

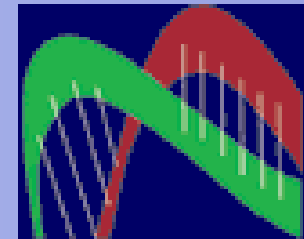


# A harmonization program for biobanks

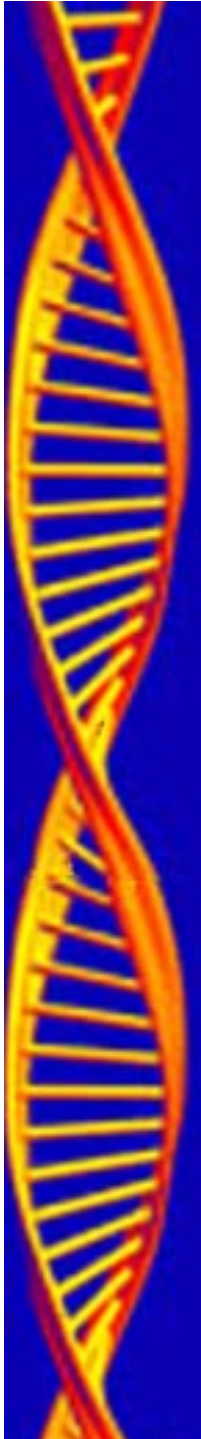
Paul Burton

Dept of Health Sciences

Dept of Genetics



University of Leicester



# A harmonization program for biobanks

- Why bother?
- What are we doing about it?

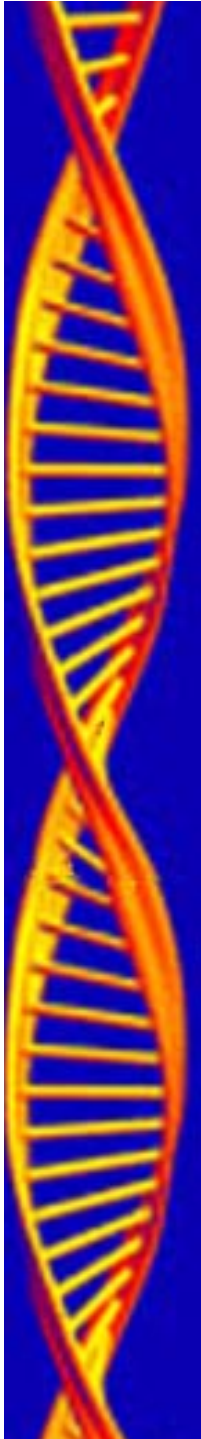


Why bother?



# We require **VERY** powerful studies

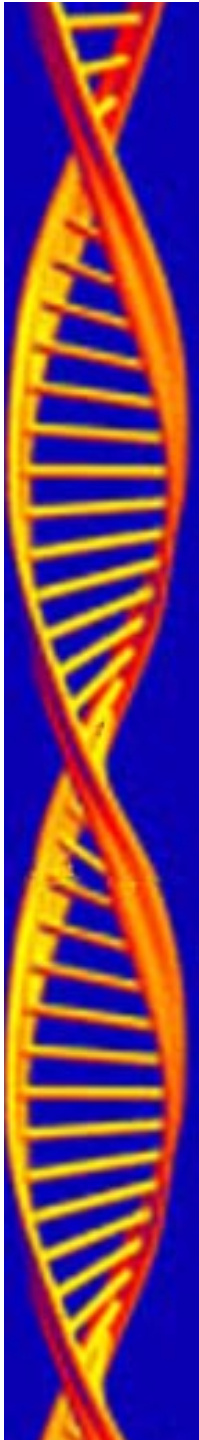
- Very large
  - Very large numbers of cases of disease
- Design must be efficient and fit for purpose
  - Case-control where appropriate
  - Cohort where necessary
- Information synthesis across studies
  - Synthesis of original data, not meta-analysis of published findings
- Study of quantitative intermediate traits\*



How large is very large?

