Variazioni nel Ensembl

Giulietta Spudich
Feb, 2007

Ordine

- Variabilità della genoma di una specie *(SNPs: single nucleotide polymorphisms)*
- Haplotypes (progetto HapMap)
- Linkage Disequilibrium
- SNPs nel ‘strains’ di *Mus musculus*

Varietà nella Genoma Umana

Mutazioni genetiche

(sostituzioni (SNPs) e insertion/deletions (InDels)

sono le cause di variabilità

nel

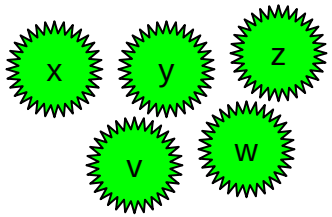
uno di ogni 300 bp (umano)

Ci sono 3 billioni ‘base pairs’ nella genoma umana.

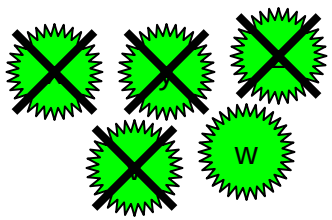
Single nucleotide polymorphisms (SNPs)

- Estimato 10 millioni SNPs nel genoma umana.
- Polymorphism: una variazione che esista in almeno 1% della popolazione.
- La maggioranza (90%) dei ‘polymorphisms’ sono SNPs: cambiano un ‘nucleotide’.

La nascita di SNPs



Mutazioni negli individui



Selezione di ‘alleles’

La frequenza di un ‘allele’ cresce nella populazione

L’allele diventa ‘fissato’ nella populazione

Adapted from Bioinformatics for Geneticists, Eds Barnes and Gray

Le conseguenze

Tipo	Conseguenze
SNPs in una regione 'coding' puo cambiare la sequenza della proteina. 'non-synonymous'	La causa della maggioranza di malattia monogenica, e.g: Cystic fibrosis (CFTR) emofilia (F8)
SNPs in una regione 'coding' che non cambiano la sequenza della proteina. 'synonymous'	Puo influire lo 'splicing'
SNPs nel 'promoter' o regioni di regolazione	Puo cambiare la quantità, locazione o tempismo del 'gene expression'
SNPs nelle altre regioni	Non sappiamo un impatto Sono utili come 'markers'

Lo studio delle variazioni- perchè?

- Prevedere il rischio di una malattia
- Prevedere le reazioni ai cambiamenti di ambiente o ai medicinali (pharmacogenomics)

- Biological markers
- Medicina forense/ criminali
- Evoluzione
- Nel lab: per progettare ‘primers’, studi del hybridization

Applicazioni Practiche

Ferro eccessivo nel sangue

Mutazione nel gene HFE

0.5% della popolazione del USA ha due mutated ‘alleles’.

Medicina ‘personalized’

Evoluzione

SNPs – tipi nel Ensembl

- **Coding SNPs:**
  Non-synonymous (cambiano la seq aa)
  Synonymous (non cambiano la seq aa)

- **Non-coding SNPs:**
  UTR (Untranslated Regions)
  Regulatory Region
  Intronic

- **Intergenic (5Kb)** (tra I geni)

dbSNP (NCBI)

- Database di SNPs (e polymorphisms come in-dels)
- 27.8 Millioni sottomissioni
  (5.6 Millioni ‘reference SNPs’ (rs)).
- http://www.ncbi.nlm.nih.gov/SNP

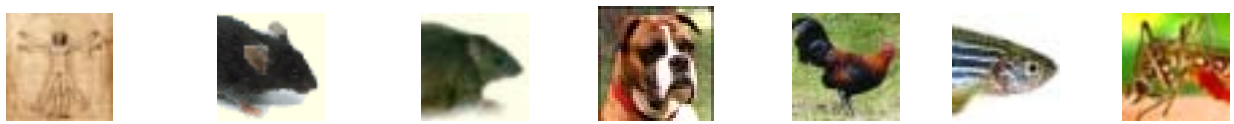
SNPs in Ensembl

- **La maggioranza di SNPs sono di dbSNP (ID: rs……):**

- Per umano anche dal:
  - HGVbase
  - TSC
  - Affy GeneChip Mapping Array (100K, 500K)

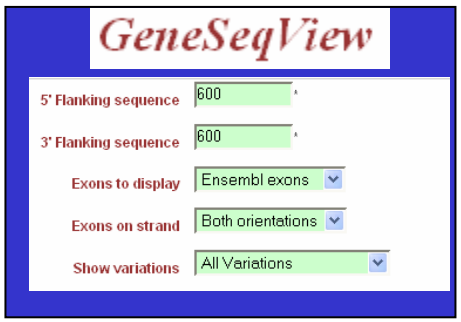
- Per topolino che anche:
  - Sanger (qualche ‘strains’)

