Published in final edited form as:
Nat Genet. 2011 June ; 43(6): 513–518. doi:10.1038/ng.840.

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

Matthew L Freedman[1,2], **Alvaro N A Monteiro**[3], **Simon A Gayther**[4,5], **Gerhard A Coetzee**[6], **Angela Risch**[7], **Christoph Plass**[7], **Graham Casey**[8], **Mariella De Biasi**[9], **Chris Carlson**[10], **David Duggan**[11], **Michael James**[12], **Pengyuan Liu**[12], **Jay W Tichelaar**[12], **Haris G Vikis**[12], **Ming You**[12], and **Ian G Mills**[13,14]

[1]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA

[2]The Eli and Edythe L. Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA

[3]Cancer Epidemiology Program, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida, USA

[4]University of Southern California, Keck School of Medicine, Los Angeles, California, USA

[5]Translational Research Laboratory, University College London Elizabeth Garrett Anderson Institute for Women's Health (EGA), London, UK

[6]Department of Urology, Norris Cancer Center, University of Southern California, Los Angeles, California, USA

[7]German Cancer Research Center, Division of Epigenomics and Cancer Risk Factors, Heidelberg, Germany

[8]Department of Preventive Medicine, University of Southern California, Los Angeles, California, USA

[9]Department of Neuroscience, Baylor College of Medicine, Houston, Texas, USA

[10]Fred Hutchinson Cancer Research Center, Seattle, Washington, USA

[11]Translational Genomics Research Institute (TGen), Phoenix, Arizona, USA

[12]Medical College of Wisconsin, Milwaukee, Wisconsin, USA

[13]Norwegian Centre for Molecular Medicine, Nordic European Molecular Biology Laboratory (EMBL) Partnership, University of Oslo, Oslo, Norway

[14]Uro-Oncology Research Group, Cancer Research UK Cambridge Research Institute, Cambridge, UK

Genome wide association studies (GWAS) have identified more than 200 mostly new common low-penetrance susceptibility loci for cancers. The predicted risk associated with each locus is generally modest (with a per-allele odds ratio typically less than 2) and so,

Correspondence should be addressed to I.G.M. (ian.mills@ncmm.uio.no).

AUTHOR CONTRIBUTIONS

All contributed collectively to the writing and drafting of the manuscript, and this paper was coordinated by I.G.M.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

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 of these loci. Biological insights can then be translated to clinical benefits, including reliable biomarkers and effective strategies for screening and disease prevention. The purpose of this article is to propose principles for the initial functional characterization of cancer risk loci, with a focus on non-coding variants, and to define 'post-GWAS' functional characterization.

By December 2010, there were 1,212 published GWAS studies[1] reporting significant ($P < 5 \times 10^{-8}$) associations for 210 traits (Table 1), and the Catalog of Published GWAS states that by March 2011, 812 publications reported 3,977 SNP associations[1]. This is likely a small fraction of the common susceptibility loci of low penetrance that will eventually be identified. Despite these successes in identifying risk loci, the causal variant and/or the molecular basis of risk etiology has been determined for only a small fraction of these associations[2–4]. Plausible candidate genes can be based on proximity to risk loci, but few have so far been defined in a more systematic manner (Supplementary Table 1).

Increased investment in post-GWAS functional characterization of risk loci[5] has now been advocated across diseases and for cardiovascular disease and diabetes[6]. For cancer biology, the complex interplay between genetics and the environment in many cancers poses a particularly exciting challenge for post-GWAS research. Here we suggest a systematic strategy for understanding how cancer-associated variants exert their effects. We mostly refer to SNPs throughout the paper, but we recognize that other types of common genetic (for example, copy number variants) or epigenetic variation may influence risk.

Our understanding of the way in which a risk variant initiates disease pathogenesis progresses from statistical association between genetic variation and trait or disease variation to functionality and causality. The functional consequences of variants in protein-coding regions causing most monogenic disorders are more readily interpreted because we know the genetic code. For non-Mendelian or multifactorial traits, most of the common DNA variants have so far mapped to non-protein–coding regions[2], where our understanding of functional consequences and causality is more rudimentary.

