

B.1 Scientific and technological objectives of the project and state of the art

(max 3 pages)

Overall objective

In this proposal, we will study the genetic underpinnings of two inherited cancers, of prostate in men and breast in women. We will probe the association between common polymorphisms in a large number of candidate cancer susceptibility genes and the risk of breast and prostate cancer. Furthermore, we will develop computational and statistical methods that center on the analysis of genetic data from studies of complex diseases. To maximize our likelihood of success, we have chosen to study two European populations with different genetic backgrounds, the homogeneous, well defined population of Iceland and the more mixed population of the Netherlands. The collection of accurate clinical information allows the definition of different clinical phenotypes which is of crucial importance in genetic studies of complex diseases like breast and prostate cancer. The long-term objective of the study is to gain an increased understanding of the genetic underpinnings of the various forms of breast and prostate cancer which may in turn lead to more effective risk assessment, increase the efficiency of screening programs and lead to improved diagnosis and treatment.

State of the art

Cancer is a complex disease where genetic and environmental factors both play a role. The genetic contribution to the development of cancer has been intensely debated in recent years. While it is clear that high-penetrance cancer genes such as *BRCA1* and *BRCA2* contribute to the development of breast and prostate cancer, it is also clear that a large component of genetic cancer risk remains unaccounted for. A recent large twin study downplayed the importance of genetic susceptibility in cancer development (Lichtenstein *et al.*, 2000). However, it has been pointed out that twin studies can yield only a lower limit of the proportion attributable to genetic factors (Peto *et al.*, 2000). Furthermore, other studies have suggested that environmental factors contribute little to familial aggregation of most cancers, supporting the idea that genetic factors contribute much more than would be suggested by twin studies (Risch, 2001). Evidence from studies in twins, tumor incidence in the contralateral breast of an affected individual and familial inheritance patterns support the idea of genetic predisposition factors (Cui *et al.*, 2001; Peto, 2001; Antoniou *et al.*, 2002).

The identification of the *BRCA1* and *BRCA2* genes raised hopes that systematic studies of large cancer-prone families would successfully identify the major genetic determinants of cancer. However, these hopes have not been fulfilled. The field of breast cancer genetics, where most of the effort has been focused, has not seen major novel cancer gene discoveries since *BRCA1* and *BRCA2* were found. However, population-based epidemiology studies have shown that only 15-20% of familial breast cancers occur in families carrying mutations in the *BRCA1* or *BRCA2* genes (reviewed in (Balmain *et al.*, 2003). Thus, the mapping of additional high-penetrance cancer genes has been unsuccessful in the majority of breast cancer families. The genetic model that best fits the observed breast cancer rate in relatives of cancer patients and in multiple-case families is a polygenic model in which the susceptibility to cancer is conferred by a large number of alleles (Cui *et al.*, 2001; Peto, 2001; Pharoah *et al.*, 2002). In this model the risk associated with each individual allele is small, but as the effects may be additive or even multiplicative, an individual with several susceptibility alleles is at high risk. Ponder has compared several hypothetical models for breast cancer susceptibility genes and concluded that if the low penetrance polygenic model is correct, there are dozens to hundreds of genes (depending on relative risk) that need to be identified in order to account for the observed familial risk in breast cancer alone (Ponder, 2001).

For any estimation of total attributable risk of low penetrance susceptibility genes it is vital to systematically evaluate them in a single study, ideally in a population where the potential bias of case-control studies can be minimized. The association approach (i.e. the case control study) has been proposed as the method of choice for identifying low penetrance genes in the polygenic model of inheritance (Cardon *et al.*, 2001). Currently, however, whole-genome association studies are too expensive and time-consuming to be seriously considered for disease-gene searches. Several approaches are currently under consideration in order to simplify this task. One approach is to reduce the number of markers needed for studying the whole genome. A major aim of the International HapMap project (The

International Hapmap Consortium, 2003) is to characterize the haplotype structure of the human genome in order to facilitate large scale association studies. Another approach is to use populations where, due to historical reasons and extensive linkage disequilibrium (LD), fewer markers are needed to get the same coverage as elsewhere. ***The population of Iceland offers such an advantage and may provide an important population for the initial genome-wide association study.***

Few convincing cancer susceptibility alleles have been identified so far using the genetic association study design. The limited success of these studies can be attributed mainly to the use of small studies sizes - which provide insufficient statistical power and give a higher rate of false positives – and limitations in the selection of candidate genes. Several approaches can be used to increase the efficiency of candidate-gene association studies, such as improving the selection of candidate genes that are likely to be associated with cancer predisposition and enriching for genetic susceptibility by studying families with a history of cancer. Large sample sizes are needed to detect and confirm, at appropriate levels of statistical significance, genetic variants that confer modest risks. The chance of success can be further improved by careful selection of both candidate genes and candidate polymorphism (Pharoah *et al.*, 2004). ***All these issues have been taken into consideration in this proposal.***

Association studies have several weaknesses which have been amply demonstrated by the large number of reports of genetic associations that have failed to be confirmed. The most prominent problems are biological and phenotypic complexity, population stratification, statistical artifacts and lack of power (Cardon *et al.*, 2003). Statistical power can be enhanced by using multi-point linkage disequilibrium mapping methods. These methods also have fewer problems related to multiple testing and recent simulations have shown that they are powerful, even for low-penetrance mutations and disease heterogeneity as long as the frequency of the mutations are relatively high. (Rafnar *et al.*, 2004). The population stratification effects and phenotypic complexity may be reduced by choosing a suitable population that is homogeneous, has good disease records and sufficiently many affected individuals. ***The Icelandic population fulfils all these criteria and provides an exceptional resource for this approach.*** In previous studies using microsatellite markers, we have shown that LD is significantly more extensive in Iceland than in other European populations. Furthermore, the complete absence of LD between markers from different regions is strong evidence that population stratification is not an issue for LD mapping in Iceland.

The major critique of using isolated populations in genetic studies is that genetic variants found may not be relevant in larger, more mixed populations and also, that mutations in important genes may not be found. ***To address this problem, we plan to perform an identical genetic analysis in a Dutch population which is genetically more heterogeneous than the population of Iceland.*** By studying in parallel and combining the results from two genetically distinct populations we will be able to distinguish population specific mutations of recent origin and disease-predisposing alleles that are common to most Caucasians populations.

Objectives

1. *Determination of the contribution of polymorphic variants in a large number of candidate genes to the risk of breast and prostate cancer.*

We propose to look for associations between SNP polymorphisms in and around candidate cancer genes and the risk of breast and prostate cancer in unselected samples from breast and prostate cancer patients in Iceland and the Netherlands. This choice of populations was rationalized in the previous paragraphs. The Icelandic study population consists of a total of 1,973 breast and 760 prostate cancer patients and an equal number of age and sex matched controls. The samples have already been collected along with detailed clinical and lifestyle information, including information on treatment, outcome and accurate family information; this data will be used as covariates in the association analysis. A comparable prospective collection of breast and prostate cancer cases will be initiated in the Netherlands. This collection is necessary in order to construct a truly population-based cohort with the complete clinical, lifestyle and family data that has already been accumulated in Iceland. In total, 500 breast cancer cases and 500 prostate cases, along with 1,000 controls will be genotyped in the Dutch population. Candidate cancer genes have been selected based on their involvement in pathways that play a role in cancer and on the existence of multiple polymorphic SNPs. Genotyping will be performed using Illumina SNP genotyping service. Odds ratios for single polymorphic variants will be determined and logistic regression used to estimate independent contributions of each susceptibility gene and of combinations of

genes in a multivariate analysis, while controlling for risk factors and stratifying by clinical variables. Since SNP ascertainment relies on public resources (dbSNP, HAPMAP) and not on resequencing of candidate regions, we may not be studying the causative mutations. Thus, fine-mapping of the disease causing variants will be an integral part of this study. The relationships between different clinical parameters (grade, histology, survival, staging, presence or absence of metastases, recurrence of disease and age at diagnosis) and disease status will be investigated. Finally, we will use our genealogy data to determine if disease associations are more prominent in cases with family history of cancer, and to derive haplotypic phase of genotypes when possible.

2. *Development of efficient statistical and computational methods for the analysis of genetic and association data.*

We will develop efficient statistical and computational methods for the analysis of genetic data central to the entire field of complex disease gene finding. We will study these methods using data originating from this study, simulated data and previously published data. Two areas of research are proposed: Multipoint methods and multi-locus methods. Multipoint analysis refers to the joint analysis of multiple neighboring marker loci, with the purpose of localizing the disease locus independently of other disease loci that might exist elsewhere in the genome. The aim is to locate the disease-causing variant. Many complex diseases are caused by the interaction between multiple loci and environmental variables which exert only weak effects by themselves. Multi-locus methods attempt to capture contributions from multiple susceptibility loci and environmental variables to a given disease. Both research areas deserve further development and are central to any attempt of unraveling the genetics of complex diseases such as cancer. The SNP genotyping data to be generated in this project is an ideal testing material for this work and the resulting methods will be crucial for pharmacogenetic studies of the future.

Caveats

We do not foresee any major problems arising in this proposal. The collection of Icelandic cases and Icelandic and Dutch controls is complete and collection of the Dutch cases should not present a problem. A list of candidate cancer genes has been assembled, together with information on SNPs; the genotyping method is tested and validated. In addition, the principles of the statistical methods proposed have been established in a previous collaboration between Iceland Genomics and Bioinformatics; no unexpected problems are foreseen in the design of computational methods. The only unpredictable feature is whether we will observe any associations. However, that is an inherent problem in any projects of this nature and therefore can not be eliminated or predicted.

B.2 Relevance to the objectives of the LifeSciHealth Priority

(max 3 pages)

(Describe the manner in which the proposed project's goals address the scientific, technical, wider societal and policy objectives of the LifeSciHealth Priority in the areas concerned.)

The relevance of this proposal

Our proposal fits the scientific, technical, societal and policy objectives of the LifeSciHealth Priority very well. The objective of call LSH-2004-2.2.0-6 is to *integrate multidisciplinary research in order to further characterize known susceptibility genes and/or identify new genes, identify their molecular signatures and develop models for early detection, diagnosis and risk prediction of two or more familial cancers, such as breast, ovary, prostate, colon and skin.* We propose to do exactly this. First, we will characterize known and suspected cancer susceptibility genes in two European populations in order to gain insight into the genetic underpinnings of cancer susceptibility. We will also develop the appropriate statistical methods for this complex analysis. Second, we will focus on breast and prostate cancer, two gender-specific cancers with an immense impact on public-health. Both breast and prostate cancer have been shown to have a considerable genetic component and both are affected by hormones. Furthermore, the co-occurrence of breast and prostate cancer in some *BRCA2* families suggests that common genetic elements may affect both cancers. However, the genetic underpinnings of breast and prostate cancer also differ significantly. In prostate cancer, three fairly penetrant genes have been found within families but the prevalence of these genes among families is too low to be of clinical relevance. Lastly, we will use the extensive clinical, treatment, family and lifestyle information available to us in order to determine associations with particular disease phenotypes. This is an essential requirement for the success of any genetic study of a complex disease and sets the stage for future pharmacogenomic studies.

The polygenic nature of cancer susceptibility

The field of breast cancer genetics, where most of the effort has been focused, has not seen major novel cancer gene discoveries since the two susceptibility genes *BRCA1* and *BRCA2* were found. However, population-based epidemiology studies have shown that only 15-20% of familial breast cancers occur in families carrying mutations in the *BRCA1* or *BRCA2* genes. Although polymorphisms in several genes have been linked to prostate cancer, the identification of highly penetrant prostate cancer genes has proven elusive, leading to the recognition that no single susceptibility gene is likely to explain a large proportion of familial prostate cancer. This has led to the currently favored hypothesis that most of the inherited prostate cancer, as well as the remaining breast cancer risk, is due to multiple, moderate genetic risk variants.

If the polygenic model is correct, linkage analysis of families with multiple cases of breast or prostate cancer will not have sufficient power to detect the low-penetrance risk variants. This is supported by the large number of linkage studies in breast and prostate cancer that have failed to come up with definitive results. In addition, people with relatively high inherited risk of cancer may not have a significant family history and thereby not be observed. The association approach (i.e. the case control study) has been proposed as the method of choice for identifying low penetrance genes and indeed polymorphisms in multiple candidate genes have been reported to be associated with cancer. However, association studies have several weaknesses which have been amply demonstrated by the large number of reports of genetic associations that have failed to be confirmed. The most prominent problems are biological and phenotypic complexity, population stratification, statistical artifacts and lack of power (Cardon and Palmer, 2003). The population stratification effects and phenotypic complexity may be reduced by choosing a suitable population that is homogeneous, has good disease records and sufficiently many affected individuals. The Icelandic population fulfils these criteria and has already proven to be highly suitable for identifying genes involved in complex disease.

An association study in European populations

Using the association approach in the homogeneous Icelandic population we will both test and characterize the associations of known susceptibility genes to breast and prostate cancer and also look for new genes that might affect cancer risk. Our hypothesis is that variations in multiple genes may have minor, yet significant, effects on expression and/or function, thereby leading to effects on susceptibility

to cancer. In our search for new cancer-gene associations, we have carefully selected genes that have been shown to participate in particular cellular functions and pathways that are relevant to cancer, such as DNA repair and cell-cycle regulation. Thus, we will have a unique opportunity to probe multiple genes in each pathway in a systematic manner which will allow us to both analyze the effect of each gene separately, the combined effect of multiple genes as well as their interactions.

The caveat of homogeneous populations is that genetic changes affecting disease risk may have occurred in isolation and findings from the Icelandic population may not be transferable to other larger, populations. This is where the European collaboration is of critical importance. For various reasons discussed in section B6, we decided to start a prospective collection of truly population-based samples of breast and prostate cancer patients in the Netherlands. In this way, we can ensure that the collection of samples and data is, from the very start, comparable in content to the extensive collection of samples and data already present in Iceland.

Genetic risk profiles

When detected in its early stages, the cure rate for both breast and prostate cancer is very high. Therefore, much emphasis has been put on the development of screening tests that can detect the disease early. In the last decade, the Prostate-Specific Antigen (PSA) test has been widely used and in some countries even recommended annually for all males over the age of 50 and over the age of 45 in case of family history. Although PSA screening can detect early-stage prostate cancer, the test has low specificity which means that a considerable number of individuals go through the anxiety and follow-up testing caused by frequent false-positive results, as well as the complications that can result from treating prostate cancers that, left untreated, might not affect the patient's health. In addition, the direct cost of unnecessary interventions relating to overtreatment of these cases is very high. Regular mammography tests for breast cancer can detect cancer in its early stages, however, such tests are expensive and lead to overdiagnosis and treatment. Also, their effectiveness for premenopausal women has been cast in doubt. Both tests have to be repeated at regular intervals throughout the person's life.

