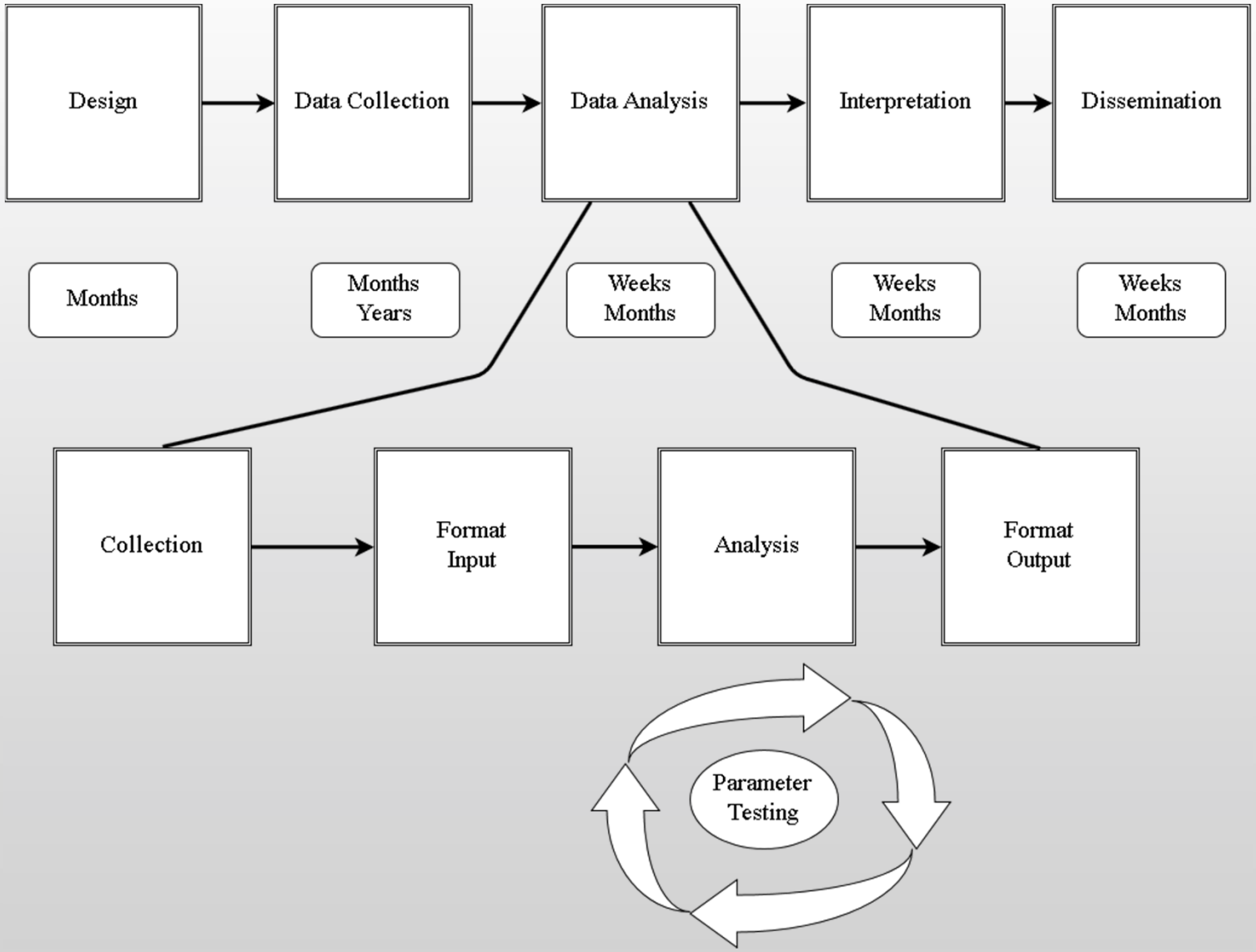
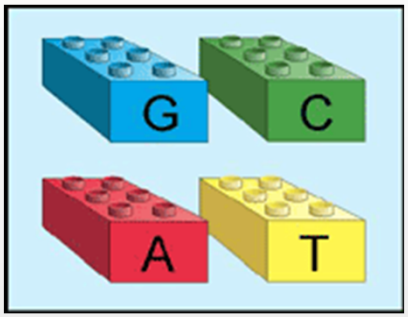
Bioinformatics At Scale

