GWA Study considerations - family-based vs. case/control

Mark Daly, PhD

Assistant Professor of Medicine
Massachusetts General Hospital/Harvard Medical School

Senior Associate Member
Broad Institute of Harvard and MIT

Issues

- Design considerations
- Data QC considerations
- Analysis considerations
- Combining results from mixed designs

Traditional debate

- Power versus robustness
  - Trio/family design robust to population substructure concerns
  - Requires more samples typed to achieve equivalent power

Neither of these are absolutes…

- Widely tested approaches to evaluating and correcting for substructure are now considered robust…

- Analytic approaches which can utilize parental phenotype data can reduce the power difference among scenarios…

Major challenge for case-control studies is acquiring a suitable control sample…

- Ideally:
  - Perfectly comparable population
  - Identical DNA quality, preparation
  - Random or interleaved evaluation of cases and controls in lab

With strict QC of SNPs, it is possible to achieve a "clean", uninflated case-control study. However, this ideal situation is not always achieved.

Concerns with borrowing controls

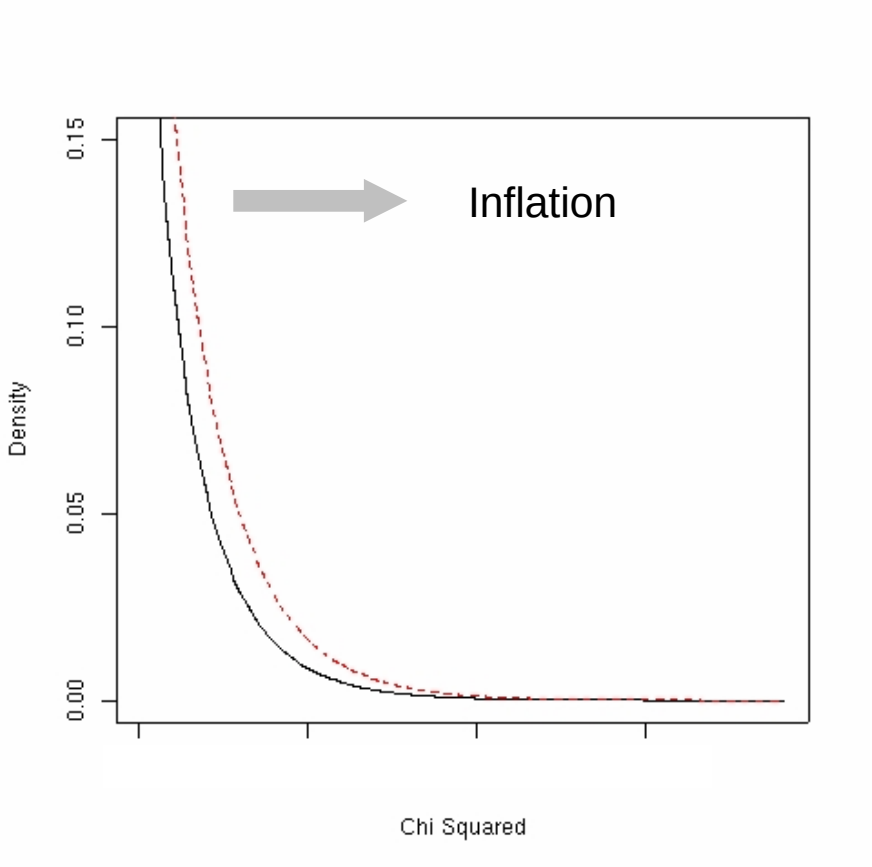
- Data quality and batch/lab effects
  - phenotype and batch correlate if controls taken from another study
- Population Differences
  - Stratification may still be present even when coarse self-reported matching done
- Phenotyping
  - Might some controls from another study be affected?
- Sample relatedness

Example 1

- Affy 100K data
- Extensively phenotyped disease cohort, followed for many years (BRASS)
- No control samples available

- Possible solution: Borrow controls from another study (FHS)

Effects of study mismatch



Capture inflation as median shift and excess hits in tail:

λ_{GC} = obs(median)/exp(median)
if >1, suggests overdispersion

P_{tail} = excess SNPs $p < 0.001$
if >1, implies more significant p-values than expected

Data quality dominates tail

- 83K SNPs
  - <10% missing data; >5% MAF

$\lambda_{GC} = 1.2$ $P_{tail} = 2.7$ (227)