Our hypothesis is that the trait-associated alleles exert their effects by influencing transcriptional output (such as transcript levels and splicing) through multiple mechanisms. We emphasize appropriate assays and models 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 modeling gene effects will emphasize measuring cancer-related phenotypes.

At some loci, multiple, independently associated risk alleles rather than single risk alleles may be functionally responsible for the occurrence of disease. Genotyping susceptibility loci (and their correlated variants) in multiple populations with different linkage disequilibrium (LD) structures may prove effective in substantially reducing the number of potentially causative variants (that is, the same causal variant may segregate in multiple populations), as shown for the *FGFR2* locus in breast cancer[7], 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.

A susceptibility locus should be re-sequenced to ascertain all genetic variation, identifying candidate functional or causal variants 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, although we propose a framework set of

analyses (Box 1), we acknowledge that the techniques and methods will continue to evolve with the field.

Box 1

Strategies to progress from tag SNP to mechanism

1. Target resequencing efforts using linkage disequilibrium (LD) structure.
2. Use other populations to refine LD regions (for example African ancestry with shorter LD and more heterogeneity).
3. Determine expression levels of nearby genes as a function of genotype at each locus (eQTL).
4. Characterize gene regulatory regions by multiple empirical techniques bearing in mind that these are tissue and context specific.
5. Combine regulatory regions with risk loci using coordinates from multiple reference genomes to capture all variation within the shorter regulatory regions that correlates with the tag SNP at each locus.
6. Multiple experimental manipulations in model systems are needed to progressively implicate transcription units (genes) in mechanisms relevant to the associated loci:
    i. Knockouts of regulatory regions in animal (difficult and may be limited by functional redundancy, but new targeting methods in rat are promising) models followed by genome-wide expression analysis.
    ii. Use chromatin association methods (3C, CHIA-PET) of regulatory regions to determine the identity of target genes (compare with eQTL data).
    iii. Targeted gene perturbations in somatic cell models.
    iv. Explore fully genome-wide eQTL and miRNA quantitative variation correlation in relevant tissues and cells.
7. Explore epigenetic mechanisms in the context of genome-wide genetic polymorphism.
8. Employ cell models and tissue reconstructions to evaluate mechanisms using gene perturbations and polymorphic variants. The human cancer cell xenograft has re-emerged as a minimal *in vivo* validation of these models.
9. Above all, resist the temptation to equate any partial functional evidence as sufficient. Published claims of functional relevance should be fully evaluated using the steps detailed above.

Fine mapping

Most GWAS identify an association between the disease trait and a surrogate marker (tag SNP) rather than a causal variant because SNP arrays were designed using SNPs chosen to capture LD structure rather than functional variants. To get to the underlying biology, a comprehensive understanding of the genetic variation of the associated regions will be necessary, starting with the most common SNPs.

The ongoing 1000 Genomes Project seeks to capture common (>5%) and less common (1–5%) variant information[8] in diverse ethnic populations using a combination of low-coverage whole-genome sequencing and deeper coverage exome sequencing. However, it remains to be determined whether it provides complete SNP coverage across the entire genome, including intergenic regions and gene deserts where the majority of the GWAS associations have been mapped. This suggests that for at least some loci, targeted sequencing remains a necessity.

The goal of targeted sequencing is to capture the causal SNP(s) that is in LD with the associated SNP(s) (assuming that the causal SNP is not the associated SNP). The likelihood of identifying the causal SNP will be affected by both how the boundaries of the region to be sequenced are defined as well as the depth of sequence coverage across the region.

The region to be sequenced can be guided by LD structure, but there are challenges to this approach, as the strength of the correlation between the associated SNP and the causal SNP may be low, suggesting that a correlation r^2 threshold value of 0.2 or even less may be needed. Incorporating GWAS information from non-European populations, such as those of African descent, 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[9]. Alternatively, LD structure can be ignored and arbitrary physical limits can be set to define boundaries, for example, by choosing to sequence 1 Mb across the risk allele. The region can be further narrowed through incorporation of biological information for the presence of a compelling candidate gene or transcript. However, note that relying on biological assumptions undermines the agnostic approach of GWAS.

