Published in final edited form as:

Trends Genet. 2014 April ; 30(4): 140–149. doi:10.1016/j.tig.2014.02.006.

Laying a solid foundation for Manhattan ‘Setting the functional basis for the post-GWAS era’

Xiaoyang Zhang^{1,*}, Swneke D. Bailey^{2,3,¶}, and Mathieu Lupien^{2,3,4,§}

¹Department of Genetics, Norris Cotton Cancer Center, Dartmouth Medical School, Lebanon, NH, USA

²The Princess Margaret Cancer Centre University Health Network, Toronto, ON, Canada

³Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada

⁴Ontario Institute for Cancer Research, Toronto, ON, Canada

Abstract

Genome-wide association studies (GWAS) have identified more than 8,900 genetic variants, mainly single nucleotide polymorphisms (SNPs), associated with hundreds of human traits and diseases, which define risk-associated loci. Variants that map to coding regions can affect protein sequence, translational rate and alternative splicing, all of which influence protein function. However, the vast majority of sequence variants map to non-coding intergenic and intronic regions, and it has been much more challenging to assess the functional nature of these variants. Recent work annotating the non-coding regions of the genome has contributed to post-GWAS studies by facilitating the identification of the functional targets of risk-associated loci. Many non-coding genetic variants within risk-associated loci alter gene expression by modulating the activity of cis-regulatory elements. Here, we review these recent findings and discuss their implication for the post-GWAS era and relate their importance to the interpretation of disease-associated mutations identified through whole-genome sequencing.

Keywords

gwas; genetic risk variant; causal variant; missing heritability; functional genomic; noncoding

Introduction

In an effort to characterize the genetic variation present within human populations the HapMap and 1,000 genome projects have identified approximately 40 million genetic

© 2014 Elsevier Ltd. All rights reserved.

[§]Address correspondence to: Mathieu Lupien, PhD, Princess Margaret Cancer Centre/University Health Network, The MaRS Centre, TMDT room 9-314, 101 College Street, Toronto, Ontario, Canada, M5G 1L7, mlupien@uhnres.utoronto.ca.

^{*}Current address: Dana-Farber Cancer Institute, Boston, MA

[¶]These authors contributed equally to this work.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

variants across the human genome [1, 2], which include structural variants such as insertion-deletions (indels), copy number variants (CNVs), and inversions as well as single nucleotide polymorphisms (SNPs). SNPs are the most abundant form of genetic variation, accounting for 95% of all known sequence variants (38 million) [2]. Genome-Wide Association Studies (GWAS) have identified more than 8,900 SNPs, referred to hereafter as leadSNPs, associated with hundreds of human traits and diseases [3, 4]. These leadSNPs capture the variation present at risk-associated loci, but are, themselves, unlikely to be the causal genetic variants that underlie the association [3]. Each risk-associated locus consists of a collection of genetic variants, all putatively causal, that are in linkage disequilibrium (LD) with the original leadSNP [5, 6], which results from the initial design of the GWAS studies. GWAS studies are array based and include a small fraction of all known SNPs (typically less than 10%) and were selected to capture, or tag, the common genetic variation present in the population by leveraging the extensive LD found across the human genome [7]. Therefore, any of the genetic variants within a risk-associated locus that are in strong LD with the leadSNP can account for the observed difference in phenotype associated with that locus. Thus, one goal for the post-GWAS era is to identify the specific genetic variant(s) from a risk-associated locus that accounts for phenotypic differences based on the functional biology it modulates.

Only ~16% of risk-associated loci harbor SNPs that affect coding sequences, absolving the majority of risk-associated loci from altering the protein sequence [8]. Population-based studies have demonstrated that genetic variants are associated with gene expression [9–12], RNA-splicing [13], transcription factor binding [14], chromatin openness measured by DNaseI hypersensitivity [15], DNA methylation [16] and histone modifications [17–19]. In addition, SNPs are more commonly associated with a particular phenotype if they fall within a DNaseI hypersensitive region from a relevant cell-type [20]. This enrichment even applies to SNPs with a probability, or p-value, of association below the genome-wide significance threshold, suggesting that there are many risk-associated loci yet to be found below the Manhattan plot “skyline”.

Until recently, the functional characterization of risk-associated loci was hindered by the limited annotation of the human genome outside of coding sequences. However, approaches to successfully characterize the functional nature of these loci are beginning to emerge, thus providing for the first time a foundation on which to build the functional validation of the risk-associated loci, “buildings”, of the Manhattan plot “skyline”. A series of large-scale genomics projects, including the ENCyclopedia of DNA Element (ENCODE), the Roadmap Epigenomics, the International Human Epigenome Consortium (IHEC) and the Functional Annotation of the Mammalian Genome (FANTOM) projects, as well as independent labs are leading an effort to comprehensively annotate the non-coding regions of the human genome in a number of diverse cell and tissue types and across several developmental stages [21–26]. These large-scale studies have benefited from recent technological advances in massively parallel sequencing-based technologies to generate genome-wide maps of functional elements, such as transcripts, regulatory elements, and origins of replication. For examples, RNA-sequencing (RNA-seq) and Cap Analysis of Gene Expression sequencing (CAGE-seq) annotate known and novel transcripts such as long noncoding RNA (lncRNA) and enhancer RNA (eRNA) [27–29]. Whole-Genome Epigenetic Mapping (WGEM) for

histone modifications through chromatin immunoprecipitation paired with massively parallel sequencing (ChIP-seq) identifies regulatory elements including promoters, enhancers, and insulators [30–34]. In addition, ChIP-seq assays reveal regulatory elements that are bound by transcription factors [22, 35, 36]. Meanwhile, high-throughput chromatin looping studies, such as Hi-C, 5C, e4C, and ChIA-PET assays, can reveal the regulatory enhancer-enhancer, enhancer-promoter or promoter-promoter combinations defining the three-dimensional nature of the regulatory element web [37–42]. Other DNA-templated processes, such as DNA damage repair or replication timing, can be annotated by WGEM of histone variants [43] and replication timing assays, repli-seq or origin recognition complex (ORC) ChIP-seq [44–46], respectively. Finally, inter-species evolutionarily conserved DNA sequences can predict functional elements and complement these maps [47, 48]. Intra-species comparison among humans also identifies conserved regions of the genome that predict potential functional DNA elements [49, 50]. Together, these layers of biological information, across the human genome, serve as the foundation for post-GWAS functional studies.

