NCBI Data Analyses (aka Pre-computes)

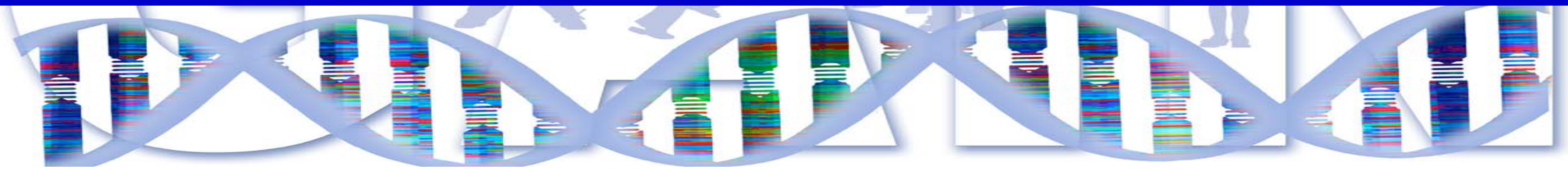
Emily L. Harris, PhD, MPH

NHGRI

Genetic Association Information Network

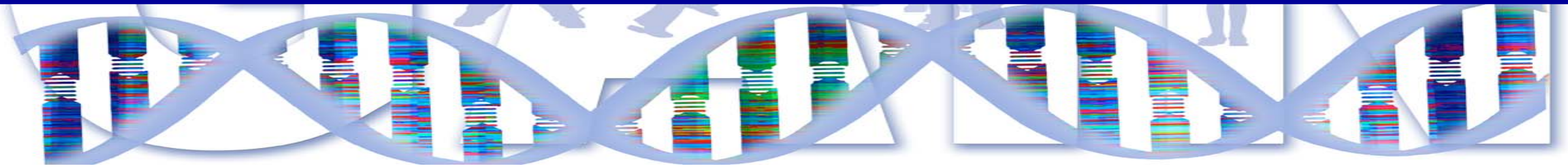
Purpose

- Provide basic results
  - Preliminary scientific results for investigators
  - Quality check for shared data
- Promote broad data use
  - Protect against claims on genotype/ allele frequency data and phenotype-genotype associations



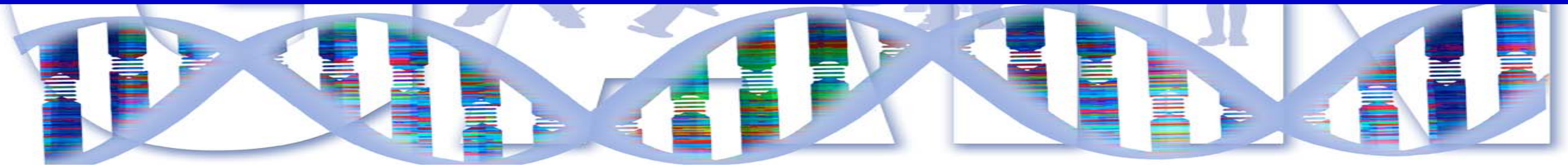
Analysis Ideas

- Each study
  - For each SNP
  - For each target trait/condition
  - For multi-SNP analyses
- Cross-study analyses



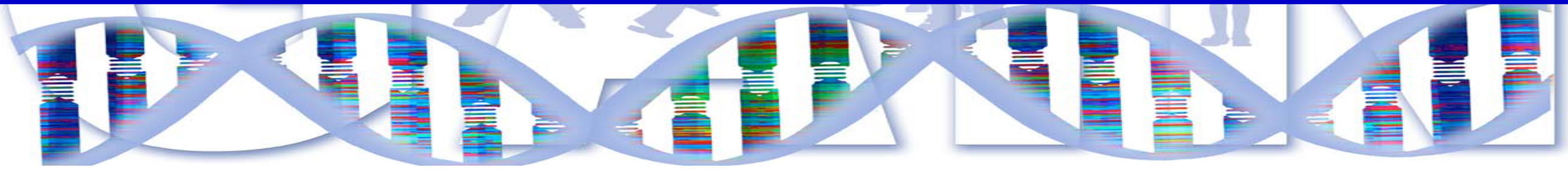
Analysis Ideas for each SNP

- Genotype call accuracy (clusters)
- Allele frequencies
- Genotype frequencies
- Tests of Hardy-Weinberg equilibrium
- Stratified on ascertainment status
  - Case-control design: Cases, controls
  - Family designs: Probands, unrelated individuals

