***PLINK**: a toolset for whole genome association analysis*

Shaun Purcell

shaun@pngu.mgh.harvard.edu

Center for Human Genetic Research, Massachusetts General Hospital

Broad Institute of Harvard & MIT

Challenges of whole-genome studies

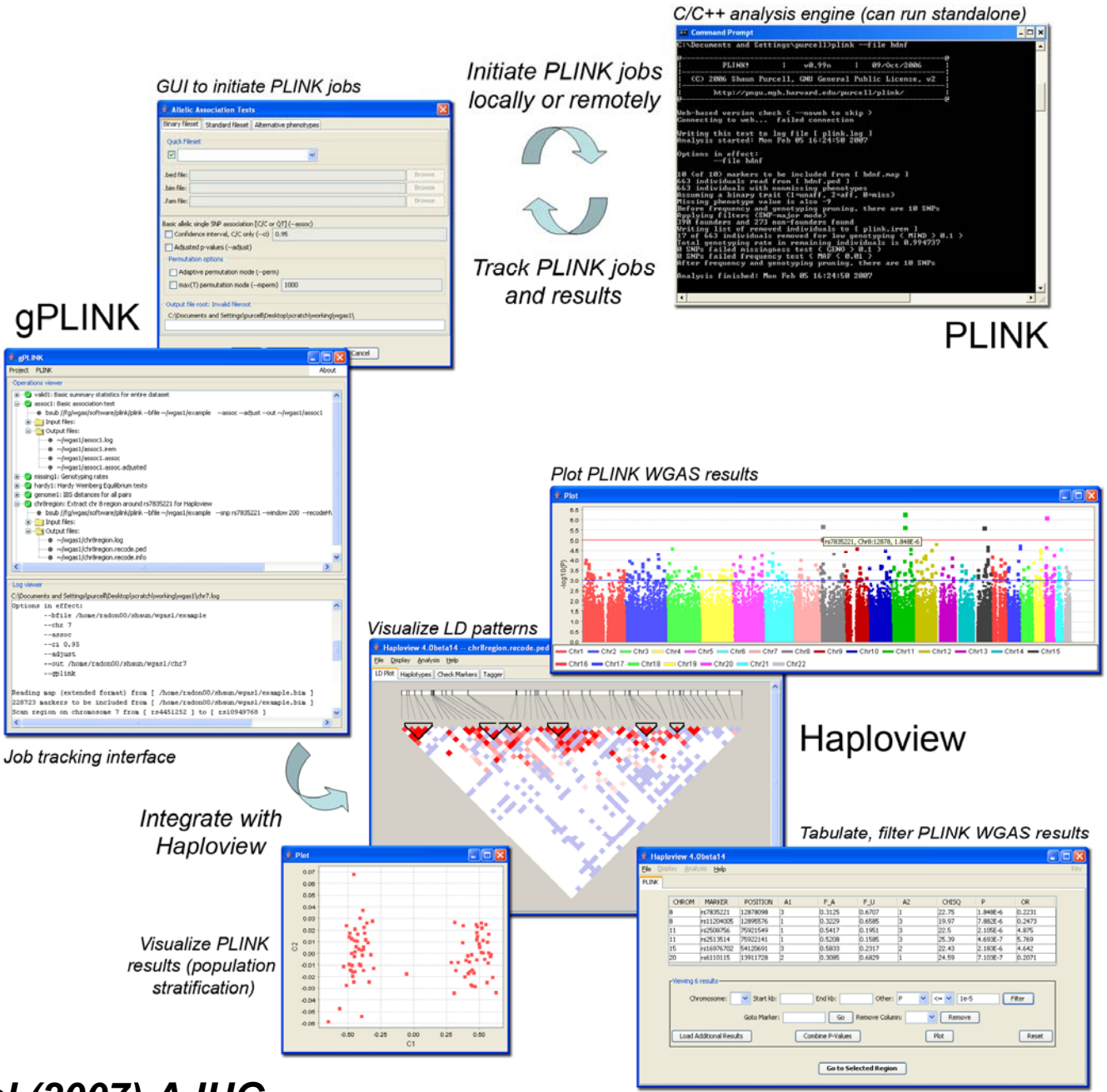
- most obviously, multiple testing burden
- computationally-intensive methods break

Opportunities

- simple methods can work well with ↑ data
- novel analyses permitted

PLINK

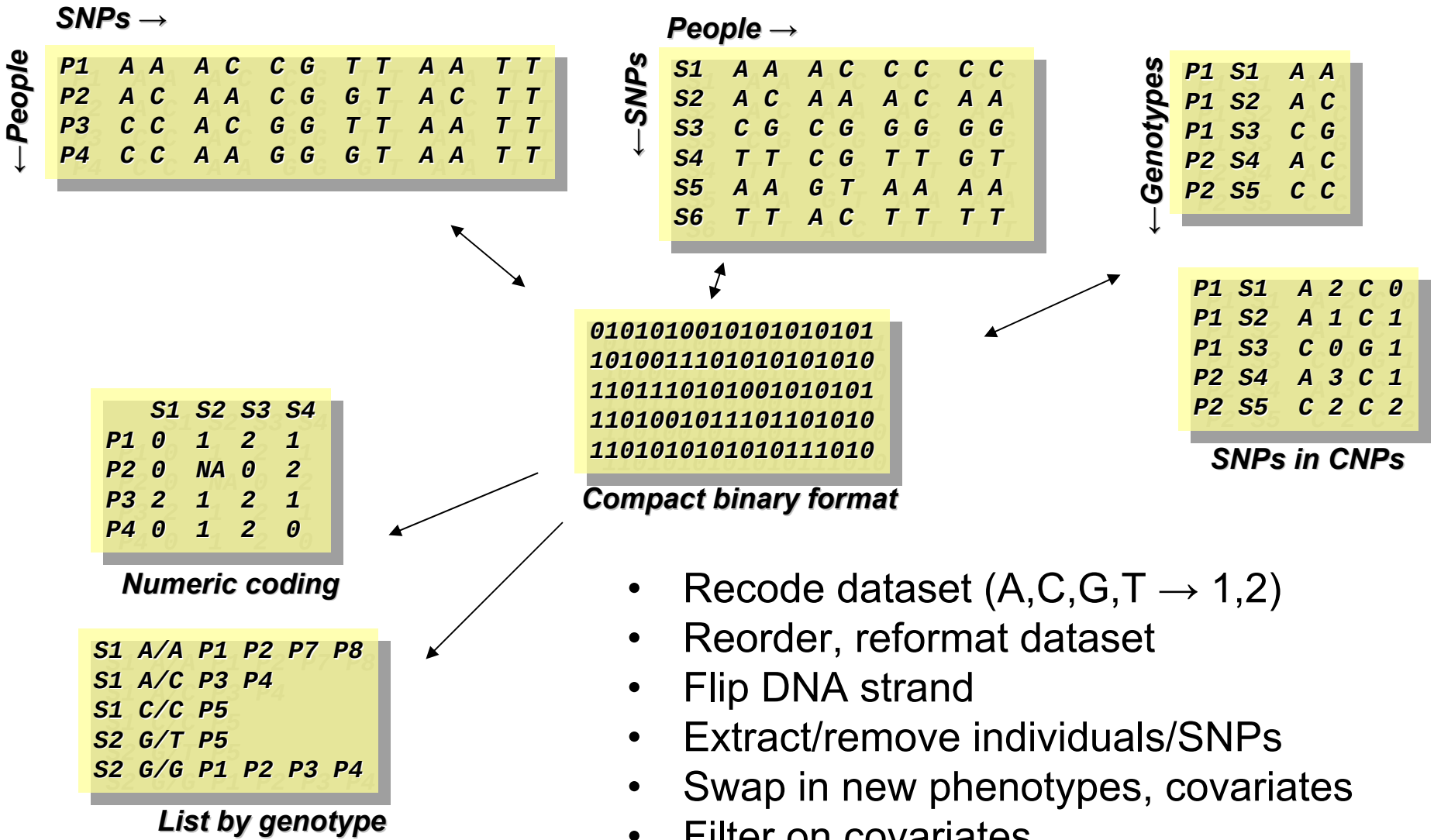
- introduce computational platform
- focus on basic data handling and testing
- “downstream analysis” (c.f. *Birdsuite*)