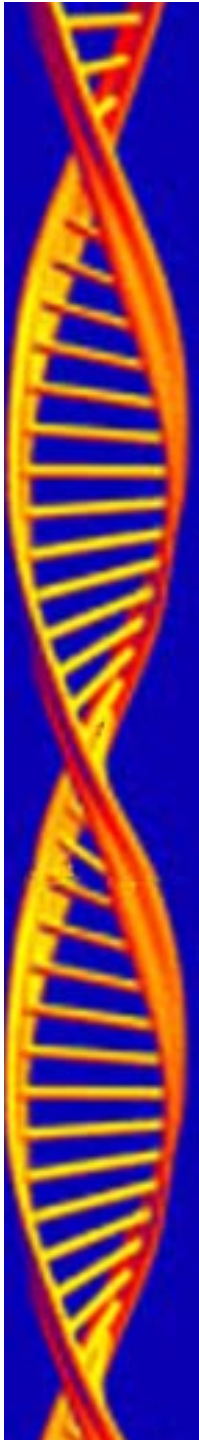
# 5,000 cases

Number of cases	Genotype prevalence	Environmental exposure prevalence	Critical P-value	Minimum detectable* OR for genetic effect	Minimum detectable* OR for environmental effect
5000	0.5	0.5	0.01	<i>1.12</i>	<i>1.12</i>
<b>5000</b>	<b>0.5</b>	<b>0.5</b>	<b><math>10^{-4}</math></b>	<b><i>1.16</i></b>	<b><i>1.16</i></b>
5000	0.5	0.5	$10^{-7}$	<i>1.22</i>	<i>1.22</i>
5000	0.25	0.25	0.01	<i>1.13</i>	<i>1.13</i>
<b>5000</b>	<b>0.25</b>	<b>0.25</b>	<b><math>10^{-4}</math></b>	<b><i>1.19</i></b>	<b><i>1.19</i></b>
5000	0.25	0.25	$10^{-7}$	<i>1.25</i>	<i>1.25</i>
5000	0.1	0.1	0.01	<i>1.19</i>	<i>1.19</i>
<b>5000</b>	<b>0.1</b>	<b>0.1</b>	<b><math>10^{-4}</math></b>	<b><i>1.27</i></b>	<b><i>1.27</i></b>
5000	0.1	0.1	$10^{-7}$	<u>1.35</u>	<u>1.35</u>
5000	0.05	0.05	0.01	<i>1.26</i>	<i>1.26</i>
<b>5000</b>	<b>0.05</b>	<b>0.05</b>	<b><math>10^{-4}</math></b>	<b><u>1.37</u></b>	<b><u>1.37</u></b>
5000	0.05	0.05	$10^{-7}$	<u>1.49</u>	<u>1.49</u>
5000	0.01	0.01	0.01	<u>1.60</u>	<u>1.60</u>
<b>5000</b>	<b>0.01</b>	<b>0.01</b>	<b><math>10^{-4}</math></b>	<b><u>1.86</u></b>	<b><u>1.86</u></b>
5000	0.01	0.01	$10^{-7}$	(2.16)	(2.16)



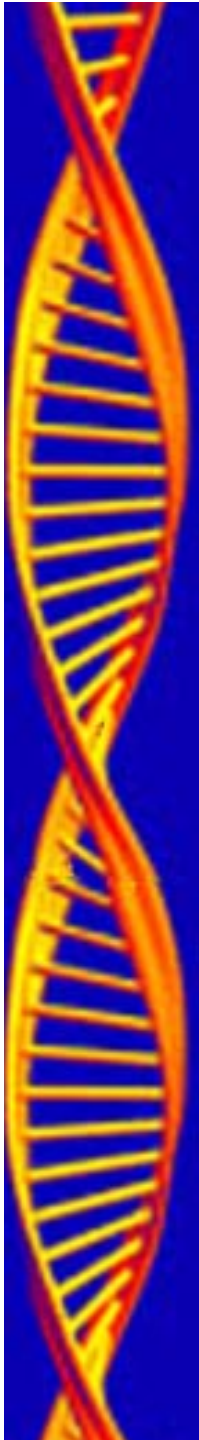
# 5,000 cases

Number of cases	Genotype prevalence	Environmental exposure prevalence	Critical P-value	Minimum detectable* OR for genetic effect	Minimum detectable* OR for environmental effect	Minimum detectable* OR for interaction effect
5000	0.5	0.5	0.01	<i>1.18</i>	<i>1.18</i>	<i>1.25</i>
<b>5000</b>	<b>0.5</b>	<b>0.5</b>	<b><math>10^{-4}</math></b>	<b><i>1.26</i></b>	<b><i>1.26</i></b>	<b><u>1.36</u></b>
5000	0.5	0.5	$10^{-7}$	<u>1.38</u>	<u>1.38</u>	<u>1.53</u>
5000	0.25	0.25	0.01	<i>1.16</i>	<i>1.16</i>	<u>1.30</u>
<b>5000</b>	<b>0.25</b>	<b>0.25</b>	<b><math>10^{-4}</math></b>	<b><i>1.22</i></b>	<b><i>1.22</i></b>	<b><u>1.43</u></b>
5000	0.25	0.25	$10^{-7}$	<i>1.30</i>	<i>1.30</i>	<u>1.57</u>
5000	0.1	0.1	0.01	<i>1.20</i>	<i>1.20</i>	<u>1.62</u>
<b>5000</b>	<b>0.1</b>	<b>0.1</b>	<b><math>10^{-4}</math></b>	<b><i>1.28</i></b>	<b><i>1.28</i></b>	<b><u>1.87</u></b>
5000	0.1	0.1	$10^{-7}$	<u>1.37</u>	<u>1.37</u>	(2.17)
5000	0.05	0.05	0.01	<i>1.27</i>	<i>1.27</i>	(2.26)
<b>5000</b>	<b>0.05</b>	<b>0.05</b>	<b><math>10^{-4}</math></b>	<b><u>1.38</u></b>	<b><u>1.38</u></b>	<b>(2.82)</b>
5000	0.05	0.05	$10^{-7}$	<u>1.50</u>	<u>1.50</u>	(3.55)



