Linking disease associations with regulatory information in the human genome

Marc A. Schaub,1 Alan P. Boyle,2 Anshul Kundaje,1 Serafim Batzoglou,1,3 and Michael Snyder2,3,4

1Department of Computer Science, Stanford University, Stanford, California 94305, USA; 2Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA

Genome-wide association studies have been successful in identifying single nucleotide polymorphisms (SNPs) associated with a large number of phenotypes. However, an associated SNP is likely part of a larger region of linkage disequilibrium. This makes it difficult to precisely identify the SNPs that have a biological link with the phenotype. We have systematically investigated the association of multiple types of ENCODE data with disease-associated SNPs and show that there is significant enrichment for functional SNPs among the currently identified associations. This enrichment is strongest when integrating multiple sources of functional information and when highest confidence disease-associated SNPs are used. We propose an approach that integrates multiple types of functional data generated by the ENCODE Consortium to help identify “functional SNPs” that may be associated with the disease phenotype. Our approach generates putative functional annotations for up to 80% of all previously reported associations. We show that for most associations, the functional SNP most strongly supported by experimental evidence is a SNP in linkage disequilibrium with the reported association rather than the reported SNP itself. Our results show that the experimental data sets generated by the ENCODE Consortium can be successfully used to suggest functional hypotheses for variants associated with diseases and other phenotypes.

[Supplemental material is available for this article.]

Genome-wide association studies (GWAS) have led to the identification of thousands of single nucleotide polymorphisms (SNPs) associated with a large number of phenotypes (Hindorff et al. 2009; Manolio 2010). These studies use genotyping platforms that measure on the order of 1 million SNPs to detect loci that have statistically significant differences in genotype frequencies between individuals who have a phenotype of interest and the general population. Although GWAS provide a list of SNPs that are statistically associated with a phenotype of interest, they do not offer any direct evidence about the biological processes that link the associated variant to the phenotype. A major challenge in the interpretation of GWAS results comes from the fact that most detected associations point to larger regions of correlated variants. SNPs that are located in close proximity in the genome tend to be in linkage disequilibrium (LD) with each other (The International HapMap Consortium 2005, 2007), and only a few SNPs per linkage disequilibrium region are measured on a given genotyping platform. Regions of strong linkage disequilibrium can be large, and SNPs associated with a phenotype have been found to be in perfect linkage disequilibrium with SNPs several hundred kilobases away. Although sequencing can be used to assess associated regions more precisely (Sanna et al. 2011), using sequence information alone is insufficient to distinguish among SNPs that are in perfect linkage disequilibrium with each other in the studied population, and thus equally associated with the phenotype.

Various approaches have been developed to identify variants that are likely to play an important biological role. Most of these approaches focus on the interpretation of coding or other SNPs in regions transcribed (Ng and Henikoff 2003; Adzhubei et al. 2010; Saccone et al. 2010). The vast majority of associated SNPs identified in GWAS, however, are in nontranscribed regions, and it is likely that the underlying mechanism linking them to the phenotype is regulatory. SNPs that influence gene expression (expression quantitative trait loci, eQTLs) (Stranger et al. 2007; Schadt et al. 2008) have been shown to be significantly enriched for GWAS associations (Nicolae et al. 2010; Zhong et al. 2010). Although eQTLs can be used to identify the downstream targets that are likely to be affected by associations identified in a GWAS, they are still based on genotyping methods and therefore also point to regions of linkage disequilibrium rather than to individual SNPs. Methods for identifying SNPs that overlap regulatory elements, such as transcription factor binding sites, are therefore necessary. Approaches based on known transcription factor binding motifs (Xu and Taylor 2009; Macintyre et al. 2010) have been successfully used to refine GWAS results and identify specific loci that have a functional role (Jarinova et al. 2009; Landers et al. 2009). However, the presence of a motif does not imply that a transcription factor is necessarily binding in vivo.

High-throughput functional assays such as chromatin immunoprecipitation assays followed by sequencing (ChIP-seq) (Johnson et al. 2007; Robertson et al. 2007) and DNase I–hypersensitive site (Gross and Garrard 1988) identification by sequencing (DNase-seq) (Crawford et al. 2006; Boyle et al. 2008) can experimentally detect functional regions such as transcription factor binding sites. Experimental evidence shows that the presence of SNPs in these regions leads to differences in transcription factor binding between individuals (Kasowski et al. 2010). A SNP that overlaps an experimentally detected transcription factor binding site and is in strong linkage disequilibrium with a SNP associated with a phenotype is thus more likely to play a biological role than other SNPs in the associated region for which there is no evidence

*These authors contributed equally to this work.
*Corresponding author
E-mail mpsnyder@stanford.edu

Article and supplemental material are at http://www.genome.org/cgi/doi/10.1101/gr.136127.111. Freely available online through the Genome Research Open Access option.
of overlap with any functional data. Several recent analyses of associated regions use these types of functional data in order to identify functional loci in individual diseases (Lou et al. 2009; Carvajal-Carnsona et al. 2011; Harismendy et al. 2011; Paul et al. 2011). A recent study of chromatin marks in nine different cell lines produced a genome-wide map of regulatory elements and showed a twofold enrichment for predicted enhancers among the associated SNPs from GWAS (Ernst et al. 2011). These examples illustrate the power of combining statistical associations between a region of the genome and a phenotype together with functional data in order to generate hypotheses about the mechanism underlying the association.

The main goal of the Encyclopedia of DNA Elements (ENCODE) project is to identify all functional elements in the human genome, including coding and noncoding transcripts, marks of accessible chromatin, and protein binding-sites (The ENCODE Project Consortium 2004, 2007, 2011). The data sets generated by the ENCODE Consortium are therefore particularly well suited for the functional interpretation of GWAS results. To date, a total of 147 different cell types have been studied using a wide variety of experimental assays (The ENCODE Project Consortium 2012). Chromatin accessibility has been studied using DNase-seq, which led to the identification of 2.89 million DNase I-hypersensitive sites that may exhibit regulatory function. DNase footprinting (Hesselberth et al. 2009; Boyle et al. 2011; Pique-Regi et al. 2011) was used to detect binding between proteins and the genome at a nucleotide resolution. ChIP-seq experiments were conducted for a total of 119 transcription factors and other DNA-binding proteins. Together these data provide a rich source of information that can be used to associate GWAS annotations with functional data.

