Assurance

Community Genetics

Community Genet 2006;9:55–61
DOI: 10.1159/000090694

PopGen: Population-Based Recruitment of Patients and Controls for the Analysis of Complex Genotype-Phenotype Relationships

Michael Krawczak[a] Susanna Nikolaus[b] Huberta von Eberstein[b]
Peter J.P. Croucher[a] Nour Eddine El Mokhtari[c] Stefan Schreiber[d]

[a]Institute of Medical Informatics and Statistics, Departments of [b]General Internal Medicine and [c]Cardiology, [d]Institute for Clinical Molecular Biology, Christian Albrechts University, Kiel, Germany

Key Words
Biobank · Complex diseases · Genetic epidemiology · Genotype-phenotype relationship · Population-based sampling · Relative risk

Abstract
Objective: Patient samples used for mapping complex human disease genes are unlikely to be representative of the phenotype spectrum of the respective population as a whole. On the other hand, most ongoing prospective studies are probably too small for evaluating polygenic disease markers. ***Design:*** Precise estimates of population-specific genotypic risks can be obtained efficiently through the complete ascertainment of patients in a geographically confined area. The PopGen project uses the most northern part of Germany as a target region for such a pursuit. ***Results:*** PopGen currently pursues recruitment, sampling and processing activities in close collaboration with a multitude of clinical partners, covering cardiovascular, neuropsychiatric and environmental diseases. ***Conclusion:*** PopGen has successfully established itself as a large-scale genetic epidemiological project of international recognition.

Introduction

For most complex human diseases, the recruitment of samples for gene finding studies has so far been retrospective in nature and limited to cases of a particularly pronounced phenotype, or with a strong family history. This strategy of 'extreme sampling' reflects the expectation that genetic variants involved in the etiology of severe or strongly heritable conditions represent so-called 'major genes'. Major genes are rare, but their individual effects are likely to be strong. This is why extreme retrospective sampling is generally thought to maximize the chances for the initial detection of a disease-associated genetic variant [1]. However, the clinical impact of a disease gene discovery is usually only made through the implementation of diagnostic and therapeutic algorithms in an unselected (i.e. 'normal') population. For such purposes, covariate-adjusted absolute and relative genotypic disease risks are the key parameters of interest [2, 3], and these figures can only be estimated from representative population samples.

For the reasons noted above, representative samples of patients are not normally gathered in the context of genetic epidemiological studies but rather have to be established anew. Furthermore, in order to be able to eval-

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

1422–2795/06/0091–0055$23.50/0

Accessible online at:
www.karger.com/cmg

Prof. Dr. Michael Krawczak
Institute of Medical Informatics and Statistics
Christian Albrechts University, Brunswiker Strasse 10
DE–24105 Kiel (Germany)
Tel. +49 431 597 3200, Fax +49 431 597 3193, E-Mail krawczak@medinfo.uni-kiel.de

