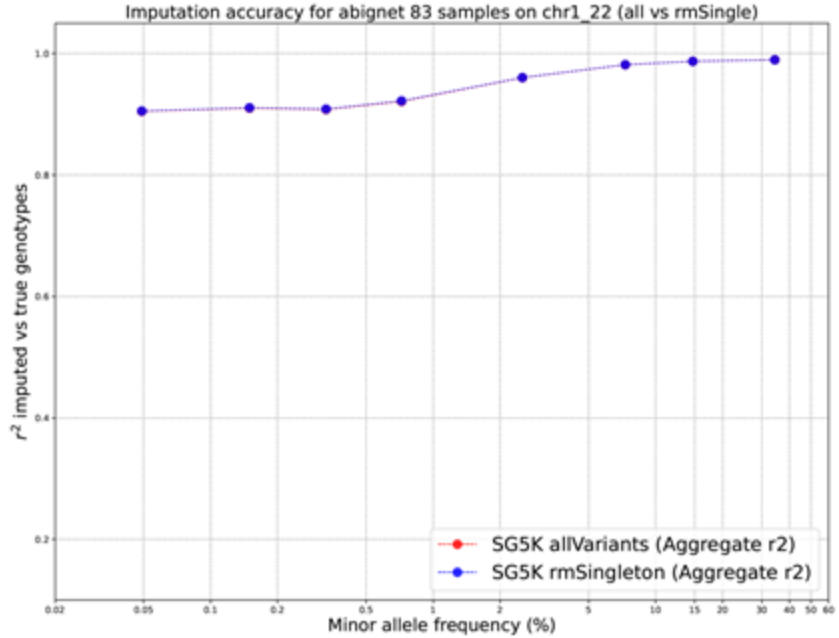
https://docs.google.com/spreadsheets/d/1asV5qImFf3q3gxjbSCIgYiW1ctV_YJmdm1arn2bP048/edit?gid=1035763666#gid=1035763666 (testing cost)

https://hail.zulipchat.com/#narrow/channel/223457-Hail-Batch-support/topic/Way.20to.20reduce.20costs.20further/near/391208898 (chat about Hail batch machine type)

2023-12-25

<BioX 18K - BGE lp-WGS Glimpse imputation using Hail batch>							
Hail batch ID	**Cohort**	**Name**	**Submitted Sample batch**	**Succeeded**	**Failed**	**Duration**	**Cost**
8092419	BioX case	phase_index: c4-standard-t4: sample200	44	44	0	1 day 13 hours	$2,695.08
8092371	BioX control	phase_index: c4-standard-t4: sample200	3	3	0	3 hours 24 minutes	$180.60
8092374	BioX control	phase_index: c4-standard-t4: sample200	41	39	2	1 day 6 hours	$2,475.69
8093229	BioX control	phase_index: **c8**-standard-t4: sample200	2	2	0	2 hours 56 minutes	$0.81
						~ 3 days	$5,352.19

1. **Completed BioX 18K imputation using Hail batch**
    a. Decreased expected cost after removing singletons from reference panel
        i. Same concordance with “all variants from ref”
        ii. 4-Standard-4 was the most efficient and time-reasonable
    b. Glimpse2 imputation for BioX 18K using Hail batch
        i. Hail batch machine type: 4core/ Standard memory/ 4 threads
        ii. Input: 18K BGE BioX (EAS) crams
        iii. Use of Glimpse2's ref-only mode
            1. if we want to use combined model (not yet released) in future, cromwell on terra would be better than hail batch since the job duration of hail batch becomes longer, the probability of failure in the hail batch increases
        iv. Reference panel: chr1-chr22 from Singapore 5k without singleton (4cM chunks)
        v. https://docs.google.com/spreadsheets/d/1asV5qImFf3q3gxjbSCIgYiW1ctV_YJmdm1arn2bP048/edit#gid=1481255298
    c. Now, running average AF & aggregated INFO



	All variants from Ref (sg5k)				Without singletons from Ref (sg5k)		
gnomad EAS maf bin	#variants	gnomad EAS maf	aggR2	gnomad EAS maf bin	#variants	gnomad EAS maf	aggR2
0~0.1%	110,618,834	0.00049065	0.904738	0~0.1%	110,618,834	0.00049065	0.905247
0.1~0.2%	49,471,896	0.00149672	0.909962	0.1~0.2%	49,471,896	0.00149672	0.910666
0.2~0.5%	75,226,220	0.0033031	0.907685	0.2~0.5%	75,226,220	0.0033031	0.908455
0.5~1%	65,801,248	0.00721831	0.921033	0.5~1%	65,801,248	0.00721831	0.921922
1~5%	172,788,193	0.0252088	0.960029	1~5%	172,788,193	0.0252088	0.960363
5~10%	90,497,845	0.0730934	0.981442	5~10%	90,497,845	0.0730934	0.98151
10~20%	122,648,580	0.14712	0.987072	10~20%	122,648,580	0.14712	0.987109
20~50%	259,198,199	0.343604	0.989548	20~50%	259,198,199	0.343604	0.989589

core		memory (core:low-mem = 1:0.9) (core:standard mem = 1:3.75)		thread				sample			
4	8	low	standard	1	2	4	8	10	50	100	200
Total = 2 x 2 x 4 x 4 = 64 tests											

