HPC in Scripps Research - Garibaldi (high-performance computing)

Shaun Chen

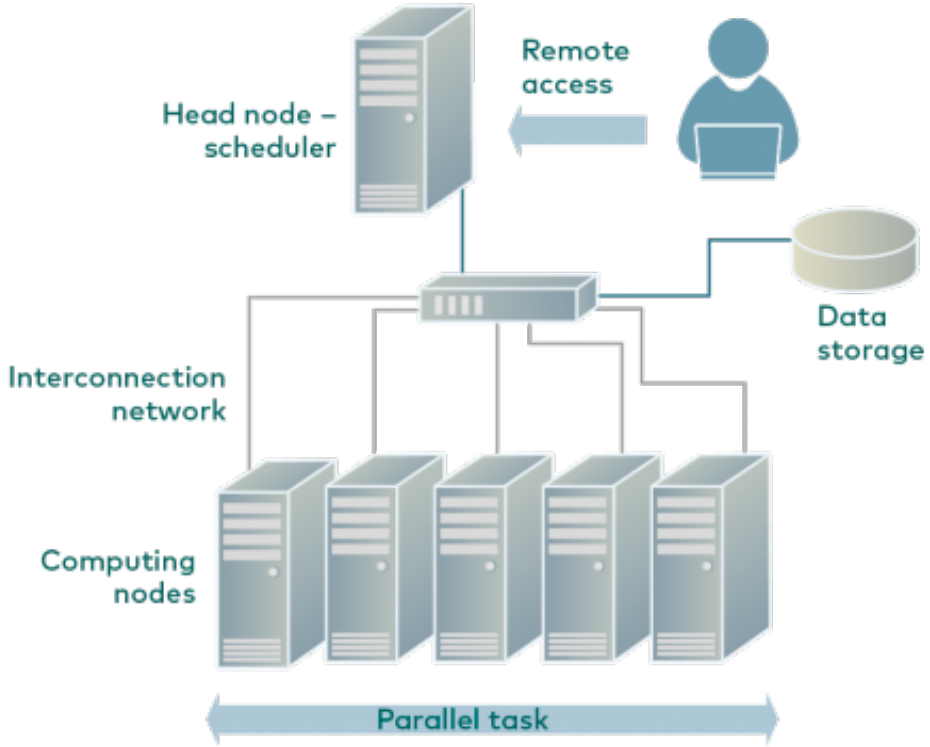
Date: June 29th, 2021

Handout https://github.com/ShaunFChen/2020_FA_HPC

Contents

Concepts of HPC

What is High-Performance Computing (HPC)?



https://hpc.rtu.lv/hpc/introduction-to-hpc/?lang=en

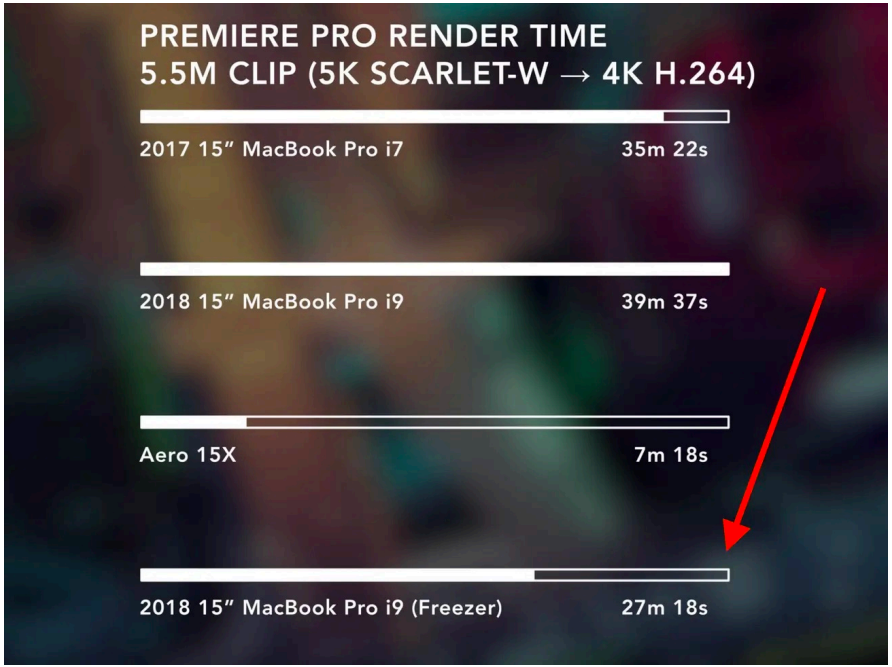
Why High-Performance Computing (HPC)?

1. **Application can not be run on a conventional (serial) computing system**
   - Insufficient memory
   - Insufficient computational power

2. **HPC generally now means:**
   - Large multi-processor system
   - Complex communications hardware
   - Specialized attached processors
   - GRID/Cloud computing

Your laptop might not design for heavy duty…

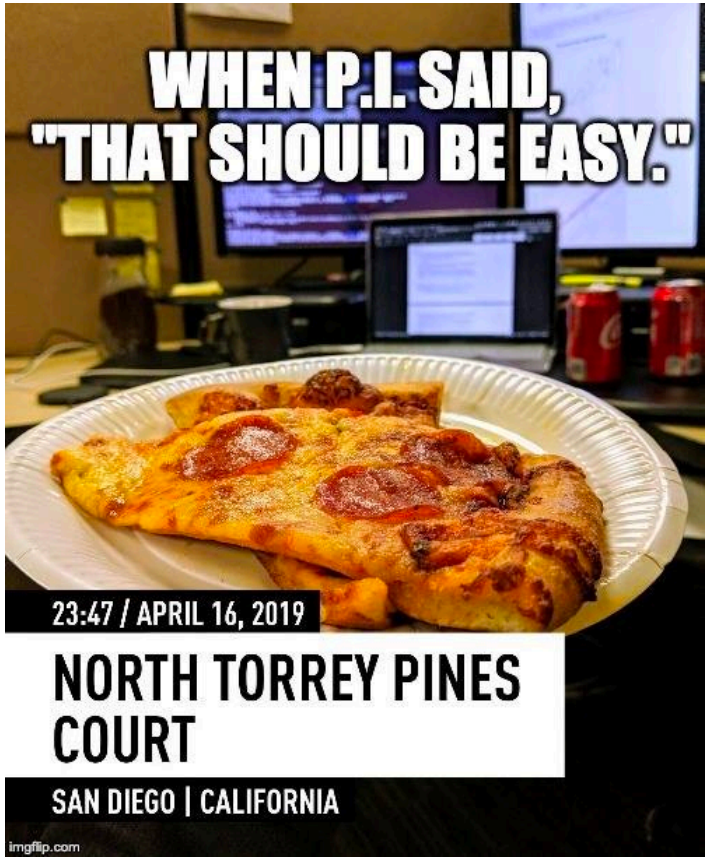
YouTube | Dave2D



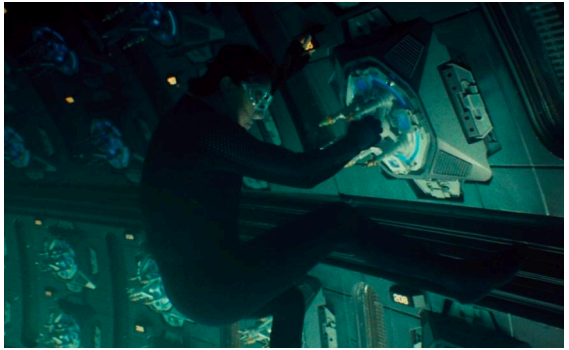
Apple fans are returning their new MacBook Pros that cost a minimum of $2,800 because they can't reach the advertised speeds
https://www.businessinsider.com/apple-macbook-pro-2018-core-i9-issues-complaints-returns-2018-7

Things took longer than you expected…

u/rueger_c, “Jumping over a snowball.” *Reddit* (2013), https://www.reddit.com/r/gifs/comments/1r84w0/jumping_over_a_snowball/



~~The evolution of supercomputer~~

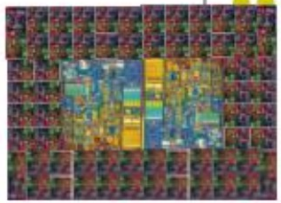


Prime Video https://www.amazon.com/Mission-Impossible-Tom-Cruise/dp/B000X4IRE4

The PlayStation Supercomputer: The Condor Cluster https://www.datacenterdynamics.com/en/analysis/the-playstation-supercomputer/

Micro2015: Evolving Processor Architecture, Intel® Developer Forum, **March 2005**

Intel's Vision: Evolutionary Configurable Architecture

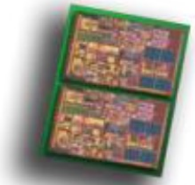


Scalar plus many core for highly threaded workloads

Large, Scalar cores for high single-thread performance

Multi-core array

- CMP with ~10 cores



Dual core

- Symmetric multithreading

Many-core array

- CMP with 10s-100s low power cores
- Scalar cores
- Capable of TFLOPS+
- Full System-on-Chip
- Servers, workstations, embedded…

Evolution

CMP = "chip multi-processor"

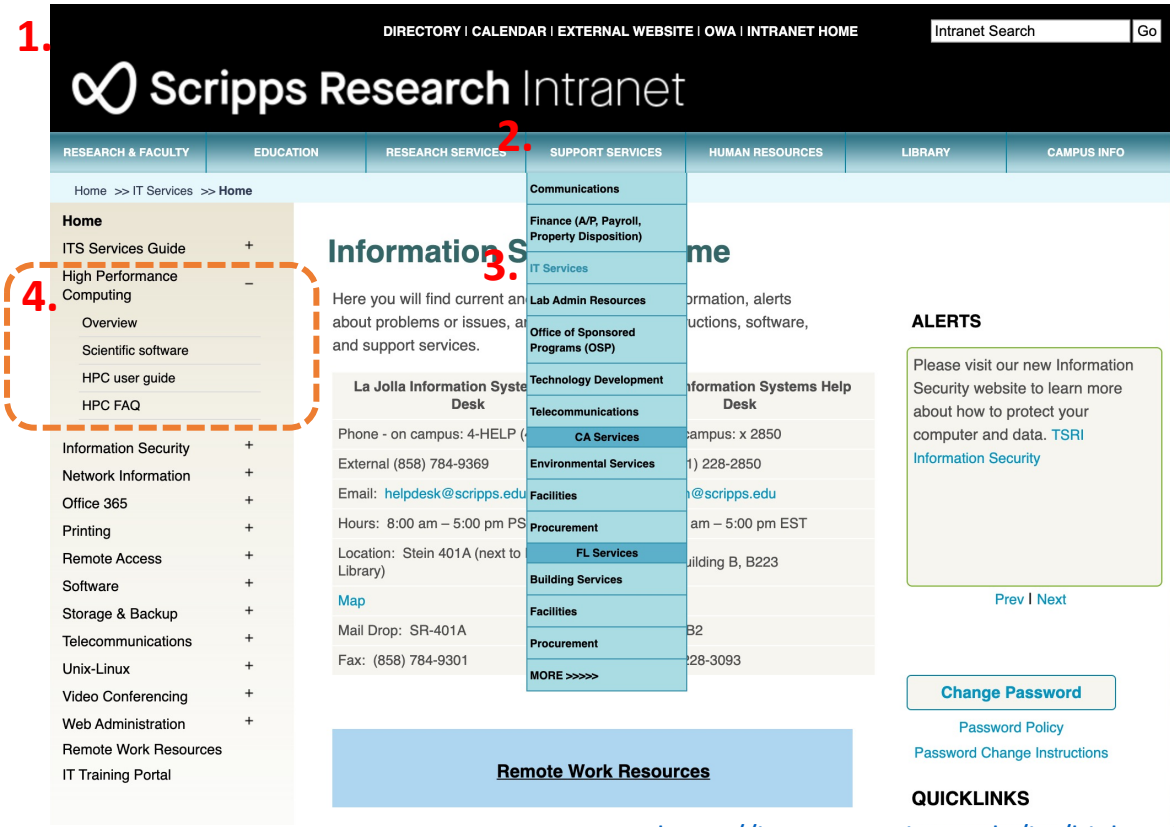
intel

Presentation Paul Petersen, Sr. Principal Engineer, Intel

Hands-on practice on Garibaldi

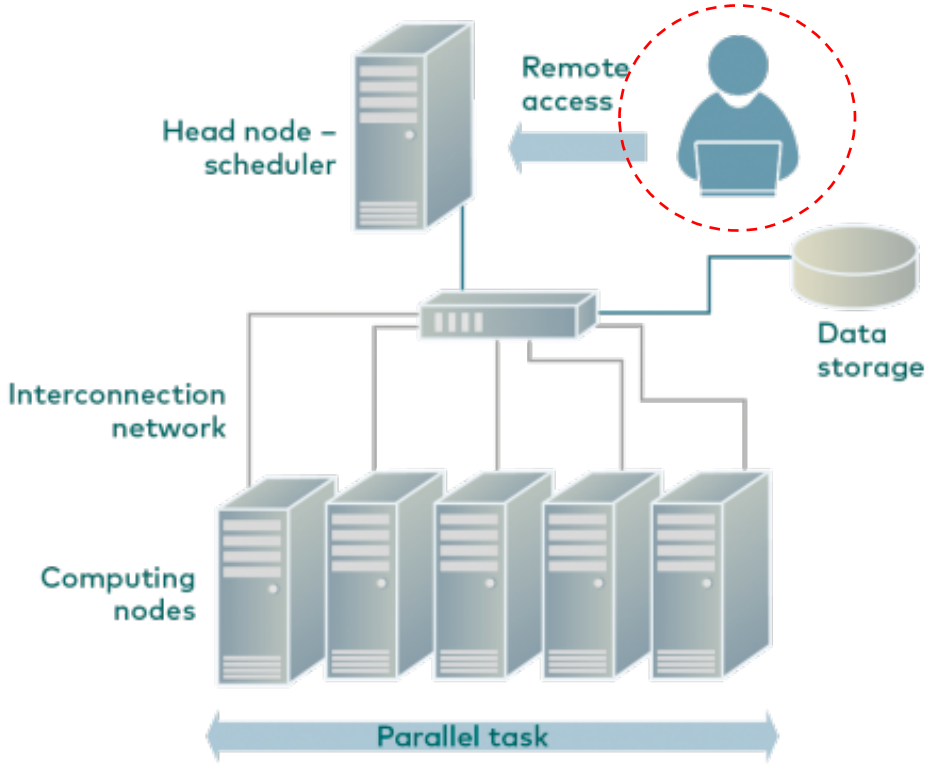
HPC in Scripps Research - Garibaldi

Documentations in Scripps intranet



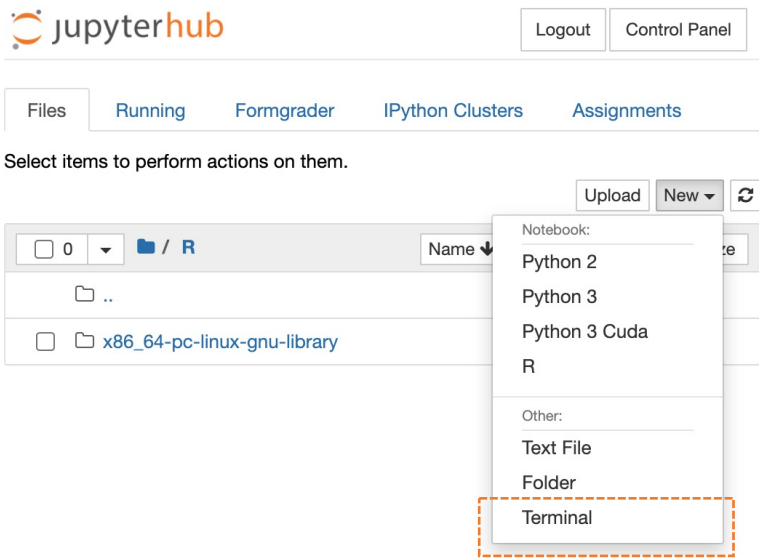
https://intranet.scripps.edu/its/highperformancecomputing/index.html

