OpenCB a next generation big data analytics and visualisation platform for the Omics revolution

Development at the University of Cambridge - Closing the Omics / Moore's law gap with Dell & Intel

Ignacio Medina, Paul Calleja, John Taylor (University of Cambridge, UIS, HPC Service (HPCS))

Abstract

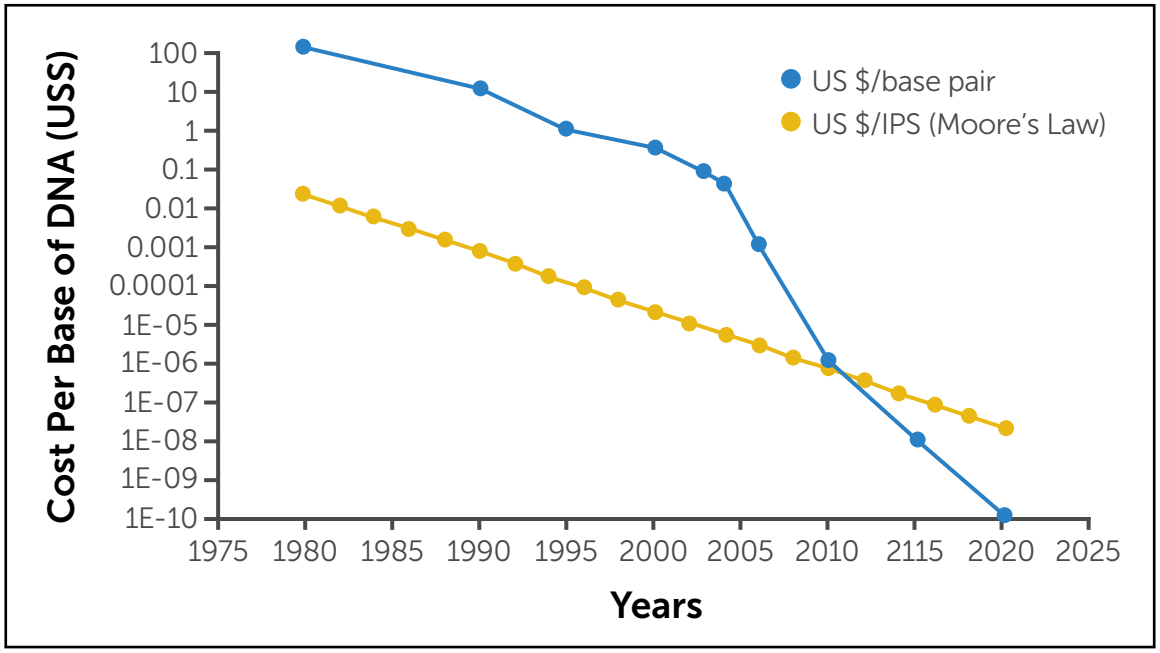
Within the life science industry we are seeing a widening omics / Moore's law gap produced by the much faster rate of improvement and consequent price drop seen in omics technologies as compared to compute & data technologies. As time goes on this is leading to an omics analysis bottle neck which is not sustainable since most of the current hardware and software platforms are not designed to work with current large data volumes. This paper introduced development work being undertaken at Cambridge to create a new state of the art omics analysis hardware and software platform utilizing the open source software framework called OpenCB and new high performance hardware from Dell & Intel. With this new OpenCB omics analysis platform current day high volume omics analysis problems become tractable. Such advancements in analytics platforms are vital in order to translate advances in population scale omics and medical informatics projects into personalized medicine technologies deployed in the clinic and finally realize a step change in human health that these technologies enable.

The Omics – Moore's Law gap

Current high-throughput technologies in genomics such as Next-Generation Sequencing (NGS) are generating "omics" data (transcriptomics, pharmacogenomics, etc) at an unprecedented scale with many clinical projects producing hundreds of TB to a few PB now being commonplace. Much of the existing software and hardware solutions in bioinformatics are not designed to work at these data volumes, thus inhibiting scalability and limiting efficiency and consequently making it very difficult for researchers to store, analyse, share and visualise data in a secure and collaborative manner.

As time moves on this problem will get more severe since the genome sequencing cost is falling faster than the cost of compute & data provision. Moore's Law states that the cost of compute and data, halves every 2 years, the cost of sequencing a full genome is halving every 4 months substantially quicker (see Figure below). In order to close this growing Omics – Moore's law gap, new Omics analytics platforms are required that combine new computational methods and new leading edge hardware and software technologies. This paper describes such computational methods being developed at Cambridge deployed on leading edge Dell and Intel hardware that offer to reduce this growing computational/data deficit for next-gen Omics.

