GenomEUtwin: A Strategy to Identify Genetic Influences on Health and Disease

Leena Peltonen on behalf of the GenomEUtwin (www.genomeutwin.org)

Department of Medical Genetics, University of Helsinki and Department of Molecular Medicine, National Public Health Institute, Finland

In this issue of Twin Research, we describe different facets of a European Community funded effort, GenomEUtwin, which capitalises on eight of the world's largest and best characterised twin registers and a multi-national population cohort, MORGAM. This international study, reaching beyond the geographical borders of Europe, is based on linkage and association strategies designed to identify genetic contributors to health and disease using integrated expertise of participating groups in genetics, epidemiology and biostatistics. By merging information from numerous epidemiological and genetic databases, GenomEUtwin will create an intellectual and technical infrastructure for future genetic epidemiological studies aiming to define genetic and life style risks for common human diseases.

New Era of Genome-guided Biomedical Research and GenomEUtwin

In the current era of biomedicine, the genome-wide technologies and vast amount of information of the genome anatomy are transforming our strategies to characterize biological processes in general and, in particular, those related to human biology. The scientific community in biomedicine aims to characterize relationships between "normal" or disease related phenotypes by utilizing comprehensive genome-wide information that is produced by the Human Genome Project. One of the major challenges in this effort is to characterize how individual variations in DNA sequence, lifestyle and environment contribute to the health and disease of human kind. Investigation of these relationships is severely hindered by often imprecise definitions and measurements of disease phenotypes, as well as the heterogeneous lifestyle and environment that characterize human populations. Also, we are limited by our lack of knowledge about the mechanisms by which variations in DNA sequence are related to the life-events and environmental changes in phenotypic expression. No well-defined or detailed information so far exists on the importance of the relationship between "genetic profiles" and environment or lifestyle for any common disease. As the tools of the genome project are now in hand (Collins et al., 2003), there is no excuse for not minutely characterizing the predisposing genetic profiles underlying common traits. Based on this information we can then start to dissect the environmental and lifestyle components with high precision, provided that suitable study populations are identified and properly exploited. The resolution level of these analyses is often insufficient due to obstacles related to statistical power, especially for the detection of gene–environment interactions. Consequently, there is a wide consensus of the necessity for large, well-characterized study samples.

Special study populations and well-designed population cohorts are often ideal for genome-wide studies of a comprehensive set of disease-related phenotypes. Cohorts, un-selected for special traits or end state diagnoses, are ideal for numerous reasons. They represent the full Gaussian variety of the trait, facilitating the maximal use of quantitative information in statistical analyses. Further, as the end state diagnosis of complex diseases poorly reflects the molecular background, many traits can be best studied in population cohorts rather than the case-control study design.

In this issue of Twin Research, we want to describe different facets of a European Community funded effort, GenomEUtwin. This international study, reaching beyond the geographical borders of Europe, is based on a strategy designed to identify genetic contributors to health and disease using integrated expertise of participating groups in genetics, epidemiology and biostatistics. We hope this effort will pave the way for other comparable large scale projects of population genetics with a potential to transform the vast amount of genetic information to a new understanding of human health and disease.

Twins as Study Populations

Very few examples so far exist for identifying genes behind common traits or diseases. A gene controlling QTL for size has been isolated in tomatoes (Frary et al., 2000) and genetic variants affecting muscle mass in pigs have been identified by positional cloning (Amarger et al., 2002). Identification of a predisposing or causative human gene in Crohn's disease and Type II diabetes represent pioneering innovations in common human disease traits (Horikawa et al., 2000; Hugot et al., 2001). Genetic studies of common diseases are complicated by ascertainment bias, population admixture, dichotomization of quantitative disease traits (affected/non-affected), and inadequate sample size. Systematic, non-disease-based, longitudinal sample collections used in

Address for correspondence: Leena Peltonen, Department of Medical Genetics, University of Helsinki and Department of Molecular Medicine, National Public Health Institute, Finland Biomedicum Helsinki, Haartmaninkatu 8, 00290 Helsinki, Finland. Email: leena.peltonen@ktl.fi

https://doi.org/10.1375/twin.6.5.354 Published online by Cambridge University Press

this proposal circumvent many of these problems and provide the maximum information for meaningful statistical analyses. Twin samples and the population cohorts facilitate the identification of disease alleles, enabling testing of the effect of any identified variant at the population level. They also allow the investigation of the effect of life-course events, since these study cohorts monitor traits and disease development during an individual's lifetime.

