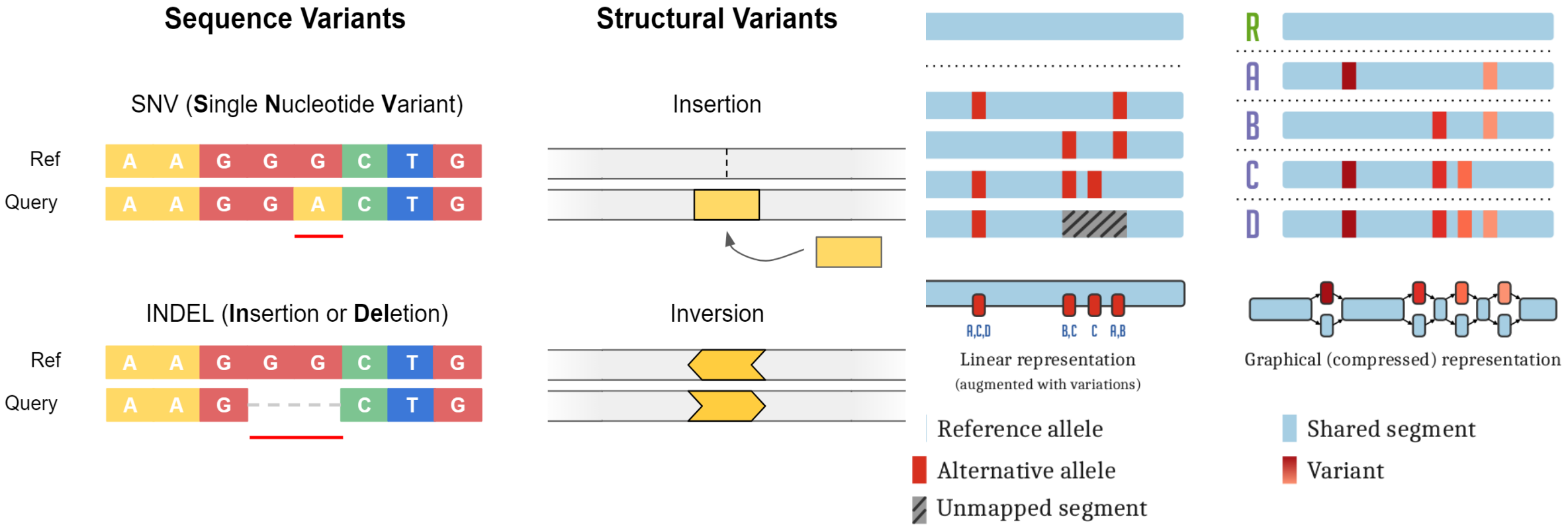
Agrobioinformatics Giessen

Linking genomic variation to gene expression using pangenome graphs

Gözde YILDIZ
Email: goezde.yildiz@agrar.uni-giessen.de

Why a pangenome graph?



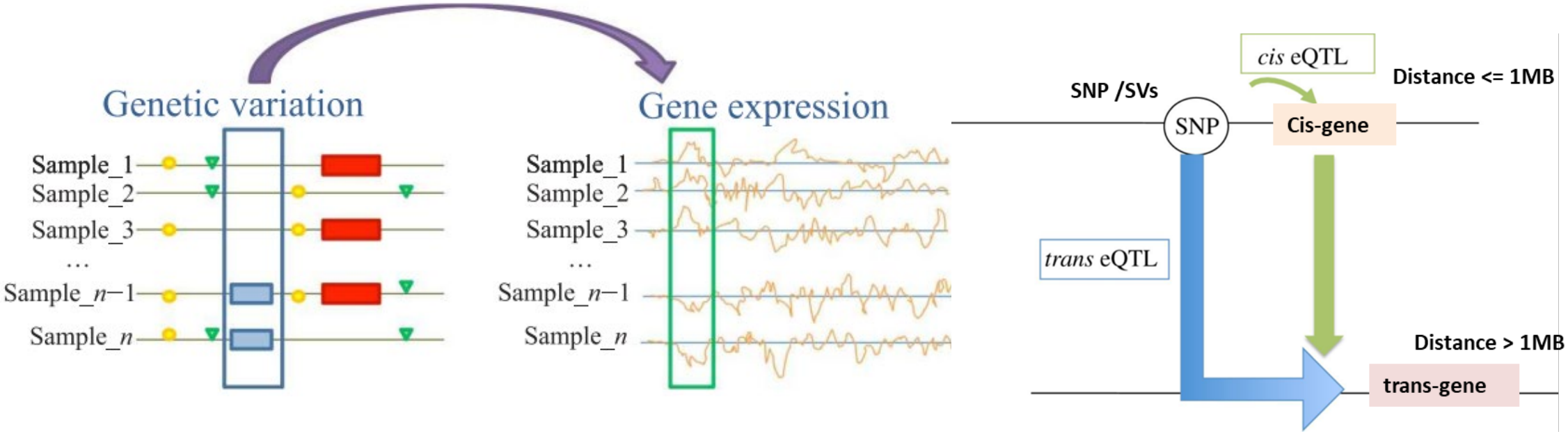
1. Encode **known variants** in a single data structure
2. **Avoid** linear **reference bias** during read mapping
3. **Improve** genotyping of SVs/SNPs

