# PLINK gPLINK Haploview

Whole genome association software tutorial

### Shaun Purcell

Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA Broad Institute of Harvard & MIT, Cambridge, MA

http://pngu.mgh.harvard.edu/purcell/plink/

http://www.broad.mit.edu/mpg/haploview/



# GUI for many **PLINK** analyses

#### Data management

#### t PLINK Rescan folder ati Data Management Summary Statistics Stratification IBD Estimation Command Line

### Summary statistics



### **Population stratification**

	PLINK	Rescan fold	ler	
ati	Data	a Management	۲	
	Summary Statistics		۰.	
	Stratification		▶	IBS distances
	Asso	Association		Clustering
	IBD Estimation		Þ	Nearest Neighbour
	Command Line			

#### Association analysis

t.

ati



### **IBD-based analysis**



### **Computational efficiency**

350 individuals genotyped on 100,000 SNPs

Load, filter and analyze	~12 seconds
1 permutation (all SNPs)	~1.6 seconds

### gPLINK / PLINK in "remote mode"





# In this practical, we will use **gPLINK**, **PLINK** and **Haploview** to...

... examine genotyping rates and look for non-random missing data ... determine SNP frequencies and test Hardy-Weinberg equilibrium ... assess population stratification via clustering, genomic control ... test for allelic, genotypic and haplotypic association ... perform stratified analyses, conditioning on population strata ... assess between-stratum heterogeneity in association signal ... examine linkage disequilibrium patterns around associated SNPs ... select tag SNPs for follow-up and replication studies

# Simulated WGAS dataset

- Real genotypes, but a simulated "disease"
- 90 Asian HapMap individuals
   10K autosomal SNPs from Affymetrix 500K product
- Simulated quantitative phenotype; median split to create a disease phenotype
- Illustrative, not realistic!

# Specific questions asked

- 1) What is the genotyping rate?
- 2) How many monomorphic SNPs?
- 3) Evidence of non-random genotyping failure?
- 4) What is the single **most associated SNP**? Does it reach genome-wide significance? What is the **most associated haplotype**?
- 5) Is there evidence of **population stratification from genomic control**?
- 6) Use genotypes to **cluster the sample** into 2 subpopulations. How well does the clustering recover the known Chinese/Japanese split?
- 7) Is there evidence for stratification conditional on the two-cluster solution?
- 8) What is the **best SNP controlling for stratification**. Is it genome-wide significant?

For the most highly associated SNP:

- 9) Does this SNP pass the Hardy-Weinberg equilibrium test?
- 10) Does this SNP **differ in frequency** between the two populations?
- 11) Is there evidence that this SNP has a **different association** between the two populations?
- 12) What are the **allele frequencies** in cases and controls? **Genotype** frequencies? What is the **odds ratio**?
- 13) Is the rate of **missing data** equal between cases and controls for this SNP?
- 14) Does an additive model well characterize the association? What about genotypic, dominant models, etc?

### Data used in this practical

• Available at http://pngu.mgh.harvard.edu/purcell/affy/purcell.zip

example.bed	Binary format genotype information (do not attempt to view in a standard text editor)
example.bim	Map file (6 fields: each row is a SNP: chromosome, RS #, genetic position, physical position, allele 1, allele 2)
example.fam	Individual information file (first 6 columns of a PED file; disease phenotype is column 6)
pop.phe	Chinese/Japanese population indicator (FID, IID, population code)
qt.phe	Alternate quantitative trait phenotype file (Family ID, Individual ID, phenotype)

### The Truth...

	Chinese	Japanese	
Case	34	7	
Control	11	38	

	"11"	"12"	"22"
Case	5	21	23
Control	16	23	2

Group difference

Single common variant rs7835221 chr8

# Agplink "project" is a folder

Arrange Icons By Refresh	•		
Paste Paste Shortcut <b>Undo Delete</b>	Ctrl+Z		
Graphics Properties Graphics Options	٠,		
New	•		Eolder
Properties		ē	Shortcut
			Briefcase Bitmap Image Microsoft Word Document DVI File Microsoft Office Access Application Microsoft PowerPoint Presentation Adobe Photoshop Image Microsoft Office Publisher Document Text Document Wave Sound Microsoft Excel Worksheet WinZip File



Right-click on the Desktop to create a project folder...

### ...and rename it "project1"

#### Copy the relevant files into this folder



Car Branne			
Project PLINK Res	can folder		About
Set project folder	>		
Save			
Exit			
Configuration			
CLog viewer			
Log viewer No view selected			
Log viewer No view selected			
Log viewer No view selected			
Log viewer No view selected			
Log viewer No view selected			
Log viewer No view selected			
Log viewer No view selected			
Log viewer No view selected			
Log viewer No view selected			

# Select the folder you previously created

👙 Open			X
Look in:	🞯 Desktop	· · · · · · · · · · · · · · · · · · ·	<b>\$ 12</b>
Recent Desktop	My Docum My Compu My Networ gplink My Compu project1 scratch sides	ents ter 'k Places ter	
My Documents			
My Computer			
My Network Places	File <u>n</u> ame: Files of <u>type</u> :	C:\Documents and Settings\purcell\Desktop\project1	Open Cancel
My Network Places	File <u>n</u> ame: Files of <u>typ</u> e:	C:\Documents and Settings\purcell\Desktop\project1 All Files	Open Cancel

# Configuring the new project

👙 Project Configura	ition 🔀
PLINK	prefix:
PLINK path: :\Docu	ments and Settings\purcell\Desktop\boulder07\bin\plink.exe Browse
Editor options	
	Default editor: write
User specified	d editor: Browse Test
Haploview .jar path:	and Settings\purcell\Desktop\boulder07\bin\Haploview_15TEST.jar Browse
Haploview optional append	
	OK Cancel

Here, we tell gPLINK ....