Twin data sets have special advantages for genetic studies, unobtainable in regular family or case-control studies, perhaps worth summarizing here. Twin pairs are perfectly correlated for age and many major environmental influences (in particular, intrauterine and childhood conditions). These are important factors in studying common traits, including chronic disorders. Members of twin cohorts have participated actively in these longitudinal studies, mostly by answering questionnaires/interviews, and their continued commitment to participating in twin cohorts reflects their favourable attitude towards research. Twins also provide easy access to collecting both information and biological samples from additional core family members, necessitated by some study designs. It should also be remembered that concordant monozygotic (MZ) twin pairs are ideal for association studies and for determination of genetic and environmental factors affecting disease development. They also offer a unique opportunity to look for genes that influence trait variability (including environment interaction) and possibly also alterations in gene expression (MacGregor et al., 2000).

Twin cohorts represent excellent population cohorts for which ascertainment-bias typically does not pose a problem because sample selection has not occurred with reference to any specific phenotype. It also seems that twin data are particularly valuable to conduct quantitative trait loci analyses (Martin et al., 1997). Population-based twin cohorts also offer the possibility of rational sample selection for initial trait-targeted genotyping efforts since the sibpairs can be selected based on the concordant or discordant strategy and based on the population-and sex-specific means and standard deviations for any measured trait.

Importantly, many additional benefits for genetic studies exist in the European populations from which these twin cohorts are recruited. Population registers and statistics through church records of deaths, births and migration have been collected continuously since the 17th century for many participating countries. A particular advantage for genetic studies is the possibility of tracking pedigree data back several generations through reliable and well-preserved population registers. Further, equal educational systems and national health care programs provide special advantages for recruitment of participants, the quality of the data collected by questionnaires and accessible health care information.

Markers to Define Individual Genome Landscapes

Precise information on the human genome sequence exists in the web and in principle the practical molecular tools to tackle disease-associated alleles and to collect information on the genetic landscape are at hand.[1] Detailed information on sequence variants and initial information on haplotype blocks should enable genome-wide studies of common phenotypes (Hirschhorn et al., 2002). However, several uncertainties need resolving before undertaking large-scale genome-wide studies. Specifically, we have to decide which type of DNA variants will be employed in various stages of the studies (multiallelic STRs vs. bi-allelic SNPs). Should we choose variants based on their map position (linkage disequilibrium-based analysis) versus variants chosen based on their hypothesized biological importance? What is the real character of haplotype blocks and how wide is the diversity in populations? What quality control measures do we need to minimise errors in genotyping? How do we combine genotype information from various types of markers and maps? The GenomEUtwin collaboration aims to also tackle these issues by pooling not only study sample from various European populations, but also the collective expertise of the consortium in molecular analyses and genetic studies.

Large-scale Investments and Goals of GenomEUtwin

The size of this integrated effort can be described in the bureaucratic language of the European Commission as a total of 4166 investigator months listed in the technical annex or € 13.4 million funding provided by EU. The total investment by individual groups and involved Institutes is at least an order of magnitude higher. Decades of data collection, various research grants for individual groups and institutional efforts to build and improve the general research infrastructure are all needed to even launch this project. The bold scientific ambitions of GenomEUtwin reflect these significant investments. GenomEUtwin will develop a framework facilitating the development and use of novel strategies to make best use of the unique features of our population cohorts, including the availability of longitudinal data and extensive information about lifestyle and environmental factors. The project aims to establish unique intellectual and technical infrastructure for research of common human traits and diseases and the training of scientists in new biology. Finally, the project will make every effort to make the epidemiological, phenotypic and genotypic information on different population cohorts accessible to investigators worldwide. If these ambitious goals are reached, the investment will have been productive beyond expectations.

