SUMMARY OF EXPERIENCE IN ANNOTATING AND ANALYSING PSEUDOGENES

Pseudogenes have long been considered non-functional elements. However, recent studies indicate that pseudogenes can be transcribed, translated and can play key regulatory roles. In particular pseudogenes can regulate the expression of functional protein coding genes by serving as a source of siRNAs, antisense transcripts, microRNA binding sites, or competing mRNAs [1, 2, 3]. The pseudogenization process is also closely linked to loss-of-function (LOF) events such as premature truncation of proteins, disruption of splicing and loss-of-functional or structural domains [4, 5, 6]. Finally, the annotation of pseudogenes is important in the analysis of personal genomes, providing a means to avoid errors in genotyping assays and variant calling.

Pseudogenes are defined as disabled copies of functional genes. Depending on their formation mechanism they can be referred to as unprocessed (originating through a gene duplication event) or processed (originating through a retrotransposition event). A functional gene may also become a pseudogene by acquiring a disabling mutation, if its function no longer confers a fitness advantage to the organism due to a change in the environment or genetic background. Such pseudogenes are called unitary pseudogenes. Pseudogenes provide valuable opportunities to study the dynamics and evolution of gene functions.

PRELIMINARY RESULTS & EXPERIENCE WITH PSEUDOGENE ANNOTATION

Pseudogene Annotation Pipelines

Gerstein lab has substantial experience in pseudogene annotation and analysis. In collaboration with the UCSC and Sanger groups, we have developed a variety of methods to identify pseudogenes [7, 8, 9].
Pseudopipe, our in house automatic annotation pipeline, is fast and accurate [9] (See Fig PG1). 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 [9]. Even more, the Pseudopipe predictions fueled the manual curation of pseudogenes in GENCODE [9].

RCPedia, our newest pseudogene annotation pipeline focuses on the annotation of retrotransposed (processed) pseudogenes [10] (see FIG PG1). 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.

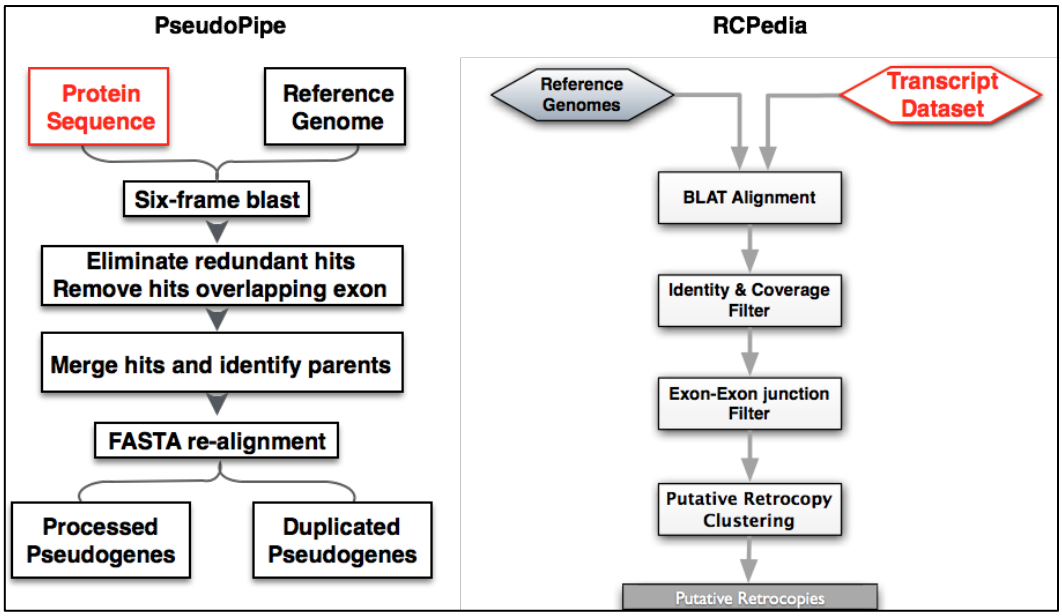


FIG PG1. Automatic pseudogene annotation pipelines.

