Genome builds, gene builds, downloading data from databases

Genome Reference Consortium

- "…working to create assemblies that better represent diversity and provide more robust substrates for genome analysis."
  - novel assembly algorithm
  - correcting assembly errors (fix patches)
  - addition of new alternate loci (patches)
  - filling in gaps

EMBL-EBI

NCBI

wellcome trust sanger institute

THE GENOME INSTITUTE at Washington University

GRCh37 or hg19?

- Ensembl/NCBI versus UCSC
- contig sequences are the same, but different naming convention (i.e. ‘chr1’ versus ‘1’)

What human genome assembly and coordinate system is Ensembl using?

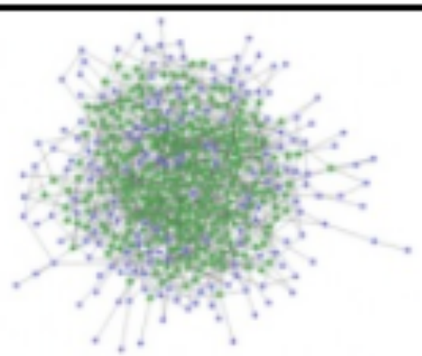
Ensembl uses a one-based coordinate system, whereas UCSC uses a zero-based coordinate system.

Ensembl uses the most recently updated human genome housed at the GRC. This current major assembly release is called GRCh38. NCBI and UCSC use the same genome. UCSC refers to the recent human genome as GRCh38/hg38.

We maintain a long-term archive of the previous assembly of the human genome, GRCh37, with BLAST/BLAT, VEP and BioMart. The data in this archive is based on the Ensembl 75 data.

The Science Web

Putting the "omic" into comical....



Home About

← Your awful, bigoted opinions are encoded in your genes

Human species advised to move to GRCh37

Posted on April 15, 2015 by jovialscientist

BOSTON. The entire human species has been advised to convert their genome to GRCh37 by the GATK Best Practices team at the Broad Institute, *The ScienceWeb* has learned.

GRCh37 is the *previous* version of the human genome reference. Last year, a rogue team of militant terrorist bioinformaticians within the Genome Reference Consortium released GRCh38, a hellish combination of core chromosomes, patches, unplaced contigs and alternate loci. In one fell swoop they broke every single bioinformatics pipeline ever written.

"Enough is enough" said Geraldine Van Damme, former martial arts expert and now head of the GATK team. "We took one look at GRCh38 and though 'that's it, we're sticking to GRCh37 and never moving'. We're therefore recommending that every human on the planet converts their genome to GRCh37. They should use CRISPR or something. It's going to make our lives a lot easier" she finished.

However, not everyone agrees. Deanna Cathedral, formerly Head of Anything Useful at the National Church of Biology Idiots (NCBI) said: "This reminds of the early days of the human genome project, when Frankie Collins suggested we try and genetically modify everyone to be haploid. It's just not realistic" she concluded.