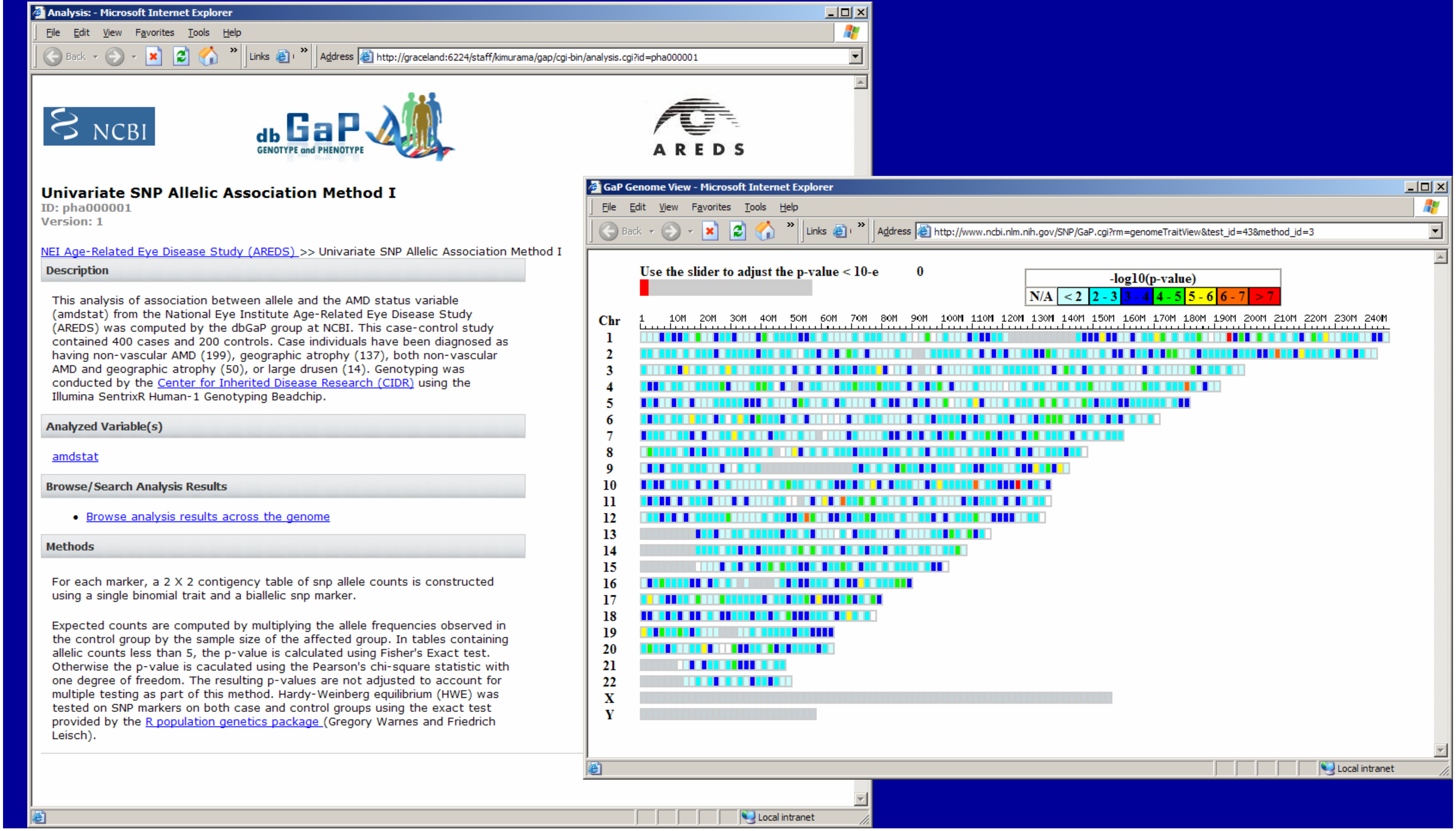


Analysis Ideas
for each Target Trait/Condition

- Association with each SNP genotype
    - Unadjusted measure of association (e.g., unadjusted odds ratio)
    - Statistical significance, with correction for multiple tests
        - By SNP
        - Genome-wide visual representation



Visual Representation: Genome-wide



Analysis: - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Links Address
NCBI
db GaP GENOTYPE and PHENOTYPE
AREDS
Local intranet

http://graceland:6224/staff/kimurama/gap/cgi-bin/analysis.cgi?id=pha000001

Univariate SNP Allelic Association Method I

ID: pha000001
Version: 1

NEI Age-Related Eye Disease Study (AREDS) >> Univariate SNP Allelic Association Method I

Description

This analysis of association between allele and the AMD status variable (amdstat) from the National Eye Institute Age-Related Eye Disease Study (AREDS) was computed by the dbGaP group at NCBI. This case-control study contained 400 cases and 200 controls. Case individuals have been diagnosed as having non-vascular AMD (199), geographic atrophy (137), both non-vascular AMD and geographic atrophy (50), or large drusen (14). Genotyping was conducted by the Center for Inherited Disease Research (CIDR) using the Illumina SentrixR Human-1 Genotyping Beadchip.

Analyzed Variable(s)

amdstat

Browse/Search Analysis Results

- Browse analysis results across the genome

Methods

For each marker, a 2 X 2 contigency table of snp allele counts is constructed using a single binomial trait and a biallelic snp marker.

Expected counts are computed by multiplying the allele frequencies observed in the control group by the sample size of the affected group. In tables containing allelic counts less than 5, the p-value is calculated using Fisher's Exact test. Otherwise the p-value is caculated using the Pearson's chi-square statistic with one degree of freedom. The resulting p-values are not adjusted to account for multiple testing as part of this method. Hardy-Weinberg equilibrium (HWE) was tested on SNP markers on both case and control groups using the exact test provided by the R population genetics package (Gregory Warnes and Friedrich Leisch).

