

# **GS01 0163**

## **Analysis of Microarray Data**

Keith Baggerly and Brad Broom  
Department of Bioinformatics and Computational Biology  
UT M. D. Anderson Cancer Center

`kabagg@mdanderson.org`

`bmbroom@mdanderson.org`

2 November 2010

# Lecture 18: Genome Browsing

- Annotation Environments in R
- AnnBuilder: Rolling Your Own Annotations
- The UCSC Genome Browser
- Chromosome Locations
- Building a Custom Track
- Viewing Your Custom Track
- Thoughts about TCGA

# Documentation for the AnnBuilder Package

The screenshot shows a Mozilla Firefox browser window displaying the AnnBuilder package documentation page. The address bar shows the URL: <http://bioconductor.org/packages/1.8/bioc/html/AnnBuilder.html>. The page content includes the following sections:

## AnnBuilder

### Bioconductor annotation data package builder

Processing annotation data from public data repositories and building annotation data packages or XML data documents using the source data.

Author J. Zhang  
Maintainer J. Zhang

**Vignettes (Documentation)**

- [ABPrimer.pdf](#)
- [AnnBuilder.pdf](#)

**Package Downloads**

Source	<a href="#">AnnBuilder 1.10.5.tar.gz</a>
Windows binary	<a href="#">AnnBuilder 1.10.5.zip</a>
OS X binary	<a href="#">AnnBuilder 1.10.5.tgz</a>

**Details**

biocViews	<a href="#">Annotation</a> , <a href="#">Microarray</a>
Depends	R, methods, Biobase, XML, annotate, utils, RSQLite
Suggests	
Imports	
SystemRequirements	

Done

## Annotation Environments in R

For most Affymetrix arrays, annotation packages are available directly (and automatically) from BioConductor whenever you need them. These packages were built using `AnnBuilder`.

You can load one of these packages as follows:

```
> require(hgu95av2.db)
```

To see what is in an annotation package, use its name as a function:

```
> hgu95av2()
```

```
Quality control information for hgu95av2:
```

This package has the following mappings:

hgu95av2ACCNUM has 12625 mapped keys (of 12625 keys)  
hgu95av2ALIAS2PROBE has 37934 mapped keys (of 37934 keys)  
hgu95av2CHR has 11957 mapped keys (of 12625 keys)  
hgu95av2CHRENGTHS has 25 mapped keys (of 25 keys)  
hgu95av2CHRLOC has 11789 mapped keys (of 12625 keys)  
hgu95av2CHRLOCEND has 11789 mapped keys (of 12625 keys)  
hgu95av2ENSEMBL has 11639 mapped keys (of 12625 keys)  
hgu95av2ENSEMBL2PROBE has 9021 mapped keys (of 9021 keys)  
hgu95av2ENTREZID has 11960 mapped keys (of 12625 keys)  
hgu95av2ENZYZME has 1978 mapped keys (of 12625 keys)  
hgu95av2ENZYZME2PROBE has 725 mapped keys (of 725 keys)  
hgu95av2GENENAME has 11960 mapped keys (of 12625 keys)  
hgu95av2GO has 11363 mapped keys (of 12625 keys)  
hgu95av2GO2ALLPROBES has 9581 mapped keys (of 9581 keys)  
hgu95av2GO2PROBE has 6774 mapped keys (of 6774 keys)

hgu95av2MAP has 11919 mapped keys (of 12625 keys)  
hgu95av2OMIM has 10350 mapped keys (of 12625 keys)  
hgu95av2PATH has 4585 mapped keys (of 12625 keys)  
hgu95av2PATH2PROBE has 203 mapped keys (of 203 keys)  
hgu95av2PFAM has 11878 mapped keys (of 12625 keys)  
hgu95av2PMID has 11898 mapped keys (of 12625 keys)  
hgu95av2PMID2PROBE has 206993 mapped keys (of 206993 keys)  
hgu95av2PROSITE has 11878 mapped keys (of 12625 keys)  
hgu95av2REFSEQ has 11883 mapped keys (of 12625 keys)  
hgu95av2SYMBOL has 11960 mapped keys (of 12625 keys)  
hgu95av2UNIGENE has 11905 mapped keys (of 12625 keys)  
hgu95av2UNIPROT has 11764 mapped keys (of 12625 keys)

Additional Information about this package:

DB schema: HUMANCHIP\_DB

DB schema version: 1.0

Organism: Homo sapiens

Date for NCBI data: 2009-Mar11

Date for GO data: 200903

Date for KEGG data: 2009-Mar10

Date for Golden Path data: 2008-Sep3

Date for IPI data: 2009-Mar03

Date for Ensembl data: 2009-Mar6

# Getting Annotations From Environments

Each of the items in the package is an *environment*, which computer scientists may recognize better if we tell them it is a hash table. The key into the probe-based hash table environments is the manufacturer's identifier (i.e., an Affymetrix probeset id such as `1854_at`).

```
> get("1854_at", hgu95av2MAP)
[1] "20q13.1"
> get("1854_at", hgu95av2CHRLOC)
      20
41729122
> get("1854_at", hgu95av2ENTREZID)
[1] "4605"
```



## More Getting Annotations From Environments

```
get("1854_at", hgu95av2REFSEQ)
[1] "NM_002466" "NP_002457"
> summary(hgu95av2REFSEQ)
REFSEQ map for chip hgu95av2 (object of class "Annot")
|
| Lkeyname: probe_id (Ltablename: probes)
|   Lkeys: "1000_at", "1001_at", ... (total=1262)
|
| Rkeyname: accession (Rtablename: refseq)
|   Rkeys: "NM_000015", "NM_000016", ... (total=2)
|
| direction: L --> R
get("NM_002466", revmap(hgu95av2REFSEQ))
[1] "1854_at"
```

# AnnBuilder: Rolling Your Own Annotations

We recently had to analyze some data from an Agilent 44K two-color glass microarray. The corresponding annotation package was not available, so we had to build our own. Finding the manufacturers basic annotations was a nontrivial task. We started at the web site (<http://www.agilent.com>), then followed the link under “Products and Services” for “Life Sciences” and “Instruments and Systems” to get to the “DNA Microarrays” page.

# Follow the Link for “Human Genome, Whole”

The screenshot shows the Agilent Technologies website for DNA Microarrays. The browser window title is "Agilent | DNA Microarrays - Mozilla Firefox". The address bar shows the URL: <http://www.chem.agilent.com/Scripts/PCol.asp?Page=494>. The page header includes the Agilent logo and the text "Life Sciences & Chemical Analysis". Below the header is a navigation menu with tabs for "Products & Services", "Technical Support", "Industries", "Buy", and "About Agilent". A search bar is located to the right of the navigation menu. The main content area is titled "DNA Microarrays" and features a product image of a microarray chip. To the right of the image is a text block: "Optimize your experimental design. Agilent Printed Microarray Solutions are based on an integrated, flexible, open approach for successful gene expression analysis. Each component is designed to work with others in our system, or with your existing set-up. Allowing you to express it your way. Buy just what you need or purchase the entire system." Below this text is a "Buy" section with links for "Request a quote", "Where to buy", and "Store Home". At the bottom of the page, there are three columns: "Related Information" with links like "Literature Library", "Applications", and "Technical Notes"; "DNA Microarrays" with a list of species including Arabidopsis 2 (V2), Arabidopsis 3, C. elegans, Canine, Human 1A (V2), M. grisea 2.0, Mouse (V2), Mouse Development 44K, Rat (V2), Rhesus Monkey, Rice, Yeast (V2), Whole Human Genome (new), Whole Mouse Genome (new), Whole Rat Genome (new), Xenopus laevis, Zebrafish, and Custom Gene; and "Announcements" with a highlighted "Agilent Multi-Pack Gene Expression Microarrays" section for Human, Mouse, and Rat, and a "Product Announcement" for "New 1.3 Version - ChIP Analytics Software".

# Try “Download Gene Lists (Specifications)”

Agilent | DNA Microarrays - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.chem.agilent.com/Scripts/PCol.asp?Page=494

Google Project Tracker Entrez-PubMed MDACC Bioinfo Microarray Core Faci... Wiki: BiomarkerJour...

**Agilent Technologies** Life Sciences & Chemical Analysis

Select a Country or Area | Contact Us

Products & Services Technical Support Industries Buy About Agilent

Search

Home Register Login

## DNA Microarrays

**Optimize your experimental design**  
Agilent Printed Microarray Solutions are based on an integrated, flexible, open approach for successful gene expression analysis. Each component is designed to work with others in our system, or with your existing set-up. Allowing you to express it your way. Buy just what you need or purchase the entire system.

**Buy**

Request a quote  
Where to buy  
Store Home

**Related Information**

- Literature Library
- Applications
- Technical Notes
- Brochures
- Posters
- Scientific Publications
- Manuals
- more...
- Technical Support
- Frequently Asked Questions
- Design with eArray
- more...

