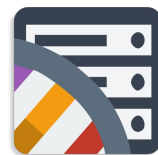


## Section 2

# Run a job, Data Preparation and Data Download



**MICHIGAN**  
IMPUTATIONSERVER



Sebastian Schönherr  
Medical University of Innsbruck  
sebastian.schoenherr@i-med.ac.at  
@seppinho

# Learning objectives

## Participants will learn

1. How to submit a job on Michigan Imputation Server (MIS)
2. How to prepare your GWAS data
3. Different ways to download final datasets

# Run your first job on MIS or TMIS

<https://imputationserver.sph.umich.edu> (MIS)

**or**

<https://imputation.biodatacatalyst.nhlbi.nih.gov> (TOPMed  
Server)

### Uploading Data

1

1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025, 2026, 2027, 2028, 2029, 2030, 2031, 2032, 2033, 2034, 2035, 2036, 2037, 2038, 2039, 2040, 2041, 2042, 2043, 2044, 2045, 2046, 2047, 2048, 2049, 2050, 2051, 2052, 2053, 2054, 2055, 2056, 2057, 2058, 2059, 2060, 2061, 2062, 2063, 2064, 2065, 2066, 2067, 2068, 2069, 2070, 2071, 2072, 2073, 2074, 2075, 2076, 2077, 2078, 2079, 2080, 2081, 2082, 2083, 2084, 2085, 2086, 2087, 2088, 2089, 2090, 2091, 2092, 2093, 2094, 2095, 2096, 2097, 2098, 2099, 2100, 2101, 2102, 2103, 2104, 2105, 2106, 2107, 2108, 2109, 2110, 2111, 2112, 2113, 2114, 2115, 2116, 2117, 2118, 2119, 2120, 2121, 2122, 2123, 2124, 2125, 2126, 2127, 2128, 2129, 2130, 2131, 2132, 2133, 2134, 2135, 2136, 2137, 2138, 2139, 2140, 2141, 2142, 2143, 2144, 2145, 2146, 2147, 2148, 2149, 2150, 2151, 2152, 2153, 2154, 2155, 2156, 2157, 2158, 2159, 2160, 2161, 2162, 2163, 2164, 2165, 2166, 2167, 2168, 2169, 2170, 2171, 2172, 2173, 2174, 2175, 2176, 2177, 2178, 2179, 2180, 2181, 2182, 2183, 2184, 2185, 2186, 2187, 2188, 2189, 2190, 2191, 2192, 2193, 2194, 2195, 2196, 2197, 2198, 2199, 2200, 2201, 2202, 2203, 2204, 2205, 2206, 2207, 2208, 2209, 2210, 2211, 2212, 2213, 2214, 2215, 2216, 2217, 2218, 2219, 2220, 2221, 2222, 2223, 2224, 2225, 2226, 2227, 2228, 2229, 2230, 2231, 2232, 2233, 2234, 2235, 2236, 2237, 2238, 2239, 2240, 2241, 2242, 2243, 2244, 2245, 2246, 2247, 2248, 2249, 2250, 2251, 2252, 2253, 2254, 2255, 2256, 2257, 2258, 2259, 2260, 2261, 2262, 2263, 2264, 2265, 2266, 2267, 2268, 2269, 2270, 2271, 2272, 2273, 2274, 2275, 2276, 2277, 2278, 2279, 2280, 2281, 2282, 2283, 2284, 2285, 2286, 2287, 2288, 2289, 2290, 2291, 2292, 2293, 2294, 2295, 2296, 