Big Biological Data (Analyses)
- NGS data

Alexander Platzer – February 2015

FNSNF

Schedule

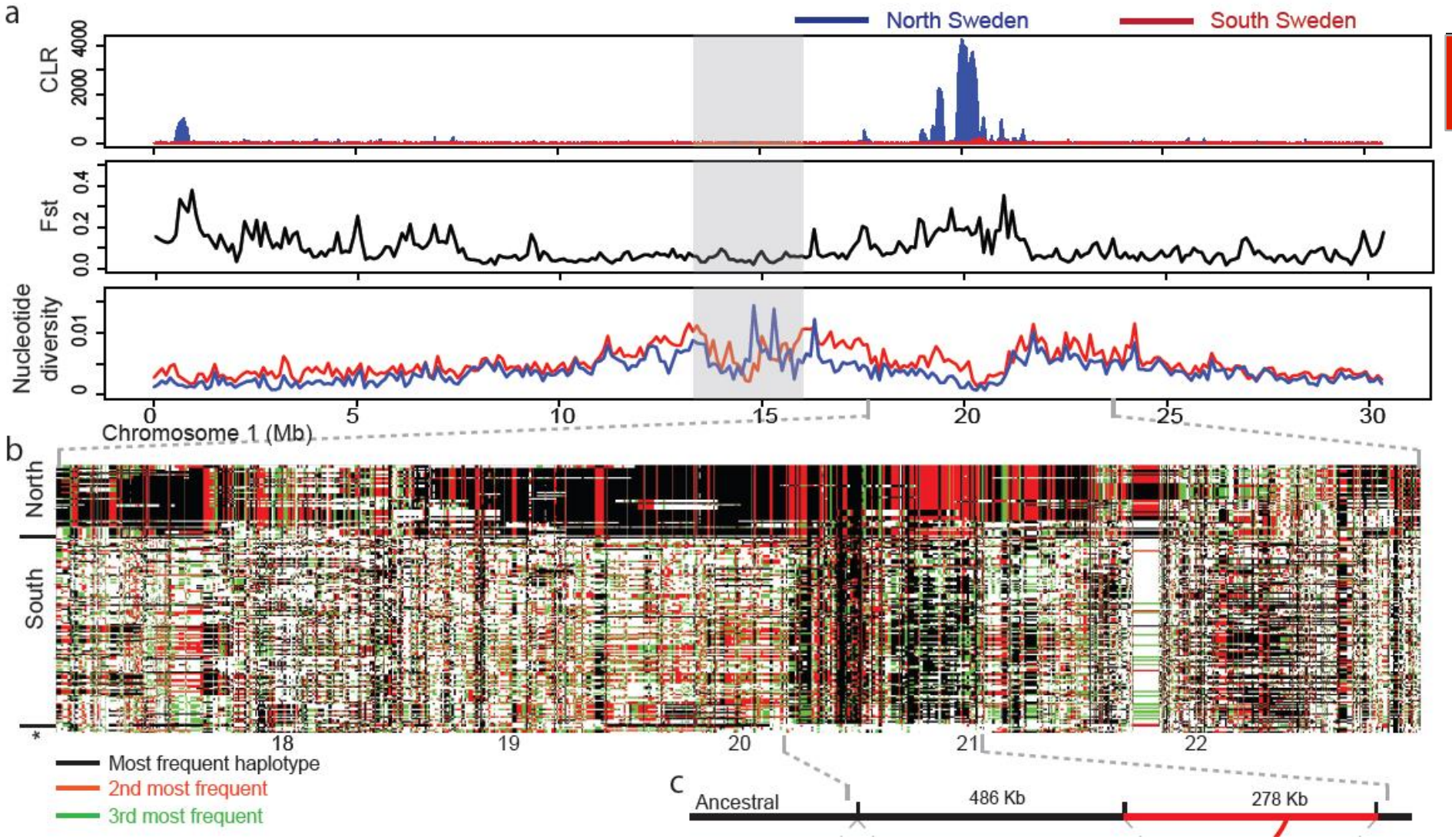
- 1001 Genomes Project
- Dimension reduction
- GWAS with transposable elements
- Indel pattern search
- Transcriptional Enhancement with Natural Antisense Transcripts

1001 Genomes Project

- More than 1000 full-genome sequenced Arabidopsis thaliana
- Related data (same accessions) from/with other projects:
  - RNAseq
  - Methylation
  - Phenotypes

