Pig sequence imputation

Cross-validation with 168 sequenced pig genomes

Sept 2016, Birgit Zumbach and Aniek Bouwman

WAGENINGEN UR
For quality of life

BREED4FOOD

Topigs Norsvin
PROGRESS IN PIGS

Questions

- How accurate can we impute from SNP chips to WGS?
  - 80k
  - 600k
  - 80k to 600k in multiple steps to WGS
- Which animals in reference population?
  - All
  - Per breed/line
- Which imputation program to use?
  - Beagle, Fimpute, MiniMac, Impute2

Questions

- How accurate can we impute from SNP chips to WGS?
  - 80k
  - 600k
  - 80k to 600k in multiple steps to WGS
- Which animals in reference population?
  - All
  - Per breed/line -> 55 Large White
- Which imputation program to use?
  - Beagle, Fimpute, MiniMac3, Impute2

Variant calling

From BAM to variants

168 TopigsNorsvin Pigs

- Sequencing format usually FASTA
- BAM files created by pipeline ABGC WUR (Juanma)
- Samtools to get index files and fix read group
- Variants called using GATK and Freebayes
- Variants filtered using VCFtools

VCF format files (Beagle, MiniMac3)

Other formats supported by VCFtools: 012, IMPUTE, Plink

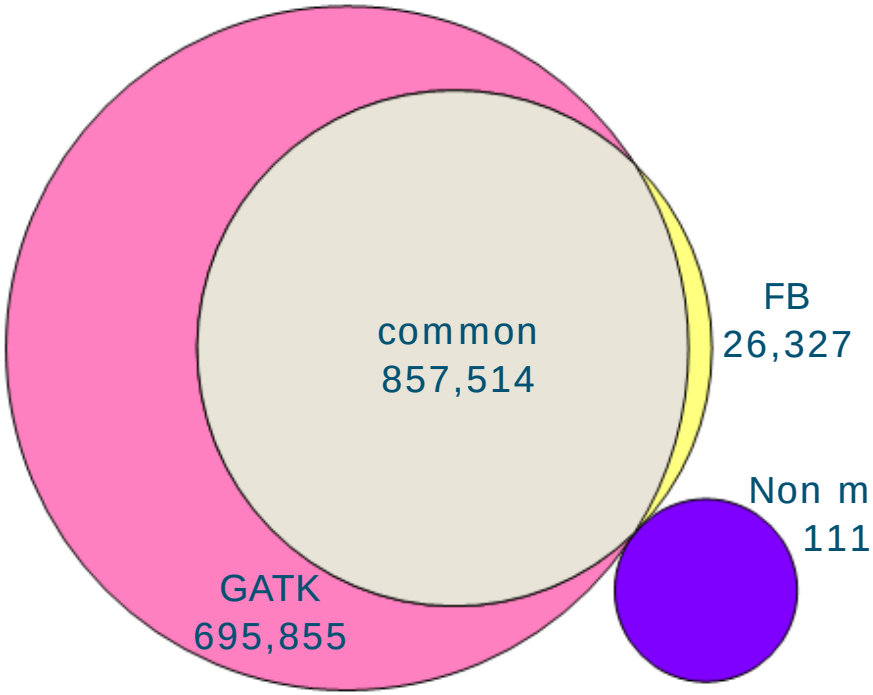
Filtering criteria

- At least 1 alt allele count
- Bi-allelic variants only
- Read depth (min: 4, max: 35 (mean+3*sd))
- Overall quality (PRED) 20
- Thinning 3bp
- Max missing 0.8

GATK: kept 1,664,543 out of 2,529,312 sites

Freebayes: kept 995,015 out of 1,877,418 sites

Variant discordance



All InDels
- Called in both
- Same position
- Different alleles

CHROM	POS1	POS2	IN_FILE	REF1	REF2	ALT1	ALT2
7	411	411	B	C	C	T	T
7	1491	.	1	C	.	T	.
7	1903	1903	O	C	CGGCGC	CGGCG	CGGCGGGCGC

