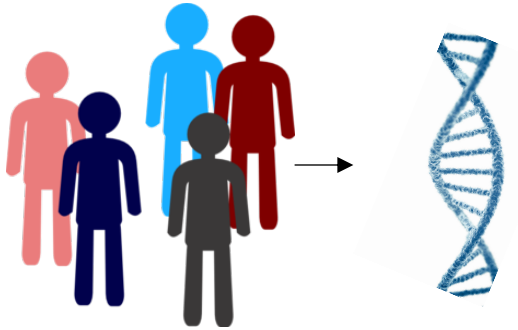
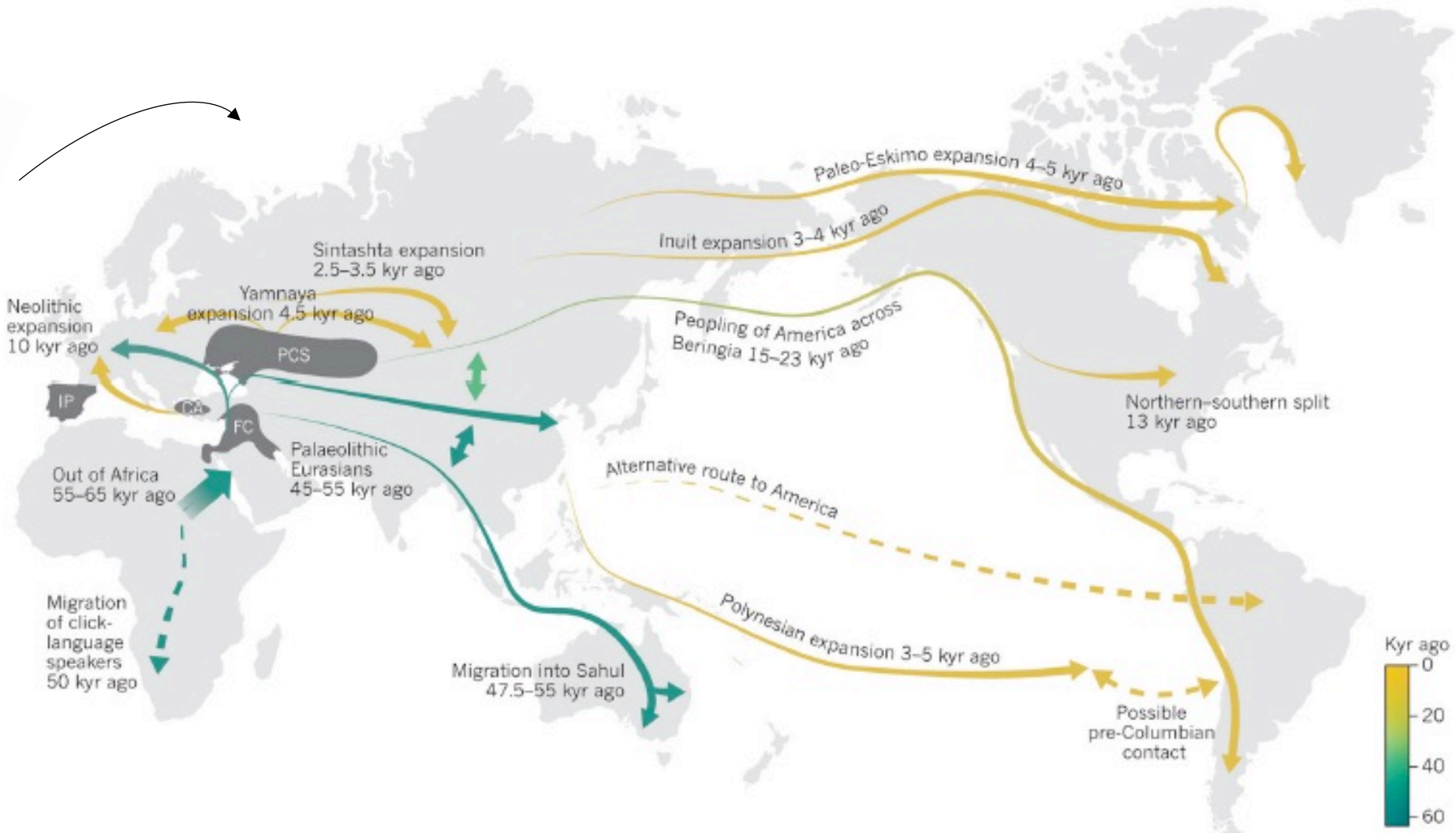
The role of OSG in advancing research in population genetics