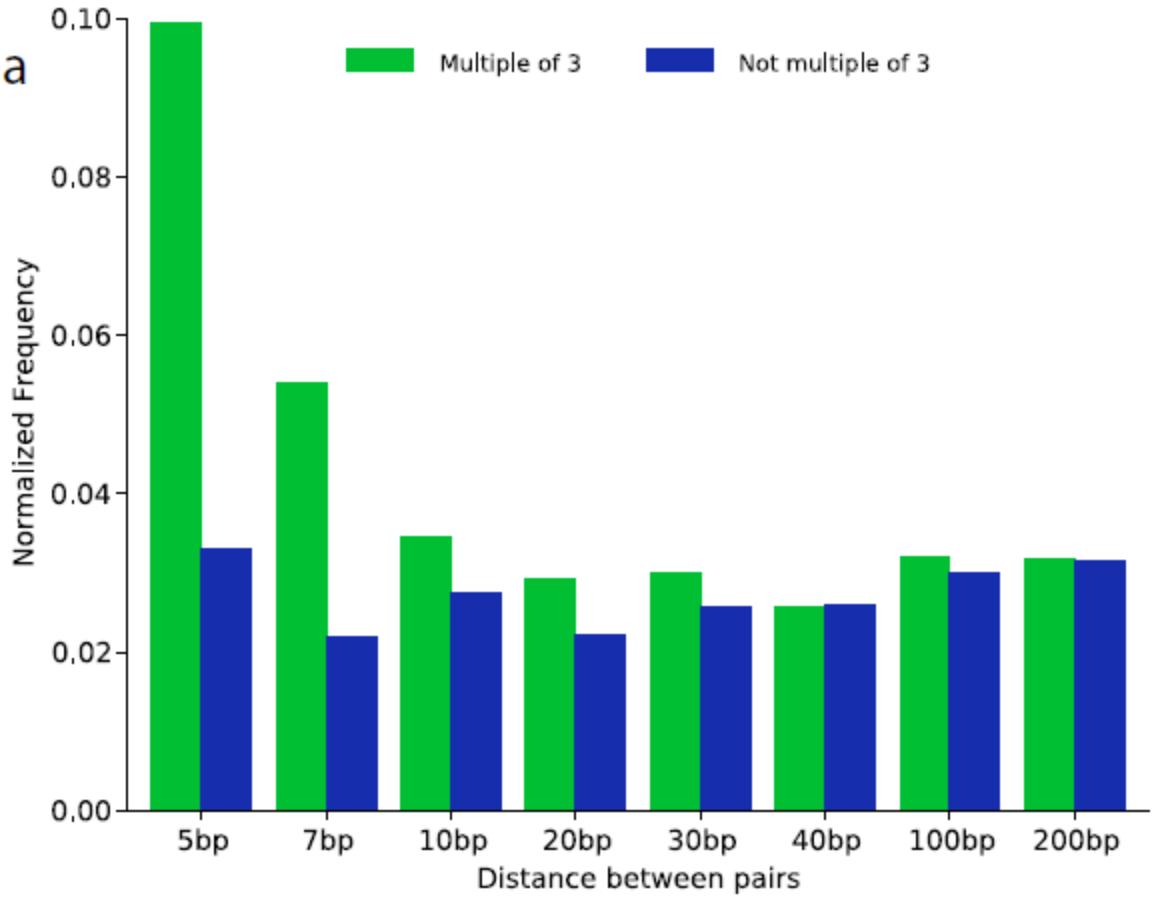
1001 Genomes Project

1001 Genomes Project

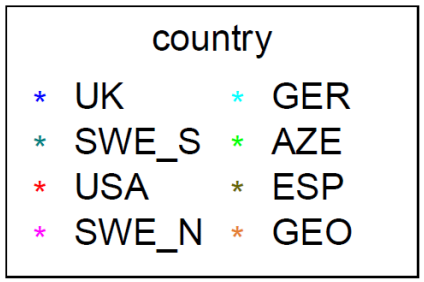


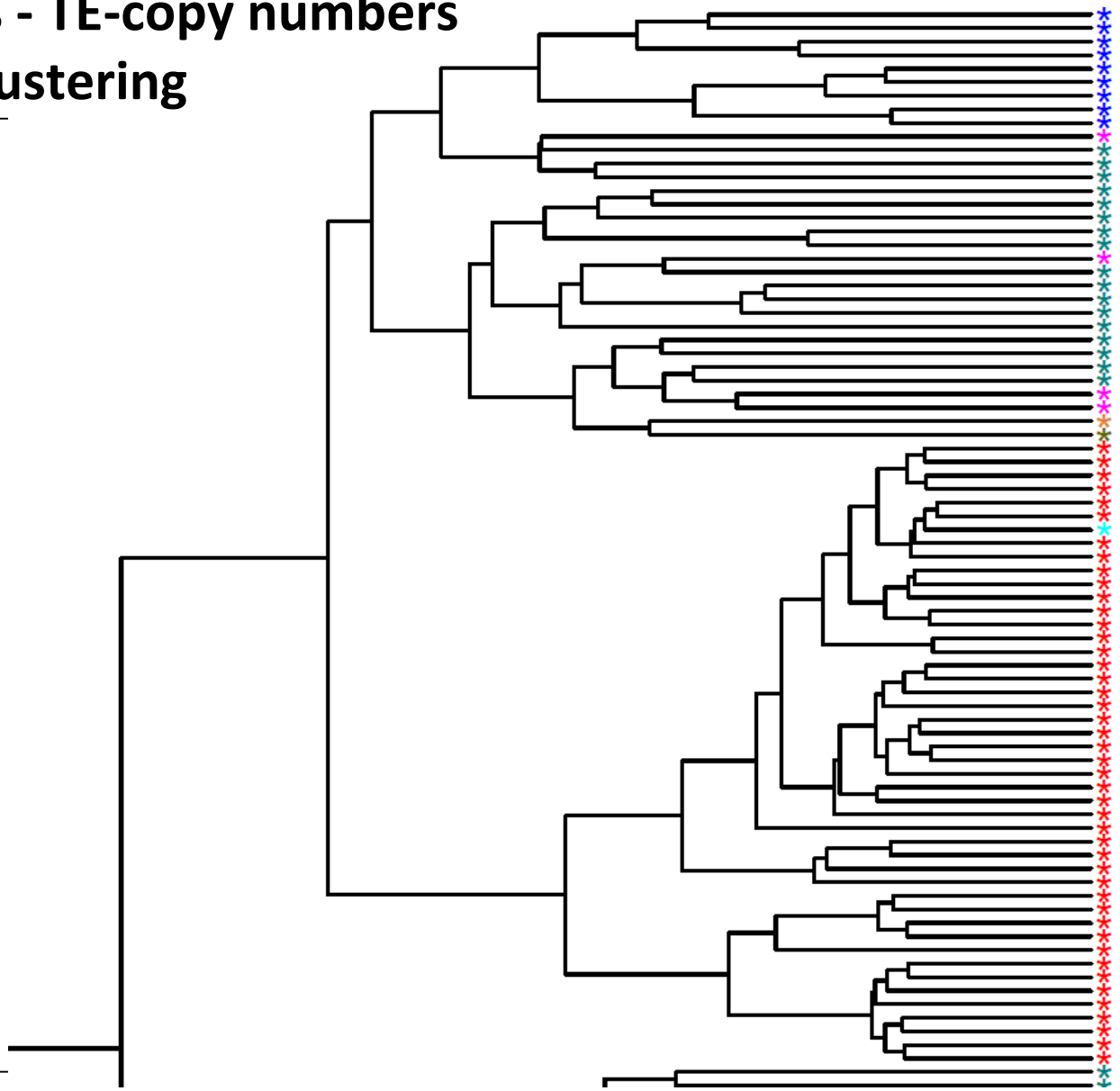
Long Q*, Rabanal FA*, Meng D*, Huber CD*, Farlow A*, **Platzer A**, Zhang Q, Vilhjalmsson BJ, Korte A, Nizhynska V, Voronin V, Korte P, Sedman L, Mandakova T, Lysak MA, Seren U, Hellmann I, Nordborg M. 2013. ***Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden***. Nature Genetics 45, 884–890

1001 Genomes Project

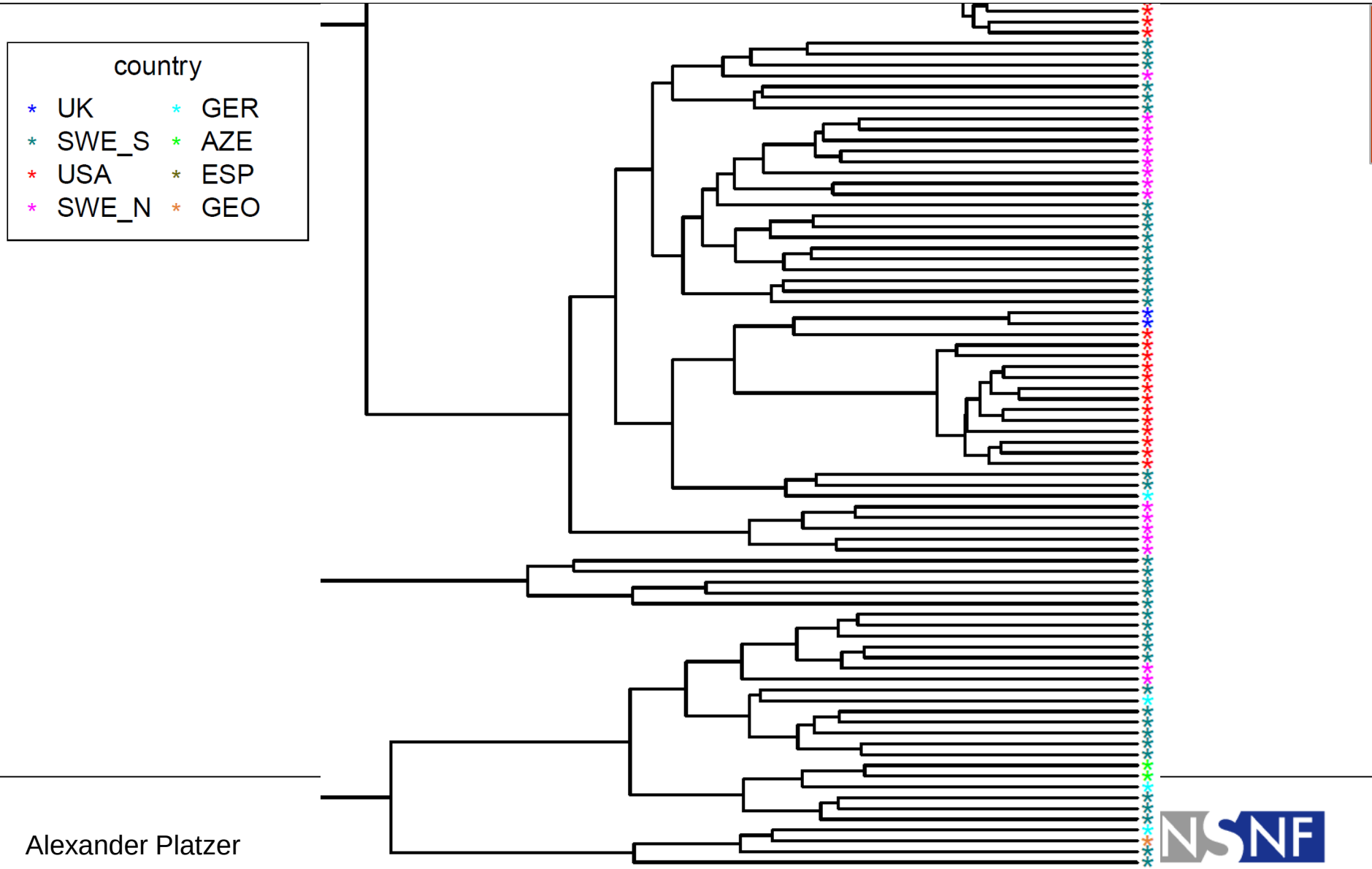


1001 genomes - TE-copy numbers
Hierarchical clustering





1001 genomes - TE-copy numbers
Hierarchical clustering



Visualization of SNPs with t-SNE

Dimension reduction

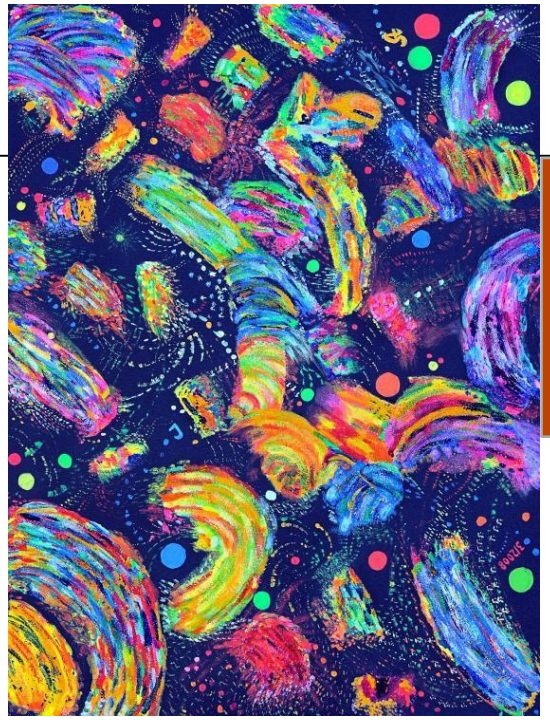
FNSNF

Problem definition - specific

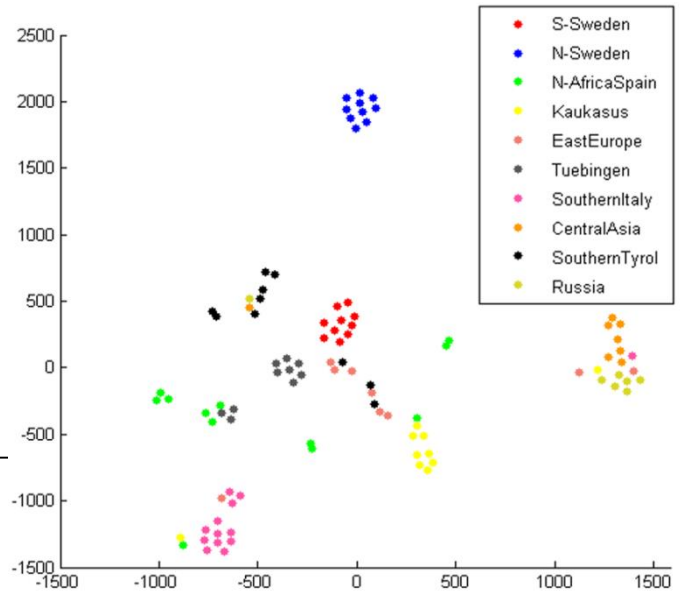
a bunch of binary/numerical variables
(a lot of SNPs)

one nominal label
(phenotype)

