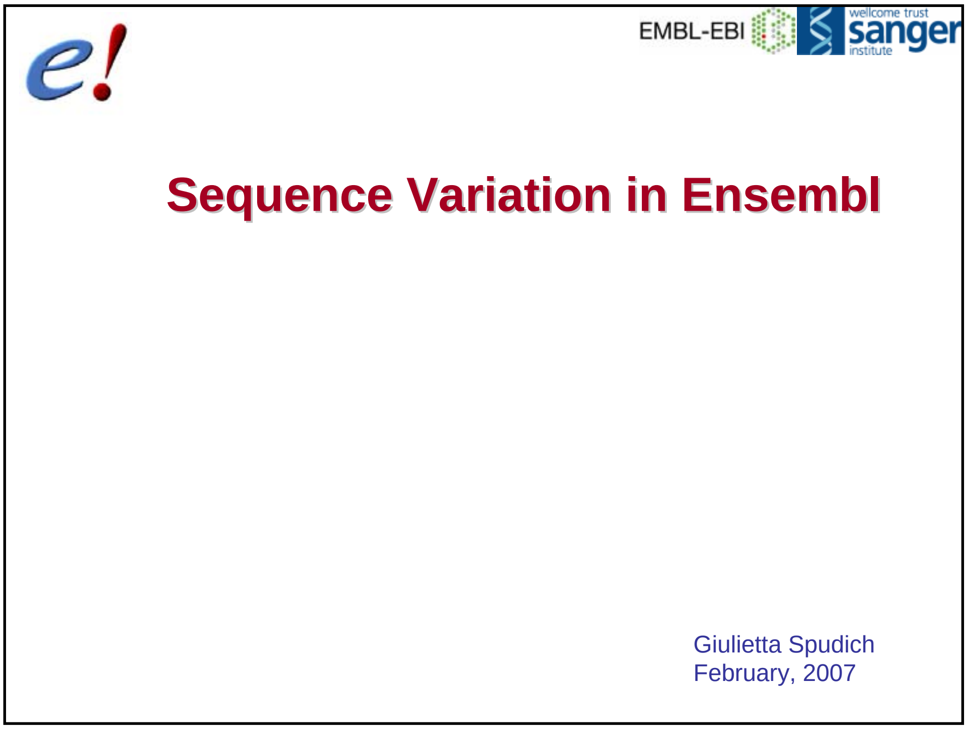
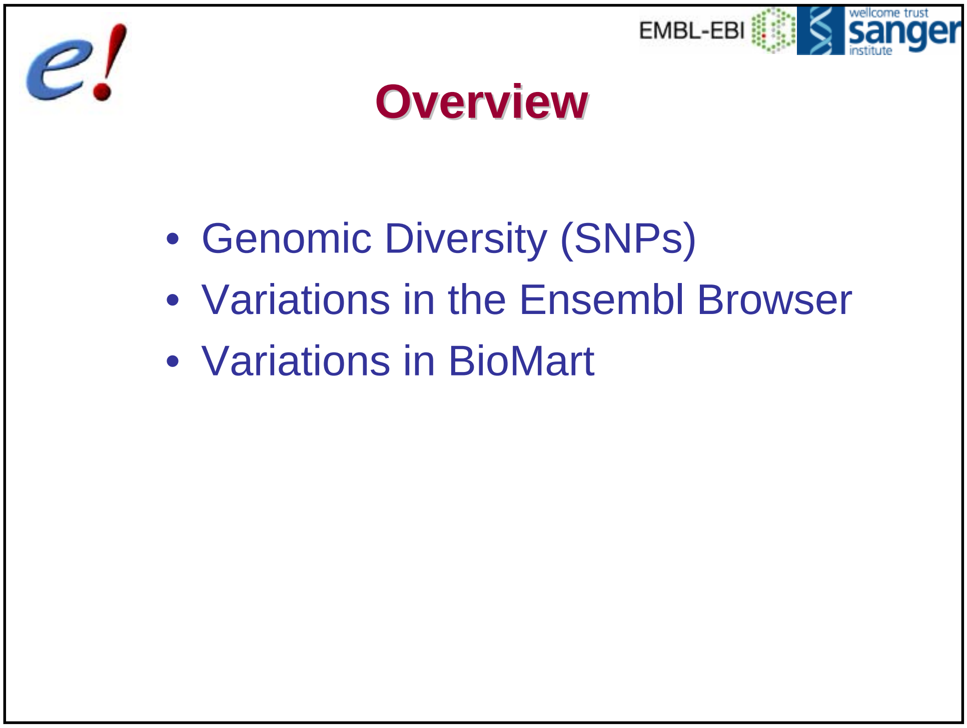


Sequence Variation in Ensembl

Giulietta Spudich
February, 2007



Overview

- Genomic Diversity (SNPs)
- Variations in the Ensembl Browser
- Variations in BioMart

Genomic Diversity

Mutations:

base pair substitutions
insertion/deletion (frameshifts)

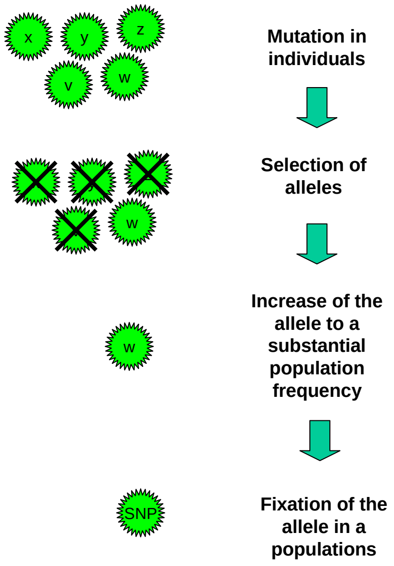
1 in every 300 bp (*human*)

~3 billion base pairs in mouse genome!