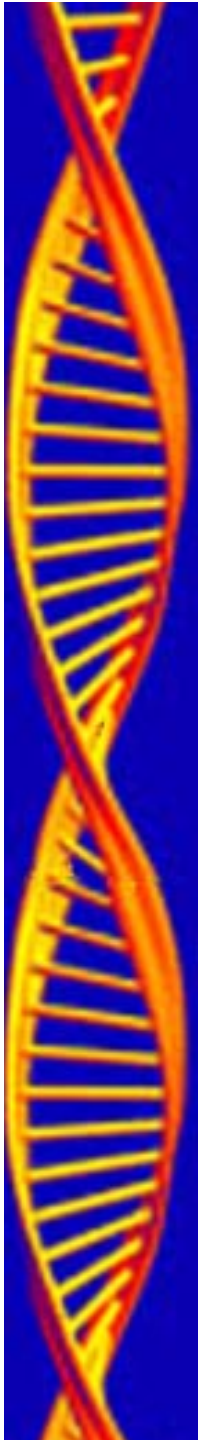
# 10,000 cases

Number of cases	Genotype prevalence	Environmental exposure prevalence	Critical P-value	Minimum detectable* OR for genetic effect	Minimum detectable* OR for environmental effect
10000	0.5	0.5	0.01	<i>1.08</i>	<i>1.08</i>
<b>10000</b>	<b>0.5</b>	<b>0.5</b>	<b><math>10^{-4}</math></b>	<b><i>1.11</i></b>	<b><i>1.11</i></b>
10000	0.5	0.5	$10^{-7}$	<i>1.15</i>	<i>1.15</i>
10000	0.25	0.25	0.01	<i>1.09</i>	<i>1.09</i>
<b>10000</b>	<b>0.25</b>	<b>0.25</b>	<b><math>10^{-4}</math></b>	<b><i>1.13</i></b>	<b><i>1.13</i></b>
10000	0.25	0.25	$10^{-7}$	<i>1.17</i>	<i>1.17</i>
10000	0.1	0.1	0.01	<i>1.13</i>	<i>1.13</i>
<b>10000</b>	<b>0.1</b>	<b>0.1</b>	<b><math>10^{-4}</math></b>	<b><i>1.18</i></b>	<b><i>1.18</i></b>
10000	0.1	0.1	$10^{-7}$	<i>1.24</i>	<i>1.24</i>
10000	0.05	0.05	0.01	<i>1.18</i>	<i>1.18</i>
<b>10000</b>	<b>0.05</b>	<b>0.05</b>	<b><math>10^{-4}</math></b>	<b><i>1.26</i></b>	<b><i>1.26</i></b>
10000	0.05	0.05	$10^{-7}$	<u>1.34</u>	<u>1.34</u>
10000	0.01	0.01	0.01	<u>1.41</u>	<u>1.41</u>
<b>10000</b>	<b>0.01</b>	<b>0.01</b>	<b><math>10^{-4}</math></b>	<b><u>1.58</u></b>	<b><u>1.58</u></b>
10000	0.01	0.01	$10^{-7}$	<u>1.78</u>	<u>1.78</u>