https://pangenome.github.io/

https://www.melbournebioinformatics.org.au/tutorials/tutorials/longread_sv_calling/longread_sv_calling/

INTRODUCTION

Expression quantitative trait locus (eQTL)

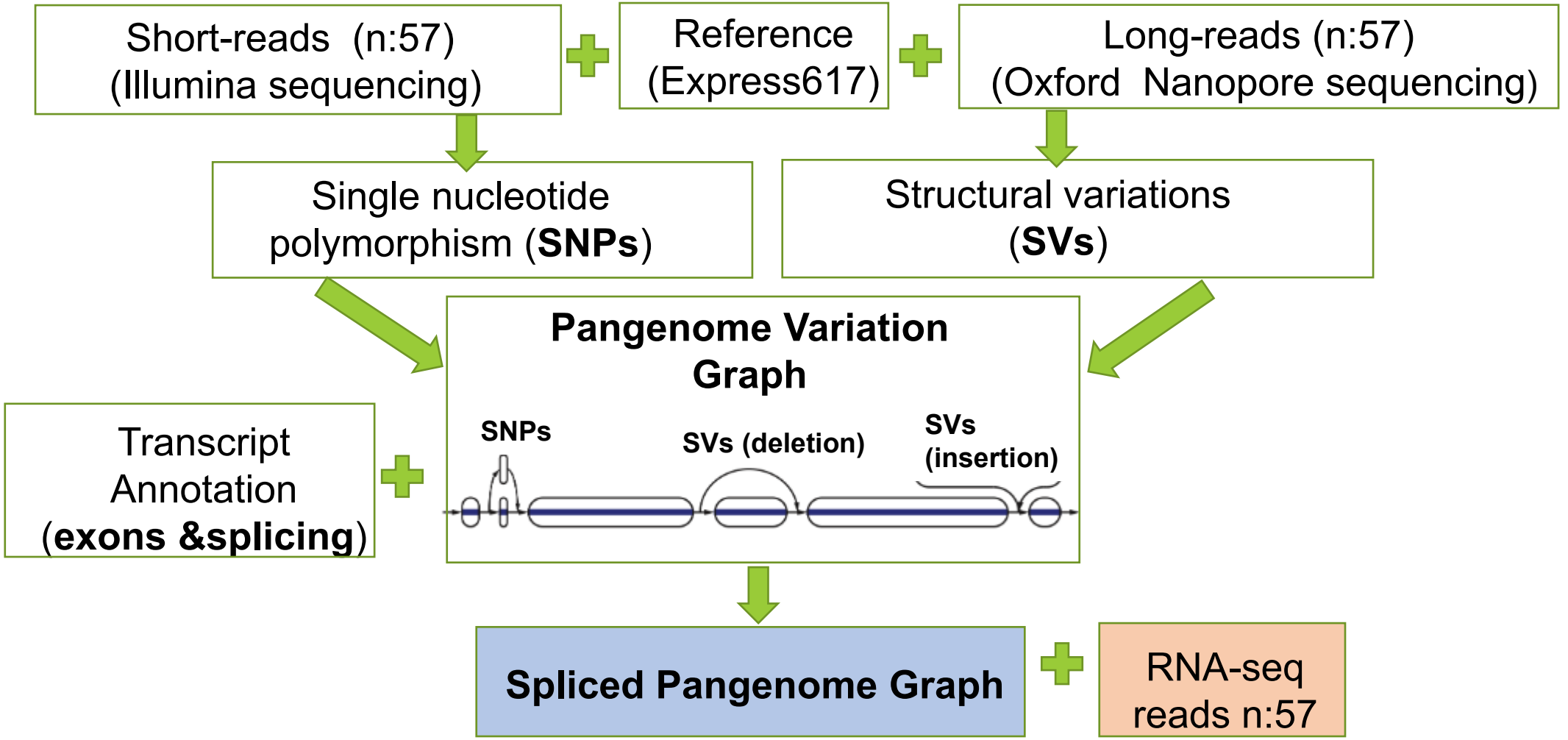


eQTL discovery from **linear-based** and **pangenome graph-based** approaches to assess the impact of SVs and SNPs on gene expression

eQTL – finds genomic variations **statistically** associated with biological traits

