User guide to upgrading Ensembl in Stemformatics

Every three to four months Ensembl upgrades to a newer version. In this update it retires, remaps and changes the Ensembl Gene IDs in it's database. In that way it is kind of like a newer model car. Most of the peices stay the same but some of the underlying parts may be replaced with newer parts or renamed (remapped genes), removed completely (retired genes) or left alone.

Stemformatics made the decision to upgrade to human v69 last year in December but unfortunately we are already out of date with v70 being released in January.

We decided to upgrade Mouse from Ensembl v63 to v67 so we could stay on the same version of mouse mm9 genome (NCBIM37) to help our collaborators.

This document gives you some insight into what these changes mean to you as a Stemformatics User.

Overview of Ensembl Gene IDs

To understand what changes have been affected you need to know a little about how we organise genes. For us, the ensembl gene id is the unique identifier of a gene. A human ensembl gene id looks like this ENSG00000115415 and a mouse one looks like this ENSMUSG00000026104. As per the car parts analogy, treat it like a unique code for your car part.

A gene symbol, other aliases (names) and even Entrez Gene IDs (from NCBI) can be mapped to these ensembl genes by tables that Ensembl provides. Here are two examples:

	Human Ensembl Gene	Mouse Ensembl Gene
Ensembl Gene ID (Unique)	ENSG00000115415	ENSMUSG00000026104
Official Gene Symbol	STAT1	Stat1
Aliases	CANDF7 ISGF-3 STAT91	2010005J02Rik AA408197
Entrez ID	6772	20846

So, in theory (and in practice) you can have multiple unique ensembl genes that have the same gene symbol, but Ensembl treats them as different genes. Search for POU5F1 to see an example: http://stemformatics.org/genes/search?gene=POU5F1 - note that when you click on

the details the Ensembl Gene ID in the right hand box changes.

How the upgrade affects you

When we upgrade Ensembl what we did was update the list of Ensembl Gene IDs. Now this means that some genes from the previous Ensembl release may have been retired or replaced.

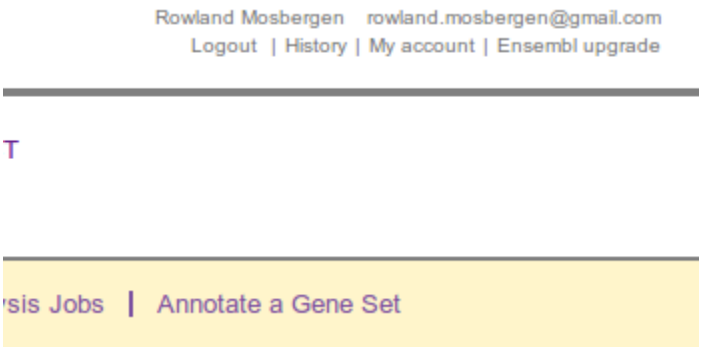
This would affect the:

- Gene Expression graph,
- workbench jobs including Gene Set Annotation,
- private gene sets,
- and public gene sets.

For the Gene Expression Graph, when we map the probes in a dataset to an Ensembl Gene ID, we might have to map them to a new Ensembl Gene ID than we had previously. So you might realise that in the Gene Expression Graph a gene for a particular dataset only used to have four “probes” but now has three. As per our analogy - in the newer car you might only need two parts that make up the radio instead of three.

