**Gerstein lab experience in general transcriptomic analysis**

We have expertise in data analytics and leading large-scale national projects whose focus is to develop and disseminate novel tools and methods. For instance, we have developed methods for normalization, analysis, and comparison of bulk RNA-seq profiles (Lu et al. 2011; Jee et al, 2011). Furthermore, we have developed multiple advanced mathematical and machine learning methods covering a variety of key tasks to analyze large-scale snRNA-seq data, including but not limited to preprocessing, imputation, quantification, visualization, clustering, and comparisons. As described in detail in the following sections, these tools and methods will be extended and adapted to develop the different modules of snRNA-seq pipeline.

**Gerstein lab experience in epigenetic profiling**

We have extensive experience processing epigenetic profiling data. For example, we have developed PeakSeq (Rozowsky et al, 2009) for the genome-wide identification of TF binding sites from ChIP-Seq data, which is used by ENCODE. MUSIC (Harmanci et al, 2014) is a peak caller that performs multiscale decomposition of ChIP-Seq signal. We have also proposed numerous models to utilize ATAC-seq data to identify functional elements (e.g. enhancers), and to construct regulatory networks from human brain tissue samples. In this proposal, we will now expand these analyses to build pipelines for scATAC-seq data. These will serve as reference annotations for further identifying cell-type-specific functional elements and gene regulatory networks.

**Gerstein lab experience in functional genomics integration with multiple machine learning models**

We have extensive experience with integration of functional genomics data via various types of machine learning models. For example, we integrated ENCODE data on TF binding, histone modifications, and target gene expression to establish regulatory relationships using a probabilistic model we named TIP (Target Identification from Profiles) (Cheng et al, 2011). We identified potential enhancers from distal gene regions and we used these modules to quantify the relationship between TF binding and gene expression (Cheng et al, 2011). We integrated these data types with protein-protein interaction and transcriptional regulatory networks, allowing us to group TFs into histone-sensitive and histone-insensitive classes, thereby refining the prediction of gene-regulation targets and effects (Dong et al, 2012).

**Gerstein lab experience within the RNA communication consortium (ERCC) and other consortia (PsychENCODE, ENCODE, and GTEx)**

We have considerable experience working in consortia, both in analysis and in data generation capacities. We have worked within consortia to develop data quality standards. Our data processing strategies are based on well-known and established standards, such as those used by other consortia (including PsychENCODE, ENCODE, and GTEx).

As part of the extracellular RNA communication consortium (ERCC), we developed a custom analysis pipeline, the extracellular RNA processing toolkit (excerpt; Rozowsky et al, 2019), which has been used over 80,000 times. It performs sequential alignment of RNA to contaminants, then to human transcriptome and genome sequences, then to human repetitive elements, and finally to exogenous sequences. The key idea behind the pipeline is to conservatively find transcripts present in trace amounts, and not be fooled by the millions of potential contaminants and errors in a typical next-generation sequencing run.

We have played a leading role in a number of consortia that integrate multi-omic datasets. Recently, we led analysis projects of the psychENCODE and Brainspan (Wang et al, 2018). We played a leading role in the integrative analysis of multi-omic datasets from the ENCODE (Djebali et al, 2012; Gerstein et al, 2014; Boyle et al, 2014) and modENCODE (Boyle et al, 2014; Gerstein et al, 2010) consortia. We developed approaches for constructing and studying biological networks that can be applied to analyze ENCODE datasets. We integrated multiple genomic datasets to construct gene regulatory networks consisting of various regulatory factors including transcription factors (TFs) and microRNAs and their target genes (Boyle et al, 2014; Gerstein et al, 2012; Cheng et al, 2011). For constructed gene regulatory networks, we developed methods to construct and analyze human and model organism gene regulatory networks (Gerstein et al, 2012; Yan et al, 2010) using ENCODE and modENCODE datasets. We also analyzed hierarchical structures of gene regulatory networks and found that hierarchy rather than centrality ("hubiness") better reflects the importance of regulators.

