DATABASES

PhenCode: Connecting ENCODE Data With Mutations and Phenotype

Belinda Giardine,[1] Cathy Riemer,[1] Tim Hefferon,[2] Daryl Thomas,[3] Fan Hsu,[3] Julian Zielenski,[4] Yunhua Sang,[4] Laura Elnitski,[2] Garry Cutting,[5] Heather Trumbower,[3] Andrew Kern,[3] Robert Kuhn,[3] George P. Patrinos,[6] Jim Hughes,[7] Doug Higgs,[7] David Chui,[8] Charles Scriver,[9] Manyphong Phommarinh,[9] Santosh K. Patnaik,[10] Olga Blumenfeld,[10] Bruce Gottlieb,[11] Mauno Vihinen,[12] Jouni Väliaho,[12] Jim Kent,[3] Webb Miller,[1] and Ross C. Hardison[1,13]*

[1]Center for Comparative Genomics and Bioinformatics, Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, Pennsylvania; [2]National Human Genome Research Institute, Bethesda, Maryland; [3]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, Santa Cruz, California; [4]Program in Genetics and Genomic Biology, The Hospital for Sick Children, Toronto, Ontario, Canada; [5]Department of Pediatrics, Johns Hopkins University School of Medicine, Johns Hopkins University, Baltimore, Maryland; [6]Erasmus Medical Center, Faculty of Medicine, Department of Cell Biology and Genetics, Rotterdam, The Netherlands; [7]Weatherall Institute of Molecular Medicine, Oxford, United Kingdom; [8]Department of Medicine, Boston University, Boston, Massachusetts; [9]Montreal Children's Hospital Research Institute, Montreal, Quebec, Canada; [10]Department of Chemistry, Albert Einstein College of Medicine, Bronx, New York; [11]Lady Davis Institute for Medical Research, Sir Mortimer B. Davis-Jewish General Hospital, Montreal, Quebec, Canada; [12]Institute of Medical Technology, University of Tampere, Tampere, Finland; [13]Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, Pennsylvania

Communicated by Richard Cotton

PhenCode (Phenotypes for ENCODE; www.bx.psu.edu/phencode) is a collaborative, exploratory project to help understand phenotypes of human mutations in the context of sequence and functional data from genome projects. Currently, it connects human phenotype and clinical data in various locus-specific databases (LSDBs) with data on genome sequences, evolutionary history, and function from the ENCODE project and other resources in the UCSC Genome Browser. Initially, we focused on a few selected LSDBs covering genes encoding alpha- and beta-globins (*HBA, HBB*), phenylalanine hydroxylase (*PAH*), blood group antigens (various genes), androgen receptor (*AR*), cystic fibrosis transmembrane conductance regulator (*CFTR*), and Bruton's tyrosine kinase (*BTK*), but we plan to include additional loci of clinical importance, ultimately genomewide. We have also imported variant data and associated OMIM links from Swiss-Prot. Users can find interesting mutations in the UCSC Genome Browser (in a new Locus Variants track) and follow links back to the LSDBs for more detailed information. Alternatively, they can start with queries on mutations or phenotypes at an LSDB and then display the results at the Genome Browser to view complementary information such as functional data (e.g., chromatin modifications and protein binding from the ENCODE consortium), evolutionary constraint, regulatory potential, and/or any other tracks they choose. We present several examples illustrating the power of these connections for exploring phenotypes associated with functional elements, and for identifying genomic data that could help to explain clinical phenotypes. Hum Mutat 0, 1–9, 2007. Published 2007 Wiley-Liss, Inc.[†]

KEY WORDS: ENCODE; mutations; phenotype; UCSC Genome Browser

INTRODUCTION

Genome browsers, such as those at the University of California, Santa Cruz (UCSC) [Kent et al., 2002] and Ensembl [Hubbard et al., 2005], provide convenient, centralized access to a wealth of genotype data, including not only sequences but also computational and experimental results in areas ranging from evolutionary history to functional studies. In particular, the ENCODE consortium [ENCODE Project Consortium, 2004], supported by the National Human Genome Research Institute, aims to identify all of the functional elements in the human genome. Results from the first phase of that project (ENCODE Project Consortium, unpublished results), covering 30 Mb of human DNA, are already in the browsers.

In contrast, detailed data on naturally-occurring human mutations and the phenotypes they cause, while obviously critical

Received 14 June 2006; accepted revised manuscript 18 December 2006.

*Correspondence to: Ross C. Hardison, Pennsylvania State University, 304 Wartik Laboratory, University Park, PA 16802.
E-mail: rch8@psu.edu

Grant sponsor: Intramural Program of the NIH; Grant sponsor: NHGRI; Grant sponsor: Réseau de Médicine Génétique Appliquée/ Fonds de la recherche en santé du Québec; Grant sponsor: Canadian Cystic Fibrosis Foundation and Genome Canada; Grant sponsor: NHGRI; Grant numbers: HG002238 and 1P41HG02371; Grant sponsor: NIH; Grant numbers: DK65806.

DOI 10.1002/humu.20484

Published online in Wiley InterScience (www.interscience.wiley.com).

from a health perspective, tend to be scattered among literature articles and/or locus-specific databases (LSDBs) dedicated to a gene or disease [Cotton et al., 1998]. These databases, while in many cases offering excellent in-depth coverage of clinical issues relating to observed mutations in their particular loci, generally do not provide an easy way to compare these data with the kinds of genotypic and functional data available at the browsers, or with similar phenotypes associated with mutations at other loci.

The PhenCode project (Phenotypes for ENCODE; www.bx.psu.edu/phencode) aims to remedy this situation by connecting these complementary data sources so users can easily navigate between them and compare their contents. This work describes the project to connect LSDBs to the UCSC Genome Browser, efforts to collect information on protein-altering variants genome-wide, and examples of the benefits of integrating the information on mutations with the functional data from ENCODE to achieve new insights into disease processes.

WEBSITES

Key websites and databases mentioned in this article are listed in Table 1.