Single nucleotide polymorphisms (SNPs)

- **Polymorphism: a DNA variation in which each possible sequence is present in at least 1% of the population**
- **Most polymorphisms (~90%) take the form of SNPs: variations that involve just one nucleotide**

Origin of SNPs



Mutation in individuals

Selection of alleles

Increase of the allele to a substantial population frequency

Fixation of the allele in a populations

Adapted from Bioinformatics for Geneticists, Eds Barnes and Gray

Functional Consequences

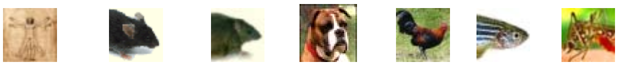
Type	Consequence
SNPs in coding area that alter aa sequence	Cause of most monogenic disorders, e.g: Cystic fibrosis (CFTR) Hemophilia (F8)
SNPs in coding areas that don’t alter aa sequence	May affect splicing
SNPs in promoter or regulatory regions	May affect the level, location or timing of gene expression
SNPs in other regions	No direct known impact on phenotype Useful as markers

Studying variation – why?

- Determine disease risk
- Predict responses to environmental changes and drugs (pharmacogenomics)

- Biological markers
- Forensics
- Evolution
- Laboratory: hybridisation studies, marker-assisted breeding

SNPs in Ensembl - Species



- Human
- Mouse
- Rat
- Dog

- Chicken
- Zebrafish
- Mosquito

EMBL-EBI wellcome trust sanger institute

SNPs in Ensembl

- Most SNPs imported from dbSNP (rs……):
  - Imported data: alleles, frequencies, flanking sequence….
  - Calculated data: synonymous status, peptide shift, SNP position….
- For mouse also:
  - Sanger

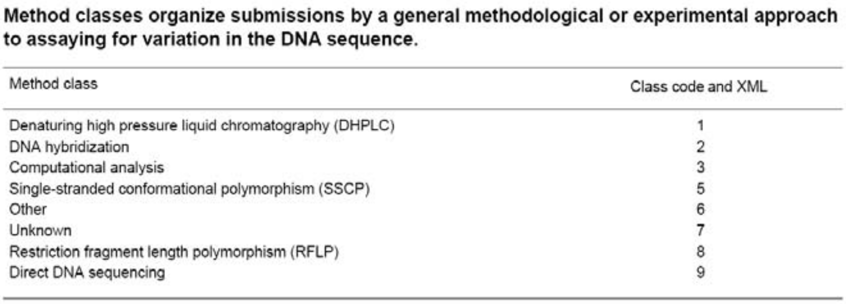
EMBL-EBI wellcome trust sanger institute

dbSNP (NCBI)

- Main database of SNPs (and short polymorphisms: in-dels)
- 6,491,554 rs (reference SNPs) in mouse (*11,961,761 in human*).
- 4,990,170 validated in mouse
- (*5,646,244 in human*).
- http://www.ncbi.nlm.nih.gov/SNP

ENTREZ SNP Single Nucleotide Polymorphism

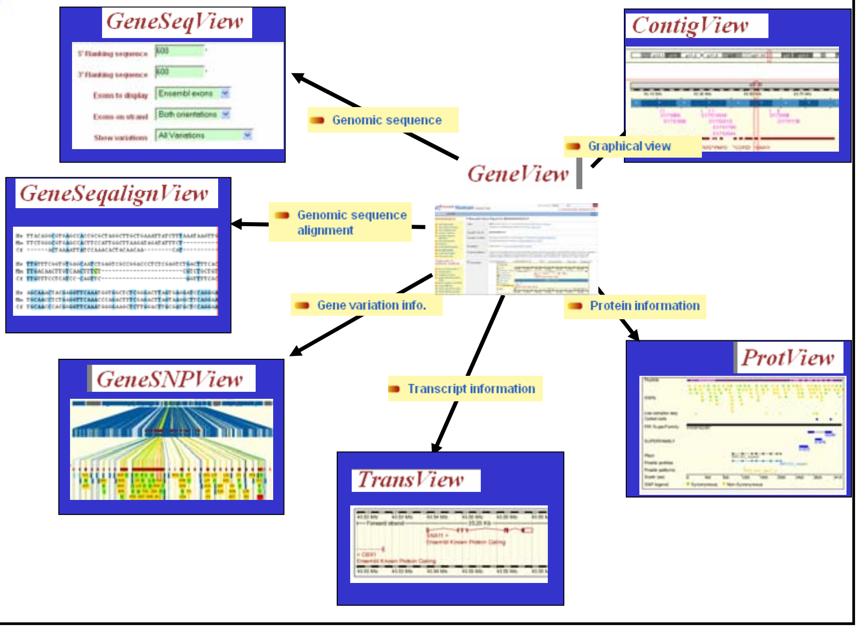
dbSNP - Validation



Method classes organize submissions by a general methodological or experimental approach to assaying for variation in the DNA sequence.