Purcell et al (2007) AJHG

- Data management
- Summary statistics
- Population stratification
- Association analysis
- Linkage disequilibrium and haplotype analysis
- Shared segment analysis
- Copy number analysis

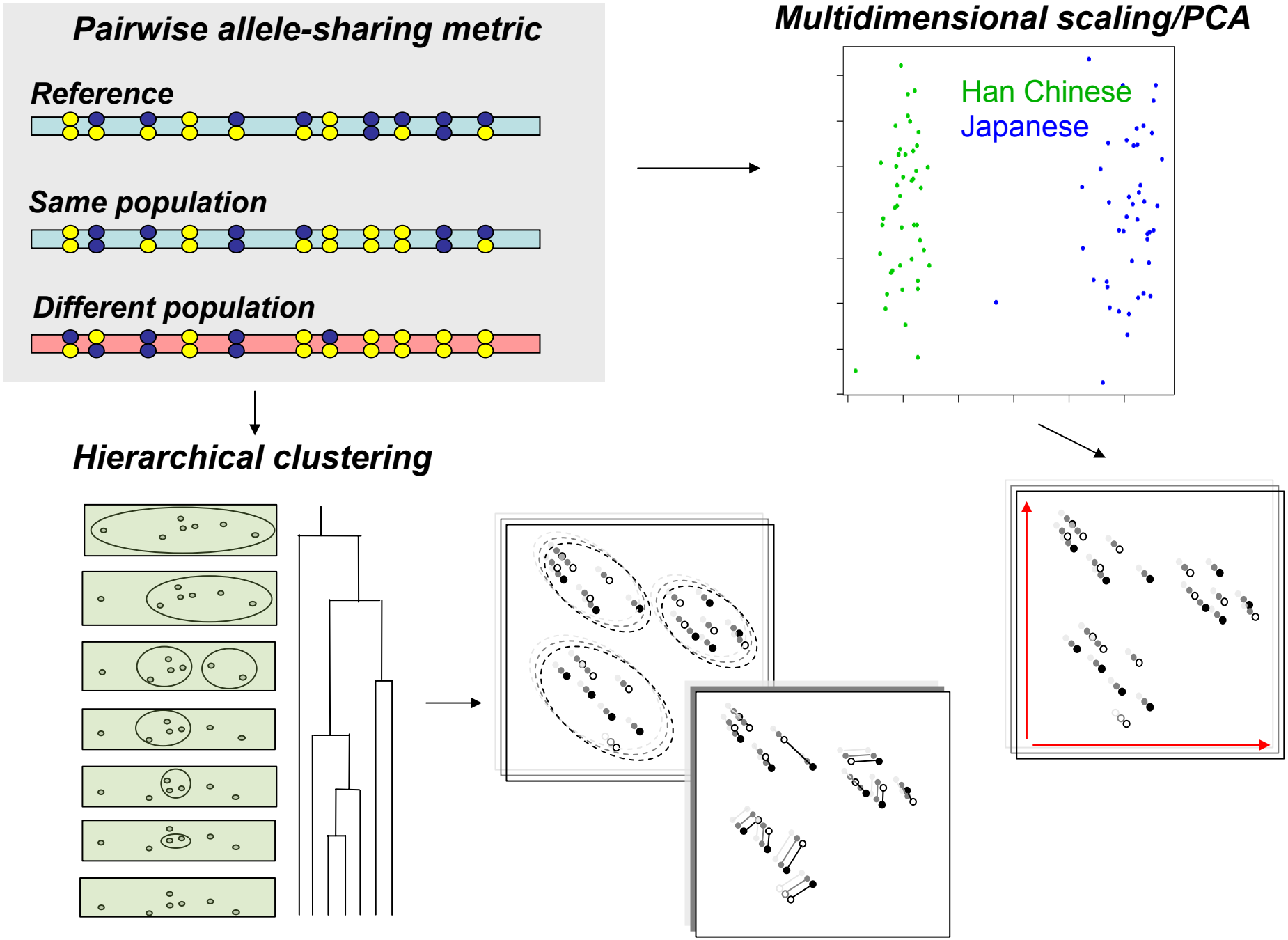
Data management



- Recode dataset (A,C,G,T → 1,2)
- Reorder, reformat dataset
- Flip DNA strand
- Extract/remove individuals/SNPs
- Swap in new phenotypes, covariates
- Filter on covariates
- Merge 2 or more filesets

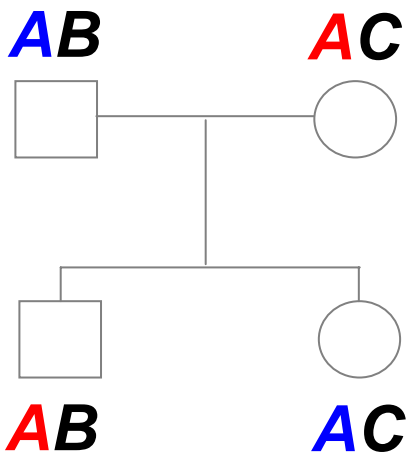
Summary statistics

- Filters and reports for standard metrics
  - Genotyping rate
  - Allele, genotype, haplotype frequencies
  - Hardy-Weinberg
  - Mendel errors
- Tests of non-random missingness
  - by phenotype and by (unobserved) genotype
- Individual homozygosity estimates
- Check/impute sex based on X chromosome
- LD-based detection of strand flips
  - A/T and C/G SNPs potentially ambiguous
- Automated search for plate effects
  - w/ subsequent masking of specific SNP/individual genotypes


