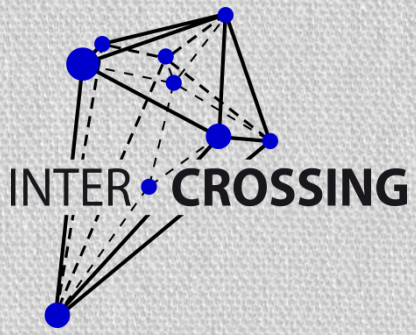
INTROGRESSION AMONG BRITISH BIRCH TREES

New method for genotyping polyploids & comparison of microsatellite and RAD loci.



Jasmin Zohren, 11/11/2015

Thanks to...

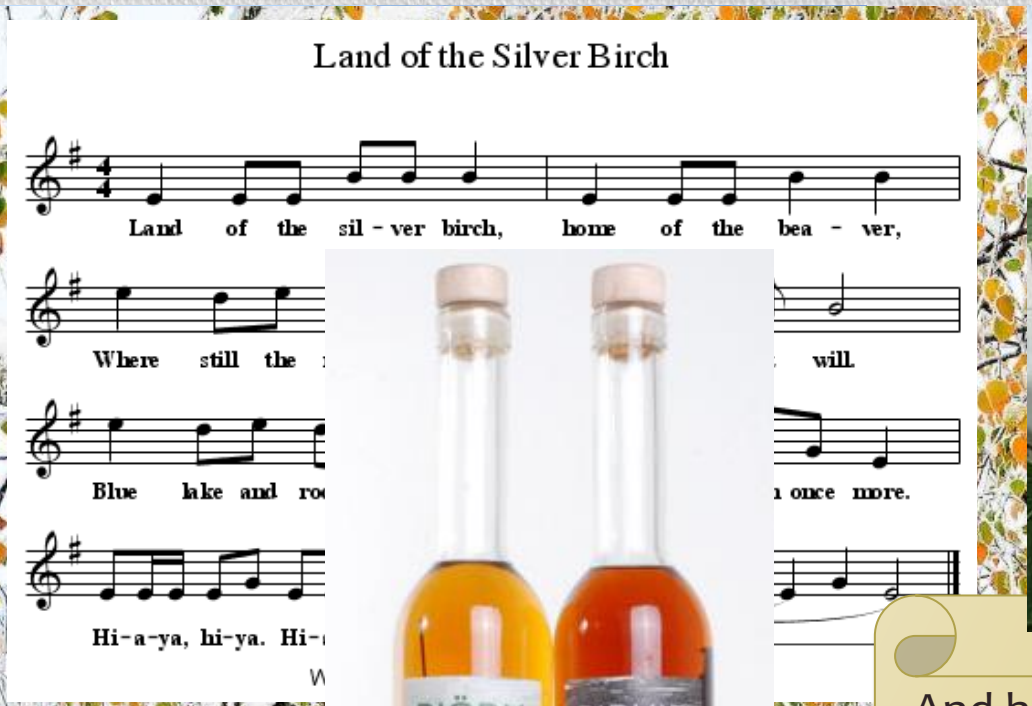
- Igor Kardailsky
- Anika Joecker
- Anne-Mette Krabbe Hein
- Lizzy Sollars

- Richard B
- Richard N
- Nian
- James

Funding:

- EU (INTERCROSSING)
- Danish Council for Strategic Research (MASPOT)

Study organism: Birch trees (*Betula*)

