High-throughput Genomics at Your Fingertips with Apache Spark

Erwin Datema
Roeland van Ham

Spark Summit EU 2016, October 26, Brussels

Overview

High-throughput Genomics at Your Fingertips with Apache Spark

- Disclaimer : I am a scientist in computational biology (*bioinformatics*)
  - I am not a *computer* scientist
  - I am not a *data* scientist

- Scope
  - KeyGene's journey into Spark to analyze genomics data
  - Goal: enable interactive genomics data processing and querying
  - Told from a *user*'s perspective

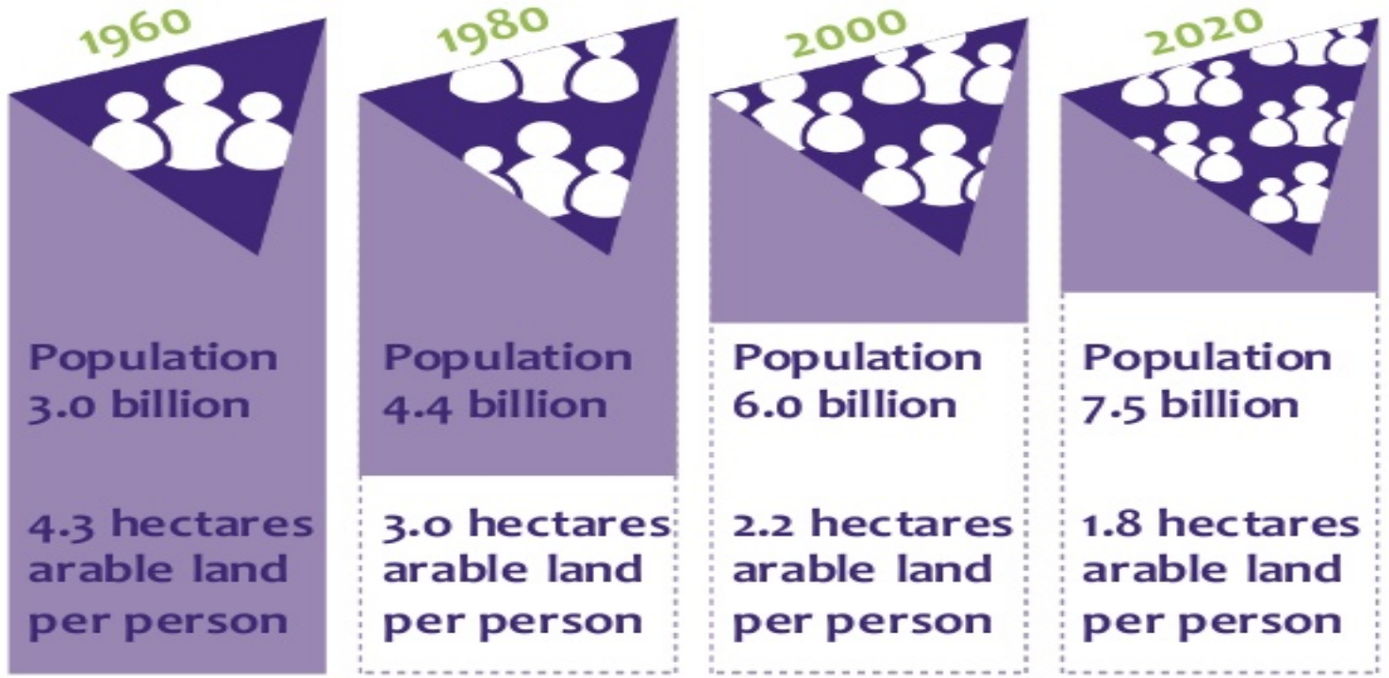
- Contents
  - Introduction to KeyGene
  - Crash Course Genomics
  - Big Data Challenges

Global trends

- World population grows from 7 to 9 billion people in 2050
- Climate change
- Limited/bad land, water and fossil fuels
- More obese people
- Malnourished people

Agricultural challenges

YIELD: Producing more food on less land



Genetic improvement of crops

Our strategy:

Use of natural genetic variation in crops

Molecular breeding
Molecular mutagenesis

Not GM:
At this moment too costly (20-100 mil €)

Regulatory, societal and technical hurdles






About KeyGene

>25 years

working for the future of global agriculture

1989

Now

Founded in 1989

Go-to Ag Biotech company for higher crop yield & quality

Current shareholders

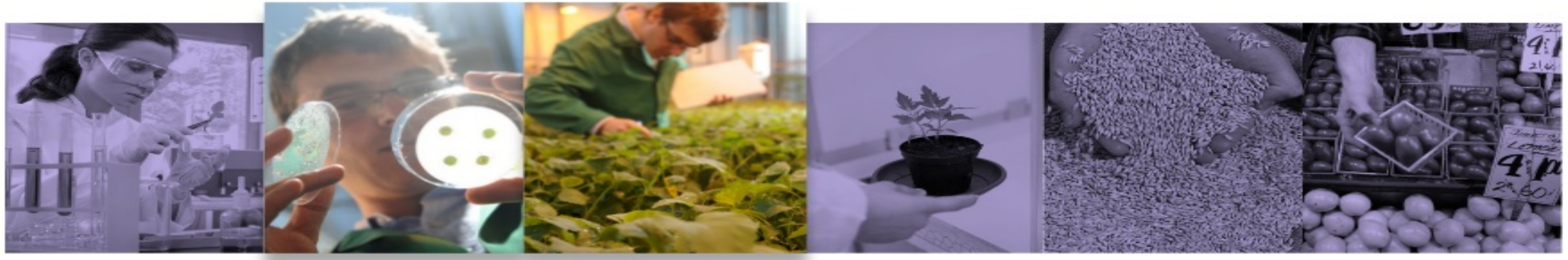
ENZA ZADEN

RIJK ZWAAN

Vilmorin & Cie

TAKII

R&D strategy



Fundamental research | Developing technologies & traits | Applying molecular breeding of crops | Breeding | Seed products | Market

