Principles for the post-GWAS functional characterisation of cancer risk loci

Matthew L. Freedman[1], Alvaro N. A. Monteiro[2], Simon A Gayther[3], Gerhard A. Coetzee[4], Angela Risch[5], Christoph Plass[5], Graham Casey[6], Mariella De Biasi[7], Chris Carlson[8], Dave Duggan[9], Michael James[10], Pengyuan Liu[10], Jay W. Tichelaar[10], Haris G.Vikis[10], Ming You[10], Ian G.Mills[11]*

*To whom correspondence should be addressed: ian.mills@ncmm.uio.no

1. The Eli and Edythe L. Broad Institute of MIT and Harvard, Cambridge MA 02142 and · Departments of Medical Oncology, Dana-Farber Cancer Institute, Boston MA 02115
2. Cancer Epidemiology Program, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida, 33612, USA
3. University of Southern California, Keck School of Medicine, Los Angeles, CA, USA and Translational Research Laboratory, University College London EGA Institute for Women's Health, London, United Kingdom.
4. Department of Urology, Norris Cancer Center, University of Southern California, Los Angeles, CA 90033
5. German Cancer Research Center, Division of Epigenomics and Cancer Risk Factors, Heidelberg, Germany
6. Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA
   Cambridge, Research Institute, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE
7. Department of Neuroscience, Baylor College of Medicine, Houston, TX 77030
8. Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N., M4-B402, P.O. Bo 1024, Seattle, WA 98109–1024
9. Translational Genomics Research Institute (TGen), Phoenix
10. Medical College of Wisconsin, Milwaukee, Wisconsin 53226, USA
11. Norwegian Centre for Molecular Medicine, Nordic EMBL Partnership, University of Oslo, Blindern N-0317 Oslo, Norway and Uro-Oncology Research Group, Cancer Research UK

Abstract

Genome wide association studies (GWAS) have identified more than 200, mostly novel common low-penetrance susceptibility loci for a range of different cancer types. The predicted risk associated with each locus is generally modest (per allele odds ratios usually ranging from 1.15 to 1.3) and so, presumably, are the functional effects of individual genetic variants conferring disease susceptibility. Perhaps the greatest challenge in the 'post-GWAS' era is to understand the functional consequences underlying these loci. In doing so, novel biological insights may be revealed leading to clinical benefits, including the development of reliable biomarkers, effective screening and disease prevention strategies. The purpose of this article is to propose some principles for the initial functional characterization of cancer risk loci, with a focus on non-coding variants, and to define 'post-GWAS' functional characterisation.

Introduction

The last few years have witnessed an explosion in the number of common, low-penetrance susceptibility alleles identified for a range of complex diseases and other common traits using genome-wide association studies (GWAS)[1]. In December 2010 there were 1212 published GWAS studies reporting associations at $p<5x10^{-8}$ for 210 traits (**Figure 1**); and the Catalog of Published Genome-Wide Association Studies states that by March 2011, 812 publications reported 3977 SNP associations[2]. It is believed that this represents a small fraction of the common low penetrance susceptibility loci that will eventually be identified. Despite these clear successes in identifying risk associated loci, the causal variant and/or the molecular basis of risk etiology has been determined for only a small fraction of these associations[3-5]. Plausible candidate genes can clearly be suggested based on proximity to risk loci but very few have so far been defined in a more systematic manner (**Supplementary Table 1**).

A recent *Nature Genetics* editorial on post-GWAS analyses began to put this problem into sharper focus and suggested that there should be a more significant investment in functional characterization of risk loci[6] across diseases. Similar guidelines have been proposed elsewhere in the context of loci that underlie non-neoplastic traits such as cardiovascular disease and diabetes[7]. Our focus here is cancer biology. The particularly complex interplay between genetics and the environment in many cancers poses a particularly exciting challenge for those interested in post-GWAS research. The goal of this perspective is to build upon this dialogue and to elaborate a systematic strategy for understanding how cancer-associated polymorphisms exert their effects. We mostly refer to single nucleotide polymorphisms (SNPs) throughout the paper, but recognize that other types of common genetic or epigenetic variation (e.g. copy number variants) may influence risk.

A central objective of human genetic research is to fully elucidate how a risk allele(s) initiates disease pathogenesis. Typically, there is progression from statistical *association* between genetic variation and trait/disease variation to *functionality* and *causality*. For Mendelian disorders where most of the DNA sequence changes occur in protein coding regions, the functional and causal consequences are more straightforward because of the knowledge of the genetic code. For non-Mendelian/multifactorial traits, most of the common DNA variants discovered to date map to non-protein coding regions (http://www.genome.gov/gwastudies/), where our understanding of the functional consequences and causality is more rudimentary. Specifically, the hypothesis we address is that the trait-associated alleles exert their effects through influencing gene expression levels (or splicing) through multiple mechanisms.

A dominant theme of this paper will be on how polymorphisms can act upon and influence the various aspects of the transcriptional machinery resulting in an alteration in transcriptional (or splicing) output. We will discuss the importance of developing appropriate assays and models that can be used to test the functional effects of both SNPs and genes mapping to cancer predisposition loci. Although much of what is written is applicable to alleles discovered for any trait, the section on modelling gene effects will emphasize measuring cancer-related phenotypes.

It should be recognized that at some loci, multiple, independently associated risk alleles rather than single risk alleles may be functionally responsible for the occurrence of the disease. Genotyping susceptibility loci (and their correlated variants) in multiple populations with different LD structures may prove effective at reducing substantially the number of potentially causative variants (i.e. the same causal variant may segregate in multiple populations), as shown by Udler and colleagues for the FGFR2 locus in breast cancer[8], but for most loci there will remain a set of potentially causative variants that cannot be separated at the statistical level from case-control genotype data.

