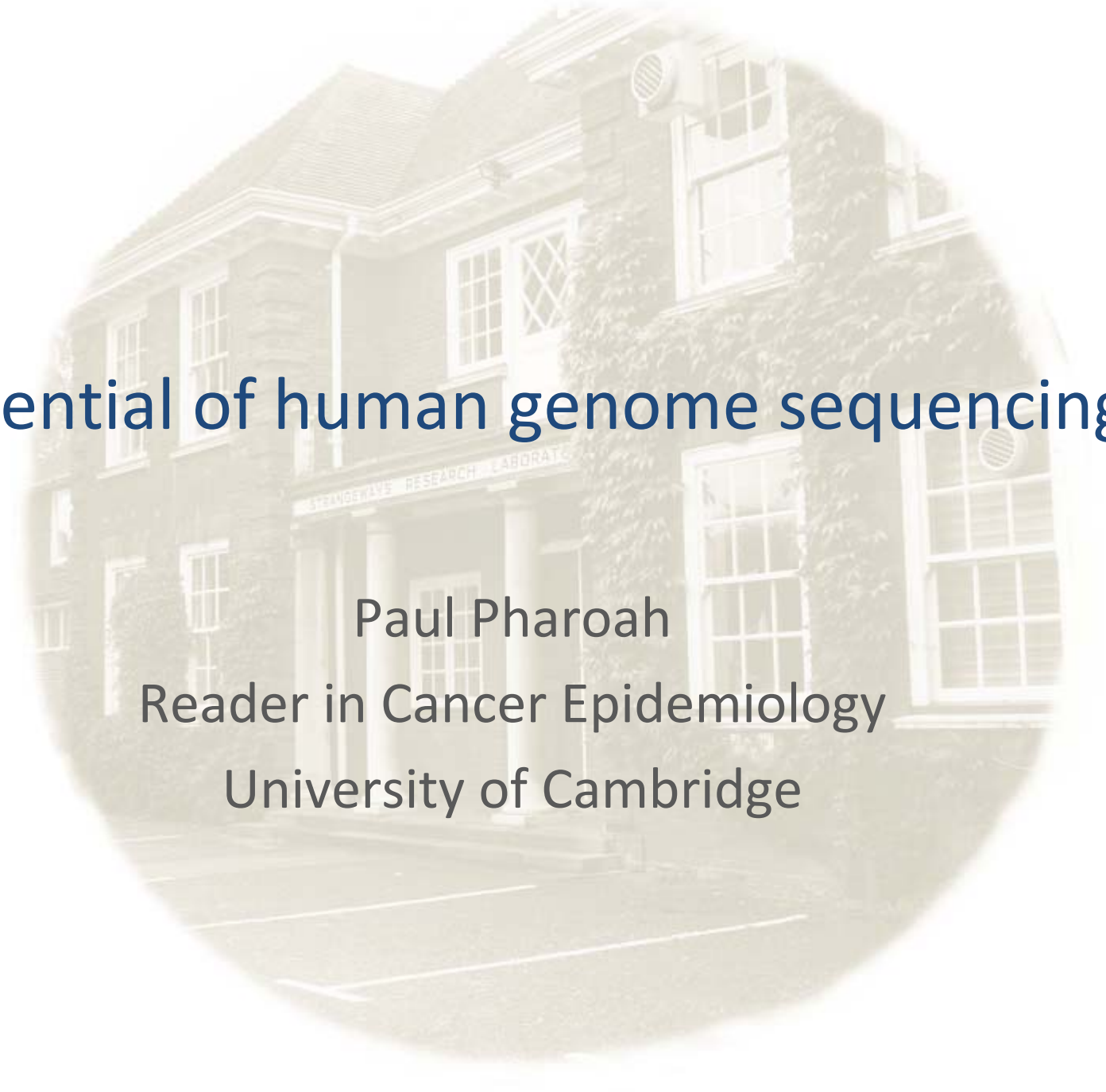


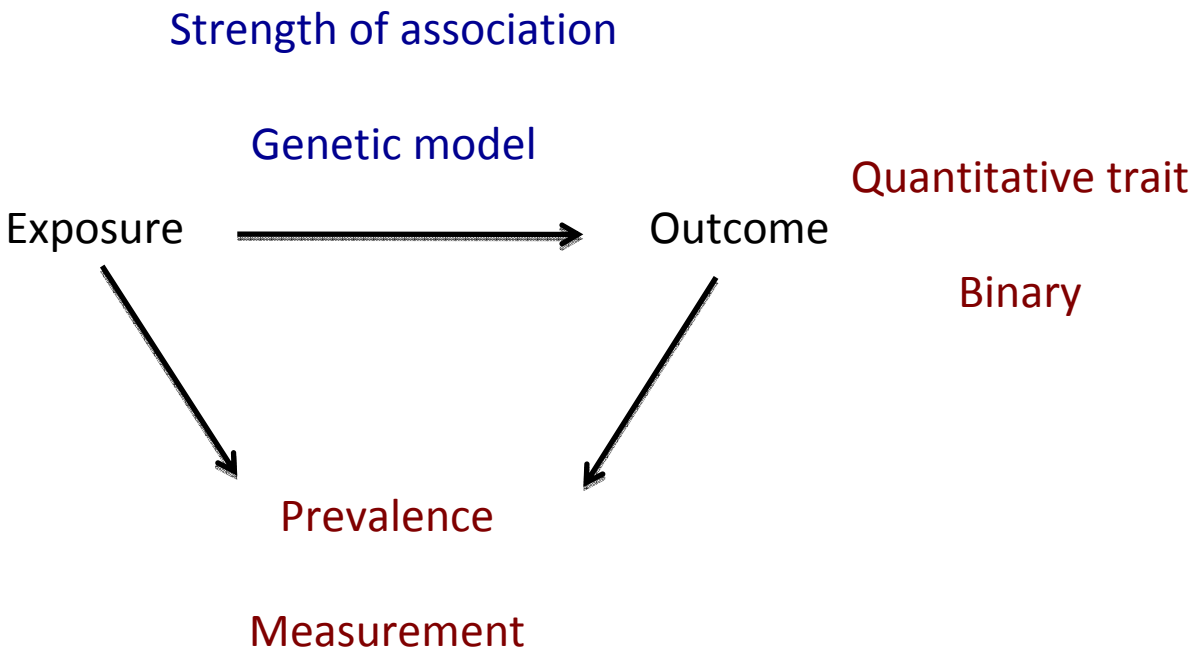
Potential of human genome sequencing

Paul Pharoah

Reader in Cancer Epidemiology

University of Cambridge

Key considerations



Timeline

Understanding of architecture of human genome

Developments in technology

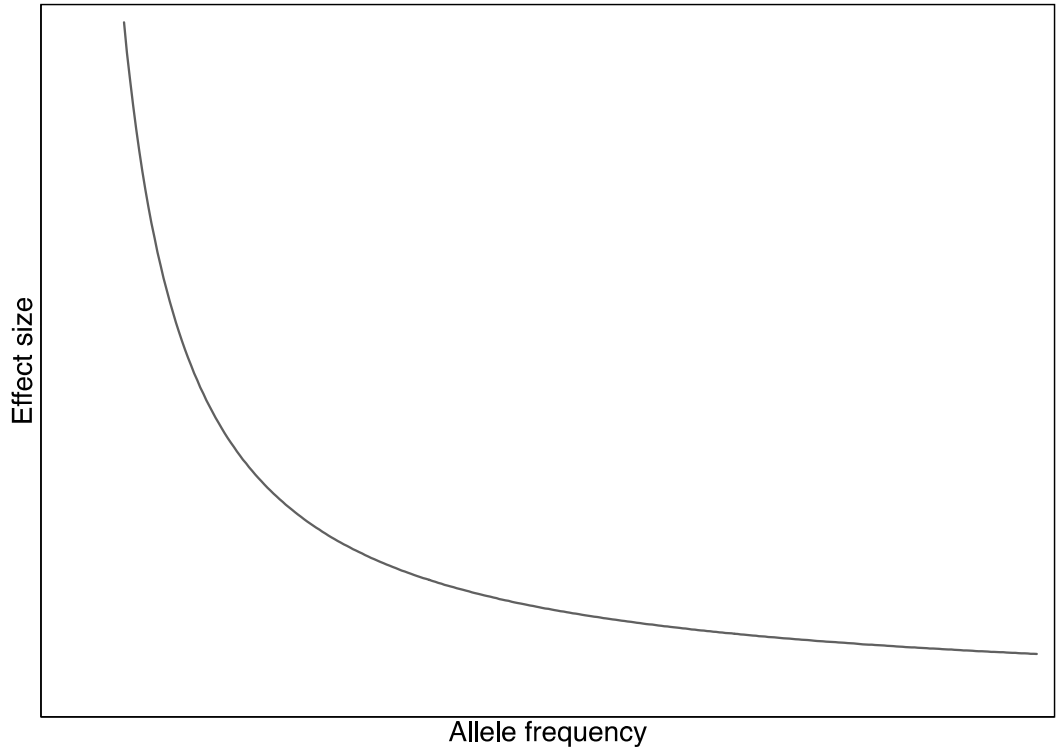
Understanding of (inherited) genetic basis of complex disease

What is meant by the genetic model?

- Allele frequency
  - Common >5%
  - Uncommon 1-5%
  - Rare 0.1 – 1%
  - Very rare <0.1%
- Mode of inheritance
  - Dominant/recessive
  - Co-dominant
- Effect size

What is likely genetic model?

- Common alleles conferring relative risk >2 not been identified for any complex phenotype
- Uncommon and rare alleles conferring modest risks (RR = 2-4) have been identified



Finding common variation for complex phenotypes

- Association study
- Compare allele frequencies in subjects with (cases) and without (controls) phenotype of interest
- Statistical methods generally straightforward

Candidate gene studies: results

- Many putative associations published
- None replicated robustly
- Inappropriate reliance on $P<0.05$ as significant

An aside on P-values

- The P-value is the probability of data IF null hypothesis is true
- It is NOT the probability that the null hypothesis is true (true negative probability)
- TNN depends on prior and power
- If prior and power are small, even a very small P is more likely to represent a false positive
- Number of tests irrelevant

Prior

- The probability of association for locus under study
- Given that a locus detectable by association must have a detectable effect size
  - explain >0.1% variance
- Thus 1,000 loci at most
- 10 million common variants
- Prior 1:10,000 at best
  Low signal to noise ratio
- May improve this by 10x with good candidate selection

Genome-wide association studies

- Possible to select a set of SNPs that tag all the common variation in genome
- Genotyping technology enable this to be done on large sample sizes
- Still too expensive to genotype everything
- Phased study design

