“Big Data” Approach to Stroke

- Individual level data (genomic and phenotypic)
- Standardized pipelines for identification, retrieval, Qcing and storage of these data
- Individual level data yields much more than summary-stats based collaborations
  - Standardized QC
  - Identification of duplicated controls
  - Genetic overlap
  - Comparison of LD structure
  - Assessment of systematic sources of bias

Psychiatric Genomics Consortium

Total Sample Size	Adult height	Crohn's Disease	Schizophrenia
	n of GW loci per 5,000 subjects	n of GW loci per 1000 cases / 1000 controls	n of GW loci per 3000 cases / 3000 controls
1x	0	2	1
2x	2	4	2
3x	7	5	6
9x	68	51	62
18x	180	-	-

Psychiatric Genomics Consortium

- Individual level data
- Minimalist approach to data requirements
  - Age, sex, case/control status
- Participating groups receive a “processed” version of their dataset
- Data stored in one place, which each group can access to conduct analyses
  - i.e. data are not downloaded

Proposal

- To develop a comprehensive repository of individual-level data, from both SiGN and non-SiGN data, which will allow analyses otherwise unfeasible with summary statistics-based approaches.
  - 1. To identify and retrieve additional sources of genetic and phenotypic data beyond SiGN, as well as publicly available data on annotation and other omics (transcriptomics, proteonomics and metabolomics)
  - 2. To harmonize quality control procedures for genomic data and outcome ascertainment for phenotypic data
  - 3. To evaluate genetic overlap between stroke types (ischemic and hemorrhagic) and subtypes
  - 4. To increase and promote access to data, analytical tools and results by making the repository available to researchers from different disciplines interested in stroke genetics

Others are doing this

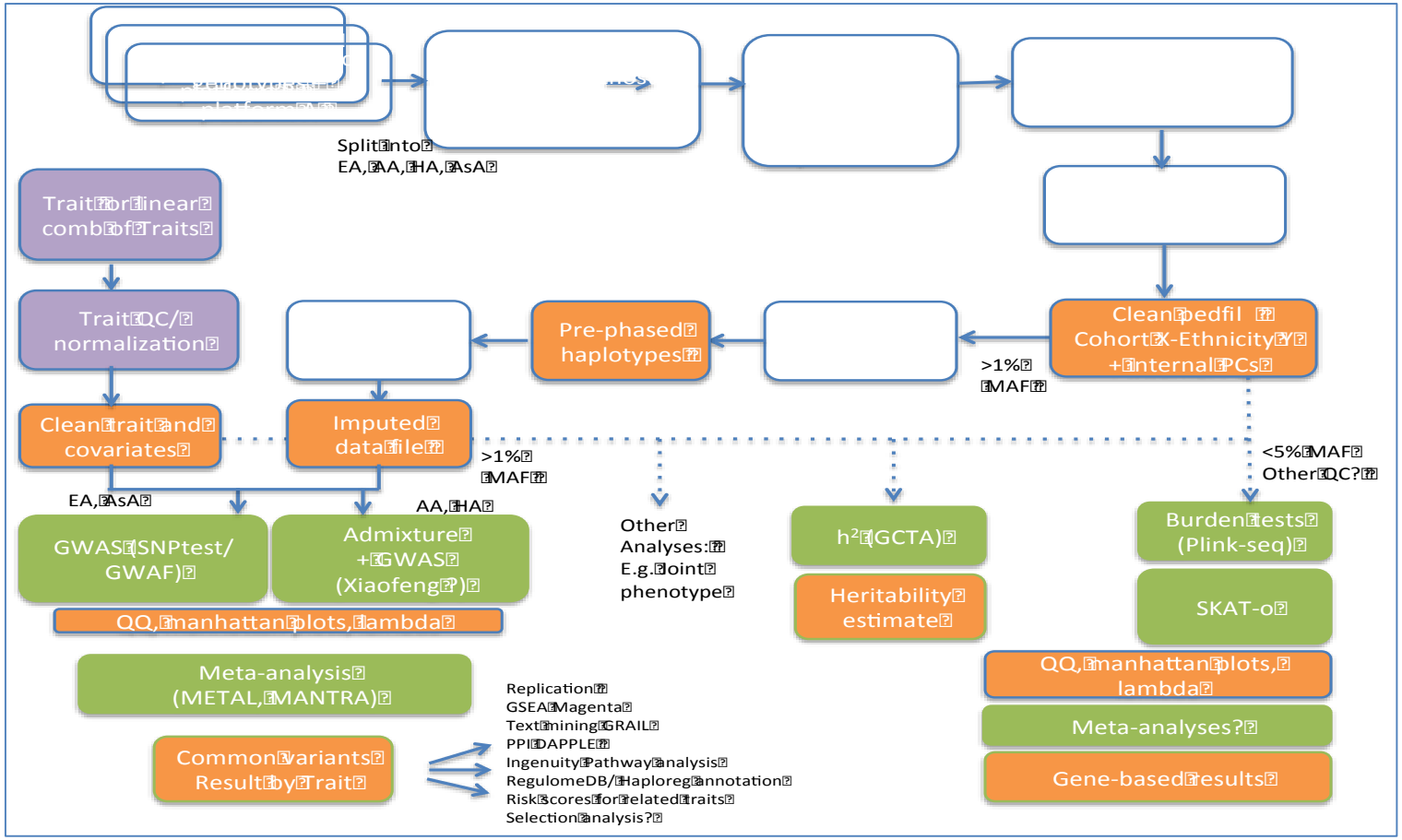
- To develop a comprehensive repository of individual-level data, from both SiGN and non-SiGN data, that will allow analyses otherwise unfeasible with summary statistics-based approaches
  - Psychiatric disorders
  - Diabetes
  - Sleep disorders
  - ICH Phase II