Thus, follow-up of a susceptibility locus is comprised of re-sequencing the region to ascertain all genetic variation, identifying candidate functional/causal polymorphisms and identifying candidate causal genes. Ideally, the identification of a *causal* SNP would be the next step to reveal the molecular mechanisms of risk modification. Practically, however, it is unclear what the criteria for causality should be, particularly in non-protein coding regions. Thus, while we propose a set of analyses, we acknowledge that it is necessarily an operational definition and the techniques and methods will continue to evolve with the field.

Genetic annotation of trait-associated regions

In the vast majority of cases, any association identified through GWAS would be predicted to be between the disease trait and a surrogate maker (i.e., tagSNP) rather than a causal variant, as SNP arrays were designed using surrogates chosen to capture LD structure on SNP arrays rather than for any functional reasons. Therefore evaluation of all common SNPs across associated regions will be desirable to fully characterize the biologic implications of disease associations.

While the current HapMap information is incomplete with respect to identifying all common variant information and may not capture the genetic diversity of the populations being used in association studies, this is a rapidly moving field. The ongoing 1000 Genomes Project seeks to address this missing data and it is expected that it will eventually capture the majority of common (>5%) and less common (1-5%) variants[9]. However, it remains to be determined whether it yet provides complete SNP coverage (>1% MAF) across the entire genome including intergenic regions and gene deserts, where the majority of associations have been mapped. That suggests that for at least some loci targeted sequencing may remain a necessity. In addition, there is a growing interest in the role of multiple rare variants (<1%) in disease risk, which will not be adequately captured by 1000 Genomes Project. It is quite likely that in the near future, full genome sequencing on large enough numbers of subjects, will have been completed, producing a reasonably complete catalogue of human genetic variation.

When considering targeted sequencing, understanding LD structure in the region across a risk locus will be critical to delimit the size of the target region and define the costs of the undertaking. It is still unclear which r^2 threshold should be set in defining LD structure as the causal SNP potentially could be in LD with the associated SNP at an r^2 of 0.2 or even less. Ideally, the LD structure should be defined using the GWAS population genotyping results used to identify the original association. Alternatively LD structure does not necessarily need to be considered and thresholds can be set. Power can be defined as a function of $1/r^2$ but it is presently difficult to be highly prescriptive about thresholds and researchers will make decisions based on cost, a priori information, and LD structure as described below. Although somewhat arbitrary these approaches represent a workable start point in the absence of data precisely defining LD structure as a start point. Additionally depth of coverage and the number of subjects to be sequenced are important considerations that are currently severely impacted by the cost of sequencing that is expected to rapidly decline in the coming years as newer technologies come on line.

Additional biological data could also be used to define the region for analysis. For example, boundaries could be reduced due to a compelling candidate gene/transcript mapping to the region. However if any of these *a priori* hypotheses based on known biological data are used to reduce the extent of sequencing it should be recognized that this decision is generally made for cost reduction purposes only. Relying on biological assumptions undermines the agnostic approach one of the main advantages of GWAS. Finally, the majority of published GWAS data comes from European

populations, but incorporating GWAS information from other ethnic groups such as African-Americans could potentially reduce the target region if a similar association was found in this population as the African-American population generally has smaller LD block structure than the European population, as exemplified by a recent study on cocaine dependence[10]. Therefore, the extent of targeted sequencing will need to be a compromise between the size of the region to be sequenced, incorporating information such as LD structure, and the overall sequencing cost.

Annotating risk regions as regulatory elements and identifying functional variants

Characterizing the regulatory landscape of susceptibility regions represents an important step in understanding how risk alleles affect function. The most abundant of these regulatory sequences are enhancers, but other regulators such as promoters, insulators and silencers may also be susceptibility targets. Unlike core promoters (at transcription start sites of genes), distal regulatory sequences such as enhancers are often cell-type specific [11] and thus may explain the tissue and disease specific nature of common susceptibility alleles. Studying histone modifications or DNase sensitivity (or hypersensitivity) has proven to be a powerful approach to annotating tissue specific regulatory elements[12,13], and is more informative than studying conservation, since regulatory elements may be unconstrained across mammalian evolution [14-16]. Using chromatin annotations to identify putative functional SNPs within regulatory sequences at known susceptibility loci has been proposed recently[17]. More precise demarcation of such regulatory regions may be achieved by assessing the association of candidate transcription factors with response elements. Both histone modifications and transcription factor occupied regions are currently identified using ChIP-seq methodologies and signals yield short DNA stretches (typically <1kb) amenable to detailed analyses. Enhancer activity in such regulatory regions can be assayed using reporter genes *in vitro*[5] and/or *in vivo*[12].

Integrating knowledge of regulatory sequences at risk loci with catalogues of risk associated SNPs at these loci may be an efficient approach to prioritizing both candidate regulatory sites and the most likely functional variants. This concept is illustrated by work on 8q24 risk loci. Two functional SNPs at chromosome 8q24 respectively have been associated with prostate and colorectal cancer. Several transcriptional enhancers were identified at 8q24. Two of them, in a prostate cancer risk region, were occupied by the androgen receptor and responded to androgen treatment; with one containing a single nucleotide polymorphism within a FoxA1 binding site[5]. The prostate cancer risk allele facilitated both stronger FoxA1 binding and stronger androgen responsiveness. In a separate study an 8q24 SNP in colorectal cancer was also found situated within a transcriptional enhancer and the enhancer activity was affected by the SNP[18]. In addition the SNP was shown to physically interact with the *MYC* proto-oncogene, with allele-dependent binding of transcription factor 7-like 2 (TCF7L2). More detailed functional follow-up of these SNPs can then be performed using biochemical approaches to study differential transcription factor binding and activity (e.g., ChIP or EMSA). Regulatory sequences containing functional SNPs determined in this way can then be matched to their physiological target genes (see below).

