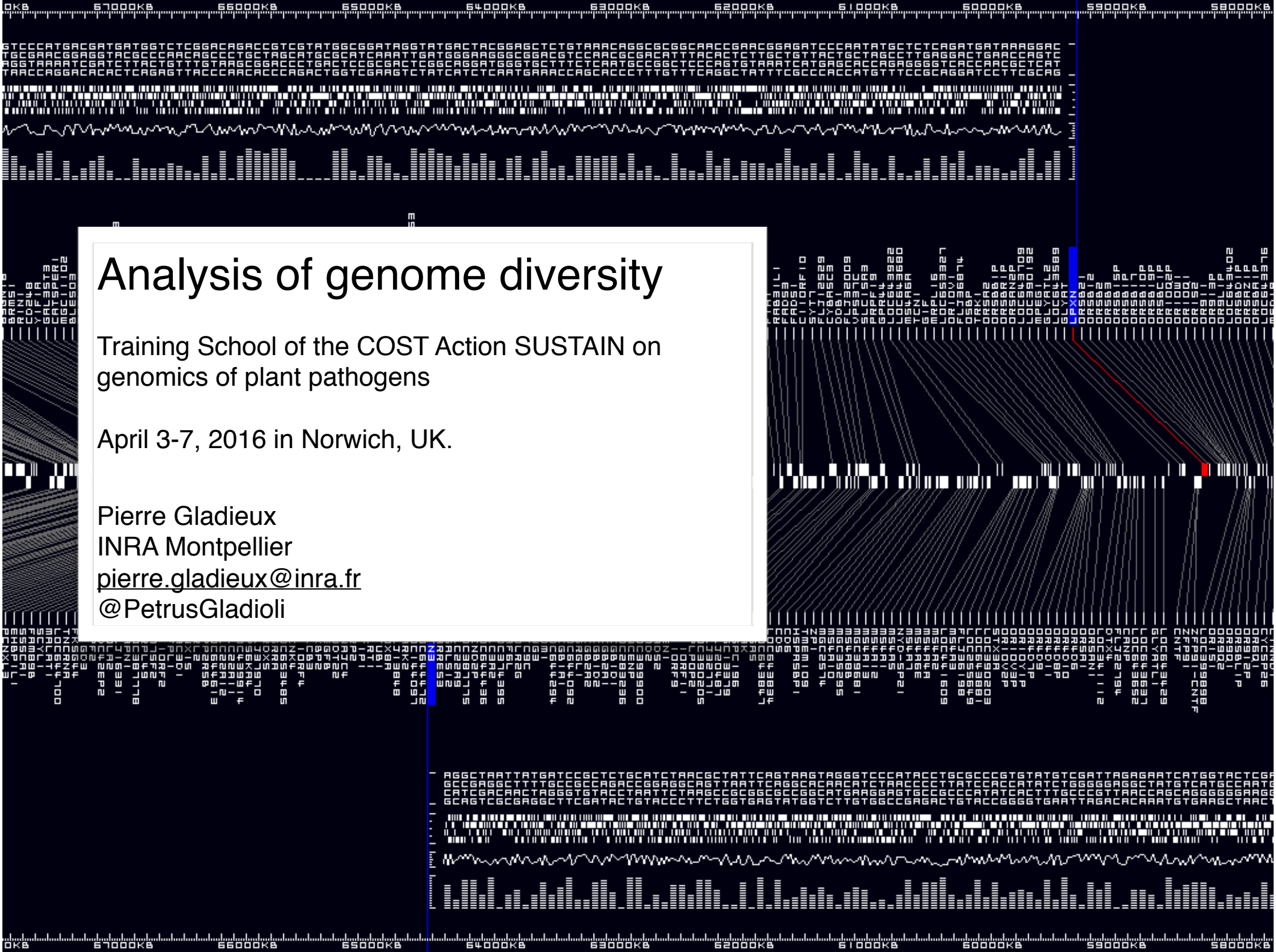


Analysis of genome diversity

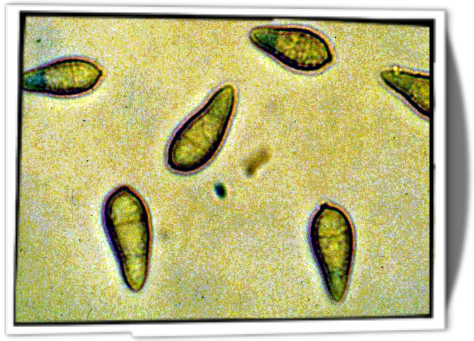
Training School of the COST Action SUSTAIN on genomics of plant pathogens

April 3-7, 2016 in Norwich, UK.

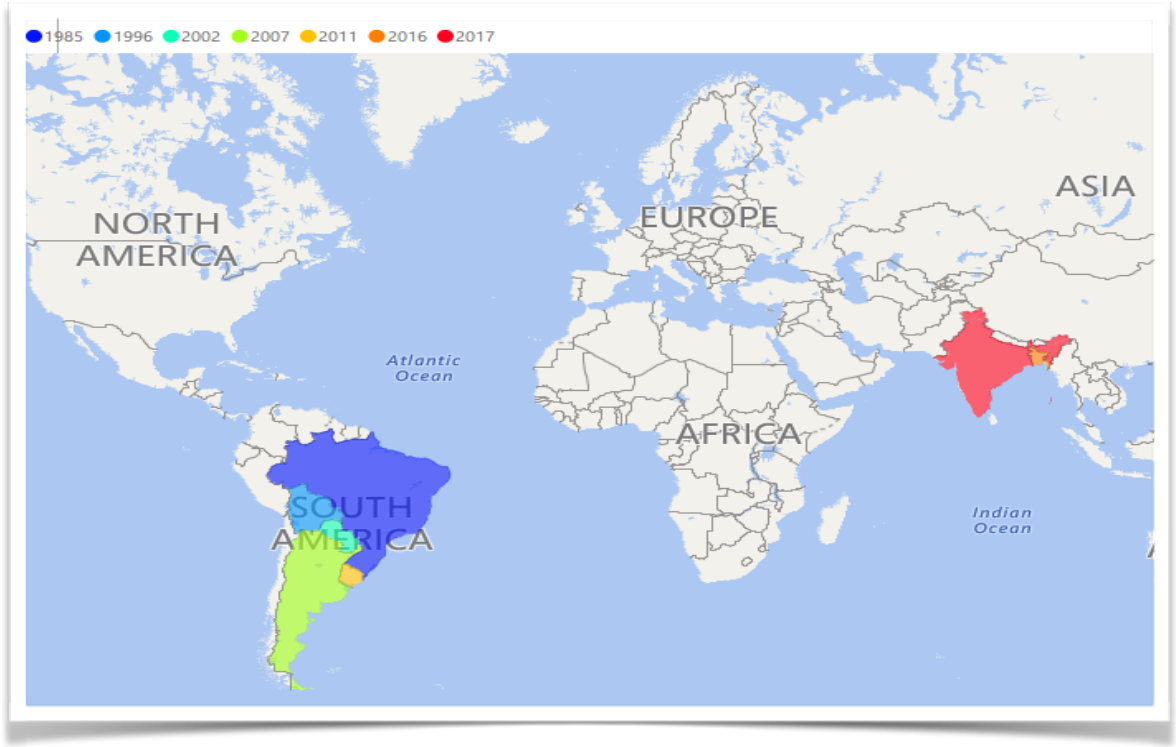
Pierre Gladieux
INRA Montpellier
pierre.gladieux@inra.fr
@PetrusGladioli