The depth of coverage and the number of subjects to be sequenced are important considerations. Current targeted enrichment technologies yield non-uniform sequencing coverage, which could increase the heterozygote false-negative rate. Sequencing coverage of 25× or greater may be required, especially if sequencing-based genotyping and not just variant discovery is a goal. The likelihood of identifying less common variants is also dependent upon the number of subjects sequenced, and often DNA from several hundreds of subjects is needed. In summary, because of the fact that both the size of the region and the number of individuals to be sequenced influence cost, the final design will likely be a compromise. Costs can be offset to some extent with the use of molecular barcoding, when individual genotypes are important, and DNA pooling, when variant discovery is important.

Annotating variable regulatory elements

Characterizing the regulatory landscape of susceptibility regions is 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[10] 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 in annotating tissue-specific regulatory elements[11,12] and is more informative than studying sequence conservation, as regulatory elements may be unconstrained across mammalian evolution[13–15]. Using chromatin annotations to identify putative functional SNPs within regulatory sequences at known susceptibility loci has recently been proposed[16]. 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 chromatin immunoprecipitation sequencing (ChIP-Seq) methodologies, and signals yield short DNA stretches (typically <1 kb) amenable to detailed analyses. Enhancer activity in such regulatory regions can be assayed using reporter genes *in vitro*[4] and/or *in vivo*[11].

Integrating knowledge of regulatory sequences at risk loci with catalogs 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 have been associated with prostate and colorectal cancer, respectively. 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 SNP within a FoxA1 binding site[4]. 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[17]. 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 (for example, ChIP or electrophoretic mobility shift assay (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: first, knocking out regulatory sequences in mouse models followed by genome-wide gene expression analyses after knockout to identify candidate targets; second, using the regulatory sequences as baits in chromatin conformation capture-based studies[18,19], including genome-wide chromatin conformation capture-based methods; third, 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 (for example, by radiation or hormones); and finally, 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.

Epigenetic regulation of gene expression

Promoter methylation, histone tail modifications and altered expression of non-coding RNAs, such as the large intergenic noncoding RNAs (lincRNAs)[20,21], which associate with chromatin-modifying complexes, also contribute to gene regulation and are obvious candidate targets of functional genetic associations[22]. Epigenetic silencing has been shown to be the predominant mechanism of gene silencing during tumor development for a subset of genes[23]. For other genes, a combination of genetic and epigenetic mechanisms can contribute to tumor suppressor gene activation[24]. Epigenetic mechanisms also play an important role in mediating environmental influences on gene expression[25]. At susceptibility loci, the key questions are: first, do common genetic variants influence the epigenetic landscape to increase disease susceptibility, and second, 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[26]. Studies of hereditary non-polyposis colorectal cancer (Lynch syndrome) suggest that germline genetic variation may affect epigenetic marks, resulting in cancer predisposition[27,28]. These changes in CpG methylation may be a consequence of *cis*- or *trans*-acting genetic variants[29]. For example, Kerkel *et al.* have shown sequence-dependent allele-specific methylation and that *cis*-regulatory variants control gene expression and affect chromatin states[30]. Further epigenetic mechanisms that modulate gene expression include microRNAs (miRNAs) and miRNA binding sites, which can be directly affected by SNPs[31], and tandem repeats that can impact gene expression by, for example, altering transcription factor binding sites but also by affecting chromatin structure (reviewed in fc ref. 32).

Risk SNPs may also be tagging variants affecting the chromatin regulation of the nucleus. Chromatin fibers dynamically explore the nuclear space to establish meta-stable, long-range interactions with other chromatin fibers[33]. The functional outcome of such interactions is largely unknown, but it has been shown that they are capable of transferring epigenetic marks to modulate transcriptional processes both in *cis*[34] and in *trans*[35,36]. 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[33]. Sequence variants can influence communication between different parts of the genome[37], 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 or chromosomal networks may therefore provide important insights into the function of germline genetic variants.