Genetic risk variants affecting coding regions

Non-synonymous genetic risk variants

Each protein has a unique sequence of amino acids specified by the coding DNA and changes to its sequence can drastically impact its function [51]. The risk associated with non-synonymous genetic variants (nonsense or missense) can readily be translated into a change in protein structure and/or function based on the amino acid change (Figure 1A). Hence, non-synonymous variants readily prioritize genes that may have important functional relevance to the trait or disease studied. For instance, the rs3812316 SNP is a non-synonymous genetic risk-variant associated with reduced plasma triglyceride that maps to the *MLXIPL* gene. This gene codes for a transcription factor promoting the expression of genes that are part of the triglyceride biosynthesis pathway [52, 53]. The risk allele of the rs3812316 SNP results in a glutamine to histidine substitution at position 241 (Q241H) of the activation domain of the *MLXIPL* protein [52]. This substitution appears to disrupt the activity of *MLXIPL* because it is associated with lower triglyceride levels. Consistent with this hypothesis, *MLXIPL*-null mice have reduced triglyceride levels [52, 53].

The rs1990760 SNP associated with type 1 diabetes (T1D), an autoimmune disease, is another example of a functional non-synonymous genetic risk-variant [54]. It maps to the *Interferon Induced with Helicase C domain 1 (IFIH1)* gene causing an alanine to threonine substitution at position 946 (A946T) of the IFIH1/MDA5 protein [54]. Although this does not affect a known catalytic domain of IFIH1, the rs1990760 SNP locates to a region of the *IFIH1* gene highly conserved across mammals, suggestive of its functionality [54]. Wild-type IFIH1 protein senses RNA virus infection by recognizing double-stranded RNA generated during virus replication. This results in interferon release from the infected cells, triggering the innate antiviral immune response [55, 56]. Accordingly, the rs1990760 SNP is thought to weaken the immune response induced by double-stranded RNA increasing the risk of virally inducing T1D.

Synonymous genetic risk variants

Synonymous genetic risk-variants by definition do not alter the codon sequence and therefore encode wild-type protein sequences. However, they can still impact protein function. For instance, synonymous genetic risk-variants can modulate translation rates with direct consequences to protein folding (Figure 1B) and is the case for the rs1045642 SNP that maps to the *Multidrug Resistant* gene *MDR1* [57, 58]. The *MDR1* gene encodes a cell-membrane transporter protein involved in drug trafficking [59] and the rs1045642 SNP alters the drug substrate specificity of MDR1 but does not impact the sequence or the expression of the MDR1 protein [58]. The rs1045642 SNP changes the frequent isoleucine (Ile) codon ATC to the rare Ile codon ATT [58]. It has been suggested that this slows down the rate of translation of the *MDR1* mRNA, which impacts protein folding [60], and that the subsequent altered MDR1 conformation decreases its drug substrate specificity [57–59]. Recently, it has been shown that a fraction of codons specify not only an amino acid, but a transcription factor binding site, providing an additional avenue through which synonymous polymorphisms may impart a functional effect [61].

Genetic risk-variants affecting RNA splicing

Splicing, in which introns are excised and exons are joined, relies on the RNA sequence [62]. Exonic splicing enhancers (ESEs) that consist of specific hexamer sequences and an AG sequence at the intron-exon boundary guide the recruitment of the splicing complex to immature RNA (pre-mRNA) leading to intron excision and exon joining [62–65]. The rs1800693 SNP located proximal to the exon 6/intron 6 boundary of the *Tumor Necrosis Factor Receptor Superfamily Member 1A TNFRSF1A* gene is associated with multiple sclerosis and affects splicing of the TNFRSF1A mRNA [66]. The reference allele allows for the production of the Tumor Necrosis Factor Receptor (TNFR1) protein. However, the risk allele induced splicing results in the shorter Δ 6-TNFR1 protein (Figure 1C) [66]. Whereas TNFR1 localizes to the cell membrane, the Δ 6-TNFR1 protein remains cytoplasmic and can antagonize TNF signaling to promote multiple sclerosis [66].

Noncoding genetic risk variants: Beyond the coding genome

Most genetic risk-variants fall outside of coding sequences [67]. Recent post-GWAS studies have demonstrated the capacity of these genetic risk-variants to regulate gene expression by modulating cis-regulatory machineries through mechanisms involving DNA methylation, transcription factor binding, chromatin looping, or miRNA recruitment.

Genetic risk variants and DNA methylation at promoters

DNA methylation consists of the addition of methyl groups to a cytosine nucleotide, which is typically part of a CpG dinucleotide. This heritable epigenetic event is involved in transcriptional regulation [68]. Aberrant DNA methylation patterns are typical of cancer cells [69]. DNA hyper-methylation near transcription start sites of tumor suppressor genes correlates with their silencing [68]. For example, the *Hepatocyte Nuclear Factor 1 Homeobox B gene HNF1B* is silenced by DNA methylation in serous ovarian tumors. The rs7405776 SNP defines a risk-locus for invasive serous ovarian cancer that is located within the promoter region of the *HNF1B* gene. This risk-associated locus is located in a CpG

island and is associated with higher DNA methylation levels at the *HNF1B* promoter [16], which suggests that this locus increases the risk for ovarian cancer through epigenetic silencing of the *HNF1B* gene (Figure 2A), although the exact mechanisms are yet to be defined.

Genetic risk variants modulate transcription factor binding to the chromatin

Transcriptional regulatory networks are required to establish lineage-specific expression programs defining cellular identity [34, 70–72]. Transcription factors bind to thousands of regulatory elements across the genome, including promoters directly upstream of their target genes and *cis*-regulatory elements such as enhancers, silencers, and insulators [73]. ChIP-seq assays for transcription factors or specific epigenetic modifications effectively annotate these *cis*-regulatory elements genome-wide. Using these annotations it was demonstrated that genetic risk variants commonly target *cis*-regulatory elements, mainly enhancers, in a disease- and tissue-specific manner [8, 20, 31, 74–76]. For example, loci associated with erythrocyte phenotypes commonly harbor enhancers that are functional in K562 erythrocyte leukemia cells, but not enhancers functional in other cell types [31]. Consistently, breast cancer risk-associated loci are associated with enhancers that are functional in T47D breast cancer cells [74]. By contrast, the genetic variants associated with erythrocyte phenotypes or breast cancer do not target active enhancers found in cells of an unrelated lineage [31, 74].

Risk-associated loci targeting enhancers commonly harbor genetic variants that map to DNA recognition motifs, also known as DNA response elements, bound by transcription factors. These genetic variants can modulate the chromatin affinity for transcription factors with direct consequences on target gene expression [74, 77–83] (Figure 2B). For instance, the variant allele of the rs1427407 SNP, which is associated with the fetal hemoglobin level, decreases the recruitment of GATA1/TAL1 to the enhancer region, resulting in the down-regulation of the *BCL11A* gene, a repressor of the fetal hemoglobin level [78]. Similarly, the rs12740374 SNP, which is associated with a lower level of plasma low-density lipoprotein cholesterol (LDL-C), up-regulates the expression level of the *SORT1* gene by increasing the binding affinity of the C/EBP transcription factor to the chromatin [79]. Over-expression of *SORT1* leads to a lower LDL-C level in livers [79]. In addition, the rs10811656 and rs10757278 SNPs, which are associated with coronary artery disease, alter DNA recognition motifs in a synergetic manner. The risk allele of these SNPs changes the same STAT DNA motif to decrease STAT1 binding in human vascular endothelial cells (HUVEC) [77]. This results in the differential expression of the target gene *ANRIL* [77].

