Introduction to Big Data Analysis and Visualization

Nacho Medina
imedina@ebi.ac.uk
http://bioinfo.cipf.es/imedina
Project Manager & Senior Software Architect
EBI Variation
EMBL-EBI
(Valencia, Spain)

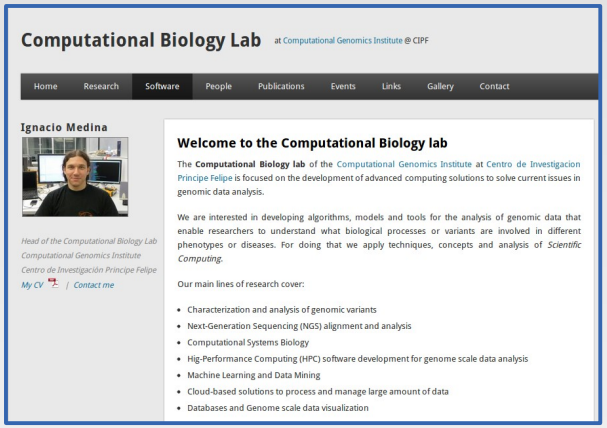
Index

- Introduction
- Genome visualization tools
- *Big data* analytics and visualization
  - Advanced computing technologies
  - OpenCGA
- OpenCB initiative

Introduction

About me

From Head Computational Biology Unit at CIPF (2011-2013) to Project Manager & Software Architect at EBI Variation (2014)
http://bioinfo.cipf.es/compbio



2013-2014

OpenCB, code released as an open-source collaborative project
http://www.opencb.org/

Two backgrounds: *Biochemistry* and *Computer Science*. More than 8 years working in Bioinformatics and ~35 papers. Main interest in *HPC* and *big data* analysis in biology.

Founder of a **Specialized Unit** consisted of **3-4 *Computer Scientists and Bioinformaticians*** such as *Scientific programmers*, *HPC developers* and *Web and Distributed developers* to produce high-performance, efficient and high quality software. Many solutions and papers released, others still under review.

Today *computational requirements* in ***Genomics*** makes necessary more advanced computing solutions. ***OpenCB*** initiative is a young and open-source software platform focused in *big data* mining and analysis (spin off the work done in Joaquin Dopazo's group at CIPF).

OpenCB technologies and solutions are used in different projects such as ***EVA and EGA (EBI)***, ***ICGC (OICR)*** or ***BRIDGE (NIHR)*** among others

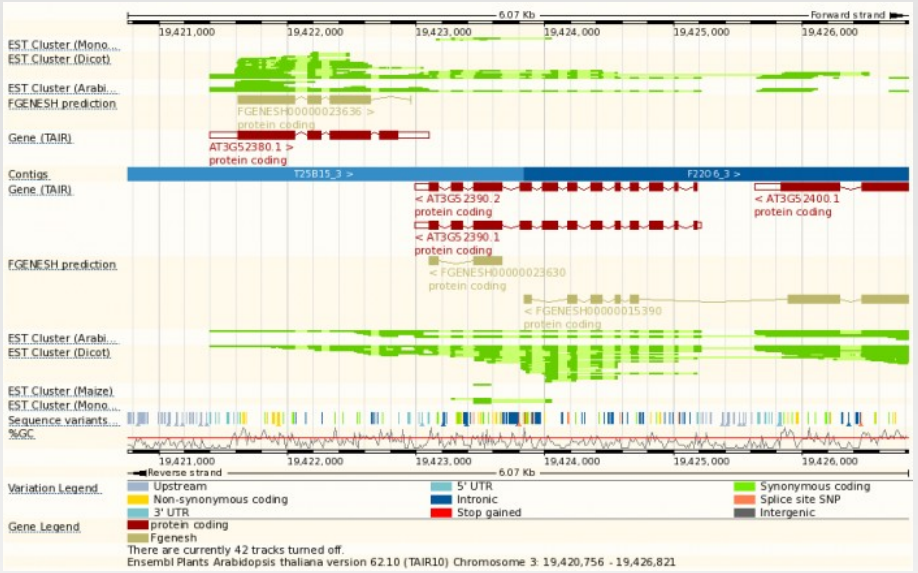
Introduction
Why do we need Genome Visualization?

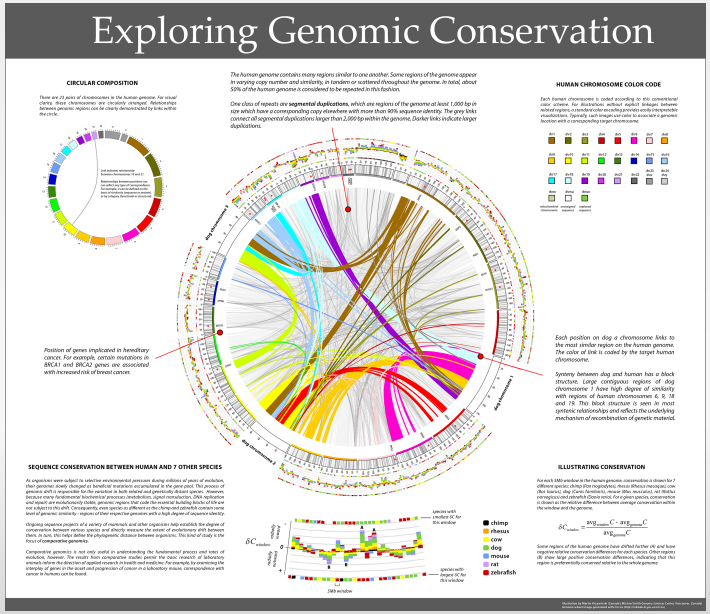
- First sequenced genomes date from ~15 years ago. During these years high-throughput technologies (*microarrays*, *NGS*) have produced a revolution. Today Biology is also now considered a **data science**.
- All these genomes and data require a specialized **visualization software mainly** for ***data analysis***. The new scientific visualization methods, together with the computing hardware and software frameworks developed during these years, have made this possible.
- Two main Genomic visualization uses:
  - the **visualization** and **exploration** of reference genomes and their annotations (gene, SNPs, regulatory elements, ...) Examples: *Ensembl Browser*, *UCSC Genome Browser*, *Genome Maps*, ... and others
  - data manipulation and interpretation during the ***data analysis*** by displaying data from experiments in a meaningful way. **This is probably the most important usage of visualization**. Examples: *IGV*, *Circos*, *Genome Maps*, *inPHAP*, ... many other tools

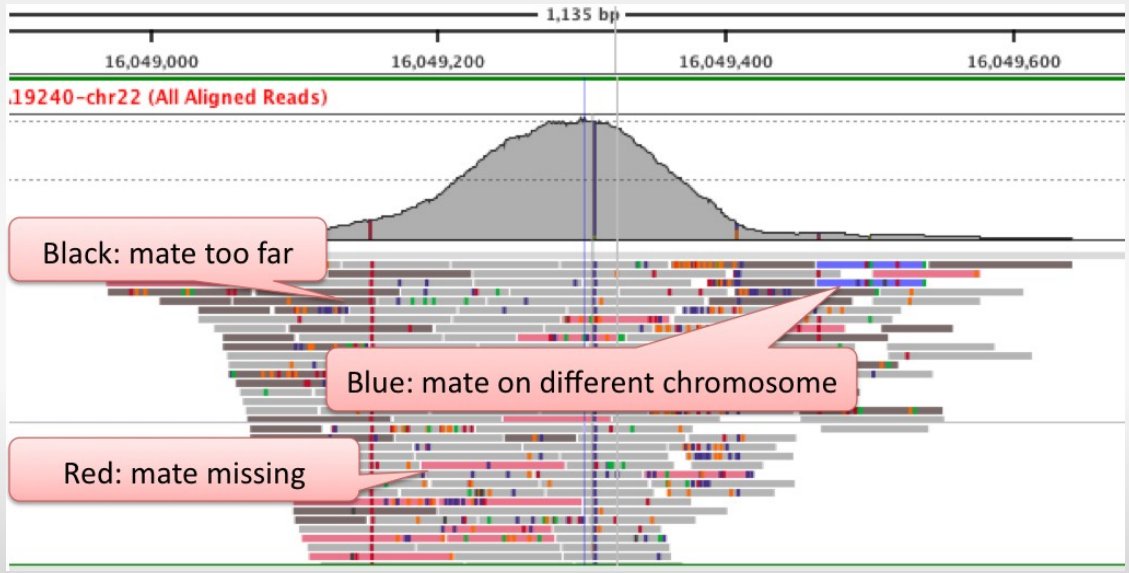
Introduction

Genome visualization applications

- Main of the applications come from ***data analysis***:
  - Genome browsing and annotation
  - Exploration, interpretation and manipulation of data
  - *de novo* sequencing assembly
  - NGS ***read alignments*** and ***variation*** data visualization
  - Comparative genomics, ...







Introduction

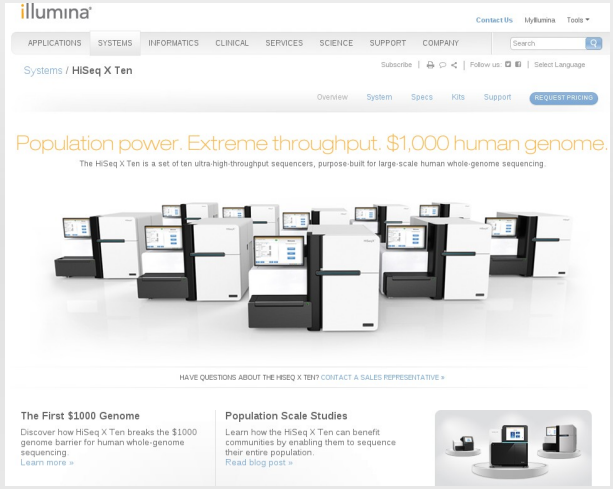
Current status of genome visualization tools

- Many different tools have been developed during the last 10 years, most developed by small teams using old technologies. In general they solve ***specific analysis*** and show ***poor performance*** and ***scalability***.
- Current challenges need bigger teams to develop **advanced** and **useful** solutions to solve current problems in ***data analysis***: *big data visualization*, *security*, *knowledge base*, ... More collaborations and standards are needed
- **Scientific Visualization** interest in biology has risen during the last few years and some groups are working on very interesting and promising projects: *new methods for data visualization*, ***genome visualization***, *big data visualization*, *data integration*, ...
- Since a few years ago **BioVis** congregates many researches interested in visualization:
  - http://www.biovis.net/year/2014/about

So, how are we doing it? Are the Bioinformaticians and Computational Biologists solving the current problems?

