PLINK / Haploview

Whole genome association software tutorial

Benjamin Neale
Leuven Belgium
12th August 2008

KING'S College LONDON
University of London

CAGR

BROAD INSTITUTE

Overview of the **PLINK** software package

↓

A simulated WGAS dataset

↓

Summary statistics and quality control

↓

Assessment of population stratification

↓

Whole genome association screen

↓

Further exploration of ‘hits’

↓

Visualization and follow-up using Haploview

Overview of the **PLINK** software package

A simulated WGAS dataset

Summary statistics and quality control

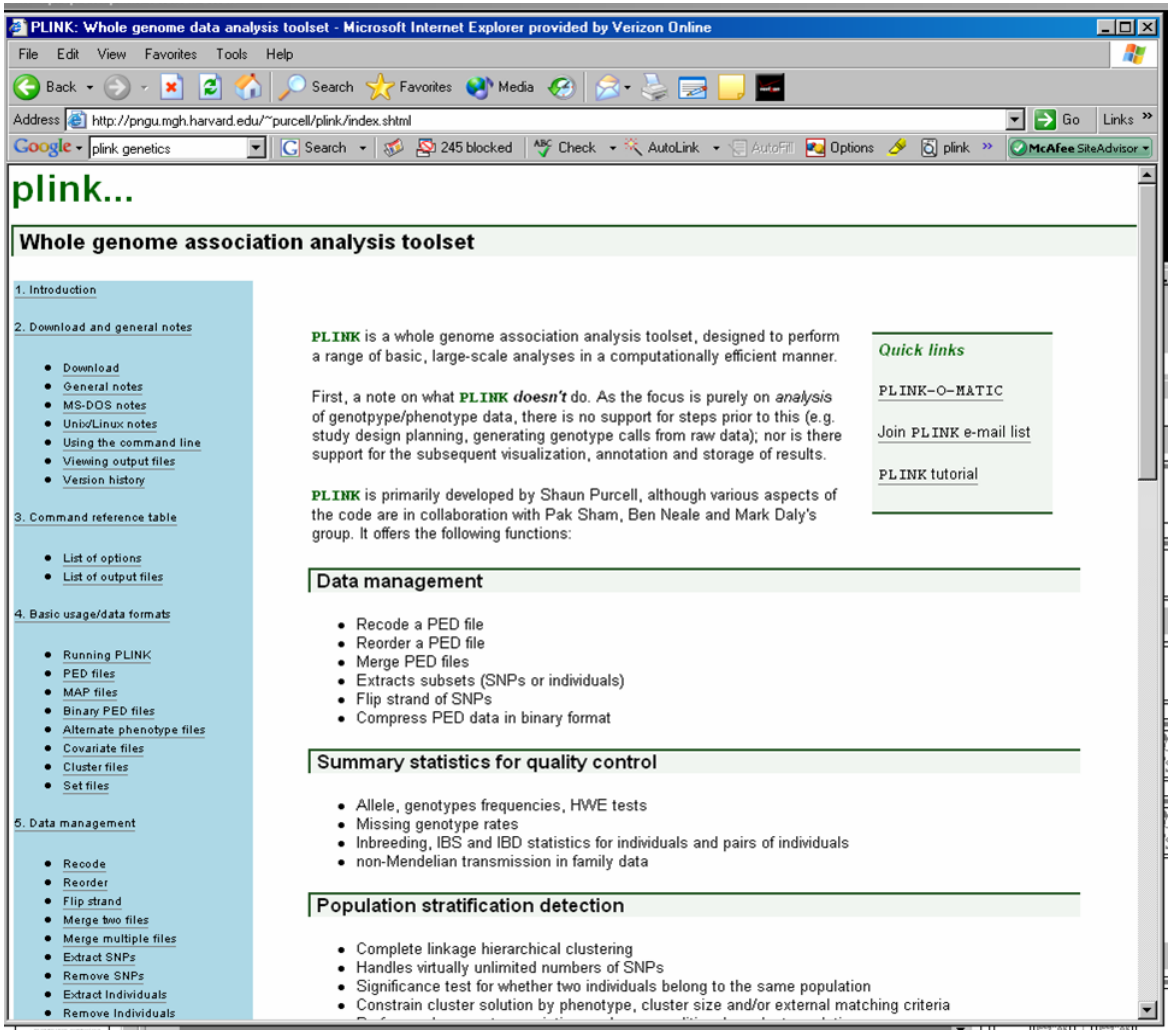
Assessment of population stratification

Whole genome association screen

Further exploration of ‘hits’

Visualization and follow-up using Haploview

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



- Data management
- Summary statistics
- Population stratification
- Association analysis
- IBD estimation

Cardinal rules

- Always consult the log file, console output
- Also consult the web documentation
  - regularly
- `PLINK` has no memory
  - each run loads data anew, previous filters lost
- Exact syntax and spelling is important
  - “minus minus” …

Overview of the **PLINK** software package

A simulated WGAS dataset

Summary statistics and quality control

Assessment of population stratification

Whole genome association screen

Further exploration of ‘hits’

Visualization and follow-up using Haploview

Simulated WGAS dataset

- Real genotypes, but a simulated “disease”
- 90 Asian HapMap individuals
  - ~228.7K autosomal SNPs
- Simulated quantitative phenotype; median split to create a disease phenotype
- Illustrative, not realistic!

