


Sequence Tracking

Understanding your sequence context

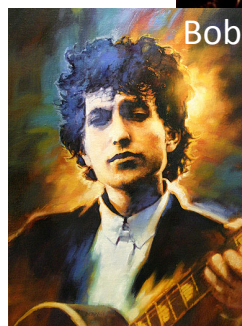
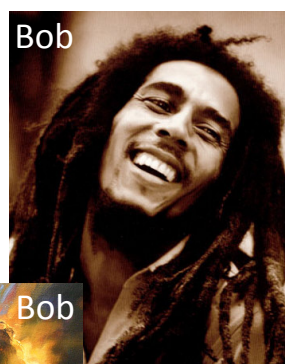
Deanna M. Church
Senior Director of Genomics and Content
Personalis, Inc



 @deannachurch

Short Course in Medical Genetics 2014

What's in a name?



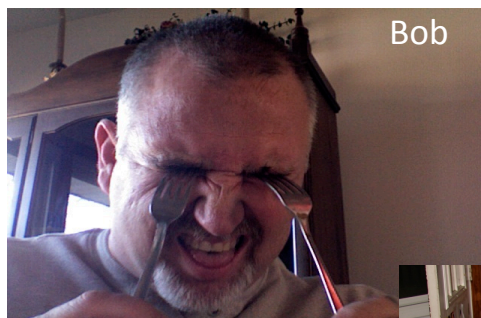
What's in a name?