2297, 2298, 2299, 2300, 2301, 2302, 2303, 2304, 2305, 2306, 2307, 2308, 2309, 2310, 2311, 2312, 2313, 2314, 2315, 2316, 2317, 2318, 2319, 2320, 2321, 2322, 2323, 2324, 2325, 2326, 2327, 2328, 2329, 2330, 2331, 2332, 2333, 2334, 2335, 2336, 2337, 2338, 2339, 2340, 2341, 2342, 2343, 2344, 2345, 2346, 2347, 2348, 2349, 2350, 2351, 2352, 2353, 2354, 2355, 2356, 2357, 2358, 2359, 2360, 2361, 2362, 2363, 2364, 2365, 2366, 2367, 2368, 2369, 2370, 2371, 2372, 2373, 2374, 2375, 2376, 2377, 2378, 2379, 2380, 2381, 2382, 2383, 2384, 2385, 2386, 2387, 2388, 2389, 2390, 2391, 2392, 2393, 2394, 2395, 2396, 2397, 2398, 2399, 2400, 2401, 2402, 2403, 2404, 2405, 2406, 2407, 2408, 2409, 2410, 2411, 2412, 2413, 2414, 2415, 2416, 2417, 2418, 2419, 2420, 2421, 2422, 2423, 2424, 2425, 2426, 2427, 2428, 2429, 2430, 2431, 2432, 2433, 2434, 2435, 2436, 2437, 2438, 2439, 2440, 2441, 2442, 2443, 2444, 2445, 2446, 2447, 2448, 2449, 2450, 2451, 2452, 2453, 2454, 2455, 2456, 2457, 2458, 2459, 2460, 2461, 2462, 2463, 2464, 2465, 2466, 2467, 2468, 2469, 2470, 2471, 2472, 2473, 2474, 2475, 2476, 2477, 2478, 2479, 2480, 2481, 2482, 2483, 2484, 2485, 2486, 2487, 2488, 2489, 2490, 2491, 2492, 2493, 2494, 2495, 2496, 2497, 2498, 2499, 2500, 2501, 2502, 2503, 2504, 2505, 2506, 2507, 2508, 2509, 2510, 2511, 2512, 2513, 2514, 2515, 2516, 2517, 2518, 2519, 2520, 2521, 2522, 2523, 2524, 2525, 2526, 2527, 2528, 2529, 2530, 2531, 2532, 2533, 2534, 2535, 2536, 2537, 2538, 2539, 2540, 2541, 2542, 2543, 2544, 2545, 2546, 2547, 2548, 2549, 2550, 2551, 2552, 2553, 2554, 2555, 2556, 2557, 2558, 2559, 2560, 2561, 2562, 2563, 2564, 2565, 2566, 2567, 2568, 2569, 2570, 2571, 2572, 2573, 2574, 2575, 2576, 2577, 2578, 2579, 2580, 2581, 2582, 2583, 2584, 2585, 2586, 2587, 2588, 2589, 2590, 2591, 2592, 2593, 2594, 2595, 2596, 2597, 2598, 2599, 2600, 2601, 2602, 2603, 2604, 2605, 2606, 2607, 2608, 2609, 2610, 2611, 2612, 2613, 2614, 2615, 2616, 2617, 2618, 2619, 2620, 2621, 2622, 2623, 2624, 2625, 2626, 2627, 2628, 2629, 2630, 2631, 2632, 2633, 2634, 2635, 2636, 2637, 2638, 2639, 2640, 2641, 2642, 2643, 2644, 2645, 2646, 2647, 2648, 2649, 2650, 2651, 2652, 2653, 2654, 2655, 2656, 2657, 2658, 2659, 2660, 2661, 2662, 2663, 2664, 2665, 2666, 2667, 2668, 2669, 2670, 2671, 2672, 2673, 2674, 2675, 2676, 2677, 2678, 2679, 26