Questions asked in this demonstration

1) What is the **genotyping rate**?

2) How many **monomorphic SNPs** are there in this sample?

3) Is there evidence of **non-random genotyping failure**?

4) Is there evidence for **stratification** in the sample? Does our knowledge about the **different populations** correct for this bias?

5) What is the single **most associated SNP not controlling for stratification**? Does it reach genome-wide significance?

6) Is there evidence for **stratification conditional on the two-cluster solution**?

7) What is the **best SNP controlling for stratification**. Is it genome-wide significant?

For the most highly associated SNP:

8) Does this SNP pass the **Hardy-Weinberg** equilibrium test?

9) Does this SNP **differ in frequency** between the two populations?

10) Is there evidence that this SNP has a **different association** between the two populations?

11) What are the **allele frequencies** in cases and controls? **Genotype** frequencies? What is the **odds ratio**?

12) Is the rate of **missing data** equal between cases and controls for this SNP?

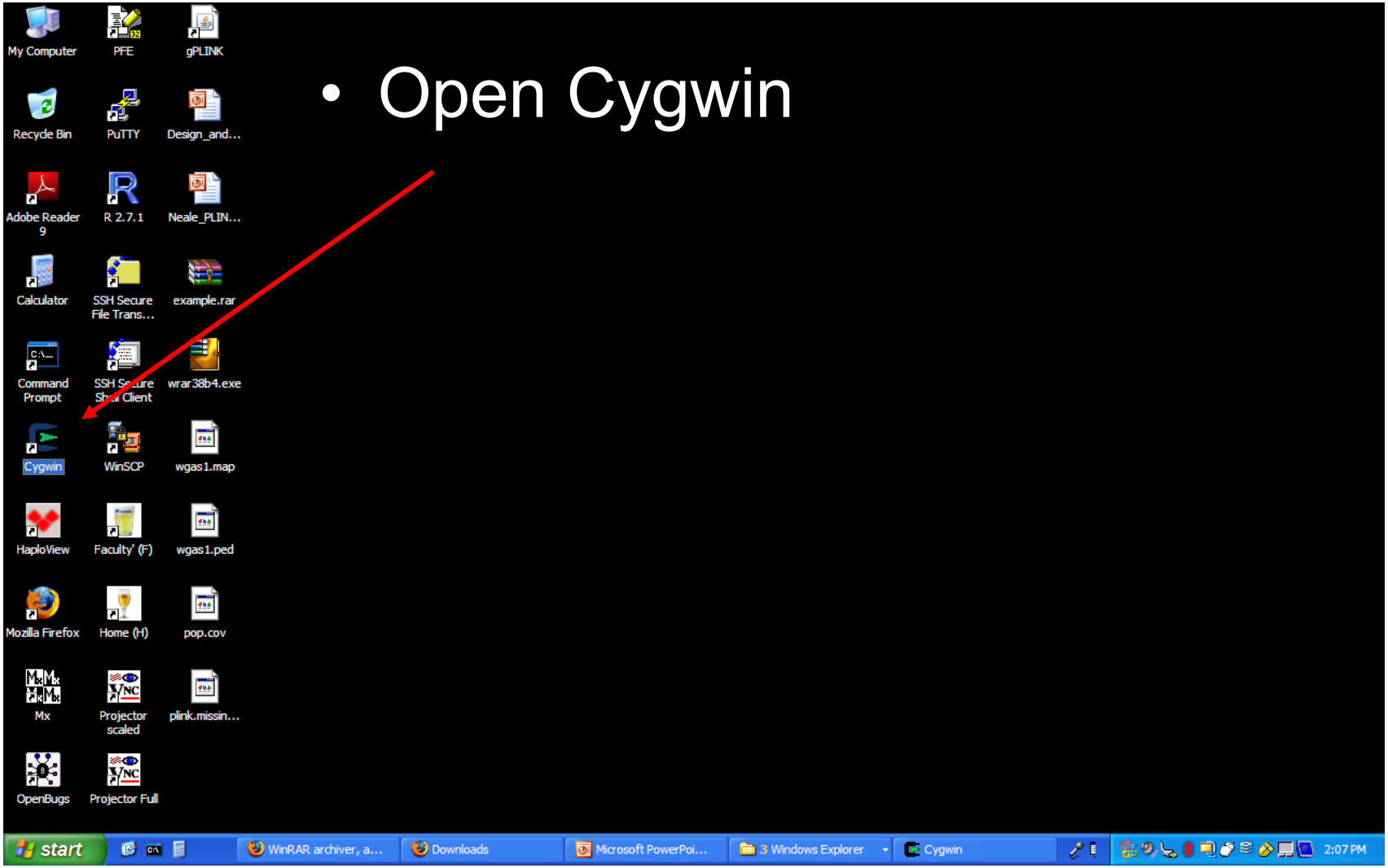
13) Does an additive model well characterize the association? What about **genotypic, dominant models**, etc?

