



Big Data & Visualization

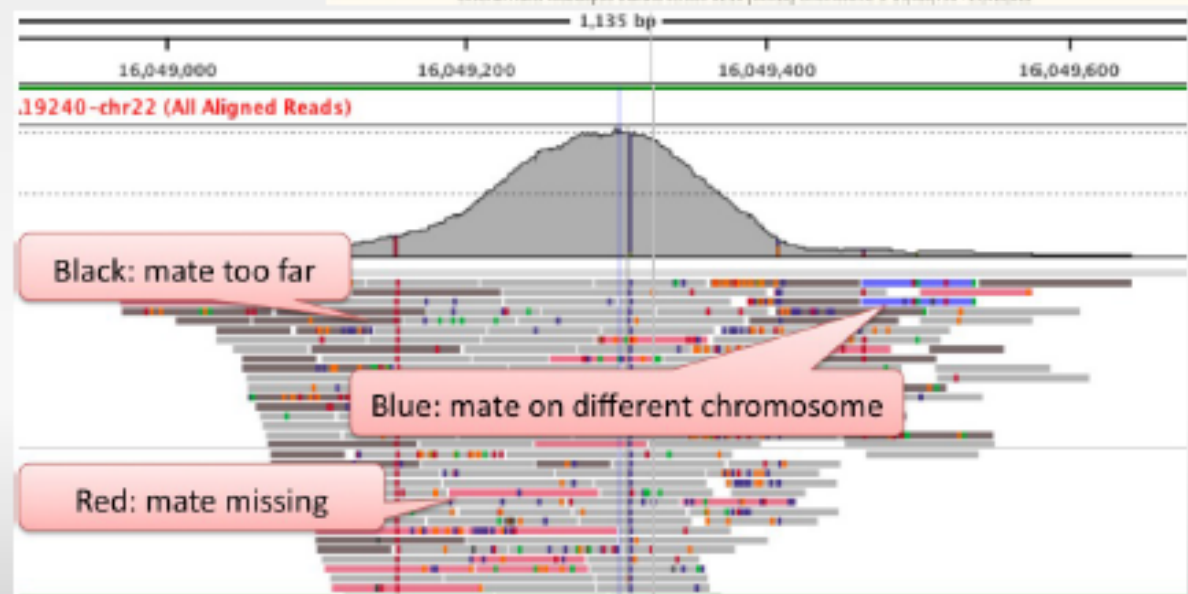
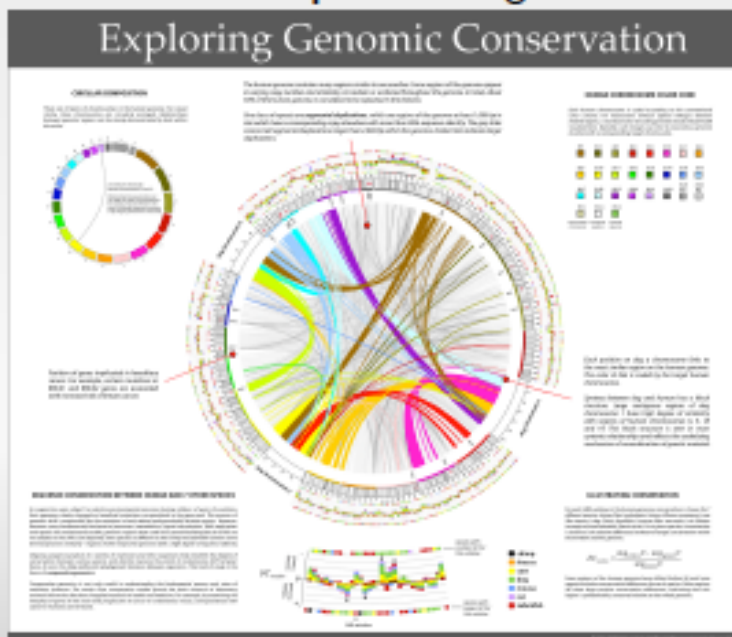
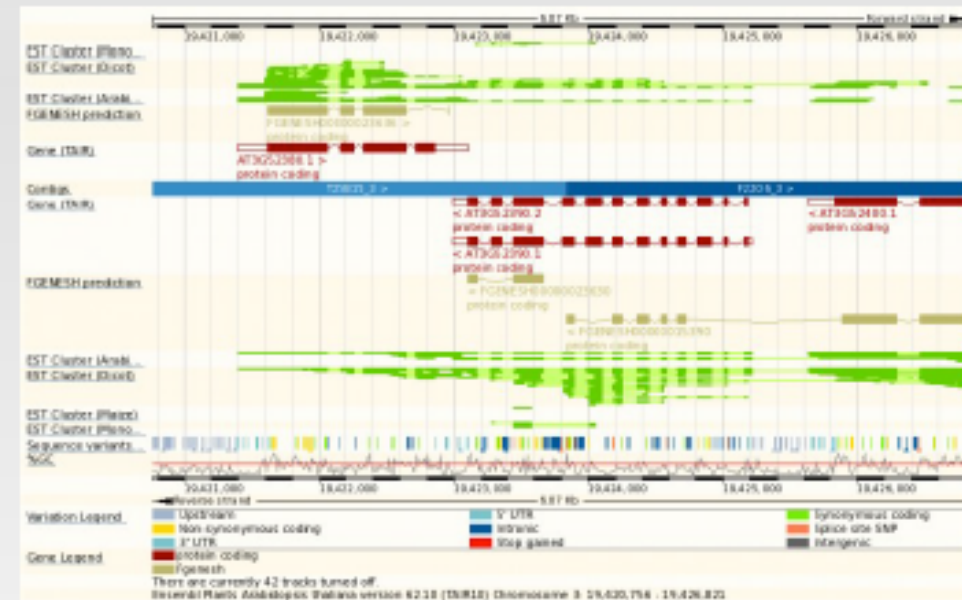
Marcel Caraciolo, CTO
marcel@genomika.com.br



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, ...