**DNA Microarrays** + expand all - close all

Gene Expression

- Arabidopsis 2 (V2)
- Arabidopsis 3
- C. elegans*
- Canine
- Human 1A (V2)
- M. grisea* 2.0
- Mouse (V2)
- Mouse Development 44K
- Rat (V2)
- Rhesus Monkey
- Rice
- Yeast (V2)
- new Whole Human Genome
- new Whole Mouse Genome
- new Whole Rat Genome
- Xenopus laevis*
- Zebrafish
- new Custom Gene Expression

**Announcements**

**Agilent Multi-Pack Gene Expression Microarrays**

Human Mouse Rat

Product Announcement  
New 1.3 Version - ChIP Analytics Software

Support  
Feature Extraction Software 9.1.3 patch

Done

## Reading the Feature Info

In any event, we finally obtained a pair of files that contained the mappings from spots to genomic material. We used the `read.table` command to get this file into R:

```
featureInfo <-  
  read.table(file.path("GeneList",  
    "014850_D_GeneList_20090416.txt"),  
    header = TRUE, row.names = NULL,  
    sep = "\t", quote = "",  
    comment.char = "")
```

## Looking at the Feature Info

Here is part of the file:

```
> colnames(featureInfo)
[1] "ProbeID"      "TargetID"      "GeneSymbol"
[4] "GeneName"     "Accessions"    "Description"
> featureInfo[1:5, c("ProbeID", "Accessions")]
      ProbeID
1 A_24_P919016      gb|A18395|thc|NP238262|t
2  A_32_P27041  gb|A18658|gb|AI819757|gb|AW299939|c
3 A_24_P693768      gb|A18658|gb|BQ367072|c
4 A_24_P475014      gb|AA004321|gb|T84046|gb|T8372
5 A_24_P456043      gb|AA004800|c
```

## What We Need

The critical information is given by the columns that contain the manufacturers identifier (`ProbeID`) and the GenBank or RefSeq accession numbers (`Accessions`). Ideally, we want one type of annotation.

```
allAnnotations <-  
  as.character(featureInfo[, "Accessions"])  
splitAnnotations <-  
  strsplit(allAnnotations, "\\|")  
firstAnnotation <-  
  lapply(splitAnnotations, function(x) {x[1]})  
table(unlist(firstAnnotation))  
   ens      gb      ref      thc  
  909   7988  26631   2441
```

## What We Need

The function we are going to use to build annotations requires only these two columns to be present in a file. So we make them available for a few genes:

```
secondAnnotation <-  
  unlist(lapply(splitAnnotations, function(x) {x[2]}  
temp <-  
  cbind(as.character(featureInfo[, "ProbeID"]),  
        secondAnnotation)  
write.table(temp[1:10, ], "agilentGenesShort.tsv",  
  sep="\t", quote=FALSE, col.names=NA)
```



# Setting Up the Annotation Package

```
> library(AnnBuilder)
> baseName <- "agilentGenes.tsv"
> baseType <- "gb"
> srcUrls <-
  getSrcUrl("all",
    organism = "Homo sapiens")
> myDir <- getwd()
```

## Building the Annotation Package

The next command takes a **very** long time, since it makes calls to databases all over the internet for every one of the 44,000 probes on the array.

```
ABPkgBuilder(baseName = baseName,  
  srcUrls = srcUrls, baseMapType = baseType,  
  pkgName = "Agilent44K", pkgPath = myDir,  
  organism = "Homo sapiens", version = "1.0",  
  author = list(authors = "krc@mdacc.tmc.edu",  
    maintainer = "krc@mdacc.tmc.edu"),  
  fromWeb = TRUE)
```

## Producing the Final Package

This command produces the **source** for a package, which must still be compiled and zipped into a binary package that can be installed easily. This task is most easily accomplished on a UNIX based machine:

```
helios% R CMD build Agilent44K
```

```
helios% R CMD build --binary Agilent44K
```

You can then convert the resulting `.tar.gz` file to a `.zip` file, which is the preferred form for distributing a Windows package.

You can check out the results by getting the annotation package from our course web site.

# The Agilent 44K Annotations

```
> library(Agilent44K)
```

```
> Agilent44K()
```

```
Quality control information for Agilent44K
```

```
Date built: Created: Sun Sep 03 07:50:38 2006
```

```
Number of probes: 41001
```

```
Probe number mismatch: None
```

```
Probe mismatch: None
```

```
Mappings found for probe based rda files:
```

```
Agilent44KACCNUM found 41001 of 41001
```

```
Agilent44KCHR found 31185 of 41001
```

```
Agilent44KCHRLOC found 28795 of 41001
```

```
Agilent44KENZYME found 3056 of 41001
```

```
Agilent44KGENENAME found 27824 of 41001
```

Agilent44KGO found 23644 of 41001

Agilent44KLOCUSID found 31224 of 41001

Agilent44KMAP found 30939 of 41001

Agilent44KOMIM found 17942 of 41001

Agilent44KPATH found 6715 of 41001

Agilent44KPMID found 30361 of 41001

Agilent44KREFSEQ found 30057 of 41001

Agilent44KSUMFUNC found 0 of 41001

Agilent44KSYMBOL found 31217 of 41001

Agilent44KUNIGENE found 31010 of 41001

Mappings found for non-probe based rda files:

Agilent44KCHRLLENGTHS found 25

Agilent44KENZYME2PROBE found 794

Agilent44KGO2ALLPROBES found 6883

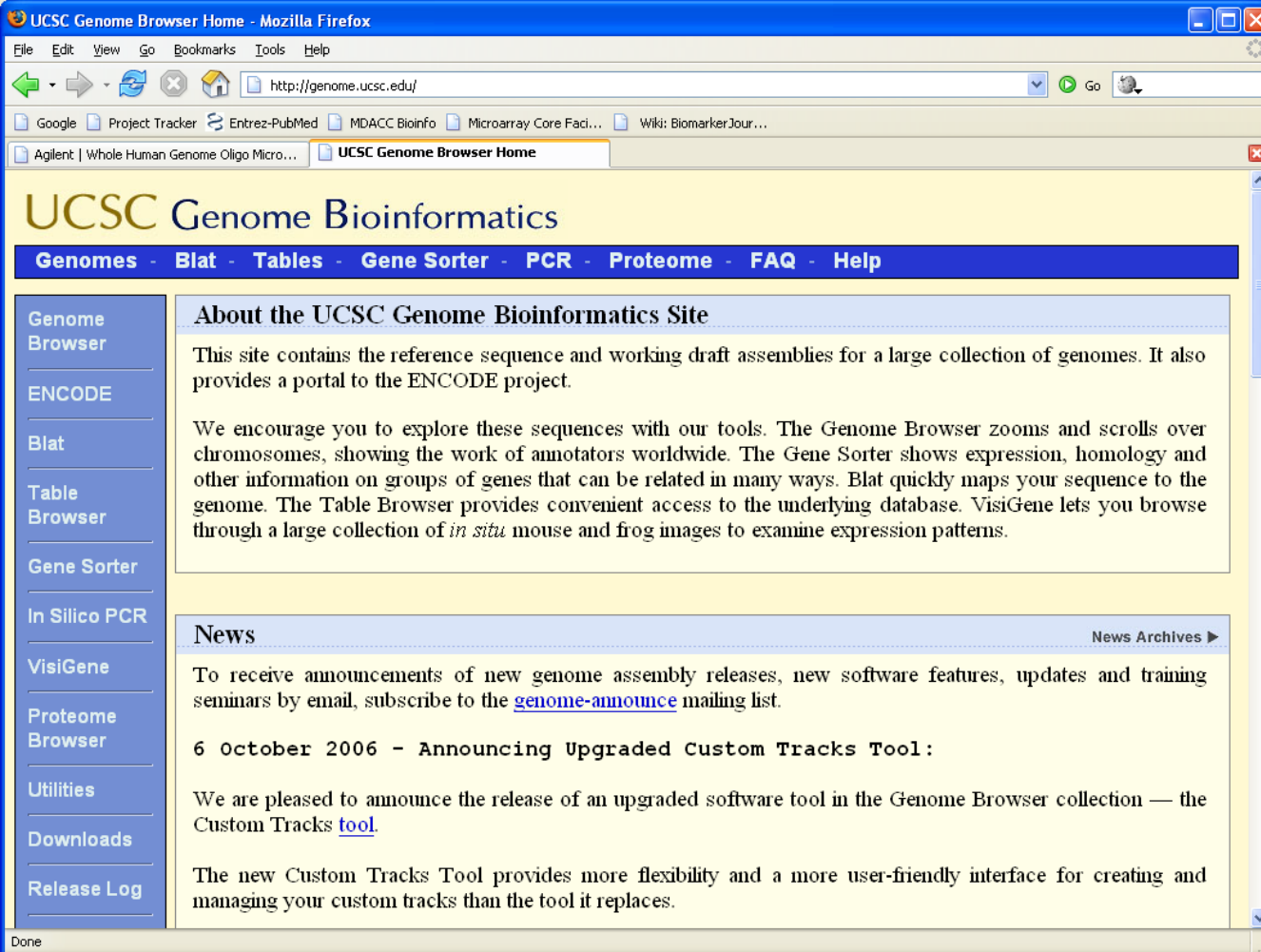
Agilent44KGO2PROBE found 5117

---

Agilent44KORGANISM found 1  
Agilent44KPATH2PROBE found 183  
Agilent44KPFAM found 21902  
Agilent44KPMID2PROBE found 131104  
Agilent44KPROSITE found 15055

# The UCSC Genome Browser

`http://genome.ucsc.edu/`



The screenshot shows the UCSC Genome Browser homepage in Mozilla Firefox. The browser window title is "UCSC Genome Browser Home - Mozilla Firefox". The address bar shows "http://genome.ucsc.edu/". The page features a navigation menu with links: Genomes, Blat, Tables, Gene Sorter, PCR, Proteome, FAQ, and Help. A sidebar on the left contains links to Genome Browser, ENCODE, Blat, Table Browser, Gene Sorter, In Silico PCR, VisiGene, Proteome Browser, Utilities, Downloads, and Release Log. The main content area includes "About the UCSC Genome Bioinformatics Site" and "News".

