The Gerstein lab has extensive experience with analyzing cancer genomes through our involvement in a number of cancer-focused consortia, including TCGA and PCAWG. In TCGA, we were involved in studies of prostate (1-3) and kidney (4) cancers, and some of this work involved detailed analyses into minor splicing in cancer contexts. The Gerstein has built software tools and developed computational methods for finding sites splice sites throughout the human genome. We have also co-lead a PCAWG sub-group to investigate the impact of non-coding mutations in cancer. We have continued to expand upon our widely-used Function-based Prioritization of Sequence Variants (FunSeq) tool to study somatic cancer variants (5). We also developed FusionSeq (7), which is a computational framework to identify fusion transcripts from paired-end RNA-sequence data. FusionSeq includes filters to remove spurious candidate fusions with artifacts such as misalignments or random pairings of transcript fragments, and it provides rankings for identified candidates. We also explored the properties and consequences of recurrent repeat expansions (rREs) spanning 29 cancer types (8). We emphasize that many of these previous studies were carried out as part of joint efforts with the other groups, so our experience in these efforts are scientific and collaborative in nature.

Both methodologically and biologically, some of these efforts parallel our work in annotating pseudogenes throughout the genome. Along these lines, we are one of the major participants of the GENCODE project. As part of our 15-year involvement in GENCODE, our focus has been on pseudogenes (9, 10) and genome annotation (11). We have published several papers on pseudogene annotation, analysis, and regulatory annotation throughout the genome (12-18). In particular, we developed the PseudoPipe tool (Fig. 1) to find the pseudogenes (19), and it is one of the major tools used to annotate pseudogenes in GENCODE (11).



**Fig. 1: Workflow of PseudoPipe.**

**References:**

1. Cancer Genome Atlas Research N. The Molecular Taxonomy of Primary Prostate Cancer. Cell. 2015;163(4):1011-25. Epub 2015/11/07. doi: 10.1016/j.cell.2015.10.025. PubMed PMID: 26544944; PMCID: PMC4695400.

2. Cancer Genome Atlas Research Network. "Comprehensive molecular characterization of papillary renal-cell carcinoma." New England Journal of Medicine 374, no. 2 (2016): 135-145.

3. Augspach, Anke, Kyle D. Drake, Luca Roma, Ellen Qian, Se Ri Lee, Declan Clarke, Sushant Kumar et al. "Minor intron splicing efficiency increases with the development of lethal prostate cancer." bioRxiv (2021): 2021-12.

4. Li S, Shuch BM, Gerstein MB. Whole-genome analysis of papillary kidney cancer finds significant noncoding alterations. PLoS genetics. 2017;13(3):e1006685. Epub 2017/03/31. doi: 10.1371/journal.pgen.1006685. PubMed PMID: 28358873; PMCID: PMC5391127.

5. Fang LT, Afshar PT, Chhibber A, Mohiyuddin M, Fan Y, Mu JC, Gibeling G, Barr S, Asadi NB, Gerstein MB, Koboldt DC, Wang W, Wong WH, Lam HY. An ensemble approach to accurately detect somatic mutations using SomaticSeq. Genome Biol. 2015;16:197. Epub 2015/09/19. doi: 10.1186/s13059-015-0758-2. PubMed PMID: 26381235; PMCID: PMC4574535.

6. Khurana, Ekta, Yao Fu, Vincenza Colonna, Xinmeng Jasmine Mu, Hyun Min Kang, Tuuli Lappalainen, Andrea Sboner et al. "Integrative annotation of variants from 1092 humans: application to cancer genomics." Science 342, no. 6154 (2013): 1235587.

7. Sboner, Andrea, Lukas Habegger, Dorothee Pflueger, Stephane Terry, David Z. Chen, Joel S. Rozowsky, Ashutosh K. Tewari et al. "FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data." Genome biology 11 (2010): 1-19.

8. Erwin, Graham S., Gamze Gürsoy, Rashid Al-Abri, Ashwini Suriyaprakash, Egor Dolzhenko, Kevin Zhu, Christian R. Hoerner et al. "Recurrent repeat expansions in human cancer genomes." Nature (2022): 1-7.

9. Sisu, Cristina, Baikang Pei, Jing Leng, Adam Frankish, Yan Zhang, Suganthi Balasubramanian, Rachel Harte et al. "Comparative analysis of pseudogenes across three phyla." Proceedings of the National Academy of Sciences 111, no. 37 (2014): 13361-13366.

10. Pei, Baikang, Cristina Sisu, Adam Frankish, Cédric Howald, Lukas Habegger, Xinmeng Jasmine Mu, Rachel Harte et al. "The GENCODE pseudogene resource." Genome biology 13 (2012): 1-26.

11. Frankish, Adam, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M. Mudge et al. "GENCODE reference annotation for the human and mouse genomes." Nucleic acids research 47, no. D1 (2019): D766-D773.

12. Lam, Hugo YK, Ekta Khurana, Gang Fang, Philip Cayting, Nicholas Carriero, Kei-Hoi Cheung, and Mark B. Gerstein. "Pseudofam: the pseudogene families database." Nucleic acids research 37, no. suppl\_1 (2009): D738-D743.

13. Liu, Yuen-Jong, Deyou Zheng, Suganthi Balasubramanian, Nicholas Carriero, Ekta Khurana, Rebecca Robilotto, and Mark B. Gerstein. "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 (2009): 1-12.

14. Balasubramanian, Suganthi, Deyou Zheng, Yuen-Jong Liu, Gang Fang, Adam Frankish, Nicholas Carriero, Rebecca Robilotto, Philip Cayting, and Mark Gerstein. "Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes." Genome biology 10 (2009): 1-10.

15. Balasubramanian, Suganthi, Lukas Habegger, Adam Frankish, Daniel G. MacArthur, Rachel Harte, Chris Tyler-Smith, Jennifer Harrow, and Mark Gerstein. "Gene inactivation and its implications for annotation in the era of personal genomics." Genes & development 25, no. 1 (2011): 1-10.

16. Abyzov, Alexej, Rebecca Iskow, Omer Gokcumen, David W. Radke, Suganthi Balasubramanian, Baikang Pei, Lukas Habegger, Charles Lee, Mark Gerstein, and 1000 Genomes Project Consortium. "Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division." Genome research 23, no. 12 (2013): 2042-2052.

17. Khurana, Ekta, Hugo YK Lam, Chao Cheng, Nicholas Carriero, Philip Cayting, and Mark B. Gerstein. "Segmental duplications in the human genome reveal details of pseudogene formation." Nucleic acids research 38, no. 20 (2010): 6997-7007.

18. Zheng, Deyou, Adam Frankish, Robert Baertsch, Philipp Kapranov, Alexandre Reymond, Siew Woh Choo, Yontao Lu et al. "Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution." Genome research 17, no. 6 (2007): 839-851.

19. Zhang, Zhaolei, Nicholas Carriero, Deyou Zheng, John Karro, Paul M. Harrison, and Mark Gerstein. "PseudoPipe: an automated pseudogene identification pipeline." Bioinformatics 22, no. 12 (2006): 1437-1439.