THE FEASIBILITY OF ESTABLISHING AN ALL-IRELAND BIOBANK

REPORT OF AN EXPERT GROUP

May 2004

TABLE OF CONTENTS

		Page
Summary		4
1.	**Introduction**	5
	1.1 Background	5
	1.2 Membership of Expert Group	6
	1.3 Aim and terms of reference of group	6
	1.4 Meetings and action plan of group	6
2.	**Scientific rationale and study designs for biobanks**	6
	2.1 Scientific rationale	6
	2.2 Study design and collection strategies	7
	2.3 Case sample and catchment population	8
3.	**Outline proposal for all-Ireland biobank**	9
	3.1 Objectives	9
	3.2 Sample size and age distribution	9
	3.3 Sampling frame	9
	3.4 Data collection by questionnaire	10
	3.5 Clinical examinations	10
	3.6 Notification of results	11
	3.7 Biochemical analyses	11
	3.8 Banking of samples	11
	3.9 Location of resource	11
4.	**Project management**	12
	4.1 Management structure	12
	4.2 Sourcing of work	12
	4.3 Equipment and infrastructure required	12
	4.4 Timetable	12
5.	**Ethical and legal considerations**	13
	5.1 Background	13
	5.2 Health Survey	13
	5.3 Collection and use of the DNA samples	13

6. Estimated costs 14
6.1 Costs of Survey 14
6.2 Project management and data analyses 14
6.3 Fieldwork 14
6.4 Laboratory analyses 14
6.5 Extraction and banking of DNA 14
6.6 Cost of proposed initial genotyping studies 14

7. Relationship with other projects 15
7.1 Disease collections with DNA in Republic of Ireland and Northern Ireland 15
7.2 International collaborations 16
7.3 Biobank UK 16
7.4 Other health monitoring surveys in Ireland 16

8. Governance of biobank 16
8.1 Custodianship of Biobank 16
8.2 Access to resource 17
8.3 Costs of DNA analyses 17

9. Cost of maintaining biobank 17

10. Benefits of an all-Ireland biobank 17

11. Conclusions 17

Tables

Table 1: Catchment populations that generate 10,000 cases of disease 8

Table 2: Estimated costs of establishing biobank 15

References 18

Appendix 1 19

Summary

This report examines the feasibility and value of establishing a collection of DNA samples together with phenotypic data from the population of the island of Ireland, as a resource for researchers studying the genetic basis of disease susceptibility and health outcomes.

We conclude that the most useful resource would be a collection based on a cross-sectional study of 10,000 adults aged 25-74 years with detailed phenotypic measurements. For genetic researchers this would provide a resource for two types of study: (1) association studies of genetic factors and quantitative traits measured at baseline or in stored plasma or urine samples; (2) case-control studies, for which it would provide a shared control group. It would also be useful as a baseline health monitoring survey, supplementing previous surveys in the Republic of Ireland and Northern Ireland, which have not included physical examination or biological samples.

We propose a three-stage sampling protocol, with 20 primary sampling units based on administrative units and secondary sampling units defined by households. We estimate that for a final sample size of 10,000 individuals with blood sampling and phenotypic measurements, it will be necessary to invite 20,000 individuals. We propose that the collection of questionnaire data and physical measurements should be based on a subset of the measurements used in the US National Health and Nutrition Examination Survey, including cardiovascular risk factors, respiratory function measurements, measurement of body composition, a blood sample and an overnight urine sample. Relevant questions from health-monitoring surveys in Ireland e.g. Northern Ireland Health and Wellbeing Survey and Slán can also be included in the questionnaire for health monitoring purposes. Some initial genetic studies of population structure and associations with traits of interest should be undertaken as soon as sample collection is complete. For subsequent studies users should be expected to meet the costs of genotyping.

To keep the costs down and to ensure that fieldwork can be completed quickly, we recommend that the fieldwork should be outsourced to a survey agency. Scientific coordination should be undertaken by a small group that is directly accountable to the funding agencies. The costs of the project depend upon what measurements are included in the study protocol; a provisional estimate is about €3.2 million. Planning for this biobank should be coordinated with plans to establish large case collections based on the all-Ireland population.

Governance of the biobank should be by an independent Board of Management which would have responsibility for arrangements for storage of data and biological samples and which, through a scientific committee, would regulate access to the resource.

1. Introduction

1.1 Background

The availability of the complete sequence of the human genome, annotated with a comprehensive catalogue of genes and sites of sequence variation (polymorphisms), opens up new possibilities for discovering genetic effects on human health. To study genetic effects on human health large collections of DNA samples and clinical data are required. The term "biobank" is now used to denote such collections. Several countries have begun to establish such biobanks. One of the largest projects is Biobank UK, first proposed in 2000 by a panel convened by the Wellcome Trust and UK Medical Research Council (MRC). To study the combined effects of genotype and environmental exposures on disease risk, the panel recommended that a prospective study should be established in which 500,000 individuals aged 45-69 would be followed over ten years.

The Research and Development Office (Northern Ireland) took part in initial discussions on the UK Biobank and considered participation in the project. However following consultation with the Health Research Board (Republic of Ireland) it was decided to examine the feasibility of establishing an all-Ireland biobank. The Health Research Board also expressed an interest in establishing a link between an all-Ireland biobank and Biobank UK, without necessarily following the same research design as the UK Biobank.

