Predicting autozygosity from runs of homozygosity

Comparison of results from next generation sequence versus high density SNP chip data for Nellore bulls

J. Sölkner, M. Milanesi, N. Khayatzadeh,
A.T. Utsunomiya, I. Curik, M. Ferencakovic,
P. Ajmone Marsan, J.F. Garcia,
Y.T. Utsunomiya

Background

Estimating autozygosity from high-throughput information: effects of SNP density and genotyping errors

Maja Ferenčaković[1], Johann Sölkner[2*] and Ino Curik[1]

Genetics Selection Evolution 2013, **45**:42

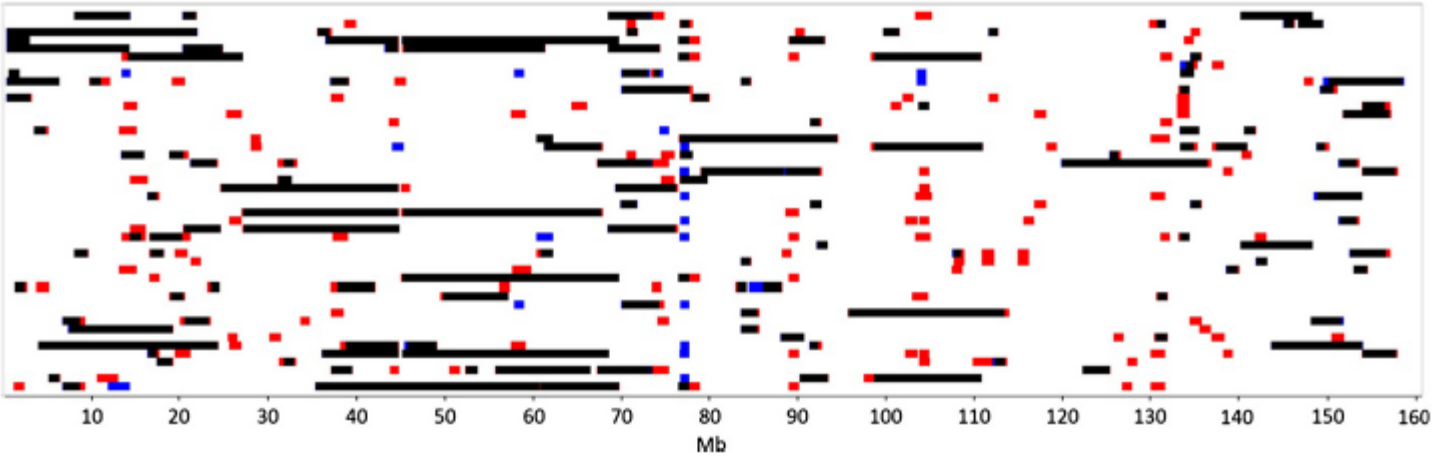


Figure 1 Overlay of ROH identified on chromosome 1 in Brown Swiss animals. ROH were identified using 50 k and HD panel data and then overlaid; each row represents one animal, and different colors were used to indicate whether ROH segments were identified using both the 50 k and HD panel (black), only the 50 k panel (red), or only the HD panel (blue).

Research question

- SNP chip data contain very few wrongly called genotypes
- Higher genotyping error rate in NGS data
- How do we call runs of homozyosity (ROH) with NGS data?
  - Allow higher rate of heterozygous call in a ROH?
  - Apply different methodology?

Data

- **21 Nellore bulls**
- Paired-end sequencing with Illumina® HiSeq2000
- Average **coverage of 9.5**±1.9x (6.1x - 13.8x)
- Quality control with FastQC
- Alignment against UMD v3.1 with BWA
- SNP calling: Samtools mpileup; lenient filtering
- **26,421,727 autosomal bi-allelic SNPs**
- **Concordance with HD data: 97.93±1.06%** (94.77% - 99.24%)
  - Highly correlated with coverage (r = -0.98)

Methods for ROH detection

- **PLINK v1.90**
  - Deterministic method based on sliding windows
  - Designed for array data
  - Genotyping error can only be modeled through number of heterozygous genotypes
- **BCFtools v1.3.1**
  - Hidden Markov Model: autozygosity/non-autozygosity are the hidden states
  - Designed for exome capture data
  - Genotype probablilities are explicitly modeled as components of emission probabilities (using phred scores)