We helped lead the structural variation (SV) analysis for the 1,000 Genomes Project (Khurana et al, 2013). We developed an annotation pipeline that maps variants to genes (Balasubramanian et al, 2017. We also developed algorithms to identify indels and SVs, and studied the mechanisms of SV generation (Zhang et al, 2017). We performed SV mechanism annotations for the 1,000 Genomes Project Phase 3 deletions using BreakSeq (Lam et al, 2010). We have been an integral part of the Data Integration and Analysis Component of the Data Management and Resource Repository for the NIH Common Fund Extracellular RNA Communication Consortium (Freedman et al, 2016; Cheung et al, 2016). In addition, we participated in the United States Department of Energy Systems Biology Knowledgebase (Arkin et al, 2018), which is an open-source software and data platform that enables data sharing, integration and analysis of microbes, plants and their communities; and the Northeast Structural Genomics Consortium (Huang et al, 2008), which employs both X-ray crystallography and nuclear magnetic resonance spectroscopy to provide novel structural information useful in modeling thousands protein domains.

Above, we have detailed our experience in working with other consortia. Here, we simply list a number of analysis tools (some of which are mentioned above) that we have developed in consortia-wide frameworks:

*Listing of example tools & integration approaches:*

* ***Identifying targets of TFs using ENCODE data***: We have extensive experience integrating of functional genomics data using various types of machine learning models. For example, we integrated data – obtained as part of our involvement with the ENCODE consortium – on TF binding, histone modifications, and target gene expression to establish regulatory relationships using **TIP** (Cheng et al, 2011), a probabilistic model that we have developed for identifying TF target genes from ChIP-seq binding profiles.
* ***Predicting enhancers using ENCODE data***: In a further step toward establishing regulatory relationships and their effects as part of our collaborative efforts with the ENCODE consortium, we used ENCODE data on gene expression and TF binding signals in a framework we developed for identifying potential enhancers from distal gene regions. We then used these predictions to quantify the relationships between TF binding and gene expression (Cheng et al, 2012).
* ***Predicting gene expression levels from chromatin features in ENCODE***: Using ENCODE data, we integrated RNA-seq data and chromatin features in multiple cell lines, DNase I hypersensitivity site information, and other data types with protein-protein interaction and transcriptional regulatory networks to group TFs into histone-sensitive and histone-insensitive classes. These classification may be used to refine the prediction of gene-regulation targets and effects (Dong et al, 2012).
* ***Finding TF binding sites using ChIP-seq data:*** We have extensive experience processing epigenetic profiling data from tissue. For example, we have developed **PeakSeq** (Rozowsky et al, 2009) for the genome-wide identification of TF binding sites from ChIP-seq data. This framework had been adopted by ENCODE, and has become the standard used by ENCODE for many years. The scoring scheme implemented and first introduced by PeakSeq enables users to optimize experimental design strategies by providing estimates of the sequencing depth needed for a given level of coverage, and by showing that more than 2 replicates may often confer limited additional information, thereby obviating the need for additional assays.
* ***Multiscale identification of enriched regions in ChIP-seq ENCODE data***: We developed **MUSIC** (Harmanci et al, 2014), a peak caller that performs multiscale decomposition of ENCODE ChIP-seq signals. Here, the term “multiscale decomposition” simply refers to the tool’s ability to identify enriched regions at multiple length scales. Thus, MUSIC is applicable to histone modifications enabling detection of broad and punctate regions, and ChIP-seq assays typically provide a very wide range of possible signal scales

