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

Alvaro N.A.Monteiro[1], Gerhard A. Coetzee[2], Matthew L. Freedman[3], Mariella De Biasi[4], Graham Casey[5], Dave Duggan[6], Angela Risch[7], Christoph Plass[7], Pengyuan Liu[8], Michael James[8], Haris G. Vikis[8], Jay W. Tichelaar[8], Ming You[8], Simon A. Gayther[5], Ian G. Mills[9*] on behalf of functional cancer genomics supported by the NIH Post-Genome Wide Association Initiative

1. Risk Assessment, Detection, and Intervention Program, H.. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida, 33612
2. Department of Preventive Medicine and Urology, Norris Cancer Center, University of Southern California, Los Angeles, CA 90033
3. The Eli and Edythe L. Broad Institute of MIT and Harvard, Cambridge MA 02142 and Department of Medical Oncology, Dana-Farber Cancer Institute, Boston MA 02115
4. Department of Neuroscience, Baylor College of Medicine, Houston, TX 77030
5. Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA
6. Translational Genomics Research Institute (TGen), Phoenix
7. German Cancer Research Center, Division of Epigenomics and Cancer Risk Factors, Heidelberg, Germany
8. Washington University, St. Louis, Missouri
9. Centre for Molecular Medicine (Norway), Nordic EMBL Partnership, University of Oslo, Blindern N-0317 Oslo, Norway

* To whom correspondence should be addressed. Email: ian.mills@ncmm.uio.no or ian.mills@cancer.org.uk



## Introduction

The last few years have seen an explosion of activity in the identification of common, low-penetrance susceptibility alleles for a range of complex diseases and other traits using genome-wide association studies (GWAS)[1]. As of June 2010 there have been 904 published genome-wide associations (at $p<5x10^{-8}$) for 165 traits[2]. Of these, 193 loci were associated with modified risk of 11 different cancer types[2]. Despite the success in identifying risk associated loci, a causal SNP or the molecular basis of risk etiology has been determined in only a small fraction of these documented associations[3-5].

A recent *Nature Genetics* editorial on post-GWAS analyses began to put this problem in sharper focus and suggested that there should be a more significant investment in functional characterization of identified risk loci[6]. Importantly, the development of reliable biomarkers and effective preventive and therapeutic agents made possible by these discoveries is predicated on a detailed understanding of biological function. The editorial made a distinction between two types of functional analysis. The first comprises a preliminary investigation that includes resequencing, association analysis of all variants within the region of linkage disequilibrium (LD), investigation of risk-associated SNPs as modifiers of monogenic traits, genomic analysis of gene expression in human tissues, screens for somatic mutations on risk haplotypes, and epigenetic analysis of human tissues in the regions of the GWAS SNPs. Similar guidelines have been proposed elsewhere, in the context of loci that underlie complex traits[7].

However, there is second stage of functional analysis geared towards understanding the biological mechanism of risk enhancement and causality. Whereas in-depth analysis cannot be the subject of systematic evaluation due to the functional diversity across loci, the

identification of general hypotheses that can be tested in a systematic way will accelerate analysis. Thus, in this paper we propose principles for the initial functional characterization of cancer risk loci to bridge this information gap.

Several challenges lie ahead in assigning functionality to susceptibility SNPs. For example, most effect sizes are small relative to effects seen in monogenic diseases, with per allele odds ratios usually ranging from 1.15 to 1.3, despite an occasional outlier such as *KITLG* in testicular germ cell cancer (OR=3.08)[8,9]. Thus, the functional effects of SNPs are likely to be subtle. It is unclear whether current molecular biology methods have enough resolution to differentiate such small effects. In addition, we anticipate that it will be difficult to address function for biological effects that are non-cell autonomous, are specific to certain developmental stages, or act at a site distant from the tissue-of-origin of the cancer. The study of such effects might benefit from *in vivo* models. An ultimate goal for any disease is to link genetic variation to causation. This is currently beyond the capability of complex disease research. Our objective here is therefore to provide a set of recommendations to optimize the allocation of effort and resources in order maximize the chances of elucidating the functional contribution of specific loci to the disease phenotype. It has been estimated that 88% of currently identified disease-associated SNP are intronic or intergenic [4].Thus, in this paper we will focus our attention on the analysis of non-coding variants and outline a hierarchical approach for post-GWAS functional studies. Connecting a risk allele to a target gene(s) is particularly tractable intermediate point in this work. The unifying hypotheses that are applicable across these studies are:

1. That there is a transcript (coding or non-coding) that is has not yet been annotated and associates with a risk locus.
2. That the risk locus is a regulatory element affecting the expression of one or more annotated transcribed regions of the genome.

