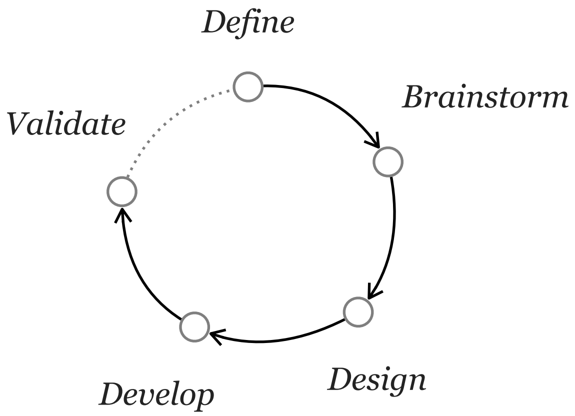
VAPP

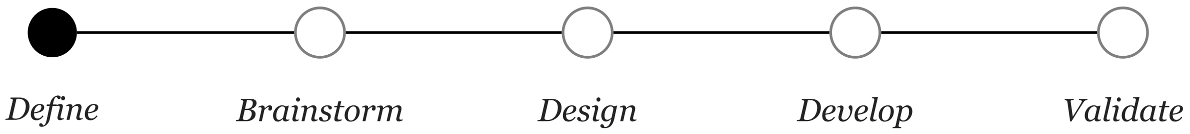
Visual Analytics for Plant Pangenomes



This design document describes the current ideas and design choices for the VAPP prototype application. The purpose of this "living" document is to keep track of our design explorations, formulate and prioritize requirements and share feedback. For the first design we focus on the analysis task of **exploring small-scale (sequence) variation of a target gene** in a plant pangenome.

The ideation process is inspired by the Five Design Sheet methodology. We start by outlining the main data and tasks that the visualization should support in exploring sequence variation. Secondly a brainstorm and the initial design sheets are presented. Furthermore we provide some mockups for a possible realisation design with a list of functionalities and requirements that are addressed. Lastly we list progress in the implementation of the views and requirements, and outline some functionalities to consider in the next iterations.

Data and abstraction



For the first design we use a small data set. The pangenome consists of 8 Arabidopsis thaliana genomes (Col-0, An-1, C24, Cvi, Eri, Kyo, Ler, Sha). For 3 genomes from geographically diverse regions (Col-0, Eri, Sha), we mapped (re)sequencing samples to their closest reference and called variants. During prototype development this data can be expanded to include more lines. In next iterations we will use a tomato pangenome.

For definition of tasks, we abstracted the relevant data into five main sets: *Source Data, Derived Data, Structural Annotations, Functional Annotations* and *Meta Data*. The Source Data consists of the 'actual' genomes and nucleotides inside them. Derived Data are the variants compared to the Source Data. Structural and Functional Annotations further describe the Source Data. Meta Data helps to put the Source data into context. All data sets have *Features*, which in turn can have *Attributes* or *Categories*. The *Cardinality* defines whether the feature is or applies to single data point (element), a region (segment) or the entire sample (sequence). Lastly *Level* describes the genomic level/scale at which Features are defined (e.g. Meta Data is defined on genome level).

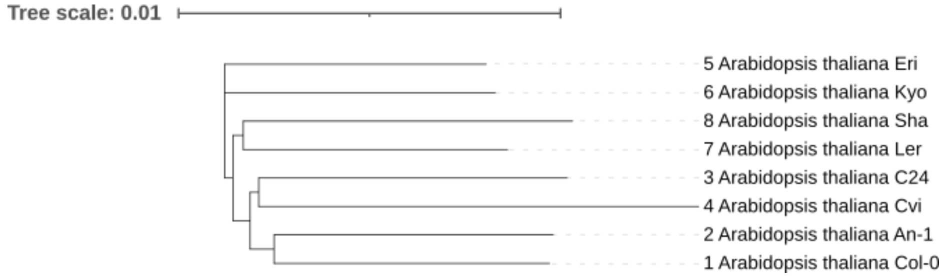

Tree scale: 0.01
5 Arabidopsis thaliana Eri
6 Arabidopsis thaliana Kyo
8 Arabidopsis thaliana Sha
7 Arabidopsis thaliana Ler
3 Arabidopsis thaliana C24
4 Arabidopsis thaliana Cvi
2 Arabidopsis thaliana An-1
1 Arabidopsis thaliana Col-0

Figure 1. K-mer distance tree of Arabidopsis pangenome

	Features	Attributes: Nr. Of Categories	Cardinality	Level
Source Data	Genome	length position	sequences	genome gene
	Nucleotide	4 (+ unknown)	element	
	Amino Acid	20		
Derived / Variant Data	SNP	length position type coverage dosage	element	genome gene
	Indel		element segment	
	SV	length position	element	genome
Structural Annotations	Gene	length position	segment	genome
	Exon			gene
	QTL			gene
Functional Annotations	Biological function		segment	gene
	Biochemical function			
	Protein domain			
Meta Data	Phenotype		sequences	genome
	Trait			
	Origin			
	Phylogeny	branch length bootstrap value	sequences	genome

Table 1. Data Abstraction

Task definition

To understand which actions a user may wants to perform on the data we describe visualization tasks. In the formulation of the the tasks we use the abstracted data terms from Table 1.

> *“VAPP is an interactive tool to help genome scientists visually analyze sequence variation of a target gene across a collection of lines in a multi-reference context to find interesting phenotype relationships for improved crops.”*

The **high-level task** is to analyze sequence variation in a target gene across multiple lines in order to examine diversity associated with agronomic traits or phenotypes. Below we define **4 low-level tasks** that facilitate the high-level task:

T1. Lookup a target gene

- Select or navigate to a known gene
  - aggregate / filter values and attributes.

T2. Locate variants in homologous genes

- Compare segments to identify variant locations;

- Identify variant attributes per line
  - navigate along the genome positions via brushing and panning;

- Compare variants between lines w.r.t. multiple references
  - navigate and rearrange genomes;

- Summarise variants globally and per reference: show frequency and/or distribution of SNP’s and indels
  - change visual encodings for genomes and variants, aggregate segments.

T3. Lookup functional annotations and meta data

- Identify values and attributes
  - show aggregations

T4. Browse or explore variants together with functional annotations / phenotype data

- Reveal agronomically important loci
  - abstract and filter genome segments;

- Compare functional annotations at a locus
  - aggregate and juxtapose functional information;

- Find phenotype-related variants
  - aggregate while zooming and brushing;

- Explore variants and their evolutionary relationships.

Design requirements and assumptions

Here we describe the **basic design requirements** that should be enabled in order to perform the previously defined tasks.

R1. Variant overview with multiple references (T2)

To represent the full diversity within the species or population of interest, multiple reference sequences and their variants are shown at once. Options are provided to adapt, collapse and expand the overview for further investigation.

R2. Multiple linked views (T2, T4)

A dashboard like system that shows multiple (connected) views show the source data, variants, annotations and metadata at several levels of scale and aggregation, facilitating different focus points/ perspectives for variant analysis.