- 43K SNPs
  - <1% missing data; >5% MAF

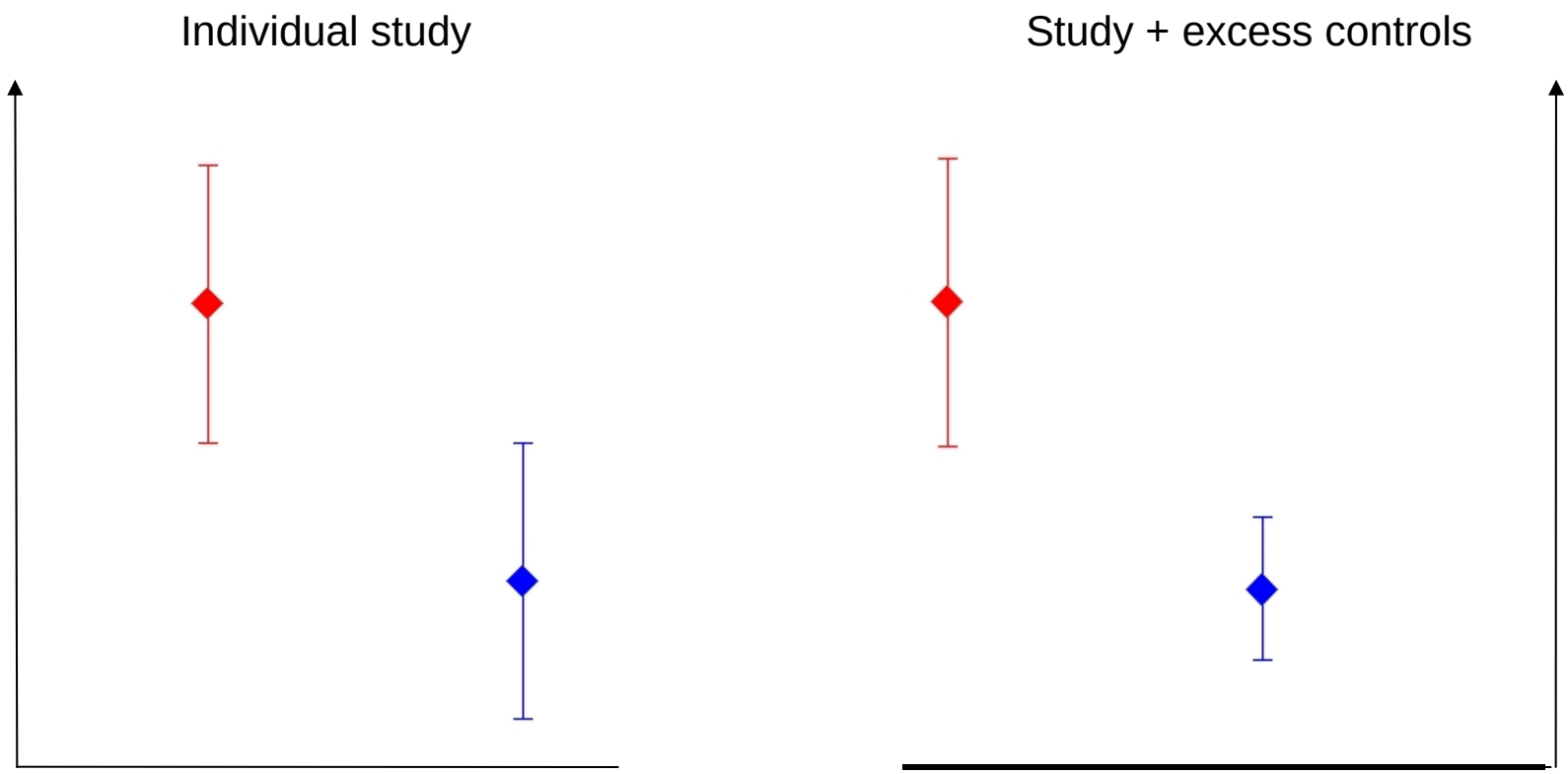
$\lambda_{GC} = 1.14$ $P_{tail} = 1.5$ (66)

The 40K lower performing SNPs had 2.5x as many p<.001 than the 43K highest performing SNPs