© 2004 Blackwell Publishing Ltd, *Journal of Internal Medicine* 255: 103–110

■ **Additional Resources:** [www.pearsoncmg.com](http://www.pearsoncmg.com) Visit the Pearson website for more information on this book and other titles in the series.

100

**Abstract** The purpose of this study was to determine if the use of a computer-based, interactive, self-paced, and self-directed learning program, the "Computerized Health Education Program" (CHEP), could improve the knowledge and attitudes of health care providers regarding the use of the computer in health care. The program was designed to provide information on the use of the computer in health care, including the benefits and limitations of the computer, the types of computers available, and the types of software available. The program was evaluated using a pre-test/post-test design. The results of the study indicated that the use of the CHEP program resulted in a significant increase in knowledge and attitudes regarding the use of the computer in health care. The program was found to be an effective tool for providing health care providers with information on the use of the computer in health care.

© 2006 The Authors  
Journal compilation © 2006 Blackwell Publishing Ltd, *Journal of Internal Medicine* 260: 105–112

© 2006 The Authors  
Journal compilation © 2006 Blackwell Publishing Ltd

# Recap

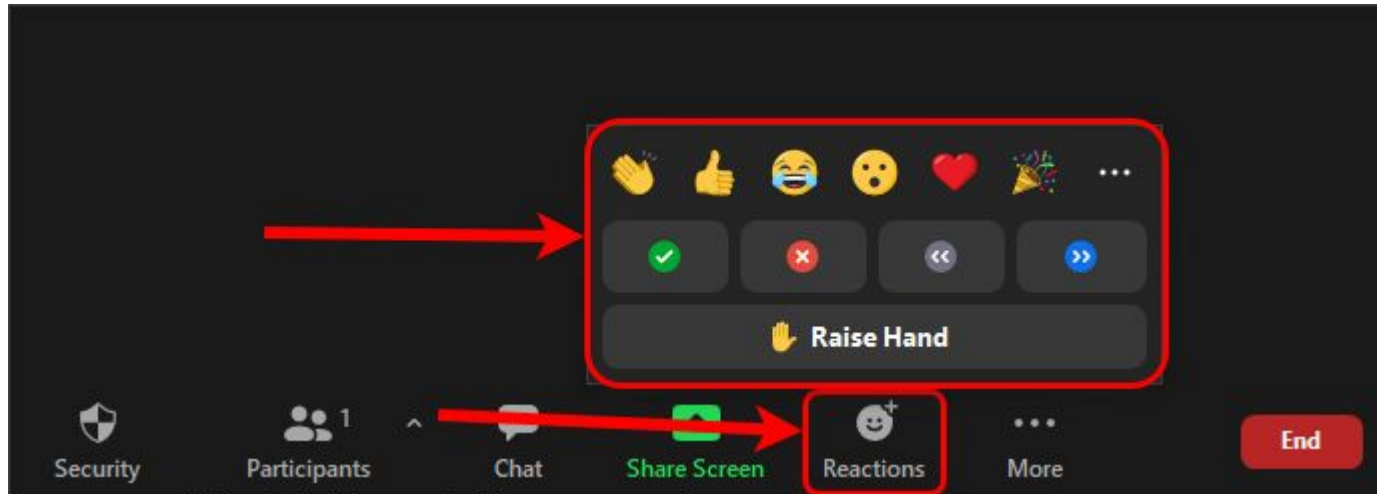
- Input Validation and Quality Control executed right after data upload
  - Immediate feedback to users
  - Jobs passing the QC are then added to a long-time queue
- MIS outputs SNP statistics and a QC Report for each job
  - Helps you to identify problems

# Have you run into QC problems so far?

Put a "Yes" in the chat

Or

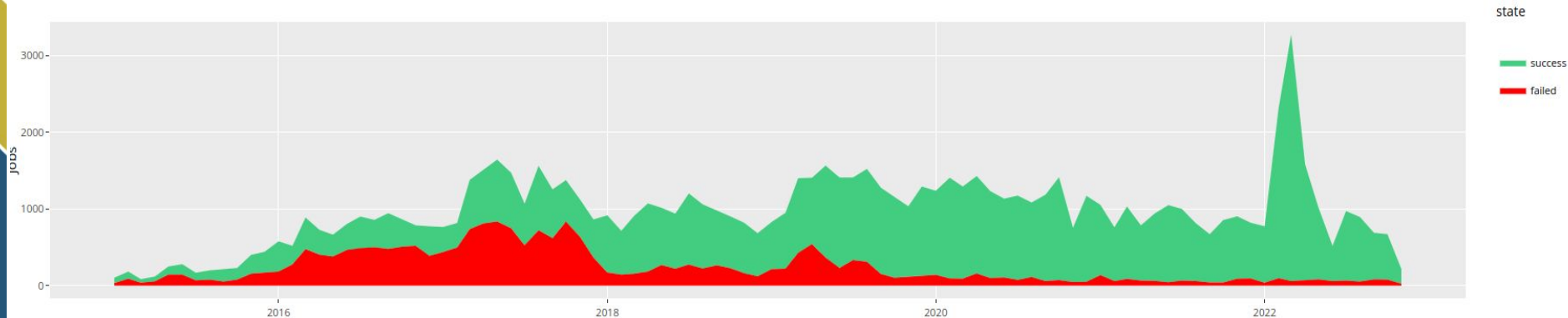
Raise Hand



# How many jobs are failing?

- 40% in 2015; 20% in 2019; 7% 2020-2022
  - Reason for job failures: Something wrong with your input data **or** phasing/imputation issue on our side

Total amount of jobs: **91,400** (Nov 22)  
> 2M in July 22



# MIS QC: Input Validation & Statistics

-

	Imputation Server
Input	<b>VCF / chromosome</b>
Output	Imputed VCF / chromosome
<b>File Validation &amp; Statistics</b>	
Basic SNP Filtering	
Lift Over	

## Input Validation

4 valid VCF file(s) found.

