Author's response to reviews

Title: The Roche Cancer Genome Database 2.0

Authors:

Jan Küntzer (jan.kuentzer@roche.com)
Daniela Eggle (daniela.eggle@roche.com)
Hans-Peter Lenhof (lenhof@bioinf.uni-sb.de)
Stefan Klostermann (stefan.klostermann@roche.com)
Helmut Burtscher (helmut.burtscher@roche.com)

Version: 2 Date: 22 February 2011

Author's response to reviews:

Dear Dr Patrick M.A. Sleiman,

Thank you for the positive response concerning our submitted paper “The Roche Cancer Genome Database 2.0”.

We would like to thank the referees for their constructive and helpful comments. All comments have been integrated in the revised version of the paper submitted to the online submission system.

Please find below a detailed list of changes to the concerns raised by the referees.

Reviewer: 1

----------

Major Compulsory Revisions

1) Is it possible to provide access to archived versions similar to other databases such as Ensembl?

We fully agree with the referee that it is sometimes interesting to have access to previous version of the data. However, we do not provide a download of the full data set and refer to the external data sources for access to old version of the data.

The current version of each data source can be found on the webpage.

2) It is nice to have the pathway enrichment analysis function. However, the method underlying enrichment analysis should be described in the manuscript. I found some problems with the enrichment analysis results. For example, using pathway enrichment analysis IV, selecting cancer subset colorectal, sorting by the p values, the result for “MAPK signaling pathway” is #signif. amplified/deleted genes in Pathway = 0, # signif. amplified/deleted genes not in Pathway = 1969, # not signif. amplified/deleted genes in Pathway = 201, # not signif. amplified/deleted genes not in Pathway = 14349, and p-Value = 7.05e-12. How
can one get such significant enrichment with #signif. amplified/deleted genes in Pathway = 0? Similarly, using pathway enrichment analysis I, querying for tissue “colon”, KEGG pathway “Drug metabolism - other enzymes” showed p value of 0.09 with 0 mutated gene. It seems that the underlying algorithm or code needs to be carefully reviewed.

The pathway enrichment analysis 4 smart search performs a Fisher's exact test, which is a two-sided statistical significance test for categorical data, measuring the association between two variables in a 2x2 contingency table. The two variables used are "number of participants of a pathway" and "number of amplification/deletions". The test is performed with the hypothesis that both categories are not associated but a low p-value indicates an association.

The p-value does not necessary contain biological but statistical relevance and could therefore give a hint for biological relevance.

We reviewed the code and the calculated p-values are correct.

A low p-value of a pathway can result from to less or to many amplifications/deletions found in this pathway as expected. In the example shown by the referee the low p-value results from to less amplifications/deletions (count was zero) as expected.

To make this more clear we changed the help on the webpage about the different smart searches: the section describes the test in more details.

3) For the cell line search, it is a useful function to show similar cell lines. However, similarity measurement needs to be described in the manuscript. Moreover, some of the links seem to be broken. For example, in the results for SW480, neither of the links to SW620 or SW620-NCI works.

If a link shows a page with no information given for the searched cell lines does not mean the link is broken. Nevertheless we checked again the links which should work if information for the cell line is in the database.

The similarity is described in details in the help section:

"On the lower left side you can find a card with a list of all cell lines with a common ancestor to the given cell line.

Over several decades, a number of cell lines have contaminated other cell cultures. The best known example is HeLa which has appeared under many other names. It is possible to identify samples derived from the same individual on the basis of their polymorphic marker genotypes ("DNA fingerprint"). The genotype information for all the cell lines and primary tumour DNAs analysed using the Affymetrix SNP arrays were compared by pair-wise analysis. All pair-wise comparisons against percentage identity which have an identity greater than 70% are included in the table. These pairs of lines are highly likely to have been derived from the same individual and therefore most are probably
derivatives of each other. The analysis has also been carried out for the NCI60 set of cancer cell lines. Within the 59 lines that make up the NCI60 set there are three pairs of lines with high identity to each other (NCI-ADR-RES/OVCAR-8, M14/MDA-MB-435 & SNB19/U251) one line (HCT-15) sourced from the NCI is different from a line of the same name from our general set of cell lines.