In this work, we show that data generated by the ENCODE Consortium can be successfully used to functionally annotate associations previously identified in genome-wide association studies. We combine multiple sources of evidence in order to identify SNPs that are located in a functional region of the genome and are associated with a phenotype. We show that a majority of known GWAS associations overlap a functional region or are in strong linkage disequilibrium with a SNP overlapping a functional region. We find that for a majority of associations, the SNP whose functional role is most strongly supported by ENCODE data is a SNP in linkage disequilibrium with the reported SNP, not the genotyped SNP reported in the association study. We show that there is significant overall enrichment for regulatory function in disease-associated regions and that combining multiple sources of evidence leads to stronger enrichment. We use information from RegulomeDB (Boyle et al. 2012), a database designed for fast annotation of SNPs that combines ENCODE data sets (ChIP-seq peaks, DNase I hypersensitivity peaks, DNase I footprints) with additional data sources (ChIP-seq data from the NCBI Sequence Read Archive, conserved motifs, eQTLs, and experimentally validated functional SNPs). Using these publicly available resources makes the approach presented herein easily applicable to the analysis of any future GWAS study.

**Results**

We use linkage disequilibrium information in order to integrate GWAS results with ENCODE data and eQTLs. We call *functional SNP* any SNP that appears in a region identified as associated with a biochemical event in at least one ENCODE cell line. Functional SNPs can be further subdivided into SNPs that overlap coding or noncoding transcripts, and SNPs that appear in regions identified as potentially regulatory, such as ChIP-seq peaks and DNase I-hypersensitive sites. We call the SNPs that are reported to be statistically associated with a phenotype lead SNPs. For each lead SNP, we first determine whether the lead SNP itself is a functional SNP. If a lead SNP is in strong linkage disequilibrium with a functional SNP, then all functional SNPs that are in strong linkage disequilibrium with the lead SNP will be identified. We integrate eQTL information in a similar way, by checking whether the lead SNP or a SNP in strong linkage disequilibrium with the lead SNP has been associated with a change in gene expression.

Figure 1 illustrates our approach by describing a scenario in which a lead SNP is in strong linkage disequilibrium with a functional SNP that overlaps a transcription factor binding site, as well as with a third SNP that is an eQTL. If neither the lead SNP nor the eQTL SNP overlaps a functional region, then the functional SNP is more likely to be the SNP that plays a biological role in the phenotype than either of the SNPs that were genotyped. An extreme example would be the case in which all three SNPs are in perfect linkage disequilibrium, but only the associated SNP was present on the genotyping platform used in the GWAS in which the association was found, and only the eQTL SNP was present on the genotyping platform used in the eQTL study. In this scenario, the functional SNP would be associated equally strongly with the disease and with the change in gene expression than the reported association and eQTL SNPs, respectively. To show the potential of this approach, we analyze a set of 5694 curated associations from the NHGRI GWAS catalog (Hindorff et al. 2009) that represent...
a total of 4724 distinct SNPs associated with a total of 470 different phenotypes (for details, see Methods).

**Lead SNP annotation**

We first annotated each lead SNP with transcription information from GENCODE v7 and regulatory information from RegulomeDB. Overall, 44.8% of all lead SNPs overlap with some ENCODE data, making them functional SNPs according to our definition, and 13.1% of the lead SNPs are supported by more than one type of functional evidence. Specifically, 223 lead SNPs (4.7%) overlap coding regions, 146 (3.1%) overlap with the noncoding part of an exon, 1714 (36.3%) overlap with a DNase I peak in at least one cell line, 355 (7.5%) overlap with a DNase I footprint, and 938 (19.9%) overlap with a ChIP-seq peak for at least one of the assessed proteins in at least one cell line. Figure 2 shows the fraction of lead SNPs supported by different sources of evidence. Thus, we find that many GWAS SNPs overlap ENCODE data.

**Linkage disequilibrium**

For each lead SNP, we next located the set of SNPs that are in strong linkage disequilibrium ($r^2 \geq 0.8$) with the lead SNP in all four HapMap 2 populations, and annotate each SNP in this set. As expected, the fraction of lead SNPs in strong linkage disequilibrium with a SNP overlapping each type of functional evidence is larger than when considering lead SNPs alone (Fig. 2), and 58% of all associations are in strong linkage disequilibrium with at least one functional SNP. A similar increase can be observed for functional SNPs supported by multiple sources of evidence. We repeated the same analysis for the 2464 lead SNPs that have been associated with a phenotype in a population of European descent, using SNPs in strong linkage disequilibrium ($r^2 \geq 0.8$) with the lead SNP in the European HapMap population only. A total of 81% of the lead SNPs are in strong LD with at least one functional SNP, and 59% of the associated SNPs are in strong linkage disequilibrium with a functional SNP supported by multiple sources of evidence (Fig. 2B). A detailed breakdown for each type of functional evidence for multiple linkage disequilibrium thresholds is provided in Supplemental Tables 2 and 3.

**Integrating gene expression data**

We integrated data from multiple eQTL studies that identified SNPs associated with changes in gene expression in several tissues. A total of 462 lead SNPs (9.8%) are also themselves an eQTL in at least one tissue, and an additional 135 lead SNPs (2.8%) are in strong LD ($r^2 \geq 0.8$ in all HapMap 2 populations) with an eQTL. When considering only associations in populations of European descent, 483 lead SNPs (19.6%) are either an eQTL, or in strong LD with an eQTL. We observe that among lead SNPs that are also eQTLs, the fraction that overlaps DNase I peaks (201, 43.5%) and ChIP-seq peaks (118, 25.5%) is significantly higher than when considering all lead SNPs ($P$-values of $7.6 \times 10^{-3}$ and $1.7 \times 10^{-3}$, respectively).

**SNP comparison within linkage disequilibrium regions**

ENCODE data can be used in order to compare multiple functional SNPs that are in LD with a given lead SNP. We used a two-step approach to compare the functional annotation of two SNPs. First, if one of the SNPs is in a coding region according to GENCODE v7 and the other one is not, the coding SNP is considered to be more likely to be functional. Similarly, a SNP in a noncoding part of an exon is considered to be more likely to be functional than a SNP in an intergenic region or an intron. Second, if both SNPs are not in exons, then we compared the amount of evidence across data sources supporting the functional role of the SNP using a scoring scheme integrated in RegulomeDB (see Supplemental Methods). We hypothesized that a SNP supported by multiple types of evidence (e.g., a ChIP-seq peak and a DNase I footprint) is more likely to be functional than a SNP supported by a single experimental modality. We find that most associations where the lead SNP is in LD with at least one other SNP, the SNP with the most strongly

![Figure 2](image-url)  
**Figure 2.** Proportions of associations for different types of functional data. Proportions are shown for individual assays (A) and for all sources of evidence combined (B). Proportions are presented separately for lead SNPs and SNPs in strong linkage disequilibrium ($r^2 \geq 0.8$) with a lead SNP. For each association, we determine which SNP in the LD region is most strongly supported by functional data in order to generate the proportions in panel B. We separately consider SNPs in strong linkage disequilibrium with a lead SNP in all HapMap 2 populations, and SNPs in strong linkage disequilibrium with a lead SNP in the CEU population. For the latter case, we use only associations identified in populations of European descent, and show that we can map 80% of these associations to a functional SNP supported by experimental ENCODE data.
supported functional SNP is not the lead SNP itself, but another SNP in the LD region (22.4% compared with 13.6% when using LD in all populations, 56.8% compared with 13.6% percent when considering CEU only) (Table 1). These results show that, in most cases, the associated SNP reported in a GWAS is not the most likely to play a biological role in the phenotype according to ENCODE data.

