Imputation using Sanger institute - draft

Jean-Tristan Brandenburg

July 8, 2022

1 Introduction : imputation genetics

Imputation in genetics refers to the statistical inference of unobserved genotypes or missing data : using Linkage disequilibrium and/or reference panel, objectives if to defined missing information and imputed position between your positions.

Common step before imputation, imputation and after imputation could be :

- Quality Control of your genetics data, for instance using qc from h3agwas piplene
- Format and prepared your file for imputation
- Phasing your data : Phasing is the task or process of assigning alleles (the As, Cs, Ts and Gs) to the paternal and maternal chromosomes)
- Imputed data using reference panel.
- post qc your data and format your data to used Genome wide associaiton software

Imputation quality depend of software and panel imputation

1.1 resource

- Genotype Imputation [1]

2 Steps of imputation

2.1 QC

for more information on Quality Control see qc from h3agwas piplene and [2]

2.2 Format data

Main of imputation pipeline used VCF as input. When you prepared your data of array intialy in Plink in VCF must be done carefully. Pipeline can rejected your vcf file if allele reference is not good for instance. h3agwas pipeline can transform your plink data in vcf and check your VCF.

For instance :

```
wget -c http://ftp.ensembl.org/pub/grch37/current/fasta/homo_sapien/dna/Homo_sapiens.GRCh37.

nextflow h3abionet/h3agwas/formatdata/plk_in_vcf_imp.nf \
 --input_dir dataqc \
 --input_pat exampledata2_qc \
 --reffasta Homo_sapiens.GRCh37.dna.primary_assembly.fa.gz \
 --output_pat exampledata2_qc \
 --output_dir ./vcfforimputation \
  -profile slurmSingularity \
 -resume
```

2.3 Phasing and imputation

Different way exist to phase and impute, using pipeline for instance

- chipimputation pipeline from h3abionet
- Sanger institute online server
- Michigan online server.

On Sanger and Michigan of online server proposed specific imputation panel as sanger with african panels.

3 Sanger imputation

3.1 Parameters

Sanger imputation, read about paramaters of choice :

- important to read about reference panels
- Important to read about pipeline

3.2 African panels from sanger

Some specificity of pipeline there are not insertion deletion.

3.3 Transfer of data

To transfer data, Sanger used globus.

3.3.1 globus installation and run

- create you need to created a account on globus

You can connect to globus or used globus information relatives to download

```
mkdir globus
 cd globus/
# download
wget https://downloads.globus.org/globus-connect-personal/linux/stable/globusconnectpersonal
# extract globus
tar xzf globusconnectpersonal-latest.tgz
# go on directory
 cd globusconnectpersonal-3.1.3/
 # connect your session follow instruction
 ./globusconnect -setup -nogui
 #open a screen
```

To open

```
 screen
 # open globus
 ./globusconnect -start  \# start globus
 # close screen ctrl-a ctrl-d
```

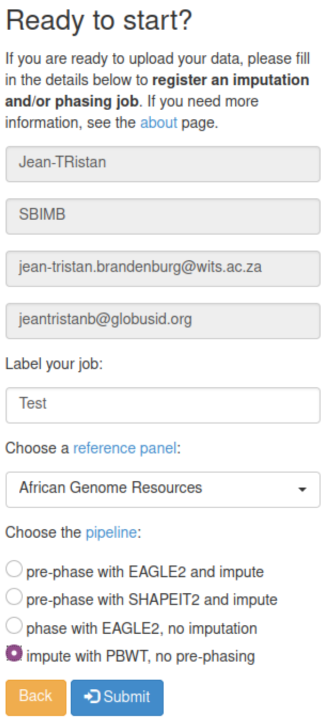
3.3.2 Created jobs and transfert data

AWIGEN have been imputed using african panels from sanger with PBWT and check what phasing.

- Sanger institute used globus to transfer file see globus section, you need to follow globus.
- created jobs for imputation on imputation sanger,see 1, 2
- You will received emel 3 of validation follow link selected your folder on the right, selected (drag and drop from your folder) your vcf file to transfert. 4
- wait confirmation of transfert by globus
- you will receive also a emel from sanger, when globus validate transfer by emel 5 click on link of sanger 6 to validate your sanger.7

3.3.3 Obtained your data imputed using globus

Waiting that you confirmation emel that well running from server and data ready from globus, click on link (be sure that you globus is on) see 7. transfers data on cluster using glob, in ”activity” of website you can follow transfers otherwise you will receive a message of confirmation 8, 9. You can deleted your data after transfer have been done, 10.