Pairwise allele-sharing metric
Reference
Same population
Different population
Multidimensional scaling/PCA
Han Chinese
Japanese
Hierarchical clustering

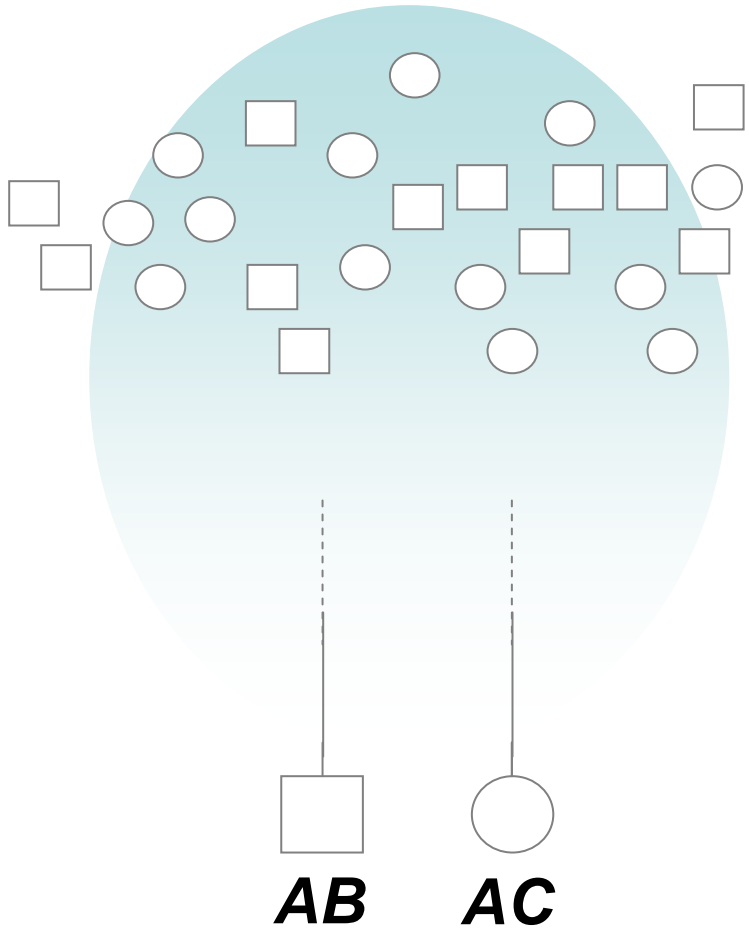
Estimation of IBD sharing (relatedness)

Parents



IBS = 1
IBD = 0

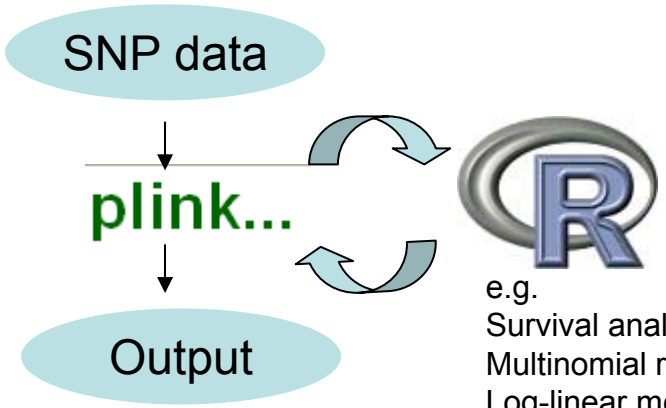
Most recent common ancestor from homogeneous random mating population



Association analysis

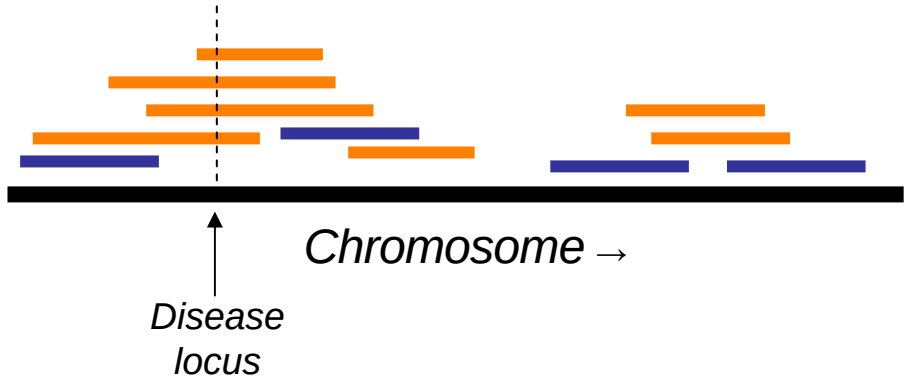
- Population-based
  - Allelic, trend, genotypic, Fisher's exact
  - Stratified tests (Cochran-Mantel-Haenszel, Breslow-Day)
  - Linear & logistic regression models
    - multiple covariates, interactions, joint tests, etc
- Family-based
  - Disease traits: TDT / sib-TDT
  - Continuous traits: QFAM (between/within model, QTDT)
- Permutation procedures
  - "adaptive", max(T), gene-dropping, between/within, rank-based, within-cluster
- Multilocus tests
  - Haplotype estimation, set-based tests, Hotelling's T^2, epistasis