Parul Johri
Postdoctoral Researcher
Jensen Lab
Arizona State University

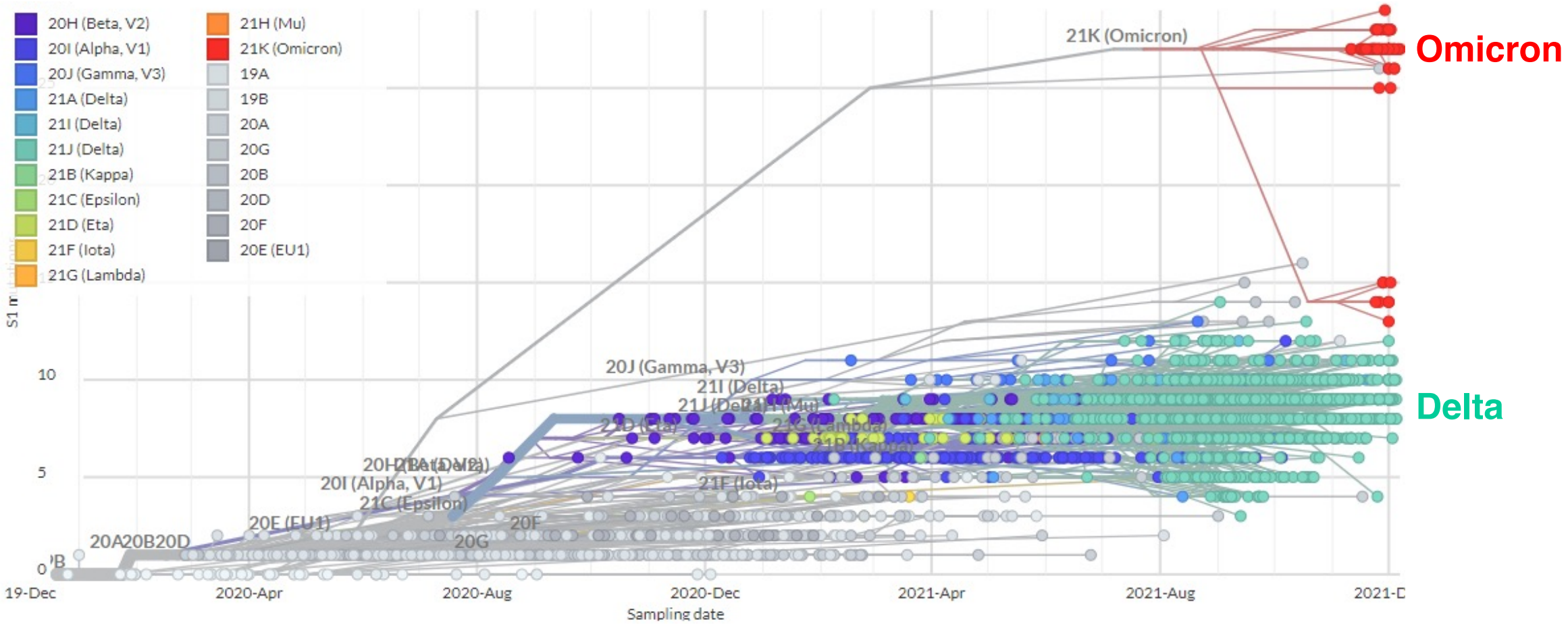
Genetic variation can yield insights into demographic history of natural populations