**Bibliography and References Cited**

Arkin AP, Cottingham RW, Henry CS, Harris NL, Stevens RL, Maslov S, Dehal P, Ware D, Perez F, Canon S, Sneddon MW, Henderson ML, Riehl WJ, Murphy-Olson D, Chan SY, Kamimura RT, Kumari S, Drake MM, Brettin TS, Glass EM, Chivian D, Gunter D, Weston DJ, Allen BH, Baumohl J, Best AA, Bowen B, Brenner SE, Bun CC, Chandonia JM, Chia JM, Colasanti R, Conrad N, Davis JJ, Davison BH, DeJongh M, Devoid S, Dietrich E, Dubchak I, Edirisinghe JN, Fang G, Faria JP, Frybarger PM, Gerlach W, Gerstein M, Greiner A, Gurtowski J, Haun HL, He F, Jain R, Joachimiak MP, Keegan KP, Kondo S, Kumar V, Land ML, Meyer F, Mills M, Novichkov PS, Oh T, Olsen GJ, Olson R, Parrello B, Pasternak S, Pearson E, Poon SS, Price GA, Ramakrishnan S, Ranjan P, Ronald PC, Schatz MC, Seaver SMD, Shukla M, Sutormin RA, Syed MH, Thomason J, Tintle NL, Wang D, Xia F, Yoo H, Yoo S, Yu D. KBase: The United States Department of Energy Systems Biology Knowledgebase. Nat Biotechnol. 2018;36(7):566-9. Epub 2018/07/07. doi: 10.1038/nbt.4163. PubMed PMID: 29979655.

Balasubramanian S, Fu Y, Pawashe M, McGillivray P, Jin M, Liu J, Karczewski KJ, MacArthur DG, Gerstein M. Using ALoFT to determine the impact of putative loss-of-function variants in protein-coding genes. Nat Commun. 2017;8(1):382. Epub 2017/08/31. doi: 10.1038/s41467-017-00443-5. PubMed PMID: 28851873; PMCID: PMC5575292.

Boyle AP, Araya CL, Brdlik C, Cayting P, Cheng C, Cheng Y, Gardner K, Hillier LW, Janette J, Jiang L, Kasper D, Kawli T, Kheradpour P, Kundaje A, Li JJ, Ma L, Niu W, Rehm EJ, Rozowsky J, Slattery M, Spokony R, Terrell R, Vafeados D, Wang D, Weisdepp P, Wu YC, Xie D, Yan KK, Feingold EA, Good PJ, Pazin MJ, Huang H, Bickel PJ, Brenner SE, Reinke V, Waterston RH, Gerstein M, White KP, Kellis M, Snyder M. Comparative analysis of regulatory information and circuits across distant species. Nature. 2014;512(7515):453-6. Epub 2014/08/29. doi: 10.1038/nature13668. PubMed PMID: 25164757; PMCID: PMC4336544.

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, Guigo R, Birney E, Weng Z, Snyder M, Gerstein M. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. Genome Res. 2012;22(9):1658-67. Epub 2012/09/08. doi: 10.1101/gr.136838.111. PubMed PMID: 22955978; PMCID: PMC3431483.

Cheng C, Min R, Gerstein M. TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles. Bioinformatics. 2011;27(23):3221-7. Epub 2011/11/01. doi: 10.1093/bioinformatics/btr552. PubMed PMID: 22039215; PMCID: 3223362.

Cheng C, Yan KK, Hwang W, Qian J, Bhardwaj N, Rozowsky J, Lu ZJ, Niu W, Alves P, Kato M, Snyder M, Gerstein M. Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. PLoS Comput Biol. 2011;7(11):e1002190. Epub 2011/11/30. doi: 10.1371/journal.pcbi.1002190. PubMed PMID: 22125477; PMCID: PMC3219617.

Cheung KH, Keerthikumar S, Roncaglia P, Subramanian SL, Roth ME, Samuel M, Anand S, Gangoda L, Gould S, Alexander R, Galas D, Gerstein MB, Hill AF, Kitchen RR, Lotvall J, Patel T, Procaccini DC, Quesenberry P, Rozowsky J, Raffai RL, Shypitsyna A, Su AI, Thery C, Vickers K, Wauben MH, Mathivanan S, Milosavljevic A, Laurent LC. Extending gene ontology in the context of extracellular RNA and vesicle communication. J Biomed Semantics. 2016;7:19. Epub 2016/04/15. doi: 10.1186/s13326-016-0061-5. PubMed PMID: 27076901; PMCID: PMC4830068.

