

Telecon 2

Visual Analytics for Plant Pangenomes (VAPP)

Astrid van den Brandt | Eindhoven University of Technology

Michel Westenberg | Eindhoven University of Technology

Sandra Smit | Wageningen University & Research



12 MARCH 2021

Agenda

1. Recap project and aims
2. Progress since T1
3. Planning and next steps
4. Questions and feedback

Project Aims

To design a **visual analytics system** that supports plant genome scientists in **analyzing genetic variation** in crop data

Visual Analytics (VA)

Computer-based visualization systems provide **visual representations** of datasets designed to help people carry out **tasks** more effectively.

VA for Genomics

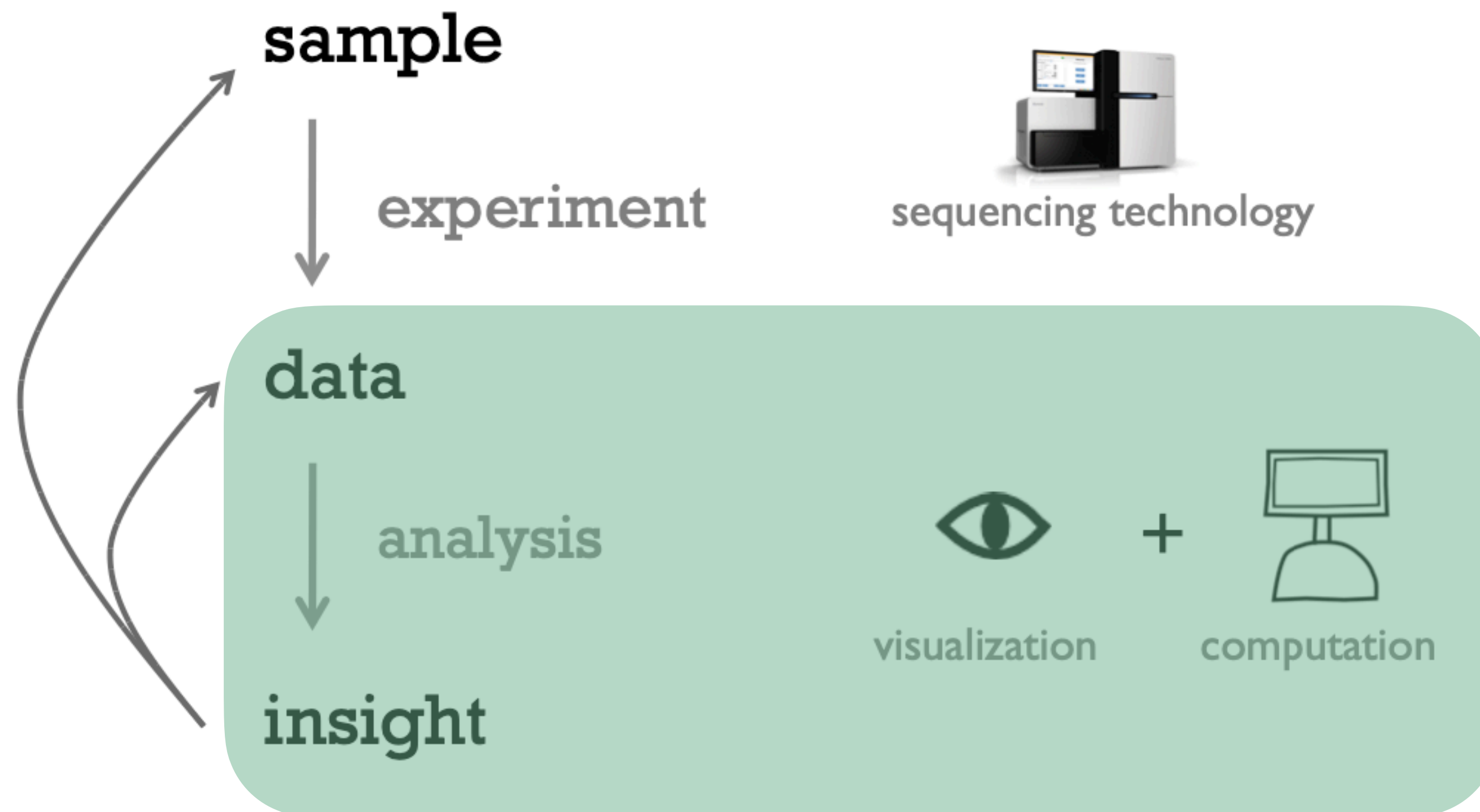


Figure: CANVAS (Nielsen, 2012)

Computational biology
Bioinformatics
Visual Analytics

Defining the Problem

1. Genomic data in general

- Growing volume of (sequence) data
- Sparse distributions of features
- Many different data types