← Universities

← KeyGene →

← - - - - Partners breeding industry →

Big Data in Genomics

Relevance

- Genomic data is being produced on an unprecedented scale
  - The cost to sequence a genome is now a few thousand dollars
  - We can routinely sequence tens to hundreds of individual plants

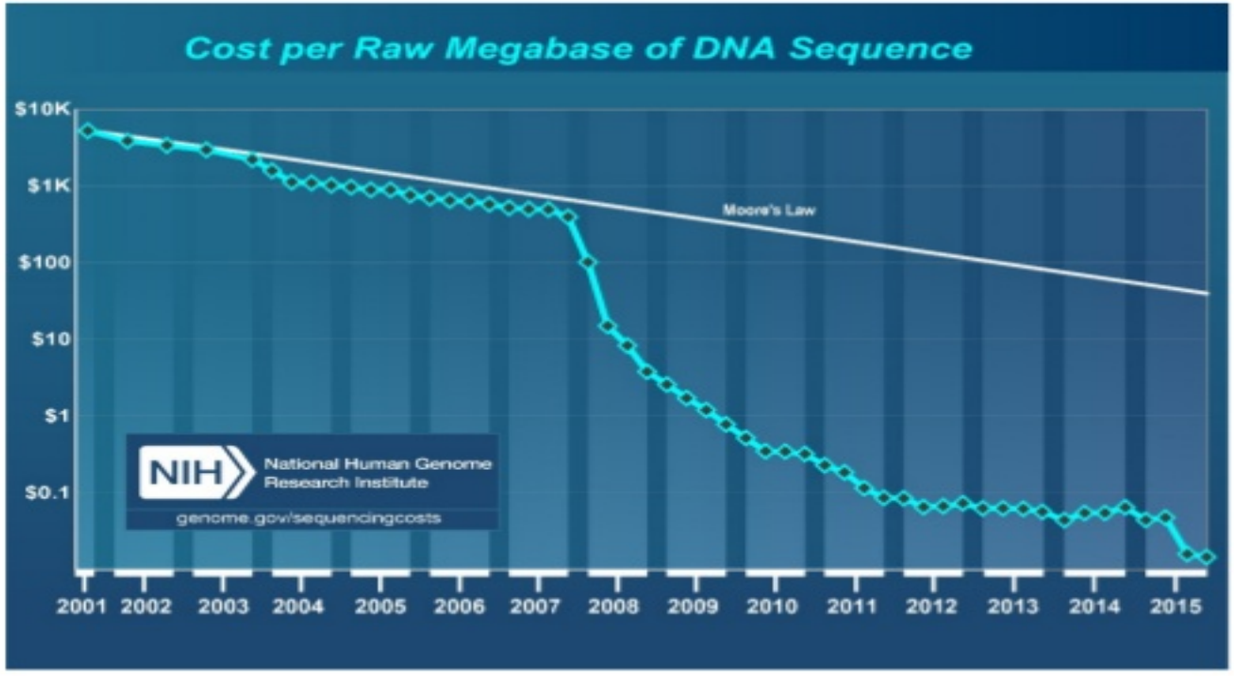
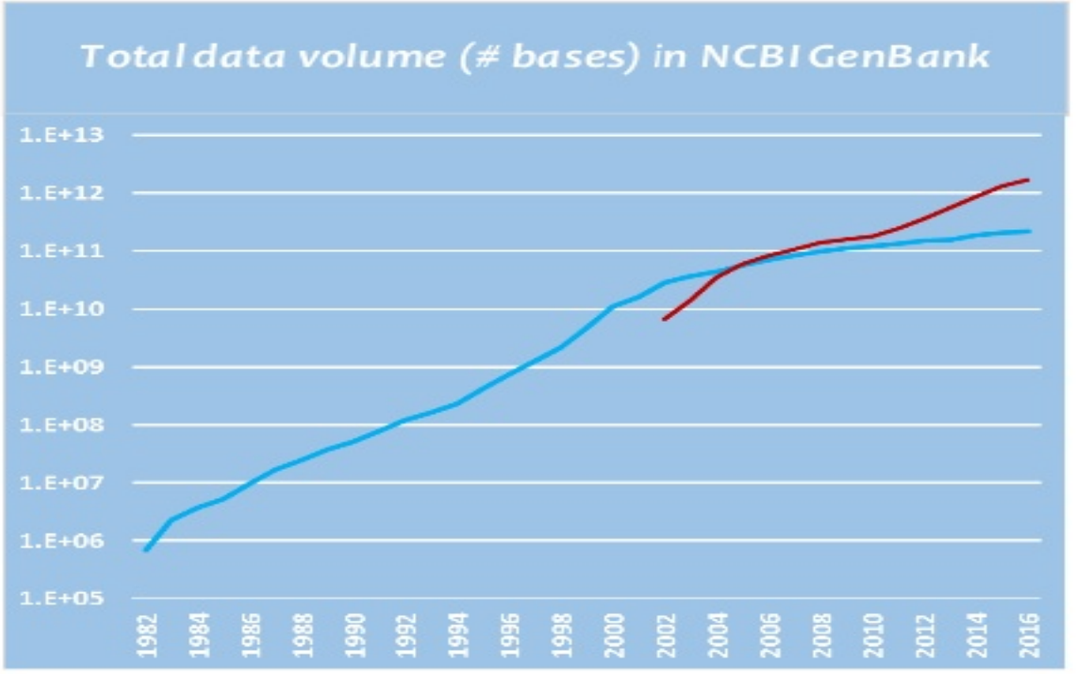


image source: https://www.genome.gov/images/content/costpermb2015_4.jpg



data source: https://www.ncbi.nlm.nih.gov/genbank/statistics/

Plant Genomics

Different from human genomics!

