NCBI News, April 2012

Peter Cooper, Ph.D.¹ and Rana Morris, Ph.D.² Created: March 30, 2012; Updated: March 30, 2012.

NCBI Discovery Workshops May 15-16 at NLM: Seats still available

NCBI will present a two-day workshop May 15 and 16, 2012, on the NIH campus in Bethesda, Maryland. The course is free and is open to anyone interested in NCBI resources. The four workshops are Sequences, Genomes, and Maps; Proteins, Domains and Structures; NCBI BLAST Services; and Human Variation and Disease Genes. These workshops provide hands-on experience exploring practical examples using tools and databases on the NCBI website. The Discovery Workshops page has more details and a link to register for the course.

Assembly: a Companion to the Genome Database

The new NCBI Assembly database provides statistics, update history and links to sequences for eukaryotic genome assemblies including assemblies for previous genome builds. Assemblies of interest can be found either by text searches on the main assembly page or through the assembly browser that provides easy access by organism. Assemblies are also linked through the Genome database main page or from a Genome record for a eukaryotic species as shown in Figure 1. Each assembly is assigned an accession and a version that unambiguously identifies the sequences in a particular version of an assembly. The database contains the placement of each scaffold in the assembly along with the name and sequence accession and version for each chromosome and scaffold. The database also organizes and provides assembly descriptive items such as assembly names and synonyms, as well as statistical reports including scaffold counts and weighted scaffold and contig length medians (N50). Figure 2 shows the page for the latest mouse genome assembly (GRCm38). This page provides access to the primary assembly and alternate loci sequences and statistics. The Assembly Help documentation provides more detailed information on using the Assembly Database.

¹ NCBI; Email: cooper@ncbi.nlm.nih.gov. ² NCBI; Email: morrisrc@ncbi.nlm.nih.gov.

NLM Citation: Cooper P, Morris R. NCBI News, April 2012. 2012 Mar 30 [Updated 2012 Mar 30]. In: NCBI News [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 1991-2012.

sembly		Assembl	ly 🔽	Browse by orga	anism			8	Search	
мттттс аттттс атт	TTCCC TTCCC C	болаасс алаасст элаасст	 	Assem	bly					
			rrrec rrrec	Genome ass	embly org	anizatior	n and addit	ional informa	ation.	
Jsing A	ssemb	bly		Submitting	g an Ass	sembly	y Re	lated Res	source	s
Assembly He	elp			Submission Infor	mation		Gen	ome		0.0.0.2
Browse by O	rganism			Submission FAQ			Gen	ome Reference	e Consort	ium
		with the title	a alau	and in mouthing hu		and comp	and a gonon	- analysis. 1110		Survice Sur
of ra	mice, provi diation- or c rganism	ding a wealth o chemically-indu Overview	of different ced, and See also Chromos	t genotypes and phe transgenic mutants p c: Genome list Organ omes	enotypes for g provide poten nelle List	enetic and itial models	i other studies s <u>More</u> As	. In addition, the sembly and Ar	ousands of	spontaneous,
of ra	mice, provi diation- or c rganism	ding a wealth o chemically-indu Overview Click on cheor 4 5 6 7	See also chromos	t genotypes and phe transgenic mutants p : <u>Genome list</u> <u>Organ</u> omes c to open MapViewer 0 11 12 13 14 15	notypes for g provide poten nelle List	enetic and	t other studies s <u>More</u> As	. In addition, the sembly and An Default asse er assemblies a	nnotation mbly are availab	spontaneous,
of ra	mice, provi diation- or c rganism	ding a wealth o chemically-indu Overview Click on cheor d 5 § 7 d 6 7	See also chromos	t genotypes and phe transgenic mutants p c Genome list Organ omes c to open MapViewer 0 11 12 13 14 15	notypes for g provide poten nelle List	Assembly	As	. In addition, the sembly and An Default asse er assemblies a	nnotation mbly are availab	spontaneous,
Assemb	mice, provi diation- or c rganism	ding a wealth o chemically-indu Overview Click on chee 4 5 6 7 1 9 9 9 9 1 9 9 9 9 1 9 9 9 9 9 1 1 1 1	of different ced, and See also Chromos Chromos Chromos 19 1 19 X rganisi	t genotypes and phe transgenic mutants p c <u>Genome list Organ</u> omes c to open MapViewer 0 11 12 13 14 15 0 1 10 10 10 10 1 10 10 10 1 10 10 10 1 10	notypes for g provide poten nelle List	Assembly Last seque	As 5 gth Name ence update	. In addition, the sembly and Au Default asse er assemblies a	nnotation mbly are availab	le 23-Feb
Assemb house mou	mice, provi diation- or o rganism	ding a wealth o chemically-indu Overview Cick on other 4 5 6 7 1	of different ced, and See also Chromos Chromos 19 1 19 X rganisi	t genotypes and phe transgenic mutants p c <u>Genome list</u> <u>Organ</u> omes c to open MapViewer 0 11 12 13 14 15 0 11 12 13 14 15	16 17 18	Assembly Last seque	As 5 gth Name anism	. In addition, the sembly and Au Default asse er assemblies a	nnotation mbly are availab	le 23-Feb
Assemb house mou Show of Results	mice, provi diation- or o rganism 1 2 3 1 1 2 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	ding a wealth o chemically-indu Overview Click on chee d 5 6 7 ation by o 0900	of different ced, and See also Chromos Base also Chromos 19 X rganisi Show all	t genotypes and phe transgenic mutants p c <u>Genome list Organ</u> omes c to open MapViewer <u>0 11 12 13 14 15</u> <u>1 12 13 14 15</u>	16 17 18	Assembly Last seque	As 5 other 5 oth Name ence update	In addition, the sembly and Ar Default asse er assemblies a	nnotation mbly are availab	ke 23-Fab
Assemb house mou Show of Results Organis	mice, providation- or or rganism	ding a wealth o chemically-indu Overview Cick on cheer 4 5 6 7 ation by o 090) ssemblies 6 Name	of different ced, and See also Chromos Moveme name 8 9 11 19 X rganisi Show all	t genotypes and phe transgenic mutants p c <u>Senome list</u> <u>Organ</u> omes c to open MapViewer 0 11 12 13 14 15 0 1 12 13 14 15 0 1 10 10 10 10 0 11 12 1 assemblies 1 assemblies Submitter	16 17 18	Assembly Last seque	As 5 gth Name ance update As As Senome epresentation	In addition, the sembly and Au Default asse er assemblies a semblies a sem	of 1 Next	<pre>> Last >> Default status</pre>
Assemb house mou Show of Results Organis Mus mus	mice, providation- or conganism	ding a wealth o chemically-indu Overview Click on chee 4 5 6 7 Click on chee 4 5 6 7 ation by o 090) ssemblies 6 Name GRCm38 UCSC Name:	of different ced, and See also thromos thromos 19 X rganisi Show all Show all	t genotypes and phe transgenic mutants p c <u>Genome list</u> Organ omes c lo open MapViewer 1 12 13 14 15 1 12 15 1 12 15 1 12 15 1 12 15 1 12 15 1 15 1	encing Consor	Assembly Last seque ch by orga	As 5 other studies As 5 oth Name ence update anism << First Senome epresentation complete	In addition, the sembly and Air Default asse er assemblies a semblies a se	nnotation mbly are availab of 1 Next Version status latest	<pre>> Last >> Default status Not default</pre>
Assemb house mou Show of Results Organis Mus mus	mice, providation- or conganism	ding a wealth o chemically-indu Overview Cick on other d 5 6 7 ation by o 090) ssemblies 6 Name GRCm38 UCSC Name: ASM216v1	of different ced, and See also Chromos moveme name 8 9 11 19 X rganisi Show all mm10	t genotypes and phe transgenic mutants p c <u>Senome list</u> <u>Organ</u> omes c to open MapViewer 0 11 12 13 14 15 0 1 1 12 15 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	encing Consor	Assembly Last seque	d other studies s More As S gth Name ance update anism << First Senome epresentation omplete	In addition, the sembly and Au Default asse er assemblies a semblies a sem	of 1 Next Version status latest	<pre>> Last >> Default status Not default Not default</pre>
Assemb house mou Show of Results Organis Mus mus Mus mus Mus mus	mice, providation- or conganism of the second secon	ding a wealth o chemically-indu Overview Cick on chee 4 5 6 7 Cick on chee 4 5 6 7 ation by o 090) ssemblies 6 Name GRCm38 UCSC Name: ASM216v1 Mm_Celera	of different ced, and See also Chromos Thromos 19 X rganisi Show all mm10	t genotypes and phe transgenic mutants p c <u>Genome list</u> Organ omes c lo open MapViewer 1 12 13 14 15 1 12 15 1 15 1	encing Consor	Assembly Last seque ch by orga	As 5 other studies As 5 oth Name ence update anism << First Senome epresentation omplete omplete omplete	In addition, the sembly and Air Default asse er assemblies a semblies a se	of 1 Next Next Version status latest latest latest	Spontaneous, Ne 23-Feb 23-Feb 23-Feb 23-Feb 23-Feb Not Status Not default Not default Not default Not default Not default
Assemb house mou Show of Results Organis Mus mus Mus mus Mus mus	mice, providation- or conganism	ding a wealth o chemically-indu Overview Cick on other 4 5 6 7 ation by o 090) ssemblies 6 Name GRCm38 UCSC Name: ASM216v1 Mm_Celera MmusALLPAT	Mainformation of different ced, and see also chromosome and the second s	t genotypes and phe transgenic mutants p c <u>Senome list</u> Organ omes c to open MapViewer 0 11 12 13 14 15 0 1 1 12 15 0 1 1 1 1 12 15 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	encing Consor	Assembly Last seque ch by orga	As A	In addition, the sembly and Automatic assembly and Automatic assemblies a semblies a	of 1 Next Version latest latest latest	spontaneous,
Assemb house mou Show of Results Organis Mus mus Mus mus Mus mus Mus mus	mice, providation- or c rganism (1 2 3 ly inform se (taxid:10 m) culus (culus (cul	ding a wealth o chemically-indu Overview Click on chee 4 5 6 7 ation by o 090) ssemblies 6 Name GRCm38 UCSC Name: ASM216v1 Mm_Celera MmusALLPATI MmusQAP1	Mainterent ced, and See also Chromos Mainterent Show all Mainterent Show all Mainterenterent Mainterenterenterenterent Mainterenterenterenteren	t genotypes and phe transgenic mutants p c <u>Genome list</u> Organ omes c lo open MapViewer <u>11</u> 12 13 14 15 <u>12</u> 13 14 15 <u>13</u> 14 15 <u>14</u> 12 13 14 15 <u>15</u> 12 13 14 15 <u>16</u> 12 13 14 15 <u>17</u> 12 13 14 15 <u>18</u> 12 12 12 13 14 15 <u>18</u> 12 12 12 12 12 12 12 12 12 12 12 12 12	encing Consor	Assembly Last seque ch by orga	As Souther studies As Souther studies As Souther studies As Complete Anism As	In addition, the sembly and Ai Default asse er assemblies a er assemblies a chromosome Chromosome Scaffold Scaffold	of 1 Next Next Next Next Next Next Next Next	Spontaneous, Ne 23-Feb 23-Feb 23-Feb 23-Feb 23-Feb 23-Feb 23-Feb 23-Feb 23-Feb 24-Feb 23-Feb 24-Feb 24-Feb 25-Feb 25-Feb 25-Feb 26-Feb