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 [11]. Candidate retrocopies overlapping by more than 50% with repeats identified by RepeatMasker [12, 13] and Tandem Repeat Finder [14] 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 [15]; and 5) Poly(A) tail insertion.

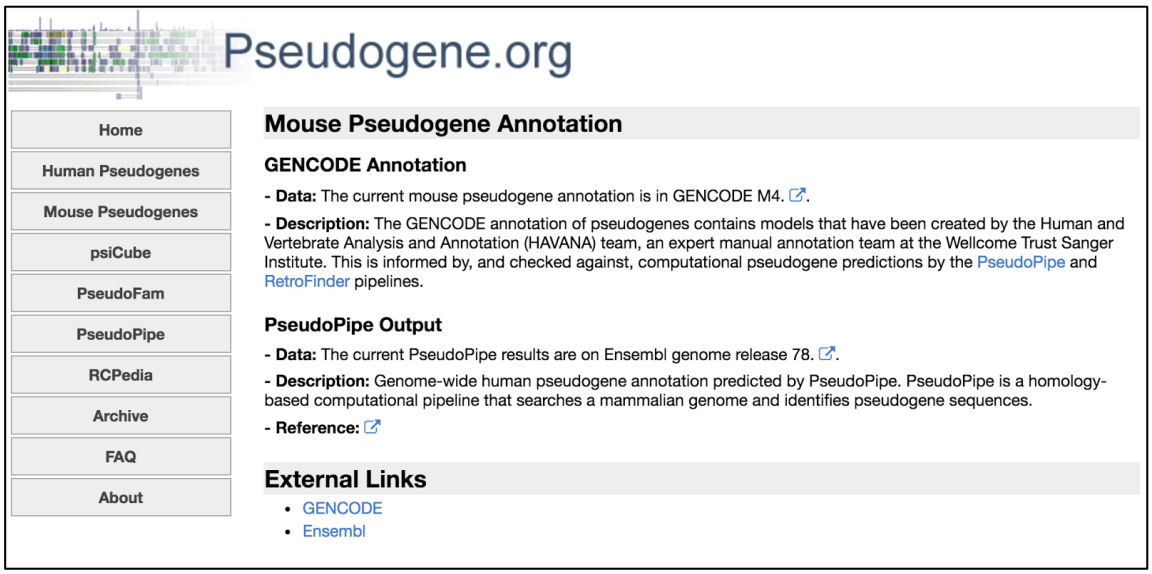
As a member of the GENCODE project, we used the pipelines to identify pseudogenes in human, mouse, worm, fly, and other model organisms [8, 9, 16]. The identified pseudogenes with related genomic and epigenomic data are available in our online databases [9, 16, 17, 18]. Moreover, using data from the 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 [19]. 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 [4].

Online Resources for Pseudogene Annotation and Analysis

Our experience in annotating and analyzing pseudogenes spans over a decade. Thus, we have built a number of tools to organize and analyze the available pseudogene data in a consistent and efficient manner.

We have built an online pseudogene repository, **pseudogene.org** [17] (see Fig PG2), that provides information regarding annotation and functional characterization of pseudogenes. Currently pseudogene.org hosts the human (**psiDR** [9]), and mouse pseudogene resources. It also provides a comparative pseudogene resource, **psiCUBE** [16], focused on cross species annotation and analysis of pseudogenes in a variety of model organisms. Both psiDR and psiCUBE also provide information regarding evolutionary and functional characterization of pseudogenes in the curated genomes.

Pseudogene.org also hosts **Pseudofam** [18], the pseudogene family database. Pseudofam resources focus on clustering pseudogenes into families based on their functional homolog protein family. Currently there are 10 eukaryotic genomes including human and mouse. Pseudofam also contains segmental duplication information associated with the human pseudogene dataset.



