Ensembl:
A New View of Genome Browsing

Giulietta M. Spudich and Xosé M. Fernández-Suárez

European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxon, Cambs, UK

www.ensembl.org

Abstract

An increasing number of methods are being developed to sequence and compare whole genomes, and to detect functionally important coding and non-coding regions. Genome browsers face the challenge of displaying an integrated picture of these data for the biological community. Ensembl allows life scientists to browse through genomic data using an extensive website, and programmers to access the same data directly through the Perl API. This paper explores the browser and underlying data, providing a walk-through of information for one gene with a focus on variation. We hope to demonstrate the power of the browser in this popular, world-wide genomic tool.

Figure 1. Ensembl data is divided into four tabs, reflecting objects in the database.

About Ensembl

Ensembl[1], a joint project between the EMBL's EBI and the Wellcome Trust Sanger Institute, was started in 2000. The focus has been on chordates, and in the last few years, Ensembl has offered a genome browser and access to underlying databases for a rapidly increasing number of vertebrate species (currently 50 species… and counting.) To extend the Ensembl platform to invertebrates, a sister project, Ensemblgenomes, at www.ensemblgenomes.org, has been recently launched.

Evolving bioinformatics methods have allowed Ensembl to increase analysis and annotation of the genome. Gene sets are determined not only for the fully-sequenced genomes through the Ensembl gene-building pipeline[2], but for low-coverage (2X) genomes, via the "low-coverage pipeline" developed at Ensembl. Comparative studies have increased, and every species in

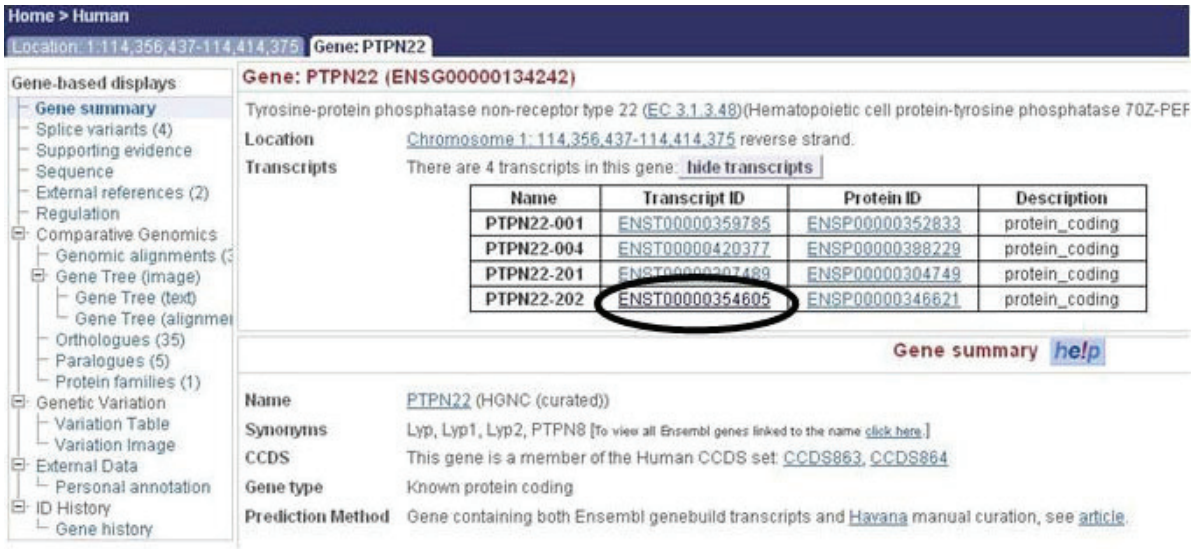

Figure 2. http://tinyurl.com/nw6vcb The gene summary page for PTPN22. Four transcripts are shown in the table, each has a unique ENST identifier. These identifiers are stable across Ensembl releases. One is circled, it provides a link to the transcript tab for PTPN22-202.