Figure 1. Accessing the Assembly database. *Top panel*. The Assembly main page with the search box and access to the Assembly Browser (Browse by Organism, red circle). *Middle panel* the mouse genome overview with showing information for the Default assembly (MGSCv37) with a link to all the assemblies. *Bottom panel*. The Assembly Browser showing the six latest assemblies for the mouse.

escription: Genome	Reference Consortium Mouse Bui	Assembly	Assembly Unit Name					
Organism name: Mus	musculus		Primary A	Assembly	(C57BL/6J)			
Submitter: Mouse Ge	nome Sequencing Consortium		129P2/OI	129P2/OlaHsd				
Synonym: mm10	and with alt losi		100000					
Assembly level: Chro	mosome		12951/50	ImJ				
Senome representatio	n: complete		129S2/Svi	Pas				
SenBank Assembly ID	: GCA_000001635.2 (latest)		12956/Sv	EvTac				
RefSeq Assembly ID:	GCF_000001635.20 (latest)		129S7/Sv	EvBrd-Hpr	t-b-m2			
RefSeq Assembly and	GenBank Assembly Identical:	yes	129X1/Sv	J				
listory (Show revision	history)		A/1					
Slobal statistics			100					
siobal statistics			AKR/J					
Number of regions with	h alternate loci or patches		72 BALB/c					
Total sequenced base	s	,140 CAST/Ei						
Gaps between scaffold	ds		191 NOD/Mrk1	Tac				
Number of scaffolds			262 NOD/ShiL	tJ				
Confield NEC		50 500	RIII					
Scattold N50		52,589	UNKNOW	/N				
Number of contigs			726 non-nucle	ar				
Contig N50		32,273	,079					
Total number of chrom	osomes and plasmids		22					
		Molecule name	GenBank ID		RefSeq ID	Unlocalized seq count		
		Molecule name	GenBank ID		RefSeq ID	Unlocalized seq count		
		Molecule name	GenBank ID CM000994.2		RefSeq ID NC_000067.6	Unlocalized seq count 5		
		Molecule name	GenBank ID CM000994.2 CM000995.2		RefSeq ID NC_000067.6 NC_000068.7	Unlocalized seq count 5 0		
		Molecule name 1 2 3	GenBank ID CM000994.2 CM000995.2 CM000996.2		RefSeq ID NC_000067.6 NC_000068.7 NC_000069.6	Unlocalized seq count 5 0 0		
		Molecule name 1 2 3 4	GenBank ID CM000994.2 CM000995.2 CM000996.2 CM000997.2		RefSeq ID NC_000067.6 NC_000068.7 NC_000069.6 NC_000070.6	Unlocalized seq count 5 0 0 6		
		Molecule name 1 2 3 4 5	GenBank ID CM000994.2 CM000995.2 CM000996.2 CM000997.2 CM000998.2		RefSeq ID NC_000067.6 NC_000068.7 NC_000069.6 NC_000070.6 NC_000071.6	Unlocalized seq count 5 0 0 6 5		
Regions		Molecule name 1 2 3 4 5 6	GenBank ID CM000994.2 CM000995.2 CM000996.2 CM000997.2 CM000998.2 CM000999.2		RefSeq ID NC_000067.6 NC_000068.7 NC_000069.6 NC_000070.6 NC_000071.6 NC_000072.6	Unlocalized seq count 5 0 0 6 5 0		
Regions	Location	Molecule name 1 2 3 4 5 6 7	GenBank ID CM000994.2 CM000995.2 CM000996.2 CM000997.2 CM000999.2 CM000999.2 CM000999.2		RefSeq ID NC_000067.6 NC_000068.7 NC_000069.6 NC_000070.6 NC_000071.6 NC_000072.6 NC_000073.6	Unlocalized seq count 5 0 0 6 5 0 1		
Regions Name Odz1	Location X-42891057-43298842	Molecule name 1 2 3 4 5 6 7 8	GenBank ID CM000994.2 CM000995.2 CM000996.2 CM000997.2 CM000999.2 CM000999.2 CM001000.2 CM001001.2		RefSeq ID NC_000067.6 NC_000068.7 NC_000069.6 NC_000070.6 NC_000071.6 NC_000072.6 NC_000073.6 NC_000074.6	Unlocalized seq count 5 0 0 6 5 0 1 0		
Regions Name Odz1 region57	Location X:42891057-43298842 X:72166181-72983206	Molecule name 1 2 3 4 5 6 7 8 9	GenBank ID CM000994.2 CM000995.2 CM000996.2 CM000996.2 CM000999.2 CM000999.2 CM001000.2 CM001001.2 CM001001.2		RefSeq ID NC_000067.6 NC_000068.7 NC_000069.6 NC_000070.6 NC_000071.6 NC_000071.6 NC_000073.6 NC_000074.6 NC_000076.6	Unlocalized seq count 5 0 0 6 5 0 1 0 1 0 0		
Regions Name Odz1 region57 Yis	Location X:42891057-43298842 X:72166181-72363206 X:400580384-100841870	Molecule name 1 2 3 4 5 6 7 8 9 10	GenBank ID CM000994.2 CM000995.2 CM000996.2 CM000996.2 CM000997.2 CM000999.2 CM000999.2 CM001000.2 CM001001.2 CM001002.2 CM001002.2		RefSeq ID NC_000067.6 NC_000068.7 NC_000069.6 NC_000070.6 NC_000071.6 NC_000073.6 NC_000073.6 NC_000075.6 NC_000075.6	Unlocalized seq count 5 0 0 6 5 0 1 1 0 0 0		
Regions Name Odz1 region57 Xic Xic Xic Xic Xic Xic Xic Xic Xic Xic	Location X.42891057.43298842 X.72166181-72363206 X:103589284-103841870 V.73002841.7314164	Molecule name 1 2 3 4 5 6 7 8 9 10 11	GenBank ID CM000994.2 CM000994.2 CM000996.2 CM000997.2 CM000999.2 CM000999.2 CM001000.2 CM001001.2 CM001001.2 CM001003.2 CM001003.2		RefSeq ID NC_000067.6 NC_000068.7 NC_000069.6 NC_000070.6 NC_000071.6 NC_000072.6 NC_000073.6 NC_000075.6 NC_000075.6 NC_000075.6	Unlocalized seq count 5 0 0 6 5 0 1 1 0 0 0 0 0		
Regions Name Odz1 region57 Xic Xir_region_1	Location X.42891057-43298842 X.72166181-72363206 X:103589284-103841870 X:73003651-73121964 V:73184167 77000000	Molecule name 1 2 3 4 5 6 7 8 9 10 11	GenBank ID CM000994.2 CM000994.2 CM000996.2 CM000998.2 CM000999.2 CM000999.2 CM001000.2 CM001001.2 CM001001.2 CM001003.2 CM001004.2		RefSeq ID NC_000067.6 NC_000068.7 NC_000069.6 NC_000070.6 NC_000071.6 NC_000072.6 NC_000073.6 NC_000075.6 NC_000075.6 NC_000075.6	Unlocalized seq count S 0 0 6 5 0 1 1 0 0 0 0 0 0 0		
Regions Name Odz1 region57 Xic Xir_region_1 Xir_region_2	Location X.42891057-43298842 X.72166181-72363206 X:103589284-103841870 X.73003651-73121964 X.73185425-73263090	Molecule name 1 2 3 4 5 6 7 8 9 10 11 12	GenBank ID CM000994.2 CM000994.2 CM000996.2 CM000998.2 CM000999.2 CM001009.2 CM001000.2 CM001001.2 CM001001.2 CM001003.2 CM001004.2 CM001005.2		RefSeq ID NC_000067.6 NC_000068.7 NC_000069.6 NC_000070.6 NC_000071.6 NC_000072.6 NC_000073.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000077.6 NC_000075.6	Unlocalized seq count 5 0 0 6 5 0 1 5 0 1 0 0 0 0 0 0 0		
Regions Name Odz1 region57 Xic Xir_region_1 Xir_region_2 If_cluster	Location X:42891057-43298842 X:72166181-72363208 X:103589284-103841870 X:73003651-73121964 X:73185425-73263090 1:173701136-174051183	Molecule name 1 2 3 4 5 6 7 8 9 10 11 12 13	GenBank ID CM000994.2 CM000994.2 CM000996.2 CM000997.2 CM000999.2 CM001009.2 CM001000.2 CM001001.2 CM001001.2 CM001003.2 CM001004.2 CM001005.2 CM001006.2		RefSeq ID NC_000067.6 NC_000068.7 NC_000069.6 NC_000070.6 NC_000071.6 NC_000072.6 NC_000073.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000075.6	Unlocalized seq count S 0 0 6 5 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0		
Regions Name Odz1 region57 Xic Xir_region_1 Xir_region_2 If_cluster NOD_mouse_Idd5.1	Location X:42891057-43298842 X:72166181-72363206 X:103589284-103841870 X:73003651-73121964 X:73185425-73263090 1:173701136-174051183 1:60640809-63588262	Molecule name 1 2 3 4 5 6 7 8 9 10 11 12 13 14	GenBank ID CM000994.2 CM000994.2 CM000996.2 CM000997.2 CM000999.2 CM001009.2 CM001000.2 CM001001.2 CM001001.2 CM001003.2 CM001004.2 CM001005.2 CM001006.2 CM001007.2		RefSeq ID NC_000067.6 NC_000068.7 NC_000069.6 NC_000070.6 NC_000071.6 NC_000072.6 NC_000073.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000076.6 NC_000076.6 NC_000079.6 NC_000079.6	Uniocalized seq count 5 0 6 5 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		
