GenomeHarvest
diversity, organization and dynamics

Exploring the mosaic structure of rice genomes

Santos J. (joao.santos@cirad.fr), Billot C., Glaszmann J. C.

Introgression among rice populations is worth studying for its role as a path to adaptation. Along human migrations, gene flow between cultivars and wild or primitive cultivated forms have generated new types which thus harbor admixed genomes with distinct components traceable to early crop history. Recent analyses based on massive sequencing efforts have enabled detailed studies of crop evolution in rice that revealed introgressions allowing the spread of domestication factors across varietal groups as well as the secondary hybrid origin of some varietal clusters.

Yet distinct views still coexist as to the global interpretation of the data, featuring one *vs* multiple domestication events and diverse scenarios for the origin of secondary varietal groups. We present the current state of our analysis of the exchanges between the major clusters of rice genetic diversity, Japonica, Indica and *circum*Aus, as can be determined by their relative differentiation, and the concomitant retrieval of the cryptic *circum*Basmati genetic signature.

Objectives

1 - Exploring methods for efficient local ancestry assignment at the individual level.

2 – Make use of the the most extensive data set of rice genomic variation to date, the 3K RG (Alexandrov et al. 2014).

3 - Identification and characterization of regions of fixed introgression.

Why?

Correct local assignment is not made easy by the increase in available data. Despite the depth of knowledge on rice population history, many movements still escape us. Introgression from wild differentiated relatives, local selection acting on segments of the genome of particular subpopulations, as well as possible crossings between domesticated species in the search of particular phenotypes, lead to intricate scenarios and evolutionary hybrid genomes. The *circum*Basmati group (*c*Basmati, the group that encompasses the famous Basmati rices) in particular, is likely to have known all three.

In this complex scenario, we expect to find a great deal of variation in ancestry among varieties of the same group, even if they all share some identity traits. This variation should increase in conserved or otherwise neutral regions, but decrease drastically in selected regions bearing the genes responsible for those identity traits.