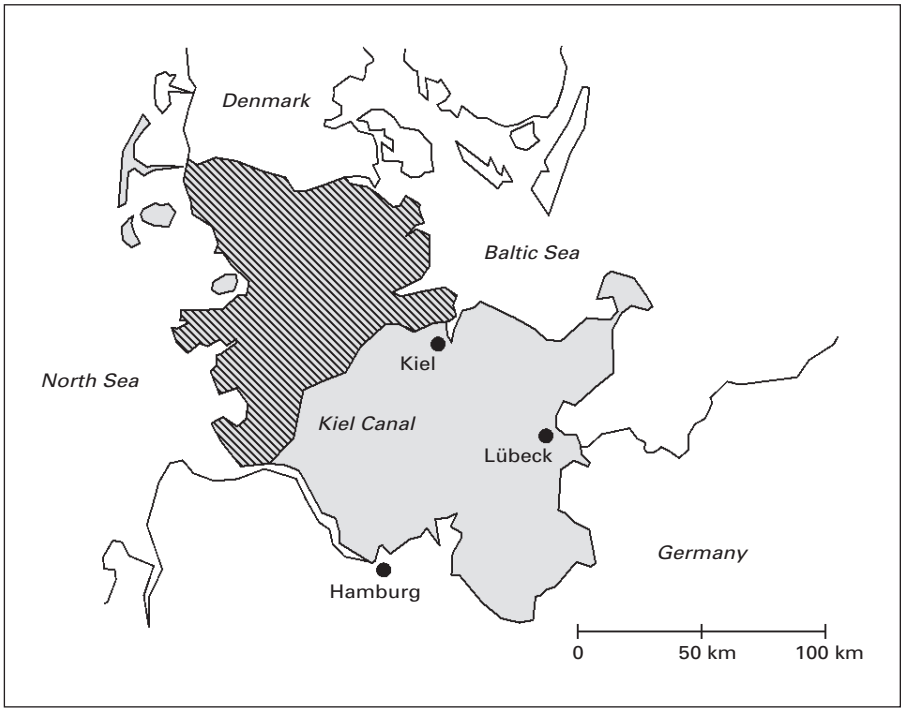

Fig. 1. Map of Schleswig-Holstein (in grey), the most northern federal state of Germany. The PopGen target region (shaded) is located north of the Kiel Canal, excluding however the North-Friesian islands.

uate the relative risk of a given genetic variant retrospectively from case-control data, its background frequency in the same population must also be known. Therefore, large samples of unselected controls from the populations of interest need to be recruited in parallel. Finally, in order to be able to examine the impact of genetic factors on both acute and chronic disease, it appears worthwhile to extend the common cross-sectional or retrospective ascertainment of phenotypes to the prospective follow-up of at least a subset of cases, defined for example by an incidence cohort.

Unbiased sampling of patients is best achieved at the population level by defining a confined geographical catchment area in which all clinically overt cases with the disease in question are recruited. A typical example of the successful, albeit prospective, pursuit of such a complete ascertainment strategy is provided by the Framingham Study [4], a survey in which the whole population of a small town in New England (USA) was analyzed for the presence of frequent conditions such as cardiovascular disease and stroke. Similar projects are underway in Iceland, Estonia and, on a particularly large scale, through 'Biobank UK' [5]. Population-representative samples require that no systematic or uncontrolled 'escape' of patients occurs from the recruitment area. In this respect, the area of Northern Schleswig-Holstein in Germany (fig. 1) provides an almost paradigmatic test case since its tight geographical borders (Denmark, North Sea and Baltic Sea) and low density of treatment facilities set stringent limits for patients seeking treatment either inside or outside its confines. For these reasons, Northern Schleswig-Holstein has been chosen as the target region of 'PopGen', a collaborative effort of the disease-driven networks of the German National Genome Research Network (NGFN; www.ngfn.de,) that seeks to define phenotypes of interest and to recruit patients from a particular geographical area. Since its foundation in May 2003, PopGen has successfully established itself as a large-scale genetic epidemiological project of international recognition.

Design

Most population-based health surveys currently underway are of cohort design, i.e. they investigate the course of diseases prospectively [5–8]. Even for frequent conditions such as coronary heart disease (CHD), however, prospective studies have to be comparatively large (and therefore expensive) in order to provide an adequate number of clinical events during follow-up. Furthermore, whilst it may be possible to assess the impact of frequent

risk factors (e.g. serum cholesterol) in this manner, incidence rates may still be too low in most instances to be able to evaluate highly specific, polygenic disease markers. In contrast, since the accuracy of proband recall is less relevant for genetic than for classical epidemiological risk factors, precise estimates of population-specific genotypic risks can be obtained more efficiently from case-control studies that involve the complete cross-sectional ascertainment of patients in a geographically confined region.