Modulation of transcription factors by genetic risk variants also applies in cancer. The colon cancer risk-associated SNP, rs6983267, maps to a distal functional enhancer upstream of the *MYC* gene. The rs6983267 SNP lies in a DNA recognition motif for the TCF4 transcription factor [82]. The G risk allele increases the binding affinity between TCF4 and the enhancer compared to the reference T allele [80–82]. This increases the enhancer activity [81] and results in higher levels of expression for the *MYC* oncogene [82]. Similarly, the breast cancer rs4784227 SNP genetic risk variant maps to a forkhead motif. The risk allele increases the binding affinity of the pioneer factor FOXA1 in breast cancer cells [74]. This favors the recruitment of the Groucho/TLE repressive complex to an enhancer to repress

expression of the *TOX3* gene [74]. In prostate cancer, the 17q24.3 risk-associated locus harbors two functional variants, namely the rs8072254 and rs1859961 SNPs, found in a double enhancer site looping to the *SOX9* gene [83]. The rs8072254 SNP increases prostate cancer risk by affecting the DNA recognition motif of the androgen receptor (AR) imposing its allele-specific recruitment [83]. The rs1859961 SNP is rare and unique in that it both disrupts a forkhead and creates an AP-1 DNA recognition motif. This decreases binding of the pioneer factor FOXA1 while increasing binding of the AP-1 transcription factor to modulate the same enhancer's activity [83]. This example also highlights that multiple different enhancers can be modulated by functional variants within a single risk locus.

Genetic risk variants found within promoters can also alter transcription factor binding to the DNA leading to differential target gene expression [84, 85]. For instance, expression in the α -globin gene locus is affected by a genetic variant associated with the α -thalassemia blood disorder [84]. The risk allele of this variant creates a GATA-1 motif at a promoter-like region that decreases the expression of the downstream α -globin genes [84]. Reduced expression of α -globin genes promotes α -thalassemia [86].

Genetic risk variants can alter chromatin loop formation bridging enhancers and promoters

The human genome is organized in a three-dimensional architecture, which is thought to regulate a diverse set of DNA-templated processes [87–91]. This allows regulatory elements, such as enhancers and promoters, to physically interact through long-range chromatin interactions, or chromatin loops, to regulate gene expression [39, 42]. The human pigmentation-associated SNP, rs12913832, imposes allele-specific chromatin loop formation [92]. The rs12913832 SNP resides in an enhancer 21 kilobases (kb) upstream of the *OCA2* pigment gene [92]. The T allele of this SNP favors chromatin loops to the *OCA2* gene compared to the C allele and is associated with a darker pigmentation in melanocytes [92]. Specific DNA binding proteins, including the cohesin and mediator complex as well as the insulator protein CTCF, promote chromatin loop formation [93–95]. Although the rs12913832 SNP is the only genetic risk-variant known to modulate chromatin loop formation, variants altering the DNA affinity for looping factors will likely also result in allele-specific chromatin loop formation (Figure 2C).

Genetic risk variants can affect miRNAs

MicroRNAs (miRNA) largely function as post-transcriptional repressors. They recruit RNA-induced silencing complex (RISC) to their target mRNAs, leading to mRNA degradation or translational repression. miRNAs target mRNAs by recognizing their complementary sequences mainly in 3' untranslated regions (3' UTR) [96]. Genetic risk variants can affect miRNA repressive functions by directly changing miRNA sequences or modifying their complementary sequence on target mRNAs (Figure 2D). This is exemplified by the Crohn's disease-associated SNP, rs10065172. It lies within the 3' UTR of the *IRGM* gene and this risk allele changes the complementary target sequence of miRNA-196 [97]. This attenuates miRNA-196 binding to the *IRGM* mRNA increasing the stability of the *IRGM* mRNA and protein levels [97], [98]. Fluctuation in *IRGM* expression results in an increase in the

number of intracellular bacteria, such as adherent invasive *E. coli* that can cause Crohn's disease-associated inflammation [97, 98].

Genetic risk variants and lncRNAs

Long non-coding RNAs (lncRNAs) are defined as non-protein coding transcripts longer than 200 nucleotides in length. They are found across intergenic regions of the human genome [27]. lncRNAs can interact with chromatin regulators to guide their recruitment to the chromatin [99, 100]. A process that relies on highly conserved lncRNA tertiary structure. For instance, the *RepA* lncRNA consists of two conserved stem-loop structures that interact with the EZH2 subunit of the Polycomb Repressive Complex 2 (PRC2) [101]. This guides PRC2 recruitment to the X-chromosome [101]. Similarly, the lncRNA *HOTAIR*, which is expressed from the *HOXC* locus, recruits PRC2 to repress the transcription from the *HOXD* locus [99]. By contrast, the lncRNA *HOTTIP* coordinates the activation of *HOXA* genes by guiding WDR1/MLL complexes to the chromatin [102]. No genetic risk variants map to the *RepA*, *HOTAIR* or *HOTTIP* lncRNAs. However, RNA tertiary structures can be altered by genetic risk-variants [103]. The 9q21.3 (coronary artery disease) and 22q12.1 (myocardial infarction) risk loci harbor SNPs mapping to the *ANRIL* and *MIAT* lncRNAs, respectively [104, 105]. The risk allele of the risk SNP rs35955962 found in the *MIAT* lncRNA increases its affinity for nuclear proteins compared to the non-risk allele [105] (Figure 2E). Further investigations are required to characterize which nuclear protein is influenced and determine the functional consequence of this risk locus on heart disease.

Integrative functional post-GWAS methodologies

Apart from the known targets of genetic risk variants (Figure 3), other DNA-templated processes, such as DNA replication or repair, may also be affected. Integrative functional genomics and bioinformatics methodologies that combine GWAS results, linkage disequilibrium, and whole-genome functional annotations can provide the means to identify the targets of risk-associated loci [8, 20, 31, 74]. The design of these integrative approaches forms the core of post-GWAS era functional validation studies. Basically, leadSNPs for a disease/trait of interest are selected and LD SNPs are imputed for each leadSNP using the appropriate reference population or ethnic group, from the HapMap or 1,000 Genomes project, to establish a genetic risk-variant set [2]. The compiled variant set is then compared to annotation tracks of functional elements from a tissue or cell-type relevant to the disease/trait. Functional elements are derived from whole-genome assays such as RNA-seq, ChIP-seq, DNase-seq, FAIRE-seq, Repli-seq, etc [22, 46].