Associations are enriched for regulatory elements

We performed randomizations in order to compare the fraction of lead SNPs that are functional SNPs or are in linkage disequilibrium with a functional SNP, to the expected fraction among all SNPs. We found that associated regions are significantly enriched for functional SNPs identified using DNase-seq and ChIP-seq. Furthermore, enrichments increased, both when integrating multiple ENCODE assays and when adding eQTL information. We used a subset of 2364 lead SNPs for which sufficient information is available and built 100 random matched SNP sets in which each lead SNP is replaced by a similar SNP (for details, see Methods). We compared the fraction of lead SNPs overlapping functional regions in the set of actual lead SNPs with the fractions observed in the random sets and computed enrichment values in order to show that the fraction of associated SNPs that overlap functional regions is higher than expected. Figure 3 provides an overview of the enrichment for different types of functional data.

When considering lead SNPs only, we observed a 1.12-fold enrichment for DNase peaks, a 1.22-fold enrichment for DNase footprints, and a 1.25-fold enrichment for ChIP-seq peaks. All enrichments are statistically significant (P-values of $1.3 \times 10^{-4}$, 0.005, and $1.3 \times 10^{-6}$, respectively). We also observed that combining multiple types of evidence increases the enrichment: There is a 1.36-fold enrichment for lead SNPs that overlap with a ChIP-seq peak, a DNase peak, a DNase footprint, and a predicted motif. Similarly, there is an 1.33-fold enrichment for eQTLs, and an even higher enrichment for eQTLs that also overlap functional regions (up to 2.4-fold).

![Figure 3. Overview of enrichment for different combinations of assays.](https://www.genome.org)

In a similar way, limiting the set of lead SNPs to the most strongly supported associations (replication in a different cohort in the original study or in multiple studies) leads to an increase in enrichment (Supplemental Fig. 1C). The enrichments can be compared with the 1.05-fold enrichment (not significant, P-value 0.887) observed when considering overlap with motif-based predictions, which do not make use of ENCODE data. When extending the set of possible functional SNPs to SNPs that are in linkage disequilibrium with a lead SNP, we observed a decrease in the enrichment (Supplemental Fig. 1A,B). At an $r^2$ LD threshold of 0.8, enrichments for most individual modalities are barely significant, but enrichment for functional SNPs supported by multiple sources of evidence remains significant (Supplemental Tables 3, 4).

Analysis at the phenotype level

In addition to considering individual associations separately, we can group associated SNPs in order to search for patterns at the phenotype level. We first assessed whether there are specific sequence binding proteins that tend to overlap functional SNPs associated with certain phenotypes more often than expected, using only associations in populations of European descent (Fig. 4). We found a strong association (P-value $= 9 \times 10^{-5}$) between height and CTCF ChIP-seq peaks. A total of 39 SNPs associated with height overlap a ChIP-seq peak or are in strong linkage disequilibrium ($r^2 \geq 0.8$ in the CEU population) with a SNP that overlaps a ChIP-seq peak, and 15 of those (38%) overlap a peak for CTCF (Supplemental Table 5), compared with 89 out of 626 SNPs (14%) when considering all phenotypes. We also found an interesting interaction between prostate cancer and the androgen receptor (AR), a transcription factor that was not assessed by ENCODE but as a control in a separate study (Wei et al. 2010). Of the nine functional SNPs for prostate cancer that overlap a ChIP-seq peak, five overlap an AR ChIP-seq peak (Supplemental Table 5). A similar analysis using DNase I assays shows that some cell line- and tissue-specific

**Table 1. Comparison of functional evidence between the lead SNP and the best SNP in the linkage disequilibrium region**

|                      | All populations | CEU only |
|----------------------|-----------------|----------|
| Only lead SNP coding | 199             | 87       | 4.21%  | 3.53%  |
| Only lead SNP transcribed, noncoding | 113   | 39       | 2.39%  | 1.88%  |
| Lead SNP supported by more regulatory evidence | 329   | 208      | 6.96%  | 8.44%  |
| **Lead better**       | 641             | 334      | 13.56% | 13.56% |
| Lead SNP and SNP in LD coding | 24    | 48       | 0.51%  | 1.95%  |
| Lead SNP and SNP in LD transcribed, noncoding | 21    | 30       | 0.44%  | 1.22%  |
| Lead SNP and SNP in LD have similar regulatory evidence | 282   | 193      | 5.97%  | 7.83%  |
| **Lead and SNP in LD equal** | 327   | 271      | 6.92%  | 11.00% |
| Lead SNP transcribed, noncoding, SNP in LD coding | 12     | 17       | 0.25%  | 0.69%  |
| Lead SNP not transcribed, SNP in LD coding | 110   | 244      | 2.33%  | 9.90%  |
| Lead SNP not transcribed, SNP in LD transcribed, noncoding | 98    | 207      | 2.07%  | 8.40%  |
| SNP in LD supported by more regulatory evidence | 356   | 456      | 7.53%  | 18.51% |
| SNP in LD annotated, lead SNP not annotated | 483   | 476      | 10.22% | 19.32% |
| **SNP in LD better**  | 1059            | 1400     | 22.40% | 56.82% |
| No annotation         | 1147            | 208      | 24.26% | 8.44%  |
| Lead SNP annotated, no SNP in LD | 1553 | 251      | 32.85% | 10.19% |

When considering a linkage disequilibrium threshold in the CEU population alone, only associations that were identified or replicated in populations of European descent are used. Boldfaced text indicates the summary of each section.
binding patterns can be observed for certain phenotypes, however, without reaching statistical significance (Supplemental Fig. 2).

 Specific examples
We identified functional SNPs in strong linkage disequilibrium for a large fraction of all reported associations. A table mapping each association to a list of candidate functional SNPs is available on our website (http://RegulomeDB.org/GWAS) and as online Supplemental Materials. Table 2 highlights the lead SNPs supported by the strongest functional evidence. These overlap a ChIP-seq peak, a DNase peak, a DNase footprint, and a predicted motif, and the transcription factor binding detected using ChIP-seq matches the conserved motif used in DNase footprinting. Table 3 provides a similar list for functional SNPs supported by the same amount of regulatory evidence, but that are in strong LD with a lead SNP. The lead SNP itself is supported by less or no evidence of a functional role. Functional SNPs in strong LD with the lead SNP are located as far as 170 kb from the reported association. Each of the functional SNPs we identify is a biological hypothesis supported by experimental regulatory data, but that still requires further validation. In this section, we describe several functional SNPs in more detail and show how ENCODE data can be used to generate interesting biological hypotheses.

