Combined Analysis of Psychiatric Studies (CAPS)

This document presents the methods for the 2012 CAPS analysis of schizophrenia 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 Schizophrenia. Where available, raw basepair allele-coding was preferred; and study-provided marker information was collected. Some datasets, e.g., SZ-D12, did not include marker allele frequencies. When sample sizes permitted, as for SZ-D1 and SZ-D11, datasets were split into major ethnic groups for genotype processing. A series of map construction and genotype processing steps were conducted to thoroughly curate the genotypic data.

The current-at-the-time distribution file (SZ 8.0) downloaded from NRGR 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 [see file CAPS_Clinical_Advisory_Board.pdf] and a conservative philosophy, we developed a diagnostic algorithm to assign each individual to one of 5 categories: Unknown, Unaffected (with respect to schizophrenia spectrum), Broad Spectrum (BS), Schizoaffective Disorder (SA), and Schizophrenia (SZ). The BS cases were set to "Unknown" and did not count as affected individuals.

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

- 1 or more narrow SZ case
- 2 or more affected (SZ or SA) 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 Schizophrenia Datasets

Data used included multiplex schizophrenia (SZ) family data with genome-wide scans available as of release HGI SZ 8.0. Of all SZ 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 1**[1,2]
- **Dataset 11**[3]
- **Dataset 12**[4]
- **Dataset 21**[5]
- **Dataset 22**[6]
- **Dataset 23**[7]
- **Dataset 24**[8]

1. Faraone, S.V., Matise, T., Svrakic, D., Pepple, J., Malaspina, D., Suarez, B., Hampe, C., Zambuto, C.T., Schmitt, K., Meyer, J., et al. (1998). Genome scan of European-American schizophrenia pedigrees: results of the NIMH Genetics Initiative and Millennium Consortium. Am J Med Genet 81, 290-295.
2. Kaufmann, C.A., Suarez, B., Malaspina, D., Pepple, J., Svrakic, D., Markel, P.D., Meyer, J., Zambuto, C.T., Schmitt, K., Matise, T.C., et al. (1998). NIMH Genetics Initiative Millenium Schizophrenia Consortium: linkage analysis of African-American pedigrees. Am J Med Genet 81, 282-289.
3. Suarez, B.K., Duan, J., Sanders, A.R., Hinrichs, A.L., Jin, C.H., Hou, C., Buccola, N.G., Hale, N., Weilbaecher, A.N., Nertney, D.A., et al. (2006). Genomewide linkage scan of 409 European-ancestry and African American families with schizophrenia: suggestive evidence of linkage at 8p23.3-p21.2 and 11p13.1-q14.1 in the combined sample. Am J Hum Genet 78, 315-333.
4. Faraone, S.V., Hwu, H.G., Liu, C.M., Chen, W.J., Tsuang, M.M., Liu, S.K., Shieh, M.H., Hwang, T.J., Ou-Yang, W.C., Chen, C.Y., et al. (2006). Genome scan of Han Chinese schizophrenia families from Taiwan: confirmation of linkage to 10q22.3. Am J Psychiatry 163, 1760-1766.
5. Almasy, L., Gur, R.C., Haack, K., Cole, S.A., Calkins, M.E., Peralta, J.M., Hare, E., Prasad, K., Pogue-Geile, M.F., Nimgaonkar, V., et al. (2008). A genome screen for quantitative trait loci influencing schizophrenia and neurocognitive phenotypes. Am J Psychiatry 165, 1185-1192.
6. Escamilla, M.A., Ontiveros, A., Nicolini, H., Raventos, H., Mendoza, R., Medina, R., Munoz, R., Levinson, D., Peralta, J.M., Dassori, A., et al. (2007). A genome-wide scan for schizophrenia and psychosis susceptibility loci in families of Mexican and Central American ancestry. Am J Med Genet B Neuropsychiatry Genet 144B, 193-199.
7. Escamilla, M., Hare, E., Dassori, A.M., Peralta, J.M., Ontiveros, A., Nicolini, H., Raventos, H., Medina, R., Mendoza, R., Jerez, A., et al. (2009). A schizophrenia gene locus on chromosome 17q21 in a new set of families of Mexican and central american ancestry: evidence from the NIMH Genetics of schizophrenia in latino populations study. Am J Psychiatry 166, 442-449.
8. Wiener, H.W., Klei, L., Irvin, M.D., Perry, R.T., Aliyu, M.H., Allen, T.B., Bradford, L.D., Calkins, M.E., Devlin, B., Edwards, N., et al. (2009). Linkage analysis of schizophrenia in African-American families. Schizophr Res 109, 70-79.