Ready to start?

If you are ready to upload your data, please fill in the details below to **register an imputation and/or phasing job**. If you need more information, see the about page.

Jean-TRistan

SBIMB

jean-tristan.brandenburg@wits.ac.za

jeantristanb@globusid.org

Label your job:

Test

Choose a reference panel:

African Genome Resources

Choose the pipeline:

○ pre-phase with EAGLE2 and impute
○ pre-phase with SHAPEIT2 and impute
○ phase with EAGLE2, no imputation
● impute with PBWT, no pre-phasing

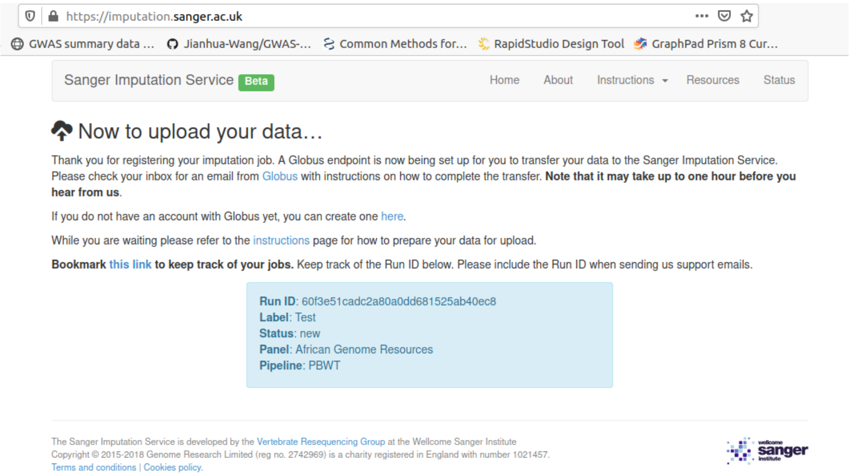
Figure 1: parameter of server

```
## in running
[globusconnectpersonal-3.1.3]$ ./globusconnect -status
Globus Online:   connected
Transfer Status: active

## finish : status is idle
[globusconnectpersonal-3.1.3]$ ./globusconnect -status
Globus Online:   connected
Transfer Status: idle
```

4 Transform your VCF imputed in plink

To transform your files vcf in plink, you can used script in h3abionet/h3agwas pipeline

```
ls imp/vcf/*.vcf.gz > listvcf
wget -c http://ftp.ensembl.org/pub/grch37/current/fasta/homo_sapiens/dna/Homo_sapiens.GRCh37
```



Now to upload your data…

Thank you for registering your imputation job. A Globus endpoint is now being set up for you to transfer your data to the Sanger Imputation Service. Please check your inbox for an email from Globus with instructions on how to complete the transfer. **Note that it may take up to one hour before you hear from us**.

If you do not have an account with Globus yet, you can create one here.

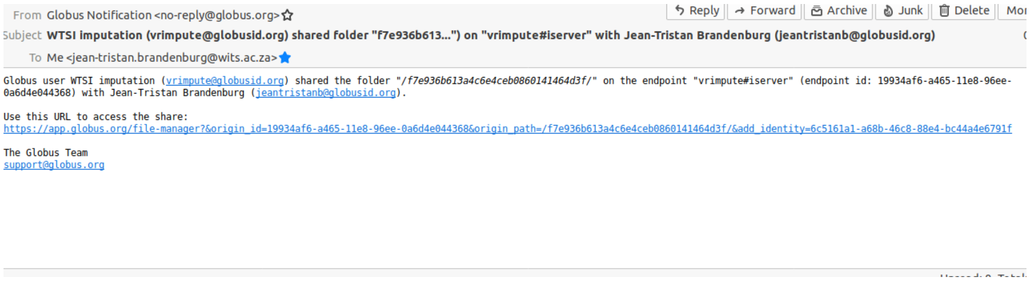
While you are waiting please refer to the instructions page for how to prepare your data for upload.

