Low-Pass Primer

Troy Rowan

BIF Genomic Prediction Workshop

December 19th, 2023

UTIA INSTITUTE OF AGRICULTURE
THE UNIVERSITY OF TENNESSEE

MacG Jul '16

DNA variants→ Genetic variation

SNPs- single base pair difference at a location (e.g. A/C allele)

– ATAGTCACTAAG–

– ATAGTCCCTAAG–

Indel - multi-base pair insertion or deletion at a location

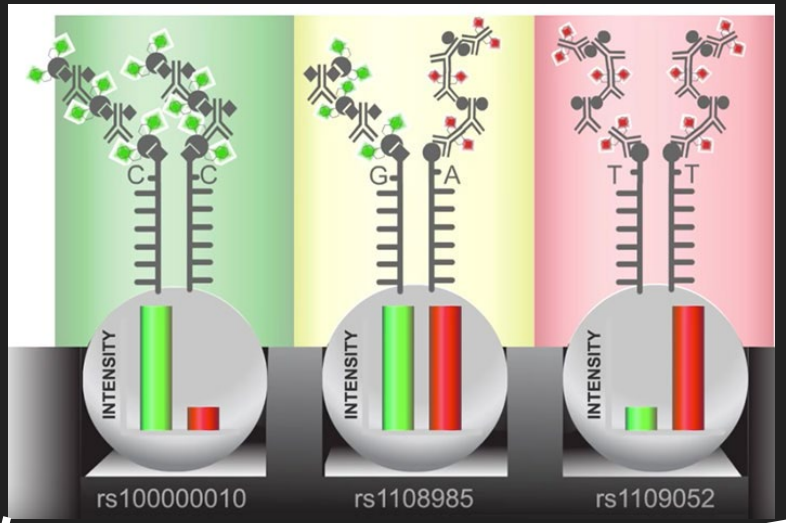
– ATAGTCCCTAAGGTCTTGCCAG–

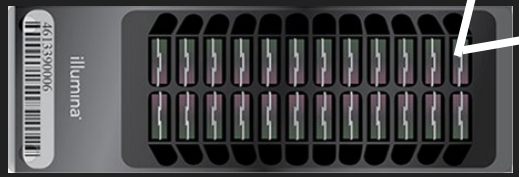
– ATAGTCGTCTTGCCAG–

Number of variants is dependent on the number (and diversity) of animals observed… Though not linear.

SNP Chip Genotyping

- Fixed locations genotyped
- Reduced representation of bovine genome (effective segments)
- SNP discovery needed (sequence large number of individuals)
- Probe design
- Ascertainment bias
- “Good enough” resolution





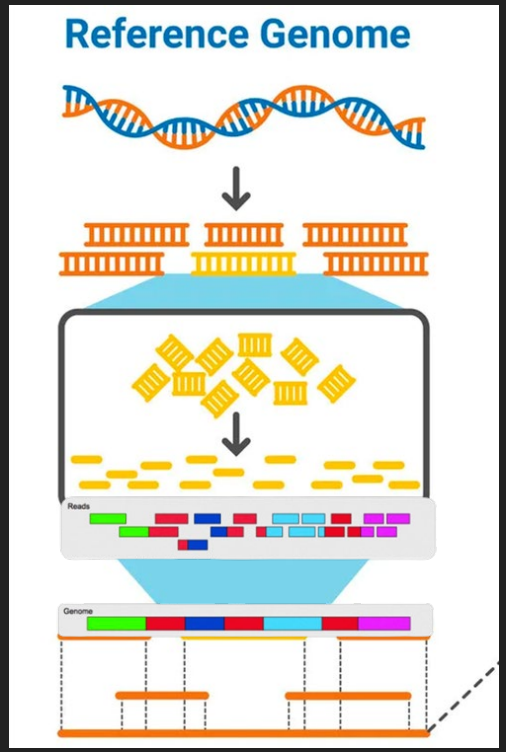
What is ~~Low Pass~~ Sequencing?

“Shotgun Sequencing”

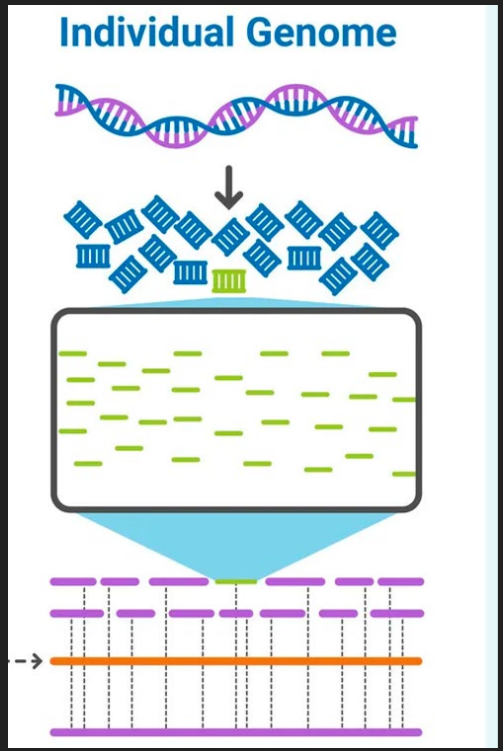
1) Break genome into small chunks

2) “Read” DNA sequence of chunks

3) Use overlapping parts of sequences to determine where things belong



4) “Reference Genome” serves as backbone for future sequencing efforts



5) Subsequent sequencing still “reads” small chunks of DNA

6) No need for *de novo* assembly once reference is available

7) Align reads and identify differences (i.e. SNPs)

Reference Genomes

Linear (for now) haploid representation of a species' genomic content

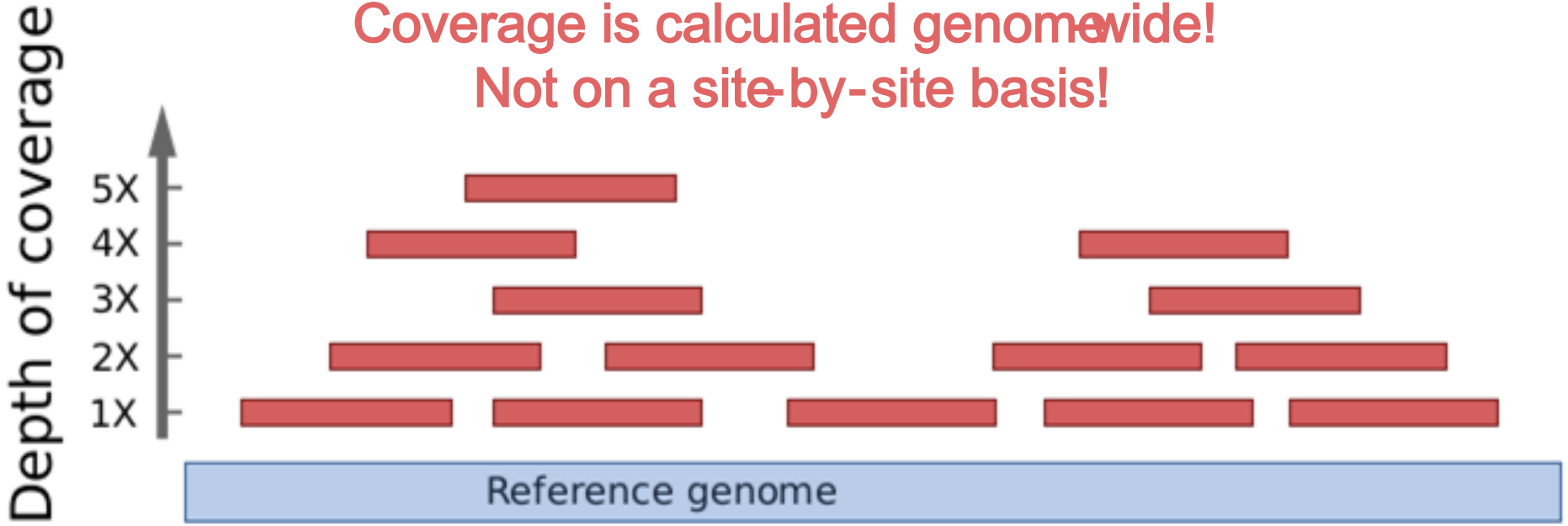
Essential for ANY position-reliant data generation/analysis