From local to remote



https://hpc.rtu.lv/hpc/introduction-to-hpc/?lang=en

Login to the head (login) node by ssh



https://hpcweb01.scripps.edu:8000/hub/login
https://ims02.scripps.edu:8000/hub/login

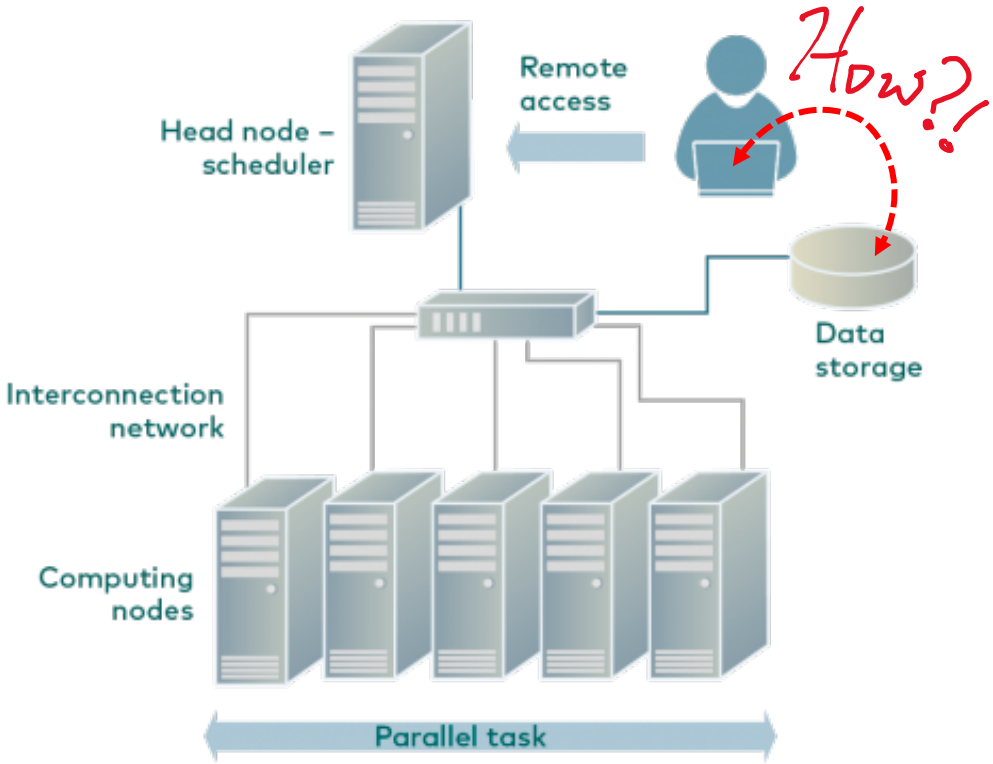
```
bash-4.1$ ssh [username]@login00.scripps.edu
bash-4.1$ ssh [username]@login02.scripps.edu
The authenticity of host 'login00.scripps.edu (172.29.83.22)'
can't be established.
RSA key fingerprint is
93:6c:ae:7a:7a:00:00:00:00:f9:f3:00:00:00:a9:00.
Are you sure you want to continue connecting (yes/no)? Yes

instructor-abcb2020@login00.scripps.edu's password:
Permission denied, please try again.

instructor-abcb2020@login00.scripps.edu's password:

Last login: Fri Oct 23 01:56:11 2020 from vpn-240-175.scripps.edu

*********************************************
*                                           *
*       Please contact hpc_ca@scripps.edu   *
*                                           *
* if you experience any problems with Garibaldi *
*                                           *
* ( applications, slowness, scheduling, etc...) *
*                                           *
*********************************************

[instructor-abcb2020@login00-adm instructor-abcb2020]$
```

Alternative: local terminal emulator

```
bash-4.1$ ssh [username]@login00.scripps.edu
bash-4.1$ ssh [username]@login02.scripps.edu
```

Make your terminal beautiful and fast with ZSH shell and PowerLevel10K
https://medium.com/@shivam1/make-your-terminal-beautiful-and-fast-with-zsh-shell-and-powerlevel10k-6484461c6efb

Work remotely… Say bye to GUI!!!



https://hpc.rtu.lv/hpc/introduction-to-hpc/?lang=en

1. scp

```
# upload single file to remote server
Local $ scp foo.txt user@login00.scripps.edu:~/

# download wildcard/folder from remote server
Local $ scp –r `user@login00.scripps.edu:~/bar/*` ./
```

2. git

```
# create a new repo
Local $ git init # follow the instructions
```

```
# add commit when own code changed
Local $ git add .

# record own commit message
Local $ git commit –m “message”

# push the own code change
Local $ git push
```

```
# load git module
Remote $ module load git-lfs
```

```
# clone the repo
Remote $ git clone https://...git
```

```
# pull change when own code updated
Remote $ git pull
```

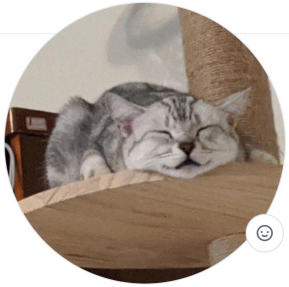
git init (I)

Search or jump to... / Pull requests Issues Marketplace Explore

1.

- New repository
- Import repository
- New gist
- New organization
- New project

Overview | Repositories 19 | Projects | Packages



Shaun Chen
ShaunFChen

Graduate student working on Human Population Genetics and Genomics

Edit profile

5 followers · 6 following · ☆ 41

@TorkamaniLab
La Jolla, CA
https://scholar.google.com/citations?...

Organizations

Popular repositories — Customize your pins

Applied-Bioinformatics-HW
Applied-Bioinformatics-HW
HTML

Applied-Bioinformatics
Forked from SuLab/Applied-Bioinformatics
Student course material for the Applied Bioinformatics course at Scripps Research
Jupyter Notebook

readme-template
Forked from dbader/readme-template
README.md template for your open-source project

CBB_Parallel_Multiprocessing
Scripps CBB CodeTopics - Basic GNU Parallel and Python Multiprocessing
Jupyter Notebook

Applied_bioinfo_nbgrader
Instructions to build nbgrader on Garibaldi @ Scripps Research
Jupyter Notebook

267 contributions in the last year — Contribution settings

Nov Dec Jan Feb Mar Apr May Jun Jul Aug Sep Oct
Mon
Wed
Fri

Learn how we count contributions. Less More

2020
2019
2018

git init (II)

Search or jump to... / Pull requests Issues Marketplace Explore

Create a new repository

A repository contains all project files, including the revision history. Already have a project repository elsewhere? Import a repository.

2.

Owner * ShaunFChen / Repository name *

Great repository names are short and memorable. Need inspiration? How about shiny-octo-memory?

Description (optional)

Public
Anyone on the internet can see this repository. You choose who can commit.

Private
You choose who can see and commit to this repository.

Initialize this repository with:
Skip this step if you're importing an existing repository.

Add a README file
This is where you can write a long description for your project. Learn more.

Add .gitignore
Choose which files not to track from a list of templates. Learn more.

Choose a license
A license tells others what they can and can't do with your code. Learn more.

Create repository

git init (III)

Search or jump to... / Pull requests Issues Marketplace Explore

Learn Git and GitHub without any code!

Using the Hello World guide, you'll start a branch, write comments, and open a pull request.

Read the guide

ShaunFChen / 2020_FA_HPC

Unwatch 1 Star 0 Fork 0

<> Code Issues Pull requests Actions Projects Wiki Security Insights Settings

Quick setup — if you've done this kind of thing before

Set up in Desktop or HTTPS SSH git@github.com:ShaunFChen/2020_FA_HPC.git

Get started by creating a new file or uploading an existing file. We recommend every repository include a README, LICENSE, and .gitignore.

...or create a new repository on the command line

3.

```
echo "# 2020_FA_HPC" >> README.md
git init
git add README.md
git commit -m "first commit"
git branch -M main
git remote add origin git@github.com:ShaunFChen/2020_FA_HPC.git
git push -u origin main
```

NOTE: Github on Garibaldi

- **Load the latest git**

```
[xxx@login02 2020_FA_HPC]$ module load git
[xxx@login02 2020_FA_HPC]$ git --version
git version 2.30.0
```

- **Generate a new SSH key**
  https://docs.github.com/en/github/authenticating-to-github/connecting-to-github-with-ssh/generating-a-new-ssh-key-and-adding-it-to-the-ssh-agent

Basic commands & variables

- `showuserjobs`
- `squeue –u [username]`
- `scontrol show job [job_id]`
- `seff –f [job_id]`
- `srun --pty bash -i`
- `sbatch --mem=`
  `--time=`
  `--job-name=`
  `--export=(var_A),(var_B),(var_C)`
- `scancel [job_id]`

- `module`
  - `module load`
  - `module unload`
  - `module purge`
  - `module av`

showuserjobs

```
[sfchen@login02 icld]$ showuserjobs
Batch job status for cluster garibaldi at Mon Jun 21 13:27:44 PDT 2021

Node states summary:
alloc     156 nodes (2708 CPUs)
down        1 nodes (20 CPUs)
idle      103 nodes (1756 CPUs)
mix        24 nodes (396 CPUs)
resv        1 nodes (16 CPUs)
Total     285 nodes (4896 CPUs)

Job summary: 420 jobs total (max=75000) in all partitions.

Username/                          Runnin            Idle
Totals            Account            Jobs   CPUs     Jobs   CPUs  Further info
================ ============== ====== ======   ====== ======  ==============================
GRAND_TOTAL       ALL                 360   2996       60    396  Running+Idle=3392 CPUs, 26 users
ACCT_TOTAL        grads               125    589       13    208  Running+Idle=797 CPUs, 6 users
sfchen            grads                20     35        0      0  Shang-Fu Chen
```

Simple PBS script template

```
#!/bin/bash
#SBATCH --time=24:00:00
#SBATCH --mem=16G
## SBATCH --nodes=1                    ### Node count required for the job
## SBATCH --ntasks=1                   ### Number of tasks to be launched per Node
## SBATCH --cpus-per-task=16

# sbatch --export=month=,day= run.slurm


# slurm will set the working directory with the job script by default.
module purge
module load R

# always print the path of your working directory
pwd

echo “Hello HPC!! Today is ${month} ${day}”
```

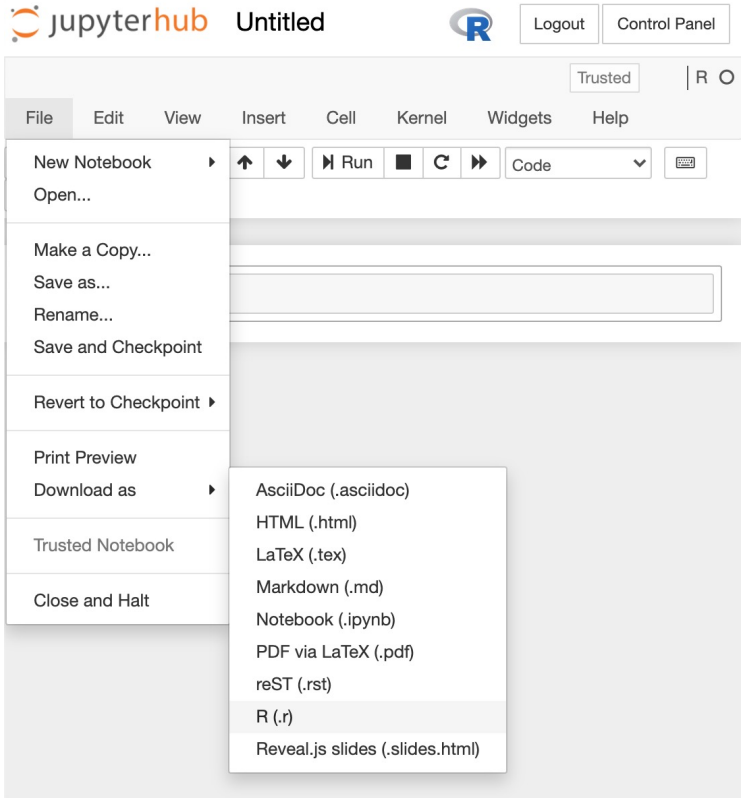
Jupyter notebook vs executable script

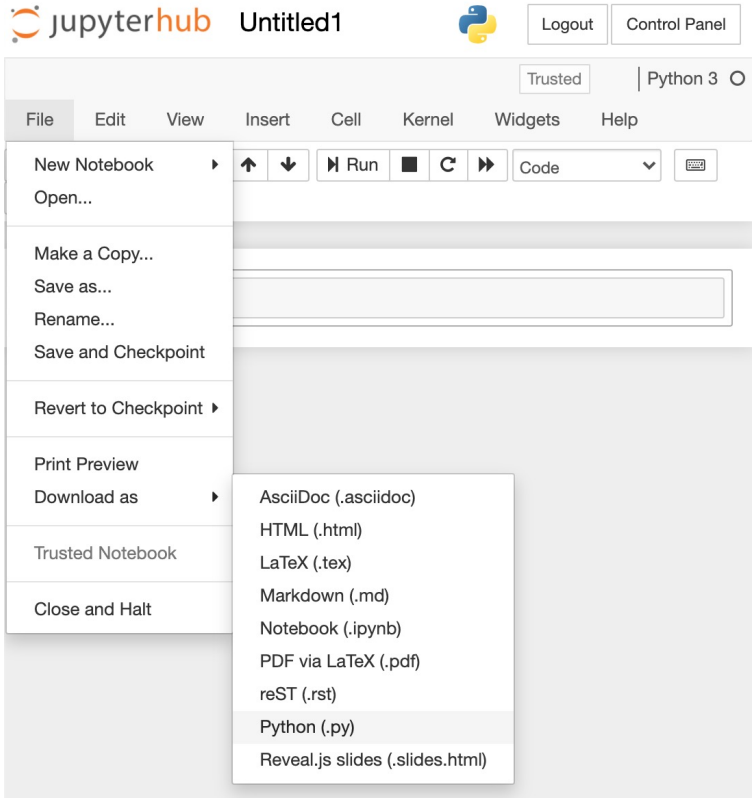

```r
library(ggplot2)

eigenvec_fp <- commandArgs()[3]
pop_fp <- commandArgs()[4]

df <- read.csv(eigenvec_fp, sep = ' ', header = 0)
pop <- read.csv(pop_fp, sep = '\t')
df_ <- merge(df, pop, by.x = 'V1', by.y = 'IID')

png("figure_tgp_pca.png", width = 6, height = 6, units = "in", res = 300)

print({
    ggplot(df_, aes(x = V3, y = V4, color = sup_pop)) +
      geom_point(size = 1, alpha = 0.5) + xlab("PC1") + ylab("PC2")
})

dev.off()


print("output generated: figure_tgp_pca.png")
write.table(df_, file = "pca_dataframe.csv", sep = ',')
```

https://towardsdatascience.com/5-reasons-why-you-should-switch-from-jupyter-notebook-to-scripts-cb3535ba9c95

```
[sfchen@hpcweb01-adm 2020_FA_HPC]$ R --no-save TGP_pca.eigenvec data/1000G.tsv < required_tools/tgp_pca.r
[sfchen@hpcweb01-adm 2020_FA_HPC]$ ls *.png
figure_tgp_pca.png
```