METHODS

PhenCode is not a new database of mutations. Rather it consists of new connections between the LSDBs and the UCSC Genome Browser. The LSDB data is analyzed by a series of scripts, and converted into several tables in the UCSC MySQL databases. The main table is in a variant of the browser extensible data (BED) format. This allows quick drawing of the regions in the main display, and filtering of what is drawn. A set of entity relationship (ER) tables stores most of the attributes that are displayed on the details page. Special cases are made for outside links, to allow more functionality. Using the ER-style tables allows flexibility in what is stored for each mutation and is easily adapted to changes in this data. The current schema can be viewed by clicking the "View table schema" link on the details page. It can also be viewed on the UCSC Table Browser, and a detailed example of doing this is provided on the frequently asked questions (FAQ) page for PhenCode. All of these tables are reloaded regularly from the LSDB sources to incorporate updates from the LSDB curators. We have also added a small amount to the UCSC Browser code to accommodate the unique features of this data.

When adding a new LSDB, the data must be mapped onto chromosome coordinates and put into a format that can be loaded into the UCSC tables. The LSDBs are encouraged to allow links from UCSC back to the original data, and to add UCSC custom tracks as an output option to their query interfaces. The links back to the individual LSDB help users to find it easily, and the LSDB can then provide more complex querying capacity and/or more detailed data. If the LSDB provides the capacity to send query results to the Genome Browser as custom tracks, then users can combine the querying ability of the LSDB with the Browser's wide range of data.

The task of loading the LSDB data into the UCSC tables can be challenging because of variations among the LSDBs in the fields, coordinate systems, and nomenclature used, but that variability is handled by altering the conversion scripts as needed. Each LSDB records fields that are interesting for that locus. Some fields, such as nucleotide changes, amino acid changes, and phenotype, are common to nearly all of the databases, while other fields are specific to a particular database, such as "gender raised as" (ARdb), and "stability" (HbVar). For maximum utility, the common fields must be identified and pooled together in the UCSC tables (even if the LSDBs use different names for them), while preserving the flexibility to also have locus-specific fields.

Because the Genome Browser presents all information in the coordinates of a genome assembly, it is critical to convert the mutations' positions from the individual LSDB's numbering system into these genomic coordinates. Most of the LSDBs use a coordinate system that is based on either a gene or a reference sequence from GenBank. However, the numbering of the bases varies. Position "1" may be the A of the ATG, the base following the G, the transcription start site, or the first base in the GenBank file. Also, for those using coding sequence numbering, intron positions can be referred to by their+/−distance from the nearest base in an exon (the Human Genome Variation Society [HGVS] standard way), or as a direct count into the intron. Alternative splicing can also complicate this, since the splice variant must be identified. Since the reference sequences used by the LSDBs are not always exactly the same as the chromosome sequence, we use BLAT [Kent, 2002] to map the coordinates. This handles small

TABLE 1. Web Sites

Database	Website address
ARdb	http://www.androgendb.mcgill.ca
BGMUT	http://www.ncbi.nlm.nih.gov/projects/mhc/xslcgi.fcgi?cmd = bgmut/home
CFMDB	http://www.genet.sickkids.on.ca/cftr
ENCODE	http://www.genome.gov/10005107/
Ensembl	http://www.ensembl.org
GenPhen	http://globin.bx.psu.edu/genphen
HbVar	http://globin.bx.psu.edu/hbvar
HGVbase	http://hgvbase.cgb.ki.se
HGVS	http://www.hgvs.org
HmutDB	http://www.ebi.ac.uk/mutations/central
IDbases	http://bioinf.uta.fi/base_root
Mammalian Phenotype Ontology	http://www.informatics.jax.org/searches/MP_form.shtml
OMIM	http://www.ncbi.nlm.nih.gov/omim
PAHdb	http://www.pahdb.mcgill.ca/
PhenCode	http://www.bx.psu.edu/phencode
SRS	http://srs.ebi.ac.uk
TRANSFAC	http://www.biobase.de
UCSC Genome Browser	http://genome.ucsc.edu
WayStation	http://www.centralmutations.org

insertions/deletions that would cause errors if we just used a simple offset from the start point. Running BLAT on the sequence as a whole is much faster and simpler than trying to run it on each mutation separately. Most LSDBs do not provide the surrounding sequence for each mutation.

Naming systems for the mutations also vary. The HGVS has developed nomenclature standards (www.hgvs.org/mutnomen) [den Dunnen and Antonarakis, 2001, 2002] for the simple mutations like substitutions and indels, but the more complex ones, such as fusion genes, are still being worked on, and these recommendations could change as more real cases are described. Meanwhile, each research/clinical community has its own traditional naming scheme: some of these are based on the nucleotide or protein change, some on the hospital where the first case was found, others on a shorthand notation that often requires knowledge of the locus to really understand. Even those that attempt to use the HGVS standards vary slightly depending on when they began, as the standards have changed over the years. We endeavor to translate the traditional name(s) for each mutation into the HGVS standard, but still maintain the others as aliases, so that users can choose the type of name they want displayed.

Last, there is a need for standardization in describing phenotypes. Mouse and rat databases have made effective use of the Mammalian Phenotype (MP) Ontology (www.informatics.jax.org/searches/MP_form.shtml), but this has not been used extensively for human mutations. We record the common terms but encourage curators to also provide standardized terms. Without such an ontology, situations where the same phenotype is produced by mutations in different loci become less apparent, since different terminology may be used.

The critical issues are the reference sequence, numbering system, and mutation description format. Once we have resolved these issues for a particular LSDB by careful examination of the data and consultation with its curators, we write an automated extraction script to convert the downloaded data files into a format that can be loaded into the UCSC track tables. Sample code for making these conversions is available from the PhenCode FAQ page, although the main script is usually modified for each new database. In the best case scenario this process may take only an hour. This main script calls other utility routines to do tasks that are common among all the conversions. For example, once the HGVS-style mutation name is extracted from the raw data (either directly or computed, depending on what is available at the LSDB), a utility routine then uses that to obtain chromosome coordinates for the mutation. In addition to writing the script, we also run the LSDB's reference sequence through BLAT (using the correct splice variant), and store the results for later access by the script. If some introns do not match, we augment the script with special code to handle them. The script and BLAT results embody all of the database-specific nuances necessary to import the data from a particular LSDB; once they are set up, the script can be rerun easily at any time to update the UCSC track with the latest version of the data from the LSDB curators. Currently these updates are initiated manually as needed, but in the future we plan to run them via an automated scheduler.

