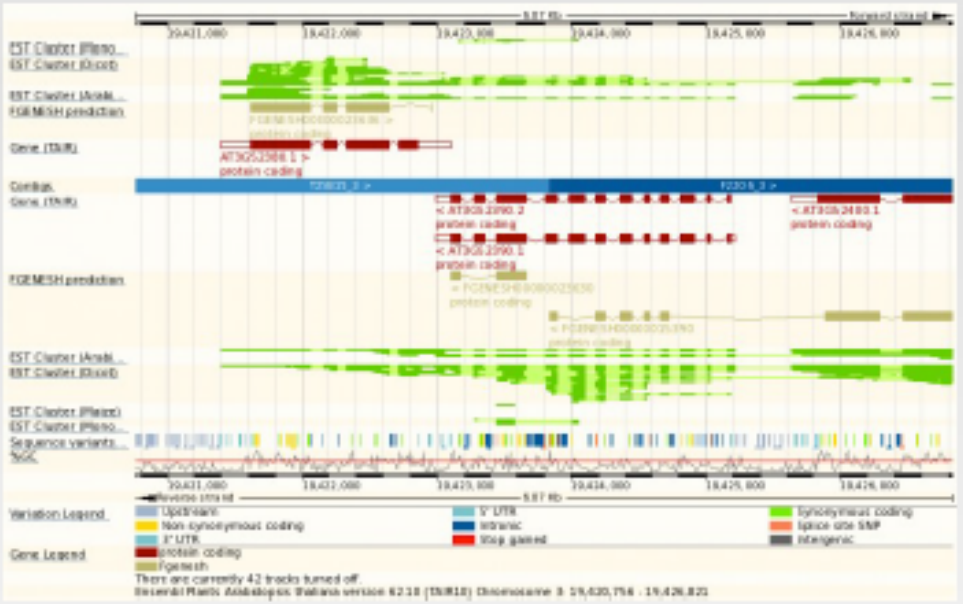
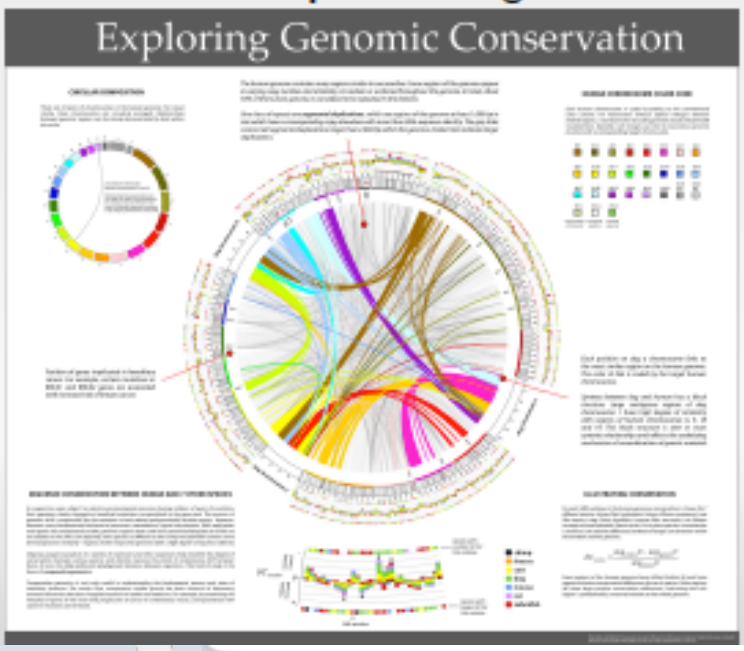
Big Data & Visualization

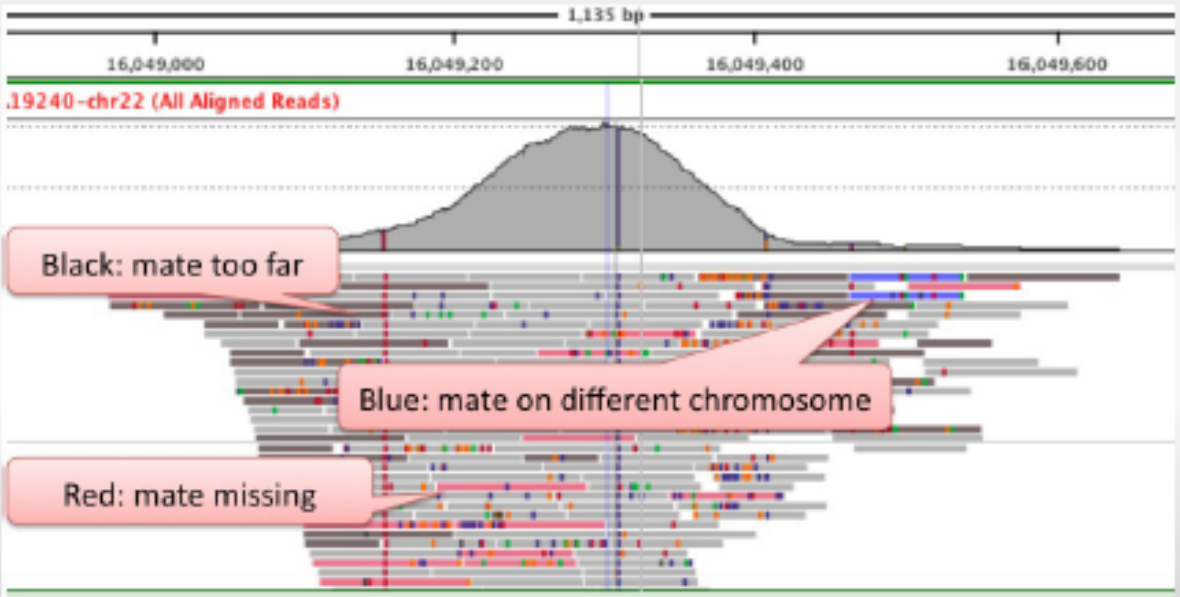
Marcel Caraciolo, CTO
marcel@genomika.com.br

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, ...