ROH detection in PLINK

- Length in Mb = (1,2], (2,4], (4,8], (8,16] and >16
- Minimum # of SNPs depending on length
  - 75 for HD, 1877 for NGS for (1,2]
- Maximum # of heterozygotes = (error rate)*(min. # SNP)

- **error rate was varied: 1, 5 and 10%**

ROH detection in BCFtools

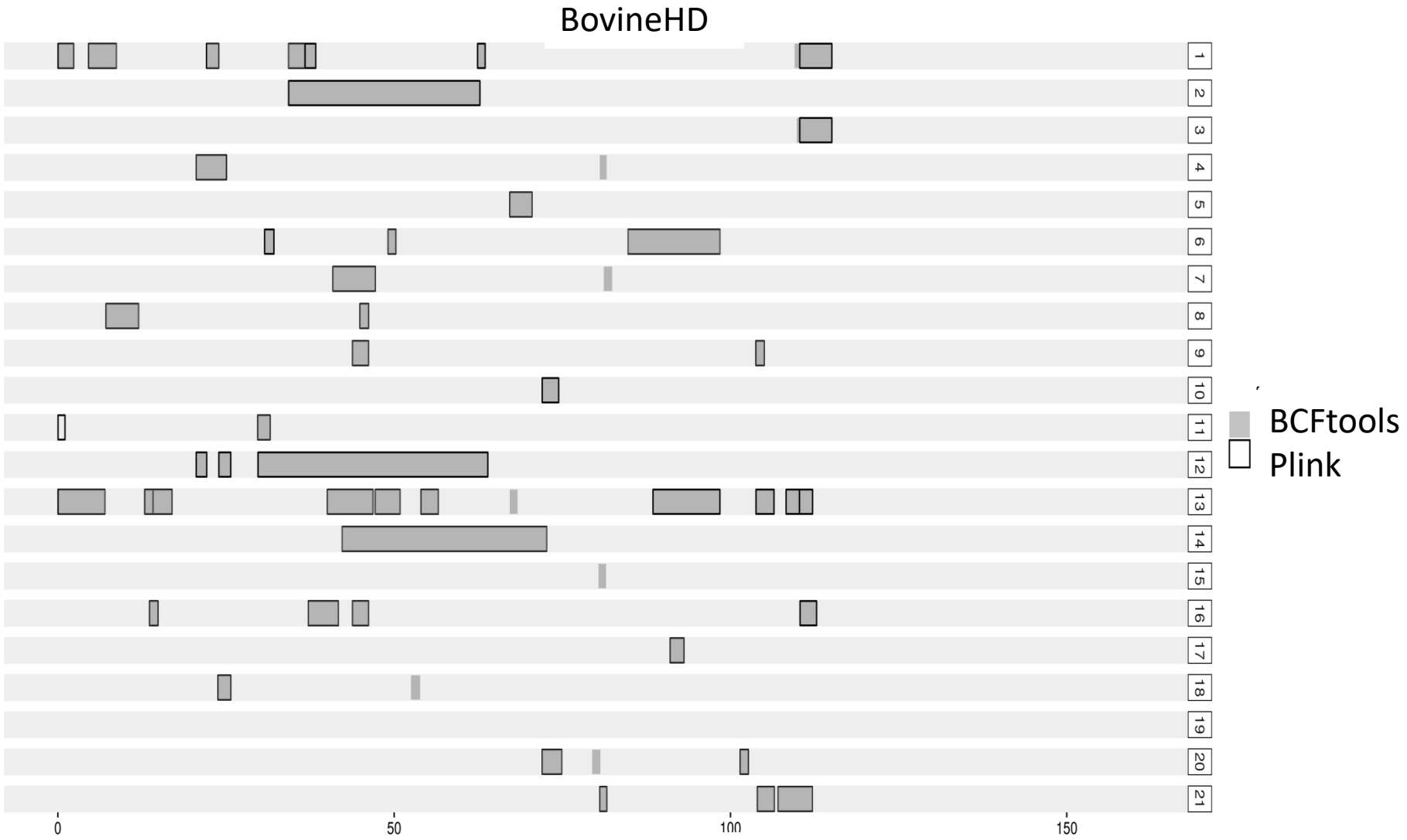
- **Emission probabilities** computed from:
  - Allele frequencies estimated from data
  - Phred-scaled genotype likelihoods (assumed 30 for BovineHD)
- Per bp recombination rate assumed constant and set at 1e-8 (1cM ~ 1Mb)
- State **transition probabilities** estimated from data via the Viterbi algorithm

