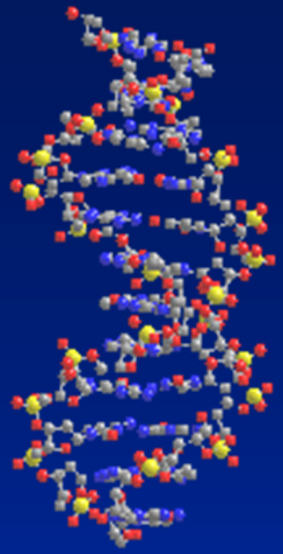
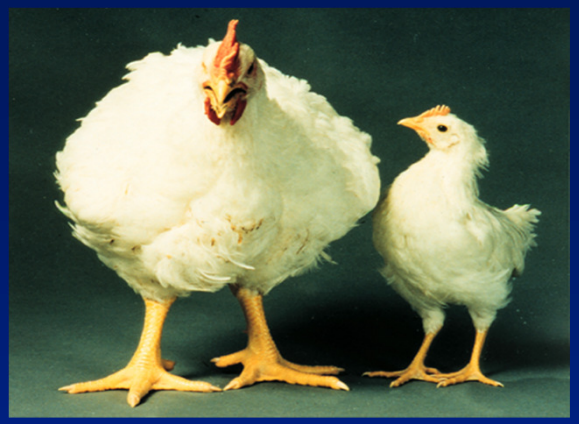
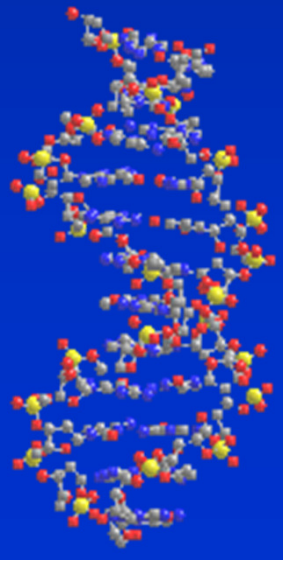




Genomic Selection in the era of Genome sequencing



Course overview

- Day 1
  - Linkage disequilibrium in animal and plant genomes
- Day 2
  - Genome wide association studies
- Day 3
  - Genomic selection
- Day 4
  - Genomic selection
- Day 5
  - Imputation and whole genome sequencing for genomic selection

Imputation

- Why impute?
- Approaches for imputation
- Factors affecting accuracy of imputation
- How can imputation give you more power?

Why impute?

- Fill in missing genotypes from the lab
- Merge data sets with genotypes on different arrays
  - Eg. Druet et al. 2010, merged two data sets in dairy cattle on alternate arrays
- Impute from low density to high density
  - 7K-> 50K (save $$$)
  - 50K->800K
  - capture power of higher density?
  - Better persistence of accuracy
- Sequence expensive, can we impute to full sequence data?

Core concept

- Identity by state (IBS)
  - A pair of individuals have the same allele at a locus
- Identity by descent (IBD)
  - A pair of individuals have the same alleles at a locus and it traces to a common ancestor
- Imputation methods determine whether a chromosome segment is IBD

Core concept 2

- Any individuals in a population may share a proportion of their genome identical by descent (IBD)
  - IBD segments are the same and have originated in a common ancestor
- The closer the relationship the longer the IBD segments
  - Pedigree relationships

Several methods for imputation

- Two main categories:
  - Family based
  - Population based
  - Or combination of the two

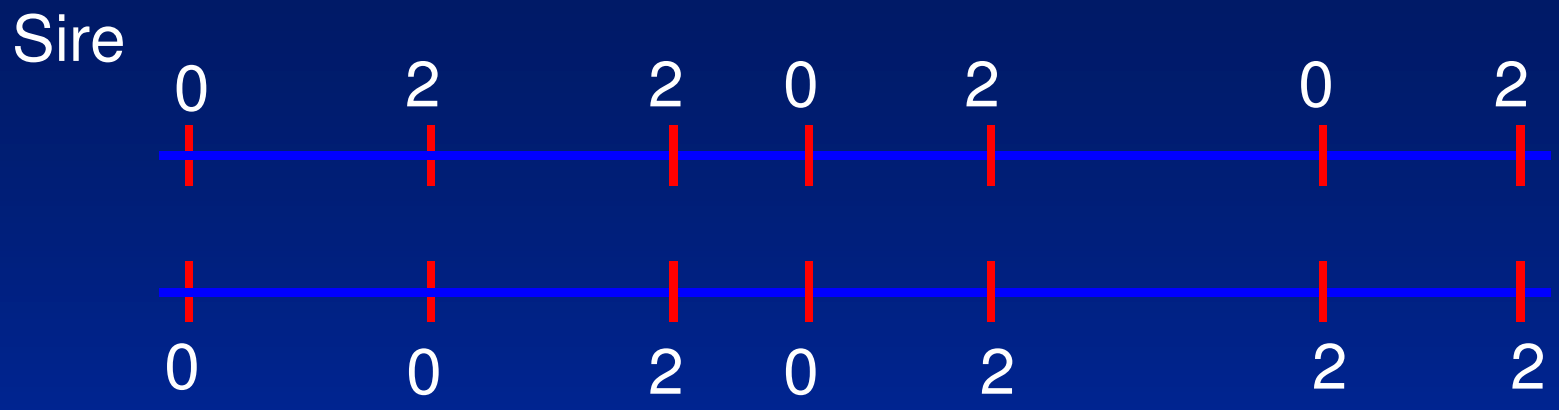
  - Some of the most effective are Beagle (Browning and Browning, 2009), MACH (Li et al., 2010), Impute2 (Howie et al., 2009), AlphaPhase (Hickey et al 2011)

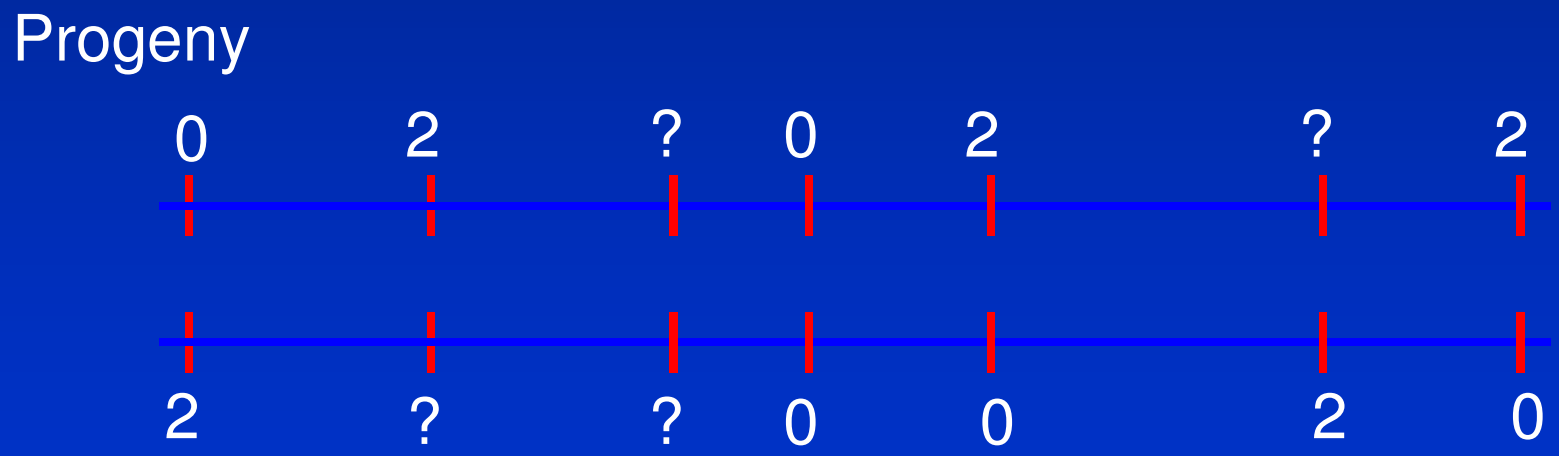
Several methods for imputation

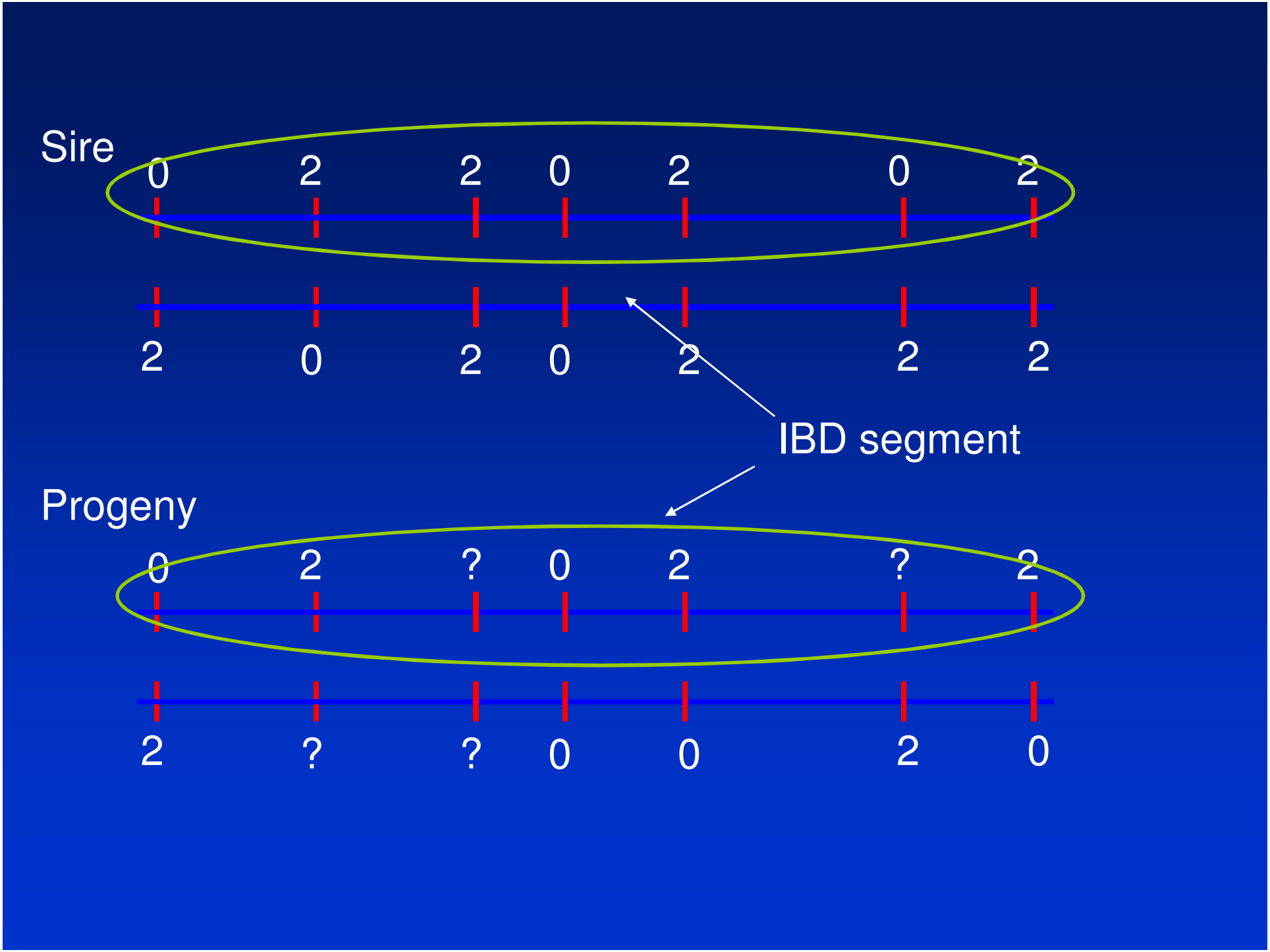
- Two main categories:
  - Family based
  - Population based
  - Or combination of the two