RESULTS

Connecting Information on Disease Variants With the UCSC Genome Browser

The PhenCode project connects LSDBs as data sources so users can easily navigate between them and compare their contents. Working closely with the LSDB staff, we collect and assemble information about the mutations into a new track at the UCSC Genome Browser called "Locus Variants," listed in the "Phenotype and Disease Associations" section. Where possible, each mutation recorded in this track provides a link back to the corresponding entry in the LSDB, which remains responsible for maintaining and curating the underlying data. In this way, users can see all of the mutations together in one place, aligned with conservation and experimental data, yet still return to the LSDB for details on phenotype. And conversely, the LSDB can offer its users the option to view their query results in the context of other tracks at the Genome Browser. Furthermore, summary attributes stored with each mutation allow users to perform basic phenotype queries across loci (e.g., all mutations that cause anemia, regardless of gene), via the UCSC Table Browser interface [Karolchik et al., 2004]. Last, we ask LSDB curators to add to Open Regulatory Annotation (ORegAnno) [Montgomery et al., 2006] a list of regulatory regions within their loci, which is assembled into a companion track called "ORegAnno" (listed in the "Expression and Regulation" section) to help guide users when viewing the mutations.

In addition to incorporating data from LSDBs, we have also imported variant data and associated OMIM links from Swiss-Prot [Bairoch and Apweiler, 2000]. This does not have as much detail as many of the LSDBs, but has the advantage of genomewide coverage. It only includes mutations that produce protein variants, and thus omits other types such as regulatory mutations. Just as with the entries from LSDBs, users can follow links back to Swiss-Prot for more information.

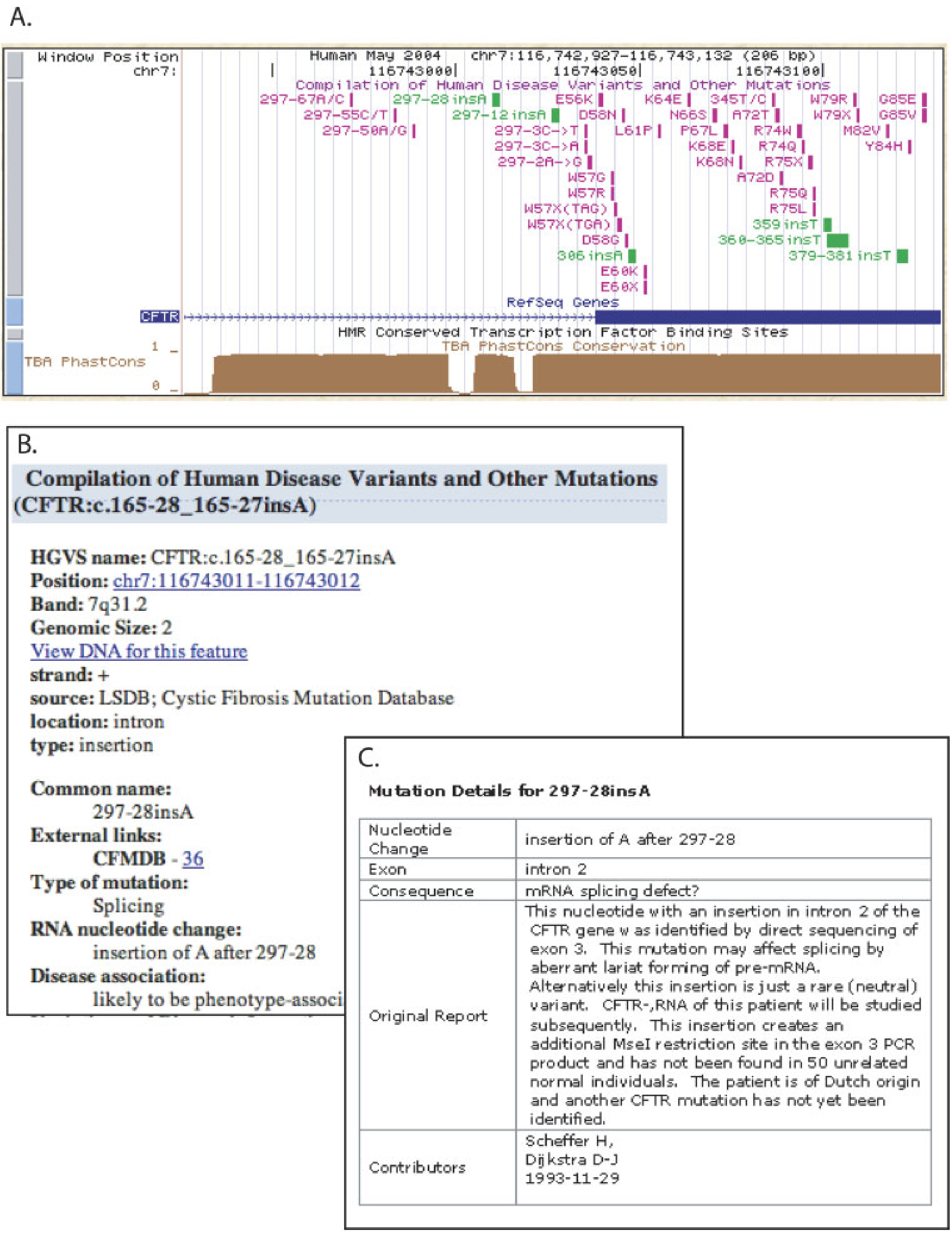
The databases currently contributing to the Locus Variants track are listed in Table 2, along with the variant counts and status of links to the source. This information is routinely updated in a table located in the FAQs at the PhenCode website (www.bx.psu.edu/phencode).

