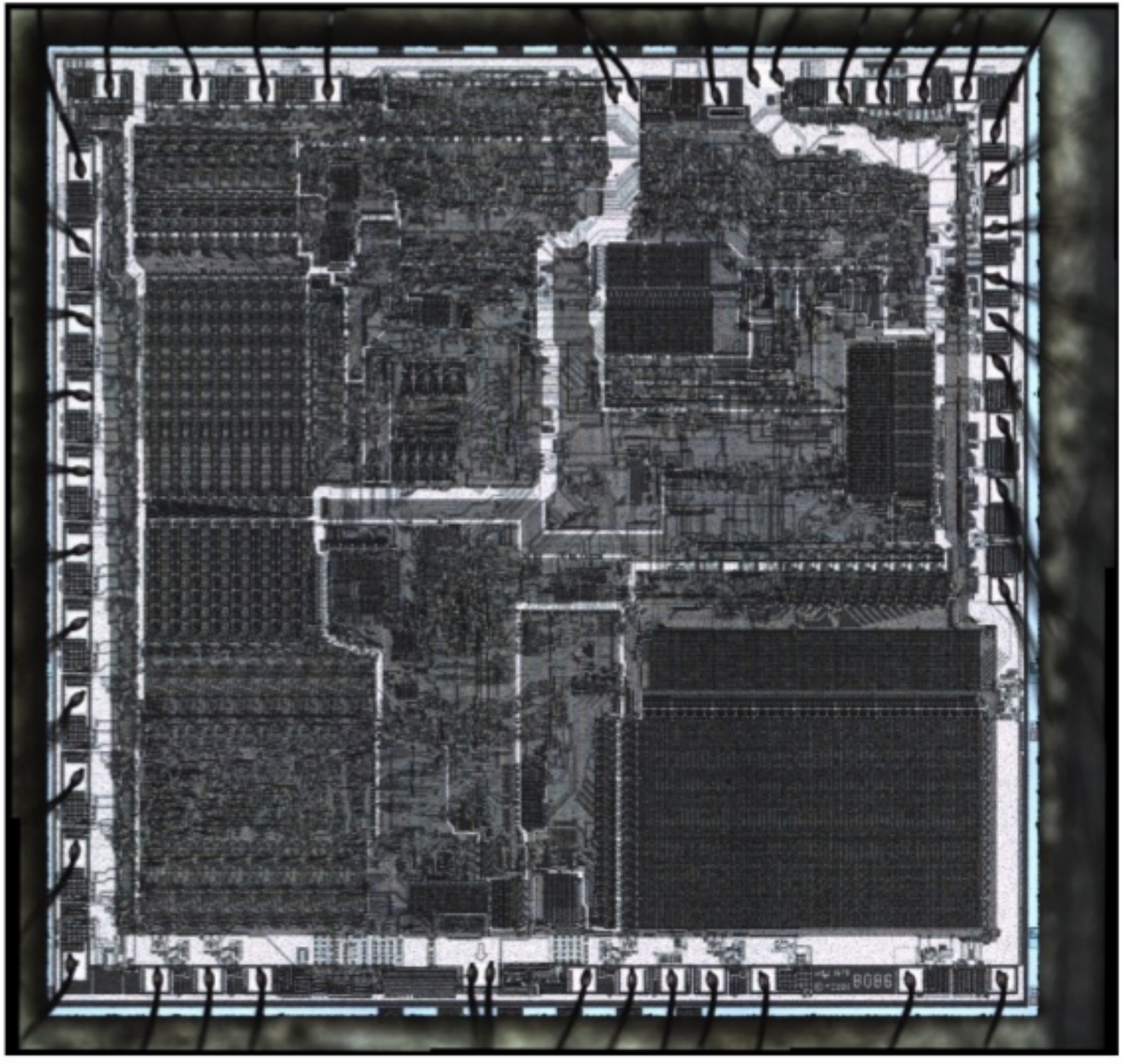
Scaling Genetic Data Analysis with Hail and Apache Spark

Jon Bloom and Tim Poterba,
Bio-curious Mathware Engineers

Initiative in Scalable Analytics
at the
Broad Institute of MIT and Harvard

…in the heart of biomedical research and technology:

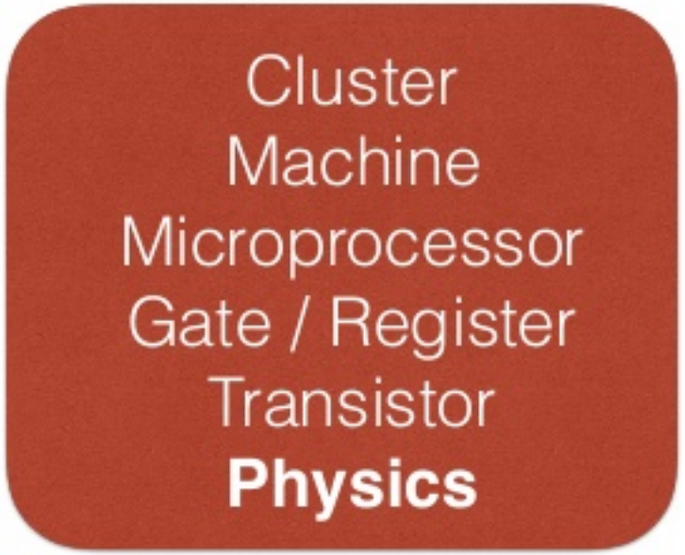
…in the heart of biomedical research and technology:

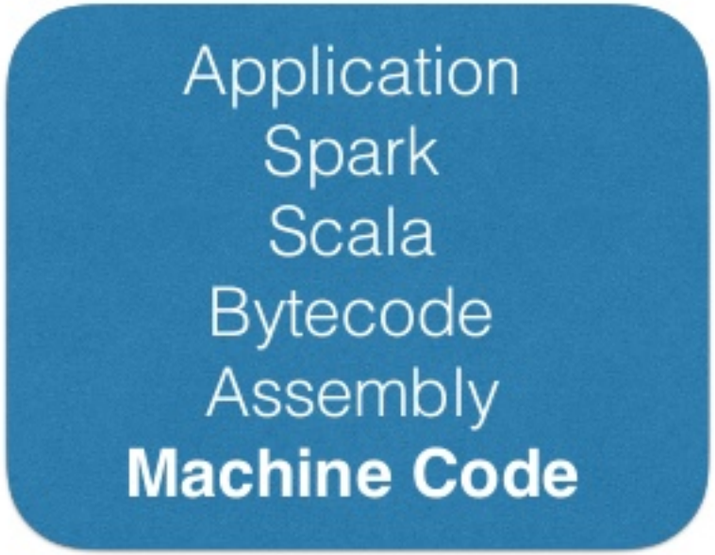


http://visual6502.org/images/pages/Intel_8086_die_shots.html

Computers are complex

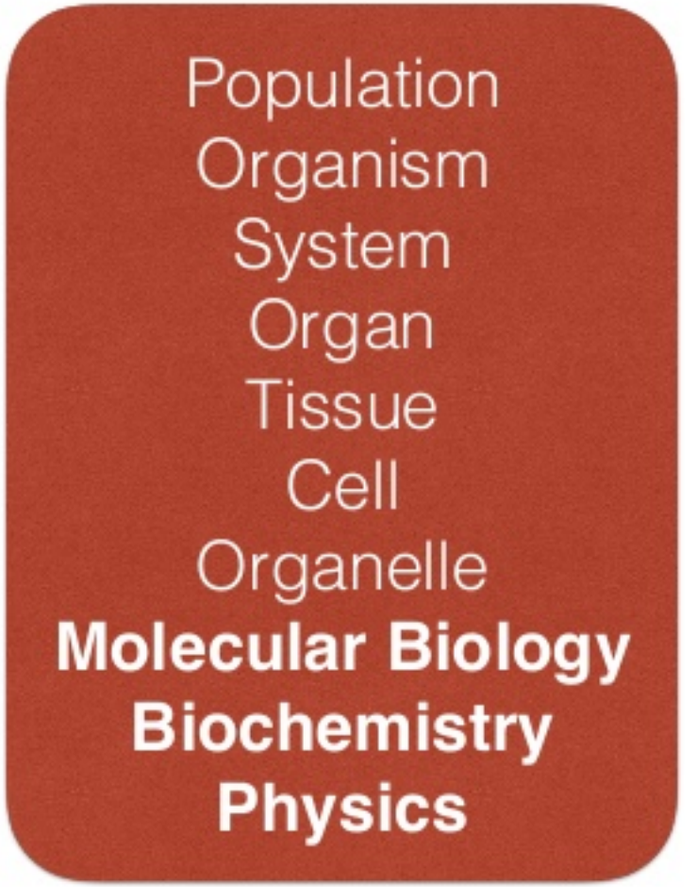
- Hardware and software
- Many levels of abstraction
- Execution at all levels

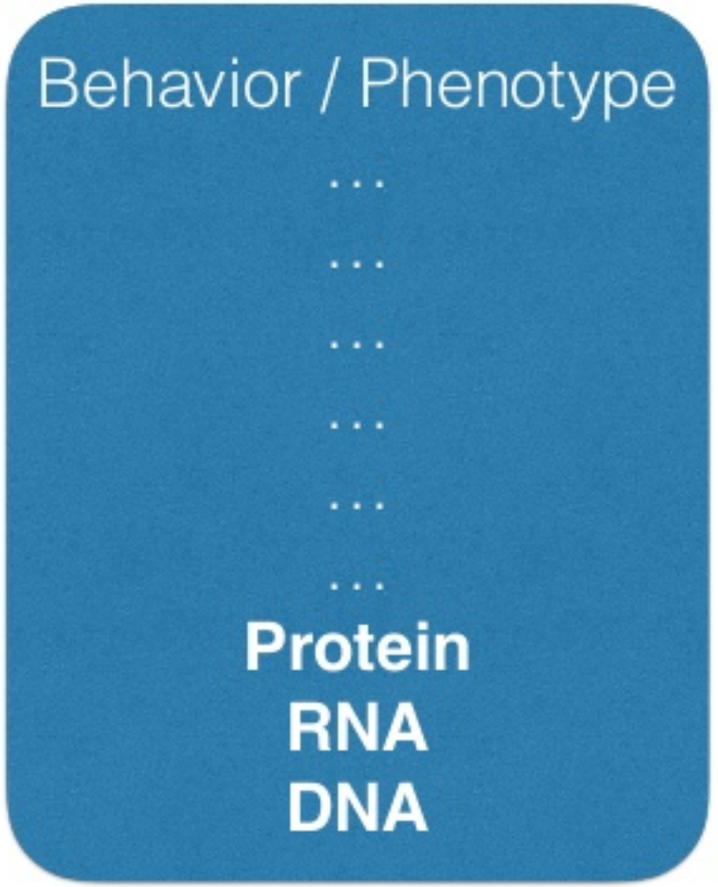




Biology is *ridiculously* complex

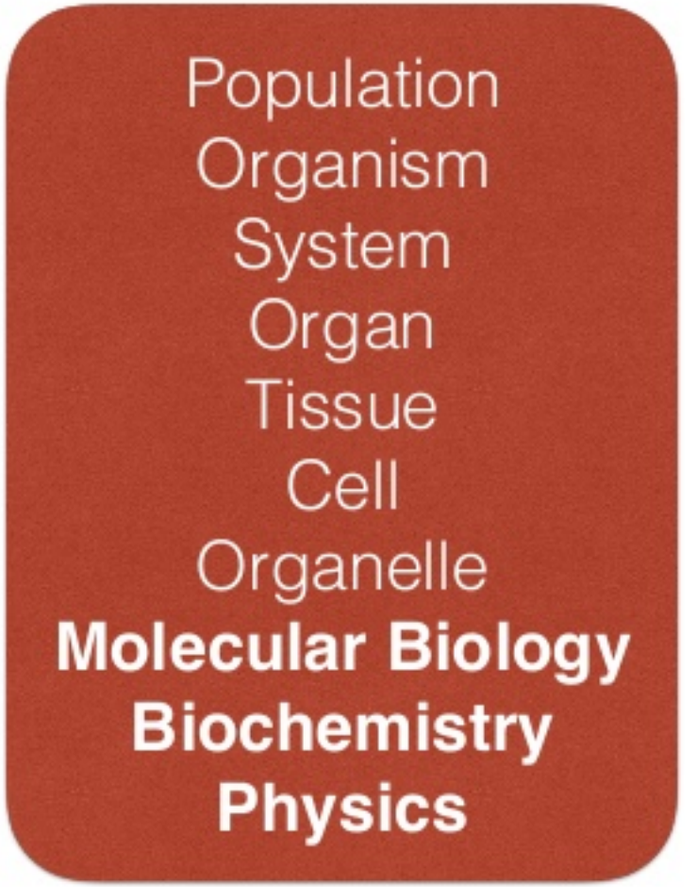
- Wetware and software
- Many levels of abstraction
- Execution at all levels

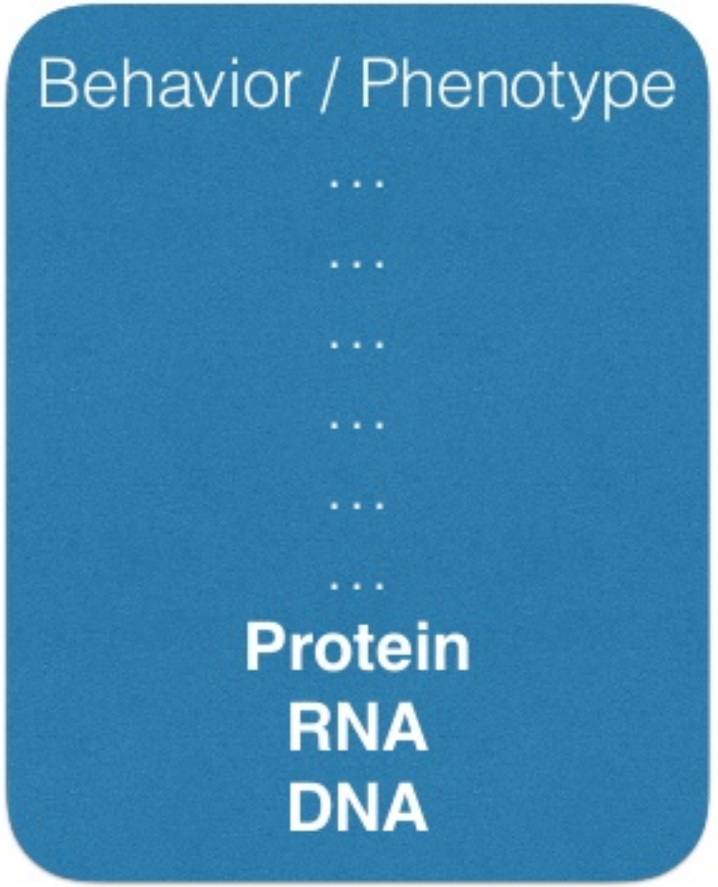




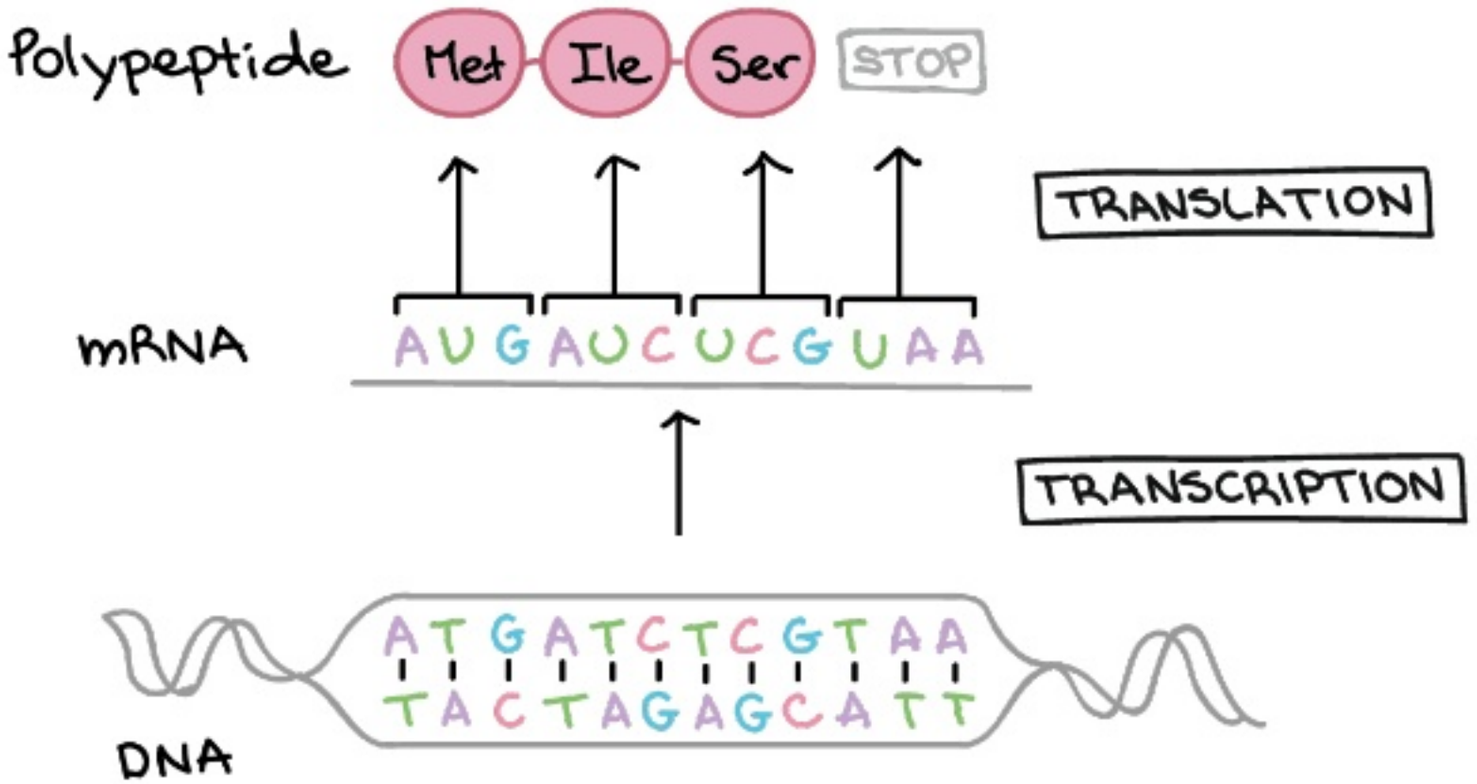
Biology is *ridiculously* complex

- Wetware and software
- Many levels of abstraction
- Execution at all levels
- We designed computers. Biology evolved us!





Biology is *ridiculously* complex



Behavior / Phenotype

...

...

...

...

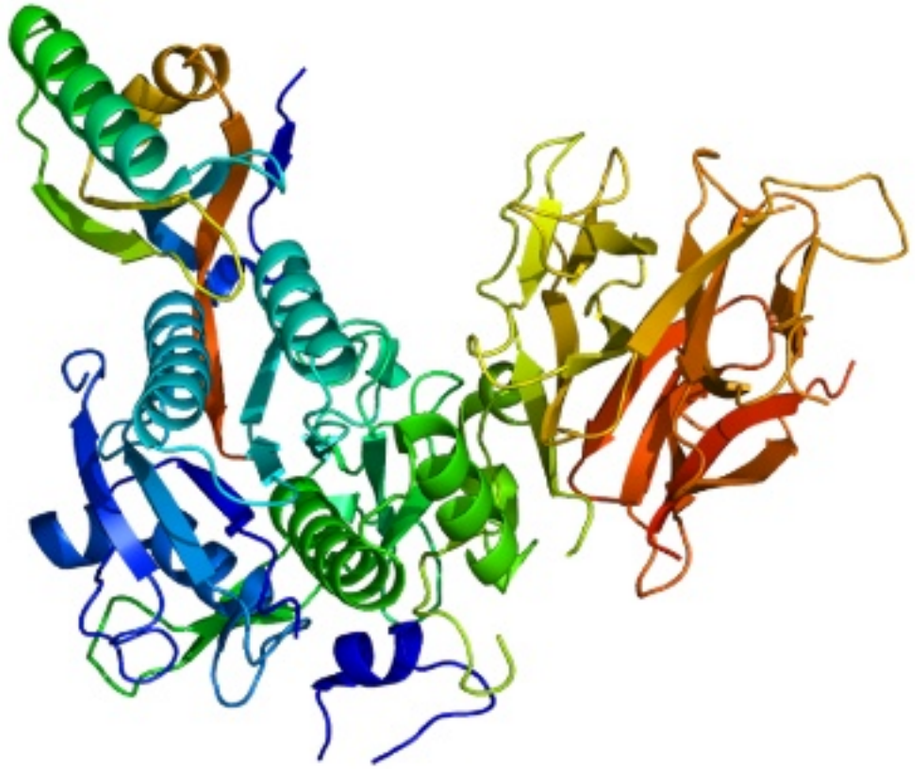
...

