Whole genome association studies of neuropsychiatric disease: An emerging era of collaborative genetic discovery

Margaret A Keller[1]
Katrina Gwinn[2]
Josefina Nash[1]
Jonathan Horsford[3]
Ran Zhang[2]
Stephen S Rich[4]
Roderick A Corriveau[1]

[1]Coriell Institute for Medical Research, Camden, NJ, USA; [2]National Institute for Neurological Disorders and Stroke-NIH, Bethesda, MD, USA; [3]National Institute of Dental and Craniofacial Research-NIH, Bethesda, MD, USA; [4]Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA

Abstract: Family history, which includes both common environmental and genetic effects, is associated with an increased risk for many neuropsychiatric diseases. Investigators have identified several disease-causing mutations for specific neuropsychiatric disorders that display Mendelian segregation. Such discoveries can lead to more rational drug design and improved intervention from a better understanding of the underlying biological mechanisms. However, a key challenge of genetic discovery in human complex diseases, including neuropsychiatric disorders, is that most diseases with genetic components display non-Mendelian patterns of inheritance. Recent advances in human population genetics include high-density genome-wide analyses of single nucleotide polymorphisms (SNPs) that make it possible to study complex genetic contributions to human disease. This approach is currently the most powerful strategy for analyzing the genetics of complex diseases. Genome-wide SNP analyses often require a large collaborative effort to collect, manage, and disseminate the numerous samples and corresponding clinical data. In this review we discuss the use of publicly available biorepositories for the collection and distribution of human genetic material, associated phenotypic information, and their use in genome-wide investigations of human neuropsychiatric diseases.

Keywords: repository, human, neurology, consent, genetics, bioinformatics

Repository science

Technical advances in molecular biology over the past 20 years, including the advent of polymerase chain reaction (PCR), the discovery of SNPs as a class of highly variable polymorphisms, and the use of multi-well format plates and associated instrumentation for many genetic analyses have provided essential building blocks for high-throughput genome-wide association studies. However, identifying, consenting and gathering relevant clinical information for thousands of individuals are challenging. Distributing human genetic materials and associated clinical information poses legal challenges and issues related to protection of human subjects. Therefore, collection and distribution of human biomaterials for genetic research is not an area into which traditional US commercial entities have ventured, and the supply of large numbers of genomic DNA samples represents a bottleneck. Evidence of the value of readily accessible and abundant human DNA comes from the study of the heritability of schizophrenia in Iceland. The relative isolation and proximity of Icelandic families with schizophrenia, together with their remarkable cooperation and consent, allowed deCODE Genetics to perform a genome-wide linkage scan that identified neuregulin I (NRG1) as a candidate gene for schizophrenia (Stefansson et al 2002). However, the specific conditions that facilitated this population-based study in the Icelandic community, and thus the efforts associated with this study are not readily reproducible in other populations. In addition, the samples used in those and most other population-based studies are not

Correspondence: Katrina A Gwinn
6001 Executive Blvd Room 2143,
NINDS-NIH, Bethesda, MD 20817, USA
Tel +1 301 496 5745
Fax +1 301 402 1501
Email gwinnk@ninds.nih.gov

© 2007 Dove Medical Press Limited. All rights reserved

publicly available. Limited access to such biomaterials and the associated clinical data is a large impediment to progress in genome-wide analyses of complex diseases, including neuropsychiatric disease. The National Institutes of Health and other contracting agencies have addressed the need for such biomaterials by funding non-profit repositories to receive, manage and distribute human biomaterials.

Repositories that generate and store human cell lines, and the genomic DNA derived from these cell lines, rely on an approach that was developed over 30 years ago in which Epstein-Barr virus (EBV) infects and transforms B lymphocytes present in whole blood (Yata et al 1975). The transformed lymphoblasts from each individual subject represent a renewable source of genetic material. Both the lymphoblasts and the DNA can be distributed to investigators, and in the case of public repositories, these materials are a valuable resource for the biomedical research community at large. Additionally, in some cases the availability of cell lines with associated genotypic and phenotypic information represents a second-generation resource for mRNA- and protein-expression analyses and other cell-based studies aimed at follow-up of genetic "hits". This strategy employs biomaterials for gene discovery, ie, the transformed lymphoblasts from which subject DNA is extracted, as a tool for further investigating the biological significance of candidate gene variation.