Big data in Genomics, a new scenario for biologists

Population power. Extreme throughput. \$1,000 human genome

The Illumina HiSeq X Ten is a set of six ultra-high-throughput sequencers, purpose-built for large-scale human whole-genome sequencing.

The First \$1000 Genome
Illumina has today's technology to sequence a human genome for \$1000. Learn how we can help you achieve this goal.

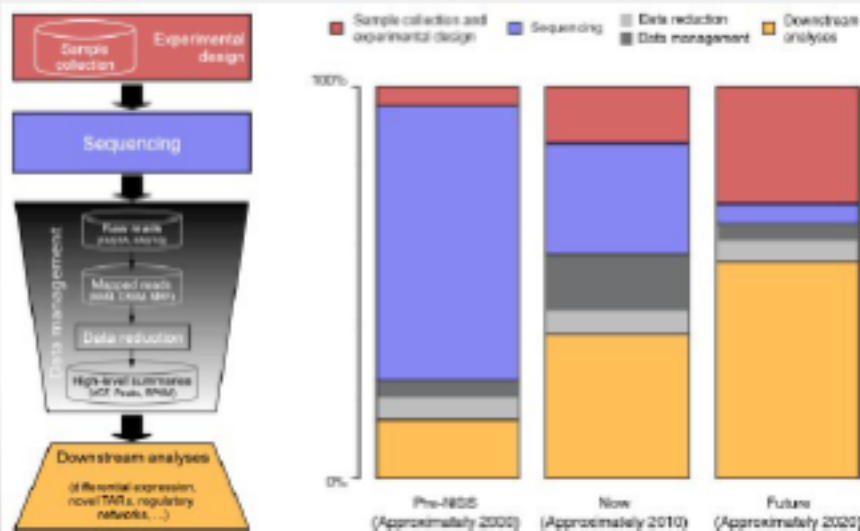
Population Scale Studies
Learn how the Illumina HiSeq X Ten can help you achieve your goals for large-scale human whole-genome sequencing.

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 Medicine **IMPACT FACTOR 3.91**

Search: Genome Medicine for

Home Articles Authors Reviewers About this journal My Genome Medicine Subscriptions

Musings **Highly accessed**

The \$1,000 genome, the \$100,000 analysis?

Elaine R Mardis

Correspondence: Elaine R. Mardis emardis@wustl.edu **Author Affiliations**

The Genome Center at Washington University School of Medicine, 4444 Forest Park Blvd, St Louis, MO 63108, USA

Genome Medicine 2010, 2:84 doi:10.1186/gm205

The electronic version of this article is the complete one and can be found online at: <http://genomemedicine.com/content/2/1/84>

Related Products and Services

- Antibodies/Proteins
- Regulatory Protein
- PRKDC
- Reagent Type

It's the analysis, stupid!



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



Visualization Challenges

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





Visualization Challenges

- 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?

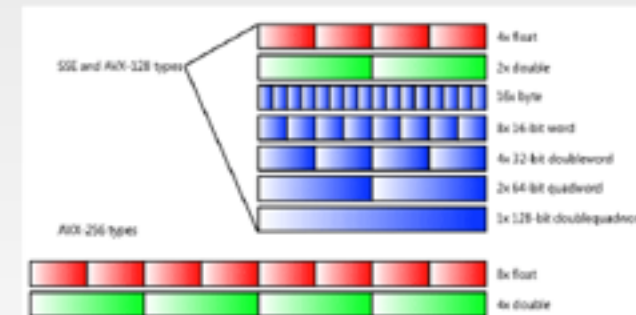
- Let's imagine next scenario: 10,000 whole genome sequenced samples with some RNA-seq, about 4PBs of data:
 - Can I easily explore and visualize the data? Can I filter and search in the data?
 - Can I filter variants by some annotations? By Stats? By Consequence type?
 - Can I perform more complex filters 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!



Visualization Challenges

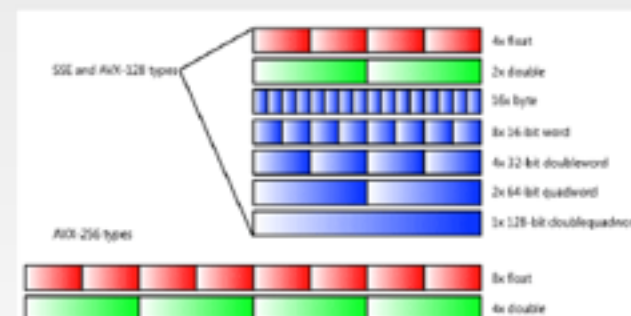
- An interesting battle between Intel and Nvidia, great for scientific research:
 - **Intel MIC architecture:** Intel Xeon and Intel Xeon Phi coprocessor, 1.01Tflops DP and more than 50 cores
 - **Nvidia Tesla:** Tesla K20X almost 1.31Tflops DP and 2688 CUDA cores
- Better HPC frameworks available:
 - **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*)





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*)