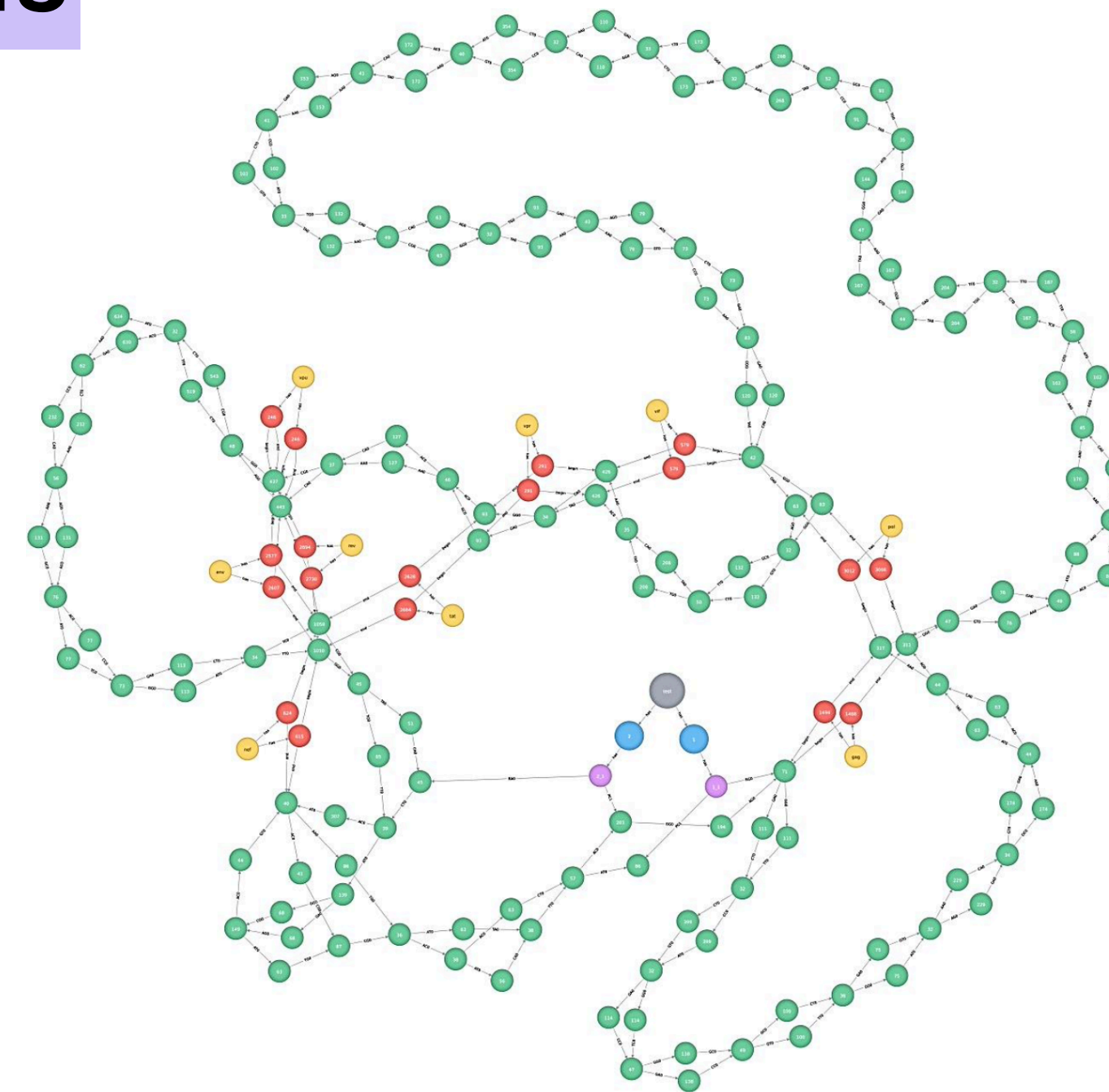
2. Plant genome data

- Large genomes
- Large variation between genomes

Defining the Problem

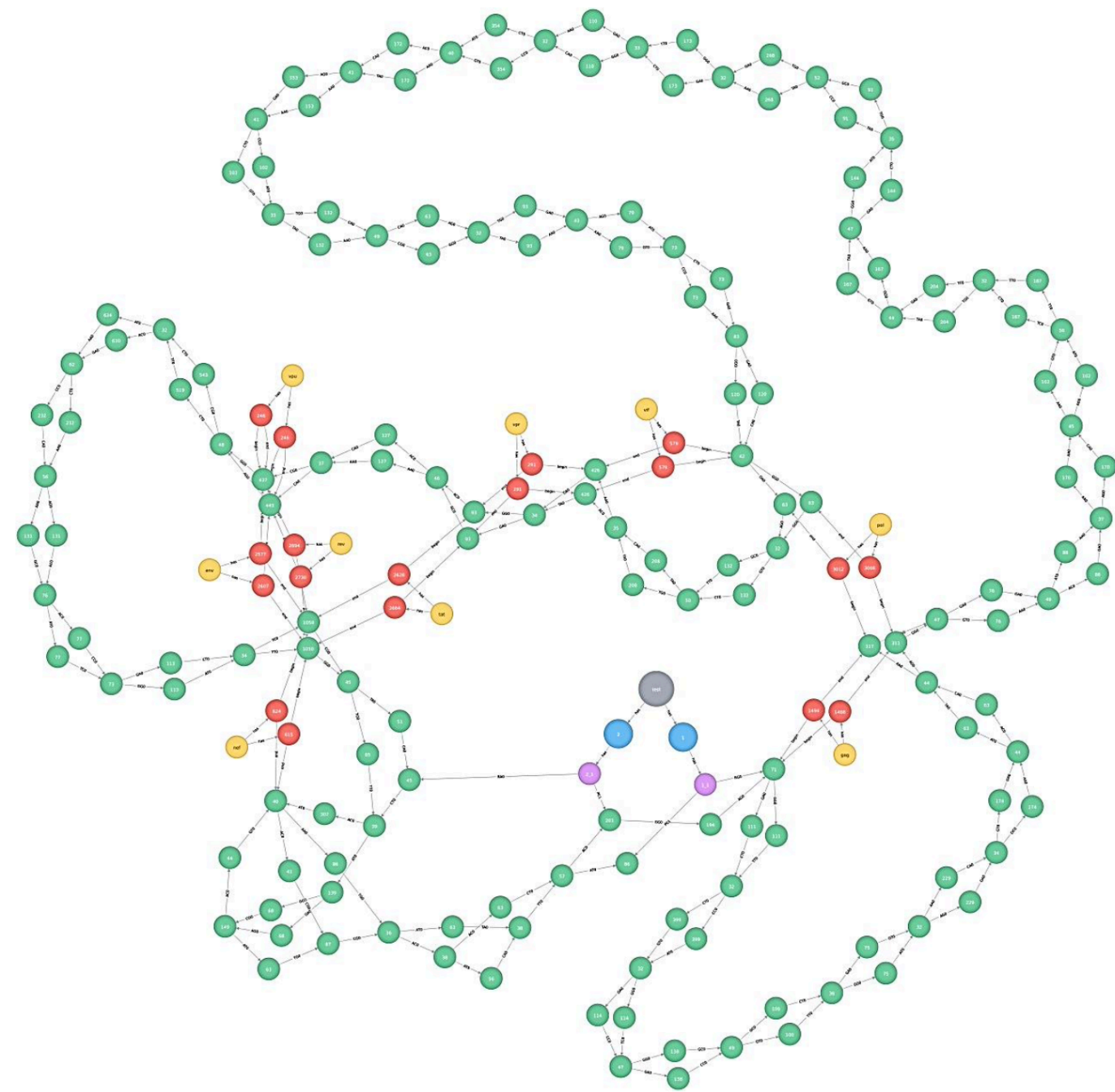
3. Pangenome (graph) representations

- Linearizing the graph structure
- Existing solutions do not scale

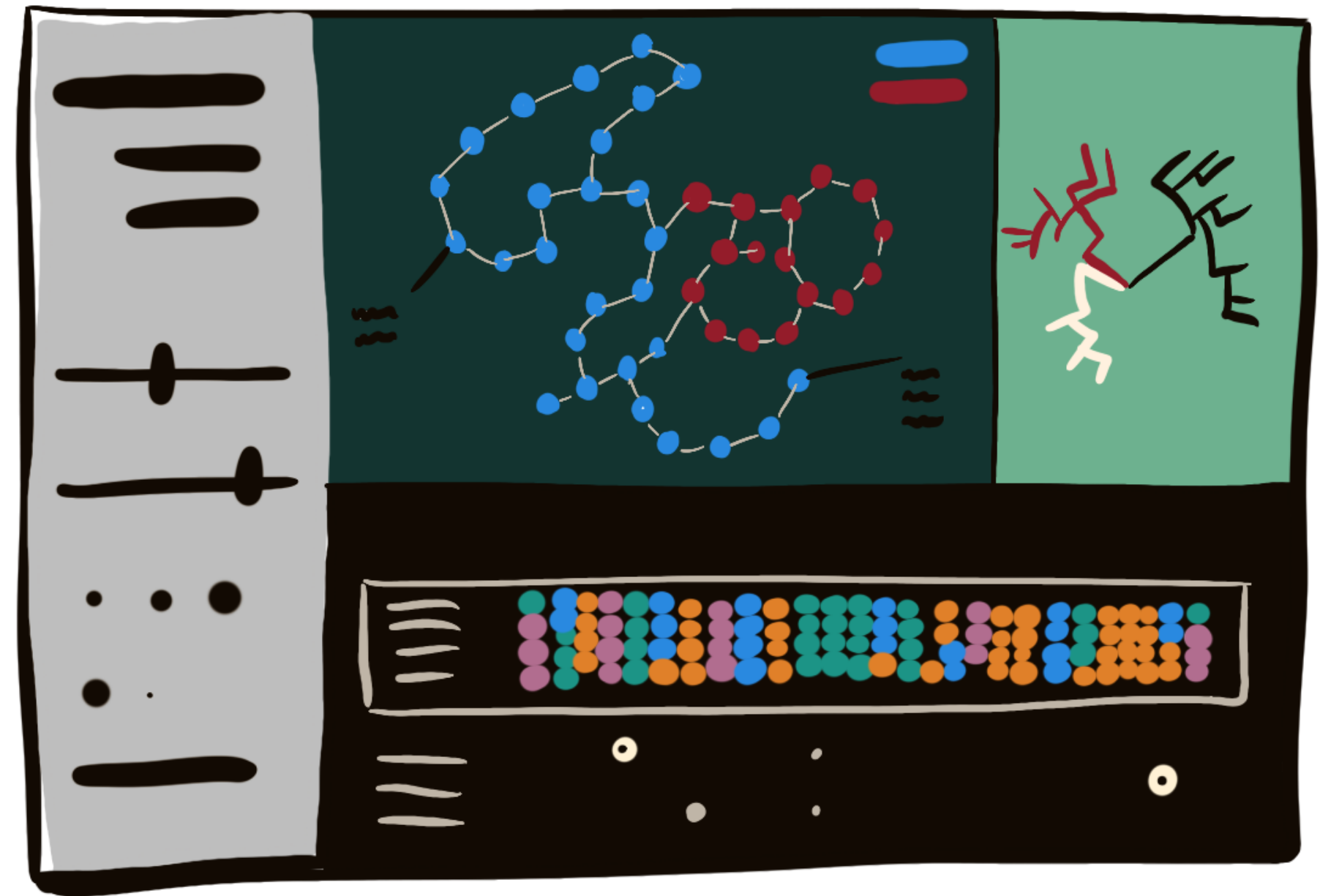
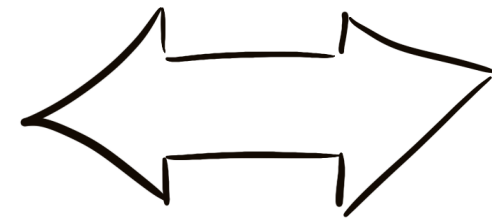


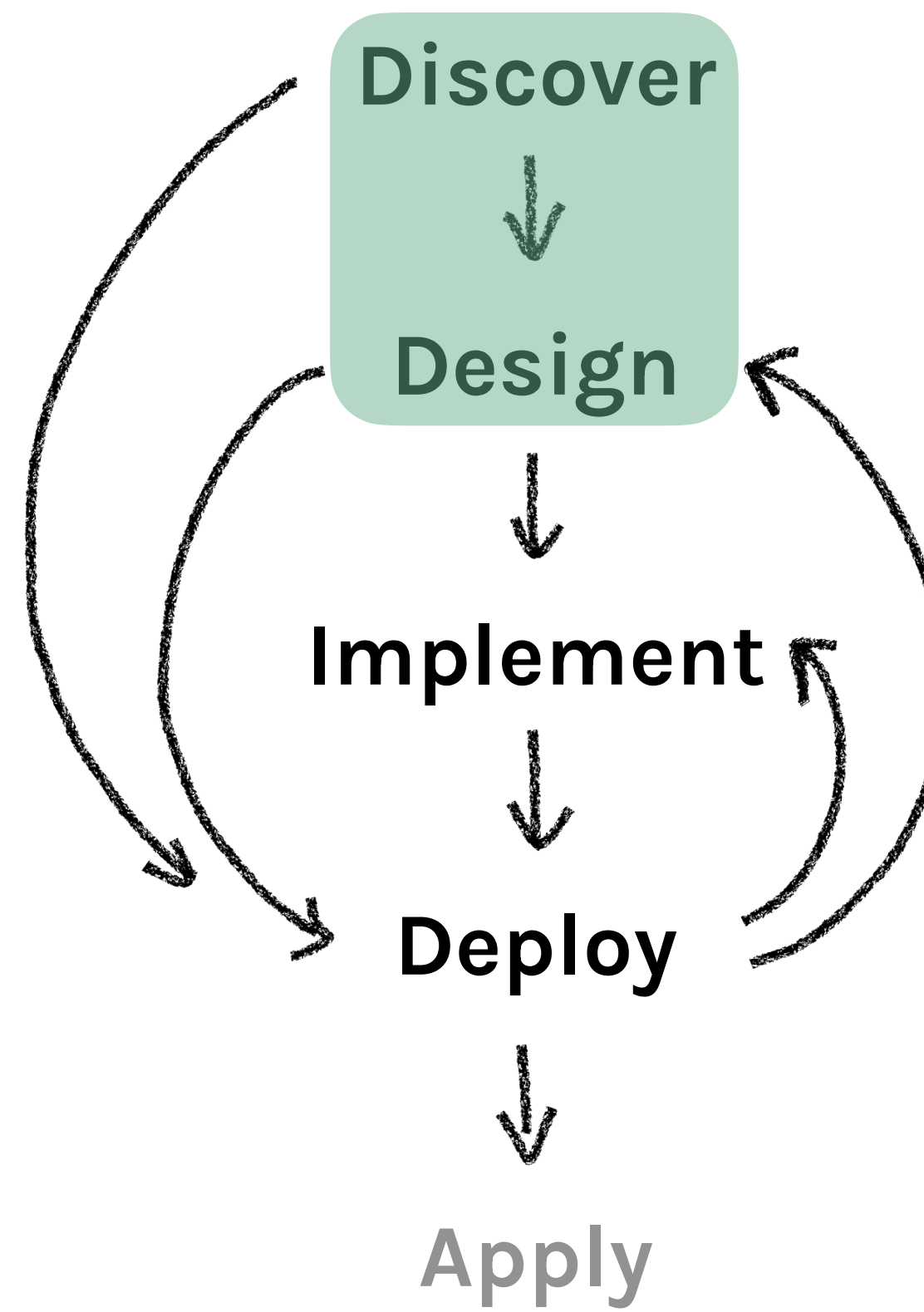
Pangenome constructed with PanTools and Neo4j

VA for Plant Pangenomics



Pangenome constructed with PanTools and Neo4j





Understanding the Target Domain

Task analysis:

- Interviews & shadowing / observation
- Nearly completed first round
 - Identify some analysis tasks
 - Understand what is important to represent

Understanding the Target Domain

Analysis of variation across 2 levels of organisation and resolution:

1. Small-scale variation (UC1): sequence - nucleotide level, markers
2. Large-scale variation (UC2): gene level and higher (structural)

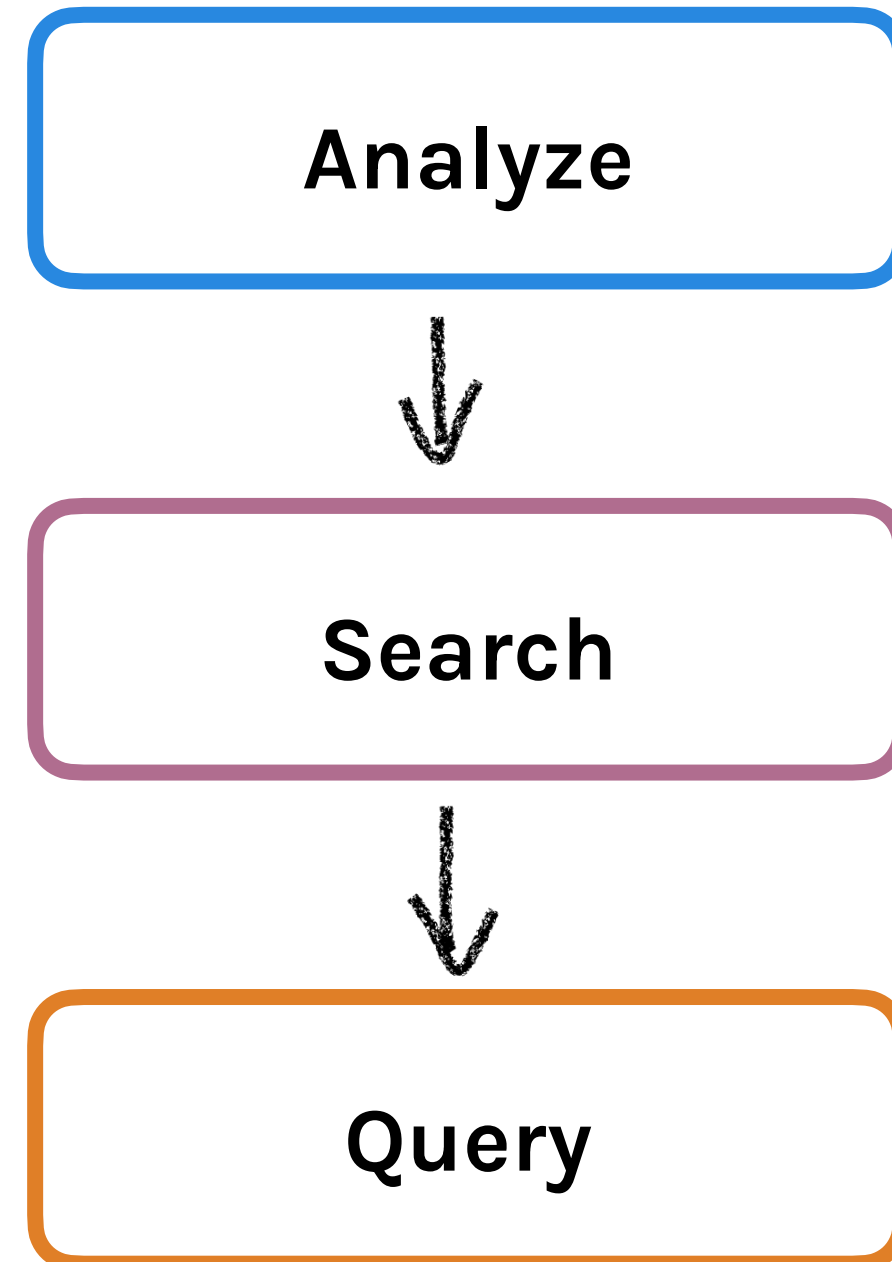
UC1: Sequence Variation

Domain specific tasks & questions

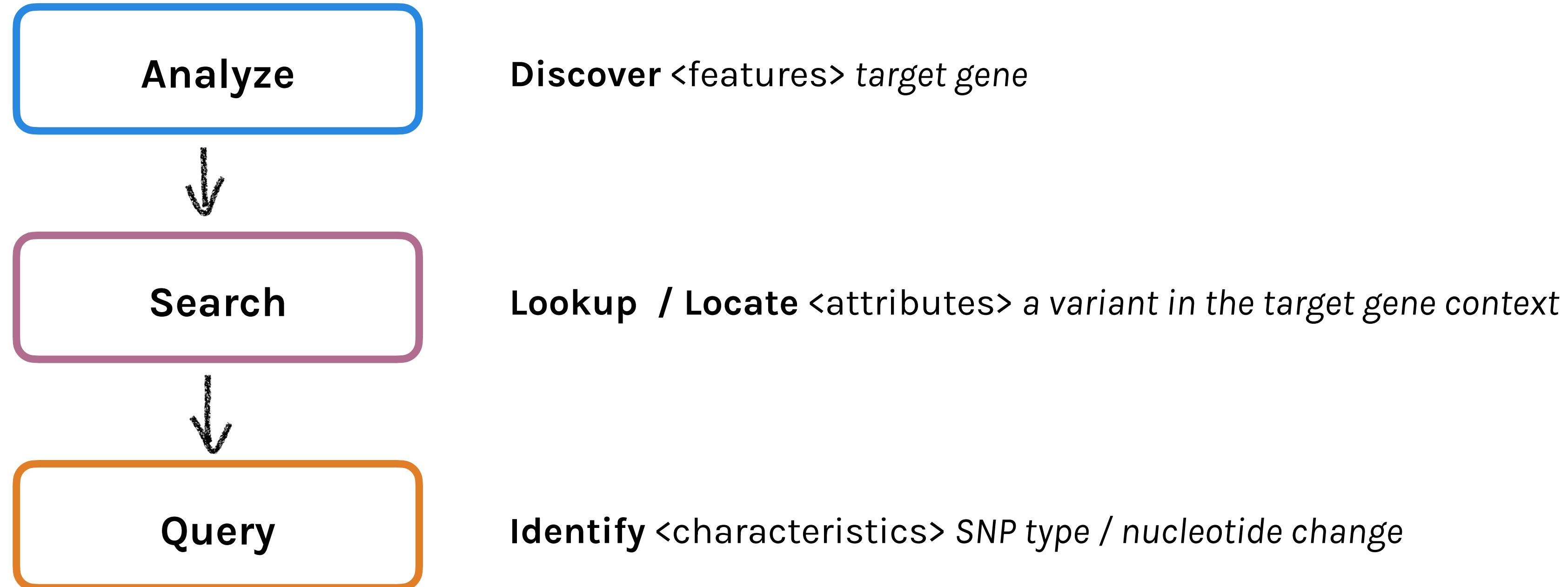


- Assess allelic variation within a **target gene**, possibly by phenotype
- Assess the genetic variation within a gene or **genomic region** in 1000s of strains
- Efficiently identify variants with a desired haplotype and their evolutionary relationships