**UCSC Genome Bioinformatics**

[Genomes](#) - [Blat](#) - [Tables](#) - [Gene Sorter](#) - [PCR](#) - [Proteome](#) - [FAQ](#) - [Help](#)

**Genome Browser**

**ENCODE**

**Blat**

**Table Browser**

**Gene Sorter**

**In Silico PCR**

**VisiGene**

**Proteome Browser**

**Utilities**

**Downloads**

**Release Log**

**About the UCSC Genome Bioinformatics Site**

This site contains the reference sequence and working draft assemblies for a large collection of genomes. It also provides a portal to the ENCODE project.

We encourage you to explore these sequences with our tools. The Genome Browser zooms and scrolls over chromosomes, showing the work of annotators worldwide. The Gene Sorter shows expression, homology and other information on groups of genes that can be related in many ways. Blat quickly maps your sequence to the genome. The Table Browser provides convenient access to the underlying database. VisiGene lets you browse through a large collection of *in situ* mouse and frog images to examine expression patterns.

**News** [News Archives](#)

To receive announcements of new genome assembly releases, new software features, updates and training seminars by email, subscribe to the [genome-announce](#) mailing list.

**6 October 2006 - Announcing Upgraded Custom Tracks Tool:**

We are pleased to announce the release of an upgraded software tool in the Genome Browser collection — the Custom Tracks [tool](#).

The new Custom Tracks Tool provides more flexibility and a more user-friendly interface for creating and managing your custom tracks than the tool it replaces.

# Follow the Link to “Genome Browser”

The screenshot shows a Mozilla Firefox browser window displaying the UCSC Genome Browser Gateway. The address bar shows the URL <http://genome.ucsc.edu/cgi-bin/hgGateway>. The browser's menu bar includes File, Edit, View, Go, Bookmarks, Tools, and Help. The page title is "Human (Homo sapiens) Genome Browser Gateway".

The website content includes a navigation menu with links for Home, Genomes, Blat, Tables, Gene Sorter, PCR, FAQ, and Help. Below the menu, the page title "Human (*Homo sapiens*) Genome Browser Gateway" is displayed. A paragraph states: "The UCSC Genome Browser was created by the [Genome Bioinformatics Group of UC Santa Cruz](#). Software Copyright (c) The Regents of the University of California. All rights reserved."

A search form is present with the following fields and values:

clade	genome	assembly	position or search term	image width	
Vertebrate	Human	Mar. 2006	chr3:110,765,251-122,424,750	620	<input type="button" value="submit"/>

Below the search form, there is a link: [Click here to reset the browser user interface settings to their defaults.](#) and three buttons: , , and .

The page also features a section titled "About the Human Mar. 2006 (hg18) assembly ([sequences](#))". The text reads: "The March 2006 human reference sequence (NCBI Build 36.1) was produced by the International Human Genome Sequencing Consortium."

Under the heading "Sample position queries", the text explains: "A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, or a cytological band, a chromosomal coordinate range, or keywords from the GenBank description of an mRNA. The following list shows examples of valid position queries for the human genome. See the [User's Guide](#) for more information."

A table provides examples of queries and their corresponding browser responses:

Request:	Genome Browser Response:
chr7	Displays all of chromosome 7

The browser status bar at the bottom shows "Done".



# Press “Submit” to Start Browsing

Human chr3:110,765,251-122,424,750 - UCSC Genome Browser v144 - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://genome.ucsc.edu/cgi-bin/hgTracks?clade=vertebrate&org=Human&db=hg18&position=chr3%3A110%2C765%2C251-122

Google Project Tracker Entrez-PubMed MDACC Bioinfo Microarray Core Faci... Wiki: BiomarkerJour...

Agilent | Whole Human Genome Oligo Micro... Human chr3:110,765,251-122,424,750

Home Genomes Blat Tables Gene Sorter PCR DNA Convert PDF/PS Help

UCSC Genome Browser on Human Mar. 2006 Assembly

move <<<< << < > >> >>>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr3:110,765,251-122,424,750 jump clear size 11,659,500 bp. configure

chr3 (q13.13-q13.33)