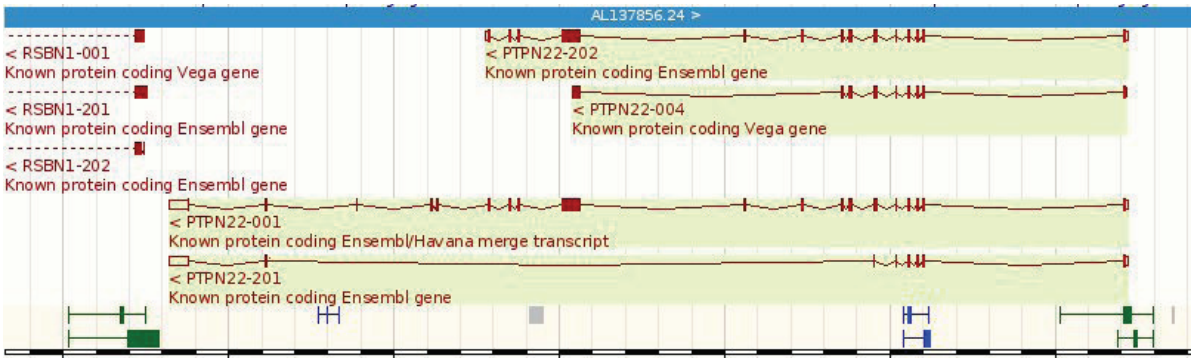

Figure 3. The transcripts are diagrammed below the genome (blue bar) in the gene summary page for PTPN22 (the url is shown in figure 2). Exons are drawn as boxes, and connecting lines show intronic sequence. Filled boxes are coding sequence. The four transcripts are on the reverse strand of the genome.

Ensembl builds phylogenetic trees to determine orthology, paralogy, and ancestral alleles [3]. Whole genome alignments between species pairs, or even multiple genomes (31 mammals) are available. Variations across populations, breeds, and strains are mapped, and a first set of disease-relationships linked to human polymorphisms is available. Functional genomics is also taking a lead, with the ENCODE project revealing potential promoter and enhancer elements in 1% of the human genome, and currently extending to a full genome analysis[4].

Ensembl Data

To cope with an expanding amount of information, Ensembl keeps evolving to enhance the user experience. Based on feedback collected at browser workshops, questions to the helpdesk, and world-wide user surveys run by Ensembl, the website has been designed to display a vast amount of information in an organised manner. The data in the website is now separated into tabs: location, gene, transcript and variation (Figure 1). Users may browse a region of the genome, or focus on homology, variations, or sequence for just one gene or even one splice variant (Ensembl transcript). Data external to the project such as expression profiles from ArrayExpress[5] are also accessible. The tabs allow easier addition of the anticipated flood of phenotypic data, variations, and regulatory regions from projects such as HapMap[6], 1,000 genomes[7] and ENCODE.

Case Study – Gene, Transcript, Variation and Location Tabs

For this investigation, we use version 55 of the Ensembl browser. To view the same pages upon future releases of the website, view the archive site for version 55:
http://Jul2009.archive.ensembl.org/index.html

Let's browse Ensembl by entering *human PTPN22* gene into the search box on the main page at www.ensembl.org

Clicking on the Ensembl gene ID ENSG00000134242 takes us to the gene summary, where we find four transcripts. These isoforms result from alternative splicing of the gene.

At the left of the gene tab are links to the sequence, whole genome alignments, gene trees, variation, and regulatory information.

The transcripts (splice variants) are drawn below the genome (the blue bar). Features below the blue bar are on the reverse strand. Ensembl transcript diagrams show exons as boxes, and

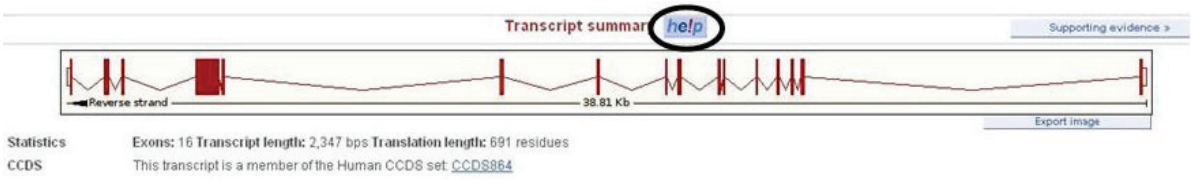

Figure 4. http://tinyurl.com/kmzmhg A larger view of the transcript structure is shown in the transcript summary. The number of exons, length of the spliced transcript in nucleotides, and number of amino acids in the corresponding protein product are displayed. The help button is circled.

