

Chapter 2

Exploring the Landscape of the Genome

Michael R. Barnes

Abstract

Genome browsers are powerful tools for biologists – offering fundamental information on genes, regulatory elements, genomic variants, genome structure, and evolution. The comprehensive range of information presented in tools such as the UCSC genome browser and Ensembl enables integrated queries of data that are otherwise reserved to the most skilled computational biologists. However, for the non-specialist user, the juxtaposition of so many different forms of data in one small space can be an information overload. Getting the most out of these tools requires some understanding of the key concepts and caveats of genome visualization and annotation. Genome analysis can be carried out at different levels of detail – at a macro level; it improves understanding of issues like genome structure and species evolution. While at a micro level, genome annotation can help to describe the full complexity of gene regulation, variation, and transcript diversity. Once demystified, it is clear that genome browsers are more than the sum of their parts – they are the most comprehensive portals available for browsing and analysis of biological data.

Key words: Genome, Bioinformatics, Variation, Gene, Regulation, FTO, Evolution

1. Introduction

To understand genes and their role in the biology and the genetics of an organism, it is necessary to understand genome sequences. A good familiarity with the landscape and mechanics of the genome can really help in the study of biology. Genomes are pertinent to the study of many different types of data, for example, in the case of genetic variation, a single sequence variant could impact function at many levels, including gene function, gene regulation, splicing, genomic stability or epigenetic modification, or indeed all, or some of these in combination. With this in mind, this review will focus on the study of genetic variation in a genomic context purely as an illustration of the range of analysis that is possible using genomic information and the tools that are used to

analyze genomic data. These principles can be generalized to any form of data that can be mapped to a genome.

Although there are many ways to access genomic information in an integrated manner, there are two primary tools that are the acknowledged leaders in the field, the UCSC human genome browser (1) and ENSEMBL (2). Although both tools have many similarities, each contains distinct information and data interpretation, and so it usually pays to consult both viewers, if only for a second opinion (both viewers provide reciprocal links). The UCSC genome browser has one great advantage over Ensembl for macro scale genome analysis as it allows detailed visualization across regions greater than 1 Mb or even whole chromosomes. This really makes the UCSC browser an exceptional tool for integrated genomic analysis, and so most examples given below focus on this tool, but almost all the examples are possible to complete using either tool.

2. Materials

All the tools described here are freely available internet web tools, which would run on any PC, Mac or Unix workstation with web access. For more sophisticated analysis of large datasets on a genomic scale, see (3). A list of genomic tools and databases mentioned and used in this review is given in Table 1.

3. Methods

3.1. Representing User Data in the UCSC Genome Browser

The UCSC genome browser allows the user to easily represent data in a genomic context with the *custom track* or *genome graph* tools. While both these mechanisms can be used to represent quite complex data, at the most basic level, a lot can be achieved with a very simple tab delimited format. So for example, the *genome graph* function can be used to represent the results of a genetic association analysis across a region by simply uploading a list of SNP ids (which are mapped to the genome by the browser) and $-\log p$ values in a tab delimited format. An example of this format is given below:

SNP ID	Log p value
RS1477196	0.824065115
RS1121980	6.490366087
RS7193144	7.841293827
RS16945088	0.737468229
RS8050136	7.698837565
RS9926289	6.682249971

SNP ID	Log p value
RS9939609	7.28029591
RS9930506	5.856133069
RS1115005	1.094787167
RS11075994	0.225325403

Table 1
Tools for genomic characterisation

Tool	URL
<i>Genome visualization</i>	
UCSC Genome Browser	genome.ucsc.edu
ENSEMBL	www.ensembl.org
NCBI MapViewer	www.ncbi.nlm.nih.gov/mapview/map_search.cgi/
<i>LD and haplotype data</i>	
HapMap website	www.hapmap.org
HapMap Genome Browser	www.hapmap.org/cgi-perl/gbrowse/gbrowse/
SNAP	www.broad.mit.edu/mpg/snap/
<i>Structural genome analysis</i>	
Db of Genomic Variants	projects.tcag.ca/variation/
Structural Variation db	humanparalogy.gs.washington.edu/
<i>Building biological rationale</i>	
GNF SymAtlas	symatlas.gnf.org/SymAtlas/
HUGE Navigator	www.hugenavigator.net/
Stanford SOURCE	source.stanford.edu
STITCH (Pathways)	stitch.embl.de/
UniProt	www.uniprot.org

The *Genome Graphs* input form can be accessed from the left hand menu on the UCSC home page (Table 1). Once the desired genome has been selected, press the upload button, enter the details of the data and paste the text into the genome graph data