## Defining regional boundaries and the assembly of layered genomic data

The general approach underlying current GWAS is to identify disease association through surrogate SNP markers that capture linkage disequilibrium (LD) relationships across the genome. This approach means that any GWAS data generated using commercial SNP arrays are limited by the depth of genomic coverage and representation of LD structure on those arrays. Unfortunately, recent data suggest that only 60% or less of the common SNP information (>5%) and less than 50% of variants with a frequency of >2.5% are found on SNP arrays used in the majority of GWAS conducted to date when compared to the 1000 Genomes Project[1]. While a new generation of SNP arrays have been developed and address this issue of genome coverage (at least for European populations), the degree (or lack thereof) of common SNP content in GWAS studies has important implications not only for missing "dark matter" signals but also for subsequent post GWAS functional characterization studies of known associations. In the vast majority of cases, any association identified through GWAS would be predicted to be between a surrogate maker (e.g., tagSNP) and the disease trait rather than a surrogate marker and 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. Unfortunately, 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. Whilst the ongoing 1000 Genomes Project seeks to address this missing

data and will capture the majority of common variants, it remains to be determined whether it will provide complete common (>5%) SNP coverage across the entire genome, including intergenic regions and gene deserts where the majority of associations have been mapped. Looking ahead it is also important to note that currently the 1000 Genomes Project will not capture adequate information on rare variants (<1%) and there is a growing interest in the role of multiple rare variants in disease risk, which will require targeted or whole genome sequencing of 1,000s - 10,000s of samples.

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 r2 threshold should be set in defining LD structure as the causal SNP potentially could be in LD with the associated SNP at an r2 of 0.2 or even less. Ideally, the LD structure should be defined using genotyping results from the GWAS population in whom original association was identified. For studies involving populations not studied in the HapMap or pilot 1000 Genomes Project, variant discovery or additional genotyping is likely to have to be undertaken prior to defining LD relationships due to the fact that it is not known how well the available SNP arrays capture the genomic diversity outside of these HapMap populations. Alternatively LD structure does not necessarily need to be considered and boundaries of sequencing could be defined by arbitrary size (such as 1Mb across the associated SNP region) or by taking the most distal and proximal SNPs with r2>0.1. Although somewhat arbitrary these approaches represent a workable start point in the absence of data precisely defining LD structure as a start point.

If planning to make use of next generation sequencing data to define regional boundaries, a second consideration is depth of sequencing coverage. For example, the pilot phase of the 1000 Genomes Project has completed low-coverage (4x) whole genome sequencing on 180 individuals (60 from the each of three populations: African, Asian, and Northern European) and deep-coverage (>30x) whole genome sequencing on 2 trios (African and Northern European; also re-sequenced at low-coverage). In addition, deep-coverage (>30x) exome re-sequencing data will soon be available on hundreds of samples and eventually 1000. Remarkably, even at low-coverage it is expected that the majority of common variants, intra- as well as inter-genic, with a minor allele frequency (MAF) >5%, will be identified. Complete coverage is expected for those common variants, intra- and inter-genic once again, with MAF >10%. These data offer a substantially improved value proposition to genetic variation discovery when compared to genome-wide arrays while at the same time pushing deeper into the minor allele frequency spectrum. That said, considerably deeper coverage will be needed and in more populations if rare variants are to be identified let alone novel hypotheses tested. Pooling of subjects for sequencing can be used to reduce costs but could limit coverage for individual samples in the pool and another genotyping step may need to be introduced to obtain individual level data. 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.

## Genetics of Gene Expression

Having defined the target region, and identified all common (>1-5% depending on depth of sequencing) variants within a risk locus by resequencing and *in silico*, a next step in progressing to functional characterisation is to explore associations between statistically significant SNPs and the expression of genes. Such associations offer a thread from which to build functional validation, although they do not necessarily mean that the genes cause the clinical trait. Gene expression is a heritable trait. Transcript abundance varies in the human population (similar to height and blood pressure) and thus can be considered a trait that is amenable to genetic mapping. A number of landmark studies have unequivocally demonstrated that a substantial fraction of transcripts in the human genome are influenced by inherited variation [11-15]. Genetic variants affecting transcript levels are often referred to as expression quantitative trait loci (eQTLs). eQTLs can be located near the gene they regulate or far away. The distinction between local and distant is often arbitrary; however, for most studies local has usually been defined as being within 1 megabase of the variant under consideration. Distant can involve interactions between an eQTL and a gene located in different non-homologous chromosomes. As has been previously pointed out, we prefer the terminology of local and distant rather than cis- and trans- which connotes mechanism [16].