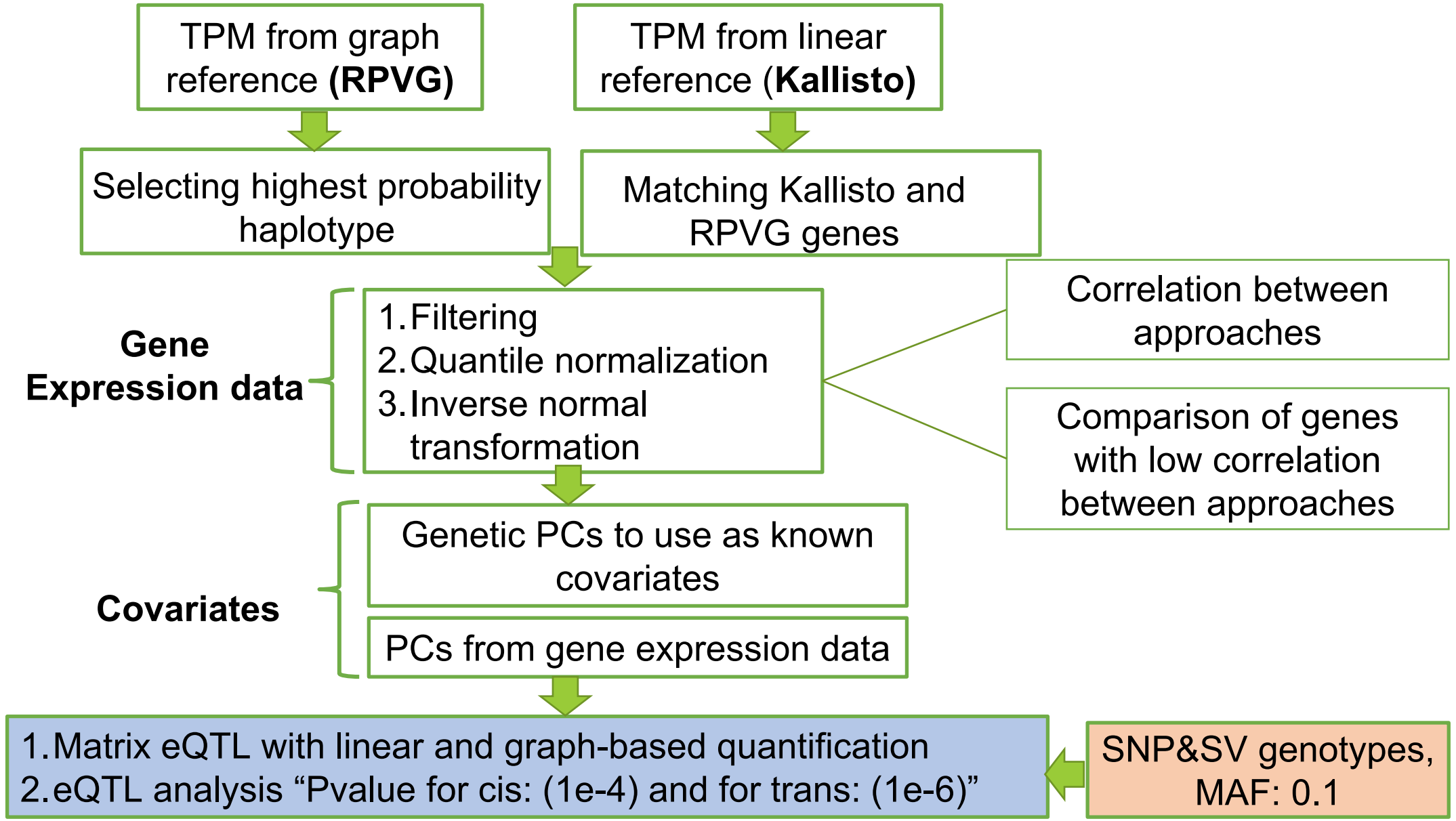
doi:10.1109/TST.2014.6961031

Building a variation graph for expression quantification



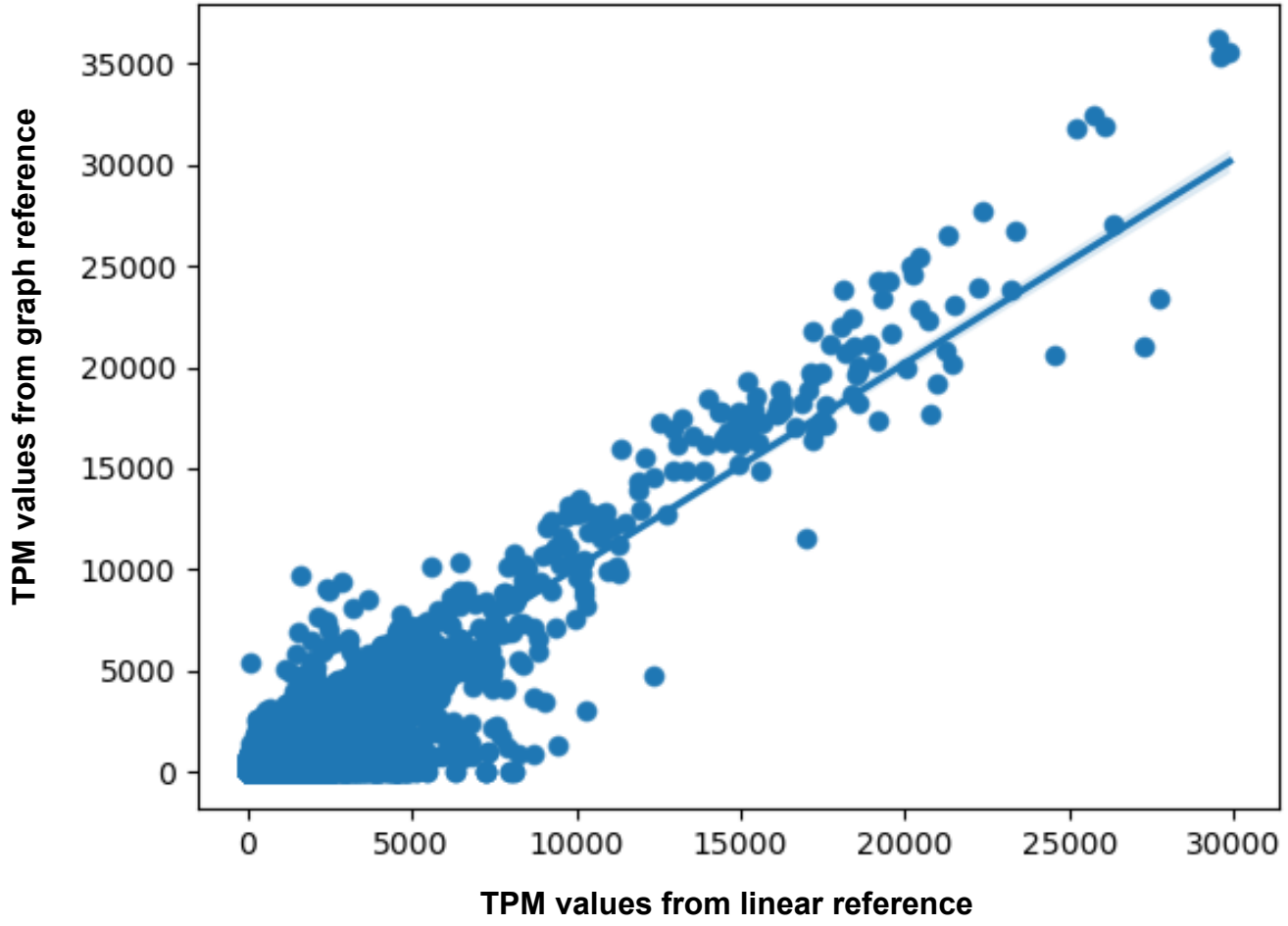
METHOD

Designing eQTL analysis



METHOD

TPM correlation between linear and pangenome graph references