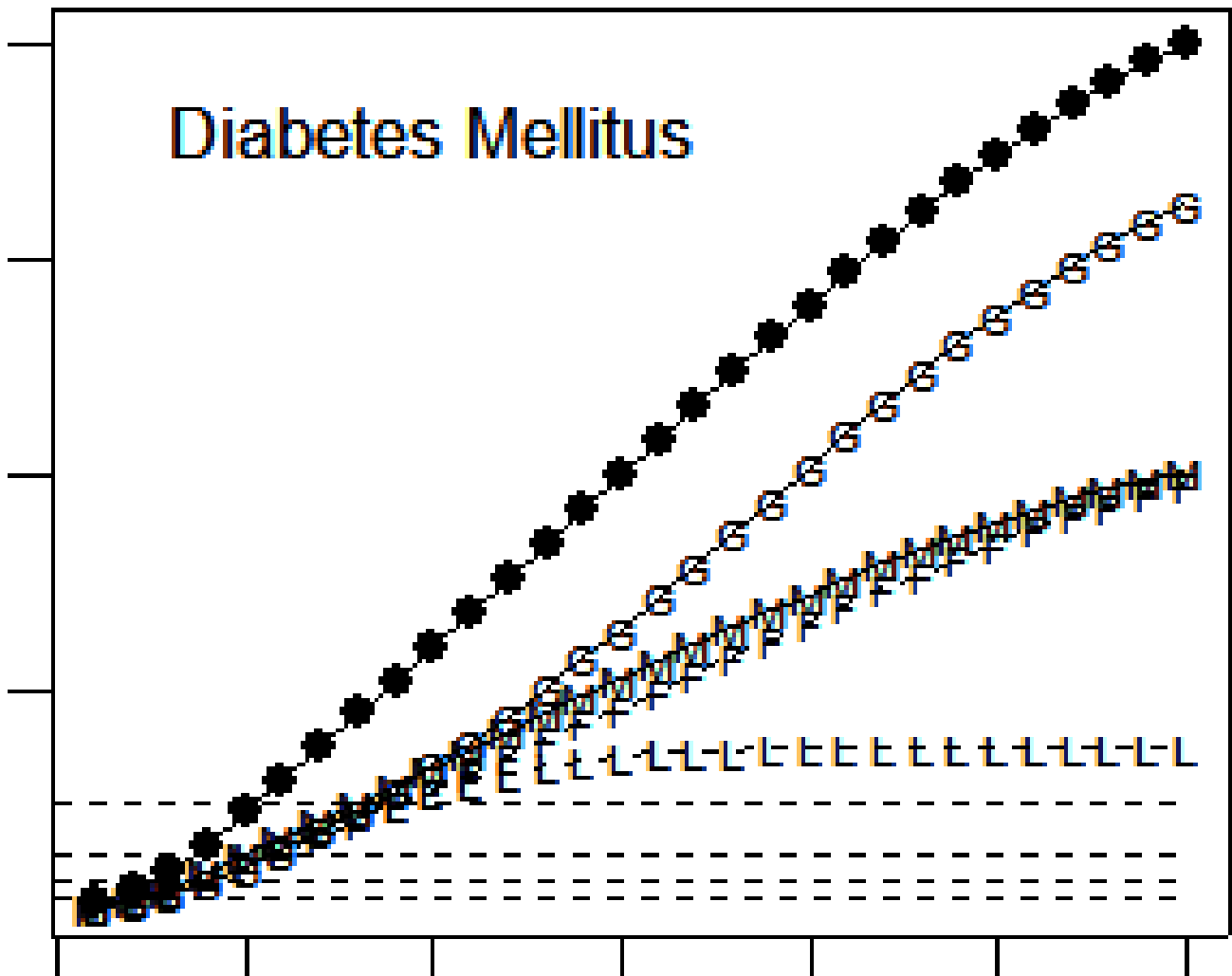
# 10,000 cases

Number of cases	Genotype prevalence	Environmental exposure prevalence	Critical P-value	Minimum detectable* OR for genetic effect	Minimum detectable* OR for environmental effect	Minimum detectable* OR for interaction effect
10000	0.5	0.5	0.01	<i>1.12</i>	<i>1.12</i>	<i>1.17</i>
<b>10000</b>	<b>0.5</b>	<b>0.5</b>	<b><math>10^{-4}</math></b>	<b><i>1.18</i></b>	<b><i>1.18</i></b>	<b><i>1.24</i></b>
10000	0.5	0.5	$10^{-7}$	<i>1.24</i>	<i>1.24</i>	<u>1.34</u>
10000	0.25	0.25	0.01	<i>1.11</i>	<i>1.11</i>	<i>1.21</i>
<b>10000</b>	<b>0.25</b>	<b>0.25</b>	<b><math>10^{-4}</math></b>	<b><i>1.15</i></b>	<b><i>1.15</i></b>	<b><i>1.30</i></b>
10000	0.25	0.25	$10^{-7}$	<i>1.20</i>	<i>1.20</i>	<u>1.40</u>
10000	0.1	0.1	0.01	<i>1.14</i>	<i>1.14</i>	<u>1.44</u>
<b>10000</b>	<b>0.1</b>	<b>0.1</b>	<b><math>10^{-4}</math></b>	<b><i>1.20</i></b>	<b><i>1.20</i></b>	<b><u>1.62</u></b>
10000	0.1	0.1	$10^{-7}$	<i>1.26</i>	<i>1.26</i>	<u>1.81</u>
10000	0.05	0.05	0.01	<i>1.19</i>	<i>1.19</i>	<u>1.85</u>
<b>10000</b>	<b>0.05</b>	<b>0.05</b>	<b><math>10^{-4}</math></b>	<b><i>1.26</i></b>	<b><i>1.26</i></b>	<b>(2.24)</b>
10000	0.05	0.05	$10^{-7}$	<u>1.35</u>	<u>1.35</u>	(2.69)