R3. Flexible arrangement and linking of views (T4)

Users have the option to rearrange or hide views depending the analysis question at hand. Views can be coupled (e.g. juxtapose phenotypes and sequences) or decoupled (e.g. showing information in a connected table).

R4. Semantic filtering and slicing of the data (T1- T4)

The source data should be filtered in real time and visual cues should guide the users to find interesting subsets.

R5. Focus and context: details on demand (T2, T4)

Linked views show different focus areas and their contexts in order to support interpretation and navigation of the data. To avoid clutter, more detailed encodings are shown on demand or based on the genomic level at which the data is displayed.

R6. Extract visualizations and alignment output

For further analysis and presentation, it the system should facilitate extraction of (visual) overviews of the data (e.g. a PNG file of the visualization or a multi-FASTA file of the sequences).

R7. Connect to PanTools data structure and functions

The proposed visualization should connect to the PanTools data structure and should enable the user to call PanTools functions if needed.

R8. Web-based tool

The proposed tool will be web-based to make it easily accessible and to minimise installation burdens.

Design assumptions

Here we report some assumptions that apply to the tasks and design requirements.

A1. Gene-centric analysis

For the first design interaction we focus on gene-centric analysis of variants. Therefore we assume that the user knows the gene ID and positions that (s)he wants to study.

A2. Display 10 columns of phenotype data at once

The design should be able to display 10 different categories of phenotype and annotation data simultaneously.

A3. Predefined reference

For the current design we assume that the researcher wants to investigate a gene and it's variants in a multi-reference context, but where variants have already called with respect to their reference of choice (ideally the closest). The added value of the visualization is to show all the references and their variants in one go, such that a complete overview of the genomic diversity is given and phenotype relations can be explored.

Brainstorm Sketches

Exploring the design space

In the initial brainstorm we explore the possible design space by sketching ideas. Many small ideas were generated, some are stand-alone and other's should be considered in a combined way. The ideas are filtered and categorised intro six groups: variants, overview, phylogeny, selection, navigation and summary. Figure 3 on the next page shows sheet 1, which is an overview of these ideas. The motivation of this overview is to document the design space, which is useful for next iterations when some encodings may need revisiting.

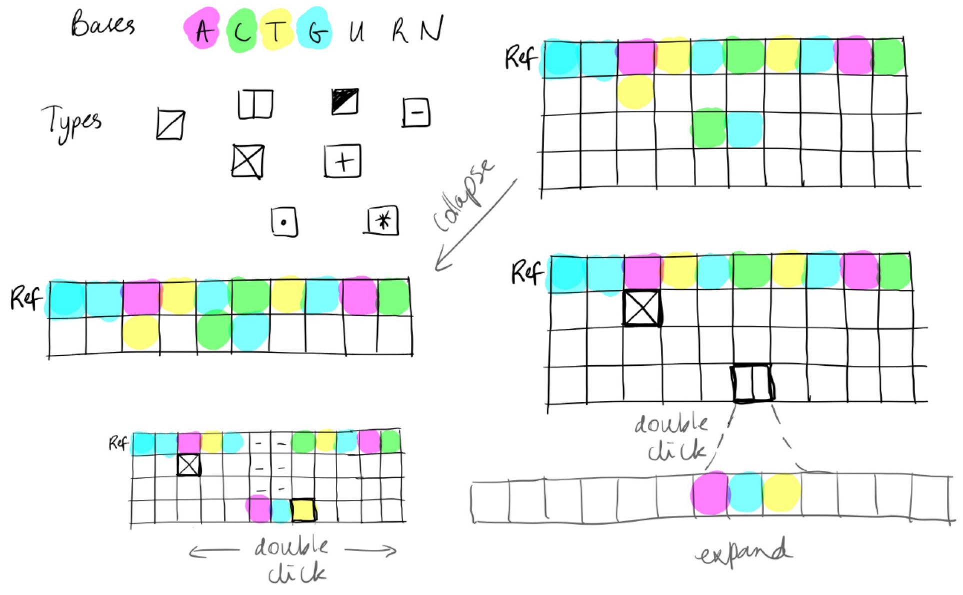
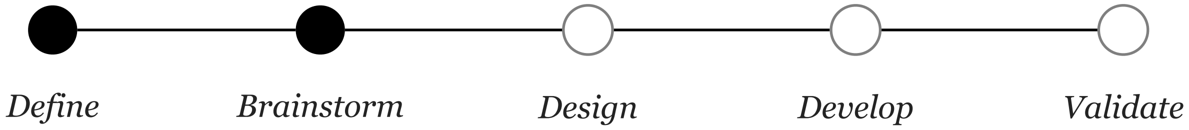


