Genotype Data QA/QC

- GAIN Genotyping Group
- HapMap samples initially
- QA samples for each study
- QC for genotyping
- NCBI QA check
- Genotype data quality standards
GAIN Genotyping Group

Gonçalo Abecasis (Chair)  
Dennis Ballinger  
John Thompson  
Stacey Gabriel, Mark Daly  
Steve Lincoln  
Elizabeth Pugh  
Peter Donnelly  
Stephen Sherry, Michael Feolo  
James Battey  
Lisa Brooks, Teri Manolio, Emily Harris  
David Wholley

Michigan  
Perlegen  
Pfizer  
Broad  
Affymetrix  
CIDR  
WTCCC  
NCBI  
NIDCD  
NHGRI  
FNIH
HapMap Samples Initially

Both centers are genotyping all 270 HapMap samples on the GAIN platforms and SNPs, to show:

• The SNPs that work.
• Genomic coverage of the SNPs.
• Completeness and concordance with HapMap genotypes.
QA Samples for Each Study

• Study trio samples (Faraone ADHD)
• QA trio samples related to study samples (some studies)
• HapMap CEPH sample(s) (all studies)
• HapMap Yoruba samples (AA studies)
• Study duplicates (all studies)
QC for Genotyping

- More QA samples for studies with unrelated samples, multiple collection sites or DNA extraction methods, more ethnic diversity.
- Cases and controls on same plates and done at same time; plates differ in sample layouts (sexes, duplicate samples).
- QC process for each genotyping center.
NCBI QA Check

• Gonçalo Abecasis is developing a software pipeline to assess genotype data quality.

• NCBI will apply it to each GAIN study.

• Any issues will be resolved between the genotyping centers, study PIs, and NCBI.
Genotype Data Quality

- Number of SNPs, genomic coverage.
- Completeness, and in HapMap QA samples by hets and homs.
- Concordance with HapMap samples and between duplicates.
- Concordance in family samples.
Data Quality Standards

Remove samples with < 80% of SNPs called.
Of > 480k for Perlegen and 500k for Broad, > 90% of SNPs will be good:
• any SNPs out of HW will not count as good,
• call rate minimum = 90% and average ≥ 97%,
• for HapMap QA samples the average call rates for hets and homs both ≥ 97%,
• concordance in duplicates of ≥ 99.5%.