Summary of Gerstein Lab Research

This Document summarizes important contributions made by the Gerstein lab in processing and annotations of various high-throughput sequencing data.

Annotation of non-coding regulatory regions of the genome

We have made a number of contributions in the analysis of the noncoding genome, as part of our extensive 10-year history with the ENCODE and modENCODE projects. Our TF work includes the development of a method called PeakSeq to define the binding peaks of TFs [1], as well as new machine learning techniques [2]. In addition, we have also proposed a probabilistic model, referred to as target identification from profiles (TIP), that identifies a given TF's target genes based on ChIP-seq data [3]. Furthermore, we have developed machine-learning methods that integrate ChIP-seq, chromatin, conservation, sequence and gene annotation data to identify gene-distal enhancers [4], which have been partially validated [5]. We have also constructed regulatory networks for human and model organisms based on the ENCODE [6] and modENCODE datasets [7], and completed many analyses on them [5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22].

Processing of RNA-seq data and annotating ncRNAs

We also have extensive experience conducting integrated analyses of large sets of RNA-seq data, such as through the ENCODE, modENCODE, BrainSpan and exRNA consortia [9, 23, 24, 25, 26]. In particular, 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 [27]. In addition, we have developed IQseq, a computationally efficient method to quantify isoforms for alternatively spliced transcripts [28]. Comparisons between RNA-Seq samples, and to other genome-wide data, will be facilitated in part by our Aggregation and Correlation Toolbox (ACT), which is a general purpose tool for comparing genomic signal tracks [29]. We have also developed a ncRNA finder [30]. Finally, we have developed statistical models relating expression levels to chromatin marks and TF binding [6, 31, 32, 33].

Analysis of Allele-specific expression

A specific class of regulatory variants is one that is related to allele-specific events. These are cis-regulatory variants that are associated with allele-specific binding (ASB), particularly of transcription factors or DNA-binding proteins, and allele-specific expression (ASE) [34, 35]. We have previously developed a tool, AlleleSeq, [20] for the detection of candidate variants associated with ASB and ASE. Using AlleleSeq, we have spearheaded allele-specific analyses in several major consortia publications, including ENCODE and the 1000 Genomes Project [6, 18, 24]. Overall, we found a substantial number of genomic elements associated with ASB and ASE [24]. We also found that these allelic elements are under differential selection from non-allelic ones [6, 18]. By constructing regulatory networks based on ASB of TFs and ASE of their target genes, we further revealed substantial coordination between allele-specific binding and

expression [6] (Fig 1). Furthermore, we have made available on the Internet the AlleleSeq tool, lists of detected allelic variants, and the personal diploid genome and transcriptome of NA12878 [36]. We continually update AlleleSeq, and the resource has been used in the scientific community, as evident in citations and publications using our data as references [37, 38].

FunSeq pipeline to prioritize somatic variants

We have extensively analyzed patterns of variation in non-coding regions along with their coding targets [5, 6, 39]. We used metrics, such as diversity and fraction of rare variants, to characterize selection on various classes and subclasses of functional annotations [39]. In addition, we have also defined variants that are disruptive to TF-binding motifs in regulatory regions [23]. Further studies by our group showed relations between selection and protein network structure, e.g. hubs vs periphery [17, 19].

In a recent study [18], we have integrated and extended these methods to develop a prototype prioritization pipeline called FunSeq (Fig 2). FunSeq identifies sensitive and ultra-sensitive regions, i.e. those annotations under strong selection pressure as determined by human population variation. It also prioritizes variants based on network connectivity and their disruptiveness (e.g. finding motif breakers), identifying deleterious variants in many non-coding functional elements, including TF binding sites, enhancer elements, and regions of open chromatin corresponding to DNase I hypersensitivity sites. In the FunSeq study, we integrated large-scale data from various resources, including ENCODE and 1000 Genomes Project, with cancer genomics data. By contrasting patterns of inherited polymorphisms from 1092 humans with somatic variants from cancer patients, FunSeq identifies candidate non-coding driver mutations [18] (Fig 4). In particular, using FunSeq, we identified ~100 non-coding candidate drivers in 90 medulloblastoma, breast and prostate cancer samples.

Development of efficient tools & calling variants on a large-scale

We have significant experience in developing high-throughput tools for bioinformatics research. Our tools take the forms of web services, distributed open source programs, annotation databases and distributed virtual machines. Many of the latter are hosted on Amazon Web Services Elastic Compute Cloud (AWS-EC2). In particular, for the analysis of high-throughput genomic experiments, we have developed pipelines for analyzing, RNA expression [40, 40, 41], alternative splicing [28], fusion transcripts [42], and copy-number variation [43]. We have developed pipelines for the analysis of regulatory networks [6, 8, 44, 44] and protein–protein interaction networks [12, 45, 46, 47, 48, 49].