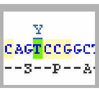
Method class	Class code and XML
Denaturing high pressure liquid chromatography (DHPLC)	1
DNA hybridization	2
Computational analysis	3
Single-stranded conformational polymorphism (SSCP)	5
Other	6
Unknown	7
Restriction fragment length polymorphism (RFLP)	8
Direct DNA sequencing	9

SNPs in Ensembl



SNPs in Ensembl (Views)

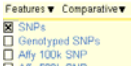
Click on a SNP



```
   Y
CAGTCCGGC
--S--P--A
```

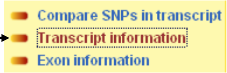
- *TransView*
- *ProtView*
- *MapView*
- *ContigView*
- *GeneSNPView*
- *SNPView*

Features: SNPs in roll-down menus

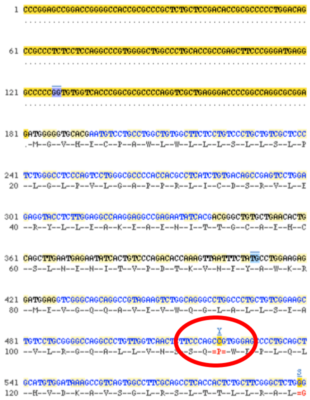


- ☒ SNPs
- ☐ Genotyped SNPs
- ☐ Affy 100k SNP

TransView



SNPs in sequence

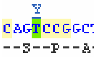


```
  1 CCCGGAGCCGGACCGGGGCCACCGCGCCCGCTCTGCTCCGACACCGCGCCCCCTGGACAG
    ............................................................

 61 CCGCCCTCTCCTCCAGGCCCGTGGGGCTGGCCCTGCACCGCCGAGCTTCCCGGGATGAGG
    ............................................................

121 GCCCCCGGTGTGGTCACCCGGCGCGCCCCAGGTCGCTGAGGGACCCCGGCCAGGCGCGGA
    ............................................................

181 GATGGGGGTGCACGAATGTCCTGCCTGGCTGTGGCTTCTCCTGTCCCTGCTGTCGCTCCC
    .-M--G--V--H--I--C--P--A--W--L--W--L--L--L--S--L--L--S--L--P
241 TCTGGGCCTCCCAGTCCTGGGCGCCCCACCACGCCTCATCTGTGACAGCCGAGTCCTGGA
 20 --L--G--L--P--V--L--G--A--P--P--R--L--I--C--D--S--R--V--L--E
301 GAGGTACCTCTTGGAGGCCAAGGAGGCCGAGAATATCACGACGGGCTGTGCTGAACACTG
 40 --R--Y--L--L--E--A--K--E--A--E--N--I--T--T--G--C--A--E--H--C
361 CAGCTTGAATGAGAATATCACTGTCCCAGACACCAAAGTTAATTTCTATGCCTGGAAGAG
 60 --S--L--N--E--N--I--T--V--P--D--T--K--V--N--F--Y--A--W--K--R
421 GATGGAGGTCGGGCAGCAGGCCGTAGAAGTCTGGCAGGGCCTGGCCCTGCTGTCGGAAGC
 80 --M--E--V--G--Q--Q--A--V--E--V--W--Q--G--L--A--L--L--S--E--A
                                            Y
481 TGTCCTGCGGGGCCAGGCCCTGTTGGTCAACTCTTCCCAGCCGTGGGAGCCCCTGCAGCT
100 --V--L--R--G--Q--A--L--L--V--N--S--S--Q-=P=-W--E--P--L--Q--L
                                                           S
541 GCATGTGGATAAAGCCGTCAGTGGCCTTCGCAGCCTCACCACTCTGCTTCGGGCTCTGGG
120 --H--V--D--K--A--V--S--G--L--R--S--L--T--T--L--L--R--A--L-=G
```

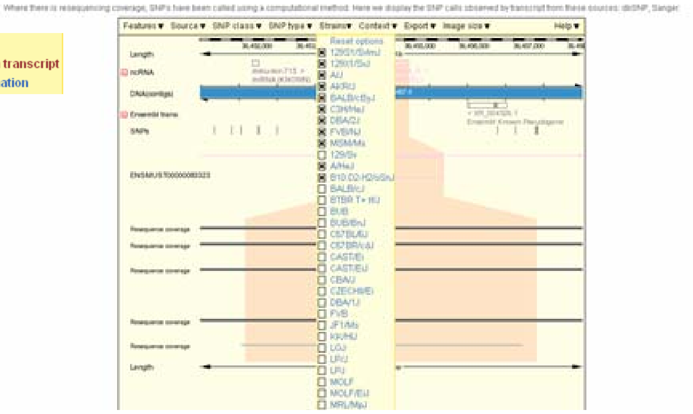


```
   Y
CAGTCCGGC
--S--P--A
```

EMBL-EBI wellcome trust sanger institute

TrancriptSNPView

- ID history
- Compare SNPs in transcript
- Transcript information



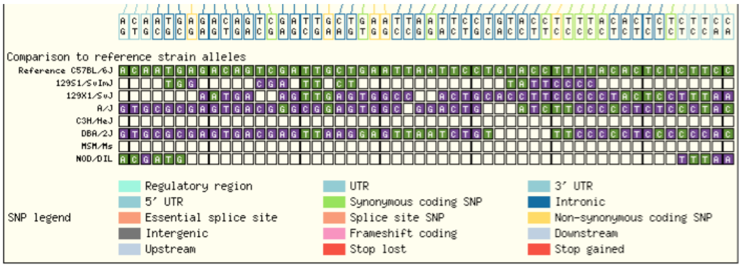
Where there is resequencing coverage, SNPs have been called using a computational method. Here we display the SNP calls observed by transcript from these sources: dbSNP, Sanger