SNPs nel Ensembl - Speci

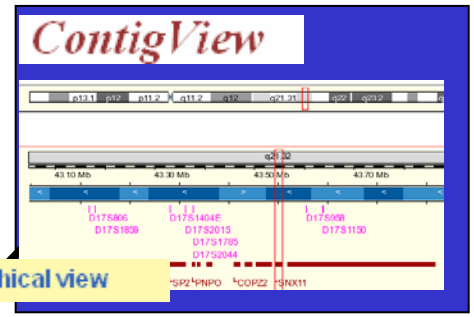


- umano
- topolino
- ratto
- cane
- gallo
- pesce zebra
- zanzare

SNPs nel Ensembl - Pagine



Genomic sequence



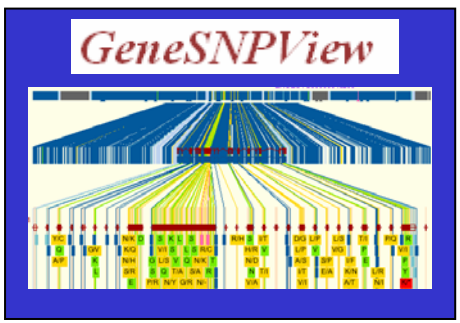
Graphical view

GeneView

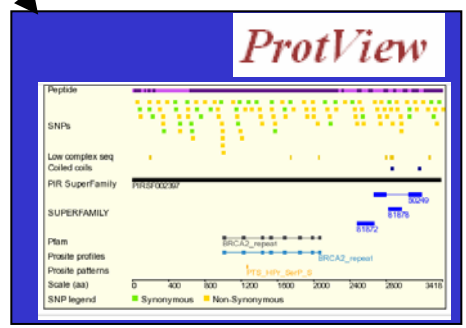
Genomic sequence alignment

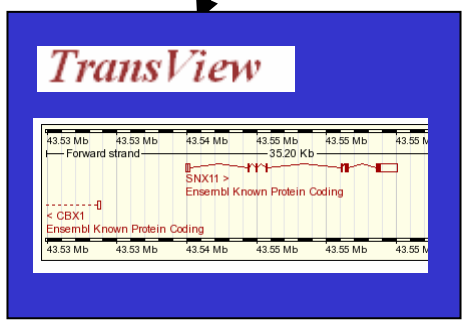
Gene variation info.

Protein information



Transcript information

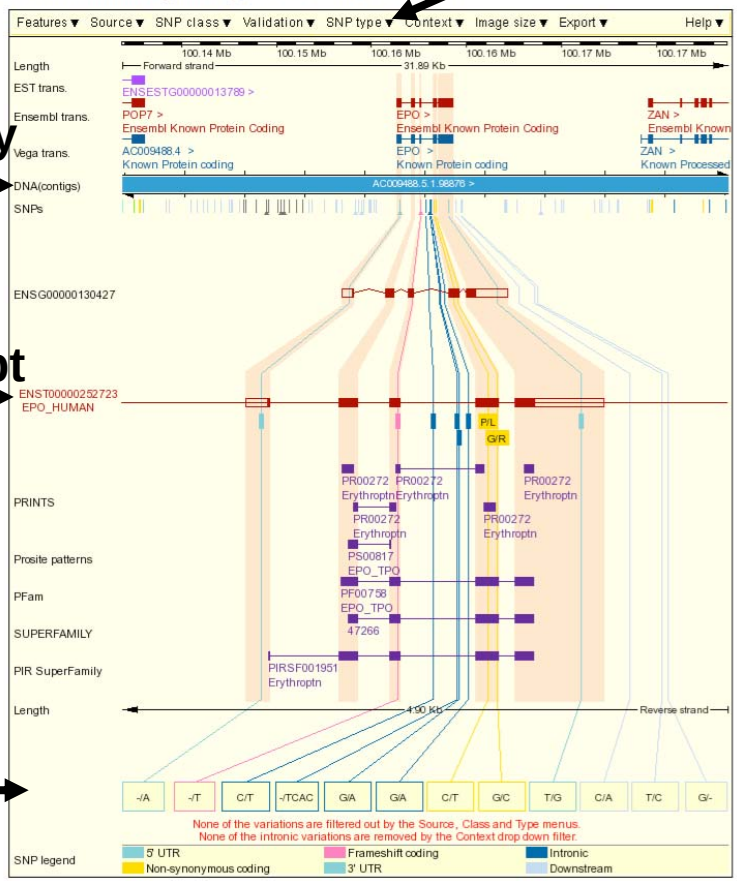




GeneSNPView

Quale SNPs sono nel mio gene?

Choose SNP type



Assembly →

Transcript zoomed →

SNPs →

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	-
rs507392	INTRONIC	7: 100157872	C/T	Y	-	-	snp	HGVbase, dbSNP	cluster, doublehit
rs33978705	INTRONIC	7: 100158068-100158067	-/TCAC		-	-	insertion	dbSNP	-
rs484199	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	-
rs564449	3PRIME_UTR	7: 100159074	T/G	K	-	-	snp	HGVbase, dbSNP, TSC, Affy GeneChip 500K Mapping Array	-
rs551238	DOWNSTREAM	7: 100159464	C/A	M	-	-	snp	HGVbase, dbSNP, TSC	cluster, freq, doublehit
rs4729606	DOWNSTREAM	7: 100159726	T/C	Y	-	-	snp	HGVbase, dbSNP	cluster, doublehit
rs35970223	DOWNSTREAM	7: 100159780	G/-		-	-	deletion	dbSNP	-