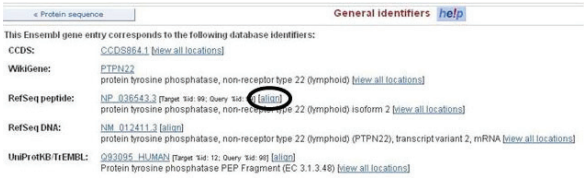

Figure 5. http://tinyurl.com/n39o6z The general identifiers section for the ENST00000345605 transcript shows IDs in other databases that contain a matching sequence to the Ensembl transcript. The extent of the match is shown as percent identity (%id). The align link, circled, shows the sequence comparison between the Ensembl transcript or protein and the external match. Links to the external databases (such as NCBI RefSeq, and UniProtKB) are encoded in the ID.

intronic sequence as connecting lines. Filled boxes are coding sequence, unfilled boxes are UnTranslated Regions (UTR). For example, PTPN22-202, the first transcript in this diagram, has sixteen exons. To learn more about the transcript, either click on the diagram and follow the link to the transcript ID, ENST00000354605, or click on the ID in the table (circled in Figure 2).

Clicking on the transcript ID opens the transcript tab. The transcript diagram is larger in this view, and a summary is available, showing the length of the mRNA (Figure 4).

Each Ensembl view has a page-specific help article, accessible by the "help" button circled above. This transcript is a member of the consensus coding sequence set (the CCDS[8]); this information is written below the diagram. CCDS sequences are agreed upon by Ensembl, Sanger's Vega/Havana[9, 10] team, UCSC[11] and NCBI[12].

View other IDs for this transcript by clicking on general identifiers at the left of the page (Figure 5). For example, the Ensembl protein sequence matches to the NCBI RefSeq[13] protein NP _ 012411.3 with a Blast Reciprocal Hit score of 99%. Click on the NP (known protein) identifier to jump to the PTPN22 protein in NCBI. Or, click the align link to view the sequence alignment between ENSP00000346621 and NP _ 012411.3.

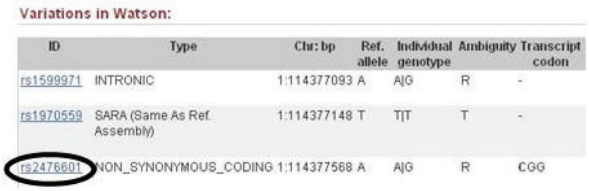

Figure 6. http://tinyurl.com/lclgnm The table above shows an excerpt from the population comparison page. The three variations (rs1599971, rs1970559, and rs2476601) are found in James Watsons' genome. The first is intronic, the second is the same as the reference sequence GRCh37, and the third shows a non-synonymous allele, indicating there is more than one possible amino acid at that position. The variations IDs provide links to the variation tab, such as rs2476601 (circled).

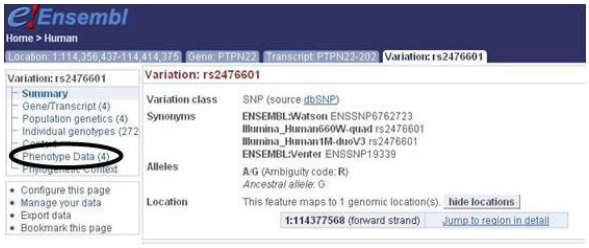

Figure 7. http://tinyurl.com/meqb64 The variation summary for rs2476601, an NCBI dbSNP identifier.

Let's explore variations such as polymorphisms mapped to this transcript. The population comparison page displays all the variations, such as SNPs and insertion-deletion mutations (indels), across populations (Figure 6). The ID and position of the variation (such as intronic, non-synonymous coding, etc.) are noted in the first two columns, and the remainder of the table includes the allele in the individual (or strain or breed for non-human species), and the source of the polymorphic data.

An image with this same information drawn graphically is available in the next link (*comparison image*). Both the table and image can be customised using the *configure this page* link at the left. This menu allows selection of individuals, variation types and/or sources to be displayed.

Click on any variation from the table or image to open the variation tab, a focused set of pages for one variation. In this example, let's click on rs2476601, a non-synonymous coding SNP found in James Watson's genome[14].

The variation summary is the first link in the tab, show in Figure 7. Here we find the SNP source (NCBI dbSNP[12]). Any other IDs that this SNP is known by are listed under "Synonyms".

The links at the left of the variation tab are specifically for rs2376601. Click on Phenotype Data (circled in Figure 7) to see that the NHGRI GWAS catalogue[15] relates this SNP to Crohn's Disease, Rheumatoid Arthritis, and Type I Diabetes (Figure 8).

You can easily jump back to the gene or transcript displays by clicking on the appropriate tab. Let's explore the fourth tab in this set, the location tab.

Figure 8. http://tinyurl.com/mcnybe The variation tab houses the "phenotype data" page. This shows any associations between a variation and disease phenotype in the NHGRI GWAS catalogue. The study in PubMed that shows the association of the variation to the phenotype is also listed.

