Mapping NGS reads for genomic studies
Where are we?

Mapping NGS reads for genomics studies

- Sequence preprocessing
- Mapping
- Variant Calling
- Variant prioritization
- Functional annotation
- Gene-Set Analysis
- GWAS Analysis

NGS pipeline
<table>
<thead>
<tr>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Introduction to OpenCB</td>
</tr>
<tr>
<td>- Intro to Git and GitHub</td>
</tr>
</tbody>
</table>

Ignacio Medina
imedina@cipf.es

Mapping NGS reads for genomic studies
Introduction to OpenCB

- OpenCB is an open-source software for Computational Biology
- Organized in 3 sub-projects
  - High-Performance Genomics (HPG): OpenMP, SSE (C language)
  - Cloud computing and NoSQL databases: Amazon AWS, MongoDB (Java language)
  - Big data visualization: HTML5, SVG (JavaScript language)
“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”

Big data is not a new scenario for other science areas: meteorology, physics, internet search, finance, business, ...

Which are the main Big data challenges?: curation, search, sharing, storage, analysis and visualization

We need to study and use new computational technologies available:

- High-Performance Computing (HPC): multi-core CPUs, SSE/AVX, GPUs
- Distributed computing: Apache Hadoop MapReduce, MPI
- Distributed and NoSQL databases: Apache Cassandra, HBase, ...
- Web apps: HTML5 (SVG, WebGL, ...), Javascript, RESTful WS, ...
- Clouds: Amazon AWS, Google Cloud, Microsoft Azure, ...
- Science: Machine learning, data mining, clustering, probabilistic graphicals models, visualization, ...
Introduction to Git and GitHub

- Git source control Manager
  - http://git-scm.com/
- GitHub as hosting
  - https://github.com/
- Projects:
  - HPG: https://github.com/opencb-hpg
  - Cloud: https://github.com/opencb-cloud
  - Visualization: https://github.com/opencb-bigdata-viz