Reset options
☒ 129S1/SvImJ
☒ 129X1/SvJ
☒ AJ
☒ AKR/J
☒ BALB/cByJ
☒ C3H/HeJ
☒ DBA/2J
☒ FVB/NJ
☒ MSM/Ms
☐ 129/Sv
☒ A/HeJ
☒ B10.D2-H2/oSnJ
☐ BALB/cJ
☐ BTBR T+ tf/J
☐ BUB
☐ BUB/BnJ
☐ C57BL/6J
☐ C57BR/cdJ
☐ CAST/Ei
☐ CAST/EiJ
☐ CBA/J
☐ CZECHII/Ei
☐ DBA/1J
☐ FVB
☐ JF1/Ms
☐ KK/HIJ
☐ LG/J
☐ LP/J
☐ LP/J
☐ MOLF
☐ MOLF/EiJ
☐ MRL/MpJ

SNP in different strains
Alleles, type (i.e. placement in relation to a gene)

EMBL-EBI wellcome trust sanger institute

TrancriptSNPView



Comparison to reference strain alleles

	A	C	A	A	T	G	A	G	A	C	A	G	T	C	G	A	T	T	G	C	T	G	A	A	T	T	A
	G	T	G	C	G	C	G	A	G	T	G	A	C	G	A	G	C	G	A	A	G	T	G	G	C	C	G
Reference C57BL/6J	A	C	A	A	T	G	A	G	A	C	A	G	T	C	G	A	T	T	G	C	T	G	A	A	T	T	A
129S1/SvImJ					T	G	G						C	G	A		T	T		C	T						
129X1/SvJ								A	A	T	G	A			A	G	T	T	G	A	G	T	G	G	C	C	
A/J	G	T	G	C	G	C	G	A	G	T	G	A	C	G	G	G	C	G	G	A	G	T	G	G	C		G
C3H/HeJ																											
DBA/2J	G	T	G	C	G	C	G	A	G	T	G	A	C	G	A	G	T	T	A	A	G	G	A	G	T	T	A
MSM/Ms																											
NOD/DIL	A	C	G	A	T	G																					

	A	T	T	C	C	T	G	T	A	C	C	T	T	T	T	A	C	A	C	T	C	T	C	T	T	C	C
	G	A	C	T	G	C	A	C	C	T	T	C	C	C	C	C	T	C	T	C	T	C	T	C	C	A	A
Reference C57BL/6J	A	T	T	C	C	T	G	T	A	C	C	T	T	T	T	A	C	A	C	T	C	T	C	T	T	C	C
129S1/SvImJ								T	A	T	T	C	C	C	C												
129X1/SvJ		A	C	T	G	C	A	C	C	T	T	C	C	C	C	C	T	A	C	T	C	C	T	T	T	A	A
A/J	G	A	C	T	G				A	T	C	T	T	C	C	C	C	C	T	C	T	C	C	C	T	A	C
C3H/HeJ																											
DBA/2J	A	T	C	T	G	T						T	T	C	C	C	C	C	T	C	C	C	C	C	C	A	C
MSM/Ms																											
NOD/DIL																						T	T	T	A	A

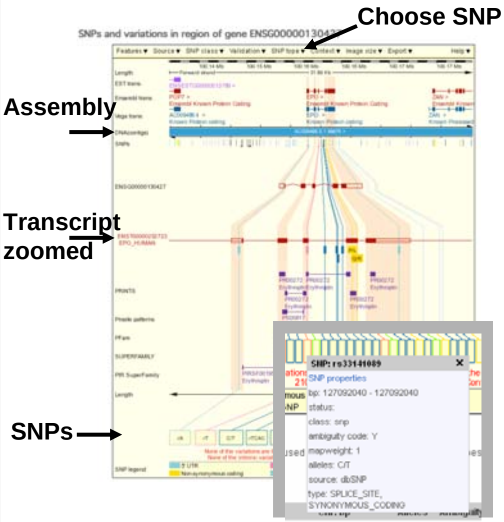
SNP legend			
	Regulatory region	UTR	3′ UTR
	5′ UTR	Synonymous coding SNP	Intronic
SNP legend	Essential splice site	Splice site SNP	Non-synonymous coding SNP
	Intergenic	Frameshift coding	Downstream
	Upstream	Stop lost	Stop gained

SARA = Same As Reference Assembly (C57BL/6J)

GeneSNPView

- Gene tree info
- Gene variation info.
- ID history

What SNPs does my gene contain?


