A machine learning approach to identify regulatory SNPs based on genotyping data in *Vicia faba*

FELIX HEINRICH

Contents

1. *Vicia faba*
2. Regulatory SNPs
3. Identification of regulatory SNPs

Vicia faba

- Faba bean (*Vicia faba*) is a grain legume
- Globally grown as livestock feed as well as for human consumption
- Several agro-ecological advantages:
  - Nitrogen symbiosis
  - Diversifying crop rotations
  - Possible replacement for imported products like soybean
- Despite advantages restricted usage due to anti-nutrients vicine and convicine

Rasbak, CC BY-SA 3.0
(https://de.wikipedia.org/wiki/Ackerbohne)

beautifulcataya, flickr, CC BY-NC-ND 2.0
(www.pflanzen-lexikon.com/index.php?a=vicia-faba)

(Cooper et al., 2017; Khazaei et al., 2021)

Vicine and Convicine (V+C)

- Vicine and convicine are anti-nutrients occurring in the seeds of *Vicia faba*
- Negative effects on livestock as well as humans
- Breeding varieties with low V+C is a major area of research
- Responsible genes and mechanisms controlling V+C have been unknown for a long time
- Research is difficult
  - Reference genome for *Vicia faba* only available since 2023
  - V+C is (nearly) exclusive to *Vicia faba*

(Cooper et al., 2017; Khazaei et al., 2021; Björnsdotter et al., 2021)

Previous work on *Vicia faba*

- Sequence reads for 20 *Vicia faba* lines through genotyping by sequencing
- Assembly of a partial genome and variant calling
  - 685,215 SNPs
- GWAS to test association of SNPs with V+C
  - 2 SNPs showed very strong association
- Next : Identify regulatory SNPs

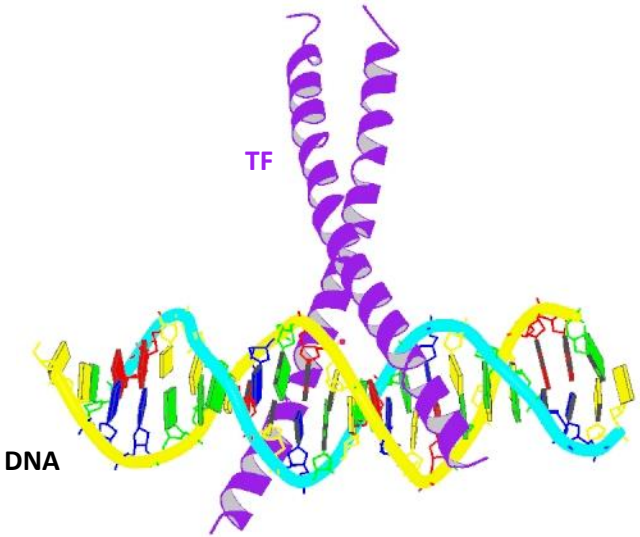
(Heinrich et al., 2020)

Regulatory SNPs (rSNPs)

- SNPs located in the non-coding, regulatory regions of genes
- Affect the phenotype through effects on the level of gene expression
- Interaction with different types of regulatory elements e.g. the binding of transcription factors
- Allow us to understand the underlying regulatory pathways leading to specific traits
- Here, we will focus on rSNPs that affect the binding of transcription factors

(Klees et al., 2021)

Transcription Factors (TFs)

