Neurogenetics on Biowulf: From GWAS to Machine Learning

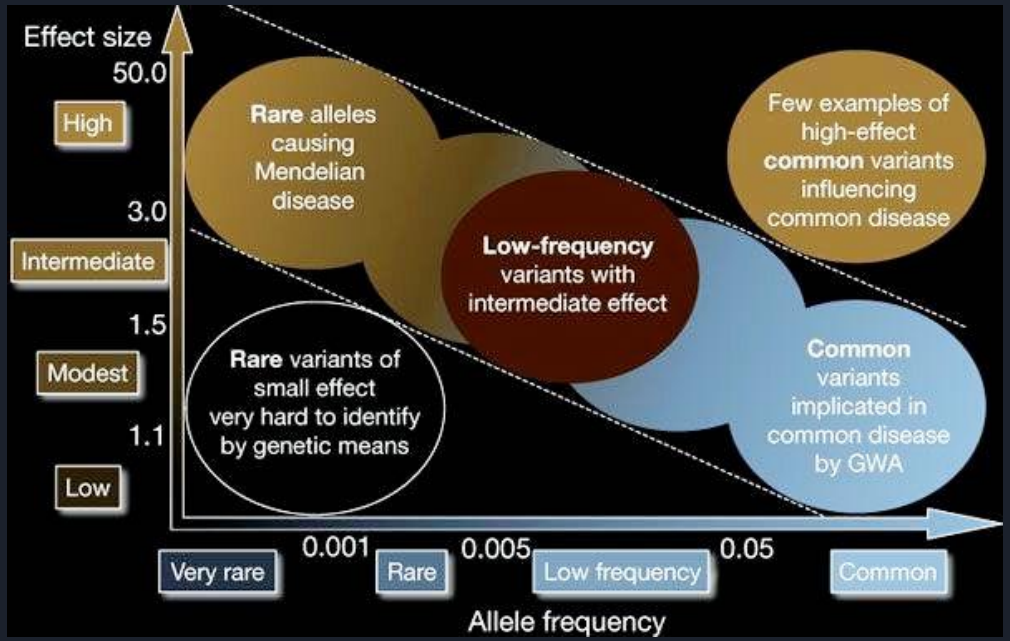
Mike A. Nalls, PhD
mike@datatecnica.com / nallsm@mail.nih.gov
Data Tecnica Int'l / NIA's Laboratory of Neurogenetics
May 7th 2019

What's going on today (in 45 min or less)...

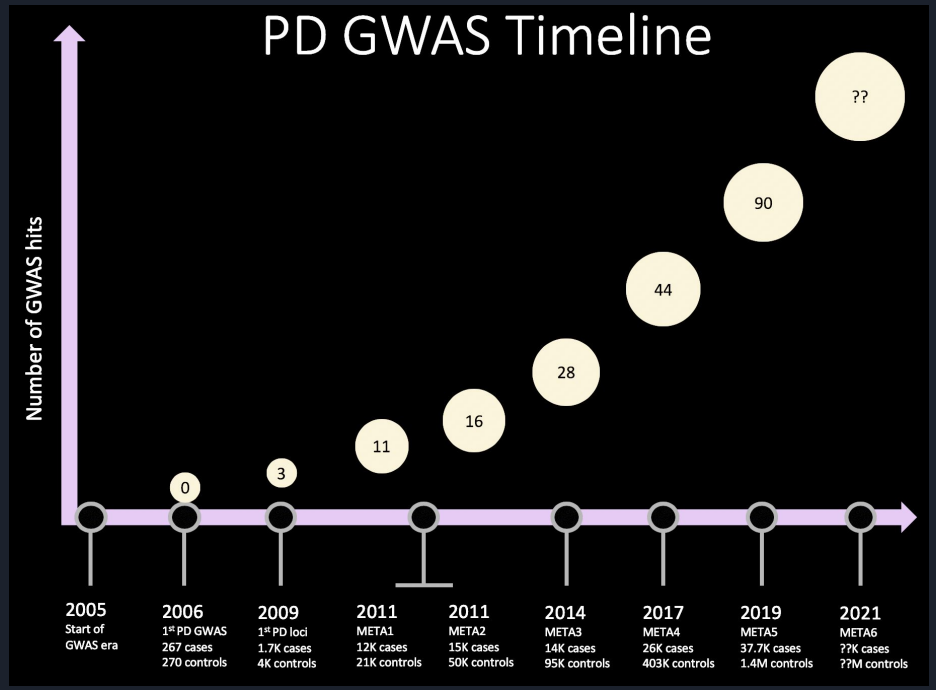
- Parkinson's disease (PD) genetics and genome-wide association studies (GWAS)
  - General overview
  - Growth of our research parallels and is facilitated by Biowulf's growth
    - "Too big for one lab"
- Massive reference data and current generation GWAS
  - Standardized pipelines
  - Multi-modality data integration
- Growing scope of genetics and getting past basic case-control analyses
  - Deep data + disease progression
  - Machine learning ecosystems
  - Analysis platforms
  - Biowulf integration, collaboration and the hybrid cloud

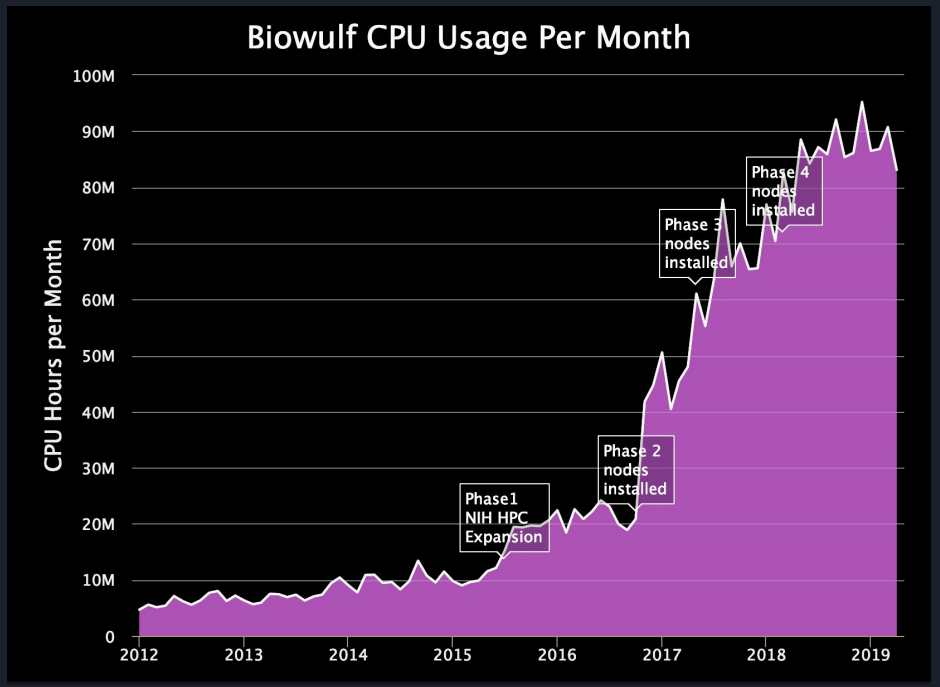
Quick note: “What is a meta-GWAS?”...

- Meta-analysis of genome-wide association studies (GWASs)
- Separate sites survey every possible variant they have data for (after QC, millions of variants included)
- Compare regression coefficients and standard errors per variant
  - Summary statistics often combined using fixed effects
  - Random effects used to account for across study heterogeneity
- Manolio et al., 2009 nature paper still a great read, also has this figure in it (right)



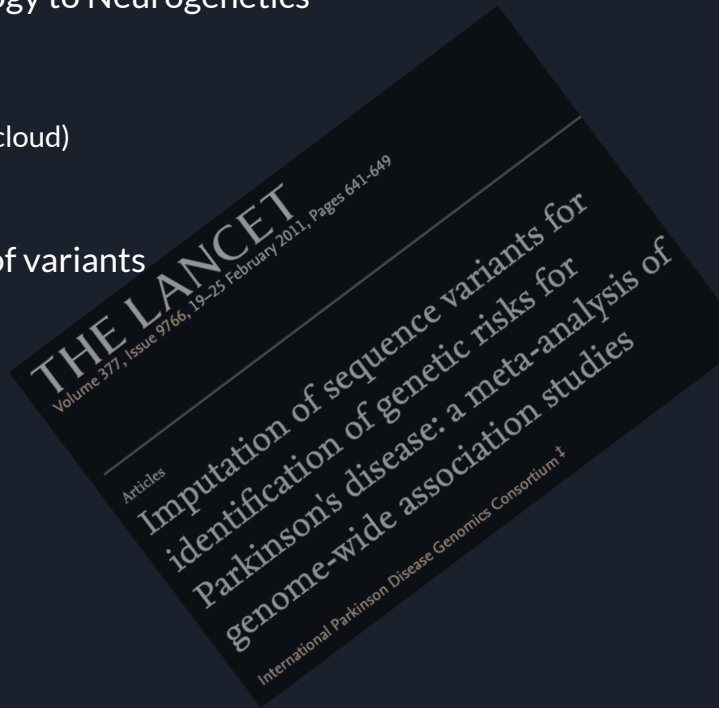
Growing together (more computers, more hits?)...





My personal beginning with Biowulf...