- Allele frequencies, genotype probabilities and recombination rate are explicitly taken into account

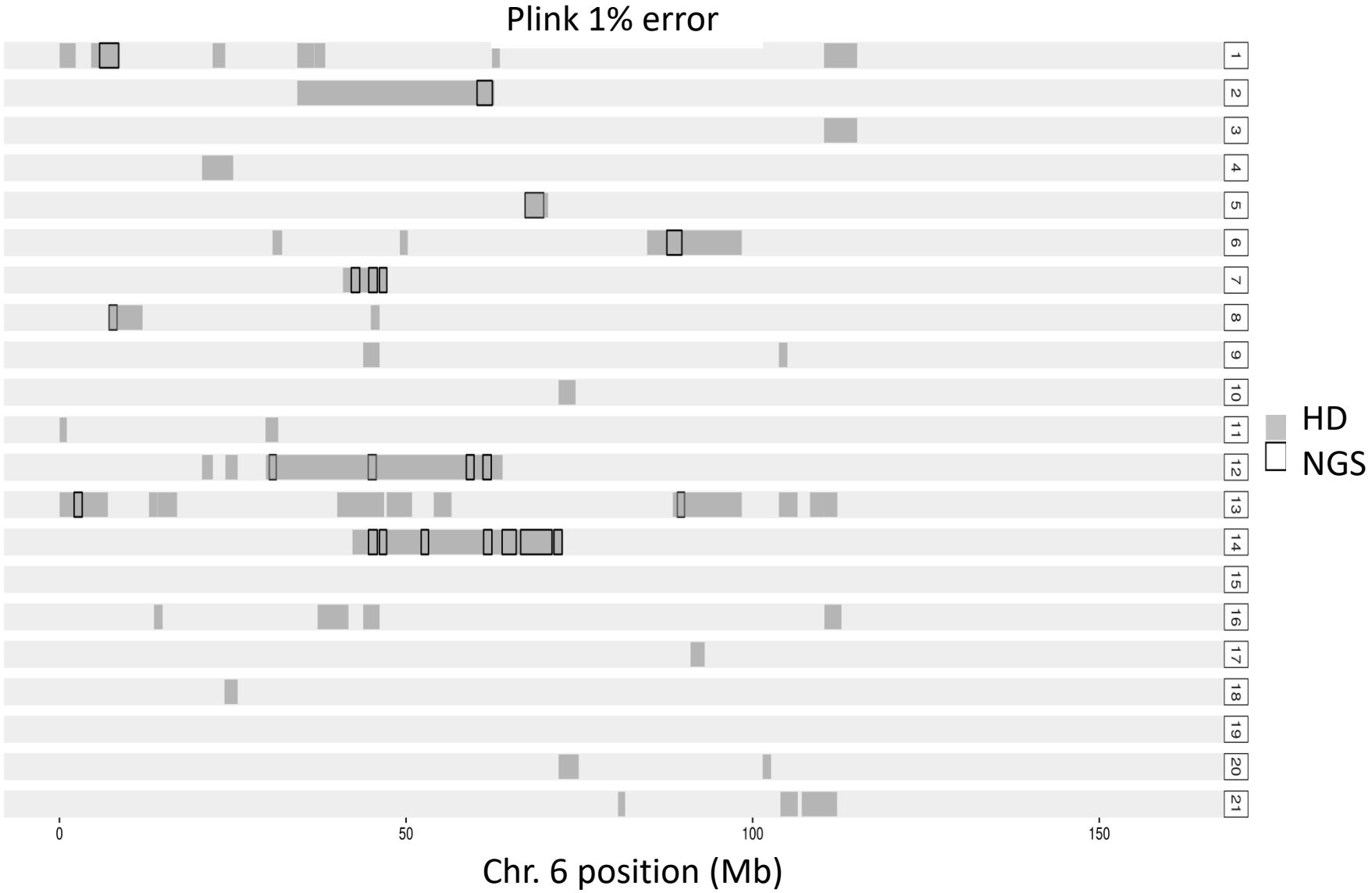
Results

- **Shown here for Chr 6**
  - HD: 34,800 SNPs