Data used in this demonstration

- Available at http://pngu.mgh.harvard.edu/~purcell/plink/hapmap1.zip/

`wgas1.ped`	Text format genotype information
`wgas1.map`	Map file (6 fields: each row is a SNP: chromosome, RS #, genetic position, physical position, allele 1, allele 2)
`chinese.set`	FID and IID for all Chinese individuals
`pop.cov`	Chinese/Japanese population indicator (FID, IID, population code)

Preparation for running PLINK



- Open Cygwin

Changing directory

```
Your group is currently "mkpasswd".  This indicates that
the /etc/passwd (and possibly /etc/group) files should be rebuilt.
See the man pages for mkpasswd and mkgroup then, for example, run
mkpasswd -l [-d] > /etc/passwd
mkgroup  -l [-d] > /etc/group
Note that the -d switch is necessary for domain users.
ben@QMSR:~$ _
```

ls [lists the contents of directory]

cd [changes the directory for work]

You need to cd into the directory where you downloaded the data off the server

Making a binary PED fileset

If no `--out` *`{filename}`* option is specified, then all new files have the form: `plink.{extension}`

```
plink --file wgas1 --make-bed --out example
```

`--file` *`{filename}`* loads in two files `wgas1.ped` and `wgas1.map`

Three files will be created:

`example.bed` (genotypes)
`example.bim` (map file)
`example.fam` (family/phenotype)

- By default, no filtering occurs at this stage
  - all individuals and SNPs are included in the binary fileset

Overview of the **PLINK** software package

A simulated WGAS dataset

Summary statistics and quality control

Assessment of population stratification

Whole genome association screen

Further exploration of ‘hits’

Visualization and follow-up using Haploview

Data management

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

Extracting the Chinese subset

Three files will be created:
`Chinese.bed` (binary ped)
`Chinese.fam` (family file)
`Chinese.map` (map file)

```
plink --bfile example --keep chinese.set --out
Chinese --make-bed
```

Extracts out the individuals listed in the chinese.set file

- Log file will display the overall genotyping rate
- All SNPs and individuals are included for making ped files

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

What is the genotyping rate?

Two files will be created:

`plink.imiss` (individual)
`plink.lmiss` (SNP)

```
plink --bfile example --missing
```

`--bfile` *`{filename}`*
loads in the binary format fileset (genotype, map and pedigree files)

- Log file will display the overall genotyping rate
- By default, low genotyping individuals are first excluded

How many monomorhpic SNPs?

Using filters to include all individuals and SNPs

- `--mind` (individual missing rate)
- `--geno` (genotype missing rate)
- `--maf` (SNP allele frequency)

```
plink --bfile example --mind 1 --geno 1 --maf 0
      --max-maf 0
```

Filter `--max-maf` sets the maximum minor allele frequency

•Command must still be entered all on a single line

Evidence for non-random genotyping failure?

Association between failure and phenotype (per SNP)?

```
plink --bfile example –-test-missing
```

```
plink --bfile example –-test-mishap [do not attempt]
```

Association between failure and genotype (per SNP)?

- These two commands generate output files `plink.missing` and `plink.missing.hap` respectively.

An example of non-random genotyping failure

--test-mishap (plink.missing.hap)

LOCUS	HAPLOTYPE	F_0	F_1	M_H1	M_H2	CHISQ	P
rs1631460	33	0.5	0.0183	7/3	7/161	56.4	5.77e-14
rs1631460	22	0.5	0.9820	7/161	7/3	56.4	5.77e-14
rs1631460	HETERO	1.0	0.0366	7/3	0/79	60.0	9.39e-15

For this particular SNP, genotyping failure consistently occurs on a particular haplotypic background …

Overview of the **PLINK** software package

A simulated WGAS dataset

Summary statistics and quality control

Assessment of population stratification

Whole genome association screen

Further exploration of ‘hits’

Visualization and follow-up using Haploview

Population stratification: confounding

- Artificially inflates test statistic distribution
- Detectable using genome-wide data
  - I'll be speaking Thursday modeling stratification
- We'll run association and correct for it

Overview of the **PLINK** software package

A simulated WGAS dataset

Summary statistics and quality control

Assessment of population stratification

Whole genome association screen

Further exploration of ‘hits’

Visualization and follow-up using Haploview

Association analysis

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

Most highly associated SNP?

