**To do list**

* High priority
  + NRSA
    - Send specific aims draft to Lori by 23rd (for meeting on 30th)
    - Who to have as collaborators / co-sponsors: check based on examples, how many do I want, what roles to fill, training plan also for classes
    - Go through Lori’s comments to try to address
  + Make stress-sensitivity PRS with GenoPred using PRS-CS
  + MDD PRS
    - Compare genopred ancestry w Mahnoor/Jinhan’s ancestry
    - PRS-CS couldn’t get reference data for CSA group?
* Medium priority
  + NRSA
    - LASSO analysis possibility
      * How does LASSO handle missingness?
      * Why haven’t people used LASSO as much?
  + MDD PRS
    - Figure out if there are any snps in common between stress-sensitivity and MDD GWAS
  + Make PTSD PRS
  + Make ADHD PRS
  + Genopred
    - Optimize workflow (parallelization)
    - See if there’s a way to find variance of proportion explained for PRS (try for basic PRS I already have vs other techniques in genopred)
* Low priority
  + Lit reviews for structural, functional differences
  + Check out enigma gradient
  + Read papers
  + Clean up notes
  + Figure out who to follow on twitter
  + CITI training (see Sam's slack message)

**7/15/2024 Jen meeting**

* NRSA
  + Haven’t heard back from Shannon or Mary, got good feedback from Lori (waiting on examples from students) and second example NRSA from Victoria
  + Meeting with Lori Tues July 30 – aim to send specific aims draft by 23rd
  + Sign up for Sept 5 training? <https://www.washington.edu/research/required-training/biomedical-research-integrity-program-nih-required-responsible-conduct-of-research-rcr-training/>
  + Concern: previous reviews talked about “high chance of significant findings” – rn don’t have anything really significant for stress-sensitivity PRS
    - This led me to try to make stress-sensitivity PRS with genopred to streamline everything, maybe I did something wrong the first time, ran into issues (see highlight below)
    - What if the preliminary data I have is based on MDD, ADHD, PTSD PRS? Could show that know enough to replicate and extend some previous findings. Would also be easier to write [instead of having to try to explain why we’re not seeing what we expected to see initially], but would look weird to not have started with stress-sensitivity PRS for preliminary data?
  + Shannon said talk to project officers at NIH early – write up a draft of specific aims and send in email then set up phone call (do this way in advance, like now)
  + Human subjects use: exception states, “Secondary research for which consent is not required: Secondary research uses of identifiable private information or identifiable biospecimens, if at least one of the following criteria is met: … ii. Information, which may include information about biospecimens, is recorded by the investigator in such a manner that the identity of the human subjects cannot readily be ascertained directly or through identifiers linked to the subjects, the investigator does not contact the subjects, and the investigator will not re-identify subjects”
  + Am I pitching this as my dissertation?