R plugin support



e.g.
Survival analysis
Multinomial regression
Log-linear models

Runs of homozygosity, extended haplotype sharing



Simple SNP simulation functions

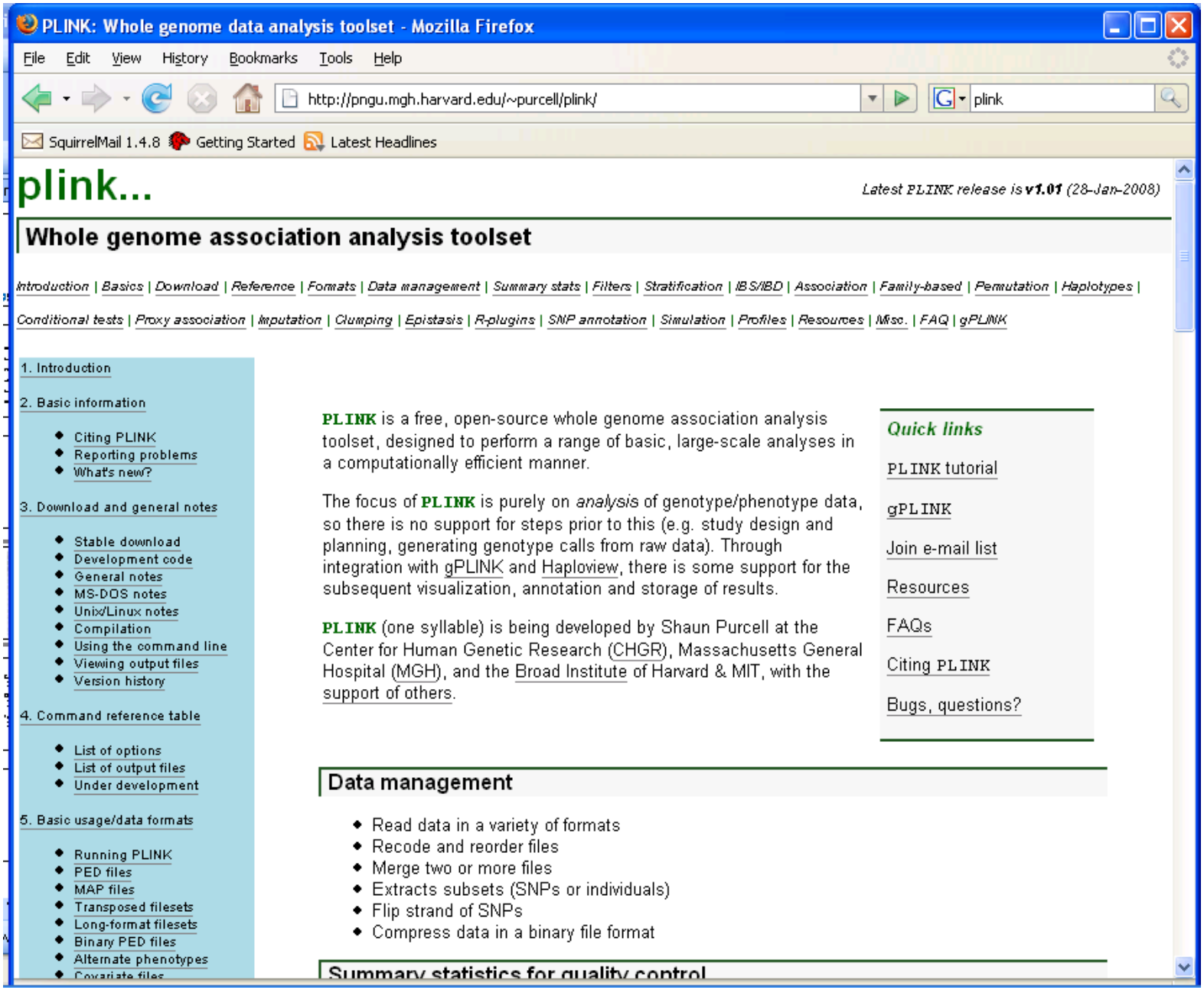
SIMULATE

P1	A	A	A	C	C	G	T	T
P2	A	C	A	A	C	G	G	T
P3	C	C	A	C	G	G	T	T
P4	C	C	A	A	G	G	G	T

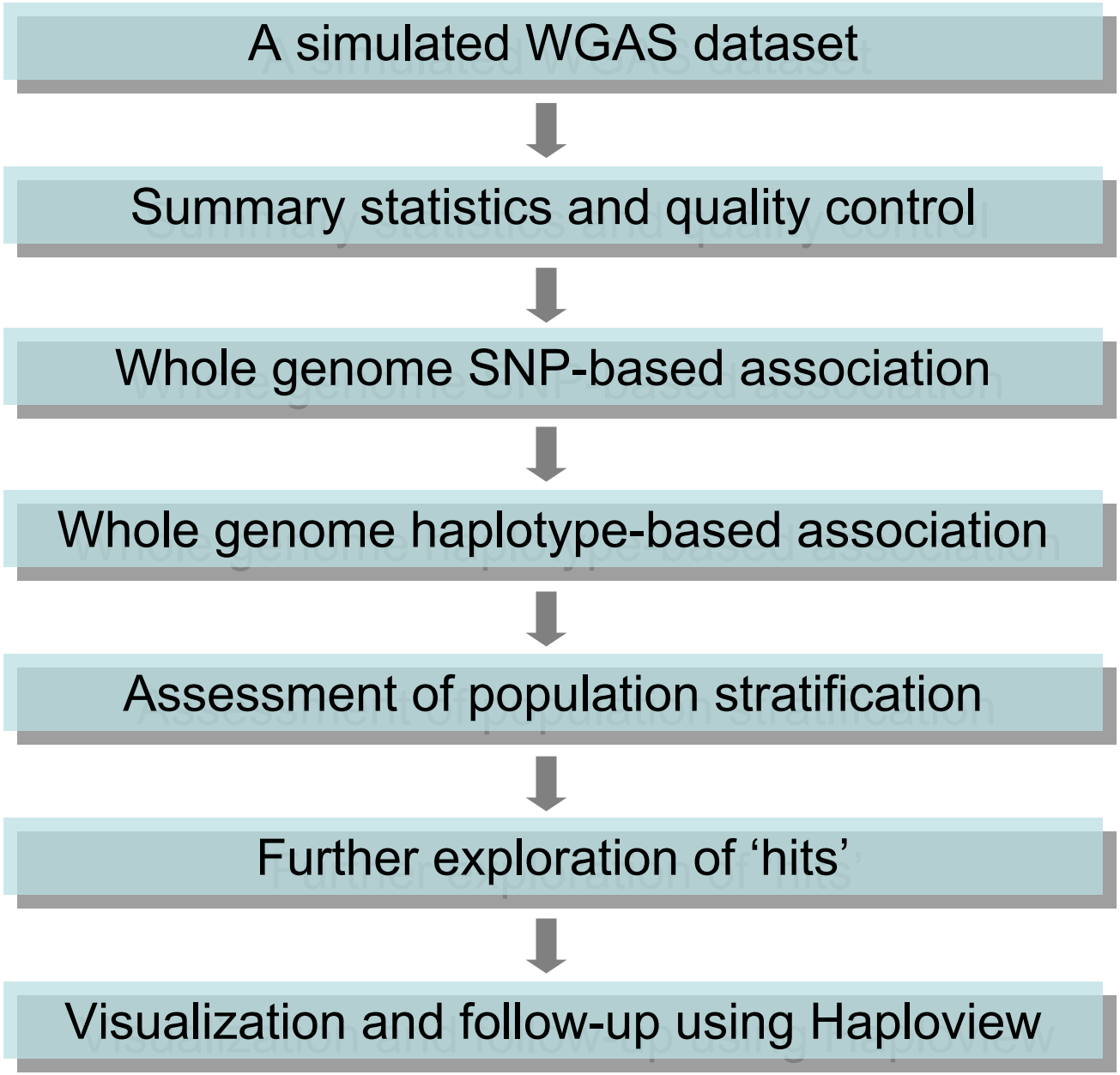
plink --lookup rs222162

```
PLINK-SNP (WGAS SNP annotation courtesy of Patrick Sullivan)
Connecting to web...

SNP ID                          : rs222162
Affy 5.0                        : no
Affy 6.0                        : no
Perlegen ID                     : 43887
Perlgen 600                     : yes
Illumina 650                    : yes
Non-syn SNP                     : no
SNP Error                       : no
Chromosome                      : 21
Strand                          : +
Position (bp)                   : 26755483
Pseudo-autosomal region?        : N/A
NCBI reference allele           : G
Human alleles                   : A/G
Chimp allele                    : A
HapMap CEU MAF                  : 0.075
HapMap ASI MAF                  : 0
HapMap YRI MAF                  : 0.5
In gene transcript              : AK125338
Nearby Genes(KB distance)       : A4(-290) CYYR1 (4) ATS1 (374) ATS5 (460)
Segmental duplication?          : no
Conservation >95% pctile?       : no
miRNA target? (TargetScan)      : no
Regulatory potential?           : yes
Promotor region? (Stanford)     : no
Transfactor binding site        : no
Enhancer?                       : no
Exon?                           : no
Consensus splice site?          : no
5' UTR?                         : no
3' UTR?                         : no
-----------------------------------------------
```



