UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

The Single Cell Transposome Hypersensitive Sites Sequencing (scTHS-seq) assay for Chromatin Accessibility and Assessment of Epigenetic States in the Human Adult Brain

Permalink

https://escholarship.org/uc/item/8kt9c2ks

Author

Sos, Brandon

Publication Date

2017

Supplemental Material

https://escholarship.org/uc/item/8kt9c2ks#supplemental

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, SAN DIEGO

The Single Cell Transposome Hypersensitive Sites Sequencing (scTHS-seq) assay for Chromatin Accessibility and Assessment of Epigenetic States in the Human Adult Brain

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Biomedical Sciences

by

Brandon Chin Sos

Committee in charge:

Professor Kun Zhang, Chair
Professor John Chang
Professor Olivier Harismendy
Professor Bing Ren
Professor Sheng Zhong

2017

Copyright

Brandon Chin Sos, 2017

All rights reserved.

The Dissertation of Brandon Chin Sos is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2017

TABLE OF CONTENTS

Signature Page iii

Table of Contents iv

List of Supplemental Files vi

List of Abbrevations vii

List of Figures viii

Acknowledgements xi

Vita xiii

Abstract of the Dissertation xv

Chapter 1: Introduction 1

Chapter 2: Characterization of chromation accessibility with a transposome hypersensitive sites sequnceing (THS-seq) assay 7

2.1 Introduction 7

2.2 Results 7

2.2.1 Experimental design and implementation 7

2.2.2 THS-seq validation 16

2.2.3 Comparison of THS-seq and ATAC-seq with two transposases: Tn5059 and EzTn5 22

2.2.4 Comparison of THS-seq/Tn5059 and ATAC-seq/EzTn5 36

2.2.5 Characterization of capture sensitivity and ATAC-seq bias 43

2.2.6 ATAC-seq improvements 49

2.3 Discussion 51

Chapter 3: Single cell THS-seq development and integrative single cell analysis by transcriptional and epigenetic states in human adult brain 53

3.1 Introduction 53

3.2 Results 53

3.2.1 snDrop-seq design and application to human brain nuclei 53

3.2.2 scTHS-seq design and application to human brain nuclei 65

3.2.3 scTHS-seq and snDrop-seq data integration 73

3.2.4 Integrative transcription factor analysis 78

3.2.5 Mapping disease risk variants to susceptible cell types 81

3.3 Discussion 85

Chapter 4: Conclusions 87

Appendix A: THS-seq materials and methods 91

Appendix B: scTHS-seq materials and methods 100

References 122

List of Supplemental Files

thesis_V5-2_all-tables.xlsx

LIST OF ABBREVATIONS

ATAC	Assay for transposase-accessible chromatin
DMF	Dimethylformamide
ENCODE	Encyclopedia of DNA Elements
FACS	Fluorescence activated cell sorting
GBM	Gradient boosting model
GWAS	Genome wide association study
PCR	Polymerase chain reaction
ROC	Receiver operating characteristic
S.E.M	Standard error of the mean
scATAC	Single cell assay for transposase-accessible chromatin
scTHS-seq	Single cell transposome hypersensitive sites sequencing assay
snDrop-seq	Single nucleus droplet sequencing
t-SNE	t-distributed stochastic neighbor embedding
TF	Transcription factor
THS-seq	Transposome hypersensitive sites sequencing

LIST OF FIGURES

Figure 2.1. Schematic overview of THS-seq..... 9

Figure 2.2. Schematic of transposon design and transposome complex generation..... 10

Figure 2.3. Validation of successful transposon insertion and DNA fragmentation. 12

Figure 2.4. Validation of T7 *in vitro* transcription..... 13

Figure 2.5. Validation of PCR amplification for barcode addition..... 14

Figure 2.6. Tn5059 concentration titrations and titration of dimethylformamide concentrations for optimal *in vitro* transcription amplification yields. 15

Figure 2.7. Validation of 100-cell THS-seq/Tn5059 data against ENCODE data and ATAC-seq data. 18

Figure 2.8. Comparison of peak region distances to transcription start sites..... 19

Figure 2.9. Significant gene ontology biological processes categories for each peaks dataset. 20

Figure 2.10. Quantitation of mitochondrial reads in all datasets 21

Figure 2.11. 100-cell THS-seq with Tn5059 read statistics analysis to determine minimal number of raw reads for quality data..... 23

Figure 2.12. Comparison between THS-seq/Tn5059, THS-seq/EzTn5, ATAC-seq/Tn5059, and ATAC-seq/EzTn5 with 500 cells of input material. 24

Figure 2.13. Peak overlap comparisons in all datasets. 25

Figure 2.14. Base pair overlap comparisons in all datasets. 26

Figure 2.15. Examination of peak overlap in a pairwise comparison of all experimental datasets. 30

Figure 2.16. Examination of base pair overlap in a pairwise comparison of all experimental datasets. 31

Figure 2.17. 200 kb view of accessible chromatin marks in 500 cell datasets. ... 32

Figure 2.18. Comparison between the two most comprehensive datasets of 500-cell THS-seq/Tn5059 and 500-cell ATAC-seq/EzTn5. ... 33

Figure 2.19. Overview of chromatin accessibility at gene loci enriched in 100- and 500-cell THS-seq with Tn5059 data and not enriched in 500-cell ATAC-seq with EzTn5 data. ... 34

Figure 2.20. Overview of chromatin accessibility at major genes implicated in cancer and in immune system function. ... 35

Figure 2.21. Peak size distribution comparisons between THS-seq, ATAC-seq and ENCODE data. ... 38

Figure 2.22. Validation of peaks based on peak length. ... 39

Figure 2.23. Validation of peaks based on peak length. ... 42

Figure 2.24. THS-seq and ATAC-seq peak capture preferences and biases. ... 46

Figure 2.25. THS-seq and ATAC-seq peak capture preferences. ... 47

Figure 2.26. Comparison of 500-cell ATAC-seq with EzTn5 data to published ATAC-seq data. ... 48

Figure 2.27. 200 kb view of accessible chromatin in ATAC-seq with EzTn5 data versus published ATAC-seq data. ... 50

Figure 3.1. snDrop-seq metrics. ... 56

Figure 3.2. snDrop-seq metrics. Transcriptional profiling with snDrop-seq resolves major cell types of the visual and frontal cortices. ... 57

Figure 3.3. Comparison of single-cell/nucleus RNA-seq methodologies. ... 58

Figure 3.4. snDrop-seq resolves neuronal and non-neuronal subpopulations. ... 59

Figure 3.5. Subtype-specific gene enrichments and tissue-specific proportions. 62

Figure 3.6. snDrop-seq identifies molecularly and spatially distinct neuronal subtypes. ... 63

Figure 3.7. Cell-type-specific expression of NAT markers and stage-specific expression of NAT and protein-coding markers. 64

Figure 3.8. Chromatin accessibility profiling resolves major cell types of the visual cortex. 67

Figure 3.9. Customized scTHS-seq transposon design and sequencing library fragments generated. 68

Figure 3.10. Confirmation of *in vitro* transcription and barcoded library generation in scTHS-seq. 69

Figure 3.11. Sequencing statistics and collisions calculations for scTHS-seq. 70

Figure 3.12. scTHS-seq Unique read statistics 71

Figure 3.13. Visualization of merged single cell data and individual single cell datasets. 72

Figure 3.14. Epigenetic heterogeneity in the frontal cortex, as measured by scTHS-seq. 72

Figure 3.15. Differential gene expression predicts differential chromatin accessibility to map corresponding transcriptional and epigenetic subtypes. 75

Figure 3.16. Probabilistic inference of differential expression and differential accessibility signals. 76

Figure 3.17. Cell type prediction based on probabilistic inference of differential expression and differential accessibility signals. 77

Figure 3.18. Mapping of transcription factor binding status and disease risk variants to specific brain cell types. 79

Figure 3.19. Gene set enrichment analysis plots to assess significance of identified TFs with expression of TFs. 80

Figure 3.20. Mapping of common disease risk variants to specific brain cell types. 83

Figure 3.21. Mapping of disease risk variants to specific brain cell types in brain diseases and non-brain diseases. 84

ACKNOWLEDGEMENTS

I would like to acknowledge Dr. Kun Zhang for his mentorship and guidance. Kun has supported me and my development as a scientist by providing a supportive lab environment and access to the tools to bring me to where I am today. For this I am very grateful.

I would also like to acknowledge the past and present members of the Zhang laboratory: Elizabeth Duong for helping to carry out and optimize the scTHS-seq assay; Shicheng Guo, Yan Wu, Dinh Diep and Chris Wei for bioinformatics advice and help with learning python and R scripting languages for data analysis; Derek Gao for keeping the lab organized and running the MiSeq as well as submitting samples for HiSeq sequencing; Molly He, Trina Osothprarop, and Amirali Kia and Illumina for collaborating and providing the Tn5 supermutant enzyme essential in my studies; Neeraj Salathia and Illumina for sequencing samples; Peter Kharchenko and Jean Fan for collaboration and analysis of THS-seq data; Andrew Richards, Athurva Gore, Jeff Gole, Song Chen, Blue Lake, and any other lab members for helpful discussions and scientific expertise; Alice Li and Nongluk Plongthongkum for welcoming me into the lab space and introducing me to the standard assays and procedures in the laboratory.

I would like to acknowledge my thesis committee: Dr. Olivier Harismendy, Dr, Bing Ren, Dr. John Chang and Dr. Sheng Zhong. Your feedback and support have been very valuable to me.

I would like to acknowledge my partner Michelle Fu and family for being supportive throughout my doctoral education.

Chapter 2, in full is an adapted version of the material that appears in *Genome Biology*. Brandon Chin Sos, Ho-Lim Fung, Derek Rui Gao, Trina Faye Osothprarop, Amirali Kia, Molly Min He, and Kun Zhang. 2016. Characterization of chromatin accessibility with a transposome hypersensitive sites sequencing (THS-seq) assay. *Genome Biology*. Published. The dissertation author was the primary author of all material.

Chapters 3, in full is an adapted version of the material that has been submitted for publication. Blue B. Lake†, Song Chen†, Brandon C. Sos†, Jean Fan†, Yun Yung, Gwendolyn E. Kaeser, Thu E. Duong, Derek Gao, Jerold Chun*, Peter Kharchenko*, Kun Zhang*. 2017. Integrative single-cell analysis by transcriptional and epigenetic states in human adult brain. In revision. The dissertation author was the co-primary author of all material.

VITA

2007 B.S., Microbiology, Immunology and Molecular Genetics; Minor, Neuroscience; University of California, Los Angeles

2007-2010 Staff Research Associate, University of California, San Francisco

2012-2013 Biomedical Sciences Student Council Co-President, University of California, San Diego

2013 Graduate Teaching Assistant, University of California, San Diego

2017 Ph.D., Biomedical Sciences, University of California, San Diego

PUBLICATIONS

Blue B. Lake†, Song Chen†, **Brandon C. Sos**†, Jean Fan†, Yun Yung, Gwendolyn E. Kaeser, Thu E. Duong, Derek Gao, Jerold Chun*, Peter Kharchenko*, Kun Zhang*. 2017. Integrative single-cell analysis by transcriptional and epigenetic states in human adult brain. In revision.

Sos BC, Fung HL, Gao DR, Osothprarop TF, Kia A, He MM, Zhang K. Characterization of chromatin accessibility with a transposome hypersensitive sites sequencing (THS-seq) assay. *Genome Biol*. 2016 Feb 4;17:20.

Nordstrom SM, Tran JL, **Sos BC**, Wagner KU, Weiss EJ. Disruption of JAK2 in adipocytes impairs lipolysis and improves fatty liver in mice with elevated GH. *Mol Endocrinol*. 2013 Aug;27(8):1333-42.

Sos BC, Harris C, Nordstrom SM, Tran JL, Balázs M, Caplazi P, Febbraio M, Applegate MA, Wagner KU, Weiss EJ. Abrogation of growth hormone secretion rescues fatty liver in mice with hepatocyte-specific deletion of JAK2. *J Clin Invest.* 2011 Apr;121(4):1412-23.

Nordstrom SM, Tran JL, **Sos BC**, Wagner KU, Weiss EJ. Liver-derived IGF-I contributes to GH-dependent increases in lean mass and bone mineral density in mice with comparable levels of circulating GH. *Mol Endocrinol.* 2011 Jul;25(7):1223-30.

Nordstrom SM, Holliday BA, **Sos BC**, Smyth JW, Levy RE, Dukes JW, Lord ST, Weiss EJ. Increased thrombosis susceptibility and altered fibrin formation in STAT5-deficient mice. *Blood*. 2010 Dec 16;116(25):5724-33.

Farber CR, van Nas A, Ghazalpour A, Aten JE, Doss S, **Sos B**, Schadt EE, Ingram-Drake L, Davis RC, Horvath S, Smith DJ, Drake TA, Lusis AJ. An integrative genetics approach to identify candidate genes regulating BMD: combining linkage, gene expression, and association. *J Bone Miner Res.* 2009 Jan;24(1):105-16.

Wong JH, Dukes J, Levy RE, **Sos B**, Mason SE, Fong TS, Weiss EJ. Sex differences in thrombosis in mice are mediated by sex-specific growth hormone secretion patterns. *J Clin Invest.* 2008 Aug;118(8):2969-78.

ABSTRACT OF THE DISSERTATION

The Single Cell Transposome Hypersensitive Sites Sequencing (scTHS-seq) assay for Chromatin Accessibility and Assessment of Epigenetic States in the Human Adult Brain

by

Brandon Chin Sos

Doctor of Philosophy in Biomedical Sciences

University of California, San Diego, 2017

Professor Kun Zhang, Chair

Chromatin accessibility captures in vivo protein-chromosome binding status, and is considered an informative proxy for protein-DNA interactions. DNase I and Tn5 transposase assays require thousands to millions of fresh cells for comprehensive chromatin mapping. Applying Tn5 tagmentation to hundreds of cells and to single cells results in sparse chromatin maps at high throughput or

dense chromatin maps at low throughput. We present a high throughput transposome hypersensitive sites sequencing assay for highly sensitive characterization of chromatin accessibility. Linear amplification of accessible DNA ends with in vitro transcription, coupled with an engineered Tn5 super-mutant, demonstrates improved sensitivity on limited input materials, and accessibility of small regions near distal enhancers, compared with ATAC-seq. Application to single cells was achieved by adapting THS-seq to combinatorial indexing to generate high sensitivity DNA accessibility maps for tens of thousands of single cells from the human adult visual and frontal cortex. Integrative analysis of snDrop-seq and scTHS-seq has allowed us to identify transcription factors and regulatory elements shaping the state of different brain cell types, and to map genetic risk factors of common human brain diseases to specific pathogenic cell types and subtypes. Overall we demonstrate the viability of scTHS-seq to quantitate DNA accessibility in single cells, and application to biological samples to shed insight on human brain diseases.

CHAPTER 1: INTRODUCTION

Chromatin accessibility assays and THS-seq

The accessibility of chromatin is a major determinant of gene regulation. The chromatin landscape defines the transcriptional regulatory networks that determine cellular identity and function as well as biological processes involved in differentiation, proliferation, development and responses to the extracellular environment. Regulatory regions can be open or poised for regulation, when DNA within the region is accessible for factors to bind to DNA binding sites. Conversely, regulatory regions can be closed with DNA binding sites inaccessible, and thus not open for regulation. Genome wide chromatin accessibility assays were first developed utilizing cloning[1], then microarrays[2], and now existing assays such as, DNase-seq[3], FAIRE-seq[4] and ATAC-seq[5] methods have been shown to be remarkably powerful in defining the binding status of transcription factors, determining nucleosome occupancy and constructing gene regulatory networks[6-14]. These assays measured DNA accessibility in bulk samples with loss of cellular resolution in treating tens of thousands to millions of cells as a single sample. In this case measurements reflect a sum of cellular heterogeneity with loss of ability to delineate chromatin accessibility in rare cell populations, or within multiple different cell populations in the same sample. However, assays have been developed to measure DNA accessibility at the single cell level, using the C1 microfluiditics chip, single cell DNase I accessibility, and combinatorial indexing[15-17]. Each method has its deficiencies, the C1 microfluidic chip and single cell DNase I assays are not high throughput,

and combinatorial indexing generates rather sparse data, such that the number of input cells and assay runs to obtain enough single cell data for analysis outweighs the monetary and time costs per experiment. Considering the proprietary nature of the C1 instrument and chip design, and no clear path to improve the throughput of the DNase I assay, we focus on further improvement on the sensitivity of the tagmentation-based approach as well as the throughput. This will facilitate routine profiling of accessible chromatin on a wide variety of samples.

The widely used DNase-Seq and ATAC-seq methods have several areas that can be improved upon to increase assay sensitivity. For DNase-seq millions of cells are required for nuclei isolation, DNase I titration, downstream enzymatic reactions and associated purification steps for DNA end polishing and adaptor ligations[3]. These inefficiencies were addressed by ATAC-seq and its usage of the Tn5 transposome system, which was originally developed for generating low input sequencing libraries[18, 19]. With this method, chromatin accessibility assays were demonstrated initially at the 50,000-cell level, and more recently with single cells with adaptation to the combinatorial indexing method[15]. However, with the ATAC-seq method, we identified three aspects that can be potentially improved to increase the assay sensitivity. Firstly, by design, the method uses PCR amplification immediately after Tn5 insertion, and only a fraction of the “tagmented” molecules can be amplified and recovered, as not all inserted adaptor pairs are in the correct orientation or have the appropriate spacing to generate molecules of a size amenable to PCR amplification. Second, buffer

conditions are critical for Tn5 activity, and critical components such as dimethylformamide (DMF) can be titrated for optimal Tn5 activity for assessing chromatin accessibility, as well as determining the optimal concoction of buffer components[20]. Thirdly, the commercial EzTn5 transposase is a mutated version of the wild-type Tn5 enzyme that has high activity for random transposition[21, 22]. There is potential room to further engineer the enzyme to achieve a more efficient and specific insertion into open chromatin.

Combinatorial indexing and application to human brain cells

Combinatorial indexing with Tn5 Transposition is a multiplexing method initially developed to process thousands of samples in a single experiment with 96 well plates. Indexing is performed at both the transposition and PCR stages, which generates 9,216 possible barcode combinations. Initially developed for resolving haplotypes in whole genome samples[23], the method has been applied to different single cell methods including DNA accessibility, Hi-C[24-26] and most recently RNA-seq. With multiple rounds of indexing and adaptor addition, it has the potential to assay tens of millions of cells, with the rate limiting factors being the cost of sequencing and computational power for data analysis. No machines are needed and the method can be carried out using PCR plates and pipetting tools. Importantly, although Tn5 transposition cuts DNA, it does not fragment DNA until Tn5 is removed since it remains tightly bound to DNA after transposition. All tags remain within nuclei, which can then be re-compartmentalized for secondary barcoding steps. THS-seq adaptation to

combinatorial indexing will allow high throughput data generation, and when combined with the higher sensitivity of THS-seq, will provide a powerful single cell chromatin accessibility method.

To our knowledge, deconvolution and identification of cell types has not been performed on single cell chromatin accessibility data obtained from a single heterogeneous sample. There has been a study performed on bulk samples that have been pre-sorted, computationally merged, and sorted into respective cell types again[27]. This provides proof of concept for deconvolution of cell types and shows there is enough heterogeneity in DNA accessibility data to do this analysis. However, this was on bulk samples and how the analysis translates to single cell chromatin accessibility data is unknown. Once deconvolution and identification of cell types is accomplished, then we can perform an integrative analysis of scTHS-seq and snDrop-seq data on human brain samples.

There have been several integrative analyses of RNA-seq data and DNA accessibility data finding novel biological insights[28-30]. However, these analyses have been performed on bulk datasets with tens of thousands cells, and a single cell integrative analysis has not been performed to our knowledge. Single cell RNA-seq technologies have made rapid advances and are able to generate tens of thousands of high quality single cell datasets, with methods including inDrop, Drop-seq and Split-seq[31-33]. Single cell chromatin accessibility assays have either had high sensitivity and low throughput, or high throughput and low sensitivity, thus not generating quality datasets to perform an integrative analysis with single cell RNA-seq data. With high quality scTHS-seq data containing thousands of

single cells, we are expecting to be able to perform an integrative analysis, generating a more complete regulatory map containing both chromatin accessibility data and transcriptional data.

There are large repositories of GWAS data available online that have curated disease risk loci for complex diseases[34, 35]. This data has proved very useful in identifying risk alleles and explaining their contribution to disease[36], however there is no direct way to predict susceptible cell types involved in disease. If an identified cell type has accessible chromatin, is expressed, or is epigenetically activated at disease risk loci, then it is more likely to be susceptible to disease, opposed to a cell under epigenetic repression at the same loci. We hypothesize that integration of scTHS-seq data and GWAS data will be able to reveal the most susceptible cell types in disease pathologies.

In this dissertation we present a systematic effort to assay chromatin accessibility in single cells at high sensitivity. Initially we greatly increase the sensitivity of transposase-based DNA accessibility assays through optimization of method, buffer and enzyme. We show that we achieve high sensitivity and capture of more distal regulatory elements. To assay chromatin accessibility in single cells in a high throughput manner, we then adapt THS-seq to combinatorial indexing and apply the method to human brain visual and frontal cortex samples. We then develop an analysis method to identify and classify cell subtypes in single cell chromatin accessibility data from a heterogonous sample of unidentified cell types. Integrative data analysis with snDrop-seq data further identifies cellular subtypes and demonstrates a new strategy to construct atlases

of single cell data from human organs. Lastly, association with disease risk factors identifies the specific cell types susceptible to common brain diseases thus demonstrating how single cell chromatin accessibility data can be used for novel biological discoveries.

CHAPTER 2: CHARACTERIZATION OF CHROMATIN ACCESSIBILITY WITH A TRANSPOSOME HYPERSENSITIVE SITES SEQUENCING (THS-SEQ) ASSAY

2.1 Introduction

DNase-seq and ATAC-seq have several inefficiencies that can be addressed to improve assay sensitivity. In order to address this, we develop a new transposome based method requiring new processing steps, optimized buffers and an engineered Tn5 that better captures accessible chromatin. We then validate THS-seq and perform a comparison with 500 cells of input material between THS-seq/Tn5059, THS-seq/EzTn5, ATAC-seq/Tn5059, and ATAC-seq/EzTn5 using the commonly used GM12878 lymphoblastoid cell line. Lastly we employ optimizations made in THS-seq to improve sensitivity in ATAC-seq.

2.2 Results

2.2.1 Experimental design and implementation

We hypothesized that the limited sensitivity of ATAC-seq is inherent in the method design, and is mainly due to three factors. First the Tn5 transposome complex inserts adaptors in random orientation, such that only half of the molecules contain the adaptors in the orientation required for PCR amplification. Second, only ~1% of the genome is accessible in typical cells, and the regions in which two adjacent transposition events are too far apart cannot be amplified by PCR. In fact, for this reason, the existing DNase-Seq method includes a fragmentation step to capture and sequence only the flanking sequences

immediately adjacent to DNase I digestion sites, which effectively capture single-digestion events. Third, accessible regions small in length would have too few transposition events, and in conjunction with losing half of the molecules due to incorrect adaptor orientation, would not produce enough molecules to form a detectible peak above background levels of transposition events. Therefore, applying such a fragmentation strategy to small numbers of cells or single cells would result in low sensitivity, especially in the small accessible regions. We therefore developed the THS-seq method, which uses a customized Tn5 transposome system to attach a T7 promoter[37-40] to the end of every DNA molecule after *in vitro* transposition. The end sequences of the insertion sites were then linearly amplified with *in vitro* transcription by roughly 1000-fold, regardless of the distance between two adjacent Tn5 insertion sites, and the resulting RNA molecules were then converted into sequencing libraries efficiently through seven enzymatic reactions (single-stranded cDNA synthesis, RNase H digestion, double-stranded cDNA synthesis, transposition, protease digestion, end fill-in and PCR amplification) (**Fig 2.1**) To this end we designed a custom transposon that, in addition to the mosaic end sequences for transposase binding, includes a T7 promoter sequence for *in vitro* transcription, and an adaptor sequence compatible with constructing Illumina sequencing libraries (**Fig. 2.2**). Additionally, to address the efficiency loss from incorrect adaptor orientations that is seen with PCR based methods, the transposome complex dimer that consists of two Tn5 molecules and two transposons, where one Tn5 molecule is bound to one transposon, was designed and generated so that every

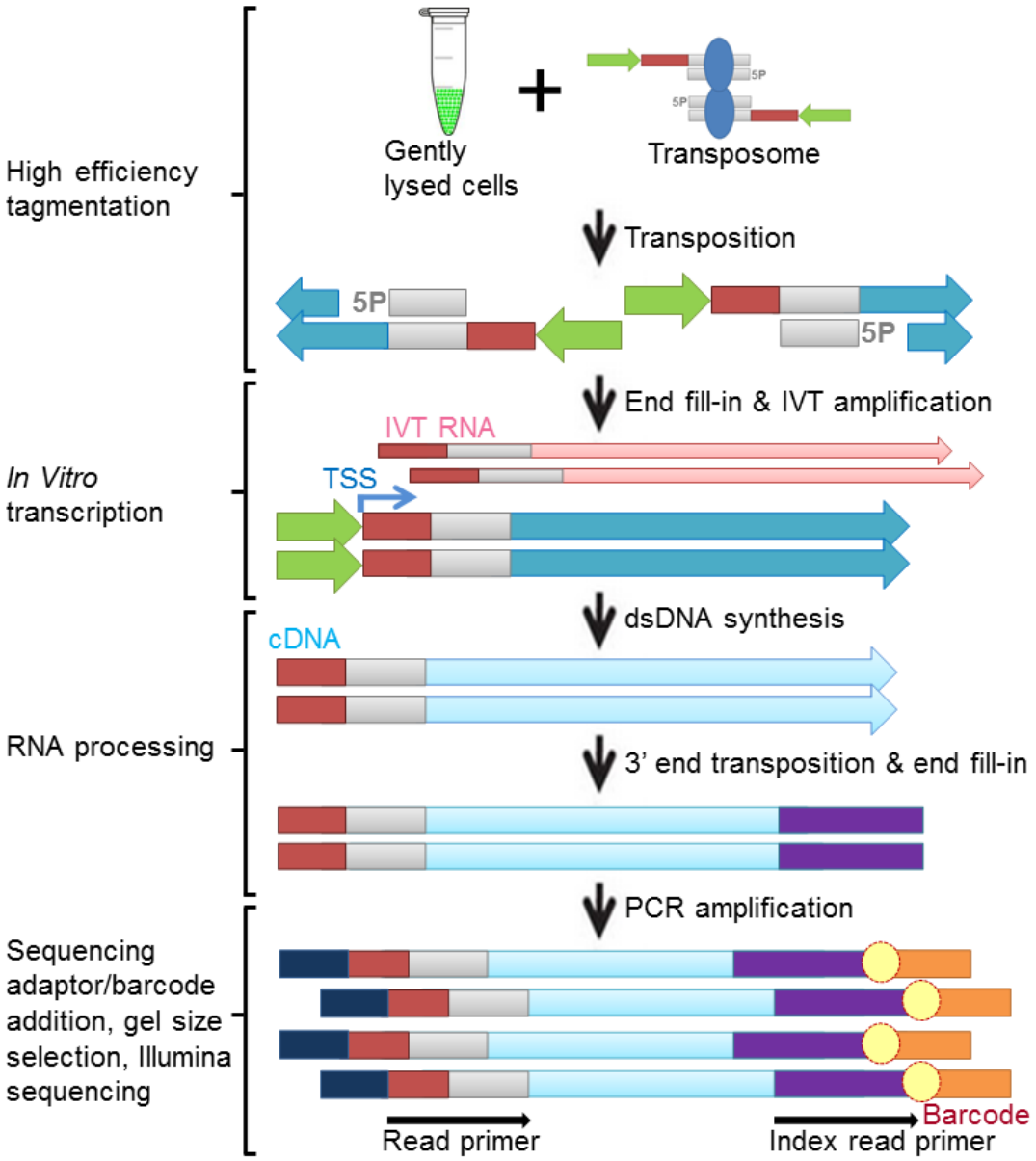


Figure 2.1. Schematic overview of THS-seq.
High efficiency tagmentation is performed on gently lysed cells, followed by in vitro transcription, and custom RNA-seq to generate barcoded sequencing libraries. The following colors depict the following segments. Light gray segments are Tn5 mosaic ends sequences, green segments are T7 promoter sequences, dark red segments are read primer sequences, dark blue segments are genomic DNA, light blue segments are cDNA sequences, purple segments are 3' adaptor sequences, orange and navy blue segments are Illumina adaptor sequences, and yellow circles are barcodes.