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



Mix of 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

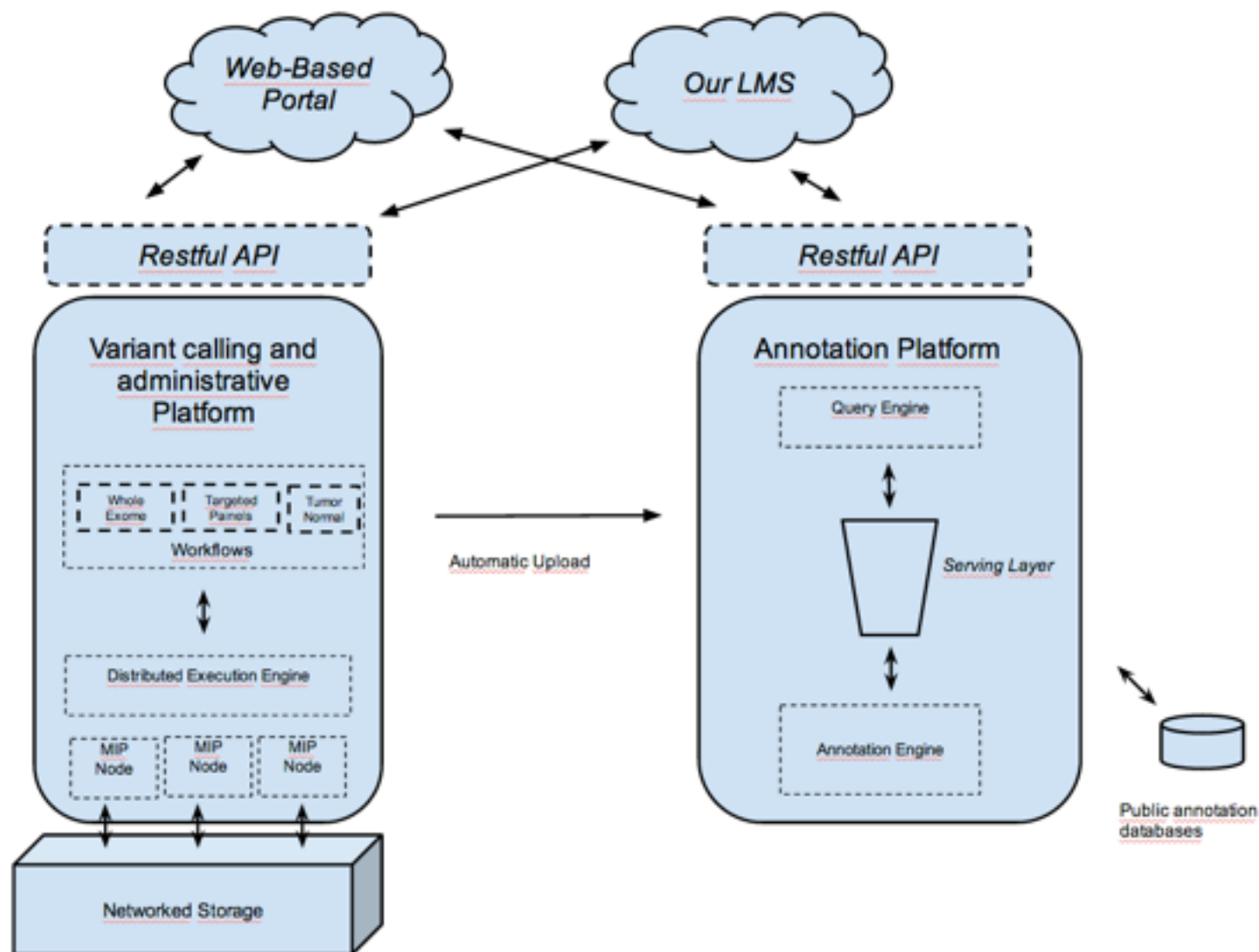


Fusion of technologies

- 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

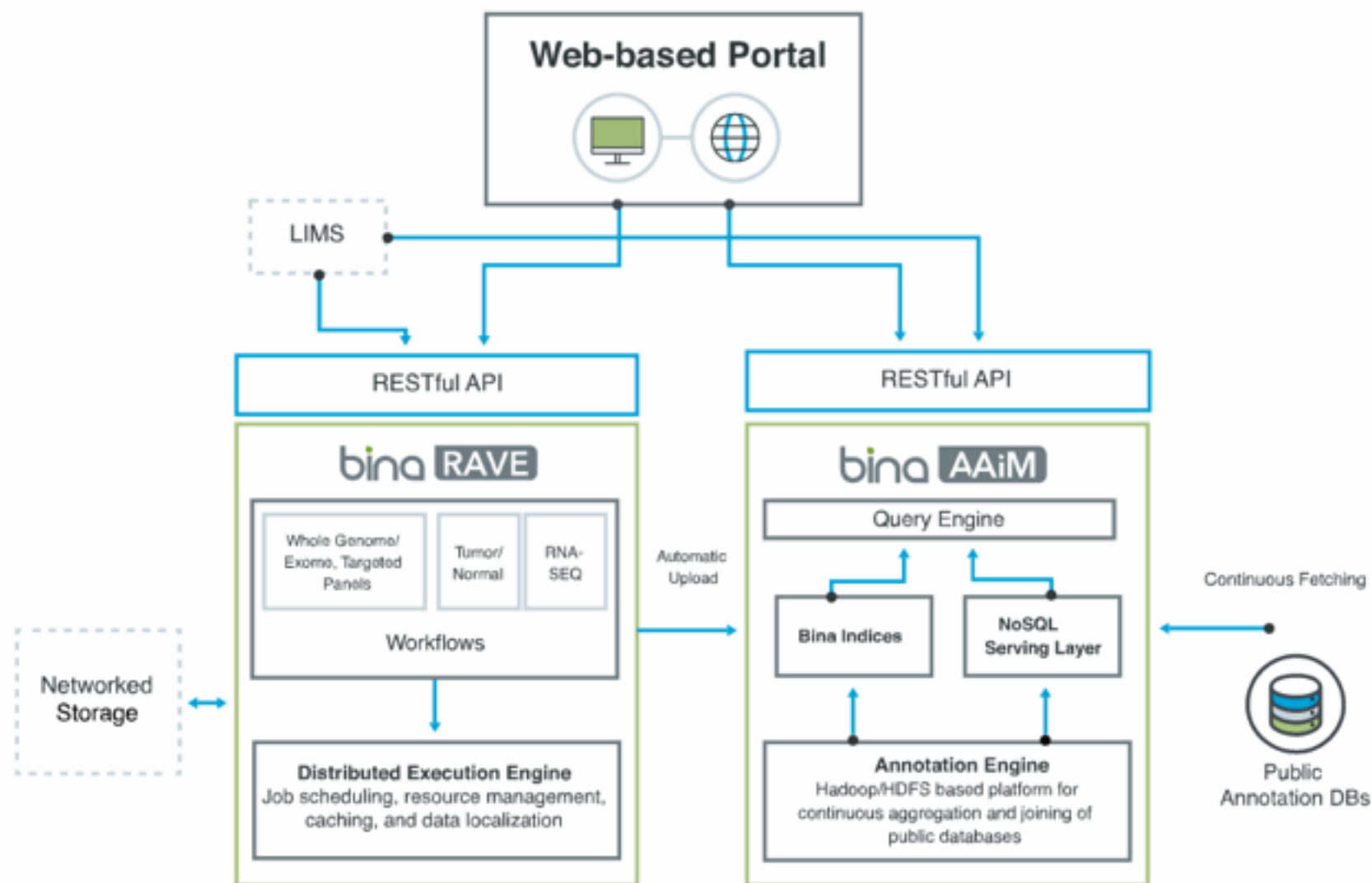


Our building approach





Bina (Full analysis platform)



<http://www.bina.com/>



Ingenuity Pathway analysis

Filter Cascade

Variants: 2564011, Genes: 20636

Common Variants: 447856, Genes: 17709

Predicted Deleterious: 5442, Genes: 4030

Genetic Analysis: 99, Genes: 96

Cancer Driver Variants: 60, Genes: 59

Biological Context: 13, Genes: 13

Add Filter

- Biological Context
- Cancer Driver Variants
- Common Variants
- Custom Annotation
- Genetic Analysis
- Pharmacogenetics
- Physical Location
- Predicted Deleterious
- Statistical Association
- User-Defined Variants

Summary | Variants | Genes | Groups/Complexes | Pathways | Processes | Diseases | Overview

Edit Columns | Export | Create List

Search for gene name/symbol

13 variants

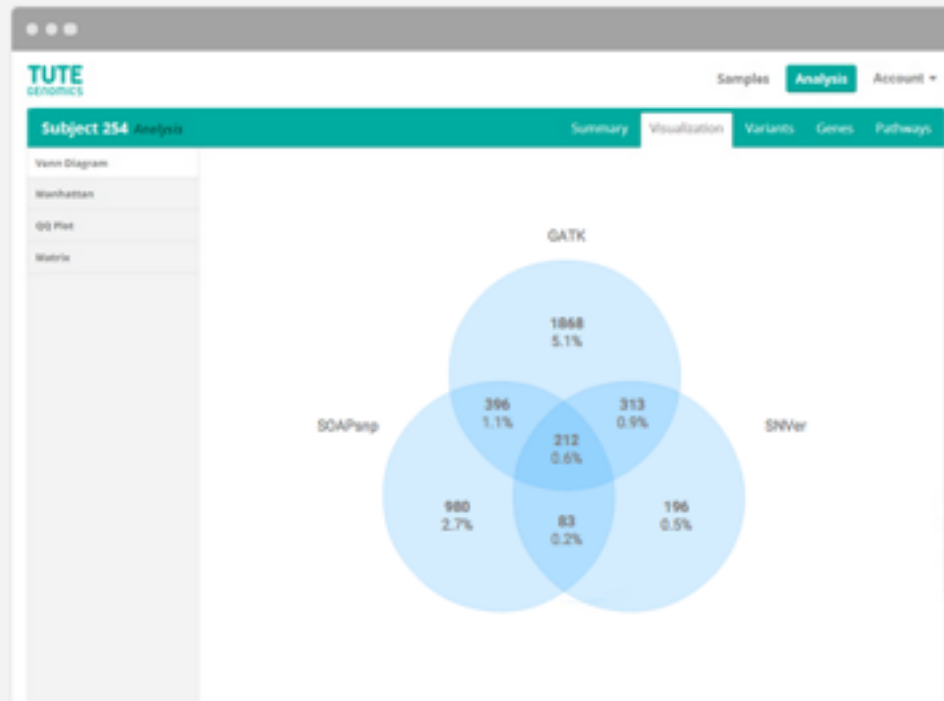
Chr...	Position	Gene Region	Gene Symbol	Protein Variant	Case Samples	Control Samples	Translation Impact	SIFT Funcio...	Regulatory Site
2	145156548	Exonic	ZEB2	M712V, M736V	--	---	missense		