Consent and patient protection

The first critical step in designing a successful genome-wide association study involves the collection of human biomaterial, most often a peripheral blood specimen. This step requires oversight by an Institutional Review Board (IRB) and creation and use of an Informed Consent document. When developing such a document for a genetic study there are three primary ethical issues related to the use of specimens and associated private medical information (PMI) in research. The first is the protection of individuals who have contributed their samples or data from research risks and to weigh these risks with the potential benefits. The principal risk to subjects is the potential breach of privacy and confidentiality, with additional risks being psychosocial harms from research or research findings, and physical risks from procedures for collecting specimens, such as biopsies or blood draws. The relative benefits to society need to be considered explicitly when establishing population-based genetic research studies that are unlikely to lead to direct benefit for any individual human subject. The second major ethical issue is respect for persons; this includes respect for individuals and groups of individuals and the autonomy of these individuals to make decisions about the use and disposition of their samples and data. Attending to respect for persons should be considered independently of the need to evaluate and minimize risks; subjects may have values or beliefs in relation to uses of samples and data that are not directly linked to concerns about risk. The third ethical issue is the importance of ensuring the responsible use of valuable research resources to advance scientific knowledge about human health for the benefit of society. The research community must engage in an open and continual dialog regarding the responsible use of human genetic materials. It is wise to create an infrastructure by which biomaterials can be collected and distributed responsibly, with minimal barriers to research or inordinate costs that prevent societal investment in research from yielding benefits to human health.

In some cases there may be risks to social groups or communities due to the release of aggregate research findings that can cause anxiety, stigma, or economic consequences, even when no individually identifiable information has been revealed. In such cases community consultation may be appropriate. For example, it may be appropriate to include a community representative on an advisory or oversight board for the repository, or other extensive consultations may be undertaken in advance of initiating a research project that focuses on a particular population.

In all cases the collection, storage, distribution and use of human specimens and data should be conducted in accordance with all applicable regulations. These include 45 Code of Federal Regulation (CFR) part 46, the Food and Drug Administration human subjects regulations 21 CFR part 50 and 56, and 812; the Health Insurance Portability and Accountability Act (HIPAA) Privacy and Security Rules (45 CFR Parts 160 and 164). There are also state and local laws, such as laws governing genetic testing, regulating genetic information and medical records privacy that must be considered when designing a genetic study of human subjects.

It is noteworthy that under 45 CFR 46, research use of specimens and data that are not readily identifiable, and for which there are no links to individually identifying information, is not considered to be human subjects research. Additionally, some repositories, such as those funded by the National Institutes of Health, are contracts to non-billable entities, ie, entities that do not provide clinical care, and in those cases, the HIPAA Privacy Rule does not apply. Nonetheless, HIPAA compliance is worthwhile, as most of the specimens are likely to originate at sites that are billable entities. Repositories should consider how the HIPAA Privacy Rule may apply to the collection

and use of any individually identifiable information that is particular to their repository operations [for further information, see: "NIH: Research Repositories, Databases and the HIPAA Privacy Rule" issued July 2, 2004 (http://privacyruleandresearch.nih.gov/research_repositories.asp) and "OHRP: Guidance on Research Involving Coded Private Information or Biological Specimens", issued August 10, 2004 (http://www.hhs.gov/ohrp/humansubjects/guidance/cdebiol.pdf)].

Sufficient sample size to address complex genetic risk

How many subjects are needed for neuropsychiatric disease gene discovery? The answer to that question depends on several factors, including the genetic architecture (number of genes, their effects, and interactions with other genes and environmental risk factors), potential disease heterogeneity (genetic and environmental), and proposed study design. Because we do not know the number of loci involved in most neuropsychiatric disease etiology, nor do we know their population prevalence and penetrance, the number of subjects needed cannot be precisely predicted. Many diseases, including neuropsychiatric disease, have multiple clinical profiles that may reflect differences in underlying pathogenesis. Sub-populations or strata defined by gender, age of onset, race, ethnicity, are all potentially valuable classification variables. However, the more stratified a population of samples, the larger the sample size needed for study (Wang et al 2005).