...

Protein
RNA
DNA

Kahn Academy

Biology is *ridiculously* complex



PCSK9

Behavior / Phenotype

…

…

…

…

…

…

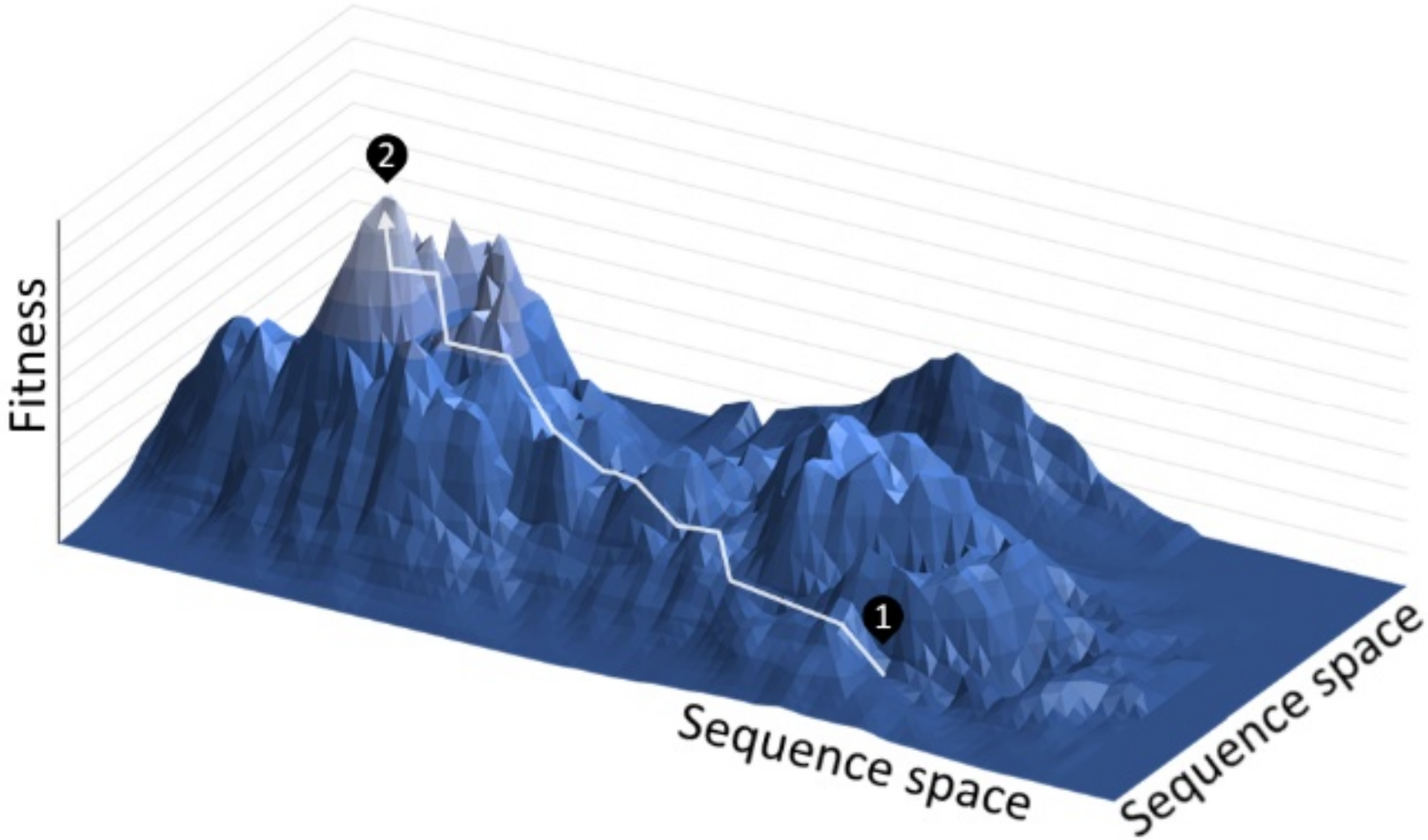
Protein
RNA
DNA

Why understand biology?

- Science!

Why understand biology?

- Science!
- Technology!



https://www.youtube.com/watch?v=U7Bh41YnfB0&list=PLIMMtIgw6qNjROoMNTBQjAcdx53kV50cS

Why understand biology?

- Science!
- Technology!
- Immortality!

Calico

Singularity
UNIVERSITY

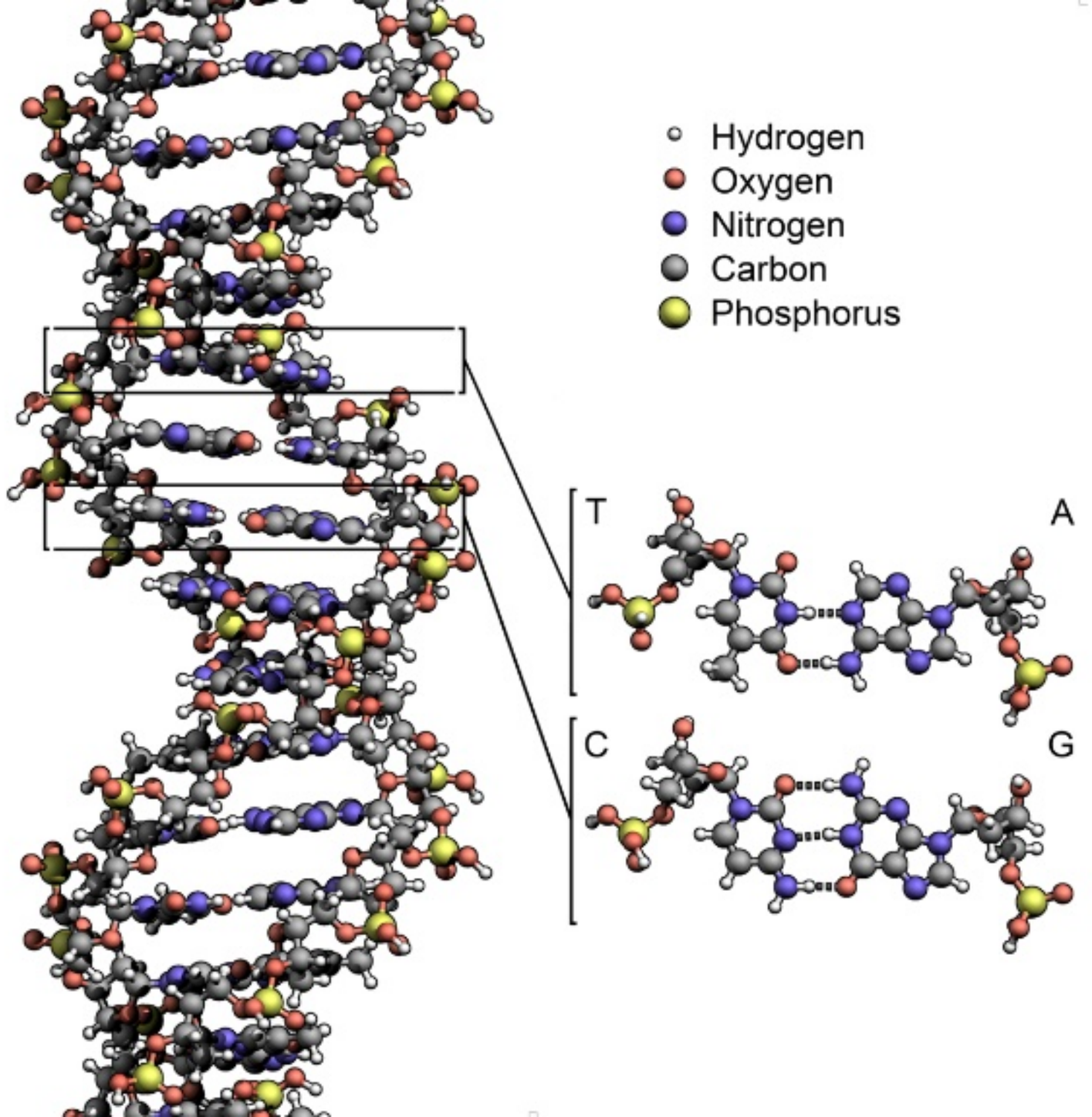
Why understand biology?

- Science!
- Technology!

- To decipher the wetware and software stack of human biology in order to prevent, diagnose, and treat disease.

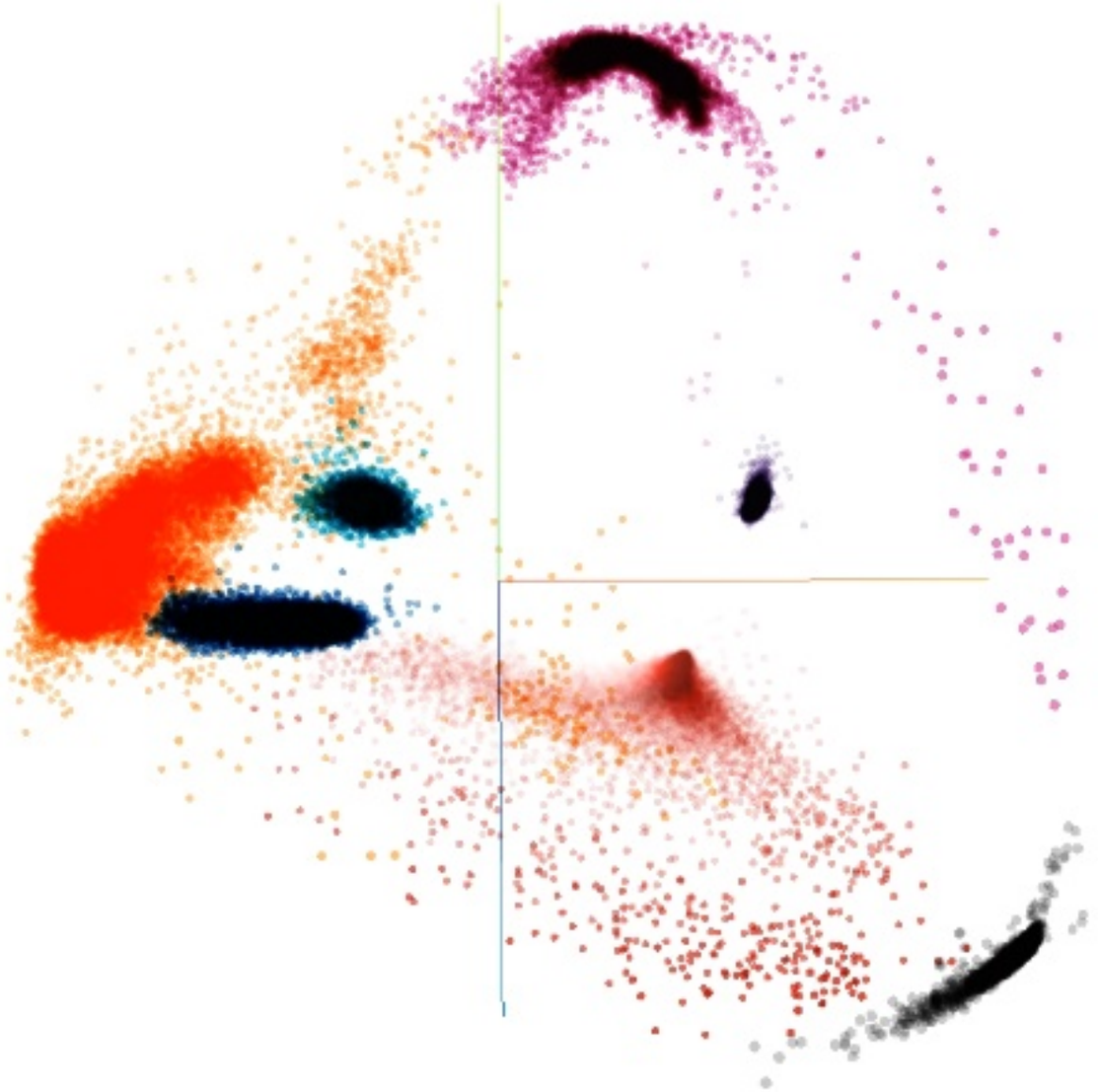
Genetics

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



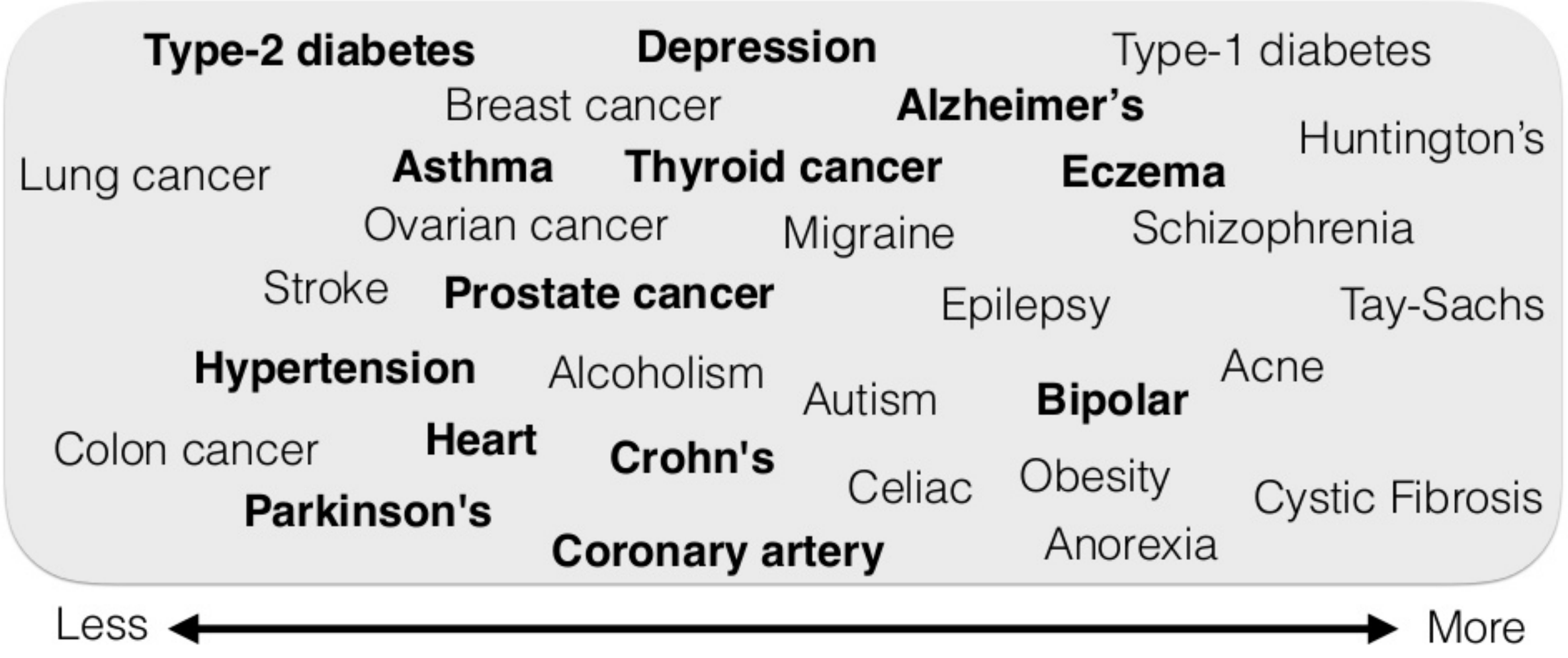
Genetics

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



gnomad.broadinstitute.org

Most diseases are heritable



https://www.snpedia.com/index.php/Heritability

The New York Times

Cholesterol-Slashing Drug Can Protect High-Risk Heart Patients, Study Finds

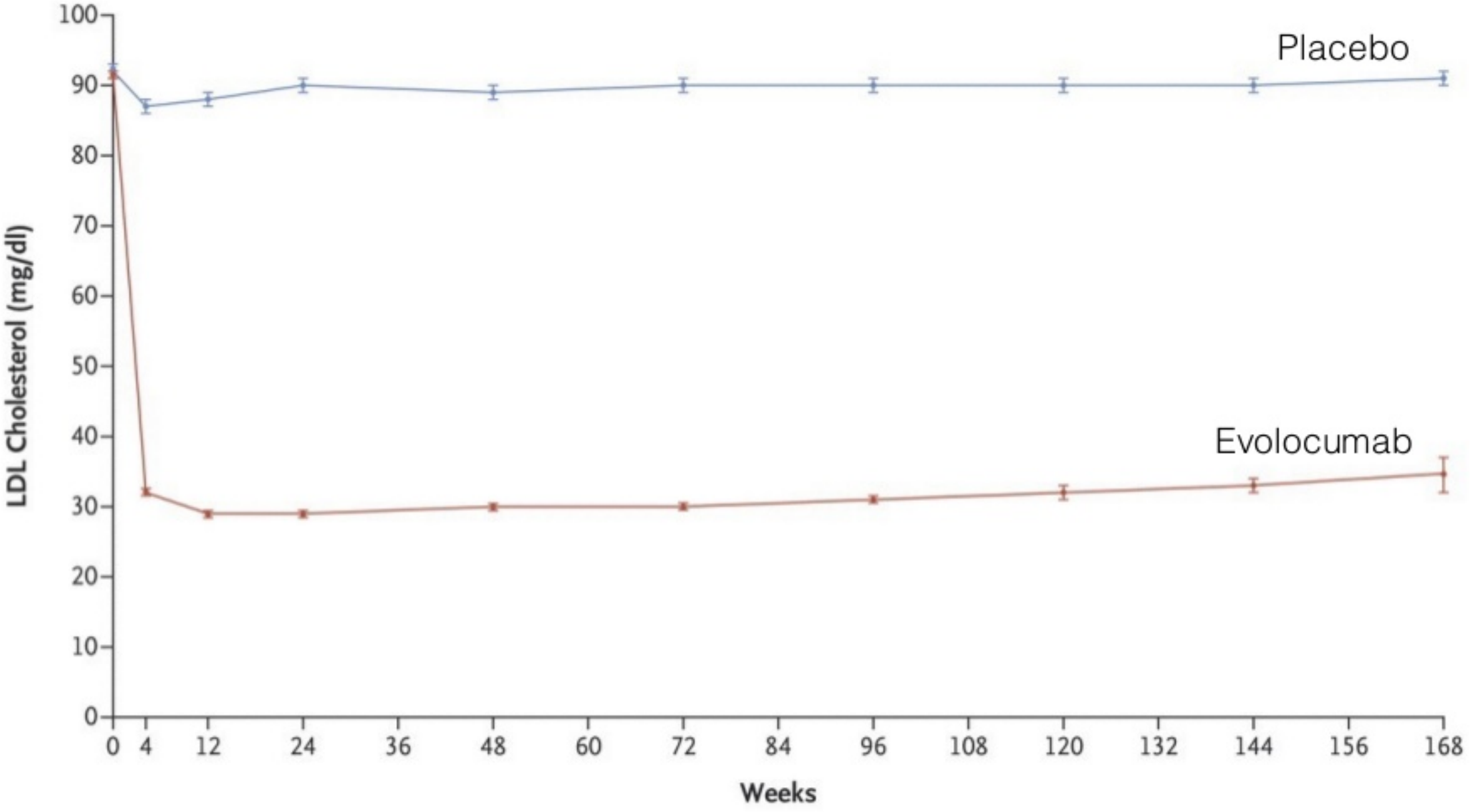
By **GINA KOLATA** MARCH 17, 2017

The first rigorous test of an expensive new drug that radically lowers cholesterol levels found that it significantly reduced the chance that a high-risk patient would have a heart attack or stroke. These were men and women who had exhausted all other options.

