Telecon 3

Visual Analytics for Plant Pangenomes (VAPP)

Astrid van den Brandt | Eindhoven University of Technology
Anna Vilanova | Eindhoven University of Technology
Sandra Smit | Wageningen University & Research

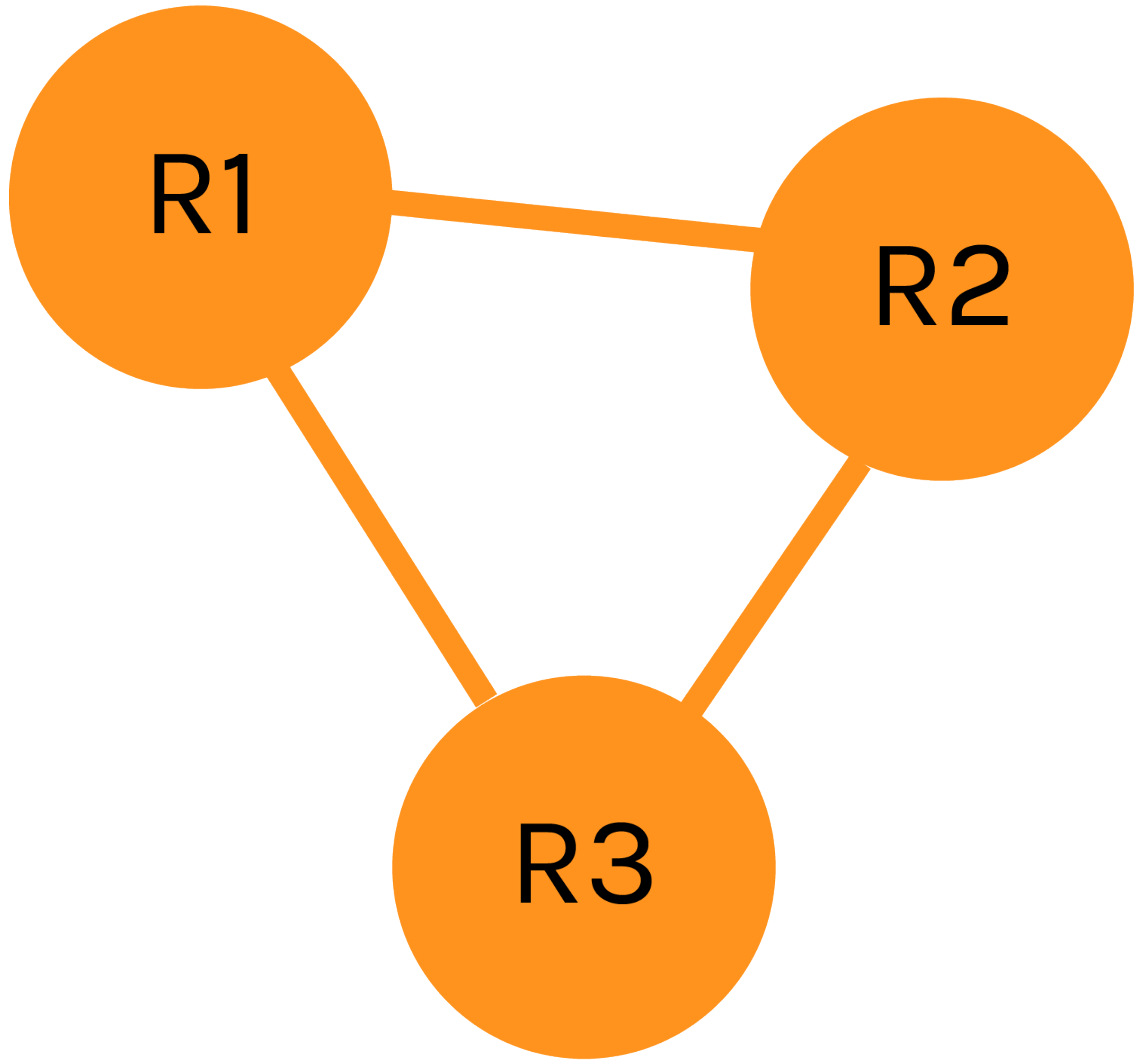
---

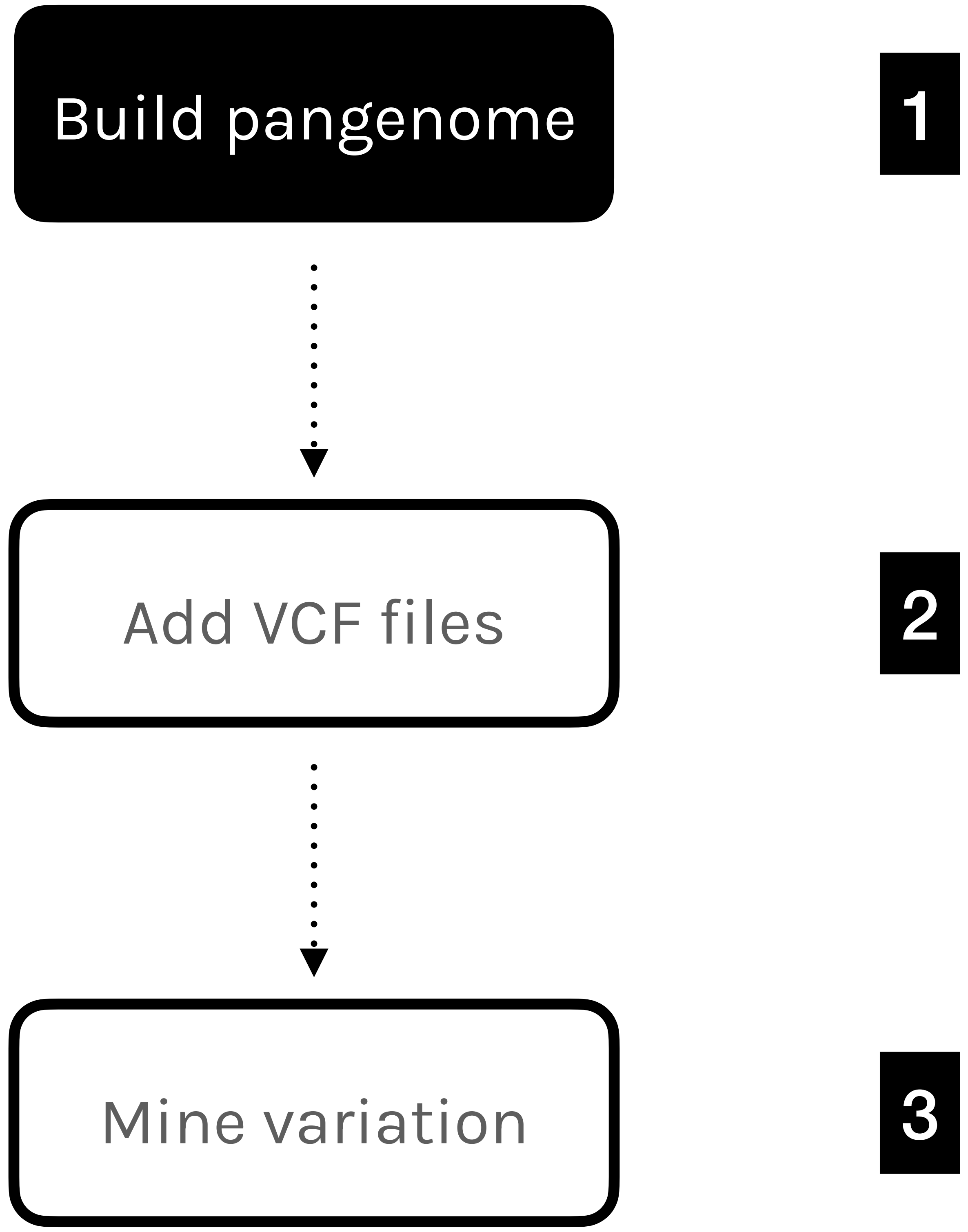
8 June 2021

Agenda

1. Analysing sequence variation (use case 1)
2. Data and encodings
3. Design
4. Implementation
5. Demo
6. Next steps

Variation in a Pangenome





Build pangenome

1

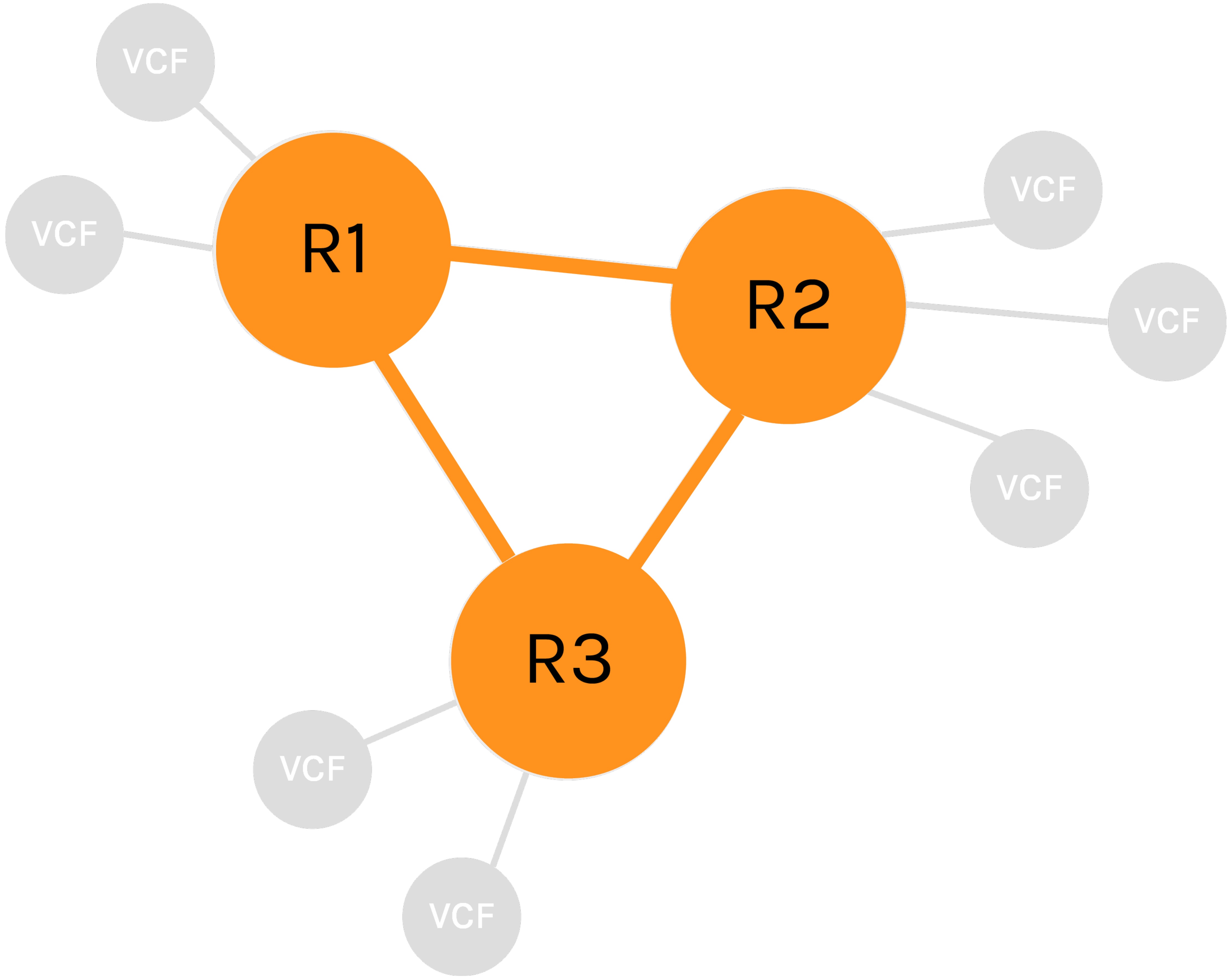
Add VCF files

2

Mine variation