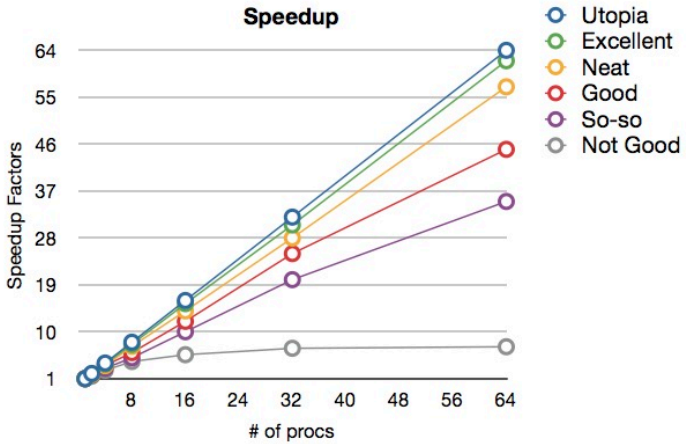
Export notebook to executable script

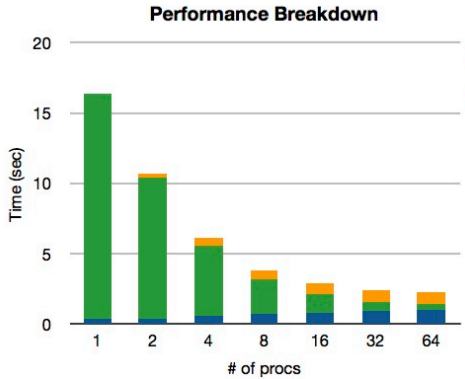


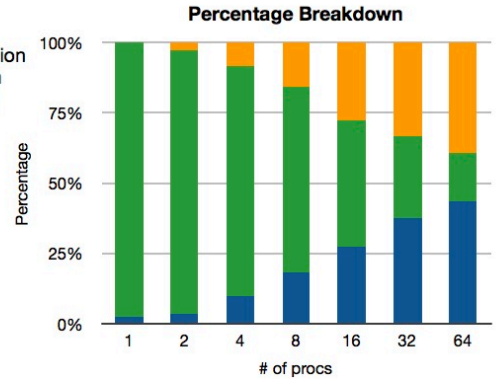


More jobs ≠ faster

- Allocate the proper resource
- Optimize your code
- Be patient…







HPC Challenges — A Perspective for General Data Analysis and Visualization
http://web.eecs.utk.edu/~huangj/hpc/hpc_intro.php

Take Home Message - Caution

- **Caution**
  - Never compute at login node
  - Never use root permission (no `sudo`)
- **Max job number**
  - started from 10
  - Put `` `sleep 3` `` between loop of job submission
- **Do**
  - Backup your projects frequently
  - Respect other users (shared space and privacy)
  - Be responsible/responsive for your behavior
  - Inform your PI for data management
- **Be friendly with hpc_ca@scripps.edu**

Take Home Message – Pipeline Development

- **First draft**
  - Interactive debugging (Jupyterhub; Rstudio; other editers, IDE, etc.)
    - Shut down the session when leaving…
  - Test case
- **Form executable script**
  - Unit test

```
If you in ["standardize multiple steps processing",
           "job couldn't finish within feasible time",
           "local machine overheated",
           "wanna go home earlier"]:
```

- **Build pipeline**
  - Print (echo) checkpoint message; Save human readable logging output
  - Allocate proper resource
  - Merge or split jobs for optimized granularity
  - Remove temp files when jobs finished
  - Fully understand the working concept of the tools
    - Do not treat the tools as black boxes
    - Read background article or even source code if needed
  - Speed: C/Java > JS/Python/Perl > R

Resources

- **2020_FA_HPC** https://github.com/ShaunFChen/2020_FA_HPC
- **GNU Parallel and Multiprocessing**

https://github.com/ShaunFChen/CBB_Parallel_Multiprocessing

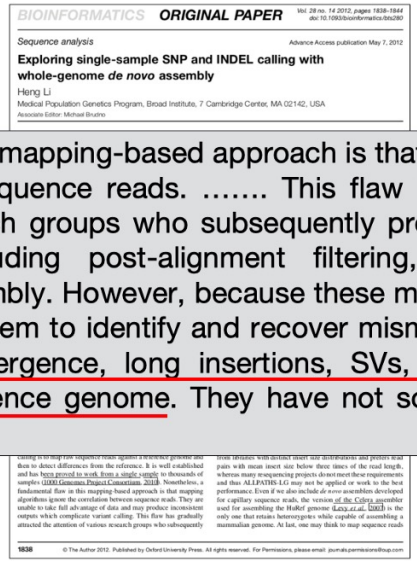
- **Scripps Intranet**

https://intranet.scripps.edu/its/highperformancecomputing/index.html

- **Scripps Libraries**

https://www.scripps.edu/science-and-medicine/cores-and-services/library/index.html

Be aware of the limitation of your tools.



BIOINFORMATICS **ORIGINAL PAPER**

Sequence analysis

Exploring single-sample SNP and INDEL calling with whole-genome *de novo* assembly

Heng Li

Medical Population Genetics Program, Broad Institute, 7 Cambridge Center, MA 02142, USA

A fundamental flaw in this mapping-based approach is that mapping algorithms ignore the correlation between sequence reads. ……. This flaw has gradually attracted the attention of various research groups who subsequently proposed several methods to alleviate the effect, including post-alignment filtering, iterative mapping, read realignment and local assembly. However, because these methods still rely on the initial mapping, it is difficult for them to identify and recover mismapped or unmapped reads due to high-sequence divergence, long insertions, SVs, copy number changes or misassemblies of the reference genome. They have not solved the problem from the root.

Medium | @chungtsai

Fast and accurate short read alignment with Burrows–Wheeler transform
H Li, R Durbin - bioinformatics, 2009 - academic.oup.com
Motivation: The enormous amount of short reads generated by the new DNA sequencing technologies call for the development of fast and accurate read alignment programs. A first generation of hash table-based methods has been developed, including MAQ, which is accurate, feature rich and fast enough to align short reads from a single individual. However, MAQ does not support gapped alignment for single-end reads, which makes it unsuitable for alignment of longer reads where indels may occur frequently. The speed of MAQ is also a …
☆ 99 Cited by 25804 Related articles All 40 versions »

Fast and accurate long-read alignment with Burrows–Wheeler transform
H Li, R Durbin - Bioinformatics, 2010 - academic.oup.com
Motivation: Many programs for aligning short sequencing reads to a reference genome have been developed in the last 2 years. Most of them are very efficient for short reads but inefficient or not applicable for reads> 200 bp because the algorithms are heavily and specifically tuned for short queries with low sequencing error rate. However, some sequencing platforms already produce longer reads and others are expected to become available soon. For longer reads, hashing-based software such as BLAT and SSAHA2 …
☆ 99 Cited by 6972 Related articles All 24 versions »

Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM
H Li - arXiv preprint arXiv:1303.3997, 2013 - arxiv.org
Summary: BWA-MEM is a new alignment algorithm for aligning sequence reads or long query sequences against a large reference genome such as human. It automatically chooses between local and end-to-end alignments, supports paired-end reads and performs chimeric alignment. The algorithm is robust to sequencing errors and applicable to a wide range of sequence lengths from 70bp to a few megabases. For mapping 100bp sequences, BWA-MEM shows better performance than several state-of-art read aligners to date …
☆ 99 Cited by 3577 Related articles All 4 versions »

Li, Heng. "Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly." *Bioinformatics* 28.14 (2012): 1838-1844.

Grill your ~~skewers~~ pipeline!!

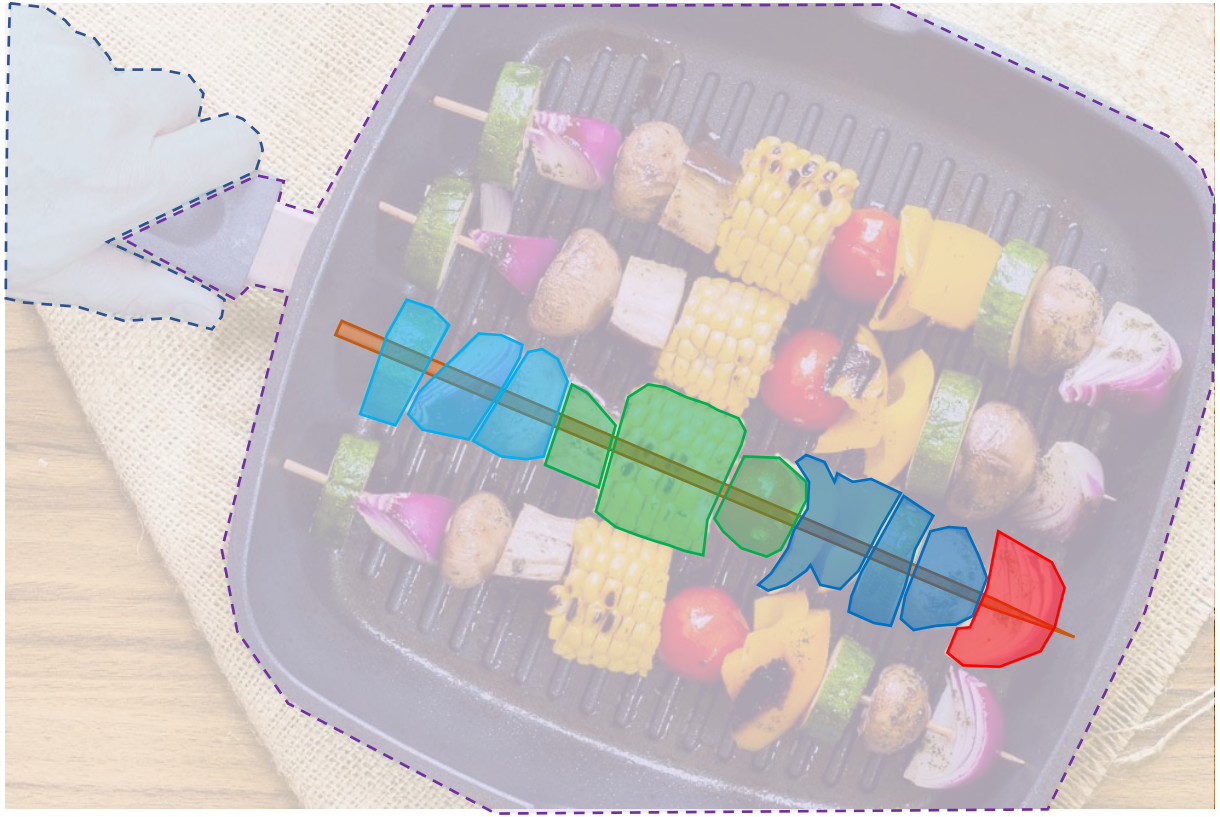
https://mypureplants.com/grilled-vegetable-kabobs

Grill your ~~skewers~~ pipeline!!

You

Container (consistent environment)

Job script



Customized Pre-QC (python/R)

Shell tools (bash)

Customized Post-QC (python/R)

Make plot (python/R)

https://mypureplants.com/grilled-vegetable-kabobs

Case study - Introduction

Genetic variation comes in many forms

Single nucleotide polymorphism (SNP)

ACGACTCGAGCG

↓

ACGACACGAGCG

μ_{SNP}: 1.20×10^{-8} /loc/gen

Short indel (1-20bp)

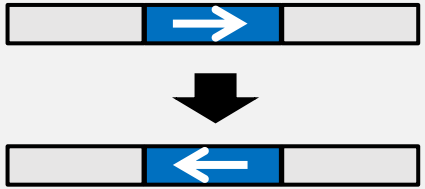
ACGACTCGAGCG

↓

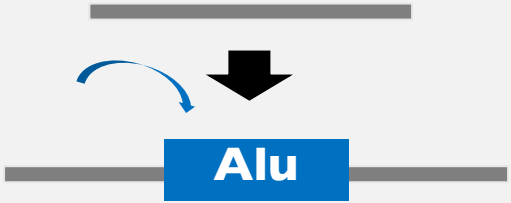
ACGAC-CGAGCG

μ_{INDEL}: 0.68×10^{-9} /loc/gen

Struct.Var /CNV (>20bp)