- 2010, fresh PhD transferred from Biostats/Epidemiology to Neurogenetics
- Few of small GWAS of PD in different labs
- Imputation to standardize variants across studies
  - Time consuming (now it's on a remote web server in the cloud)
  - Large reference data (for the time)
  - Local compute intensive (not anymore)
- Meta-analyze 12K cases and 20K controls at millions of variants
- This is not happening on laptops in your lab
- No commercial cloud at the time



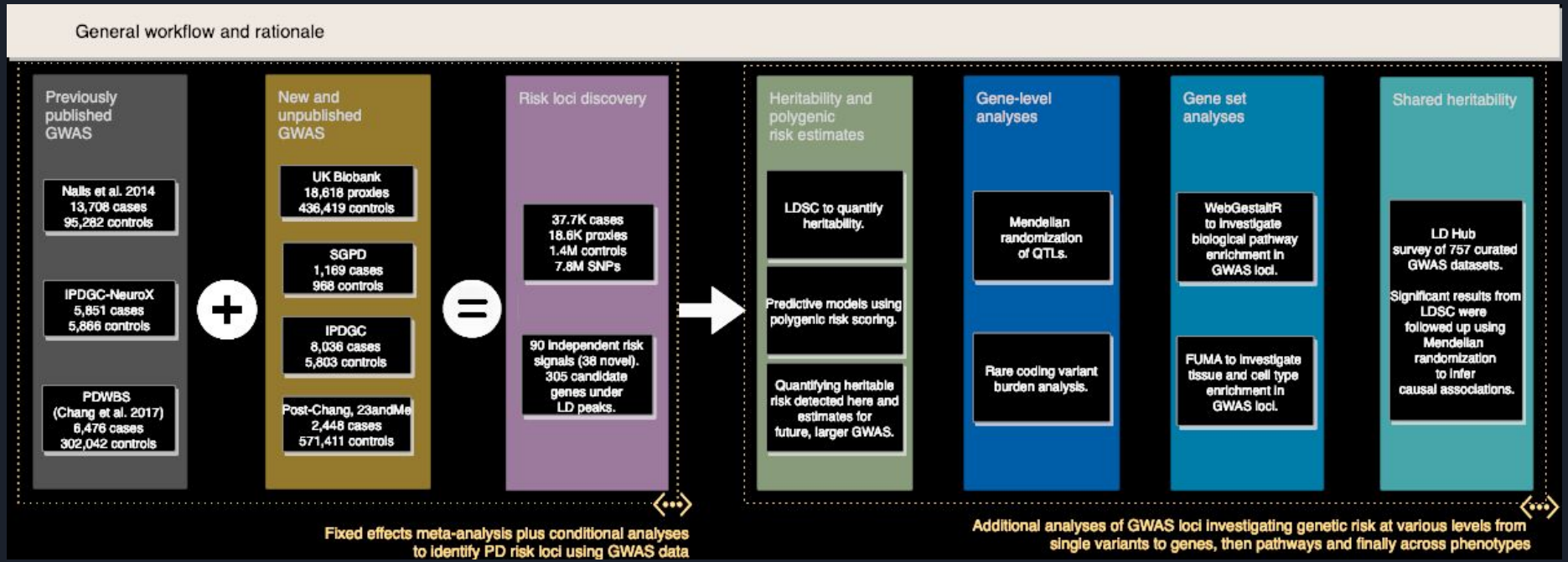
8 years of growth → more samples, more risk loci, more collaborators, more data…

Meta-GWAS of ~42K cases and 1.4M controls at over 7M variants each (can't happen without some very smart friends and alot of compute)...

Mike A. Nalls, Cornelis Blauwendraat, Costanza L. Vallerga, Karl Heilbron, Sara Bandres-Ciga, Diana Chang, Manuela Tan, Demis A. Kia, Alastair J. Noyce, Angli Xue, Jose Bras, Emily Young, Rainer von Coelln, Javier Simón-Sánchez, Claudia Schulte, Manu Sharma, Lynne Krohn, Lasse Pihlstrom, Ari Siitonen, Hirotaka Iwaki, Hampton Leonard, Faraz Faghri, J. Raphael Gibbs, Dena G. Hernandez, Sonja W. Scholz, Juan A. Botia, Maria Martinez, Jean-Christophe Corvol, Suzanne Lesage, Joseph Jankovic, Lisa M. Shulman, The 23andMe Research Team, System Genomics of Parkinson's Disease (SGPD) Consortium, Margaret Sutherland, Pentti Tienari, Kari Majamaa, Mathias Toft, Ole A. Andreassen, Tushar Bangale, Alexis Brice, Jian Yang, Ziv Gan-Or, Thomas Gasser, Peter Heutink, Joshua M Shulman, Nicolas Wood, David A. Hinds, John A. Hardy, Huw R Morris, Jacob Gratten, Peter M. Visscher, Robert R. Graham, Andrew B. Singleton for the **International Parkinson's Disease Genomics Consortium**

Special thanks to Susan Chacko, Mark Patkus and Steve Fellini @ Biowulf HPC.

Basic workflow...



What it takes to accomplish that workflow...

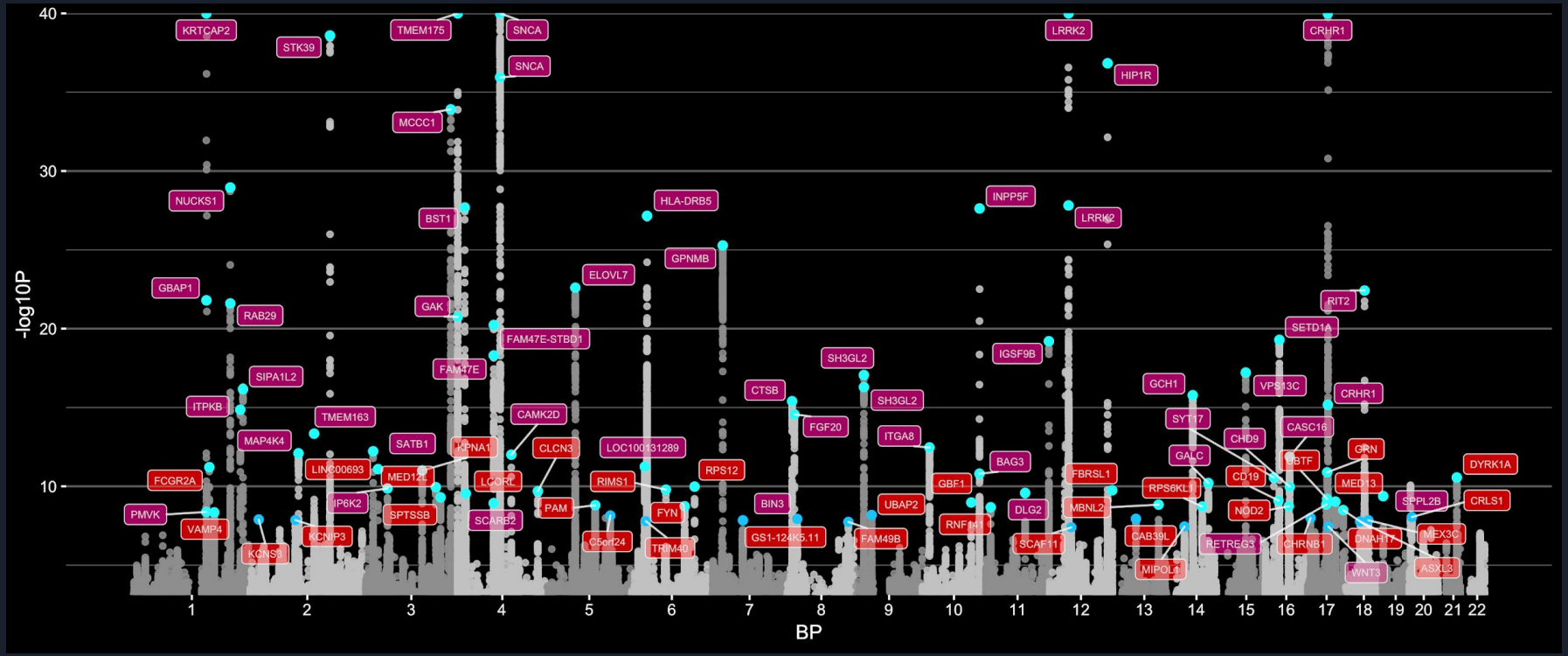
- Huge amounts of genotype data (imputed)
- Lots or reference data from various public sources (Mendelian randomization)
  - QTL datasets
  - External GWAS of other traits
- Web services
  - Pathways
  - GWAS catalogs
  - Imputation servers
- Standardized workflows (GitHub / markdowns / notebooks, more on this later)

This includes over 2.4TB or genotype data at our site alone.

Additional derived and reference data includes another 4.2TB.

In year 2000's terms, this would be a 300ft tall stack of CD-ROMs (w/o cases) on Biowulf.

90 independent common risk factors for PD...



