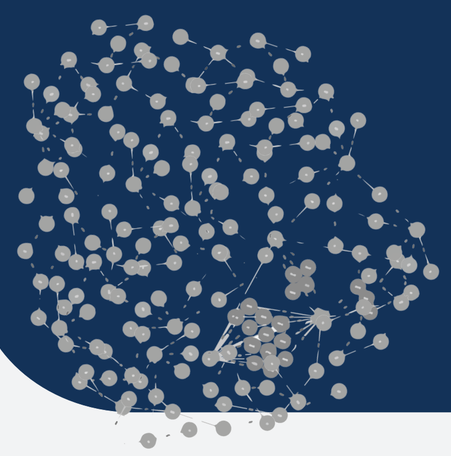
VISUAL EXPLORATION OF GENETIC SEQUENCE VARIANTS IN PANGENOMES

Astrid van den Brandt[1], Eef M. Jonkheer[2], Dirk-Jan M. van Workum[2], Sandra Smit[2] and Anna Vilanova[1]

[1]Eindhoven University of Technology, The Netherlands
[2]Bioinformatics Group, Wageningen University, The Netherlands



Motivation

Genome scientists increasingly use **pangenomes** to examine genetic variation underlying traits of interest.

Pangenomes are useful as they capture a species' full set of genetic material and **avoid bias** toward a single reference.

Pangenomes' size and complex data structure **hinder contextualization** & interpretation of analysis results.

Current visualizations fall short because they are created for **single references or do not show links to metadata**.

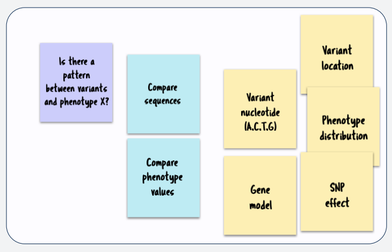
User Centered Design Process

Our collaborators want to interactively explore genetic sequence variants in pangenomes in the context of metadata.

Discovery

USER INTERVIEWS & OBSERVATION

VISUALIZATION OPPORTUNITIES WORKSHOP



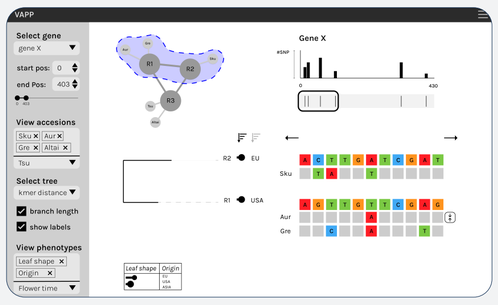
We defined a **shared workflow** to refine our goal into 4 **analysis tasks**.

Design

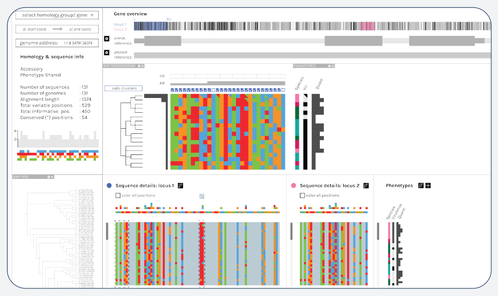
DESIGN SHEETS

We discussed **strengths and weaknesses** of several design sheets to inform the mockup designs.

MOCK-UPS V1



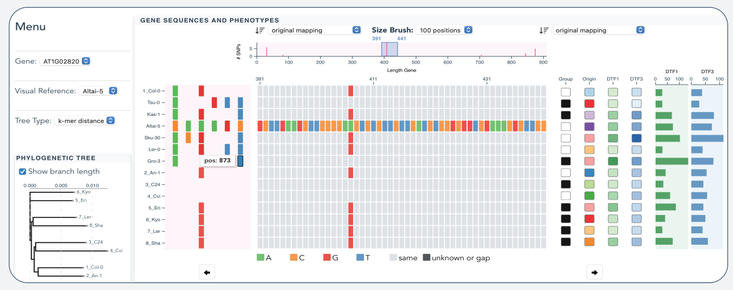
MOCK-UPS V2



Implement

With an online implementation of our early prototype we gather feedback and ideas over **a longer period of time**.

EARLY PROTOTYPE



Feedback & Review

TASKS

- **T1** Explore genomic sequence context of a gene
- **T2** Analyze relations between sequences, phylogeny & traits
- **T3** Define and analyze groups of similar samples
- **T4** Explore variant-trait associations within groups

Current Design

The design has **two main views** that use **established representations combined with linked interactions** such as sorting, clustering and selection with aggregation, which allow an overview+details exploration of variants in a gene.

1. The Gene Overview visualizes conservation across the whole gene and allows slicing an area for further inspection in the Locus View **T1**

2E. Filtering and sorting options

2F. Interactively calculate clustering by a selection of positions **T3**

2H. View MSA and metadata sorting linked to another dendrogram or tree (e.g., gene tree, core SNP tree of all genomes or, k-mer tree) derived from this pangenome to compare trees **T3**

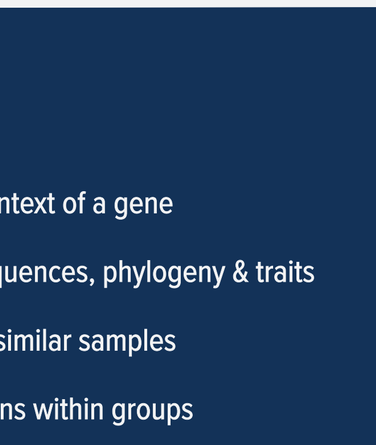
2. The **Locus View** contains four subviews: (A) a dendrogram or tree of the samples, (B) bipartite graphs connecting tree, sequence and trait orderings, (C) the multiple sequence alignment, (D) traits of interest. Interaction options (E, F, G,H).

2A. Phylogenetic tree or customized cluster dendrogram showing (evolutionary) relations **T2**

2B. Bipartite graph to connect various tree, sequence, and trait orderings **T2**

2C. Multiple sequence alignment (MSA) of the selected region by a Heatmap **T1**

2D. Traits of interest + sorting **T2** **T4**



2G. Samples can be collapsed through the inner tree nodes to fit more information on the screen (scale vertically)

Outlook

We aim to extend interactions by:

- Aggregation of samples with similar metadata
- Filtering by structural & functional annotations
- Side views with group statistics