  - NGS: 1,274,485 SNPs

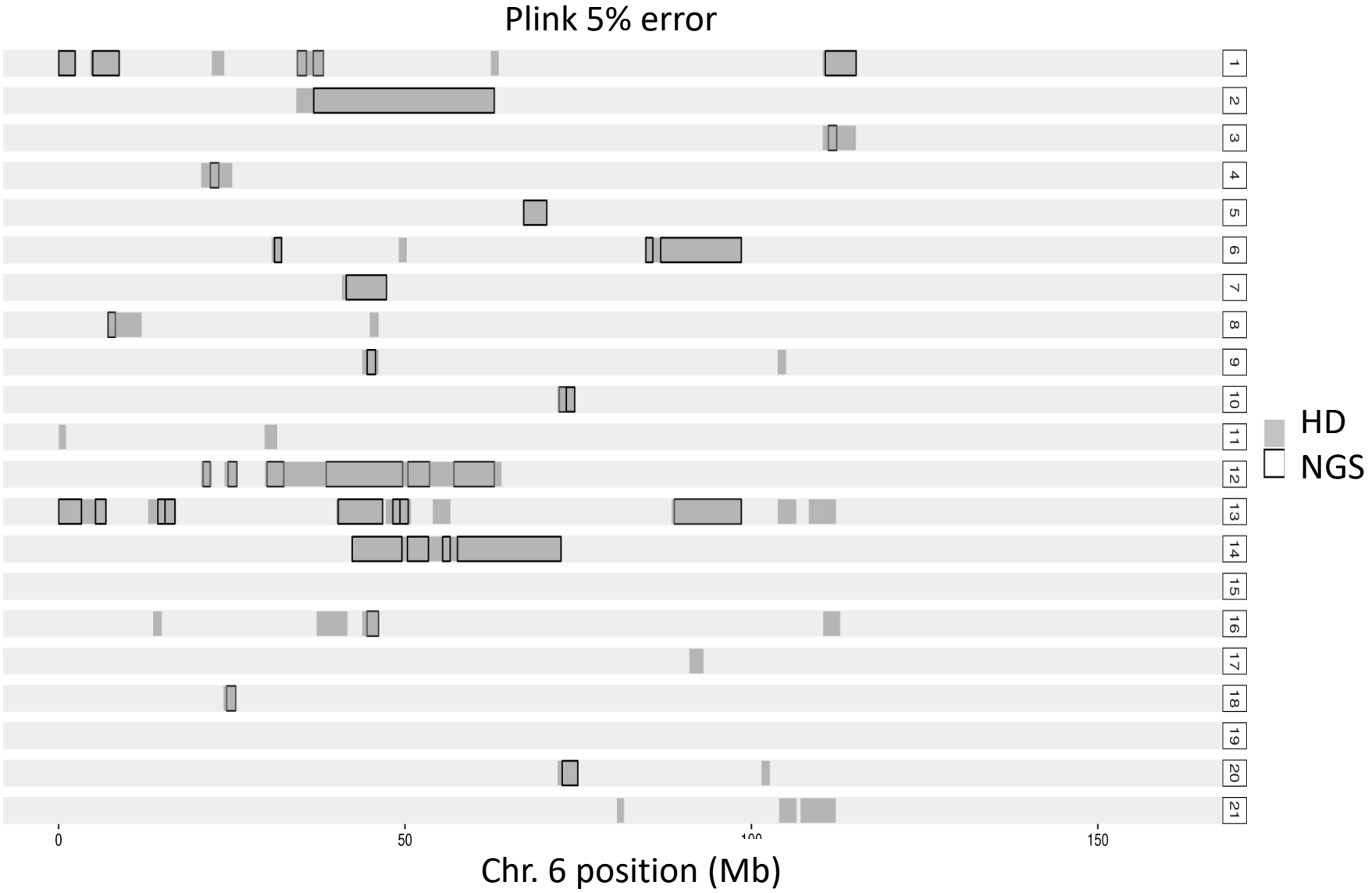
Results


BovineHD
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
0
50
100
150
Chr. 6 position (Mb)
BCFtools
Plink