Introduction

Big data in Genomics, a new scenario for biologists

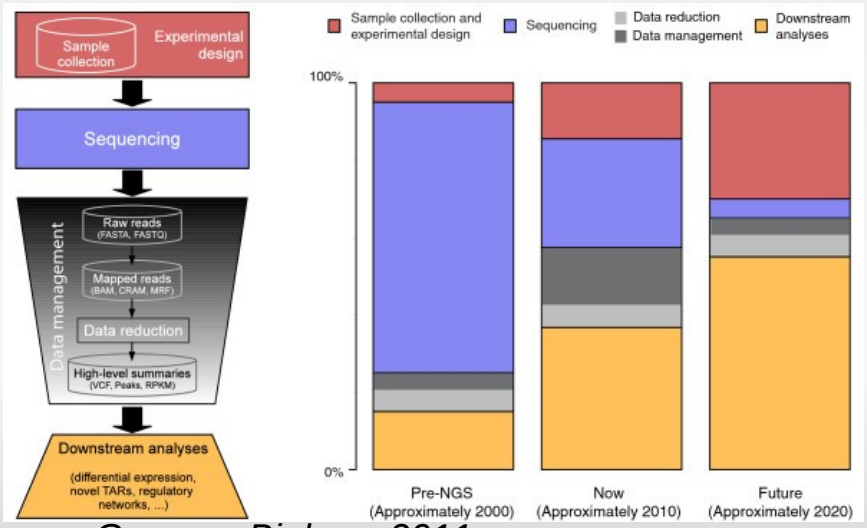


Next-Generation Sequencing (*NGS*) technology is changing the way how researchers perform experiments. Many new experiments are being conducted by sequencing: *re-sequencing, RNA-seq, Meth-seq, ChIP-seq, ...*

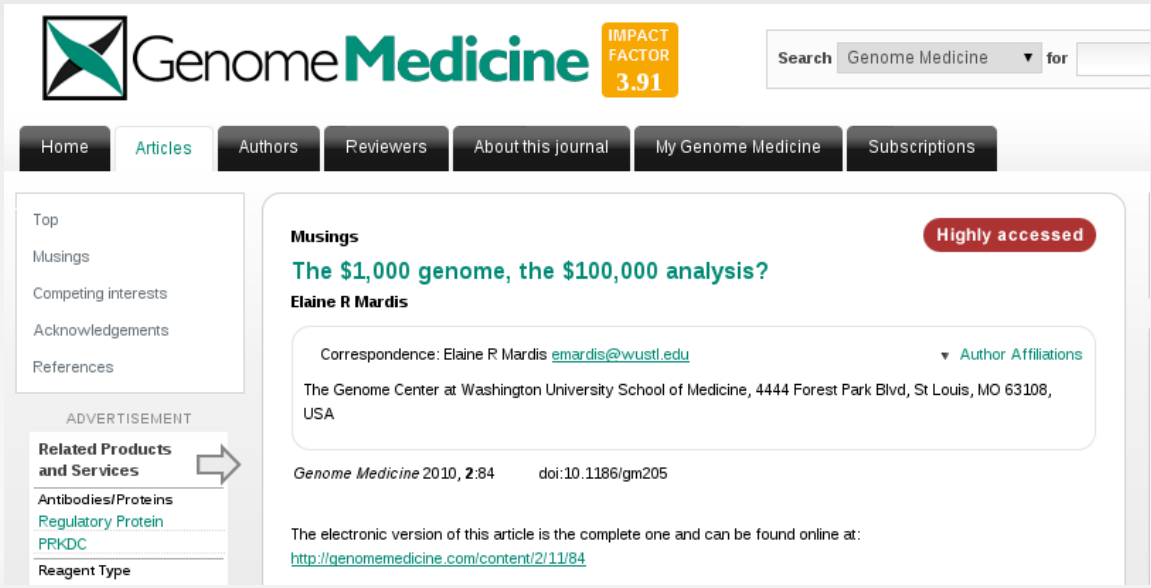
*Experiments **have increased data size by more than 4000x** when compared with old microarrays or first sequencers. Surprisingly, many software solutions are not very different.*

Sequencing costs keep falling, today a whole genome can be sequenced by **$1000**, so much more data is expected

Data processing and analysis are today a **bottleneck** and a nightmare, **from days or weeks with microarrays to months with NGS,** and it will be worse as more data become available



Genome Biology, 2011



It's the analysis, stupid!

Introduction

Big data in Genomics, some projects and data

- At **EMBL-EBI** (http://www.ebi.ac.uk/)
  - ***European Genome-phenome Archive (EGA)***: stores human datasets under ***controlled access*** https://www.ebi.ac.uk/ega/home and ***encrypted***. Current size about **1.5PB**, it is expected to increase 2x-3x over the next few years. No new functionality provided or being implemented.
  - ***European Variation Archive (EVA)***: public and open archive for all genomic variation data for all species http://www.ebi.ac.uk/eva/ A new project with only a few TB of data yet.
  - ***1000G Phase 3***: 2535 individuals from 26 populations, a few hundreds of Tbs
- Other ***big*** projects
  - ***NIHR BRIDGE***: 10,000 whole genomes from rare diseases, ~1-2PB of data expected
  - ***Genomics England (GEL)***: also known as UK100K, is sequencing 100K whole genomes from UK, several diseases and cancers to study, 20-30PB of data expected
  - ***International Cancer Genome Consortium (ICGC)***: store more than 10,000 sequenced cancers, few PB of data
- And many other medium sized projects
  - ***BIER at Ciberer***
  - ...

Introduction
Visualization challenges

- ***Big data***: low prices and new NGS technology are producing high volumes of data, new projects size are in PB scale.
- ***Security concerns***: much of these data must be kept secure (authentication, authorization, encryption, ...)
- ***Data Analysis***: visualization must be useful for data analysis. Real-time and Interactive graphical data analysis.
- ***Performance and scalability***: software must be high-performance and scalable. Take advantage of cloud computing.
- ***Data Integration***: different types of data such as variation, expression, ChIP, ...
- ***Collaboration***: many projects require the collaboration among different groups. Moving data not possible.
- ***Knowledge base and sample annotations***: many of the visual analytic tools need genome and sample annotations

Introduction

Desired features and goals

Which is your feature selection? Do you have the right tools for your research or analysis?

Genome visualization tools

Current browsers for different scenarios

- Many different Genome Visualization applications with different features and usages to solve **specific** problems
- In general
  - developed by small teams using, usually, old technologies
  - poor performance and scalability
  - difficult to install and use
  - no security implemented
  - no collaborative (nor data analysis nor development)
  - ...
- All these leads users to combine several genome browsers, and sometimes to some frustration...

Genome visualization tools

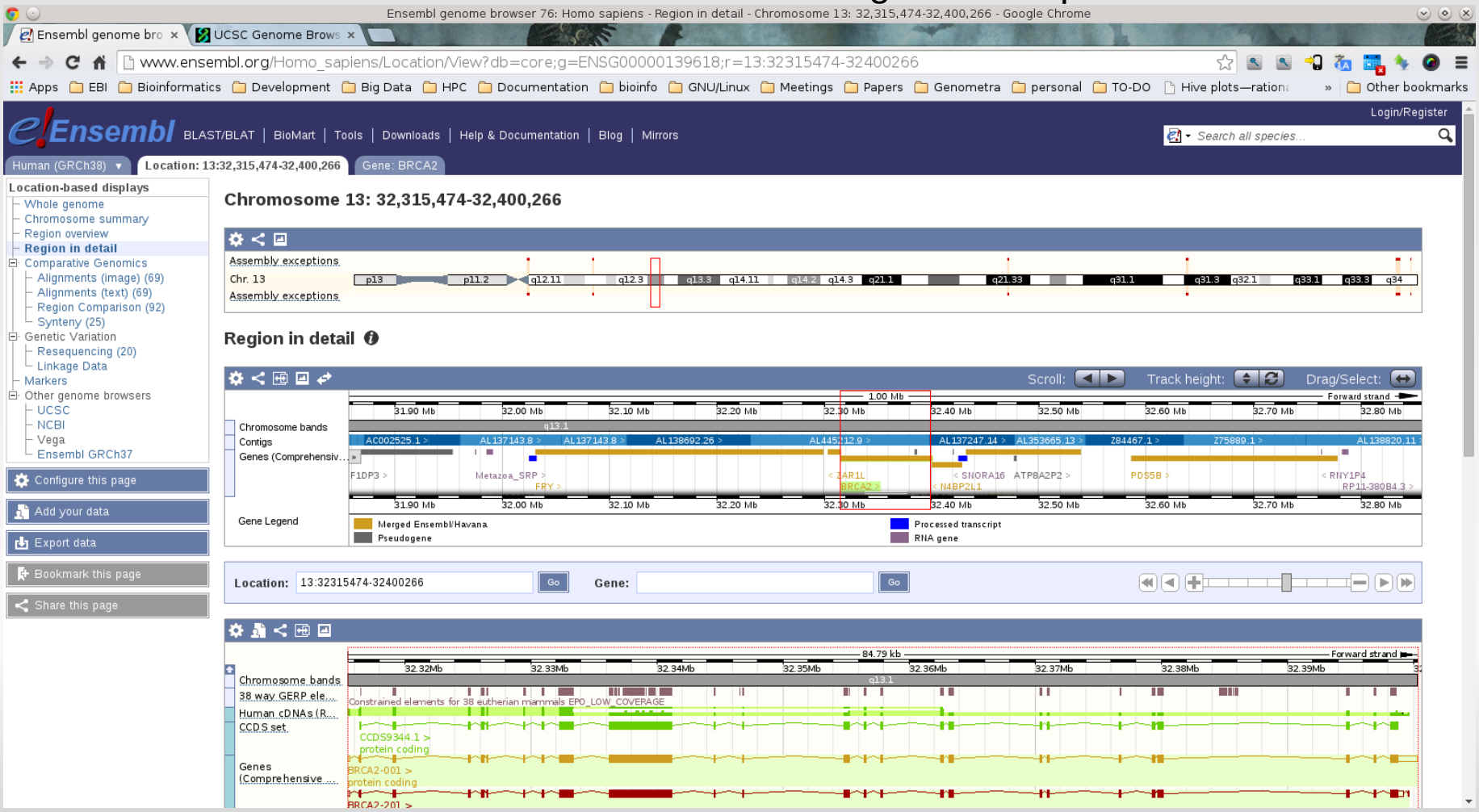
Ensembl Browser

http://www.ensembl.org/

Probably one of the most known and old projects:

- Positive: great knowledge base, many annotations
- Negative: slow, no data analysis, no installable, old technologies, no collaborative, NGS, ...

