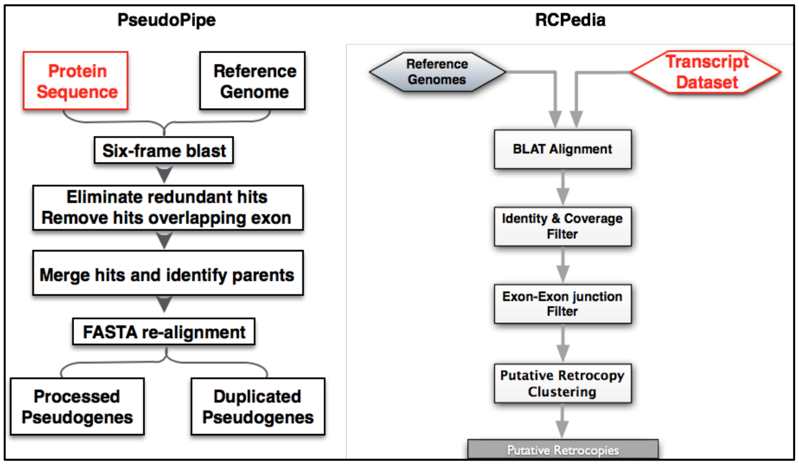
Page 19: [1] Deleted **Jin-rui Xu** **5/7/16 5:57 PM**

Figure 43: Automatic pseudogene annotation pipelines.



The Yale group has substantial experience in pseudogene annotation and analysis. In collaboration with the UCSC and Sanger Havana group, we have developed a variety of methods to identify pseudogenes \cite{16574694,16925835,22951037}.

Pseudopipe, Yale's in house automatic annotation pipeline, is fast and accurate \cite{22951037} (**Figure 3**). The pipeline takes as input all known protein sequences in the genome and using an homology search is able to identify disabled copies of functional paralogs (referred to as pseudogene parents). Based on their formation mechanism pseudogenes are classified into 3 different biotypes: processed, unprocessed and ambiguous. There is a good consensus overlap between the human pseudogene prediction set obtained with Pseudopipe and the set manually curated by the GENCODE annotators \cite{22951037}. Even more, the Pseudopipe predictions fueled the manual curation of pseudogenes in GENCODE \cite{22951037}.

RCPedia, Yale's newest pseudogene annotation pipeline focuses on the annotation of retrotransposed (processed) pseudogenes \cite{23457042} (**Figure 3**). This pipeline takes as input all known protein coding RNA transcripts and using sequence alignment is able to identify all possible retrocopies of functional genes. In the human genome there is an over 85% consensus between processed pseudogenes predicted by RCPedia and those annotated using Pseudopipe.

Retrofinder is the UCSC retrocopies annotation pipeline. Retrocopies can be functional genes that have acquired a promoter, non-functional pseudogenes, or transcribed pseudogenes. Retrofinder finds retroposed messenger RNAs (mRNAs) in genomic DNA \cite{18842134}. Candidate retrocopies overlapping by more than 50% with repeats identified by RepeatMasker \cite{16093699,Smit} and Tandem Repeat Finder \cite{9862982} are removed. Retrocopies are identified based on a score function using a weighted linear combination of features indicative of retrotransposition. These include: 1) Multiple contiguous exons with the parent gene introns removed; 2) Negatively scored introns as distinguished from repeat insertions (SVAs, LINEs, SINEs, Alus); 3) Lack of conserved splice sites; 4) Breaks in synteny with mouse and dog genomes (syntenic net alignments \cite{14500911}; and 5) Poly(A) tail insertion.

As a member of the GENCODE project, we used the pipelines to identify pseudogenes in human, mouse, worm, fly, and other model organisms \cite{16925835,22951037,25157146}. The identified pseudogenes with related genomic and epigenomic data are available in our online databases

\cite{17099229,18957444,22951037,25157146}. Moreover, using data from the 1000 Genomes Project in addition to the pseudogene annotation resulting from our pipelines, we were able to study the impact of pseudogene in human population variation. To this end we evaluated 2,504 individuals across 26 human populations and we investigated the impact of coding and non-coding structural variants in the human genome \cite{26432246}. We described processed pseudogenes as a novel class of gene copy number polymorphism that creates variability across populations. We were also able to associate their origin mechanism to cell division \cite{24026178}.