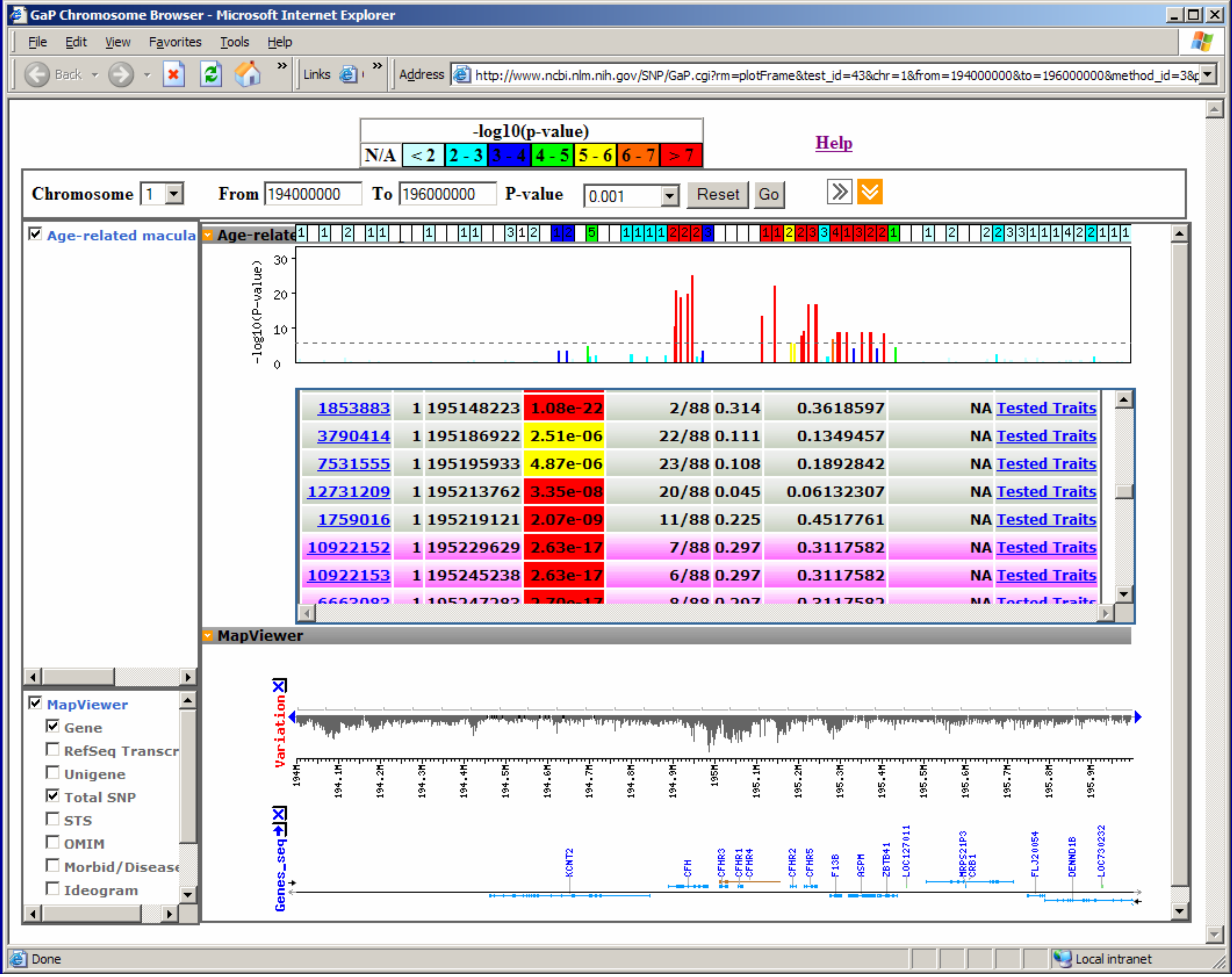
GaP Genome View - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Links Address
Local intranet

http://www.ncbi.nlm.nih.gov/SNP/GaP.cgi?rm=genomeTraitView&test_id=43&method_id=3

Use the slider to adjust the p-value < 10-e 0

-log10(p-value)							
N/A	< 2	2 - 3	3 - 4	4 - 5	5 - 6	6 - 7	> 7

Visual Representation: A More Detailed Look



Address: http://www.ncbi.nlm.nih.gov/SNP/GaP.cgi?rm=plotFrame&test_id=43&chr=1&from=194000000&to=196000000&method_id=3&p