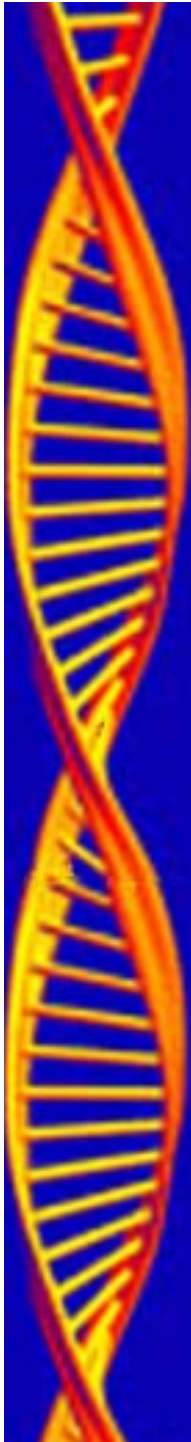
NUMBER OF CASES

80K  
60K  
40K  
20K



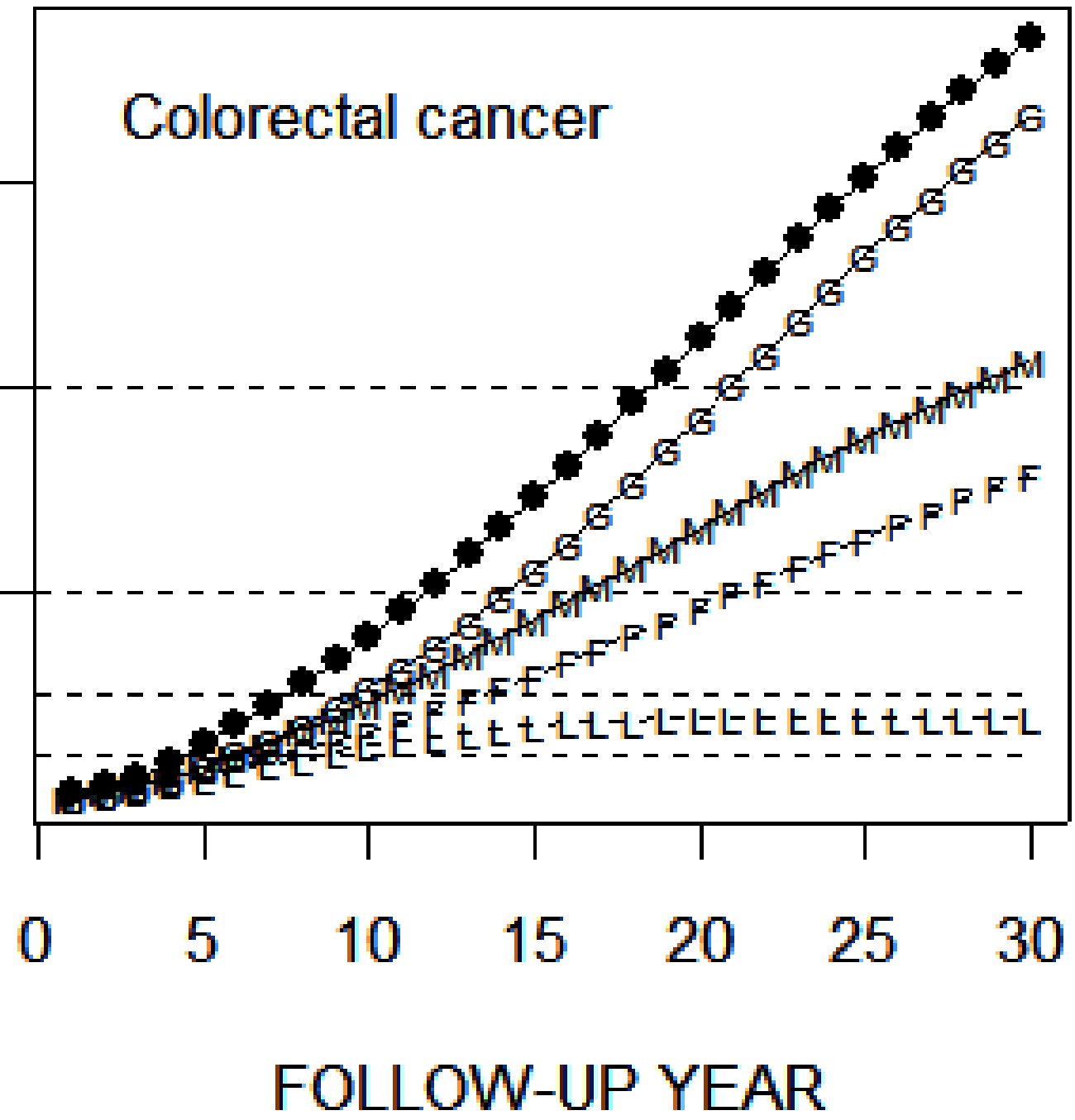
0 5 10 15 20 25 30

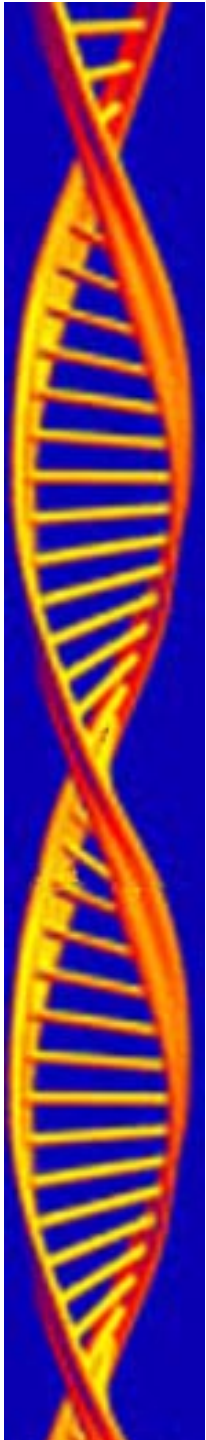
FOLLOW-UP YEAR



NUMBER OF CASES

15K  
10K  
5K





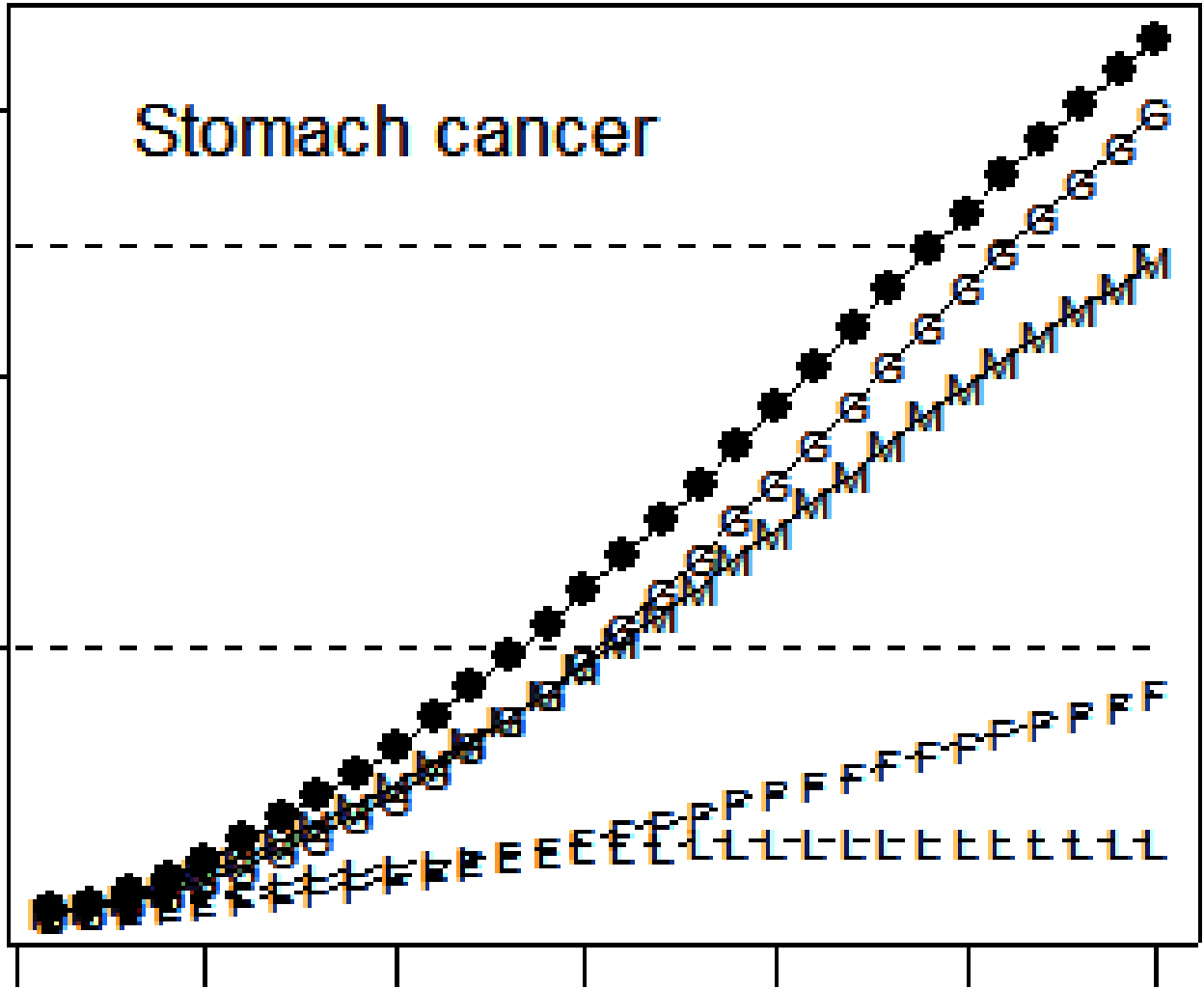
NUMBER OF CASES

3K  
2K  
1K

Stomach cancer

0 5 10 15 20 25 30

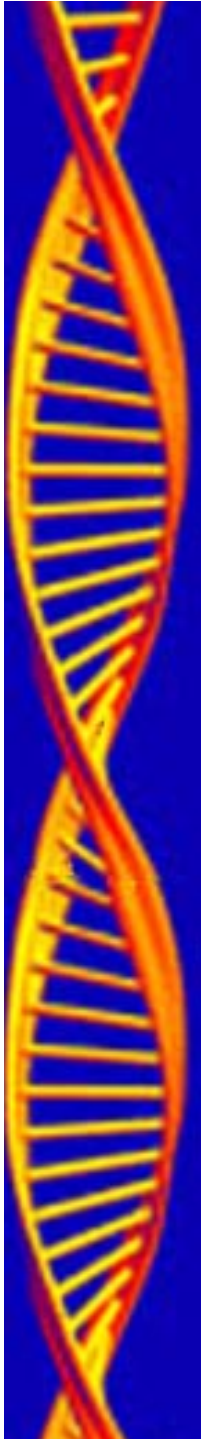
FOLLOW-UP YEAR



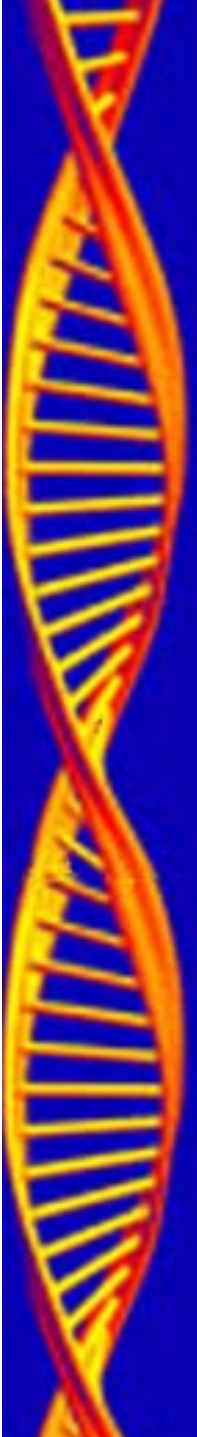
# With misclassification

Number of cases	GENO prevalence	ENV exposure prevalence	GENO error rate	ENV error rate	Disease to non-diseased error rate	Non-diseased to diseased error rate	Critical P-value	MDOR GENO	MDOR ENV	MDOR INT
<b>5,000</b>	0.1	0.1	0%	0%	0%	0%	$10^{-4}$	<b>1.28</b>	<b>1.28</b>	<b>1.90</b>
<b>5,000</b>	0.1	0.1	5%	5%	0%	0%	$10^{-4}$	<b>1.37</b>	<b>1.37</b>	<b>2.31</b>
<b>5,000</b>	0.1	0.1	0%	0%	50%	0.5%	$10^{-4}$	<b>1.52</b>	<b>1.52</b>	<b>2.29</b>
