Combined Arms: A Full Spectrum Approach to Variation Detection in Livestock

Genetic Variation

How genomes change over time
- Single nucleotide variations – SNP (human millions of variants)
- Indels – Insertions/Deletions (1 bp – 100 bp)
- Mobile Elements – SINE, LINE Transposition (100bp - 6 kb)
- Genomic structural variation (1 kb – 5 Mb)
  - Large-scale Insertions/Deletions (Copy Number Variation: CNV)
  - Segmental Duplications (> 1kb, > 90% sequence similarity)
  - Chromosomal Inversions, Translocations, Fusions.

NGS Variant detection is not straightforward
- Align billions of short reads per animal
- Sequence Errors
- Misaligned reads
- Need different methods

Combined arms
- Use all resources
- Smart merger
- Pipeline implementation
  - Alignment is slow
  - Efficiency with strategies
  - Existing tools hard-coded for human

A combined variant detection pipeline

GATK “Best Practices” Workflow

- Massively parallel
- Self-cleaning
- Config file based

Align:
- BWA
- MrsFAST

Call:
- GATK

Final Data:
- SNPs
- INDELS

BWA Alignment
BAM Processing
GATK Walkers
Indel Detection
Indel Realigner
Filtration
Genotyper
Genetic Feature Annotation
A combined variant detection pipeline

- Massively parallel
- Self-cleaning
- Config file based

Align: 
- BWA
- MrsFAST

Call: 
- GATK

Final Data: 
- SNPs
- INDELs

Three NGS Methods

CNV detection by NGS

Reference Genome

Read Pair (RP)

Concordant
Deletion
Insertion
Inversion

Read Depth (RD)

Split Read (SR)

Average/Baseline
Duplication
Deletion

An Ideal Merger Situation

Precision Aware Merger

- Remove dups; bad maps
- RD: Least precise; surrounding windows
- RP: Fairly precise; within outer coords
- SR: Most precise; bp around breakpoint

Problems with alignment give false signals

Low GC%

High GC%
Simple Repeats

High GC%
Simple Repeats

Low GC%

Low GC%

Confounding alignments confuse the signal.

Poor quality reads overlap region

Shorter split read mappings are incorrect
Our “100 Bulls” dataset

<table>
<thead>
<tr>
<th>Number of Animals</th>
<th>Number of Breeds</th>
<th>Total Gigabases</th>
</tr>
</thead>
<tbody>
<tr>
<td>107</td>
<td>8</td>
<td>~ 4300</td>
</tr>
</tbody>
</table>

71 Low Coverage (~4-5X) individuals: 2000 Gb
36 High Coverage (7-30X) individuals: 2300 Gb

Results: nearly all genetic variants in a genome are interrogated

<table>
<thead>
<tr>
<th>Method</th>
<th>Count</th>
<th>Base pairs (Mb)</th>
<th>Genome percentage</th>
<th>Average length</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP calls</td>
<td>1,989,637</td>
<td>1.9</td>
<td>0.07%</td>
<td>1 bp</td>
</tr>
<tr>
<td>RD calls</td>
<td>1057</td>
<td>40.9</td>
<td>1.5%</td>
<td>40 kb</td>
</tr>
<tr>
<td>RP calls</td>
<td>570</td>
<td>0.2</td>
<td>&lt; 0.01%</td>
<td>350 bp</td>
</tr>
<tr>
<td>SR calls</td>
<td>2730</td>
<td>1.1</td>
<td>0.04%</td>
<td>400 bp</td>
</tr>
<tr>
<td>Total Variants</td>
<td>1,993,994</td>
<td>44.1</td>
<td>1.6%</td>
<td></td>
</tr>
</tbody>
</table>

Multiple CNV methods are complimentary

PAM resolves contradictions

Potential functional impacts of CNVs

SNP functional impacts are annotated

<table>
<thead>
<tr>
<th>Effect</th>
<th>Note</th>
<th>Count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXON</td>
<td>Variant hits an exon</td>
<td>41,129</td>
<td>5.91%</td>
</tr>
<tr>
<td>NON_SYNTHONM</td>
<td>Variant causes a codon that produces a different amino acid</td>
<td>17,299</td>
<td>2.79%</td>
</tr>
<tr>
<td>SYNTHONM</td>
<td>Variant causes a codon that produces the same amino acid</td>
<td>22,298</td>
<td>3.03%</td>
</tr>
<tr>
<td>STOP_GAINED</td>
<td>Variant causes STOP codon</td>
<td>160</td>
<td>0.02%</td>
</tr>
<tr>
<td>STOP_LOST</td>
<td>Variant causes stop codon to be mutated into a non-stop codon</td>
<td>10</td>
<td>&lt; 0.001%</td>
</tr>
<tr>
<td>SPLICE_SITE</td>
<td>Variant hits a splicing site</td>
<td>265</td>
<td>0.012%</td>
</tr>
</tbody>
</table>

Lipid metabolism
Lipid Transport
Conclusions

- Efficiency via pipeline
- Big picture from multiple methods
  - Complimentary
  - Precision aware merger
- Goal: release data and pipeline to public

Acknowledgements

- George Liu's lab
  - Lingyang Xu
  - Yali Hou
  - Ruben Anderson
  - Alexandre Dmitrov
- USDA BFGL
  - Tad Sonstegard
  - Curtis Van Tassell
  - Steven Schroeder
  - Erin Connor
  - Heather Huson
  - Matthew McClure
- Evan Eichler's lab at the University of Washington
- Jeremy Taylor and Robert Schnabel's lab
- Jose Fernando Garcia from UNESP of Brazil
- USDA AIPL
  - George Wiggans
  - Paul VanRaden
  - John Cole
  - Anna Hutchinson
  - Tabatha Cooper
  - Kay Megenigal
  - Quin Hull
  - Lillian Backeller
  - Suzanne Hubbard
  - Other Support Staff

- Funded by National Research Initiative (NRI) Grant No. 2007-35205-17869 and 2011-67015-30183 from USDA-NIFA

Combining multiple methods of detection

Aligned Sequence → Calculate PE Stats

Broad's GATK → mrsFAST + WSSD → VariationHunter → SplitRead

SNPs and INDELs → RD CNV calls → RP CNV calls → SR CNV calls

Indel realignment fixes common errors

Figure from: DePisto et al. 2011. Nature Genetics. 43, 491-498

CNV detection by NGS

A. Read Pair (RP) or PEM

Concordant Deletions Insertions Inversions Translocations

Reference

B. Read Depth (RD)

Reference

C. Split read (SR)

Reference

D. Local assembly

Reference

• Chimeric read

ACGTG
GGTACATACGA
GACAGATGGG
AACCACACA
GAGAGG
GGAGATAGAG

ACGTG
GGTAGATAGA
GACAGATGGG
AACCACACA
GAGAGG
GGAGATAGAG