CI Day
October 04, 2017

Project Life Cycle

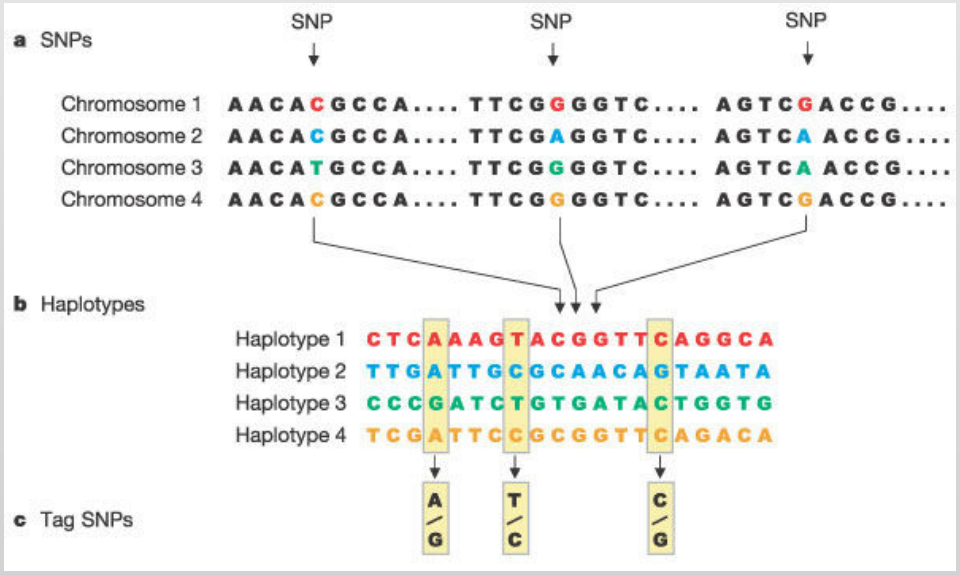


Genome Sequence & Markers



- Mammalian genome ~ 3,000,000,000 bases
- Humans 23 chromosome pairs
- Cows 30 chromosome pairs

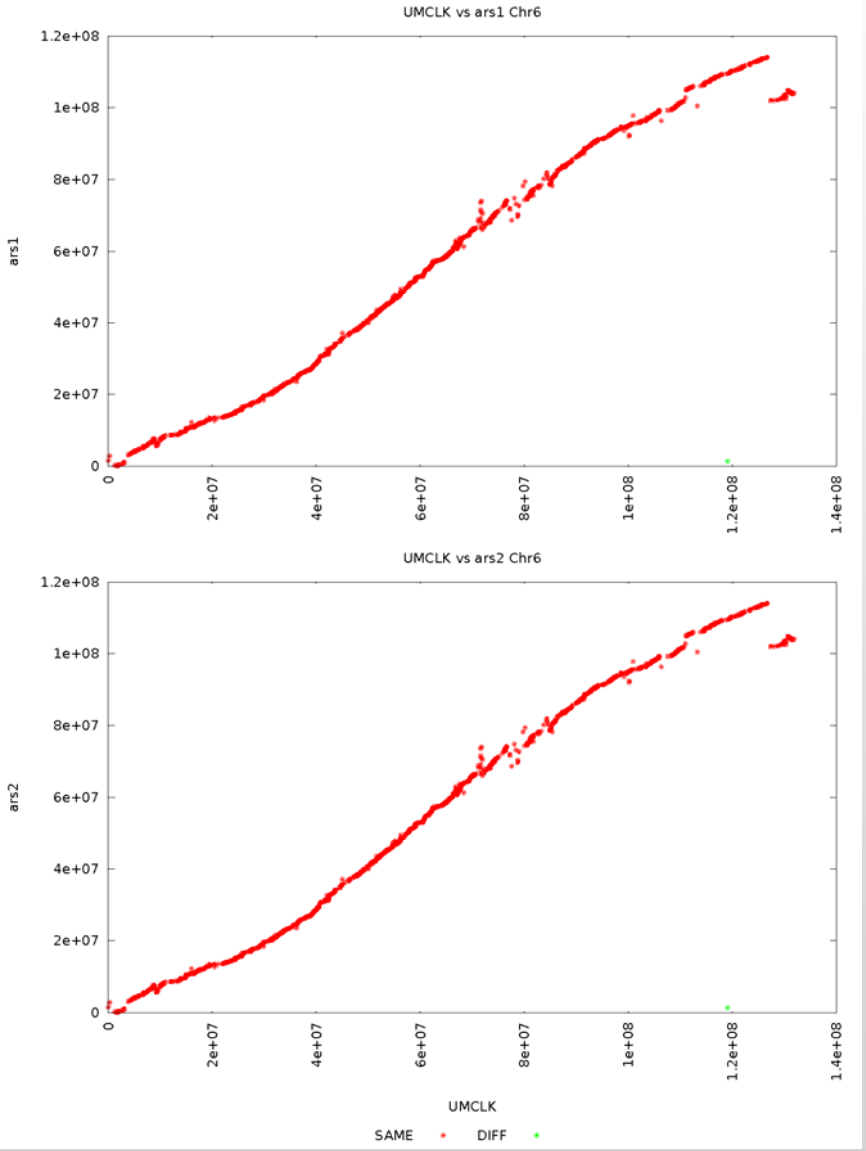
ARS-BFGL-BAC-10365 Chr14:26,821,443
GGCTTGTTGGCAGACTCACAGCACTAATTTCTCATCTTCTCCCAGCCAATGTTTATTCTC
[A/C]
TTCTGCTTAACTCTTCCTCAGGAGTGTAATATAAACGGGATAATACACTTGAAGTTTTTG