chr3: 115000000 120000000

UCSC Known Genes Based on UniProt, RefSeq, and GenBank mRNA

PVRL3 GCET2 BOC VS109 GPR43 IGSF11 FSTL1

PVRL3 CD200 BOC VS109 LSAMP IGSF11 NDUFB4

PVRL3 BTLA BOC AF119886 IGSF11 HGD

CD96 ATG3 MARK3 C3orf30 HGD

CD96 ATG3 MARK3 UPK1B RABL3

CD96 URB ZDHHC23 B4GALT4 GTF2E1

ZBED2 URB ZDHHC23 CDGAP GTF2E1

ZBED2 CD200R1 DRD3 TMEM39A STXBPSL

PLCKD2 GTFBP8 DRD3 TMEM39A

PHLDB2 BOC DRD3 C3orf9

AY553772 GTFBP8 DRD3 C3orf9

PHLDB2 BOC C3orf1

PHLDB2 WDR52 CD89

PHLDB2 WDR52 ADPRH

AK074525 CCDC52 FLA1A

ABHD10 SIDT1 FLA1A

TAGLN3 KIAA2018 FLA1A

TMFRS7 ATP6V1A POPDC2

C3orf52 GRAMD1C COX17

C3orf52 GRAMD1C C3orf15

BC041355 KIAA1487 C3orf15

SLC9A1 QTRTD1 C3orf15

SLC9A1 DRD3 C3orf15

SLC35A5 DRD3 C3orf15

SLC35A5 ZNF80 NR112

CD200R2 ZBTB20 NR112

CD200R2 ZBTB20 NR112

CD200R1 ZBTB20 NR112

GTFBP8 NR112

GTFBP8 NR112

C3orf17 GSK3B

C3orf17 GSK3B

AK128280 AK128198

RefSeq Genes

LOC151760 GCET2 BOC VS109 IGSF11 GPR156

PVRL3 GCET2 WDR52 GPR43 IGSF11 FSTL1

CD96 ATG3 NAT13 C3orf30 NDUFB4

CD96 CD200R1 DRD3 UPK1B HGD

ZBED2 BTLA SIDT1 B4GALT4 RABL3

PLCKD2 GTFBP8 DRD3 B4GALT4 GTF2E1

PHLDB2 CCDC52 CDGAP

ABHD10 KIAA2018 TMEM39A

Done

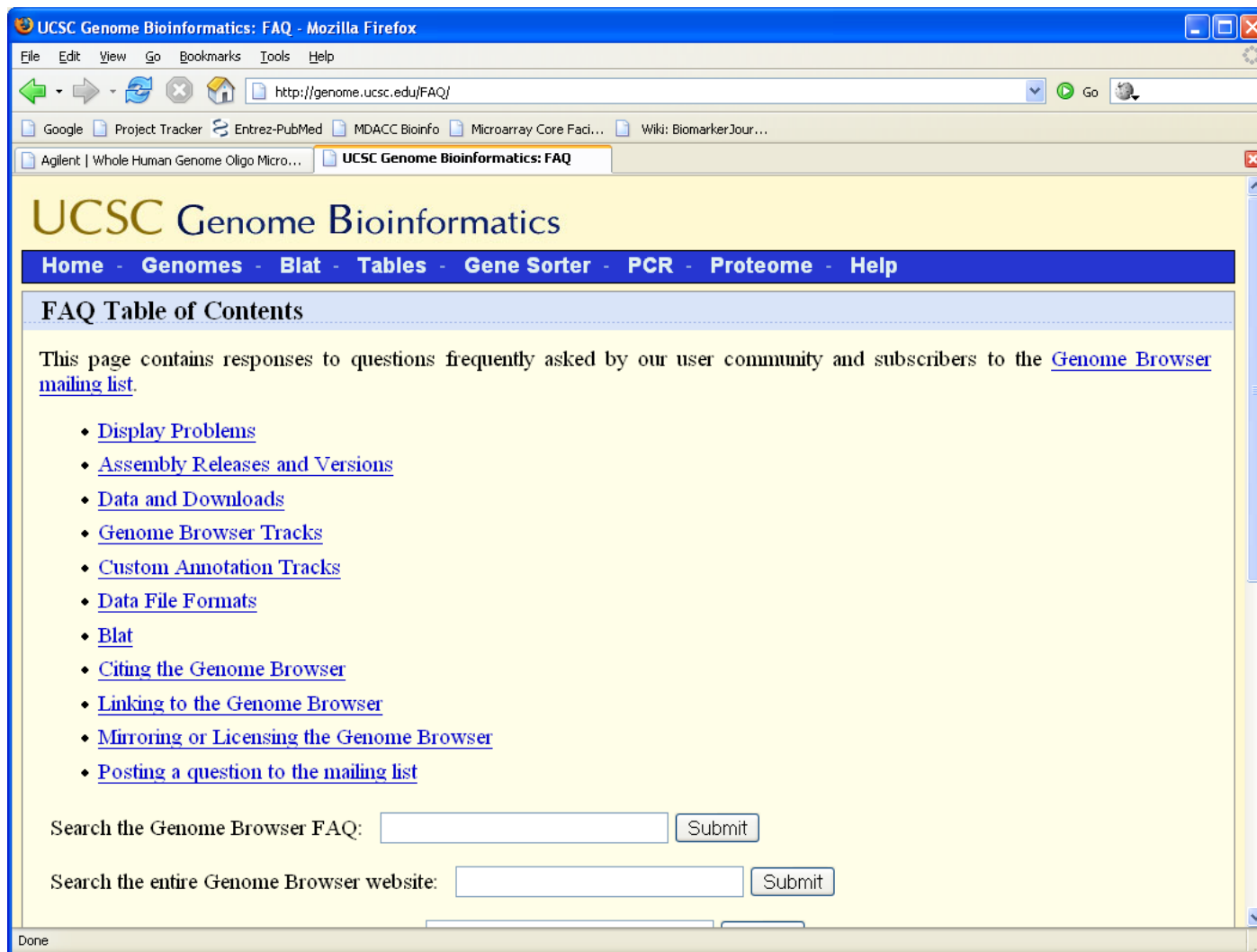
## About the Genome Browser

The genome browser lets you see a great deal of information laid out along the latest completed build of the human genome. The most obvious thing to look at are the known genes, which are typically displayed in such a way that you can see the individual introns and exons (provided you zoom in closely).

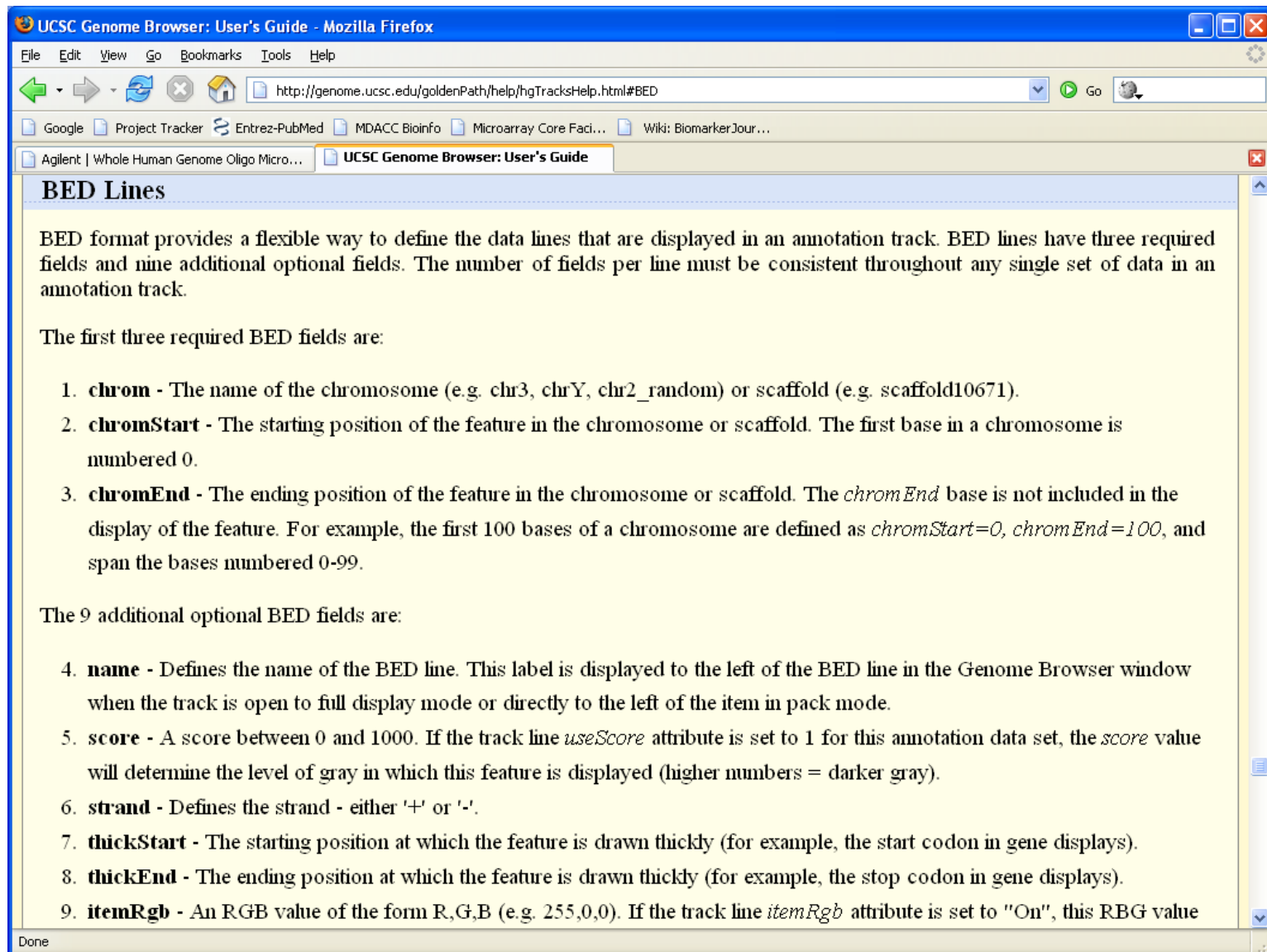
For our purposes (as people who analyze microarray data), an extremely interesting feature of the Genome Browser is that it lets you add your own “Custom Tracks”, which is their name for a set of annotations you can define.

# Custom Tracks

To learn about the genome (custom) tracks, go to the FAQ.



# BED Format



The screenshot shows a Mozilla Firefox browser window displaying the UCSC Genome Browser User's Guide page for BED format. The page title is "BED Lines". The content explains that BED format provides a flexible way to define data lines in an annotation track, with three required fields and nine optional fields. The required fields are: 1. **chrom** - The name of the chromosome or scaffold. 2. **chromStart** - The starting position of the feature. 3. **chromEnd** - The ending position of the feature. The optional fields are: 4. **name** - Defines the name of the BED line. 5. **score** - A score between 0 and 1000. 6. **strand** - Defines the strand. 7. **thickStart** - The starting position at which the feature is drawn thickly. 8. **thickEnd** - The ending position at which the feature is drawn thickly. 9. **itemRgb** - An RGB value of the form R,G,B.

UCSC Genome Browser: User's Guide - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html#BED

Google Project Tracker Entrez-PubMed MDACC Bioinfo Microarray Core Faci... Wiki: BiomarkerJour...

Agilent | Whole Human Genome Oligo Micro... UCSC Genome Browser: User's Guide

## BED Lines

BED format provides a flexible way to define the data lines that are displayed in an annotation track. BED lines have three required fields and nine additional optional fields. The number of fields per line must be consistent throughout any single set of data in an annotation track.

The first three required BED fields are:

1. **chrom** - The name of the chromosome (e.g. chr3, chrY, chr2\_random) or scaffold (e.g. scaffold10671).
2. **chromStart** - The starting position of the feature in the chromosome or scaffold. The first base in a chromosome is numbered 0.
3. **chromEnd** - The ending position of the feature in the chromosome or scaffold. The *chromEnd* base is not included in the display of the feature. For example, the first 100 bases of a chromosome are defined as *chromStart=0*, *chromEnd=100*, and span the bases numbered 0-99.