Bookmark this link to keep track of your jobs. Keep track of the Run ID below. Please include the Run ID when sending us support emails.

Run ID: 60f3e51cadc2a80a0dd681525ab40ec8
Label: Test
Status: new
Panel: African Genome Resources
Pipeline: PBWT

The Sanger Imputation Service is developed by the Vertebrate Resequencing Group at the Wellcome Sanger Institute
Copyright © 2015-2018 Genome Research Limited (reg no. 2742969) is a charity registered in England with number 1021457.
Terms and conditions | Cookies policy.

Figure 2: after validation



From Globus Notification <no-reply@globus.org>
Subject **WTSI imputation (vrimpute@globusid.org) shared folder "f7e936b613...") on "vrimpute#iserver" with Jean-Tristan Brandenburg (jeantristanb@globusid.org)**
To Me <jean-tristan.brandenburg@wits.ac.za>

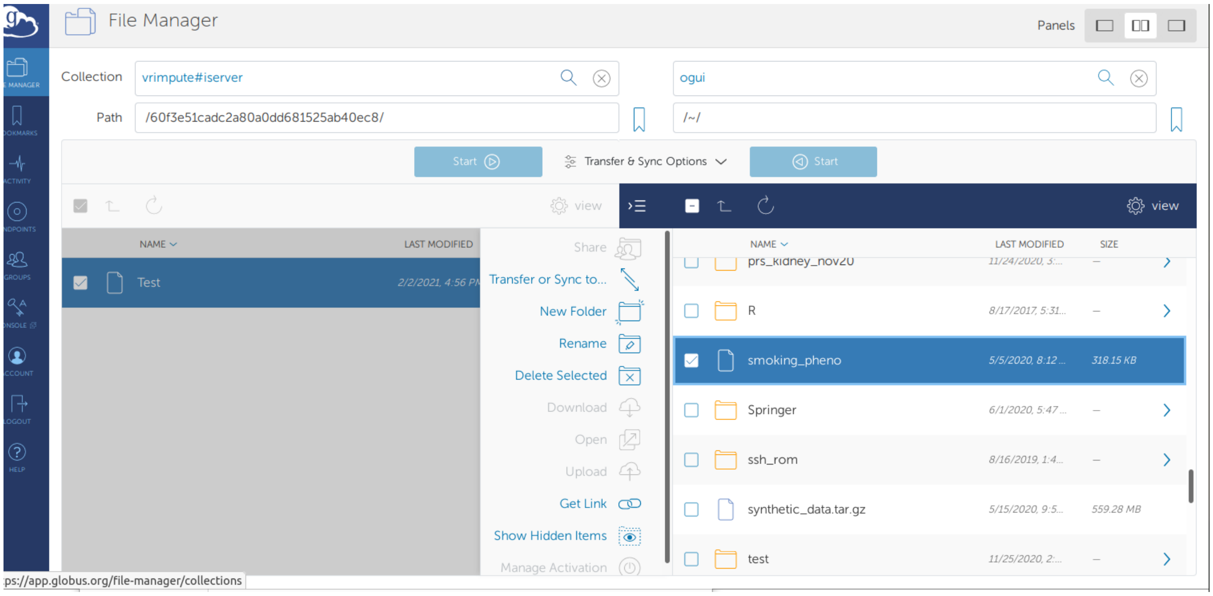
```
Globus user WTSI imputation (vrimpute@globusid.org) shared the folder "/f7e936b613a4c6e4ceb0860141464d3f/" on the endpoint "vrimpute#iserver" (endpoint id: 19934af6-a465-11e8-96ee-
0a6d4e044368) with Jean-Tristan Brandenburg (jeantristanb@globusid.org).

Use this URL to access the share:
https://app.globus.org/file-manager?&origin_id=19934af6-a465-11e8-96ee-0a6d4e044368&origin_path=/f7e936b613a4c6e4ceb0860141464d3f/&add_identity=6c5161a1-a68b-46c8-88e4-bc44a4e6791f

The Globus Team
support@globus.org
```

Figure 3: Emel receive from globus to select job, follow job

```
nextflow run h3abionet/h3agwas/formatdata/vcf_in_plink.nf \
  --file_listvcf listvcf \
  --output_pat exampledata2_imp \
  --output_dir fileimputation/ \
  --reffasta Homo_sapiens.GRCh37.dna.primary_assembly \
-profile slurmSingularity
```

References

[1] Yun Li et al. “Genotype Imputation”. In: *Annual review of genomics and human genetics* 10 (2009), pp. 387–406. ISSN: 1527-8204. DOI: 10.1146/annurev.genom.9.081307.164242. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2925172/ (visited on 07/06/2022).