Content is based on one (inbred) Hereford animal (Dominette)

```
>NG_008679.1:5001-38170 Homo sapiens paired box 6 (PAX6)
ACCCTCTTTTCTTATCATTGACATTTAAACTCTGGGGCAGGTCCTCGCGTAGAACGCGGCTGTCAGATCT
GCCACTTCCCCTGCCGAGCGGCGGTGAGAAGTGTGGGAACCGGCGCTGCCAGGCTCACCTGCCTCCCCGC
CCTCCGCTCCCAGGTAACCGCCCGGGCTCCGGCCCCGGCCCGGCTCGGGGCCCGCGGGGCCTCTCCGCTG
CCAGCGACTGCTGTCCCCAAATCAAAGCCCGCCCCAAGTGGCCCCGGGGCTTGATTTTTGCTTTTAAAAG
GAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGGTGGAGGAGGGACTTGTCTT
TGCCGAGTGTGCTCTTCTGCAAAAGTAGCAAAATGTTCCACTCCTAAGAGTGGACTTCCAGTCCGGCCCT
GAGCTGGGAGTAGGGGGCGGGAGTCTGCTGCTGCTGTCTGCTAAAGCCACTCGCGACCGCGAAAAATGCA
```

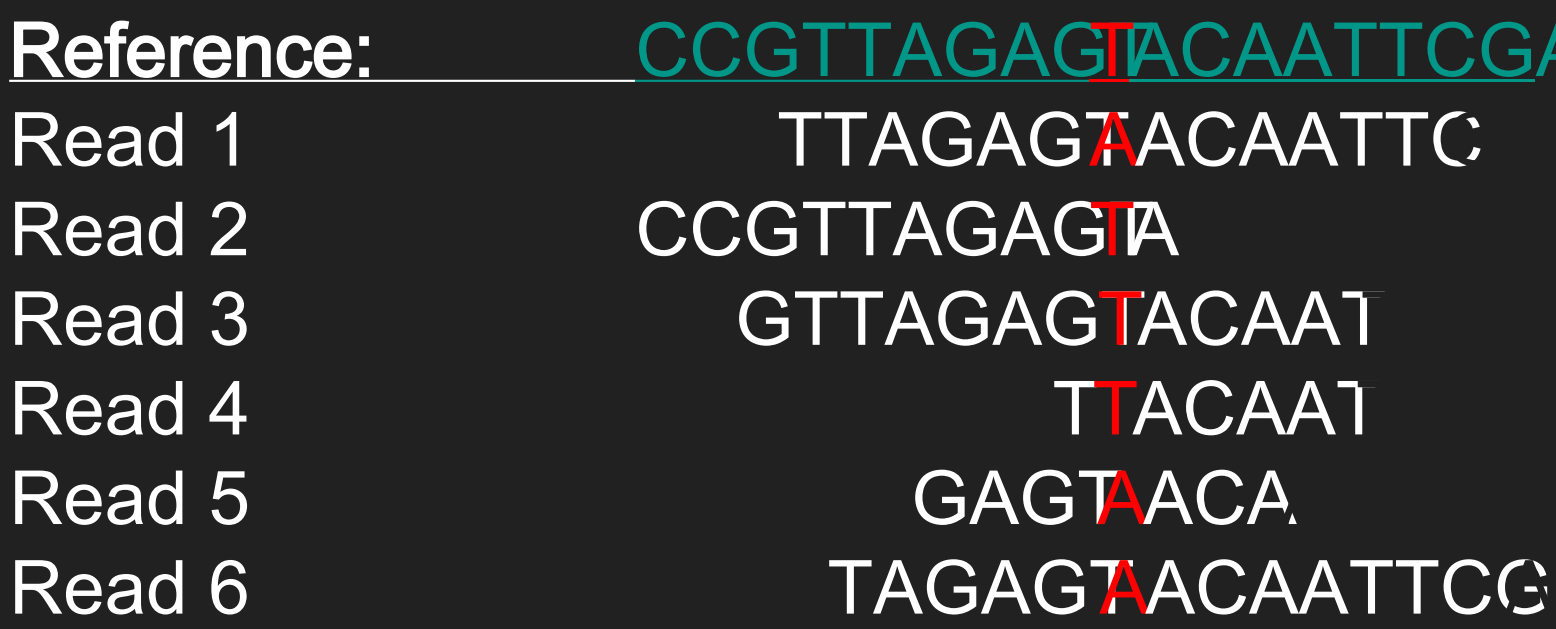
$$Coverage = \frac{nReads \times len(read)}{Genome \quad Size}$$

Coverage is calculated genome-wide!
Not on a site-by-site basis!