Inherited variation and gene expression

Both empirical and computational data support the notion that a considerable proportion of trait-associated loci will harbor variants that influence the abundance of specific transcripts. These variants are often referred to as expression quantitative trait loci (eQTLs)[38–42]. Several landmark studies have unequivocally shown that many transcripts in the human genome are influenced by inherited variation[43–47]. 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. Importantly, and as is the case in 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 at considerable distances away from it. The distinction between local and distant is often arbitrary, however, as in most studies, local has often been defined as being within 1 Mb 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 connote mechanism[48]. It should be noted that not only mRNA transcripts but also miRNA and non-coding RNA (ncRNA) transcripts should be considered as candidates.

Certain principles have emerged from eQTL studies: first, 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). Second, local eQTLs tend to have larger effects on gene expression than distant eQTLs and are therefore easier to discover. Third, there are likely to be a larger number of distant than local eQTLs[49].

Many of the initial successful eQTL studies relied on available lymphoblastoid cell lines[39,50]. More recently, eQTL studies have been performed in primary human tissues and have shown that at least some associations are tissue specific[40,42,51]. Although large sample sizes are needed in order to achieve sufficient power to detect eQTL associations, they are typically smaller than those used in GWAS to identify risk alleles. Consequently, comprehensive biobanks 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 major 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 US National Institutes of Health Common Fund.

A complementary and powerful approach to defining local eQTLs is to measure allelic imbalance (also called allele-specific gene expression) in individuals that are heterozygous for a risk allele. Any transcript with a deviation from a 1:1 ratio (as typically measured by a transcribed heterozygous marker) becomes a strong candidate gene[52–54]. 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 show that a gene is directly involved with disease development.

False negatives (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. Also, 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 or 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: first, defining the appropriate target tissues to examine. Risk alleles may act in a non-cell or non-tissue autonomous fashion and therefore may exert their effect through other cell types that act upon the target tissue under consideration. Second, defining the importance of eQTL analysis in tumor 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. Third, using 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 network dependency using tools such as ARACNE[55] to Bayesian network approaches to identify predictive relationships between genes from a combination of expression and eQTL data[56]. Although these tools are elegant, the ability to translate their outputs into biological importance is heavily dependent on the availability of manipulable and relevant model systems with which to test the predicted connectivity. These and many other approaches clearly pose validation challenges for many diseases, however, the field of computational biology is a powerful and essential catalyst for post-GWAS studies.

Cell and tissue models

Once there is sufficient evidence in support of a candidate susceptibility gene, more detailed functional studies will be required to characterize 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 importance 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 *in vitro* cancer models are the most accessible way to test the function of candidate genes at susceptibility loci in tumor 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 or 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 are good models of normal breast tissue (for example, MCF10A cells and immortalized HMECs). Three-dimensional 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[57]; this link has been further highlighted by Proia *et al.*[58]. By using such three-dimensional 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 small interfering RNA or short hairpin RNA knockdown.

The GWAS community has arrived at an important crossroads. As resources are limited and as the variants found so far operate within their genotypic context, the debate revolves around whether enough progress has been made toward identifying the variants that are likely to contribute most to disease causation to invest in functional follow up. As sequencing technologies become cheaper and more accessible and 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 show 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 or multi-group level, just as effective international collaborations led to the identification of 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 populations (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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank F. Bunz and all the members of the NIH Post-Genome Wide Association Initiative for helpful discussions and, in particular, I. 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 (http://grants.nih.gov./grants/oer.htm). 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: (http://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 R. 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).