one genome	many genomes
"simple" genetics	complex genetics
high quality, shared resources	variable quality, fragmented resources
genomics as a diagnostic tool	genomics as a tool to direct breeding
understand and cure diseases	improve crop yield and quality

Crash Course Genomics

Plant genome sizes

~450 Mb	~850 Mb	~3.5 Gb	~5.5 Gb	~16 Gb
				
Melon	Potato	Pepper	Wheat	Onion
diploid	tetraploid	diploid	hexaploid	diploid

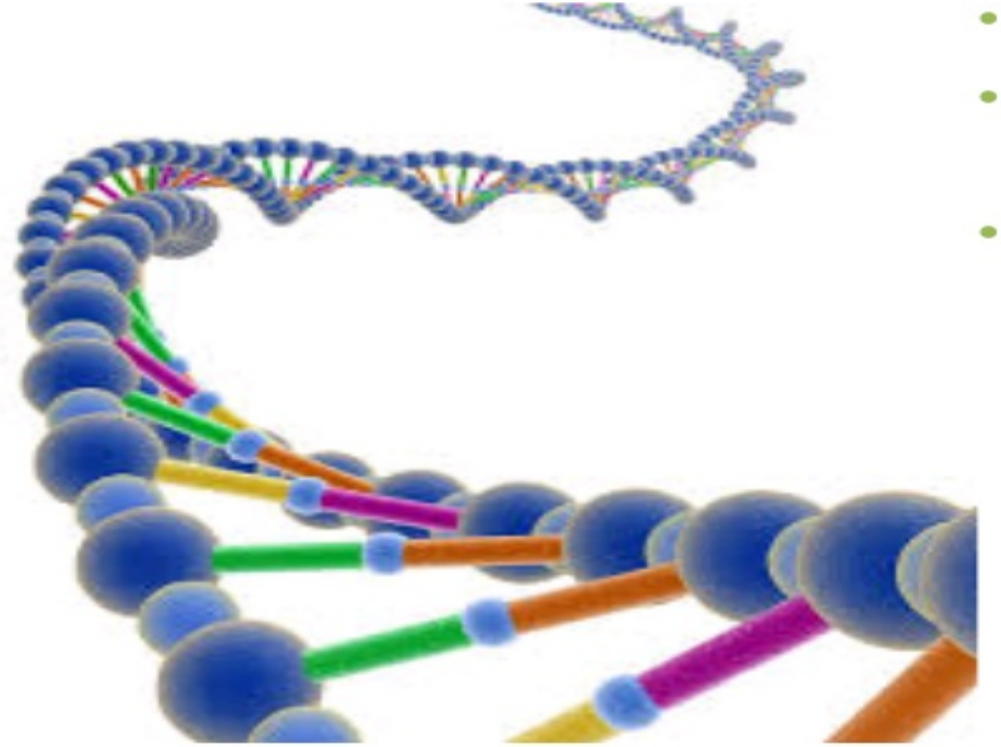
diploid = two sets of chromosomes
tetraploid = four sets of chromosomes

Mb = megabases
Gb = gigabases

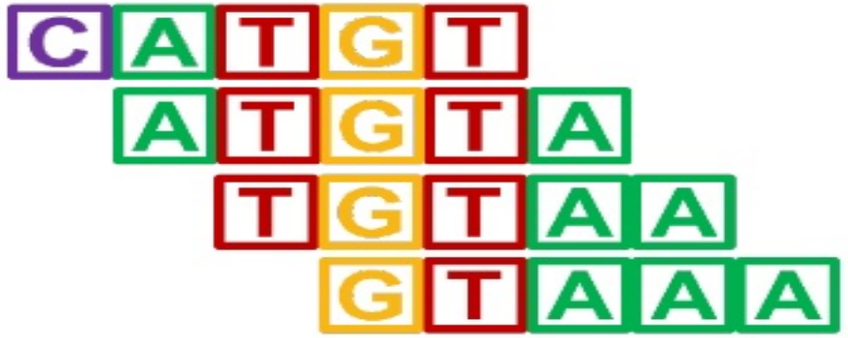
Crash Course Genomics

DNA, chromosomes, nucleotides

- DNA consists of four different elements (nucleotides or bases)
  - We represent DNA as strings of characters from the alphabet **A C G T**