3

Variation in a Pangenome

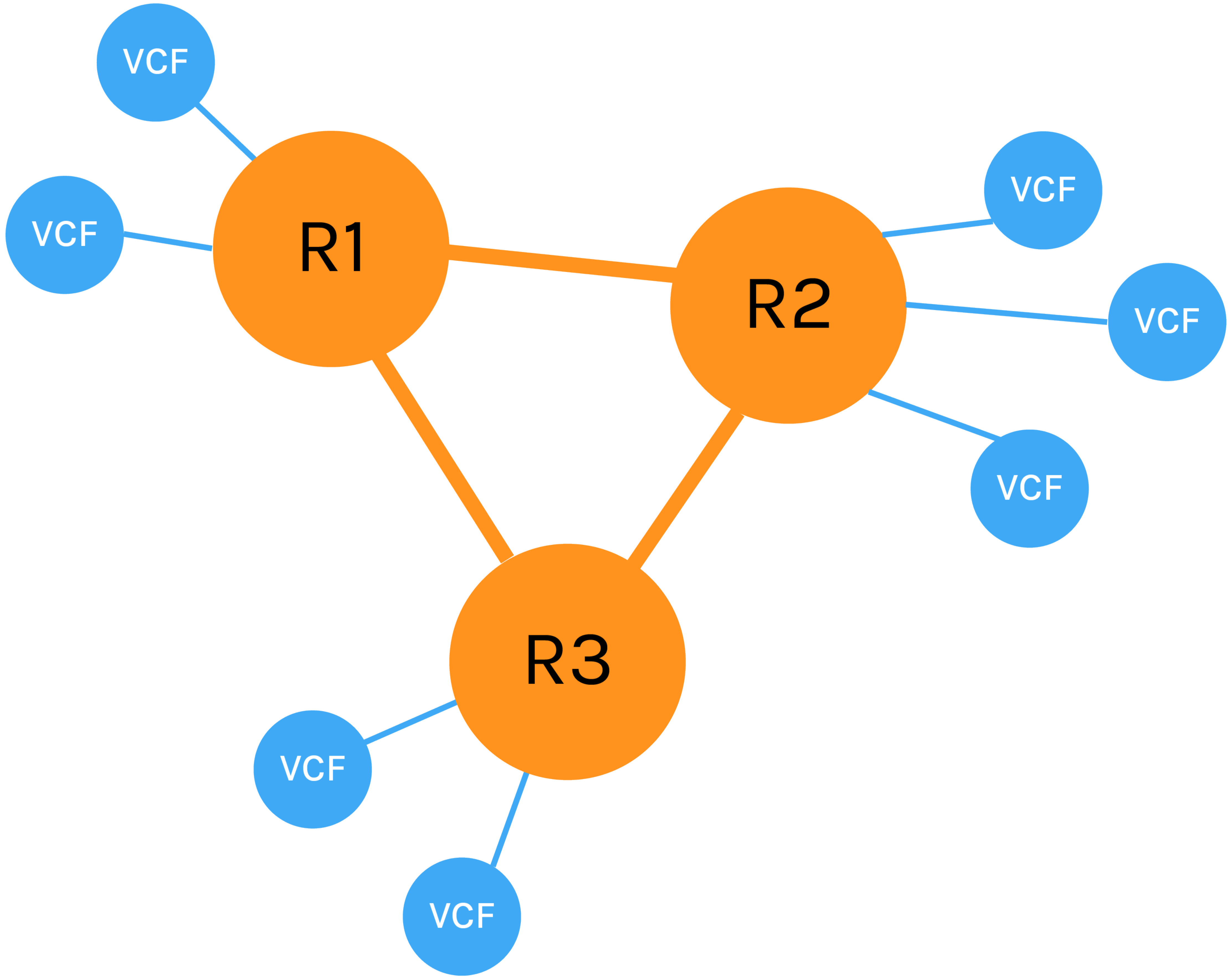


Build pangenome 1

Add VCF files 2

Mine variation 3

Variation in a Pangenome



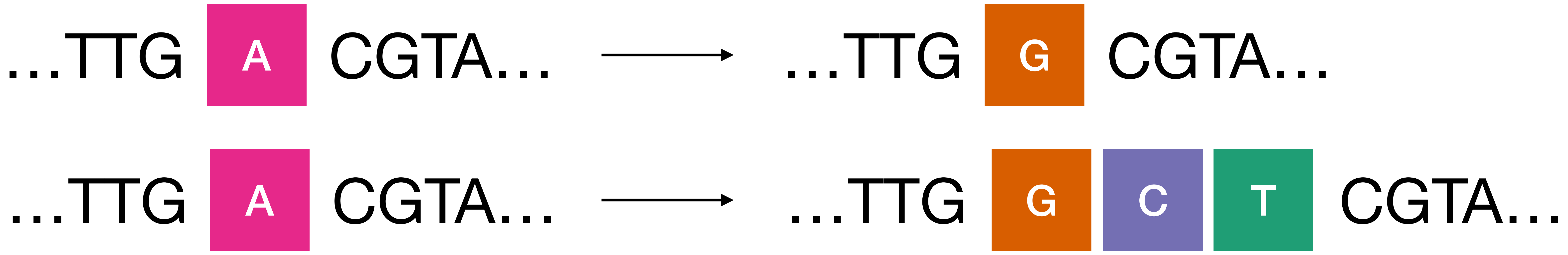
Build pangenome 1

Add VCF files 2

Mine variation 3

Sequence Variants: data

SNPs and small indels:



...TTG A CGTA... → ...TTG G CGTA...

...TTG A CGTA... → ...TTG G C T CGTA...

Annotation data:

- Meta data (e.g. phenotypes, traits and evolutionary context)
- Genome annotation data (e.g. domain, catalytic site)

Sequence Variants: tasks

- Single locus [gene]:

