UNIVERSITÄT DES SAARLANDES

ZBI ZENTRUM FÜR BIOINFORMATIK

mpii

# Day 1: Computational Pan-Genomics: Status, Promises and Challenges

Jordan Eizenga, Erik Garrison and Tobias Marschall

CPANG18 @ Instituto Gulbenkian de Ciência
March, 2018

# Rebooting the Human Genome*

*„The Human Genome Project was one of mankind's greatest triumphs. But the official gene map that resulted in 2003, known as the "reference genome," is no longer up to the job."*

(Antonio Regalado, MIT Technology Review, June 3, 2015)

*https://www.technologyreview.com/s/537916/rebooting-the-human-genome

Lorentz center

# Future Perspectives in Computational Pan-Genomics

Workshop: 8 – 12 June 2015, Leiden, the Netherlands

**Scientific Organizers**

- Victor Guryev, ERIBA Groningen
- Tobias Marschall, Saarland U / MPII
- Alexander Schönhuth, CWI Amsterdam
- Fabio Vandin, SDU Odense
- Kai Ye, St. Louis, Washington U

**Invited Speakers**

- Can Alkan, Bilkent U Ankara
- Paul De Bakker, UMC Utrecht
- Valentina Boeva, Institut Curie Paris
- Francesca Chiaromonte, Penn State U
- Francesca Ciccarelli, King's College London
- Evan Eichler, U Washington
- Eleazar Eskin, UCLA
- Paul Kersey, EMBL EBI
- Jan Korbel, EMBL Heidelberg
- Jens Lagergren, Karolinska Institute
- Ben Langmead, Johns Hopkins U
- Veli Mäkinen, U Helsinki
- Manja Marz, U Jena
- Paul Medvedev, Penn State U
- Sven Rahmann, U Duisburg-Essen
- Ben Raphael, Brown U
- Knut Reinert, FU Berlin
- Cenk Sahinalp, Simon Fraser U
- Ole Schulz-Trieglaff, Illumina UK

Workshop Paper

Briefings in Bioinformatics Advance Access published October 21, 2016

OXFORD

Briefings in Bioinformatics, 2016, 1–18

doi: 10.1093/bib/bbw089
Paper

# Computational pan-genomics: status, promises and challenges

The Computational Pan-Genomics Consortium*

Corresponding author: Tobias Marschall, Center for Bioinformatics at Saarland University and Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken, Germany. Tel.: +49 681 302 70880; E-mail: t.marschall@mpi-inf.mpg.de
*The Computational Pan-Genomics Consortium formed at a workshop held from 8 to 12 June 2015, at the Lorentz Center in Leiden, the Netherlands, with the purpose of providing a cross-disciplinary overview of the emerging discipline of Computational Pan-Genomics. The workshop was organized by Victor Guryev, Tobias Marschall, Alexander Schönhuth (chair), Fabio Vandin, and Kai Ye. Consortium members are listed at the end of this article.



## Abstract

Many disciplines, from human genetics and oncology to plant breeding, microbiology and virology, commonly face the challenge of analyzing rapidly increasing numbers of genomes. In case of Homo sapiens, the number of sequenced genomes will approach hundreds of thousands in the next few years. Simply scaling up established bioinformatics pipelines will not be suff cient for leveraging the full potential of such rich genomic data sets. Instead, novel, qualitatively different computational methods and paradigms are needed. We will witness the rapid extension of computational pan-genomics, a new sub-area of research in computational biology. In this article, we generalize existing def nitions and understand a pan-genome as any collection of genomic sequences to be analyzed jointly or to be used as a reference. We examine already available approaches to construct and use pan-genomes, discuss the potential benef ts of future technologies and methodologies and review open challenges from the vantage point of the above-mentioned biological disciplines. As a prominent example for a computational paradigm shift, we particularly highlight the transition from the representation of reference genomes as strings to representations as graphs. We outline how this and other challenges from different application domains translate into common computational problems, point out relevant bioinformatics techniques and identify open

# Table of Content

# Table of Content

# What is a Pan-Genome?

Term **pan-genome** popularized in microbiology in 2005.