Alu retrotransposition



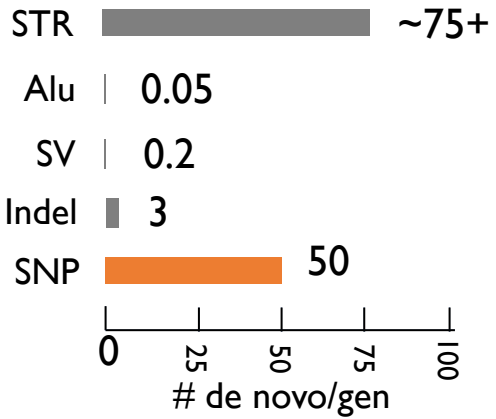
Short tandem repeat

CAGCAG---CAGCAGCA

↓

CAGCAGCAGCAGCAGCA

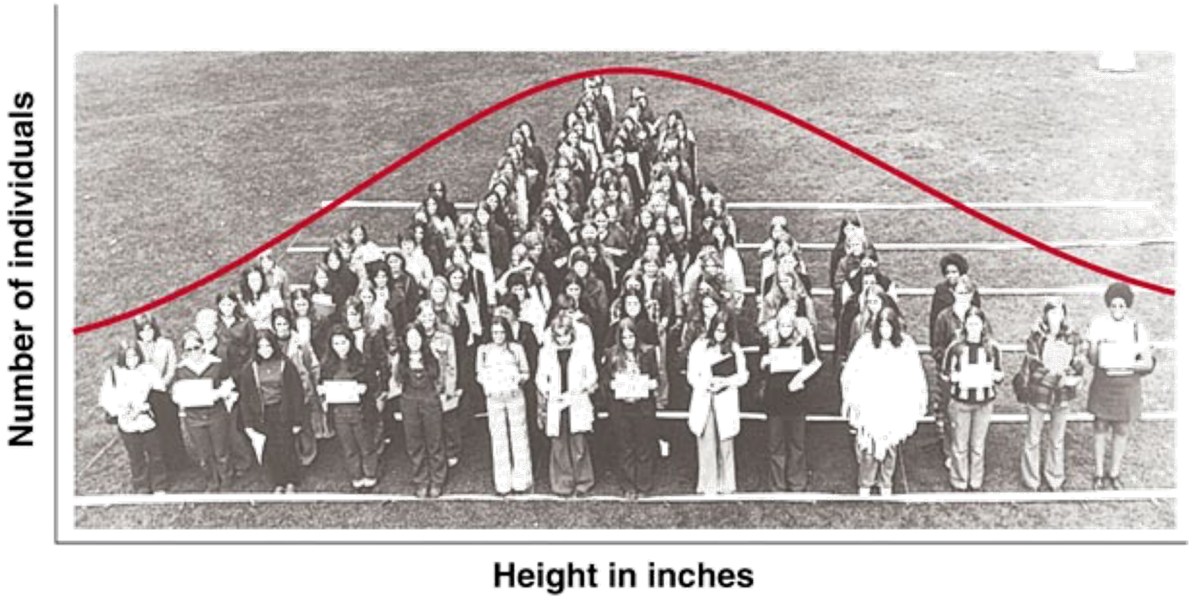
μ_{STR}: 10^{-2}-10^{-5} /loc/gen



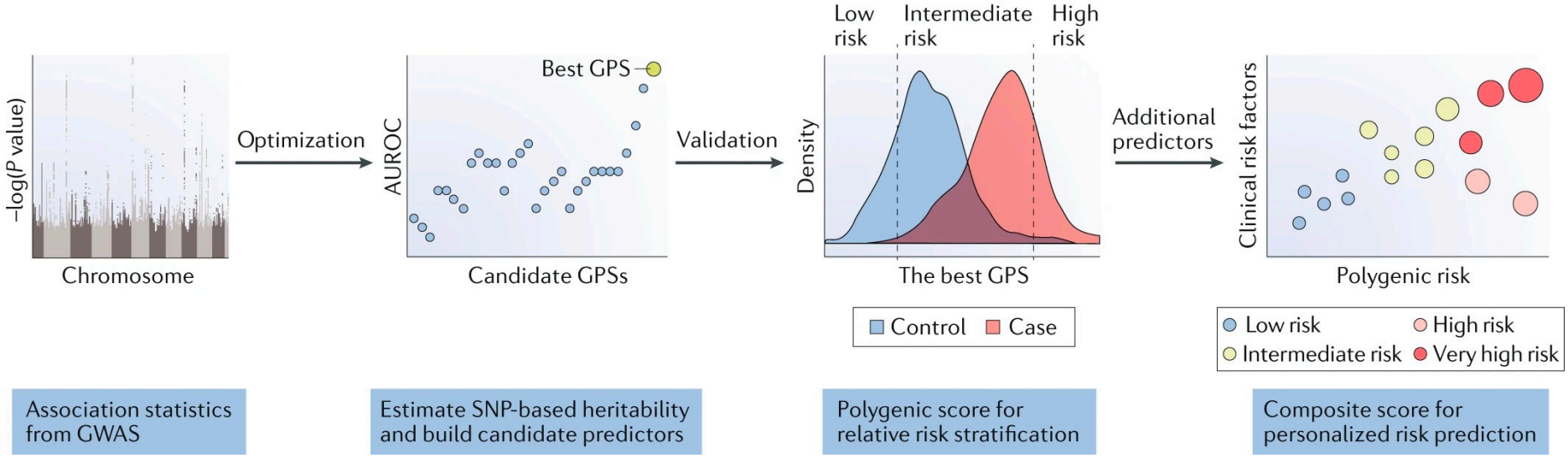
Melissa Gymrek "Characterizing the short tandem repeat mutation process at every locus in the genome" *ASHG* (2016)

Beyond Mendelian – polygenic traits

Tobin/Dusheck, Asking About Life, 2/e
Figure 16.6



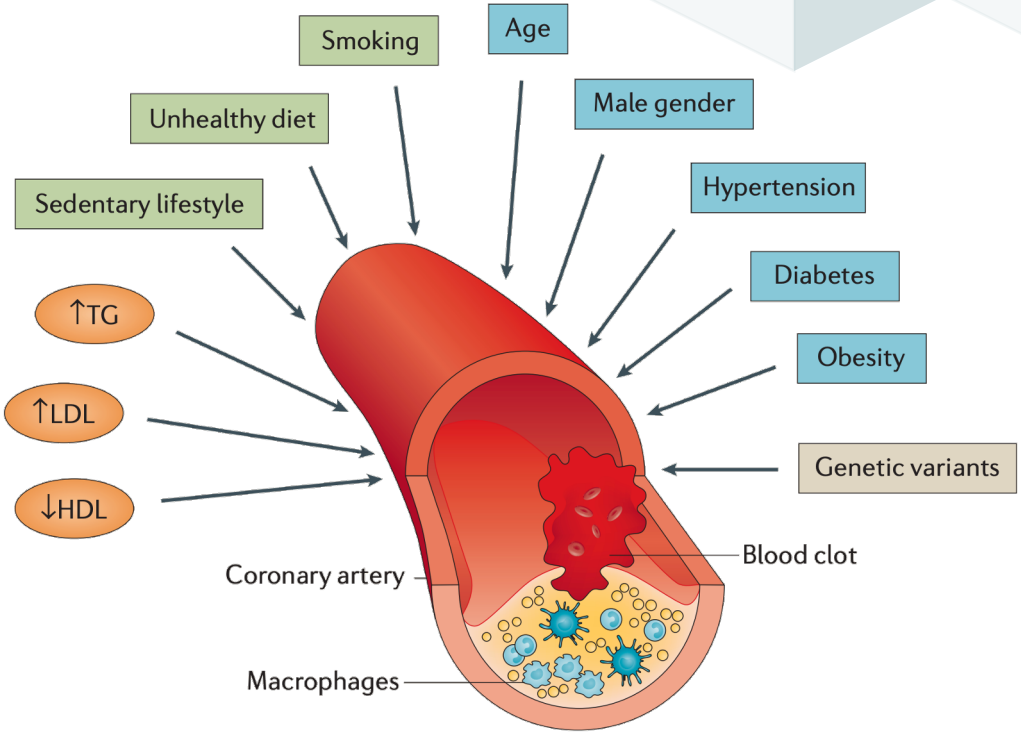
Construction of a Genome-wide Polygenic Risk Score (PRS)



Liu, Lili, and Krzysztof Kiryluk. "Genome-wide polygenic risk predictors for kidney disease." *Nature Reviews Nephrology* 14.12 (2018): 723-724.

Coronary artery disease (CAD)

- High morbidity & mortality
- Framingham Risk Score (FRS)
  - 10-year risk
- Genetic factor
  - Estimated 50-60% heritability
  - Stable from birth



Khera, Amit V., and Sekar Kathiresan. **"Genetics of coronary artery disease: discovery, biology and clinical translation."** *Nature Reviews Genetics* 18.6 (2017): 331.

Polygenic risk score (PRS)

- Genetic prediction of an individual's phenotype
- Sum the products of genotypes effect size estimates from a GWAS across the genome

Fundamental choices:
- Which SNPs to include
- What weights to apply

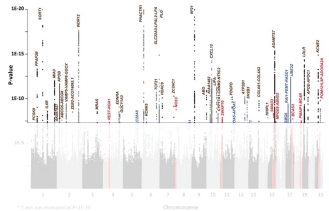
Considerations
- LD
- p-value thresholds

Methods:
- **Meta-GWAS**
- Clumping + thresholding = pruning + thresholding (C+T, P+T)
- Stacked C+T (SCT)
- **LDpred**, SBayesR, PRS-CS (continuous shrinkage)
- Lassosum, crosspred
- Annopred

$$\hat{y} = \sum_{i=1}^{m} \hat{\beta}_i \, x_i$$

GWAS

Individual genome



Slide modified from Nikolas Baya, "Fast Methods for Genome Analysis." *ASHG19:* Session #70

Slide modified from Omer Weissbrod & Alicia Martin, "Trans-ethnic polygenic risk scores: challenges and opportunities" *MPG Primer* (2020)

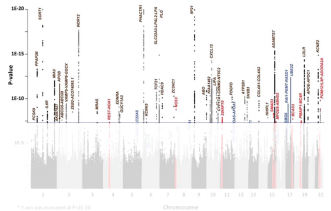
Polygenic risk score (PRS)

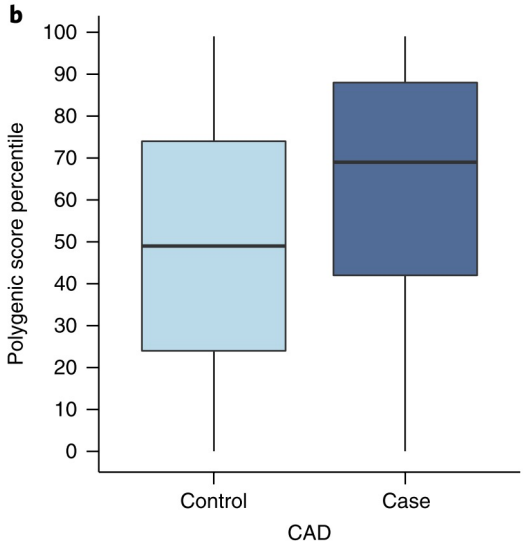
- For common diseases like coronary artery disease (CAD), we can **observe the difference between cases/controls**

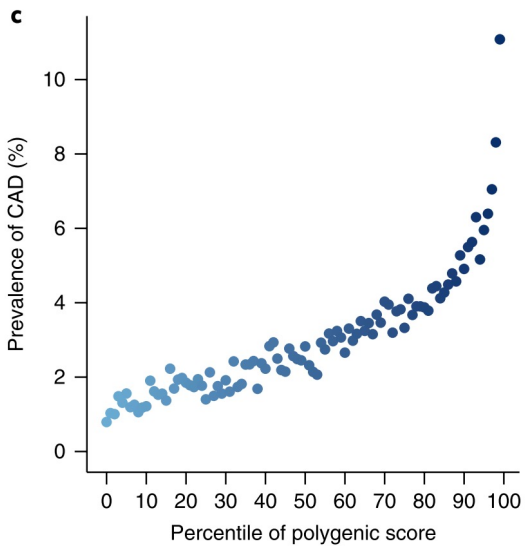
$$\hat{y} = \sum_{i=1}^{m} \hat{\beta}_i \, x_i$$

GWAS

Individual genome







Khera, Amit V., et al. *Nature genetics* 50.9 (2018): 1219-1224.

Slide modified from Nikolas Baya, "Fast Methods for Genome Analysis." *ASHG19:* Session #70

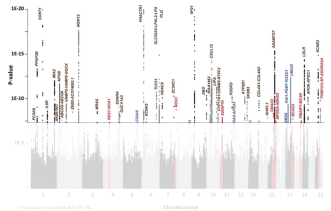
Polygenic risk score (PRS)

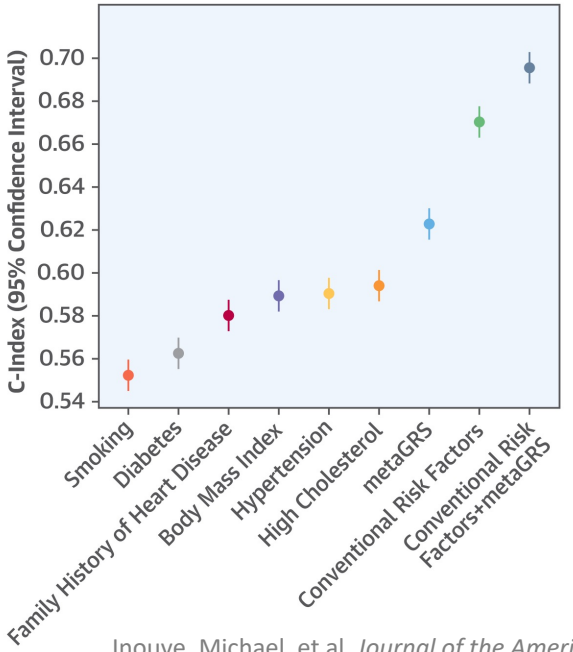
- For common diseases like coronary artery disease (CAD), we can **stratify individuals with different trajectories of risk.**

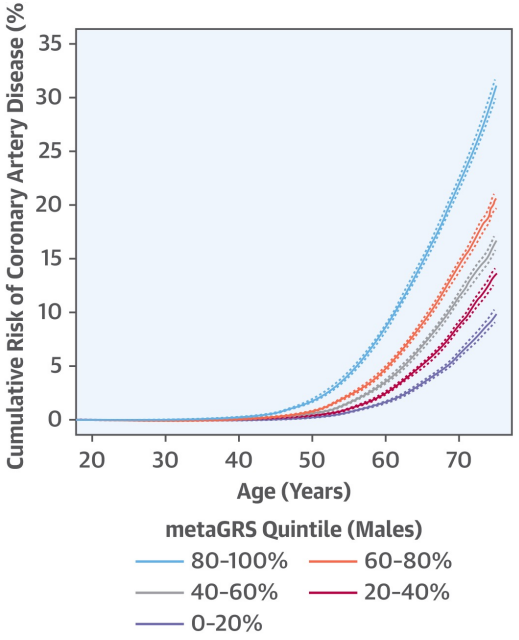
$$\hat{y} = \sum_{i=1}^{m} \hat{\beta}_i \, x_i$$

GWAS

Individual genome







Inouye, Michael, et al. *Journal of the American College of Cardiology* 72.16 (2018): 1883-1893.
Slide modified from Nikolas Baya, "Fast Methods for Genome Analysis." *ASHG19:* Session #70

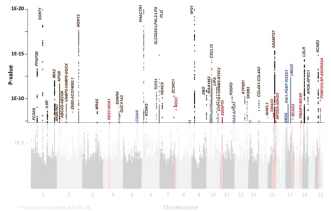
Polygenic risk score (PRS)