Missing data is often heavily biased

- SNPs missing even 1-5% of genotype calls are frequently missing them non-randomly
  - Both SNPs preferentially/exclusively losing heterozygotes and others losing homozygotes are seen

Excess controls add power



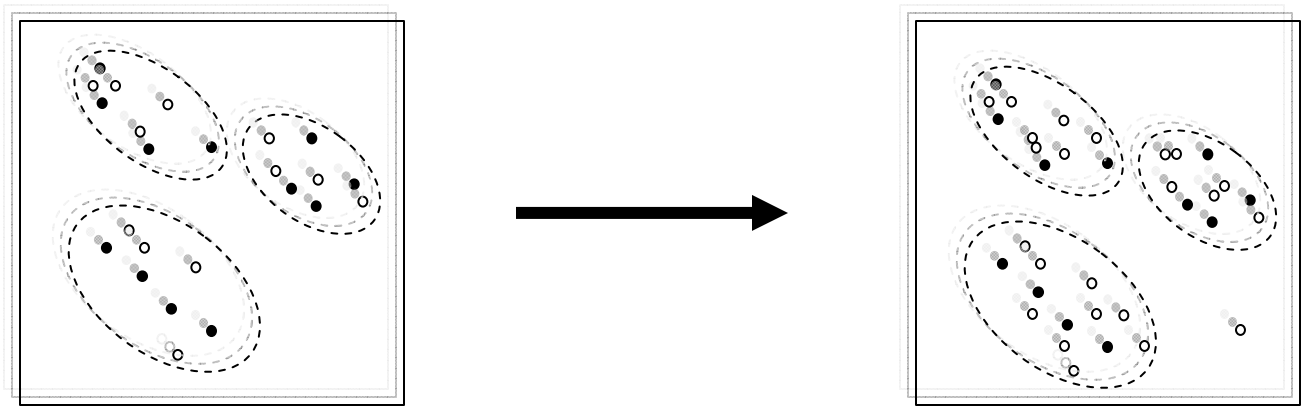
Power gains from excess controls

				Ratio cases : controls					
gene	**disease**	**allele freq**	**OR**	**1:1**	**1:2**	**1:3**	**1:5**	**1:10**	**1:20**
PTPN22	RA	0.08	1.75	0.31	0.52	0.62	0.72	0.79	0.83
TCF7L2	T2D	0.25	1.50	0.69	0.88	0.93	0.96	0.98	0.99
IRF5	SLE	0.40	1.50	0.77	0.92	0.95	0.98	0.99	0.99
PPARG	T2D	0.85	1.25	0.02	0.04	0.04	0.05	0.06	0.06
CTLA4	T1D	0.60	1.20	0.05	0.07	0.09	0.11	0.13	0.14

power (p=0.001)
n=400 cases

Chris Cotsapas, Robert Plenge

Population differences can be managed with many approaches
(structured association, PCA analyses, …)



Matching can also be done on sample quality to
Simultaneously reduce inflation due to technical artifacts

Data quality challenges are not limited to case-control design

Missing data is often heavily biased

- SNPs missing even 1-5% of genotype calls are frequently missing them non-randomly
  - Both SNPs preferentially/exclusively losing heterozygotes and others losing homozygotes are seen

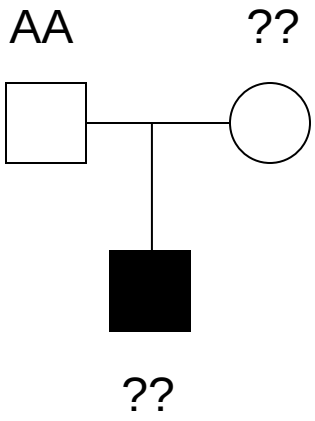
Example from TDT study

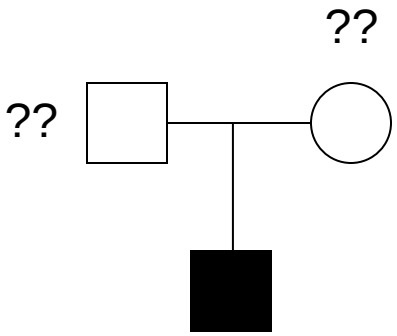
rs10086956 **40-138** chisq=53.96 MAF=.065 missing=3.8% HWp=.003

Observed genotypes AA=1579 AB=236 **BB=0**

SNP passes reasonably standard thresholds but appears to be systematically losing all minor allele homozygotes (and in this case also some hets)
and is thus likely falsely associated with strong undertransmission of the rare allele

Why is this a problem?