Genotype discordance

GATK / FB	0 / 0	0 / 1	1 / 1	./.
0 / 0	**91,343,200**	832,416	54	946,949
0 / 1	145,169	**22,563,151**	54,186	292,821
1 / 1	26	145,613	**15,122,944**	179,850
./.	1,730,860	678,555	714,989	**9,371,713**

Imputation

Chip data

- WGS masked to represent chip
  - 600k Axiom
- No allele coding issue, nor format differences

From real chip data:

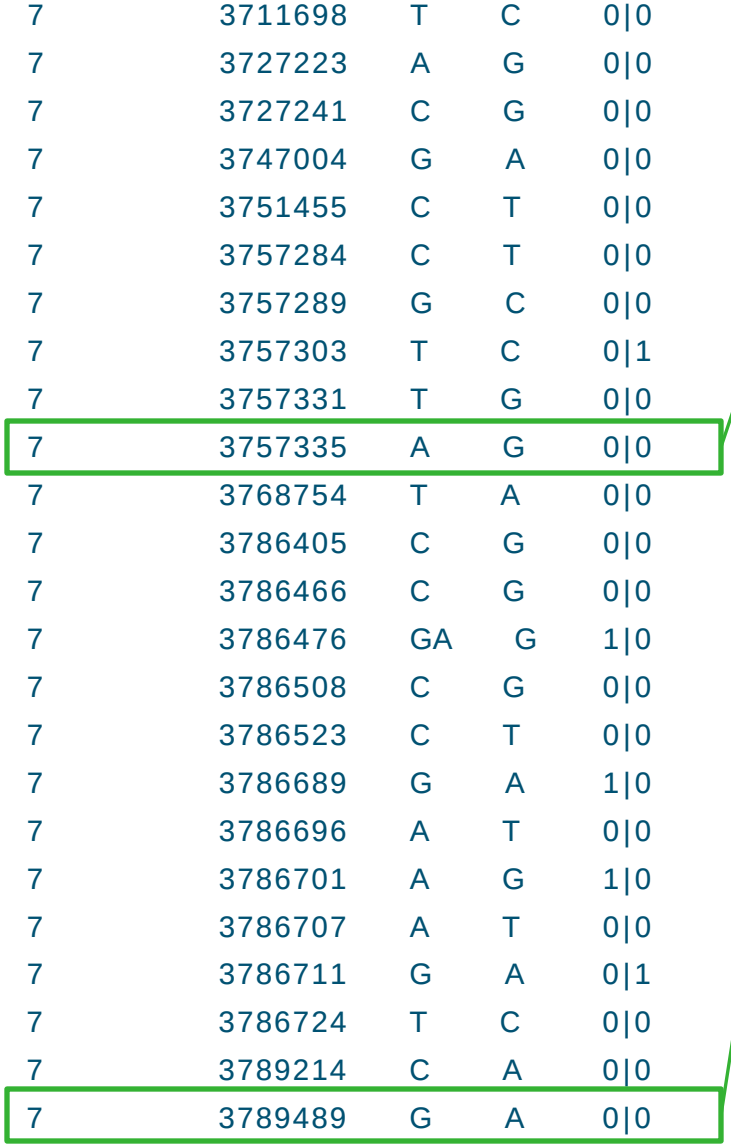
- **allele coding chip to ref / alt**
- **format readable by imputation program**

WGS

#CHROM	POS	REF	ALT	162405
7	3711698	T	C	0|0
7	3727223	A	G	0|0
7	3727241	C	G	0|0
7	3747004	G	A	0|0
7	3751455	C	T	0|0
7	3757284	C	T	0|0
7	3757289	G	C	0|0
7	3757303	T	C	0|1
7	3757331	T	G	0|0
7	3757335	A	G	0|0
7	3768754	T	A	0|0
7	3786405	C	G	0|0
7	3786466	C	G	0|0
7	3786476	GA	G	1|0
7	3786508	C	G	0|0
7	3786523	C	T	0|0
7	3786689	G	A	1|0
7	3786696	A	T	0|0
7	3786701	A	G	1|0
7	3786707	A	T	0|0
7	3786711	G	A	0|1
7	3786724	T	C	0|0
7	3789214	C	A	0|0
7	3789489	G	A	0|0

