Phasing haplotypes in rabbit using long reads technology
Julie Demars

To cite this version:
Julie Demars. Phasing haplotypes in rabbit using long reads technology. Journées Plateforme Get-Plage, Nov 2017, Toulouse, France. hal-02784960

HAL Id: hal-02784960
https://hal.inrae.fr/hal-02784960
Submitted on 4 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License
Phasing haplotypes in rabbit using long reads technology

Julie Demars
GenPhySE-Genetics, Physiology and Livestock Systems

Long reads: Dream or Reality?
28th of November 2017

http://get.genotoul.fr
get@genotoul.fr
@get_genotoul
Genome phasing

A/A  T/C  G/C  G/A

A  T  G  G
A  C  C  A

Fundamental aspect of genetics that is relevant in many applied problems
Interest of genome phasing

• Allele-specific expression or methylation
  Genomic imprinting

• Highly heterozygous regions
  HLA genes

• Population genetics and genome-wide association studies
  Polygenic traits, Allelic heterogeneity
Detecting local haplotype sharing and haplotype association.

Xu H\textsuperscript{1}, Guan Y\textsuperscript{2}.
General principle of genome phasing

Unphased genome:
- A/A
- T/C
- G/C
- G/A

Sequence reads:
- C
- C
- A
- T
- T
- G
- C
- A

Phased results:
- A
- T
- G
- G
- A
- C
- C
- A

http://get.genotoul.fr
Genome phasing: methods and developments

• **Scale of datasets**
  From SNP beadchips to whole genome

• **Statistical models and computational approaches**
  Time consuming: balance quality of phasing/computational cost

• **Laboratory-based experimental methods**
  Long reads produce virtual multi-kilobases reads on regular sequencers
Evaluation of rabbit genome phasing

1001 line
♂
15405156
♀
15404125

1777 line
♀
16005189
♀
16005195

Rabbit Reference Genome (Oryctolagus cuniculus)

<table>
<thead>
<tr>
<th>Version</th>
<th>OryCun2 (GCA_000003625.1), 1st assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference rabbit</td>
<td>New Zealand</td>
</tr>
<tr>
<td>Sequencing Depth</td>
<td>7X</td>
</tr>
<tr>
<td>Chromosomes</td>
<td>21</td>
</tr>
<tr>
<td>Size (Mb)</td>
<td>2247.75</td>
</tr>
<tr>
<td>Genes</td>
<td>20999</td>
</tr>
<tr>
<td>Scaffolds</td>
<td>3318</td>
</tr>
<tr>
<td>Size (Mb)</td>
<td>489.69</td>
</tr>
<tr>
<td>Genes</td>
<td>8099</td>
</tr>
</tbody>
</table>
## Summary of phasing results

<table>
<thead>
<tr>
<th>Phased WGS</th>
<th>15405156</th>
<th>15404125</th>
<th>16005189</th>
<th>16005195</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>15405156</td>
<td>15404125</td>
<td>16005189</td>
<td>16005195</td>
</tr>
<tr>
<td>Female</td>
<td>15405156</td>
<td>15404125</td>
<td>16005189</td>
<td>16005195</td>
</tr>
<tr>
<td>Father</td>
<td>15405156</td>
<td>15404125</td>
<td>16005189</td>
<td>16005195</td>
</tr>
<tr>
<td>Mother</td>
<td>15405156</td>
<td>15404125</td>
<td>16005189</td>
<td>16005195</td>
</tr>
<tr>
<td>Offspring</td>
<td>15405156</td>
<td>15404125</td>
<td>16005189</td>
<td>16005195</td>
</tr>
<tr>
<td>Offspring</td>
<td>15405156</td>
<td>15404125</td>
<td>16005189</td>
<td>16005195</td>
</tr>
</tbody>
</table>

### Fold coverage

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Female</th>
<th>Female</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>32</td>
<td>28</td>
<td>26</td>
<td>28</td>
<td></td>
</tr>
</tbody>
</table>

### % aligned

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Female</th>
<th>Female</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>81</td>
<td>80</td>
<td>80</td>
<td>77</td>
<td></td>
</tr>
</tbody>
</table>

### % duplication

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Female</th>
<th>Female</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4,42</td>
<td>6,09</td>
<td>6,30</td>
<td>6,66</td>
<td></td>
</tr>
</tbody>
</table>

### Relative genomic equivalents per partition

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Female</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absolute genomic equivalents</td>
<td>0,362</td>
<td>0,321</td>
<td>0,342</td>
<td>0,508</td>
</tr>
<tr>
<td>Relative genomic equivalents</td>
<td>0,362</td>
<td>0,321</td>
<td>0,342</td>
<td>0,508</td>
</tr>
<tr>
<td>Absolute genomic equivalents</td>
<td>0,362</td>
<td>0,321</td>
<td>0,342</td>
<td>0,508</td>
</tr>
<tr>
<td>Relative genomic equivalents</td>
<td>0,362</td>
<td>0,321</td>
<td>0,342</td>
<td>0,508</td>
</tr>
</tbody>
</table>

