New advanced solutions for Genomic *big data* Analysis and Visualization

Ignacio Medina (*Nacho*)
im411@cam.ac.uk
http://bioinfo.cipf.es/imedina
Head of Computational Biology Lab
HPC Service, University of Cambridge
Cambridge, UK

UNIVERSITY OF CAMBRIDGE

Genomics england

About me

- Currently
  - Head of Computational Biology Lab at HPC Service, University of Cambridge, UK
  - Team lead for Bioinformatics Software Development at Genomics England (*aka* UK100K), London, UK
  - External Scientific Collaborator at EBI Variation, EMBL-EBI, Hinxton, UK
- Formerly
  - About 1 year as a Project Manager at EMBL-EBI Variation Team
  - About 8 years at CIPF (Spain) working in genomic data analysis (algorithms and tools development)
- Education
  - BSc in Biochemistry and MSc in Genetics (Universitat de Valencia)
  - BSc in Computer Science and Machine Learning (Universitat Politecnica de Valencia)
- Interests
  - Genomics, Data Analytics, Big Data, Visualization, HPC

Index

- Introduction
- OpenCB initiative
- Advanced Computing Technologies
- OpenCB Projects
  - CellBase
  - Genome Maps, *Big data* visualization
  - High-Performance Genomics (HPG)
  - OpenCGA
- Q&A

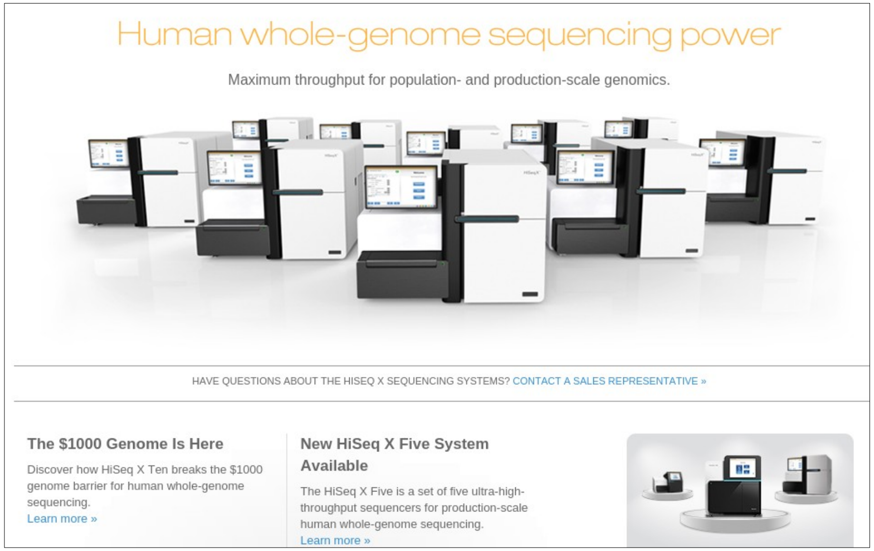
Introduction

Big data in Genomics, a new scenario in biology

Next-Generation Sequencing (*NGS*) high-throughput technology for DNA sequencing that is changing the way how genomic researchers are perform experiments. Many new experiments are being conducted by sequencing: *re-sequencing, RNA-seq, Meth-seq, ChIP-seq, …*
*Experiments **have increased data size by more than 5000x** when compared with microarrays.*
***Surprisingly**, many existing software solutions are not very different.*

Sequencing costs keep falling, today a whole genome can be sequenced by just **$1000**, so much more data is expected. Data acquisition is ***highly distributed*** and involves heterogeneous formats.

A single HiSeq X Ten System can sequence ~20,000 human genomes a year



Human whole-genome sequencing power

Maximum throughput for population- and production-scale genomics.

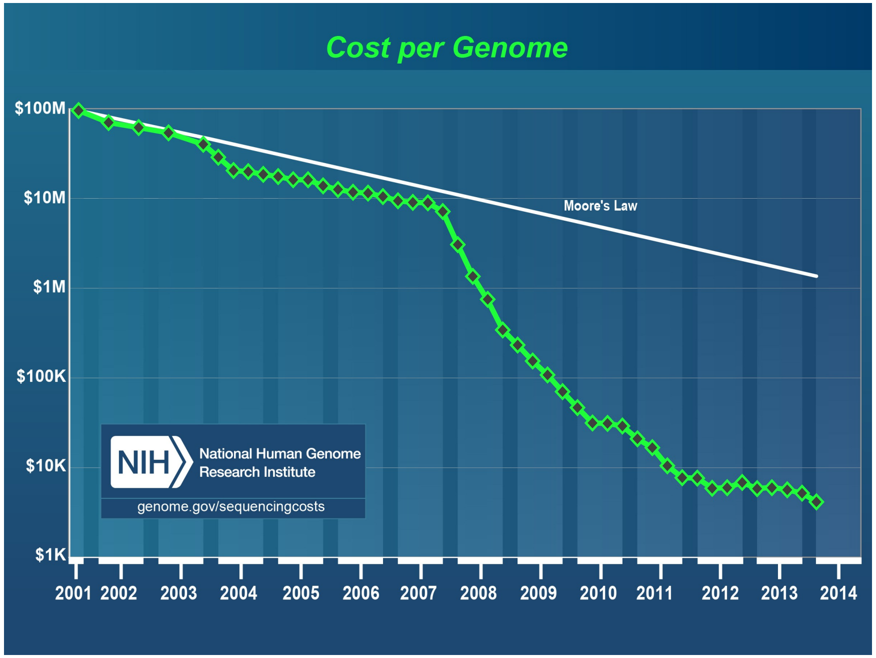
HAVE QUESTIONS ABOUT THE HISEQ X SEQUENCING SYSTEMS? CONTACT A SALES REPRESENTATIVE »

The $1000 Genome Is Here

Discover how HiSeq X Ten breaks the $1000 genome barrier for human whole-genome sequencing.
Learn more »

New HiSeq X Five System Available

The HiSeq X Five is a set of five ultra-high-throughput sequencers for production-scale human whole-genome sequencing.
Learn more »

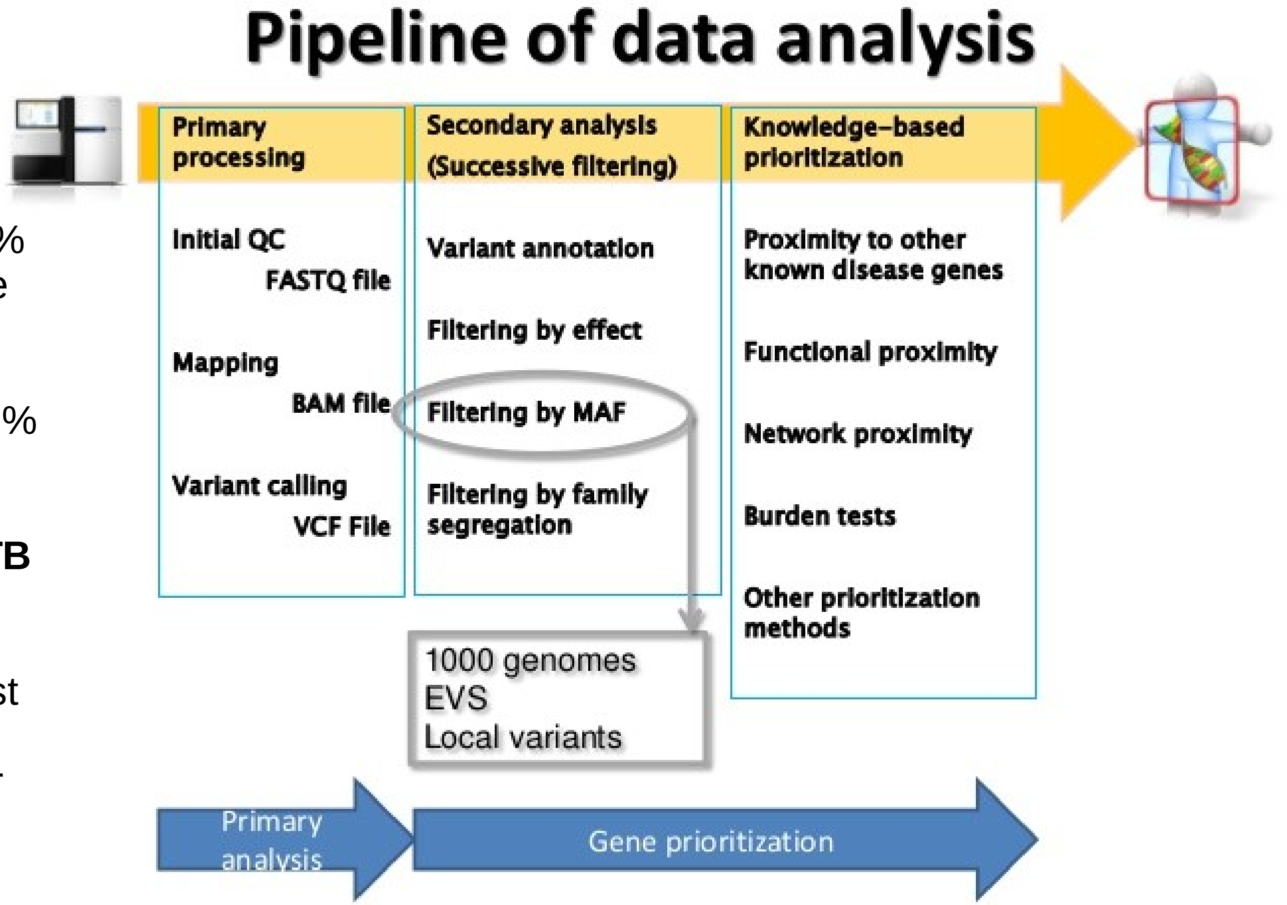


Introduction

Standard NGS experiment

- Illumina NGS sequencer series:
  - **HiSeq 2500** provides high-quality 2x125bp: 50-1000Gb in 1-6 days, 90.2% bases above Q30. One human genome at ~60x coverage
  - **HiSeq 4000** provides high-quality 2x150bp: 125-1500Gb in 1-4 days, >75% bases above Q30. Up to 12 human genomes at ~40x coverage
- ***Each sample*** produces a ***FASTQ*** file ~**1TB** size containing ~**1-2B** reads
- New **Illumina X Ten**: Consists if 10 ultra-high-throughput HiSeq X sequencers. First $1000 human genome sequencer, it can sequence up to 20,000 genomes per year

Real flexibility.
Real throughput.
Real data quality.
The HiSeq 2500 is ready for any application, any sample size—today.