Population Cohorts of GenomEUtwin

The GenomEUtwin project relies on unique European (and Australian) epidemiological resources. Twin cohorts and registries facilitate studies on dizygotic twins and their families, and are amenable to a wide range of statistical strategies going beyond simple association studies. The inclusion of MZ twins in linkage studies immediately enables separate estimation of the contribution of background genes and shared family environment; it also provides a much better estimate of unique individual environment, and places an upper bound on estimates of variance due to a linked QTL. Association studies with MZ twins are more powerful since one obtains two phenotypes for every one genotype, so halving the error variance. Furthermore, longitudinal life-course event data collection and the inclusion of monozygotic twins provide the unique

https://doi.org/10.1375/twin.6.5.354 Published online by Cambridge University Press

possibility of analysing gene–environment interactions, since the MZ intrapair difference is a direct estimate of E, the unique environment. Similar advantages concerning cardiovascular traits are provided by the MORGAM-study, a representative population cohort of multiple European populations. The desire of the investigators involved is that with the study strategy outlined in this proposal we can contribute to the global understanding of the genetic background of common human traits and diseases. Importantly, the information of the interactions between genes and environment in the development of common diseases and their trait components has a potential to leave an impact on future national and international health care and prevention programs.

The Danish, Dutch, Finnish, Italian, Norwegian and Swedish national twin cohorts form an amazing collection of over 0.8 million twins, the largest epidemiological study cohort in the world. Two more recently joined cohorts, St Thomas' twin cohort in Great Britain with 10,000, and the Australian twin cohort with over 30,000 extremely well characterized twin participants are great additions to the GenomEUtwin study populations. DNA samples with informed consents for genetic studies of common diseases have already been stored from over 50,000 twins in these population-based twin cohorts and collection is continuing in all centres. Many of the participating twin cohorts have been proceeding with systematic data collection for decades, and the principal investigators of these cohort have built a unique infrastructure for mounting genetic epidemiological studies. It should be especially emphasized that the DNA collections for the cohorts reflects a natural extension of the long-standing practice in these population-wide genetic epidemiological studies of the European institutions involved.

Similar to the twin studies, the MORGAM population cohort of GenomEUtwin (a continuation of the WHO MONICA[2] collaboration), has a long and successful history of conducting multicentre research (participating countries in MONICA were: Australia, Belgium, Canada, China, Czech Republic, Denmark, Finland, France, Germany, Iceland, Italy, Lithuania, New Zealand, Poland, Russian Federation, Spain, Sweden, Switzerland, United Kingdom, USA, Yugoslavia), which makes it ideal for embarking on genetic studies of complex, multifactorial diseases. MORGAM (Monica, Risk, Genetics, Archiving and Monograph) has its focus on genetic and environmental predictors of chronic diseases, in particular coronary heart disease (CHD) and stroke. The study has a prospective case-cohort design enabling the sampling of cohort controls as the comparison group for various trait phenotypes. Epidemiological and phenotype data collected for over 350,000 participants provide a necessary background for studying the association between disease events, trait components and environmental and life style risk factors. The value of this cohort for the research described herein is in the immediate validation of any gene(s) identified as influencing a trait in our genetic studies of twin cohorts. This holds for any common trait, especially cardiovascular traits and their components, and any trait with a major population impact, such as cancer. Most MORGAM cohorts also have stored frozen sera or plasma. This allows for later determination of additional phenotypes deemed important on the basis of genetic findings or new hypotheses.

The Danish Twin Registry. The Danish Twin Registry[3] was established in 1954. It covers birth cohorts from 1870 through 1996 with a total of 66,412 identified twin pairs, of whom 27,900 are intact living pairs. The basic information on all of these pairs includes date of birth and national identification number, name, address, vital status, sex and zygosity. Data are gathered by means of questionnaires, clinical investigations, interviews and record linkage with other registries. The registry contains questionnaire and interview data on the most common diseases as well as common phenotypes. Furthermore, data on all discharge diagnosis since 1977 are obtained. The possibilities for registry linkage are excellent within the Danish Act on Processing of Personal Data, as long as the linkage is done in search of health information. DNA samples exist for over 10,000 twins.

The Netherlands Twin Registry. The Netherlands Twin Registry (NTR)[4] was established in 1987 and consists of over 20,000 young twin pairs (and higher-order multiples) aged between 0 and 13 years, and over 6000 twin families with adolescent and (young) adult twin pairs. In addition to the adult twins themselves, the NTR registers the siblings, parents and spouses of twins. Recruitment of new-born twins is ongoing with between 50 and 60% of all twins born in The Netherlands participating. DNA samples exist for over 5000 twins and their parents.