NAMED COLLABORATORS

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, Trentino, Italy), Dan Bolser (Dundee University, Dundee, UK), David F. Burke (Department of Zoology, University of Cambridge, Cambridge UK), Yari Ciribilli (CIBIO, University of Trento, Trentino, Italy), Lucia Conde (Environmental Health Sciences, University of California Berkeley, Berkeley, California, USA), Giovanni Marco Dall'Olio (Institut de Biologia Evolutiva (CSIC-UPF), Barcelona, Spain), Doug Easton (CR-UK Genetic Epidemiology Unit, University of Cambridge, Cambridge, UK), Rosalind Eeles (Translational Cancer Genetics Team, The Institute of Cancer Research and Royal Marsden National Health Service (NHS) Foundation Trust, London, UK), Johannes Engelken (Instituto de Biologia Evolutiva (CSIC), Barcelona, Spain), Marta Ramirez Gaite (WikiGenes), Evgeny A. Glazov (Diamantina Institute, The University of Queensland, Brisbane, Queensland, Australia), Jeremy Leighton John (The British Library, London, UK), Kevin L. Keys (Universitat Pompeu Fabra, Barcelona,

Spain), Anchit Khanna (Institute of Medical Technology, University of Tampere and University Hospital, Tampere, Finland), Georgios D. Kitsios (Institute for Clinical Research and Health Policy Studies, Boston, Massachusetts, USA), S. Lillioja (Illawarra Health and Medical Research Institute, University of Wollongong, Wollongong, New South Wales, Australia), Mary Mangan (OpenHelix LLC, Belleview, Washington, USA; Vancouver, British Columbia, Canada), Christopher Maxwell (Child and Family Research Institute, University of British Columbia, Vancouver, British Columbia, Canada), Sumit Middha (Mayo Clinic, Biomedical Informatics and Statistics, Rochester, Minnesota, USA), Pooja Mohan (University of British Columbia, Vancouver, British Columbia, Canada), Paulo Nuin (Queen's University and Ontario Cancer Biomarker Network, Kingston, Ontario, Canada), Rolf Ohlsson (Department of Microbiology, Tumor and Cell Biology, Karolinska Institute, Stockholm, Sweden), Mingxiong Pang (Section of Molecular Cell and Developmental Biology, The University of Texas at Austin, Austin, Texas, USA), Chilakamarti V. Ramana (Department of Medicine, Dartmouth Medical School, Hanover, New Hampshire, USA), Amiya Sarkar (B.S. Medical College, West Bengal Medical Education Service, Bankura, India), Khader Shameer (Mayo Clinic, Rochester, Minnesota, USA), Christine F Skibola (School of Public Health, University of California Berkeley, Berkeley, California, USA), Rayna Stamboliyska (Evolutionary Biology, Ludwig-Maximilians Universität (LMU), Munich, Germany), Muy-Teck (Barts & the London School of Medicine & Dentistry, Queen Mary University of London, London, UK), Tao Zhang (The Research Institute for Children, Children's Hospital, New Orleans, Louisiana, 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). We hope that this will be the first of many valuable examples of increased engagement with scientists through these avenues.

References

1. Hindorff LA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009; 106:9362–9367. [PubMed: 19474294]
2. Easton DF, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature. 2007; 447:1087–1093. [PubMed: 17529967]
3. Hindorff LA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA. 2009; 106:9362–9367. [PubMed: 19474294]
4. Jia L, et al. Functional enhancers at the gene-poor 8q24 cancer-linked locus. PLoS Genet. 2009; 5:e1000597. [PubMed: 19680443]
5. On beyond GWAS. Nat Genet. 2010; 42:551. [PubMed: 20581872]
6. Glazier AM, Nadeau JH, Aitman TJ. Finding genes that underlie complex traits. Science. 2002; 298:2345–2349. [PubMed: 12493905]
7. Udler MS, et al. *FGFR2* variants and breast cancer risk: fine-scale mapping using African American studies and analysis of chromatin conformation. Hum Mol Genet. 2009; 18:1692–1703. [PubMed: 19223389]
8. Via M, Gignoux C, Burchard EG. The 1000 Genomes Project: new opportunities for research and social challenges. Genome Med. 2010; 2:3. [PubMed: 20193048]
9. Saccone NL, et al. In search of causal variants: refining disease association signals using cross-population contrasts. BMC Genet. 2008; 9:58. [PubMed: 18759969]
10. Heintzman ND, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature. 2009; 459:108–112. [PubMed: 19295514]
11. Visel A, et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature. 2009; 457:854–858. [PubMed: 19212405]