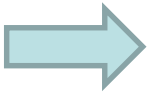
SouthernItaly
SouthernItaly
CentralAsia
CentralAsia
CentralAsia
CentralAsia
CentralAsia
CentralAsia
CentralAsia
CentralAsia
SouthernTyrol
SouthernTyrol
SouthernTyrol
SouthernTyrol
EastEurope
SouthernTyrol
SouthernTyrol
SouthernTyrol
SouthernTyrol
SouthernTyrol
EastEurope
EastEurope
EastEurope
N-AfricaSpain
N-AfricaSpain
Russia
Russia



deMim: Das lachende Wühaha



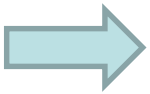
Problem definition - general

a huge amount of | numerical
ordinal
nominal | dimensions  2 or 3 numerical dimensions

(for plotting)

+ maybe 1-some nominal effect dimensions

Problem definition - general

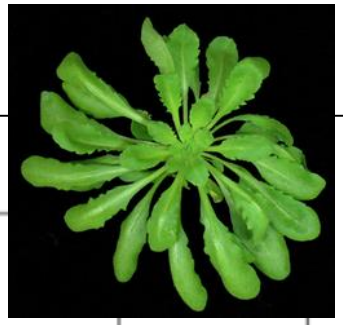
a huge amount of | numerical
ordinal
nominal | dimensions  2 or 3 numerical dimensions

... there is a subfield ‘dimension reduction’ for this.

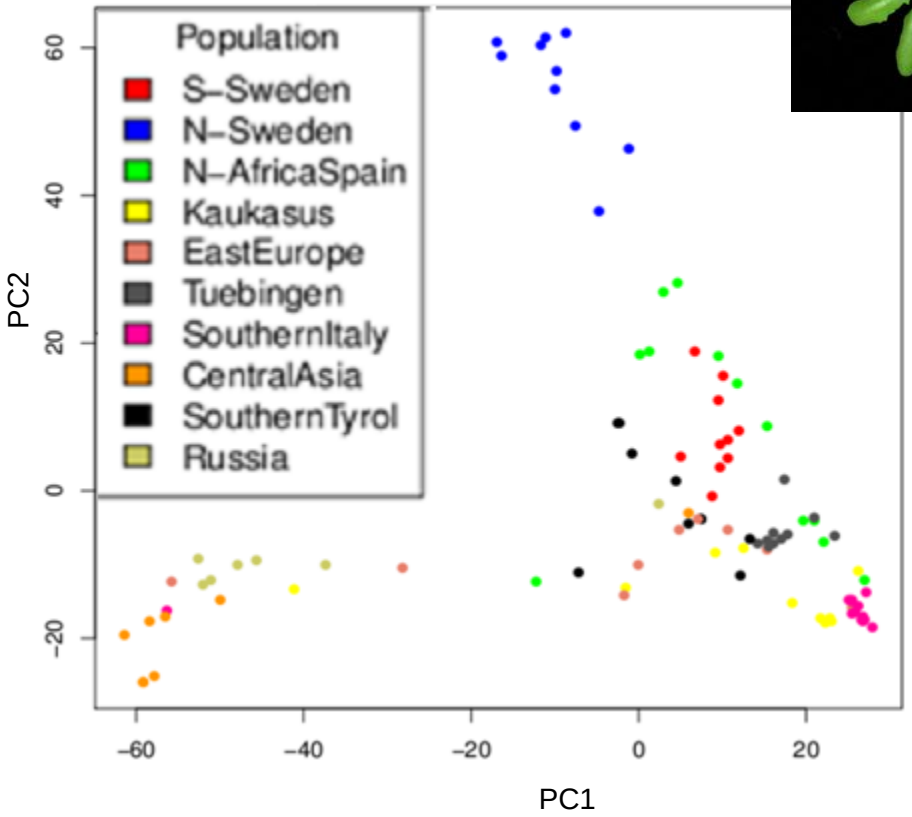
Beside the widely used PCA, the alternatives are e.g.:

Sammon mapping, Isomap, Locally Linear Embedding, Classical multidimensional scaling, Laplacian Eigenmap, m-SNE, t-SNE, ...