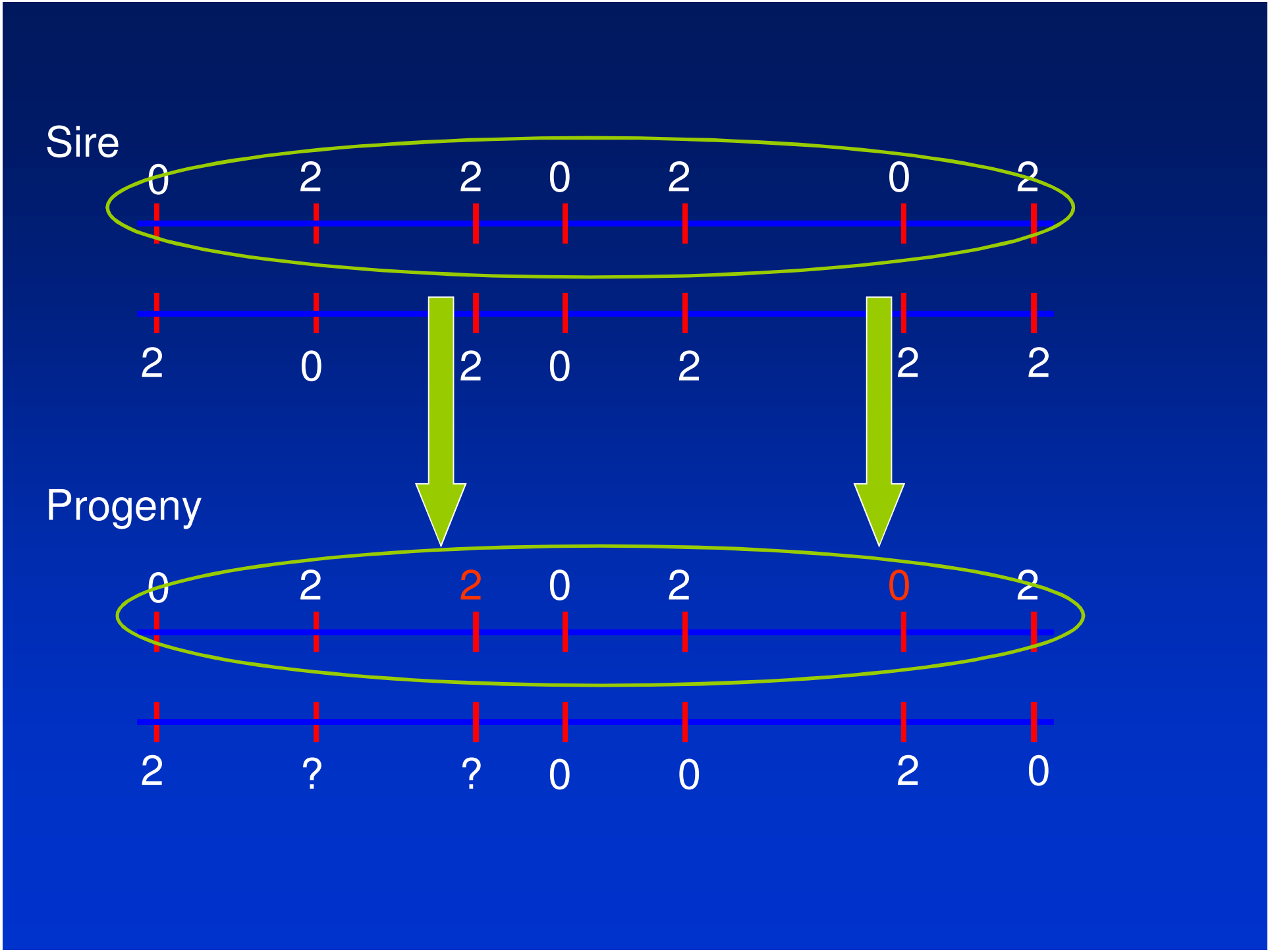
  - Some of the most effective are Beagle (Browning and Browning, 2009), MACH (Li et al., 2010), Impute2 (Howie et al., 2009), AlphaPhase (Hickey et al 2011)

Finding an IBD segment

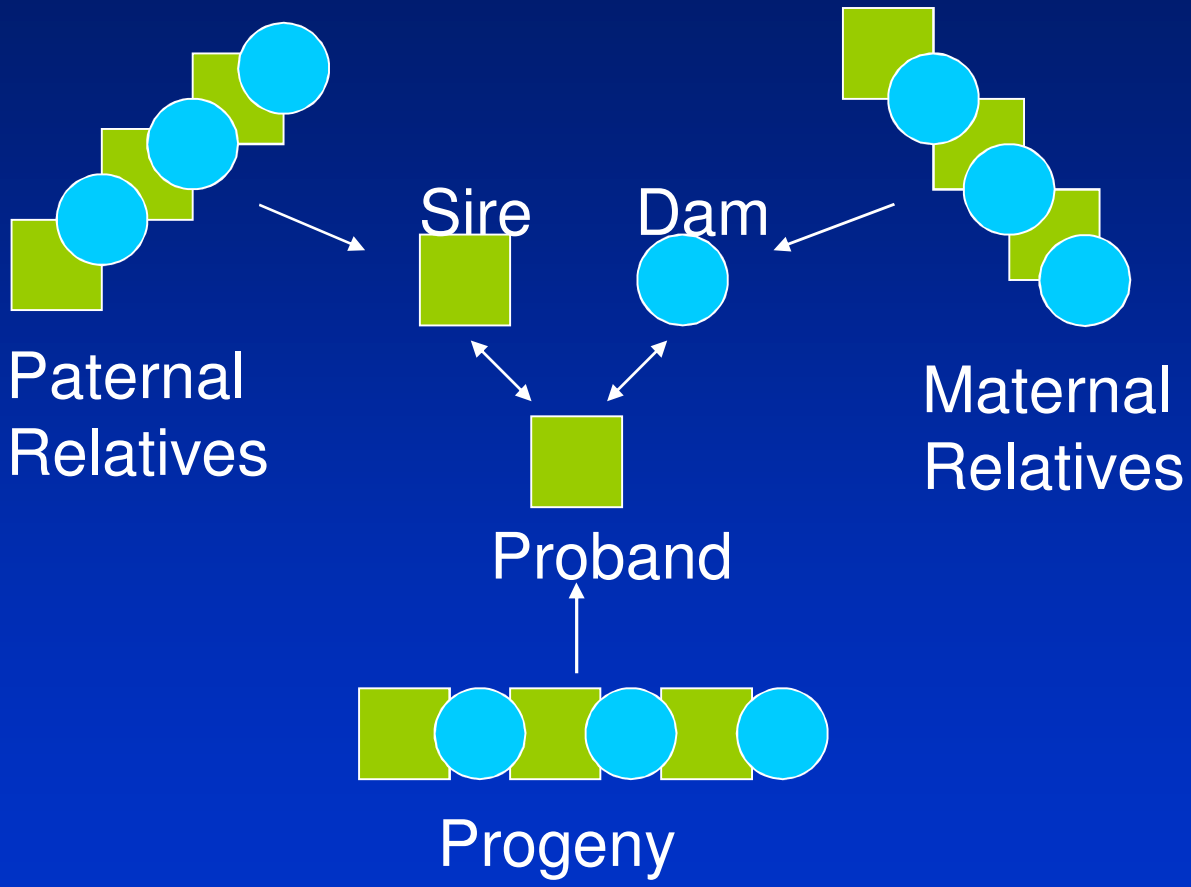




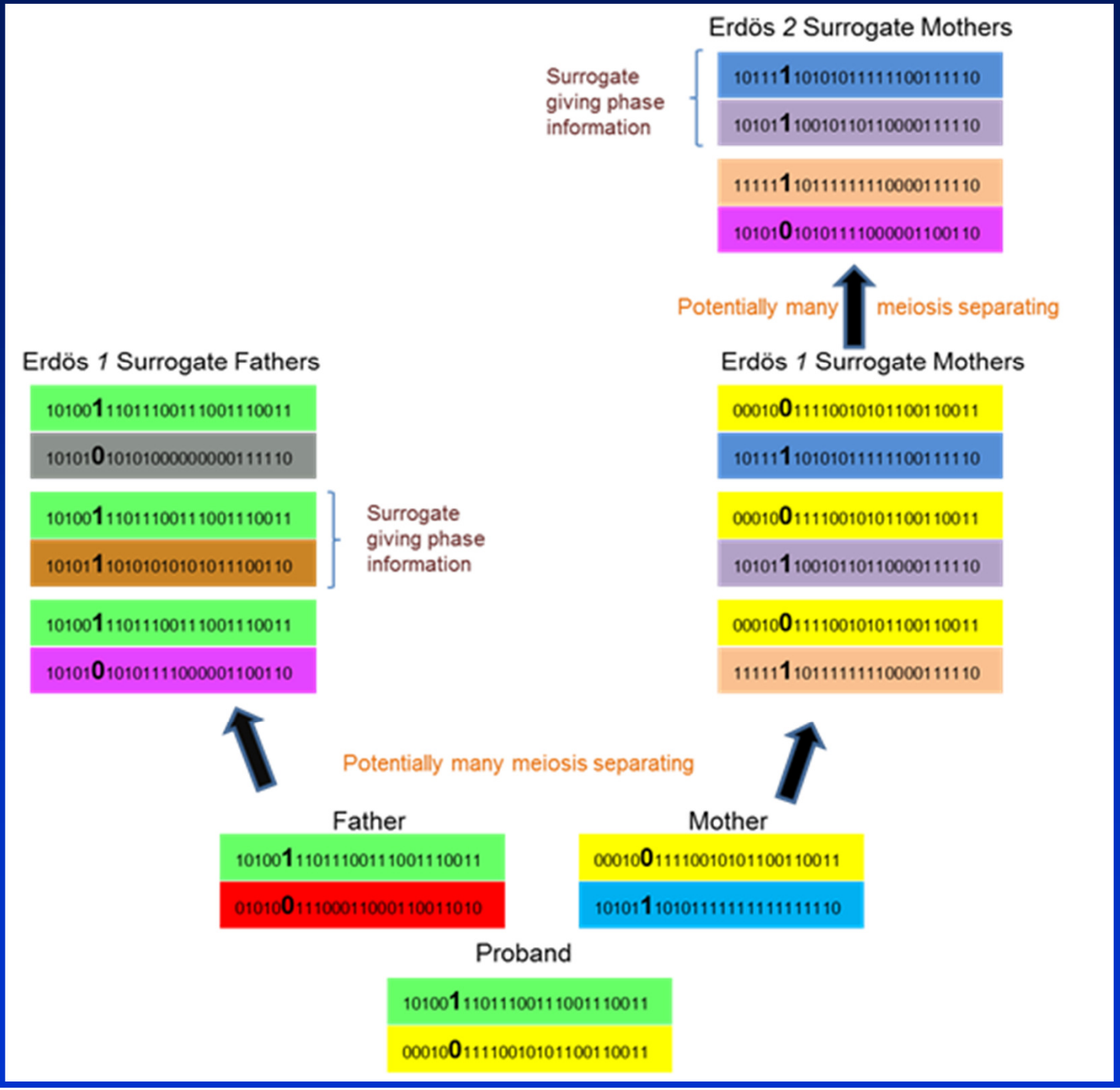

Sire
0 2 2 0 2 0 2
2 0 2 0 2 2 2
IBD segment
Progeny
0 2 ? 0 2 ? 2
2 ? ? 0 0 2 0


Sire
0 2 2 0 2 0 2
2 0 2 0 2 2 2
Progeny
0 2 2 0 2 0 2
2 ? ? 0 0 2 0

Relationships



Long range phasing


Erdös 2 Surrogate Mothers
Surrogate giving phase information
Potentially many meiosis separating
Erdös 1 Surrogate Fathers
Erdös 1 Surrogate Mothers
Surrogate giving phase information
Potentially many meiosis separating
Father
Mother
Proband

Several methods for imputation

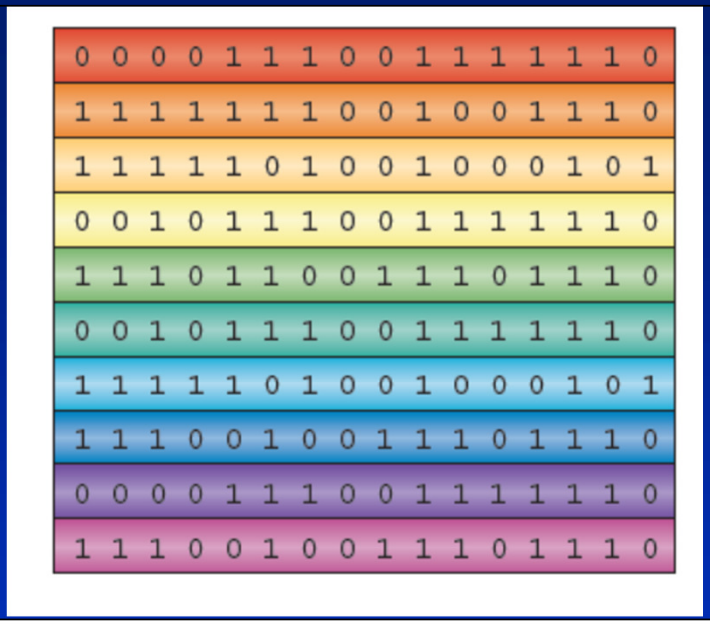
- Two main categories:
  - Family based
  - Population based (exploits LD)
  - Or combination of the two