If all the genetic elements that predispose to either breast or prostate cancer were known, it might be possible to construct genetic risk profiles to help delineate the subpopulations that would benefit most from cancer screening programs. Genetic diagnostic tests that identify individuals at increased risk may also lead to direct medical or lifestyle interventions, e.g. if a cancer-susceptibility gene belongs to a biochemical pathway that could be targeted by preventive drugs. This risk stratification method could thus improve intervention strategies by minimizing the screening of low risk individuals while focusing on individuals at higher risk.

Merikangas et al. (Merikangas *et al.*, 2003) have suggested that in searching for susceptibility genes, efforts should primarily be spent on diseases with a considerable genetic factor that can not be easily modified with environmental changes and where phenotyping is unequivocal. Furthermore, the disease should have high prevalence in the population and a high impact on public health. Prostate and breast cancer fulfill all these criteria and thus represent diseases with great potential for genetic testing and intervention. The definition of genetic risk factors for breast and prostate cancer is also key to further our understanding of cancer initiation and progression and may lead to effective intervention strategies, including prevention, early diagnosis and treatment. We propose to use the unique resources available to members of the consortium to perform a comprehensive study on known genetic factors and previously untested candidate genes in breast and prostate cancer in two European populations of different composition. The results cast light on how the presence of genetic variants affect the risk of cancer initiation and cancer phenotype such as progression and response to treatment. Finally, the results will serve as a starting point for the building of models of genetic risk of these cancers.

B.3 Potential impact

(max 3 pages)

(Describe the strategic impact of the proposed project, for example in reinforcing competitiveness or on solving societal problems. Describe the overall innovation aspects. Describe the exploitation and/or dissemination plans which are foreseen to ensure use of the project results. Describe the added-value in carrying out the work at a European level. Indicate what account is taken of other national or international research activities. For projects proposing the generation of exploitable deliverables, size and importance of the potential market should be considered and initial elements of a business plan presented.)

Innovation aspects

Breast cancer and prostate cancer are the most common cancers among European women and men, respectively. The annual incidence has been rising for the past two decades, reaching a rate of 92 per 100,000 individuals for breast cancer and 79 per 100,000 individuals for prostate cancer, translating into close to 210,000 and 145,000 new cases, respectively, every year (www.encl.com.fr). While non-genetic risk factors are important, family history is one of the most significant risk factors for these diseases, suggesting an important role for inheritance. Tremendous effort and resources have been spent on linkage studies of families with multiple cases of breast or prostate cancer in the hope of identifying the genes involved. However, only two genes that strongly increase the risk of breast (and prostate) cancer, *BRCA1* and *BRCA2*, have so far emerged from these studies. It should be noted that three HPC genes have been identified so far that have relatively high penetrance, however, they are too rare to be relevant at the population level. Although the identification of *BRCA1* and *BRCA2* presented major breakthroughs, they only explain the increased cancer risk in a minority of families and subsequent efforts at identifying genetic causes of breast and prostate cancer by linkage analysis have failed.

Statistical modeling suggests that the majority of inherited cancer risk not due to *BRCA1* or *BRCA2* is due to multiple risk alleles, each with moderate to low risk. If this is the case, linkage analysis of families with high rates of cancer is not a useful method for identifying the risk alleles. On the other hand, genetic association studies provide an efficient design for identifying common genetic risk variants that confer modest disease risk (Pharoah *et al.*, 2004). As described in section B1, there are many problems and pitfalls that need to be avoided in the case-control study design. A major innovation aspect of this proposal is that we will use the unique properties of the Icelandic population, along with the valuable sample and data collection constructed in conjunction with the Icelandic Cancer Project, to perform an extensive genetic association study of breast and prostate cancer. Our hypothesis is that this study setup provides an unprecedented opportunity to find cancer alleles that are associated with various forms of these complex diseases. Importantly, ICPs population based clinical genomics database and biobank provide an opportunity to correlate genetic findings with extensive clinical, epidemiological and genealogical data.

The qualities of the Icelandic populations for genetic research also carry certain limitations (see B1) and to address the possible problems arising from this, we will set up an identical case-control study in the Dutch population which is more representative of a typical Northern European population. This parallel study setup has never been done before to our knowledge and presents great opportunity for comparing these two populations, both with respect to genetic structure and genetic cancer risk.

The third innovation aspect of our proposal involves the development of statistical methods for analysis of complex genomic data. Genetic association studies have still to reach their full potential and analysis methods are still catching up with the tremendous amounts of data provided by high-throughput genotyping efforts. We will build on our previous efforts aimed at producing efficient statistical and computational methods for the analysis of genetic data central to the entire field of complex disease gene finding. The combination of two different but comparable populations, extensive data on disease phenotype and the association methodology may prove very useful in the hunt for breast and prostate cancer genes. Furthermore, our study will aid in determining the efficacy of this gene-hunting method in general and put novel statistical methods to the test.

Strategic impact

The strategic impact of this proposal is significant. If all the genetic elements that predispose to breast and prostate cancer were known, it might be possible to construct genetic risk profiles to help delineate the subpopulations that are at greatest risk of developing the disease. Genetic diagnostic tests that identify individuals at increased risk could lead to direct medical or lifestyle interventions, particularly if a cancer-susceptibility gene belongs to a biochemical pathway that could be targeted by preventive drugs. The identification of novel cancer susceptibility genes is key to further our understanding of cancer initiation and progression and may lead to effective interventions strategies, including prevention, early diagnosis and treatment of the disease. Furthermore, definition of inheritable risk factors in breast and prostate cancer will allow the construction of genetic risk profiles, which again may help in delineating the subpopulation that would benefit most from cancer screening programs.

Early diagnosis and treatment are key factors in determining survival of breast and prostate cancer patients. Regular screening of individuals at high risk for the disease greatly improve disease outcome; however, in both disease, screening programs have led to overdiagnosis and excessive treatment of localized lesions that might never progress to invasive cancer. In prostate cancer, the screening tests used have low specificity, leading to a great number of false positives. Each false positive test is costly in terms of direct monetary costs but no less in human terms as it involves unnecessary anxiety and follow-up testing that may have adverse side effects. The drastic increases in prostate cancer diagnosis in countries where the PSA screening test has come into routine use is reflective of this overdiagnosis issue and the usefulness of the test for prostate cancer screening has been hotly debated. Taken together, measures aimed at focusing the screening effort towards individuals at highest risk at developing breast or prostate cancer are needed. One way of doing this is to stratify individuals based on genetic risk in order to minimize the screening of low risk individuals while focusing on individuals at higher risk.

Exploitation/dissemination plans

There are two possible exploitable outcomes of this research. First, the identification of multiple genetic risk alleles in breast and prostate cancer are the first steps in the development of tests to assess the genetic risk profile of individuals. The decision whether to start the development of such an assay will be taken at the end of the project and the rights to such a test will be shared by the consortium members.

Second, the data analysis methods will be incorporated into software packages that can either be licensed or made available for free to the scientific community. This decision will also be made once the project is underway.

In any case, the results of the genetic association study will be published in peer-reviewed journals in collaboration by all consortium members. The collection of the Dutch population will continue after this project is completed and it will serve as material for future genetic and epidemiological studies.

Added value at European level

Breast and prostate cancer are primarily diseases of the Western world. Both are diseases of older age, their incidence is rising and the social impact and healthcare cost are escalating. Both diseases have a strong, yet largely unexplained genetic component, which, if known, might help provide more effective measures in the fight against them.

European populations offer many advantages for human genetic research, the most pronounced being advanced healthcare infrastructures, extensive population- and disease registries, clear national legislation on medical research and high education level, allowing for ease of recruitment and unequivocal informed consent process. Iceland and the Netherlands both have a rich history of genetic studies and both populations have contributed considerably to our knowledge of disease genetics. Although the social structures are similar in these countries, the populations have different genetic characteristics that will be of great value in the proposed study.

The major aim of Iceland Genomics is to set up a population-based, clinical genomics database and biobank that can be used to study all aspects of cancer. As this resource has reached a critical mass, the major problem for the company has been the lack of highly qualified biostatisticians, bioinformaticians

and epidemiologists in Iceland. Because of this, the Icelandic consortium members looked to Europe to identify strong research groups that complement the expertise available in Iceland. The Danish and UK researchers had been working on the development of analytical methods for complex genomic data and the three groups had already formed a very productive collaboration which was the spark of this proposal. The Dutch group's expertise in cancer epidemiology and access to Dutch cancer patients and controls is the ideal complement to the collaboration. We believe that this four-way collaboration is exceptionally complementary with limited overlap and that it provides opportunities for novel research not available in the respective countries alone.

A very important aspect of this collaboration is the cross-training of young scientists from the different groups. At least two Ph.D. students and 2 postdoctoral fellows will work on the project and spend time with other members of the consortium. Thus, one Ph.D. student and one fellow will split their time between the Danish and UK groups and one Dutch Ph.D. student will spend time with other consortium members. Scientists from the Icelandic group will consult with the Dutch on the collection of cancer cases and will spend considerable time working with the Danish and UK groups on data analysis, providing clinical insights into disease phenotypes and learning analysis methods.

B.3.1 Contributions to standards

(max 1 page)

(Describe contributions to national or international standards which may be made by the project, if any.)

Does not apply.

B.4 The consortium and project resources

(max 5 pages)

The participants on this project are Iceland Genomics Corporation of Reykjavik, Iceland (Drs. Steingrímsson, Rafnar, Thorlacius), Bioinformatics ApS of Aarhus Denmark (Drs. Schauer, Schierup), the University of Oxford (Dr. Hein) and the Radboud University Nijmegen Medical Centre in the Netherlands (Drs. Kiemeny, Schalken Swinkels, Van Dijck). In addition to these individuals, each site will designate a number of employees to the project, including nurses, laboratory technicians, statisticians, computer programmers, cancer clinicians and other staff, as needed.

The role of the participants, the specific skills of each and complementarity

The participants on this project complement each other in skills and experiences. Drs. Steingrímsson, Rafnar and Thorlacius have been instrumental in setting up the Icelandic Cancer Project, a major undertaking involving the collection of biological samples and detailed clinical, family and lifestyle information on practically all cancer patients in Iceland. The project and the resulting sample and data collection has been carefully designed to include relevant information on the different cancers as well as to fit international ethics standards. Steingrímsson, Rafnar and Thorlacius also have extensive experience in the genetics and biology of cancer. Steingrímsson will lead the consortium, Thorlacius will coordinate the association study of Icelandic patients and Rafnar will oversee the legal, ethical and contractual aspects of the project. Two clinical researchers work with the Icelandic team; Dr. Thorvaldur Jonsson who is a surgical oncologist with clear insight into the molecular aspects of the disease and Dr. Eiríkur Jonsson who is the Chief of Urology at the Landspítali University Hospital. These highly qualified clinicians serve essential roles in defining clinical data variables to be collected, defining disease phenotypes and solving any controversies in the interpretation of clinical data.

Drs. Schierup and Schauer are population geneticists who have worked with IGC previously on several different projects. They have experience in the analysis of population genetic data, particularly with respect to determining linkage disequilibrium and associations and their involvement in the project is therefore essential for its success. In addition to Drs. Schauer and Schierup, three statisticians and computer scientists in the Danish group will participate in the project, Drs. Wiuf, Mailund and Pedersen.

Dr. Jotun Hein of the University of Oxford is an expert in the coalescent theory and in applications thereof. The coalescent is an important method for the analysis of population structure and history and is essential for the meaningful determination and interpretation of linkage disequilibrium data. Together with the IGC group, Drs. Schierup, Schauer and Hein will select the SNPs to be used in the study. They will also develop methods for LD and candidate gene interaction analysis and will integrate their findings with data originating from this study.

Dr. Bart Kiemeny from the Radboud University Nijmegen Medical Centre is a cancer epidemiologist with extensive experience in the epidemiology of prostate cancer, including the genetic aspects of the disease. Together with Dr. Jack Schalken, head of the Laboratory for Experimental Urology (and PI of the FP6 Priority I project PRIMA: Prostate Cancer Integral Management Approach), he maintains an extensive collection of clinical data, family history, and biosamples from Dutch prostate cancer patients as well as from a large random sample of the general population. Dr. Kiemeny is also PI of a national project in which all families with hereditary prostate cancer are registered and verified and in which unaffected men are invited for screening. Dr. Kiemeny's group has a formal alliance with the population-based cancer registry of the Comprehensive Cancer Centre East-Netherlands (IKO). Dr. Jos van Dijck, a breast cancer epidemiologist and head of the cancer registry department of the IKO, will coordinate the recruitment, and collection of clinical data, of a population-based sample of breast and prostate cancer patients from the catchment area of the IKO. In collaboration with Dr. Kiemeny, Dr. Dorine Swinkels, vice-head of the Department of Clinical Chemistry, is keeping a large biobank from a random sample of the general population (from which the controls will be taken). Dr. Swinkels will also coordinate the DNA collection and storage from the two patient groups.

We feel that the team assembled is interdisciplinary in nature and therefore appropriate for taking on this important project which ranges from clinical aspects of cancer to genetics and genomics, to population genetics and epidemiology.

Human resources

I. Iceland Genomics Corporation

Eiríkur Steingrímsson, PhD. Chief Scientific Officer, Iceland Genomics Corporation, Reykjavik, Iceland and Research Professor, University of Iceland, Faculty of Medicine, Reykjavik, Iceland

Selected peer reviewed publications:

1. Kristinsson, S.Y., Thorolfsson, E.T., Talseth, B., **Steingrímsson, E.**, Thorsson, A.V., T. Helgason, Hreidarsson, A.B., Arngrimsson, R. 2001. *Diabetologia*, 44:2098-2103.
2. **Steingrímsson, E.**, Arnheiter, H., Hallsson, J.H., Lamoreux, M.L., Copeland N.G., and Jenkins, N.A. 2003. *Genetics*, 163:267-276.
3. Rafnar, T., Thorlacius, S., **Steingrímsson, E.**, Schierup, M.H., Madsen, J.N., Calian, V., Eldon, B.J., Jonsson, T., Hein, J., and Thorgeirsson, S.S. 2004. *Nature Reviews Cancer*, 4:488-492.
4. **Steingrímsson, E.**, Copeland, N.G., and Jenkins, N.A. 2004. *Annual Review of Genetics*, 38.
5. Rafnar, T., Benediktsdóttir, K.R., Eldon, B.J., Gestsson, T., Saemundsson, H., Olafsson, K., Salvarsdóttir, A., **Steingrímsson, E.**, and Thorlacius, S. 2004. *European Journal of Cancer*, in press.