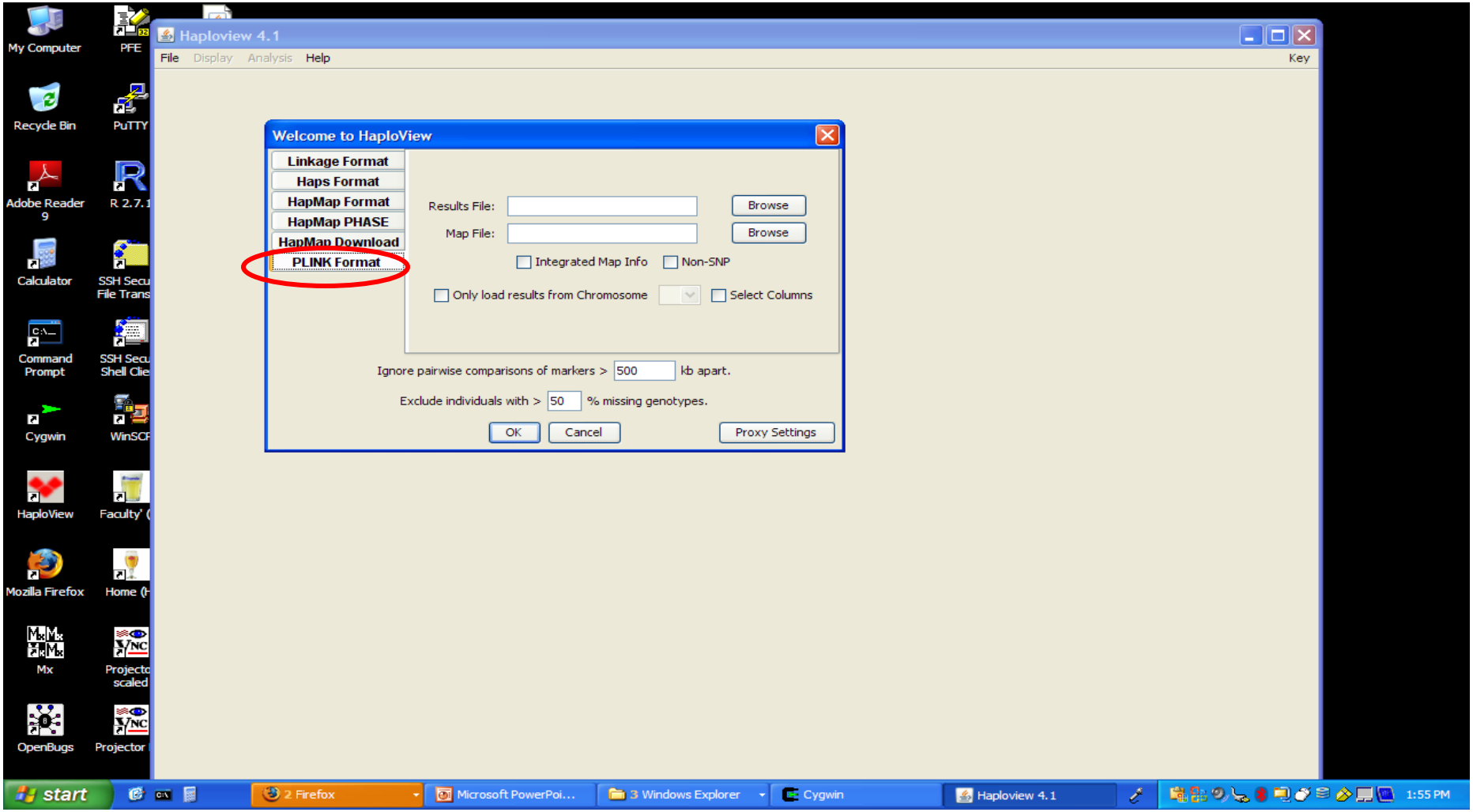
Standard case/control association

```
plink --bfile example –-assoc --adjust
```

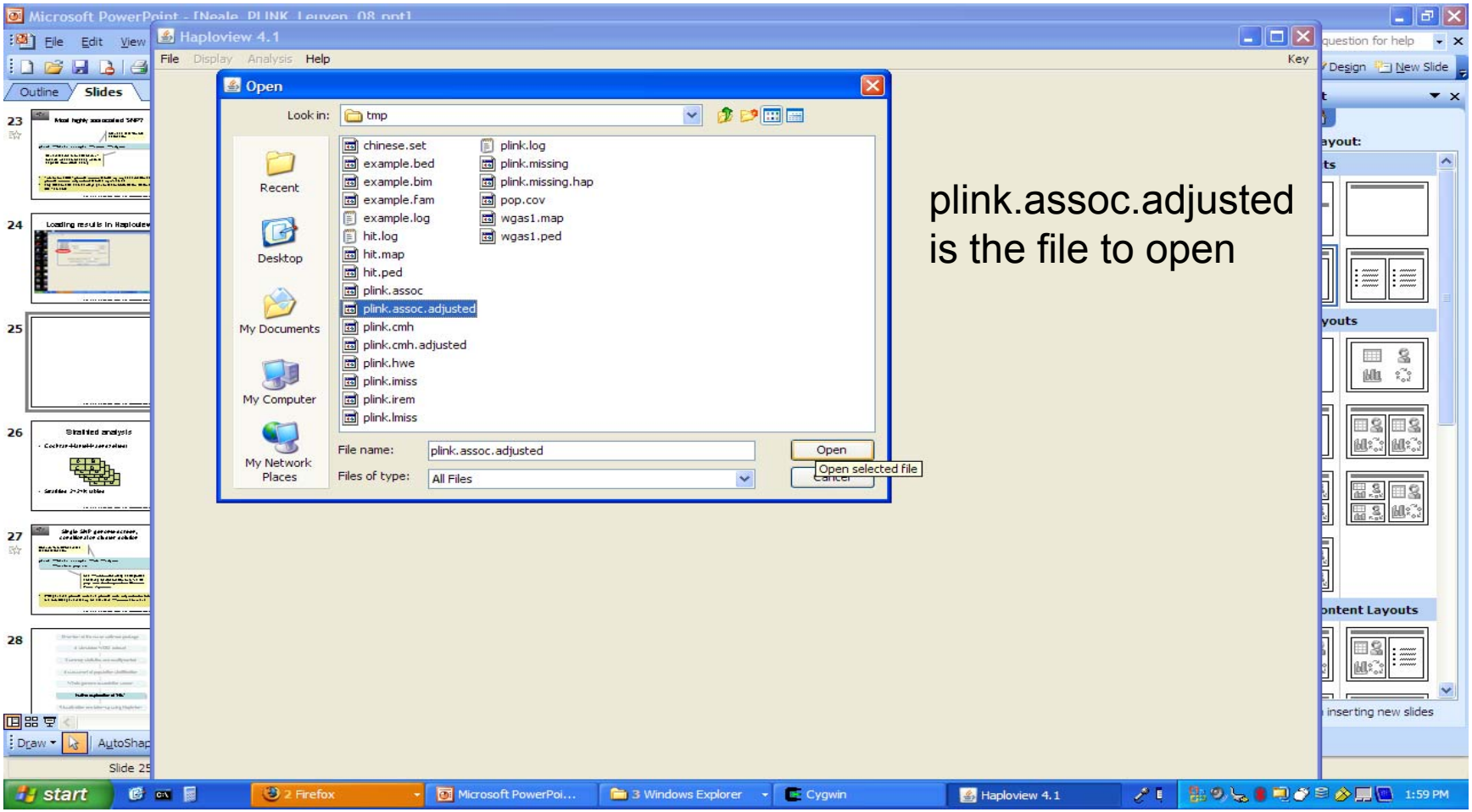
Generate extra output file of rank-ordered p-values, including p-values adjusted for multiple testing