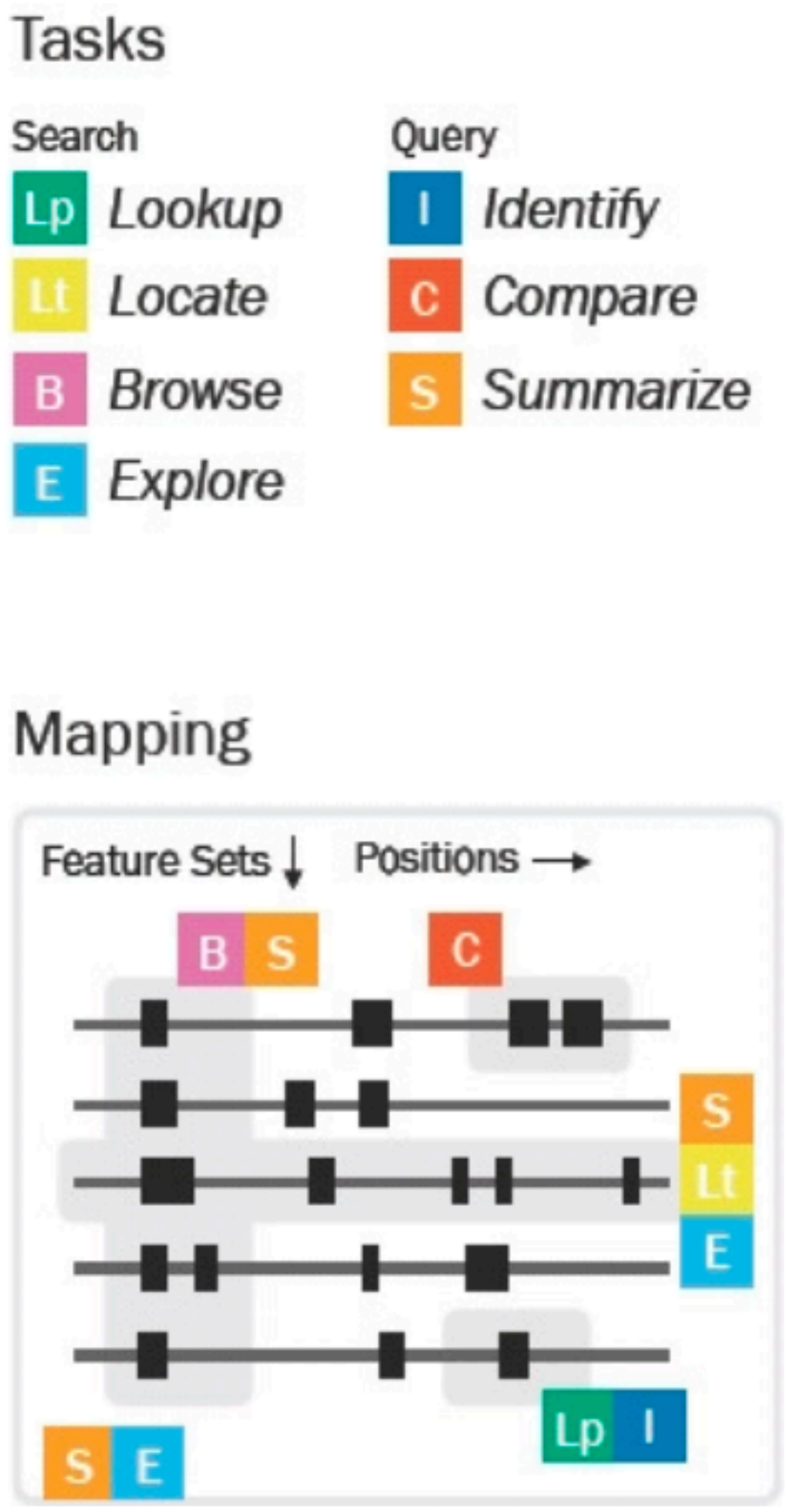
  Lookup / navigate to a known gene

  Browse / identify mutations at position of a known gene (w.r.t multiple references)

  Explore variation in a gene: repeatedly browse features at known positions / locate features

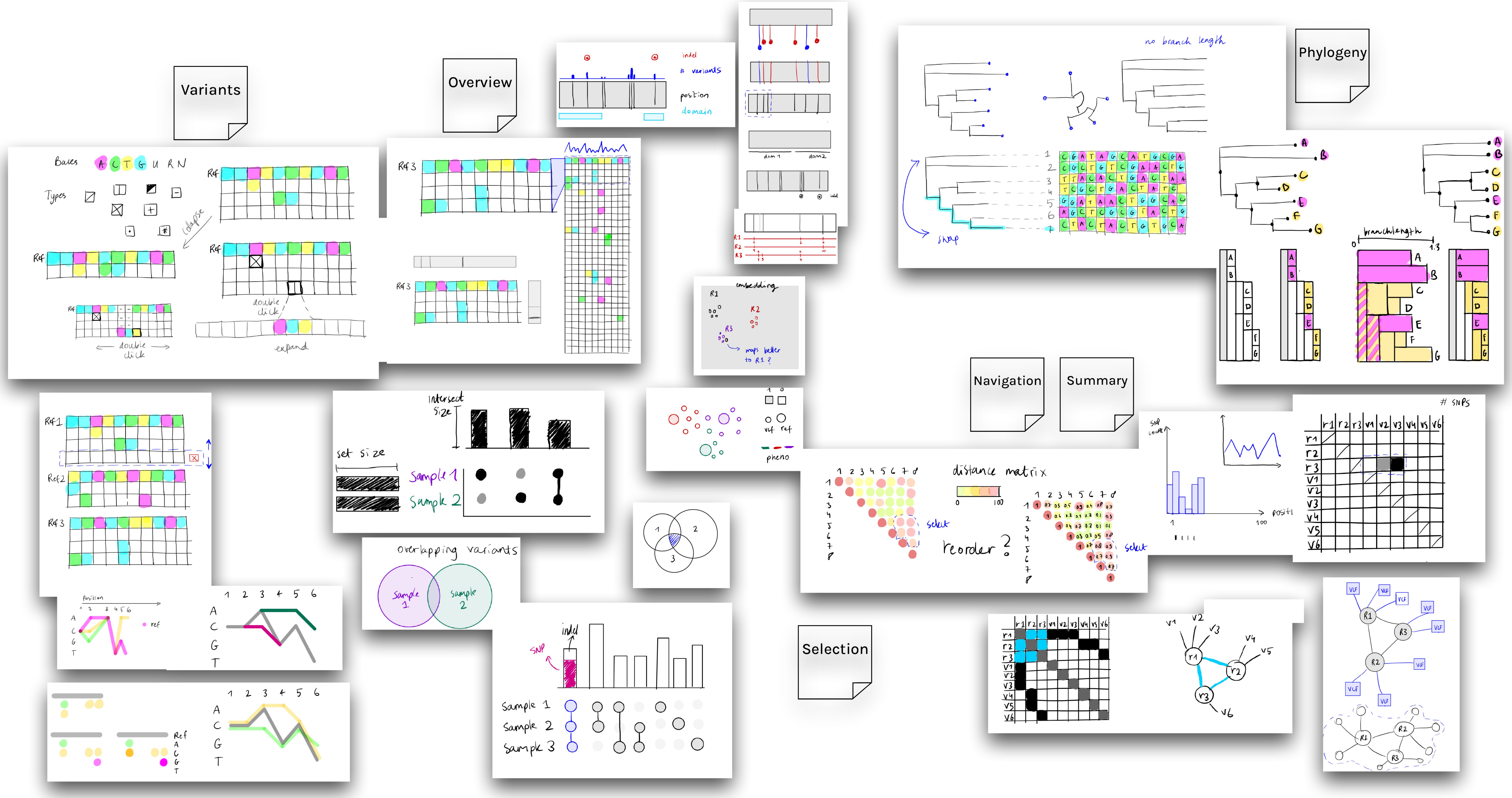
  Summarize variable regions and annotations

- Multi locus [gene + region around it]



S Nusrat, T Harbig, N Gehlenborg. "Tasks, Techniques, and Tools for Genomic Data Visualization" Computer Graphics Forum