The Finnish Twin Cohort. The Finnish Twin Cohort[5] was established in 1974. Twins and their families have been ascertained in three stages from the Central Population Register in 1974 (older like-sexed pairs), in 1987 (multiple births 1968-1987) and in 1995 (opposite-sex pairs 1938-1957), with a total of 35,850 pairs with both members currently alive. The older part consists of same sex twin pairs born before 1958 with both twins alive in 1975, and three surveys have been carried out in 1975, 1981 and 1990 with the response rates varying between 77 and 89%. In addition to mortality follow-up, morbidity follow-up of the twins has been performed through nationwide computerized medical registries: incident malignancies through the Cancer registry, hospitalisations through the Hospital Discharge Register, and information of selected diseases based on the national register for fully-reimbursable medications. Numerous sub-sample studies have been performed with DNA collection being a routine feature since the mid-1990s. For the younger cohort, two 5-year birth cohorts born 1975–1979 and 1983–1987 have been studied in an intensive longitudinal manner. Together they compromise about 5500 families with twins (twins, parents and sibs), who have participated in at least three waves of questionnaires and interviews with more intense subset studies. DNA samples exist for 10,000 twins and their parents.

The Italian Twin Registry. The Italian Twin Registry[6] was established in 1997 when all potential twins (approximately 1,200,000 persons with the same last name, same place and date of birth, alive at the end of the previous year) were identified. The Italian Ministry of Health has recently

https://doi.org/10.1375/twin.6.5.354 Published online by Cambridge University Press

provided funds for a comprehensive national research program on twin studies. As a first step, all twin pairs belonging to three age ranges: 5–14, 35–44, 65–74 years are currently being contacted by mail. By the end of the year 2002, 120,000 twin pairs are expected to be enrolled: 36,000 among children, 54,000 among adults, 30,000 among the elderly. General background data will be soon available on zygosity, education, occupation, weight and height family size and family history of selected diseases. A system for linking mortality data, cancer registers (25% of the Italian population), dementia hospital discharges will soon be implemented. DNA samples already exist for hundreds of twins and the collection is ongoing.

The Norwegian Twin Panel. The National Institute of Public Health in Oslo[7] maintains an active program of research in genetic epidemiology based on the population-based twin panel. This is a cohort sequential design, the current database includes information on 15,370 twins born from 1967 through 1979, and new cohorts will be added. The first wave of data collection was conducted in 1992 and recruited all twin pairs born between 1967 and 1974. The information collected includes zygosity items, demographic data, measures of physical and mental health, and lifestyle factors. The response rate was 75% and included 2570 twin pairs. A second, greatly extended questionnaire was sent in 1998 to the first sample of twins from 1992 and to 5 new birth cohorts (born 1975–1979). The response rate was 69% and includes 3334 pairs. The data are already linked to medical information from the National Birth Registry, and linkage to information collected as part of the National Health Examination should also be possible soon. DNA exists for 4000 twins, and several national and international sub-projects, some which collect clinical measures, are completed or underway.

The Swedish Twin Registry. The Swedish Twin Registry (STR)[8] is comprised of three cohorts, each of which differs in its method of ascertainment and extent of data collected. The cohort of twins born from 1886–1925 consists of 10,503 same-sex twin pairs who were alive in 1961. Three health surveys have been performed in 1963, 1967 and 1970. Recently, information from about 11,500 opposite-sex twins has been computerized. In 1970, a new cohort of twins born between 1926 and 1967 was compiled. This time it was compiled by use of nationalized birth registrations. A birth register consisting of all 50,000 twin births was established. In 1972, all like-sexed twins born between 1926 and 1958 and living in Sweden responded to a questionnaire similar to that given to the older cohort. Responses were received from 36,536 individuals in 13,863 twin pairs. There is currently a major data collection effort that entails telephone interview screening of all living twins born from 1886 to1958 (N = 52,080). This effort, in which 32,000 individuals have already been contacted, was completed in June 2002. Twins are screened for several diseases, symptoms, medical and environmental exposures. Twins born from 1959–1968, identified in the birth register in 1972, and twins born from 1969–1990, identified by record linkage to the National Birth Register, are also included in the registry (N = 29,960). From this cohort, only the parents of twin pairs born between 1985 and 1986 have been contacted. The STR is regularly updated with current addresses and marital status, as well as linkages to the cause of death, cancer, inpatient discharge and medical birth registries. DNA exists for numerous sub-cohorts of almost 5,000 twins.