Magnaporthe/Pyricularia oryzae
Rice blast

- Widely distributed
- Rapidly adapting
- Most damaging rice disease
- Model for plant pathology



Magnaporthe/Pyricularia oryzae
Wheat blast



- Emergence in South America in the 80s
- Introduction in Bangladesh in 2016
- Introduction in India in 2017

Wheat blast in Bangladesh

In February 2016, wheat blast was spotted in Bangladesh– its first report in Asia. Wheat is the second major food source in Bangladesh, after rice. The blast disease has, so far, caused up to 90% yield losses in more than 15000 hectares. Scientists fear that the pathogen could spread further to other wheat growing areas in South Asia.

MAKING DATA INSTANTLY ACCESSIBLE

LET'S MAKE A DIFFERENCE

RECENT POSTS

- A community article published using data contributed on

Emergence of wheat blast in Bangladesh was caused by a South American lineage of *Magnaporthe oryzae*

M. Tofazzal Islam [1]*, Daniel Croll [2], Pierre Gladieux [3], Darren M. Soanes [4], Antoine Persoons [5], Pallab Bhattacharjee [1], Shaid Hossain [1], Dipali Rani Gupta [1], Md. Mahbubur Rahman [1], M. Golam Mahboob [6], Nicola Cook [5], Moin U. Salam [7], Vanessa Bueno Sancho [5], João Leodato Nunes Maciel [8], Antonio Nhani Júnior [8], Vanina Lilián Castroagudín [9], Juliana T. de Assis Reges [9], Paulo Cezar Ceresini [9], Sebastien Ravel [10], Ronny Kellner [11, 12], Elisabeth Fournier [3], Didier Tharreau [10], Marc-Henri Lebrun [13], Bruce A. McDonald [2], Timothy Stitt [5], Daniel Swan [5], Nicholas J. Talbot [4], Diane G.O. Saunders [5, 14], Joe Win [11], and Sophien Kamoun [11]*

[1] Department of Biotechnology, Bangabandhu Sheikh Mujibur Rahman Agricultural University, Gazipur 1706, Bangladesh

[2] Plant Pathology, Institute of Integrative Biology, ETH Zurich, 8092 Zurich, Switzerland

[3] INRA, UMR 385 Biologie et génétique des interactions plantes-pathogènes BGPI, Montpellier, France

BioMed Central

BMC Biology

RESEARCH ARTICLE | OPEN ACCESS

Emergence of wheat blast in Bangladesh was caused by a South American lineage of *Magnaporthe oryzae*

M. Tofazzal Islam ✉, Daniel Croll, Pierre Gladieux, Darren M. Soanes, Antoine Persoons, Pallab Bhattacharjee, Md. Shaid Hossain, Dipali Rani Gupta, Md. Mahbubur Rahman, M. Golam Mahboob, Nicola Cook, Moin U. Salam, Musrat Zahan Surovy, Vanessa Bueno Sancho, João Leodato Nunes Maciel, Antonio NhaniJúnior, Vanina Lilián Castroagudín, Juliana T. de Assis Reges, Paulo Cezar Ceresini, Sebastien Ravel, Ronny Kellner, Elisabeth Fournier, Didier Tharreau, Marc-Henri Lebrun, Bruce A. McDonald, Timothy Stitt, Daniel Swan, Nicholas J. Talbot, Diane G. O. Saunders, Joe Win and Sophien Kamoun ✉

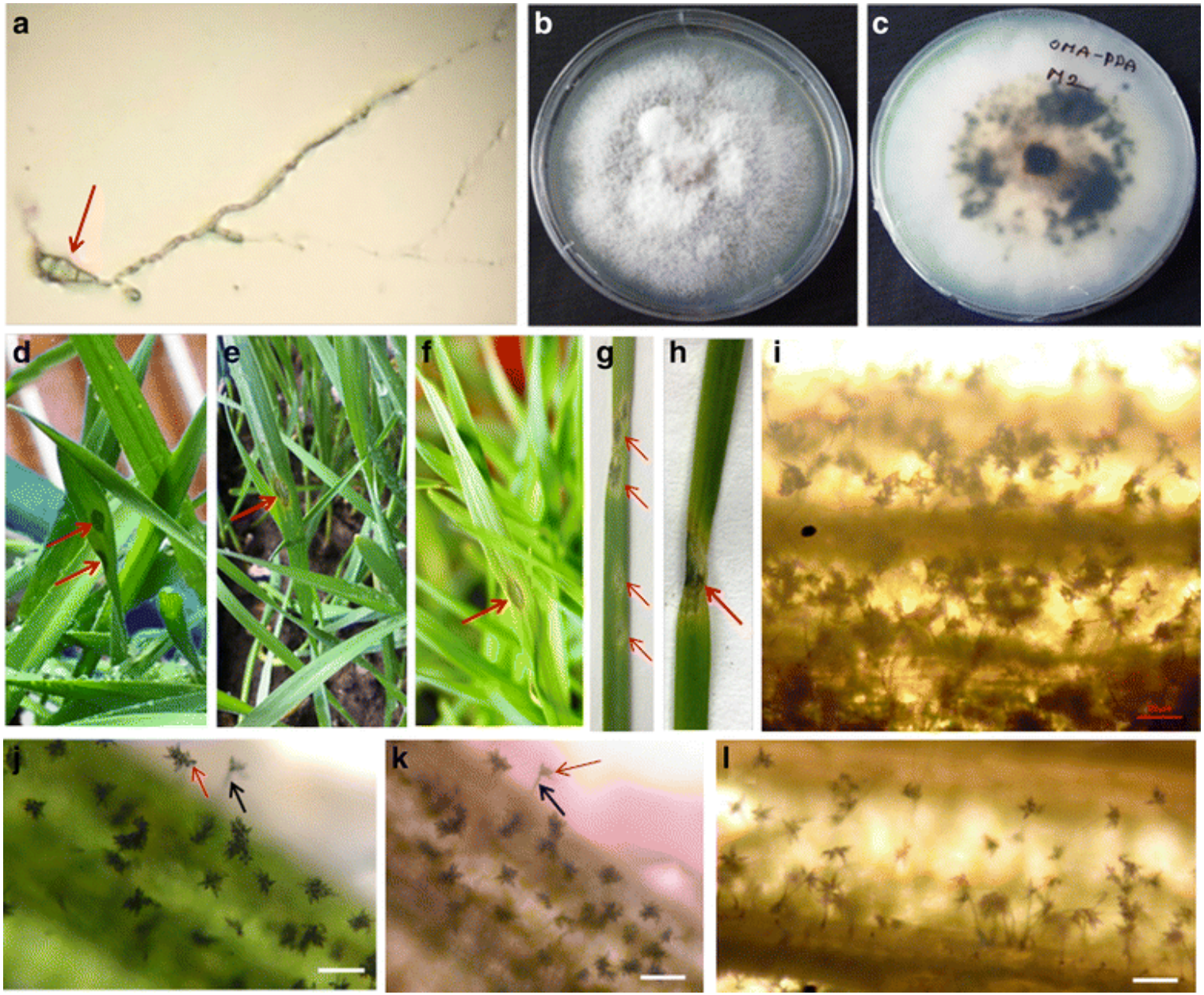
BMC Biology 2016 14:84 | DOI: 10.1186/s12915-016-0309-7