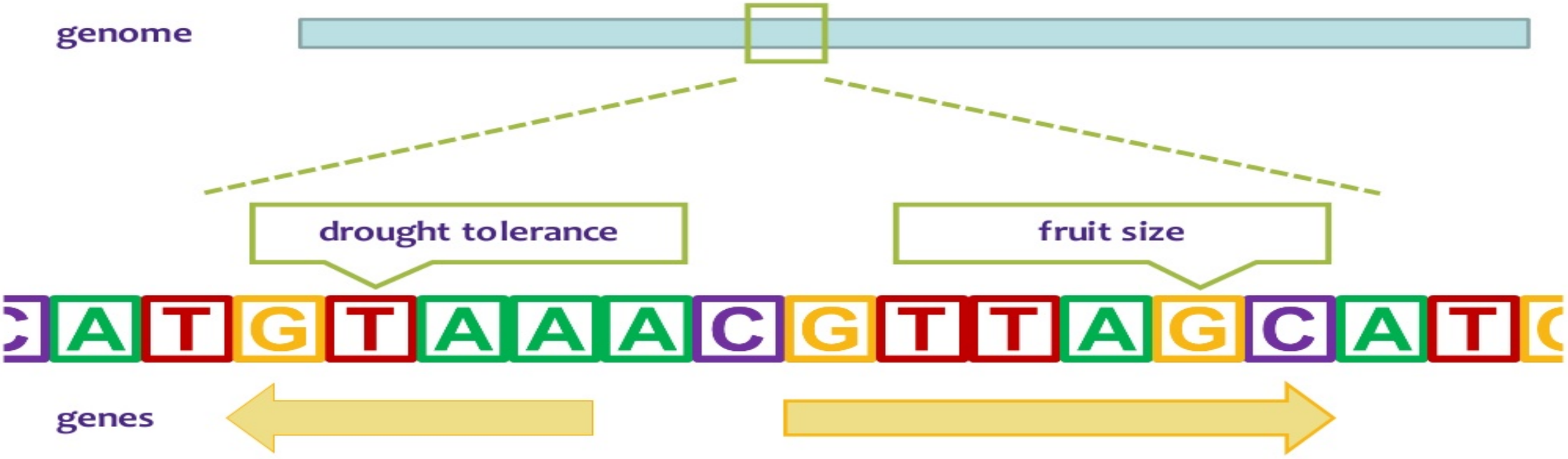


- DNA is organized into chromosomes
- Each chromosome contains millions of nucleotides (characters)
- We can only 'read' short pieces of DNA (hundreds to thousands of nucleotides)



Crash Course Genomics

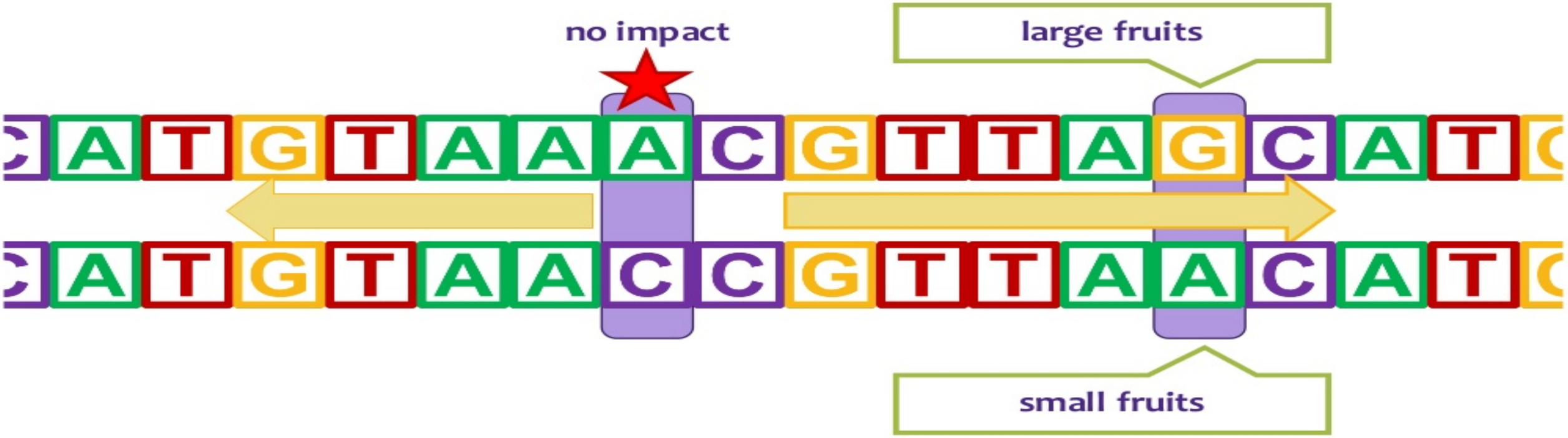
Genes and traits



- Genes are (some of) the functional elements in the genome
  - We represent genes as an interval on a chromosome