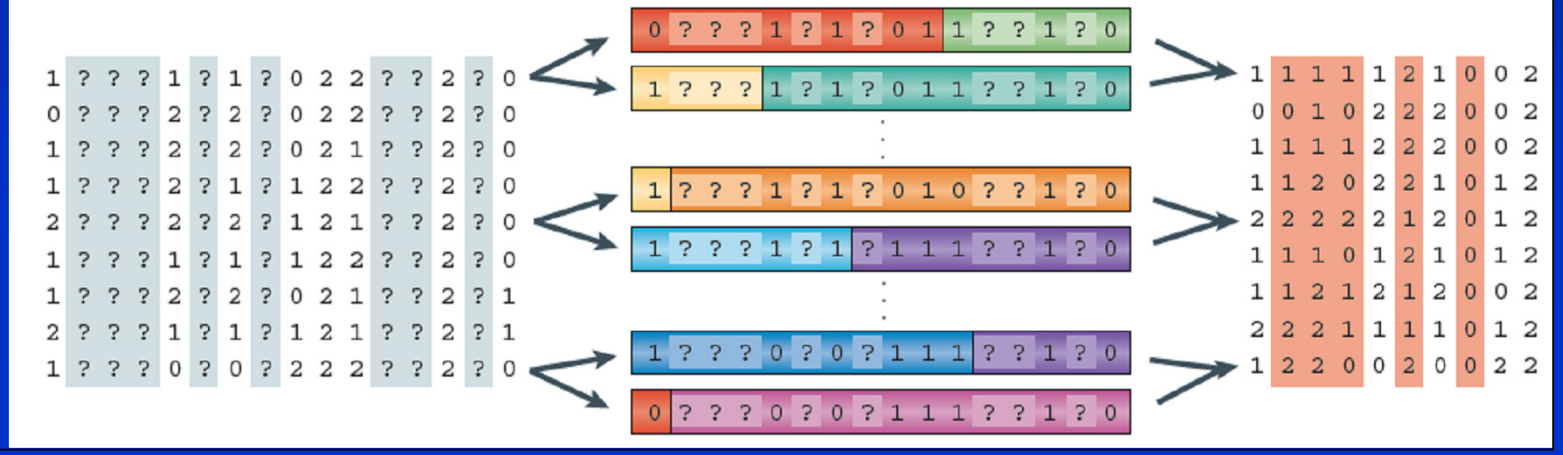
  - Some of the most effective are Beagle (Browning and Browning, 2009), MACH (Li et al., 2010), Impute2 (Howie et al., 2009), AlphaPhase (Hickey et al 2011)

Population based imputation

Reference population

Target population





Population based imputation

- Hidden Markov Models
  - Has "hidden states"
  - For target individuals these are "map" of reference haplotypes that have been inherited
  - Imputation problem is to derive genotype probabilities given hidden states, sparse genotypes, recombination rates, other population parameters

$$P(Gi|H,\theta,\rho) = \sum_{S} P(Gi|S,\theta)P(S|H,\rho)$$

Population based imputation

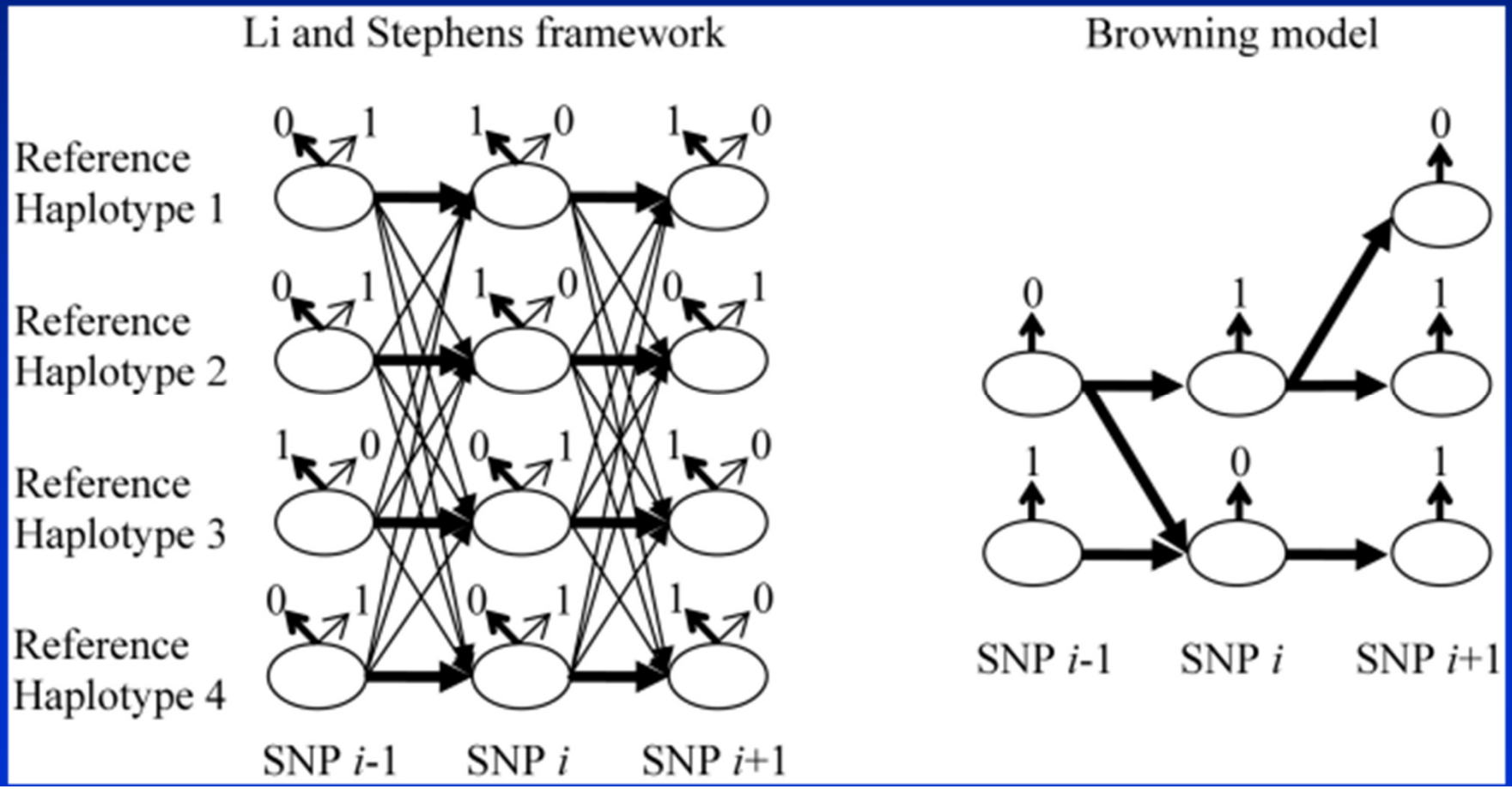
- Hidden Markov Models
  - Example with reference haplotypes
    - 011
    - 010
    - 101
    - 001
  - What are possible genotypes?

Population based imputation

- Hidden Markov Models

fastPHASE

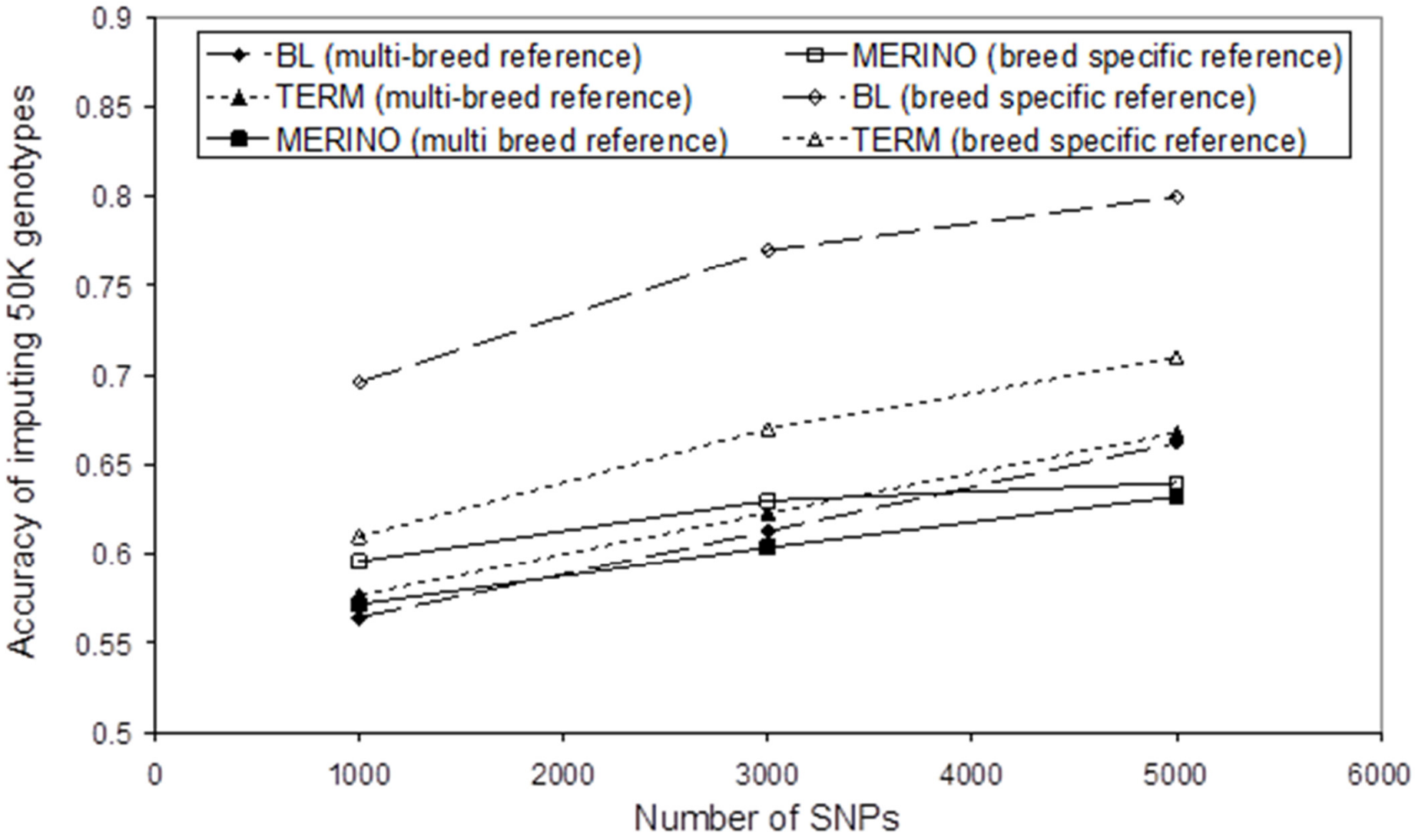
BEAGLE



Imputation accuracy

- Depends on
  - Size of reference set
    - bigger the better!
  - Density of markers
    - extent of LD, effective population size
  - Frequency of SNP alleles
  - Genetic relationship to reference

Imputation accuracy sheep



Imputation accuracy

- Density of markers (extent of LD)
  - In Holstein Dairy cattle
    - 3K -> 50K accuracy 0.93
    - 7K -> 50K accuracy 0.98

Illumina Bovine HD array

- We genotyped
  - 898 Holstein heifers
  - 47 Holstein Key ancestor bulls
  - 67 Jersey Key ancestor bulls
- After (stringent) QC **634,307** SNPs

Imputation 50K -> 800K

- Holsteins

	Cross validation	% Correct
Heifers only	1	96.7%
	2	96.7%
	Average	96.7%
Heifers	1	97.8%
using key	2	97.7%
ancestors	Average	97.7%