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, Roder 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, Guigo R, Gingeras TR. Landscape of transcription in human cells. Nature. 2012;489(7414):101-8. Epub 2012/09/08. doi: 10.1038/nature11233. PubMed PMID: 22955620; PMCID: PMC3684276.

Dong X, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, Gingeras TR, Gerstein M, Guigo R, Birney E, Weng Z. Modeling gene expression using chromatin features in various cellular contexts. Genome Biol. 2012;13(9):R53. Epub 2012/09/07. doi: 10.1186/gb-2012-13-9-r53. PubMed PMID: 22950368; PMCID: PMC3491397.

Freedman JE, Gerstein M, Mick E, Rozowsky J, Levy D, Kitchen R, Das S, Shah R, Danielson K, Beaulieu L, Navarro FC, Wang Y, Galeev TR, Holman A, Kwong RY, Murthy V, Tanriverdi SE, Koupenova-Zamor M, Mikhalev E, Tanriverdi K. Diverse human extracellular RNAs are widely detected in human plasma. Nat Commun. 2016;7:11106. Epub 2016/04/27. doi: 10.1038/ncomms11106. PubMed PMID: 27112789; PMCID: PMC4853467.

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 JJ, 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. Architecture of the human regulatory network derived from ENCODE data. Nature. 2012;489(7414):91-100. Epub 2012/09/08. doi: 10.1038/nature11245. PubMed PMID: 22955619; PMCID: PMC4154057.

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, Dose 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, 3rd, Muroyama A, Murray JI, Ooi SL, Pham H, Phippen T, Preston EA, 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 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, mod EC, 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. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science. 2010;330(6012):1775-87. Epub 2010/12/24. doi: 10.1126/science.1196914. PubMed PMID: 21177976; PMCID: PMC3142569.

Gerstein MB, Rozowsky J, Yan KK, Wang D, Cheng C, Brown JB, Davis CA, Hillier L, Sisu C, Li JJ, Pei B, Harmanci AO, Duff MO, Djebali S, Alexander RP, Alver BH, Auerbach R, Bell K, Bickel PJ, Boeck ME, Boley NP, Booth BW, Cherbas L, Cherbas P, Di C, Dobin A, Drenkow J, Ewing B, Fang G, Fastuca M, Feingold EA, Frankish A, Gao G, Good PJ, Guigo R, Hammonds A, Harrow J, Hoskins RA, Howald C, Hu L, Huang H, Hubbard TJ, Huynh C, Jha S, Kasper D, Kato M, Kaufman TC, Kitchen RR, Ladewig E, Lagarde J, Lai E, Leng J, Lu Z, MacCoss M, May G, McWhirter R, Merrihew G, Miller DM, Mortazavi A, Murad R, Oliver B, Olson S, Park PJ, Pazin MJ, Perrimon N, Pervouchine D, Reinke V, Reymond A, Robinson G, Samsonova A, Saunders GI, Schlesinger F, Sethi A, Slack FJ, Spencer WC, Stoiber MH, Strasbourger P, Tanzer A, Thompson OA, Wan KH, Wang G, Wang H, Watkins KL, Wen J, Wen K, Xue C, Yang L, Yip K, Zaleski C, Zhang Y, Zheng H, Brenner SE, Graveley BR, Celniker SE, Gingeras TR, Waterston R. Comparative analysis of the transcriptome across distant species. Nature. 2014;512(7515):445-8. Epub 2014/08/29. doi: 10.1038/nature13424. PubMed PMID: 25164755; PMCID: PMC4155737.

Harmanci A, Rozowsky J, Gerstein M. MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework. Genome Biol. 2014;15(10):474. Epub 2014/10/09. doi: 10.1186/s13059-014-0474-3. PubMed PMID: 25292436; PMCID: PMC4234855.