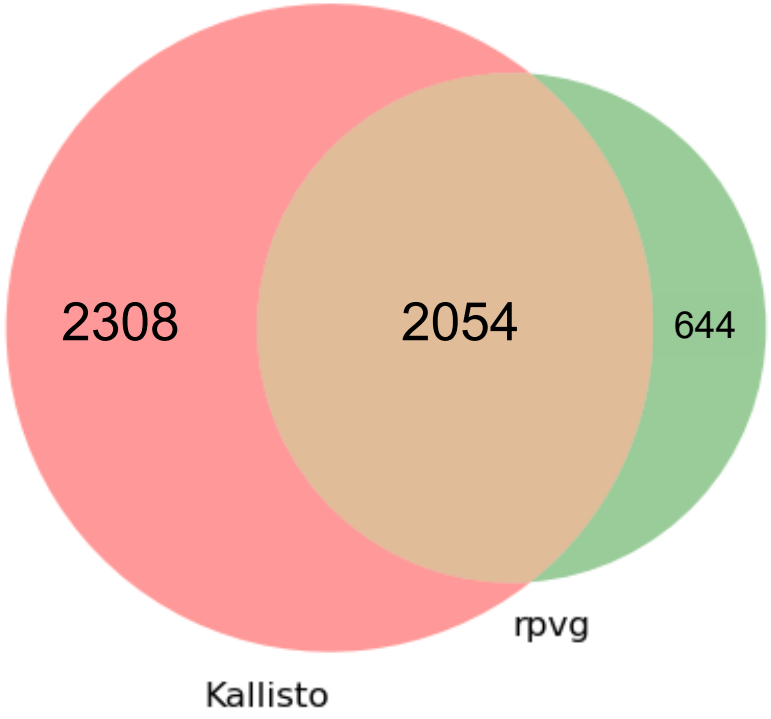
Comparison of gene expression (TPM) from graph (rpvg, highest probability haplotype) and linear (Kallisto) references

- **Pearson's r** : 0.938749954810782
- **Spearman's rho**: 0.912368168914421 4
- **Kendall's tau**: 0.809163904096016 2