Figure 2. Some ideas for encoding variant types

Brainstorm ideas



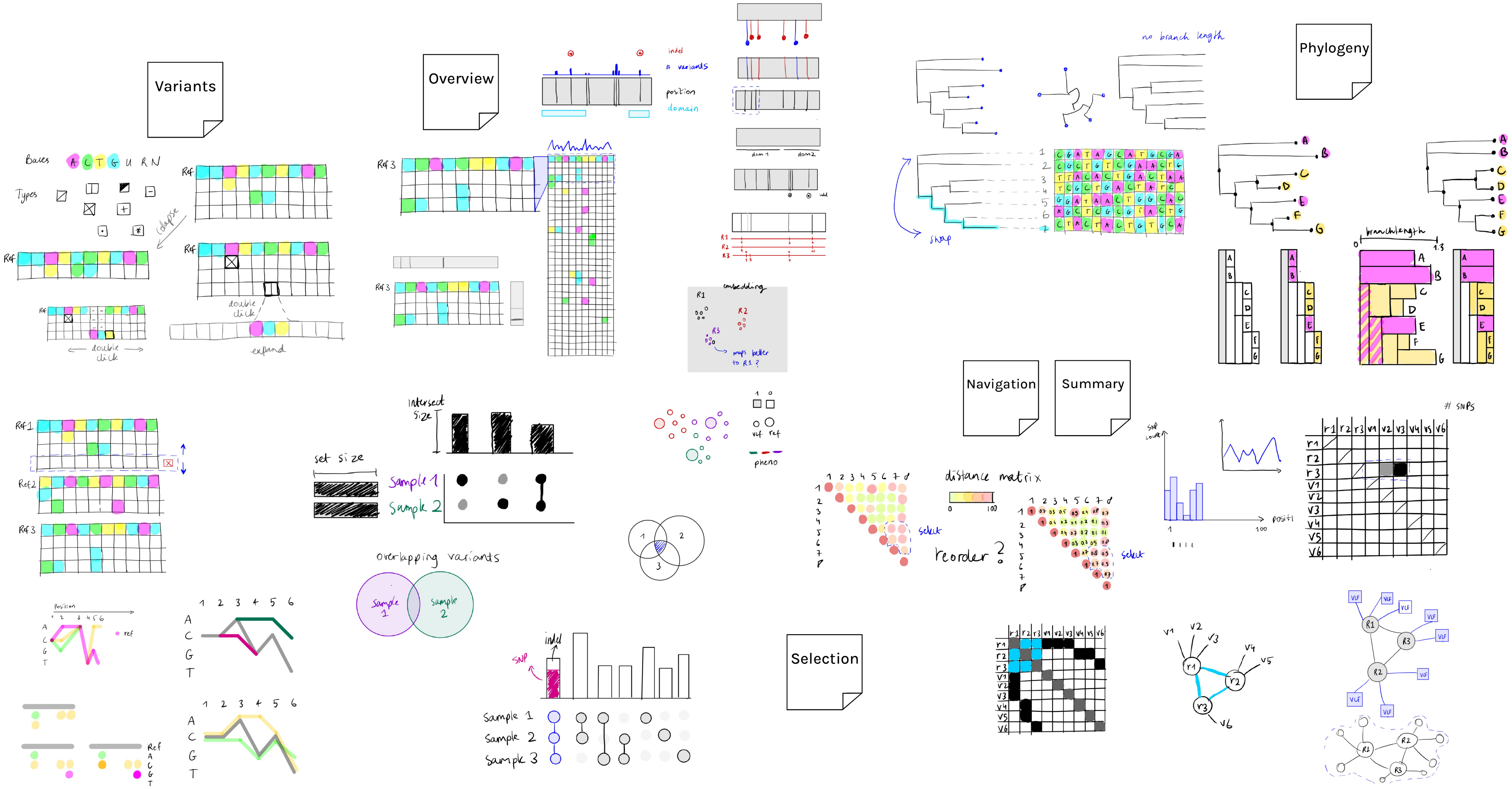


Figure 3. Brainstorm with categorised ideas: variants, overview, phylogeny, selection, navigation and summary

The ideas represented above show two different aspects of the data: (1) variants and their sequences and (2) phylogeny. In some explorations these are tightly coupled (e.g. a phylogenetic tree that can reorder a multiple sequence alignment that is shown in the ‘phylogeny’ group). In both aspects we should also be able to add phenotype information (e.g. by colouring labels or showing small glyphs). In this brainstorm this was not explored in much detail yet. Next to the aspects of the data, some interaction methods are explored: selection and navigation. Linked to this are some methods that show overviews or summaries of the data, for example a matrix where cells are color-coded to how similar the sequences are in order to select a subset for the other views.

Initial Designs

Creating design concepts

After the brainstorm three design sheets (2,3,4) are created. These design sheets focus on different ideas, of which strengths and weaknesses can be discussed for the final realisation design. Every sheet describes a design concept by 5 pointers: (1) its global layout, (2) focus - a key technique of the visualization, (3) the interaction options, (4) discussion section for advantages and disadvantages of this design and (5) meta-information (shown at the top).

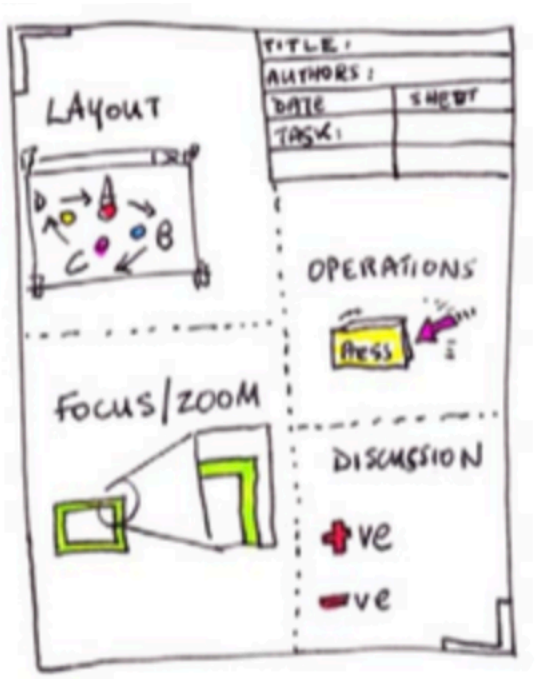
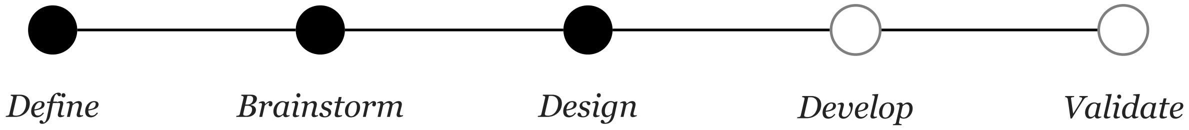
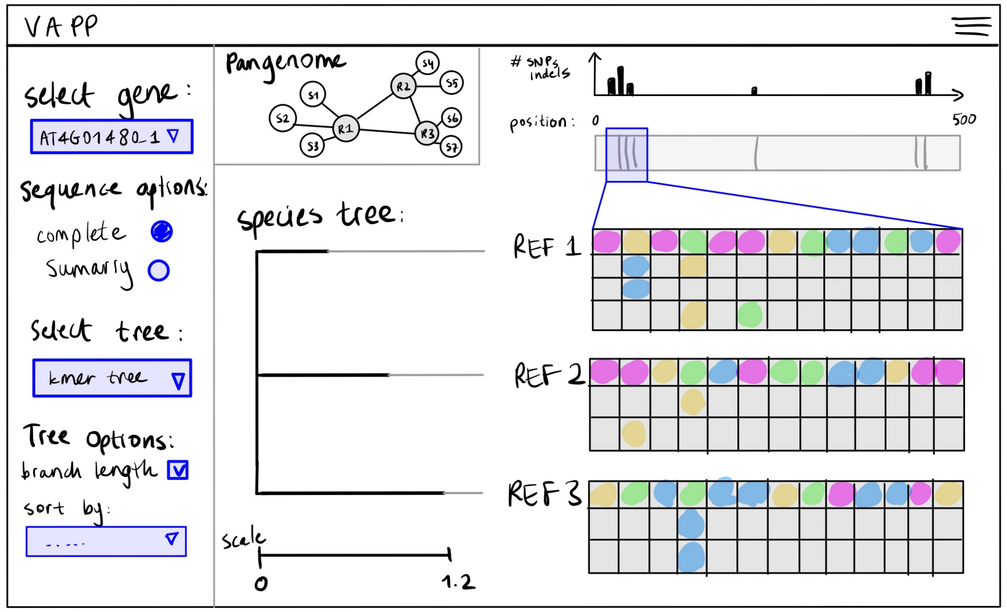


Figure 4. Design sheet example (created by Five Design Sheets)