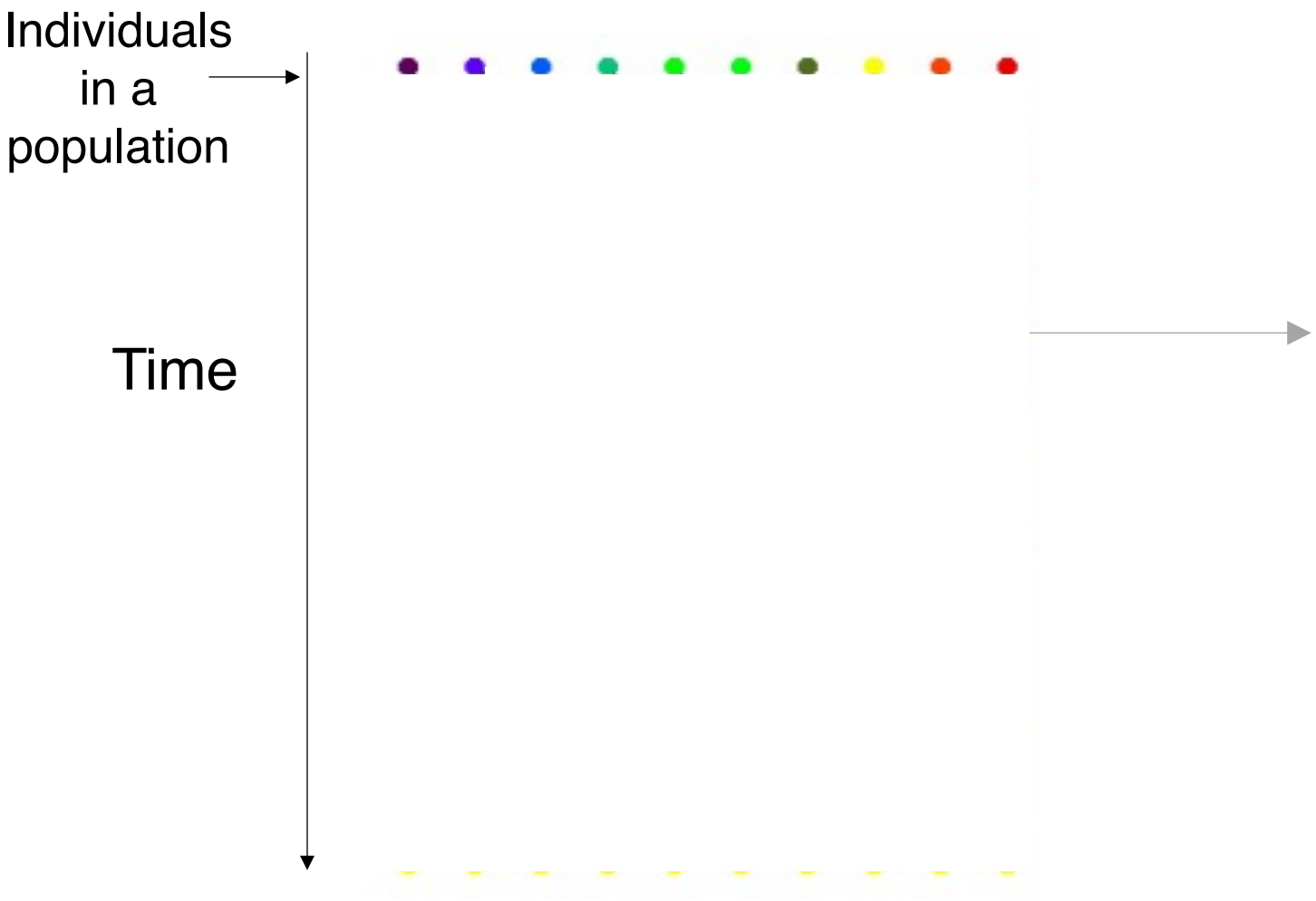


Nielsen *et al.*, 2017

Evolution of SARS-CoV-2 in human populations



Evolutionary forces that affect trajectories of natural populations



Evolutionary processes

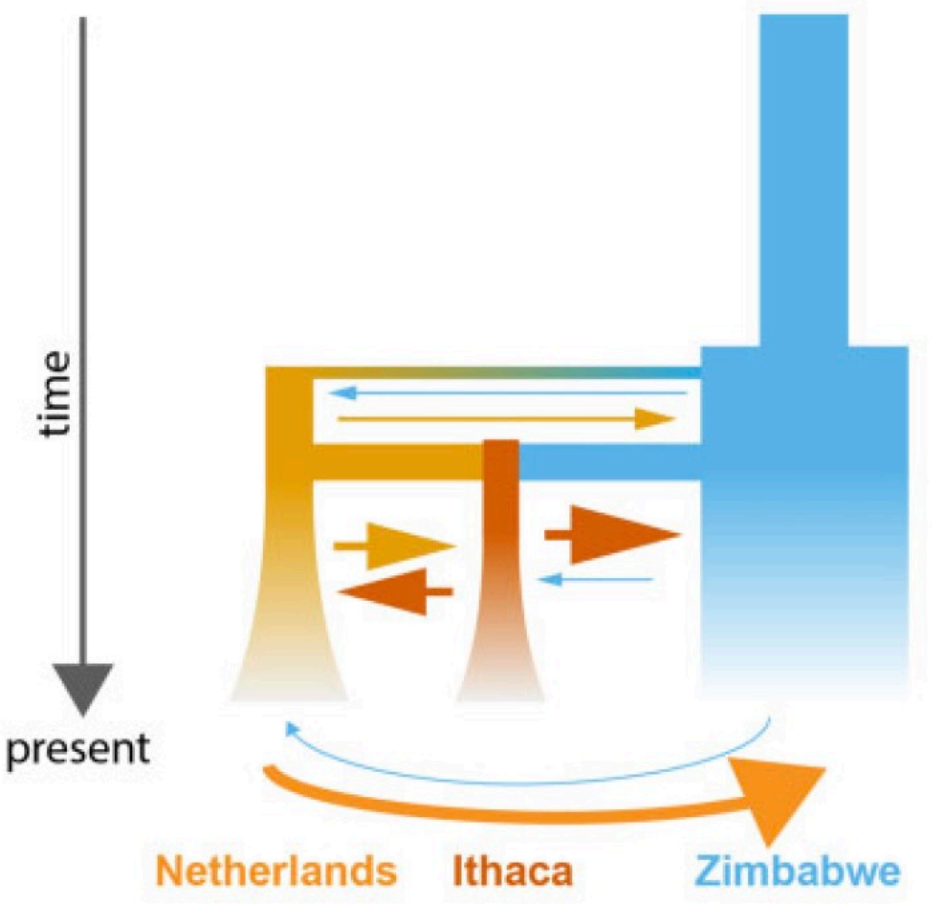
Selection

Mutation

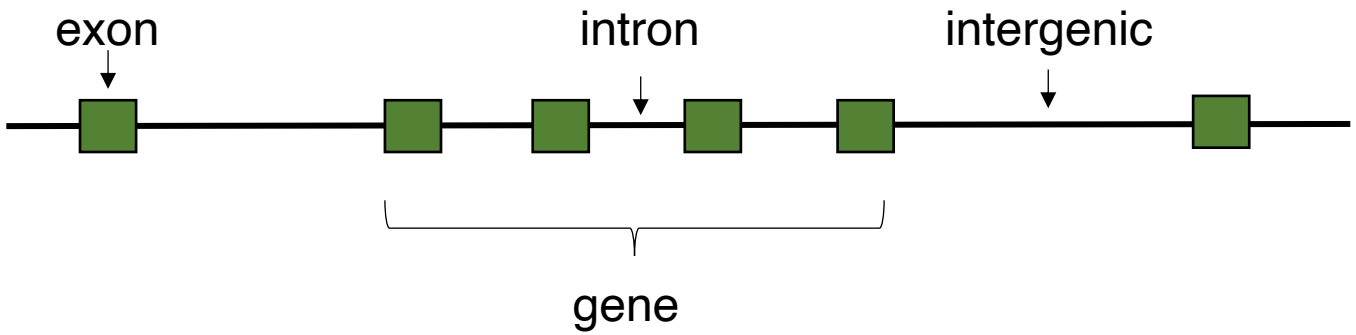
Recombination

Population size changes

Modeling realistic evolutionary dynamics in natural populations requires simulations

