European Journal of Human Genetics (2009) 17, 533 – 536
© 2009 Macmillan Publishers Limited All rights reserved 1018-4813/09 $32.00
www.nature.com/ejhg

SHORT REPORT

GENESTAT: an information portal for design and analysis of genetic association studies

Samuli Ripatti*[,1,2], Tim Becker[1,3], Heike Bickeböller[4], Annica Dominicus[5], Christine Fischer[6], Keith Humphreys[1], Gudrun Jonasdottir[1], Yves Moreau[7], Marita Olsson[1,8], Alexander Ploner[1], Nuala Sheehan[9,10], Kristel Van Steen[11,12,13], Max Baur[3], Cornelia van Duijn[14] and Juni Palmgren[1,15]

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; [2]FIMM Institute for Molecular Medicine and Department of Molecular Medicine and National Public Health Institute, Helsinki, Finland; [3]Institute for Medical Biometry, Informatics and Epidemiology, University of Bonn, Bonn, Germany; [4]Department of Genetic Epidemiology, Medical School, University of Göttingen, Göttingen, Germany; [5]Department of Biostatistics, AstraZeneca R&D, Södertälje, Sweden; [6]Institute for Human Genetics, University of Heidelberg, Heidelberg, Germany; [7]Department of Electrical Engineering ESAT – SDC, KU Leuven, Leuven, Belgium; [8]Department of Mathematical Sciences, Chalmers University of Technology, Gothenburg, Sweden; [9]Department of Health Sciences, University of Leicester, Leicester, UK; [10]Department of Genetics, University of Leicester, Leicester, UK; [11]Department of Applied Mathematics and Computer Science, Ghent University, Ghent, Belgium; [12]StepGen cvba, Merelbeke, Belgium; [13]Department of electrical engineering and computer science – Montefiore Institute, University of Liège, Liège, Belgium; [14]Department of Epidemiology and Biostatistics, Erasmus University Medical Center, Rotterdam, The Netherlands; [15]Department of Mathematical Statistics, Stockholm University, Stockholm, Sweden

We present the rationale, the background and the structure for version 2.0 of the GENESTAT information portal (www.genestat.org) for statistical genetics. The fast methodological advances, coupled with a range of standalone software, makes it difficult for expert as well as non-expert users to orientate when designing and analysing their genetic studies. The ultimate ambition of GENESTAT is to guide on statistical methodology related to the broad spectrum of research in genetic epidemiology. GENESTAT 2.0 focuses on genetic association studies. Each entry provides a summary of a topic and gives links to key papers, websites and software. The flexibility of the internet is utilised for cross-referencing and for open editing. This paper gives an overview of GENESTAT and gives short introductions to the current main topics in GENESTAT, with additional entries on the website. Methods and software developers are invited to contribute to the portal, which is powered by a Wikipedia-type engine and allows easy additions and editing.

European Journal of Human Genetics (2009) **17,** 533–536; doi:10.1038/ejhg.2008.216; published online 12 November 2008

Keywords: statistical genetics; genetic software; internet

Introduction

The twenty-first century biomedical and health sciences have seen a movement from study of rare monogenic disorders to common, multifactorial diseases,[1] with high-throughput technologies that enhance dissection of the underlying complex aetiological processes.

Elucidation of multifactorial disease aetiology is challenging because these diseases are, as a rule, causally

*Correspondence: Dr S Ripatti, FIMM, P.O. Box 20, 00014 Helsinki, Finland. Tel: +358 9 4744 8159; Fax: +358 9 47448480;
E-mail: Samuli.ripatti@ktl.fi

Received 25 September 2007; revised 2 October 2008; accepted 15 October 2008; published online 12 November 2008

heterogeneous, and they are driven by a large number of small, additive or synergistic effects, representing consequences of genetic predisposition, lifestyle and the environment. Success in revealing these complex interactions will depend critically on the availability of large-scale documented, up-to-date epidemiological, clinical, biological and molecular sources of information.[2–8]

Infrastructures for large-scale population-based research are a focus in the EU sixth (FP6) and seventh (FP7) framework programmes and the European Strategy Forum of Research Infrastructures (http://cordis.europa.eu/esfri/) preparatory phase. As European countries have strong national health-care systems and they have accumulated epidemiological data throughout decades, often complemented with biological samples, it is important that the process of harmonising data and methods takes place within the region and is coordinated with similar efforts in other parts of the world.