We have much experience in large-scale germline variant calling through being active members of the 1000 Genomes Consortium, especially in the analysis working group and the structural variant (SV) and functional interpretation (FIG) subgroups of the consortium where the majority of the variant calling tools are developed and used [50, 51, 52, 53]. We have extensively used [18] Broad Institute's Genome Analysis Toolkit (GATK) [54] for variant calling. We have developed a number of SV calling algorithms,

including BreakSeq, which compares raw reads with a breakpoint library (junction mapping) [55], CNVnator, which measures read depth [56], AGE, which refines local alignment [57], and PEMer, which uses paired ends [58]. We have also developed array-based approaches [59] and a sequencing-based bayesian model [60].

Analysis of recurrent germline & somatic variants (LARVA module)

We have developed a computational framework for identifying these types of recurrent variation, called Large-scale Analysis of Recurrent Variants and Annotations (LARVA). Given a set of cancer whole genome variant calls, and a set of genome annotations, LARVA will pick out the recurrent variants, recurrently mutated annotations, and recurrently mutated subsets of annotations.

LARVA also has a module for computing the statistical significance of its results by simulating the creation of WGS variant calls with randomized variant positions. These random datasets, which otherwise contain the same number of samples and variants, are used to determine the null distribution of variants across the annotation set for comparison with the actual variant data. LARVA determines the positions of variants for its random variant datasets using a null mutation model designed to reflect factors affecting the neutral mutation rates of different genome regions, and represents an extension of an exome null mutation model developed for MutSig [61]. These factors include the (1) genome-wide DNA replication timings, since later-replicating regions are more error-prone due to the depletion of free nucleotides [62], (2) histone marks for H3K4me1 and H3K4me3, because they are anti-correlated with SNV density [63], (3) whole genome RNA-seq data from the ENCODE project [23], representing the connection between expression and transcription-coupled repair [64], and (4) SNV density data from the 1000 Genomes Project [52], which is a proxy for the amount of variation one normally sees in a genomic region.

References

1. Kimchi-Sarfaty, C., Oh, J. M., Kim, I.-W., Sauna, Z. E., Calcagno, A. M., Ambudkar, S. V., & Gottesman, M. M. (2007) A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science* 315, 525-8, PMCID: 17185560.
2. Yip, K. Y. & Gerstein, M. (2009) Training set expansion: an approach to improving the reconstruction of biological networks from limited and uneven reliable interactions. *Bioinformatics* 25, 243-50, PMCID: 19015141.
3. Cheng, C., Min, R., & Gerstein, M. (2011) TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles. *Bioinformatics* 27, 3221-7, PMCID: 22039215.
4. Yip, K. Y., Alexander, R. P., Yan, K.-K., & Gerstein, M. (2010) Improved reconstruction of in silico gene regulatory networks by integrating knockout and perturbation data. *PLoS One* 5, e8121, PMCID: 20126643.
5. Yip, K. Y., Cheng, C., Bhardwaj, N., Brown, J. B., Leng, J., Kundaje, A., Rozowsky, J., Birney, E., Bickel, P., Snyder, M., & Gerstein, M. (2012) Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol* 13, R48, PMCID: 22950945.
6. 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, PMCID: 22955619.
7. Nègre, N., Brown, C. D., Ma, L., Bristow, C. A., Miller, S. W., Wagner, U., Kheradpour, P., Eaton, M. L., Loriaux, P., Sealfon, R., Li, Z., Ishii, H., Spokony, R. F., Chen, J., Hwang, L., Cheng, C., Auburn, R. P., Davis, M. B., Domanus, M., Shah, P. K., Morrison, C. A., Zieba, J., Suchy, S., Senderowicz, L., Victorsen, A., Bild, N. A., Grundstad, A. J., Hanley, D., MacAlpine, D. M., Mannervik, M., Venken, K., Bellen, H., White, R., Gerstein, M., Russell, S., Grossman, R. L., Ren, B., Posakony, J. W., Kellis, M., & White, K. P. (2011) A cis-regulatory map of the Drosophila genome. *Nature* 471, 527-31, PMCID: 21430782.
8. 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, PMCID: 22125477.
9. 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., Dosé, 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, 3rd, D. M., Muroyama, A., Murray, J. I., Ooi, S.-L., Pham, H., Phippen, T., Preston, E. A., 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, 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., modENCODE Consortium, 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, 1775-87, PMCID: 21177976.
10. Yan, K.-K., Fang, G., Bhardwaj, N., Alexander, R. P., & Gerstein, M. (2010) Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks. *Proc Natl Acad Sci U S A* 107, 9186-91, PMCID: 20439753.
11. Yu, H., Greenbaum, D., Xin Lu, H., Zhu, X., & Gerstein, M. (2004) Genomic analysis of essentiality within protein networks. *Trends Genet* 20, 227-31, PMCID: 15145574.
12. Yu, H., Zhu, X., Greenbaum, D., Karro, J., & Gerstein, M. (2004) TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. *Nucleic Acids Res* 32, 328-37, PMCID: 14724320.
13. 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, PMCID: 17447836.
14. Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A., & Gerstein, M. (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* 431, 308-12, PMCID: 15372033.
15. Gianoulis, T. A., Raes, J., Patel, P. V., Bjornson, R., Korbel, J. O., Letunic, I., Yamada, T., Paccanaro, A., Jensen, L. J., Snyder, M., Bork, P., & Gerstein, M. B. (2009) Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proc Natl Acad Sci U S A* 106, 1374-9, PMCID: 19164758.
16. Yu, H., Paccanaro, A., Trifonov, V., & Gerstein, M. (2006) Predicting interactions in protein networks by completing defective cliques. *Bioinformatics* 22, 823-9, PMCID: 16455753.
17. Kim, P. M., Korbel, J. O., & Gerstein, M. B. (2007) Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. *Proc Natl Acad Sci U S A* 104, 20274-9, PMCID: 18077332.
18. Khurana, E., Fu, Y., Colonna, V., Mu, X. J., Kang, H. M., Lappalainen, T.,