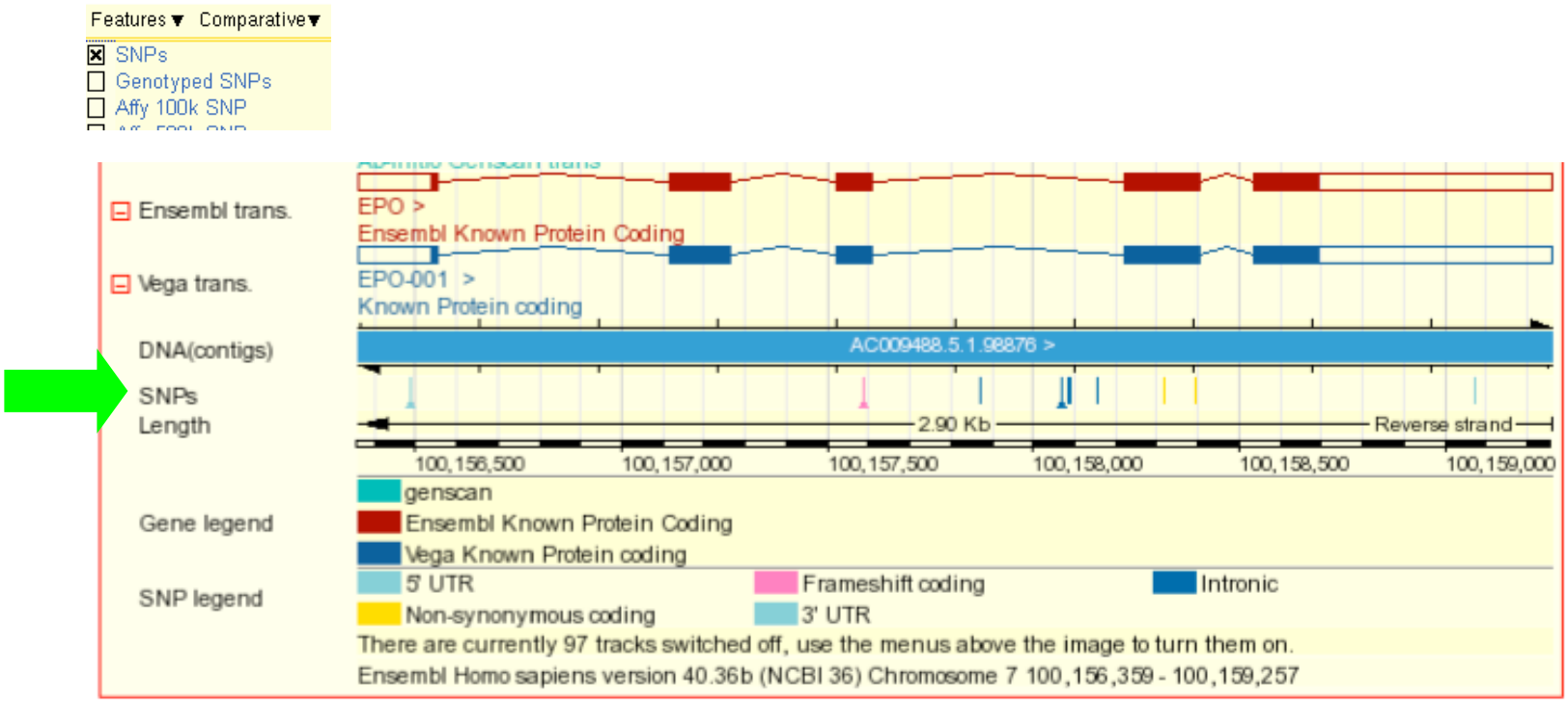
- Gene tree info
- Gene variation info.
- ID history

TransView

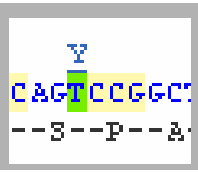
SNPs nella sequenza

ContigView

SNPs in una regione cromosomale

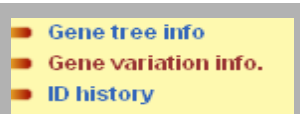


SNPs nel Ensembl (Pagine)



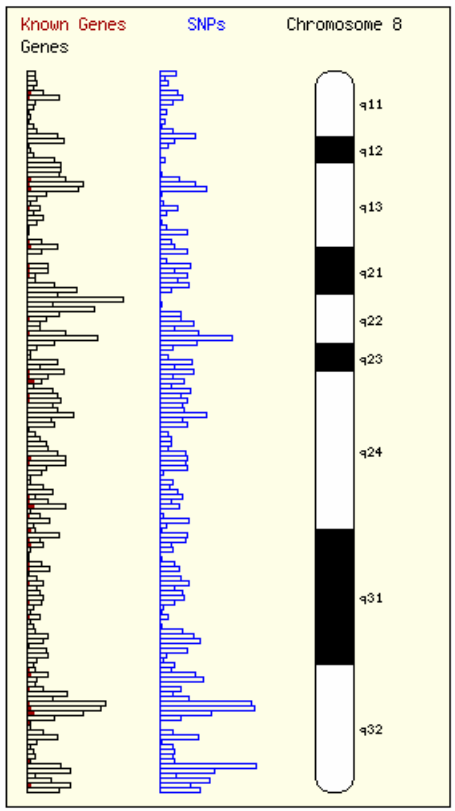
- *TransView*
- *ProtView*
- *MapView*
- *ContigView*
- *GeneSNPView*
- *SNPView*

Features: SNPs in roll-down menus



MapView

La densita dei SNPs su una cromosoma



Click on the image above to zoom into that point

Chromosome 8

Length:	129,041,809	bps
Known Protein-coding Genes:	58	
Novel Protein-coding Genes:	1,196	
Pseudogene Genes:	94	
miRNA Genes:	20	
rRNA Genes:	5	
snRNA Genes:	48	
snoRNA Genes:	38	
Misc RNA Genes:	5	
SNPs:	1,905	