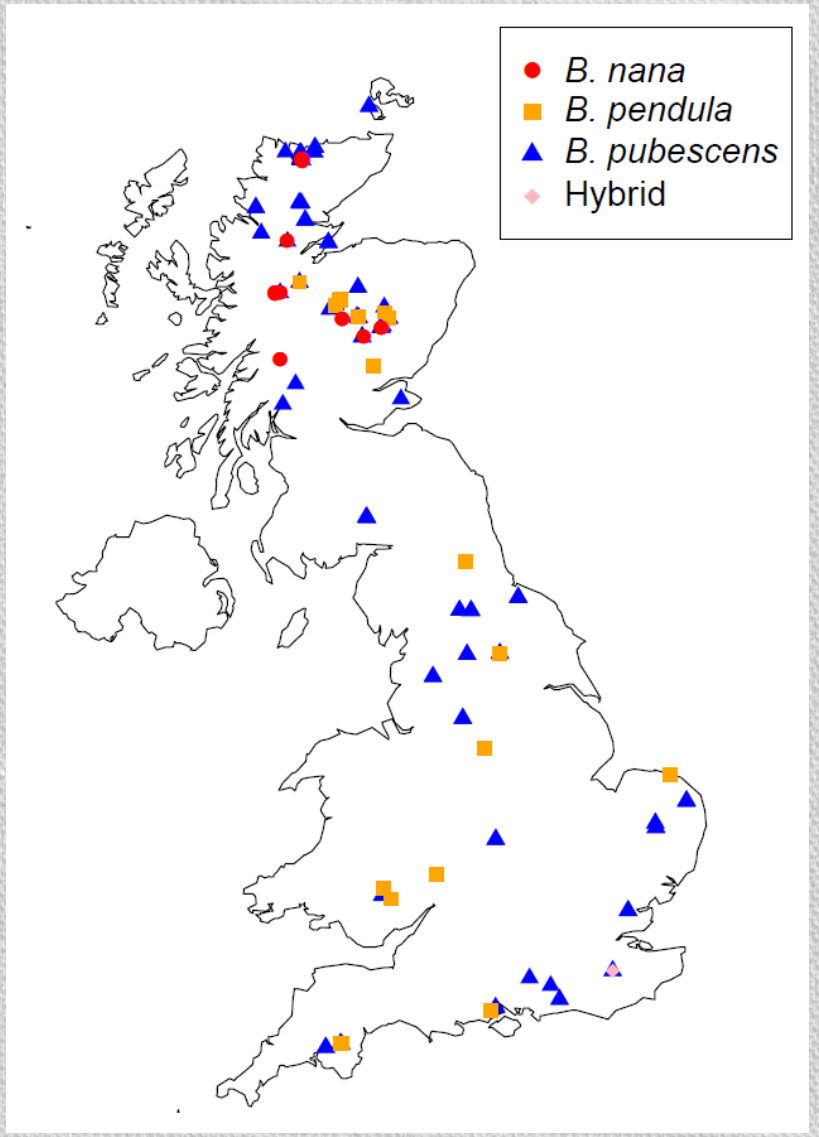


And hark, the noise of a near waterfall!
I pass forth into light - I find myself
Beneath **a weeping birch (most beautiful
Of forest trees, the lady of the woods)**
Hard by the brink of a tall, weedy rock
That overflows the cataract.

"The Picture or The Lover's Resolution"
by Samuel Taylor Coleridge (1802)

Data

- 205 individuals from the UK
- Three species:
  - *Betula nana* (dwarf birch), diploid, restricted to Scotland
  - *B. pendula* (downy birch), diploid, widespread
  - *B. pubescens* (silver birch), tetraploid, widespread



Genomic data

- DNA from dried cambium and leaves
- Restriction-site associated DNA (RAD) PstI libraries, 96 bp paired-end and 42 bp single-end
- 1.4 billion raw reads, ~14x coverage

- Reference sequence: RAD sites in 12 *Betula* species (2x to 12x)

Workflow

- Read mapping and variant calling in CLC Genomics Workbench (GWB)
- Genotyping and filtering using R

- PolyTypeR script:
  - handles various ploidy levels
  - Calculates most probably allelic configuration (e.g. “AAA”, “AAB”, “ABB” in a triploid)
  - Uses Log-likelihood model
  - Computes Bayes factor as quality measurement