Professor Bernadette Herity, Emeritus Professor of Public Health Medicine and Epidemiology at University College Dublin (UCD) and a former Board Member of the Health Research Board, was invited to convene and chair a small expert group, which would examine aspects of establishing an all-Ireland biobank and make a report to the Health Research Board and Research and Development Office. The first meeting of the Group took place on 14 April 2003.

1.2 Membership of expert group

Professor Bernadette Herity	Chair and Convenor
Dr Ruth Barrington	CEO, Health Research Board, Dublin
Professor Hugh Brady	Department of Therapeutics and Medicine, University College Dublin
Dr Michael Boland	Irish College of General Practitioners
Professor Leslie Daly	Department of Public Health Medicine and Epidemiology, University College Dublin
Professor Frank Kee	Department of Epidemiology and Public Health, The Queen's University, Belfast
Dr Teresa Maguire	Secretary to Group, Health Research Board, Dublin
Professor Peter Maxwell	Faculty of Medicine, The Queen's University, Belfast
Professor Paul McKeigue	Conway Institute, University College Dublin
Dr Pierre Meulien	Dublin Molecular Medicine Centre, Dublin
Professor Tony McGleenan	School of Law, University of Ulster, Jordanstown
Professor Ivan Perry	Department of Epidemiology and Public Health, University College Cork
Professor Philip Reilly	Department of General Practice, The Queen's University, Belfast
Professor Bob Stout	Director, Research and Development Office, Belfast

1.3 Aim and Terms of Reference of the Group

Aim:
To advise on the feasibility of establishing an all-Ireland biobank and to contribute to a report to the HRB and the RDO on such feasibility.

Terms of Reference:
To report on

1. The objective of and scientific rationale for an all-Ireland biobank for research into genetically related diseases, including multi factorial diseases
2. The sample size and its distribution by age, sex, and geography
3. The number and nature of the biological samples to be collected and stored in the biobank
4. The baseline data collection
5. Arrangements for handling and analysing the data and for statistical support
6. A possible location or locations for a biobank
7. The key ethical issues involved in establishing a biobank and how these might be addressed
8. The cost of establishing the biobank over a five-year period and how the cost might be met
9. The feasibility of bringing existing banks of biological material within the biobank
10. The most productive relationship between the all-Ireland biobank and Biobank UK
11. The governance arrangements for the biobank
12. The conditions under which researchers would be given access to the biological samples in the biobank

1.4 Meetings and action plan of group

The group met on three occasions alternately in Dublin and Belfast. There was consensus at an early stage that the establishment of a biobank on the island of Ireland was feasible and desirable. It was agreed that the group should develop a plan for an all-Ireland biobank which would address the issues contained in the terms of reference. It was further agreed that the preparation of a detailed protocol for the establishment of a biobank would require the commitment of significant resources not currently available to the group. It was therefore decided to present an outline protocol for the establishment of an all-Ireland biobank which could be further developed if funding for the project were to be forthcoming.

Draft papers on various aspects of the biobank were produced for the group by Professors Leslie Daly, Frank Kee, Peter Maxwell, Paul McKeigue and Ivan Perry; Professor McKeigue wrote the final consolidated draft paper. Professor Tony McGleenan contributed to the section on ethical and legal considerations. These drafts were considered and discussed by the group, various members of which suggested amendments and the final report was edited by Professor Bernadette Herity and Dr Teresa Maguire. Ms Brónagh O'Connor gave valuable administrative assistance to the group.

2. Scientific rationale and study designs for biobanks

2.1 Scientific rationale

Advances in biology and medicine are likely to depend upon exploiting the opportunities arising from developments in genetic knowledge and technology[1,2]. One of the key objectives is to discover how genetic variation influences disease susceptibility. This is because discovering an effect of variation in a gene upon disease susceptibility establishes an unequivocal causal relationship between the gene product (a protein) and the disease or trait under study. This advances understanding of the molecular basis of disease susceptibility, and opens up possibilities for developing new preventive or therapeutic

measures, for instance by developing drugs that target the gene product or other proteins in the same pathway.

The assembly of the finished human genome sequence is nearing completion. The sequence is now being annotated to show all known and putative genes and most common polymorphisms (sites of sequence variation), and the haplotypes formed by alleles at polymorphisms that are close together on the genome. This information, although valuable in itself, does not tell us anything about the effects of genetic variation on disease risk or traits of importance to human health. To study such effects requires access to collections of human DNA and clinical data, known as biobanks. Once such collections have been created, they provide an almost inexhaustible resource for further research as DNA extracted from a single blood sample is sufficient (with careful management) for many thousands of genes to be studied in detail. The effects of common genetic variants on multifactorial diseases are likely to be of modest size (risk ratios associated with genotype < 2). To study these modest effects requires very large collections.

2.2 Study design and collection strategies

Two epidemiological study designs may be used to investigate the association between genotype and risk of disease, cohort and case-control designs. For detailed discussions of epidemiological methods for studying genetic effects on disease risk see references 3 and 4.