Samples: 51471

Chromosomes: 11 12 13 14

SNPs: 72808

Chunks: 26

Datatype: unphased

Build: hg19

Reference Panel: apps@1000g-phase-3-v5 (hg19)

Population: eur

Phasing: eagle

Mode: imputation



# MIS QC: Basic SNP Filtering

-

	Imputation Server
Input	<b>VCF / chromosome</b>
Output	Imputed VCF / chromosome
File Validation & Statistics	
<b>Basic SNP Filtering</b>	
Lift Over	

## Statistics:

Alternative allele frequency > 0.5 sites: 2,308

Reference Overlap: 99.95 %

Match: 7,816

Allele switch: 0

Strand flip: 0

Strand flip and allele switch: 0

A/T, C/G genotypes: 0

## Filtered sites:

Filter flag set: 0

Invalid alleles: 0

Multiallelic sites: 0

Duplicated sites: 0

NonSNP sites: 0

Monomorphic sites: 0

Allele mismatch: 4

SNPs call rate < 90%: 0

# MIS QC: Lift Over Step

-

	Imputation Server
Input	<b>VCF / chromosome</b>
Output	Imputed VCF / chromosome
File Validation & Statistics	
Basic SNP Filtering	
<b>Lift Over</b>	

## Quality Control

Uploaded data is hg38 and reference is hg19.

Lift Over

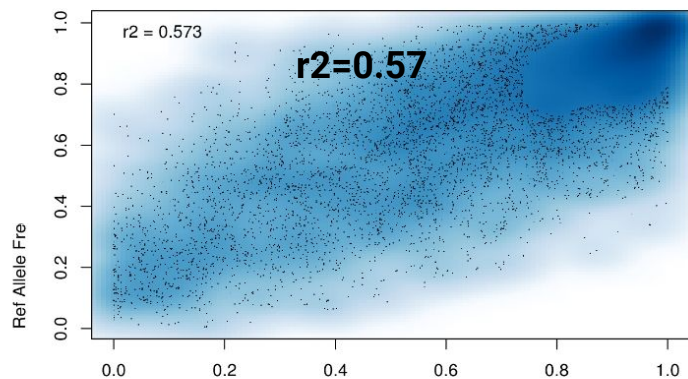
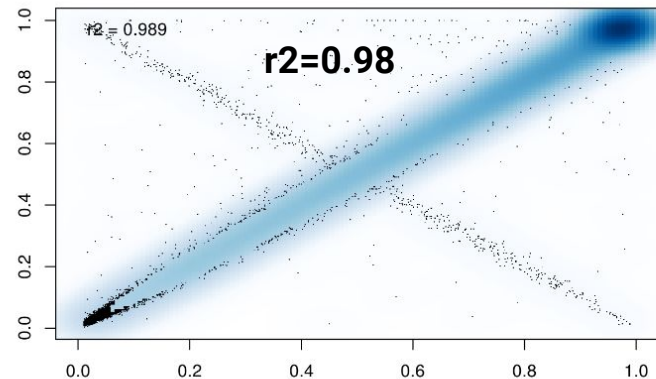
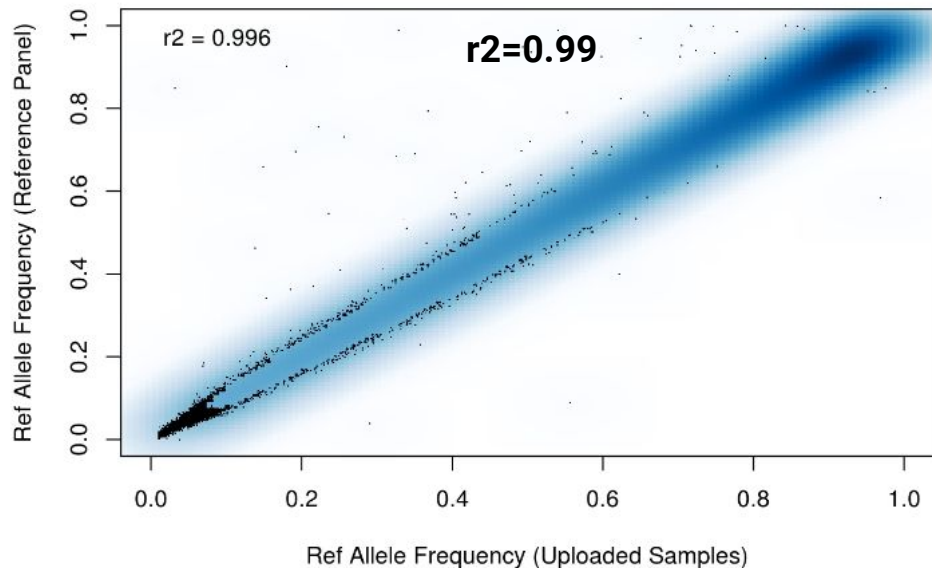
Calculating QC Statistics