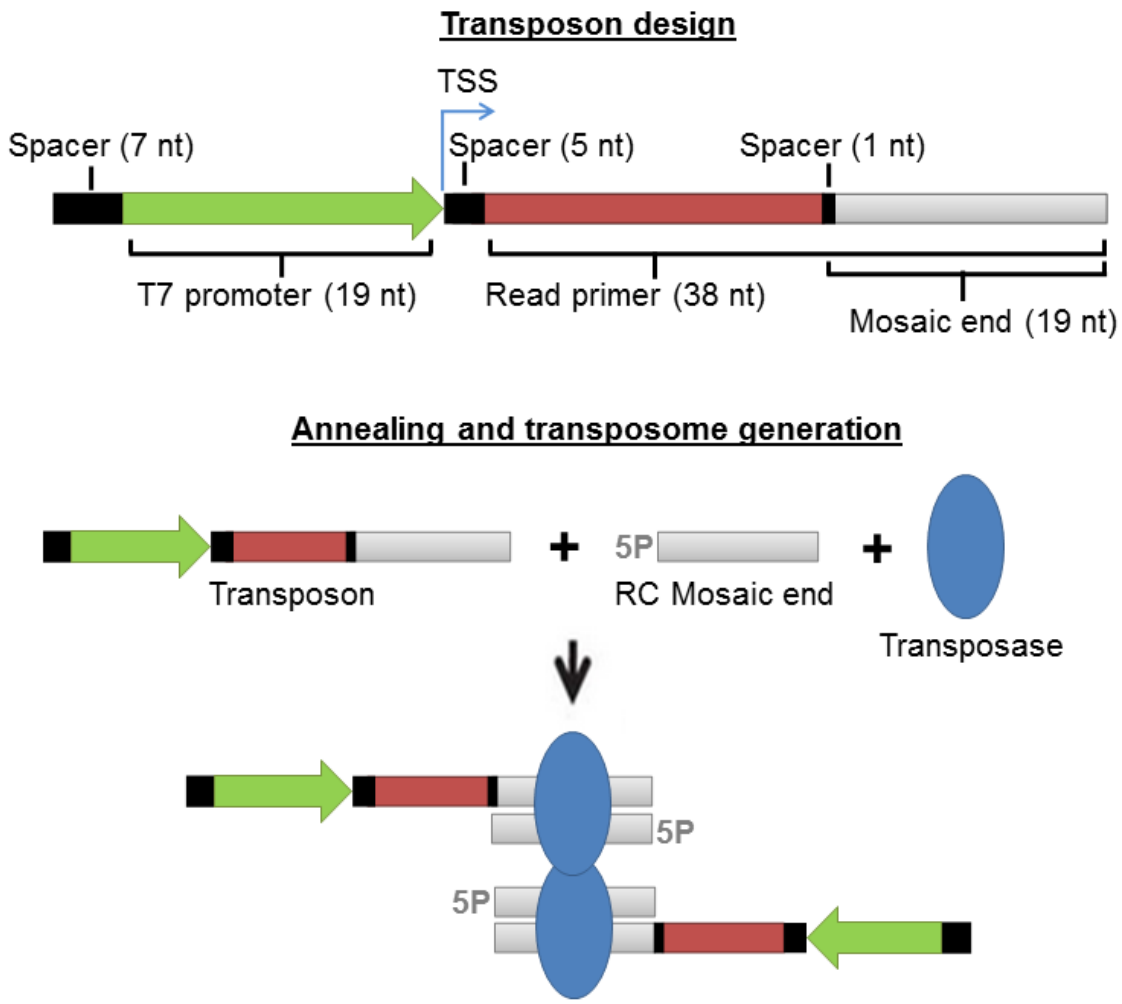


Figure 2.2. Schematic of transposon design and transposome complex generation.
Transposon consists of the T7 promoter, Read primer sequence, and the mosaic end sequence for Tn5 binding that is contained within the read sequence.

single insertion yields usable ends regardless of the orientation of the transposon after insertion. Successful insertion of our customized transposon was confirmed by electrophoresis (**Fig. 2.3**). Next, all proteins were removed by treatment with guanidine hydrochloride, followed by end fill-in and linear *in vitro* transcription amplification by T7 RNA polymerase which in theory has a lower bias (propagation of errors introduced early in PCR) compared with exponential PCR

amplification. The resulting RNA molecules (**Fig. 2.4a-d**) were converted to double stranded cDNA, followed by a second round of transposition to tag 3'-ends of double stranded cDNA. Transposase was digested by protease to release DNA fragments and then to generate sequencing libraries, PCR amplification was performed that added sequence containing Illumina adaptors and a sample barcode (**Fig. 2.5a-d**). Multiple samples were tagged with sample-specific DNA barcodes, pooled and size selected for Illumina sequencing. Finally, we replaced the standard EzTn5 transposase with a novel Tn5 super-mutant (Tn5059), which resulted from a semi-rational design and has a higher activity and less sequence-specific bias than EzTn5 on purified DNA. Concentration titrations were performed on Tn5059 for optimal activity for THS-seq with cellular input (**Fig. 2.6a-d**), as well as development of an optimized tagmentation reaction buffer that includes 16% DMF (**Fig. 2.6e-f**).

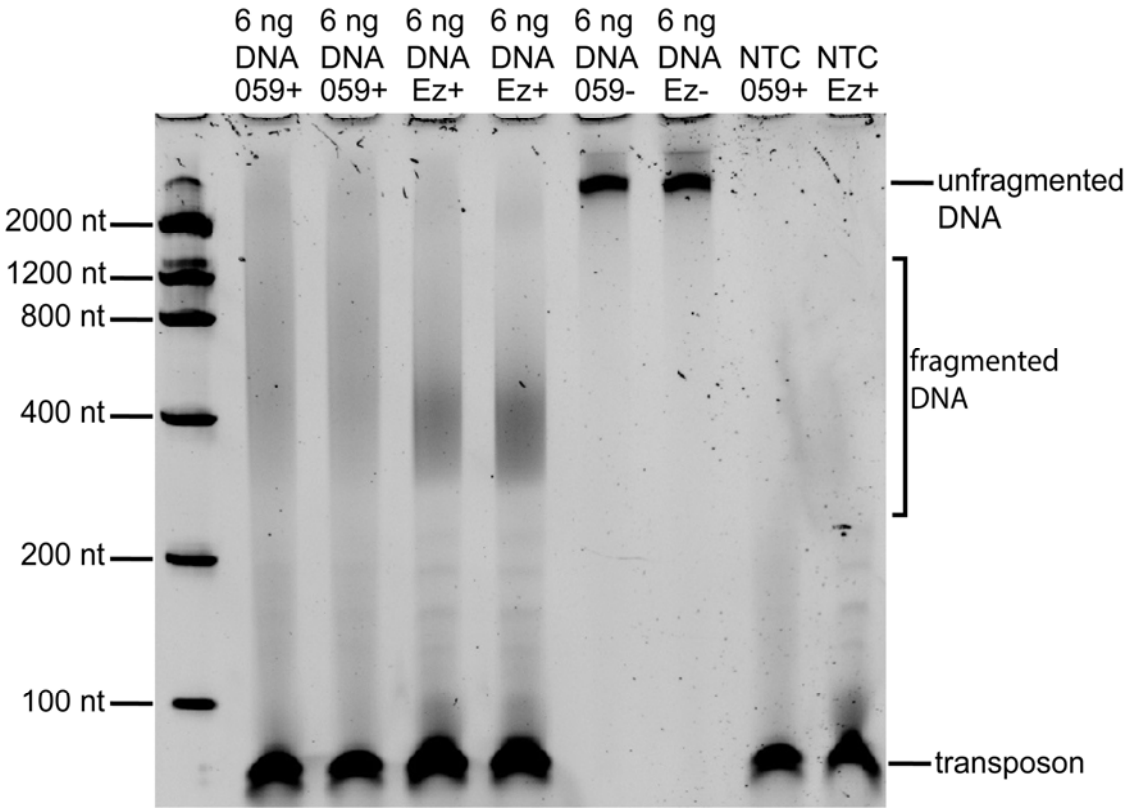


Figure 2.3. Validation of successful transposon insertion and DNA fragmentation.
The first two lanes are reactions using 0.7 μM Tn5059, the third and fourth lanes are reactions containing 0.5 μM EzTn5, the fifth and sixth lanes have Tn5 replaced with nuclease free water, the seventh lane is a reaction with 0.7 μM Tn5059 and with water substituted for DNA in the reaction, and the eighth lane is a reaction with 0.5 μM EzTn5 with water substituted for DNA in the reaction. The customized EzTn5 and Tn5059 transposomes are designed to fragment DNA upon transposon incorporation. By fragmenting DNA, this implies the mosaic end DNA oligo and transposon oligo anneal correctly, the transposome complex is forming correctly, the transposome complex is active, and the transposome complex can successfully fragment DNA and insert the transposon. Successful attachment of the transposon is indicated by generation of in vitro transcribed RNA, which is shown in Figure 2.4. It is not expected to have dark smears for fragmented DNA, since 6 nanograms of DNA were used as input and no PCR was done after fragmentation. Experimentally a transposition reaction was performed on samples, cleaned with zymo DNA clean and concentrator, and then run on a 6% Tris-borate-EDTA (TBE) gel.

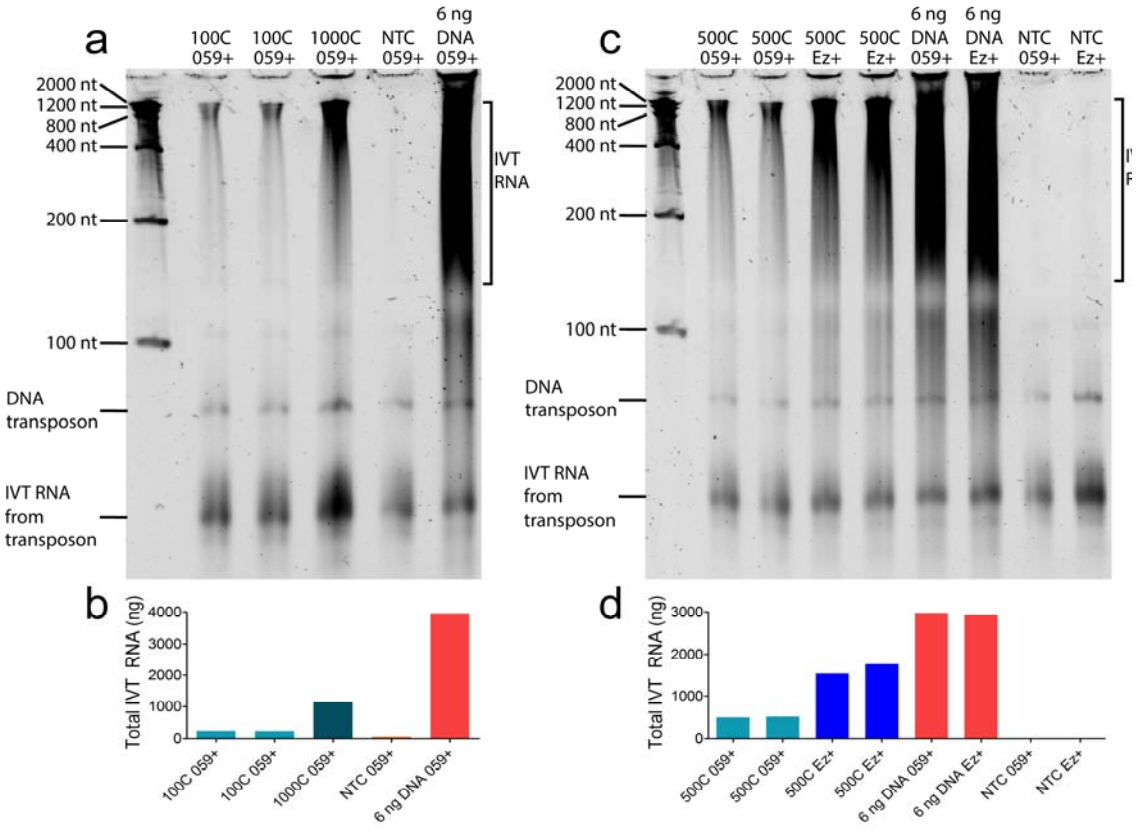


Figure 2.4. Validation of T7 *in vitro* transcription.
6% Tris-Borate-Urea (TBU) gels assaying for single stranded nucleic acids, in this case RNA, are used for validation of RNA amplification from *in vitro* transcription. Generally, smears are seen from ~150 base pairs to >2000 base pairs. (a) RNA smears from THS-seq with Tn5059 using 100 cells, 1000 cells and 6 nanograms of pure DNA. No RNA smear is generated for the NTC control. (b) Quantitation of the *in vitro* transcribed RNA smears present in (a). (c) RNA smears are also present for assaying with both 500-cell THS-seq with Tn5059 and 500-cell THS-seq with Ez-Tn5 and 6 nanograms of pure DNA with each respective enzyme. No RNA smears are generated for the NTC controls for each respective enzyme. (d) Quantitation of the *in vitro* transcribed RNA smears present in (c).

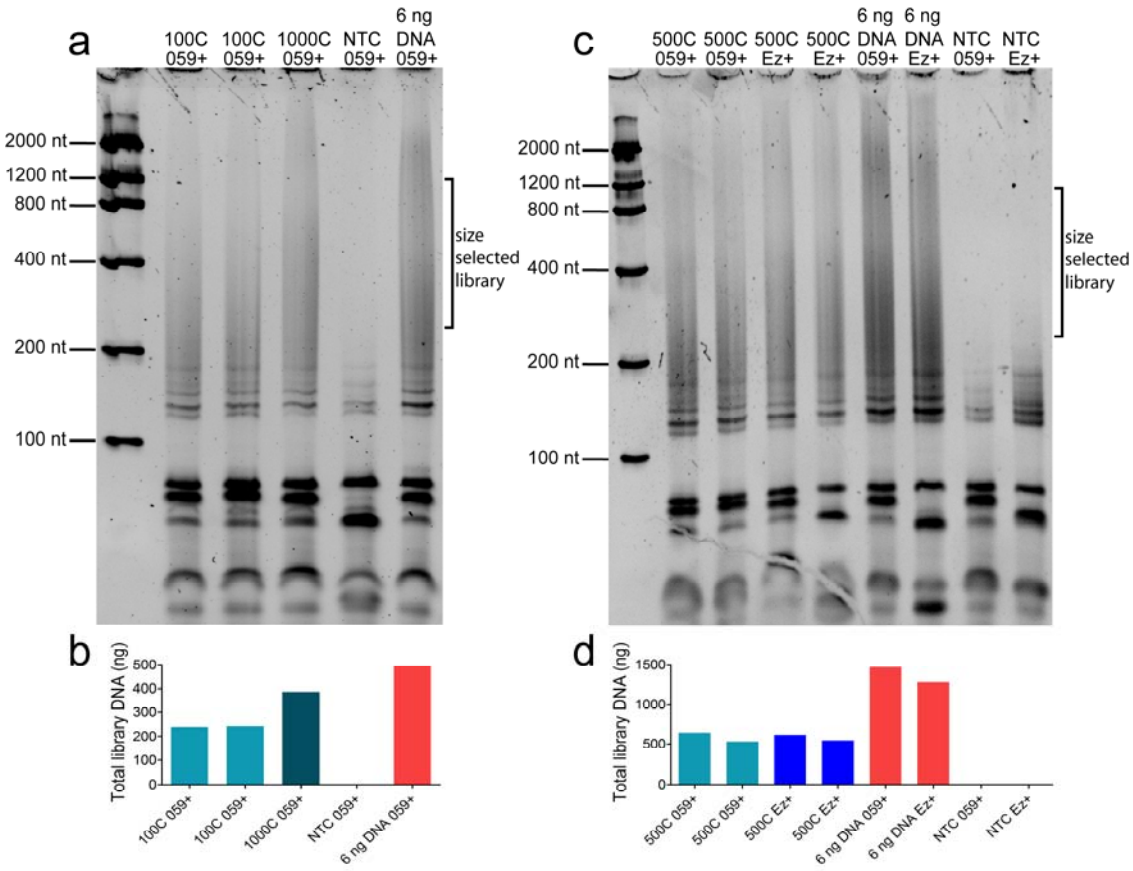


Figure 2.5. Validation of PCR amplification for barcode addition.
6% TBE gels for assaying double stranded DNA are used here for confirming generation of barcoded libraries. (a) Libraries have been successfully generated for THS-seq with Tn5059 using 100 cells, 1000 cells and with 6 nanograms of pure DNA. No DNA smears are generated for the NTC control. (b) Quantitation of size selected library smears present in (a). (c) DNA smears are also present for assaying with both 500-cell THS-seq with Tn5059 and 500-cell THS-seq with Ez-Tn5 and 6 nanograms of pure DNA with each respective enzyme. No DNA smears are generated for the NTC controls for each respective enzyme. (d) Quantitation of size selected library smears present in (c). Subsequently libraries are pooled, and gel size selected between ~220-250 to ~1000 base pairs to remove primers and primer dimers below the 220-250 base pair range.

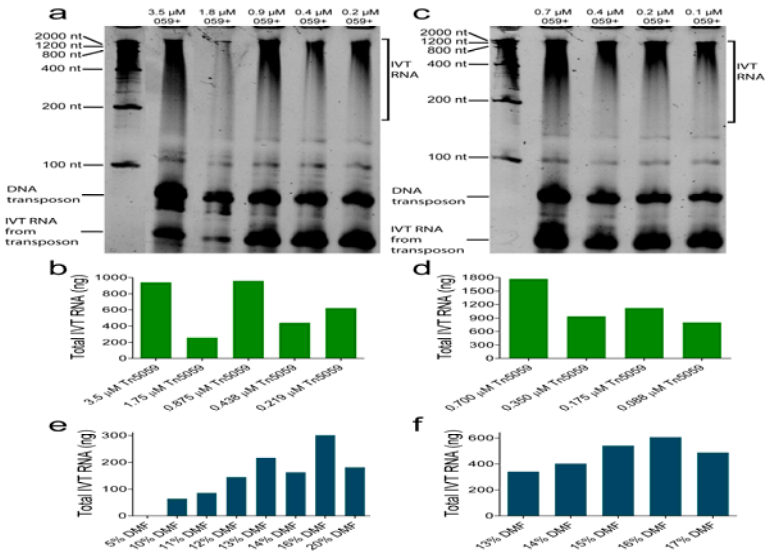


Figure 2.6. Tn5059 concentration titrations and titration of dimethylformamide concentrations for optimal *in vitro* transcription amplification yields.

(a-d) Concentration titrations were performed to determine the optimal concentration of Tn5059 for assaying accessible chromatin on 1000 GM12878 lymphoblastoid cells that gives the most *in vitro* transcription amplification product. 6% TBU gels were used for visualization. (a,b) Concentration titration of Tn5059 was performed at a range of 3.5 μM to 0.2 μM Tn5059 final concentration in reaction. Here 0.9 μM Tn5059 gave the best amplification and the reaction with 3.5 μM Tn5059 was similar, however the smear looks washed out and not how IVT smears usually appear, also with 0.9 μM Tn5059 reagents are saved. The 1.8 μM Tn5059 reaction possibly had pipetting errors. (c,d) The concentration titration of Tn5059 was performed with 0.7 μM to 0.1 μM Tn5059 final in reaction, and here 0.7 μM gave the strongest amplification. The quantitation of IVT RNA differs greatly between the gels since differences in how the ladders ran affected quantitation, however the trends are similar. We also performed a titration on DMF to determine the optimal concentration for tagmentation that gives the most *in vitro* transcription amplification product. Reactions used 1000 GM12878 lymphoblastoid cells. Titration was performed on a transposon that was 17 base pairs longer than the one used in the study, T7tspn-top2, however the result of this titration should be applicable to any transposon that is being used to assay accessible chromatin. The concentration of other tagmentation reaction buffer components that was kept consistent while DMF concentrations were varied was, 165 mM Tris-OAc, pH 7.8, 330 mM K-OAc, 50 mM Mg-OAc. All reactions were carried out in 5 μL solutions. (e) First a wider range of DMF concentrations were titrated, and once an optimal concentration was determined, (f) a narrower range was titrated to confirm the results.

2.2.2 THS-seq validation

To validate THS-Seq, we applied the method to 100 GM12878 lymphoblastoid cells, and generated 71.5 and 83.4 million Illumina sequencing reads per replicate. After sequence alignment and clonal read removal, we obtained roughly 11 million uniquely mapped reads for each library. Each THS-seq unique read by design represents a unique transposon insertion event on a single chromosomal molecule. Therefore, ~110,000 unique transposition events per cell was captured in our data set. We then performed peak calling with Dfilter[41], and observed consistency in the base pair overlap called between the two technical replicates (**Fig. 2.7b-c)**. Chromatin accessibility profiles mimicked published ENCODE DNase I data of GM12878 (**Fig. 2.7a**). In comparison with ENCODE reference data there was 61% base pair overlap with Duke data, and 70% overlap with UW data from the same cell line[42], and this was comparable to base pair overlap between the ENCODE datasets from different labs when overlapped against each other (**Fig. 2.7d, Table 1**). Additionally, base pair overlap between published 50,000-cell ATAC-seq data and 100-cell THS-seq/Tn5059 data was 61%, which is comparable to ENCODE data base pair overlaps against themselves as well as the number of peaks called and peak size distributions from published 50,000-cell ATAC-seq data, indicating a ~500-fold improvement of sensitivity (**Fig. 2.7d, e-f, Table 2**). This increase in sensitivity is accompanied by a 30% increase in the total base pairs called

significant by 100-cell THS-seq/Tn5059 over published 50,000-cell ATAC-seq data (**Table 3**). In contrast, the control data generated from 6 nanograms of purified genomic DNA, where all proteins have been removed and DNA is completely open, yielded mostly non-specific peaks that have less than 1% overlap with ENCODE datasets (**Fig. 2.7a, Table 1**). Furthermore, an analysis by GREAT[43] shows accessible regions form a bimodal distribution of peaks around transcription start sites, suggesting that accessible regions largely lie over regulatory non-coding regions (**Fig. 2.8a**). Annotation of accessible regions identified enrichment of immune system and B lymphocyte biological processes, including interferon signaling, B cell receptor signaling pathway, and B cell homeostasis, which are known to be upregulated in EBV transformed lymphoblastoid cells[44-46] (**Fig. 2.9)**. There is a well-known limitation with ATAC-seq where 30-50% of sequenced reads are clonal reads from the mitochondrial genome, however with 100-cell THS-seq/Tn5059, only 6% of mapped reads came from the mitochondrial genome (**Fig. 2.10a-c**).

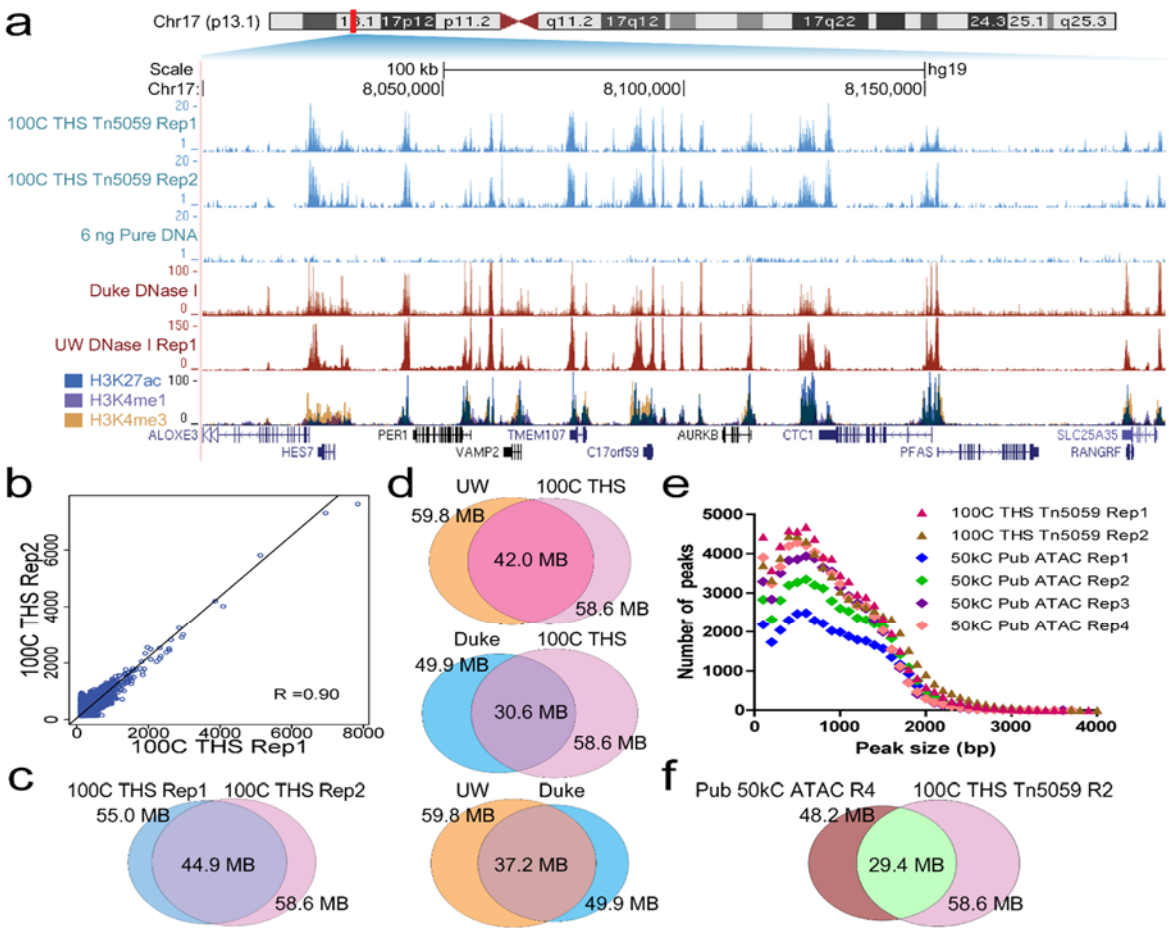


Figure 2.7. Validation of 100-cell THS-seq/Tn5059 data against ENCODE data and ATAC-seq data.
(a) 200 kb view of accessible chromatin marks in GM12878 lymphoblastoid cells in 100-cell THS-seq/Tn5059 data, a purified DNA control, published ENCODE accessibility data from Duke and UW, and ENCODE histone modifications which are often found near regulatory elements and promoters. (b) Correlation of 100-cell THS-seq/Tn5059 replicate one data and 100-cell THS-seq/Tn5059 replicate two data. (c) Base pair overlap of 100-cell THS-seq/Tn5059 replicate one data and 100-cell THS-seq/Tn5059 replicate two data. (d) Base pair overlap between 100-cell THS-seq/Tn5059 replicate two data and ENCODE datasets, and ENCODE datasets base pair overlap amongst themselves. 100-cell THS-seq/Tn5059 replicate two data was used in ENCODE comparisons because it had the most base pairs in peaks called significant. (e) Peak size distributions between 100-cell THS-seq/Tn5059 datasets, and published 50,000-cell ATAC-seq datasets. (f) Base pair overlap between 100-cell THS-seq/Tn5059 and published 50,000-cell ATAC-seq replicate 4 data. Published 50,000-cell ATAC-seq replicate 4 data was used since it had the most base pairs called significant compared to the other published 50,000-cell ATAC-seq datasets.