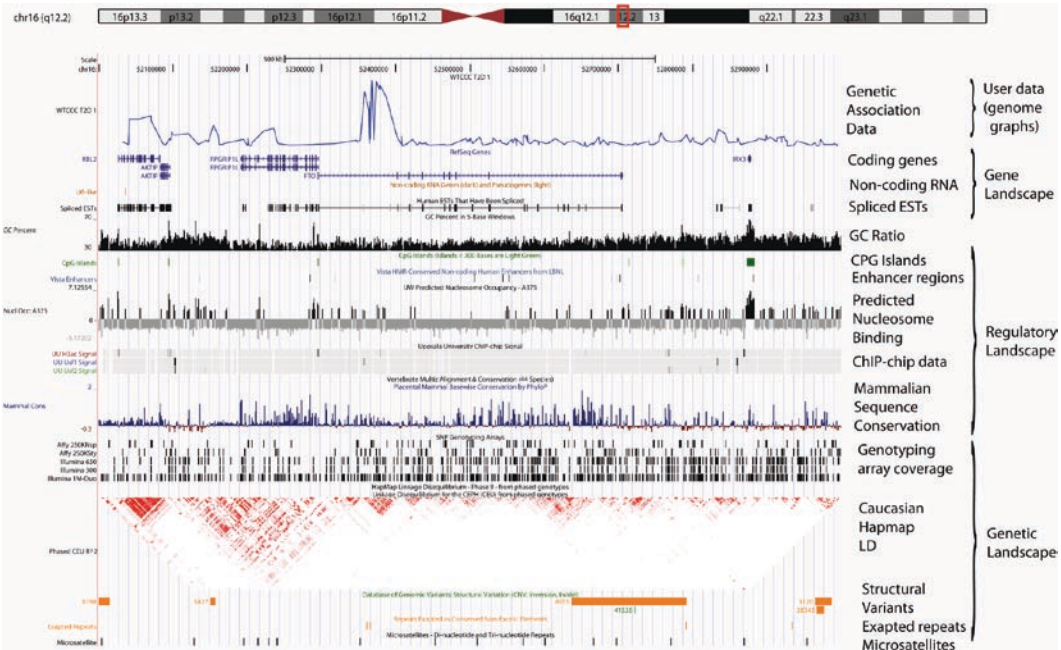


Fig. 1. The Genomic Macro-Environment. A view of the 1 Mb region (52–53 Mb) around the FTO gene, generated with the UCSC human genome browser. User submitted genetic association data is displayed using the UCSC Genome Graph and Custom Track functions. Genetic association with type II diabetes is plotted across the region ($-\log p$ values) and shows that the association is restricted to the FTO gene. The macro HapMap LD structure across the region also supports this. Descriptive information for each UCSC dataset can be accessed by pressing the grey button in the UCSC browser to the left of each track

window. After loading the data, a chromosome ideogram is returned. Select your graph from the pull down menu and the $-\log p$ values are annotated as a graph across the Chr. 16 ideogram. If the “browse regions2 button” is followed, then the data is displayed in the genome view as shown in Fig. 1. For more information about the Genome Graph function, see the UCSC help documentation (see Note 1).

Similarly, the UCSC *Custom Track* feature can be used to annotate a list of SNPs or any other genomic features by providing chromosome number, start and end genome coordinates and optionally a name. For example to view all the SNPs that are in LD with an associated SNP across a genomic region (see Note 2), the following format can be used:

Chr	Chr start	Chr end	Name
Track name=LD_SNPs description="SNPs_in_LD_with_associated_SNP"			
chr22	20100000	20100001	RS5346536
chr22	20100011	20100012	RS346976
chr22	20100215	20100216	RS2658758

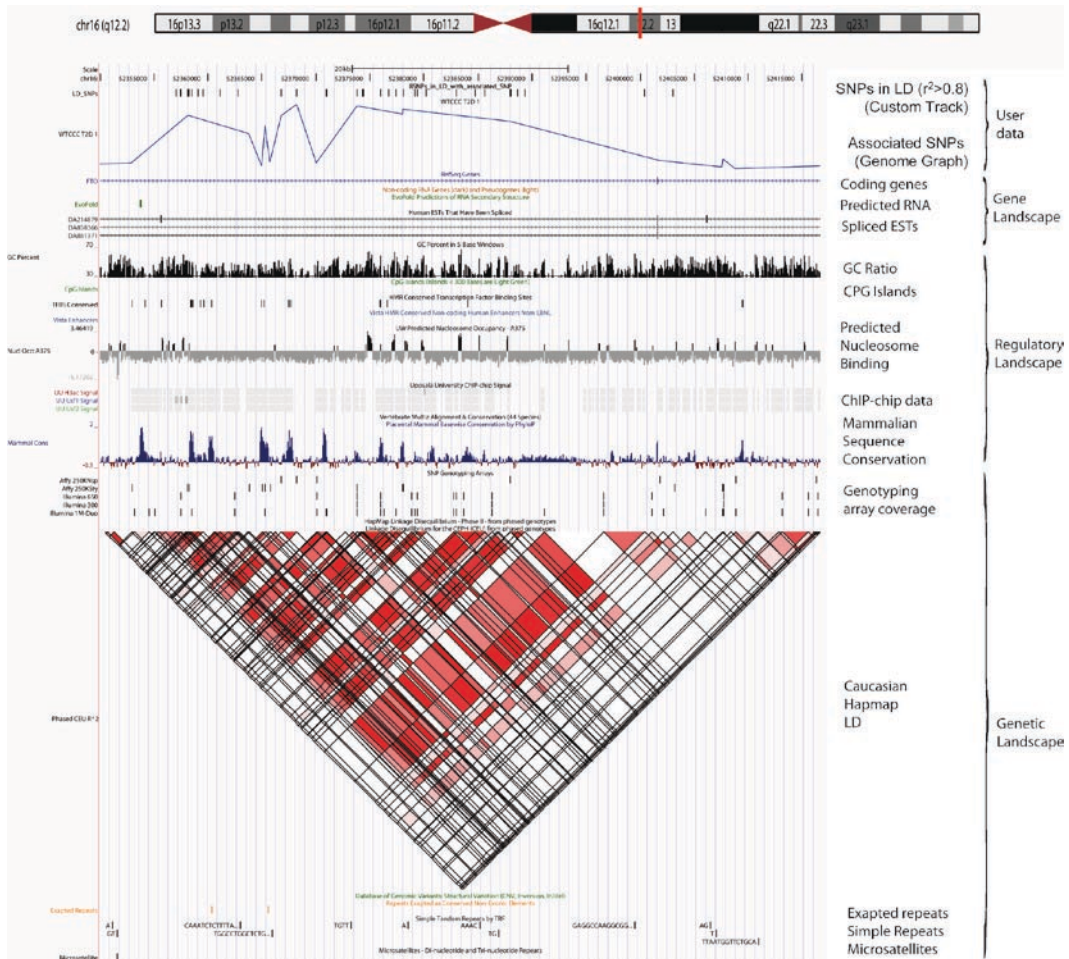


Fig. 2. The Genomic Micro-Environment. A closer view of the 100 Kb associated region in intron 1 of the FTO gene, generated with the UCSC Human genome browser. User submitted genetic association data is displayed using the UCSC Genome Graph and Custom Track functions. This view shows more detail of known regulatory elements across the region and allows the user to identify variants in these regions

After pasting this text into the browser window, the SNPs of interest are annotated in the genome view (see Fig. 2). For more information about custom track formats, see the UCSC help documentation (see Note 3).