Collection vrimpute#iserver

Path /60f3e51cadc2a80a0dd681525ab40ec8/

ogui

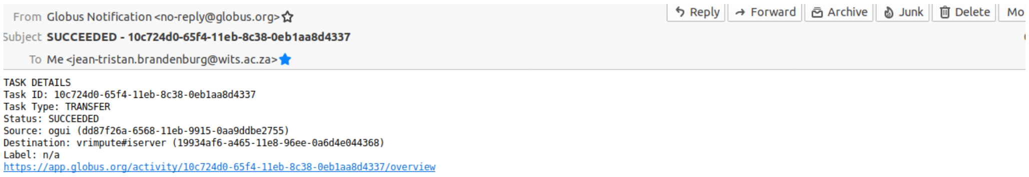
/~/

	NAME	LAST MODIFIED
☑	Test	2/2/2021, 4:56 PM

	NAME	LAST MODIFIED	SIZE
□	prs_kidney_nov20	11/24/2020, 3:...	–
□	R	8/17/2017, 5:31...	–
☑	smoking_pheno	5/5/2020, 8:12 ...	318.15 KB
□	Springer	6/1/2020, 5:47 ...	–
□	ssh_rom	8/16/2019, 1:4...	–
□	synthetic_data.tar.gz	5/15/2020, 9:5...	559.28 MB
□	test	11/25/2020, 2:...	–

Figure 4: transfert data

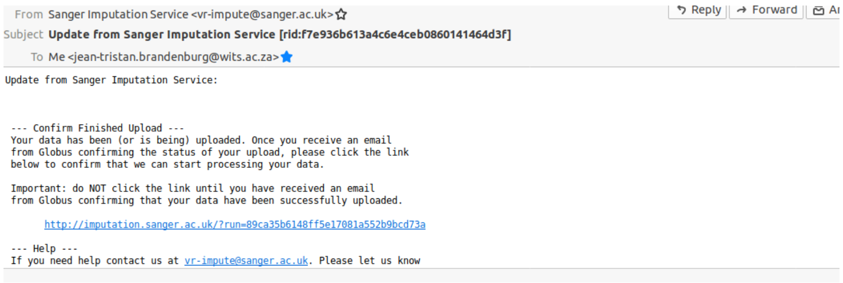
[2] Stephen Turner et al. “Quality Control Procedures for Genome Wide Association Studies”. In: *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]* CHAPTER (Jan. 2011), Unit1.19. ISSN: 1934-8266. DOI: 10.1002/0471142905.hg0119s68. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3066182/ (visited on 07/06/2022).



From Globus Notification <no-reply@globus.org>
Subject **SUCCEEDED - 10c724d0-65f4-11eb-8c38-0eb1aa8d4337**
To Me <jean-tristan.brandenburg@wits.ac.za>

```
TASK DETAILS
Task ID: 10c724d0-65f4-11eb-8c38-0eb1aa8d4337
Task Type: TRANSFER
Status: SUCCEEDED
Source: ogui (dd87f26a-6568-11eb-9915-0aa9ddbe2755)
Destination: vrimpute#iserver (19934af6-a465-11e8-96ee-0a6d4e044368)
Label: n/a
https://app.globus.org/activity/10c724d0-65f4-11eb-8c38-0eb1aa8d4337/overview
```

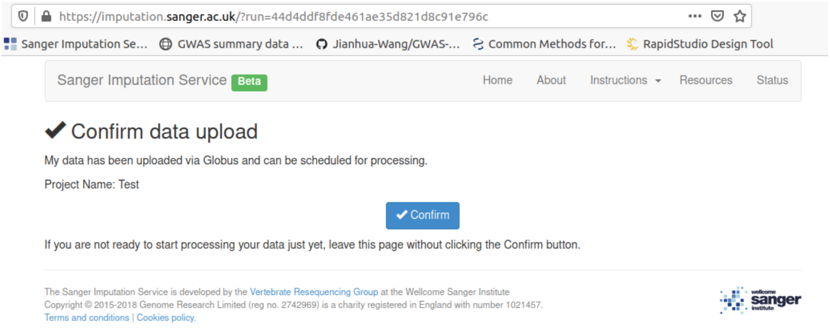
Figure 5: Emel of globus validate download