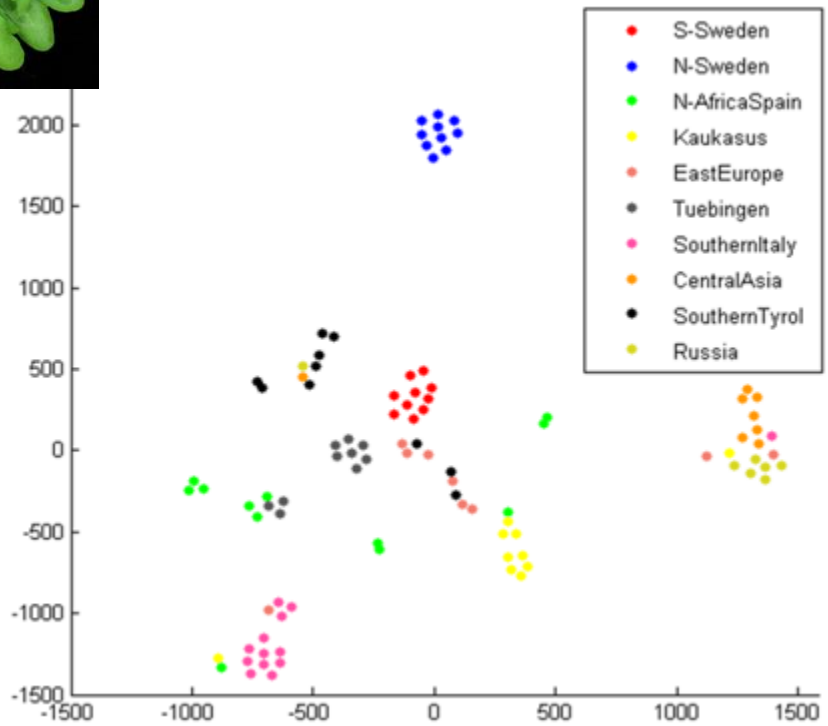
Visualizations



PCA



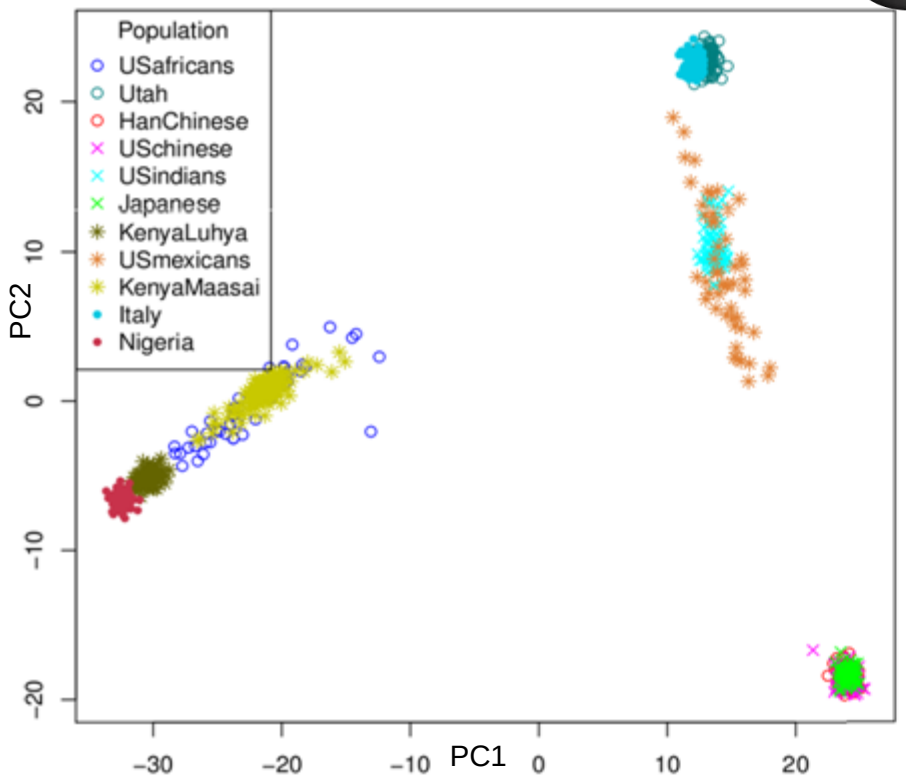
t-SNE



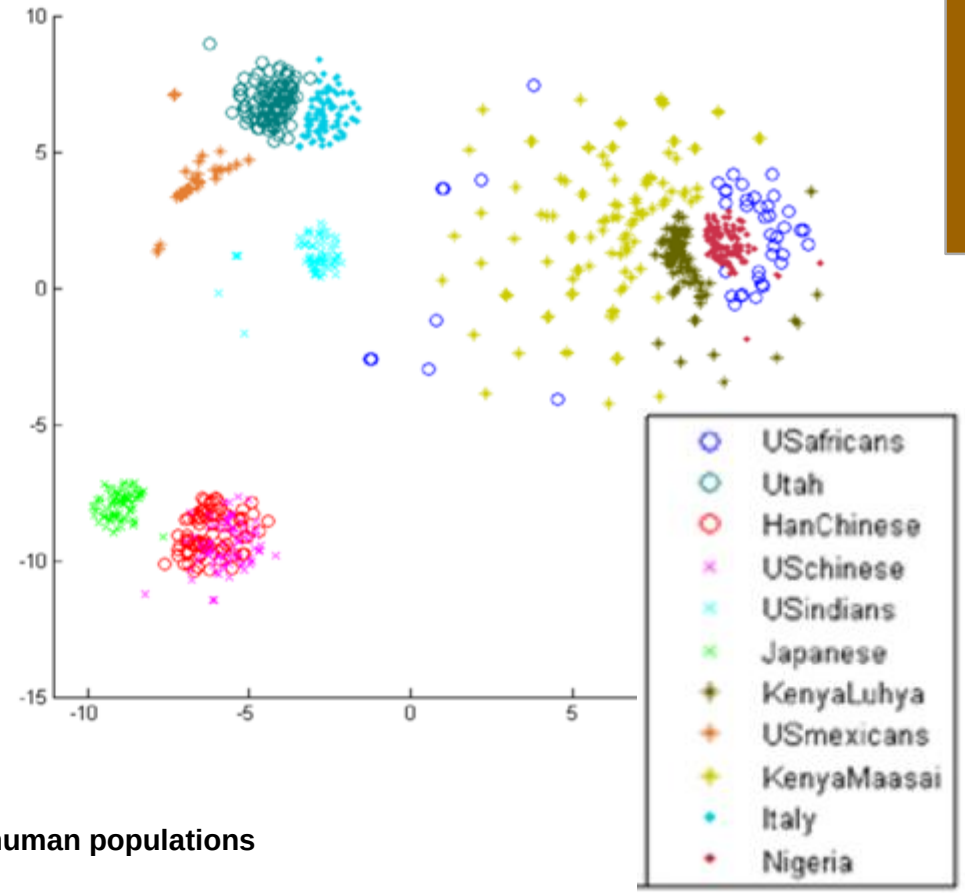
data from:
Long Q*, Rabanal FA*, Meng D*, Huber CD*, Farlow A*, **Platzer A**, Zhang Q, Vilhjalmsson BJ, Korte A, Nizhynska V, Voronin V, Korte P, Sedman L, Mandakova T, Lysak MA, Seren U, Hellmann I, Nordborg M. 2013. ***Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden***. Nature Genetics 45, 884–890

Visualizations

PCA



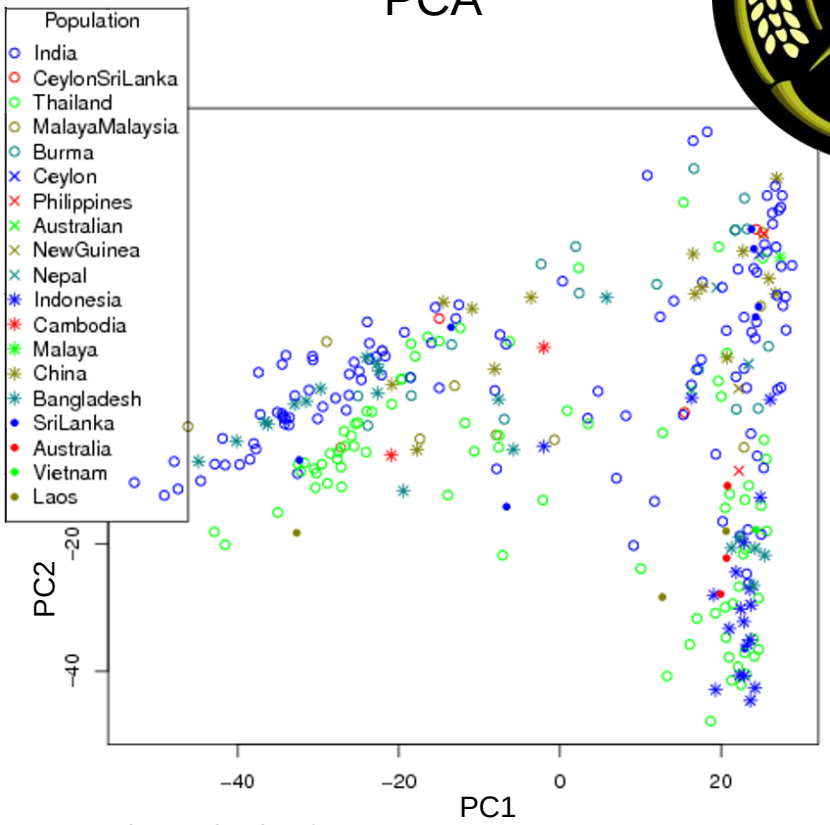
t-SNE



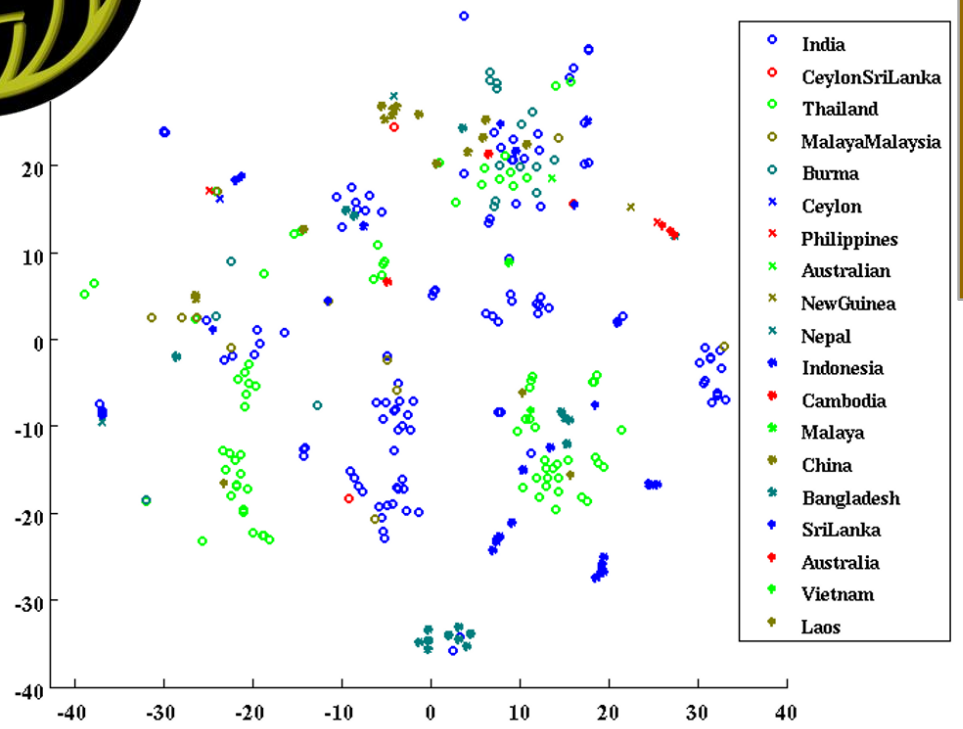
data selection from
The International **HapMap** Consortium.
Integrating common and rare genetic variation in diverse human populations
Nature 467, 52-58. 2010.