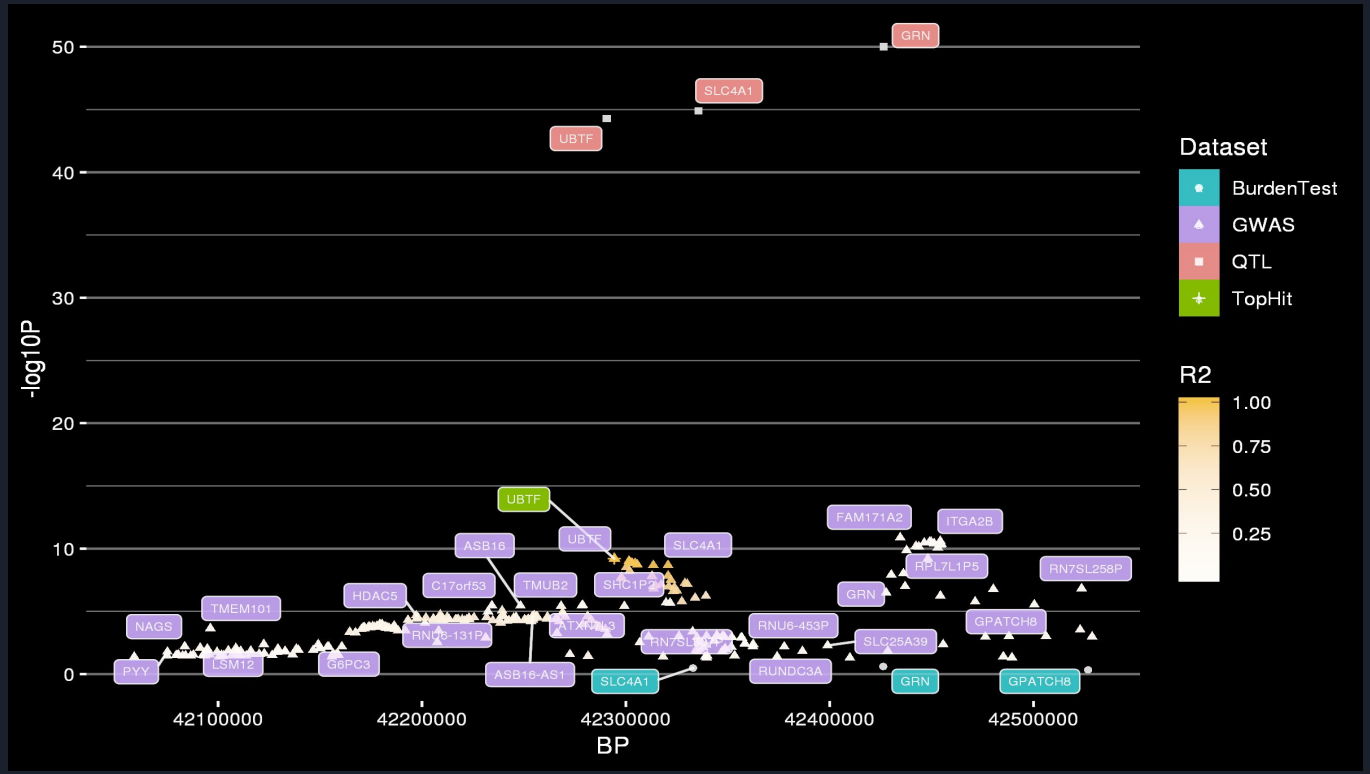
38 risk factors identified are novel hits...

SNP	CHR	BP	Nearest Gene	Effect allele	Other allele	Effect allele frequency	OR	Beta	SE	P, fixed-effects	P, random-effects	P, conditional	I2, %
rs6658353	1	161469054	FCGR2A	c	g	0.501	1.07	0.065	0.009	6.10E-12	3.71E-05	1.38E-05	40.2
rs11578699	1	171719769	VAMP4	t	c	0.195	0.93	-0.070	0.012	4.47E-09	1.09E-07	2.63E-03	5.1
rs76116224	2	18147848	KCNS3	a	t	0.904	1.12	0.110	0.019	1.27E-08	1.27E-08	3.75E-07	0
rs2042477	2	96000943	KCNIP3	a	t	0.242	0.94	-0.066	0.012	1.38E-08	1.38E-08	3.49E-05	0
rs6808178	3	28705690	LINC00693	t	c	0.379	1.07	0.066	0.010	8.09E-12	8.09E-12	8.84E-05	0
rs55961674	3	122196892	KPNA1	t	c	0.172	1.09	0.086	0.013	9.98E-12	9.98E-12	2.80E-06	0
rs11707416	3	151108965	MED12L	a	t	0.367	0.94	-0.063	0.010	1.13E-10	1.77E-07	2.66E-04	10.9
rs1450522	3	161077630	SPTSSB	a	g	0.674	0.94	-0.062	0.010	5.01E-10	2.27E-05	3.51E-04	24.6
rs34025766	4	17968811	LCORL	a	t	0.159	0.92	-0.084	0.013	2.87E-10	2.87E-10	7.43E-06	0
rs62333164	4	170583157	CLCN3	a	g	0.326	0.94	-0.064	0.010	2.00E-10	2.17E-05	5.10E-05	21.3
rs26431	5	102365794	PAM	c	g	0.703	1.06	0.062	0.010	1.57E-09	2.36E-07	6.00E-03	7.9
rs11950533	5	134199105	C5orf24	a	c	0.102	0.91	-0.092	0.016	7.16E-09	2.68E-08	5.08E-04	1.9
rs9261484	6	30108683	TRIM40	t	c	0.245	0.94	-0.064	0.011	1.62E-08	1.62E-08	1.26E-06	0
rs12528068	6	72487762	RIMS1	t	c	0.284	1.07	0.066	0.010	1.63E-10	1.63E-10	9.80E-06	0
rs997368	6	112243291	FYN	a	g	0.805	1.07	0.071	0.012	1.84E-09	1.84E-09	2.61E-05	0
rs75859381	6	133210361	RPS12	t	c	0.967	0.80	-0.221	0.034	1.04E-10	1.04E-10	1.09E-06	0
rs76949143	7	66009851	GS1-124K5.11	a	t	0.051	0.87	-0.143	0.025	1.43E-08	2.04E-06	5.47E-09	12.3
rs2086641	8	130901909	FAM49B	t	c	0.723	0.94	-0.061	0.011	1.81E-08	1.81E-08	6.07E-06	0
rs6476434	9	34046391	UBAP2	t	c	0.734	0.94	-0.062	0.011	6.58E-09	6.58E-09	2.74E-04	0
rs10748818	10	104015279	GBF1	a	g	0.851	0.92	-0.079	0.013	1.05E-09	1.05E-09	7.47E-06	0
rs7938782	11	10558777	RNF141	a	g	0.878	1.09	0.087	0.015	2.12E-09	2.12E-09	2.17E-07	0
rs7134559	12	46419086	SCAF11	t	c	0.404	0.95	-0.054	0.010	3.96E-08	1.84E-05	1.69E-02	25.2
rs11610045	12	133063768	FBRSL1	a	g	0.490	1.06	0.060	0.009	1.77E-10	8.79E-07	3.57E-05	19.5
rs9568188	13	49927732	CAB39L	t	c	0.740	1.06	0.062	0.011	1.15E-08	2.46E-04	4.29E-06	21.4
rs4771268	13	97865021	MBNL2	t	c	0.230	1.07	0.068	0.011	1.45E-09	1.45E-09	1.41E-04	0
rs12147950	14	37989270	MIPOL1	t	c	0.438	0.95	-0.053	0.010	3.54E-08	3.54E-08	1.06E-03	0
rs3742785	14	75373034	RPS6KL1	a	c	0.787	1.07	0.071	0.012	1.92E-09	8.18E-06	2.22E-06	24.8
rs2904880	16	28944396	CD19	c	g	0.309	0.94	-0.065	0.011	7.87E-10	7.87E-10	1.39E-05	0
rs6500328	16	50736656	NOD2	a	g	0.599	1.06	0.059	0.010	1.82E-09	1.82E-09	1.43E-03	0
rs12600861	17	7355621	CHRNB1	a	c	0.648	0.95	-0.057	0.010	1.01E-08	1.01E-08	5.10E-03	0
rs2269906	17	42294337	UBTF	a	c	0.653	1.07	0.063	0.010	6.24E-10	6.24E-10	1.17E-05	0
rs850738	17	42434630	FAM171A2	a	g	0.606	0.93	-0.071	0.011	1.29E-11	2.17E-07	4.18E-04	17
rs61169879	17	59917366	BRIP1	t	c	0.164	1.09	0.082	0.013	9.28E-10	6.21E-06	9.07E-07	16.4
rs666463	17	76425480	DNAH17	a	t	0.833	1.08	0.076	0.013	3.20E-09	4.17E-04	1.62E-05	41
rs1941685	18	31304318	ASXL3	t	g	0.498	1.05	0.053	0.009	1.69E-08	1.69E-08	1.64E-08	0
rs8087969	18	48683589	MEX3C	t	g	0.550	0.94	-0.058	0.010	1.41E-08	1.41E-08	1.09E-04	0
rs77351827	20	6006041	CRLS1	t	c	0.128	1.08	0.080	0.014	8.87E-09	4.38E-07	1.84E-05	11.2
rs2248244	21	38852361	DYRK1A	a	g	0.283	1.07	0.071	0.011	2.74E-11	8.78E-06	6.31E-05	34.3

Gene-level analyses (summary of QTL Mendelian randomization)...