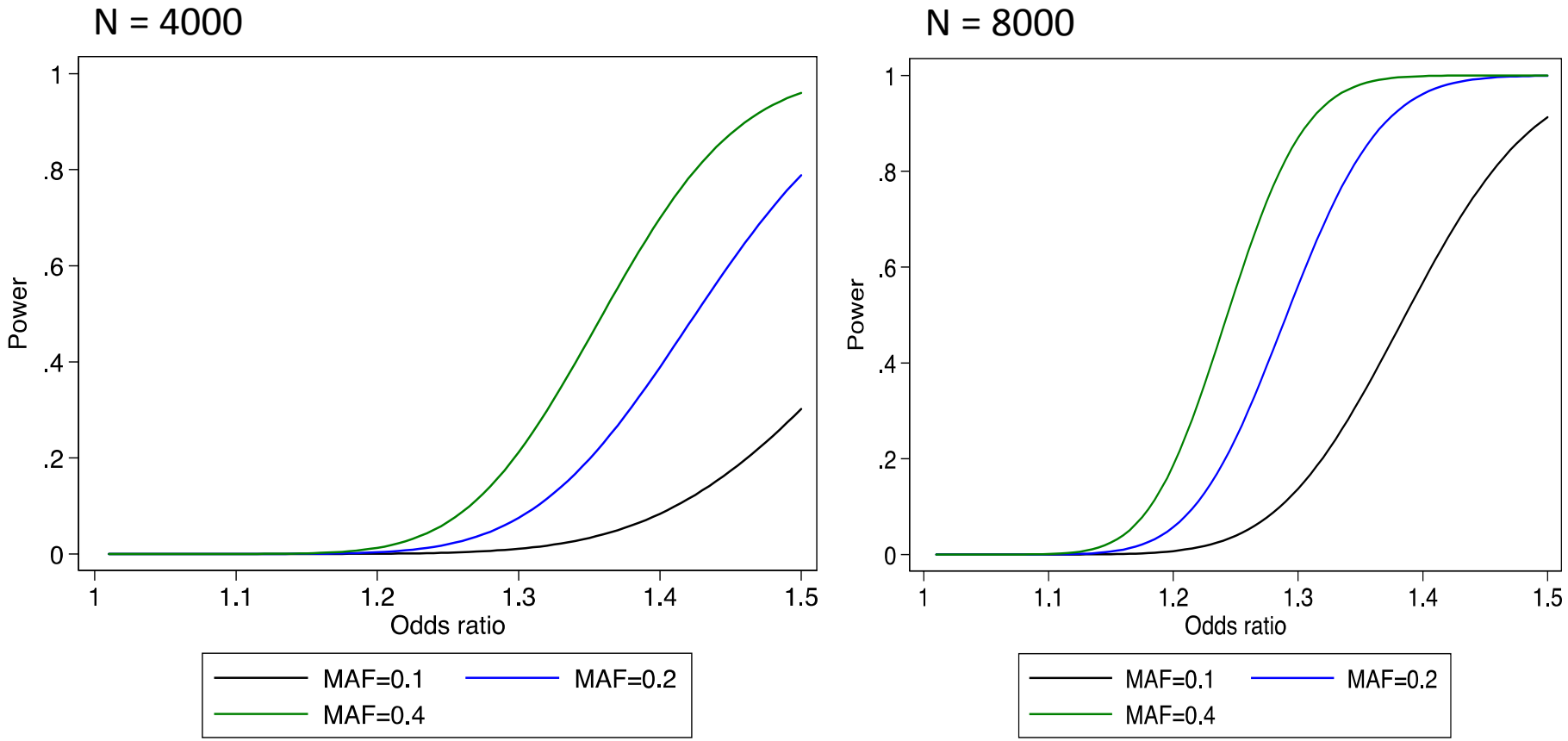
Data analysis

- Simple test of association SNP by SNP
- Genome-wide significance
  - Based on the principle of adjusting for multiple testing
  - 10M SNPs are correlated, so not 10M independent tests
  - 5×10^{-8} is approx equal to 0.05 corrected
- However, small biases can have disproportionate effects at the extremes of the distribution of test statistics
- Large samples sizes are needed