3	127325493	Exonic	MCM2	V312L	--	---	missense	Damaging	
3	170110147	Exonic	SKL	M620K, M646K	--	---	missense	Activating	
4	109084821	Exonic	LEF1	P106L, P38L	--	---	missense		
4	114294537	Exonic	ANK2	E1837K, E1846I	--	---	missense	Tolerated	
5	172662014	Exonic	NKX2-5	R25C	--	---	missense		
6	15496930	Exonic	JARID2	R320C, R492C	--	---	missense	Damaging	
9	35056950	3'UTR	VCP		--	---			microRNA Bind
14	62204808	Exonic	HIF1A	T418I, T442I	--	---	missense	Tolerated	
15	25616257	Exonic	UBE3A	S335T, S355T, S	--	---	missense		
16	3807376	Exonic	CREBBP	Y1166F, Y1204F	--	---	missense		
17	66042639	Exonic	KPNA2	V506D	--	---	missense	Damaging	
20	1895955	Exonic	SIRPA	T97K	--	---	missense	Damaging	

<http://www.ingenuity.com/products/variant-analysis#/>



TuteGenomics

time for the user:



TUTE
GENOMICS

Tute Clinical Genomics Report *BETA*

Generated Fri 23rd, May 2014 - 7:22pm (UTC)

File Format: VCF
Report Version: 1.2-beta

File Upload Date: 05-22-2014 - 10:20pm (UTC)
Report Generation Time: 05-23-2014 - 7:22pm

Summary

The variants file (size: 24.56 MB, format: VCF) was uploaded into Tute web server. Genome annotation and report in hg19 coordinate were requested. The variant file meets Tute standard QC threshold for Genome annotation and report generation.