Regions Name Odz1 region57 Xic Xir_region_1 Xir_region_2 If_cluster NOD_mouse_Idd5.1 NOD_mouse_Idd5.3	Location X:42891057-43298842 X:72166181-72363206 X:103589284-103841870 X:73003651-73121964 X:73185425-73263090 1:173701136-174051183 1:60640809-63588262 1:65597865-69265254	Molecule name 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15	GenBank ID CM000994.2 CM000994.2 CM000996.2 CM000997.2 CM000999.2 CM001009.2 CM001000.2 CM001001.2 CM001001.2 CM001003.2 CM001004.2 CM001005.2 CM001006.2 CM001007.2 CM001008.2		RefSeq ID NC_000067.6 NC_000068.7 NC_000069.6 NC_000070.6 NC_000071.6 NC_000072.6 NC_000073.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000076.6 NC_000078.6 NC_000078.6 NC_000089.6 NC_000081.6	Uniocalized seq count S 0 0 6 5 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		
Regions Name Odz1 region57 Xic Xir_region_1 Xir_region_2 If_cluster NOD_mouse_Idd5.1 NOD_mouse_Idd5.3	Location X:42891057-43298842 X:72166181-72363206 X:103589284-103841870 X:73003651-73121964 X:73185425-73263090 1:173701136-174051183 1:60640809-63588262 1:65597865-69265254	Molecule name 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16	GenBank ID CM000994.2 CM000994.2 CM000996.2 CM000997.2 CM000999.2 CM001009.2 CM001000.2 CM001001.2 CM001001.2 CM001003.2 CM001004.2 CM001005.2 CM001006.2 CM001007.2 CM001008.2 CM001009.2		RefSeq ID NC_000067.6 NC_000068.7 NC_000069.6 NC_000070.6 NC_000071.6 NC_000072.6 NC_000073.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000077.6 NC_000078.6 NC_000078.6 NC_000089.6 NC_000081.6 NC_000081.6	Uniocalized seq count S 0 0 6 5 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		
Regions Name Odz1 region57 Xic Xir_region_1 Xir_region_2 If_cluster NOD_mouse_Idd5.1 NOD_mouse_Idd5.3 NOD_mouse_Idd5.4	Location X:42891057-43298842 X:72166181-72363206 X:103589284-103841870 X:73003651-73121964 X:73185425-73263090 1:173701136-174051183 1:60640809-63588262 1:65597865-69265254 4:43034066,130853062	Molecule name 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17	GenBank ID CM000994.2 CM000994.2 CM000996.2 CM000997.2 CM000998.2 CM001099.2 CM001000.2 CM001001.2 CM001001.2 CM001004.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2		RefSeq ID NC_000067.6 NC_000068.7 NC_000069.6 NC_000070.6 NC_000071.6 NC_000072.6 NC_000073.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000077.6 NC_000077.6 NC_000078.6 NC_000078.6 NC_000081.6 NC_000081.6 NC_000081.6	Uniocalized seq count S 0 0 6 5 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		
Regions Name Odz1 region57 Xic Xir_region_1 Xir_region_2 If_cluster NOD_mouse_Idd5.1 NOD_mouse_Idd5.3 NOD_mouse_Idd5.3	Location X:42891057-43298842 X:72166181-72363206 X:103589284-103841870 X:73003651-73121964 X:73185425-73263090 1:173701136-174051183 1:60640809-63588262 1:65597865-69265254 4:400044064,100650047	Molecule name 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18	GenBank ID CM000994.2 CM000994.2 CM000996.2 CM000996.2 CM000998.2 CM000998.2 CM001009.2 CM001000.2 CM001001.2 CM001001.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2		RefSeq ID NC_000067.6 NC_000068.7 NC_000069.6 NC_000070.6 NC_000071.6 NC_000072.6 NC_000073.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000077.6 NC_000075.6 NC_000055.6 NC_000055.6 NC_000055.6 NC_000055.6 NC_000055.6 NC_00055.6 NC_00055.6 NC_00055.6 NC_00055.6 NC_00055.6 NC_00055.6 NC_00055.6 NC_00055.6 NC_00055.6 NC_00055.6 NC_00055.6 NC_00055.6 NC_00055.6 NC_00055.6 NC_00055.6 NC_00055.6 NC_00055.6 NC_0055.6 NC_00055.6 NC_00055.6 NC_0055.6 NC_0055.6 NC_0055.6 NC_0055.6 N	Uniocalized seq count S 0 0 6 5 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0		
Regions Name Odz1 region57 Xic Xir_region_1 Xir_region_2 If_cluster NOD_mouse_Idd5.1 NOD_mouse_Idd5.3 AIOD_mouse_Idd5.3	Location X:42891057-43298842 X:72166181-72363206 X:103589284-103841870 X:73003651-73121964 X:73185425-73263090 1:173701136-174051183 1:60640809-63588262 1:65597865-69265254 4:490944064,490659042	Molecule name 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19	GenBank ID CM000994.2 CM000996.2 CM000996.2 CM000998.2 CM000998.2 CM001099.2 CM001000.2 CM001000.2 CM001001.2 CM001001.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001015.2 CM001015.2		RefSeq ID NC_000067.6 NC_000068.7 NC_00007.6 NC_00007.6 NC_00007.6 NC_00007.6 NC_00007.6 NC_00007.6 NC_00007.6 NC_000076.6 NC_000077.6 NC_000077.6 NC_000078.6 NC_000078.6 NC_00008.6	Uniocalized seq count S 0 0 6 5 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0		
Regions Name Odz1 region57 Xic Xir_region_1 Xir_region_2 Ifj_cluster NOD_mouse_Idd5.1 NOD_mouse_Idd5.3 AIOD_mouse_Idd5.3	Location X:42891057-43298842 X:72166181-72363206 X:103589284-103841870 X:73003651-73121964 X:73185425-73263090 1:173701136-174051183 1:60640809-63588262 1:65597865-69265254 4:490944084,490859042	Molecule name 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X	GenBank ID CM000994.2 CM000994.2 CM000996.2 CM000996.2 CM000998.2 CM000998.2 CM001009.2 CM001000.2 CM001001.2 CM001001.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001008.2 CM001008.2 CM001010.2 CM001011.2 CM001011.2 CM001011.2 CM001011.2		RefSeq ID NC_000067.6 NC_000068.7 NC_000068.7 NC_000070.6 NC_000071.6 NC_000072.6 NC_000073.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000081.6 NC_000085.6 NC_000085.6 NC_000085.6	Uniocalized seq count 5 0 0 6 5 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		
Regions Name Odz1 region57 Xic Xir_region_1 Xir_region_2 Ifj_cluster NOD_mouse_Idd5.1 NOD_mouse_Idd5.3 XIOD_mouse_Idd5.3	Location X:42891057-43298842 X:72166181-72363206 X:103589284-103841870 X:73003651-73121964 X:73185425-73283090 1:173701136-174051183 1:60640809-63588262 1:655597865-69265254	Molecule name 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X Y	GenBank ID CM000994.2 CM000994.2 CM000995.2 CM000996.2 CM000998.2 CM000998.2 CM001009.2 CM001000.2 CM001001.2 CM001001.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001005.2 CM001010.2 CM001011.2 CM001011.2 CM001011.2 CM001011.2 CM001011.2 CM001011.2 CM001011.2		RefSeq ID NC_000067.6 NC_000068.7 NC_000068.7 NC_000070.6 NC_000071.6 NC_000071.6 NC_000073.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000075.6 NC_000085.6 NC_000085.6 NC_000085.7 NC_000085.7	Uniocalized seq count 5 0 0 6 5 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		