Brainstorm



Variants

Overview

Phylogeny

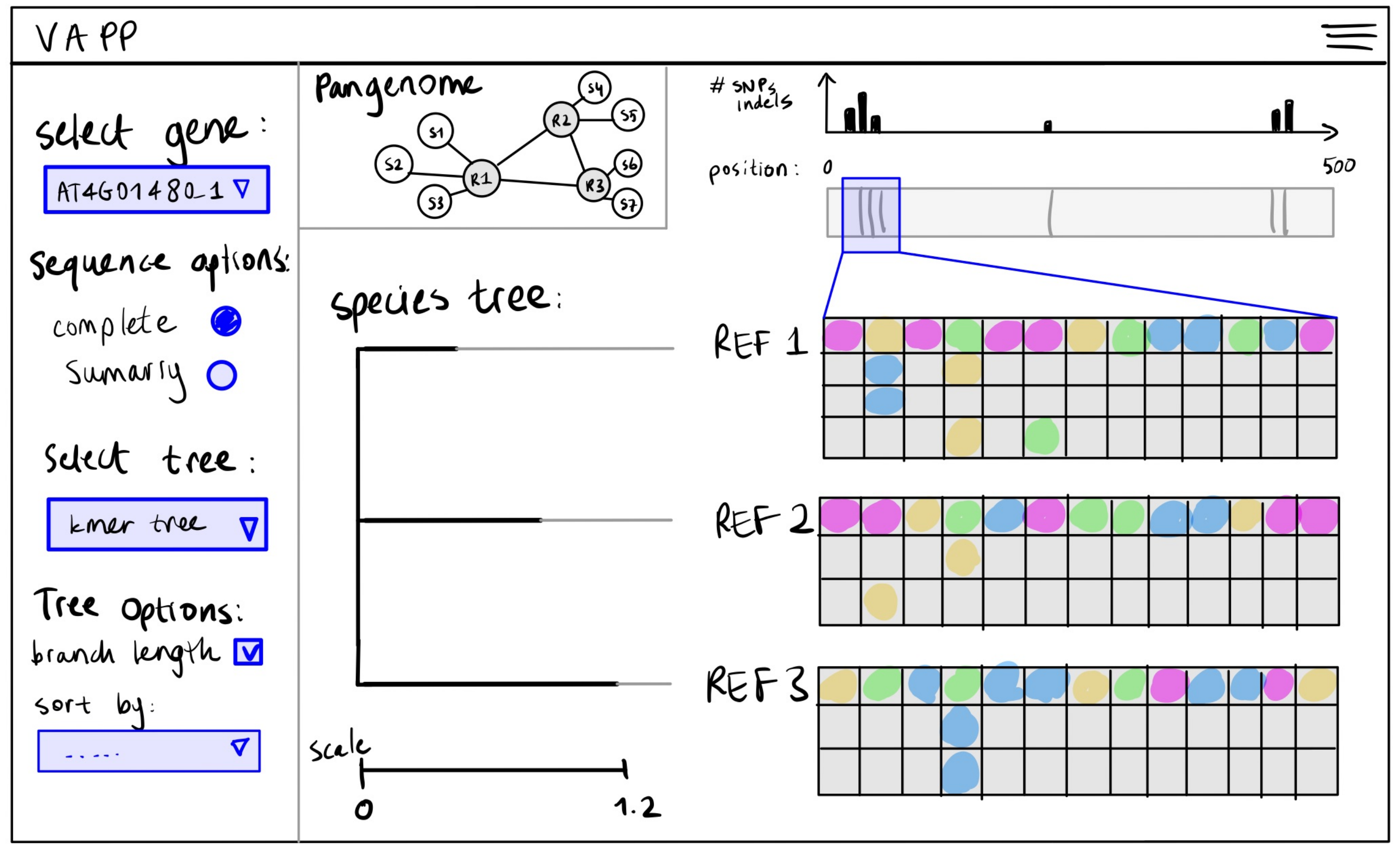
Navigation

Summary

Selection

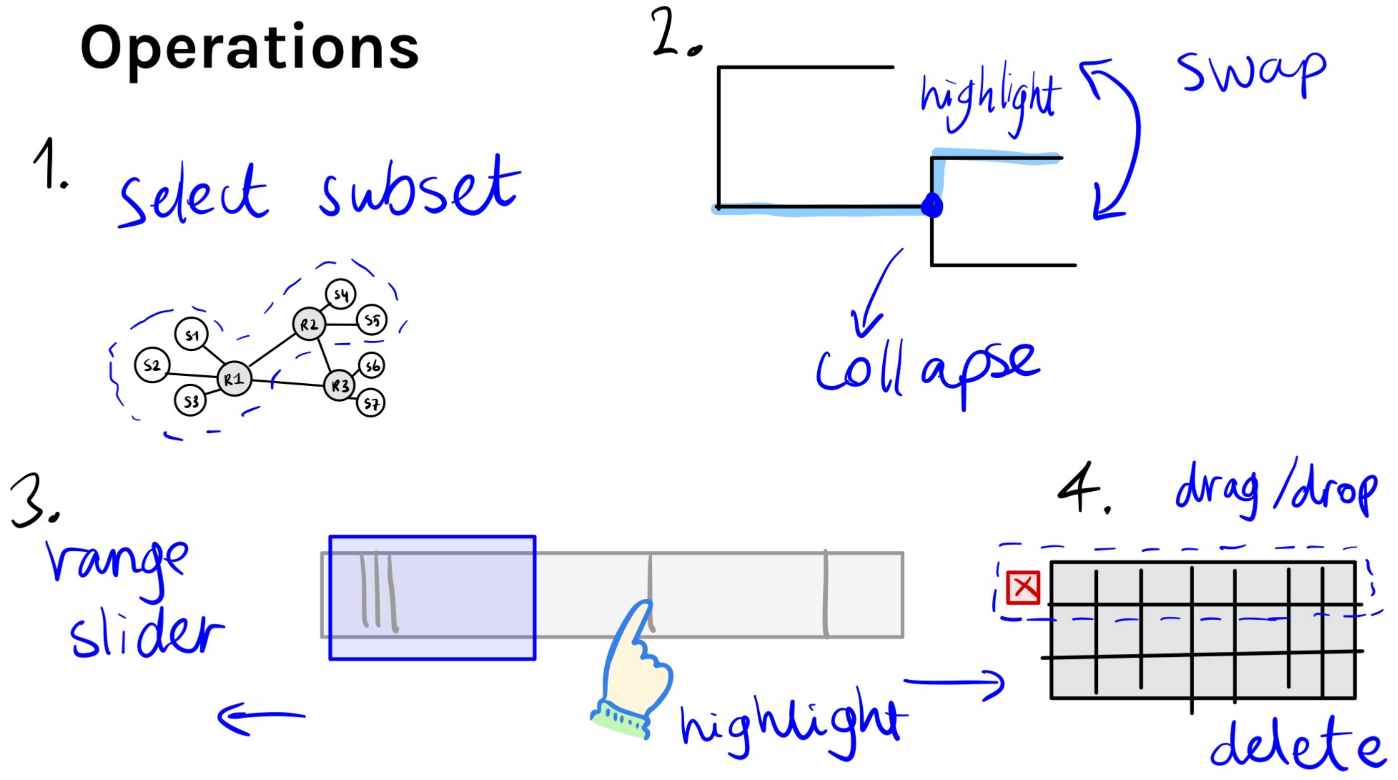
Initial Designs

Layout

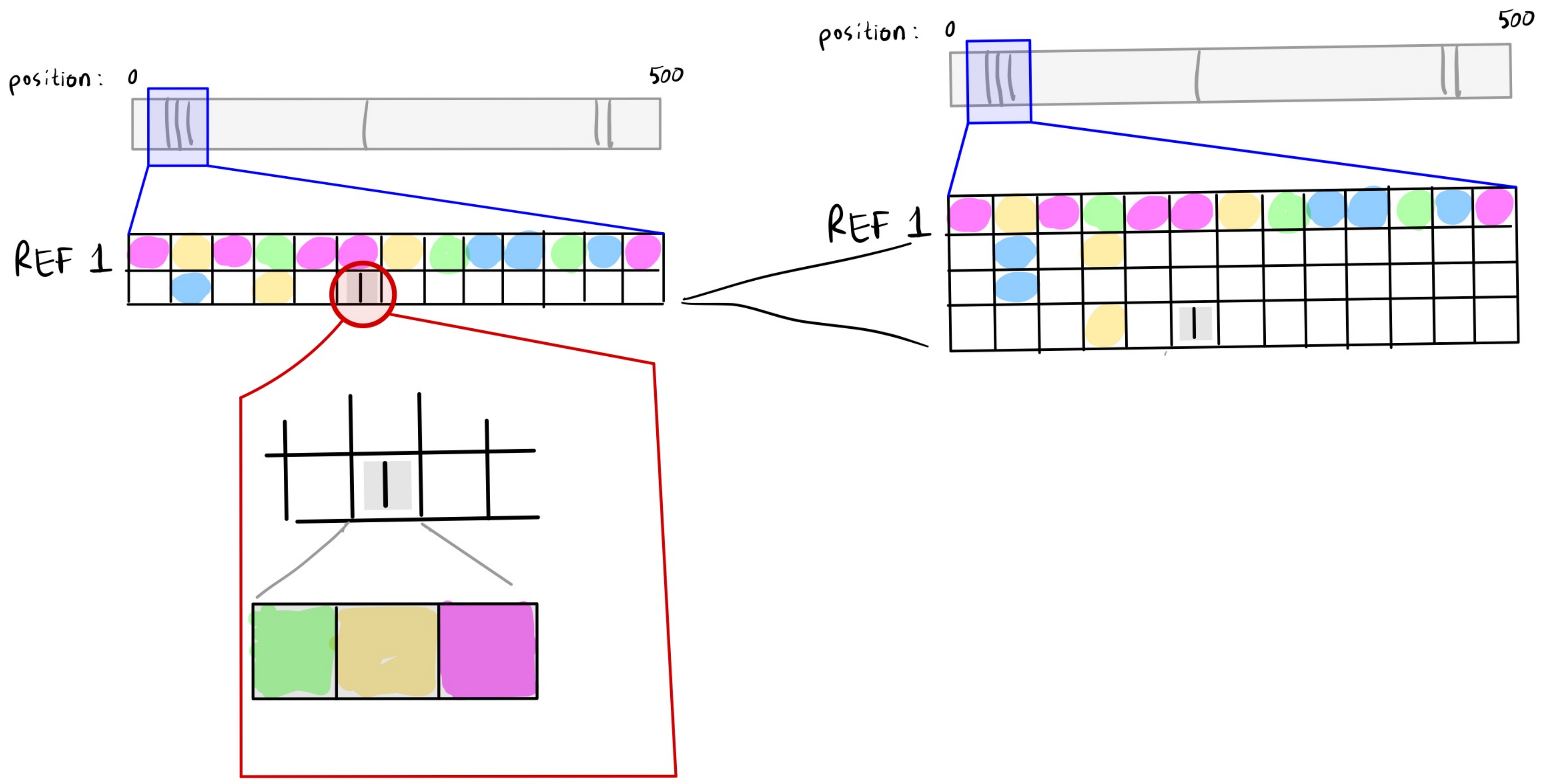


Title: Sequence variation
Author: Astrid van den Brandt
Date: May 2021
Sheet: 1

Operations



Focus

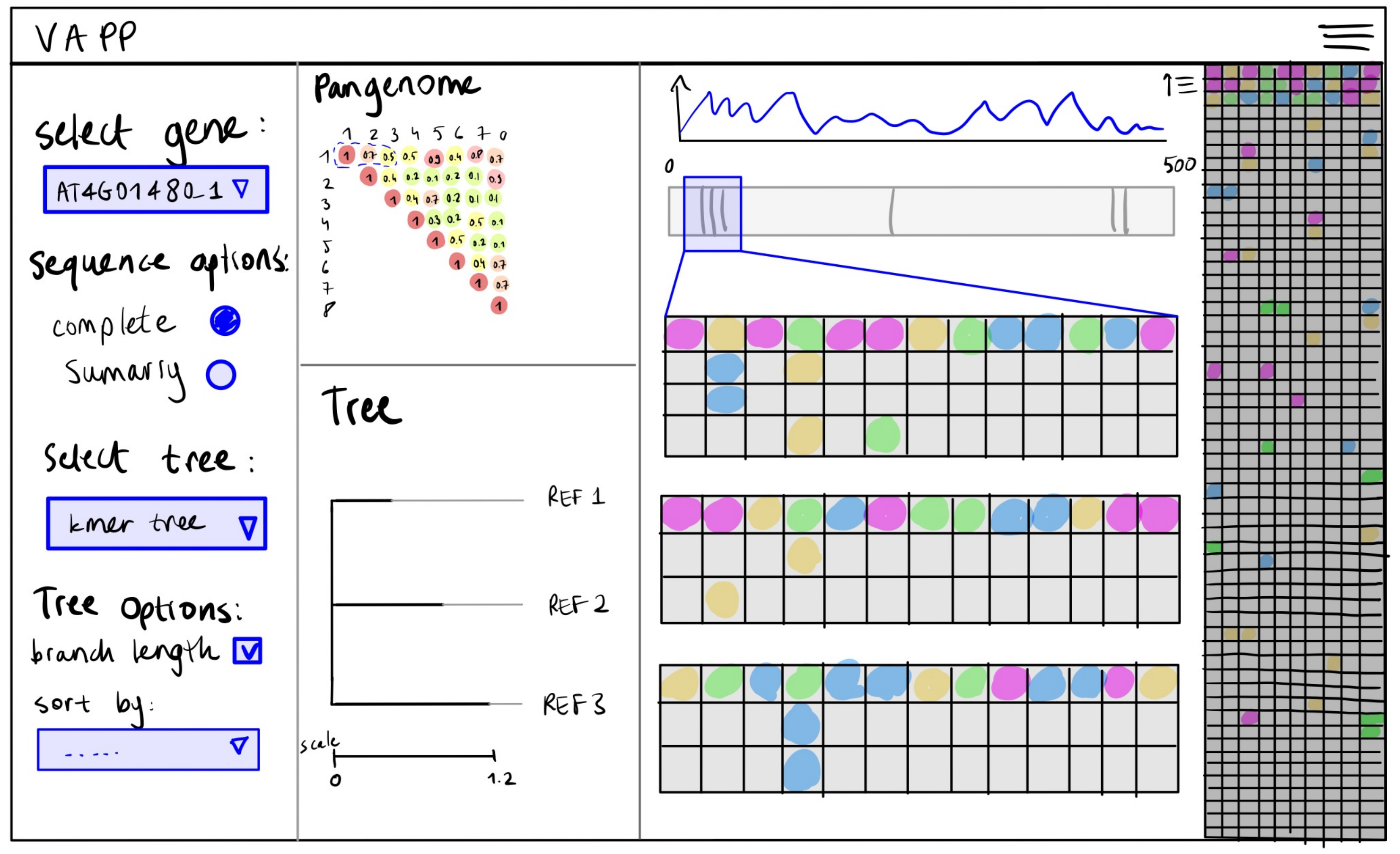


Discussion

- Scalability? e.g. 500 lines