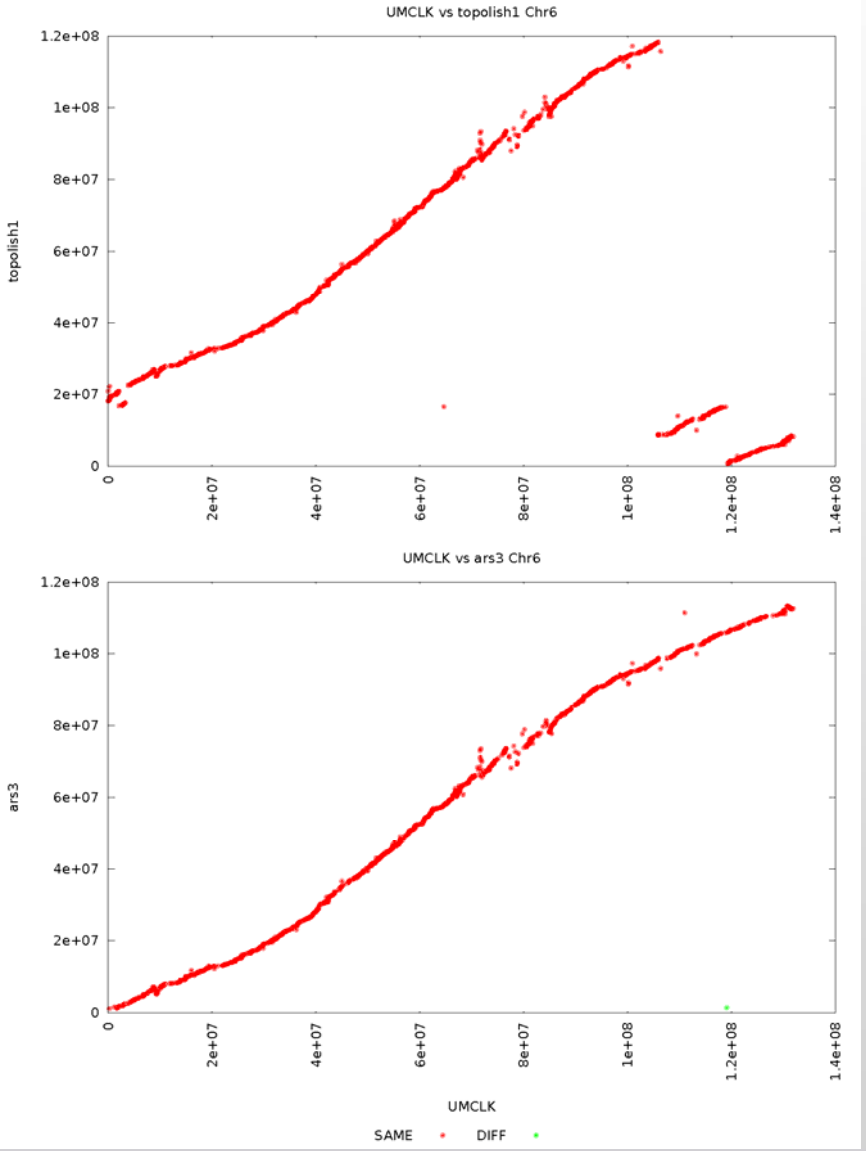


Map loci between genomes

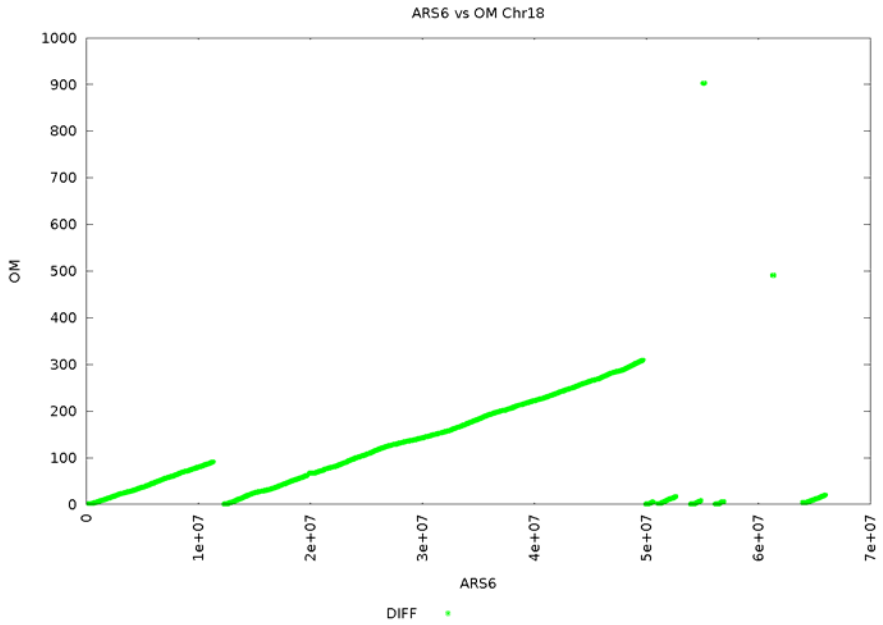
- Coordinate on genome A
- Extract flanking sequence from A
- BLAT left & right sequence from A on B
  - BLAT operations = 2X number of loci (66M)
- Check left = right
- Coordinate on A vs coordinate on B