In the case of Parkinson's disease (PD), there are examples of disease-causing alleles that are inherited in a Mendelian fashion (Simon-Sanchez et al 2005; Mizuno et al 2006), but these do not represent the majority of cases (see reviews (Gwinn-Hardy 2002, Wood-Kaczmar et al 2006). Substantial statistical genetic and empirical data suggest that 5,000 PD cases would be sufficient for discovery of individual gene effects underlying risk of complex human disease under most genetic models with 80% statistical power. Banking specimens from a minimum of 2,000 cases allows identification of SNPs with approximately a 1.5-fold or greater relative risk. For common SNPs (with minor allele frequency $>10\%$), this represents an allele frequency difference of at least 10%. In addition, replication of the experimental findings depends upon the availability of populations independent of the original cohort.

To design a study capable of detecting gene-gene or gene-environment interactions, the sample size required to maintain power would be increased by at least 4-fold. Thus, there is a clear rationale and need for large specimen collections. Public availability of such biomaterials is vital for accelerating the rate at which large-scale genetic studies can reach sufficient power to detect disease-causing genes and interacting loci.

Collecting clinical data

The quality of genotype-phenotype characterizations involving many thousands of samples acquired over a period of years depends on the quality of phenotypic definitions, uniform data acquisition, and rigorous data management. Standardized clinical criteria are essential to a large-scale sample comparison within and across clinical collection sites. Furthermore, use of these criteria can allow stratification by associated traits (endophenotypes or sub-phenotypes). Fortunately, in the neuropsychiatric disease community, a large number of clinical trials have accustomed many clinicians to standardized clinical criteria, thus facilitating collection of large, well-characterized populations. Also, the strong clinical tradition in neuropsychiatric disease sets a foundation for pharmacogenetic studies, and, when possible, pharmacological response or significant exposure should be included in clinical datasets. Additionally, well-designed collections of control subjects are crucial and in some cases may allow for the use of control subjects for multiple neuropsychiatric disorders. This is particularly important because most genome-wide association studies (GWAS) use the case/control paradigm. Identifying and enrolling large numbers of control subjects in these studies requires significant dedication and application by a collaborative team. This is due to the fact that "apparently healthy", "neurologically normal" individuals, who would be suitable for use as control subjects, are not likely to be seen in clinical neurology and clinical neuropsychiatric settings. Some case/control studies utilize unaffected spouses as control subjects in part due to their accessibility and to the fact that they are likely to share some environmental exposure history.

It is unlikely that individual researchers are capable of enrolling the large number of subjects necessary for GWAS of complex diseases. In addition, the most appropriate subjects for genetic studies are sometimes those who are the most difficult to enroll. "Typical" sporadic cases of a disease are often desirable in GWAS, yet these are the cases that are least likely to present in a subspecialty clinic. Therefore, a major collaborative effort involving a large number of clinicians is often required to enroll sufficient subjects for successful GWAS. An approach that has been used successfully in the cancer research community is one in which private practitioners that enroll subjects maintain an academic interest

and involvement in the collaborative project (Kaluzny et al 1993). This approach has catalyzed the expansion of cancer clinical trials to include a much broader and larger collection of patients. In 2003, the National Cancer Institute proposed a framework for NCI-funded large-scale biospecimen and data collection and distribution entitled, The National Biospecimen Network Blueprint (http://prostatenbnpilot.nci.nih.gov/FINAL_NBN_Blueprint.pdf).

Using the experiences gained from the structure of NIH-funded cancer clinical trials networks, the Clinical Research Collaboration (CRC) at the National Institute of Neurological Disorders and Stroke (NINDS) has undertaken a project to develop of an infrastructure to efficiently execute NINDS-sponsored clinical research (see http://www.ninds.nih.gov/news_and_events/proceedings/2002_clinical_research_workshop.htm). The CRC was established to expedite subject recruitment; it involves a wide spectrum of investigators, including practicing physicians. The involvement of the clinic-based neurologist has the additional benefit of making participation in NINDS-sponsored clinical research more accessible to patients. One of the projects that utilizes the CRC initiative enrolls subjects and collects specimens to build a sizeable, well-characterized cohort with which to perform GWAS of neuropsychiatric disorders. Additionally, the CRC at NINDS aims to engender a tradition of participation in clinical research and to educate primary physicians regarding neurogenetics research. The NINDS Human Genetics Repository (http://ccr.coriell.org/ninds), established in 2002, benefits from the CRC Sample Collection and Submission Study. The Repository, located at the Coriell Institute for Medical Research, currently has specimens from over 14,000 subjects, over 4,000 of which are already publicly available, including those with Parkinson's disease, epilepsy, cerebrovascular disease and motor neuron diseases. The NINDS Repository has detailed neuropsychiatric clinical data both for cases and controls. In addition, The NINDS Repository is similar in framework and management of processes, quality and data to the large-scale system of biospecimen and data collection and distribution proposed in the NBNB.