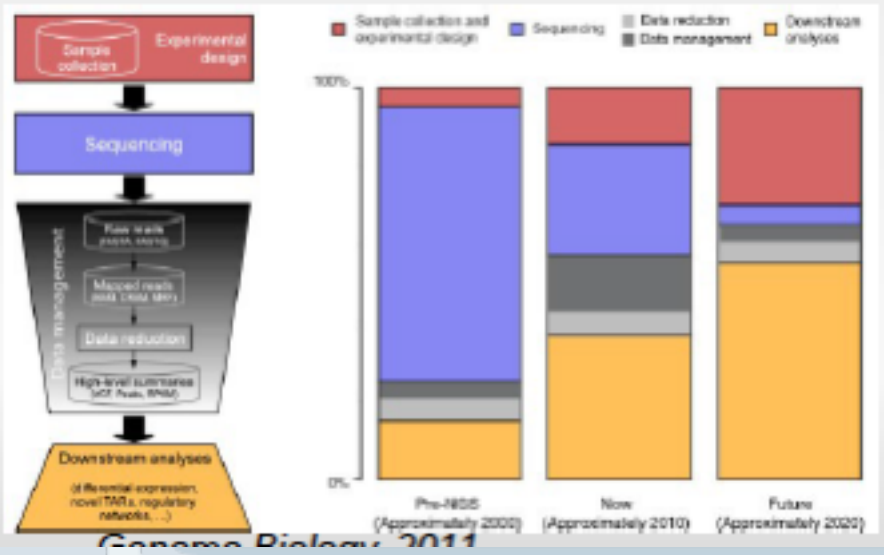
Big data in Genomics, a new scenario for biologists

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, ...*

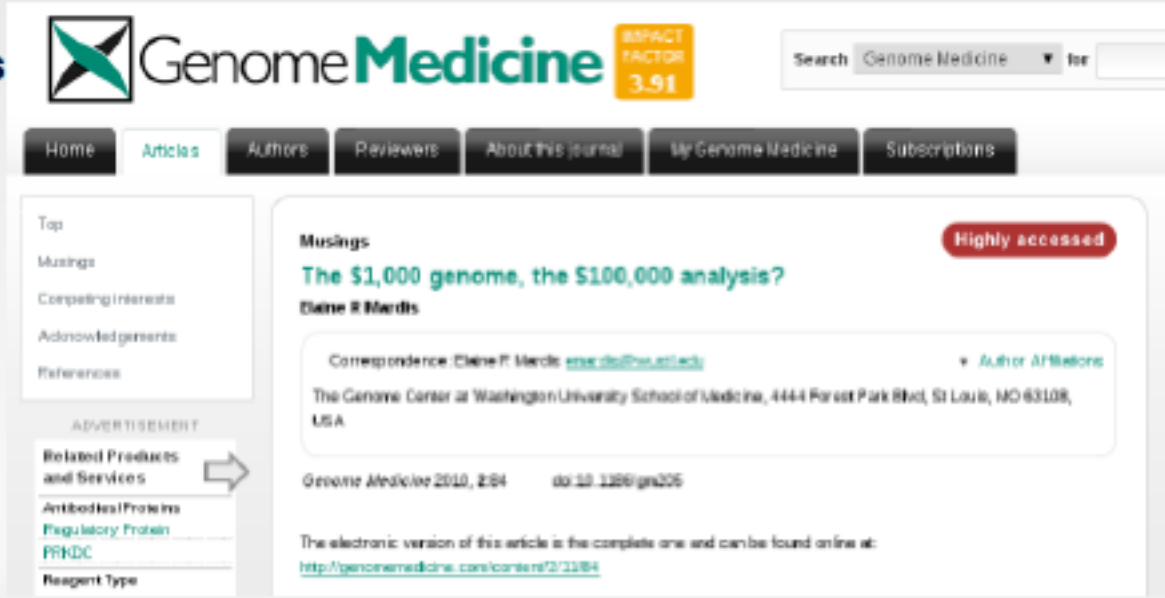
*Experiments **have increased data size by more than 4000x** when compared with old microarrays or first sequencers. Surprisingly, many software solutions are not very different.*

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!

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...

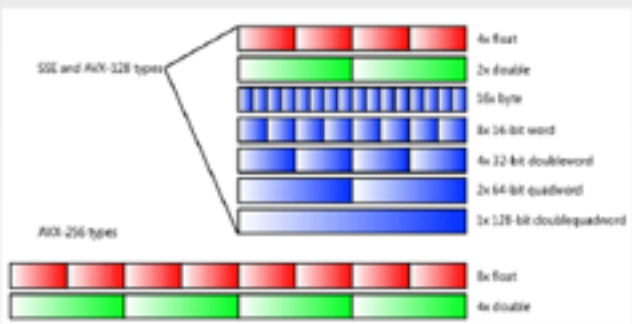
So, how are we doing it? Are the Bioinformaticians and Computational Biologists solving the current problems?

- 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

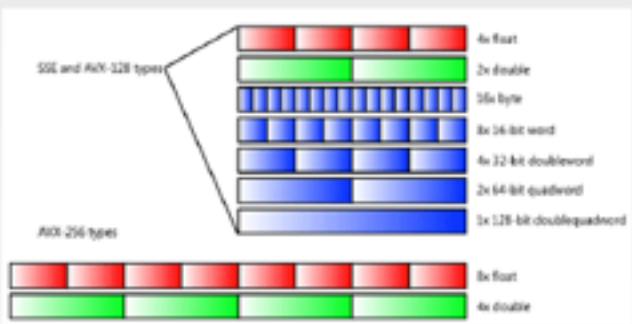
- An interesting battle between Intel and Nvidia, great for scientific research:
  - **Intel MIC architecture**: Intel Xeon and Intel Xeon Phi coprocessor, 1.01Tflops DP and more than 50 cores
  - **Nvidia Tesla**: Tesla K20X almost 1.31Tflops DP and 2688 CUDA cores
- Better HPC frameworks available: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