Variant calling pipeline

Introduction

Big data in Genomics, some current projects

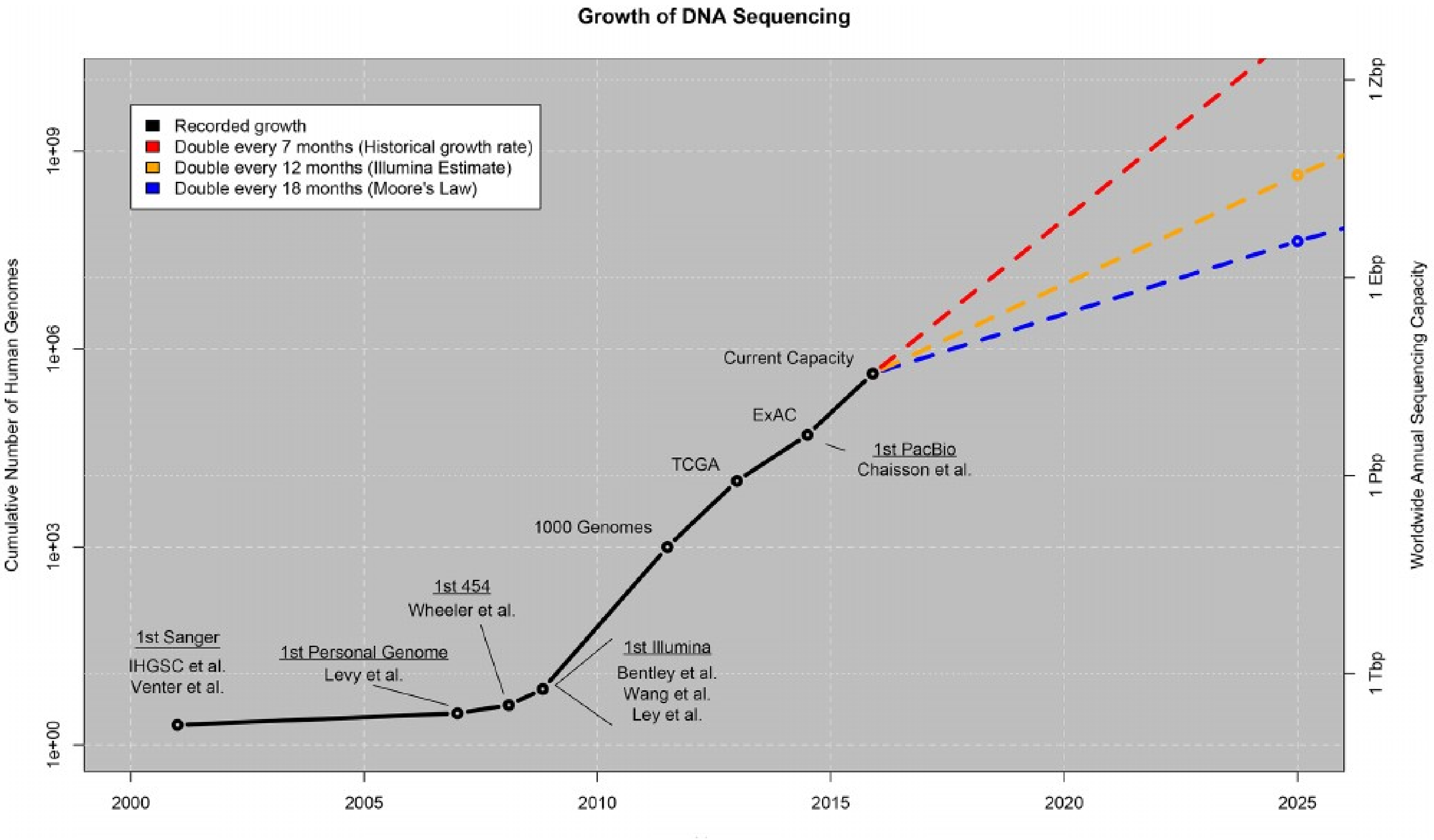
- At **EMBL-EBI and Sanger**
  - ***European Genome-phenome Archive (EGA)***: stores human datasets under controlled access https://www.ebi.ac.uk/ega/home Current size about 2PB, it is expected to increase 2x-3x over the next few years. Now new functionality is being implemented.
  - ***European Variation Archive (EVA)***: open archive for all public genomic variation data for all species http://www.ebi.ac.uk/eva/ A new project with only a few TBs of data so far
  - ***1000G Phase 3***: about 2500 individuals from 26 populations, a few hundreds of TBs
- Other ***big data*** projects
  - ***NIHR BRIDGE***: 10,000 whole genomes from rare diseases, ~1-2PB of data expected
  - ***Genomics England (GEL)***: is sequencing **100K** whole genomes from UK, several rare diseases and cancers being studied, data estimation: ~20PB of BAM and ~400TB of VCF data are expected! About 100 whole genomes/day, ~5-10TB/day
  - ***International Cancer Genome Consortium (ICGC)***: store more than 10,000 sequenced cancers, few PB of data
- And of course **many** “medium-sized” projects. Data acquisition is ***highly distributed, not a single huge project***
- ***Data encryption and security*** is a concern in many of them

Introduction

Estimated sequencing data size projection for human

Big Data: Astronomical or Genomical?
Stephens ZD. et al., PLOS Biology, July 2015

ABSTRACT - Genomics is a Big Data science and is going to get much bigger, very soon, but it is not known whether the needs of genomics will exceed other Big Data domains. ***Projecting to the year 2025****, we compared genomics with three other major generators of Big Data:* ***astronomy, YouTube, and Twitter****. Our estimates show that* ***genomics is a "four-headed beast"—it is either on par with or the most demanding of the domains analyzed here in terms of data acquisition, storage, distribution, and analysis***



Introduction

NGS genomic variation data, big and complex

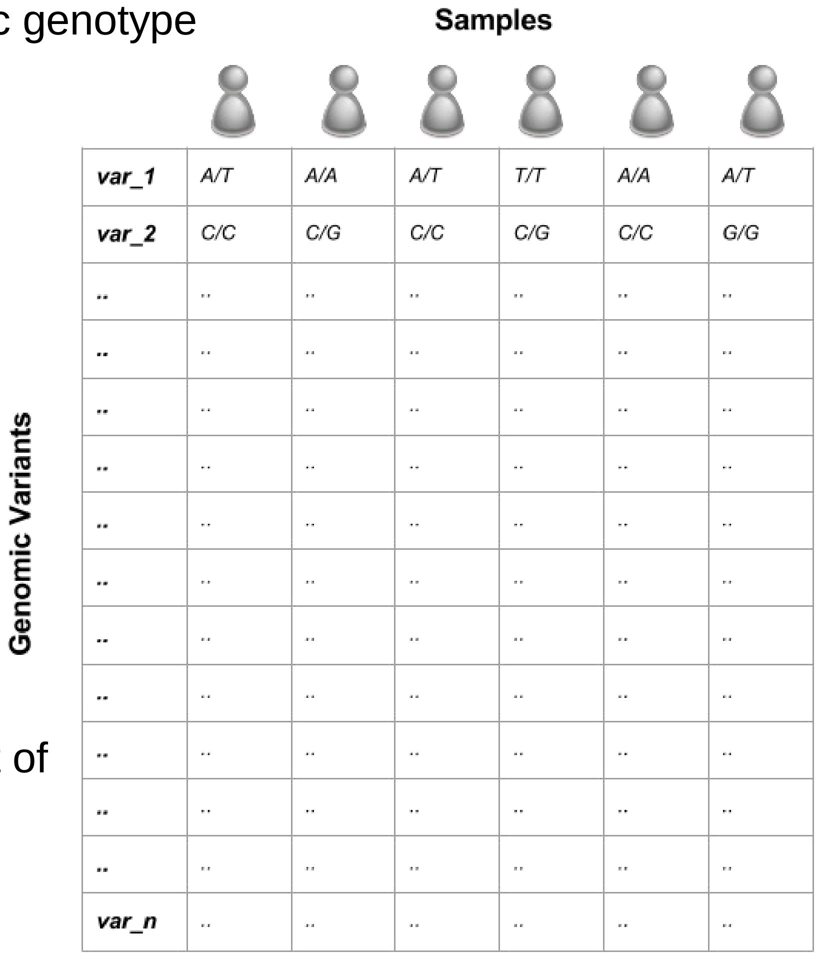
Logical view of genomic variation data, real data comes in **different VCF files**.

Each cell represents one specific genotype for one mutation in one sample

Hundreds of millions of mutations, some meta data needed: **Variant annotation**

- Clinical info
- Consequence type
- Conservation scores
- Population frequencies
- ...

Genomics England project: 200M variants x 100K samples. About **20 trillion** points with a lot of meta data. About **500-1000TB** to be indexed.



Samples

Genomic Variants

var_1	A/T	A/A	A/T	T/T	A/A	A/T
var_2	C/C	C/G	C/C	C/G	C/C	G/G
..	..	..	..	..	..	..
..	..	..	..	..	..	..
..	..	..	..	..	..	..
..	..	..	..	..	..	..
..	..	..	..	..	..	..
..	..	..	..	..	..	..
..	..	..	..	..	..	..
..	..	..	..	..	..	..
..	..	..	..	..	..	..
..	..	..	..	..	..	..
..	..	..	..	..	..	..
var_n	..	..	..	..	..	..

Meta data: **Sample annotation**

- Phenotype
- Family pedigree, Population
- Clinical variables
- ...

Heterogeneous data analysis and algorithms, different technologies and solutions required:

- Search and filter using data and meta data
- Data mining, correlation
- Statistic tests
- Machine learning
- Interactive analysis
- Network-based analysis
- Visualization
- Encryption
- ...

Applications:

- Personalized medicine
- ...

Introduction

Big data analysis challenges

- ***Data Analysis and visualization***: Real-time and Interactive graphical data analysis and visualization is needed.
- ***Data mining***: Complex queries, aggregations and correlations are needed
- ***Security***: sometimes data access require authentication, authorization, *encryption*, ...
- ***Performance and scalability***: software must be high-performance and scalable
- ***Data Integration***: different types of data such as variation, expression, ChIP, ...
- ***Share and collaboration***: many projects require the collaboration among different groups. Moving data is not a good idea.
- ***Knowledge base and sample annotations***: many of the visual analytic tools need genome and sample annotations

Do current bioinformatic tools solve these problems?

Introduction

Current status in bioinformatics

- Bioinformatic tools are in general not designed for processing and analyzing big data.
- Most of the them are written in R, Python, Java, … They usually don't exploit the parallelism of modern hardware and current technologies. Poor performance and scalability.
- We need to develop new generation of software and methodologies to:
  - Improve performance and scalability of analysis
  - Store data efficiently and secured to be queried and visualized
  - Easy to adapt to new technologies and uses cases to be useful to researchers

- So, are bioinformaticians doing things right? Are researchers happy with their current tools?
- Is needed a ***paradigm change*** in bioinformatics? Most tools are designed and implemented to run in a single ***workstation***, but PB scale data require more ***advanced computing technologies***

OpenCB

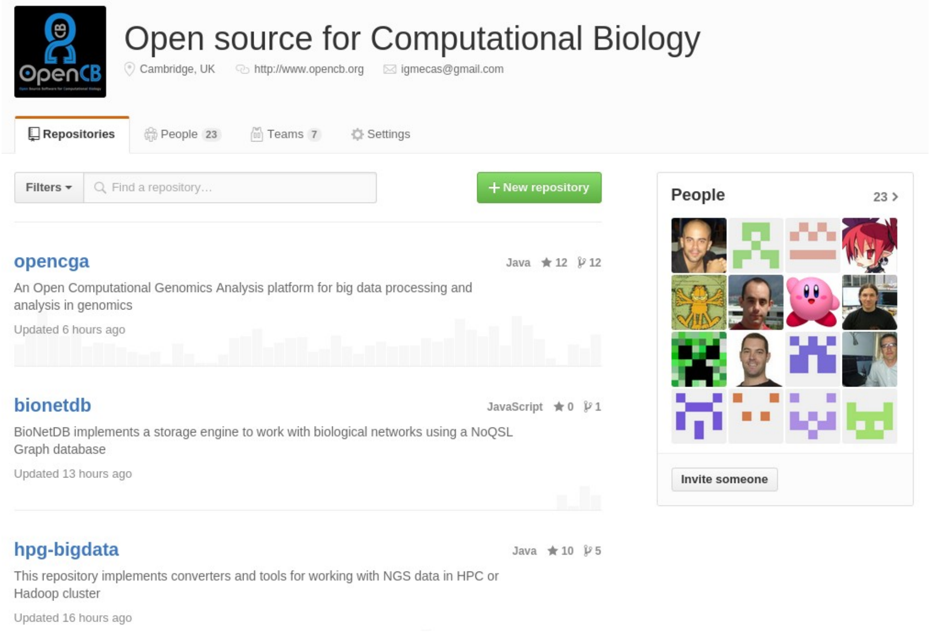
Open source initiative for Computational Biology

- Software in Biology is still usually developed in small teams, we must learn to collaborate in bigger projects to solve bigger problems. OpenCB tries to engage people to work in big problems: **http://www.opencb.org**
- Shared, collaborative and well designed platforms to build more **advanced solutions to solve current biology problems** are needed
- No one computing programming language oriented! No like BioPerl, BioPython, Bioconductor, ... Good software solutions may use different languages and technologies to solve different problems and use cases
- So far, is where all the software we develop is being released. About 15 active committers. Available as open-source at GitHub **https://github.com/opencb**

OpenCB is a collaborative project with more than 16 actives data analysts and developers and more than 12 repositories

Many papers published during last two years, very good adoption

Many different technologies used: HPC, Hadoop, web applications, NoSQL databases, ...