Chromosome 1 From 194000000 To 196000000 P-value 0.001

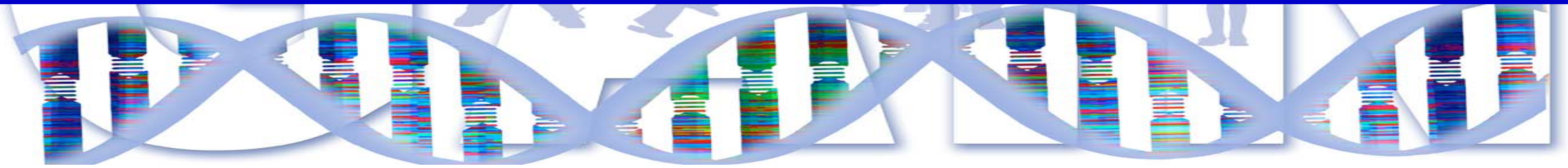
☑ Age-related macula

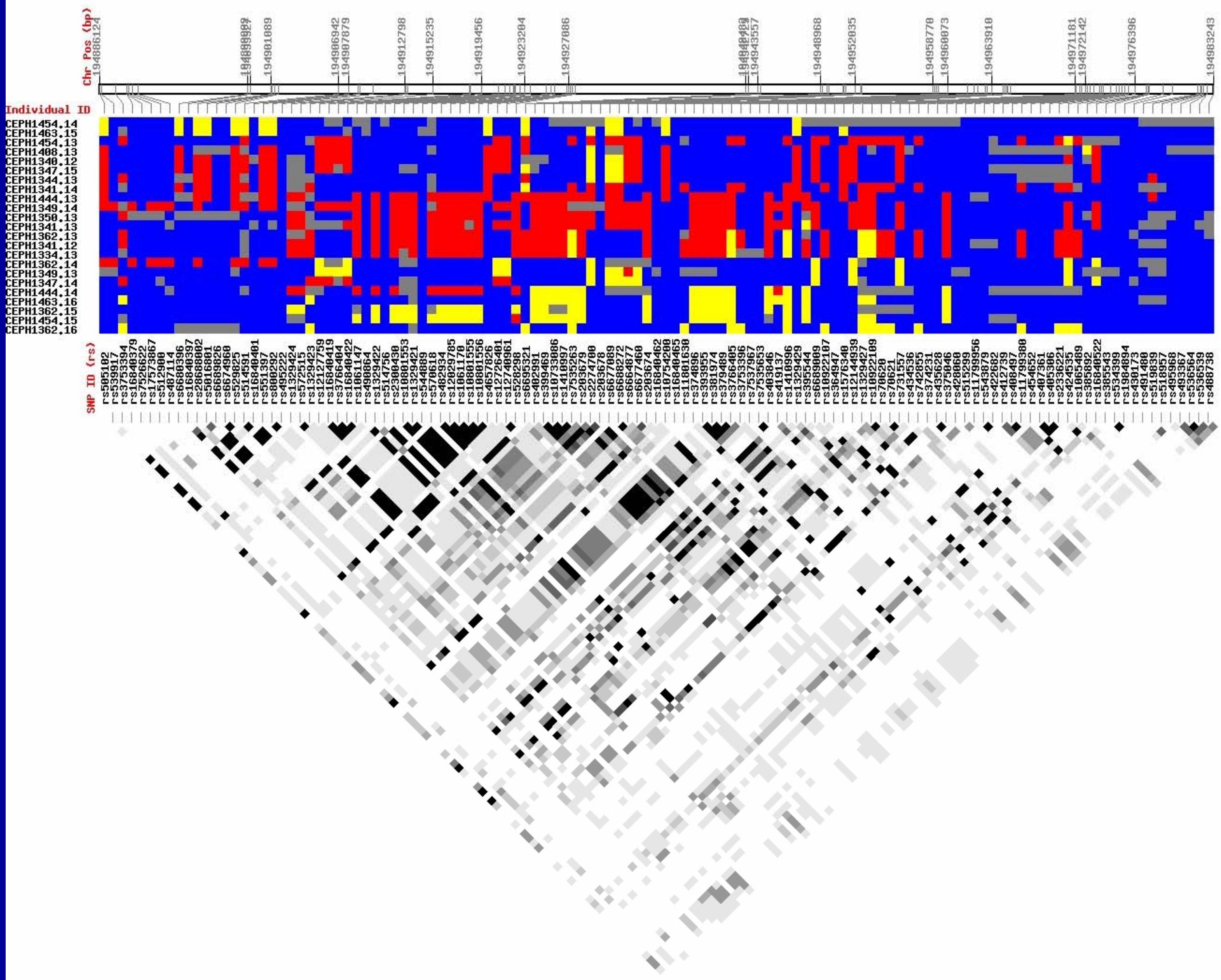
☑ MapViewer
☑ Gene
☐ RefSeq Transcr
☐ Unigene
☑ Total SNP
☐ STS
☐ OMIM
☐ Morbid/Disease
☐ Ideogram

1853883	1	195148223	1.08e-22	2/88	0.314	0.3618597	NA	Tested Traits
3790414	1	195186922	2.51e-06	22/88	0.111	0.1349457	NA	Tested Traits
7531555	1	195195933	4.87e-06	23/88	0.108	0.1892842	NA	Tested Traits
12731209	1	195213762	3.35e-08	20/88	0.045	0.06132307	NA	Tested Traits
1759016	1	195219121	2.07e-09	11/88	0.225	0.4517761	NA	Tested Traits
10922152	1	195229629	2.63e-17	7/88	0.297	0.3117582	NA	Tested Traits
10922153	1	195245238	2.63e-17	6/88	0.297	0.3117582	NA	Tested Traits
6663083	1	195247283	2.70e-17	8/88	0.297	0.3117582	NA	Tested Traits

Analysis Ideas for Multi-SNP Analyses

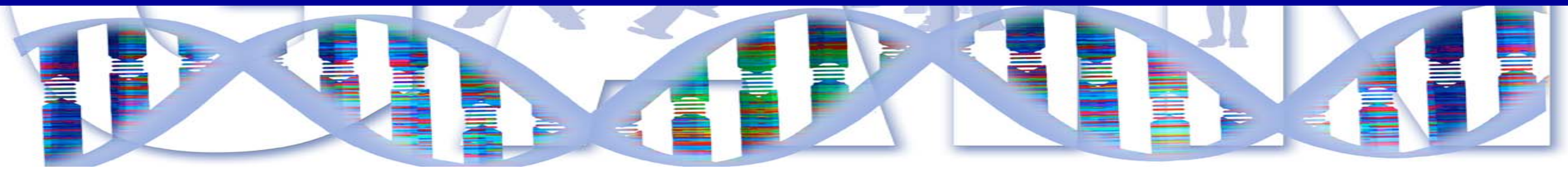
- Pairwise linkage disequilibrium
  - By pairs of SNPs
  - Visual representation
- Stratified on ascertainment status
  - Case-control design: Cases, controls
  - Family designs: Probands, unrelated individuals