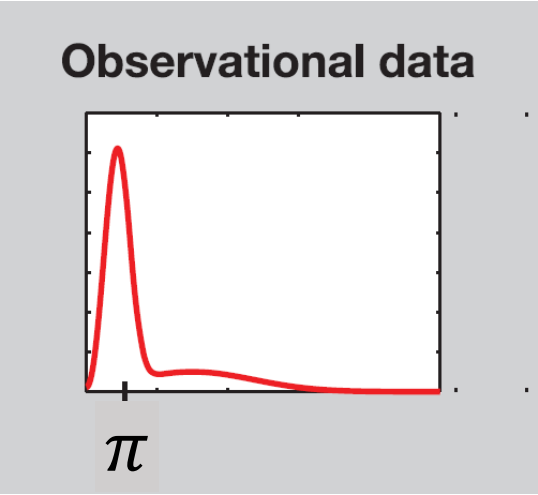


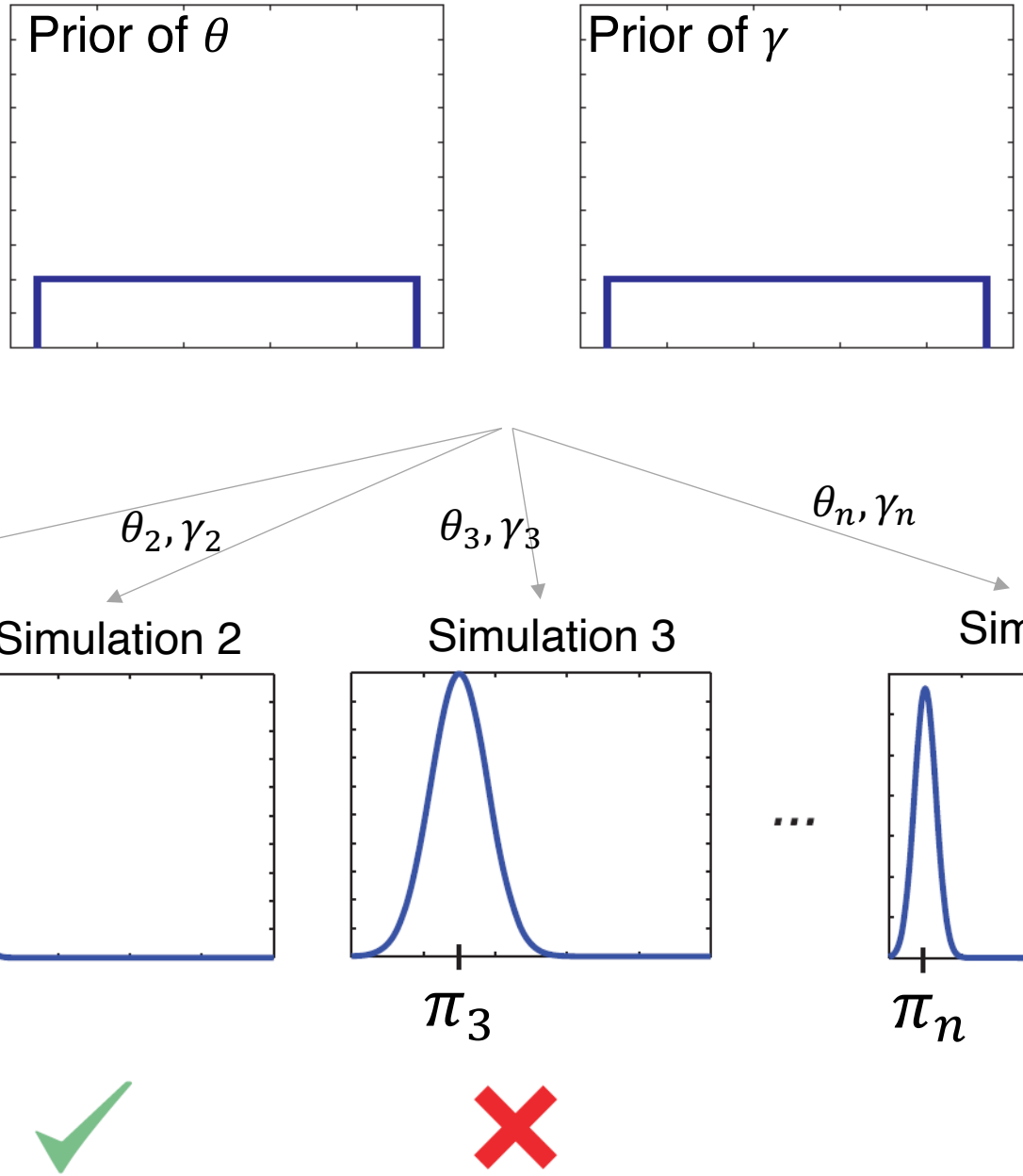
Arguello et al., 2019



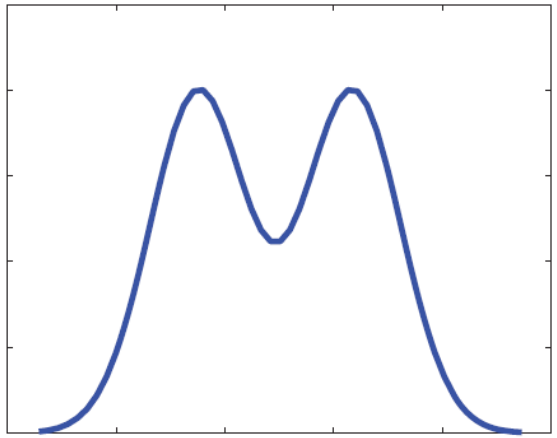
- Complex natural history: multiple populations with migration and population size changes
- Complex genomes: Non-uniform distribution of functionally important elements along our genomes.