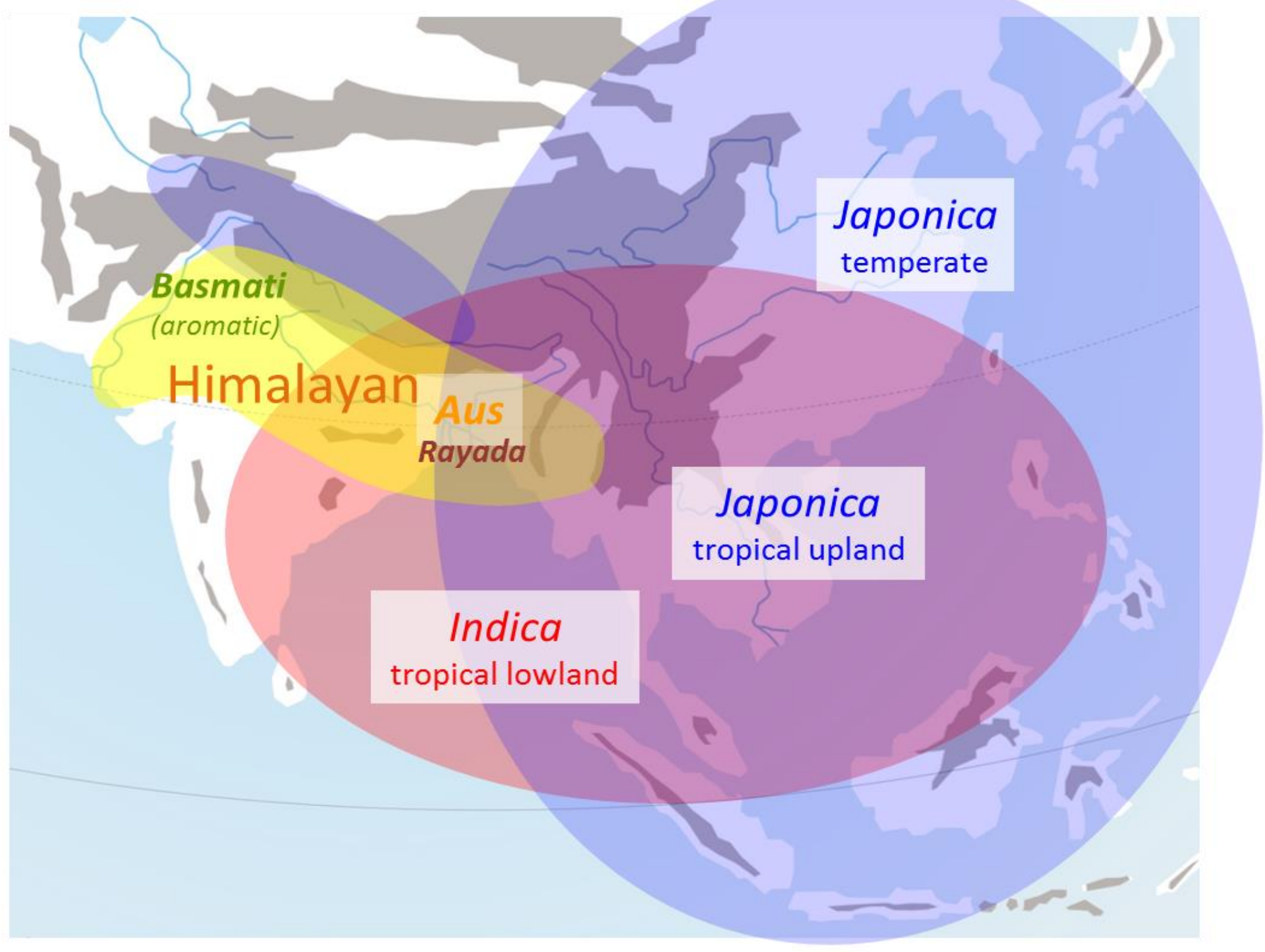