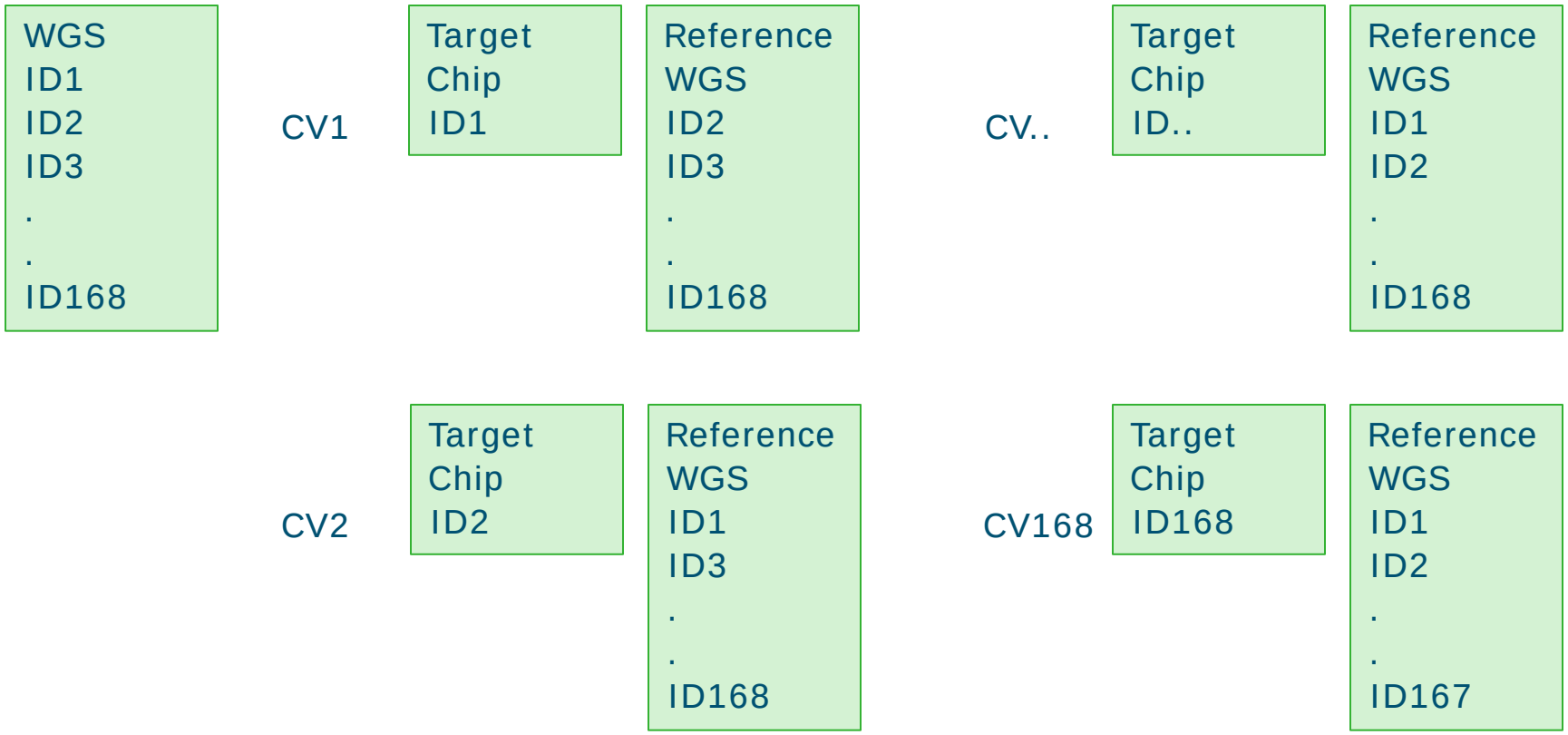
Chip

#CHROM	POS	REF	ALT	162405
7	3757335	A	G	0|0
7	3789489	G	A	0|0



WAGENINGEN UR
For quality of life

Leave-one-out cross-validation

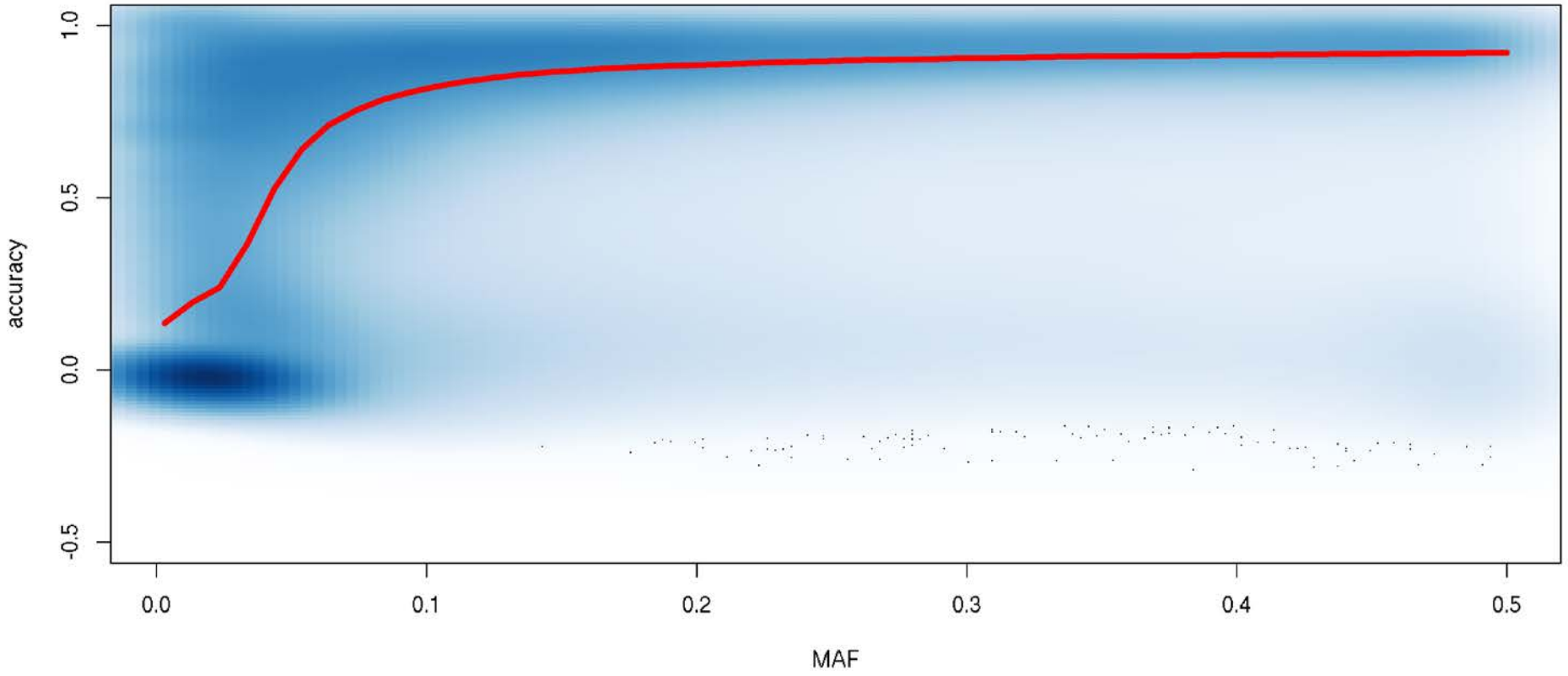


Imputation accuracy (GATK; SSC7)

Imputation program	600k2WGS GATK	600k2WGS FB
Beagle4.1 gt	0.754	0.827
FImpute gt	0.611	-
MiniMac3 gt	0.700	0.810
Beagle4.1 ds	0.571	0.769
MiniMac3 ds	0.443	0.758