Several countries are now establishing collections of DNA and clinical data on a national basis. One of the largest is Biobank UK, a study of 500,000 people aged 45-69 in Great Britain, for which support has been awarded by the Wellcome Trust and UK MRC. This is based on a **cohort** design in which participants are studied at baseline and then followed for 10 years. This design has limitations:

- with a two-year pilot phase, five years of baseline collection and ten years of follow-up, the resource will not achieve its targets for case accrual until ~ 2020
- because samples cannot be irreversibly anonymised compliance with ethical guidelines will necessitate strict limits on access to and use of the resource

The Republic of Ireland (RoI) and Northern Ireland (NI) have the opportunity to adopt an alternative **case-control** approach which would be more efficient, would allow collection of more detailed phenotypic data and would not suffer from the same ethical restrictions.

It is our view that **case-control** designs have the following advantages over **cohort** designs for studying genetic effects on disease risk:

- for a given outlay of resources far greater statistical power can be obtained with case-control designs than with cohort designs
- case-control designs yield results more quickly than cohort designs
- in a case-control design DNA samples and clinical data can be irreversibly anonymised. This minimises ethical difficulties: consent can be obtained for unrestricted use of the anonymised samples and the resource can be made freely accessible

With modern statistical methods to control for population stratification in genetic case-control studies it is unnecessary to match cases and controls for demographic factors. A single large control group can thus be used for multiple case-control studies and this collection of controls can serve two other uses:

- as a resource for studying associations with quantitative traits such as obesity, blood pressure, plasma lipids and glucose tolerance
- if based on a representative population sample it can provide data for monitoring the health of the population and for studying social and regional variation in health status

2.3 Size of case sample and catchment population required for case-control studies

Calculations of the required sample size for genetic case-control studies depend upon somewhat arbitrary assumptions but most statisticians now agree that a target sample size of 3,000 to 5,000 cases, plus up to three controls per case, is a minimum for serious study of modest effects (genotype risk ratios less than 2). To recruit 3,000 cases we need a catchment population large enough to generate at least 10,000 cases (based on 50% ascertainment and a 60% response rate).

The table below shows the estimated size of the catchment population required to generate 10,000 cases of a few diseases that have been suggested for collection. This is based on combining prospective case ascertainment over 1-2 years with retrospective ascertainment of surviving cases.

Table 1: Catchment populations that generate 10,000 cases of disease

Disease	Prevalence /1000	Incidence /100 000 /year	Catchment population that generates 10,000 cases
Colorectal cancer		60	4 million (4 years' cases)
Oesophageal cancer		9	56 million (2 years' cases)
Rheumatoid arthritis	10		1 million
Type 1 diabetes	2		20 million
Type 2 diabetes	20		2 million
Neural tube defects	1.5 (at birth)		20 million (5 years' cases, crude annual fertility rate 70/1000)
Hepatitis C		1.6	13 million (5 years' cases)

Resources for establishing case collections have already been allocated in RoI under the Programme for Research in Third Level Institutions (PRTLI). Establishing a shared control group will allow these resources to be used more efficiently. For rarer diseases, collaboration with other European countries will be necessary to establish a resource of adequate size.

Genetic structure of the Irish population

A recent study [5,] shows that there is hidden stratification in the population of Ireland, with a gradient of allele frequencies from east to west. This is a continuum of a gradient across Europe from south-east to north-west and probably reflects varying degrees of admixture between the earliest inhabitants and later settlers. It has been suggested that this genetic stratification underlies geographic variation in risk of coeliac disease within Ireland. [6] With a collection that covers all regions of Ireland, it will be possible to measure population stratification and to control for it where necessary when studying genetic associations.

3. Outline proposal for all-Ireland Biobank

3.1 Objectives

The objective of this project is to establish a freely available resource for studying genetic effects on health and disease based on the population of the island of Ireland. This resource will consist of a bank of DNA samples linked to clinical and demographic data, irreversibly anonymised so that a link to individual subjects is not possible. A cross-sectional study of 10,000 individuals sampled from the general population will provide:

(i) a resource for studying genetic effects on quantitative traits such as obesity
In a short physical examination combined with analysis of plasma many quantitative traits of importance to health can be measured.

(ii) a shared control group for genetic case-control studies
Establishment of a shared control group will allow resources allocated for other DNA collections in Ireland to be used more efficiently. To establish a case-control bank for a disease of interest, researchers need only collect cases.

(iii) a health monitoring survey
As a by-product, collection of clinical and demographic data from a representative population sample will provide a baseline for monitoring the health of the population, complementing previous surveys based on questionnaire data only.

The resource will ensure that researchers in the Republic of Ireland (RoI) and Northern Ireland (NI) have rapid and unimpeded access to a resource from which it will be possible to discover new associations between genetic variation and health outcomes and establish whether associations reported elsewhere are valid in the Irish population. This could facilitate discovery of possible drug targets through detecting genetic associations with disease susceptibility and would allow tests of disease susceptibility to be validated in the Irish population.

3.2 Sample size and age distribution

From experience with the Health Survey for England (http://www.dh.gov.uk/PublicationsAndStatistics/PublishedSurvey/HealthSurveyForEngland/fs/en), we estimate that a response rate of 70% can be achieved at the interview stage and that about 70% of these respondents will attend for physical examination including blood sampling, giving a final response rate of about 50%. The interview data can be used to adjust for any selection bias that arises from non-participation in physical examination among those interviewed. Thus for 10,000 people to be examined, it will be necessary to invite 20,000 and to interview 14,000.