- Output in STRUCTURE format

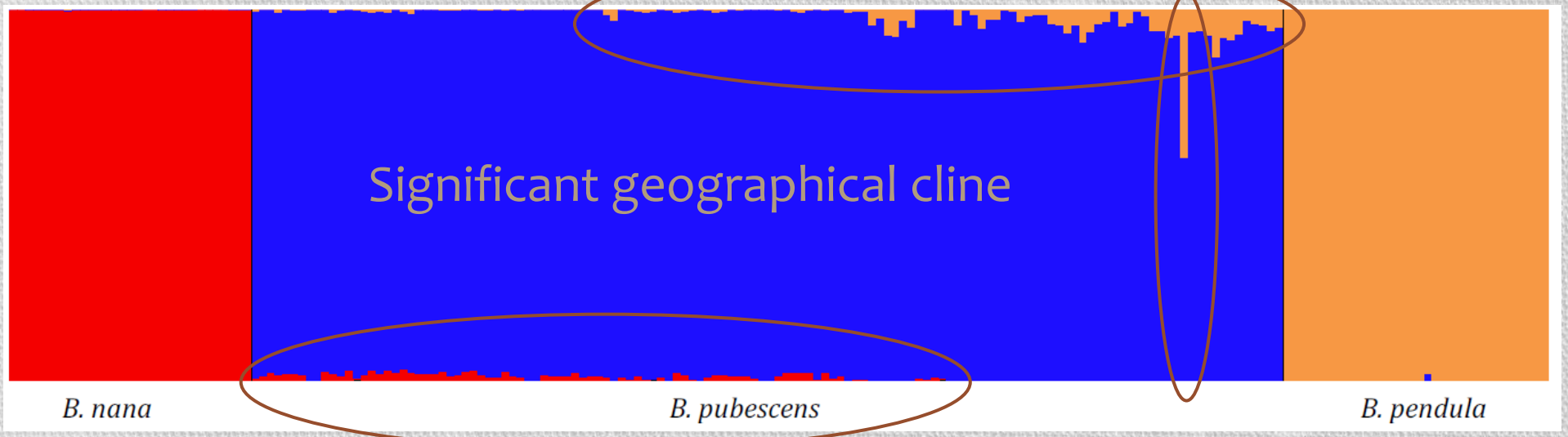



Results

- Almost four million ‘raw’ variants
- Filters:
  - >1 mio raw reads per individual
  - only SNVs were kept (including deletions)
  - <50% missing data allowed
- Remaining variants: 645,175
- Variants in 80%: 76,587
- Variants in all: 9,528