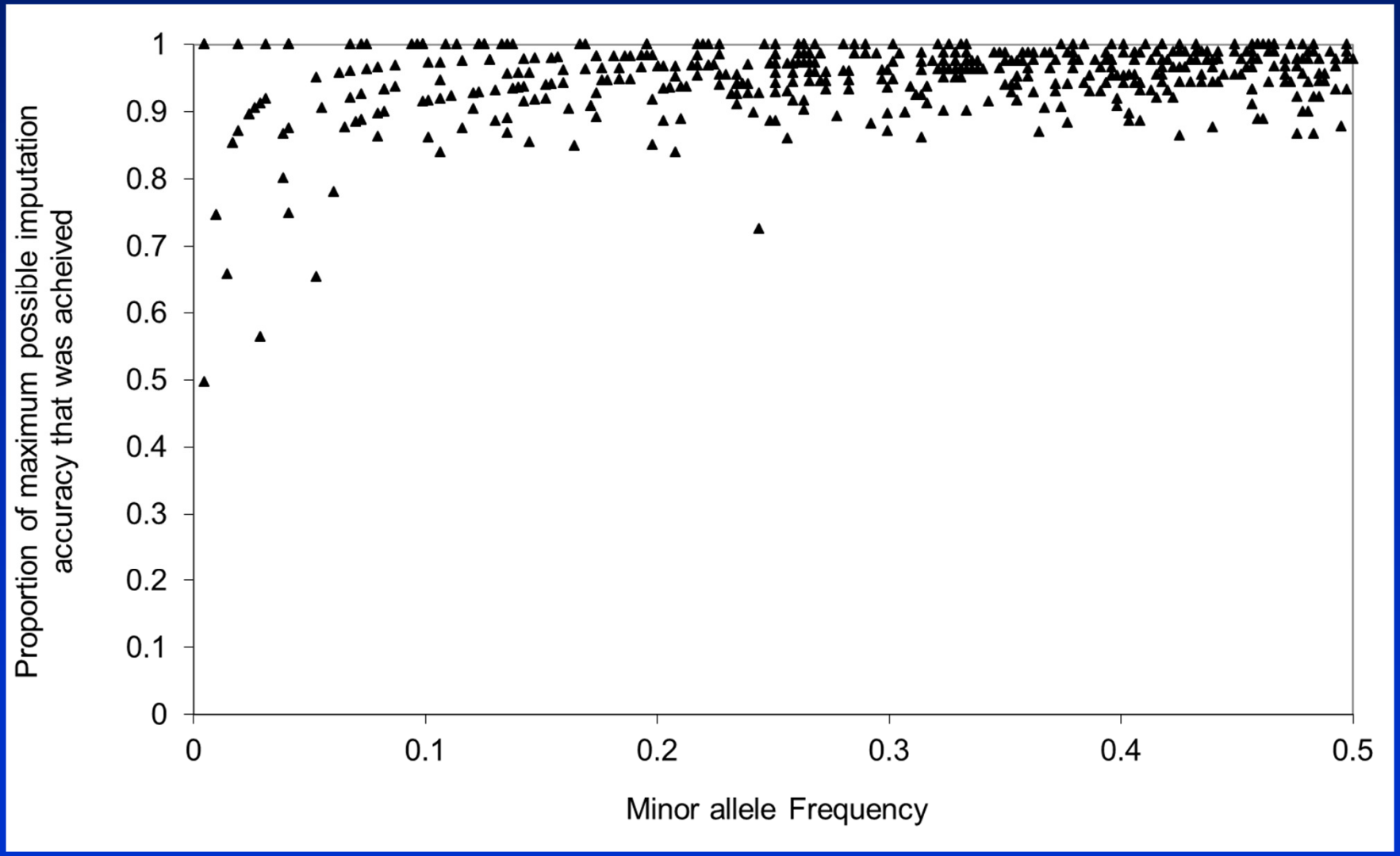
Imputation 50K -> 800K

- Jerseys

Cross validation	% Correct
1	95.2%
2	95.5%
3	95.3%
4	95.6%
5	96.2%
Average	95.6%

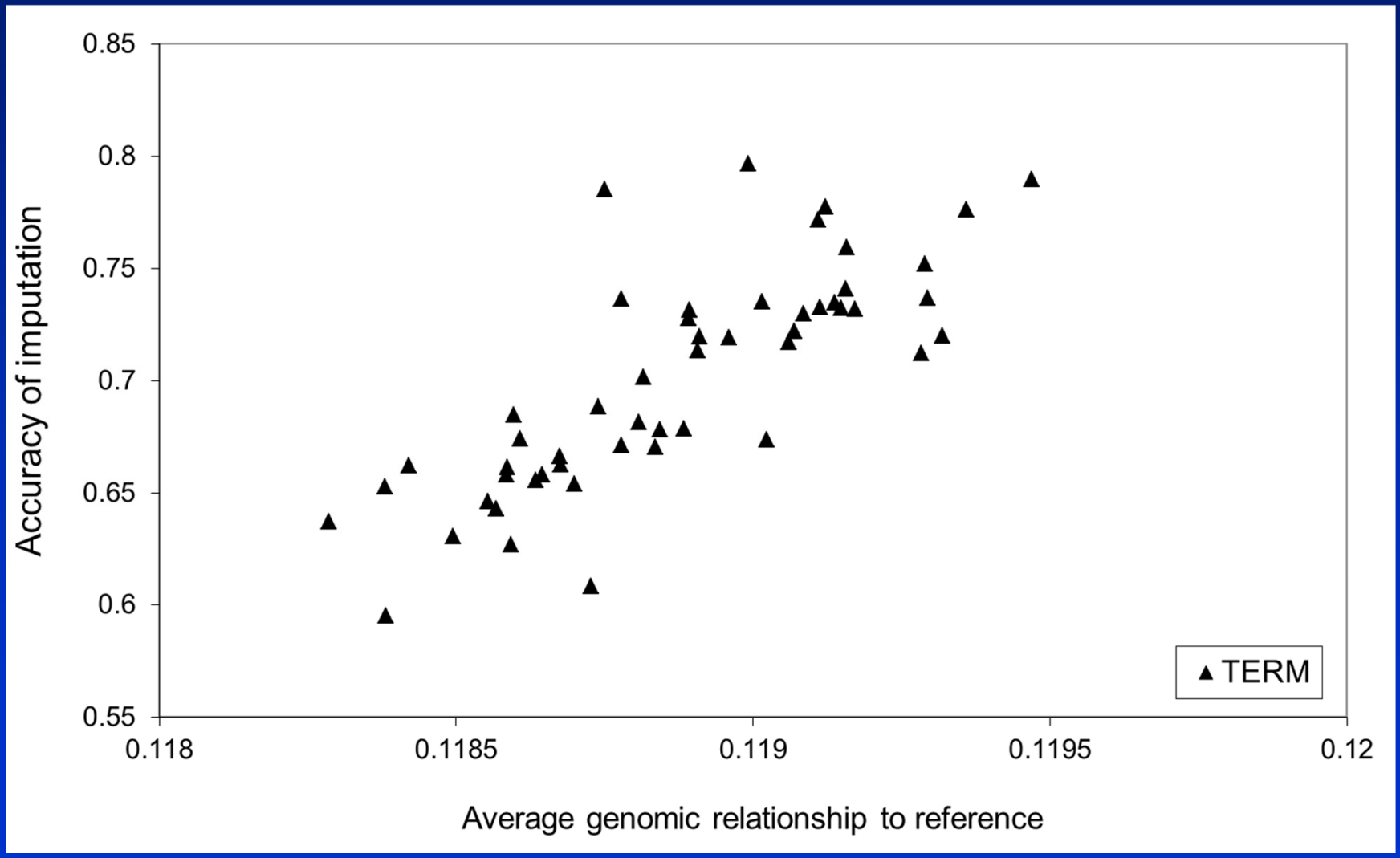
Imputation accuracy

- Rare alleles?



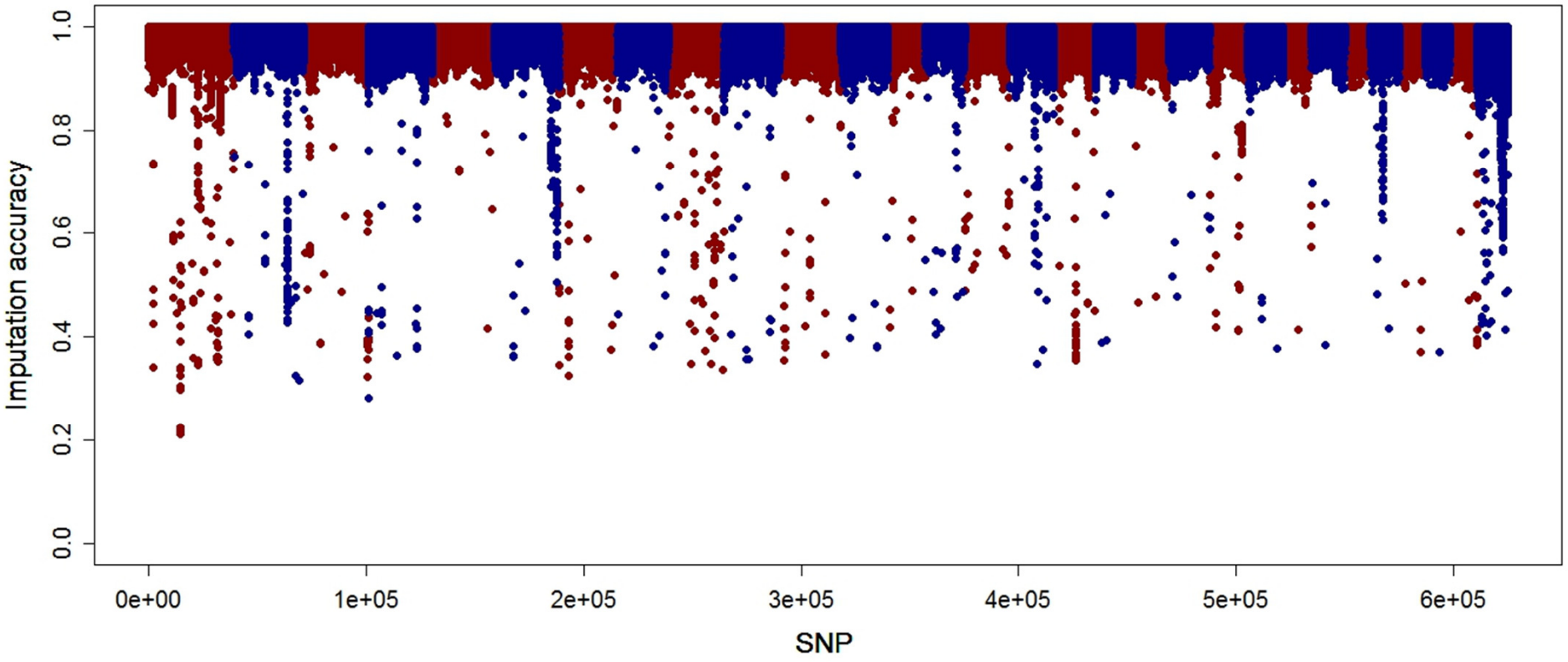
Imputation accuracy

- Relationship to reference?



Imputation of full sequence data

- Effect of map errors?



Why more power with imputation

- High accuracies of imputation demonstrate that we can infer haplotypes of animal genotyped with e.g. 3K accurately
- But potentially large number of haplotypes
- With imputed data can test single snp, only use 1 degree of freedom, rather than number of haplotypes

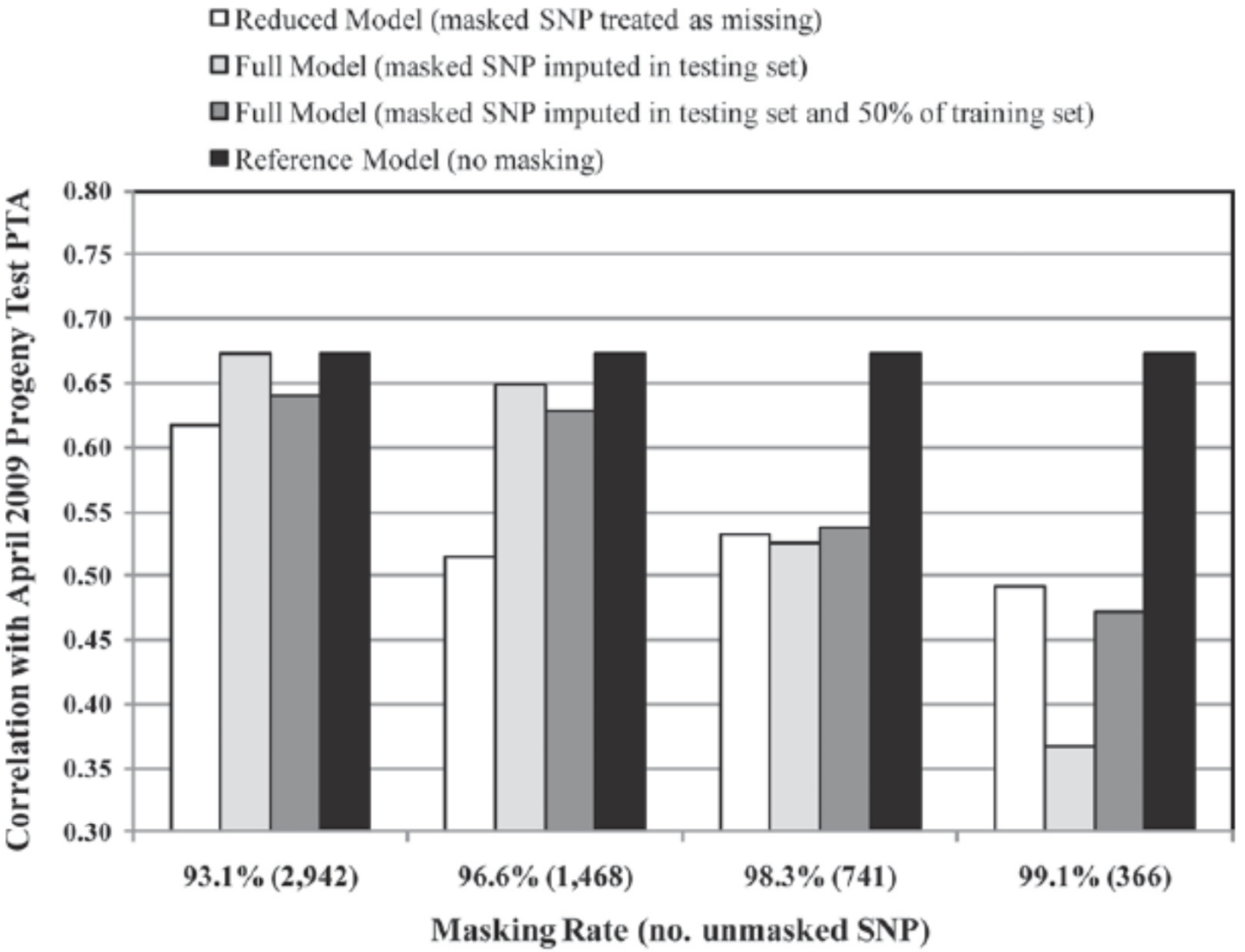


Figure 2. Correlations between predicted direct genomic values for milk yield and corresponding April 2009 progeny-test PTA using full or reduced models with 42,552 or 366, 741, 1,468, or 2,942 single SNP covariates, respectively, with or without imputation of masked genotypes for bulls in the testing set or bulls in the testing set and a randomly chosen 50% of bulls in the training set. The bars denoted as "reference" correspond to correlations from a full model in which all 42,552 SNP genotypes were left as unmasked in both the training and testing sets.