- Two output files: `plink.assoc` (sorted by physical position) and `plink.assoc.adjusted` (sorted by p-value)
- Log file/console also displays genomic control inflation factor in log file / console

Loading results in Haploview



File Selection



plink.assoc.adjusted is the file to open

Results file

Haploview 4.1 -- plink.assoc.adjusted

File Display Analysis Help

PLINK

CHROM	MARKER	POSITION	UNADJ	GC	BONF	HOLM	SIDAK_SS	SIDAK_SD	FDR_BH	FDR_BY
11	rs2513514	75922141	4.693E-7	7.131E-6	0.08508	0.08508	0.08156	0.08156	0.06439	0.8168
20	rs6110115	13911728	7.103E-7	9.938E-6	0.1288	0.1288	0.1208	0.1208	0.06439	0.8168
11	rs2508756	75921549	2.105E-6	2.373E-5	0.3816	0.3816	0.3172	0.3172	0.09895	1.0
15	rs16976702	54120691	2.183E-6	2.443E-5	0.3958	0.3958	0.3268	0.3268	0.09895	1.0
8	rs11204005	12895576	7.882E-6	6.841E-5	1.0	1.0	0.7604	0.7604	0.2446	1.0
9	rs16910850	94478347	1.216E-5	9.688E-5	1.0	1.0	0.8897	0.8897	0.2446	1.0
12	rs1195747	129970575	1.427E-5	1.102E-4	1.0	1.0	0.9248	0.9248	0.2446	1.0
17	rs7207095	77933018	1.682E-5	1.257E-4	1.0	1.0	0.9526	0.9526	0.2446	1.0
15	rs16971118	77672467	1.907E-5	1.391E-4	1.0	1.0	0.9685	0.9685	0.2446	1.0
20	rs1570484	14139687	2.014E-5	1.452E-4	1.0	1.0	0.974	0.974	0.2446	1.0
20	rs6074704	14115283	2.014E-5	1.452E-4	1.0	1.0	0.974	0.974	0.2446	1.0
17	rs9944528	77894039	2.166E-5	1.54E-4	1.0	1.0	0.9803	0.9803	0.2446	1.0
3	rs636006	32426349	2.279E-5	1.604E-4	1.0	1.0	0.9839	0.9839	0.2446	1.0
9	rs17534370	70297172	2.307E-5	1.62E-4	1.0	1.0	0.9848	0.9847	0.2446	1.0
21	rs2178836	39421114	2.41E-5	1.678E-4	1.0	1.0	0.9873	0.9873	0.2446	1.0
11	rs898311	112128501	2.488E-5	1.721E-4	1.0	1.0	0.989	0.989	0.2446	1.0
11	rs7931135	112116789	2.488E-5	1.721E-4	1.0	1.0	0.989	0.989	0.2446	1.0
11	rs12418173	112133479	2.488E-5	1.721E-4	1.0	1.0	0.989	0.989	0.2446	1.0
15	rs16971120	77673011	2.82E-5	1.903E-4	1.0	1.0	0.994	0.994	0.2446	1.0
19	rs3844444	46125887	2.834E-5	1.911E-4	1.0	1.0	0.9941	0.9941	0.2446	1.0
12	rs4445711	103139068	2.834E-5	1.911E-4	1.0	1.0	0.9941	0.9941	0.2446	1.0
10	rs12098374	98947420	3.021E-5	2.012E-4	1.0	1.0	0.9958	0.9958	0.249	1.0
9	rs1548299	3640174	3.349E-5	2.186E-4	1.0	1.0	0.9977	0.9977	0.2632	1.0
19	rs7251418	46033429	3.485E-5	2.256E-4	1.0	1.0	0.9982	0.9982	0.2632	1.0
6	rs544263	49439359	3.868E-5	2.453E-4	1.0	1.0	0.9991	0.9991	0.2743	1.0
8	rs10098173	79972581	4.258E-5	2.651E-4	1.0	1.0	0.9996	0.9996	0.2743	1.0
8	rs9298320	79972713	4.258E-5	2.651E-4	1.0	1.0	0.9996	0.9996	0.2743	1.0
11	rs12418739	112122631	4.43E-5	2.736E-4	1.0	1.0	0.9997	0.9997	0.2743	1.0