Chr Pos (bp)
194886124
194901089
194906942
194907879
194912798
194915235
194919456
194923204
194927086
194943557
194948968
194952035
194958770
194960073
194963910
194971181
194972142
194976396
194983243
Individual ID
CEPH1454.14
CEPH1463.15
CEPH1454.13
CEPH1408.13
CEPH1340.12
CEPH1347.15
CEPH1344.13
CEPH1341.14
CEPH1444.13
CEPH1349.14
CEPH1350.13
CEPH1341.13
CEPH1362.13
CEPH1341.12
CEPH1334.13
CEPH1362.14
CEPH1349.13
CEPH1347.14
CEPH1444.14
CEPH1463.16
CEPH1362.15
CEPH1454.15
CEPH1362.16
SNP ID (rs)
rs505102
rs539917
rs3753394
rs16840379
rs7526622
rs17573867
rs512900
rs487114
rs6680396
rs16840397
rs28688002
rs5016801
rs6689826
rs6674960
rs529825
rs514591
rs16840401
rs551397
rs800292
rs495222
rs1329424
rs572515
rs1329423
rs12127759
rs16840419
rs3766404
rs16840422
rs1061147
rs490864
rs1329422
rs514756
rs2300430
rs10801553
rs1329421
rs544889
rs570618
rs482934
rs12029785
rs1061170
rs10801555
rs10801556
rs4657826
rs12726401
rs12740961
rs528298
rs6695321
rs402991
rs399469
rs10733086
rs1410997
rs7535263
rs203679
rs2274700
rs203678
rs6677089
rs6688272
rs6664877
rs6677460
rs203674
rs16840462
rs10754200
rs16840465
rs11801630
rs374896
rs393955
rs381974
rs379489
rs3766405
rs3753396
rs7537967
rs7535653
rs403846
rs419137
rs1410996
rs1329429
rs395544
rs6689009
rs10922107
rs364947
rs1576340
rs12144939
rs1329427
rs10922109
rs70620
rs70621
rs731557
rs434536
rs742855
rs374231
rs435628
rs375046
rs428060
rs515299
rs11799956
rs543879
rs422682
rs412739
rs408497
rs11799380
rs454652
rs407361
rs405306
rs2336221
rs424535
rs1065489
rs385892
rs16840522
rs385543
rs534399
rs1984894
rs430173
rs491480
rs519839
rs518957
rs495968
rs493367
rs536564
rs536539
rs488738

Analysis Ideas
for Multi-SNP Analyses

- Haplotyping for chromosomal regions that surround SNPs with significant associations



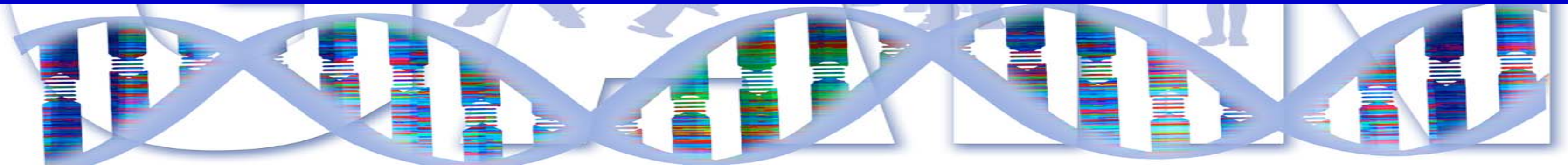
Analysis Ideas

What else would be useful?

⇨ Cross-study analyses?

⇨

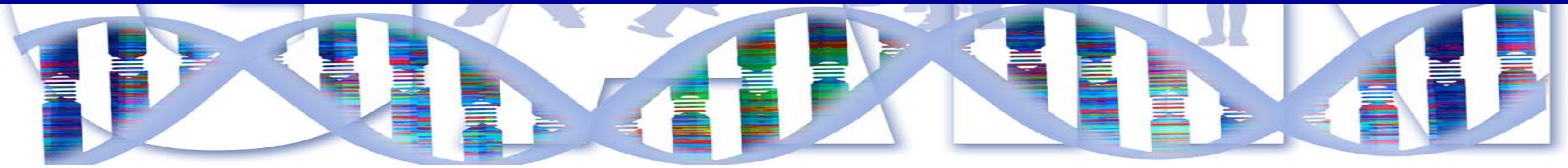
⇨



Analysis Ideas

Questions: Analysis

- ⇨ How many/which genetic models should we test?
- ⇨ How important to consider multiple testing for p-values in the pre-computed analyses, given that it is a public, unpublished resource?
- ⇨ What multiple testing corrections should be included for the pre-computed analyses?