Thorunn Rafnar, PhD. Chief Operating Officer, Iceland Genomics Corporation, Reykjavik, Iceland.

Selected peer reviewed publications

1. Ragnarsson GB, Mikaelssdóttir EK, Vidarsson H, Jónasson JG, Ólafsdóttir K, Kristjánsdóttir K, Kjartansson J, Ögmundsdóttir HM and **Rafnar T.** 2000 *Br. J. Cancer* 83, 1715-1721
2. Mikaelssdóttir EK, Benediktsdóttir KR, Olafsdóttir K, Arnadóttir T, Ragnarsson GB, Olafsson K, Sigurdsson K, Kristjánsdóttir GS, Imsland AK, Ögmundsdóttir HM and **Rafnar T.** 2003 *Gynecologic Oncology* 89, 22-30
3. **Rafnar T.**, Thorlacius S, Steingrímsson S, Schierup MH, Madsen JN, Calian V, Eldon BJ, Jonsson T, Hein J and Thorgeirsson S. 2004 *Nature Reviews Cancer* 4, 488-492
4. Mikaelssdóttir EK, Valgeirsdóttir S, Eyfjord JE and **Rafnar T.** 2004 *Breast Cancer Research* 6, R284-290
5. **Rafnar T.**, Benediktsdóttir KR, Eldon BJ, Gestsson T, Saemundsson H, Olafsson K, Salvarsdóttir A, Steingrímsson E and Thorlacius S. 2004 *European Journal of Cancer* (in press).

Steinunn Thorlacius, PhD. Director of Cancer Research, Iceland Genomics Corporation, Reykjavik, Iceland.

Selected peer-reviewed publications.

1. **Thorlacius, S.**, Struewing, J.P., Hartge, P., Olafsdóttir, G.H., Sigvaldason, H., Tryggvadóttir, L., Wacholder, S., Tulinius, H., and Eyfjord, J.E. (1998). Population-based study of risk of breast cancer in carriers of BRCA2 mutation. *Lancet* 352, 1337-1339.
2. Antoniou, et al.. (2003). *Am J Hum Genet* 72, 1117-1130.
3. Gudmundsdóttir, K., **Thorlacius, S.**, Jonasson, J.G., Sigfusson, B.F., Tryggvadóttir, L., and Eyfjord, J.E. (2003). *Br J Cancer* 88, 933-936.
4. Rafnar, T., **Thorlacius, S.**, Steingrímsson, E., Schierup, M.H., Madsen, J.N., Calian, V., Eldon, B.J., Jonsson, T., Hein, J. and Thorgeirsson, S. (2004) *Nature Reviews Cancer* 4, 488-492.
5. Rafnar T, Benediktsdóttir KR, Eldon BJ, Gestsson T, Saemundsson H, Olafsson K, Salvarsdóttir A, Steingrímsson E and **Thorlacius S.** (2004) *European Journal of Cancer* (in press)

II. Bioinformatics ApS

Mikkel H Schierup, PhD. Associate professor, Bioinformatics Research Center, University of Aarhus, Denmark and Chief Scientific Officer, Bioinformatics ApS.

Selected peer-reviewed publications.

1. Eskildsen, S., J. Justesen, **M. H. Schierup** and R. Hartmann (2003). *Nucleic Acids Res*, 31: 3166-3173.
2. Charlesworth, D., C. Bartolomé, **M. H. Schierup** and B. K. Mable (2003). *Molecular biology and evolution*, 20 (9): 1741-1753.
3. Charlesworth, D., B. K. Mable, **M. H. Schierup**, C. Bartolomé-Husson and P. Awadalla (2003). *Genetics*, 164: 1519-1535.
4. Bechgaard, J., T. Bataillon and **M. H. Schierup** (2004). *Journal of Evolutionary Biology*, 17:554-61.
5. Hein, J., **Schierup, M.** and Wiuf, C. (2004) Oxford University Press 0-19-852995-3

Leif Schauser, PhD. Associate professor, Bioinformatics Research Center, University of Aarhus, Denmark and Chief Executive Officer, Bioinformatics ApS.

Selected peer-reviewed publications.

1. **Schauser, L.**, Roussis, A., Stiller, J., and Stougaard, J. (1999) *Nature* 402 : 191-195.

2. Peart, J. R., R. Lu, A. Sadanandom, I. Malcuit, P. Moffett, D. C. Brice, **L. Schauser**, D. A. W. Jaggard, S. Xiao, M. J. Coleman, M. Dow, J. D. G. Jones, K. Shirasu and D. C. Baulcombe (2002). *Proc. Natl. Acad. Sci. USA*, 99 (16): 10865-10869.
3. Borisov, A. Y., L. H. Madsen, V. Tsyganov, Y. Umehara, V. Voroshilova, A. Batagov, N. Sandal, A. Mortensen, **L. Schauser**, N. Ellis, I. Tikhonovich and J. Stougaard (2003). *Plant Physiol*, 131 (3): 1009-1017.
4. Rivas S, Rougon-Cardoso A, Smoker M, **Schauser L**, Yoshioka H, Jones JD. (2004) *EMBO J*.23:2156-65.
5. **Schauser, L**, Wieloch, W. and Stougaard, J. (2004). *Journal of Molecular Evolution* (in press)

III. Oxford University

Jotun Hein, PhD. Professor of Bioinformatics, Dept. of Statistics, University of Oxford, UK.

Selected peer reviewed publications:

1. **Hein, J.**, C.Wiuf, B.Knudsen, Møller, M., and G.Wibling (2000): *J. Molecular Biology* 302.265-279.
2. Schierup, M. and **J.Hein** (2000): *Genetics* 156.897-91.
3. **Hein, J**, J.Jensen and C. Storm (2003) *PNAS* 100(25):14960-14965.
4. **Hein, J** (2001) *Pac.Symp.Biocompu.* 2001 p179-190
5. **Hein, J**, MH Schierup and C. Wiuf (2004) *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*". Oxford University Press 350 pages.

IV. Radboud University Nijmegen Medical Centre

Lambertus A. (Bart) Kiemeney, PhD. Head Cancer epidemiology research programme, Radboud University, Nijmegen Medical Centre, Nijmegen, the Netherlands.

Selected peer reviewed publications:

1. Guo Z, Linn JF, Wu G, Anzick SL, Eisenberger CF, Halachmi S, Cohen Y, Fomenkov A, Obaidul Hoque M, Okami K, Steiner G, Engles JM, Osada M, Moon C, Ratovitski E, Trent JM, Meltzer PS, Westra WH, **Kiemeney LA**, Schoenberg MP, Sidransky D, Trink B *Nat Med* 2004; 10(4): 374-81.
2. Verhage BAJ, Aben KKH, Witjes JA, Straatman H, Schalken JA, **Kiemeney LALM** *Int J Cancer* 2004; 109(4): 611-617.
3. Verhage BAJ, van Houwelingen K, Ruijter TEG, **Kiemeney LA**, Schalken JA. *The Prostate* 2003; 54: 50-57.
4. Verhage BAJ, van Houwelingen K, Ruijter TEG, **Kiemeney LA**, Schalken JA. *Int J Cancer* 2002; 100: 683-5.
5. Zeegers MP, **Kiemeney LALM**, Nieder AM, Ostrer H. *Cancer Epi Biomarker Prev* 2004; 13(11): 1765-71.

Jack A. Schalken, PhD, Professor of experimental urology, Radboud University, Nijmegen Medical Centre, Nijmegen, the Netherlands.

Selected peer reviewed publications

1. **Schalken JA**, Hessels D, Verhaegh G. *Urology*. 2003 Nov;62(5 Suppl 1):34-43.
2. de Kok JB, Verhaegh GW, Roelofs RW, Hessels D, Kiemeney LA, Aalders TW, Swinkels DW, **Schalken JA**. *Cancer Res*. 2002 May 1;62(9):2695-8.
3. van der Poel HG, McCadden J, Verhaegh GW, Kruszewski M, Ferrer F, **Schalken JA**, Carducci M, Rodriguez R. *Cancer Gene Ther*. 2001 Dec;8(12):927-35.
4. van Leenders GJ, Aalders TW, Hulsbergen-van de Kaa CA, Ruiters DJ, **Schalken JA**. *J Pathol*.2001;195(5):563-70
5. Verhaegh GW, van Bokhoven A, Smit F, **Schalken JA**, Bussemakers MJ. *J Biol Chem*. 2000 Dec 1;275(48): 37496-503.

Jos A.A.M. van Dijck, PhD. Head Department of Cancer Registry, Comprehensive Cancer Center IKO, Nijmegen, the Netherlands.

Selected peer reviewed publications

1. **Van Dijck JA**, Verbeek AL, Beex LV, Hendriks JH, Holland R, Mravunac M, Straatman H, Werre JM. *Int J Cancer*. 1996 Jun 11;66(6):727-31.
2. Fracheboud J, Otto SJ, **van Dijck JA**, Broeders MJ, Verbeek AL, de Koning HJ; National Evaluation Team for Breast cancer screening (NETB). *Br J Cancer*. 2004 Aug 31;91(5):861-7.
3. Siesling S, **van Dijck JA**, Visser O, Coebergh JW; Working Group of The Netherlands Cancer Registry. *Eur J Cancer*. 2003 Nov;39(17):2521-30.
4. Jacobs HJ, **van Dijck JA**, de Kleijn EM, Kiemeney LA, Verbeek AL. *Ann Oncol*. 2001 Aug;12(8):1107-13.
5. **Van Dijck JA**, Verbeek AL, Beex LV, Hendriks JH, Holland R, Mravunac M, Straatman H, Werre JM. *Int J Cancer*. 1997 Jan 17;70(2):164-8.

Dorine W Swinkels, MD, PhD. Vice-Head and Associate Professor, Dept. of Clinical Chemistry, Radboud University, Nijmegen Medical Centre, Nijmegen, the Netherlands.

Selected peer reviewed publications

1. **Swinkels DW**, Wiegerinck E, Steegers EA, de Kok JB. *Clin Chem*. 2003 Mar;49(3):525-6.

2. de Kok JB, Verhaegh GW, Roelofs RW, Hessels D, Kiemeny LA, Aalders TW, **Swinkels DW**, Schalken JA. *Cancer Res.* 2002 May 1;62(9):2695-8.
3. de Kok JB, Wiegerinck ET, Giesendorf BA, **Swinkels DW**. *Hum Mutat.* 2002 May;19(5):554-9.
4. **Swinkels DW**, deKok JB, Hendriks JC, Wiegerinck E, Zusterzeel PL, Steegers EA. *Clin Chem.* 2002;48(4):650-3.
5. de Kok JB, Ruers TJ, van Muijen GN, van Bokhoven A, Willems HL, **Swinkels DW**. *Clin Chem.* 2000 Mar;46(3):313-8.

Material resources

Sample resources at IGC

In 2001, Iceland Genomics Corporation (IGC) and its collaborators launched the Icelandic Cancer Project (ICP), a cancer research initiative aimed at creating a population-based clinical genomics database and biobank for the study of cancer. The structure of the ICP is described in Rafnar et al., (Rafnar *et al.*, 2004); the project is collaboration between IGC and all the entities in Iceland who deal with cancer.

All Icelandic cancer patients (and their relatives) are invited to participate, along with a randomly selected but matched control population. To date, blood samples from over 20,000 individuals have been collected; of those over 5,000 are from cancer patients, including approximately 1,700 breast cancer patients and 1,000 prostate cancer patients. All samples and data are collected with written, informed consent (for the method of collection and ethical issues, see section B7). Data collected includes an extensive lifestyle questionnaire, detailed clinical data on diagnosis, treatment and outcome and information on family history of cancer. Furthermore, where possible, fresh biopsies are obtained from cancer surgeries.

Sample resources at Radboud University Nijmegen Medical Centre

At the Radboud University Nijmegen Medical Centre, a close collaboration exists between the Departments of Epidemiology & Biostatistics, Urology, Pathology, Clinical Chemistry, and Genetics, for the study of the genetics of urological tumors. Large family case control studies were conducted on familial bladder cancer and familial prostate cancer for cluster and segregation analyses, LOH, and CGH studies. Also, several studies were done on specific genetic polymorphisms in bladder and prostate cancer candidate genes. In order to facilitate a more high-throughput approach for the latter type of studies, investments were made along three lines. First, the routine collection was started of clinical data, a lifestyle questionnaire, and DNA from all adult new patients at the Department of Urology. Thousands of patients have already been included. Second, in 2003, a random sample of the general population of the catchment area of the Medical Centre was asked to fill out a detailed questionnaire and to give 30ml of blood for serum/plasma and DNA. Questionnaire information was collected from 9,500 participants and blood from 6,700. Pilot studies are now underway based on these investments. As with the study of prostate cancer, the Department of Epidemiology & Biostatistics has a strong history in the collaborative study of breast cancer, more specifically screening of breast cancer. As is the case with prostate cancer, within this research collaboration, the Department has access to blood samples of hundreds of patients with breast cancer. However, because both collections of patients are referral-based (university clinic) instead of population-based, it has been considered necessary to collect new samples. For this, all breast and prostate cancer patients diagnosed in the last 2 years and registered by the cancer registry of the Comprehensive Cancer Centre IKO will be selected (approx. 600 PC and 1,100 of BC patients per year). All clinical information and pathology records are already available in the cancer registry. More than 80% of these patients are diagnosed and treated in the Radboud University Medical Centre and three large community hospitals with which close collaborations exist. It is expected that DNA will be collected from 600 prostate cancer patients and 600 patients with sporadic breast cancer.

Laboratory facilities

IGC

Laboratory facilities The laboratory is equipped with all the necessary equipment for processing DNA and plasma from blood and for storing the resulting materials. The laboratory and the freezer room are also equipped with computers and bar-code scanners in order to keep track of samples and sample processing. The sample processing unit and biobank operate under strict quality control procedures and the principles of Good Laboratory Practice. The laboratory has standard equipment for DNA analysis and molecular biology work, and a Tecan Genesis RSP 150 robot and a Robbins Scientific Hydra96 pipetting station for aliquoting samples.