Using sequence data in genomic selection and GWAS

- Motivation
- Characteristics of sequence data
- Which individuals to sequence?
- Imputation of full sequence data
- Methods for genomic prediction with full sequence data
- Examples
  - GWAS in Rice, Cattle

Using sequence data in genomic selection and GWAS

- Motivation
  - Genome wide association study
    - Straight to causative mutation
  - Genomic selection (all hypotheses!)
    - No longer have to rely on LD, causative mutation actually in data set
      - Higher accuracy of prediction?
    - Better prediction across breeds?
      - Assumes same QTL segregating in both breeds
      - No longer have to rely on SNP-QTL associations holding across breeds
    - Better persistence of accuracy across generations

Using sequence data in genomic selection and GWAS

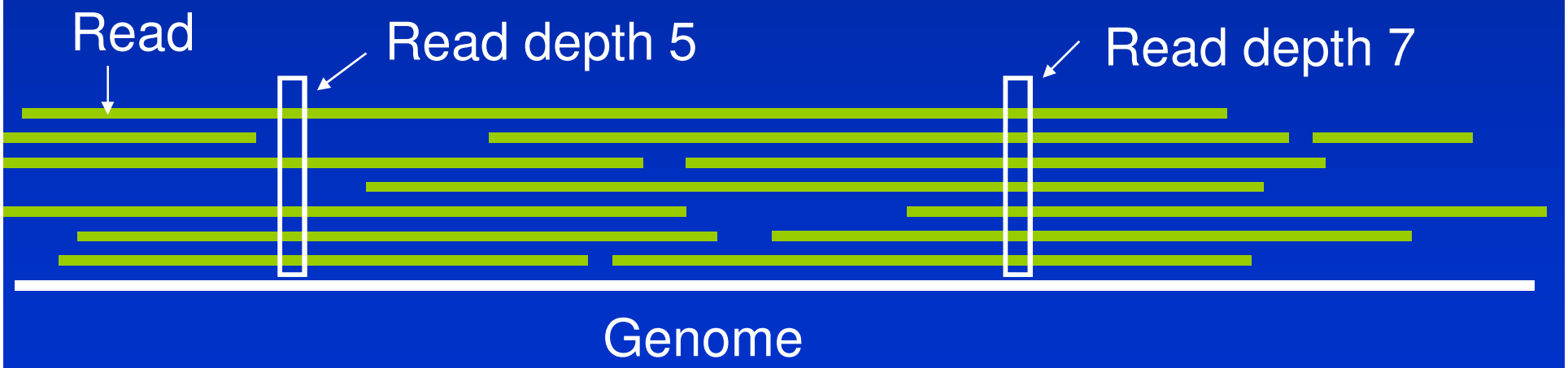
- Motivation
- Characteristics of sequence data
- Which individuals to sequence?
- Imputation of full sequence
- Methods for genomic prediction with full sequence data
- Examples
  - GWAS in Rice, Cattle

Sequence data

- Generates reads of DNA approx. 100 base pair (bp) length
- Reads are aligned to a reference genome
  - Or they could be assembled *de novo*
  - Assigns each read a location on genome
- Reads have an error rate!
  - One error per read
- Information is base pair (ACTG) + Quality score for each base
  - PHRED score = -10*log10(error rate)
    - 0.01 error rate = Q20
    - 0.001 error rate = Q30
    - 0.0001 error rate = Q40

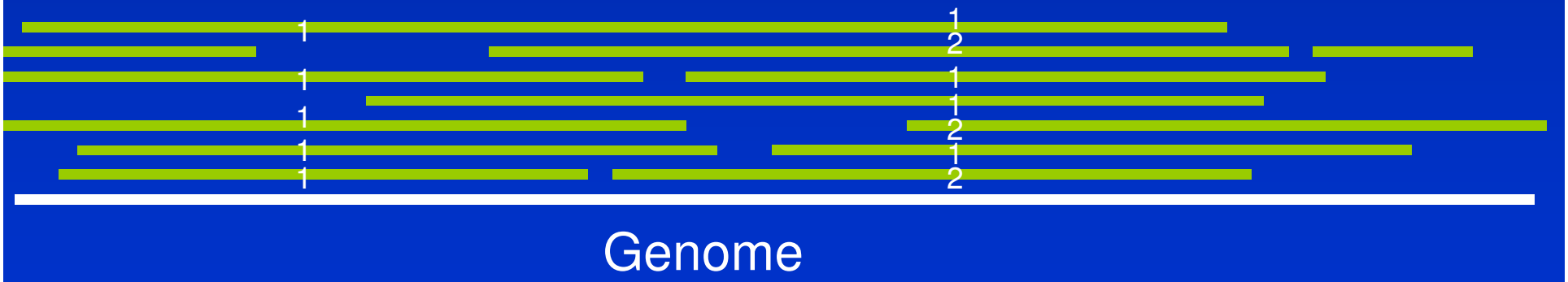
Read depth

- Each sequenced animal is aligned separately to reference
  - .bam files are created
- Read depth or fold coverage



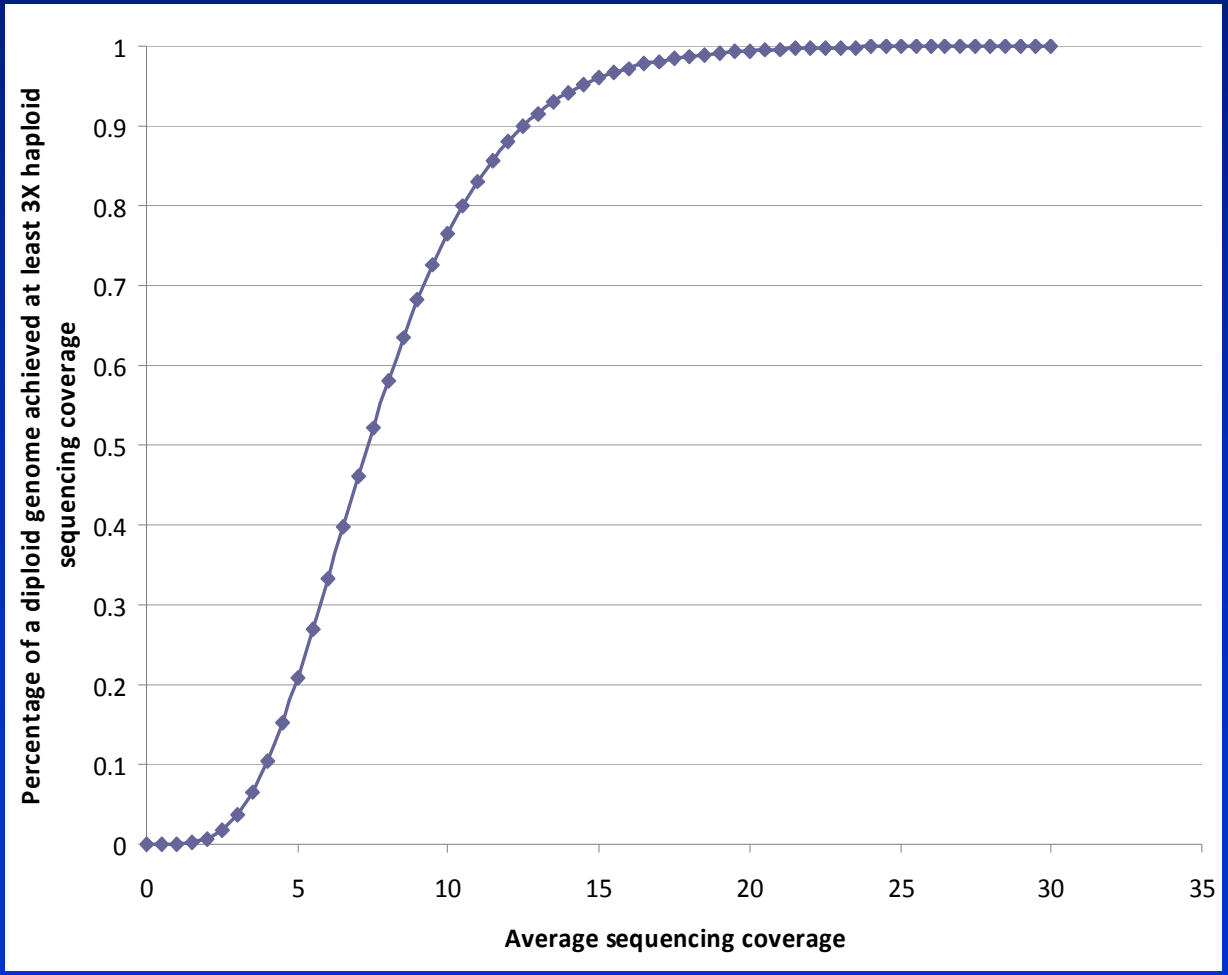
Importance of read depth

- Consider a heterozygous locus (animal carries 2 different alleles)
  - 50/50 chance of observing each allele in every read
- If read depth is low, it is possible to not observe an allele and therefore call a het locus homozygous
  - Read depth 5 → $0.5^5 = 0.03125$



What read depth is sufficient?

- Proportion of genome achieving at least 6x diploid coverage
- 12.5x achieves 90% in simulation below (Shen et al. 2010, Suppl. Material)

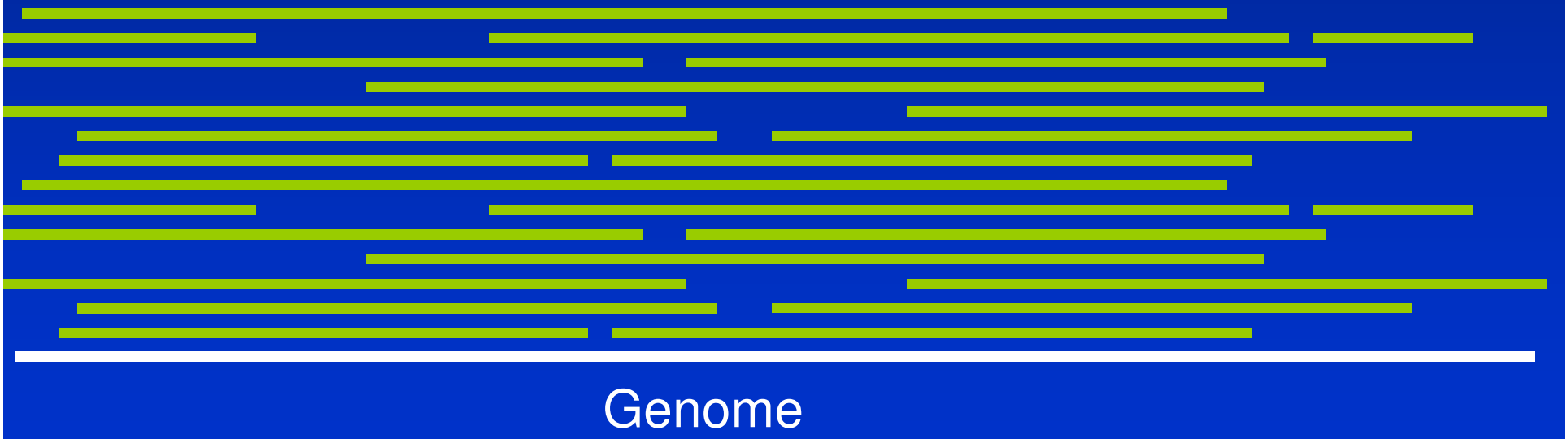


Heterozygosity and read depth

- SNP discovery
  - Missing some heterozygotes is not critical
    - Hopefully picked up in other animals
  - Just do more animals to identify SNP
  - Animal genotype not used directly
- Genotype calling
  - Missing heterozygotes a problem because incorrect genotype in downstream analysis
  - Statistical methods can be used to correct incorrect genotype calls
  - *Use genotype probabilities, not best guess!*

Identification of variants

- Program SAMtools
- stacks aligned bam files of multiple animals
- Calls variants and calculates quality/confidence statistics for calls
- http://samtools.sourceforge.net/mpileup.shtml



Variants in sequence

- SNP
- INDEL
  - INsertions and DELetions of DNA sections
- Copy number variants (CNV)
  - Repeated sections of DNA of various lengths

- Most studies to date have concentrated on SNP

Filtering of variants

- Reasons for filters:
- Number of artefacts of the sequencing process that lead to falsely identified variants
- Little evidence for a variant
  - Quality scores low
- Reasons against filters:
- Real variants may be lost
  - Low frequency SNP often have lower quality scores

Variant filters we use (vcf)

1. Read depth
   - Minimum read depth
     - Individual genotype calls will be low quality
   - Maximum read depth
     - Short reads of repetitive regions may be mapped to same locations causing massive read depth
2. Mapping quality
   - Low quality calls
3. Quality
   - Phred score
4. Multiple variants within 5bp window
   - Alignment errors and indels can cause shifts → call 2 SNP close together instead of 1
   - Remove SNP close to indels

Phred quality scores (Q)

- Related to base-calling error probabilities. Expressed in a range from 0 to 999 in our data.
- Probabilities are calculated by the following formula: $P = 10^{\frac{-Q}{10}}$
- e.g. Phred of 30 = error rate of 0.001
- Phred of 20 = error rate of 0.01

- Result is probability of each genotype at each variant eg. AA=0.95 AT=0.05 TT=0.00
- Use these in BEAGLE!