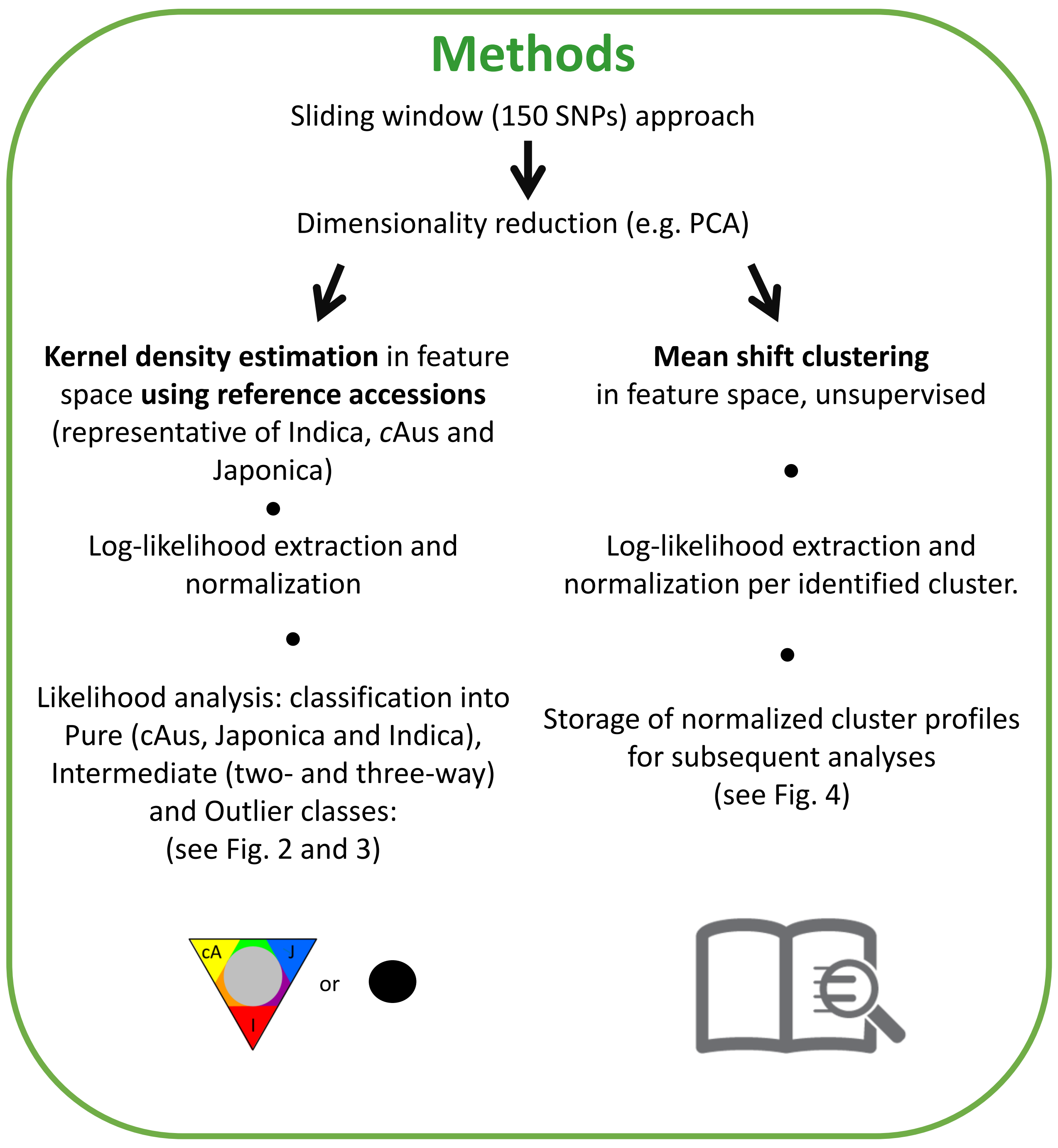