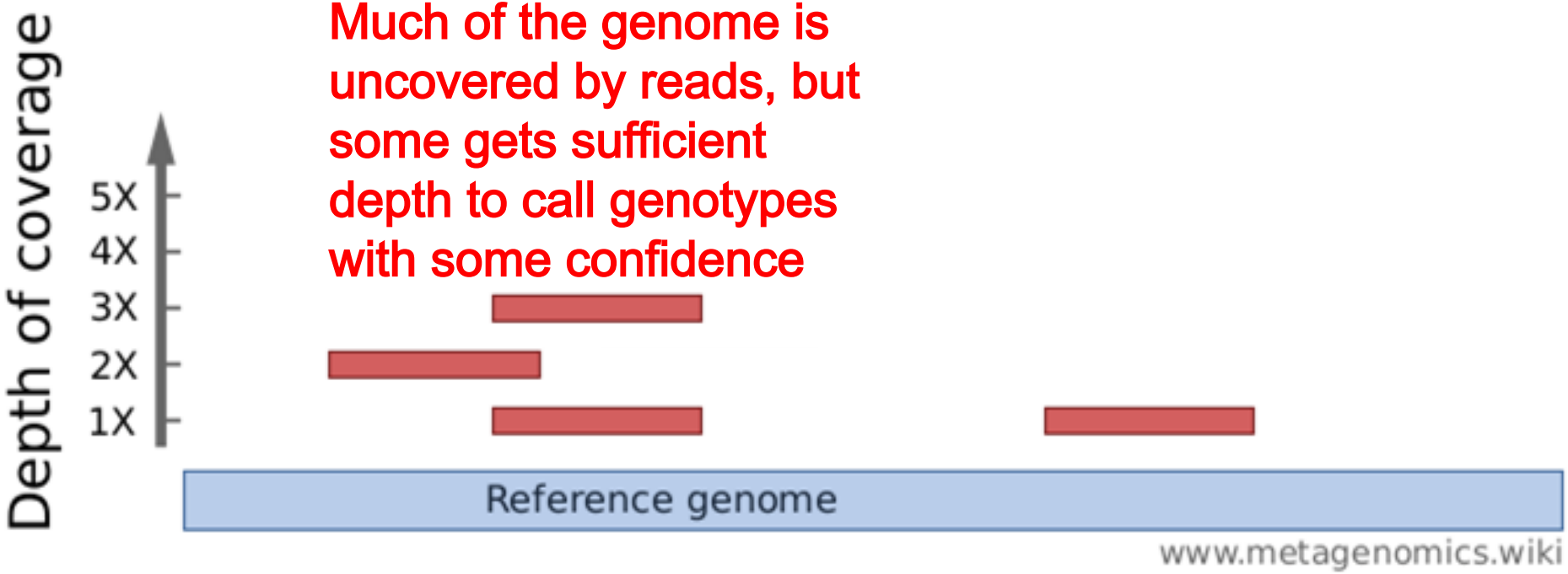
www.metagenomics.wiki

Genotype Calling: More reads = More confidence

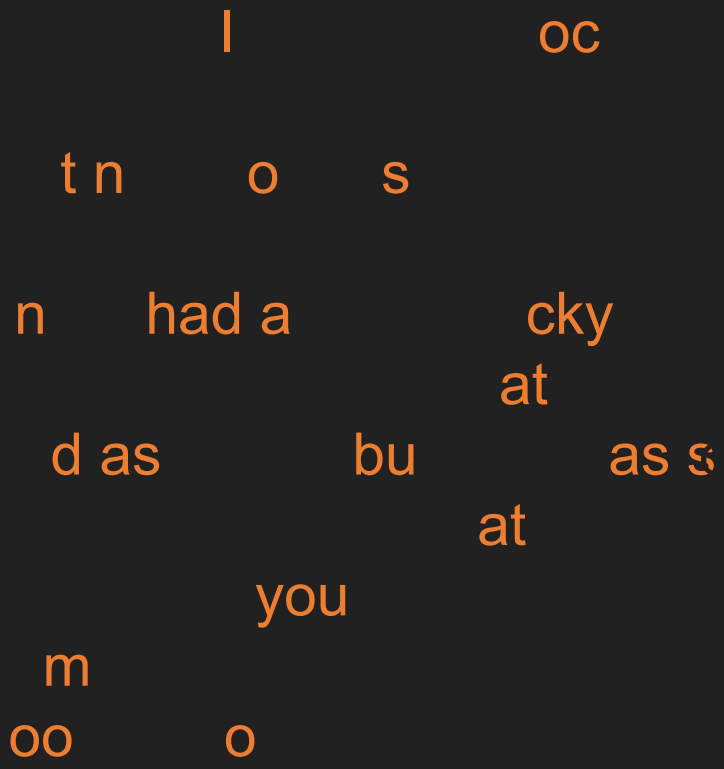


What is Low-Pass Sequencing?

Low-Pass sequencing



Low pass by itself may not be useful in genomic prediction…



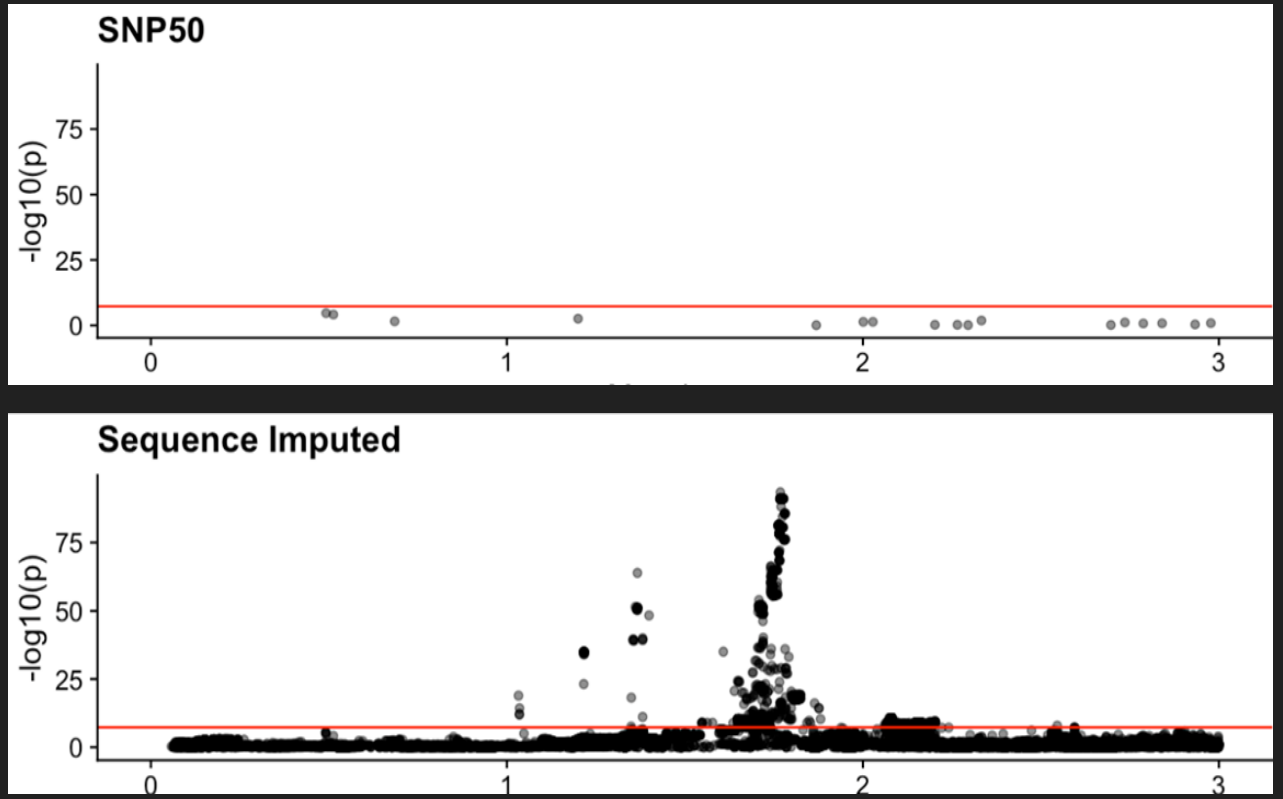
… but imputation helps us fill in the missing variants