Imputation of full sequence data

Create BAM files

1. Filter reads on quality score, trim ends
2. Remove PCR duplicates
3. Align with BWA

BAM

Variant calling

SamTools mPileup
Vcf file -> filter (*number forward /reverse reads of each allele, read depth, quality, filter number of variants in 5bp window*)

Beagle Phasing in Reference
Input genotype probs from Phred scores
QC with 800K

Differences between SNP chip and sequence

- SNP chip
  - Sample of SNP
  - Higher minor allele frequency
  - Limited linkage disequilibrium depending on number of SNP
- Sequence
  - Contains most variants
    - SNP, indels, CNVs, etc
  - Allele frequency matches underlying causative variant frequency
  - Causative variants included
  - High linkage disequilibrium between variants

Using sequence data in genomic selection and GWAS

- Motivation
- Characteristics of sequence data
- Which individuals to sequence?
- Imputation of full sequence
- Methods for genomic prediction with full sequence data
- Examples
  - GWAS in Rice, Cattle

Which individuals to sequence?

- Those which capture greatest genetic diversity?
- Select set of individuals which are likely to capture highest proportion of unique chromosome segments

Which individuals to sequence?

- Let total number of individuals in population be *n*, number of individuals that can be sequenced be *m*.
- **A** = average relationship matrix among *n* individuals, from pedigree

- An example A matrix……..

Pedigree

Animal	Sire	Dam
1	0	0
2	0	0
3	0	0
4	1	2
5	1	2
6	1	3

Animals 6 is a half sib of 4 and 5

	Animal 1	Animal 2	Animal 3	Animal 4	Animal 5	Animal 6
Animal 1	1					
Animal 2	0	1				
Animal 3	0	0	1			
Animal 4	0.5	0.5	0	1		
Animal 5	0.5	0.5	0	0.5	1	
Animal 6	0.5	0	0.5	0.25	0.25	1

Which individuals to sequence?

- Let total number of individuals in population be *n*, number of individuals that can be sequenced be *m*.
- **A** = average relationship matrix among n individuals, from pedigree
- **c** is a vector of size *n*, which for each animal has the average relationship to the population (eg. Sum up the elements of **A** down the column for individual i)

Which individuals to sequence?

- If we choose a group of *m* animals for sequencing, how much of the diversity do they capture

- $\mathbf{p}_m = \mathbf{A}_m^{-1}\mathbf{c}_m$
  - Where $\mathbf{A}_m$ is the sub matrix of **A** for the *m* individuals, and $\mathbf{c}_m$ is the elements of the **c** vector for the *m* individuals
- Proportion of diversity = $\mathbf{p}_m\text{'}\mathbf{1n}$

Which individuals to sequence?

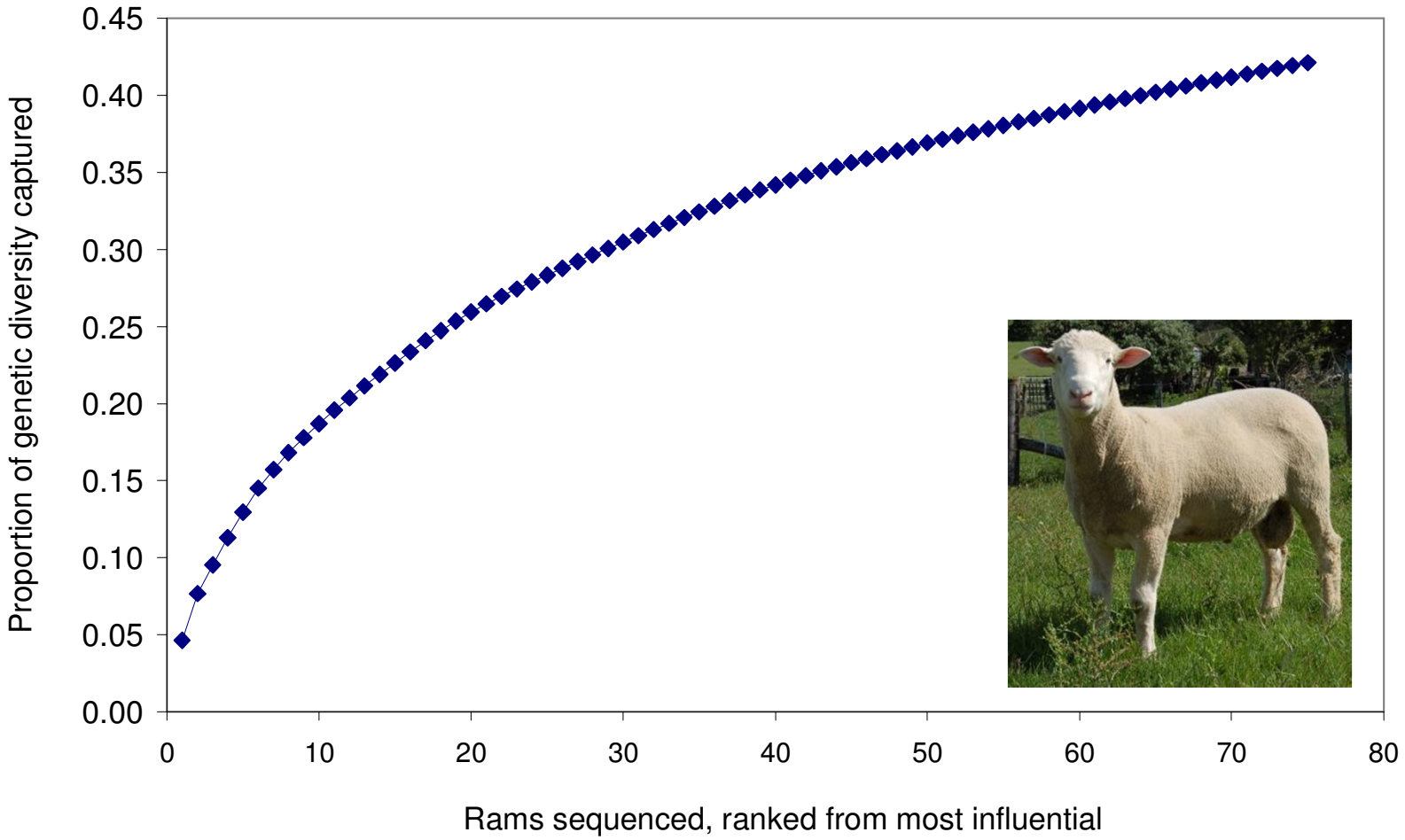
- Example

Which individuals to sequence?

- Then choose set of individuals to sequence (*m*) which maximise $\mathbf{p_m'1n}$
- Step wise regression
  - Find single individual with largest p_i, set c_i to zero, next largest p_i, set c_i to zero…..
- Genetic algorithm

Which individuals to sequence?

- Poll Dorset sheep



Which individuals to sequence?

- Then choose set of individuals to sequence (m) which maximise $\mathbf{p_m'1n}$
- Step wise regression
  - Find single individual with largest p_i, set c_i to zero, next largest p_i, set c_i to zero…..
- Genetic algorithm

- No **A**? Use **G**

Using sequence data in genomic selection and GWAS

- Motivation
- Characteristics of sequence data
- Which individuals to sequence?
- Imputation of full sequence data
- Methods for genomic prediction with full sequence data
- Examples
  - GWAS in Rice, Cattle

Imputation of full sequence data

- Two groups of individuals
  - Sequenced individuals: reference population
  - Individuals genotyped on SNP array: target individuals

Imputation of full sequence data

- Steps:
  - Step 1. Find polymorphisms in sequence data
  - Step 2. Genotype all sequenced animals for polymorphisms (SNP, Indels)
  - Step 3. Phase genotypes (eg Beagle) in sequenced individuals, create reference file
  - Step 4. Impute all polymorphisms into individuals genotyped with SNP array

Imputation of full sequence data

Create BAM files

1. Filter reads on quality score, trim ends
2. Remove PCR duplicates
3. Align with BWA

BAM

Variant calling

SamTools mPileup
Vcf file -> filter (*number forward /reverse reads of each allele, read depth, quality, filter number of variants in 5bp window*)

Beagle Phasing in Reference
Input genotype probs from Phred scores
QC with 800K

Reference file for imputation

Beagle Imputation in Target

SNP array data in target population

Genotype probabilities

Analysis

Genome wide association

Genomic selection

Imputation of full sequence data

- How accurate?

Imputation 50K -> 800K

- Holsteins

	Cross validation	% Correct
Heifers only	1	96.7%
	2	96.7%
	Average	96.7%
Heifers	1	97.8%
using key	2	97.7%
ancestors	Average	97.7%

Using sequence data in genomic selection and GWAS

- Motivation
- Characteristics of sequence data
- Which individuals to sequence?
- Imputation of full sequence data
- Methods for genomic prediction with full sequence data
- Examples
  - GWAS in Rice, Cattle

Methods for genomic prediction with full sequence

- 14 million SNPs in Holstein Friesian cattle?
- Which method is most appropriate
- Priors
  - BLUP (GBLUP) -> all SNPs in LD with QTL, very small effects
  - BayesA -> some SNPs have moderate to large effects, rest very small
  - BayesB -> many SNPs have zero effect, some have small to moderate effect?

Methods for genomic prediction with full sequence

- Meuwissen and Goddard 2010
  - Simulated population with full sequence data, ~ 900 mutations chosen to be QTL
  - Used BLUP and BayesB to predict GEBV

The accuracy of the predictions of total genetic value (±SE) in the TEST1 data set when the training data contained $T = 200$ individuals and GWBLUP or BayesB is used to estimate the marker effects

	Causative SNPs			
	GWBLUP		BayesB	
Data	Excluded	Included	Excluded	Included
3 QTL	0.503 ± 0.011	0.508 ± 0.011	0.938 ± 0.013	0.973 ± 0.004
30 QTL	0.491 ± 0.016	0.493 ± 0.010	0.806 ± 0.023	0.826 ± 0.019

Meuwissen, Goddard (2010) Genetics 185:623

Methods for genomic prediction with full sequence

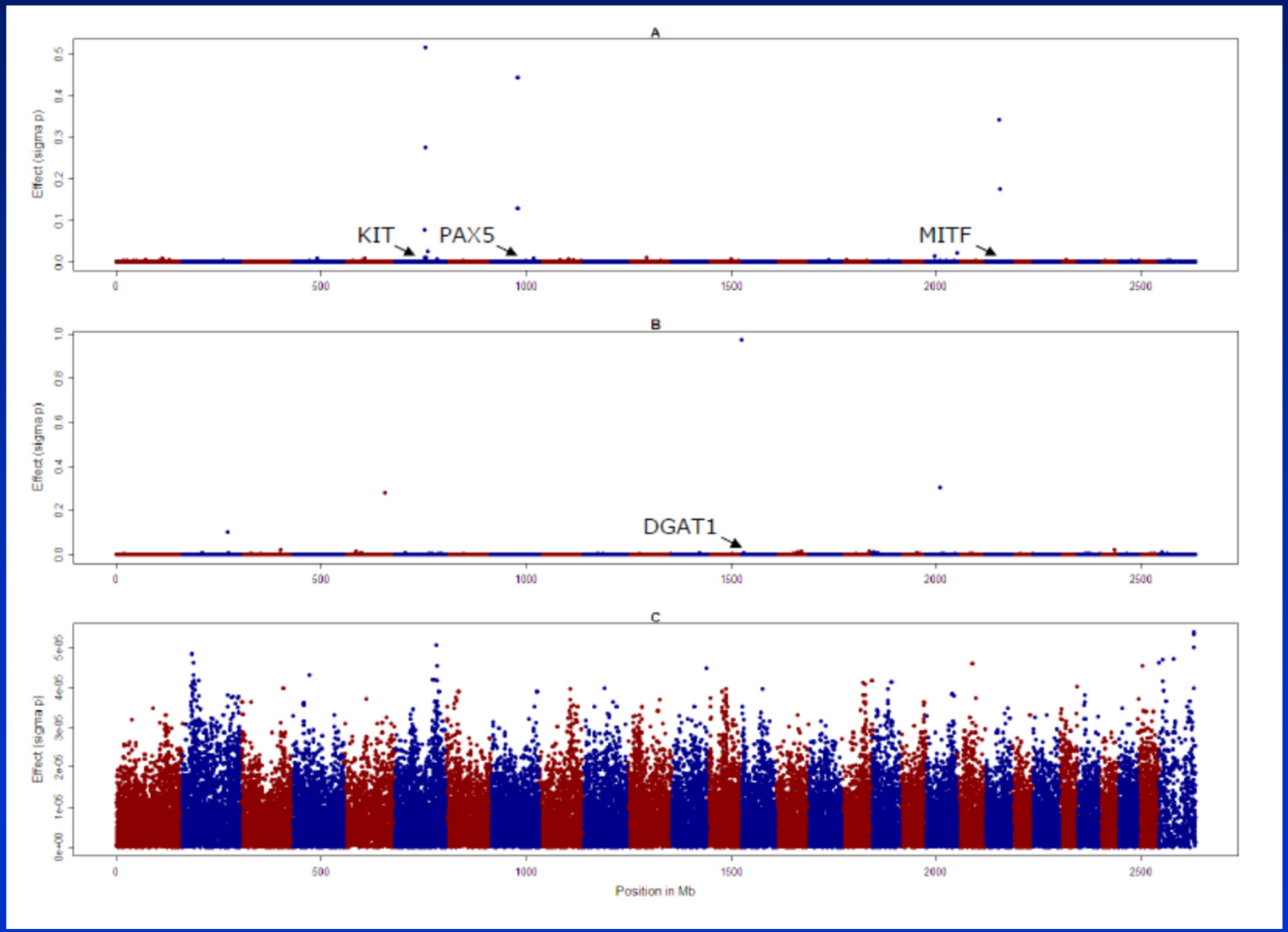
- Meuwissen and Goddard 2010
  - Simulated population with full sequence data, ~ 900 mutations chosen as QTL
  - Used BLUP and BayesB to predict GEBV
  - Large advantage of BayesB over BLUP
    - Prior matches their simulated data -> only 900 QTL amongst millions of SNP
  - 3% advantage of having mutation in data
  - Real data??