Example 1: Finding Specific Information on Mutations Starting at the Genome Browser

Opening the Locus Variants track at the *CFTR* gene shows many variants (Fig. 1A). These were obtained from CFMDB (www.genet.sickkids.on.ca/cftr), which is a frequently updated, comprehensive database of *CFTR* variants reported in the literature and exchanged between researchers in the Cystic Fibrosis Genetic Analysis Consortium. The CFMDB currently contains 1,448 sequence variants, most of which were detected in cystic fibrosis (CF) patients and are presumed to be disease-causing. Entries include the specific nucleotide change, its predicted consequence to the transcript and/or protein, a brief

TABLE 2. **Locus-Specific Databases and Other Data Sources for the Locus Variants Track**

Database	Number of variants	Number of track items	Links to source
ARdb	329	329	No
BGMUT	630	1605	Yes
BTKbase	508	512	Yes
CFMDB	1,400	1,400	Yes
HbVar	1,220	1,531	Yes
PAHdb	508	513	Yes
SRD5A2	42	42	No
Swiss-Prot	22,577	22,454 (hg18)	Yes
Totals	27,212	28,382	



Compilation of Human Disease Variants and Other Mutations (CFTR:c.165-28_165-27insA)

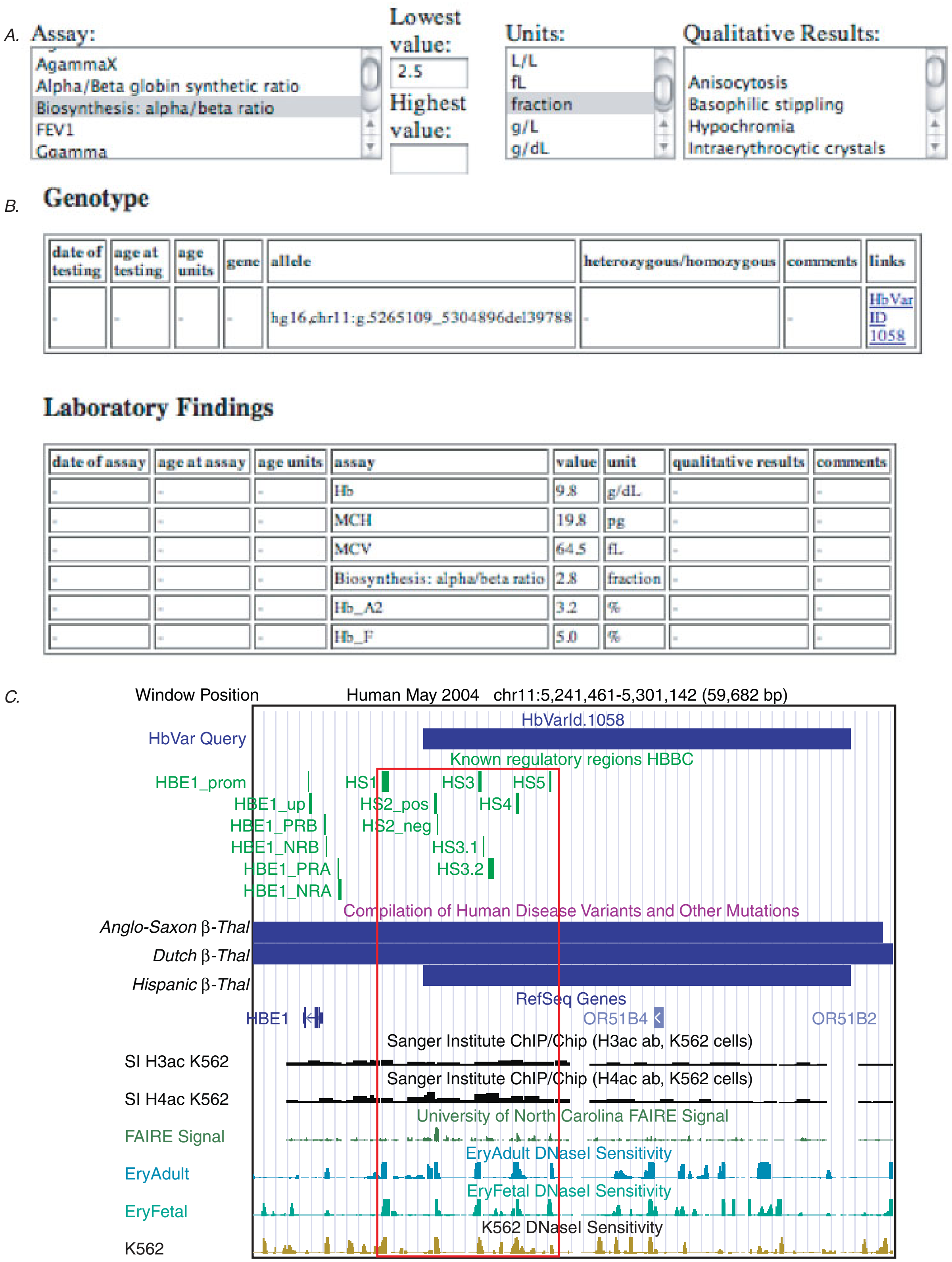
HGVS name: CFTR:c.165-28_165-27insA
Position: chr7:116743011-116743012
Band: 7q31.2
Genomic Size: 2
View DNA for this feature
strand: +
source: LSDB; Cystic Fibrosis Mutation Database
location: intron
type: insertion

Common name:
297-28insA
External links:
CFMDB - 36
Type of mutation:
Splicing
RNA nucleotide change:
insertion of A after 297-28
Disease association:
likely to be phenotype-associ

Mutation Details for 297-28insA

Nucleotide Change	insertion of A after 297-28
Exon	intron 2
Consequence	mRNA splicing defect?
Original Report	This nucleotide with an insertion in intron 2 of the CFTR gene w as identified by direct sequencing of exon 3. This mutation may affect splicing by aberrant lariat forming of pre-mRNA. Alternatively this insertion is just a rare (neutral) variant. CFTR-,RNA of this patient will be studied subsequently. This insertion creates an additional MseI restriction site in the exon 3 PCR product and has not been found in 50 unrelated normal individuals. The patient is of Dutch origin and another CFTR mutation has not yet been identified.
Contributors	Scheffer H, Dijkstra D-J 1993-11-29

FIGURE 1. **Interesting mutations identified on the LocusVariants track can lead to insights into phenotype. A: UCSC Genome Browser display centered on exon 3 of the *CFTR* gene, showing a cluster of noncoding mutations in a highly conserved region of intron 2. Purple = substitution; Green = insertion. B: Details page for mutation c.165-28insA, including a link to the disease-specific mutation database CFMDB. The nomenclature used on this page follows the recommendation of HGVS that base "1" corresponds to the "A" of the first codon ATG. The common names (shown on panels A and C) reflect the traditional mutation nomenclature for *CFTR*, which begins with base "1" as the start of the 5′UTR. C: Mutation details at the Cystic Fibrosis Mutation Database. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com]**



A.