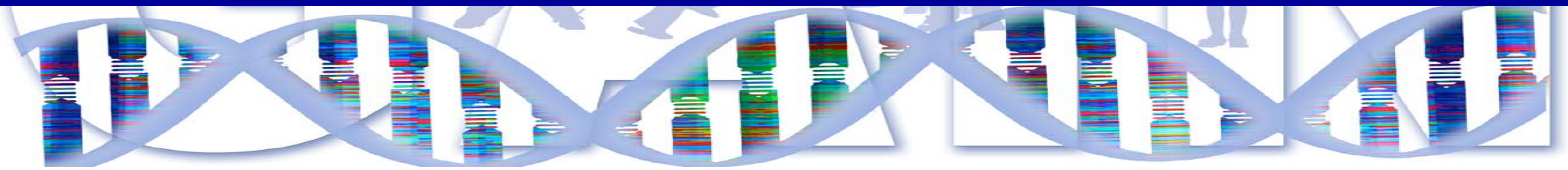


Analysis Ideas

Questions: Links to genomic information

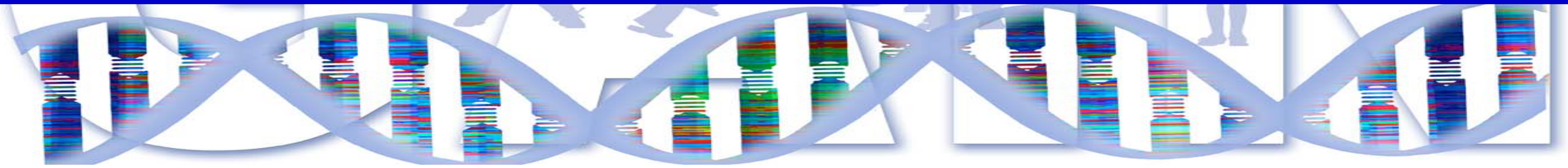
For SNPs with a significant association

- ⇨ Should we include in the report information about location of each SNP within a known gene?
- ⇨ Should we annotate the genome with these significant associations?



Availability

- Results initially available only through the controlled access process for approved users to download
- After the 9-month protected period for a specific project, results available on the GAIN public web site



Thanks!

NCBI

Al Graeff

Matt Mailman

Jim Ostell

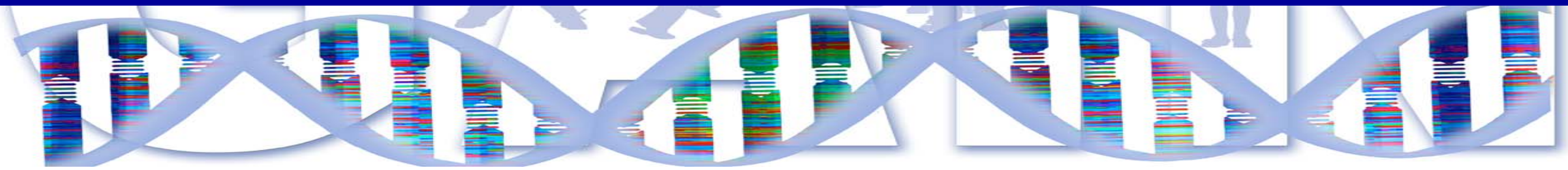
Steve Sherry

NHGRI

Lisa Brooks

Teri Manolio

Laura Lyman Rodriguez



GWA data availability

- dbGaP – NCBI
- CGEMS – NCI
- NINDS Open Access Repository

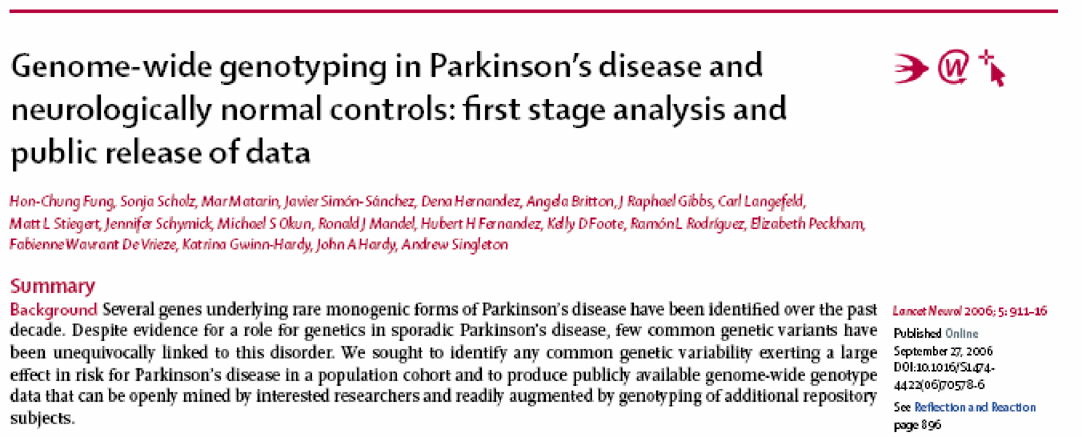
Genome-wide SNP genotyping

- Initial genome wide genotyping in:
    - 276 PD
    - 276 Stroke
    - 276 ALS
    - 276 Controls

    FROM NINDS OPEN ACCESS REPOSITORY

    - 200 African Americans

    M.E. & A.Z. NIA
- 109k exon-centric assay (phase I)
- 317k HapMap assay (phase II)
- >99.8% call rate, >450,000,000 unique genotypes, >99.9% reproducibility (over 19,000,000 replicate genotypes)



Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data

Hon-Chung Fung, Sonja Scholz, Mar Matarin, Javier Simón-Sánchez, Dena Hernandez, Angela Britton, J Raphael Gibbs, Carl Langefeld, Matt L Stiegert, Jennifer Schymick, Michael S Okun, Ronald J Mandel, Hubert H Fernandez, Kelly D Foote, Ramón L Rodríguez, Elizabeth Peckham, Fabienne Wavrant De Vrieze, Katrina Gwinn-Hardy, John A Hardy, Andrew Singleton