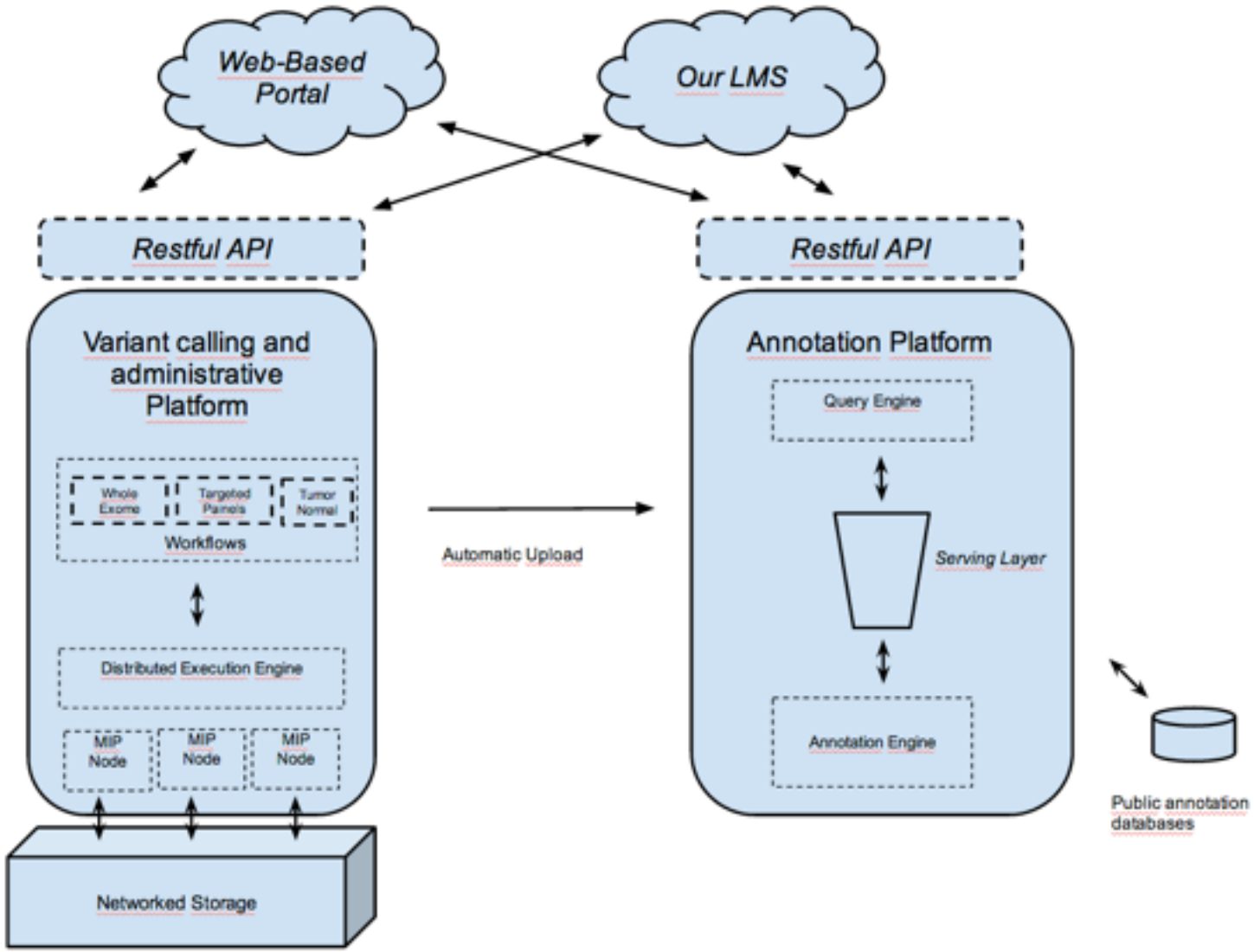
Mix of databases

- **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

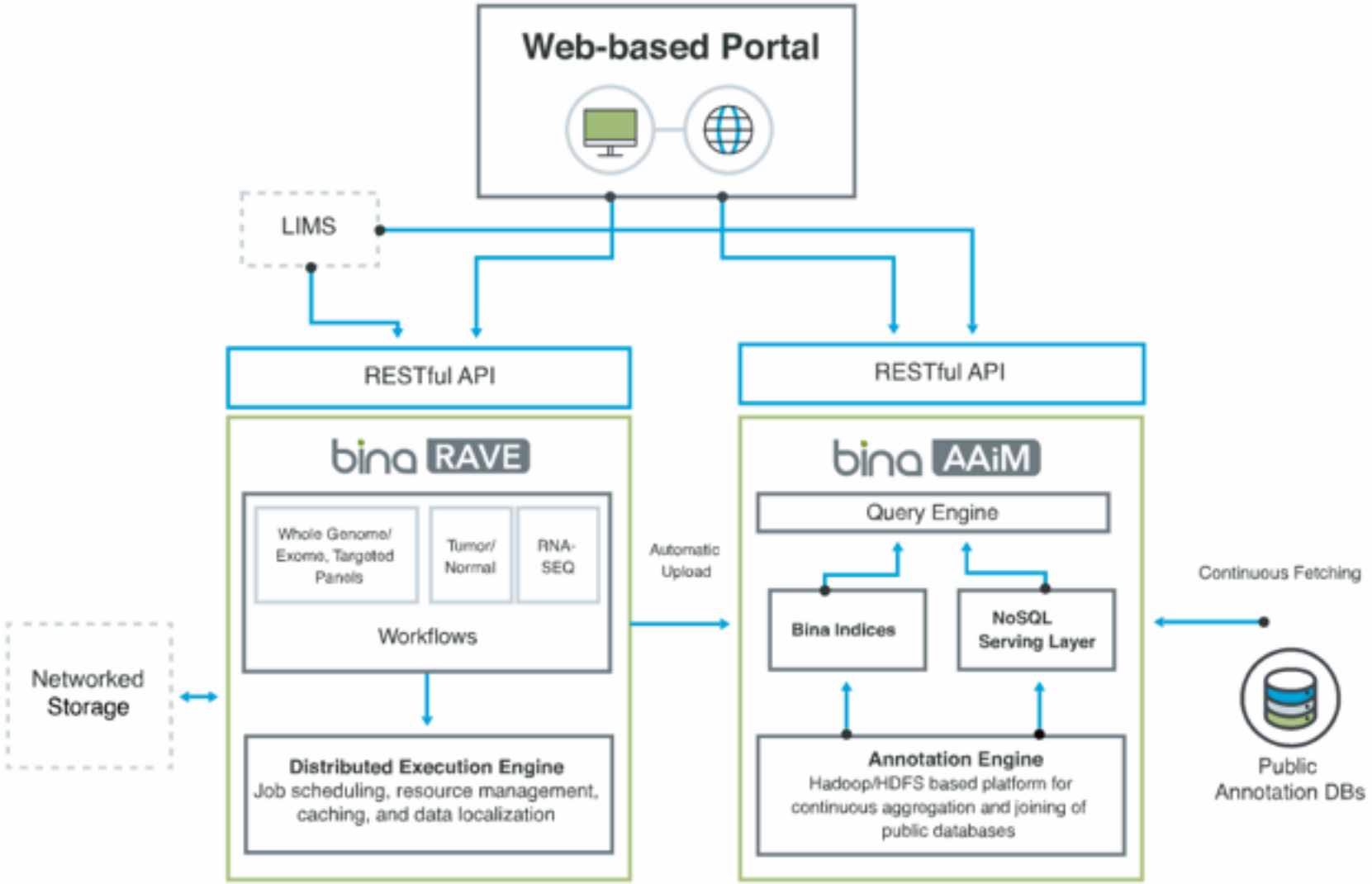
