needs
  - Performance and memory usage is not acceptable in most cases, software democratization
  - HPC technologies can be applied: multi-core, SSE (SIMD), GPU, ...
  - Software not ready for scientific clusters: MPI
  - Software must be designed for working on *clouds*
  - Poor software engineering: lack of libs
  - HPG project released: http://www.opencb.org/technologies/hpg
  - HPG is part of the OpenCB initiative released to the community http://www.opencb.org

Ignacio Medina
imedina@ebi.ac.uk

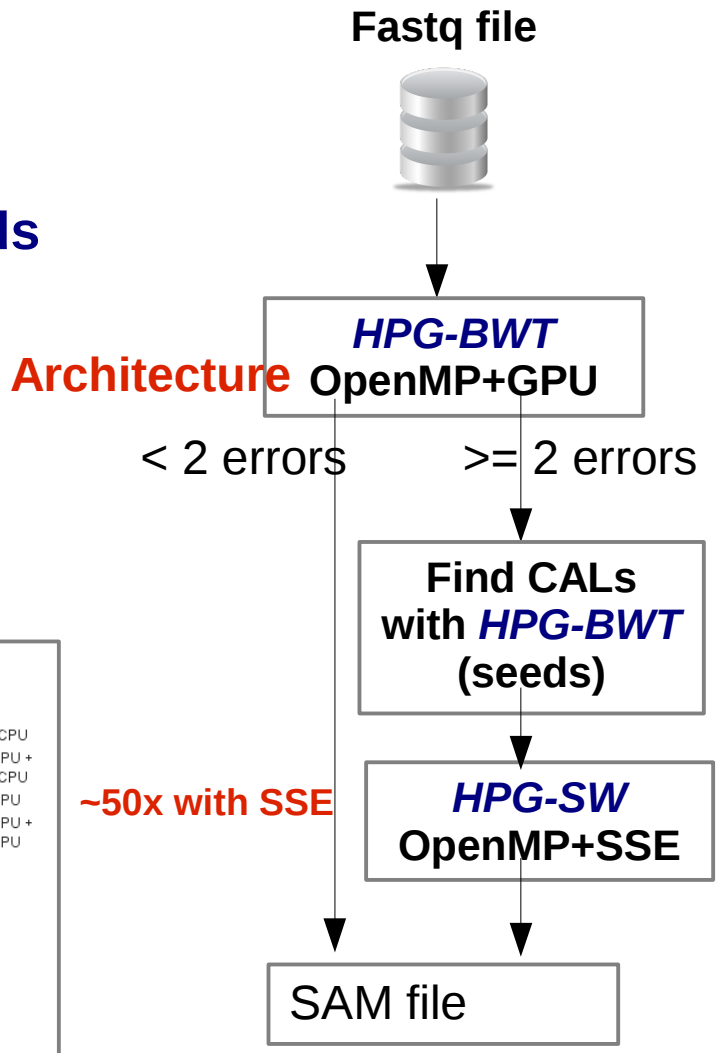
HPG Aligner
Architecture and features

- Current read aligners software tend to fit in one of these groups:
  - *Very fast, but no too sensitive*: no gaps, no indels, rna-seq...
  - *Slow, but very sensitive*: up to 1 day by sample
- Current aligners show **bad performance** with **long reads**
- Current read Aligner algorithms
  - *Burrows-Wheeler Transform (BWT)*: very fast! No sensitive
  - *Smith-Waterman (SW)*: very sensitive but very slow
- Hybrid approach (*papers in preparation*):
  - ***HPG-BWT*** implemented with *OpenMP and Nvidia CUDA*
  - ***HPG-SW*** implemented using *OpenMP and SSE (~26x in 8-core)*



Testing BWT-GPU

Architecture



Ignacio Medina
imedina@ebi.ac.uk

Mapping NGS reads for genomic studies

HPG Aligner
Benchmarks and results: *DNA* alignment

- First results show an amazing ***performance*** and the best ***sensitivity***

***DNA* 40M simulated datasets**

Program	100nt %mapped Time(min)	150nt %mapped Time(min)	400nt %mapped Time(min)
HPG Aligner 2.0 *dna* mode	98.77% 20.57min	99.54% 22.90min	99.93% 31.35min
BWA MEM 0.7.5	96.99% 29.34min	98.09% 43.35min	99.12% 124.16min
Bowtie2 2.1.0	94.67% 29.40min	96.71% 47.61min	98.82% 209.26min