Certain principles have emerged from these 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 eQTLs and gene expression traits that can be discovered with smaller sample sizes than association studies between inherited variation and 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) expression phenotypes are primarily regulated by distant eQTLs[17].

Closing the gap between genotype and phenotype in complex diseases is proving complicated because, unlike Mendelian disorders, a large fraction of the associated loci are located outside known protein-coding regions. Both empirical and computational data support the notion that a considerable proportion of these loci will be eQTLs[18] [19-21](and Pomerantz *et al.,* PLoS Genetics In Press). The strategy of applying the genetics of gene expression approach offers an appealing and straightforward way to initiate the complicated task of connecting risk variants to their target genes. Importantly, this strategy does not require knowledge of the actual causal allele.

Many of the initial successful eQTL studies relied on lymphoblastoid cell lines largely due to their availability[22],[19]. More recently, eQTL studies have been performed in primary human tissues and demonstrate that some of the associations are tissue-specific [20, 23] and Pomerantz *et al (* PLoS Genetics, In Press). A complementary and powerful approach to defining local eQTLs is to measure allelic imbalance 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. [24-26]

What if the risk allele is *not* associated with the expression trait? False negatives can occur because gene expression varies in time and space; therefore, the developmental time point and/or the cellular population(s) that one is examining may not be appropriate. As a rule of

thumb, associations between the risk allele and expression should initially be tested in the cell lineage that is believe to give rise to the tumor type (and subtype) under study and contrasted with the expression patterns and associations in other tissues. 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. Transcript abundance is usually evaluated under steady-state conditions. Lastly, effects may only be revealed in certain contexts, such as the activation of a particular pathway. In these cases, alternative assays will be required to implicate these genes.

Future questions for the field include: What are 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. Should the diseased tissue or normal tissue or both be evaluated? We will address some of these points later in the article. Recent elegant research demonstrates that network analysis using risk variant and gene expression data is proving to be a powerful and fruitful tool in dissecting 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)[27], through to Bayesian network approaches to identify predictive relationships between genes from a combination of expression and eQTL data[28]. 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. Finally, RNA-sequencing (RNA-seq) using next generation sequencing platforms has the potential to reveal the transcriptome in its entirety. The RNA-seq platform also possesses desirable characteristics such as increased sensitivity for low abundance transcripts and a wider dynamic range than microarrays.

**Epigenetics represents an additional tier of regulatory control relevant to understanding the transcriptional effects of SNPs in gene-distal or –proximal regions.**

Promoter methylation, histone tail modifications and altered expression of non-coding RNAs, such as the large intergenic noncoding RNAs (lincRNAs)[29,30] which associate with chromatin modifying complexes, contribute to gene regulation in normal development as well as to aberrant gene expression in tumorigenesis [31]. Furthermore, epigenetic mechanisms play an important role in mediating environmental influences on gene expression [32]. Epigenetic silencing has been shown to be the predominant mechanism of gene silencing for a subset of genes [33], while for other genes genetic as well as epigenetic mechanisms have been shown to jointly contribute to tumor suppressor gene activation [34]. An important research direction in the field is to understand the interplay between environmental, genetic and epigenetic factors, both in healthy tissues as well as in the initiation and progression of malignancies. Here the focus is on two questions: (1) Do genetic variants alter the epigenetic landscape and in this way increase the susceptibility to develop cancer? (2) Do genetic variants increase the risk of a locus to become epigenetically silenced in the tumor? Technologies, in particular for the assessment of DNA methylation are now advanced and affordable enough to screen large tissue collections at a single CpG resolution [35]. Methylation profiling in and around haplotype blocks associated with cancer risk thus represents one appropriate strategy. Mechanisms by which genetic variants have the ability to affect epigenetic marks are known from studies in hereditary non-polyposis colorectal cancer. Here constitutional epimutations exist, that predispose to early onset colorectal cancer [36,37]. The *MSH2* and *MLH1* constitutional epimutations, present in somatic cells including peripheral blood lymphocytes, are examples

of highly penetrant epigenetic predisposing factors. For example, mutation of *TACSTD1* leads to transcriptional inactivation through promoter methylation of a single *MSH2* allele in normal somatic tissues. It has been proposed that many epimutations are a consequence of *cis* or *trans* acting genetic variants (reviewed in [38]). In an elegant experiment Kerkel et al [39] showed sequence dependence of allele-specific methylation and demonstrated that *cis*-regulatory polymorphisms control gene expression and affect chromatin states. Recent Chip-Seq data provides further compelling evidence that SNPs and structural variants frequently coincide with allele specific differences in transcription factor binding and chromatin structure. This can now be investigated using allele-specific sequence analysis approaches [40], although the depth of sequencing required for this approach is still an issue [41]. Genome-wide maps of allelic asymmetries are expected to identify functional regulatory polymorphisms [42]. The fact that SNPs can affect allelic imbalance in a tissue specific manner, as has been shown for *UGT2B15* [43], is important in the experimental design. Further epigenetic mechanisms modulating gene expression include miRNAs and miRNA binding sites which can directly be affected by SNPs [44], and tandem repeats that can impact gene expression e.g. by altering transcription factor binding sites, but also by affecting chromatin structure (reviewed in [45]).