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 [see file CAPS_T1.xlsx]

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 [see file CAPS_SP1.pdf]

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 [see file CAPS_T1.xlsx]
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) [see file CAPS_H1.xlsx]

2. Missingness

a. compute % missingness for individuals; zero-out individuals above cutoff (e.g., 20%) [see file CAPS_H3.xlsx]
b. compute % missingness for markers for remaining individuals; remove markers above cutoff (e.g., 10%) [see file CAPS_H2.xlsx]

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 [see file CAPS_SP1.pdf]

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 [see file CAPS_BC2.pdf] to require multiplex families based on phenotype for analysis [see file CAPS_BC1.pdf], 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. We applied exclusionary criteria [GLOBAL_EXCLUDE] 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 SZ or that cause ancillary psychiatric symptoms (delusions, delirium, hallucinations, depressed mood, anxiety disorder) were excluded. Individuals with any exclusionary diagnosis were coded as phenotype “unknown.”
4. Remaining SZ disorders were divided into two levels: (i) narrow SZ (Schizophrenia Disorder, including Disorganized, Catatonic, Paranoid, and Residual types) [SZ_CODES]; (ii) Schizoaffective (SA) (Schizoaffective Disorder or any SZ Disorder with a significant affective component) [SA_CODES]. We also classified as SZ Spectrum anyone meeting criteria for a Delusional Disorder, Brief Psychotic Disorder, Psychotic Disorder NOS, Schizophreniform Disorder, or Cluster A Personality Disorder [BS_CODES]. Only 195 (7% of affected individuals) were classified as meeting criteria for SZ Spectrum, and these subjects were recoded to “unknown.”
5. Based on these classifications, problematic comorbid diagnoses of Recurrent Major Depressive and Bipolar Disorder [SZ_DEMOTE1] were examined, in order to reliably account for affective presentation while taking into consideration diagnostic uncertainty due to lack of temporal data. SZ individuals with these comorbid diagnoses were recoded as SA. All individuals with Major Depressive or Bipolar Disorder coded as either “severe, with psychotic features” or as “severity unknown” [SZ_EXCLUDE; SA_EXCLUDE; BS_EXCLUDE] were recoded as phenotypically “unknown.”
6. Individuals were classified at the highest known level even if they were unknown at higher level(s) of the SZ, SA, BS hierarchy. Individuals meeting neither SZ, SA, SZ Spectrum, nor exclusionary diagnoses (as above) were tentatively coded as “unaffected;” “unaffected” individuals with Bipolar I or II Disorder NOS, Mood Disorder NOS, Depressive Disorder NOS, Personality Disorder NOS, or Diagnosis Deferred on Axis II [UNAFF_EXCLUDE] were recoded as “unknown.”

SZ	SA	BS	Classification	DX4
2	[0,1]	[0,1,2]	Schizophrenia	4
[0,1]	2	[0,1,2]	Schizoaffective	3
[0,1]	[0,1]	2	Broad Spectrum	2
1	1	1	Unaffected	1
0	0	0	Unknown	0

7. Analyses were restricted to multiplex families with at least one case of SZ and one additional case of either SZ or schizoaffective disorder (SA), with at least two affected genotyped individuals. Families were also characterized by the presence or absence of any SA individuals. This distinction was not made in any of the original studies.