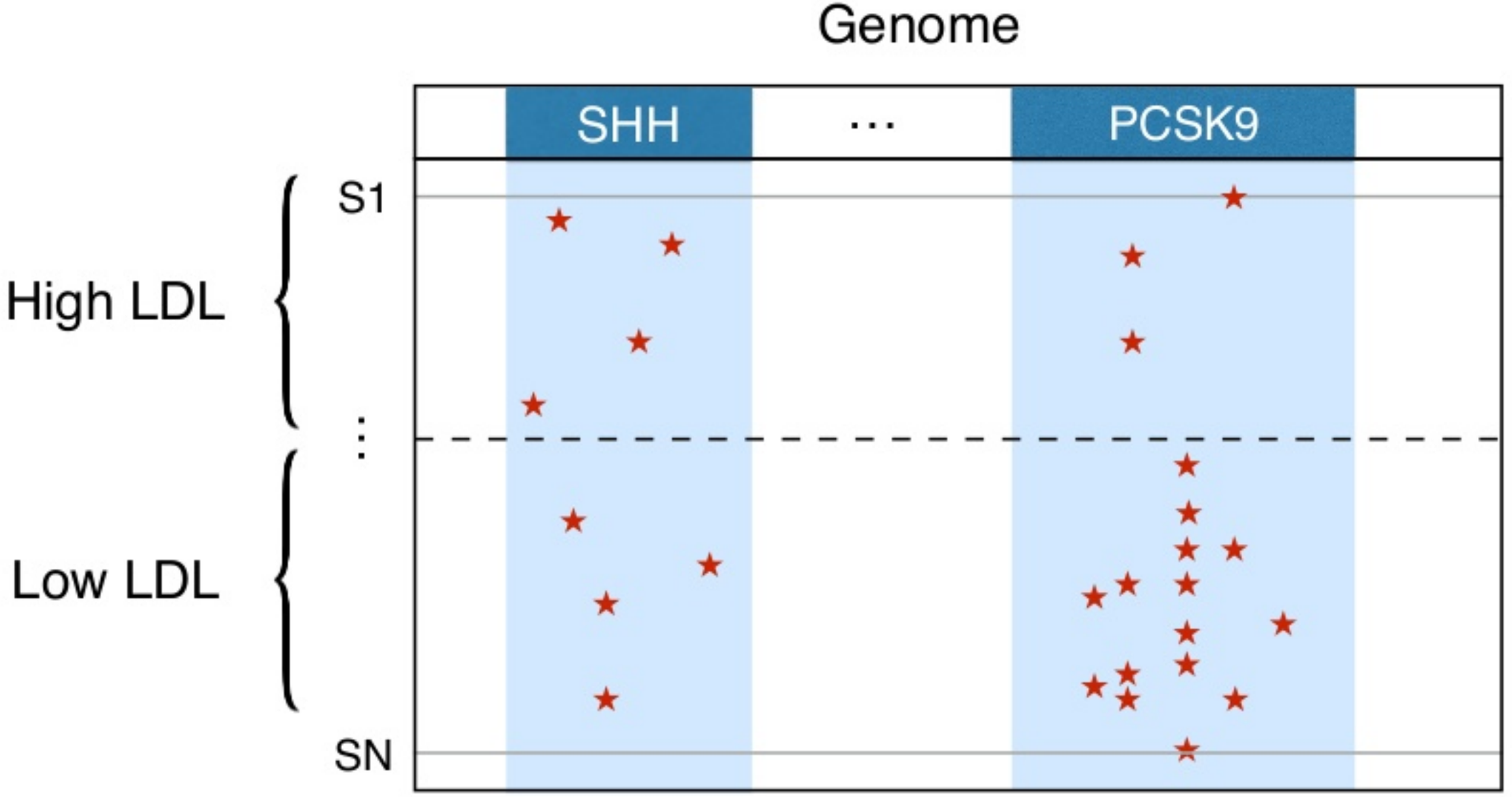
The drug, Repatha, is called a PCSK9 inhibitor and can make cholesterol tumble to levels almost never seen naturally in adults, or even in people taking cholesterol-lowering statins.

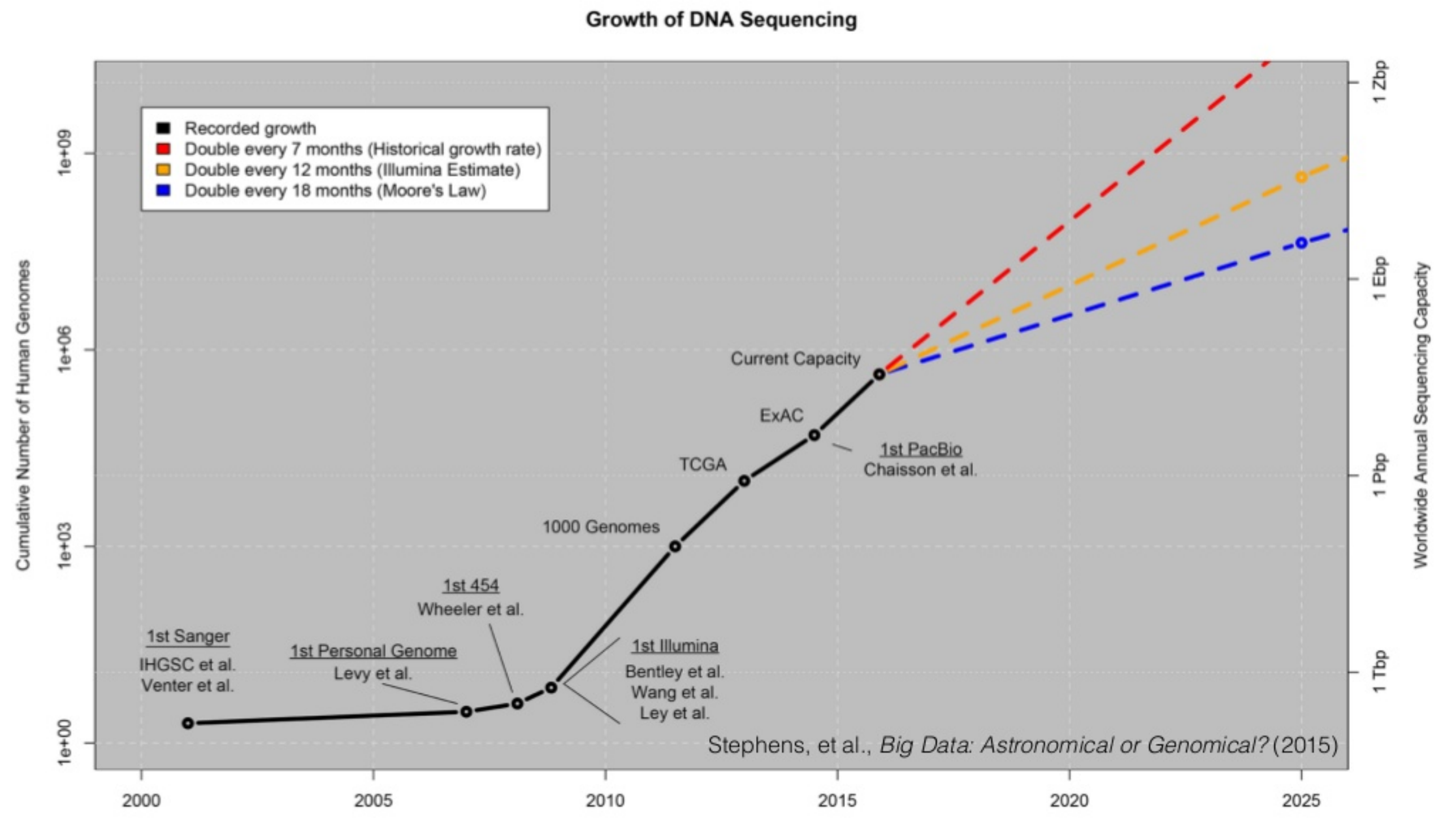
The NEW ENGLAND JOURNAL of MEDICINE

Evolocumab and Clinical Outcomes in Patients with Cardiovascular Disease



Statistical genetics




Growth of DNA Sequencing
Recorded growth
Double every 7 months (Historical growth rate)
Double every 12 months (Illumina Estimate)
Double every 18 months (Moore's Law)
Cumulative Number of Human Genomes
1e+00
1e+03
1e+06
1e+09
Worldwide Annual Sequencing Capacity
1 Tbp
1 Pbp
1 Ebp
1 Zbp
2000
2005
2010
2015
2020
2025
1st Sanger
IHGSC et al.
Venter et al.
1st Personal Genome
Levy et al.
1st 454
Wheeler et al.
1st Illumina
Bentley et al.
Wang et al.
Ley et al.
1000 Genomes
TCGA
ExAC
Current Capacity
1st PacBio
Chaisson et al.
Stephens, et al., Big Data: Astronomical or Genomical? (2015)

Broad Institute data

- The Broad sequences **1 genome every 10 minutes**.
- The Broad generates **17 TB** of new genomes per day.
- The Broad manages **45 PB** of scientific data.

Metabolome

Proteome

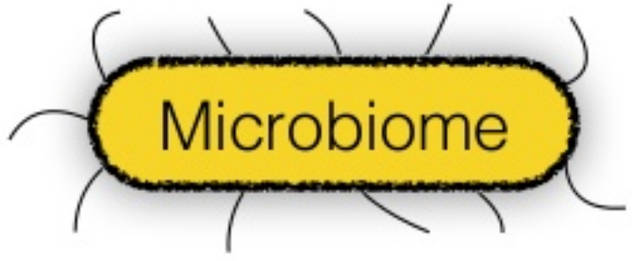
Transcriptome

Epigenome

Genome

Broad Institute data

- The Broad sequences **1 genome every 10 minutes**.
- The Broad generates **17 TB** of new genomes per day.
- The Broad manages **45 PB** of scientific data.



Computational Experiments

Science

Implementation

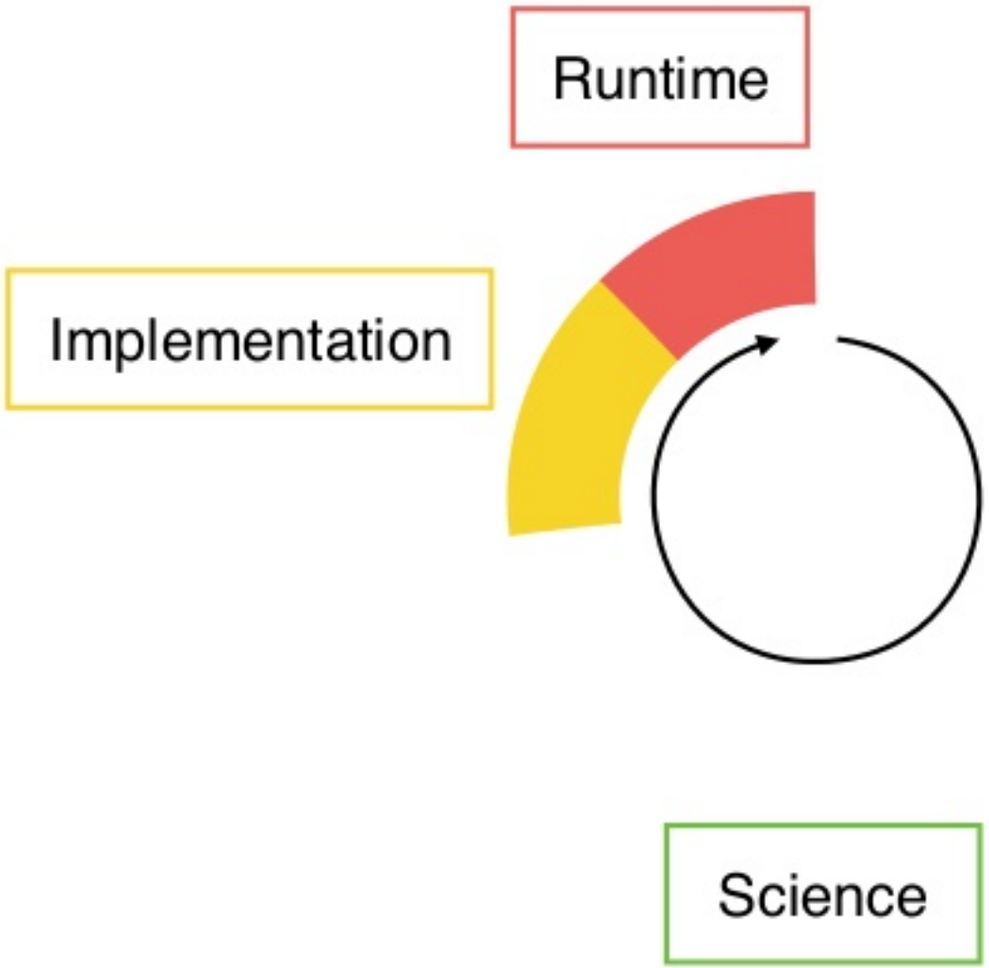
Runtime

Failure to Scale

Science

Implementation

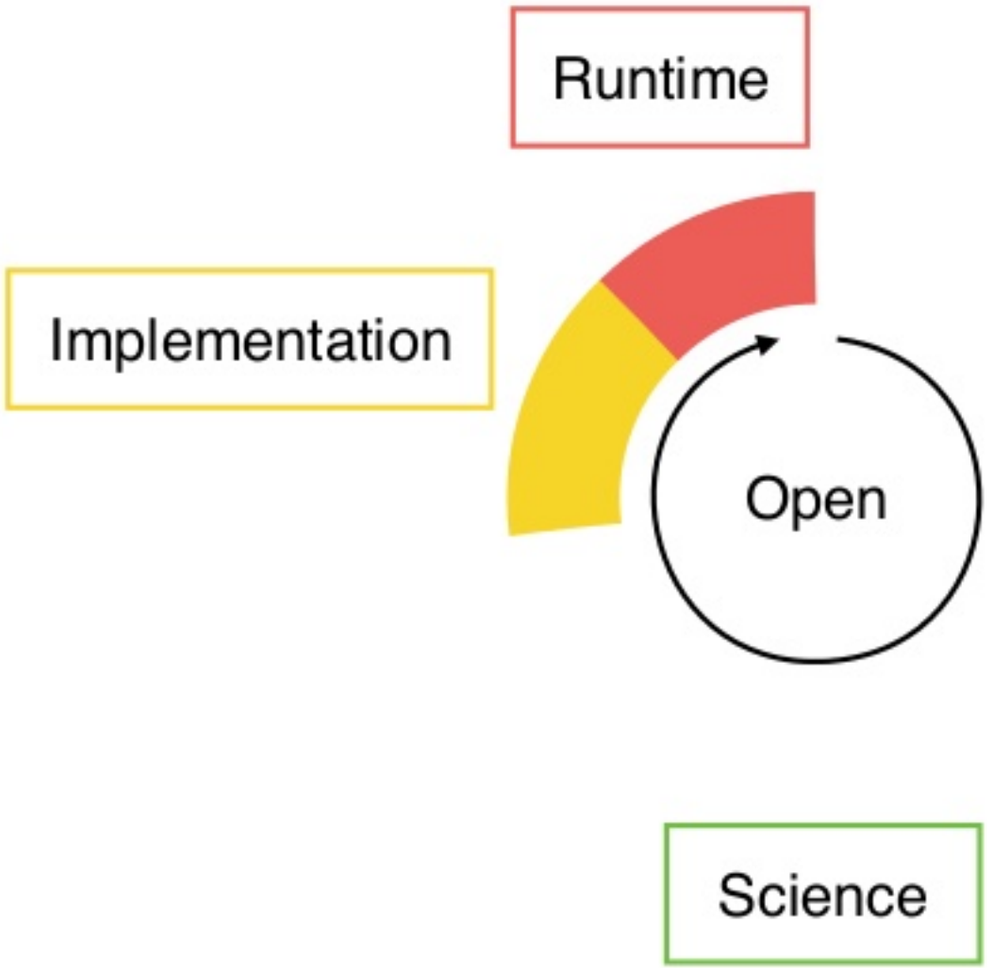
Runtime



Powerful and flexible

Domain-relevant, easy to use

Fast and scalable



Powerful and flexible

Domain-relevant, easy to use

Fast and scalable

Open-source and open-dev!

Drop the latency of computational experiments to rev the engine of biomedical science!

Runtime

Implementation

Science

Powerful and flexible

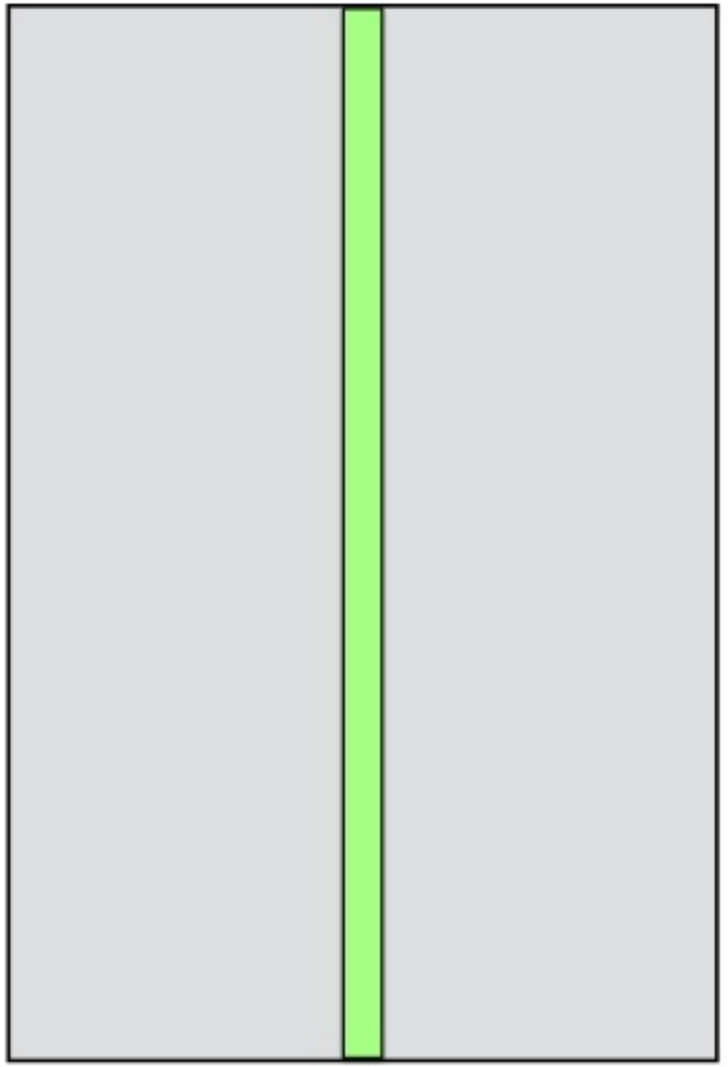
Domain-relevant, easy to use

Fast and scalable

Open-source and open-dev!