Bioinformatics

The utility of a sample biorepository used for GWAS is maximized by a scalable and extensible informatics infrastructure. The information management system should meet the requirements of real-time data capture, collection site management, chain of custody handling, and operational efficiency for a large number of samples, each of which is linked to individual data. Moreover, as an integrated solution, the system must manage the genotypic and phenotypic data associated with biospecimens under compliance with all relevant privacy laws.

The information systems design must not only consider the quality but also the accessibility of the biospecimens and associated data. The system should have the ability to integrate with other databases as both a source and a recipient of data. For example, the informatics infrastructure of the NINDS Human Genetics Repository provides open, internet-based access to both phenotypic and genotypic data in several data formats, such as comma-separated values text, Excel, and XML. Additionally, curated links to external data sources such as the dbSNP (http://www.ncbi.nlm.nih.gov/SNP) and Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) supplements the information available about the biospecimens. The scale of GWAS in complex human disorders demands that the data management system be secure, robust, and flexible in order to accommodate the complex and collaborative nature of such projects.

The future

The cost of high-throughput genome-wide SNP genotyping is approaching less than 1 cent per genotype, and public application of genome-wide SNP analysis has begun in earnest. Biorepositories that can supply large numbers of human specimens and associated clinical datasets are becoming increasingly important to the advancement of translational research and to the achievement of improved medical care.

Other federally-funded repositories have focused on collections of specimens for distinct missions. For example, one of the oldest NIH-funded repositories, the National Institute of General Medical Sciences (NIGMS) repository (http://ccr.coriell.org/nigms), collects biomaterials from subjects with a large number of genetic conditions, many of them displaying Mendelian inheritance. Structurally, this type of repository tends to be built from many non-interacting individual collection sites, rather than from multiple interacting sites and to have a small number of specimens from each of a large number of disorders. Collections such as these continue to be integral to the success of research aimed at identifying disease-causing mutations for Mendelian disorders (eg, Kapetanaki et al 2006; Matsuura et al 2006; Rivolta et al 2006). Most recently, the National Human Genome Research Institute (NHGRI) established a cell repository (http://ccr.coriell.org/nhgri) aimed at banking NHGRI-sponsored projects, including the International HapMap Project specimens. These samples are being collected from distinct populations

for studies of human genetic variation and are associated with no clinical information.

Additionally, other NIH-funded repositories are collecting samples for large-scale studies aimed at identifying genes that have a complex non-Mendelian influence on susceptibility. Included in this group are those dedicated to banking samples for the study of a number of neurological disorders (http://neuroscienceblueprint.nih.gov/neuroscience_resources/cell_tissue.htm). The National Heart Lung and Blood Institute (NHLBI)-sponsored Framingham Heart Study, established 50 years ago to investigate cardiovascular disease by collecting general medical information on a population of individuals from Framingham, Massachusetts (Namboodiri 1984) expanded its goals in recent years to include genetic studies and offers DNA from individuals with associated medical information to researchers. One of the first biorepositories to make genome-wide genotyping data completely and publicly accessible, in combination with detailed clinical data, is the NINDS Human Genetics Repository in collaboration with researchers at the NIA (see Fung et al 2006). This genotype/phenotype dataset (https://queue.coriell.org/Q/snp_index.asp), initially posted in March 2006, has generated considerable interest in the scientific community, having been accessed by researchers from across the globe more than 560 times. The Wellcome Trust Case Control Consortium (WTCCC), made up of 24 United Kingdom(UK)-based geneticists, is performing high-density SNP genotyping on 2000 cases for each of 7 diseases along with a common set of 3000 controls (1500 from the 1958 British Birth Cohort and 1500 UK Blood donors); similar analyses in four additional disease groups are planned. Genotype and phenotype data are stored at WTCCC database. Access to data and materials will be considered by an independent access committee 6 months after completion of the study (http://www.wtccc.org.uk/info/overview.shtml).