<b>5,000</b>	0.1	0.1	5%	5%	50%	0.5%	$10^{-4}$	<b>1.66</b>	<b>1.66</b>	<b>2.62</b>

Number of cases	GENO prevalence	ENV exposure prevalence	GENO error rate	ENV error rate	Disease to non-diseased error rate	Non-diseased to diseased error rate	Critical P-value	MDOR GENO	MDOR ENV	MDOR INT
<b>20,000</b>	0.1	0.1	0%	0%	0%	0%	$10^{-4}$	<b>1.14</b>	<b>1.14</b>	<b>1.42</b>
<b>20,000</b>	0.1	0.1	5%	5%	0%	0%	$10^{-4}$	<b>1.18</b>	<b>1.18</b>	<b>1.65</b>
<b>20,000</b>	0.1	0.1	0%	0%	50%	0.5%	$10^{-4}$	<b>1.26</b>	<b>1.26</b>	<b>1.62</b>
<b>20,000</b>	0.1	0.1	5%	5%	50%	0.5%	$10^{-4}$	<b>1.30</b>	<b>1.30</b>	<b>1.98</b>



How can we achieve the  
numbers required?



# Large case-control series

- Large case series that can be compared to a large “universal” control series
  - Wellcome Trust Case Control Consortium
    - MRC Biomedical DNA Collections
    - 1958 Birth Cohort
- Very cost effective provided interest does not include joint effect of genes with an environmental determinant that must be assessed pre-morbidly



