Sequencing Strategies to Enhance the Next Generation of Genetic Evaluations

Troy Rowan
BIF Annual Symposium
Advances in Selection Decisions Breakout
July 4, 2023
Our low-pass sequencing future

Potential for further cost reduction

Rare variation

No need for chip redesign or updates

SNP Discovery

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

\[ \text{Coverage} = \frac{n \text{Reads} \times \text{len(read)}}{\text{Genome Size}} \]
Low-Pass sequencing
Low-Pass Imputation

Rubinacci et al. 2021
The Power of Imputation

Rowan et al. 2021
What does accurate imputation need?

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

Marchini et al. 2010
Imputation opportunity & challenge: Rare variation
We can’t impute what we don’t observe

As such, rare variation is a challenge

Rowan et al., 2019
The big questions:

- Who do we sequence?
- How deep do we sequence?
- How often do we update?
  - Reference
  - Imputed samples in evaluation
The big questions:

- Who do we sequence?
- How deep do we sequence?
- How often do we update?
  - Reference
  - Imputed samples in evaluation
Before we talk about sequencing for imputation...

We should be sequencing ALL sires with even moderate levels of AI usage!
Why sequence all AI sires?

- “Insurance Policy”: Accelerate abnormality mapping and management
- Proactive monitoring of *de novo* mutations
- Lethal haplotype mapping at sequence resolution
- Enable haplotype-aware analyses
- Increased imputation qualities
- Current costs make this tenable!
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!

![Total imputation errors](chart.png)

<table>
<thead>
<tr>
<th>Breed</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gelbvieh</td>
<td>0.998</td>
<td>0.994</td>
<td>0.999</td>
</tr>
<tr>
<td>Hereford</td>
<td>0.997</td>
<td>0.991</td>
<td>0.999</td>
</tr>
<tr>
<td>Holstein</td>
<td>0.997</td>
<td>0.995</td>
<td>0.998</td>
</tr>
<tr>
<td>Simmental</td>
<td>0.996</td>
<td>0.984</td>
<td>0.999</td>
</tr>
<tr>
<td>Angus</td>
<td>0.995</td>
<td>0.959</td>
<td>0.999</td>
</tr>
<tr>
<td>Jersey</td>
<td>0.995</td>
<td>0.981</td>
<td>0.997</td>
</tr>
<tr>
<td>Limousin</td>
<td>0.989</td>
<td>0.930</td>
<td>0.996</td>
</tr>
<tr>
<td>Nelore</td>
<td>0.981</td>
<td>0.977</td>
<td>0.984</td>
</tr>
<tr>
<td>Brahman</td>
<td>0.941</td>
<td>0.932</td>
<td>0.961</td>
</tr>
<tr>
<td>Gir</td>
<td>0.903</td>
<td>0.869</td>
<td>0.948</td>
</tr>
<tr>
<td>Romagnola</td>
<td>0.874</td>
<td>0.855</td>
<td>0.896</td>
</tr>
<tr>
<td>N'Dama</td>
<td>0.783</td>
<td>0.743</td>
<td>0.803</td>
</tr>
</tbody>
</table>

Rowan et al. 2019

850K Chip Imputation

n =50 Gelbvieh
Ascertainment Bias: Human Example

Imputation $r^2$ in Africans across technologies

Imputation $r^2$ in Europeans across technologies

- Gencove 0.5x
- Gencove 1x
- Illumina GSA
We have to move past sequencing only most common animals

Case Study from ASA:

- Top 150 sires cluster very closely together in PCA of full genomic dataset
- Sequencing only heavily-used bulls will sample only a small portion of haplotype-space
So how might we select sires to sequence?

- Use chip genotype data!
- Iteratively search for sequencing candidates
The big questions:

- Who do we sequence?
- How deep do we sequence?
- How often do we update?
  - Reference
  - Imputed samples in evaluation
The quality of information coming out cannot be better than the quality of information that went in.

GIGO is used in IT and mathematics

Garbage In, Garbage Out
Sequencing depth = greater genotype confidence

>10X coverage resulted in substantially more non-reference discordances (i.e. wrongly called genotypes)

% Bob Schnabel
We should generate at least 15X genomes for reference individuals!

Element & Illumina are both generating sequence at $2/GB

Miniaturized library preparations

Marginal cost increase between 5X and 15X is between $60-$150
Coverage Impacts on Genotype Call Quality

Mendelian Errors Duos Chr25

H = > 20X
M = 10-20X
L = < 10X

% Bob Schnabel
Sequencing is just the first step… Phasing matters!

Genotypes = 0, 1, 2

2N haplotypes = 0, 1

Unphased Variation collapsed into a single mixed sequence

Phased Each haplotype assembled separately

Our pedigrees can do much of this phasing for us (>60%)

Read-backed phasing

Various HMM algorithms (SHAPEIT5)
And so does continually empirically evaluating accuracy

Testing Individuals: True high-coverage calls

Downsampled to low-coverage reads or chip genotypes

Evaluate imputation accuracy: On per-SNP and per-individual basis
Other things to consider:

- Pangenomes
- Moderate-coverage sequencing
- Storage of genomes and imputed genotypes
The Million Dollar Question: How do we use this information to improve predictions?
Haplotype reference panels must be representative of the target populations.

Multi-breed > Within-breed

High-quality reference sequences should be at least 10X to be optimally useful.

It is important to regularly evaluate imputation accuracy for both individuals and SNPs.

Reach out with questions!

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