Structure results

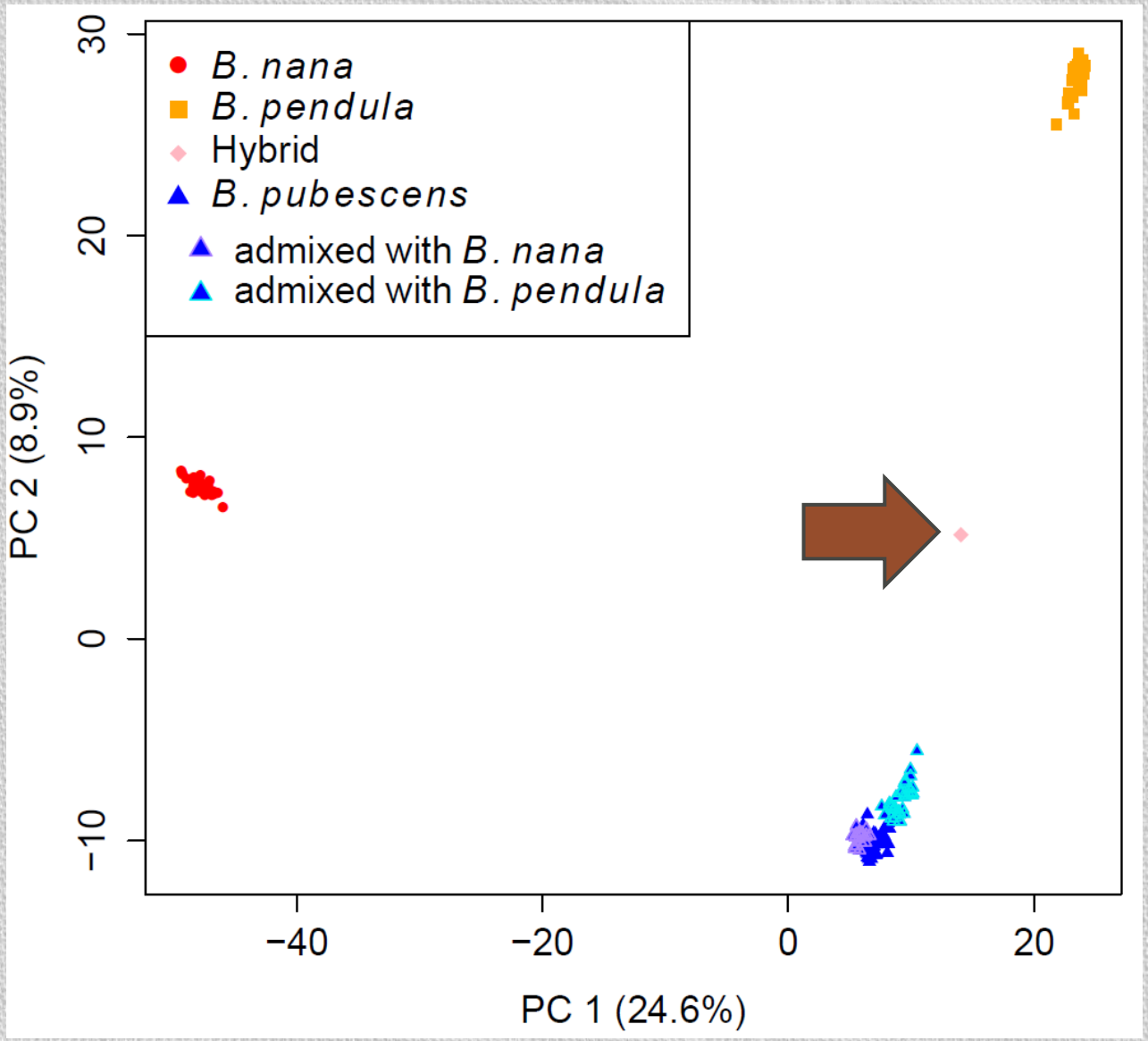
- 203 samples, 76,587 variants ("80% data set")
- 10,000 burn-ins and 100,000 repeats, K = 3
- Within species ordered from ~~north (left) to south~~ (right)