Gene	Probe	CHR	Probe, BP	Top SNP, BP	Top SNP	N SNPs	QTL reference	Effect	SE	Odds ratio	P	Bonferroni adjusted P
VAMP4	ENSG00000117533	1	171,690,343	171,717,417	rs10913587	98	Võsa et al. 2018 - blood expression	-0.272	0.05	0.762	5.67E-07	1.19E-04
KCNIP3	ENSG00000115041	2	96,007,438	95,989,766	rs3772034	14	Qi et al. 2018 - brain expression	-0.161	0.04	0.851	1.12E-05	1.15E-03
MAP4K4	ENSG00000071054	2	102,410,880	102,338,377	rs6733355	3	Võsa et al. 2018 - blood expression	1.119	0.24	3.063	2.32E-06	4.87E-04
TMEM163	ENSG00000152128	2	135,344,950	135,248,544	rs598668	28	Qi et al. 2018 - brain expression	0.074	0.02	1.077	3.55E-07	3.65E-05
KPNA1	ENSG00000114030	3	122,187,294	122,201,610	rs73190142	110	Võsa et al. 2018 - blood expression	0.310	0.05	1.363	1.56E-06	3.28E-04
GAK	ENSG00000178950	4	884,612	906,131	rs11248057	1	Qi et al. 2018 - brain expression	0.508	0.10	1.663	7.47E-07	7.69E-05
CAMK2D	ENSG00000145349	4	114,527,635	114,730,260	rs115671064	146	Võsa et al. 2018 - blood expression	-0.006	0.05	0.994	5.74E-06	1.21E-03
PAM	ENSG00000145730	5	102,228,247	102,118,633	rs2432162	679	Võsa et al. 2018 - blood expression	-0.031	0.01	0.970	2.08E-06	4.36E-04
LOC100131289	cg21339923	6	27,636,378	27,636,378	rs78149975	2	Qi et al. 2018 - brain methylation	-0.094	0.02	0.911	1.53E-06	3.06E-04
TRIM40	cg01641092	6	30,094,300	30,094,315	rs9261443	8	Qi et al. 2018 - brain methylation	0.072	0.01	1.075	6.15E-06	1.23E-03
HLA-DRB5	cg26036029	6	32,552,443	32,570,311	rs34039593	8	Qi et al. 2018 - brain methylation	-0.153	0.02	0.858	7.53E-10	1.51E-07
GPNMB	ENSG00000136235	7	23,295,156	23,294,668	rs858274	74	Qi et al. 2018 - brain expression	0.090	0.01	1.094	2.73E-21	2.81E-19
CTSB	ENSG00000164733	8	11,713,495	11,699,279	rs4631423	33	Qi et al. 2018 - brain expression	-0.150	0.04	0.861	4.37E-09	4.50E-07
BIN3	ENSG00000147439	8	22,502,296	22,456,517	rs71513892	32	Qi et al. 2018 - brain expression	0.046	0.01	1.047	1.43E-06	1.48E-04
SH3GL2	ENSG00000107295	9	17,688,103	17,684,784	rs10756899	15	Qi et al. 2018 - brain expression	0.252	0.05	1.287	5.83E-08	6.00E-06
ITGA8	ENSG00000077943	10	15,659,036	15,548,925	rs7910668	6	Qi et al. 2018 - brain expression	-0.201	0.05	0.818	6.13E-05	6.32E-03
RNF141	ENSG00000110315	11	10,548,001	10,553,355	rs4910153	120	Võsa et al. 2018 - blood expression	-0.054	0.05	0.947	6.25E-07	1.31E-04
IGSF9B	cg25790212	11	133,800,774	133,800,477	rs11223626	1	Qi et al. 2018 - brain methylation	-0.172	0.04	0.842	3.24E-06	6.48E-04
FBRSL1	cg03621470	12	133,137,479	133,138,334	rs10781619	16	Qi et al. 2018 - brain methylation	-0.057	0.01	0.944	6.35E-05	1.27E-02
CAB39L	ENSG00000102547	13	49,950,524	49,918,175	rs35214871	30	Qi et al. 2018 - brain expression	0.097	0.02	1.102	3.51E-08	3.62E-06
GCH1	ENSG00000131979	14	55,339,148	55,348,837	rs3825611	6	Qi et al. 2018 - brain expression	0.113	0.03	1.120	2.76E-04	2.85E-02
SYT17	ENSG00000103528	16	19,229,472	19,273,554	rs727747	4	Qi et al. 2018 - brain expression	0.177	0.05	1.193	1.54E-04	1.58E-02
SETD1A	ENSG00000099381	16	30,982,526	30,950,352	rs7206511	34	Võsa et al. 2018 - blood expression	-0.710	0.09	0.492	2.75E-13	5.77E-11
CHRNB1	ENSG00000170175	17	7,354,703	7,373,595	rs60488855	18	Qi et al. 2018 - brain expression	0.115	0.03	1.122	1.67E-05	1.72E-03
UBTF	ENSG00000108312	17	42,290,697	42,297,631	rs113844752	34	Võsa et al. 2018 - blood expression	-0.466	0.09	0.628	5.68E-06	1.19E-03
MAPT	ENSG00000186868	17	44,038,724	44,862,347	rs199502	6	Qi et al. 2018 - brain expression	0.265	0.03	1.304	7.13E-24	7.35E-22
WNT3	ENSG00000108379.5	17	44,875,148	44,908,263	rs9904865	2	GTEx v7 - substantia nigra brain expression	-0.082	0.02	0.921	4.01E-06	4.81E-05
DNAH17	cg09006072	17	76,425,972	76,427,732	rs589582	3	Qi et al. 2018 - brain methylation	0.100	0.02	1.106	2.44E-05	4.88E-03
MEX3C	ENSG00000176624	18	48,722,797	48,731,131	rs12458916	40	Võsa et al. 2018 - blood expression	-0.291	0.05	0.748	5.28E-05	1.11E-02

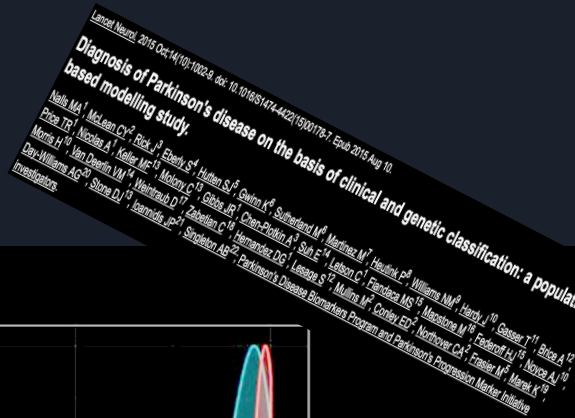
Gene-level analyses (cont'd)…



QTL-MR analyses nominate GRN under this peak, a possible link to FTD.

Quick note: “What is a PRS?”...

- Polygenic risk score (PRS)
- Summary of aggregate independent genetic risk variants for a disease per individual
- Weighted by external GWAS effect estimates
- PRS = B_p*SNP_p + … + B_q*SNP_q
- Consider overall disease heritability, prevalence and sample sizes
- Can be combined with clinical + demographic data for better results (see image to the right)

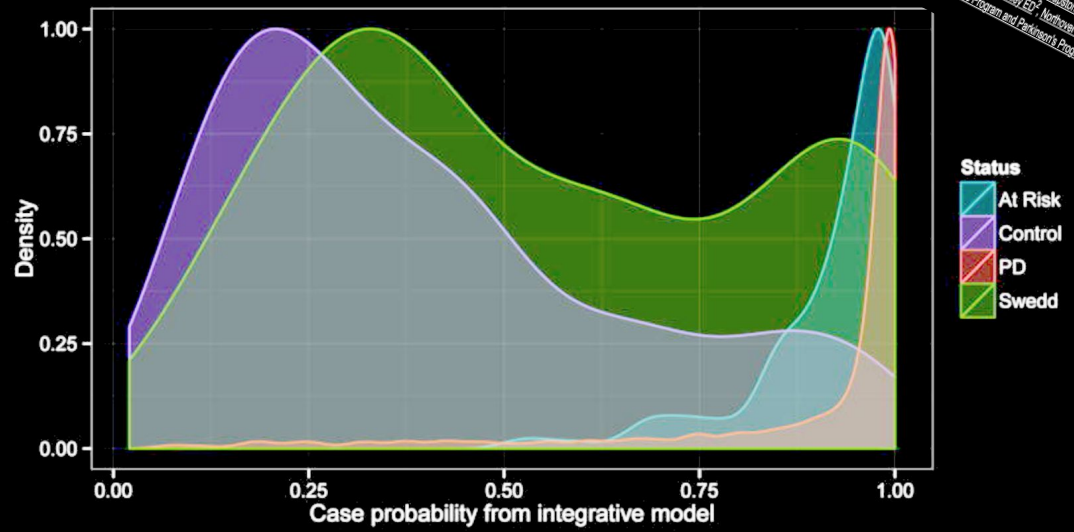

Lancet Neurol. 2016 Oct;14(10):1002-9. doi: 10.1016/S1474-4422(15)00178-7. Epub 2015 Aug 10.

Diagnosis of Parkinson's disease on the basis of clinical and genetic classification: a popular

based modelling study.

Nalls MA[1], McLean CY[2], Rick J[3], Eberly S[4], Hutten SJ[5], Gwinn K[6], Sutherland M[6], Martinez M[7], Hardy J[1,8], Gasser T[11], Brice A[12], Price TR[1], Nicolas A[1], Keller MF[13], Molony CJ[13], Gibbs JR[1], Chen-Plotkin A[3], Sun E[14], Letson C[1], Pandolfo MS[15], Macrorie M[16], Frederick H[1], Novce AJ[16], Morris H[16], Van Deerlin VM[14], Weintraub D[17], Zabetian C[18], Hernandez DG[1], Lesage S[12], Mullins M[6], Conley ED[2], Northover CA[2], Eraser M[6], Marek K[19], Day-Williams AG[20], Stone DJ[13], Ioannidis JP[21], Singleton AB[22], Parkinson's Disease Biomarkers Program and Parkinson's Progression Marker Initiative investigators

Figure 4



Probability of Parkinson’s disease

Density plots of the predicted probability of Parkinson’s disease estimated from the integrative model by participant status. Participant-level data from all studies analyzed except for 23andMe were included in this figure. Densities were smoothed using a gaussian kernel and no particpants had probability estimates of zero.