Sheet 1

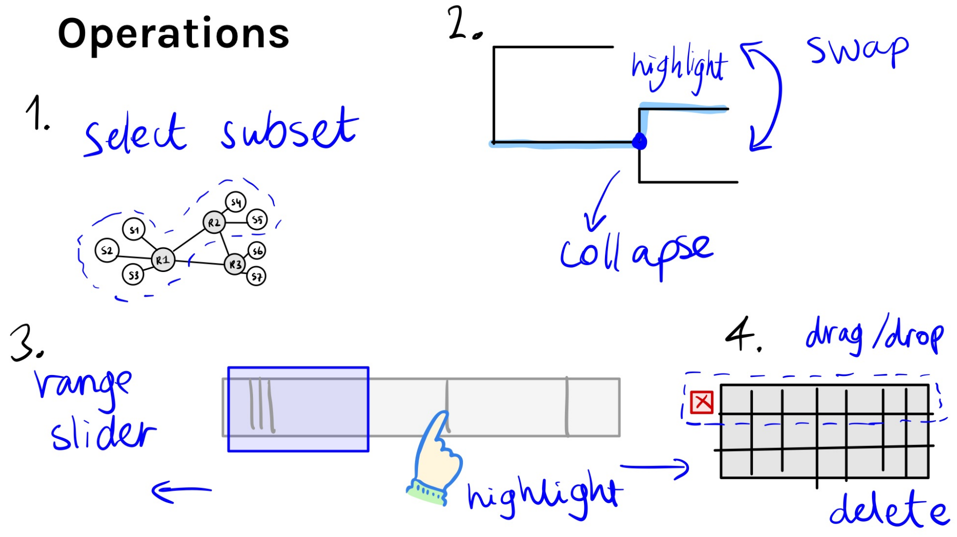


Layout

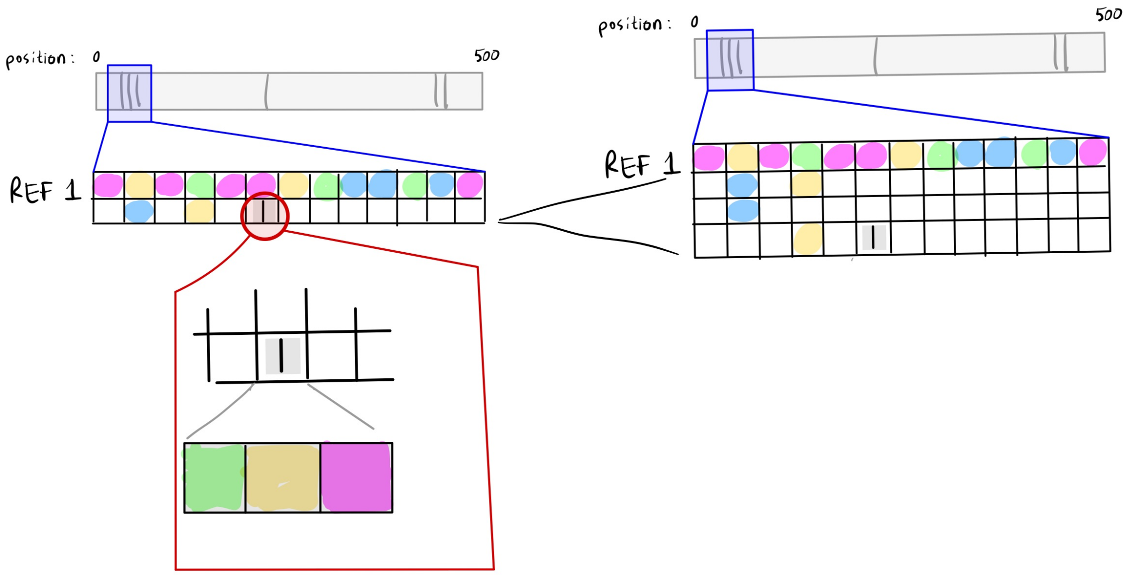


Title:	Sequence variation
Author:	Astrid van den Brandt
Date:	May 2021
Sheet:	1

Operations



Focus

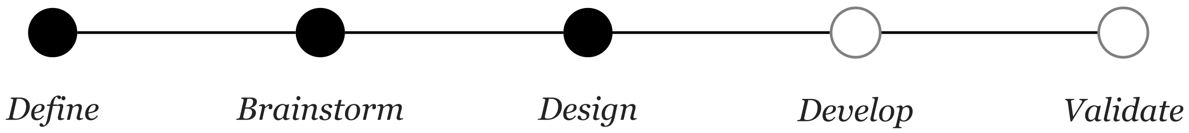


Discussion

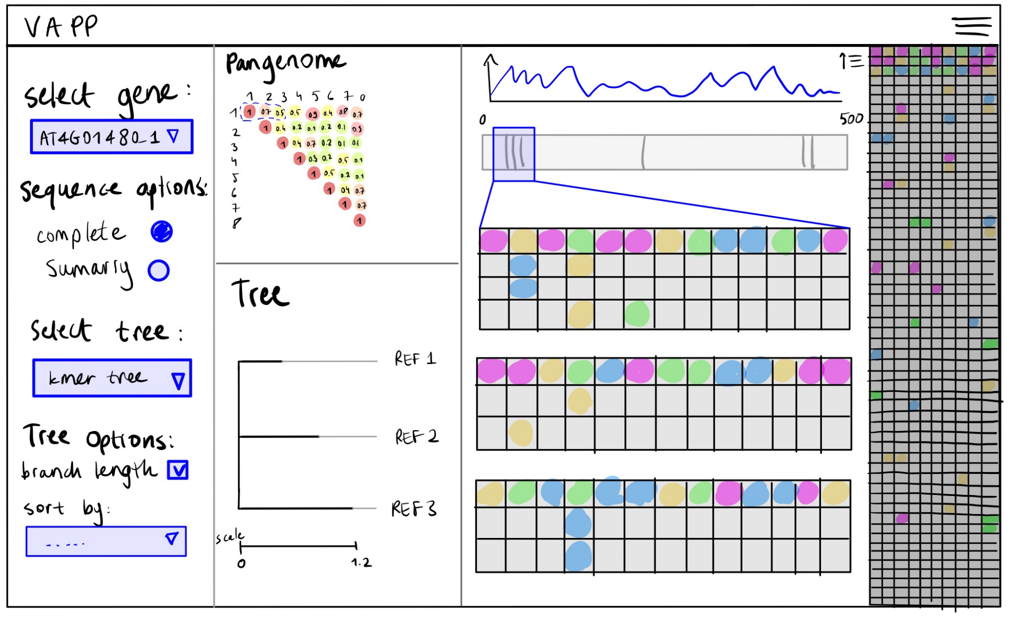
- Scalability? e.g. 500 lines
- summary views of variants
  - per haplotype?
  - compressed ?
- Indels

Above we show the first design concept (sheet 1). On the left side of the layout a menu with different interaction / control options is shown. The two main views are a phylogenetic tree and a sequence alignment, showing variants. The sequence alignment includes a summary view at the top which can be used to brush the sequence for details at the bottom. The focus section shows a possible way to encode indels. A point of discussion is whether this layout is scalable.