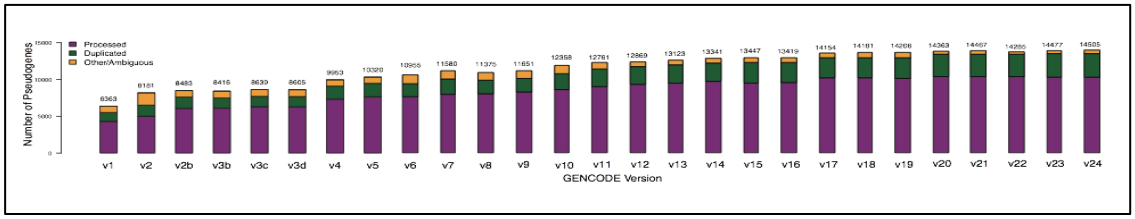
PG2. Pseudogene.org interface linking the available pseudogene tools and resources

In order to record the structural and functional relationship between the pseudogenes within a family, we developed a **pseudogene ontology** [20]. The pseudogene ontology is used in the generation of the GENCODE genomes annotation resource.

Current Results on Pseudogenes as Part of GENCODE in Human and Model Organisms

Our experience with annotating pseudogenes spans more than fifteen years. Over time we have annotated and reviewed pseudogenes in a variety of species ranging from prokaryotic organisms (archaea and bacteria) [21, 22], to yeast [23, 24], plants [25], worm [26], fly [27, 28], and a wide range of vertebrates (e.g. zebrafish, mouse, rat, chimp, and human) [24, 29, 30, 31, 32]. Our involvement in the GENCODE project started over a decade ago and ever since we have led and contributed to the identification and characterization of pseudogenes in human and model organisms (see Fig PG3).

Leveraging on the completed annotation of protein coding genes in human, worm and fly we were able to provide the complete and comprehensive set of pseudogenes in these organisms. In order to elucidate the role played by pseudogene in genome biology we integrated the annotation data with variation and functional genomics information.



PG3. GENCODE human pseudogene distribution in various releases.

In this respect we identified 14505 pseudogenes in human, 911 in worm, and 145 in fly [9, 16]. A close comparison of the three genomes shows that pseudogenes complements do not follow the genome size or the number of protein coding genes in the genome, highlighting the species specific evolution of pseudogenes. This specificity is also reflected at pseudogene

biotype level, where processed pseudogenes resulting from the burst of retrotransposition events that occurred at the dawn of primate lineage dominate the mammalian genomes, while the smaller fraction of duplicated pseudogenes hints at shared ancestry with more distant species [16].

We conducted a systematic analysis of human pseudogenes focusing on large groups of pseudogenes such as ribosomal pseudogenes [24, 29, 33], unitary [34] and polymorphic pseudogenes. The latter are a peculiar class of pseudogenes with a dual behaviour – their sequence is disabled in the reference genome but in some individuals, it encodes a functional gene. We conducted a comprehensive review of polymorphic pseudogenes [6].

Despite the presence of disabling mutations such as premature stop codons, loss of promoters in the upstream sequence, numerous studies have shown that pseudogenes can be transcribed and even translated [35, 36, 37, 38, 39].

PRELIMINARY RESULTS AND EXPERIENCE WITH FUNCTIONAL CHARACTERIZATION OF PSEUDOGENES

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 1,441, 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.

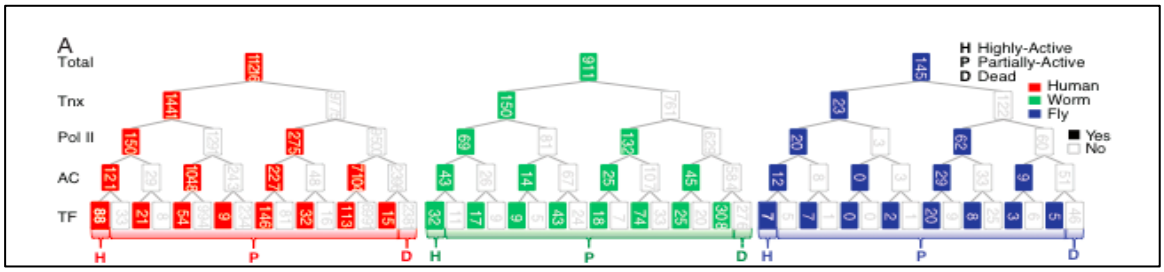


Fig PG4. Pseudogene activity. Distribution of pseudogenes as a function of various activity features: transcription (Tnx),
active chromatin (AC), and presence of active Pol II and TF binding sites in the upstream region.