Choose SNP type
Assembly
Transcript zoomed
SNPs
SNPs and variations in region of gene ENSG00000130427
Features ▾ Source ▾ SNP class ▾ Validation ▾ SNP type ▾ Context ▾ Image size ▾ Export ▾ Help ▾

SNP: rs33141089

SNP properties

bp: 127092040 - 127092040

status:

class: snp

ambiguity code: Y

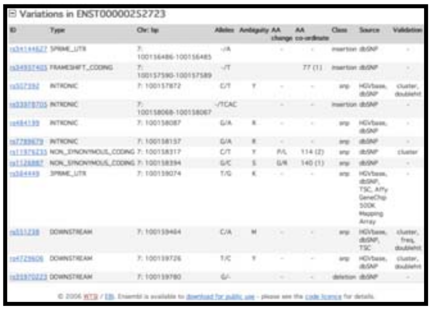
mapweight: 1

alleles: C/T

source: dbSNP

type: SPLICE_SITE, SYNONYMOUS_CODING

Table of Variations

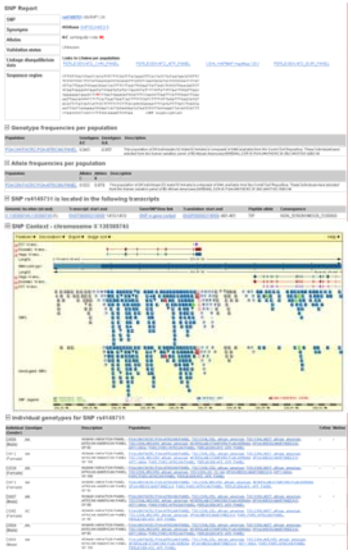


⊟ Variations in ENST00000252723

ID	Type	Chr: bp	Alleles	Ambiguity	AA change	AA co-ordinate	Class	Source	Validation
rs34144627	5PRIME_UTR	7: 100156486-100156485	-/A		-	-	insertion	dbSNP	-
rs34937405	FRAMESHIFT_CODING	7: 100157590-100157589	-/T			77 (1)	insertion	dbSNP	-
rs557392	INTRONIC	7: 100157872	C/T	Y	-	-	snp	HGVbase, dbSNP	cluster, doublehit
rs33978705	INTRONIC	7: 100158068-100158067	-/TCAC		-	-	insertion	dbSNP	-
rs484139	INTRONIC	7: 100158087	G/A	R	-	-	snp	HGVbase, dbSNP	-
rs7789679	INTRONIC	7: 100158157	G/A	R	-	-	snp	dbSNP	-
rs11976235	NON_SYNONYMOUS_CODING	7: 100158317	C/T	Y	P/L	114 (2)	snp	dbSNP	cluster
rs1126887	NON_SYNONYMOUS_CODING	7: 100158394	G/C	S	G/R	140 (1)	snp	dbSNP	-
rs584449	3PRIME_UTR	7: 100159074	T/G	K	-	-	snp	HGVbase, dbSNP, TSC, Affy GeneChip 500K Mapping Array	-
rs551238	DOWNSTREAM	7: 100159484	C/A	M	-	-	snp	HGVbase, dbSNP, TSC	cluster, freq, doublehit
rs4729606	DOWNSTREAM	7: 100159726	T/C	Y	-	-	snp	HGVbase, dbSNP	cluster, doublehit
rs33970223	DOWNSTREAM	7: 100159780	G/-		-	-	deletion	dbSNP	-

© 2006 WTSI / EBI. Ensembl is available to download for public use - please see the code licence for details.

SNPView



Info about one specific SNP:

- **SNP Report:**

  Imported from… dbSNP, TSC, HGVbase, Affy Chip Data

  Validation

  Genotype and allele frequencies per population

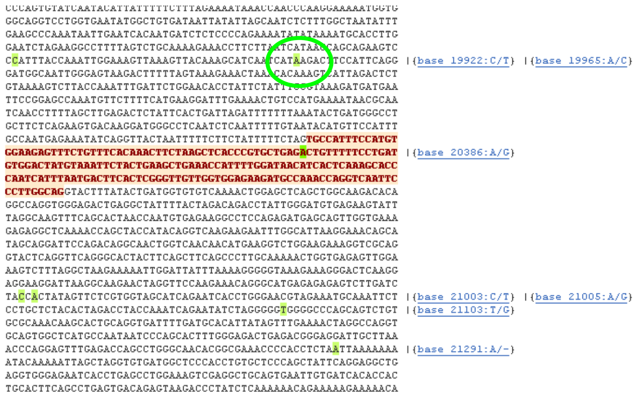
  Location in transcripts

  Type:coding/noncoding

GeneSeqView

- Gene information
- Genomic sequence
- Genomic sequence alignment

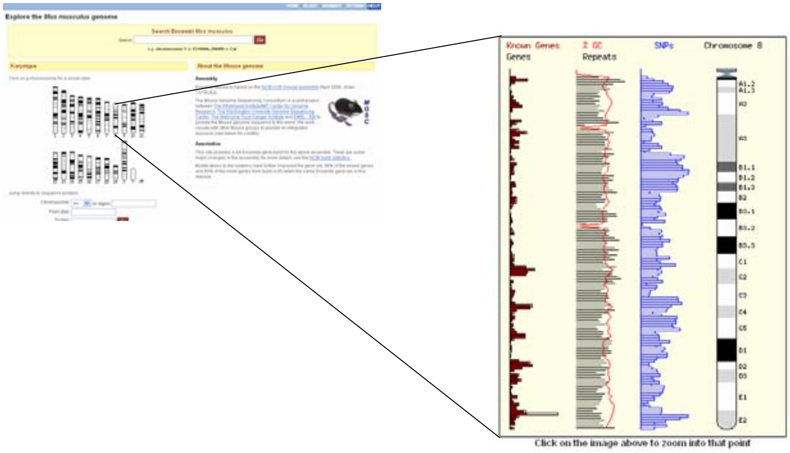
SNPs in genomic sequence



```
CCCAGTGTATCAATACATTATTTTTCTTTAGAAAATAAACCAACCCAAGGAAAAATGGTG
GGCAGGTCCTGGTGAATATGGCTGTGATAATTATATTAGCAATCTCTTTGGCTAATATTT
GAAGCCCAAATAATTGAATCACAATGATCTCTCCCCAGAAAATATATAAAATGCACCTTG
GAATCTAGAAGGCCTTTTAGTCTGCAAAAGAAACCTTCTTAATCATAACCAGCAGAAGTC
CCATTTACCAAATTGGAAAGTTAAAGTTACAAAGCATCAATCATAAGACTTCCATTCAGG |{base 19922:C/T} |{base 19965:A/C}
GATGGCAATTGGGAGTAAGACTTTTTAGTAAAGAAACTAAACACAAAGTCATTAGACTCT
GTAAAAGTCTTACCAAATTTGATTCTGGAACACCTATTCTATTTCGGTAAAGATGATGAA
TTCCGGAGCCAAATGTTCTTTTCATGAAGGATTTGAAAACTGTCCATGAAAATAACGCAA
TCAACCTTTTAGCTTGAGACTCTATTCACTGATTAGATTTTTTTAAATACTGATGGGCCT
GCTTCTCAGAAGTGACAAGGATGGGCCTCAATCTCAATTTTTGTAATACATGTTCCATTT
GCCAATGAGAAATATCAGGTTACTAATTTTTCTTCTATTTTTCTAGTGCCATTTCCATGT
GGAAGAGTTTCTGTTTCACAAACTTCTAAGCTCACCCGTGCTGAGACTGTTTTTCCTGAT |{base 20386:A/G}
GTGGACTATGTAAATTCTACTGAAGCTGAAACCATTTTGGATAACATCACTCAAAGCACC
CAATCATTTAATGACTTCACTCGGGTTGTTGGTGGAGAAGATGCCAAACCAGGTCAATTC
CCTTGGCAGGTACTTTATACTGATGGTGTGTCAAAACTGGAGCTCAGCTGGCAAGACACA
GGCCAGGTGGGAGACTGAGGCTATTTTACTAGACAGACCTATTGGGATGTGAGAAGTATT
TAGGCAAGTTTCAGCACTAACCAATGTGAGAAGGCCTCCAGAGATGAGCAGTTGGTGAAA
GAGAGGCTCAAAACCAGCTACCATACAGGTCAAGAAGAATTTGGCATTAAGGAAACAGCA
TAGCAGGATTCCAGACAGGCAACTGGTCAACAACATGAAGGTCTGGAAGAAAGGTCGCAG
GTACTCAGGTTCAGGGCACTACTTCAGCTTCAGCCCTTGCAAAAACTGGTGAGAGTTGGA
AAGTCTTTAGGCTAAGAAAAATTGGATTATTTAAAAGGGGGTAAAGAAAGGGACTCAAGG
AGGAAGGATTAAGGCAAGAACTAGGTTCCAAGAAACAGGGCATGAGAGAGAGTCTTGATC
TACCACTATAGTTCTCGTGGTAGCATCAGAATCACCTGGGAACGTAGAAATGCAAATTCT |{base 21003:C/T} |{base 21005:A/G}
CCTGCTCTACACTAGACCTACCAAATCAGAATATCTAGGGGGTGGGGCCCAGCAGTCTGT |{base 21103:T/G}
GCGCAAACAAGCACTGCAGGTGATTTTGATGCACATTATAGTTTGAAAACTAGGCCAGGT
GCAGTGGCTCATGCCAATAATCCCAGCACTTTGGGAGACTGAGACGGGAGGATTGCTTAA
ACCCAGGAGTTTGAGACCAGCCTGGGCAACACGGCGAAACCCCACCTCTAATTAAAAAAA |{base 21291:A/-}
ATACAAAAATTAGCTAGGTGTGATGGCTCCCACCTGTGCTCCCAGCTATTCAGGAGGCTG
AGGTGGGAGAATCACCTGAGCCTGGAAAGTCGAGGCTGCAGTGAATTGTGATCACACCAC
TGCACTTCAGCCTGAGTGACAGAGTAAGACCCTATCTCAAAAAACAGAAAAAGAAAAACA
```

MapView

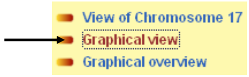
SNP density on a chromosome



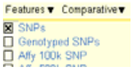
Click on the image above to zoom into that point