 First, we show that we can re-identify a previously validated functional SNP. Lead SNP rs1541160 is associated with amyotrophic lateral sclerosis (ALS) in a GWAS, and there is no evidence that this SNP overlaps a functional region. However, it is in perfect LD with rs522444, a functional SNP overlapping DNase hypersensitivity regions and ChIP-seq peaks in a large number of ENCODE

Figure 4. Phenotype level overview of the overlap between associations and ChIP-seq binding. This matrix view shows phenotypes vertically and DNA binding proteins assessed using ChIP-seq horizontally. Each cell represents the number of lead SNPs for the respective phenotype that overlap with a ChIP-seq peak for the respective DNA binding protein or are in strong LD ($r^2 \geq 0.8$ in the CEU HapMap 2 population) with a SNP that overlaps such a peak. Only phenotypes with at least 20 lead SNPs and DNA binding proteins overlapping at least 20 functional SNPs are shown, but totals are computed over the entire data set. The significant interaction between height-associated functional SNPs and CTCF, as well as the association between prostate cancer–associated functional SNPs and androgen receptor (AR), are represented in bold font.

Table 2. Overview of the lead SNPs that are most strongly supported by functional evidence

| Phenotype                                | Total lead SNPs | PubMed ID | Rep | P-value |
|------------------------------------------|-----------------|-----------|-----|---------|
| chr1 rs1967017                            | 2a              | Serum urate                   | 20884846 | Yes | $4 \times 10^{-8}$ |
| chr5 rs2188962                            | 2a              | Crohn’s disease               | 20570966 | Yes | $1 \times 10^{-7}$ |
| chr6 rs9491696                            | 2a              | Waist–hip ratio               | 20935629 | Yes | $2 \times 10^{-12}$ |
| chr6 rs9483788                            | 2a              | Hematocrit                    | 19862010 | Yes | $3 \times 10^{-15}$ |
| chr11 rs2074238                           | 2a              | Other erythrocyte phenotypes  | 19305408 | Yes | $1 \times 10^{-47}$ |
| chr11 rs7940646                           | 2a              | Platelet aggregation          | 20526338 | Yes | $1 \times 10^{-6}$ |
| chr12 rs902274                            | 2a              | Prostate cancer               | 21743057 | Yes | $5 \times 10^{-9}$ |
| chr14 rs1256531                           | 2a              | Conduct disorder (symptom count) | 20585324 | Yes | $4 \times 10^{-10}$ |
| chr15 rs17293632                          | 2a              | Crohn’s disease               | 21102463 | Yes | $3 \times 10^{-19}$ |
| chr16 rs4788084                           | 2a              | Type 1 diabetes               | 19430480 | Yes | $3 \times 10^{-13}$ |
| chr17 rs9303029                           | 2a              | Protein quantitative trait loci | 18464913 | Yes | $4 \times 10^{-7}$ |
| chr19 rs10411210                          | 2a              | Colorectal cancer             | 19011631 | Yes | $5 \times 10^{-9}$ |
| chr19 rs3764650                           | 2a              | Alzheimer’s disease           | 21460840 | Yes | $5 \times 10^{-17}$ |

Each of these lead SNPs overlaps a ChIP-seq peak, matched DNase footprint, matched motif, and a DNase I-seq peak.
## Table 3. Overview of all strongly supported functional SNPs in linkage disequilibrium with an associated lead SNP

| Lead SNP | Phenotype | score | functional SNP in LD | CEU | CHB | JPT | YRI |
|----------|-----------|-------|----------------------|-----|-----|-----|-----|
| chr1 rs6686634 6 | Height | 2 | $r^2 = 0.8$ in all populations | 1.00 | 1.00 | 1.00 | 1.00 |
| chr1 rs380390 6 | Age-related macular degeneration | 2 | $r^2 = 0.8$ in all populations | 1.00 | 1.00 | 1.00 | 1.00 |
| chr3 rs6806528 6 | Celiac disease | 2 | $r^2 = 0.8$ in all populations | 1.00 | 1.00 | 1.00 | 1.00 |
| chr4 rs1800789 7 | Fibrinogen | 2 | $r^2 = 0.8$ in all populations | 1.00 | 1.00 | 1.00 | 1.00 |
| chr5 rs3776331 7 | Serum uric acid | 2 | $r^2 = 0.8$ in all populations | 1.00 | 1.00 | 1.00 | 1.00 |
| chr6 rs7747361 5b eQTL | Ankylosing spondylitis | 2 | $r^2 = 0.8$ in all populations | 1.00 | 1.00 | 1.00 | 1.00 |
| chr6 rs642858 6 | Type 2 diabetes | 2 | $r^2 = 0.8$ in all populations | 1.00 | 1.00 | 1.00 | 1.00 |
| chr7 rs3718685 7a eQTL | Systemic lupus erythematosus | 2 | $r^2 = 0.8$ in all populations | 1.00 | 1.00 | 1.00 | 1.00 |

This table represents associations for which there is more evidence supporting a regulatory role for a functional SNP in linkage disequilibrium with the lead SNP than for the lead SNP itself. Each functional SNP in this table overlaps a ChIP-seq peak, matched DNase footprint, matched motif, and a DNase I-seq peak. The table is separated into cases in which the functional SNP is in strong linkage disequilibrium ($r^2 < 0.8$) with the lead SNP in all HapMap 2 populations, and cases in which the functional SNP is in strong linkage disequilibrium with the lead SNP in the HapMap 2 CEU population only (and the association was identified and replicated in a population of European descent).
cell lines. The investigators in the original study identified rs522444 due to its position in a putative SP1 binding site and experimentally validated its functional role (Landers et al. 2009) in altering the expression of the gene KIFAP3.

One novel functional SNP that we identify is rs7163757 (Fig. 5). This SNP is in strong LD with rs7172432, a SNP recently shown to be associated with type 2 diabetes in the Japanese population and replicated in a European population (Yamauchi et al., 2010), and associated with insulin response in the Danish population (Grarup et al., 2011). This functional SNP is supported by evidence from both DNase I hypersensitivity and ChIP-seq assays. DNase footprinting indicates that the functional SNP overlaps a potential NFAT binding site. Interestingly, the risk allele at rs7172432 is the common allele in the population (53%), and there is a single haplotype with frequency above 1% that includes the risk allele between the associated SNP and the functional SNP, but several alleles with high frequency that include the protective allele.