Aim 1

- To identify and retrieve additional sources of genetic and phenotypic data beyond SiGN, as well as publicly available data on annotation and other omics (transcriptomics, proteonomics and metabolomics)
    - SiGN
    - ISGC
    - METASTROKE
    - Welcome trust
    - db-Gap
    - Kaiser Permanente

Aim 2

- To harmonize quality control procedures for genomic data and phenotypic data



Aim 3

- To evaluate genetic overlap between stroke types (ischemic and hemorrhagic) and subtypes (TOAST categories of ischemic stroke and lobar/nonlobar categories for intracerebral hemorrhage)

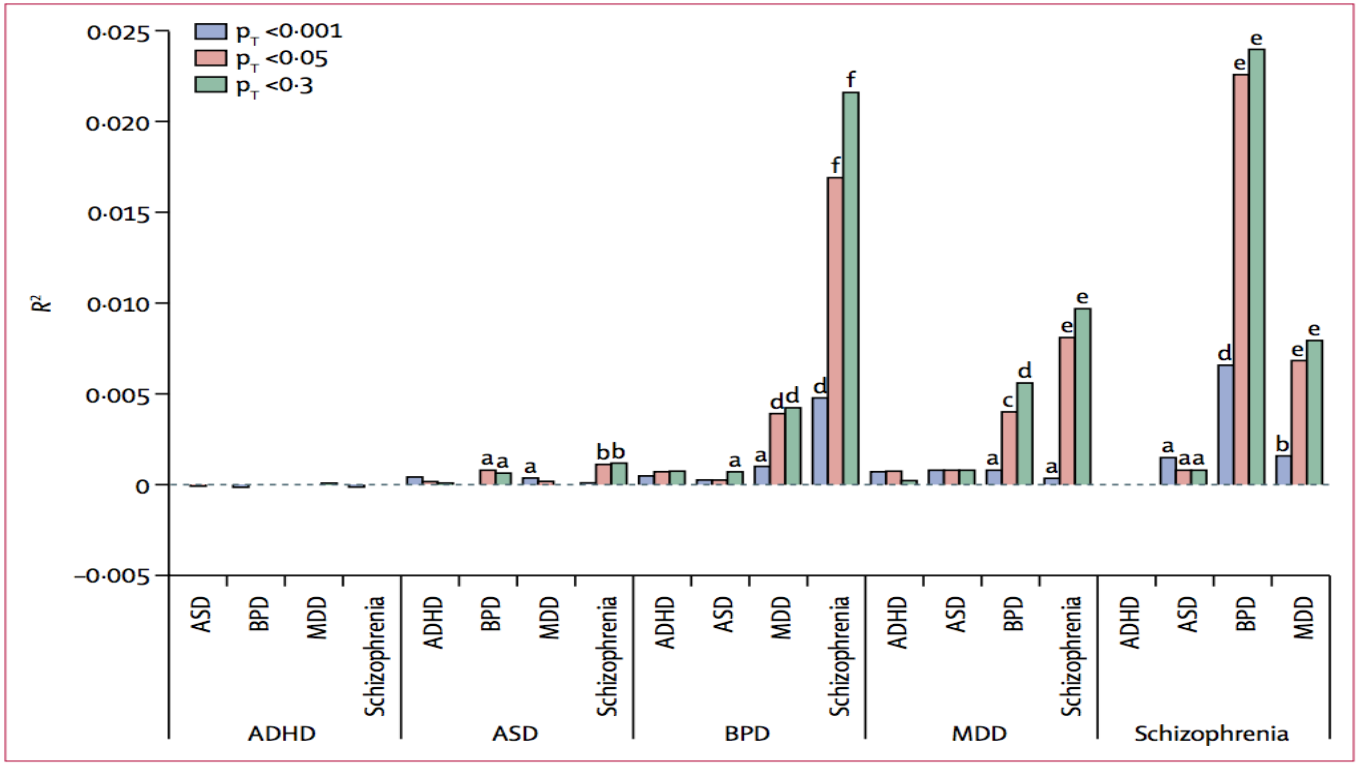


Figure 3: **Pair-wise cross-disorder polygene analysis**

Lancet. 2013 Apr 20;381(9875):1360.

Aim 4