**7/9/2024 – 7/14/2024**

* NRSA
  + Started gathering info from sharepoint and examples from other students
  + Fully due Dec 8, due to psych grants team 10 days ahead of time
  + Tips doc says need to check with department grant person on whether consultants/mentors can write recommendation letters? (Sponsors will definitely write letters)
  + Tips doc recommends sending potential mentors/consultants specific aims: Some students have wondered how to ask someone to serve as a consultant/personnel on their NRSA. Previous students have recommended sending an email with the request and including a description of what they specifically hope the individual with provide consultation with as well as a copy of the Specific Aims. It should be clear why you are picking this person and what you hope they will be training you in.
  + Human subjects use: exception states, “Secondary research for which consent is not required: Secondary research uses of identifiable private information or identifiable biospecimens, if at least one of the following criteria is met: … ii. Information, which may include information about biospecimens, is recorded by the investigator in such a manner that the identity of the human subjects cannot readily be ascertained directly or through identifiers linked to the subjects, the investigator does not contact the subjects, and the investigator will not re-identify subjects”
  + Google doc with tips: <https://docs.google.com/document/d/1oJJiA0_0m9NxJAs4u16ecSBWVe5O0BVdi89EXuhfgfs/edit>
  + Team
    - Jen and Lori as co-mentors (like mentor to have history of supervising other NRSA students per fellowship tips doc on intranet)
    - Matt Conamos or Kevin as stat consultant (from tips google doc on intranet, but also see p 7 of that doc – do/do not use word consultant?: “Faculty and previous students have also found that clinically-focused NRSAs have received better scores if the mentorship team includes a statistical consultant, unless the mentor or co-mentor have significant statistical expertise.”)
    - Someone as imaging consultant?
    - Someone as environmental variables consultant?
    - Note that don’t want too many people – from google doc w tips: One thing to consider is the size of your mentorship team. You want to ensure you have sufficient mentorship throughout the process, however proposing too many mentors has received negative commentary from reviewers. Ultimately they are concerned that the project, including your training, be sustainable and doable.
* Genopred
  + Jinhan got running on klone, will work on optimizing next week
  + Stress-sensitivity PRS
    - Based only on SNPs from MR (stress1 folder)
      * To make sumstat table based on snps shown to be linked to psych disorders by mendelian randomization by penner-goeke, took the 11 SNPs from table S6 that were responsive to dex only
      * Cross-referenced with table S1 and found that column named “Ref” in table S6 corresponded to “Ref” in table S1 and column names “Effect” in table S6 corresponded to “Alt” in table S1, so changed column names from (Ref, Effect, FDR) to (REF, ALT, P) to match expected genopred headers
      * In gwas\_list set n to be 160, sampling to be EUR because Arloth eQTL based on 160 Caucasian men
      * From gwas sumstats cleaned log: “after matching variants to the reference, 2 variants remain” → can I do anything about this?
    - Stress2 folder: same as stress1 but using UK biobank as reference to see if can keep more variants, didn’t work → still get 2 variants remain from gwas sumstats cleaned log
    - In original stress-sensitivity PRS with just basic PRS (not genopred or PRS-CS) skipped 3 SNPs from MR and 10 total
    - Stress 3 folder: same as stress 1 but instead using ptclump instead of PRS-CS, still get log saying only 2 variants remain
    - Stress 4 folder: same as stress 1 but now using output > renamed\_scz\_ABCD\_imputed > ancestry and geno folders from mdd3, ran okay but still get 2 variants remain message
    - Stress 5 folder: same as stress 1 but now using gwas\_list with all 78 SNPs from S3 in Penner-Goeke, output says only kept 21 total after matching to reference
    - Looks like the XX variants remain error is maybe due to not findings the variants in the reference data in pipeline > resources > data > prscs\_ref > ldblk\_1kg\_eur?
    - Asked JInhan for reference data she and Mahnoor used to maybe sub in for default ref in genopred, Jinhan said she thinks they’re the same ref data
    - Variants not kept by genopred ie not 21 kept in stress 5 folder (eg rs1000464, see S3 modified for genopred tab in notes from Penner-Goeke 2023 supp tables.xlsx file for full list of snps or compare cleaned gwas sumstats and ref gwas sumstats in genopred files) should be in thousand genomes reference data set because in USCS genome browser which uses 1kg data (double check with Jen)
    - Looks like getting removed as part of sumstat\_cleaner.R script, specifically ref\_harmonise or maybe filter\_info commands, for details see: <https://github.com/opain/GenoUtils/blob/50ac8a2078226c8c2349064f904031576fbfe606/R/sumstat_cleaner.R#L419>
* Data behind stress-sensitivity PRS
  + Arloth: performed eQTL which has 320 loci (LD bins) corresponding to 3662 eSNPs linked to dex responsiveness
  + Penner-Goeke:
    - said the 3662 eSNPs were enriched in enhancer regions table S1
    - performed fine-mapping with STARR-Seq using two different cell lines (U2OS cells transfected with glucocorticoid receptors ie U2OS-GR and U138MG)
    - after 4h incubation with 100nm dex found 547 dex-responsive regulatory element ie DREs (but based on supp tables I think that’s a typo and they meant 574), basically trying to answer in these cell lines does which elements change activity in response to dex (ie which are DREs, does the dex actually make a difference), see table S2
    - of the 547 (or 574) DREs there were different levels of expression for the alt vs ref allele (ie answering does the SNP actually make a difference in gene expression) for 154 DREs in the combined veh only and both veh and dex condition; 174 (but I think it’s a typo and they meant 172) DREs in the combined dex only and both veh and dex condition; 93 DREs in both the veh and dex condition; 81 DREs in the dex only condition
    - Of the 81 regulatory elements which responded to dex (ie DREs) in which the ref vs alt allele (ie SNP) actually made a diff in gene expression, 78 were used to make the functional gene score because two were sig for dex only in one cell line but sig for both dex and vehicle in other cell line and didn’t give reason for why excluded third SNP (I checked and it wasn’t in the MHC)
* ABCD
  + Gender: conversation with Carly – discrepancies in youth vs parent-reported gender, youth data available from GISH survey starting at year 3 follow-up