F1 hybrid

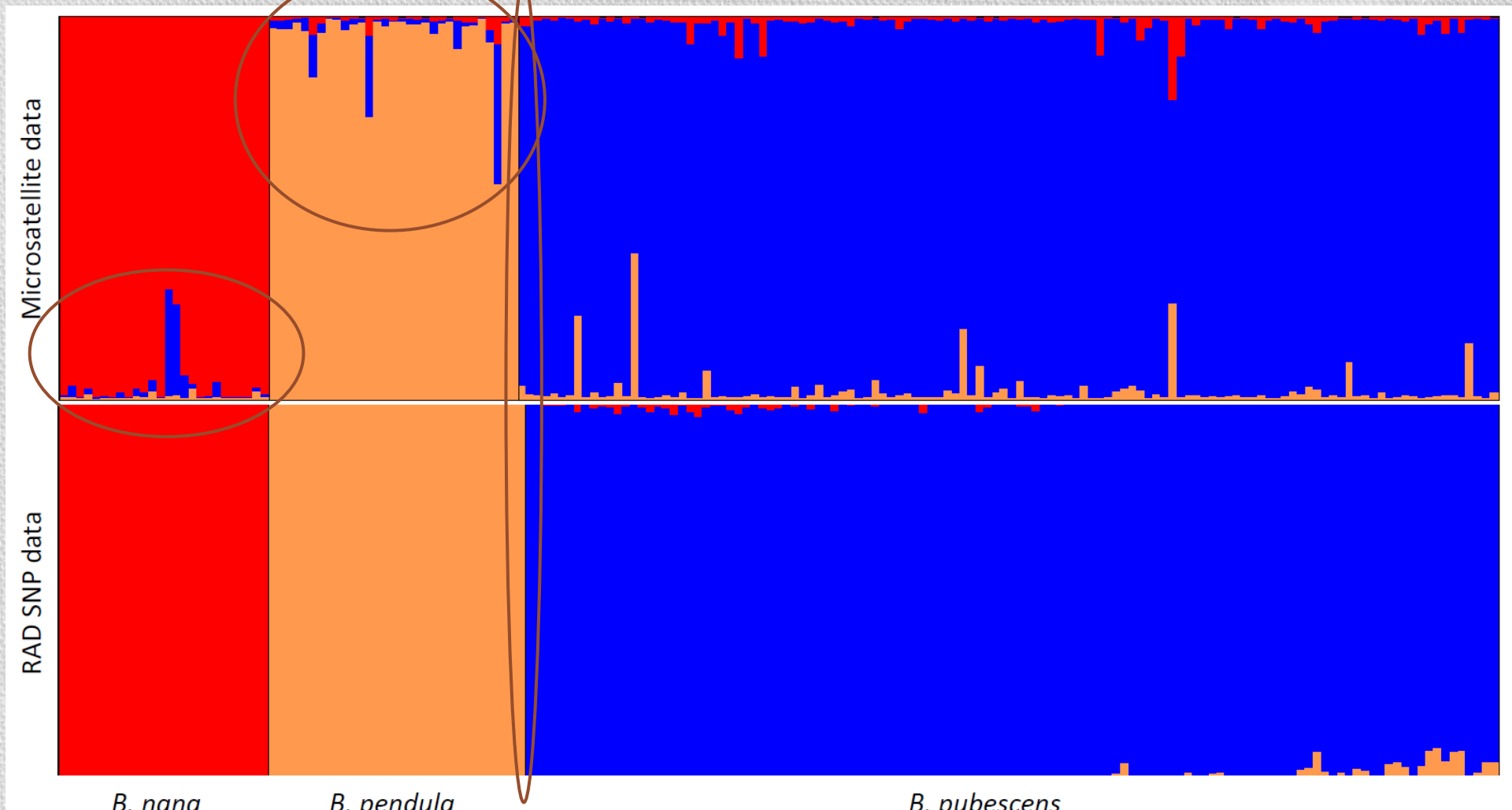
Principal Component Analysis

- 203 samples
- 74,084 variants (biallelic 80% set)
- "admixed" individuals at least 2% (from STRUCTURE)
- Using "adegenet" package in R, missing data imputed with "missMDA" package in R