Having used the approaches described above to generate a focused list of polymorphisms for functional follow-up, the subsequent challenge is to examine their impact in appropriate *in vivo* model systems. The principal criterion for taking a polymorphism forward is the association of the eQTL with disease.

**Models for testing function**

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 (**Table 1**). 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 data 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. Models are however generally limited to studying one variant/gene at a time. It remains technically even more challenging to study cofounder effects, the possibility that multiple genes in a locus cooperate. This mirrors challenges of assessing individual risk based on a statistical model in which SNPs interact, and indeed the statistical significance of associations between SNPs and disease is predicated on the idea of independent contributions from each SNP rather than a complex interplay.

The basic hypothesis underpinning these studies is that genetic variation at susceptibility loci influences the initiation of the disease phenotype. Although the expression of several highly penetrant disease genes (e.g. *BRCA1*, *BRCA2*, *APC*) is ubiquitous, the functional effects of genetic variation are reflected in a tissue-specific manner. Therefore, it will be necessary to evaluate the functional effects of disease-associated SNPs in the precursor tissue.

Another issue is the size of the functional effects expected for common genetic variants. Whereas the functional effects of several sequence alterations in genes such as *BRCA1* and *APC* are clearly detrimental (they are usually coding sequence changes leading to defective protein products or abrogated expression), the functional effects of SNPs are likely to be subtle, and therefore likely to require large sample sizes in order to achieve sufficient power to detect associations. Thus, large bio-banks of normal tissues may need to be established to evaluate functional 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 bio-repositories, it has been less common to do so for normal tissues from the cells of origin of cancers. This issue is particularly problematic for tumor subtypes in which the cell of origin is still debated.

Studies of normal tissues will need to be complemented by *in vitro* analyses using cell culture models. The location of SNPs with respect to candidate genes will provide multiple testable hypotheses about the functional consequence of genetic variants- for example whether or not SNPs lie in coding sequences, intronic regions, in well-defined gene promoters or stretches of chromatin enhancer sequences that may suggest a role in transcriptional regulation. Progress in establishing good 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*. By using such 3D models it is possible to dissect subtle phenotypes, such as changes associated with gene dosage.

The development of *in vitro* models of human tumors, which represent the quickest and most accessible way to test the function of candidate genes located at susceptibility loci, are more advanced, but functional effects may be masked by an aberrant genetic background. Importantly most SNPs are reported to confer a predisposition to a particular disease and consequently we must expect that the greatest functional impact will be achieved in an essential healthy, non-aberrant tissue/background. Such a context in which to study function 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.

Evaluating the functional effects of SNPs using *in vivo* models represents an even greater challenge. Even with genetically engineered animal models, several significant limitations in the biological validation of candidate genes remain including:

- Relatively short duration of experimental models compared to human tumorigenesis that typically develops over several decades.
- Important differences in human vs. animal physiology
- Important differences in the structure and sequence of non-coding regions
- Limited modelling of gene-environment interactions
- Sensitivity of animal modelling to confirm function of low penetrance alleles when studied in isolation.

Despite these limitations, animal models remain a vital tool for post-GWAS validation, particularly when considering quantitative phenotypes. The multi-tumour APC-mutant mice for example provide power to detect quite subtle effects of variants.