A second novel functional SNP is in the 9p21 region, a gene desert that contains multiple SNPs that are strongly associated with several common diseases. Lead SNP rs1333049 has been associated with coronary artery disease in multiple studies in populations of Europeans (Samani et al., 2007; The Wellcome Trust Case Control Consortium, 2007; Broadbent et al., 2008; Wild et al., 2011) as well as Japanese and Korean descent (Hinohara et al., 2008; Hiura et al., 2008). In the HapMap 2 CEU population, this SNP is part of a haplotype block that includes rs10757278 and rs1333047, both of which are in perfect LD with rs1333049. There is no evidence in ENCODE supporting a functional role for rs1333049. However, both rs10757278 and rs1333047 overlap a DNase hypersensitivity peak as well as ChIP-seq peaks for STAT1 and STAT3 in HeLa-S3 cells. Furthermore, rs10757278 lies in a STAT1 binding site, and rs1333047 lies in a binding site and a DNase I footprint for Interferon-stimulated gene factor 3 (ISGF3). Figure 6 provides an overview of this region. Although the functional role of rs10757278 has been previously reported (Harismendy et al., 2011), evidence of the functional role of rs1333047 is novel. Interestingly, while only 27 bp separates the two SNPs, they are in perfect linkage disequilibrium in the CEU population only. The frequency of the A allele at rs1333047 in the Yoruba in Ibadan, Nigeria (YRI) HapMap 2

**Figure 5.** Functional SNP rs7163757. Multiple sources of evidence indicate that SNP rs7163757 is functional. (A) Overview of the region between genes C2CD4A and C2CD4B. (Blue vertical line) Functional SNP rs7163757; (green vertical line) lead SNP rs7172432. Multiple ChIP-seq and DNase-seq peaks can be seen, including one that overlaps rs7163757. (B) Vicinity of functional SNP rs7163757. ChIP-seq binding is observed for multiple transcription factors in multiple cell lines. Due to space, DNase peaks are represented only for a subset of the peaks overlapping the region. (C) Sequence around rs7163757 and motif for the NFAT binding site that overlaps the functional SNP. The minor allele is T. (D) Linkage disequilibrium region between the functional SNP and the lead SNP in the HapMap 2 CEU population. The two SNPs are in perfect LD (r² = 1.0). (E) Haplotypes between the functional SNP and the lead SNP. There is a single haplotype with frequency above 1% that carries the identified risk allele (A at rs7172432), whereas there are multiple haplotypes that include the protective allele. Haplotypes with frequency of <1% are not shown.
population is only 0.8%, compared with 50.8% in the CEU population. This allele is part of the protective haplotype found in GWAS performed in populations of European descent. The \( A \) allele is part of the motif for \( ISGF3 \) binding, whereas the \( T \) allele is not.

**Discussion**

In this study, we used data generated by the ENCODE Consortium to identify regulatory and transcribed functional SNPs that are associated with a phenotype, either directly in a genome-wide association study or indirectly through linkage disequilibrium with a GWAS association. We further added eQTL information, thus identifying SNPs that are associated with a phenotype, for which there is evidence that they affect a regulatory region or a transcribed region, and for which a downstream target affected by the SNP is known. This approach therefore has the potential to provide putative mechanistic explanations for GWAS associations. We showed that this method is successful in identifying a functional SNP for a majority of previously reported GWAS associations (up to 81% when considering association studies performed in populations of European descent, and using the CEU population to obtain linkage disequilibrium information).

The fraction of associated SNPs for which we can provide a functional annotation is similar to the one reported in the ENCODE integrative analysis paper (The ENCODE Project Consortium 2012). The integrative analysis uses both DNase-seq and formaldehyde-assisted isolation of regulatory elements (FAIRE) (Giresi et al. 2007) data to identify regions of open chromatin, and thus finds a slightly larger fraction of the associated SNP to overlap or be in LD with open chromatin regions compared with our approach, which does not use FAIRE data. We found that GWAS associations are significantly enriched for DNase hypersensitivity peaks, DNase I footprints, and ChIP-seq peaks even when accounting for most features of associated SNPs. Our results are consistent with chromatin state-based methods (Ernst et al. 2011), in which a segmentation approach was used in order to identify enrichment for disease associations in predicted enhancers. Segmentation-based approaches use machine learning methods to predict chromatin state at every position in the genome based mostly on histone information. These predictions are then compared with GWAS results, thus showing enrichment for predicted states. A major difference of our work is that we directly used ChIP-seq and DNase I-seq functional data in our analysis, and show enrichment for observed ChIP-seq peaks or DNase I–hypersensitive regions. In this study, we demonstrated that there is significant
enrichment of GWAS associations for these types of data. Furthermore, we found that (1) integrating multiple types of functional data and expression information identifies more likely candidate causal SNPs within an LD region, and (2) phenotypic information from GWAS studies can be associated with biochemical data.

Existing methods for prioritizing SNPs based on their functional role focused on transcribed regions (Ng and Henikoff 2003; Adzhubei et al. 2010; Saccone et al. 2010), whereas we focused on regulatory regions. In the context of regulatory regions, most approaches are based on motif information (Xu and Taylor 2009; Macintyre et al. 2010), and approaches using experimental data have generally been limited to individual associations (Harismendy et al. 2011). The comprehensive data sets generated by the ENCODE Consortium are the first to offer sufficient information to allow for genome-wide methods that rely on experimental information. We used enrichment to compare the sensitivity of our approach with motif-based methods. We found that there is no significant enrichment for GWAS associations among conserved motifs.

Identifying functional SNPs in linkage disequilibrium with lead SNPs

We found that, in most cases, there is more evidence supporting another SNP in strong LD with the lead SNP than the lead SNP itself. This is consistent with results from fine-mapping analyses that indicate that multiple variants in the linkage disequilibrium region surrounding a lead SNP appear to play a role in the phenotype of interest (Chung et al. 2011; Sanna et al. 2011). This result is of particular importance for the interpretation of GWAS results, because LD patterns differ markedly between populations. If the functional SNP is in strong LD with the lead SNP in the population in which the GWAS was performed, but not in a different population, then the lead SNP will not be associated with the phenotype in this second population. An example of this situation is functional SNP rs1333047, which lies in a region associated with coronary artery disease. This SNP is in perfect LD with two lead SNPs in populations of European descent in which the studies identifying the associations were performed, but not in populations of African descent, in which the associations could not be replicated (Assimes et al. 2008; Kral et al. 2011; Lettre et al. 2011; see Supplemental Material).

Comparison of functional assays

We integrated data from multiple types of functional assays in order to identify functional SNPs.