Dropping families with AA parents introduces no bias – these are not counted in the TDT!

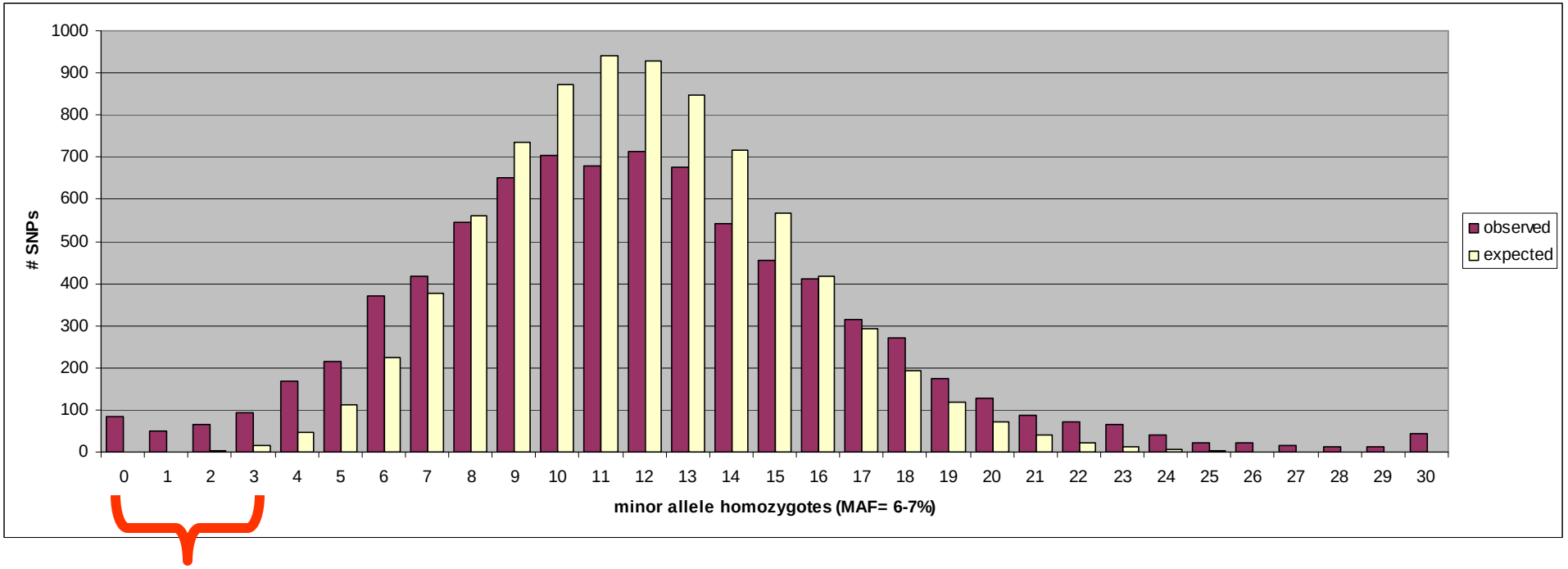
Dropping families with AA children introduces strong bias against allele A – only A could have been transferred in this family and when A is rare, usually 2 transmissions have been removed

AA

Similarly, systematic loss of hets strongly biases against rare alleles

Scope of the problem:

8124 SNPs from ~2400 recent Affy (BRLMM) runs with 280–320 heterozygotes (MAF ~ 6 to 7%)



0-1: observe 133 SNPs, expect < 1
0-3: observe 291 SNPs, expect 21

For low MAF, data appears consistent with a few percent of SNPs dropping most or all minor allele homozygotes

