Sequence Tracking

Understanding your sequence context

Deanna M. Church
Senior Director of Genomics and Content
Personalis, Inc

Personalis.

@deannachurch

Short Course in Medical Genetics 2014

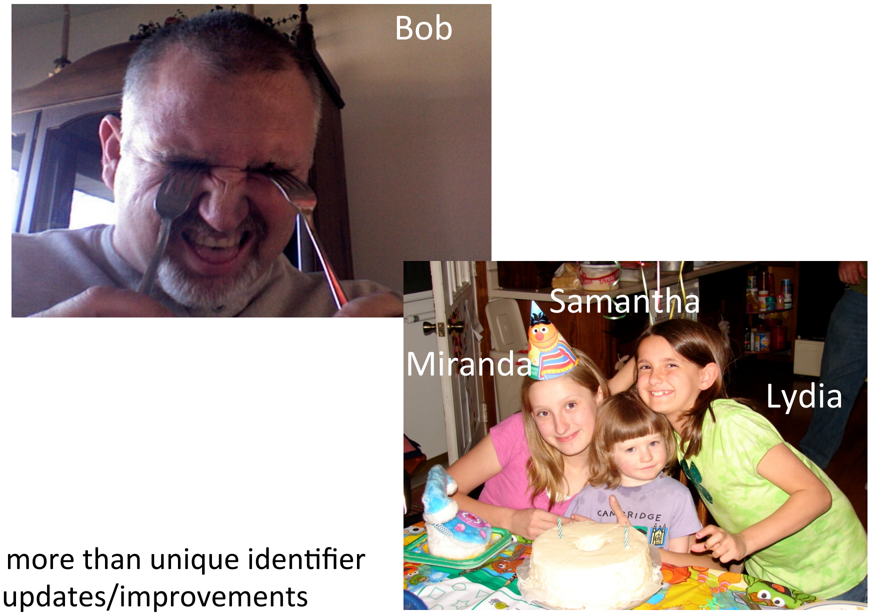
What's in a name?

What's in a name?

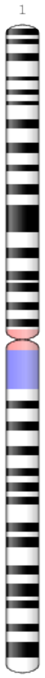
123-45-6789

*http://howmanyofme.com

What's in a name?