3.2. Evaluating a Genetic Association at a Genomic Level

The custom track and genome graph facilities make the UCSC genome browser a powerful tool for evaluating the genomic context of a genetic association. A new generation of genome-wide association studies (GWAS) is revolutionizing our understanding of human disease genetics (4), and so it seems fitting to use the output of such a study as an example. To demonstrate the process in Figs. 1 and 2, a genome graph is used to plot type II diabetes

(T2D) association data across the fat mass-and obesity-associated gene (FTO) (5). This region of chromosome 16 has been reproducibly associated with fat mass and body mass index (BMI), risk of obesity, and adiposity, however no clear molecular mechanism for the action of FTO in obesity has been determined (see Note 4). By placing the association data into a genomic context, a clearer picture of the nature of this association emerges. As the amount of information available in a genome browser can be bewildering, it is often beneficial to consider a genomic region in both terms of the macro and micro-environment. The genomic *macro-environment* (Fig. 1) informs on the overall physical structure of the region, including the GC ratio, long range LD, recombination rate, structural variants and overall gene content. Gaining an understanding of the wider region can help with further study design and the interpretation of data generated from the region, e.g. the presence of large structural variants in a region would need to be factored into primer design or the interpretation of expression or association data. The genomic *micro-environment* (Fig. 2) encompasses all the features present in the immediate region around the most strongly associated SNPs (and SNPs in LD with these SNPs). These features can usually be defined at a sequence level and are immediately relevant to the regulation or function of a gene. In the sections below and in Figs. 1 and 2, the Type II diabetes association across the FTO gene is considered at both levels.

3.3. Evaluating Genomic Information: The Genomic Macro-Environment

In Fig. 1, a UCSC genome browser view of the 1 Mb region (52–53 Mb) around the FTO gene is presented with a number of tracks that highlight the properties of the macro-environment of the genetic association with type II diabetes. Firstly, the genome graph function is used to plot the $-\log p$ value of the T2D GWAS. A custom track is also used to annotate all SNPs in LD ($r^2 > 0.5$) (see Note 2) with the most strongly associated SNP in the region (RS7193144). Descriptive information for each UCSC dataset can be accessed by pressing the grey button in the UCSC browser to the left of each track. A great deal of configurable extra information is also available but not shown here for brevity (see Note 5).

In order to evaluate the initial association, the UCSC has a track which shows the SNP coverage of the major genotyping panels. The WTCCC diabetes study was completed with the affymetrix 500 K platform, which is actually composed of two chips each with 250K SNPs, these chips are presented in the “Affy 250KNsp” and “Affy 250KSty” tracks. This shows relatively good coverage across the entire region with the notable exception of a large gap in coverage immediately to the 5' (left) of the association signal. This lack of coverage should be considered when the association is evaluated – as there is no marker coverage over the first exon or the promoter region of FTO – further follow-up

studies of this association would need to provide better coverage of this region. Incidentally, coverage by the Illumina 550 K genotyping panel is also displayed and this seems to provide adequate physical coverage of the entire region.

Marker coverage across a region also needs to be considered in terms of capture of variation by LD (so called SNP tagging) rather than physical spacing alone. There have been several good comparisons of the capture of variation by commercial marker panels (6,7). Tools on the HapMap website (Table 1) also present comprehensive web-based views of LD and haplotype structure; however, they offer limited genomic information. For the purposes of an initial evaluation of the LD around a genetic association, the UCSC genome browser excels on a number of levels. Both Ensembl and the UCSC genome browser offer an integrated view of HapMap LD data, however, the UCSC browser allows visualization of LD across regions of greater than 1 Mb or even whole chromosomes. This is demonstrated in Fig. 1, where the Macro LD structure across the 1 Mb region containing the FTO gene is shown. This clearly delineates LD into Blocks of high LD punctuated by recombination hotspots, which are also displayed, based on calculations from HapMap data. From this data, it looks likely that the FTO association is restricted to an LD block in intron 1 of the gene (see Note 6). Zooming in closer to a 100 Kb view in Fig. 2, this correlation is even clearer and is also backed up by the map locations of the SNPs that are known to be in LD with the most highly associated SNP. All this information points strongly to the involvement of a genetic variant that is present in a restricted region shown in Fig. 2.

3.4. The Genomic Micro-Environment: The Nuts and Bolts of Gene Function

After defining a locus of interest, one of the key questions to ask is – what genes are located in the locus? A genome viewer is the best tool to ask this question, if known genes are all that is required then the answer is routine, but if a comprehensive answer is needed – all known and novel genes, and all transcript variants of these genes – then analysis is non-trivial. The UCSC human genome browser and Ensembl both run the human genome sequence through sophisticated gene prediction and sequence mapping pipelines (1,2). Genome viewers offer a comprehensive view of supporting evidence for genes, such as ESTs, CPG islands and both predicted and experimentally determined regulatory regions. Homology with the constantly expanding collection of genomes from other species is also presented. At the time of writing (March 09), 43 vertebrate genomes were mapped against human sequence in the UCSC genome browser. It is important to be aware of the provenance of the data presented – in effect genome annotations can be viewed as a hierarchy of evidence, with known genes at the top, hypothetical genes, spliced ESTs, sequence conservation and finally unspliced ESTs at the bottom.