### Statistics:

Alternative allele frequency > 0.5 sites: 2

Reference Overlap: 100.00 %

# MIS QC Report: Allele Frequency Check



# Failing Validation - Obvious Problems

## Input Validation

The provided VCF file is malformed. Error during index creation: [tabix] was bgzip used to compress this file? (see [Help](#)).

## Input Validation

The provided VCF file contains more than one chromosome. Please split your input VCF file by chromosome (see [Help](#)).

## Input Validation

Unable to parse header with error: Your input file has a malformed header: We never saw the required CHROM header line (starting with one #) for the input VCF file (see [Help](#)).

# Failing QC - Trickier Problems

Excluded sites in total: 695

Remaining sites in total: 185,791

See [snps-excluded.txt](#) for details

Typed only sites: 397

See [typed-only.txt](#) for details

**Warning:** 2 Chunk(s) excluded: reference overlap < 50.0% (see [chunks-excluded.txt](#) for details).

Remaining chunk(s): 40

**Error:** More than 100 obvious strand flips have been detected. Please check strand. Imputation cannot be started!



## Send Notification on Failure

We have sent an email to [sebastian.schoenherr@i-med.ac.at](mailto:sebastian.schoenherr@i-med.ac.at) with the error message.

# How to fix input files?

# Imputation Preparation Tool

- developed by W. Rayner
- Works for all major reference panels (HRC, TOPMed, Asia, CAAPA, 1000G)
- Checks for consistency between input data and a reference panel
- Updates/removes SNPs, Updates strand, position and ref/alt assignment
- Input Data in PLINK Binary Format (bim, bed, fam)

<https://www.well.ox.ac.uk/~wrayner/tools/HRC-1000G-check-bim-v4.3.0.zip>

# Execute Imputation Tool before uploading data

	Imputation Server	Preparation Tool
Input	<b>VCF / chromosome</b>	PLINK binary data
Output	Imputed VCF / chromosome	<b>VCFs / chromosome</b>
File Validation & Statistics		
Basic SNP Filtering		
Lift Over		
<b>Fixes Strand Errors, Updating Ref / Alt Assignment</b>		
<b>Removes SNPs with allele freq difference, A/T &amp; G/C SNPs if MAF &gt; 0.4</b>		

```
seb@seb-genepi:/data3/projects/ashg-imputation-tool$
```

```
seb@seb-genepi:/data3/projects/ashg-imputation-tool$ perl HRC-1000G-check-bim.pl -b study-raw-filtered.bim -f study.frq -r HRC.r1-1.GRCh37.wgs.mac5.sites.tab.gz -h
```

Script to check plink .bim files against HRC/1000G for strand, id names, positions, alleles, ref/alt assignment

William Rayner 2015-2020

wrayner@well.ox.ac.uk

Version 4.3

Options Set:

Reference Panel: HRC  
Bim filename: study-raw-filtered.bim  
Reference filename: HRC.r1-1.GRCh37.wgs.mac5.sites.tab.gz  
Allele frequencies filename: study.frq  
Plink executable to use: plink

