Guidelines on combining genomic data across populations

R. Bonifazi, S. Aivazidou, A. Bowman, F. Macedo, R.F. Veerkamp, K. Quigley, T. Pabiou, R. Evans, H. Bunning, G.M. Tarekegn, P. Davoudi, K. Matthews, J. Layton, R.G.M.F. Damai, R. Finocchiaro, M. Cassandro, A. Quaglia, D. Vicario, L. Degano, M. Piazza, T. Roozen, M. Coffey, M. Burke, J. Vandenplas

Interbull AM, Verona, June 1st, 2026

M3GE project

- **Goal**: Develop multi-breed multi-trait international genomic beef cattle evaluations
- Potential advantages:
  - Improve GEBVs accuracy by including data on (other) purebreed and crossbred
  - Leverage international collaboration for novel traits
  - Involve new (numerically-small and local) breeds (better usage of genetic resources, e.g., transboundary breeds)
  - Deliver GEBVs to breeders for traits not evaluated at national level

3 countries

7 national organizations

WAGENINGEN
UNIVERSITY & RESEARCH

Data collected

- Two trait groups:
  - FEED: Feed + indicator traits
  - LONG: Longevity + indicator traits
- Beef purebred, Crosses (Beef x Beef & Beef x Dairy), Dairy purebred (AI sel. candidates)
- Genotypes:
  - ~3.2M individual (imputed) genotypes
  - medium/high density (+30K SNPs)

Pedigree and (repeated) phenotypes

Genotypes

GENOEX
International Genotype Exchange Platform
International Genotype Exchange platform for Animal Breeding

WAGENINGEN
UNIVERSITY & RESEARCH

Chips available

- 11 chips in total (28k – 777k) from 2 assemblies (UMD 3.1, ARS1.2)
- Different chips for **different breeds**
- Different overlap between chips
- ~815k unique SNPs across all chips & ~160k unique SNPs across 2+ chips
- **~6.5k SNP in common** across all chips

	COU1_chip1*	COU1_chip2*	COU1_chip3	COU1_chip4	COU2_chip1	COU2_chip2	COU2_chip3	COU2_chip4	COU2_chip5	COU3_chip1	COU3_chip2
COU1_chip1*	132,665										
COU1_chip2*	26,952	28,289									
COU1_chip3	60,874	14,812	64,888								
COU1_chip4	37,420	9,582	37,555	42,172							
COU2_chip1	34,670	11,716	23,964	14,091	41,855						
COU2_chip2	37,863	10,491	38,040	37,685	14,745	44,460					
COU2_chip3	39,280	11,016	38,940	38,640	15,154	44,165	48,862				
COU2_chip4	39,280	11,016	38,940	38,640	15,154	44,165	48,862	48,862			
COU2_chip5	37,863	10,491	38,040	37,685	14,745	44,460	44,165	44,165	51,828		
COU3_chip1	37,796	10,511	37,990	37,564	14,689	44,210	43,922	43,922	50,390	50,493	
COU3_chip2	116,260	24,961	60,118	35,716	37,420	38,064	41,877	41,877	38,064	38,065	667,044
Unique SNPs	491	139	44	127	-	-	-	-	1,188	-	535,040

WAGENINGEN UNIVERSITY & RESEARCH

*Same chip for ITA breeds, LIM&CHA

Genomic data harmonization

Information collected:

- Maps: `SNP name, chr, bp` (`rsID` if available)
- Genotypes: `0/1/2/5` following Illumina AB coding (GenoEX = `A` counts)

To combine genomic data across populations, ensure that:

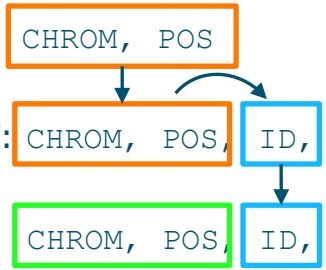
1. Maps are aligned across chips
2. Genotype information is aligned with corresponding map

1. Maps alignment

- Ensure genomic data is on same assembly version (UMD 3.1)
- Program to convert map information between assemblies
- "rs-ID bridging" approach:

	Chip1	Chip2
unmatched in source VCF	2%	8%
unmatched in target VCF	6%	6%
matched in both VCF	93%	86%

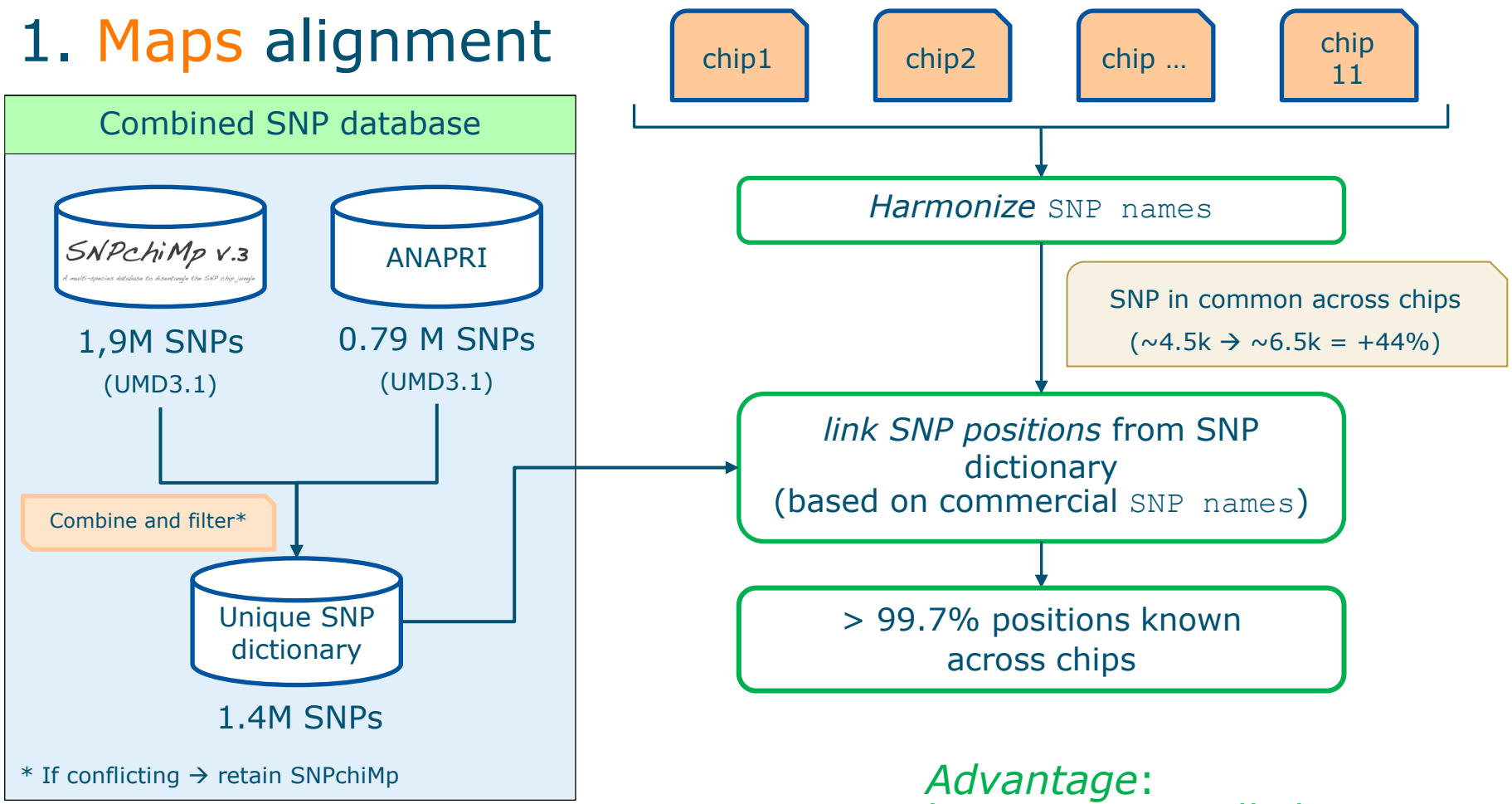
1. Map file: `SNP name, CHROM, POS`
2. Source assembly VCF: `CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO`
3. Target assembly VCF: `CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO`