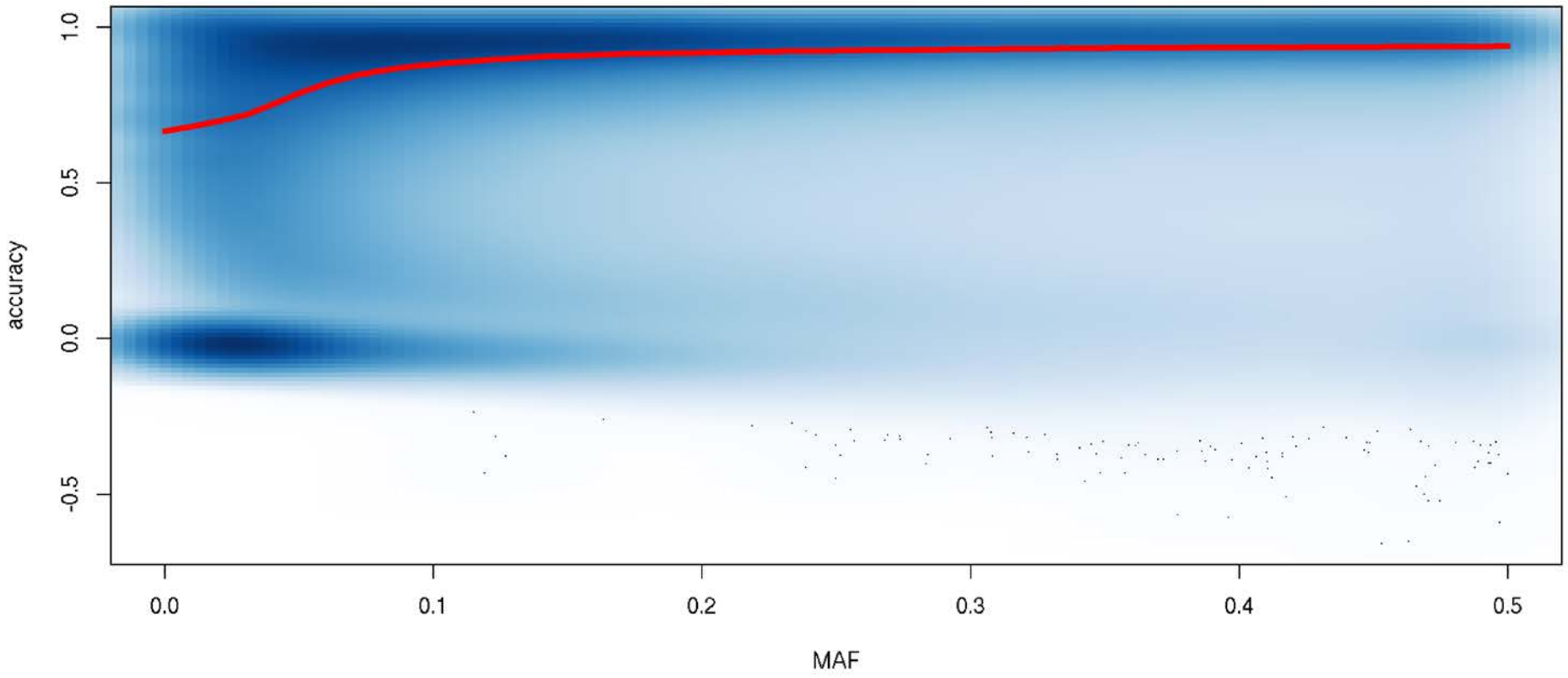
Imputation accuracy vs MAF (GATK; SSC7)