Sheet 2

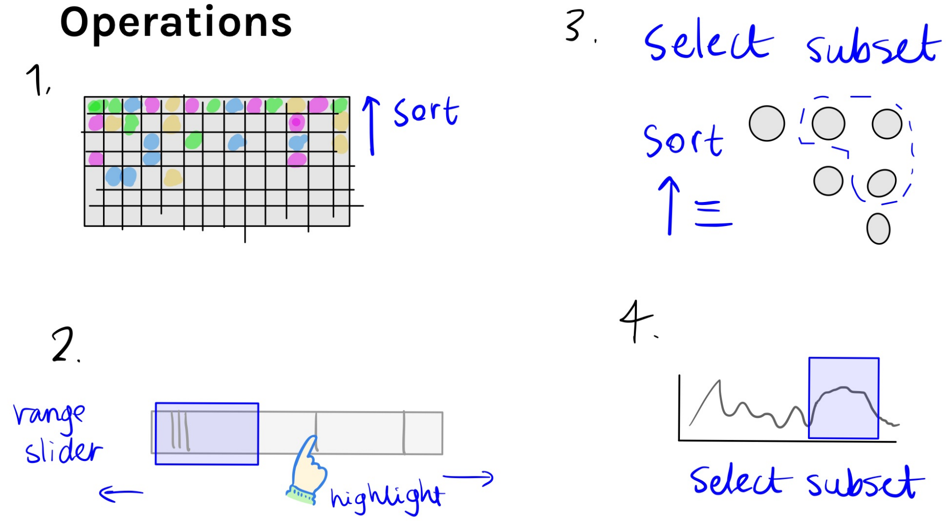


Layout

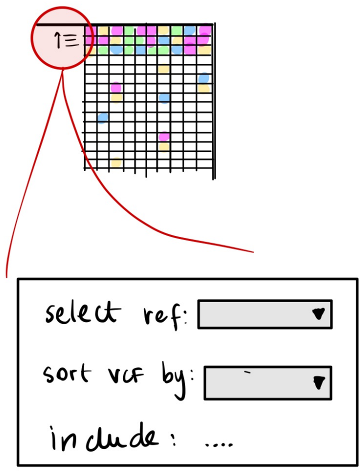


Title: Sequence variation
Author: Astrid van den Brandt
Date: May 2021
Sheet: 2

Operations



Focus

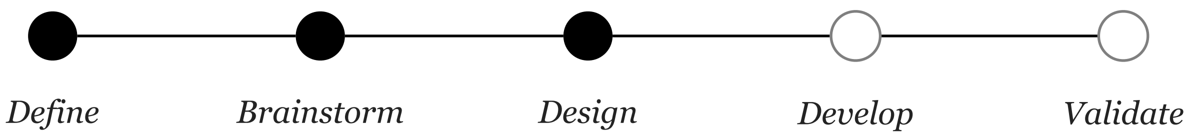


Discussion

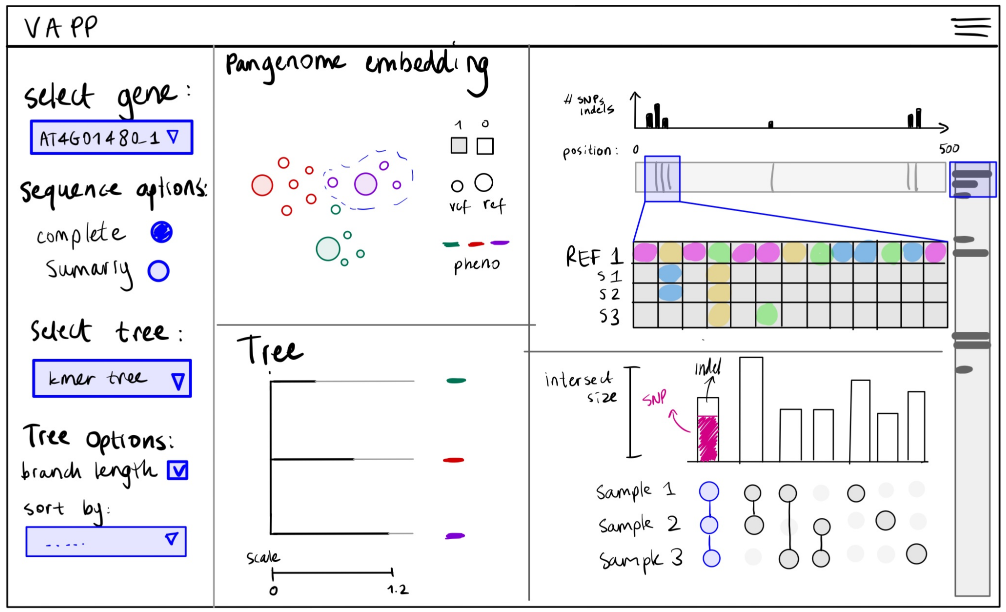
- Overview VCF
- Select genomes
- tree decoupled

This page shows the second design concept (sheet 2). It differs from sheet 1 in the overview encodings and subset options. Instead of using a pangenome graph and variant calls, we can select a subset based on their sequence similarity. Next we can choose which lines to include per variant by the list overview in the right (which has a pop-up that can sort and filter on criteria). In this design the three can also be decoupled from the MSA.