Good for reference genome exploration and visualization



Genome visualization tools

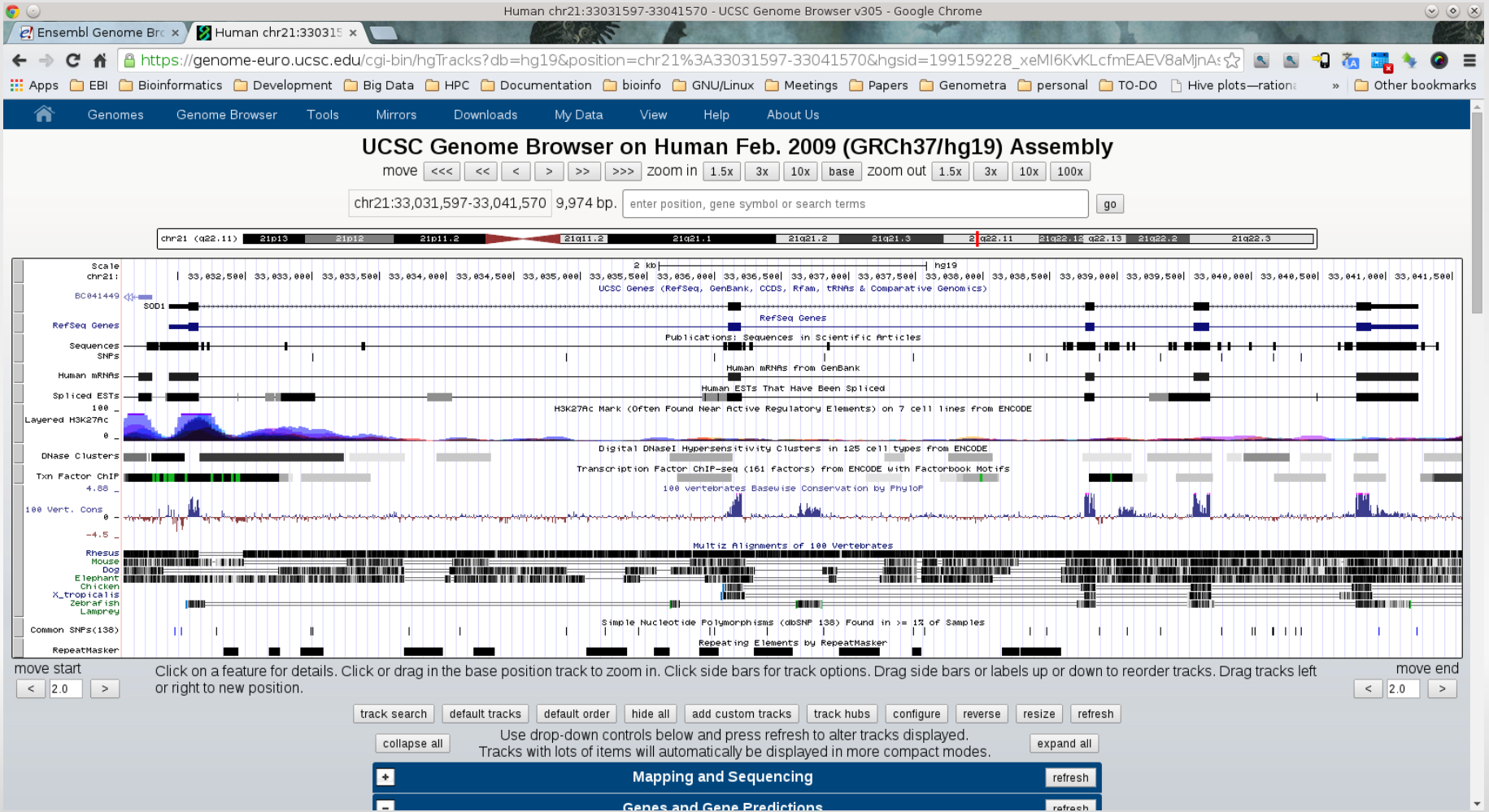
UCSC Genome Browser

https://genome.ucsc.edu/

Probably one of the most known and old projects:

- Positive: great knowledge base, many annotations
- Negative: slow, no data analysis, no installable, old technologies, no collaborative, NGS, ...

Good for reference genome exploration and visualization



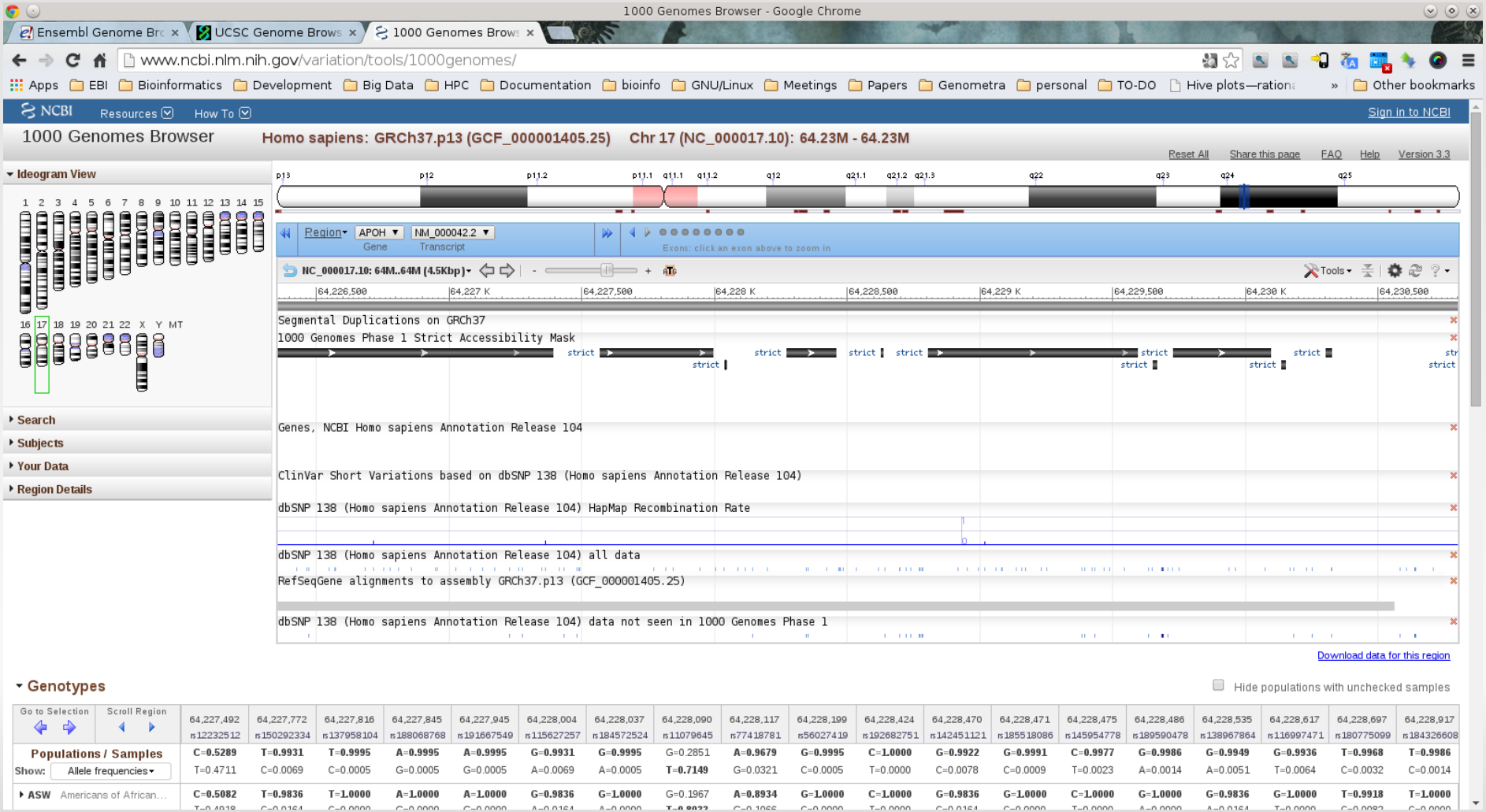
Genome visualization tools

1000G tool

http://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/

1000 genomes project data:

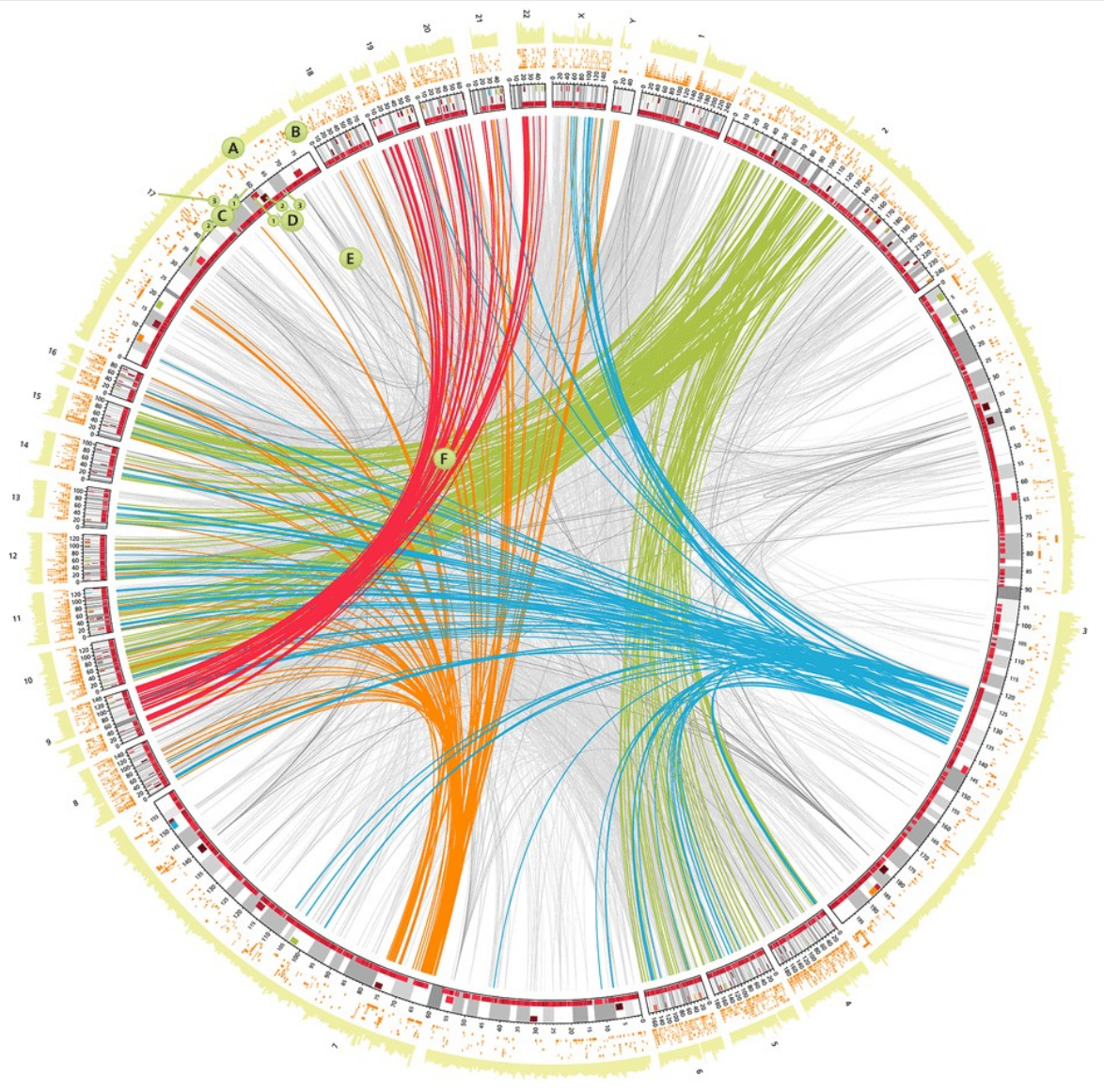
- Positive: ... it can display 1000G data, web based
- Negative: slow, no data analysis, not too many annotations, no collaborative, no installable, ...



Genome visualization tools

Circos

http://circos.ca/



Circos visualization:

- Positive: allows to describe relationships or multi-layered annotations at different scales
- Negative: slow, not interactive data analysis, not too many annotations, no collaborative, hard to use and install, ...

Genome visualization tools

IGV tool

http://www.broadinstitute.org/igv/

Good tool for NGS data analysis:

- Positive: basic data analysis, NGS data, easy to install (Java), ...
- Negative: no scalable, slow, no complex data analysis, not too many annotations, no collaborative, ...

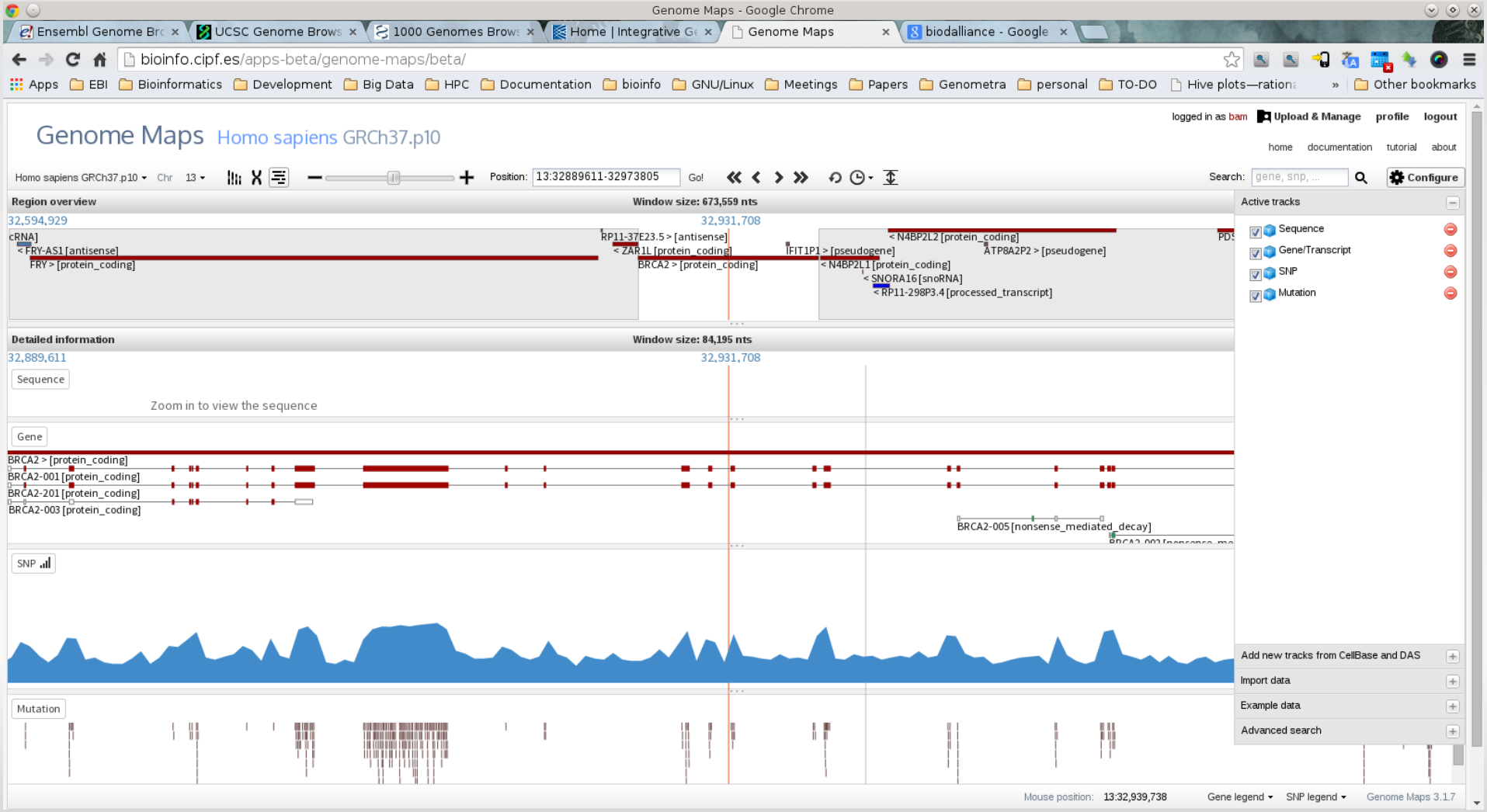
Genome visualization tools

Genome Maps

http://genomemaps.org/

Good tool for exploration and NGS data analysis:

- Positive: basic and complex data analysis, NGS data, easy to install (server in Java), fast, scalable, collaborative, encryption, ...
- Negative: not ready yet



Genome visualization tools
Conclusion

- In general many visualization tools, but most solve specific problems. No high performance, integrative, collaborative, ... poor analysis integration.
  All of them are valuable and are based on very good ideas!!
  But...

So, how are we doing it? Are the Bioinformaticians and Computational Biologists solving the current problems?

- Let's imagine next scenario: 10,000 whole genome sequenced samples with some RNA-seq, about 4PBs of data:
  - Can I easily explore and visualize the data? Can I filter and search in the data?
  - Can I filter variants by some annotations? By Stats? By Consequence type?
  - Can I perform more complex filters and queries? For example: give me all those variants enriched in regulatory elements in the cases over the controls
  - Can I perform some data analysis such as eQTLs or epistasis?
  - Can I share my data? Encrypt? Sample annotation?

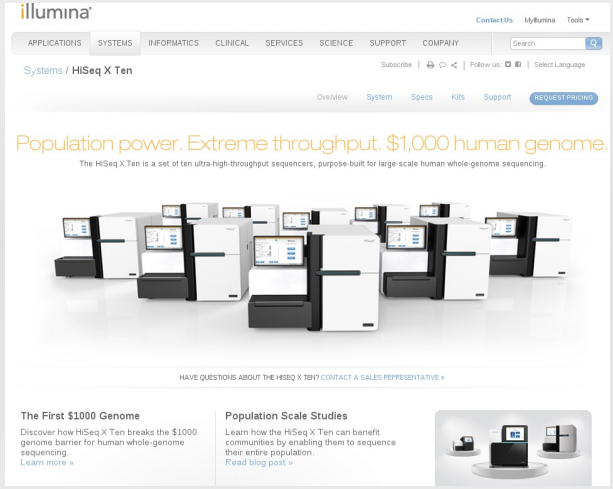
It's the server side!