Viewing 181302 results

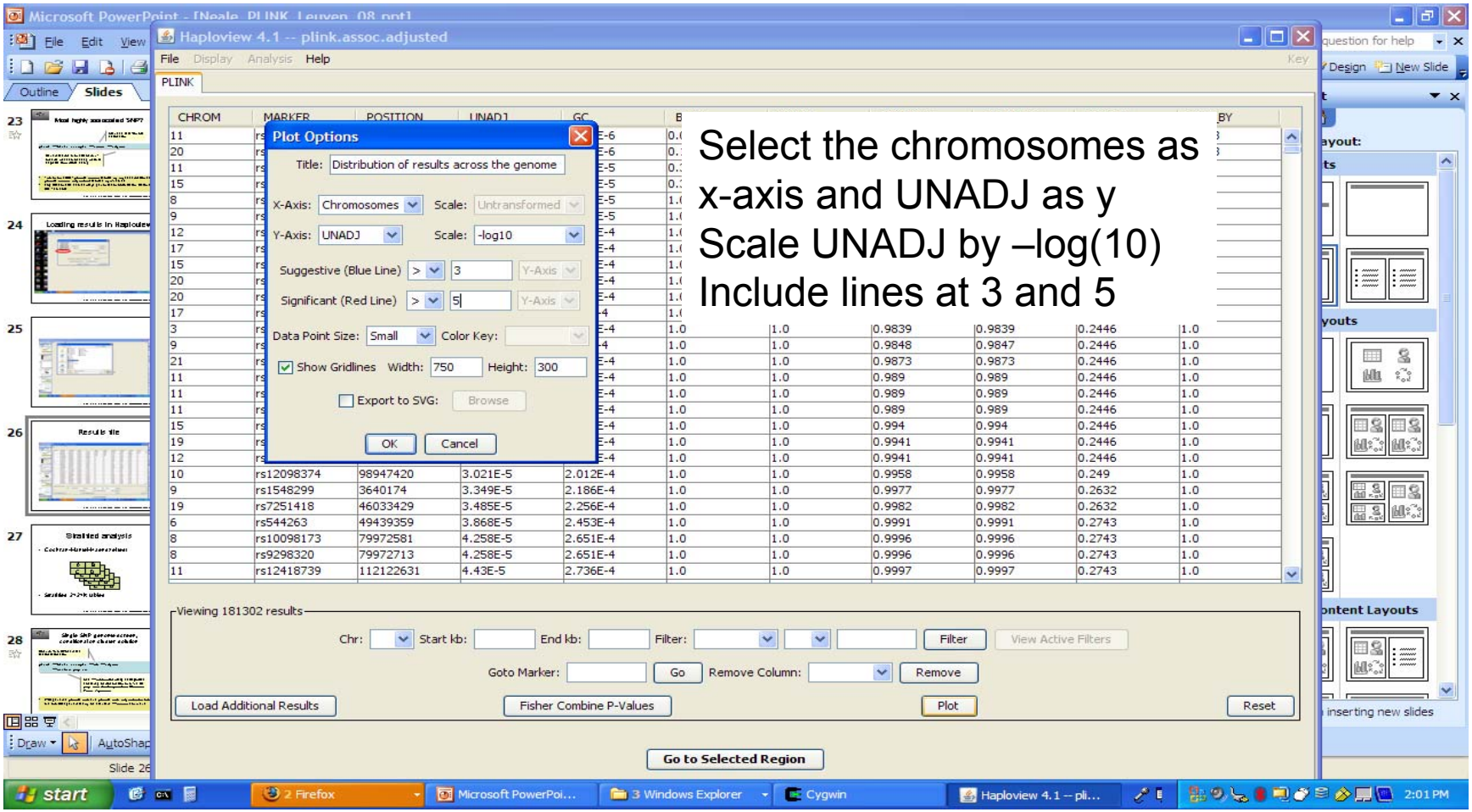
Chr: Start kb: End kb: Filter: Filter View Active Filters