Received: 8 July 2016 | Accepted: 12 September 2016 | Published: 3 October 2016

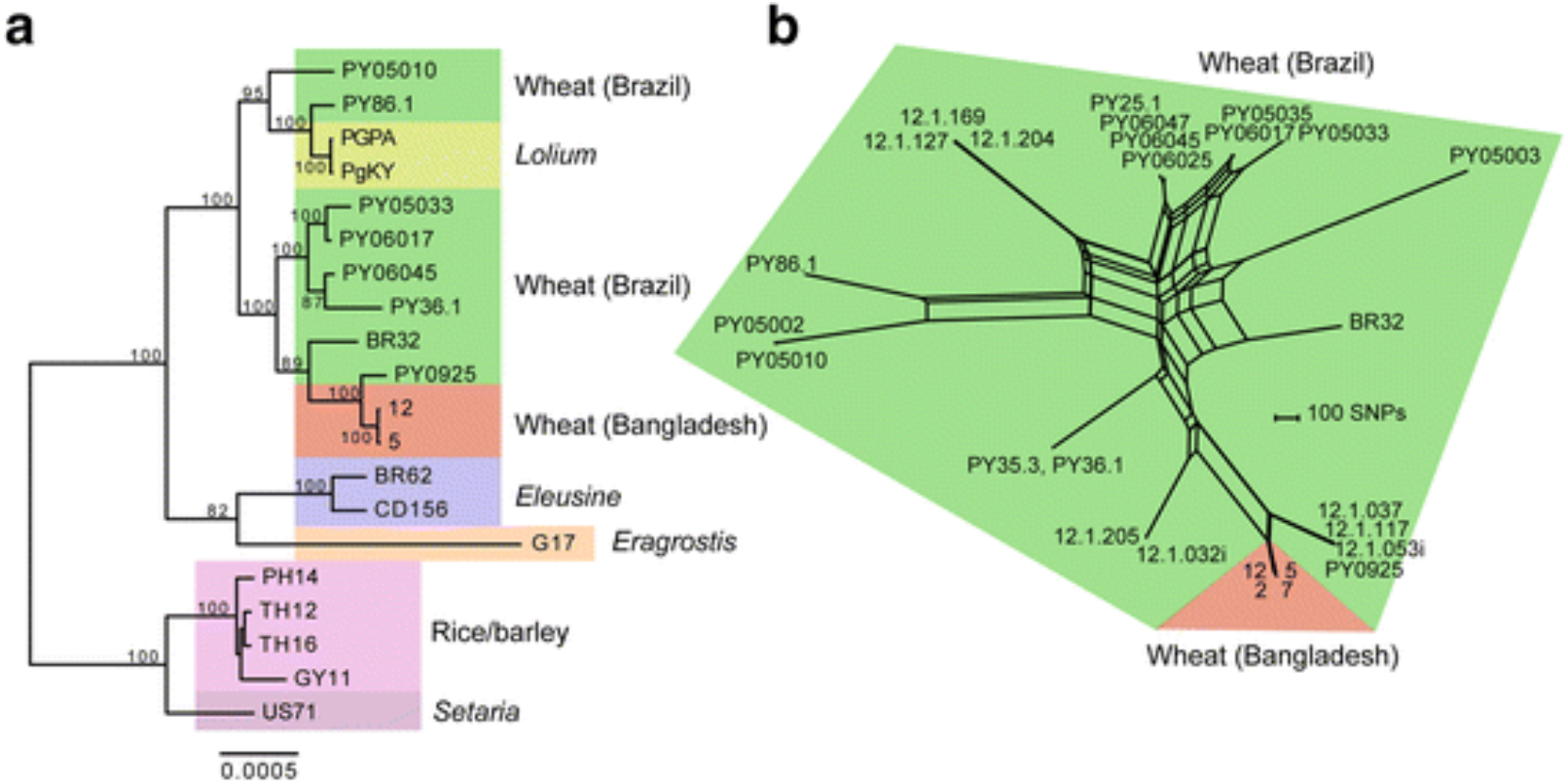
Symptoms

- complete bleaching of wheat spike
- typical eye-shaped lesions
- severely infected rachis and damaged spikelets
- pyriform conidia

Koch’s postulate



Phylogenomic and population genomic analyses



Maximum likelihood genealogy inferred from the concatenation of aligned genomic data at 2193 orthologous groups of predicted transcript sequences. Scale bar represents the mean number of nucleotide substitutions per site

Population genomic analyses of transcriptomic single nucleotide polymorphisms among M. oryzae isolates from wheat in Brazil and Bangladesh. The network was constructed using the Neighbor-Net algorithm. The scale shows the number of informative sites

Population genomics of wheat blast

Aims

- Origin of Bangladeshi samples
- Population structure of wheat-infecting Magnaporthe oryzae
- Genomic changes and evolutionary factors underlying the emergence of the wheat-infecting lineage

Approach

- SNP Data collection
- Analysis of population subdivision
- Analysis of polymorphism and divergence and neutrality tests

Disclaimer

- The proposed exercises are not necessarily the optimal way to do things. For example we will work on fasta files (Sanger type of datasets), whereas it would be possible to directly work on the VCFs (NGS type of datasets).

- The purpose of the exercises is to manipulate data and write code in Python, and to use Biopython and Egglib, not to write the fastest or the most elegant code.

- The proposed solutions are not designed to be optimal either, because: (1) it's not the goal of this workshop, (2) we are biologists, not bioinformaticians. The advantage of having code written by a biologist, is that you can actually read it!

SNP calling: Islam et al. pipeline

GitHub, Inc. [US] https://github.com/crolllab/wheat-blast

The origin of wheat blast in Bangladesh

What is wheat blast? Is it dangerous?

Wheat blast is a fungal disease that leads to large yield losses. You can find some more information at Open Wheat Blast and from Kansas State University

What happened in February 2016?

