Gerstein Lab experience in the development of multi-omic processing pipelines and data integration

Created date: 28 Mar 2016

Created by: DS

The Gerstein lab has experience developing ‘omics’ pipelines, computational methods and tools for large consortia, particularly the ENCODE and modENCODE projects. The following describes expertise for different ‘omics’ types and their integration.

Transcriptome Analysis.

We have extensive experience in developing RNA-Seq processing pipelines[[103](#_ENREF_103)] as part of the mod/ENCODE consortia [[26](#_ENREF_26), [34](#_ENREF_34)]. We have developed tools for identifying non-coding transcription and novel transcribed elements[[9](#_ENREF_9), [22](#_ENREF_22), [34](#_ENREF_34), [65](#_ENREF_65), [87](#_ENREF_87)], and contributed to the ExceRpt[[2](#_ENREF_2)] pipeline for the analysis of extracellular small RNA-Seq experiments. We have also developed a number of tools and data formats to handle large quantities of data generated by RNA-Seq experiments[[41](#_ENREF_41), [103](#_ENREF_103)].

ChIP-Seq analysis of TFs and chromatin marks.

We have developed PeakSeq[[86](#_ENREF_86)], a tool for the genome-wide identification of TF binding sites from ChIP-Seq data, which is used by ENCODE. More recently, we developed MUSIC[[32](#_ENREF_32)], a peak caller that performs multiscale decomposition of ChIP-Seq signal, which is applicable to studies of histone modifications enabling detection of broad and punctate regions of enrichment.

DNA Methylation Analysis.

We developed criteria for the design of epigenome-wide association studies[[44](#_ENREF_44), [67](#_ENREF_67)] and algorithms for whole-genome bisulfite sequencing[[24](#_ENREF_24), [60](#_ENREF_60)], and applied them to analyze over 100 methylomes as part of the NIH Roadmap Epigenomics project[[83](#_ENREF_83)]. We have performed integrative extensive analyses of the methylome, genetic variation[[23](#_ENREF_23), [62](#_ENREF_62)], and chromatin marks[[83](#_ENREF_83)].

Proteomics Analysis.

We have experience with the analysis of proteomic data[[90](#_ENREF_90), [97](#_ENREF_97), [101](#_ENREF_101)] and its integration with genomic data such as the combination of mass spectrometry (MS) proteomic and transcriptomic data[[56](#_ENREF_56), [104](#_ENREF_104)]. We developed miBAT on top of RNA-Seq to improve the reference proteome [[55](#_ENREF_55)].

# Metabolomics Studies.

We analyzed Gas Chromatrography Mass Spectrometry profiles for unknown metabolites, which can constitute as much as 50% of spectral features[16]. By integrating these data with transcriptomics profiles from different experimental conditions[36] and defining statistics to correlate the co-occurrence patterns of metabolites and genes we generated hypotheses about the identities of unannotated biosynthetic pathways.

# Normalization and deconvolution

We have developed a number of advanced methods for normalization, analysis, and comparison of RNA-seq profiles. In particular: (1) incRNA, a method that predicts novel ncRNAs using known ncRNAs of various biotypes as a training set[[65](#_ENREF_65)]; (2) FusionSeq, a pipeline to detect transcripts that arise due to trans-splicing or chromosomal translocations[[77](#_ENREF_77), [89](#_ENREF_89)]; (3) IQSeq, a transcript isoform quantification tool that uses an EM algorithm to resolve the maximum likelihood expression level of individual transcript isoforms; (4) Pseudo-seq[[75](#_ENREF_75)], which addresses the issue of quantification of pseudogene and repetitive region expression; and (5) Aggregation and Correlation Toolbox (ACT), which is a general purpose tool for comparing genomic signal tracks[[52](#_ENREF_52)]. In addition, we contributed to the development of a classification and analysis scheme for “spike” event patterns in omics data with longitudinal profiles[[16](#_ENREF_16)].

To identify both intracellular and tissue composition changes in profiled cells and tissues it is highly desirable to deconvolute the “omics” profiles of heterogeneous tissue samples. Toward this end, we have developed and experimentally validated Epigenomic Deconvolution (in review), a novel *in silico* deconvolution method that provides estimates of genomic CpG methylation, gene transcription, and other “omic” profiles within a diversity of constitutive cell types. The method employs an iterative algorithm for constrained matrix factorization using quadratic programming and extends the related method of Houseman[[45](#_ENREF_45) ] by deconvoluting gene expression and other “omic” profiles in addition to just CpG methylation profiles.