Sboner, A., Lochovsky, L., Chen, J., Harmanci, A., Das, J., Abyzov, A., Balasubramanian, S., Beal, K., Chakravarty, D., Challis, D., Chen, Y., Clarke, D., Clarke, L., Cunningham, F., Evani, U. S., Flicek, P., Fragoza, R., Garrison, E., Gibbs, R., Gümüs, Z. H., Herrero, J., Kitabayashi, N., Kong, Y., Lage, K., Liluashvili, V., Lipkin, S. M., MacArthur, D. G., Marth, G., Muzny, D., Pers, T. H., Ritchie, G. R. S., Rosenfeld, J. A., Sisu, C., Wei, X., Wilson, M., Xue, Y., Yu, F., 1000 Genomes Project Consortium, Dermitzakis, E. T., Yu, H., Rubin, M. A., Tyler-Smith, C., & Gerstein, M. (2013) Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* 342, 1235587, PMCID: 24092746.
19. Khurana, E., Fu, Y., Chen, J., & Gerstein, M. (2013) Interpretation of genomic variants using a unified biological network approach. *PLoS Comput Biol* 9, e1002886, PMCID: 23505346.
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, PMCID: 21811232.
21. Lörcher, U., Peters, J., & Kollath, J. (1990) [Changes in the lungs and pleura following chemoembolization of liver tumors with mitomycin-lipiodol]. *Rofo* 152, 569-73, PMCID: 2160691.
22. 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, PMCID: 21253555.
23. ENCODE Project Consortium, Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., & Snyder, M. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57-74, PMCID: 22955616.
24. Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Bar, N. S., 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, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O. J., Park, E., Persaud, K., Preall, J. B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A. M., 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, S. E., Hannon, G., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigó, R., & Gingeras, T. R. (2012) Landscape of transcription in human cells. *Nature* 489, 101-8, PMCID: 22955620.
25. http://brainspan.org () Last accessed on 15th January 2014. , , PMCID: .
26. http://exRNA.org () Last accessed on 15th January 2014. , , PMCID: .
27. Habegger, L., Sboner, A., Gianoulis, T. A., 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, 281-3, PMCID: 21134889.