- summary views of variants
  - per haplotype?
  - compressed?
- Indels

Initial Designs

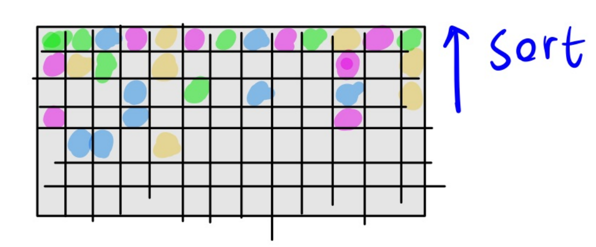
Layout



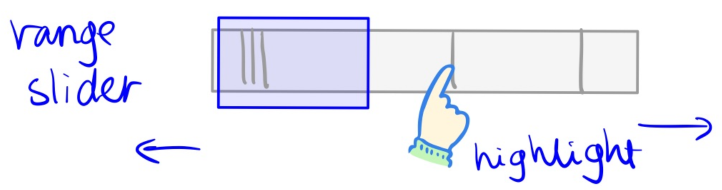
Title: Sequence variation
Author: Astrid van den Brandt
Date: May 2021
Sheet: 2

Operations

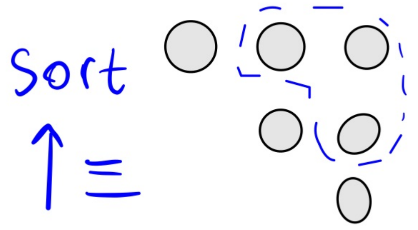
1.



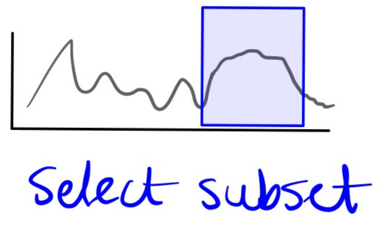
2.



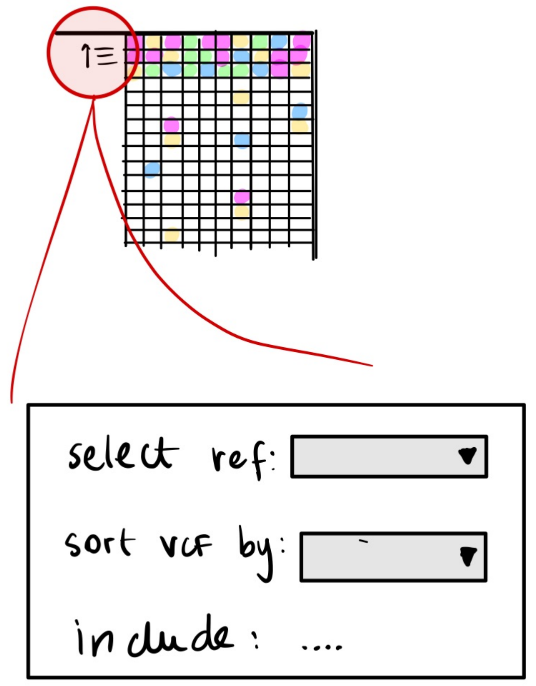
3. Select subset



4.