Approximate Bayesian Computation (ABC)





Posterior distribution of θ and γ

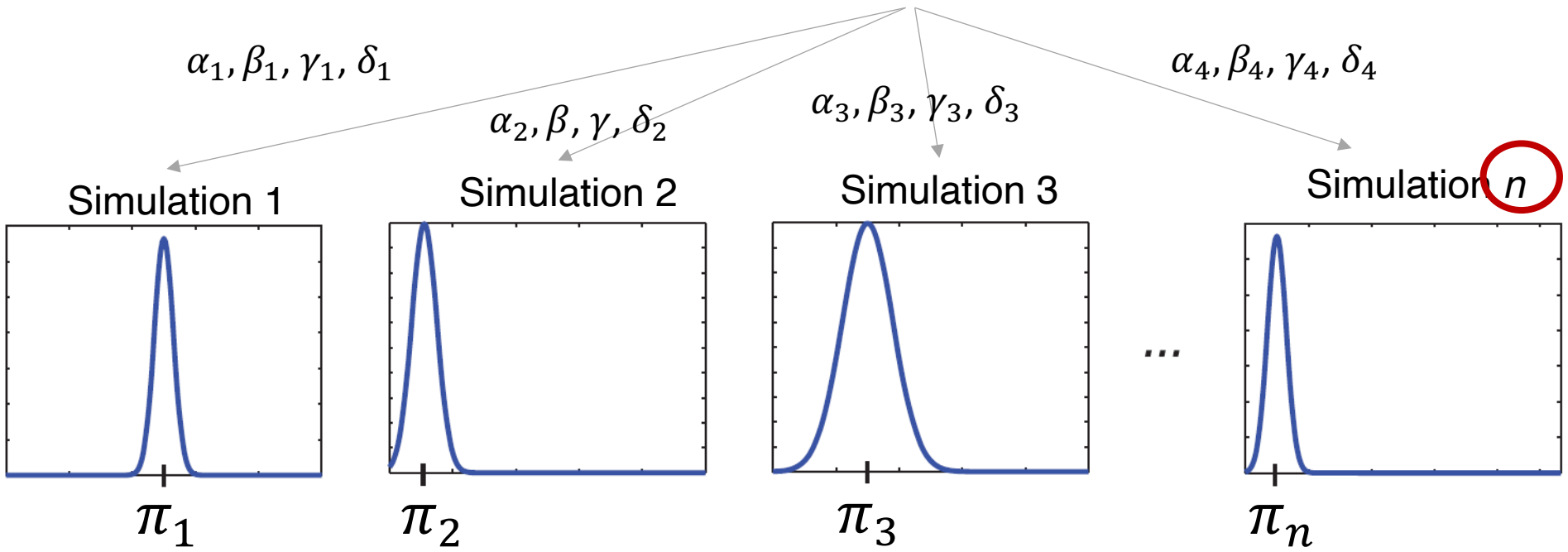


Adapted from Sunnaker *et al.* 2013

10 minutes per simulation (1000 reps, 3k combinations) = 500,000 hours!

Open Science Grid
HT Condor

Parameters of the model: $\alpha, \beta, \gamma, \delta$



3k – 1 Million

Simulation 1	Simulation 2	Simulation 3	Simulation n
Replicate 1	Replicate 1	Replicate 1	Replicate 1
.	.	.	.
.	.	.	.
Replicate 1000	Replicate 1000	Replicate 1000	Replicate 1000

How I discovered OSG

Gil Speyer
Arizona State University
Research Computing Staff

Lauren Michael
OSG Research Computing Facilitator
OSG

Quickstart page is very resourceful

https://support.opensciencegrid.org/support/solutions/articles/5000633410-osg-connect-quickstart

OSG Help Desk

Home | Solutions

How can we help you today?

Enter your search term here... SEARCH

Solution home / Managing HTC Workloads on OSG Connect / Submitting HTC Workloads with HTCondor

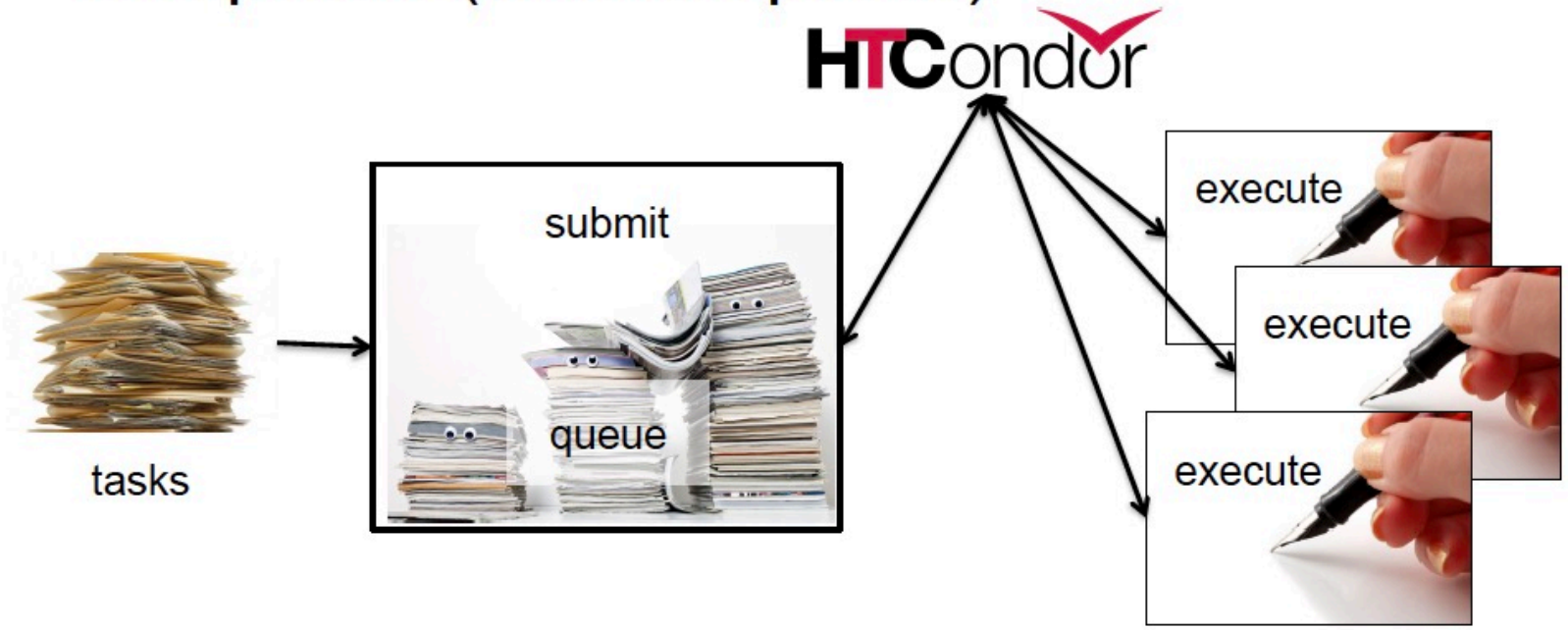
Quickstart - Submit Example HTCondor Jobs