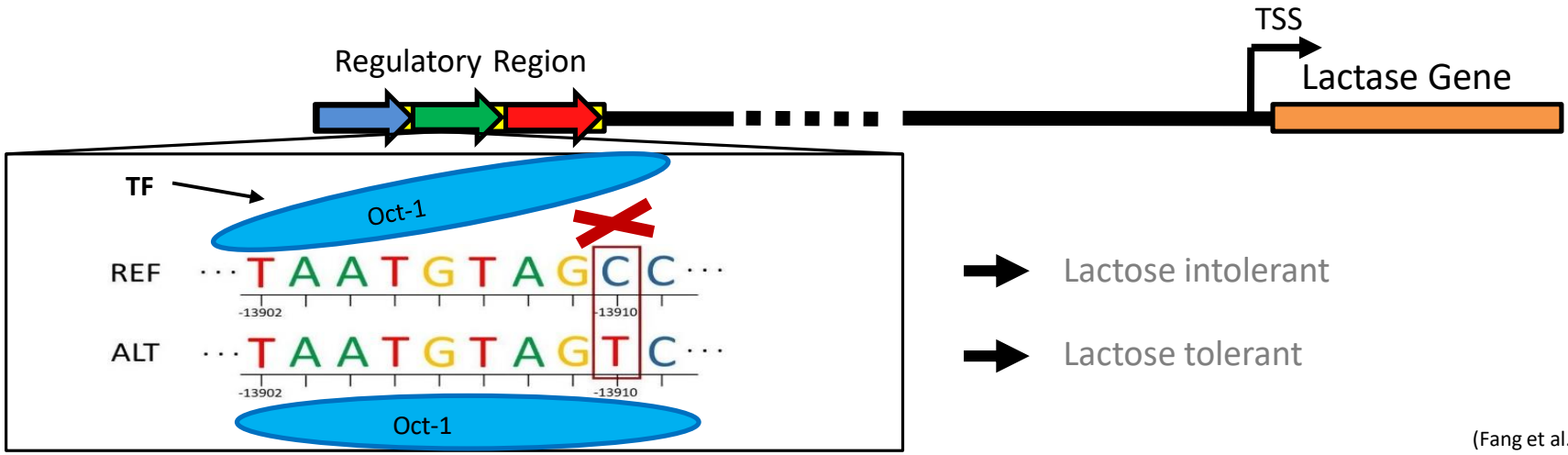


- TFs are a special class of gene regulatory proteins
- They bind to specific motifs in the DNA sequence, so called transcription factor binding sites
- TFs influence the transcription of a gene

(Maston et al., 2006)

Example for gene regulation through rSNP

- Most adults cannot metabolize lactose since they no longer produce lactase
- Those who can, have the alternative allele of a specific SNP
- The change in the DNA sequence allows the binding of a TF and with that the production of lactase



(Fang et al., 2012)

Identification of regulatory SNPs

1. Identify SNPs that are located in regulatory regions / promoters
2. Predict the transcription factor binding sites surrounding these SNPs for both alleles
3. If the binding sites differ with respect to the predicted binding affinity between the alleles, then the SNP is a rSNP

   ➢ Potential regulatory effect on gene expression

(Heinrich et al., 2020)

Identification of regulatory SNPs

0. **Classify genome sequences as promoter or non-promoter using a Convolutional Neural Network**

1. Identify SNPs that are located in regulatory regions / promoters

2. Predict the transcription factor binding sites surrounding these SNPs for both alleles

3. If the binding sites differ with respect to the predicted binding affinity between the alleles, then the SNP is a rSNP

   ➢ Potential regulatory effect on gene expression

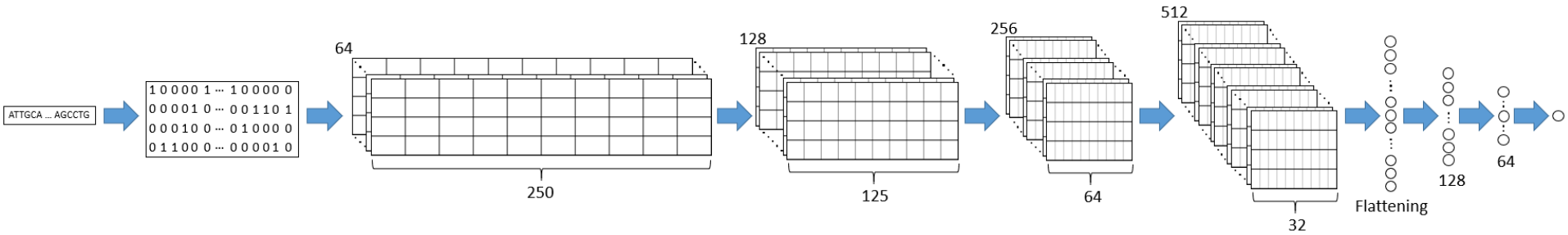
(Heinrich et al., 2020)

Convolutional Neural Networks (CNNs)

- Special type of neural networks
- DNA is encoded as 4-dimensional vector
- Multiple filter matrices are moved across the input
- At each position a value is calculated
- Suitable for data with temporal/spatial structures
- No manual feature preparation necessary

(Umarov et al., 2017)

CNN for Promoter Classification



- Input is a 250bp sequence
- Output is
  - 1 for promoter
  - 0 for non-promoter

(Heinrich et al., 2020)

Training the network

- Training requires sequences of known promoters and non-promoters
- *Vicia faba* did not have an annotated reference genome
  - ➢ No known promoters available

Training the network

- Training requires sequences of known promoters and non-promoters
- *Vicia faba* did not have an annotated reference genome
  - ➢ No known promoters available
- Promoters of closely related species share similar structures
  - Monocot plants have similar promoters
  - Dicot plants have similar promoters
- 7 species belonging to the Leguminosae family have annotated reference genomes

(Kumari et al., 2013)

Promoter classification across species

Evaluated / Trained	Glycine max	Lupinus angustifolius	Medicago truncatula	Phaseolus vulgaris
Glycine max	0.864	0.915	0.847	0.863
Lupinus angustifolius	0.611	0.981	0.720	0.586
Medicago truncatula	0.797	0.959	0.876	0.789
Phaseolus vulgaris	0.842	0.888	0.834	0.898

(Heinrich et al., 2020)