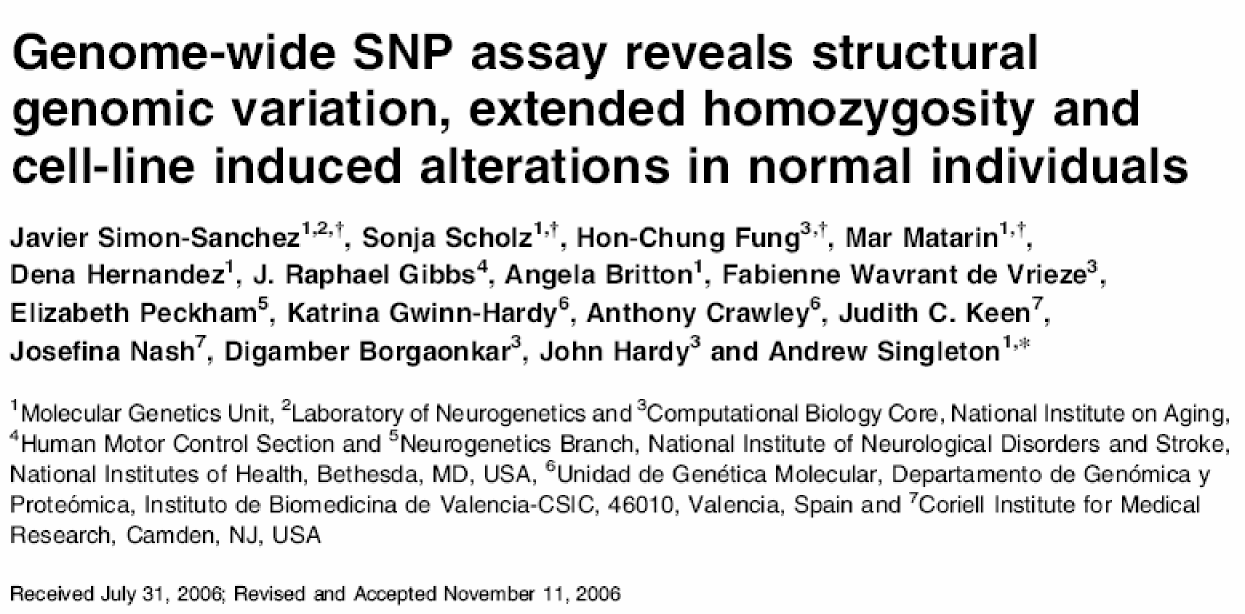
Summary

Background Several genes underlying rare monogenic forms of Parkinson's disease have been identified over the past decade. Despite evidence for a role for genetics in sporadic Parkinson's disease, few common genetic variants have been unequivocally linked to this disorder. We sought to identify any common genetic variability exerting a large effect in risk for Parkinson's disease in a population cohort and to produce publicly available genome-wide genotype data that can be openly mined by interested researchers and readily augmented by genotyping of additional repository subjects.

Lancet Neurol 2006; 5: 911–16

Published Online
September 27, 2006
DOI:10.1016/S1474-4422(06)70578-6

See Reflection and Reaction page 896



Human Molecular Genetics, 2006, Vol. 16, No. 1 *1–14*
doi:10.1093/hmg/ddl436
Advance Access published on xxx

Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals

Javier Simon-Sanchez[1,2,†], Sonja Scholz[1,†], Hon-Chung Fung[3,†], Mar Matarin[1,†], Dena Hernandez[1], J. Raphael Gibbs[4], Angela Britton[1], Fabienne Wavrant de Vrieze[3], Elizabeth Peckham[5], Katrina Gwinn-Hardy[6], Anthony Crawley[6], Judith C. Keen[7], Josefina Nash[7], Digamber Borgaonkar[3], John Hardy[3] and Andrew Singleton[1,*]

[1]Molecular Genetics Unit, [2]Laboratory of Neurogenetics and [3]Computational Biology Core, National Institute on Aging, [4]Human Motor Control Section and [5]Neurogenetics Branch, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA, [6]Unidad de Genética Molecular, Departamento de Genómica y Proteómica, Instituto de Biomedicina de Valencia-CSIC, 46010, Valencia, Spain and [7]Coriell Institute for Medical Research, Camden, NJ, USA