Ideally, most genes should be evidenced by several of these features, e.g. a spliced EST supported by vertebrate sequence conservation is fairly reliable supporting evidence for a novel gene. Improvement on the quality of annotation provided by Ensembl and the UCSC requires an in-depth understanding of the intricacies of gene prediction, which are not within the scope of this review. Instead, it is probably best to focus on the available data to build gene models based on existing annotation. Once all the genes in the locus have been identified, the next logical step would be to investigate each gene for involvement in the phenotype being studied (see Note 7). Searching the literature can give some clues about gene function and the likelihood of involvement in a specific phenotype.

Returning to the FTO case study, let us review the genetic association signal (Fig. 2). This should be considered to encompass the associated SNPs plotted in the genome graph and also the SNPs in LD with the associated SNPs – so called proxy SNPs. Comparison of the location of the associated SNPs and the proxy SNPs against genes, ESTs and non-coding RNA, appears to restrict the association to intron 1, part of intron 2 and possibly exon 2 of the FTO gene. Reviewing the known gene information, all the associated SNPs and proxy SNPs are intronic. It is also worth reviewing EST data for evidence of novel splice variants. In this case, there is an EST (DA214879) that may represent a novel FTO exon leading to a novel splice variant. Again there are no SNPs in this EST. The magnitude and repeated replication of the association signal in the FTO region, suggest the involvement of a common variant (8). As there are no exonic variants showing association or LD with associated SNPs, it seems reasonable to assume that the causal variant(s) are likely to be intronic or alternatively there may be as yet un-characterized variants, for example copy-number variants or repeat sequences. Genome browsers are ideal tools to enable further exploration of some of these possible hypotheses to explain the functional nature of the FTO association, helping to formulate lab testable hypotheses.

3.5. The Regulatory and Epigenetic Landscape

If the FTO causative alleles are most likely to be restricted to introns, then it is clearly important to evaluate the regulatory landscape of the associated region. The traditional view of gene regulation usually focuses on the promoter region of a gene. However, regulatory sequences can be located throughout a gene, in the 5' regions, the introns, exons, splice boundaries and 3' untranslated region (Fig. 3). Regulatory elements may also take many forms, including highly specific transcription factor binding sites, or extended enhancer regions controlling tissue-specific expression or alternative splicing (9).

A key field which is helping to shed light on the basis of gene regulation is the emergent science of Epigenomics – the study of

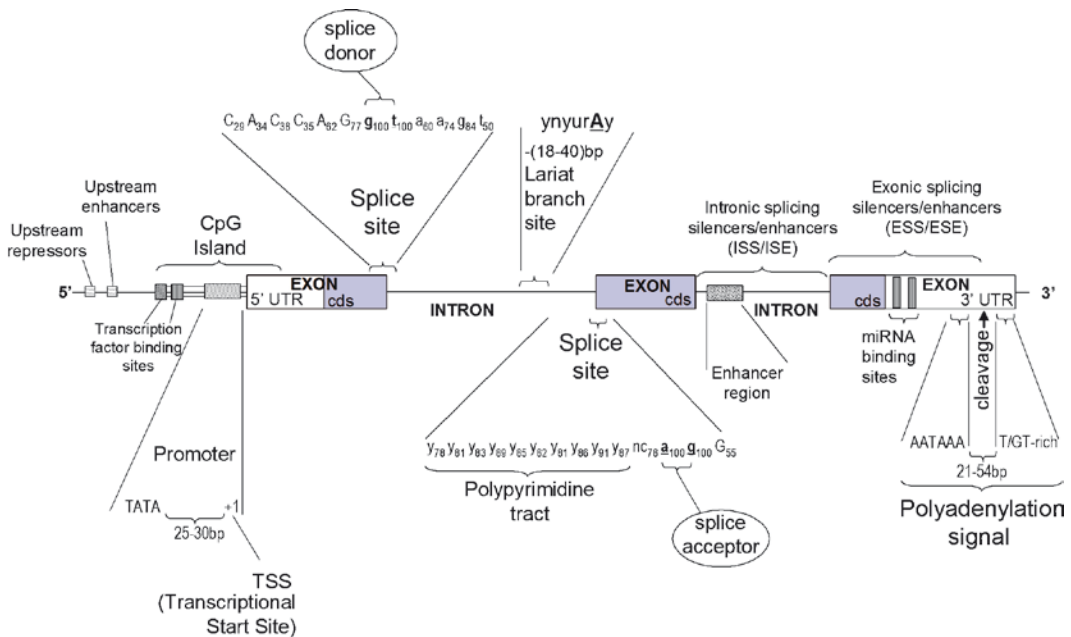


Fig. 3. The anatomy of a gene. This figure illustrates some of the key regulatory regions, which control the transcription, splicing, and post-transcriptional processing of genes and transcripts. Polymorphisms in these regions should be investigated for functional effects

epigenetic modification on a genome-wide scale. Epigenetics is concerned with the study of heritable changes other than those in the DNA sequence and encompasses two major modifications of DNA or chromatin: DNA methylation, the covalent modification of cytosine, and post-translational modification of histones including methylation, acetylation, phosphorylation and sumoylation (10). In terms of function, epigenetic modifications act to regulate gene expression and stabilize adjustments of gene dosage, as seen in X inactivation, gene silencing and genomic imprinting.