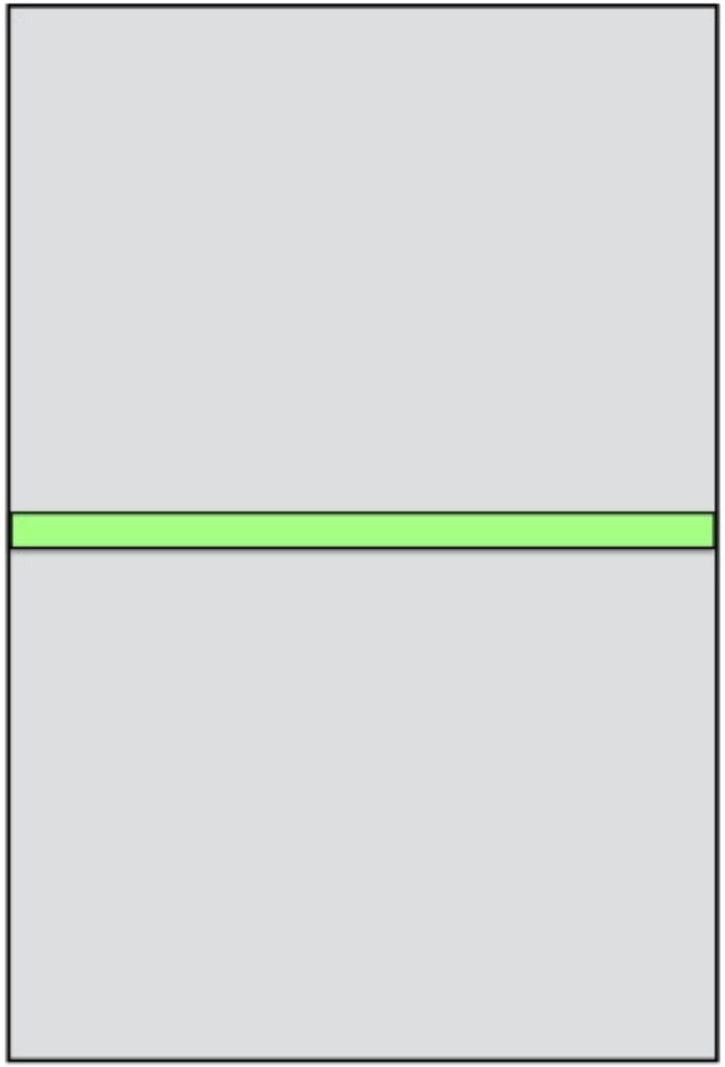
Individual ID

`"NA12878"`



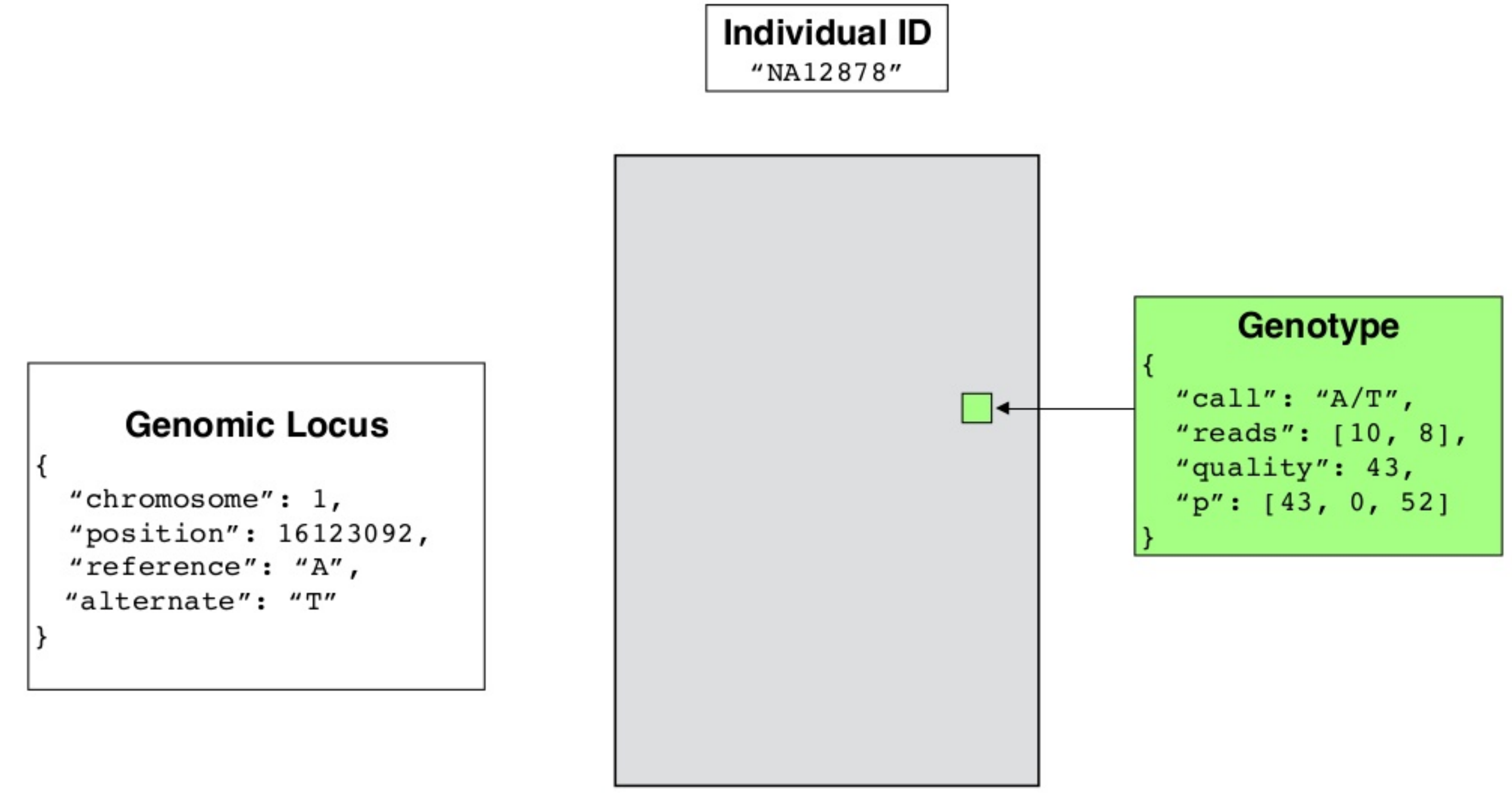
Individual ID

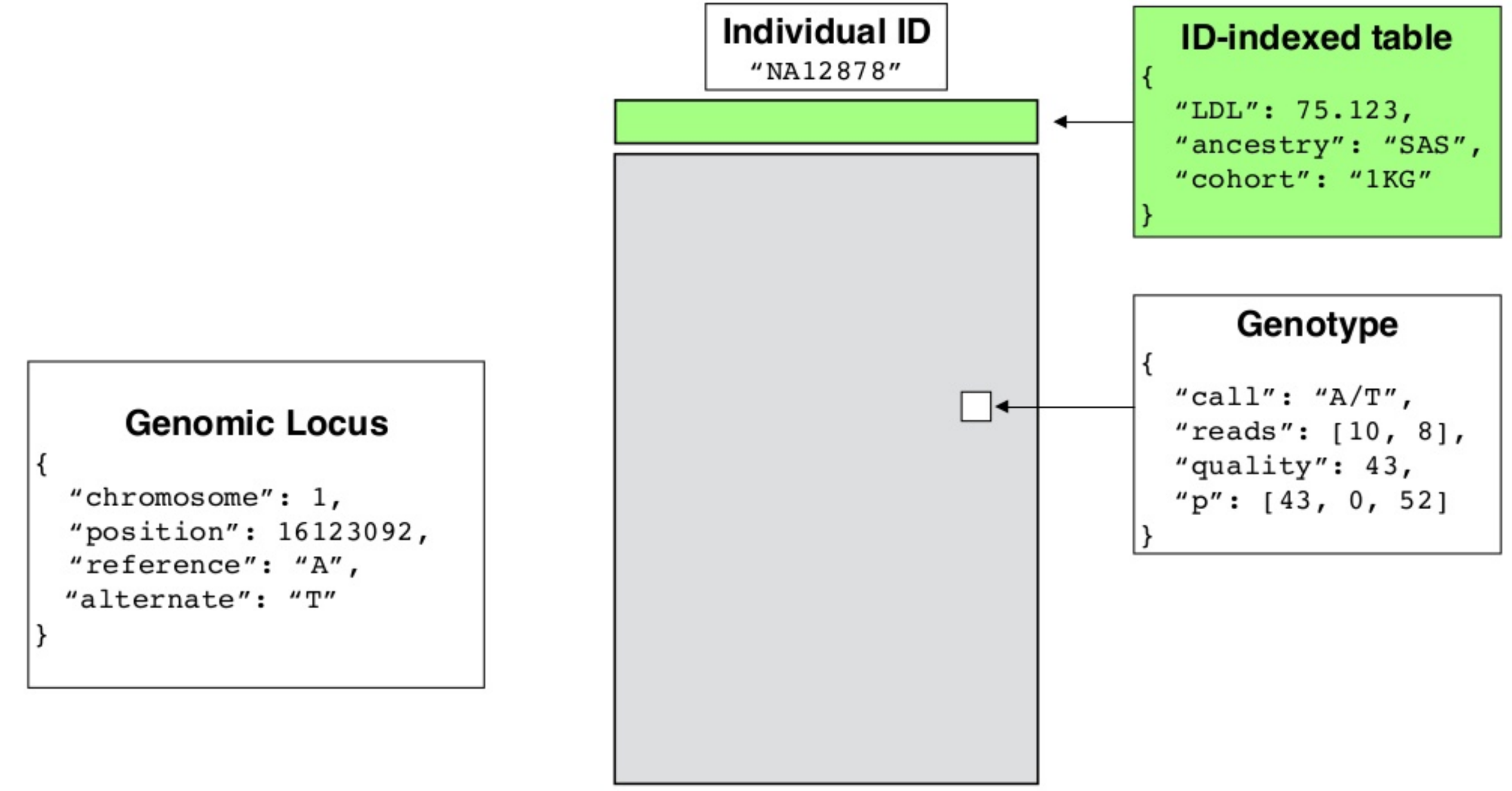
"NA12878"

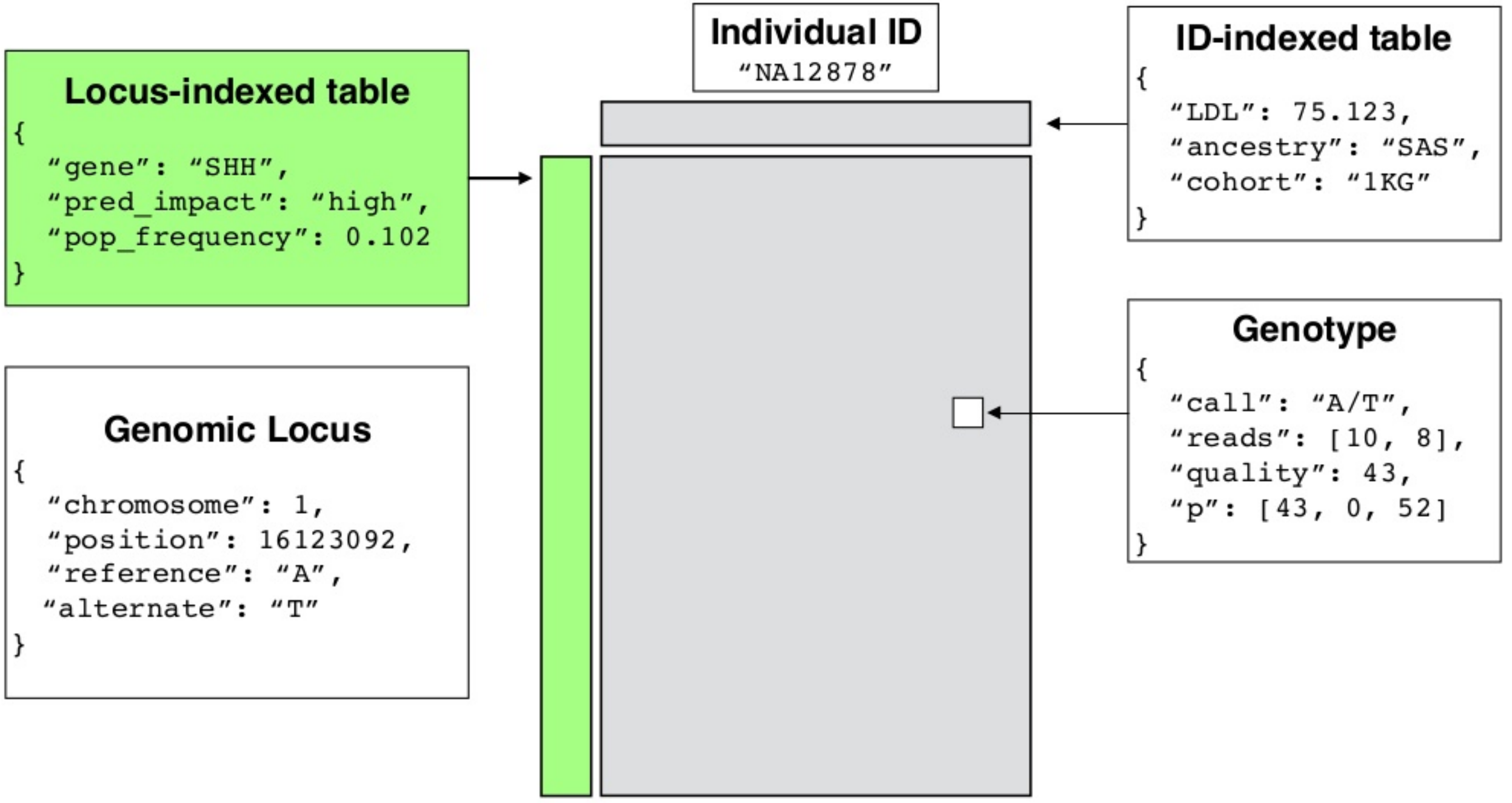


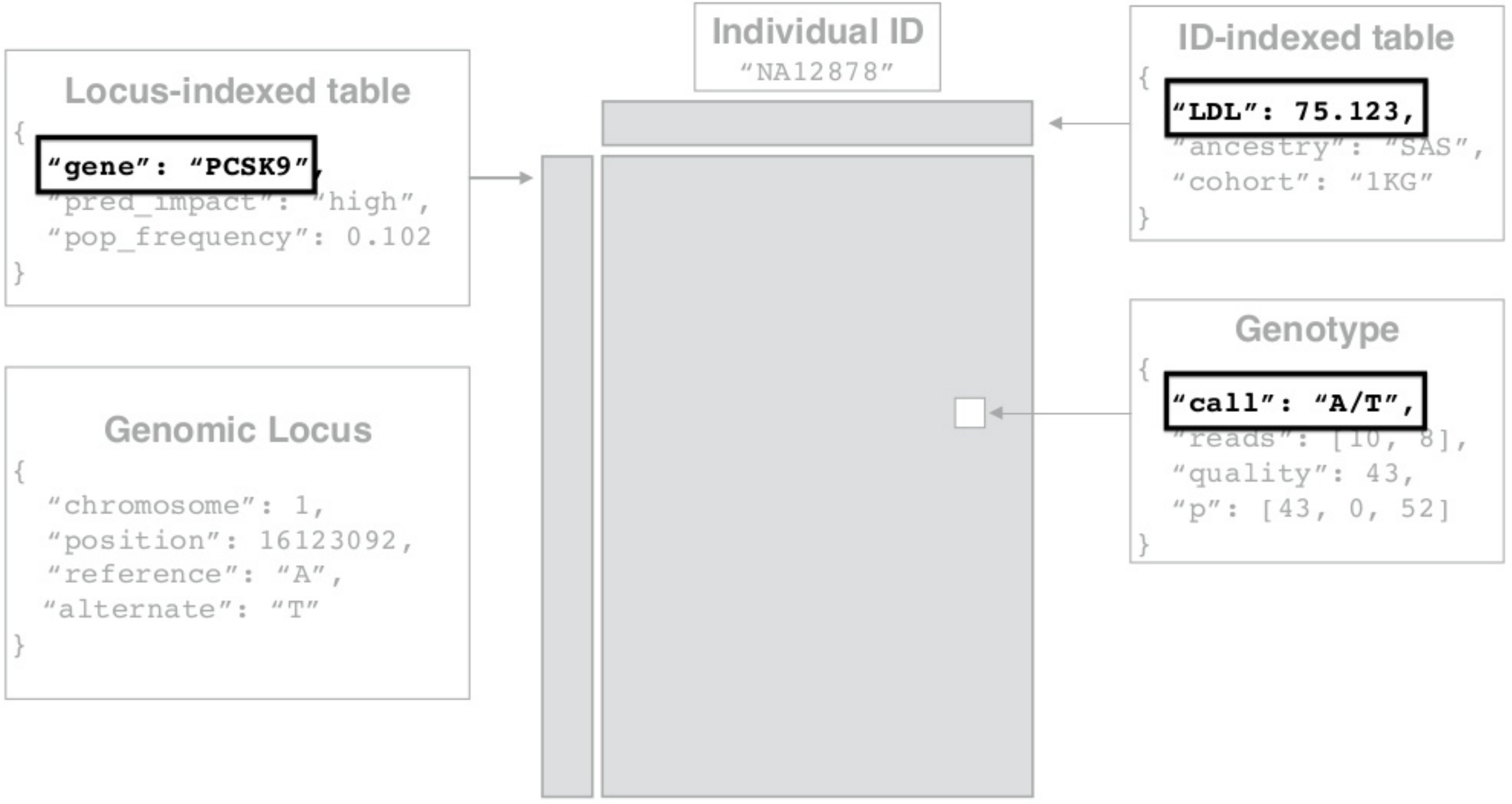
Genomic Locus

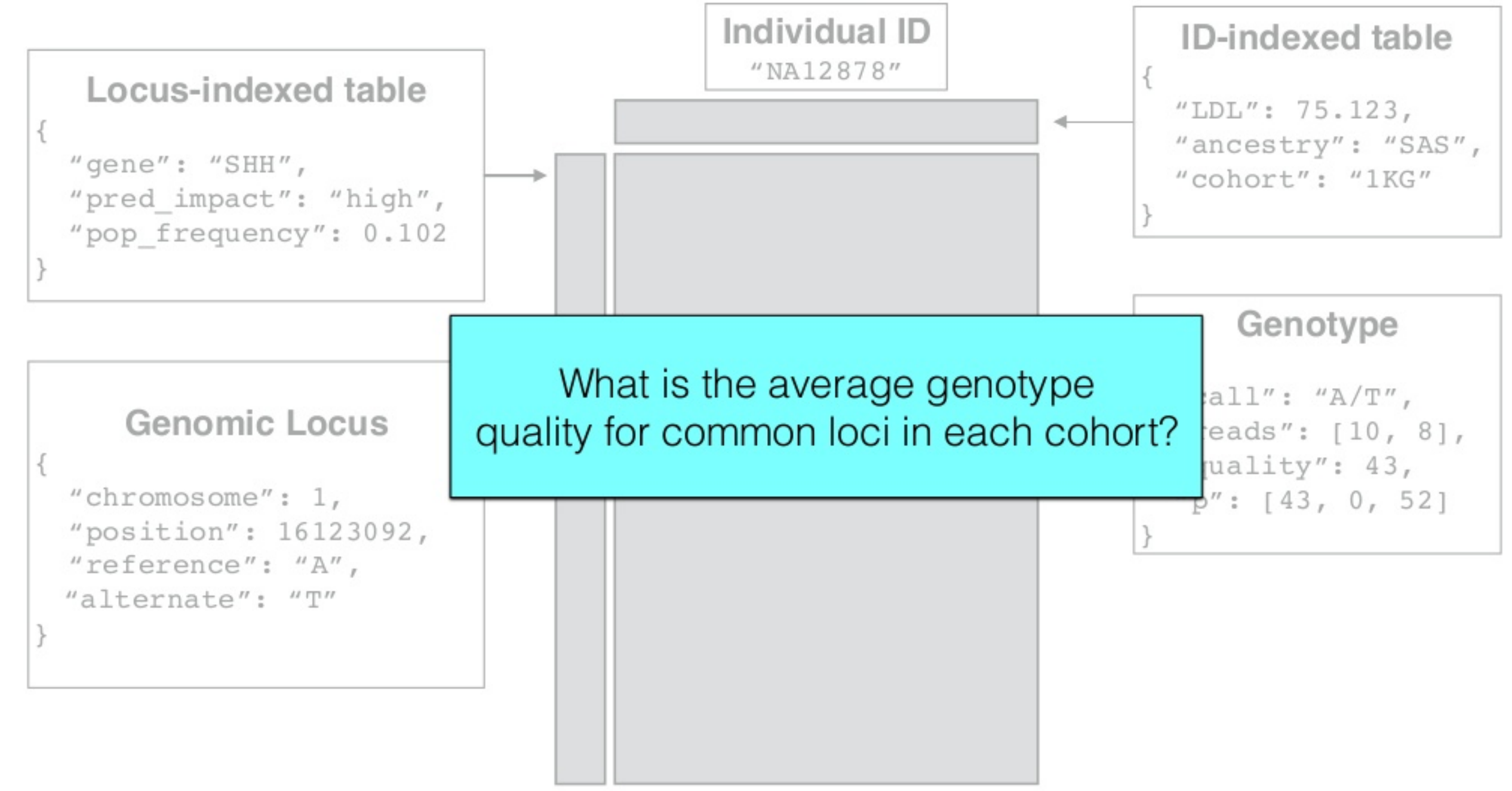
```
{
  "chromosome": 1,
  "position": 16123092,
  "reference": "A",
  "alternate": "T"
}
```