Goto Marker: Go Remove Column: Remove

Load Additional Results Fisher Combine P-Values Plot Reset

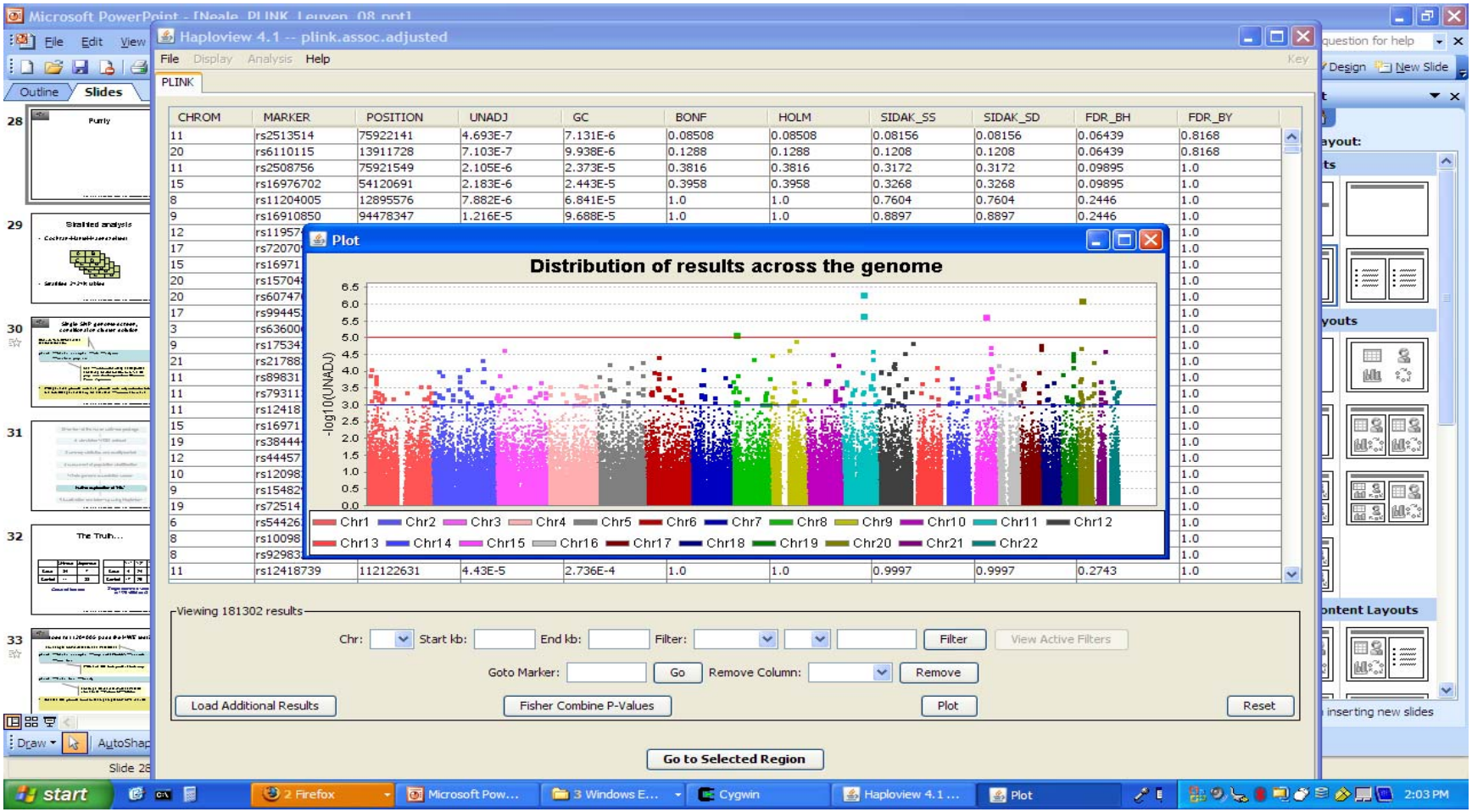
Go to Selected Region

Plotting



Select the chromosomes as x-axis and UNADJ as y
Scale UNADJ by –log(10)
Include lines at 3 and 5