Huang YJ, Hang D, Lu LJ, Tong L, Gerstein MB, Montelione GT. Targeting the human cancer pathway protein interaction network by structural genomics. Mol Cell Proteomics. 2008;7(10):2048-60. Epub 2008/05/20. doi: 10.1074/mcp.M700550-MCP200. PubMed PMID: 18487680; PMCID: PMC2559933.

Jee J, Rozowsky J, Yip KY, Lochovsky L, Bjornson R, Zhong G, Zhang Z, Fu Y, Wang J, Weng Z, Gerstein M. ACT: aggregation and correlation toolbox for analyses of genome tracks. Bioinformatics. 2011;27(8):1152-4. Epub 2011/02/26. doi: 10.1093/bioinformatics/btr092. PubMed PMID: 21349863; PMCID: 3072554.

Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, 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 US, Flicek P, Fragoza R, Garrison E, Gibbs R, Gumus ZH, Herrero J, Kitabayashi N, Kong Y, Lage K, Liluashvili V, Lipkin SM, MacArthur DG, Marth G, Muzny D, Pers TH, Ritchie GRS, Rosenfeld JA, Sisu C, Wei X, Wilson M, Xue Y, Yu F, Genomes Project C, Dermitzakis ET, Yu H, Rubin MA, Tyler-Smith C, Gerstein M. Integrative annotation of variants from 1092 humans: application to cancer genomics. Science. 2013;342(6154):1235587. Epub 2013/10/05. doi: 10.1126/science.1235587. PubMed PMID: 24092746; PMCID: PMC3947637.

Lam HY, Mu XJ, Stutz AM, Tanzer A, Cayting PD, Snyder M, Kim PM, Korbel JO, Gerstein MB. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. Nat Biotechnol. 2010;28(1):47-55. Epub 2009/12/29. doi: 10.1038/nbt.1600. PubMed PMID: 20037582; PMCID: PMC2951730.

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. 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):276-85. Epub 2010/12/24. doi: 10.1101/gr.110189.110. PubMed PMID: 21177971; PMCID: PMC3032931.

Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol. 2009;27(1):66-75. Epub 2009/01/06. doi: 10.1038/nbt.1518. PubMed PMID: 19122651; PMCID: PMC2924752.

Rozowsky J, Kitchen RR, Park JJ, Galeev TR, Diao J, Warrell J, Thistlethwaite W, Subramanian SL, Milosavljevic A, Gerstein M. exceRpt: A Comprehensive Analytic Platform for Extracellular RNA Profiling. Cell Syst. 2019;8(4):352-7 e3. Epub 2019/04/09. doi: 10.1016/j.cels.2019.03.004. PubMed PMID: 30956140.

Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FCP, Clarke D, Gu M, Emani P, Yang YT, Xu M, Gandal MJ, Lou S, Zhang J, Park JJ, Yan C, Rhie SK, Manakongtreecheep K, Zhou H, Nathan A, Peters M, Mattei E, Fitzgerald D, Brunetti T, Moore J, Jiang Y, Girdhar K, Hoffman GE, Kalayci S, Gumus ZH, Crawford GE, Psych EC, Roussos P, Akbarian S, Jaffe AE, White KP, Weng Z, Sestan N, Geschwind DH, Knowles JA, Gerstein MB. Comprehensive functional genomic resource and integrative model for the human brain. Science. 2018;362(6420). Epub 2018/12/14. doi: 10.1126/science.aat8464. PubMed PMID: 30545857.

Yan KK, Fang G, Bhardwaj N, Alexander RP, Gerstein M. 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. 2010;107(20):9186-91. Epub 2010/05/05. doi: 10.1073/pnas.0914771107. PubMed PMID: 20439753; PMCID: PMC2889091.

Zhang Y, Li S, Abyzov A, Gerstein MB. Landscape and variation of novel retroduplications in 26 human populations. PLoS Comput Biol. 2017;13(6):e1005567. Epub 2017/07/01. doi: 10.1371/journal.pcbi.1005567. PubMed PMID: 28662076; PMCID: PMC5510864.