Figure 2. Aspects of the mouse GRCm38 assembly. *Top left panel*. General Assembly Definition showing names, synonyms, and assembly identifiers. *Top right panel*. Assembly units including the primary assembly for C57BL/6J and alternate loci for other mouse strains. These alternate loci are also available by region (*Lower left panel*). *Lower right panel*. Listing of the molecules and corresponding nucleotide accessions making up the selected assembly unit, the C57BL/6J primary assembly in this case. Detailed statistics are available for each molecule in the assembly in a separate tab (not shown).

New Videos on NCBI's YouTube Channel

Eleven new tutorial videos have been added to the NCBI YouTube channels in the past few months. To make topics of interest easier to find, the Tutorials playlist now provides special playlists for certain resources. The channel now features separate tutorial playlists for Genome Workbench (7 videos), Sequence Viewer (4 videos), My NCBI (4 videos), Genetic Testing Registry (2 videos) and General (22 videos).

Five of the recent videos are in the General playlist and include tutorials on using the new Advanced Search Builder in PubMed (video, advanced search page), an overview of RefSeqGene reference standard records for selected human genes (video, resource page), an introduction to the E-utilities, the programming interface to the Entrez system (video, E-Utilities Help Manual), a demonstration of the highlight sequence features tool in sequence databases (video, NCBI News), and a video on how to use Genome Remapping Tool (video, tool page) that can map coordinates of genes and other markers from one genome build to another.

Three of the new videos are about My NCBI, the service that allows registered users to customize their experience and to save and share results, searches, and preferences through their accounts. New titles in the My NCBI playlist are My Bibliography, Save Searches and Set E-mail Alerts, and Save Search Results in Collections.

One new video demonstrating how load a genome into Genome Workbench was recently added to the playlist for Genome Workbench, NCBI's standalone sequence analysis and annotation platform.

Most recently a new playlist was created for two tutorials (GTR: Homepage and Basic Search Functions and GTR: Locate a Test in Under Three Minutes) featuring the newly launched Genetic Testing Registry, a repository of information about available genetic tests. Additional information about the GTR is provided in the following section of this newsletter.



The Genetic Testing Registry: Finding Genetic Tests and Related Information

The NCBI recently released the Genetic Testing Registry (GTR). This new resource is a voluntary registry of genetic tests and laboratories with detailed information about the tests and their providers. The initial scope of GTR includes single gene tests for Mendelian disorders, as well as arrays, panels and pharmacogenetic tests. The registry includes detailed information about the purpose of the test, methodology, analytical and clinical validity, and information on clinical usefulness. GTR provides access to information from the GeneReviews book on the NCBI Bookshelf – peer reviewed descriptions of genetic diseases and information on genetics tests and NCBI molecular databases such as Gene. GTR is a central hub for information about genetic conditions and also provides context-specific links to a variety of resources, including practice guidelines, published literature, and genetic information. As mentioned in the previous section of this newsletter, two new

videos on the NCBI YouTube channel provide quick introductions to the GTR. The original NIH press release has more information about the GTR.