Big data analytics and visualization

OpenCGA

Introduction

Big data in Genomics, a new scenario for biologists

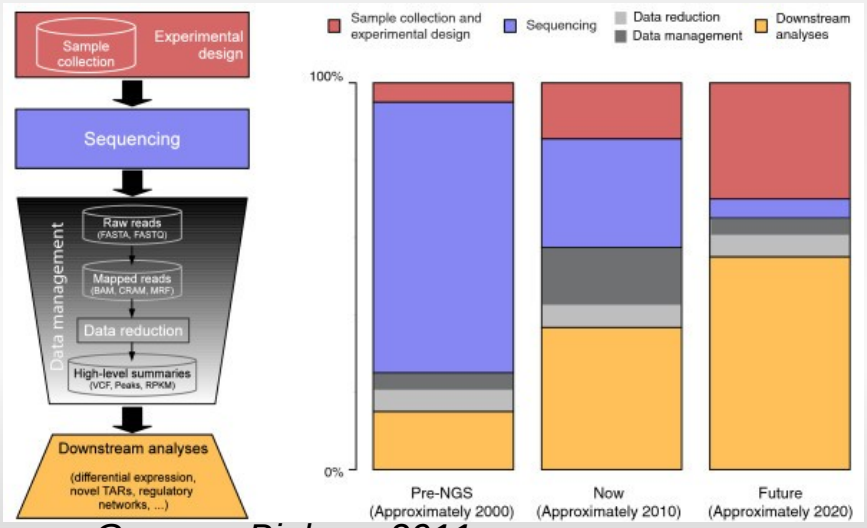


Next-Generation Sequencing (*NGS*) technology is changing the way how researchers perform experiments. Many new experiments are being conducted by sequencing: *re-sequencing, RNA-seq, Meth-seq, ChIP-seq, ...*

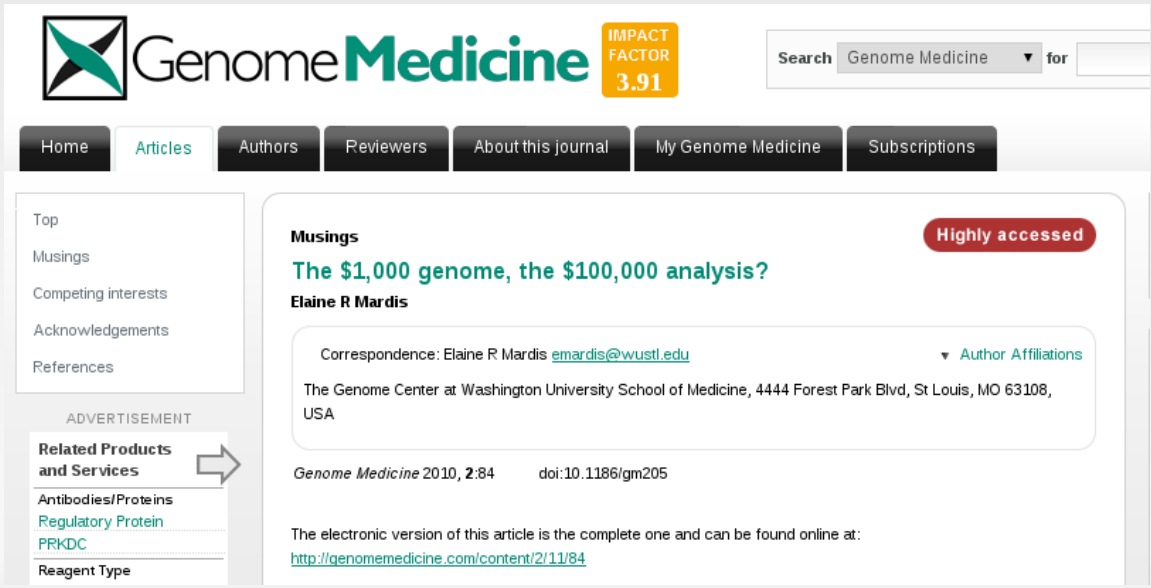
*Experiments **have increased data size by more than 4000x** when compared with old microarrays or first sequencers. Surprisingly, many software solutions are not very different.*

Sequencing costs keep falling, today a whole genome can be sequenced by **$1000**, so much more data is expected

Data processing and analysis are today a **bottleneck** and a nightmare, **from days or weeks with microarrays to months with NGS,** and it will be worse as more data become available



Genome Biology, 2011



It's the analysis, stupid!

Introduction
Big data considerations

- What is *big data*? It's not only big, it's also complex
  - *"Big data is a collection of data sets **so large and complex** that it becomes **difficult to process using** on-hand database management tools or **traditional data processing applications**"*
- How big is *big data*? TBs? PBs?
- *Big data* is ***not a new scenario*** for other science areas: *meteorology, physics, internet search, finance, business, ...*, some "outdated" examples:
  - CERN, at some points about 1 TB/s, about 25PB a year
  - NASA Center for Climate Simulation (NCCS) about 35PB
  - Large Synoptic Survey Telescope (LSST): ~30TB a night
  - Facebook: more than 300PB
    https://www.facebook.com/notes/facebook-engineering/presto-interacting-with-petabytes-of-data-at-facebook/10151786197628920

How did they process and analyze all these data? How do they work?

- Which are the ***main Big data challenges***? Typically *curation, search, sharing, storage, analysis and **visualization***, but also ***Interactive analysis***
- *Genomics* is a new player in *big data*. Some research in *Genomics* is now data-driven. We need better computing solutions and more data mining.

Advanced Computing Technologies

A quick overview

- ***Web technologies and visualization*: *HTML5*** standard brings some new features and possibilities to web browsers: SVG, FileReader, IndexedDB, WebGL … New Javascript libraries to build RIA applications JQuery, Backbone, Nodejs, Sencha, d3js, … Mobile applications, make data and applications available anywhere: *Google Android*, *Apple iOS*

- ***High-Performance Computing (HPC):*** Multi-core and new *many-core* CPUs (Xeon PHI, GPGPUs (Nvidia), SIMD (AVX2), … allow to speed-up analysis. Several frameworks and libraries: OpenMP, MPI, Nvidia CUDA, OpenCL, SSE, AVX2, ...

- ***Big data analysis and NoSQL databases:* *Big Data*** forces us to work in a distributed environment. Databases with biological information can no longer be stored in a single machine. We need to make data available through web services. Some solutions available: Hadoop MapReduce, Google Dremel (BigQuery), NoSQL databases (MongoDB, HBase, Solr), RESTful web services, ...

- ***Cloud computing:*** Computing needs can change and scalability is needed. Data can not be transfer over Internet, “move computation no data”. More *cloud-based* solutions are needed to analyze data. Clouds like Amazon AWS, Microsoft Azure or Google Cloud can be used

- ***Machine learning and data mining:*** these data can be used to train predictors for diagnose or for data mining to knowledge discovery

Our Biology lab

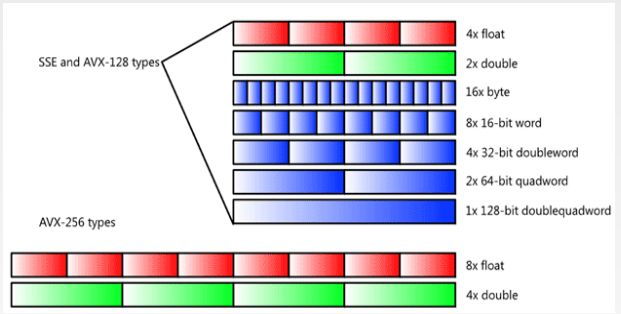
Advanced Computing Technologies
Web technologies, HTML5 standard and RESTful WS

- Web applications are slowly becoming the new “desktop applications”, more *cloud* and mobile-friendly applications, ie: *Gmail*, *Twitter*, *Facebook*, ...
- HTML5 brings many new standards and libraries to Javascript allowing developing amazing web applications in a very easy to 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. Demo: http://www.chromeexperiments.com/webgl/
  - **Others***: WebCL*, *Google NaCl* (not in HTML5, yet), Web components, …
- Many new web tools, frameworks and libraries being developed to build great web applications:
  - User interfaces and libs: JQuery, Ext JS, Bootstrap, ...
  - MVC: AngularJS, Backbone, EXT JS, …
  - Build: Grunt, Bower, Yeoman, …
  - Visualization: d3.js, Three.js, Highcharts, …
- RESTful web services and JSON ease the development of light and fast RPC

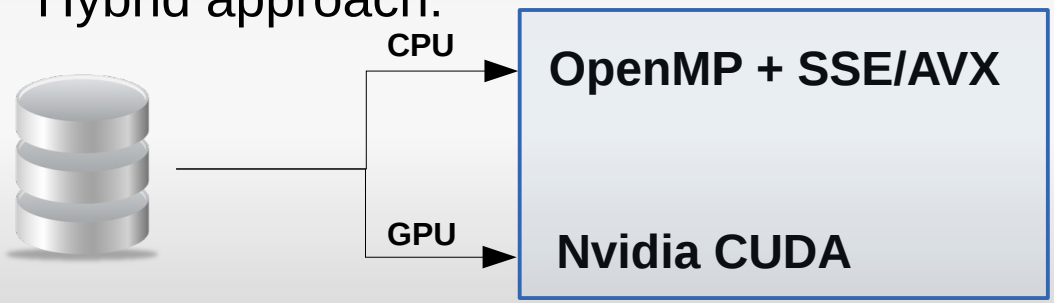
Advanced Computing Technologies
High-Performance Computing (HPC)

- An interesting battle between Intel and Nvidia, great for scientific research:
  - **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

- Better HPC frameworks available:
  - **Shared-memory parallel**: OpenMP, OpenCL, Intel Cilk Plus
  - **GPGPU computing**: CUDA, OpenCL, OpenACC
  - **Message passing Interface (MPI)**: MPI v3.0 coming soon with many new features (fault tolerant, remote memory access, ...)
  - **SIMD**: SSE instructions extended to AVX2 with a 512-bit SIMD



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



Advanced Computing Technologies
Big data analysis and NoSQL databases

- **Apache Hadoop** (**http://hadoop.apache.org/**) is *de facto* standard for ***big data analysis***. It's a Java framework library that allows distributed processing of ***large data sets*** across a cluster of nodes using a simple programming models such as *MapReduce*, or the new *Spark* and *Tez* execution engines
  - Core: HDFS, MapReduce and HBase
  - Also in the framework: Hive, Pig, Mahout, Spark, ...
  - Some distributions available: Hortonworks, Cloudera, MapR
- **NoSQL databases**, distributed and scalable, not normalized databases, 4 families
  - *Column store*: Apache Hadoop HBase/Cassandra, Hypertable, ...
  - *Document store*: MongoDB, CouchDB, Solr, ElasticSearch, ...
  - *Key-Value*: DynamoDB, Riak, Redis, ...
  - *Graph*: Neo4J, OrientDB, ...
- New solutions for PB scale **interactive analysis**:
  - *Google Dremel* (Google BigQuery) and similar implementations: *HortonWorks Stinger+Tez (now Hive 0.13), Apache Drill, Cloudera Impala, Facebook Presto*
  - Nested data, and comma and tab-separated data, SQL queries allowed

Advanced Computing Technologies
Cloud computing

- Many interesting features such as ***scalability*** and ***elasticity***
- We need to change the bioinformatic analysis model: ***Move computing, not data***
- Some commercial solutions available:
  - Amazon AWS: many services such as Hadoop, NoSQL, ...
  - Google Cloud: less services but BigQuery available, also Hadoop
  - Microsoft Azure: it's not that mature yet
- Open solutions:
  - OpenStack (Sahara project provides Hadoop over OpenStack https://wiki.openstack.org/wiki/Sahara)
  - OpenNebula
- Ease the administration of big clusters for big data analysis and services