In order to record the structural and functional relationship between the pseudogenes within a family, we developed a **pseudogene ontology** \cite{20529940}. The pseudogene ontology is used in the generation of the GENCODE genomes annotation resource and is available, alongside many other tools for pseudogene analysis at the online pseudogene repository, **pseudogene.org** \cite{17099229}

Functional characterization We integrated ENCODE functional genomics data to obtain a comprehensive map of pseudogenes activity in human and other model organisms. We found that transcription signals have been observed for some pseudogenes and that the majority of pseudogenes (75% in human and 60% in worm and fly) have a large range in biochemical activity (e.g. presence of transcription factor or polymerase II binding sites in the upstream region, active chromatin, etc) (see Fig PG4). Moreover, we found 1441, 143, and 23 transcribed pseudogenes in human, worm, and fly, respectively. We also identified 878 transcribed pseudogenes in mouse and 31 in zebrafish. These numbers represent a fairly uniform fraction (~15%) of the total pseudogene complement in each organism reflecting the similarity across phyla observed in their transcriptomes.

Among transcribed pseudogenes, ~13% in human and ~30% in worm and fly have a discordant transcription pattern with their parent genes over multiple samples. A large fraction of pseudogenes are associated with a few highly expressed gene families, e.g. the ribosomal proteins in human \cite{25157146}.

The parent genes of broadly expressed pseudogenes tend to be broadly expressed as well, but the reciprocal statement is not valid. Specifically, only 5.1%, 0.69%, and 4.6% of the total number of pseudogenes are broadly expressed in human, worm, and fly, respectively. However, in general, transcribed pseudogenes show higher tissue specificity than protein-coding genes \cite{25157146}.

[[move to the personal annotation]] Loss of function analysis A loss-of-function (LOF) event is a genetic event that results in a severe disruption of the protein coding gene. Some LOFs can impact only one individual, resulting in the inactivation of an essential gene, which may lead to disease, while other LOFs can become fixed in the population as nonfunctional relics, through the pseudogenization process of the affected gene. The identification, analysis, and characterization of LOFs as either disease related or pseudogenization precursors is especially important in the era of personal genome annotation \cite{21205862}.

Moreover, the identification of pseudogenization/LOF events in mouse provides a very useful resource for understanding LOF in humans, by using mouse LOF phenotypes as proxy for human LOF events. To this end, the identification of mouse-specific unitary pseudogenes (regions that are functional in human and non functional in mouse) is important in highlighting human genes that can (have functional paralogs in mouse) or cannot (are paralogs to unitary pseudogenes in mouse) be studied in mouse models \cite{12909341,14746985}.

Taking advantage of the rich 1000 Genomes data, we have developed a tool, called Variant Annotation Tool (VAT) \cite{22743228}, to systematically annotate and catalogue LOF events in the human genome. This pipeline enables rapid and efficient annotation of genomic variations (SNPs, indels and SVs) with respect to a reference genome and a gene annotation model. VAT can be used to identify pseudogenization events such as premature STOPs as well as polymorphic pseudogenes where a pseudogene in the reference genome becomes functional in another genome due to genetic variability at the stop codon.

We applied our tools to the 1000 Genomes Phase 3 data and we were able to characterize putative LOF events from individuals belonging to 26 different populations. While earlier studies have suggested that on average the human genome contains ~100 genuine LOF variants resulting in the total disablements of ~20 genes \cite{22344438}, we found this number to be higher. On average the human genome contains 149–182 sites with protein truncating variants, ~11000 sites with peptide-sequence-altering variants, and around 500000 variant sites overlapping known regulatory regions (untranslated regions, promoters, enhancers, etc.) \cite{26432245}. Even more we were able to identify 24-30 sites per genome that are predicted severe disease-causing variants.

In a similar manner we surveyed the impact of LOFs on personal genome annotation \cite{21205862}. We found that LOFs variants that introduce premature STOPs resulting in a gene truncation in the reference genome can lead to an incorrect annotation of the gene. This highlights the importance of correct LOF identification for an accurate annotation. Finally we have studied the LOF events that results in a pseudogenization process. It is known that the loss-of-function in duplicated pseudogenes happens right after the gene duplication processes \cite{20615899}. To this end we have developed a pipeline to identify unitary pseudogenes in human \cite{20210993} and we explored the functional constraints faced by different species and the timescale of functional gene loss \cite{20210993}. All these results together with fully annotated sets of pseudogenes are deposited in our repository at pseudogene.org.