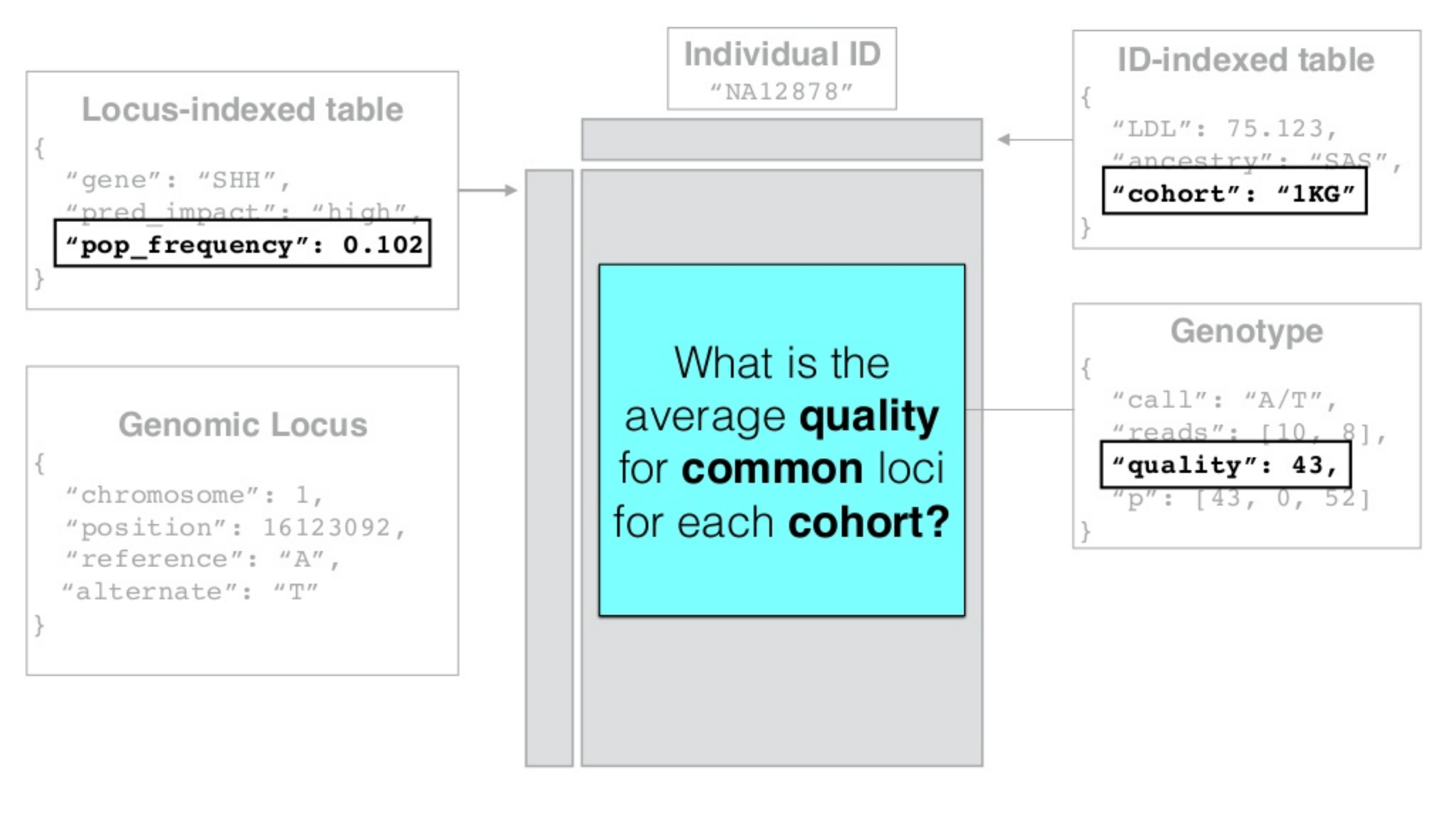

Individual ID
"NA12878"
Genomic Locus
{
"chromosome": 1,
"position": 16123092,
"reference": "A",
"alternate": "T"
}
Genotype
{
"call": "A/T",
"reads": [10, 8],
"quality": 43,
"p": [43, 0, 52]
}


Individual ID
"NA12878"
ID-indexed table
{
"LDL": 75.123,
"ancestry": "SAS",
"cohort": "1KG"
}
Genotype
{
"call": "A/T",
"reads": [10, 8],
"quality": 43,
"p": [43, 0, 52]
}
Genomic Locus
{
"chromosome": 1,
"position": 16123092,
"reference": "A",
"alternate": "T"
}

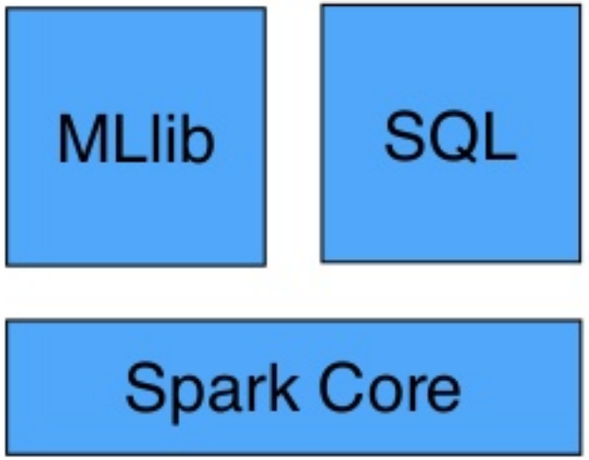

Individual ID
"NA12878"
Locus-indexed table
{
"gene": "SHH",
"pred_impact": "high",
"pop_frequency": 0.102
}
ID-indexed table
{
"LDL": 75.123,
"ancestry": "SAS",
"cohort": "1KG"
}
Genomic Locus
{
"chromosome": 1,
"position": 16123092,
"reference": "A",
"alternate": "T"
}
Genotype
{
"call": "A/T",
"reads": [10, 8],
"quality": 43,
"p": [43, 0, 52]
}

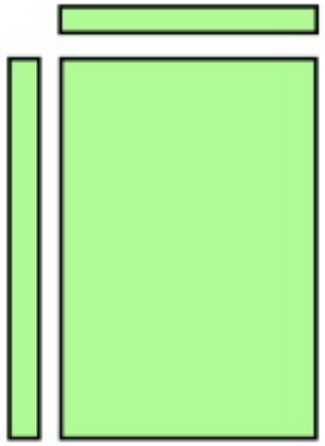

Individual ID
"NA12878"
Locus-indexed table
{
"gene": "PCSK9",
"pred_impact": "high",
"pop_frequency": 0.102
}
ID-indexed table
{
"LDL": 75.123,
"ancestry": "SAS",
"cohort": "1KG"
}
Genomic Locus
{
"chromosome": 1,
"position": 16123092,
"reference": "A",
"alternate": "T"
}
Genotype
{
"call": "A/T",
"reads": [10, 8],
"quality": 43,
"p": [43, 0, 52]
}


Individual ID
"NA12878"
Locus-indexed table
{
"gene": "SHH",
"pred_impact": "high",
"pop_frequency": 0.102
}
ID-indexed table
{
"LDL": 75.123,
"ancestry": "SAS",
"cohort": "1KG"
}
Genomic Locus
{
"chromosome": 1,
"position": 16123092,
"reference": "A",
"alternate": "T"
}
Genotype
all": "A/T",
eads": [10, 8],
uality": 43,
": [43, 0, 52]
}
What is the average genotype quality for common loci in each cohort?


Individual ID
"NA12878"
Locus-indexed table
{
"gene": "SHH",
"pred_impact": "high",
"pop_frequency": 0.102
}
ID-indexed table
{
"LDL": 75.123,
"ancestry": "SAS",
"cohort": "1KG"
}
Genomic Locus
{
"chromosome": 1,
"position": 16123092,
"reference": "A",
"alternate": "T"
}
What is the average quality for common loci for each cohort?
Genotype
{
"call": "A/T",
"reads": [10, 8],
"quality": 43,
"p": [43, 0, 52]
}

- scalability
- high-level programming APIs
- linear algebra, MLlib
- Scala, Python, R

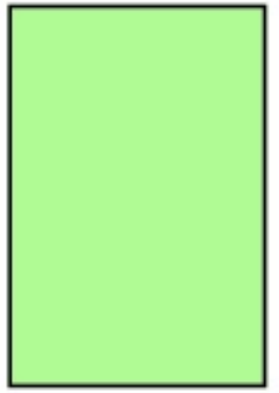




=

+

+



DataFrame 1


```
{
  “ID”: “NA12878”,
  “LDL”: 75.123,
  “ancestry”: “SAS”,
  “cohort”: “1KG”
}
```


DataFrame 2


```
{
  “chromosome”: 1,
  “position”: 16123092,
  “reference_allele”: “A”,
  “alternate_allele”: “T”,
  “gene”: “SHH”,
  “pred_impact”: “high”,
  “pop_frequency”: 0.102
}
```


DataFrame 3


```
{
  “ID”: “NA12878”,
  “chromosome”: 1,
  “position”: 16123092,
  “reference_allele”: “A”,
  “alternate_allele”: “T”,
  “call”: “A/T”,
  “reads”: [10, 8],
  “quality”: 43,
  “p”: [43, 0, 52]
}
```

What is the
average genotype quality
for common loci
in each cohort?

```
df3
JOIN df1 ON ID
JOIN df2 ON {chr, pos, ref, alt}
WHERE pop_frequency > 0.10
GROUP BY cohort
AGGREGATE mean(quality)
```

DataFrame 1

```
{
  "ID": "NA12878",
  "LDL": 75.123,
  "ancestry": "SAS",
  "cohort": "1KG"
}
```

DataFrame 2

```
{
  "chromosome": 1,
  "position": 16123092,
  "reference_allele": "A",
  "alternate_allele": "T",
  "gene": "SHH",
  "pred_impact": "high",
  "pop_frequency": 0.102
}
```

DataFrame 3

```
{
  "ID": "NA12878",
  "chromosome": 1,
  "position": 16123092,
  "reference_allele": "A",
  "alternate_allele": "T",
  "call": "A/T",
  "reads": [10, 8],
  "quality": 43,
  "p": [43, 0, 52]
}
```

What is the
average genotype quality
for common
in each cohort?

```
df3
JOIN df1 ON ID
JOIN df2 ON {chr, pos, ref, alt}
WHERE po      quency > 0.10
GROUP       hort
AGGR        mean(quality)
```

DataFrame 1

```
{
  "ID": "NA12878",
  "LDL": 75.123,
  "ancestry": "SAS",
  "cohort": "1KG"
}
```

DataFrame 2

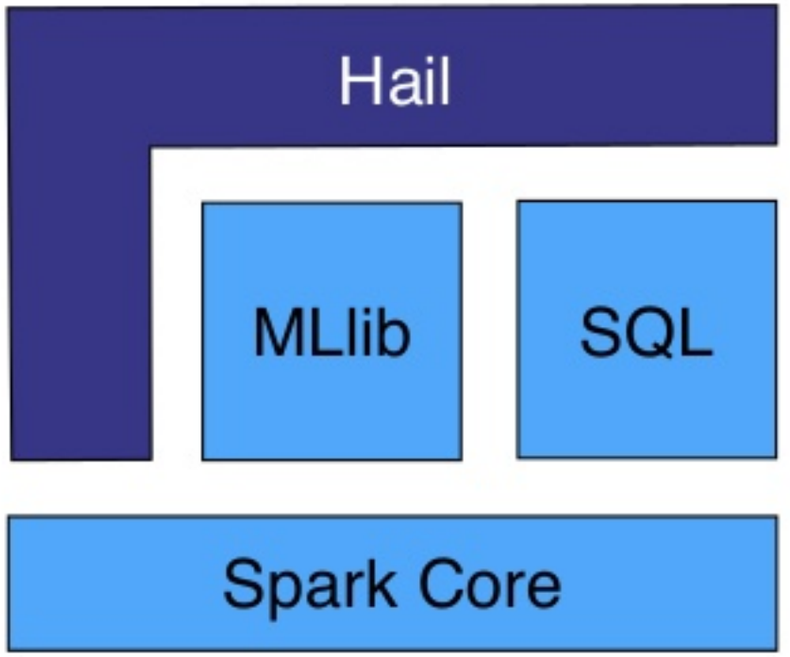
```
{
  "ch
  "          092,
  "ence_a          "A",
  "ernate_all          T",
  "gene": "SHH",
  "pred_impact": "hig
  "pop_frequency": 0.1
}
```

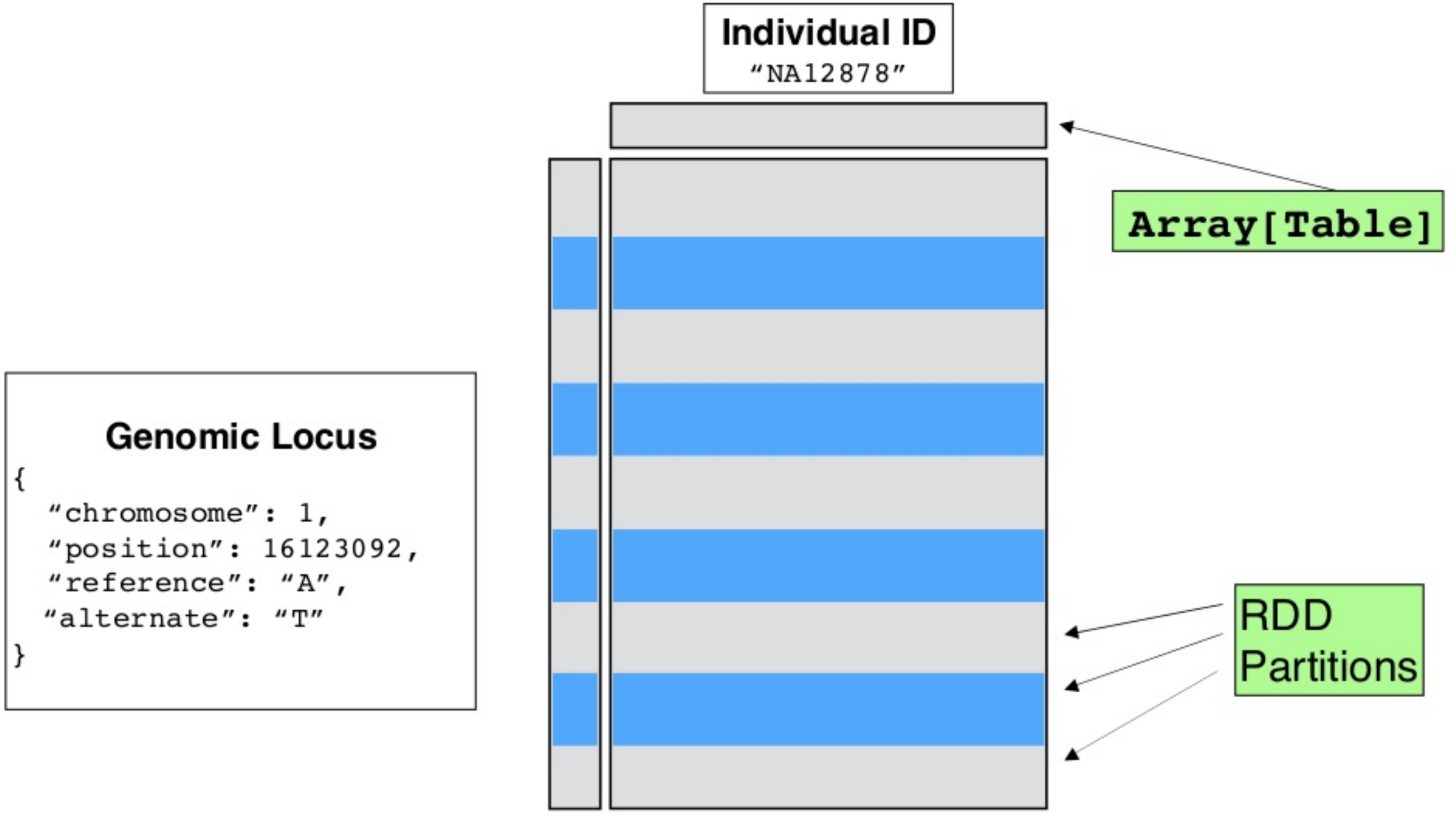
DataFrame 3

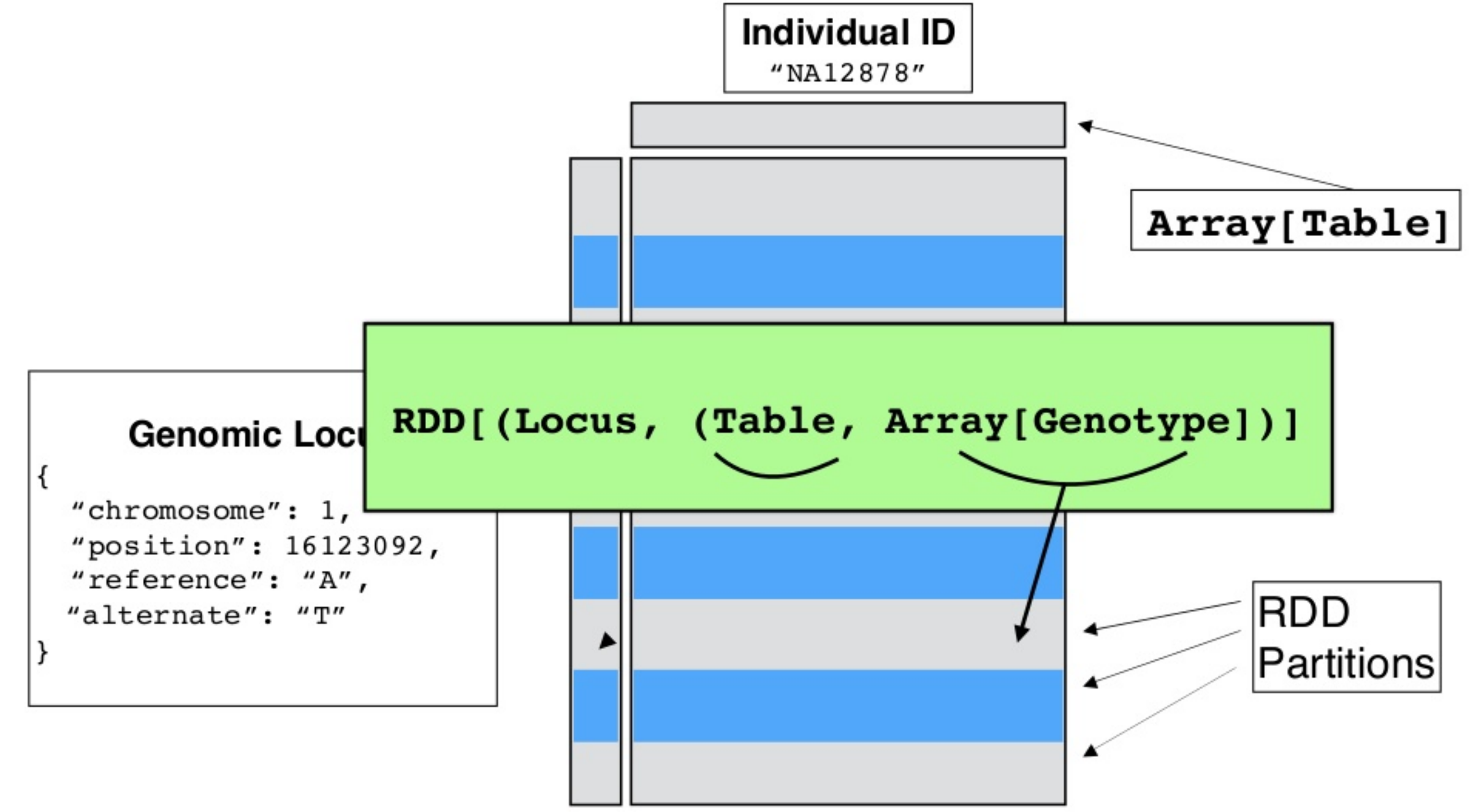
```
{
  "ID": "NA12878",
  "chromosome": 1,
  "position": 16123092,
  "reference_allele": "A",
  "alternate_allele": "T",
  "call": "A/T",
  "reads": [10, 8],
  "quality": 43,
  "p": [43, 0, 52]
}
```