Power to detect association

Prevalence of radiosensitivity phenotype = 10 percent

$P < 5 \times 10^{-8}$



Phased GWAS designs

Genome-wide array 200k – 1M
2000 cases and 2000 controls

↓

Test for association → P>0.05

↓ P<0.05

Custom array
4000 cases and 4000 controls

↓

Test for association → P>0.00001

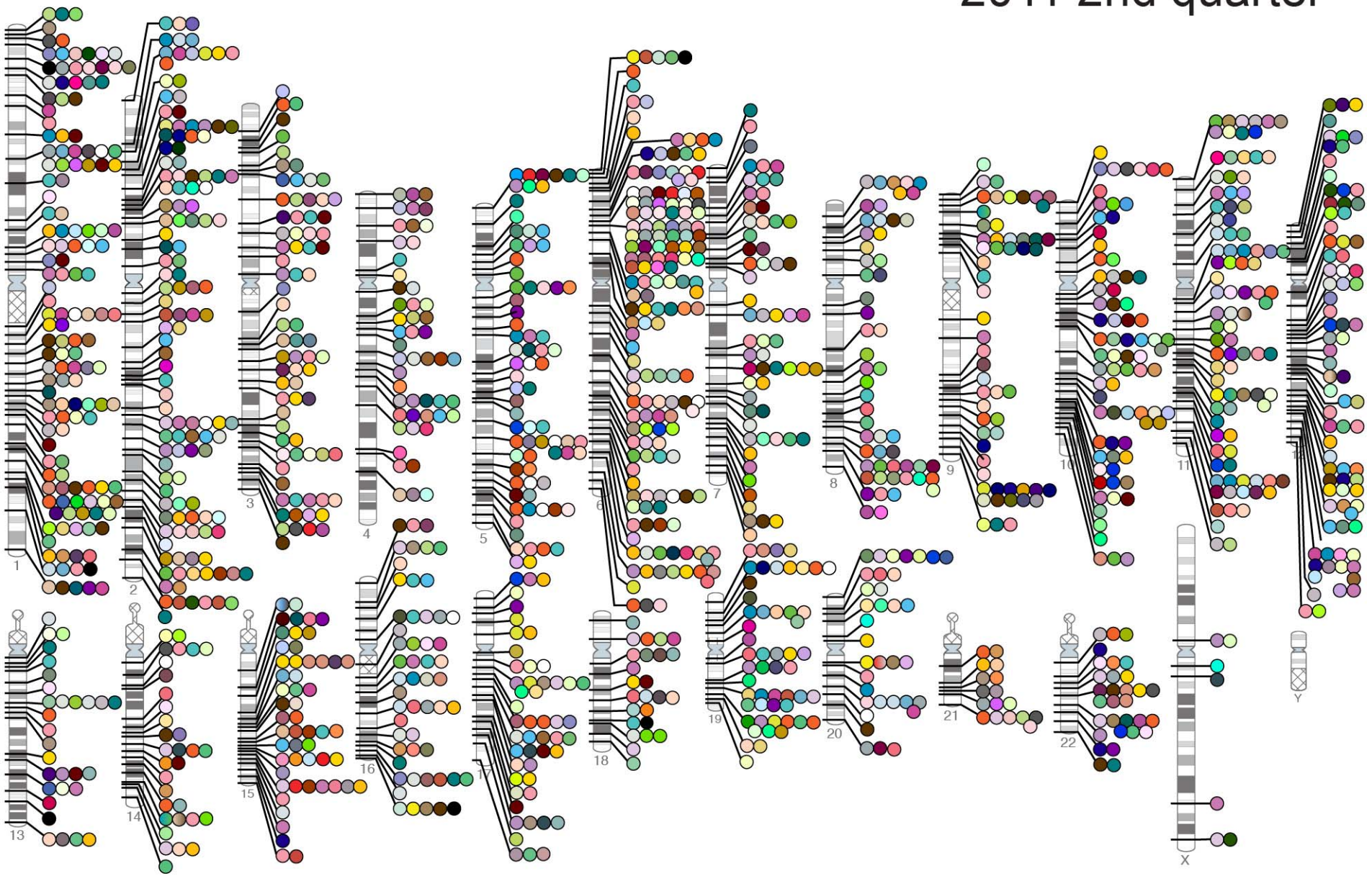
↓ P<0.00001

Top “hits”
10,000 cases and 10,000 controls

↓

SNPs with $P<10^{-8}$ declared significant

2011 2nd quarter



GWAS results

- Over 1,400 loci for 250 complex phenotypes
- Relative risks modest
- Proportion of genetic component of phenotypic variance explained <20%
- “Missing heritability”
  - Range of underlying genetic models possible
- Large scale GWAS for radiosensitivity phenotypes just being established
  - Radiogenomics Consortium

Searching for uncommon and rare variants

- Understanding of human genomic architecture continues to increase
- 1000 Genomes Project
  - Resequencing different populations
- Many studies now sequencing exomes of subjects with selected phenotypes
- ~40 million different variants detected by 1000GP