The 9 additional optional BED fields are:

4. **name** - Defines the name of the BED line. This label is displayed to the left of the BED line in the Genome Browser window when the track is open to full display mode or directly to the left of the item in pack mode.
5. **score** - A score between 0 and 1000. If the track line *useScore* attribute is set to 1 for this annotation data set, the *score* value will determine the level of gray in which this feature is displayed (higher numbers = darker gray).
6. **strand** - Defines the strand - either '+' or '-'.
7. **thickStart** - The starting position at which the feature is drawn thickly (for example, the start codon in gene displays).
8. **thickEnd** - The ending position at which the feature is drawn thickly (for example, the stop codon in gene displays).
9. **itemRgb** - An RGB value of the form R,G,B (e.g. 255,0,0). If the track line *itemRgb* attribute is set to "On", this RGB value

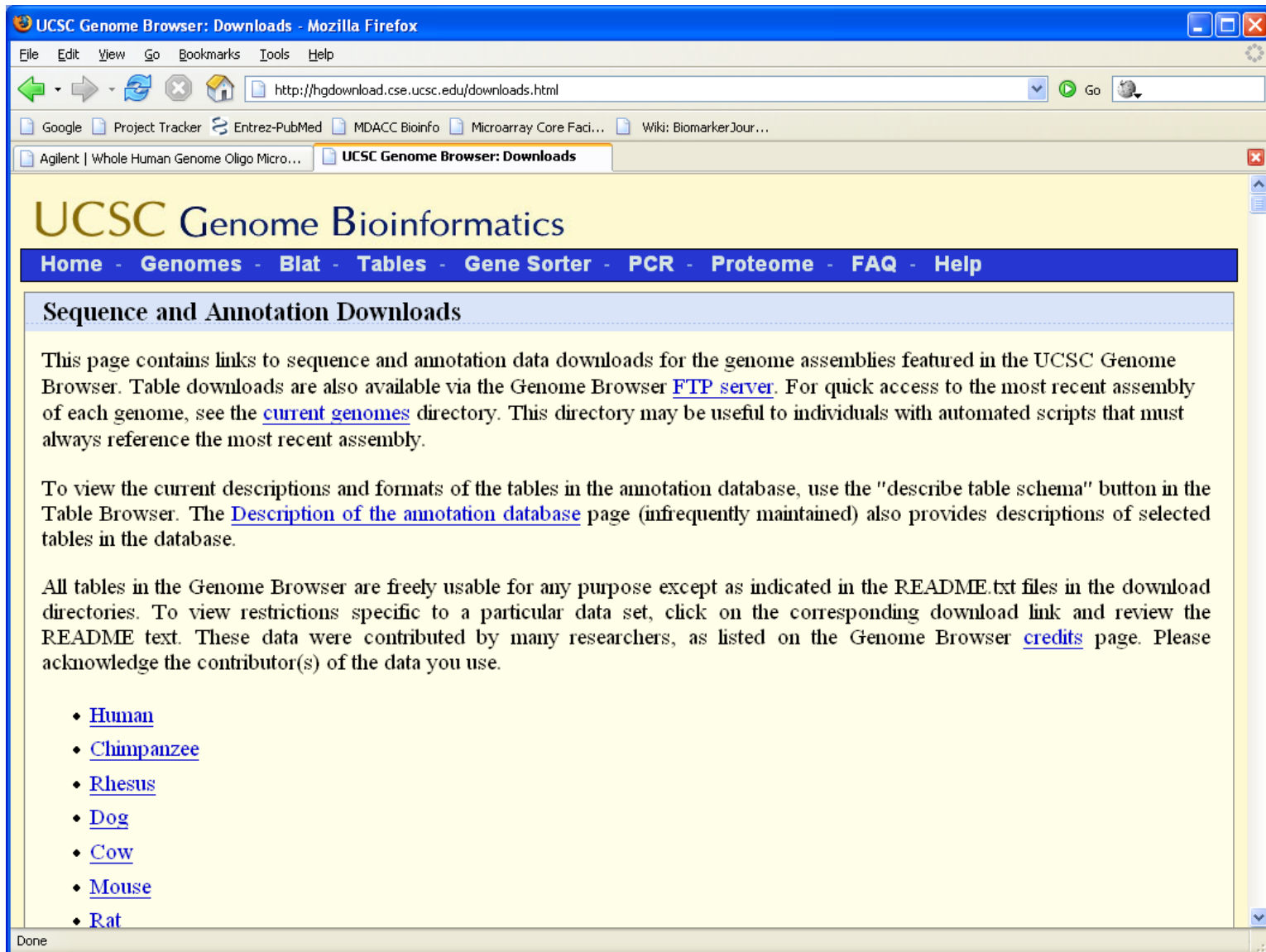
Done

# Chromosome Locations

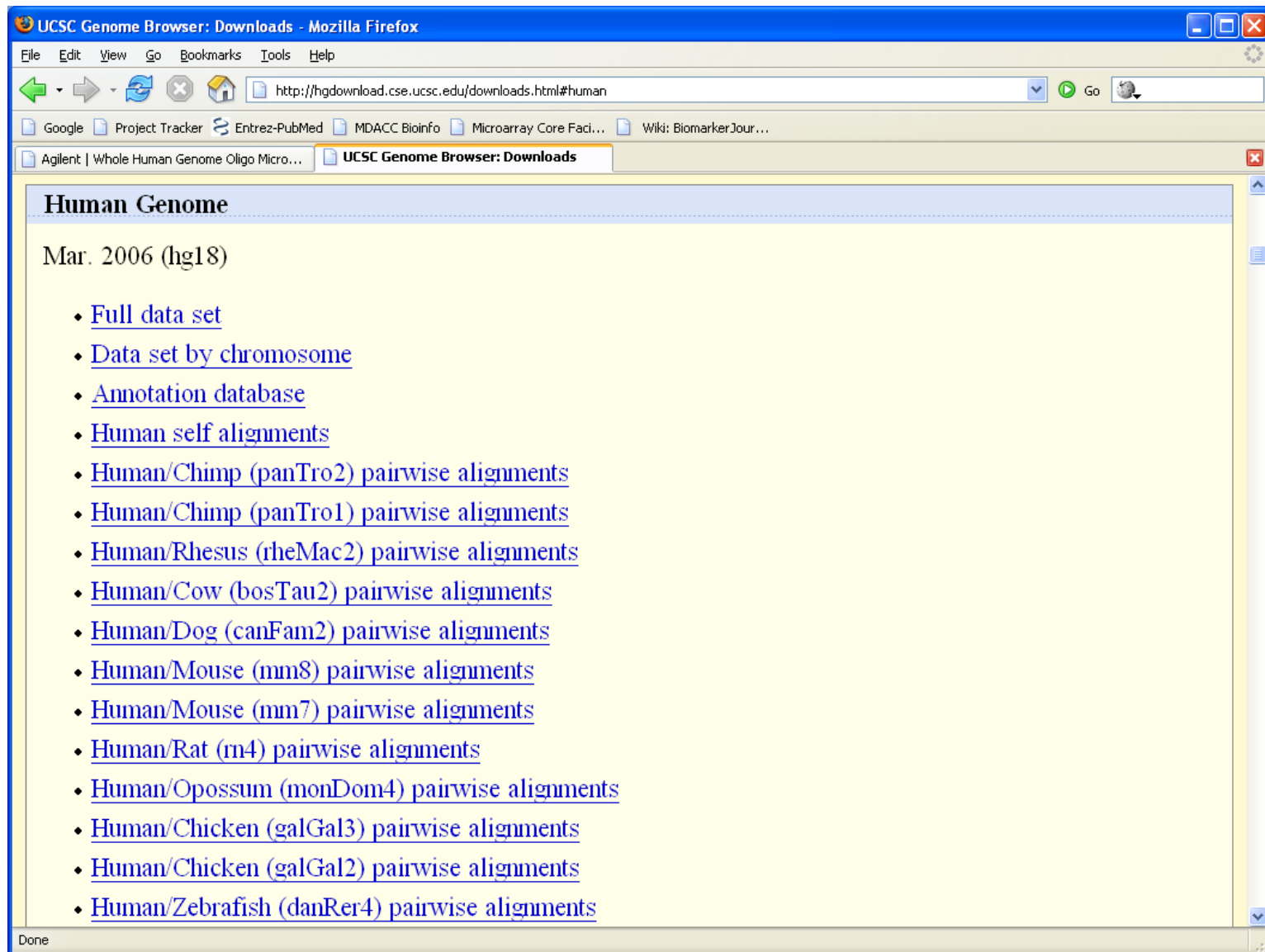
You can read more of the custom track documentation on your own; here, we are going to focus on how to build a custom track in R. The first thing we want to point out is that we need to know both the starting base location and the ending base location in order to build a custom track. Thus, the `CHRLOC` annotations that `AnnBuilder` constructs are not adequate.

Fortunately, we can get start and end points directly from the folks at the UCSC Genome Browser. Go back to the main page, then follow the link for “Downloads”.