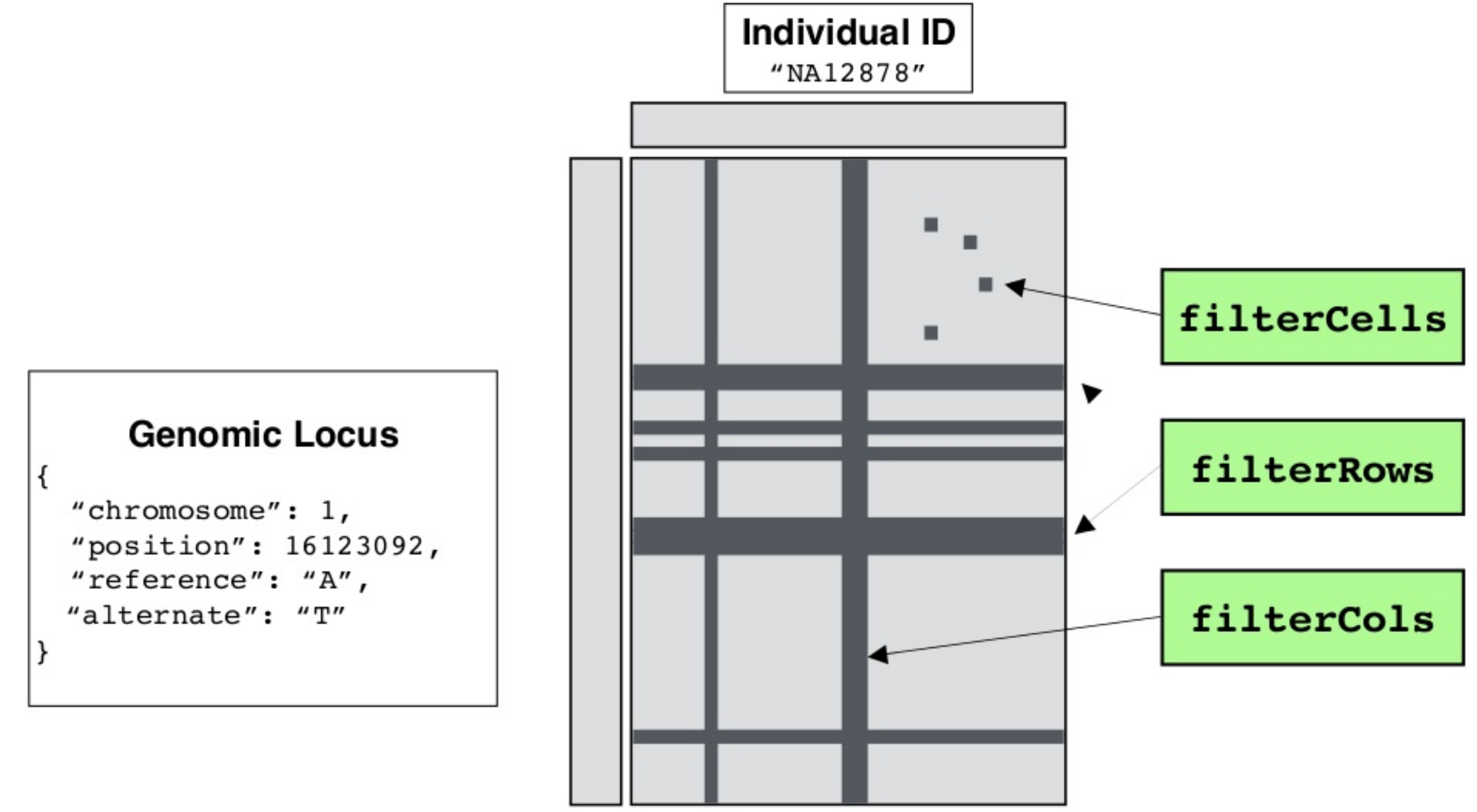
- genomic data ETL
- high-level APIs for multi-dimensional data query
- stats and ML methods
- Scala, Python

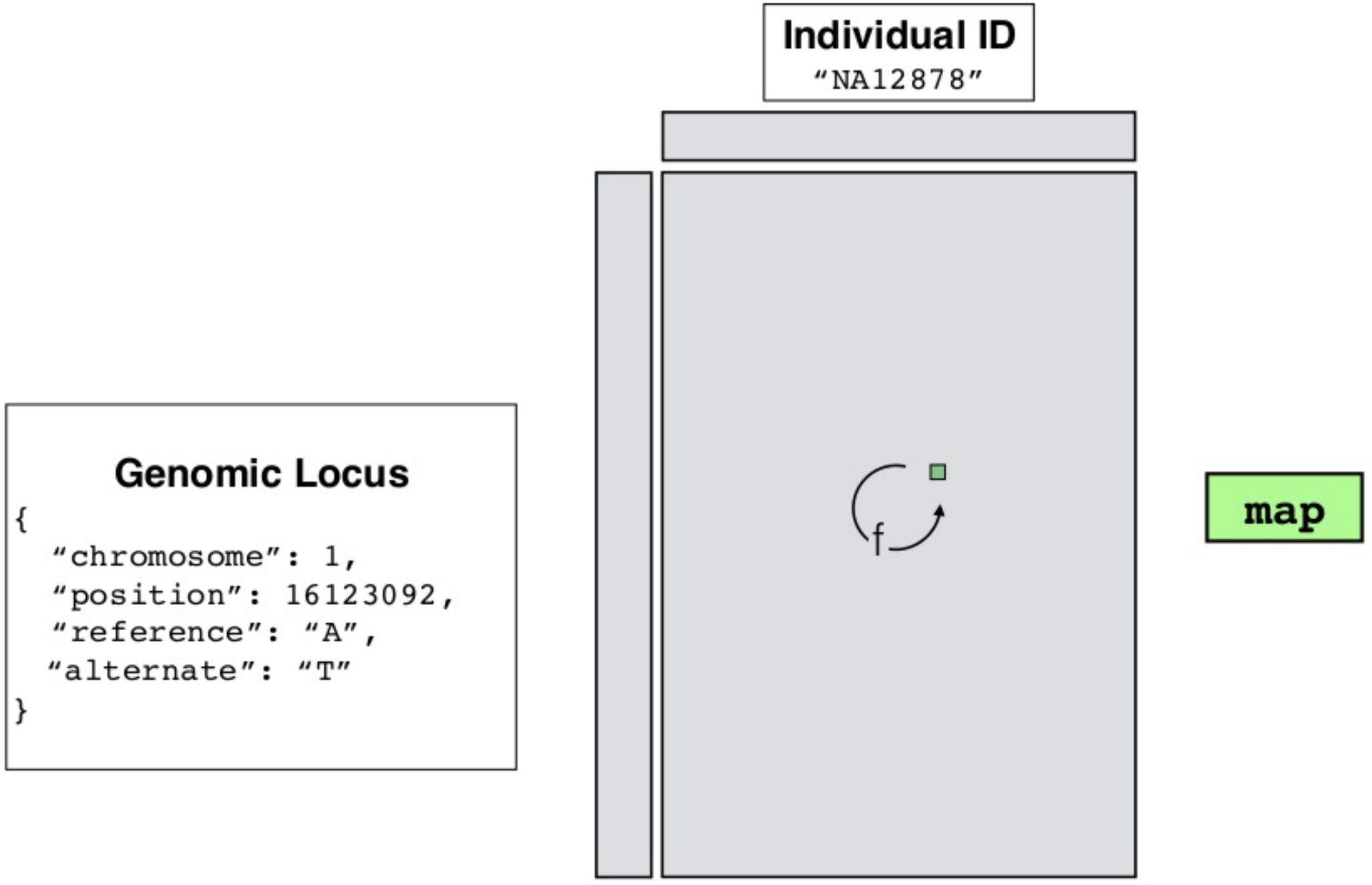
- scalability
- high-level programming APIs
- linear algebra, MLlib
- Scala, Python, R

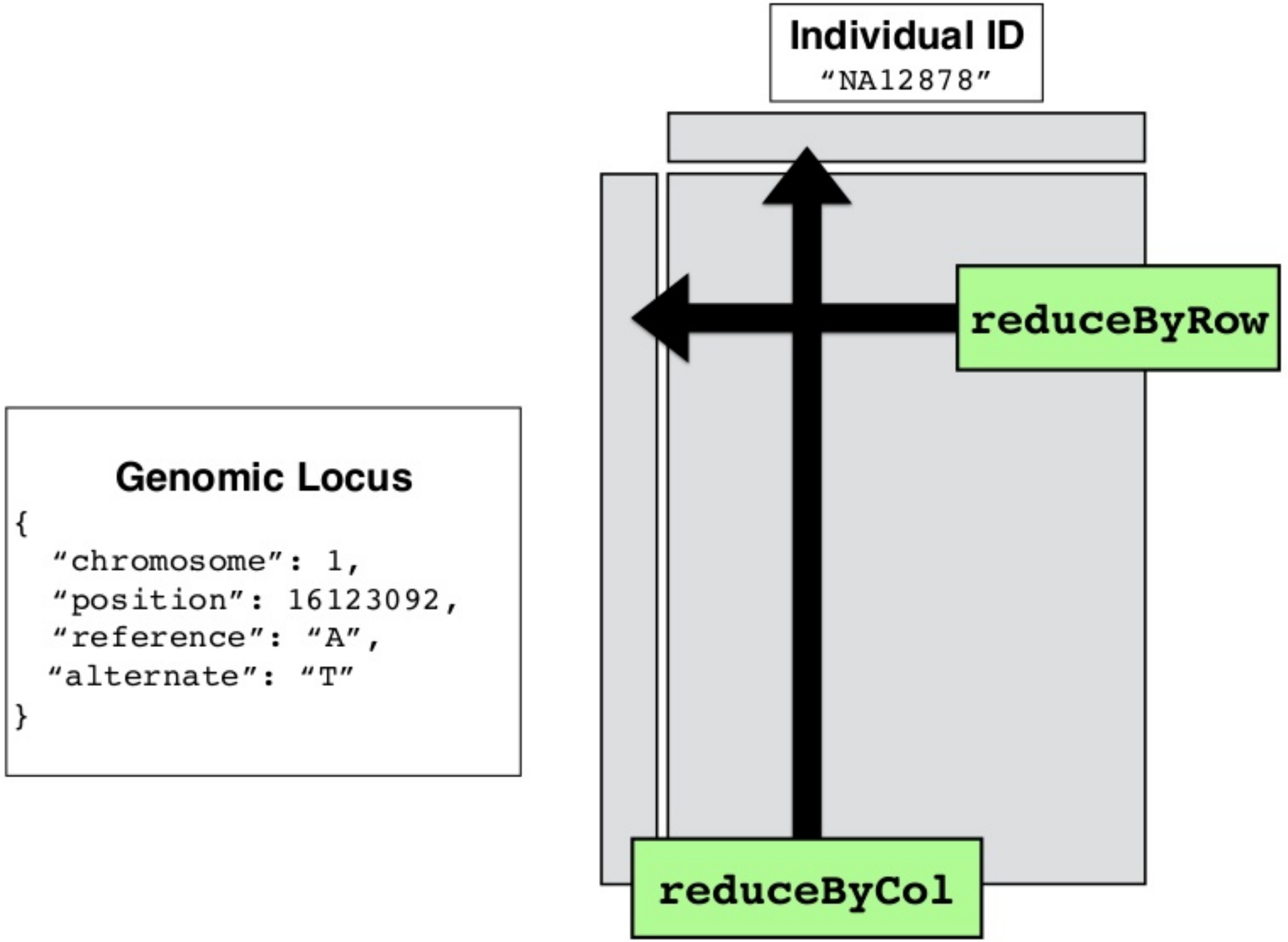


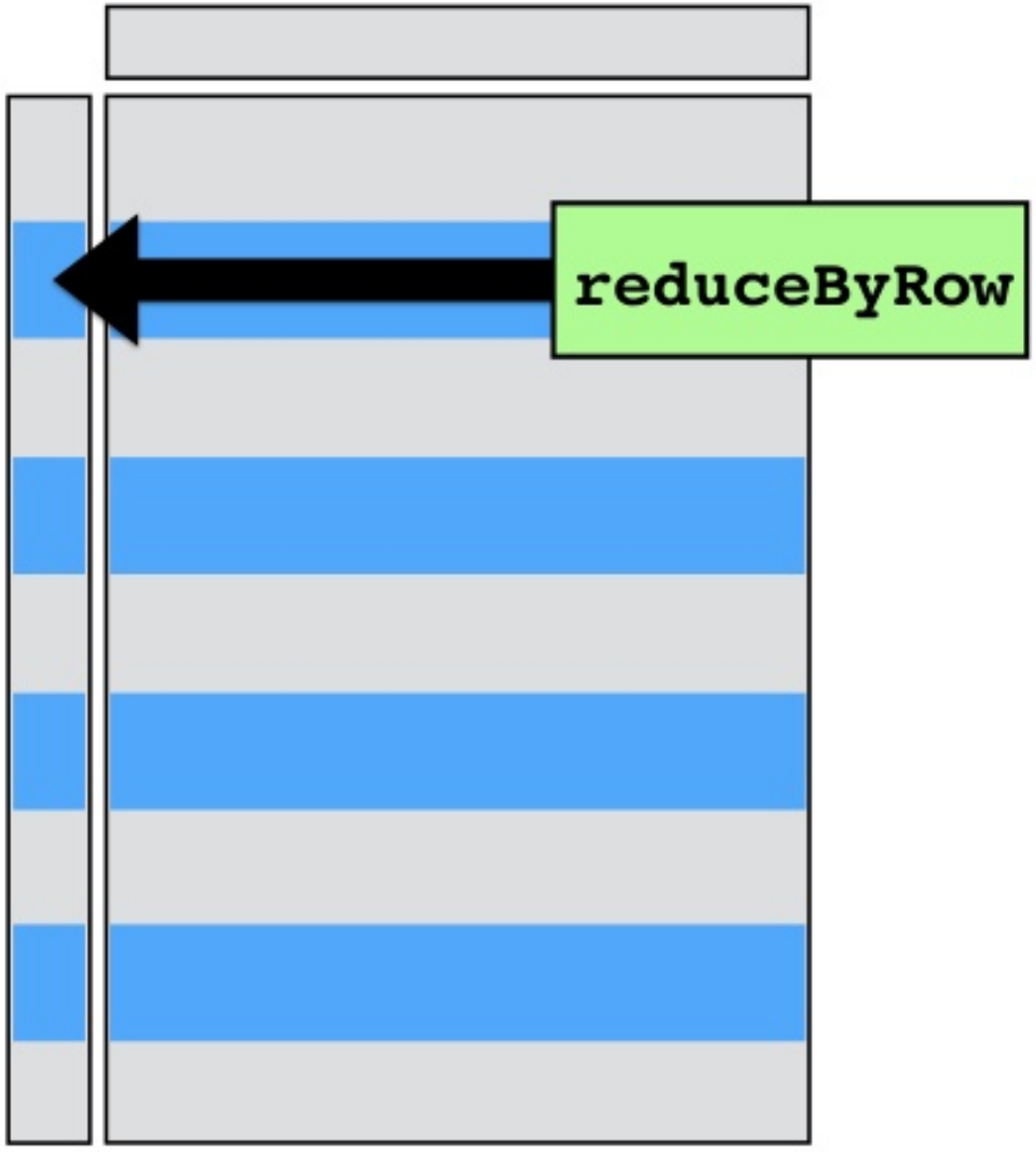

Individual ID
“NA12878”
Array[Table]
Genomic Locus
{
“chromosome”: 1,
“position”: 16123092,
“reference”: “A”,
“alternate”: “T”
}
RDD
Partitions

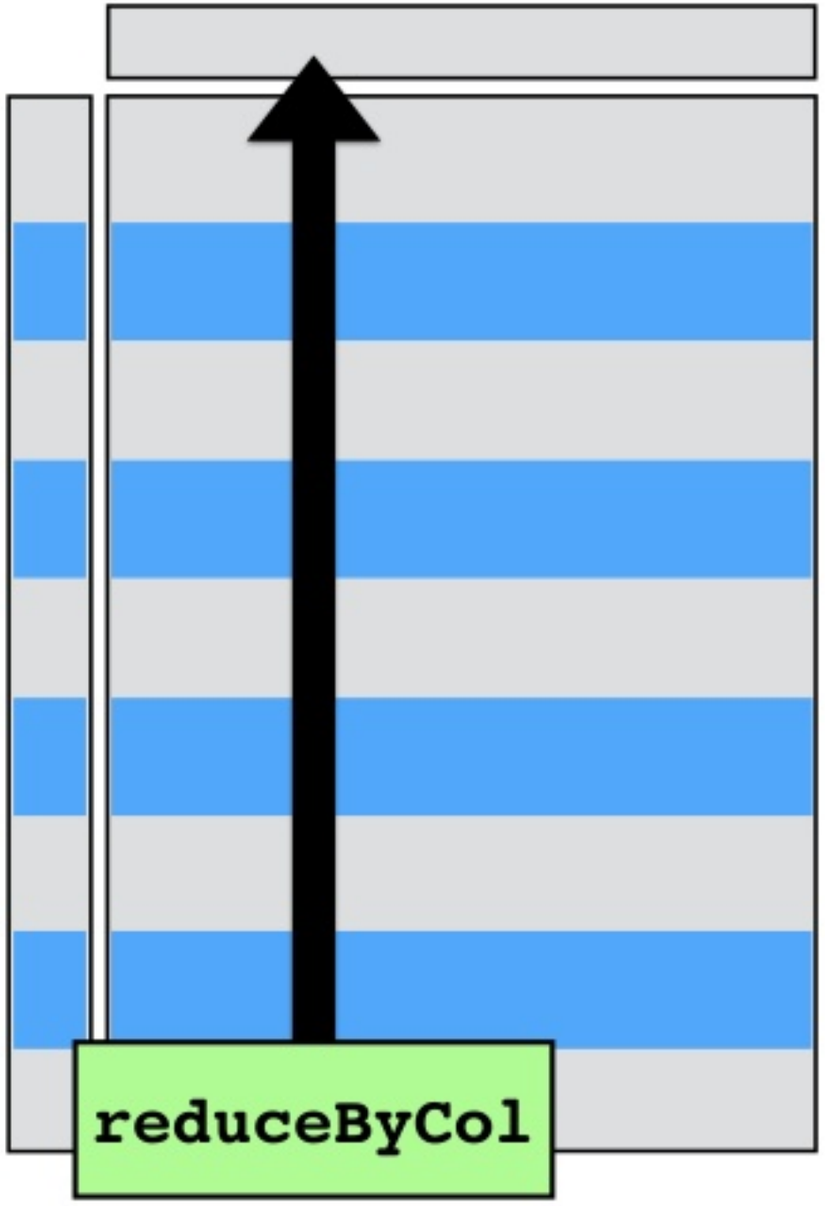

Individual ID
"NA12878"
Array[Table]
RDD[(Locus, (Table, Array[Genotype]))]
Genomic Locu
{
"chromosome": 1,
"position": 16123092,
"reference": "A",
"alternate": "T"
}
RDD
Partitions