Visualizations

PCA

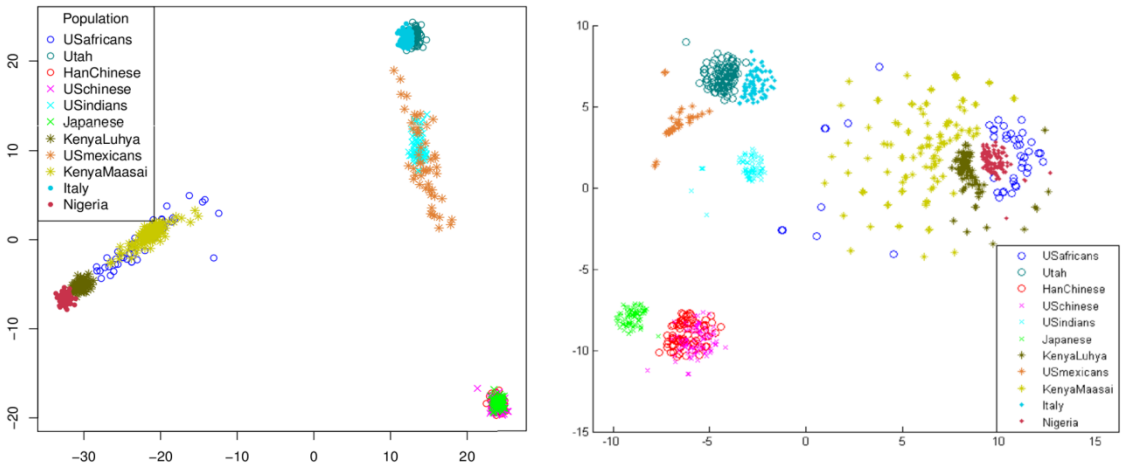


t-SNE



data selection from
Huang X, Kurata N, Wei X, Wang ZX, Wang A, et al. (2012)
A map of rice genome variation reveals the origin of cultivated rice.
Nature 490: 497-501

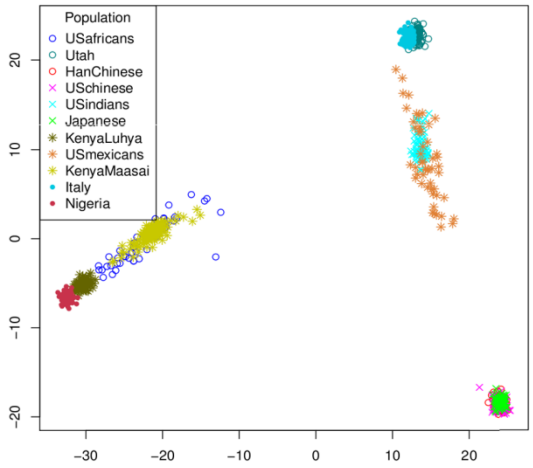
Measurements - Question

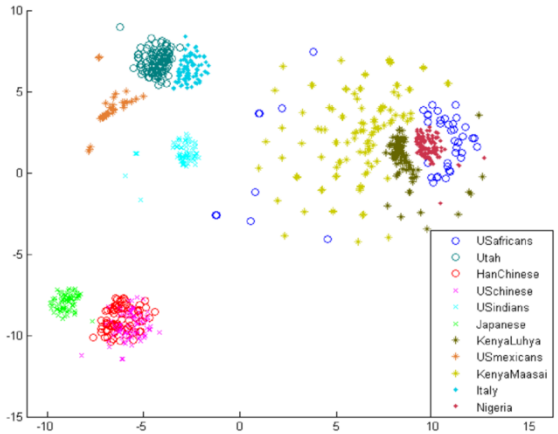


- how well is the data structured?
- *how much (correct) insight can be obtained from it?*

Measurements - Structuredness - As classification problem

transformed data / diagrams





Coordinates

classification problem

7.0717	15.357	Burma
-9.5674	14.878	Bangladesh
15.314	-12.017	Bangladesh
14.705	-8.3615	Bangladesh
3.1356	-34.54	Bangladesh
15.119	-9.0675	Bangladesh
15.634	-9.2958	Bangladesh
27.437	11.955	Bangladesh
-0.34996	-33.382	Bangladesh
-28.565	-1.9802	Bangladesh
-0.30935	-35.71	Bangladesh

C4.5, PART, Perceptron, Naive Bayes

Measurements - As classification problem

1001 genomes project			**RegMap**		**hapmap3 r2**		**hapmap3 r3**		**Rice**	
10x cross-val.%	*PCA*	*t-SNE*	*PCA*	*t-SNE*	*PCA*	*t-SNE*	*PCA*	*t-SNE*	*PCA*	*t-SNE*
C4.5	55.6	72.7	79.2	89.7	72.9	90.5	72.9	87.5	41.3	66.6
PART	60.6	76.8	77.6	89.1	72.7	90.9	73.3	87.6	39.7	64.9
Perceptron	67.7	76.8	80.7	85.8	70.3	85.1	72.2	84.8	50.5	56.4
Naive Bayes	62.6	75.8	75.2	80.3	74.6	87.2	71.8	84.1	40.7	42.3
Mean diff.	13.9		8.1		15.8		13.5		14.5	
St.dev.	3.6		3.4		2.6		1.2		12.5	

Platzer, A., *Visualization of SNPs with t-SNE*. **PLoS One, 2013. 8(2): p. e56883.**

Transposable Elements

FNSNF

TE-locate

TE-Locate: A Tool to Locate and Group Transposable Element Occurrences Using Next-generation Sequencing Data

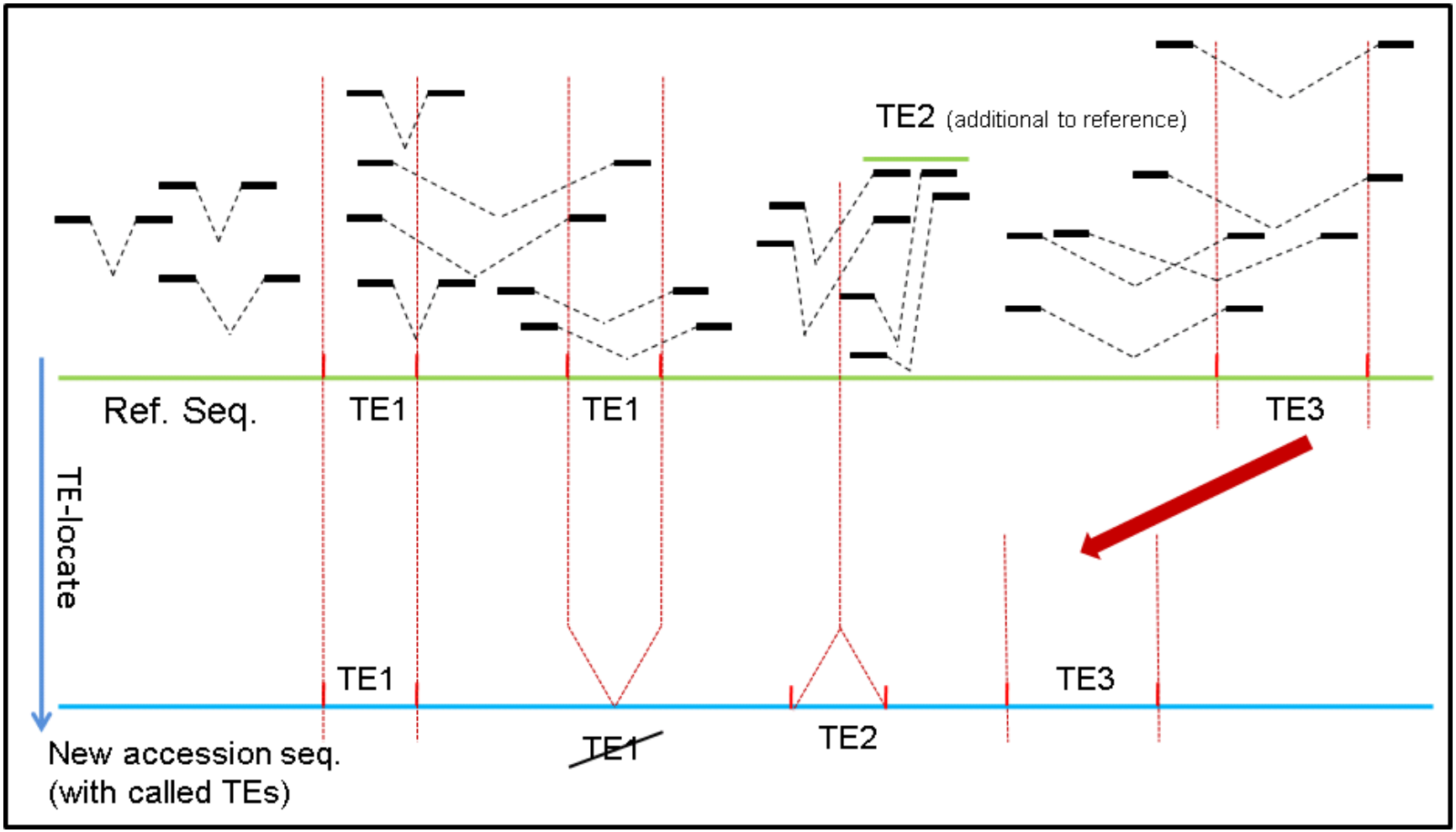
Input:

- (Read pairs)
- A reference genome
- Annotated TE elements therein
- SAM (dependency: A aligner)
- Optional: Mapping from annotated TEs to a higher hierarchical level)

TE – Hierarchy
(as in The Gypsy Database (GyDB) of Mobile Genetic Elements)

- Superfamily (e.g. LTR/Gypsy)
  - Systems (e.g. LTR retroelements)
    - Families (Ty1/Copia)
      - Elements (Hydra1-1)
        - Annotated reference loci (AT1TE09970)