For more information on gene statistics, see the MapView help page

Change Chromosome

Chromosome 8

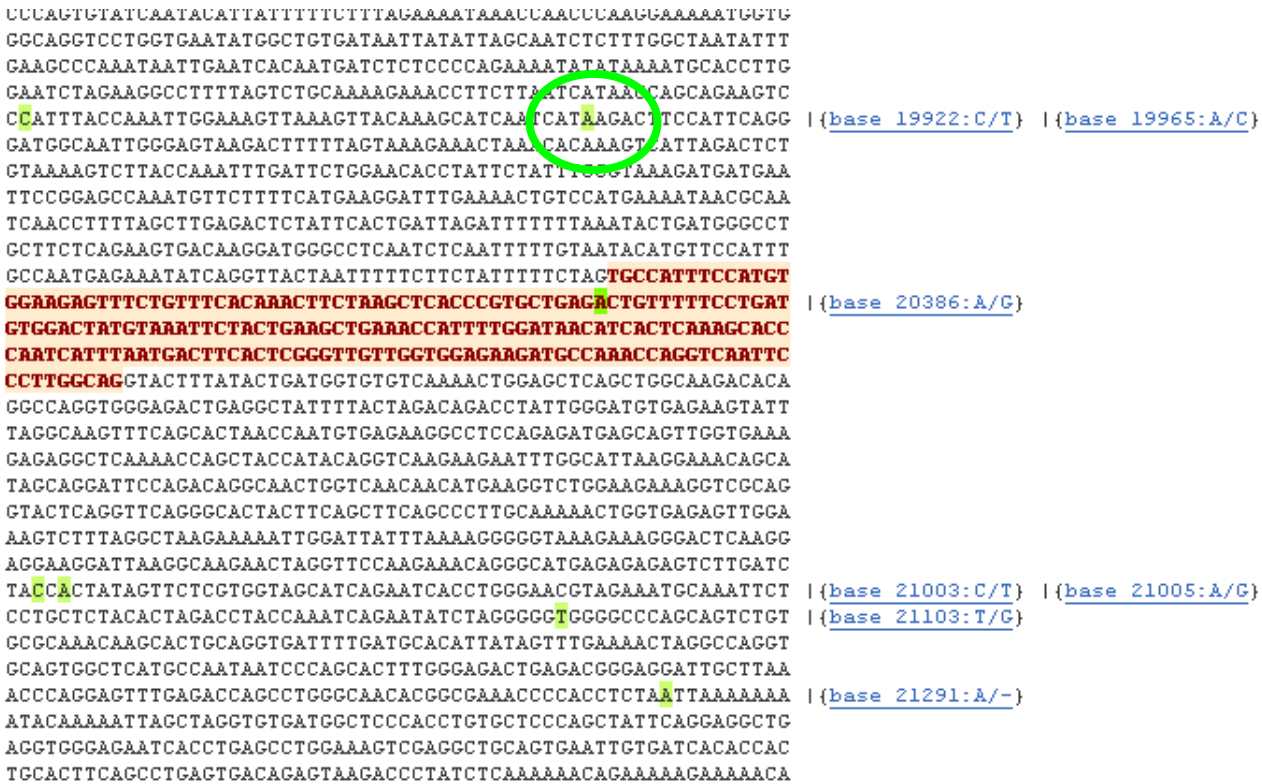
Jump to ContigView

Choose two features from this chromosome as anchor points and display the region between t
mapped to the current Ensembl assembly. If you select "None" for the second feature, the displ

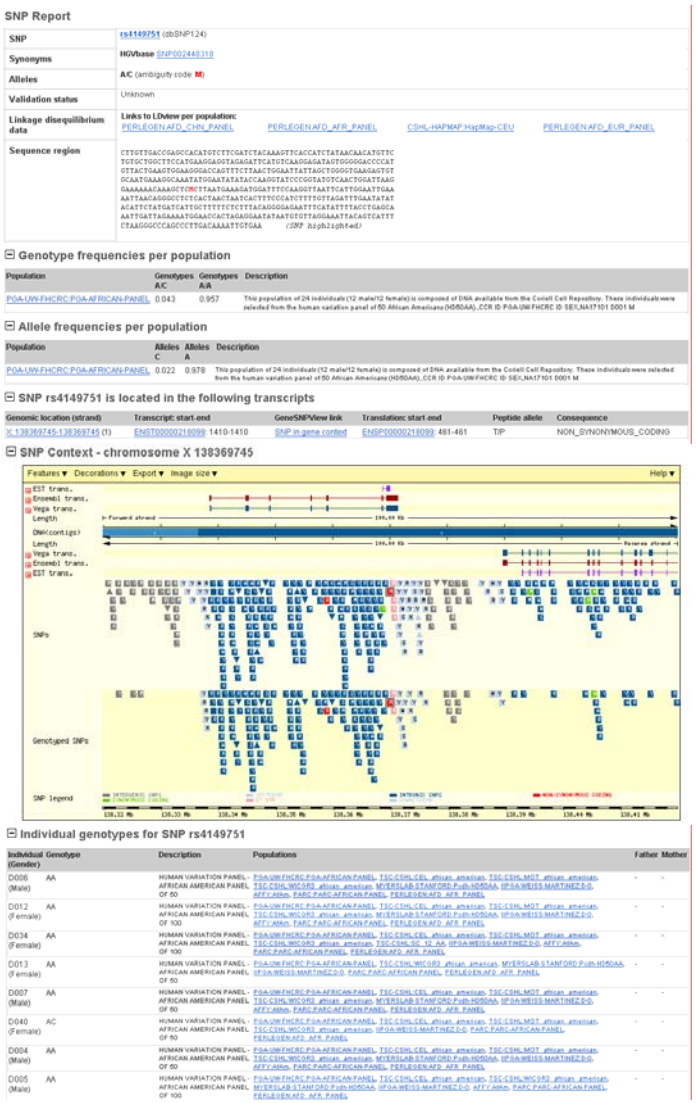
Esempio: Ratto cromosoma 8

GeneSeqView

SNPs nella sequenza genomica



SNPView



Informazione di un SNP specifico:

Di dov'è?... dbSNP, TSC, HGVbase, Affy Chip Data

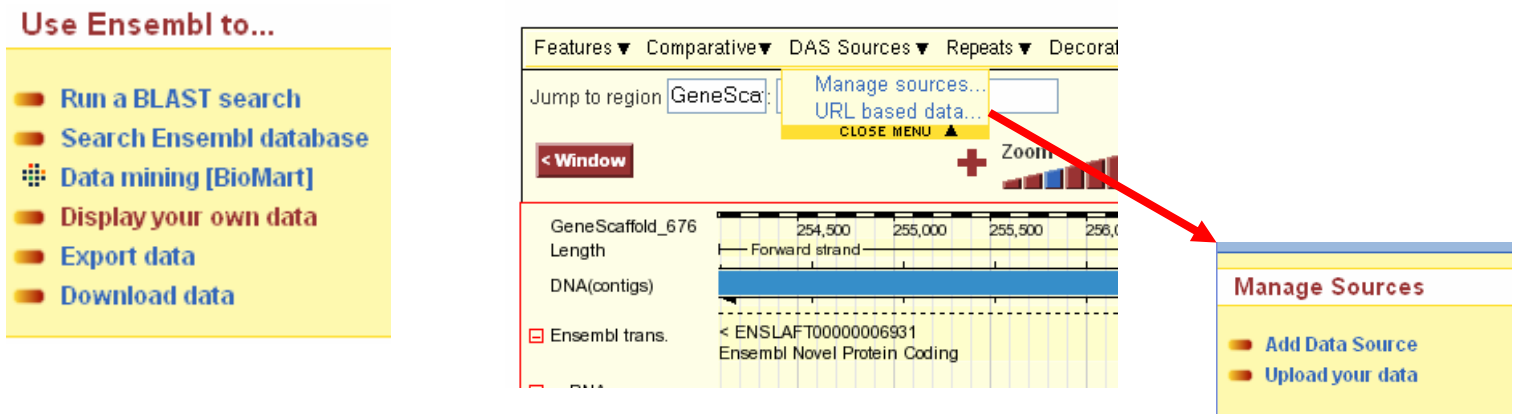
Com'era confermato 'validated'? .. Lo stesso di un un altro SNP…

Genotype e frequenza di un 'allele' in una popolazione.

Locazioni nei transcripts (un SNP puo essere in piu di uno transcript.)

Tipo:coding/noncoding

I Vostri Dati



Mostrate i vostri dai nelle pagine:
ContigView, CytoView, GeneView, TransView, ProtView and KaryoView

O fate un consegna (submission) a dbSNP
http://www.ncbi.nih.gov/SNP/

Ordine

- Variabilità nella genoma di una specie *(SNPs: single nucleotide polymorphisms)*
- **Haplotypes (progetto HapMap)**
- Linkage Disequilibrium
- SNPs nel ‘strains’ di *Mus musculus*

HapMap

- Una campagna internazionale per identificare e catalogare cos'è simile e cos'è diverso nella genoma umana.
- Collaborazione tra scienziati e le agenzie del Giappone, Grand Bretagna, Canada, Cina, Nigeria, and l'USA.
- Tutta l'informazione del progetto è accessibile dal publico.
- http://www.hapmap.org/

Haplotype

- una raccolta di SNPs (di solito ~25 kb)

- ereditato insieme durante il tempo

HapMap

- Fase I:

  1 millione SNPs umani

  Densità: 1/3,000 ‘bases’

- Fase II:

  Un altro 4.6 millioni SNPs dai database publici

  Densità: 1/600 ‘bases’

HapMap

- Sequenze delle popolazioni africane, asiatiche e con progenitori europei

- 270 DNA sequenze di 4 popolazioni:
  - 30 ‘trio’ (2 genitori e figlio/a adulto/a) della populazione Yoruba di Ibadan, Nigeria
  - 45 giapponesi dal regione di Tokyo
  - 45 cinesi (Han) di Beijing
  - 30 ‘trio’ di Utah (USA) con progentori di Europa (nord ed ovest) (gruppo CEPH)

SNPView

Population:	CSHL-HAPMAP:HapMap-CEU [size: unknown]
Description:	30 mother-father-child trios from the CEPH collection (Utah residents with ancestry,from northern and western Europe), representing one of the populations studied,in the International HapMap project (http://www.hapmap.org)
SNP in tagged set for this population:	Yes
Super-population:	EUROPE [size: unknown]
Description:	Samples from Europe north and west of Caucasus Mountains, Scandinavia, Atlantic Islands.

Linkage Disequilibrium

Linkage disequilibrium (LD)

Una misura della frequenza in cui due ‘SNPs’ sono ereditati insieme.

(Casuale? O collegato?)

Freq A * freq B = Freq(AB)?

La misura ‘LD’

$D = p(AB) - p(A)p(B)$

- $D' = D / D_{max}$
  - $D' = 1$

    LD ‘perfetto’

