Experience Creating RNAseq processing pipelines

We have built an RNA-seq processing pipeline based on the software suite, RSEQtools, that we have largely developed. These tools consist of a set of modules that perform common tasks such as calculating gene and exon expression values, generating signal tracks of mapped reads and segmenting that signal into actively transcribed regions. Also, implemented within RSEQtools are more specialized analysis pipelines that we have developed (e.g. FusionSeq for fusion transcript detection [2], IQSeq for transcript quantification [3], and DupSeq for analyzing expression patterns of highly homologous genomic regions [4]), as well as thoroughly validated tools such as Bowtie and Tophat [5, 6]. These tools are implemented using Mapped Read Format (MRF), a compact data summary format for short, long and paired-end read alignments that enables the anonymization of confidential sequence information. With this set of tools, we have provided custom processing pipelines.

# Non-coding RNA (ncRNA) and pseudogene analysis

A fraction of transcription comes from genomic regions not associated with standard annotations, representing ‘non-canonical transcription’. These transcripts are observable even when experimental protocols use poly-A enrichment [7]. A class of non-canonical transcripts of particular interest is the pseudogene, which recent studies have shown are useful biomarkers to distinguish different cell types. Despite their low abundance, pseudogenes and ncRNAs have been shown to exhibit a greater degree of cell-type specific expression than mRNAs [8]. However, the quantification of pseudogene expression is challenging because of the sequence similarity with its parent genes. To address the issue, we developed DupSeq, which solves this problem by focusing only on those reads and regions that are uniquely mappable [4].

Several other classes of non-coding RNAs have been shown to play regulatory or other roles in the cell. To identify these loci we will apply incRNA, a method that predicts novel ncRNAs using known ncRNAs of various biotypes as a gold standard training set and a minimum-run–maximum-gap algorithm to process reads mapping outside of protein-coding transcripts, pseudogenes and annotated non-coding RNAs [9, 10].

# Functional annotation through clustering and network analyses

We have extensive experience in characterizing the functions of genes and non-coding elements via expression data through clustering and network analyses.A group of genes in a co-expression cluster are often responsible for a common function [11]. While there are well known algorithms for expression clustering such as hierarchical clustering, spectral clustering and K-means, we developed several novel methods. We developed a spectral biclustering method for co-clustering genes and conditions. More recently, we developed a new clustering framework, OrthoClust, for simultaneously clustering network data across different contexts [12]. OrthoClust is able to identify conserved and specific components across different networks. We applied OrthoClust in the comparative transcriptome analysis, and discovered co-expression modules shared in animals and enriched in their developmental genes. Furthermore, expression clusters can be used for annotating functions of unknown transcripts. For example, in modENCODE analysis, by mapping the expression profiles of various ncRNAs to expression clusters, we have used identified functions of various ncRNAs.

The functional relationships between co-expressed genes can further be understood via various molecular networks. Over the past decade, we have identified many relationships between topological properties of genes in networks and their functional genomics features. In terms of local properties, for instance, we found the “essentiality” of a given node is closely related to its tendency to act as a hub or bottleneck [13, 14]. In terms of system-wide organization, we found that gene-regulatory networks are composed of hierarchical structures dominated by downward information flow from master regulators [15]. We further developed methods to determine the hierarchical organization of various regulatory networks, including networks constructed from ENCODE, modENCODE and MCF7 data [16, 17, 18, 19]. Apart from studying individual networks, we also introduced a framework to quantify differences between networks and found a consistent ordering of rewiring rates of different network types. [20].