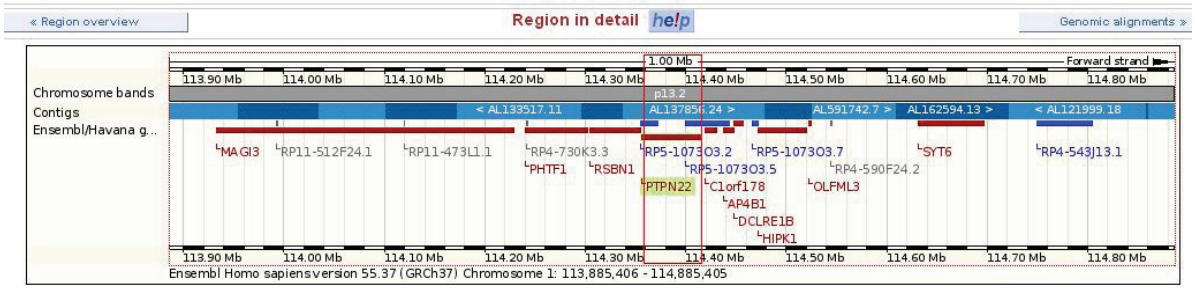

Figure 9. http://tinyurl.com/lfnnna The top panel of the region in detail page, centred on the PTPN22 gene (highlighted). Protein coding genes from Ensembl are shown in red, genes from the Vega/Havana project are blue. Grey blocks and identifiers show pseudogenes.

Figure 10. The main panel of the region in detail page is centred on PTPN22. The four transcripts are drawn as in the gene summary page (Figure 3). Coding sequence in the CCDS set along with human cDNA sequences in EMBL-Bank are aligned to the genome. Regions of alignment are displayed in green, filled boxes. Gaps in the alignment are shown by empty boxes. Aligned sequenced support the (red) exons in PTPN22 transcripts.

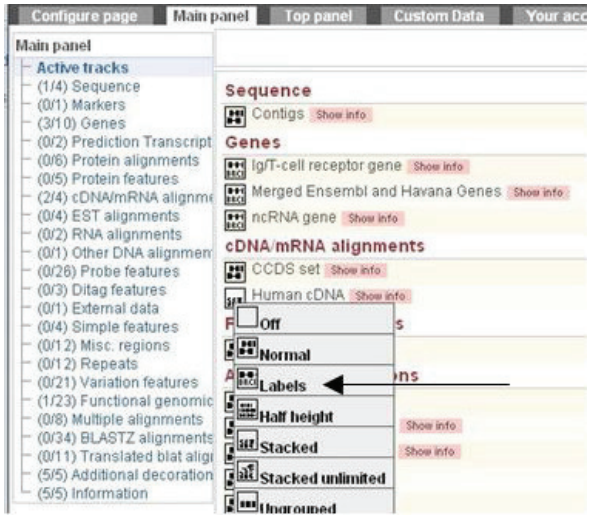

Figure 11. The active tracks (i.e. selected tracks) in the region in detail page. The menu is revealed using the configure this page link. Tracks may be selected in the main or top panel (see tabs) or user data may be uploaded, along with DAS sources (custom data tab). Menus of displayable tracks are shown at the left. The human cDNA track is collapsed; expand the track by clicking on 'normal' or 'labels' (shown by the arrow in the figure.)

Clicking on the location tab (circled in Figure 8 opens the "region in detail" page. For long-time users of Ensembl, this was the ContigView page (Ensembl versions 50 and previous).

The top panel shows a large (1Mb) region of the genome centered on the PTPN22 gene (highlighted in figure 9). Neighboring genes on either side of PTPN22 include RSBN1 and AP4B1. Any gene can be clicked on to view the ID, and/or jump to its gene or transcript tab.

The red box outlining the PTPN22 gene is expanded in the panel below (the main panel, figure 10).

All transcripts (four) of the PTPN22 gene are displayed. These transcripts can be from the Ensembl genebuild, the Vega/Havana manual curators, or they may be a merged transcript, agreed upon by both projects. This is the case for PTPN22-001.

The CCDS set is also drawn in Figure 10, along with the human cDNA alignments. The cDNA track is collapsed by default. To expand the cDNA alignments, click *configure this* page (Figure 11). The active tracks menu will appear (these are the tracks displayed in the "region in detail" page.)