\+ straightforward, fast, flexible, works well for arrays with dbSNP `IDs`

\- relies on `rsID` availability and consistency across assemblies, needs exact position matching (→ liftover approaches)

1. Maps alignment

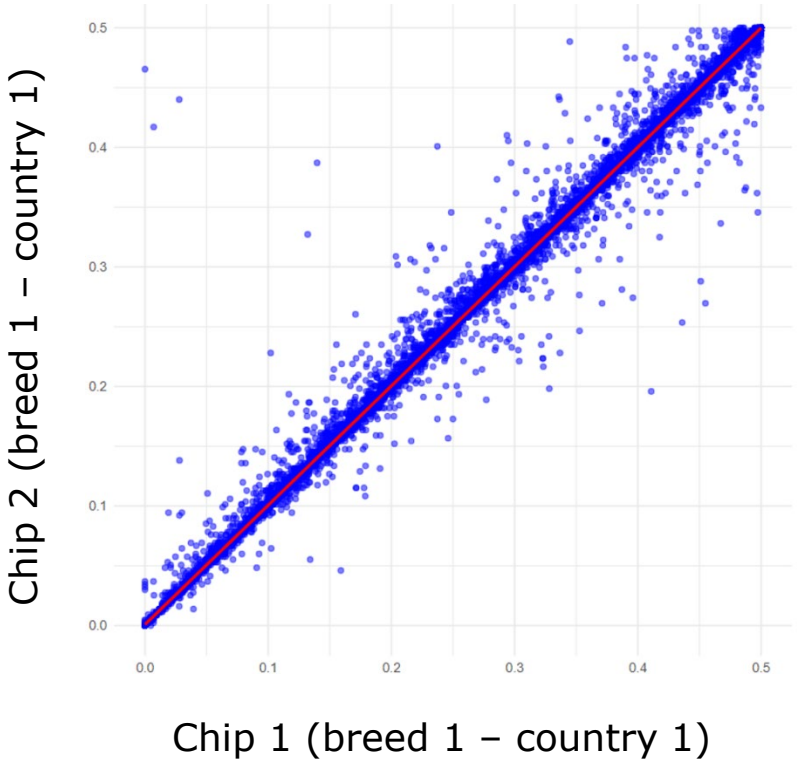


Advantage:
same coordinates across all chips

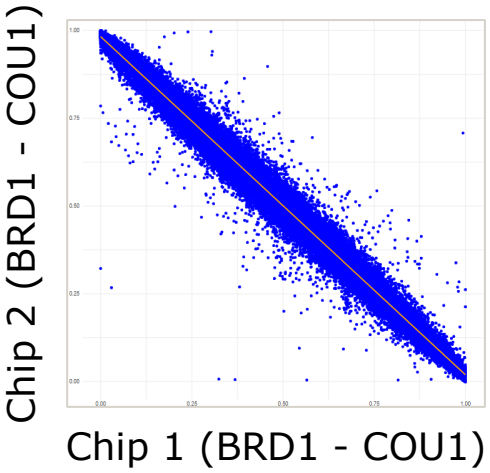
WAGENINGEN
UNIVERSITY & RESEARCH

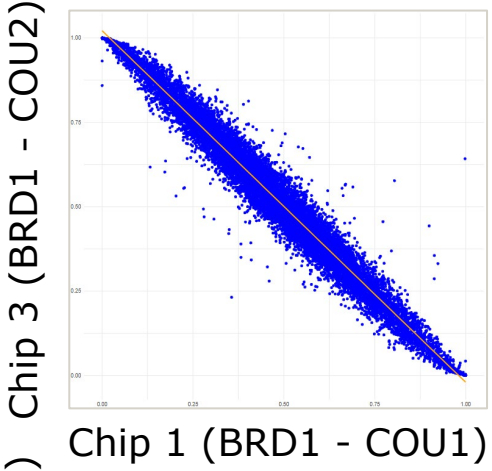
2. Genotype information alignment

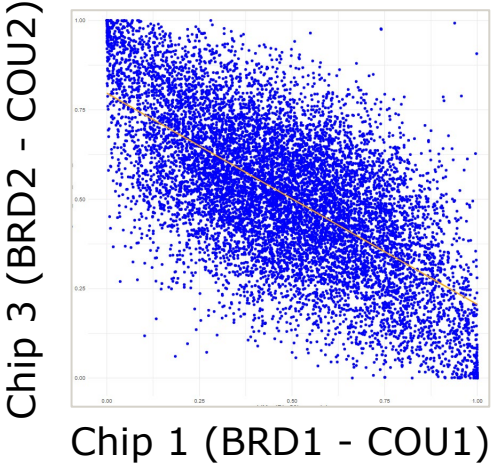
- Validate agreement:
  - between genotypes and corresponding map information
  - allele coding across chips
- Allele frequencies as a diagnostic tool
- Comparing chips (overlapping SNPs), two levels:
  - Within-breed within-country (possibly same animals), within-breed across-country
  - Across-breed across-country (more noisy but still useful)