... where the PLINK executable is

... specify any PLINK prefixes (advanced option for grid computing)

... where the Haploview (version 4.0) executable is

... which text editor to use to view files, e.g. WordPad (write.exe)

### Data management

- Recode dataset (A,C,G,T  $\rightarrow$  1,2)
- Reorder dataset
- Flip DNA strand
- Extract subsets (individuals, SNPs)
- Remove subsets (individuals, SNPs)
- Merge 2 or more filesets
- Compact binary file format

# Summarizing the data

- Hardy-Weinberg
- Mendel errors
- Missing genotypes
- Allele frequencies
- Tests of non-random missingness
  - by phenotype and by (unobserved) genotype
- Individual homozygosity estimates
- Stretches of homozygosity
- Pairwise IBD estimates

# Validating the fileset

# Doesn't do anything, except (attempt to) load the data and report basic statistics

👉 gPLI	NK				
Project	PLINK	Rescan fol	der		
Operati	Data	Management	•		
	Sumi	mary Statistics	•	Validate fileset	
	Stra	tification	•	Missingness	
	Asso	ciation	•	Hardy Weinberg equilibrium	
	IBD I	Estimation	•	Mendel error rates	
	Com	mand Line		Allele frequencies	
				Haplotype frequencies	
				Haplotype phases	

👙 Validate fileset	X						
Binary fileset Standard fileset Alternative phenotypes							
Quick Fileset							
example							
.bed file:	Browse						
.bim file:	Browse						
.fam file:	Browse						
.fam file:     Browse       Output file root: Invalid fileroot     C:\Documents and Settings\purcell\Desktop\project1\       Filter     Threshold     OK							
Need to enter a unique root fil	lename						

Output file root: Valid fileroot
C:\Documents and Settings\purcell\Desktop\project1\
valid1

### Then add a description (for logging)

👙 Execute Command						
PLINK command:	"C:\Documents and Settings\purcell\plink\plink.exe"bfile "example"	out "valid)				
Command description:	Validate the fileset and check the number of SNPs, individuals, etc					
Run						



### Q1) What is the genotyping rate?

Clicking on the tree to expand or contract it; individual input or output files can be selected here

The log file always gives a lot of useful information: it is good practice always to check it to confirm that an analysis has run okay.

Default filters applied here

#### 🖆 gPLINK Project PLINK About Operations viewer 🖃 🐶 🚽 🔄 🖃 🖃 🖃 🖃 🖃 🖃 🖃 🖃 🖃 🖃 🖃 🖃 "C:\Documents and Settings\purcell\Desktop\boulder07\bin\plink.exe" --bfile example --out valid1 🖃 🖂 Input files: C:\Documents and Settings\purcell\Desktop\boulder07\data\example.bed C:\Documents and Settings\purcell\Desktop\boulder07\data\example.bim C:\Documents and Settings\purcell\Desktop\boulder07\data\example.fam 😋 Output files: C:\Documents and Settings\purcell\Desktop\boulder07\data\valid1.log C:\Documents and Settings\purcell\Desktop\boulder07\data\valid1.irem Log viewer C:\Documents and Settings\purcell\Desktop\boulder07\data\valid1.log Reading map (extended format) from [ example.bim ] 228723 markers to be included from [ example.bim ] Reading pedigree information from [ example.fam ] 90 individuals read from [ example.fam ] 90 individuals with nonmissing phenotypes Assuming a disease phenotype (1=unaff, 2=aff, 0=miss) Missing phenotype value is also -9 49 cases, 41 controls and 0 missing 45 males, 45 females, and 0 of unspecified sex Reading genotype bitfile from [ example.bed ] Detected that binary PED file is v1.00 SNP-major mode Before frequency and genotyping pruning, there are 228723 SNPs Applying filters (SNP-major mode) 90 founders and 0 non-founders found Writing list of removed individuals to [ validl.irem ] 1 of 90 individuals removed for low genotyping (MIND > 0.1) Total genotyping rate in remaining individuals is 0.995473 623 SNPs failed missingness test ( GENO > 0.1 ) 46834 SNPs failed frequency test ( MAF < 0.01 ) After frequency and genotyping pruning, there are 181331 SNP: **Overall genotyping rate** <

# Viewing an output file



### Filters and thresholds



Most forms have Filter and Thresholds buttons

### Filters exclude people or SNPs based on prespecified lists, or genomic location

👙 Filter SNPs and/or Individuals 🛛 🛛 🔀
~ Ву Мар
Chromosome (chr) 1
🔿from SNP 💙 🛛to SNP 💙
O Specific SNP (snp) Optional kb window (window)
By List
O SNP set-file (set) Browse Specific gene (gene)
O SNPs extract 🗸 Browse
🔿 Individuals keep 🖌 Browse
OK Cancel

### Thresholds exclude people or SNPs based on genotype data

🖆 Threshold 🛛 🚺	<
Minor allele frequency (maf) 0.01	
Maximum minor allele frequency (max-maf) 1.0	
Maximum SNP missingness rate (geno) 0.1	
Maximum individual missingness rate (mind) 0.1	
Hardy Weinberg equilibrium (hwe)	
Mendel errors (me)	
OK Restore Default Cancel	

### **Q2) How many monomorphic SNPs?**

### We can use thresholds and the Validate fileset option to answer this:

👙 Validate fileset		
Binary fileset Standard fileset Alternative phenotypes		
Quick Fileset		
.bed file:	Browse	
.bim file:	Browse	
.fam file:	Browse	
Filter Threshold OK Cancel		<ul> <li>Minor allele frequency (maf)</li> <li>Maximum minor allele frequency (max-maf)</li> <li>Maximum SNP missingness rate (geno)</li> <li>Maximum individual missingness rate (mind)</li> <li>Maximum individual missingness rate (mind)</li> <li>Hardy Weinberg equilibrium (hwe)</li> <li>Mendel errors (me)</li> </ul>