Page 36: [2] Deleted **Jin-rui Xu** **5/7/16 6:02 PM**

Currently we are in the midst of completing the mouse reference genome pseudogene annotation, with plans to develop customised pseudogene annotations for the available 18 strains. Using Pseudopipe we are able to identify 18627 putative pseudogenes in the reference genome (MM8, Ensembl 83) that are classified into three groups based on their biotypes as follows: 9748 processed pseudogenes, 1940 duplicated and 6939 ambiguous. Using RCPedia we are able to identify 9755 processed pseudogenes in the reference genome. Retrofinder predicts 18467 processed pseudogenes in the mouse reference genome. The tri-way consensus between the three pipelines with respect to the processed pseudogenes is ~80%. Integrating the automatic predictions with the manually curated pseudogenes we were able to annotate a comprehensive set of pseudogenes in the mouse reference genome. Table PG1 summarises the current state of pseudogene annotation.

Table PG1. Mouse reference genome MM8 pseudogene annotation. PSSD = Processed pseudogenes, DUP = duplicated pseudogenes, FRAG = ambiguous pseudogenes

Pipeline	**PSSD**	**PSSD Parents**	**DUP**	**DUP Parents**	**FRAG**	**FRAG Parents**	**Total**

Pseudopipe	9748	2581	1940	1146	6939	2884	18627
RCPedia	9755	2731	-	-	-	-	9755
Retrofinder	18467	-	-	-	-	-	18467

6.1.2 Status on Human-Mouse Pseudogene Comparison

Preliminary comparative analysis of human and mouse genomes have shown that while they are divergent enough to permit a reliable identification of species specific elements, they are also similar enough to allow a reliable comparative analysis \cite{14746985,25157146}. Comparing the two organisms we found that they exhibit a similar number of pseudogenes. The pseudogene complements of both human and mouse are dominated by processed pseudogenes. At family level we found that most of the pseudogenes are lineage specific and the majority of them arise from housekeeping genes (e.g. ribosomal proteins). Also, the age distribution of mouse processed pseudogenes closely resembles that of LINEs, in contrast to human, where the age distribution closely follows Alus (SINEs).

Annotating loss of function events in mouse Building on our previously developed human unitary pseudogene annotation pipeline \cite{20210993} we aim to develop a reliable framework for the identification of unitary pseudogenes across the 18 available strains. The unitary mouse pseudogene pipeline can be summarized as follows. First we will create a global inventory of orthologs between the mouse strains using the available multi sequence alignment data from UCSC. Next we are going to identify homologous regions between any two strains and annotate the syntenic ones. Finally, we will conduct a survey of gene disablements in the syntenic regions. We are going to use the available mouse genome variation data to filter our false positives.

In order to create a comprehensive set of polymorphic pseudogenes in mouse we will focus on annotating variants that change the strain genome stop codon to a functional allele across another mouse strains. For this we will use our previously developed pipeline VAT to annotate the SNPs of interests. Next we will extend the VAT to annotate frame shifts that revert a stop codon across two mouse strains.

We will build on our experience in developing variant annotation tools to create a pipeline that will provide an extensive annotation of putative LOF variants. We will include variants causing premature stop codons, canonical splice-site disruption and frameshift-causing indels as putative LOF variants. The pipeline will feature (1) function-based annotations; (2) evolutionary conservation; and (3) biological network data. For comprehensive functional annotation we will integrate several annotation resources such as PFAM and SMART functional domains, signal peptide and transmembrane annotations, post-translational modification sites, NMD prediction, and structure-based features such as SCOP domains and disordered residues. For evolutionary conservation, our pipeline will output variant position-specific GERP scores, which is a measure of evolutionary conservation and dN/dS values. In addition, we will evaluate if the region removed due to the truncation of the coding sequence is evolutionarily conserved based on GERP and PhyloCSF constrained elements. Our model will also include network features to predict disease causing variants: we will use a proximity parameter that gives the number of disease genes connected to a gene in a protein-protein interaction network and the shortest path to the nearest disease gene. The pipeline will also include features to help identify erroneous LOF calls, potential mismapping, and annotation errors, because LOF variant calls have been shown to be enriched for annotation and sequencing artifacts.

To understand the impact of putative LOF variants on gene function we will develop a prediction model to classify premature stop causing variants into those that are benign, those that lead to recessive disease and those that lead to dominant disease using the annotation output as predictive features. To build the classifier, we will use benign homozygous premature stop variants, dominant heterozygous and recessive homozygous disease premature stop mutations. We will build the classifier to distinguish among the three classes and will provide class probability estimates for each mutation. To validate our classifier we plan to use LOF from Mendelian Diseases, Cancer samples and healthy control datasets such as 1000 Genomes and ExAC. The classifier will also be applied to variants in mouse strains and will be used to classify variations into Loss-of-Function, Gene Death and Pseudogenization.