FImpute



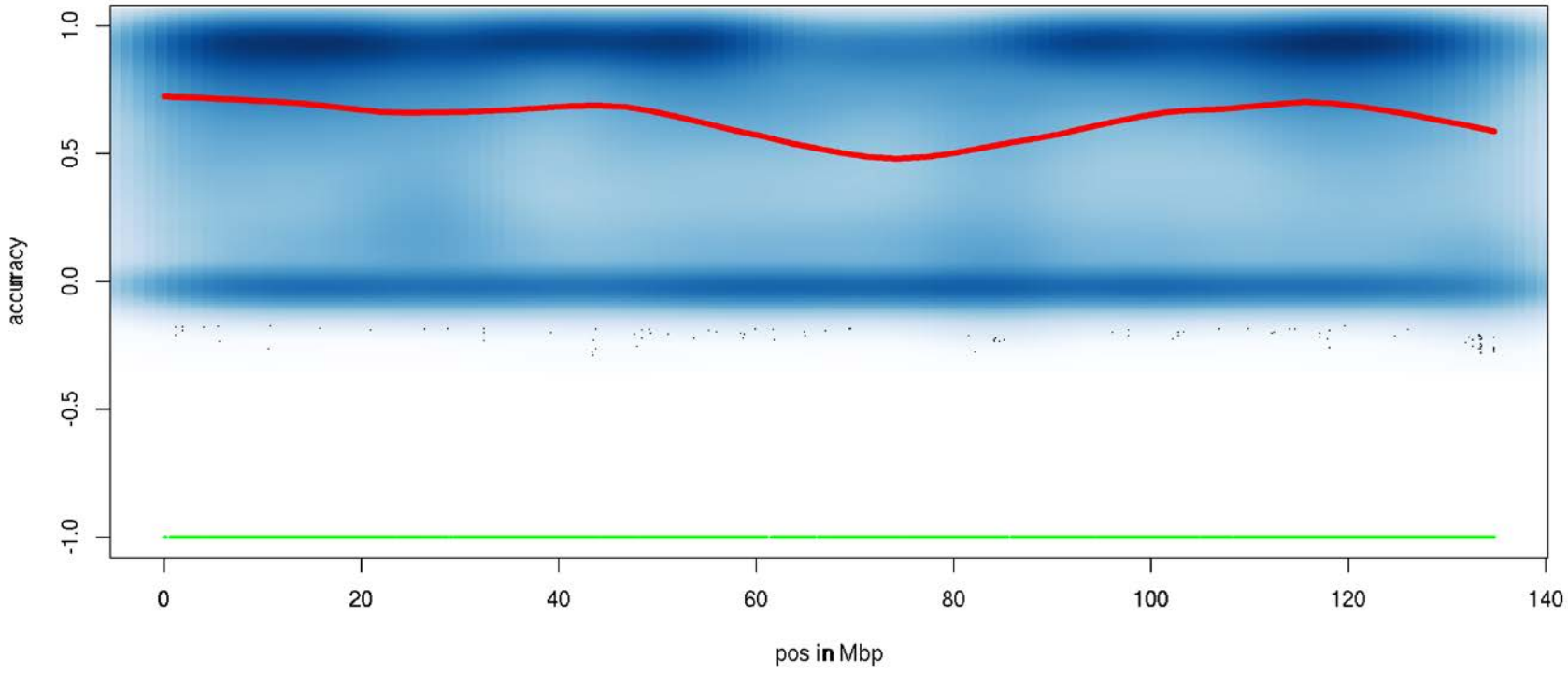
Imputation accuracy vs MAF (GATK; SSC7)