Recent Posts

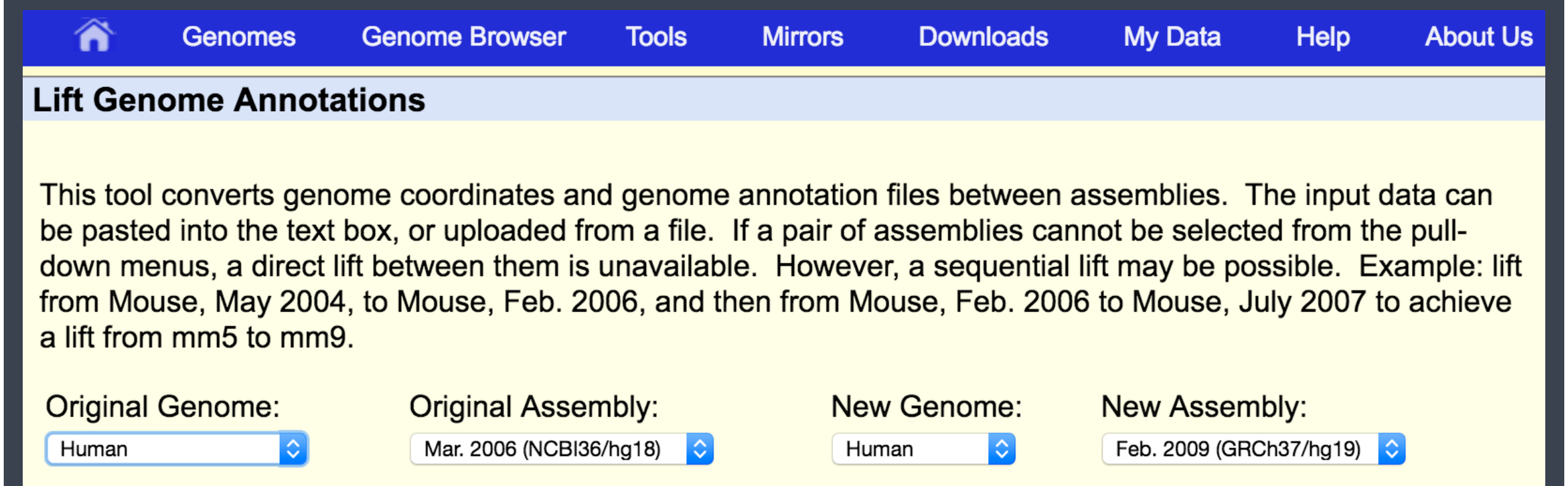
- Human species advised to move to GRCh37
- Your awful, bigoted opinions are encoded in your genes
- Only three gel images ever made, admit scientists
- Bacteria will pay you to sequence them by 2016, analysis reveals
- SGM held at Birmingham to allow scientists to collect filthy new diseases

Meta

- Register
- Log in
- Entries RSS
- Comments RSS
- WordPress.com

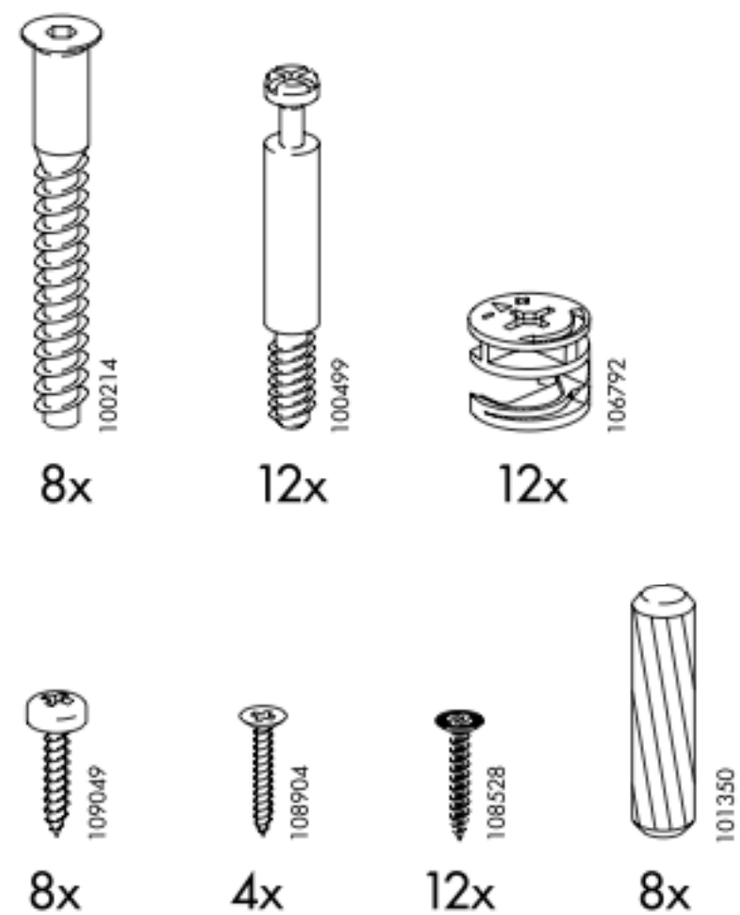
LiftOver at UCSC

You can obtain corresponding coordinates of a different genome build, if you have a set of coordinates from a known build using the **LiftOver tool (UCSC)**