Now each cDNA can be clearly seen (Figure 12). The dark boxes within each entry (such as BC0716701) show the alignment of the cDNA to the genome. Unfilled boxes are gaps in the alignment. (Gaps are expected for cDNA and protein alignments to the genome, as intronic sequence will not be present in cDNA and protein sequences). Click on any green cDNA diagram for an information box, showing the source, as in the example above. These cDNA alignments are updated with every new release of Ensembl, for human.

Case Study Summary

We started this walk-through by searching for a gene symbol (PTPN22). However, Ensembl also allows search by genomic region, accession number, variation ID, clone ID, or disease or phenotype.

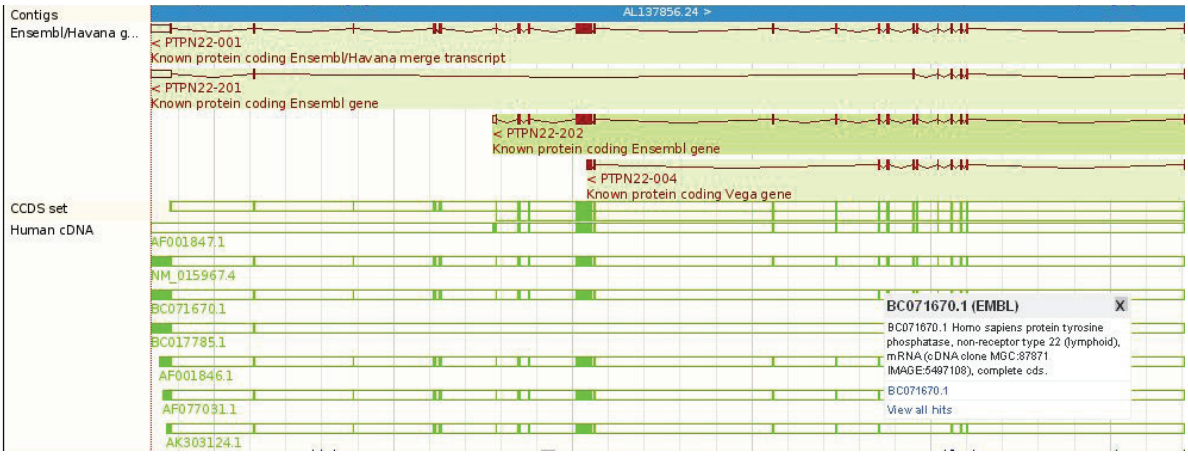

Figure 12. The region in detail page with the human cDNA track expanded. Click on any cDNA alignment to view a pop-up box of information. For example, BC071670.1 is a human tyrosine phosphatase record in EMBL-Bank. The CCDS track has been turned off in this view.

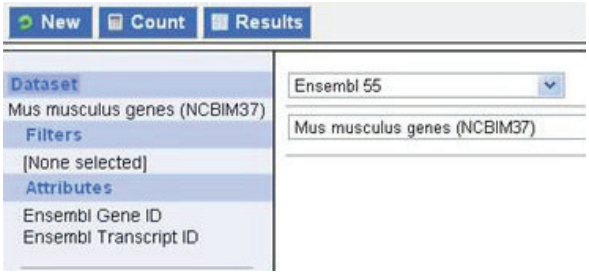

Figure 13. http://www.ensembl.org/biomart/martview The BioMart tool allows fast access of data in Ensembl. Sequences and annotation may be exported in FASTA or tabular format. In this example, all mouse genes in Ensembl version 55 have been selected. Output information appears as columns in the table, and are selected in "attributes" (in this case, the Ensembl gene and transcript IDs.).

We quickly learned there are four isoforms (splice variants) that come from the Ensembl genebuild, and/or the Vega/Havana project. The location tab, "region in detail" page displayed human cDNA alignments for the gene locus. This allows users to view support for each exon by entries in databases such as EMBL-Bank[16]. Protein alignments can also be viewed in the "region in detail" page.

One transcript was also explored (PTPN22-202, also named ENST00000354605 in Ensembl). The protein product for this transcript, ENSP00000346621, matched well to the RefSeq peptide NP036543.3, which was seen in the *general identifiers* view.

Finally, variations in different individuals were compared in the population comparison page and the comparison image. One specific variation, rs2476601, was explored in depth. Ensembl views showed the source of this variation was dbSNP, though it is also present in other variation sets, and that three diseases are associated with variation at this nucleotide position.