MLL3 NOTCH2 MPRIP BRIP1
SERPINE8 CAP1 NUP50 NOTCH1
CAP1 NUP50 BRIP1 NUP50

On the right were detected in this sample variants that may cause disease.

Interpretation

Top Variant List

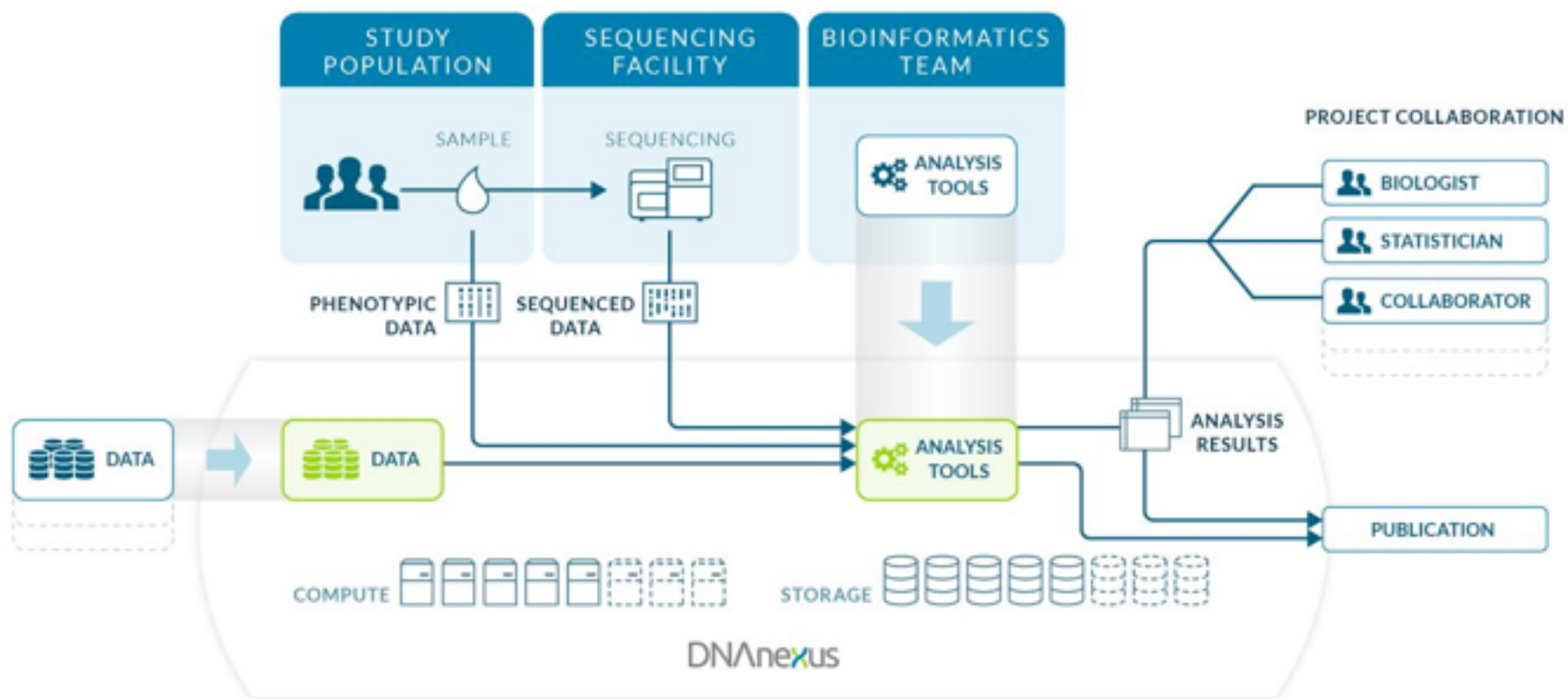
These variants identified in the sample may be responsible for existing disease or the development of disease

Gene	Variant	OMIM Link	Tute Prediction	ACMG Category
SDX1	NM_001079653:p.11_191del	MIM:607216	Damaging	Previously unreported, may or may not be causative
COL7A1	NM_000094:p290.1s	MIM:120120	Damaging	Previously unreported, may or may not be causative
STAB1	NM_015136:p.3_4d	MIM:608560	Damaging	Previously unreported, may or may not be causative