Task Abstraction

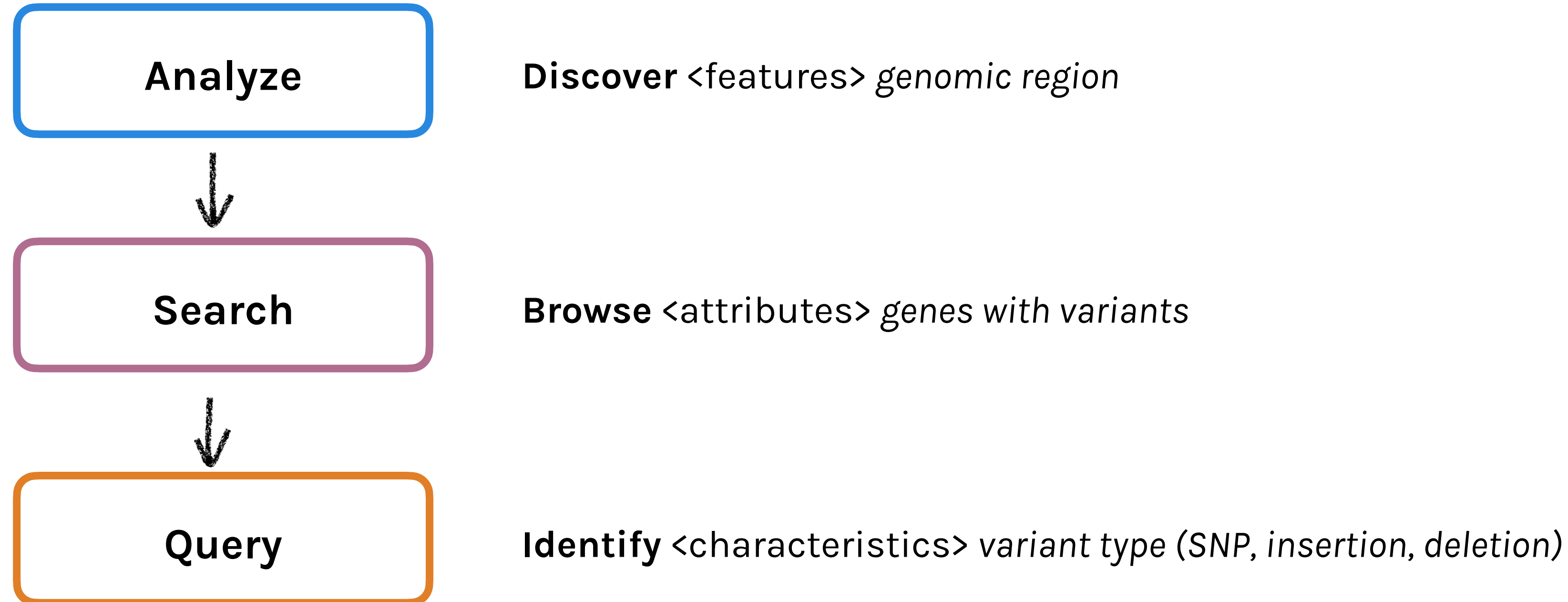


Task Abstraction (UC1)



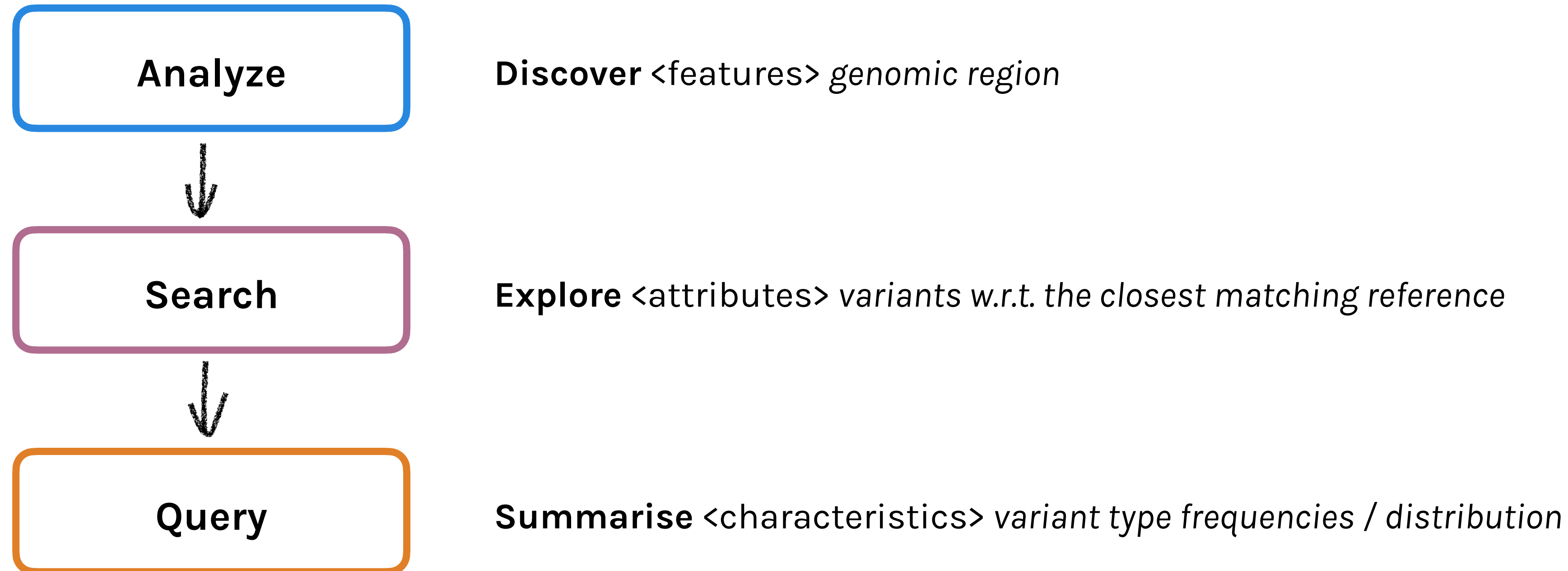
Q: Considering a target gene in x accessions, where are variants and what are their types?

Task Abstraction (UC1)



Q: Within a genomic region, are there genes with variants and what are the types?

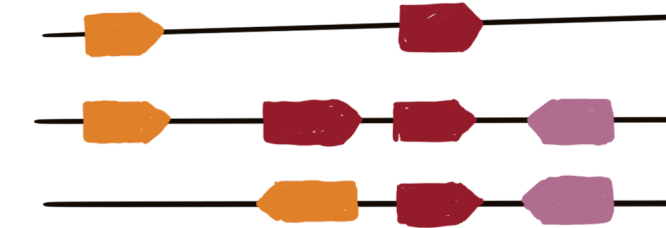
Task Abstraction (UC1)



Q: Within a genomic region, which variants are known for which genes and what is the closest reference?

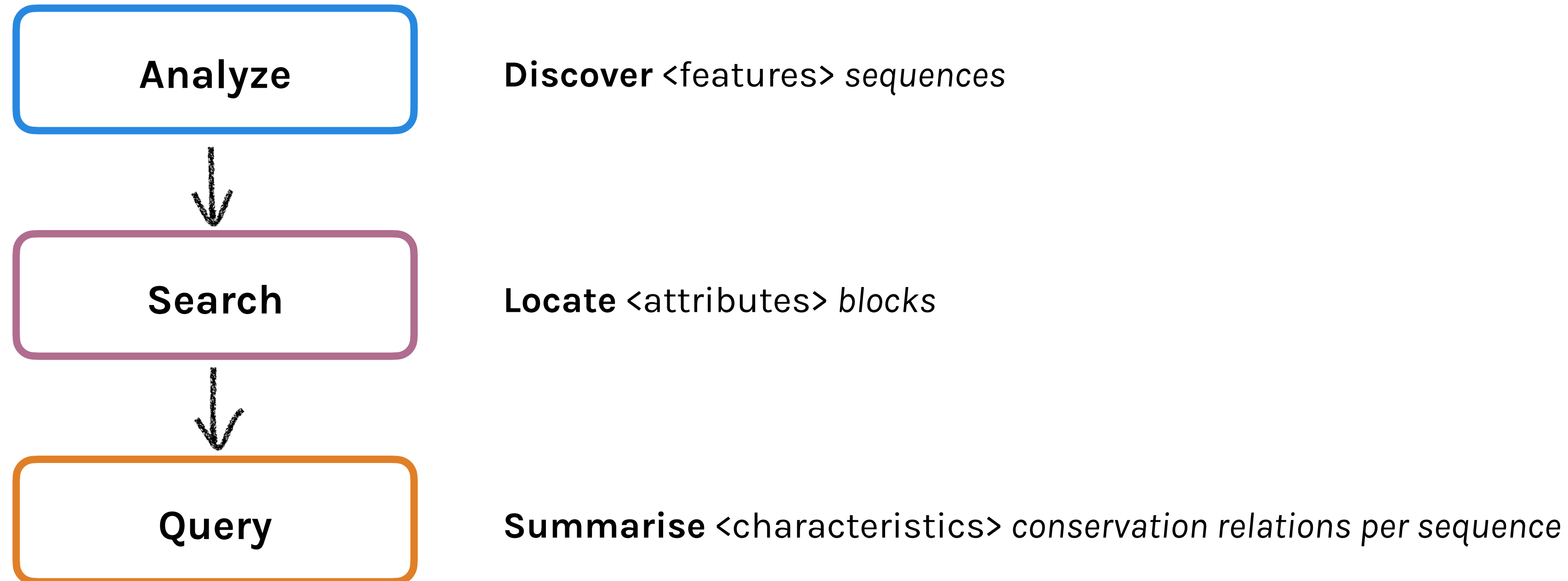
UC2: Structural Variation

Domain specific tasks & questions



- Show **collinearity / syntenic** to assess if recombination is likely to occur from wild
- Reveal organisation of resistance **genes** across multiple accessions
- Show presence/absence patterns of genes
- Support diagnosis of marker-segregation problems caused by **SV**

Task Abstraction (UC2)



Q: For one sequence, how many other sequences does it share syntenic blocks with?

Synteny (UC2)

→ Mizbee (Meyer et al., 2009)

	question	scale	relation
Q1	Which genomes share conserved blocks?	p, g	p
Q2	For one genome, how many other genomes does it share blocks with?	p, g	p
Q3	Which chromosomes share conserved blocks? In one genome? Between genomes?	g, c	p
Q4	For one chromosome, how many other chromosomes does it share blocks with?	g, c	p
Q5	Where are the blocks: on genomes? Around a specific location on a genome?	p, g	p
Q6	Where are the blocks: on chromosomes? Around a specific location on a chromosome?	g, c	p
Q7	What are sizes and locations of other genomic features near a block?	c	p, z
Q10	Are orientations forward or reverse for block pairs? gene pairs?	c, b	o
Q11	Do neighbouring blocks have matched orientations?	c	o
Q12	Are scores similar within a block? Between neighbouring blocks? Between genomes?	p, g, c, b	s
Q13	How large is a gene relative to other genes/features within a block?	b	z
Q14	What are size, name and location of genes in a block?	b	p, z

Task Abstraction (UC2)

Analyze

Discover <features> *genomic region*

Search

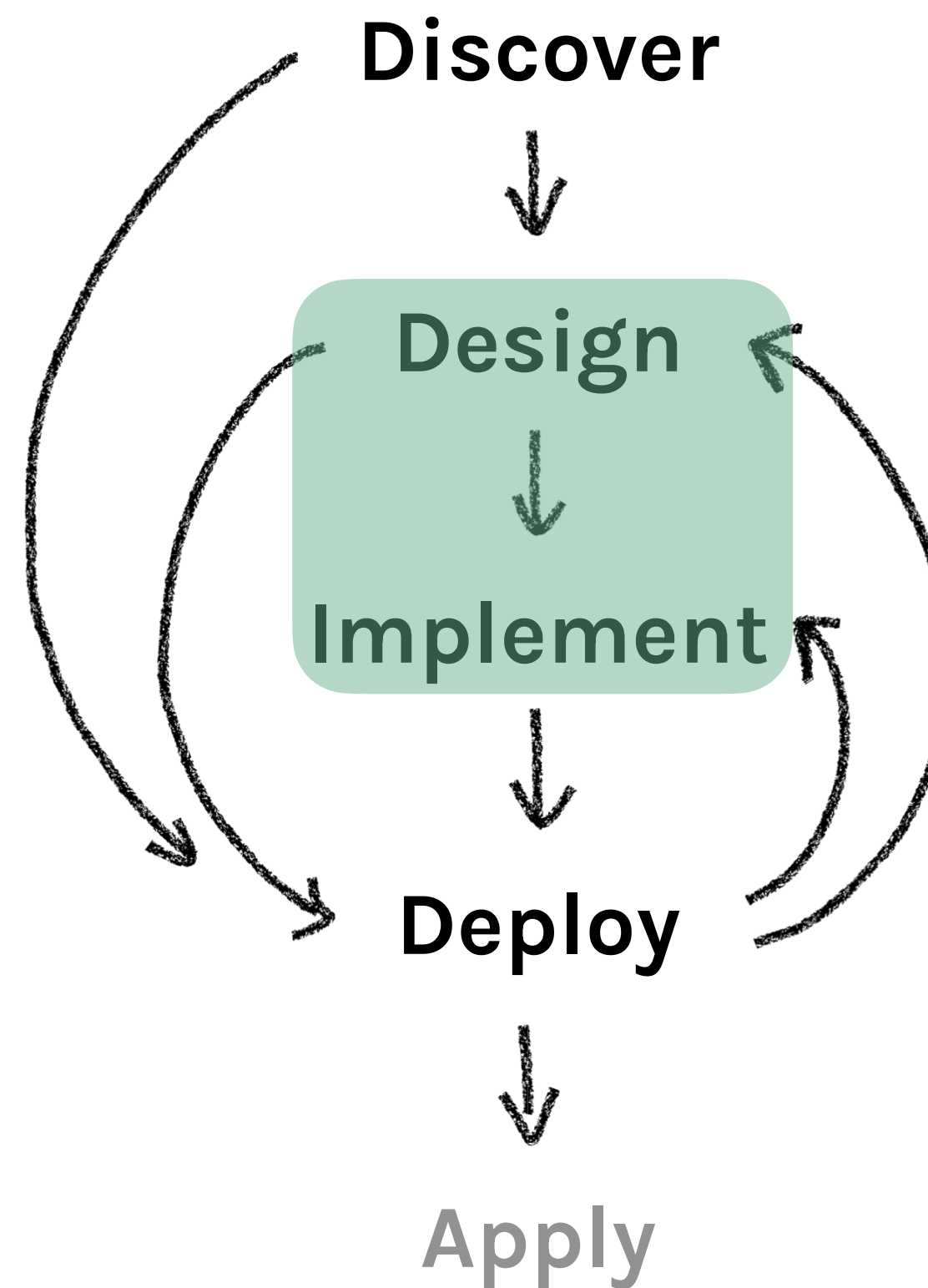
Explore <attributes> *organisation of homologous genes*