👙 gPLINK		
Project PLINK	Rescan folder	About
Operations viewe	r	
🗉 💼 valid1: V	lidate the fileset and check the number of SNPs, ind	lividuals, etc
🗄 🔄 mono1: 0	ount of monomorphic SNPs (using thresholds)	
CLog viewer		
C:\Documents an	settinas\purcell\Desktop\project1\mono1.loa	
Assuming a d	isease phenotype (1=unaff, 2=aff, 0=	=miss)
Missing phen	otype value is also -9	-
49 cases and	41 controls	
Reading geno	type bitfile from [ example.bed ]	
Detected tha	t binary PED file is vl.00 SNP-major	r mode
Before frequ	ency and genotyping pruning, there a	are 11207 SNPs
Applying fil	ters (SNP-major mode)	
90 founders	and U non-founders found	
lotal genoty	ping rate in remaining individuals i d viscing the second second	13 0.99544
O SMDe feile	iled for more that ( WAE ( ) or WA	E > 0 >
0 SNPs fails		1 2 0 1
0 SNPs faile 9197 SNPs fa	lied frequency test ( MAF < 0 or MA.	ro 2010 SMDa
0 SNPs faile 9197 SNPs fa After freque	iled frequency test ( MAF < 0 of MA ncy and genotyping pruning, there as	re 2010 SNPs
0 SNPs faile 9197 SNPs fa After freque Analysis fin	ifed frequency test ( MAF < 0 of MA ncy and genotyping pruning, there a ished: Wed Nov 22 16:43:55 2006	re 2010 SNPs
0 SNPs faile 9197 SNPs fa After freque Analysis fin	iled frequency test ( MAF < 0 of MA ncy and genotyping pruning, there a ished: Wed Nov 22 16:43:55 2006	re 2010 SNPs

### Q3) Evidence of non-random genotyping failure?

#### The Summary Statistics/Missingness option can answer this:



### Missing rate in cases (A) and controls (U) and a test for whether rate differs

lapl	oview 4.0I	beta11							
₽is	play <u>A</u> nalys	is <u>H</u> elp							ł
чĸ									
	15								
	Result	Chrom	Marker	Position	F_MISS_A	F_MISS_U	CHISQ_MISS	P_MISS	
	1	1	rs3094315	792429	0.02083	0.0	0.8639	0.3527	~
	2	1	rs4040617	819185	0.0	0.0	0.0	1.0	
	3	1	rs4075116	1043552	0.0	0.0	0.0	1.0	
	4	1	rs9442385	1137258	0.02083	0.0	0.8639	0.3527	
	5	1	rs11260562	1205233	0.04167	0.0	1.748	0.1862	
	6	1	rs6685064	1251215	0.0	0.0	0.0	1.0	
	7	1	rs3766180	1563420	0.0	0.0	0.0	1.0	
	8	1	rs6603791	1586208	0.0	0.0	0.0	1.0	
	9	1	rs7519837	1596068	0.02083	0.0	0.8639	0.3527	
	10	1	rs3737628	1755094	0.0	0.0	0.0	1.0	
	11	1	rs7511905	1825948	0.0	0.0	0.0	1.0	
	12	1	rs3855951	1836464	0.0	0.02439	1.184	0.2765	
	13	1	rs6603803	1844850	0.02083	0.0	0.8639	0.3527	
	14	1	rs2803285	1920531	0.0	0.0	0.0	1.0	
	15	1	rs7513222	2060063	0.0	0.0	0.0	1.0	
	16	1	rs3107146	2079746	0.0	0.0	0.0	1.0	
	17	1	rs3107157	2094131	0.0	0.0	0.0	1.0	
	18	1	rs3753242	2101843	0.0	0.0	0.0	1.0	
	19	1	rs385039	2109571	0.0	0.0	0.0	1.0	~
	Filters-	Chromosome:	Start kb: View top 1	End kb:	Other:	<b>~ ~</b> Go	]	Filter Reset Fil	ters
				Go to Se	lected Region				

### Non-random genotyping failure



"Mishap" test

~10% (30,824) of SNPs with >5 missing genotypes fail mishap test at *p* < 1e-8

For example: rs7524558 has 68 missing genotypes (~2.6% missing)

Flanking haplotypes	GENO	MISSING
НОМ	2340	0
HET	49	68

## Association analysis

- Case/control
  - allelic, trend, genotypic
  - general Cochran-Mantel-Haenszel
- Family-based TDT
- Quantitative traits
- Haplotype analysis
  - focus on "multimarker predictors"
- Multilocus tests, covariates, epistasis, etc

### Standard association tests

	🖆 Allelic Association Tests	
	 Binary fileset Standard fileset Alternative phenotypes	
PLINK Rescan folder	⊂ Ouick Fileset	
j Data Management 🕨		
Summary Statistics 🕨		
Stratification	.bed file:	Browse
IBD Estimation Genotypic tests (C/C)	.bim file:	Browse
Command Line Family-based tests (TDT)	.fam file:	Browse
Stratified tests		
Quantitative trait interaction	Basic allelic single SNP association [C/C or QT] (assoc)	
Set-based tests	Confidence interval, C/C only (ci) 0.95	
Haplotype-based C/C tests	Adjusted p-values (adjust)	
Haplotype-based TDT tests	Permutation options	
	Adaptive permutation mode (perm)	
	max(T) permutation mode (mperm) 1000	
	<ul> <li>Output file root: Valid fileroot</li> </ul>	
	C:\Documents and Settings\purcell\Desktop\project1\ essoc1	
		]
	Filter Threshold OK Cancel	