The St Thomas' UK Adult Twin Registry (TwinsUK). The St Thomas UK twin registry[9] was started in 1992 as an adult volunteer database recruited by media campaigns without mentioning specific diseases. The twin database includes over 9000 twins with an average age of around 45, and is predominantly female and same sex, since the diseases the cohort initially focused on were more common in women. The ratio of identical to non-identical twins is approximately 50:50, and the volunteers came from throughout the UK and Ireland following multiple media campaigns for twins to help with research. One thousand five hundred DZ and 500 MZ pairs have been studied in considerable detail with hundreds of clinical and biochemical phenotypes and genotyped for a genome-wide marker panel with up to 737 markers on each of the twins. DNA and blood is available on over 5000 twins.

The Australian Twin Cohort. The only non-European group we will collaborate with is the Australian twin cohort. The Australian Twin Register[10] was founded by Nick Martin and John Mathews in 1978 and by 1983, 15,000 pairs had been enrolled. It now stands at about 30,000 pairs. Of the ~6000 twins in the older adult cohort, blood samples were obtained from 3354 during 1993–95 in a major collection drive funded by NIH. In the next 5 years, around 3000 twins and 7000 of their siblings drawn from both adult cohorts will be interviewed and bled. The vast majority of participants are of North European origin. The cohort has detailed phenotype data for nearly all the target traits of GenomEUtwin.

The MORGAM Population Cohorts. A vast array of phenotypic data has been collected from MONICA participants.[11] For example, the CHD studies in the MORGAM cohorts include approximately 6000 individuals, drawn from population-based cohorts consisting of more than 140,000 participants who have donated DNA samples in the first half of 1990s. The average follow-up time is approximately 6 years. It should be noted that many of the MORGAM cohorts are drawn from the same countries as the twin cohorts listed above, complementing the synergy between the twin approach and the population approach.

Specific Research Objectives of GenomEUtwin

1. To develop novel strategies to maximally utilize the unique features of twin cohorts, including the availability of longitudinal data and ample information about lifestyle and environmental factors, in the characterization of complex traits.
2. To utilize the synergy between the twin cohorts and the representative population cohorts (MORGAM) from the same countries, in studies of genetic and environmental predictors of traits.
3. To develop (in collaboration with the biotechnology industry) new molecular methods for high-throughput

https://doi.org/10.1375/twin.6.5.354 Published online by Cambridge University Press

genotyping by analysing sub-samples of twin cohorts selected for specific traits (stature, BMI, migraine, coronary heart disease, stroke and longevity), requiring genotyping of hundreds of multiallelic and SNP markers in thousands of subjects.

4. To develop novel mathematical strategies to combine information on genetic profiles underlying common traits (multiple genes and their various alleles with different impacts on disease outcome) and to estimate the role of genetic and environmental factors in the disease process in selected traits using dizygotic (DZ) and MZ twins and population cohorts.
5. To develop statistical methods for longitudinal data analyses of these life-span data sets to identify genes and/or environmental risk factors expressing themselves only at certain ages.
6. To compare major genetic factors predisposing to migraine, coronary heart disease and stroke in different European populations.

Building the Research Infastructure

GenomEUtwin research operations are carried out by intellectual core facilities, established to integrate the expertise in epidemiology, database structure, molecular techniques and statistical analyses. The cores operate as facilitators for both the research carried out in individual participating centres and for the massive research enterprise of the integrated project. They also provide a top-level environment for the training of the European researchers, clinicians and students in all components of genetic epidemiology.

The Epidemiology and Phenotype Core. (Dr. Kaare Christensen) provides the guidelines for ideal study samples for various analyses, as well as agrees upon harmonized criteria for traits and quality control of diagnostic instruments. The outcome of some of the work produced so far using the expertise of this core is provided in six articles of this issue describing the primary analysis of population prevalence and heritability of target traits of GenomEUtwin: height (Silventoinen et al.). BMI (Schousboe et al.). coronary heart disease (Evans et al.); stroke (Gaist et al.); longevity (Skytthe et al.) and migraine (Mulder et al.).