3.3 Sampling frame

We propose choosing 20 primary sampling units each covering an equal catchment population, stratified to sample urban and rural districts within each province of Ireland and sampling households as the secondary sample points. Similar protocols have been used successfully in national health monitoring surveys in England, Scotland and Northern Ireland. The alternative would be a sampling frame based on lists of general practitioners: with this, incorrect address records could cause difficulties and selection bias would be greater. Although selection bias is not a serious problem for genetic studies, it would lessen the value of the collection for health monitoring.

For any variable of interest the "design effect" – the increase in sample size required to compensate for the reduced information obtained from using a cluster sample rather than a simple random sample of the same size – can be calculated from the intra-class correlation coefficient. Experience in other health monitoring surveys is that for most of the variables of interest, this design effect is small and that it is far outweighed by the lower unit costs of using a cluster sample rather than a simple random sample. For instance, if two adults are interviewed per household and the intra-household correlation coefficient for obesity is 0.15,

the design effect is only 15% [(2 -1) × 0.15]. The ability to measure geographic and household clustering of variables of interest is also of relevance for public health.

With a residential sampling frame, the most efficient approach is for an interviewer to visit each sampled household, ascertain the number of adults living there, invite them to participate, obtain consent and administer a questionnaire. The electoral register can be used to draw a sample of households (but not individuals) and a letter of invitation addressed to the occupiers can be delivered in advance, accompanied by publicity in local and national media. Sampled individuals will be invited to complete the interview even if they do not wish to attend subsequently for physical examination and blood sampling.

3.4 Data collection by questionnaire

The questionnaire will include demographic items, medical history and environmental/behavioural variables such as smoking and alcohol intake. Standard instruments can measure anxiety and depression. A simple food frequency questionnaire can be included, but dietary survey measurements have severe limitations. A food frequency questionnaire can be relied on only to distinguish gross variation in dietary pattern, e.g. it will discriminate between vegetarians and non-vegetarians.

Additional items from the questionnaires used in the SLÁN [7] and NI Health and Wellbeing survey [8] such as items covering health service utilisation, could be included for health monitoring purposes. If the respondent agrees to attend for physical examination, an appointment will be arranged and (with the participant's consent) the general practitioner will be informed.

3.5 Clinical examination of participants

At each sampling point office space for examinations would be rented in a nearby health centre or community centre. Alternatively, trailers can be equipped as mobile examination centres. Examinations will be undertaken by a team of two nurses or possibly four in urban areas where a single examination station can serve two primary sampling units.

A detailed protocol for the examination will have to be decided at a later stage. We note that the US NHANES health monitoring survey has already undertaken detailed studies of what to measure in a survey of this type, and the measurement protocols are freely available (www.cdc.gov/nchs/nhanes.htm). The following is a possible list of measurements and samples to be obtained at examination or from subsequent recording.

- Visual acuity and refractive error
- Hearing in those aged over 60 years
- Tooth count
- Anthropometry
- Blood pressure measurement: for increased reliability, measurements could be obtained both at interview and when participants attend for examination
- Blood sample (either fasting, or at 2 h after an oral glucose load taken at home)
- Respiratory function measurements
- Body impedance (an indirect measure of body water and fat content)
- Vascular function studies in those aged over 40 years
- Autonomic function in those aged over 40 years
- Timed overnight urine sample
- Resting ECG in those aged over 40 years
- Energy expenditure by doubly-labelled water (on a 5% sub-sample of participants)

- Physical activity monitors (calibrated for energy expenditure by the doubly-labelled water method)

We estimate that a resting ECG will yield only about 100 cases of probable coronary disease (Minnesota ECG codes 1-1 or 1-2) and that at least half of these individuals will have been previously diagnosed.

Monitoring energy expenditure in the population would make it possible to distinguish the contributions of changes in energy expenditure and energy intake to the increase in obesity in the population, and to define the possibilities for public health measures to reverse this trend. The reference standard for measuring energy expenditure is the doubly-labelled water procedure. [9]

Although this procedure is too expensive to undertake on all participants, measurements on a sub-sample will be valuable for health monitoring purposes, and can be used to validate measurements of physical activity that are undertaken on all participants. Where genetic associations with obesity are detected in the entire sample, the mechanism of these associations can be explored using the sub-sample of participants on whom doubly-labelled water measurements have been obtained.

Plasma from at least one aliquot of blood will be separated at the site of examination and transported within 24 h to a central lab for biochemical measurements.

3.6 Notification of results

Participants will be informed of their clinical examination and laboratory test results and (with the participant's consent) a copy of these results will be sent to the general practitioner. After notification is complete and any follow-up queries have been dealt with, the records will be irreversibly anonymised so that it would be impossible for anyone to link them to identifiable participants. To ensure that all copies of the key linking ID numbers to patient identities can be destroyed, this key file will be held on a separate volume with separate backup facilities. The protocol for this irreversible anonymisation will be reviewed by an independent data security expert.

3.7 Biochemical analyses

Plasma lipid and glucose levels will be measured on the plasma samples; measurements of each analyte will be outsourced to a single laboratory. At least six aliquots of plasma from each participant will be stored for subsequent analysis. Urine analyses should include albumin (as an early indicator of hypertensive damage) and creatinine (to standardise the measurements of other analytes).