Data quality

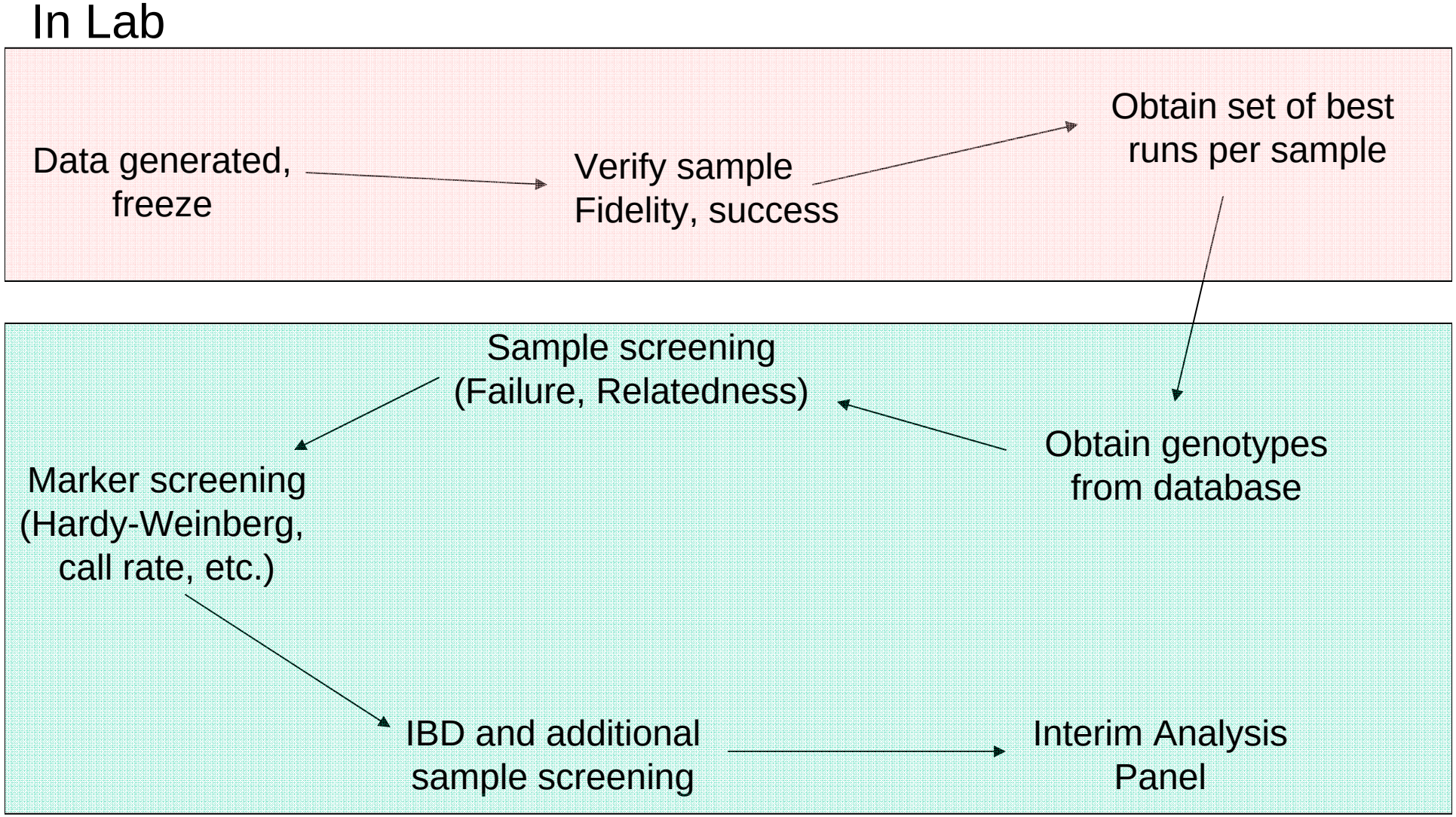
- Lower quality SNPs, even considering only SNPs passing reasonably strict thresholds, often dominate most significant false positives
  - Unlike substructure, relatively little methodologic attention paid to date
  - Affects family and case-control studies

Combining unrelated and family studies

Example 2:

Diabetes Genetics Initiative: Broad/Lund/Novartis Whole Genome Scan in Type 2 Diabetes:

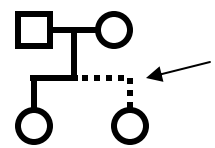
Analysis QC Pipeline



Paul de Bakker, Ben Voight

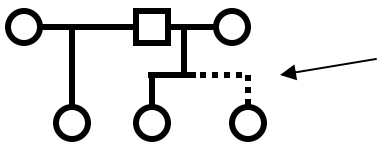
GW-IBD estimate screening

Identify cryptic relatedness



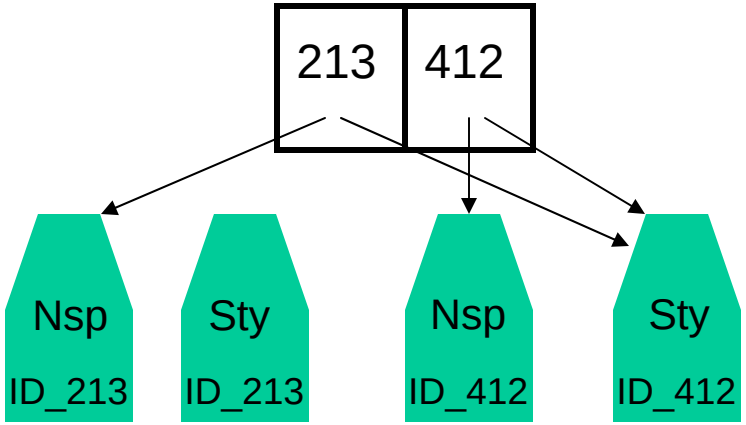
Unknown relationship discovered

Verify existing relationships



Existing relationship consistent with an alternative configuration

Check for sample swapping



DNA Fingerprinting

Master FP (Sequenom)

ID_213	AA TT CT NN GA TT …
ID_412	AG TT TT CG GG NN …
ID_567	GG TC NN GG AA TT …
ID_871	AA TC CT CG GA TC …
…	

Nsp Fragment FP

ID_213 AA TT CT NN GA TT …
ID_412 AG TT TT CG GG CT …

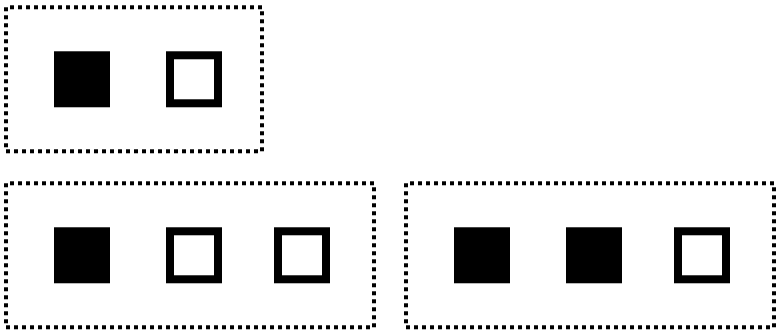
→ Nsp frags match!

Sty Fragment FP

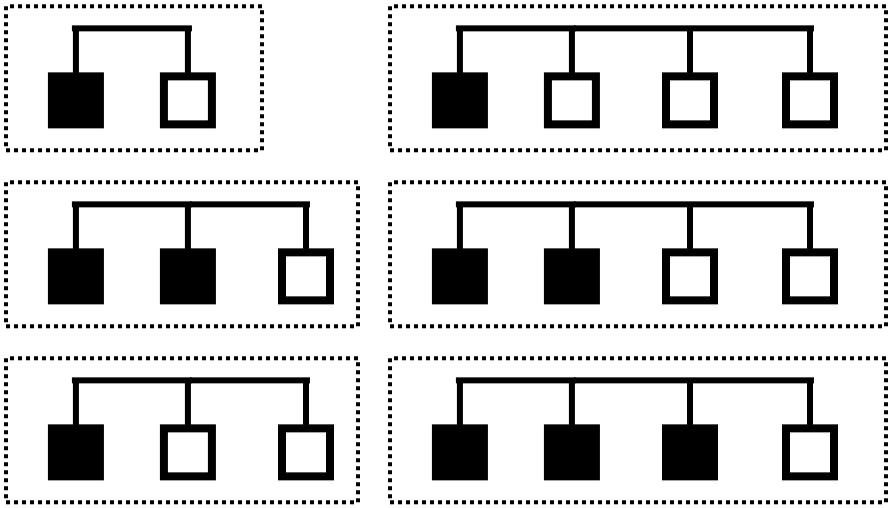
ID_213 AG TT TT CG GG CT …
ID_412 AG TT TT NN GG CT …

→ID_412 consistent with a sample swap with 213

DGI: Study Design



Clusters with
matched cases and controls
[~72% of data]



Clusters with
matched discordant siblings
[~28% of data]