Open source for Computational Biology

Cambridge, UK   http://www.opencb.org   igmecas@gmail.com

opencga Java ★ 12 12

An Open Computational Genomics Analysis platform for big data processing and analysis in genomics

Updated 6 hours ago

bionetdb JavaScript ★ 0 1

BioNetDB implements a storage engine to work with biological networks using a NoQSL Graph database

Updated 13 hours ago

hpg-bigdata Java ★ 10 5

This repository implements converters and tools for working with NGS data in HPC or Hadoop cluster

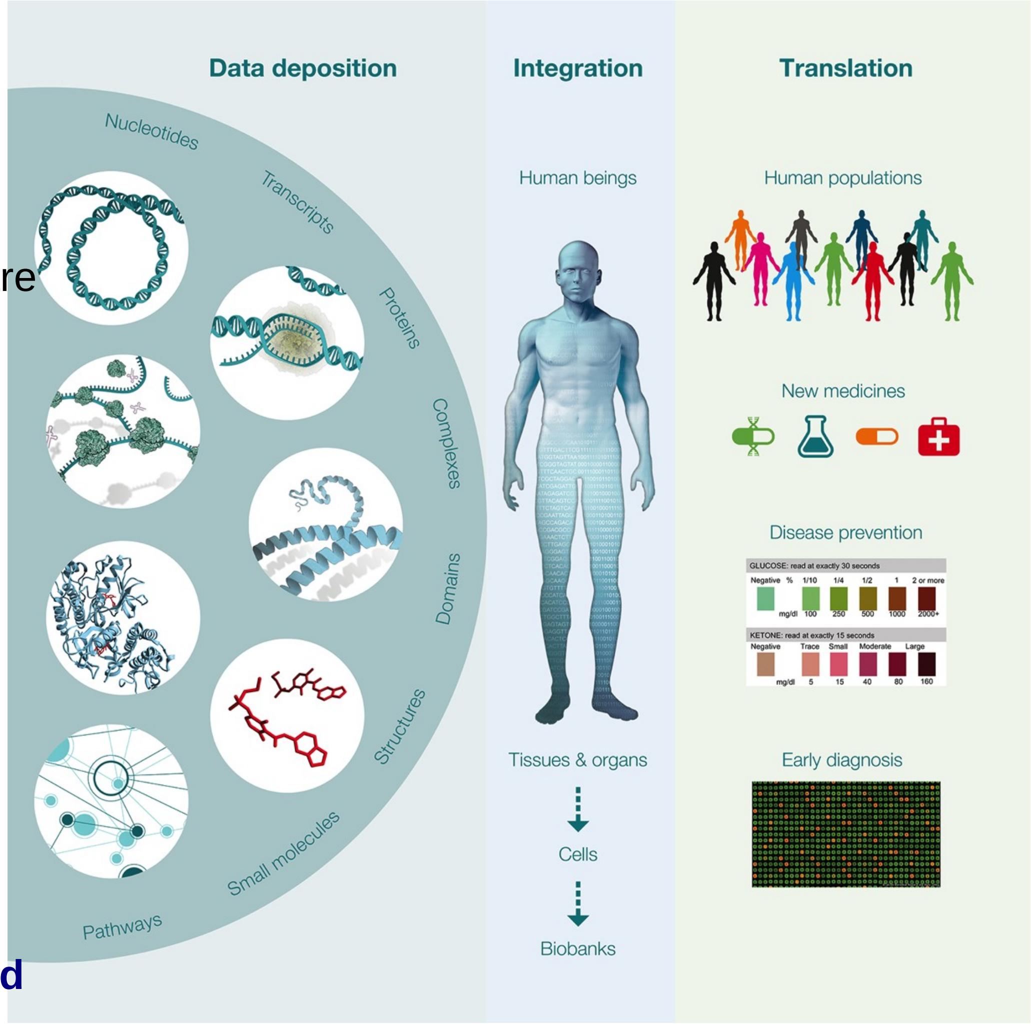
Updated 16 hours ago

OpenCB

Motivation and goals

- *Poor performance*, **Software** in Bioinformatics is generally not designed for processing big datasets in big clusters *like in other science areas*.
- As data analysis group we focused in **analysis***.* Typically Few weeks of sample **sequencing and preprocessing.** Then **several months of data analysis**
- ***Software Goals***: to develop new generation of software and methodologies in bioinformatics to:
  - High-performance software with a **low memory** footprint and **useful analysis**
  - be able to handle and analyze TB of data, **index** data efficiently to be queried and visualized
  - distribute computation, no data. Make uses of ***clouds***
  - develop tools and pipelines for ***diagnosing***
  - ***Maintainability and adaptative***
- Software ***must*** exploit current hardware and infrastructures and be **fast and efficient in a standard workstation or in a Cloud or HPC cluster**

Data analysis software must be ***efficient and useful*** for researchers and clinicians



OpenCB

A *big data* friendly architecture

Many **analysis and visualization** tools in Bioinformatics are still **desktop-based** applications: ***It's the server!***

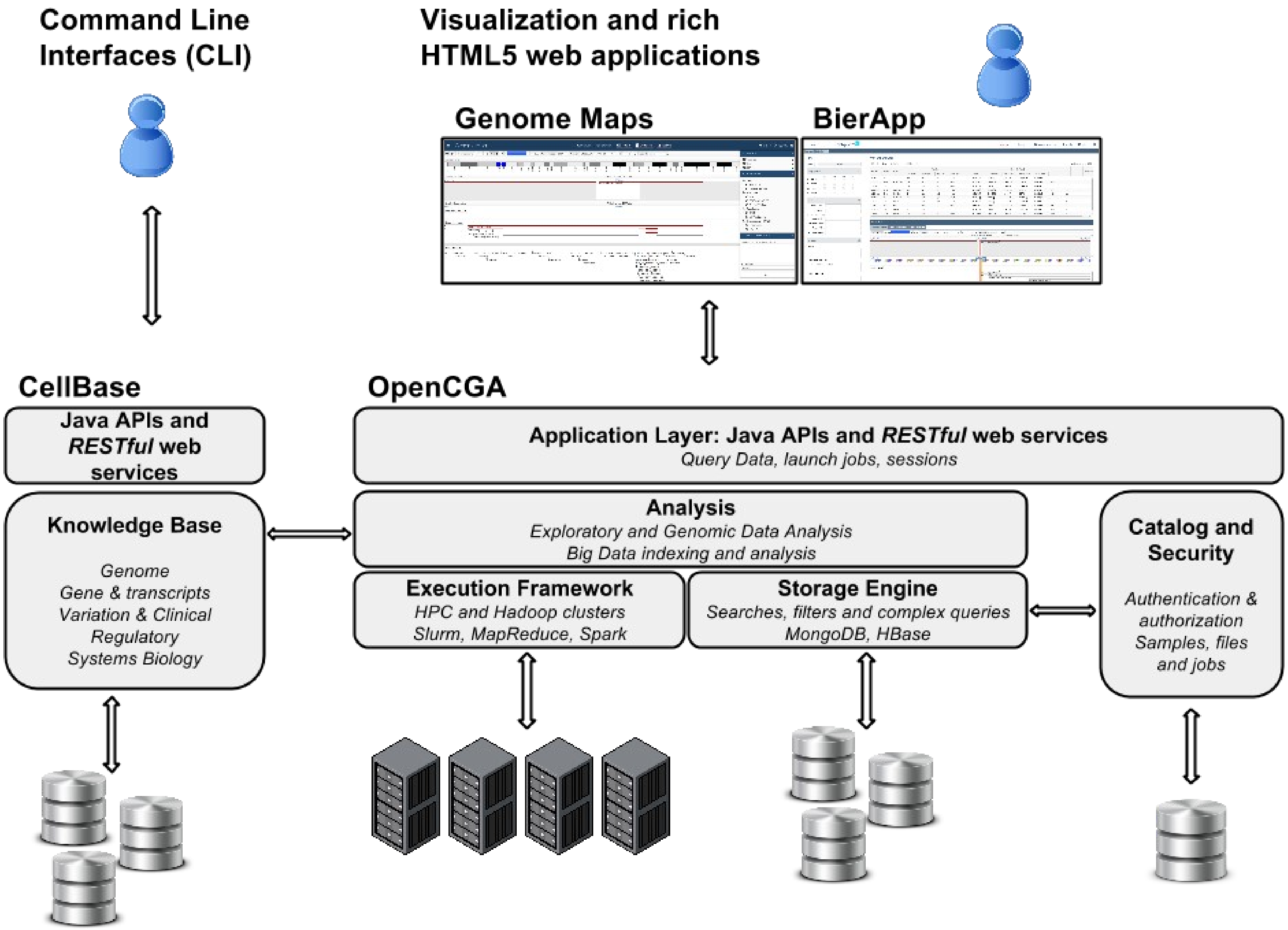
Client

Rich Web applications and visualization. New HTML5 web technologies: SVG, IndexedDB, WebWorkers, WebGL, SIMD.js, ...