All GTR	Tests	Conditions/Phenotypes	Genes	Labs	GeneReviews	
SCID			Search	All GTR		
Adenosi F RS-SCIE SCID du SCIDX1	ne Deamin): Severe o e to absen : X-linked s	ase-Deficient Severe C combined immunodefici t class II HLA antigens evere combined immun	ditions/pher	notypes, ge	enes, and labs.	
IMPORTAN submitters to tests or labo consumers genetics pro	T NOTE: In provide in ratories list with spece fessional	NIH does not independentl Information that is accurate sted in the GTR. GTR is no ific questions about a gene	y verify info and not m at a substitu atic test sho	ormation hisleading ute for me ould conta	submitted to the GTR NIH makes no endo edical advice. <i>Patien</i> a act a health care prov	R; it relies orsements ts and vider or a

BLAST News

BLAST 2.26+ Release

The latest version of the C++ build of BLAST+ (2.2.26) is now available from the BLAST FTP area and is running on the NCBI BLAST Web service. This new BLAST+ release contains a number of important changes and improvements including the three listed below.

Domain Enhanced Lookup Time Accelerated BLAST (DELTA-BLAST) is a new BLAST algorithm that can be more sensitive than standard protein-protein BLAST. DELTA-BLAST identifies conserved domains in the query sequence using Reverse PSI BLAST and then uses this information to construct a Position Specific Score Matrix (PSSM) then performs a PSSM search against the BLAST protein database. DELTA-BLAST can be invoked on the Protein-protein BLAST Web Service by selecting the DELTA-BLAST radio button in the "Program Selection" area of the submission form. The standalone BLAST package has DELTA-BLAST as a separate program (deltablast). Running DELTA-BLAST locally requires a special version of CDD database (cdd_delta) available from the BLAST db directory on the FTP site.

A new **Finite Size Correction** has been added to the to the blastp algorithm to improve the accuracy of BLAST statistics (Expect values). The new finite size correction especially improves statistics for matches for short query or short database sequences. Standalone BLAST now contains the program **makeprofiledb**, a C++ coded replacement for the NCBI C toolkit program formatrpsdb. Makeprofiledb can generate search sets for RPS-BLAST, including the specialized data needed by DELTA-BLAST.

Final version of C-toolkit BLAST Package

Version 2.2.26 is the final version of NCBI C language toolkit BLAST. The source code for these applications will no longer be developed, but will continue to be available. Users of these legacy programs should migrate to the BLAST+ applications that are being actively developed. The BLAST Command Line Applications User Manual provides help on transitioning to the BLAST+ applications.

Netblast (blastcl3) Service Discontinued: Replaced by remote Option in BLAST+

The Netblast client (blastcl3) that has provided batch search access to the NCBI Web BLAST service will be discontinued in the near future. The BLAST+ applications replace and improve upon the functions provided by blastcl3. Blastcl3 users should switch to BLAST+ as soon as possible. Locally installed BLAST+ applications can perform remote searches using the NCBI Web service when the 'remote' option is included on the command line. The BLAST+ remote service has a number of advantages over the blastcl3 application. Blastcl3 requires a persistent connection during the entire search, can only submit one query at a time, and is unable to return the BLAST Request ID (RID) used in the search. The BLAST+ remote service can submit multiple queries (from FASTA input) at once, poll for the results using the BLAST RID, and also print the RID in the BLAST report. Using the BLAST RID, it is possible to reformat the search locally with the blast_formatter application, reformat the search at the NCBI web site, or use analysis tools such as the BLAST treeview or the taxonomy report.

Changes in the BLAST Database List on the NCBI Web Services

A new **microbial 16S ribosomal** RNA sequence database is now available on nucleotidenucleotide BLAST search page. This database contains Archaeal and Bacterial 16S sequences from the Archaeal 16S Ribosomal RNA and Bacterial 16S Ribosomal RNA Targeted Loci Projects. This database should be helpful in classifying unknown microbial 16S sequences from a wide range of sources.

Sequences from **environmental samples** formerly available in the env_nr and the env_nt databases are now available in the Metagenomic proteins database and, for nucleotide sequences, through the Whole Genome Shotgun Contigs (WGS) database by selecting "metagenomes (taxid: 408169)" as an Organism limit.

The following image shows the selections needed on the BLAST submission form to search these three new or modified databases.

Choose Searc Database	How Set OHuman genomic + transcript OHuman genomic + transcript In the sequences (Bacteria and Archaea)
Choose Database	Search Set Metagenomic proteins(env_nr) Search Set
Choose Searc	ch Set
Database	OHuman genomic + transcript OMouse genomic + transcript Others (nr etc.):
	Whole-genome shotgun contigs (wgs)
Organism Optional	metagenomes (taxid:408169) Exclude + Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

CDD Results Now Shown for Translated BLAST (blastx) Searches

Conserved Domain Search results are now provided for all translated (blastx) searches with query sequences shorter than 10,000 bases. Conserved domain searches are performed with all six reading frames of the query sequence and results are reported for each frame that has matches. This is very useful for helping to characterize coding regions on genomic regions as shown immediately below from the results for a blastx search with a human endogenous retrovirus (AF164611).

Cons	served domains	On [gi 5802813 gb AF10	i4611.1]					View concise result	
Graphical su	mmary show options >	nekv-kros, complete sequel							
(+) RF #1	1599	catalytic motif Catalytic motif Actine site flap inhibiter binding site Ala active site	4500	4	. 6910		7536	0101	
Specific hits		Les HIV.							
Non-specific hits		duttra du Dus PCHR2 D PR d d			Ber				
Superfamilies		and and			Bonn				
Multi-domains		Pesi			and/2				
<pre>(+) RF #2 Non-specific hits Superfamilies Multi-domains</pre>	1500 Gog Gog DDF 15	3010 Gag_p24 Deg_p24 av	*505		6910 1		, 7500	9181	
	1 1508	2010	4500		6000		7500	9181	
(+) RF #3		ectiv nucleic acid bind NTP bindin Pano	e site Ling site active g site	site <u>A A</u>		imunosuppressi honotrimer i	HR1 ive resion interface		
Non-specific hits		- KHIL	RT_Rtv RT_2FREv_ RT_LTR RT_L RT_L	FNase FNase_	PB	HERV-K_e	HI GP4		
Superfamilies			RT_like su	Phose		HERV-K_A	Ebo		
Hulti-domains			DUT 4	and the second second	and the second	Control Street.	and all		

Remap and Variation Reporter: Two New Services for Mapping Locations onto Genome Builds

The Genome Remapping Service (Remap) and the Variation Reporter are related tools that find locations on current and past genome builds.

The Remap tool translates or projects the coordinates of genes, variants (SNPs), and other sequence-based markers from one genome assembly (build) to another for human, mouse, rat, zebrafish and sea urchin (*Strongylocentrotus purpuratus*). It also includes a Clinical Remap version that performs coordinate remapping between genome assemblies and the reference standard RefSeqGene records. Figure 3 and Figure 4 show the submission and results for the Remap service. Locations to be projected can be in a variety of common genome annotation formats such as UCSC Browser Extensible Data (BED) format, Gene Transfer Format (GTF), Generic Feature Format (GFF and GFF3), Human Genome Variation Society (HGVS) nomenclature, and Genome Variation Format (GVF) among others. When projection of features is successful, the service reports the new locations with the submitted annotations in the selected format for downloading and also provides output in a format suitable for loading into Genome Workbench, the NCBI's standalone sequence analysis and annotation platform. A programming interface (API) is