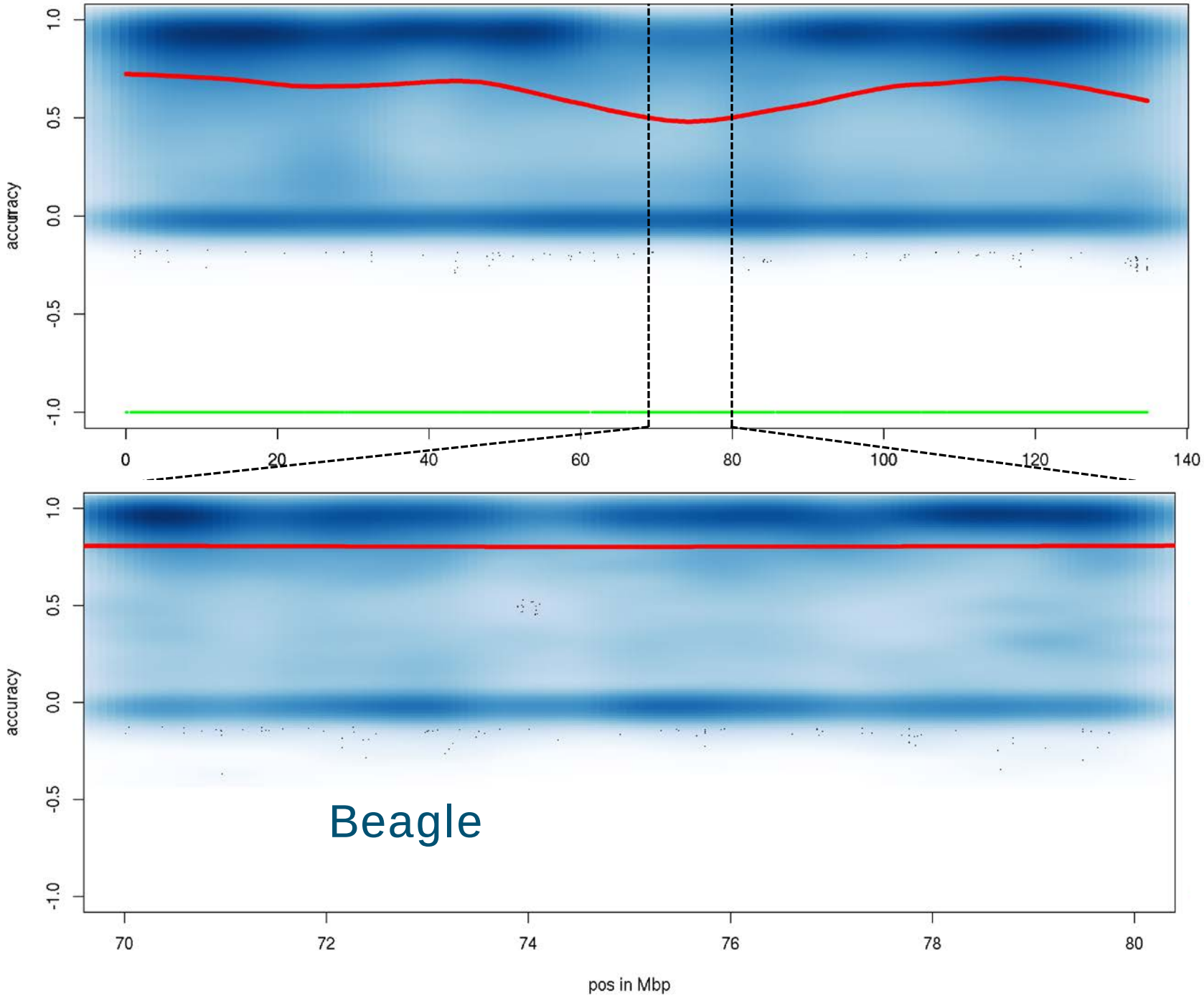
Beagle



Imputation accuracy vs position (GATK; SSC7)

FImpute




accuracy
1.0
0.5
0.0
-0.5
-1.0
0
20
40
60
80
100
120
140
accuracy
1.0
0.5
0.0
-0.5
-1.0
70
72
74
76
78
80
pos in Mbp
Beagle

Within breed imputation

55 Large White animals

#variants GATK: 1,332,695, #variants Freebayes: 938,298

Reference population	Imputation program	GATK 600k2WGS	FB 600k2WGS
55LW	Beagle4.1 gt	0.765	0.795
55LW	Beagle4.1 ds	0.510	0.619
168MB	Beagle4.1 gt		
168MB	Beagle4.1 ds		

Keep in mind

- Cross-validation done among founders
- No pedigree used
- Only ~12 founders per line
- Sequence data far from ‘clean’ genotypes
- Build issues (on SSC7) ?