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 [16].

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 [16].

We have also investigated pseudogene transcription by using the RNA-Seq data from the Illumina Human BodyMap data across 16 different tissues. Amongst all the transcribed pseudogenes identified, only a tiny proportion (~3%) are transcribed in all the 16 tissues,

while the transcription of all the other pseudogenes show different degrees of tissue specificity. Furthermore, more than 50% of the transcribed pseudogenes are transcribed in one tissue only. While testis contained the largest number of transcribed pseudogenes, skeletal muscle contained the least [9].

PRELIMINARY RESULTS & EXPERIENCE WITH ANNOTATION AND ANALYSIS OF LOSS-OF-FUNCTION EVENTS

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 [6].

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 [31, 40].

Taking advantage of the rich 1000 Genomes data, we have developed a tool, called Variant Annotation Tool (VAT) [41] (see Fig PG5), 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.

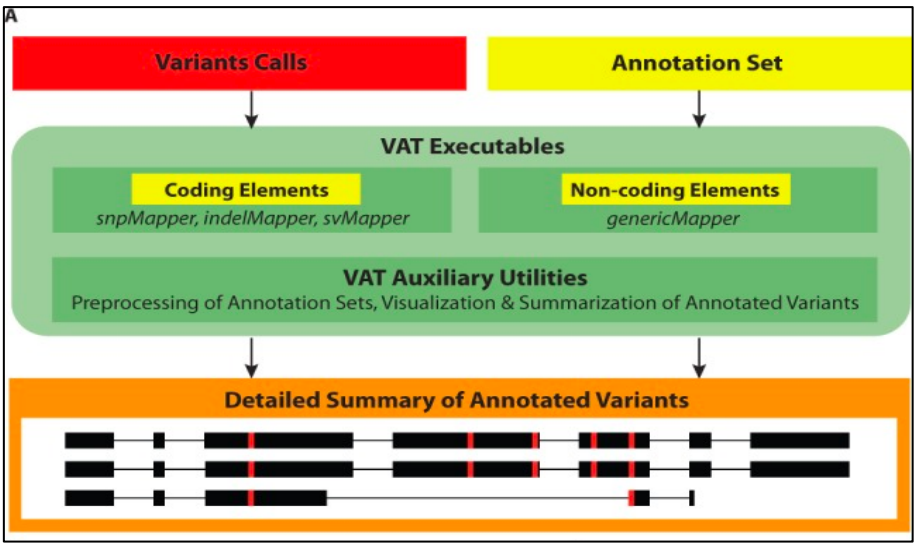


FIG PG5. Variant annotation tool (VAT) architecture.

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 [5], we found this number to be higher. On average the human genome contains 149–182 sites with protein truncating variants, ~11,000 sites with peptide-sequence-altering variants, and around 500,000 variant sites overlapping known regulatory regions (untranslated regions, promoters, enhancers, etc.)

[42]. 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 [6]. 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 [43]. To this end we have developed a pipeline to identify unitary pseudogenes in human [34] and we explored the functional constraints faced by different species and the timescale of functional gene loss [34]. All these results together with fully annotated sets of pseudogenes are deposited in our repository at pseudogene.org.