After generating data that implicate a functional mechanism, the next challenge will be to identify genes that are regulated by these elements. Possible approaches for identifying targets of regulatory sequences include: (i) Knocking out regulatory sequences in mouse models followed by genome-wide gene expression analyses after knockout to identify candidate targets; (ii) Using the regulatory sequences as baits in chromatin conformation capture (3C) based studies[19,20], including genome-wide 3C-based methods; (iii) targeted editing using somatic cell knock-in technology; for example, allelic series in isogenic settings may be created and gene expression differences measured, either in naturally growing cells or in cells that are perturbed (e.g., by radiation or hormones); (iv) Identifying correlations between the different genotypes of trait-associated SNPs and variations in the transcript abundance of candidate genes at those loci. Of these, the last approach represents a straightforward method to identify putative target genes.

Susceptibility loci and the regulation of gene expression through epigenetic mechanisms

Promoter methylation, histone tail modifications and altered expression of non-coding RNAs, such as the large intergenic noncoding RNAs (lincRNAs)[21,22] which associate with chromatin modifying complexes, also contribute to gene regulation and represent obvious candidate targets of functional genetic associations[23]. Epigenetic silencing has been shown to be the predominant mechanism of gene silencing during tumor development for a subset of genes [24]. For other genes, a combination of genetic and epigenetic mechanism can contribute to tumor suppressor gene activation [25]. Epigenetic mechanisms also play an important role in mediating environmental influences on gene expression [26]. At susceptibility loci, the key questions are: (1) Do common genetic variants influence the epigenetic landscape to increase disease susceptibility? (2) Do susceptibility variants within the epigenetic landscape affect the likelihood of gene silencing during tumor development?

The ability to perform such studies has been made possible through the development of platforms that enable high throughput DNA methylation profiling at single CpG resolution [27]. Studies of hereditary non-polyposis colorectal cancer (Lynch syndrome) suggest that germline genetic variation may affect epigenetic marks resulting in cancer predisposition [28,29]. Epimutations may be a consequence of *cis* or *trans* acting genetic variants [30]. For example, Kerkel et. al have shown sequence-dependent allele-specific methylation and demonstrated that *cis*-regulatory polymorphisms control gene expression and affect chromatin states[31]. Further epigenetic mechanisms that modulate gene expression include miRNAs and miRNA binding sites which can be directly affected by SNPs [32], and tandem repeats that can impact gene expression e.g. by altering transcription factor binding sites, but also by affecting chromatin structure (reviewed in [33]).

Another target of SNP associated regulation is the chromatin network. Chromatin fibres dynamically explore the nuclear space to establish meta-stable, long-range interactions with other chromatin fibres[34]. The functional outcome of such interactions is largely unknown, but it has been demonstrated that they are capable of transferring epigenetic marks to modulate transcriptional processes both in cis [35] and in trans[36,37]. In this way, chromosome crosstalk sets the stage for the spreading and propagation of pleiotropic epigenetic effects in a manner that reflects the topology of the network involved [34]. Sequence polymorphisms can influence communication between different parts of the genome [38] and so SNPs can probably influence chromatin networks in a genotype-specific manner. For example, single SNPs or combinations of SNPs may confer disease susceptibility by promoting or antagonizing the formation of chromatin networks. The functional annotation of susceptibility loci with respect to chromatin/chromosomal networks may therefore provide important insights into the function of germline genetic variants.

Inherited Variation and Gene Expression at Susceptibility Loci

Both empirical and computational data support the notion that a considerable proportion of trait-associated loci will harbour variants that influence the abundance of specific transcripts. These polymorphisms are often referred to as expression quantitative trait loci (eQTLs)[39] [40-43]. Several landmark studies have unequivocally demonstrated that many transcripts in the human genome are influenced by inherited variation [44-48]. Studying the associations between genetic variation and gene expression offers a straightforward way to begin the complicated task of connecting risk variants to their putative target genes or transcripts. Significantly, and like GWAS, an agnostic approach can be taken to these analyses, which does not require the disease causing allele to be known.

eQTLs can be located either near the gene they regulate or considerable distances away. The distinction between local and distant is often arbitrary; however, for most studies local has often been defined as being within 1 megabase of the variant under consideration. 'Distant' can involve interactions between an eQTL and a gene located on different non-homologous chromosomes. The terminology of local and distant in this context, is preferred to cis- and trans-, which connotes mechanism [49]. It should be noted that not only mRNA transcripts but also micro RNA (miRNA) and non-coding RNA (ncRNA) transcripts should be considered as candidates.

Certain principles have emerged from eQTL studies: i) eQTLs tend to explain a greater proportion of trait variance than is typically seen for risk alleles and clinical traits; this observation translates into the ability to perform an eQTL study with smaller sample sizes than association studies for clinical traits (such as disease risk), ii) local eQTLs tend to have larger effects on gene expression than distant eQTLs and are therefore easier to discover, iii) there are likely more distant than local eQTLs [50]

Many of the initial successful eQTL studies relied on lymphoblastoid cell lines largely due to their availability[51] [40]. More recently, eQTL studies have been performed in primary human tissues and demonstrate that at least some associations are tissue-specific [41,43,52]. Reasonably large sample sizes (although typically smaller than GWAS to identify risk alleles) are needed in order to achieve sufficient power to detect eQTL associations. Consequently, comprehensive bio-banks of normal tissues will need to be established to evaluate expression differences between the different alleles of a SNP. Establishing such biobanks will be a significant part of the challenge; whereas extensive efforts within the cancer research community have established tumor tissue biorepositories, it has been less common to do so for normal tissues from the cells representing the origin of cancers. This issue is particularly problematic for tumor subtypes in which the cell of origin is still debated. This challenge is now being recognized and addressed through funding initiatives such as the 'Genotype-Tissue Expression (GTEx)' supported by the NIH Common Fund.