Crash Course Genomics

Polymorphisms and their impact



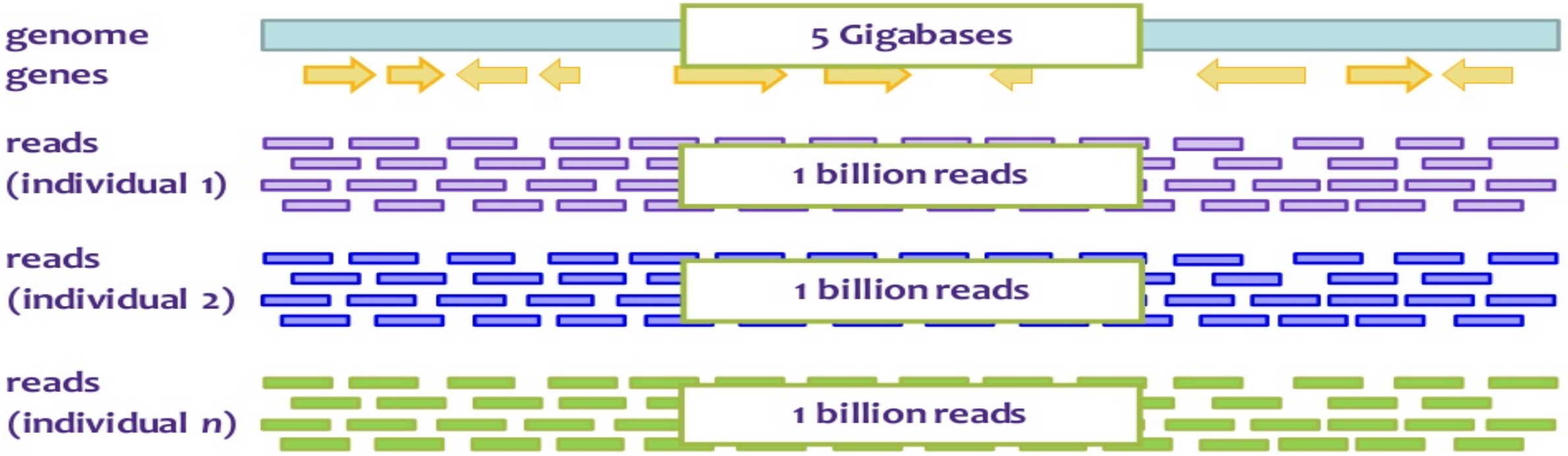
Crash Course Genomics

Read alignment and variant calling

- The 'reference genome' represents the known sequence of a species
- Reads are aligned against the reference genome (string similarity search)
  - Complex: sequence variation, repetitive regions and data errors
- Variants are called from differences observed in 'pile-ups' of reads

Crash Course Genomics

Population-scale genome sequencing



- Align a billion reads x 1,000 individuals to a 5 Gb genome
- Call hundreds of millions (up to potentially billions) of sequence variants

Crash Course Genomics

Recap

- Genome sequences are represented as strings of **A C G T**
- Variation between genome sequences underlies differences in traits
- We can "read" the genome in little pieces
  - High throughput, massively parallel sequencing technologies
  - Up to thousands of individual plants from a given species
- Genomics data analysis is challenging
  - Rapid increase in data generation
  - Rapid turnover of sequencing technologies and their outputs
  - Scientific software is (often) bad (*Nature News, Oct 13 2010*)

Genomics data analysis

High Performance Computing

- Computational challenges
  - Align billions of reads to the reference genome
  - Call millions or billions of sequence variants
  - Determine the small number of variants that impact a given trait

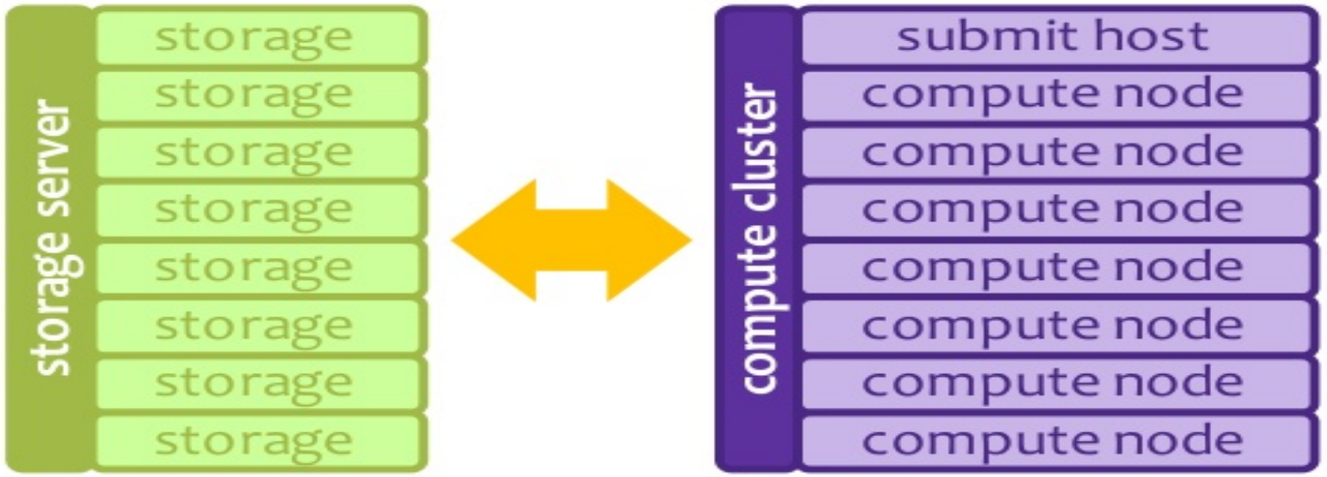
- HPC infrastructure (e.g. SGE clusters) are the *de facto* standard
  - Manually split large datasets (to accommodate the job scheduler)
  - Manually deal with failures: check logs, resubmit jobs…

- Many software tools are in fact large, monolithic "pipelines"
  - No fine-grained control over resource usage
  - A single error often implies a complete re-run of the analysis

High Performance Computing

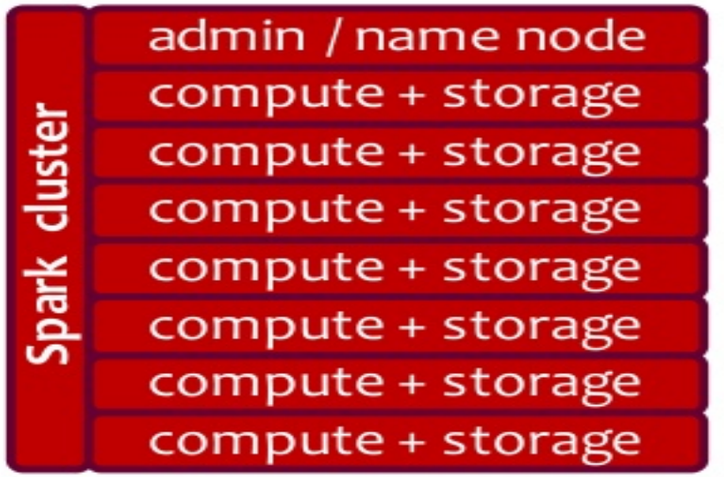
Big Data technologies

Conventional Compute Cluster



Expensive, proprietary storage
Expensive network connections
Expensive, high-reliability hardware