# Clustering and module formation

We have extensive experience in analyzing gene expression data to identify co-regulated modules. For example, we recently developed a new method, OrthoClust, to simultaneously cluster cross-species gene networks into gene modules[[105](#_ENREF_105)]. We applied OrthoClust to human and cross-species gene co-expression networks from the mod/ENCODE projects and discovered novel human developmental transcriptional programs.

Gene regulatory factors work cooperatively, forming a complex regulatory circuit controlling gene expression. We developed Loregic, a general-purpose method to characterize the cooperativity of regulatory factors[[102](#_ENREF_102)]. Using ENCODE ChIP-Seq and TCGA RNA-Seq data, we demonstrated how Loregic characterizes complex circuits involving TFs and miRNAs in human cancer. We developed continuous model-based approaches such as DREISS[[4](#_ENREF_4)] to identify gene expression dynamics driven by external and internal regulatory modules, helping dissect the temporal dynamic effects of different regulatory subsystems on gene expression.

# Building Integrative models

We have comprehensive experience integrating transcriptomic, metabolomics, and proteomic data. We integrated unknown metabolites, which can constitute as much as 50% of spectral features[16], with transcriptomics profiles from different experimental conditions[36]. By defining statistics to correlate the co-occurrence patterns of metabolites and genes we generated hypotheses about the identities of unannotated biosynthetic pathways. In addition, we have experience with the analysis of proteomic data and its integration with transcriptomics [56, 90, 97, 104]. This allowed us to identify previously uncharacterized proteins in a temporally and spatially resolved manner[104].

We have extensively used machine-learning to generate models from integrated datasets. For example, we integrated ENCODE data on transcription factor (TF) binding, histone modifications, and target gene expression to establish regulatory relationships using a probabilistic model we named TIP (Target Identification from Profiles)[19]. We identified potential enhancers from distal gene regions and we used these modules to quantify the relationship between TF binding and gene expression [17, 18, 25, 106]. We integrated these data types with protein-protein interaction and transcriptional regulation networks [20, 21, 27, 33]. This allowed us to group TFs into histone-sensitive and -insensitive classes that refined the prediction of gene-regulation targets and effects. Finally, we were able to build cross-organism integrative chromatin models [105].

# BIBLIOGRAPHY

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

2. extra-cellular RNA processing toolkit (exceRpt). (URL). http://github.gersteinlab.org/exceRpt/.

3. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature, 2008. 455(7216): p. 1061-8.

4. DREISS. PLoS Comput Biol, (in revision). https://github.com/gersteinlab/Dreiss.

5. Ahmad, T., et al., Prognostic Implications of Long-Chain Acylcarnitines in Heart Failure and Reversibility With Mechanical Circulatory Support. J Am Coll Cardiol, 2016. 67(3): p. 291-9.

6. Amin, V., et al., Epigenomic footprints across 111 reference epigenomes reveal tissue-specific epigenetic regulation of lincRNAs. Nat Commun, 2015. 6: p. 6370.

7. Belhajjame, K., et al., PROV-O: The PROV Ontology. W3C Recommendation, 2013. Apr.

8. Bernstein, B.E., et al., The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol, 2010. 28(10): p. 1045-8.

9. Bertone, P., et al., Global identification of human transcribed sequences with genome tiling arrays. Science, 2004. 306(5705): p. 2242-6.

10. Bhattacharya, S., et al., Validation of the association between a branched chain amino acid metabolite profile and extremes of coronary artery disease in patients referred for cardiac catheterization. Atherosclerosis, 2014. 232(1): p. 191-6.

11. Booth, F.W., et al., Waging war on modern chronic diseases: primary prevention through exercise biology. J Appl Physiol (1985), 2000. 88(2): p. 774-87.

12. Bouchard, C., et al., Adverse metabolic response to regular exercise: is it a rare or common occurrence? PLoS One, 2012. 7(5): p. e37887.

13. Bouchard, C., et al., Genomic predictors of the maximal O(2) uptake response to standardized exercise training programs. J Appl Physiol (1985), 2011. 110(5): p. 1160-70.

14. Celniker, S.E., et al., Unlocking the secrets of the genome. Nature, 2009. 459(7249): p. 927-30.

15. Challis, D., et al., An integrative variant analysis suite for whole exome next-generation sequencing data. BMC Bioinformatics, 2012. 13: p. 8.