Query

Explore <characteristics> *distribution and similarity*

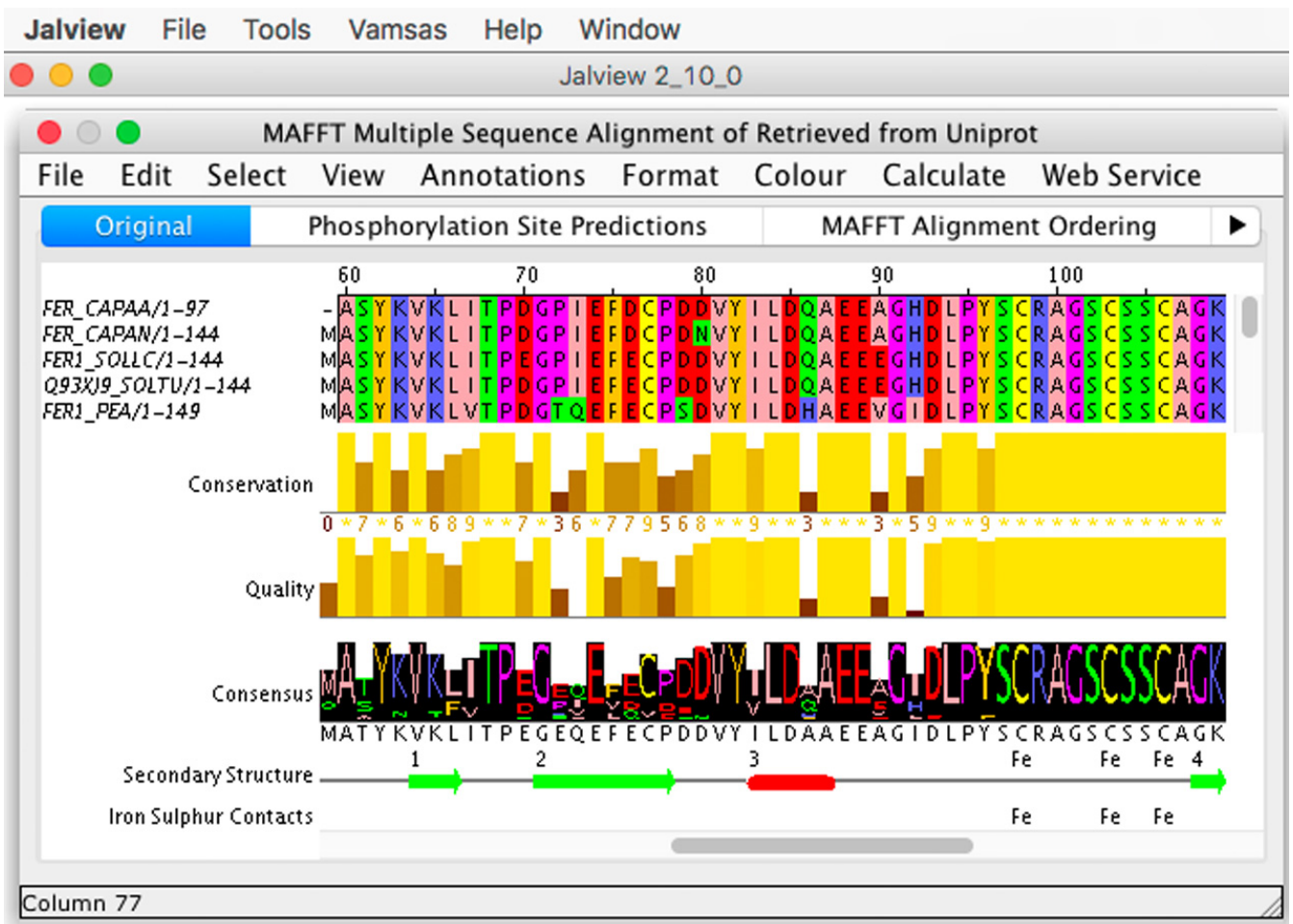
Q: Considering x accessions, how are homologous genes organized/ordered? What are functions of neighboring genes?

Design Process

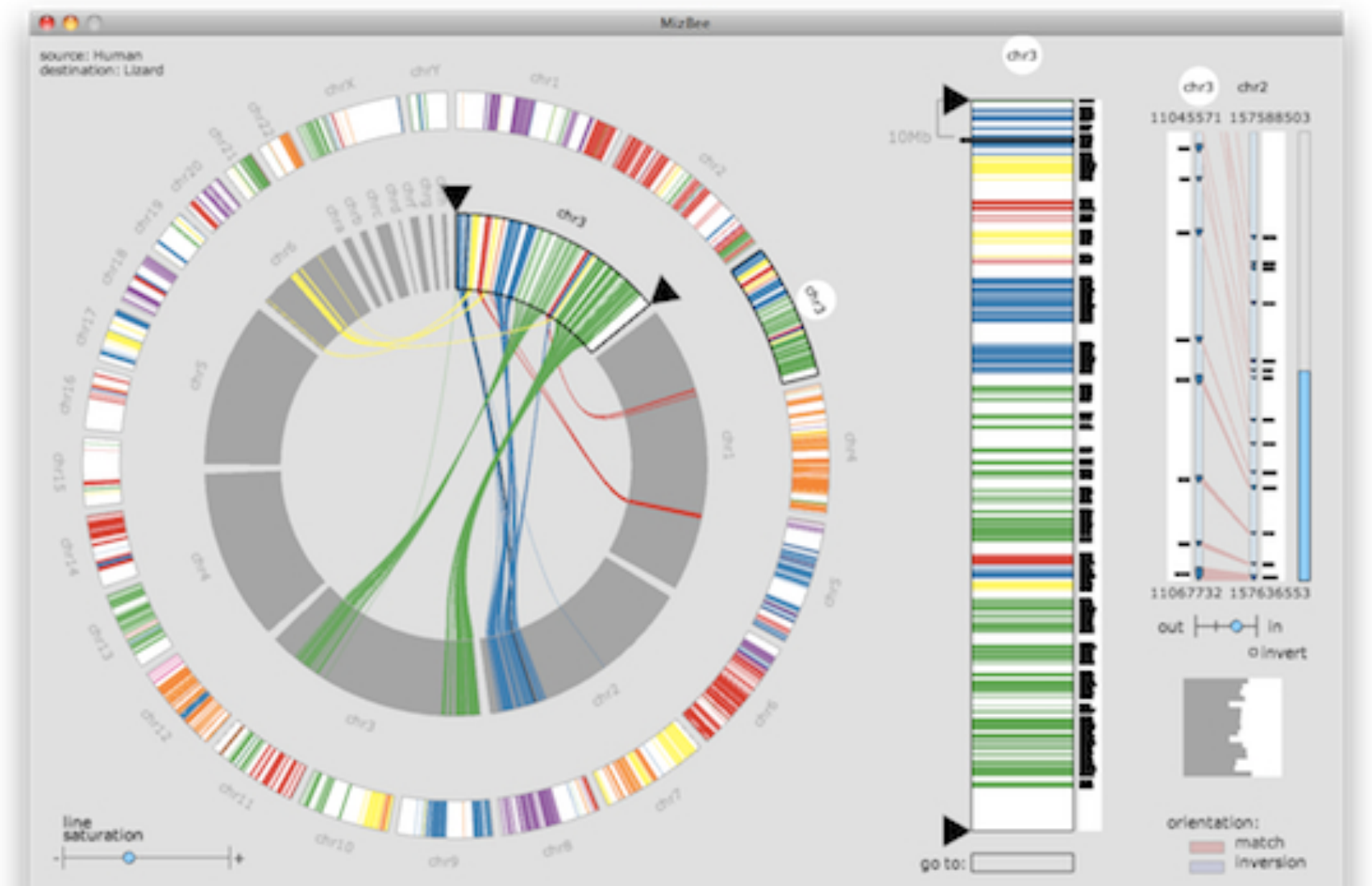


Design study (Sedlmair et al., 2012)