A complementary and powerful approach to defining local eQTLs is to measure allelic imbalance (also called allele specific gene expression or ASE) in individuals that are heterozygous for a risk allele. Any transcript demonstrating a deviation from a 1:1 ratio (as typically measured by a transcribed heterozygous marker) becomes a strong candidate gene. [53-55] It is critical to note that even if a transcript is associated with a risk allele, it does not necessarily mean that the gene is definitively involved in the trait of interest; functional follow-up with assays relevant to the trait are still needed to demonstrate that a gene is directly involved with disease development.

False negatives (i.e. where the risk associated allele is *not* associated with an expression trait) can occur, because gene expression varies in time and space. Therefore, the developmental time point and/or the tissue being studied may not be appropriate. Effects on transcript abundance may be subtle and therefore below the sensitivity threshold of a particular platform and/or sample size may not be adequate. In addition, transcript abundance is usually evaluated under steady-state conditions. Lastly, effects may only be revealed in certain contexts, such as perturbation of a particular pathway and may occur through changes in gene transcripts mediated by alterations in microRNAs/non-coding RNAs rather than through direct effects on genes. In these cases, alternative assays will be required to implicate these genes.

Future areas of exploration for the field include: (i) Defining the appropriate target tissues to examine. Risk alleles may act in a non-cell or -tissue autonomous fashion and therefore may exert their effect through other cell types that act upon the target tissue under consideration; (ii) Defining the significance of eQTL analysis in tumour as well as normal tissue. We advocate that both tissue states should be studied until a clearer picture of the relationship between the two emerges; (iii) Utilizing higher order computational methods, such as network analysis using risk variant and gene expression data to dissect the pathways driving disease pathogenesis. This ranges from transcriptomic analysis to predict the regulatory influence of transcription factors over gene networks dependency using tools such as ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks)[56], through to Bayesian network approaches to identify predictive relationships between genes from a combination of expression and eQTL data[57]. Whilst these tools are elegant, the ability to translate their outputs into biological significance is heavily dependent on the availability of manipulable and relevant model systems with which to test the predicted connectivity. These approaches clearly pose validation challenges for many diseases, however the computational biology field is a powerful and essential catalyst for post-GWAS studies. Extensive discussion of the challenges and opportunities created by computational biology lie beyond the scope of this article.

Complex Disease Modeling to Characterize the Role of Candidate Susceptibility Genes in Disease Development

Once there is sufficient evidence in support of a candidate susceptibility gene, more detailed functional studies will be required to characterise the gene's role in the pathogenesis of the trait under consideration. Gaining a better understanding of the biological mechanisms of cancer development often relies on the analysis of models that reflect the human disease and the application of technologies that facilitate the analysis of these models (**Supplementary Table 2**). It is likely that establishing a functional rationale underlying the significance of allelic variation and candidate genes at common low penetrance susceptibility loci in biologically relevant disease models will become a major component of following-up the genes emerging from GWAS. Disease models can be based on either the *in vitro* characterization of human tissues (primary tissues or cells in culture) or *in vivo* models of disease development.

Human cancer *in vitro* models represent the most accessible way to test the function of candidate genes at susceptibility loci in tumour development, but functional effects may be masked by an aberrant genetic background. Most GWAS to date have focused on genetic susceptibility to disease and so the greatest functional impact may be observed in an essentially healthy, non-aberrant tissue/background. This is perhaps the hardest context to replicate and maintain in a laboratory situation meaning that there will be a continuous drive for improvements in the models used.

Progress in establishing suitable *in vitro* models of normal tissues has been hampered by difficulties in accessing specimens, and the challenges of culturing primary cells. For example prostate epithelial cells are dependent on the presence of a co-cultured stromal component for establishing the secretory cell phenotype and functional differentiation. For the normal colon, most commercially available normal epithelial cell lines are fetal in origin, and differences in fetal and adult cell biology limits the translational potential of work using fetal cells to model adult epithelial cancer genesis. There are exceptions - in breast, well-characterized commercially available cell lines exist that represent good models of normal breast tissue (e.g. MCF10A cells and immortalized HMECs). Three-dimensional (3D) cultures of MCF10As form polarized cystic structures that closely reflect the architecture and molecular features of breast acini *in vivo*. Using this system, a link between loss of BRCA1 function and impaired luminal differentiation of mammary epithelia was established[58]; this link has been further highlighted in Proia et al.,[59]. By using such 3D models it is therefore possible to dissect subtle phenotypes, such as changes associated with gene dosage.

As a first step, we recommend measuring cancer related traits in these more 'traditional' models. Targeting genes under the control of functional SNP-containing regulatory regions may have important roles in characteristics of developing cancer phenotypes, such as proliferation, migration and apoptosis. Endpoints of the cancer phenotype, such as cell division, migration and apoptosis rates and protease secretion may be measured in cultured cells and mouse xenografts after the overexpression of the genes of interest, or their selected siRNA/shRNA knockdown.

Conclusion

The GWAS community has arrived at an important crossroads. As resources are limited the debate revolves around whether enough progress has been made towards identifying the polymorphisms that are likely to contribute most to disease causation to invest in functional follow-up. As sequencing technologies become cheaper and more accessible, as datasets expand, we argue that this will evolve rapidly and will afford greater certainty in defining both the spectrum of inherited variation and the LD structure within the regions in which they lie. This will require a detailed mapping and annotation of epigenetic and transcriptomic landscapes within which a major limiting factor may prove to be the sample collections themselves. While this progresses it is vital that proof-of-principle studies develop the methodologies and take forward the strongest candidate SNPs identified so far, not necessarily to test their causative association with disease but to understand their functional impact. 'Strong' candidate SNPs are those that demonstrate significant associations with transcript expression (eQTL analysis and chromosome conformation capture), tissue specificity and the phenotypic impacts of these transcript associations on model systems in downstream experiments. It is therefore far too soon

in this emerging field, to make definitive recommendations of what unequivocally proves a correlation between genotype and phenotype at common low penetrance susceptibility loci. Successfully making the transition to progress experimentally through this process will require collective thinking at a consortia/multi-group level, just as effective international collaborations led to the identification susceptibility loci through GWAS. It is essential for the field that this overrides the temptation to publish fragmentary work capturing only sub-steps in this sequence. Over time, integration of the re-sequenced, epigenetic and molecular-epidemiological data within different ethnic groups (and thus within different linkage disequilibrium structures) will help localize causal variants. If we begin considering how to explore the functional impact of variants now we will, as a community, be well positioned to rise to the challenge of testing causation in the future.