Chromosome flag set: No  
Allele frequency threshold: 0.2

Path to plink bim file: /data3/projects/ashg-imputation-tool



```
seb@seb-genepi:/data3/projects/ashg-imputation-tool$  
seb@seb-genepi:/data3/projects/ashg-imputation-tool$ sh Run-plink.sh  
PLINK v1.90b3.40 64-bit (16 Aug 2016)      https://www.cog-genomics.org/plink2  
(C) 2005-2016 Shaun Purcell, Christopher Chang  GNU General Public License v3  
Logging to /data3/projects/ashg-imputation-tool/TEMP1.log.  
Options in effect:  
  --bfile /data3/projects/ashg-imputation-tool/study-raw-filtered  
  --exclude /data3/projects/ashg-imputation-tool/Exclude-study-raw-filtered-HRC.txt  
  --make-bed  
  --out /data3/projects/ashg-imputation-tool/TEMP1
```

```
32074 MB RAM detected; reserving 16037 MB for main workspace.  
1453472 variants loaded from .bim file.  
5034 people (3027 males, 2007 females) loaded from .fam.  
--exclude: 1392377 variants remaining.  
Using 1 thread (no multithreaded calculations invoked).  
Before main variant filters, 5034 founders and 0 nonfounders present.  
Calculating allele frequencies... done.  
Total genotyping rate is 0.997701.  
1392377 variants and 5034 people pass filters and QC.  
Note: No phenotypes present.  
--make-bed to /data3/projects/ashg-imputation-tool/TEMP1.bed +  
/data3/projects/ashg-imputation-tool/TEMP1.bim +  
/data3/projects/ashg-imputation-tool/TEMP1.fam ... 30%
```


seb@seb-genepi:/data3/projects/ashg-imputation-tool\$

seb@seb-genepi:/data3/projects/ashg-imputation-tool\$ bgzip study-raw-filtered-updated-chr15.vcf

# Life Cycle of an Imputation Job



- Job passed Quality Control
- Job scheduled in imputation queue

**job-20221111-082200-989**  
🕒 Fri Nov 11 2022 14:22:01 ⌚ 17 sec 👤 admin 📁 Genotype Imputation (Minimac4) 1.6.8

Job is in queue on position **24**.

- Waits until resources are available

# Life Cycle of an Imputation Job



- Phasing and Imputation starts



- Waiting
- Running
- Complete

# Life Cycle of an Imputation Job



- Data is encrypted
- Email with one time password is sent to user

Dear Lukas,  
the password for the imputation results is: pp09Z0KeQvQM

The results can be downloaded from <https://imputationserver.sph.umich.edu/start.html#!jobs/job-20190919-112230-581/results>

# Life Cycle of an Imputation Job



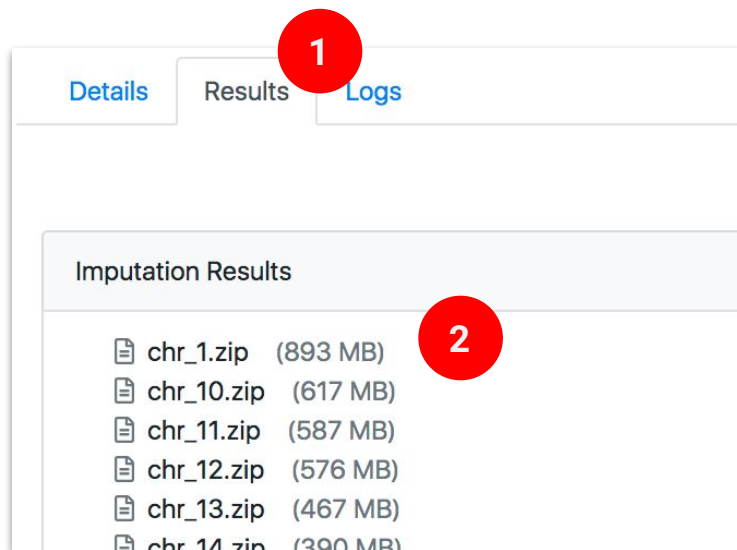
- After 7 days the job is retired
- All results are deleted
- We will send you an email 2 days before

Dear Lukas Forer,  
Your job retires in 2 days! All imputation results will be deleted at that time.

Please ensure that you have downloaded all results from  
<https://imputationserver.sph.umich.edu/start.html#!jobs/job-20191011-124306-370>







How to download  
the imputed genotypes?

