**A. Specific Aims**

The main objective of the pseudogenes study is to annotate and analyze the mouse pseudogenes in light of the sequence data available. The pseudogenes will be annotated using a rigorous approach combining manual annotation with automatic pipelines (PseudoPipe and Retrofinder). The resulting annotation will be integrated with functional genomics and evolutionary data to provide a full genetic map of mouse pseudogenes. The results will be collected into a mouse pseudogene resource and will be available online.

**B. Studies and Results**

We preformed a first round of annotations of mouse pseudogenes using the MM10 genome assembly, Ensembl 65 using an automatic pipeline annotation. Using 22,706 protein coding genes as the starting material we annotated 15,887 pseudogenes. Subsequent to the release of a newer Ensembl build (74), we improved our annotation using two automatic pipelines: PseudoPipe and Retrofinder. Both pipelines identified a comparable number of pseudogenes (18,029 PseudoPipe and 18,854 RetroFinder) with an agreement of 65% between the two annotated sets using at least 1 base pair overlap. While the number of protein coding genes did not vary substantially, the number of pseudogenes shows large fluctuations between the two releases. The sequence decay at pseudogene loci makes it challenging to rightly identify authentic pseudogenes and accurately define their boundaries. To this end we highlight the importance of combining automatic and manual annotation to correctly identify the mouse pseudogenes.

Using the PseudoPipe accuracy for the human pseudogene annotation (83%), we estimated the total number of pseudogenes in the mouse genome to be around 15,000.

Based on their mechanism of formation, we classified the pseudogenes into two categories: processed and duplicated. We found that 11% of the mouse pseudogenes are duplicated, while the rest are processed. This result is comparable with the biotype distribution obtained for the human pseudogenes (19/81 duplicated/processed pseudogenes). We expect that refining the automatic annotation will result in an even similar distribution between the biotypes of pseudogenes in mouse and human.

Next we looked at the distribution of pseudogene biotypes as function of age. We inferred the pseudogene age using the sequence similarity to parents as a timescale. We observed that regardless of their time of apparition the majority of pseudogenes are processed (Figure 1).

For the recently formed pseudogenes (>80% similarity to parents) we noticed a constant increase in the fraction of processed pseudogenes. This increase is comparable to the one observed for the human pseudogenes, marking the burst of retrotransposition events. While this event took place about 40Mya, (50 Million years after the human/mouse speciation), our result suggests potentially that the burst of retrotranspositon event is not human/primate specific but might be a more general, mammalian characteristic.

Figure 1. Distribution of the fraction of processed pseudogene as function of pseudogene age (CDS similarity to parents) for mouse and human.

Following this analysis, we looked at evolution of pseudogenes by identifying human-mouse pseudogene orthologs. We annotated two pseudogenes as orthologs if they are located in syntenic regions in the two species and if their respective parent genes are orthologs. Using these criteria we identified 129 orthologous pseudogenes in human and mouse. Surprisingly the majority of pseudogenes (127), are processed and share a high sequence similarity to their parents. This result suggests that these peculiar pseudogenes might have been created prior to the human/mouse speciation and were under some sort of a selection pressure. Another explanation would be that these pseudogenes are resulting from former paralogous genes that subsequent to human/mouse speciation suffered loss of function. A more detailed analysis of this pseudogene set is necessary for determining their provenience and evolutionary history.

Next we studied the pseudogene family distribution in mouse and human. While both species share the Ribosomal proteins as the main family type of pseudogenes, the top families are distinct for the two organisms. A detailed Pfam analysis shows that the top human family is 7TM (trans-membrane protein), reflecting perhaps the evolution of olfactory receptors while the top family in mouse is GAPDH.

Finally we integrated the annotation with functional genomics data in order to identify potentially transcribed pseudogenes. To this end we mapped RNA-seq data from 30 mouse tissues to the pseudogenes exons filtering out regions less than 100bp or with low mappability. We calculated the RPKM values for the resulting regions using only uniquely mapped reads. With a human RPKM pseudogene transcription threshold set at 2, we set the mouse transcription RPKM threshold at 3.28 based on the quantile normalization factor between the human and mouse protein coding expression profiles. Thus we obtained 878 pseudogene candidates for transcription in mouse. This equates to ~6% of the total mouse pseudogenes, half of the number obtained for human transcribed pseudogenes. The difference might be related to a higher rate in the loss of function events in the mouse genome.

**C. Significance**

The identification and analysis of pseudogenes is important not only in helping the annotation of the protein coding elements but also in understanding the evolution of the genome. While coding elements are conserved, the pseudogenes, most of them free of the selection pressure, are able to capture the key moments that shape the genome evolution.

**D. Plans**

Next year we aim to refine the pseudogene set using manual annotation. Also we aim to extend our pseudogene biotype differentiation creating a pipeline for the identification of unitary and polymorphic pseudogenes. To this end we will use an improved annotation of the mouse genome combined with the finished annotation of the human genome. Subsequently we plan to develop a variant annotation tool for mouse, making use of the improved annotation and the availability of the 17 published mouse strains and thus study the pseudogene variation in various mouse populations. Finally we aim to integrate this analysis with more functional data in order to identify potentially biologically active pseudogenes.