Wheat blast was first found in Bangladesh. This was the first report of the disease in Asia. The disease already caused large yield losses and there is a significant worry that the disease will rapidly spread to wheat production areas in India and beyond.

Why was there an outbreak of wheat blast in Bangladesh?

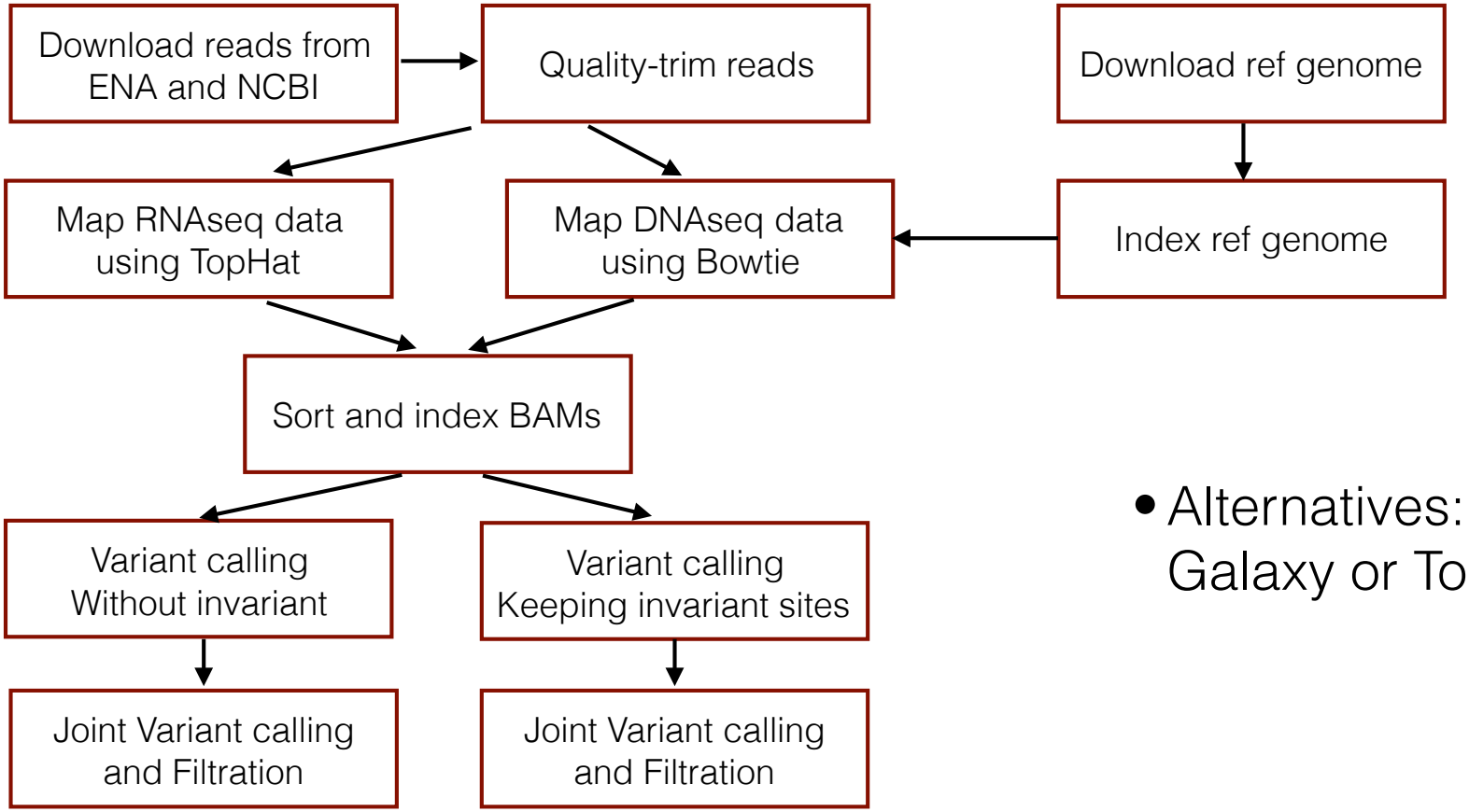
Currently, nobody knows. The key question is whether wheat blast was introduced to Bangladesh from elsewhere or whether it originated locally (e.g. through natural mutations in local fungal strains).

Wheat blast and rice blast are closely related pathogen lineages and currently identified as one species. All rice blast pathogen strains are genetically very similar. Wheat blast pathogen strains are genetically diverse and distinct from rice blast, as shown here.

Who are we?

We are plant pathologists at the ETH Zurich (Switzerland) using genomics tools to identify how plant pathogens cause disease. You can read more about Daniel Croll's research here and Bruce McDonald's research here.

SNP calling using Tophat, bowtie, GATK et al.



- Alternatives: Galaxy or Toggle

Datasets

- SNPs in wheat blast isolates from Brazil and Bangladesh
  wheat_blast_brazil_bangladesh.snp-only.filters_maxmissing30.vcf
- Variants (SNPs + indels) and invariant positions in wheat blast isolates from Brazil.
  wheat_blast_brazil.with_inv.filters.vcf
- Variants (SNPs + indels) and invariant positions in wheat blast isolates from Brazil, including an isolate of M. grisea as outgroup.
  wheat_blast_brazil_outgroup.with_inv.filters.vcf

Population subdivision: Structure software

Aim

- Infer the number of populations ('clusters') in the dataset
- Assign Bangladeshi genotypes to inferred clusters