#### Q4) What is the most associated SNP?

### **Q5) Evidence of stratification from genomic control?**

Project	PLINK	Rescan folder		About
Operat	tions view	er		
	valid1: V mono1: nonrand ibs1: Ge strat1: 1 assoc1: 	Validate the fileset and Count of monomorphi lom1: Test for non-rai lom2: Test for non-rai nerate IBS distance fi Two class stratification Standard association Documents and Settir ut files: put files: C:\Documents and Set C:\Documents and Set C:\Documents and Set	check the number of SNPs, individuals, etc : SNPs (using thresholds) dom missingness with respect to phenotype e for all pairs of individuals solution :est (alleleic, all SNPs) gs\purcell\plink\plink.exe"bfile "example"asso ttings\purcell\Desktop\project1\assoc1.log ttings\purcell\Desktop\project1\assoc1.assoc	ocadjustout "assoc1"
	÷	C:\Documents and Se	ttings\purcell\Desktop\project1\assoc1.assoc.adjus	sted
Log vie		C:\Documents and Se	ttings\purcell\Desktop\project1\assoc1.assoc.adjus	sted
C:\Doct	wer uments ar 90 ind:	C:\Documents and Se nd Settings\purcell\De ividuals remove	<pre>ttings\purcell\Desktop\project1\assoc1.assoc.adjus ktop\project1\assoc1.log for low genotyping ( MIND &gt; 0.1 )</pre>	sted
C:\Doct C:\Doct 1 of Total	ewer uments ar 90 ind: . genot;	C:\Documents and Se nd Settings\purcell\De ividuals remove yping rate in r	<pre>ktop\project1\assoc1.assoc.adjus ktop\project1\assoc1.log for low genotyping ( MIND &gt; 0.1 ) emaining individuals is 0.995473 s test ( GENO &gt; 0.1 )</pre>	sted
C:\Doct 1 of Total 623 S 46834	ewer uments ar 90 ind: . genoty NPs fa: L SNPs :	C:\Documents and Se nd Settings\purcell\De ividuals remove yping rate in r iled missingnes failed frequenc	<pre>ttings\purcell\Desktop\project1\assoc1.assoc.adjus ktop\project1\assoc1.log for low genotyping ( MIND &gt; 0.1 ) emaining individuals is 0.995473 s test ( GENO &gt; 0.1 ) v test ( MAF &lt; 0.01 )</pre>	sted
Log vie C:\Doct 1 of Total 623 S 46834 After	. genot NPs fa: SNPs : SNPs :	C:\Documents and Se d Settings\purcell\De ividuals remove yping rate in r iled missingnes failed frequenc ency and genoty	<pre>ktop\project1\assoc1.log tings\purcell\Desktop\project1\assoc1.log for low genotyping ( MIND &gt; 0.1 ) emaining individuals is 0.995473 s test ( GEN0 &gt; 0.1 ) y test ( MAF &lt; 0.01 ) ping pruning, there are 181331 SNPs</pre>	sted
Log vie C:\Doct 1 of Total 623 S 46834 After Writi	ewer uments ar 90 ind: NPs fa: SNPs fa: SNPs fa: freque ng main	C:\Documents and Se d Settings\purcell\De ividuals remove yping rate in r iled missingnes failed frequenc ency and genoty n association r	<pre>ktop\project1\assoc1.log f for low genotyping ( MIND &gt; 0.1 ) maining individuals is 0.995473 s test ( GENO &gt; 0.1 ) y test ( MAF &lt; 0.01 ) ping pruning, there are 181331 SNPs esults to [ assoc1.assoc ]</pre>	sted
Log vie C:\Doct 1 of Total 623 S 46834 After Writi Compu	ewer uments ar 90 ind: NPs fa: 1 SNPs : 1 SNPs : 1 freque ng main	C:\Documents and Se d Settings\purcell\De ividuals remove yping rate in r iled missingnes failed frequenc ency and genoty n association r orrected signif	<pre>ktop\project1\assoc1.log d for low genotyping ( MIND &gt; 0.1 ) emaining individuals is 0.995473 s test ( GENO &gt; 0.1 ) y test ( MAF &lt; 0.01 ) ping pruning, there are 181331 SNPs esults to [ assoc1.assoc ] icance values (FDP, Sidak, etc)</pre>	sted
Log vie C:\Doct 1 of Total 623 S 46834 After Writi Compu Genom	. genoty NPs fa: SNPs fa: SNPs fa: SNPs fa: SNPs fa: SNP	C:\Documents and Se ividuals remove yping rate in r iled missingnes failed frequenc ency and genoty n association r orrected signif lation factor (	<pre>ktop\project1\assoc1.log tings\purcell\Desktop\project1\assoc1.log for low genotyping ( MIND &gt; 0.1 ) emaining individuals is 0.995473 s test ( GENO &gt; 0.1 ) y test ( MAF &lt; 0.01 ) ping pruning, there are 181331 SNPs esults to [ assoc1.assoc ] icance velues (FDP, Sidak, etc) based on median chi-squared) is 1.2'</pre>	7915
Log vie C:\Doct 1 of Total 623 S 46834 After Writi Genom Mean Writi	. genoty SNPs fa: SNPs fa: SNPs fa: SNPs fa: SNPs fa: SNPs fa: freque ng main ting c iting c iting c iting c iting c iting c iting c	C:\Documents and Se ividuals remove yping rate in r iled missingnes failed frequenc ency and genoty n association r orrected signif lation factor ( uared statistic tiple-test corr	<pre>ktop\project1\assoc1.log tings\purcell\Desktop\project1\assoc1.log for low genotyping ( MIND &gt; 0.1 ) emaining individuals is 0.995473 s test ( GENO &gt; 0.1 ) y test ( MAF &lt; 0.01 ) ping pruning, there are 181331 SNPs esults to [ assoc1.assoc ] icance values (FDP, Sidak, etc) based on median chi-squared) is 1.2 18 1.22947 ected significance values to [ assoc</pre>	7915 cl.assoc.adjusted ]
Log vie C:\Docu 1 of Total 623 S 46834 After Writi Compu Genom Mean Writi	E wer 90 ind: 90 ind: 1 SNPs fa: 1 S	C:\Documents and Se ad Settings\purcell\De ividuals remove yping rate in r iled missingnes failed frequenc ency and genoty n association r orrected signif lation factor ( uared statistic tiple-test corr	<pre>ktop\project1\assoc1.log f for low genotyping ( MIND &gt; 0.1 ) maining individuals is 0.995473 s test ( GENO &gt; 0.1 ) y test ( MAF &lt; 0.01 ) ping pruning, there are 181331 SNPs esults to [ assoc1.assoc ] icance values (FDP, Sidak, etc) based on median chi-squared) is 1.27 13 1.22947 ected significance values to [ assoc 22 17:45:28 2006</pre>	7915 cl.assoc.adjusted ]