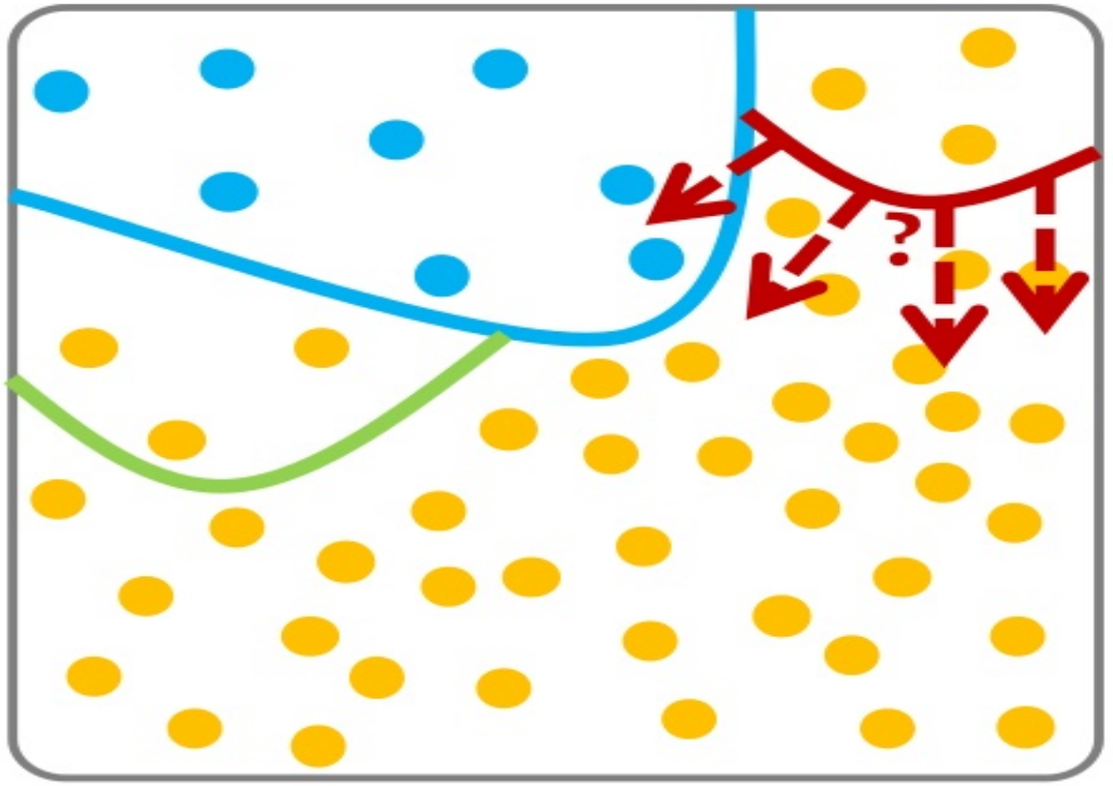
Spark Cluster



Commodity hardware
Linear scalability
Fault tolerance

Genomics application landscape

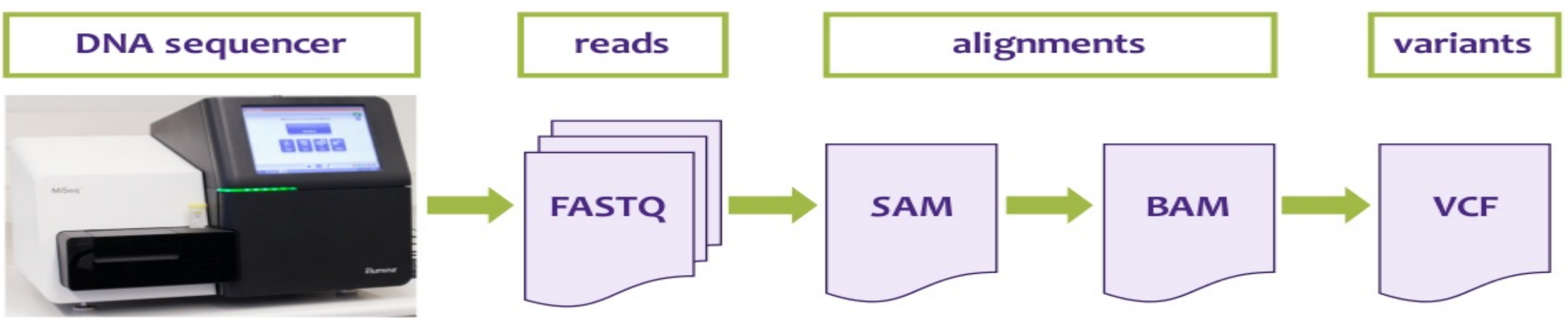
Opportunities for Big Data technologies



Big Data in Genomics

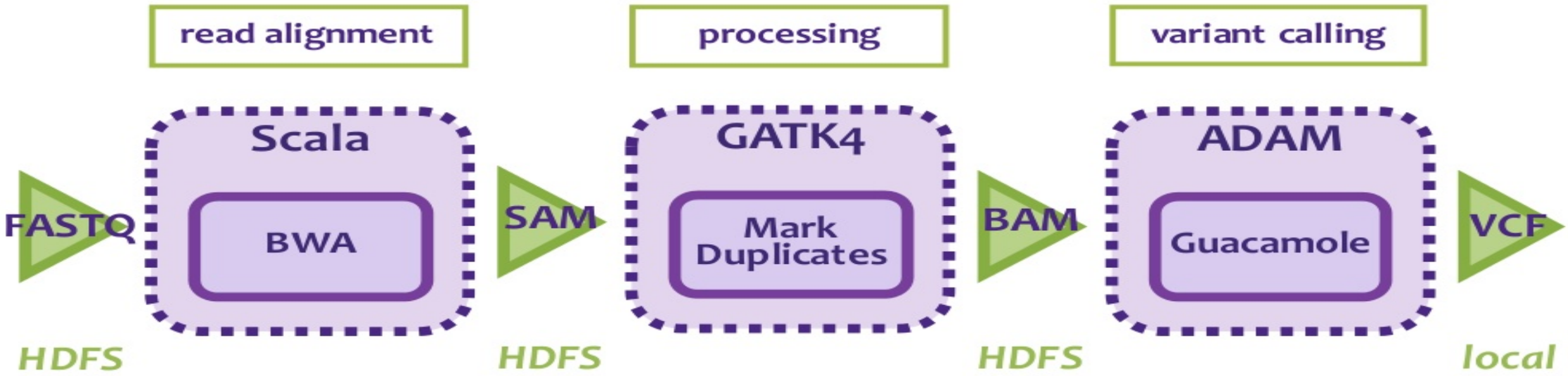
Challenges

- Genomics is a dynamic, rapidly changing field
  - Data generators and analysis algorithms are in constant flux
  - Tools are generally built around flat, text-based file formats
  - Workflow is file-centric (POSIX file system; no streaming…)



Big Data in Genomics

Our 'Sparkified' pipeline



Genomics on Spark

Solutions to legacy designs