**Exploring the functions of SNPs in regulatory sequences/regions: drawing the themes together.**

Functional SNPs found within large, non-coding intergenic intervals highlight the challenges of defining LD structure and pursuing functional validation. One of the hypotheses to explain this is that the risk locus contains regulatory region(s) with element(s) affecting the expression of one or more distantly transcribed region(s) of the genome. Here the starting point is to explore whether they are components of regulatory elements and what their distal effects may be. Although the most abundant of these regulatory sequences are enhancers, they may also include other regulators such as insulators and silencers. Unlike promoters (at transcription start sites of genes), distal regulatory sequences, such as enhancers, are often cell-type specific [46] and thus may be targets for tissue-specific risk-SNP effects. Annotating such regulatory sequences using chromatin marks or DNase sensitivity has proven to be a powerful method[47,48], and is more informative than the alternative approach of using evolutionary conservation, since regulatory elements tend to be unconstrained across mammalian evolution [49-51]. Specifically, identifying regulatory sequences containing functional SNPs within response elements by using chromatin annotations has been proposed recently[52]. Additionally, demarcation of such regulatory regions is even more precisely 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 (~1kb) amenable to detailed analyses. Enhancer activity in such regulatory regions can be assayed using reporter genes *in vitro*[5] and/or *in vivo*[47]. Additionally, resequencing the target regulatory regions should then be prioritized to capture all the variation within them. Once activity is demonstrated, identified SNPs within the regions, especially ones within known transcription factor response elements and in LD with a tag-SNP, may be analyzed using biochemical methods for differential transcription factor binding and activity. Resequencing, targeting the regulatory regions, then should be prioritized to capture all the variation within them. Finally, regulatory sequences containing functional SNPs determined in this way can be matched with their physiological target genes to probe functional significance (see below). It is possible that multiple SNPs may function co-ordinately at a particular locus and each should be taken forward in subsequent mechanistic analyses.

Three approaches are conceptually available to identify targets of regulatory sequences: (i) Target genes may be identified in regulatory sequences knockout mouse models (using cre-lox for tissue-specific and timing purposes) by genome-wide gene expression analyses after the knockout; and (ii) Target genes may also be identified by using the regulatory sequences as baits in chromatin conformation capture (3C) based studies[53,54], including genome-wide 3C-seq; and (iii) Targeted editing using somatic cell knock-in technology, although technically demanding, is another approach. Allelic series in isogenic settings may be created and gene expression differences measured - either in naturally growing cells or in cells that are perturbed in some manner (e.g., radiation, hormones, etc.). Finally, these screening approaches should be followed by matching results from them, creating a list of robust targets that can each specifically be studied further.

Finally, these screening approaches should allow the matching of results from and between them, creating a priority list of robust potential targets that might each specifically be studied further. Thus, target genes under control of functional SNP-containing regulatory regions may have important roles in the cancer phenotype, 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 cancer cells and mouse xenografts

after the overexpression of the genes of interest, or their selected siRNA/shRNA knockdown. It will also be important to develop ways to make this fine functional annotation data generated in these analyses freely accessible, because in many cases while they may not constitute publishable data they could be useful for other post-GWAS efforts.

*In summary the order of investigation is: identify chromatin architecture in tag-SNPS (+LD blocks) → assay for regulatory activity, such as enhancers → determine how SNPs in LD with the tag-SNP effect such activity → identify regulated target genes → understand biology/cancer risk → understand causal predisposition.*

## 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 SNPs that are likely to contribute most to disease causation to invest in functional follow-up. As sequencing technologies become cheaper and more accessible, we argue that this will evolve rapidly as datasets expand and will afford greater certainty in defining both SNPs and LD structure within the region in which they lie. This will require a detailed mapping and annotation of epigenetic and transcriptomic landscapes within which a major limiter may prove to be the sample collections themselves. Whilst this progresses, and hopefully does so increasingly through consortia assembled from academia and industry, it is vital that proof-of-principle studies take forward the strongest candidate SNPs available so far, not in this case necessarily to test their causative association with disease but to understand their functional impact. What makes for strong candidates are 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. Successfully making the experimental transitions to progress through this process will require collective working at a consortia/multi-group level and a clear decision tree. 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 localizing causal variants. If we begin considering how to explore the functional impact of SNPs now we will, as a community, be in good shape 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. Nonetheless our view is that causality can only be inferred if the eQTL is associated with disease and a SNP leads to expression differences in reliable *in vitro* and perhaps *in vivo* assays. 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 but we hope that this article will help to frame the developing debate and the emerging research that seeks to rise 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 RFA-CA-09-002. 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/pgwas/.