Clusters represent fine-scale matching with respect to: Age, Sex, BMI, Sample Collection site
Cases/controls interleaved and blinded in lab process

Primary Analysis Design

- Cochran-Mantel-Haenszel Stratified Test
  - Testing { SNP x phenotype | Cluster }
  - Intuitively similar to standard χ^2 tests for association
- Clusters represent fine-matching criteria:
  - Kinship [sibships vs. unrelated individuals]
  - BMI, Age, Sex, Collection Locale.
  - 'Orphans' lacking a match pooled into a single cluster (due to interim analysis)
- Significance assessment via permutation

Positive controls

Some examples of true positive results as seen in the WGAS

- rs4506565 (TCF7L2) for T2D ($p \approx 3 \times 10^{-6}$)
- rs4420638 (apoE) for LDL ($p \approx 10^{-8}$)
- rs693 (apoB) for LDL ($p \approx 10^{-7}$)
- rs1800775 (CETP) for HDL ($p \approx 10^{-6}$)

- rs17410962 (LPL) for LDL, TG ($p \approx 0.001$)
- rs5215 (KCNJ11) for T2D ($p \approx 0.001$)
- rs481843 (APOA5) for triglycerides ($p \approx 0.001$)

True associations will require follow-up confirmation in most cases

Summary

- Possible to combine study designs within a study effectively
- Careful experimental design can largely control inflation

Combining distinct studies

- Different levels of combination possible
  - Fisher's p-value method
    - Operates purely on p-values
  - Combined score/Z approach
    - Evaluates pure excess of associated alleles across multiple studies (e.g., SUM obs, exp and var of #alleles in affecteds), better accomodates different sized studies
  - Mantel-Haenszel statistics
    - Estimate OR assuming homogeneity, dovetails with Breslow-Day test of homogeneity

Goncalo will next present on the practical challenges of combining data across studies, particularly when different SNP sets are used

IL23R-Crohn's association

Table 1. Non-Jewish and Jewish ileal Crohn's disease (CD) case-control association study results for IL23R region markers with P-values < 0.0001 in the non-Jewish cohort. Minor allele frequencies (MAF), allelic test P-values, and odds ratios (OR) with 95% confidence intervals (CI) are shown for each case-control cohort (*8*). The ORs shown are for the minor allele. Combined Cochran-Mantel-Haenszel P-values are also shown (*8*).

		Non-Jewish case-control cohort				Jewish case-control cohort				
		CD (n = 547)	Control (n = 548)			CD (n = 401)	Control (n = 433)			Combined
marker	location	MAF	MAF	P-value	OR [95% CI]	MAF	MAF	P-value	OR [95% CI]	P-value
rs1004819	intron	0.374	0.280	3.79E-06	1.53 [1.27,1.84]	0.426	0.334	1.00E-04	1.48 [1.21,1.82]	1.54E-09
rs7517847	intron	0.331	0.443	1.09E-07	0.62 [0.52,0.74]	0.240	0.352	5.84E-07	0.58 [0.47,0.72]	3.36E-13
rs10489629	intron	0.378	0.475	4.27E-06	0.67 [0.56,0.80]	0.355	0.465	5.79E-06	0.63 [0.52,0.77]	1.14E-10
rs2201841	intron	0.385	0.291	4.57E-06	1.52 [1.27,1.83]	0.414	0.315	2.92E-05	1.53 [1.25,1.89]	5.46E-10
rs11465804	intron	0.020	0.063	7.52E-07	0.30 [0.18,0.51]	0.048	0.096	1.39E-04	0.47 [0.31,0.71]	5.97E-10
rs11209026	Arg381Gln	0.019	0.070	5.05E-09	0.26 [0.15,0.43]	0.033	0.070	7.95E-04	0.45 [0.27,0.73]	3.55E-11
rs1343151	intron	0.275	0.370	2.26E-06	0.65 [0.54,0.78]	0.229	0.336	1.69E-06	0.59 [0.47,0.73]	1.64E-11
rs10889677	exon-3'UTR	0.385	0.288	1.82E-06	1.55 [1.29,1.86]	0.419	0.316	1.51E-05	1.56 [1.27,1.91]	9.58E-11
rs11209032	inter-genic	0.393	0.293	1.03E-06	1.56 [1.30,1.87]	0.382	0.298	3.49E-04	1.45 [1.18,1.79]	1.60E-09
rs1495965	inter-genic	0.498	0.412	2.93E-05	1.44 [1.21,1.71]	0.469	0.412	0.0204	1.26 [1.03,1.53]	2.55E-06

Replication seen in family-based studies

Table 2. Family-based and combined (case-control and family-based) association results. Family-based association P-values were computed using the empirical variance estimator implemented in the FBAT software package (*8*). Combined Fisher P-values for all case-control (Table 1) and nuclear family cohorts are also shown (*8*).