Methods for genomic prediction with full sequence

- Meuwissen and Goddard 2010
  - Better persistence of accuracy over generations

Causal SNPs	TEST1: $T = 200$, $L = 1$: 30 QTL	TEST2: $T = 200$, $L = 1$: 30 QTL
Excluded	0.806 ± 0.023	0.806 ± 0.022
Included	0.826 ± 0.019	0.824 ± 0.019

Genomic selection methods for GWAS?



Using sequence data in genomic selection and GWAS

- Motivation
- Characteristics of sequence data
- Which individuals to sequence?
- Imputation of full sequence data
- Methods for genomic prediction with full sequence data
- Examples
  - GWAS in Rice, Cattle

GWAS with sequence

ARTICLES

nature genetics

Genome-wide association studies of 14 agronomic traits in rice landraces

Xuehui Huang[1,2,10], Xinghua Wei[3,10], Tao Sang[4,10], Qiang Zhao[1,2,10], Qi Feng[1,10], Yan Zhao[1], Canyang Li[1], Chuanrang Zhu[1], Tingting Lu[1], Zhiwu Zhang[5], Meng Li[5,6], Danlin Fan[1], Yunli Guo[1], Ahong Wang[1], Lu Wang[1], Liuwei Deng[1], Wenjun Li[1], Yiqi Lu[1], Qijun Weng[1], Kunyan Liu[1], Tao Huang[1], Taoying Zhou[1], Yufeng Jing[1], Wei Li[1], Zhang Lin[1], Edward S Buckler[5,7], Qian Qian[3], Qi-Fa Zhang[8], Jiayang Li[9] & Bin Han[1,2]

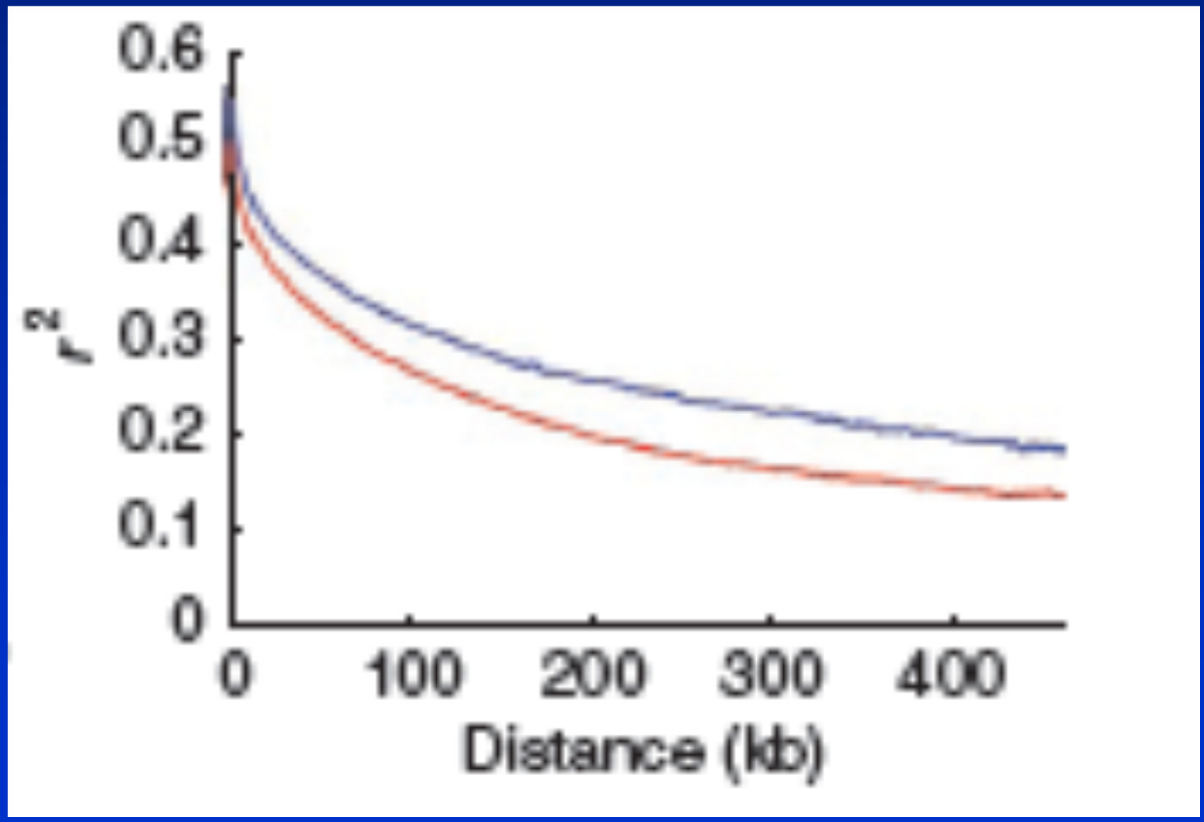
Uncovering the genetic basis of agronomic traits in crop landraces that have adapted to various agro-climatic conditions is important to world food security. Here we have identified ~3.6 million SNPs by sequencing 517 rice landraces and constructed a high-density haplotype map of the rice genome using a novel data-imputation method. We performed genome-wide association studies (GWAS) for 14 agronomic traits in the population of *Oryza sativa indica* subspecies. The loci identified through GWAS explained ~36% of the phenotypic variance, on average. The peak signals at six loci were tied closely to previously identified genes. This study provides a fundamental resource for rice genetics research and breeding, and demonstrates that an approach integrating second-generation genome sequencing and GWAS can be used as a powerful complementary strategy to classical biparental cross-mapping for dissecting complex traits in rice.

GWAS with sequence

- Huang et al. (2010)
  - Sequenced 517 rice landraces (inbred lines!) at 1x coverage
  - Represent ~ 82% of diversity in worlds rice cultivars
  - Called SNP in sequence pileups
    - 3.6 million SNP
  - With 1x coverage, could only call genotypes at ~ 20% of SNP
  - Therefore use imputation to fill in missing genotype
  - Example

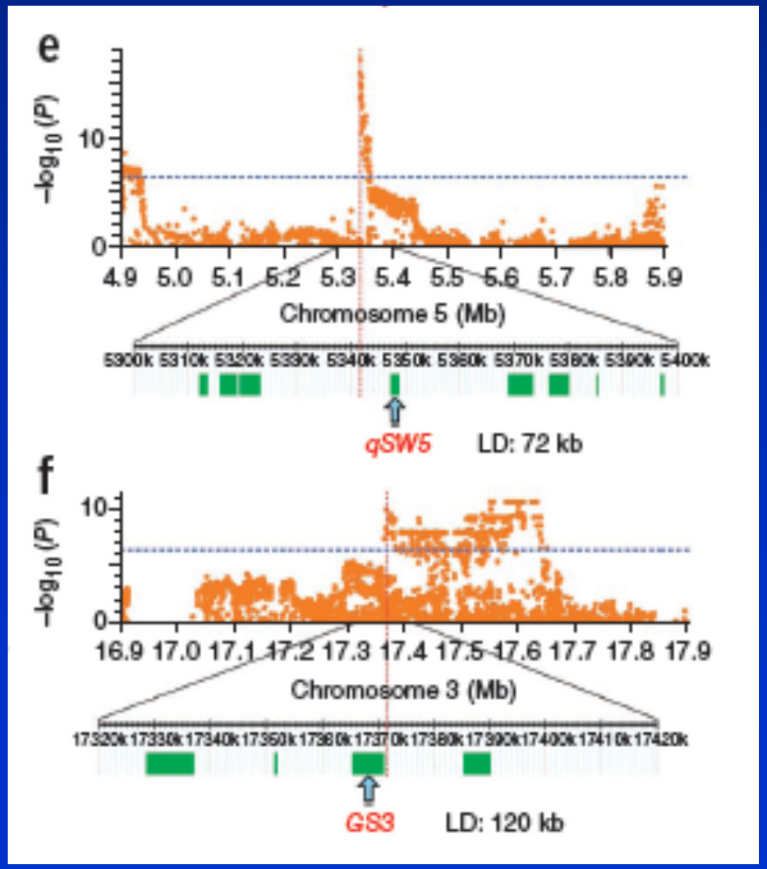
GWAS with sequence

- Huang et al. (2010)
  - Extent of LD



GWAS with sequence

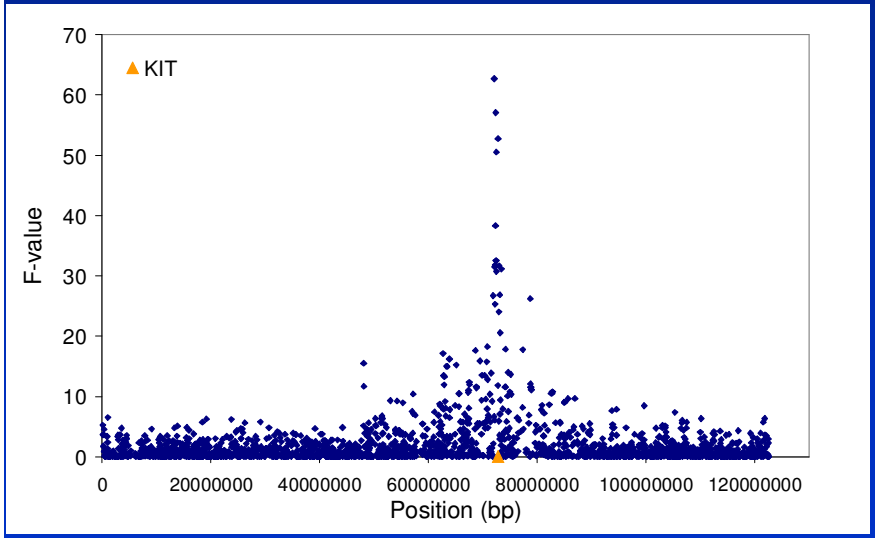
- Huang et al. (2010)
  - Now have 517 lines with 3.6 million SNP genotyped
  - Well characterised phenotypes for 14 agronomic traits
    - Grain size, flowering date, etc



  - Perform GWAS!
  - Confirmed known mutations
  - Many new mutations

GWAS with sequence

- KIT example
  - Earlier genome wide association study for proportion of black in Holsteins found association with SNP in KIT locus



  - Can we impute sequence in this region and re-run association study?

GWAS with sequence

	Average fold coverage	Filtered SNPs	Concordance with 800K
PICKARD-ACRES VIC KAI	10.4	3,061,950	
GLENAFTON ENHANCER	10.9	2,934,805	99.9%
BUSHLEA WAVES FABULON	11.3	4,249,998	97.4%
HANOVERHILL STARBUCK	12.5	3,237,681	97.9%
BIS-MAY S-E-L MOUNTAIN ET	12.6	3,009,463	98.5%
SHOREMAR PERFECT STAR	13.6	2,985,205	
ROYBROOK STARLITE	14.9	3,421,859	97.6%
TOPSPEED H POTTER	15.0	3,839,627	
LOCHAVON RAMESES	16.2	3,986,520	
BRAEDALE GOLDWYN	17.2	3,559,227	97.9%
CARENDA GRAVITY	17.8	4,331,849	96.8%
ONKAVALE GRIFFLAND MIDAS	22.5	3,742,799	

Imputation of full sequence data

Create BAM files

1. Filter reads on quality score, trim ends
2. Remove PCR duplicates
3. Align with BWA

BAM

Variant calling

SamTools mPileup
Vcf file -> filter (*number forward /reverse reads of each allele, read depth, quality, filter number of variants in 5bp window*)