Advanced Computing Technologies
Software and technology combination

- No single technology or solution solves current *big data* problems
- Advanced solutions need the proper combination of some technologies
- Not even a single NoSQL database, many problems require combination of some different databases
- HPC vs Big Data processing
  - HPC: fast computation
  - Big data processing
- Better software engineering to build up bigger and better solutions:
  - ETL (Cascading, Oozie, ...)
  - TDD (JUnit, Mockito, ...), Design patterns (DI, ...),
  - HPC and Distributed computing
  - Cloud-based solutions

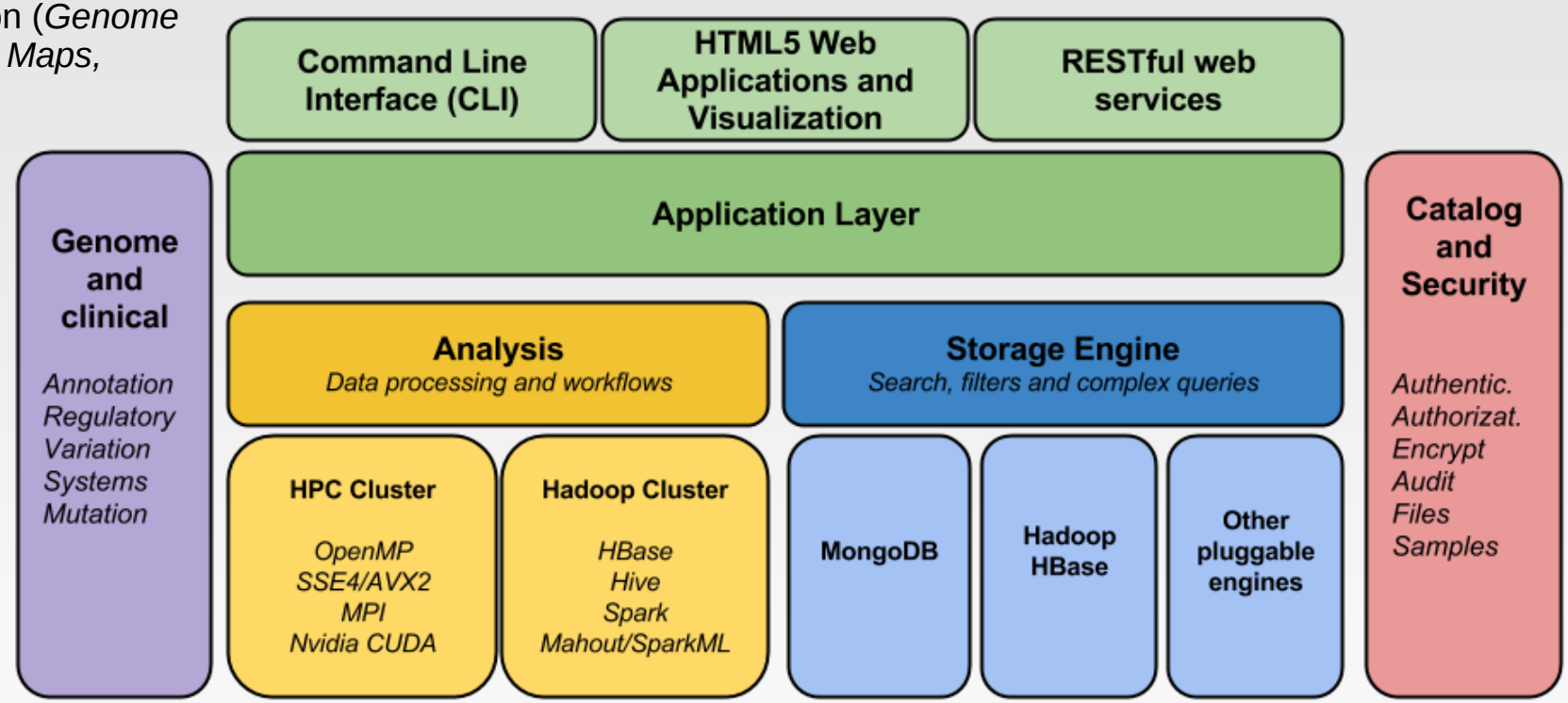
OpenCGA
Overview and goals

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

OpenCGA puts together all pieces: CellBase, Genome Maps, Cell Maps, HPG Aligner, HPG Variant, Variant annotation

Do not move data

HTML5 apps and visualization (*Genome Maps, Cell Maps, BierApp*)



Genome annotations and clinical information (*CellBase, Variant Annotation*)

Genomic and functional analysis (*HPG Aligner, HPG Variant, OpenCGA Analysis*)

Data (BAM,VCF, ...) indexes (*OpenCGA Storage*)

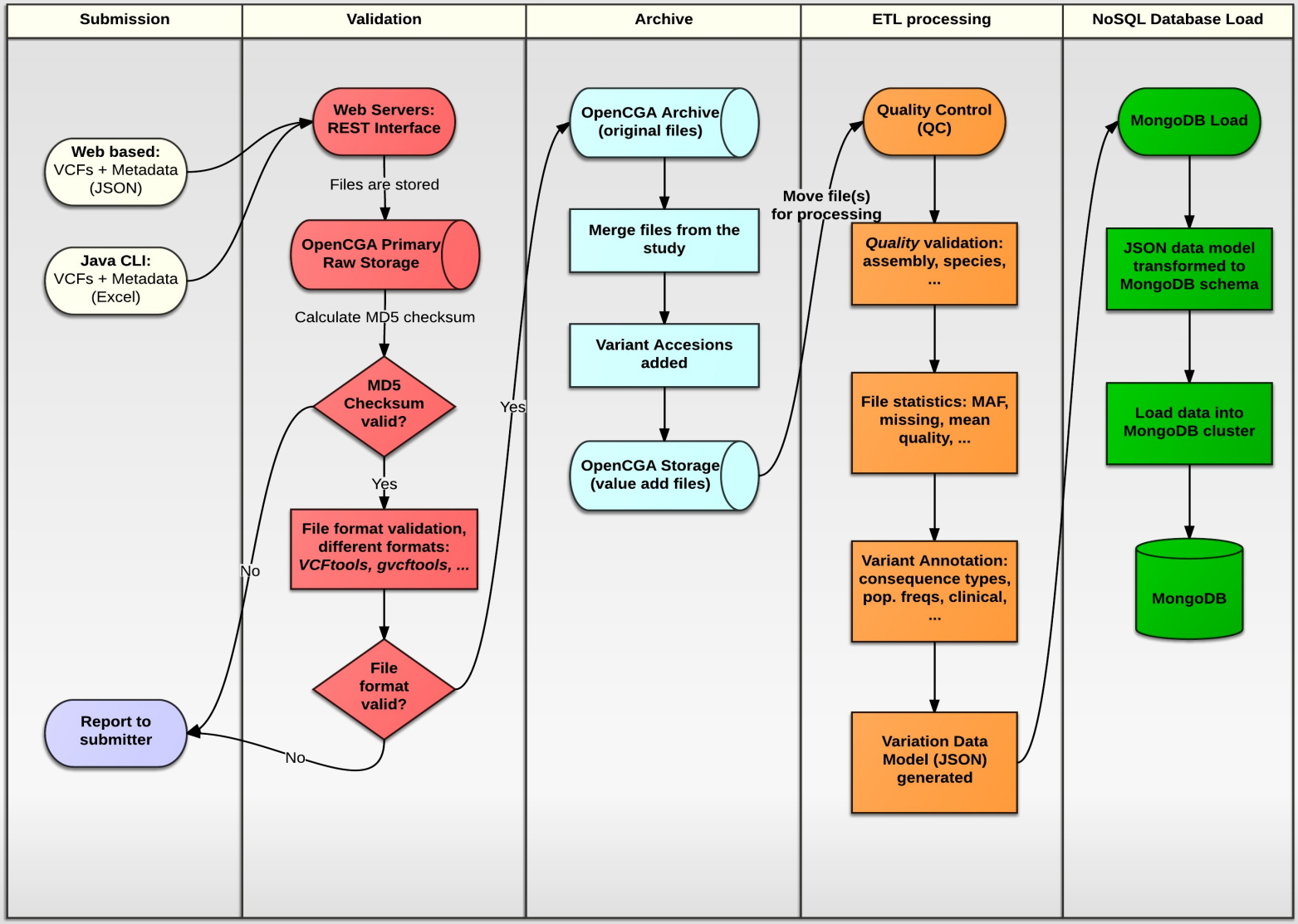
Metadata and AA (*OpenCGA Catalog*)