The PopGen Catchment Area

Northern Schleswig-Holstein, an area that is home to approximately 1.1 million people, is enclosed by the Danish border (North), the North Sea and Elbe River (West), the Baltic Sea (East), and the Kiel Canal (South). The latter can only be crossed by a limited number of bridges and ferries. Denmark has no tertiary or other specialized referral centers near the border that would be attractive to foreigners. Therefore, the only easily accessible tertiary referral center in the region is the University Hospital Schleswig-Holstein Campus Kiel, located on its southern fringe (fig. 1). There is no historical, demographic or genetic evidence suggesting that etiological factors relevant in the present context differ substantially between this and other regions of Germany.

The PopGen project (www.popgen.de) was initiated by clinical and non-clinical partners at the Christian Albrechts University Kiel in 2002 to provide disease-orientated projects from the NGFN with a unique interdisciplinary platform for the identification and cross-sectional recruitment of all locally prevalent cases with the disease in question. Its establishment was originally facilitated through financial support by the Optimization and Networking Fund of the NGFN. Confinement of patient and control sampling activities to the PopGen target region ensures that

- all individuals classified as being affected by a disease can be identified using the resources of the German public health care system
- diagnostic accuracy and ascertainment efficiency can be monitored in a standardized fashion
- patients can easily be enrolled in a follow-up scheme
- a DNA bank can be established and made available to partners from within and outside the NGFN
- processing of phenotype data is centralized and consistent
- standard operating procedures and quality control measures can be implemented at all stages of the recruitment process.

Patient and Proband Recruitment

From the outset, the choice of target diseases for PopGen has been based upon

- an active interest expressed by one of the NGFN networks
- the (likely) availability of a susceptibility genotype for the disease in question
- a prevalence that matches the PopGen design (i.e. high enough to allow reliable risk assessment, but below the limits of suitability for a cohort study)
- a phenotype that can be assessed from existing files without additional examinations.

Diseases covered by PopGen, or intended to be covered in the near future, include autism, essential tremor, seizure, Parkinson's disease (PD), bipolar disorder, arteriosclerosis/CHD, dilated cardiomyopathy, atopic eczema/asthma, inflammatory bowel disease, sarcoidosis, periodontitis, colonic adenocarcinoma, and juvenile pneumonia. Patients are being identified in several different ways, including contact with regional hospitals, general practitioners, specialist inpatient clinics and health insurers. Patients are first contacted by mail, sent through their health care provider or insurer, respectively, on behalf of PopGen. Interested individuals are then asked to respond directly to the PopGen recruitment center and to give permission for PopGen personnel to obtain complete health care records, detailing the diagnostic and follow-up procedures. No information regarding the willingness, or other, of a contacted person to participate is provided to third parties. Diagnoses are verified on the basis of the available documentation using preestablished criteria as defined by the clinical partners from the respective NGFN disease-orientated network. After phenotype assessment, a representative proportion of patients are asked to participate in a follow-up scheme.

Healthy control individuals are identified through official population registries ('Einwohnermeldeamt') and contacted by mail. The control group is intended to ultimately comprise 7,200 individuals, with 2,400 people in each of three age groups (18–30, 30–50, 50–80 years). The declarations of written informed consent that are used for both patients and controls fully comply with current ICH standards for the conduct of clinical research, with some Biobank-specific items added (available from the authors upon request). A minimum 24-hour grace period for withdrawal from the study is granted to all participants prior to the pseudonymization or anonymization of their data, respectively. All recruitment and data management procedures have been approved by the Ethics Committee of the Medical Faculty of Kiel and by the data protection