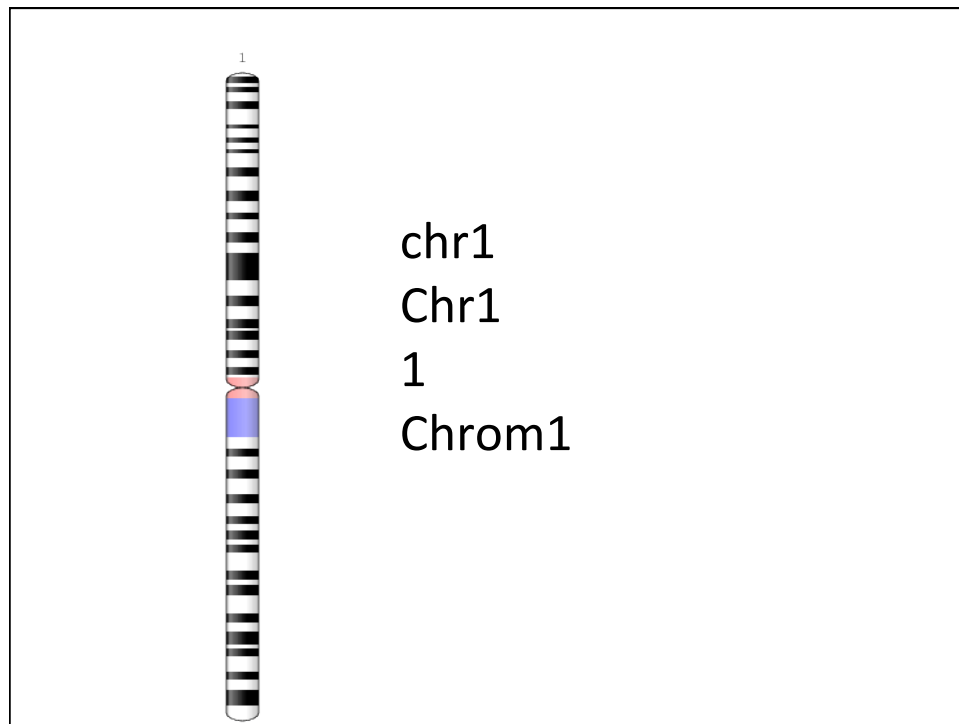
123-45-6789

*<http://howmanyofme.com>

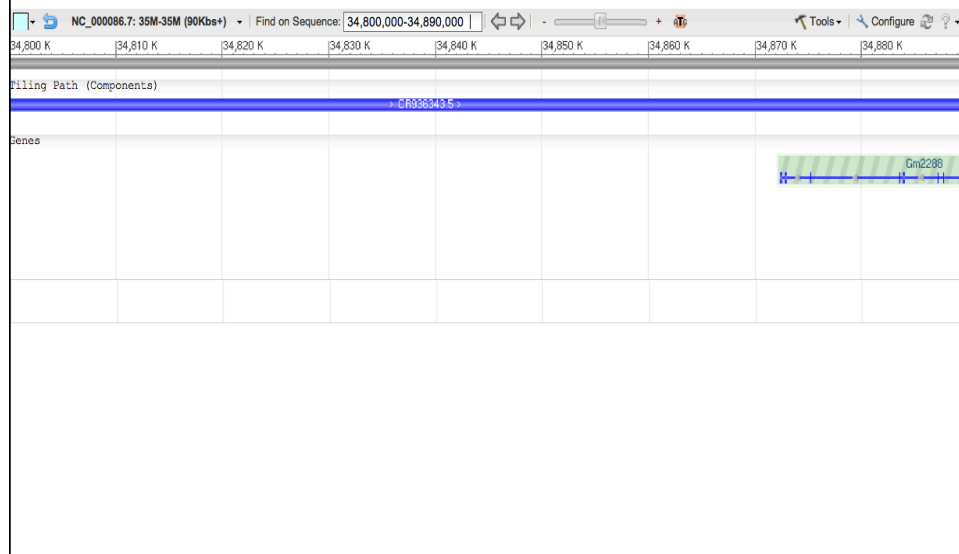
What's in a name?



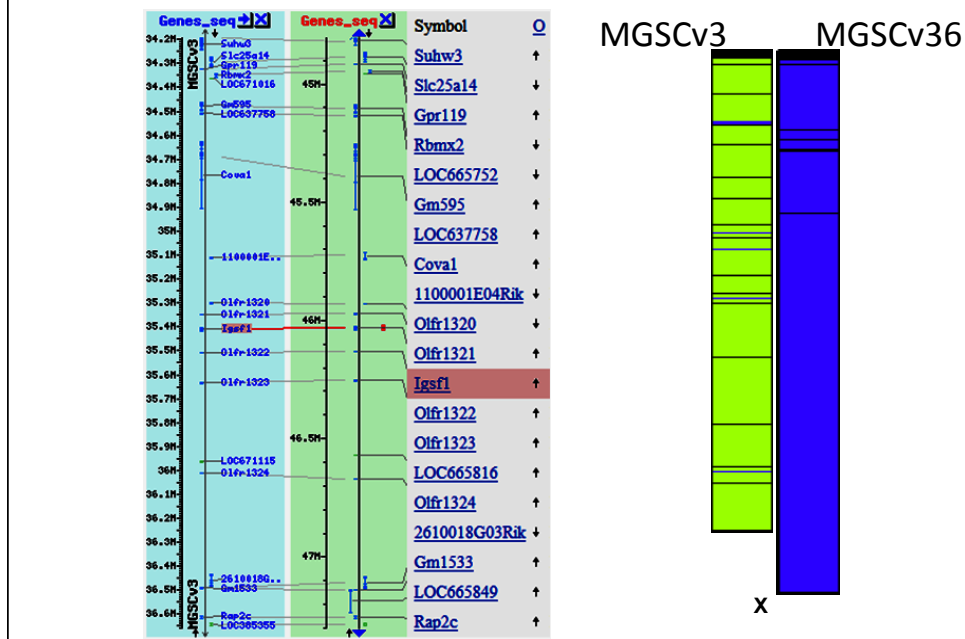
Need more than unique identifier
track updates/improvements



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



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



Data Archives



- 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
<u>FP565796.1</u>	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](#)

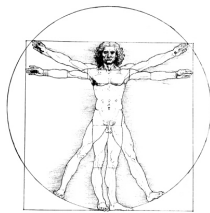
178,933 bp linear DNA

Accession: GL000088.1 GI: 224183254

[GenBank](#) [FASTA](#) [Graphics](#)