Received July 31, 2006; Revised and Accepted November 11, 2006

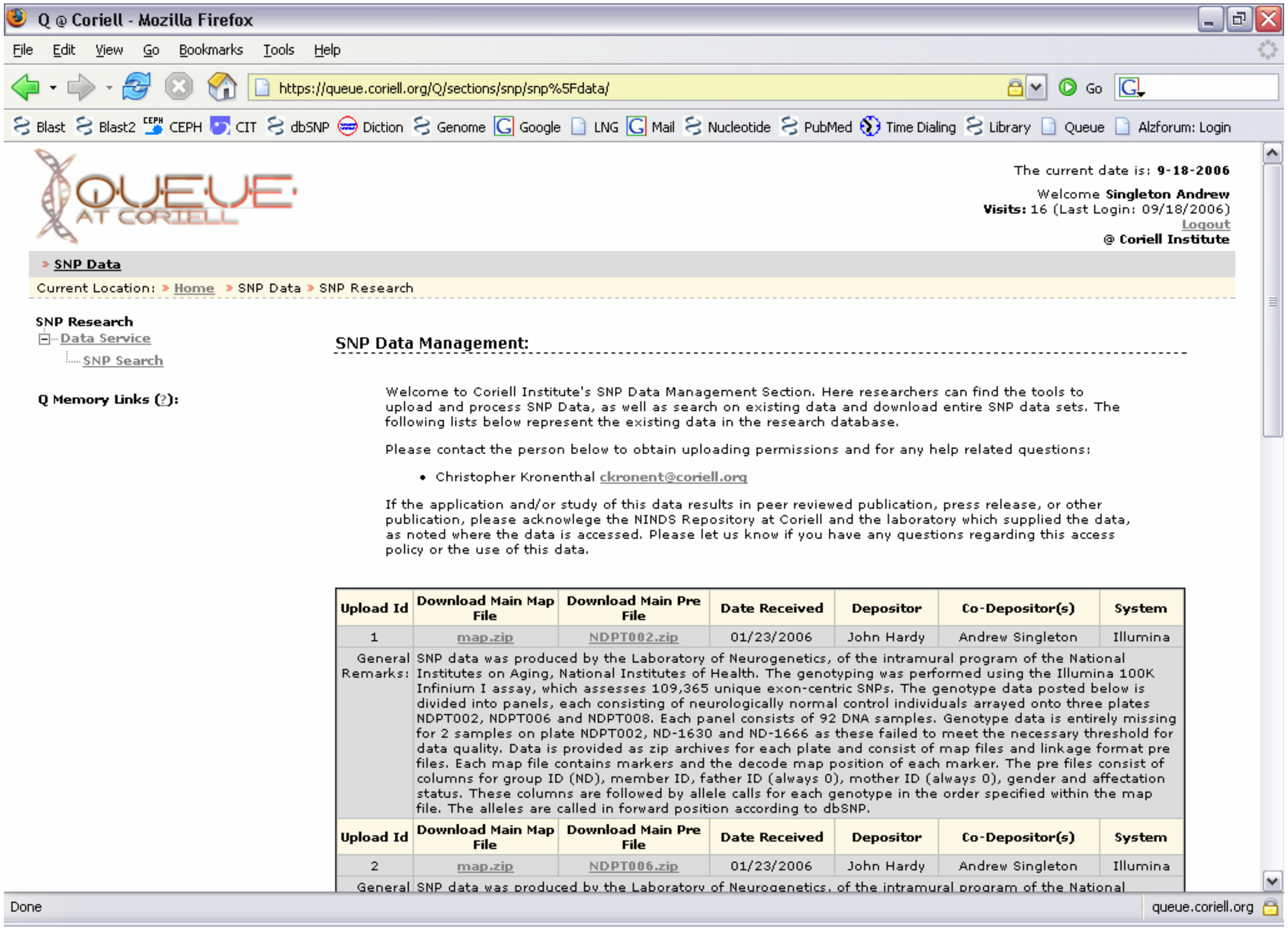
Genome-wide SNP genotyping

PD and control genotype data posted publicly

downloaded by >300 unique visitors.

Stroke and ALS data available Jan 2007.

Allows entire data download or search for specific SNPs


Q @ Coriell - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
https://queue.coriell.org/Q/sections/snp/snp%5Fdata/
Go
Blast Blast2 CEPH CIT dbSNP Diction Genome Google LNG Mail Nucleotide PubMed Time Dialing Library Queue Alzforum: Login
QUEUE AT CORIELL
Done
queue.coriell.org

The current date is: **9-18-2006**

Welcome **Singleton Andrew**
Visits: 16 (Last Login: 09/18/2006)
Logout
@ Coriell Institute

» **SNP Data**

Current Location: » Home » SNP Data » SNP Research

SNP Research
- Data Service
  - SNP Search

Q Memory Links (?):

SNP Data Management:

Welcome to Coriell Institute's SNP Data Management Section. Here researchers can find the tools to upload and process SNP Data, as well as search on existing data and download entire SNP data sets. The following lists below represent the existing data in the research database.

Please contact the person below to obtain uploading permissions and for any help related questions:

- Christopher Kronenthal ckronent@coriell.org

If the application and/or study of this data results in peer reviewed publication, press release, or other publication, please acknowlege the NINDS Repository at Coriell and the laboratory which supplied the data, as noted where the data is accessed. Please let us know if you have any questions regarding this access policy or the use of this data.

Upload Id	Download Main Map File	Download Main Pre File	Date Received	Depositor	Co-Depositor(s)	System
1	map.zip	NDPT002.zip	01/23/2006	John Hardy	Andrew Singleton	Illumina
General Remarks:	SNP data was produced by the Laboratory of Neurogenetics, of the intramural program of the National Institutes on Aging, National Institutes of Health. The genotyping was performed using the Illumina 100K Infinium I assay, which assesses 109,365 unique exon-centric SNPs. The genotype data posted below is divided into panels, each consisting of neurologically normal control individuals arrayed onto three plates NDPT002, NDPT006 and NDPT008. Each panel consists of 92 DNA samples. Genotype data is entirely missing for 2 samples on plate NDPT002, ND-1630 and ND-1666 as these failed to meet the necessary threshold for data quality. Data is provided as zip archives for each plate and consist of map files and linkage format pre files. Each map file contains markers and the decode map position of each marker. The pre files consist of columns for group ID (ND), member ID, father ID (always 0), mother ID (always 0), gender and affectation status. These columns are followed by allele calls for each genotype in the order specified within the map file. The alleles are called in forward position according to dbSNP.					
Upload Id	**Download Main Map File**	**Download Main Pre File**	**Date Received**	**Depositor**	**Co-Depositor(s)**	**System**
2	map.zip	NDPT006.zip	01/23/2006	John Hardy	Andrew Singleton	Illumina
General	SNP data was produced by the Laboratory of Neurogenetics, of the intramural program of the National					

http://ccr.coriell.org/ninds/