In 2006, the NIH launched two large GWAS initiatives with open access to genotype data at their core. The first, Genetic Association Information Network or GAIN, was established in 2006 through a public/private partnership between the Foundation of the National Institutes of Health, Pfizer Global Research and Development, Perlegen Sciences, Affymetrix Inc. and the Broad Institute; currently, it funds six studies aimed at determining the genetic factors that influence susceptibility to psoriasis, attention deficit hyperactivity disorder, schizophrenia, bipolar disorder, major depression and type 1 diabetes in a total of 18,000 individuals. High-density SNP genotype information is housed in a database managed by the National Center for Biotechnology Information (NCBI) and is available to the research community (http://www.ncbi.nlm.nih.gov/WGA/programs/GAIN). The second, Cancer Genetic Markers of Susceptibility or CGEMS, is funded by the National Cancer Institute, and is aimed at determining the genetic factors that influence susceptibility to breast and prostate cancer (http://cgems.cancer.gov). The data generated is made publicly available at the Cancer Biomedical Informatics Grid (caBIG) (http://caintegrator.nci.nih.gov/cgems). The first significant finding of this initiative is the identification of a new prostate cancer susceptibility locus as well as confirmation of a previously identified locus, both on Chromosome 8 (Yeager et al 2007).

The role of gene dosage has been shown to be an important causal factor in neurological disorders, including Alzheimer's and Parkinson's disease (eg, Martins et al 2005, Bonifati et al 2005). Also, there is renewed interest in copy number variation and its influence on human disease (Khaja et al 2006; Redon et al 2006; White and den Dunnen 2006) including neuropsychiatric disease (eg, Eriksen et al 2005; Nishioka et al 2006) as well as its role in pharmacogenomics (eg, Ouahchi et al 2006). This work is supported by new approaches for measuring such variation (Fiegler et al 2006; Ragoussis et al 2006; Tchinda and Lee 2006). Recently, the use of genome-wide SNP genotyping has been shown to increase the potential to augment or replace current methods aimed at defining chromosomal abnormalities (Simon-Sanchez et al 2006). Public availability of genome-wide SNP genotyping will greatly facilitate the rapid accumulation of equivalent data to produce an encyclopedia of normal and abnormal structural genomic variation, similar to previous efforts for gross chromosomal abnormalities (Mitelman 1998), in addition to enabling standardized cross-comparison among laboratories.

Summary

In summary, the field of neuropsychiatric disease has entered an era of complex gene discovery research. As with all genome-wide approaches to complex disease, there are continued challenges regarding determination of optimal sample sizes for affected and control populations. It is clear that biorepositories play an important role in this effort. Such repositories, involved in the collection and dissemination of specimens and associated data, must supply high quality biomaterials and implement powerful data management and bioinformatics tools to meet the needs of a complex collaborative network of collection sites and scientists. Currently, this niche is being filled in large part by federally-sponsored repositories with distinct scientific missions, but a shared

interest in supporting collaborative research and public availability of data. There is the potential for participation by the commercial sector, specifically the pharmaceutical industry, in such large-scale efforts aimed at improving human health. However, the societal investment in population-based studies of complex disease, including neuropsychiatric disease, is best served by repositories that offer broad access to results and material that can propel these efforts from bench to bedside.

Disclosure

Drs. Gwinn, Zhang, and Horsford are government (National Institutes of Health) employees.

References

Bonifati V, Rohe CF, Breedveld GJ, et al. 2005. Early-onset parkinsonism associated with PINK1 mutations: frequency, genotypes, and phenotypes. *Neurology*, 65:87–95.

Eriksen JL, Przedborski S, Petrucelli L. 2005. Gene dosage and pathogenesis of Parkinson's disease. *Trends Mol Med*, 11:91–6.

Fiegler H, Redon R, Andrews D, et al. 2006. Accurate and reliable high-throughput detection of copy number variation in the human genome. *Genome Res*, 16:1566–74.

