Combined Analysis of Psychiatric Studies (CAPS)

This document presents the methods for the 2012 CAPS analysis of bipolar data. The datasets used for analysis are available for download by authorized investigators in the Download Data section of www.nimhgenetics.org.

I. Data Acquisition

All available genetic data for the CAPS datasets were downloaded with permission from the NRGR Downloads section for Bipolar Disorder. Where available, raw basepair allele-coding was preferred; and study-provided marker information was collected. A series of map construction and genotype processing steps were conducted to thoroughly curate the genotypic data.

The current-at-the-time distribution file (BP 6.01) downloaded from HGI consists of useful pedigree, demographic, and clinical information. First, we standardized the DSM-IIIR and DSM-IV codes across studies, correcting obvious errors, such as typographical and case variations. Then, based on the expertise of our Clinical Advisory Board and a conservative philosophy, we developed a diagnostic algorithm to assign each individual to one of 5 categories: Unknown, Unaffected (with respect to bipolar spectrum), Recurrent Unipolar Depression (RUD), Bipolar II Disorder (BPII), and Bipolar I Disorder (BPI).

II. Data Curation

Detailed protocols for these steps are provided on the following pages:

- Map Construction
- Genotype Processing
- Phenotype Processing

III. Criteria for Inclusion of Families

Once the genotypic and phenotypic BP data were curated, we assessed the families based on the following inclusion criteria for analysis. Pedigrees were also distinguished by presence or absence of psychosis for analysis subsetting (in addition to cleaning group).

- 1 or more narrow BPI case
- 2 or more affected (BPI or BPII or RUD) cases with clean genotypic data
- not a genetic-trio (if 3 or less genotypes in pedigree)

Third-party software

The primary software package used by CAPS is KELVIN. We also use several third-party software tools during our genotype cleaning process. References to these tools are listed in relevant protocol.

CAPS Bipolar Datasets

Data used included multiplex bipolar (BP) family data with genome-wide scans available as of release BP 6.01. Of all BP datasets available as of April 2011, we selected those studies with family-based designs and genome-wide data.

The qualifying datasets are listed here and are available for download by authorized investigators in the Download Data section of www.nimhgenetics.org.

- **Dataset 4 (waves 1-4)[1-8]**

1. Dick, D.M., Foroud, T., Edenberg, H.J., Miller, M., Bowman, E., Rau, N.L., DePaulo, J.R., McInnis, M., Gershon, E., McMahon, F., et al. (2002). Apparent replication of suggestive linkage on chromosome 16 in the NIMH genetics initiative bipolar pedigrees. Am J Med Genet 114, 407-412.
2. Dick, D.M., Foroud, T., Flury, L., Bowman, E.S., Miller, M.J., Rau, N.L., Moe, P.R., Samavedy, N., El-Mallakh, R., Manji, H., et al. (2003). Genomewide linkage analyses of bipolar disorder: a new sample of 250 pedigrees from the National Institute of Mental Health Genetics Initiative. Am J Hum Genet 73, 107-114.
3. Foroud, T., Castelluccio, P.F., Koller, D.L., Edenberg, H.J., Miller, M., Bowman, E., Rau, N.L., Smiley, C., Rice, J.P., Goate, A., et al. (2000). Suggestive evidence of a locus on chromosome 10p using the NIMH genetics initiative bipolar affective disorder pedigrees. Am J Med Genet 96, 18-23.
4. Goes, F.S., Zandi, P.P., Miao, K., McMahon, F.J., Steele, J., Willour, V.L., Mackinnon, D.F., Mondimore, F.M., Schweizer, B., Nurnberger, J.I., Jr., et al. (2007). Mood-incongruent psychotic features in bipolar disorder: familial aggregation and suggestive linkage to 2p11-q14 and 13q21-33. Am J Psychiatry 164, 236-247.
5. Group, T.N.G.I.B. (1997). Genomic survey of bipolar illness in the NIMH genetics initiative pedigrees: a preliminary report. Am J Med Genet 74, 227-237.
6. McQueen, M.B., Devlin, B., Faraone, S.V., Nimgaonkar, V.L., Sklar, P., Smoller, J.W., Abou Jamra, R., Albus, M., Bacanu, S.A., Baron, M., et al. (2005). Combined analysis from eleven linkage studies of bipolar disorder provides strong evidence of susceptibility loci on chromosomes 6q and 8q. Am J Hum Genet 77, 582-595.
7. Willour, V.L., Zandi, P.P., Huo, Y., Diggs, T.L., Chellis, J.L., MacKinnon, D.F., Simpson, S.G., McMahon, F.J., Potash, J.B., Gershon, E.S., et al. (2003). Genome scan of the fifty-six bipolar pedigrees from the NIMH genetics initiative replication sample: chromosomes 4, 7, 9, 18, 19, 20, and 21. Am J Med Genet B Neuropsychiatr Genet 121B, 21-27.
8. Zandi, P.P., Willour, V.L., Huo, Y., Chellis, J., Potash, J.B., MacKinnon, D.F., Simpson, S.G., McMahon, F.J., Gershon, E., Reich, T., et al. (2003). Genome scan of a second wave of NIMH genetics initiative bipolar pedigrees: chromosomes 2, 11, 13, 14, and X. Am J Med Genet B Neuropsychiatr Genet 119B, 69-76.