12. Visel A, Rubin EM, Pennacchio LA. Genomic views of distant-acting enhancers. Nature. 2009; 461:199–205. [PubMed: 19741700]
13. Birney E, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007; 447:799–816. [PubMed: 17571346]
14. Blow MJ, et al. ChIP-Seq identification of weakly conserved heart enhancers. Nat Genet. 2010; 42:806–810. [PubMed: 20729851]
15. Kunarso G, et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet. 2010; 42:631–634. [PubMed: 20526341]
16. Coetzee GA, et al. A systematic approach to understand the functional consequences of non-protein coding risk regions. Cell Cycle. 2010; 9:47–51.
17. Pomerantz MM, et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nat Genet. 2009; 41:882–884. [PubMed: 19561607]
18. Ahmadiyeh N, et al. 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. Proc Natl Acad Sci USA. 2010; 107:9742–9746. [PubMed: 20453196]
19. Wasserman NF, Aneas I, Nobrega MA. An 8q24 gene desert variant associated with prostate cancer risk confers differential in vivo activity to a MYC enhancer. Genome Res. 2010; 20:1191–1197. [PubMed: 20627891]
20. Gupta RA, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature. 2010; 464:1071–1076. [PubMed: 20393566]
21. Khalil AM, et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci USA. 2009; 106:11667–11672. [PubMed: 19571010]
22. Jones PA, Baylin SB. The epigenomics of cancer. Cell. 2007; 128:683–692. [PubMed: 17320506]
23. Raval A, et al. Downregulation of death-associated protein kinase 1 (DAPK1) in chronic lymphocytic leukemia. Cell. 2007; 129:879–890. [PubMed: 17540169]
24. Smith LT, et al. Epigenetic regulation of the tumor suppressor gene *TCF21* on 6q23-q24 in lung and head and neck cancer. Proc Natl Acad Sci USA. 2006; 103:982–987. [PubMed: 16415157]
25. Jirtle RL, Skinner MK. Environmental epigenomics and disease susceptibility. Nat Rev Genet. 2007; 8:253–262. [PubMed: 17363974]
26. Lister R, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009; 462:315–322. [PubMed: 19829295]
27. Chan TL, et al. Heritable germline epimutation of *MSH2* in a family with hereditary nonpolyposis colorectal cancer. Nat Genet. 2006; 38:1178–1183. [PubMed: 16951683]
28. Suter CM, Martin DI, Ward RL. Germline epimutation of *MLH1* in individuals with multiple cancers. Nat Genet. 2004; 36:497–501. [PubMed: 15064764]
29. Hesson LB, Hitchins MP, Ward RL. Epimutations and cancer predisposition: importance and mechanisms. Curr Opin Genet Dev. 2010; 20:290–298. [PubMed: 20359882]
30. Kerkel K, et al. Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. Nat Genet. 2008; 40:904–908. [PubMed: 18568024]
31. Pelletier C, Weidhaas JB. MicroRNA binding site polymorphisms as biomarkers of cancer risk. Expert Rev Mol Diagn. 2010; 10:817–829. [PubMed: 20843204]
32. Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu Rev Genet. 2010; 44:445–477. [PubMed: 20809801]
33. Zhao Z, et al. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet. 2006; 38:1341–1347. [PubMed: 17033624]
34. Sandhu KS, et al. Nonallelic transvection of multiple imprinted loci is organized by the H19 imprinting control region during germline development. Genes Dev. 2009; 23:2598–2603. [PubMed: 19933149]
35. 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. 2007; 117:2611–2620. [PubMed: 17694175]