Databases and Software: Oracle 8.1.7 database installed on an IBM E-server Pseries running AIX. The database is segmented into separate schemas containing: i) data regarding biological samples, ii) participant self-reported data, iii) genealogical data, iv) clinical information, and v) lab experiments and results. Query tools can access multiple schemas and all personal identifiers are encrypted with the same key, thus allowing cross-reference of the various types of data. For security reasons, programs that update the database are limited to a single schema. The Laboratory Information Management System (LIMS), as well as other software used to record and update information in the database, were developed by the IGC Software Department.

Bioinformatics ApS

Hardware and Software. Bioinformatics ApS has developed software for the analysis of genetic data from case-control studies. This software is designed for the finemapping of disease mutations on the basis of a coalescence model of cases. The parameters are estimated using a Markov-Chain Monte Carlo algorithm. Bioinformatics ApS has also developed software for the simulation of chromosomes. A 15 CPU Linux-cluster is in place to do the intensive calculations.

Radboud University Nijmegen Medical Centre

Laboratory facilities Both the laboratory of the RUN-MC Department of Clinical Chemistry and the laboratory of the Department of Genetics' subdepartment of DNA diagnostics, is equipped with all the necessary equipment for processing DNA and plasma from blood and for storing the resulting materials. The laboratory and the freezer room are also equipped with computers and bar-code scanners in order to keep track of samples and sample processing. The sample processing unit and biobank operate under strict quality control procedures and the principles of Good Laboratory Practice. The subdepartment of DNA diagnostics offers facilities for sequencing and mutation detection using a 16-capillary automated sequencer (ABI 3100), two 48-capillary automated sequencers (ABI 3730), and a pyrosequencing instrument (Pyrosequencing).

Involvement of SME's

Two SMEs are involved in this project. *The Iceland Genomics Corporation (IGC)* is a privately held cancer biology company using an innovative 'clinical genomics' approach to understand the underlying mechanisms of cancer, isolate and characterize new therapeutic targets for cancer, assess outcomes of specific therapies in genetically defined subcategories of cancer patients and optimize clinical trials. To this end, the company has established the Icelandic Cancer Project, a comprehensive survey of cancer in the entire nation. *Bioinformatics ApS* is a bioinformatics company developing software for the analysis of population genetic data. The tools developed by the company enable the identification of genetic factors that influence disease risk and history. Common diseases, such as cancer, involve, in addition to environmental factors, multiple genetic components, each of which contribute to the disease risk. The analysis of these complex data sets must build on powerful statistical methods; the company has used its knowledge and experience to build such a model and algorithms for putting them to practical use. The collaboration of the companies proposed here involves the integration of clinical and genetic data on cancer in Iceland and the Netherlands, with the algorithms and models provided by Bioinformatics. This collaboration is therefore an essential component of this application.

Participation of women

Women are involved in all aspects of this proposal, from management to laboratory work and sample collection. Drs. Rafnar, Thorlacius, van Dijck and Swinkels are women. Furthermore, the project involves the analysis of the genetics of breast cancers, a cancer which primarily affects women and is a major health concern among European women.

Overall financial plan

We feel that the budget of this application is appropriate for the task proposed. In addition to personnel cost and the cost of sample collection and management, a major component of the budget involves genotyping cost, an essential part of an application dealing with the genetics of cancer. All genotyping will be contracted to Illumina, a company providing a high-throughput genotyping service at the lowest cost available. We plan to generate 8.1 million SNP genotypes in this project, each at a cost of US\$ 0.08. Thus, the total cost of genotyping will be approximately €850.000.

STREP Project Effort Form
Full duration of project (Total workforce*)
(insert person-months for activities in which partners are involved)

	IGC	Bioinformatics	Oxford	Radboud	Partner 5 short name	etc	TOTAL PARTNERS
Research/innovation activities							
WP1 Sample collection	3	0	0	80			83
WP2 Association studies	75	24	9	13			121
WP3 Bioinformatics	4	54	32	8			98
Total research/innovation	82	78	41	101			302
Demonstration activities							
WP1 Sample collection	0	0	0	0			0
WP2 Association studies	0	0	0	0			0
WP3 Bioinformatics	0	6	0	0			6
Total demonstration	0	6	0	0			6
Management activities							
WP1 Sample collection	1	0	0	2			3
WP2 Association studies	6	3	1	3			13
WP3 Bioinformatics	1	6	3	1			11
Total management	8	9	4	6			27
TOTAL ACTIVITIES	90	93	45	107			335

* EU funded plus non-EU funded

B.4.1 Sub-contracting***(max 1 page)***

(If any part of the work is foreseen to be sub-contracted by the participant responsible for it, describe the work involved and explain why a sub-contract approach has been chosen for it.)

All high-throughput genotyping will be subcontracted to Illumina, a company providing genotyping services on a fee-for-services basis. This is the cheapest and fastest method available for SNP genotyping at the scale proposed in this proposal. Although this is a major cost for the project, it does not call for investment in capital equipment or the hiring of many technicians to perform laboratory work. The project will therefore focus on getting the necessary results in the fastest possible manner and analyzing the resulting data. Thus, the proposed use of Illumina genotyping will increase the competitiveness of the project since we will not lose the time involved in purchasing the necessary equipment or setting up the genotyping process involved.

The Netherlands group will subcontract part of the sample collection process to a service provider.

B.4.2 Other countries

(max 1 page)

(If one or more of the participants is based outside of the EU Member and Associated States, INCO target countries or countries having an RTD co-operation agreement with the European Community, explain in terms of the project's objectives why this/these participants have been included, describe the level of importance of their contribution to the project.)

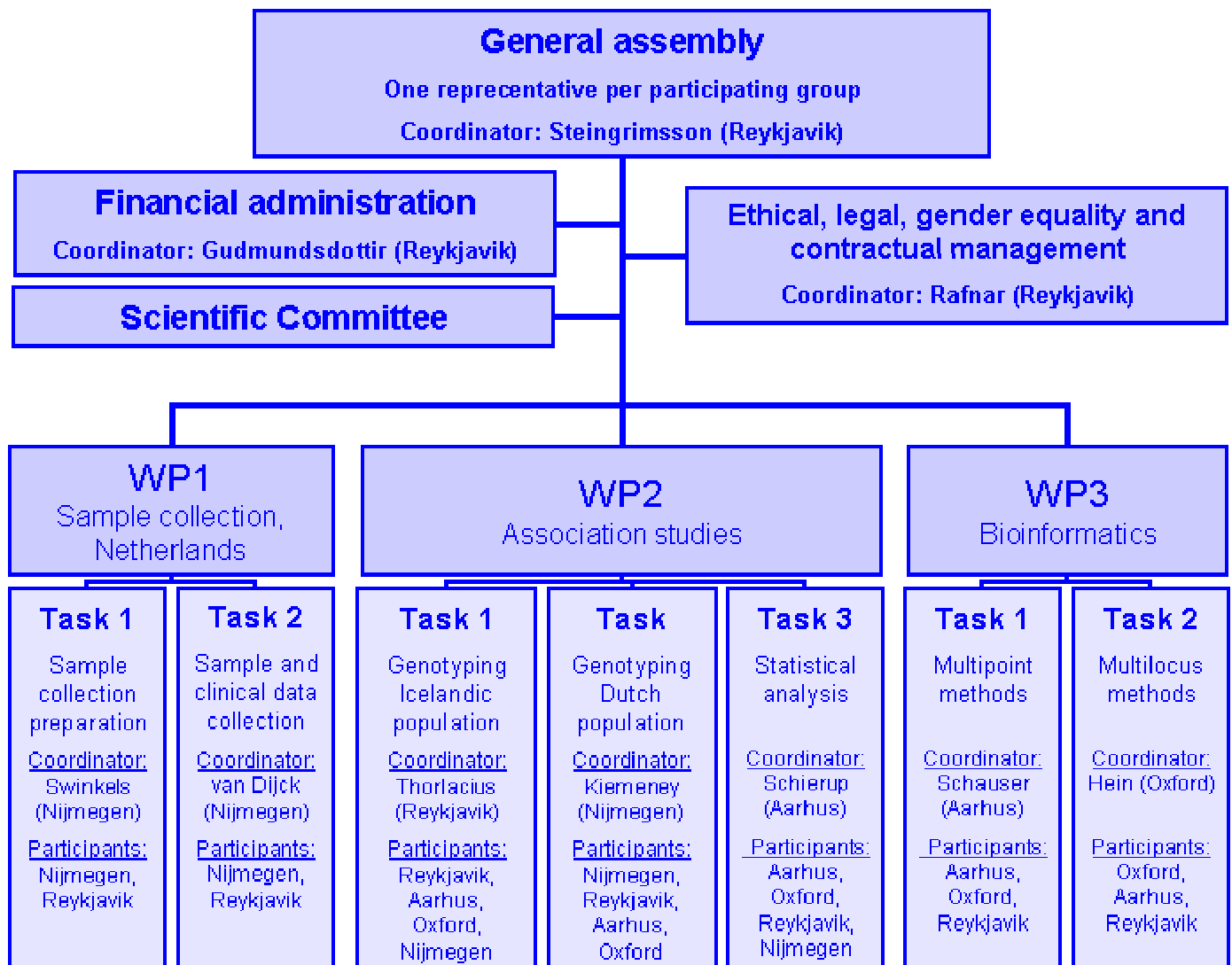
Does not apply

B.5 Project management

(max 3 pages)

(Describe the organisation, management and decision making structures of the project. Describe the plan for the management of knowledge, of intellectual property and of other innovation-related activities arising in the project.)

The organizational structure and management of the project, including the responsibilities for each task is shown below. The General Assembly consists of Drs. Steingrímsson (chair), Hein, Schauser and Kiemeny and interacts with the Commission, coordinates all aspects of the project and solves any disputes that may arise. The General Assembly has regular contact as needed. The financial management also consists of one financial administrative person from each group, chaired by Mrs. Gudmundsdóttir. The scientific committee consists of Drs. Steingrímsson, Thorlacius, Rafnar, Hein, Schauser, Schierup, Kiemeny and Schalken. After an initial organizational meeting, the Scientific Committee will have monthly meetings (telephone conference) to discuss the progress of the projects, decide the next steps and solve problems.



INTELLECTUAL PROPERTY or Ownership of Research Results

If the project is funded, the consortium members will sign a Research Agreement that describes the roles of each group and addresses the rights to intellectual property and publication. In the meantime, the consortium members have agreed to the following:

1.1 All information, data, results of research or other technology owned or controlled by IGC prior to the initiation of the Research Agreement and that are necessary and useful for the Research Agreement and any worldwide intellectual property rights arising out of any of the foregoing, are and shall remain the sole and exclusive property of IGC.

1.2 All information, data, results of research or other technology owned or controlled by Bioinformatics ApS prior to the initiation of the Research Agreement and that are necessary and useful for the Research Agreement and any worldwide intellectual property rights arising out of any of the foregoing, are and shall remain the sole and exclusive property of Bioinformatics.

1.3 All information, data, results of research or other technology owned or controlled by Radboud University prior to the initiation of the Research Agreement and that are necessary and useful for the Research Agreement and any worldwide intellectual property rights arising out of any of the foregoing, are and shall remain the sole and exclusive property of Radboud University.

1.4 All information, data, results of research or other technology owned or controlled by Oxford University prior to the initiation of the Research Agreement and that are necessary and useful for the Research Agreement and any worldwide intellectual property rights arising out of any of the foregoing, are and shall remain the sole and exclusive property of Oxford University.

1.5 All intellectual property rights derived from the analysis of genetic data under this agreement shall be considered the joint property of the Consortium, subject to any third party encumbrances, which shall be disclosed prior to the execution of the agreement.

1.6 Should novel associations be discovered, the Consortium in conjunction with intellectual property counsel, shall determine if novel IP filings are necessary. The Consortium shall submit an IP filing on a novel association discovered within sixty (60) days of the initial discovery.

1.7 The Consortium shall be responsible for all IP filings and share any cost evenly. If any party decides not to participate in the IP filing then they shall forego their participation in any downstream commercial benefit that results from direct or indirect commercial use of said IP.

1.8 Upon completion of any novel IP filings, The Consortium shall jointly publish their findings. In addition, The Consortium shall actively seek out third party collaborators to replicate their findings in independent cohorts and publish the resultant data.

1.9 IGC shall also have the sole responsibility of commercializing such novel IP through partnerships with third parties. Should IGC decide not to commercialize such novel IP or should IGC be unable to commercialize such IP after nine (9) months from the date that such IP has been filed, Bioinformatics ApS, Radboud University and Oxford University) shall have the right to commercialize such IP independently of IGC. Commercialization shall be defined as an agreement for the licensing of the Consortium IP for further development or commercial use. Any proceeds from the commercialization of the jointly owned IP will be shared evenly amongst the Consortium, excluding the cost of commercialization.

1.10 Bioinformatics ApS has previously developed software "GeneRecon", which is proposed to be enhanced as part of WP3. Bioinformatics ApS reserves the sole right for marketing and selling of any enhanced version of "GeneRecon" that might be released as a result of this project.

1.11 Should no meaningful associations be found, all data generated by the Consortium, but not any pre-existing underlying data, shall be jointly owned, and may be used by each party for only the purpose as laid out in the research plan and consented to by the Icelandic and Dutch regulatory authorities.

B.6 Workplan

(up to max 15 pages, excluding forms specified below)

(This section describes in detail the work planned to achieve the objectives for the full duration of the proposed project. An introduction should explain the structure of this workplan plan and how the plan will lead the participants to achieve the objectives. The workplan should be broken down according to types of activities: Research, technological development and innovation related activities, demonstration activities and project management activities. It should identify significant risks, and contingency plans for these. The plan must for each type of activity be broken down into workpackages (WPs) which should follow the logical phases of the project, and include management of the project and assessment of progress and results.