In a similar manner we surveyed the impact of LOFs on personal genome annotation \cite{21205862}. We found that LOFs variants that introduce premature STOPs resulting in a gene truncation in the reference genome can lead to an incorrect annotation of the gene. This highlights the importance of correct LOF identification for an accurate annotation. Finally we have studied the LOF events that results in a pseudogenization process. It is known that the loss-of-function in duplicated pseudogenes happens right after the gene duplication processes \cite{20615899}. To this end we have developed a pipeline to identify unitary pseudogenes in human \cite{20210993} and we explored the functional constraints faced by different species and the timescale of functional gene loss \cite{20210993}. All these results together with fully annotated sets of pseudogenes are deposited in our repository at pseudogene.org.

Annotating pseudogene activity We are going to leverage our experience in pseudogene transcription analysis in human, worm and fly to study the pseudogene transcription in mouse, significantly improving on previous efforts. We are going to focus on identifying tissue and strain specific transcriptionally active pseudogenes. In particular we will highlight pseudogenes that have a high coexpression correlation with their parents or are differentially transcribed with respect to their functional paralogs.

Our approach is summarized in the following steps: 1) calculate the genome mappability in each mouse strain; 2) remove low mappability regions from the pseudogene annotation in each mouse strain; 3) calculate the RPKM value of each pseudogene based on the RNA-Seq reads mapped to the remaining high mappability regions in that pseudogene locus; 4) quantile normalize the transcriptome signals across all the mouse strains and identify transcribed pseudogenes uniformly in each strain. We will combine the pseudogene transcription results with their annotation to study the strain-specificity of pseudogenes. For example, a pseudogene may exist and be transcribed in only one mouse strain, or a pseudogene may exist in all mouse strains but be transcribed in only one or a few closely related strains. Such information and data resources will benefit the evolutionary studies on mouse and the comparative studies between mouse, human, and other model organisms.

We aim to integrate tissue specific transcription information and regulatory data with the pseudogene annotation in order to characterize pseudogene activity. In particular, we will focus on the transcriptomics (ENCODE, BrainSpan, TCGA), epigenomics (ENCODE, Roadmap Epigenomics) and cis-regulatory interactions data (GTEx, PsychENCODE). These datasets will allow us to provide annotation on tissue-specific pseudogene transcription and tissue-specific pseudogene regulation. Such information will be valuable for understanding the biological consequences from the pseudogene activities, such as the regulatory mechanisms of pseudogene transcription, and whether transcribed pseudogenes may perform regulatory roles through interaction with their functional paralogs.

Page 36: [3] Deleted Jin-rui Xu 5/7/16 6:11 PM

The human reference genome is a haploid sequence derived as a composite from multiple individuals. Current genome annotations are based on this reference and as such do not provide an accurate representation for the large genomic diversity of the human population. We have developed a computational tool, *vcf2diploid*\cite{21811232}, which integrates an individual's variation data (SNVs, indels, and SVs) into the reference genome producing the maternal and paternal haplotypes of the individual's *personal genome* (see Fig PG6).

The tool's versatility to account for coordinate offsets between the reference and personal genome and to convert between them facilitates mapping of genomic annotated regions between the genomes. Thus, *personal annotation* can be generated by mapping GENCODE annotations against the individual's personal genome. Using personal annotation in downstream analyses allows to account for differences due to impact of the personal variation on genes and other genomic elements between individuals as well as between haplotypes of the same individual.

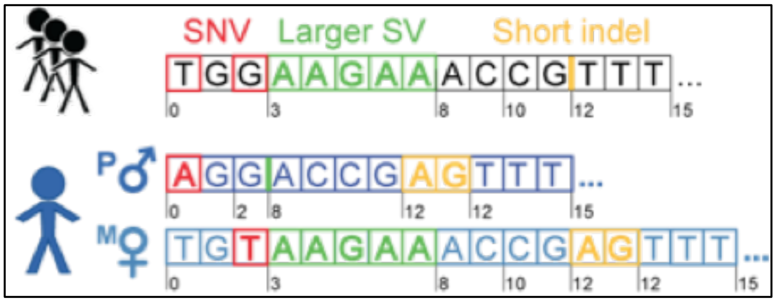


Figure PG6. Each haplotype in the diploid personal genome is derived by incorporating phased or unphased variants (SNVs, indels and SVs) into the human reference genome. The coordinates can be mapped back to the human reference coordinates to facilitate comparisons with other reference-based resources, such as gene annotations from GENCODE.