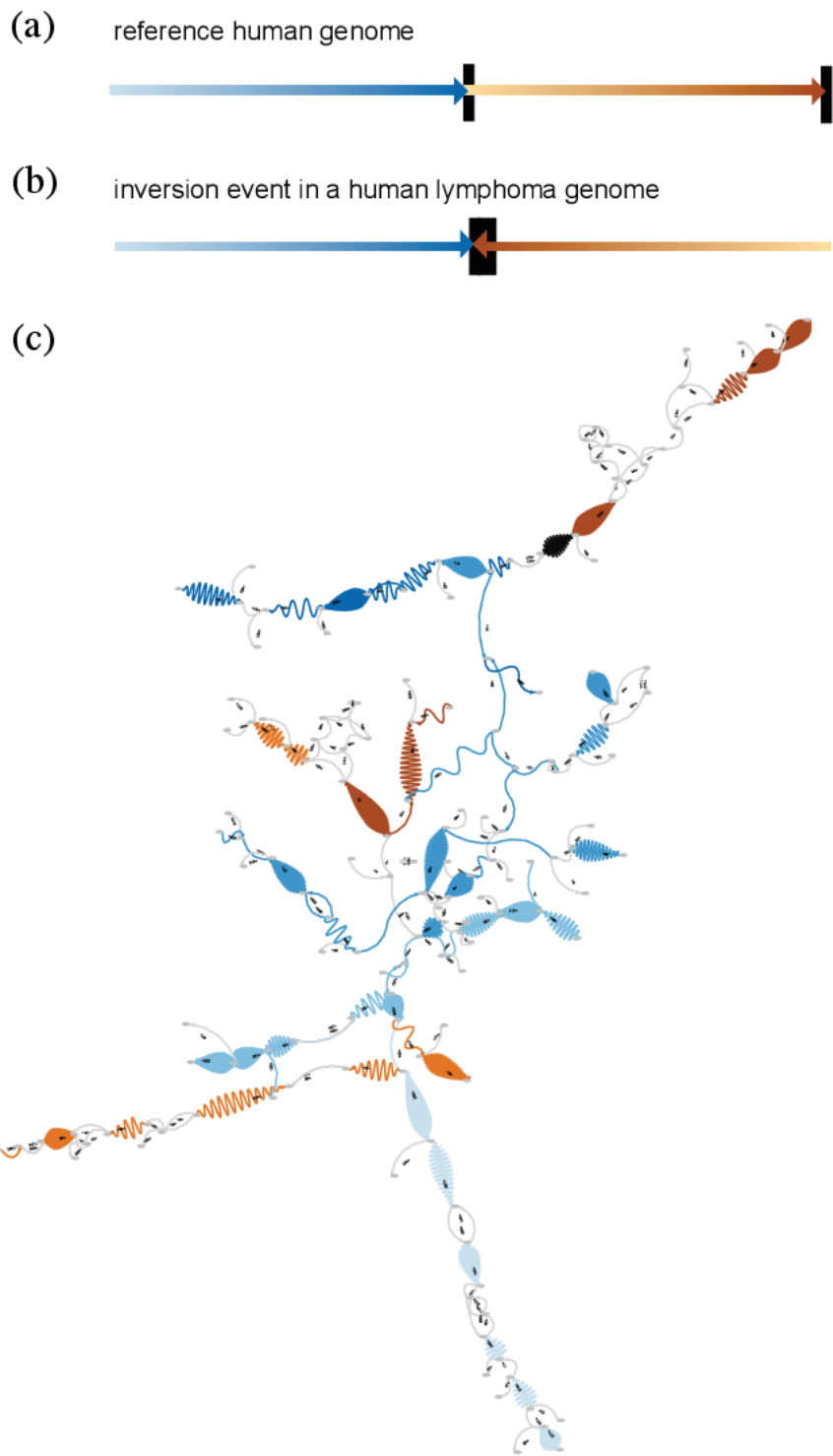
Related Work



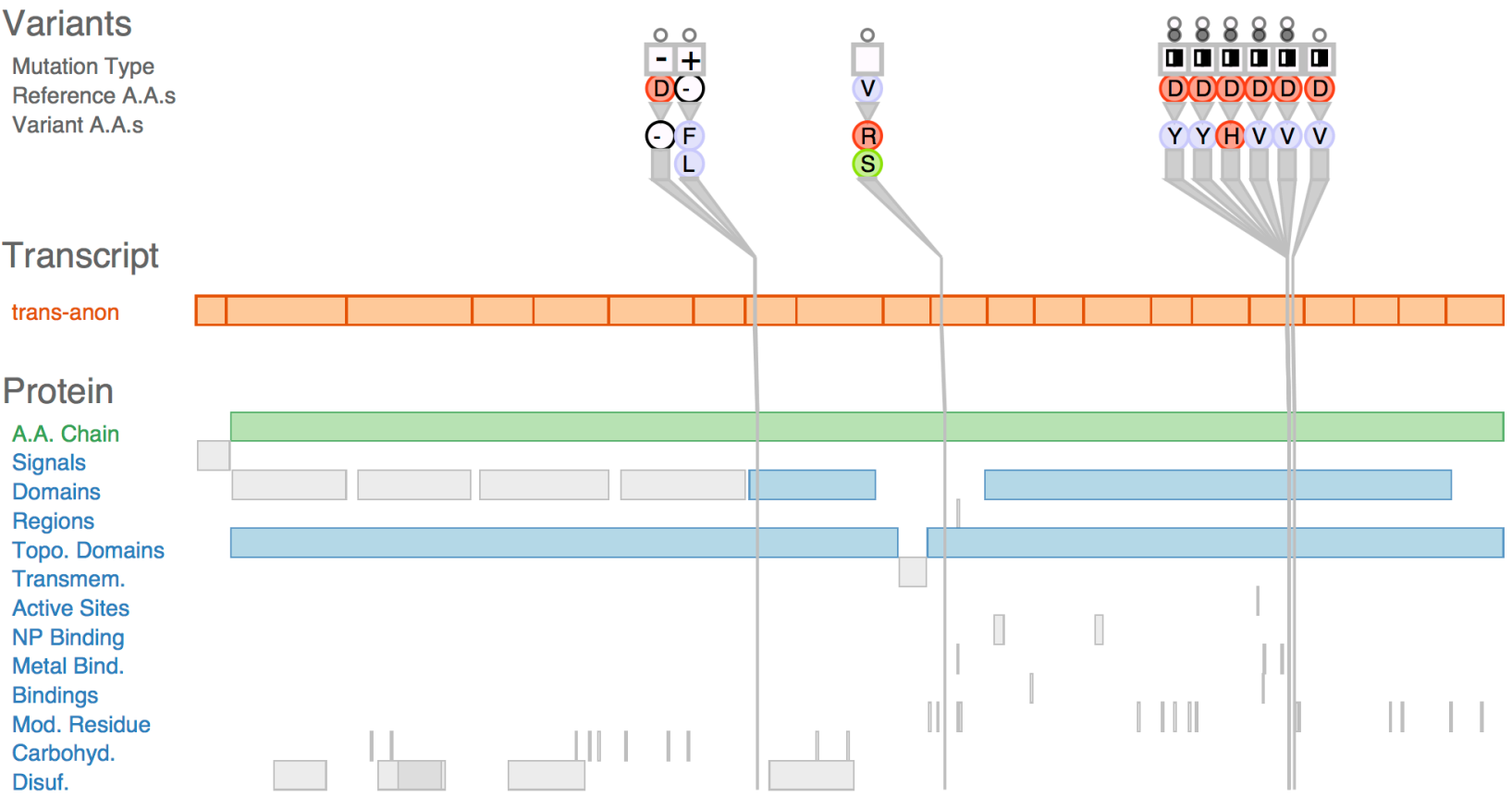
Jalview



Mizbee



ABYSS-Explorer

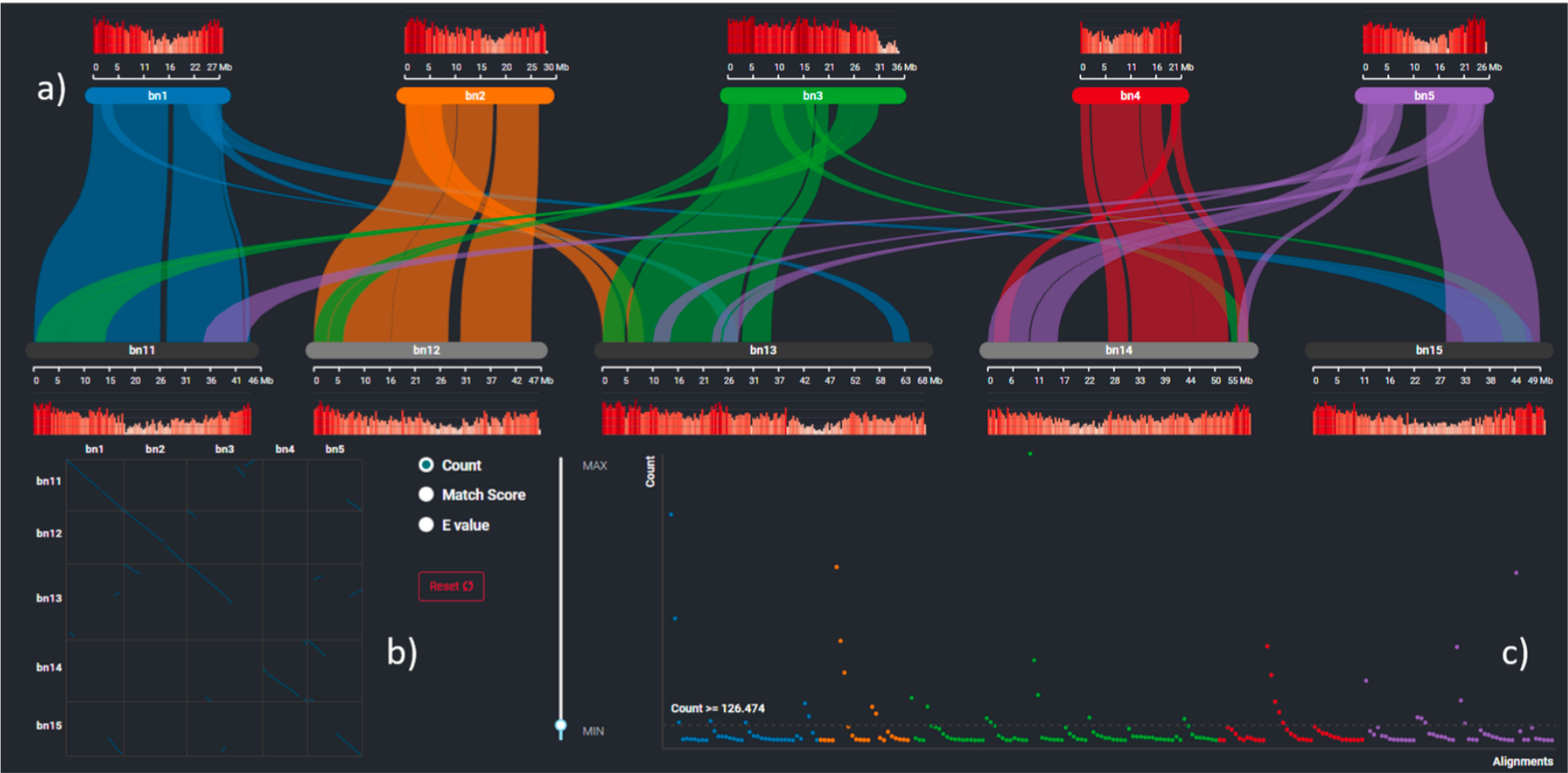


Variant-View

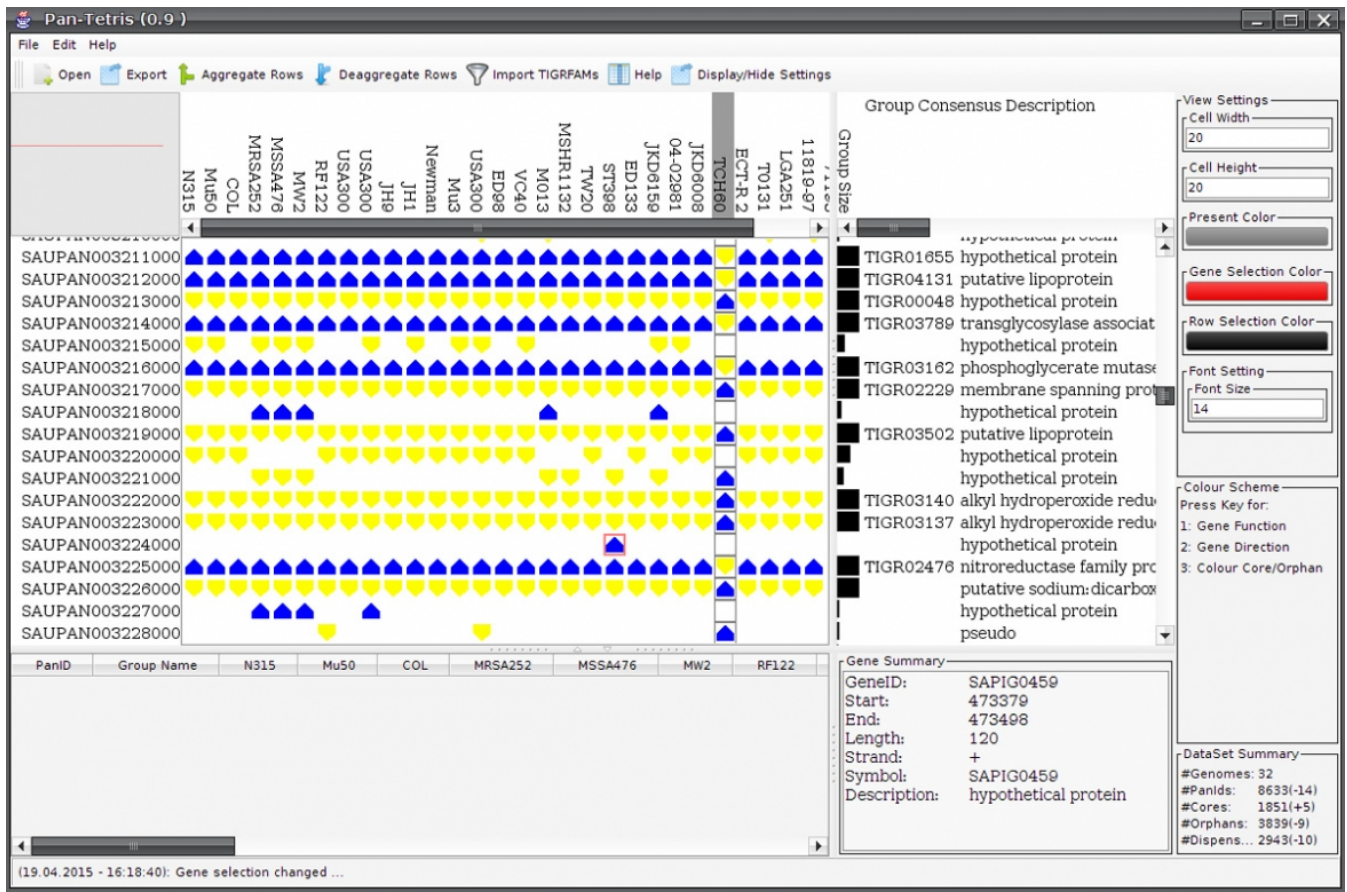


IGV

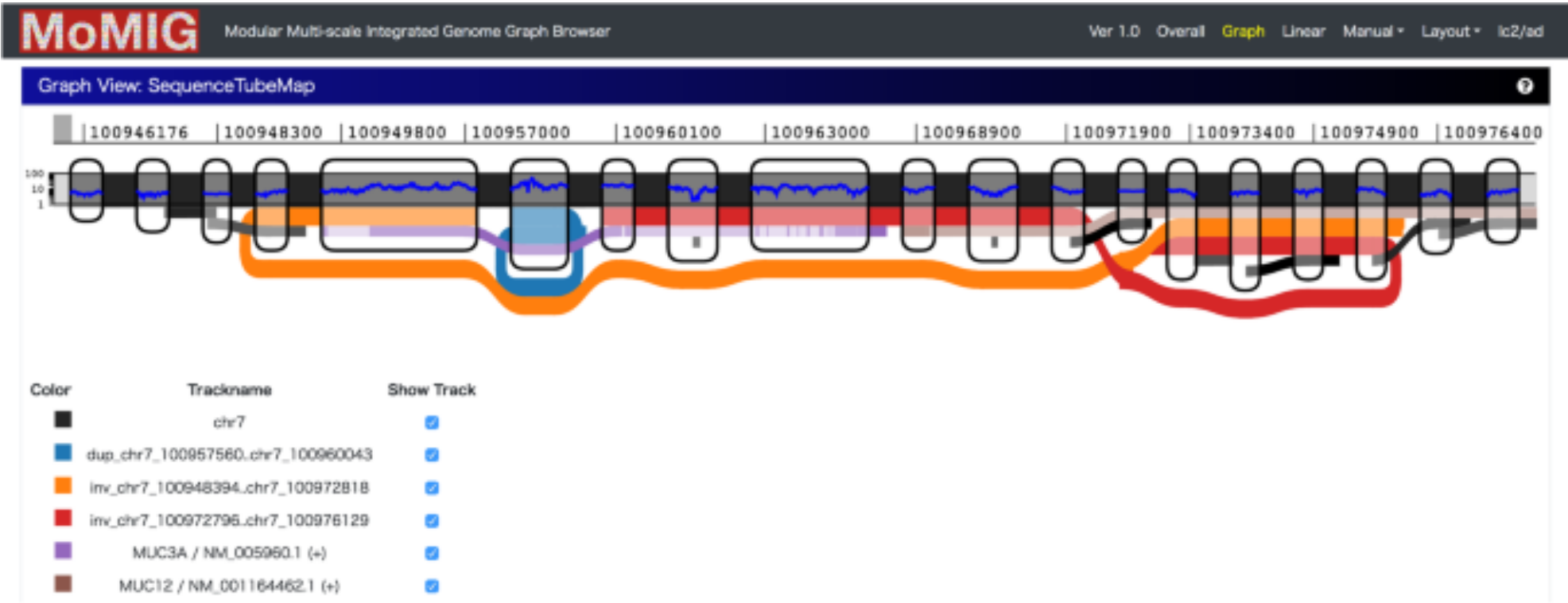
Related Work



SynVisio



PanTetris



Sequence Tube Maps



PanX

Problems Existing Tools

- Cannot visualize much at the same time
- Switching between tools & recalculation
- Phenotype relations and grouping not visible

