

# Genome-Wide Association Studies

---

Sanjay Shete, Ph. D.

M. D. Anderson Cancer Center

Houston, Texas

[sshete@mdanderson.org](mailto:sshete@mdanderson.org)

# Genome Wide Association Studies

---

A genome-wide association study:

To scan several thousand SNPs on many individuals to find genetic variations associated with a particular disease.

Help develop better strategies to detect, treat and prevent the disease.

GWAS are particularly useful in finding genetic variations that contribute to common, complex diseases, such as asthma, cancer, diabetes, heart disease and mental illnesses.

Source: <http://www.genome.gov/20019523#1>

# Genome Wide Association Studies

---

- Begins a new paradigm in genetic epidemiology
  - Hypothesis free
  - Study biology AFTER association is documented, unlike traditional genetic epidemiology where biology is done first

# Genome Wide Association Studies

---

- Advantages over linkage analysis:
  - Increased precision to localize a disease susceptibility locus
  - Association study may have more power than a linkage study, particularly for genes with modest individual effects
  - Association studies can be performed on unrelated individuals, simplifying recruitment and enabling larger samples.

# Genome Wide Association Studies

---

- “Not only is it argued that we need know basically nothing substantial about the biology of a trait to do a mapping study, but it need not even aggregate in families, and to the contrary, the study design is to compare unrelated cases with controls. Often this is now proposed as an attraction of a study design. A strange way to do science.” Terwilliger and Weiss, 2003

# What causes association that enables us to perform GWAS?

---

- Linkage Disequilibrium (LD) is an association between the genotypes at two or more loci that enables us to perform GWAS
- LD is typically observed as a disease phenotype and marker genotype(s) association due to proximity of putative disease locus and the marker loci

# What causes association that enables us to perform GWAS?

---

- When a disease mutation first occurs at a locus, it is associated with all variants at loci nearby on the chromosome
- After many generations of random mating, equilibrium is attained, but if the two loci are tightly linked, the LD between them will remain.
- This is the basis of fine mapping using LD – searching for a population association between a disease and a linked marker variant

# GWAS: What and Why?

---

- Why perform an association study?
  - Locate causal variants in the genome
  - Estimate attributable risk due to causal variants
  - To predict clinical outcomes using associated variant → prediction, treatment response



# GWAS: What and Why?

---

- What kinds of traits?
  - Binary, ordinal, continuous
  - Univariate or multivariate
- Type of Sample?
  - Random cohorts (unrelated, trios, nuclear families, extended pedigrees)
  - Selected cohorts (case/control, trios, nuclear, multiplex)

# Allelic association

- ❖ A number of generation ago, an allele D1 (with a marker allele M1 at a nearby locus) mutated to a disease allele D2.
- ❖ Chromosome is passed down through generations and in current generation there are many copies.
- ❖ If the distance between A and D is very small (fewer recombinations) then most of the D1 chromosomes will also have M1.

# Linkage disequilibrium (allelic association)

For loci D and M with alleles  $D_1, D_2, M_1, M_2 \dots$

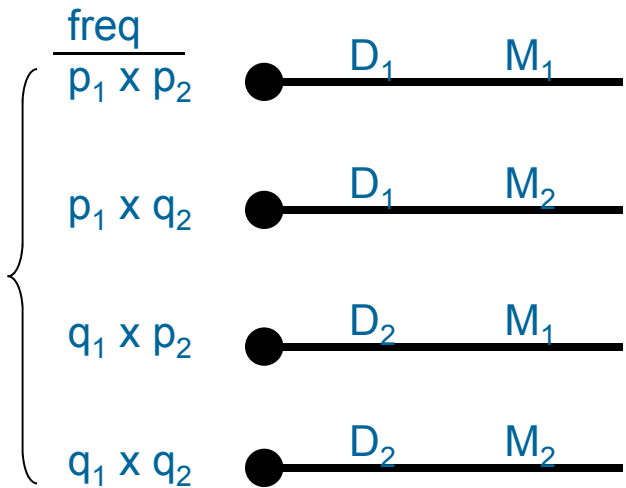
freq  $D_1 = p_1$

freq  $M_1 = p_2$

freq  $D_2 = q_1 = 1 - p_1$

freq  $M_2 = q_2 = 1 - p_2$

If allele frequencies are independent of linkage relationship, i.e. linkage equilibrium (same as no linkage disequilibrium)



	$M_1$	$M_2$	
observed values $D_1$	$p_1 \times p_2 + \delta$ = freq ( $D_1M_1$ )	$p_1 \times (1 - p_2) - \delta$ = freq ( $D_1M_2$ )	Sum = $p_1$
$D_2$	$(1 - p_1) \times p_2 - \delta$ = freq ( $D_2M_1$ )	$(1 - p_1) \times (1 - p_2) + \delta$ = freq ( $D_2M_2$ )	Sum = $1 - p_1$

$\delta$  is the disequilibrium coefficient = freq ( $D_1M_1$ ) -  $p_1 \times p_2$

# Some properties of allelic association

	M <sub>1</sub>	M <sub>2</sub>	
D <sub>1</sub>	p <sub>1</sub> × p <sub>2</sub> + δ	p <sub>1</sub> × (1 - p <sub>2</sub> ) - δ	← Sum = p <sub>1</sub>
D <sub>2</sub>	(1 - p <sub>1</sub> ) × p <sub>2</sub> - δ	(1 - p <sub>1</sub> ) × (1 - p <sub>2</sub> ) + δ	← Sum = 1 - p <sub>1</sub>

No linkage disequilibrium at  $\delta = 0$   
 Consider  $p_1 = 0.5$  and  $p_2 = 0.5$

	M <sub>1</sub>	M <sub>2</sub>	
D <sub>1</sub>	0.25 + δ	0.25 - δ	← Sum = p <sub>1</sub>
D <sub>2</sub>	0.25 - δ	0.25 + δ	← Sum = 1 - p <sub>1</sub>

$$\delta_{\min} = -0.25;$$

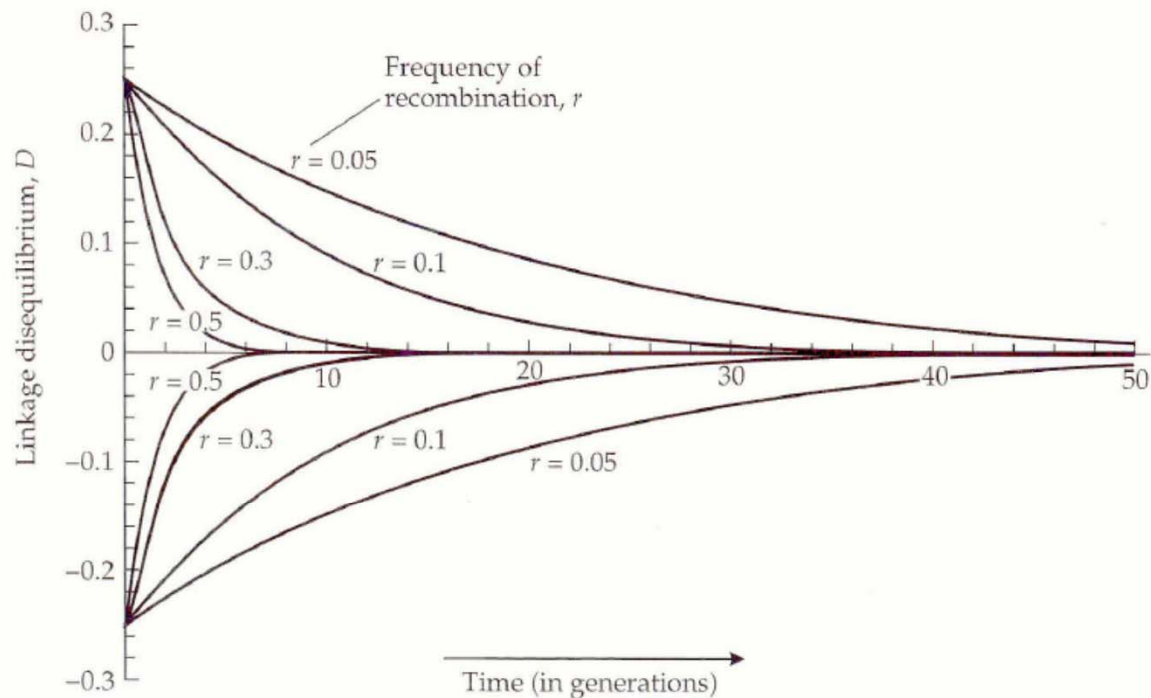
$$\delta_{\max} = 0.25$$

← The range of  $\delta$  depends on allele frequencies (maximal for 0.5)