OpenCGA

Example: Variant Storage ETL

OpenCGA provides a *pluggable* Java framework for storing and querying variants

Two implementations are delivered with OpenCGA: MongoDB and Hadoop HBase



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

Security integrated in the framework, ie. Data can be encrypted with AES

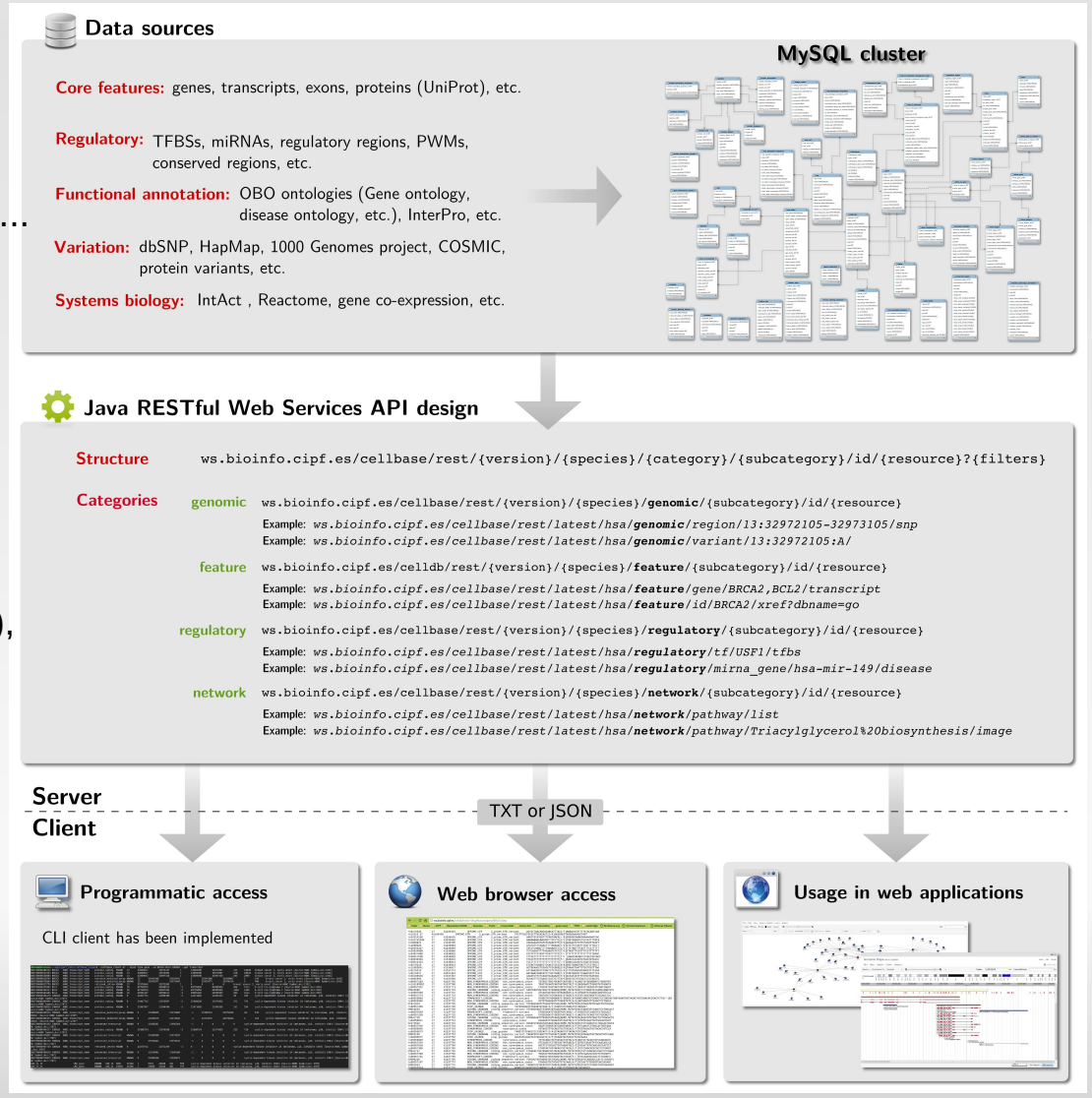
CellBase

An integrative database and RESTful WS API

- ***CellBase***, a comprehensive integrative database and *RESTful Web Services API*, current version 2.0 contains more than 200GB of data and 90 SQL tables, data exported in TXT and JSON:
  - *Core features*: genome sequence, genes, transcripts, exons, cytobands, proteins (UniProt),...
  - *Variation*: dbSNP and Ensembl SNPs, HapMap, 1000Genomes, Cosmic, ...
  - *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 ICGC and other projects (through *Genome Maps*)

Project: http://bioinfo.cipf.es/compbio/cellbase

Wiki: http://wiki.opencb.org/projects/cloud/doku.php?id=cellbase:overview



CellBase

New features coming in version 3.0

- About 15 species and much more data available, a database more scalability and flexibility needed: **NoSQL** databases
- Moving CellBase to NoSQL databases:
  - Version **v3.0**: release for ***MongoDB*** (~0.5TB data): Current version installed at EBI
  - Version **v4.0**: in ***HBase***? Not only genomes? (~TBs data)
- Clients: Python, R and HTML5 web based application (similar to *Biomart*)
- More coming features
  - Many more species
  - Aggregation stats
  - Feature clustering
  - Richer and scalable API
  - Amazon AWS release, different regions
  - Data format specification (similar to *DAS*)
  - JSON and Protocol Buffer encoding
- v3.0 released: 3Q14

Big data Visualization
Using HTML5 standards

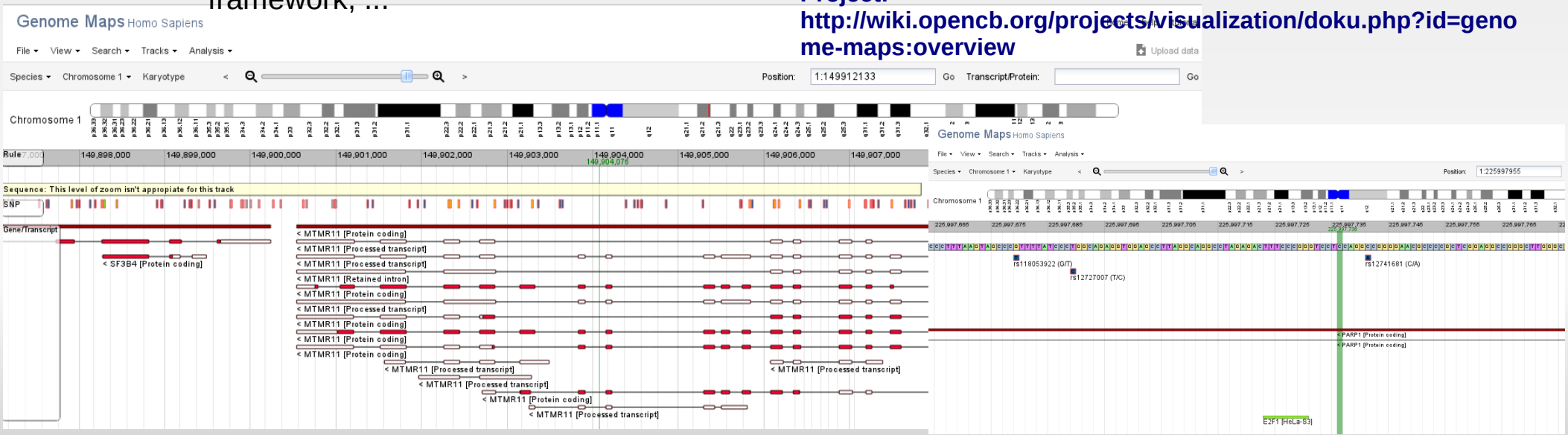
- Some visualization tools in Bioinformatics are desktop based applications, commonly written in Java or Perl, ie: IGV, Cytoscape, Circos, …
  - ***It's the server side!*** poor server-side development, how can I browse TB scale data? What about PBs?
- Others are based on *old* web technologies, ie: HTML4, based on images from a remote server, …
- HTML5 standards (boosted by Google among others) bring all the necessary to build cloud based applications for *big data*:
  - SVG for static and interactive visualization
  - WebGL for 3D visualization
  - IndexedDB for offline browsing and cache data (less remote accessions)
  - WebWorkers for more interactive applications
  - …
- HTML5 developed more than 5 years ago, but poor implantation in Bioinformatics

Big data Visualization

Genome Maps, a 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 or installation**
  - Genome data consumed from ***CellBase*** database through RESTful WS (JSON data) and other sources such as local files or DAS servers: genes, transcripts, exons, SNPs, TFBS, miRNA targets, … Data is parsed and SVG elements injected into DOM, this makes server side so light and improve network transfers
  - NGS data viewer. BAMs and VCFs file formats supported in remote and local way.
  - Other features: Multi species, Feature caches, API oriented, embeddable, key navigation, plugin framework, ...

Project:
http://wiki.opencb.org/projects/visualization/doku.php?id=genome-maps:overview



Big data Visualization
Genome Maps, new features coming

- Next version **v3.5**, tentative release date in September *2014*
  - 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 http://bioinfo.cipf.es/apps-beta/genome-maps/bam/
  - 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 *1Q15*
  - *JSCircos* for structural variation and other visualizations (http://bioinfo.cipf.es/apps-beta/circular-genome-viewer/)
  - 3D *WebGL* visualization for data sample comparison
  - *RNA-seq* visualization improvements
  - … many more performance improvements
- New server features being added
- Used by ICGC and other projects:
  - ICGC data portal: http://icgc.org/
  - Lens Beta website: http://patseq.dev.lens.org/lens/

