Gerstein Lab Contributions to RNA-Seq Analysis

(12/02/2013)

Here we summarize work in the Gerstein lab over the last decade developing analysis tools for RNA-Seq, and conducting integrative analyses of large scale RNA-seq data sets.

For general RNA-Seq analysis, we have developed RSEQtools, a computational package that enables expression quantification of annotated RNAs and identification of splice sites and gene models [1]. In addition, we have developed IQseq, a computationally efficient method to quantify isoforms for alternatively spliced transcripts [2]. Both of these tools employ a special sequence read format that we have developed that can dissociate genome sequence information from RNA-Seq signal, maintaining the privacy of test subjects. Comparisons between RNA-Seq samples, and to other genome-wide data, are be facilitated in part by our Aggregation and Correlation Toolbox (ACT), which is a general purpose tool for comparing genome signal tracks [3].

An important challenge in RNA-Seq analysis is detecting unannotated transcription that may be hard to distinguish from noise. This topic has been central to many of our expression analysis tools for both microarrays and RNA-Seq [4-10]. Our Database of Annotated Regions with Tools (DART) package contains tools for identifying unannotated genomic regions that are enriched for transcription, as well as a framework for storing and querying this information [11]. To investigate newly transcriptionally active regions further, we developed incRNA, a method that predicts novel ncRNAs using known ncRNAs of various biotypes as a gold standard training set [12].

We have also developed specific tools to identify types of transcripts that are difficult to detect using standard analysis pipelines. We recently developed FusionSeq, a pipeline to detect transcripts that arise due to trans-splicing or chromosomal translocations [13,14]. These transcripts have been implicated in numerous diseases, most famously in chronic myeloid leukemia, and this tool aids searches for other fusion transcripts with potentially important functions [15]. We also developed Pseudo-seq, which addresses the issue of quantification of pseudogene expression, which is difficult to separate from the transcription of parent genes with similar sequences. Pseudoseq solves this problem by calculating the expression in terms of RPKM for pseudogenes by focusing only on those reads and regions that are uniquely mappable [16]. Though pseudogenes have long been thought to be non-functional, recent studies have revealed their roles in cancer, X-chromosome inactivation and intercellular signaling [17-19].

Another major area of interest in RNA-Seq analysis is linking expression variation to genotype. We have expertise in this subject in the form of allelic analysis. Our AlleleSeq tool [20] combines diploid genomic information with RNA-Seq data to identify transcripts showing allele specific expression.

We also have extensive experience conducting integrated analyses of large sets of RNA-seq data, primarily through the ENCODE project [21]. We played a lead role in the analysis of model organisms (C.Elegans) and human transcriptome studies within the consortium, two of the largest RNA-Seq studies to date [7,8]. We have also conducted extensive studies of the relationship between ChIP-Seq data for localization of transcription factors and histone modifications and gene expression through RNA-Seq [22,23]. Currently, we are active participants of the Brainspan project, which profiles RNAs in different parts of the human brain (http://www.brainspan.org). We are involved in the coordination of the RNA-seq working group activities for the ENCODE project [20]; and we lead the data integration and analysis component (DIAC) of the data management and resource repository (DMRR) for the NIH extracellular RNA communication program consortium (http://commonfund.nih.gov/Exrna/).

References:

[1] Habegger, L, Sboner, A, Gianoulis, TA, Rozowsky, J, Agarwal, A, Snyder, M, Gerstein, M (2011). RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. Bioinformatics, 27, 2:281-3.
[2] 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, 1:e29175.
[3] Jee, J, Rozowsky, J, Yip, KY, Lochovsky, L, Bjornson, R, Zhong, G, Zhang, Z, Fu, Y, Wang, J, Weng, Z, Gerstein, M **(2011)**. ACT: aggregation and correlation toolbox for analyses of genome tracks. *Bioinformatics*, 27, 8:1152-4.
9, 7:e1000625; discussion e1001102.
[4] Nagalakshmi, U, Wang, Z, Waern, K, Shou, C, Raha, D, Gerstein, M, Snyder, M **(2008)**. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320, 5881:1344-9.
[5] - Bertone, P, Stolc, V, Royce, TE, Rozowsky, JS, Urban, AE, Zhu, X, Rinn, JL, Tongprasit, W, Samanta, M, Weissman, S, Gerstein, M, Snyder, M **(2004)**. Global identification of human transcribed sequences with genome tiling arrays. *Science*, 306, 5705:2242-6.
[6] - Wang, Z, Gerstein, M, Snyder, M **(2009)**. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10, 1:57-63.
[7] - Djebali, S, Davis, CA, Merkel, A, Dobin, A, Lassmann, T, Mortazavi, A, Tanzer, A, Lagarde, J, Lin, W, Schlesinger, F, Xue, C, Marinov, GK, Khatun, J, Williams, BA, Zaleski, C, Rozowsky, J, Röder, M, Kokocinski, F, Abdelhamid, RF, Alioto, T, Antoshechkin, I, Baer, MT, Bar, NS, Batut, P, Bell, K, Bell, I, Chakrabortty, S, Chen, X, Chrast, J, Curado, J, Derrien, T, Drenkow, J, Dumais, E, Dumais, J, Duttagupta, R, Falconnet, E, Fastuca, M, Fejes-Toth, K, Ferreira, P, Foissac, S, Fullwood, MJ, Gao, H, Gonzalez, D, Gordon, A, Gunawardena, H, Howald, C, Jha, S, Johnson, R, Kapranov, P, King, B, Kingswood, C, Luo, OJ, Park, E, Persaud, K, Preall, JB, Ribeca, P, Risk, B, Robyr, D, Sammeth, M, Schaffer, L, See, LH, Shahab, A, Skancke, J, Suzuki, AM, Takahashi, H, Tilgner, H, Trout, D, Walters, N, Wang, H, Wrobel, J, Yu, Y, Ruan, X, Hayashizaki, Y, Harrow, J, Gerstein, M, Hubbard, T, Reymond, A, Antonarakis, SE, Hannon, G, Giddings, MC, Ruan, Y, Wold, B, Carninci, P, Guigó, R, Gingeras, TR **(2012)**. Landscape of transcription in human cells. *Nature*, 489, 7414:101-8.
[8] Gerstein, MB, Lu, ZJ, Van Nostrand, EL, Cheng, C, Arshinoff, BI, Liu, T, Yip, KY, Robilotto, R, Rechtsteiner, A, Ikegami, K, Alves, P, Chateigner, A, Perry, M, Morris, M, Auerbach, RK, Feng, X, Leng, J, Vielle, A, Niu, W, Rhrissorrakrai, K, Agarwal, A, Alexander, RP, Barber, G, Brdlik, CM, Brennan, J, Brouillet, JJ, Carr, A, Cheung, MS,

Clawson, H, Contrino, S, Dannenberg, LO, Dernburg, AF, Desai, A, Dick, L, Dosé, AC, Du, J, Egelhofer, T, Ercan, S, Euskirchen, G, Ewing, B, Feingold, EA, Gassmann, R, Good, PJ, Green, P, Gullier, F, Gutwein, M, Guyer, MS, Habegger, L, Han, T, Henikoff, JG, Henz, SR, Hinrichs, A, Holster, H, Hyman, T, Iniguez, AL, Janette, J, Jensen, M, Kato, M, Kent, WJ, Kephart, E, Khivansara, V, Khurana, E, Kim, JK, Kolasinska-Zwierz, P, Lai, EC, Latorre, I, Leahey, A, Lewis, S, Lloyd, P, Lochovsky, L, Lowdon, RF, Lubling, Y, Lyne, R, MacCoss, M, Mackowiak, SD, Mangone, M, McKay, S, Mecenas, D, Merrihew, G, Miller, DM, Muroyama, A, Murray, JI, Ooi, SL, Pham, H, Phippen, T, Preston, EA, Rajewsky, N, Rätsch, G, Rosenbaum, H, Rozowsky, J, Rutherford, K, Ruzanov, P, Sarov, M, Sasidharan, R, Sboner, A, Scheid, P, Segal, E, Shin, H, Shou, C, Slack, FJ, Slightam, C, Smith, R, Spencer, WC, Stinson, EO, Taing, S, Takasaki, T, Vafeados, D, Voronina, K, Wang, G, Washington, NL, Whittle, CM, Wu, B, Yan, KK, Zeller, G, Zha, Z, Zhong, M, Zhou, X, Ahringer, J, Strome, S, Gunsalus, KC, Micklem, G, Liu, XS, Reinke, V, Kim, SK, Hillier, LW, Henikoff, S, Piano, F, Snyder, M, Stein, L, Lieb, JD, Waterston, RH **(2010)**. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science*, 330, 6012:1775-87.

[9] - Agarwal, A, Koppstein, D, Rozowsky, J, Sboner, A, Habegger, L, Hillier, LW, Sasidharan, R, Reinke, V, Waterston, RH, Gerstein, M **(2010)**. Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. *BMC Genomics*, 11:383.