3.8 Banking of samples

Cells and whole blood for DNA extraction will be shipped to a central laboratory. A backup specimen of whole blood on FTA paper will be stored separately to ensure that mislabelled samples can be identified subsequently. Each 10 ml EDTA blood sample should yield at least 300 μg of DNA. Extracted DNA will be stored in a format that allows individual samples to be easily retrieved for preparation of microplates. Storage will be either in liquid phase (requiring automated freezers) or in solid phase at room temperature if specimen handling can be automated. It is not necessary to prepare immortalised cell lines, which would cost at least €100 per participant. Current genotyping protocols use only 1-2 ng of DNA per genotype, and even less if whole genome amplification is used.

3.9 Location of the resource

Storage of the DNA samples should be outsourced to any lab that has appropriate facilities. A set of aliquots will be transferred to whichever lab is used for routine SNP genotyping. The main database of clinical and genetic data should be maintained by the biobank management centre.

4. Project management

4.1 Management structure

A steering committee of researchers should be responsible for strategic direction of the project and for preparing a detailed protocol. The plan outlined here is based on keeping costs to a minimum and on completing all collections within 18 months. There is no cost saving in extending fieldwork over 2 – 3 years. By recruiting more staff on shorter contracts fieldwork can be concentrated into an eighteen month period. Another advantage of setting a shorter time frame for fieldwork is that it would be compatible with a three-year research training fellowship for a junior epidemiologist to work on details of the protocol, coordinate the project on a day-to-day basis and write up the results.

4.2 Sourcing of work

Once the protocol has been agreed, we recommend that the collection be outsourced to a survey agency, with a full-time project manager who reports to a small executive committee (rather than the full steering committee). A model for this is the Health Survey for England, for which fieldwork (~17,000 adults interviewed and ~12,000 examined over a twelve-month period) is undertaken by a semi-commercial survey agency (National Centre for Social Research) and data analysis is undertaken by a university-based group. Even if no survey agencies in Ireland currently have capacity for a project on this scale, consortia can be formed if bids are invited. To minimise competing interests, researchers on the steering committee should step down if they become involved as contractors in the fieldwork or laboratory analyses.

4.3 Equipment and infrastructure requirements

The survey itself will not require much new infrastructure: a survey agency can recruit and train interviewers and nurses and can rent temporary space in local community centres and hospitals. In rural areas trailers could be equipped as mobile examination centres. Equipment for each field station will cost about €2,000, depending on what is included in the examination protocol. DNA extraction can be outsourced to a commercial lab; currently the rate for this is about €15 per specimen.

Freezer storage will require –80^{o}C freezers for plasma storage and –20^{o}C freezers for storing DNA in solution. Storage space can be rented on any site that has secure storage, backup power supplies and facilities for connecting an alarm to the telephone network. The problem is to retrieve samples from freezers. Freezers with automated retrieval of individual specimens are currently expensive. DNA can be stored on solid media (FTA paper) at room temperature, but currently no automated equipment is available for transferring aliquots from paper to microplate format. The availability of commercial facilities for storage and automated retrieval of specimens for robotic liquid handling would have to be investigated. This is the only stage of the project that is likely to require investment in infrastructure.

Database management for the project will require full-time staff, but will not impose any special infrastructure requirements. At this stage there would be no need to invest in genotyping facilities – commercial facilities are currently competitive with any academic centres and the technology is rapidly changing so that any new investment is likely to become obsolete.

4.3 Timetable

A timetable for the project is shown below:

Recruitment of staff, training, finalising protocol	months 1-9
Fieldwork	months 9-18
Sample banking, genotyping and analysis	Months 16-24

5. Ethical and legal considerations

5.1 Background

The group acknowledges that research on databases of biological materials raises ethical and legal issues and has considered relevant legal and ethical guidelines on the collection and use of DNA and human tissue samples for research purposes. At present there are no specific laws in this regard in either the UK or ROI. However a draft for a legal instrument on use of archived human material in biomedical research has been drawn up by the Council of Europe's Steering Committee on Bioethics (www.coe.int/T/E/Legal%5Faffairs/Legal%5Fco%2Doperation/Bioethics/Activities/Biomedical_research/CDBI-INF(2002)6E.asp#TopOfPage). This draft sets out general principles for research using human tissue samples and regulations for ethical review, information and consent. The draft specifically exempts from these regulations collections which have been irreversibly anonymised with the proviso that the anonymisation process should be reviewed by the "competent body" (an ethics committee). This draft is at present out for consultation by member states. In the UK a Human Tissue Bill has been drafted (http://www.parliament.the-stationery-office.co.uk/pa/cm200304/cmbills/009/2004009.pdf) that will make it an offence to posses human tissue with the intent to analyse DNA without consent. Detailed ethical guidelines for research using human DNA samples have been issued by the Medical Research Council (www.mrc.ac.uk/pdf-tissue_guide_fin.pdf). A working group on Human Biological Collections, established by the Irish Council for Bioethics in June 2002, has not yet produced guidelines. The Irish Council on Bioethics is represented on this Expert Group. In the protocol proposed for an all-Ireland biobank, results from genetic studies will only be entered into the database after the samples and clinical data have been irreversibly anonymised. Therefore, many of the more complex legal and ethical issues do not arise. Specific ethical issues in this project are:

- those arising in studies that screen apparently well people
- those related to banking of samples for genetic studies

5.2 Health Screening Survey

The ethical issues arising in health screening surveys are mainly those of informed consent to testing and the participant's right to decide who should be informed of the test results. Potential participants will be provided with information about the proposed research in a comprehensible form. Participants who agree to take part in the research will be invited to indicate on the consent form whether they wish to be notified of their test results, and whether they wish their general practitioner to be informed of their physical examination and test results. Where consent is given to notification, letters will be sent within 2-3 weeks to the participant and to the general practitioner, with a recommendation to consult where appropriate. This stage will be completed before the records are irreversibly anonymised

5.3 Collection and use of DNA samples

The protocol complies fully with current and proposed legal regulations and ethical guidelines described above (5.1). No blood samples will be taken before written consent is obtained from participants following the provision of full and comprehensible information about the proposed research. The samples will be irreversibly anonymised and the anonymisation process will be independently reviewed in both NI and RoI.