The field is still making the first forays into the functional characterization of SNPs and is many steps away from proving causality for the vast majority of risk associated genetic variants that have been found. Naturally our ability to get close to this goal will need to be assessed in the context of current technologies and knowledge on a disease-by-disease basis. We hope that this article will help to frame the developing debate and the emerging research that seeks to rise to this great challenge.

Acknowledgements

We would like to thank Dr. Fred Bunz and all the members of the NIH Post-Genome Wide Association Initiative for helpful discussions and in particular Dr. Ian Tomlinson (Wellcome Trust Centre for Human Genetics, Oxford). The contributing groups are supported by funding made available through the NIH Post-Genome Wide Association Initiative in response to Call (grants.nih.gov...). This Call sustains research across five cancer organ sites (prostate: 1U19CA148537-01, breast: 1U19CA148065-01, ovarian: 1U19CA148112-01, colorectal: 1U19CA148107-01 and lung: 1U19CA148127-01). For further information on this Initiative please refer to the website: (epi.grants.cancer.gov...). In addition this article is the product of the first attempt to engage the entire scientific community in the drafting of a scientific paper through open-access websites. We would like to thank Robert Hoffmann at Wikigenes for hosting the pre-submission version of this submission (http://www.wikigenes.org/e/pub/e/84.html) and his unstinting energy and enthusiasm for this project and also Nature Precedings for hosting the same version (http://precedings.nature.com/documents/5162/version/1). A number of people made contributions through the Wikigenes website or in response to the posting of a draft version these additional contributors are Alessandra Bisio (Centre for Integrative Biology (CIBIO), University of Trento, Italy), Dan Bolser (Dundee University, UK), David F Burke (Department of Zoology, University of Cambridge, UK), Yari Ciribilli (Centre for Integrative Biology (CIBIO), University of Trento, Italy), Lucia Conde (Environmental Health Sciences, UC Berkeley, USA), Giovanni Marco Dall'Olio (Institut de Biologia Evolutiva (CSIC-UPF), Barcelona, Spain), Doug Easton (CR-UK Genetic Epidemiology Unit, University of Cambridge,UK), Rosalind Eeles (Translational Cancer Genetics Team, The Institute of Cancer Research and Royal Marsden NHS Foundation Trust, UK), Johannes Engelken (Instituto de Biologia Evolutiva (UPF-CSIC), Barcelona, Spain), Marta Ramirez Gaite (Wikigenes), Evgeny A. Glazov (Diamantina Institute, The University of Queensland, Australia),Jeremy Leighton John (The British Library, UK), Kevin L. Keys (Universitat Pompeu Fabra, Barcelona, Spain), Anchit Khanna (Institute of Medical Technology, University of Tampere and Univesity Hospital, Finland), Georgios D. Kitsios (Institute for Clinical Research and Health Policy Studies, Boston, USA), S. Lillioja (Illawarra Health and Medical Research Institute, University of Wollongong, Australia),Mary Mangan (OpenHelix LLC, USA), Christopher Maxwell (Child and Family Research Institute, University of British Columbia, Canada), Sumit Middha, Pooja Mohan (University of British Columbia, Canada), Paulo Nuin (Queen's University/Ontario Cancer Biomarker Network, Canada), Rolf Ohlsson (Department of Microbiology, Tumor and Cell Biology, Karolinska Institute, Sweden),Mingxiong Pang (Section of Molecular Cell and Developmental Biology, The University of Texas at Austin, USA),Chilakamarti V. Ramana (Department of Medicine, Dartmouth Medical School, USA),Amiya Sarkar (B.S. Medical College, Bankura, West Bengal Medical Education Service, India),Khader Shameer (Mayo Clinic, USA), Christine F. Skibola (School of Public Health, UC Berkeley, USA), Rayna Stamboliyska (Evolutionary biology, Ludwig-Maximilians Universität (LMU), Munich), Muy-Teck The (Barts & the London School of Medicine & Dentistry, Queen Mary University of London, UK), Tao Zhang (The research institute for children, children's hospital, New Orleans, USA). We regret that owing to space constraints and thematic continuity we could not include all of these contributions in this version but all may still be viewed through Wikigenes using the link provided (http://www.wikigenes.org/e/pub/e/84.html). Finally, we would like to thank Myles Axton at Nature Genetics who proposed and commissioned this article and promoted the concept of community outreach through these websites. We hope that this will be the first of many valuable examples of increased engagement with scientists through these avenues.

Glossary

3C, 4C, 5C, Hi-C, 3C-seq: chromosome conformation capture (3C) is a technique is used to identify interactions between genes and long-range regulatory made possible by chromosome loops that bring the two regions to physical proximity. Developments on this method include circularization (4C) of the genomic fragments with the use of inverse PCR primers, carbon copy (5C) technology for multiplexed ligation-mediated amplification, and high throughput analysis by massively parallel sequencing between many baits and targets (Hi-C), or many targets from a single locus (3C-seq).

Supporting references: [34,60-62]

Causal variant: In the context of GWAS it represents the SNP that is mechanistically linked to risk enhancement. This is distinct from SNPs that do not have any functional impact but are statistically associated with the disease phenotype because it is in linkage disequilibrium with the causal variant.