Assay:

AgammaX
Alpha/Beta globin synthetic ratio
Biosynthesis: alpha/beta ratio
FEV1
Ggamma

Lowest value: 2.5

Highest value:

Units:

L/L
fL
fraction
g/L
g/dL

Qualitative Results:

Anisocytosis
Basophilic stippling
Hypochromia
Intraerythrocytic crystals

B.

Genotype

date of testing	age at testing	age units	gene	allele	heterozygous/homozygous	comments	links
-	-	-	-	hg16.chr11:g.5265109_5304896del39788	-	-	HbVar ID 1058

Laboratory Findings

date of assay	age at assay	age units	assay	value	unit	qualitative results	comments
-	-	-	Hb	9.8	g/dL	-	-
-	-	-	MCH	19.8	pg	-	-
-	-	-	MCV	64.5	fL	-	-
-	-	-	Biosynthesis: alpha/beta ratio	2.8	fraction	-	-
-	-	-	Hb_A2	3.2	%	-	-
-	-	-	Hb_F	5.0	%	-	-

C.
Window Position
Human May 2004 chr11:5,241,461-5,301,142 (59,682 bp)
HbVar Query
HbVarId.1058
Known regulatory regions HBBC
HBE1_prom
HS1
HS3
HS5
HBE1_up
HS2_pos
HS4
HBE1_PRB
HS2_neg
HBE1_NRB
HS3.1
HBE1_PRA
HS3.2
HBE1_NRA
Compilation of Human Disease Variants and Other Mutations
Anglo-Saxon β-Thal
Dutch β-Thal
Hispanic β-Thal
RefSeq Genes
HBE1
OR51B4
OR51B2
Sanger Institute ChIP/Chip (H3ac ab, K562 cells)
SI H3ac K562
Sanger Institute ChIP/Chip (H4ac ab, K562 cells)
SI H4ac K562
University of North Carolina FAIRE Signal
FAIRE Signal
EryAdult DNaseI Sensitivity
EryAdult
EryFetal DNaseI Sensitivity
EryFetal
K562 DNaseI Sensitivity
K562

FIGURE 2. **Phenotypes discovered at an LSDB can be related to mechanisms using other tracks at the UCSC Genome Browser. A: GenPhen query form. B: GenPhen details page. C: Genome Browser display showing the thalassemia deletion in comparison with ENCODE results. The locus control region is boxed in red. Long blue rectangles show the extent of the deletions, and tracks below them show RefSeq gene annotations followed by tracks from the ENCODE Project Consortium (unpublished results) displaying features associated with gene regulatory regions. The custom track of known regulatory regions in the *HBB* complex is from www.bx.psu.edu/~ross/dataset/ReglRegHBBhg17CusTrk.txt [King et al., 2005]. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com]**

report of the original discovery of the mutation including brief clinical and molecular data, names of the original contributors and their institutional affiliations, any recent updates, and a reference to the original paper or personal communication describing the mutation, including an automatic PubMed search for the mutation by name.

The Genome Browser enables the user to browse for interesting mutations. For example, examining the Locus Variants track immediately upstream of *CFTR* exon 3 reveals a cluster of noncoding mutations (Fig. 1A). Comparison with the Conservation track shows that these intronic mutations are in a highly constrained region, as measured by the PhastCons score [Siepel et al., 2005]. Thus these mutations could be altering a function that is under intense purifying selection in mammals, perhaps a splicing enhancer, which could lead to the pathological phenotype. To investigate this further, the user clicks on one of the mutation icons in the Genome Browser window to go to a details page containing some information about the mutation (Fig. 1B). This page includes a link directly to the CFMDB entry for a more complete description and corresponding references (Fig. 1C). The description at CFMDB states that this particular variant could affect splicing or it may be a neutral variant.

Example 2: Using GenPhen To Find Candidate Mutations for a Thalassemia Patient, Then Viewing Them in Register With ENCODE Functional Data at the Genome Browser

A user can start with information on mutations or patients in LSDBs and go to the Genome Browser for information that may help in understanding the phenotype. Consider a patient with anemia whose blood tests reveal an excess synthesis of alpha-globin compared to beta-globin, which is characteristic of beta-thalassemia. Suppose you are interested in the regulatory mutations that cause this, especially deletions, and in particular, you would like to see how they correspond with the experimental data on function coming from the ENCODE project. This can be accomplished by combining the query capabilities of GenPhen (http://globin.bx.psu.edu/genphen) with the Genome Browser's ability to display ENCODE data. GenPhen is a prototype database of human hemoglobinopathy genotypes and phenotypes that records anonymized information from individual patients, including laboratory findings and clinical presentation. We begin by going to the GenPhen query form (Fig. 2A) to search for patients with a high alpha/beta ratio (greater than 2.5). This query finds six patients, one of whom has the mutation known as the "Hispanic deletion." The details page for this patient (Fig. 2B) shows that he/she is anemic and has the diagnostic biosynthetic chain imbalance. Following the link to the corresponding mutation entry in the hemoglobin variant database (HbVar) [Hardison et al., 2002; Patrinos et al., 2004] leads to a link to the UCSC Genome Browser, with the current mutation displayed as a user track (Fig. 2C). Examination alongside several tracks of functional data (ENCODE Project Consortium, unpublished results) shows that the deleted interval overlaps DNA segments with hallmarks of gene regulatory regions, including DNase hypersensitive sites (DNaseI Sensitivity), nucleosome-depleted regions (FAIRE Signal), and histone modifications (Sanger Institute ChIP/Chip). These correspond to known *cis*-regulatory regions in the locus control region (LCR) [reviewed in Li et al., 2002] indicated in another custom track. These results suggest that the loss of critical LCR regions accounts for the low beta-globin production in the patient carrying the Hispanic deletion. Indeed, mapping [Driscoll et al., 1989] and analyzing [Forrester et al., 1990] this deletion provided major contributions to our understanding of the LCR [Grosveld et al., 1987; Tuan et al., 1989].

Example 3: Leveraging Swiss-Prot Annotations

Currently almost all LSDBs focus on rare mutations in single genes that cause severe disease phenotypes. However, many diseases with a genetic component involve multiple genes and less severe phenotypes, and in some cases the "disease" allele of any particular gene is relatively common. If there is evidence associating a particular region of the genome with the disease, it can be fruitful to look for SNPs in coding regions for possible candidates. In this case generally an LSDB is not available, but there may be useful mutation annotations derived from Swiss-Prot.

