***Gerstein Lab Experience in Transcriptome Analysis and Cancer Genomics***

**Pipelines for advanced molecular profiling.** Gerstein lab has extensive experience in developing advanced pipelines for extracting multi-layered molecular information of relevance for immuno-oncology such as DNAseq, RNAseq, and ChIPseq data. The lab is co-leading data analysis for ENCODE[31], and has previously led data analysis for mod/ENCODE[27, 32] and played a key role within the 1000 Genomes consortium[30]. Gerstein has also developed advanced analysis methods for proteomics[47, 49, 53] and metabolomics[39] data. Molecular profiling of cancer. Gerstein is currently co-leading the International Cancer Genomics Consortium (ICGC) pan-cancer analysis-working group (PCAWG)-2 (analysis of mutations in regulatory regions) group. In addition, we have participated in two The Cancer Genome Atlas (TCGA) studies on comprehensive molecular characterizations of 333 primary prostate carcinomas[20] and 161 primary papillary renal-cell carcinomas[21].

**Cancer genomics tools developed in Gerstein laboratory.** Gerstein lab has developed Variant Annotation Tool (VAT)[33] that annotates the impact of protein sequence mutations; ALoFT tool that predicts the impact of potential loss of function (LOF) variants in protein-coding genes; STRESS[25] tool that employs models of conformational change to predict allosteric residues. The lab developed methods to predict variants that are disruptive to a TF-binding motif in a regulatory region[27] and has integrated these methods into a prioritization pipeline for variants from WGS profiling called FunSeq[29, 36] that identified ~100 non-coding candidate drivers in ~90 WGS medulloblastoma, breast, and prostate cancer samples[36].

**Transcriptome analysis tools developed in Gerstein laboratory.** RNA-Seq provides a further layer of data which can provide valuable information, for instance the contribution of alternative splicing and non-coding RNAs to tumor and immune gene regulation. Gerstein lab has extensive experience in developing RNA-Seq processing pipelines as part of the mod/ENCODE consortia[27, 32]; has developed tools for identifying noncoding transcription and novel transcribed elements[18, 24, 32, 40, 46]; has developed the exceRpt[37] pipeline for extracellular small RNA-Seq profiling; and has developed tools and data formats for extremely large quantities of RNA-Seq data[34, 54].

**Liquid biopsy.** Gerstein lab has co-led data analysis and coordination for the Extracellular RNA

Communication Consortium, constructed the exRNA Atlas[2], and developed informatics methods, tools, and pipelines for the analysis of circulating RNA in human body fluids[52].

**Network modeling in Gerstein laboratory.** Gerstein lab has pioneered network frameworks for integrating a great variety of genomic data[22, 31, 55] and investigated the dynamics of networks[19, 48], thus laying a methodological groundwork for modeling tissue-level heterotypic tumor-immune interactions as well as modeling systemic effects of immuno-therapy.

**Development of data resource sharing policies.** Serving on the steering committees of several consortia, Gerstein helped drafting data sharing policies, led the effort to collect feedback from the Consortium members, modify and adopt the policies.

**References**

1. DreamHost. (URL). https://www.dreamhost.com.

2. The exRNA Atlas. (URL). http://exrna-atlas.org.

3. exRNA Research Portal. (URL). http://www.exrna.org.

4. HIPC. (URL). https://www.immuneprofiling.org.

5. HITRUST CSF. (URL). https://hitrustalliance.net/hitrust-csf/.

6. IEDB. (URL). http://www.iedb.org/.

7. ImmGen. (URL). https://www.immgen.org/.

8. ImmPort. (URL). http://www.immport.org/immport-open/public/home/home.

9. Internap. (URL). http://www.internap.com/.

10. ITN. (URL). https://www.itntrialshare.org/.

11. National Kidney Registry. (URL). http://www.kidneyregistry.org.

12. Open Science Data Cloud. (URL). https://www.opensciencedatacloud.org.

13. OpenStack. (URL). https://www.openstack.org/), .

14. Rackspace. (URL). https://www.rackspace.com/.

15. Altar, C.A., The Biomarkers Consortium: on the critical path of drug discovery. Clin Pharmacol Ther, 2008.

83(2): p. 361-4.

16. Amin, V., et al., Epigenomic footprints across 111 reference epigenomes reveal tissue-specific epigenetic

regulation of lincRNAs. Nat Commun, 2015. 6: p. 6370.

17. Bandrowski, A., et al., The Ontology for Biomedical Investigations. PLoS One, 2016. 11(4): p. e0154556.

18. Bertone, P., et al., Global identification of human transcribed sequences with genome tiling arrays.

Science, 2004. 306(5705): p. 2242-6.

19. Bhardwaj, N., P.M. Kim, and M.B. Gerstein, Rewiring of transcriptional regulatory networks: hierarchy,

rather than connectivity, better reflects the importance of regulators. Sci Signal, 2010. 3(146): p. ra79.

20. Cancer Genome Atlas Research, N., The Molecular Taxonomy of Primary Prostate Cancer. Cell, 2015.

163(4): p. 1011-25.

21. Cancer Genome Atlas Research, N., et al., Comprehensive Molecular Characterization of Papillary Renal-

Cell Carcinoma. N Engl J Med, 2016. 374(2): p. 135-45.

22. Cheng, C., R. Min, and M. Gerstein, TIP: a probabilistic method for identifying transcription factor target

genes from ChIP-seq binding profiles. Bioinformatics, 2011. 27(23): p. 3221-7.