Modified on: Tue, 8 Mar, 2022 at 1:58 PM

- Login to OSG Connect
  - Pretyped setup
  - Manual setup
- Tutorial jobs
- Job 1: A simple, nonparallel job
  - Run the job locally
  - Create an HTCondor submit file
  - More about projects
  - Submit the job
  - Check the job status
  - Job history
  - Check the job output
- Job 2: Passing arguments to executables
- Job 3: Submitting jobs concurrently
  - Where did jobs run?
- Removing jobs
- What's next?

Christina Koch's tutorial for introduction to OSG

- Submit tasks to a queue (on a submit point)
- HTCondor schedules them to run on computers (execute points)

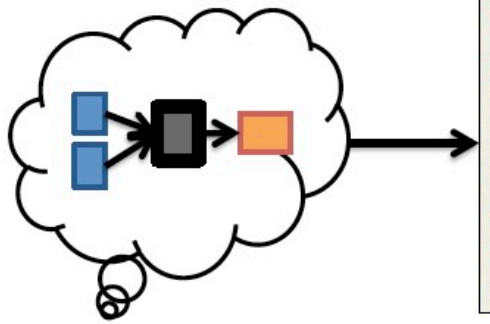


https://agenda.hep.wisc.edu/event/1325/contributions/5469/attachments/1421/1594/2019-Koch-User-Tutorial.pdf

Learning resources helpful for submitting jobs using HT Condor

Christina Koch's tutorial

- Submit file: communicates everything about your job(s) to HTCondor



```
executable = compare_states
arguments = wi.dat us.dat wi.dat.out

should_transfer_files = YES
transfer_input_files = us.dat, wi.dat
when_to_transfer_output = ON_EXIT

log = job.log
output = job.out
error = job.err

request_cpus = 1
request_memory = 20MB
request_disk = 20MB

queue 1
```

HTCondor

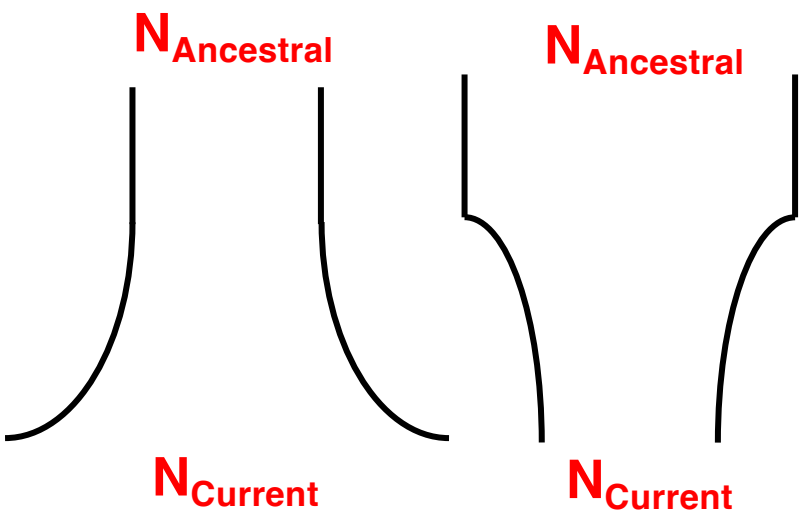
Other Links

- https://support.opensciencegrid.org/support/solutions/articles/5000633410-quickstart-submit-example-htcondor-jobs
- http://www.iac.es/sieinvens/siepedia/pmwiki.php?n=HOWTOs.CondorSubmitFile