Fig. 1 - Distribution of various varietal groups in Asia. The current synthesis between studies based on isozymes and all generations of molecular markers led to the recognition of several groups. These are currently refered as: Japonica, with tropical and temperate forms; Indica; cAus, or *circum*Aus, a group of varieties including the Aus ecotype in Northeast India and Bangladesh and cBasmati, or *circum*Basmati, a group including famous aromatic varieties such as the basmati from India and Pakistan.

Methods



Sliding window (150 SNPs) approach

Dimensionality reduction (e.g. PCA)

Kernel density estimation in feature space **using reference accessions** (representative of Indica, *c*Aus and Japonica)

Log-likelihood extraction and normalization

Likelihood analysis: classification into Pure (cAus, Japonica and Indica), Intermediate (two- and three-way) and Outlier classes:
(see Fig. 2 and 3)

Mean shift clustering in feature space, unsupervised

Log-likelihood extraction and normalization per identified cluster.

Storage of normalized cluster profiles for subsequent analyses
(see Fig. 4)

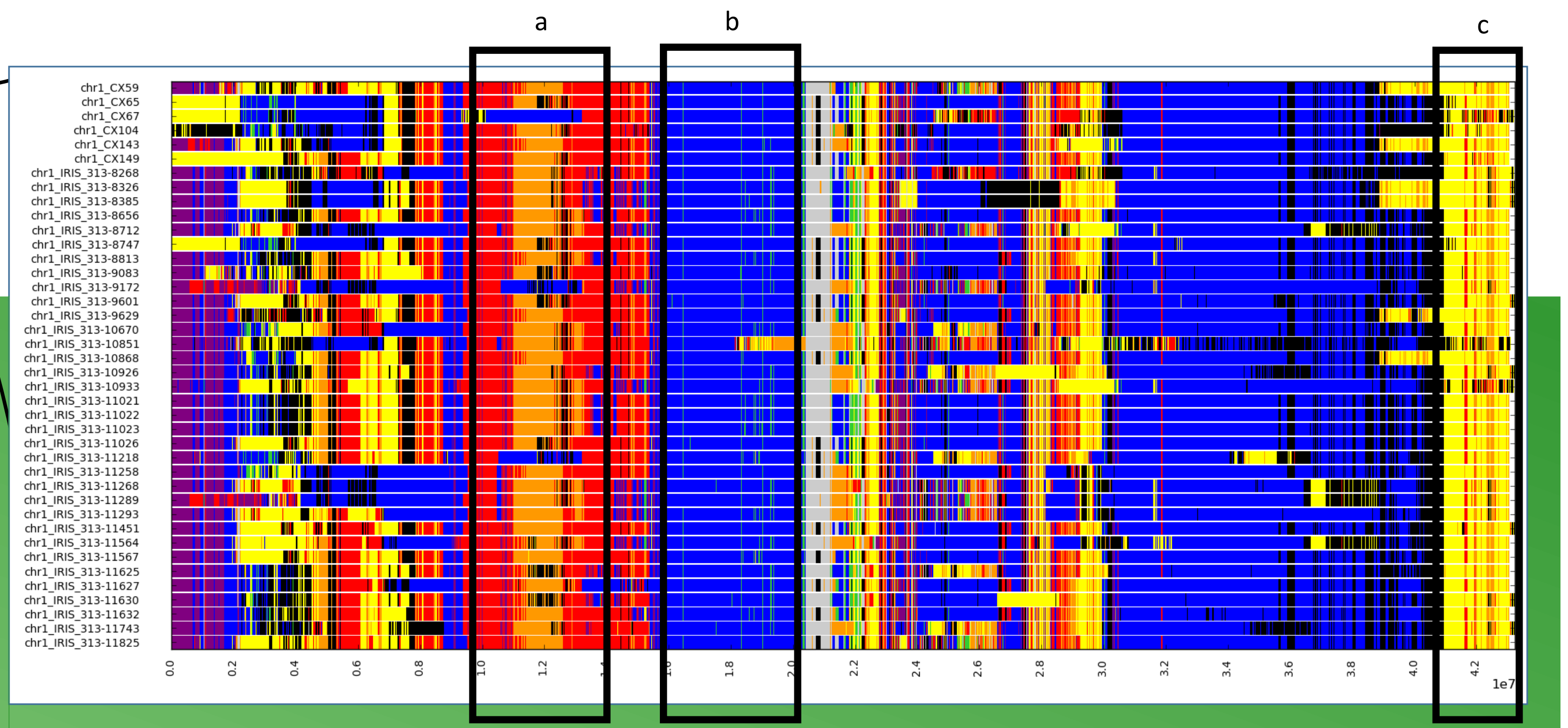


Fig. 3 - Population assignment along chromosome 1 for 40 *circum*Basmati accessions. Rectangles a, b, and c outline regions extracted for plots A, B and C in Fig. 4 respectively.