***DNA* 4M simulated INDEL datasets**

Program	100nt %mapped I10	150nt %mapped I10	400nt %mapped I15
HPG Aligner 2.0 *dna* mode	84.69%	88.23%	88.94%
BWA MEM 0.7.5	83.37%	81.92%	77.36%
Bowtie2 2.1.0	68.50%	66.28%	52.13%

- Correctly mapped results
- No GPUs were used
- Other tools were benchamarked: GEM, SOAP, BFAST,... but no positive result were obtained

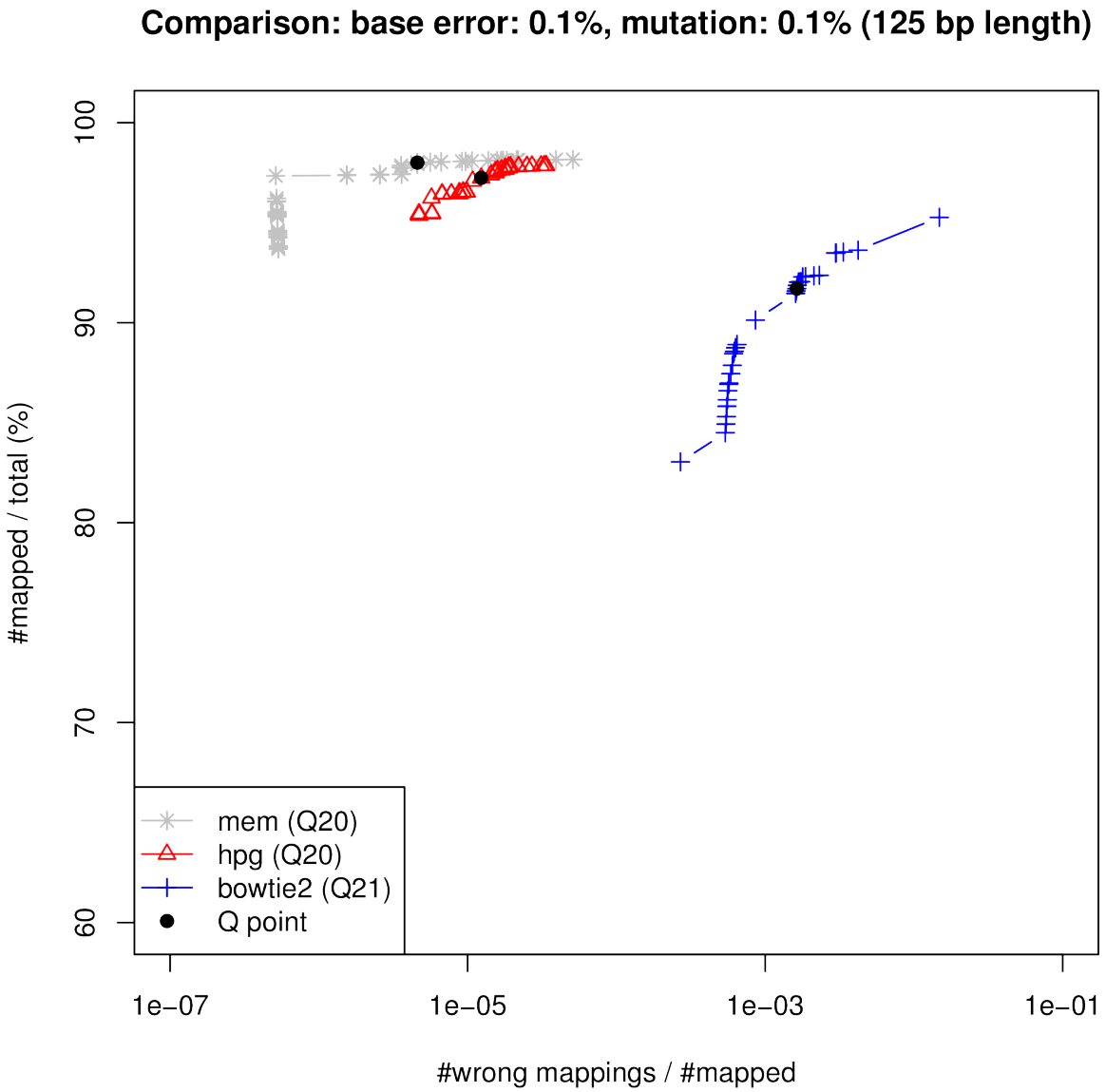
Published, Bioinformatics 2014
http://bioinformatics.oxfordjournals.org/content/early/2014/09/01/bioinformatics.btu553.long