Supporting evidence: [63]

ChIP-Seq: Chromatin immunoprecipitation (ChIP) is a method to study protein-DNA interactions. It identifies genomic regions that are binding sites for a known protein. Analysis of these regions is typically performed by PCR, when there is a hypothesized known binding site, or through the use of genomic microarrays (ChIP-chip). Alternatively, analysis can be done using next-generation sequencing (Seq) technology to analyze DNA fragments.

Supporting references: [64,65]

CNV: Copy number variation is a type of structural variation in which a particular segment of the genome, typically larger than 1kb, is found to have a variable copy number from a reference genome.

Supporting references: [66-68]

Deep sequencing: a sequencing strategy used to reveal variations present at extremely low levels in a sample. For example, to identify rare somatic mutations found in a small number of cells in a tumor, or low abundance transcripts in transcriptome analysis.

Supporting references: [69]

DNA Methylation: A modification of the DNA that involves predominantly the addition of a methyl group to the 5 position of the pyrimidine ring of a cytosine found in a CpG dinucleotide sequence.

Supporting references: [70]

EMSA: Electrophoretic Mobility Shift Assays assess the ability of a protein factor (or a complex) to bind to a specific DNA sequence

Supporting references: [71].

Epigenetic markers: an array of modifications to DNA and histones independent of changes in nucleotide sequence but rather the addition of a methyl group to cytosine and a series of post-translation modifications of histones including methylation, acetylation, and phosphorylation.

Supporting references: [72]

eQTL: Expression Quantitative Trait Loci refer to regions in the genome that control a quantitative trait, in this case mRNA (or protein) expression. These loci can be located close (sometimes referred to as "local" or "*cis*-eQTL") or far away (sometimes referred to as "distant" or "*trans*-eQTL") from a target gene.

Supporting references: [50]

Fine mapping: a strategy to identify other lower frequency variants in a disease-associated region (typically spanning a haplotype block) not represented in the initial genotyping platform with the goal of uncovering candidate causal variants. It can include data mining of publically available sequencing efforts, such as the 1000 Genomes Project and targeted re-sequencing.

Supporting references: [63,73]

Functional variant: a variant that confers a detectable functional impact on the locus. It can represent a change in coding region but also changes in regulatory regions that have an impact on function.

Supporting references: [5]

GWAS: genome-wide association study is a case-control study design in which most loci in the genome are interrogated for association with a trait (disease) through the use of SNPs by comparing allele frequencies in cases and controls.

Supporting references: [1]

Haplotype block: linear segments of the genome comprising coinherited alleles in the same chromosome.

Supporting references: [74,75]

Homologous recombination: an error-free recombination mechanism that exchanges genetic sequences between homologous loci during meiosis, and utilizes homologous sequences such as the sister-chromatid to promote DNA repair during mitosis.

Supporting references: [75]

Linkage disequilibrium: a nonrandom association between two markers (*e.g.* SNPs), which are typically close to one another due to reduced recombination between them. *Supporting references:* [74,76-78].

MicroRNAs: endogenous short (~23 nt) RNAs involved in gene regulation by pairing to mRNAs of protein coding mRNAs.

Supporting references: [79]

Next gen sequencing: a technology to sequence DNA in a massively parallel fashion, therefore sequencing is achieved at a much faster speed and lower cost than traditional methods.

Supporting references: [72]

Non-coding variant: a variant that is located outside of the coding region of a certain locus.

Tagging variant: a variant (SNP) that defines most of the haplotype diversity of a haplotype block.

Supporting references: [74]

Transcriptome: The complete set of transcripts in a cell. In some cases it can also include quantitative data about the amount of individual transcripts.

Supporting references: [69]

RNA-Seq: a method to obtain genome-wide transcription map using deep sequencing technologies to generate short sequence reads (30-400 bp). It reveals a transcriptional profile and levels of expression for each gene.

Supporting references: [69]

SNP: single nucleotide polymorphism

Figure Legend

Figure 1. The genomic context in which a variant is found can be used as preliminary functional analysis. The figure shows the genomic context distribution of disease- and trait-associated SNPs annotated in the Catalog of Genome-Wide Association Studies (http://www.genome.gov/gwastudies/) as of December 9th, 2010. Most of the SNPs are located in intergenic and intronic positions, but a small percentage are located upstream and downstream of genes, as well as in regulatory regions and splice sites. SNPs in these locations can be analyzed in more detail using more specific bioinformatics tools.

Supplementary Information

Supplementary Table 1. Summary of more than 80 susceptibility loci identified for breast, lung, ovarian, colorectal and prostate cancer. The table shows the diversity in genetic architecture of these loci with respect to the most significantly associated SNP at each region, the proximity of the risk associated SNPs with respect to nearest annotated genes, the known gene content within 1MB spanning each region, and details of plausible candidate genes within those regions.