Note: The number of workpackages used must be appropriate to the complexity of the work and the overall value of the proposed project. Each workpackage should be a major sub-division of the proposed project and should also have a verifiable end-point (normally a deliverable or an important milestone in the overall project). The planning should be sufficiently detailed to justify the proposed effort and allow progress monitoring by the Commission – the day-to-day management of the project by the consortium may require a more detailed plan.)

a) Implementation plan introduction

(explaining the structure of this plan and the overall methodology used to achieve the objectives)

The overall aim of this research proposal is to search for and characterize common variants that affect the risk of breast or prostate cancer. To try to maximize the likelihood of success, our research design involves several novel strategies. First, we will take advantage of the unique qualities of the Icelandic population for genetic association studies (explained in section B1), while at the same time performing analogous studies in the Dutch population. This is an essential step in determining if associations found in Iceland are replicated in a more mixed Caucasian population and not just a chance occurrence in Iceland and may also identify genetic elements important in the mixed population that are not found in Iceland. Second, we will focus on genes that belong to specific pathways that are known to be important in cancer, specifically steroid hormone metabolism, carcinogen metabolism, DNA repair, cell growth and differentiation, cell cycle, apoptosis, transcription or signaling. Third, we will focus on genes where SNPs have been found and validated and therefore bypass the need to embark upon a costly and time-consuming SNP validation project. Fourth, we will use advanced statistical methodology implemented by members of the consortium for the fine-mapping of disease-associated alleles and develop methods for the discovery of interactions between variants of candidate genes. Finally, we will resequence the finemapped regions from genes showing high association with the disease state from selected cases, in order to discover the disease-predisposing alleles.

Work Package 1 - collection of a population-based sample of breast and prostate cancer cases in the Netherlands

The major objective of WP1 is to acquire a truly population-based sample of Dutch breast and prostate cancer cases with complete clinical and lifestyle information. Such a collection has already been established in Iceland and for the project to reach its full potential, it is essential that the study populations and types of information gathered be comparable. It should be emphasized that although several breast and prostate cancer cohorts are already in existence in the Netherlands, many of those are not appropriate for this study. A number of study populations have focused on familial cases, in other instances there is lack of sufficient clinical information and/or enough high-quality DNA for large-scale genotyping, the informed consent does not cover the use of samples for new studies, or the sampling frame is not population-based which makes a valid case-control comparison almost impossible. Therefore, we believe it is crucial for the success of this study to start a new collection of cases that is specifically designed to include all the important variables required for a genetic study of a complex disease. The Icelandic study design will be used as a prototype, i.e. it will be ensured that all the information gathered on the Icelandic patients will also be collected for the Dutch cohort. In this manner, the comparison between the two populations will be much more reliable than if previously collected material was used.

In the Netherlands, life-style information, family history of cancer, reproductive and medical history as well as blood samples are available from a group of 6,700 population controls. These controls were collected in a survey in 2002-2003 based on a random sample of the population registration. From this group 500 female and 500 male controls will be selected, frequency age-matched to the patient population.

WP1 - Task 1: Preparation of collection.

In this task, we will obtain all necessary permissions for the study and define the clinical and lifestyle information that will be collected. Next we will identify 1,000 prostate cancer patients and 1,000 breast cancer patients diagnosed in the last two years, from the Cancer Registry to ensure that we will reach at least 500 participating patients in each group. At the same time, we will train the personnel responsible for patient contact, sample and data collection and DNA isolation and set up the database infrastructure to handle the data. We expect that this preparation will take about 6 months from the time funding is available.

WP1 - Task 2. Collection of 500 breast cancer and 500 prostate cancer cases.

In order to obtain a population-based sample of prostate cancer patients we have decided to recruit patients (N=500) from the regional population-based cancer registry of the Comprehensive Cancer Center IKO. In the registry, clinical and pathology data of all patients have been extracted from the medical files in the hospitals where the patients were treated. IKO personnel will invite the patients for participation by a letter signed by the treating physicians. If the patients want to participate, they will identify themselves to the researchers by a preposted answering card. Subsequently, they will receive a questionnaire and specific information regarding the blood collection. It will be suggested to the patients that the Interregional Red Cross Thrombosis Service in Nijmegen (or one of the other services in the region) can draw blood. In that case, plastic EDTA tubes and identifying stickers will be sent to the patient. He or she will be requested to visit one of the addresses of the Thrombosis Service, to have a blood sample taken by the service and to send this back to the PI (overnight) using preposted and preaddressed material. The Dept. of Epidemiology & Biostatistics has proved this procedure feasible in two other large-scale studies. Alternatively, if the patient is not able to visit one of the addresses of the Thrombosis Service or if they prefer to be visited at home, the blood samples will be collected during a home visit by the Thrombosis Service. A similar routine will be used for the patient group with breast cancer (N=500). We expect that the collection of 500 prostate and 500 breast cancer cases will start after about 6 months and be completed 24 months after the initiation of the project.

Milestones: At the end of 30 months we will have constructed an extremely valuable collection of samples and data for genomic research of breast and prostate cancer that can be built on in the future.

Consortium members involved: This work package will necessarily be mostly the responsibility of the Dutch group. However, the experience of the Icelandic group will be utilized wherever possible, including the informed consent process and data protection issues, the definition of data collected and methods for processing and storing material.

Work package 2. Genetic association studies of Icelandic and Dutch breast and prostate cancer patients and controls

In this work package, we will determine the association of polymorphisms in and around candidate cancer genes to cancer susceptibility, using SNP genotyping. The polymorphisms will be characterized in breast and prostate cancer patients and will be compared to unaffected, randomly selected controls. Both the Icelandic and Dutch populations will be studied, the Icelandic samples and the Dutch control samples are ready to be genotyped immediately but the Dutch patient samples will be genotyped as they accumulate in year 2.

Our list of candidate cancer susceptibility genes was generated by searching the literature for cancer genes and pathways and by taking advantage of several different cancer gene databases. The databases used are Cancer Gene Census (Futreal *et al.*, 2004), Cancer Gene Database (<http://caroll.vjfi.cnrs.fr/cancergene/HOME.html>), the NCI Cancer Gene List Data (<http://lpgws.nci.nih.gov/html-cgap/cgl/>) and the Breast Cancer Database

(<http://condor.bcm.tmc.edu/ermb/bcgd/bcgd.html>). In addition, we have searched the RTCGD database (<http://rtcgd.ncifcrf.gov>) for the most common retroviral integration sites in several different mouse model systems. The list of selected genes contained several hundred genes belonging to various cellular pathways relevant to cancer. Given the current cost of SNP genotyping, we will be able to select and analyze 100 of those genes within the budget of this grant. In light of the fact that new knowledge accumulates rapidly, we will not finalize our list of candidate genes until right before genotyping starts. At that time we will select 100 genes based on the following premises: i) that the gene or surrounding region contains at least 5 validated SNPs, ii) that they belong to one of the following pathways, DNA repair, cell-cycle regulation or steroid-hormone metabolism. Genes known to interact will be included preferably.

The SNP variants to be studied will be selected (in order of priority) from SNP500, Illuminas Knowledge Resource (a proprietary database), or from the HapMap and dbSNP databases. We aim for studying 5–15 SNPs per gene, depending on the size of the gene. If possible, we will study coding SNPs, or SNPs in regulatory regions. If known, we will use tagging SNPs which identify “Haploblocks”, as defined by the HapMap project. We aim for spacing the SNPs at equal distances in and around the candidate genes. This procedure should maximize the potential for fine-mapping the disease-causing variant. After finemapping, we will resequence the mapped regions from genes showing high association with the disease state from selected cases, in order to discover the disease-predisposing alleles.

WP2 - Task 1 - Genotyping of the Icelandic study population.

The Icelandic cancer patient samples will be selected from the Biobank of the Iceland Genomics Corporation (IGC). Samples from 1,700 breast cancer patients and 1,000 prostate cancer patients are available. However, since the sample collection is retrospective for patients diagnosed before 2000, there is a survival bias in this sample set. This should be minimal among cases diagnosed during or after year 2000, when sample collection started. Thus, in this proposal, we will study all participating patients diagnosed since 2000, to get a nearly population-based sample. In addition we will study all breast, and prostate cancer patients who were diagnosed with disease stage higher than 1, before year 2000. In total this amounts to 973 breast cancer patients and 760 prostate cancer patients. For breast cancer, this covers over 70% of all breast cancer patients diagnosed in Iceland in 2000–2003. All patients have a clinically verified disease and clinical information such as tumor stage, laterality, grade, histology, treatment and disease recurrence is available for each patient. All participants in the ICP fill out a life-style questionnaire where information on a number of potential risk factors, including height, weight, smoking history and family history of cancer is given. Patients with breast cancer and female controls give also thorough information on menstrual history, reproductive history and hormone use. In addition to the above-mentioned clinical data, information on all first, second and third degree relatives is obtained for each patient from the Genealogical Committee at the University of Iceland. This information is linked to data from the Icelandic Cancer Registry, to get information on cancer in relatives. This data can be used to draw pedigrees, estimate family history of cancer and check for relatedness of study participants. Houlston and Peto (Houlston *et al.*, 2003) have shown that the power of association studies can be significantly enhanced by basing them on familial cases. In addition, Antoniou *et al.* (Antoniou *et al.*, 2003) have shown that cases with bilateral disease are as powerful as cases with two affected relatives. We will take advantage of this data in our association studies.

In addition to the patients, 1,733 control samples will be genotyped for the same markers. The controls were randomly selected from the National Registry of Persons, and are age- and sex matched to the cases. DNA samples from the 1,733 patients and 1,733 controls as well as 2 individuals with known genotypes (CEPH individuals) will be plated on 384-well plates for genotyping. The plates will be sent to Illumina for genotyping using the BeadArray method. Genotyping results will be downloaded into the IGC database and the quality of the data checked by comparing genotypes of CEPH individuals with their published genotypes. For each SNP, allele frequencies will be determined. We will then use the procedure described in Guo and Thompson (Guo *et al.*, 1992) to identify departures from the Hardy-Weinberg equilibrium.

The genotyping of the Icelandic samples can start immediately and the first phase is expected to be completed in about 4 months. At this timepoint, we will analyze the data with respect to allele frequencies, adherence to Hardy-Weinberg equilibrium and determine if we need to add more SNPs to cover all the genes under study. If this is the case, we will select additional SNPs and send the samples for further genotyping to Illumina. We expect to have complete coverage of the candidate genes at the end of 6 months from the initiation of the project.

Milestones: At the end of 6 months we will have genotyped 1,733 breast and prostate cancer patients and 1,733 controls for polymorphisms in 100 genes.

Consortium members involved: The Icelandic group will be responsible for this task and send the data to other consortium members for analysis. The Danish, UK and Dutch groups will be actively involved in the SNP selection process, in particular, the Danish and UK groups will advise on the distribution and density of SNPs with regards to subsequent data analysis.

WP2 - Task 2. Genotyping of the Dutch study population.

The breast and prostate cancer cases collected in the Netherlands will be genotyped for the same polymorphisms and using the same procedures as described in WP2-Task 1. For practical purposes, it is most efficient to genotype cases in multiplicities of 92 and SNPs in multiplicities of 384. The Dutch samples will be sent for genotyping as they accumulate, based on this criterion. We expect that genotyping can begin early in year 2 and will be completed 30 months after the initiation of the project.

Milestones: At the end of 30 months we will have genotyped 1,000 breast and prostate cancer patients and 1,000 controls from the Netherlands for multiple polymorphisms in 100 genes.

Consortium members involved: The Dutch group will be responsible for this task. The Dutch group will prepare the samples for genotyping and receive the data for distribution to other consortium members. The Icelandic group will be actively involved in the SNP selection, using the experience gained in the completed genotyping of Icelandic samples. The Danish and UK groups will consult on any additional SNPs required for covering all the genes under study.

WP2 - Task 3 Statistical analysis of genotyping data

Analysis of variance will be conducted in order to estimate the effect of the measured environmental and lifestyle variables. These effects will not be considered in the following statistical analysis, where only the residuals will be associated with the genetic data. The data will be analyzed all together (metaanalysis), or separated according to origin (population wise) or in various combinations after stratification of the cases according to their *BRCA* state, family history of cancer, clinical phenotype, age of onset and other variables predictive of heritability. Simple analysis to be performed for testing for association with the disease state include single marker association using contingency table tests (Weir, 1996).

Many multi-point association tests are based on haplotype information. In order to estimate the haplotypes for each gene from the genotype data, we will employ the "Phase" software (Stephens & Donnelly, 2003). Haplotype association of individual genes will be tested in order to increase power and reduce multiple testing problems. Permutation tests can be used in order to calculate the significance of association. The Haplotype Pattern-Mining algorithm will be used in order to identify ancestral segments (identical by descent, (Toivonen *et al.*, 2000)). For genes showing significant single marker or haplotype associations, the finemapping of disease causing variants will be accomplished by using GeneRecon (Rafnar *et al.*, 2004), which will be further developed in WP3. This software is able to analyze genotype data, and hence is robust for/against errors introduced during phase inference.

Multi-locus methods:

For identification of interaction between variants of different genes, standard logistic regression will be used to identify the models best explaining the disease state. The multifactor-dimensionality reduction method (Hahn *et al.*, 2003) will also be applied to the data. These findings will be compared to the software we propose to implement and improve in WP3, namely Monte Carlo Logic Regression.

Milestones: By the end of the first 12 months we will have completed the initial analysis of the Icelandic cases which can then be submitted for publication (month 14). By the end of month 24, we will have analyzed the Icelandic data with respect to multiple clinical and lifestyle variables. And by the end of month 33, we will have completed the analysis of all the Dutch data and written a manuscript detailing that part of the study. The comparison of the two populations will be completed and ready for publication by the end of month 36.

Consortium members involved: The Danish and UK groups are responsible for this task. The Icelandic and Dutch groups will work closely with the Danish and UK groups in the data analysis, providing biological and clinical insights into the data sets. To disseminate statistical experience and know-how, members of all groups will meet regularly (monthly or bimonthly) to go over data.

Caveats and alternative approaches

We do not see any major caveats in this part of the proposal. The genotyping technology is robust and of high quality. We can not predict the frequency of polymorphic SNPs in the Icelandic population and this may affect the ability to confirm our conclusions in the Dutch population. We will type SNPs regardless of whether they are likely to have effects on protein function or not; we expect that few of the SNP's will directly alter protein function. It may prove difficult to find the appropriate number of SNPs per gene of the desired frequency in order to achieve statistically significant results. Our starting list of genes to study, however, contains many genes and we believe our results will not suffer significantly by abandoning a few genes due to lack of appropriate SNPs for genotyping. We expect that at least 50% of the SNPs validated in a Caucasian population will be found in Iceland. This is a very conservative estimate when our previous experience with SNP genotyping in Iceland is taken into account. This should be enough to get statistically significant results in each of the gene regions.