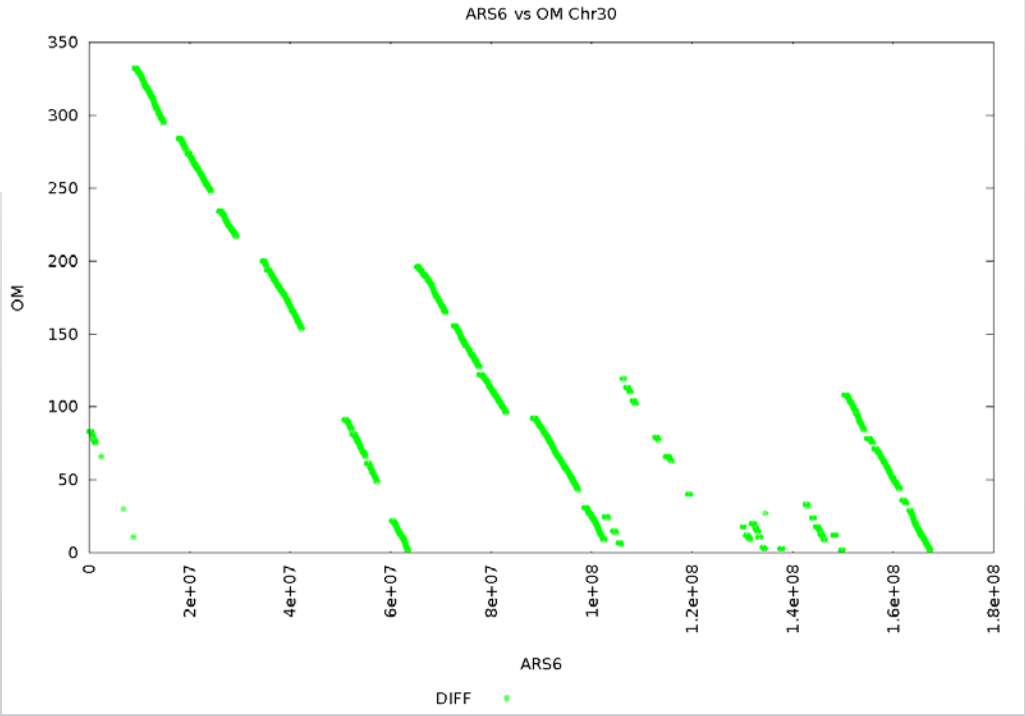
Fix reference genome





Order optical map contigs





Phasing/Imputation

Input data:
1) List 1
2) List 2
3) List 3

Phasing Software:
1) Eagle (S/C)
2) ShapeIt3 (S/C)