16. Chen, R., et al., Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell, 2012. 148(6): p. 1293-307.

17. Cheng, C., et al., Understanding transcriptional regulation by integrative analysis of transcription factor binding data. Genome Res, 2012. 22(9): p. 1658-67.

18. Cheng, C. and M. Gerstein, Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. Nucleic Acids Res, 2012. 40(2): p. 553-68.

19. 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.

20. Cheng, C., et al., Genome-wide analysis of chromatin features identifies histone modification sensitive and insensitive yeast transcription factors. Genome Biol, 2011. 12(11): p. R111.

21. Cheng, C., et al., A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. Genome Biol, 2011. 12(2): p. R15.

22. Clark, M.B., et al., The reality of pervasive transcription. PLoS Biol, 2011. 9(7): p. e1000625; discussion e1001102.

23. 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.

24. Coarfa, C., et al., Pash 3.0: A versatile software package for read mapping and integrative analysis of genomic and epigenomic variation using massively parallel DNA sequencing. BMC Bioinformatics, 2010. 11: p. 572.

25. Consortium, E.P., An integrated encyclopedia of DNA elements in the human genome. Nature, 2012. 489(7414): p. 57-74.

26. Djebali, S., et al., Landscape of transcription in human cells. Nature, 2012. 489(7414): p. 101-8.

27. Dong, X., et al., Modeling gene expression using chromatin features in various cellular contexts. Genome Biol, 2012. 13(9): p. R53.

28. Evani, U.S., et al., Atlas2 Cloud: a framework for personal genome analysis in the cloud. BMC Genomics, 2012. 13 Suppl 6: p. S19.

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., A map of human genome variation from population-scale sequencing. Nature, 2010. 467(7319): p. 1061-73.

31. Genomes Project, C., et al., A global reference for human genetic variation. Nature, 2015. 526(7571): p. 68-74.

32. Gerstein, M.B., et al., Architecture of the human regulatory network derived from ENCODE data. Nature, 2012. 489(7414): p. 91-100.

33. Gerstein, M.B., et al., Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science, 2010. 330(6012): p. 1775-87.

34. Gerstein, M.B., et al., Comparative analysis of the transcriptome across distant species. Nature, 2014. 512(7515): p. 445-8.

35. Giaever, G., et al., Functional profiling of the Saccharomyces cerevisiae genome. Nature, 2002. 418(6896): p. 387-91.

36. Gianoulis, T.A., et al., Genomic analysis of the hydrocarbon-producing, cellulolytic, endophytic fungus Ascocoryne sarcoides. PLoS Genet, 2012. 8(3): p. e1002558.

37. Gibbs, R.A., et al., Evolutionary and biomedical insights from the rhesus macaque genome. Science, 2007. 316(5822): p. 222-34.

38. Gibbs, R.A., et al., Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature, 2004. 428(6982): p. 493-521.

39. Glynn, E.F., J. Chen, and A.R. Mushegian, Detecting periodic patterns in unevenly spaced gene expression time series using Lomb-Scargle periodograms. Bioinformatics, 2006. 22(3): p. 310-6.

40. Glynn, E.L., et al., Impact of combined resistance and aerobic exercise training on branched-chain amino acid turnover, glycine metabolism and insulin sensitivity in overweight humans. Diabetologia, 2015. 58(10): p. 2324-35.

41. 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.

42. Harmanci, A. and M. Gerstein, Quantification of private information leakage from phenotype-genotype data: linking attacks. Nat Methods, 2016. 13(3): p. 251-6.

43. Harris, R.A., J. Rogers, and A. Milosavljevic, Human-specific changes of genome structure detected by genomic triangulation. Science, 2007. 316(5822): p. 235-7.

44. Harris, R.A., et al., Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. Nat Biotechnol, 2010. 28(10): p. 1097-105.

45. Houseman E A, et al., Reference-free deconvolution of DNA methylation data and mediation by cell composition effe. bioRxiv, 2016.

46. Huffman, K.M., et al., Metabolite signatures of exercise training in human skeletal muscle relate to mitochondrial remodelling and cardiometabolic fitness. Diabetologia, 2014. 57(11): p. 2282-95.

47. Huffman, K.M., et al., Caloric restriction alters the metabolic response to a mixed-meal: results from a randomized, controlled trial. PLoS One, 2012. 7(4): p. e28190.