3.6. Epigenetic Insight into Gene Regulation

At the most basic level, the sequence composition of a specific region of DNA can give some clues about its regulatory potential. Inside the nucleus, DNA is wrapped into a complex molecular structure called chromatin, which is composed of a fundamental unit of approximately 150 bp of DNA organized around an eight-histone protein complex known as the nucleosome. The local organization of nucleosomes defines the accessibility of DNA to protein binding and hence the regulatory potential of a region. An excellent example of the role of the nucleosome in gene regulation was reviewed by Costello and Vertino (11) based on the work of Futscher et al. (12). This example is based on studies of the tissue-specific regulation of SERPINB5 which is controlled at the level of the nucleosome by the methylation and acetylation state of the promoter region of the gene. This regulatory mechanism

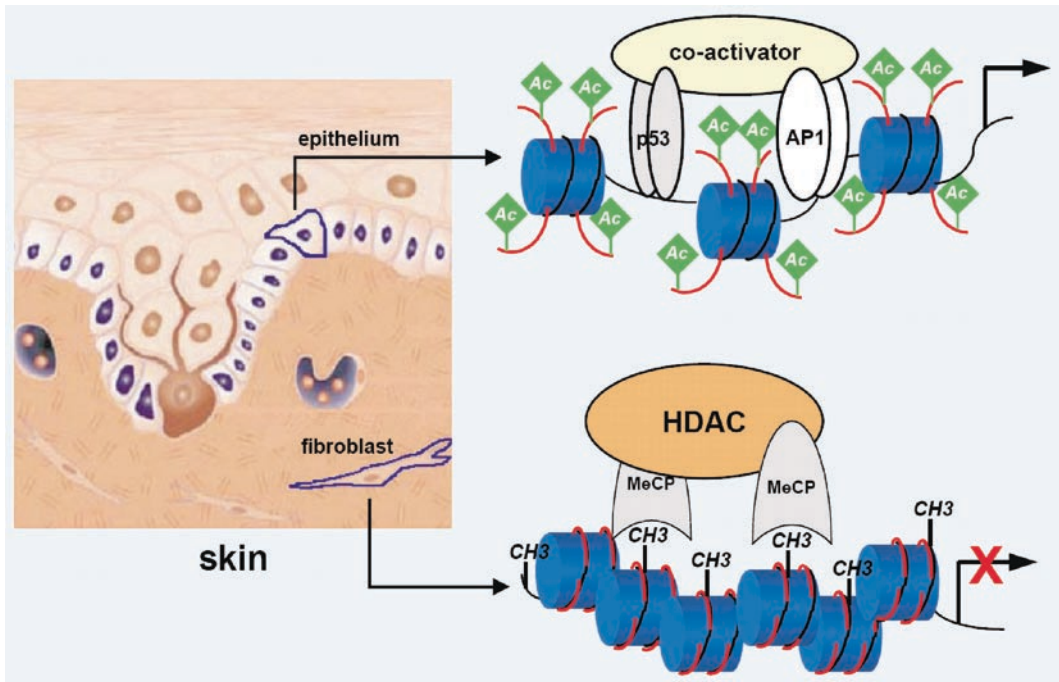


Fig. 4. Epigenetic control of SERPINB5 tissue-specific expression. Expression is mediated by methylation leading to the opening and closing of chromatin structure. (reproduced from (11) with permission from Nature publishers)

is equally applicable to regulatory elements in introns. Figure 4 shows a model of the tissue-specific control of SERPINB5 expression by methylation leading to the opening and closing of chromatin structure. The SERPINB5 promoter is unmethylated in skin epithelial cells allowing the sequence specific occupation by the transcription factors AP1 and p53. In addition, the histones bound in that region are acetylated (Ac), limiting histone–histone interactions and opening up the chromatin structure to allow binding by other transcription factors required for SERPINB5 expression. By contrast, in skin fibroblasts, the promoter is completely methylated (CH₃), this is associated with hypoacetylated histones and adopts a tighter inaccessible state that is transcriptionally inactive. DNA Methylation is a key to this model as it allows the binding of methyl CpG-binding proteins (MeCP), which mediate histone deacetylase (HDAC) and chromatin remodeling complexes to direct the compression of the chromatin structure into the transcriptionally inactive state. In this model, methylation is a primary impediment to SERPINB5 expression and thus determines the cell type-specificity. This is a good example, where the consideration of epigenetics could help genetic analysis. SNPs in CpG sites could lead to loss or gain of cytosine–guanine dinucleotide (CpG) methylation sites – and hence an

indirect impact on regulation at nearby sites. Rakyan et al. (13) suggested that CpG polymorphisms might affect the overall methylation profile of a locus and, consequently, promoter activity and gene expression. Alternatively, a non-CpG SNP located within an epigenetically sensitive regulatory element could also influence the epigenetic makeup of that region. Therefore, mutations in regulatory sequences could influence epigenetic profiles, resulting in altered phenotypes.

Moving back to the FTO case study, several UCSC tracks included in Figs. 1 and 2 give some indication of the epigenetic environment and hence the regulatory potential of the associated region. Examining Fig. 1, G/C nucleotide ratio is plotted across the region. Extended GC rich regions of the genome, known as CpG islands, are also shown. These usually correlate with gene promoter regions – this region is no exception, and it is possible to see a clear correlation between CpG islands and the start of genes in the region. As Fig. 1 shows, the GC ratio of a DNA region is also somewhat predictive of nucleosome occupancy, but GC ratio alone is a crude measure, so the UCSC browser also has a track with predicted nucleosome occupancy scores produced by a cell-line trained model (14). Aside from the predicted data, the UCSC browser also presents several valuable epigenomic data sets. These include a number of ChIP on chip data sets, representing laboratory observed nucleotide binding by specific transcription factors. In Fig. 1, data is displayed for three transcription factors generated by the University of Uppsala (15). It is notable that binding sites for M3ac and Usf1 are present in the FTO associated region. Data is also presented on known enhancer elements in the Vista enhancers track (16). This is a fascinating genome-wide set of enhancer regions that show super-conservation (>99% conservation) over 100–250 bp in human, mouse and rat. Pennacchio et al. (16), showed that when inserted upstream of a lacZ construct these enhancer regions drove highly tissue-specific expression. Enhancers in grey showed no activity in constructs, while enhancers in black drive tissue-specific expression. By clicking on each of the enhancer elements, it is possible to view the in situ expression information for each enhancer. From Fig. 1, it is clear that none of these enhancers fall in the region of association, however there are a remarkably large number of active (black) enhancers across the larger region. Three are located within the FTO gene. The closest to the association is in intron 7 of the FTO gene. Interestingly, this ultra-conserved enhancer element was shown to drive hindbrain specific expression in mouse embryos (see Note 8). Regions arising from the embryonic hindbrain in adults are known as a key region for mediation of appetite and satiety. In genome-wide terms, these enhancers are quite rare and so although there is no direct evidence that the association extends across this region, further investigation would clearly be sensible.