Documentation also available as PDF (>200 pages)


A simulated WGAS dataset
Summary statistics and quality control
Whole genome SNP-based association
Whole genome haplotype-based association
Assessment of population stratification
Further exploration of 'hits'
Visualization and follow-up using Haploview

Acknowledgements

- **PLINK development**
  - Kathe Todd-Brown
  - Douglas Ruderfer
  - Lori Thomas
  - Manuel Ferreira
  - Pak Sham
- **ENDGAME (NIH)**
- **Broad Institute Medical & Population Genetics Program**
  - Julian Maller
  - Dave Bender
  - Ben Neale
  - Andrew Kirby
  - Paul de Bakker
  - Itsik Pe'er
  - Ben Voight
  - David Altshuler
  - Pamela Sklar
  - Mark Daly

BROAD INSTITUTE

MGH 1811

CNGR

PLINK practical 1

Data cleaning and
association testing

- In this practical we will analyse a simulated dataset using PLINK
- 15 single nucleotide polymorphisms, in a candidate gene region spanning ~30kb
- Case/control design: 1000 cases and 1000 controls
- Specifically, we will:
  - examine the format of the raw data (PED and MAP files)
  - perform an initial association analysis for each SNP
  - perform basic QC steps, including tests for HWE and looking at genotyping rate statistics
  - repeat the association analysis
  - consider genotypic as well as allelic tests
  - perform sex-specific tests
  - perform conditional and haplotypic tests

Preliminary association analysis (allelic)

```
./plink --file mygene
        --assoc
```

Output file plink.assoc

CHR	SNP	BP	A1	F_A	F_U	A2	CHISQ	P	OR
1	rs00001	0	G	0.13	0.1524	C	4.015	0.04509	0.8314
1	rs00002	2013	C	0.1489	0.1459	A	0.07039	0.7908	1.024
1	rs00003	4367	A	0.4612	0.5225	T	11.92	0.0005542	0.7825
1	rs00004	6473	G	0.4286	0.4279	C	0.002133	0.9632	1.003
1	rs00005	8887	C	0.454	0.4636	T	0.3656	0.5454	0.9619
1	rs00006	11054	T	0.1725	0.1619	A	0.7824	0.3764	1.079
1	rs00007	13413	G	0.03846	0.04239	A	0.3887	0.533	0.9036
1	rs00008	15820	T	0.139	0.1014	G	13.14	0.0002883	1.43
1	rs00009	18125	T	0.2391	0.2024	C	7.681	0.005582	1.238
1	rs00010	20253	A	0.423	0.4117	C	0.5109	0.4748	1.048
1	rs00011	22633	C	0.1925	0.1926	G	8.647e-05	0.9926	0.9992
1	rs00012	24739	C	0.152	0.1326	T	2.995	0.08353	1.172
1	rs00013	26762	G	0.2829	0.5571	A	302.3	1.041e-67	0.3136
1	rs00014	28833	A	0.1985	0.2161	G	1.856	0.1731	0.8982
1	rs00015	30974	A	0.2071	0.225	C	1.849	0.1739	0.8996

Allele frequencies

```
./plink --file mygene
        --freq
```

Output file plink.frq

CHR	SNP	A1	A2	MAF	NCHROBS
1	rs00001	G	C	0.1412	3902
1	rs00002	C	A	0.1474	3914
1	rs00003	A	T	0.486	3296
1	rs00004	G	C	0.4283	3918
1	rs00005	C	T	0.4588	3908
1	rs00006	T	A	0.1672	3924
1	rs00007	G	A	0.04043	3908
1	rs00008	T	G	0.1202	3936
1	rs00009	T	C	0.2208	3936
1	rs00010	A	C	0.4174	3896
1	rs00011	C	G	0.1925	3932
1	rs00012	C	T	0.1423	3892
1	rs00013	G	A	0.4201	3918
1	rs00014	A	G	0.2073	3922
1	rs00015	A	C	0.2161	3906

Genotyping rate per individual and per marker

```
./plink --file mygene
        --missing
```

Per-individual genotyping/missing rate, plink.imiss

FID	IID	MISS_PHENO	N_MISS	N_GENO	F_MISS
per0	per0	N	0	15	0
per1	per1	N	0	15	0
per2	per2	N	0	15	0
per3	per3	N	0	15	0
per4	per4	N	0	15	0
per5	per5	N	1	15	0.06667
per6	per6	N	0	15	0
per7	per7	N	2	15	0.1333
...	...				