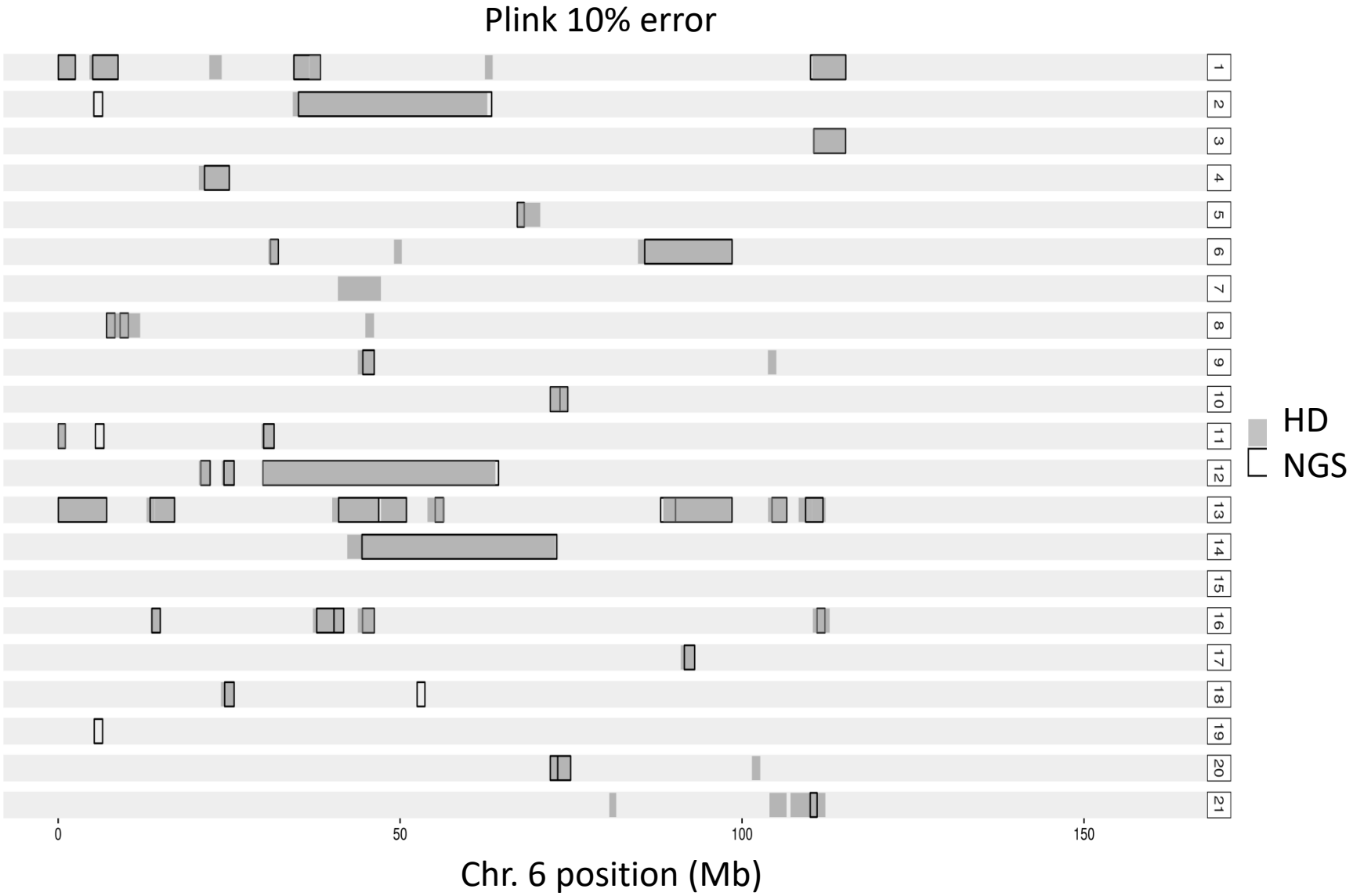
Results


Plink 1% error
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
HD
NGS
0
50
100
150
Chr. 6 position (Mb)