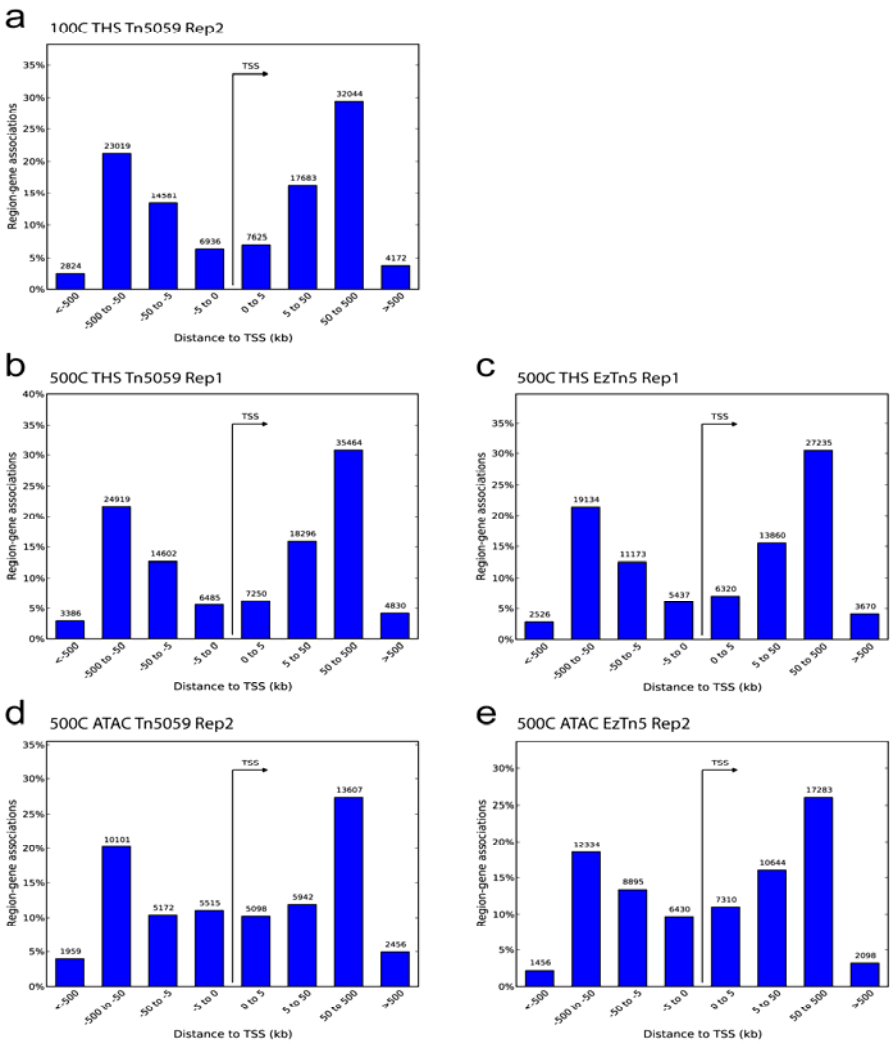


Figure 2.8. Comparison of peak region distances to transcription start sites.

For peak calling all samples were down sampled to 8,351,125 unique reads. Peak files generated by Dfilter were inputted into GREAT for analysis. Peak regions that caused database errors were removed, which was usually one to ten peaks per file, and thus should not impact the analysis. The replicate represented in each analysis was chosen based on if they had more base pairs called significant, and more base pair overlap with ENCODE data. Datasets represented are: (a) 100-cell THS-seq with Tn5059 Rep2, (b) 500-cell THS-seq with Tn5059 Rep1, (c) 500-cell THS-seq with EzTn5 Rep1, (d) 500-cell ATAC-seq with Tn5059 Rep2, (e) 500-cell ATAC-seq with EzTn5 Rep2.

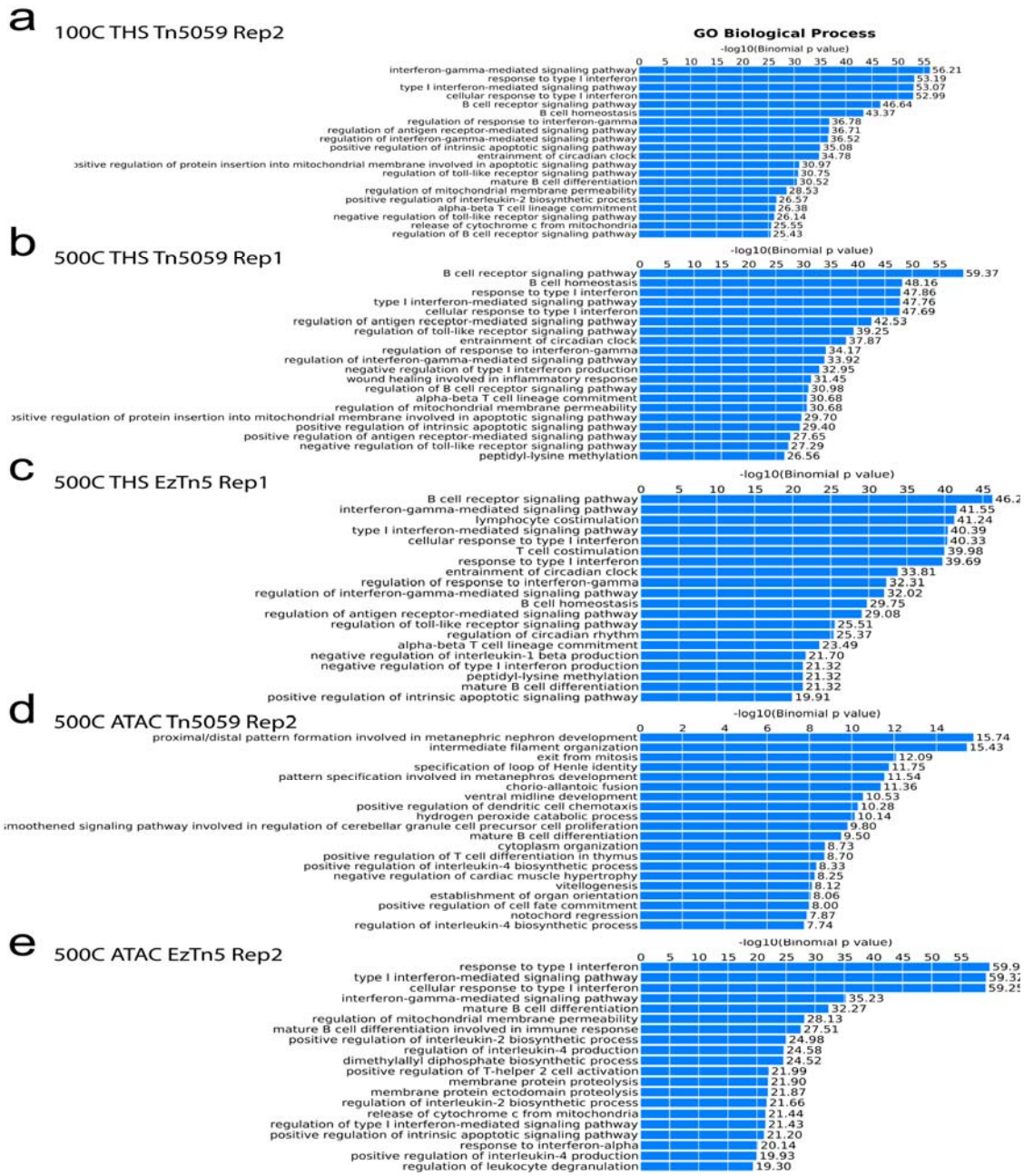


Figure 2.9. Significant gene ontology biological processes categories for each peaks dataset.

This is part of the same analysis performed by GREAT that was used in Figure 2.8, and thus the same peaks files are used as input. The replicate represented was chosen based on if they had more base pairs called significant, and more base pair overlap with ENCODE data. Datasets represented are: (a) 100-cell THS-seq with Tn5059 Rep2, (b) 500-cell THS-seq with Tn5059 Rep1, (c) 500-cell THS-seq with EzTn5 Rep1, (d) 500-cell ATAC-seq with Tn5059 Rep2, (e) 500-cell ATAC-seq with EzTn5 Rep2.

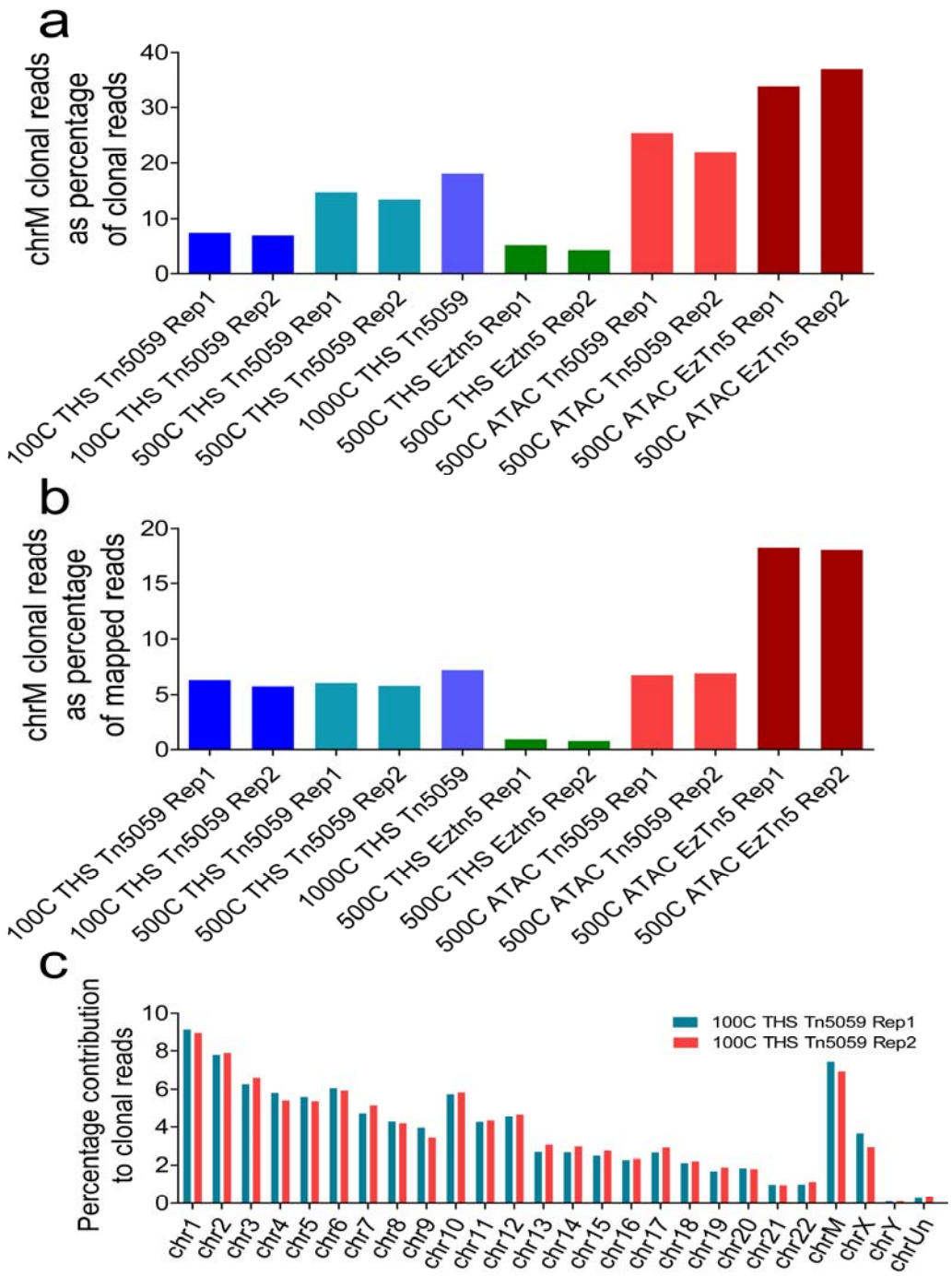


Figure 2.10. Quantitation of mitochondrial reads in all datasets
The number of mitochondrial reads (chrM) were counted and represented as a percentage of (a) the total number of clonal reads, and (b) the total number of mapped reads in each dataset. (c) In 100-cell THS-seq with Tn5059, the proportion of clonal reads from each chromosome was calculated as a percentage of the total clonal reads for each sample.

2.2.3 Comparison of THS-seq and ATAC-seq with two transposases: Tn5059 and EzTn5

We next sought to identify the contributions to the improved sensitivity by individual components of THS-seq. More specifically, we asked whether a direct replacement of the transposase in ATAC-seq, which had fewer processing steps, would lead to a similar level of improvement, and whether Tn5059 led to a substantial improvement over EzTn5 with the THS-seq protocol. We used an input of 500 cells in order to directly compare with published 500-cell ATAC-seq data[5]. We performed ATAC-seq and THS-seq with EzTn5 and Tn5059, generating data for all four combinations in two replicates. To eliminate the effects of variable read depth among different data sets, we down-sampled all data sets to 8,351,125 uniquely aligned reads per sample, which matched the sample with the lowest number of unique alignments, and was adequate for generating accessibility data metrics (**Fig. 2.11, Table 3, Table 4)**. Technical replicates consistently have high base pair overlap, peak overlap and correlations with each other **(Fig. 2.12a, Fig. 2.13a, 2.14a).** 500-cell THS-seq/Tn5059 consistently performs the best by generating 61% more significant peaks, calling 54% more base pairs significant, having 45% more peak overlap with ENCODE UW peaks, having 16% more peak overlap with ENCODE Duke peaks, and having 30% more base pairs overlapped with UW base pairs, when compared with 500-cell ATAC-seq/EzTn5 data, which is the next most comprehensive accessibility dataset (**Fig. 2.12b-d, Fig. 2.13b-c, Fig. 2.14b,**

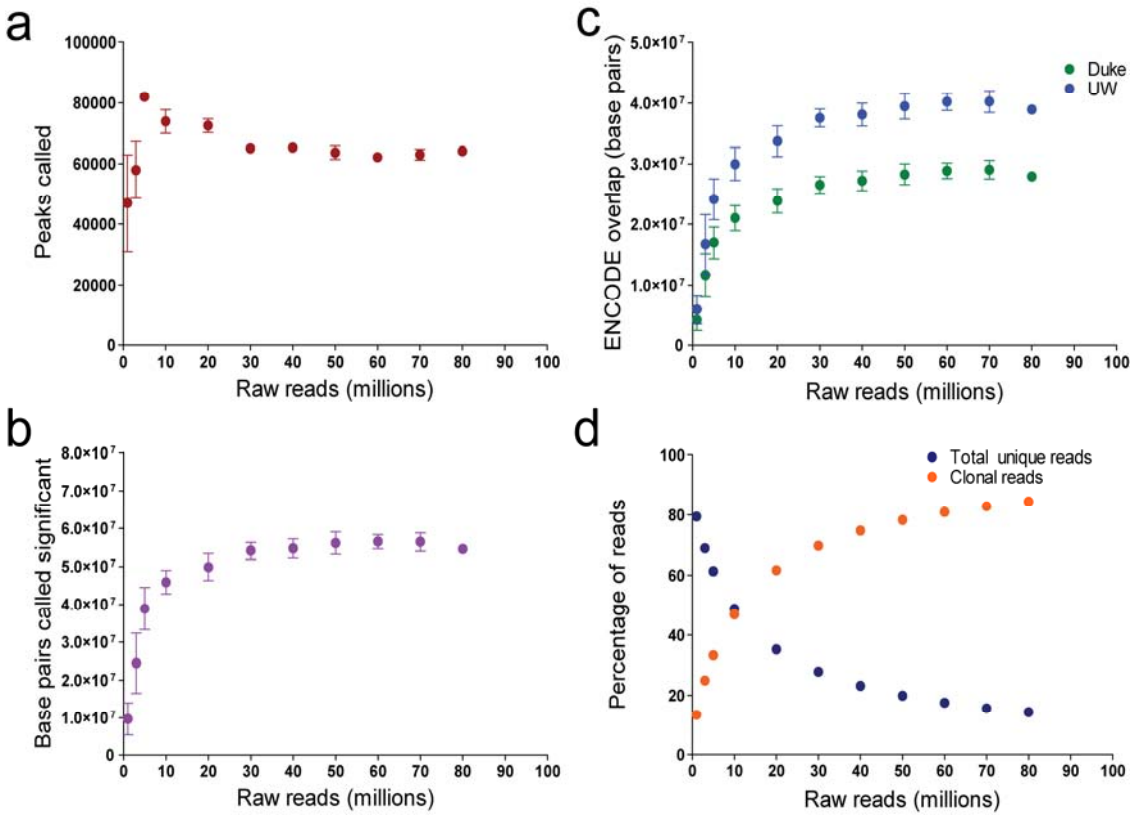


Figure 2.11. 100-cell THS-seq with Tn5059 read statistics analysis to determine minimal number of raw reads for quality data.
Raw reads were down sampled in increments of 10 million reads with in house scripts, run through BWA with default parameters, and then were run through Dfilter to (a) call peaks, (b) determine total number of base pairs called significant, (c) calculate ENCODE data overlap, and (d) determine the total number of unique and clonal reads. The mean and standard deviation between the two 100-cell THS-seq with Tn5059 replicates was calculated and displayed.

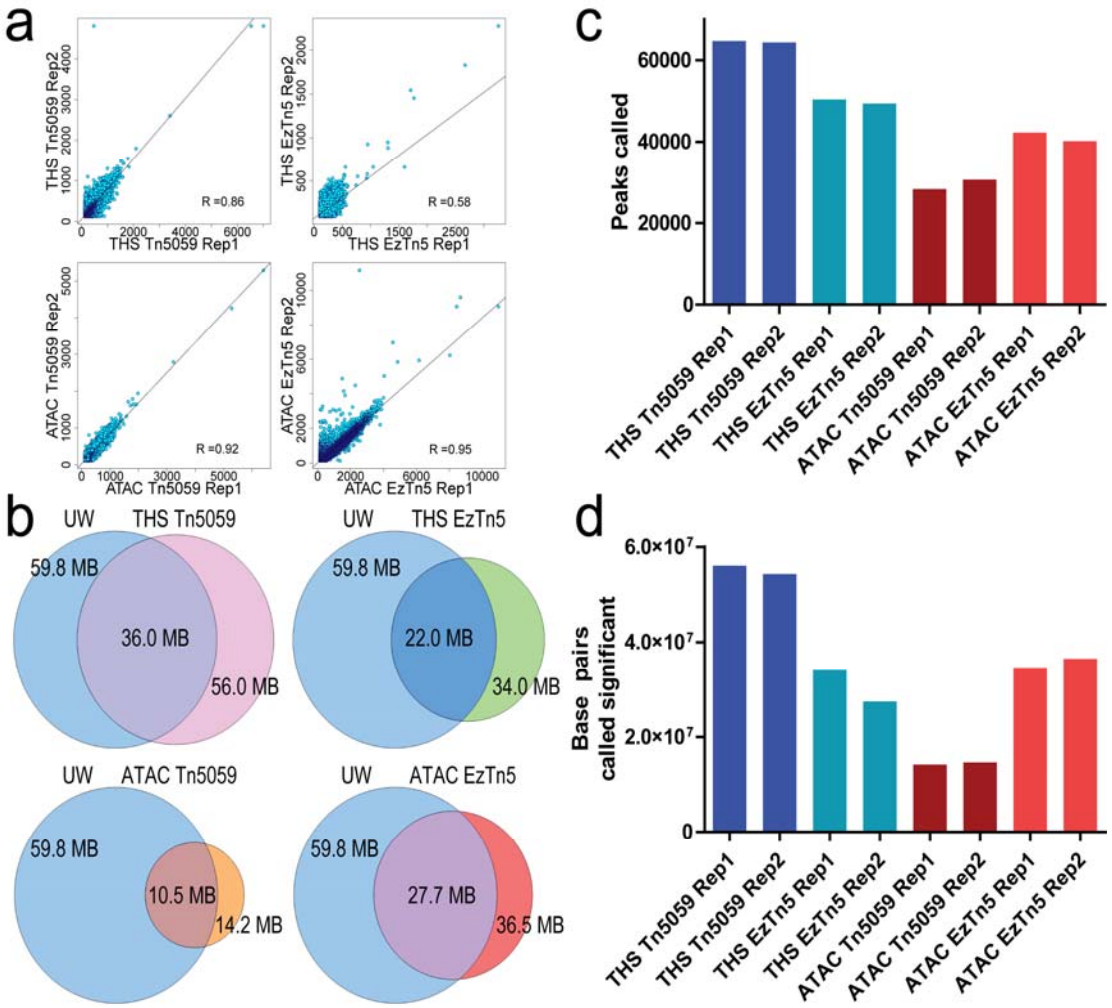


Figure 2.12. Comparison between THS-seq/Tn5059, THS-seq/EzTn5, ATAC-seq/Tn5059, and ATAC-seq/EzTn5 with 500 cells of input material. All datasets and replicates were down sampled to 8,351,125 unique alignments before analysis. (a) Correlation between replicates for each experimental condition. (b) Base pair overlap of each experimental condition with UW data. The replicate with the most base pairs called significant was used in analysis and represented in each condition. UW data was chosen since it had the most base pairs called significant of the ENCODE datasets. (c) Total number of peaks called by Dfilter for each condition. (d) Total number of base pairs under peaks called significant by Dfilter

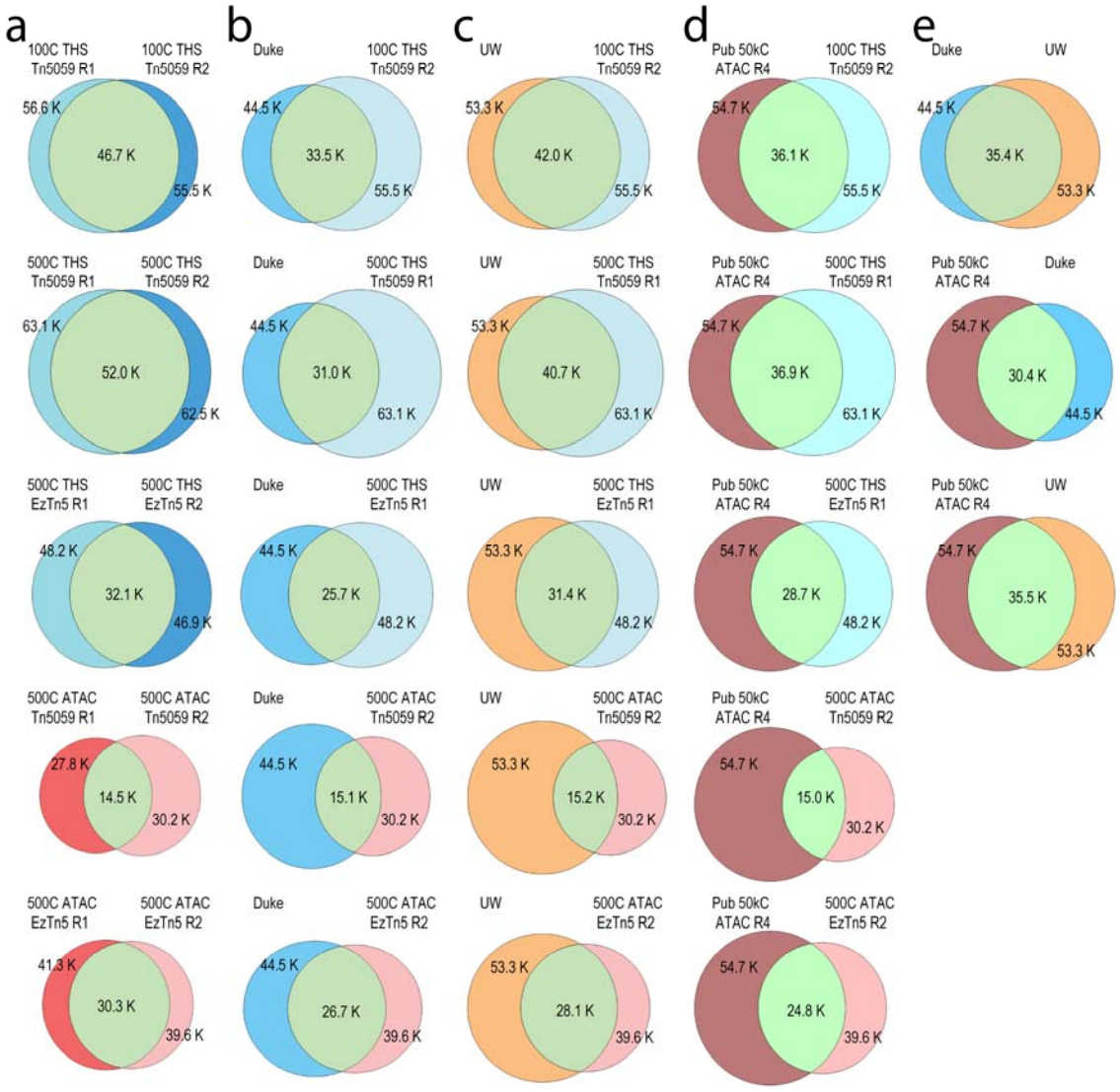


Figure 2.13. Peak overlap comparisons in all datasets.
Peaks were counted as overlapped if at least one base pair is overlapped in two significant peaks that are overlapped against each other in the datasets being compared. All samples except ENCODE Duke and UW samples were down sampled to the same number of unique reads, 8,351,125, before peak calling. (a) Peak overlap for all experimental replicates when overlapped against each other. (b) Peak overlap for the experimental replicate with the most base pairs called significant against ENCODE Duke data. (c) Peak overlap for the experimental replicate with the most base pairs called significant against ENCODE UW data. (d) Peak overlap for the experimental replicate with the most base pairs called significant against published 50,000-cell ATAC-seq replicate 4 data. (e) Peak overlap of ENCODE UW and Duke data against themselves, and ENCODE UW and Duke data against published 50,000-cell ATAC-seq replicate 4 data.

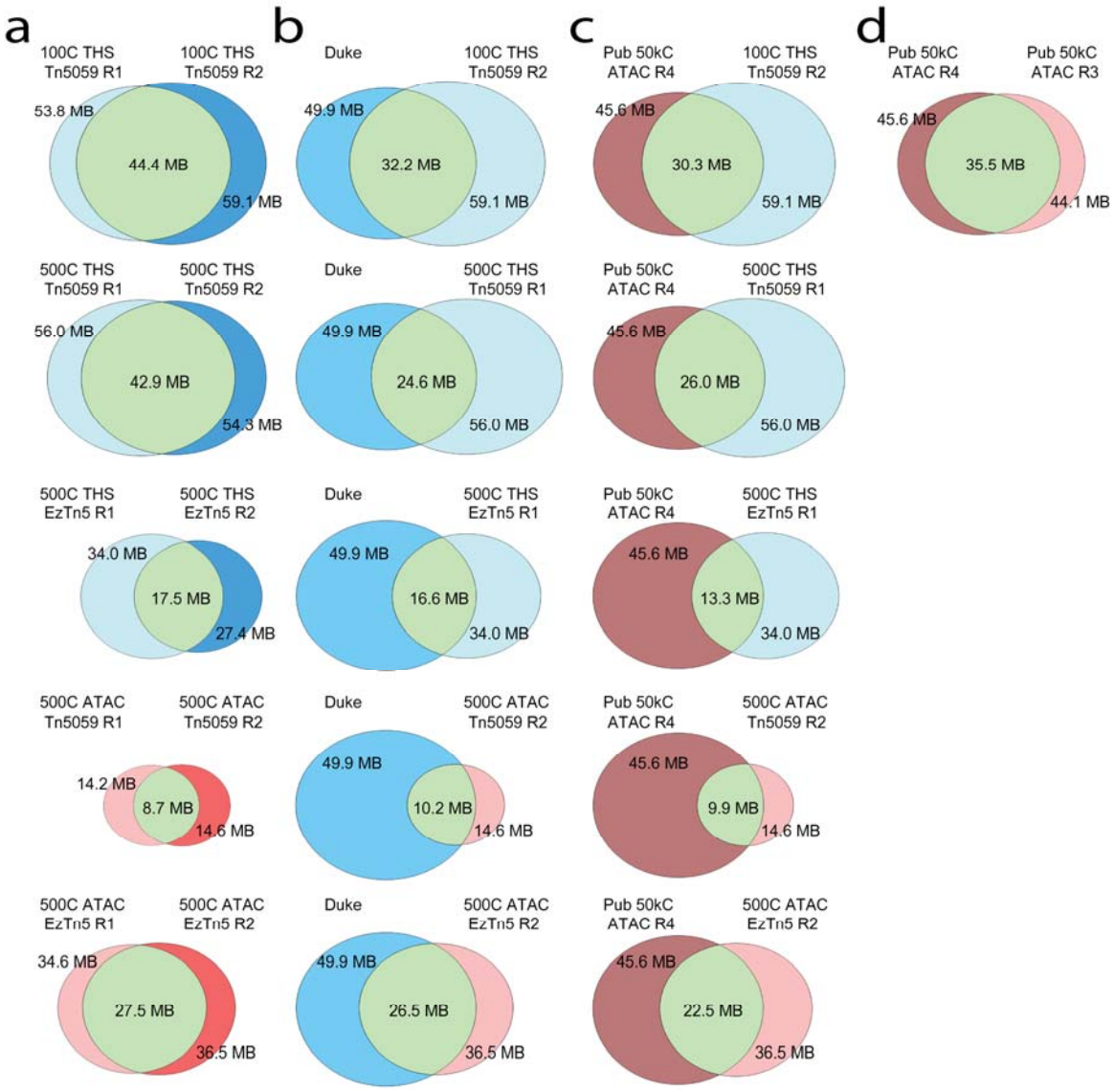


Figure 2.14. Base pair overlap comparisons in all datasets.
If a base pair that is within a significant peak is overlapped with the same base pair in a significant peak in the dataset being compared to, then it is called significant and counted. All samples except ENCODE Duke samples were down sampled to the same number of unique reads, 8,351,125, before peak calling. (a) Base pair overlap for all experimental replicates when overlapped against each other. (b) Base pair overlap for the experimental replicate with the most base pairs called significant against ENCODE Duke data. (c) Base pair overlap for the experimental replicate with the most base pairs called significant against published 50,000-cell ATAC-seq replicate 4 data. (d) Base pair overlap of the two most comprehensive published 50,000-cell ATAC-seq datasets, replicate 3 and replicate 4.