The consent form will be based on the model consent form that is included in the MRC guidelines and will explain the potential use and storage of the clinical data and samples and the irreversible anonymisation process. Where consent is withheld the blood sample will not be taken. Participants will also have the right to have their samples and clinical data withdrawn from the bank up to the point where the record has been irreversibly anonymised.

6. Estimated Costs

6.1 Costs of survey

The Health Survey for England achieves a unit cost of about €120 per participant for a basic protocol that includes an interview, anthropometry, blood pressure measurement, blood sampling and biochemical measurements. The all-Ireland biobank as proposed will have a more complex protocol, for which a lower limit for the estimated cost is about €300 per case, giving a total cost of about €3 million over two years. These costs are outlined in Table 2.

6.2 Project management and data analysis

Contract for a small university-based group to supervise the project and to maintain the database would cost about €200k per year (three full-time salaries plus overheads).

6.3 Fieldwork

The main cost will be the salaries of nurses and interviewers: employing 40 half-time nurses and 40 half-time interviewers for 12 months will cost about €1.6 million. In each primary sampling unit, a team of two interviewers and two nurses would be expected to interview about 700 participants and to examine about 500 of them in one year: equivalent to 70 interviews and 50 examinations/month.

6.4 Laboratory analyses

Laboratory analyses will cost about €30 per case.

6.5 Extraction and banking of DNA

Costs for extraction and banking of DNA would be approximately €150k.

6.6 Cost of proposed initial genotyping studies

Although the primary objective of the biobank project is to establish a resource for further research we propose that the project should include an extra funding provision for initial genotyping studies. An allocation of €100k for genotyping would allow up to 100 loci (at €0.10/genotype) to be typed in the entire collection, or more loci to typed in comparisons of subgroups of interest (for instance e.g. the top decile versus the bottom decile for obesity). This initial genotyping study will be valuable for several reasons:-

- it will allow all components of the resource to be tested as they are assembled . Any problems with DNA yield, sample storage and retrieval, quality of clinical data and database management can be identified while the project is still under way, rather than at a later date when staff have moved on and errors cannot be corrected.
- it will make the post of project co-ordinator more attractive to a junior clinical researcher wishing to undertake an MD or PhD. We consider that appointment of a highly motivated project co-ordinator is crucial to the success of the project.
- an initial study typing a panel of markers informative for ancestry can be used to measure the genetic structure of the Irish population: for instance to define the gradient of admixture that has been inferred previously from a study using Y chromosome haplotypes[5]. This will contribute to understanding the basis of any regional variation in risk of genetic diseases within Ireland and will allow case-control studies to control for hidden population stratification by selecting matched controls. This could be undertaken in collaboration with a similar study now planned in Britain.
- It would also allow some hypotheses to be tested immediately e.g. in relation to lipid levels, obesity, hypertension etc in the Irish population. Appendix I briefly outlines a plan for a specific study of obesity, which could be undertaken within the resources.

Table 2: Estimated costs of establishing biobank

Cost heading	**€k**
Project management group	
3 salaries for 2 years: project manager @ €70k, 2 assistants @ €45k	320
Expenses and equipment	80
Total cost of project management group	**400**
Contract for survey	
40 half-time nurses for 12 months @ €42k/year full time	840
40 half-time interviewers for 12 months @ €38k/year full time	760
Expenses, field equipment and management costs	650
Total cost of fieldwork contract	**2250**
Contract for laboratory analyses @ €30 a sample	**300**
Contract for extraction and banking of DNA	**150**
Total cost	**3.1 million**

7. Relationship with other projects

7.1 Existing collections of DNA with clinical data in Ireland

The value of this project as a shared control collection is dependent on the availability of large, freely available, irreversibly anonymised case collections. With the exception of acute coronary syndromes, most existing case collections in NI and RoI consist of only a few hundred cases, rather than the thousands of cases required for serious genetic association studies. The list below is based on a Medline search of recent publications and on personal enquiries (numbers of cases are not necessarily up to date).