## Definition (gene-based pan-genome)

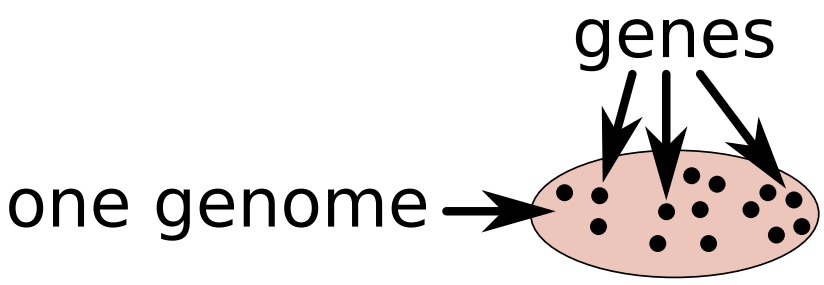
The **pan-genome** of a species (or other taxonomic unit) is the **union of all sets of genes** across all individuals.

# What is a Pan-Genome?

Term **pan-genome** popularized in microbiology in 2005.

## Definition (gene-based pan-genome)

The **pan-genome** of a species (or other taxonomic unit) is the **union of all sets of genes** across all individuals.

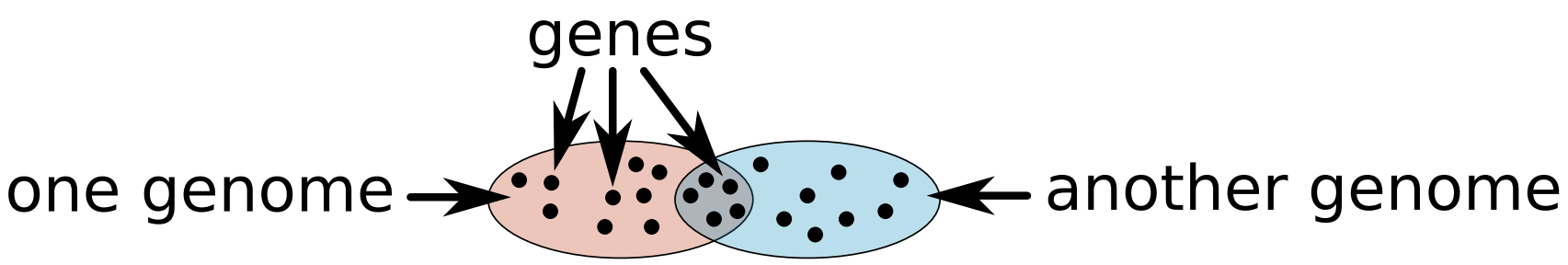


# What is a Pan-Genome?

Term **pan-genome** popularized in microbiology in 2005.

## Definition (gene-based pan-genome)

The **pan-genome** of a species (or other taxonomic unit) is the **union of all sets of genes** across all individuals.