TE-Locate : Method

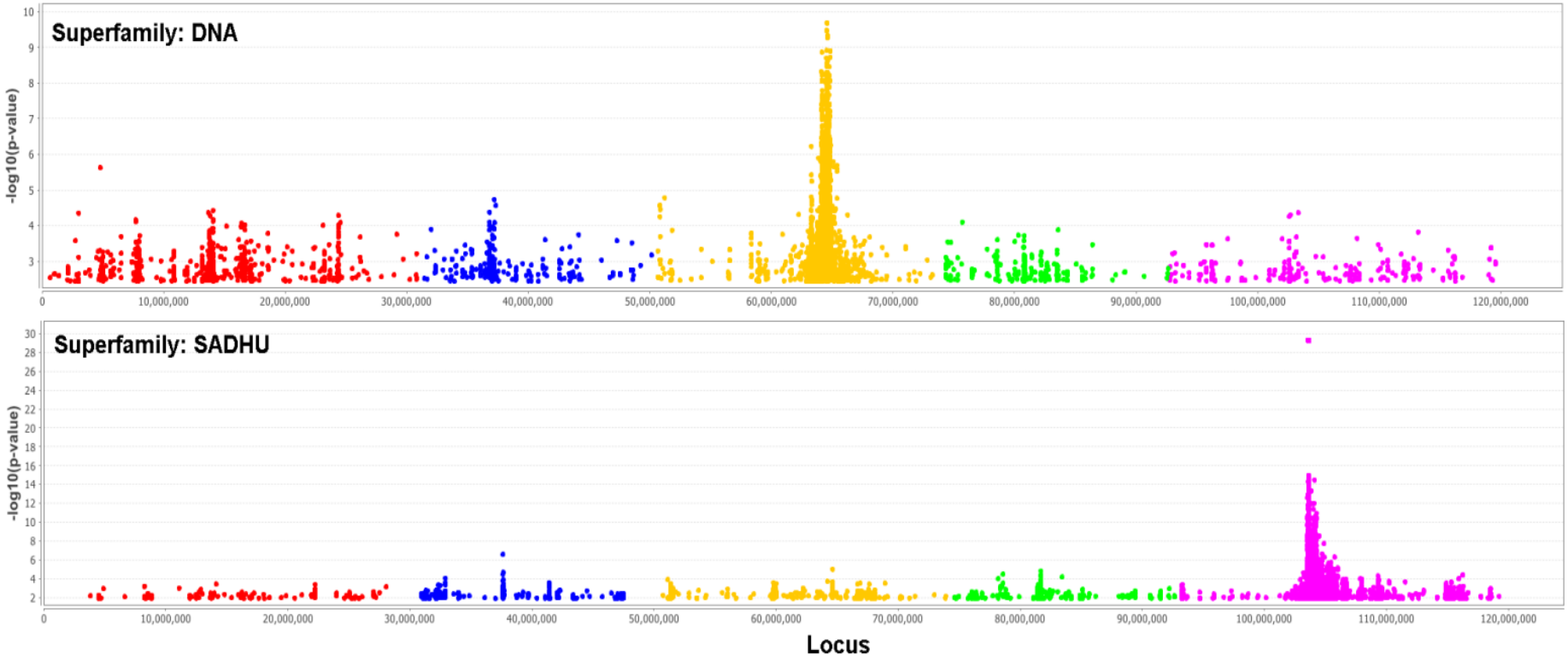


Results: GWAS of copy numbers

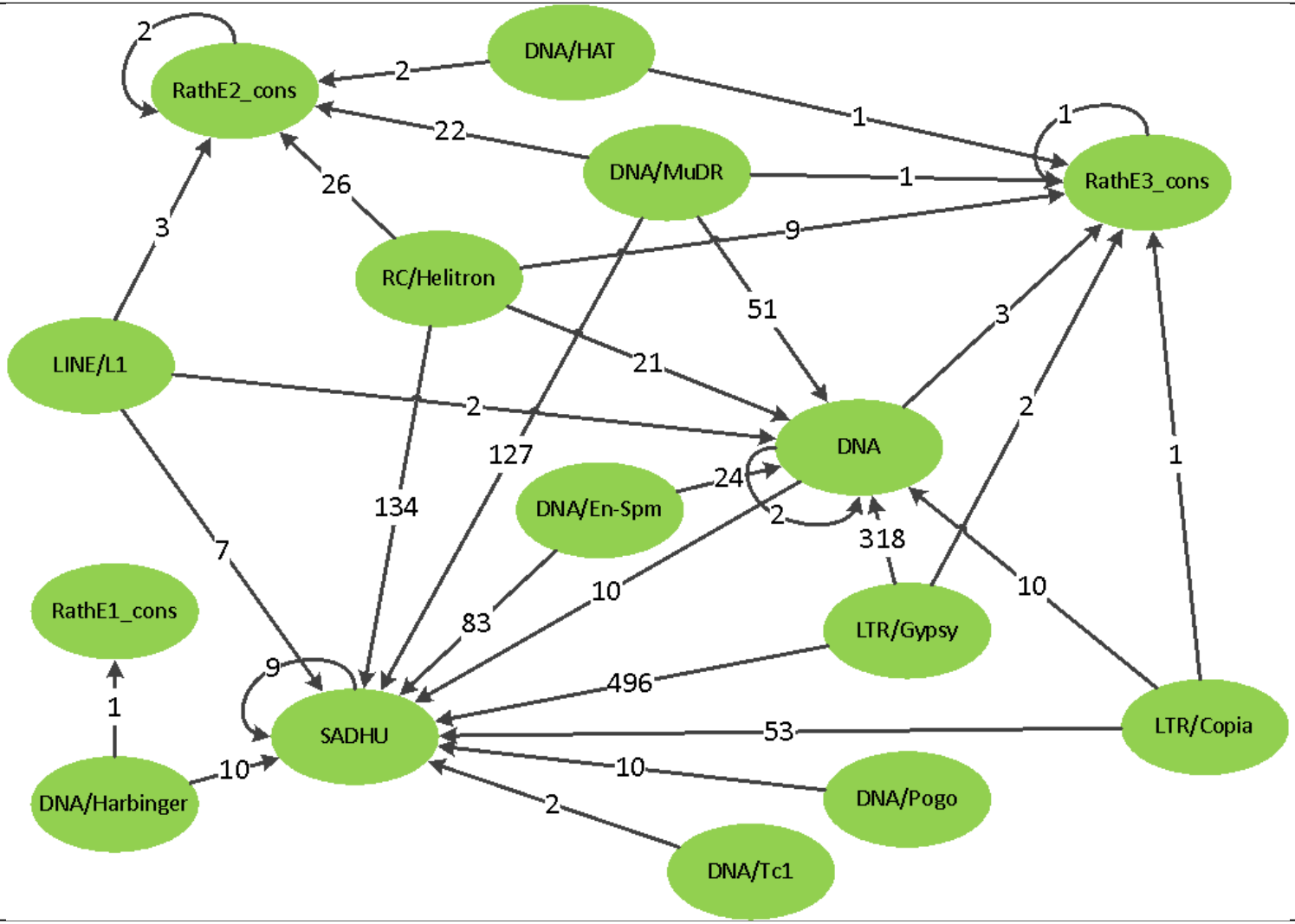
Sources:

- New SNP release
- TE superfamily copy number as phenotype

e.g.:



Results: GWAS of copy numbers
Dependency of GWASes



Indel patterns

FNSNF

Indels pattern search

- +/- 30 bp are extracted
- from the loci of called indels (Q30) and
- randomly from the reference (TAIR10)

- Classified with C5.0

- as all rules are starting with the bp just before the event, this is declared as BP1 (the BP2, BP3, ...), the pattern strings are always starting with BP1

Indels patterns – MSH2$^-$ vs Swedish

MSH2$^-$ (MA line)
186 indels vs. 2000 random
zero ratio = 91.5%
PCC = 97.7%

Rule 1
0.923 (48%)
NAAAAAA

Rule 2
0.871 (53%)
NTTTTT

Population (part of 1001 genomes)
1031959 indels vs. 1000000 random
zero ratio = 50.8%
PCC = 67.2%