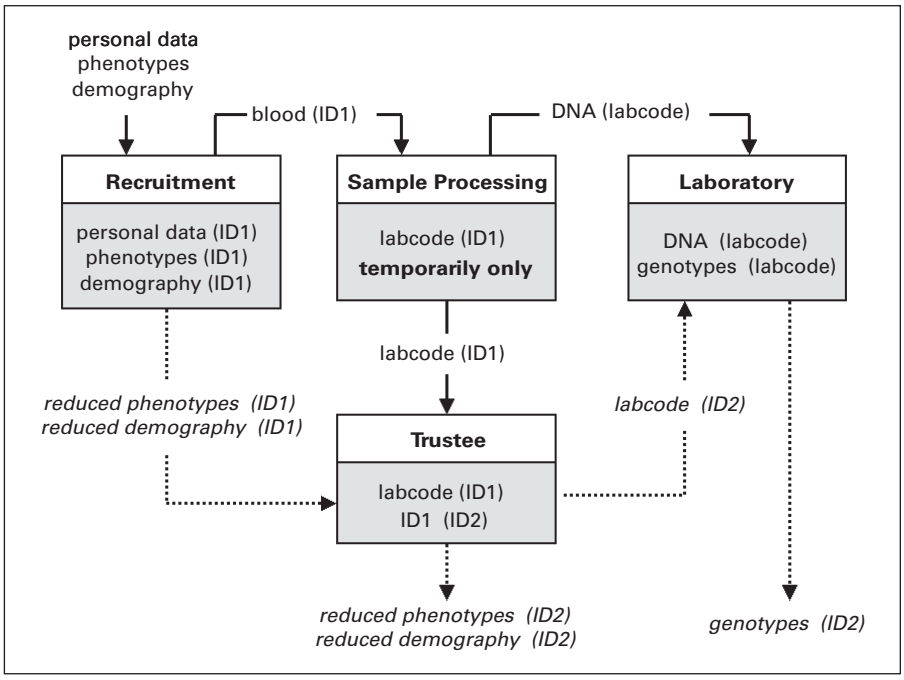

Fig. 2. Schematic representation of the PopGen data security concept. Solid arrows mark the regular flow of data and biological specimen associated with a newly recruited control or patient. No personal data or phenotypes are being stored for controls. Dotted arrows depict the information flow (in italics) associated with the scientific analysis of PopGen data. Upon request, a reduced and analysis-specific set of information is transferred from the recruitment database to the trustee server, which relabels it using identifier ID2 and forwards the information to the data analysis platform. At the same time, the trustee transfers the respective ID2/labcode relationships to the laboratory data management system with a request to sent selected genotypes to the data analysis platform, labelled by ID2 as well. All pathways are implemented such that they can only be used in the indicated, unidirectional way.

officer of the University Hospital Schleswig-Holstein, prior to commencement of the study.

All PopGen participants undergo venipuncture to obtain 30 ml of EDTA blood (yielding 600–1,000 μg DNA). This amount of blood will allow a sufficiently large number of genetic tests to be performed. Venipuncture is usually performed by local physicians, and blood is mailed directly to the PopGen recruitment office. DNA is extracted using standard techniques, followed by storage at –20°C. The DNA bank is accessible to all NGFN members for the purpose of exploring or verifying genotype-phenotype relationships, provided that promising disease-associated genetic variants have been detected for the condition in question. Access to the DNA bank must be approved by the local ethics committee and is free of charge to NGFN members.

Management and Steering

All information and biological material is being collected and maintained by a dedicated PopGen recruitment team, based at the Christian Albrechts University Kiel, have been established for the sampling and phenotyping process, including patient identification, recruitment, sample and phenotype ascertainment, and sample processing. The practical implementation of these standard operating procedures (SOPs) is constantly being improved. The PopGen group is now a self-contained structure, with a trained sociologist as managing director, drawing medical, bioinformatic, statistical, epidemiological and field-working expertise from its own staff members. The involvement of an independent medical director has served to guarantee that professional standards are adhered to in all disease- and patient-related matters.