Genome Information	Source Assembly *	Tarnet Assembly *		
Homo sapiens	HuRef	HuRef		
Start typing to get a list of available organisms	GRCh37.p5 NCBI36 (hg18) NCBI35 (hg17)	NCBI36 (hg18) NCBI35 (hg17) NCBI34 (hg16)		
Alignments performed: Jan	uary 16, 2012			
First Pass	Assembly-Assembly Clinical Remap			E
GRCh37.p5 Coverage:				-
NCBI36 (ho18) Cover	Genome Information			
Rement Identify 0.000	Available only for human			
Percent Identity. 0.99	Available only for human			
	GRCh37 (ho19)	RefSeoGene	a to "	
Remapping Options	GRCh37.p7			
Minimum ratio of bases tha	RefSeqGene			
Maximum ratio for difference				
Maximum ratio for dillerend	Remapping Options			
Allow multiple locations to I	Define RefSeqGenes		Not all regions of the genome have	
Marga Fragmants: 54 9	Map to any available RefSec	Gene sequence	RefSeqGenes.	
merge i ragments. 💽 🥌	Map only to the RefSeqGen	es I specify	You can choose to get data for any av	vailable
			RefSeqGene or only specific ones. To	0
			RefSeqGene or only specific ones. To request a RefSeqGene for a gene clic	o ck <u>here</u>
			RefSeqGene or only specific ones. To request a RefSeqGene for a gene clic	o ck <u>here</u>
	Define Transcripts/Proteins		RefSeqGene or only specific ones. To request a RefSeqGene for a gene clic	o ck <u>here</u>
	Define Transcripts/Proteins	Ps associated with F	RefSeqGene or only specific ones. To request a RefSeqGene for a gene clic RefSeqGenes	o ck <u>here</u>
	Define Transcripts/Proteins ✓ Provide locations on NMs/Ni ✓ Provide locations on NMs/Ni	Ps associated with F	RefSeqGenes	o ck <u>here</u>
ata	Define Transcripts/Proteins Provide locations on NMs/Ni Provide locations on NMs/Ni	Ps associated with F Ps even if there is no	RefSeqGene or only specific ones. To request a RefSeqGene for a gene clic RefSeqGenes	o ck <u>here</u>
ata Input format:	Define Transcripts/Proteins Provide locations on NMs/Ni Provide locations on NMs/Ni Output format: CFF3	Ps associated with F Ps even if there is no	RefSeqGene or only specific ones. To request a RefSeqGene for a gene clic RefSeqGenes <u>a RefSeaGene</u>	o ck <u>here</u>
ata Input format: BED : 6	Define Transcripts/Proteins Provide locations on NMs/Ni Provide locations on NMs/Ni Output format: GFF3	Ps associated with F Ps even if there is no	RefSeqGene or only specific ones. To request a RefSeqGene for a gene clic RefSeqGenes <u>D RefSeaGene</u>	o ck <u>here</u>
ata Input format: BED : 6	Define Transcripts/Proteins Provide locations on NMs/Ni Provide locations on NMs/Ni Output format: CFF3	Ps associated with F Ps even if there is no even i	RefSeqGenes	o ck <u>here</u>
ata Input format: BED 0	Define Transcripts/Proteins Provide locations on NMs/NI Provide locations on NMs/NI Output format: CFF3	Ps associated with F Ps even if there is no e Browse	RefSeqGene or only specific ones. To request a RefSeqGene for a gene clic RefSeqGenes D RefSeaGene	o ck <u>here</u>
ata Input format: BED 0	Define Transcripts/Proteins Provide locations on NMs/NI Provide locations on NMs/NI Output format: GFF3	Ps associated with F Ps even if there is no Browse)	RefSeqGene or only specific ones. To request a RefSeqGene for a gene clic RefSeqGenes <u>c RefSeqGene</u>	o ck <u>here</u>
ata Input format: BED • • • Upload a file: OR Paste data here: Chr19 577423	Define Transcripts/Proteins Provide locations on NMs/Ni Provide locations on NMs/Ni Output format: GFF3 Output format: GFF3	Ps associated with F Ps even if there is no Browse)	RefSeqGene or only specific ones. To request a RefSeqGene for a gene clic RefSeqGenes D. RefSeqGene C	o ck <u>here</u>
ata Input format: BED • • • Upload a file: OR Paste data here: Chr19 577423 Chr19 108657 Chr19 10857	Define Transcripts/Proteins Provide locations on NMs/Ni Provide locations on NMs/Ni Output format: GFF3 Output format: GFF3 77 57746915 AURKC 8 1095391 POLR2E 8 1228434 STK11	Ps associated with F Ps even if there is not Browse	RefSeqGene or only specific ones. To request a RefSeqGene for a gene clic RefSeqGenes <u>D RefSeqGene</u>	o ck <u>here</u>
ata Input format: BED Upload a file: OR Paste data here: Chr19 577423 Chr19 108657 Chr19 120579 Chr19 120579 Chr19 125745	Define Transcripts/Proteins Provide locations on NMs/Ni Provide locations on NMs/Ni Output format: GFF3 Output format: GFF3 1095391 POLR2E 8 1095391 POLR2E 8 1228434 STK11 50 45808541 MARK4	Ps associated with F Ps even if there is not Browse)	RefSeqGene or only specific ones. To request a RefSeqGene for a gene clic RefSeqGenes <u>D RefSeaGene</u>	o ck <u>here</u>
ata Input format: BED © © Upload a file: OR Paste data here: Chr19 577423 Chr19 108657 Chr19 120579 Chr19 120579 Chr19 457545 Multiple lines	Define Transcripts/Proteins Provide locations on NMs/Ni Provide locations on NMs/Ni Output format: GFF3 Output format: GFF3 77 57746915 AURKC 8 1095391 POLR2E 8 1228434 STK11 50 45808541 MARK4	Ps associated with F Ps even if there is not Browse)	RefSeqGene or only specific ones. To request a RefSeqGene for a gene clic RefSeqGenes <u>D RefSeaGene</u>	o ck <u>here</u>
ata Input format: BED © © Upload a file: OR Paste data here: Chr19 577423 Chr19 108657 Chr19 108657 Chr19 120579 Chr19 120579 Chr19 457545 Multiple lines into the text	Define Transcripts/Proteins Provide locations on NMs/Ni Provide locations on NMs/Ni Output format: GFF3 Output format: GFF3 77 57746915 AURKC 8 1095391 POLR2E 8 1228434 STK11 50 45808541 MARK4	Ps associated with F Ps even if there is not Browse)	RefSeqGene or only specific ones. To request a RefSeqGene for a gene clic RefSeqGenes <u>D RefSeaGene</u>	o ck <u>here</u>
ata Input format: BED © © Upload a file: OR Paste data here: You can paste multiple lines into the text area	Define Transcripts/Proteins Provide locations on NMs/Ni Provide locations on NMs/Ni Output format: CFF3 Output format: CFF3 77 57746915 AURKC 8 1095391 POLR2E 8 1228434 S7K11 50 45808541 MARK4	Ps associated with F Ps even if there is no Browse	RefSeqGene or only specific ones. To request a RefSeqGene for a gene clic RefSeqGenes D RefSeaGene	o ck <u>here</u>
ata Input format: BED © © Upload a file: OR Paste data here: Chr19 577423 Chr19 108657 Chr19 108657 Chr19 120579 Chr19 120579 Chr19 120579 Chr19 457545 Wou can paste multiple lines into the text area Chr19 . Variat	Define Transcripts/Proteins Provide locations on NMs/Ni Provide locations on NMs/Ni Output format: CFF3 Output format: CFF3 77 57746915 AURKC 8 1095391 POLR2E 8 1228434 STK11 50 45808541 MARK4 ion 196079 196079 .	Ps associated with F Ps even if there is no Browse)	RefSeqGene or only specific ones. To request a RefSeqGene for a gene clic RefSeqGenes D RefSeaGene C ID=rs4046282; gbkoy=Variatio ID=rs4046282; gbkoy=Variatio	
ata Input format: BED © © Upload a file: OR Paste data here: You can paste multiple lines into the text ared chr19 Variat. chr19 Variat. chr19 Variat. chr19 Variat.	Define Transcripts/Proteins Provide locations on NMs/Ni Provide locations on NMs/Ni Output format: CFF3 Output format: CFF3 V0100000000000000000000000000000000000	Ps associated with F Ps even if there is not Browse	RefSeqGene or only specific ones. To request a RefSeqGene for a gene clic RefSeqGenes D RefSeqGenes C C ID=rs4046282;gbkey=Variatio ID=rs4046282;gbkey=Variatio ID=rs4046286;gbkey=Variatio ID=rs4046286;gbkey=Variatio	

Figure 3. Submission forms for the Genome Remapping Service. **A**. Genome Remap set to map a set of locations from human build 37 to build 36. **B**. The Clinical Remap tab set to map a set of locations from build 37 to RefSeqGene records. **C**. BED format for gene position shown in the data text area for the Genome Remap. **D**. Data in GFF3 format showing the positions of variations to be projected on to RefSeqGene records in Clinical Remap.

also available for the Remap service. A demonstration PERL script (remap_api.pl) that accesses the service is available from the NCBI FTP site.