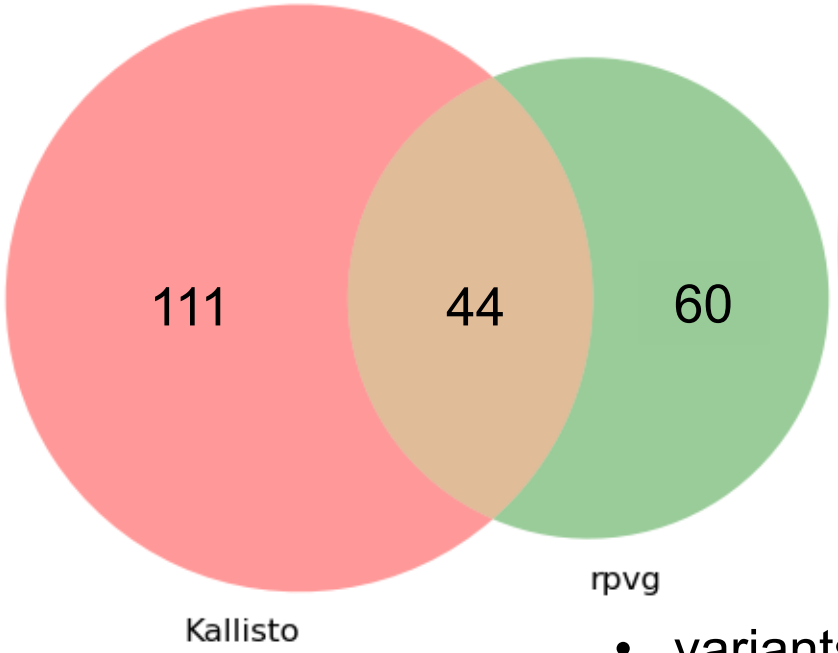
RESULTS

Comparison of genes with eQTL identified by two approaches

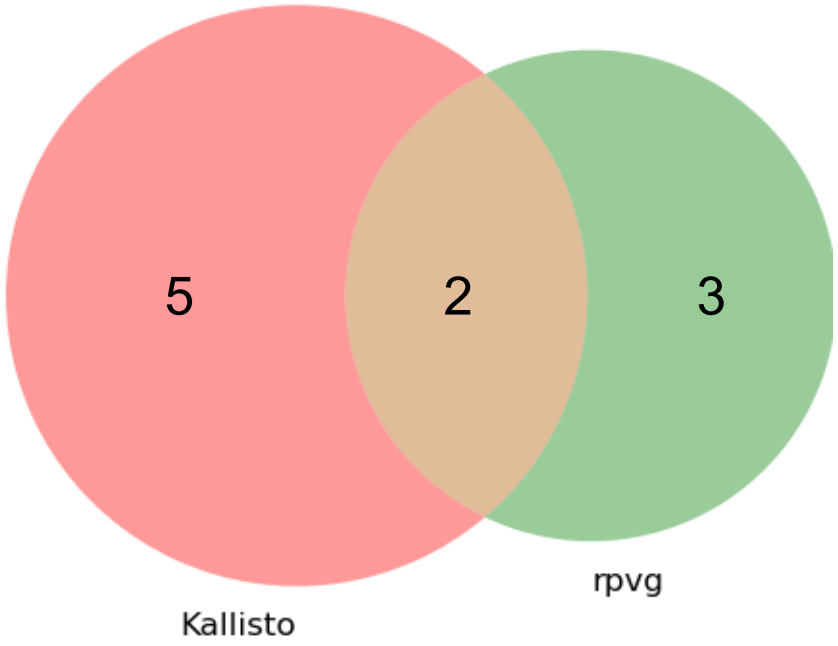
Genes with SNP eQTL only



Genes with SV eQTL only