The variant set enrichment (VSE) approach is amongst a set of first generation integrative tools developed [74]. It is a permutation-based method that compares the enrichment of genetic risk-variant sets within any functional element to randomly generated matched genetic risk-variant sets [74, 106]. Using this method, it has been demonstrated that colorectal cancer risk variants preferentially map to enhancers inactivated in cancer compared to normal cells [106]. It has also revealed that breast cancer risk variants preferentially target enhancers bound by the FoxA1 and ESR1 transcription factors within breast cancer cells [74]. Similar methodologies have associated genetic risk-variants from

various diseases with specific chromatin states defined by WGEM [31] and regions of open chromatin [8, 20].

Bioinformatics tools can be employed to predict the biological impact of genetic risk variants and identify putative causal genetic variant responsible for risk loci (Table 1). Tools such as PolyPhen and MuTIP predict changes in protein structure imposed by genetic risk-variants mapping to coding regions [107, 108]. Motif-prediction tools, such as HaploReg, RegulomeDB, FunSeq and SnpEff, identify genetic variants that significantly alter DNA recognition motifs to modulate transcription factor binding [109–112]. The Intra-Genomic Replicates (IGR) method provides an alternative and can predict changes in chromatin binding affinity of transcription factors caused by risk variants without the use of position-weighted matrices (PWM) [74]. This allows IGR to determine changes in the chromatin binding affinity of transcription factors caused by genetic variants located outside of known DNA recognition motifs [74]. The drawback of IGR is that it relies on the availability of ChIP-seq data and is therefore limited to transcription factors previously investigated by ChIP-seq assays.

SNPs that disrupt the coding sequence of a gene can clearly implicate the corresponding gene in the manifestation of the associated phenotype or disease. However, identifying the target gene of non-coding variants remains a significant challenge. Genetic variation associated with gene expression, known as expression quantitative trait loci (eQTL), can identify the target genes of risk loci [9–12, 113]. If a risk locus coincides with an eQTL it can clearly indicate a regulatory mechanism and provide the impetus for its functional validation. Online tools such as SCAN and the eQTL browser are publically available to query eQTL data [15, 113] and several reviews regarding the application of eQTL studies are available [114, 115]. eQTL analysis can also complement Pathway-based association approaches that apply prior biological knowledge of genes and pathways to the interpretation of GWAS data [117–121]. Pathway-based tools, such as the Gene Relationships Among Implicated Loci (GRAIL), can also identify candidate target genes by identifying genes that are part of a pathway(s) that is enriched within multiple risk-associated loci identified for the same disease [122]. However, pathways are constantly evolving and adapting in parallel with our knowledge of them. Networks created using gene expression data from patient samples can also model the underlying molecular machinery [118] and can be exploited to bridge GWAS results with an underlying disease mechanism, as exemplified in Autism spectrum disorder [123]. Finally, tools such as RegRNA can predict how genetic variants impact miRNA target specificity [97, 116].

Acknowledgments

The National Cancer Institute (NCI) of the National Institutes of Health (NIH) under Award Number R01CA155004 (M.L.) and the Princess Margaret Cancer Foundation (M.L.) supported the research reported in this publication. The research content reported is the sole responsibility of the authors and does not necessarily represent the official views of the funding sources. M.L. holds a young investigator award from the Ontario Institute for Cancer Research and a new investigator salary award from the Canadian Institute of Health Research (CIHR). S.D.B. is supported by a Knudson postdoctoral fellowship from the Princess Margaret Cancer Centre.

Glossary

Genetic variants	DNA sequence differences found across human individuals. These include single nucleotide polymorphisms (SNPs) and structural variants such as insertion-deletions, block substitutions, inversions and copy number variants. The variants that are statistically associated with human diseases or traits are called genetic risk variants
Genome-Wide Association Studies (GWAS)	Studies designed to identify genetic variants, such as SNPs, statistically associated with a human trait or disease
Linkage Disequilibrium (LD)	A non-random association of alleles of multiple genetic variants
LeadSNP	A SNP defining a risk locus identified through GWAS
Manhattan plot	A scatter plot used to present GWAS results. In a Manhattan plot, DNA coordinates of SNPs are displayed on the X-axis, while their negative logarithm <i>P</i> -value associated with a specific trait or disease is displayed on the Y-axis. The name of the plot is derived from its similarity with the Manhattan skyline, a profile of buildings with a range of heights
Regulatory elements	DNA regions that regulate expression of target genes. These include elements such as promoters, enhancers and insulators
Non-synonymous and synonymous genetic variants	Non-synonymous genetic variants are variants that can alter the amino acids sequence of a protein. By contrast, a synonymous genetic variant also resides in a coding exon, but has no effect on the amino acid sequence
DNA methylation	A biochemical process where a methyl group is added to a cytosine nucleotide, typically found in a CpG sequence. DNA methylation represses gene expression when found in CpG rich (CpG island) promoter regions
Chromatin loops	Higher-order chromatin architecture that brings in close physical proximity DNA regions separated by hundred or more nucleotides
Open chromatin regions	Correspond to nucleosome-depleted regions of the chromatin, allowing the binding of proteins to the chromatin. These regions can be identified by DNaseI-seq or FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements)-seq
Chromatin states	Chromatin states represent different spatial combinations of histone modifications. Distinct chromatin states can define different functional units of the genome