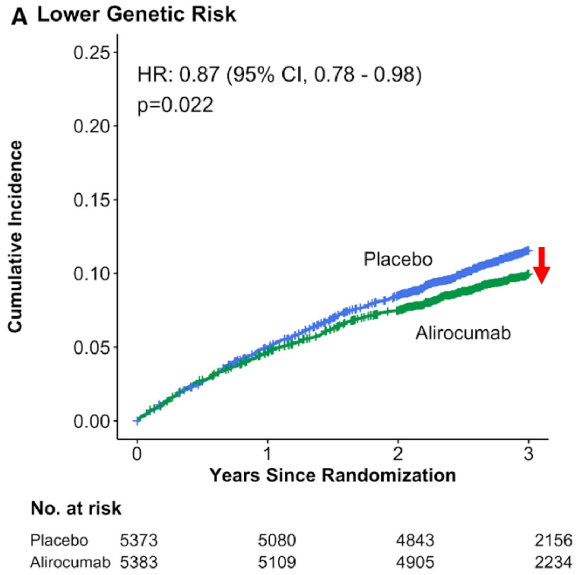
- For common diseases like coronary artery disease (CAD), we can **prioritize patients who derive increased benefit from treatments.**

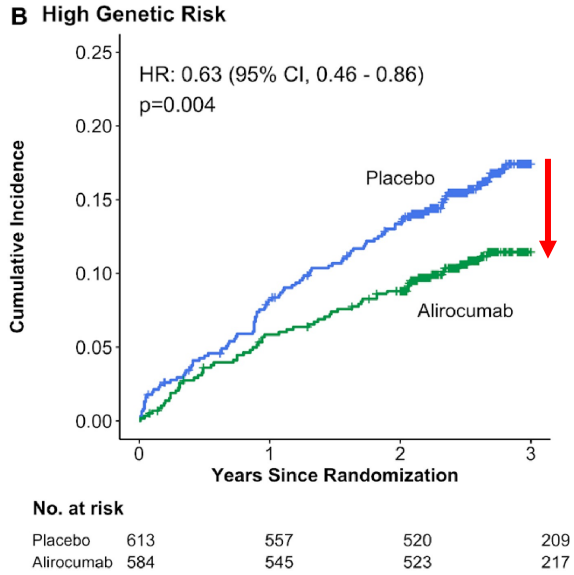
$$\hat{y} = \sum_{i=1}^{m} \hat{\beta}_i \, x_i$$

GWAS

Individual genome







Damask, Amy, et al. *Circulation* 141.8 (2020): 624-636.

Slide modified from Nikolas Baya, "Fast Methods for Genome Analysis." *ASHG19:* Session #70

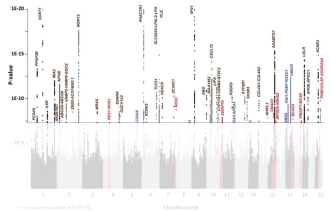
Polygenic risk score (PRS)

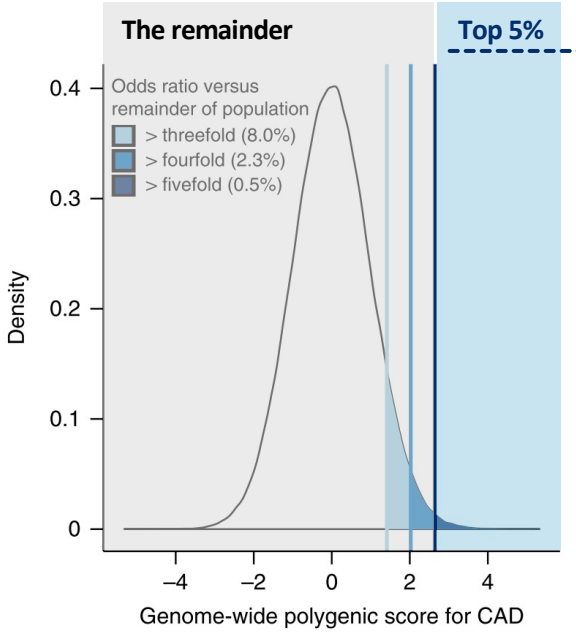
- For common diseases like coronary artery disease (CAD), we can **predict the fold change of inherited risk**.

$$\hat{y} = \sum_{i=1}^{m} \hat{\beta}_i \, x_i$$

GWAS

Individual genome





3.3-fold increased risk of developing CAD, comparing to the remainder.

Khera, Amit V., et al. *Nature genetics* 50.9 (2018): 1219-1224.

Slide modified from Nikolas Baya, "Fast Methods for Genome Analysis." *ASHG19:* Session #70

The PGS Catalog – Oct. 19, 2020

Latest release: **Oct. 19, 2020**

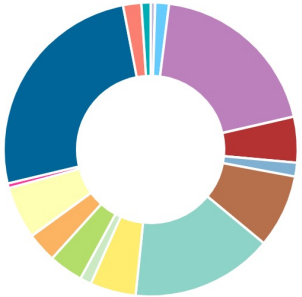
The Polygenic Score (PGS) Catalog

An open database of polygenic scores and the relevant metadata required for accurate application and evaluation.

Polygenic Scores
334

Traits
123

Publications
100



Biological process	2 PGS
Body measurement	6 PGS
Cancer	77 PGS
Cardiovascular disease	20 PGS
Cardiovascular measurement	6 PGS
Digestive system disorder	32 PGS
Hematological measurement	62 PGS
Immune system disorder	20 PGS
Inflammatory measurement	5 PGS
Lipid or lipoprotein measurement	15 PGS
Metabolic disorder	13 PGS
Neurological disorder	23 PGS
Other disease	2 PGS
Other measurement	102 PGS
Other trait	8 PGS
Sex-specific PGS	4 PGS

The PGS Catalog – June 11, 2021

Latest release: **June 11, 2021**

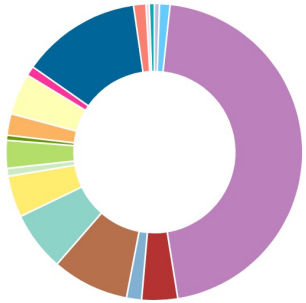
The Polygenic Score (PGS) Catalog

An open database of polygenic scores and the relevant metadata required for accurate application and evaluation.

Polygenic Scores
806

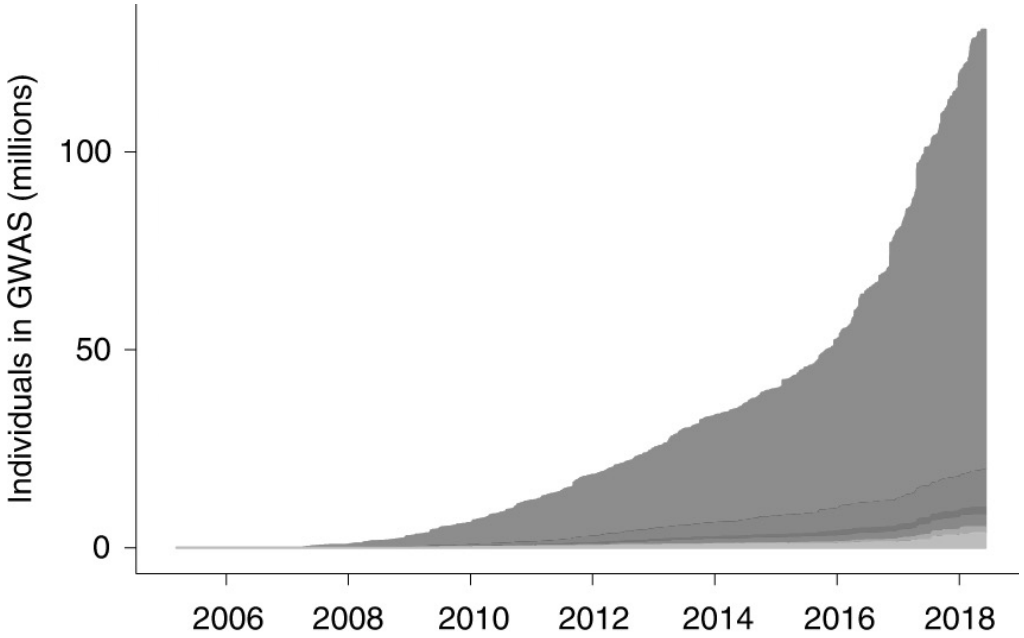
Traits
209

Publications
190



Biological process	4 PGS
Body measurement	10 PGS
Cancer	461 PGS
Cardiovascular disease	38 PGS
Cardiovascular measurement	15 PGS
Digestive system disorder	83 PGS
Hematological measurement	64 PGS
Immune system disorder	44 PGS
Inflammatory measurement	7 PGS
Lipid or lipoprotein measurement	29 PGS
Liver enzyme measurement	4 PGS
Metabolic disorder	22 PGS
Neurological disorder	44 PGS
Other disease	9 PGS
Other measurement	130 PGS
Other trait	12 PGS
Response to drug	2 PGS
Sex-specific PGS	4 PGS

Genetic studies are increasingly powerful



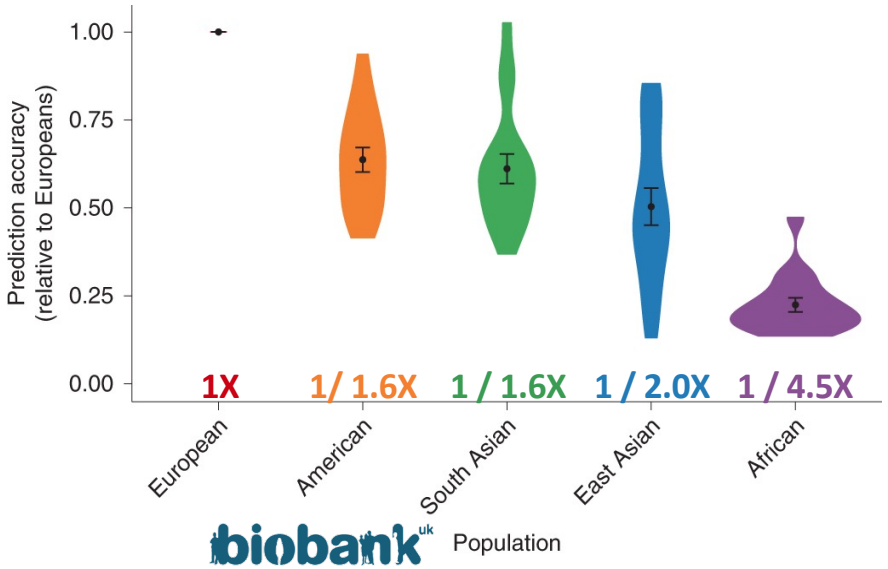
- Previous criticism: limited sample size
- Recently: cheap test for insights into many diseases or traits

- Future: integrate with other clinical factors for therapeutic decision-making

Slide modified from Omer Weissbrod & Alicia Martin, “Trans-ethnic polygenic risk scores: challenges and opportunities” *MPG Primer* (2020)

Martin, Alicia R., et al. "Clinical use of current polygenic risk scores may exacerbate health disparities." *Nature genetics* 51.4 (2019): 584-591.

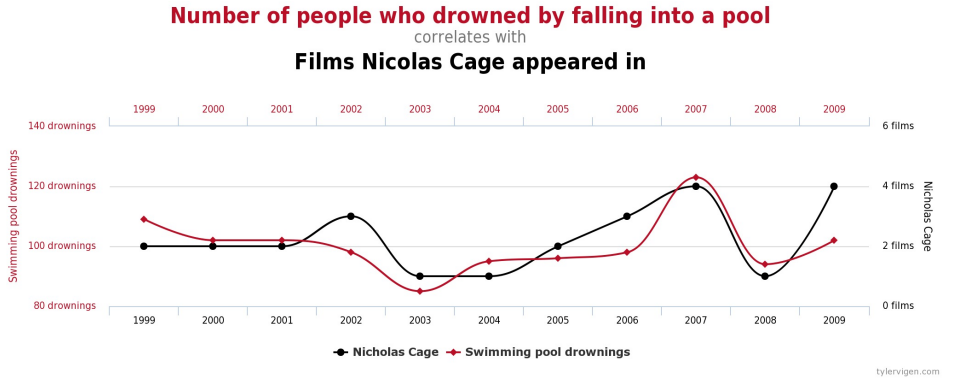
Current GWAS summary statistics were Eurocentric

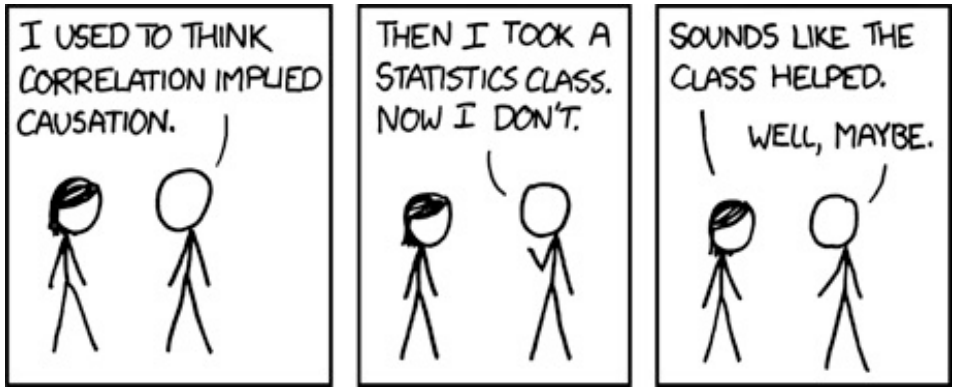


- Prediction accuracy decayed across populations
  - 17 quantitative traits in the UKBB
  - EUR-derived summary statistics

Martin, Alicia R., et al. "Clinical use of current polygenic risk scores may exacerbate health disparities." *Nature genetics* 51.4 (2019): 584-591.

GWASs derived PRSs – while correlation does not imply causation.