Next generation association studies

- High throughput genotyping arrays now being designed to capture rarer variation
  - Illumina Omni 5M
    - >1% variation across the genome
  - Illumina and Affymetrix exome arrays
    - ~270K variants
    - Selected from an analysis of 12,000 exomes
    - Variant present in at least 3 exomes
- Testing for association as with common variants

Problems

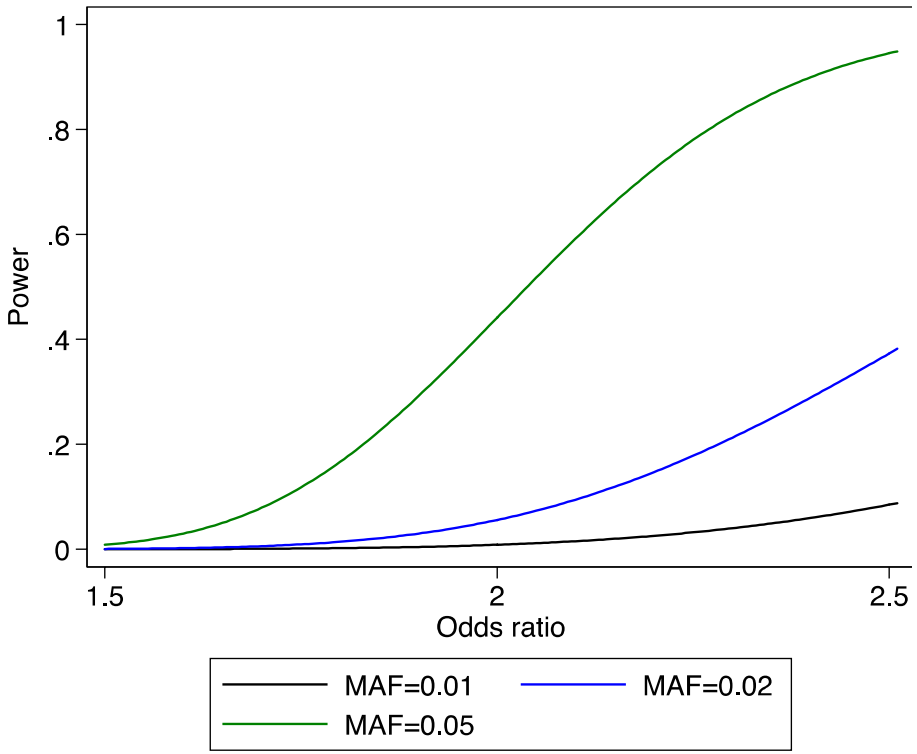
- Population substructure may cause substantial bias that is difficult to correct for using standard methods developed for analysis of common variants
- Small errors in genotype calling may result in bias
- Sample size remains a major problem