Server

Distributed and HPC technologies for real-time and interactive data analysis and visualization: NoSQL, Hadoop, HPC, …

Search, filter and aggregate the ***data needed*** by researchers



Command Line Interfaces (CLI)

Visualization and rich HTML5 web applications

Genome Maps

BierApp

CellBase

Java APIs and *RESTful* web services

Knowledge Base

Genome
Gene & transcripts
Variation & Clinical
Regulatory
Systems Biology

OpenCGA

Application Layer: Java APIs and *RESTful* web services
Query Data, launch jobs, sessions

Analysis
Exploratory and Genomic Data Analysis
Big Data indexing and analysis

Execution Framework
HPC and Hadoop clusters
Slurm, MapReduce, Spark

Storage Engine
Searches, filters and complex queries
MongoDB, HBase

Catalog and Security

Authentication & authorization
Samples, files and jobs

Advanced Computing Technologies

Web technologies, HTML5 standard and RESTful WS

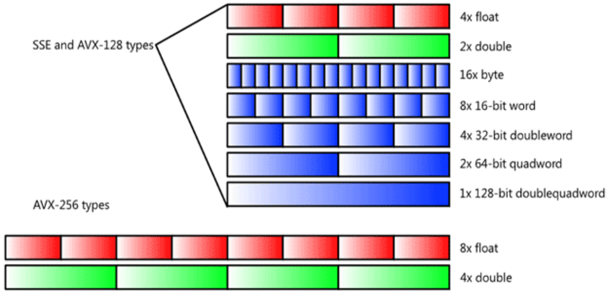
- Web applications are becoming the new “desktop applications”
- **HTML5** brings many new standards and libraries to JavaScript allowing developing amazing web applications in a easy way
  - **Canvas/SVG inline**: browsers can render dynamic content and inline static SVG easily
  - **IndexedDB**: client-side indexed storage for high-performance query
  - **WebWorkers**: simple mean for creating OS threads
  - **WebGL**: hardware-accelerated 3D graphics to web, based on OpenGL ES
- Many new web tools, frameworks and libraries being developed to build great and bigger web applications:
  - Performance: asm.js, SIMD.js, WebCL, pNaCl, ...
  - Web Components: Polymer
  - Build: Grunt, Bower, Yeoman, …
  - Visualization: d3.js, Three.js, Highcharts, …
- RESTful web services and JSON ease the development of light and fast RPC

Advanced Computing Technologies

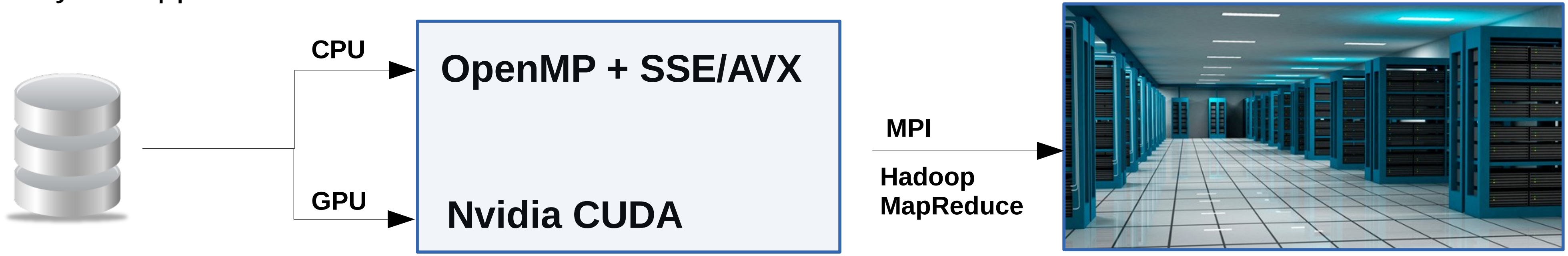
High-Performance Computing (HPC)

- HPC hardware:
  - **Intel MIC architecture**: Intel Xeon and Intel Xeon Phi coprocessor, 1.01Tflops DP and more than 50 cores
  - **Nvidia Tesla**: Tesla K20X almost 1.31Tflops DP and 2688 CUDA cores

- Some HPC frameworks available:
  - **Shared-memory parallel**: OpenMP, OpenCL
  - **GPGPU computing**: CUDA, OpenCL, OpenACC
  - **Message passing Interface (MPI)**
  - **SIMD**: SSE4 instructions extended to AVX2 with a 256-bit SIMD



- Heterogeneous HPC in a shared-memory
  - CPU (*OpenMP+AVX2*) + GPU (*CUDA*)
- Hybrid approach:



Advanced Computing Technologies

Big data analysis and NoSQL databases

- **Apache Hadoop (http://hadoop.apache.org/)** is currently *de facto* standard for ***big data* processing and analysis:**
  - **Core**: HDFS, MapReduce, HBase, Pig, ...
  - **Spark**: SparkML, SparkR, Zeppelin
- **NoSQL databases**, four main families of **high-performance distributed and scalable** databases:
  - *Column store*: Apache Hadoop HBase/Cassandra, Hypertable, ...
  - *Document store*: MongoDB, Solr, ElasticSearch, ...
  - *Key-Value*: Redis, DynamoDB, Riak, ...
  - *Graph*: Neo4J, OrientDB, ...
- New solutions for PB scale **interactive analysis**:
  - *Google Dremel* (Google BigQuery) and similar implementations: new *Hive, Cloudera Impala*
  - Nested data, and comma and tab-separated data, **SQL queries allowed**

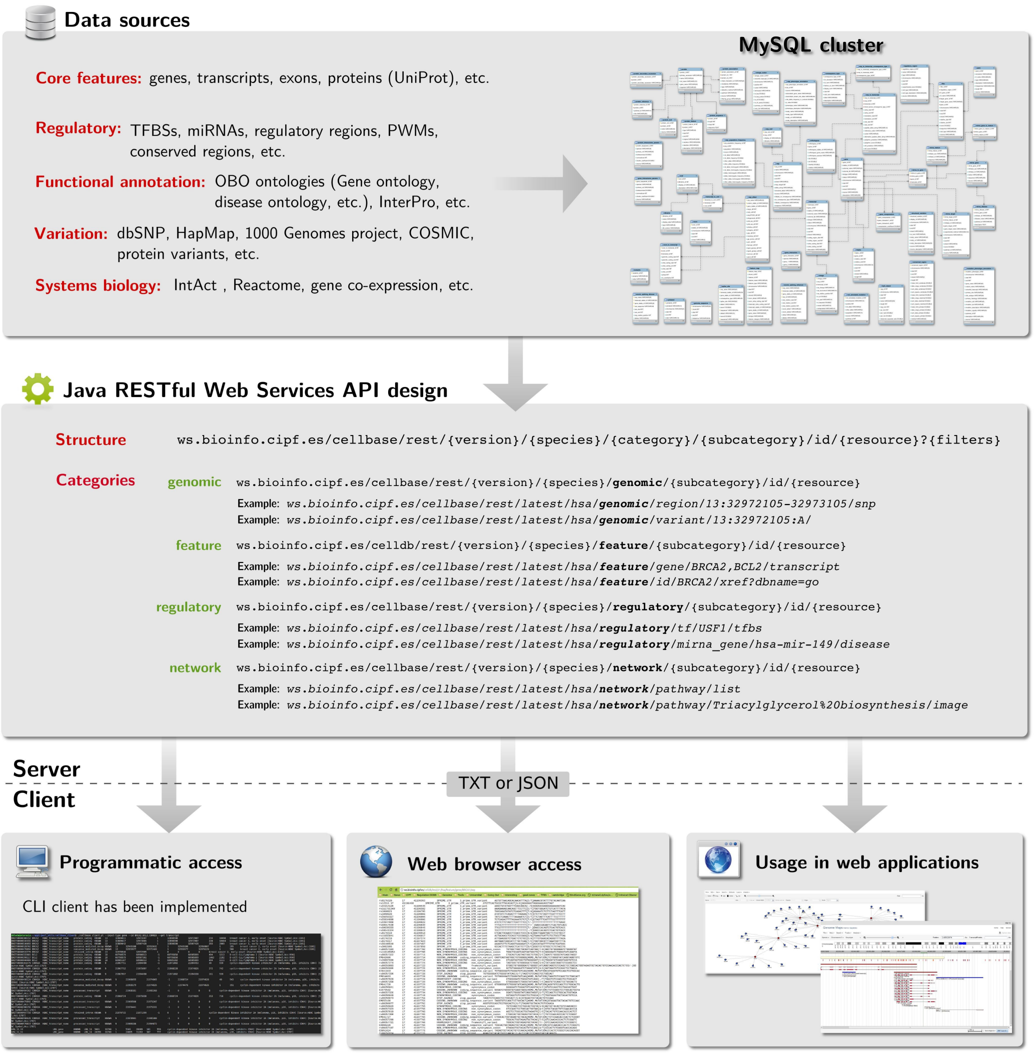
CellBase

An integrative database and RESTful WS API

- ***CellBase*** is a comprehensive integrative NoSQL database and a *RESTful Web Service API*, designed to provide a ***high-performance a scalable*** solution. Currently contains more than 1TB of data:
  - *Ensembl Core features*: genome sequence, genes, transcripts, exons, gene expression, ...
  - Protein: UniProt, Interpro
  - *Variation*: dbSNP and Ensembl Variation, HapMap, 1000Genomes, Cosmic, ClinCar, ...
  - *Functional*: 40 OBO ontologies(Gene Ontology), Interpro domains, ...
  - *Regulatory*: TFBS, miRNA targets, conserved regions, CTCF, histones, …
  - *Systems biology*: Reactome, Interactome (IntAct)
- Published at NAR 2012:
  - http://nar.oxfordjournals.org/content/40/W1/W609
- Used by EMBL-EBI, ICGC, GEL, … among others

Project: https://github.com/opencb/cellbase

Wiki: https://github.com/opencb/cellbase/wiki



Data sources

MySQL cluster

Core features: genes, transcripts, exons, proteins (UniProt), etc.

Regulatory: TFBSs, miRNAs, regulatory regions, PWMs, conserved regions, etc.

Functional annotation: OBO ontologies (Gene ontology, disease ontology, etc.), InterPro, etc.

Variation: dbSNP, HapMap, 1000 Genomes project, COSMIC, protein variants, etc.

Systems biology: IntAct , Reactome, gene co-expression, etc.

Java RESTful Web Services API design

Structure `ws.bioinfo.cipf.es/cellbase/rest/{version}/{species}/{category}/{subcategory}/id/{resource}?{filters}`

Categories

genomic `ws.bioinfo.cipf.es/cellbase/rest/{version}/{species}/genomic/{subcategory}/id/{resource}`

Example: `ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/genomic/region/13:32972105-32973105/snp`

Example: `ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/genomic/variant/13:32972105:A/`

feature `ws.bioinfo.cipf.es/celldb/rest/{version}/{species}/feature/{subcategory}/id/{resource}`

Example: `ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/feature/gene/BRCA2,BCL2/transcript`

Example: `ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/feature/id/BRCA2/xref?dbname=go`

regulatory `ws.bioinfo.cipf.es/cellbase/rest/{version}/{species}/regulatory/{subcategory}/id/{resource}`

Example: `ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/regulatory/tf/USF1/tfbs`

Example: `ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/regulatory/mirna_gene/hsa-mir-149/disease`

network `ws.bioinfo.cipf.es/cellbase/rest/{version}/{species}/network/{subcategory}/id/{resource}`

Example: `ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/network/pathway/list`

Example: `ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/network/pathway/Triacylglycerol%20biosynthesis/image`

Server

TXT or JSON

Client

Programmatic access

CLI client has been implemented

Web browser access

Usage in web applications

CellBase
New features in version 3.x

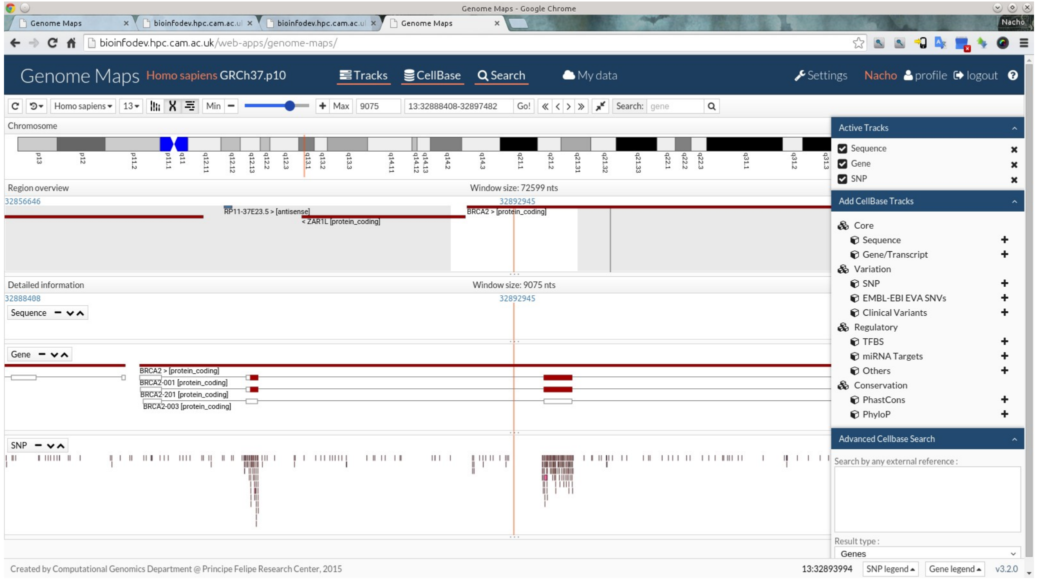
- CellBase v3.0 moved to **MongoDB** NoSQL database to provide a much higher performance and scalability, most queries run in <40ms. Installed at University of Cambridge and EMBL-EBI, some links:
  - GitHub https://github.com/opencb/cellbase
  - Wiki https://github.com/opencb/cellbase/wiki
  - Official domain and Swagger http://bioinfo.hpc.cam.ac.uk/cellbase/webservices/
- **Variant annotation** integrated:
  - http://bioinfo.hpc.cam.ac.uk/cellbase/webservices/rest/v3/hsapiens/genomic/variant/19:45411941:T:C/full_annotation
- About 30 species and much more data available. Coming features in v3.2:
  - Focus on Clinical data
  - Many more species
  - Aggregation and stats
  - R and Python clients
  - Richer and scalable API