The Database Core. (Drs Jan-Eric Litton and Nancy Pedersen) aims to harmonise and, to some extent, standardize the epidemiological, phenotypic and genotype databases in participating centres to facilitate extensive pooled analyses of genetic, phenotype and epidemiological data. The basis and initial strategy of this effort, so crucial for the success of GenomEUtwin is described in the paper by Litton et al.

The DNA Extraction and Genotyping Core. (Drs Ann-Christine Syvänen, Aarno Palotie and Markus Perola) provides centralized DNA extraction and genotyping for collected DNA samples as well as quality control steps for data produced and databases formed. Descriptions of molecular analyses of DNA variants are provided in the paper of Silander et al.

The Statistical Analysis Core. (Dr Hans van Houwelingen) creates the intellectual and computational infrastructure for statistical analyses needed by the project. Examples of this effort are provided by the paper by Putter et al., and the general overview of theory and practice in quantitative genetics is given in the paper by Posthuma et al.

The Ethics and Education Core. (Drs Jennifer Harris and Jaakko Kaprio) works on ethical guidelines for the study, including issues like informed consent and quality control of data security systems for the databases, to provide the highest possible level of data confidentiality. A range of cross-country ethical aspects are presented in the paper by Harris et al. This core operates in close collaborations with involved academic institutions to arrange the education of young scientists in different elements of modern genetic epidemiology.

How Much Genotyping Should we do and How Much can we Afford?

Selection of variants for genotyping will be critical for GenomEUtwin. We plan to identify new loci for studied traits by expanding twin families to core pedigrees and larger family units and perform genotyping using a set of 400 multiallelic markers to provide a skeletal map for our twin-cohort based studies. The multiallelic markers can be selected with reasonable confidence to provide a good genome-wide coverage with even distribution and known genetic intervals between the markers. Selection of the relevant set of SNPs is a more challenging task. We are initially genotyping sets of densely and evenly spaced SNPs in the pilot study of the biologically relevant genes or linked genomic regions. In combination with the multiallelic genotypes this data set will facilitate analysis for linkage disequilibrium (LD) and chromosomal haplotypes for different genomic regions — as well as their comparison in different study populations. Recent data on the extent of LD and SNP haplotypes in the human genome will soon help to shape our thinking regarding the selection of optimal SNPs for genotyping studies (Daly et al., 2001; Jeffreys et al., 2001; Johnson et al., 2001; Rioux et al., 2001). Our analyses, targeted to different populations included in our twin cohorts will provide data for optimal SNP selection strategies for different populations, and will guide the scaling-up for genome-wide investigations using SNPs. GenomEUtwin will be well-prepared for the time when such efforts will also be economically feasible.

Genetic or Environmental Influences?

The sharing of a common environment by the twins in conjunction with the enormous study sample size provides a unique opportunity to dissect the genetic and environmental risk factors in European populations. The development of statistical methods will surely be one of its most important contributions to basic science. Analytical approaches are available to probe pleiotropy and polygeny using multivariate quantitative genetic twin models. No other projects have the advantage of combining quantitative and molecular levels of analysis allowing both reductionist and integrationist approaches to the analysis of complex traits. We can determine if there are common sets of genes or environments explaining the relationships between the outcomes and the intermediate phenotypes. Known QTL:s or genetic polymorphisms can be incorporated into the

https://doi.org/10.1375/twin.6.5.354 Published online by Cambridge University Press