### Genomic control



Test locus Unlinked 'null' markers



Stratification  $\rightarrow$  adjust test statistic



harvard.edu

assoc1: Standard association test (alleleic, all SNPs)

"C:\Documents and Settings\purcell\plink\plink.exe" --bfile "example" --assoc --adjust --out "assoc1"

🗄 💼 Input files:

🖮 🧰 Output files:

💼 💼 C:\Documents and Settings\purcell\Desktop\project1\assoc1.log

C:\Documents and Settings\purcell\Desktop\project1\assoc1.irem

C:\Documents and Settings\purcell\Desktop\project1\assoc1.assoc
 C:\Documents and Settings\purcell\Desktop\project1\assoc1.assoc

s and Settings\py	Open in Editor	oc1.assoc.adjusted
	Open in Haploview	
	Edit	
	Create New	
	Delete	

#### 🔳 assoc1.assoc.adjusted - WordPad

<u>File E</u>dit <u>V</u>iew Insert Format <u>H</u>elp

#### 🗅 🚅 🖬 🎒 🕵 🛤 🐰 🖻 🛍 🗠 🧕

CHR	SNP	UNADJ	GC	BONF	HOLM	SIDAK SS	SIDAK SD	FDR BH	FDR BY	~
11	rs2513514	4.693e-007	7.131e-006	0.0851	0.0851	0.08158	0.08158	0.0644	0.817	_
20	rs6110115	7.103e-007	9.938e-006	0.1288	0.1288	0.1209	0.1209	0.0644	0.817	
8	rs7835221	1.848e-006	2.138e-005	0.335	0.335	0.2847	0.2847	0.07917	1	
11	rs2508756	2.105e-006	2.373e-005	0.3817	0.3817	0.3173	0.3173	0.07917	1	
15	rs16976702	2.183e-006	2.443e-005	0.3958	0.3958	0.3269	0.3269	0.07917	1	
8	rs11204005	7.882e-006	6.841e-005	1	1	0.7605	0.7605	0.2336	1	
9	rs16910850	1.216e-005	9.688e-005	1	1	0.8898	0.8898	0.2336	1	
12	rs1195747	1.427e-005	0.0001102	1	1	0.9248	0.9248	0.2336	1	
17	rs7207095	1.682e-005	0.0001257	1	1	0.9526	0.9526	0.2336	1	
15	rs16971118	1.907e-005	0.0001391	1	1	0.9685	0.9685	0.2336	1	
20	rs6074704	2.014e-005	0.0001452	1	1	0.974	0.974	0.2336	1	
20	rs1570484	2.014e-005	0.0001452	1	1	0.974	0.974	0.2336	1	
17	rs9944528	2.166e-005	0.000154	1	1	0.9803	0.9803	0.2336	1	
3	rs636006	2.279e-005	0.0001604	1	1	0.984	0.9839	0.2336	1	
9	rs17534370	2.307e-005	0.000162	1	1	0.9848	0.9848	0.2336	1	
21	rs2178836	2.41e-005	0.0001678	1	1	0.9873	0.9873	0.2336	1	
11	rs12418173	2.488e-005	0.0001721	1	1	0.989	0.989	0.2336	1	
11	rs898311	2.488e-005	0.0001721	1	1	0.989	0.989	0.2336	1	
11	rs7931135	2.488e-005	0.0001721	1	1	0.989	0.989	0.2336	1	
15	rs16971120	2.82e-005	0.0001903	1	1	0.994	0.994	0.2336	1	
12	rs4445711	2.834e-005	0.0001911	1	1	0.9941	0.9941	0.2336	1	
19	rs3844444	2.834e-005	0.0001911	1	1	0.9941	0.9941	0.2336	1	*
or Help, pr	ress F1									

### Haplotype based association



#### Q4b) What is the most associated haplotype?

# Specifying haplotype tests

#### Specify specific haplotypes

Predicted							Predictors					
ID	chr	сМ	bp	alle	les		Haplo	type	SNPs (in	data file)		
i_rs2906364	8	0	158484	1	2		14	'rs'	7000519	rs10488	370	
i_rs3750097	8	0	187042	1	2		23	rs	2906334	rs11988	064	
i_rs10105400	8	0	188546	1	2		23	rs	2906334	rs11988	064	
i_rs13258954	8	0	211039	1	2		34	rs	13265571	rs30082	257	
etc												

### Or, specify the locus (i.e. only specify predicting SNPs)

- \* rs7000519 rs10488370
- \* rs2906334 rs11988064
- \* rs2906334 rs13265571 rs3008257
- ... etc ...

### Or, specifying a sliding window of fixed SNPs with:

```
e.g. -- hap-window 4
```

### Haplotype-based tests



# Identity-by-state (IBS) sharing

### Pair from same population



### Pair from different population

Individual 3	A/C	G/G	A/A	A/A	G/ <b>G</b>
Individual 4	C/ <b>C</b>	T/T	G/G	C/C	A/G
IBS	1	0	0	0	1

### **Empirical assessment of ancestry**



### **Q6)** Use genotypes to cluster the sample into 2 subpopulations

Pro 0

### Step 1) Generate IBS distances for all pairs (may take a few minutes)

	👙 IBS distances 🛛 🔀
t PLINK Rescan folder	Binary fileset Standard fileset Alternative phenotypes
ati Data Management 🕨	- Ouick Eileast
Summary Statistics	
Association Clustering	
IBD Estimation Nearest Neighbour	.bed file: Browse
Command Line	.bim file: Browse
	.fam file: Browse
Project       PLINK       Rescan folder         Operations viewer       Image: SNPs, individuals, etc.         Image: SNPs       Image: SNPs, individuals, etc.         Image: SNPs       Image: SNPs (using thresholds)         Image: SNPs       Image: SNPs (using thresholds)	Create IBS distance file (genome)  IBS distance matrix (matrix)  Output file root: Valid fileroot  C:\Documents and Settings\purcell\Desktop\project1\ ibs1  Filter Threshold OK Cancel
<ul> <li>"C:\Documents and Settings\purcell\plink\plink.exe"bfile "example"q</li> <li>Input files:</li> <li>Output files:</li> <li>C:\Documents and Settings\purcell\Desktop\project1\ibs1.log</li> <li>C:\Documents and Settings\purcell\Desktop\project1\ibs1.genome</li> <li>C:\Documents and Settings\purcell\Desktop\project1\ibs1.irem</li> </ul>	