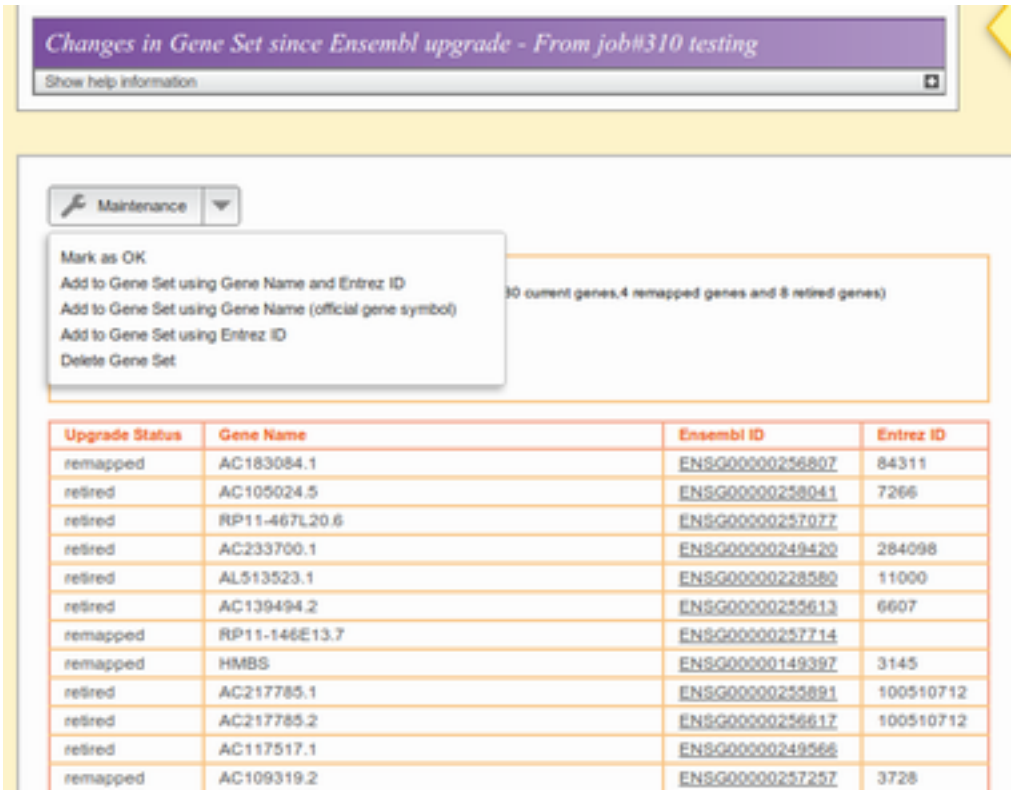
In the workbench jobs, it is better to re-run your jobs to get updated results after the Ensembl Upgrade. This is because the values that were calculated pre-upgrade may be inconsistent with the genes that are currently part of Stemformatics post-upgrade.

For private gene sets - you might find that your gene lists have less genes in them than before. We keep all the details of what has changed in your private gene sets - you can view them by clicking on “Ensembl upgrade” when you are logged in on the top right of the page near “My account”



You can then see your private gene sets that have been affected and can manipulate them to your desire. You can simply “Mark as OK” if you choose not to do anything or you can click on view to learn more.

In the view - you can try and use the old Entrez ID and/or Gene Name to find any genes that maybe associated with these “retired” Ensembl Gene IDs. This will take you to the Bulk Import Manager to help you add these genes to your gene set if possible. This is the equivalent of not being able to look up the car part unique number, but looking up the everyday name in the hope that you can match it to one of the new unique numbers.



Below is a table of all the Ensembl Gene IDs that have been retired that might be mapped via ogs (official gene symbol) or Entrez ID. Notice that no retired mouse gene ids can be mapped via ogs or Entrez and that very few ogs humans can be mapped.

	Total	Matches	Ambiguous	No Matches
ogs human	374	5	0	369
entrez human	88	3	66	19
ogs mouse	24	0	0	24
entrez mouse	4	0	0	4
Total	490	8	66	416

With public gene sets - all retired Ensembl Gene IDs have been removed and any remapped Ensembl Gene IDs have been updated to the new Ensembl Gene ID. This is nothing for you to do, but it is something to recognise.

Conclusion and where to find help

Upgrading to the new version of Ensembl is a work in progress for us. We are trying to provide a more streamlined and informed way of letting you know what has changed as part of our Ensembl upgrades. If you have any questions, comments or suggestions, please email us at info@stemformatics.org