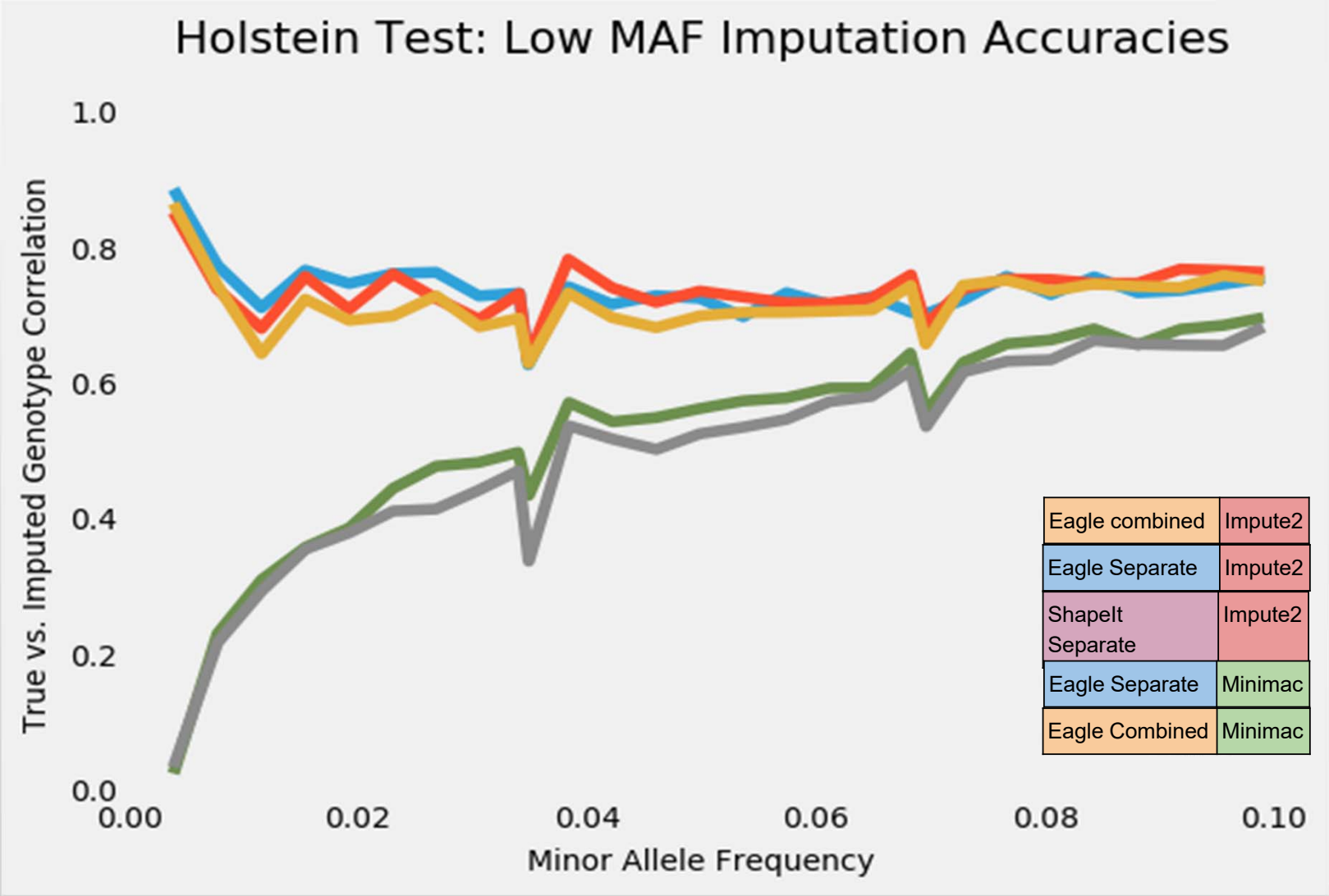
Imputation Software:
1) Impute2
2) Impute4
3) Minimac

Troy Rowan

	Phasing Method	Imputation Method	Animal List	snp50_Accuracy
Run 1	Eagle combined	Impute2	List 1	0.9697877503
Run 2	Eagle Seperate	Impute2	List 1	0.9696363368
Run 4	ShapeIt Separate	Impute2	List 1	0.9688384399
Run 5	Eagle Seperate	Minimac	List 1	0.9519533042
Run 6	Eagle Combined	Impute2	List 2	0.9723829201
Run 7	Eagle Seperate	Impute2	List 2	0.9701853428
Run 8	Eagle Combined	Minimac	List1	0.9508772631
Run 9	ShapeIt Separate	Impute2	List 2	0.9673005011
Run 10	Eagle Separate	Minimac	List 2	0.9540561055
Run 11	Eagle Combined	Minimac	List 2	0.954044613
Run 12	Eagle Combined	Impute2	List 3	0.9734224999
Run 13	Eagle Seperate	Impute2	List 3	0.9731905528
Run 14	Shape It Separate	Impute2	List 3	0.9673530407
Run 15	Eagle Separate	Minimac	List 3	0.9578527182
Run 16	Eagle Combined	Minimac	List 3	0.9584204624
Run 17	ShapeIt3 Separate	Impute4	List1	0.9682929529
Run 18	ShapeIt3 Separate	Impute4	List2	0.9682162699
Run 19	ShapeIt3 Separate	Impute4	List3	0.9682602866
Run 30	Eagle combined	Impute4	List 1	0.9690416594