# Step 2) Cluster individuals based on IBS distances and other constraints

	Binary fileset Standard fileset Alternative phenotypes
t PLINK Rescan folder	- Ouide Electet
ati Data Management 🕨	
Summary Statistics 🕨	
Stratification	had film
Association Clustering	
IBD Estimation	.bim file: Browse
Command Line	.fam file: Browse
	IBS distance file (read-genome) ;\purcell\Desktop\project1\ibs1.genome Browse
🗄 Open 📉	Optional clustering constraints
Look jn: 📄 projecti V 💟 🗊 🖃 🖃	Pairwise population concordance threshold (pcc) 0.0ppc-gap 500.0
Recent	Identity by missingness (ibm)
	Phenotype constraint (-cc)
Desitop	
My Documents	
My Computer	Number of clusters (K) 2
Wy Network     Places     bs1.genome     Qpen       Files of type:     GENCME     Cand Open st	External categorical matching (match) Browse
Spacify proviously constant	Positive/negative matches(match-type) Browse
specify previously-generated	External quantitative matching (qmatch) Browse
IBS file (*.genome)	Thresholds (qt) Browse
	Output file root: Valid fileroot
Constrain alustar solution	C:\Documents and Settings\purcell\Desktop\project1\
Constrain cluster solution	Strat1
to two classes (K=2)	Filter Threshold OK Cancel

PLINK tutorial, December 2006; Shaun Purcell, shaun@pngu.mgh.harvard.edu

X



#### 🗒 strat2.cluster1 - WordPad

<u>File Edit View Insert Format Help</u>

#### 🗅 🚄 🖨 🗟 🛤 🐰 🖻 🛍 🗠 🤜

SOL-O CH18526 NA18526 CH18637 NA18637 CH18561 NA18561 CH18566 NA18566 CH18540 NA18540 CH18563 NA18563 CH18573 NA18573 CH18545 NA18545 CH18609 NA18609 CH18577 NA18577 CH18550 NA18550 CH18582 NA18582 CH18636 NA18636 CH18555 NA18555 CH18571 NA18571 CH18558 NA18558 CH18532 NA18532 CH18622 NA18622 CH18623 NA18623 CH18547 NA18547 CH18612 NA18612 CH18524 NA18524 JA18976 NA18976 CH18562 NA18562 CH18620 NA18620 CH18593 NA18593 CH18537 NA18537 CH18635 NA18635 CH18529 NA18529 CH18603 NA18603 CH18570 NA18570 CH18632 NA18632 CH18572 NA18572 CH18579 NA18579 CH18621 NA18621 CH18633 NA18633 CH18605 NA18605 CH18594 NA18594 CH18552 NA18552 CH18624 NA18624 CH18542 NA18542 CH18611
N&18611 CH18564 N&18564 CH18608 N&18608 CH18576 N&18576 CH18592 N&18592
_NA18980 JA18948_NA18948 JA18975_NA18975 JA18943_NA18943 JA18969_NA18969 JA18945_NA18945 JA18973
NA18973 JA18972 NA18972 JA19007 NA19007 JA18974 NA18974 JA18991 NA18991 JA18978 NA18978 JA18994
NA18994 JA18990 NA18990 JA18998 NA18998 JA18992 NA18992 JA18940 NA18940 JA18967 NA18967 JA18959
NA18959 JA18960 NA18960 JA18966 NA18966 JA18970 NA18970 JA19005 NA19005 JA18951 NA18951 JA18947

For Help, press F1

PLINK tutorial, December 2006; Shaun Purcell, shaun@pngu.mgh.harvard.edu



strat2.cluster	2 - Word	Pad				
<u>File E</u> dit <u>V</u> iew In:	sert F <u>o</u> rm	at <u>H</u> elp				
D 🖻 🔒 🎒	À 🖊	X Ba	🛍 🗠	<b>e</b> 9		
CH18576 NA18	3576	0				~
CH18577 NA18	3577	0				
CH18624 NA18	3624	0				
CH18579 NA18	3579	0				
CH18632 NA18	3632	0				
CH18582 NA18	3582	0				
CH18633 NA18	3633	0				
CH18635 NA18	3635	0				_
CH18592 NA18	3592	0				
CH18636 NA18	3636	0				_
CH18593 NA18	3593	0				
CH18637 NA18	3637	0				
JA18942 NA18	3942	1				
JA18940 NA18	3940	1				
JA18951 NA18	3951	1				
JA18943 NA18	3943	1				
JA18947 NA18	3947	1				
JA18944 NA18	3944	1				
JA18945 NA18	3945	1				
JA18949 NA18	3949	1				
JA18948 NA18	3948	1				~
For Help, press F1						1.1

### Stratified analysis

Cochran-Mantel-Haenszel test



Stratified 2×2×K tables

Binary fileset Standard fileset Alternative phenotypes
Quick Fileset       example       .bed file:   Browse
.bim file: Browse .fam file: Browse
Cochran-Mantel-Haenszel, SNP-DISEASE   CLUSTER (mh) Cochran-Mantel-Haenszel, SNP-CLUSTER   DISEASE (mh2) Test of homogeneous association (homog)
Options         □ Cluster by family (family)         ☑ Cluster file (within) ]s\purcell\Desktop\project1\strat1.cluster2 Browse         □ Breslow-Day test of heterogeneous CMH ORs (bd)         ☑ Adjusted p-values (adjust)         □ Confidence intervals (ci) 0.95         Output file root: Valid fileroot         C:\Documents and Settings\purcell\Desktop\project1\