# Magnitude of LD relative to theoretical maximum or minimum

$$D' = \begin{cases} \frac{D}{\min(p_1q_2, p_2q_1)} & \text{if } D > 0 \\ \frac{D}{\min(p_1q_1, p_2q_2)} & \text{if } D < 0 \end{cases}$$

# LD: distance and generations



**Figure 3.9** Linkage disequilibrium between genes gradually disappears when mating is random, provided there is no countervailing force building it up. The rate of approach to linkage equilibrium depends on the recombination frequency between the genes. The disappearance of linkage disequilibrium is gradual even with free recombination ( $r = 1/2$ ). In these examples, the frequencies of both alleles at both loci equal  $1/2$ , and the initial linkage disequilibrium is either at its maximum ( $D = 0.25$ ) or minimum ( $D = -0.25$ ) value, given these allele frequencies.

# Why Case-Control?

---

- Case-Control
  - Has been work horse of association studies
- Cohorts
  - Multiple end points can be considered
  - A common set of controls can be used for several phenotype (“Universal controls”)

# GWAS: What and Why?

---

- What are we looking for?
  - Effects of one locus genotypes
  - Effects of alleles (haplotypes)
  - Interactions with environment
  - Joint effects of multiple factors?
    - Additive on some scale
    - Necessarily require interactions (epistasis)
    - Transformation



# Designs for GWAS

---

- One stage
  - All markers typed on all samples
  - Replication left up to others
- Replication
  - Replication of entire scans not a good use of resources except as a protection for false negatives

# Designs for GWAS

---

## ■ What is replicated

- Scientific replication: Different investigators studying different populations with different study designs, each with potentially different strengths and weaknesses (*multiple, one stage studies*)
- Statistical replication: Multistage sampling designs have built-in replication if same study design and population are used in different stages (*two stage studies*)

# Designs for GWAS

---

- Two stage design is one way to increase efficiency
  - SNP genotyping costs decreasing
  - Increasing sample sizes typed
- Two stage designs can retain near full power at much reduced cost compared to one stage

# Two Stage Design

---

## ■ What is Two stages?

- Two independent analyses (same markers, same individuals)
- Two sets of markers, same individuals
- Two sets of individuals, same markers

## ■ Why Two stages?

- ↑ Power
- ↓ Cost
- ↑ Accuracy (location, prediction)

# Two Stage Design

---

- What should a Stage 1 sample be?
  - Samples with linkage information
    - DNA available
    - Haplotypes easily determined
    - Essential if allelic heterogeneity
    - Disadvantages = cost, family members, poor for predicting risk

# Two Stage Design

---

- What should a Stage 1 sample be?
  - Samples without linkage info (case-control)
    - Pooled samples vs non-pooled
    - Disadvantages = cost of prepping pools, less haplotype info, accuracy of alleles/haplotype measures, accuracy of calling

# Stage 1 Sample Composition

---

- Cases with family history can enrich genetic susceptibility

Issues: Introduces cryptic relatedness

- Cases with high severity

Issues: enrich genetic and also environmental factors

- Strategy depends upon intention. (a) To find common polymorphism having main effect on disease or (b) To find polymorphisms with modifying effects on other genes and environment

# Two Stage Design

---

- What should Stage 1 do?
  - Subset of samples ( $\pi_{\text{samples}}$ ) typed on large number of markers (M)
  - Determine what method of analysis is best (take 1/3 data to figure out what statistic to use)
  - Determine markers to go forward
  - Determine what individuals to go forward
  - Determine best hypothesis to try and replicate



# Two Stage Design

---

- What should Stage 2 do?
  - Replicate Stage 1
    - New individuals, same markers
    - New markers, same individuals
- Power depends on...
  - How many markers? How are samples divided between two stages? What proportion of markers typed in stage 2? What method to test for association?

# Designs for GWAS

---

Joint analysis (Skol et al. 2006)

- Recommended if more than 30% of participants are in Stage 1 and more than 1% of markers are followed-up in Stage 2

# Joint Analysis

---

- Split data into two groups:  $N=N_1+N_2$
- In first group (i.e., stage 1), genotype all markers and calculate a test statistic at each marker

$$z_{1M} = \frac{\hat{p}'_{1M} - \hat{p}_{1M}}{\sqrt{[\hat{p}'_{1M}(1-\hat{p}'_{1M}) + \hat{p}_{1M}(1-\hat{p}_{1M})]/(2N_1)}}$$

where  $\hat{p}'_{1M}; \hat{p}_{1M}$  are estimated allele frequencies at marker M in cases & controls, respectively

# Joint Analysis

---

- Then pick the number of markers to be evaluated in group 2
- Define threshold  $C_1$  such that

$$P(|z_{1.}| > C_1) = \pi_{\text{markers}}$$

- Note, under the null of no association,  $z_{1.}$  follows a  $N(0,1)$  distribution
- $C_1 = 1.96$ ,  $\pi_{\text{markers}} = 0.05$
- Correction for multiple testing:

$$\alpha_{\text{genome}} / M = \pi_{\text{markers}}$$

# Joint Analysis

---

- Calculate a test statistic at each marker in group 2 (i.e., Stage 2)

$$z_{2M} = \frac{\hat{p}'_{2M} - \hat{p}_{2M}}{\sqrt{[\hat{p}'_{2M}(1 - \hat{p}'_{2M}) + \hat{p}_{2M}(1 - \hat{p}_{2M})]/(2N_2)}}$$

# Joint Analysis

---

- Calculate the JOINT test statistic for each marker in stage 2

$$z_M = \sqrt{N_1} z_{1M} + \sqrt{N_2} z_{2M}$$

- Joint analysis combines evidence of association without assuming equal effect sizes or allele frequencies between the two stages (i.e., accounts for between stage heterogeneity)
- $z_1$  and  $z_M$  are not independent, so false positive rate calculated by integration

# Joint Analysis

---

## ■ Power

- Power for stage 1 =  $\Pr(\text{disease variant selected for stage 2})$
- Power calculator for arbitrary sample sizes and genetic models: <http://csg.sph.umich.edu>
- A two stage design using joint analysis can achieve nearly the same power as the one stage design in which all the samples are genotyped on all markers
- Joint strategy is more powerful than replication strategy except when the association is greater in stage 2 compared to stage 1

# Power Comparison of Replication vs Joint

---

- For example, Skol et al. 2006 compared power of replication based and joint analysis strategies with  $\alpha_{\text{genome}}=0.05$ 
  - for a wide range of sample sizes, proportions of samples used in Stage 1, and proportions of markers selected for follow-up in Stage 2 and under different genetic models, effect sizes and disease variant frequencies



# Power Comparison of Replication vs Joint

---

- In the case where 1000 cases and 1000 controls were divided equally among the two stages ( $\pi_{\text{samples}} = 50\%$ ), 10% of stage 1 markers were followed up in stage 2, disease prevalence was 0.10, control allele frequency was 0.50, multiplicative disease model and genotype relative risk = 1.4
  - Replication-based analysis power = 26%
  - Joint analysis power = 74%

(See Figure 2 in Skol, 2006)

# Power Comparison of One Stage vs Joint

---

- In the case where 1000 cases and 1000 controls typed on all markers (300K markers = 600M genotypes),  $GRR=1.4$ , prevalence 0.10 and risk allele freq in controls=0.50. One stage power = 75%
- For comparison, a joint analysis can achieve 72% power with only a third as many genotypes using 30% of samples in stage 1 and following up 5% of samples in stage 2

(See Table 1 Skol, 2006)

# Optimal Design

---

- Examine the influence of the ratio of  $R$  on stage 2 to 1 per genotype cost
  - What is the proportion of Stage 1 power retained?
  - What is the impact of false positives?
- Stage 1: cost of standard chip
- Stage 2: per genotype cost higher, but fewer markers typed

# Optimal Design

---

- Optimal design depends on the proportion of each stage's power retained... can sacrifice some power or false positive rate to save money if necessary
- Joint analysis is more powerful than replication except when heterogeneity in Stage 2 is high.