Table 3). Peak overlaps are comparable to ENCODE and published 50,000 ATAC-seq data overlap amongst themselves (**Fig. 2.13e**). Pairwise comparisons of peak overlap and base pair overlap between all four combinations show that THS-seq or ATAC-seq against each other, regardless of enzyme used, have the best overlap (**Fig. 2.15, Fig. 2.16**). The two most comprehensive datasets, 500-cell THS-seq/Tn5059 and 500-cell ATAC-seq/EzTn5 when overlapped against each other have 62% of the same peaks covered, and 50% of the same base pairs covered, of the total from the 500-cell ATAC/EzTn5 dataset (**Fig. 2.15, Fig. 2.16**). Interestingly 500-cell ATAC-seq/EzTn5 has 7.7% more base pair overlap with Duke data than 500-cell THS-seq/Tn5059, possibly because 500-cell ATAC-seq/EzTn5 data captures larger portions of peaks, where THS-seq instead captures individual peaks in the same region that are smaller (**Fig. 2.13b, Fig. 2.14b)**. This can be explained by the increased sensitivity of THS-seq/Tn5059, which in this case would result in less base pair coverage and more overlapping peaks called. It could also be that the Duke data was generated differently due to protocol differences, and is capturing larger peaks[6]. Compared with published 50,000-cell ATAC-seq data, 500-cell THS-seq/Tn5059 has 49% more peak overlaps and 16% more base pair overlaps than 500-cell ATAC-seq/EzTn5. 500-cell THS-seq/Tn5059 data also has more comparable peak overlap of 36,846 peaks to published 50,000-cell ATAC-seq data and Duke data peak overlap of 30,417 peaks, and 50,000-cell ATAC-seq and UW data peak overlap of 35,501 peaks. 500-cell ATAC-seq/EzTn5 peak overlap with 50,000-cell published ATAC-seq data is comparably less than ENCODE datasets overlap with 50,000-cell

published ATAC-seq data, with 19% less Duke data peak overlap, and 30% less UW data overlap, with 24,759 peaks. (**Fig. 2.13d-e, Fig. 2.14c**). A higher fraction of overlapping peaks and base pairs between THS-seq and ENCODE reference data indicates that THS-seq not only has a higher sensitivity, but also a higher specificity, which could be due to either less sequence-specific bias or less off-target transposition or both.

THS-seq with EzTn5 provides an obvious improvement for all metrics over the published ATAC-seq data from 500 cells[5]. Compared with the 500-cell data we generated in this study using ATAC-seq/EzTn5 in our optimized protocol, THS-seq/EzTn5 identified 25% more peaks that are smaller in size, such that 6.7% less base pairs were covered (**Fig. 2.12b-d, Table 3**). Unexpectedly, ATAC-seq/Tn5059 results in the least number of peaks called, base pairs called significant and base pair overlap with ENCODE reference data (**Fig. 2.12b-d, Table 3**). This is also reflected in the gene ontology analysis by GREAT[43], which reported ontology categories mostly not involved with immune system function for ATAC-seq/Tn5059 peaks (**Fig. 2.9d**). Furthermore, the data is reflected when visually inspecting the data tracts, where THS-seq/Tn5059, and ATAC-seq/EzTn5 have the most well defined peaks (**Fig. 2.17)**.

THS-seq and ATAC-seq, regardless of which enzyme was used, identified similar number of peaks, 5,098-7,310, within 5 kilobases of transcription start sites. However, THS-seq called 1.5 to two times more peaks in distal regions to transcription start sites, suggesting additional regulatory regions are being captured by THS-seq that are not being captured with ATAC-seq (**Fig. 2.18e)**.

This leads to gene ontology descriptions that appear more descriptive of GM12878 lymphoblastoids, such as the B cell receptor signaling pathway, and B cell homeostasis (**Fig. 2.8, Fig. 2.9**). Further inspection of these gene enrichments reveals genes where 100- and 500-cell THS-seq/Tn5059 was able to identify accessible regions, and 500-cell ATAC-seq/EzTn5 did not. Some notable examples include *ABCA12*, *FCRL4*, *IFI44*, *PDGFD*, and *PHLDB2* (**Fig. 2.19a-e**). Genes including *BCL2*, *CASP3*, *CD38*, the interferon gene cluster and specifically *IFNA2*, and *MAPK1* that have major roles in cancer and in immune system function were enriched by both 100- and 500-cell THS-seq/Tn5059 and 500-cell ATAC-seq/EzTn5, however 100- and 500-cell THS-seq/Tn5059 called two to six times as many peaks per gene region, and more accurately mirrored UW DNase-seq data from the same gene regions (**Fig. 2.20a-e**). Consistent with published results, ATAC-seq/EzTn5 mitochondrial reads account for more than 30% of the total reads, and account for three times the percentage of mapped reads when compared to THS-seq. ATAC-seq/Tn5059 mitochondrial read percentages are similar to THS-seq, probably due to the difference of sequence preference between EzTn5 and Tn5059 (**Fig. 2.10a-b**).

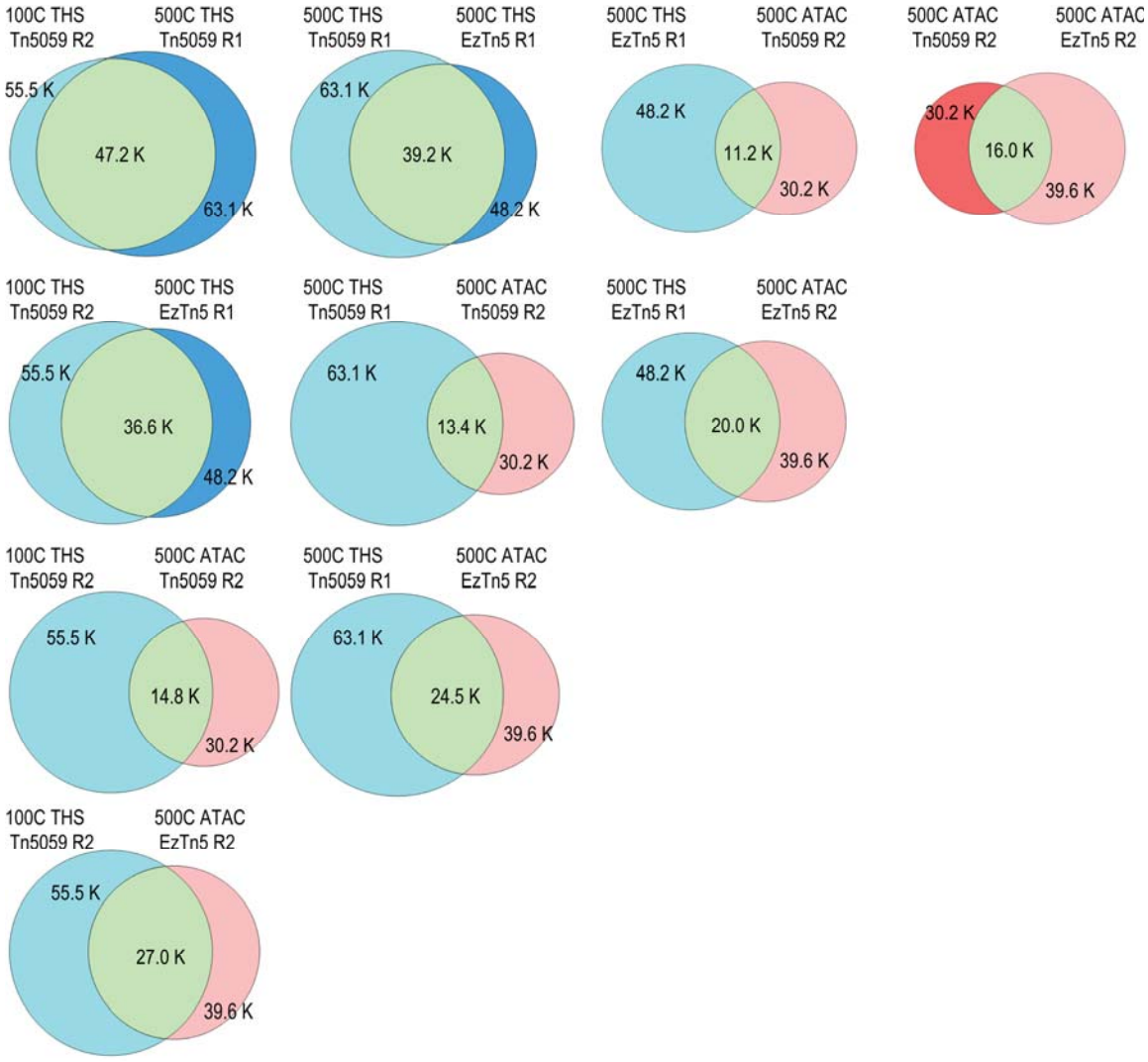


Figure 2.15. Examination of peak overlap in a pairwise comparison of all experimental datasets.

Peaks were counted as overlapped if at least one base pair is overlapped in two significant peaks that are overlapped against each other, in the datasets being compared. All samples were down sampled to the same number of unique reads, 8,351,125, before peak calling.

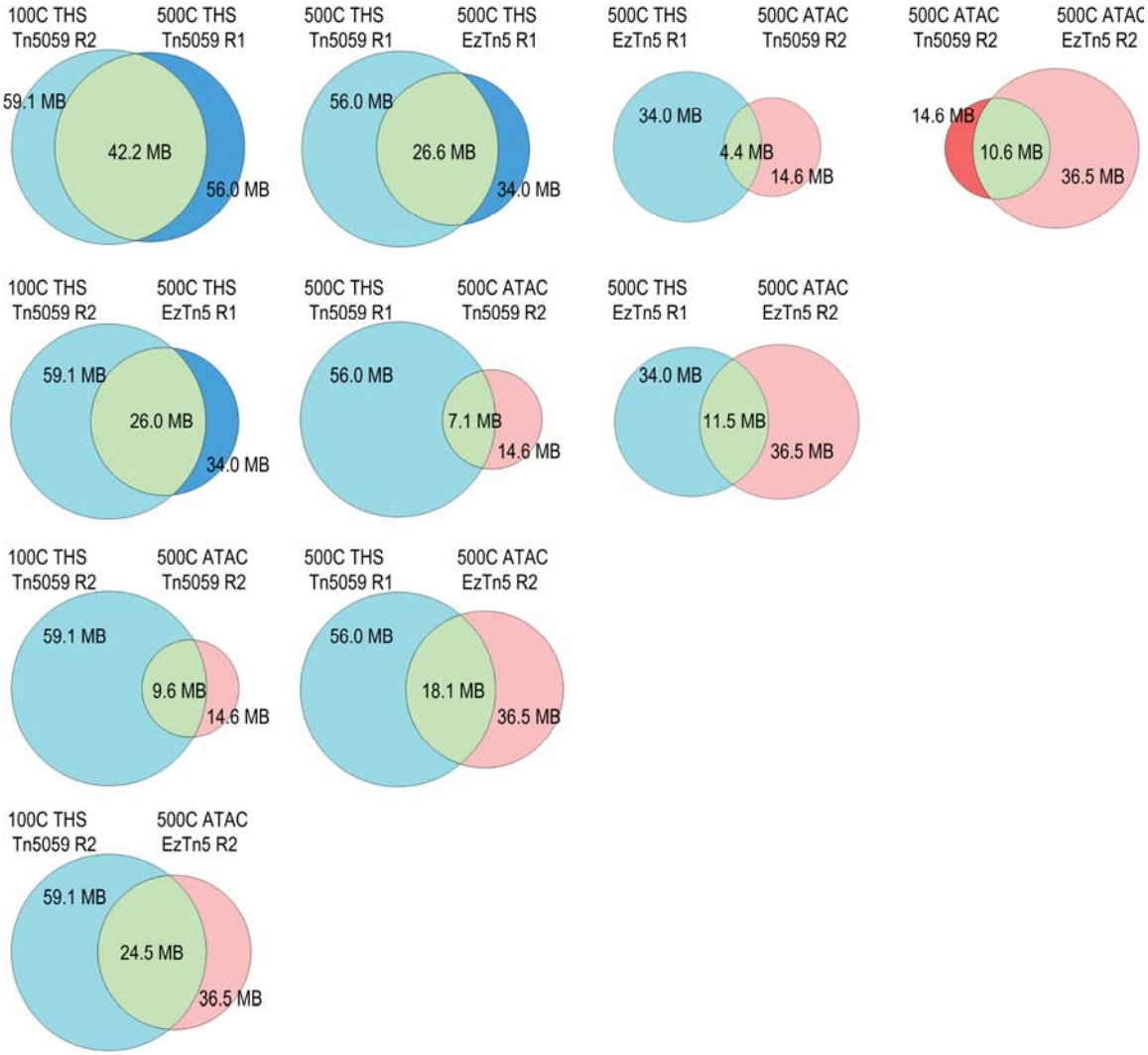


Figure 2.16. Examination of base pair overlap in a pairwise comparison of all experimental datasets.
If a base pair that is within a significant peak is overlapped with the same base pair in a significant peak in the dataset being compared to, then it is called significant and counted. All samples were down sampled to the same number of unique reads, 8,351,125, before peak calling.

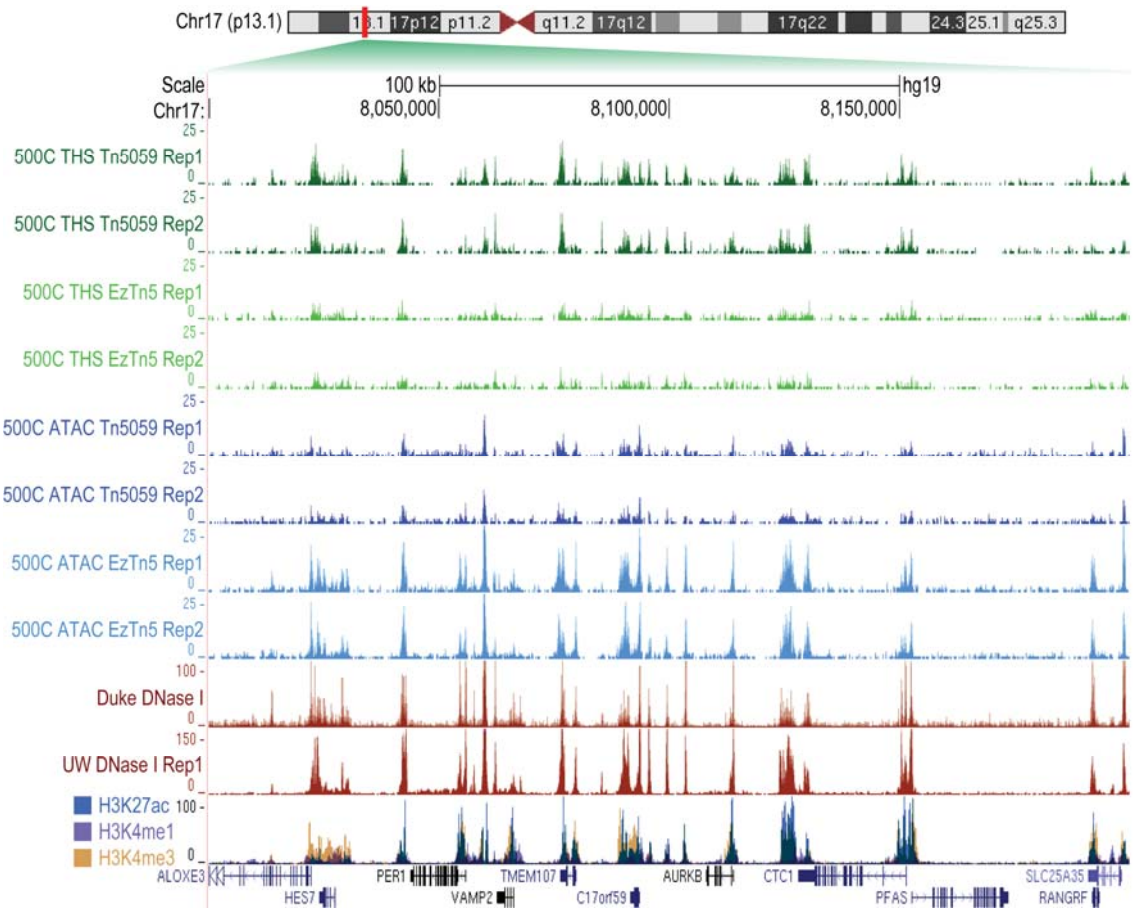


Figure 2.17. 200 kb view of accessible chromatin marks in 500 cell datasets.
Comparison between THS-seq with Tn5059, THS-seq with EzTn5, ATAC-seq with Tn5059, and ATAC-seq with EzTn5 against ENCODE GM12878 lymphoblastoid cell chromatin accessibility data from Duke and UW, and against histone modifications which are often found near regulatory elements and promoters. All samples were down sampled to the same number of unique reads, 8,351,125, before viewing tracts.

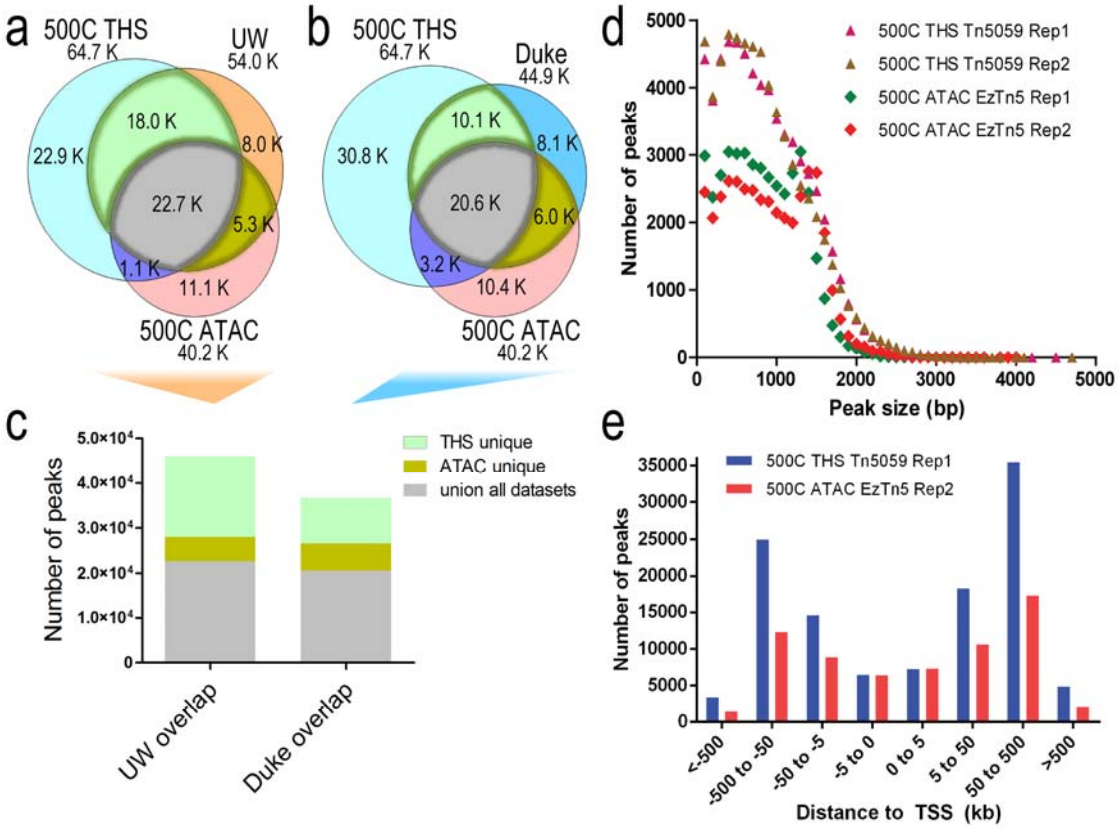


Figure 2.18. Comparison between the two most comprehensive datasets of 500-cell THS-seq/Tn5059 and 500-cell ATAC-seq/EzTn5.
(a) Venn diagram depicting peak overlap between 500-cell THS-seq/Tn5059, 500-cell ATAC-seq/EzTn5 and ENCODE UW data. A peak is shared if one base pair or more overlaps with a peak in the dataset being compared to. (b) Venn diagram depicting peak overlap between 500-cell THS-seq/Tn5059, 500-cell ATAC-seq/EzTn5 and ENCODE Duke data. A peak is shared if one base pair or more overlaps with a peak in the dataset being compared to. (c) Representation of the number of peaks that are shared between all three datasets for UW data, peaks that are found by 500-cell THS-seq/Tn5059 data and ENCODE UW data and not 500-cell ATAC-seq/EzTn5 data, and peaks that are found by 500-cell ATAC-seq/EzTn5 data and ENCODE UW data and not 500-cell THS-seq/Tn5059 data. Also a representation of the number of peaks that are shared between all three datasets for Duke data, peak regions that are found by 500-cell THS-seq/Tn5059 data and ENCODE Duke data and not 500-cell ATAC-seq/EzTn5 data, and peaks that are found by 500-cell ATAC-seq/EzTn5 data and ENCODE Duke data and not 500-cell THS-seq/Tn5059 data. (d) Comparison of peak size distributions. (e) Peak distances from transcription start sites as determined by GREAT[43].

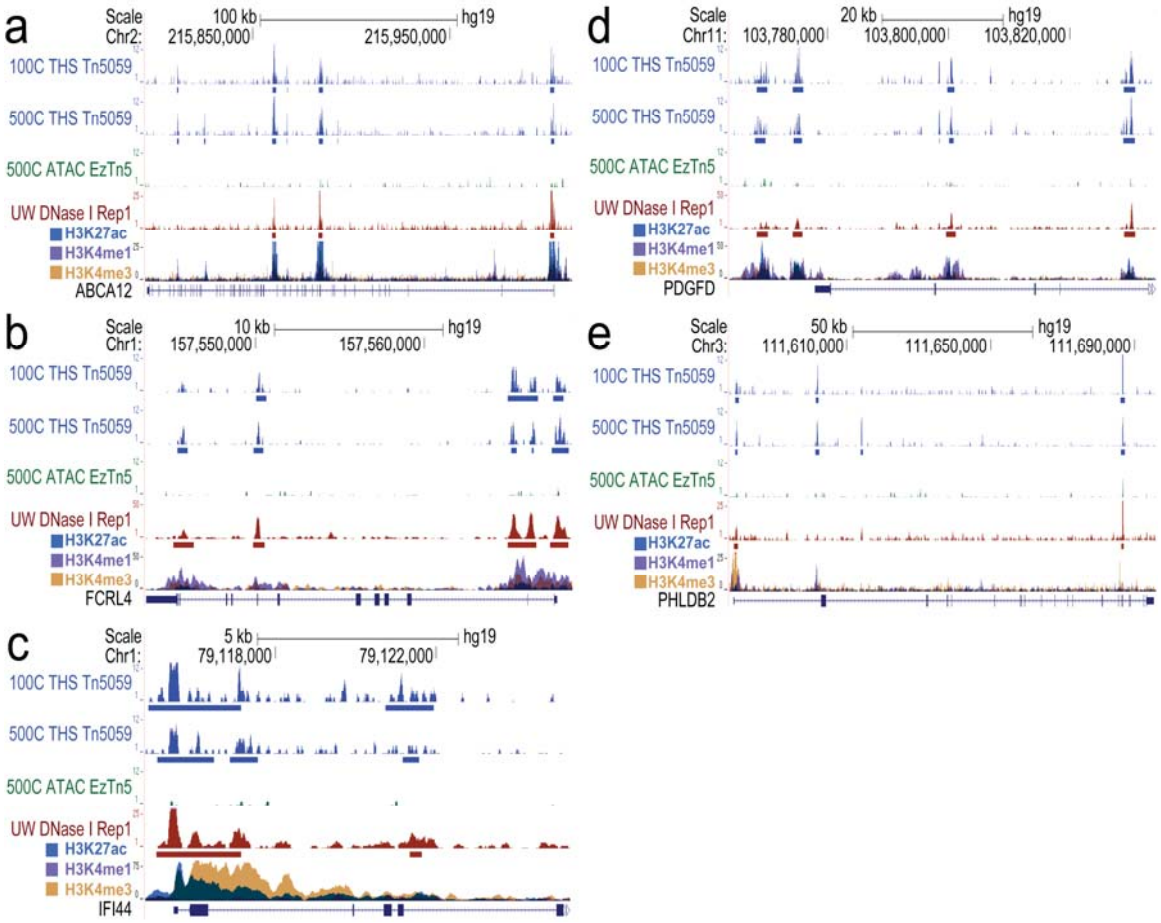


Figure 2.19. Overview of chromatin accessibility at gene loci enriched in 100- and 500-cell THS-seq with Tn5059 data and not enriched in 500-cell ATAC-seq with EzTn5 data.

Enriched genes from GREAT were used to determine which genes were enriched in THS-seq and not ATAC-seq. 100- and 500-cell THS-seq with Tn5059 and 500-cell ATAC-seq with EzTn5 were down sampled to the same number of unique reads, 8,351,125, before analysis. UW DNase chromatin accessibility data was used for comparison since it was the most comprehensive of the GM12878 ENCODE datasets. ENCODE histone modifications which are often found near regulatory elements and promoters are also depicted. The replicate with the highest number of base pairs called significant was used in the read density data tract for each sample and for analysis. Also under each tract, the regions called significant by Dfilter are represented by bars. The gene loci for (a) ABCA12, (b) FRCL4, (c) IFI44, (d) PDGFD, and (e) PHLDB2 are represented.

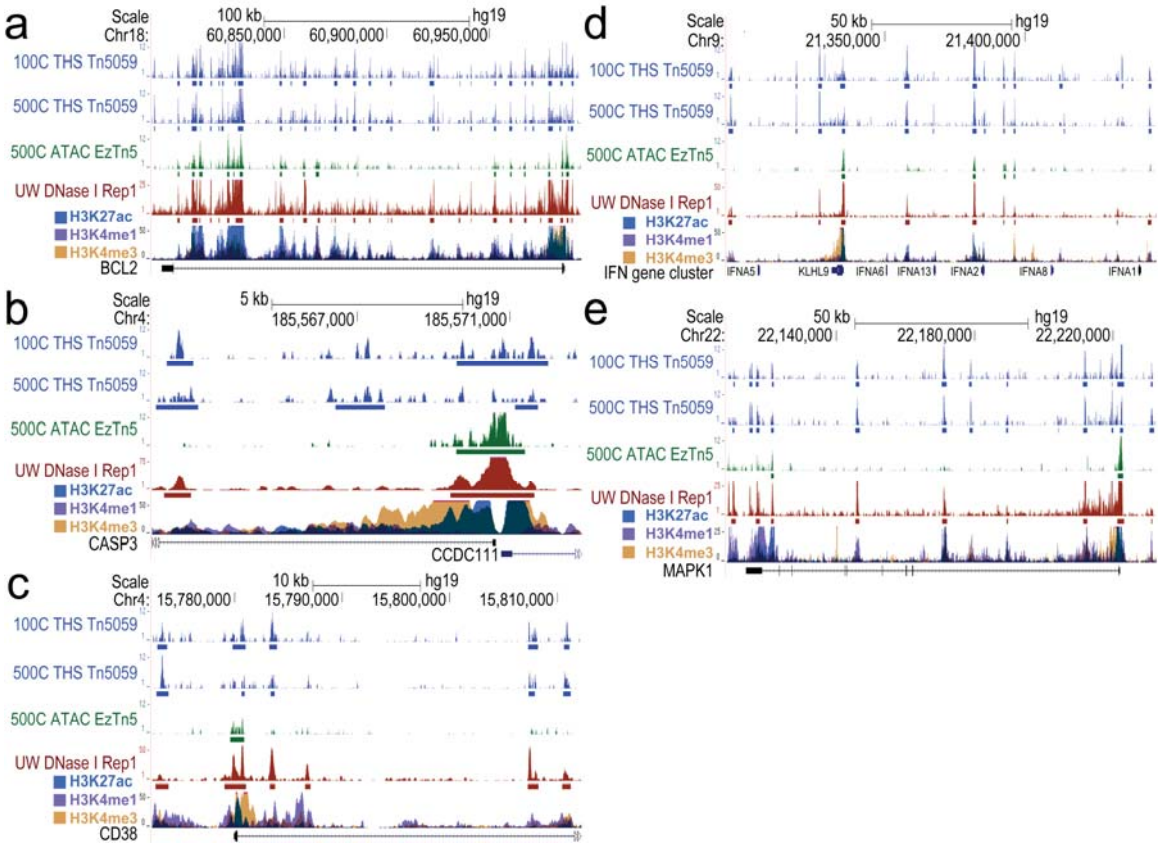


Figure 2.20. Overview of chromatin accessibility at major genes implicated in cancer and in immune system function.
100- and 500-cell THS-seq with Tn5059 data and 500-cell ATAC-seq with EzTn5 data were down sampled to the same number of unique reads, 8,351,125, before analysis. UW DNase chromatin accessibility data was used for comparison since it was the most comprehensive of the GM12878 ENCODE datasets. ENCODE histone modifications which are often found near regulatory elements and promoters are also depicted. The replicate with the highest number of base pairs called significant was used in the read density data tract for each sample and for analysis. Also under each tract, the regions called significant by Dfilter are represented by bars. The gene loci for (a) BCL2, (b) CASP3, (c) CD38, (d) the interferon gene cluster and specifically IFNA2, and (e) MAPK1 are represented. Enrichments performed by GREAT reported BCL2, CD38 and MAPK1 in the B cell receptor signaling pathway gene ontology biological process category. BCL2, CD38 and MAPK1 were reported for the B cell homeostasis gene ontology biological process category. IFNA2, IFNA1, IFNA5, IFNA6 and IFNA8 were reported for the response to type I interferon gene ontology biological process category.