Setting for Hail batch				Imputation (chr1-22) experiment using Singapore 5K reference panel							
Machine type			Sample Batch	All variants from Ref				Without singletons from Ref			
Core	Memory	Threads	Size	Duration per 2 x (sample size)	Cost per 2 x (sample size)	Expected cost for a sample	Expected cost for 18K	Duration per 2 x (sample size)	Cost per 2 x (sample size)	Expected cost for a sample	Expected cost for 18K*
c4	standard	t4	200	~~52 minutes 54s~~	~~$0.23~~	Out of memory	Out of memory	2 hours 7 minutes	$0.68	$0.49	$8,811.73
c4	standard	t4	100	~~1 hour 3 minutes~~	~~$0.25~~	Out of memory	Out of memory	1 hour 6 minutes	$0.34	$0.49	$8,841.49
c4	standard	t4	50	~~15 minutes 58s~~	~~$0.07~~	Out of memory	Out of memory	35 minutes 26s	$0.18	$0.50	$9,071.78
c8	standard	t8	200	~~1 hour 2 minutes~~	~~$0.51~~	Out of memory	Out of memory	1 hour 15 minutes	$0.73	$0.53	$9,502.59
c8	standard	t8	100	~~33 minutes 27s~~	~~$0.28~~	Out of memory	Out of memory	42 minutes 10s	$0.40	$0.58	$10,445.74
c4	standard	t2	200	3 hours 19 minutes	$1.01	$0.73	$13,075.93	2 hours 50 minutes	$0.88	$0.63	$11,427.69
c8	standard	t4	200	1 hour 45 minutes	$1.09	$0.78	$14,108.34	1 hour 29 minutes	$0.89	$0.64	$11,498.85
c8	standard	t8	50	~~19 minutes 24s~~	~~$0.16~~	Out of memory	Out of memory	23 minutes 54s	$0.22	$0.65	$11,628.23
c8	standard	t4	100	55 minutes 6s	$0.54	$0.77	$13,894.88	47 minutes 33s	$0.47	$0.67	$12,101.74
c8	standard	t4	50	31 minutes 46s	$0.30	$0.86	$15,499.13	26 minutes 9s	$0.25	$0.71	$12,777.08
c4	standard	t2	50	55 minutes 29s	$0.28	$0.80	$14,453.78	49 minutes 45s	$0.25	$0.72	$12,953.03
c4	standard	t4	10	~~7 minutes~~	~~$0.03~~	Out of memory	Out of memory	11 minutes 20s	$0.05	$0.73	$13,222.13
c4	standard	t2	100	1 hour 38 minutes	$0.50	$0.72	$12,919.39	2 hours 9 minutes	$0.60	$0.86	$15,563.81
c4	standard	t2	10	13 minutes 58s	$0.07	$0.95	$17,155.13	14 minutes 9s	$0.07	$0.94	$16,974.00
c8	standard	t8	10	10 minutes 41s	$0.08	$1.22	$21,942.00	9 minutes 22s	$0.07	$1.05	$18,914.63
c8	standard	t4	10	10 minutes 36s	$0.09	$1.27	$22,925.25	9 minutes 34s	$0.08	$1.13	$20,286.00
c4	standard	t1	100	3 hours 20 minutes	$1.01	$1.45	$26,157.04	2 hours 48 minutes	$0.85	$1.23	$22,050.68
c8	standard	t2	200	3 hours 17 minutes	$2.01	$1.44	$25,940.98	2 hours 47 minutes	$1.71	$1.23	$22,147.71
c8	standard	t2	100	1 hour 41 minutes	$1.02	$1.46	$26,304.53	1 hour 25 minutes	$0.86	$1.23	$22,211.10
c4	standard	t1	50	1 hour 31 minutes	$0.47	$1.34	$24,156.90	1 hour 25 minutes	$0.43	$1.24	$22,309.43
c8	standard	t2	50	55 minutes 35s	$0.55	$1.57	$28,348.65	45 minutes 52s	$0.45	$1.30	$23,406.53
c4	standard	t1	200	6 hours 24 minutes	$1.95	$1.40	$25,246.24	6 hours 10 minutes	$1.84	$1.32	$23,777.83
c4	standard	t1	10	24 minutes 8s	$0.12	$1.69	$30,506.63	20 minutes 28s	$0.10	$1.47	$26,496.00
c8	standard	t2	10	14 minutes 50s	$0.13	$1.91	$34,362.00	14 minutes 1s	$0.12	$1.74	$31,334.63
c8	lowmem	t1	200	6 hours 34 minutes	$3.61	$2.60	$46,710.84	5 hours 30 minutes	$3.10	$2.23	$40,059.68
c8	lowmem	t1	50	1 hour 44 minutes	$0.95	$2.74	$49,385.03	1 hour 26 minutes	$0.77	$2.23	$40,101.08
c8	lowmem	t1	100	6 hours 7 minutes	$2.90	$4.18	$75,161.70	2 hours 53 minutes	$1.56	$2.24	$40,341.71
c8	standard	t1	200	6 hours 44 minutes	$4.09	$2.94	$52,913.08	5 hours 26 minutes	$3.34	$2.40	$43,219.01
c8	standard	t1	100	3 hours 15 minutes	$1.97	$2.83	$51,025.50	2 hours 46 minutes	$1.68	$2.42	$43,563.15
c8	standard	t1	50	2 hours 57 minutes	$1.57	$4.52	$81,288.90	1 hour 25 minutes	$0.85	$2.45	$44,137.58
c8	lowmem	t1	10	24 minutes 30s	$0.20	$2.90	$52,267.50	22 minutes 29s	$0.18	$2.62	$47,092.50
c8	standard	t1	10	22 minutes 43s	$0.22	$3.20	$57,623.63	20 minutes 15s	$0.19	$2.77	$49,809.38

c4-standard-t4 is the most efficient, but it's still expensive because I calculated this expected cost based on the assumption of using bins with maximum size.

So, the actual cost has been lower than expected!
-> $5,353.19