# Power and Type 1 Error for GWAS

---

- Genetic Power Calculator: comparison of number of tests versus sample size requirements
  - <http://pngu.mgh.harvard.edu/~purcell/gpc/>
- Genome-wide (Per Marker) alpha level
  - Bonferroni is conservative when the tests are not independent of each other because of LD between markers
  - Alpha  $\sim 10^{-7}$  regardless of number of tests based on Bayesian- alpha-level calculation
  - Adaptive significance level: Benjamini and Hochberg (1995)
  - More recently accepted threshold is  $5 \times 10^{-8}$

# Platforms for GWAS

---

- Affymetrix
  - Essentially random set of SNPs:
    - Affy 100K
    - Affy 500K
- Illumina
  - Designed using HapMap
    - Illumina 317K
    - Illu 550K
    - Illu 650K (550K + 100K YRI fill-in)
- More recently you can type more than million SNPs

# Comparison of Platforms

---

- Choices: maximal power, sample size, which SNPs to genotype, analysis method
- Constraints: cost, sample size often has upper limit, which SNPs?: commercially available chips (Affymetrix/Illumina), analysis

# Comparison of Platforms

---

- Comparisons of chips on the basis of coverage can be misleading as a surrogate for power comparison (power depends on coverage + sample size, allele frequency, magnitude of effect and analysis method)
  - Comparing coverage and power: Difference between different chips is a few percent, except Affy100
  - Comparing power over different effect sizes: All except Affy100 are reasonably close to theoretical limit



# General Thoughts

---

- As a general rule, put resources into larger sample sizes rather than more SNPs/coverage per chip” leads to greater payoff by increased sample size
- If sample sizes are limited, and if affordable, it’s obviously better to use chips with more coverage
- Save intermediate files (call rates can be improved) and assign alleles for the entire sample with the best available algorithm

# Resources

---

## GAIN=Genetic Association Information Network

- Public-private partnership of the Foundation for the NIH which will include corporations, private foundations, advocacy groups, concerned individuals, and the National Institutes of Health
- [http://www.fnih.org/GAIN/GAIN\\_home.shtml](http://www.fnih.org/GAIN/GAIN_home.shtml)
- Support for at least seven studies using Perlegen and Affymetrix platforms
- <http://grants.nih.gov/grants/gwas/>

# Thoughts

---

- Correlations (LD) suggest that you need to only genotype roughly 300 to 400 thousand SNPs to obtain nearly all genotypic variation for all 9-10 million possible SNPs.
- Don't accept only commercially available platforms. If you only type bins with multiple SNPs (and ignore bins with only 1 SNP) then you will miss 50% of all possible LD bins.

# Sample Sizes and MAF

---

Number of cases/control pairs to significance level  $\alpha=10^{-7}$  with 95% power in a single stage study assuming multiplicative genetic model. Thomas, CEBP, 2006

RR	MAF=5%	MAF=10%	MAF=20%
1.2	28000	15000	8700
1.5	5200	2800	1700
2.0	1600	870	540
2.5	830	470	300
3.0	540	310	200

# Sample sizes

---

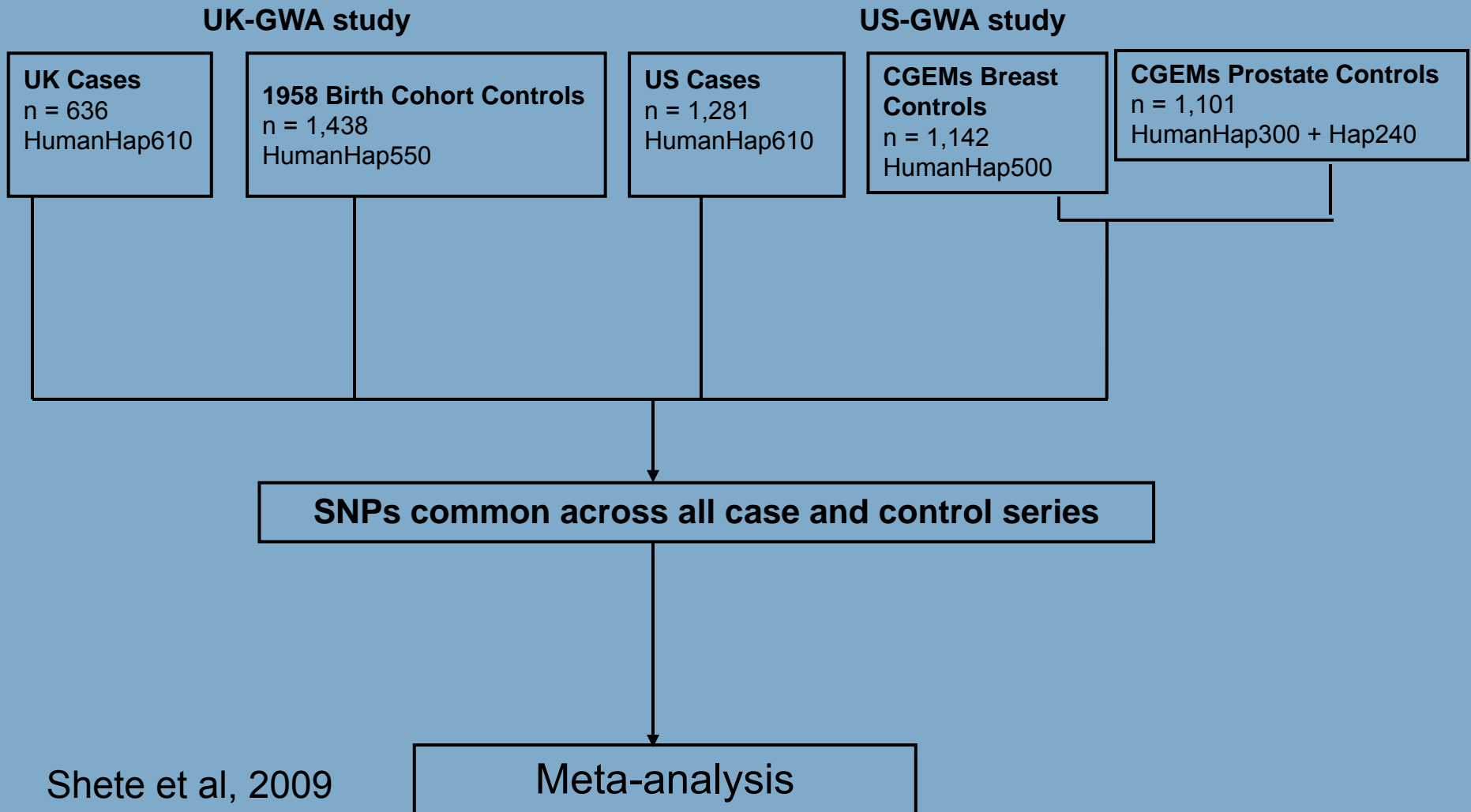
- These numbers may be reduced by half in a multistage design!
- However, testing multiple genetic models, additional SNPs or haplotypes, subgroups or interactions would require an even stricter significance level and larger sample sizes!