DNA sequencer → **reads** + **reads** → **'interleaved' reads**

FASTQ + FASTQ → **RDD**[String]

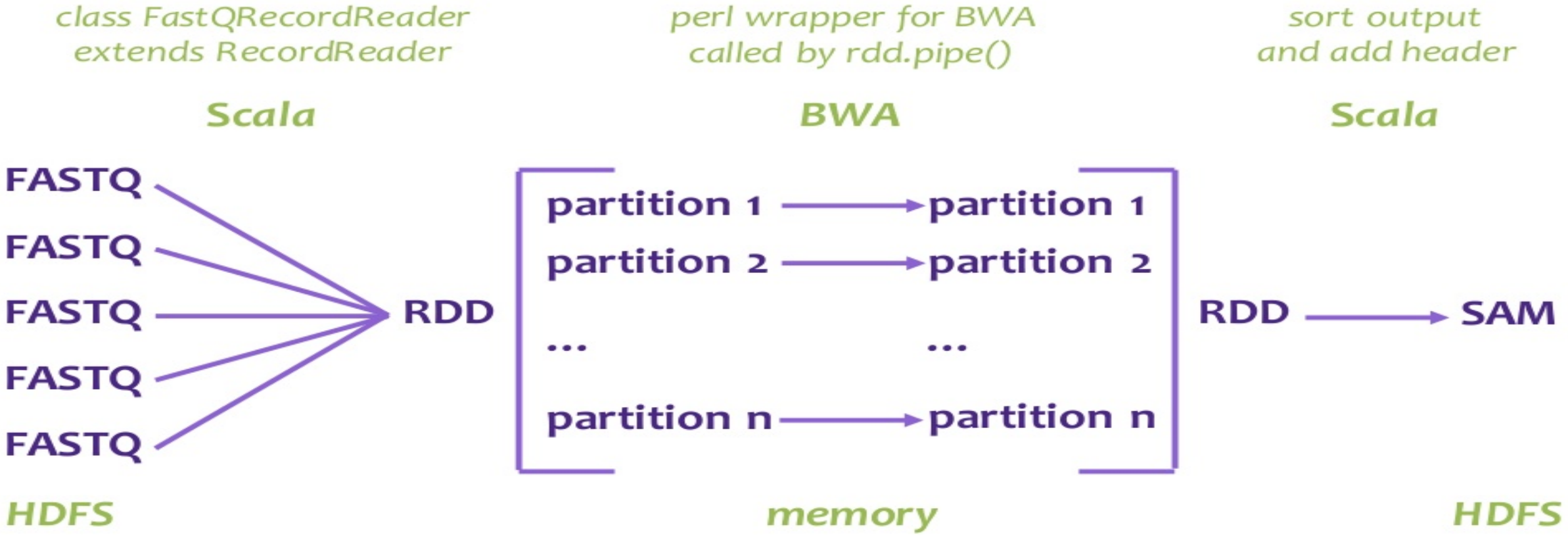
f1.zip(f2).map
(e => e._1+ e._2)

class FastQRecordReader extends RecordReader

identifier	@SEQ_ID
sequence	GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
separator	+
quality	!''*(((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65

Genomics on Spark

Read alignment



Genomics on Spark

Processing and variant calling

gatk

Broad Institute, Cambridge

- Alignment processing
- Variant calling (*in development*)

APACHE Spark Parquet AVRO

AMPlab, University of California

- Data schemas + APIs
- Variant calling (Guacamole)