1. Kalyana-Sundaram, S., Kumar-Sinha, C., Shankar, S., Robinson, D. R., Wu, Y.-M., Cao, X., Asangani, I. A., Kothari, V., Prensner, J. R., Lonigro, R. J., Iyer, M. K., Barrette, T., Shanmugam, A., Dhanasekaran, S. M., Palanisamy, N., & Chinnaiyan, A. M. (2012) Expressed pseudogenes in the transcriptional landscape of human cancers. *Cell* 149, 1622-34, PMID:22726445.
2. Swami, M. (2010) Epigenetics: Demethylation links cell fate and cancer. *Nat Rev Cancer* 10, 740, PMID:21080588.
3. Poliseno, L. (2012) Pseudogenes: newly discovered players in human cancer. *Sci Signal* 5, re5, PMID:22990117.
4. Abyzov, A., Iskow, R., Gokcumen, O., Radke, D. W., Balasubramanian, S., Pei, B., Habegger, L., 1000 Genomes Project Consortium, Lee, C., & Gerstein, M. (2013) Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. *Genome Res* 23, 2042-52, PMID:24026178.
5. MacArthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J. K., Montgomery, S. B., Albers, C. A., Zhang, Z. D., Conrad, D. F., Lunter, G., Zheng, H., Ayub, Q., DePristo, M. A., Banks, E., Hu, M., Handsaker, R. E., Rosenfeld, J. A., Fromer, M., Jin, M., Mu, X. J., Khurana, E., Ye, K., Kay, M., Saunders, G. I., Suner, M.-M., Hunt, T., Barnes, I. H. A., Amid, C., Carvalho-Silva, D. R., Bignell, A. H., Snow, C., Yngvadottir, B., Bumpstead, S., Cooper, D. N., Xue, Y., Romero, I. G., Wang, J., Li, Y., Gibbs, R. A., McCarroll, S. A., Dermitzakis, E. T., Pritchard, J. K., Barrett, J. C., Harrow, J., Hurles, M. E., Gerstein, M. B., & Tyler-Smith, C. (2012) A systematic survey of loss-of-function variants in human protein-coding genes.. *Science* 335, 823-828, PMID:22344438.
6. Balasubramanian, S., Habegger, L., Frankish, A., MacArthur, D. G., Harte, R., Tyler-Smith, C., Harrow, J., & Gerstein, M. (2011) Gene inactivation and its implications for annotation in the era of personal genomics.. *Genes Dev* 25, 1-10, PMID:21205862.
7. Zhang, Z., Carriero, N., Zheng, D., Karro, J., Harrison, P. M., & Gerstein, M. (2006) PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics* 22, 1437-9, PMID:16574694.
8. Zheng, D. & Gerstein, M. B. (2006) A computational approach for identifying pseudogenes in the ENCODE regions. *Genome Biol* 7 Suppl 1, S13.1-10, PMID:16925835.
9. Pei, B., Sisu, C., Frankish, A., Howald, C., Habegger, L., Mu, X. J., Harte, R., Balasubramanian, S., Tanzer, A., Diekhans, M., Reymond, A., Hubbard, T. J., Harrow, J., & Gerstein, M. B. (2012) The GENCODE pseudogene resource. *Genome Biol* 13, R51, PMID:22951037.
10. Navarro, F. C. P. & Galante, P. A. F. (2013) RCPedia: a database of retrocopied genes. *Bioinformatics* 29, 1235-7, PMID:23457042.
11. Baertsch, R., Diekhans, M., Kent, W. J., Haussler, D., & Brosius, J. (2008) Retrocopy contributions to the evolution of the human genome. *BMC Genomics* 9, 466, PMID:18842134.

12. Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110, 462-7, PMID:16093699.
13. Smit, A. F. A., Hubley, R., & Green, P. (1996-2010) RepeatMasker Open-3.0. *http://www.repeatmasker.org* , , PMID:.
14. Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27, 573-80, PMID:9862982.
15. Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W., & Haussler, D. (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A* 100, 11484-9, PMID:14500911.
16. Sisu, C., Pei, B., Leng, J., Frankish, A., Zhang, Y., Balasubramanian, S., Harte, R., Wang, D., Rutenberg-Schoenberg, M., Clark, W., Diekhans, M., Rozowsky, J., Hubbard, T., Harrow, J., & Gerstein, M. B. (2014) Comparative analysis of pseudogenes across three phyla. *Proc Natl Acad Sci U S A* 111, 13361-6, PMID:25157146.
17. Karro, J. E., Yan, Y., Zheng, D., Zhang, Z., Carriero, N., Cayting, P., Harrrison, P., & Gerstein, M. (2007) Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Res* 35, D55-60, PMID:17099229.
18. Lam, H. Y. K., Khurana, E., Fang, G., Cayting, P., Carriero, N., Cheung, K.-H., & Gerstein, M. B. (2009) Pseudofam: the pseudogene families database. *Nucleic Acids Res* 37, D738-43, PMID:18957444.
19. Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Paolo Casale, F., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., Malig, M., Chaisson, M. J. P., Walter, K., Meiers, S., Kashin, S., Garrison, E., Auton, A., Lam, H. Y. K., Jasmine Mu, X., Alkan, C., Antaki, D., Bae, T., Cerveira, E., Chines, P., Chong, Z., Clarke, L., Dal, E., Ding, L., Emery, S., Fan, X., Gujral, M., Kahveci, F., Kidd, J. M., Kong, Y., Lameijer, E.-W., McCarthy, S., Flicek, P., Gibbs, R. A., Marth, G., Mason, C. E., Menelaou, A., Muzny, D. M., Nelson, B. J., Noor, A., Parrish, N. F., Pendleton, M., Quitadamo, A., Raeder, B., Schadt, E. E., Romanovitch, M., Schlattl, A., Sebra, R., Shabalin, A. A., Untergasser, A., Walker, J. A., Wang, M., Yu, F., Zhang, C., Zhang, J., Zheng-Bradley, X., Zhou, W., Zichner, T., Sebat, J., Batzer, M. A., McCarroll, S. A., 1000 Genomes Project Consortium, Mills, R. E., Gerstein, M. B., Bashir, A., Stegle, O., Devine, S. E., Lee, C., Eichler, E. E., & Korbel, J. O. (2015) An integrated map of structural variation in 2,504 human genomes. *Nature* 526, 75-81, PMID:26432246.
20. Holford, M. E., Khurana, E., Cheung, K.-H., & Gerstein, M. (2010) Using semantic web rules to reason on an ontology of pseudogenes. *Bioinformatics* 26, i71-8, PMID:20529940.
21. Liu, Y., Harrison, P. M., Kunin, V., & Gerstein, M. (2004) Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes.. *Genome Biol* 5, R64, PMID:15345048.
22. Harrison, P. M., Carriero, N., Liu, Y., & Gerstein, M. (2003) A "polyORFomic" analysis of prokaryote genomes using disabled-homology filtering reveals conserved but undiscovered short ORFs.. *J Mol Biol* 333, 885-892, PMID:14583187.
23. Harrison, P., Kumar, A., Lan, N., Echols, N., Snyder, M., & Gerstein, M. (2002) A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution.. *J Mol Biol* 316, 409-419, PMID:11866506.
24. Zhang, Z. L., Harrison, P. M., & Gerstein, M. (2002) Digging deep for ancient relics: a survey of protein motifs in the intergenic sequences of four eukaryotic genomes.. *J Mol Biol* 323, 811-822, PMID:12417195.
25. Harrison, P. M. & Gerstein, M. (2002) Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. *J Mol Biol* 318, 1155-74, PMID:12083509.
26. Harrison, P. M., Echols, N., & Gerstein, M. B. (2001) Digging for dead genes: an analysis of the characteristics of the pseudogene population in the Caenorhabditis elegans genome.. *Nucleic Acids Res* 29, 818-830, PMID:11160906.
27. Echols, N., Harrison, P., Balasubramanian, S., Luscombe, N. M., Bertone, P., Zhang, Z., & Gerstein, M. (2002) Comprehensive analysis of amino acid and nucleotide composition in