# Some Recent Successes

---

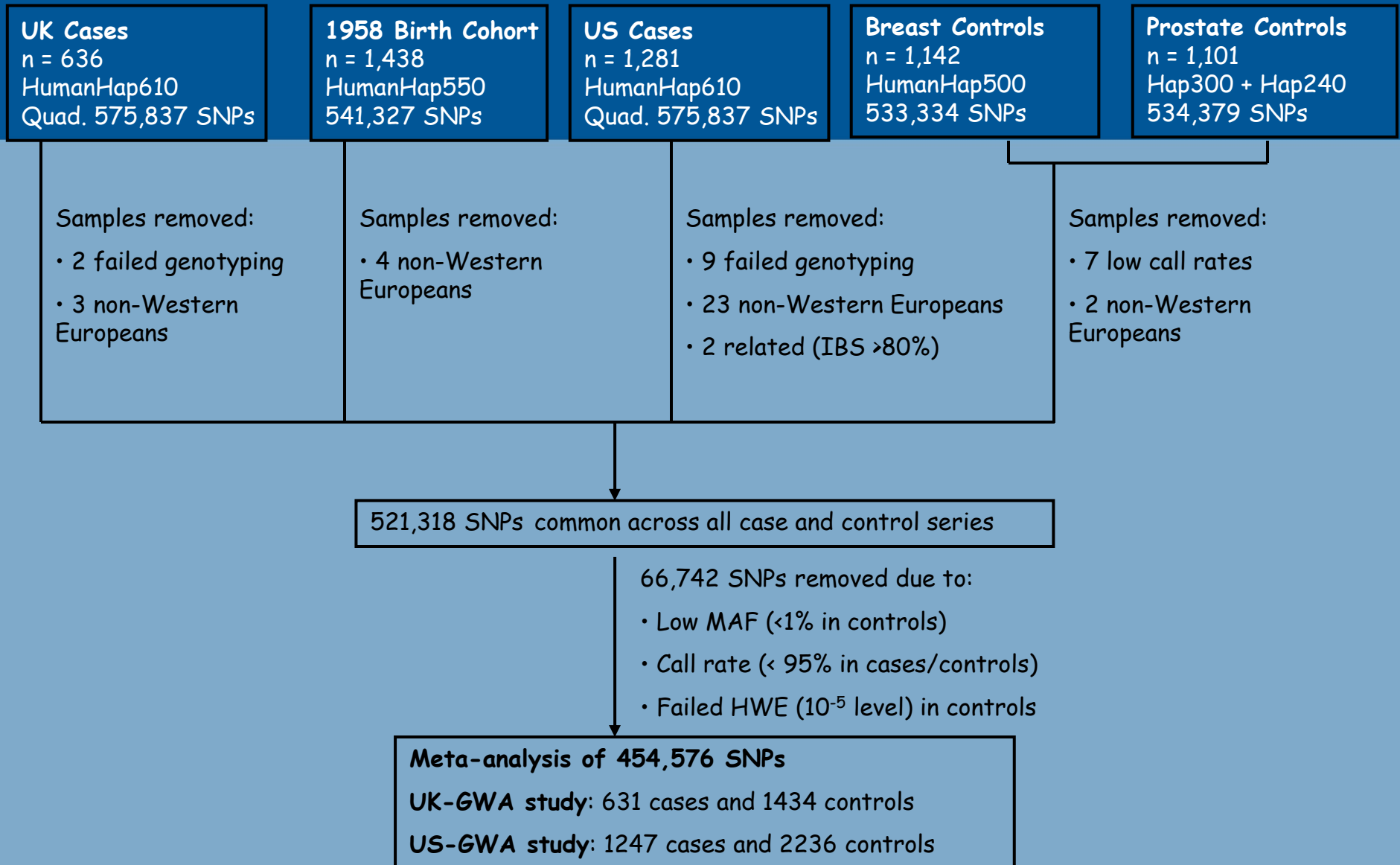
- PTPN22 Predicts risk for autoimmune diseases
- Used 475 cases/controls in discovery pool
- Used 463 probands/controls in confirmation
- PTPN22 confers 1.9 fold increased risk to heterozygotes of prevalent risk allele (about 85% of individuals carry risk allele)
- Very strong decline in allele frequencies with minor allele frequency (MAF) approaching 20% in Northern European and near 0% in Southern European populations

# Schema for the GWAS



## UK-GWA study

## US-GWA study



# Patient and SNP exclusion schema



# Replication analyses

- 34 SNPs had p-value less than  $10^{-5}$  and these SNPs were fast tracked in 3 independent case-control series

French series 1392 cases, 1602 controls

German series 504 cases, 573 controls

Swedish series 649 cases, 778 controls

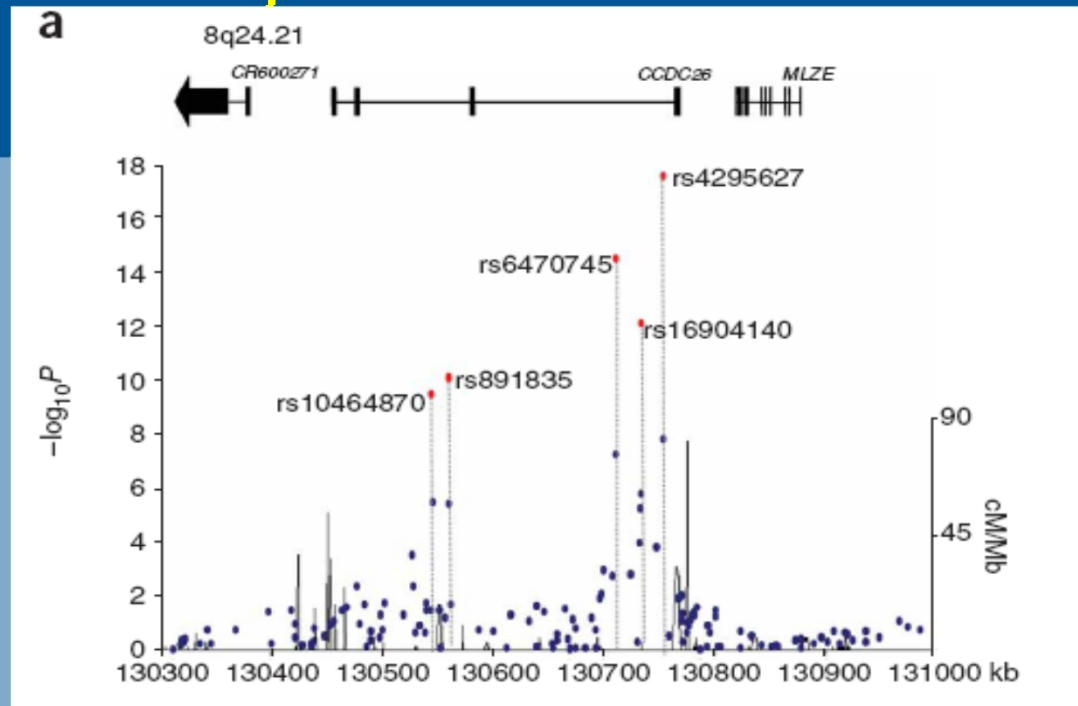
- 14 of the 31 SNPs representing 5 genomic regions satisfied the accepted threshold for genome-wide statistical significance  $5 \times 10^{-8}$