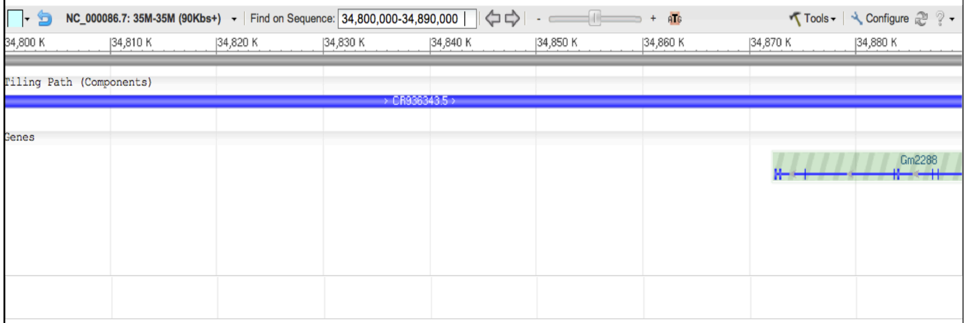


Need more than unique identifier
track updates/improvements



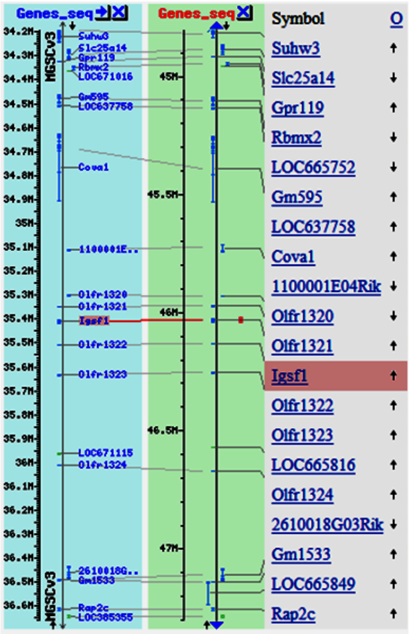
chr1
Chr1
1
Chrom1

Mouse chrX: 34,800,000-34,890,000

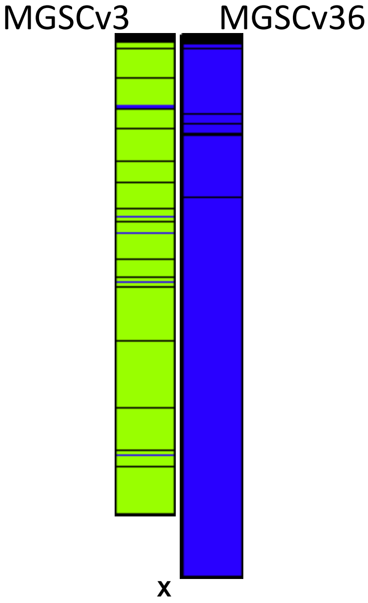

NC_000086.7: 35M-35M (90Kbs+)

Find on Sequence: 34,800,000-34,890,000

Mouse chrX: 35,000,000-36,000000



Symbol	O
Suhw3	↑
Slc25a14	↓
Gpr119	↑
Rbmx2	↓
LOC665752	↓
Gm595	↑
LOC637758	↑
Coval	↑
1100001E04Rik	↓
Olfr1320	↓
Olfr1321	↑
Igsf1	↑
Olfr1322	↑
Olfr1323	↑
LOC665816	↑
Olfr1324	↑
2610018G03Rik	↓
Gm1533	↑
LOC665849	↑
Rap2c	↑



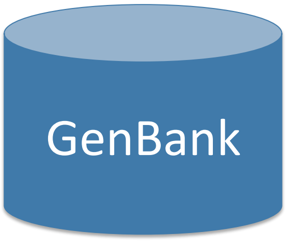
Data Archives

INSDC

NCBI

DDBJ

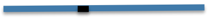
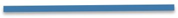
ENA
European Nucleotide Archive



- Data in a common format
- Data in a single location (and mirrored)
- Most quality checked prior to deposition
- Robust data tracking mechanism (accession.version)
- Data owned by submitter

Data tracking

ABC14-1065514J1

	Date	Phase	Gaps	Length
FP565796.1	21-Oct-2009	1	1	
FP565796.2	14-Oct-2010	1	0	
FP565796.3	07-Nov-2010	3	0	

Data Archives

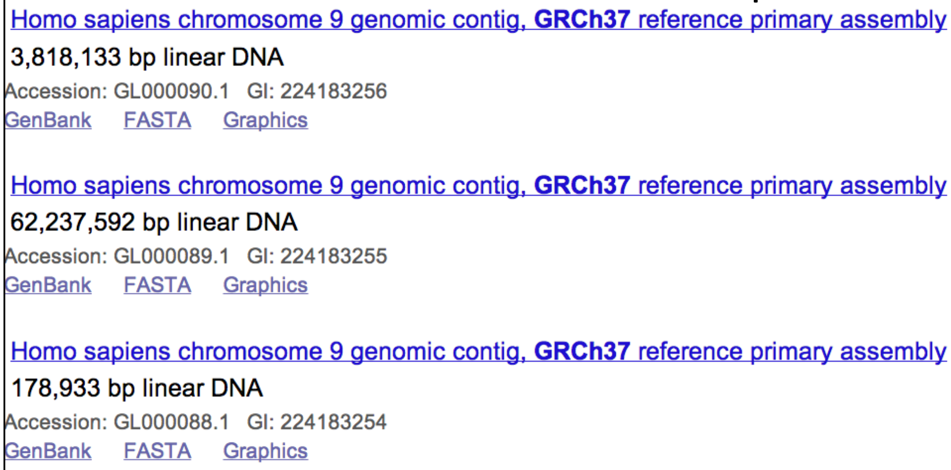
Initial versions of human and mouse reference assemblies not in INSDC!!*

First human version in INSDC: GRCh37
First mouse version in INSDC: NCBI36

* But were tracked by RefSeq

Data Archives

INSDC archives track INDIVIDUAL sequences



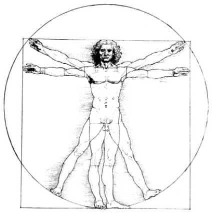
Homo sapiens chromosome 9 genomic contig, **GRCh37** reference primary assembly
3,818,133 bp linear DNA
Accession: GL000090.1 GI: 224183256
GenBank FASTA Graphics

Homo sapiens chromosome 9 genomic contig, **GRCh37** reference primary assembly
62,237,592 bp linear DNA
Accession: GL000089.1 GI: 224183255
GenBank FASTA Graphics