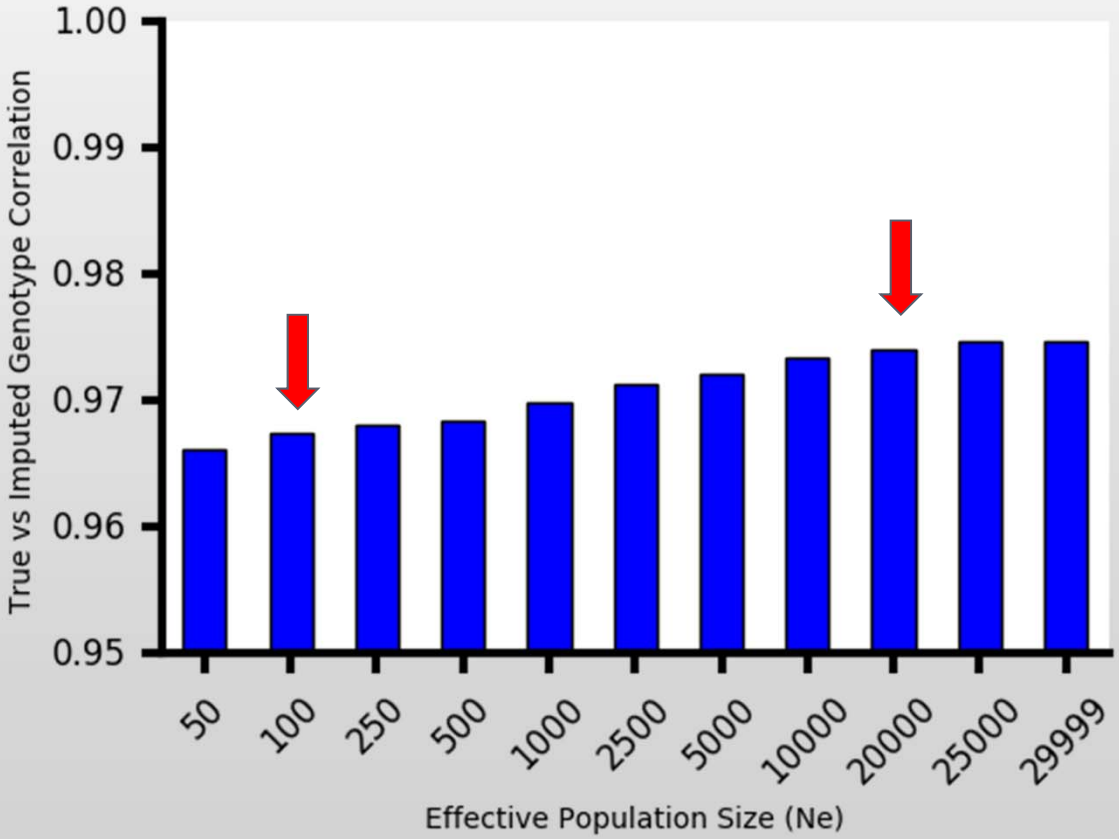
...Run128

Phasing/Imputation

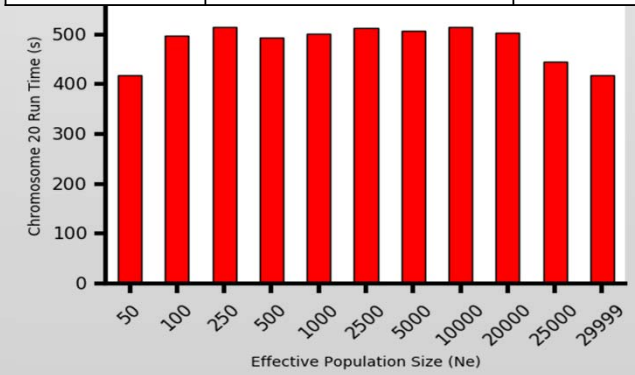


Phasing/Imputation

Effective population size: Run time and imputation accuracy



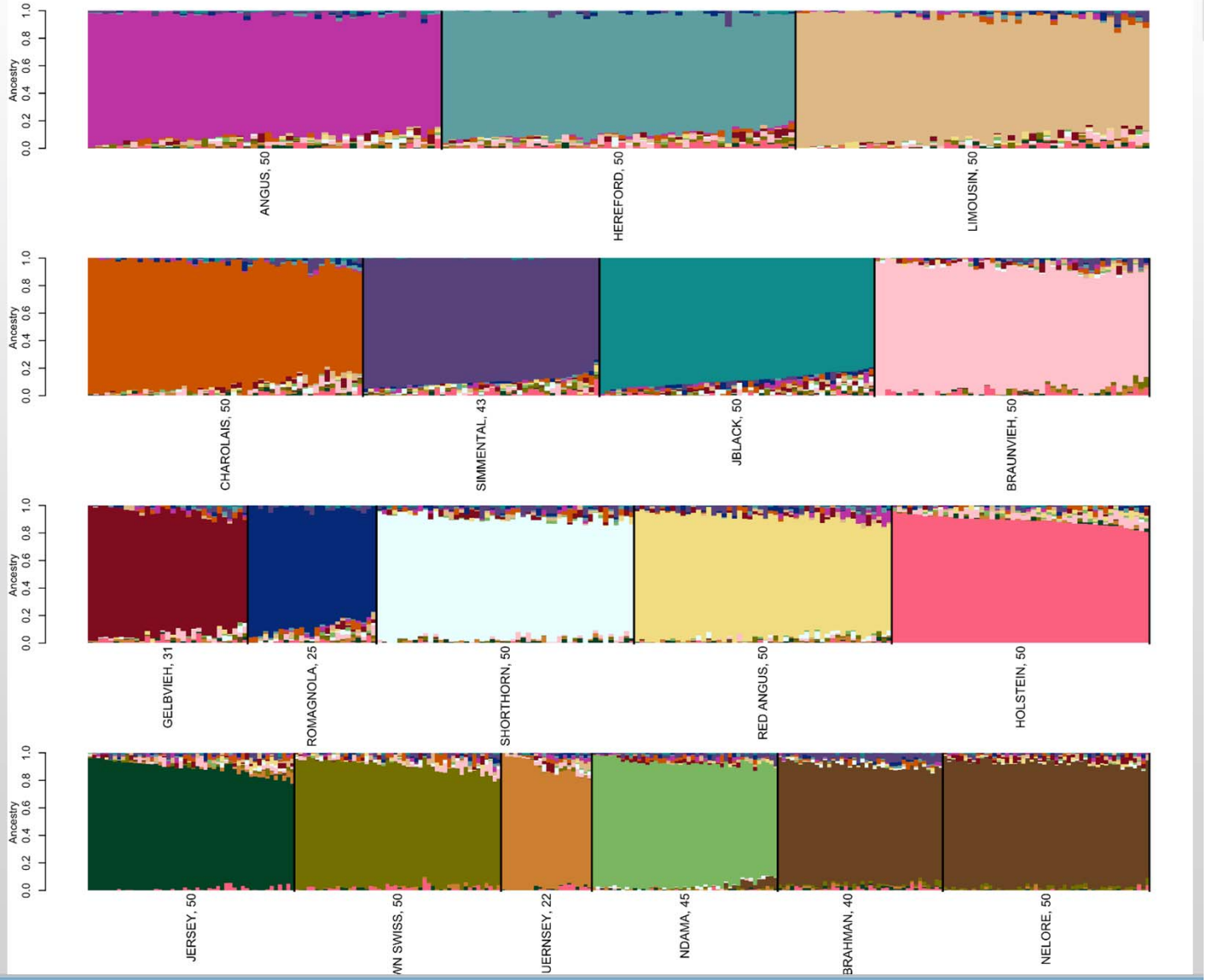
Run	Impute Option Change	SNP50 Accuracy
run32	Ne = 50	0.9661228048
run33	Ne = 100	0.9674798069
run34	Ne = 250	0.9680432685
run35	Ne = 500	0.9684722483
run36	Ne = 1000	0.969832221
run37	Ne = 2500	0.9713246254
run38	Ne = 5000	0.9720878882
run39	Ne = 10,000	0.9734445241
run40	Ne = 20,000	0.974057906
run41	Ne = 25,000	0.9746797499
run42	Ne = 29,999	0.9747018007