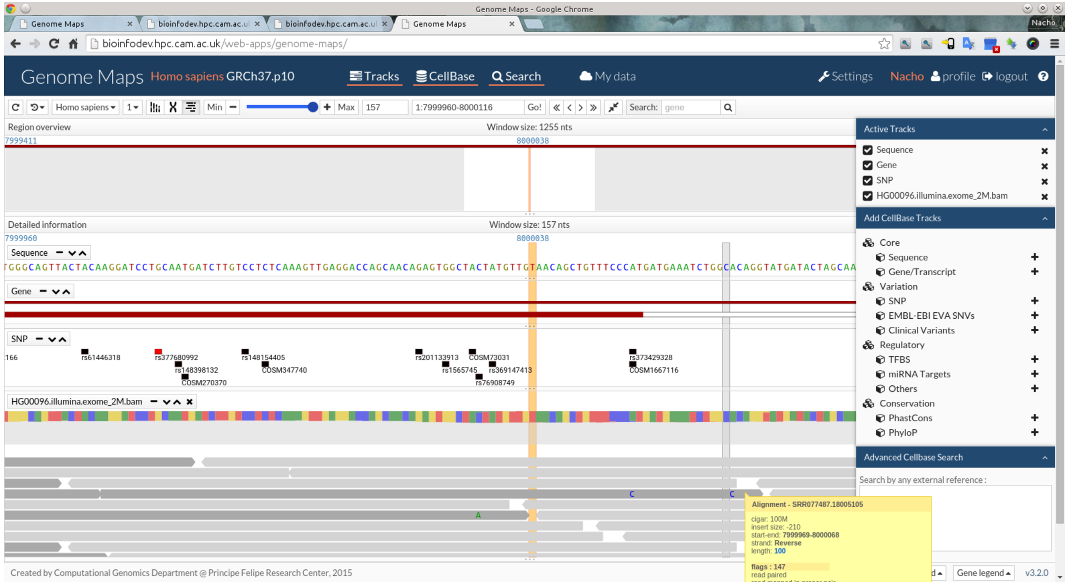
Genome Maps

A *big data* HTML5+SVG genome browser

- Genome scale data ***visualization*** is an important part of the data analysis: *Do not move data!*

- Main features of ***Genome Maps*** (www.genomemaps.org, *published at NAR 2013*)
  - 100% HTML5 web based: ***HTML5+SVG, and other JavaScript libraries.*** Always updated, **no browser plugins needed**
  - Genome data is mainly consumed from ***CellBase and OpenCGA*** database through **RESTful web services**. **JSON** data is parsed and SVG is rendered, making server lighter and improve network transfers
  - Other features: NGS data viewer, Multi species, Feature caches, API oriented, embeddable, key navigation, …
  - Beta: http://bioinfodev.hpc.cam.ac.uk/web-apps/genome-maps/



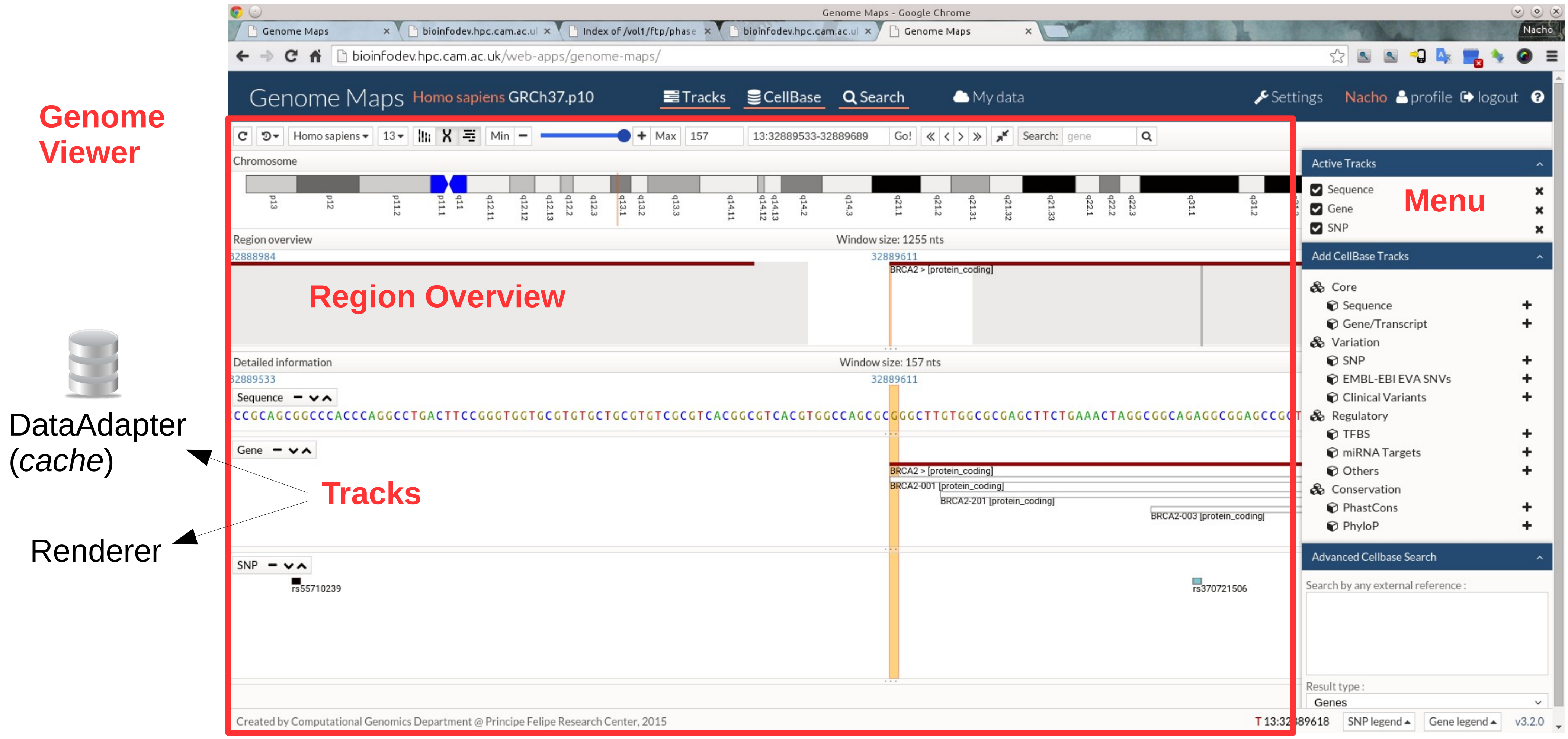


Genome Maps

Software Architecture and Design

Genome Maps design goals are

- provide a high-performance visualization → **JSON data** fetched remotely: CellBase, OpenCGA, EVA, … **no images** sent from the server
- easy to integrate: **event system** and **JavaScript API** developed
- memory efficient: integrated cached (**IndexedDB**), reduces the number of calls
- OpenCB JSorolla JS library



Genome Viewer

DataAdapter (*cache*)

Renderer

Region Overview

Tracks

Menu

Genome Maps

New features and coming releases

- Next version **v3.2** for Summer *2015*
  - More species (~20) and a more efficient ***IndexedDB*** based ***FeatureCache***, less memory footprint and less remote queries
  - More NGS data friendly, better rendering and features for BAM and VCF files
  - More secure, uses HTTPS and can read and **cache encrypted data** (ie. AES) http://bioinfo.cipf.es/apps-beta/genome-maps/encryption/
- Future version **v4.0**, tentative release date during 4*Q15*
  - *JSCircos* for structural variation and other visualizations (http://bioinfo.cipf.es/apps-beta/circular-genome-viewer/)
  - *RNA-seq and* 3D *WebGL* visualization
  - Local data browsing using Docker
  - New Hadoop-based server features being added
- Used by some projects:
  - EMBL-EBI EVA: http://www.ebi.ac.uk/eva/?Home
  - ICGC data portal: http://icgc.org/
  - New genome browser at Peer Bork group http://ct.bork.embl.de/ctbrowser/
  - Lens Beta website: http://patseq.dev.lens.org/lens/

HPG Aligner Suite

NGS pipeline, a HPC implementation

NGS sequencer

HPG suite

High-Performance Genomics

More info at:

http://bioinfo.cipf.es/docs/compbio/projects/hpg/doku.php

Fastq file, up to hundreds of GB per run

QC and preprocessing

QC stats, filtering and preprocessing options

HPG Aligner, short read aligner

Double mapping strategy:
Burrows-Wheeler Transform (*GPU Nvidia CUDA*)
+
Smith-Waterman (*CPU OpenMP+SSE/AVX*)

SAM/BAM file

QC and preprocessing

QC stats, filtering and preprocessing options

Other analysis (*HPC4Genomics consortium*)

RNA-seq (*mRNA sequenced*)
DNA assembly (*not a real analysis*)
Meth-seq
Copy Number
Transcript isoform
...