For example, the association might be linked to a structural variant which could extend across the enhancer element.

Perhaps the most fundamental source of information which can be used to infer genome function is conservation. In Fig. 1, mammalian conservation determined from alignment of 43 different species is plotted across the genome. Sequence conservation is a universal measure of preserved function caused by evolutionary constraint. Conservation is usually highest in coding exons of genes; however, high levels of conservation are also seen in promoter regions and other regulatory regions, like the enhancer regions discussed above. A quick scan across the sequence conservation in Fig. 1 reveals high conservation across exons, but there are also conserved sequences across the entire region. A review of the conservation across the associated region in Fig. 2 shows several intronic regions that appear to be more highly conserved than exon 2. These are clearly of interest and might be considered for further *in silico* and laboratory investigation.

3.7. The Variant Landscape

Once all the genes and regulatory features in the region have been identified, the next step is to determine how variants in the region might impact function, explaining the association. As SNP genotyping is the technology of choice for most genetic association studies, accordingly a large amount of the information presented at the UCSC relates to known SNPs and HapMap LD data. However, SNPs are not the only form of variation and a great deal of information is also available relating to non-SNP variants, such as structural variants, microsatellites and other repeat sequences. One track is available which maps all published structural variants in the *Database of Genomic Variants* (17). Until recently, structural variants in the human genome were rarely reported, but several studies help us to appreciate the contribution that copy-number variants (CNVs) may be making to clinical phenotypes (18,19). Identifying and evaluating the impact of a CNV is quite a complex process, and determining the true impact of CNVs is likely to be a big challenge for genetics in the coming years (20). In the case of the FTO case study, examining the wider FTO region, several structural variants are present, although none appear to be located in the region of association. Some information is also presented on other types of repeat sequences, in the context of the FTO association. One of the most interesting is the exapted repeat track. This track displays conserved non-exonic elements that have been deposited by mobile elements, these regions were identified during a genome-wide survey (21) with the expectation that regions of this type may act as distal transcriptional regulators for nearby genes. A previous case study experimentally verified an exapted mobile element acting as a distal enhancer (22). It is tempting to speculate that exapted repeats in the FTO locus may also play some sort of enhancer role.

3.8. Dealing with “Personal Genome” Data

One of the weaknesses of the sequence based view of the genome is that a single sequence does not effectively represent the full dynamic range of variation that may be seen within and between populations. The human genome sequence represented in genome browsers like the UCSC and Ensembl is actually a composite sequence generated from several individuals. With the rise of next-generation sequencing technologies (23), there are now several projects completed or underway that are resequencing individual genomes (24). The most high profile “individual genome” sequences have been those generated for James Watson (25) and Craig Venter (26). These projects are now being overshadowed by the “1000 genomes” project which seeks to re-sequence the genomes of 1,000 individuals around the world (<http://www.1000genomes.org/>). The Ensembl and UCSC genome browsers have already developed views of individually sequenced genomes. In the case of Ensembl, a *Resequencing Alignment View* is available which presents the sequences of James Watson, Craig Venter and four other anonymous individuals across a user defined genomic region (http://www.ensembl.org/Homo_sapiens/sequencealignview?). In Fig. 5, a small region of the FTO gene is shown with an SNP highlighted in grey. Intriguingly, this shows a tri-allelic SNP position that is not represented in dbSNP, the human genome reference sequence (REF:36) shows a T allele shared with two of the anonymous individuals. The other two anonymous individuals have an A allele, while Craig Venter carries a C allele and James Watson has the A/C ambiguity call, M, showing a heterozygote A/C call at this base. As more individual sequence data becomes available, this type of view may become an increasingly important consideration in the study of any genomic region.

3.9. Using UCSC Custom Tracks and Table Browser to Intersect Genomic Features and Identify Potentially Functional Variants

In addition to visualization, the UCSC browser is also a powerful tool for large scale analysis of the genomic context of a given list of genome features, such as SNPs. A causal SNP is unlikely to be tested directly in a genome scan, but in principle it may be in LD with markers that have been genotyped (This is the principle underlying association analysis). After creation of a custom track containing the SNPs of interest (see above), the SNPs can be queried using the UCSC *Table browser* (27). Table Browser, which is accessed from the “Tables” link in the main browser, is an excellent tool that effectively allows the user to perform complex queries between data sets, including custom tracks loaded by the user. For example, it is possible to identify all SNPs (your custom track) that overlap with conserved transcription factor binding sites (TFBS). To do this, take the following steps:

1. Entering the Table Browser and select the “Custom Tracks” from the pull down “group” menu

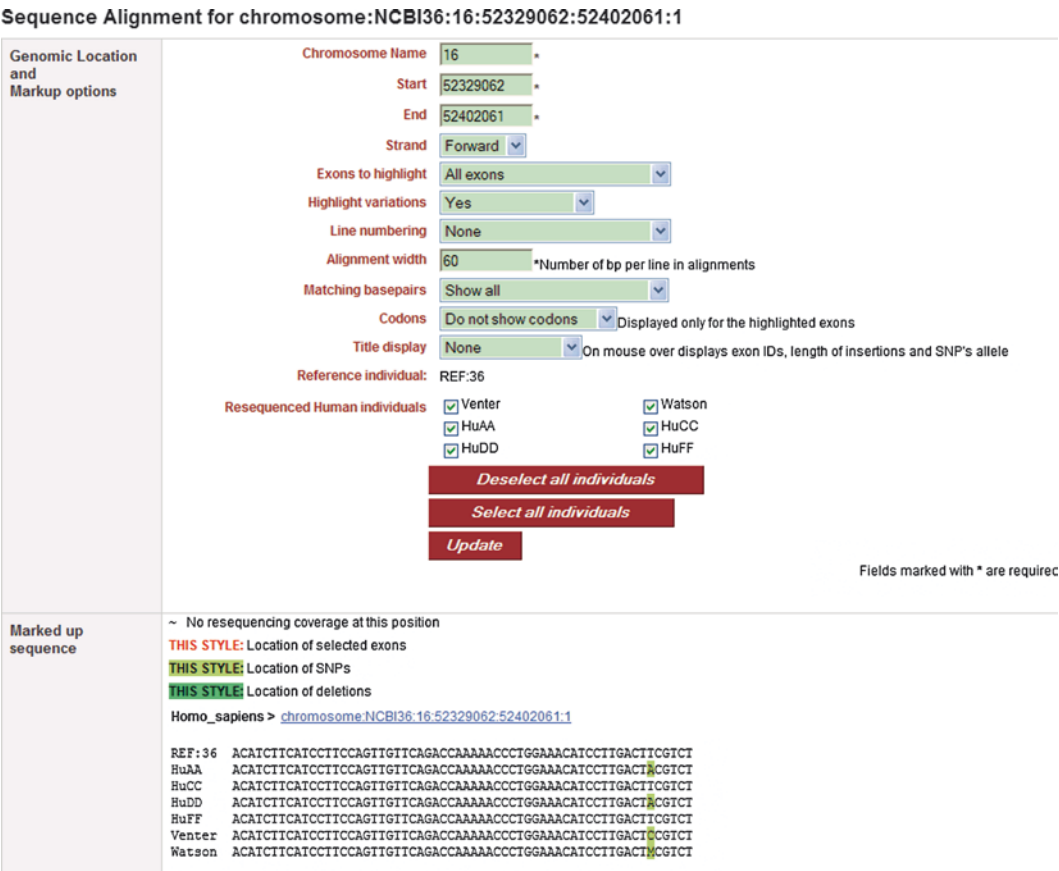


Fig. 5. Individual genome sequence data. An Ensembl *Resequencing Alignment View* of six individual genome sequences, including sequences from James Watson, Craig Venter and four other anonymous individuals across a user defined genomic region in the FTO gene. (http://www.ensembl.org/Homo_sapiens/sequencealignview?). SNPs are highlighted in grey, in this case, a tri-allelic SNP position is shown that is not represented in dbSNP

- 2. Select your custom track of choice from the “track” menu
- 3. Press the [create] intersection button
- 4. Select the group and track you are interested in, e.g. The “Regulation” group and the “TFBS Conserved” track. Press [submit]
- 5. To view a summary of overlaps, press the [summary/statistics] button
- 6. To view SNPs overlapping TFBS sites, press the [get output] button.

This basic process can be used for very large complex queries, making the UCSC table browser one of the most useful tools available for biologists and geneticists. It is possible to take this type of analysis to an even higher level of sophistication using UCSC data focused workflow tools such as Galaxy (3).

3.10. Conclusion

In the case study used in this review, the association seen between Type II diabetes and the FTO locus has been evaluated at a molecular level. Analysis of the associated locus in the full context of the data annotated by tools like the UCSC genome browser, supported several hypotheses which might explain the association. LD appeared to restrict the association to Intron 1 of the FTO gene, suggesting a possible regulatory element. Review of the data across the region identified an associated variant in an exapted repeat sequence, which is known to show regulatory function. This might warrant further investigation. An ultra-conserved element, directing specific hindbrain expression was also identified neighboring the associated markers, this may also be worth further investigation. As this example illustrates, mastering the in silico data to build a biological rationale around an association is not a trivial process, but it is achievable using publicly available web resources. Ultimately, good in silico analysis may help to align an association to a molecular mechanism, but as a general rule, it will raise more questions than it answers, returning the focus to the experimentalist.

4. Notes

1. The UCSC genome graphs help documentation: (<http://genome.ucsc.edu/goldenPath/help/hgGenomeHelp.html>).
2. *Retrieving a set of SNPs in Linkage Disequilibrium (LD)*: The SNP Annotation and Proxy Search tool, *SNAP* (Table 1) is a useful tool for identifying SNPs in LD with an SNP of interest. The output of the tool can be rapidly converted into a custom track using a text editor. The r^2 LD threshold is set by default to 0.8, this can be modified to increase or reduce stringency of LD.
3. The UCSC Custom track help documentation: <http://genome.ucsc.edu/goldenPath/help/customTrack.html>
4. The FTO region case study directly addresses one of the most challenging problems for complex disease genetics. Although an SNP association may be localized to a particular gene, association mapping also needs to take LD into account. An SNP showing association may be in strong LD with an ungenotyped marker nearby or in some cases at a considerable distance from the associated marker. This means that genetic associations need to evaluate the LD across a region, and each marker in LD with the associated SNP needs to be evaluated as a candidate for the molecular basis of the association. Genome browsers are supremely effective tools to assist this search.