# UCSC Download Page



# Follow the link for “Human”



The screenshot shows a Mozilla Firefox browser window titled "UCSC Genome Browser: Downloads - Mozilla Firefox". The address bar contains the URL "http://hgdownload.cse.ucsc.edu/downloads.html#human". The browser's menu bar includes "File", "Edit", "View", "Go", "Bookmarks", "Tools", and "Help". The address bar also shows a "Go" button and a search icon. Below the address bar, there are several bookmarks: "Google", "Project Tracker", "Entrez-PubMed", "MDACC Bioinfo", "Microarray Core Faci...", and "Wiki: BiomarkerJour...". The main content area of the browser displays the "UCSC Genome Browser: Downloads" page for the "Human Genome". The page title is "Human Genome" and the version is "Mar. 2006 (hg18)". A list of links is provided, each preceded by a blue diamond bullet point:

- [Full data set](#)
- [Data set by chromosome](#)
- [Annotation database](#)
- [Human self alignments](#)
- [Human/Chimp \(panTro2\) pairwise alignments](#)
- [Human/Chimp \(panTro1\) pairwise alignments](#)
- [Human/Rhesus \(rheMac2\) pairwise alignments](#)
- [Human/Cow \(bosTau2\) pairwise alignments](#)
- [Human/Dog \(canFam2\) pairwise alignments](#)
- [Human/Mouse \(mm8\) pairwise alignments](#)
- [Human/Mouse \(mm7\) pairwise alignments](#)
- [Human/Rat \(rn4\) pairwise alignments](#)
- [Human/Opossum \(monDom4\) pairwise alignments](#)
- [Human/Chicken \(galGal3\) pairwise alignments](#)
- [Human/Chicken \(galGal2\) pairwise alignments](#)
- [Human/Zebrafish \(danRer4\) pairwise alignments](#)

The status bar at the bottom of the browser window shows "Done".

# In “Annotation Database”, Scroll To “refGene”

The screenshot shows a Mozilla Firefox browser window with the address bar displaying `http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/`. The browser's address bar and tabs are visible at the top. The main content area shows a directory listing of files. The file `refGene.txt.gz` is highlighted in blue, indicating it is the current selection. The listing includes file names, their last modified dates and times, and their sizes.

File Name	Last Modified	Size
<a href="#">productName.txt.gz</a>	08-Oct-2006 13:14	4.4M
<a href="#">recombRate.sql</a>	13-Apr-2006 12:46	877
<a href="#">recombRate.txt.gz</a>	13-Apr-2006 12:46	88K
<a href="#">refFlat.sql</a>	08-Oct-2006 13:26	1.6K
<a href="#">refFlat.txt.gz</a>	08-Oct-2006 13:26	2.0M
<a href="#">refGene.sql</a>	08-Oct-2006 12:54	1.9K
<a href="#">refGene.txt.gz</a>	08-Oct-2006 12:54	2.2M
<a href="#">refLink.sql</a>	08-Oct-2006 13:26	1.6K
<a href="#">refLink.txt.gz</a>	08-Oct-2006 13:26	4.9M
<a href="#">refSeqAli.sql</a>	08-Oct-2006 12:45	2.1K
<a href="#">refSeqAli.txt.gz</a>	08-Oct-2006 12:45	2.4M
<a href="#">refSeqStatus.sql</a>	08-Oct-2006 13:30	1.2K
<a href="#">refSeqStatus.txt.gz</a>	08-Oct-2006 13:30	681K
<a href="#">refSeqSummary.sql</a>	08-Oct-2006 13:30	1.3K
<a href="#">refSeqSummary.txt.gz</a>	08-Oct-2006 13:30	2.0M
<a href="#">regPotential7X.sql</a>	24-Jun-2006 07:48	1.7K
<a href="#">regPotential7X.txt.gz</a>	24-Jun-2006 07:49	59M
<a href="#">rnBlastTab.sql</a>	22-Jul-2006 11:29	1.6K
<a href="#">rnBlastTab.txt.gz</a>	22-Jul-2006 11:30	586K
<a href="#">scBlastTab.sql</a>	13-Apr-2006 12:46	797
<a href="#">scBlastTab.txt.gz</a>	13-Apr-2006 12:46	350K
<a href="#">seq.sql</a>	13-Apr-2006 10:13	586
<a href="#">seq.txt.gz</a>	13-Apr-2006 10:14	29M
<a href="#">sov.sql</a>	08-Oct-2006 12:45	1.2K

The status bar at the bottom of the browser window shows the URL: `http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/refGene.txt.gz`.



## Using the RefGene locations in R

Load the file.

```
> refgene <- read.table("refGene.txt", header = FALSE,
  sep = "\t", comment.char = "", quote = "")
```

Add the column names, which are not included.

```
> colnames(refgene) <- c("bin", "name", "chrom",
  "strand", "txStart", "txEnd", "cdsStart",
  "cdsEnd", "exonCount", "exonStarts", "exonEnds",
  "id", "name2", "cdsStartStat", "cdsEndStat",
  "exonFrames")
```

We are going to ignore the intron and exon boundaries. We are also going to remove duplicate entries, which seem for some reason to exist; the search to identify these is long.

## More RefGene

```
> temprg <- refgene[, c(1:9, 13:15)]
> omit <- unlist(lapply(levels(temprg$name),
  function(x, n) {
    which(n == x)[1]
  }, as.character(temprg$name)))
> summary(omit)
> refgene <- temprg[omit, ]
> rownames(refgene) <-
  as.character(refgene[, "name"])
```

Finally, we save this as a binary object that we can load later.

```
> save(refgene, file = "refgene.rda")
```

## Linking the Agilent Array to RefGene locations

First, convert the environment in the `AnnBuilder` package for the Agilent 44K arrays to a list.

```
> temp2 <- as.list(Agilent44KREFSEQ)
```

Next, we produce a list that maps the annotations to the spots. This code works because the `Accessions` column of the `featureInfo` object contains RefSeq IDs (primarily), which are the names of the rows in the `temp2` object we just created.

```
> ag.annoList <- temp2[as.character(featureInfo[,  
(filtering for RefSeq here),
```

# Alternative Splicing

```
> ag.annoList[1]
$A_23_P80353
[1] "NM_001003689" "NP_001003689" "NM_031488"
[4] "NP_113676"
```

Notice that some probes are associated with more than one RefSeq gene; this happens because different isoforms (produced by alternative splicing) of the same gene have different RefSeq identifiers. That is, the same piece of DNA can give rise to different mRNA molecules. So, we now search through and select just the first annotation for each spot.

## Grabbing the First

```
> agilent.lc <- unlist(lapply(ag.annoList, length,
> agilentREFSEQ <- unlist(lapply(ag.annoList, func
    if (length(x) == 0) {
      return(NA)
    }
    if (length(x) == 1) {
      return(x)
    }
    idx <- 1
    while (idx <= length(x)) {
      if (x[[idx]] == "") {
        idx <- idx + 1
        next
      }
    }
  })
```

```
        return(x[[idx]])
    }
    return(NA)
}))
> agilentREFSEQ[agilentREFSEQ == ""] <- NA
```

## Checking the Output

```
> length(agilentREFSEQ)
[1] 41675
> sum(!is.na(agilentREFSEQ))
[1] 30612
```

Finally, we use the updated RefSeqs (that we just constructed in the `agilentREFSEQ` object) as indices into the `refgene` chromosome locations above. This computation is also slow, since it uses a search in a list instead of in a hash.

## Checking More Output

```
> agilent2refgene <- refgene[agilentREFSEQ, ]
> agilent2refgene[1:3, ]
```

	bin	name	chrom	strand	txStart	txEnd	cdsStart	cdsEnd	exonCount
NM_001003689	889	NM_001003689	chr22	+	3993125	39957220	39931312	39953547	18
NM_005503	98	NM_005503	chr15	+	2700114	27197806	27133379	27196628	14
NM_004672	795	NM_004672	chr1	-	2755425	27565924	27554468	27565675	29
							cdsStartStat	cdsEndStat	
NM_001003689							compl	compl	
NM_005503							compl	compl	
NM_004672							compl	compl	



## Building a Custom Track

We analyzed the Agilent 44K microarray data using a linear model. The results are contained in an object called `ourResults`:

```
> summary(ourResults)
```

UntreatedMeanLog	Beta	PValue
Min. : 4.870	Min. : -3.15530	Min. : 2.024
1st Qu.: 6.907	1st Qu.: -0.19572	1st Qu.: 8.142
Median : 8.058	Median : -0.05431	Median : 2.749
Mean : 8.742	Mean : -0.04300	Mean : 3.511
3rd Qu.: 9.982	3rd Qu.: 0.10075	3rd Qu.: 5.823
Max. : 16.523	Max. : 3.27672	Max. : 1.000

## Computing a Displayable Score

We are going to use the p-values to decide which genes to display, and we are going to use the coefficient (`Beta`) to compute a score that shows the amount of differential expression. The allowed scores for a custom track range from 0 to 1000. Since the true values of `Beta` range between  $-3$  and  $+3$  (more or less), we are going to multiply by 300 to get a useful score.

```
score <- 300 * ourResults[, "Beta"]
score[score > 1000] <- 1000
score[score < -1000] <- -1000
score <- abs(score)
```

## A Track Data Frame

Now we build a data frame that includes the information we need for a custom track in the desired order:

```
> temp <- data.frame(agilent2refgene[, c("chrom",  
    "txStart", "txEnd", "name2")], score = score,  
    strand = agilent2refgene[, "strand"])
```

```
> temp[1:3, 1:5]
```

	chrom	txStart	txEnd	name2	score
NM_001003689	chr22	39931258	39957220	L3MBTL2	96.90
NM_005503	chr15	27001144	27197806	APBA2	74.41
NM_004672	chr1	27554256	27565924	MAP3K6	2.28