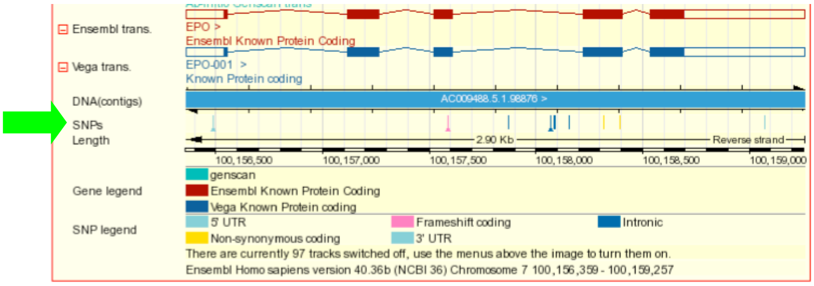
Example: Mouse chromosome 8

ContigView



SNPs along a chromosomal region

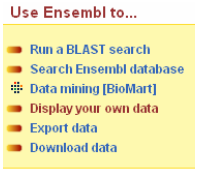


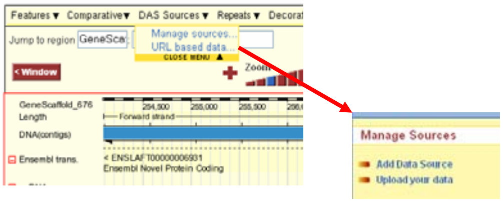


There are currently 97 tracks switched off, use the menus above the image to turn them on.

Ensembl Homo sapiens version 40.36b (NCBI 36) Chromosome 7 100,156,359 - 100,159,257

Your own SNP data



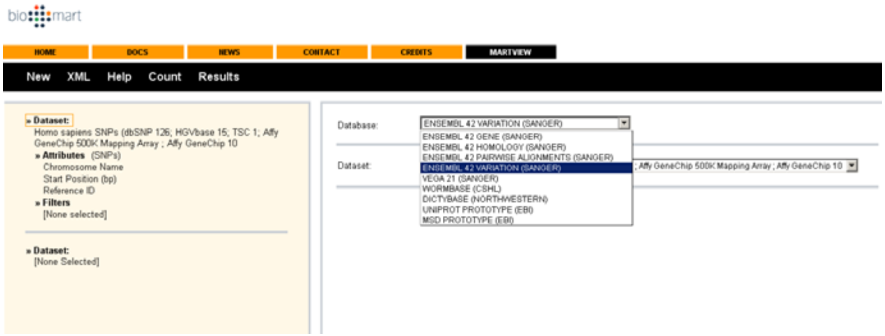


Display your own data in:
ContigView, CytoView, GeneView, TransView, ProtView and KaryoView

OR Submit to dbSNP
http://www.ncbi.nih.gov/SNP/

SNPs in BioMart

Obtain a table of genes corresponding to SNPs



» Dataset:
Homo sapiens SNPs (dbSNP 126; HGVbase 15; TSC 1; Affy GeneChip 500K Mapping Array ; Affy GeneChip 10
» Attributes (SNPs)
Chromosome Name
Start Position (bp)
Reference ID
» Filters
[None selected]

» Dataset:
[None Selected]

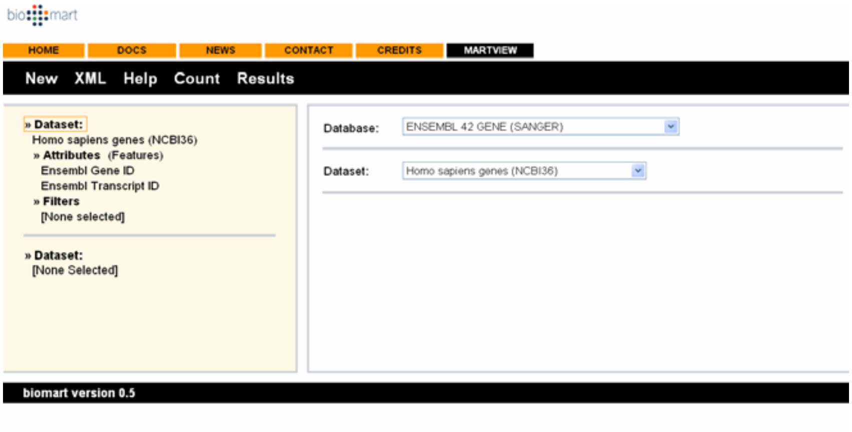
Database: ENSEMBL 42 VARIATION (SANGER)

- ENSEMBL 42 GENE (SANGER)
- ENSEMBL 42 HOMOLOGY (SANGER)
- ENSEMBL 42 PAIRWISE ALIGNMENTS (SANGER)
- ENSEMBL 42 VARIATION (SANGER)
- VEGA 21 (SANGER)
- WORMBASE (CSHL)
- DICTYBASE (NORTHWESTERN)
- UNIPROT PROTOTYPE (EBI)
- MSD PROTOTYPE (EBI)

Dataset: ; Affy GeneChip 500K Mapping Array ; Affy GeneChip 10

SNPs in BioMart

Obtain a table of SNPs for Ensembl genes



» Dataset:
Homo sapiens genes (NCBI36)
» Attributes (Features)
Ensembl Gene ID
Ensembl Transcript ID
» Filters
[None selected]

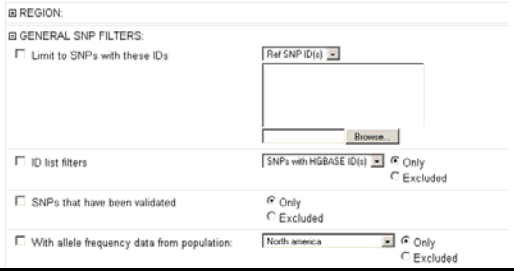
» Dataset:
[None Selected]

Database: ENSEMBL 42 GENE (SANGER)