Genes with SNP & SV eQTL



- variants with the lowest pvalues :

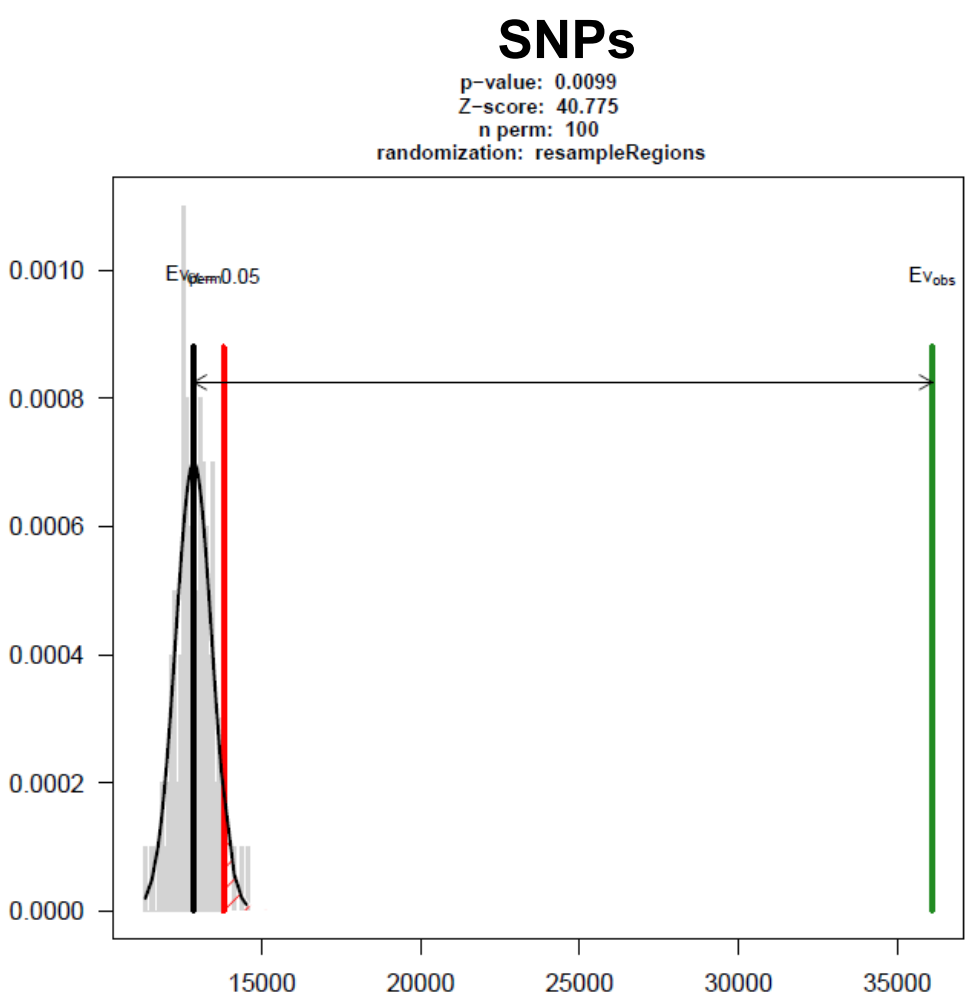
(4e-23) <= pvalues <= (1e-11) for Kallisto
(9e-33) <= pvalues <= (5e-12) for rpvg