References

1. Altshuler DM, et al. Integrating common and rare genetic variation in diverse human populations. *Nature*. 2010; 467:52–58. [PubMed: 20811451]
2. Abecasis GR, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012; 491:56–65. [PubMed: 23128226]
3. 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]
4. Hindorff, LA., et al. A catalog of published genome-wide association studies. 2012. <http://www.genome.gov/gwastudies>
5. McClellan J, King MC. Genetic heterogeneity in human disease. *Cell*. 2010; 141:210–217. [PubMed: 20403315]
6. Raychaudhuri S. Mapping rare and common causal alleles for complex human diseases. *Cell*. 2011; 147:57–69. [PubMed: 21962507]
7. Grant SF, Hakonarson H. Microarray technology and applications in the arena of genome-wide association. *Clin Chem*. 2008; 54:1116–1124. [PubMed: 18499899]
8. Schaub MA, et al. Linking disease associations with regulatory information in the human genome. *Genome Res*. 2012; 22:1748–1759. [PubMed: 22955986]
9. Dimas AS, et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science*. 2009; 325:1246–1250. [PubMed: 19644074]
10. Grisanzio C, et al. Genetic and functional analyses implicate the NUDT11, HNF1B, and SLC22A3 genes in prostate cancer pathogenesis. *Proc Natl Acad Sci U S A*. 2012; 109:11252–11257. [PubMed: 22730461]
11. Li Q, et al. Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell*. 2013; 152:633–641. [PubMed: 23374354]
12. 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]
13. Kwan T, et al. Genome-wide analysis of transcript isoform variation in humans. *Nat Genet*. 2008; 40:225–231. [PubMed: 18193047]
14. Kasowski M, et al. Variation in transcription factor binding among humans. *Science*. 2010; 328:232–235. [PubMed: 20299548]
15. Degner JF, et al. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*. 2012; 482:390–394. [PubMed: 22307276]
16. Shen H, et al. Epigenetic analysis leads to identification of HNF1B as a subtype-specific susceptibility gene for ovarian cancer. *Nat Commun*. 2013; 4:1628. [PubMed: 23535649]
17. McVicker G, et al. Identification of Genetic Variants That Affect Histone Modifications in Human Cells. *Science*. 2013
18. Kasowski M, et al. Extensive Variation in Chromatin States Across Humans. *Science*. 2013
19. Kilpinen H, et al. Coordinated Effects of Sequence Variation on DNA Binding, Chromatin Structure, and Transcription. *Science*. 2013
20. Maurano MT, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science*. 2012; 337:1190–1195. [PubMed: 22955828]
21. 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]
22. Dunham I, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012; 489:57–74. [PubMed: 22955616]
23. Bernstein BE, et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol*. 2010; 28:1045–1048. [PubMed: 20944595]
24. Tang Q, et al. A Comprehensive View of Nuclear Receptor Cancer Cistromes. *Cancer Res*. 2011; 71:6940–6947. [PubMed: 21940749]
25. Ravasi T, et al. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*. 2010; 140:744–752. [PubMed: 20211142]

26. Yan J, et al. Transcription Factor Binding in Human Cells Occurs in Dense Clusters Formed around Cohesin Anchor Sites. *Cell*. 2013; 154:801–813. [PubMed: 23953112]
27. Guttman M, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*. 2009; 458:223–227. [PubMed: 19182780]
28. Djebali S, et al. Landscape of transcription in human cells. *Nature*. 2012; 489:101–108. [PubMed: 22955620]
29. Plessy C, et al. Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. *Nat Methods*. 2010; 7:528–534. [PubMed: 20543846]
30. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol*. 2010; 28:817–825. [PubMed: 20657582]
31. Ernst J, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*. 2011; 473:43–49. [PubMed: 21441907]
32. Heintzman ND, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*. 2009; 459:108–112. [PubMed: 19295514]
33. Heintzman ND, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet*. 2007; 39:311–318. [PubMed: 17277777]
34. Lupien M, et al. FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell*. 2008; 132:958–970. [PubMed: 18358809]
35. Gerstein MB, et al. Architecture of the human regulatory network derived from ENCODE data. *Nature*. 2012; 489:91–100. [PubMed: 22955619]
36. Ross-Innes CS, et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*. 2012; 481:389–393. [PubMed: 22217937]
37. Lieberman-Aiden E, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. 2009; 326:289–293. [PubMed: 19815776]
38. Schoenfelder S, et al. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat Genet*. 2010; 42:53–61. [PubMed: 20010836]
39. Sanyal A, et al. The long-range interaction landscape of gene promoters. *Nature*. 2012; 489:109–113. [PubMed: 22955621]
40. Fullwood MJ, et al. An oestrogen-receptor- α -bound human chromatin interactome. *Nature*. 2009; 462:58–64. [PubMed: 19890323]
41. Handoko L, et al. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet*. 2011; 43:630–638. [PubMed: 21685913]
42. Li G, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*. 2012; 148:84–98. [PubMed: 22265404]
43. Seo J, et al. Genome-wide profiles of H2AX and gamma-H2AX differentiate endogenous and exogenous DNA damage hotspots in human cells. *Nucleic Acids Res*. 2012; 40:5965–5974. [PubMed: 22467212]
44. Ryba T, et al. Genome-scale analysis of replication timing: from bench to bioinformatics. *Nat Protoc*. 2011; 6:870–895. [PubMed: 21637205]
45. Hansen RS, et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc Natl Acad Sci U S A*. 2010; 107:139–144. [PubMed: 19966280]
46. Dellino GI, et al. Genome-wide mapping of human DNA-replication origins: levels of transcription at ORC1 sites regulate origin selection and replication timing. *Genome Res*. 2013; 23:1–11. [PubMed: 23187890]
47. Kolaczowski, B.; Kern, AD. Does conservation imply function? Chapter 6, In *Evolution since Darwin: The first 150 years*. Bell, MA., et al., editors. Vol. Chap. 156. Sinauer; Sunderland, MA: 2010.
48. Siepel A, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*. 2005; 15:1034–1050. [PubMed: 16024819]
49. Ward LD, Kellis M. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science*. 2012; 337:1675–1678. [PubMed: 22956687]
50. Ward LD, Kellis M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol*. 2012; 30:1095–1106. [PubMed: 23138309]