CAPS Map Construction Protocol

1. Reference Map Acquisition

a. **KNOWN GENOTYPING ARRAY:** If we already have a map table (containing both physical and genetic positions) for the genotyping array, make sure it is still current with the Rutgers Maps. If current, use it directly; If not, get update from Rutgers. Skip step 1b – 1e

b. **DETERMINE BUILD:** Document which NCBI build Rutgers currently uses for physical locations

c. **PHYSICAL LOCATIONS:** Determine physical locations for all the markers in the dataset from the appropriate build in local database (or table) downloaded from UCSC. Available databases: hg18 NCBI36 and hg19 NCBI37. Available tables: snp130 dbSNP 130 build, snp131 dbSNP 131 build, stsAlias, stsMaps

d. **MARKER NOT FOUND:** If a marker is not in our database (searching both truename and alias variables), utilize following options (again careful to choose correct build)
   - OPTION 1: Recheck UCSC genome browser
   - OPTION 2: Search Map-o-Mat [archive link; site is nonfunctional]
   - OPTION 3: Search for UniSTS marker name (without hyphenated suffices) in NCBI records
   - OPTION 4: If not found, search for name (may indicate CIDR primer pair) in CIDRmarkers.xls
   - OPTION 5: If multiple disparate regions returned, must determine pcr primer set used by investigators (genotyping lab)
   - OPTION 6: To convert physical coordinates to an earlier assembly (such as hg36), use one of these sites: UCSC In-Silico PCR; UCSC GenBank BLAT

e. **CONVERT ALL PHYSICAL LOCATIONS TO GENETIC POSITIONS:** If genetic positions not already obtained (OPTION 7), use Rutgers tool. Use female_cM output for the X chromosome (unless pseudo-autosomal with male data). If NULL returned by interpolator (usually at chromosome tails), must extrapolate from nearby markers with returned values.

f. **ORDER CHECK:** Physical and genetic position orders are in agreement; no markers on the same chromosome with the same cM position or physical position.

2. Study Map Construction (execute separately for each cleaning group)

a. **SORT GENOTYPE DATA:** Order genotypes, mapfiles, and datafiles according to this reference map. Record any instances of order disagreement between the study data and the reference map

b. **θ_REF:** Convert the inter-marker distance from cM to θ_ref for each adjacent marker pairs in the study using their genetic positions in the reference map

c. **KELVIN M2M:** Run marker-to-marker option on all adjacent marker pairs to get (θ^, lod_max) output for each pair

d. **GENETIC DISTANCE:** Choose final genetic distance according to M2M output using using one of the options in 2e. Note: the 2-lod-unit support interval is the range of θ values such that lod(θ) > lod_max – 2

e. **CASES**
   - CASE I. LOW LOD_MAX (WITH LINKAGE): (θ^ < 0.5 lod_max < 2; or no_Inf); use θ_ref (from step 2b)
   - CASE II. COLLOCATED MARKER PAIRS: (θ^ = 0.0; lod_max ≥ 2) Rerun M2M (forcing br_out) to get LOD profile over θ θ' = upper bound of the 2-lod-unit support interval; use min(θ', θ_ref)
   - CASE III. UNLINKED MARKER PAIRS: (θ^ = 0.5 lod_max = 0) Rerun M2M (forcing br_out) to get LOD profile over θ θ' = lower bound of the 2-lod-unit support interval; use max(θ', θ_ref)
   - CASE IV. STANDARD ESTIMATED PAIRS: (WITH LINKAGE) θ^ > 0.0; < 0.5 lod_max ≥ 2; use θ^ (from step 2c)

f. **CONVERT TO KOSAMBI:** Convert resulting θ recombination fractions to kosambi genetic distances. Ensure no two markers have identical genetic positions; change 0 distance to 0.0001 if necessary

g. **MAP POSITIONS:** Sum inter-marker kosambi distances to construct marker map positions.