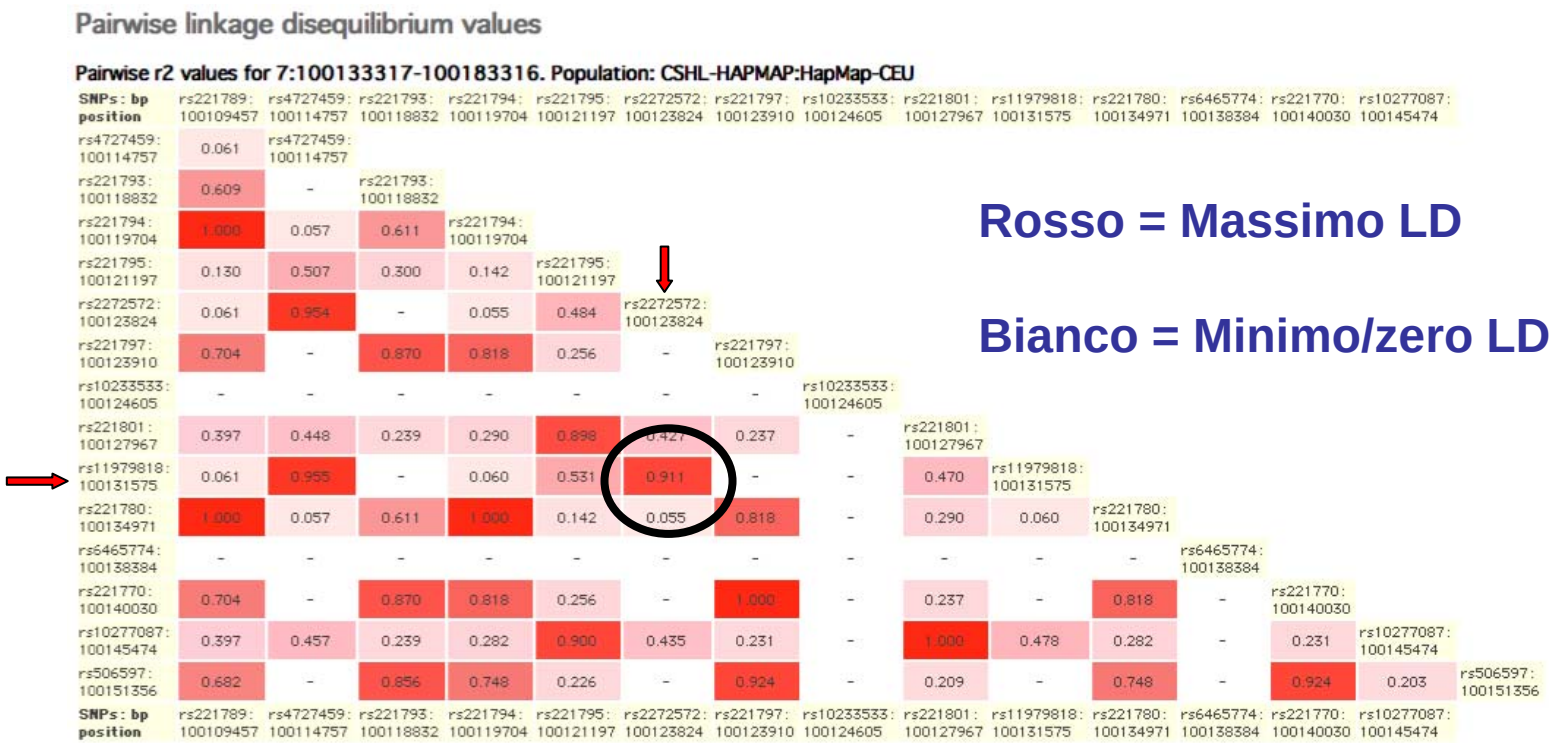
    ~~(‘allele’ A e B sono sempre insieme)~~

- $r^2 = D^2 / p(A)p(B)p(a)p(b)$
  - $r^2 = 1$

    LD ‘perfetto’

Linkage Disequilibrium

LDTableView

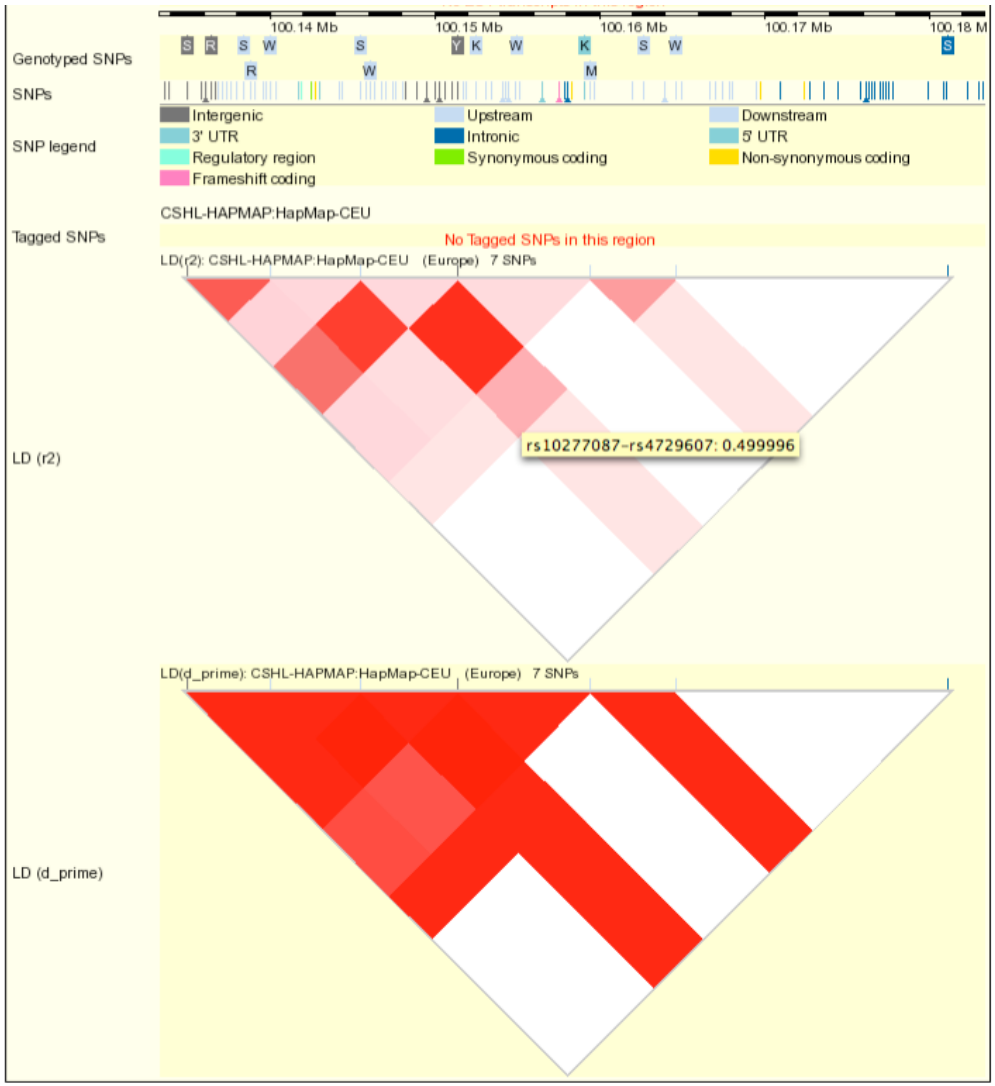


Rosso = Massimo LD

Bianco = Minimo/zero LD

Linkage Disequilibrium

LDView



Rosso (LD alto)