51. Nelson, DL.; MMC. Lehninger's principles of biochemistry. 4. 2005.
52. Kooner JS, et al. Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides. *Nat Genet.* 2008; 40:149–151. [PubMed: 18193046]
53. Uyeda K, Repa JJ. Carbohydrate response element binding protein, ChREBP, a transcription factor coupling hepatic glucose utilization and lipid synthesis. *Cell Metab.* 2006; 4:107–110. [PubMed: 16890538]
54. Smyth DJ, et al. A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. *Nat Genet.* 2006; 38:617–619. [PubMed: 16699517]
55. Andrejeva J, et al. The V proteins of paramyxoviruses bind the IFN-inducible RNA helicase, mda-5, and inhibit its activation of the IFN-beta promoter. *Proc Natl Acad Sci U S A.* 2004; 101:17264–17269. [PubMed: 15563593]
56. Kato H, et al. Differential roles of MDA5 and RIG-I helicases in the recognition of RNA viruses. *Nature.* 2006; 441:101–105. [PubMed: 16625202]
57. Hoffmeyer S, et al. Functional polymorphisms of the human multidrug-resistance gene: multiple sequence variations and correlation of one allele with P-glycoprotein expression and activity in vivo. *Proc Natl Acad Sci U S A.* 2000; 97:3473–3478. [PubMed: 10716719]
58. Kimchi-Sarfaty C, et al. A “silent” polymorphism in the MDR1 gene changes substrate specificity. *Science.* 2007; 315:525–528. [PubMed: 17185560]
59. Fung KL, Gottesman MM. A synonymous polymorphism in a common MDR1 (ABCB1) haplotype shapes protein function. *Biochim Biophys Acta.* 2009; 1794:860–871. [PubMed: 19285158]
60. Komar AA. Genetics. SNPs, silent but not invisible. *Science.* 2007; 315:466–467. [PubMed: 17185559]
61. Stergachis AB, et al. Exonic transcription factor binding directs codon choice and affects protein evolution. *Science.* 2013; 342:1367–1372. [PubMed: 24337295]
62. Matlin AJ, et al. Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol.* 2005; 6:386–398. [PubMed: 15956978]
63. Blencowe BJ. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. *Trends Biochem Sci.* 2000; 25:106–110. [PubMed: 10694877]
64. Cartegni L, et al. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nat Rev Genet.* 2002; 3:285–298. [PubMed: 11967553]
65. Fairbrother WG, et al. Predictive identification of exonic splicing enhancers in human genes. *Science.* 2002; 297:1007–1013. [PubMed: 12114529]
66. Gregory AP, et al. TNF receptor 1 genetic risk mirrors outcome of anti-TNF therapy in multiple sclerosis. *Nature.* 2012; 488:508–511. [PubMed: 22801493]
67. Frazer KA, et al. Human genetic variation and its contribution to complex traits. *Nat Rev Genet.* 2009; 10:241–251. [PubMed: 19293820]
68. Jones PA, Baylin SB. The epigenomics of cancer. *Cell.* 2007; 128:683–692. [PubMed: 17320506]
69. Sharma S, et al. Epigenetics in cancer. *Carcinogenesis.* 2010; 31:27–36. [PubMed: 19752007]
70. Boyer LA, et al. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell.* 2005; 122:947–956. [PubMed: 16153702]
71. Graf T, Enver T. Forcing cells to change lineages. *Nature.* 2009; 462:587–594. [PubMed: 19956253]
72. Son CG, et al. Database of mRNA gene expression profiles of multiple human organs. *Genome Res.* 2005; 15:443–450. [PubMed: 15741514]
73. Ong CT, Corces VG. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat Rev Genet.* 2011; 12:283–293. [PubMed: 21358745]
74. Cowper-Sal Lari R, et al. Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat Genet.* 2012; 44:1191–1198. [PubMed: 23001124]
75. Hnisz D, et al. Super-Enhancers in the Control of Cell Identity and Disease. *Cell.* 2013
76. Parker SC, et al. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc Natl Acad Sci U S A.* 2013

77. Harismendy O, et al. 9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response. *Nature*. 2011; 470:264–268. [PubMed: 21307941]
78. Bauer DE, et al. An Erythroid Enhancer of BCL11A Subject to Genetic Variation Determines Fetal Hemoglobin Level. *Science*. 2013; 342:253–257. [PubMed: 24115442]
79. Musunuru K, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*. 2010; 466:714–719. [PubMed: 20686566]
80. Tuupanen S, et al. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat Genet*. 2009; 41:885–890. [PubMed: 19561604]
81. 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]
82. Wright JB, et al. Upregulation of c-MYC in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. *Mol Cell Biol*. 2010; 30:1411–1420. [PubMed: 20065031]
83. Zhang X, et al. Integrative functional genomics identifies an enhancer looping to the SOX9 gene disrupted by the 17q24.3 prostate cancer risk locus. *Genome Res*. 2012; 22:1437–1446. [PubMed: 22665440]
84. De Gobbi M, et al. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science*. 2006; 312:1215–1217. [PubMed: 16728641]
85. Huang Y, et al. A functional SNP of interferon-gamma gene is important for interferon-alpha-induced and spontaneous recovery from hepatitis C virus infection. *Proc Natl Acad Sci U S A*. 2007; 104:985–990. [PubMed: 17215375]
86. Higgs DR, et al. A review of the molecular genetics of the human alpha-globin gene cluster. *Blood*. 1989; 73:1081–1104. [PubMed: 2649166]
87. Bickmore WA. The spatial organization of the human genome. *Annu Rev Genomics Hum Genet*. 2013; 14:67–84. [PubMed: 23875797]
88. Fraser P, Bickmore W. Nuclear organization of the genome and the potential for gene regulation. *Nature*. 2007; 447:413–417. [PubMed: 17522674]
89. Gibcus JH, Dekker J. The hierarchy of the 3D genome. *Mol Cell*. 2013; 49:773–782. [PubMed: 23473598]
90. Misteli T. Beyond the sequence: cellular organization of genome function. *Cell*. 2007; 128:787–800. [PubMed: 17320514]
91. Roix JJ, et al. Spatial proximity of translocation-prone gene loci in human lymphomas. *Nat Genet*. 2003; 34:287–291. [PubMed: 12808455]
92. Visser M, et al. HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. *Genome Res*. 2012; 22:446–455. [PubMed: 22234890]
93. Splinter E, et al. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes Dev*. 2006; 20:2349–2354. [PubMed: 16951251]
94. Mishiro T, et al. Architectural roles of multiple chromatin insulators at the human apolipoprotein gene cluster. *The EMBO journal*. 2009; 28:1234–1245. [PubMed: 19322193]
95. Jing H, et al. Exchange of GATA factors mediates transitions in looped chromatin organization at a developmentally regulated gene locus. *Mol Cell*. 2008; 29:232–242. [PubMed: 18243117]
96. Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell*. 2009; 136:215–233. [PubMed: 19167326]
97. Brest P, et al. A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. *Nat Genet*. 2011; 43:242–245. [PubMed: 21278745]
98. Singh SB, et al. Human IRGM induces autophagy to eliminate intracellular mycobacteria. *Science*. 2006; 313:1438–1441. [PubMed: 16888103]
99. Rinn JL, et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*. 2007; 129:1311–1323. [PubMed: 17604720]