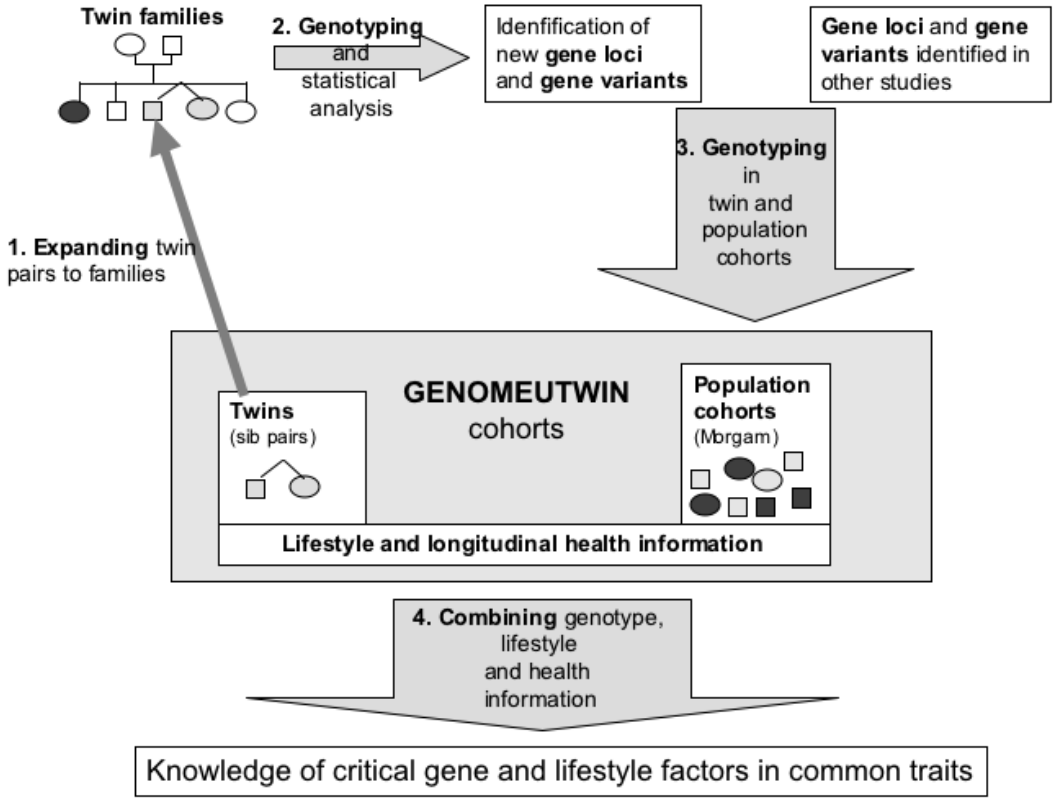

Figure 1

A schematic presentation of the current strategy of GenomEUtwin.

quantitative models. These types of analyses help determine the genetic and environmental architecture of the trait and complement the procedures (Martin et al., 1997; MacGregor et al., 2000).

Different statistical approaches will be used and novel ones developed within the consortium. With common diseases, we can expect to encounter genetic heterogeneity, pleiotropism, epistasis, genotype-by-environment interaction, and correlated environments. Classical penetrance-based linkage methods are simply not up to the task of analysing common diseases. The trend toward less model dependent methods of statistical analysis will only accelerate. A great deal of work remains to be done, both in theoretical development and translation of that theory into practical software.

What are our Chances of Success?

The strategy of GenomEUtwin is presented schematically in Figure 1. Will the results of this effort pave the way towards improved understanding of human health and disease? We believe they will. The success of studies like GenomEUtwin depend on the molecular basis of human disease or its absence, something we have very limited knowledge of so far. Let us consider a trivial parameter to exemplify our ignorance: the frequency of DNA variants that underlie specific phenotypic variation. High frequency variants are likely to be old and present in a wide range of populations, while low frequency variants may be relatively new and specific to one or a few populations. It is likely that some high-frequency (old) SNPs contribute to the disease related phenotypes that we are studying (coronary heart disease, stroke, migraine). Within a few years, most such variants will be available from public databases and thus will be available 'off the shelf' for association studies of disease related phenotypes.

The major current value of the phenotypically well-characterized population cohorts of GenomEUtwin lies in the instant ability to address the real significance of any reported DNA variant. Such variants can be tested for any of the traits on which data are collected in one or several of the cohorts. However, it is also most likely that much of the phenotypic variation that we seek to understand derives from unidentified sequence variants. Such variants could either be specific to certain populations or could be infrequent in normal populations, resembling mutations for Mendelian traits (Glatt et al., 2001). We therefore anticipate that achieving the goals of the collaboration will still require identification of new variants associated with the studied traits. Different variants will probably be identified in different populations and their real impact will have to be tested against the background of the environmental and life style factors, equally local and population specific as rare DNA variants.