# Summary Results in GWAS and Replication Series

SNP	Chr.	Gene <sup>a</sup>	Location (bp)	Ancestral allele frequency	Risk allele <sup>b</sup>	GWA studies		Replication studies		Combined		
						OR (95% CI)	<i>P</i>	OR (95% CI)	<i>P</i>	OR (95% CI)	<i>P</i>	<i>P</i> <sub>het</sub>
rs2736100	5	<i>TERT</i>	1,339,516	0.51	G	1.20 (1.10–1.33)	$2.21 \times 10^{-6}$	1.33 (1.20–1.49)	$2.87 \times 10^{-13}$	1.27 (1.19–1.37)	$1.50 \times 10^{-17}$	0.18
rs2853676	5	<i>TERT</i>	1,341,547	0.27	A	1.22 (1.14–1.31)	$5.30 \times 10^{-6}$	1.30 (1.21–1.38)	$1.06 \times 10^{-9}$	1.26 (1.20–1.32)	$4.21 \times 10^{-14}$	0.67
rs10464870	8	<i>CCDC26</i>	130,547,005	0.21	C	1.24 (1.15–1.34)	$3.90 \times 10^{-6}$	1.22 (1.13–1.31)	$1.77 \times 10^{-5}$	1.23 (1.17–1.30)	$3.04 \times 10^{-10}$	0.05
rs891835	8	<i>CCDC26</i>	130,560,934	0.22	G	1.24 (1.15–1.33)	$3.92 \times 10^{-6}$	1.24 (1.15–1.33)	$4.43 \times 10^{-6}$	1.24 (1.17–1.30)	$7.54 \times 10^{-11}$	0.01
rs6470745	8	<i>CCDC26</i>	130,711,103	0.20	G	1.30 (1.20–1.39)	$5.79 \times 10^{-8}$	1.31 (1.22–1.41)	$9.09 \times 10^{-9}$	1.30 (1.24–1.37)	$2.77 \times 10^{-15}$	0.01
rs16904140	8	<i>CCDC26</i>	130,734,825	0.21	A	1.25 (1.16–1.35)	$1.41 \times 10^{-6}$	1.28 (1.19–1.37)	$1.14 \times 10^{-7}$	1.27 (1.20–1.33)	$7.88 \times 10^{-13}$	0.01
rs4295627	8	<i>CCDC26</i>	130,754,639	0.17	G	1.33 (1.23–1.42)	$1.47 \times 10^{-8}$	1.39 (1.30–1.49)	$2.20 \times 10^{-11}$	1.36 (1.29–1.43)	$2.34 \times 10^{-18}$	0.01
rs1063192	9	<i>CDKN2A/B</i>	21,993,367	0.44	C	1.21 (1.13–1.29)	$1.44 \times 10^{-6}$	1.21 (1.14–1.29)	$6.97 \times 10^{-7}$	1.21 (1.16–1.27)	$4.61 \times 10^{-12}$	0.81
rs2157719	9	<i>CDKN2A/B</i>	22,023,366	0.57	G	1.22 (1.11–1.35)	$6.80 \times 10^{-7}$	1.22 (1.11–1.33)	$4.42 \times 10^{-7}$	1.22 (1.14–1.30)	$1.41 \times 10^{-12}$	0.68
rs1412829	9	<i>CDKN2A/B</i>	22,033,926	0.42	C	1.22 (1.14–1.30)	$7.23 \times 10^{-7}$	1.23 (1.15–1.30)	$1.80 \times 10^{-7}$	1.22 (1.17–1.28)	$6.23 \times 10^{-13}$	0.67
rs4977756	9	<i>CDKN2A/B</i>	22,058,652	0.40	G	1.25 (1.17–1.32)	$2.39 \times 10^{-8}$	1.24 (1.16–1.31)	$5.90 \times 10^{-8}$	1.24 (1.19–1.30)	$7.24 \times 10^{-15}$	0.94
rs498872	11	<i>PHLDB1</i>	117,982,577	0.31	T	1.26 (1.17–1.34)	$1.03 \times 10^{-7}$	1.12 (1.04–1.20)	$4.56 \times 10^{-3}$	1.18 (1.13–1.24)	$1.07 \times 10^{-8}$	0.04
rs6010620	20	<i>RTEL1</i>	61,780,283	0.23	G	1.28 (1.18–1.38)	$8.38 \times 10^{-7}$	1.28 (1.18–1.38)	$6.49 \times 10^{-7}$	1.28 (1.21–1.35)	$2.52 \times 10^{-12}$	0.38
rs2297440	20	<i>RTEL1</i>	61,782,743	0.22	C	1.28 (1.18–1.38)	$1.01 \times 10^{-6}$	1.26 (1.16–1.35)	$4.44 \times 10^{-6}$	1.27 (1.20–1.34)	$2.06 \times 10^{-11}$	0.40

Shete et al. 2009 Nature Genetics

# 8q24.21 association



- $P=2.34 \times 10^{-18}$  ; OR=1.36 (1.29-1.43;)

- intron-3 of *CCDC26* - RA (Retinoic acid) modulator of differentiation & death

- RA induces caspase-8 transcription through phosphorylation of CREB & increases apoptosis to death stimuli in neuroblastoma cells and in glioblastoma cells with down regulation of telomerase activity

- This region is also implicated in colorectal, prostate, bladder, breast cancer risk and in cleft lip (a risk factor for primary brain tumor)