The *APOA* cluster lies within several quantitative trait loci (QTLs) for blood pressure, body weight, and rheumatoid arthritis

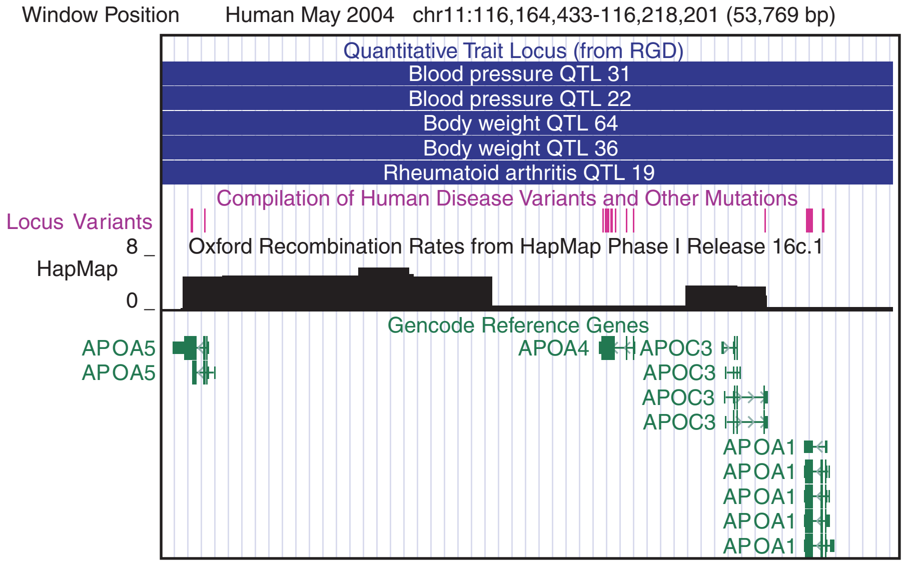

Window Position
Human May 2004 chr11:116,164,433-116,218,201 (53,769 bp)
Quantitative Trait Locus (from RGD)
Blood pressure QTL 31
Blood pressure QTL 22
Body weight QTL 64
Body weight QTL 36
Rheumatoid arthritis QTL 19
Compilation of Human Disease Variants and Other Mutations
Locus Variants
8 _
HapMap
0 _
Oxford Recombination Rates from HapMap Phase I Release 16c.1
Gencode Reference Genes
APOA5
APOA5
APOA4
APOC3
APOC3
APOC3
APOC3
APOA1
APOA1
APOA1
APOA1
APOA1

FIGURE 3. **Mutations, quantitative trait loci, and recombination frequency in the *APOA* cluster. The default display in the UCSC Genome Browser excludes variants from Swiss-Prot, so in order to see these, users will need to uncheck the "Exclude" box for Swiss-Prot on the track settings page for Locus Variants. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com]**

identified in the rat (Fig. 3). The Locus Variants track shows that there are many nonsynonymous mutations in the cluster. Following the links back to Swiss-Prot shows that the *APOA5* gene has extensive literature regarding lipid regulation that could be related to cardiovascular disease [Pennacchio et al., 2002; Vrablik et al., 2003; Kao et al., 2003]. However, association studies for cardiovascular disease have had low power to ascribe function to individual genes due to the extensive linkage disequilibrium across the four genes in this locus [reviewed in Lai et al., 2005]. The Genome Browser view shows a region of elevated recombination rate between the *APOA5* gene and the other *APO* genes in the locus (corroborated by Olivier et al.