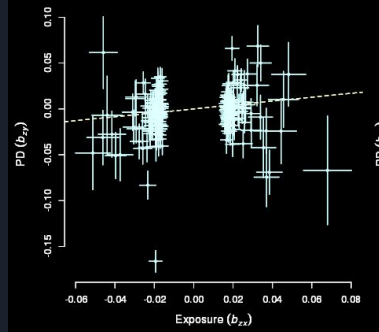
Shared heritability across phenotypes...

Initial nominations of phenotypes via LDSC.

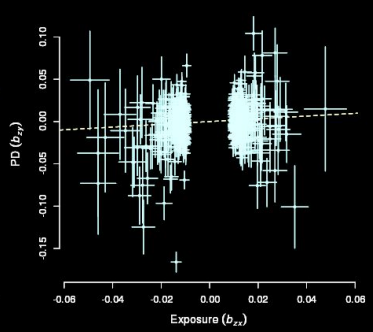
Trait of interest	PMID	correlation,	SE, RG	Z, RG	P, RG	P, FDR adjusted	Observed H2
Intracranial volume	25607358	0.351	0.077	4.580	4.64E-06	3.51E-03	0.166
Current tobacco smoking	Not available (UKB)	-0.134	0.034	-3.947	7.92E-05	2.41E-02	0.055
Mean Putamen	25607358	0.248	0.064	3.902	9.55E-05	2.41E-02	0.282
Qualifications: NVQ or HND or HNC or equivalent	Not available (UKB)	-0.169	0.045	-3.726	2.00E-04	3.79E-02	0.015

Follow-up phenotypes of interest via Mendelian randomization.

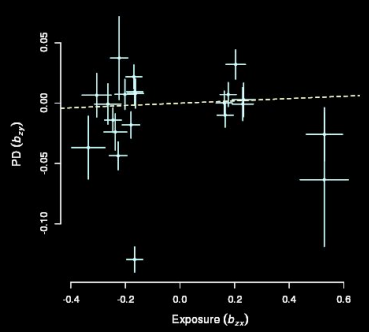
Cognitive performance*



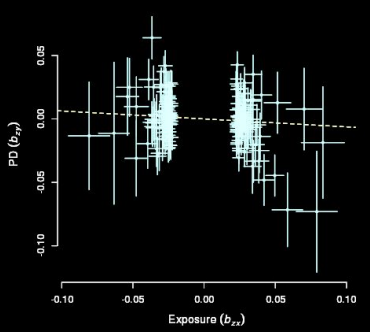
Educational attainment*



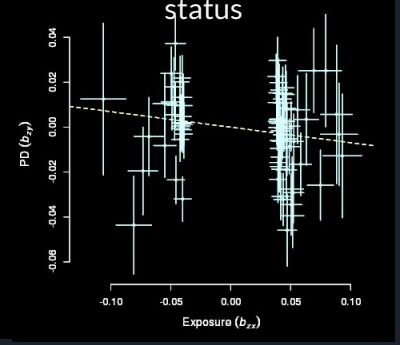
Putamen Volume



Ever smoker

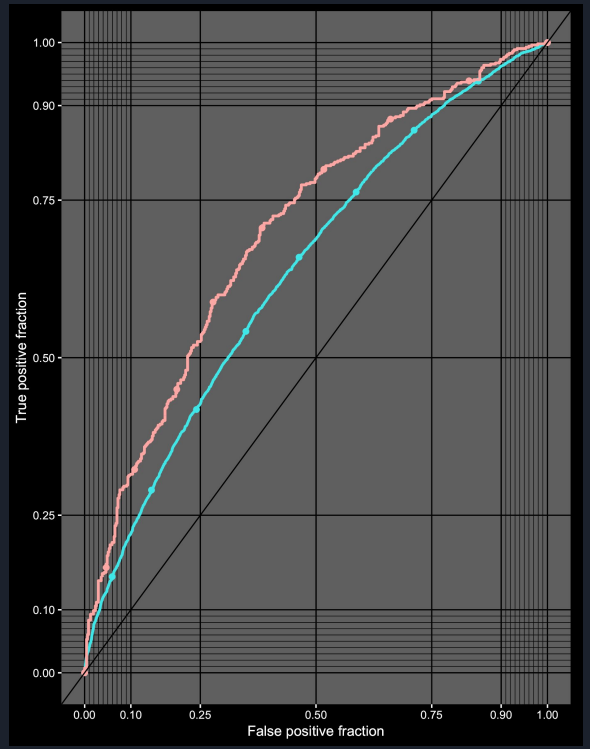


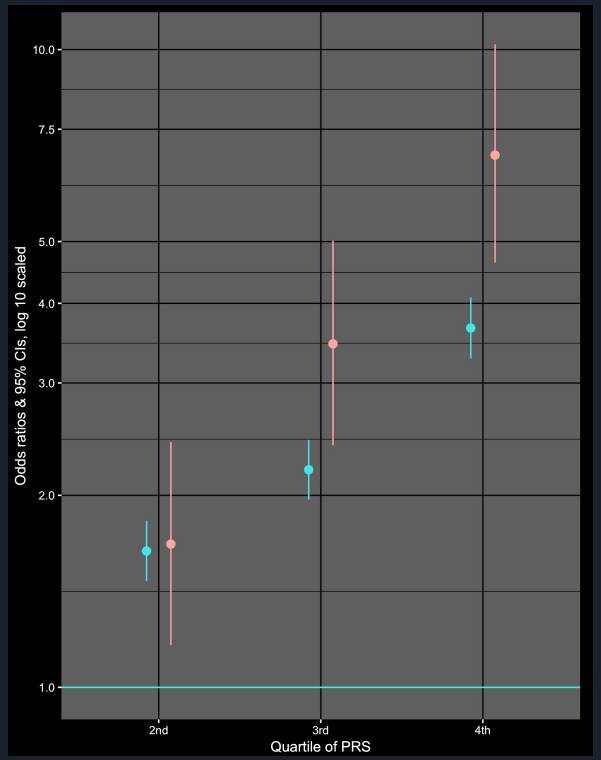
Current smoking status

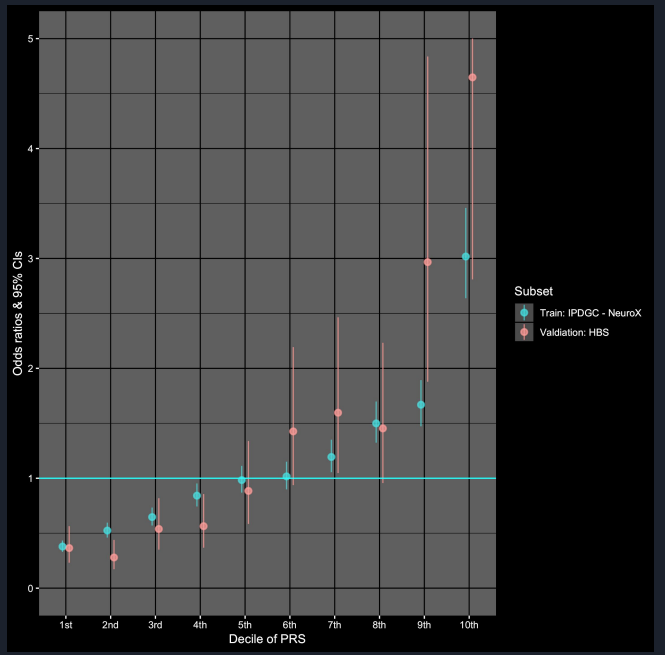


Heritability and polygenic risk estimates...

LDSC heritability ~21%, including 26-36% of heritable risk via GWAS.
Up to 70% AUC at validation with PRS comprising > 1800 variants.







Where are we now with risk locus discovery in PD?

- More loci we have, the more biological insight we gain
- Power calculations based on sub-significant variants from PRS analysis point to 99K cases
  - P < 5E-3 to P < 5E8
  - Once alleles get very rare with moderate effect sizes, hard to build into PRS
- Larger reference data for prioritization and colocalization
- Better imputation panels
- Actively recruiting more diverse populations of PD cases and controls
  - Most important next step
  - Trans-ethnic fine mapping and locus discovery
  - PRS in diverse populations?

Next steps in NDD genetics...

Topics

- Predictors of progression
  - Single outcome like cognitive scores
  - General progression trajectories
- More data from diverse sources and multiple modalities (iPS etc)
  - Huge numbers of collaborators, sites and analysts
- Improved disease predictors
  - Population specific
  - Lower prevalence diseases
- General concerns
  - Applicability
  - Reproducibility
  - Real time sharing
  - Scalability

Tools

- Deeply phenotyped studies
- Machine learning pipelines
  - Supervised for prediction
  - Unsupervised for disease subtyping
  - Federated across silos
- Hybrid cloud for cost efficient collaboration, analyses and reporting
  - Biowulf
  - Google cloud
  - Terra
  - Anthos
- Public code, data and resources
  - Jupyter
  - GitHub
  - Biorxiv

Most of my work is helping to build these tools and infrastructure to facilitate this next evolution of our work in NDDs, particularly as a collaborator w/in the AMP-PD project.
https://fnih.org/what-we-do/programs/accelerating-medicines-partnership-parkinsons

Quick intro to key concepts in machine learning (ML)...