The shear pace of advancement in omics and medical informatics deployment within the health care industry demands the creation of these next generation omics analytics platforms. This is vital in order to translate advances in population scale omics and medical informatics projects into personalized medicine technologies deployed in the clinic and finally realize a step change in human health that these technologies enable.



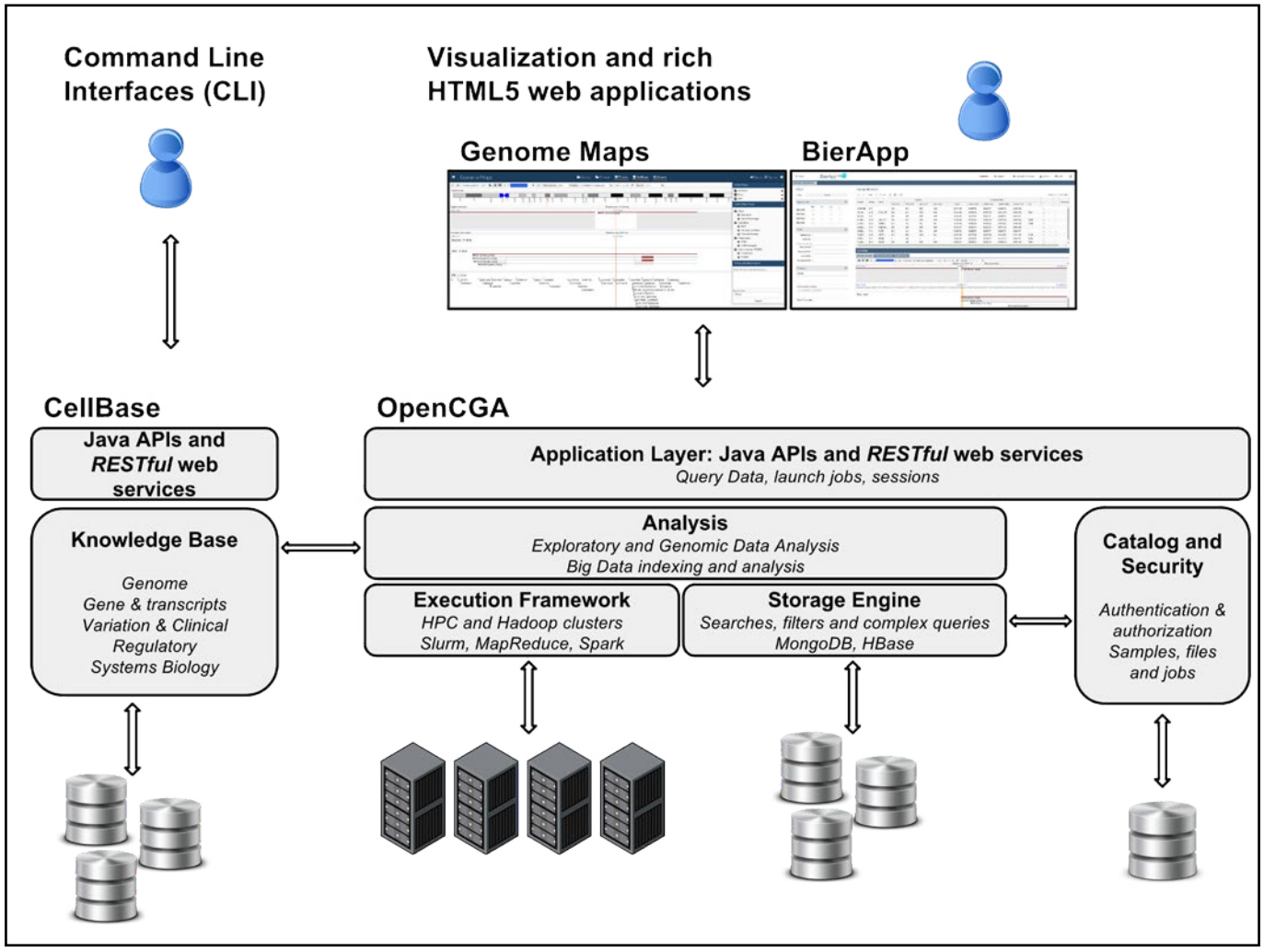
The increasing divide between Cost per Base of DNA vs. Moore's Law

The open-source software for Computational Biology (OpenCB) platform initiative offers a new capability in genomics analytics: to provide an extensible platform which aims to redress these challenges and provide a complete stack for big data in genomics. OpenCB is implemented using the state-of-the-art advanced High-Performance Computing (HPC), and Big Data technologies from Dell & Intel and is actively being developed at the University of Cambridge and other research institutes. With the OpenCB platform additional challenges relating to the integration of diverse "omics" data can be realized providing even more insight for clinicians and practitioners.