# RNA-seq pipeline development for large-scale projects

We have worked on the development and analysis of multiple RNA-seq flows in the context of large consortia, including the implementation of tools we developed and other popular tools such as Bowtie and Tophat. For example, we have been playing a role in such activities for the ENCODE consortium [21], including a recent publication involving the processing and integration of all ENCODE and modENCODE data, which involved 575 experiments and more than 65 billion reads from three organisms. [10]. We are the data integration hub in the Extracellular RNA consortium [22] that generates hundreds of RNA-seq and small RNA-seq samples. Other notable consortia for which we have processed large quantities of data include the BrainSpan project [23] which collected RNA-seq data for 8-16 brain structures in each of 13 developmental stages [24], as well as the PsychENCODE project [25].

2. Sboner, A., Habegger, L., Pflueger, D., Terry, S., Chen, D. Z., Rozowsky, J. S., Tewari, A. K., Kitabayashi, N., Moss, B. J., Chee, M. S., Demichelis, F., Rubin, M. A., & Gerstein, M. B. (2010) FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data.. *Genome Biol* 11, R104.3. Du, J., Leng, J., Habegger, L., Sboner, A., McDermott, D., & Gerstein, M. (2012) IQSeq: integrated isoform quantification analysis based on next-generation sequencing.. *PLoS One* 7, e29175.4. Habegger, L. (2012) Computational methodologies for transcript analysis in the age of next-generation DNA sequencing. *Thesis (Ph. D.)--Yale University, 2012.* , .5. Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10, R25.6. Trapnell, C., Pachter, L., & Salzberg, S. L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105-11.7. Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., Huarte, M., Zuk, O., Carey, B. W., Cassady, J. P., Cabili, M. N., Jaenisch, R., Mikkelsen, T. S., Jacks, T., Hacohen, N., Bernstein, B. E., Kellis, M., Regev, A., Rinn, J. L., & Lander, E. S. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458, 223-7.8. 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, 13361-6.9. Lu, Z. J., Yip, K. Y., Wang, G., Shou, C., Hillier, L. W., Khurana, E., Agarwal, A., Auerbach, R., Rozowsky, J., Cheng, C., Kato, M., Miller, D. M., Slack, F., Snyder, M., Waterston, R. H., Reinke, V., & Gerstein, M. B. (2011) Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data. *Genome Res* 21, 276-85.10. Gerstein, M. B., Rozowsky, J., Yan, K.-K., Wang, D., Cheng, C., Brown, J. B., Davis, C. A., Hillier, L., Sisu, C., Li, J. J., Pei, B., Harmanci, A. O., Duff, M. O., Djebali, S., Alexander, R. P., Alver, B. H., Auerbach, R., Bell, K., Bickel, P. J., Boeck, M. E., Boley, N. P., Booth, B. W., Cherbas, L., Cherbas, P., Di, C., Dobin, A., Drenkow, J., Ewing, B., Fang, G., Fastuca, M., Feingold, E. A., Frankish, A., Gao, G., Good, P. J., Guigó, R., Hammonds, A., Harrow, J., Hoskins, R. A., Howald, C., Hu, L., Huang, H., Hubbard, T. J. P., Huynh, C., Jha, S., Kasper, D., Kato, M., Kaufman, T. C., Kitchen, R. R., Ladewig, E., Lagarde, J., Lai, E., Leng, J., Lu, Z., MacCoss, M., May, G., McWhirter, R., Merrihew, G., Miller, D. M., Mortazavi, A., Murad, R., Oliver, B., Olson, S., Park, P. J., Pazin, M. J., Perrimon, N., Pervouchine, D., Reinke, V., Reymond, A., Robinson, G., Samsonova, A., Saunders, G. I., Schlesinger, F., Sethi, A., Slack, F. J., Spencer, W. C., Stoiber, M. H., Strasbourger, P., Tanzer, A., Thompson, O. A., Wan, K. H., Wang, G., Wang, H., Watkins, K. L., Wen, J., Wen, K., Xue, C., Yang, L., Yip, K., Zaleski, C., Zhang, Y., Zheng, H., Brenner, S. E., Graveley, B. R., Celniker, S. E., Gingeras, T. R., & Waterston, R. (2014) Comparative analysis of the transcriptome across distant species. *Nature* 512, 445-8.11. D'haeseleer, P., Liang, S., & Somogyi, R. (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16, 707-26.12. Yan, K.-K., Wang, D., Rozowsky, J., Zheng, H., Cheng, C., & Gerstein, M. (2014) OrthoClust: an orthology-based network framework for clustering data across multiple species. *Genome Biol* 15, R100.13. Yu, H., Greenbaum, D., Xin Lu, H., Zhu, X., & Gerstein, M. (2004) Genomic analysis of essentiality within protein networks.. *Trends Genet* 20, 227--231.14. Yu, H., Kim, P. M., Sprecher, E., Trifonov, V., & Gerstein, M. (2007) The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics.. *PLoS Comput Biol* 3, e59.15. Yu, H. & Gerstein, M. (2006) Genomic analysis of the hierarchical structure of regulatory networks.. *Proc Natl Acad Sci U S A* 103, 14724--14731.16. Cheng, C., Andrews, E., Yan, K.-K., Ung, M., Wang, D., & Gerstein, M. (2015) An approach for determining and measuring network hierarchy applied to comparing the phosphorylome and the regulome. *Genome Biol* 16, 63.17. Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., Mu, X. J., Khurana, E., Rozowsky, J., Alexander, R., Min, R., Alves, P., Abyzov, A., Addleman, N., Bhardwaj, N., Boyle, A. P., Cayting, P., Charos, A., Chen, D. Z., Cheng, Y., Clarke, D., Eastman, C., Euskirchen, G., Frietze, S., Fu, Y., Gertz, J., Grubert, F., Harmanci, A., Jain, P., Kasowski, M., Lacroute, P., Leng, J., Lian, J., Monahan, H., O'Geen, H., Ouyang, Z., Partridge, E. C., Patacsil, D., Pauli, F., Raha, D., Ramirez, L., Reddy, T. E., Reed, B., Shi, M., Slifer, T., Wang, J., Wu, L., Yang, X., Yip, K. Y., Zilberman-Schapira, G., Batzoglou, S., Sidow, A., Farnham, P. J., Myers, R. M., Weissman, S. M., & Snyder, M. (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature* 489, 91-100.18. Cheng, C., Yan, K.-K., Hwang, W., Qian, J., Bhardwaj, N., Rozowsky, J., Lu, Z. J., Niu, W., Alves, P., Kato, M., Snyder, M., & Gerstein, M. (2011) Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. *PLoS Comput Biol* 7, e1002190.19. Gerstein, M. B., Lu, Z. J., Van Nostrand, E. L., Cheng, C., Arshinoff, B. I., Liu, T., Yip, K. Y., Robilotto, R., Rechtsteiner, A., Ikegami, K., Alves, P., Chateigner, A., Perry, M., Morris, M., Auerbach, R. K., Feng, X., Leng, J., Vielle, A., Niu, W., Rhrissorrakrai, K., Agarwal, A., Alexander, R. P., Barber, G., Brdlik, C. M., Brennan, J., Brouillet, J. J., Carr, A., Cheung, M.