### **Q7)** Evidence of stratification conditional on cluster solution?

👙 gPL	INK									
Project	PLINK	Rescan folder		About						
Operat	ions viewe	r								
÷	nonrando	m2: Test for non-rar	ndom missingness with respect to genotype	^						
📄 🖻 ··· 💼	ibs1: Gene	erate IBS distance fi	le for all pairs of individuals							
strat1: Two class stratification solution										
Emergence Standard association test (alleleic, all SNPs)										
	condasso(	curvents and Settin	iation, conditional on K=2 solution (strat1) using CMH test .ashpursell\plipk\plipk eye"bfile "eyemple"mbediustwithin "C\Doci	ments and Sett						
	- Ci (Di - Ci (Di	files:	gs(purce)(pinik(pinik,exebrile exampleminiau)astwithin C.(bott	intents and bett						
	- Outpu	ut files:								
	i 👿 💼 c	:\Documents and Se	ttings\purcell\Desktop\project1\condassoc1.log							
	🗄 🛅 C	:\Documents and Se	ttings\purcell\Desktop\project1\condassoc1.cmh.adjusted							
	🗄 💼 C	:\Documents and Se	ttings\purcell\Desktop\project1\condassoc1.cmh							
	🗄 💼 C	:\Documents and Se	ttings\purcell\Desktop\project1\condassoc1.irem	<u> </u>						
				>						
⊂Loa vie	wer									
C:\Doci	ments and	Settings\purcell\De	sktop/project1/condassoc1.log							
Annly	ing filt	ters (SNP-maio	r mode)	~						
90 fo	unders a	and O non-foun	ders found							
Writi	ng list	of removed in	dividuals to [ condassocl.irem ]							
l of	90 indiv	viduals remove	d for low genotyping ( MIND > 0.1 )							
Total	genoty	ping rate in r	emaining individuals is 0.995473							
623 S	NPs fai	led missingnes	s test ( GENO > 0.1 )							
46834	SNPs fa	ailed frequenc	y test ( $MAF < 0.01$ )							
After	freque	ncy and genoty	ping pruning, there are 181331 SNPs							
Cochr	an-Mante	el-Haenszel 2x	2xK test, K = 2							
Writi	ng resul	lts to [ conda	ssocl.cmh ]							
Compu	ting cou	rrected cignif	icance values (FDR, Sidek, etc)							
Genom	ic infla	ation factor (	based on median chi-squared) is 1.01355							
Mean	cni-squ	red statistic	is 0.997675	=						
Writi	ng mult:	iple-test corr	ected significance values to [ condassocl.cmh.adjust(	ed ]						
Analy	sis fin:	ished: Wed Nov	22 17:57:22 2006	~						
<				>						

#### **Q8)** What is the best SNP controlling for stratification?



ploview 4.	Obeta11							
<u>)</u> isplay <u>A</u> na	lysis <u>H</u> elp							
1.5								
Result	Chrom	Marker	Position	CHISQ	P_CMH	OR_CMH	L95	U95
1	8	rs7835221	2878098	31.08	2.481E-8	0.05766	0.0167	0.1991
2	8	rs11204005	12895576	24.59	7.081E-7	0.1031	0.03723	0.2858
3	8	rs2460338	12914531	21.3	3.933E-6	0.103	0.03415	0.3107
Chror	nosome: 🔤 💌	Start kb:	End kb:	Other:	P_CMH	✓ <= ✓ :	le-5	Filter
-							-	
Load	Additional Result	s						Reset Filters

### Making a Haploview fileset

👙 Generate fileset	
Binary fileset Standard fileset Alternative phenotypes	
Quick Fileset	
example	Select 200kb region
.bed file:	Browse around our Dest mit
.bim file:	Browse
.fam file:	Browse
<ul> <li>Standard fileset w/ allele recoding (recode12)</li> <li>Raw genotype file (recodeAD)</li> <li>Haploview fileset (recodeHV)</li> <li>Binary fileset (make-bed)</li> <li>Output file root: Valid fileroot</li> <li>C:\Documents and Settings\purcell\Desktop\project1\</li> <li>chr8region</li> <li>Filter</li> <li>Threshold</li> <li>OK</li> <li>Cancel</li> </ul>	By Map   Chromosome (chr)   Image: state of the
	OK Cancel

b gPL	INK	5 (II								
Project	PLINK	Rescan folder			About					
Operations viewer										
🖅 🧰 valid1: Validate the fileset and check the number of SNPs, individuals, etc										
mono1: Count of monomorphic SNPs (using thresholds)										
nonrandom1: Test for non-random missingness with respect to phenotype										
terminia in the second										
	strat1: T	wo class stratification	solution							
÷	assoc1:	Standard association t	est (alleleic, all SNPs)							
÷	condass	oc1: Single SNP associ	ation, conditional on K=2	solution (strat1) using CMH test						
	chr8regi   הסיייייייייייייייייייייייייייייייייייי	on: Create a Haplovie Decuments and Settin	w fileset of 200kb surroun	ding best SNP (chr8, rs7835221) bfile "example"space rs7825221indew 200	recodel-U out "					
		Documents and Setting It files:	gs(purce)(p)ink(p)ink.exe	bnie examplesnprs7835221window 200	recouenvouc					
	- Cutr	out files:								
		C:\Documents and Sel	tings\purcell\Desktop\pro	ject1\chr8region.log						
	主 ··· 💼	C:\Documents and Sel	tings\purcell\Desktop\pro	ject1\chr8region.recode.info						
	±	C:\Documents and Se	Open in Editor	ect1\chr8region.recode.ped						
<			Open in Haploview		>					
			Edit							
Log vie	wer		Create New							
C:\Doci	uments an	d Settings\purcell\Des	Delete	og						
Readi	ng geno	type bitfile fi	com [ example.bed	1	-					
Detec Rofor	tea tha	at binary PED E:	LIE 13 VI.UU SNP-M ming pruning the	ajor mode re ere 59 gNDa						
Annly	ina fil	ters (SNP-main)	(ping praning, and mode)	TE die 30 JMrs						
90 fo	unders	and 0 non-found	lers found							
Total	genoty	ping rate in re	emaining individua	ls is 0.994253						
O SNPs failed missingness test ( GENO > 1 )										
O SNP	s faile	a frequency tes	st ( MAF < 0 )							
After	freque	ency and genoty	oing pruning, ther	e are 58 SNPs						
Writi Writi	ng rec( ng Heri	ouea pea Ille t( oView_formet ma	) [ cnroregion.rec on file [ chr8regi	on recode info l						
WIICI	ng nap.	CALEM-LOLMAC W	ip rife ( curoregi	on.recode.into j						
Analy	sis fir	nished: Wed Nov	22 18:08:01 2006							