Wish that I was on ol' Rocky Top
Down in the Tennessee hills
Ain't no smoggy smoke on Rocky Top
Ain't no telephone bills
Once I had a girl on Rocky Top
Half bear, other half cat
Wild as a mink, but sweet as soda pop
I still dream about that
Rocky Top, you'll always be
Home sweet home to me
Good ol' Rocky Top
Rocky Top, Tennessee
Rocky Top, Tennessee

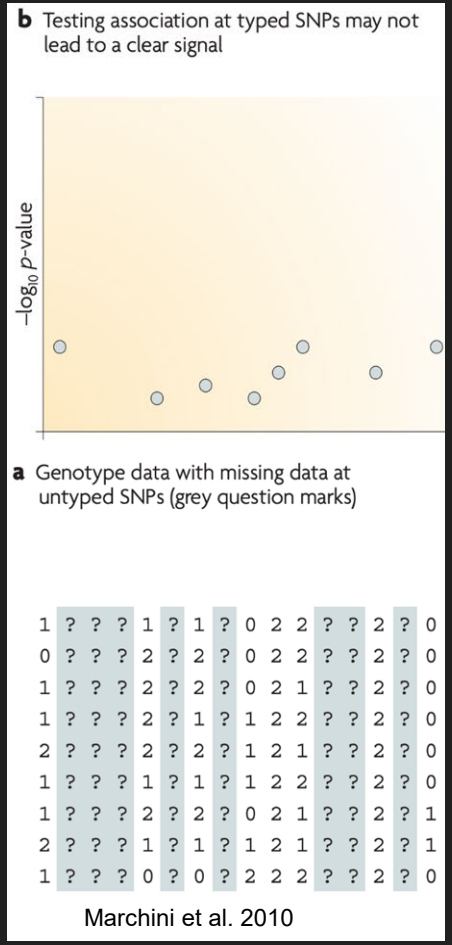
The Power of Imputation

Sufficient for resolving relationships (i.e., making a GRM)

Necessary for mapping "causal" variants in large-scale datasets



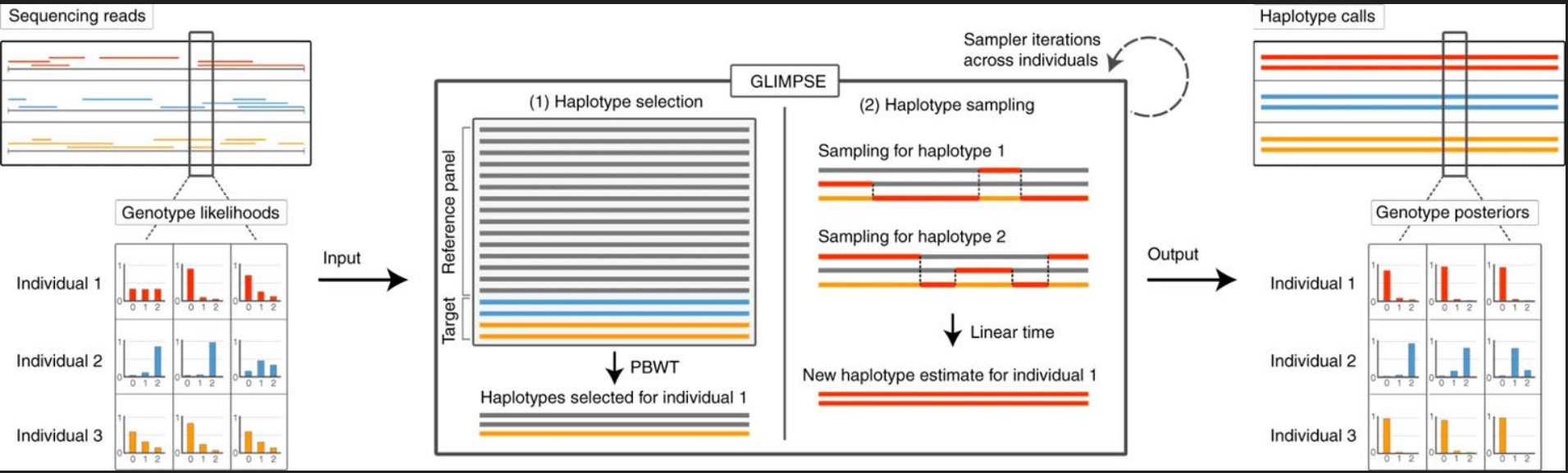
Rowan et al. 2021



Marchini et al. 2010

Whole genome density imputation ~30 million SNPs

GLIMPSE LowPass Imputation Algorithm



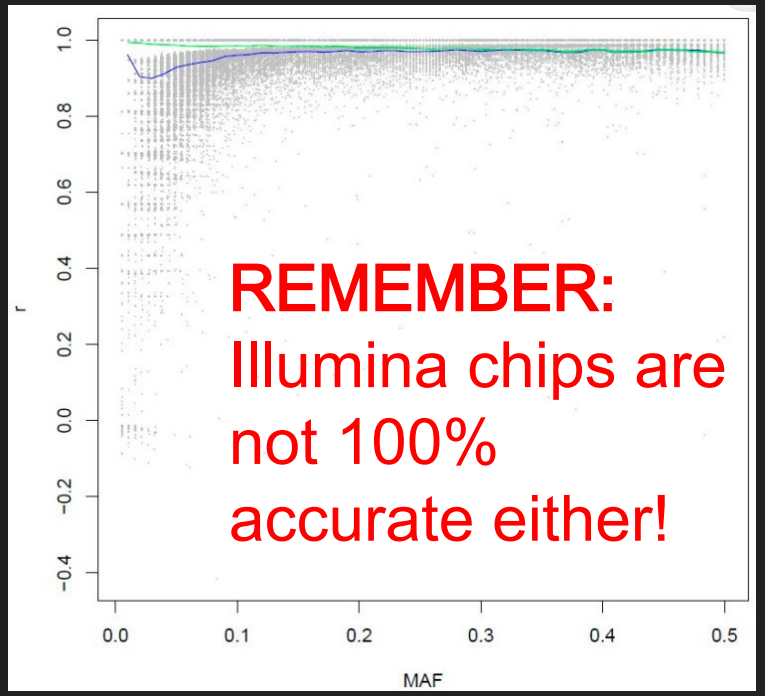
Rubinacci et al. 2023

What does accurate imputation need?