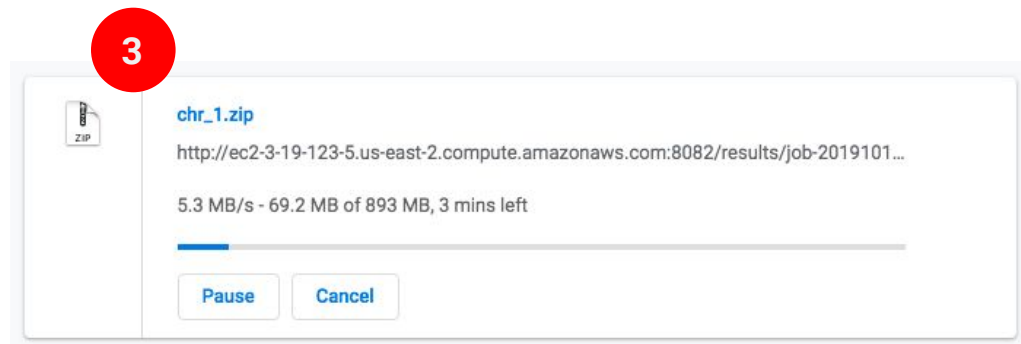
# Option 1: Web-Interface




Details Results **1** Logs

Imputation Results

-  chr\_1.zip (893 MB) **2**
-  chr\_10.zip (617 MB)
-  chr\_11.zip (587 MB)
-  chr\_12.zip (576 MB)
-  chr\_13.zip (467 MB)
-  chr\_14.zip (390 MB)




 **3** chr\_1.zip

<http://ec2-3-19-123-5.us-east-2.compute.amazonaws.com:8082/results/job-2019101...>

5.3 MB/s - 69.2 MB of 893 MB, 3 mins left



# Option 2: Batch Download

Imputation Results  wget

- chr\_1.zip (469 MB)
- chr\_10.zip (287 MB)
- chr\_11.zip (281 MB)
- chr\_12.zip (269 MB)
- chr\_13.zip (195 MB)
- chr\_14.zip (192 MB)
- chr\_15.zip (181 MB)
- chr\_16.zip (203 MB)
- chr\_17.zip (187 MB)
- chr\_18.zip (160 MB)
- chr\_19.zip (170 MB)
- chr\_2.zip (471 MB)
- chr\_20.zip (129 MB)


## Download data

wget (22) [URLs \(22\)](#)

```
wget https://imputationserver.sph.umich.edu/share/results/1fc3d1b  
wget https://imputationserver.sph.umich.edu/share/results/3d9f5f6  
wget https://imputationserver.sph.umich.edu/share/results/528e41f  
wget https://imputationserver.sph.umich.edu/share/results/ed598ab  
wget https://imputationserver.sph.umich.edu/share/results/7c818b4  
wget https://imputationserver.sph.umich.edu/share/results/1c1e65d
```

Use the following command to download all results at once:

```
curl -sL https://imputationserver.sph.umich.edu/get/1584555,
```



OK

fantasia:~> █

```
fantasia:~> curl -sL https://imputationserver.sph.umich.edu/get/1584555/675233d69  
db57b793589b916f2a81cb8 | bash
```

```
fantasia:~> curl -sL https://imputationserver.sph.umich.edu/get/1584555/675233d69
db57b793589b916f2a81cb8 | bash
```

```
Downloading file chr_1.zip (1/22)...
```

% Total		% Received		% Xferd		Average Speed		Time	Time	Time	Current
						Dload	Upload	Total	Spent	Left	Speed
100	185	100	185	0	0	21087	0	--:--:--	--:--:--	--:--:--	30833
100	469M	100	469M	0	0	116M	0	0:00:04	0:00:04	--:--:--	167M

```
Downloading file chr_10.zip (2/22)...
```

% Total		% Received		% Xferd		Average Speed		Time	Time	Time	Current
						Dload	Upload	Total	Spent	Left	Speed
100	185	100	185	0	0	23886	0	--:--:--	--:--:--	--:--:--	37000
100	287M	100	287M	0	0	87.4M	0	0:00:03	0:00:03	--:--:--	138M

```
Downloading file chr_11.zip (3/22)...
```



Downloading file chr\_7.zip (20/22)...

% Total		% Received		% Xferd		Average	Speed	Time	Time	Time	Current
						Dload	Upload	Total	Spent	Left	Speed
100	185	100	185	0	0	24189	0	--:--:--	--:--:--	--:--:--	37000
100	337M	100	337M	0	0	101M	0	0:00:03	0:00:03	--:--:--	160M

Downloading file chr\_8.zip (21/22)...