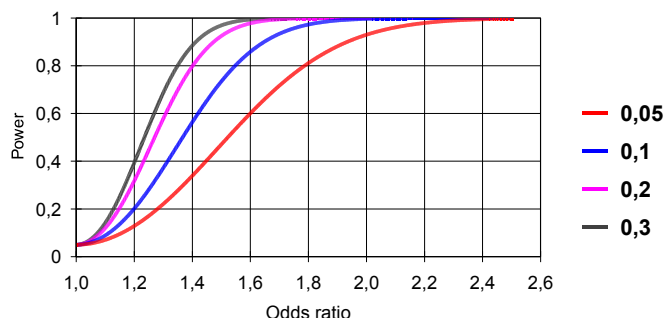


Figure: Power of the proposed Case-Control Study to detect associations with risk allele of varying frequencies and with a Type I error rate of 5%. The sample size is 500 cases and 500 controls. Abbreviations: p_0 , frequency of the predisposing allele; The Graph was plotted with the PS power and sample-size program (available at <http://www.mc.vanderbilt.edu/prevm/ps; Dupont and Plummer, 1998>).

Power analysis shows that using 500 cases and controls, this study has power of 90% of detecting association to risk allele with odds ratio as low as 1.5 if it occurs at a frequency of $> 0.2\%$ in the population when using a homogeneity test. Caution should be taken, since we might not study the causative SNP itself, but some marker in LD with it. This raises the number of individuals needed in the study depending on the amount of LD in the region. Enrichment for family history and excluding *BRCA1* and *BRCA2* mutation carriers might enhance the power of this approach considerably. Antoniou and Easton (2003) (Antoniou and Easton, 2003) showed that, relative to a standard case-control association study with cases unselected for family history, the sample size required to detect a common disease-susceptibility allele was typically reduced by more than twofold if cases with an affected first-degree relative were selected, and more than fourfold if cases with two affected first-degree relatives were used. Based on earlier observations of breast and prostate cancer in Iceland, we expect the proportion of cases with a family history to be in the order of 10%. Bilateral breast cancer cases might offer a similar gain in power to cases with two affected first-degree relatives.

Work Package 3 - Development and application of bioinformatic solutions

This Workpackage has two major objectives: method development and application thereof. Several processes will be assisted by this WP: (1) study design, (2) SNP selection and (3) the statistical analysis of the data originating from this study. The focus of this Workpackage is the development of

efficient statistical and computational methods for the analysis of genetic data central to the entire field of complex disease gene finding. We will study these methods using data originating from this study, simulated data and previously published data. Two areas of research are proposed: Multipoint methods and multi-locus methods. The term ***multipoint analysis*** refers to the joint analysis of multiple neighboring marker loci, with the purpose of localizing the disease locus independently of other disease loci that might exist elsewhere in the genome. This is also termed "fine mapping". Many complex diseases are caused by the interaction between multiple loci and environmental variables which exert only weak effects by themselves. ***Multi-locus methods*** attempt to capture contributions from combinations of multiple susceptibility loci to a given disease. Both research areas deserve further development and are central to any attempt of unraveling the genetics of complex diseases.

WP3 - Task 1. Development of multipoint methods

Multipoint methods take all marker information at a locus into account simultaneously and have potentially increased resolution and power as compared to the more traditional single marker association tests. The pattern of linkage disequilibrium around a susceptibility locus is determined by the genealogical process relating the individual subjects of a study. Present day chromosomes have been shaped by mechanisms determined partly by processes related to reproduction (mutation, recombination and gene conversion), by selection, demographic history, and by drift, the random change in allele frequencies. These processes are mathematically described by "coalescent theory". Chromosomes that share a disease-influencing variant are likely to be more closely related at the disease gene locus and some surrounding region determined by LD than two other chromosomes that do not share the variant, and thus are likely to have more markers in common. The coalescent provides a probabilistic framework for quantifying such statements.

Bioinformatics ApS has previously developed software for the finemapping of disease genes called GeneRecon (Rafnar *et al.*, 2004). This software conducts multipoint analysis by modeling the unknown genealogy that relates diseased individuals, examined in a genetic study of the case – control type. The modeling of the *Coalescent with recombination* is a computationally hard problem which GeneRecon approaches by Markov-Chain Monte-Carlo technology in a Bayesian setting.

Case - control studies are the design of choice for diseases with genetic components of low-penetrance and low inheritance, to which most common disease belong. Genetic heterogeneity and environmental effects are integrated aspects modeled by GeneRecon. We propose to extend the capabilities of the current software to *incorporate the emerging knowledge* about local variation in recombination gene conversion rates. The incorporation of knowledge about these parameters into the analysis will further enhance the precision of the prediction of the mutation position. The HapMap project is designed to provide this knowledge. Potentially, the HapMap project might also provide knowledge about the phasing of the genotype data. Such knowledge should also be included into the analysis, as the unknown phase problem is a large computational burden, and the gain in speed and accuracy by having good initial haplotype guesses might be substantial. The Bayesian formulation of the problem makes the incorporation of prior knowledge about the parameters straightforward.

To enhance the rate of convergence and the mixing of the Markov Chains, we propose to develop *heuristics that make good initial guesses of the parameters*. The most relevant is given by the tree-space of the coalescent. Different tree building algorithms will be tested for this purpose. By considering blocks of markers around a proposed mutation location the mixing problem can be addressed efficiently. However, it is unclear at the moment, what the optimal number of neighboring markers should be.

We also propose to expand the utility of our software by developing methods that can not only efficiently analyze dichotomous disease states (healthy – diseased) but also *quantitative traits*, which should result in further increase in power and resolution, since no information is discarded. This can be achieved by weighing the contribution of a case to the genetic signal according to its effect size. For this purpose, assayable phenotypes with continuous traits should be used. For example, women with extensive dense breast tissue visible on a mammogram have a risk of breast cancer that is 1.8 – 6.0 times that of women of the same age with little or no density, and breast density has been shown to

be heritable (Boyd *et al.*, 2002). Aggressiveness and age-of-onset are other continuous traits that could be used in such an analysis.

The precise *limits* of GeneRecon with regard to the effect size need to be explored by simulations and compared to that of simple methods such as homogeneity tests. Our preliminary results show that GeneRecon is able to detect signals generated by the genetic contributions of as few as 10% of 1000 cases. It appears that the genotypic relative risk (GRR) is not important for the ability to detect this signal (Rafnar *et al.*, 2004). These findings deserve further study. On the contrary, simple methods are sensitive to GRR and need large sample sizes in order to detect significant signals below a GRR of 1.5. Simulation studies can determine the limits and advantages of both (and other) methods.

A final proposition is to use the sampling from the Markov Chain in order to *partition the cases* in order to determine those individuals that carry a given disease mutation (vs. genetic heterogeneity and environmental effects). Again the cases should be weighted according to a quantitative trait, if possible.

WP3 - Task 2 Development of multi-locus methods

The focus of this task is the complex interplay between variants of genes acting on the same pathway as parts of multi-protein complexes. An alternative to parametric statistical methods such as logistic regression is to implement statistical and computational methods that have improved power to identify multivariate effects. One such method, the Multifactor-dimensionality reduction method, was able to reveal in a case control study, high-order interactions between three estrogen-metabolizing genes, without significant main effects (Ritchie *et al.*, 2001). Recently, a method called "Monte Carlo Logic Regression" has been proposed as an efficient way of identifying variants in multi-gene interactions that best explain the disease state (Kooperberg *et al.*, 2004). This explorative tool makes use of an adaptive regression method that attempts to construct predictors as Boolean combinations of SNPs. Like GeneRecon, Monte Carlo Logic Regression employs Markov-Chain Monte-Carlo technology in a Bayesian setting to generate a collection of interacting SNPs that may be associated with the disease outcome.

We propose to reimplement this algorithm in order to *speed up the computation time* by using memorization techniques. Using simulation tools developed by us earlier (Rafnar *et al.*, 2004), we will *test the feasibility* of this approach. Also, the choice of *summary statistics* is important, and needs careful study. *Prior knowledge about protein-protein interactions* between the proteins studied should be incorporated into the analysis, which again is straight-forward in a Bayesian framework. Another heuristic that reduces the combination-space could be to *weigh* the loci according to their single association with the phenotype (and by extension two-loci combinations). The opportunity of applying this method on data originating from the targeted study of a pathway is unique, and will benefit both the cancer study and development of the general analysis methods.

Milestones for WP3: By month 18, the multipoint and multi-locus methods will be implemented and evaluated against other methods using simulated and real data. Two manuscripts describing the technical details and results of the comparisons will be submitted. In the period from year one until the end of the study, iterative cycles of data analysis (both simulated and originating from this study) and method improvement will optimize the efficiency of the methods. Two articles reporting the improvements and performance of the methods on the data originating from this study will be submitted by the end of year 2. Year 3 will see the release of a commercial software package for the multi-locus method, and a PhD thesis will be submitted. Finally, the main results will be published in a high impact journal and followed by reviews on the candidate gene approach and methods of fine mapping and methods to discover multi-locus interactions.

Consortium members involved: The Danish and UK groups are responsible for this task under the direction of Drs. Schauer, Schierup and Hein. The data will be supplied by Iceland and The Netherlands Groups. A PhD student from Oxford will spend one year in Denmark, whereas the Postdocs from Denmark will spend one year in Oxford each. All three people will visit and be in close

contact with the experimental environments in Iceland and The Netherlands. The Dutch PhD student will spend 12 months in Denmark.

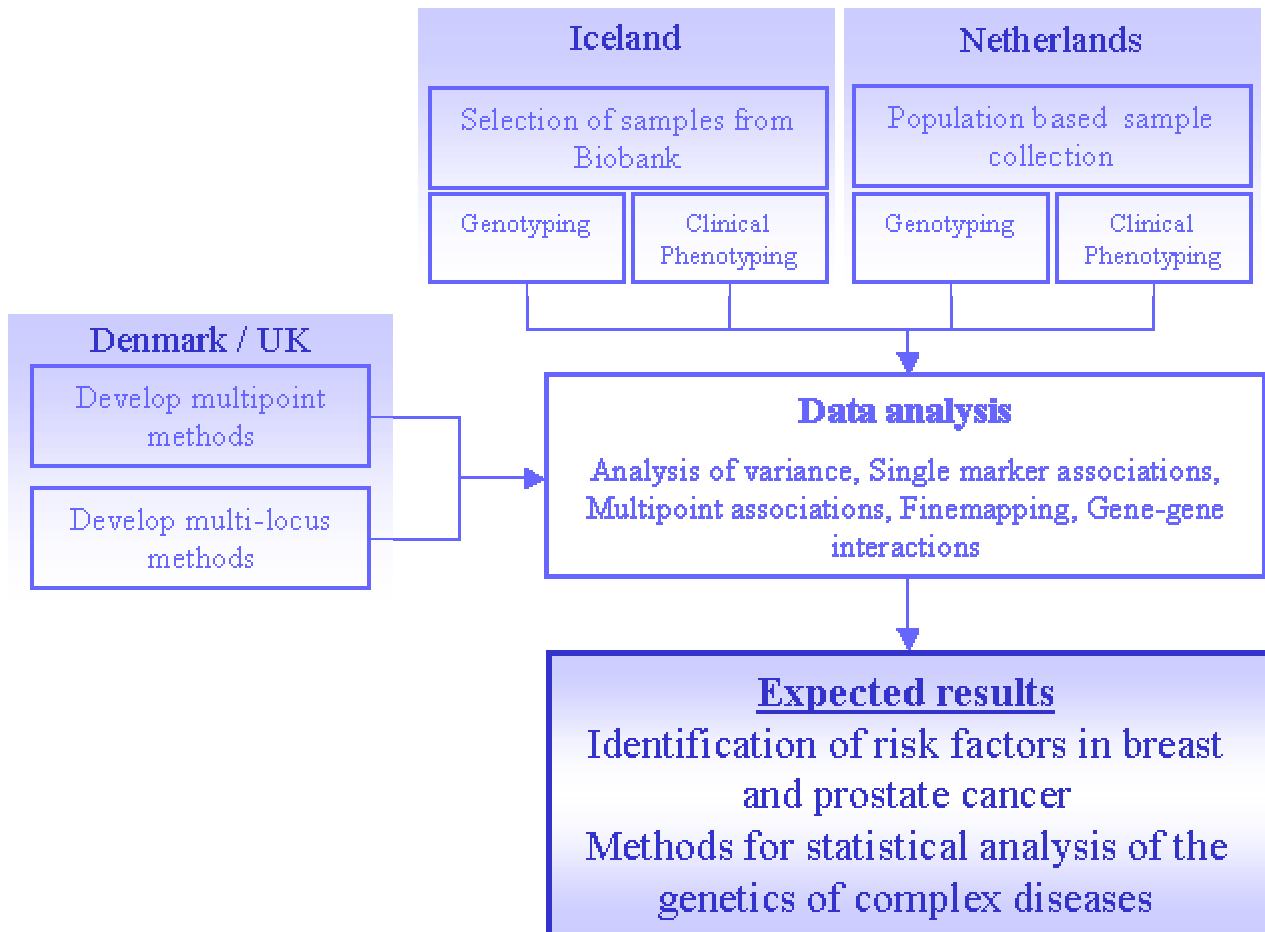
References:

- Antoniou, A.C., Pharoah, P.D., McMullan, G., Day, N.E., Stratton, M.R., Peto, J., Ponder, B.J., and Easton, D.F. (2002). A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. *Br J Cancer* **86**, 76-83.
- Antoniou, A.C., and Easton, D.F. (2003). Polygenic inheritance of breast cancer: Implications for design of association studies. *Genet Epidemiol* **25**, 190-202.
- Balmain, A., Gray, J., and Ponder, B. (2003). The genetics and genomics of cancer. *Nat Genet* **33**, 238-244.
- Boyd, N.F., Dite, G.S., Stone, J., Gunasekara, A., English, D.R., McCreddie, M.R., Giles, G.G., Trichler, D., Chiarelli, A., Yaffe, M.J., and Hopper, J.L. (2002). Heritability of mammographic density, a risk factor for breast cancer. *N Engl J Med* **347**, 886-894.
- Cardon, L.R., and Bell, J.I. (2001). Association study designs for complex diseases. *Nat. Rev. Genet.* **2**, 91-99.
- Cardon, L.R., and Palmer, L.J. (2003). Population stratification and spurious allelic association. *Lancet* **361**, 598-604.
- Cui, J., Antoniou, A.C., Dite, G.S., Southey, M.C., Venter, D.J., Easton, D.F., Giles, G.G., McCreddie, M.R., and Hopper, J.L. (2001). After BRCA1 and BRCA2-what next? Multifactorial segregation analyses of three-generation, population-based Australian families affected by female breast cancer. *Am J Hum Genet* **68**, 420-431.
- Dupont, W.D., and Plummer, W.D., Jr. (1998). Power and sample size calculations for studies involving linear regression. *Control Clin Trials* **19**, 589-601.
- Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. *Nat Rev Cancer* **4**, 177-183.
- Guo, S.W., and Thompson, E.A. (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* **48**, 361-372.
- Hahn, L.W., Ritchie, M.D., and Moore, J.H. (2003). Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics* **19**, 376-382.
- Houlston, R.S., and Peto, J. (2003). The future of association studies of common cancers. *Hum Genet* **112**, 434-435.
- Kooperberg, C., and Ruczinski, I. (2004). Identifying interacting SNPs using Monte Carlo logic regression. *Genet Epidemiol* **5**, 5.
- Lichtenstein, P., Holm, N.V., Verkasalo, P.K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A., and Hemminki, K. (2000). Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.* **343**, 78-85.
- Merikangas, K.R., and Risch, N. (2003). Genomic priorities and public health. *Science* **302**, 599-601.
- Peto, J., and Mack, T.M. (2000). High constant incidence in twins and other relatives of women with breast cancer. *Nat Genet* **26**, 411-414.
- Peto, J. (2001). Cancer epidemiology in the last century and the next decade. *Nature* **411**, 390-395.
- Pharoah, P.D., Antoniou, A., Bobrow, M., Zimmern, R.L., Easton, D.F., and Ponder, B.A. (2002). Polygenic susceptibility to breast cancer and implications for prevention. *Nat Genet* **31**, 33-36.
- Pharoah, P.D., Dunning, A.M., Ponder, B.A., and Easton, D.F. (2004). Association studies for finding cancer-susceptibility genetic variants. *Nat Rev Cancer* **4**, 850-860.
- Ponder, B.A. (2001). Cancer genetics. *Nature* **411**, 336-341.
- Rafnar, T., Thorlacius, S., Steingrimsdottir, E., Schierup, M.H., Madsen, J.N., Calian, V., Eldon, B.J., Jonsson, T., Hein, J., and Thorgeirsson, S.S. (2004). The Icelandic Cancer Project--a population-wide approach to studying cancer. *Nat Rev Cancer* **4**, 488-492.
- Risch, N. (2001). The genetic epidemiology of cancer: interpreting family and twin studies and their implications for molecular genetic approaches. *Cancer Epidemiol. Biomarkers Prev.* **10**, 733-741.
- Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F., and Moore, J.H. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* **69**, 138-147.
- The International Hapmap Consortium. (2003). The International HapMap Project. *Nature* **426**, 789-796.
- Toivonen, H.T., Onkamo, P., Vasko, K., Ollikainen, V., Sevón, P., Mannila, H., Herr, M., and Kere, J. (2000). Data mining applied to linkage disequilibrium mapping. *Am J Hum Genet* **67**, 133-145.

b) Work planning, showing the timing of the different WPs and their components
(Gantt chart or similar)

		Months					
		0-6	6-12	12-18	18-24	24-30	30-36
WP1	Task 1						
	Task 2						
WP2	Task 1						
	Task 2						
	Task 3						
WP3	Task 1						
	Task 2						

c) Graphical presentation of the components showing their interdependencies
(*Pert diagram or similar*)



*(Deliverables list, use Deliverables list form below)***Deliverables list (full duration of project)**

Deliverable No⁷	Deliverable title	Delivery date⁸	Nature⁹	Dissemination level¹⁰
D1	Population-based biorepository for breast and prostate cancer in Nijmegen	30	O	PP
D2	Database with genotypes of 3466 Icelandic cancer patients and controls	6	O	PP
D3	Database with genotypes of 2000 Dutch cancer patients and controls	30	O	PP
D4	Publication(s) of results from Association studies	34	R	PU
D5	Two articles on the implementation and comparison of the multipoint and multi-locus methods	18	R	PU
D6	Two articles reporting the improvements and performance of the methods on the data originating from this study	24	R	PU
D7	Commercial software package for the multi-locus method	24	P	PP
D8	Ph.D. thesis on statistical analysis submitted	36	O	PU
D9	Two review articles on finemapping strategies and multi-locus methods	36	P	PU

⁷ Deliverable numbers in order of delivery dates: D1 – Dn.

⁸ Month in which the deliverables will be available. Month 0 marking the start of the project, and all delivery dates being relative to this start date.

⁹ Please indicate the nature of the deliverable using one of the following codes:

- R** = Report
- P** = Prototype
- D** = Demonstrator
- O** = Other

¹⁰ Please indicate the dissemination level using one of the following codes:

- PU** = Public
- PP** = Restricted to other programme participants (including the Commission Services)
- RE** = Restricted to a group specified by the consortium (including the Commission Services)
- CO** = Confidential, only for members of the consortium (including the Commission Services)

(Description of each workpackage, use Workpackage description form below, one per workpackage)

Workpackage description (full duration of project)

Workpackage number	1	Start date or starting event:				Month 1
Participant id	RUN-MC	IGC				
Person-months per participant	74	2				

Objectives

- Construction of a population-based repository of Dutch breast (500) and prostate (500) cancer cases with complete clinical and lifestyle information.

Description of work

- Obtaining of necessary permissions, preparation of questionnaires and definition of clinical data to be collected.
- Selection of breast and prostate cancer cases from the regional population-based cancer registry of the Comprehensive Cancer Center IKO. Matching of 500 male and 500 female controls to the two patient groups.
- Collection of questionnaire data from the recruited patients.
- Collection of clinical information from the cancer registry.
- Collection of blood samples from the recruited patients.
- Isolation of DNA from the blood samples.
- Banking of samples and data.

Deliverables

A population-based biorepository of high quality DNA samples and clinical and lifestyle data from Dutch breast and prostate cancer patients

Milestones¹¹ and expected result

All preparatory work will be completed by month 6. The repository will contain samples and data from 500 breast cancer patients and 500 prostate cancer patients, as well as controls by month 30.

¹¹ Milestones are control points at which decisions are needed; for example concerning which of several technologies will be adopted as the basis for the next phase of the project.

(Description of each workpackage, use Workpackage description form below, one per workpackage)

Workpackage description (full duration of project)

Workpackage number	2	Start date or starting event:				Month 1
Participant id	IGC	RUN-MC	Bioinf	Oxford		
Person-months per participant	66	10	18	6		

Objectives

- Analysis of the association of 100 candidate cancer susceptibility genes to breast and prostate cancer in the Icelandic and Dutch populations.

Description of work

- 1733 cancer patients and 1733 matched control individuals from the Icelandic population will be genotyped for 100 candidate genes, using the SNP analysis technology from Illumina. Each gene will be probed with 15 SNPs.
- 1000 cancer patients and 1000 matched control individuals from the Dutch population will be genotyped for 100 candidate genes, using the SNP analysis technology from Illumina. Each gene will be probed with 15 SNPs.
- The genotypes will be analyzed in association with available data on family history of cancer, disease phenotype and lifestyle information.

Deliverables

Databases containing a large number of genotypes from the Icelandic and Dutch populations.
Associations between breast and prostate cancer and 100 candidate genes confirmed or rejected.
Publication(s) describing the results of the analysis.

Milestones¹² and expected result

Genotyping of the Icelandic samples will be completed by month 6. Genotyping of the Dutch samples will be completed by month 30. Analysis of data will start as soon as the first data is in-house and is expected to be completed by month 36.

¹² Milestones are control points at which decisions are needed; for example concerning which of several technologies will be adopted as the basis for the next phase of the project.

(Description of each workpackage, use Workpackage description form below, one per workpackage)

Workpackage description (full duration of project)

Workpackage number	3	Start date or starting event:				Month 1		
Participant id	Bioinform.	Oxford	IGC	RUN-MC				
Person-months per participant	54	30	4	6				

Objectives

- Development of efficient statistical and computational methods for the analysis of genetic data central to the entire field of complex disease gene finding. Two areas of research are proposed: Multipoint methods and multi-locus methods.

Description of work

- The current finemapping software "GeneRecon" will be extended in order to be able to handle QTLs and variation in recombination rates.
- The algorithm behind Monte Carlo Logic Regression will be implemented and enhanced in order to maximally meet the opportunities arising from this study.
- The data originating from this study will be analyzed.

Deliverables

Two articles on the implementation and comparison of the multipoint and multi-locus methods
 Two articles reporting the improvements and performance of the methods on the data originating from this study
 Commercial software package for the multi-locus method
 Ph.D. thesis on statistical analysis
 Two review articles on finemapping strategies and multi-locus methods

Milestones¹³ and expected result

Two articles on the implementation and comparison of the multipoint and multi-locus methods will be submitted by month 18
 Two articles reporting the improvements and performance of the methods on the data originating from this study will be submitted by month 24
 Commercial software package for the multi-locus method will be ready for marketing by month 24
 Ph.D. thesis on statistical analysis will be submitted by month 36
 Two review articles on finemapping strategies and multi-locus methods by month 36

¹³ Milestones are control points at which decisions are needed; for example concerning which of several technologies will be adopted as the basis for the next phase of the project.

B.7 Ethical, safety and other EC-policy related issues

Ethical, legal, social and safety issues

(Identify the ethical, legal, social and safety issues that may be raised by the subject and activities of the proposal, show they have been adequately taken into account - indicate which national and international regulations are applicable and explain how they will be respected. Explore potential ethical aspects of the implementation of project results¹⁴.)

Other EC-policy related issues

Identify relevant EC-policy related issues (e.g. Life sciences and biotechnology – A strategy for Europe (COM(2002) 27)), and show how they are taken into account. Demonstrate a readiness to engage with actors beyond the research to help spread awareness and knowledge and to explore the wider societal implications of the proposed work; if relevant set out synergies with education at all levels.)

B.7.1 Ethical aspects

(Due to the particular relevance of ethical issues for the LifeSciHealth Priority, include the Ethical issues form given below)

Ethical issues form

A) Specify if your project involves:

Does your proposed research involve:	YES	NO
• Human beings	X	
Persons not able to give consent		X
Children		X
Adult healthy volunteers	X	
• Human embryos		X
• Human biological samples	X	
Human embryonic stem cells		X
Human foetal tissue/cells		X
• Human genetic information	X	
• Other personal data	X	
Sensitive data about health, sexual lifestyle, ethnicity, political opinion, religious or philosophical conviction	X	
• Animals (any species)		X
Non-human primates		X
Transgenic small laboratory animals		X
Transgenic farm animals		X
Cloning of farm animals		X
• Developing countries (e.g. clinical trials, use of human and animal genetic resources...)		X
• Dual use		X

If you answer 'YES' to any of these, please refer to the following web address ("Crucial information") for guidelines on addressing ethical issues:

http://europa.eu.int/comm/research/science-society/ethics/rules_en.html

B) Confirm that the proposed research does not involve:

¹⁴ Further information on ethics requirements and rules are given in Annex 3 of the Guide for Proposers and on the science and ethics website at: http://europa.eu.int/comm/research/science-society/ethics/ethics_en.html.

- research activity aiming at human cloning for reproductive purposes,
- research activity intended to modify the genetic heritage of human beings which could make such changes heritable¹⁵,
- research activities intended to create human embryos solely for the purpose of research or for the purpose of stem cell procurement, including by means of somatic cell nuclear transfer.

The proposed research does not involve any of the issues listed in point B.7.1.b)	CONFIRM
	No Involvement
	X

C) The applicants are requested to address the ethical issues related to the proposed research: *Describe the potential ethical aspects of the proposed research regarding its objectives, the methodology and the possible implications of the results; Explain and justify the research design; Indicate the relevant national legislation or requirements of the Member State(s) where the research takes place.*

The ethical aspects of genetic studies have been the focus of intense debate for decades. The great successes in the identification of the causative factors in various diseases are hailed but, with the exception of some metabolic disorders, they have not resulted in effective preventive measures, therapy or cure. The ethical aspects of genetic research are particularly important in monogenic diseases, where the possession of a single disease-causing allele carries a high risk of developing disease. A striking case of this are mutations in the HD gene in Huntingtons Disease where the effect of the mutations appears later in life and no preventive or therapeutic measures can be offered. Another example, closer to this proposal, are mutations in the BRCA genes that carry a high risk of cancer, in particular breast cancer where it has been estimated that a lifetime risk of breast cancer in a mutation carrier is over 80%. However, in this case, mutation carriers can be offered some measures of protection, such as frequent screening, chemoprevention (e.g. Tamoxifen) and in more drastic cases prophylactic surgeries. However, the personal, psychological and social implications of genetic screening are complex and need to be approached with great care.

Genetic research in complex, late-onset diseases are much less contentious than in monogenic disorders for several reasons. Most importantly, in these diseases it is assumed that many genes are involved, each carrying a slightly elevated risk. Therefore, it can be assumed that all individuals carry some predisposing alleles and that disease risk is not excessively high even where several of these alleles come together. In this case, there is less stigma associated with genetic research and testing and the benefits of assessing the genetic risk profile of an individual. Cancer is a disease that has a high cure rate if detected early, resulting in the implementation of public screening programs for various cancers including breast cancer. It can be envisioned that by tuning the frequency of observation to the level of genetic risk, it may be possible to greatly reduce the cost of those programs without reducing efficiency.

Prostate cancer is a good example of a complex, late-onset disease where overwhelming evidence suggests that many genes, each with a relatively small increase or decrease in risk, affect the total inherited risk of the disease. The same is also true for the great majority of breast cancer, only excluding the minority of cases where mutations in BRCA1, BRCA2 or genes underlying known cancer syndromes are the major effectors. Thus, for the two cancers under study in this proposal, a case can be made that the finding of genetic variants that increase or decrease the risk of the respective cancers will not have decisive adverse implications for the carrier but may, with time, result in a general risk assessment that will

¹⁵ Research relating to cancer treatment of the gonads can be financed.

help in monitoring those at highest risk. This development is parallel to what is taking place in cardiovascular disease research, where polymorphisms in genes in metabolic pathways have been shown to affect the risk of stroke and heart attack and are subsequently used as indicators for preventive intervention.

Research design

The research design for finding low-penetrance disease-associated genetic variants is reflective of the nature of the disease. Instead of collecting families with multiple cases with the disease, the search for low-risk genes has to occur in the population at large. This is because many individuals who never get the disease will carry one or more disease-associated variants and many individuals with relatively high inherited risk may not have a family history of the disease. Thus, in our study, we will use the genetic association (case-control) approach, where we search for difference in allele frequencies between cases and controls. This means that the only criteria for inviting a participant to join the study is that he or she has had prostate or breast cancer or that they are healthy controls. No inferences are made to family history before the participant has given their explicit consent that family history of cancer can be documented and no disease information is collected about any live individuals who have not given their consent. All samples and data are encrypted before analysis and the flow of information is strictly one-way, i.e. no genetic information is ever associated with personally identifiable information. Results are presented for groups only, not individuals.