2.2.4 Comparison of THS-seq/Tn5059 and ATAC-seq/EzTn5

We next sought to further characterize and directly compare the two most comprehensive datasets of 500-cell THS-seq/Tn5059 and 500-cell ATAC-seq/EzTn5. In order to determine the number of peaks that were found unique between ENCODE UW data and 500-cell THS-Seq/Tn5059 data, and not found by 500-cell ATAC-seq/EzTn5 data, a union dataset was made. Here we found 500-cell THS-seq/Tn5059, 500-cell ATAC-seq/EzTn5 and UW had 22,700 peaks in common, while 500-cell THS-seq/Tn5059 data had 240% more uniquely identified peaks that overlap with UW data than 500-cell ATAC-seq/EzTn5 data (**Fig. 2.18a-c)**. The same analysis was performed with Duke data, though 9.3% less peaks were found in common between all datasets, and 500-cell THS-seq/Tn5059 had 68% more uniquely identified peaks that overlap with Duke data than 500-cell ATAC-seq/EzTn5 data (**Fig. 2.18b-c**). Next, we examined the distribution of peak sizes for each dataset, and found peaks in the smaller size range of 100–1200 base pairs and larger size range of 1300-3000 base pairs, relative to 500-cell ATAC-seq/EzTn5, are gained with 500-cell THS-seq/Tn5059 (**Fig. 2.18d, Table 2**). These peaks are present in more proximal and distal regions from the transcription start site, suggesting involvement in regulatory regions (**Fig. 2.18e**). The best published 50,000-cell ATAC-seq data sets approach the capture rate of smaller peaks by THS-seq/Tn5059, suggesting higher sensitivity Tn5 based methods are able to capture these peaks (**Fig. 2.21f-g)**.

We next sought to confirm that these extra peaks 100-1200 base pairs in length are not artifact generated by THS-seq, but instead has biological significance. A possibility is that these smaller peaks overlap in portions of larger peaks in 500-cell ATAC-seq/EzTn5 data, and thus are being counted as parts of single large peaks, or could be a unique set of called peaks. To investigate this, we examined the total base pair overlap with ENCODE Duke and UW DNase-Seq data for each individual peak length. Since 75% and 125% more peaks in the size range of 100-1200 base pairs and 1700-3000 base pairs respectively are identified by 500-cell THS-seq/Tn5059 (**Fig. 2.18d**), we would expect a marked increase in the number of significant base pairs that overlap with ENCODE data at these peak lengths. Indeed with ENCODE UW data as the reference we observed an increase in overlapping base pairs in peaks 100-1200 base pairs and 1700-3000 base pairs (**Fig. 2.22a**), suggesting the additional small peaks detected by THS-seq are also present in ENCOCE UW data. However, when examining the Duke data, we found no difference among all three data sets in overlapping base pairs with peaks 100-1200 base pairs, but more base pair overlap with THS-seq data and Duke data in the range of 1700-3000 base pairs (**Fig. 2.22b**). Therefore, there is a difference in the peaks called by the two ENCODE data sets (Duke versus UW) that needs to be taken into account during analysis. Interestingly, the spike in base pair overlap seen from 1300-1600 base pairs in 500-cell ATAC-seq/EzTn5 data is due to both ATAC-seq/EzTn5 and ENCODE Duke and UW data better capturing peaks in the 1300-1600 base pair size range (**Fig. 2.22a-b, Fig. 2.21a-b, e**). Moreover, to further validate the

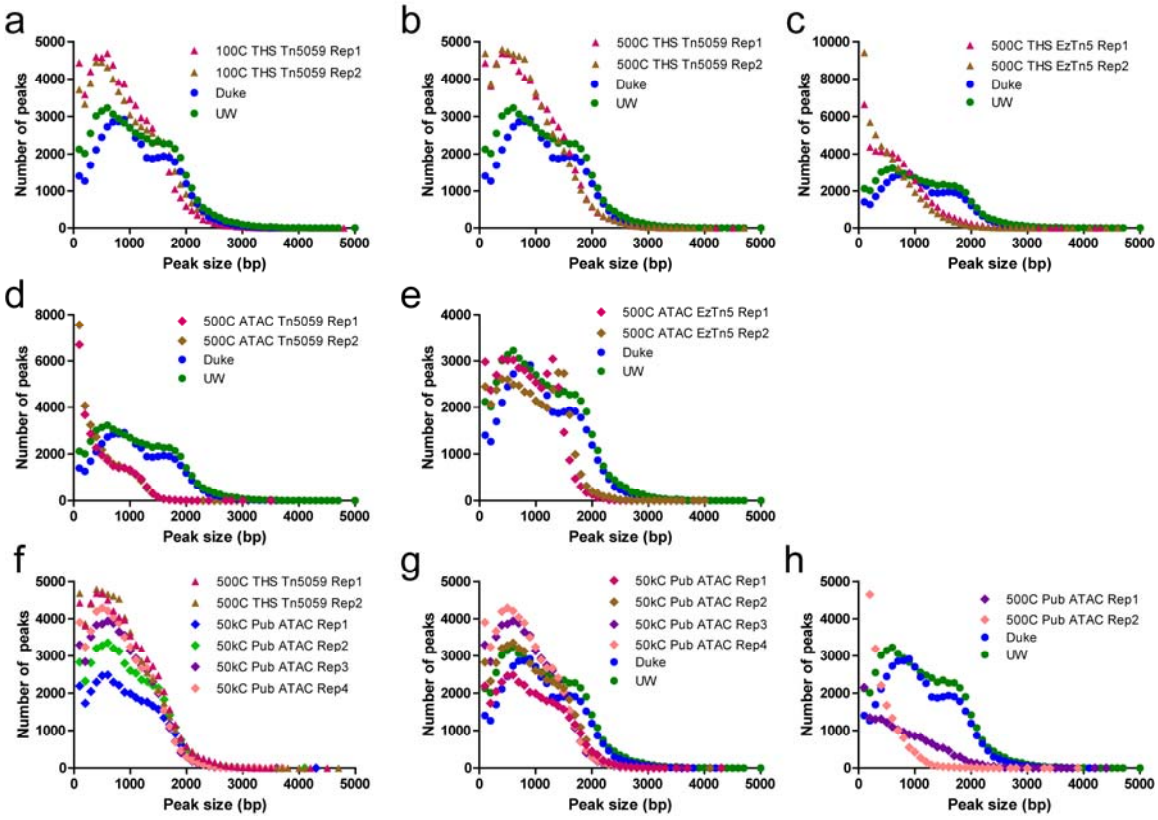


Figure 2.21. Peak size distribution comparisons between THS-seq, ATAC-seq and ENCODE data.

ENCODE data from Duke and UW for chromatin accessibility on GM12878 lymphoblastoid cells was processed as described in methods. All THS-seq and ATAC-seq datasets were down sampled to the same number of unique reads, 8,351,125, before analysis, except for 100-cell THS-seq with Tn5059 datasets, and the published 500-cell ATAC-seq replicate two dataset, which had less than 8,351,125 unique reads. For all datasets, counts of the number of peaks for each peak size were performed in 100 base pair increments, using the peaks file generated by Dfilter. Some data points were excluded from the graphs because values were beyond the axis, and the number of data points excluded for each graph is: (a) 14, (b) 8, (c) 8, (d) 8, (e) 8, (f) none, (g) 8, (h) 11. Samples compared to ENCODE data include: (a) 100-cell THS-seq with Tn5059, (b) 500-cell THS-seq with Tn5059, (c) 500-cell THS-seq with EzTn5, (d) 500-cell ATAC-seq with Tn5059, (e) 500-cell ATAC-seq with EzTn5, (g) Published 50,000-cell ATAC-seq data, and (h) published 500-cell ATAC-seq data. (f) A peaks distribution comparison of 500-cell THS-seq with Tn5059 was also done against published 50,000-cell ATAC-seq data.

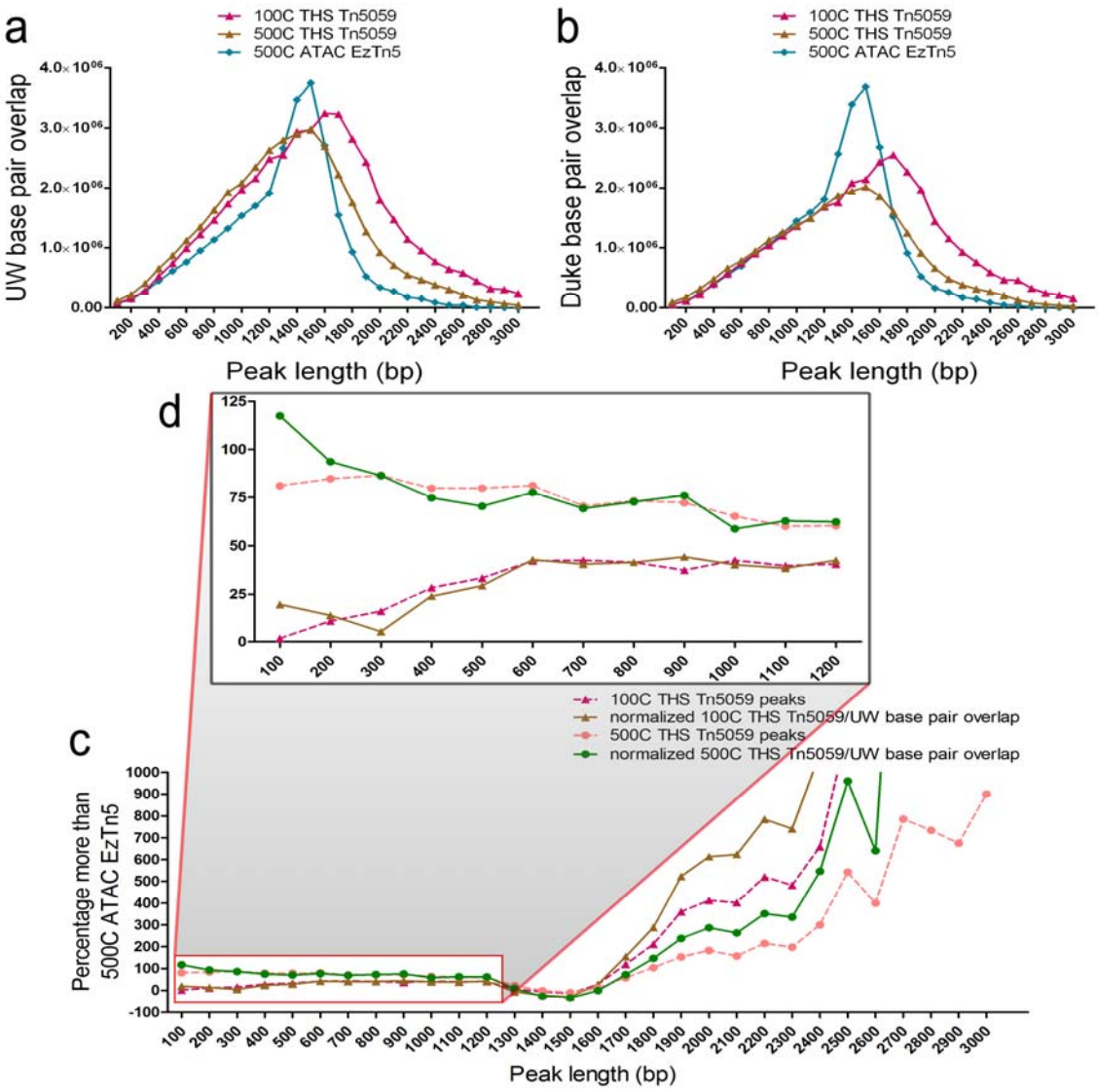


Figure 2.22. Validation of peaks based on peak length.

(a) Base pair overlap with ENCODE UW data based on peak lengths for 100-cell and 500-cell THS-seq/Tn5059 data, and 500-cell ATAC-seq/EzTn5 data. (b) Base pair overlap with ENCODE Duke data based on peak lengths for 100-cell and 500-cell THS-seq/Tn5059 data, and 500-cell ATAC-seq/EzTn5 data. (c) The percentage more peaks found in 100-cell and 500-cell THS-seq/Tn5059 than in 500-cell ATAC-seq/EzTn5 data, and the percentage more normalized 100-cell and 500-cell THS-seq/Tn5059 and UW base pair overlap than in 500-cell ATAC-seq/EzTn5. Normalizing was performed using the global base pair overlap values for each ENCODE dataset. (d) Zoom in on graph (c) showing the peak lengths between 100-1200 base pairs.

extra peaks found with THS-seq/Tn5059, we examined the percentage more identified peaks than 500-cell ATAC-seq/EzTn5, and the normalized THS-seq/Tn5059 and ENCODE UW overlap. We would expect that when we see 60-80% more peaks called in 500-cell THS-seq/Tn5059 than ATAC-seq/EzTn5, we would see 40% less base pairs overlap, which is what is actually seen for peaks 100-1200 base pairs in length (**Fig. 2.23a-b**). Global ENCODE UW overlap is 60% for 500-cell THS-seq/Tn5059 data, and 70% for 100-cell THS-seq/Tn5059 data. Normalizing THS-seq/Tn5059 and UW base pair overlap with those values shows the extra peaks called by THS-seq/Tn5059 in the 100-1200 base pair size range overlap at the expected percentages. (**Fig. 2.22c-d**). Additionally, there are more normalized base pair overlaps with 100 base pair peaks and 1700-3000 base pair peaks in 100- and 500-cell THS-seq/Tn5059, suggesting these peaks have higher base pair overlap then the average global ENCODE UW overlap percentage (**Fig. 2.22c-d, Fig. 2.23a**). This further suggests these THS-seq peaks represent true accessible regions, not artifacts.

Interestingly the same results are not reflected with ENCODE Duke data when examining normalized THS-seq/Tn5059 and ENCODE Duke base pair overlaps, when normalized to ENCODE Duke global overlap of 61% for 100 cell data, and 49% for 500 cell data (**Fig. 2.23c-f**). The normalized overlap percentage is greatly below the percentage of peaks called, for peaks 100-1200 base pairs in length, suggesting these peaks are not present in ENCODE Duke data. The exception is with normalized 500-cell THS-seq/Tn5059 ENCODE Duke base pair overlaps, where peaks 100-300 base pairs in length are above and

closely match the percentage more peaks, suggesting these additional peaks are also present in ENCODE Duke data (**Fig. 2.23e-f**). Interestingly there are 70% more normalized THS-seq/Tn5059 and ENCODE Duke base pair overlaps with 100 base pair peaks, and more normalized base pair overlaps with peaks greater than 1700 base pairs, suggesting these peaks have higher base pair overlap with ENCODE Duke data than the average global overlap percentage (**Fig. 2.23e-f**). Overall, we see some differences between THS-seq data and ENCODE Duke data that are not seen between THS-seq and ENCODE UW data, which could be due to factors such as protocol differences, sample preparation or stochastic differences in the GM12878 cells assayed.

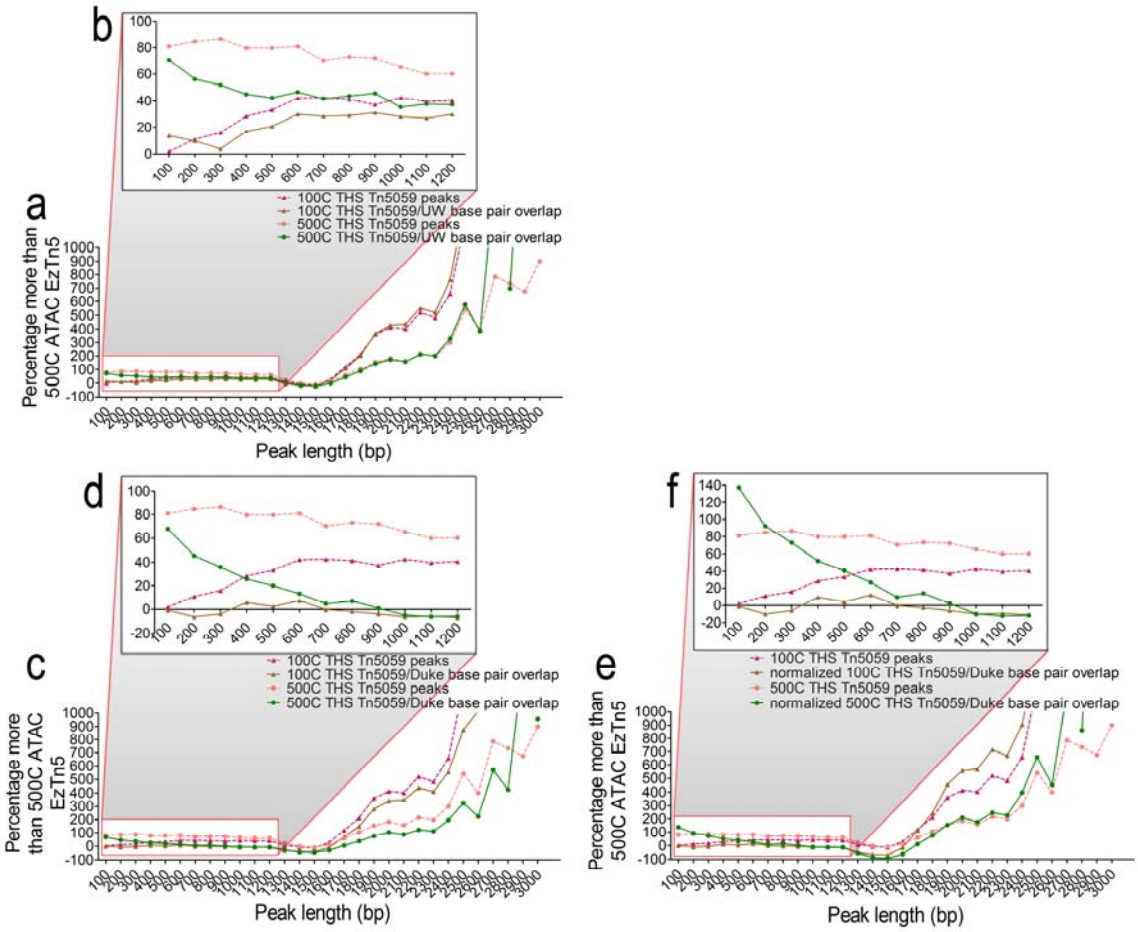


Figure 2.23. Validation of peaks based on peak length.
(a) The percentage more peaks found in 100-cell and 500-cell THS-seq with Tn5059 data then in 500-cell ATAC-seq with EzTn5 data, and the percentage more 100-cell and 500-cell THS-seq with Tn5059 and UW base pair overlap then in 500-cell ATAC-seq with EzTn5. (b) Zoom in on graph (a) showing the peak lengths between 100-1200 base pairs. (c) The percentage more peaks found in 100-cell and 500-cell THS-seq with Tn5059 data then in 500-cell ATAC-seq with EzTn5 data, and the percentage more 100-cell and 500-cell THS-seq with Tn5059 and Duke base pair overlap then in 500-cell ATAC-seq with EzTn5. (d) Zoom in on graph (c) showing the peak lengths between 100-1200 base pairs. (e) The percentage more peaks found in 100-cell and 500-cell THS-seq with Tn5059 data then in 500-cell ATAC-seq with EzTn5 data, and the percentage more normalized 100-cell and 500-cell THS-seq with Tn5059 and Duke base pair overlap then in 500-cell ATAC-seq with EzTn5. Normalizing was performed using the global base pair overlap values for the ENCODE Duke dataset. (f) Zoom in on graph (e) showing the peak lengths between 100-1200 base pairs

2.2.5 Characterization of capture sensitivity and ATAC-seq bias

We next investigated further the open chromatin regions, or peaks, called with the four combinations. Both THS-seq/Tn5059 and ATAC-seq/EzTn5 have high percentages of alignments in peaks, with 25% and 27% respectively. Replacement of EzTn5 with Tn5059 in THS-seq results in 2.0 to 2.8 times more unique alignments in peaks (**Fig. 2.24a**). From this it can be conjectured that Tn5059 has higher preference for areas of open chromatin and/or EzTn5 has a higher background sporadic transposition activity in more compact chromatin, which explains how THS-seq with EzTn5 has significantly more background noise. For ATAC-seq with EzTn5, this is actually beneficial when considering the areas directly adjacent to an accessible region are semi-accessible. EzTn5 would insert more into regions around an open region than would Tn5059 and would generate more PCR viable fragments. This then semi-compensates for the 50% sample loss from using ATAC-seq. Moreover, this explains how ATAC-seq with Tn5059 has the least number of unique alignments in peaks. Coupled with Tn5059's more selective insertion properties, ATAC-seq's inherent 50% sample loss upon transposition dilutes the number of alignments in any peaks by at least 50%, and hinders capture of any smaller regions of open chromatin. This is reflected in peak size distributions, where both ATAC-seq/Tn5059 and THS-seq/EzTn5, have 49% and 30% of peaks respectively in the size range of 100-300 base pairs long, which is 2-3 times the number of peaks found in that size range for THS-seq/Tn5059 or ATAC-seq/EzTn5. (**Fig. 2.21c-d, Fig. 2.25f, g, Table 2**). We hypothesized that higher background transposition events with

THS-seq/EzTn5, and sample loss with ATAC-seq/Tn5059 raised the thresholds for calling significant peaks, thus resulting in only the most significant portions of peaks being called significant and leading to smaller peak sizes. Further evidence reinforcing this interpretation is seen when examining the number of unique alignments within significantly called peaks, where THS-seq/EzTn5 and ATAC-seq/Tn5059 have the lowest with 11% and 6.4% of unique alignments in peaks, respectively (**Fig. 2.24a**). Here, the percentage of alignments in peaks is indicative of peak capture efficiency.

We next characterized peak capture bias in ATAC-seq, where preferentially peaks of longer length were captured, and peaks of shorter length were lost. ATAC-seq/EzTn5 had 20% more alignments in the larger 30% of peaks than did THS-seq/Tn5059, and concomitantly 27% less alignments in the smaller 70% of peaks respectively (**Fig. 2.24b-c**). Moreover, this is further illustrated as THS-seq/Tn5059 with both 100 and 500 cells of input material has an even distribution of normalized alignments over all peak lengths when compared to ATAC-seq/EzTn5, where normalized alignments increase as peak length increases beginning around a peak length of 1300 base pairs (**Fig. 2.24d, Fig. 2.25a, b, d-e, h**). The trend of more alignments in larger peaks is method consistent irrespective of enzyme used, and is seen in published 50,000-cell ATAC-seq data, though less so with published 500-cell ATAC-seq data due to sparse peak capture (**Fig. 2.24b-f, Fig. 2.25a-c, Fig. 2.26a-b**). We hypothesize the reason for THS-seq's more uniform capture of accessible regions is due to method design, usage of Tn5059, and is a result of the length of amplified

molecules. The importance of fragment length in determining the type of biological data captured has been demonstrated with DNase-FLASH[47], and with THS-seq the majority of fragments generated from *in vitro* transcription were 400-2000+ base pairs, while the majority of fragments from PCR amplification with ATAC-seq were 200-800 base pairs (**Fig. 2.4a, c**). Additionally, with THS-seq, the downstream transposition step for addition of 3' adaptors favors longer double stranded cDNA fragments. With longer fragment sizes and the hypothesis Tn5059 has higher preference for open chromatin, it is probable that THS-seq/Tn5059 is losing sample due to insertions being too close to each other in accessible regions of longer length, which does not happen with ATAC-seq since PCR amplification favors smaller fragment sizes. This would lead to ATAC-seq's preference for capturing larger accessible regions. However, this does not affect THS-seq peak capture, as this would only occur in large regions of accessible chromatin. It is possible to further modify the RNA processing method in THS-seq to capture shorter fragments. Taken together, these data demonstrated that THS-seq/Tn5059 is feasible and improves upon current Tn5 based methods for measuring chromatin accessibility with limited inputs. This is achieved by reducing peak capture bias introduced by ATAC-seq, while providing data from 100 cells comparable to DNase-seq on millions of cells, or ATAC-seq on 50,000 cells.

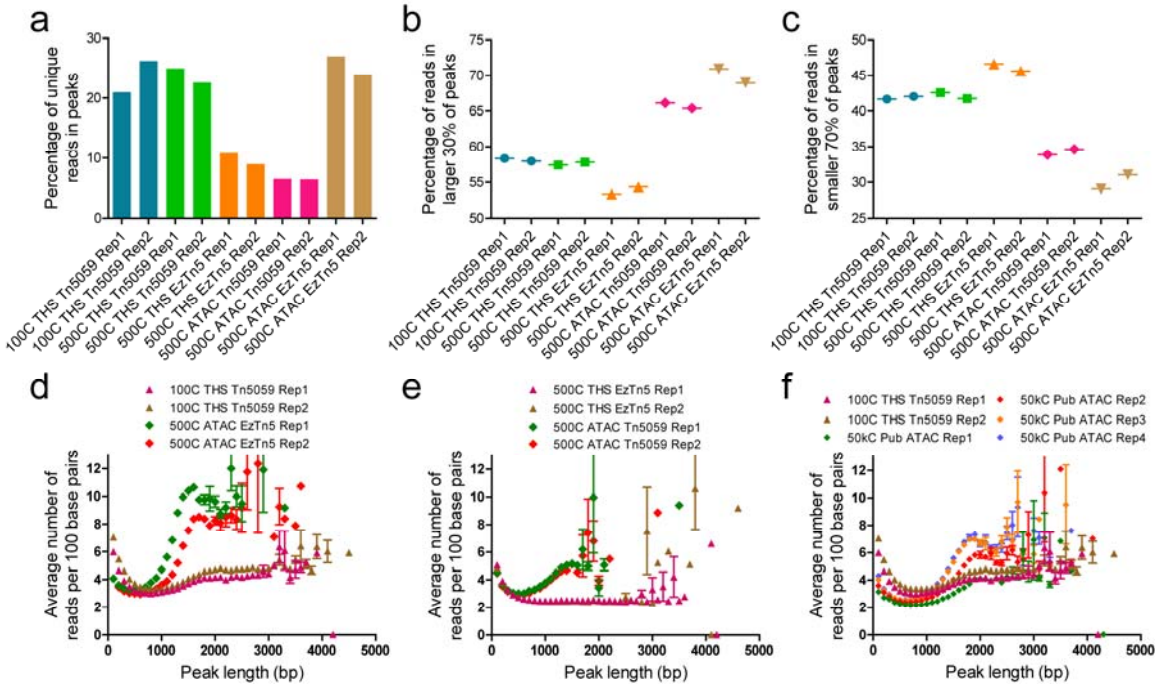


Figure 2.24. **THS-seq and ATAC-seq peak capture preferences and biases.**
All datasets and replicates were down sampled to 8,351,125 unique alignments before analysis. (a) Total percentage of unique alignments in peaks out of the 8,351,125 unique alignments for each dataset. (b) The percentage of alignments that are in the larger 30% of peaks called significant for each individual sample and replicate. (c) The percentage of alignments that are in the smaller 70% of peaks called significant for each individual sample and replicate. (d-f) For all peaks in each individual dataset, the normalized number of alignments in each peak length, with peak lengths in increments of 100 base pairs, represented by mean ± SEM in (d) 100-cell THS-seq/Tn5059 data and 500-cell ATAC-seq/EzTn5 data, (e) 500-cell THS-seq/EzTn5 data and 500-cell ATAC-seq/Tn5059 data, and (f) 100-cell THS-seq/Tn5059 data and published 50,000-cell ATAC-seq data. Some data points were excluded from the graphs because values were beyond the axis, and the number of data points excluded for each graph is: (d) 13, (e) 16, and (f) 8.