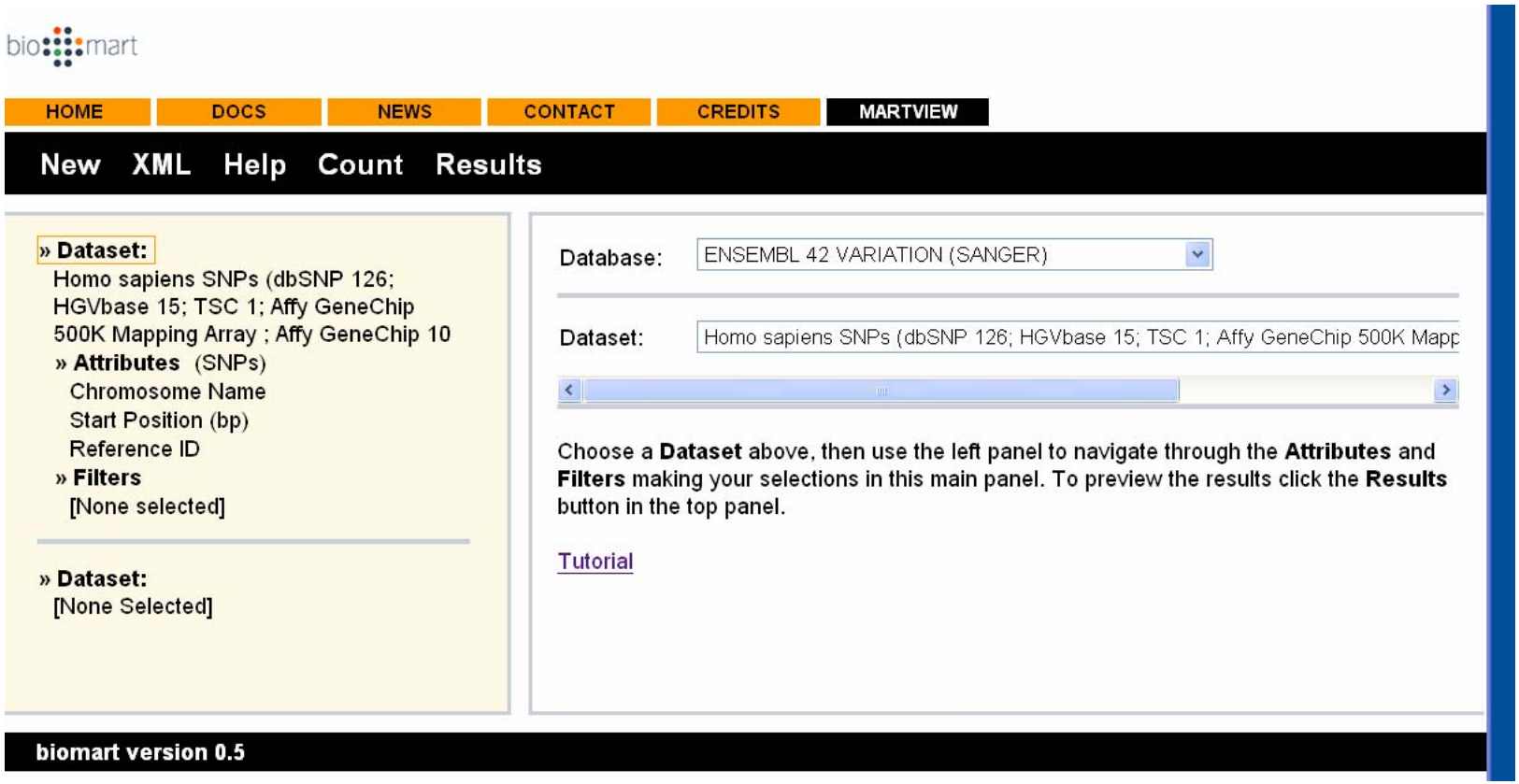
Bianco (LD basso/zero)

Ordine

- Variabilità dell genoma di una specie *(SNPs: single nucleotide polymorphisms)*
- Haplotypes (progetto HapMap)
- Linkage Disequilibrium
- **SNPs nel BioMart e ‘strains’ di *Mus musculus***

SNPs nel BioMart

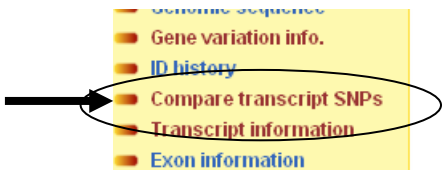
Scegli la database ‘Ensembl Variation’



Anche il sito di HapMap ha un BioMart!