Homo sapiens chromosome 9 genomic contig, **GRCh37** reference primary assembly
178,933 bp linear DNA
Accession: GL000088.1 GI: 224183254
GenBank FASTA Graphics

An assembly is a COLLECTION of sequences

More naming issues



GRCh38

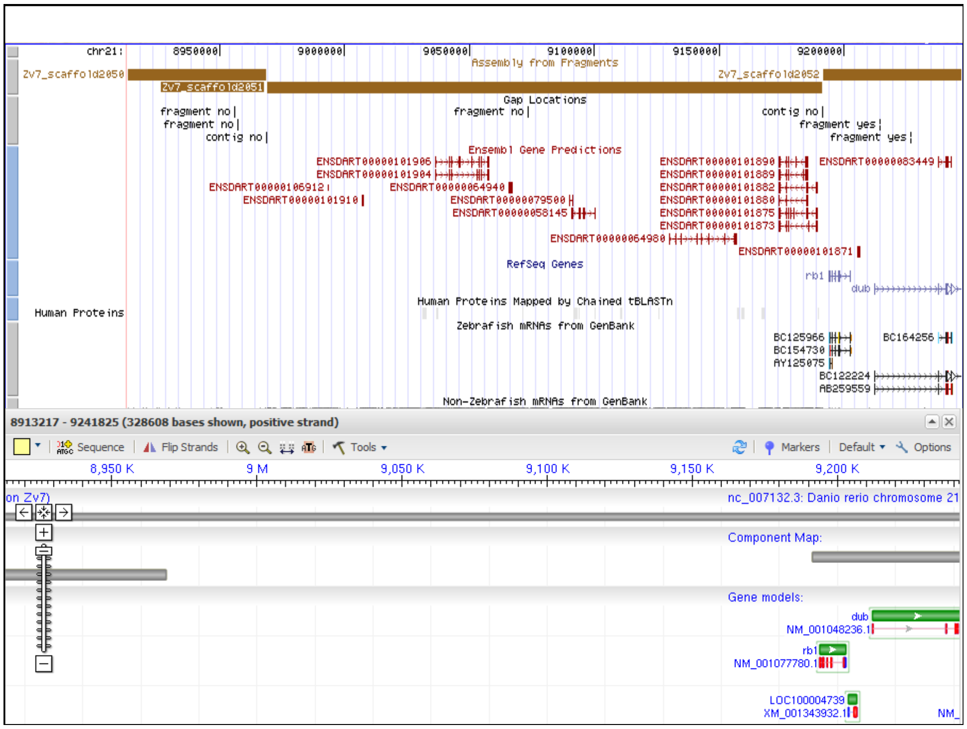
hg18

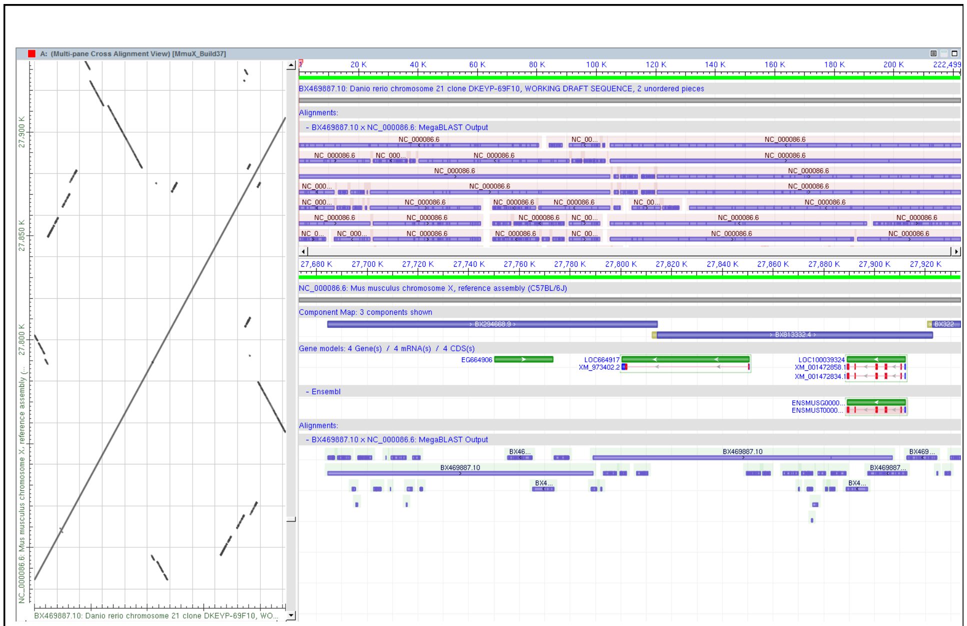
Zv9

danRer7

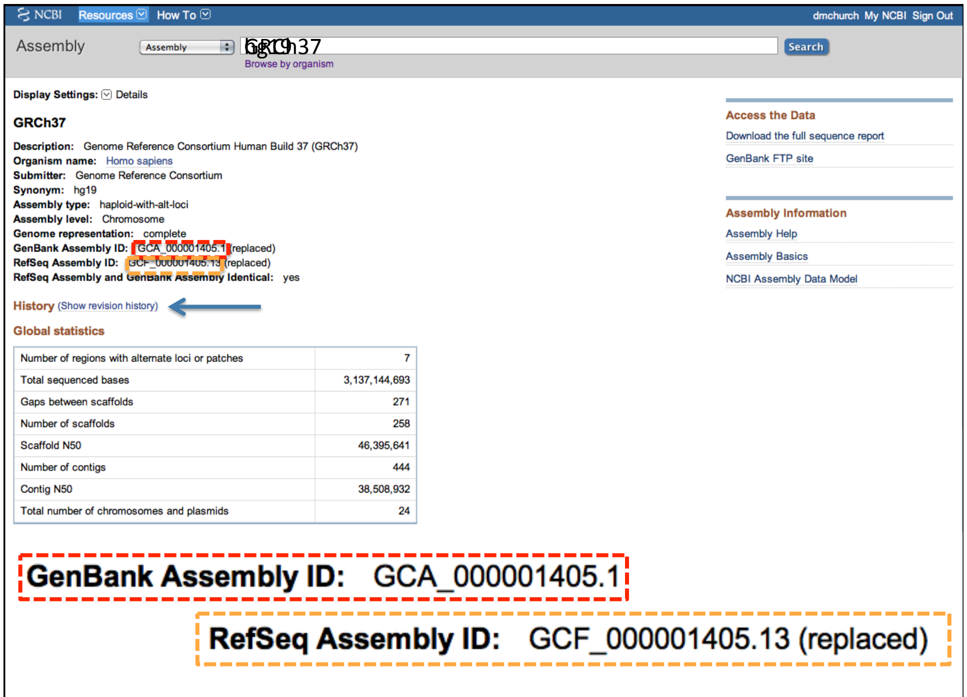
GRCm38

mm10




A: (Multi-pane Cross Alignment View) [MmuX_Build37]
27,900 K
27,850 K
27,800 K
NC_000086.6: Mus musculus chromosome X, reference assembly (...
BX469887.10: Danio rerio chromosome 21 clone DKEYP-69F10, WO...
20 K
40 K
60 K
80 K
100 K
120 K
140 K
160 K
180 K
200 K
222,499
BX469887.10: Danio rerio chromosome 21 clone DKEYP-69F10, WORKING DRAFT SEQUENCE, 2 unordered pieces
Alignments:
- BX469887.10 x NC_000086.6: MegaBLAST Output
NC_000086.6
27,680 K
27,700 K
27,720 K
27,740 K
27,760 K
27,780 K
27,800 K
27,820 K
27,840 K
27,860 K
27,880 K
27,900 K
27,920 K
NC_000086.6: Mus musculus chromosome X, reference assembly (C57BL/6J)
Component Map: 3 components shown
BX294668.9
BX813332.4
BX322...
Gene models: 4 Gene(s) / 4 mRNA(s) / 4 CDS(s)
EG664906
LOC664917
XM_973402.2
LOC100039324
XM_001472858.1
XM_001472834.1
- Ensembl
ENSMUSG0000...
ENSMUST0000...
Alignments:
- BX469887.10 x NC_000086.6: MegaBLAST Output
BX469887.10
BX46...
BX4...
BX469...
BX469887...