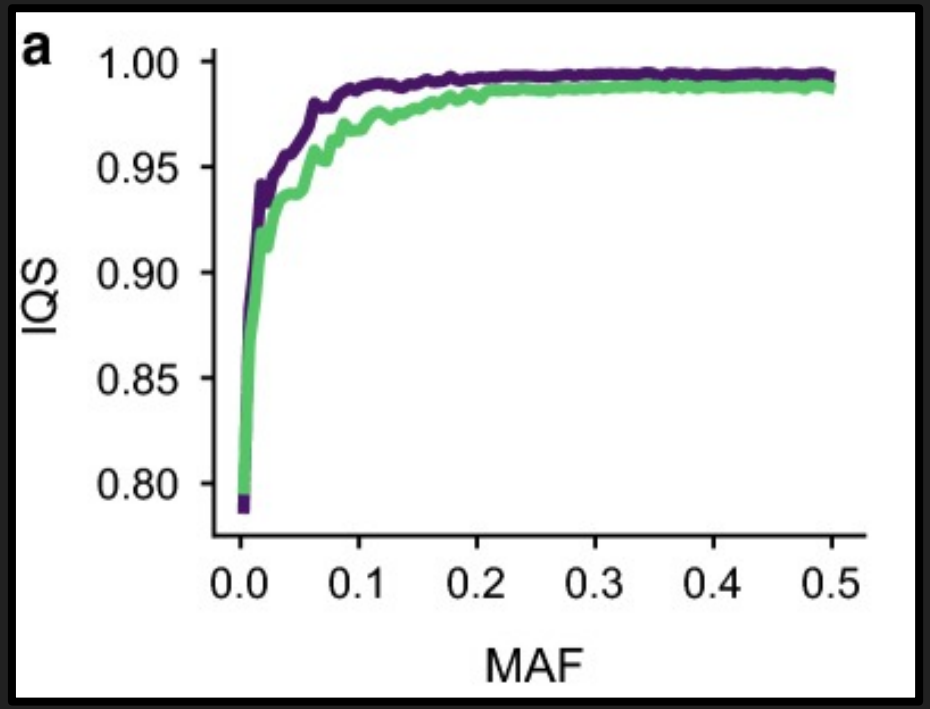
- **A large reference set of haplotypes**
  - High-coverage resequenced haplotypes
  - Representative of target population haplotypes
- High-quality reference genome
  - Physical positions matter
- Recombination map

Marchini et al. 2010

How accurate is imputation? It depends, largely on allele frequency!

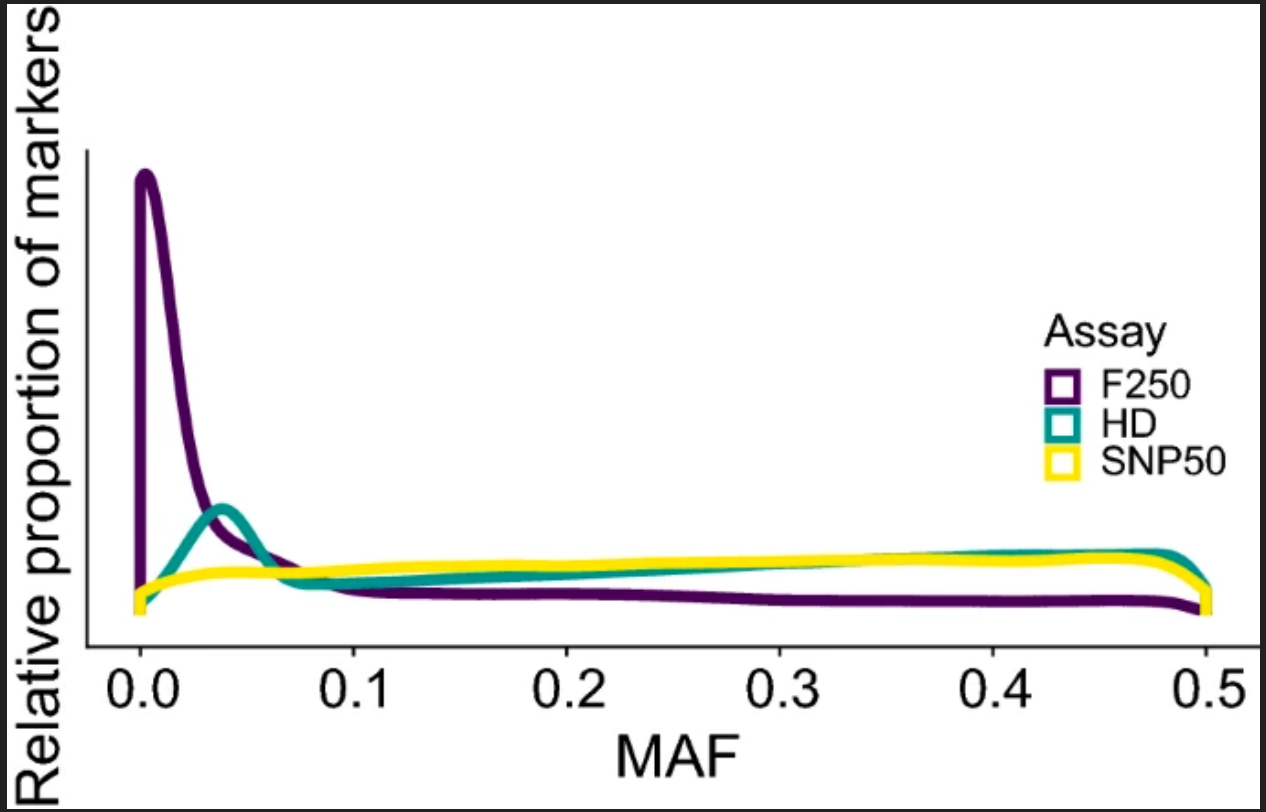


Snelling et al. 2020



Rowan et al. 2019

Imputation opportunity & challenge: Rare variation



Imputation will only impute what it “sees” in a reference panel

Representation matters!

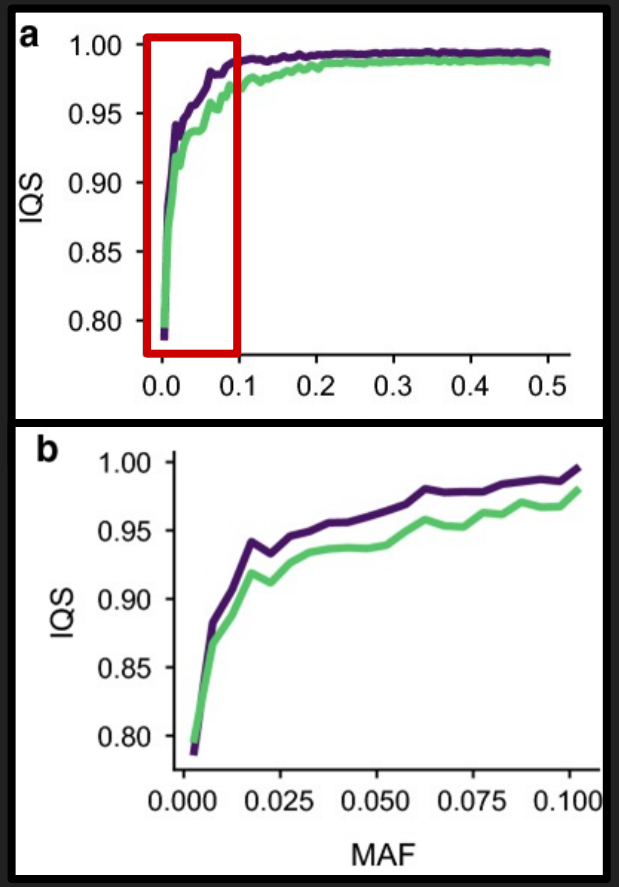
How to build a reference panel?