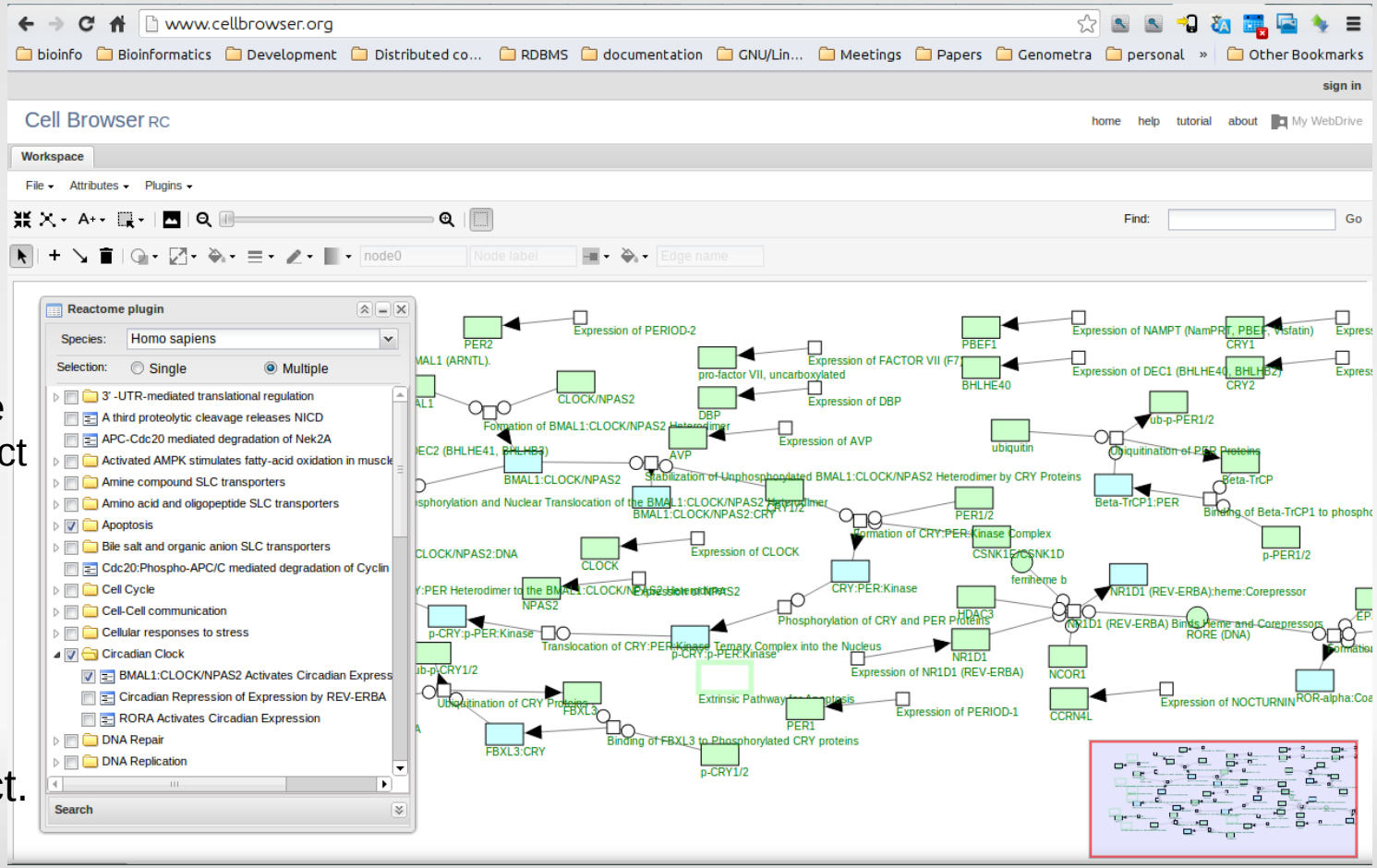
Big data Visualization

Cell Maps, HTML5+SVG biological network viewer

- Systems Biology ***analysis*** and ***visualization*** tool, similar to Cytoscape.
- Work in progress project. Release 1.0 for end 2014
- Used by NetworkMiner and RENATO tools
- Main features:
  - Graph edition
  - Plugin oriented
  - Connected to CellBase for Reactome and IntAct data
  - Main formats
  - Embeddable
  - ...
- It's a collaborative project. Paper in preparation.

Available at:
http://cellmaps.babelomics.org/

Documentation:
https://github.com/opencb/cell-maps/wiki



Big data Visualization
JSorolla, a JS library for visualization

- **JSorolla** is a high quality JavaScript library for genomic data visualization. Named in honor of Joaquin Sorolla, probably the most important Valencian (Spanish) painter (Wikipedia http://en.wikipedia.org/wiki/Joaqu%C3%ADn_Sorolla)
- All OpenCB visualization libraries are available under JSorolla for free at GitHub: https://github.com/opencb/jsorolla
- Some projects already using JSorolla such as ICGC or Lens. But many other integrating it right now!
- Currently Genome Maps and CellMaps libraries available. Many more visualization coming soon:
  - ***3D genome data***
  - ***Circos*** *(http://bioinfo.cipf.es/apps-beta/circular-genome-viewer/)*
  - Phylogenetic trees
  - ...
- It's a collaborative project. Paper in preparation.

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*)
  - association, TDT, Hardy-Weinberg, ...
  - ***Epistasis***: HPC implementation, 2-way 420K SNPs epistasis in 9 days in a 12-core node. MPI implemented.
  - 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 VARIANT

HPG Variant, a *cloud*-based variant *effect* predictor

- ***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*)
- *HPG-variant **effect*** combines *CellBase* WS and *Genome Maps* technology to build a *cloud*-based variant effect predictor, **no installation, no data download, always updated**
- Variant ***effect*** include not only genomic regions:
  - *Genomic location*: intronic, non-synonymous, upstream, …
  - *Regulatory location*: TFBS, miRNA targets, conserved regions, …
  - Protein functional prediction, conserved region
- Three interfaces: RESTful, web application and CLI

MySQL clusters
+200GB of data

Java RESTful WS API

Three clients available

Fast!
- GZip
- Multi Thread

WEB browser (AJAX) | **WEB application** | **CLI program**

HPG VARIANT

Coming soon

- Complete web environment for working with VCF:
  - http://bioinfo.cipf.es/apps-beta/variant/2.0.5/
- TB scale database server and RESTful web services for variants and genotypes
- HTML5 web app for Variation Data Mining tool (*aka* biomart)
- New **Variation Annotator**: HBase based resource with all the info for variants: consequence type, CADD, phenotype, conservation, protein function, ...
- Genomic analysis implemented in Hadoop, some HPC code imported as JNI
- ...

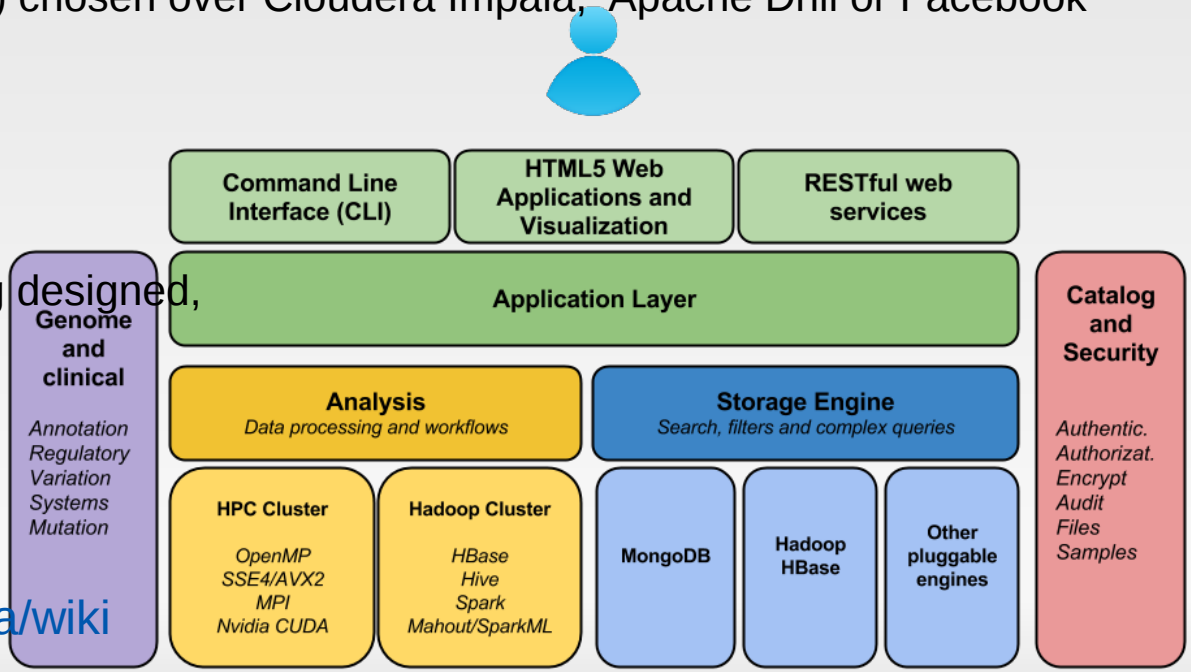
OpenCGA

A computing resource for *big data* analysis

- *HPC and Hadoop*-based environment for *big data* storage and analysis being developed: *OpenCGA*
  - *Storage*: efficient storage and data retrieval of hundreds TB, transparent connection to others clouds such as Amazon AWS or Microsoft Azure. Using *Hadoop HBase and Hive for scaling to petabytes*
  - *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

OpenCGA
Architecture and status

- Currently working in BAM and VCF files storage
- Pluggable storage engine, 2 back-ends implemented so far. Focused in **analysis** and **functionality**:
  - *MongoDB*: only for VCF, up to few TBs
  - *HBase*: for BAM and VCF, up to few PBs. Efficient storage for BAMs and VCFs.
- A new storage engine being designed for BAMs and VCFs for interactive analysis over PBs:
  - Hortonworks Stinger+Tez (Hive 0.13) chosen over Cloudera Impala, Apache Drill or Facebook Presto
- Analysis
  - Some HPC analysis are ready
  - Hadoop based implementation being designed, recently started
- Some groups are showing real interest: ICGC, CRG, EBI, ...
- Documentation at https://github.com/opencb/opencga/wiki (not ready yet)



OpenCGA HTML5 web applications

BierApp and Variant

- BierApp
  - A gene and variant prioritization tool built on OpenCGA
  - Published at NAR in 2014
  - http://bierapp.babelomics.org/
- Variant web
  - http://bioinfo.cipf.es/apps-beta/variant/2.0.5/
  - Paper in preparation
- Others coming soon

Other projects
New variant analysis tools

- These two tools are the last development of Joaquin Dopazo's group.
  - TEAM: Targeted Enrichment Analysis & Management
  - http://bioinfo.cipf.es/apps-beta/panels/1.0.1/
- VARIANT Effect 2.0: New Variant Annotation tool
  - Most of possible annotations in one single tool
  - Hadoop based and accessible through RESTful web services
- EGA and EVA at EBI, new huge and exciting projects to continue all this work

OpenCB

Open source initiative for Computational Biology

- Software in Biology is usually developed in small teams or independent way, we must learn from other science fields. OpenCB try to gather people:
  - http://www.opencb.org
  - http://wiki.opencb.org/doku.php
- We need more common and well designed platforms to build more **advanced solutions to solve current biology problems**:
  - Better engineering and more integrated solutions, better standards, ...
- No computing programming language oriented!
- OpenCB is a very young project and faces many challenges. However few people is starting to contribute during the last months.
- So far, is where all the software we develop is being released. About 10-12 active committers. Available as open-source at GitHub
  - https://github.com/opencb

Acknowledgements

- Current boss
  - Justin Paschall, EBI Variation
- Former boss at CIPF:
  - Joaquín Dopazo, Computational Genomic Department
- Three people in my former unit:
  - Joaquín Tárraga (*coming soon to Cambridge*)
  - Francisco Salavert (*coming soon to EBI*)
  - Cristina González (*currently at EBI with me*)
- Other outstanding collaborators from J. Dopazo's group:
  - Marta Bleda (*currently at Addenbrooke's with Stefan Graf*)
  - Alejandro Alemán
- Collaborators
  - UPV: Ignacio Blanquer, José Salavert, Andres Tomas
  - UJI: Enrique Quintana, Héctor Martinez, Sergio Barrachina, Maribel Castillo
- Bull
  - Nvidia Tesla and formation
  - Chair and support
  - Intel Xeon Phi accession
  - Hadoop cluster at BRDIGE