Focus

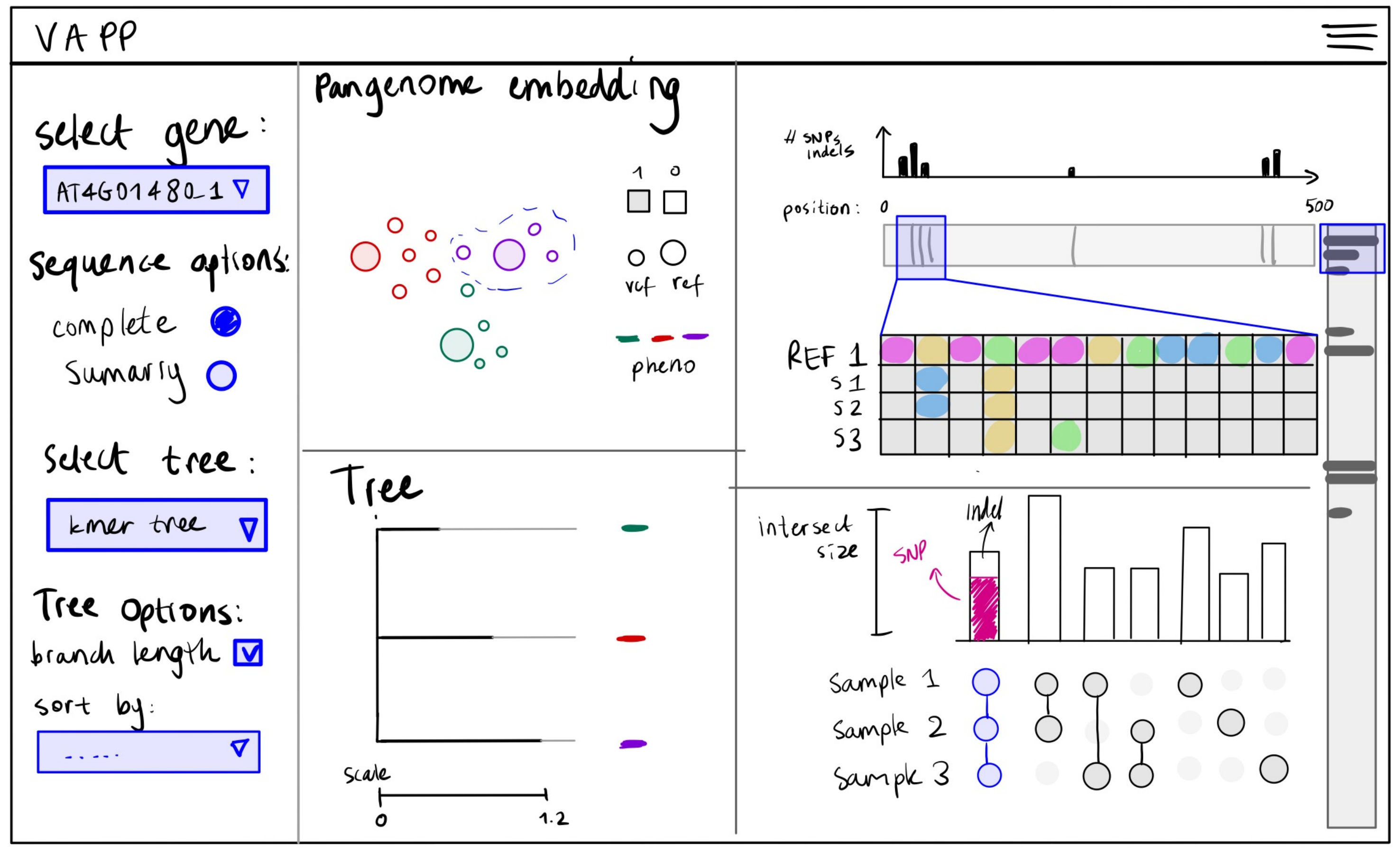


Discussion

- Overview VCF
- Select genomes
- tree decoupled

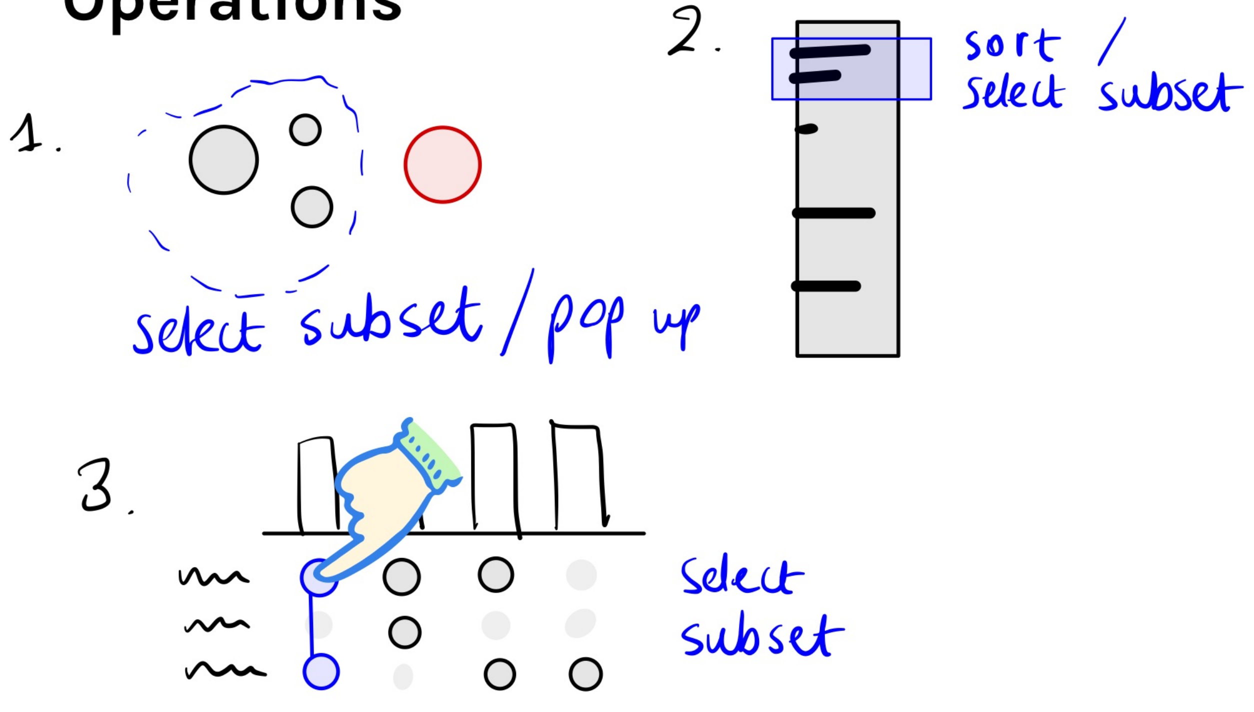
Initial Designs

Layout

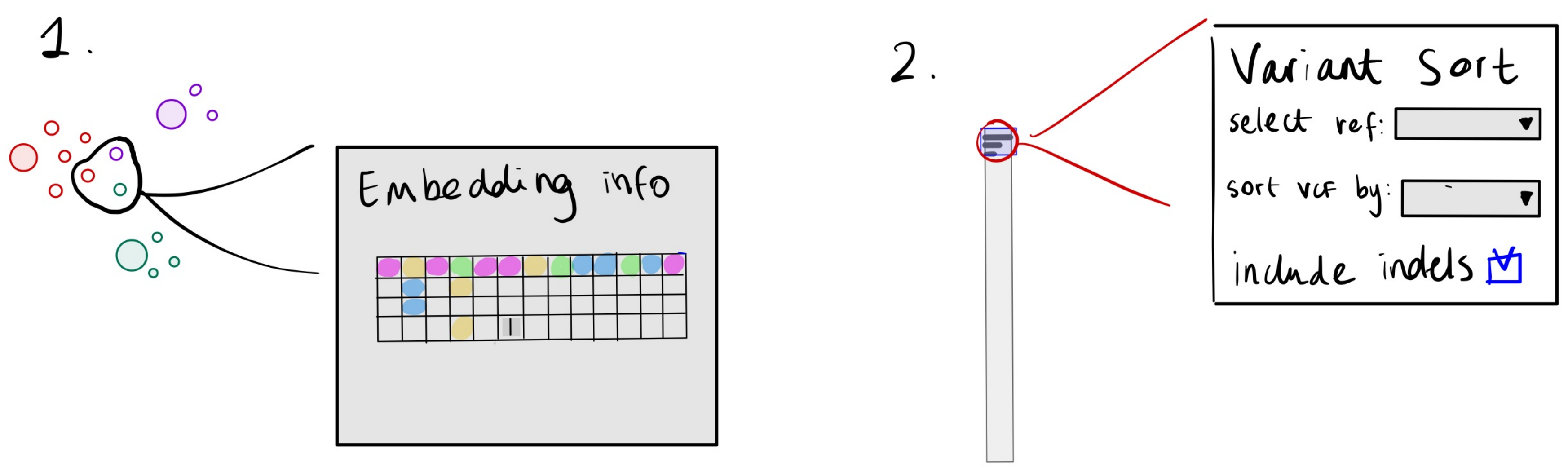


Title: Sequence variation
Author: Astrid van den Brandt
Date: May 2021
Sheet: 3

Operations



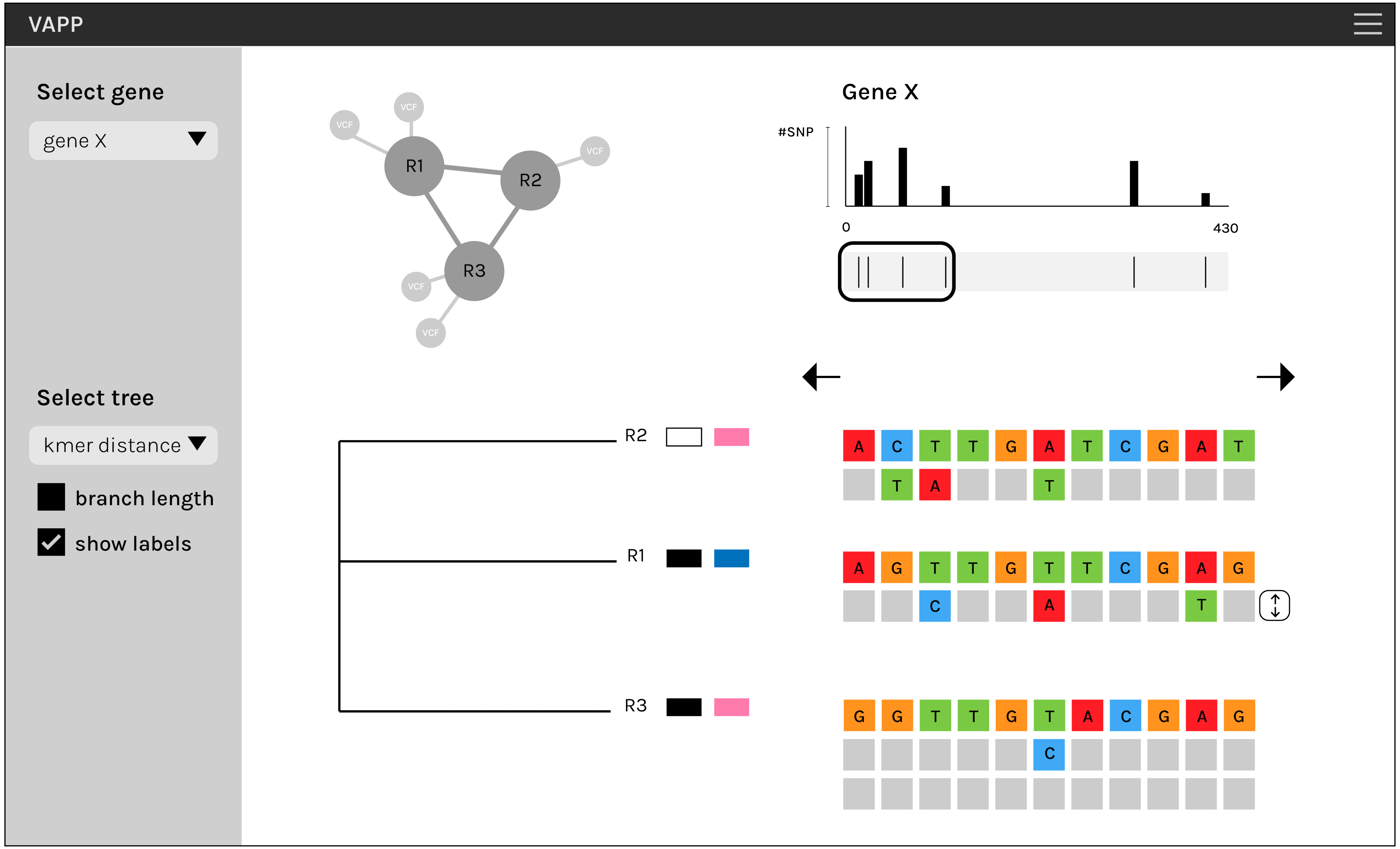
Focus



Discussion

- Explore sets of variants
- Meta data
- Pangenome embedding

Mockups



VAPP

Select gene

gene X ▼

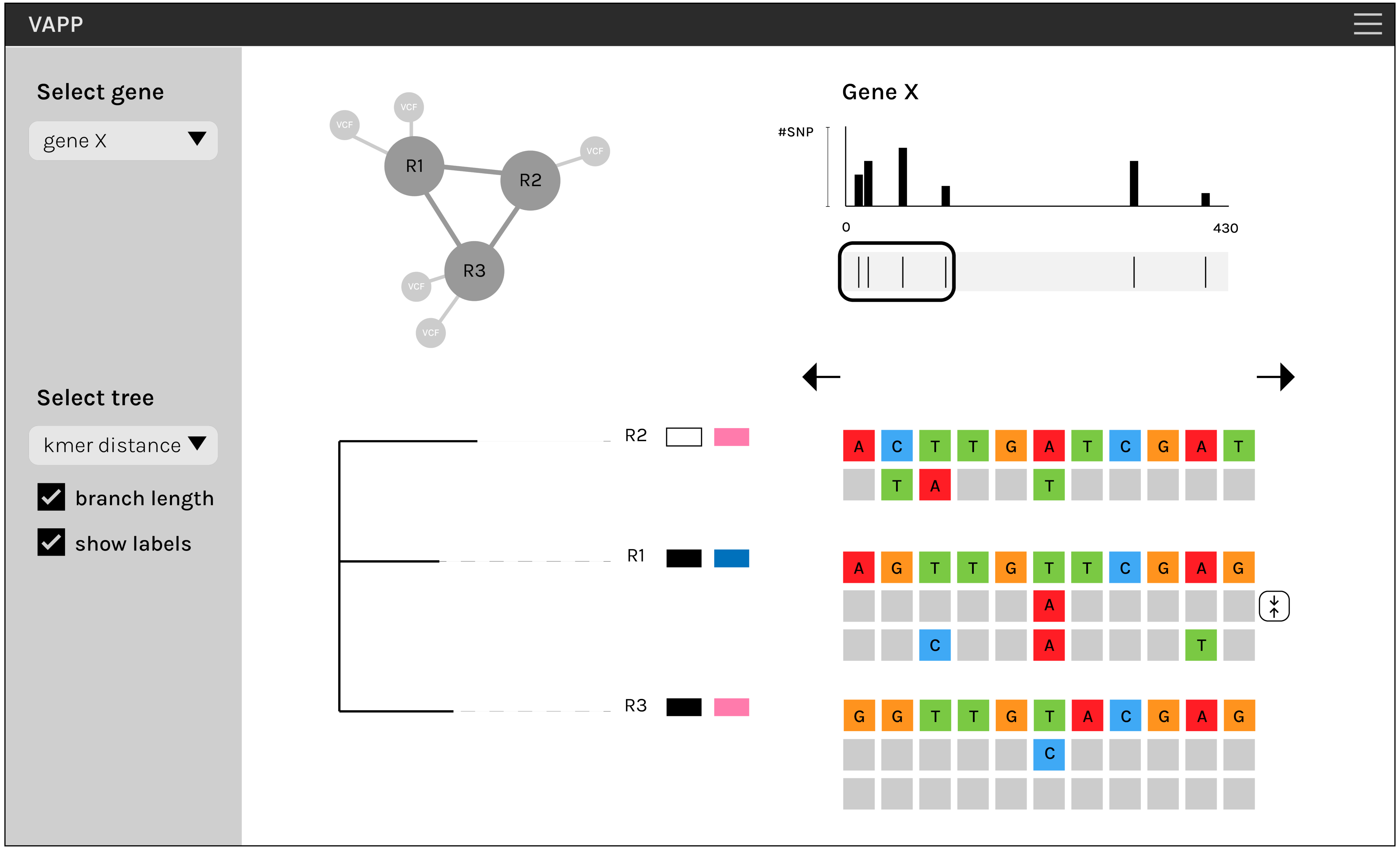
Select tree

kmer distance ▼

■ branch length

☑ show labels

Mockups



Select gene

gene X

Select tree

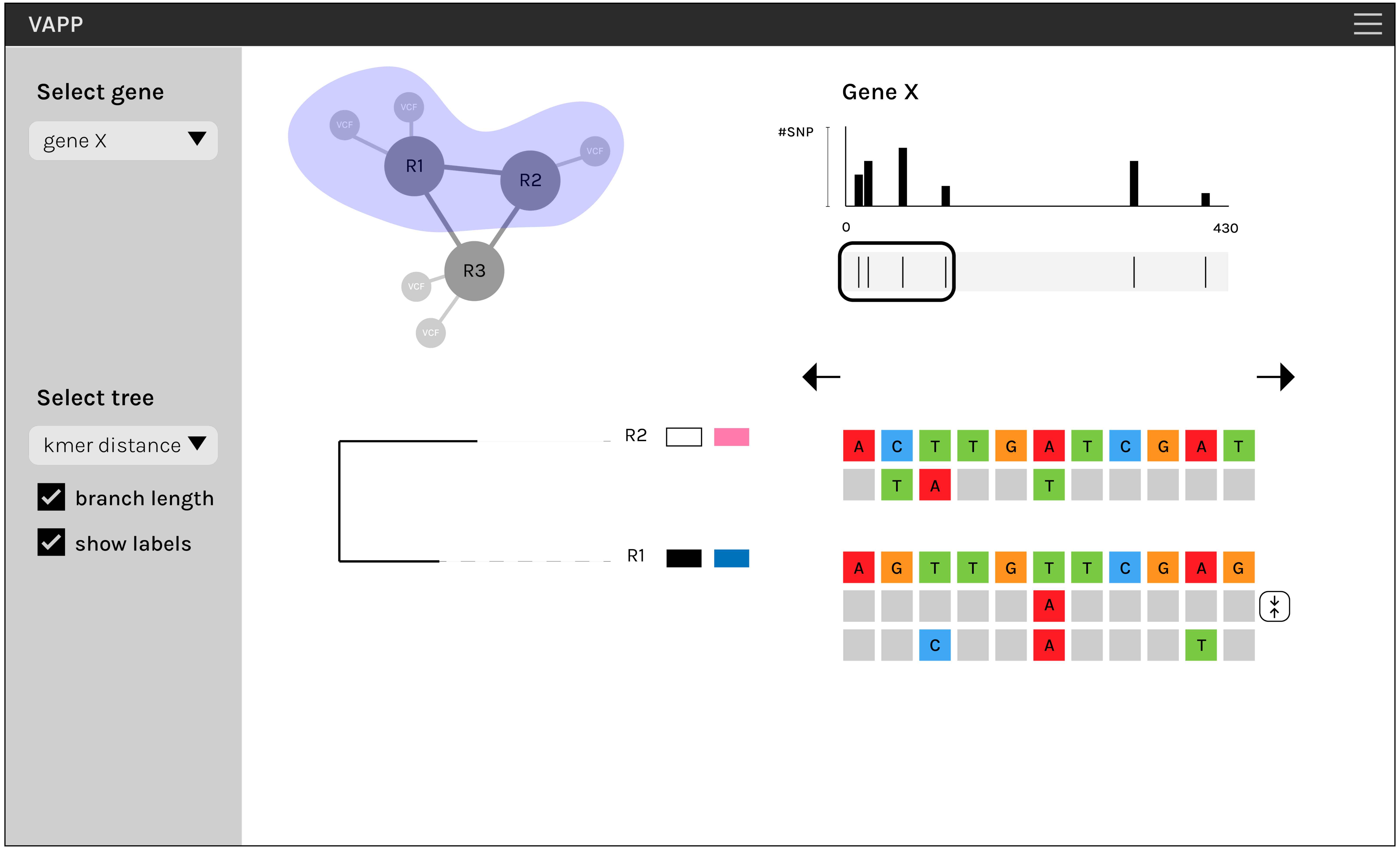
kmer distance

☑ branch length

☑ show labels

A	C	T	T	G	A	T	C	G	A	T