Zv7 chr21:8,913,216-9,246,964 vs MGSCv36 chrX



Display Settings: Details

GRCh37

Description: Genome Reference Consortium Human Build 37 (GRCh37)
Organism name: Homo sapiens
Submitter: Genome Reference Consortium
Synonym: hg19
Assembly type: haploid-with-alt-loci
Assembly level: Chromosome
Genome representation: complete
GenBank Assembly ID: GCA_000001405.1 (replaced)
RefSeq Assembly ID: GCF_000001405.13 (replaced)
RefSeq Assembly and GenBank Assembly Identical: yes

History (Show revision history)

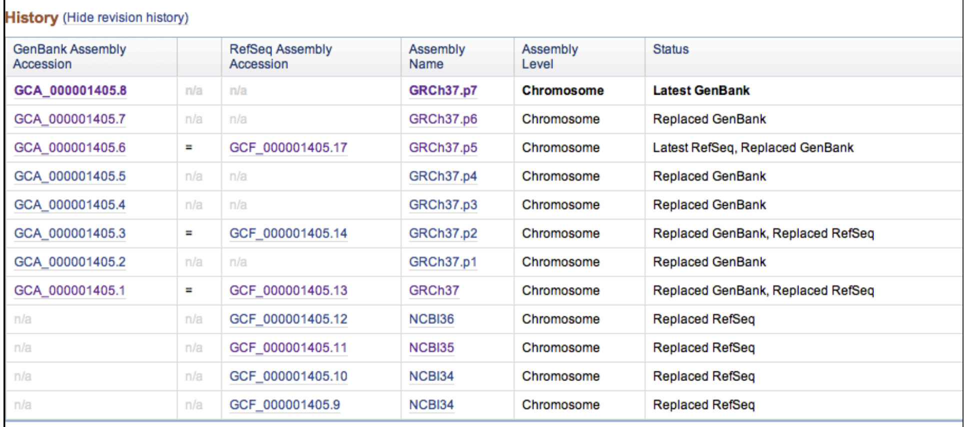
Global statistics

Number of regions with alternate loci or patches	7
Total sequenced bases	3,137,144,693
Gaps between scaffolds	271
Number of scaffolds	258
Scaffold N50	46,395,641
Number of contigs	444
Contig N50	38,508,932
Total number of chromosomes and plasmids	24

GenBank Assembly ID: GCA_000001405.1

RefSeq Assembly ID: GCF_000001405.13 (replaced)

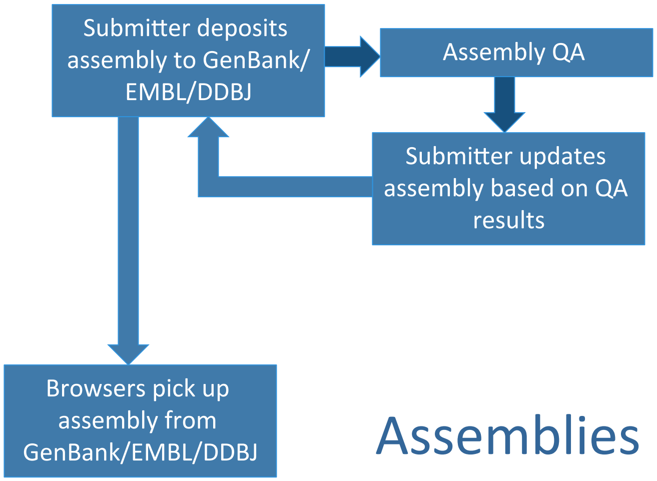
http://www.ncbi.nlm.nih.gov/genome/assembly



History (Hide revision history)

GenBank Assembly Accession		RefSeq Assembly Accession	Assembly Name	Assembly Level	Status
GCA_000001405.8	n/a	n/a	**GRCh37.p7**	**Chromosome**	**Latest GenBank**
GCA_000001405.7	n/a	n/a	GRCh37.p6	Chromosome	Replaced GenBank
GCA_000001405.6	=	GCF_000001405.17	GRCh37.p5	Chromosome	Latest RefSeq, Replaced GenBank
GCA_000001405.5	n/a	n/a	GRCh37.p4	Chromosome	Replaced GenBank
GCA_000001405.4	n/a	n/a	GRCh37.p3	Chromosome	Replaced GenBank
GCA_000001405.3	=	GCF_000001405.14	GRCh37.p2	Chromosome	Replaced GenBank, Replaced RefSeq
GCA_000001405.2	n/a	n/a	GRCh37.p1	Chromosome	Replaced GenBank
GCA_000001405.1	=	GCF_000001405.13	GRCh37	Chromosome	Replaced GenBank, Replaced RefSeq
n/a	n/a	GCF_000001405.12	NCBI36	Chromosome	Replaced RefSeq
n/a	n/a	GCF_000001405.11	NCBI35	Chromosome	Replaced RefSeq
n/a	n/a	GCF_000001405.10	NCBI34	Chromosome	Replaced RefSeq
n/a	n/a	GCF_000001405.9	NCBI34	Chromosome	Replaced RefSeq