Supplementary Table 2. Methods for functional validation/enabling technologies

References

1. Manolio, T.A. Genomewide association studies and assessment of the risk of disease. *N Engl J Med* **363**, 166-76 (2010).
2. Hindorff LA, J.H., Hall PN, Mehta JP, and Manolio TA. A Catalog of Published Genome-Wide Association Studies. . Vol. 2010.
3. Easton, D.F. et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087-1093 (2007).
4. Hindorff, L.A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**, 9362-7 (2009).
5. Jia, L. et al. Functional enhancers at the gene-poor 8q24 cancer-linked locus. *PLoS Genet* **5**, e1000597 (2009).
6. On beyond GWAS. *Nat Genet* **42**, 551 (2010).
7. Glazier, A.M., Nadeau, J.H. & Aitman, T.J. Finding genes that underlie complex traits. *Science* **298**, 2345-9 (2002).
8. Udler, M.S. et al. FGFR2 variants and breast cancer risk: fine-scale mapping using African American studies and analysis of chromatin conformation. *Hum Mol Genet* **18**, 1692-703 (2009).
9. Via, M., Gignoux, C. & Burchard, E.G. The 1000 Genomes Project: new opportunities for research and social challenges. *Genome Med* **2**, 3.
10. Saccone, N.L. et al. In search of causal variants: refining disease association signals using cross-population contrasts. *BMC Genet* **9**, 58 (2008).
11. Heintzman, N.D. et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108-12 (2009).
12. Visel, A. et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854-8 (2009).
13. Visel, A., Rubin, E.M. & Pennacchio, L.A. Genomic views of distant-acting enhancers. *Nature* **461**, 199-205 (2009).
14. Birney, E. et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816 (2007).
15. Blow, M.J. et al. ChIP-Seq identification of weakly conserved heart enhancers. *Nat Genet* **42**, 806-10 (2010).
16. Kunarso, G. et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* **42**, 631-4 (2010).
17. Coetzee, G.A. et al. A systematic approach to understand the functional consequences of non-protein coding risk regions. *Cell Cycle* **9**, 47-51 (2010).
18. Pomerantz, M.M. et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet* **41**, 882-4 (2009).
19. Ahmadiyeh, N. et al. 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. *Proc Natl Acad Sci U S A* **107**, 9742-6 (2010).
20. Wasserman, N.F., Aneas, I. & Nobrega, M.A. An 8q24 gene desert variant associated with prostate cancer risk confers differential in vivo activity to a MYC enhancer. *Genome Res* (2010).
21. Gupta, R.A. et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071-6.
22. Khalil, A.M. et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* **106**, 11667-72 (2009).
23. Jones, P.A. & Baylin, S.B. The epigenomics of cancer. *Cell* **128**, 683-92 (2007).
24. Raval, A. et al. Downregulation of death-associated protein kinase 1 (DAPK1) in chronic lymphocytic leukemia. *Cell* **129**, 879-90 (2007).
25. Smith, L.T. et al. Epigenetic regulation of the tumor suppressor gene TCF21 on 6q23-q24 in lung and head and neck cancer. *Proc Natl Acad Sci U S A* **103**, 982-7 (2006).
26. Jirtle, R.L. & Skinner, M.K. Environmental epigenomics and disease susceptibility. *Nat Rev Genet* **8**, 253-62 (2007).

27. Lister, R. et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315-22 (2009).
28. Chan, T.L. et al. Heritable germline epimutation of MSH2 in a family with hereditary nonpolyposis colorectal cancer. *Nat Genet* **38**, 1178-83 (2006).
29. Suter, C.M., Martin, D.I. & Ward, R.L. Germline epimutation of MLH1 in individuals with multiple cancers. *Nat Genet* **36**, 497-501 (2004).
30. Hesson, L.B., Hitchins, M.P. & Ward, R.L. Epimutations and cancer predisposition: importance and mechanisms. *Curr Opin Genet Dev* **20**, 290-8 (2010).
31. Kerkel, K. et al. Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. *Nat.Genet.* **40**, 904-908 (2008).
32. Pelletier, C. & Weidhaas, J.B. MicroRNA binding site polymorphisms as biomarkers of cancer risk. *Expert Rev Mol Diagn* **10**, 817-29 (2010).
33. Gemayel, R., Vinces, M.D., Legendre, M. & Verstrepen, K.J. Variable Tandem Repeats Accelerate Evolution of Coding and Regulatory Sequences. *Annu Rev Genet* (2010).
34. Zhao, Z. et al. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet* **38**, 1341-7 (2006).
35. Sandhu, K.S. et al. Nonallelic transvection of multiple imprinted loci is organized by the H19 imprinting control region during germline development. *Genes Dev* **23**, 2598-603 (2009).
36. Steidl, U. et al. A distal single nucleotide polymorphism alters long-range regulation of the PU.1 gene in acute myeloid leukemia. *J Clin Invest* **117**, 2611-20 (2007).
37. Blaydon, D.C. et al. The gene encoding R-spondin 4 (RSPO4), a secreted protein implicated in Wnt signaling, is mutated in inherited anonychia. *Nat Genet* **38**, 1245-7 (2006).
38. Kelsell, D.P. et al. Mutations in ABCA12 underlie the severe congenital skin disease harlequin ichthyosis. *Am J Hum Genet* **76**, 794-803 (2005).
39. Nicolae, D.L. et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* **6**, e1000888.
40. Moffatt, M.F. et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* **448**, 470-3 (2007).
41. Musunuru, K. et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714-9 (2010).
42. Zhong, H. et al. Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. *PLoS Genet* **6**, e1000932 (2010).
43. Pomerantz, M.M. et al. Analysis of the 10q11 cancer risk locus implicates MSMB and NCOA4 in human prostate tumorigenesis. *PLoS Genet* **6**, e1001204.
44. Monks, S.A. et al. Genetic inheritance of gene expression in human cell lines. *Am J Hum Genet* **75**, 1094-105 (2004).
45. Morley, M. et al. Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743-7 (2004).
46. Stranger, B.E. et al. Genome-wide associations of gene expression variation in humans. *PLoS Genet* **1**, e78 (2005).
47. Schadt, E.E. et al. Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297-302 (2003).
48. Johnson, J.M. et al. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302**, 2141-4 (2003).
49. Rockman, M.V. & Kruglyak, L. Genetics of global gene expression. *Nat Rev Genet* **7**, 862-72 (2006).
50. Cheung, V.G. & Spielman, R.S. Genetics of human gene expression: mapping DNA variants that influence gene expression. *Nat Rev Genet* **10**, 595-604 (2009).
51. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M. Mapping complex disease traits with global gene expression. *Nat Rev Genet* **10**, 184-94 (2009).
52. Schadt, E.E. et al. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol* **6**, e107 (2008).
53. Pastinen, T. Genome-wide allele-specific analysis: insights into regulatory variation. *Nat Rev Genet* **11**, 533-8.