Approach

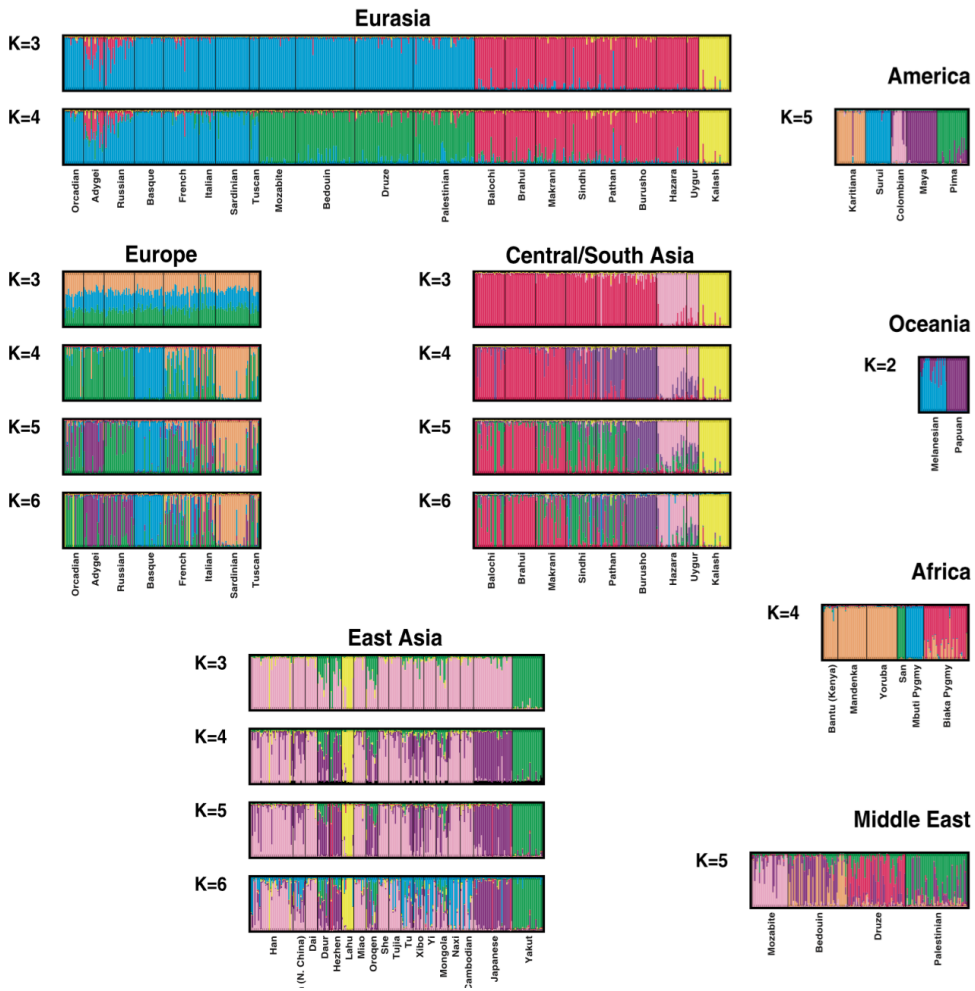
- Builds K clusters that minimize linkage disequilibrium and Hardy-Weinberg disequilibrium
- Infer individuals ancestry in K clusters

Method

- Prepare dataset, subsampling SNPs
- Run program
- Check convergence of MCMC

Population subdivision: Structure software

Inference of population **structure** using multilocus genotype data
JK **Pritchard**, M Stephens, P **Donnelly** - Genetics, 2000 - Genetics Soc America
Abstract We describe a model-based clustering method for using multilocus genotype data to infer population structure and assign individuals to populations. We assume a model in which there are K populations (where K may be unknown), each of which is characterized by
Cited by 19185 Related articles All 60 versions Cite Saved



Exercise 1: Make Structure dataset

input: wheat_blast_brazil_bangladesh.snp-only.filters.vcf

1. Write script to convert vcf to tab file

see demo

output:

SNP1 genotype_indiv1 genotype_indiv2 … genotype_indivn
SNP2 genotype_indiv1 genotype_indiv2 … genotype_indivn
…
SNPk genotype_indiv1 genotype_indiv2 … genotype_indivn

genotype_indiv_i in [A,T,C,G,N]

2. Write script to convert tab to Structure format

keep 1SNP/gene, ~1000 SNPs final

output:

indiv1 genotype_SNP1 genotype_SNP2 … genotype_SNPk
indiv2 genotype_SNP1 genotype_SNP2 … genotype_SNPk
…
indivn genotype_SNP1 genotype_SNP2 … genotype_SNPk

genotype_indiv_j in [1,2,3,4,-9]

Exercise 2: Run Structure

input: wheat_blast_brazil_bangladesh.snp-only.filters_maxmissing30_one_SNP_per_gene.str

Population structure model:

- admixture: estimate membership proportions, not membership probabilities
- linkage: models correlations in ancestry along chromosomes
  - Mixture LD: « The first source is variation in ancestry (*q*) among the sampled individuals. Variation in *q* leads to correlations among markers across the genome, even if they are unlinked, because individuals with a large component of ancestry in population *k* have an excess of alleles that are common in *k*. »
  - Admixture LD: « The second source is correlations in ancestry along each chromosome, which cause additional LD between linked markers. »
  - Background LD: « usually decays on a much shorter scale »

Command line:

edit mainparams (50000 burnin, then 150000 steps)

then run in the terminal: ./structure -K $k -o output

Population subdivision: Structure software

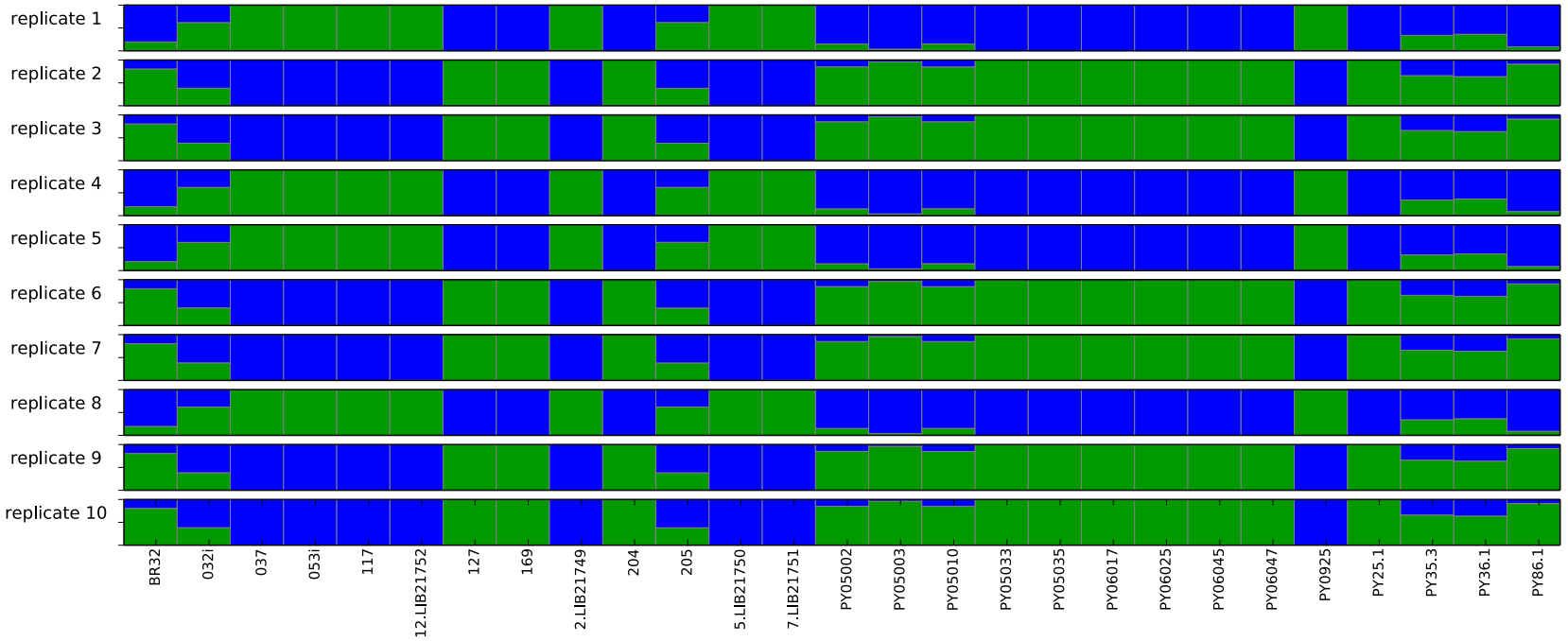
Check convergence:

- Burnin length
- Run length

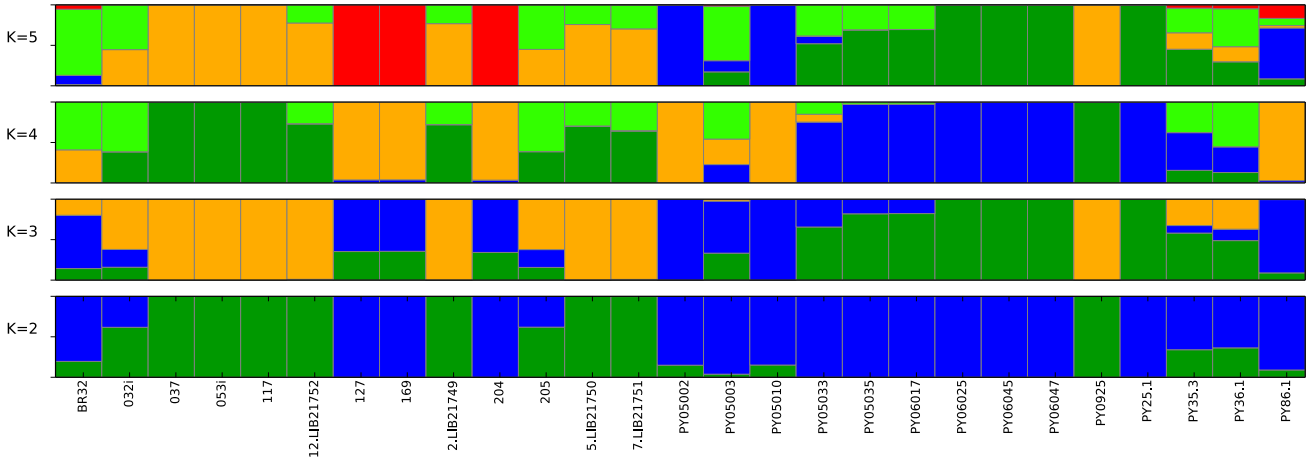
Alternative: CLUMPP program

- alpha, F, the divergence distances among populations Di,j , and the likelihood) to see whether they appear to have converged
- clustering patterns:

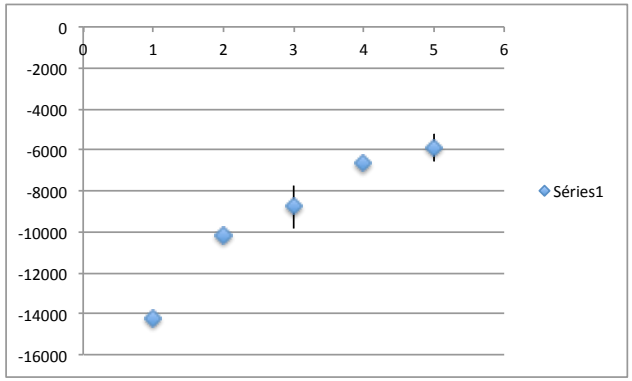
K=2



Population subdivision: Structure software



Likelihood



- K=1 to K=2, strongest increase in likelihood
- Alternative: Evanno's DeltaK statistic

K=2 random mating populations?

- Impact of clonality
- Need method showing hierarchical relationships, and taking into account heterogeneity in reproductive mode

Population subdivision: Neighbor-net network

« Neighbor-Net, a distance based method for constructing phylogenetic networks that is based on the Neighbor-Joining (NJ) algorithm of Saitou and Nei. »

First construct a collection of weighted splits (bipartitions of the taxa set), then represent these splits using a splits graph.

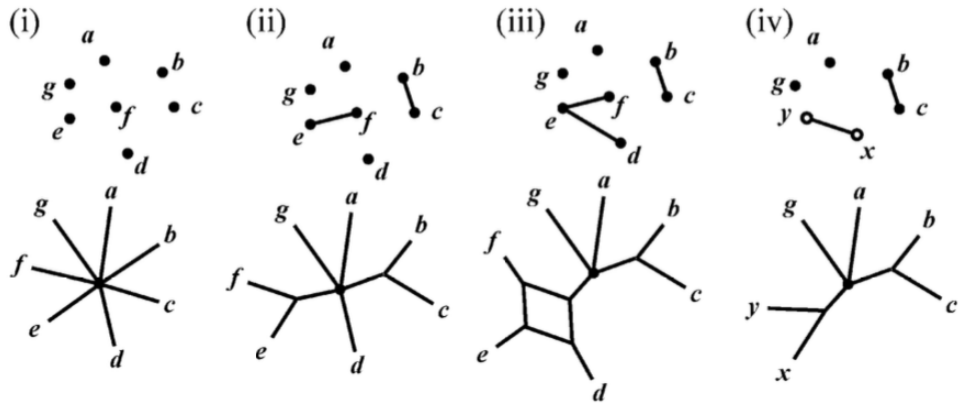


FIG. 2.—The agglomerative process for Neighbor-Net. (i) We begin with each node representing a single taxon. (ii) Using the selection criterion, we identify *b* and *c* as neighbors, as well as *e* and *f*. Unlike NJ, we do not amalgamate immediately. (iii) We have identified *e* as a neighbor of *d* (as well as *f*). Notice how the splits *ef*|*abcdg* and *de*|*acdfg* are both represented in the splits graph. (iv) As *e* has two neighbors, we perform a reduction, replacing *d*, *e*, *f* by *x*, *y*.