Power to detect rare variants

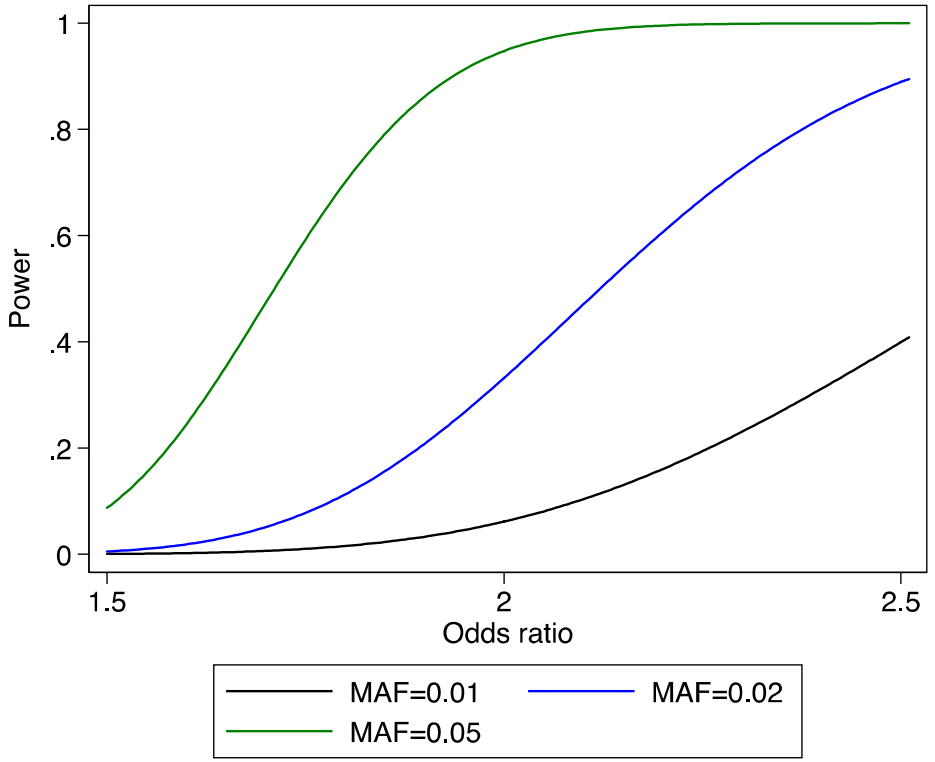
Prevalence of radiosensitivity phenotype = 10 percent

$P < 5 \times 10^{-8}$

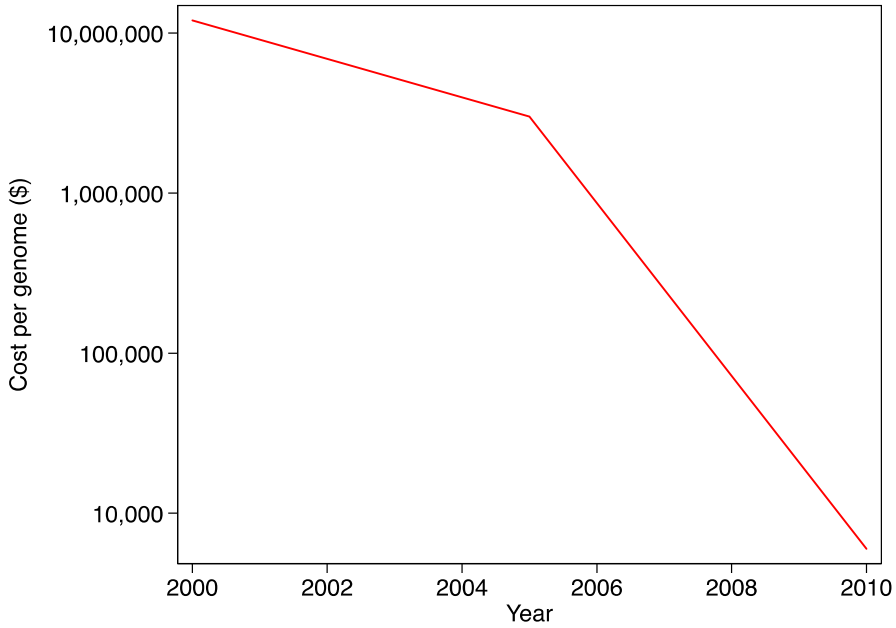
N = 4000



N = 8000



Resequencing



- Alternative to genotyping
- Whole genome sequencing now ~$5,000
- Exome sequencing $500
- Targeted sequencing 10 genes $50

Problems

- Data management
  - Very large data files
  - Exome 1-5 Gbytes
- Data analysis
  - Variant calling
  - Each genome will include very large number of variants including “private” variants of unknown function
  - Methods for statistical analysis not yet developed
- Sample size

Conclusions

- Only fraction of the inter-individual variation in radiation sensitivity explained by well known sensitivity syndromes
- Technological advances have made it possible to genotype/sequence thousands of subjects making large scale genetic epidemiological studies feasible
- Considerable obstacles still to be overcome

Questions?