[1] Interpreted here as the cost per operation halving every two years

Closing the Gap with OpenCB

OpenCB was started in 2012 by Ignacio Medina now head of the Computational Biology Lab UIS Cambridge and is now used by many projects and research institutes. Currently it is developed by more than 10 active developers and researchers from University of Cambridge, EMBL-EBI and Genomics England among others. More information can be found at http://www.opencb.org/, OpenCB consists of different projects that solve different problems in current genomics, each of these projects constitutes a standalone solution than can be easily imported into existing projects. The projects have been designed to provide a scalable and high-performance solution for storing, processing, analysing, sharing and visualizing big data in genomics and clinics in a secure and efficient manner. To achieve this, OpenCB uses the most advanced computing technologies in HPC (such as SSE4/AVX2, GPUs, OpenMP) and Big Data (Hadoop, Spark) for data processing and analysis; NoSQL databases (MongoDB, HBase) for data indexing or HTML5 (SVG, IndexedDB) for interactive data visualisation. An overview of all the projects can be seen at https://github.com/opencb



OpenCB architecture. Server side stores, indexes and executes all the analysis. Client side HTML5 applications and CLI use RESTful web services to interact with the server

OpenCB Components

OpenCB consists of a number of projects as listed below:

High-Performance Genomics (HPG)

HPG projects make use of standard HPC and big data technologies to provide a scalable and efficient solution for several genomic analysis. The main HPG projects are:

a) HPG Aligner (https://github.com/opencb/hpg-aligner) is an ultra-fast and sensitive HPC Next-Generation Sequencing (NGS) read aligner. It combines advanced data structures and novel algorithms implemented with multi-threading and AVX2. Current work at Cambridge is being performed to explore Intel Xeon PHI.

b) HPG Variant (https://github.com/opencb/hpg-variant) is HPC software to process and analyse genomic variant data, several algorithms have been developed and implemented.

c) HPG BigData (https://github.com/opencb/hpg-bigdata) is a Hadoop MapReduce and Spark implementation of several genomics analyses for working with big data.

CellBase

CellBase (https://github.com/opencb/cellbase) constitutes the knowledge-based database for all OpenCB projects. CellBase is a NoSQL database that integrates the most relevant biological information about genomic features and proteins, gene expression, regulation, functional annotation, genomic variation and systems biology information. Its knowledge base relies on the most relevant repositories such as ENSEMBL, Uniprot, ClinVar, COSMIC or IntAct among others. CellBase has also a variant annotation built-in component that provides an Ensembl VEP compatible annotation. All data is available through a command line or by RESTful web services.

OpenCGA

OpenCGA (https://github.com/opencb/opencga) provides a scalable and high-performance solution for big data analysis and visualisation in a shared environment. OpenCGA integrates some of the OpenCB projects and implements, in addition, other components:

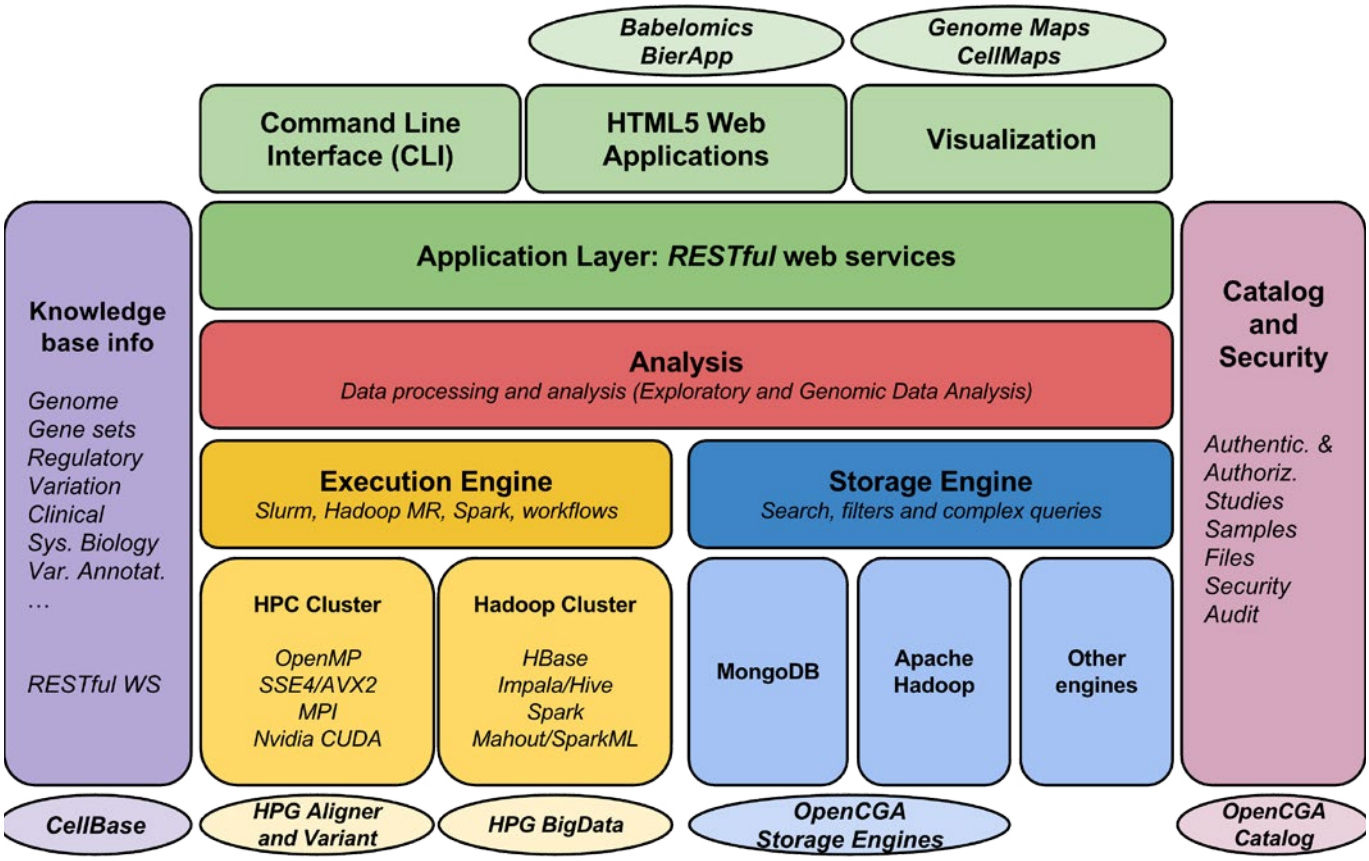
a) A storage engine framework to store and index alignments and genomic variants into different NoSQL such as MongoDB or Hadoop HBase - the current implementation can store efficiently thousands of gVCF files while remaining responsive when querying data.

b) A catalogue which keeps track of users, projects, files, samples, annotations, etc, and also provides authentication and authorisation capabilities.

c) Analysis engine to execute genomic analysis in a traditional HPC cluster or in Hadoop. OpenCGA has implemented a command line and a RESTful web services to manage and query all the data.

Visualisation with Genome Maps and CellMaps

Finally in OpenCB, a genome browser called Genome Maps (https://github.com/opencb/genome-maps) and a systems biology tool called CellMaps (https://github.com/opencb/cell-maps) provide a high-performance HTML5 SVG-based genome browser to interactively display CellBase data and OpenCGA indexed data such as BAM and VCF files. Users can also easily extend Genome Maps to display their own data and formats. OpenCB projects are compliant with the new GA4GH data models and formats.



An overview of main OpenCB components. Some OpenCB projects and tools in ovals

Who is using it

Many projects within research institutes around the world are using some OpenCB technologies demonstrating the success of this initiative. For instance ICGC, EMBL-EBI or Genomics England are using and contributing to some of this projects. Source code is open and it is freely available in GitHub at https://github.com/opencb

Future Developments

In order to ensure OpenCB maintains a cutting-edge platform for large-scale Genomics processing, access to state-of-the-art technologies is important. At Cambridge, the OpenCB team is exploring future processing capabilities offered by Xeon PHI and GPU technologies, non-volatile RAM as a means to effect larger in-memory processing capability as well as enhanced MapReduce capability to ensure that scalability is maintained. These technologies will be coupled with enhanced statistical analysis techniques to provide practitioners with even more insight into "omics". The University of Cambridge under the auspices of the UIS is working with a number of industry partners in this respect.