<u>ل</u>	Haploview 4.0beta11 chr8region.recode.ped												
<u>F</u> ile	jile <u>D</u> isplay <u>A</u> nalysis <u>H</u> elp Key												
LD P	LD Plot Haplotypes Check Markers Tagger												
	#	Name	Position	ObsHET	PredHET	HWpval	%Geno	FamTrio	MendErr	MAF	Alleles	Rating	
	1	rs4644261	12707252	0.0	0.0	1.0	100.0	0	0	0.0	C:C		~
	2	rs4831834	12731869	0.337	0.495	0.0041	98.9	0	0	0.449	A:⊂	Image: A start of the start	
	3	rs7833301	12737416	0.0	0.0	1.0	100.0	0	0	0.0	G:G		
	4	rs7812965	12737472	0.3	0.396	0.0393	100.0	0	0	0.272	G:T	Image:	
	5	rs6981317	12739561	0.438	0.488	0.4267	98.9	0	0	0.421	C:T	Image: A start of the start	
	6	rs10102302	12745345	0.3	0.299	1.0	100.0	0	0	0.183	G:A	Image: A start of the start	
	7	rs13282410	12745511	0.311	0.358	0.3153	100.0	0	0	0.233	T:C	Image: A start of the start	
	8	rs12677284	12752271	0.303	0.316	0.8914	98.9	0	0	0.197	C:T	Image: A start of the start	
	9	rs12547628	12761935	0.311	0.411	0.0379	100.0	0	0	0.289	C:T	Image: A start of the start	
	10	rs17121059	12763195	0.1	0.095	1.0	100.0	0	0	0.05	C:G	<b>~</b>	
	11	rs7840130	12763279	0.089	0.085	1.0	100.0	0	0	0.044	T:A	<b>~</b>	
	12	rs10105014	12764400	0.567	0.498	0.2999	100.0	0	0	0.472	T:⊂	<b>~</b>	
	13	rs11778591	12764720	0.2	0.231	0.3622	100.0	0	0	0.133	C:A	<b>~</b>	
	14	rs7828117	12772943	0.494	0.424	0.2063	96.7	0	0	0.305	G:T	<b>~</b>	
	15	rs6991079	12777125	0.189	0.206	0.6647	100.0	0	0	0.117	T:G		~
L	15       rs6991079       12777125       0.189       0.206       0.6647       100.0       0       0       0.117       T:G       V         16       rs6991079       12700010       0.400       0.400       0.507       0.677       0       0.466       0.67         16       rs6991079       12700010       0.400       0.400       0.507       0.677       0       0.466       0.67         HW p-value cutoff:       0.0010         Min genotype %:       75         Max # mendel errors:       1         Minimum minor allele freq.       0.0010         Select All       Deselect All       Reset Values       Rescore Markers												

👙 Haploview 4.0beta11 chr8region.recode.ped									
<u>File D</u> isplay <u>A</u> nalysis <u>H</u> elp Key									
LD Plot Haplotypes Check Markers Tagger									
Configuration Results									
#	Name	Position	Design Score	Force Include	Force Exclude	Capture this Allele?			
2	rs4831834	12731869	0			<ul> <li>Image: A set of the set of the</li></ul>	~		
4	rs7812965	12737472	0			Image: A start and a start			
5	rs6981317	12739561	0			Image: A start of the start			
6	rs10102302	12745345	0			Image: A start of the start			
7	rs13282410	12745511	0						
8	rs12677284	12752271	0						
9	rs12547628	12761935	0						
10	rs17121059	12763195	0						
11	rs7840130	12763279	0						
12	rs10105014	12764400	0						
13	rs11778591	12764720	0						
14	rs7828117	12772943	0						
15	rs6991079	12777125	0						
16	rs4831378	12783013	0				~		
Include All Exclude All Reset Table									

# In the remaining time (if any...)

- Extract as a new PLINK fileset just the single best SNP (rs7835221)
- Using this new file, attempt questions 9-14.
  - Here are some clues
    - 9) Summary statistics → Hardy Weinberg
    - 10) Standard association test, with an alternate phenotype
    - 11) Stratified association with Breslow-Day test
    - 12) You've already calculated these (i.e. \*.assoc, \*.hwe)
    - 13) This is already calculated also (i.e. \*.missing)
    - 14) Use genotypic association test

Consult the PLINK documentation (http://pngu.mgh.harvard.edu/purcell/plink/)

# In summary

- We performed whole genome
  - summary statistics and QC
  - stratification analysis
  - conditional and unconditional association analysis
- We found a single SNP rs7835221 that...
  - is genome-wide significant
  - has similar frequencies and effects in Japanese and Chinese subpopulations
  - shows no missing or HW biases
  - is consistent with an allelic, dosage effect
  - has common T allele with strong protective effect (~0.05 odds ratio)

### Acknowledgements

Haploview development Julian Maller (g)PLINK development

Shaun Purcell

Dave Bender Jeff Barrett Mark Daly Kathe Todd-Brown Ben Neale Mark Daly Pak Sham