Visualization Challenges

1. Large number of samples
2. Genomic features are small and sparse
- 3. Navigation between levels of detail**
4. Comparison n-to-n relations
5. Large number of datatypes

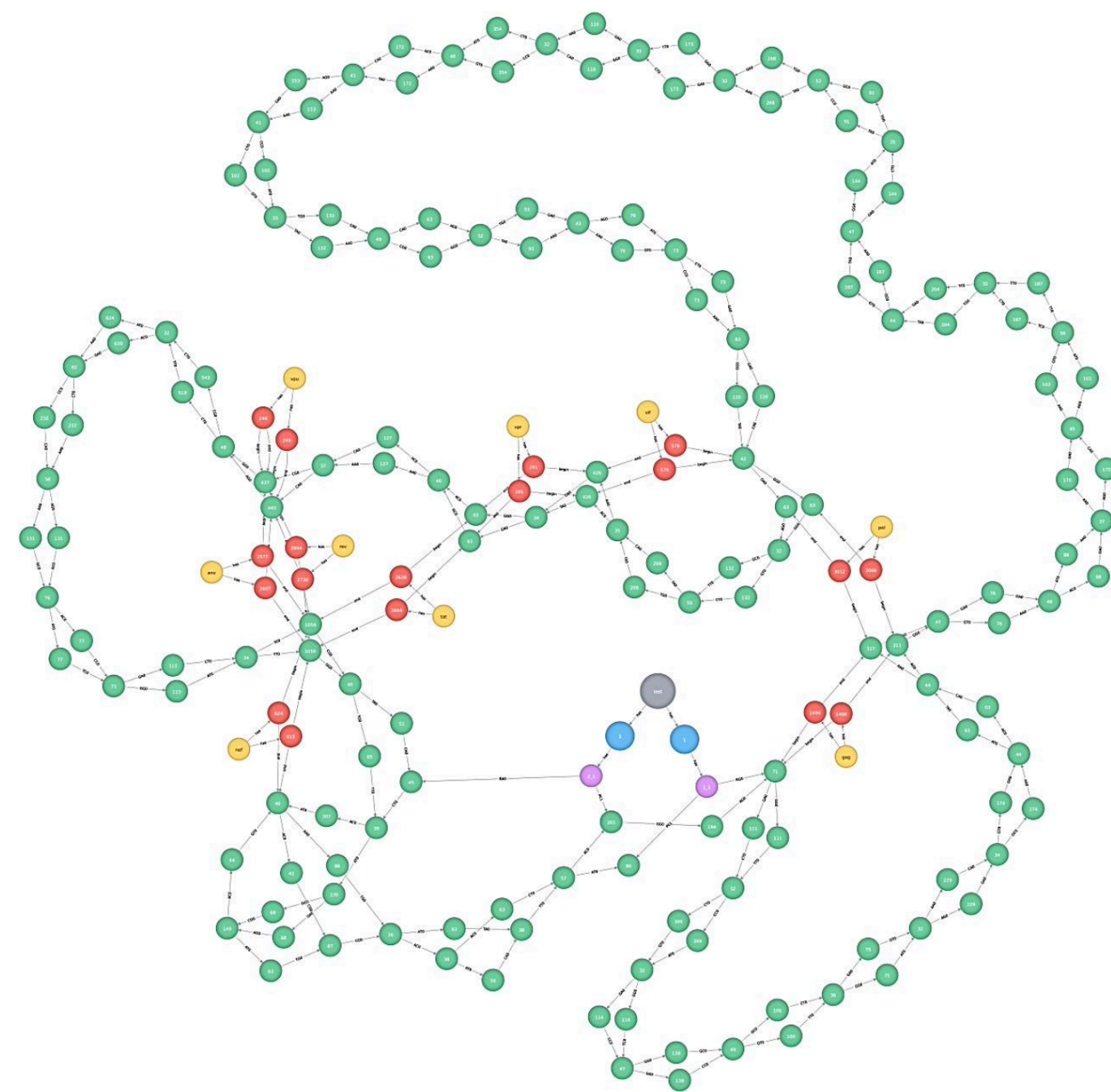


Addressing Issues of Scale

1. **Scan sequentially:** the user will examine items serially;
2. **Select subset:** the user will examine a smaller set of items;
3. **Summarise:** the user will examine an abstraction that concisely describes the items

(Gleicher, 2018)

PanTools Backend

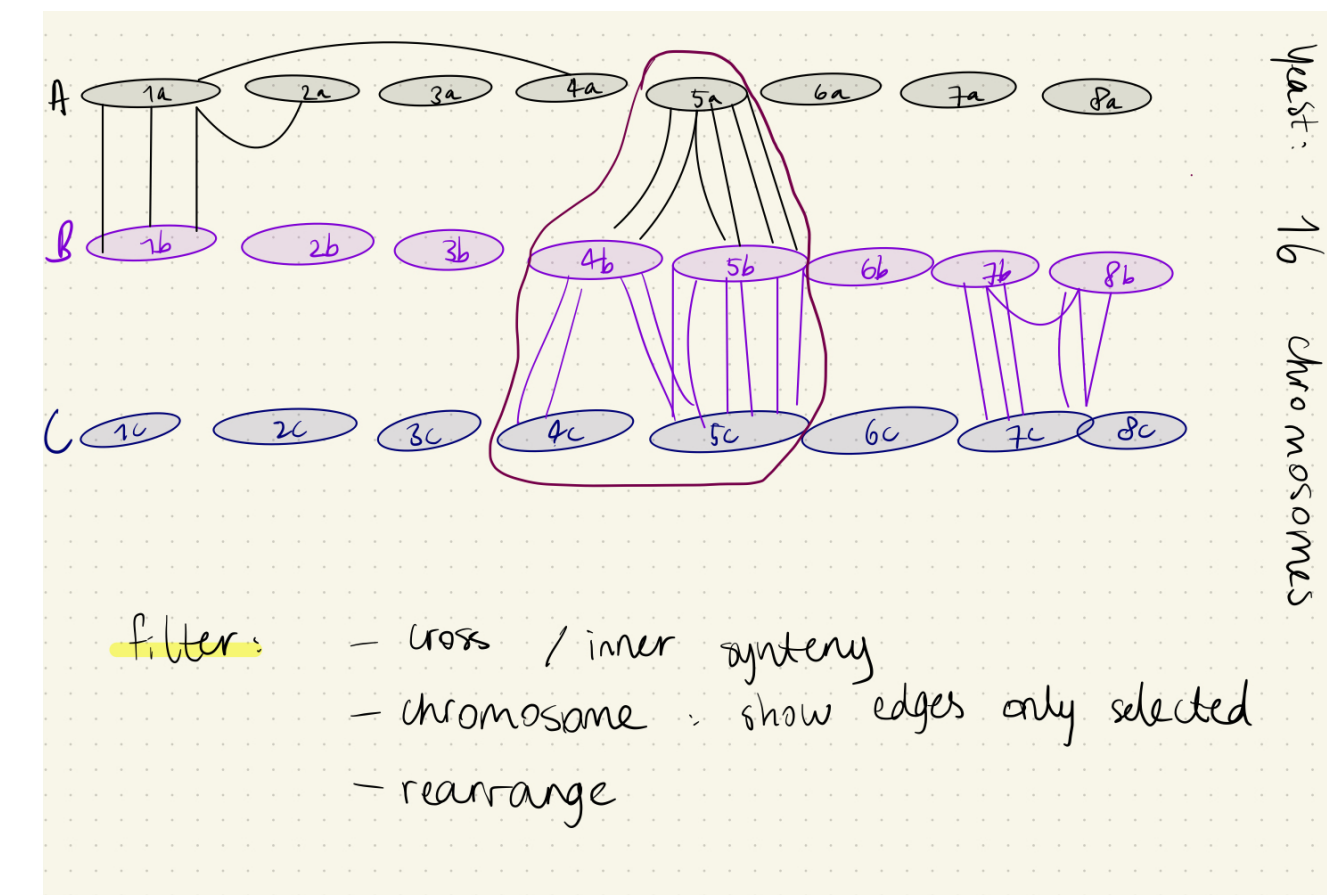
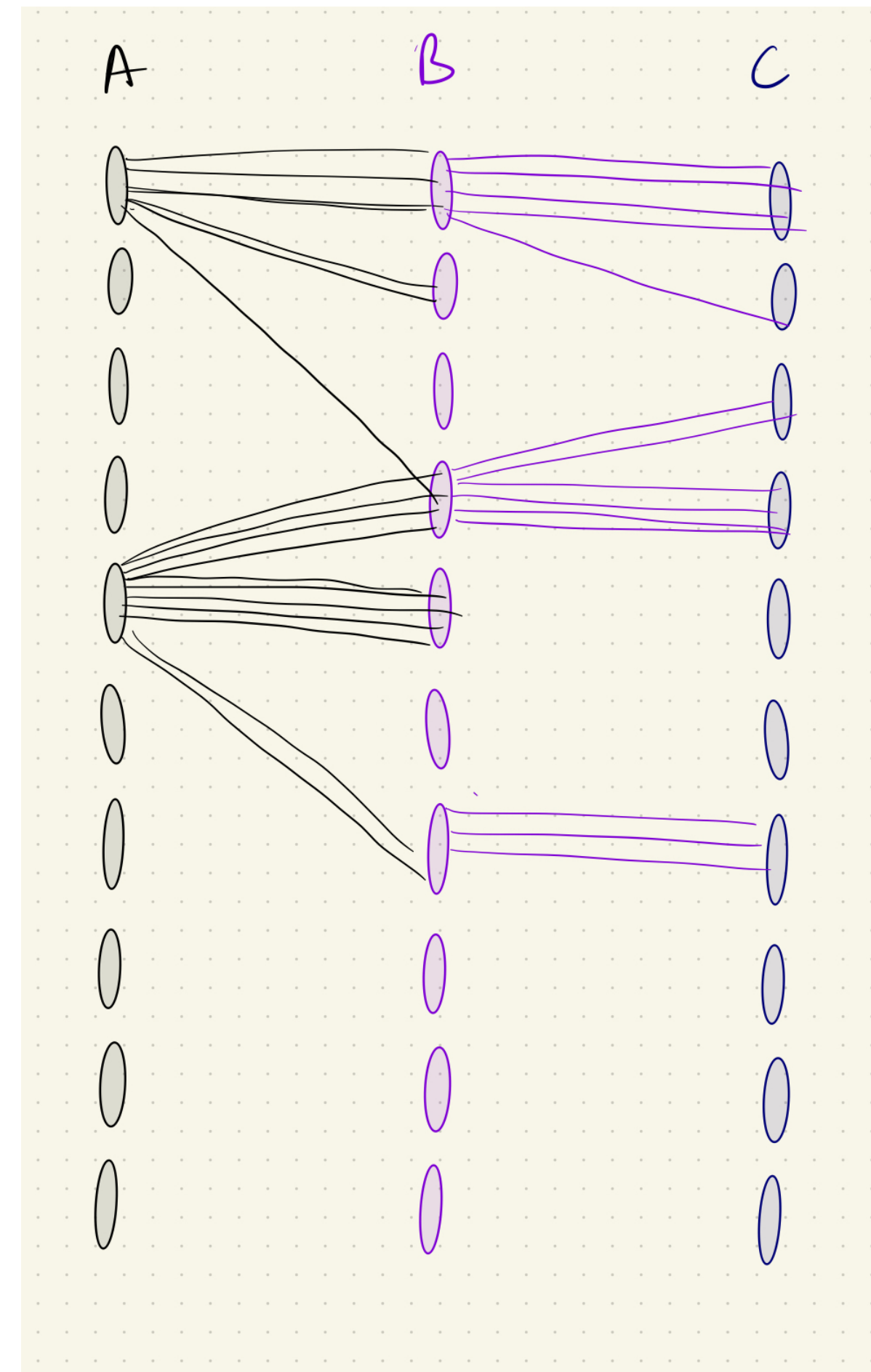
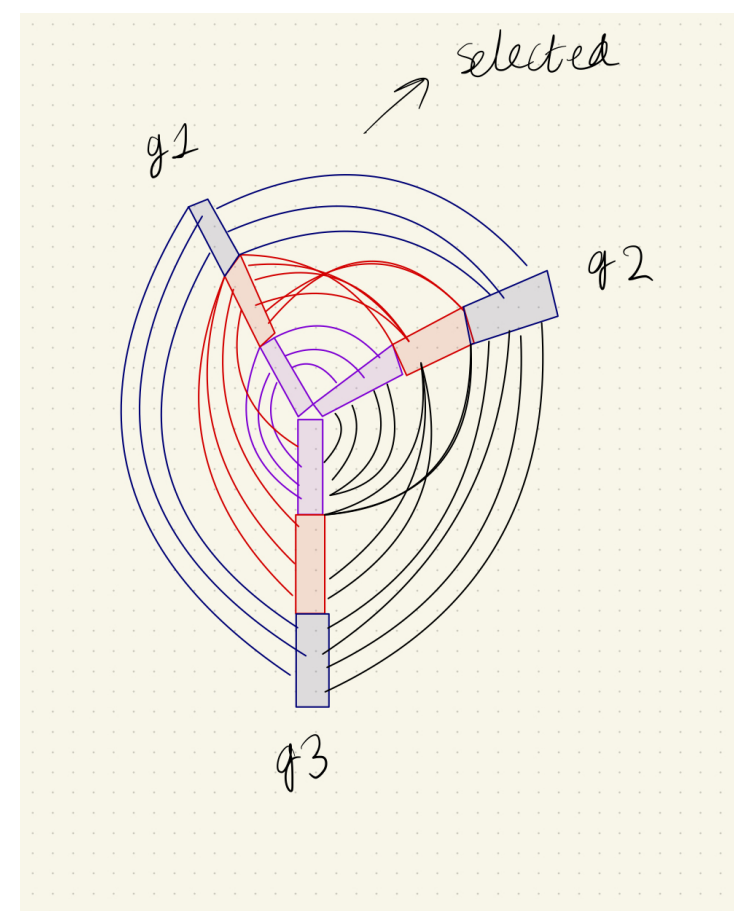
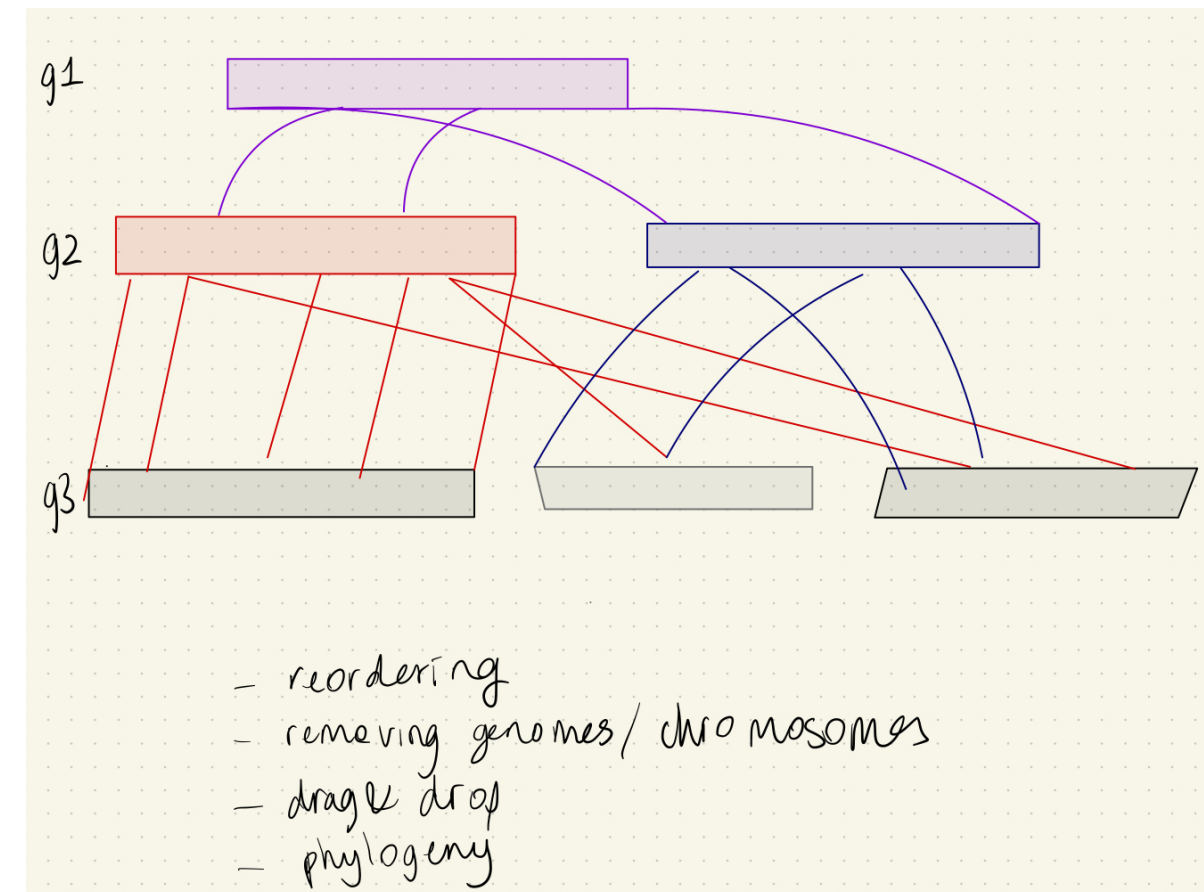