Fung HC, Scholz S, Matarin M, et al. 2006. Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol*, 5:911–16.

Gwinn-Hardy K. 2002. Genetics of parkinsonism. *Mov Disord*, 17:645–56.

Kaluzny AD, Lacey LM, Warnecke R, et al. 1993. Predicting the performance of a strategic alliance: an analysis of the Community Clinical Oncology Program. *Health Serv Res*, 28:159–82.

Kapetanaki MG, Guerrero-Santoro J, Bisi DC, et al. 2006. The DDB1-CUL4ADDB2 ubiquitin ligase is deficient in xeroderma pigmentosum group E and targets histone H2A at UV-damaged DNA sites. *Proc Natl Acad Sci USA*, 103:2588–93.

Khaja R, Zhang J, Macdonald JR, et al. 2006. Genome assembly comparison identifies structural variants in the human genome. *Nat Genet*, 38:1413–18.

Martins CA, Oulhaj A, de Jager, et al. 2005. APOE alleles predict the rate of cognitive decline in Alzheimer disease: a nonlinear model. *Neurology*, 65:1888–93.

Matsuura S, Matsumoto Y, Morishima K, et al. 2006. Monoallelic BUB1B mutations and defective mitotic-spindle checkpoint in seven families with premature chromatid separation (PCS) syndrome. *Am J Med Genet A*, 140:358–67.

Mitelman F, ed. 1998. Catalog of Chromosome Aberrations in Cancer '98.

Mizuno Y, Hattori N, Yoshino H, et al. 2006. Progress in familial Parkinson's disease. *J Neural Transm Suppl*, 191–204.

Namboodiri KK. 1984. Framingham Heart Study: review of genetic data and design, limitations and prospects. *Prog Clin Biol Res*, 147:65–78.

Nishioka K, Hayashi S, Farrer MJ, et al. 2006. Clinical heterogeneity of alpha-synuclein gene duplication in Parkinson's disease. *Ann Neurol*, 59:298–309.

Ouahchi K, Lindeman N, Lee C. 2006. Copy number variants and pharmacogenomics. *Pharmacogenomics*, 7:25–9.

Ragoussis J, Elvidge GP, Kaur K, et al. 2006. Matrix-assisted laser desorption/ionisation, time-of-flight mass spectrometry in genomics research. *PLoS Genet*, 2:e100.

Redon R, Ishikawa S, Fitch KR, et al. 2006. Global variation in copy number in the human genome. *Nature*, 444:444–54.

Rivolta C, McGee TL, Rio Frio T, et al. 2006. Variation in retinitis pigmentosa-11 (PRPF31 or RP11) gene expression between symptomatic and asymptomatic patients with dominant RP11 mutations. *Hum Mutat*, 27:644–53.

Simon-Sanchez J, Hanson M, Singleton A, et al. 2005. Analysis of SCA-2 and SCA-3 repeats in Parkinsonism: evidence of SCA-2 expansion in a family with autosomal dominant Parkinson's disease. *Neurosci Lett*, 382:191–4.

Simon-Sanchez J, Scholz S, Fung HC, et al. 2006. Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. *Hum Mol Genet*.

Stefansson H, Sigurdsson E, Steinthorsdottir V, et al. 2002. Neuregulin 1 and susceptibility to schizophrenia. *Am J Hum Genet*, 71:877–92.

Tchinda J, Lee C. 2006. Detecting copy number variation in the human genome using comparative genomic hybridization. *Biotechniques*, 41:385, 387, 389 passim.

Wang WY, Barratt BJ, Clayton DG, et al. 2005. Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet*, 6:109–18.

White SJ, den Dunnen JT. 2006. Copy number variation in the genome; the human DMD gene as an example. *Cytogenet Genome Res*, 115:240–6.

Wood-Kaczmar A, Gandhi S, Wood NW. 2006. Understanding the molecular causes of Parkinson's disease. *Trends Mol Med*, 12:521–8.

Yata J, Desgranges C, Nakagawa T, et al. 1975. Lymphoblastoid transformation and kinetics of appearance of viral nuclear antigen (EBNA) in cord-blood lymphocytes infected by Epstein-Barr Virus (EBV). *Int J Cancer*, 15:377–84.

Yeager M, Orr N, Hayes RB, et al. 2007. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat Genet*.