Sheet 3

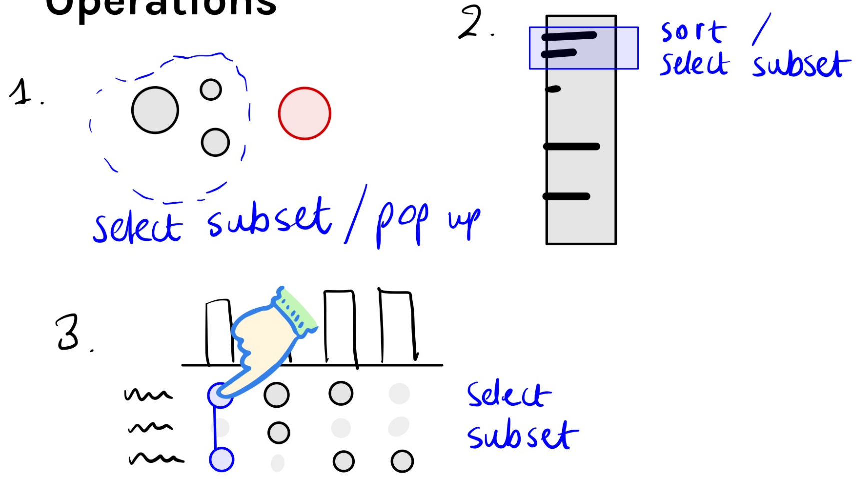


Layout

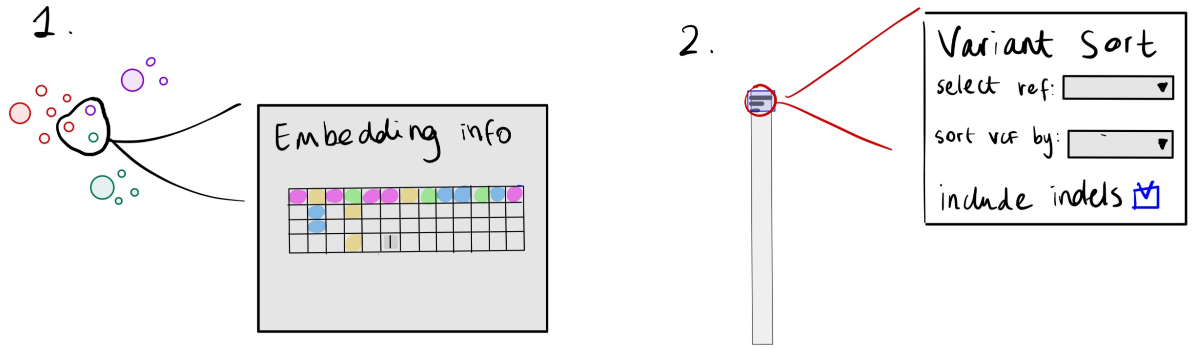


Title: Sequence variation
Author: Astrid van den Brandt
Date: May 2021
Sheet: 3

Operations



Focus



Discussion

- Explore sets of variants
- Meta data
- Pangenome embedding

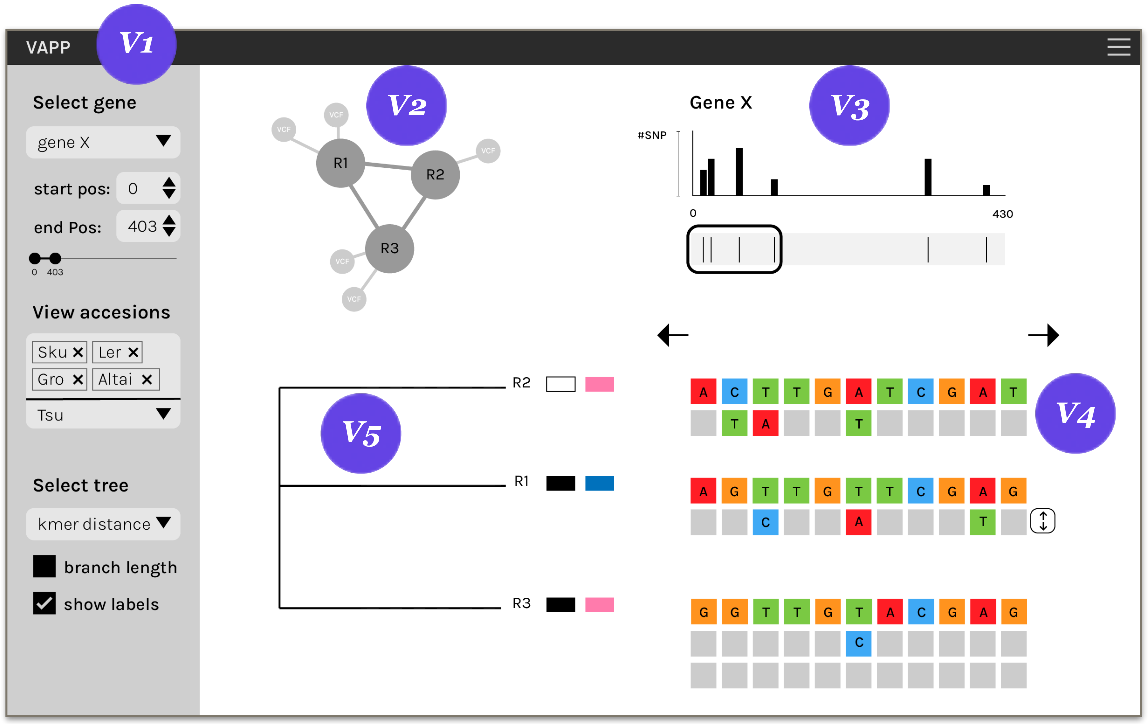
The third design sheet shows again two different ideas for the pangenome (data) and lines overviews. The sequences can now be selected by looking at the embedding, which includes phenotype information in addition to the sequence similarity.

The alignment view now only shows one sequence and its variants, which are selected by sorting and brushing the right most rectangle (bars show SNP freq). The upset plot below it shows further similarities between the lines.

Realization Design

Showing the first mockups

The realization design shows more concretely what the visualization tool may look like at current stage of development. The mockups serve as a guidance for prototype development. The design consists of five views: utility view (V1), pangenome overview (V2), gene variant overview (V3), gene variant alignment view (V4) and phylogeny view (V5).