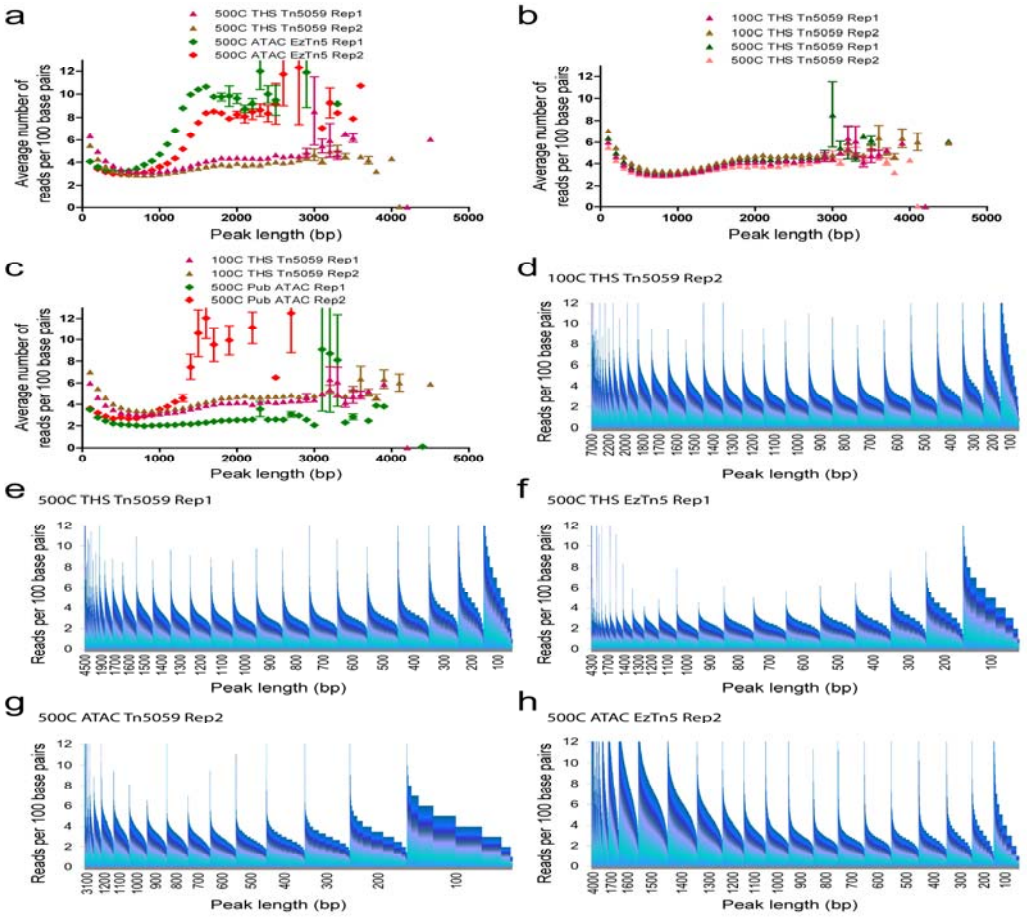


Figure 2.25. THS-seq and ATAC-seq peak capture preferences.
All datasets were down sampled to 8,351,125 unique reads before analysis. Normalization was performed by calculating the number of reads per 100 base pairs in each peak, and any peaks with zero alignments were removed before analysis. (a-c) For all peaks in each individual dataset, the normalized number of alignments in each peak represented by mean ± SEM in (a) 500-cell THS-seq with Tn5059 and 500-cell ATAC-seq with EzTn5 (b) 100-cell and 500-cell THS-seq with Tn5059 (c) 100-cell THS-seq with Tn5059 and published 500-cell ATAC-seq data. Some data points were excluded from the graphs because values were beyond the axis, and the number of data points excluded for each graph is: (a) 13, (b) 4, and (c) 15. (d-h) For all peaks in each individual dataset, the normalized number of unique alignments in each peak, visualized from highest peak length to lowest, and with the most to the least number of normalized unique alignments for each peak length from left to right. Datasets represented had the most base pairs called significant by Dfilter of the two replicates for each condition. Datasets include (d) 100-cell THS-seq with Tn5059 Rep2, (e) 500-cell THS-seq with Tn5059 Rep1, (f) 500-cell THS-seq with EzTn5 Rep1, (g) 500-cell ATAC-seq with Tn5059 Rep2, and (h) 500-cell ATAC-seq with EzTn5 Rep2.

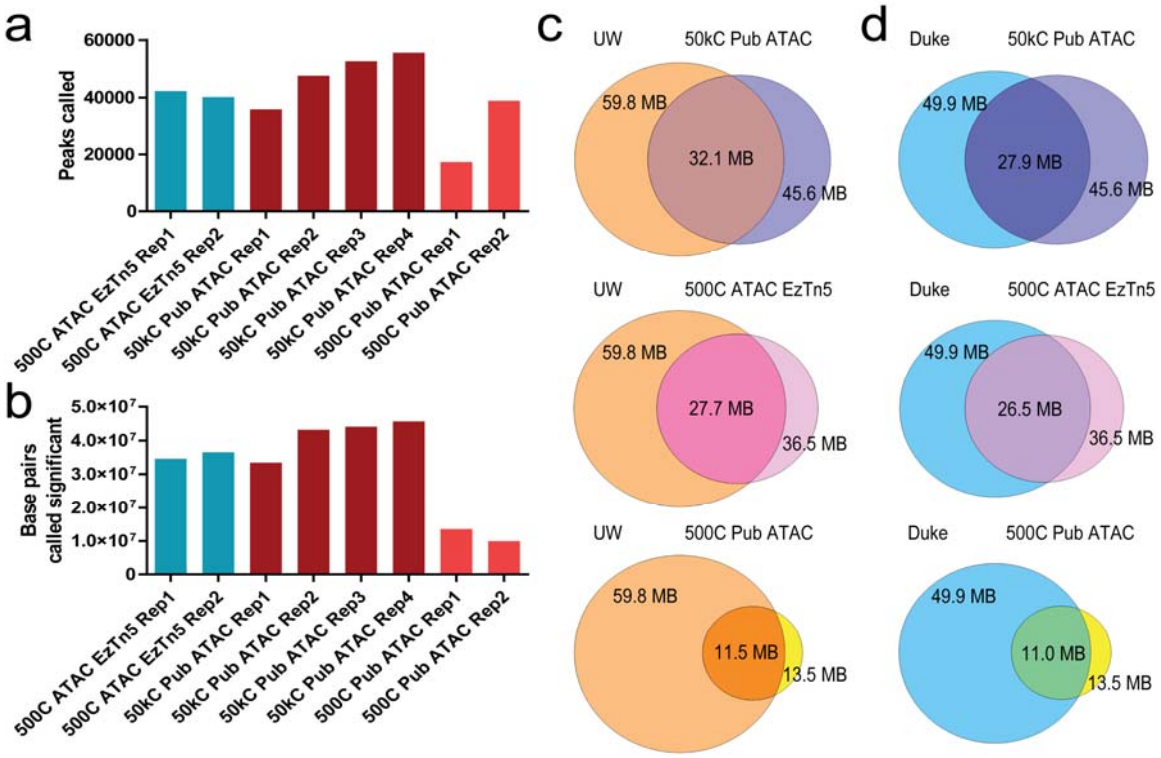


Figure 2.26. Comparison of 500-cell ATAC-seq with EzTn5 data to published ATAC-seq data.
A modified protocol that included an optimized tagmentation buffer and protease digestion of proteins and transposase were used for generating 500-cell ATAC-seq with EzTn5 data, which was then compared to published 50,000-cell ATAC-seq data, and published 500-cell ATAC-seq data. All samples were down sampled to the same number of unique reads, 8,351,125, before analysis. The replicate with the highest number of base pairs called significant was used in the venn diagram representations for each sample and in analysis. Comparisons were made examining (a) the total number of peaks called, (b) total number of base pairs called significant, (c) base pair overlap of published 50,000-cell ATAC-seq data, 500-cell ATAC-seq with EzTn5 data and published 500-cell ATAC-seq data against ENCODE accessibility data from UW and (d) base pair overlap of published 50,000-cell ATAC-seq data, 500-cell ATAC-seq with EzTn5 data and published 500-cell ATAC-seq data against ENCODE accessibility data from Duke.

2.2.6 ATAC-seq improvements

Through optimizations to THS-seq, we inadvertently have made substantial improvements to ATAC-seq/EzTn5. Published 500-cell ATAC-seq data had on average 69% less peaks called, 70% less total base pairs called significant, and 61%-64% less ENCODE base pair overlap compared with published 50,000-cell ATAC-seq data (**Fig. 2.26a-d Table 3**). Upon examination of peak size distributions, published 500-cell ATAC-seq data replicate two had 78% of all peaks in the 100-300 base pair range, thus indicating high background noise, while replicate one had 17,331 peaks called (**Fig. 2.21h, Table 2, Table 3**). When we performed ATAC-seq/EzTn5, two changes were made to the published protocol. First, the tagmentation reaction contained 16% DMF as final concentration in reaction, and second, after tagmentation protease digestion was performed on tagmented DNA and any cellular proteins in solution. These changes were able to generate data from 500 cells that had 131% more peaks called, 169% more base pairs called significant, and 141% more ENCODE data base pair overlap than published 500-cell ATAC-seq data. This in turn generates data of higher quality, identifying more discernable chromatin accessibility peaks than published 500-cell ATAC-seq data (**Fig. 2.21e, g, h, Fig. 2.27**). For 500 cells, our optimized ATAC-seq protocol with EzTn5 produced data at a quality approaching that of published 50,000-cell ATAC-seq data, albeit with 28% less peaks called, 20% less base pairs called significant, and 5-14% less ENCODE base pair overlap (**Fig. 2.26a-d, Table 3**). Together these data illustrate how minor optimizations in protocol can substantially improve ATAC-seq for

assessing chromatin accessibility, however still to a lesser degree than THS-seq/Tn5059.

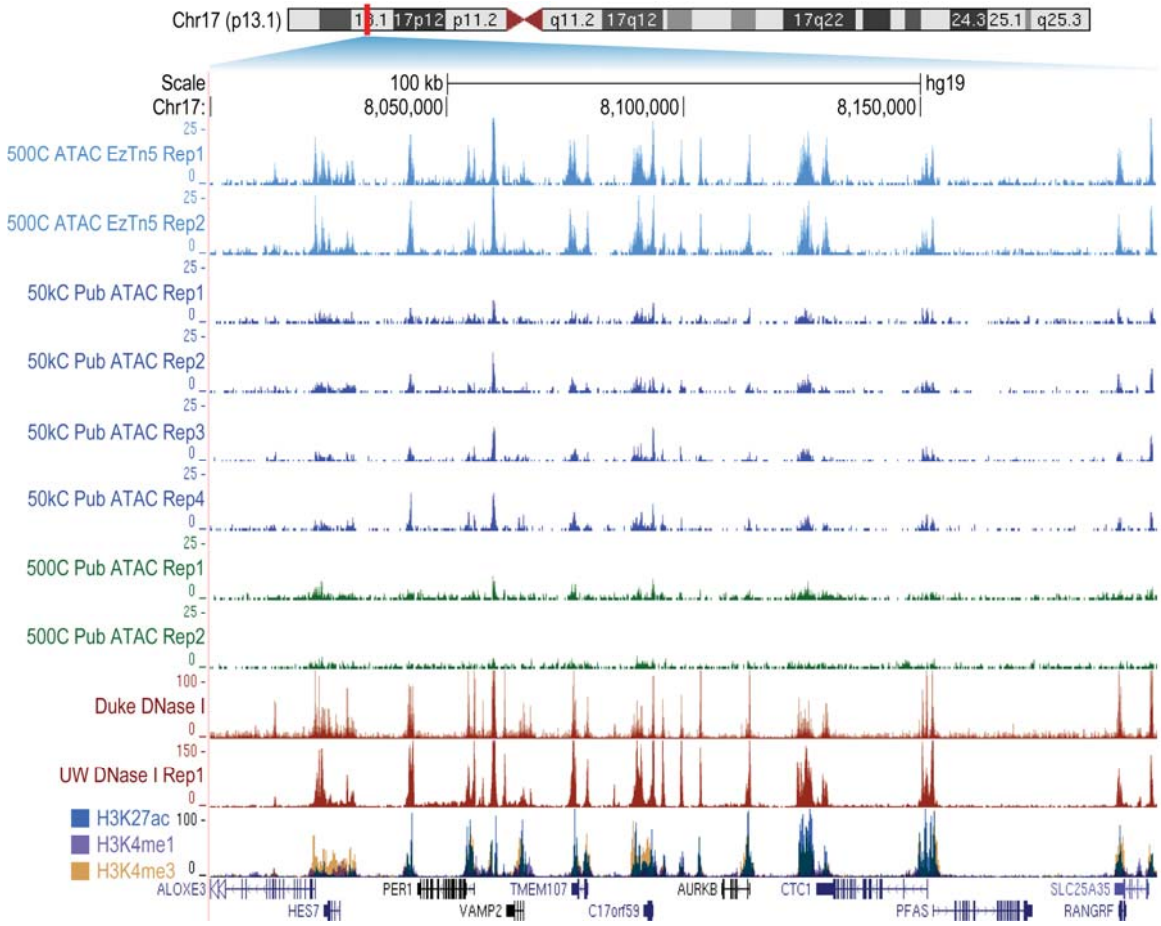


Figure 2.27. 200 kb view of accessible chromatin in ATAC-seq with EzTn5 data versus published ATAC-seq data.
500-cell ATAC-seq with EzTn5 data was compared to published 50,000-cell ATAC-seq data, and published 500-cell ATAC-seq data. 500-cell ATAC-seq with EzTn5 data was generated using modified protocols with optimized tagmentation buffer and protease digestion of proteins and transposase. ENCODE accessibility data and histone modifications which are often found near regulatory elements and promoters are also depicted. All samples were down sampled to the same number of unique reads, 8,351,125, before viewing tracts

2.3 Discussion

In summary, THS-seq achieves a high sensitivity and specificity in detecting accessible chromatin by efficient *in vitro* transposition with a customized transposon, an optimized tagmentation buffer and an engineered Tn5 super-mutant, as well as linear RNA amplification of tagmented DNA molecules. We have validated THS-seq/Tn5059 against ENCODE data, and have shown it captures many peaks ATAC-seq/EzTn5 does not, including a large number of smaller peaks with probable roles in more proximal and distal regulatory regions to transcription start sites. The superior performance of THS-seq with Tn5059 over ATAC-seq with either enzymes, or THS-seq with EzTn5, represents an interesting case where the contributions of individual components are not additive, since changing the method from ATAC-seq to THS-seq while keeping the original EzTn5 provides only a subtle improvement, whereas replacing EzTn5 with Tn5059 is detrimental to ATAC-seq. Previously uncharacterized ATAC-seq bias, where peaks of longer length are preferentially captured, and peaks of shorter length are lost, reduces the comprehensiveness of ATAC-seq generated chromatin accessibility maps. Furthermore, as an unexpected result of this study, an optimized ATAC-seq protocol, is widely applicable to many current and future studies on small numbers of cells or single cells[16]. Although the lowest amount of input material used in this study was 100 cells, THS-seq is compatible with the high throughput combinatorial cell barcoding strategy[15] and when applied to single cells should lead to substantially denser maps of single-cell chromatin accessibility. With THS-seq, we captured

>110,000 insertion events per cell. This compares favorably than what was recently reported with combinatorial ATAC-seq, with a median of 1,685 reads per cell, and scATAC-seq using the C1 device to isolate single cells, with an average of 73,000 fragments per cell of those cells that pass filter[15, 16]. THS-seq is compatible with the combinatorial barcoding scheme, which will facilitate a much more comprehensive characterization of accessible chromatin landscape in tens of thousands of single cells.

Chapter 2, in full is an adapted version of the material that appears in *Genome Biology*. Sos BC, Fung HL, Gao DR, Osothprarop TF, Kia A, He MM, Zhang K. Characterization of chromatin accessibility with a transposome hypersensitive sites sequencing (THS-seq) assay. *Genome Biol.* 2016 Feb 4;17:20. Published. The dissertation author was the primary author of all material.

CHAPTER 3: SINGLE CELL THS-SEQ DEVELOPMENT AND INTEGRATTIVE SINGLE CELL ANALYSIS BY TRANSCRIPTIONAL AND EPIGENETIC STATES IN HUMAN ADULT BRAIN.

3.1 Introduction

Previously it was shown THS-seq is a viable method to assay chromatin accessibility in small number of cells. Direct comparisons with ATAC-seq found higher sensitivity and more robust capture of open chromatin, resulting in more peaks called, smaller peaks being called and more peaks called in distal elements. This required designing a new method ensuring each Tn5 insertion would results in captured fragments. With the improved sensitivity in THS-seq we next wanted to apply the methods to single cells. For this, we adapted THS-seq to combinatorial indexing. After proof of concept and optimization of scTHS-seq, we then applied the method to biological samples. While methods for characterizing chromatin accessibility have been reported[16, 17, 26], they remain to be demonstrated on highly heterogeneous archived human tissues, like the brain, at scale. Here we perform scTHS-seq and snDrop-seq, a new method for quantifying nuclear transcripts, on human archived tissue, and then perform an integrative analysis of single-nucleus transcriptomes and chromatin maps of human tissues.

3.2 Results

3.2.1 snDrop-seq design and application to human brain nuclei

To overcome the limitations associated with single-nucleus RNA sequencing, we have adapted a droplet-based methodology[32] to analyze single nuclei, termed snDrop-seq, that we find can provide a higher-scale assessment of neuronal and non-neuronal diversity in human post-mortem brain samples. The application of this methodology to isolated nuclei necessitated certain modifications (see **Methods**) that included coating all inner surfaces with bovine serum albumen (BSA) to prevent non-specific binding by nuclei, and heat treatment of droplet-encapsulated nuclei to ensure complete lysis without introducing excessive RNA degradation (**Fig. 3.1a**). Subsequently, we applied the snDrop-seq pipeline (**Fig. 3.2A**) to both the human adult visual cortex (Brodmann Area 17 (BA17) or V1) and frontal cortex samples (BA10 and BA6) from four different individuals (**Table 6)**. We generated 16,262 visual cortex single-nucleus data sets and 9,794 frontal cortex data sets, of which 15,819 and 8,755, respectively, were resolved into neuronal and non-neuronal cell types (**Fig. 3.2B-E, Table 7**). These libraries were sequenced to a median of 5,247 usable reads (2,483 for frontal cortex), with the majority of mapped reads falling within intronic regions (**Fig. 3.1D-E**), and predominantly to the 3' ends of transcripts (**Fig. 3.1F**), consistent with poly-A capture and 3'-end counting for not only mRNA transcripts, but also the pre-mRNA abundant in nuclei[48]. In comparison with other RNA-seq methodologies (**Fig. 3.3**), both snDrop-seq and scDrop-seq[32] methods showed highly comparable, albeit lower, median UMI counts and gene detection rates. However, nuclear data was slightly biased for longer genes (**Fig**. **3.3Q-R**), likely reflecting differential transcript processing and

export rates associated with genic length and intron fraction[48]. Overall, we detected a median of 1,072 unique transcripts and 826 genes per visual cortex nucleus (**Fig. 3.3)** and 7,422 total genes detected on average per cell type **(Table 7**). Analysis of transcriptional heterogeneity (see **Methods**) resolved 30 distinct cellular clusters representing 10 excitatory (**Ex**) and 14 inhibitory (**In**) neuronal subtypes, and six non-neuronal cells, including: endothelial cells (**End**), smooth muscle cells or pericytes (**Per**), astrocytes (**Ast**), oligodendrocytes (**Oli**) and their precursor cells (**OPCs**) and microglia (**Mic**) (**Fig. 3.2B, Fig. 3.4**). These subpopulations showed cell-type or subtype specific expression profiles (**Fig. 3.2C, Table 8**), expected marker genes (**Fig. 3.2D**), and were highly consistent with pooled cell populations from the mouse[49] and human (temporal lobe)[50] cerebral cortex (**Fig. 3.2F**). Comparison with single-cell data generated from the mouse visual cortex[51] and human temporal lobe[52] confirmed broad cell type classification and the consistency between nuclear and whole cell data (**Fig. 3.2G**). Neuronal clusters were annotated based on their correlation values with subtypes previously identified from SNS in six cortical regions[53] (**Fig. 3.2H**). In addition to the high correspondence, snDrop-seq permitted finer resolution into sub-populations (e.g. **Ex1** to **Ex1a,b** of the visual cortex) while showing little representation from subtypes not previously observed in these regions (e.g. rostral-specific **Ex2** found only in frontal cortex and caudal-specific **Ex3** found only in visual cortex)[53]. Otherwise, subtypes resolved were found to be highly consistent between these two cortical regions (**Fig. 3.2I**). This demonstrates the high accuracy of snDrop-seq in resolving neuronal subtype diversity in the

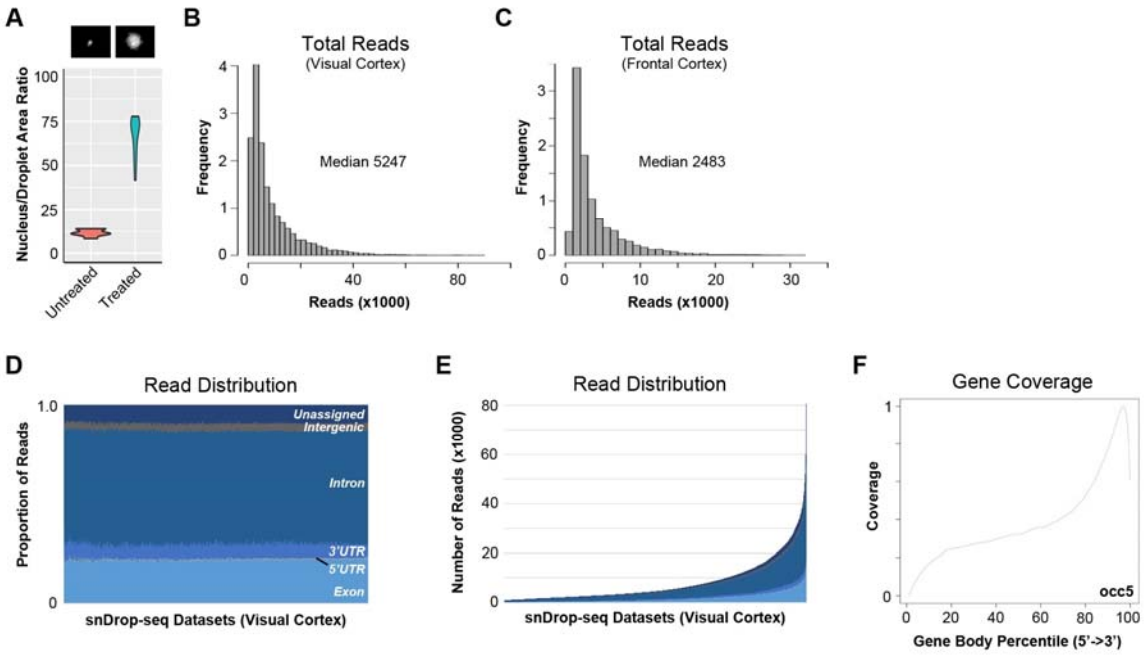


Figure 3.1. snDrop-seq metrics.
A. Violin plot showing the ratio of area size of DAPI staining versus droplet size of untreated and heat treated samples, confirming efficient nuclear lysis in snDrop-seq. Left panel: 293T cells pre-stained with DAPI showing cellular, but not nuclear lysis inside droplets. Right panel: DAPI staining in droplets after applying heat-treatment showing efficient lysis as evidenced by the overspreading of DNA stain inside the droplets. B. Histogram showing frequency of total reads (unmapped, multiple mapped and uniquely mapped) associated with all quality filtered data sets from the visual cortex. C. Histogram showing frequency of total reads associated with all quality filtered data sets from the frontal cortex. D. Proportion of reads mapped to exons, 5' untranslated regions (UTR) 3' UTRs, introns, intergenic regions and that are unassigned across all quality filtered data sets from the visual cortex. E. Number of reads associated with (D). F. Gene body coverage for visual cortex data sets (library occ5).

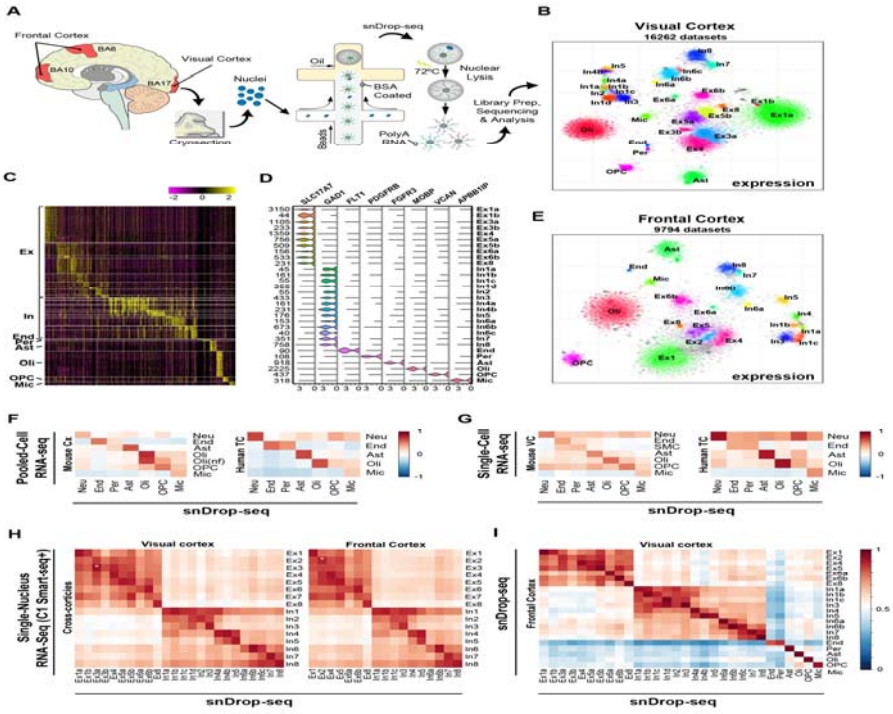


Figure 3.2. snDrop-seq metrics. Transcriptional profiling with snDrop-seq resolves major cell types of the visual and frontal cortices.
A. Overview of single nucleus isolation from the visual cortex (occipital lobe, BA17) and frontal cortex (BA6 and BA10) and subsequent snDrop-seq procedure. B. Single nucleus visual cortex data sets (16,262) showed distinct clustering, visualized here using t-distributed Stochastic Neighbor Embedding (t-SNE). Unannotated data sets are shown in gray. C. Heatmap of expression of top differentially expressed genes identified across cell types and subtypes of the visual cortex (Table 8). D. Violin plots of expression values for type-specific marker genes for the visual cortex. Number of data sets contributing to each cluster is listed. E. Distinct clustering of the frontal cortex data sets using t-SNE. Unannotated data sets are shown in gray. F. Correlation heatmaps comparing averaged visual cortex snDrop-seq data with average expression values from RNA-seq data from: mouse pooled cortical cell types (*12*) (left, nf = newly formed); human pooled temporal lobe cell types (*13*) (right). G. Correlation heatmaps comparing average visual cortex data with average expression values from single-cell RNA-seq data from the mouse visual cortex (*14*) (left) and human temporal lobe (*15*) (right). H. Correlation heatmap of snDrop-seq-identified neuronal subtypes (visual cortex or frontal cortex) compared with subtypes previously identified using the C1 single nucleus Smart-seq+ pipeline (SNS, across cortical regions) (*7*). Star indicates distinct region-specific Ex2 and Ex3 sub-populations. I. Correlation of average expression values for cell types and subtypes resolved from the visual and frontal cortices.