## Significant Overexpressed Genes

We built this data frame for all genes; now we are going to select the ones that are significant (p-value  $< 0.02$ ) and are overexpressed in response to the treatment ( $\beta > 0$ ). We further restrict to those genes that we are able to map.

```
> trackInfo <- temp[!is.na(temp[, "chrom"]) & ourResults[  
  "PValue"] < 0.02 & ourResults[, "Beta"] >  
  0, ]
```

We also have to create a header line that tells the browser to make use of the scores.

```
> trackheader <- paste("track name=upNormal",  
  "description=\"Increased in Normal Cells\"",  
  "useScore=1 color=0,60,120")
```

## Writing the Track Info to a File

We can now write the header line followed by the track data:

```
> write(trackheader, file = "upNormalRNA.tsv",  
        append = FALSE)  
> write.table(trackInfo, file = "upNormalRNA.tsv",  
              append = TRUE, quote = FALSE, sep = "\t",  
              row.names = FALSE, col.names = FALSE)
```

Finally, we do the same thing for the genes that are underexpressed.

```
> trackInfo <- temp[!is.na(temp[, "chrom"]) & ourResults[
  "PValue"] < 0.02 & ourResults[, "Beta"] <
  0, ]
> trackheader <- paste("track name=downNormal",
  "description=\"Decreased in Normal Cells\"",
  "useScore=1 color=100,50,0")
> write(trackheader, file = "dnNormalRNA.tsv",
  append = FALSE)
> write.table(trackInfo, file = "dnNormalRNA.tsv",
  append = TRUE, quote = FALSE, sep = "\t",
  row.names = FALSE, col.names = FALSE)
```

# Viewing Your Custom Track

Now we can return to the genome browser and look at our custom tracks.

**Add Custom Tracks**

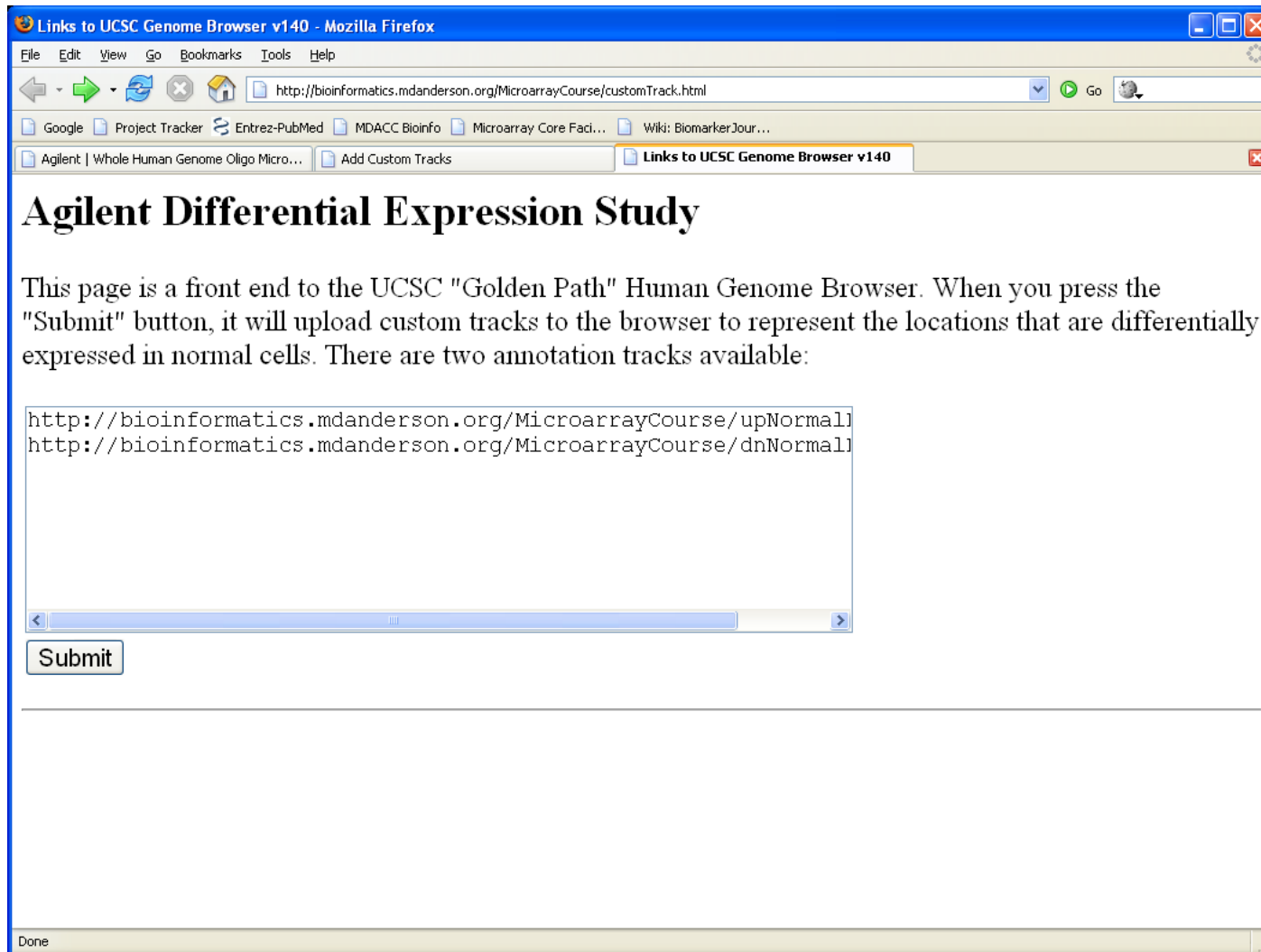
Display your own data as custom annotation tracks in the browser. Data must be formatted in [BED](#), [GFF](#), [GTF](#), [WIG](#) or [PSL](#) formats. To configure the display, set [track](#) and [browser](#) line attributes as described in the [User's Guide](#). Publicly available custom tracks are listed [here](#). Examples are [here](#).

Paste URLs or data:  Or upload:

Optional track documentation:  Or upload:

Done

**[http://bioinformatics.mdanderson.org/  
MicroarrayCourse/customTrack.html](http://bioinformatics.mdanderson.org/MicroarrayCourse/customTrack.html)**





# Displaying Our Tracks

Human chr3:110,765,251-122,424,750 - UCSC Genome Browser v144 - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://genome.ucsc.edu/cgi-bin/hgTracks?position=chr3:110765251-122424750&hgsid=79152112&ct\_downNormal=pac

Google Project Tracker Entrez-PubMed MDACC Bioinfo Microarray Core Faci... Wiki: BiomarkerJour...

Agilent | Whole Human Genome Oligo Micro... Add Custom Tracks Human chr3:110,765,251-122,424,75...

Home Genomes Blat Tables Gene Sorter PCR DNA Convert PDF/PS Help

## UCSC Genome Browser on Human Mar. 2006 Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr3:110,765,251-122,424,750 jump clear size 11,659,500 bp. configure

chr3 (q13.13-q13.33)

chr3: 115000000 120000000

URB  
URB  
ATP6V1A  
PHLDB2 CDC52  
ATG3 QTRTD1  
SLC35A5 ZBTB20  
KIAA1407  
RefSeq Genes  
LOC151760 GCET2 BOC VSI09 GPR43  
PVRL3 GCET2 HDR52 LSAMP  
CD96 ATG3 NAT13  
CD96 CD200R1 DRD3  
ZBED2 BTLA SIDT1  
PLCKD2 GTPBP8 DRD3  
PHLDB2 CDC52  
ABHD10 KIAA2015  
TAGLN3 ATP6V1A  
TAGLN3 GRAMD1C  
TAGLN3 ZDHHC23  
TMFR37 KIAA1407  
C3orf52 QTRTD1  
SLC9A10 ZNF80  
CD200 ZBTB20  
CD200  
SLC35A5  
CCDC80  
CCDC80  
CD200R2  
CD200R1  
CD200R1  
CD200R1  
GTFBP8  
C3orf17  
C3orf17  
C3orf17

Increased in Normal Cells  
Decreased in Normal Cells  
NDUFB4  
RABL3  
GTF2E1  
IGSF11 GPR156  
IGSF11 FSTL1  
C3orf38 NDUFB4  
UFK1B HGD  
B4GALT4 RABL3  
B4GALT4 GTF2E1  
CDGAP  
TMEM39A  
C3orf9  
C3orf9  
C3orf1  
CD80  
ADPRH  
FLA1A  
POFDC2  
COX17  
C3orf15  
NR112  
NR112  
NR112  
GSK3B

Click on a feature for details. Click on base position to zoom in around cursor. Click on left mini-buttons for

move start < 2.0 > move end < 2.0 >

Done

# Searching for a Gene

Human chr3:110,765,251-122,424,750 - UCSC Genome Browser v144 - Mozilla Firefox

http://genome.ucsc.edu/cgi-bin/hgTracks?position=chr3:110765251-122424750&hgsid=79152112&ct\_downNormal=pac