23. Cheung, K.H., et al., Extending gene ontology in the context of extracellular RNA and vesicle

communication. J Biomed Semantics, 2016. 7: p. 19.

24. Clark, M.B., et al., The reality of pervasive transcription. PLoS Biol, 2011. 9(7): p. e1000625; discussion

e1001102.

25. Clarke, D., et al., Identifying Allosteric Hotspots with Dynamics: Application to Inter- and Intra-species

Conservation. Structure, 2016. 24(5): p. 826-37.

26. Coarfa, C., et al., Analysis of interactions between the epigenome and structural mutability of the genome

using Genboree Workbench tools. BMC Bioinformatics, 2014. 15 Suppl 7: p. S2.

27. Consortium, E.P., An integrated encyclopedia of DNA elements in the human genome. Nature, 2012.

489(7414): p. 57-74.

28. Diehl, A.D., et al., The Cell Ontology 2016: enhanced content, modularization, and ontology

interoperability. J Biomed Semantics, 2016. 7(1): p. 44.

29. Fu, Y., et al., FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome Biol,

2014. 15(10): p. 480.

30. Genomes Project, C., et al., A global reference for human genetic variation. Nature, 2015. 526(7571): p.

68-74.

31. Gerstein, M.B., et al., Architecture of the human regulatory network derived from ENCODE data. Nature,

2012. 489(7414): p. 91-100.

32. Gerstein, M.B., et al., Comparative analysis of the transcriptome across distant species. Nature, 2014.

512(7515): p. 445-8.

33. Habegger, L., et al., VAT: a computational framework to functionally annotate variants in personal

genomes within a cloud-computing environment. Bioinformatics, 2012. 28(17): p. 2267-9.

34. Habegger, L., et al., RSEQtools: a modular framework to analyze RNA-Seq data using compact,

anonymized data summaries. Bioinformatics, 2011. 27(2): p. 281-3.

35. Jonquet, C., N.H. Shah, and M.A. Musen, The open biomedical annotator. Summit on Translat Bioinforma,

2009: p. 56-60.

36. Khurana, E., et al., Integrative annotation of variants from 1092 humans: application to cancer genomics.

Science, 2013. 342(6154): p. 1235587.

37. Kitchen, R., M. Gerstein, and e. al., The extra-cellular RNA processing toolkit (in preparation). (URL), 2017.

http://github.gersteinlab.org/exceRpt/.

38. Lang, J.-P., Redmine. (URL), 2017. http://www.redmine.org.

39. Li, X., et al., Extensive in vivo metabolite-protein interactions revealed by large-scale systematic analyses.

Cell, 2010. 143(4): p. 639-50.

40. Lu, Z.J., et al., Prediction and characterization of noncoding RNAs in C. elegans by integrating

conservation, secondary structure, and high-throughput sequencing and array data. Genome Res, 2011.

21(2): p. 276-85.

41. Mungall, C.J., et al., Uberon, an integrative multi-species anatomy ontology. Genome Biol, 2012. 13(1): p.

R5.

42. Musen, M.A., et al., The center for expanded data annotation and retrieval. J Am Med Inform Assoc, 2015.

22(6): p. 1148-52.

43. Onuchic, V., et al., Epigenomic Deconvolution of Breast Tumors Reveals Metabolic Coupling between

Constituent Cell Types. Cell Rep, 2016. 17(8): p. 2075-2086.

44. Riehle, K., et al., The Genboree Microbiome Toolset and the analysis of 16S rRNA microbial sequences.

BMC Bioinformatics, 2012. 13 Suppl 13: p. S11.

45. Roadmap Epigenomics, C., et al., Integrative analysis of 111 reference human epigenomes. Nature, 2015.

518(7539): p. 317-30.

46. Rozowsky, J.S., et al., The DART classification of unannotated transcription within the ENCODE regions:

associating transcription with known and novel loci. Genome Res, 2007. 17(6): p. 732-45.

47. Sboner, A., et al., Robust-linear-model normalization to reduce technical variability in functional protein

microarrays. J Proteome Res, 2009. 8(12): p. 5451-64.

48. Shou, C., et al., Measuring the evolutionary rewiring of biological networks. PLoS Comput Biol, 2011. 7(1):

p. e1001050.

49. Smith, A., et al., Leveraging the structure of the Semantic Web to enhance information retrieval for

proteomics. Bioinformatics, 2007. 23(22): p. 3073-9.

50. Spackman, K., SNOMED RT and SNOMEDCT. Promise of an international clinical terminology. MD

Comput, 2000. 17(6): p. 29.

51. Strom, B.L., et al., Data sharing, year 1--access to data from industry-sponsored clinical trials. N Engl J

Med, 2014. 371(22): p. 2052-4.

52. Subramanian, S.L., et al., Integration of extracellular RNA profiling data using metadata, biomedical

ontologies and Linked Data technologies. J Extracell Vesicles, 2015. 4: p. 27497.

53. Vidal, M., et al., The human proteome - a scientific opportunity for transforming diagnostics, therapeutics,

and healthcare. Clin Proteomics, 2012. 9(1): p. 6.

54. Wang, Z., M. Gerstein, and M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet,

2009. 10(1): p. 57-63.

55. Yan, K.K., et al., OrthoClust: an orthology-based network framework for clustering data across multiple

species. Genome Biol, 2014. 15(8): p. R100.