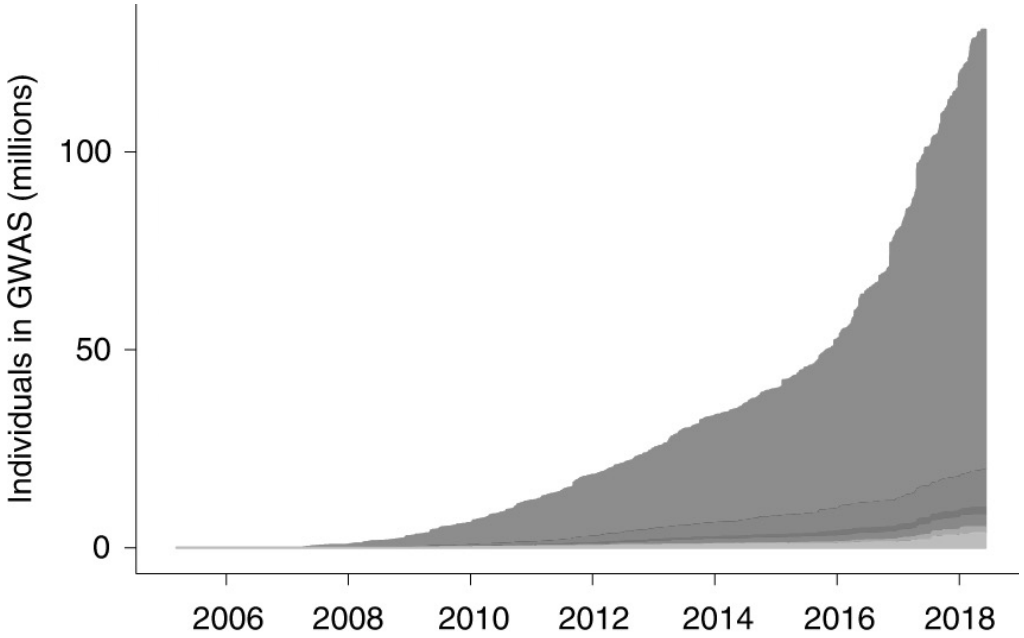




xkcd “Correlation" https://xkcd.com/552/

Spurious correlations https://www.tylervigen.com/spurious-correlations

Genetic studies are increasingly powerful

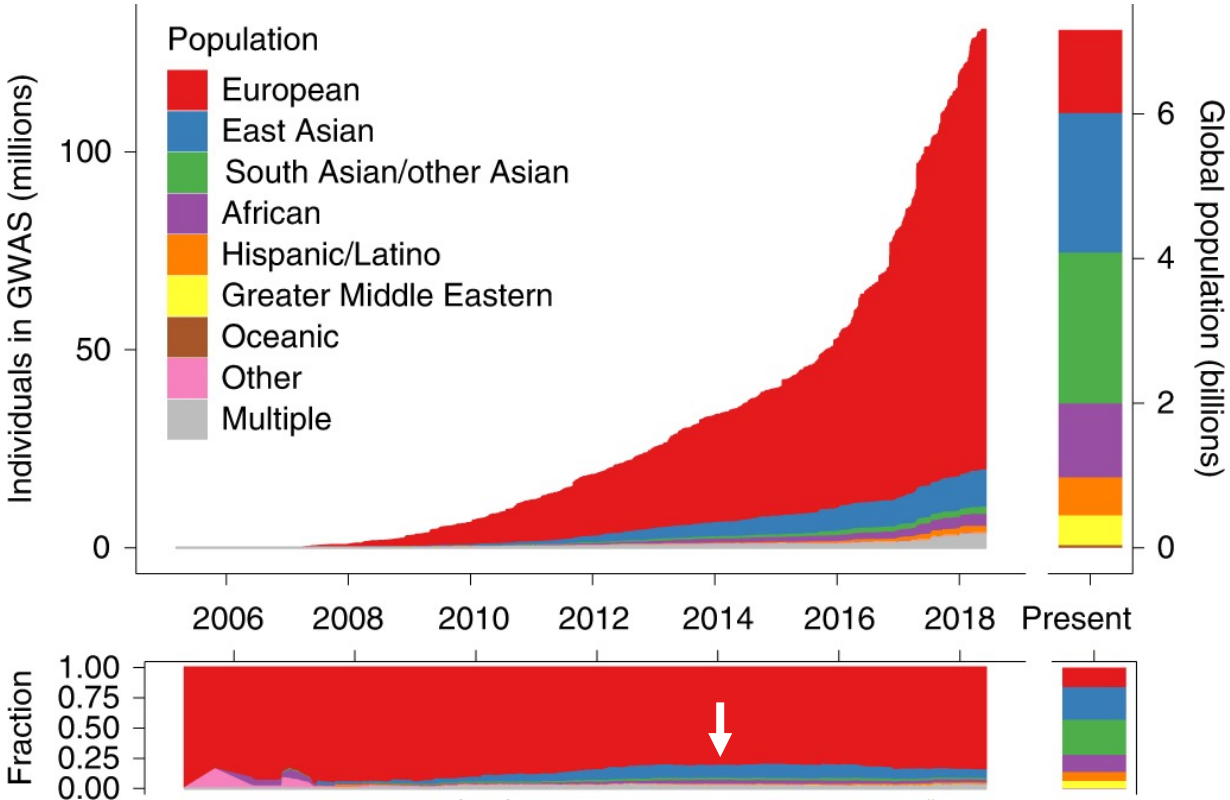


- Previous criticism: limited sample size
- Recently: cheap test for insights into many diseases or traits
- Future: integrate with other clinical factors for therapeutic decision-making

Slide modified from Omer Weissbrod & Alicia Martin, “Trans-ethnic polygenic risk scores: challenges and opportunities” *MPG Primer* (2020)

Martin, Alicia R., et al. "Clinical use of current polygenic risk scores may exacerbate health disparities." *Nature genetics* 51.4 (2019): 584-591.

…but genetics has a HUGE diversity problem

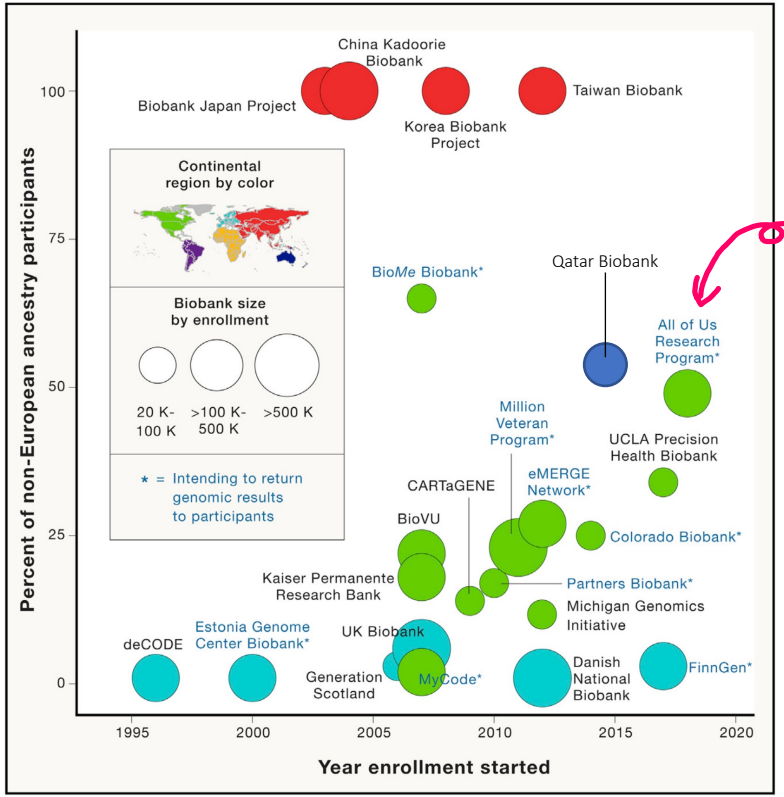


- Previous criticism: limited sample size
- Recently: cheap test for insights into many diseases or traits
- **Currently: may exacerbate health disparities for multi-ancestral population**
- Future: integrate with other clinical factors for therapeutic decision-making

Slide modified from Omer Weissbrod & Alicia Martin, "Trans-ethnic polygenic risk scores: challenges and opportunities" *MPG Primer* (2020)

Martin, Alicia R., et al. "Clinical use of current polygenic risk scores may exacerbate health disparities." *Nature genetics* 51.4 (2019): 584-591.

Biobanks with Genomic Data Linked to non-EHRs Participants



Abul-Husn, Noura S., and Eimear E. Kenny. "Personalized medicine and the power of electronic health records." *Cell* 177.1 (2019): 58-69.

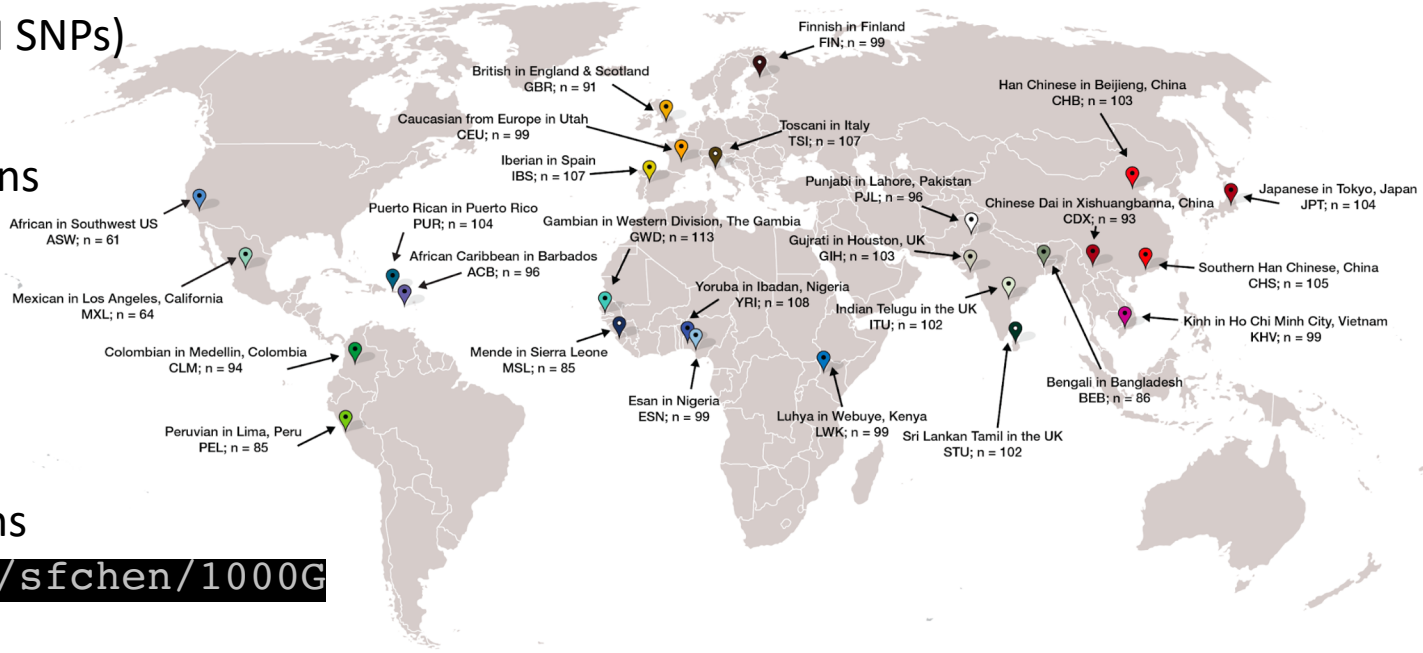
https://www.scripps.edu/science-and-medicine/translational-institute/translational-research/precision-medicine/index.html

Case study – Ancestry inference

Ancestry Reference Panel - 1000 Genomes (TGP)

1000G Phase 3 (GRCh37)

- 2.3M sites (1.92M SNPs)
- 2,504 individuals
- 5 super populations
  - 661 AFR
  - 347 AMR
  - 504 EAS
  - 503 EUR
  - 489 SAS
- 26 sub-populations

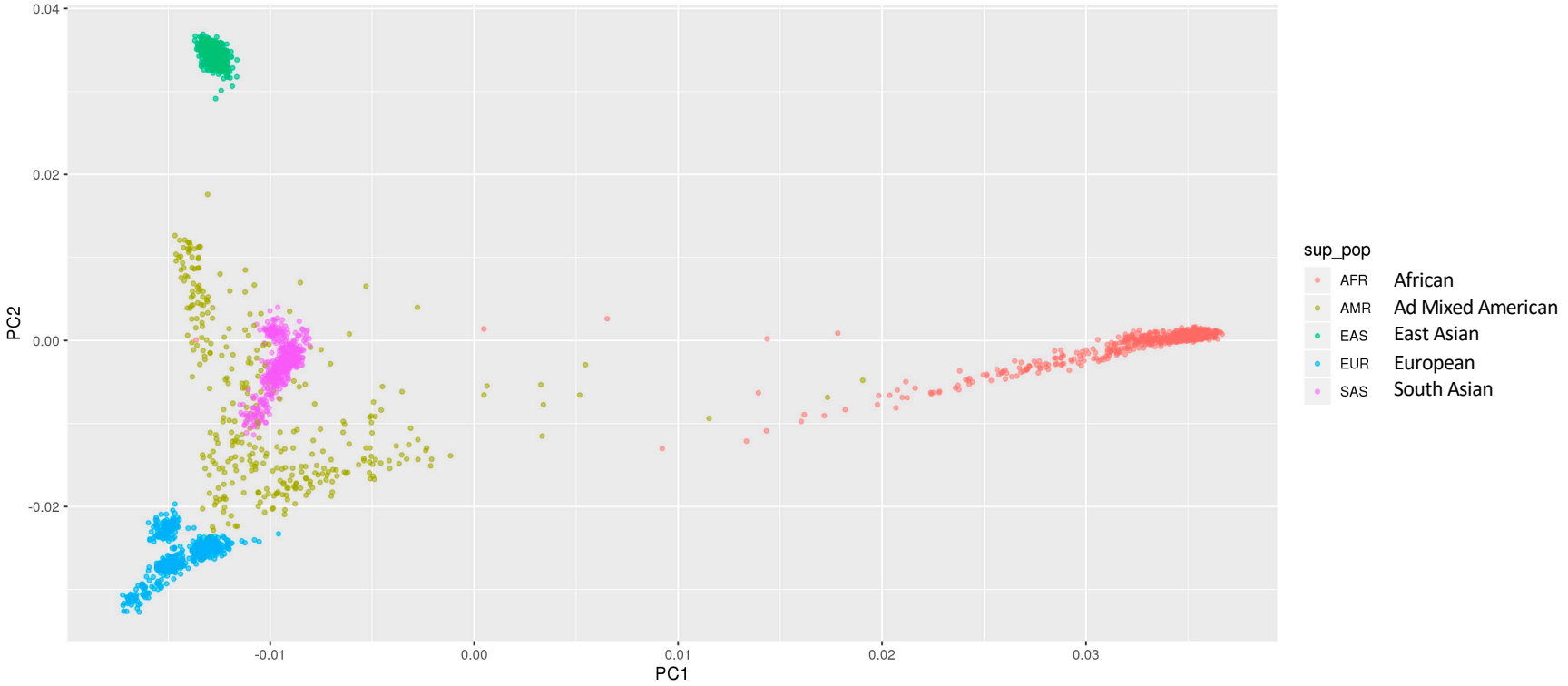



```
~ $ ls /gpfs/work/sfchen/1000G
```