eukaryotic genomes, comparing genes and pseudogenes.. *Nucleic Acids Res* 30, 2515-2523, PMID:12034841.
28. Harrison, P. M., Milburn, D., Zhang, Z., Bertone, P., & Gerstein, M. (2003) Identification of pseudogenes in the Drosophila melanogaster genome.. *Nucleic Acids Res* 31, 1033-1037, PMID:12560500.
29. Liu, Y.-J., Zheng, D., Balasubramanian, S., Carriero, N., Khurana, E., Robilotto, R., & Gerstein, M. B. (2009) Comprehensive analysis of the pseudogenes of glycolytic enzymes in vertebrates: the anomalously high number of GAPDH pseudogenes highlights a recent burst of retrotrans-positional activity.. *BMC Genomics* 10, 480, PMID:19835609.
30. Balasubramanian, S., Harrison, P., Hegyi, H., Bertone, P., Luscombe, N., Echols, N., McGarvey, P., Zhang, Z., & Gerstein, M. (2002) SNPs on human chromosomes 21 and 22 - analysis in terms of protein features and pseudogenes.. *Pharmacogenomics* 3, 393-402, PMID:12052146.
31. Zhang, Z. & Gerstein, M. (2003) The human genome has 49 cytochrome c pseudogenes, including a relic of a primordial gene that still functions in mouse.. *Gene* 312, 61-72, PMID:12909341.
32. Zhang, Z. D., Cayting, P., Weinstock, G., & Gerstein, M. (2008) Analysis of nuclear receptor pseudogenes in vertebrates: how the silent tell their stories.. *Mol Biol Evol* 25, 131-143, PMID:18065488.
33. Balasubramanian, S., Zheng, D., Liu, Y.-J., Fang, G., Frankish, A., Carriero, N., Robilotto, R., Cayting, P., & Gerstein, M. (2009) Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes. *Genome Biol* 10, R2, PMID:19123937.
34. Zhang, Z. D., Frankish, A., Hunt, T., Harrow, J., & Gerstein, M. (2010) Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates.. *Genome Biol* 11, R26, PMID:20210993.
35. Harrison, P. M., Zheng, D., Zhang, Z., Carriero, N., & Gerstein, M. (2005) Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. *Nucleic Acids Res* 33, 2374-83, PMID:15860774.
36. Svensson, O., Arvestad, L., & Lagergren, J. (2006) Genome-wide survey for biologically functional pseudogenes. *PLoS Comput Biol* 2, e46, PMID:16680195.
37. Zheng, D., Zhang, Z., Harrison, P. M., Karro, J., Carriero, N., & Gerstein, M. (2005) Integrated pseudogene annotation for human chromosome 22: evidence for transcription. *J Mol Biol* 349, 27-45, PMID:15876366.
38. Zheng, D., Frankish, A., Baertsch, R., Kapranov, P., Reymond, A., Choo, S. W., Lu, Y., Denoeud, F., Antonarakis, S. E., Snyder, M., Ruan, Y., Wei, C.-L., Gingeras, T. R., Guigo, R., Harrow, J., & Gerstein, M. B. (2007) Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution.. *Genome Res* 17, 839-851, PMID:17568002.
39. Frith, M. C., Wilming, L. G., Forrest, A., Kawaji, H., Tan, S. L., Wahlestedt, C., Bajic, V. B., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., Bailey, T. L., & Huminiecki, L. (2006) Pseudo-messenger RNA: phantoms of the transcriptome. *PLoS Genet* 2, e23, PMID:16683022.
40. Zhang, Z., Carriero, N., & Gerstein, M. (2004) Comparative analysis of processed pseudogenes in the mouse and human genomes.. *Trends Genet* 20, 62-67, PMID:14746985.
41. Habegger, L., Balasubramanian, S., Chen, D. Z., Khurana, E., Sboner, A., Harmanci, A., Rozowsky, J., Clarke, D., Snyder, M., & Gerstein, M. (2012) VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment.. *Bioinformatics* 28, 2267-2269, PMID:22743228.
42. 1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015) A global reference for human genetic variation. *Nature* 526, 68-74, PMID:26432245.
43. Khurana, E., Lam, H. Y. K., Cheng, C., Carriero, N., Cayting, P., & Gerstein, M. B.

(2010) Segmental duplications in the human genome reveal details of pseudogene formation. *Nucleic Acids Res* 38, 6997-7007, PMID:20615899.