[10] Clark, MB, Amaral, PP, Schlesinger, FJ, Dinger, ME, Taft, RJ, Rinn, JL, Ponting, CP, Stadler, PF, Morris, KV, Morillon, A, Rozowsky, JS, Gerstein, MB, Wahlestedt, C, Hayashizaki, Y, Carninci, P, Gingeras, TR, Mattick, JS **(2011)**. The reality of pervasive transcription. *PLoS Biol.*,

[11] Rozowsky, JS, Newburger, D, Sayward, F, Wu, J, Jordan, G, Korbel, JO, Nagalakshmi, U, Yang, J, Zheng, D, Guigó, R, Gingeras, TR, Weissman, S, Miller, P, Snyder, M, Gerstein, MB **(2007)**. The DART classification of unannotated transcription within the ENCODE regions: associating transcription with known and novel loci. *Genome Res.*, 17, 6:732-45.

[12] Lu, ZJ, Yip, KY, Wang, G, Shou, C, Hillier, LW, Khurana, E, Agarwal, A, Auerbach, R, Rozowsky, J, Cheng, C, Kato, M, Miller, DM, Slack, F, Snyder, M, Waterston, RH, Reinke, V, Gerstein, MB **(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, 2:276-85.

[13] Sboner, A, Habegger, L, Pflueger, D, Terry, S, Chen, DZ, Rozowsky, JS, Tewari, AK, Kitabayashi, N, Moss, BJ, Chee, MS, Demichelis, F, Rubin, MA, Gerstein, MB **(2010)**. FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. *Genome Biol.*, 11, 10:R104.

[14] Pflueger, D, Terry, S, Sboner, A, Habegger, L, Esgueva, R, Lin, PC, Svensson, MA, Kitabayashi, N, Moss, BJ, MacDonald, TY, Cao, X, Barrette, T, Tewari, AK, Chee, MS, Chinnaiyan, AM, Rickman, DS, Demichelis, F, Gerstein, MB, Rubin, MA (2011). Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. Genome Res., 21, 1:56-67.

[15] Druker, BJ, Tamura, S, Buchdunger, E, Ohno, S, Segal, GM, Fanning, S, Zimmermann, J, Lydon, NB (1996). Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells. Nat. Med., 2, 5:561-6.

[16] Pei, B, Sisu, C, Frankish, A, Howald, C, Habegger, L, Mu, XJ, Harte, R, Balasubramanian, S, Tanzer, A, Diekhans, M, Reymond, A, Hubbard, TJ, Harrow, J, Gerstein, MB (2012). The GENCODE pseudogene resource. Genome Biol., 13, 9:R51.

[17] Poliseno, L, Salmena, L, Zhang, J, Carver, B, Haveman, WJ, Pandolfi, PP (2010). A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature, 465, 7301:1033-8.
[18] Duret, L, Chureau, C, Samain, S, Weissenbach, J, Avner, P (2006). The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. Science, 312, 5780:1653-5.
[19] Rapicavoli, NA, Qu, K, Zhang, J, Mikhail, M, Laberge, RM, Chang, HY (2013). A mammalian pseudogene lncRNA at the interface of inflammation and anti-inflammatory therapeutics. Elife, 2:e00762.
[20] Rozowsky, J, Abyzov, A, Wang, J, Alves, P, Raha, D, Harmanci, A, Leng, J, Bjornson, R, Kong, Y, Kitabayashi, N, Bhardwaj, N, Rubin, M, Snyder, M, Gerstein, M (2011). AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol. Syst. Biol., 7:522.
[21] ENCODE Consortium **(2012)**. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489, 7414:57-74.
[22] Gerstein, MB, Kundaje, A, Hariharan, M, Landt, SG, Yan, KK, Cheng, C, Mu, XJ, Khurana, E, Rozowsky, J, Alexander, R, Min, R, Alves, P, Abyzov, A, Addleman, N, Bhardwaj, N, Boyle, AP, Cayting, P, Charos, A, Chen, DZ, 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, EC, Patacsil, D, Pauli, F, Raha, D, Ramirez, L, Reddy, TE, Reed, B, Shi, M, Slifer, T, Wang, J, Wu, L, Yang, X, Yip, KY, Zilberman-Schapira, G, Batzoglou, S, Sidow, A, Farnham, PJ, Myers, RM, Weissman, SM, Snyder, M **(2012)**. Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489, 7414:91-100.
[23] Cheng, C, Alexander, R, Min, R, Leng, J, Yip, KY, Rozowsky, J, Yan, KK, Dong, X, Djebali, S, Ruan, Y, Davis, CA, Carninci, P, Lassman, T, Gingeras, TR, Guigó, R, Birney, E, Weng, Z, Snyder, M, Gerstein, M **(2012)**. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res.*, 22, 9:1658-67.