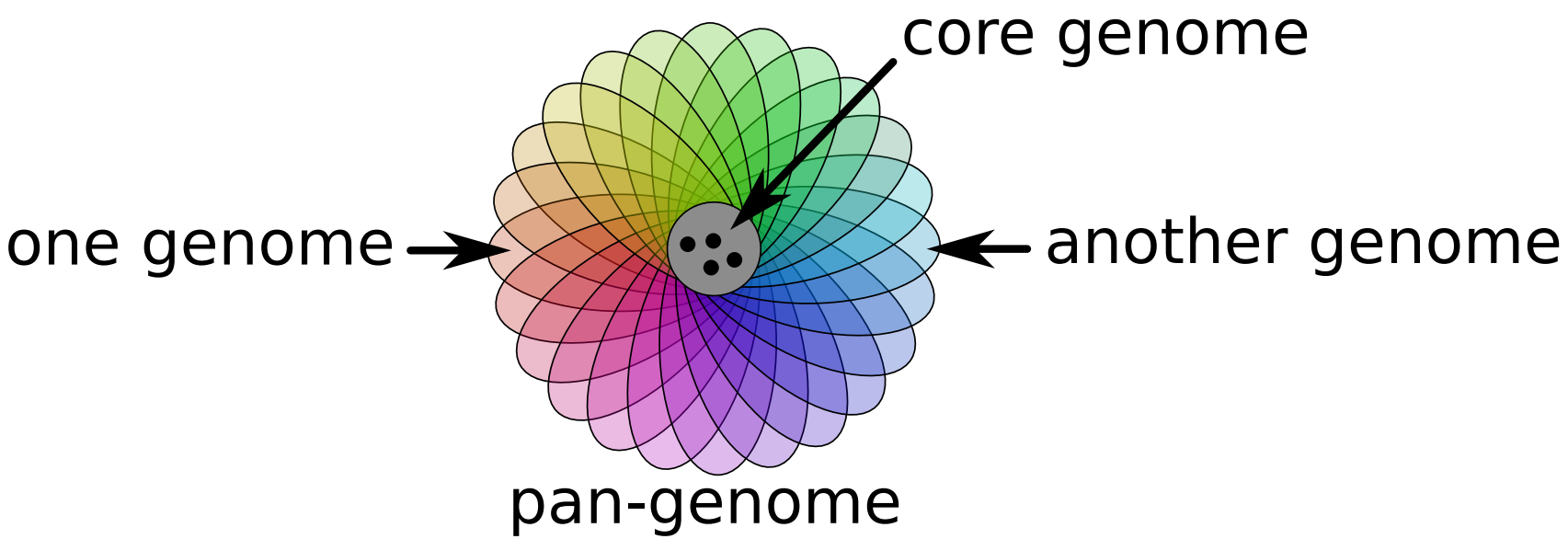


# What is a Pan-Genome?

Term **pan-genome** popularized in microbiology in 2005.

## Definition (gene-based pan-genome)

The **pan-genome** of a species (or other taxonomic unit) is the **union of all sets of genes** across all individuals.



# Pan-Genome Definition

*"... and use the term* **pan-genome** *to refer to any collection of genomic sequences to be analyzed jointly or to be used as a reference. These sequences can be linked in a graph-like structure, or simply constitute sets of (aligned or unaligned) sequences. Questions about efficient data structures, algorithms and statistical methods to perform bioinformatic analyses of pan-genomes give rise to the discipline of* **computational pan-genomics**."

# Pan-Genome Definition

*"... and use the term* **pan-genome** *to refer to any collection of genomic sequences to be analyzed jointly or to be used as a reference. These sequences can be linked in a graph-like structure, or simply constitute sets of (aligned or unaligned) sequences. Questions about efficient data structures, algorithms and statistical methods to perform bioinformatic analyses of pan-genomes give rise to the discipline of* **computational pan-genomics***."*

## Notes

- Not restricted to taxonomic units
- Not restricted to full genomes
- Not tied to graphs
- Intentionally intersects with **metagenomics**, **comparative genomics**, and **population genetics**

# Table of Content

# (High-Level) Goals of Computational Pan-Genomics

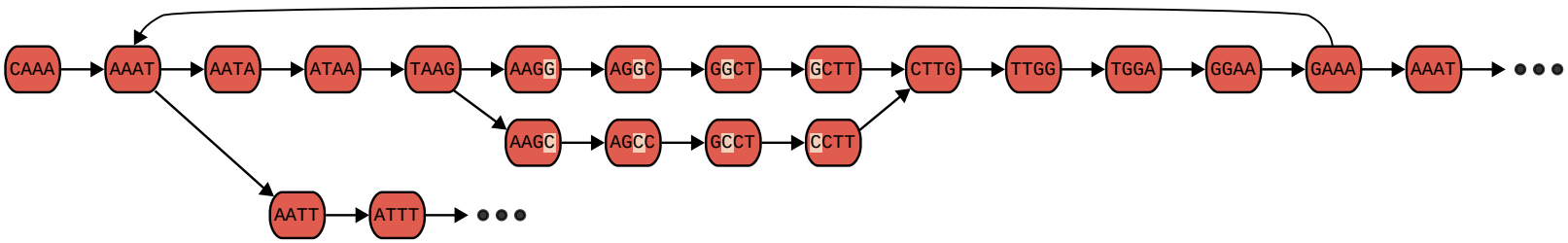
- **completeness**: containing all functional elements and enough of the sequence space to serve as a reference for the analysis of additional individuals,
- **stability**: having uniquely identifiable features that can be studied by different researchers and at different points in time,
- **comprehensibility**: facilitating understanding of the complexities of genome structures across many individuals or species,
- **efficiency** organizing data in such a way as to accelerate downstream analysis.