Breed-specific

Multi-breed

Admixed populations will benefit from a multi-breed reference

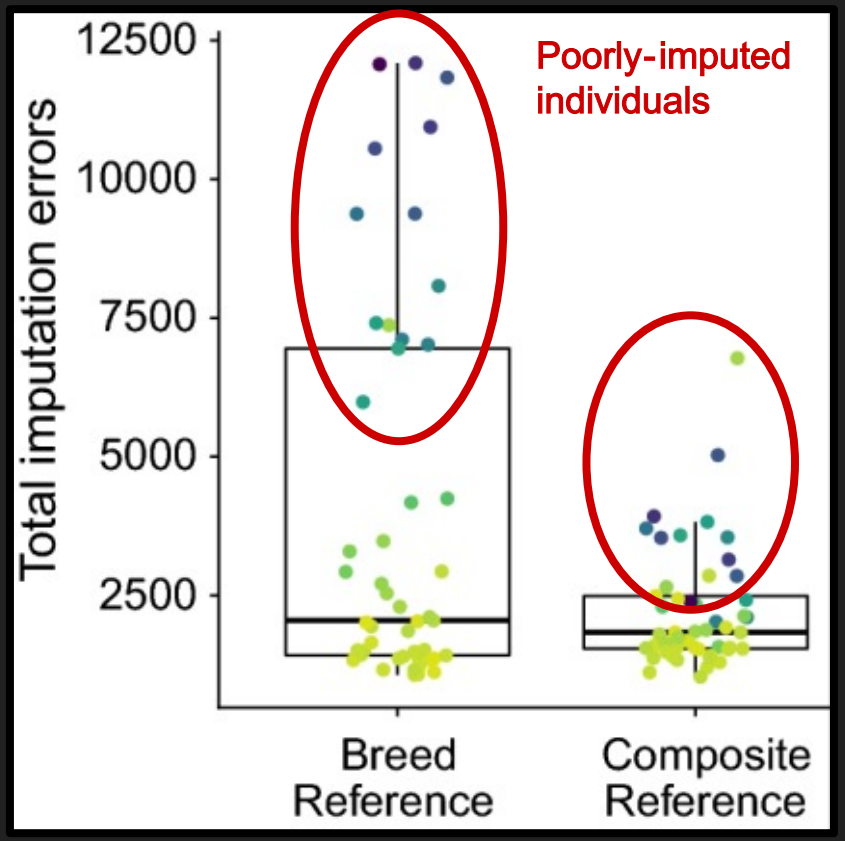
- Admixed populations need representation across diversity of individuals
- Labelled population ≠ Actual population
- Draw on haplotype diversity from other population in imputation reference
- **Using multi-population reference significantly improves per-SNP and per-individual imputation accuracy across samples!**



Rowan et al. 2019

Imputation is just pattern matching!

n =50
Gelbvieh

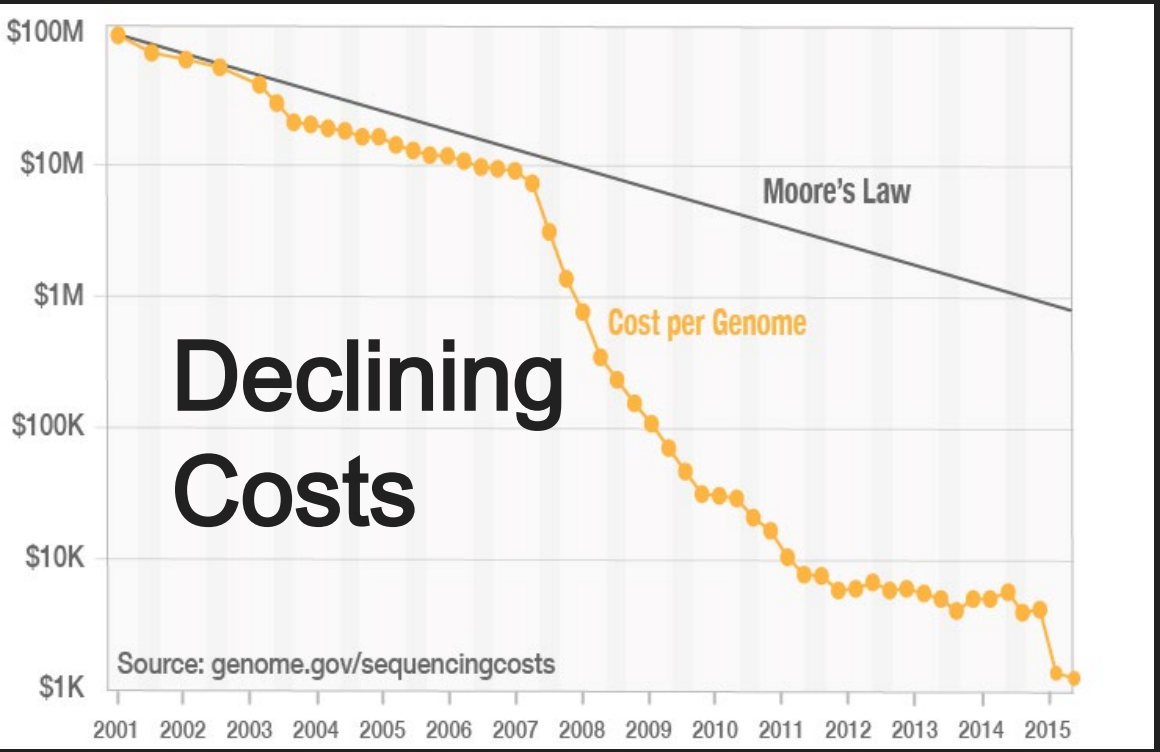


850K Chip Imputation

Breed	Mean	Min	Max
Gelbvieh	0.998	0.994	0.999
Hereford	0.997	0.991	0.999
Holstein	0.997	0.995	0.998
Simmental	0.996	0.984	0.999
Angus	0.995	0.959	0.999
Jersey	0.995	0.991	0.997
Limousin	0.989	0.930	0.996
Nelore	0.981	0.977	0.984
Brahman	0.941	0.932	0.961
Gir	0.903	0.869	0.948
Romagnola	0.874	0.855	0.896
N'Dama	0.763	0.747	0.803