We found that the highest enrichments are obtained when requiring functional SNPs to be supported by multiple sources of experimental evidence rather than only one. The highest enrichments are observed when using both eQTL information and ENCODE data, and when considering associations that have been replicated. A similar trend can be observed when examining individual assays. The more specific the assay, the higher is the enrichment for overlap among GWAS associations: The DNase hypersensitivity peaks, which broadly capture regions in which chromatin is accessible, do overlap with a large fraction of SNPs in general, thus leading to relatively weak enrichments, whereas the enrichment is much higher for ChIP-seq peaks, which experimentally identify the binding of specific transcription factors and other molecules. There is a clear trade-off between the more significant enrichment we observe, and the lower fraction of associations annotated with ChIP-seq peaks. The ChIP-seq data generated so far by the ENCODE Consortium only assesses 119 transcription factors, a fraction of the 1800 known ones (The ENCODE Project Consortium 2012). Most transcription factors are assessed in a small subset of the ENCODE cell lines, whereas DNase-seq has been performed on most ENCODE cell lines. DNase footprinting, which combines DNase-seq data with sequence and motif information, is useful to identify potential binding sites for transcription factors not assessed using ChIP-seq. An example of this situation is functional SNP rs7163757, which is in LD with a lead SNP associated with type 2 diabetes. DNase footprinting identifies a nuclear factor of activated T-cells (NFAT) footprint that overlaps rs7163757. NFAT is part of the calcineurin/NFAT pathway (Crabtree and Olson 2002), which has been involved in the regulation of growth and function of the insulin-producing pancreatic beta cells, and linked to the expression of genes known to be associated with type 2 diabetes (Heit et al. 2006).

Differences between tissue types

Transcription factor binding patterns are heterogeneous and differ between tissue types. Assessing this heterogeneity has been a main motivation for the ENCODE Project. One concern is that the cell lines from which the functional information is derived do not necessarily correspond to the tissue type that is most relevant to the phenotype of interest. A similar approach has been successfully used to identify functional SNPs that play a role in coronary artery disease based on a ChIP-seq assay performed in the immortalized HeLa cell line (Harismendy et al. 2011). By choosing to use functional data across all tissues, we purposefully favor sensitivity over specificity. An example illustrating the benefits of this trade-off is rs2074238, a functional SNP associated with long QT syndrome. A ChIP-seq experiment identifies the binding of estrogen receptor alpha at this location in an epithelial cell line. Long QT syndrome is more prevalent in women (Hashiba 1978; Locati et al. 1998), the menstrual cycle affects the QT interval (Nakagawa et al. 2006), and estrogen therapy has been shown to affect the duration of the QT interval in postmenopausal women (Kadish et al. 2004; Gökc et al. 2005). ChIP-seq data for this transcription factor are only available for two cell lines, neither of cardiac origin. By limiting our approach to functional data obtained in cardiac tissues, we would have excluded a transcription factor whose role in the phenotype is supported by extensive prior evidence. When examining all associations, the significant enrichments we report demonstrate that our current approach improves specificity compared with using motif information only.

Although the ChIP-seq data generated so far by the ENCODE Consortium are sparse, especially in terms of the number of different tissues in which a transcription factor is assessed, the number of available data sets is growing rapidly. We expect that it will soon become possible to refine this approach by considering the most relevant tissue types only, thus further improving its specificity. A remaining challenge is the identification of specific tissue types that are relevant for a given phenotype. A specific example is a functional SNP we identify in the context of Alzheimer’s disease: In cell lines of hepatic origin, rs3764650 overlaps a binding site for HNF4A, a transcription factor known to mediate hepatic gene expression and linked to Alzheimer’s disease (Crabtree and Olson 2002), which has been implicated in the regulation of growth and function of the insulin-producing pancreatic beta cells, and linked to the expression of genes known to be associated with type 2 diabetes (Heit et al. 2006).
Functional SNPs beyond reported associations

In this study, we focused on using ENCODE information in order to identify functional SNPs in strong LD with previously reported associations. It is, however, important to note that these SNPs only represent a small fraction of all the SNPs that overlap functional regions identified by ENCODE. SNPs that alter transcription factor binding sites are likely to have some biologically important effect and have an impact on some phenotype. Such a SNP, however, will only be found in a GWAS if the specific phenotype it affects is assessed. Given this fundamental limitation of association studies, an orthogonal approach would be to study the functional effects of common SNPs regardless of their association with a phenotype. Furthermore, this effect explains why the enrichments we observe, while significant, are relatively modest. We used a stringent null model in which a lead SNP is matched to a random SNP that is similar to the lead SNP, and in particular located at a similar distance to the nearest transcription start site. Associated SNPs are located more closely to genes than SNPs in general, and therefore null sets are also biased toward SNPs that are likely to have some biological effect. Relaxing the null model leads to higher enrichments (Supplemental Fig. 1B,C).

Conclusion

We show that genome-wide experimental data sets generated by the ENCODE Consortium can be successfully used to provide putative functional annotations for the majority of the GWAS associations reported in the literature. The use of these experimental assays outperforms the use of in silico binding predictions based on sequence motifs when trying to identify functional SNPs associated with a phenotype in a GWAS. We demonstrate that an integrative approach combining genome-wide association studies, gene expression analysis, and experimental evidence of regulatory activity leads to the identification of loci that are involved in common diseases, and generates hypotheses about the biological mechanism underlying the association. In the majority of cases, the SNP most likely to play a functional role according to ENCODE evidence is not the reported association, but a different SNP in strong linkage disequilibrium with the reported association. Our approach, which builds directly on the publicly available RegulomeDB database, provides a simple framework that can be applied to the functional analysis of any genome-wide association study.

Methods

Data

We use the NHGRI GWAS catalog (Hindorff et al. 2009) (http://www.genome.gov/gwastudies downloaded on August 10, 2011) to obtain a list of GWAS associations. We use HapMap version 2 (The International HapMap Consortium 2007) and version 3 (The International HapMap Consortium 2010) in order to obtain linkage disequilibrium information between SNPs. HapMap data can be downloaded from http://hapmap.ncbi.nlm.nih.gov/. We use the list of SNPs that appear on genotyping arrays from the SNP Genotyping Array track of the UCSC Genome Browser (Kent et al. 2002). We use the function information generated by the UCSC Genome Browser for each SNP in dbSNP 132 (Sayers et al. 2012). We used the November 7, 2011 version of RegulomeDB (Boyle et al. 2012) in order to annotate SNPs with regulatory information and to obtain a list of eQTLs. The RegulomeDB server is available at http://www.regulomedb.org, and all ENCODE data sets used in RegulomeDB can be accessed via the ENCODE portal at http://encodeproject.org. We use GENCODE v7 (Harrow et al. 2012) to identify SNPs that overlap transcribed regions. The GENCODE v7 track can be accessed on the UCSC Genome Browser at http://genome.ucsc.edu. These data sets are described in more detail in the Supplemental Material.