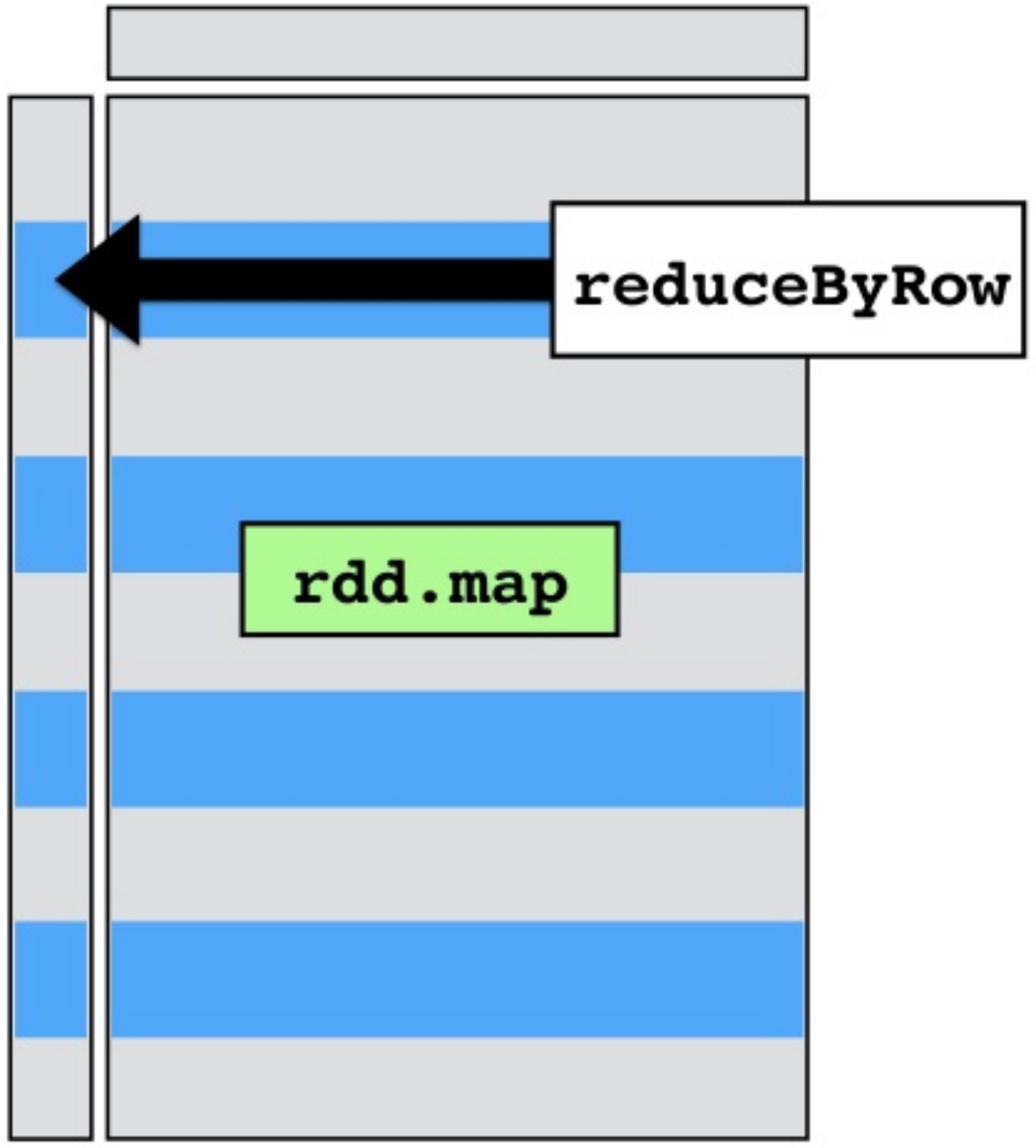

Individual ID
“NA12878”
filterCells
filterRows
filterCols
Genomic Locus
{
“chromosome”: 1,
“position”: 16123092,
“reference”: “A”,
“alternate”: “T”
}

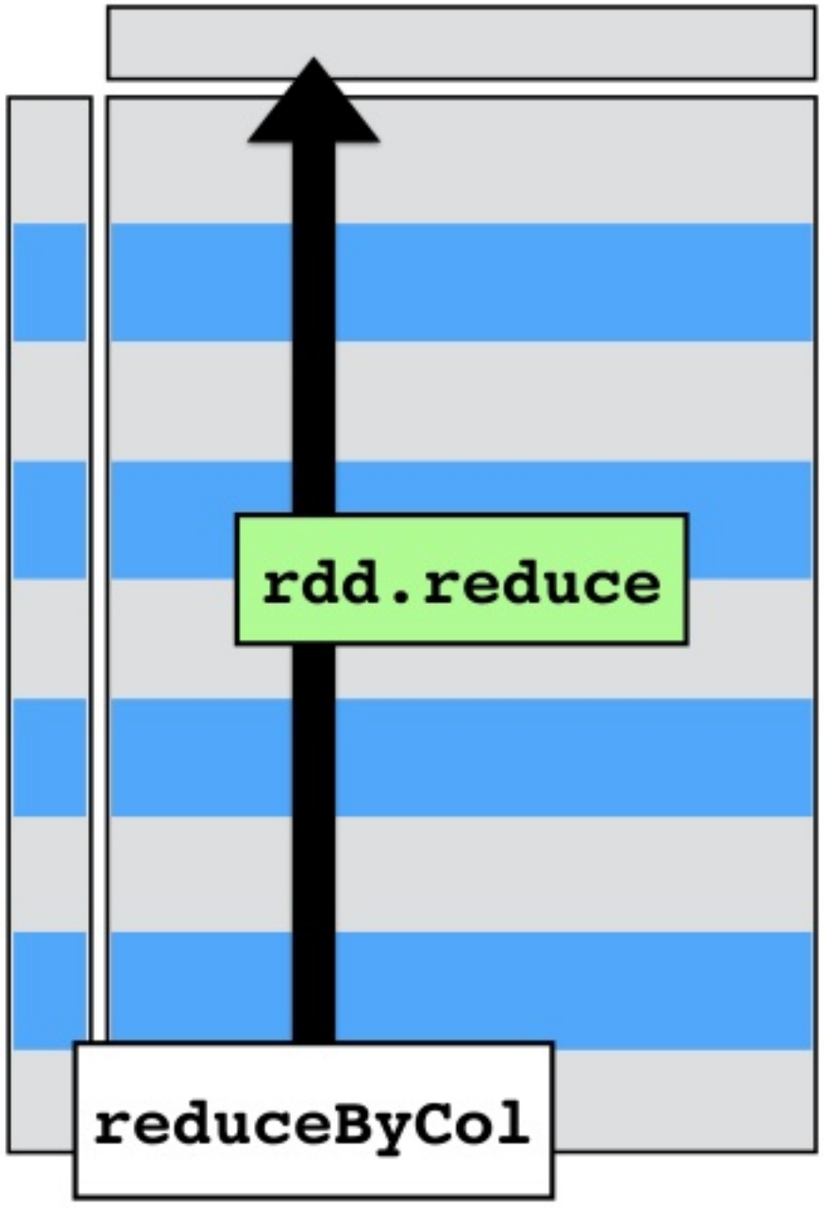

Individual ID
“NA12878”
Genomic Locus
{
“chromosome”: 1,
“position”: 16123092,
“reference”: “A”,
“alternate”: “T”
}
f
map


Individual ID
"NA12878"
reduceByRow
reduceByCol
Genomic Locus
{
"chromosome": 1,
"position": 16123092,
"reference": "A",
"alternate": "T"
}


reduceByRow


reduceByCol

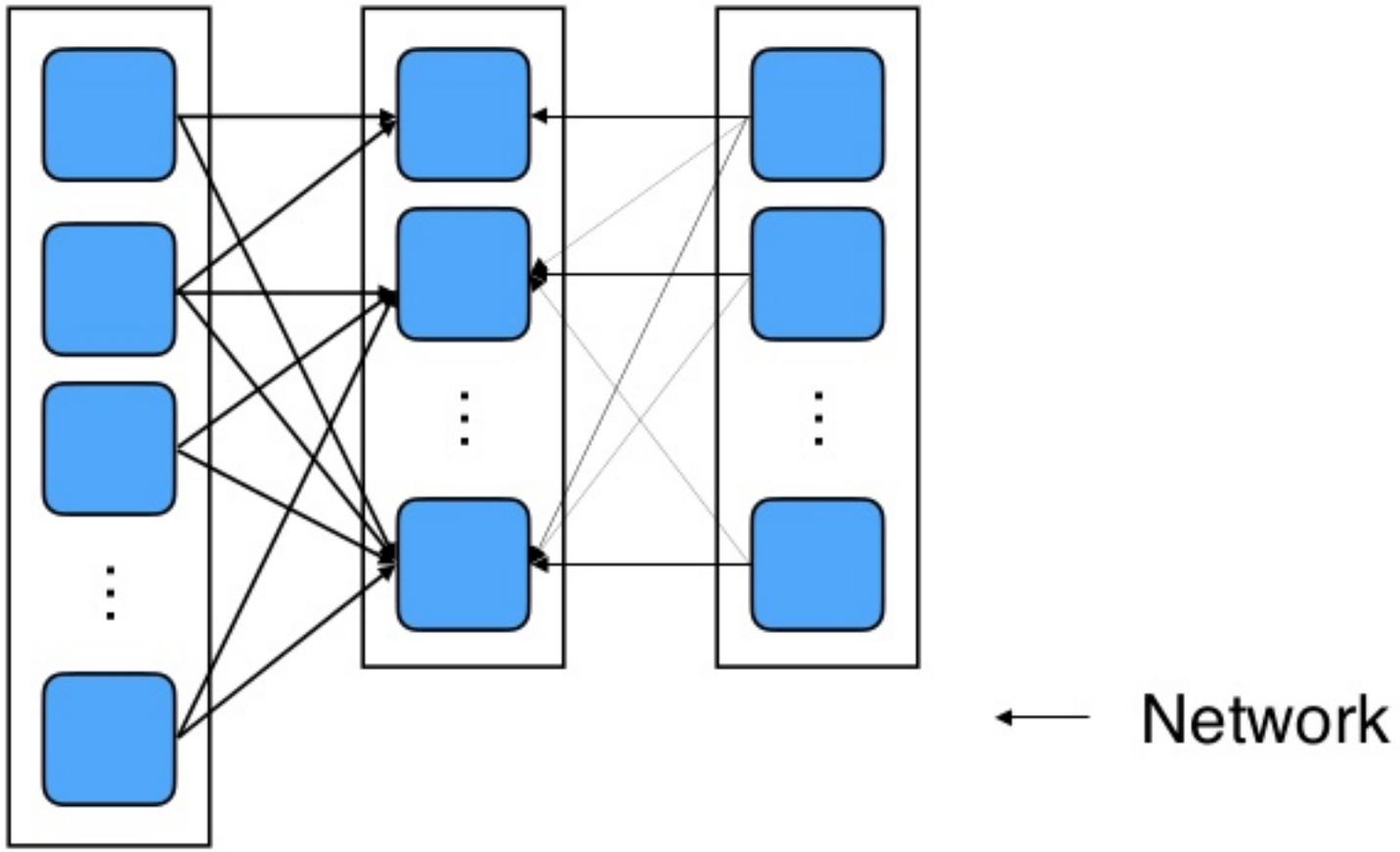

reduceByRow
rdd.map


rdd.reduce
reduceByCol

`join`: the troublemaker

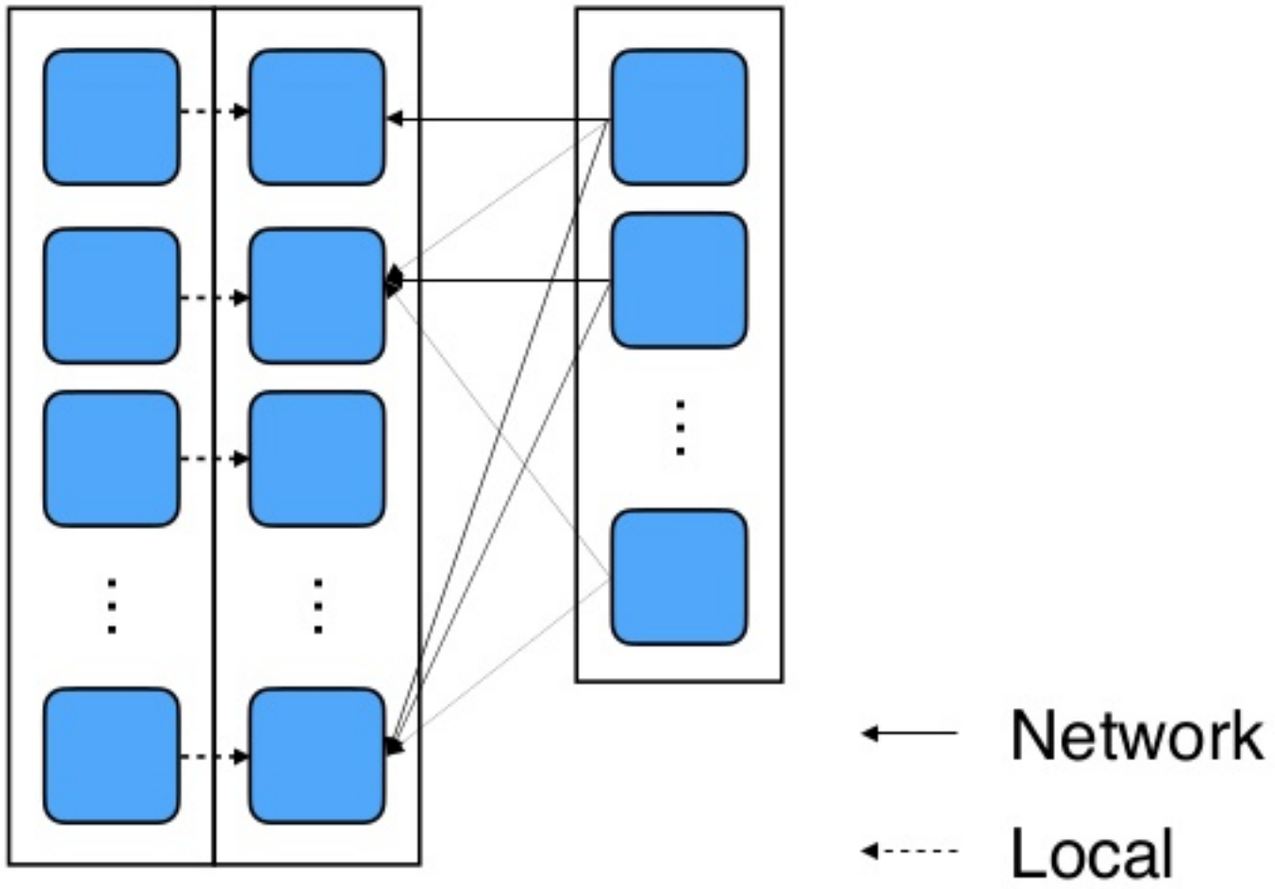
- Need to join by genomic locus **all the time**
  - Machine learning models for effect of mutation
  - Clinical annotation databases
  - Combine information across genetic datasets

Typical `join`



← Network

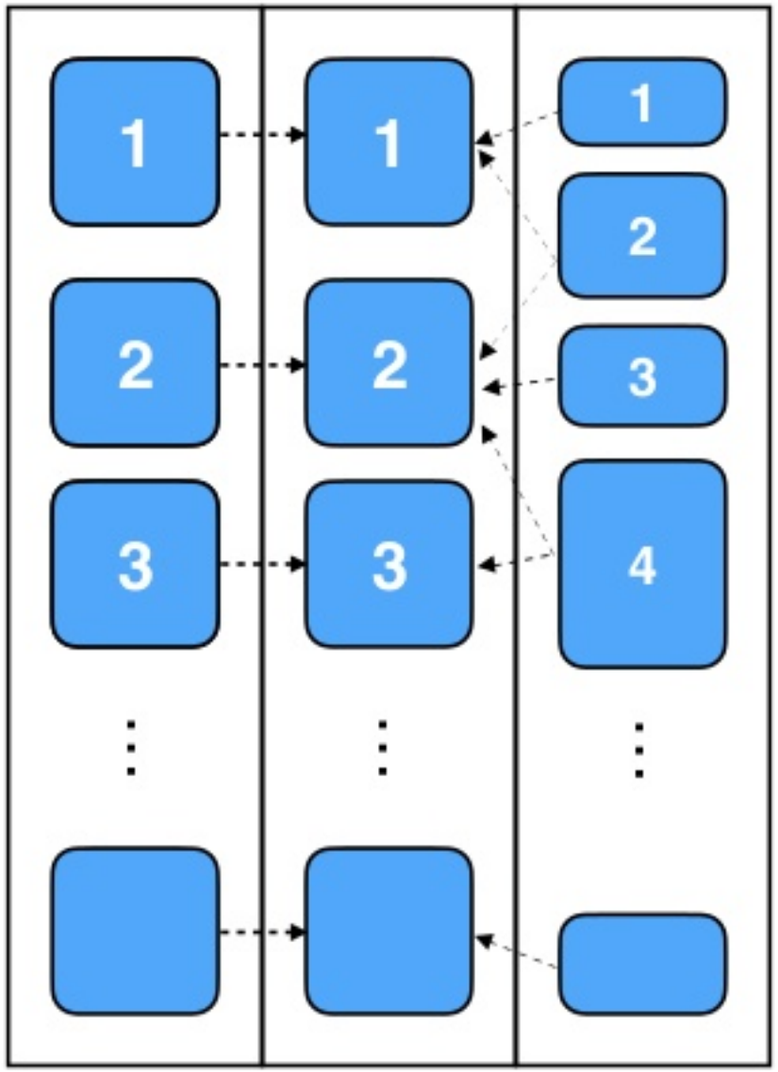
Partitioned `join`



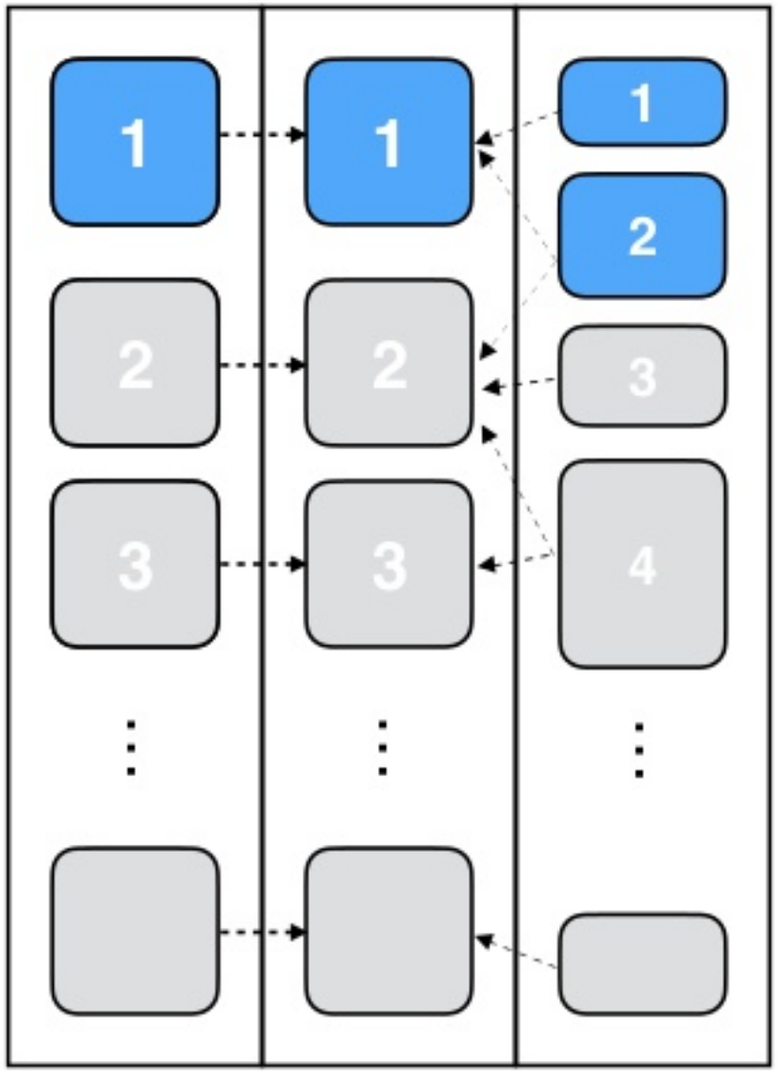
`OrderedRDD`

- Generalizes Spark **`RangePartitioner`**
- Shuffle-free joins
- Preserves partitioner on disk