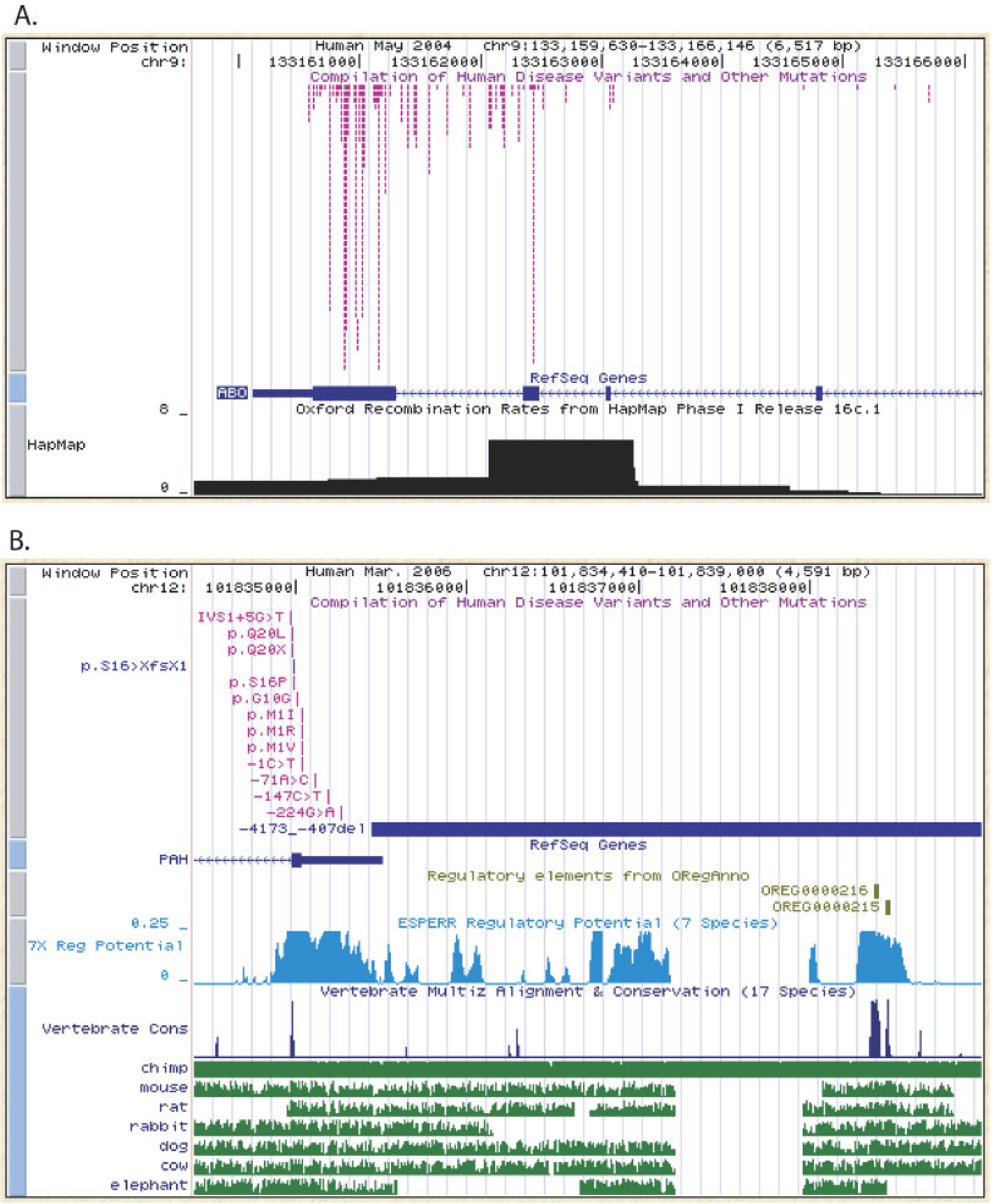

A.
Window Position
chr9:
Human May 2004 chr9:133,159,630-133,166,146 (6,517 bp)
133161000| 133162000| 133163000| 133164000| 133165000| 133166000|
Compilation of Human Disease Variants and Other Mutations
RefSeq Genes
ABO
8 _
Oxford Recombination Rates from HapMap Phase I Release 16c.1
HapMap
0 _
B.
Window Position
chr12:
Human Mar. 2006 chr12:101,834,410-101,839,000 (4,591 bp)
101835000| 101836000| 101837000| 101838000|
Compilation of Human Disease Variants and Other Mutations
IVS1+5G>T
p.Q20L
p.Q20X
p.S16>XfsX1
p.S16P
p.G10G
p.M1I
p.M1R
p.M1V
-1C>T
-71A>C
-147C>T
-224G>A
-4173_-407del
RefSeq Genes
PAH
Regulatory elements from ORegAnno
OREG0000216
OREG0000215
0.25 _
7X Reg Potential
0 _
ESPERR Regulatory Potential (7 Species)
Vertebrate Multiz Alignment & Conservation (17 Species)
Vertebrate Cons
chimp
mouse
rat
rabbit
dog
cow
elephant

FIGURE 4. **Examples of insights into phenotype gained by examining information from LSDBs along with functional information in the UCSC Genome Browser. On the Locus Variants track, purple = substitutions, blue = deletions. A:** ***ABO*** **recombination example. B:** ***PAH*** **deletion example. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com]**

[2004]). The ability of these two parts of the locus to recombine allows the disentanglement of the *APOA5* gene from the rest of the locus when searching for causative elements for the QTLs.

Example 4: The Second Phase: Extending to the Rest of the Genome

Phase 2 of the ENCODE project will extend the detailed biochemical analyses to full genome coverage. Thus, virtually all loci will be represented, and it will be important to incorporate information from all LSDBs and any available genome-wide sources, such as OMIM. The power of looking at data from LSDBs in combination with information already available throughout the human genome is illustrated in this section.

The Blood Group Antigen Gene Mutation Database (BGMUT) compiles information on variation in blood group antigens [Blumenfeld and Patnaik, 2004]. Sometimes "unresolved paternity issues" arise when a child has a different ABO phenotype from either parent. For instance, both parents may be blood group O, resulting from loss of function of the *ABO* gene product. However, one parent could be a compound heterozygote, and a meiotic recombination between the two homologs could restore function (e.g., giving blood group A) in one of the recombination products. The Genome Browser view reveals a hotspot for recombination (Fig. 4A) that helps to explain the appearance of unexpected blood group phenotypes.

A deficiency in phenylalanine hydroxylase, encoded by *PAH*, can lead to phenylketonuria or other problems with phenylalanine metabolism. Mutations in *PAH* and their resultant phenotypes are recorded in PAHdb [Scriver et al., 2003]. Many of these cause amino acid substitutions, but some are deletions. Examination of the Locus Variants track for the *PAH* locus in the Genome Browser (Fig. 4B) shows that one such deletion removes a segment of 5′ flanking DNA, including a highly conserved noncoding region (Conservation track) that is annotated in the ORegAnno track as a regulatory region. Information in PAHdb shows that this region is a liver-specific enhancer of *PAH* expression.

DISCUSSION

A number of projects related to PhenCode are also underway. The HGVS has initiated the WayStation (www.centralmutations.org) that will serve as a central distribution point for submission of new mutations into affiliated LSDBs. The Society also is working on a central composite database to collect and store LSDB data. Coordination of efforts will provide future connections between the WayStation data and/or central repository and the Genome Browser. As is true for most LSDB-related projects, limited funding remains a fundamental problem that slows realization of the potential of this field [Patrinos and Brookes, 2005].

OMIM [Hamosh et al., 2002] provides extensive phenotype data, but it fills a different role than PhenCode. It provides users with rich, detailed information in a prose form. However, it does not use a controlled vocabulary or uniform base numbering, which impedes automated analysis. HmutDb [Lehväslaiho, 2000] is a subset of SRS [Zdobnov et al., 2002] that incorporates phenotype data, and a prototype is currently available. It provides a reference sequence for each entry but not standardized chromosome coordinates. HGVbase [Fredman et al., 2002] is planning to add phenotype data. None of these resources currently provide connections to the Genome Browser (to our knowledge). Future developments of PhenCode and these other resources should strive for coordination to optimize the utility of their information.

PhenCode provides a seamless, bidirectional connection between LSDBs and ENCODE data at the UCSC Genome Browser, which allows users to easily explore phenotypes associated with functional elements and look for genomic data that could explain clinical phenotypes, thus helping to fulfill the promise of the Human Genome Project to improve human health. Close comparison of the tracks shown in the examples provides additional insights and suggests more experiments. Thus, PhenCode not only is helpful to clinicians for diagnostics, it also serves biomedical researchers by integrating multiple types of information and facilitating the generation of testable hypotheses to improve our understanding of both the functions of genomic DNA and the mechanisms by which it achieves those functions. Other rich data sources, such as genomewide gene expression profiles under many different conditions, are currently available for integration with the genotype and phenotype information described here. Genome-wide data on sites occupied by nuclear proteins are being published, and it is reasonable to expect many more in the near future. These and other types of data provide new opportunities to better explain phenotypes. Although the challenges of formalizing descriptions of phenotypes and fully integrating disparate data types are daunting, progress in addressing the challenges will likely provide exciting new insights into genetic functions.