## 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:* [55-5855-58]

**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:* [59]

**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:* [60,61]

**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:* [62-64]

**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:* [65]

**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:* [66]

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

*Supporting references:* [67]

**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:* [59,68]

**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:* [55]

**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:* [69,70]

**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:* [70]

**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:* [69,71-73].

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

*Supporting references:* [74]

**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:* [67]

**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:* [69]

**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:* [65]

**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:* [65]

**SNP:** single nucleotide polymorphism

**Table 1. Methods for functional validation/enabling technologies**

| **Approach** | **Description** | **Host/Applications** | **Advantages/disadvantages** | **Technical challenges, References, and Resources** |
|---|---|---|---|---|
| Homologous recombination | Modification of cell's or organism's genotype to assess allele-specific effects using targeted recombination vectors | *Cell lines:* Manipulation of cell's genotype by introduction/changes in specific SNPs into cell lines (*e.g.* colorectal cancer HCT116 cell line). | Cell lines can be used for biochemistry studies, drug response screens, short and long term cell biological assays (*e.g.* apoptosis, survival fraction assays, soft agar growth assays).<br><br>Limitations include the inability to assess effects from the in vivo milieu or developmentally-restricted effects. | Relatively low frequency of integration necessitating the screening of a large number of clones. Efficiency of integration may be gene specific as it depends on the structure of the homology segments flanking the target locus. While a targeting vector with a $G418^r$ (or $Hyg^r$) cassette driven by the phosphoglycerate kinase promoter can be effectively used to target regulatory regions, it is expected to increase the number of random events<br><br>*References and resources:* [75-79] _ENREF_2 [80-83] |
| | | *Mouse models:* Generation of transgenic mice for the tissue-specific and/or doxycycline regulated expression of target genes. Gene replacement and knockouts in animals (Cre recombinase-mediated). | Allows study of effects of associated SNPs or associated genes in development and in the context of the whole organisms.<br>Ability to assess effects of the microenvironment and paracrine effects.<br>Mouse physiology not always similar to human. In particular | Recombineering techniques for the production of BAC transgenes and gene targeting constructs allow faster production of mouse models for functional genomic studies. Tissue specific expression depends on availability of |

| | | | | |
|---|---|---|---|---|
| | | | some tumor types and precursor lesions may be different in humans. | tissue-specific promoters to drive Cre expression.<br>*References and resources:*<br>http://www.eucomm.orghttp://www.eucomm.org<br>http://www.komp.org |
| Gene targeting by zinc finger nucleases | DNA-binding nucleases that can be used for genomic editing. It creates DSBs at specified sites which are normally repaired by error-prone NHEJ. A transfected template including a mutation favors homology-directed recombination leading to a knock-in of the desired mutation. | *Cell lines:* Manipulation of cell's genotype by introduction/changes in specific SNPs into cell lines (*e.g.* myelogenous leukaemia K562 cell line). | *Same as described for cell lines (see above)* | The creation of a DSB leads to a high frequency of modified clones. Challenges include limited availability of off-the-shelf cell lines and very high cost.<br>*References and resources:* [84,85]<br>http://www.zincfingers.org<br>http://www.sigmaaldrich.com/life-science/zinc-finger-nuclease-technology.html |
| | | *Rat models:* Rapid and targeted gene knockouts in ES cells enables the development of transgenic rats. | Advantages include several listed for mouse models. In addition, the anatomy and physiology of certain organs (*e.g.* prostate) is closer to humans than the mouse. | Expensive to make and to maintain.<br>*References and resources:* [86] |
| RNA | Silencing of | *Cell lines:* | Easier and faster than | There are now |

| | | | | |
|---|---|---|---|---|
| interference | gene expression using plasmid-encoded short hairpin RNAs (shRNAs) or double stranded RNA (dsRNAs) | Manipulation of expression levels of associated genes to monitor its effects. | homologous recombination. Lentiviral/adenoviral delivery generates high transduction efficiencies and can be done transiently or stably. Results may depend on the effectiveness of shRNA. While there are clear advantages of using stable cell lines they may adapt to the knock down by acquiring additional mutations or the stable knock down might be incompatible with viability. Thus, both approaches should be exploited. | several commercially available sources of validated shRNAs, developed by the RNAi Consortium, cloned in lentiviral vectors that can be packaged by 293FT into replication-defective lentivirus. Thoughtful controls should be applied to rule out off-target effects.<br>*References and resources:* [87]<br>http://www.broadinstitute.org/rnai/trc/lib |
| Ectopic gene overexpression | Transient or stable transfection mediated by plasmid or lentiviral expression vectors. | *Cell lines:* Modulation of expression levels achieved often using constructs with strong promoters (*e.g.* CMV) but tissue-specific and inducible promoters can also be used to modulate gene expression. | Fast, easy, economical and reproducible. However, it is often difficult to obtain expression comparable to physiologic levels or in a cell cycle-specific pattern. There may be artifactual effects resulting from overexpression. | Analysis should be performed in pools as well as in multiple clonal isolates to control for clonal variation or adaptation.<br>*References and resources:* [88] |