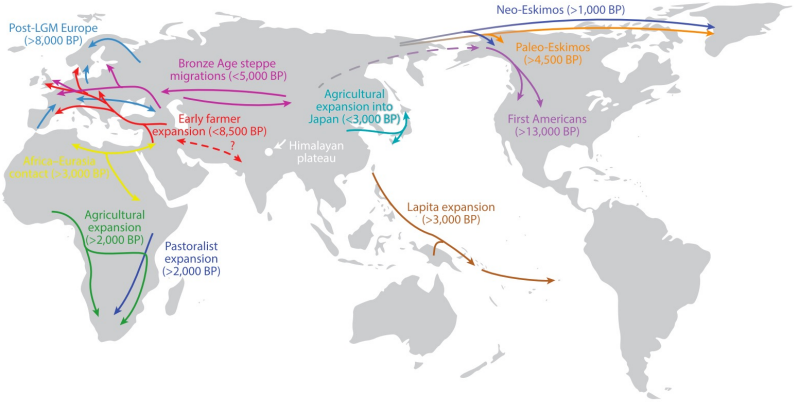
```
~ $ sbatch --job-name=PCA_TGP 0_TGP_PCA.slurm.sh
```

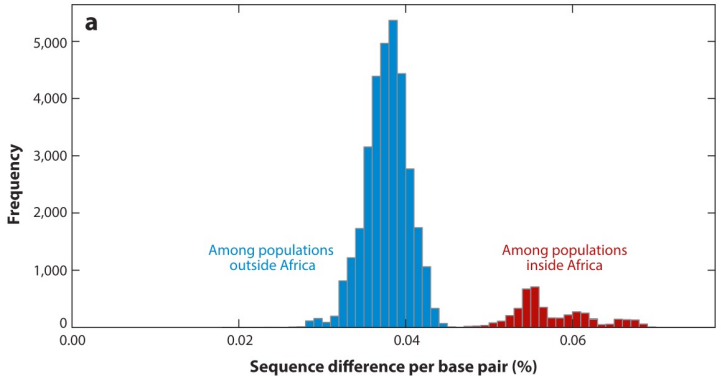
IGSR: The International Genome Sample Resource https://www.internationalgenome.org/

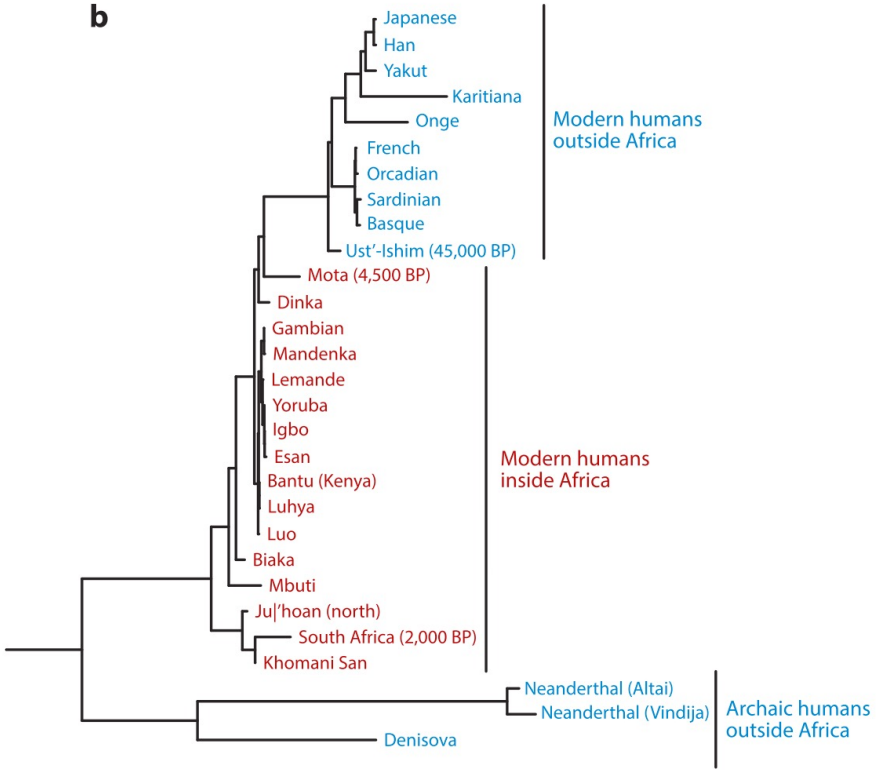
Ancestry inference using PCA – 1000G



Patterns of modern human genomic diversity – Out-of-Africa model

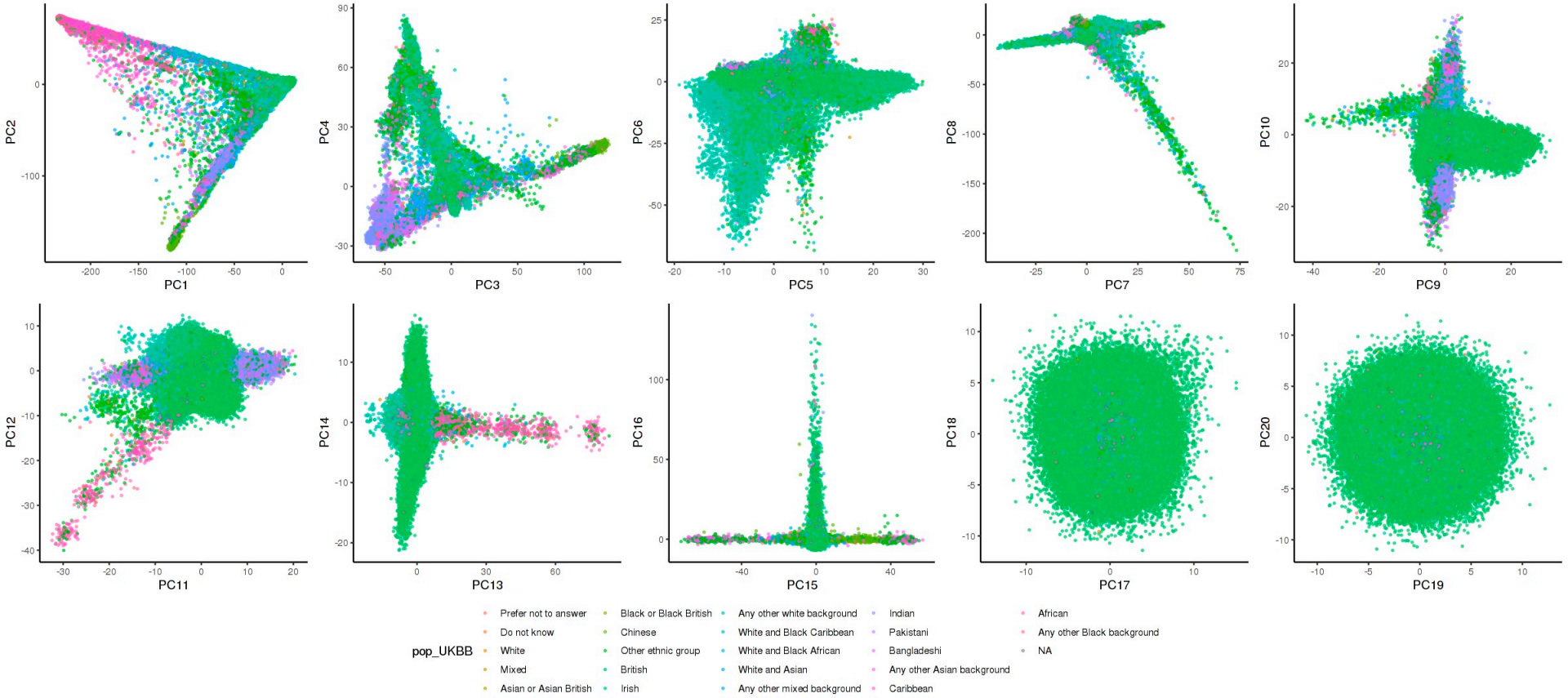




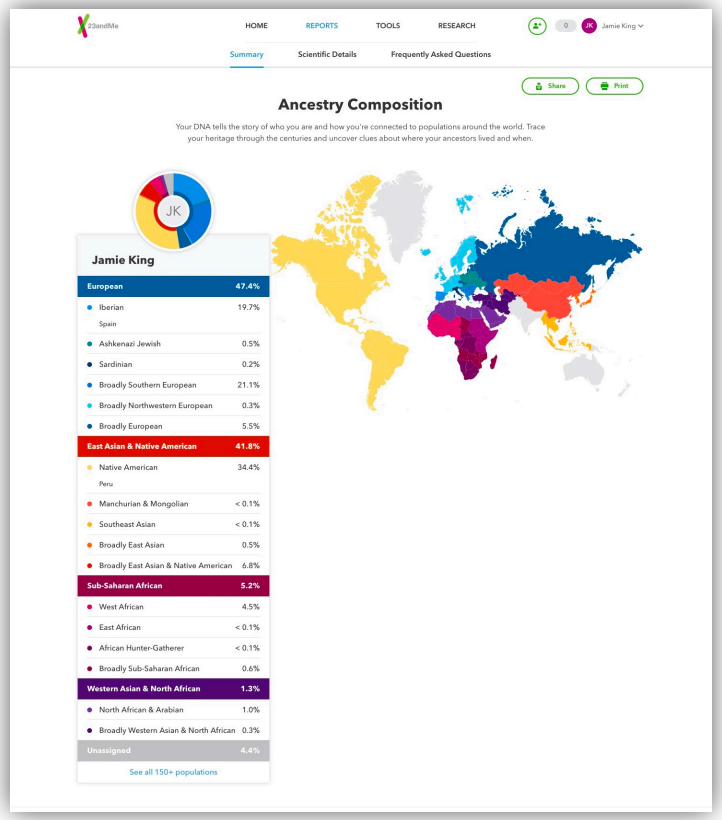


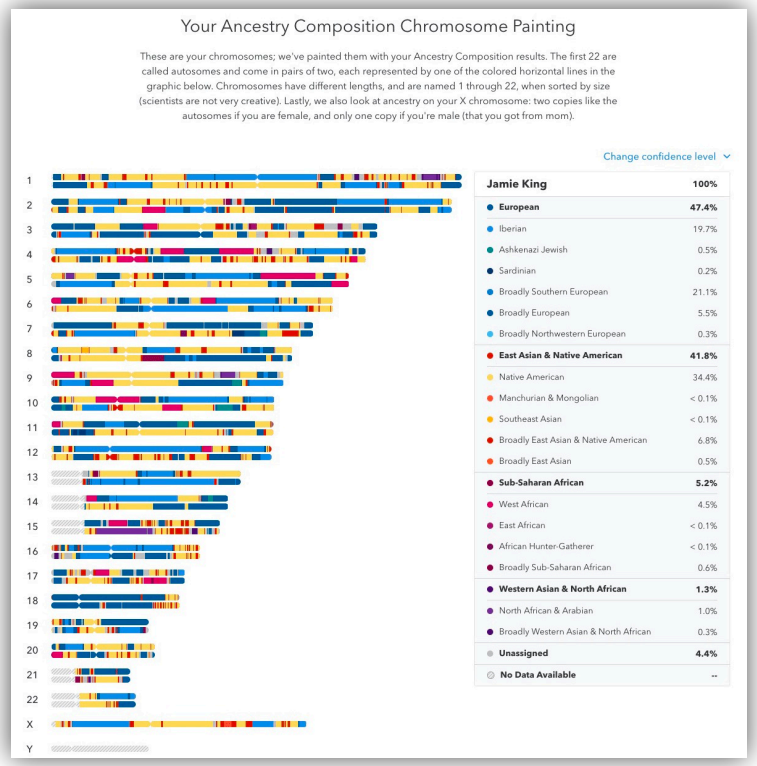
Skoglund, Pontus, and Iain Mathieson. "Ancient genomics of modern humans: the first decade." *Annual review of genomics and human genetics* 19 (2018): 381-404.

Ancestry inference using PCA – UK Biobank



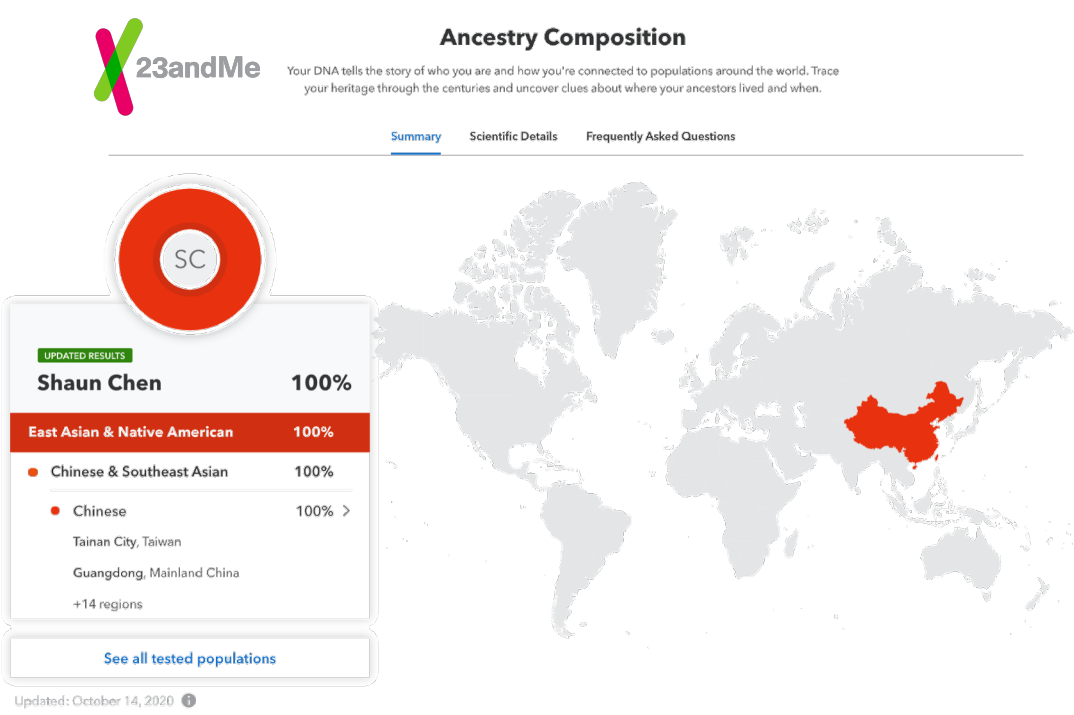
Personal ancestry composition report

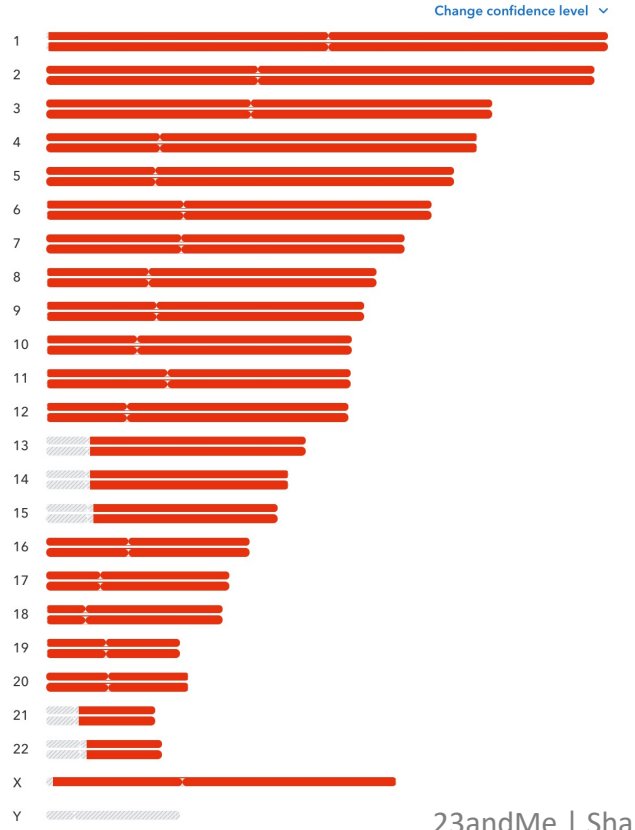




Ancestry Composition report, 23andMe, https://blog.23andme.com/ancestry-reports/23andme-updates-ancestry-composition/