Ignacio Medina
imedina@ebi.ac.uk

HPG Aligner

Benchmarks and results: *DNA* alignment

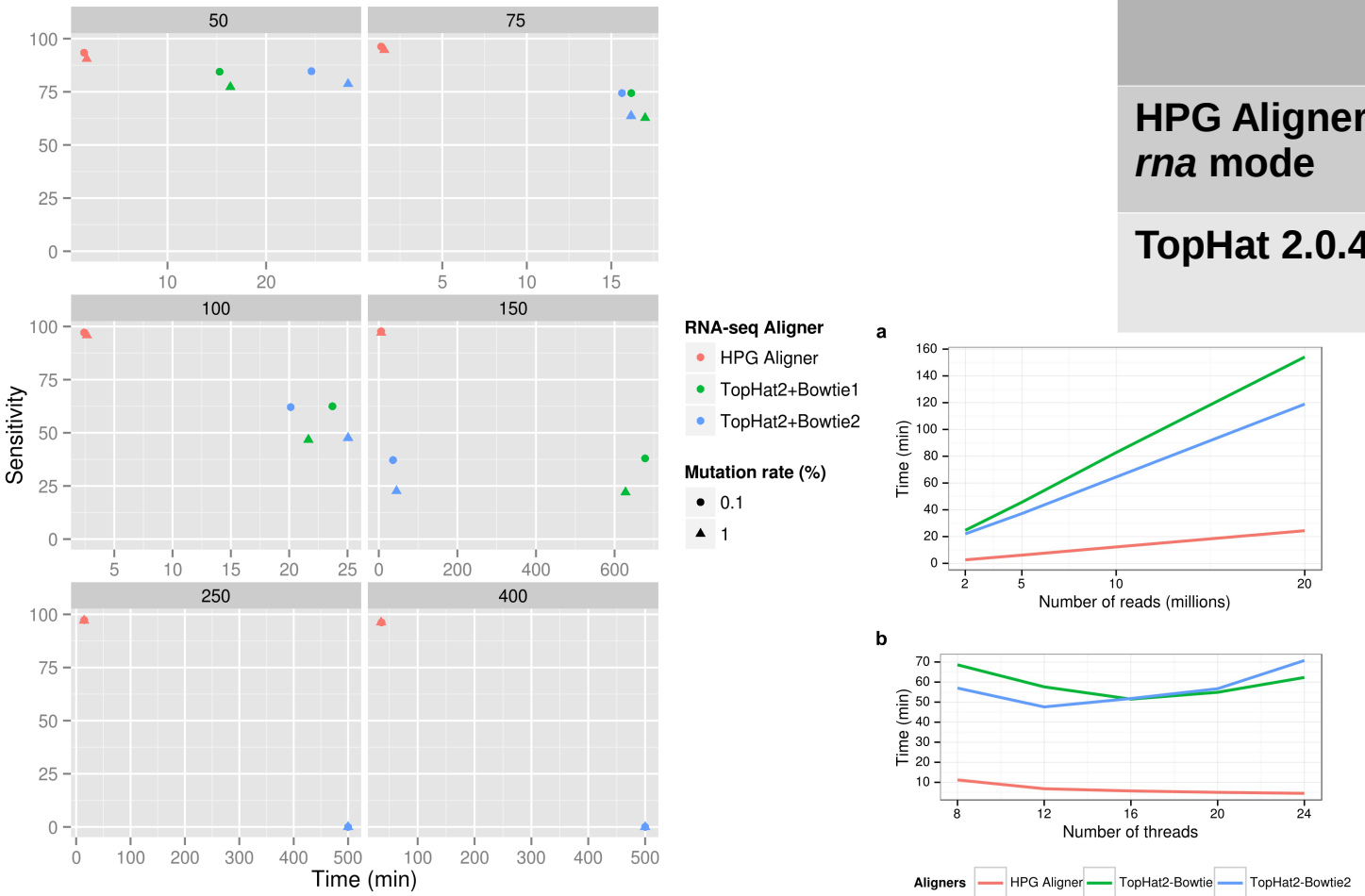
- First results show an amazing ***performance*** and the best ***sensitivity***
- Simulation studies are very valuable to set up a analysis pipeline
- Several simulators for DNA and RNA-seq available