-S., Clawson, H., Contrino, S., Dannenberg, L. O., Dernburg, A. F., Desai, A., Dick, L., Dose, A. C., Du, J., Egelhofer, T., Ercan, S., Euskirchen, G., Ewing, B., Feingold, E. A., Gassmann, R., Good, P. J., Green, P., Gullier, F., Gutwein, M., Guyer, M. S., Habegger, L., Han, T., Henikoff, J. G., Henz, S. R., Hinrichs, A., Holster, H., Hyman, T., Iniguez, A. L., Janette, J., Jensen, M., Kato, M., Kent, W. J., Kephart, E., Khivansara, V., Khurana, E., Kim, J. K., Kolasinska-Zwierz, P., Lai, E. C., Latorre, I., Leahey, A., Lewis, S., Lloyd, P., Lochovsky, L., Lowdon, R. F., Lubling, Y., Lyne, R., MacCoss, M., Mackowiak, S. D., Mangone, M., McKay, S., Mecenas, D., Merrihew, G., Miller, D. M. r., Muroyama, A., Murray, J. I., Ooi, S.-L., Pham, H., Phippen, T., Preston, E. A., Rajewsky, N., Ratsch, G., Rosenbaum, H., Rozowsky, J., Rutherford, K., Ruzanov, P., Sarov, M., Sasidharan, R., Sboner, A., Scheid, P., Segal, E., Shin, H., Shou, C., Slack, F. J., Slightam, C., Smith, R., Spencer, W. C., Stinson, E. O., Taing, S., Takasaki, T., Vafeados, D., Voronina, K., Wang, G., Washington, N. L., Whittle, C. M., Wu, B., Yan, K.-K., Zeller, G., Zha, Z., Zhong, M., Zhou, X., Ahringer, J., Strome, S., Gunsalus, K. C., Micklem, G., Liu, X. S., Reinke, V., Kim, S. K., Hillier, L. W., Henikoff, S., Piano, F., Snyder, M., Stein, L., Lieb, J. D., & Waterston, R. H. (2010) Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project.. *Science* 330, 1775--1787.20. Shou, C., Bhardwaj, N., Lam, H. Y. K., Yan, K.-K., Kim, P. M., Snyder, M., & Gerstein, M. B. (2011) Measuring the evolutionary rewiring of biological networks.. *PLoS Comput Biol* 7, e1001050.21. Washietl, S., Pedersen, J. S., Korbel, J. O., Stocsits, C., Gruber, A. R., Hackermüller, J., Hertel, J., Lindemeyer, M., Reiche, K., Tanzer, A., Ucla, C., Wyss, C., Antonarakis, S. E., Denoeud, F., Lagarde, J., Drenkow, J., Kapranov, P., Gingeras, T. R., Guigó, R., Snyder, M., Gerstein, M. B., Reymond, A., Hofacker, I. L., & Stadler, P. F. (2007) Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res* 17, 852-64.22. http://exRNA.org.23. http://brainspan.org.24. Miller, J. A., Ding, S.-L., Sunkin, S. M., Smith, K. A., Ng, L., Szafer, A., Ebbert, A., Riley, Z. L., Royall, J. J., Aiona, K., Arnold, J. M., Bennet, C., Bertagnolli, D., Brouner, K., Butler, S., Caldejon, S., Carey, A., Cuhaciyan, C., Dalley, R. A., Dee, N., Dolbeare, T. A., Facer, B. A. C., Feng, D., Fliss, T. P., Gee, G., Goldy, J., Gourley, L., Gregor, B. W., Gu, G., Howard, R. E., Jochim, J. M., Kuan, C. L., Lau, C., Lee, C.-K., Lee, F., Lemon, T. A., Lesnar, P., McMurray, B., Mastan, N., Mosqueda, N., Naluai-Cecchini, T., Ngo, N.-K., Nyhus, J., Oldre, A., Olson, E., Parente, J., Parker, P. D., Parry, S. E., Stevens, A., Pletikos, M., Reding, M., Roll, K., Sandman, D., Sarreal, M., Shapouri, S., Shapovalova, N. V., Shen, E. H., Sjoquist, N., Slaughterbeck, C. R., Smith, M., Sodt, A. J., Williams, D., Zöllei, L., Fischl, B., Gerstein, M. B., Geschwind, D. H., Glass, I. A., Hawrylycz, M. J., Hevner, R. F., Huang, H., Jones, A. R., Knowles, J. A., Levitt, P., Phillips, J. W., Sestan, N., Wohnoutka, P., Dang, C., Bernard, A., Hohmann, J. G., & Lein, E. S. (2014) Transcriptional landscape of the prenatal human brain. *Nature* 508, 199-206.25. http://psychencode.org.