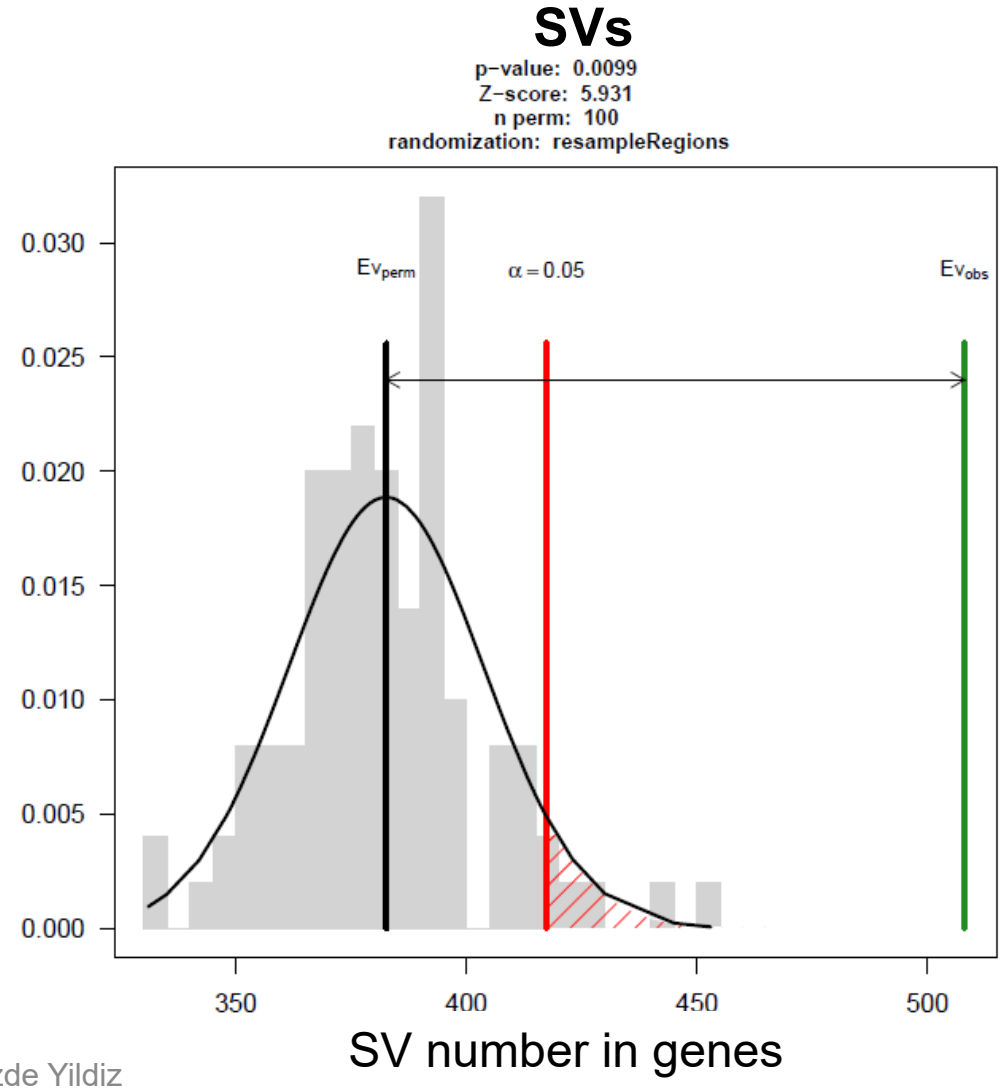
- FDR threshold : **1e-6**

RESULTS

Genes with eQTL found by linear approach only are overrepresented in SNPs & SVs

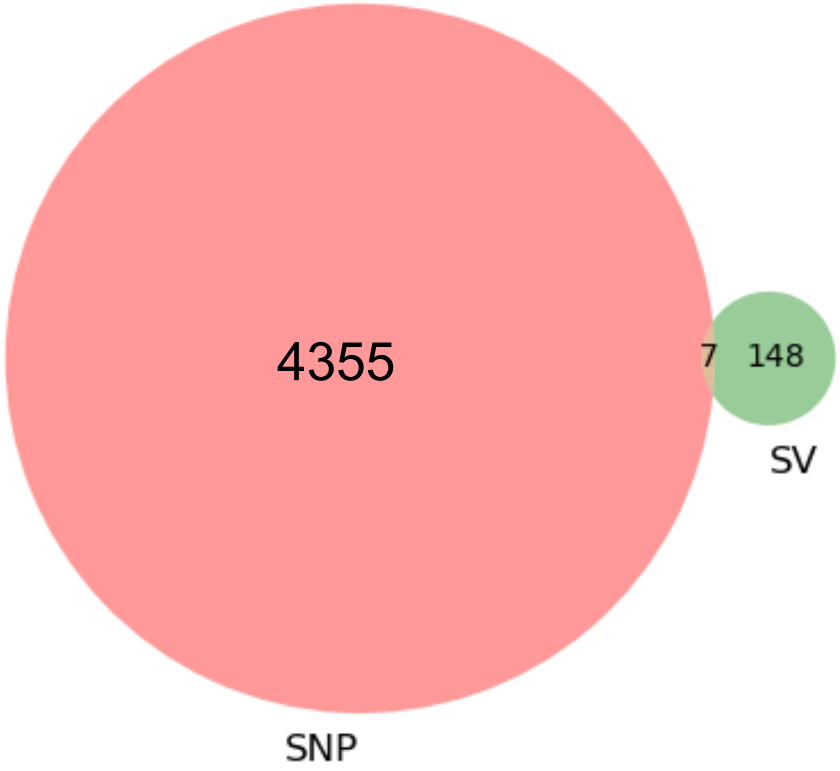
RESULTS



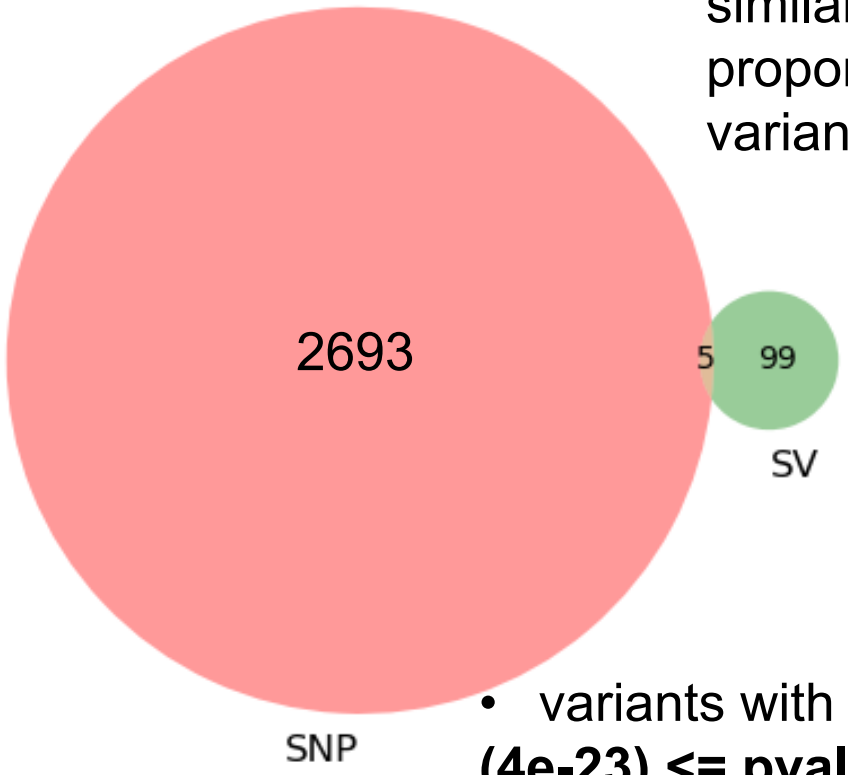


Number of genes with SNP and SV eQTL

Linear based-approach



Graph based-approach



Proportion of eQTL SVs is similar to the overall proportion of SVs among all variants **(~3.5%)**

- variants with the lowest pvalues :

(4e-23) <= pvalues <= (1e-11) for Kallisto
(9e-33) <= pvalues <= (5e-12) for rpvg

- FDR threshold : **1e-6**

RESULTS

Limitations and future work

- Limited samples size (57) allowing for detection of large effects only

- To expand dataset, SV genotyping from short reads can be performed (initial benchmarks suggest that ~50% of variants can be confidently genotyped)

- Functional analysis of eQTL genes under way

ONGOING WORKS

Acknowledgments

Agrobioinformatics group (JLU)

Dr. Agnieszka Golicz[1] (PI)

Dr. Silvia Zanini[2]

Jose Antonio Montero Tena[3]

Kevin Rockenbach[4]

Kübra Arslan[5]

Venkataramana Kopalli[6]

Rishi Srivastava[7]

Plant Breeding Department (JLU)

Prof. Dr. Rod Snowdon[8]

goezde.yildiz@agrar.uni-giessen.de

agnieszka.golicz@agrar.uni-giessen.de