Whilst the PopGen managing director is responsible for the day-to-day upkeep of the project, three local university institutions (the Department of General Internal Medicine, the Institute of Clinical Molecular Biology and the Institute of Medical Informatics and Statistics) have assumed joint responsibility for its supervision. In view of its central importance to all patient-related research in the NGFN, the NGFN project committee decided to inaugurate an external advisory board for PopGen in January 2004. This board involves the speakers of the disease-orientated networks and of the central methodological platforms (genotyping and genetic epidemiology) within the NGFN, with bylaws regulating the mutual responsibilities and interactions of the advisory board and the PopGen management.

Data Security

At the core of the PopGen data security concept lies the maintenance of two geographically separated databases (‘Recruitment’ and ‘Laboratory’) that dissociate

personal and phenotype data from genotypes (fig. 2). The two data types are labelled by independent identifiers (ID1 and labcode) which can only be connected via a dedicated, specifically protected trustee server. All data transfer from the recruitment office to the laboratory management system proceeds through the trustee server. This computer exclusively holds the table encoding the ID1/labcode relationship. In addition, the trustee server automatically generates a second temporary identifier, ID2, on request. This temporary identifier may be used to re-label pertinent reduced sets of personal and phenotypic data before forwarding them to the data analysis platform. The respective ID2/labcode relationships are sent to the laboratory data management system at the same time, which labels the respective genotypes by ID2 and forwards them to the analysis platform. Internal SOPs regulate the immediate destruction of these temporary genotype/phenotype files after use for the respective statistical analyses. Data traffic is protected by the firewall of the University Clinic and an additional mini-firewall around the trustee server.

Early Results, Future Plans and Discussion

For a number of diseases, identification and/or contacting of patients to be included in PopGen has already been completed by the time of writing this article. Other recruitment activities are currently in the planning phase, involving both PopGen staff and representatives of the respective NGFN disease-orientated networks.

Cardiovascular Diseases

The Arteriosclerosis/CHD project is conducted on behalf of the Cardiovascular Disease Network ('CardioNet') of the NGFN. All patients included are below 55 years of age and show significant CHD, as validated by cardiac catheterization. To achieve maximum recruitment efficiency, a close collaboration has been initiated between PopGen and the six centers performing coronary angiography in the PopGen catchment area. All cardiac catheterizations performed at these units between January 1997 and June 2003 were scrutinized (25,000 in total) and 2,200 patients identified as matching the PopGen inclusion criteria. These patients have been contacted, and 1,400 have agreed to participate in the study. Given a response rate of over 60% upon single contact, past experience from NGFN collections established under similar conditions suggests that the final recruitment rate will exceed 80% once additional approaches by mail and telephone have been made. From the treatment records, a detailed disease history is ascertained, including information on myocardial infarctions, surgery, heart and kidney function, glucose and fat metabolism for example. A standardized questionnaire is sent out to the participating patients in order to obtain additional demographic, phenotypic and environmental information. Under the same recruitment scheme as used for CHD (i.e. age $\leq$ 55 years), some 400 patients with dilated cardiomyopathy, but without CHD, could be identified. These patients have been contacted for inclusion in a separate project.

Neuropsychiatric Diseases

PopGen projects on PD and essential tremor are currently in the late planning phase. PD is a key phenotype in the Neuronal Disease Network ('NeuroNet') of the NGFN and has a prevalence of approximately 100–200 per 100,000, with an age-related increase in all populations. As yet, however, genetic risk factors for PD have been identified mostly in early-onset patients. From the PopGen target population, between 1,100 and 2,500 PD patients are expected to be identifiable, irrespective of age. From these candidates, PD patients will be selected and recruited in a two-tiered fashion. First, all neurologists and psychiatrists in the PopGen catchment area will be contacted (approximately 70, including four hospitals) and patients identified by searching the registers of board-certified physicians. Second, a cross-sectional sample of 40,000 people of at least 60 years of age will be taken from a defined subregion of the PopGen catchment area. This strategy will target PD patients who are only treated by general physicians and who would therefore escape a specialist-based recruitment scheme. A PopGen project on bipolar affective disorder (BPAD) is also in the planning phase. Since BPAD affects up to 5% of the general population, an assumed response rate of 50% will result in the inclusion of some 3,000–5,000 individuals in the PopGen BPAD sample.