Images from this great review I had no part of.

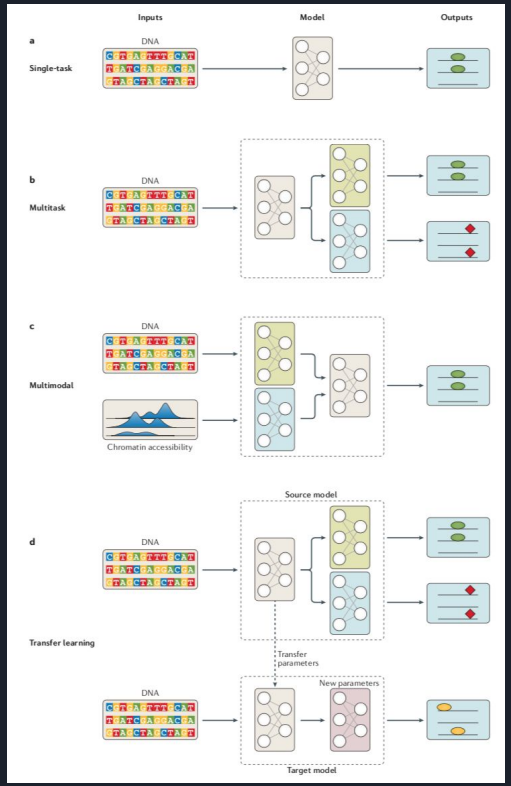
REVIEWS

Deep learning: new computational modelling techniques for genomics

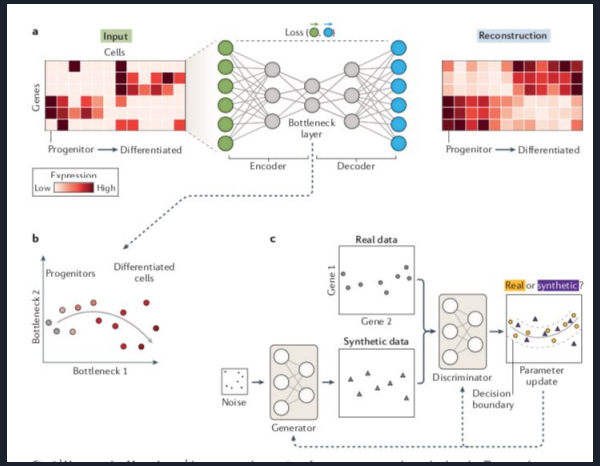
Gökcen Eraslan[1,2,5], Žiga Avsec[3,5], Julien Gagneur[3] and Fabian J. Theis[1,2,4]**

Abstract | As a data-driven science, genomics largely utilizes machine learning to capture dependencies in data and derive novel biological hypotheses. However, the ability to extract new insights from the exponentially increasing volume of genomics data requires more expressive machine learning models. By effectively leveraging large data sets, deep learning has transformed fields such as computer vision and natural language processing. Now, it is becoming the method of choice for many genomics modelling tasks, including predicting the impact of genetic variation on gene regulatory mechanisms such as DNA accessibility and splicing.

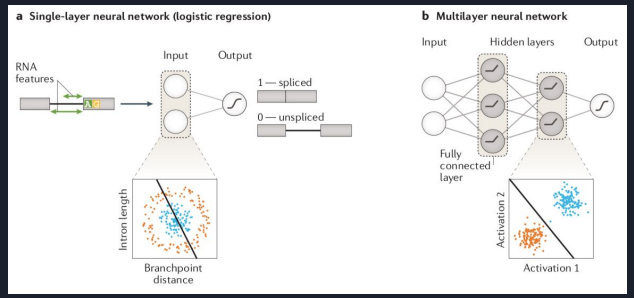
Supervised learning



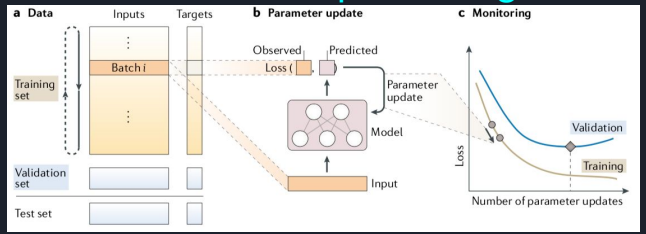
Unsupervised learning



Linear models (regression) are ML.



Cross-validation plus tuning



Case study in supervised ML applications in GWAS (predictions across PD cohorts)...

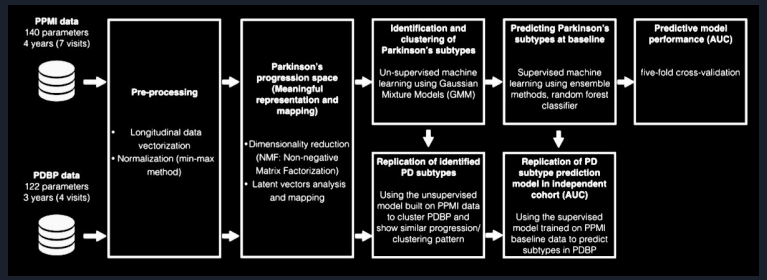
	PRS	ML	PRS plus clinical*	ML plus clinical*
AUC estimate	0.6534	0.6999	0.9218	0.9463
95% CI (DeLong)	0.6437-0.6631	0.6905-0.7092	0.8984-0.9452	0.9283-0.9644
Sensitivity	0.5864	0.6627	0.8310	0.8255
Specificity	0.6346	0.6407	0.9030	0.9515
Positive predictive value (PPV)	0.6861	0.5794	0.9494	0.9739
Negative predictive value (NPV)	0.5298	0.7178	0.7095	0.7136
Balanced accuracy	0.6105	0.6517	0.8670	0.8885

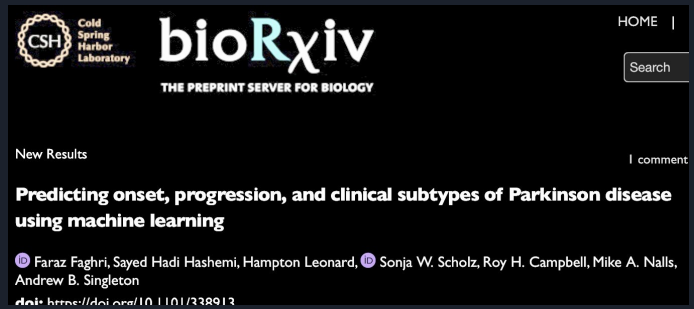
*Note: genetically ascertained sex is included in models, * denotes inclusion of smell test, age and family history in PPMI from previously published discovery series but with PRS from this iteration of results.*

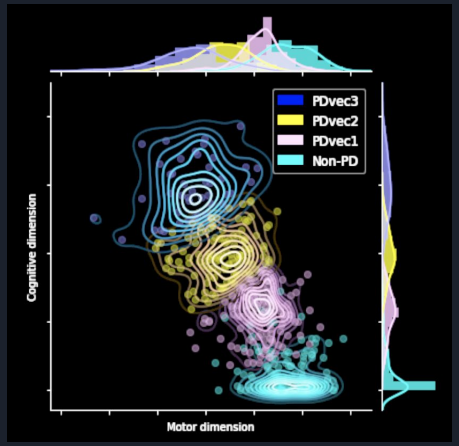
Individual AUC estimates for consituent cohorts include: Finland 0.6870, Germany 0.6447, HBS 0.7102, McGill 0.6606, MF 0.6959, NIA 0.6350, Oslo 0.6827, PDBP 0.6407, PPMI 0.6167, Baylor 0.7481, SPAIN 0.6754, Vance0.6712, and NeuroX - dbGaP 0.6265.

- In a head to head comparison at validation, supervised ML methods tend to outperform PRS methods often by > 1% AUC on the same samples as PRS
  - PRS at best P thresholds from external GWAS.
  - No external GWASs needed for ML
    - Better suited for custom phenotypes with small sample sizes and little to no external GWAS.
  - Can be more computationally intensive in some instances
  - Advantageous when solutions to problems are non-linear
- Detailed analysis provided in upcoming Spanish GWAS cohort
  - Unique risk structure in Spain changes variant weights in standard PRS
  - Juan Botia and Sara Bandres-Ciga for IPDGC