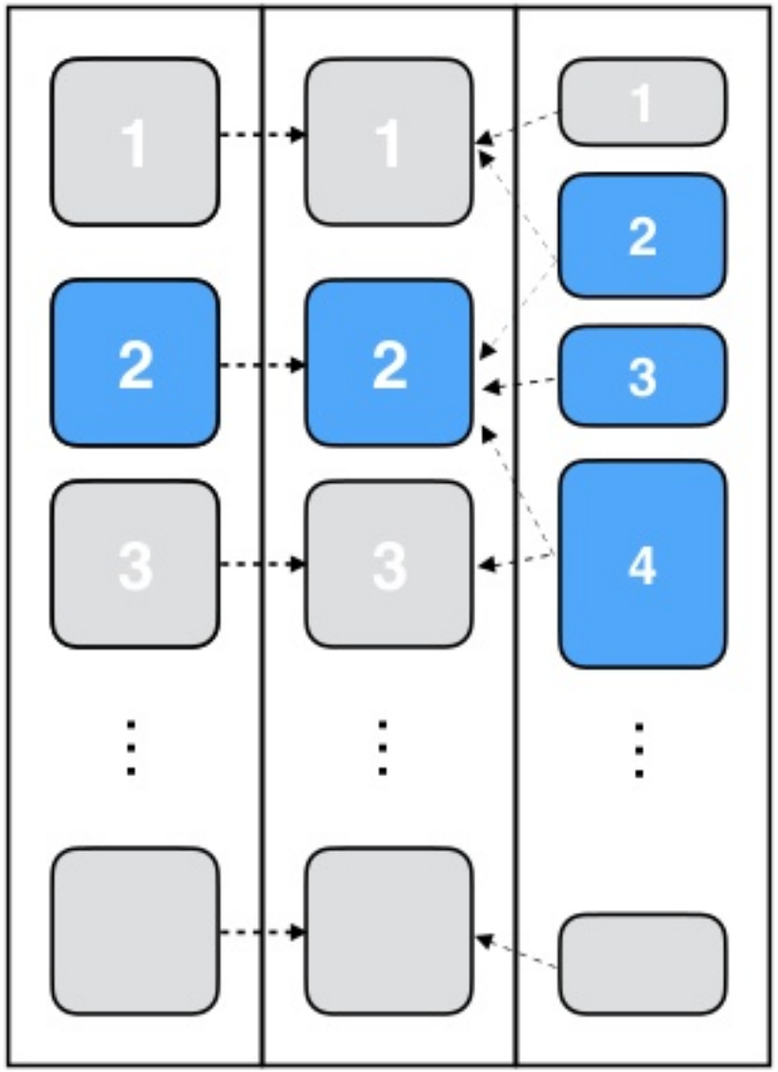
`OrderedRDD` range `join`



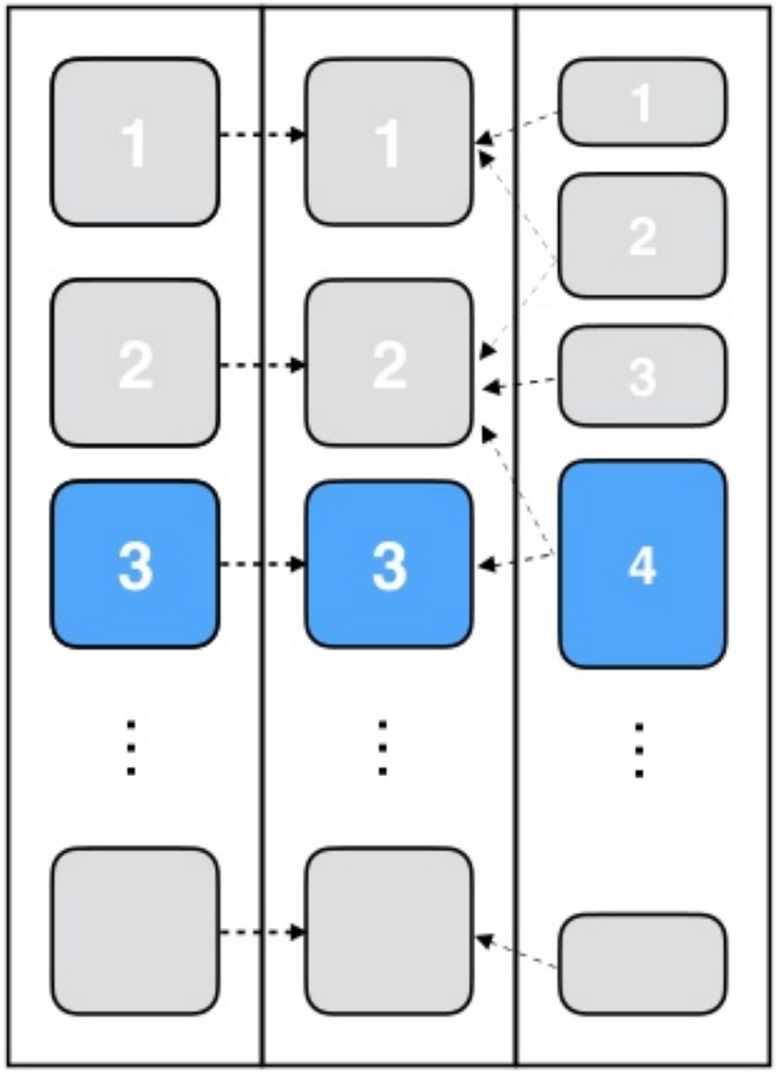
OrderedRDD range **join**



OrderedRDD range **join**

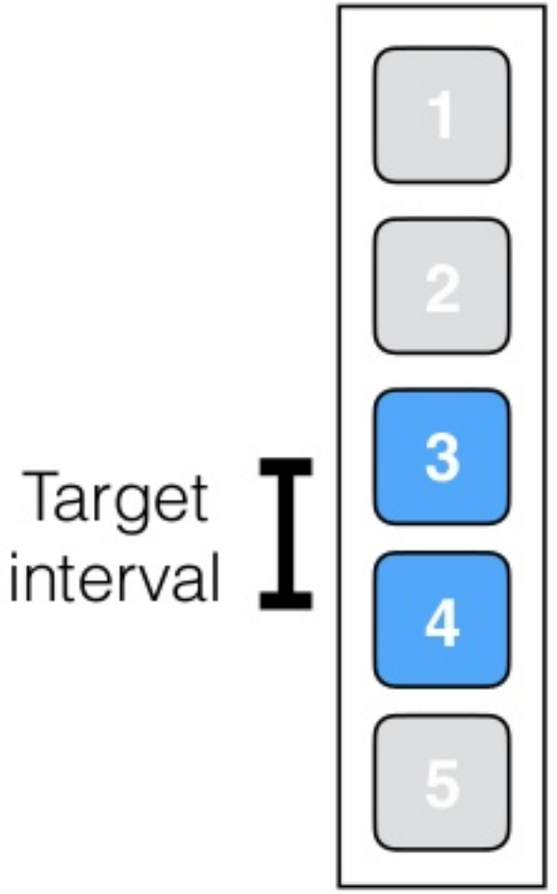


OrderedRDD range **join**



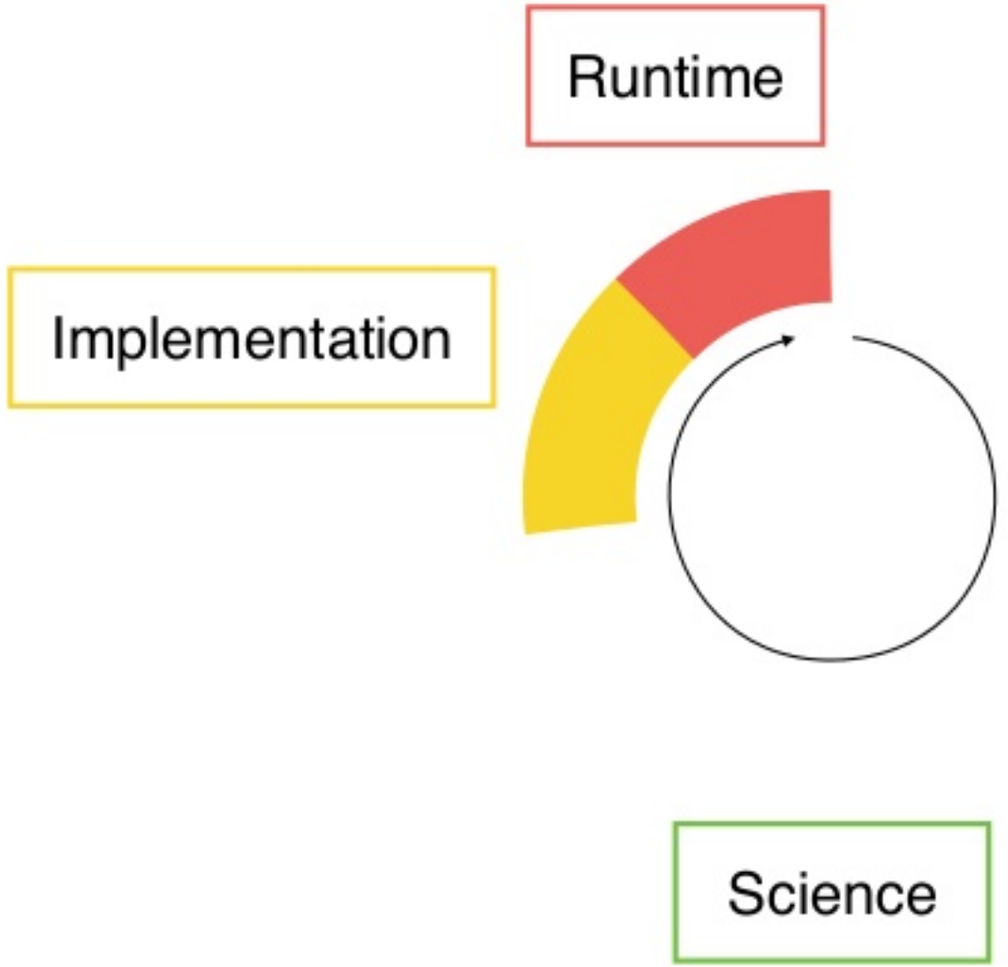
`OrderedRDD` predicate pushdown

- Interval filter is O(partitions kept)
- 20G dataset: 100ms on a laptop



Where we are today

- Stable interface: 0.1 release
- Transformed iterative QC
- Rebuilt most standard genetics tools as methods in Hail



Hail Science

- L. Francioli, MacArthur lab, Analysis of whole-genome sequencing from 15,139 individuals
- A. Ganna et al., Ultra-rare disruptive and damaging mutations influence educational attainment in the general population, Nature Neuroscience
- A. Ganna et al., The impact of ultra-rare variants on human diseases and traits
- A. Ganna et al., The impact of rare variants on schizophrenia: whole genome sequencing of 10,000 individuals from the WGSPD consortia
- M. Kurki, Palotie Lab, Alzheimer's Disease Rare Variant Association Study in Finnish Founder Population
- M. Kurki, Palotie Lab, Genetic Architecture of Idiopathic Intellectual Disability in a Northern Finnish founder population cohort
- M. Kurki, P. Gormley, Palotie Lab, Genetic Architecture of Familial Migraine in a Family collection of 9000 Individuals in 2000 Families
- K. Karczewski, MacArthur Lab, The Human Knockout Project: analyzing loss-of-function variants across 126,216 individuals

Hail Science

- X. Li et al., Developing and optimizing a whole genome and whole exome sequencing quality control pipeline with 652 Genotype-Tissue Expression donors
- M. A. Rivas et al., Insights into the genetic epidemiology of Crohn's and rare diseases in the Ashkenazi Jewish population
- K. Satterstrom, iPSYCH-Broad Consortium, Rare variants conferring risk for autism identified by whole exome sequencing of dried bloodspots
- C. Seed et al., Neale Lab, Hail: An Open-Source Framework for Scalable Genetic Data Analysis
- G. Tiao, Pan-Cancer Analysis of Whole Genomes, Analysis of rare variation in 2,818 whole-genome germline samples from cancer patients
- S. Maryam Zekavat, P. Natarajan, Kathiresan Lab. An analysis of deep, whole-genome sequences and plasma lipids in ~16,000 multi-ethnic samples.
- S. Maryam Zekavat, Kathiresan Lab. An analysis of deep, whole-genome sequences and coronary artery disease in ~7,000 multi-ethnic samples.
- S. Maryam Zekavat, Kathiresan Lab. Analyzing the full spectrum of genomic variation with Lp(a) Cholesterol: Novel insights from deep, whole genome sequence data in 5,192 Europeans and African Americans from Estonia and from the Jackson Heart Study

Hail Engagement

Academia

Aarhus University
Garvan Institute
John Hopkins
Mayo Clinic
University of Michigan
National Cancer Institute
Oxford
Sanger
Stanford
UCSF
Wash U Genome Institute
…

Industry

Databricks
Illumina
Cloudera
- Children's of Atlanta
- Rush University MC
- Dignity Hospital
- Quest Labs
- Labcorp

Digital China Health
Cray
Metistream
…

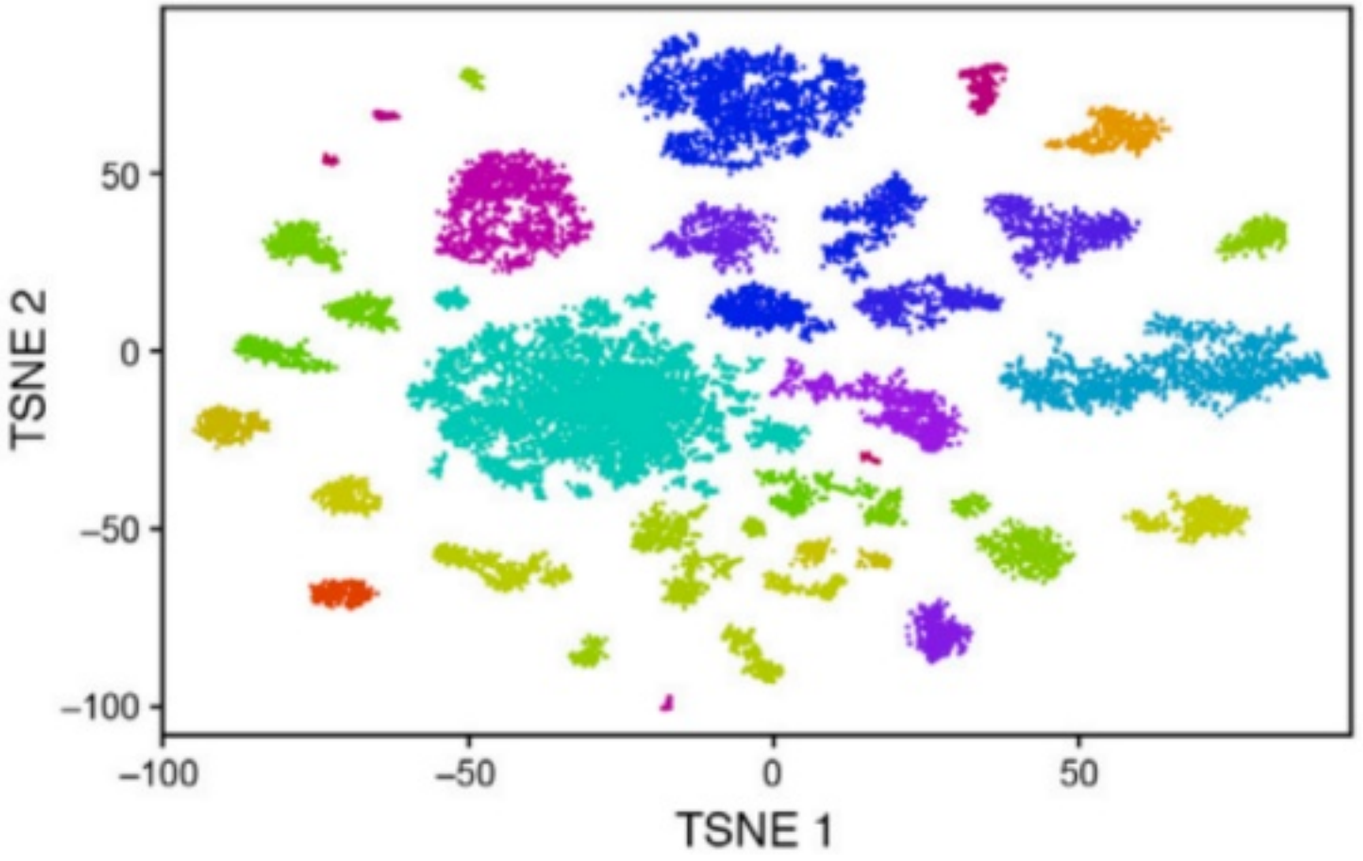
New code

- **10-100x** performance improvements in the next year
  - compile DSL to JVM bytecode
  - query optimizer
  - explore alternatives to Parquet

- Build new models and algorithms for big genetic data

New applications

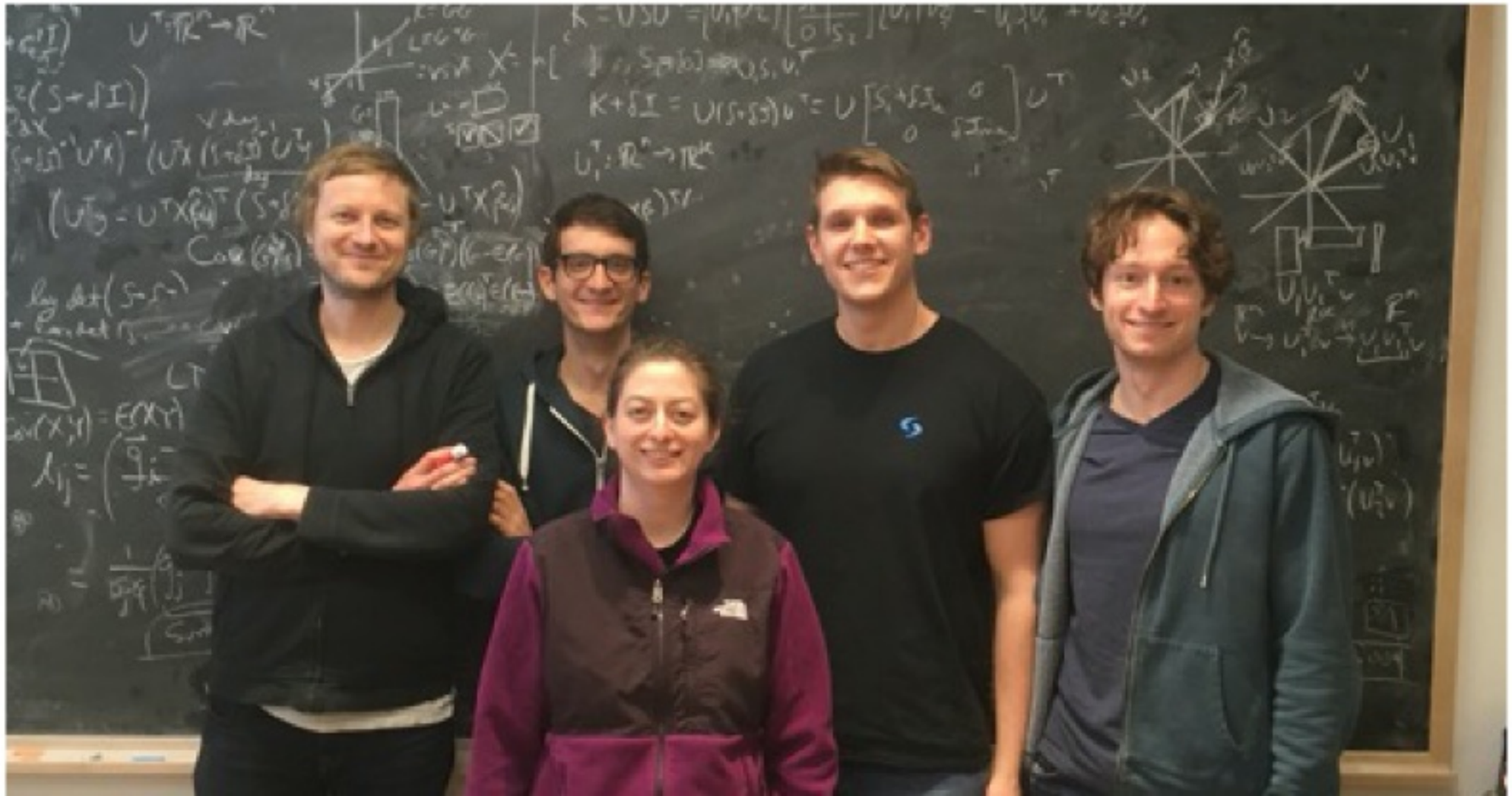
- Single-cell RNA profiles
- Deep phenotypes
- Best to analyze it all together!



We need …

Data Query MapReduce / SQL on Tensorial Structured Data	**Distributed Linear Algebra** Matrix Algebra Numerical Methods
Bayesian Inference Graphical Models MCMC Variational Inference	**Deep Learning** NN, CNN, RNN, GAN Representational Learning Reinforcement Learning

We need help!



Demo at Cray booth:

Tue 12:20 - 1:20
Wed 3:20 - 4:20

www.hail.is

hail@broadinstitute.org

www.broadinstitute.org/mia

Hail Team

Cotton Seed
Jackie Goldstein
Dan King
John Compitello

Contributors

Szabolcs Berecz
Alex Bloemendal
John Compitello
Laurent Francioli
Konrad Karczewski
Jack Kosmicki
Mitja Kurki
Mark Pinese
Ben Weisburd
Tom White
Alex Zamoshchin
@shusson
Patrick Schultz
Duncan Palmer

Broad Institute

Ben Neale
Mark Daly
Daniel MacArthur

Too many other scientific collaborators to list…