54. Montgomery, S.B. et al. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773-7 (2010).
55. Pickrell, J.K. et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768-72 (2010).
56. Margolin, A.A. et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* **7 Suppl 1**, S7 (2006).
57. Bumgarner, R.E. & Yeung, K.Y. Methods for the inference of biological pathways and networks. *Methods Mol Biol* **541**, 225-45 (2009).
58. Furuta, S. et al. Depletion of BRCA1 impairs differentiation but enhances proliferation of mammary epithelial cells. *Proc Natl Acad Sci U S A* **102**, 9176-81 (2005).
59. Proia, T.A. et al. Genetic predisposition directs breast cancer phenotype by dictating progenitor cell fate. *Cell Stem Cell* **8**, 149-63.
60. Dekker, J. The three 'C' s of chromosome conformation capture: controls, controls, controls. *Nat Methods* **3**, 17-21 (2006).
61. Simonis, M. et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat Genet* **38**, 1348-54 (2006).
62. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-93 (2009).
63. Todd, J.A. Statistical false positive or true disease pathway? *Nat Genet* **38**, 731-3 (2006).
64. Mardis, E.R. ChIP-seq: welcome to the new frontier. *Nat Methods* **4**, 613-4 (2007).
65. Robertson, G. et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* **4**, 651-7 (2007).
66. Iafrate, A.J. et al. Detection of large-scale variation in the human genome. *Nat Genet.* **36**, 949-951 (2004).
67. Lee, C., Iafrate, A.J. & Brothman, A.R. Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nat Genet* (2007).
68. Sebat, J. et al. Large-Scale Copy Number Polymorphism in the Human Genome. *Science* **305**, 525-528 (2004).
69. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57-63 (2009).
70. Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet* **33 Suppl**, 245-54 (2003).
71. Garner, M.M. & Revzin, A. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic Acids Res* **9**, 3047-60 (1981).
72. Chi, P., Allis, C.D. & Wang, G.G. Covalent histone modifications--miswritten, misinterpreted and mis-erased in human cancers. *Nat Rev Cancer* **10**, 457-69 (2010).
73. Ioannidis, J.P. Common genetic variants for breast cancer: 32 largely refuted candidates and larger prospects. *J Natl.Cancer Inst.* **98**, 1350-1353 (2006).
74. Jobling, M.A., Hurles, M. & Tyler-Smith, C. *Human evolutionary genetics : origins, peoples & disease*, xx, 523 p. (Garland Science, New York, 2004).
75. Griffiths, A.J.F. et al. *Introduction to Genetic Analysis*, (W.H. Freeman and Company, New York, 2005).
76. Altshuler, D., Daly, M.J. & Lander, E.S. Genetic mapping in human disease. *Science* **322**, 881-8 (2008).
77. Lackie, J.M., Dow, J.A.T. & Blackshaw, S.E. *The dictionary of cell biology*, 390 p. (Academic Press, London ; San Diego, 1995).
78. Reich, D.E. et al. Linkage disequilibrium in the human genome. *Nature* **411**, 199-204 (2001).
79. Bartel, D.P. MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215-33 (2009).

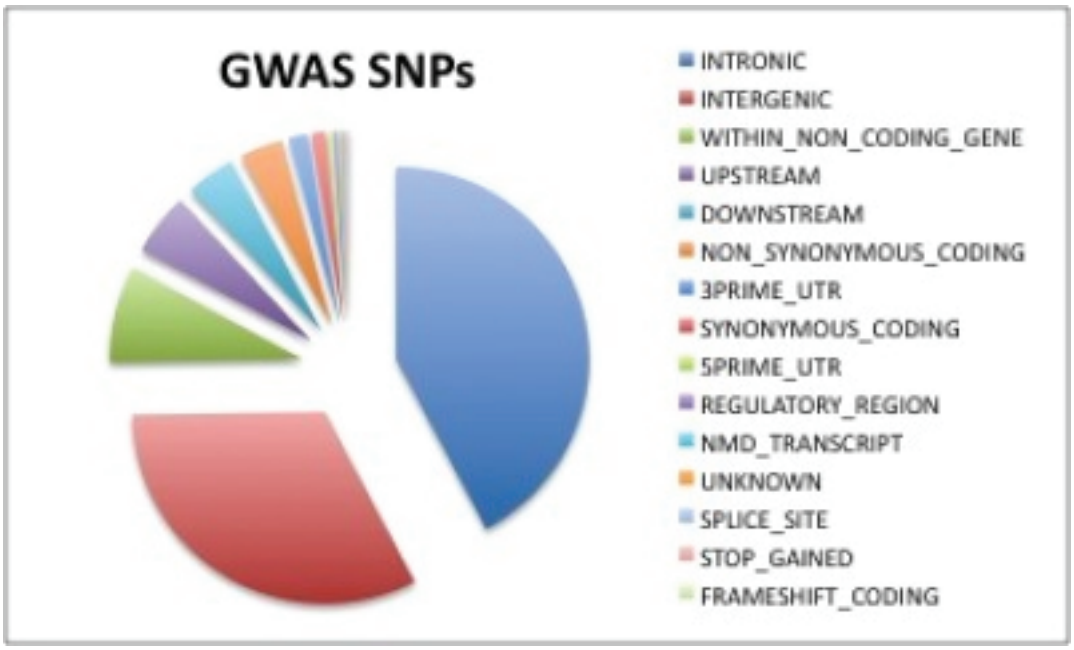
GWAS SNPs
INTRONIC
INTERGENIC
WITHIN_NON_CODING_GENE
UPSTREAM
DOWNSTREAM
NON_SYNONYMOUS_CODING
3PRIME_UTR
SYNONYMOUS_CODING
5PRIME_UTR
REGULATORY_REGION
NMD_TRANSCRIPT
UNKNOWN
SPLICE_SITE
STOP_GAINED
FRAMESHIFT_CODING

Figure 1