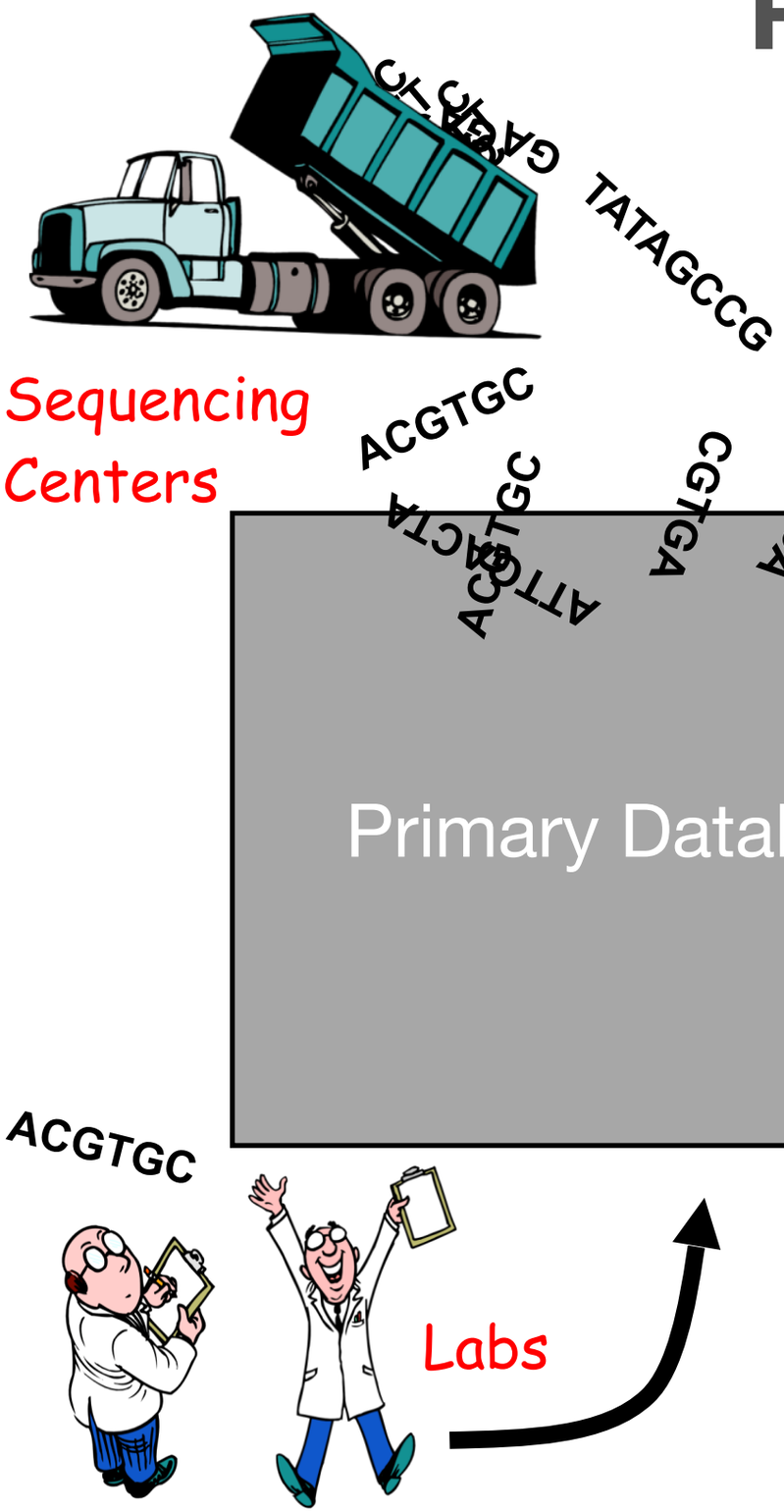


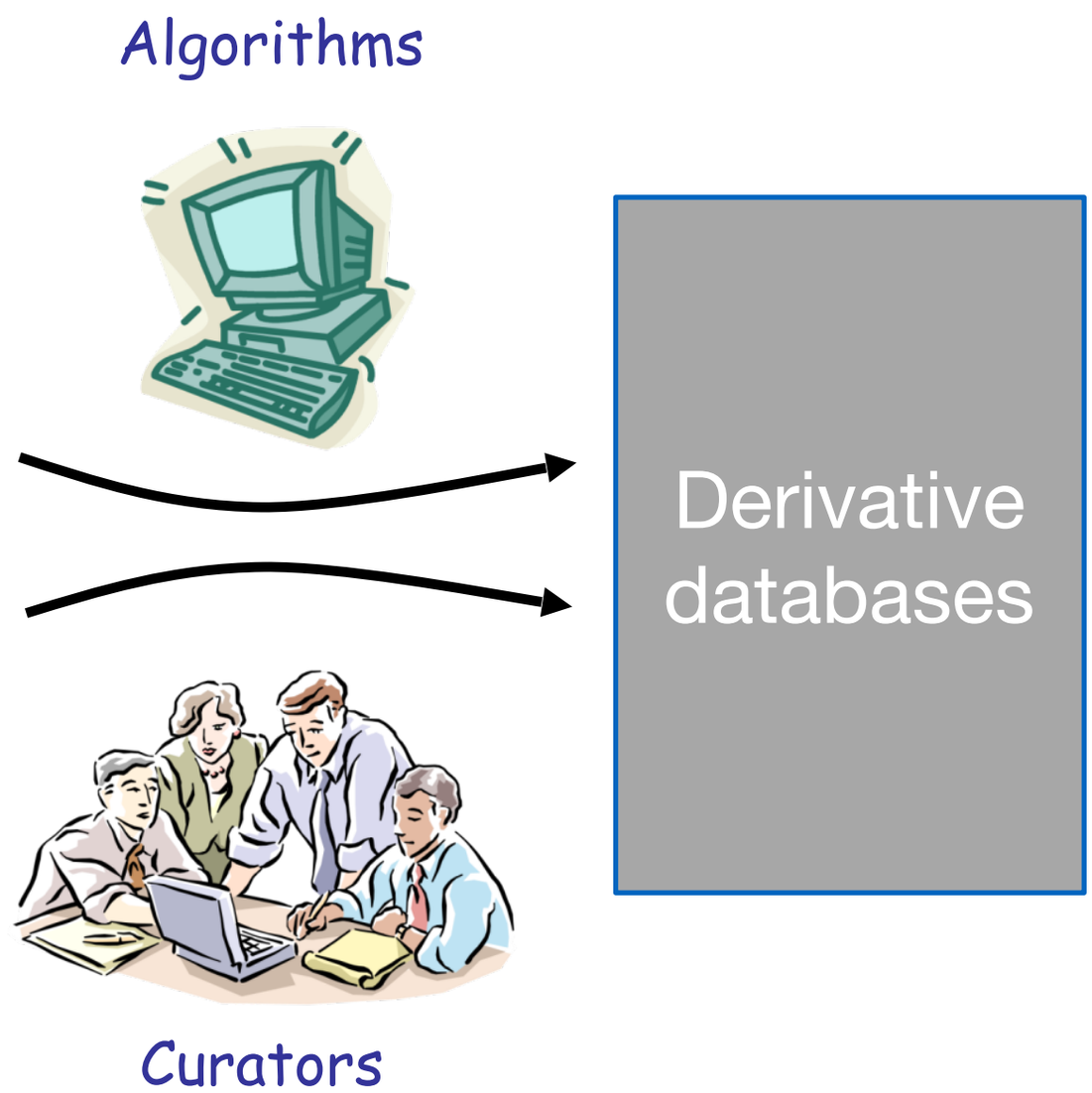
Gene Builds
(not to be confused with *genome* builds)