Bryant & Moulton 2003 MBE

Neighbor-net network using Splitstree

Takes into account possibility recombination and incomplete lineage sorting

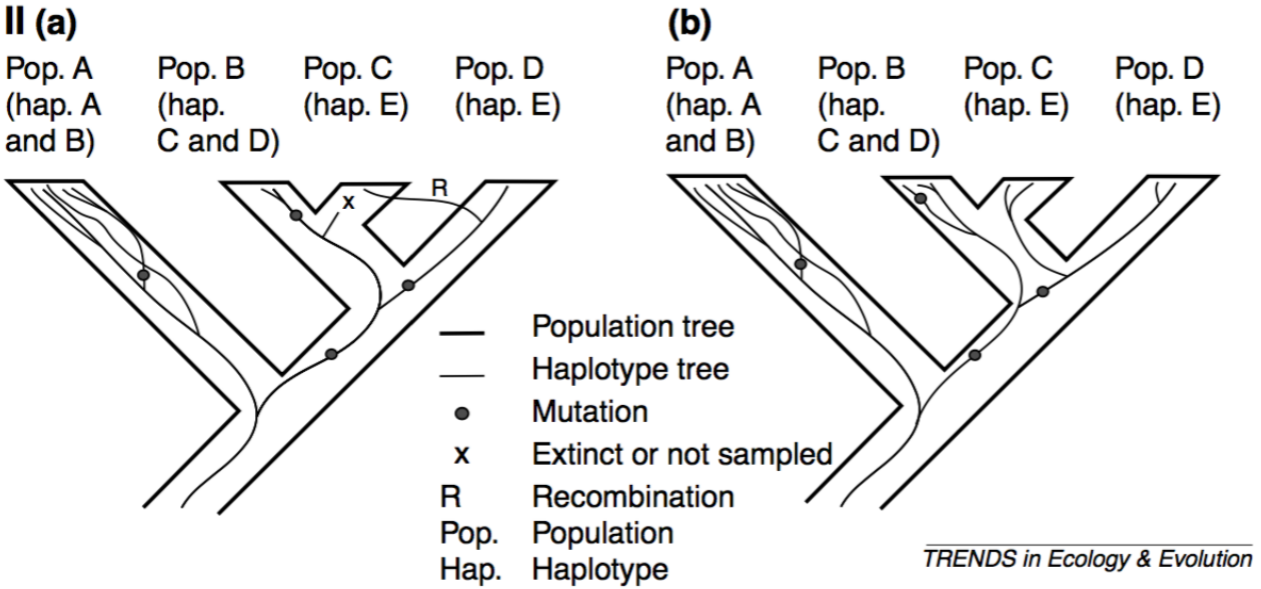


Figure: Disagreement among haplotype trees and population trees

Homoplasy:

same mutation in « unrelated » genomes

Test for recombination using PHI-test, assumptions:

all sites equivalent

limited level of homoplasy

Posada & Crandall 2001 TREE

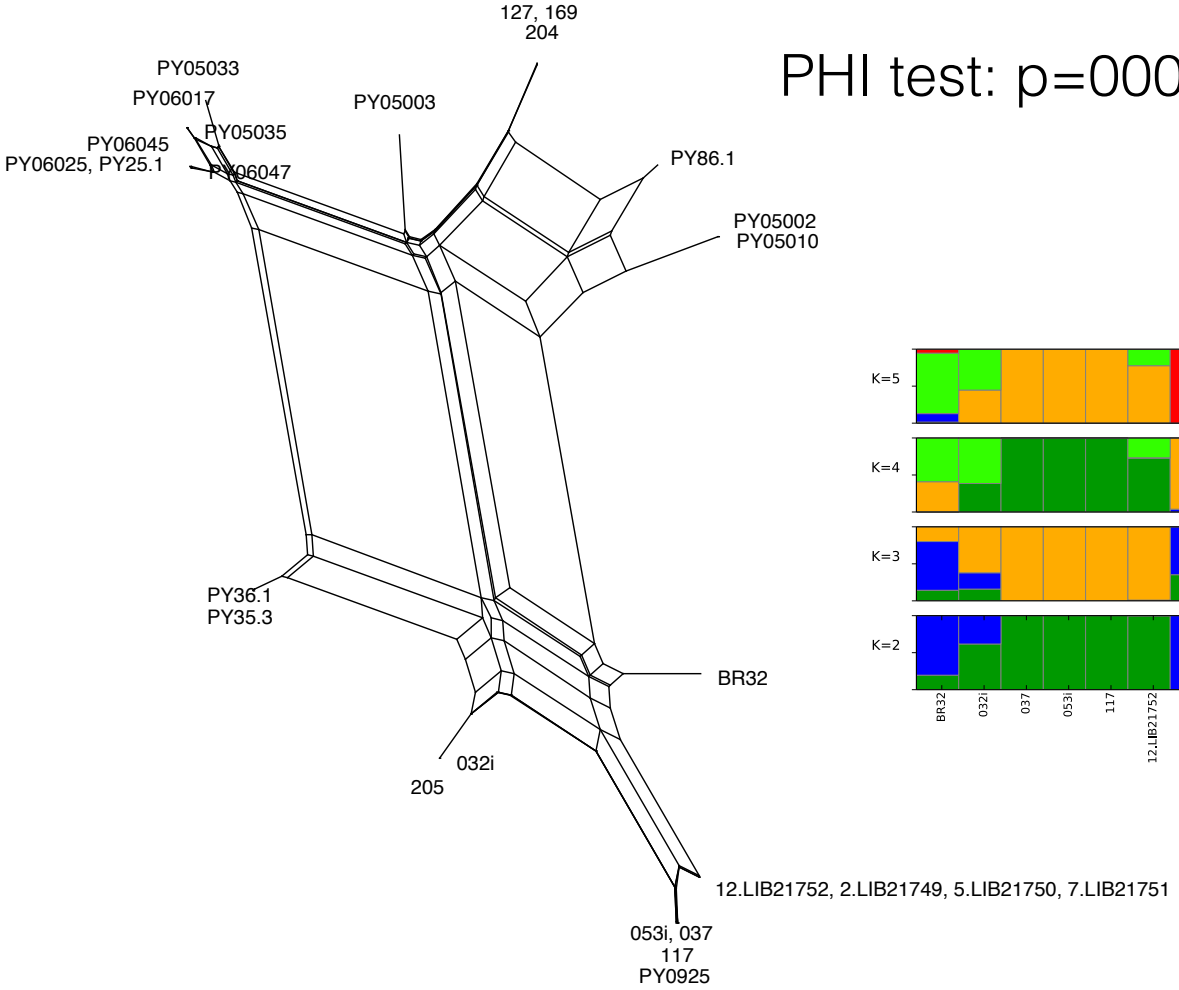
Exercise3: Make Splitstree dataset

input: wheat_blast_brazil_bangladesh.snp-only.filters.vcf

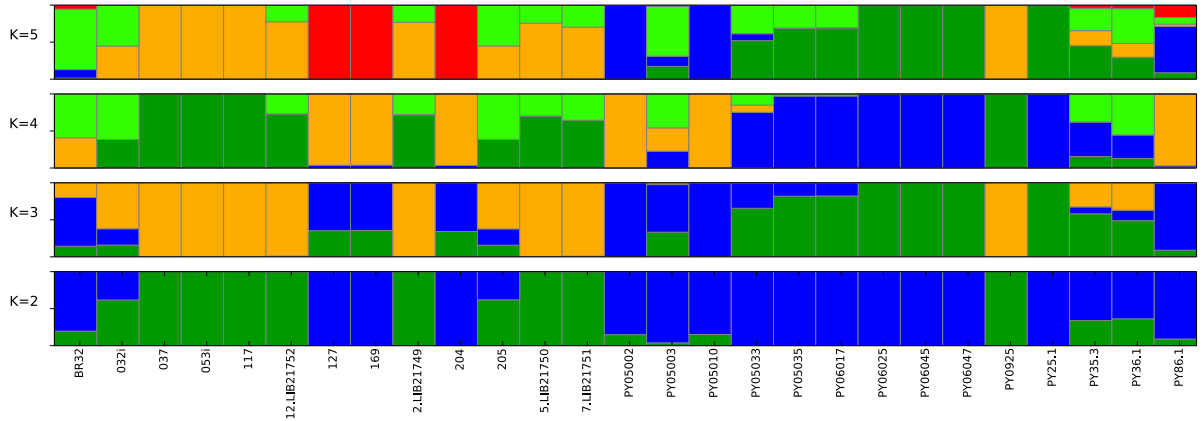
Edit « tab to structure » script to
make a fasta file with only SNPs