CAPS Genotype Processing

0. Pre-Processing

a. count families by study + site + ethnicity to decide cleaning groups and to inform eventual pooling for analysis subsets in (10a); if groups, decide whether to use pooled or group-specific allele frequencies and marker maps
b. find physical positions for all markers according to map construction and document current Rutgers NCBI build (0b,c,d can be done at any point prior to 7); decide whether to construct M2M map for analysis (7) or use either the provided or reference maps
c. adjust marker order in pedigree, map, and data files if study-provided order in disagreement with reference map; record out-of-order markers.
d. verify pedigree integrity (protocol software will fail without necessary dummy parents) and genotype/pedigree file agreement

1. Hardy-Weinberg

a. run PEDSTATS[1] to test for HWE; remove markers with p-value < cutoff (e.g., 0.0001)

2. Missingness

a. compute % missingness for individuals; zero-out individuals above cutoff (e.g., 20%)
b. compute % missingness for markers for remaining individuals; remove markers above cutoff (e.g., 10%)

3. Relatedness

a. use MENDEL[2] to estimate MK allele frequencies within cleaning group
b. run RELCHECK[3] to verify relatedness within family

4. Mendel Errors

a. use MENDEL[2] to determine first order Mendel errors; count errors by family & marker
b. remove markers above cutoff; zero-out families at markers with error

5. Verify Changes & Gender

a. repeat (4) Mendel Errors (MENDEL[2])
b. repeat (2) Missingness
c. repeat (1) Hardy-Weinberg (PEDSTATS[1])
d. review any cases of unexpected sex data, i.e., males with heterozygosity or females with all homozygous markers (taking into account presence of genotyped offspring)

6. Duplicates & Extended Pedigrees

a. run RELCHECK[3] to identify duplicates across families; i.e., look for MZ, par/offspring, or full sibs
b. reconstruct any extended pedigrees detected

7. Marker Positions

a. if constructing own map(s), run M2M in KELVIN to produce (θ^, lod_max) for each adjacent marker pair; otherwise, skip to (8)
b. handle 3 cases (lod_max<2, θ^=0.5, θ^=0) according to map construction to arrive at final Kosambi cM map positions

8. Unlikely Genotypes

   a. convert final linkage mapfile distances (7b or 0b) to HALDANE cM and sum to create converted genetic map
   b. run MERLIN[4] to detect higher order recombination events; record marker positions with errors

9. Final Pedigrees

   a. apply filter to require multiplex families based on phenotype for analysis, i.e., with at least 1 most-narrow case and at least 2 affected+genotyped members
   b. remove genotype trios, i.e., pedigrees with only 2 parents and their single offspring genotyped
   c. trim extraneous dummies (algorithm may be developed); produce pedigree drawings for families with 6 or more dummies
   d. count sizes, genotypes, and phenotypes of remaining families by study + site + ethnicity to decide subsetting and liability classes for analysis

10. Linkage Analysis

   a. pool data for analysis subsets and run likelihood-server-directed KELVIN, preserving the phenotypes, pedigree filters, marker maps, & allele frequencies of each cleaning group
   b. project subset-specific results onto a common 2cM genome map using the reference markers in (0b) and sequentially updating across subsets

1. Wigginton, J.E., and Abecasis, G.R. (2005). PEDSTATS: descriptive statistics, graphics and quality assessment for gene mapping data. Bioinformatics 21, 3445-3447.
2. Lange, K., et al. 2001, MENDEL version 4.0: a complete package for the exact genetic analysis of discrete traits in pedigree and population data sets. Am. J. Hum. Genet. 69Suppl, 504.
3. Broman, K.W., and Weber, J.L. (1998). Estimation of pairwise relationships in the presence of genotyping errors. Am J Hum Genet 63, 1563-1564.
4. Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30: 97– 101.