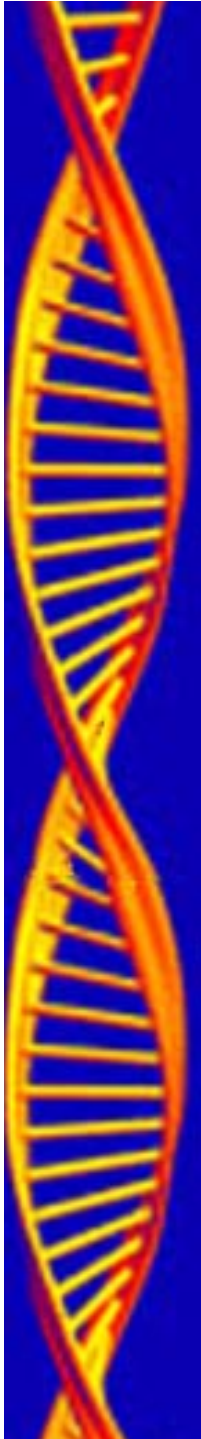
# Biobanks

- The systematic collection of biological tissue from a large number of individuals usually associated with additional information about each subject from whom the tissue was derived
  - Disease focused biobanks
  - Exposure focused biobanks
  - *Population-based biobanks*
  - Health care and *research objectives*



# Population-based biobanks

- Minimalist biobanks
  - Store biological tissue from which DNA can be extracted. Track subjects using routine health surveillance mechanisms. Possibly link data sets to other information sources.
    - deCODE Iceland
    - Estonian Genome Project
- Cost effective provided interest does not include joint effect of genes with an environmental determinant that must be assessed pre-morbidly and cannot be obtained from linkage to other sources of data.