Important and helpful HTCondor features and OSG support structures

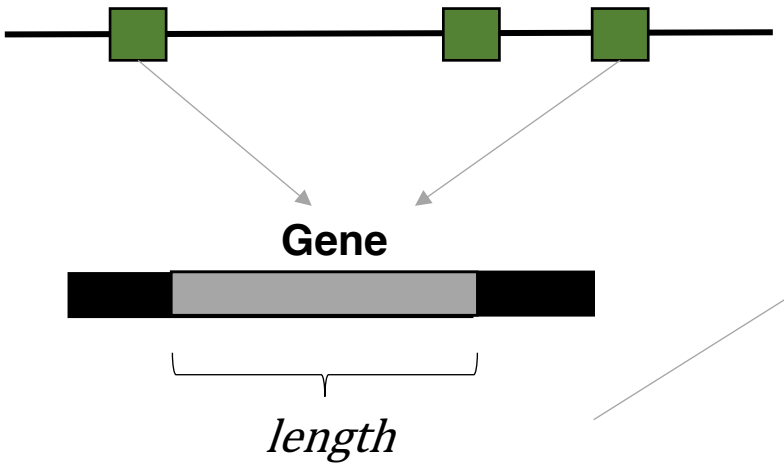
1 - High throughput simulations are made easy on HTCondor

Population parameters



SimNum

Genes across the genome



-mutation rate
-recombination rate

GeneNum

Evolutionary replicates

Replicate 1
Replicate 2
Replicate 3
.
.
.
Replicate 100

RepNum

How I run multiple jobs - my submit files

```
arguments = $(SimNum) $(GeneNum) $(RepNum) $(seed) $(scaling_factor) $(Nanc) $(growth_factor) $(s_f0) $(s_f1) $(s_f2) $(s_f3) $(RecRate) $(InterLen) $(ExonLen)

transfer_input_files = sim$(SimNum).csv, SingExon_slim_commands_human.sh, script_sim$(SimNum).slim

Queue SimNum,GeneNum,RepNum,seed,scaling_factor,Nanc,growth_factor,s_f0,s_f1,s_f2,s_f3,RecRate,InterLen,ExonLen from sim478.csv
```

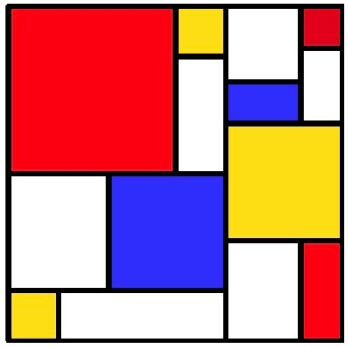
sim478.csv

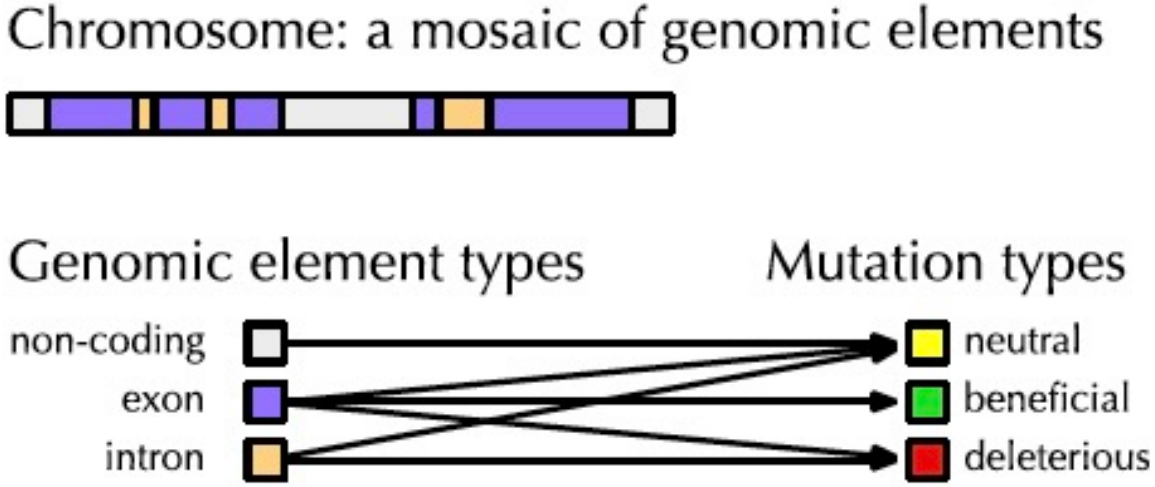
```
478,1,1,1,1.32,32207,0.990686270461,10,0,0,90,1.742,7424,2088
478,2,1,2,1.32,32207,0.990686270461,10,0,0,90,1.423,10044,2084
478,3,1,3,1.32,32207,0.990686270461,10,0,0,90,3.225,2872,2500
478,4,1,4,1.32,32207,0.990686270461,10,0,0,90,2.879,3084,2041
```

2 - OSG allows accessibility to containers

https://support.opensciencegrid.org/support/solutions/articles/12000024676-use-containers-on-the-osg

SLiM: a simulator of population-genetic data





- Uses a single core
- Does not require an input file
- Requires only the set of parameters as options to the script.

