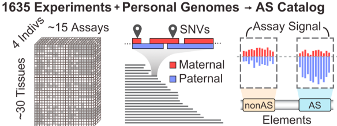
***Gerstein lab experience with non-coding genomic annotations***: We have extensive experience in leading large-scale consortium projects to develop and disseminate novel tools for genome annotation. Many of our tools are tailored for processing functional genomics data to produce annotations for non-coding elements. For example, we developed **PeakSeq** for the genome-wide identification of TF binding sites from ChIP-seq data24. This framework was adopted by the ENCODE Consortium in the uniform ChIP-Seq processing pipeline8. Furthermore, our group has led efforts in the GENCODE consortium to annotate pseudogenes, inactive counterparts of functional protein-coding genes, in new reference genome releases. Over the years, using a combination of manual curation and computational analysis, we have identified and characterized pseudogenes in humans25,26 and multiple species27,28 using our **PseudoPipe** software package29. Through this package, we have identified over 14,000 pseudogenes in the human genome30 and uncovered specific biology linking pseudogenes to fixed loss-of-function mutations in the genome. We also contributed to the Pan-cancer Analysis of Whole Genomes (PCAWG) Consortium by creating custom annotations for cancer genomes, leveraging eCLIP, Hi-C, and whole-genome STARR-seq assays for data-rich ENCODE cell types, including networks of transcription factors and RNA-binding proteins31. Similarly, we designed standardized pipelines for uniform processing of bulk and single-cell RNA-Seq and ATAC-Seq data in human brain tissues for the PsychENCODE Consortium, which allowed us to find tissue- and cell-type-specific regulatory elements and QTLs for individuals with psychiatric diseases32,33.

We recently led an effort within the ENCODE and GTEx consortia to develop the EN-TEx resource, consisting of linked sets of over 1600 functional genomics assays (RNA-Seq, DNAse-Seq and ATAC-Seq, ChIP-Seq, Hi-C, proteomics, and methylation assays) across 30 tissues in four individuals22 (**Fig. 2**). We performed an extensive evaluation to determine the best methods for mapping and quantifying various types of genomic data22. Specifically, we mapped 1,635 sequencing experiments (15 assays in 30 tissues) to phased diploid and reference genomes, resulting in three signal tracks for each assay (two haplotypes and the reference). We discovered that mapping sequences to the haplotypes of each sample instead of the reference genome led to an overall improvement in mapping accuracy for each assay. By applying conventional mapping criteria, we noted a rise in the number of mapped reads by approximately 0.5–1% in the haplotype-mapped data; with stricter filtering criteria for high-quality and uniquely mapping sequences, we observed a significantly larger improvement of 2–4% more mapped reads across the four individuals. This enabled us to assess variant effects on gene regulation more accurately, particularly in structural variant regions. In fact, we identified 300 SV-QTLs that often co-occurred with disrupted chromatin peaks, and further observed depletion of open chromatin regions near transposable elements22**.** These findings underscore the value of gaining insights into genomic function by annotating non-coding elements on individual haplotypes rather than the reference genome. We hypothesize that aligning these data to the human pangenome graph would yield even higher read mapping rates and identify more non-coding annotations on non-reference sequences (i.e., structural variants on individual haplotypes).

**Fig. 2.** Catalog of AS events identified in the EN-TEx resource22.

***Gerstein lab experience with Allele-specific analysis***: We have spearheaded AS analyses in several major consortia publications, including ENCODE and the 1000 Genomes Project22,56–60. For example, we annotated allele-specific expression (ASE) and binding (ASB) events by aligning functional sequencing data57 (955 RNA-Seq and 165 ChIP-Seq in total) to individual haplotypes61. Overall, we detected >6K and 63K SNVs associated with ASB and ASE, respectively, which we made available as an online resource, AlleleDB57. We further constructed a high-resolution map of allelic imbalances in DNA methylation, histone marks, and transcription in 71 epigenomes from 36 distinct cell and tissue types from 13 donors60. To measure ASE, ASB, and AS chromatin activity (ASCA) in diploid genomes, we developed a pipeline dubbed AlleleSeq2 to incorporate all variants present on each haplotype to account for reference bias in a straightforward way56,57,62, and included additional filters to mitigate ambiguous read mapping57,63. To account for the over-dispersed nature of functional genomics read count data, the significance of allelic imbalance is assessed using the beta-binomial test57. In our analysis of the EN-TEx resource, we found that between 1.1-7.3% of assay reads across all samples that preferentially aligned to a haplotype overlapped with heterozygous SNVs22. Using AlleleSeq2, we generated an annotation catalog of >1 million allele-specific events from EN-TEx experiments22 (**Fig. 2**). These events exhibit coordinated activity along haplotypes and are less conserved than the corresponding non-allele-specific SNVs. Combining these data with existing annotations also revealed strong associations between allele-specific and GWAS loci22.