Subtyping disease with unsupervised learning…







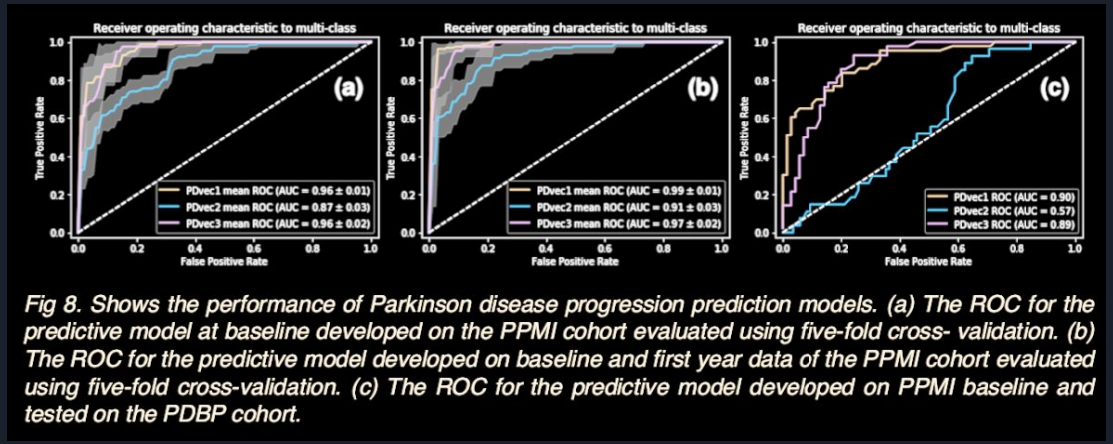
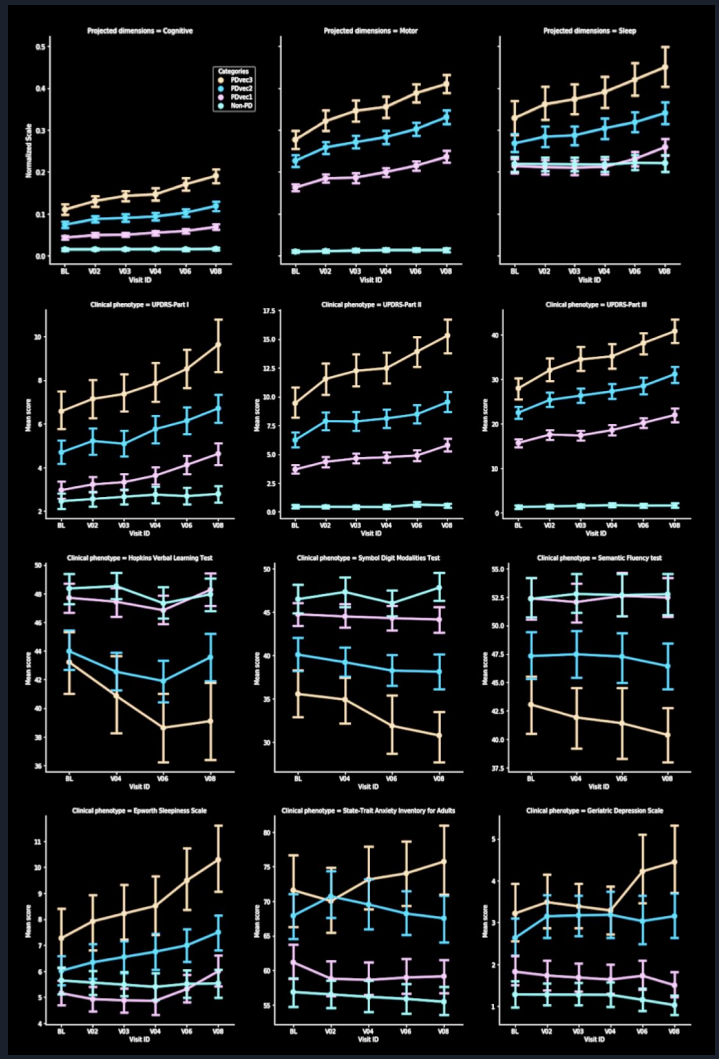


Fig 8. Shows the performance of Parkinson disease progression prediction models. (a) The ROC for the predictive model at baseline developed on the PPMI cohort evaluated using five-fold cross- validation. (b) The ROC for the predictive model developed on baseline and first year data of the PPMI cohort evaluated using five-fold cross-validation. (c) The ROC for the predictive model developed on PPMI baseline and tested on the PDBP cohort.

Subtyping disease with unsupervised learning...

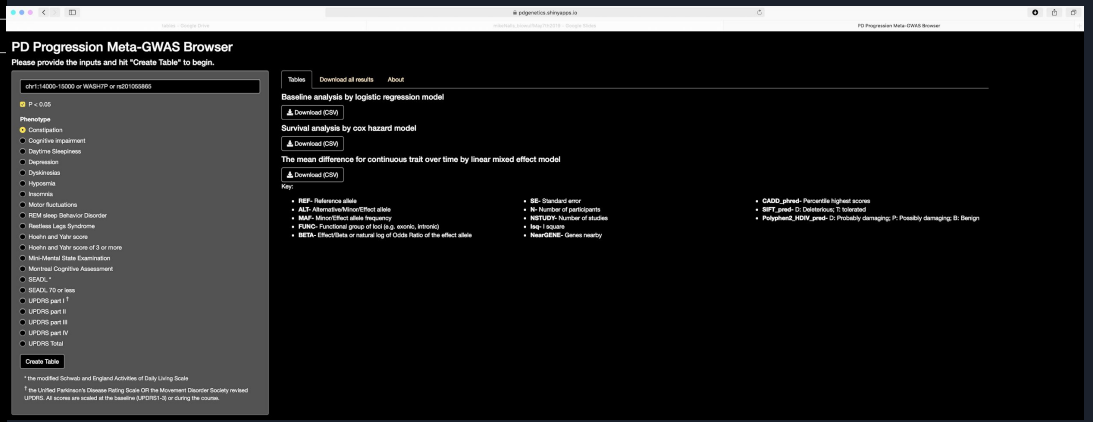
Predicting progression is important for drug dev:

- Trial readouts
- Patient recruitment
  - Fast progressers
  - Same “flavor” of disease
- Rescue failed trials in more homogenous samples

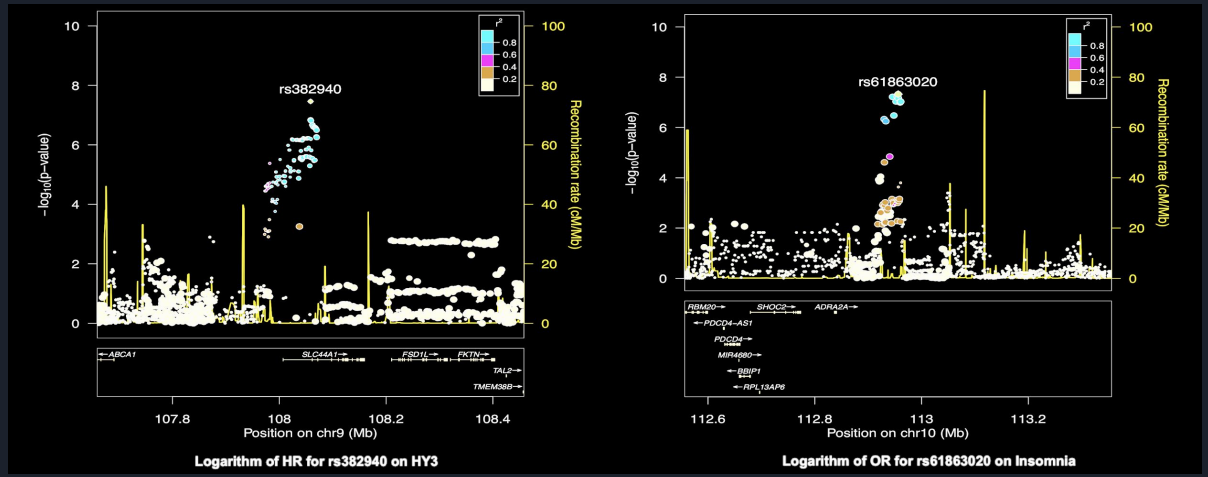


Case study: PD progression meta-GWASs ...

- Well powered GWAS exist for case-control status
- Not much exists for progression markers in cases
- Integrating many data silos is a growing problem
- 12 longitudinal cohorts with 4,093 patients
  - 25,254 observations
  - Median of 3.81 years