- These SNPs may be defining a common disease locus in this region

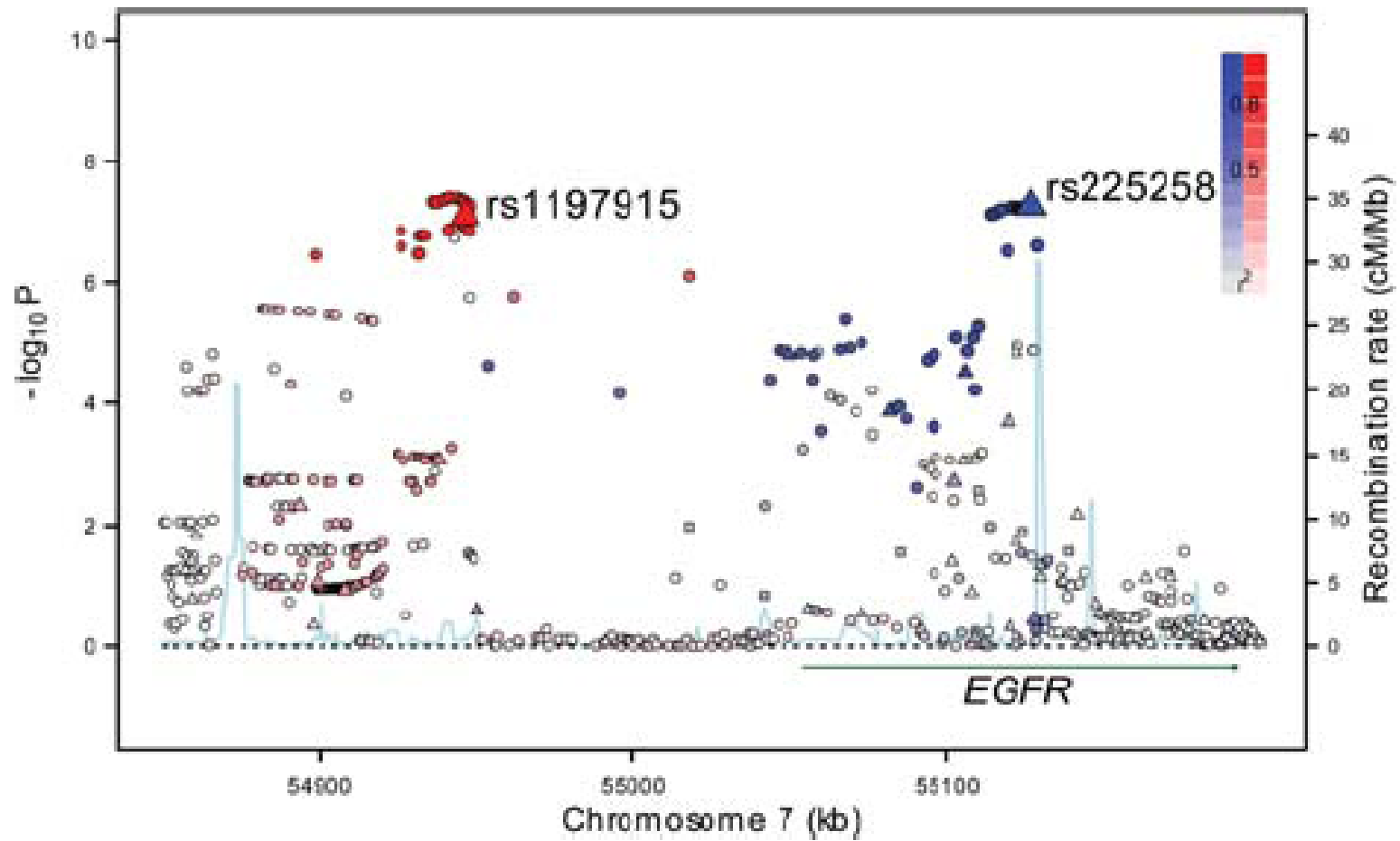
# Follow-up Findings-EGFR

GWA data	Samples removed	Final data
<b>UK cases</b> (n = 636) HumanHap610 Quad	- 2 failed genotyping - 3 non-Western Europeans	<b>UK GWA study</b> 631 cases (270 GBM) 2,699 controls
<b>1958 WTCCC controls</b> (n = 2930) Human 1M Duo		
<b>US cases</b> (n = 1,281) HumanHap610 Quad	- 9 failed genotyping - 23 non-Western Europeans - 2 closely related - 1 non-Western European	<b>US GWA study</b> 1,247 cases (655 GBM) 2,236 controls
<b>CGEMs breast cancer controls</b> (n = 1,143) HumanHap500		
<b>CGEMs prostate cancer controls</b> (n = 1,102) HumanHap240&300		
<b>French cases</b> (n = 1,495) HumanHap660	- 20 failed genotyping - 9 duplicates, 2 closely related - 39 non-Western Europeans - 2 sex discrepancies - 23 non-Western Europeans	<b>French GWA study</b> 1,423 cases (430 GBM) 1,190 controls
<b>French controls</b> (n = 1,213) HumanHap660		
<b>German cases</b> (n = 880) HumanHap660	- 6 failed genotyping - 8 duplicates, 1 closely related - 19 non-Western Europeans - 36 cancer history/ parents birth place - 8 low call rate - 1 non-Western European - 108 cancer history/grandparents ethnicity - 8 low call rate - 72 cancer history/parents birthplace - 1 non-Western European - 2 closely related	<b>German GWA study</b> 846 cases (431 GBM) 1,310 controls (344 Heinz Nixdorf, 371 KORA, 595 PopGen)
<b>Heinz Nixdorf Recall study controls</b> (n = 380) HumanHap550		
<b>KORA controls</b> (n = 488) HumanHap550		
<b>PopGen controls</b> (n=678) HumanHap550		

# Follow-up Study

- **The four studies combined samples size:**
  - 4147 glioma cases
  - 7435 controls
  - 424,460 common tagged SNPs
  - Corrected for population substructure using principal-components analyses-Eigenstrat
  - Resulting lambda value  $<1.05$  for all studies

# EGFR



# EGFR

- **SNP rs11979158 (location 55126843) yielded p-value  $7.03 \times 10^{-8}$** 
  - **Population corrected p-value  $7.72 \times 10^{-8}$**
  - **OR = 1.23 (95% CI 1.15-1.35)**
- **SNP rs2252586 (location 54946418) yielded p-value  $7.89 \times 10^{-8}$** 
  - **Population corrected p-value  $2.09 \times 10^{-8}$**
  - **OR = 1.18 (95% CI 1.11-1.25)**

# EGFR-Two SNPs are Independent

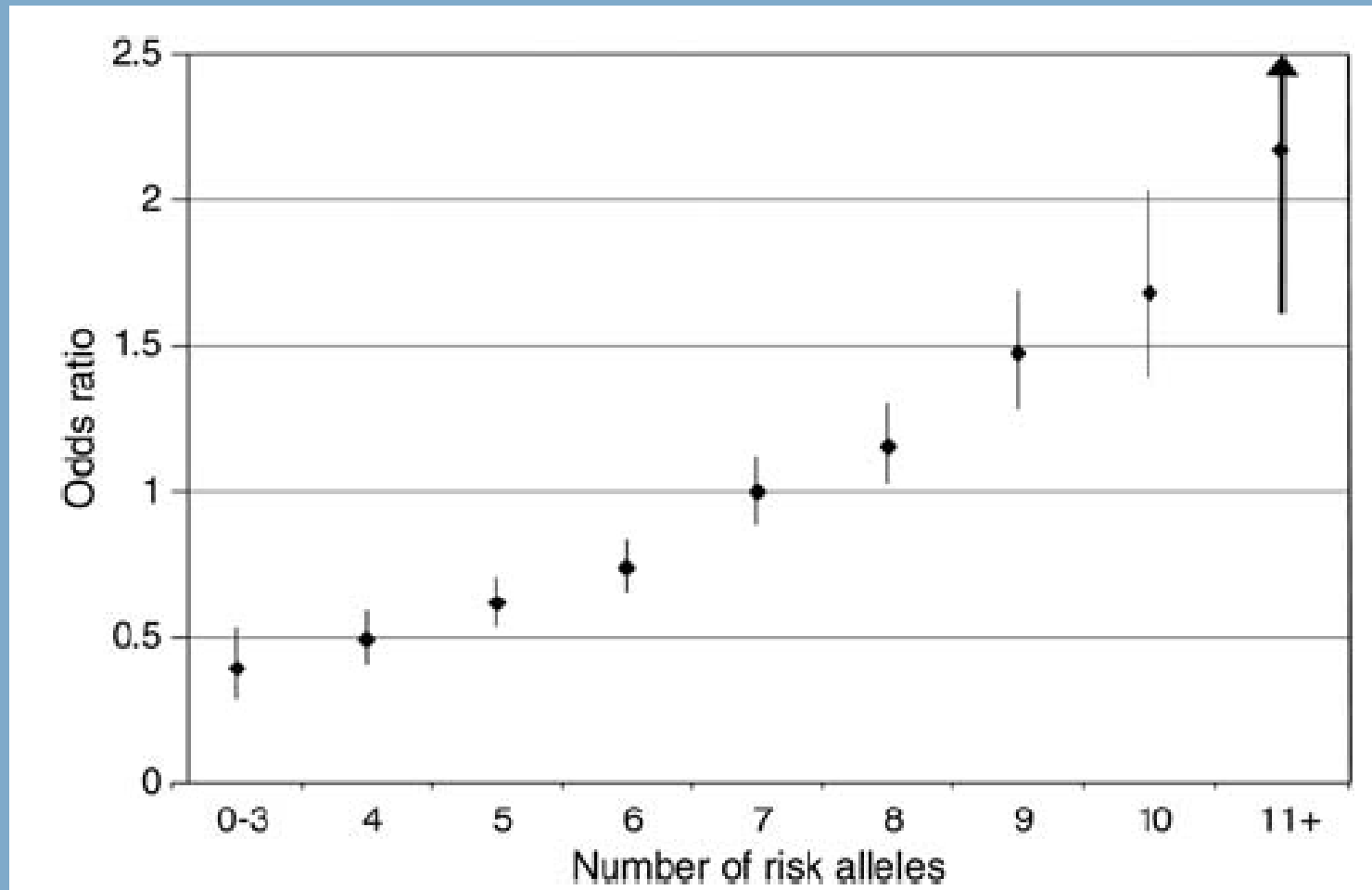
- LD between two SNPs is low ( $r^2 = 0.02$  and  $D' = 0.62$ )
- To address the question that these two SNPs may be correlated with an untyped variant: We imputed untyped SNPs from HAPMAP3 and 1000 Genomes data. No SNPs with significant better evidence of significance >> evidence of two independent risk loci at 7p11.2



## **EGFR-Two SNPs are Independent**

- **Adjusting r7s1199158 for rs2253586 and rs2253586 for r7s1199158 still provided evidence of association**