Disease	***Number***	***Type of collection***	***Institution***
Rheumatoid arthritis	251	Unrelated cases	QUB
Type 1 diabetes	427	Unrelated cases and parent-offspring trios	QUB
Osteoporosis	311	Unrelated cases	Belfast City Hospital
Parkinson's disease	90	Unrelated cases	Belfast City Hospital
Multiple sclerosis	304	Unrelated cases	Belfast City/Royal Hospital
Alzheimer disease	242	Unrelated cases	QUB
Schizophrenia	219	Unrelated cases	TCD

Attention-deficit hyperactivity disorder	119	Unrelated cases	TCD
Neural tube defects	276	Parent-offspring trios	TCD
Crohn's disease	131	Multiplex families	UCC
Acute coronary syndromes	1400	Unrelated cases	RCSI

Support has already been awarded under the PRTLI scheme to the Dublin Molecular Medicine Centre (DMMC) in Dublin for the establishment of large case collections under the Programme for Human Genomics. The establishment of collections of up to 3,000 cases for each of four diseases is proposed. A provisional list of these diseases is: prostate cancer, coronary heart disease, schizophrenia and inflammatory bowel disease. This selection of diseases to study is based upon relevance to existing research programmes, availability of registries or other means of rapid case ascertainment, feasibility of collection in Ireland and the possibilities for obtaining diseased and normal tissue samples for functional genomics/proteomics studies.

7.2 International collaborations

The all-Ireland biobank will be a source of controls for other multicentre DNA case-control collections in which Ireland is participating:

ECTIM (QUB): European case-control collection for myocardial infarction
EUDRAGENE (UCD): European case-control collection for adverse drug reactions
Irish Schizophrenia Families' Study (HRB collaboration with Virginia Commonwealth University)

7.3 Biobank UK

Since samples in Biobank UK will not be irreversibly anonymised, they cannot be made freely available to researchers. Any proposals for genotyping the Biobank UK collection will therefore be subject to ethical review and possibly to a requirement for further informed consent. Although this will impose delays, it will be possible for researchers to use Biobank UK to replicate studies undertaken using the all-Ireland biobank, where the same diseases or phenotypes have been studied.

7.4 Other health monitoring surveys in Ireland: SLÁN, NI Health and Wellbeing Survey

Where relevant, questionnaire items from these earlier surveys (which were almost entirely questionnaire-based) can be included in this project.

8. Governance of all-Ireland biobank

8.1 Custodianship of biobank

Formal responsibility for custodianship and control of use of biobank samples should rest with a Board of Management, which is representative of all interests and has an independent chair and some independent membership. It would be expected that the Board of Management would have responsibility for the central administration of the biobank and for arrangements for storage of data and biological samples. The Board would also have to establish a scientific committee to develop policy on the use of the biobank and to oversee applications from researchers for access to the resource. Arrangements for such a management structure could be best developed and implemented by the Health Research Board and the Research and Development Office in consultation with interested bodies and individuals.

8.2 Access to resource

Access to the resource will be freely available to researchers in Ireland and to researchers outside Ireland where there is a reciprocal agreement on sharing samples and clinical data. Prospective users of the resource will be expected to submit an outline of their research plan to a scientific committee, and to undertake that (after they have written up results of interest for publication) all their genotype data will be submitted for merging with the main database.

8.3 Costs of DNA analyses

Costs of DNA analyses should be borne by users of the resource. The main use of the resource will be for typing known variant sites (usually single-nucleotide polymorphisms) in candidate genes. Occasionally users may request control DNA for resequencing one or more candidate genes, although more usually they will resequence a small panel of cases, then type all cases and controls for the variants detected in the cases. We expect that as the costs of genotyping fall, these genotyping studies will extend to the entire genome. To minimise the losses in preparing separate aliquots, banking of samples and genotyping should if possible be undertaken in the same lab. Users can then make their own contracts with the genotyping lab. As resequencing and genotyping are now largely routine tasks, they should be outsourced to a lab that can offer competitive prices (currently ~ €0.10/genotype, new assay setup ~ €100) and acceptable quality (missing genotypes <5%, error rate < 1%). If users have special requirements that cannot be met in the main genotyping lab, they can be supplied with aliquots to analyse in their own lab.

9. Cost of maintaining the biobank

To maintain the biobank, with storage, retrieval and aliquoting of samples and data management, we estimate that, based on current costs, funding of €100-150k per annum would be needed.

10. Benefits of an all-Ireland biobank

Studying genetic effects on human health will advance understanding of the molecular basis of disease susceptibility, and lead to development of new therapeutic measures. The biobank will provide a freely accessible resource for such studies, based on the population of the island of Ireland. Quantitative traits such as obesity and common diseases such as asthma can be studied using only the samples in the biobank. For case-control studies, the biobank will provide a shared control group representative of the total population. This will enable researchers to use funding more effectively by concentrating their resources on collecting cases for diseases of interest. Case collections on the island of Ireland have already been established, and more are proposed (see 7.1). Appendix I gives examples of the types of studies which could be undertaken using the proposed biobank either alone or in combination with large case collections for which funding has already been awarded.

11. Conclusions

The Group concludes that the establishment of a biobank from the all-Ireland population is feasible and has real potential to advance research into the genetic basis of health and disease on the island of Ireland. Preparation of a more detailed research protocol with precise costings, submissions for ethical approval etc. would require further resources.

References

1. Bell J. The new genetics in clinical practice. Br Med J 1998: **316**: 618-620.

2. Fears R., Roberts D., Poste G. Rational or rationed medicine? The promise of genetics for improved clinical practice. Br Med J 2000: **320**: 933-935.

3. Clayton D., McKeigue P.M. Epidemiological methods for studying genes and environmental factors in complex diseases. Lancet 2001: **358**:1356-60.

4. Hoggart C.J., Parra E.J., Shriver M.D., Bonilla C., Kittles R.A., Clayton D.G., McKeigue P.M. Control of confounding of genetic associations in stratified populations. Am J Hum Genet 2003: **72**:1492-504.