**REFERENCES**

8. Hitz BC, Jin-Wook L, Jolanki O, Kagda MS, Graham K, Sud P, Gabdank I, Strattan JS, Sloan CA, Dreszer T, Rowe LD, Podduturi NR, Malladi VS, Chan ET, Davidson JM, Ho M, Miyasato S, Simison M, Tanaka F, Luo Y, Whaling I, Hong EL, Lee BT, Sandstrom R, Rynes E, Nelson J, Nishida A, Ingersoll A, Buckley M, Frerker M, Kim DS, Boley N, Trout D, Dobin A, Rahmanian S, Wyman D, Balderrama-Gutierrez G, Reese F, Durand NC, Dudchenko O, Weisz D, Rao SSP, Blackburn A, Gkountaroulis D, Sadr M, Olshansky M, Eliaz Y, Nguyen D, Bochkov I, Shamim MS, Mahajan R, Aiden E, Gingeras T, Heath S, Hirst M, Kent WJ, Kundaje A, Mortazavi A, Wold B, Cherry JM. The ENCODE Uniform Analysis Pipelines. bioRxiv [Preprint]. 2023 Apr 6:2023.04.04.535623. doi: 10.1101/2023.04.04.535623. PMID: 37066421; PMCID: PMC10104020.

22. Rozowsky J, Gao J, Borsari B, Yang YT, Galeev T, Gürsoy G, Epstein CB, Xiong K, Xu J, Li T, Liu J, Yu K, Berthel A, Chen Z, Navarro F, Sun MS, Wright J, Chang J, Cameron CJF, Shoresh N, Gaskell E, Drenkow J, Adrian J, Aganezov S, Aguet F, Balderrama-Gutierrez G, Banskota S, Corona GB, Chee S, Chhetri SB, Cortez Martins GC, Danyko C, Davis CA, Farid D, Farrell NP, Gabdank I, Gofin Y, Gorkin DU, Gu M, Hecht V, Hitz BC, Issner R, Jiang Y, Kirsche M, Kong X, Lam BR, Li S, Li B, Li X, Lin KZ, Luo R, Mackiewicz M, Meng R, Moore JE, Mudge J, Nelson N, Nusbaum C, Popov I, Pratt HE, Qiu Y, Ramakrishnan S, Raymond J, Salichos L, Scavelli A, Schreiber JM, Sedlazeck FJ, See LH, Sherman RM, Shi X, Shi M, Sloan CA, Strattan JS, Tan Z, Tanaka FY, Vlasova A, Wang J, Werner J, Williams B, Xu M, Yan C, Yu L, Zaleski C, Zhang J, Ardlie K, Cherry JM, Mendenhall EM, Noble WS, Weng Z, Levine ME, Dobin A, Wold B, Mortazavi A, Ren B, Gillis J, Myers RM, Snyder MP, Choudhary J, Milosavljevic A, Schatz MC, Bernstein BE, Guigó R, Gingeras TR, Gerstein M. The EN-TEx resource of multi-tissue personal epigenomes & variant-impact models. Cell. 2023 Mar 30;186(7):1493-1511.e40. doi: 10.1016/j.cell.2023.02.018. PMID: 37001506; PMCID: PMC10074325.

24. 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 Jan;27(1):66-75. doi: 10.1038/nbt.1518. Epub 2009 Jan 4. PMID: 19122651; PMCID: PMC2924752.

25. 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. The GENCODE pseudogene resource. Genome Biol. 2012 Sep 26;13(9):R51. doi: 10.1186/gb-2012-13-9-r51. PMID: 22951037; PMCID: PMC3491395.

26. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, Barnes I, Berry A, Bignell A, Carbonell Sala S, Chrast J, Cunningham F, Di Domenico T, Donaldson S, Fiddes IT, García Girón C, Gonzalez JM, Grego T, Hardy M, Hourlier T, Hunt T, Izuogu OG, Lagarde J, Martin FJ, Martínez L, Mohanan S, Muir P, Navarro FCP, Parker A, Pei B, Pozo F, Ruffier M, Schmitt BM, Stapleton E, Suner MM, Sycheva I, Uszczynska-Ratajczak B, Xu J, Yates A, Zerbino D, Zhang Y, Aken B, Choudhary JS, Gerstein M, Guigó R, Hubbard TJP, Kellis M, Paten B, Reymond A, Tress ML, Flicek P. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019 Jan 8;47(D1):D766-D773. doi: 10.1093/nar/gky955. PMID: 30357393; PMCID: PMC6323946.

27. 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 MB. Comparative analysis of pseudogenes across three phyla. Proc Natl Acad Sci U S A. 2014 Sep 16;111(37):13361-6. doi: 10.1073/pnas.1407293111. Epub 2014 Aug 25. PMID: 25157146; PMCID: PMC4169933.