Ignacio Medina
imedina@ebi.ac.uk

HPG Aligner
Benchmarks and results: *RNA-seq*

Comparative results with TopHat2 using both Bowtie and Bowtie2



***RNA-seq* 1M simulated datasets**

Program	75nt %mapped Time(min)	100nt %mapped Time(min)	150nt %mapped Time(min)
HPG Aligner *rna* mode	96.6% 1.03min	97.47% 1.2min	98.35% 2.28min
TopHat 2.0.4	74.4% 15.6min	62.1% 20.1min	37.1% 36.2min

Notes:

15x faster!!

- Max errors allowed
- Similar results are obtained with real datasets
- TopHat doubles disk space and big memory needs

Hardware scalability tests

Under review in Bioinformatics

Ignacio Medina
imedina@ebi.ac.uk

Mapping NGS reads for genomic studies

HPG Aligner
Main and coming features

- Part of the HPG suite (http://www.opencb.org/technologies/hpg) with other tools: *hpg-fastq*, *hpg-bam*, *hpg-aligner*, *hpg-variant*
- Only one execution is needed to generate the BAM output file (saves disk)
- Faster index creator, multi-core implementation
- Designed to provide the better sensitivity
- Soft clipping of adaptors
- HPC technologies used to provide the fastest runtime: multicore, SSE, GPUs, …
- Open-source and open development, code at GitHub https://github.com/opencb-hpg
- Part of the OpenCB project: http://www.opencb.org
- Coming features
  - DNA: INDEL realignment (GATK algorithm)
  - BS-seq: for methylation analysis (being testing)
  - RNA-seq: suport for no canonical splices
  - Hadoop implementation will allow to run it in a distributed environment
  - Performance improvements