Annotation

Lead SNPs

We call the associated SNP reported in a GWAS the lead SNP. For each lead SNP, we retrieve the regulatory annotation from RegulomeDB and the transcriptional annotation from GENCODE v7. We determine the fraction of lead SNPs that are coding, in noncoding parts of exons, that overlap DNase I peaks, DNase I footprints, and ChIP-seq peaks independently of each other. This means that if, for example, a SNP overlaps both a DNase peak and a ChIP-seq peak, then it will be counted for both types of assays. We consider that there is an overlap between the SNP and the type of assay if there is one ENCODE cell line in which there is, respectively, a DNase peak, a DNase footprint for at least one motif, or a ChIP-seq peak for at least one binding protein that overlaps the SNP. To determine a score for lead SNPs, we first assess whether the SNP is in an exon. If the SNP is not in an exon, then we assign the modified RegulomeDB score to this SNP (see the Supplemental Material). We use Fisher's exact test on a 2 × 2 table to compute a P-value for the difference in the fraction of functionally annotated SNPs between all lead SNPs and lead SNPs that are eQTLs.

Linkage disequilibrium

For each lead SNP, we compute the set of all SNPs in LD with that lead SNP. We first use an r^2 threshold in order to limit the LD set to SNPs in strong LD with the lead SNP. To add a SNP to the LD set, we require that the r^2 is above the threshold in all four HapMap 2 populations. We then look separately at associations found in populations of European descent. For each of these lead SNPs, we obtain a set of SNPs in LD with the lead SNP when considering the HapMap 2 CEU population only. We separately analyze the set of all lead SNPs, and the subset of European-descent lead SNPs.

To compute the fraction of SNPs in LD with a lead SNP that overlap a type of functional data, we do count every lead SNP at most once, namely, when one or more SNPs in the LD set overlap with the functional data type. To compute a score, we find the best candidate in the LD set corresponding to each lead SNP. We consider that a coding SNP had more functional evidence than a SNP in a noncoding part of an exon, and that a SNP in an exon has more functional evidence than a regulatory SNP. If no SNP in the LD set is transcribed, then we find the SNP with the best RegulomeDB score. We consider an associated region to be an eQTL if there is at least one eQTL in the set of SNPs in LD with the lead SNP.

Randomization

We create n = 100 null sets in which each lead SNP is matched to a random SNP that has a similar minor allele frequency, is present on the same genotyping platform as the lead SNP, has the same predicted function (using UCSC gene predictions), and is located at a similar distance from the nearest transcription start site. To perform these randomizations, we filter out lead SNPs for which insufficient information is available, lead SNPs that are not assessed in one or more HapMap populations, and lead SNPs that are in linkage disequilibrium with another lead SNP that is more strongly associated with a phenotype. The filtering and randomization steps are described in more detail in the Supplemental Material.
We then repeat the annotation steps on each null set and obtain an empirical distribution of the fraction of functional SNPs expected for matched SNPs, and of the score distribution among matched SNPs. We obtain a P-value for the difference between the lead SNPs and the null sets using a Student’s t distribution with \( n - 1 \) degrees of freedom and the same mean and standard deviation from the empirical distribution of the counts overlapping the feature in the \( n \) randomized null sets. This distribution is used to estimate the probability of having a null set (which is by construction of the same size as the set of lead SNPs) with a fraction of SNPs overlapping the feature that is as extreme or more extreme than the fraction observed for the lead SNP set, which results in a two-tailed P-value.

### Analysis at the phenotype level

We group all lead SNPs per phenotype using the GWAS catalog phenotype classification. We do not further group phenotypes, even though some are similar. We use only associations identified or replicated in populations of European descent. For each lead SNP, we count how many times the lead SNP or at least one SNP in strong LD \((r^2 \geq 0.8\) in the HapMap 2 CEU population) overlaps with a ChIP-seq peak for a given DNA binding protein. Each lead SNP is counted at most once for each DNA binding protein, and we ensure that no two lead SNPs are in LD with each other. We then add the totals for all of the lead SNPs associated with each phenotype. We use a Fisher’s exact test on a 2 \( \times \) 2 table to show that the fraction of lead SNPs associated with heights that are in strong LD with at least one SNP overlapping with a \( CTCF \) ChIP-seq peak is higher than the same fraction for all associated lead SNPs.

### Analysis of individual loci

We use Haploview (Barrett et al. 2005) to analyze linkage disequilibrium data and haplotype frequencies in individual regions. We obtain transcription factor binding motifs from TRANSFAC (STAT1, NFAT) and JASPAR (ISGF3). The motif representations in Figures 5 and 6 were created using WebLogo 3 (Crooks et al. 2004).

### Data access

The list of all functional SNP predictions we generate is available at http://RegulomeDB.org/GWAS and as online Supplemental Material.

### Competing interest statement

M.S. is a consultant for Illumina, a founder and member of the scientific advisory board for Personalis, and a member of the scientific advisory board for 23andMe, GigaGen, and Moleculo.

### Acknowledgments

We thank Ross Hardison, Ewan Birney, Jason Ernst, Konrad Karczewski, Manoj Harharian, and the members of the Batzoglou laboratory for suggestions and comments. We thank the anonymous reviewers for valuable feedback and suggestions. We thank the ENCODE Consortium, the Office of Population Genomics at the National Human Genome Research Institute, the HapMap Consortium, and the Genome Bioinformatics Group at the University of California–Santa Cruz for generating the data and tools used in this work. This work was supported in part by the ENCODE Consortium under Grant No. NH 5U54 HG 004558, by the National Science Foundation under Grant No. 0640211, funding from the Beta Cell Consortium, and by a King Abdullah University of Science and Technology research grant. M.A.S. was supported in part by a Richard and Naomi Horowitz Stanford Graduate Fellowship. A.K. was partially supported by an ENCODE analysis subcontract.

### References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nat Methods 7: 248–249.

Ahearn TL, Knowles JW, Basu A, Ibarraen C, Southwick A, Tang H, Absher D, LiJ, Fair JM, Rubin GD, et al. 2008. Susceptibility locus for clinical and subclinical coronary artery disease at chromosome 9p21 in the multi-ethnic ADVANCE study. Hum Mol Genet 17: 2320–2328.

Broadbent HM, Peden JF, Lorkowski S, Goel A, Ongen H, Green F, Clarke R, Barlow J, Gong J, Bittner V, et al. 2012. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res (this issue). doi: 10.1101/gr.137235.112.

Broadbent HM, Peden JF, Lorkowski S, Goel A, Ongen H, Green F, Clarke R, Collins R, Frangouzi MG, Tognoni G, et al. 2008. Susceptibility to coronary artery disease and diabetes is encoded by distinct, tightly linked SNPs in the ANRIL locus on chromosome 9p. Hum Mol Genet 17: 806–814.

Carvajal-Carmona LG, Cazier JB, Jones AM, Howarth K, Broderick P, Pittman A, Dobbins S, Tenesa A, Farrington S, Prendergast J, et al. 2011. Fine-mapping of colorectal cancer susceptibility loci at 8q23.3, 16q22.1 and 19q13.11: Refinement of association signals and use of in silico analysis to suggest functional variation and unexpected candidate target genes. Hum Mol Genet 20: 2879–2888.