Rule 1
0.874 (50.0%)
NNAAAA

Rule 4
0.776 (70.1%)
BP2 = T
BP4 = T
BP8 = T

Rule 2
0.788 (74.6%)
BP3 = T
BP5 = T
BP7 = T

Rule 3
0.776 (42.9%)
TAA

Rule 5
0.682
(11.6%)
NATAT

Rule 6
0.678 (6.5%)
NGGG

Rule 7
0.668 (6.5%)
NCCC

Rule 8
0.597
(2.0%)
NGAGAA

Rule 9
0.560
(277.3%)
BP3 = A

Rule 10
0.597
(276.3%)
BP3 = T

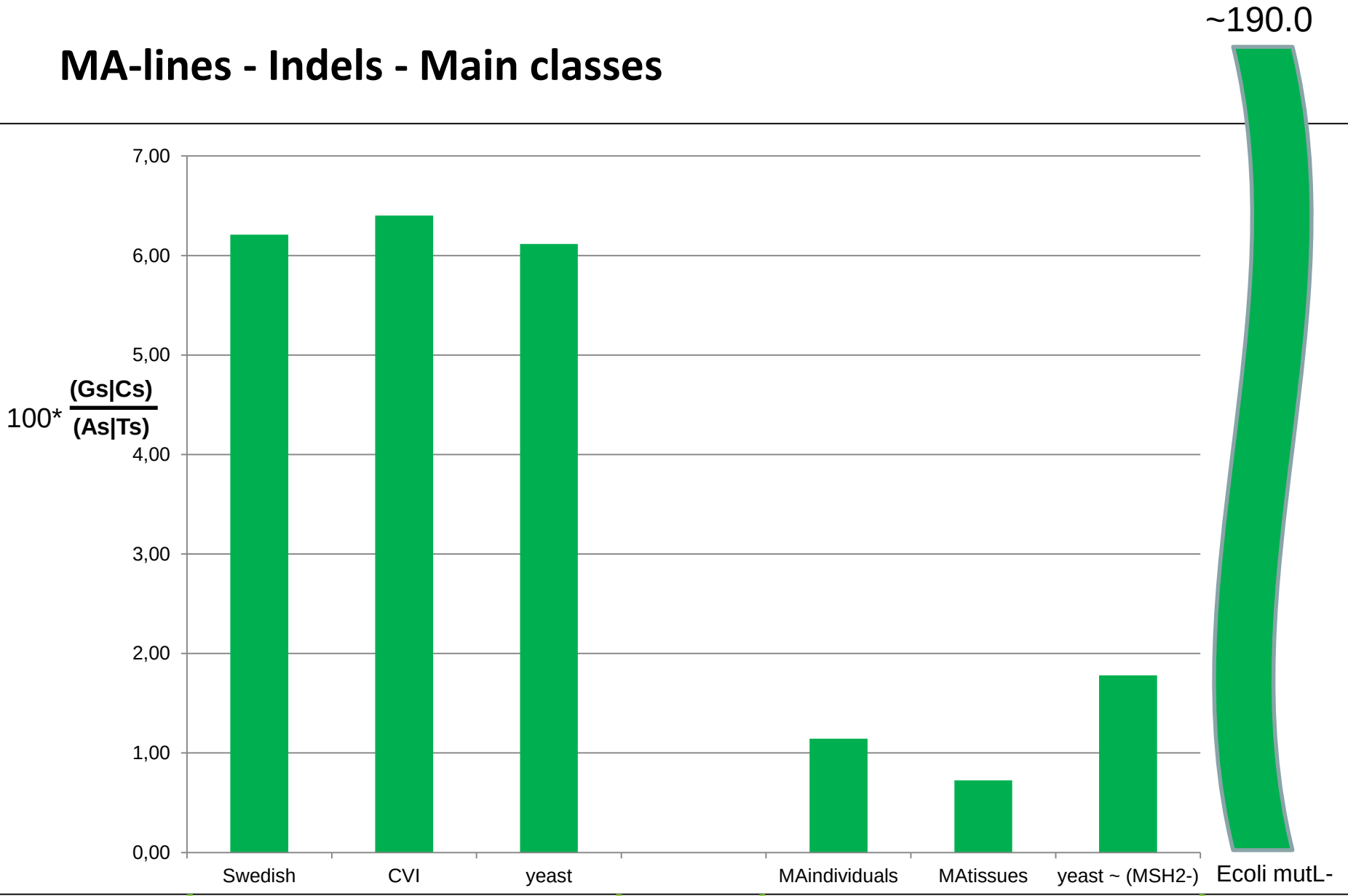
Indels patterns

- Class 2 ... ‚MSH2^{-}‘-alike rules: 1, 2 and 5
- Class 1 ... [CCC|GGG]
- Class 0 ... the remainder

Several patterns, but ~two main classes:

- As|Ts
- Gs|Cs

MA-lines - Indels - Main classes



Alexander Platzer

MLH in Arabidopsis

The Plant Journal (2007) **51**, 431–440 doi: 10.1111/j.1365-313X.2007.03145.x

An *Arabidopsis MLH1* mutant exhibits reproductive defects and reveals a dual role for this gene in mitotic recombination

Éric Dion[†], Liangliang Li[†,‡], Martine Jean and François Belzile[*]
Département de phytologie, 1243 Pavillon C.-E. Marchand, Université Laval, Québec, QC, G1K 7P4, Canada

Received 12 February 2007; revised 20 March 2007; accepted 26 March 2007.
[*]For correspondence (fax +1 418 656 7176; e-mail fbelzile@rsvs.ulaval.ca).
[†]These authors contributed equally to this work.
[‡]Present address: Medicago Inc., 1020 route de l'Église, Bureau 600, Québec, QC, G1V 3V9, Canada.

Summary

The eukaryotic DNA mismatch repair (MMR) system contributes to maintaining genome integrity and DNA sequence fidelity in at least two important ways: by correcting errors arising during DNA replication, and also by preventing recombination events between divergent sequences. This study aimed to investigate the role of one key MMR gene in recombination. We obtained a mutant line in which the *AtMLH1* gene has been disrupted by the insertion of a T-DNA within the coding region. Transcript analysis indicated that no full-length transcript was produced in mutant plants. The loss of a functional *AtMLH1* gene led to a significant reduction in fertility in both homozygotes and heterozygotes, and we observed a strong bias against transmission of the mutant

Transcriptional Enhancement with Natural Antisense Transcripts

FNSNF

Transcriptional Enhancement with Natural Antisense Transcripts

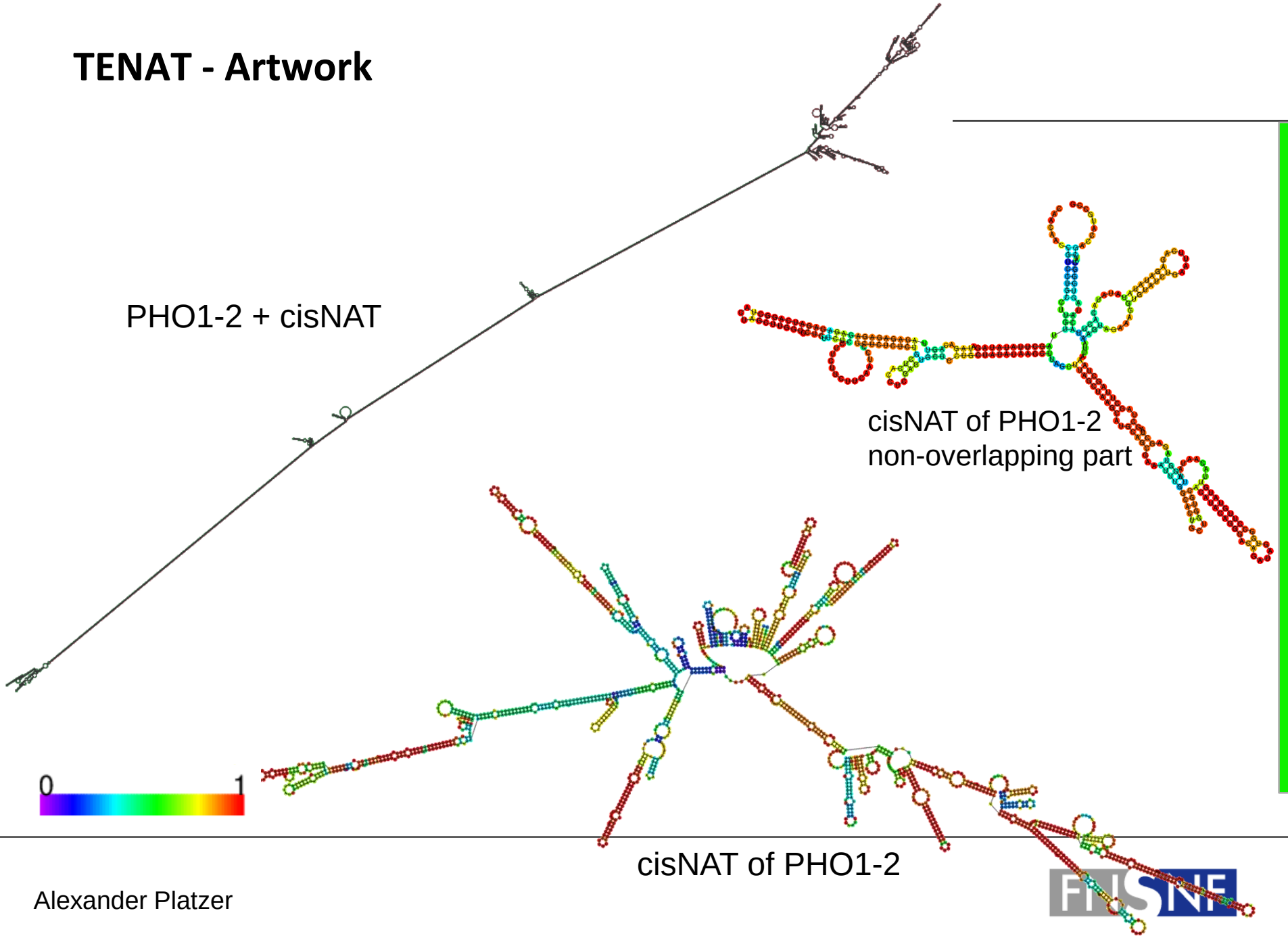
The vast majority of known NATs are lowering mRNA expression and/or translation.

Beside that there are two strongly validated examples of enhancing translation.