Christina Koch
Open Science Grid

3 - Data storage at OSG is fantastic

Data Locations and Quotas

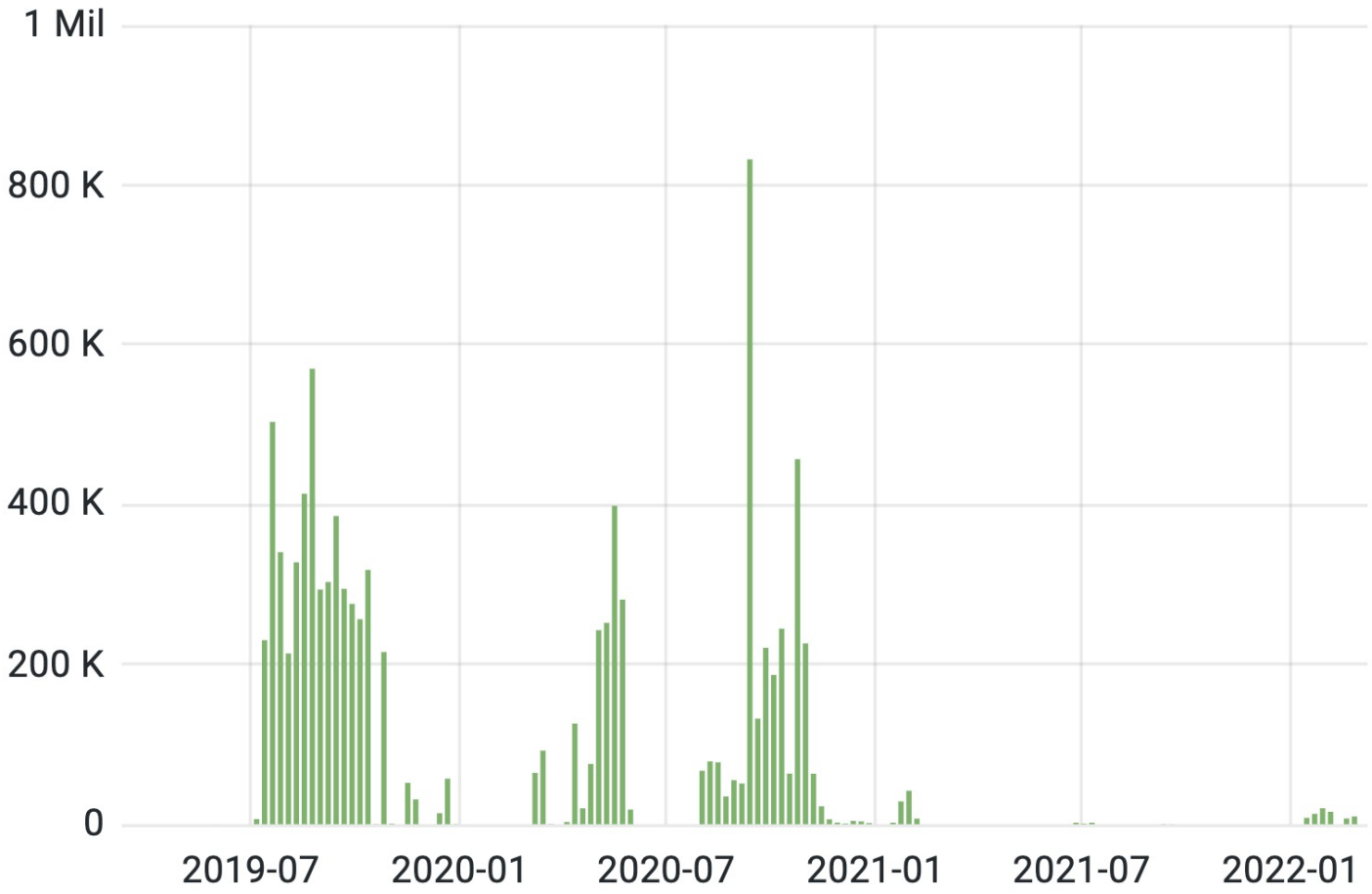
Your OSG Connect account includes access to two data storage locations: `/home` and `/public`. **Where you store your files and how your files are made accessible to your jobs depends on the size of the file and how much data is needed or produced by your jobs.**

Location	Storage Needs	Network mounted	Backed Up?	Initial Quota
`/home`	Storage of submit files, input files <100MB each, and per-job output up to a 1GB. Jobs should ONLY be submitted from this folder.	No	No	50 GB
`/public`	Staging ONLY for large input files (100MB-50GB, each) for publicly-accessible download into jobs (using HTTP or stashcp, see below) and large output files (1-10GB).	Yes	No	500 GB

Required computing time for my projects

	Time for 1 replicate	Number of replicates	Number of parameter combinations	Total time
Drosophila	~20 min	1000	3000	1 million hours
Influenza	~3 hours	100	500	150,000 hours
Humans	~10 min	500	3000	250,000 hours

OSG usage in the last 3 years



9.66 million
Wall Hours

Manuscripts published where OSG was a key computing resource

- Ana Yansi Morales-Arce*, **Parul Johri***, Jeffrey D. Jensen. 2022. **Inferring the distribution of fitness effects in influenza A virus and human cytomegalovirus**. *Heredity* 128, 79–87 (2022).

- **Parul Johri**, Kellen Riall, Hannes Becher, Laurent Excoffier, Brian Charlesworth, Jeffrey D. Jensen. 2021. **The impact of purifying and background selection on the inference of population history: problems and prospects**. *Molecular Biology and Evolution*. 38(7): 2986-3003.

- **Parul Johri**, Brian Charlesworth, Jeffrey D. Jensen. 2020. **Towards an evolutionarily appropriate null model: jointly inferring demography and purifying selection**. *Genetics*. 215: 173-192.

The importance of OSG in the future for population genetics

- We do not yet have mathematical expressions describing patterns of genetic variation in populations with complex population history and selection acting on linked genes (on a chromosome).
- Simulations are a great way to understand complex evolutionary scenarios.
- Simulation frameworks in population genetics are becoming faster and more efficient.
- OSG can play a central role in performing evolutionary inferences.

THANK YOU

Acknowledgements

Open Science Grid

Advisor:
Jeffrey Jensen, Arizona State University

Gil Speyer
Arizona State University
Research Computing Staff

Lauren Michael
OSG Research Computing Facilitator
OSG

Christina Koch
Open Science Grid