Hail

Scalable genetic data analysis

http://hail.is

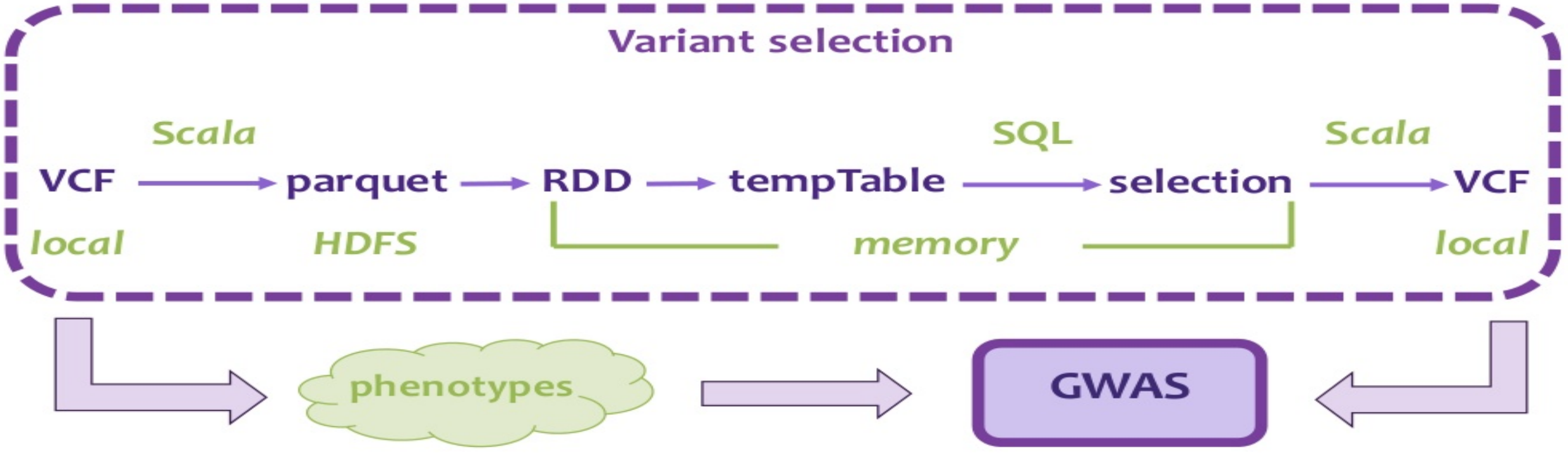
Broad Institute, Cambridge

- Variant analysis
- You've all attended the Keynote talk…

Genomics on Spark

Variant selection and analysis

- Interactive, "real-time" selection of variant data with simple SQL queries
- GWAS analysis on Spark (e.g. hail) or conventional infra (e.g. PLINK)



Big Data in Genomics

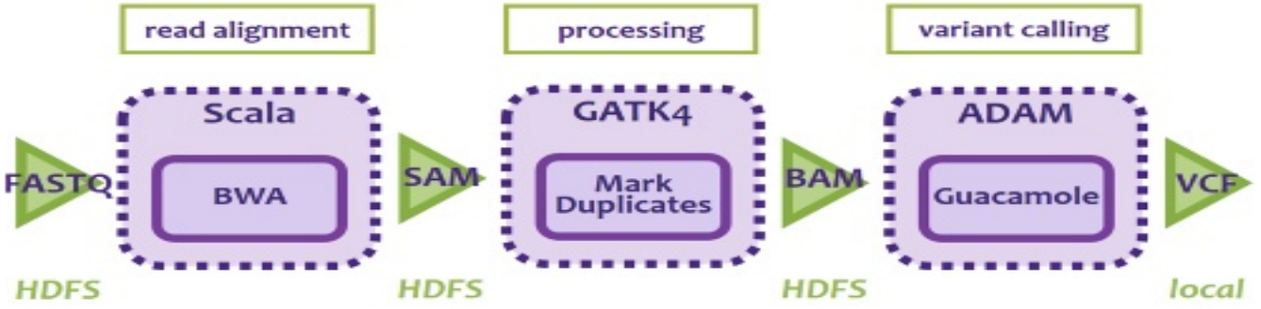
KeyGene's ambition

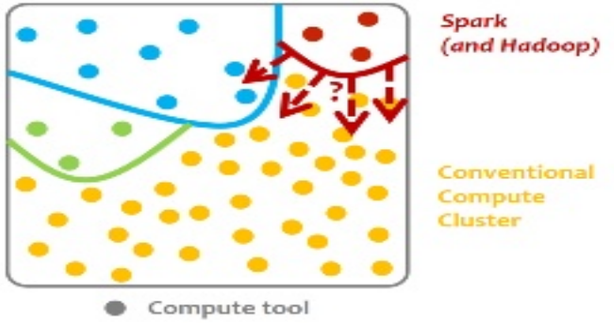
Wrap-up

Conclusions and lessons learnt

- Initial success in applying Spark to plant genomics
  - Proof-of-concept for enabling interactive GWAS analysis on Spark

- Spark appears to be a good fit for (some of our current) Genomics problems
  - Developer community needed to translate core Genomics applications!
  - Paradigm shift required to move away from flat POSIX files…
  - Opportunities for streaming data analysis

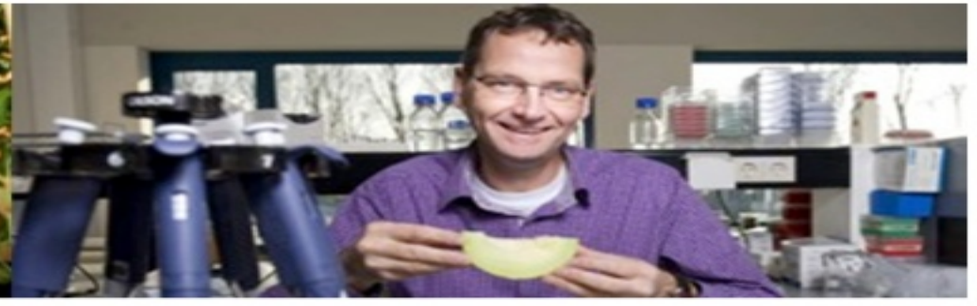




The End

High-throughput Genomics at Your Fingertips with Apache Spark


Thank you for your attendance!

- Erwin Datema erwin.datema@keygene.com
- Roeland van Ham roeland.van-ham@keygene.com