28. Balasubramanian S, Zheng D, Liu YJ, Fang G, Frankish A, Carriero N, Robilotto R, Cayting P, Gerstein M. Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes. Genome Biol. 2009;10(1):R2. doi: 10.1186/gb-2009-10-1-r2. Epub 2009 Jan 5. PMID: 19123937; PMCID: PMC2687790.

29. Zhang Z, Carriero N, Zheng D, Karro J, Harrison PM, Gerstein M. PseudoPipe: an automated pseudogene identification pipeline. Bioinformatics. 2006 Jun 15;22(12):1437-9. doi: 10.1093/bioinformatics/btl116. Epub 2006 Mar 30. PMID: 16574694.30. Tutar Y. Pseudogenes. Comp Funct Genomics. 2012;2012:424526. doi: 10.1155/2012/424526. Epub 2012 May 7. PMID: 22611337; PMCID: PMC3352212.

31. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature. 2020 Feb;578(7793):82-93. doi: 10.1038/s41586-020-1969-6. Epub 2020 Feb 5. PMID: 32025007; PMCID: PMC7025898.

32. 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, Gümüş ZH, Crawford GE; PsychENCODE Consortium; 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 Dec 14;362(6420):eaat8464. doi: 10.1126/science.aat8464. PMID: 30545857; PMCID: PMC6413328.

33. Emani PS, Liu JJ, Clarke D, Jensen M, Warrell J, Gupta C, Meng R, Lee CY, Xu S, Dursun C, Lou S, Chen Y, Chu Z, Galeev T, Hwang A, Li Y, Ni P, Zhou X; PsychENCODE Consortium‡; Bakken TE, Bendl J, Bicks L, Chatterjee T, Cheng L, Cheng Y, Dai Y, Duan Z, Flaherty M, Fullard JF, Gancz M, Garrido-Martín D, Gaynor-Gillett S, Grundman J, Hawken N, Henry E, Hoffman GE, Huang A, Jiang Y, Jin T, Jorstad NL, Kawaguchi R, Khullar S, Liu J, Liu J, Liu S, Ma S, Margolis M, Mazariegos S, Moore J, Moran JR, Nguyen E, Phalke N, Pjanic M, Pratt H, Quintero D, Rajagopalan AS, Riesenmy TR, Shedd N, Shi M, Spector M, Terwilliger R, Travaglini KJ, Wamsley B, Wang G, Xia Y, Xiao S, Yang AC, Zheng S, Gandal MJ, Lee D, Lein ES, Roussos P, Sestan N, Weng Z, White KP, Won H, Girgenti MJ, Zhang J, Wang D, Geschwind D, Gerstein M; PsychENCODE Consortium. Single-cell genomics and regulatory networks for 388 human brains. Science. 2024 May 24;384(6698):eadi5199. doi: 10.1126/science.adi5199. Epub 2024 May 24. PMID: 38781369; PMCID: PMC11365579.

56. 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. AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol Syst Biol. 2011 Aug 2;7:522. doi: 10.1038/msb.2011.54. PMID: 21811232; PMCID: PMC3208341.

57. Chen J, Rozowsky J, Galeev TR, Harmanci A, Kitchen R, Bedford J, Abyzov A, Kong Y, Regan L, Gerstein M. A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. Nat Commun. 2016 Apr 18;7:11101. doi: 10.1038/ncomms11101. PMID: 27089393; PMCID: PMC4837449.

58. 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 Sep 6;489(7414):91-100. doi: 10.1038/nature11245. PMID: 22955619; PMCID: PMC4154057.

59. 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. Landscape of transcription in human cells. Nature. 2012 Sep 6;489(7414):101-8. doi: 10.1038/nature11233. PMID: 22955620; PMCID: PMC3684276.

60. Onuchic V, Lurie E, Carrero I, Pawliczek P, Patel RY, Rozowsky J, Galeev T, Huang Z, Altshuler RC, Zhang Z, Harris RA, Coarfa C, Ashmore L, Bertol JW, Fakhouri WD, Yu F, Kellis M, Gerstein M, Milosavljevic A. Allele-specific epigenome maps reveal sequence-dependent stochastic switching at regulatory loci. Science. 2018 Sep 28;361(6409):eaar3146. doi: 10.1126/science.aar3146. Epub 2018 Aug 23. PMID: 30139913; PMCID: PMC6198826.

61. 1000 Genomes Project Consortium; Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012 Nov 1;491(7422):56-65. doi: 10.1038/nature11632. PMID: 23128226; PMCID: PMC3498066.

62. Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, Pritchard JK. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics. 2009 Dec 15;25(24):3207-12. doi: 10.1093/bioinformatics/btp579. Epub 2009 Oct 6. PMID: 19808877; PMCID: PMC2788925.

63. van de Geijn B, McVicker G, Gilad Y, Pritchard JK. WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nat Methods. 2015 Nov;12(11):1061-3. doi: 10.1038/nmeth.3582. Epub 2015 Sep 14. PMID: 26366987; PMCID: PMC4626402.