An assembly is a **COLLECTION** of sequences

More naming issues



GRCh37

hg19



Zv7

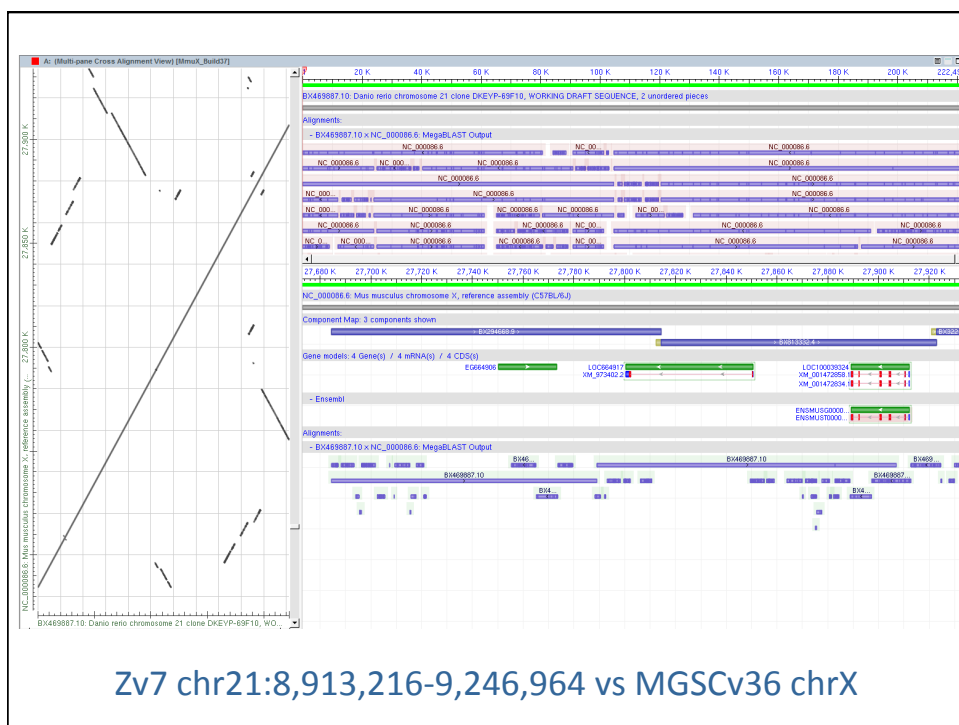
danRer5



MGSCv37

mm8

NCBIM37



NCBI

Resources

How To

dmchurch My NCBI Sign Out

Assembly

Assembly

GRCh37

Search

Browse by organism

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

Access the Data

[Download the full sequence report](#)

[GenBank FTP site](#)

Assembly Information

[Assembly Help](#)

[Assembly Basics](#)

[NCBI Assembly Data Model](#)

GenBank Assembly ID: [GCA_000001405.1](#)

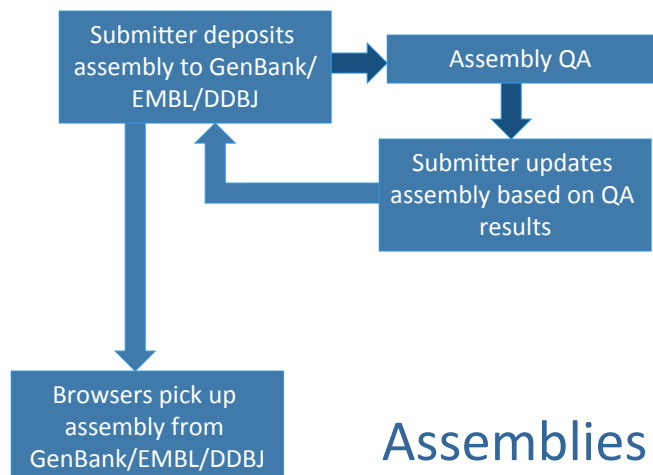
RefSeq Assembly ID: [GCF_000001405.13](#) (replaced)

<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



Assemblies must be in
GenBank/EMBL/DDBJ



GenBank

vs

RefSeq

Submitter Owned

RefSeq Owned

Redundancy

Non-Redundant

Updated rarely

Curated

INSDC

Not INSDC

BRCA1

83 genomic records

3 genomic records

31 mRNA records

5 mRNA records

27 protein records

1 RNA record

5 protein records

Homo sapiens chr **Related information** **Primary Assembly**

NCBI Reference Sequence: NC_010192.3

COMMENT **REFSEQ INFORMATION** **Gene** have been produced

for build 37 **GeneView in dbSNP** notation [see

[documentation](#) **Genome** tical to [CM000663.1](#).

On Jun 10, 20 **Identical GenBank Sequence** gi:[89161185](#).

Assembly Name **HomoloGene**

The DNA sequence finished clone **Map Viewer**

Project. PCR where necessa **mRNA**

are manually <http://genome> **Protein**

PubMed

PubMed (Weighted)

Taxonomy

UniSTS

GEO Profiles

BioAssay by RNA target

BioProject

Component Of

Components (Core)

Full text in PMC

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/>

Take home messages

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