48. Huffman, K.M., et al., Relationships between circulating metabolic intermediates and insulin action in overweight to obese, inactive men and women. Diabetes Care, 2009. 32(9): p. 1678-83.

49. Huffman, K.M., et al., Exercise-induced changes in metabolic intermediates, hormones, and inflammatory markers associated with improvements in insulin sensitivity. Diabetes Care, 2011. 34(1): p. 174-6.

50. International, E., ECMA-404 The JSON Data Interchange Standard. (URL), 2013. http://json.org

(October).

51. J, C. and e. al., A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. Nat Commun. (in press).

52. Jee, J., et al., ACT: aggregation and correlation toolbox for analyses of genome tracks. Bioinformatics, 2011. 27(8): p. 1152-4.

53. Joyner, M.J. and D.J. Green, Exercise protects the cardiovascular system: effects beyond traditional risk factors. J Physiol, 2009. 587(Pt 23): p. 5551-8.

54. Khurana, E., et al., Integrative annotation of variants from 1092 humans: application to cancer genomics. Science, 2013. 342(6154): p. 1235587.

55. Kitchen, R. and e. al., miBAT: Multi-modal profiling of the translatome at isoform resolution. (in preparation).

56. Kitchen, R.R., et al., Decoding neuroproteomics: integrating the genome, translatome and functional anatomy. Nat Neurosci, 2014. 17(11): p. 1491-9.

57. Kovalik, J.P., et al., Metabolic remodeling of human skeletal myocytes by cocultured adipocytes depends on the lipolytic state of the system. Diabetes, 2011. 60(7): p. 1882-93.

58. Kraus, W.E., T.S. Bernard, and R.S. Williams, Interactions between sustained contractile activity and beta-adrenergic receptors in regulation of gene expression in skeletal muscles. Am J Physiol, 1989. 256(3 Pt 1): p. C506-14.

59. Kraus, W.E., et al., Metabolomic Quantitative Trait Loci (mQTL) Mapping Implicates the Ubiquitin Proteasome System in Cardiovascular Disease Pathogenesis. PLoS Genet, 2015. 11(11): p. e1005553.

60. Kunde-Ramamoorthy, G., et al., Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing. Nucleic Acids Res, 2014. 42(6): p. e43.

61. Kutmon, M., et al., WikiPathways: capturing the full diversity of pathway knowledge. Nucleic Acids Res, 2016. 44(D1): p. D488-94.

62. Li, J., et al., Genomic hypomethylation in the human germline associates with selective structural mutability in the human genome. PLoS Genet, 2012. 8(5): p. e1002692.

63. Lochovsky, L., et al., LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations. Nucleic Acids Res, 2015. 43(17): p. 8123-34.

64. Love, M.I., W. Huber, and S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol, 2014. 15(12): p. 550.

65. 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.

66. Lum, H., et al., Plasma acylcarnitines are associated with physical performance in elderly men. J Gerontol A Biol Sci Med Sci, 2011. 66(5): p. 548-53.

67. Michels, K.B., et al., Recommendations for the design and analysis of epigenome-wide association studies. Nat Methods, 2013. 10(10): p. 949-55.

68. Milosavljevic, A., Putting epigenome comparison into practice. Nat Biotechnol, 2010. 28(10): p. 1053-6.

69. Muoio, D.M., et al., Fatty acid homeostasis and induction of lipid regulatory genes in skeletal muscles of peroxisome proliferator-activated receptor (PPAR) alpha knock-out mice. Evidence for compensatory regulation by PPAR delta. J Biol Chem, 2002. 277(29): p. 26089-97.

70. Muoio, D.M., et al., Muscle-specific deletion of carnitine acetyltransferase compromises glucose tolerance and metabolic flexibility. Cell Metab, 2012. 15(5): p. 764-77.

71. Muoio, D.M., et al., Peroxisome proliferator-activated receptor-alpha regulates fatty acid utilization in primary human skeletal muscle cells. Diabetes, 2002. 51(4): p. 901-9.

72. Muzny, D.M., et al., The DNA sequence, annotation and analysis of human chromosome 3. Nature, 2006. 440(7088): p. 1194-8.

73. Neufer, P.D., et al., Understanding the Cellular and Molecular Mechanisms of Physical Activity-Induced Health Benefits. Cell Metab, 2015. 22(1): p. 4-11.

74. Overby, C.L., et al., Providing Access to Genomic Variant Knowledge in a Healthcare Setting: A Vision for the ClinGen Electronic Health Records Workgroup. Clin Pharmacol Ther, 2016. 99(2): p. 157-60.