**7/8/2024 Jen Meeting**

* NRSA
  + Sent emails discussed last week, waiting to hear from Lili, Lori and Mary
  + Thao said: department point person for grants starting this week is Yuichi Fukuda (he/him)
  + Neuroacademy – look at genetics
  + Email Shannon for examples
* GenoPred
  + Put pipeline itself in packages folder in Hyak, put my data under ABCD still; check sharepoint notes for troubleshooting on klone, can also reach out to Altan if stuck and Jinhan’s not here
  + PRS-CS LD reference from 1KG (default) or UK Biobank? Use 1KG (Thousand Genomes)
  + PRS-CS phi: can set own list of thresholds or use auto? “In the second version of the algorithm, which we call PRS-CS-auto, we use a fully Bayesian approach and place a standard half-Cauchy prior on the global shrinkage parameter[19](https://www.nature.com/articles/s41467-019-09718-5#ref-CR19),[20](https://www.nature.com/articles/s41467-019-09718-5#ref-CR20): *ϕ*1/2 ~ C+(0, 1), such that *ϕ* is automatically learnt from data and no validation data set is needed.” – from (Ge et al., 2019)
* MDD PRS
  + Generated scores with PRS-CS using GenoPred pipeline with some help from Jinhan
  + Based on paper mentioned last week, looks like best option (better than PRS-CSx actually) is PRS-CS
  + Good news is that Meng 2024 has separate summary stats for African, Hispanic, East Asian, South Asian ancestries; still using Howard 2019 for Eur ancestry
  + Genopred breaks down ancestry into African (AFR, n=1716), admixed American (AMR, n=1693), east Asian (EAS, n=106), European (EUR, n=5789), central and south Asian (CSA, n=55), middle eastern (MID, n=5)
  + Cut CSA and MID, double check EAS to see if usable

**7/2/2024 – 7/7/2024**

* NRSA
  + Department point person for grants starting this week is Yuichi Fukuda (he/him)
  + Fellowship application website: <https://uwnetid.sharepoint.com/sites/uw_psych/graduate/fellowships/SitePages/Home.aspx>
  + Need to submit within department a week before deadline
  + Who to have as collaborators / co-sponsors: check based on examples, how many do I want, what roles to fill, training plan also for classes
  + LASSO analysis possibility
    - How does LASSO handle missingness?
    - Why haven’t people used LASSO as much?
* Make MDD PRS
  + GenoPred
    - Make sure to use --use-conda when try to run genopred command to get PRS so it finds plink
    - Mdd1 folder just contains stuff needed to make cleaned up GWAS summary stat file (for use in klone)
    - Mdd2 folder contains ABCD data and GWAS summary stat for all ancestry (only one file), PRS created with ptclump
    - Mdd3 folder: now trying to use PRS-CS, data for African, east Asian, south asian and Hispanic/latino ancestries from Meng 2024 and data for European ancestry from Howard 2019; note that PRS-CS was unable to get reference data for CSA group so took out line: [MDD\_SAS\_MENG2024 mdd/mdd\_data/reference/gwas\_sumstats/mdd2023diverse\_SAS\_Neff CSA 31681 0.142198794 0.1 NA NA "MDD\_SAS"]
  + See if there’s a way to find variance of proportion explained for PRS (try for basic PRS I already have vs other techniques in genopred)
  + ABCD paper comparing PRS methods (Ahern et al., 2023) found single ancestry PRS-CS (PRS-CS for African, European, admixed ancestry) performs much better than PRS-CSx or PRS-CSx Meta even if single ancestry isn’t same as target eg CS with European ancestry better than CSx for admixed target group, similar pattern for depression or schizophrenia PRS scores using KSADS or CBCL total problems as outcomes