<http://www.tutegenomics.com/product>



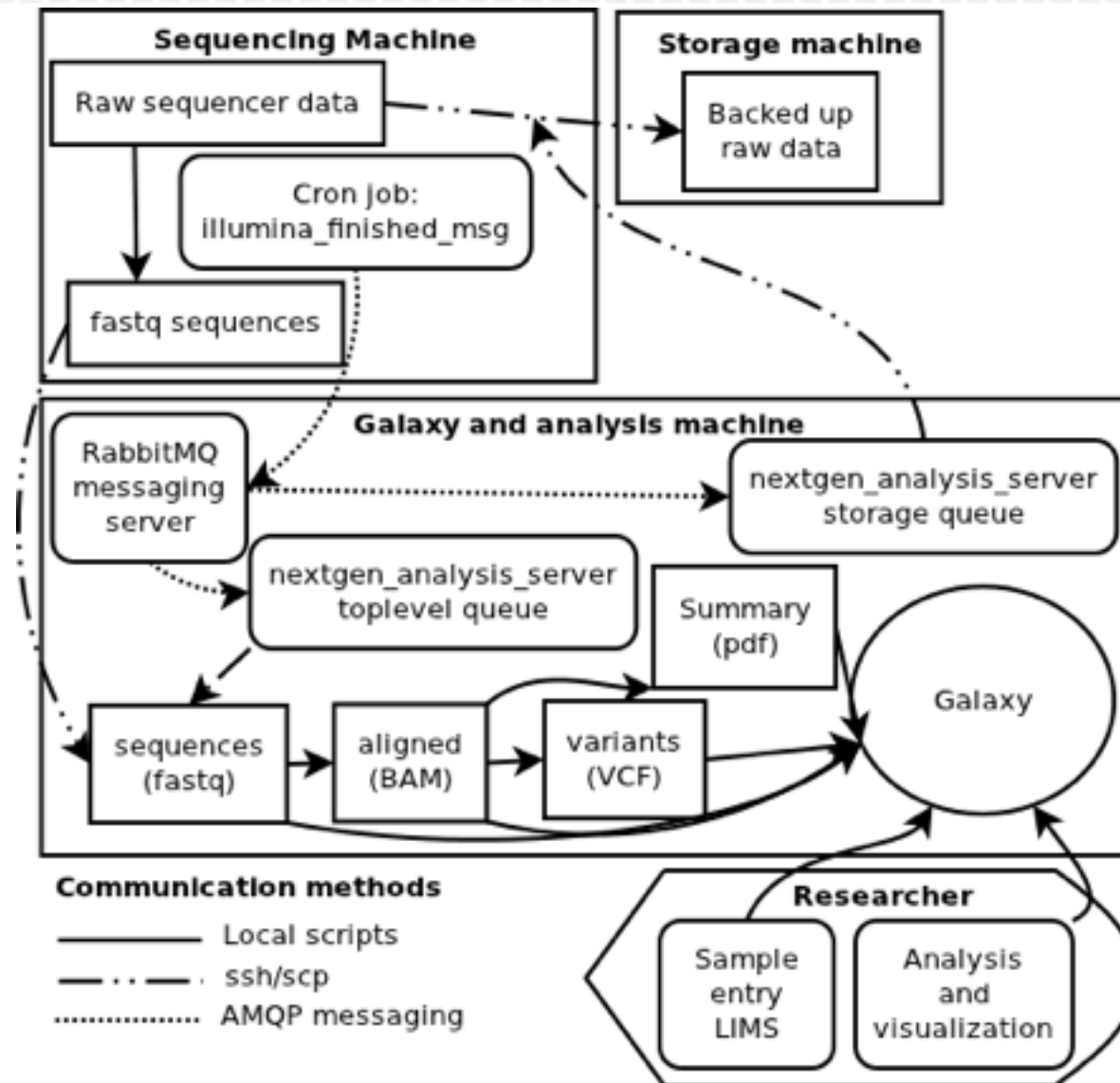
DNAnexus



<https://www.dnanexus.com/>



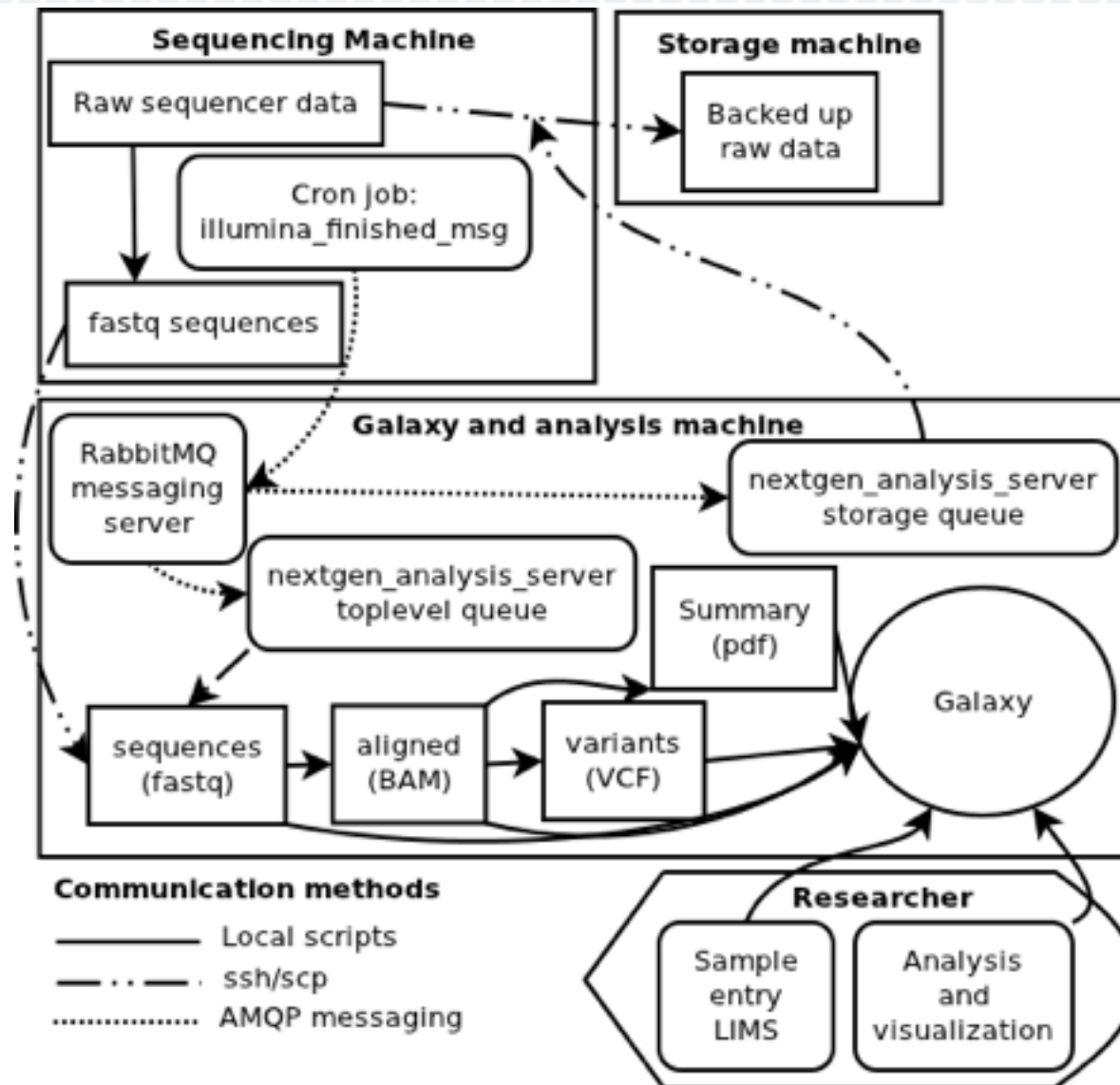
Bcbio-Nextgen (open-source)