- A set of annotations for the assembled genome
- Database specific
- Predicted genes based on varying levels of evidence

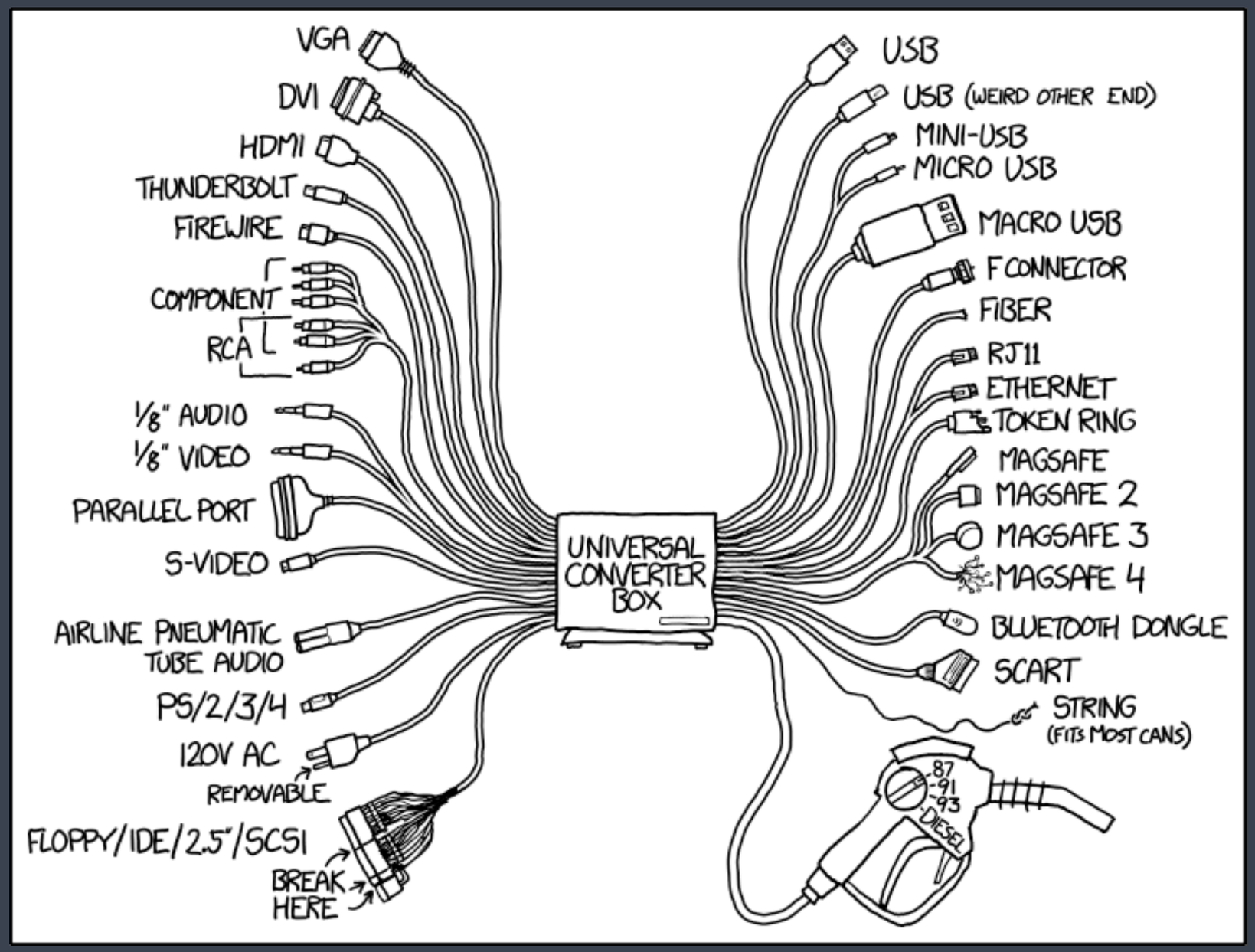


Primary versus Derivative Databases





ID conversions: **biomaRt**



https://xkcd.com/1406/

Genome builds at Illumina: **iGenomes**

http://support.illumina.com/sequencing/sequencing_software/igenome.html

Ready-To-Use Reference Sequences and Annotations

The iGenomes are a collection of reference sequences and annotation files for commonly analyzed organisms. The files have been downloaded from Ensembl, NCBI, or UCSC, and chromosome names have been changed to be simple and consistent with their download source. Each iGenome is