Per-marker (locus) genotyping/ missing rate, plink.lmiss

CHR	SNP	N_MISS	N_GENO	F_MISS
1	rs00001	49	2000	0.0245
1	rs00002	43	2000	0.0215
1	rs00003	352	2000	0.176
1	rs00004	41	2000	0.0205
1	rs00005	46	2000	0.023
1	rs00006	38	2000	0.019
1	rs00007	46	2000	0.023
1	rs00008	32	2000	0.016
1	rs00009	32	2000	0.016
1	rs00010	52	2000	0.026
1	rs00011	34	2000	0.017
1	rs00012	54	2000	0.027
1	rs00013	41	2000	0.0205
1	rs00014	39	2000	0.0195
1	rs00015	47	2000	0.0235

Check for Hardy-Weinberg disequilibrium

```
./plink --file mygene
        --hardy
```

Output file plink.frq

CHR	SNP	TEST	A1	A2	GENO	O(HET)	E(HET)	P
1	rs00001	ALL	G	C	44/463/1444	0.2373	0.2425	0.3501
1	rs00001	AFF	G	C	17/219/737	0.2251	0.2262	0.8871
1	rs00001	UNAFF	G	C	27/244/707	0.2495	0.2583	0.2674
1	rs00002	ALL	C	A	34/509/1414	0.2601	0.2514	0.1495
1	rs00002	AFF	C	A	17/257/703	0.2631	0.2535	0.3114
1	rs00002	UNAFF	C	A	17/252/711	0.2571	0.2493	0.3707
1	rs00003	ALL	A	T	415/772/461	0.4684	0.4996	0.0119
1	rs00003	AFF	A	T	215/474/291	0.4837	0.497	0.4038
1	rs00003	UNAFF	A	T	200/298/170	0.4461	0.499	0.006616
1	rs00004	ALL	G	C	363/952/644	0.486	0.4897	0.7467
1	rs00004	AFF	G	C	178/485/318	0.4944	0.4898	0.7945
1	rs00004	UNAFF	G	C	185/467/326	0.4775	0.4896	0.4339
...	...							
1	rs00013	ALL	G	A	567/512/880	0.2614	0.4872	5.262e-96
1	rs00013	AFF	G	A	277/0/702	0	0.4058	7.409e-254
1	rs00013	UNAFF	G	A	290/512/178	0.5224	0.4935	0.07019

Check for differential genotyping rate (case vs. control)

```
./plink --file mygene
        --test-missing
```

Output file plink.missing

CHR	SNP	F_MISS_A	F_MISS_U	P
1	rs00001	0.027	0.022	0.5633
1	rs00002	0.023	0.02	0.7582
1	rs00003	0.02	0.332	6.43e-87
1	rs00004	0.019	0.022	0.7527
1	rs00005	0.022	0.024	0.8816
1	rs00006	0.023	0.015	0.2513
1	rs00007	0.025	0.021	0.655
1	rs00008	0.018	0.014	0.5936
1	rs00009	0.015	0.017	0.8589
1	rs00010	0.026	0.026	1
1	rs00011	0.018	0.016	0.863
1	rs00012	0.023	0.031	0.3343
1	rs00013	0.021	0.02	1
1	rs00014	0.02	0.019	1
1	rs00015	0.027	0.02	0.376

Remove bad SNPs

```
echo "rs00003" > bad.snps
echo "rs00013" >> bad.snps

./plink --file mygene
        --exclude bad.snps
        --recode
        --out cleaned
```

Equivalent command, using filters

```
./plink --file mygene
        --hwe 1e-3
        --hwe-all
        --geno 0.1
        --recode
        --out cleaned
```

Normally, HWE filters on controls; the* --hwe-all *flag implies all individuals

Re-run association analysis (allelic)

```
./plink --file cleaned
        --assoc
```

Output file plink.assoc

CHR	SNP	BP	A1	F_A	F_U	A2	CHISQ	P	OR
1	rs00001	0	G	0.13	0.1524	C	4.015	0.04509	0.8314
1	rs00002	2013	C	0.1489	0.1459	A	0.07039	0.7908	1.024
1	rs00004	6473	G	0.4286	0.4279	C	0.002133	0.9632	1.003
1	rs00005	8887	C	0.454	0.4636	T	0.3656	0.5454	0.9619
1	rs00006	11054	T	0.1725	0.1619	A	0.7824	0.3764	1.079
1	rs00007	13413	G	0.03846	0.04239	A	0.3887	0.533	0.9036
1	rs00008	15820	T	0.139	0.1014	G	13.14	0.0002883	1.43
1	rs00009	18125	T	0.2391	0.2024	C	7.681	0.005582	1.238
1	rs00010	20253	A	0.423	0.4117	C	0.5109	0.4748	1.048
1	rs00011	22633	C	0.1925	0.1926	G	8.647e-05	0.9926	0.9992
1	rs00012	24739	C	0.152	0.1326	T	2.995	0.08353	1.172
1	rs00014	28833	A	0.1985	0.2161	G	1.856	0.1731	0.8982
1	rs00015	30974	A	0.2071	0.225	C	1.849	0.1739	0.8996

Corrections for multiple testing

```
./plink --file cleaned
        --assoc
        --adjust
        --mperm 10000
```

Output file plink.assoc.adjust: rs00008 is significant after Bonferroni correction

CHR	SNP	UNADJ	GC	BONF	HOLM	SIDAK_SS	SIDAK_SD	FDR_BH	FDR_BY
1	rs00008	0.0002883	0.005643	0.003748	0.003748	0.003742	0.003742	0.003748	0.01192
1	rs00009	0.005582	0.03437	0.07256	0.06698	0.07018	0.06496	0.03628	0.1154
1	rs00001	0.04509	0.1261	0.5862	0.496	0.4511	0.398	0.1954	0.6214
1	rs00012	0.08353	0.1864	1	0.8353	0.6782	0.582	0.2715	0.8633
1	rs00014	0.1731	0.2983	1	1	0.9155	0.8192	0.3768	1
1	rs00015	0.1739	0.2992	1	1	0.9166	0.8192	0.3768	1
1	rs00006	0.3764	0.4995	1	1	0.9978	0.9633	0.699	1
1	rs00010	0.4748	0.5853	1	1	0.9998	0.979	0.709	1
1	rs00007	0.533	0.6341	1	1	0.9999	0.979	0.709	1
1	rs00005	0.5454	0.6444	1	1	1	0.979	0.709	1
1	rs00002	0.7908	0.8395	1	1	1	0.9908	0.9345	1
1	rs00004	0.9632	0.9719	1	1	1	0.9986	0.9926	1
1	rs00011	0.9926	0.9943	1	1	1	0.9986	0.9926	1

Corrections for multiple testing

```
./plink --file cleaned
        --assoc
        --adjust
        --mperm 10000
```