## 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. Easton, D.F. & Eeles, R.A. Genome-wide association studies in cancer. *Human Molecular Genetics* **17**, R109-R115 (2008).
9. Kanetsky, P.A. et al. Common variation in KITLG and at 5q31.3 predisposes to testicular germ cell cancer. *Nat Genet* **41**, 811-5 (2009).
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. Monks, S.A. et al. Genetic inheritance of gene expression in human cell lines. *Am J Hum Genet* **75**, 1094-105 (2004).
12. Morley, M. et al. Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743-7 (2004).
13. Stranger, B.E. et al. Genome-wide associations of gene expression variation in humans. *PLoS Genet* **1**, e78 (2005).
14. Schadt, E.E. et al. Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297-302 (2003).
15. Johnson, J.M. et al. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302**, 2141-4 (2003).
16. Rockman, M.V. & Kruglyak, L. Genetics of global gene expression. *Nat Rev Genet* **7**, 862-72 (2006).
17. 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).
18. Nicolae, D.L. et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet* **6**, e1000888.
19. Moffatt, M.F. et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* **448**, 470-3 (2007).
20. Musunuru, K. et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714-9 (2010).
21. Zhong, H. et al. Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. *PLoS Genet* **6**, e1000932 (2010).
22. 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).
23. Schadt, E.E. et al. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol* **6**, e107 (2008).
24. Pastinen, T. Genome-wide allele-specific analysis: insights into regulatory variation. *Nat Rev Genet* **11**, 533-8.
25. Montgomery, S.B. et al. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773-7 (2010).
26. Pickrell, J.K. et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768-72 (2010).

27. 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).
28. Bumgarner, R.E. & Yeung, K.Y. Methods for the inference of biological pathways and networks. *Methods Mol Biol* **541**, 225-45 (2009).
29. Gupta, R.A. et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071-6.
30. 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).
31. Jones, P.A. & Baylin, S.B. The epigenomics of cancer. *Cell* **128**, 683-92 (2007).
32. Jirtle, R.L. & Skinner, M.K. Environmental epigenomics and disease susceptibility. *Nat Rev Genet* **8**, 253-62 (2007).
33. Raval, A. et al. Downregulation of death-associated protein kinase 1 (DAPK1) in chronic lymphocytic leukemia. *Cell* **129**, 879-90 (2007).
34. 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).
35. Lister, R. et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315-22 (2009).
36. Chan, T.L. et al. Heritable germline epimutation of MSH2 in a family with hereditary nonpolyposis colorectal cancer. *Nat Genet* **38**, 1178-83 (2006).
37. Suter, C.M., Martin, D.I. & Ward, R.L. Germline epimutation of MLH1 in individuals with multiple cancers. *Nat Genet* **36**, 497-501 (2004).
38. Hesson, L.B., Hitchins, M.P. & Ward, R.L. Epimutations and cancer predisposition: importance and mechanisms. *Curr Opin Genet Dev* **20**, 290-8 (2010).
39. Kerkel, K. et al. Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. *Nat.Genet.* **40**, 904-908 (2008).
40. Ge, B. et al. Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. *Nat Genet* **41**, 1216-22 (2009).
41. Pastinen, T. Genome-wide allele-specific analysis: insights into regulatory variation. *Nat Rev Genet* **11**, 533-8 (2010).
42. Tycko, B. Allele-specific DNA methylation: beyond imprinting. *Hum Mol Genet* (2010).
43. Sun, C., Southard, C., Witonsky, D.B., Olopade, O.I. & Di Rienzo, A. Allelic imbalance (AI) identifies novel tissue-specific cis-regulatory variation for human UGT2B15. *Hum Mutat* **31**, 99-107 (2010).
44. Pelletier, C. & Weidhaas, J.B. MicroRNA binding site polymorphisms as biomarkers of cancer risk. *Expert Rev Mol Diagn* **10**, 817-29 (2010).
45. Gemayel, R., Vinces, M.D., Legendre, M. & Verstrepen, K.J. Variable Tandem Repeats Accelerate Evolution of Coding and Regulatory Sequences. *Annu Rev Genet* (2010).
46. Heintzman, N.D. et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108-12 (2009).
47. Visel, A. et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854-8 (2009).
48. Visel, A., Rubin, E.M. & Pennacchio, L.A. Genomic views of distant-acting enhancers. *Nature* **461**, 199-205 (2009).
49. 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).