Fusion of technologies

- 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

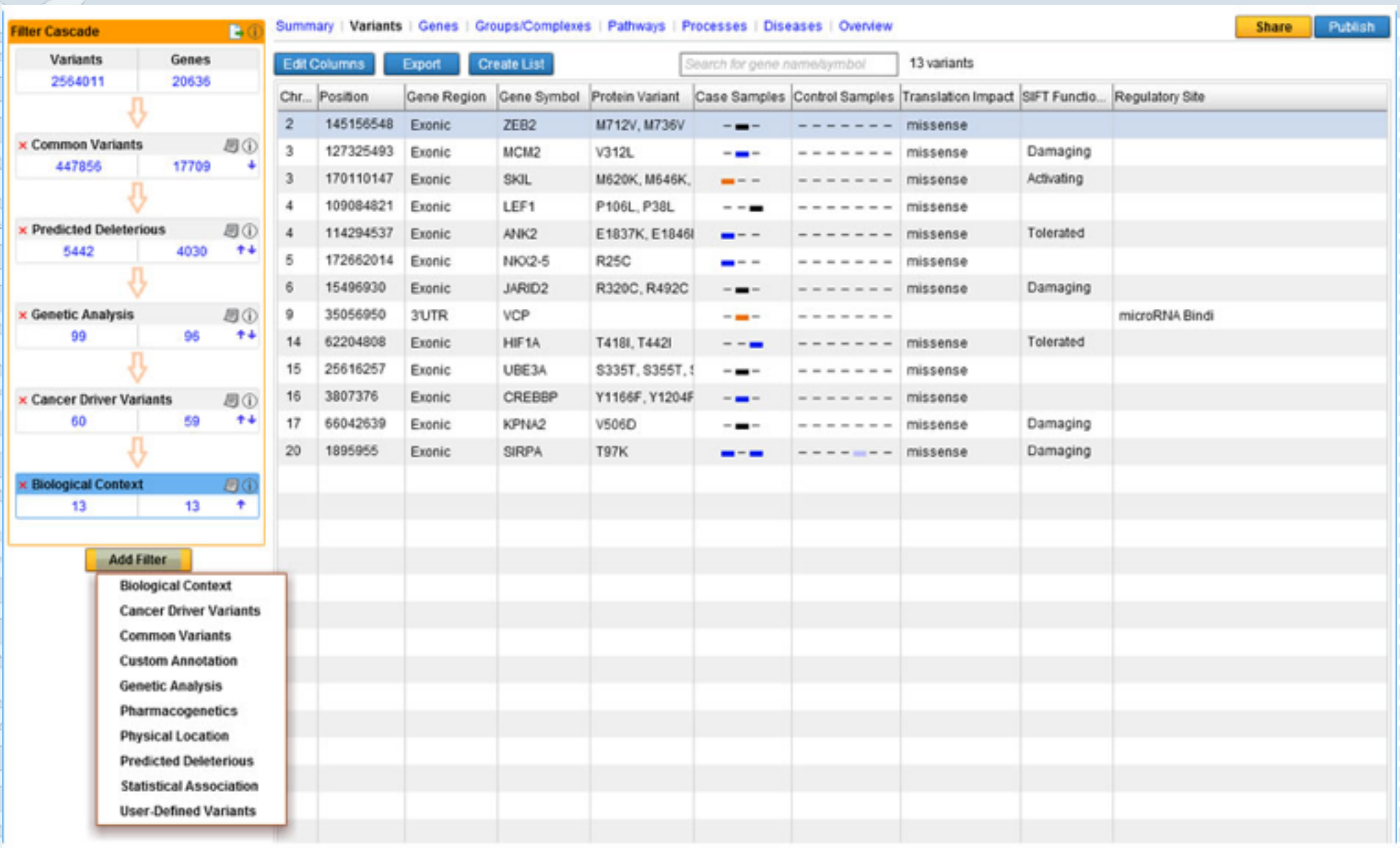


Bina (Full analysis platform)