100. Tsai MC, et al. Long noncoding RNA as modular scaffold of histone modification complexes. *Science*. 2010; 329:689–693. [PubMed: 20616235]
101. Zhao J, et al. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science*. 2008; 322:750–756. [PubMed: 18974356]
102. Wang KC, et al. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature*. 2011; 472:120–124. [PubMed: 21423168]
103. Shen LX, et al. Single-nucleotide polymorphisms can cause different structural folds of mRNA. *Proc Natl Acad Sci U S A*. 1999; 96:7871–7876. [PubMed: 10393914]
104. Broadbent HM, et al. Susceptibility to coronary artery disease and diabetes is encoded by distinct, tightly linked SNPs in the ANRIL locus on chromosome 9p. *Hum Mol Genet*. 2008; 17:806–814. [PubMed: 18048406]
105. Ishii N, et al. Identification of a novel non-coding RNA, MIAT, that confers risk of myocardial infarction. *J Hum Genet*. 2006; 51:1087–1099. [PubMed: 17066261]
106. Akhtar-Zaidi B, et al. Epigenomic enhancer profiling defines a signature of colon cancer. *Science*. 2012; 336:736–739. [PubMed: 22499810]
107. Adzhubei IA, et al. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010; 7:248–249. [PubMed: 20354512]
108. Niknafs N, et al. MuPIT interactive: webserver for mapping variant positions to annotated, interactive 3D structures. *Hum Genet*. 2013
109. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res*. 2012; 40:D930–934. [PubMed: 22064851]
110. Boyle AP, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res*. 2012; 22:1790–1797. [PubMed: 22955989]
111. Khurana E, et al. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science (New York, N Y)*. 2013; 342:1235587.
112. Cingolani P, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)*. 2012; 6:80–92. [PubMed: 22728672]
113. 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]
114. Gilad Y, et al. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet*. 2008; 24:408–415. [PubMed: 18597885]
115. Montgomery SB, Dermitzakis ET. From expression QTLs to personalized transcriptomics. *Nat Rev Genet*. 2011; 12:277–282. [PubMed: 21386863]
116. Huang HY, et al. RegRNA: an integrated web server for identifying regulatory RNA motifs and elements. *Nucleic Acids Res*. 2006; 34:W429–434. [PubMed: 16845041]
117. Wang K, et al. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet*. 2010; 11:843–854. [PubMed: 21085203]
118. Califano A, et al. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nat Genet*. 2012; 44:841–847. [PubMed: 22836096]
119. Segre AV, et al. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet*. 2010; 6.
120. Torkamani A, et al. Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics*. 2008; 92:265–272. [PubMed: 18722519]
121. Wang K, et al. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet*. 2007; 81:1278–1283. [PubMed: 17966091]
122. Raychaudhuri S, et al. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet*. 2009; 5:e1000534. [PubMed: 19557189]
123. Voineagu I, et al. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature*. 2011; 474:380–384. [PubMed: 21614001]

124. Khurana E, et al. Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. *Science*. 2013; 342:1235587. [PubMed: 24092746]
125. Cong L, et al. Multiplex genome engineering using CRISPR/Cas systems. *Science*. 2013; 339:819–823. [PubMed: 23287718]
126. Sanjana NE, et al. A transcription activator-like effector toolbox for genome engineering. *Nat Protoc*. 2012; 7:171–192. [PubMed: 22222791]
127. Soldner F, et al. Generation of isogenic pluripotent stem cells differing exclusively at two early onset Parkinson point mutations. *Cell*. 2011; 146:318–331. [PubMed: 21757228]
128. Onder TT, Daley GQ. New lessons learned from disease modeling with induced pluripotent stem cells. *Curr Opin Genet Dev*. 2012; 22:500–508. [PubMed: 22749051]
129. Soldner F, Jaenisch R. Medicine. iPSC disease modeling. *Science*. 2012; 338:1155–1156. [PubMed: 23197518]
130. Voight BF, et al. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet*. 2012; 8:e1002793. [PubMed: 22876189]
131. Sakoda LC, et al. Turning of COGS moves forward findings for hormonally mediated cancers. *Nat Genet*. 2013; 45:345–348. [PubMed: 23535722]
132. Bahcall OG. iCOGS collection provides a collaborative model. Foreword. *Nat Genet*. 2013; 45:343. [PubMed: 23535721]
133. Morris AP, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet*. 2013; 44:981–990. [PubMed: 22885922]
134. Michailidou K, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet*. 2013; 45:353–361. 361 e351–352. [PubMed: 23535729]
135. Eeles RA, et al. Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. *Nat Genet*. 2013; 45:385–391. 391 e381–382. [PubMed: 23535732]
136. Pharoah PD, et al. GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. *Nat Genet*. 2013; 45:362–370. 370e361–362. [PubMed: 23535730]
137. Garcia-Closas M, et al. Genome-wide association studies identify four ER negative-specific breast cancer risk loci. *Nat Genet*. 2013; 45:392–398. 398e391–392. [PubMed: 23535733]
138. Bojesen SE, et al. Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. *Nat Genet*. 2013; 45:371–384. 384e371–372. [PubMed: 23535731]
139. Yang J, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*. 2010; 42:565–569. [PubMed: 20562875]
140. Dickson SP, et al. Rare variants create synthetic genome-wide associations. *PLoS Biol*. 2010; 8:e1000294. [PubMed: 20126254]
141. Wray NR, et al. Synthetic associations created by rare variants do not explain most GWAS results. *PLoS Biol*. 2011; 9:e1000579. [PubMed: 21267061]
142. Nejentsev S, et al. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*. 2009; 324:387–389. [PubMed: 19264985]
143. Network TCGAR. Integrated genomic analyses of ovarian carcinoma. *Nature*. 2011; 474:609–615. [PubMed: 21720365]
144. Network TCGAR. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 2012; 487:330–337. [PubMed: 22810696]
145. Imielinski M, et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell*. 2012; 150:1107–1120. [PubMed: 22980975]
146. Michaelson JJ, et al. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell*. 2012; 151:1431–1442. [PubMed: 23260136]
147. Baca SC, et al. Punctuated evolution of prostate cancer genomes. *Cell*. 2013; 153:666–677. [PubMed: 23622249]
148. Huang FW, et al. Highly recurrent TERT promoter mutations in human melanoma. *Science*. 2013; 339:957–959. [PubMed: 23348506]
149. Horn S, et al. TERT promoter mutations in familial and sporadic melanoma. *Science*. 2013; 339:959–961. [PubMed: 23348503]

150. Reijnen MJ, et al. Disruption of a binding site for hepatocyte nuclear factor 4 results in hemophilia B Leyden. *Proc Natl Acad Sci U S A*. 1992; 89:6300–6303. [PubMed: 1631121]
151. Bosma PJ, et al. The genetic basis of the reduced expression of bilirubin UDP-glucuronosyltransferase 1 in Gilbert's syndrome. *N Engl J Med*. 1995; 333:1171–1175. [PubMed: 7565971]
152. Benko S, et al. Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. *Nat Genet*. 2009; 41:359–364. [PubMed: 19234473]
153. Weedon MN, et al. Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nat Genet*. 2014; 46:61–64. [PubMed: 24212882]

