BGE pipeline

Comments and Questions

- What issues have you had with running GLIMPSE2 imputation in your samples?
- Do you have slides on finding the optimum batch size along with CPU/Memory requirements?
- What samples and reference panels have you used?
- Have you discovered any unique problems with low pass BGE data?
- How can things be improved (both within your control and outside your control) to make a better pipeline?

Comments and Questions

- What issues have you had with running GLIMPSE2 imputation in your samples?
- Do you have slides on finding the optimum batch size along with CPU/Memory requirements?
- What samples and reference panels have you used?
- Have you discovered any unique problems with low pass BGE data?
- How can things be improved (both within your control and outside your control) to make a better pipeline?

What issues have you had with running GLIMPSE2 imputation in your samples?

- How do you choose a proper window size?
  - GLIMPSE2_chunk
    - Default parameter:
    - --window-cm arg (=2.5) Minimal window size in cM
    - --window-mb arg (=2) Minimal window size in Mb
    - **1,048 chunks**
  - GLIMPSE2 github (resources/chunks/b38)
    - 4cM
    - The 4cM lengths are designed for SHAPEIT5 phase_rare and GLIMPSE2.
    - This chunking has been used to phase the UK Biobank interim release of 200k WGS.
    - **575 chunks**

Do you have slides on finding the optimum batch size along with CPU/Memory requirements?

- HGDP+KGP
  - (Koenig Z, Yohannes MT, Nkambule LL, et al. A harmonized public resource of deeply sequenced diverse human genomes. Preprint. *bioRxiv*.)

- **2.5cM** window size
  - Less memory required

HGDP+KGP (2.5cM chunks)

Batch size	Time (s)	Peak mem (G)	Cost ($)	Estimated total cost (n=921)
5	210	5.86	0.0126	2416.4784
10	317	8.86	0.019	1831.904
25	510	10.33	0.0305	1182.668
50	911	10.36	0.0545	1085.204
100	1574	14.12	0.0942	987.216
912	6720	18.12	0.4027	422.0296

8 CPU, Standard (30G MEM) Region: 10_101140021_107819005

HGDP+KGP (2.5cM chunks)

- 912 samples
  - Including all samples in the analysis
  - 1,048 chunks
    - 1,046 jobs on 8CPU-Standard (30G)
    - 2 jobs on 16CPU-Highmem (104G)
  - Cost: 384.2289 + 2.8959 = $387.1

How can things be improved (both within your control and outside your control) to make a better pipeline?

- Is it possible to set the automatic resubmission to high memory machine if the job failed due to insufficient memory?