75. Pei, B., et al., The GENCODE pseudogene resource. Genome Biol, 2012. 13(9): p. R51.

76. Pette, D. and R.S. Staron, Myosin isoforms, muscle fiber types, and transitions. Microsc Res Tech, 2000. 50(6): p. 500-9.

77. Pflueger, D., et al., Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. Genome Res, 2011. 21(1): p. 56-67.

78. Pico, A.R., et al., WikiPathways: pathway editing for the people. PLoS Biol, 2008. 6(7): p. e184.

79. Psych, E.C., et al., The PsychENCODE project. Nat Neurosci, 2015. 18(12): p. 1707-12.

80. Redman, L.M., et al., Effect of caloric restriction with and without exercise on metabolic intermediates in nonobese men and women. J Clin Endocrinol Metab, 2011. 96(2): p. E312-21.

81. Rehm, H.L., et al., ClinGen--the Clinical Genome Resource. N Engl J Med, 2015. 372(23): p. 2235-42.

82. Riehle, K., et al., The Genboree Microbiome Toolset and the analysis of 16S rRNA microbial sequences. BMC Bioinformatics, 2012. 13 Suppl 13: p. S11.

83. Roadmap Epigenomics, C., et al., Integrative analysis of 111 reference human epigenomes. Nature, 2015. 518(7539): p. 317-30.

84. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 2010. 26(1): p. 139-40.

85. Rozowsky, J., et al., AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol Syst Biol, 2011. 7: p. 522.

86. Rozowsky, J., et al., PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol, 2009. 27(1): p. 66-75.

87. 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.

88. Rubin, R., Despite Potential Health Benefits of Maternity Leave, US Lags Behind Other Industrialized Countries. JAMA, 2016. 315(7): p. 643-5.

89. Sboner, A., et al., FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. Genome Biol, 2010. 11(10): p. R104.

90. 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.

91. Scherer, S.E., et al., The finished DNA sequence of human chromosome 12. Nature, 2006. 440(7082): p. 346-51.

92. Schiaffino, S. and C. Reggiani, Molecular diversity of myofibrillar proteins: gene regulation and functional significance. Physiol Rev, 1996. 76(2): p. 371-423.

93. Science, A.I.f.B., BrainSpan Atlas of the Developing Human Brain. (URL), 2015. http://brainspan.org.

94. Shah, A.A., et al., Metabolic profiles predict adverse events after coronary artery bypass grafting. J Thorac Cardiovasc Surg, 2012. 143(4): p. 873-8.

95. Shah, S.H., et al., Association of a peripheral blood metabolic profile with coronary artery disease and risk of subsequent cardiovascular events. Circ Cardiovasc Genet, 2010. 3(2): p. 207-14.

96. Shah, S.H., et al., Baseline metabolomic profiles predict cardiovascular events in patients at risk for coronary artery disease. Am Heart J, 2012. 163(5): p. 844-850 e1.

97. Smith, A., et al., Leveraging the structure of the Semantic Web to enhance information retrieval for proteomics. Bioinformatics, 2007. 23(22): p. 3073-9.

98. Steve Speicher, et al., Linked Data Platform 1.0 (Working Draft 16 September 2014). http://www.w3.org/TR/ldp/, 2014. Sep.

99. 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.

100. Van Dongen, H.P., et al., Searching for biological rhythms: peak detection in the periodogram of unequally spaced data. J Biol Rhythms, 1999. 14(6): p. 617-20.

101. Vidal, M., et al., The human proteome - a scientific opportunity for transforming diagnostics, therapeutics, and healthcare. Clin Proteomics, 2012. 9(1): p. 6.

102. Wang, D., et al., Loregic: a method to characterize the cooperative logic of regulatory factors. PLoS Comput Biol, 2015. 11(4): p. e1004132.

103. Wang, Z., M. Gerstein, and M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet, 2009. 10(1): p. 57-63.

104. Wu, L., et al., Global survey of human T leukemic cells by integrating proteomics and transcriptomics profiling. Mol Cell Proteomics, 2007. 6(8): p. 1343-53.

105. Yan, K.K., et al., OrthoClust: an orthology-based network framework for clustering data across multiple species. Genome Biol, 2014. 15(8): p. R100.

106. Yip, K.Y., et al., Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. Genome Biol, 2012. 13(9): p. R48.