# BMI 705 Precision Medicine II: Integrating Clinical and Genomic Data

# Using Hail, a genomic variant store in the Cloud

Alba Gutiérrez-Sacristán - Inès Krissaane

December 13, 2018

# Contents

What you need to know about genetic data

The Human Genome

What is a variant?

Genome-wide association studies

Perform scalable genomic data analysis

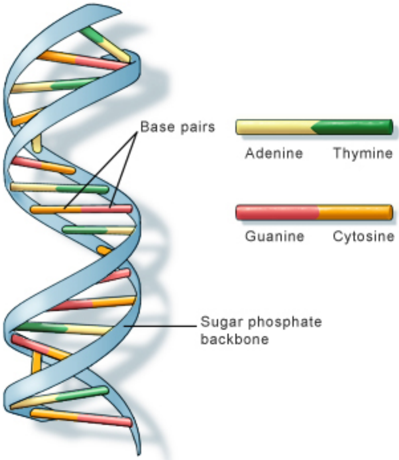
A Big Data Science

Google Cloud Deployment

Hail, a genomic Variant Store

# The Human Genome

- 3.2 billion DNA bases pairs
- 23 chromosomes pairs
- 1.5% codes for 20k genes
- 98.5% is non coding



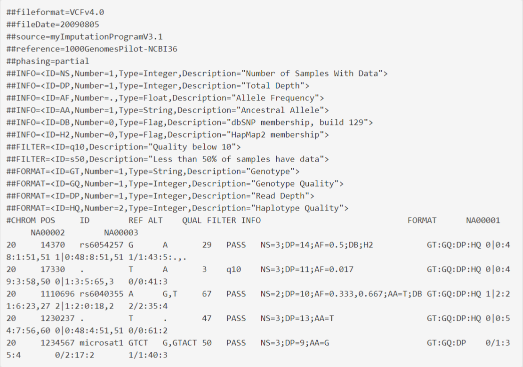
U.S. National Library of Medicine

## What is a variant?

- Genetic variations, or **variants**, are the differences that make each person's genome unique. DNA sequencing identifies an individual's variants by comparing the DNA sequence of an individual to the DNA sequence of a reference genome.
- Some contribute to differences between humans like eye color and blood type. A small number of variants have been linked with disease.

# Variant Call Format

**Variant Call Format** is a text file format with meta information lines, a header line, and data lines each containing information about a position in the genome.

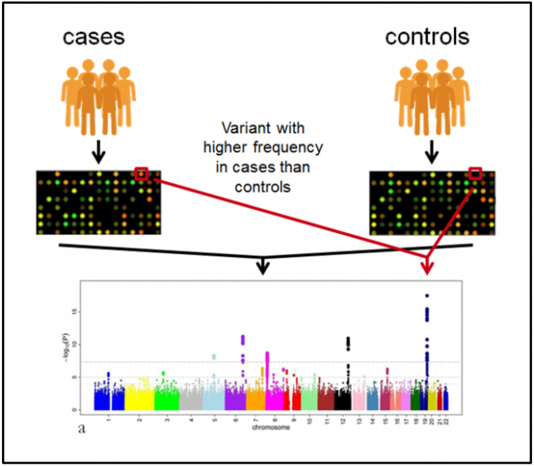


```
##fileformat=VCFv4.0
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=1000GenomesPilot-NCBI36
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF ALT    QUAL FILTER INFO                              FORMAT      NA00001
     NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:4
8:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:4
9:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:2
1:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:5
4:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTCT   G,GTACT 50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:3
5:4       0/2:17:2       1/1:40:3
```

Example of VCF from 1000 Genomes Project

# What is a genome-wide association studies (GWAS)?

- Genome-wide association studies (GWAS) are hypothesis free methods to identify associations between genetic regions (loci) and traits (including diseases).
- Variants associated with a trait will be found at a higher frequency in cases than controls.
- Statistical analysis is carried out to indicate how likely a variant is to be associated with a trait.