Pangenome constructed with PanTools and Neo4j

Discussing output specification

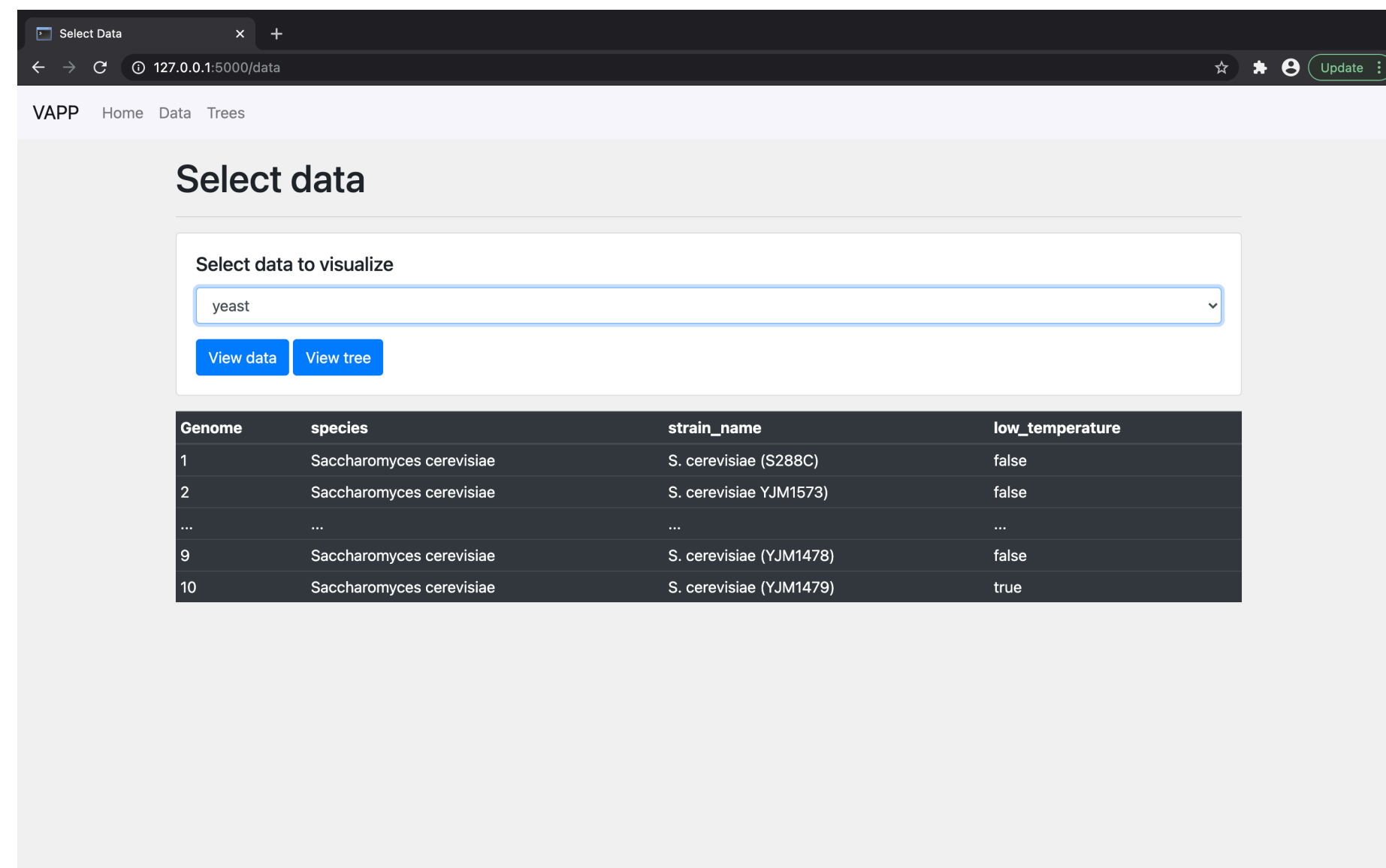
- File format
- Indexing
- Coordinates
- Annotations

Paper Prototypes (Synteny)



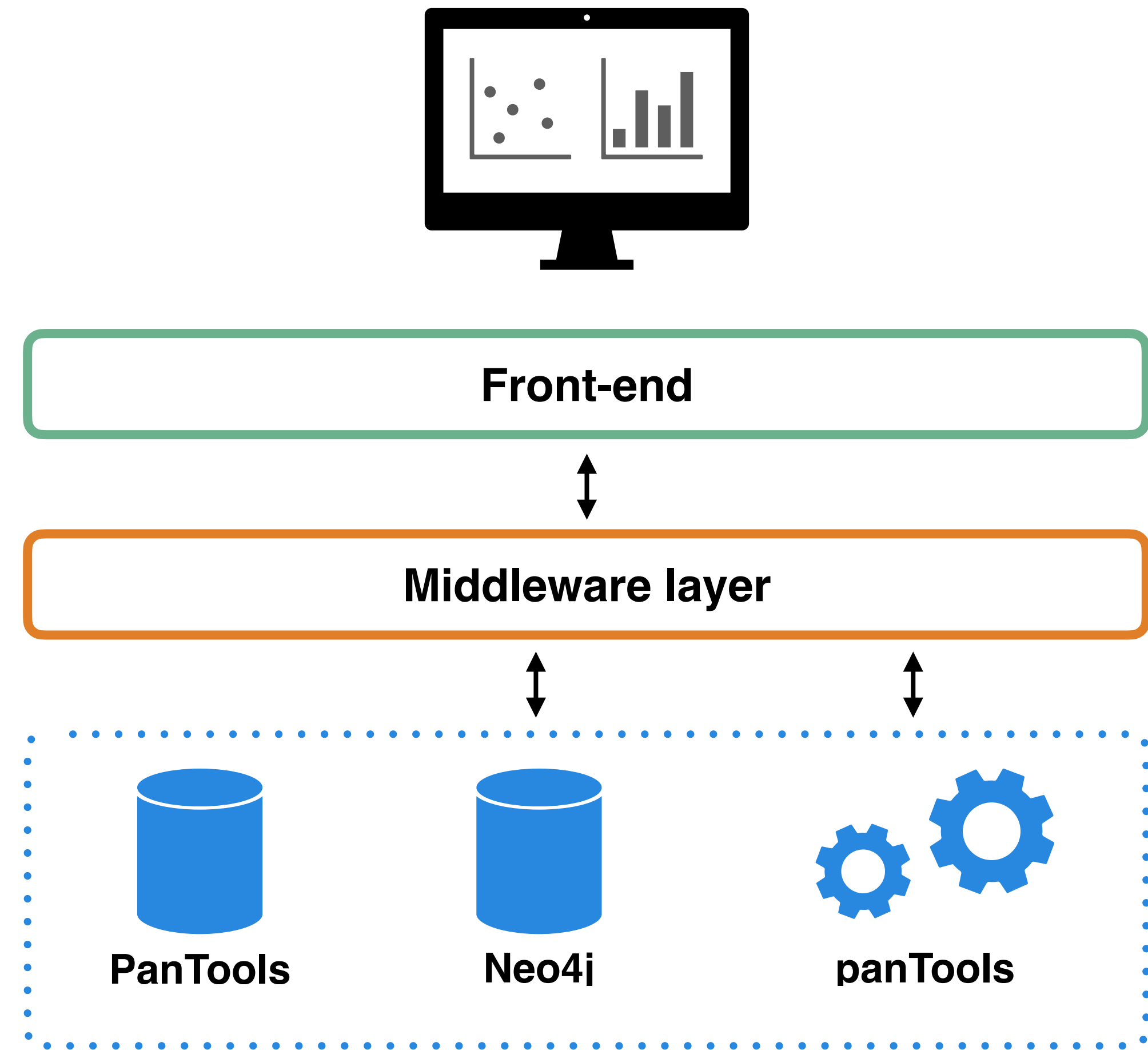
First Prototype

- Yeast data set
- Exploring data structure
- **Flask** + D3

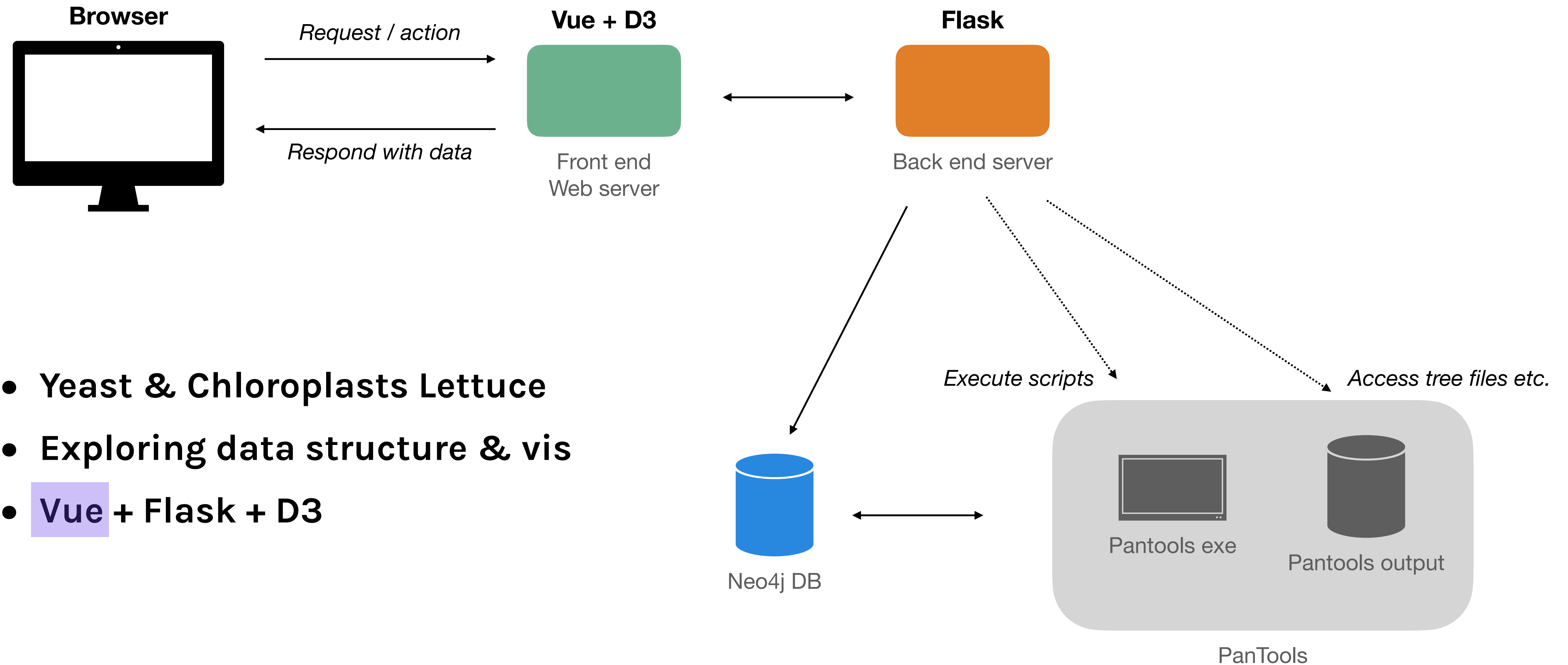


The screenshot shows a web browser window with the address bar displaying '127.0.0.1:5000/data'. The page title is 'Select Data'. Below the title, there is a search bar labeled 'Select data to visualize' with the word 'yeast' entered. Below the search bar are two buttons: 'View data' and 'View tree'. Below these buttons is a table with four columns: 'Genome', 'species', 'strain_name', and 'low_temperature'. The table contains five rows of data, with the first row showing '1', 'Saccharomyces cerevisiae', 'S. cerevisiae (S288C)', and 'false'.