Environmental Diseases

A close collaboration has been established between PopGen and the Environmental Disease Network ('EnviroNet') of the NGFN. Within the network, PopGen-based patient sampling is in progress for four chronic diseases: bronchial asthma, juvenile periodontitis, inflammatory bowel diseases (IBD; for further information, see www.mucosa.de), and sarcoidosis. The first three diseases covered by these projects are of primary importance to public health owing to their consistently high prevalence (5–10% for asthma, 1% for juvenile periodontitis and

0.5% for IBD). Generally accepted etiological concepts maintain that environmental factors are important for all of the diseases to develop but that individuals have to be genetically predisposed in order for an external stimulus to trigger disease onset. Using highly selected patient samples, multiple predisposing genes have already been identified for some of the diseases in question [9–17]; however their impact at the population level and the degree of interaction with environmental factors are still unknown.

Controls

Recruitment of PopGen controls has been nearly completed for the City of Kiel (>95%). In total, 4,000 probands have agreed to participate. Blood samples have been obtained from 3,250 of the 4,000 controls, and 3,600 samples were available by July 2005. Recruitment of a second, equally sized set of controls from rural areas surrounding the City of Kiel commenced in spring 2005. Data from control individuals are totally anonymized before inclusion in the PopGen database.

Ascertainment Efficiency and Follow-Up

The response rate is constantly monitored in each phase of the PopGen recruitment process and, if necessary, recruitment is promoted by repeated mail or telephone contacts. Cumulative prevalence figures per age group and geographical region have been provided by health insurers, thus allowing the determination of the ascertainment efficiency. Our own previous experience suggests that the local set-up in Schleswig-Holstein leads to an extremely high response rate. This is exemplified by the envisaged population-wide recruitment of over 80% of diagnosed cases with CHD. Long-term clinical research performed at the Department of Internal Medicine, Kiel, on similar instances of chronic illness, further indicates that more than 50% of patients can be subjected to long-term follow-up. Patients and their doctors will be contacted once a year to obtain information about the development of the patient's disease. This scientifically important follow-up scheme will nevertheless be confined to a set of key features and is primarily intended to help in characterizing the natural course of a disease. It will not resolve any details related to, for example, pharmacological response.

In summary, PopGen provides a unique population-based resource, focused upon a specific set of diseases, that is indispensable for the future development of genetic medicine. PopGen will thus be of pivotal importance for patient-based studies within the confines of the NGFN, but at the same time is open to external collaborators with an active interest in disease-orientated genomic research.

Acknowledgments

PopGen is currently funded by the German Ministry of Science and Education (grant numbers 1GS0121, 01GS0171, 01GR0468). As of October 2004, PopGen is funded through a 2-year grant awarded to M.K. and S.S. in the second round of NGFN funding). The authors wish to thank Timothy H. Lu, Kiel, for helpful discussion.