# Table of Content

# Approaches I

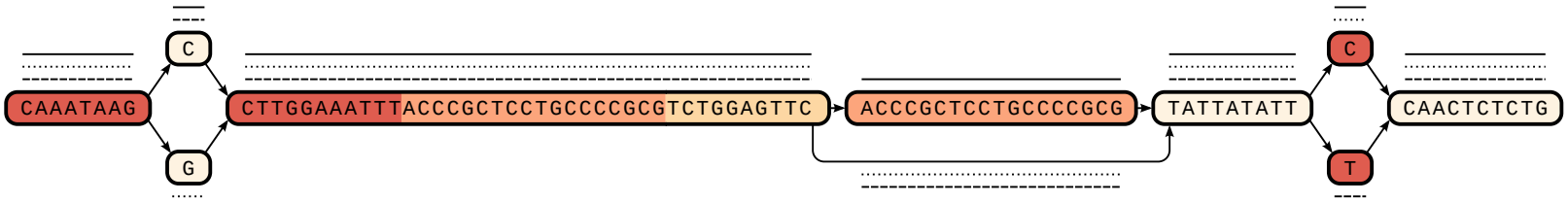

(a) Unaligned sequences


(b) Multiple sequence alignment

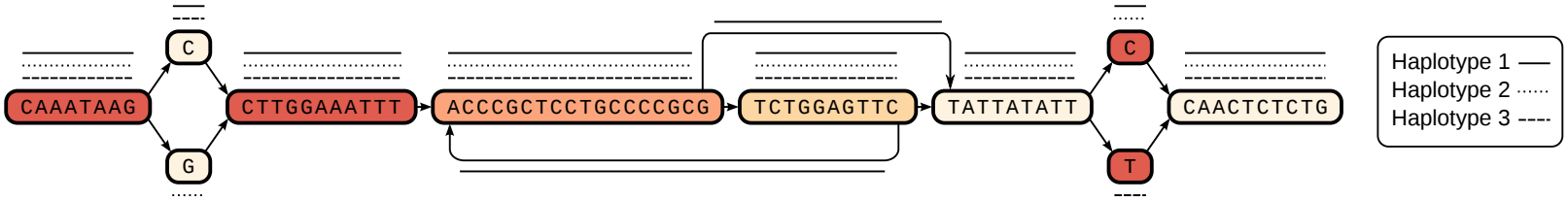



(c) De Bruijn graph

# Approaches II

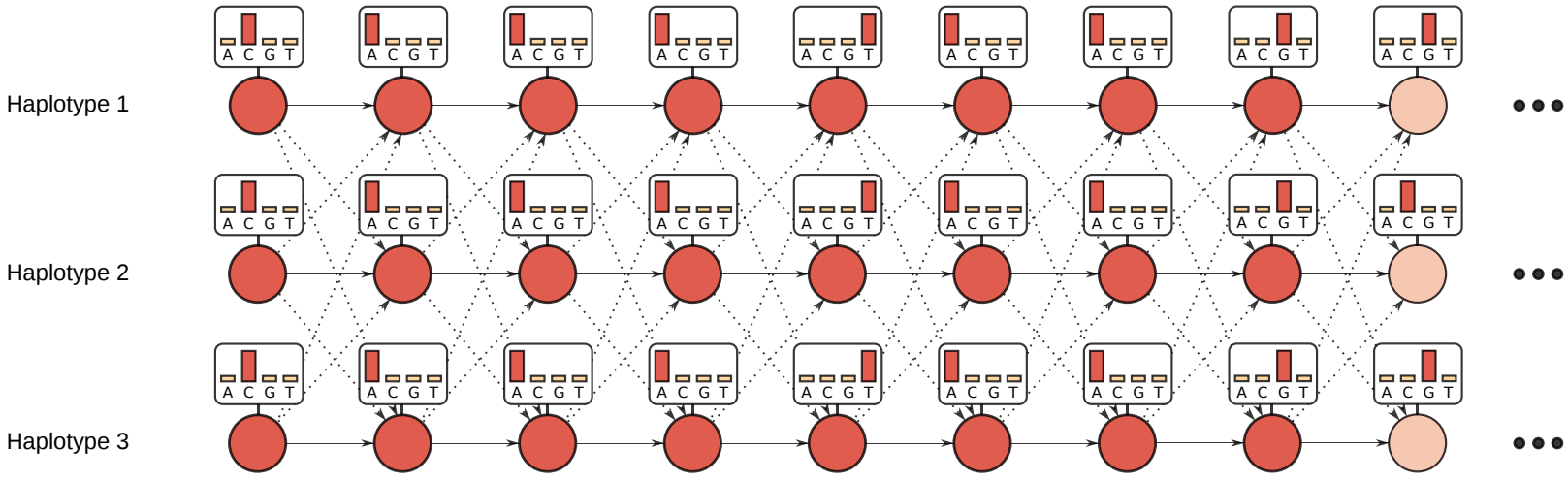



(d) Acyclic sequence graph




(e) Cyclic sequence graph

# Approaches III




(f) Li-Stephens model

# Course Objectives

- Understand Pan-Genomics concepts and appreciate the **limitations of linear reference genomes**,
- Learn to build corresponding analysis pipelines in the **VG framework**,
- We focus on putting you in a position to **design (and debug) pipelines** tailored to your use case,