http://www.bina.com/

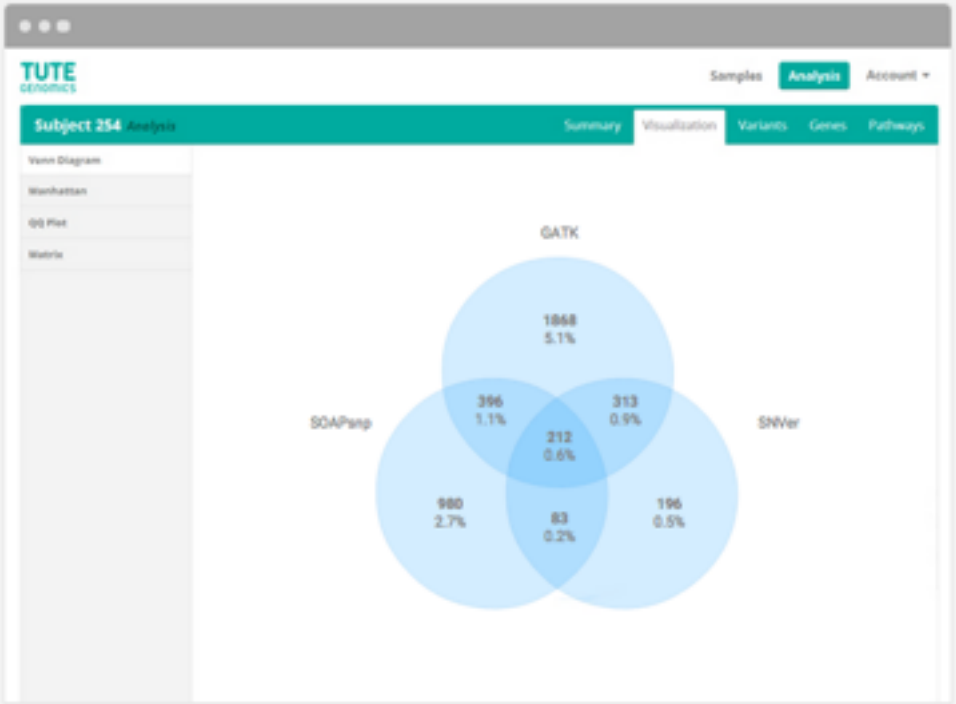
Ingenuity Pathway analysis

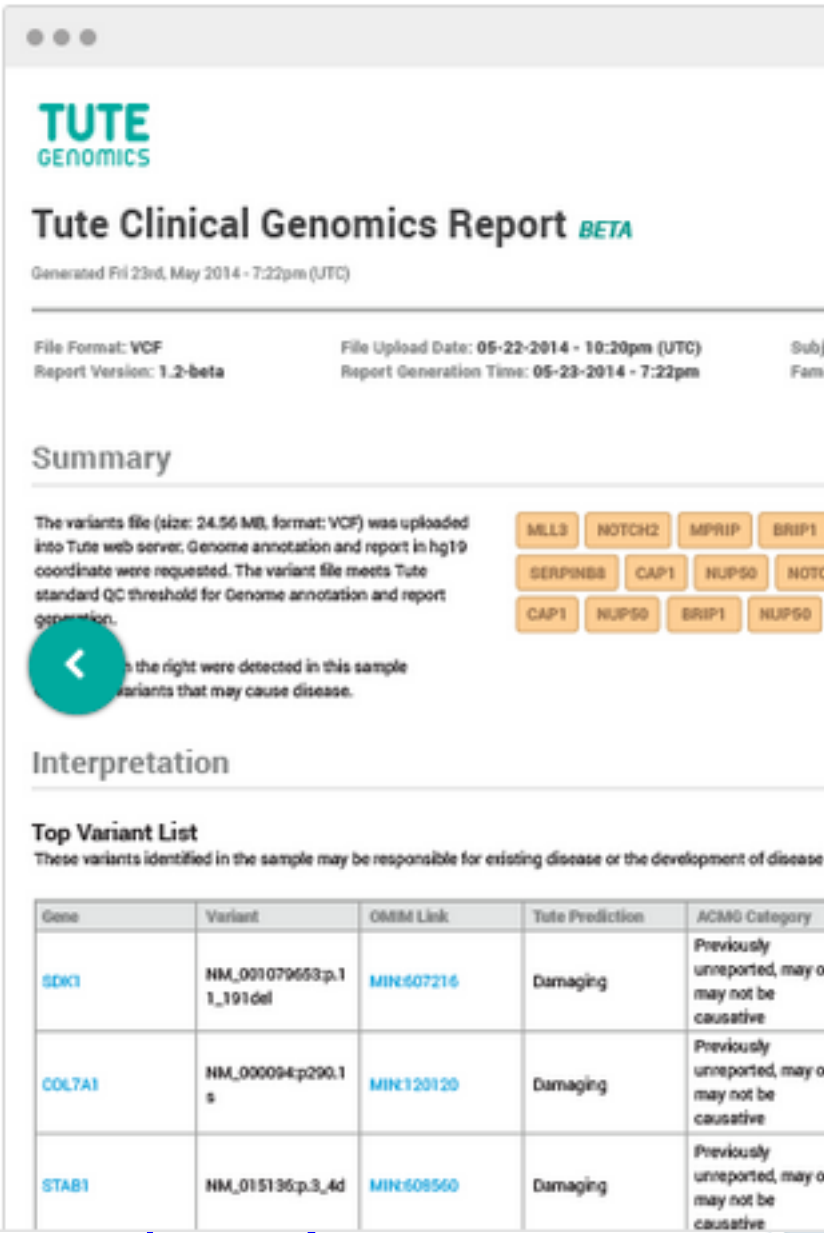


http://www.ingenuity.com/products/variant-analysis#/

TuteGenomics

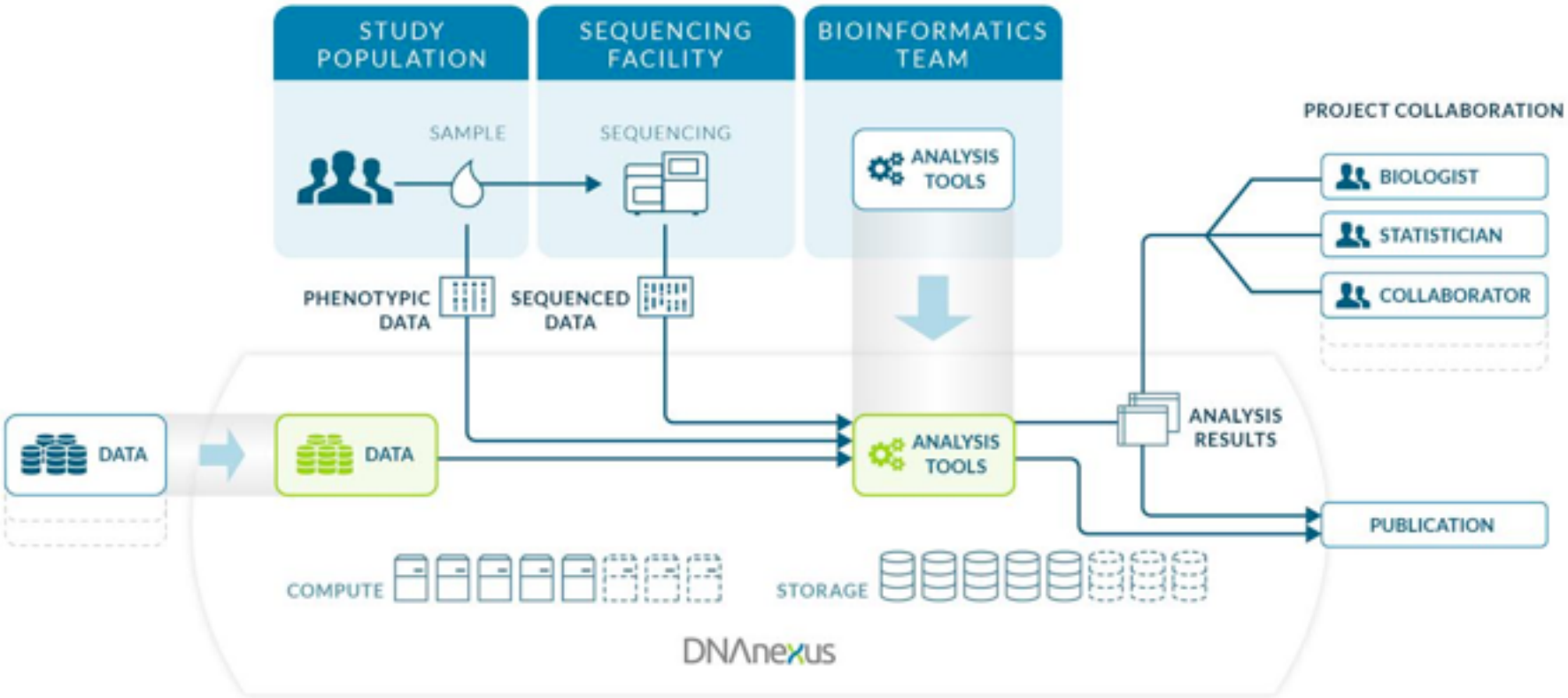
time for the user:





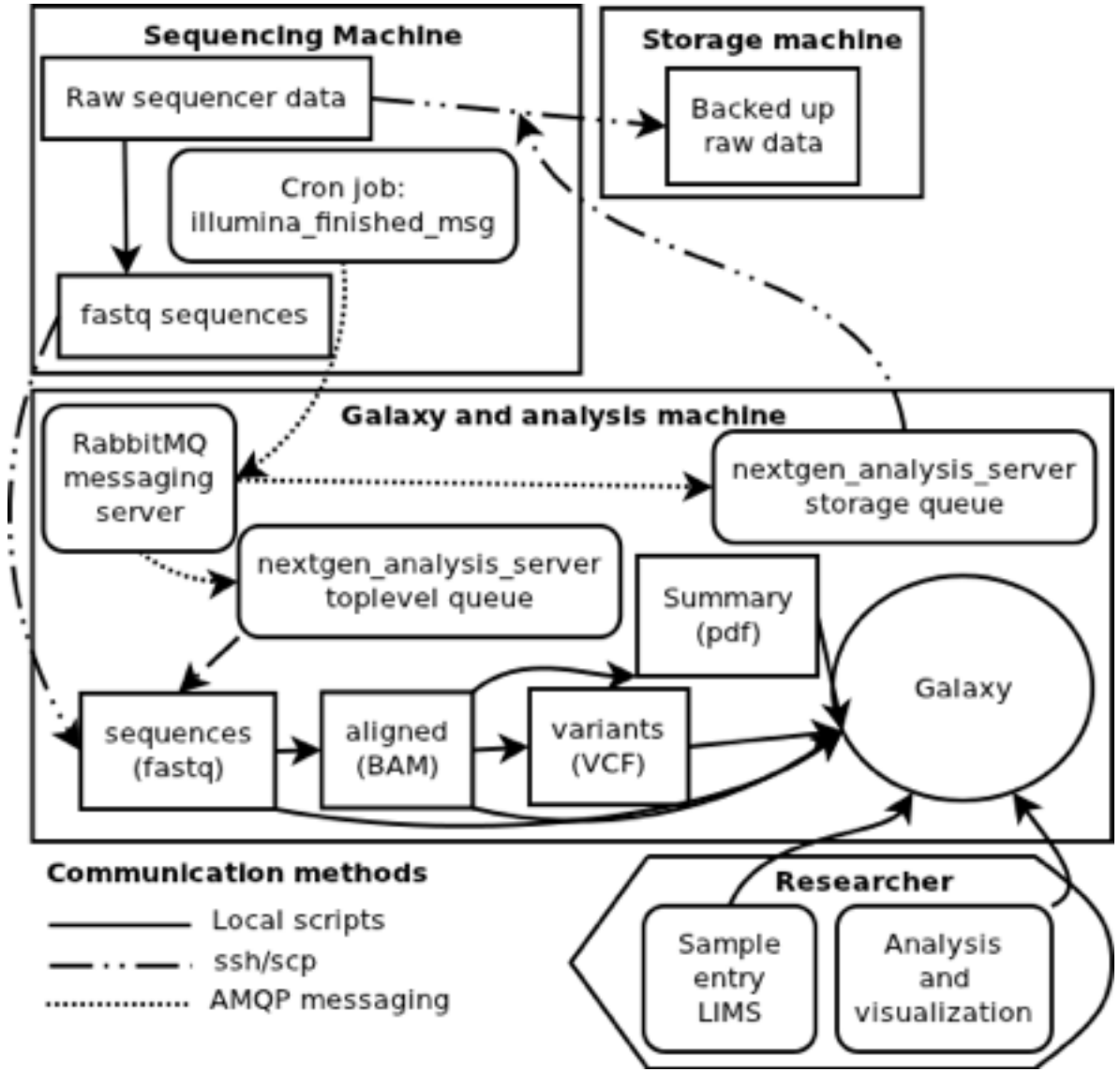
http://www.tutegenomics.com/product

DNAnexus



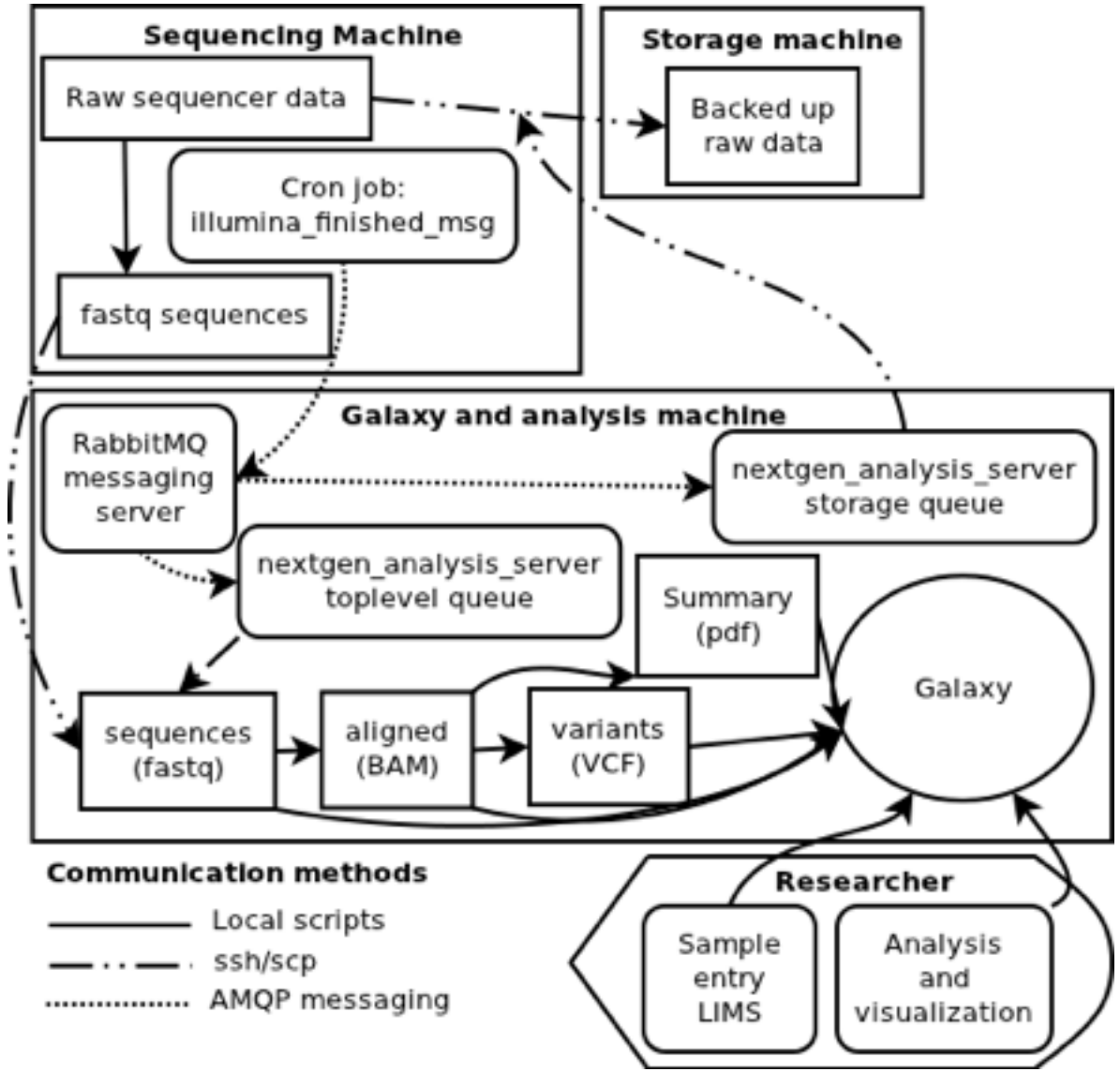
https://www.dnanexus.com/

Bcbio-Nextgen (open-source)