<https://bcbio-nextgen.readthedocs.org/en/latest/>



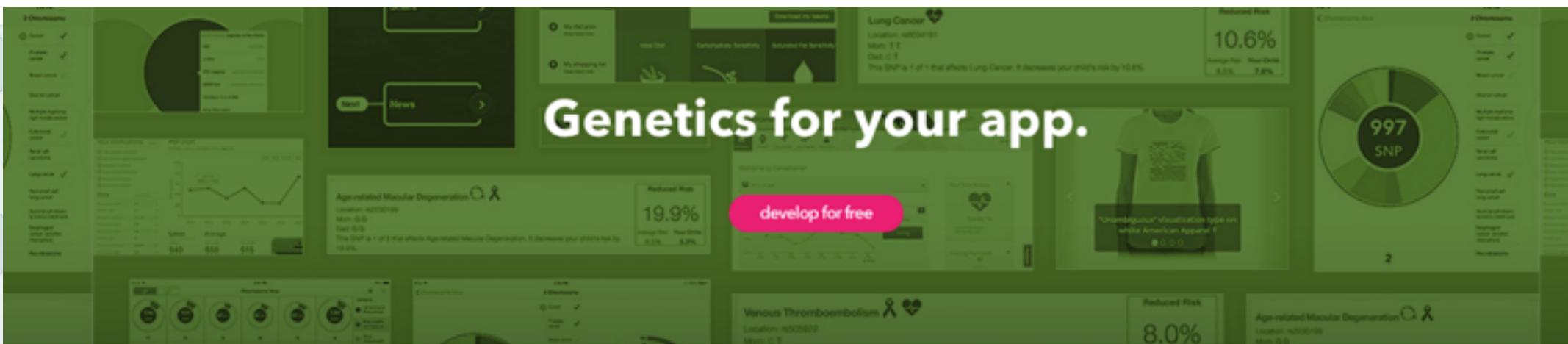
Bcbio-Nextgen (open-source)



<https://bcbio-nextgen.readthedocs.org/en/latest/>

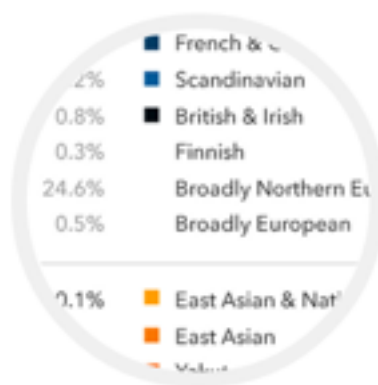


23AndMe (Genome data APIs)



REST-ful genes

Our customers are genotyped for hundreds of thousands of SNPs, conveniently accessible through our free REST API. Not genotyped? We have demo endpoints.



No need for a PhD

Our scientists have calculated [ancestry](#) and found [relatives](#) for genotyped customers. You could use this data without even knowing what a gene is!



Build novel apps

We have [Github](#) examples in Python, JavaScript, and others, to get you started quickly, as well as a [forum](#) to ask questions. Build novel apps on the human genome.