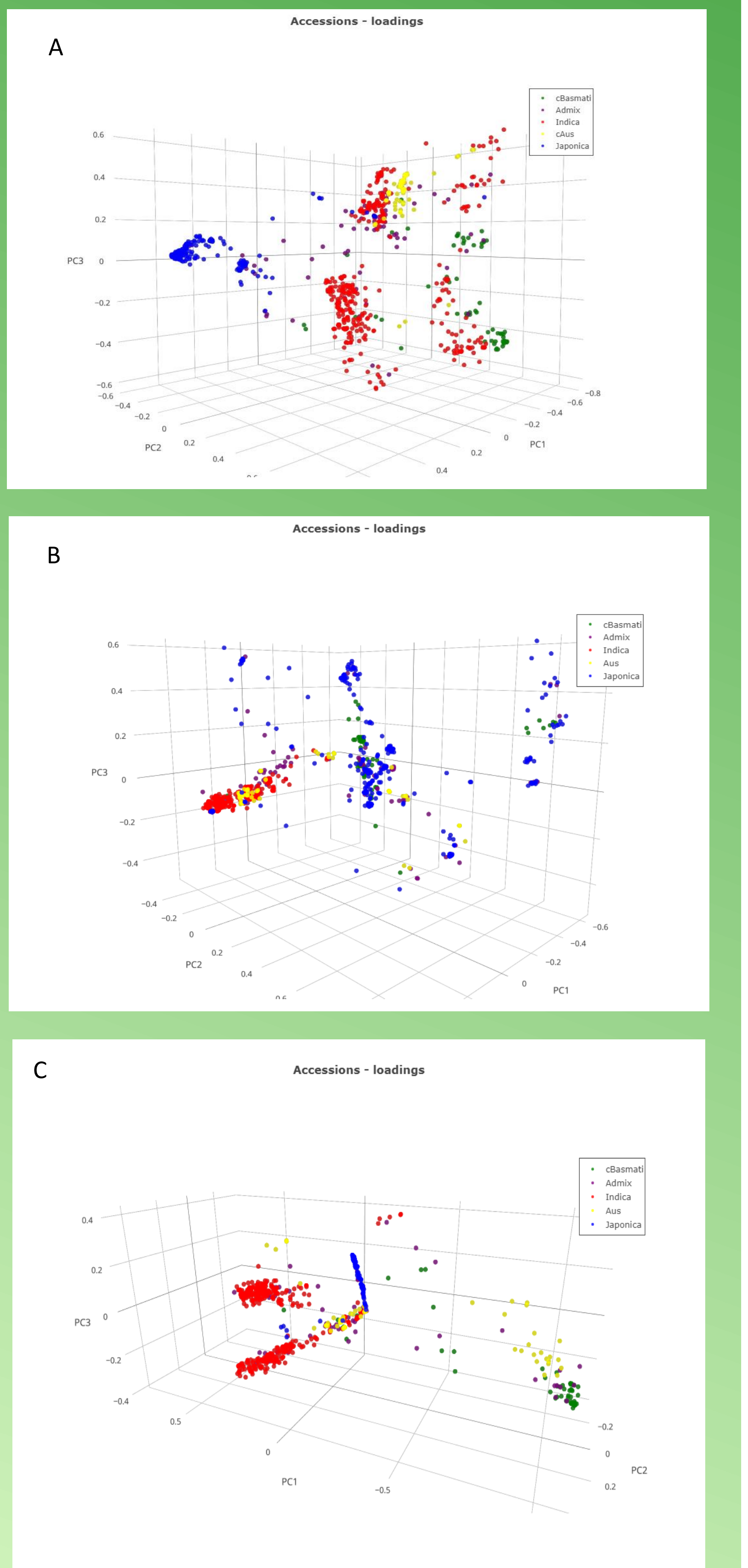


Fig. 4 - Principal component analysis of relative cluster association along regions of common ancestry - Chromosome 1
Extraction of normalized cluster profiles at windows in selected regions if over 80% of cBasmati are predicted to be assigned to target reference population. A) Indica reference; B) Japonica reference; C) *c*Aus reference. Clusters are identified using Mean Shift clustering

A – PCA on cluster profiles of Indica assigned *c*Basmati from 10 to 14 Mb of chromosome 1. The majority of *c*Basmati appear closest to a subcluster of Chinese accessions (ind1A – IRRI classification)
B – PCA on cluster profiles of Japonica assigned *c*Basmati from 16 to 20 Mb of chromosome 1. Associations to both a wider cluster of *Japonica* and Indonesian tropical Japonica are evident
C – PCA of cluster profiles of *c*Aus assigned *c*Basmati from 41 to 42 Mb of chromosome 1