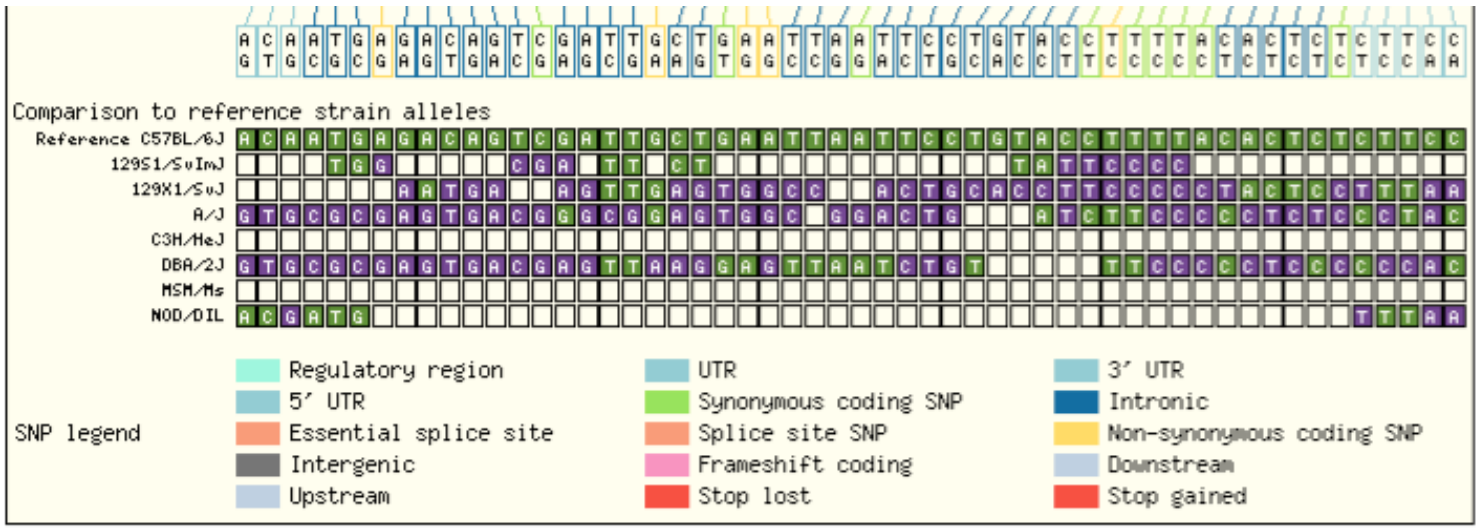
SNPs specifichi per un ‘strain’

- La sequenza genomica del *Mus musculus* nel Ensembl corresponde al strain C57BL/6J
- Però c’è l’informazione per I SNPs nei piu di 30 altri ‘strains’.
- Per vederli nel sito web, ‘click’ qui:



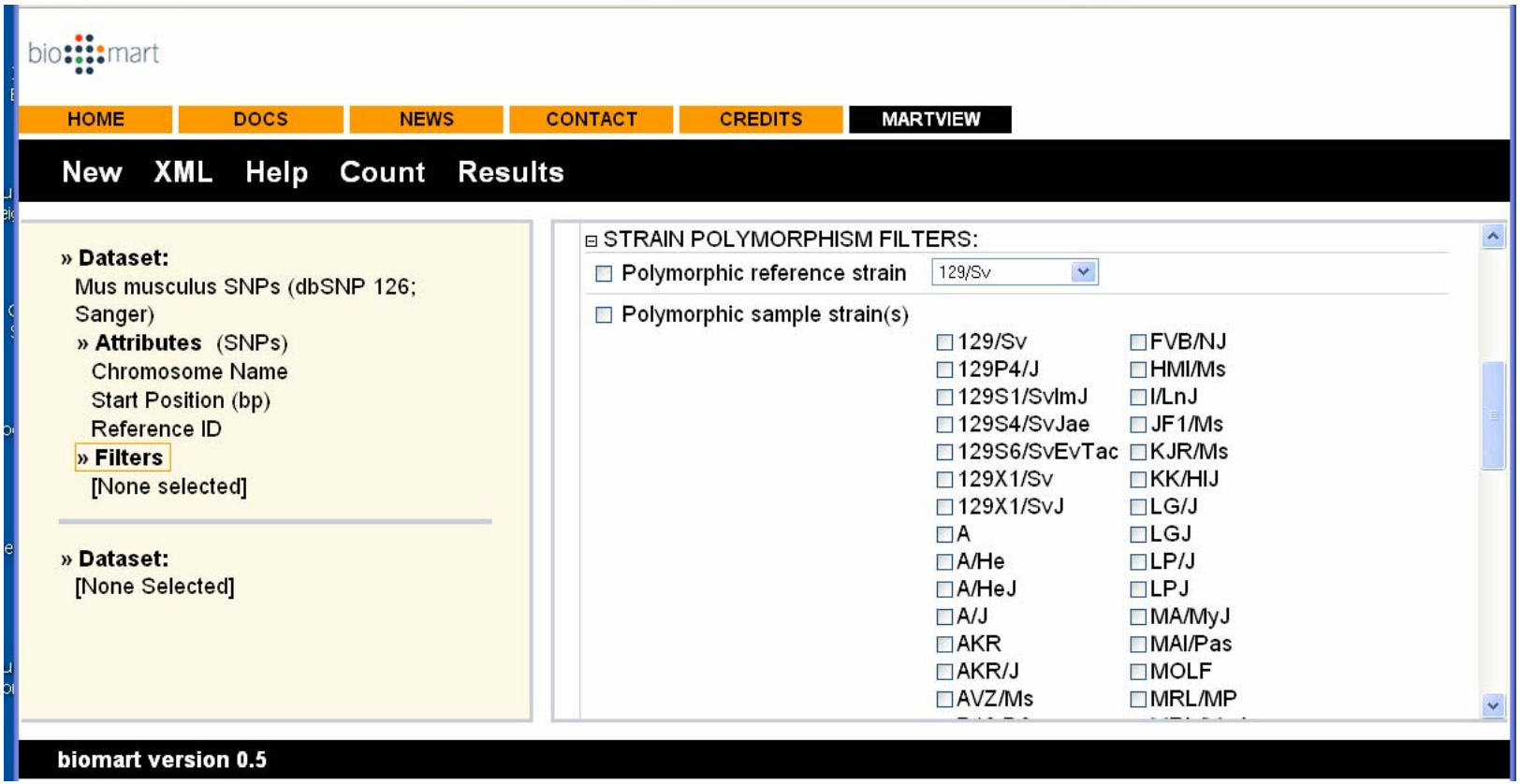
SNPs per i ‘strains’

TranscriptSNPView



SARA = Same As Reference Assembly

SNPs per i ‘strains’ nel BioMart



Gruppo Ensembl

Leaders	**Ewan Birney (EBI), Tim Hubbard (Sanger Institute)**
	Glenn Proctor, Ian Longden, Patrick Meidl, Andreas Kähäri
	Arek Kasprzyk, Syed Haider, Richard Holland, Damian Smedley, Benoit 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