**7/1/2024 Jen Meeting**

* NRSA
  + Lori said will review, may not get back by today
  + How does LASSO handle missingness?
  + Why haven’t people used LASSO as much?
  + Responsible conduct of research training, neuroacademy and ariel rokem, matt conamos (biostat, created analysis packages)
* GenoPred
  + Found newer MDD meta-GWAS with summary stats broken down by ancestry group
  + Got full tutorial to run with sample data
  + Got pipeline to run through cleaning up GWAS summary stat
  + Put cleaned up summary stats into basic PRS, did get some warnings but just due to SNPs present in summary stats not in ABCD data
  + Now working on calculating PRS with more advanced methods using full GenoPred pipeline – which option?
  + See if there’s a way to find variance of proportion explained for PRS (try for basic PRS I already have vs other techniques in genopred),
  + Look for ABCD paper that came out a year or two ago from genetics folks in ABCD group, talk about genesis package, see if can find because they compared methods

**6/24/2024 – 7/1/2024**

* Second-year project proposal: July 1 (may take longer for Lori to review)
* NRSA
  + Meet with Lori
  + Mary Larimer mock review: see email from Ange
  + Ask Meyer (and JP?) for letter
* General PRS stuff
  + QC
    - GenoPred
      * To get to files in linux/ubuntu part with normal file explorer put \\wsl$ into the search bar, go to home > kscheuer > genopred > GenoPred > pipeline
      * Don’t change config.yaml file in pipeline folder, instead make copy and put in new folder for each project so can customize
      * Don’t think there’s a way to use chromosome and position instead of rsid, but doesn’t seem to actually matter in this case because EA and NEA are D and I if no rsid for MDD summary stats?
    - Use chromosome and position instead of rsid?
  + BridgePRS: Jacquelyn is working on making instructions list
  + Check that all PRS are normally distributed
* General analysis stuff
  + For interactions with genesis package either put both in quotes "a\*b" or either "a"\*"b"
  + Try Z-scored and CMC vs not for all continuous variables
  + Suggestion from kevin - equivalence testing to determine whether the effects you're seeing are consistent with the null - see ref 115 from baldwin 2022
  + Mahnoor and Jinhan did not use site as a covariate
  + Meet with stats (see email from sam), start working on prs for other disorders, look at reri ie interaction contrast ratio see guloksuz citation 48 and 49 venderweele tutorial on interaction (and look at daskalakis suicide paper)
* MDD PRS
  + GenoPred
    - Now using data from Meng 2024 instead of Howard 2019
    - Cleaned summary stats
    - Decided to use 10% as prevalence but can vary – see Marx et al 2023 “Major depressive disorder”
    - Used EUR as reference because recommended that for mixed ancestry use largest ancestry group as reference
    - Put summary stats into normal/basic PRS in plink to make MDD PRS scores, note: used effect allele column when put into plink
  + Output log info for making PRS just normal/basic way using Meng 2024 summary stats
    - 9670438 variants in bim file
    - 11101 from fam file
    - Total genotyping rate is in [0.9999995, 1)
    - 9670438 variants and 11101 people pass filters and QC.
    - Warning: 16872 lines skipped in --score file (16788 due to variant ID mismatch, 84 due to allele code mismatch) → I think these are SNPs that aren’t present in ABCD data because no overlapping rsids between SNPs in nopred file and SNPs in ABCD bim file? Jinhan says this is true)
    - 1118258 valid predictors loaded
  + Problems with Howard 2019 version
    - Flipped alleles
      * Need to do this in plink
      * Compared to ABCD bim file some alleles are flipped (ie MDD rs4040617)
      * Plan: create list of ids with flipped alleles, put that variant list into flip command, replace remove with flip <- this didn't work, didn't recognize flip as command
    - Mismatch between rsid in ABCD bim file and summary stats? Drop if ambiguous whether same SNP
* ADHD PRS
  + Mismatch between loaded alleles for variant and allele in file
  + Variant ID mismatch
* Anxiety PRS
  + Check papers Jen sent instead of preprint I found to try to find cross ancestry info
* Stress-sensitivity PRS
  + look for list of 542 or whatever genes to see if can make broader
  + match rsid to hg38 from summary stats
  + effect vs reference vs alt allele column?