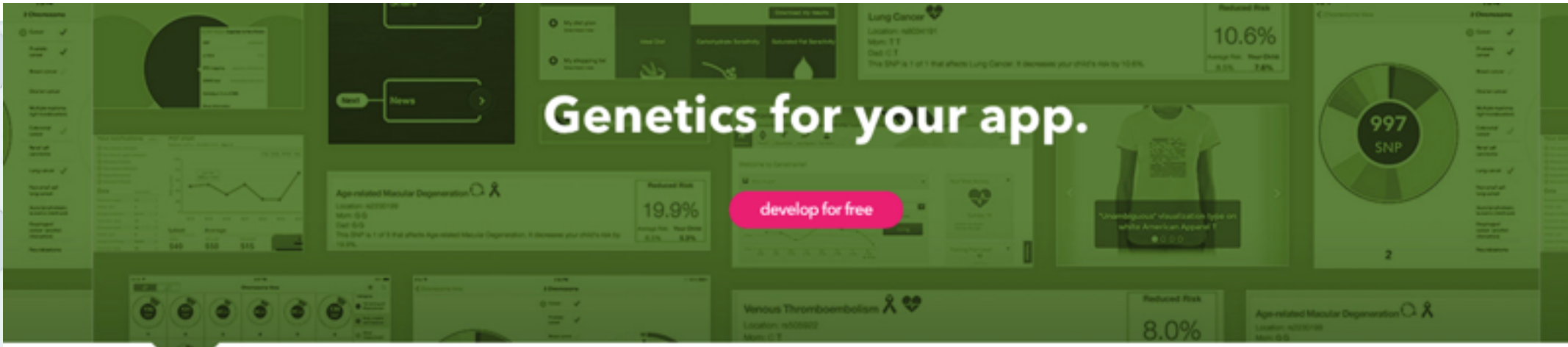
https://bcbio-nextgen.readthedocs.org/en/latest/

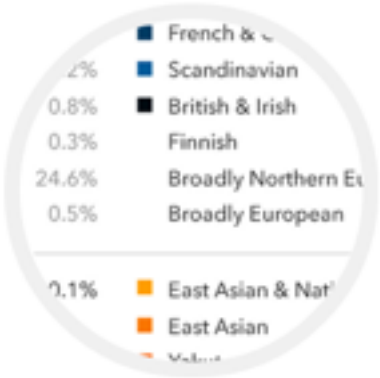
Bcbio-Nextgen (open-source)

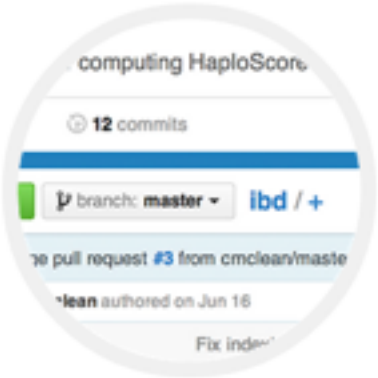


https://bcbio-nextgen.readthedocs.org/en/latest/

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

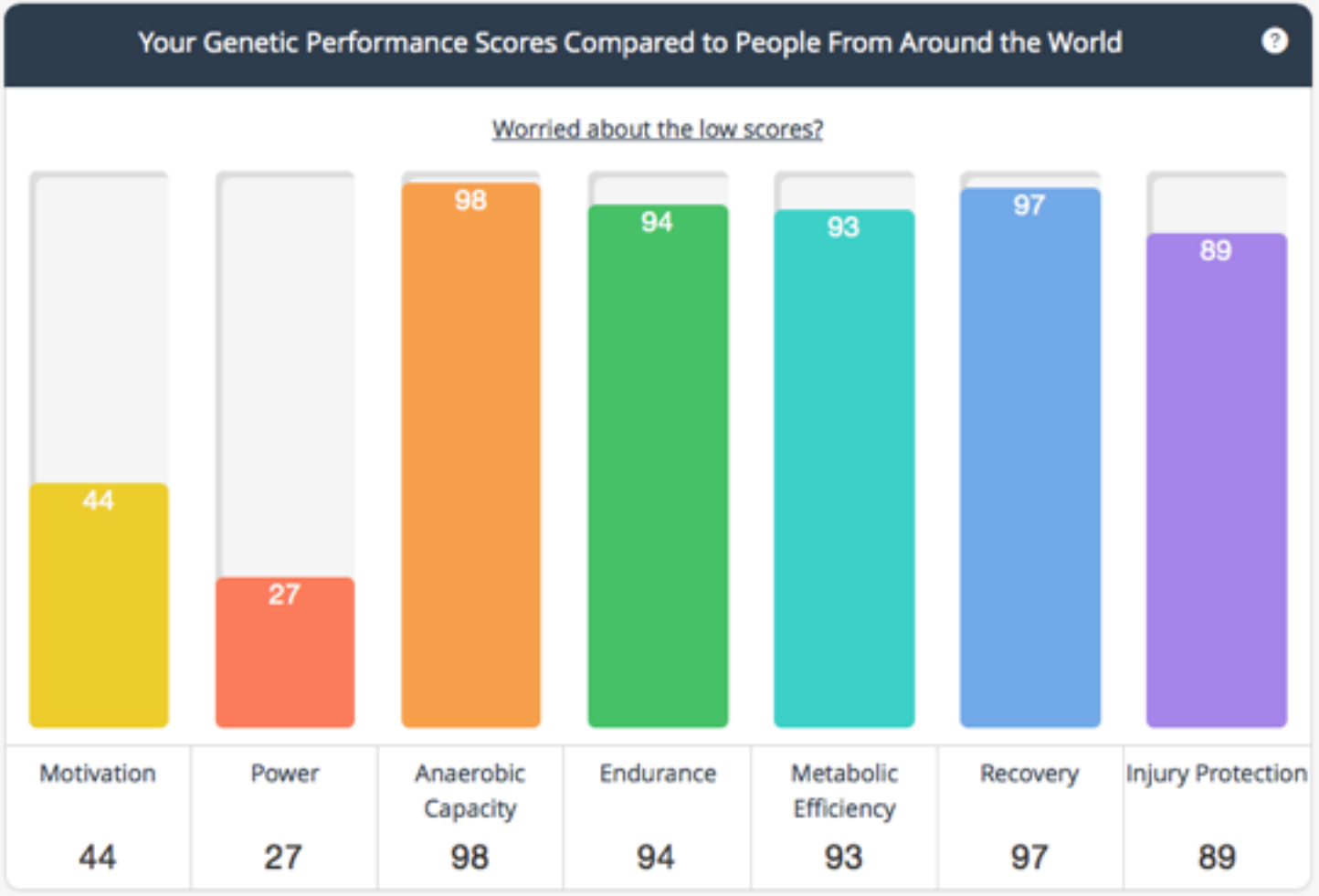
Our scientists have calculated ancestry and found relatives for genotyped customers. You could use this data without even knowing what a gene is!