What I really got…





qsub to sbatch translation			
To specify the:	**qsub option**	**sbatch option**	**Comments**
Queue/partition	-q *QUEUENAME*	-p *QUEUENAME*	Torque "queues" are called "partitions" in slurm. **Note:** the partition/queue structure has been simplified, see below.
Account/allocation to be charged	-A *ACCOUNTNAME*	-A *ACCOUNTNAME*	Gold uses "allocations", Slurm uses "accounts"
Number of nodes/ cores requested	-l nodes=*NUMBERCORES*	-n *NUMBERCORES*	See below
	-l nodes=*NUMBERNODES*:*CORESPERNODE*	-N *NUMBERNODES* -n *NUMBERCORES*	
Wallclock limit	-l walltime=*TIMELIMIT*	-t *TIMELIMIT*	*TIMELIMIT* should have form of *HOURS:MINUTES:SECONDS*. Slurm supports some other time formats as well.
Memory requirements	-l mem=*MEMORY*mb	--mem=*MEMORY*	Moab: This is Total memory used by job Slurm: This is memory per node
	-l pmem=*MEMORY*mb	--mem-per-cpu=*MEMORY*	This is per CPU/core. *MEMORY* in MB
Stdout file	-o *FILENAME*	-o *FILENAME*	This will combine stdout/stderr on slurm if `-e` not given also
Stderr file	-e *FILENAME*	-e *FILENAME*	This will combine stderr/stdout on slurm if `-o` not given also
Combining stdout/stderr	-j oe	-o *OUTFILE* and no `-e` option	stdout and stderr merged to stdout/*OUTFILE*
	-j eo	-e *ERRFILE* and no `-o` option	stdout and stderr merged to stderr/*ERRFILE*
Email address	-M *EMAILADDR*	--mail-user=*EMAILADDR*	
Email options	-mb	--mail-type=BEGIN	Send email when job starts
	-me	--mail-type=END	Send email when job ends
	-mbe	--mail-type=BEGIN --mail-type=END	Send email when job starts and ends
Job name	-N *NAME*	--job-name=NAME	
Working directory	-d *DIR*	--workdir=*DIR*	

Slurm vs Torque (UMD) https://hpcc.umd.edu/hpcc/help/slurm-vs-moab.html

Moab/Torque to Slurm Environment Correlations			
Function	**Moab/Torque Variable**	**Slurm Variable**	**Comments**
Job ID	$PBS_JOBID	$SLURM_JOBID	
Job Name	$PBS_JOBNAME	$SLURM_JOB_NAME	
Submit Directory	$PBS_O_WORKDIR	$SLURM_SUBMIT_DIR	
Node List	cat $PBS_NODEFILE	$SLURM_JOB_NODELIST	See below
Host submitted from	$PBS_O_HOST	$SLURM_SUBMIT_HOST	
Number of nodes allocated to job	$PBS_NUM_NODES	$SLURM_JOB_NUM_NODES	
Number of cores/node	$PBS_NUM_PPN	$SLURM_CPUS_ON_NODE	
Total number of cores for job???	$PBS_NP	$SLURM_NTASKS	Uncertain about these
Index to node running on relative to nodes assigned to job	$PBS_O_NODENUM	$SLURM_NODEID	
Index to core running on within node	$PBS_O_VNODENUM	$SLURM_LOCALID	
Index to task relative to job	$PBS_O_TASKNUM	$SLURM_PROCID + 1	

Slurm vs Torque (UMD) https://hpcc.umd.edu/hpcc/help/slurm-vs-moab.html

Take Home Message - Caution

- **Caution**
  - Never compute at login node
  - Never use root permission (no `sudo`)
- **Max job number**
  - started from 10
  - Put `` `sleep 3` `` between loop of job submission
- **Do**
  - Backup your projects frequently
  - Respect other users (shared space and privacy)
  - Be responsible/responsive for your behavior
  - Inform your PI for data management
- **Be friendly with hpc_ca@scripps.edu**

Take Home Message – Pipeline Development

- **First draft**
  - Interactive debugging (Jupyterhub; Rstudio; other editers, IDE, etc.)
    - Shut down the session when leaving…
  - Test case
- **Form executable script**
  - Unit test

```
If you in ["standardize multiple steps processing",
           "job couldn't finish within feasible time",
           "local machine overheated",
           "wanna go home earlier"]:
```

- **Build pipeline**
  - Print (echo) checkpoint message; Save human readable logging output
  - Allocate proper resource
  - Merge or split jobs for optimized granularity
  - Remove temp files when jobs finished
  - Fully understand the working concept of the tools
    - Do not treat the tools as black boxes
    - Read background article or even source code if needed
  - Speed: C/Java > JS/Python/Perl > R

Resources

- **2020_FA_HPC** https://github.com/ShaunFChen/2020_FA_HPC
- **Scripps Intranet**

https://intranet.scripps.edu/its/highperformancecomputing/index.html

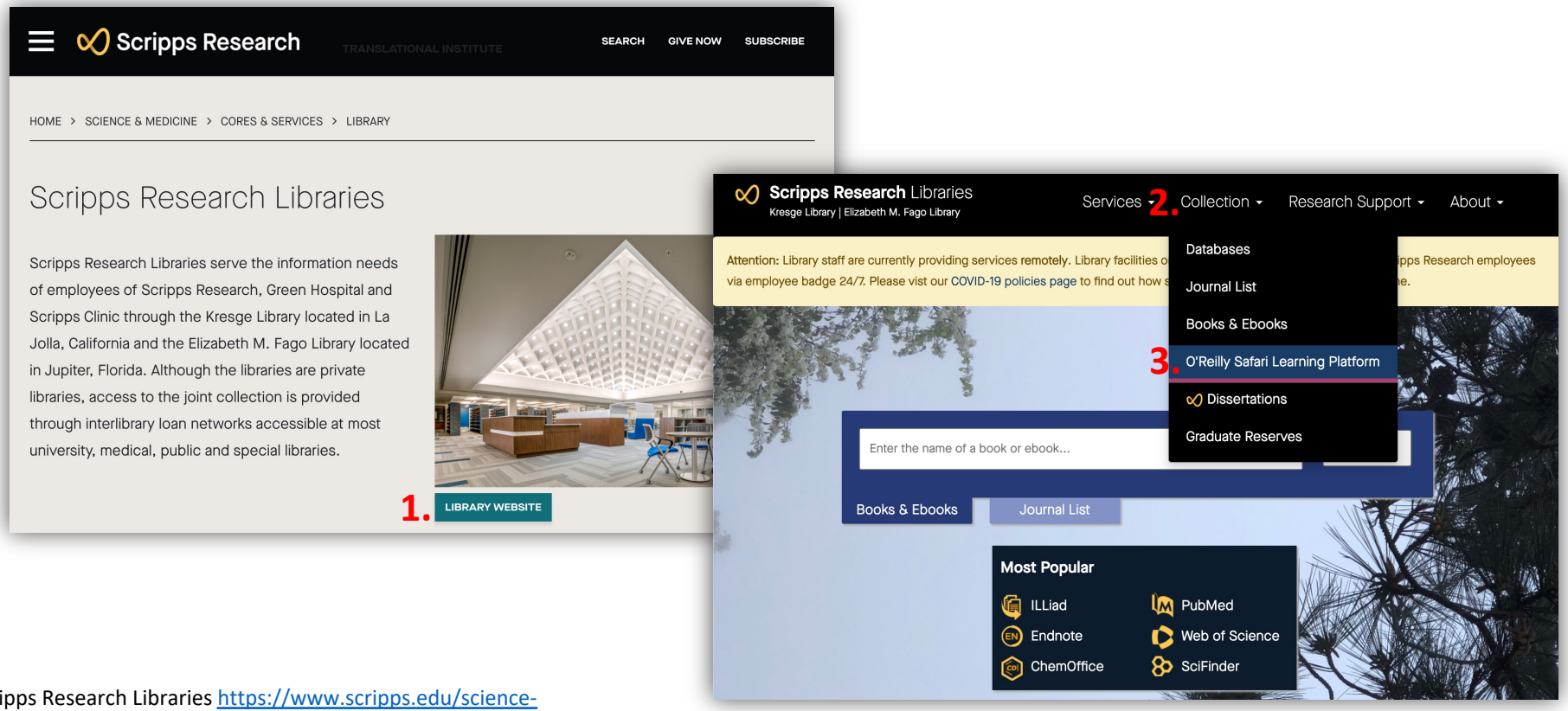
- **Scripps Libraries**

https://www.scripps.edu/science-and-medicine/cores-and-services/library/index.html

- **pbs2slurm** https://github.com/bjpop/pbs2slurm
- **GNU Parallel and Multiprocessing**

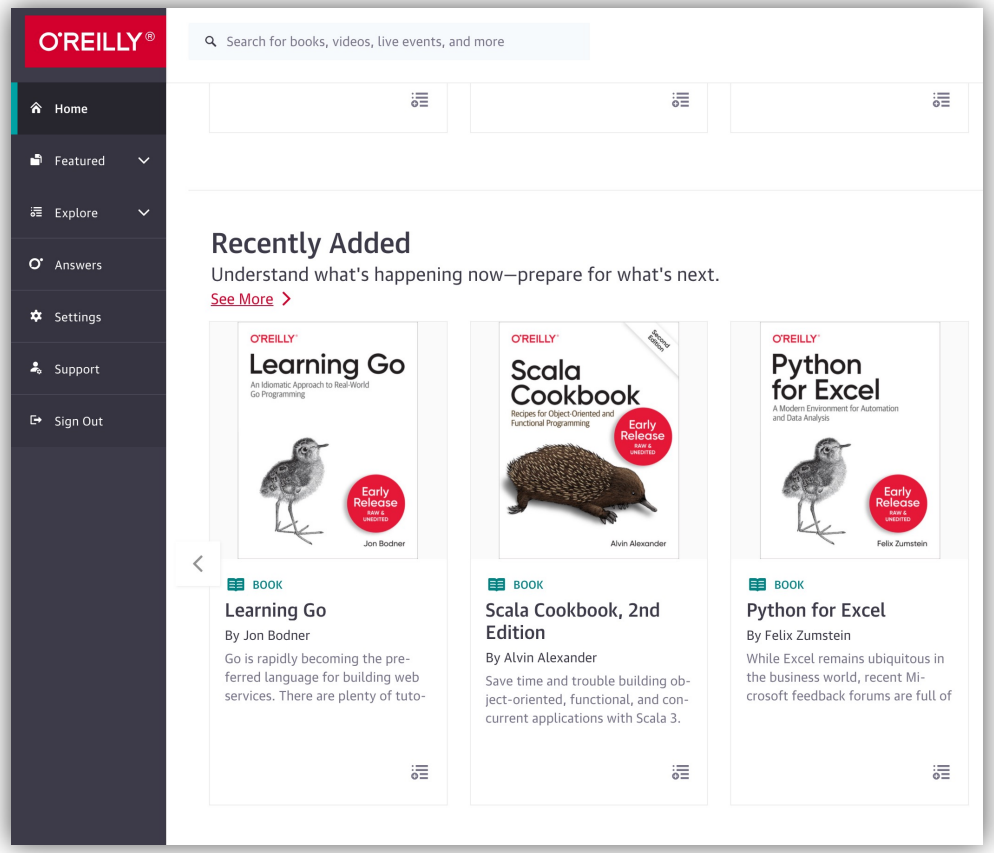
https://github.com/ShaunFChen/CBB_Parallel_Multiprocessing

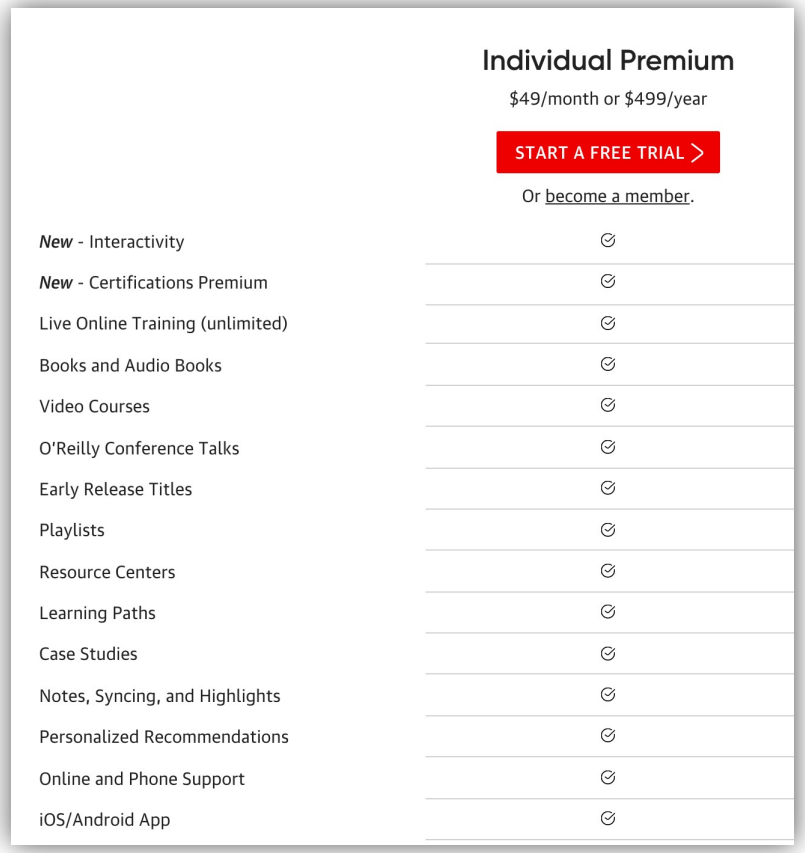
O'Reilly Online Learning Platform - I



Scripps Research Libraries https://www.scripps.edu/science-and-medicine/cores-and-services/library/index.html

O'Reilly Online Learning Platform - II





Thanks for your attention!! Wish you have…