1 PERMIT PROVIDE TABLE 1	Cummon Data								
bonnous	Summary Data								
ID	Source Features		Remapped Features		e Intervals	Remapped Intervals			
Chr19	4		4	4		4			
Mappin	g Report (sa I Full Mapping Re	mple)							
Feature	Src Intervals	Remap Intervals	Src Location	Src Length	Map Location	Map Length	Coverage		
AURKC	1	1	Chr19:57742377 -57746915	4539	Chr19:62434189 -62438727	4539	1.00000		
POLR2E	: 1 1		Chr19:1086578 -1095391	8814	Chr19:1037578 -1046391	8814	1.00000		
STK11	1 1 1		Chr19:1205798 -1228434	22637	Chr19:1156798 -1179434	22637	1.00000		
MARK4	1	1	Chr19:45754550 -45808541	53992	Chr19:50446390 -50500381	53992	1.00000		
Annotal	tion Data	8							
Annotal Download Genome Download	tion Data Annotation Data e Workbenct	Files							
Annotat Download Genome Download	tion Data Annotation Data e Workbench Genome Workbe	Files							
Annotal Download Genome Download eature	tion Data Annotation Data Workbench Genome Workbe Src Intervals	E Files ench Files E Remap Intervals	Src Location	Src Length	Map Location	Map Leng	th Cover		
Annotal Download Genome Download eature #40463	tion Data Annotation Data Workbench Genome Workbe Src Intervals	ench Files	Src Location chr19:195999	Src Length 1	Map Location NG_028701.1:2	Map Leng	th Covera 1.0000		
Annotat Download Genome Download eature 40463	tion Data Annotation Data Workbench Genome Workbe Src Intervals 1 1	En Files ench Files Remap Intervals 1 1	Src Location chr19:195999 chr19:196020	Src Length 1 1	Map Location NG_028701.1:2 NG_028701.1:2	Map Leng 1845 1 1866 1	th Cover 1.000 1.000		
Annotal Download Oownload Download eature eature eature eature eature eature eature eature eature eature	tion Data Annotation Data Workbench Genome Workbe Genome Workbe Src Intervals	Remap Intervals	Src Location chr19:195999 chr19:196020 chr19:196021	Src Length 1 1 1	Map Location NG 028701.1:2 NG 028701.1:2 NG 028701.1:2	Map Leng 1845 1 1866 1 1867 1	th Cover 1.000 1.000		
Annotal Download Download eature s40463 s40463 s40463	tion Data Annotation Data e Workbench Genome Workbe Src Intervals 1 1 1 1 1	Ench Files Remap Intervals 1 1 1 1 1	Src Location chr19:195999 chr19:196020 chr19:196021 chr19:196043	Src Length 1 1 1 1	Map Location NG 028701.1:2 NG 028701.1:2 NG 028701.1:2 NG 028701.1:2	Map Leng 1845 1 1866 1 1867 1 1889 1	th Cover 1.000 1.000 1.000		
Annotal Download Download eature eature 940463 940463 940462 940462	tion Data Annotation Data e Workbench Genome Workbe Src Intervals 1 1 1 1 1 1 1	En Files ench Files Remap Intervals 1 1 1 1 1 1 1 1	Src Location chr19:195999 chr19:196020 chr19:196021 chr19:196043 chr19:196054	Src Length 1 1 1 1 1 1	Map Location NG 028701.1:2 NG 028701.1:2 NG 028701.1:2 NG 028701.1:2 NG 028701.1:2	Map Leng 845 1 866 1 1 867 1 1 889 1 900 1	th Cover 1.0000 1.0000 1.0000 1.0000 1.0000		
Annotal Download Genome Download eature \$40463 \$40463 \$40463 \$40462 \$40462 \$40462	tion Data Annotation Data Workbench Genome Workbe Src Intervals 1 1 1 1 1 1 1 1	Remap Intervals	Src Location chr19:195999 chr19:196020 chr19:196021 chr19:196043 chr19:196054 chr19:196072	Src Length 1 1 1 1 1 1 1	Map Location NG 028701.1:2 NG 028701.1:2 NG 028701.1:2 NG 028701.1:2 NG 028701.1:2 NG 028701.1:2 NG 028701.1:2	Map Leng 845 1 866 1 867 1 889 1 900 1 900 1 918 1	th Covera 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000		
Annotal Download Genome Download eature s40463 s40463 s40463 s40462 s40462 s40462	tion Data Annotation Data Workbench Genome Workbe Intervals	Remap Intervals	Src Location chr19:195999 chr19:196020 chr19:196021 chr19:196043 chr19:196054 chr19:196072 chr19:196079	Src Length 1 1 1 1 1 1 1 1 1 1	Map Location NG 028701.1:2 NG 028701.1:2 NG 028701.1:2 NG 028701.1:2 NG 028701.1:2 NG 028701.1:2 NG 028701.1:2 NG 028701.1:2	Map Leng 845 1 866 1 889 1 900 1 918 1 925 1	th Covera 1.000 1.000 1.000 1.000 1.000 1.000 1.000		
Annotal Download Download eature eature e40463 e40463 e40462 e40462 e40462 e40462 e40462 e40462 e40462 e40462 e40462 e40462 e40462	tion Data Annotation Data e Workbench Genome Workbe Intervals	Remap Intervals	Src Location chr19:195999 chr19:196020 chr19:196021 chr19:196043 chr19:196054 chr19:196072 chr19:196079 chr19:196107	Src Length 1 1 1 1 1 1 1 1 1 1 1	Map Location NG 028701.1:2 NG 028701.1:2 NG 028701.1:2 NG 028701.1:2 NG 028701.1:2 NG 028701.1:2 NG 028701.1:2 NG 028701.1:2 NG 028701.1:2 NG 028701.1:2	Map Leng 845 1 866 1 889 1 900 1 918 1 925 1 925 1	th Cover 1.0000 1.00		
Annotal Download Download eature s40463 s40463 s40462 s40462 s40462 s40462 s40462 s40462 s40462	tion Data Annotation Data e Workbench Genome Workbe Src Intervals 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	En Files ench Files Intervals	Src Location chr19:195999 chr19:196020 chr19:196021 chr19:196043 chr19:196054 chr19:196072 chr19:196079 chr19:196107 chr19:196158 -196162	Src Length 1 1 1 1 1 1 1 1 1 1 5	Map Location NG 028701.1:2 NG 028701.1:2 NG 028701.1:2 NG 028701.1:2 NG 028701.1:2 NG 028701.1:2 NG 028701.1:2 NG 028701.1:2 NG 028701.1:3 -3008	Map Leng 845 1 866 1 889 1 900 1 918 1 925 1 925 1 953 1 004 5	th Covera 1.0000 1.0		

Figure 4. Output from the Remap service. *Top panel*. Results of projecting gene locations from human build 36 onto build 37. The output provides downloadable results in the form of spreadsheets (Mapping Report and Annotation Data). Annotation data are also available in a format that can be loaded into NCBI's Genome Workbench, a standalone sequence analysis and annotation platform. *Bottom panel*. Mapping Report from Clinical Remap showing the projection of variations onto RefSeqGene records. The Clinical Remap Service also produces the Summary Data, Annotation Data and Genome Workbench files.