Animal Genomics @ MIZZOU

Ancestry

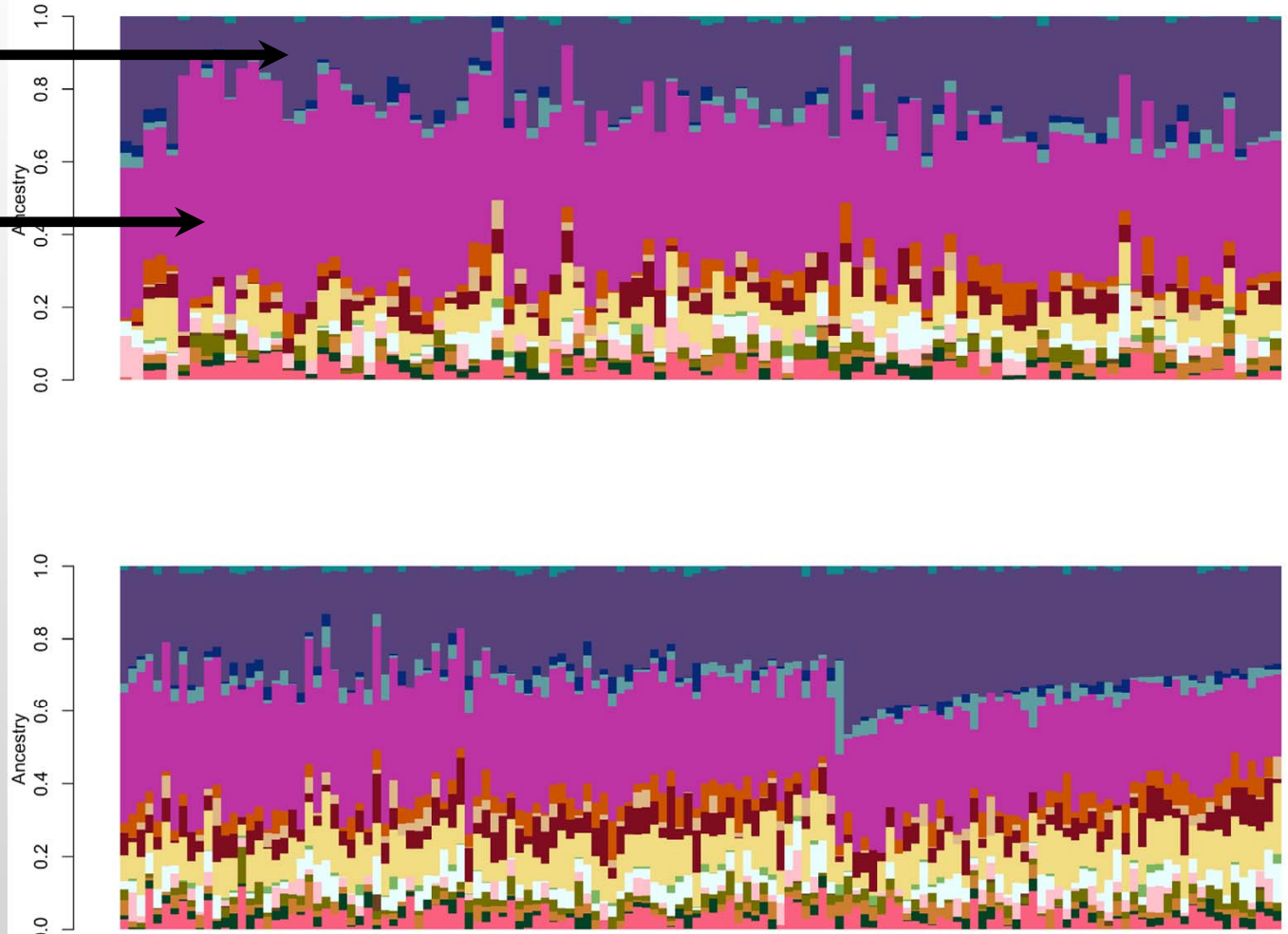
Tamar Crum



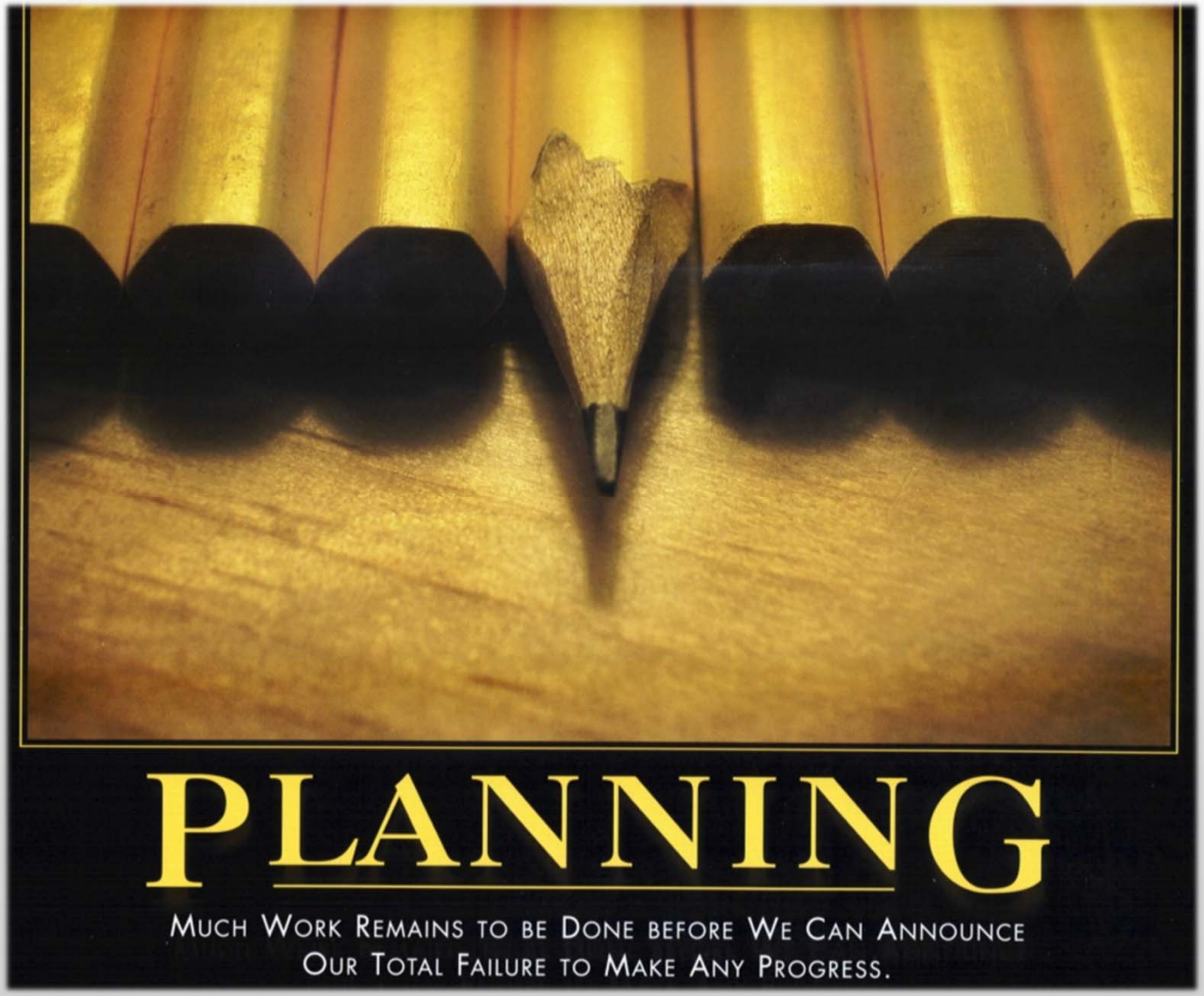
Ancestry

Simmental

Angus



Resources + Vision = Competitive Advantage


PLANNING
Much Work Remains to be Done before We Can Announce
Our Total Failure to Make Any Progress.