28. 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, PMCID: 22238592.
29. Jee, J., Rozowsky, J., Yip, K. Y., 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, 1152-4, PMCID: 21349863.
30. 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, PMCID: 21177971.
31.Cheng, C., Alexander, R., Min, R., Leng, J., Yip, K. Y., Rozowsky, J., Yan, K.-K., Dong, X., Djebali, S., Ruan, Y., Davis, C. A., Carninci, P., Lassman, T., Gingeras, T. R., 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, 1658-67, PMCID: 22955978.
32. Cheng, C., Shou, C., Yip, K. Y., & Gerstein, M. B. (2011) Genome-wide analysis of chromatin features identifies histone modification sensitive and insensitive yeast transcription factors. *Genome Biol* 12, R111, PMCID: 22060676.
33. Cheng, C. & Gerstein, M. (2012) Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Res* 40, 553-68, PMCID: 21926158.
34. Pastinen, T. (2010) Genome-wide allele-specific analysis: insights into regulatory variation. *Nat Rev Genet* 11, 533-8, PMCID: 20567245.
35. Birney, E., Lieb, J. D., Furey, T. S., Crawford, G. E., & Iyer, V. R. (2010) Allele-specific and heritable chromatin signatures in humans. *Hum Mol Genet* 19, R204-9, PMCID: 20846943.
36. http://alleleseq.gersteinlab.org () Last accessed on 15th January 2014. , , PMCID: .
37. Ji, H., Li, X., Wang, Q.-f., & Ning, Y. (2013) Differential principal component analysis of ChIP-seq. *Proc Natl Acad Sci U S A* 110, 6789-94, PMCID: 23569280.
38. Younesy, H., Möller, T., Heravi-Moussavi, A., Cheng, J. B., Costello, J. F., Lorincz, M. C., Karimi, M. M., & Jones, S. J. M. (2013) ALEA: a toolbox for allele-specific epigenomics analysis. *Bioinformatics* , , PMCID: 24371156.
39. Mu, X. J., Lu, Z. J., Kong, Y., Lam, H. Y. K., & Gerstein, M. B. (2011) Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. *Nucleic Acids Res* 39, 7058-76, PMCID: 21596777.
40. Greenbaum, D., Colangelo, C., Williams, K., & Gerstein, M. (2003) Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol* 4, 117, PMCID: 12952525.
41. Wang, Z., Gerstein, M., & Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10, 57-63, PMCID: 19015660.
42. 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, PMCID: 20964841.
43. Kim, P. M., Lam, H. Y. K., Urban, A. E., Korbel, J. O., Affourtit, J., Grubert, F., Chen, X., Weissman, S., Snyder, M., & Gerstein, M. B. (2008) Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. *Genome Res* 18, 1865-74, PMCID: 18842824.
44. Qian, J., Lin, J., Luscombe, N. M., Yu, H., & Gerstein, M. (2003) Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics* 19, 1917-26, PMCID: 14555624.
45. Yip, K. Y., Yu, H., Kim, P. M., Schultz, M., & Gerstein, M. (2006) The tYNA platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks. *Bioinformatics* 22, 2968-70, PMCID: 17021160.
46. Clarke, D., Bhardwaj, N., & Gerstein, M. B. (2012) Novel insights through the integration of structural and functional genomics data with protein networks. *J Struct Biol* 179, 320-6, PMCID: 22343087.
47. Bhardwaj, N., Clarke, D., & Gerstein, M. (2011) Systematic control of protein interactions for systems biology. *Proc Natl Acad Sci U S A* 108, 20279-80, PMCID: 22160691.
48. Bhardwaj, N., Abyzov, A., Clarke, D., Shou, C., & Gerstein, M. B. (2011) Integration of protein motions with molecular networks reveals different mechanisms for permanent and transient interactions. *Protein Sci* 20, 1745-54, PMCID: 21826754.
49. Fasolo, J., Sboner, A., Sun, M. G. F., Yu, H., Chen, R., Sharon, D., Kim, P. M., Gerstein, M., & Snyder, M. (2011) Diverse protein kinase interactions identified by protein microarrays reveal novel connections between cellular processes. *Genes Dev* 25, 767-78, PMCID: 21460040.
50. Zhang, Z. D., Du, J., Lam, H., Abyzov, A., Urban, A. E., Snyder, M., & Gerstein, M. (2011) Identification of genomic indels and structural variations using split reads. *BMC Genomics* 12, 375, PMCID: 21787423.
51. Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., Abyzov, A., Yoon, S. C., Ye, K., Cheetham, R. K., Chinwalla, A., Conrad, D. F., Fu, Y., Grubert, F., Hajirasouliha, I., Hormozdiari, F., Iakoucheva, L. M., Iqbal, Z., Kang, S., Kidd, J. M., Konkel, M. K., Korn, J., Khurana, E., Kural, D., Lam, H. Y. K., Leng, J., Li, R., Li, Y., Lin, C.-Y., Luo, R., Mu, X. J., Nemesh, J., Peckham, H. E., Rausch, T., Scally, A., Shi, X., Stromberg, M. P., Stütz, A. M., Urban, A. E., Walker, J. A., Wu, J., Zhang, Y., Zhang, Z. D., Batzer, M. A., Ding, L., Marth, G. T., McVean, G., Sebat, J., Snyder, M., Wang, J., Ye, K., Eichler, E. E., Gerstein, M. B., Hurles, M. E., Lee, C., McCarroll, S. A., Korbel, J. O., & 1000 Genomes Project (2011) Mapping copy number variation by population-scale genome sequencing. *Nature* 470, 59-65, PMCID: 21293372.
52. 1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., & McVean, G. A. (2010) A map of human genome variation from population-scale sequencing. *Nature* 467, 1061-73, PMCID: 20981092.
53. 1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D.,

DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., & McVean, G. A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56-65, PMCID: 23128226.
54. DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43, 491-8, PMCID: 21478889.
55. Lam, H. Y. K., Mu, X. J., Stütz, A. M., Tanzer, A., Cayting, P. D., Snyder, M., Kim, P. M., Korbel, J. O., & Gerstein, M. B. (2010) Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat Biotechnol* 28, 47-55, PMCID: 20037582.
56. Abyzov, A., Urban, A. E., Snyder, M., & Gerstein, M. (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* 21, 974-84, PMCID: 21324876.
57. Abyzov, A. & Gerstein, M. (2011) AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. *Bioinformatics* 27, 595-603, PMCID: 21233167.
58. Korbel, J. O., Abyzov, A., Mu, X. J., Carriero, N., Cayting, P., Zhang, Z., Snyder, M., & Gerstein, M. B. (2009) PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol* 10, R23, PMCID: 19236709.
59. Wang, L.-Y., Abyzov, A., Korbel, J. O., Snyder, M., & Gerstein, M. (2009) MSB: a mean-shift-based approach for the analysis of structural variation in the genome. *Genome Res* 19, 106-17, PMCID: 19037015.
60. Zhang, Z. D. & Gerstein, M. B. (2010) Detection of copy number variation from array intensity and sequencing read depth using a stepwise Bayesian model. *BMC Bioinformatics* 11, 539, PMCID: 21034510.
61. Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., Kiezun, A., Hammerman, P. S., McKenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., Stransky, N., Helman, E., Kim, J., Sougnez, C., Ambrogio, L., Nickerson, E., Shefler, E., Cortés, M. L., Auclair, D., Saksena, G., Voet, D., Noble, M., DiCara, D., Lin, P., Lichtenstein, L., Heiman, D. I., Fennell, T., Imielinski, M., Hernandez, B., Hodis, E., Baca, S., Dulak, A. M., Lohr, J., Landau, D.-A., Wu, C. J., Melendez-Zajgla, J., Hidalgo-Miranda, A., Koren, A., McCarroll, S. A., Mora, J., Lee, R. S., Crompton, B., Onofrio, R., Parkin, M., Winckler, W., Ardlie, K., Gabriel, S. B., Roberts, C. W. M., Biegel, J. A., Stegmaier, K., Bass, A. J., Garraway, L. A., Meyerson, M., Golub, T. R., Gordenin, D. A., Sunyaev, S., Lander, E. S., & Getz, G. (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499, 214-8, PMCID: 23770567.
62. Chen, C.-L., Rappailles, A., Duquenne, L., Huvet, M., Guilbaud, G., Farinelli, L., Audit, B., d'Aubenton-Carafa, Y., Arneodo, A., Hyrien, O., & Thermes, C. (2010) Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Res* 20, 447-57, PMCID: 20103589.

63. Schuster-Böckler, B. & Lehner, B. (2012) Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* 488, 504-7, PMCID: 22820252.
64. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M. F., Monahan, J. E., Morais, P., Meltzer, J., Korejwa, A., Jané-Valbuena, J., Mapa, F. A., Thibault, J., Bric-Furlong, E., Raman, P., Shipway, A., Engels, I. H., Cheng, J., Yu, G. K., Yu, J., Aspesi, Jr, P., de Silva, M., Jagtap, K., Jones, M. D., Wang, L., Hatton, C., Palescandolo, E., Gupta, S., Mahan, S., Sougnez, C., Onofrio, R. C., Liefeld, T., MacConaill, L., Winckler, W., Reich, M., Li, N., Mesirov, J. P., Gabriel, S. B., Getz, G., Ardlie, K., Chan, V., Myer, V. E., Weber, B. L., Porter, J., Warmuth, M., Finan, P., Harris, J. L., Meyerson, M., Golub, T. R., Morrissey, M. P., Sellers, W. R., Schlegel, R., & Garraway, L. A. (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483, 603-7, PMCID: 22460905.