Harmonisation in the use of human samples and subject-specific information from large population groups must cover guidelines for study design and statistical analysis. Besides theoretical advances in statistical inference, computational statistics and algorithm development *per se*, we see a flood of methodological development in statistical genetics and genetic epidemiology in leading journals. Pieces of software, often in the form of standalone programs, are found on individual www pages and/or collected in archives such as http://linkage.rockefeller.edu/soft/list.html. Although good search engines are generally available in the internet, the practising clinicians and geneticists, as well as non-expert statisticians and bioinformaticians, find it difficult to absorb and make use of the rapid methodological development. This in turn results in a suboptimal use of resources.

In 2002, the national biobank programme in Sweden (www.wcn.se) initiated an internet-based statistical genetics information portal, GENESTAT, to provide tutorials, reviews and a discussion forum, related to genetic association studies, with links to key websites and computer programs for the analysis of genetic data.

During 2007, a European GENESTAT working group was assembled with the aim to oversee content, provide an editorial function and make the portal known. This paper by the working group is one such step. The GENESTAT portal has been given its own domain (www.genestat.org), a number of new entries have been added and, most importantly, the portal is set up in an interactive mode, in the hope of attracting interest and attention not only from users but also from experts and methods developers.

This paper introduces the GENESTAT information portal version 2.0 (www.genestat.org), which is the first version put forward to the broad international scientific community. The ultimate aim of GENESTAT is to address design and statistical analysis issues in genetic epidemiology. As a first step, this paper and version 2 of GENESTAT focus on genetic association studies, both candidate gene studies and genome-wide association studies. The engine is a Wikipedia style of information exchange to allow the best mix of sustainability and up-to-date quality. Initially, GENESTAT supports free entries and editing, with only weak supervision. Stricter supervision may be considered later. Each GENESTAT entry is structured to provide a broad overview, with references to key papers and software. In the next section, we give examples of GENESTAT entries and end the paper with a discussion of possible extensions, and with an invitation to methods and software developers to add information to the portal and to edit existing entries.

Current GENESTAT entries

GENESTAT is targeted towards researchers conducting and analysing data from genetic association studies. The portal includes two sections, ‘Genetic Association Studies’, with subsections ‘Planning’, ‘Quality Control’, ‘Population Stratification’, ‘Testing and Estimating Association’ and ‘Statistical Modelling’, which is divided into ‘Modelling Genotypic Information’, ‘Pathways’, ‘Replication’, ‘Meta-analysis’ and ‘Mendelian Randomisation’. Navigation in the portal is simple and the Wikipedia-like structure allows for augmentation of the available information using simple editing tools.

To convey the flavour of the portal, we give short introduction to the current main sections of GENESTAT. The actual entries are found on www.genestat.org.

Before genotyping

This section in GENESTAT discusses questions in genetic study design. These include elaboration of the underlying biological mechanism and about the structure of the study population, choice of markers and phenotype and family- *vs* population-based independent individual designs.
http://www.genestat.org/index.php?n = GeneStat.PlanningStage

Genotype data quality control

This section gives guidance on procedures for gender checks and relatedness checks, quality control based on call rates and Hardy–Weinberg Equilibrium and discusses combining of data across different studies and platforms.
http://www.genestat.org/index.php?n = GeneStat.GenotypingQualityControl

Population stratification

A thorough section on population stratification discusses genetic confounding caused by the underlying population structure and potentially leading to both false-positive and false-negative results in genetic association studies. This section also presents the current methods for solving the problem in both candidate gene studies and in genome-wide association studies.

http://www.genestat.org/index.php?n = GeneStat.PopulationStratification

Testing and estimating association

The largest section in GENESTAT describes association testing and estimation under different study designs and different kinds of phenotypes. Testing for single-marker associations as well as for haplotypes, interactions and model selection procedures are discussed in this section. In addition, more advanced topics such as controlling for multiple testing and modelling associations in copy number variation along with power comparisons between different tests are presented in this section.
http://www.genestat.org/index.php?n = GeneStat.TestingAndEstimatingAssociation

Modelling genotypic information

This section discusses more advanced topics on structuring genotypic information beyond single-marker analyses. In particular, methods for haplotype estimation, identification of haplotype blocks, measures of linkage disequilibrium and methods for capturing most of the genetic variation in a gene through tag SNPs are discussed.
http://www.genestat.org/index.php?n = GeneStat.MeasuringLinkageDisequilibriumAndHaplotypeEstimation