	T	A			T					
A	G	T	T	G	T	T	C	G	A	G
					A					
		C			A				T	
G	G	T	T	G	T	A	C	G	A	G
					C					

Mockups



VAPP

Select gene

gene X

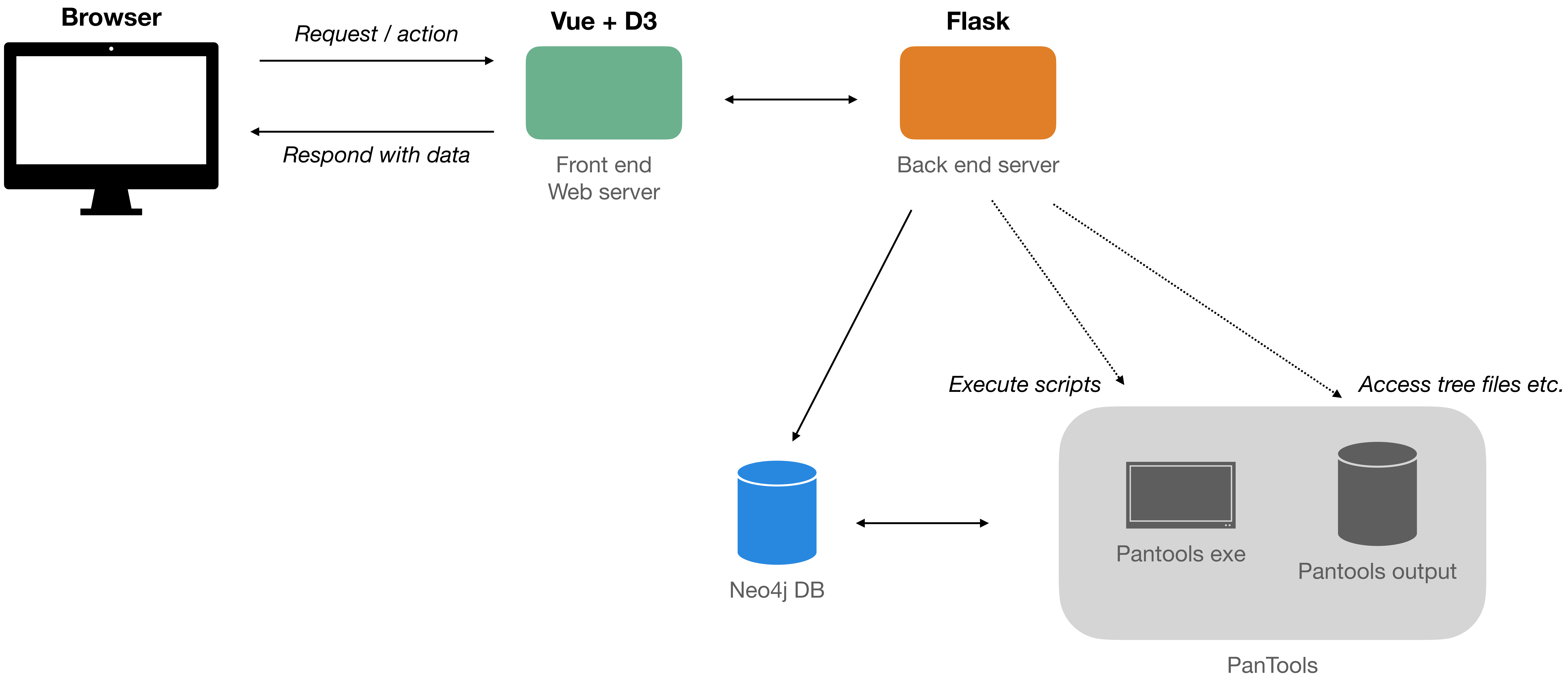
Select tree

kmer distance

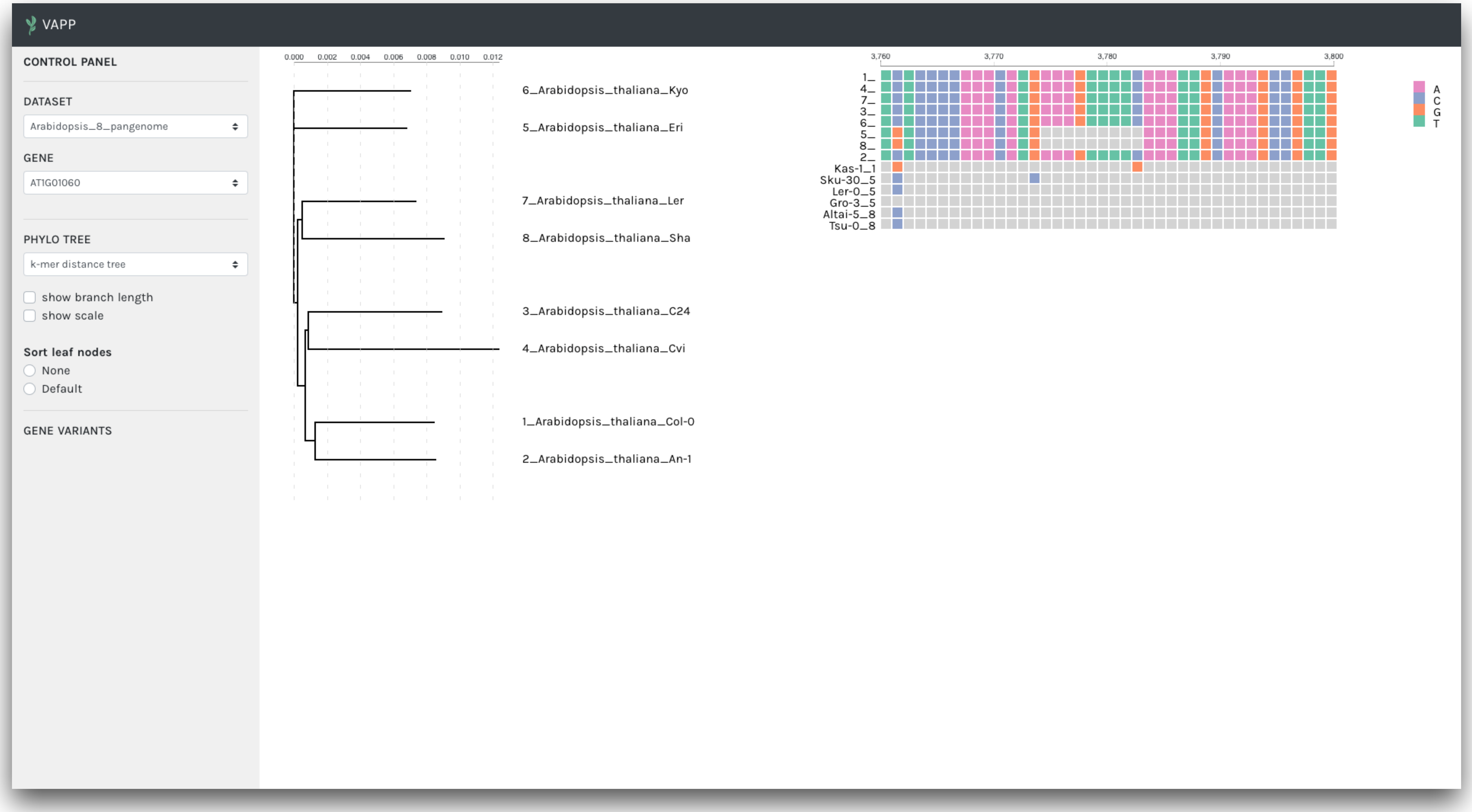
☑ branch length

☑ show labels

System Setup



Demo



CONTROL PANEL

DATASET

Arabidopsis_8_pangenome

GENE

AT1G01060

PHYLO TREE

k-mer distance tree

☐ show branch length
☐ show scale

Sort leaf nodes
○ None
○ Default

GENE VARIANTS

Unadressed Complexities

1. Homozygous vs heterozygous variants
2. Indels encoding
3. Polyploidy

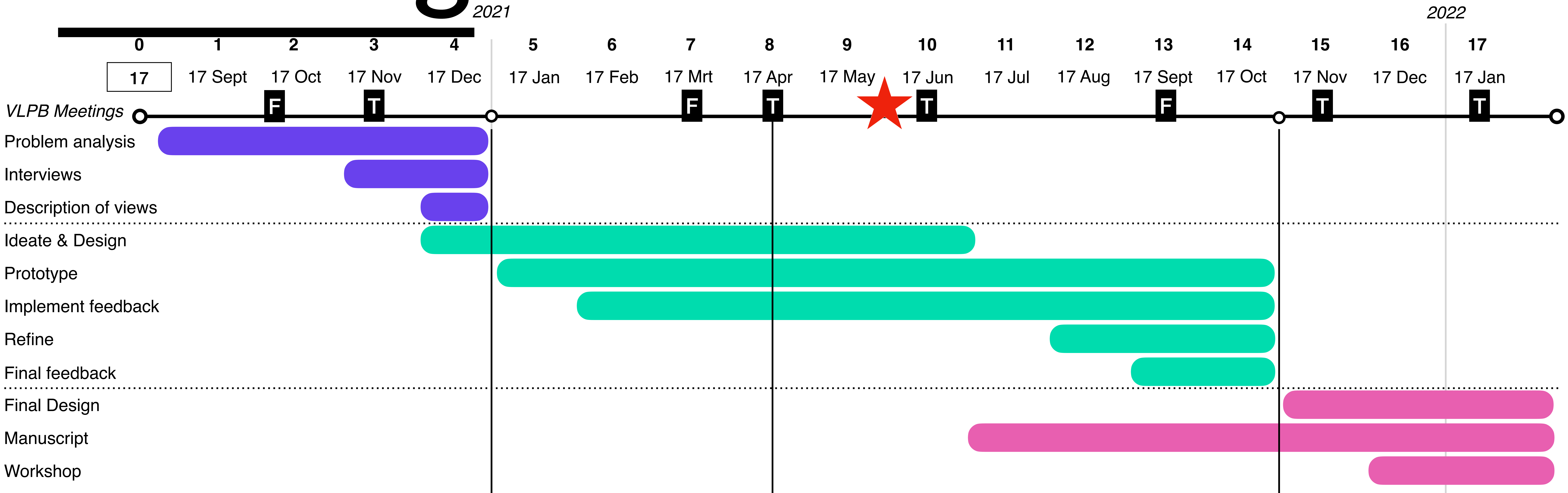
Next Steps

1. Prototype: focus+context, filter and select options, indels and hom/het
2. Upgrade to real crop data (e.g. tomato, cucumber)
3. Feedback current design: “design document”
    - Feedback functionality
    - Requirement list

3. Include structural variation (UC2)

Planning

We are here!



Problem analysis
Interviews
Literature review

Stage 1: Problem & Views

DESCRIPTION VIEWS

Small-scale variation
Large-scale variation

Stage 2: Design & Prototype

DRAFT PROTOTYPE

Final prototype
Workshop
Manuscript

Stage 3: Refine & Implement

FINAL PROTOTYPE & DOCUMENTATION

Questions?

Questions

1. How important is the type of the variant (e.g. SNP vs indel)?
2. Deciding which gene to analyze: which summary info is useful?
    - Genes with many variants
    - Genes with many variants that are close to each other
3. Coloring conventions?