# Trend in OR with increasing number of risk alleles

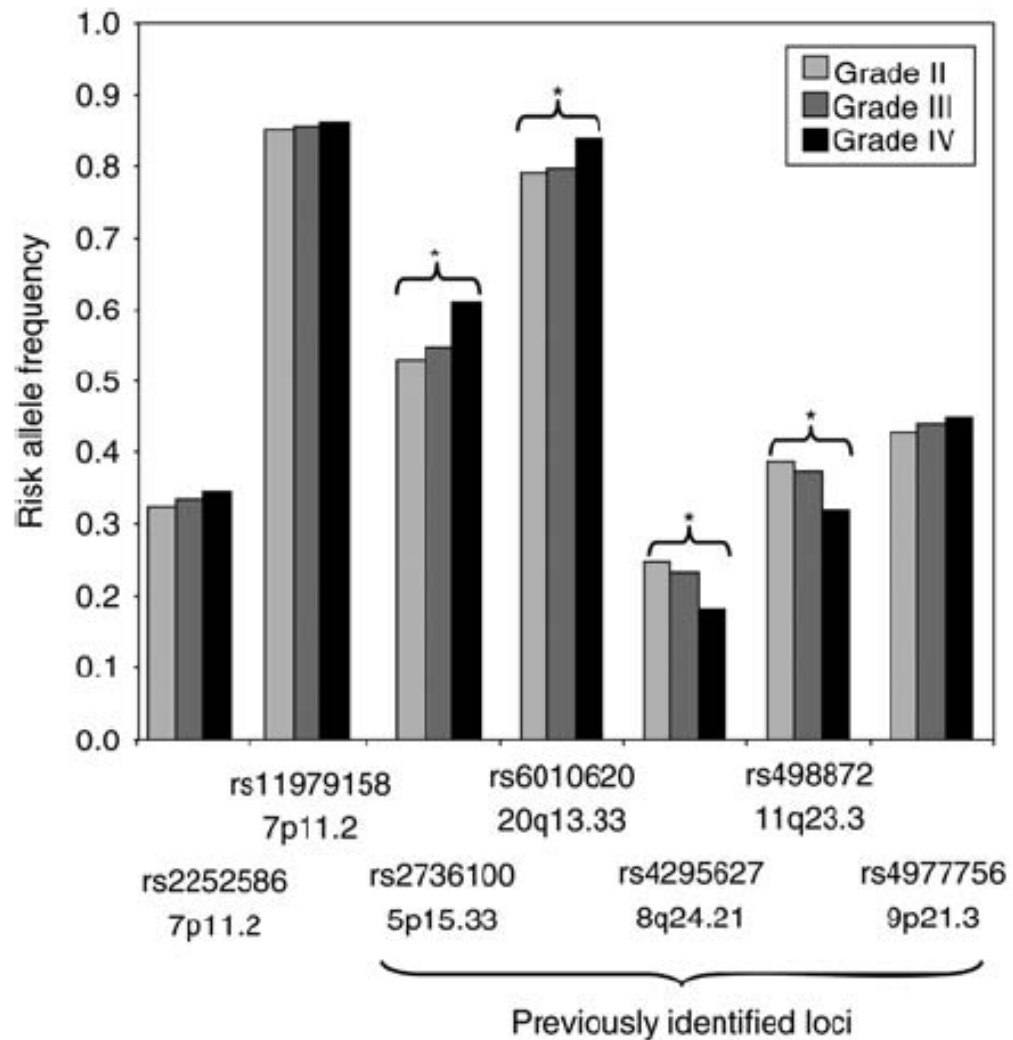


# Trend in OR

Number of risk alleles	Controls (%)	Cases (%)	OR (95% CI)
0-3	225 (3.0)	53 (1.3)	0.39 (0.29 - 0.53)
4	589 (7.9)	174 (4.2)	0.49 (0.41 - 0.59)
5	1114 (15.0)	412 (9.9)	0.62 (0.54 - 0.71)
6	1655 (22.3)	738 (17.8)	0.74 (0.66 - 0.83)
7	1637 (22.0)	984 (23.7)	1.00 (0.89 - 1.12)
8	1255 (16.9)	872 (21.0)	1.16 (1.03 - 1.30)
9	626 (8.4)	554 (13.4)	1.47 (1.28 - 1.69)
10	259 (3.5)	262 (6.3)	1.68 (1.39 - 2.03)
11+	98 (1.0)	98 (2.4)	2.17 (1.59 - 2.97)
Total	7,435	4,147	1.24 (1.21 - 1.27) $P_{\text{trend}} = 2.89 \times 10^{-72}$

Hanson et al. Human Molecular Genetics, 2011

# Tumor subtypes-n=4002



# GWAS: p-value and odds ratio

- Generally, in GWA studies, the most significant single-nucleotide polymorphisms (SNPs) associated with top-ranked p values are selected in stage one, with follow-up in stage two.
- The value of selecting SNPs based on statistically significant p values is obvious.
- However, when minor allele frequencies (MAFs) are relatively low, less-significant p values can still correspond to higher odds ratios (ORs), which might be more useful for prediction of disease status.
- Therefore, if SNPs are selected using an approach based only on significant p values, some important genetic variants might be missed.
- Wang and Shete (2011) A powerful hybrid approach to select top single-nucleotide polymorphisms for genome-wide association study *BMC Genetics* 12:3.

# Some Recent Successes

---

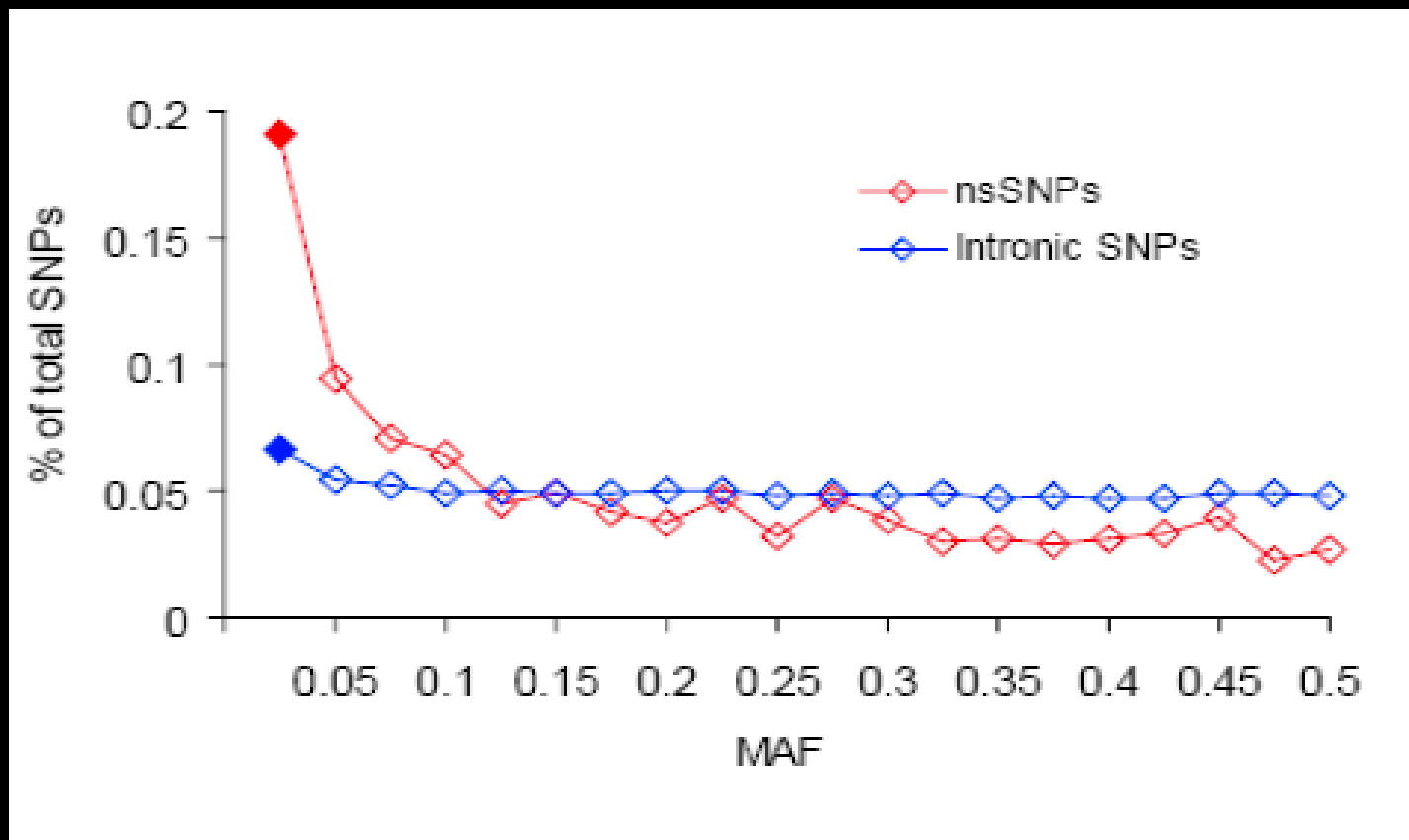
- IL23R predicts risk for Crohn's disease
- Identified via whole genome association study using 317K Illumina SNPs in 500 cases and 500 controls.
- Several SNPs in strong LD including rs11209026, c.1142G>A, Arg381Gln show strong protective (?) associations
- Replicated in additional data sets
- Duer et al., 2006 Science. 2006 314(5804):1461-3