4) Queries from different paths for the same question may end up with different answers. For example, using gene query for KRAS, clicking on Number of samples with somatic mutations in this gene, and searching for colon under primary tissue, I got 4 mutated samples. Using sample query for colon and searching for KRAS under gene, I got 7 mutated samples. These problems need to be fixed to avoid misleading information.

We fixed the problem and the counts are now the same for both ways of querying.

Minor Essential Revisions
1) Table 1. Please provide release versions of the databases.

We change Table 1 and added the information

2) Publication-derived mutations comprise a major class of mutations in the database. It would be helpful to provide a brief description on how these mutations were collected.

The mutations are collected by hand mostly from the suppl of publications.

3) Data collection section, paragraph 4, “The cancer genome data is further enriched by pathway information from KEGG, BioCarta, and Roche internal networks”. I don’t see information on the Roche internal networks on the website.

The Roche pathway are available if the user searches by pathway. You can find three different tabs, one labeled by Roche. These are the Roche internal networks.

4) It might be helpful to provide a pull-down menu for cell lines and diseases using controlled vocabularies.

We do not provide a pull-down menu for cell lines and diseases but rather provide the auto-suggestion functionality on the search page as used for genes.

5) In the cell line search output, I don’t understand what are the numbers of samples without somatic mutation and with somatic mutation. Moreover, the amplification section is difficult to understand too.

The number of samples without and with somatic mutations are the count of all samples with the given name. We cannot merge cell lines by name since the samples come from different labs and batches and the same name does not necessary mean the same cell line.
If you look e.g. at ht-29 we have 43 cell-lines in the database where in 27 somatic mutations were found and in 16 no mutations were found.

Reviewer: 2

1. For manuscript

The previous version of this database was described in Human Mutation and Nucleic Acids Research, 2010. Has the previous version been already abolished? If the previous version is completely included in the new version, it would be better to be mentioned in a text about that. If not, isn’t it worth keeping older version on line with the new version? Otherwise, readers of the papers of the previous version might be confused a little.

The previous version of the database was published in Human Mutation which is cited in the paper. However the database was not published in NAR 2010. The old version offered only 3 simple searches which are still available. The result pages contain much more information which a similar structure as the old version.

We think that a user of version 1 can easily work with version 2. For users who have problems with the new version we offer a detailed help section.

2. Points to be desired for improvement and further development for database contents and functions:

(1) In spite that the database name has the word "genome", "genomic" functions look not enough. For example, the following data contents and functions are desirable.

a. Graphical genomic view for mutations/polymorphisms in introns, UTRs, promoters and spacer regions.

b. Graphical wide genomic view to show CNVs which affect one or more genes.

We fully agree with the referee that this content and functionality is desirable and can be found on our wish list for the next version of the database.

(2) Browsing functions for entire contents are desired. Users cannot know if the information of their interests is covered until they actually search it. Especially, number of contents is not available for genes, cell lines and tissues. For these, approximate scale of this database is not clear (e.g. this database includes 100, 500, 1000, or 5000 genes?)

All of this information can be found on the "About" page linked on the main page in the upper right corner.

(3) Currently, domain information is not displayed for any gene as far as I tried. Is that OK? Is the function highly browser-dependent? I used Firefox ver.3.6 and IE
ver.6. on WindowsXP.

There was a problem with the public version. We fixed it and the domain information should be displayed correctly now (see e.g. ERBB2)

(4) Search functions using domain information is desired. For example, search for genes/proteins with domain(s) of interest and mutations/polymorphisms within the domain(s) would be very useful.

This is a very interesting search functionality. We would like to add a smart search using domain information in the next version of the database.

(5) Regarding the current "Smart search function":
a. If each information in result table has hyper-link, it would be useful.
b. When users wish to add more keywords such as additional genes after some search, it would be better that they can be added in the result view. Currently, it looks no such function. Even, clicking backward button of browser totally abolishes previously typed keywords. It’s very sad.

We think that going back from the result view to the smart search interface should show the old search performed. This works perfectly fine in IE6, but unfortunately not in Firefox 3.6.

We are currently checking to overcome the problem, but it seems to be a problem of the browser.

c. In the Pure Somatic Mutation function in Smart Search, the result with 2 or more genes is displayed without gene names. If it’s within the planned functions, how readers should use this?

We added the gene name in the result page.