D) The applicants are requested to address the questions 1 to 11 where relevant to their research:

D.1. National legislation and international codes of conduct

Iceland.

The intense public debate that took place in conjunction with the legislation of the Health Sector Database in 1998 resulted in highly increased awareness of the ethical, legal, social and safety issues concerning genetic and medical research. What emerged from this open, and often heated, discussion is legislation that tries to accommodate such research while strictly adhering to international standards and agreements and a population that is unusually favorably disposed towards participation in biomedical research. IGC only uses samples and data from living individuals who have signed an informed consent form and the use of these resources is limited to research on cancer.

All of IGC's research protocols have been reviewed and approved by the National Bioethics Committee (NBC). The NBC's rules of procedure accord with the applicable Recommendations of the Committee of Ministers of the Council of Europe to the Member States, the Declaration of Helsinki made by the International Medical Association, the Recommendations Guiding Medical Doctors in Biomedical Research Involving Human Subjects and the International Ethical Recommendations on Medical Scientific Research on Humans.

IGC operates within a regulatory framework, which includes the following acts, regulations, directives, standards and guidelines:

1. Acts

- Health Services (No. 97/1990)
- Rights of Patients (No. 74/1997)
- Health Sector Database (No. 139/1998)
- Biobanks (No. 110/2000)

- Data Protection (Nos. 77/2000, 90/2001, 81/2002, 77/2000)
 - Information (No. 50/1996)
 - National Civil Protection (No. 94/1962)
2. **Regulations Issued by Ministry of Health and Social Security**
 - Keeping and Utilisation of Biological Samples in Biobanks (No. 134/2001)
 - Scientific Research in the Health Sector (No. 552/1999)
 3. **Regulations Issued by the Privacy and Data Protection Authority**
 - Notification of Processing of Personal Data (No. 90/2001)
 - Obtaining Informed Consent for Scientific Research in the Health Sector (No. 170/2001)
 - Personal Data Security (No. 299/2001)
 - Processing and Storage of Samples in Biobanks (No. 918/2001)
 4. **EU Directives**
 - Directive 95/46/EC (“The Data Protection Directive”) of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data
 5. **Standards**
 - IST BS7799 Management of Information Security
 - ISO/FDIS 9001: 2000 Quality Management Systems - Requirements
 6. **Guidelines Issued by the National Bioethics Committee**
 - Informed Consent for Participation in Genetic Research Projects Utilising Biological Samples
 - Rejecting Participation in a Research Project

Netherlands

In the Netherlands, all health sciences research has to comply with a number of laws and regulations. The most important of these are:

- WBP: Law on protection of personal information
- WGBO: Law on medical treatment agreement
- WMO: Law on medical scientific research

These laws have been translated by the Federation of Medical Scientific Associations (www.fmwv.nl) into a practical research code “Code Goed Gedrag” (Code proper behavior in scientific research). The second version of this code of April 1, 2004 has been approved by the responsible governmental agencies as a legal working document for ethical, privacy, social, and safety issues in the conduct of medical scientific research. The code is in agreement with Directive 95/46/EC (“The Data Protection Directive”) of the European Parliament of the Council of 24 October 1995, with Directive 2001/20/EG of April 4, 2001 on Good Clinical Practice, and with the Declaration of Helsinki. The code has been adopted by all University Medical Centers in the Netherlands.

In addition, because of a lack of formal legislation with respect to research on (remaining) biomaterial collected for diagnostic or therapeutic reasons, the Federation of Medical Scientific Associations developed the “Code Goed Gebruik” (Code proper secondary use of human tissue) in 2002. This code has also been approved by the responsible governmental

agencies and formal legislation is being prepared now based on the Code. This code has been adopted as well by the University Medical Centers.

Every new research proposal is being submitted to a local Institutional Review Board (CMO). These CMO's examine whether a research proposal, patient information, and informed consent forms comply with the Codes and the legislation. The overall National Central Committee on Research Involving Human Subjects (known by its Dutch initials, CCMO) oversees medical research involving human subjects in the Netherlands.

D.2. Use of banked or isolated human embryonic stem cells in culture, human fetuses, and human fetal tissue.

None of this material is used in the proposed research.

D.3. Use of other human biological samples

DNA from the blood sample of study participants is used. In Iceland, the subjects donate 50 ml of blood when they enter the Icelandic Cancer Project, however, only a small portion of the resulting DNA or about 10 micrograms are needed for this particular study. The rest of the sample is stored in IGC's Biobank. In the Netherlands, study subjects' DNA is being collected from 10 ml blood samples. For a detailed description of the informed consent process and protection of personal data see the next section.

D.4. Use of personal data in bio-banking (including gene-banking)

Iceland

All individuals in Iceland who have been, or are being, diagnosed with prostate or breast cancer are invited to participate in the project. The current number of prostate cancer patients is about 1000 and 1800 for breast cancer patients. This includes all age groups. Seriously ill or institutionalized patients are not contacted.

Informed consent process. Cancer patients are identified through the collaborating cancer clinicians or the National Cancer Registry. Previously diagnosed patients receive an invitation to participate in the ICP by letter, which is signed by one or more of that particular patient's treating physicians. Newly diagnosed patients are informed about the ICP during an interview with their physician. Those patients who are willing to consider entering the study then make an appointment with a staff member at IGC's patients recruitment centre, Skulaver. To date, 92% of patients invited have agreed to participate in the study. Patients are asked to name relatives who may be contacted for participation. These relatives are subsequently sent a letter of invitation and are asked to make an appointment with Skulaver. Control individuals, randomly selected from the Icelandic Registry of Persons, are invited by mail to participate in the ICP. Individuals are selected to match the patient population with regard to age and sex.

The study is explained in further detail to prospective participants during their visit to Skulaver. Those who choose to enter the study sign an informed consent form and a separate form waiving all rights to financial gain from the studies. Participants select one of three different consent forms that allow the use of samples and information to a varying degree.

Upon entering the study, the participants:

- donate a blood sample
- fill out a lifestyle questionnaire
- permit access to genealogical data

- permit access to all clinical data relevant to their disease, including pathological records and information on treatment (surgery, radiation, medication) and outcome (patients only)
- permit access to archived tissue samples that may be stored in biobanks elsewhere (patients only)
- allow access to fresh biopsies from tissue to be removed in a forthcoming surgery (newly-diagnosed patients only)

Information on patient recruitment is kept in the Skulaver database. The database includes features that prevent the re-contacting of an individual who has declined participation, types of consent given by each individual etc.

Protection of personal data. At Skulaver every sample tube is labeled with a barcode which is linked to the participant's social security number in Skulaver's database. Once the samples arrive at IGC, the Skulaver label is removed from the tube and replaced by a "IGC barcode" label. Both barcodes are entered into the IGC database. The social security numbers of participants are encrypted by the Director of Skulaver, using an encryption process defined and overseen by the Privacy and Data Protection Authority (PDPA). A file containing the encrypted social security numbers (PN) and their corresponding Skulaver sample barcodes is brought to IGC on a removable disk by a staff member of Skulaver. The managing director of IGC's BioBank oversees processing this file to link the PNs to the samples in the IGC database. The Skulaver barcodes are then no longer needed and are erased from the IGC system; thus eliminating the possibility of them being used to trace the identity of the sample's owner.

Samples are kept in IGC's licensed Biobank and can only be accessed for studies that have been approved by the NBC. IGC's BioBank operates on firm national and international legal foundations within a well-refined and developed local regulatory environment. A description of the licenses, laws and regulations under which the BioBank operates can be found in section **D.1** above.

IGC has exclusive rights to the samples in the BioBank. Researchers outside IGC may request access to samples for well-defined studies that have been approved by the PDPA and NBC. The requests are reviewed by the board of the BioBank, which may approve or deny access; denial does not have to be justified formally to the applicants, although IGC considers it common courtesy to do so. Charges for access to samples can range from nil for studies done in collaboration with IGC to the estimated cost of collecting, processing and storing a sample. Samples can under no circumstances be sold for profit.

All participants donating samples in the ICP sign a form waiving the rights to financial gain. Patients retain the right to have samples and accompanying data removed from the BioBank upon written request. Formal procedures for removal of samples and data are in place.

Netherlands

In the Netherlands, DNA has already been collected from all controls. Samples of hundreds of prostate cancer and breast cancer patients are available as well. However, it has been decided to collect samples from a new group of patients in order to ensure that the selection of patients is population-based and that every participant signs an appropriate informed consent. The controls were invited to participate in a study on gene-environment interactions in multifactorial diseases such as cancer. All participants were fully informed about the goals and the procedures of the study. Informed consent was obtained for

- the collection of questionnaire data on life-style, medical history and family history
- the collection of DNA, serum and plasma
- the possibility to link the (encrypted) personal data to disease registries
- the verification of clinical data with medical records kept by the treating physicians
- the reporting of any clinically relevant information from DNA or serum analyses through the participants' general practitioners
- the keeping of identifying information for 25 years.

Similar informed consents will be obtained from the patients. The patients will be identified through the Regional Cancer Registry of the Comprehensive Cancer Center IKO. IKO personnel will invite the patients for participation by a letter signed by the treating physicians. If the patients want to participate, they will identify themselves to the researchers by a preposted answering card. It will be suggested to the patients that the Interregional Red Cross Thrombosis Service in Nijmegen (or one of the other services in the region) can withdraw blood. In that case, plastic EDTA tubes and identifying stickers will be sent to the patient. He or she will be requested to visit one of the addresses of the Thrombosis Service, to have a blood sample taken by the service and to send this back to the PI (overnight) using preposted and preaddressed material. The Dept. of Epidemiology & Biostatistics has proved this procedure feasible in 2 other large-scale studies. Alternatively, if the patient is not able to visit one of the addresses of the Thrombosis Service or if they prefer to be visited at home, the blood samples will be collected during a home visit by the Thrombosis Service.

All participants will be given a unique number / barcode. The informed consents with this number will be kept at the Department of Epidemiology and Biostatistics. The ID information together with the unique numbers will be entered into a database on a server of the Department of Epidemiology and Biostatistics. Identifying information will be removed from the questionnaire data and clinical data (with exception of the unique number and an encrypted form of part of the identifying information for validation reasons) after which these data will be entered into another database that is kept at another server. The same holds for the blood samples, which will be administered and stored in freezers at the Department of Clinical Chemistry, similar to the samples of the controls.

D.5. Research involving persons (individuals or populations) in particular children or persons unable to give consent, pregnant women or healthy volunteers for clinical trials

None of the above are included in the study.

D.6. Protection of personal data.

The process for obtaining informed consent and encryption of data is described in section **D.2**. The anonymisation of data is overseen by the privacy and Data Protection Authority of Iceland, and by the Institutional Review Board / CMO of the Radboud University Nijmegen Medical Centre.

D.7. Use of animals

Not applicable

D.8. Research in co-operation with developing countries

Not applicable

D.9. Local ethics committees opinions and authorizations of competent bodies

The permissions for this study have been obtained from the NBC of Iceland and will be provided in English translation before contract signature.

For the Dutch part of the study, IRB/CMO permission has been obtained already for the control group. An extension of this permission will be requested for the patient groups.

D.10. Conflict of interest

No conflict of interest is declared.

D.11. Ethical implications of research results

See introduction to this section.

B.8 Gender issues

(for further explanation see Annex 4 (General approach across the programme) and Box 2 (Specific approach for the LifeSciHealth Priority) in the Guide for Proposers)

B.8.1 Participation of women

Answer the following questions:

- Are there women directly involved:

- in the scientific management of the project?	Yes <input checked="" type="checkbox"/>	No
- in the scientific partnership as scientific team leader in the project?	Yes <input checked="" type="checkbox"/>	No
- % of women scientists involved in the project¹⁶:

⇒ Early researchers (less than 4 years after graduate)? 10..%
⇒ Experienced researchers (minimum 4 years after graduate or having a PhD)?	...30...%
- Comment and justify if necessary
- Do you plan specific measure(s) regarding women role/participation in your project?
Which? How? When?

The scientific team of the consortium consists of 4 females and 6 males. The slight preponderance of males in the consortium is largely due to the fact that the ratio of women in biostatistics is still quite low. It is our plan to recruit 2 Ph.D. students in conjunction with the project and we hope that at least one of them will be a female.

¹⁶ Definitions according to the FP6 mobility & Marie Curie activities.

B.8.2 Gender aspects in research.

(If there are gender issues associated with the subject of the proposal, show how they have been adequately taken into account.)

(Recommended length - one page)

Answer the following questions:

	Yes	No
• Does the project involve human subjects?	X	
• Does the project use human cells / tissues / other specimens?	X	
• If human subjects are not involved or human materials not used, does the research involve animal subjects or animal tissues / cells / other specimens (<i>as models of human biology/physiology</i>) in such a way that it is expected that may have implications for humans?		X
• Does the project use collection of data related to human subjects, human materials, animal subjects or animal materials	X	

A positive answer to any of these questions implies that gender/sex aspect should be taken into consideration in the research proposal.

	Yes	No
Are gender/sex differences with respect to the research documented in the literature?	NA	

If yes please give details.

A negative answer to this question may imply some innovation in the proposal towards this issue that will be taken into account in the evaluation process.

If there are gender/sex aspects in your project:

- Detail the questions addressed in their proposal related to gender/sex aspects in research.
- Comment on the expected outcome.
- Describe how the gender/sex aspects will be taken into account in the research, methodology and interpretation of their results.

If you do not consider gender/sex differences, provide justification.

- *The evaluation panel will assess the relevance of the justifications provided.*
- *Neither additional costs, nor difficulties in obtaining female cells, female tissues, female specimens, or recruiting female subjects, would not normally be considered as a valid reason for excluding gender/sex aspects ("female" includes both animal and human subjects).*

Our proposal focuses on breast and prostate cancer which are the most important, gender-specific cancers in females and males, respectively. Both diseases will be analyzed with the same methodology and the results should have similar impact. The major difference between males and female subjects in this study is that the female participants (cases and controls) answer a more detailed lifestyle questionnaire than the males. The female questionnaire includes questions on menstrual and birthing history, hormone use (contraceptive and hormone replacement therapy) and other variables that are of high relevance to cancers affected by hormones. Thus, the dataset on females is expected to yield even more information than the corresponding dataset for males.