Select gene: gene X

start pos: 0

end Pos: 403

0 403

View accesions: Sku × Ler × Gro × Altai ×

Tsu

Select tree: kmer distance

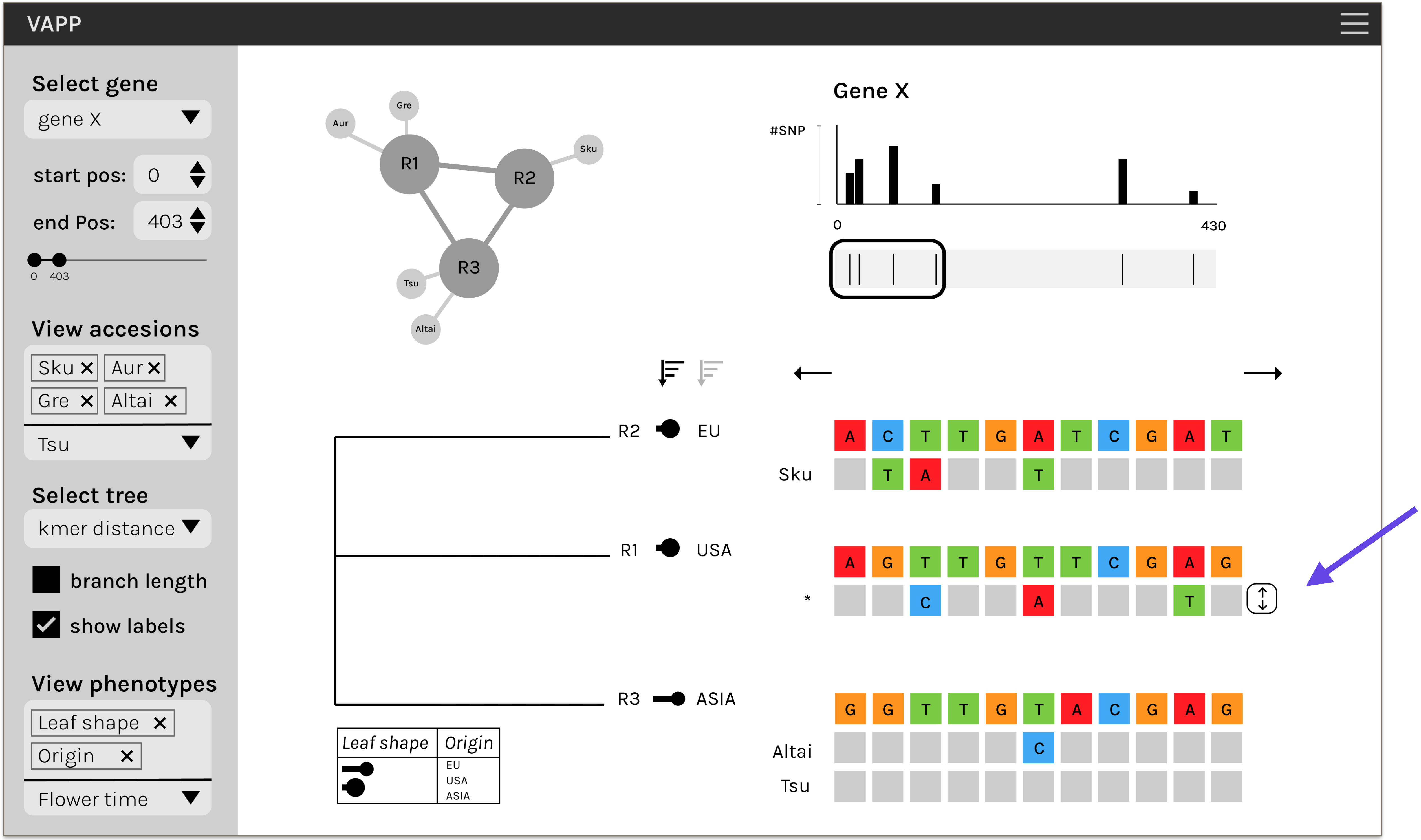
■ branch length

☑ show labels

Figure 5. Example mockup showing five views (V1-V5)

Realization design: compressed variant

● Define ● Brainstorm ● Design ● Develop ○ Validate



Select gene: gene X

start pos: 0

end Pos: 403

View accessions: Sku ✕ Aur ✕ Gre ✕ Altai ✕

Tsu

Select tree: kmer distance

■ branch length

☑ show labels

View phenotypes: Leaf shape ✕ Origin ✕

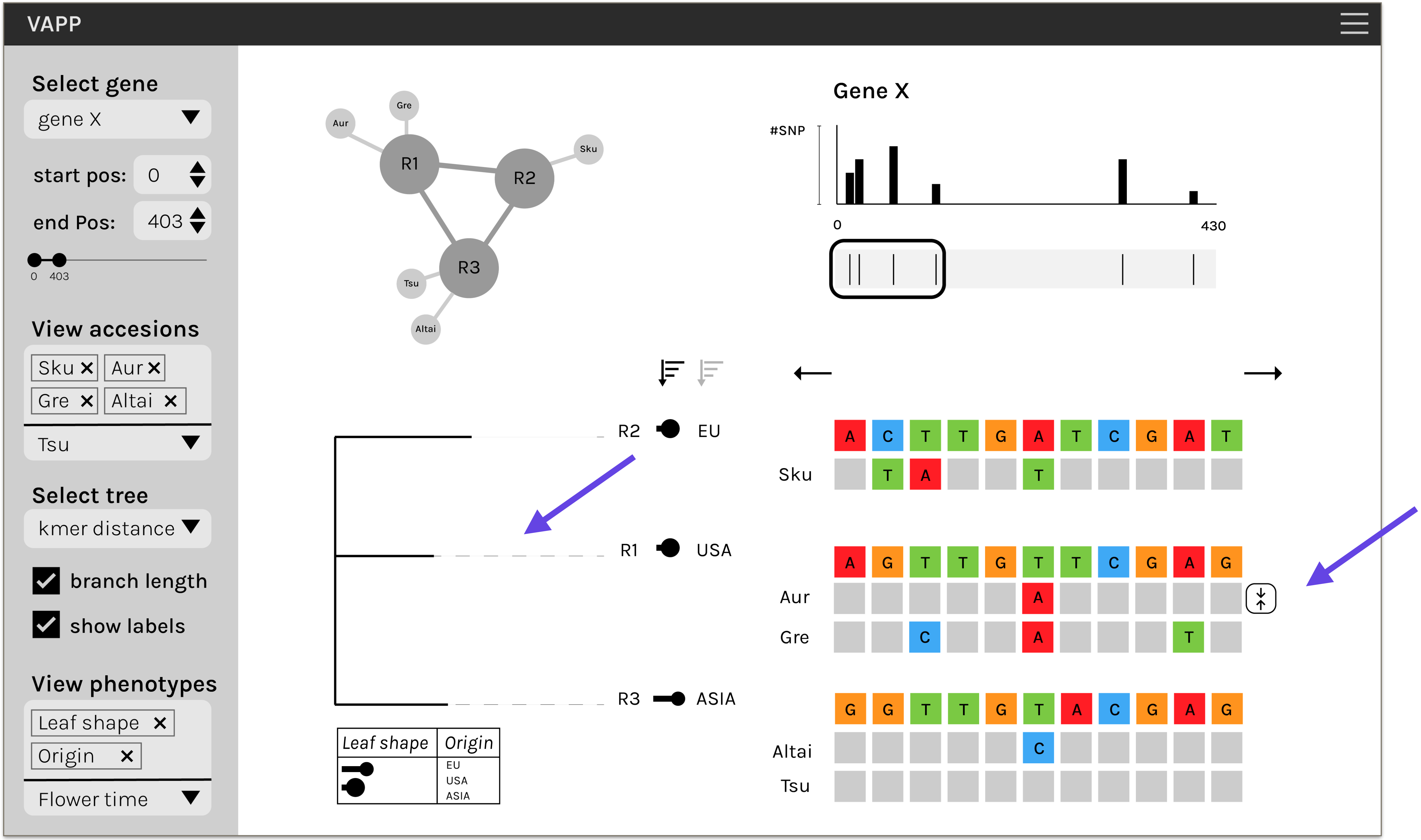
Flower time

	A	C	T	T	G	A	T	C	G	A	T
Sku		T	A			T					
	A	G	T	T	G	T	T	C	G	A	G
*			C			A				T	
	G	G	T	T	G	T	A	C	G	A	G
Altai						C					
Tsu											

Figure 6. Mockup showing a compressed variant representation for R1 (purple arrow)