Variant calling analysis

GATK and SAM mPileup HPC Implementation.
Statistics genomic tests

VCF file

QC and preprocessing

QC stats, filtering and preprocessing options

Variant VCF viewer

HTML5+SVG Web based viewer

HPG Variant, Variant analysis

Consequence type, GWAS, regulatory variants and system biology information

HPG Aligner Suite

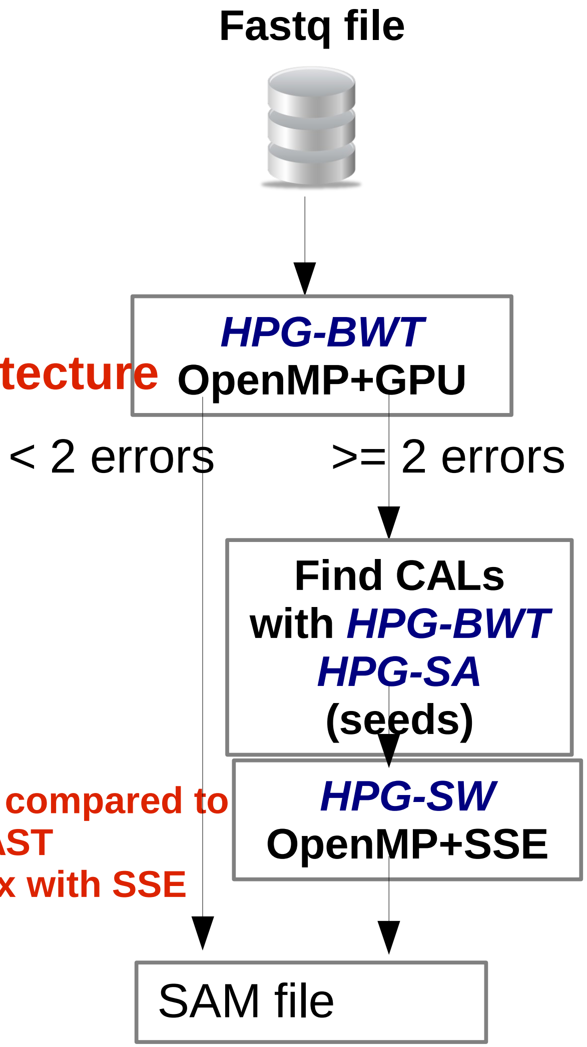
Why another NGS read aligner

- There are more than 70 aligners
  - http://wwwdev.ebi.ac.uk/fg/hts_mappers/
- This project began as a *experimental project,* could a well designed algorithm implemented using modern HPC technologies speed up NGS data mapping?
  - *“It's hardware that makes a machine fast. It's software that makes a fast machine slow”* Craig Bruce
- Focus on ***performance*** and ***sensitivity***. Current software is designed for standard workstations, but ***new projects are processed in clusters***. During these two years we have developed the fastest and one of the most sensitive NGS aligners for DNA and RNA.
- Some computing papers
  - http://dl.acm.org/citation.cfm?id=2223945
  - http://arxiv.org/abs/1304.0681
- Bioinformatics paper: http://bioinformatics.oxfordjournals.org/content/early/2014/09/01/bioinformatics.btu553.long

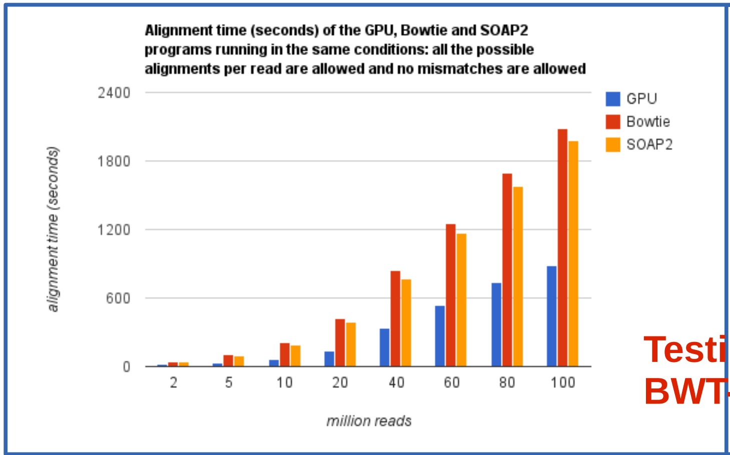
HPG Aligner Suite

The first HPC *DNA* and *RNA-seq* aligner

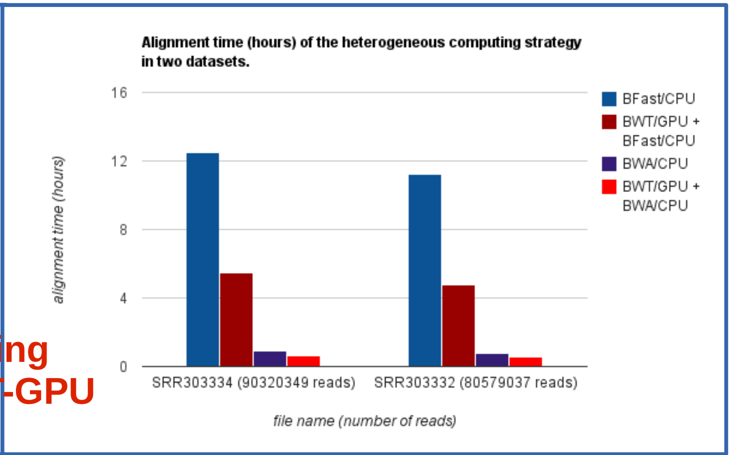
- Many read aligners software tend to fit in one of these groups:
  - *Very fast, but no too sensitive*: no gaps, no indels, rna-seq...
  - *Slow, but very sensitive*: up to 1 day by sample
- Current aligners show **bad performance** with **long reads**
- Current read Aligner algorithms
  - *Burrows-Wheeler 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)*



Architecture



Testing BWT-GPU



10x compared to BFAST
~50x with SSE

HPG Aligner Suite

DNA aligner results, 40 million reads simulation

40 million reads		HPG Aligner 2.0		BWA MEM 0.7.5a		Bowtie2 2.1.0	
Read length (nt)	**Mutation rate (%)**	**Sensitivity (%)**	**Time (min)**	**Sensitivity (%)**	**Time (min)**	**Sensitivity (%)**	**Time (min)**
100	0.1	98.77	20.57	96.99	29.34	94.67	29.40
	1	98.22	19.66	96.65	33.34	92.98	29.15
150	0.1	99.54	22.90	98.09	43.35	96.71	47.61
	1	99.29	22.09	97.96	49.12	95.93	46.50
400	0.1	99.93	31.35	99.12	124.16	98.82	209.26
	1	99.78	30.49	99.06	142.81	98.71	221.92
800	0.1	99.95	35.57	99.42	279.54	99.29	4,604.90
	1	99.74	35.00	99.38	312.55	99.24	2,750.26
2000	0.1	99.93	51.77	99.68	761.37	-	-
	1	99.65	54.08	99.63	814.44	-	-
5000	0.1	99.91	108.81	99.85	1,914.43	-	-
	1	99.64	143.41	99.80	2,034.93	-	-

Sensitivity shows correct alignments

- **Suffix Arrays** used instead of BWT
- Similar results are obtained with real datasets
- Tests performed in a 12-*cores* Intel Xeon E5645, No GPUs were used
- Long reads (~KBs) and *INDELS* supported

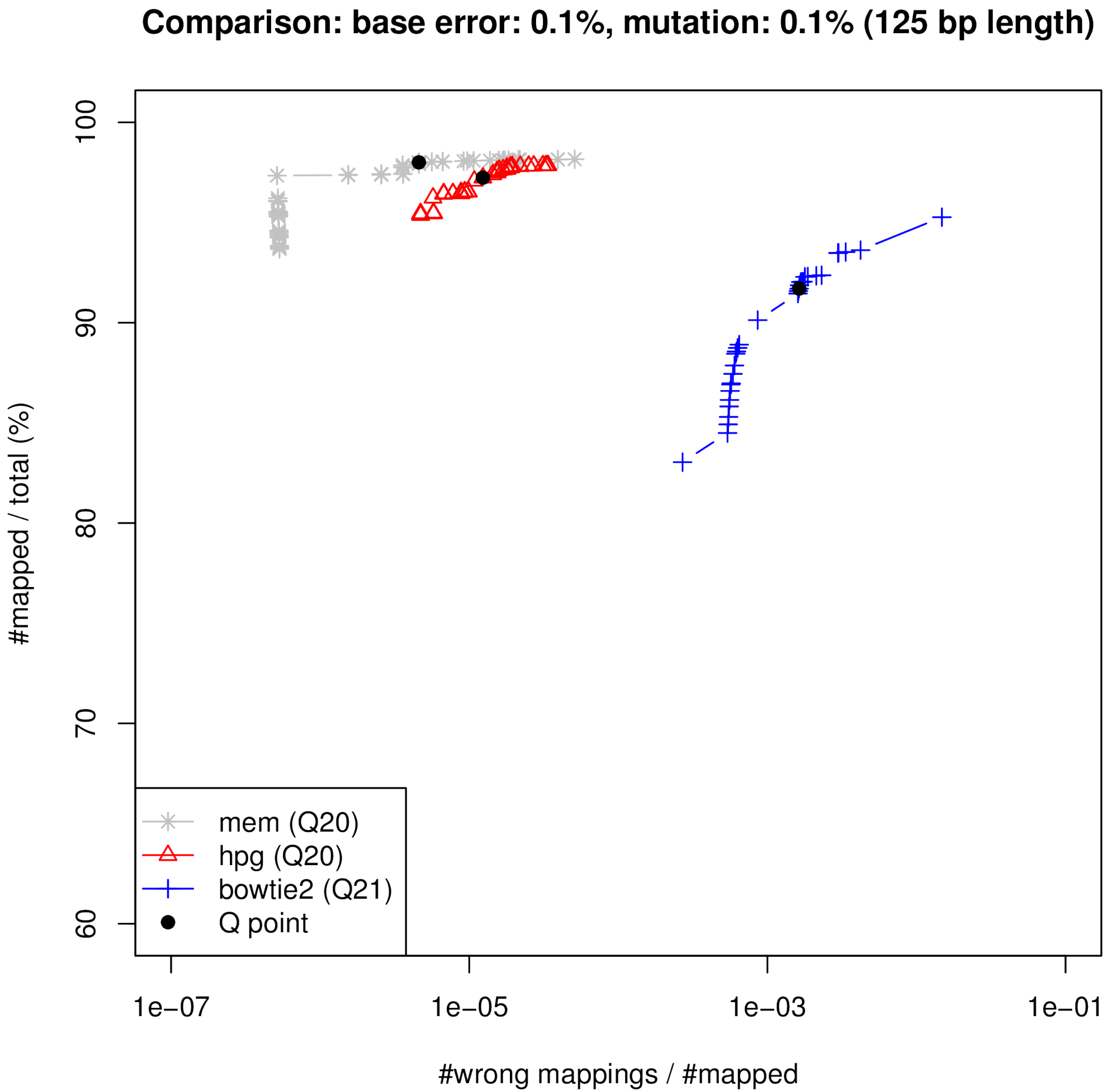
Paper accepted in Bioinformatics!