available as a compressed file that contains sequences and annotation files for a single genomic build of an organism.

For more information, see the iGenomes Overview and Change Log.

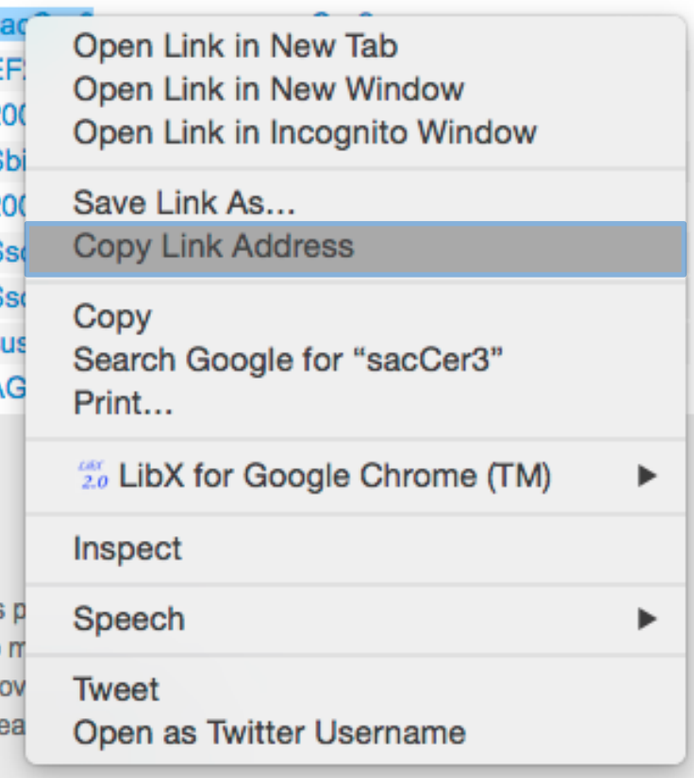
Species	Source	Build(s)			
Arabidopsis thaliana	Ensembl	TAIR10	TAIR9		
	NCBI	TAIR10	build9.1		
Bacillus_cereus strain ATCC 10987	NCBI	2003-02-13			
Bacillus_subtilis strain 168	Ensembl	EB2			
Bos taurus (Cow)	Ensembl	UMD3.1	Btau_4.0		
	NCBI	UMD_3.1.1	UMD_3.1	Btau_4.6.1	Btau_4.2
	UCSC	bosTau8	bosTau7	bosTau6	bosTau4
Caenorhabditis elegans	Ensembl	WBcel235	WBcel215	WS220	WS210
	NCBI	WS195	WS190		
	UCSC	ce10	ce6		

On Orchestra, `/n/groups/shared_databases/igenome/` has some of these.