Genome	species	strain_name	low_temperature
1	Saccharomyces cerevisiae	S. cerevisiae (S288C)	false
2	Saccharomyces cerevisiae	S. cerevisiae YJM1573	false
...
9	Saccharomyces cerevisiae	S. cerevisiae (YJM1478)	false
10	Saccharomyces cerevisiae	S. cerevisiae (YJM1479)	true

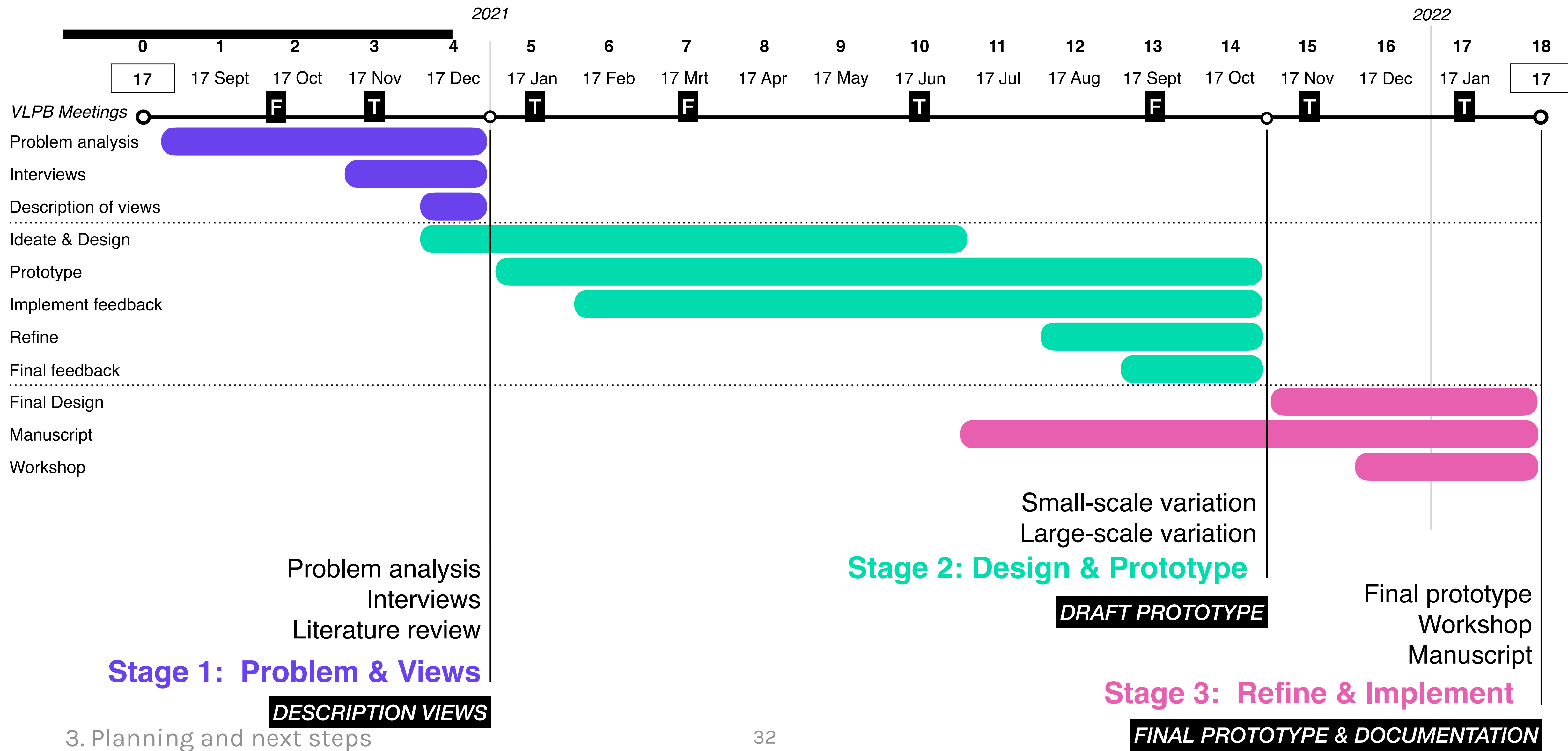


Second Prototype



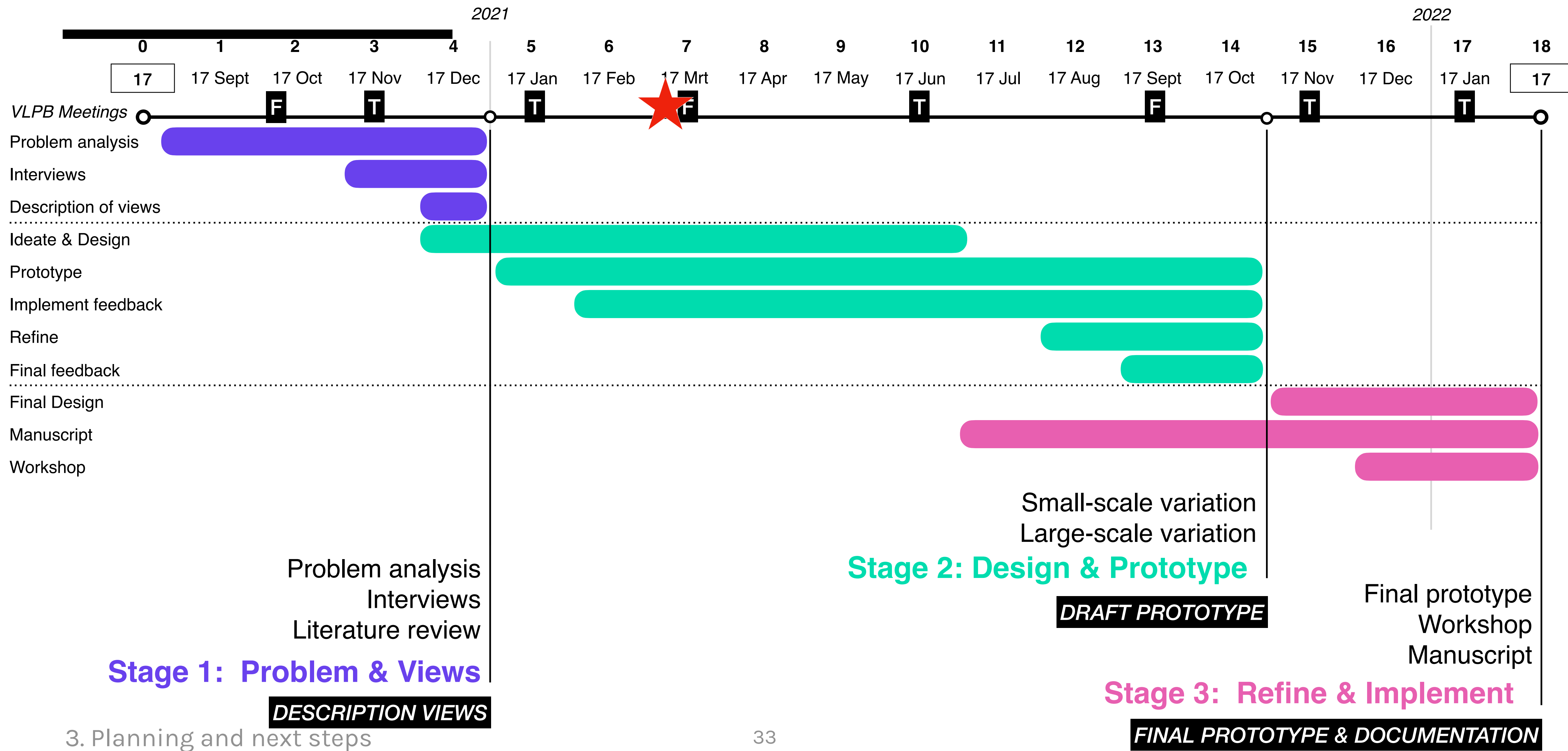
- Yeast & Chloroplasts Lettuce
- Exploring data structure & vis
- **Vue** + Flask + D3

Timeline



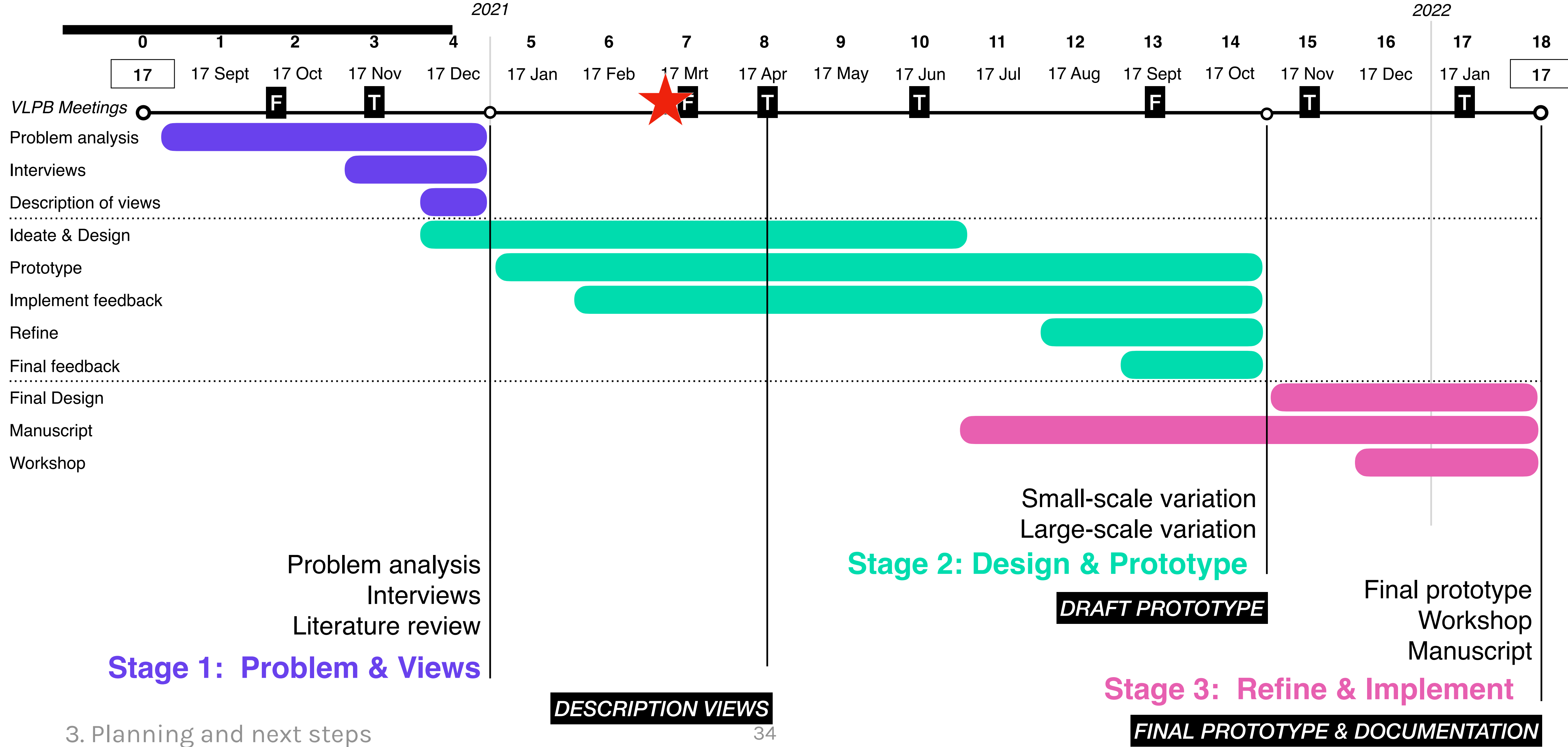
Timeline

We are here!



Timeline

We are here!



Next Steps

1. Complete and refine task abstraction for UC1 and UC2
2. Design of views: mockups
3. Implement design in prototype
4. Feedback: first group and plenary



Questions?