Coming soon:

- Hadoop MR based version using JNI
- Better INDEL support
- Support GRCh38

HPG Aligner Suite

DNA aligner results, quality similar to new BWA-MEM



Similar sensitivity than BWA-MEM but **3-5x times faster**

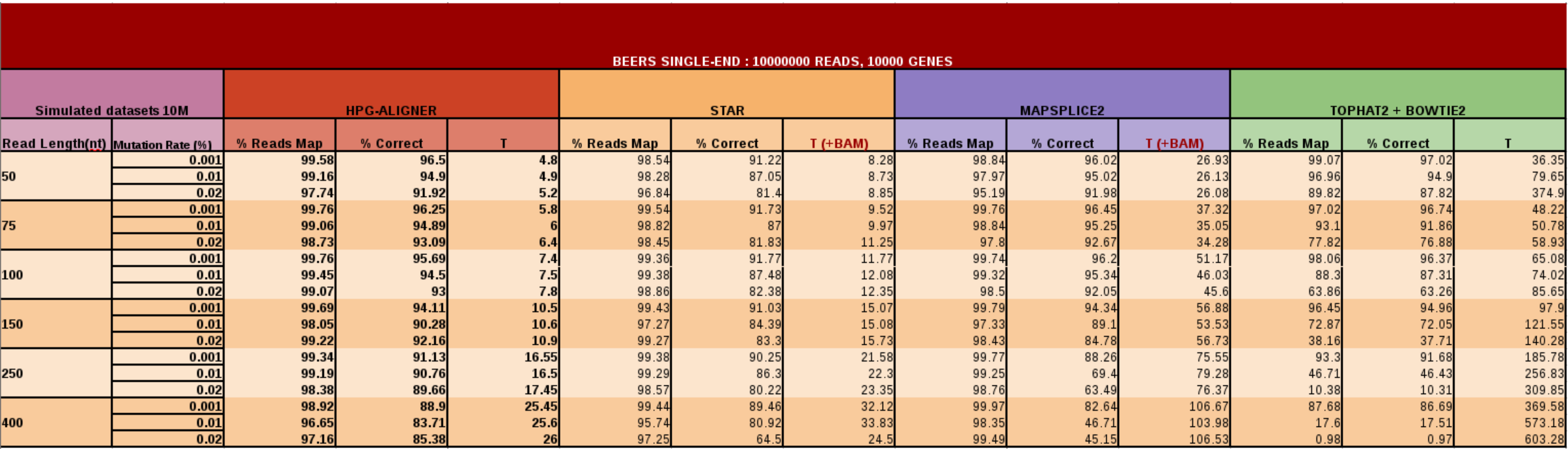
Other features:

- Adaptor support
- INDEL realignment
- Base recalibration

HPG Aligner Suite

RNA-seq gap aligner results, 10M reads simulation

10M reads simulated with *BEERS*, a RNA-seq simulator. Great ***sensitivity*** and ***performance***



BEERS SINGLE-END : 10000000 READS, 10000 GENES														
Simulated datasets 10M		HPG-ALIGNER			STAR			MAPSPLICE2			TOPHAT2 + BOWTIE2		
Read Length(nt)	Mutation Rate (%)	% Reads Map	% Correct	T	% Reads Map	% Correct	T (+BAM)	% Reads Map	% Correct	T (+BAM)	% Reads Map	% Correct	T
50	0.001	99.58	96.5	4.8	98.54	91.22	8.28	98.84	96.02	26.93	99.07	97.02	36.35
	0.01	99.16	94.9	4.9	98.28	87.05	8.73	97.97	95.02	26.13	96.96	94.9	79.65
	0.02	97.74	91.92	5.2	96.84	81.4	8.85	95.19	91.98	26.08	89.82	87.82	374.9
75	0.001	99.76	96.25	5.8	99.54	91.73	9.52	99.76	96.45	37.32	97.02	96.74	48.22
	0.01	99.06	94.89	6	98.82	87	9.97	98.84	95.25	35.05	93.1	91.86	50.78
	0.02	98.73	93.09	6.4	98.45	81.83	11.25	97.8	92.67	34.28	77.82	76.88	58.93
100	0.001	99.76	95.69	7.4	99.36	91.77	11.77	99.74	96.2	51.17	98.06	96.37	65.08
	0.01	99.45	94.5	7.5	99.38	87.48	12.08	99.32	95.34	46.03	88.3	87.31	74.02
	0.02	99.07	93	7.8	98.86	82.38	12.35	98.5	92.05	45.6	63.86	63.26	85.65
150	0.001	99.69	94.11	10.5	99.43	91.03	15.07	99.79	94.34	56.88	96.45	94.96	97.9
	0.01	98.05	90.28	10.6	97.27	84.39	15.08	97.33	89.1	53.53	72.87	72.05	121.55
	0.02	99.22	92.16	10.9	99.27	83.3	15.73	98.43	84.78	56.73	38.16	37.71	140.28
250	0.001	99.34	91.13	16.55	99.38	90.25	21.58	99.77	88.26	75.55	93.3	91.68	185.78
	0.01	99.19	90.76	16.5	99.29	86.3	22.3	99.25	69.4	79.28	46.71	46.43	256.83
	0.02	98.38	89.66	17.45	98.57	80.22	23.35	98.76	63.49	76.37	10.38	10.31	309.85
400	0.001	98.92	88.9	25.45	99.44	89.46	32.12	99.97	82.64	106.67	87.68	86.69	369.58
	0.01	96.65	83.71	25.6	95.74	80.92	33.83	98.35	46.71	103.98	17.6	17.51	573.18
	0.02	97.16	85.38	26	97.25	64.5	24.5	99.49	45.15	106.53	0.98	0.97	603.28

Notes:

- *Correct* mapped results shown
- Similar results are obtained with other NGS simulators and real datasets
- HPG Aligner show the less memory usage
- High accuracy in splice junction detection, *metaexon* structure
- Long reads (~Kbs) and *INDELS* supported
- Tests performed in a 12-*cores* Intel Xeon E5645, No GPUs were used

Coming soon:

- New Suffix Array index in version 2.0, first results show and imprtant speed up and better sensitivity
- Hadoop MR based version using JNI

HPG Aligner Suite

Main and coming features

- Part of the HPG Aligner suite (http://www.opencb.org/technologies/hpg) with other tools: *hpg-fastq*, *hpg-bam and hpg-aligner*
- Usability: the two aligners under the same binary, only one execution is needed to generate the BAM output file, faster index creator, multi-core implementation
- Focused on providing the best sensitivity and the best performance by using HPC technologies: multicore, SSE4/AVX2, GPUs, MPI, …
- Part of OpenCB (http://www.opencb.org). Open-source code and open development, code at GitHub https://github.com/opencb
- Coming features
  - Smith-Waterman in AVX2 implementation and studying Xeon Phi
  - Initial support for **GRCh38** with ALTs
  - BS-seq released: for methylation analysis (being tested)
  - **Apache Hadoop implementation** will allow to run it in a distributed environment
  - **New SA index** (not BWT) in version 2.0 for performance improvements, more memory needed but first results show a speed-up of 2-4x. i.e. more than 10 billion exact reads of 100nt aligned in 1 hour in a 12-cores node.

HPG VARIANT

A suite of tools for variant analysis

- *HPG Variant*, a suite of tools for HPC-based genomic variant analysis
  - **VARIANT** = ***VARI***ant ***AN***alysis ***T***ool
- Three tools are already implemented: ***vcf***, ***gwas*** and ***effect.*** Implemented using *OpenMP*, SSE/AVX, *Nvidia CUDA* and *MPI* for large clusters. Hadoop version coming soon.
- ***VCF****: C library and tool*: allows to analyze large VCFs files with a low memory footprint: stats, filter, split, ***merge***, … (*paper in preparation*)
  - Example: *hpg-variant vcf –stats –vcf-file ceu.vcf*
- ***GWAS***: suite of tools for gwas variant analysis (~*Plink*)
  - Genetic tests: association, TDT, Hardy-Weinberg, ...
  - ***Epistasis***: HPC implementation with SSE4 and MPI, 2-way 420K SNPs epistasis in 9 days in a 12-core node.
  - Example: *hpg-variant gwas –tdt –vcf-file tumor.vcf*
- ***EFFECT***: A CLI and web application, it's a *cloud*-based genomic variant ***effect*** predictor tool has been implemented (http://variant.bioinfo.cipf.es, *published in NAR 2012*)

HPG BigData

A suite of tools for working in genomics with Hadoop

- Data formats and serialization for all data formats in Avro and Parquet
- Most of the common functionality and analysis implemented with Mapreduce and Spark
- HBase, Hive and Impala also used
- C code embedded using JNI
- Available at https://github.com/opencb/hpg-bigdata

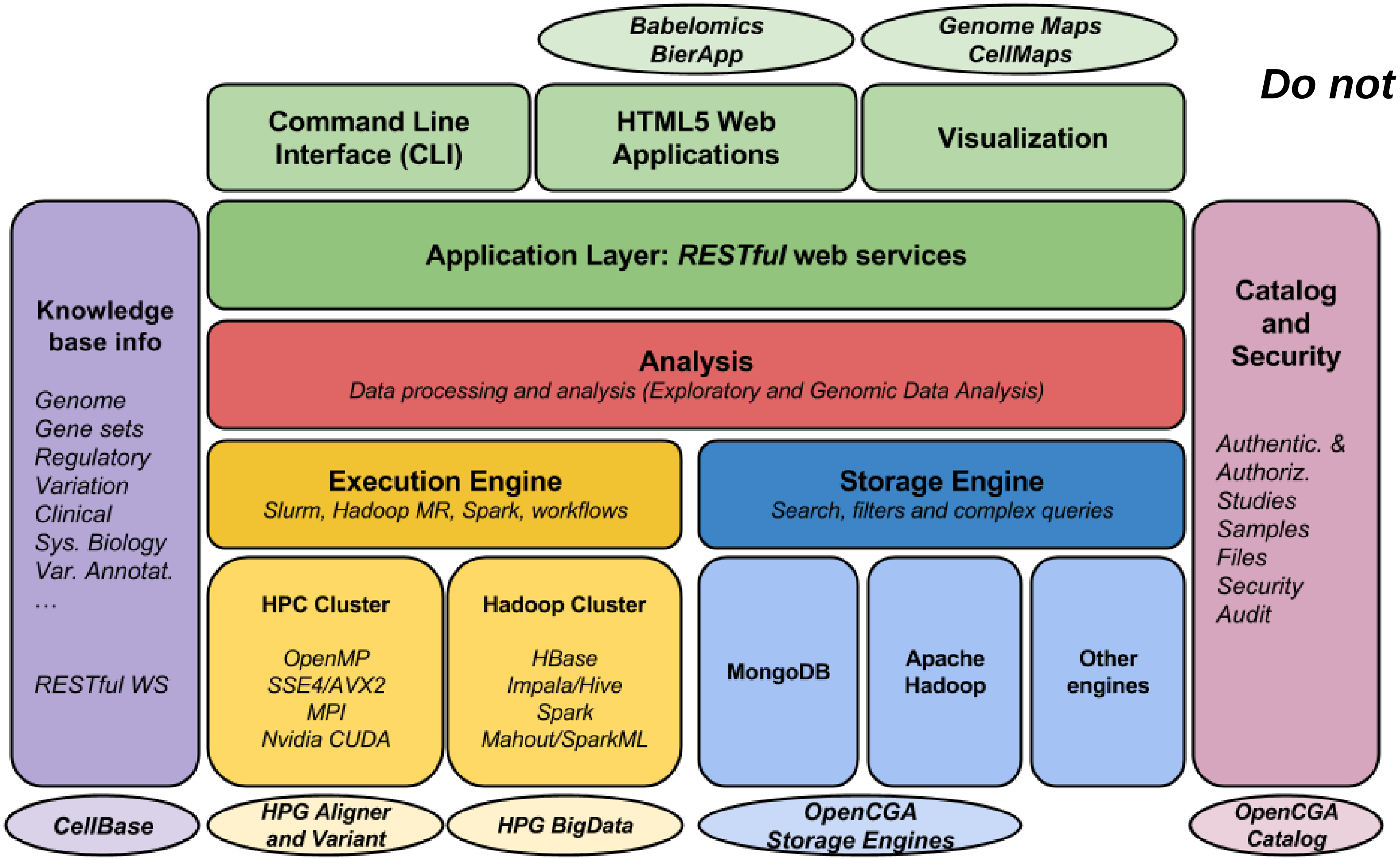
OpenCGA
Overview and goals

Open-source Computational Genomics Analysis (**OpenCGA**) aims to provide to researchers and clinicians a ***high performance and scalable solution*** for genomic big data processing and analysis

OpenCGA is built on OpenCB: CellBase, Genome Maps, Cell Maps, HPG Aligner, HPG BigData, Variant annotation