Genome Browser Agreement


Submitter deposits assembly to GenBank/EMBL/DDBJ

Assembly QA

Submitter updates assembly based on QA results

Browsers pick up assembly from GenBank/EMBL/DDBJ

Assemblies must be in GenBank/EMBL/DDBJ

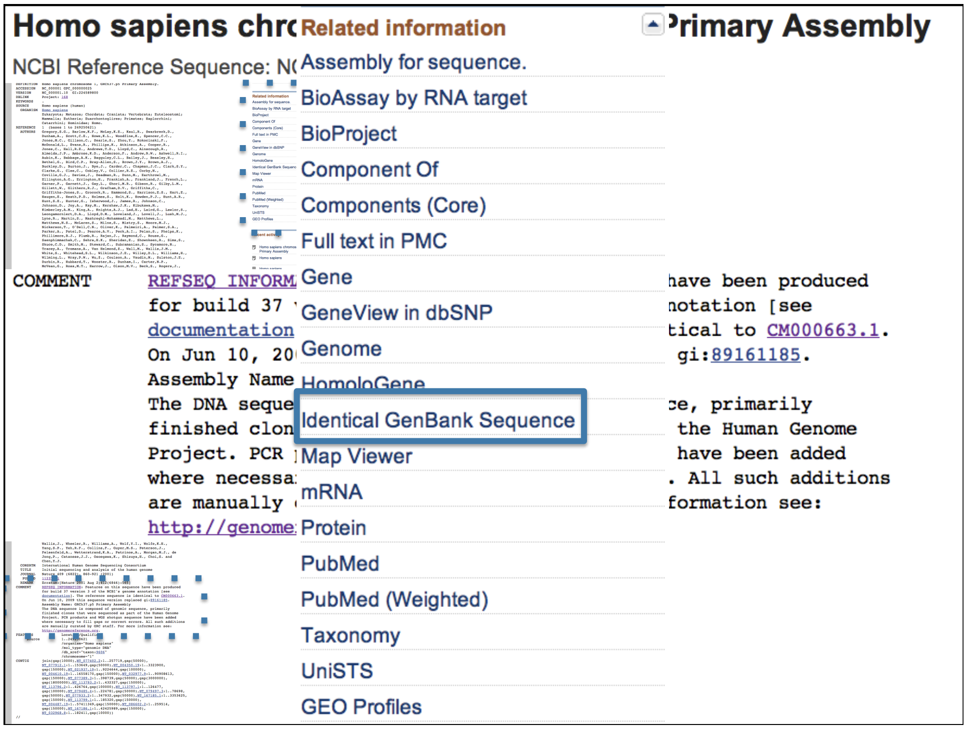
Ensembl NCBI UCSC Genome Bioinformatics

GenBank vs RefSeq

GenBank	RefSeq
Submitter Owned	RefSeq Owned
Redundancy	Non-Redundant
Updated rarely	Curated
INSDC	Not INSDC

BRCA1

GenBank	RefSeq
83 genomic records	3 genomic records
31 mRNA records	5 mRNA records
27 protein records	1 RNA record
	5 protein records



Homo sapiens chrc Primary Assembly

NCBI Reference Sequence: N(

COMMENT REFSEQ INFORM have been produced
for build 37 notation [see
documentation tical to CM000663.1.
On Jun 10, 20 gi:89161185.
Assembly Name
The DNA seque ce, primarily
finished clon the Human Genome
Project. PCR have been added
where necessa . All such additions
are manually formation see:
http://genome