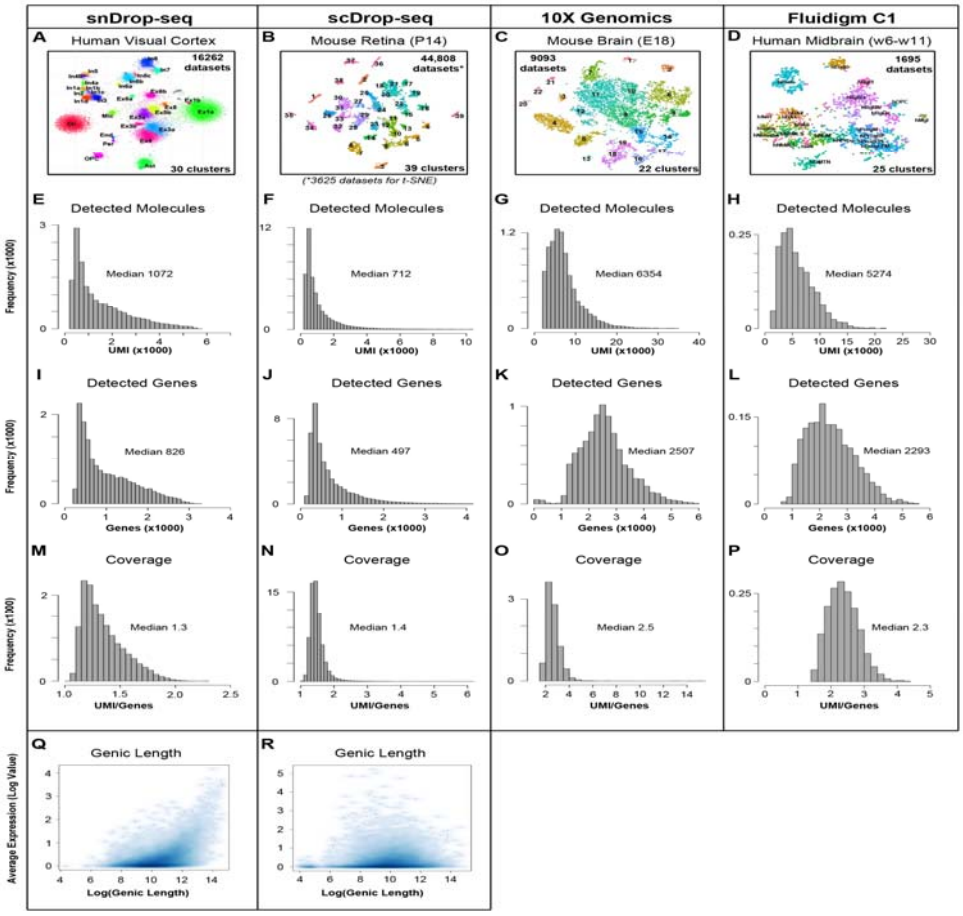


Figure 3.3. Comparison of single-cell/nucleus RNA-seq methodologies. A. snDrop-seq clusters from the human adult visual cortex visualized by t-SNE (from Fig. 1B). B. Mouse retina data generated from single cells (sc) using the Drop-seq method and previously annotated[54] were visualized using a maximum of 100 cells per cluster by t-SNE. C. Mouse embryonic brain (E18) data sets generated on the 10X Genomics platform were clustered and visualized by t-SNE. D. Embryonic human midbrain data sets generated on the Fluidigm C1 platform and previously annotated[55] were visualized by t-SNE. E-H. Histograms of detected molecules (number of UMI) for associated quality filtered data sets. I-L. Histograms of detected genes for associated quality filtered data sets. M-P. Coverage (number of UMI per gene detected) for associated quality filtered data sets. Q. Scatter plot showing average expression values for protein-coding genes across all quality filtered snDrop-seq samples (log transformed) and their associated genic length. R. Scatter plot showing average expression values for protein-coding genes across all quality filtered scDrop-seq samples (log transformed) and their associated genic length.

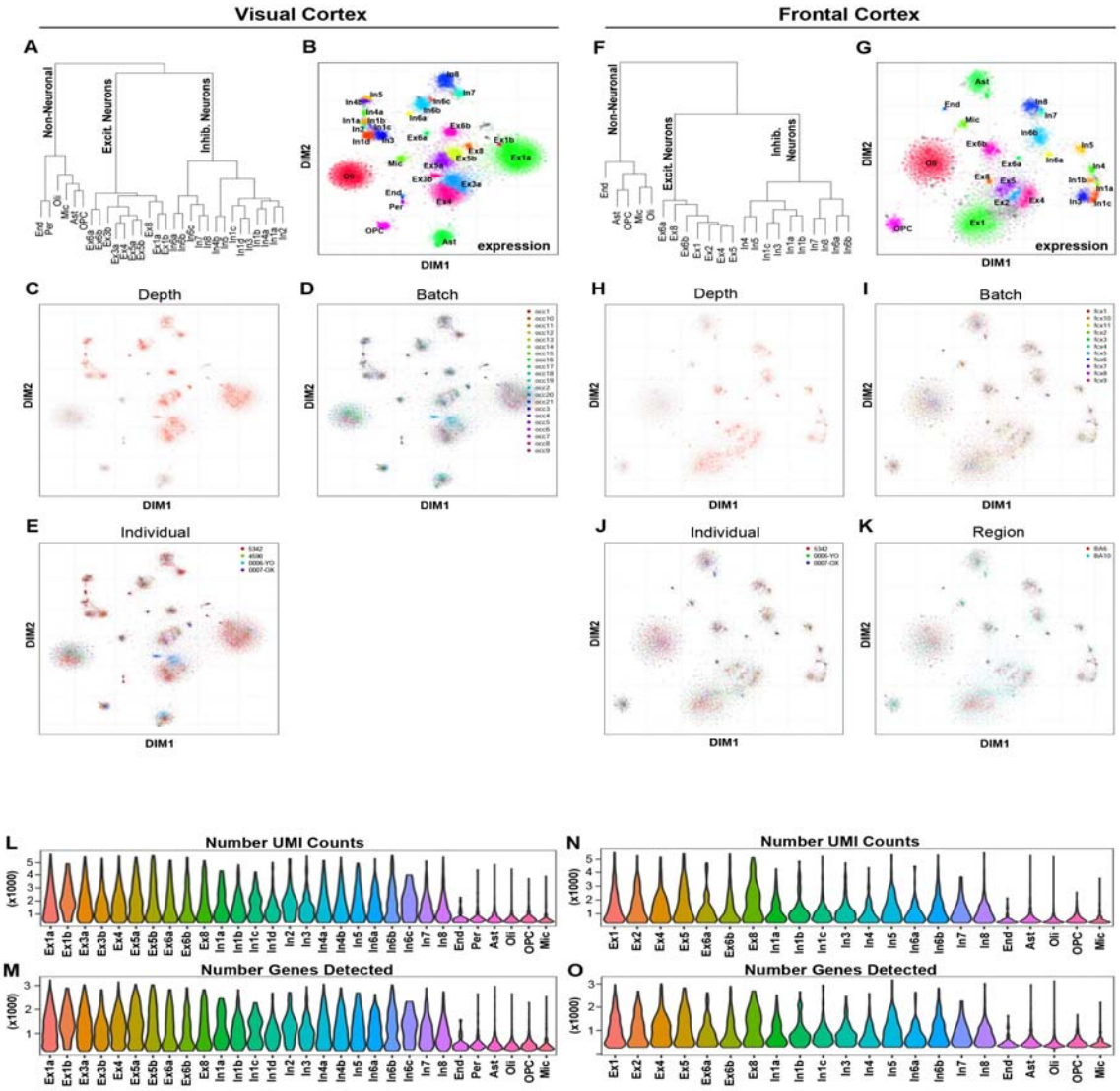


Figure 3.4. snDrop-seq resolves neuronal and non-neuronal subpopulations.

A. Cluster dendrogram showing unsupervised grouping of visual cortex non-neuronal, excitatory and inhibitory neuronal cell types or subtypes. B. t-SNE plot as shown in Fig. 1B. C. t-SNE plot as in (B) showing UMI read depth. Non-neuronal cell types show lower coverage than neuronal subtypes. D. t-SNE plot as in (B) showing batch identity (see Table 6) and the absence of any significant batch-dependent drivers for nuclear data clustering. E. t-SNE plot as in (B) showing individual patient identity (see Table 6). F-J. Same as for B-E except showing frontal cortex data. K. t-SNE plot as in (G) showing BA region (see Table 6). L. Total number of UMI counts for each cell type or subtype from the visual cortex. M. Total number of genes detected for each cell type or subtype from the visual cortex. N-O. Same as for L-M except instead showing data from the frontal cortex.

cerebral cortex though profiling a larger cell cohort compared with our previous SNS efforts, albeit with ~1500-fold lower per cell sequencing depth.

Excitatory neurons of the visual cortex marked by expression of *SLC17A7* (**Fig. 3.2D**) were resolved into 10 distinct subtypes, showing enriched marker gene profiles (**Fig. 3.5A, Table 9**) and that could be distinguished by their spatial orientation within the cortex[56] (**Fig. 3.6A, B**). In addition, we identified specific subpopulations located within cortical layers, including the *CARTPT*$^{+}$*RASGRF2*$^{+}$ **Ex1b** of upper layer 2; the distinct *HS3ST5*$^{+}$*PCP4*$^{-}$ (**Ex5b**) and *HTR2C*$^{+}$*PCP4*$^{+}$*TLE4*$^{+}$ (**Ex6a**) subpopulations in layer 5, the latter bordering on a *HTR2C*$^{-}$*TLE4*$^{+}$ (**Ex6b**) layer 6 population; and the *NR4A2*$^{+}$*SYNPR*$^{+}$ **Ex8** subpopulation in layer 6b (**Fig. 3.6B**). Inhibitory neurons of the VC, marked by shared expression of *GAD1* (**Fig. 3.2D**), were resolved into 14 subtypes showing enriched marker gene expression (**Fig. 3.5B, Table 10**), distinct profiles of canonical interneuron markers (e.g. *VIP, RELN, PVALB, SST*) as well as sub-type restricted expression (e.g. *SHISA8, CA3, CA8*) (**Fig. 3.6C**). We were further able to resolve spatially distinct inhibitory neuron subpopulations, including: layer 1 *RELN*$^{+}$*CCK*$^{+}$*CNR1*$^{+}$ **In1a**; upper layer *VIP*$^{+}$*CALB2*$^{+}$*TAC3*$^{+}$ **In1d**; *PVALB*$^{+}$*CA8*$^{+}$ **In6a** concentrated around layer 4, as well as the more peripheral *PVALB*$^{+}$*TAC1*$^{+}$ **In6b**; and two distinct SST positive subtypes, including the upper layer *SST*$^{+}$*CALB1*$^{+}$ (**In7**) and lower layer *SST*$^{+}$*CALB1*$^{-}$ (**In8**) subpopulations (**Fig. 3.6D**). Furthermore, given the strand-specificity of snDrop-Seq, we characterized the expression patterns of natural antisense transcripts (NATs), which was not possible with SNS[53]. We identified 26 NATs to be differentially expressed across

cell clusters (**Table 12, Fig. 3.7**), including inhibitory neuron-specific expression of DLX6-AS1 potentially involved in interneuron specification during development[57, 58] (**Fig. 3.7B**). Furthermore, we detect a higher proportion of neuronal data sets in the visual over frontal cortex consistent with a higher neuronal density in this region (**Fig. 3.6E**)[59]. Cell type proportions detected by snDrop-seq were somewhat consistent with those from single-cell analyses on the temporal cortex, with the exception of astrocytes and endothelial cells which appear under-represented. This may reflect a bias in sample processing or a potential limited detection based on lower transcript levels for these cell types (**Fig. 3.4L-O**). Therefore, snDrop-seq provides not only a more comprehensive cellular profile of human postmortem tissues (**Fig. 3.6E**), but also insights into region-specific cell-type proportion differences and strand-specific transcriptomic dynamics, attributes not achievable using previous methodologies[53].

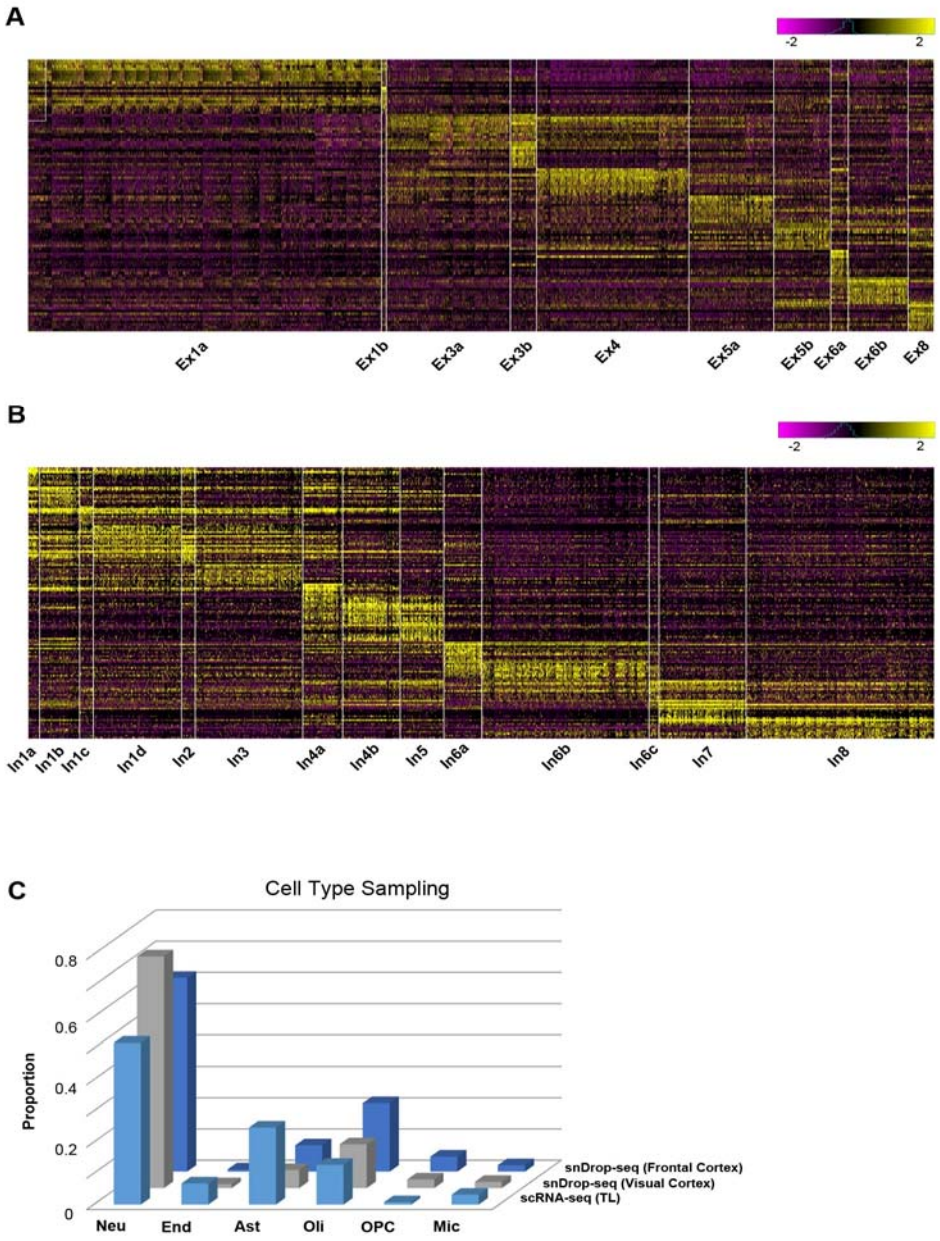


Figure 3.5. Subtype-specific gene enrichments and tissue-specific proportions.
A. Heatmap of top 10 differentially expressed protein-coding marker genes enriched across excitatory neuron subtypes (Table 9). B. Heatmap of the top 10 differentially expressed protein-coding marker genes enriched across inhibitory neuron subtypes (Table 10). C. Cell type sampling rates from single-cell RNA-seq on the temporal lobe[52] compared with snDrop-seq sampling rates from the visual and frontal cortices.

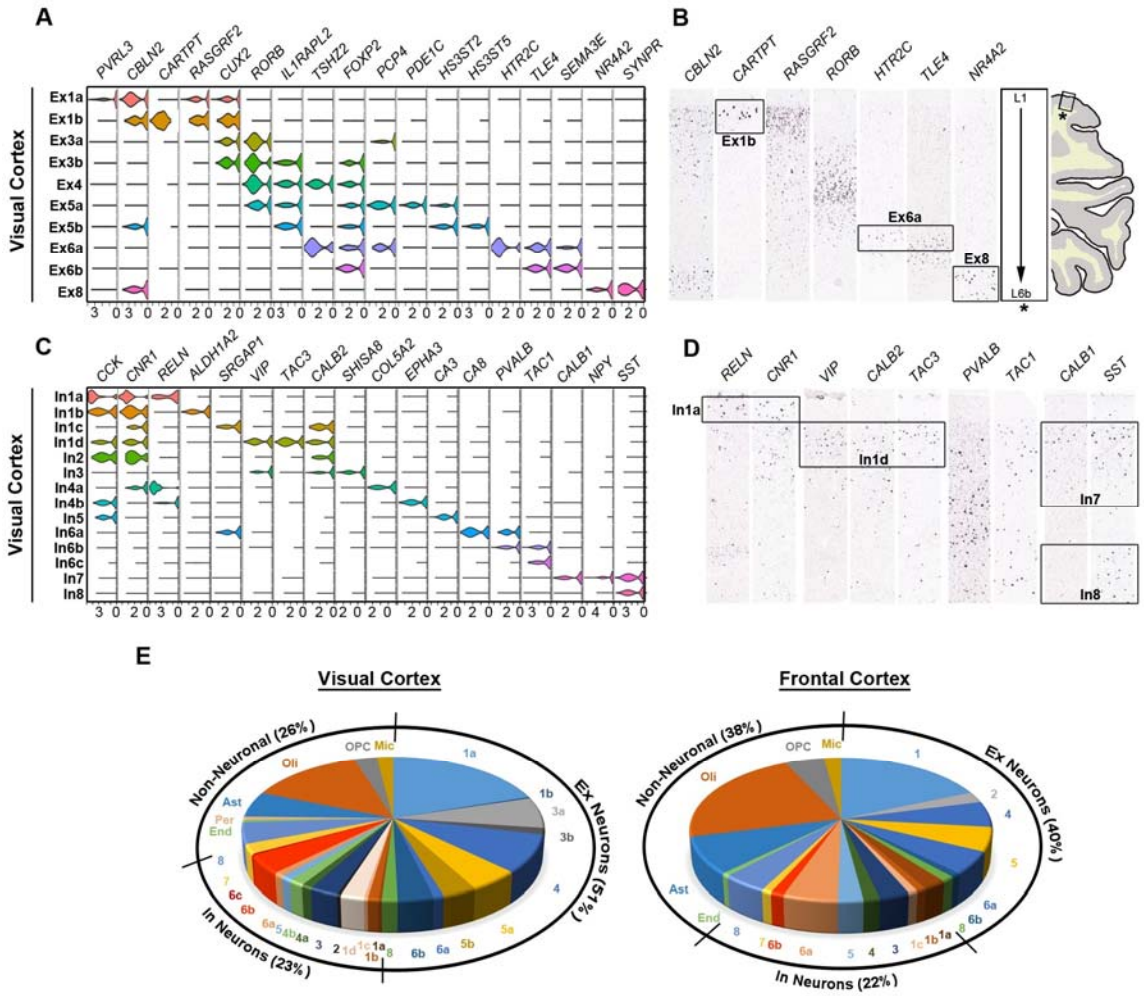


Figure 3.6. snDrop-seq identifies molecularly and spatially distinct neuronal subtypes.

A. Violin plots showing gene expression values of layer specific (*7, 16*) and subtype-enriched markers for excitatory neuronal subtypes. B. RNA *in situ* hybridization (ISH) stains (Table 11) of the visual cortex for select marker genes shown in (A). C. Violin plots showing expression values for classical interneuron marker genes (*7*) and subtype-enriched transcripts. D. RNA ISH stains (Table 11) of the visual cortex showing select stains for markers in (C) and predicted spatial distribution of associated inhibitory neuron subtypes. E. Proportion of cell types and subtypes detected from snDrop-seq interrogation of the human visual cortex compared with the frontal cortex.

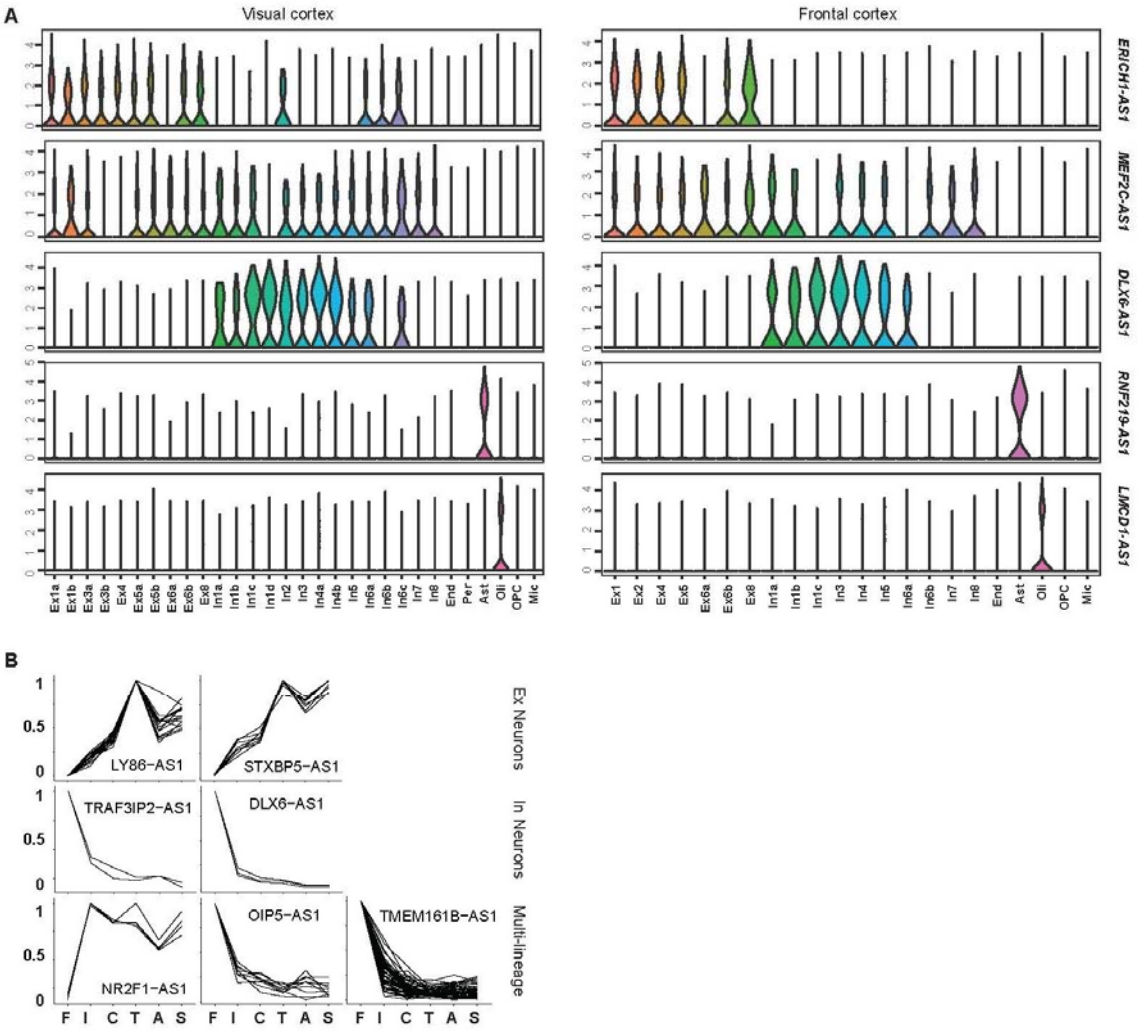


Figure 3.7. Cell-type-specific expression of NAT markers and stage-specific expression of NAT and protein-coding markers.
A. Violin plots of expression values for excitatory neuronal, pan-neuronal, inhibitory neuronal, astrocyte and oligodendrocyte NAT marker genes. B. The expression of 7 differentially expressed NAT marker genes in brains from different ages. Each curve represents one differential expression region of NAT gene. All values are normalized to the highest expression level of each region among all development stages. F, fetal; I, Infant; C, child; T, teen; A, adult; S, senior or age ≥ 50.

3.2.2 scTHS-seq design and application to human brain nuclei

The epigenetic state of a cell can provide valuable information about its overall phenotypic potential, and reveal gene regulatory processes responsible for establishing and maintaining its function. To investigate the epigenetic configuration of different cells in the human adult cortex at scale, we developed a single-cell DNA accessibility assay by integrating THS-Seq[60], which uses *in vitro* transcription and an engineered super-mutant of Tn5 transposase[61] to achieve higher sensitivity and better coverage of distal enhancers than ATAC-seq[5], with combinatorial cellular indexing[26] based on a set of 384 customized barcoded transposomes (**Fig. 3.8A, Fig. 3.9**). scTHS-seq, confirmed to have a low doublet rate and generating high quality data (**Fig. 3.10, Fig. 3.11, Table 13**), identified 287,381 peaks associated with DNA accessibility regions in combined data sets, covering 144 million base pairs with unique genomic alignments, at the clonal rate of ~60% and doublet rate of 11.7%. In total, we generated 14,870 quality-filtered single-nucleus data sets having a median of 11,889 unique reads per cell that, from comparison with merged datasets, were confirmed to detect accessible regions (**Fig. 3.12, Fig 3.13**). To identify epigenetically-distinct subpopulations within this scTHS-seq dataset, we first used an unbiased approach. We modeled the probability of observing reads from a genomic site in a given cell as a censored Poisson process (see **Methods**), which accounts for the fact that the scTHS-seq signal from even the most accessible site will saturate after only a few reads. This approach revealed multiple distinct subpopulations in both the visual and frontal cortex (**Fig. 3.8B-D, Fig. 3.14**), with proportions that are more

favorable to non-neuronal cell types than snDrop-seq **(Fig. 3.6E)**, suggesting that snDrop-Seq is less efficient in sampling non-neuronal cells, either due to data quality filtering that is biased against cells transcribing fewer transcripts or lower efficiency in packaging smaller nuclei into droplets. Characterizing the identity of epigenetically-defined subpopulations, however, is more challenging than in the case of transcriptionally-defined subsets, as functional roles of most regulatory sequences remains poorly annotated. Based on the functional annotation of the genes neighboring differentially-accessible sites, we could distinguish five glial cell populations and three neuronal cell populations having 97,672 differential accessibility peaks (**Table 13**) that when aggregated into accessibility profiles could be used to identify putative regulatory regions within a given locus (**Fig. 3.8D**).

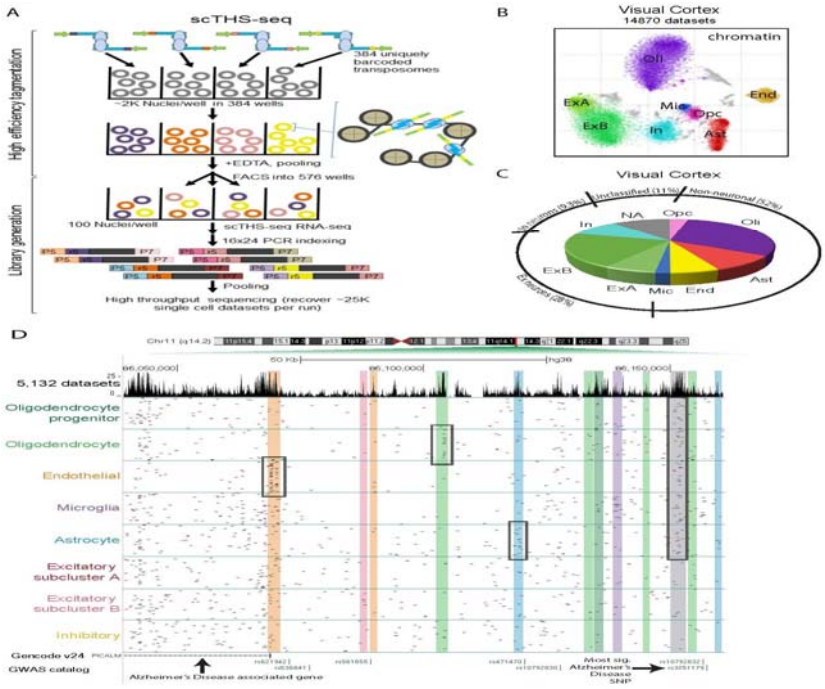


Figure 3.8. Chromatin accessibility profiling resolves major cell types of the visual cortex.