### Number of molecules (millions)

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Female</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of molecules</td>
<td>1,59</td>
<td>1,59</td>
<td>1,40</td>
<td>1,34</td>
</tr>
</tbody>
</table>

### Length-weighted mean molecule length (kb)

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Female</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length-weighted mean molecule</td>
<td>63,17(±41,95)</td>
<td>73,90(±48,02)</td>
<td>45,73(±30,30)</td>
<td>54,06(±43,98)</td>
</tr>
<tr>
<td>Length-weighted mean molecule</td>
<td>63,17(±41,95)</td>
<td>73,90(±48,02)</td>
<td>45,73(±30,30)</td>
<td>54,06(±43,98)</td>
</tr>
</tbody>
</table>

### % SNPs phased

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Female</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>% SNPs phased</td>
<td>97</td>
<td>97</td>
<td>98</td>
<td>97</td>
</tr>
</tbody>
</table>

### N50 phase block (kb)

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Female</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>N50 phase block</td>
<td>875,49</td>
<td>1086,73</td>
<td>609,92</td>
<td>838,05</td>
</tr>
</tbody>
</table>

### Longest phase block (kb)

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Female</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Longest phase block</td>
<td>4 417,55</td>
<td>9 930,61</td>
<td>3 725,93</td>
<td>4 599,55</td>
</tr>
</tbody>
</table>

### Long structural variants (number)

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Female</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long structural variants</td>
<td>162</td>
<td>183</td>
<td>197</td>
<td>155</td>
</tr>
</tbody>
</table>

### Short structural variants (number)

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>Female</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short structural variants</td>
<td>49 846</td>
<td>45 901</td>
<td>46 265</td>
<td>47 723</td>
</tr>
</tbody>
</table>
Phase the « full » spectrum of called variants

**a**

- Graphs showing total bases and fraction of genome phased against molecule and phase block length.
- Key variants: 15405156, 15404125, 16005189, 16005195.

**b**

- Bar charts or sequence analysis showing variants 15405156, 15404125, 16005189, 16005195.

**c**

- Diagram of chromosome KIT, chr 15: 93,911,537-93,952,727 – 41kb.
- Haplotypes 1 and 2 visualized.
- Molecule length and total bases (x10^9) indicated.

Additional notes:
- a: Haplotype 1 and Haplotype 1
- b: Haplotype 1 and Haplotype 2
- c: Haplotype 1 and Haplotype 1

http://get.genotoul.fr
Phase structural variants: large deletion

Homozygous deletion

Genome assembly

Reference

Homozygous 50kb deletion

chr7
Phase structural variants: large deletion

Homozygous deletion
15404125 (mother)

chr7

Genome assembly

http://get.genotoul.fr
Phase structural variants: large deletion

Heterozygous deletion
15405156 (father)

Genome assembly
Barcode overlap
Linked reads
Phase structural variants: tandem duplication

Duplication

chr8

References

Homozygous 50kb duplication

Genome assembly

http://get.genotoul.fr
Heterozygous duplication
15404125 (mother)

Genome assembly
Barcode overlap
Linked reads

Genomic position
Density
Barcode count

Phase structural variants: tandem duplication
Rescue repetitive regions

• Problem to assign reads in repetitive elements

• Information from 10X barcodes rescue unmapped reads
Recover variants in repetitive regions
Linked reads and the 10X technology

• **Resolve the genome into long (megabases) phase blocks**
  Phase the “full” spectrum of called variants

• **Identify structural variants and breakpoints**
  Insertions, deletions, duplications, translocations…

• **Rescue variants in inaccessible parts of the genome**
  Confidently map reads in repetitive regions

• **Improve genome assembly**
Improve genome assembly: gap resolution

- **Real or fake gap in the genome assembly?**
  Analysis of barcodes of linked reads
Improve genome assembly: inversion identification

Complex rearrangement including inversion

Genome assembly

IGV window (pair orientation)
Acknowledgments

- Genotoul GeT-PlaGe
  - Olivier Bouchez
  - Cécile Donnadieu
  - Pauline Heuillard
  - Claire Kuchly
  - Sophie Valière

- GenPhySE
  - Isabelle Hochu
  - Sophie Leroux
  - Juliette Riquet
  - Julien Ruesche

- PECTOUL
  - Patrick Aymard
  - Elodie Balmisse
  - Virginie Helies