- Some of the practicals address **serious research questions** (that we do not completely solve in class), the aim is rather to give you the tools to attack them.

# Getting to know each other

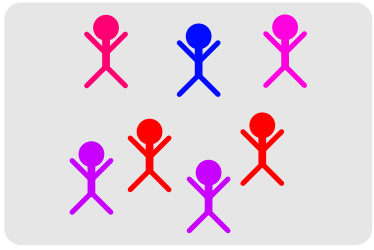
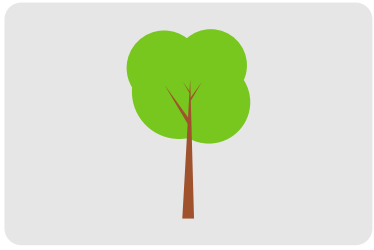
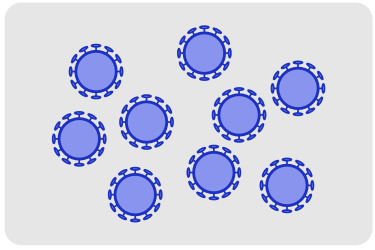
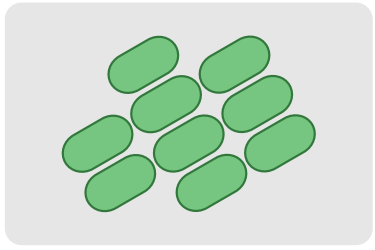
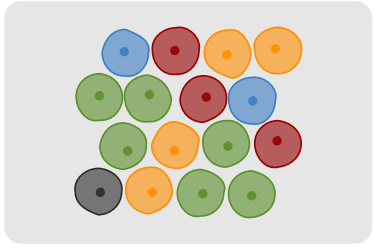
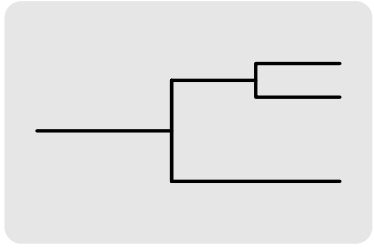
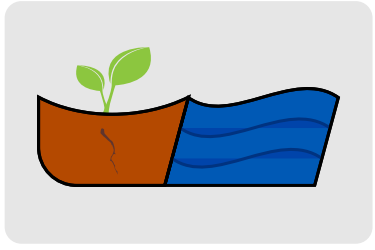
- What is your background?
- What are your expectations?
- Which data / use cases do you work on (or will work on)?