% Total		% Received		% Xferd		Average	Speed	Time	Time	Time	Current
						Dload	Upload	Total	Spent	Left	Speed
100	185	100	185	0	0	24529	0	--:--:--	--:--:--	--:--:--	37000
100	306M	100	306M	0	0	94.9M	0	0:00:03	0:00:03	--:--:--	152M

Downloading file chr\_9.zip (22/22)...

% Total		% Received		% Xferd		Average	Speed	Time	Time	Time	Current
						Dload	Upload	Total	Spent	Left	Speed
100	185	100	185	0	0	22486	0	--:--:--	--:--:--	--:--:--	37000
100	245M	100	245M	0	0	82.0M	0	0:00:02	0:00:02	--:--:--	84.8M

All 22 file(s) downloaded.

fantasia:~> █

## Option 3: Use Imputation Bot

- Run everything on the command line
- Checkout Session 4

# Data Decryption

- All imputed genotypes are in **encrypted zip files** (e.g. chr\_1.zip)
- We send you an email with a password

Dear Lukas,  
the password for the imputation results is: pp09Z0KeQvQMc

The results can be downloaded from <https://imputationserver.sph.umich.edu/start.html#!jobs/job-20190919-112230-581/results>

- You need this password to **decrypt** your genotypes
- Decryption with standard zip programs (e.g. WinZip, 7zip or gunzip)
- AES Encryption: Needs additional software to decrypt (e.g. 7z)

# What is in each zip file?

chr\_20.zip

└─ chr20.dose.vcf.gz

└─ chr20.info.gz



# What is in each zip file?

chr\_20.zip

└─ chr20.dose.vcf.gz

└─ chr20.info.gz

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
20	61795	20:61795:G:T	G	T	.	PASS	AF=0.26318;MAF=0.26318;R2=0.54658;IMPUTED
20	63231	20:63231:T:G	T	G	.	PASS	AF=0.03843;MAF=0.03843;R2=0.67736;IMPUTED
20	63244	20:63244:A:C	A	C	.	PASS	AF=0.16132;MAF=0.16132;R2=0.49907;IMPUTED

# What is in each zip file?

chr\_20.zip

├── chr20.dose.vcf.gz  
└── chr20.info.gz

#CHROM	POS	ID	REF	ALT
20	61795	20:61795:G:T	G	T
20	63231	20:63231:T:G	T	G
20	63244	20:63244:A:C	A	C

...

...

...

FORMAT	Sample1
GT:DS:GP	1 0:1.126:0.100,0.673,0.226
GT:DS:GP	0 0:0.002:0.998,0.002,0.000
GT:DS:GP	0 0:0.285:0.723,0.270,0.008

# What is in each zip file?

chr\_20.zip

└─ chr20.dose.vcf.gz

└─ **chr20.info.gz**

```
SNP      REF(0)  ALT(1)  ALT_Frq  MAF      AvgCall  Rsq      Genotyped ...  
20:61795:G:T  G      T      0.26318  0.26318  0.88455  0.54658  Imputed ...  
20:63231:T:G  T      G      0.03843  0.03843  0.98342  0.67736  Imputed ...  
20:63244:A:C  A      C      0.16132  0.16132  0.91761  0.49907  Imputed ...
```

# What is in each zip file?

md5 checksum file

chr\_20.zip

- chr20.dose.vcf.gz
- chr20.info.gz

```
(base) seb@seb-laptop:~/ashg22$ cat results.md5
3ea13c00d323117e0b4648a683175d39 chr_11.zip
9ecb19e40d3f8a55f128c640333ab2ef chr_22.zip
161918ed598f32bcd88536399695b398 chr_12.zip
2709ee09f353c0b332686fdf40e9d062 chr_13.zip
```

SNP	REF(0)	ALT(1)	ALT_Frq	MAF	AvgCall	Rsq	Genotyped	...
20:61795:G:T	G	T	0.26318	0.26318	0.88455	0.54658	Imputed	...
20:63231:T:G	T	G	0.03843	0.03843	0.98342	0.67736	Imputed	...
20:63244:A:C	A	C	0.16132	0.16132	0.91761	0.49907	Imputed	...

# Summary

- MIS Web Interface provides a fast and reliable way to impute data
- MIS applies a strict Quality Control with the goal to return high quality imputation data
- Pre-Imputation tools available for data preparation
- Different options to download data

More info and FAQ can be found here:  
<https://imputationserver.readthedocs.io>