We have a large experience with building personal genomes and annotations and using them in functional genomic analyses. We have previously constructed the personal diploid genome, splice-junction libraries and personalized gene annotations for NA12878. We have made this assembly available as a resource - alleleseq.gersteinlab.org - and have been updating it as new versions of the human reference genome, genomic annotations, and NA12878 genetic variation data are released.

It has been demonstrated that using the diploid genome with individual's variants improves both mappability of the reads \cite{21811232} and the results of the downstream analyses \cite{26432246}. In particular, it was shown that using personal genome and annotation for NA12878 as opposed to the standard reference affected estimated expression of hundreds (525) of exons \cite{21811232}.

Using personal genomes in analyses involving mapping of functional assay reads alleviates known biases associated with short read alignment to the reference genome: reduced mappability in regions with higher genomic variation and the preferential mapping of reads bearing the reference allele. Allele-specific analyses are particularly sensitive to these biases. For this purpose, the initial step of our *AlleleSeq* pipeline\cite{21811232} involves construction of the personal diploid genome. We have spearheaded allele-specific analyses in several major consortia publications, including ENCODE and the 1000 Genomes Project \cite{22955620,22955619,24092746}. The availability of an efficient computational tool enables construction of personal genomes and annotation in a high throughput fashion, as demonstrated in a recent publication \cite{27089393} where we provided allele-specific annotation of variants in 382 individuals.

Page 38: [4] Deleted **Jin-rui Xu** **5/7/16 6:16 PM**

Annotation based on the current human reference genome does not account for variation in the number of functional genes between people and does not provide an accurate and complete set of an individual's genes and other genomic elements. We propose to use the diploid personal genome and personal annotation as a more accurate representation of an individual's genome. Individual variations can affect gene annotations and thus sequence variations need to be taken into account for annotation purposes.

For this we will develop a personal genome annotation resource containing a number of tools and utilities for constructing a personal genome and for creating a personal GENCODE annotation set. We will incorporate a personal genome construction step into our genome and variant annotation pipelines. Using well-characterized public genomes as well as matching variant calls and functional datasets, we will evaluate the impact of using personal genome and annotations for various types of genomic analyses. For these genomes we will generate a reference set of personal diploid (haplotype-resolved) annotations and make them publicly available.

In particular, given an individual's variation data, the proposed annotation resource can be used to identify and further analyse GENCODE-annotated features characteristic to the individual, such as his/her distinctive set of functional genes or structures of variant-affected transcripts. Using our automated in house annotation pipelines we are going to create a comprehensive personal pseudogene complement. We will use the newly constructed personal annotations to identify LOF and pseudogenization events by comparison with the reference genome. We are going to assess the annotated personal SNPs for allele specific expression using the data from AlleleDB \cite{27089393}, an online repository that provides genomic annotation of cis-regulatory single nucleotide variants associated with allele-specific binding and expression.

Next, by integrating Mendelian disease and cancer data we will be able to filter the LOF and pseudogenization variants and characterize them with respect to their disease driver potential. In particular we are going to use VAT and the newly proposed LOF analysis pipeline as described above (See Sections 4 and 8.2). Further, we are going to test the presence of mouse orthologs for all the curated genes affected by LOFs and pseudogenization variants, determining whether or not the mouse is a suitable model organism to study them.

We aim to integrate all the personal annotation tools in an online framework that can easily be applied to newly sequenced individual genomes.

Page 38: [5] Moved to page 43 (Move #1)Cristina Sisu **5/8/16 9:55 PM**

Improving the mouse strain pseudogene annotation The relatively small divergence time frame between the mouse strains \cite{25038446} allows us to map the reference mouse annotation on each of the strains using the UCSC LiftOver tool. In addition, we will develop extensions to the available annotation pipeline to use the UCSC strain dependent protein coding annotation as input in order to draft each strain's pseudogene complement. Using these two annotation sets will allow us to produce an accurate map of pseudogenes and loss-of-function events in mouse strains.

Page 38: [6] Deleted	Cristina Sisu	5/8/16 9:56 PM

Improving the mouse strain pseudogene annotation The relatively small divergence time frame between the mouse strains \cite{25038446} allows us to map the reference mouse annotation on each of the strains using the UCSC LiftOver tool. In addition, we will develop extensions to the available annotation pipeline to use the UCSC strain dependent protein coding annotation as input in order to draft each strain's pseudogene complement. Using these two annotation sets will allow us to produce an accurate map of pseudogenes and loss-of-function events in mouse strains.