https://doi.org/10.1375/twin.6.5.354 Published online by Cambridge University Press

The foundation of GenomEUutwin lies in well-established and professionally collected epidemiological study samples, on which vast amounts of data relevant for human health already exist. Some elegant landmark studies of the role of genetics in human traits have already been published using the data collected from these cohorts (for a review, see Boomsma et al., 2002). In GenomEUtwin, modern genome-wide tools and an immense amount of genetic information can be adapted to extend and deepen these studies. Traditionally, twin research has for decades been a driving force in human genetics and it has served as inspiration for the strategy designed for GenomEUtwin. The task of this international project is to use modern molecular methods to carry the torch further, increasing the accuracy of our knowledge of the fascinating interplay between genetics, environment and life events in human biology and disease.

Endnotes

1 Available from http://www.ncbi.nlm.nih.gov/genome/guide/human/

2 Available from www.ktl.fi/monica

3 Available from http://www.sdu.dk/med/iph/EPID/TwinReg/html/index-gb.htm

4 Available from http://www.psy.vu.nl/ntr/

5 Available from http://kate.pc.helsinki.fi/twin/twinhome.html

6 Available from http://www.gemelli.iss.it

7 Available from http://www.folkehelsa.no/

8 Available from http://www.mep.ki.se/twin/index_en.html

9 Available from http://www.twin-research.ac.uk

10 Available from http://www.twins.org.au

11 Available from http://www.ktl.fi/monica

References

Amarger, V., Nguyen, M., Van Laere, A-S., Braynschweig, M., Nezer, C., Georges, M., & Andersson, L. (2002). Comparative sequence analysis of the INS-IGF2-H19 gene cluster in pigs. *Mammalian Genome, 13*, 388–398.

Boomsma, D., Busjahn, A., & Peltonen, L. (2002). Classical twin studies and beyond. *Nature Genetics Reviews, 3*, 872–882.

Collins, F. S., Green, E. D., Guttmache, A. E., & Guyer, M. S. (2003). A vision for the future of genomics research. *Nature, 422,* 835–847.

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., & Lander, E. S. (2001). High-resolution haplotype structure in the human genome. *Nature Genetics, 29,* 229–232.

Frary, A., Nesbitt, T. C., Grandillo, S., Knaap, E, Cong, B., Liu, J., et al. (2000). A quantitative trait locus key to the evolution of tomato fruit size. *Science, 289*, 85–88.

Glatt, C. E., DeYoung, J. A., Delgado, S., Service, S. K., Giacomini, K. M., Edwards, R. H., et al. (2001). Screening a large reference sample to identify very low frequency sequence variants: Comparisons between two genes. *Nature Genetics, 27*, 435–438.

Hirschhorn, J. N., Lohmueller, K., Byrne, E., & Hirschhorn, K. (2002). A comprehensive review of genetic association studies. *Genetics in Medicine, 4*, 45–61.

Horikawa, Y., Oda, N., Cox, N. J., Li, X., Orho-Melander, M., Hara, M., et al. (2000). Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nature Genetics, 26,* 163–175.

Hugot, J. P., Chamaillard, M., Zouali, H., Lesage, S., Cezard, J. P., Belaiche, J., et al. (2001). Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature, 411*, 599–603.

Jeffreys, A. J., Kauppi, L., & Neumann, R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genetics, 29*, 217–222.

Johnson, G. C., Esposito, L., Barratt, B. J., Smith, A. N., Heward, J., Di Genova, G., Ueda, H., Cordell, H. J., et al. (2001). Haplotype tagging for the identification of common disease genes. *Nature Genetics, 29*, 233–237.

MacGregor, A. J., Snieder, H., Schork, N. J., & Spector, T. D. (2000) Twins. Novel uses to study complex traits and genetic diseases. *Trends in Genetics, 16*, 131–134.

Martin, N. G., Boomsma, D. I., & Machin, G. (1997). A twin-pronged attack on complex traits. *Nature Genetics, 17*, 387–392.

Rioux, J. D., Daly, M. J., Silverberg, M. S., Lindblad, K., Steinhart, H., Cohen, Z., et al. (2001). Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohns disease. *Nature Genetics, 29*, 223–228.

https://doi.org/10.1375/twin.6.5.354 Published online by Cambridge University Press