Genome builds at Illumina: **iGenomes**

http://support.illumina.com/sequencing/sequencing_software/igenome.html

Rhodobacter sphaeroides strain 2.4.1	NCBI	2005-10-07			
Saccharomyces cerevisiae (Yeast)	Ensembl	R64-1-1	EF4	EF3	EF2
	NCBI	build3.1	build2.1		
	UCSC	sac			
Schizosaccharomyces pombe	Ensembl	EF			
Sorangium cellulosum strain So_ce_56	NCBI	20			
Sorghum bicolor	Ensembl	Sbi			
Staphylococcus aureus strain NCTC 8325	NCBI	20			
Sus scrofa (Pig)	Ensembl	Ss			
	NCBI	Ss			
	UCSC	sus			
Zea mays (Corn)	Ensembl	AG			



nologies

l is to apply innovative technologies to the analysis of genetic variation and function, making studies p
t a few years ago. It is mission critical for us to deliver innovative, flexible, and scalable solutions to m
bal company that places high value on collaborative interactions, rapid delivery of solutions, and prov
to meet this challenge. Illumina innovative sequencing and array technologies are fueling groundbrea
h, translational and consumer genomics, and molecular diagnostics.

To download it, copy link address and use with `wget` on Orchestra.

Downloading data from igenomes or NCBI

Start a new interactive session in O2

Use wget to download:

```
$ wget <copied FTP link>
```

If it is tar compressed, uncompress it as follows:

```
$ tar -xf <file.tar.gz>
```

GEO, SRA, FTP downloads at NCBI

- GEO: Gene Expression Omnibus
- SRA: Sequence Read Archive
- Data is available for download on the **NCBI FTP site**
- Related NCBI databases are linked together
  - Select “GEO datasets” from the pull-down menu and search for **Mov10** on the NCBI main page

GEO and SRA at NCBI

Search results

Items: 1 to 20 of 27

<< First < Prev Page 1 of 2 Next > Last >>

1. FMRP-associated **MOV10** facilitates and antagonizes miRNA-mediated regulation

(Submitter supplied) The fragile X mental retardation protein FMRP is an RNA binding protein that regulates translation of its bound mRNAs through incompletely defined mechanisms. FMRP has been linked to the microRNA pathway and we show here that it is associated with **MOV10**, a putative helicase that is also associated with the microRNA pathway. We show that FMRP associates with **MOV10** in an RNA-dependent manner and facilitates **MOV10**-association with RNAs in brain. more...

Organism: Homo sapiens
Type: Expression profiling by high throughput sequencing
Platform: GPL11154 8 Samples
Download data: GEO (TXT), SRA SRP029367
Series Accession: GSE50499 ID: 200050499
PubMed Full text in PMC Similar studies

2. Identification of the cellular RNAs bound by **MOV10**

(Submitter supplied) Using the iCLIP protocol we have identified the cellular RNA entities that are bound by **MOV10**. We report the location and sequence of the **MOV10** binding region on each RNA entity.

Organism: Homo sapiens
Type: Expression profiling by high throughput sequencing
Platform: GPL11154 4 Samples
Download data: GEO (BED), SRA SRP031507
Series Accession: GSE51443 ID: 200051443
PubMed Full text in PMC Similar studies

GEO and SRA at NCBI

NCBI

GEO
Gene Expression Omnibus

HOME | SEARCH | SITE MAP | GEO Publications | FAQ | MIAME | Email GEO

NCBI > GEO > **Accession Display**

Not logged in | Login

Scope: Self Format: HTML Amount: Quick GEO accession: GSE50499 GO

Series GSE50499 Query DataSets for GSE50499

Status	Public on Nov 20, 2014
Title	FMRP-associated MOV10 facilitates and antagonizes miRNA-mediated regulation
Organism	Homo sapiens
Experiment type	Expression profiling by high throughput sequencing
Summary	The fragile X mental retardation protein FMRP is an RNA binding protein that regulates translation of its bound mRNAs through incompletely defined mechanisms. FMRP has been linked to the microRNA pathway and we show here that it is associated with MOV10, a putative helicase that is also associated with the microRNA pathway. We show that FMRP associates with MOV10 in an RNA-dependent manner and facilitates MOV10-association with

GEO and SRA at NCBI

Platforms (1)	GPL11154 Illumina HiSeq 2000 (Homo sapiens)
Samples (8) ± More...	GSM1220262 MOV10 knockdown 2 GSM1220263 MOV10 knockdown 3 GSM1220264 MOV10 overexpression 1