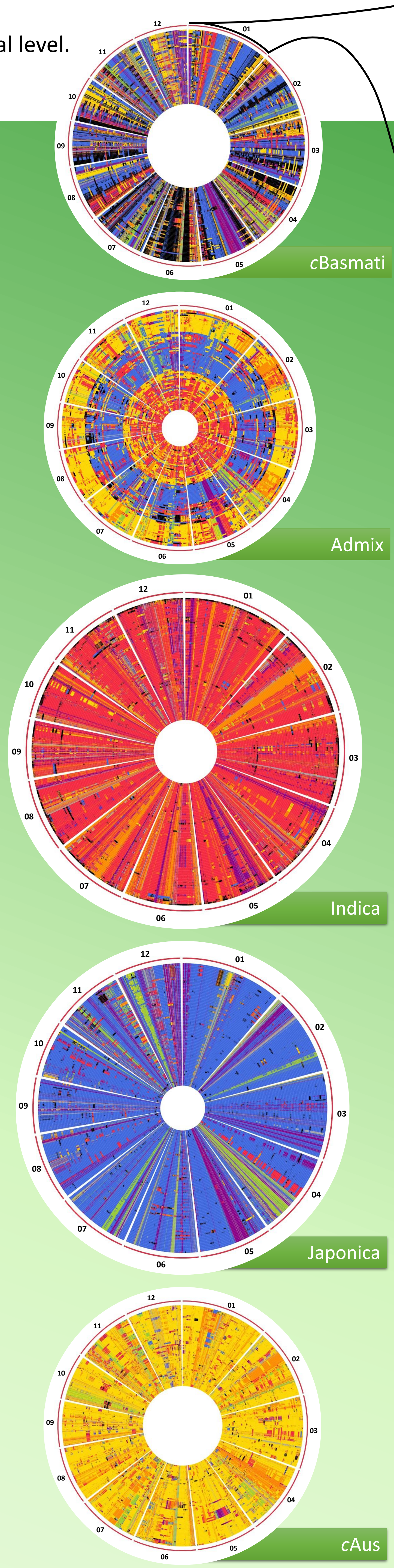


Fig. 2 - Overview of genome wide assignment across cultivated groups.
CIRCOS whole genome representation. Each segment of the genome of each accession is assigned to either one of four "pure classes" - Japonica (blue); Indica (red); *c*Aus (yellow) or X (outlier, black), or an intermediate class – allowing for two-way uncertainty: Japonica-Indica (purple); Japonica-cAus (green) and Indica-*c*Aus (orange); and three-way uncertainty: Japonica-*c*Aus-Indica (gray) (see Methods section)

Interactive Genome Exploration

Genome wide classification is only the first step in understanding the history of modern cultivated rices. The future of population genomics lies in integrating our analysis of increasingly large data sets with modern data visualisation tools.

Go to https://github.com/Joaos3092/PAG_2018 to explore further.

Take away

- *c*Basmati varieties share ancestry across large segments of chromosome 1.
- Dissecting the genome into segments of relative ancestry provides a useful way to further explore connections in the complex scenario of rice domestication.
- From a Functional Genomics standpoint, assessing local ancestry can increase the power of future association study designs.

Reference Alexandrov N., Tai S., Wang W., Mansueto L., Palis K., Fuentes R. R., ... & Mauleon, R. (2014). SNP-Seek database of SNPs derived from 3000 rice genomes. *Nucleic Acids Research*, *43*(D1), D1023-D1027.

(1) CIRAD, UMR AGAP, F-34398 Montpellier, France.
(2) AGAP, Univ Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France.

cirad AGRICULTURAL RESEARCH FOR DEVELOPMENT
agropolis fondation
IRIC International Rice Informatics Consortium
The 3000 Rice Genome Project
BGI IRRI GRiSP CAAS
Rice Science for a Better World