ACKNOWLEDGMENTS

This work was supported by the NHGRI (HG002238 to W.M. and 1P41HG02371 to W.J.K.), NIH (DK65806 to R.H.), the Intramural Program of the NIH, NHGRI (to L.E.), Réseau de Médicine Génétique Appliquée/Fonds de la recherche en santé du Québec (PAHdb) (to C.S.), and the Canadian Cystic Fibrosis Foundation and Genome Canada through the Ontario Genomics Institute (to J.Z.).

REFERENCES

Bairoch A, Apweiler R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 30: 45–48.

Blumenfeld OO, Patnaik SK. 2004. Allelic genes of blood group antigens: a source of human mutations and cSNPs documented in the Blood Group Antigen Gene Mutation Database. Hum Mutat 23:8–16.

Cotton RG, McKusick V, Scriver CR. 1998. The HUGO Mutation Database Initiative. Science 279:10–11.

den Dunnen JT, Antonarakis SE. 2001. Nomenclature for the description of human sequence variations. Hum Genet 109:121–124.

den Dunnen JT, Antonarakis SE. 2002. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. Hum Mutat 15:7–12.

Driscoll MC, Dobkin CS, Alter BP. 1989. Gamma delta beta-thalassemia due to a de novo mutation deleting the 5′ beta-globin gene activation-region hypersensitive sites. Proc Natl Acad Sci USA 86:7470–7474.

ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306:636–640.

Forrester WC, Epner E, Driscoll MC, Enver T, Brice M, Papayannopoulou T, Groudine M. 1990. A deletion of the human beta-globin locus activation region causes a major alteration in chromatin structure and replication across the entire beta-globin locus. Genes Dev 4:1637–1649.

Fredman D, Siegfried M, Yuan YP, Bork P, Lehväslaiho H, Brookes AJ. 2002. HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. Nucleic Acids Res 30: 387–391.

Grosveld F, van Assendelft GB, Greaves DR, Kollias G. 1987. Position-independent, high-level expression of the human beta-globin gene in transgenic mice. Cell 51:975–985.

Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA. 2002. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 30:52–55.

Hardison RC, Chui D, Giardine B, Riemer C, Patrinos G, Anagnou N, Miller W, Wajcman H. 2002. HbVar: a relational database of human hemoglobin variants and thalassemia mutations at the globin gene server. Hum Mut 19:225–233.

Hubbard T, Andrews D, Caccamo M, Cameron G, Chen Y, Clamp M, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, Down T, Durbin R, Fernandez-Suarez XM, Gilbert J, Hammond M, Herrero J, Hotz H, Howe K, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Keenan S, Kokocinsci F, London D, Longden I, McVicker G, Melsopp C, Meidl P, Potter S, Proctor G, Rae M, Rios D, Schuster M, Searle S, Severin J, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Trevanion S, Ureta-Vidal A, Vogel J, White S, Woodwark C, Birney E. 2005. Ensembl 2005. Nucleic Acids Res 33:D447–D453.

Kao J-T, Wen H-C, Chien K-L, Hsu H-C, Lin S-W. 2003. A novel genetic variant in the apolipoprotein A5 gene is associated with hypertriglyceridemia. Hum Mol Genet 12:2533–2539.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC Table Browser data retrieval tool. Nucleic Acids Res 32:D493–D496.

Kent WJ. 2002. BLAT: the BLAST-like alignment tool. Genome Res 12: 656–664.

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.

King DC, Taylor J, Elnitski L, Chiaromonte F, Miller W, Hardison RC. 2005. Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. Genome Res 15:1051–1060.

Lai C-Q, Parnell LD, Ordovas JM. 2005. The APOA1/C3/A4/A5 gene cluster, lipid metabolism, and cardiovascular disease risk. Curr Opin Lipidology 16:153–166.

Lehväslaiho H. 2000. Sequence variation and mutation databases. Brief Bioinform 1:161–166.

Li Q, Peterson KR, Fang X, Stamatoyannopoulos G. 2002. Locus control regions. Blood 100:3077–3086.

Montgomery SB, Griffith OL, Sleumer MC, Bergman CM, Bilenky M, Pleasance ED, Prychyna Y, Zhang X, Jones SJM. 2006. OregAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites, and regulatory variation. Bioinformatics 22:637–640.

Olivier M, Wang X, Cole R, Gau B, Kim J, Rubin EM, Pennacchio LA. 2004. Haplotype analysis of the apolipoprotein gene cluster on human chromosome 11. Genomics 83:912–923.

Patrinos GP, Giardine B, Riemer C, Miller W, Chui DHK, Anagnou NP, Wajcman H, Hardison RC. 2004. Improvements in the HbVar database of human hemoglobin variants and thalassemia mutations for population and sequence variation studies. Nucleic Acids Res 32:D537–D541.

Patrinos GP, Brookes AJ. 2005. DNA, diseases and databases: disastrously deficient. Trends Genet 21:333–338.

Pennacchio LA, Olivier M, Hubacek JA, Krauss RM, Rubin EM, Cohen JC. 2002. Two independent apolipoprotein A5 haplotypes influence human plasma triglyceride levels. Hum Mol Genet 11:3031–3038.

Scriver CR, Hurtubise M, Konecki D, Phommarinh M, Prevost L, Erlandsen H, Stevens R, Waters PJ, Ryan S, McDonald D, Sarkissian C. 2003. PAHdb 2003: what a locus-specific knowledgebase can do. Hum Mutat 21:333–344.

Siepel A, Bejerano G, Pederson JS, Hinrichs A, Hou M, Rosenbloom K, Clawson J, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15:1034–1050.

Tuan DY, Solomon WB, London IM, Lee DP. 1989. An erythroid-specific, developmental-stage-independent enhancer far upstream of the human "beta-like globin" genes. Proc Natl Acad Sci USA 86:2554–2558.

Vrablik M, Horinek A, Ceska R, Adamkova V, Poledne R, Hubacek JA. 2003. Ser19→Trp polymorphism within the apolipoprotein AV gene in hypertriglyceridaemic people. J Med Genet 40:E105.

Zdobnov EM, Lopez R, Apweiler R, Etzold T. 2002. The EBI SRS server: new features. Bioinformatics 18:1149–1150.