Analysis of pathways

The pathway section discusses methods for incorporating biological *a priori* knowledge to the association testing. This can be done, for example, by jointly testing the effects of markers selected from the same biochemical pathway, or by combining information of intermediate and end phenotypes for association testing.
http://www.genestat.org/index.php?n = GeneStat.Pathways

Replication and meta-analysis

Sections about replication and meta-analysis discuss strategies for scientifically meaningful replication of a *de novo* gene association finding and for combining data and statistical inference across association studies. It also discusses the origin and impact of between-study heterogeneity in association studies.
http://www.genestat.org/index.php?n = GeneStat.Meta-analysis

Mendelian randomisation: inferring causality in observational epidemiology

This section of GENESTAT discusses Mendelian randomisation; a special design for using genetic markers for inferring causality between modifiable risk factors and disease. Inferring causality from observational data is difficult as it is not always clear which of the two associated variables is the cause, which the effect, or whether both are common effects of a third unobserved variable or confounder. Mendelian randomisation is a method that allows to test for, or in certain cases to estimate, a causal effect between modifiable risk factor and disease from observational data in the presence of confounding factors by using common genetic polymorphisms with well-understood effects on exposure patterns.
http://www.genestat.org/index.php?n = GeneStat.MendelianRandomisation

Discussion

The usefulness of GENESTAT will be proven over time. In its current state, groups applying association methods in their daily work benefit most from GENESTAT. A partial aim of GENESTAT is also to improve the quality of statistical analyses of complex disease, and this would be beneficial for the scientific community as a whole.

There are several directions towards which the current GENESTAT information portal could be extended. Differential measurement errors in SNPs and measured lifestyle factors are worth exploring. Harmonisation of SNP measurements from different platforms calls for imputation techniques using the available HapMap data. Novel designs are needed for studying genes and the environment jointly, and with proper meta-analytic methods, the heterogeneity in the phenotype definitions and measurements and strengths of association may be addressed. An increased interest in the design and analysis of population-based studies involving epigenome, transcriptome or proteome data is also expected. The current open content management system, with a Wikipedia type of 'edit this page' link on every page, is trivially open for these extensions, in principle, but relies heavily on the commitment of the scientific community with expertise in these areas.

We emphasise that GENESTAT does not cover all the possible statistical methods related to genetic association studies and has no ambition to be complete at any point in time, but rather to develop and evolve over time. The aim today is to provide an interesting embryo for further development that can adapt to a variety of needs from scientists who use human samples and subject-specific information from large population groups. We welcome the broad genetic research community to visit the portal, and we specifically invite the community of statistical genetics methods developers to contribute to its content. The ultimate aim is to create a growing and constantly updated information repository for statistical genetics. GENESTAT success will be manifest by the number of visits to the portal and by the new contributions.

Acknowledgements

This study was supported by FP6 coordinated action PHOEBE (Promoting Harmonisation of Epidemiological Biobanks in Europe). Version 1.0 of GENESTAT was built with support from The Wallenberg Consortium North, Sweden. HB was supported by BMBF – German National Genome Research Network NGFN.

We thank Staffan Nilsson and Jaana Wessman for discussions in the planning stage.

References

1 Collins FS, Green ED, Guttmacher AE, Guyer MS: A vision for the future of genomics research. *Nature* 2003; **422**: 835–847.

2 Herbert A, Gerry NP, McQueen MB *et al*: A common genetic variant is associated with adult and childhood obesity. *Science* 2006; **312**: 279–283.

3 Frayling TM, Timpson NJ, Weedon MN *et al*: A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 2007; **316**: 889–894.

4 Helgadottir A, Thorleifsson G, Manolescu A *et al*: A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science* 2007; **316**: 1491–1493.

5 McPherson R, Pertsemlidis A, Kavaslar N *et al*: A common allele on chromosome 9 associated with coronary heart disease. *Science* 2007; **316**: 1488–1491.

6 Diabetes Genetics Initiative: Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 2007; **316**: 1331–1336.

7 Scott LJ, Mohlke KL, Bonnycastle LL *et al*: A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 2007; **316**: 1341–1345.

8 The Wellcome Trust Case Control Consortium: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007; **447**: 661–683.