Storing and analysing a million genomes on a desktop computer

Gregor Gorjanc, Jana Obsteter, Gabriela Mafra Fortuna, Roger Ros-Freixedes, Martin Johnsson, Ivan Pocrnic

InterBull & EAAP
Lyon, 2023-03-24
Where have we been & where are we going?

• Started career with pedigree-based mixed models

• Rode on the excitement of introducing genomic selection
  – 2010s $\sim 10^3$ individuals
  – 2015s $\sim 10^5$ individuals
  – 2020s $\sim 10^6$ individuals
  – 2030s $\sim 10^9$ individuals???

• Contributed to the whole-genome sequencing “craze”
  – 2015s $\sim 10^3$ individuals
  – 2020s $\sim 10^{4-6}$ individuals
  – 2030s $\sim 10^{6-9}$ individuals???
Handling MEGA-SCALE through data generation process

Kelleher et al. (2019, Nature Genetics; …)
Tracking chromosome segments

n1
n2
n3
n4
n5
n6
n7
n8
Ancestral recombination graph & Local trees
Tree sequence – Nodes

<table>
<thead>
<tr>
<th>Node</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>n1</td>
<td>0</td>
</tr>
<tr>
<td>n2</td>
<td>0</td>
</tr>
<tr>
<td>n3</td>
<td>1</td>
</tr>
<tr>
<td>n4</td>
<td>1</td>
</tr>
<tr>
<td>n5</td>
<td>2</td>
</tr>
<tr>
<td>n6</td>
<td>2</td>
</tr>
<tr>
<td>n7</td>
<td>2</td>
</tr>
<tr>
<td>n8</td>
<td>2</td>
</tr>
</tbody>
</table>
Tree sequence – Nodes & Edges (pedigree)

Node Time
n1 0
n2 0
n3 1
n4 1
n5 2
n6 2
n7 2
n8 2

Edge Desc Anc Start Stop
e1 n2 n7 0 100
e6 n4 n8 0 100
### Tree sequence – Nodes & Edges (coalescent)

#### Nodes & Times

<table>
<thead>
<tr>
<th>Node</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>n9</td>
<td>100</td>
</tr>
<tr>
<td>n10</td>
<td>150</td>
</tr>
<tr>
<td>n11</td>
<td>500</td>
</tr>
</tbody>
</table>

#### Edges & Descendants

<table>
<thead>
<tr>
<th>Edge</th>
<th>Desc</th>
<th>Anc</th>
<th>Start</th>
<th>Stop</th>
</tr>
</thead>
<tbody>
<tr>
<td>e7</td>
<td>n7</td>
<td>n9</td>
<td>0</td>
<td>27</td>
</tr>
<tr>
<td>e8</td>
<td>n8</td>
<td>n9</td>
<td>0</td>
<td>27</td>
</tr>
<tr>
<td>e9</td>
<td>n5</td>
<td>n10</td>
<td>0</td>
<td>18</td>
</tr>
<tr>
<td>e10</td>
<td>n6</td>
<td>n10</td>
<td>0</td>
<td>18</td>
</tr>
<tr>
<td>e11</td>
<td>n9</td>
<td>n11</td>
<td>0</td>
<td>27</td>
</tr>
<tr>
<td>e12</td>
<td>n10</td>
<td>n11</td>
<td>0</td>
<td>18</td>
</tr>
</tbody>
</table>
Computational power of tree sequences

**STORAGE**

Kelleher et al. (2019, Nature Genetics; …)

**COMPUTE**

Ralph et al. (2020, Genetics)
Roslin-Genus/PIC ~1 million pig genomes project

- **Pedigree & SNP array data**
  - 9 lines with a total of ~450K pigs
  - ~15K-50K markers

- **Whole-genome sequence**
  - ~8K pigs (a mix of ~1x and ~30x)
  - ~46M variants passed quality control across lines

- **Accurate imputation of whole-genomes**
  - ~450K diploid pigs * 2 = ~900,000 haploid genomes
  - ~450K pigs * ~46M sites * 8 bytes / 2^40 = ~152 TiB of memory
    (2-bit storage → ~5TiB of memory)
Roslin-Genus/PIC ~1 million pig genomes project

- Pedigree & SNP array data
  - 9 lines with a total of ~450K pigs
  - ~15K-50K markers

- Whole-genome sequence
  - ~8K pigs (a mix of ~1x and ~30x)
  - ~46M variants passed quality control across lines

- Accurate imputation of whole-genomes
  - ~450K diploid pigs * 2 = ~900,000 haploid genomes
  - ~450K pigs * ~46M sites * 8 bytes / 2^40 = ~152 TiB of memory
    (2-bit storage → ~5 TiB of memory)

~15+ GiB vs. ~152 TiB
~99.9% “compression”!!!
Real application: 1000 Bull Genomes data

• 2,716 samples & 157 groups
  – Bos taurus
  – Bos indicus
  – Crossbred & African
  – Bos taurus coreanae (nat. Korean)
  – Bos primigenius (auroch)
  – Bos grunniens (yak)

• 29 autosomal chromosomes with ~116M variants

• ShapeIt (phase), tsinfer (infer tree sequence), & tskit (analyse)

• VCFs: HUGE SIZE → Tree sequences: DECENT SIZE → ”Compression”: LARGE → Analysis: FAST
Conclusion

• MEGA-SCALE genomic datasets are here & growing
• Tree sequence data format to the rescue!?
  – PROS
    • Succinctly encodes the inheritance process
    • Combines pedigree, coalescent, phylogenetics (gene & species trees), segregation, recombination, gene conversion, mutations, IBS, IBD, ...
    • Significant storage reduction & fast analyses
    • Novel insights & modelling
  – CONS (on-going work)
    • Novel way of thinking
    • Ancestral alleles, inference from real data, and inputs still HUGE
    • Need to develop specialised algorithms
Join us at The University of Edinburgh, Roslin Lab

4-year PhD studentship available (start in autumn 2024)!!!
for (Year in 1:10) {
    Pop = randCross2(males = Sires,
                    females = Dams,
                    nCrossovers = 750,
                    nProgeny = 100)
    Dams = selectInd(Pop,
                     nInd = 750,
                     sex = "f")
    Sires = selectInd(Pop,
                     nInd = 25,
                     sex = "m")
    for (Year in 1:10) {
        Variety = selectInd(EYT, nInd = 1)
        EYT = selectInd(AYT, nInd = 10)
        AYT = selectInd(PYT, nInd = 50)
        PYT = selectInd_HDRW, nInd = 500)
        HDRW = makeDH(F1, nDH = 100)
        Parents = c(EYT, AYT)
        F1 = randCross(Parents, nCrossores = 100)
    }
}
Storing and analysing a million genomes on a desktop computer

Gregor Gorjanc, Jana Obsteter, Gabriela Mafra Fortuna, Roger Ros-Freixedes, Martin Johnsson, Ivan Pocrnic

InterBull & EAAP
Lyon, 2023-03-24