Promoter classification for *Vicia faba*

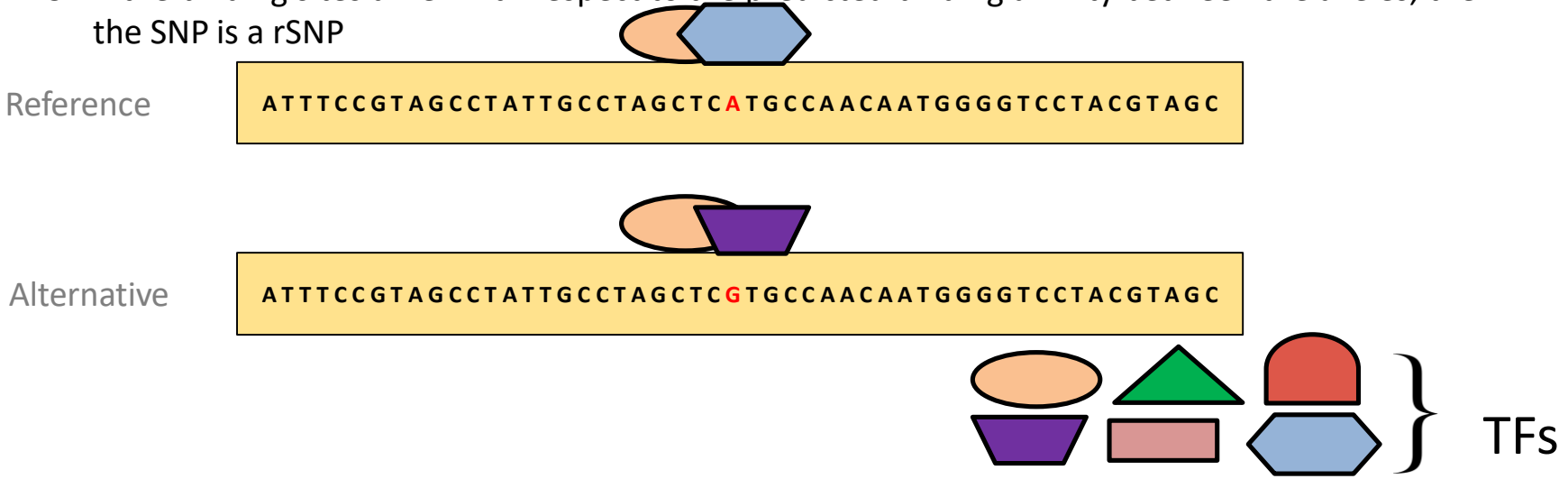
- Final model trained on
  - *Medicago truncatula*
  - *Lupinus angustifolius*
  - Additional negative sets
- Classification of the *Vicia faba* sequences
  - 2.46% of the sequences are classified as promoters
  - 19% of SNPs are located in promoters

(Heinrich et al., 2020)

Identification of regulatory SNPs

2. Predict the transcription factor binding sites surrounding these SNPs for both alleles

3. If the binding sites differ with respect to the predicted binding affinity between the alleles, then the SNP is a rSNP



Regulatory SNPs in *Vicia faba*

- 14 regulatory SNPs in region associated with V+C
- Both previously identified strongly associated SNPs are rSNPs
  - First SNP has a weak effect on TF binding
  - Second SNP has a very strong effect causing different TFs to bind
- TF which binding is disrupted by second SNP is related to seed coat development
  - Suggested site of biosynthesis for V+C

(Heinrich et al., 2020)

Summary

- Regulatory SNPs are an important aspect to understand the regulatory elements like TFs leading to specific traits
- Identification of rSNPs is difficult if an annotated reference genome is missing
- Known promoters/non-promoters from related species solve this problem
- Identified multiple rSNPs for *Vicia faba*
  - Strongly with V+C associated rSNP influences binding of TF related to potential site of V+C biosynthesis

Thank you for your attention!

Questions?

References

- Heinrich, Felix, et al. "Identification of regulatory SNPs associated with vicine and convicine content of Vicia faba based on genotyping by sequencing data using deep learning." Genes 11.6 (2020): 614.
- Klees, Selina, et al. "agReg-SNPdb: A Database of Regulatory SNPs for Agricultural Animal Species." Biology 10.8 (2021): 790.
- Umarov, Ramzan Kh, and Victor V. Solovyev. "Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks." PloS one 12.2 (2017): e0171410.
- Kumari, Sunita, and Doreen Ware. "Genome-wide computational prediction and analysis of core promoter elements across plant monocots and dicots." PloS one 8.10 (2013): e79011.
- Fang, Lin, et al. "The human lactase persistence-associated SNP− 13910* T enables in vivo functional persistence of lactase promoter–reporter transgene expression." Human genetics 131.7 (2012): 1153-1159.
- Maston, Glenn A., Sara K. Evans, and Michael R. Green. "Transcriptional regulatory elements in the human genome." Annu. Rev. Genomics Hum. Genet. 7 (2006): 29-59.
- Cooper, James W., et al. "Enhancing faba bean (Vicia faba L.) genome resources." Journal of Experimental Botany 68.8 (2017): 1941-1953.
- Khazaei, Hamid, et al. "Recent advances in faba bean genetic and genomic tools for crop improvement." Legume Science 3.3 (2021): e75.
- Björnsdotter, Emilie, et al. "VC1 catalyses a key step in the biosynthesis of vicine in faba bean." *Nature plants* 7.7 (2021): 923-931.