[[CSDS2MG I think that the paragraph above ~Improving the mouse strain pgene annotation~ will be more suitable in the **Plans to scale up the curation process (~2.5 pages)**

Extending Automatic Pseudogene Annotation Pipelines section below]][[OK make the change]]

Page 38: [7] Deleted	Jin-rui Xu	5/7/16 6:17 PM

Improving the Mouse Reference and Mouse Strain Pseudogene Annotation Using the GENCODE manually annotated pseudogenes as a gold standard we plan to determine the annotation accuracy of our automatic pipelines Pseudopipe and RCPedia. Next we are going to use the false positives to refine and improve the pseudogene identification process as well as the biotype assignment.

Leveraging on the availability of the improved 18 mouse strain assemblies, we are going to extend our automatic annotation pipelines to annotate the pseudogenes in these strains. The 18 available mouse strains are 129S1_SvImJ, AKR_J, A_J, BALB_cJ, C3H_HeJ, C57BL_6NJ, CAROLI_EiJ, CAST_EiJ, CBA_J, DBA_2J, FVB_NJ, LP_J, NOD_ShiLtJ, NZO_HILtJ, PWK_PhJ, Pahari_EiJ, SPRET_EiJ, WSB_EiJ.

The evolutionary distance between these strains ranges from 400,000 to 2,000 years \cite{25038446}. The relatively small divergence time frame will allow us to map the reference mouse annotation on each of the strains. To this end we are going to use the UCSC LiftOver tool to align the reference genome annotated pseudogenes to each of the strains. In addition, the UCSC group is currently in the midst of completing a first draft of the mouse strain specific annotation of protein coding genes. We are going to use these two draft annotation sets as a support structure in building each strain's pseudogene complement. Overall this will allow us to produce a better and more accurate pseudogene annotation and will facilitate our identification of conserved elements and loss-of-function events in mouse strains.

Next, we are going to develop extensions to the available in house automatic annotation pipelines in order to use them in predicting pseudogene models in the mouse strains. The details of the proposed pipeline updates are described below.

Page 43: [8] Deleted	Jin-rui Xu	5/7/16 6:19 PM

Given the close evolutionary time scale between the strains, we expect that the expressed protein amino acid sequence to be preserved across all strains for the conserved protein coding genes. As such we are going to use the conserved protein coding genes between each strain and the reference genome as input for identifying pseudogenes. The extended Pseudopipe workflow is summarized in the following steps:

1) Identify the consensus protein coding genes between strain and reference; 2) extract the amino acid sequence of the conserved proteins from the available ENSEMBL peptide database for the mouse reference genome; 3) mark and identify the coordinates of the consensus protein coding genes in the analyzed strain; 4) use a six frame blast homology search to match the consensus peptides to the strain sequence; 5) refine results and eliminate redundant hits (e.g. remove matches that overlap protein coding exons); 6) merge hits and identify parents; 7) align parents and pseudogenes and check for the presence of disablements (e.g. frameshifts, premature stop codons); 7) assign pseudogene biotype.

Similarly to the protein sequence approach, transcript sequence is expected to be conserved at short evolutionary time scales. RCPedia will be adapted to integrate gold standard transcript annotation, such as GENCODE mouse annotation, and annotations based on the strain genome. The extended RCPedia pipeline is summarized as follow:1) Merge multiple annotations of parental transcripts using an hierarchical prioritization; 2) align transcripts sequences to the target genome and extract alignments blocks and their distances; 3) select alignments containing mostly intronless blocks; 4) refine results removing alignments with most of the parental introns and remove putative genomic duplications; 5) merge call sets and select the most likely parent transcript; 7) calculate properties of the putative pseudogene such as target site duplication sequence, identity and polyA length.

Both Pseudopipe and RCPedia pipelines are broadly used by the pseudogene research community and both are available through our online resource pseudogene.org. Collectively they use many different standalone tools such as aligners, toolsets and well established annotation software such as repeatmasker. The invariably complex environment necessary to install and configure these pipelines can create difficulties to the end user. In order to mitigate dependency and compatibility issues, we plan to create docker images for both pipelines and make them publicly available after the mentioned extension. Docker images will contain all dependencies necessary to set up Pseudopipe and RCPedia as well as all in house scripts. Parameterization and fine tuning will be made by a single configuration file editable by user. We will also create amazon machine images (AMI) compatible with Amazon AWS and other major cloud services so users can easily annotate additional genomes.