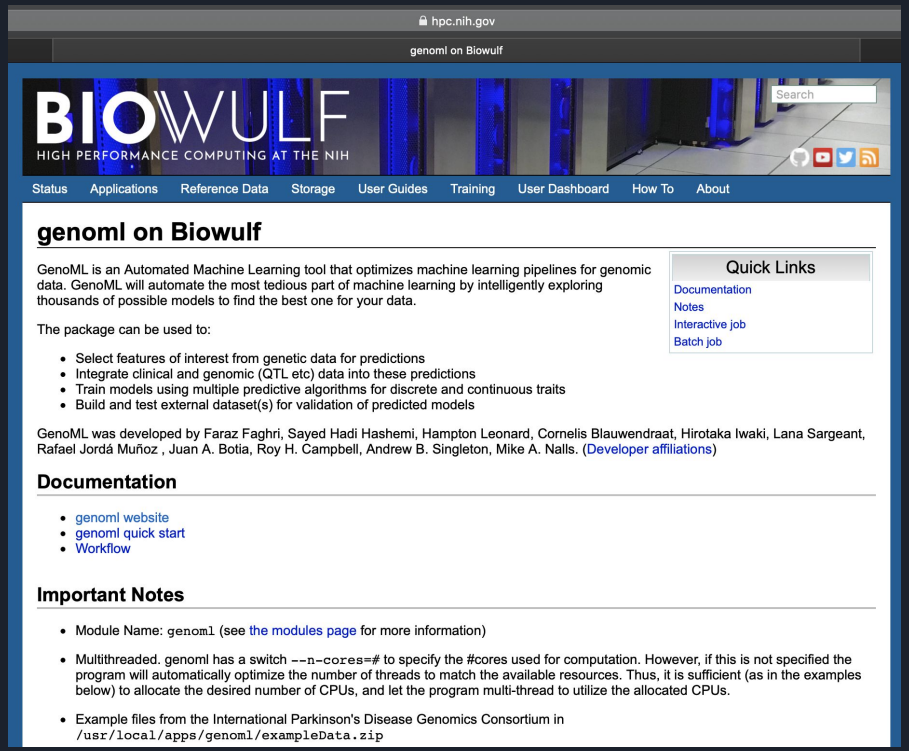


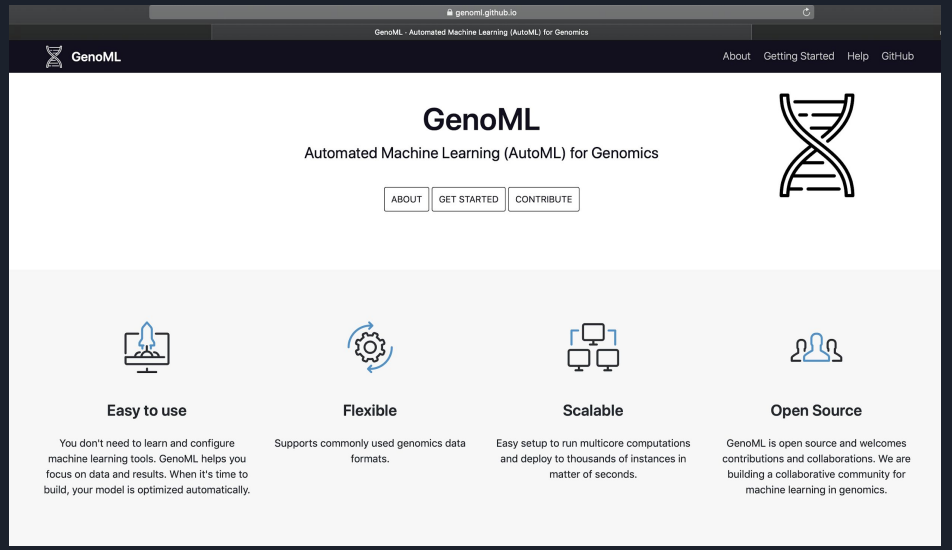
Case study: PD progression meta-GWASs ...



- In this instance, ML + cross-validation makes more sense than PRS
- High translational value in future predictors of specific outcomes
- Studies like this and future EMRs / data silos
  - Federated learning efforts
  - Learn across data silos by sharing aggregate summary data
  - Centralized analysis server
  - Your phone already does this

Tools we are working on (GenoML)...

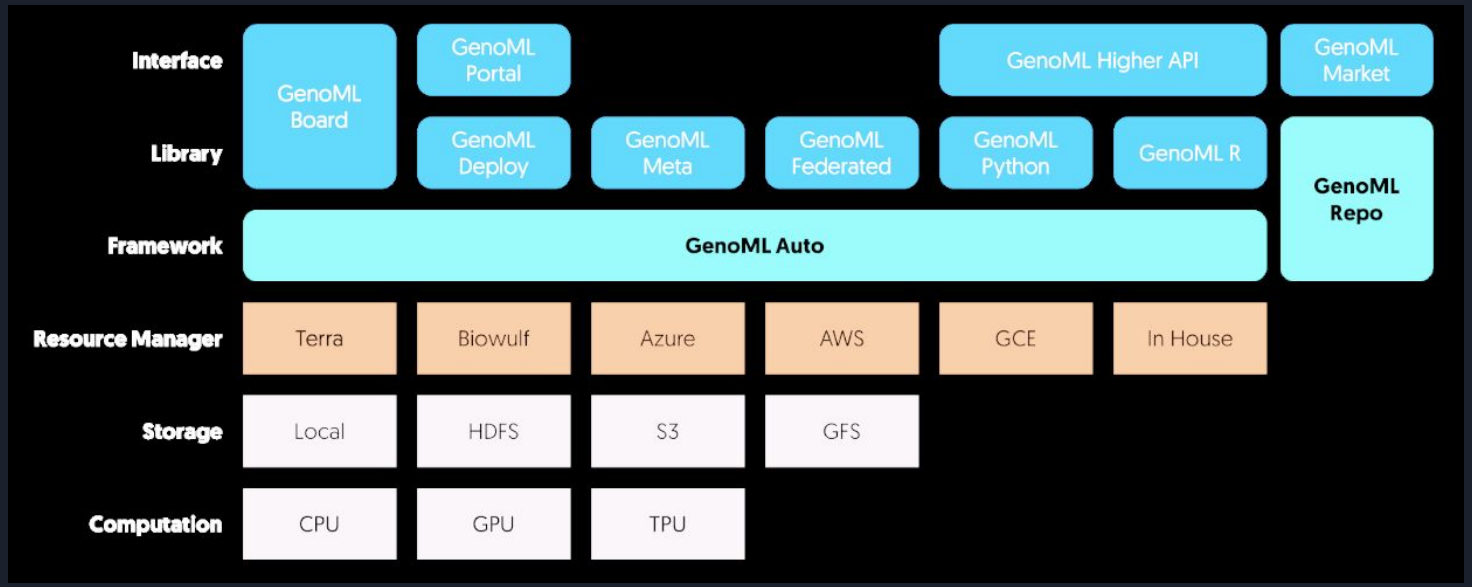




Team

Faraz Faghri (NIA-NIH, UIUC), Sayed Hadi Hashemi (UIUC), Hampton Leonard (NIA-NIH / DT), Cornelis Blauwendraat (NIA-NIH), Hirotaka Iwaki (NIA-NIH / MJFF / DT), Lana Sargeant (VCU), Susan Chacko (Biowulf-NIH), Rafael Jordá Muñoz (UM), Juan A. Botia (UM), Roy H. Campbell (UIUC), Andrew B. Singleton (NIA-NIH), Mike A. Nalls (NIA-NIH / DT)

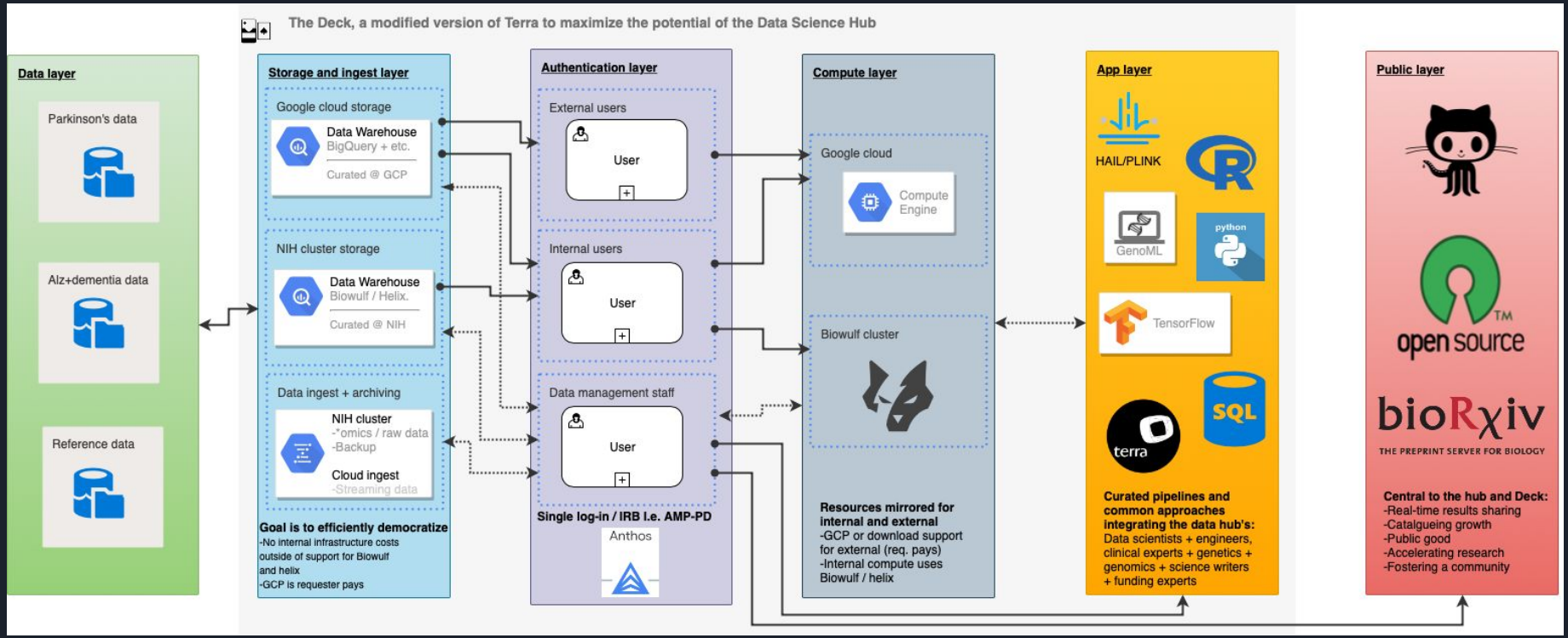
Tools we are working on (GenoML)...



May 2019	v 1.0	GenoML Auto
Sep 2019	v 2.0	GenoML Deploy, Repo
Dec 2019	v 3.0	GenoML Meta, Federated
Mar 2020	v 4.0	GenoML Python, R, Higher API
June 2020	v 5.0	GenoML Market, Portal, Board

Concepts we are exploring (Terra / TheDeck)...

Inspired by work going on at Global Alliance for Genomics & Health (Collaborate. Innovate. Accelerate.) and similar initiatives.



Thanks!

This was alot to cover today!

Slides available at https://github.com/neurogenetics/talks

If you are interested in collaborating on any of the topics I have just discussed please get in touch...

- Analysis projects?
- Contributing to GenoML's ecosystem?
- NDD sample genotyping / sequencing?

Also, we are recruiting 8 data scientists and computer scientists for different NDD initiatives at NIH.

Email: mike@datatecnica.com / nallsm@mail.nih.gov