From Sanger Imputation Service <vr-impute@sanger.ac.uk>
Subject **Update from Sanger Imputation Service [rid:f7e936b613a4c6e4ceb0860141464d3f]**
To Me <jean-tristan.brandenburg@wits.ac.za>

```
Update from Sanger Imputation Service:


 --- Confirm Finished Upload ---
 Your data has been (or is being) uploaded. Once you receive an email
 from Globus confirming the status of your upload, please click the link
 below to confirm that we can start processing your data.

 Important: do NOT click the link until you have received an email
 from Globus confirming that your data have been successfully uploaded.

      http://imputation.sanger.ac.uk/?run=89ca35b6148ff5e17081a552b9bcd73a

 --- Help ---
 If you need help contact us at vr-impute@sanger.ac.uk. Please let us know
```

Figure 6: Emel to validate well transfert to server



✓ Confirm data upload

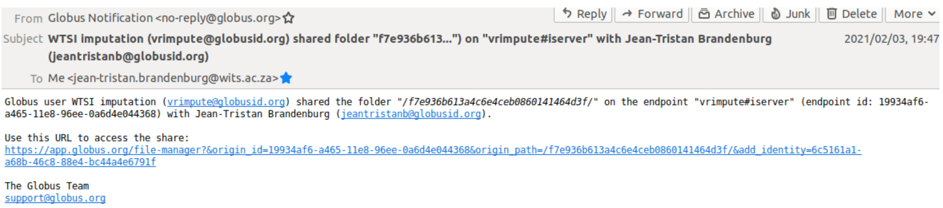
My data has been uploaded via Globus and can be scheduled for processing.

Project Name: Test

If you are not ready to start processing your data just yet, leave this page without clicking the Confirm button.

The Sanger Imputation Service is developed by the Vertebrate Resequencing Group at the Wellcome Sanger Institute
Copyright © 2015-2018 Genome Research Limited (reg no. 2742969) is a charity registered in England with number 1021457.
Terms and conditions | Cookies policy.

Figure 7: Validate your data after link emel



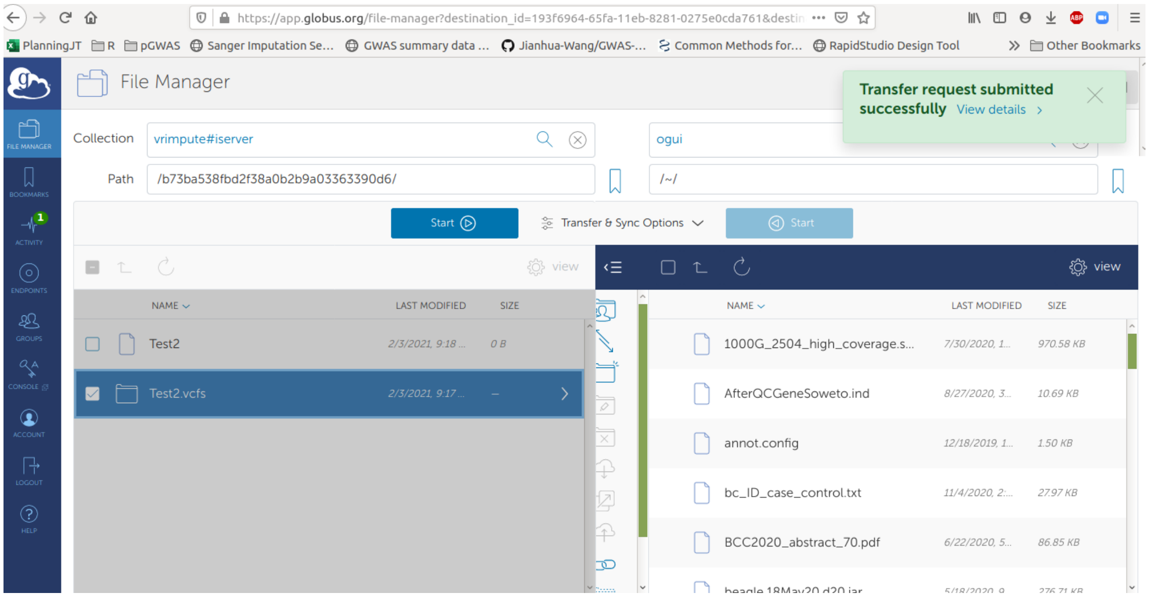
From Globus Notification <no-reply@globus.org>