Project at GitHub: **https://github.com/opencb/opencga**

Do not move data

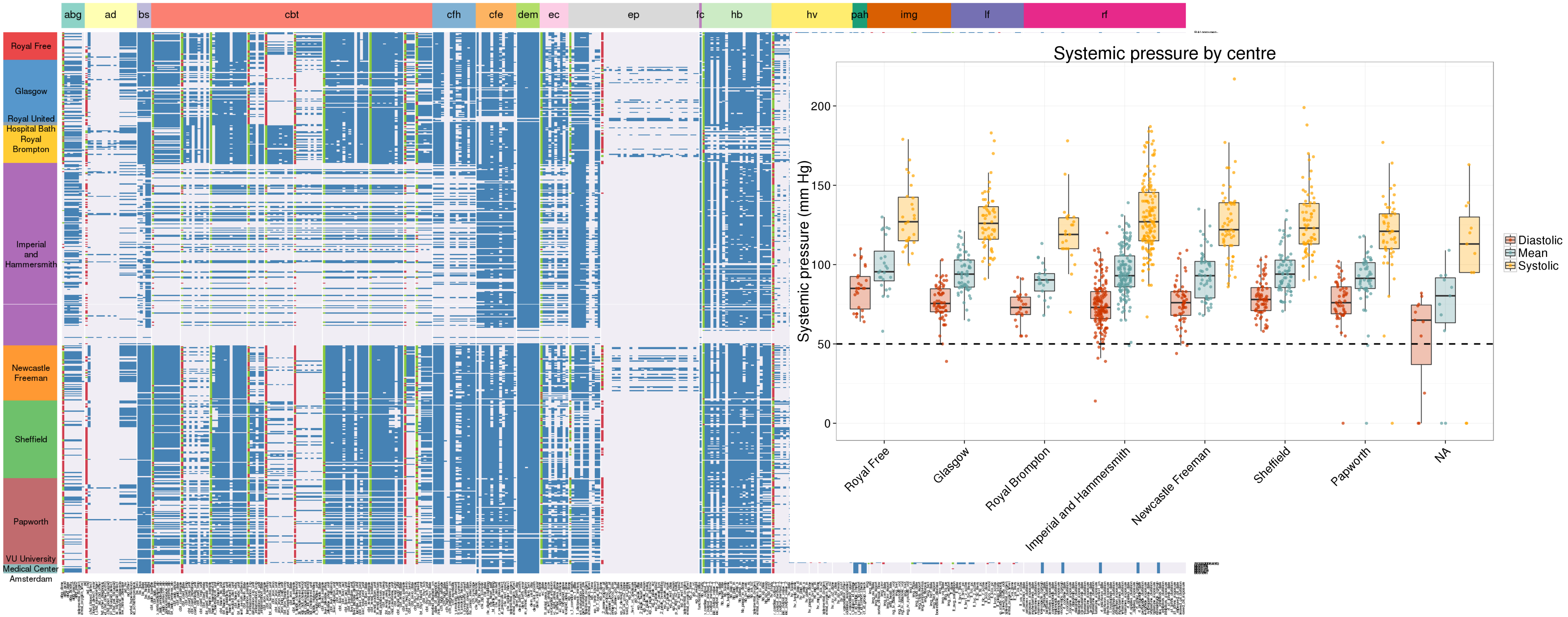


OpenCGA

Metadata Catalog

- **OpenCGA Catalog** provides user authentication, authorization, *sample annotation*, file and job tracking, audit, …
- Allow to the different storage engines to perform optimizations
- ***Sample annotations*** is probably the main feature:
  - Allow complex queries and aggregations
  - Allow to detect bias and other problems with the data

R prototypes, interactive web-based plots being implemented



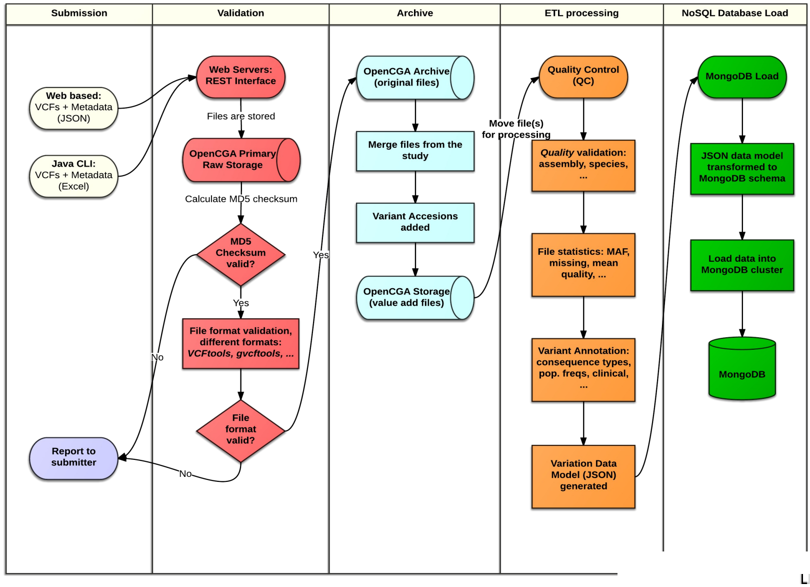
OpenCGA
Storage Engines

OpenCGA Storage provides a *pluggable* Java framework for storing and querying alignment and variant data

- Data is processed, normalized, **annotated and indexed**, also some **stats are precomputed and indexed**

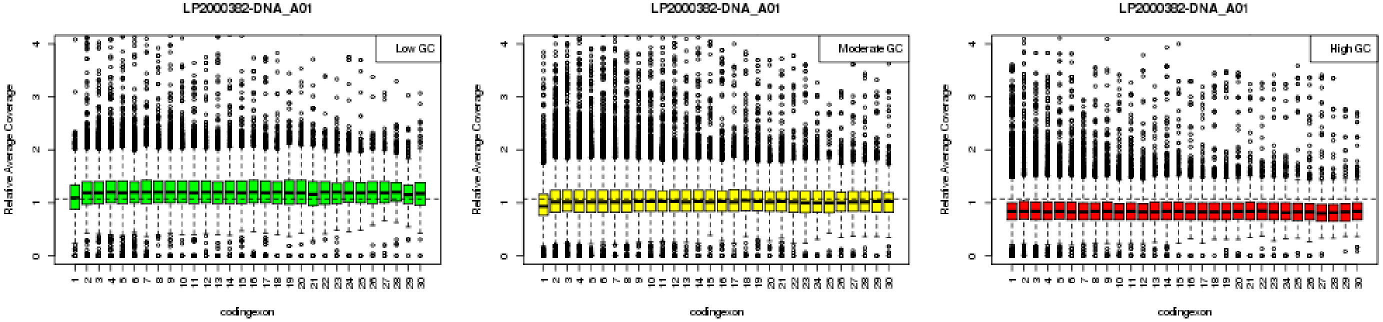
Two implementations are delivered by default with OpenCGA: **MongoDB** and **Hadoop**. Huge performance and scalability. Hadoop-based solution **~500k inserts/second**

Highly customizable and easy to extend (ie. only two Java classes are needed to be implemented)


Submission
Validation
Archive
ETL processing
NoSQL Database Load
Web based: VCFs + Metadata (JSON)
Java CLI: VCFs + Metadata (Excel)
Web Servers: REST Interface
Files are stored
OpenCGA Primary Raw Storage
Calculate MD5 checksum
MD5 Checksum valid?
Yes
No
File format validation, different formats: VCFtools, gvcftools, ...
File format valid?
Yes
No
Report to submitter
OpenCGA Archive (original files)
Merge files from the study
Variant Accesions added
OpenCGA Storage (value add files)
Move file(s) for processing
Quality Control (QC)
Quality validation: assembly, species, ...
File statistics: MAF, missing, mean quality, ...
Variant Annotation: consequence types, pop. freqs, clinical, ...
Variation Data Model (JSON) generated
MongoDB Load
JSON data model transformed to MongoDB schema
Load data into MongoDB cluster
MongoDB

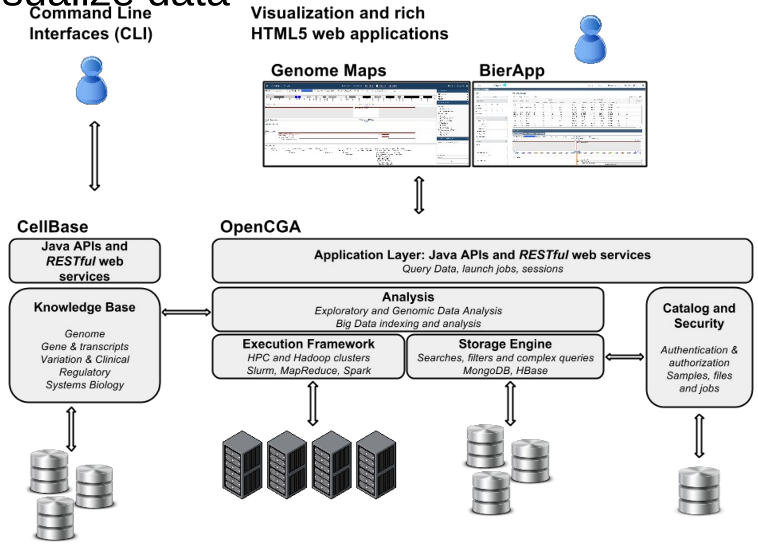
Variant pipeline ETL

Precomputed data allows interactive analysis



OpenCGA
Summary

- *HPC and Hadoop*-based environment for *big data* storage and analysis being developed: *OpenCGA*
  - *Storage*: efficient storage and data retrieval of hundreds TB. Moving to *Hadoop for scaling to PBs*
  - *Analysis and workflows*: many tools already packaged (aligners, GATK, …), users can upload their tools to extend functionality, SGE queue, … *Hadoop MR and HPC framework*
  - *Search and access*: data is indexed and can be queried efficiently, RESTful WS allows users to access to data and analysis programatically
  - *Sharing and security*: users can share their data and analysis, public and private data. Encryption is also integrated.
  - *Visualization*: HTML5-SVG based web applications to visualize data
- Used by Genome Maps, EVA, GEL or BierApp


Command Line Interfaces (CLI)
Visualization and rich HTML5 web applications
Genome Maps
BierApp
CellBase
Java APIs and RESTful web services
Knowledge Base
Genome
Gene & transcripts
Variation & Clinical
Regulatory
Systems Biology
OpenCGA
Application Layer: Java APIs and RESTful web services
Query Data, launch jobs, sessions
Analysis
Exploratory and Genomic Data Analysis
Big Data indexing and analysis
Execution Framework
HPC and Hadoop clusters
Slurm, MapReduce, Spark
Storage Engine
Searches, filters and complex queries
MongoDB, HBase
Catalog and Security
Authentication & authorization
Samples, files and jobs

HPCS-Dell Genomics

New projects coming soon

- Current collaboration with Dell to port some big data processing and algorithms to Dell Statistica
- Proof of concept being developed for next year

Acknowledgements

- Bosses
  - **Paul Calleja**, Director of HPC Service, University of Cambridge
  - **Augusto Rendon**, Director of Bioinformatics of Genomics England
- Team and core OpenCB developers:
  - Jacobo Coll (*Genomics England*)
  - Javier Lopez (*EMBL-EBI*)
  - Joaquin Tarraga (*CIPF*)
  - Cristina González (*EMBL-EBI*)
  - Jose M. Mut (*EMBL-EBI*)
  - Francisco Salavert (*CIPF*)
  - Matthias Haimel (*Addenbrooke's*)
  - Marta Bleda (*Addenbrooke's*)
- Collaborators
  - **Joaquin Dopazo** (*CIPF*)
  - Justin Paschall (*EMBL-EBI*)
  - Stefan Gräf (*Addenbrooke's*)
  - UJI: Enrique Quintana, Héctor Martinez, Sergio Barrachina, Maribel Castillo
- Dell
- Cloudera
  - Tom White