<https://api.23andme.com/>

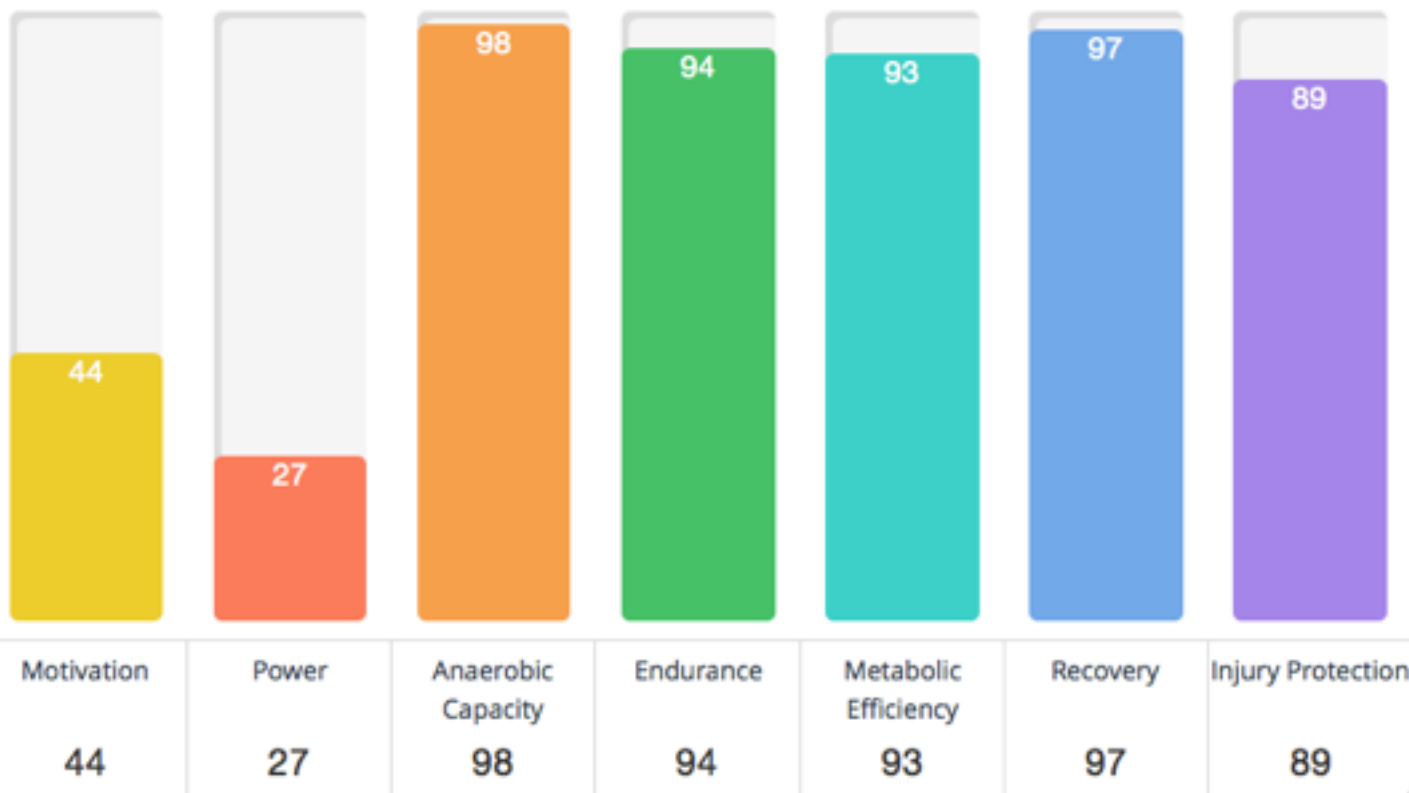


Athletigen (genome consumer apps)

Your Genetic Performance Scores Compared to People From Around the World



Worried about the low scores?



Your Gene Map

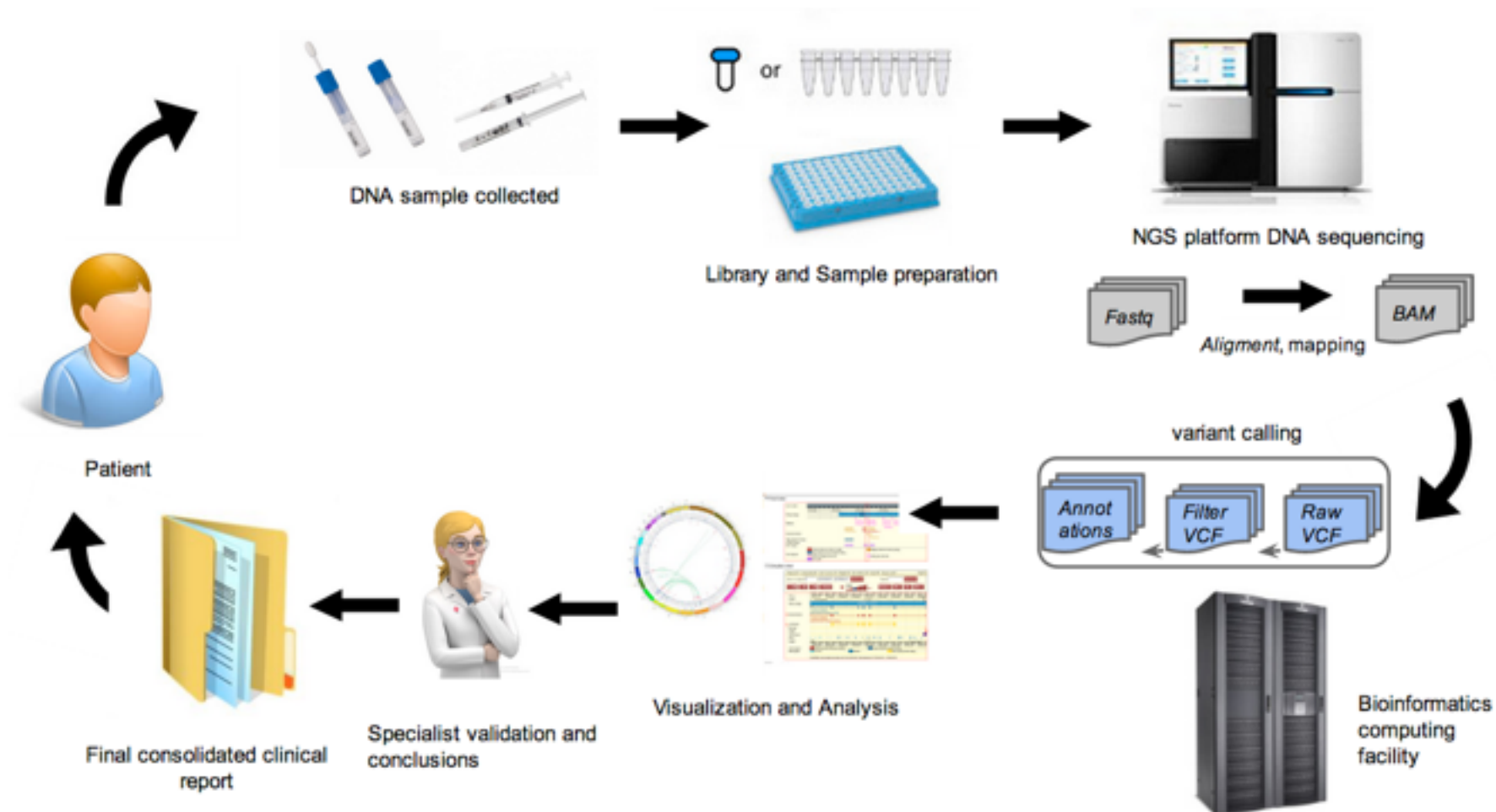


<https://www.athletigen.com/>



Isto é apenas 1 componente!

Pipeline Overview





Conclusões

- › Bioinformatics and Big Data:
 - › Interesting area
 - › Many open problems
 - › opportunities to acquire and apply knowledge
 - › opportunities from projects
 - › opportunities for researchers and novel high impact health business



Big Data & Visualization

Marcel Caraciolo, CTO
marcel@genomika.com.br