Rowan et al. 2019

Advantages of Low-Pass Sequencing



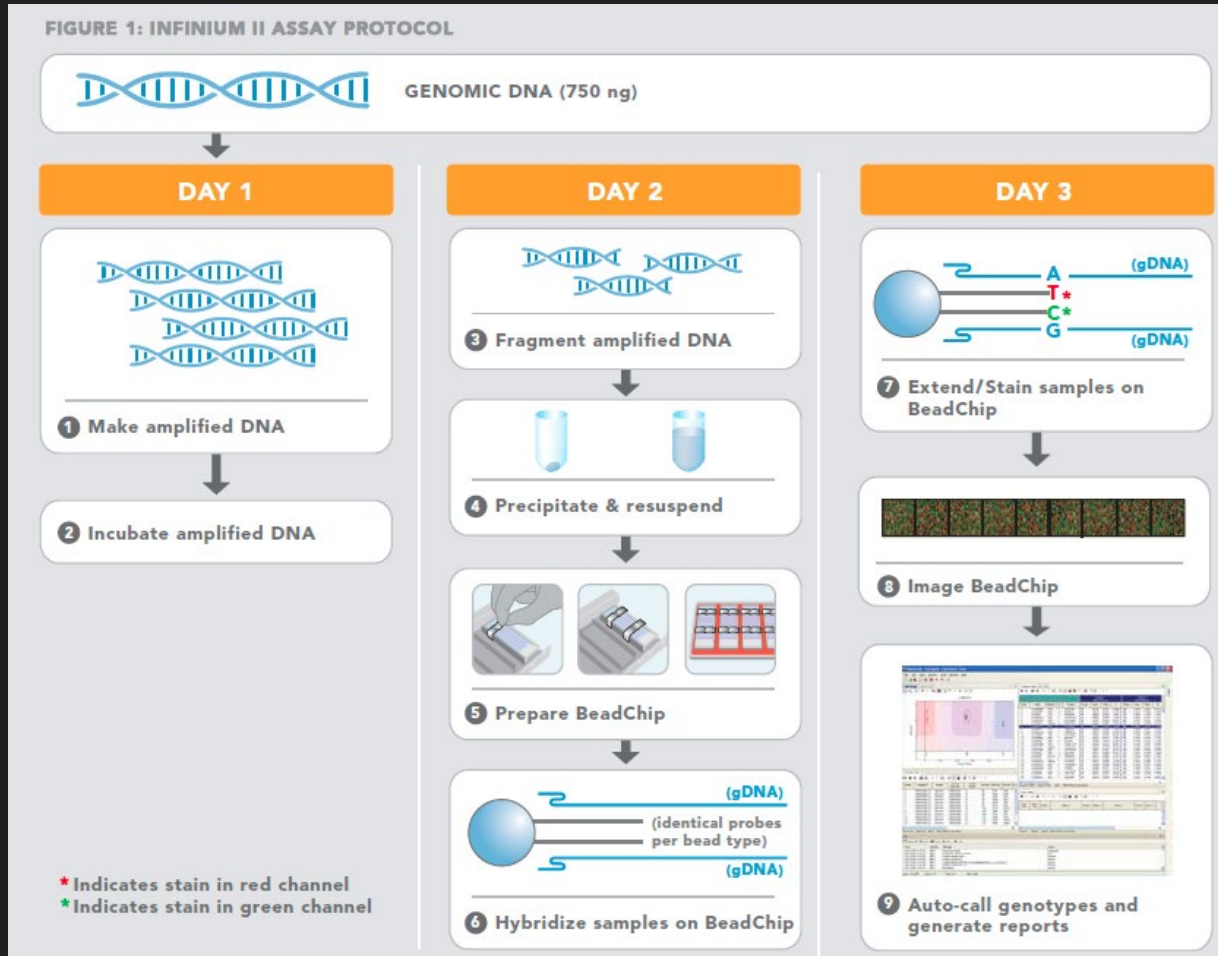
- Potential for further cost reduction
- Rare variation
- No need for chip redesign or updates
- SNP Discovery
- CNV detection

Comparing and Contrasting Chip & Low-Pass

Sample Processing: Chip vs. Sequencing

Chip Processes:

1) Amplify DNA
2) Fragment DNA
3) Precipitate
4) Put on chip
5) Image chip

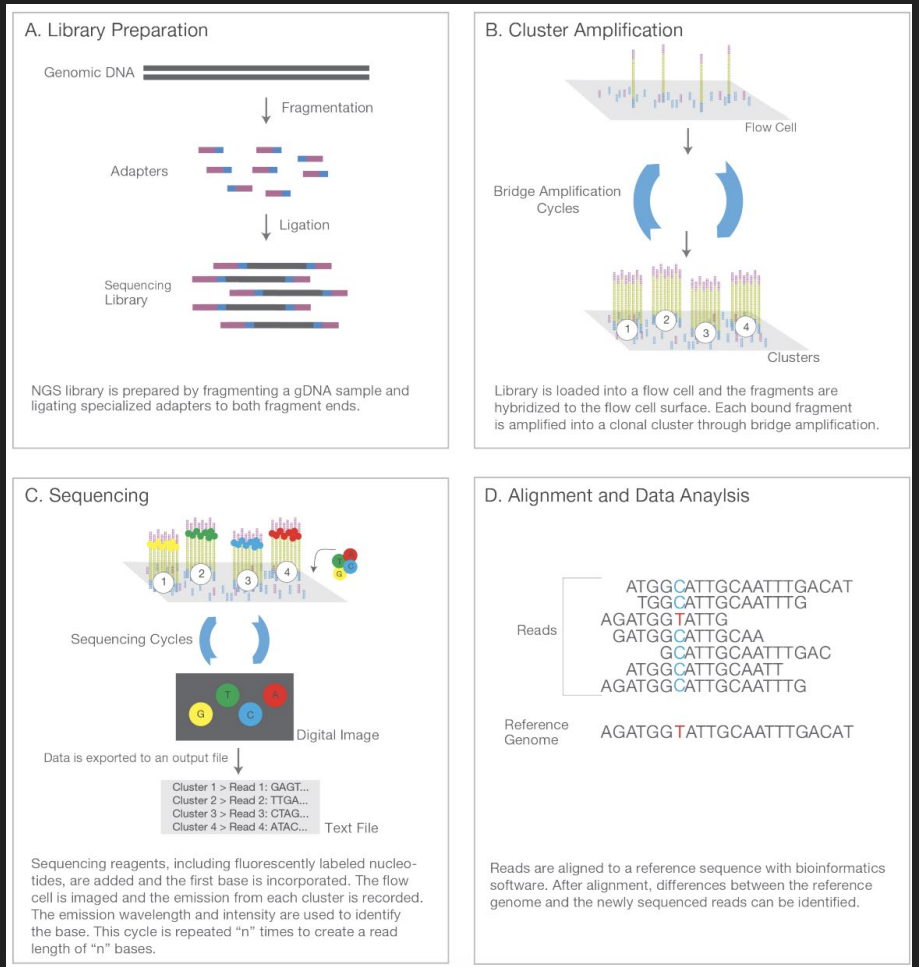


UTAGRESEARCH
INSTITUTE OF AGRICULTURE
THE UNIVERSITY OF TENNESSEE

Sample Processing: Chip vs. Sequencing

Sequencing Processes:

1) Fragment DNA
2) Add & ligate adapters
3) Library Preparation (labor intensive + technically difficult)
4) Cluster amplification
5) Sequencing
6) Data processing & Bioinformatics



Data scale in low-pass sequence data

Raw Data: FASTQ file

Every base pair sequenced (large file)

```
>NODE_1_length_1401924_cov_73.350607
AAAGTCTCCTCACGCAAACCCTTCTTGGTTGGGTCACGGAAGATAACCTTGGCCAAAGGA
GCACCTCTACCAGGGTCGTGAATGATCTGCTTAACATATCCACGGATCTGTCCTTCACGT
TCGGAAGAGTCCAAAGCTCTAAGCTTGGCAGCTCCCTTTCTCAGACGAGTGTGAGCGGTG
AAGATGGAACCAGCACCCTTTCTTTGAGCACGAATAACTCTACCCATGTTGTGTTAATGT
TTCTTGCTGTAAGAGGACTTGAAATTTTTTATTGGTTTTTTTTTTTGGGGAGTATGAGGG
TTCTTGTGTTCGCGGGTTAACCCTAGTGCTGGTCACGTGCCTTATTGGGCAAGCTGTGTG
AGGTATCATAAGGTGGTAGTTGAAAGGTACCTTATGGAAGACTTCGTTAGGAAGGTGTCT
GTATGATTAGAGTGGCGTAGGGTGAATGATTTAATTCTTCTTCG......
>NODE_2_length_1392447_cov_73.757244
TGTACCTACTAGCTTGAATAACAAGTTTATCTTTGAGGAAACTTGGTTTCAGAGACAAAA
GTTAAGTACTTGACATTGGGAGCTAAGCTTCTGCATTGCTCCTCTCTGAAAGACTTCAAG
ACTTACCATTTGGAAACAAGTGAGTTTGCATTAGTATCAAAAGGTTGGTATGATATAGGT
```

Imputation

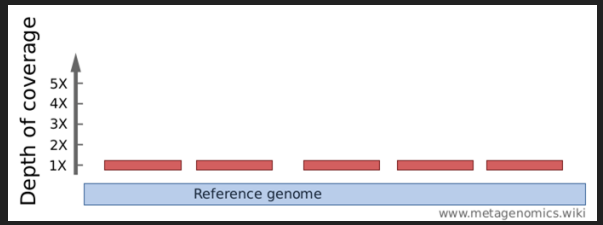
Processed Data: VCF File

ONLY called & imputed SNPs w/ metadata (much smaller file)– Maybe ~30 M SNPs

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:
```

Raw Data Scale: Chip vs. Sequencing

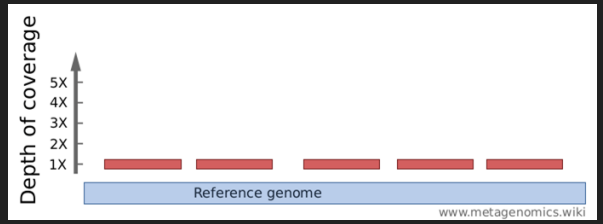
100K Array Final Report = 0.003 GB



1 X FASTQ = 2.9 GB

Processed Data Scale: Chip vs. Sequencing

25,000 animals on 50K Array BCF= 17.7 MB



Imputed BCF w/ 30M variants = 10.6 GB

Storage Questions

1) What do we store?
    a) FASTQs (raw data) BIG files,
    b) Imputed sequence – Takes time & resources to impute
    c) “Core” SNP set(s)
2) How long do we store it?
3) Do we re-impute? How Often?
4) How do we integrate with chip data?

How do we use this “extra” data?

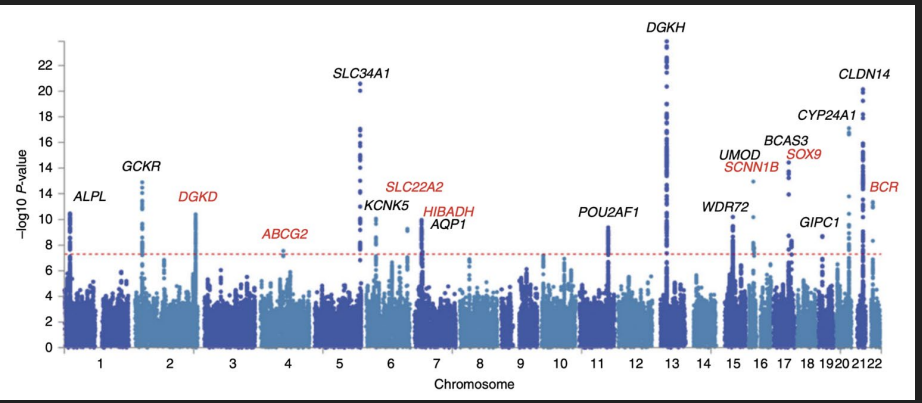
1) Same as we always have – Extract same markers that we use with SNP chips
   a) Immediately allows low-pass to be congruent with existing evaluations
   b) No extra value extracted from low-pass data

2) Throw more variants in the mix – If 50,000 is good, 30 million is better
   a) Limited evidence that this actually helps in ssGBLUP settings
   b) Var_G captured by 50K is largely sufficient for genomic prediction

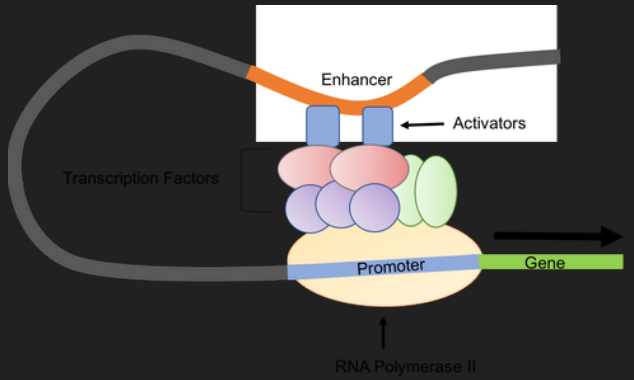
3) Prioritize variants – Find biologically important variants (e.g., causal/functional) use these to (hopefully) improve predictions
   a) Varying results – May not improve prediction accuracy but may improve portability
   b) Requires that modeling can handle non-normal variant effects (e.g., BayesRC, etc.)

Improving genomic predictions with biological knowledge

Variant discovery & prioritization



Trait associations



Functional classification

UTAGRESEARCH INSTITUTE OF AGRICULTURE THE UNIVERSITY OF TENNESSEE

UTCVM INSTITUTE OF AGRICULTURE THE UNIVERSITY OF TENNESSEE

Low-pass sequencing and imputation is the next evolution of cattle genotyping technologies

Quality imputation relies on representative reference panels

Genomic prediction machinery will need to adapt to take full advantage of low-pass imputed genotypes

Reach out with questions!

trowan@utk.edu
(865) 974-3190
@TroyNRowan