		Non-Jewish CD (518 families, 651 affected offspring)	Non-Jewish UC (215 families, 251 affected offspring)	Jewish CD (77 families, 99 affected offspring)	Jewish UC (80 families, 91 affected offspring)	All IBD (883 families, 1,119 affected offspring)
marker	location	P-value	P-value	P-value	P-value	P-value
rs1004819	intron	3.60E-05	1.20E-03	1.24E-02	5.47E-01	6.06E-08
rs7517847	intron	2.30E-05	2.71E-01	3.50E-02	5.00E-01	1.80E-05
rs10489629	intron	1.87E-03	2.70E-01	4.33E-01	8.21E-01	1.27E-03
rs2201841	intron	5.80E-04	3.21E-04	3.50E-02	5.69E-01	1.04E-07
rs11465804	intron	1.32E-04	2.70E-03	8.90E-05	3.71E-01	3.46E-09
rs11209026	Arg381Gln	8.00E-06	2.97E-04	9.41E-04	4.91E-01	1.32E-10
rs1343151	intron	9.63E-02	8.51E-02	3.30E-02	1.89E-01	1.24E-05
rs10889677	exon-3'UTR	2.60E-03	3.35E-04	5.88E-02	7.32E-01	1.65E-06
rs11209032	inter-genic	2.68E-03	3.57E-04	3.48E-02	7.50E-01	2.41E-06
rs1495965	inter-genic	4.07E-04	1.74E-02	3.93E-02	9.21E-01	1.72E-05

Replication seen in family-based studies

Table 2. Family-based and combined (case-control and family-based) association results. Family-based association P-values were computed using the empirical variance estimator implemented in the FBAT software package (*8*). Combined Fisher P-values for all case-control (Table 1) and nuclear family cohorts are also shown (*8*).

		Non-Jewish CD (518 families, 651 affected offspring)	Non-Jewish UC (215 families, 251 affected offspring)	Jewish CD (77 families, 99 affected offspring)	Jewish UC (80 families, 91 affected offspring)	All IBD (883 families, 1,119 affected offspring)	Combined (family-based and case-control)
marker	location	P-value	P-value	P-value	P-value	P-value	P-value
rs1004819	intron	3.60E-05	1.20E-03	1.24E-02	5.47E-01	6.06E-08	1.78E-14
rs7517847	intron	2.30E-05	2.71E-01	3.50E-02	5.00E-01	1.80E-05	9.99E-16
rs10489629	intron	1.87E-03	2.70E-01	4.33E-01	8.21E-01	1.27E-03	1.62E-11
rs2201841	intron	5.80E-04	3.21E-04	3.50E-02	5.69E-01	1.04E-07	1.10E-14
rs11465804	intron	1.32E-04	2.70E-03	8.90E-05	3.71E-01	3.46E-09	3.33E-16
rs11209026	Arg381Gln	8.00E-06	2.97E-04	9.41E-04	4.91E-01	1.32E-10	6.62E-19
rs1343151	intron	9.63E-02	8.51E-02	3.30E-02	1.89E-01	1.24E-03	2.74E-12
rs10889677	exon-3'UTR	2.60E-03	3.35E-04	5.88E-02	7.32E-01	1.65E-06	3.40E-14
rs11209032	inter-genic	2.68E-03	3.57E-04	3.48E-02	7.50E-01	2.41E-06	5.50E-13
rs1495965	inter-genic	4.07E-04	1.74E-02	3.93E-02	9.21E-01	1.72E-05	3.55E-09

Acknowledgments

HapMap & Genetic Analysis

Paul de Bakker Itsik Pe'er
Ben Voight Roman Yelensky
Julian Maller Todd Green
Chris Cotsapas Finny Kuruvilla
Robert Plenge

David Altshuler Shaun Purcell

Stacey Gabriel

DGI study: Leif Groop, Tom Hughes, David Altshuler