- PHO1-2 in rice
- Uchl1 in mouse

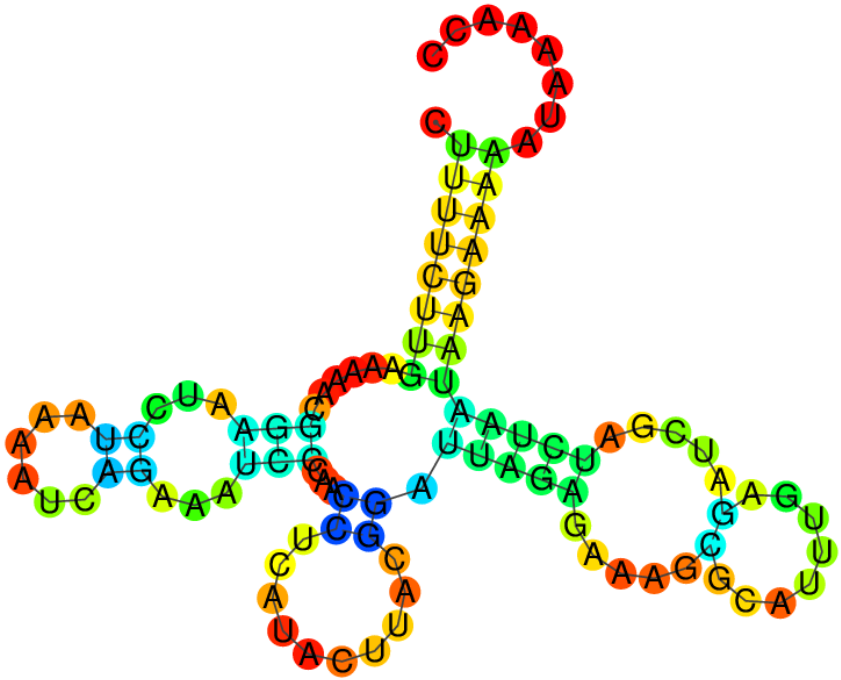
More examples? Classification and prediction? Mechanism?

TENAT - Artwork

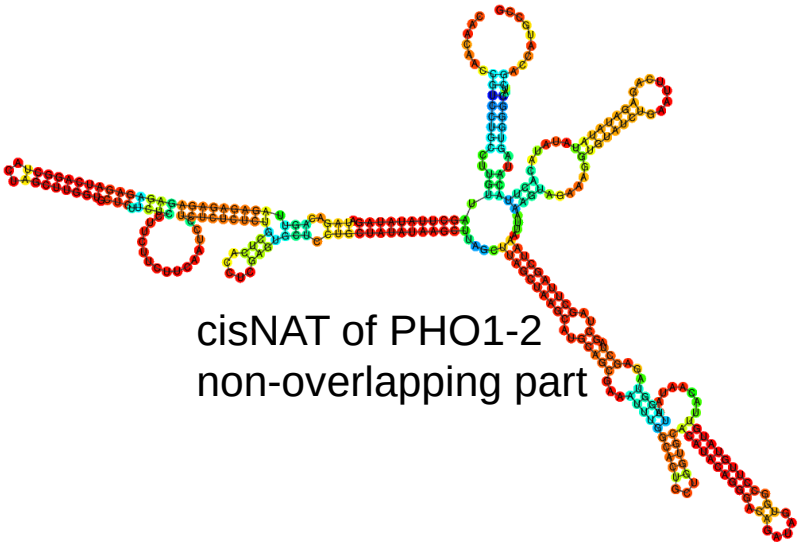


Alexander Platzer

TENAT - Artwork



somehow conserved sequence of cisNAT of PHO1-2 non-overlapping part
(combined with Arabidopsis thaliana)



cisNAT of PHO1-2
non-overlapping part

Alexander Platzer

Summary

- 1000´s fully sequenced genomes
- Dimension reduction methods
- GWAS with transposable elements
- Sequence pattern search
- Transcriptional Enhancement with Natural Antisense Transcripts

Acknowledgements

Quan Long

Irina Druzhinina

Laurens van der Maaten

Ivo Hofacker

GREGOR MENDEL INSTITUTE OF MOLECULAR PLANT BIOLOGY

FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

universität wien

Backup slides

FNSNF

t-SNE

constraints:

- Dimension reduction for not more than 3 dimensions
- Local linearity (violated by highly-varying manifolds)
- Non-convexity of the t-SNE cost function

t-SNE - Parameters

- *Initial dimensions*: The number of dimensions the data is reduced in advance (with PCA), because t-SNE works only up to quadratic matrices; this can be set to the maximum (= number of data records = number of individuals)

- *perplexity*: As defined in information theory, it is a measure how confused a model might be; can be interpreted as smoothing of the data and/or as expected noise in the data, or seen as the number of neighbors taken into account at looking to the distances
  -> lower than the number of individuals,
  range: 5 to #individuals/2

t-SNE - Perplexity

$2^{H(p,q)}$ with $\mathrm{H}(p,q) = -\sum_{x} p(x) \log q(x).$

The parameter is used to normalize the conditional probability distributions to this perplexity.

t-SNE - Algorithm

momentum : 0.5 -> 0.8

init. learning rate : 500
(then adaptive)

Algorithm 1: Simple version of t-Distributed Stochastic Neighbor Embedding.

Data: data set $\mathcal{X} = \{x_1, x_2, ..., x_n\}$,
cost function parameters: perplexity *Perp*,
optimization parameters: number of iterations T, learning rate η, momentum $\alpha(t)$.
Result: low-dimensional data representation $\mathcal{Y}^{(T)} = \{y_1, y_2, ..., y_n\}$.
begin
  compute pairwise affinities $p_{j|i}$ with perplexity *Perp* (using Equation 1)
  set $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$
  sample initial solution $\mathcal{Y}^{(0)} = \{y_1, y_2, ..., y_n\}$ from $\mathcal{N}(0, 10^{-4}I)$
  for $t=1$ **to** T **do**
    compute low-dimensional affinities q_{ij} (using Equation 4)
    compute gradient $\frac{\delta C}{\delta \mathcal{Y}}$ (using Equation 5)
    set $\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta \frac{\delta C}{\delta \mathcal{Y}} + \alpha(t)\left(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)}\right)$
  end
end

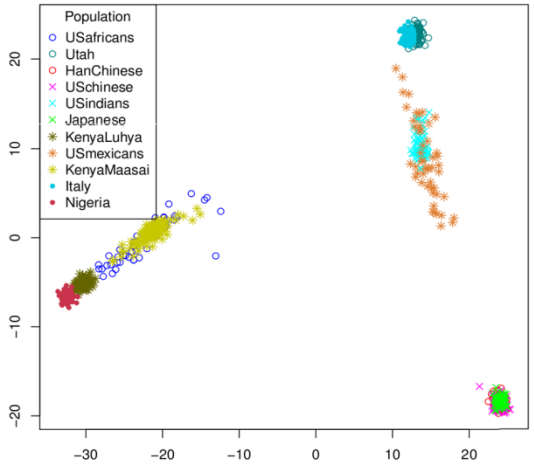
$$(1)\quad p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}$$

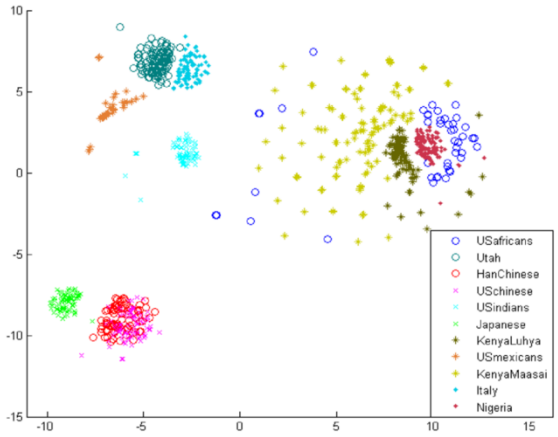
$$(4)\quad q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|y_k - y_l\|^2\right)^{-1}}$$

$$(5)\quad \frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)\left(1 + \|y_i - y_j\|^2\right)^{-1}$$

Measurements – Structuredness - As a clustering result

transformed data / diagrams





Coordinates

as clustering result

distance	label1	label2

Dunn's Validity Index,
Silhouette Validation
Method

Measurements - Indices of cluster validity

Data	**Dunn's Validity Index**			**Silhouette Validation Method**		
	PCA	*t-SNE*	*Diff*	*PCA*	*t-SNE*	*Diff*
1001 genomes	0.52 (0.09)	0.61 (0.07)	0.09	0.07 (0.04)	0.22 (0.04)	0.15
RegMap	0.50 (0.06)	0.50 (0.04)	0.00	0.08 (0.02)	0.15 (0.02)	0.07
Hapmap3R2	0.16 (0.01)	0.25 (0.02)	0.09	0.27 (0.02)	0.31 (0.02)	0.04
Hapmap3R3	0.16 (0.01)	0.35 (0.01)	0.19	0.26 (0.02)	0.32 (0.02)	0.06
Rice	0.06 (0.07)	0.10 (0.10)	0.04	-0.54 (0.04)	-0.46 (0.04)	0.08

Backup