36. Blaydon DC, et al. The gene encoding R-spondin 4 (RSPO4), a secreted protein implicated in Wnt signaling, is mutated in inherited anonychia. Nat Genet. 2006; 38:1245–1247. [PubMed: 17041604]
37. Kelsell DP, et al. Mutations in *ABCA12* underlie the severe congenital skin disease harlequin ichthyosis. Am J Hum Genet. 2005; 76:794–803. [PubMed: 15756637]
38. Nicolae DL, et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010; 6:e1000888. [PubMed: 20369019]
39. Moffatt MF, et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature. 2007; 448:470–473. [PubMed: 17611496]
40. Musunuru K, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature. 2010; 466:714–719. [PubMed: 20686566]
41. Zhong H, et al. Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. PLoS Genet. 2010; 6:e1000932. [PubMed: 20463879]
42. Pomerantz MM, et al. Analysis of the 10q11 cancer risk locus implicates MSMB and NCOA4 in human prostate tumorigenesis. PLoS Genet. 2010; 6:e1001204. [PubMed: 21085629]
43. Monks SA, et al. Genetic inheritance of gene expression in human cell lines. Am J Hum Genet. 2004; 75:1094–1105. [PubMed: 15514893]
44. Morley M, et al. Genetic analysis of genome-wide variation in human gene expression. Nature. 2004; 430:743–747. [PubMed: 15269782]
45. Stranger BE, et al. Genome-wide associations of gene expression variation in humans. PLoS Genet. 2005; 1:e78. [PubMed: 16362079]
46. Schadt EE, et al. Genetics of gene expression surveyed in maize, mouse and man. Nature. 2003; 422:297–302. [PubMed: 12646919]
47. Johnson JM, et al. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science. 2003; 302:2141–2144. [PubMed: 14684825]
48. Rockman MV, Kruglyak L. Genetics of global gene expression. Nat Rev Genet. 2006; 7:862–872. [PubMed: 17047685]
49. Cheung VG, Spielman RS. Genetics of human gene expression: mapping DNA variants that influence gene expression. Nat Rev Genet. 2009; 10:595–604. [PubMed: 19636342]
50. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. Nat Rev Genet. 2009; 10:184–194. [PubMed: 19223927]
51. Schadt EE, et al. Mapping the genetic architecture of gene expression in human liver. PLoS Biol. 2008; 6:e107. [PubMed: 18462017]
52. Pastinen T. Genome-wide allele-specific analysis: insights into regulatory variation. Nat Rev Genet. 2010; 11:533–538. [PubMed: 20567245]
53. Montgomery SB, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. 2010; 464:773–777. [PubMed: 20220756]
54. Pickrell JK, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010; 464:768–772. [PubMed: 20220758]
55. Margolin AA, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics. 2006; 7 (Suppl 1):S7. [PubMed: 16723010]
56. Bumgarner RE, Yeung KY. Methods for the inference of biological pathways and networks. Methods Mol Biol. 2009; 541:225–245. [PubMed: 19381545]
57. Furuta S, et al. Depletion of BRCA1 impairs differentiation but enhances proliferation of mammary epithelial cells. Proc Natl Acad Sci USA. 2005; 102:9176–9181. [PubMed: 15967981]
58. Proia TA, et al. Genetic predisposition directs breast cancer phenotype by dictating progenitor cell fate. Cell Stem Cell. 2010; 8:149–163. [PubMed: 21295272]

Table 1

The genomic context in which a variant is found can be used as preliminary functional analysis

Classification	Approximate percentages[a]	Approximate numbers[a]
Intronic	40	1,047
Intergenic	32	838
Within non-coding sequence of a gene	10	262
Upstream	8	210
Downstream	4	105
Non-synonymous coding	3	79
3′ untranslated region	~1	26
Synonymous coding	~1	26
5′ untranslated region		
Regulatory region		
Nonsense-mediated decay transcript		
Unknown	~1	26
Splice site		
Gained stop codon		
Frameshift in a coding sequence		

The table broadly summarizes the genomic context 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: 1,212 published genome-wide associations with $P < 5 \times 10^{-8}$ for 210 traits totaling 2,619 SNPs. 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.

[a]Values are indicative and dependent on genomic boundaries used.