- To increase and promote access to data, analytical tools and results by making the repository available to researchers from different disciplines interested in stroke genetics

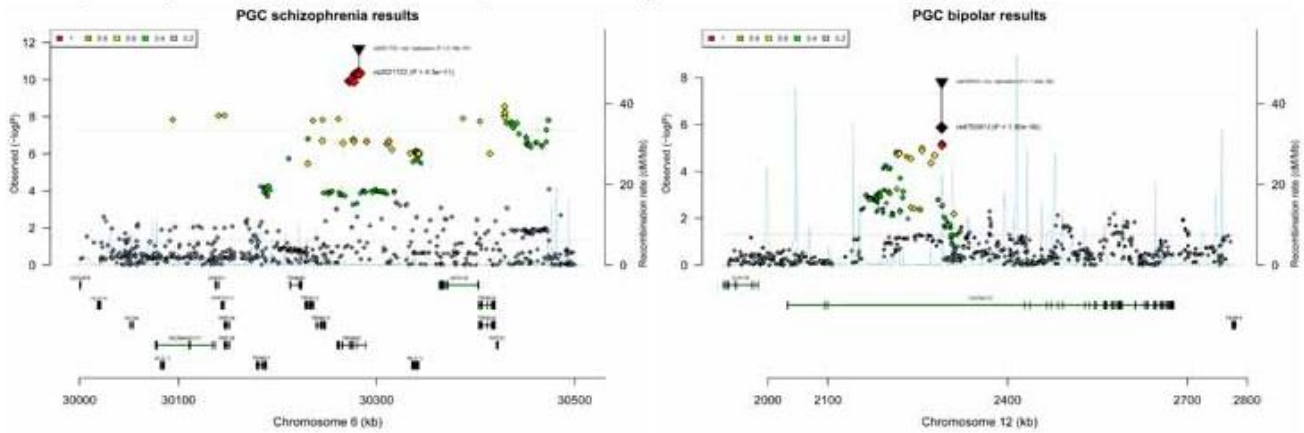
Psychiatric Genomics Consortium

Data Sharing

Data Visualization

Published PGC results can be viewed using "**ricopili**" , a web site that generates high-resolution images of PGC results. This web resource takes as input a gene name or genomic region, and produces a plot of PGC findings in genomic context. Thanks to Stephan Ripke and Brett Thomas of the Broad Institute.

Example output from ricopili, MHC region in schizophrenia and CACNA1C in bipolar disorder.



We are thinking along the same lines

- NINDS SiGN
- METASTROKE
- ICH GWAS
- METASTROKE
- ISGC Analysis Committee—synergy
- MRC proposal led by Steve Bevan
- Others