Empirical p-values in plink.assoc.mperm: rs00008 is experiment-wide significant

CHR	SNP	STAT	EMP1	EMP2
1	rs00001	4.015	0.0489	0.4197
1	rs00002	0.07039	0.7965	1
1	rs00004	0.002133	0.9682	1
1	rs00005	0.3656	0.5391	0.9999
1	rs00006	0.7824	0.3678	0.9965
1	rs00007	0.3887	0.5423	0.9998
1	rs00008	13.14	0.0003	0.0031
1	rs00009	7.681	0.006699	0.06269
1	rs00010	0.5109	0.4799	0.9997
1	rs00011	8.647e-05	0.9961	1
1	rs00012	2.995	0.07449	0.6378
1	rs00014	1.856	0.1748	0.8973
1	rs00015	1.849	0.1724	0.8981

Determine pattern of linkage disequilibrium in the region

```
./plink --file cleaned
        --r2
```

Calculates pairwise LD (r^2) between all SNPs;
by default, only output only pairs with $r^2 > 0.2$,
to file plink.ld

CHR_A	BP_A	SNP_A	CHR_B	BP_B	SNP_B	R2
1	15820	rs00008	1	18125	rs00009	0.482748
1	28833	rs00014	1	30974	rs00015	0.948877

Genotypic tests of association at rs00008

```
./plink --file cleaned
        --snp rs00008
        --model
```

Calculates allelic, trend, genotypic, dominant and recessive tests: plink.model

CHR	SNP	A1	A2	TEST	AFF	UNAFF	CHISQ	DF	P
1	rs00008	T	G	GENO	19/235/728	14/172/800	13.89	2	0.0009615
1	rs00008	T	G	TREND	273/1691	200/1772	12.86	1	0.0003354
1	rs00008	T	G	ALLELIC	273/1691	200/1772	13.14	1	0.0002883
1	rs00008	T	G	DOM	254/728	186/800	13.89	1	0.0001934
1	rs00008	T	G	REC	19/963	14/972	0.7913	1	0.3737

Trend test uses Cochran-Armitage test

Genotypic tests of association at rs00008, using logistic regression

```
./plink --file cleaned
        --snp rs00008
        --logistic
        --genotypic
```

Reports results in plink.assoc.logistic

CHR	SNP	BP	A1	TEST	NMISS	OR	STAT	P
1	rs00008	15820	T	ADD	1968	1.221	1.123	0.2615
1	rs00008	15820	T	DOMDEV	1968	1.229	1.011	0.312
1	rs00008	15820	T	GENO_2DF	1968	NA	13.8	0.001007

Fits a single model logit(P) ~ A + D + e

Reports three tests:

$H_0 : A = 0$	*ADD (1df)*
$H_0 : D = 0$	*DOMDEV (1df)*
$H_0 : A = D = 0$	*GENO (2df)*

Coding of genotypes:

	ADD (A)	*DOMDEV (D)*
GG	*0*	*0*
TG	*1*	*1*
TT	*2*	*0*

Genotypic tests of association at rs00008, alternate parameterization

```
./plink --file cleaned
        --snp rs00008
        --logistic
        --genotypic
        --hethom
```

Reports results in plink.assoc.logistic

CHR	SNP	BP	A1	TEST	NMISS	OR	STAT	P
1	rs00008	15820	T	HOM	1968	1.491	1.123	0.2615
1	rs00008	15820	T	HET	1968	1.501	3.607	0.0003095
1	rs00008	15820	T	GENO_2DF	1968	NA	13.8	0.001007

Fits a single model logit(P) ~ TT + TG + e

Reports three tests:

$H_0 : TT = 0$	*HOM (1df)*
$H_0 : TG = 0$	*HET (1df)*
$H_0 : TT = TG = 0$	*GENO (2df)*

Coding of genotypes:

	HOM	*HET*
GG	*0*	*0*
TG	*0*	*1*
TT	*1*	*0*

Test for sex-specific effects, e.g. a male-only analysis

```
./plink --file cleaned
        --filter-males
        --assoc
```

Or formally test SNP-by-sex interaction, using logistic model

```
./plink --file cleaned
        --snp rs00008
        --logistic
        --interaction
        --sex
```

Reports results in plink.assoc.logistic

CHR	SNP	BP	A1	TEST	NMISS	OR	STAT	P
1	rs00008	15820	T	ADD	1968	1.321	2.087	0.03685
1	rs00008	15820	T	SEX	1968	0.9723	-0.2762	0.7824
1	rs00008	15820	T	ADDxSEX	1968	1.174	0.8119	0.4169

Both rs00008 and rs00009 are associated P<0.01 and are also in moderately high LD with each other. Are these two associations independent?

```
./plink --file cleaned
        --logistic
        --condition rs00008
```

Includes genotype at rs00008 as a covariate; results in plink.assoc.logistic

CHR	SNP	BP	A1	TEST	NMISS	OR	STAT	P
1	rs00001	0	G	ADD	1919	0.8403	-1.89	0.05878
1	rs00001	0	G	rs00008	1919	1.436	3.621	0.0002939
1	rs00002	2013	C	ADD	1927	1.013	0.1432	0.8862
1	rs00002	2013	C	rs00008	1927	1.433	3.616	0.0002996
...	...							

Often desirable to extract out only the terms for the SNP (ADD)

```
fgrep -w ADD plink.assoc.logistic
```

1	rs00001	0	G	ADD	1919	0.8403	-1.89	0.05878
1	rs00002	2013	C	ADD	1927	1.013	0.1432	0.8862
1	rs00004	6473	G	ADD	1927	1.015	0.2337	0.8152
1	rs00005	8887	C	ADD	1923	0.9542	-0.7148	0.4747
1	rs00006	11054	T	ADD	1930	1.092	1.012	0.3114
1	rs00007	13413	G	ADD	1922	0.8844	-0.7346	0.4626
1	rs00008	15820	T	ADD	1968	NA	NA	NA
1	rs00009	18125	T	ADD	1936	1.051	0.4677	0.64
1	rs00010	20253	A	ADD	1917	1.054	0.8027	0.4221
1	rs00011	22633	C	ADD	1934	0.9934	-0.08065	0.9357
1	rs00012	24739	C	ADD	1916	1.182	1.793	0.07298
1	rs00014	28833	A	ADD	1929	0.8983	-1.33	0.1834
1	rs00015	30974	A	ADD	1921	0.8954	-1.386	0.1656

```
./plink --file cleaned
        --logistic
        --condition rs00009
```