Microsatellite (MS) vs. RAD data 1/3

- 179 samples; 12 MS loci; 9,523 RAD variants ("in all" data set); K = 3

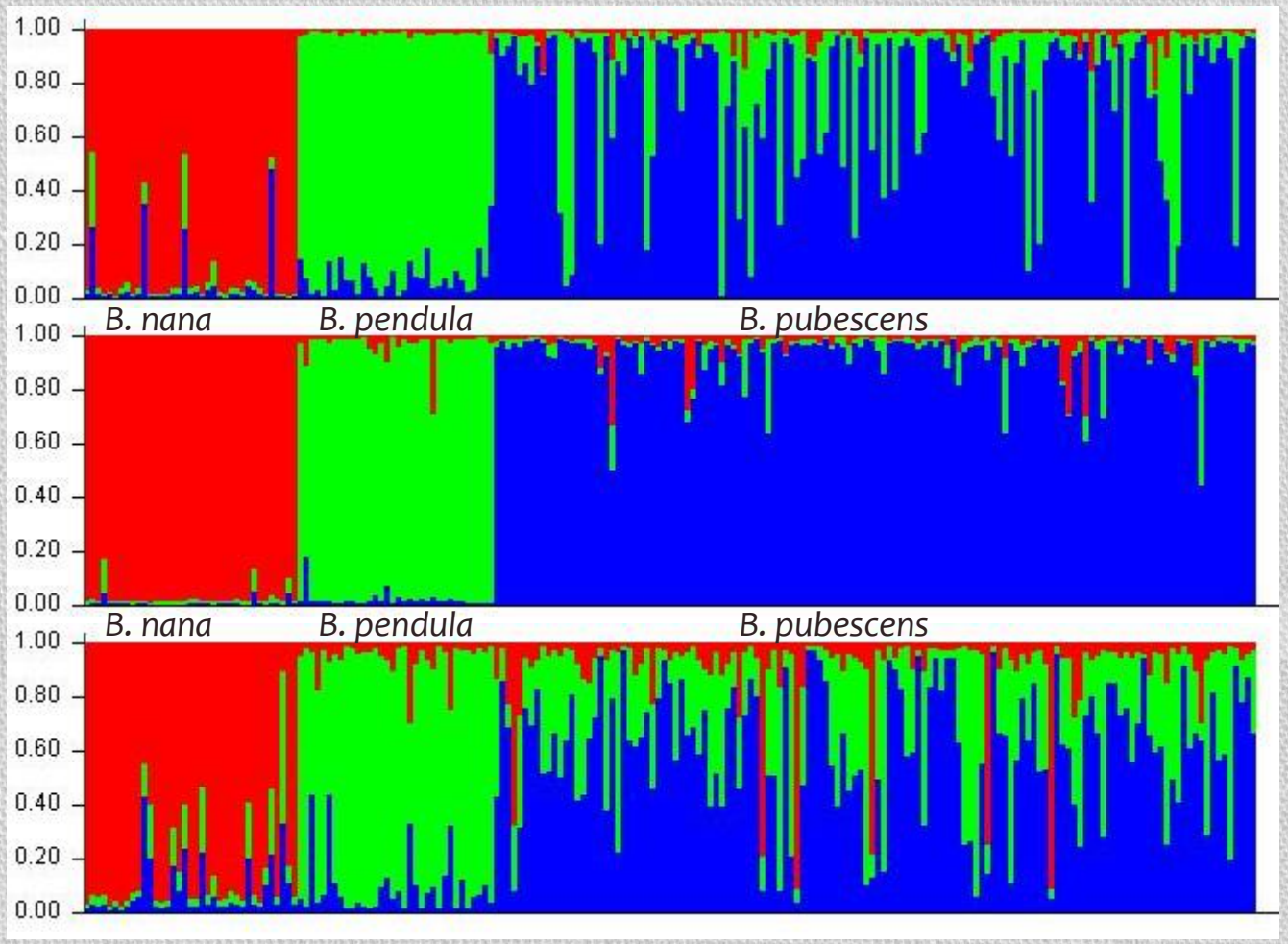


Microsatellite vs. RAD data 2/3

- RAD variants more accurate? (more and widely distributed)
- RAD variants linked to loci under selection?
- MS loci mutate faster, thus reflect more recent hybridisation?
- RAD variants distinguish better between *B. pendula* and *B. pubescens*
- Homoplasy more common in MS markers?

Microsatellite (MS) vs. RAD data 3/3

- 203 samples; 24 RAD variants (randomly from “in all” data set); K = 3



Thank you!

- Summary
  - Variant calling in CLC
  - Genotyping with PolyTypeR
  - Comparison with microsatellite markers
  - Structure
  - PCA
  - Cline analysis

- Outlook:
  - Detailed investigation of hybrid (ploidy and MS)
  - Identification of introgressed loci
  - Improvement and annotation of *B. nana* genome

Possible allele dosage

Triploid (3n)

- 3:0:0
- 2:1:0
- 1:1:1

Tetraploid (4n)

