CUMACH - A Fast GPU-based Genotype Imputation Tool

Agatha Hu
ahu@nvidia.com
Term explanation

- **Allele:**
  - one of two or more forms of a gene or a genetic locus.

- **Genotype**
  - the genetic makeup of a cell, an organism, or an individual

- **Locus (plural loci)**
  - the position of a gene on a chromosome

Figure resource: http://en.wikipedia.org/wiki/Genotype
Term explanation

- **Haplotype**
  - Combination of alleles (DNA sequences) at adjacent locations (loci) on the chromosome that are transmitted together.

- **SNP (single-nucleotide polymorphism)**
  - A DNA sequence variation occurring when a single nucleotide in the genome differs between members of a biological species or paired chromosomes in an individual.

- **Imputation**
  - Estimation of unmeasured genotypes.
Genome-wide association studies is a powerful analysis approach which have already found a lot of risk loci of complex diseases.

In a typical GWA study, there’ll be ~500,000 to ~1,000,000 variants, several hundred cases and controls.
And next?

Find common shared gene area

Analyze data from platform 1

Analyze data from platform 2

Analyze data from platform 3

If data from different platforms doesn’t overlap each other?
There are >10 million genetic variants in human genome, but even the most recent technology only provides 100,000~1,000,000 variants. Moreover, different genotyping platforms may provide different information which doesn’t overlap each other.

Genotype imputation allows geneticists to accurately evaluate the evidence for association at genetic markers that are not directly genotyped. It is useful in combining results across platforms, or estimating un-genotyped loci by existing GWAS data.
What is genotype imputation
Hidden Markov Model

- States: $X_1, X_2, \ldots, X_i$
- Transition probability: $a_{12}, a_{13}, \ldots, a_{ij}$
- Emission probability: $b_{11}, b_{12}, \ldots, b_{ij}$
- Observations: $y_1, y_2, \ldots, y_j$
Why choosing HMM for imputation?

- Easily map the variants (alleles, genotypes, haplotypes) to HMM parameters.
- Stage 1
  - One haplotype in reference: observation
  - The rest haplotypes: hidden states
- Stage 2
  - Genotypes of study samples: observation
  - Haplotype combinations: hidden states
HMM for imputation: Stage 1

- Haplotype 0 (N loci)
- Haplotype 1 (N loci)
- Haplotype 2 (N loci)
- Haplotype 3 (N loci)
- ... (continues to Haplotype M (N loci))

Reference panel

Observation

M+1 states, each state has only two phases: 0/1, as there are only two different alleles at one locus in biallelic data.
Map Stage 1 to GPU

\[ i = 0, 1, 2, 3 \ldots N, \text{ left to right} \]

\[ P(H_{obs}^i | H_j), \text{ standing for the probability where the } i\text{-th allele on observing haplotype comes from the } j\text{-th reference haplotype.} \]

**Block 0:** Haplotype 0 as observation

\[
\begin{align*}
P(H_0^0 | H_1) & \quad P(H_0^1 | H_2) & \quad P(H_0^1 | H_3) & \quad P(H_0^1 | H_4) & \quad \ldots & \quad P(H_0^1 | H_k) \\
\end{align*}
\]

**Block 1:** Haplotype 1 as observation

\[
\begin{align*}
P(H_1^0 | H_0) & \quad P(H_1^1 | H_2) & \quad P(H_1^1 | H_3) & \quad P(H_1^1 | H_4) & \quad \ldots & \quad P(H_1^1 | H_k) \\
\end{align*}
\]

**Block 2:** Haplotype 2 as observation

\[
\begin{align*}
P(H_2^0 | H_0) & \quad P(H_2^1 | H_1) & \quad P(H_2^1 | H_3) & \quad P(H_2^1 | H_4) & \quad \ldots & \quad P(H_2^1 | H_k) \\
\end{align*}
\]

\[ \ldots \ldots \]
HMM for imputation: Stage 2

Genotypes

<table>
<thead>
<tr>
<th></th>
<th>00</th>
<th>11</th>
<th>01</th>
<th>01</th>
<th>11</th>
<th>01</th>
<th>00</th>
</tr>
</thead>
</table>

H1H1 (00/01/11) → H1H2 (00/01/11) → H2H2 (00/01/11) → H2H3 (00/01/11) → H1H3 (00/01/11) → H1H1 (00/01/11)
Map Stage2 to GPU

Block 0: Study sample 0 as observation

\[
\begin{array}{c|c|c|c}
H_{00} - H_{0M} & H_{11} - H_{1M} & H_{22} - H_{2M} & \ldots & H_{MM}
\end{array}
\]

Block 1: Study sample 1 as observation

\[
\begin{array}{c|c|c|c}
H_{00} - H_{0M} & H_{11} - H_{1M} & H_{22} - H_{2M} & \ldots & H_{MM}
\end{array}
\]

Block 2: Study sample 2 as observation

\[
\begin{array}{c|c|c|c}
H_{00} - H_{0M} & H_{11} - H_{1M} & H_{22} - H_{2M} & \ldots & H_{MM}
\end{array}
\]

... ...
HMM for imputation: Final Stage

- Fill the missing sites with reference haplotypes
  - If no crossover between two typed sites of two strands, copy corresponding allele data from two reference haplotypes.
  - If there is crossover on one of the strands (1st or 2nd), or on both strands, estimate the occurring position by Monte Carlo sampling, and copy data from corresponding reference haplotypes.
CPU vs. GPU

Samples = 1000

- Femi based GPU
- Intel CPU

<table>
<thead>
<tr>
<th>Type</th>
<th>CPU Time (ms)</th>
<th>GPU Time (ms)</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 reference</td>
<td>182</td>
<td>2430</td>
<td>13.5x</td>
</tr>
<tr>
<td>200 reference</td>
<td>325</td>
<td>4800</td>
<td>14.77x</td>
</tr>
<tr>
<td>100 reference</td>
<td>218</td>
<td>500</td>
<td>11.46x</td>
</tr>
<tr>
<td>200 reference</td>
<td>365</td>
<td>4850</td>
<td>13.28x</td>
</tr>
<tr>
<td>200 reference</td>
<td>663</td>
<td>9550</td>
<td>14.40x</td>
</tr>
<tr>
<td>100 reference</td>
<td>237</td>
<td>2630</td>
<td>11.09x</td>
</tr>
<tr>
<td>200 reference</td>
<td>696</td>
<td>10455</td>
<td>15.02x</td>
</tr>
</tbody>
</table>
Limitation for the current implementation

- Configuring kernel with 48K shared memory and 16K L1 cache can get higher SM occupancy, and better performance.
- Can provide good results if the number of reference haplotypes is small (~120 or less).
- With the increasing of reference panel, active reference pool is limited by the available shared memory size, so the overall imputation quality will decrease.
  - Load all states into shared memory → resource insufficient!
  - Load all states into global memory → VERY poor performance!
THE END

THANKS FOR LISTENING!