Beagle Phasing in Reference
Input genotype probs from Phred scores
QC with 800K

Reference file for imputation

Beagle Imputation in Target

SNP array data in target population

Genotype probabilities

Analysis

Genome wide association

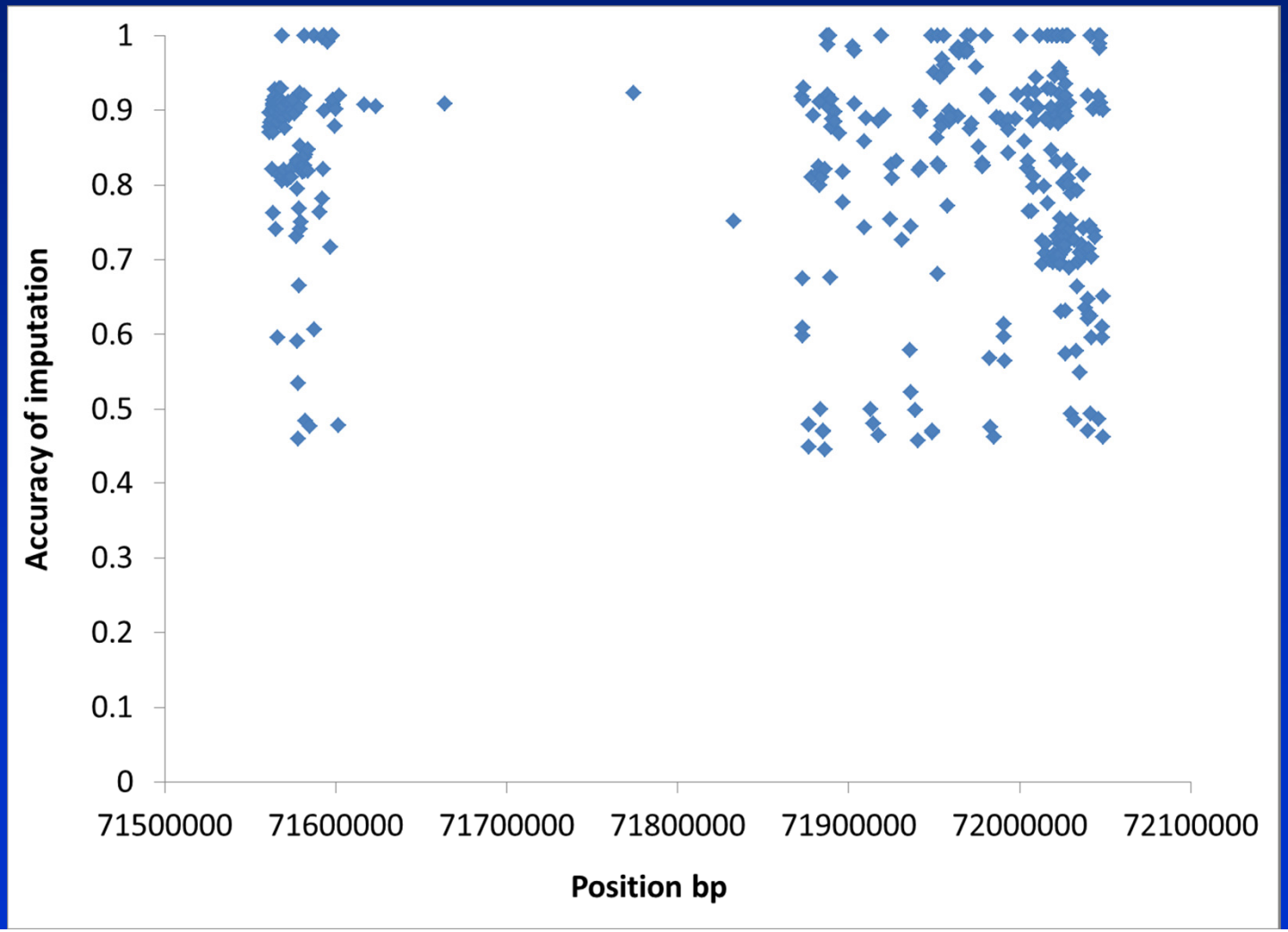
Genomic selection

GWAS with sequence

- KIT example
  - In sequenced bulls, compile list of SNPs/Indels in KIT region (352/20)
  - Call genotypes for the 372 variants in the 12 bulls
  - Use this as reference file for imputing the 372 variants in 697 bulls with % black phenotype (from 800K) data
  - Run association study on the 372 variants imputed in 697 bulls

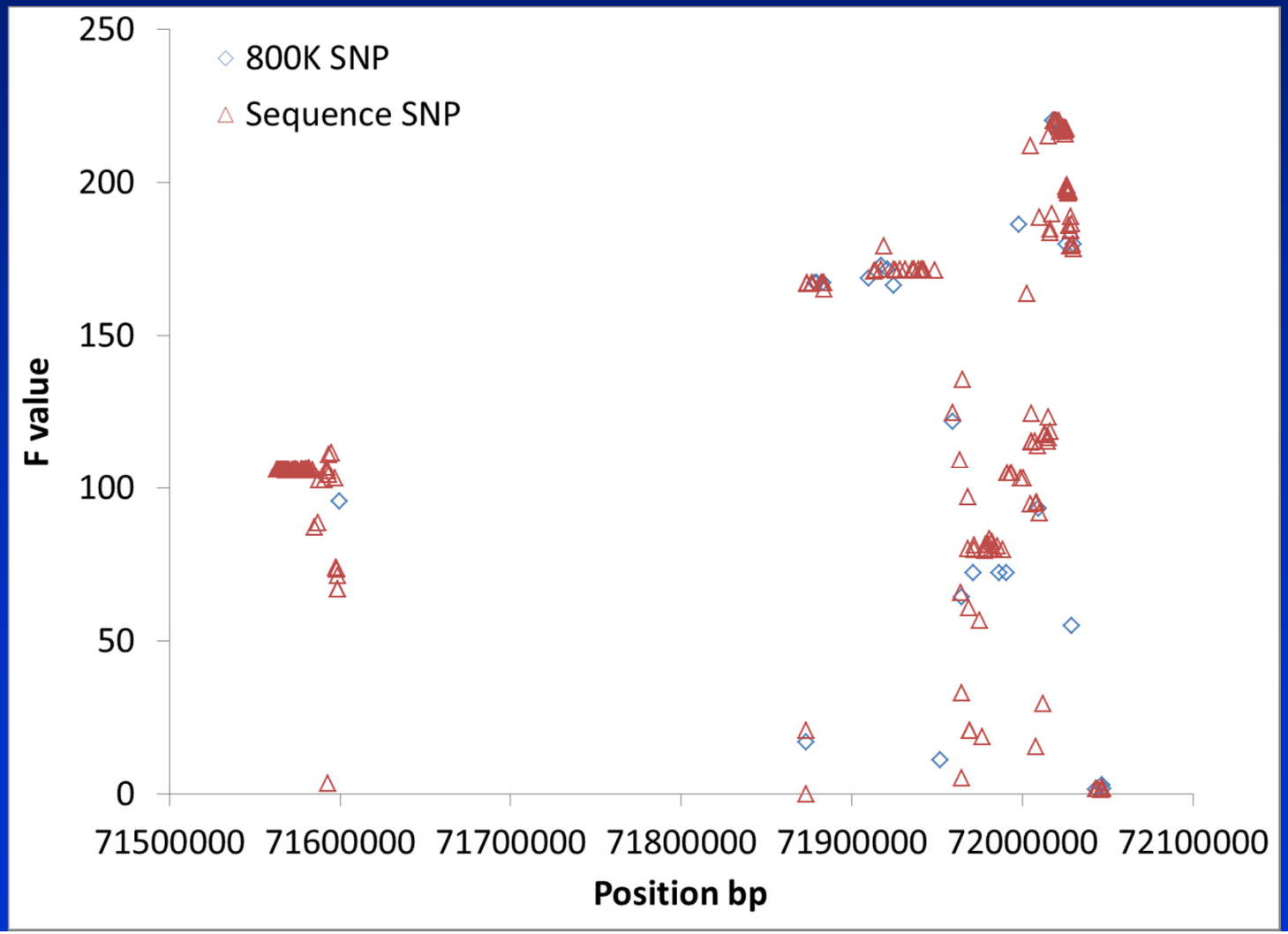
GWAS with sequence

- KIT example



GWAS with sequence

- KIT example



1000 bull genomes on the cloud

- We will all need "reference" population of many sequenced bulls to impute from
  - SNP, indel and CNV genotypes
  - The more bulls the better!
- We propose a project where we each upload our sequence files (BAM) for each bull to a shared server
- Run SNP/indel/CNV calling software every new 100 bulls uploaded

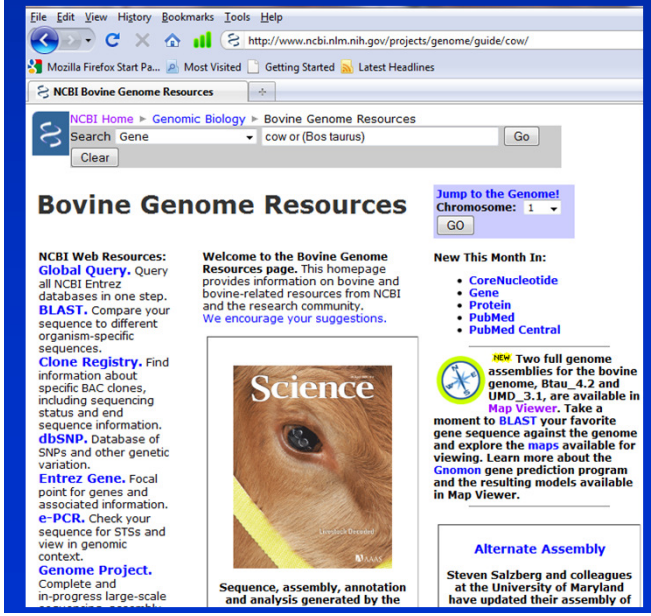
- Contributors can download SNP/indel/CNV genotype file on all bulls to use for imputation anytime
- Partners welcome!

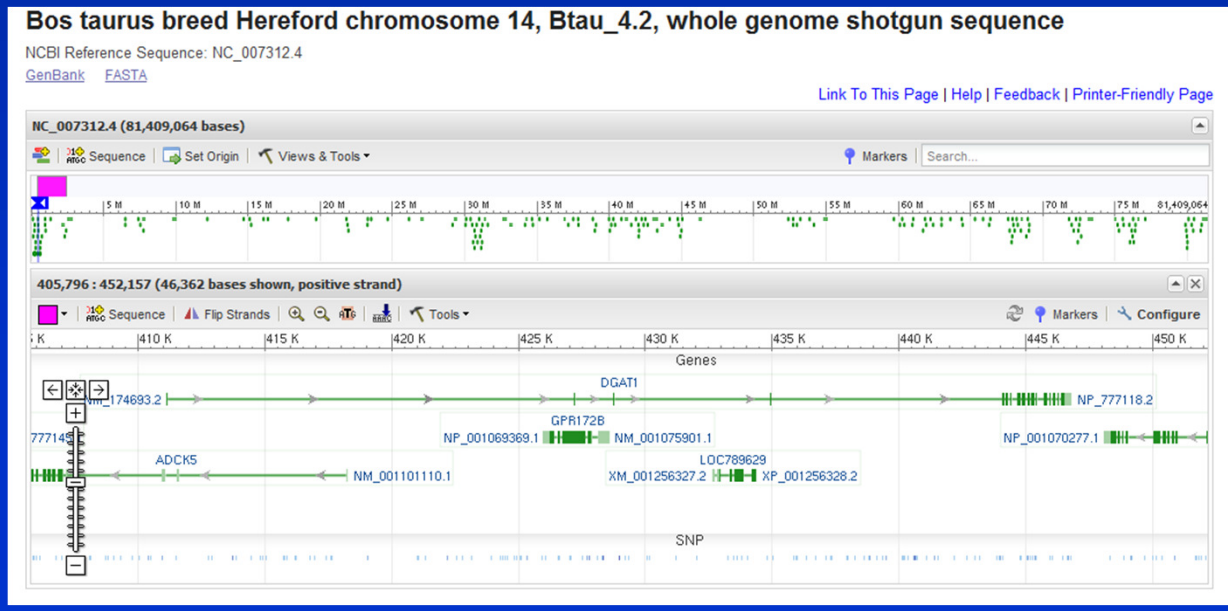
GWAS with sequence

- An alternative approach to GWAS?
  - For a target QTL region, sequence bulls of known QTL genotype (eg QQ,Qq,qq)
  - Have converted complex trait into a Mendelian trait
  - Far fewer individuals required for same power
  - Requires knowledge from linkage studies/previous GWAS!
  - Which method is more successful?

Quality of reference genomes?

- Cattle
  - Bovine build 4.2
  - Annotated
    - But many genes no assigned function
  - No Y chromosome yet, X is messy
  - ~ 9.5 million putative SNP in dbSNP





Quality of reference genomes?

- Cattle
  - Bovine build 4.2
  - Annotated
    - But many genes no assigned function
  - No Y chromosome yet, X is messy
  - ~ 9.5 million putative SNP in dbSNP

  - Map of copy number variation?
  - Kijas et al. (2010) – 51 CNV detected, 82% spanned at least one gene
  - Hou et al. (2011) – 682 CNV from SNP array intensity data

Conclusions

- Potential of whole genome sequence data
  - Enable genome wide association study -> straight to causative mutation
  - Genomic selection
    - No longer have to rely on LD, causative mutation actually in data set, Higher accuracy of prediction?, Better persistence of accuracy across generations

Conclusions

- Potential of whole genome sequence data
  - Enable genome wide association study -> straight to causative mutation
  - Genomic selection
    - No longer have to rely on LD, causative mutation actually in data set, Higher accuracy of prediction?, Better persistence of accuracy across generations
- Choose individuals to sequence based on genetic contribution to population?
- Imputation of target population genotyped with SNP arrays
  - Caution with low frequency alleles, relationship to reference
- Large collaborative projects required for bovine/plant communities?