Realization design: variants extended

Define — Brainstorm — Design — Develop — Validate



VAPP

Select gene: gene X

start pos: 0

end Pos: 403

View accessions: Sku ✕ Aur ✕ Gre ✕ Altai ✕ | Tsu

Select tree: kmer distance

☑ branch length

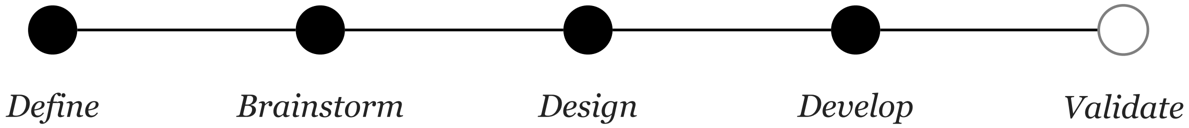
☑ show labels

View phenotypes: Leaf shape ✕ Origin ✕ | Flower time

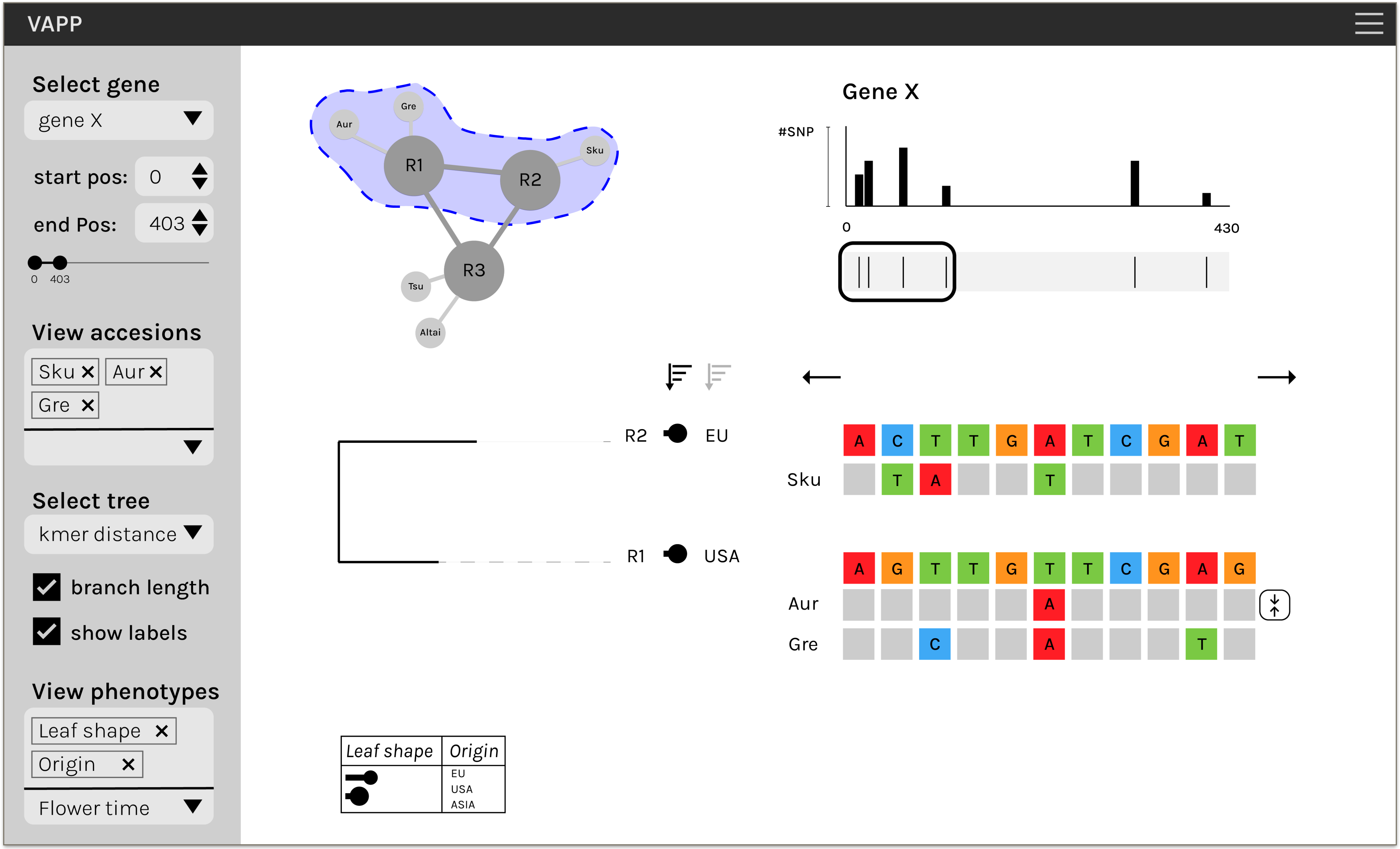
	A	C	T	T	G	A	T	C	G	A	T
Sku		T	A			T					
	A	G	T	T	G	T	T	C	G	A	G
Aur						A					
Gre			C			A				T	
	G	G	T	T	G	T	A	C	G	A	G
Altai						C					
Tsu											

Figure 7. Mockup showing a all variants for R1 and branch lengths for phylogenetic tree (purple arrows)

Realization design: select subset



Define — Brainstorm — Design — Develop — Validate



VAPP

Select gene: gene X

start pos: 0

end Pos: 403

View accessions: Sku ✕ Aur ✕ Gre ✕

Select tree: kmer distance

☑ branch length

☑ show labels

View phenotypes: Leaf shape ✕ Origin ✕

Flower time

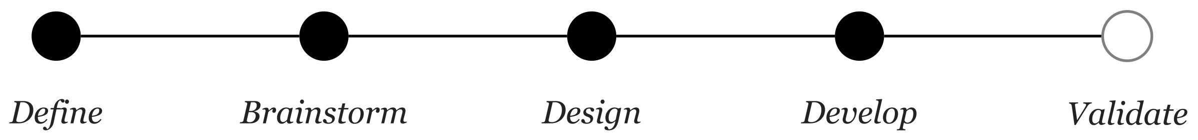
	A	C	T	T	G	A	T	C	G	A	T
Sku		T	A			T					

	A	G	T	T	G	T	T	C	G	A	G
Aur						A					
Gre			C			A				T	

Leaf shape	Origin
● ●	EU USA ASIA

Figure 8. Mockup showing subset selection (indicated by blue area in pangenome overview)

Included and future functionalities



Here we list the currently included **views**, **requirements** and future **functionalities** or unaddressed complexities for next iterations of the design. However, for some of the addressed requirements/functionalities only a basic version is established. Based on feedback sessions, the functionalities are continuously improved to deal with more complexity.

Views (V), Requirements (R) and upcoming functionalities (F)	Priority	T 1	T 2	T 3	T 4	Status	Complete
V1. Utility view: menu with filters and controls	High	x	x	x	x	In progress	100%
V2. Pangenome overview	High	x				In progress	100%
V3. Gene variant overview	High		x		x	In progress	100%
V4. Gene variant alignment view	High		x		x	In progress	100%
V5. Phylogeny view	High			x	x	Finished	100%
R1.Variant overview with multiple references	High		x			In progress	30%
R2. Multiple linked views	High		x		x	In progress	20%
R3. Flexible arrangement and linking of views	Low				x	Not started	0%
R4. Semantic filtering and slicing of the data	High	x	x	x	x	In progress	20%
R5. Focus and context: details on demand	High		x		x	In progress	10%
R6. Extract visualizations and alignment output	Medium					Not started	0%
F1. Design an encoding for showing InDels	Medium		x		x	Not started	0%
F2. Include information on amino acid changes				x	x		
F3. Integrate expression data				x	x		
F4. Show allelic diversity - heterozygosity		x	x		x		