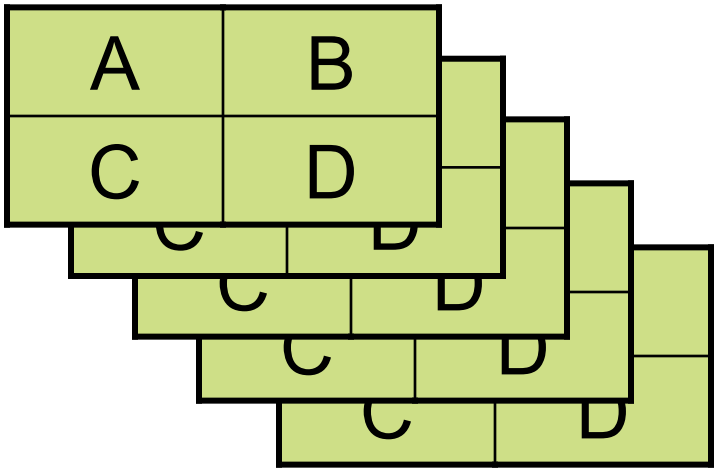
Purrty



CHROM	MARKER	POSITION	UNADJ	GC	BONF	HOLM	SIDAK_SS	SIDAK_SD	FDR_BH	FDR_BY
11	rs2513514	75922141	4.693E-7	7.131E-6	0.08508	0.08508	0.08156	0.08156	0.06439	0.8168
20	rs6110115	13911728	7.103E-7	9.938E-6	0.1288	0.1288	0.1208	0.1208	0.06439	0.8168
11	rs2508756	75921549	2.105E-6	2.373E-5	0.3816	0.3816	0.3172	0.3172	0.09895	1.0
15	rs16976702	54120691	2.183E-6	2.443E-5	0.3958	0.3958	0.3268	0.3268	0.09895	1.0
8	rs11204005	12895576	7.882E-6	6.841E-5	1.0	1.0	0.7604	0.7604	0.2446	1.0
9	rs16910850	94478347	1.216E-5	9.688E-5	1.0	1.0	0.8897	0.8897	0.2446	1.0
12	rs11957									1.0
17	rs720709									1.0
15	rs16971									1.0
20	rs157048									1.0
20	rs607470									1.0
17	rs99445									1.0
3	rs636000									1.0
9	rs17534									1.0
21	rs21788									1.0
11	rs89831									1.0
11	rs79311									1.0
11	rs12418									1.0
15	rs16971									1.0
19	rs38444									1.0
12	rs44457									1.0
10	rs12098									1.0
9	rs15482									1.0
19	rs72514									1.0
6	rs54426									1.0
8	rs10098									1.0
8	rs92983									1.0
11	rs12418739	112122631	4.43E-5	2.736E-4	1.0	1.0	0.9997	0.9997	0.2743	1.0

Stratified analysis

- Cochran-Mantel-Haenszel test



- Stratified 2×2×K tables

Single SNP genome screen, conditional on cluster solution

Cochran-Mantel-Haenszel test of association

```
plink --bfile example --mh --adjust
      --within pop.cov
```

Use `--within` to specify a categorical clustering (i.e. to condition on). The file `pop.cov` distinguishes Chinese from Japanese

- Will generate `plink.cmh` and `plink.cmh.adjusted`, mirroring the two files generated by the standard `--assoc` command

Overview of the **PLINK** software package

A simulated WGAS dataset

Summary statistics and quality control

Assessment of population stratification

Whole genome association screen

Further exploration of ‘hits’

Visualization and follow-up using Haploview

The Truth…

	Chinese	Japanese
Case	34	7
Control	11	38

Group difference

	“11”	“12”	“22”
Case	4	24	21
Control	17	20	4

Single common variant
rs11204005 chr8

Does rs11204005 pass the HWE test?

For a single SNP, create standard PED fileset

```
plink --bfile example --snp rs11204005 --recode
      --out hit
```

Will name files `hit.ped` and `hit.map`

```
plink --file hit --hardy
```

Loading a standard text-based PED file now so use `--file`, not `--bfile`

- Creates file `plink.hwe` containing single SNP HWE results

Does rs11204005 differ in frequency between the two populations?

```
plink --file hit --assoc --pheno pop.cov
```

Use an alternate phenotype – instead of disease status, the outcome variable for the case/control analysis is now Chinese versus Japanese subpopulation membership

- The file `pop.cov` is the same file that we used for the purposes of splitting the sample into the two groups for stratified anaylsis

Does rs2513514 differ in frequency between the two populations?

Selecting out a different single SNP from the original WGAS binary fileset

```
plink --bfile example --snp rs2513514
      --assoc --pheno pop.cov
```

rs2513514 is the most significant SNP prior to correction for population stratification

Does rs11204005 show different effects between the two populations?

Specify the Breslow-Day test for homogeneity of odds ratio as well as the Cochran-Mantel-Haenszel procedure

```
plink --file hit --mh --bd
      --within pop.cov
```

Estimates of the allele, genotype frequencies and odds ratio for rs11204005 ?

Generates simple association statistics for the single SNP, in `plink.assoc`

```
plink --file hit --assoc
```

- Allele frequencies in `plink.assoc`
- Genotypes counts are in `plink.hwe` (previously calculated)
- Odds ratio in `plink.cmh` (previously calculated)

Similar case/control genotyping rates for rs11204005?

```
plink --file hit –-test-missing
```

Test of phenotype / genotype failure association, in `plink.missing`

Additive, genotypic models for rs11204005?

Force genotype tests, irrespecitve of genotype counts

```
plink --file hit --model --cell 0
```

Genotypic tests, reported in `plink.model`

- Also includes the Cochran-Armitage trend test in `plink.model`

In summary

- We performed whole genome
  - summary statistics and QC
  - stratification analysis
  - conditional and unconditional association analysis

- We found a single SNP rs11204005 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)

Overview of the **PLINK** software package

A simulated WGAS dataset

Summary statistics and quality control

Assessment of population stratification

Whole genome association screen

Further exploration of ‘hits’

Visualization and follow-up using Haploview

Acknowledgements

Haploview development

Dave Bender
Julian Maller

Jeff Barrett
Mark Daly

PLINK development

Shaun Purcell

Lori Thomas
Kathe Todd-Brown

Mark Daly
Pak Sham