# Ways to maximize genetic signals

---

- Genotype the functional SNPs
  - These are unknown, but SNPs with suggested functional relevance can be identified using bioinformatic tools to assess predicted impact on protein, splicing, regulation, or evolutionary conservation

# Functional SNPs tend to be rare





# Ways to maximize genetic signals

---

- Genotype the functional SNPs
- Genotype SNPs strongly associated with functional SNPs – requires a large number
  - For ‘tagging’ SNPs, Illumina suggests 317K in Caucasians, 550K for Asian and 650K for Africans. Tagging SNPs identify common SNPs but not rare SNPs. LD patterns are complex for identifying causal variants.

# Ways to Maximize Genetic Signals

---

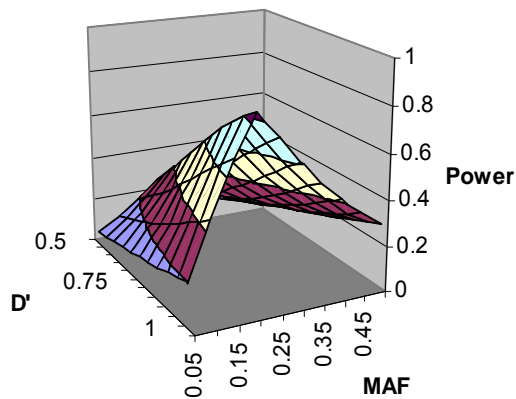
- Select genetically homogeneous subsets
  - Sample size requirements increase proportionally to square of the mixture of nongenetic or different genetic causes.
  - Presence of recurrent mutation for disease reduces the disequilibrium.
  - Studying isolated populations may lead to a more homogeneous genetic etiology.
  - Obtaining data on ancestry can protect against spurious association due to ethnic stratification
  - Type probands of families with linkage in a region

# Ways to Maximize Genetic Signals

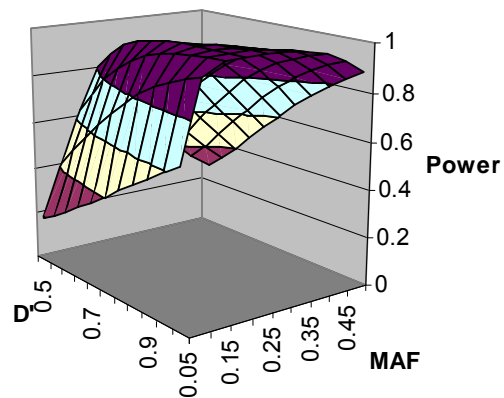
---

- Select genetically homogeneous subsets
- Select cases to be enriched for genetic causes – by sampling through cases with affected relatives

# Sample Sizes with family history



Power when selecting  
Without respect to family history  
For an additive effect  
Causal SNP variant allele  
Frequency is 0.2



Power when selecting cases with  
An affected sibling  
For an additive effect  
Causal SNP variant allele  
Frequency is 0.2

# Ways to Maximize Genetic Signals

---

- Select genetically homogeneous subsets
- Select cases to be enriched for genetic causes – by sampling through cases with affected relatives
- Evaluate quantitative traits that have high heritability
- Select controls to show less genetic influence than cases

# Design Issues

---

- Need to maximize signal due to genetic causes associated with genotyped markers
- Need to minimize experimental costs, largely reflecting genotyping costs
  - Genome wide association analysis often 15-30 times cheaper per genotyped sample than custom SNP arrays, which are usually cheaper than other polymorphisms

# Minimizing Costs

---

- Two stage designs can reduce costs – if
  - Samples in both stages are comparable
  - Costs for genotyping of custom markers in second stage are not too high compared with genome-wide analysis (about 30 fold higher may be upper limit)
  - Only a single phenotype is of interest (else how to select markers from first to second stage)

# Identifying Causal variants

---

- Because not all SNPs have been uncovered can be beneficial to perform resequencing of cases (perhaps fewer controls need resequencing)
- Role for investigating copy number variation- direct measure of genomic association (rather than indirect which occurs when using tagging SNPs).



# Statistical Methods for GWAS

---

- Main emphasis on comparison of allele frequencies comparing cases to controls
- May need to infer SNPs from several tagging SNP genotypes
- Haplotypes can provide additional information for SNPs not in strong LD with any single SNP

# Statistical Methods for GWAS

---

- Need to identify true signals from multiple tests – requires large sample sizes
- Correlation among tests can be accounted for by permutation analyses by fixing the covariance among the tests and then resampling test statistics under a null hypothesis

# Summary

---

- Large sample sizes are likely to be needed for GWAS because of the need to identify true signals from large amount of noise
- Putative functional SNPs should be included
- Need to balance costs while maintaining power currently suggests two-stage designs (may be obviated by decreasing genotyping costs)

# References

---

- Wang H, Thomas DC, Pe'er I, Stram DO. 2006. Optimal two-stage genotyping designs for genome-wide association scans. *Genet. Epidemiol.* 30:356-368.
- Zuo Y, Zou G & Zhao H (2006). Two-stage designs in case-control association analysis. *Genetics* 173: 1747-1760.
- Wang T, Elston RC. 2006. A quantitative linkage score for an association study following a linkage analysis. *BMC Genet.* 7:5.
- Roeder K, Bacanu SA, Wasserman L, Devlin B. 2006. Using linkage genome scans to improve power of association in genome scans. *Am J Hum Genet* 78:243-52.

# References

---

- Skol AD, Scott LJ, Abecasis GR, Boehnke M. 2006. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.* 38: 209-213.
- Van Steen K, McQueen MB, Herbert A, Raby B, Lyon H, DeMeo DL, Murphy A, Su J, Datta S, Rosenow C, Christman M, Silverman EK, Laird NM, Weiss ST, Lange C. 2005. Genomic screening and replication using the same data set in family-based association testing. *Nat. Genet.* 37: 683-691.
- Evans DM, Marchini J, Morris AP, Cardon LR (2006) Two-stage two-locus models in genome-wide association. *PLoS Genet* 2(9): e157.

# References

---

- Shete et al. (2009) Genome-wide association study identifies five susceptibility loci for glioma. *Nature Genetics* 41 (8):899-907
- Sanson\*, Hosking\*, Shete\* et al. (2011) Chromosome 7p11.2 (EGFR) variation influences glioma risk. *Human Molecular Genetics*, 15;20(14):2897-904.
- Wang and Shete (2010) Using Both Cases and Controls for Testing Hardy-Weinberg Proportions in a Genetic Association Study. *Human Heredity* 69:212–218
- Wang and Shete (2011) A powerful hybrid approach to select top single-nucleotide polymorphisms for genome-wide association study *BMC Genetics* 12:3.
- Shete et al. (2012) Genome-wide high-density SNP linkage search for glioma susceptibility loci: results from the Gliogene Consortium. *Cancer Research*, 71(24) 7568-7575