Includes genotype at rs00009 as a covariate; results in plink.assoc.logistic

CHR	SNP	BP	A1	TEST	NMISS	OR	STAT	P
1	rs00001	0	G	ADD	1919	0.8541	-1.707	0.08775
1	rs00002	2013	C	ADD	1926	1.068	0.7016	0.4829
1	rs00004	6473	G	ADD	1929	1.01	0.1552	0.8767
1	rs00005	8887	C	ADD	1923	0.9601	-0.6208	0.5347
1	rs00006	11054	T	ADD	1930	1.065	0.7292	0.4659
1	rs00007	13413	G	ADD	1923	0.8862	-0.728	0.4666
1	rs00008	15820	T	ADD	1936	1.349	2.19	0.02854
1	rs00009	18125	T	ADD	1968	NA	NA	NA
1	rs00010	20253	A	ADD	1917	1.041	0.6117	0.5407
1	rs00011	22633	C	ADD	1934	0.9656	-0.4269	0.6695
1	rs00012	24739	C	ADD	1914	1.171	1.693	0.0905
1	rs00014	28833	A	ADD	1929	0.8848	-1.522	0.128
1	rs00015	30974	A	ADD	1921	0.8867	-1.513	0.1304

Given these are in high LD, often useful to explicitly model the haplotypic associations instead

```
./plink --file cleaned
        --chap
        --hap-snps rs00008,rs00009
```

```
+++ PLINK conditional haplotype test results +++

2 SNPs, and 3 common haplotypes ( MHF >= 0.01 ) from 4 possible

 CHR           BP          SNP   A1   A2          F
   1        15820      rs00008    T    G     0.1202
   1        18125      rs00009    T    C     0.2208

Haplogrouping: each {set} allowed a unique effect
Alternate model
   { TT }  { GT }  { GC }
Null model
   { TT, GT, GC }

  HAPLO         FREQ        OR(A)        OR(N)
-------       ------      -------      -------
     TT       0.1206      (-ref-)      (-ref-)
     GT      0.09969       0.7412         |
     GC       0.7797        0.705         |
-------       ------      -------      -------

Model comparison test statistics:

                          Alternate       Null
                -2LL :         2671       2684

  Likelihood ratio test: chi-square = 12.44
                         df = 2
                         p = 0.001992
```

The --chap command means "conditional haplotype tests". Output is written to plink.chap

Test rs00008 against haplotypic background

```
./plink --file cleaned
        --chap
        --hap-snps rs00008,rs00009
        --independent-effect rs00008
```

```
Haplogrouping: each {set} allowed a unique effect
Alternate model
   { TT }  { GT }  { GC }
Null model
   { TT, GT }  { GC }

  HAPLO         FREQ        OR(A)        OR(N)
-------       ------      -------      -------
     TT       0.1206      (-ref-)      (-ref-)
     GT      0.09969       0.7412         |

     GC       0.7797        0.705       0.8086
-------       ------      -------      -------

Model comparison test statistics:

                          Alternate       Null
                -2LL :         2671       2676

  Likelihood ratio test: chi-square = 4.815
                         df = 1
                         p = 0.02821
```

Test rs00009 against haplotypic background

```
./plink --file cleaned
        --chap
        --hap-snps rs00008,rs00009
        --independent-effect rs00009
```

```
Alternate model
   { TT }  { GT }  { GC }
Null model
   { TT }  { GT, GC }

  HAPLO         FREQ        OR(A)        OR(N)
-------       ------      -------      -------
     TT       0.1206      (-ref-)      (-ref-)

     GT      0.09969       0.7412       0.7092
     GC       0.7797        0.705          |
-------       ------      -------      -------

Model comparison test statistics:

                          Alternate       Null
                -2LL :         2671       2672

  Likelihood ratio test: chi-square = 0.2187
                         df = 1
                         p = 0.64
```

Output in Haploview-friendly format, to confirm LD structure

```
./plink --file cleaned
        --recodeHV
```

Produces two files that can be loaded into Haploview: plink.ped

```
per0 per0 0 0 1 2 C C A A C C C T A A A A G G T C A A G G C C G G C C
per1 per1 0 0 1 2 C C C A G C T T A A A A T G T C A C G G T T G G C C
per2 per2 0 0 2 2 G G C A C C C C A A A A G G C C A C C G T T G G C C
per3 per3 0 0 1 2 C C A A G G C T T A A A G G C C A C G G C C G G C C
per4 per4 0 0 2 2 C C A A C C C C T A A A G G C C A C C G T T G G C C
per5 per5 0 0 1 2 C C A A G G T T A A A A T G 0 0 A C C G T T A G A C
per6 per6 0 0 1 2 C C A A G C C C A A G A T G T C A C C G C T G G C C
```

and plink.info

```
rs00001 0
rs00002 2013
rs00004 6473
rs00005 8887
rs00006 11054
rs00007 13413
rs00008 15820
rs00009 18125
rs00010 20253
rs00011 22633
rs00012 24739
rs00014 28833
rs00015 30974
```

Output in R-friendly format, to confirm SNP-by-sex analysis

```
./plink --file cleaned
        --recodeA
```

Produces a single file that can be loaded into R: plink.raw

```
FID IID PAT MAT SEX PHENOTYPE rs00008_T
per0 per0 0 0 1 2 0
per1 per1 0 0 1 2 1
per2 per2 0 0 2 2 0
per3 per3 0 0 1 2 0
per4 per4 0 0 2 2 0
per5 per5 0 0 1 2 1
per6 per6 0 0 1 2 1
per7 per7 0 0 2 2 NA
...  ...
```

```
R
d <- read.table("plink.raw",header=T)
str(d)
```

Output in R-friendly format, to confirm SNP-by-sex analysis

```
summary(glm( PHENOTYPE-1 ~ rs00008_T * SEX ,
             data=d , family="binomial" ) )
```

Same result for interaction test (P = 0.4169) as PLINK

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.1308     0.1627  -0.804   0.4212
rs00008_T       0.5999     0.3220   1.863   0.0624 .
SEX             0.0281     0.1017   0.276   0.7824
rs00008_T:SEX  -0.1608     0.1981  -0.812   0.4169
```

Note effect arbitrary coding of sex term (M/F): in PLINK (0/1) versus (1/2) here

```
summary(glm( PHENOTYPE-1 ~ rs00008_T * I(SEX==1),
             data=d , family="binomial" ) )
```

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -0.07465    0.07056  -1.058   0.2901
rs00008_T                  0.27827    0.13331   2.087   0.0369 *
I(SEX == 1)TRUE           -0.02810    0.10174  -0.276   0.7824
rs00008_T:I(SEX == 1)TRUE  0.16084    0.19811   0.812   0.4169
```

In summary

- We have performed basic QC and association analysis on a candidate gene case/control dataset
- The SNP rs00008 showed a significant association (P=3x10^{-4}) and was significant after correction for multiple testing by permutation (P=0.003)
- The T (versus G) allele has a 12% sample frequency and an allelic odds ratio of 1.43
- An additive model fits the data well versus a 2df genotypic model (P=0.31 for genotypic vs allelic)
- There is no indication of sex-specific effects (P=0.42)
- Haplotype-based tests shows the weaker association at nearby rs00009 does not represent an independent signal

Questions?