Home Genomes Blat Tables Gene Sorter PCR DNA Convert PDF/PS Help

## UCSC Genome Browser on Human Mar. 2006 Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search TP53 jump clear size 11,659,500 bp. configure

chr3 (q13.13-q13.33)

chr3: 115000000 120000000

Increased in Normal Cells

Decreased in Normal Cells

RefSeq Genes

URB  
URB  
ATP6V1A  
PHLDB2  
ATG3  
SLC35A5  
CADC52  
QTRTD1  
ZBTB20  
KIAA1407  
LOC151760  
PVRL3  
CD96  
ZBED2  
PLCKD2  
PHLDB2  
ABHD10  
TAGLN3  
TAGLN3  
TMPRSS7  
C3orf52  
SLC9A10  
CD200  
CD200  
SLC35A5  
CCDC80  
CCDC80  
CD200R2  
CD200R1  
CD200R1  
CD200R1  
GTFBF8  
C3orf17  
C3orf17  
C3orf17

BOC  
VSIG9  
WDR52  
NAT13  
DRD3  
SIRT1  
DRD3  
KIAA2015  
ATP6V1A  
GRAMD1C  
ZDHHC23  
KIAA1407  
QTRTD1  
ZNF80  
ZBTB20

GRF43  
LSAMP

IGSF11  
IGSF11  
C3orf38  
UFK1B  
B4GALT4  
B4GALT4  
CDGAP  
TMEM39A  
C3orf9  
C3orf9  
C3orf9  
CD80  
ADPRH  
FLA1A  
POFDC2  
COX17  
C3orf15  
NR112  
NR112  
NR112  
GSK3B

GPR156  
FSTL1  
NDUFB4  
HGD  
RABL3  
RABL3  
GTF2E1

Click on a feature for details. Click on base position to zoom in around cursor. Click on left mini-buttons for

move start < 2.0 > move end < 2.0 >

http://genome.ucsc.edu/cgi-bin/hgTracks?position=chr6:138453619-138470280&hgsid=79152112&knownGene=pac&hgFind.matches=NM\_022121,

# Searching for a Gene

The screenshot shows a web browser window titled "Human TP53 - UCSC Genome Browser v144 - Mozilla Firefox". The address bar displays "http://genome.ucsc.edu/cgi-bin/hgTracks". The page content is titled "Known Genes" and lists various genes with their accession numbers and descriptions. The list includes:

- TP53 (NM 000546) at chr17:7512465-7531642 - tumor protein p53
- TP53 (DQ186649) at chr17:7512447-7531524 - Del133 p53 gamma isoform.
- TP53 (DQ186649) at chr17:7512447-7531524 - Del133 p53 gamma isoform.
- TP53 (DQ186648) at chr17:7512447-7531524 - Del133 p53 beta isoform.
- TP53 (DQ186648) at chr17:7512447-7531524 - Del133 p53 beta isoform.
- C20orf10 (NM 014477) at chr20:43435935-43440371 - TP53-target gene 5 protein
- TP53AP1 (NM 007233) at chr7:86792477-86812767 - TP53 activated protein 1
- TP53RK (BC019621) at chr20:44747581-44751486 - TP53 regulating kinase.
- PERP (NM 022121) at chr6:138453619-138470280 - PERP, TP53 apoptosis effector
- RPRM (NM 019845) at chr2:154042098-154043568 - reprimo, TP53 dependant G2 arrest mediator
- RRM2B (NM 015713) at chr8:103285907-103320522 - ribonucleotide reductase M2 B (TP53 inducible
- TP53 (BC003596) at chr17:7512465-7531511 - Del133 p53 isoform.
- VRK1 (BC103761) at chr14:96333459-96417696 - vaccinia related kinase 1
- TP53INP1 (NM 033285) at chr8:96007377-96030767 - tumor protein p53 inducible nuclear protein
- TP53INP1 (AF409114) at chr8:96007377-96030767 - tumor protein p53 inducible nuclear protein 1
- UBE2L6 (BC032491) at chr11:57075712-57091756 - ubiquitin-conjugating enzyme E2L 6
- PPP1R13B (NM 015316) at chr14:103269842-103383555 - protein phosphatase 1, regulatory (inhibi
- TP53BP2 (BC058918) at chr1:222034413-222100255 - tumor protein p53 binding protein, 2
- ENC1 (NM 003633) at chr5:73958991-73972273 - ectodermal-neural cortex (with BTB-like domain)
- PPM1D (BC042418) at chr17:56032412-56096616 - protein phosphatase 1D magnesium-dependent, del
- TP53BP2 (NM 001031685) at chr1:222034413-222100255 - tumor protein p53 binding protein, 2 iso
- GNL3 (NM 206825) at chr3:52694976-52703548 - guanine nucleotide binding protein-like 3
- ING5 (NM 032329) at chr2:242290129-242317563 - inhibitor of growth family, member 5
- PDRG1 (NM 030815) at chr20:29996420-30003544 - p53 and DNA damage-regulated protein
- ING4 (NM 016162) at chr12:6629707-6642565 - inhibitor of growth family, member 4 isoform 1

# Searching for a Gene

The screenshot displays the UCSC Genome Browser interface. The browser title is "Human chr17:7,512,465-7,531,642 - UCSC Genome Browser v144 - Mozilla Firefox". The URL in the address bar is "http://genome.ucsc.edu/cgi-bin/hgTracks?position=chr17:7512465-7531642&hgtsid=79152112&knownGene=pack&hgFind.match...". The page features a navigation menu with options like "Home", "Genomes", "Blat", "Tables", "Gene Sorter", "PCR", "DNA", "Convert", "PDF/PS", and "Help". Below the menu, the title "UCSC Genome Browser on Human Mar. 2006 Assembly" is displayed. Navigation controls include "move" buttons (left arrow, right arrow, double left, double right), "zoom in" buttons (1.5x, 3x, 10x), and "zoom out" buttons (1.5x, 3x, 10x). A "position/search" field contains "chr17:7,512,465-7,531,642" with "jump" and "clear" buttons. A "size 19,178 bp." label and a "configure" button are also present. A chromosome map shows the location of the gene on chromosome 17 (p13.1 to q25.3). The main track displays "UCSC Known Genes Based on UniProt, RefSeq, and GenBank mRNA" for the TP53 gene. Below this, "RefSeq Genes" and "Increased in Normal Cells" / "Decreased in Normal Cells" tracks are visible. At the bottom, there are buttons for "default tracks", "hide all", "manage custom tracks", "configure", and "refresh". Instructions encourage clicking on features for details and using zoom controls. A "Custom Tracks" section is partially visible at the bottom.

# Searching for a Gene

Human chr17:6,563,154-8,480,953 - UCSC Genome Browser v144 - Mozilla Firefox

http://genome.ucsc.edu/cgi-bin/hgTracks?hgid=79152112&hgt.in2=+3x+&position=chr17%3A6563154-8480953

Home Genomes Blat Tables Gene Sorter PCR DNA Convert PDF/PS Help

## UCSC Genome Browser on Human Mar. 2006 Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr17:7,202,421-7,841,687 jump clear size 639,267 bp. configure

chr17 (p13.1) p12 p11.2 11.2 17q12 17q22 q25.3

chr17: 7300000 | 7400000 | 7500000 | 7600000 | 7700000 | 7800000

UCSC Known Genes: TNK1, PLSCR3, C17orf61, NLGN2, TMEM182, FGF11, CHRNB1, TNFSF12-TNFSF13, ZBTB4, POLR2A, TNFSF13, CD68, EIF4A1, TNFSF12, TNFSF13, SAT2, SHBG, ATP1B2, DNHD3, TNFSF13, MDR79, TMEM88, KCNAB3, LSMD1, TRAPPC1, CYB5D1, CNTR0B, CHDS, AK097045

RefSeq Genes: TNK1, PLSCR3, C17orf61, NLGN2, TMEM182, FGF11, CHRNB1, TNFSF12-TNFSF13, ZBTB4, POLR2A, TNFSF13, CD68, EIF4A1, TNFSF12, TNFSF13, SAT2, SHBG, ATP1B2, DNHD3, TNFSF13, MDR79, TMEM88, KCNAB3, LSMD1, TRAPPC1, CYB5D1, CNTR0B, CHDS, AK097045

mRNA: TRAPPC1, KCNAB3, TRAPPC1, CNTR0B, CHDS, CHDS

Increased in Normal Cells: CD68

Decreased in Normal Cells: EIF4A1, TNFSF13

Based on UniProt, RefSeq, and GenBank

Click on a feature for details.

Click on base position to zoom in around cursor. Click

move start move end

Done

# Comments on TCGA

What's there?

from the *Broad*:

ht\_hg\_u133a

from *Harvard*:

hg-cgh-244a

hg-cgh-415k-g4124a

illumina-mrna-dge

from *Johns Hopkins*:

humanmethylation27

## Comments on TCGA (2)

from *Memorial Sloan-Kettering*:

hg\_cgh\_244a

cgh\_1x1m\_g4447a

from *U North Carolina*:

agilent4502a-07-2

agilent4502a-07-3

h\_mirna\_8x15K