# Practicals

- This is a hands-on course. We collect praticals in this git repository: `https://github.com/Pfern/PANGenomics`
- We add new practicals for each day in the course of this week
- Ppracticals are a starting point for **exploring what VG can do**. So please don't only copy paste commands, but try to understand their meaning, modify them, etc.
- Help each other
- Present your results in class
- Don't hesitate to send pull requests ;)

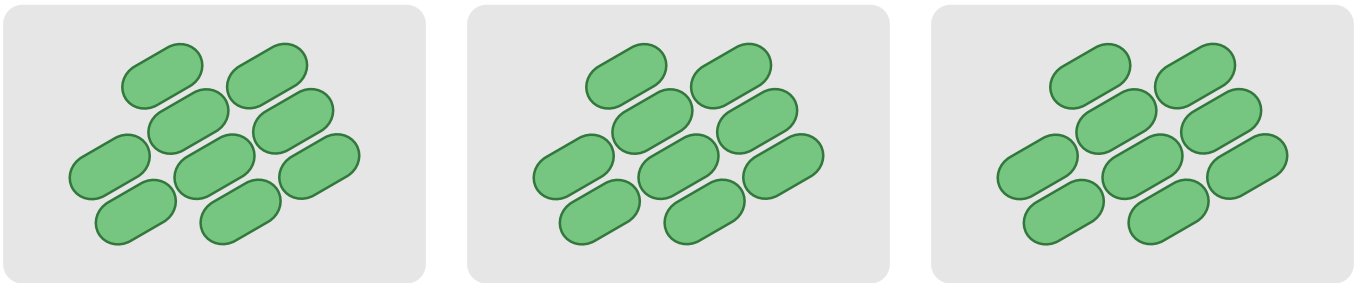
# Table of Content

# Applications

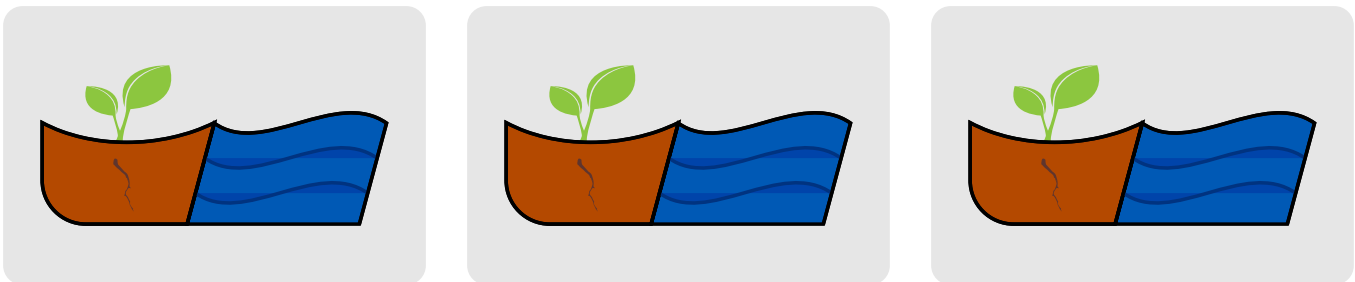
# Application Domain: Microbes

- Pan-genomics at the **gene level**: established workflows and mature software are available
- For a number of microorganisms, **pan-genome sequence data is already available**
- Microbial pan-genomes support **comparative genomics studies** (especially given horizontal gene exchange)
- **Genome-wide association studies (GWAS)** for microbes is an emerging field