RefSeq for Assemblies

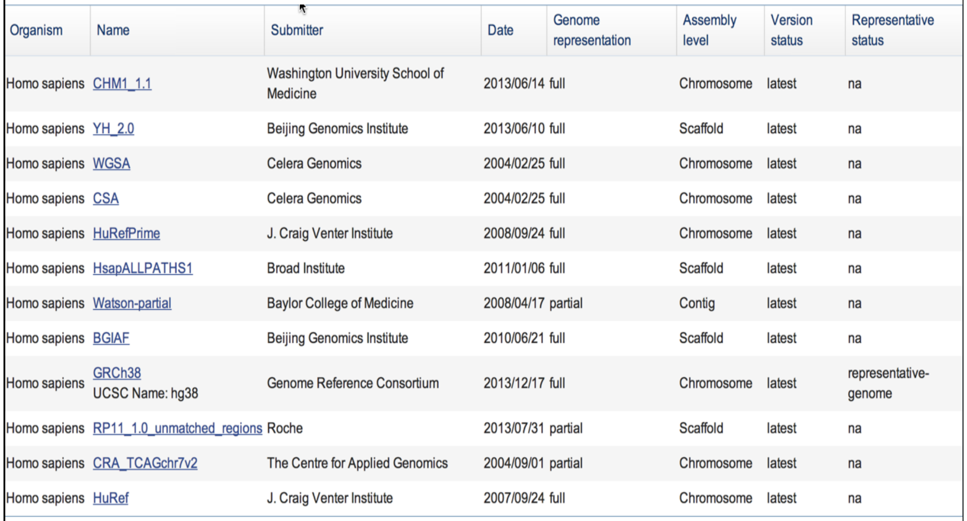
Typical assembly edits

Addition of non-nuclear (e.g. MT) assembly units

Removal of contamination

Drop unlocalized/unplaced scaffolds
Mask contamination that is placed on chromosome (while preserving coordinate space)

Human assemblies in assembly database



Organism	Name	Submitter	Date	Genome representation	Assembly level	Version status	Representative status
Homo sapiens	CHM1_1.1	Washington University School of Medicine	2013/06/14	full	Chromosome	latest	na
Homo sapiens	YH_2.0	Beijing Genomics Institute	2013/06/10	full	Scaffold	latest	na
Homo sapiens	WGSA	Celera Genomics	2004/02/25	full	Chromosome	latest	na
Homo sapiens	CSA	Celera Genomics	2004/02/25	full	Chromosome	latest	na
Homo sapiens	HuRefPrime	J. Craig Venter Institute	2008/09/24	full	Chromosome	latest	na
Homo sapiens	HsapALLPATHS1	Broad Institute	2011/01/06	full	Scaffold	latest	na
Homo sapiens	Watson-partial	Baylor College of Medicine	2008/04/17	partial	Contig	latest	na
Homo sapiens	BGIAF	Beijing Genomics Institute	2010/06/21	full	Scaffold	latest	na
Homo sapiens	GRCh38 UCSC Name: hg38	Genome Reference Consortium	2013/12/17	full	Chromosome	latest	representative-genome
Homo sapiens	RP11_1.0_unmatched_regions	Roche	2013/07/31	partial	Scaffold	latest	na
Homo sapiens	CRA_TCAGchr7v2	The Centre for Applied Genomics	2004/09/01	partial	Chromosome	latest	na
Homo sapiens	HuRef	J. Craig Venter Institute	2007/09/24	full	Chromosome	latest	na

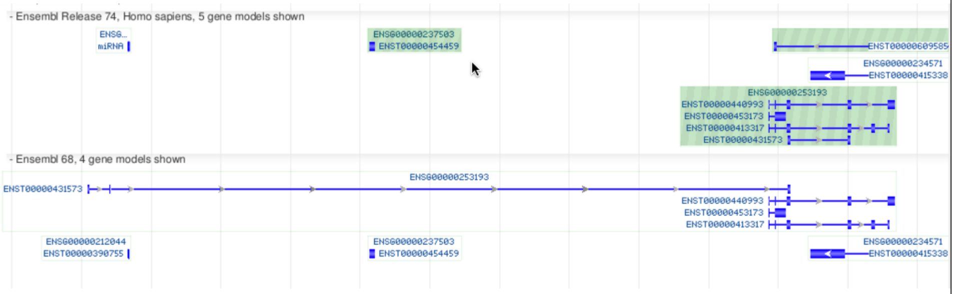
http://www.ncbi.nlm.nih.gov/assembly/organism/9606/

Annotation should have versions too!

- NCBI
  - NCBI Homo sapiens annotation 105
  - NCBI Homo sapiens annotation 106
- Ensembl
  - Ensembl 74
  - Ensembl 75

Annotation should have versions too!

Ensembl 74



Ensembl 68

Take home messages

- Assemblies can (and do) update!
- Know what assembly your are working on
  - Track by accession.version, not just name
- Data in INSDC databases are mirrored
- RefSeq is NCBI specific
- Track annotation too!