Dataset: Homo sapiens genes (NCBI36)

biomart version 0.5

SNPs in BioMart



⊞ REGION:

⊟ GENERAL SNP FILTERS:

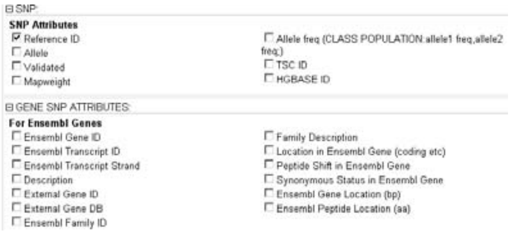
☐ Limit to SNPs with these IDs — Ref SNP ID(s)

☐ ID list filters — SNPs with HGBASE ID(s) ● Only ○ Excluded

☐ SNPs that have been validated — ● Only ○ Excluded

☐ With allele frequency data from population: North america ● Only ○ Excluded

SNP FILTER options



⊟ SNP:

SNP Attributes

☑ Reference ID
☐ Allele
☐ Validated
☐ Mapweight
☐ Allele freq (CLASS POPULATION:allele1 freq,allele2 freq;)
☐ TSC ID
☐ HGBASE ID

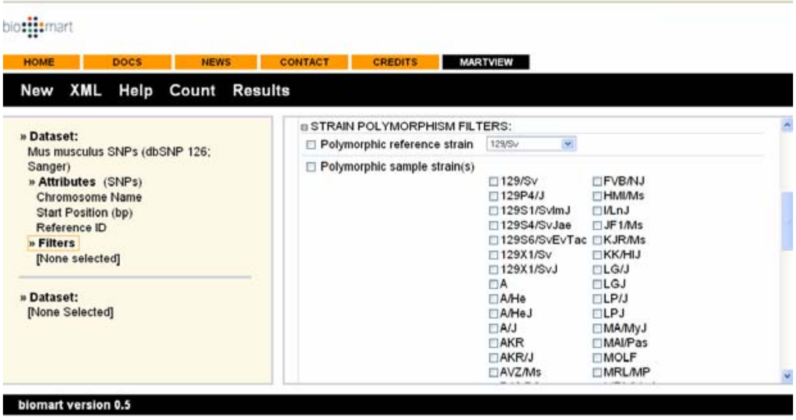
⊟ GENE SNP ATTRIBUTES:

For Ensembl Genes

☐ Ensembl Gene ID
☐ Ensembl Transcript ID
☐ Ensembl Transcript Strand
☐ Description
☐ External Gene ID
☐ External Gene DB
☐ Ensembl Family ID
☐ Family Description
☐ Location in Ensembl Gene (coding etc)
☐ Peptide Shift in Ensembl Gene
☐ Synonymous Status in Ensembl Gene
☐ Ensembl Gene Location (bp)
☐ Ensembl Peptide Location (aa)

SNP Attribute options

Strain-specific SNPs in BioMart



» **Dataset:**
Mus musculus SNPs (dbSNP 126; Sanger)
» **Attributes** (SNPs)
Chromosome Name
Start Position (bp)
Reference ID
» **Filters**
[None selected]

» **Dataset:**
[None Selected]

⊟ STRAIN POLYMORPHISM FILTERS:

☐ Polymorphic reference strain — 129/Sv

☐ Polymorphic sample strain(s)

☐ 129/Sv
☐ 129P4/J
☐ 129S1/SvImJ
☐ 129S4/SvJae
☐ 129S6/SvEvTac
☐ 129X1/Sv
☐ 129X1/SvJ
☐ A
☐ A/He
☐ A/HeJ
☐ A/J
☐ AKR
☐ AKR/J
☐ AVZ/Ms
☐ FVB/NJ
☐ HMI/Ms
☐ I/LnJ
☐ JF1/Ms
☐ KJR/Ms
☐ KK/HIJ
☐ LG/J
☐ LGJ
☐ LP/J
☐ LPJ
☐ MA/MyJ
☐ MAI/Pas
☐ MOLF
☐ MRL/MP

biomart version 0.5

EMBL-EBI wellcome trust sanger institute

Summary

- Genomic Diversity (SNPs)
- Variations in the Ensembl Browser
- Variations in BioMart

EMBL-EBI wellcome trust sanger institute

Ensembl Team

Leaders **Ewan Birney (EBI), Tim Hubbard (Sanger Institute)**

Glenn Proctor, Ian Longden, Patrick Meidl, Andreas Kähäri

Arek Kasprzyk, Syed Haider, Richard Holland, Damian Smedley, Benoît Ballester

Eugene Kulesha

Xosé M Fernández, Bert Overduin, Michael Schuster, Giulietta Spudich

James Smith, Fiona Cunningham, Anne Parker, Bethan Pritchard, Stephen Rice, Steve Trevanion, Matt Wood

Abel Ureta-Vidal, Benoit Ballester, Kathryn Beal, Stephen Fitzgerald, Javier Herrero Sánchez, Albert Vilella

Val Curwen, Steve Searle, Bronwen Aken, Julio Banet, Laura Clarke, Sarah Dyer, Kevin Howe, Felix Kokocinski, Jan-Hinnerck Vogel, Simon White

Paul Flicek, Yuan Chen, Stefan Gräf, Nathan Johnson, Daniel Rios

Martin Hammond, Dan Lawson, Karyn Megy