Ignacio Medina
imedina@ebi.ac.uk

SAM/BAM specification
Mapping output: SAM/BAM format

SAM Specification: http://samtools.sourceforge.net/SAM1.pdf

Take a quick look:

```
@PG     ID:HPG-Aligner  VN:1.0
@SQ     SN:20   LN:63025520

HWI-ST700660_138:2:2105:7292:79900#2@0/1        16      20      76703   254     76=     *       0       0
GTTTAGATACTGAAAGGTACATACTTCTTTGTAGGAACAAGCTATCATGCTGCATTTCTATAATATCACATGAATA
GIJGJLGGFLILGGIEIFEKEDELIGLJIHJFIKKFELFIKLFFGLGHKKGJLFIIGKFFEFFEFGKCKFHHCCCF    AS:i:254       NH:i:1  NM:i:0

HWI-ST700660_138:2:2208:6911:12246#2@0/1        16      20      76703   254     76=     *       0       0
GTTTAGATACTGAAAGGTACATACTTCTTTGTAGGAACAAGCTATCATGCTGCATTTCTATAATATCACATGAATA
HHJFHLGFFLILEGIKIEEMGEDLIGLHIHJFIKKFELFIKLEFGKGHEKHJLFHIGKFFDFFEFGKDKFHHCCCF    AS:i:254       NH:i:1  NM:i:0

HWI-ST700660_138:2:1201:2973:62218#2@0/1        0       20      76655   254     76M     *       0       0
AACCCCAAAAATGTTGGAAGAATAATGTAGGACATTGCAGAAGACGATGTTTAGATACTGAAAGGGACATACTTCT
FEFFGHHHGGHFKCCJKFHIGIFFIFLDEJKGJGGFKIHLFIJGIEGFLDEDFLFGEIIMHHIKL$BBGFFJIEHE    AS:i:254       NH:i:1  NM:i:1

HWI-ST700660_138:2:1203:21395:164917#2@0/1      256     20      68253   254     4M1D72M *       0       0
NCACCCATGATAGACCAGTAAAGGTGACCACTTAAATTCCTTGCTGTGCAGTGTTCTGTATTCCTCAGGACACAGA
#4@ADEHFJFFEJDHJGKEFIHGHBGFHHFIICEIIFFKKIFHEGJEHHGLELEGKJMFGGGLEIKHLFGKIKHDG    AS:i:254       NH:i:3  NM:i:1

HWI-ST700660_138:2:1105:16101:50526#6@0/1       16      20      126103  246     53M4D23M        *       0       0
AAGAAGTGCAAACCTGAAGAGATGCATGTAAAGAATGGTTGGGCAATGTGCGGCAAAGGGACTGCTGTGTTCCAGC
   FEHIGGHIGIGJI6FCFHJIFFLJJCJGJHGFKKKKGIJKHFFKIFFFKHFLKHGKJLJGKILLEFFLIHJIEIIB    AS:i:368       NH:i:1  NM:i:4
```

Ignacio Medina
imedina@ebi.ac.uk

SAM/BAM specification
Mapping output, mandatory fields

First columns are mandatory

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	`[0,2^16-1]`	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	`[0,2^29-1]`	1-based leftmost mapping POSition
5	MAPQ	Int	`[0,2^8-1]`	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next segment
8	PNEXT	Int	`[0,2^29-1]`	Position of the mate/next segment
9	TLEN	Int	`[-2^29+1,2^29-1]`	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

Flags

Bit	Description
0x1	template having multiple segments in sequencing
0x2	each segment properly aligned according to the aligner
0x4	segment unmapped
0x8	next segment in the template unmapped
0x10	SEQ being reverse complemented
0x20	SEQ of the next segment in the template being reversed
0x40	the first segment in the template
0x80	the last segment in the template
0x100	secondary alignment
0x200	not passing quality controls
0x400	PCR or optical duplicate

CIGAR

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

Ignacio Medina
imedina@ebi.ac.uk

SAM/BAM specification
Mapping output, optional fields

Some optionals fields, in the aligner section

SAM specification is part of **SAMtools** package. More info at:
http://samtools.sourceforge.net/

A binary SAMtools is distributed freely to:
- SAM ↔ BAM
- Depth
- Merge
- Sort
- ...