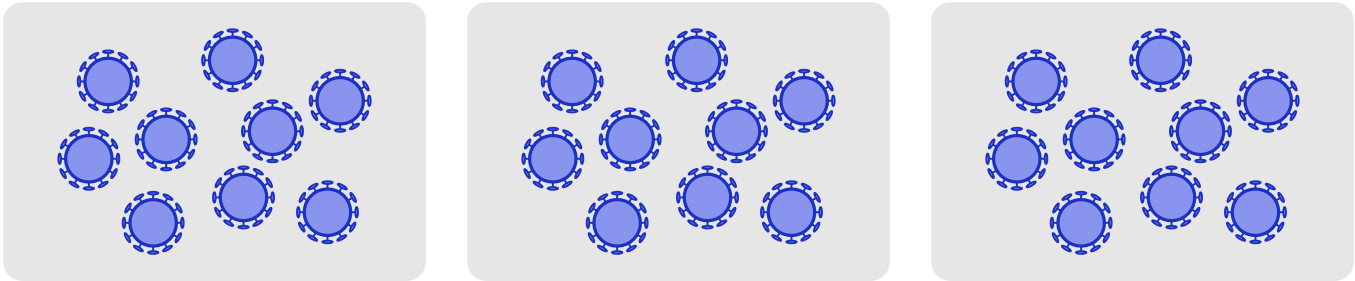
# Application Domain: Metagenomics

- Metagenomics: set of genomic sequences **co-occurrence in an environment**
- Questions: **taxonomic composition** of the sample, **presence of certain gene products** or whole pathways, and determining **which genomes these functional genes** are associated with.
- Pan-Genome data structures present the chance to reveal **common adaptations** to the environment as well as **co-evolution** of interactions.



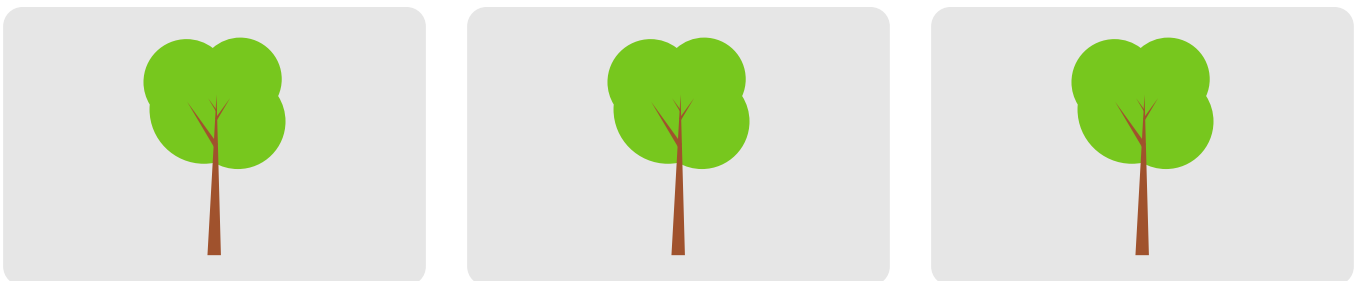
# Application Domain: Viruses

- Reliable **viral haplotype reconstruction** is not fully solved
- **Patient's viral pan-genome $\rightarrow$ diagnosis, staging, and therapy selection**
- **Virus-host interactions:** pan-genome structure of a viral population to be directly compared with that of a susceptible host population



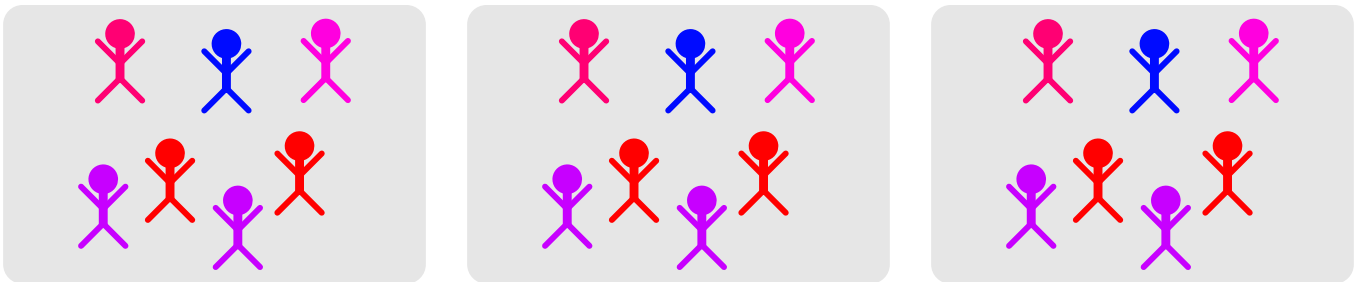
# Application Domain: Plants

- **Large-scale genomics projects** completed / under way: *Arabidopsis thaliana*, rice, maize, sorghum, and tomato
- Plant genomes are **large**, **complex** (containing many repeats) and often **polyploid**
- Having a pan-genome available for a given crop that includes its wild relatives provides a **single coordinate system** to anchor all known variation and phenotype information