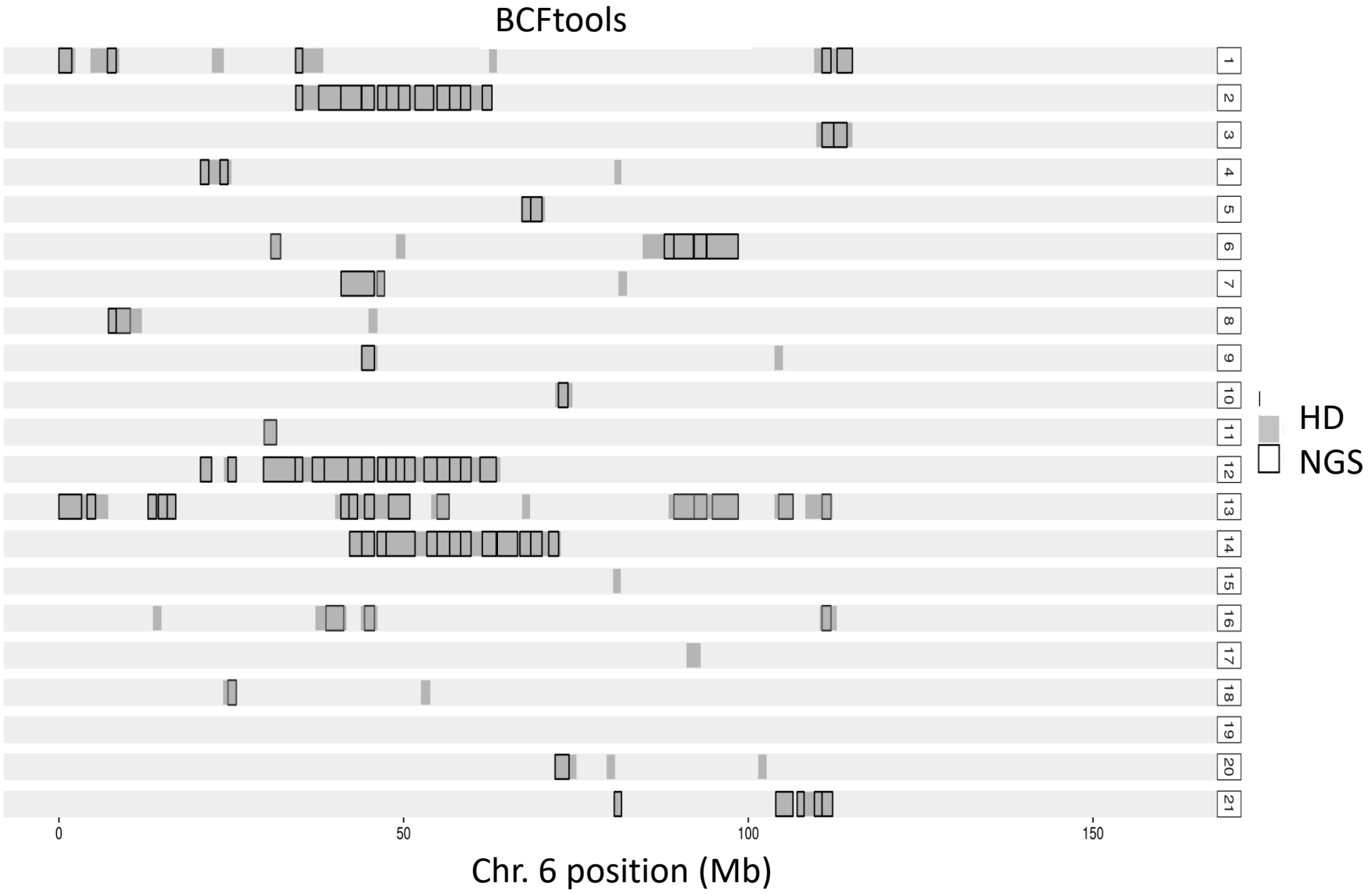
Results


Plink 5% error
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
HD
NGS
0
50
150
Chr. 6 position (Mb)

Results


Plink 10% error
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
HD
NGS
0
50
100
150
Chr. 6 position (Mb)

Results


BCFtools
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
HD
NGS
0
50
100
150
Chr. 6 position (Mb)

Comparison for HD data

- BCFtools and PLINK in high agreement
- **Very low genotyping error for this type of data**
- PLINK uses a heuristic approach, easy to understand; and is very fast
- BCFtools works equally well, based on a genetic model

Comparison for NGS data

- **Much higher genotyping error rate**
  - Allowing many heterozygous calls in a ROH (as PLINK does right now) is not optimal
  - Explicit modelling of genetic states (as BCFtools does) provides a more sensible framework
- For this data set, BCFtools cut long ROH segments into several smaller ROH
  - Artefacts due to assumed constant recombination rate?
  - Artefacts due to lenient variant calling?

Conclusions

- ROH calling from NGS data is still not mature
- Alternative algorithms are becoming available:
  - Bosse et al., 2014, PLoS Genetics
  - H3M2 (Magi et al., 2014, Bioinformatics)
  - BCFtools (Narasimhan et al., 2016, Bioinformatics)
  - ngsF-HMM (Vieira et al., 2016, Bioinformatics)
  - Druet and Gautier (this conference)

    please follow up regularly!
- More work is planned, comparing algorithms
  - Check effects of filtering for SNP calling
  - Use a proper genetic map (recombination)