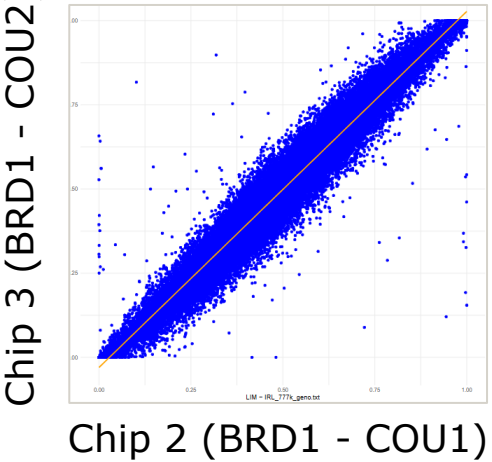


Examples of detected misalignments



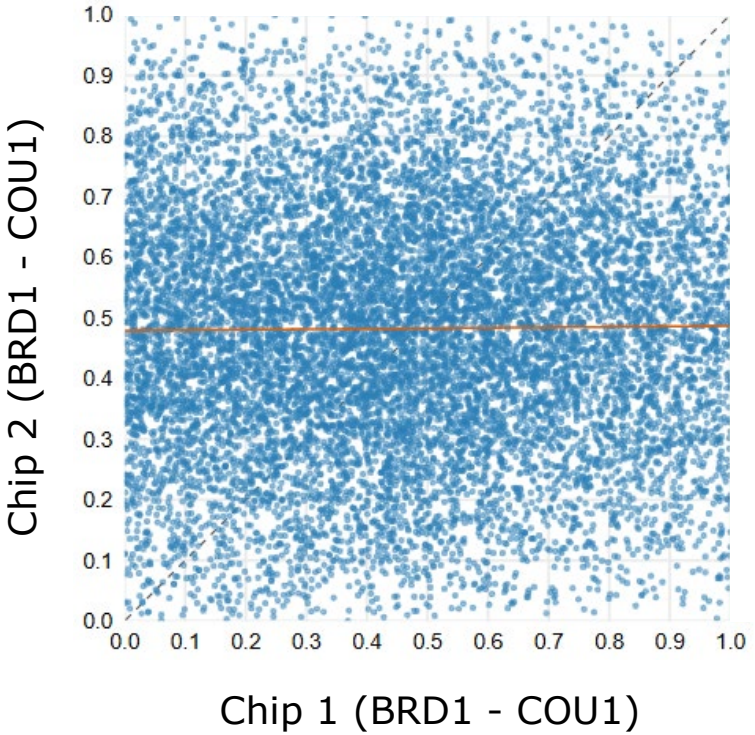






Allele coding switch in chip 1

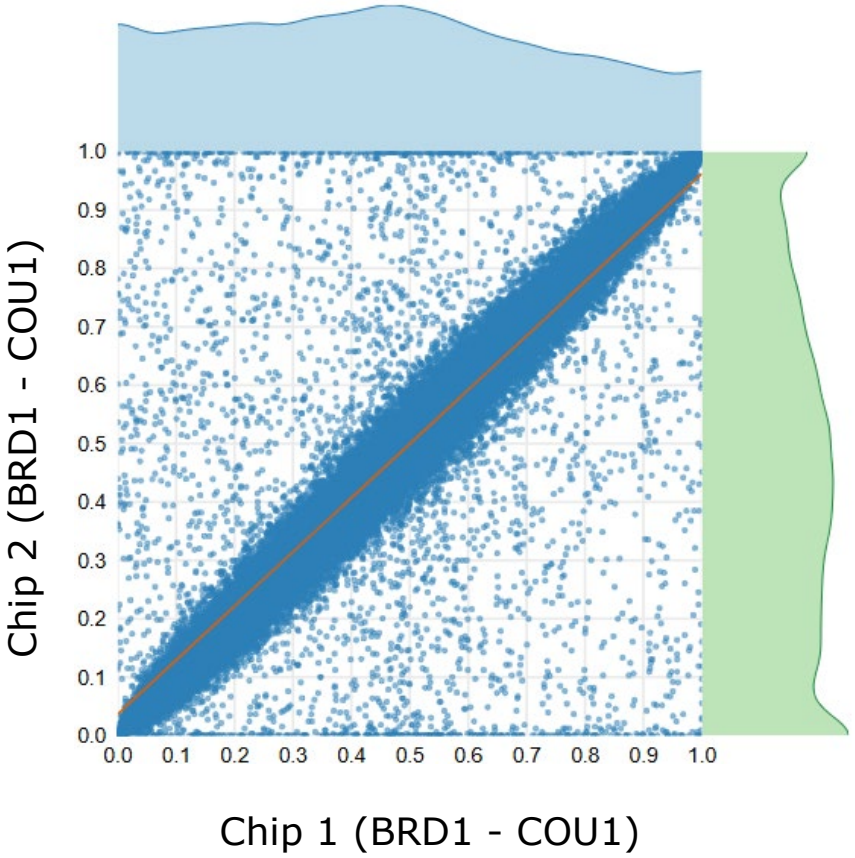
Examples of detected misalignments



Misaligned chip 2

ρ = 0.01
RMSE = 0.34

Examples of detected misalignments



Shift of +1 on SNP position

ρ = 0.90
RMSE = 0.12

Guidelines and recommendations

1. **Maps alignment**
   - Use public resources when possible (SNPchiMp)
   - Store `rsID` and assembly version
   - Keep complete manifest files (older chips not in public domain)
   - Add support for metadata to GenoEX (`rsID`, assembly version, manifest)

2. **Genotypes alignment**
   - Use consistent Illumina AB coding
   - Use AF checks as a diagnostic tool
   - Automated AF checks (ρ and RMSE) → detect issues on submission
   - "Already-approved" genotypes as reference

Implications

Unravelling the Genetic Structure of Local and Mainstream Red-Pied Cattle Breeds Using Genomics and Extended Pedigree Analysis

Margaux Vandewalle[1] | Renzo Bonifazi[2] | Jérémie Vandenplas[2] | Sipke-Joost Hiemstra[3] | Gerben de Jong[4] | Dirk Hinrichs[5] | Nicolas Gengler[1] | Hélène Wilmot[6]

- At international level
  - Projects needing collaboration (e.g., M3GE)
  - International evaluations (e.g., InterGenomics, SNP-MACE)
  - Studies across countries/organizations
  - Genotype exchange (e.g., GenoEX) for genomic evaluations or parentage analysis
- Public resources are of great help (SNPchiMp; Schnabel 2018) but are phasing out or incomplete
  - Need for new and updated tools → Let's talk!

Take-home messages

- Pooling and aligning genomic data can be challenging and prone to mistakes
- Guidelines and recommendations to avoid pitfalls
- AF diagnostic for quality control
- Public resources useful but phasing out

Thank you for your attention

WAGENINGEN
UNIVERSITY & RESEARCH

renzo.bonifazi@wur.nl