Reviewer’s report

Title: The Roche Cancer Genome Database 2.0

Version: 1 Date: 24 November 2010

Reviewer: Bing Zhang

Reviewer’s report:

General

In this manuscript, Kuntzer et al. describe an updated version of the Roche Cancer Genome Database (RCGDB2.0). The new version has significantly more data. It also provides a more user-friendly query interface with smart search and advanced search options. Overall, it is useful to have such a central repository for the cancer related mutations. However, there are several issues that should be considered to further improve the manuscript and the website.

-------------------------------------------------------------------------------------------------------------------------------------

Major Compulsory Revisions

1. A clear description of improvements over the first version of the database is needed. Moreover, although it is nice to have the most up-to-date information, some users may prefer to have access to previous versions (e.g. for reference data in publications). Is it possible to provide access to archived versions similar to other databases such as Ensembl?

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 siginf. 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.

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.

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.

Minor Essential Revisions

1. Table 1. Please provide release versions of the databases.
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.
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.
4. It might be helpful to provide a pull-down menu for cell lines and diseases using controlled vocabularies.
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.

Discretionary Revisions (which the author can choose to ignore)

None

Level of interest: An article of importance in its field

Quality of written English: Acceptable

Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests:

I declare that I have no competing interests