Relations

BioProject	PRJNA217781
SRA	SRP029367

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE50499_GEO_Ceman_counts.txt.gz	320.2 Kb	(ftp)(http)	TXT
SRP/SRP029/SRP029367		(ftp)	SRA Study

Raw data provided as supplementary file

Processed data is available on Series record

GEO and SRA at NCBI

NCBI

GEO
Gene Expression Omnibus

HOME | SEARCH | SITE MAP | GEO Publications | FAQ | MIAME | Email GEO

NCBI > GEO > **Accession Display** Not logged in | Login

GEO help: Mouse over screen elements for information.

Scope: Self Format: HTML Amount: Quick GEO accession: GSM1220263 GO

Sample GSM1220263 Query DataSets for GSM1220263

Status	Public on Nov 20, 2014
Title	MOV10 knockdown 3
Sample type	SRA
Source name	Human Embryonic Kidney cell lines
Organism	Homo sapiens
Characteristics	cell type: Human Embryonic Kidney cells cell line: HEK293F treatment: MOV10 knockdown mov expression: low
Treatment protocol	MOV10 and irrelevant siRNA treatments were performed at 24 hr intervals three times, overexpression studies involved a single myc-MOV10 transfection
Growth protocol	Cells were grown in serum containing DMEM at 37C
Extracted molecule	total RNA

GEO and SRA at NCBI

Platform ID	GPL11154	
Series (1)	GSE50499	FMRP-associated MOV10 facilitates and antagonizes miRNA-mediated regulation

Relations

BioSample	SAMN02340011
SRA	SRX342247

Supplementary file	Size	Download	File type/resource
SRX/SRX342/SRX342247		(ftp)	SRA Experiment

Raw data provided as supplementary file

Processed data is available on Series record

GEO and SRA at NCBI

Full ▾ Send to: ▾

SRX342247: GSM1220262: MOV10 knockdown 2; Homo sapiens; RNA-Seq
2 ILLUMINA (Illumina HiSeq 2000) runs: 52.7M spots, 5.3G bases, 3.6Gb downloads

Submitted by: Gene Expression Omnibus (GEO)

Study: FMRP-associated MOV10 facilitates and antagonizes miRNA-mediated regulation
PRJNA217781 • SRP029367 • All experiments • All runs
show Abstract

Sample: MOV10 knockdown 2
SAMN02340011 • SRS475153 • All experiments • All runs
Organism: Homo sapiens

Library:
Instrument: Illumina HiSeq 2000
Strategy: RNA-Seq
Source: TRANSCRIPTOMIC
Selection: cDNA
Layout: SINGLE
Construction protocol: Cells were lysed and RNA was extracted using Trizol Illumina's TruSeq Stranded RNAseq Sample Prep kit was used with 1 ug of total RNA for the construction of sequencing libraries. Indices (barcodes) were included to be able to differentiate the sequences from each sample. The adapter sequence used was AGATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNATCTCGTATGCCGTCTTCTGCTTG (NNNNNN = 6 nt barcode index in .fastq file name)

Experiment attributes:
GEO Accession: GSM1220262

Links:
External link: GEO Sample GSM1220262

Runs: 2 runs, 52.7M spots, 5.3G bases, 3.6Gb

Run	# of Spots	# of Bases	Size	Published
SRR960455	27,426,242	2.7G	1.9Gb	2014-11-20
SRR960456	25,254,091	2.5G	1.7Gb	2014-11-20

Downloading data from SRA

Start a new interactive session in O2

Load the sratoolkit module

```
$ module load sratoolkit/2.8.1
```

#Download the dataset of interest

```
$ prefetch -v SRR390728
```

convert the .sra file to fastq format

```
$ fastq-dump -h
$ fastq-dump <options> <SRR390728.sra>
```

https://www.ncbi.nlm.nih.gov/books/NBK242621/#SRA_Download_Guid_BK.Download_with_Prefe