50. Blow, M.J. et al. ChIP-Seq identification of weakly conserved heart enhancers. *Nat Genet* **42**, 806-10 (2010).
51. Kunarso, G. et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* **42**, 631-4 (2010).
52. 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).
53. 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).
54. 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).
55. Dekker, J. The three 'C' s of chromosome conformation capture: controls, controls, controls. *Nat Methods* **3**, 17-21 (2006).
56. 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).
57. 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).
58. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-93 (2009).
59. Todd, J.A. Statistical false positive or true disease pathway? *Nat Genet* **38**, 731-3 (2006).
60. Mardis, E.R. ChIP-seq: welcome to the new frontier. *Nat Methods* **4**, 613-4 (2007).
61. Robertson, G. et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* **4**, 651-7 (2007).
62. Iafrate, A.J. et al. Detection of large-scale variation in the human genome. *Nat Genet.* **36**, 949-951 (2004).
63. Lee, C., Iafrate, A.J. & Brothman, A.R. Copy number variations and clinical cytogenetic diagnosis of constitutional disorders. *Nat Genet* (2007).
64. Sebat, J. et al. Large-Scale Copy Number Polymorphism in the Human Genome. *Science* **305**, 525-528 (2004).
65. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57-63 (2009).
66. Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet* **33 Suppl**, 245-54 (2003).
67. 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).
68. Ioannidis, J.P. Common genetic variants for breast cancer: 32 largely refuted candidates and larger prospects. *J Natl.Cancer Inst.* **98**, 1350-1353 (2006).
69. Jobling, M.A., Hurles, M. & Tyler-Smith, C. *Human evolutionary genetics : origins, peoples & disease*, xx, 523 p. (Garland Science, New York, 2004).
70. Griffiths, A.J.F. et al. *Introduction to Genetic Analysis*, (W.H. Freeman and Company, New York, 2005).
71. Altshuler, D., Daly, M.J. & Lander, E.S. Genetic mapping in human disease. *Science* **322**, 881-8 (2008).
72. Lackie, J.M., Dow, J.A.T. & Blackshaw, S.E. *The dictionary of cell biology*, 390 p. (Academic Press, London ; San Diego, 1995).

73. Reich, D.E. et al. Linkage disequilibrium in the human genome. *Nature* **411**, 199-204 (2001).
74. Bartel, D.P. MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215-33 (2009).
75. Di Nicolantonio, F. et al. Replacement of normal with mutant alleles in the genome of normal human cells unveils mutation-specific drug responses. *Proc Natl Acad Sci U S A* **105**, 20864-9 (2008).
76. Bunz, F. et al. Requirement for p53 and p21 to sustain G2 arrest after DNA damage. *Science* **282**, 1497-1501 (1998).
77. Bunz, F. et al. Targeted inactivation of p53 in human cells does not result in aneuploidy. *Cancer Res* **62**, 1129-1133 (2002).
78. Hurley, P.J., Wilsker, D. & Bunz, F. Human cancer cells require ATR for cell cycle progression following exposure to ionizing radiation. *Oncogene* **26**, 2535-2542 (2007).
79. Jallepalli, P.V., Lengauer, C., Vogelstein, B. & Bunz, F. The Chk2 tumor suppressor is not required for p53 responses in human cancer cells. *J Biol.Chem.* **278**, 20475-20479 (2003).
80. Rago, C., Vogelstein, B. & Bunz, F. Genetic knockouts and knockins in human somatic cells. *Nat Protoc.* **2**, 2734-2746 (2007).
81. Topaloglu, O., Hurley, P.J., Yildirim, O., Civin, C.I. & Bunz, F. Improved methods for the generation of human gene knockout and knockin cell lines. *Nucleic Acids Res* **33**, e158 (2005).
82. Wang, P., Yu, J. & Zhang, L. The nuclear function of p53 is required for PUMA-mediated apoptosis induced by DNA damage. *Proc Natl Acad Sci U S A* **104**, 4054-9 (2007).
83. Kohli, M., Rago, C., Lengauer, C., Kinzler, K.W. & Vogelstein, B. Facile methods for generating human somatic cell gene knockouts using recombinant adeno-associated viruses. *Nucleic Acids Res* **32**, e3 (2004).
84. Miller, J.C. et al. An improved zinc-finger nuclease architecture for highly specific genome editing. *Nat Biotechnol.* **25**, 778-785 (2007).
85. Moehle, E.A. et al. Targeted gene addition into a specified location in the human genome using designed zinc finger nucleases. *Proc Natl Acad Sci U S A* **104**, 3055-60 (2007).
86. Geurts, A.M. et al. Knockout rats via embryo microinjection of zinc-finger nucleases. *Science* **325**, 433 (2009).
87. Kaiser, J. Biobanks. Population databases boom, from Iceland to the U.S. *Science* **298**, 1158-61 (2002).
88. Dull, T. et al. A Third-Generation Lentivirus Vector with a Conditional Packaging System. *The Journal of Virology* **72**, 8463-8471 (1998).