Extracted from https://www.ebi.ac.uk/training/online/course/gwas-catalog-exploring-snp-trait-associations/why-do-we-need-gwas-catalog/what-are-genome

# A Big Data Science

| 1000 Genomes Releases | Variants | Individuals | Populations |
|---|---|---|---|
| Phase 3 | 84.4 million | 2504 | 26 |
| Phase 1 | 37.9 million | 1092 | 14 |
| Pilot 1 | 14.8 million | 179 | 4 |

Too many Data = A failure to scale

http://www.internationalgenome.org/

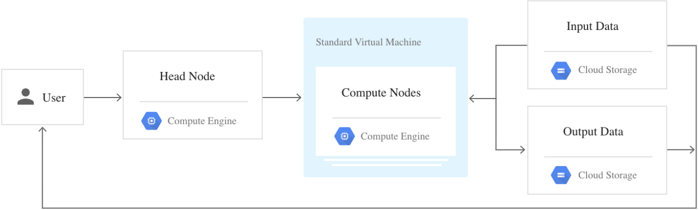
# The Big Data Problem

- Single machine can no longer process or store all this data !
- Only solution is to distribute over large clusters.
- This involves virtual cluster deployment, monitoring and managing large-scale clusters on the cloud.

# Using clusters for large-scale computing in the cloud

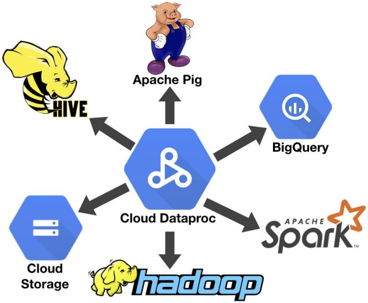
Cluster computing aggregates and coordinates a collection of machines to work together to solve a task. Clusters typically have a single head node and some number of compute nodes. The head node is the brains of the system and is responsible for:

1. Registering compute nodes into the system.
2. Monitoring the nodes.
3. Allocating jobs to particular nodes.



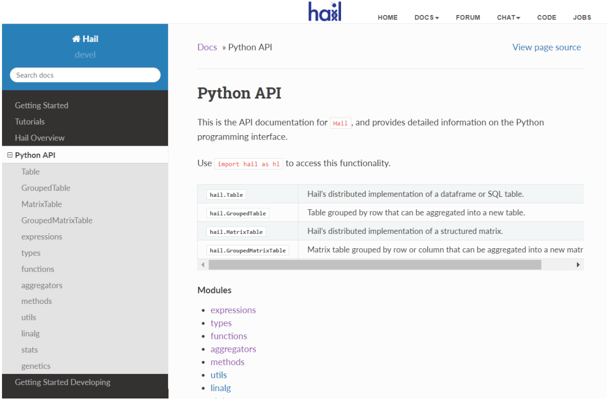
# Google Cloud Dataproc

Cloud Dataproc is a fast, easy-to-use, fully-managed cloud service for running Apache Spark and Apache Hadoop clusters in a simpler and more cost-efficient way.



# Hail, a genomic Variant Store

- Open-source, modular, scalable platform for statistical genetics in developed by the Neale lab at Broad Institute
- Exposed through Python and backed by distributed algorithms built on top of Apache Spark



## Python API

This is the API documentation for `Hail`, and provides detailed information on the Python programming interface.

Use `import hail as hl` to access this functionality.

| | |
|---|---|
| `hail.Table` | Hail's distributed implementation of a dataframe or SQL table. |
| `hail.GroupedTable` | Table grouped by row that can be aggregated into a new table. |
| `hail.MatrixTable` | Hail's distributed implementation of a structured matrix. |
| `hail.GroupedMatrixTable` | Matrix table grouped by row or column that can be aggregated into a new matr |

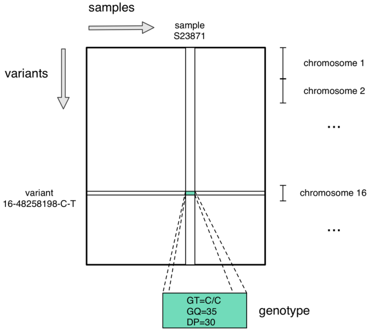
Modules

- expressions
- types
- functions
- aggregators
- methods
- utils
- linalg

Python API for Hail

# From VCF to MT file

A **Matrix Table (MT)** is a huge matrix, where rows are keyed by variant, and columns by sample. Each sample is from an individual and an individual may have many samples taken from them for sequencing.



Hail Matrix Table Format

# Documentations

| | |
|---|---|
| Docs, tutorials, code | hail.is |
| Forum, chat | discuss.hail.is |
| Hail Deployment on Google Cloud | github.com/hms-dbmi/Hail-on-Google-Cloud |