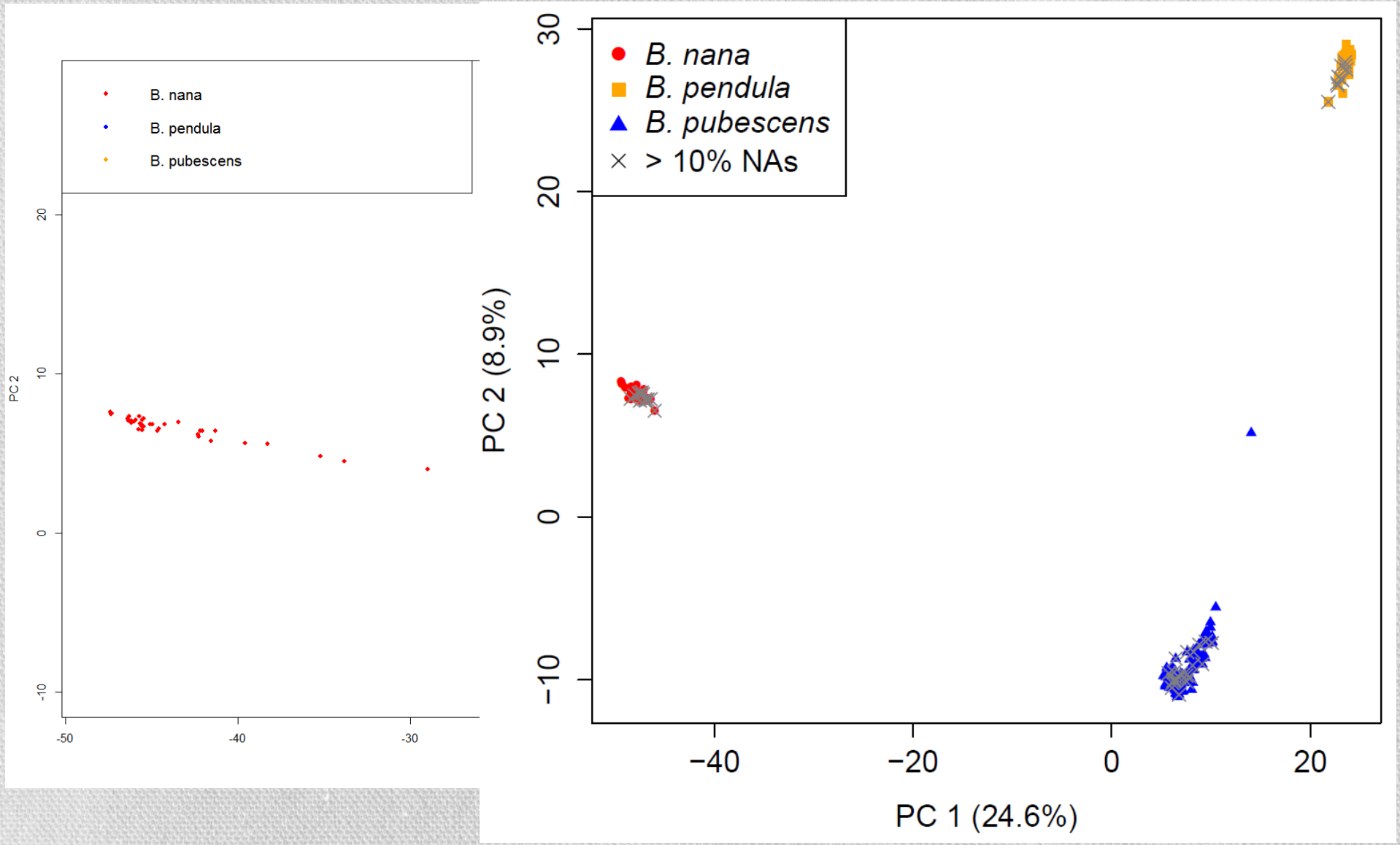
4:0:0:0
3:1:0:0
2:2:0:0
2:1:1:0
1:1:1:1

Pentaploid (5n)