5. *Selection and configuration of track information in the UCSC genome browser*: Over 100 tracks of information are available to view in the UCSC human genome browser. These tracks contain highly specific information across many fields. However, for general applications, 20–30 tracks are likely to see the most regular use. More importantly, selection of more than 10–15 tracks is likely to slow the browser down considerably, so it is worth turning off tracks which are not being used. In order to determine the best track for the job, it is worth reading the track documentation to check the provenance and age of the data.
6. *A Caveat to consider when dealing with LD “blocks”*: Although the traditional triangular block structure of an LD plot (Fig. 1) is a useful and intuitive guide to the extent of LD across a region, it is important to be aware that LD may extend across greater distances than the block structure suggests. This may be due to many factors, including the presence of longer rare haplotypes in the population or differences between LD structure in the study population and the HapMap population. Consequently, LD blocks should be taken as guides only and further analysis of the extent of LD should always be carried out.
7. *Building Biological Rationale around genes*: It is important to preface the consideration of biological rationale for genes in phenotypes or diseases, with an acknowledgement that a convincing rationale can be made for almost any gene in almost any phenotype if enough sources of information are mined. However, there are some simple principles and tools (listed in Table 1) that may help to identify genes with good links to a specific phenotype. Firstly, is the gene expressed in the relevant tissue? This can be reviewed with the *SymAtlas* tool. Secondly, is the gene linked to the phenotype in the literature? *Huge Navigator* is a good tool enabling rapid review of the literature around a gene. Finally, does the gene fall into a pathway or interact with other genes with a known involvement in the phenotype? In this case, the EMBL *STITCH* tool is a good place to start. Once these areas have been considered and wishful thinking has been purged, then further investigation can be planned.
8. *Vista Enhancers*: Three ultra-conserved enhancer regions which have been demonstrated to drive tissue specific within the FTO gene, intron 7 (http://enhancer.lbl.gov/cgi-bin/imagedb.pl?form=presentation&show=1&experiment_id=element_155).

References

1. Kuhn, R.M., Karolchik, D., Zweig, A.S., Wang, T., Smith, K.E., Rosenbloom, K.R., *et al.* (2009) The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res.*, **37**, D755–D761.
2. Hubbard, T.J., Aken, B.L., Ayling, S., Ballester, B., Beal, K., Bragin, E., *et al.* (2009) Ensembl 2009. *Nucleic Acids Res.*, **37**, D690–D697.
3. Woollard, P. (2010) Asking complex questions of the genome without programming. *Methods Mol. Biol.*, 39–52.
4. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.
5. Frayling, T.M., Timpson, N.J., Weedon, M.N., Zeggini, E., Freathy, R.M., Lindgren, C.M., *et al.* (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, **316**, 889–894.
6. de Bakker, P.I., Yelensky, R., Pe'er, I., Gabriel, S.B., Daly, M.J. and Altshuler, D. (2005) Efficiency and power in genetic association studies. *Nat. Genet.*, **37**, 1217–1223.
7. Anderson, C.A., Pettersson, F.H., Barrett, J.C., Zhuang, J.J., Ragoussis, J., Cardon, L.R. and Morris, A.P. (2008) Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. *Am. J. Hum. Genet.*, **83**, 112–119.
8. Dina, C. (2008) New insights into the genetics of body weight. *Curr. Opin. Clin. Nutr. Metab. Care*, **11**, 378–384.
9. Sandelin, A. (2008) Prediction of regulatory elements. *Methods Mol. Biol.*, **453**, 233–244.
10. Callinan, P.A. and Feinberg, A.P. (2006) The emerging science of epigenomics. *Hum. Mol. Genet.*, **15 Spec No 1**, R95–R101.
11. Costello, J.F. and Vertino, P.M. (2002) Methylation matters: a new spin on maspin. *Nat. Genet.*, **31**, 123–124.
12. Futscher, B.W., Oshiro, M.M., Wozniak, R.J., Holtan, N., Hanigan, C.L., Duan, H. and Domann, F.E. (2002) Role for DNA methylation in the control of cell type specific maspin expression. *Nat. Genet.*, **31**, 175–179.
13. Rakyan, V.K., Hildmann, T., Novik, K.L., Lewin, J., Tost, J., Cox, A.V., *et al.* (2004) DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. *PLoS Biol.*, **2**, e405.
14. Ozsolak, F., Song, J.S., Liu, X.S. and Fisher, D.E. (2007) High-throughput mapping of the chromatin structure of human promoters. *Nat. Biotechnol.*, **25**, 244–248.
15. Rada-Iglesias, A., Ameur, A., Kapranov, P., Enroth, S., Komorowski, J., Gingeras, T.R. and Wadelius, C. (2008) Whole-genome maps of USF1 and USF2 binding and histone H3 acetylation reveal new aspects of promoter structure and candidate genes for common human disorders. *Genome Res.*, **18**, 380–392.
16. Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., Nobrega, M.A., Shoukry, M., *et al.* (2006) In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, **444**, 499–502.
17. Zhang, J., Feuk, L., Duggan, G.E., Khaja, R. and Scherer, S.W. (2006) Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. *Cytogenet. Genome Res.*, **115**, 205–214.
18. Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.
19. Cooper, G.M., Zerr, T., Kidd, J.M., Eichler, E.E. and Nickerson, D.A. (2008) Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat. Genet.*, **40**, 1199–1203.
20. McCarroll, S.A. (2008) Extending genome-wide association studies to copy-number variation. *Hum. Mol. Genet.*, **17**, R135–R142.
21. Lowe, C.B., Bejerano, G. and Haussler, D. (2007) Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 8005–8010.
22. Bejerano, G., Lowe, C.B., Ahituv, N., King, B., Siepel, A., Salama, S.R., Rubin, E.M., Kent, W.J. and Haussler, D. (2006) A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature*, **441**, 87–90.
23. Mardis, E.R. (2008) Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, **9**, 387–402.
24. Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., *et al.* (2008) The diploid genome sequence of an Asian individual. *Nature*, **456**, 60–65.

25. Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872–876.
26. Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol.*, **5**, e254.
27. Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D. and Kent, W.J. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.



<http://www.springer.com/978-1-60327-366-4>

Genetic Variation

Methods and Protocols

Barnes, M.R.; Breen, G. (Eds.)

2010, XI, 388 p., Hardcover

ISBN: 978-1-60327-366-4

A product of Humana Press