Exercise4: Make Neighbor-net network and test for recombination (PHI test) in Splitstree

Population subdivision: Structure & Network



PHI test: p=000 (null hypothesis: clonality)



- No long branches separating clusters of isolates
- Groups in Structure are groups of related isolates
- No population structure

Summary statistics for polymorphism

Exercise5: Make fasta file for all coding sequences

input: wheat_blast_brazil.with_inv.filters.vcf

1. Edit vcf to tab script to make fasta sequences for all isolates and all scaffolds

2. Write scripts that extracts coding sequences from genomic sequences, using a GFF annotation file

Exercise6: Compute summary stats for polymorphism using Egglib

tools and recipes for
evolutionary genetics and genomics

EggLib's documentation »

Welcome to EggLib's home page!

EggLib is a C++/Python library and program package for evolutionary genetics and genomics. Main features are sequence data management, sequence polymorphism analysis, and coalescent simulations. EggLib is a flexible Python module with a performant underlying C++ library and allows fast and intuitive development of Python programs and scripts.

Citation: De Mita S. and M. Siol. 2012. EggLib: processing, analysis and simulation tools for population genetics and genomics. *BMC Genet.* **13**:27. Open access.

Current version: 3.0.0b12

Exercise6: Compute summary stats for polymorphism using Egglib

Compute diversity stats on fasta files using Align objects and ComputeStats class

```
import egglib
aln = egglib.io.from_fasta(path+file, groups=False)
cs = egglib.stats.ComputeStats()
cs.configure(max_missing_freq=0.3)
cs.add_stat('ls')
stats = cs.process_align(aln)
print stats['ls']
```

summary_stats_fasta_egglib.py

Exercise6: Compute summary stats for polymorphism using Egglib

Alternative: compute directly on VCFs
using object egglib.io.VcfParser

```
vcf = egglib.io.VcfParser('my_genome.vcf')
print 'list of samples:', [vcf.get_sample(i) for i in range(vcf.num_samples)]
sites = vcf.get_genotypes(get_genotypes=True)
cs.configure()
cs.add_stat("He")
print cs.process_site(sites)

for i in range(50):
    chrom, pos, nall = vcf.next()
    sites = vcf.get_genotypes(get_genotypes=True)
    print chrom, pos, nall, sites.freqs()[0], cs.process_site(sites)
```

McDonald-Kreitman tests

Exercise8: compute Pn and Ps

NB: handout: Dn & Ds (mistake)

input: wheat_blast_brazil.with_inv.filters.vcf

script: summary_stats_fasta_egglib_syn_nonsyn.py

Define reading frame, create an egglib.stats.CodingDiversity object, create an egglib.stats.Filter object to analyze alignments representing synonymous and non-synonymous variation

```
rf = egglib.tools.ReadingFrame([(0,full_length_stats['ls']-1)])
cdiv = egglib.stats.CodingDiversity(aln, frame=rf, max_missing=int(0.3*aln.ns))
align_S = cdiv.mk_align_S()
align_NS = cdiv.mk_align_NS()
numS=cdiv.num_sites_S
numNS=cdiv.num_sites_NS
codon_filter = egglib.stats.Filter(rng=(0, 63), missing=64)
cs = egglib.stats.ComputeStats()
cs.configure(filtr=codon_filter, max_missing_freq=0.3)
…
```

McDonald-Kreitman tests

Exercise9: compute Dn and Ds

i.e. compute Dn & Ds on a pair of ingroup/ outgroup sequences (e.g. BR32 vs BR29)

input: wheat_blast_brazil_outgroup.with_inv.filters.vcf

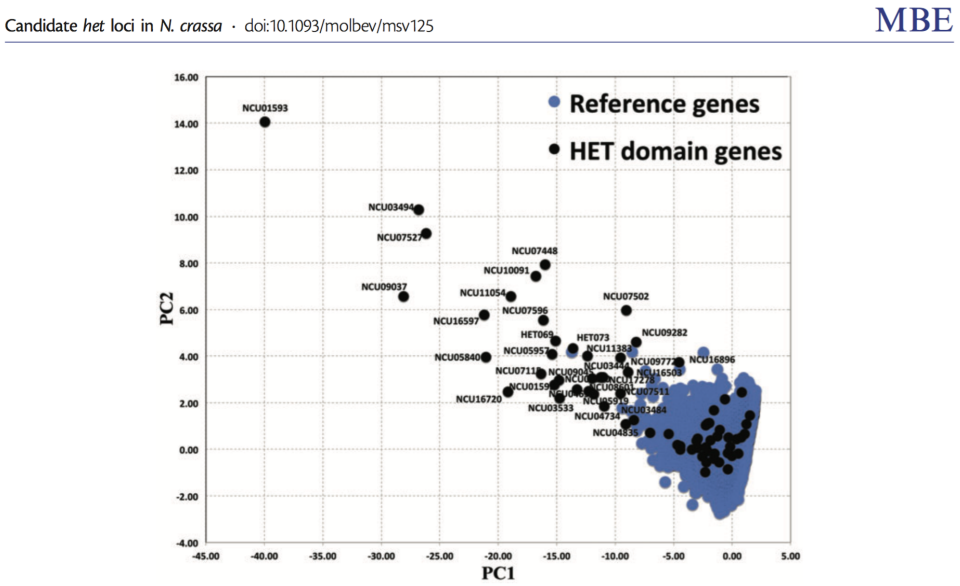
Workflow: Build scaffold genomic sequences including outgroup, Make fasta files for all genes, Compute stats

Alternative: Annotate VCF using SNPEff, write your own script to compute numbers of syn and non-syn differences between ingroup and outgroup

java -jar path_to_snpeff_jar_file/snpEff.jar -upDownStreamLen 500 -v wheat_blast path_to_vcf_file/ wheat_blast_brazil.with_inv.filters.vcf > path_to_output_SNPEff/ wheat_blast_brazil.with_inv.filters.ANN.vcf

Exercise 11: Balancing selection

- Compute S per bp, Pi and Tajima's D using Egglib
- Compute number of protein variants using Biopython
  - Use .translate() in Biopython
  - Get number of elements in my_list: len(list(set(my_list)))
- Compute maximum number of differences between sequence pairs
- Summarize results using PCA in R



http://little-book-of-r-for-multivariate-analysis.readthedocs.io/en/latest/src/multivariateanalysis.html