Chung CC, Ciampa J, Yeager M, Jacobs KB, Broderick P, Pittman A, Dobbins S, Tenesa A, Farrington S, Prendergast J, et al. 2011. Fine-mapping of colorectal cancer susceptibility loci at 8q23.3, 16q22.1 and 19q13.11: Refinement of association signals and use of in silico analysis to suggest functional variation and unexpected candidate target genes. Hum Mol Genet 20: 2879–2888.

Crabtree GR, Olson EN. 2002. NFAT signaling: Chromoenging the social lives of cells. Cell Suppl 109: S67–S79.

Crawford GE. 2008. High-resolution mapping and characterization of DNA elements in the human genome. Nature 447: 636–640.

Crawford GE, Holt IE, Whitfield ML, Brown CJ, Ross-Innes CS, Whittle J, Webb BD, Tai D, Davis S, Margulies EH, Chen Y, Bernat JA, Ginsburg D, et al. 2006. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res 16: 123–131.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: A sequence logo generator. Genome Res 14: 1188–1190.

The ENCODE Project Consortium. 2004. The ENCODE (ENCYCLOPEDIA Of DNA Elements) Project. Science 306: 663–664.

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799–816.

The ENCODE Project Consortium. 2011. A user’s guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol 9: e1001046. doi: 10.1371/journal.pbio.1001046.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature (in press).

Ernst J, Kheradpour P, Stamatoyannopoulos JA, Kellis M, Garber M, Love MI, Meyers MA, Young J, Chakravarti A,0, et al. 2006. A source for regulatory elements from human chromatin. Nature 441: 171–176.

Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. 2007. FAIRE (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. Genome Res 17: 877–885.

Gökçe M, Karahan B, Yilmaz R, Orem C, Erdol C, Ozdemir S. 2005. Long term effects of hormone replacement therapy on heart rate variability, QT interval, QT dispersion and frequencies of arrhythmia. Int J Cardiol 99: 373–379.

Granup N, Övervad M, Sparzo T, Witte DR, Pröinger C, Jorgensen T, Yamauchi T, Hass K, Maeda S, Kadowaki T, et al. 2011. The diabetogenic VPS13C/C2CD4A/C2CD4B rs7172432 variant impairs glucose-stimulated insulin release. PLoS ONE 21: 1001046.
response in 5,722 non-diabetic Danish individuals. *Diabetologia* **54**: 789–794.

Gross DS, Garrard WT. 1988. Nucleus hypersensitive sites in chromatin. *Annu Rev Biochem* **57**: 139–197.

Hartfuss M, O, Naundorf D, Sonn A, Rahim GN, Tanasa A, Heintz N, Ren B, Fu XD, Topol EJ, Rosenfeld MG, et al. 2011. 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. *Nature* **470**: 264–268.

Harro J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski E, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* (this issue). doi: 10.1101/gr.135350.111.

Hashiba K. 1978. Hereditary QT prolongation syndrome in Japan: Genetic analysis and pathological findings of the conducting system. *Ipn Circ J** **42**: 1133–1150.

Heit J, Apelqvist AG, Gu X, Winslow MM, Neilson JR, Crabtree GR, Kim SK. 2006. Calcineurin/NFAT signalling regulates pancreatic β-cell growth and function. *Nature* **443**: 345–349.

Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS, et al. 2009. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods* **6**: 283–289.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association studies for human diseases and traits. *Proc Natl Acad Sci USA* **106**: 9362–9367.

Hinojara K, Nakajima T, Takahashi M, Hohda S, Sasaoka T, Nakahara K, Chida K, Sawahee M, Aitman T, Sato A, et al. 2008. Replication of the association between a chromosome 9p21 polymorphism and coronary artery disease in Japanese and Korean populations. *J Hum Genet** **53**: 357–359.

Hiura Y, Fukushima Y, Yuri M, Sawamura H, Kokubo Y, Okamura T, Tomoike H, Goto Y, Noroigi H, Takahashi R, et al. 2008. Validation of the association of genetic variants on chromosome 9p21 and 1q41 with myocardial infarction in a Japanese population. *Circ** **72**: 1213–1217.

The International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437**: 1299–1302.

The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**: 851–861.

The International HapMap Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–1073.

Jartinova O, Stewart AF, Roberts R, Wells G, Lau P, Naing T, Buerki C, McLean BW, Cook RC, Parker JS, et al. 2009. Functional analysis of the human genome. *Science** **323**: 1497–1502.

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**: 1497–1502.

Kadish AH, Greenland P, Limacher MC, Frishman WH, Daugherty SA, Schwartz JB. 2004. Estrogen and progestin use and the QT interval in postmenopausal women. *Ann Noninvasive Electrocardiol*** **9**: 366–374.

Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, The International HapMap Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061–1073.

Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, Hhabegger L, Rozowsky J, Shi M, Urban AE, et al. 2010. Variation in regulatory information for disease associations. *Proc Natl Acad Sci USA* **107**: 9860–9865.

Kashuk JA, Stashenko P, Davis BL, Brown MT. 1988. Estrogen receptor stimulation of uterine estradiol-binding protein messenger RNA synthesis and estradiol-induced protein-DNA binding. *J Biol Chem*** **263**: 3533–3540.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The Human Genome Browser at UCSC. *Genome Res* **12**: 996–1006.

Kernohan JW, Smith M, Semathunmo J, Mistry SD, Feigenson D, Blackford AL, et al. 2010. A genome-wide association study of 14,000 cases of seven common diseases and 3,000 matched controls. *Nature Genet*** **42**: 3812–3818.

Kernohan JW, Smith M, Semathunmo J, Mistry SD, Feigenson D, Blackford AL, et al. 2010. A genome-wide association study of 14,000 cases of seven common diseases and 3,000 matched controls. *Nature Genet*** **42**: 3812–3818.

Kernohan JW, Smith M, Semathunmo J, Mistry SD, Feigenson D, Blackford AL, et al. 2010. A genome-wide association study of 14,000 cases of seven common diseases and 3,000 matched controls. *Nature Genet*** **42**: 3812–3818.

Kernohan JW, Smith M, Semathunmo J, Mistry SD, Feigenson D, Blackford AL, et al. 2010. A genome-wide association study of 14,000 cases of seven common diseases and 3,000 matched controls. *Nature Genet*** **42**: 3812–3818.

Kernohan JW, Smith M, Semathunmo J, Mistry SD, Feigenson D, Blackford AL, et al. 2010. A genome-wide association study of 14,000 cases of seven common diseases and 3,000 matched controls. *Nature Genet*** **42**: 3812–3818.