5:0:0:0:0
4:1:0:0:0
3:2:0:0:0
3:1:1:0:0
2:2:1:0:0
2:1:1:1:0
1:1:1:1:1

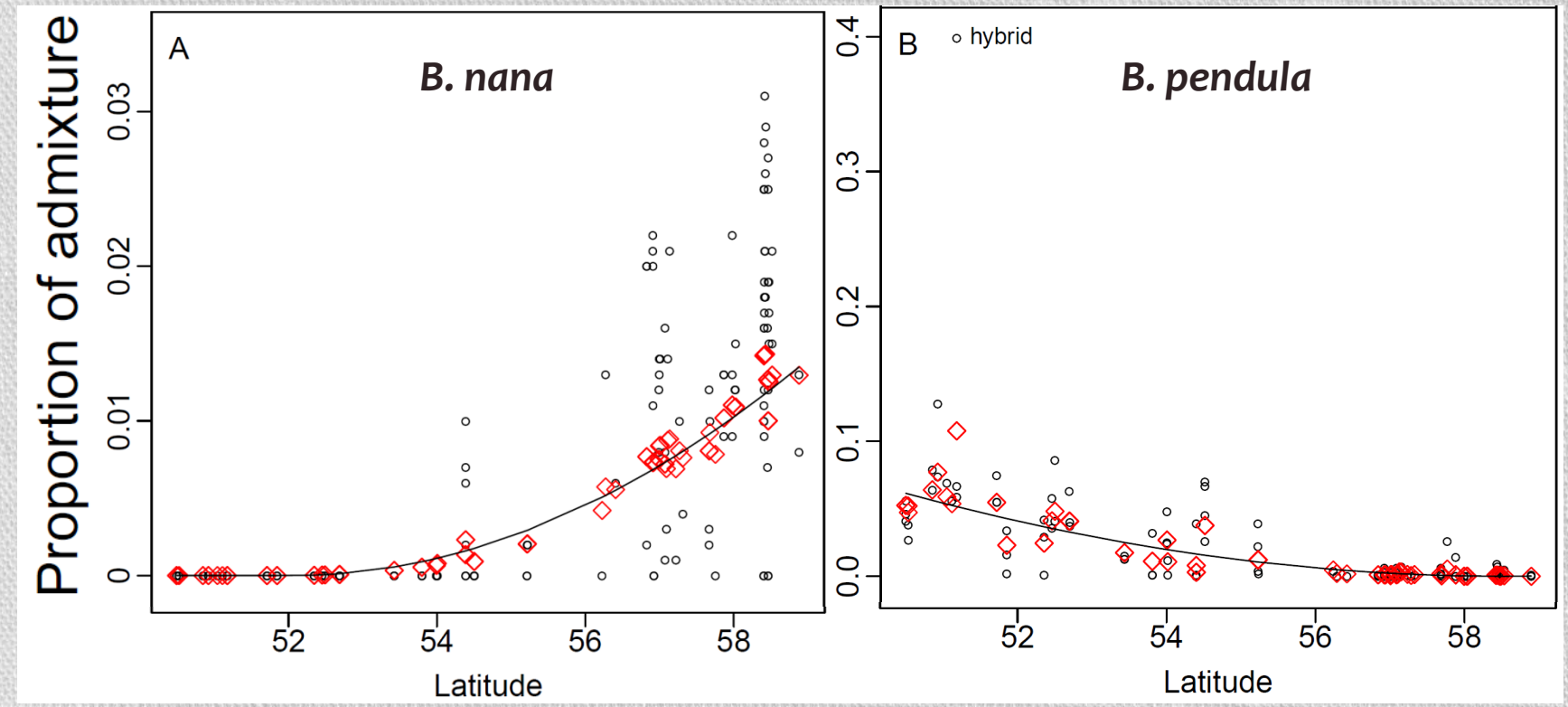
Hexaploid (6n) etc. likewise

Missing data in PCA



Cline analysis

- Arcsine transformed admixture values
- Mixed effects model with population as random effect
- Red diamonds = population means



Microsatellite vs. RAD data 2/3

- 179 samples
- 12 MS loci
- 9,523 RAD variants
- Q-values from STRUCTURE