Emerging functional tools

The bioinformatic predictions require functional validation. Finding the appropriate in vivo model system is a challenge. Cell lines that are naturally heterozygous for a genetic risk variant(s) can be used in validation experiments to attribute allele-specific effects to the risk allele(s) while controlling for the environmental effects and genetic background. However, it may be difficult to find cell lines heterozygous for low-frequency risk alleles. In addition, many cellular phenotypes, such as cell proliferation, differentiation efficiency or drug response, cannot be assessed in an allele-specific manner. Genome-editing technologies such as TALEN and CRISPR/Cas can be used to artificially generate a cell line(s) or mouse model(s) of genetic risk variants by introducing mutations into the genome [125, 126]. Functional assays aimed at assessing relevant phenotypes, such as protein structure, gene expression, transcription factor recruitment or cell proliferation/differentiation in these model systems will aid in determining the disease causality of genetic risk variants. In addition, the CRISPR/Cas system allows mutagenesis at multiple loci in one cell [125], providing the tools needed to investigate the joint effects of multiple risk variants. Another challenge is to assess the role of risk variants across disease development. While animal models can be used for this purpose, a large proportion of the functional human genome is not conserved across species. The genome-editing system can be integrated with the induced pluripotent stem cell (iPSC) technology and will allow the correction of risk alleles in patient-derived iPSC cells or the creation risk alleles in normal embryonic stem cells. Differentiation of these edited iPSC or normal stem cells into disease-relevant cell types may uncover the role of the genetic risk variants in disease development [127–129].

Perspectives

Recently, the use of custom genotyping-arrays that prioritize SNPs with promising but not compelling evidence of association from previous GWASs has led to the identification of additional trait and disease-associated variants [130, 131]. These collectively increase the proportion of the risk explained, supporting the design of follow-up candidate-SNP arrays for GWAS [132–138]. For example, the Collaborative Oncological Gene-environment Study (COGS) project used a custom genotyping array that included over 200,000 SNPs prioritized for follow-up and identified more than 70 new genetic risk variants surpassing genome-wide significance for three hormone-related cancers: breast, prostate, and ovarian [131]. However, common SNPs that fail to reach the level of association required to be considered genome-wide significant, rare variants or “synthetic” associations may also contribute to the heritability of human traits and diseases [20, 139, 140, 141]. This highlights the need of re-sequencing risk loci implicated by GWAS [142]. Integrative functional post-GWAS methodologies can be adapted to identify functional genetic variants within the population, regardless of their level of association or frequency. The identification of functional sequence variants would provide a comprehensive and effective list of candidate variants for the design of arrays for the next generation of association studies. Thus, the goal of the post-GWAS era should not only be to identify the functional consequences of known risk variants, but also to assist the identification of the additional risk- and/or trait-associated genetic variants, which may account for a portion of the missing heritability. Furthermore, a portion of the missing heritability may also be revealed through the identification of the mostly likely causal variant(s) underlying each risk locus by helping to clarify the current effect estimates, which may be underestimated.

Lessons learned through post-GWAS studies can also provide a framework to investigate the role of somatic mutations associated with diseases. Whole-genome sequencing (WGS) has identified thousands of mutations in diverse diseases [143–147], the vast majority of which map outside coding regions [143–147]. Recently, two recurrent mutations in the promoter of the *TERT* gene were described in melanomas and glioblastoma to create DNA recognition motifs for ETS transcription factors and to increase the transcriptional activity of the *TERT* promoter [148, 149]. This is reminiscent of the allele-specific changes in the activity of regulatory elements reported for genetic risk variants [74, 77–83]. Given that several inherited non-coding sequence changes are the reported genetic lesions responsible for what appear to be “single gene” disorders [150–153]. Many more similarities are anticipated for the role of non-coding somatic mutations in cancer.

In light of the systematic functional annotation of the human genome, we have witnessed the first wave of post-GWAS studies addressing the causal nature of common genetic risk variants. From these initial studies the methods required to validate the observed GWAS associations are beginning to emerge. These methodologies and tools will help us achieve a more complete understanding of the genomic alterations underlying common complex diseases and help usher forward personalized genomic medicine.

Highlights

1. Noncoding genetic risk-variants affect regulatory elements, chromatin architecture or ncRNA.
2. Integrative functional methodologies inferring the function of risk-loci are emerging
3. Functional GWAS analysis can guide next generation association studies
4. Integrative functional GWAS methods inform on how to study noncoding mutations

Figure 1. Genetic risk-variants mapping to coding regions

A. Non-synonymous genetic risk-variants alter structure and activity of encoded proteins.

A1 and A2 are the two alleles of a genetic risk variant. In all figures, A1 is the normal allele (or risk allele), and A2 is the risk allele (or normal allele).

B. Synonymous genetic risk-variants alter translational rate, further influencing protein folding. R: Ribosome.

C. Genetic risk-variants disrupt exonic splicing enhancers (ESE), further leading to exon skipping. SC: splicing complex.

Figure 2. Noncoding genetic risk variants

- A. Genetic risk variants influence DNA methylation level at promoter regions.
- B. Genetic risk variants modulate transcription factor binding to chromatin. TF: transcription factor; PF: pioneer factor; LF: chromatin looping factor.
- C. Genetic risk variants alter chromatin loop formation bridging enhancers and promoters.
- D. Genetic risk variants influence the repression effect of miRNAs. RISC: RNA-induced silencing complex.
- E. Genetic risk variants influence the interaction of lncRNAs with target proteins. NP: nuclear proteins.

Figure 3. Categories of functional genetic risk variants

Non-Sym: Non-synonymous; Sym: Synonymous; DNA methyl: DNA methylation.

Table 1

Computational tools for the post-GWAS era.

Analysis	Tools	Notes	Accessibility	Ref
Identify the functional elements targeted by genetic risk-variants	VSE; Other VSE-like tools	Permutation-based methods to identify functional elements enriched of genetic risk variant sets	Publically available software in development Methodology described in the references	[74] [8] [20] [31]
Predict protein structure changes imposed by genetic risk-variants	PolyPhen MuTIP	Predict and compare the protein structures associated with the reference and the risk allele of a genetic variant mapping to the coding region.	http://genetics.bwh.harvard.edu/pph2/ http://mupit.icm.jhu.edu/	[107] [108]
Predict changes in chromatin affinity for transcription factors caused by genetic risk-variants	IGR HaploReg; RegulomeDB; FunSeq; SnpEff	Predict chromatin affinity for transcription factors based on the raw ChIP-seq signal. Identify altered DNA recognition motifs based on position-weighted matrices (PWM) scores	Publicly available software in development www.broadinstitute.org/mammals/haploreg/ http://regulomedb.org/ http://funseq.gersteinlab.org/ Program available for download	[74] [109] [110] [124] [112]
Predict target genes for genetic risk-variants targeting regulatory elements	SCAN eQTL browser	Predict target genes by correlating expression level of candidate genes with SNPs genotypes.	http://www.scandb.org/newinterface/ http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/	[113] [15]
Predict miRNA target specificity altered by genetic risk-variants	RegRNA	Predict the binding affinity of a given miRNA in an output mRNA sequence.	http://regrna2.mbc.nctu.edu.tw/	[116]