CAPS Phenotype Processing

1. Individuals with no clinical data were considered “unknown” phenotypically.
2. For assessed individuals, the NRGR provided diagnoses in the form of Diagnostic and Statistical Manual of Mental Disorders Third Ed. Revised (DSM-IIIR) and Fourth Ed. (DSM-IV) [Spitzer] codes. These codes represent lifetime diagnoses, although no temporal data were available. Therefore it was not possible to distinguish comorbid conditions from conditions that occurred over the course of illness or due to disease progression. In view of this, we opted to take a conservative approach to diagnostic classification.
3. Each individual pedigree member was ultimately assigned as affected, unaffected or unknown on two dimensions, bipolar disorder and psychosis. Based on various diagnostic codes (DSM-IV, DSMIIR, RDC and MRDC) available from the NRGR Bipolar Distribution 6.01, we first classified potential cases along the first dimension as Bipolar I (BPI), Bipolar II (BPII), or Recurrent Unipolar Depression (RUD), favoring the most severe qualifying category.
4. We applied exclusionary criteria involving disorders that complicate clinical presentation, including all diagnostic spectrums for dementia, as well as amnestic and cognitive disorders, and codes for unknown/unspecified or deferred diagnoses on Axis I. Additionally, substance related disorders that have been linked to BP or that cause ancillary psychiatric symptoms (delusions, delirium, hallucinations, depressed mood, or anxiety disorder) were excluded. Individuals with any schizophrenia spectrum diagnosis were excluded. Based on the advice of our clinical advisory board (CAB) and the conservative phenotype approach of this overall project, only BPI and BPII were considered affected cases for analysis, and individuals with RUD were excluded. Individuals with any exclusionary diagnosis were coded as phenotype “unknown.”
5. Only individuals with clear data indicating no psychiatric illness were called “unaffected”; all others were classified as “unknown”.
6. The second dimension of Psychosis was not used to distinguish individual cases, but rather to classify family types. To diagnose the second dimension of psychosis, we combined the NRGR codes with seven variables from the John Hopkins Phenome Database. Any family member having an NRGR codes indicating psychosis, including schizophrenia and schizoaffective disorder, or showing positive values for Phenome variables such as “ever psychotic” or “mania psychosis”, was labelled as affected for psychosis. Data availability determined the assignment of unaffected versus unknown, although this did not affect family characterization. Prior to analysis, the families were subset by the presence or absence of (known) psychosis, based on the hypothesis of locus heterogeneity in these pedigree types.

CAPS Clinical Advisory Board

The following individuals served as the Clinical Advisory Board for the initial CAPS projects:

Name	Affiliation
Anne Bassett, MD, FRCPC	Canada Research Chair in Schizophrenia Genetics and Genomics Disorders Professor of Psychiatry Associate Member, Canadian College of Medical Genetics Director, Clinical Genetics Research Program, Centre for Addiction & Mental Health University of Toronto
Prudence Fisher, PhD	Research Scientist, Division of Child Psychiatry Assistant Professor of Clinical Psychiatric Social Work New York State Psychiatric Institute & Columbia University
Deborah Levy, PhD	Director, Psychology Research Laboratory Associate Professor of Psychology Department of Psychiatry McLean Hospital, Harvard Medical School
Ellen Leibenluft, MD	Senior Investigator Chief, Section on Bipolar Spectrum Disorders, Emotion and Development Branch National Institute of Mental Health
Kathleen Merikangas, PhD	Senior Investigator Chief, Genetic Epidemiology Research Branch National Institute of Mental Health
Michel Maziade, MD, FRCPC, CQ	Scientific Director, Centre de recherche Canada Research Chair in Genetics of Neuropsychiatric Disorders Université Laval Robert-Giffard
Joseph Piven, MD	Sarah Graham Kenan Professor of Psychiatry, Pediatrics and Psychology Director, Carolina Institute for Developmental Disabilities University of North Carolina at Chapel Hill
Peter Szatmari, MD	Chedoke Child Health Chair in Child Psychiatry Professor and Head, Division of Child Psychiatry, Department of Psychiatry and Behavioural Neurosciences Director, Offort Centre for Child Studies McMaster University