References

1 Terwilliger JD, Weiss KM: Linkage disequilibrium mapping of complex disease: fantasy or reality? Curr Opin Biotechnol 1998;9:578–594.
2 Schork NJ, Cardon LR, Xu X: The future of genetic epidemiology. Trends Genet 1998;14:266–272.
3 Dekker MC, van Duijn CM: Prospects of genetic epidemiology in the 21st century. Eur J Epidemiol 2003;18:607–616.
4 Wolf PA: Fifty years at Framingham: contributions to stroke epidemiology. Adv Neurol 2003;92:165–172.
5 Austin MA, Harding S, McElroy C: Genebanks: a comparison of eight proposed international genetic databases. Community Genet 203;6:37–45.
6 Barbour V: UK Biobank: a project in search of a protocol? Lancet 2003;361:1734–1738.
7 Filipiak B, Heinrich J, Schäfer T, Ring J, Wichmann HE: Farming, rural lifestyle and atopy in adults from southern Germany – results from the MONICA/KORA study Augsburg. Clin Exp Allergy 2001;31:1829–1838.
8 Klimis D, Gnardellis C, Trichopoulou A: Gender differences in blood lipids in a Greek island population. The EPIC study. Nutrition Res 2000;20:35–45.
9 Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M, Binder V, Finkel Y, Cortot A, Modigliani R, Laurent-Puig P, Gower-Rousseau C, Macry J, Colombel JF, Sahbatou M, Thomas G: Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature 2001;411:599–603.
10 Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, Ramos R, Britton H, Moran T, Karaliuskas R, Duerr RH, Achkar JP, Brant SR, Bayless TM, Kirschner BS, Hanauer SB, Nunez G, Cho JH: A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature 2001;411:603–606.
11 Hampe J, Cuthbert A, Croucher PJ, Mirza MM, Mascheretti S, Fisher S, Frenzel H, King K, Hasselmeyer A, MacPherson AJ, Bridger S, van Deventer S, Forbes A, Nikolaus S, Lennard-Jones JE, Foelsch UR, Krawczak M, Lewis C, Schreiber S, Mathew CG: Association between insertion mutation in NOD2 gene and Crohn's disease in German and British populations. Lancet 2001;357:1925–1928.

12 Stoll M, Corneliussen B, Costello CM, Waetzig GH, Mellgard B, Koch WA, Rosenstiel P, Albrecht M, Croucher PJ, Seegert D, Nikolaus S, Hampe J, Lengauer T, Pierrou S, Foelsch UR, Mathew CG, Lagerstrom-Fermer M, Schreiber S: Genetic variation in DLG5 is associated with inflammatory bowel disease. Nat Genet 2004;36:476–480.
13 Peltekova VD, Wintle RF, Rubin LA, Amos CI, Huang Q, Gu X, Newman B, Oene MV, Cescon D, Greenberg G, Griffiths AM, St George-Hyslop PH, Siminovitch KA: Functional variants of OCTN cation transporter genes are associated with Crohn disease. Nat Genet 2004;36:471–475.
14 Valentonyte R, Hampe J, Huse K, Rosenstiel P, Albrecht M, Stenzel A, Nagy M, Haesler R, Gaede KI, Platzer M, Franke A, Lengauer T, Seegert D, Schwinger E, Krawczak M, Müller-Quernheim J, Schürmann M, Schreiber S: Sarcoidosis is associated with a truncating splice site mutation in BTNL2. Nat Genet 2005;37:357-364.
15 Van Eerdewegh P, Little RD, Dupuis J, Del Mastro RG, Falls K, Simon J, Torrey D, Pandit S, McKenny J, Braunschweiger K, Walsh A, Liu Z, Hayward B, Folz C, Manning SP, Bawa A, Saracino L, Thackston M, Benchekroun Y, Capparell N, Wang M, Adair R, Feng Y, Dubois J, FitzGerald MG, Huang H, Gibson R, Allen KM, Pedan A, Danzig MR, Umland SP, Egan RW, Cuss FM, Rorke S, Clough JB, Holloway JW, Holgate ST, Keith TP: Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness. Nature 2002;418:426–430.
16 Oguma T, Palmer LJ, Birben E, Sonna LA, Asano K, Lilly CM: Role of prostanoid DP receptor variants in susceptibility to asthma. N Engl J Med 2004;351:1752–1763.
17 Nicolae D, Cox NJ, Lester LA, Schneider D, Tan Z, Billstrand C, Kuldanek S, Donfack J, Kogut P, Patel NM, Goodenbour J, Howard T, Wolf R, Koppelman GH, White SR, Parry R, Postma DS, Meyers D, Bleecker ER, Hunt JS, Solway J, Ober C: Fine mapping and positional candidate studies identify HLA-G as an asthma susceptibility gene on chromosome 6p21. Am J Hum Genet 2005;76:349–357.