Other Access

Ensembl data need not be accessed through the browser. The tabs reflect the organisation of the data in our publicly accessible databases, which are queried by a large number of bioinformaticians. A Perl API is supported, which is heavily accessed by our user community, and is kept current with Ensembl releases (every two months).

http://www.ensembl.org/info/docs/api/

In addition, BioMart allows fast mining of Ensembl data for programmers and non-programmers alike (Figure 13) [17] [18].

Ensembl Scope

Who's using us? A recent heat map of page impressions on our website shows a worldwide community of users (Figure 14). Countries shaded darkly show highest usage, and lighter countries access the browser less. Interestingly, this reflects locations of our worldwide workshops.

Ensembl offers workshops in the browser for free. Get to grips with our data by hosting a workshop. Details are here:

http://www.ensembl.org/info/about/outreach/

In addition to workshops, we have tutorials on our website, and a YouTube channel of videos providing task-based walk-throughs of the browser.

http://www.youtube.com/user/EnsemblHelpdesk

Blog statistics also show a worldwide readership (Figure 15). Posts range from the direction of genomics to details about upcoming species or variation sets in Ensembl.

Find out about upcoming workshops, species, and more on our blog.

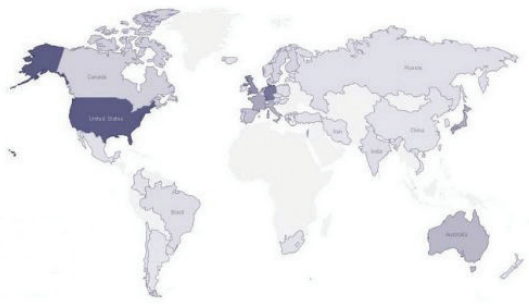

Figure 14. A heat map showing page impressions to the Ensembl browser in May, 2009. Europe, the USA, Australia and Japan show heaviest use.

Figure 15. Blog readership is marked by red "pins" on the map. "Recent visitor map" and statistics are from StatCounter (http://www.statcounter.com/).

http://ensembl.blogspot.com/

Despite the relatively high number of queries to the Ensembl helpdesk, answers are returned within one to two days. Questions or comments may be submitted through a form on the Ensembl browser, or directly emailed to helpdesk@ensembl.org.

Conclusions

The Ensembl project aims to provide high-quality genome annotation for vertebrate genomes. Users can freely access all data from various sources, using the extensive pages of the browser at www.ensembl.org, or through BioMart or the Perl API. Ensembl provides displays under four separate tabs to allow flexible addition of the new genomics data to come. Our worldwide users access both the website and the blog.

References

1. Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, et al. (2009) Ensembl 2009. Nucleic Acids Res 37(Database issue): D690-7.
2. Potter SC, Clarke L, Curwen V, Keenan S, Mongin E, et al. (2004) The ensembl analysis pipeline. Genome Res 14(5): 934-941.
3. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, et al. (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res 19(2): 327-335.
4. ENCODE Project Consortium. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447(7146): 799-816.
5. Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, et al. (2009) ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Res 37(Database issue): D868-72.
6. International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449(7164): 851-861.
7. [Anonymous]. 1,000 genomes. .
8. Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, et al. (2009) The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Res 19(7): 1316-1323.
9. Ashurst JL, Chen CK, Gilbert JG, Jekosch K, Keenan S, et al. (2005) The vertebrate genome annotation (vega) database. Nucleic Acids Res 33(Database issue): D459-65.
10. Wilming LG, Gilbert JG, Howe K, Trevanion S, Hubbard T, et al. (2008) The vertebrate genome annotation (vega) database. Nucleic Acids Res 36(Database issue): D753-60.
11. Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, et al. (2009) The UCSC genome browser database: Update 2009. Nucleic Acids Res 37(Database issue): D755-61.
12. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, et al. (2009) Database resources of the national center for biotechnology information. Nucleic Acids Res 37(Database issue): D5-15.
13. Pruitt KD, Tatusova T, Maglott DR. (2007) NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35(Database issue): D61-5.
14. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, et al. (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452(7189): 872-876.
15. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106(23): 9362-9367.
16. Stoesser G, Baker W, van den Broek A, Camon E, Garcia-Pastor M, et al. (2002) The EMBL nucleotide sequence database. Nucleic Acids Res 30(1): 21-26.
17. Haider S, Ballester B, Smedley D, Zhang J, Rice P, et al. (2009) BioMart central portal--unified access to biological data. Nucleic Acids Res 37(Web Server issue): W23-7.
18. Mullan L. (2006) Mining ensembl. EMBnet.News 12(1): 12-13.