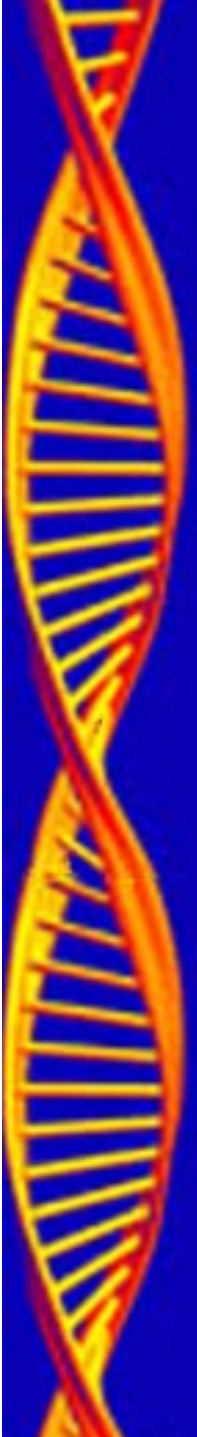
# Population-based biobanks

- Cohort study biobanks
  - Store biological tissue from which DNA can be extracted. Track subjects using routine health surveillance mechanisms. Possibly link data sets to other information sources. ***Attempt to optimise collection of life-style and environmental information at recruitment (and possibly at repeated reassessments at a later stage)***



# Cohort study biobanks

- Danish Birth Cohort
- EPIC
- Norwegian Cohorts
  - HUNT
  - MOBA
  - CONOR
- ProtecT
- UK Biobank
- Western Australian Genome Project



# Cohort study → biobank conversions

- Busselton Health Study
- Framingham Heart Study
- GenomeEUtwin
  - Twin registries
  - MORGAM
- 1958, 1946 Birth Cohorts in UK



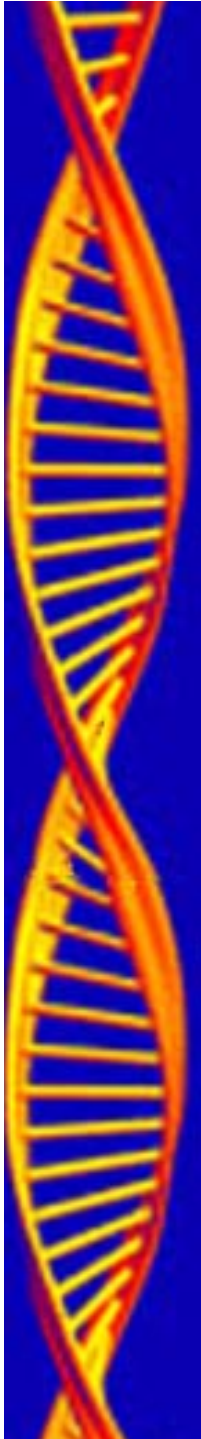
# Cohort study biobanks

- Expensive but essential if primary interest includes direct effects of environmental determinants or joint effect of genes with environmental determinants
  - Enable pre-morbid life-style assessment
  - Minimise recall bias
  - Avoid issue of reverse causality



# Cohort study biobanks

- Nested case-control studies
  - Focus on joint effects of genes and environment, but once set up also provide a cheap infrastructure to support studies of genes only
- Genotype-based studies
  - Recall on genotype and study continuous intermediate phenotypes
  - Even rare genotypes may be studied. In UK Biobank, 1,000 recruits will be expected to exhibit a genotype as rare as 1/500.

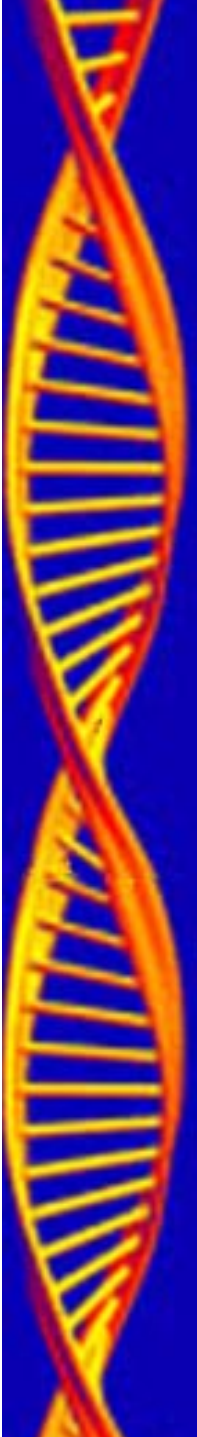


# The biobank harmonization program



# Why bother?

- Clear benefit of synthesising information between population-based biobanks
  - Hit critical target numbers more quickly
  - Hit critical target numbers in population subgroups
  - Hit critical target numbers for rarer diseases
  - Enable more powerful selection of homogeneous disease subtypes
  - More powerful genotype-based studies



# Coordination action (CA) under EU framework 6

- “Harmonising population-based biobanks and cohort studies to strengthen the foundation of European biomedical science in the post-genome era”
- Camilla Stoltenberg, Leena Peltonen, Paul Burton
  - 18 institutions from 13 European nations + Canada
- Direct outcome of EU-funded “COGENE”



# Work packages

- Opportunities for Future Biobanking in Europe
  - Database and Biobank Information Systems
  - DNA/SNPs and Genotyping
  - Questionnaires and Clinical Measures
  - Ethical and Societal Issues
  - Statistical Methods
- 
- Scope out key issues and identify solutions (if any) to problems that may emerge.



# Statistical methods

- Formal strategies for data-synthesis in the genetic epidemiological setting
- Single Nucleotide Polymorphisms (SNPs) and the construction and analysis of haplotypes
- Problems arising from population stratification and admixture
- Approaches to analysis that allow for non-independence of sampling units
- Modifications and extensions of conventional nested case-control designs
- Flexible platforms for developing and implementing new statistical methodologies.