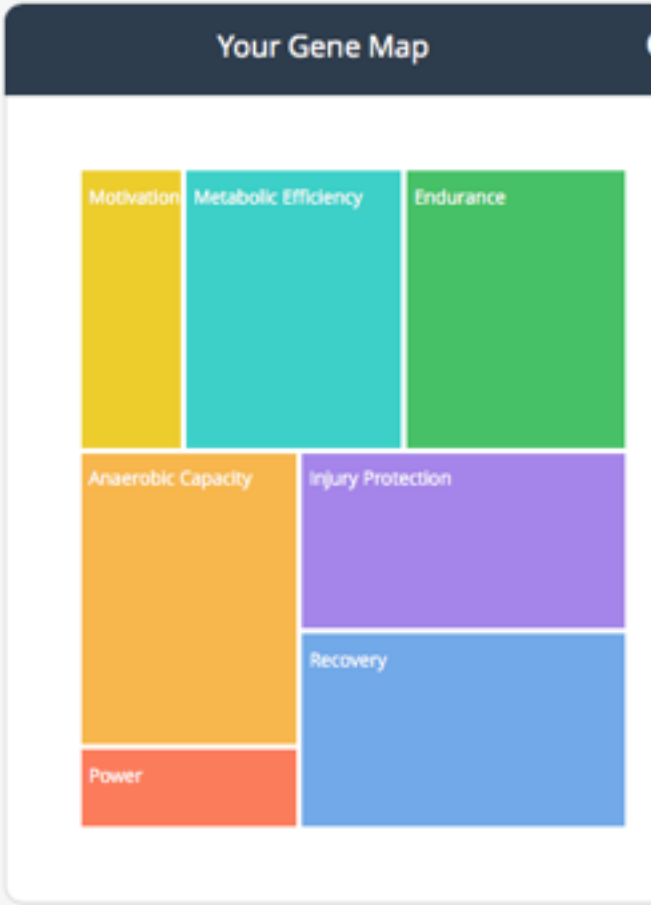
Build novel apps

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/

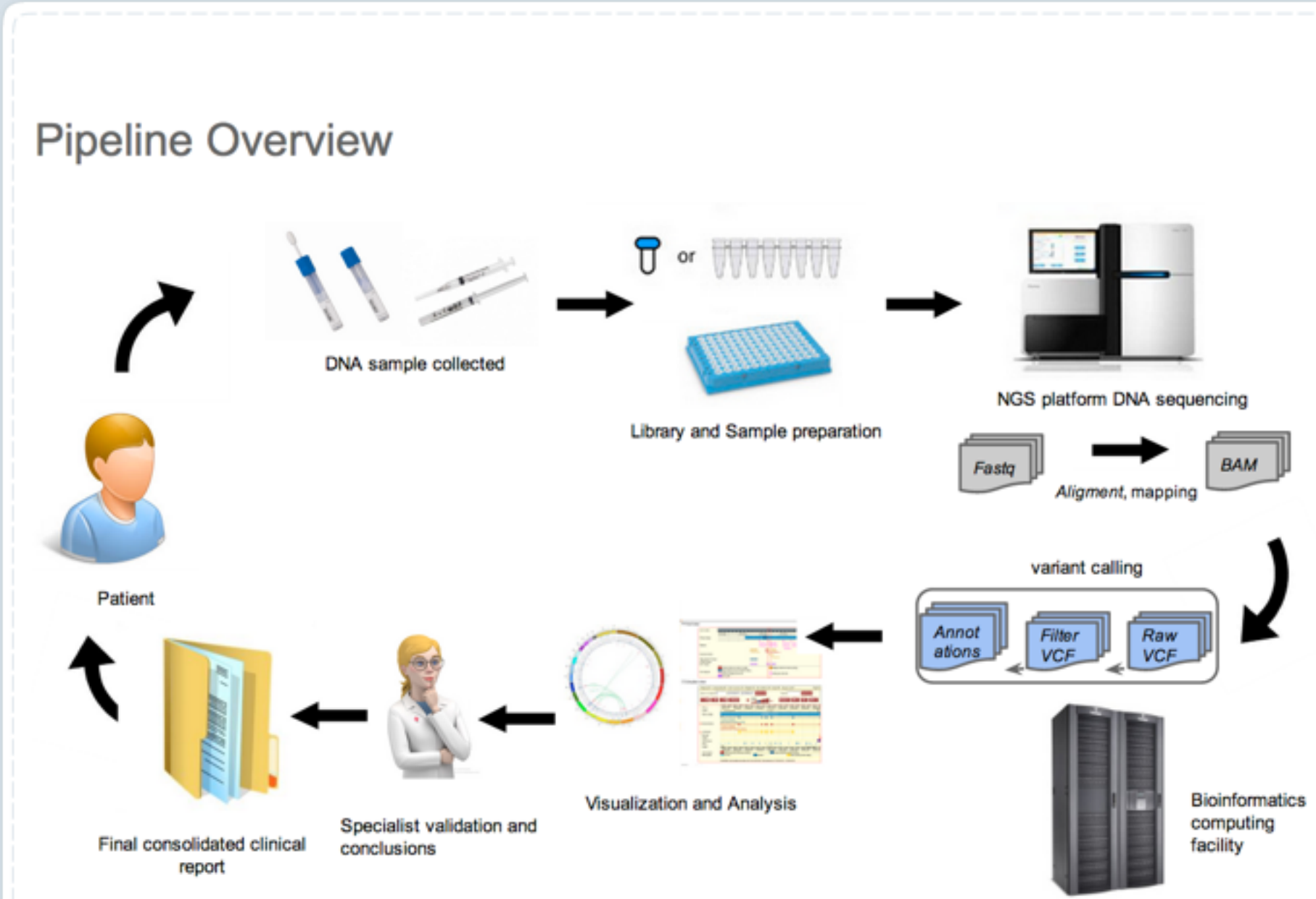
Athletigen (genome consumer apps)





https://www.athletigen.com/

Isto é apenas 1 componente!



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