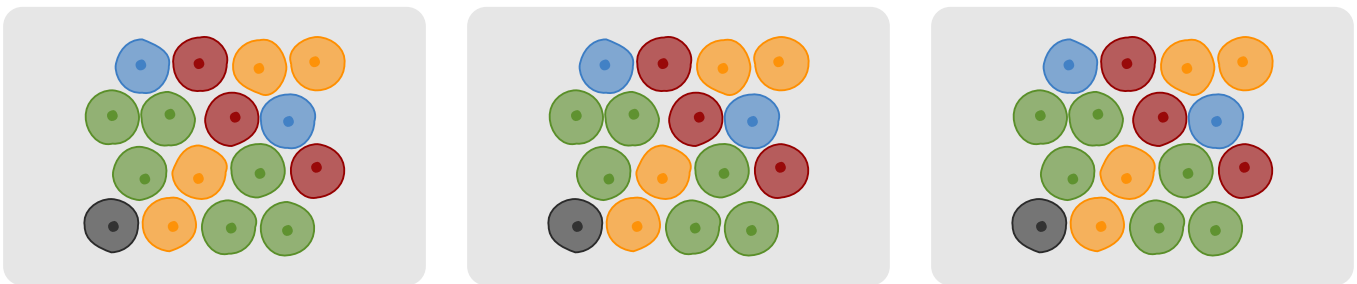
# Application Domain: Human Genetic Diseases

- Numerous genes have been successfully mapped for **rare monogenic diseases**
- **Common diseases** ← GWAS ← imputation ← catalogs of human genome variation, their linkage disequilibrium (LD) properties
- Pan-Genomics can help to achieve this, especially in **difficult genomic regions**



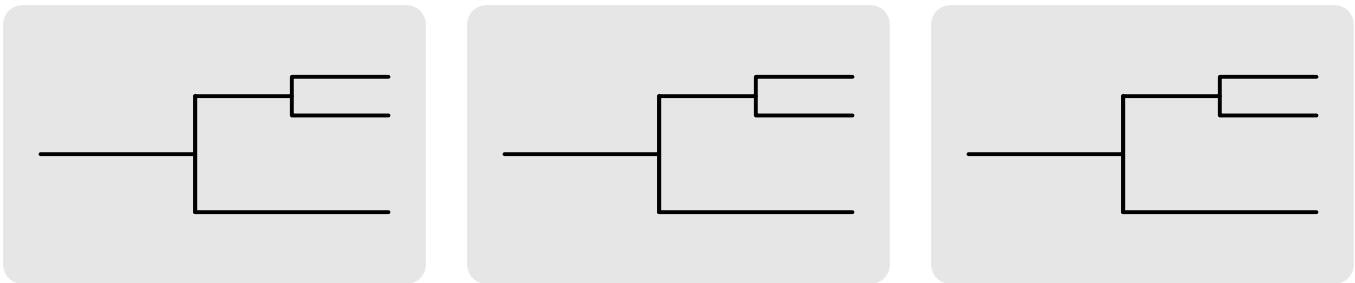
# Application Domain: Cancer

- Improved **detection of somatic mutations**, through improved quality of read mapping to polymorphic regions
- Somatic pan-genome describing the **general somatic variability** in the human population, $\rightarrow$ accurate baseline for assessing the impact of somatic alterations.
- Vision: **personal cancer pan-genome** to be built for **each tumor patient**: single-cell data, haplotype information, sequencing data from circulating tumor cells and DNA, etc.



# Application Domain: Phylogenomics

- Computational pan-genomics: **repidly extrcat evolutionary signals**, such as gene content tables, sequence alignments of shared marker genes, genome-wide SNPs, or internal transcribed spacer (ITS) sequences
- Move beyond only using the **best aligned, and most well behaved residues of a multiple sequence alignment** (often used in "traditional" phylogenomics)



# Operations on Pan-Genomes