Please so	lect the org	anism and	assembly upon v	which your variants are	annotated.	ć.					
Select Or	ganism *		Select Assem	ibly • 😣							
Human			GRCh37 (hg19) GRCh37 (hg19) NC8I36 (hg18)	patch release							
Data											
Input form	hat: EH	cvs i	90								
Upload a	file:				Bro	wse)					
OR											
Paste dat	a here:	NC_00001 NC_00000 NC_00000 NC_00000	9:g.46201890C> 7:g.150883955T 9:g.45411941T> 7:g.113528849T	T A C C C C C C C C C C C C C C C C C C							
multiple	Click on the va	lue in the Sub	mitted Loc column to s	how it in Sequence Viewer.							Download Repo
into the area.	Submitted Id		Submitted Loc	Reported Allele	Cytoband	NCBI Id	Somatic Observed?	GMAF	Clinical Information	PubMed	Consequences
	NC_000019:g.4	45201890C>T	NC 000019.9 46201890-46201890	NC_000019.9/g.46201890C>T	19913.3	m151211223	no				non_synonymous_co
vomit) (NC_000007:g.1	150883955T>A	NC 000007.13 150883955-150883955	NC_000007.13/g.150583955T	A 7q36	rs151344617	no			10	non_synonymous_coo
	NC_000007;g.1	150883955T>A	NC 000007.13 150883955-150883955	NC_000007.13.g.1508839557	A 7q36	cs151344617	no			10	intron_variant
	NC_000019:g.4	15411941T>C	45411941-45411941	NC_000019.9:g.45411941T>C	19913.3	0429358	yes	C: 0.154		<u>97</u> 🖬	non_synonymous_coo
	NC_000007.g.1	113528849T>C	113528549-113528549 NC 000011 9		7q31.1						
	NC_000011:g.6	6828674A>G	66828674-66828674	•	11q13.2						
	D NC_000011	9.9: 46M_46M (1	(08b+) • Find on Sequer	ND#: 148.201.820 148.201.880	(¢¢) -	1993 🖬 🛛 146 2	🕕 • 💩	#6 201 910	146 201 9	× Tool	• Configure @ 3
	RECACTOR				0110000 CARCOOO	CACCIOC GIGGACCC	CCACCTO	ATCOACT INCOTOR	CTATATION	enenese ellenese	
	SNP		rs106227446		1223						
	Genes		, ,			-	,				
		<u>}</u>	19106227441	R5141204641 m		total ra	DIS QPCTL	9.9 (46,195,7	4146,207,248)		
	- 451 GA 6	20 020 020	40 000 WO 000 0	00 AND ONC 100 CHI 140	001 100 cgo	CAC (total le strand: Links i	ngth: 11,508 plus & Tools			ene age e	pe age ape ape ape
	001 004 0	40 000 CTO	110622744	te the one the fu the	001 100 100	View G View H CAC View H	enelD: <u>54814 (</u> GNC: <u>25752</u> PRD: <u>09612</u>	090711		CAO AGO C	te obe etc yte yte
				P141204641		004540-201	12.00			-0.1 1020	
	Clinical Var	rianta			_						

Figure 5. The Variation Reporter submission form and results. Top panel. Submission form maps locations of variations onto human genome builds. The input data in this case are variations in Human Genome Variation Society (HGVS) notation. Bottom panel. Results of mapping the variations onto build 37. The first four of the six variations map to NCBI Reference SNP locations. The corresponding identifiers and other information from dbSNP is shown for each of these. The location and genomic context for each mapped location is available for each of the mapped locations in the graphical sequence viewer. Clicking the linked location (red arrow) loads that marker and surrounding region in the sequence viewer.

The Variation Reporter, shown in Figure 5, takes a set of locations in a human genome assembly and identifies known human variations (NCBI Reference SNPs) at those positions. This service is particularly helpful for identifying experimentally or clinically determined variants. Like the Remap service, the Variation Reporter accepts a variety of genome annotation formats – HGVS, GVF and BED. The results provide the location of

the variants in the selected build and important information about any identified known variants including the dbSNP ID, the known allele, and, if available, clinical information, minor allele frequency, links to literature, and functional consequences. The results also provide the genomic context by displaying the mapped locations in the graphic sequence viewer (Figure 5, bottom panel). The Remap Service and the Variation Reporter are useful for interconverting annotations between genome builds and mapping and identifying experimentally determined variants.

NCBI Aspera Download Site Available for NCBI Databases and Tools

An Aspera protocol download site is available as an alternative to FTP for all NCBI downloads. The Aspera protocol provides a much faster transfer rate and is most important for downloading very large data sets such as those from next-generation sequencing studies, but can be used to improve download performance for any public NCBI data files or software packages. The Aspera protocol site requires the free AsperaConnect client application available from Aspera Connect. The Aspera Transfer Guide, available on the NCBI Bookshelf, provides additional information on using the fast download site.

1000 Genomes Project Data Now on Amazon Cloud Service

As announced in the recent NIH press release, data from the 1000 Genomes project - the world's largest set of data on human genetic variation produced by the international 1000 Genomes Project — are now publicly available on the Amazon Web Services (AWS) cloud. 1000 genomes data may also be downloaded from the NCBI though FTP or through the Aspera protocol site.

Microbial Genomes Update

One hundred ninety nine finished microbial (archaeal and bacterial) genomes were released from November 2011 through March 2012. The original sequence data files submitted to the International Sequence Database Collaboration (INSDC) are available in the Bacteria directory in the genomes area of the GenBank FTP site. RefSeq provisional versions were released for a selected set of 118 of the complete INSDC microbial genomes during the same period. These are available from the /genomes/Bacteria directory on the FTP site.

In addition, data from 1,135 microbial whole genome-shotgun (WGS) sequencing projects were added to the INSDC during this period. The original submitted files are available in the Bacteria_DRAFT directory in the GenBank genomes area. RefSeq provisional versions of 210 WGS microbial projects were released in the /genomes/Bacteria_DRAFT area of the FTP site.

All GenBank and RefSeq microbial genomes are incorporated in the NCBI integrated Entrez search and retrieval system and the BLAST sequence similarity search service.

NCBI Articles in Nucleic Acids Research Database Issue

The Nucleic Acids Research 2011 Database Issue contains 10 articles about NCBI resources, tools, and databases including BioAssay, SRA, GEO, BioProject / BioSample, Taxonomy Epigenomics, MMDB (Structure), RefSeq and GenBank. Free full-text articles from the database issue are available from PubMed Central and the publisher's site and are linked to the summaries and abstracts in PubMed.

GenBank News

GenBank release 189 is available through Entrez, BLAST and from the GenBank FTP area. The current release incorporates data available as of April 15, 2011 and, with the whole-genome shotgun portion, contains 411,959,832,946 bases from 232,729,719 sequence records. Release notes describe the current state of data and upcoming changes. The GenBank page provides more information on the database content and scope as well as submission information.

RefSeq News

RefSeq Release 52

RefSeq Release 52 is available through Entrez, BLAST, and from the RefSeq FTP area. The current release includes 20.2 million Reference Sequence records from 16,923 different species or strains. The RefSeq release notes provide more detailed information.

RefSeq Genome Annotation Files in GFF3 Format

NCBI now offers Reference Sequence (RefSeq) genome annotation files in the latest Generic Feature Format (GFF3) specification (1.20). RefSeq genome data can be downloaded from the genomes area of the NCBI FTP site. GFF3 files are in the GFF directory within each organism directory. Currently GFF3 files are available for the NCBI annotations of the latest assemblies for human, cow, dog, chicken, and many others.

Keeping Up with NCBI

Seventeen topic-specific mailing lists are available that provide email announcements about changes and updates to NCBI resources including dbGaP, BLAST, GenBank, and Sequin. The various lists are described on the Announcement List summary page. Subscribe to the NCBI Announce list to receive updates on the NCBI News.

Twenty-five RSS feeds are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce.

NCBI's Facebook page and Twitter feed also provide updates on NCBI resources.

Send comments and questions about NCBI resources to info@ncbi.nlm.nih.gov, or call 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.