Subject **WTSI imputation (vrimpute@globusid.org) shared folder "f7e936b613...") on "vrimpute#iserver" with Jean-Tristan Brandenburg (jeantristanb@globusid.org)** 2021/02/03, 19:47

To Me <jean-tristan.brandenburg@wits.ac.za>

```
Globus user WTSI imputation (vrimpute@globusid.org) shared the folder "/f7e936b613a4c6e4ceb0860141464d3f/" on the endpoint "vrimpute#iserver" (endpoint id: 19934af6-
a465-11e8-96ee-0a6d4e044368) with Jean-Tristan Brandenburg (jeantristanb@globusid.org).

Use this URL to access the share:
https://app.globus.org/file-manager?&origin_id=19934af6-a465-11e8-96ee-0a6d4e044368&origin_path=/f7e936b613a4c6e4ceb0860141464d3f/&add_identity=6c5161a1-
a68b-46c8-88e4-bc44a4e6791f

The Globus Team
support@globus.org
```

Figure 8: Mel validate that you data is ready on globus



Transfer request submitted successfully View details

Collection vrimpute#iserver

Path /b73ba538fbd2f38a0b2b9a03363390d6/

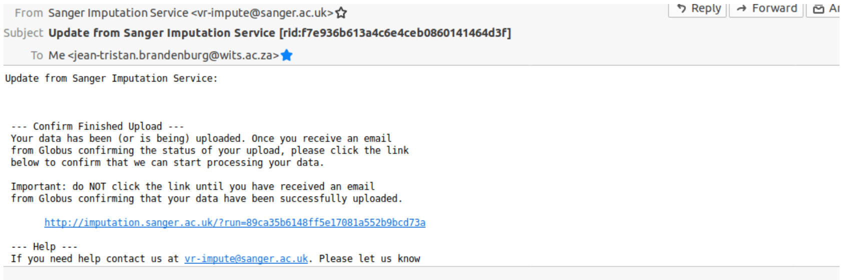
Collection ogui

Path /~/

NAME	LAST MODIFIED	SIZE
□ Test2	2/3/2021, 9:18 ...	0 B
☑ Test2.vcfs	2/3/2021, 9:17 ...	–

NAME	LAST MODIFIED	SIZE
1000G_2504_high_coverage.s...	7/30/2020, 1...	970.58 KB
AfterQCGeneSoweto.ind	8/27/2020, 3...	10.69 KB
annot.config	12/18/2019, 1...	1.50 KB
bc_ID_case_control.txt	11/4/2020, 2:...	27.97 KB
BCC2020_abstract_70.pdf	6/22/2020, 5...	86.85 KB
beagle.18May20.d20.jar	5/18/2020, 9...	276.71 KB

Figure 9: Transfer data from server to cluster



From Sanger Imputation Service <vr-impute@sanger.ac.uk>

Subject **Update from Sanger Imputation Service [rid:f7e936b613a4c6e4ceb0860141464d3f]**

To Me <jean-tristan.brandenburg@wits.ac.za>

```
Update from Sanger Imputation Service:



 --- Confirm Finished Upload ---
 Your data has been (or is being) uploaded. Once you receive an email
 from Globus confirming the status of your upload, please click the link
 below to confirm that we can start processing your data.

 Important: do NOT click the link until you have received an email
 from Globus confirming that your data have been successfully uploaded.

       http://imputation.sanger.ac.uk/?run=89ca35b6148ff5e17081a552b9bcd73a

 --- Help ---
 If you need help contact us at vr-impute@sanger.ac.uk. Please let us know
```

Figure 10: Validate your data after link emel