5. Hill E.W., Jobling M.A., Bradley D.G. Y-chromosome variation and Irish origins. Nature 2000: **404**: 351-2.

6. Cronin C.C., Shanahan F. Why is celiac disease so common in Ireland? Perspect Biol Med 2001: **44**: 342-52.

7. Friel S., Nic Gabhainn S. and Kelleher C. The National Health and Lifestyle Surveys SLÁN survey and HBSC). Galway: Centre for Health Promotion Studies, National University of Ireland 1999.

8. Northern Ireland Statistics and Research Agency. Northern Ireland Health and Social Wellbeing Survey. Belfast: Northern Ireland Statistics and Research Agency, 2001

9. Coward WA. Stable isotopic methods for measuring energy expenditure. The doubly-labelled-water (2H2(18)O) method: principles and practice. Proc Nutr Soc. 1988: **47**:209-18.

Appendix I

Examples of the types of studies which could be undertaken using the all-Ireland biobank

Preliminary study of the genetic structure of the Irish population

A study of Y chromosome haplotype frequencies in Ireland shows that there is a gradient of haplotype frequencies from east to west. This appears to be one extreme of a gradient of allele frequencies across Europe from south-east to north-west, resulting from varying degrees of admixture between the earliest Mesolithic settlers and the spread of Neolithic farmers from the Near East during the last 10,000 years. It is important to measure this population stratification because it may confound genetic associations, and because it may be relevant to understanding geographical variation in risks of diseases such as coeliac disease.

To study population structure and to measure the admixture of each individual so that confounding can be controlled, we require a panel of about 100 markers that have been selected to be informative for ancestry. To identify such markers, it will be necessary to screen at least 100,000 marker loci to identify those that show geographic variation in allele frequencies within Europe (from southeast to northwest), within the UK and Ireland. This screening can be undertaken efficiently using pooled DNA samples with currently available genotyping chips that score 10,000 loci on each chip. These markers can then be typed in the Irish biobank samples, and analysed using programs such as ADMIXMAP or STRUCTURE to define the genetic structure of the population and to measure the admixture proportions of each individual. Such a study could be undertaken in collaboration with a group of UK-based researchers led by Professor Walter Bodmer and Professor Stephen Donnelly at Oxford, who are planning a systematic study of the genetic structure of the British population.

Studies of quantitative traits: a study of obesity as an example

The increasing prevalence of obesity in developed countries underlies many other health problems, generating increased rates of diabetes, hypertension, and disability. Several study designs may be used to achieve a better understanding of the molecular basis of obesity using the proposed biobank. The most efficient design is to contrast two groups at the extremes of the distribution in the population: for instance one could compare allele frequencies or haplotype frequencies for 1,000 individuals in the top decile and 1,000 individuals in the bottom decile of body fat percent, stratifying by age and sex. Markers informative for ancestry can be typed so that confounding for hidden population stratification can be controlled, either in the design by matching the two groups or in the analysis by adjustment.

Two alternative approaches are available for discovering associations of a trait with variation in a gene, direct and indirect. In the direct approach, polymorphisms that affect the function of the gene are tested directly for association with the trait. This approach relies on being able to determine in advance which polymorphisms are likely to affect gene expression or the function of the gene product. The indirect approach relies on detecting associations with marker polymorphisms that are in allelic association with functional polymorphisms in the same gene. Using the data on haplotype frequencies that are accumulating through the International HapMap Project (www.hapmap.org), it will be possible to select a panel of about 200,000 "haplotype-tagging" SNPs that capture most of the genetic variation in all 30,000 human genes. This will make it possible to study associations with any candidate gene systematically, using the indirect approach. Assays are now available that allow allele frequencies to be measured in pooled DNA samples from large numbers of individuals. This makes it feasible to exploit massively parallel assays, such as chip-based genotyping methods, that are expensive when used to score individuals but cheap when used to estimate allele frequencies in pooled DNA samples. The optimal design is to use about 20

pools of DNA from 50 individuals in each of the two groups of 1,000 individuals being compared. [1]

With chip-based assays to measure allele frequencies for 200,000 haplotype-tagging SNPs in pooled DNA samples, it will be feasible to test every gene in the human genome for association with a quantitative trait such as obesity, if this trait has been measured in the individuals who have contributed DNA to the biobank. Where association with haplotype-tagging SNPs in a gene is detected in an initial screen using pooled DNA samples, this can be investigated further with typing of individual samples and additional polymorphisms in the same gene.

Where genetic effects on obesity are identified, it will be possible to use the biobank to explore the mechanism of these effects. For instance, we can use measurements of energy expenditure and physical activity to test whether a genetic association with obesity is mediated through these factors.

Case-control studies

The DMMC Programme for Human Genomics plans to collect 3000 cases of each of the following diseases (provisional list): prostate cancer, acute coronary syndromes, inflammatory bowel disease and schizophrenia. For each of these diseases it will be possible to undertake case-control studies with three controls per case, using a research design similar to that outlined above for studies of obesity.

1. Barratt B.J., Payne F., Rance H.E., Nutland S., Todd J.A., Clayton D.G. Identification of the sources of error in allele frequency estimations from pooled DNA indicates an optimal experimental design. Ann Hum Genet 2002 : **66**:393-405.