Tag[1]	Type	Description
X?	?	Reserved fields for end users (together with Y? and Z?)
AM	i	The smallest template-independent mapping quality of segments in the rest
AS	i	Alignment score generated by aligner
BC	Z	Barcode sequence
BQ	Z	Offset to base alignment quality (BAQ), of the same length as the read sequence. At the i-th read base, $\mathrm{BAQ}_i = Q_i - (\mathrm{BQ}_i - 64)$ where Q_i is the i-th base quality.
CC	Z	Reference name of the next hit; "=" for the same chromosome
CM	i	Edit distance between the color sequence and the color reference (see also NM)
CP	i	Leftmost coordinate of the next hit
CQ	Z	Color read quality on the original strand of the read. Same encoding as QUAL; same length as CS.
CS	Z	Color read sequence on the original strand of the read. The primer base must be included.
E2	Z	The 2nd most likely base calls. Same encoding and same length as QUAL.
FI	i	The index of segment in the template.
FS	Z	Segment suffix.
FZ	B,S	Flow signal intensities on the original strand of the read, stored as `(uint16_t) round(value * 100.0)`.
LB	Z	Library. Value to be consistent with the header RG-LB tag if @RG is present.
H0	i	Number of perfect hits
H1	i	Number of 1-difference hits (see also NM)
H2	i	Number of 2-difference hits
HI	i	Query hit index, indicating the alignment record is the i-th one stored in SAM
IH	i	Number of stored alignments in SAM that contains the query in the current record
MD	Z	String for mismatching positions. *Regex*: `[0-9]+(([A-Z]|\^[A-Z]+)[0-9]+)*`[2]
MQ	i	Mapping quality of the mate/next segment
NH	i	Number of reported alignments that contains the query in the current record
NM	i	Edit distance to the reference, including ambiguous bases but excluding clipping
OQ	Z	Original base quality (usually before recalibration). Same encoding as QUAL.
OP	i	Original mapping position (usually before realignment)
OC	Z	Original CIGAR (usually before realignment)
PG	Z	Program. Value matches the header PG-ID tag if @PG is present.
PQ	i	Phred likelihood of the template, conditional on both the mapping being correct
PU	Z	Platform unit. Value to be consistent with the header RG-PU tag if @RG is present.
Q2	Z	Phred quality of the mate/next segment. Same encoding as QUAL.
R2	Z	Sequence of the mate/next segment in the template.
RG	Z	Read group. Value matches the header RG-ID tag if @RG is present in the header.
SM	i	Template-independent mapping quality
TC	i	The number of segments in the template.

Ignacio Medina
imedina@ebi.ac.uk

Best practices
Take home messages

- Choose the best aligner for your analysis and hardware
- Remove duplicated and low qualities reads from FASTQ
- Try to use paired-end datasets for variant calling and structural variation analysis. In RNA-seq paired-end can detect gene fusions
- Do **not allow multiple hits** for variant calling analysis. RNA-seq depending on read size and the analysis to perform
- Realign INDELS and recalibrate mapping quality for variant calling analysis
- **Simulation** can be very useful for choosing the right aligner

Data repositories
Open and controlled access repositories

- **1000 Genome** project
  - http://www.1000genomes.org/
- **SRA**, *Short Read Archive*
  - http://www.ncbi.nlm.nih.gov/sra
- **EGA**, European Genome-Phenome Archive
  - https://www.ebi.ac.uk/ega
- … and many others

Hands-on
Tutorial

Go to:

http://ngs-course.github.io/Course_Materials/alignment/tutorial/example.html

Ignacio Medina
imedina@ebi.ac.uk

QC alignment
Motivation

- We need to know how well the alignment process went
- Hundreds of million of mapped reads
- Some biases can occur
- Some useful information
  - % reads mapped
  - Mean average error
  - Error distribution
  - Length distribution
  - Coverage
  - …
- Not many software for QC available, sometimes you have to use more than one

QC alignment
Download QC software

- **SAMstat**, download from http://samstat.sourceforge.net/
  - Uncompress
    - tar zxvf samstat.tgz
    - cd samstat/src
    - make
    - Move the binary *userhome/bin* folder
- **HPG-BAM**, download from http://wiki.opencb.org/projects/hpg/doku.php?id=utilities:bam
  - Uncompress
    - gunzip hpg-bam.gz
    - Move the binary *userhome/bin* folder

QC alignment
SAMstat

- **SAMstat**, easy to execute:
  - ./samstat hq-test_pe.bam
- Produces a HTML5 page, use Google Chrome or Firefox to open it, some useful info:
  - % mapped (grouped by mapping quality)
  - Error distribution
  - Length distribution
  - ...
- Download BAM file from 1000Genomes to study real data

QC alignment
HPG-BAM

- **HPG-BAM**, we want to use '*stats*' command, read tutorial from http://wiki.opencb.org/projects/hpg/doku.php?id=utilities:bam
  - ./hpg-bam stats -b hq-test1.bam -o /tmp
- Produces some plots and txt files with stats:
  - Coverage
  - Error distribution
  - GC content
  - …
- Download BAM file from 1000Genomes to study real data
- Cloud based version being developed