A. Implementation of scTHS-seq with a Tn5 supermutant and cellular combinatorial indexing. First, 384 uniquely barcoded transposomes are added to 500,000 to 1 million mixed mouse and human nuclei that have been aliquoted into a 384 well plate, with one uniquely barcoded transposon added per well. Species nuclei are mixed to reduce and calculate collision rates, and for quality control. Next, high efficiency tagmentation is performed, followed by EDTA inactivation, pooling, and FACS redistribution of 100 nuclei/well into 96 well plates. scTHS-seq RNA-seq is performed with each well treated as one reaction, followed by PCR indexing, pooling and high throughput sequencing of libraries. Optimally ~25,000 total single cell datasets are recovered per run, the sum of mouse and human nuclei recovered. B. tSNE plot of major subpopulations of visual cortex nuclei, clustered based on the overall similarity of the chromatin accessibility profiles. C. Pie chart depicting major visual cortex subpopulations identified in B. D. Identification of cell specific DNA accessibility peaks over the promoter and regulatory region of Alzheimer's disease associated gene PICALM. Glial specific peaks are over-represented. The top tract is all cells merged to generate peaks. Each cell subpopulation tract is represented by 100 randomly selected single cells of cells that had reads in the depicted region, where each row represents a single cell and each dot is a read. The color of highlighted peak regions corresponds to the cell type specificity of the peak, with each subpopulation tract title a specific color. Boxed regions highlight specific cell type specific peaks. The gray box highlights a glial cell specific peak. Nonspecific peaks are not highlighted.

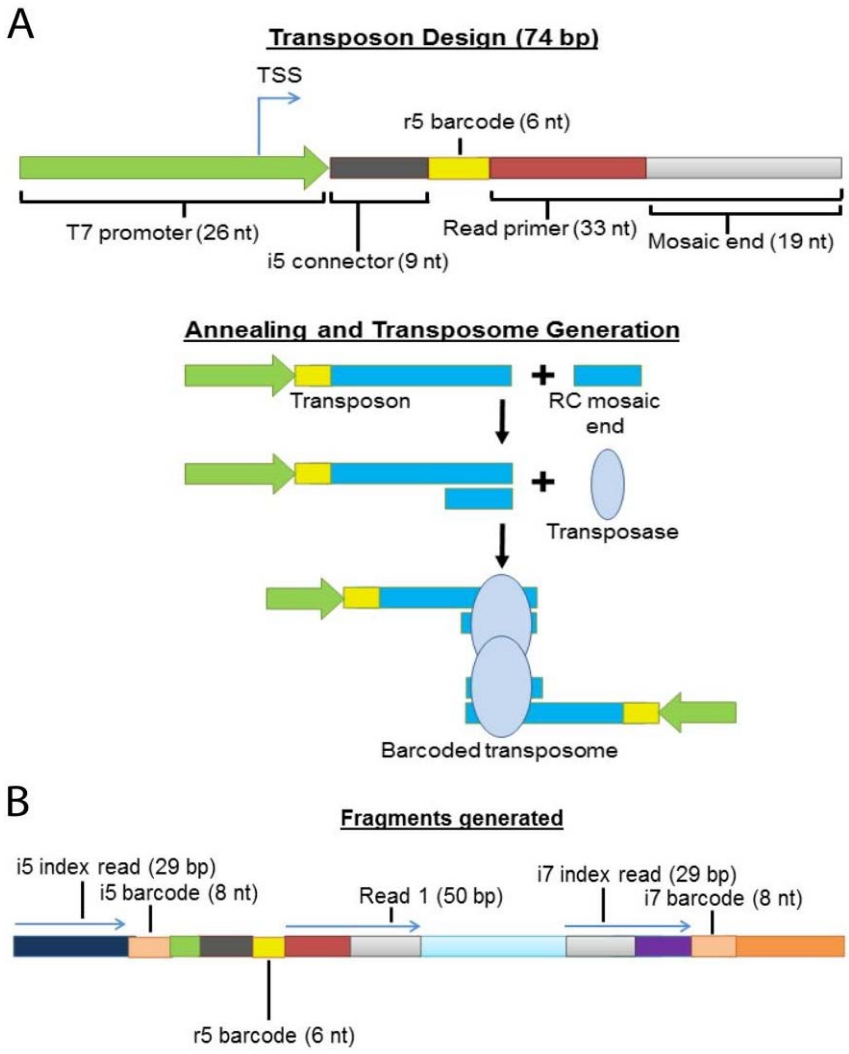


Figure 3.9. Customized scTHS-seq transposon design and sequencing library fragments generated.
A. The design of the transposon, which contains a mosaic end for Tn5 binding, read primer sequence for sequencing, the r5 barcode which is unique to every transposon, an i5 connector sequence used in PCR amplification and a T7 promoter for *in vitro* transcription amplification. To generate barcoded transposome complexes, first transposon is annealed to the RC mosaic end, then transposase is added. B. Depiction of sequencing ready fragments generated after library generation. Fragments contain i5/i7 adaptors and barcodes, the r5 transposon barcode, Read1 sequencing primer, and captured genomic DNA. The i5 and r5 barcodes are read in one sequencing read, with sequencing progressing through the linker region and ending with reading the r5 barcode.

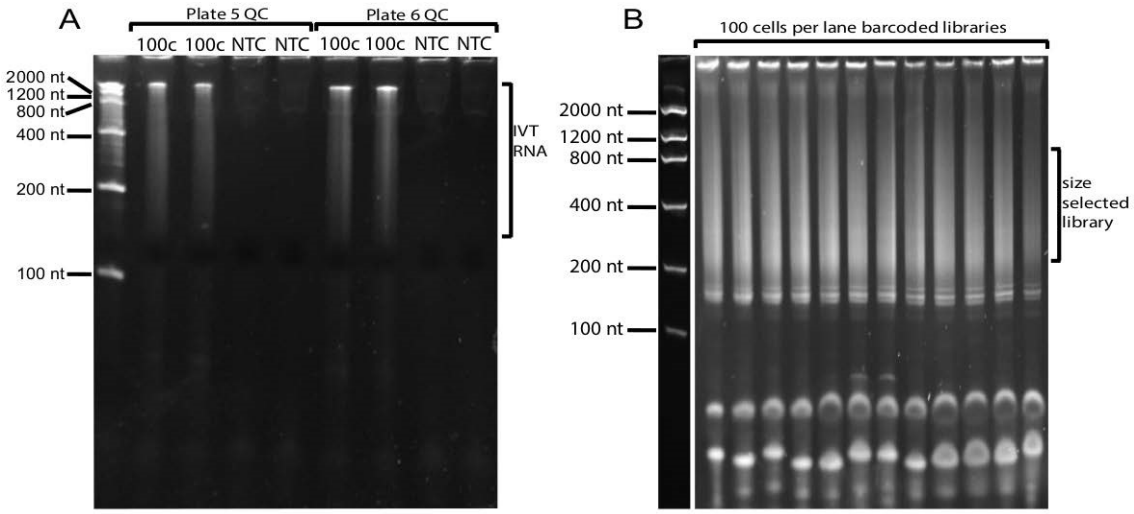


Figure 3.10. Confirmation of *in vitro* transcription and barcoded library generation in scTHS-seq.
A. Whole reactions were purified with SPRI binding buffer, since beads were already in solution. Then whole reactions were run on a 6% TBU gel and visualized with a UV-transilluminator. Plate 5 and plate 6 quality control (QC) samples consisted two 100 cell samples, and two NTC samples. Cells were FACS sorted, and NTCs consisted of only PBS with no FACS sorting. B. Each 96 well plate had 12-6% TBE gels run for validation of library generation with visualization on a UV-transilluminator. Shown is one example of 12 wells from a 96 well plate, where we can see library amplification in the expected size range that is size selected after pooling all samples.

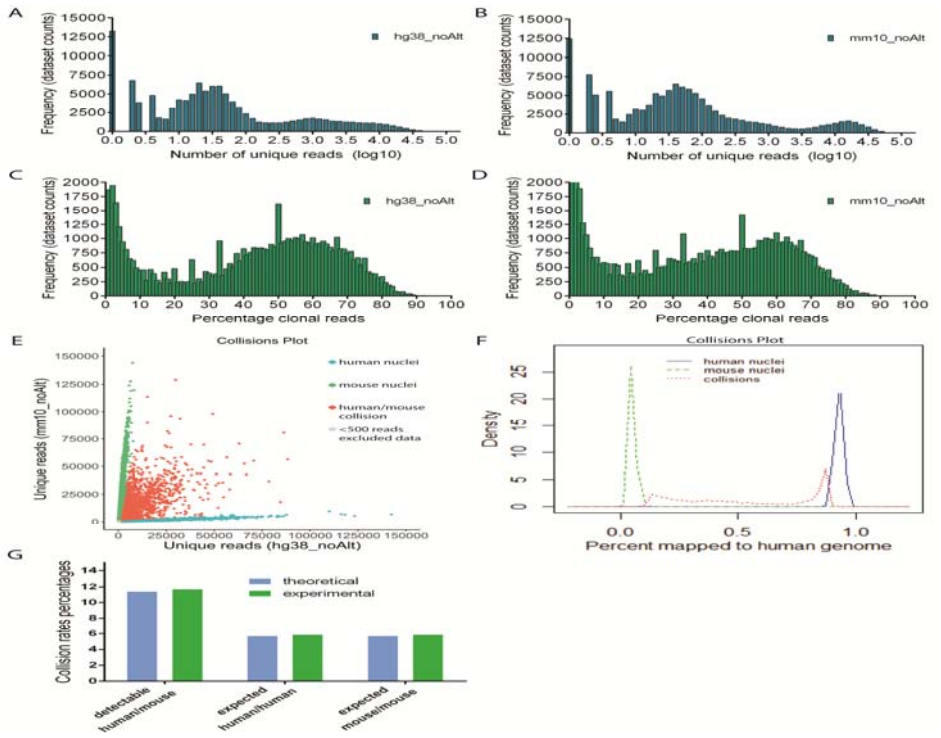


Figure 3.11. Sequencing statistics and collisions calculations for scTHS-seq.

A. Frequency of the log10 number of unique reads in each possible barcode combination in the hg38 noAlt dataset. B. Frequency of the log10 number of unique reads in each possible barcode combination in the mm10 noAlt dataset. C. Frequency of the number of percentage clonal reads in each possible barcode combination in the hg38 noAltdataset. Clonal read statistics were rounded to the nearest whole percent before calculating. D. Frequency of the number of percentage clonal reads in each possible barcode combination in the mm10 noAlt dataset. Clonal read statistics were rounded to the nearest whole percent before calculating. E. Collisions plot examining the ratio of unique reads mapped to human and unique reads mapped to mouse in each unique barcode combination. Plot of human nuclei in blue, mouse nuclei in green, human mouse collision nuclei in red, and <500 reads excluded data. See Methods for determination of which data sets are considered collisions. F. Density plot of the ratio of unique reads mapped to human and unique reads mapped to mouse in each unique barcode combination. Human nuclei are in blue, mouse nuclei are in a green dotted line, and collisions in a red dotted line. <500 reads excluded data was excluded in this analysis. G. Theoretical and experimentally derived collision rate percentages. Graphed are the detectable human/mouse collision rate, expected human/human collision rate, and expected mouse/mouse collision rate. The expected collision rates cannot be directly measured since they are the same genome and would appear as only mouse nuclei or only human nuclei on a collisions plot.

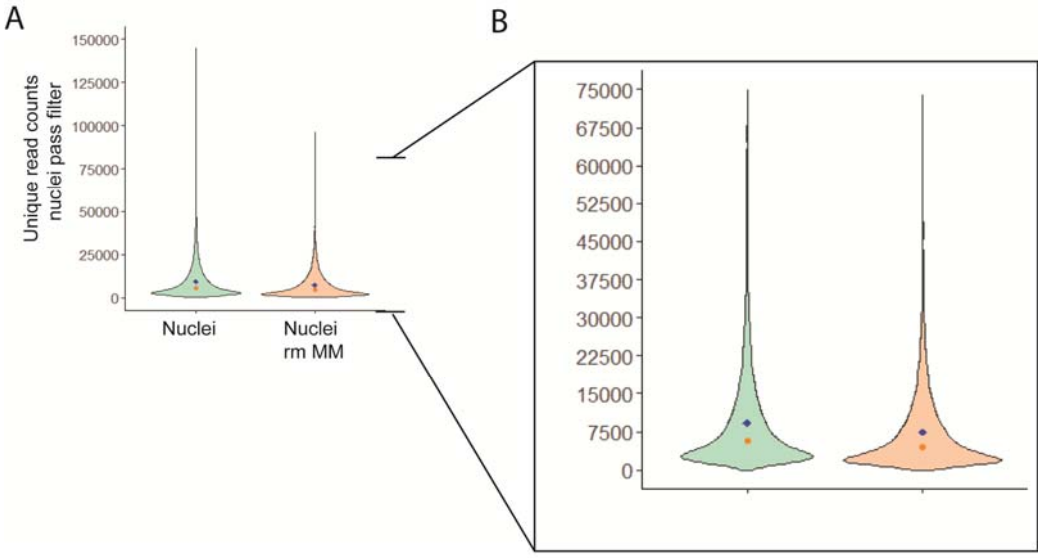


Figure 3.12. scTHS-seq Unique read statistics.
A. Violin plot of unique read counts for nuclei that pass filter and nuclei that pass filter with multi-mappers removed. The blue dot in the middle of the violin plot is the average number of unique reads and the red dot in the middle of the violin plot is the median number of unique reads. B. Zoom in graphic from 0 to 80,000 unique read counts for nuclei that pass filter.

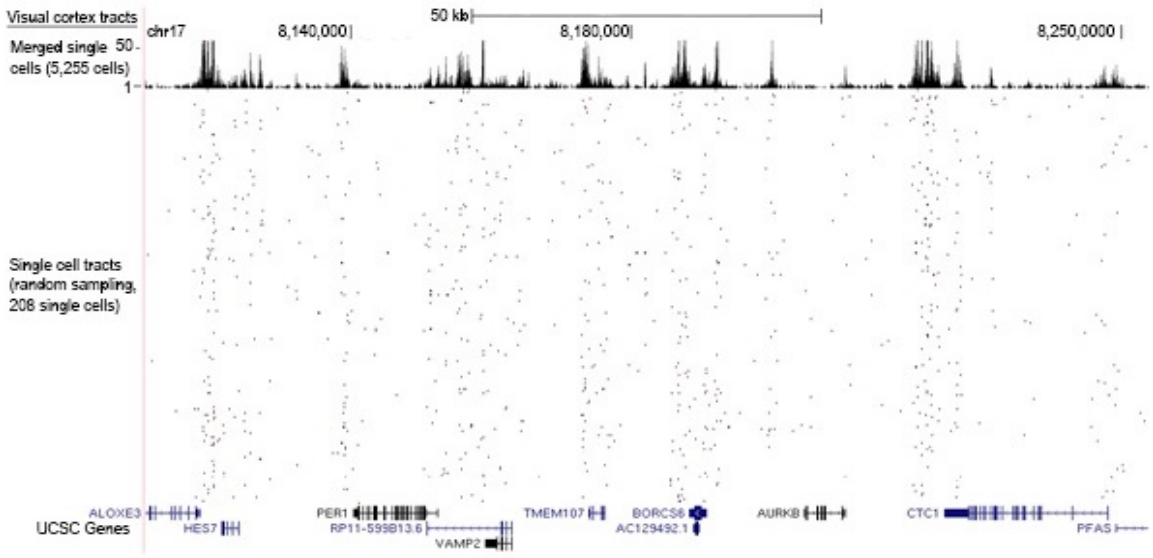


Figure 3.13. Visualization of merged single cell data and individual single cell datasets.
All unique reads in single cell datasets that passed filter when multi-mappers were removed were merged to generate the merged tract. Single cell datasets that had reads present in the chromosome 17 from 8,100,000 to 8,250,000 base pairs locus were randomly sampled and plotted in the single cell tracts section. Each black dot represents a unique read. UCSC genes are overlaid in the on the bottom tract.

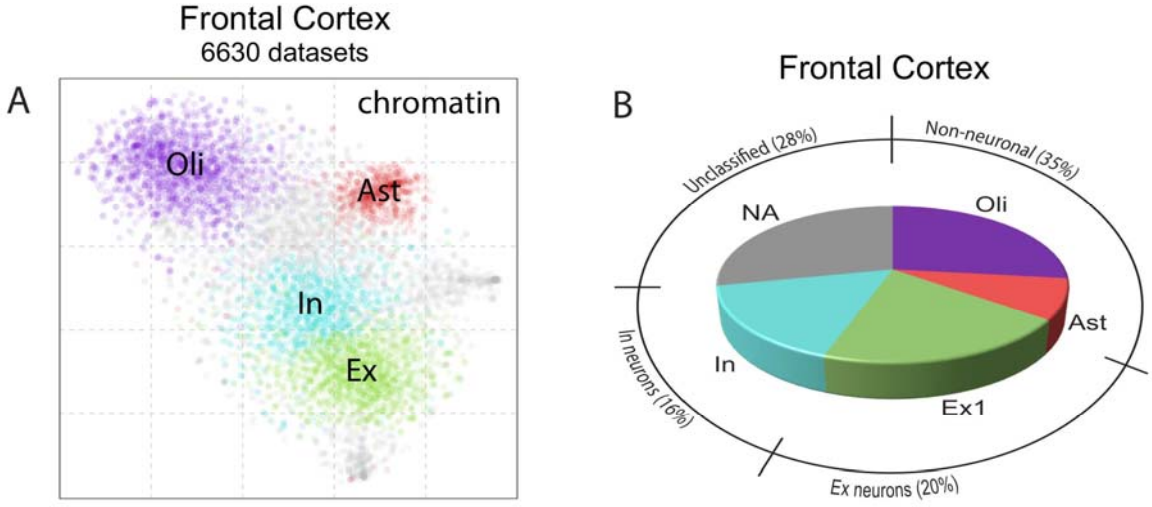


Figure 3.14. Epigenetic heterogeneity in the frontal cortex, as measured by scTHS-seq.
A. The tSNE embedding showing the classified sub populations. B. Pie chart visualizing the distribution of cell types.

3.2.3 scTHS-seq and snDrop-seq data integration

To establish more precise correspondence between transcriptional and epigenetic states of different subpopulations, we sought to identify cells corresponding to transcriptional subpopulations in the scTHS-seq data as well as cells corresponding to epigenetic subpopulations in the snDrop-seq data. We trained a gradient boosting model (GBM) to predict differentially accessible genomic sites based on the differential expression patterns (**Fig. 3.15A, Fig. 3.16**) and a separate GBM to predict differential expression based on differential accessibility (**Fig. 3.16, Fig. 3.17**) using features such as the distance of a site to a gene, degree of differential expression or accessibility of the site or gene, among others (see **Methods**). While the ability to predict differential expression or differential accessibility of any individual gene or site is limited (**Fig. 3.16B, E**), joint consideration of multiple genes or sites allows for confident cell type classifications (**Fig. 3.15B, C, Fig. 3.16C, F, Fig. 3.17B, C**). Applying such models to classify scTHS-seq cells based on the differentially expressed genes between major cell types, we were able to resolve astrocytes, oligodendrocytes, inhibitory and excitatory neurons with high sensitivity and precision in both the visual cortex (**Fig. 3.15B**) and frontal cortex (**Fig. 3.17D**). Similarly, applying such models to classify snDrop-seq cells based on the differentially accessible sites between major cell types, we were likewise able to resolve these cell types with high sensitivity and precision in both the visual cortex (**Fig. 3.17B, C**) and frontal cortex (**Fig. 3.17E**). Although layer 4 excitatory neurons (L4 = **Ex2-4**) were not distinguishable from layer 5 and 6 excitatory neurons (L5/6 = **Ex5-8**) from an

unbiased analysis of scTHS-seq data alone (**ExB in Fig 3.8B**), integrating differential expression information from the higher resolution snDrop-seq data allowed us to identify relevant differentially accessible sites to further partition scTHS-seq clusters (**Fig. 3.15D**). Similarly, such integrated analysis allowed us to identify epigenetic differences relevant to inhibitory neuron subtypes that appear distinguished by their developmental origin from subcortical regions of the medial or lateral/caudal ganglionic eminences[53, 62, 63] (**Fig. 3.15E**). Thus, despite lower intrinsic cell type resolution of accessibility data compared to transcription, computational integration of both scTHS-seq and snDrop-seq allowed us to reconstruct detailed epigenetic profiles of fine-grained cell types within the brain, enabling investigations of the regulatory processes active within each cell type.

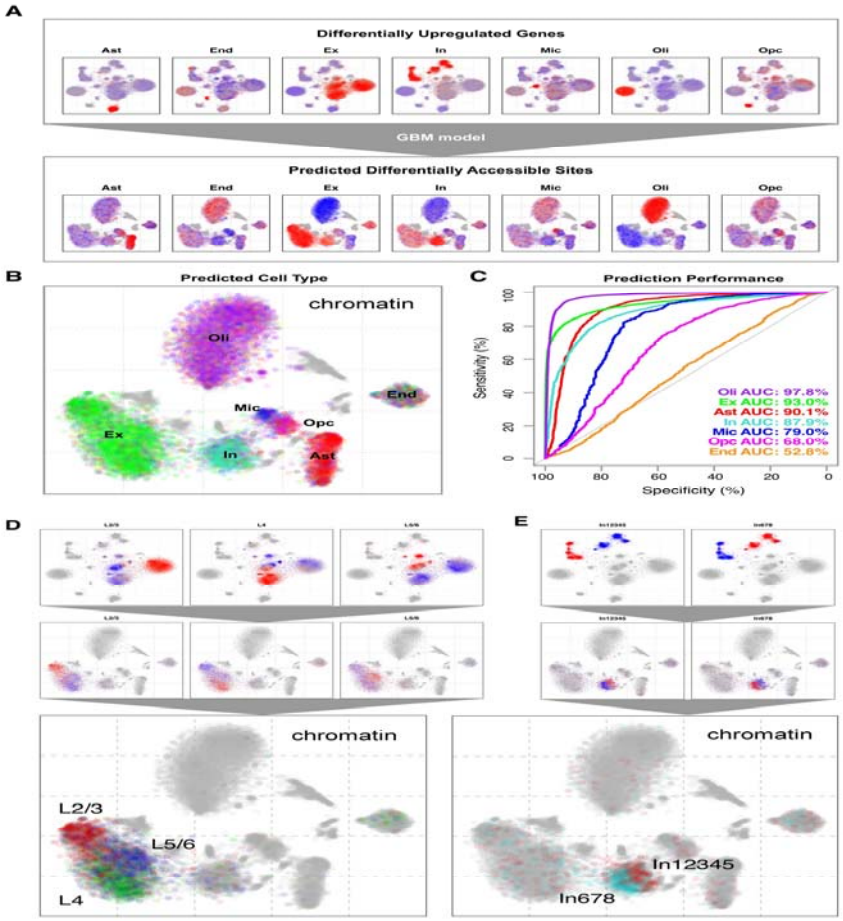


Figure 3.15. Differential gene expression predicts differential chromatin accessibility to map corresponding transcriptional and epigenetic subtypes.

A. Average expression genes significantly upregulated (Z > 1.96) in each major cell type in the visual cortex are shown, with red corresponding to high expression of blue for low expression. A GBM model predicts differential accessibility of sites observed in scTHS-seq based on the predicted association of sites with the differentially expressed genes (see Methods). Predicted differentially accessible sites are visualized with red corresponding to high accessibility and blue for low accessibility. B. Predicted cell type identities are assigned based on the cell type with the maximal joint score on predicted differentially accessible sites. C. Prediction performance is measured by ROC curves and AUC comparing joint scores and previously inferred cell type identities identified from scTHS-seq analysis alone. D. Differential expression among excitatory neuronal subtypes predicts differentially accessible regions to further divide excitatory neuronal subtypes in the scTHS-seq data. E. Differential expression among inhibitory neuronal subtypes predicts differentially accessible regions to further divide inhibitory neuronal subtypes in the scTHS-seq data.

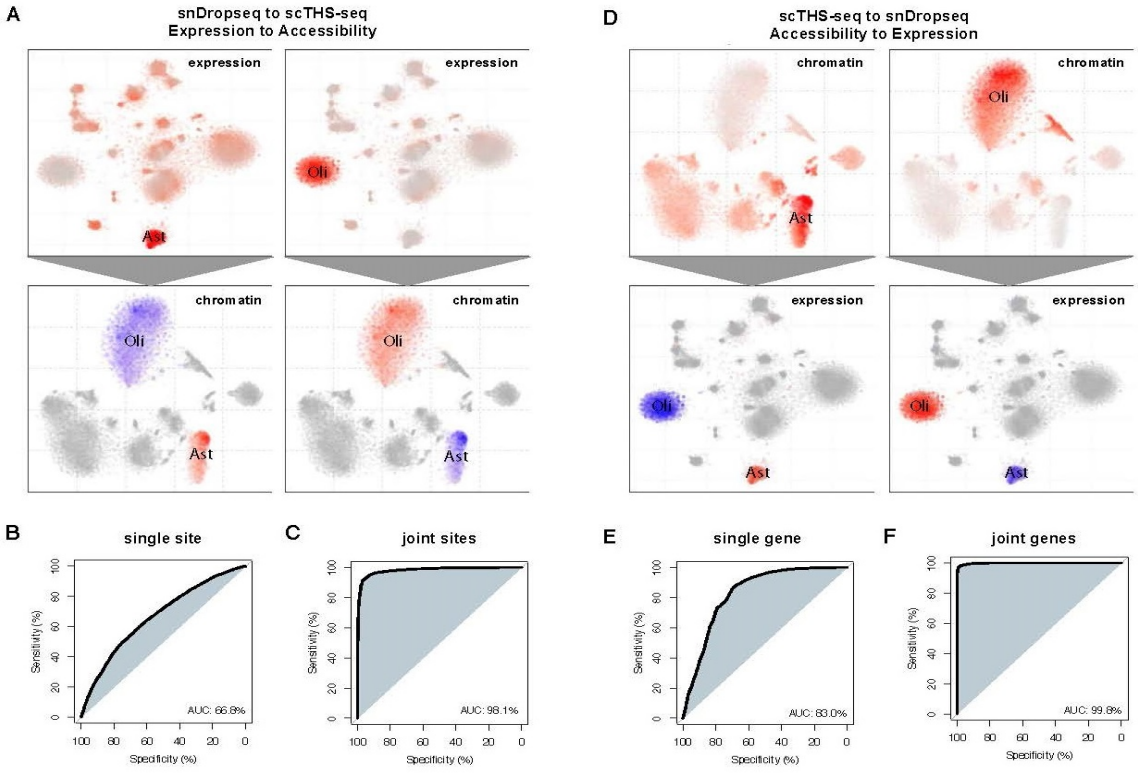


Figure 3.16. Probabilistic inference of differential expression and differential accessibility signals.
A. Predicting differential accessibility of sites from differential expression. Astrocytes (Ast) and Oligodendrocytes (Oli) identified independently from snDrop-seq and scTHS-seq data analysis are used for training and testing. Significantly differentially upregulated genes ($Z > 1.96$) in Ast and Oli are visualized (top) with red indicating higher aggregate expression of genes. Predicted differential accessibility is visualized (bottom) with red indicating higher aggregate predicted accessibility and blue indicated lower aggregate predicted accessibility. True cell population assignments are labeled. Genes highly upregulated in Ast snDrop-seq population predict sites highly accessibility in the Ast scTHS-seq population and similarly for Oli. B. A receiver operating characteristic (ROC) curve is used to illustrate the accuracy of predicting the differentially accessible sites within the proximity of the differentially-expressed genes. C. The ROC curve shows the performance of cell classification based on a multi-site predicted differential accessibility signature. In this case, we assess the ability of the average expected differential accessibility inferred from the differential expression signature to be able to distinguish Ast and Oli groups in the scTHS-seq data. D. Predicting differential expression from differential accessibility profiles. E. ROC of predicting differential of an individual gene is shown. F. ROC curve and visualization of the combined differential expression scores inferred from the differential accessibility signature comparing Ast and Oli groups.

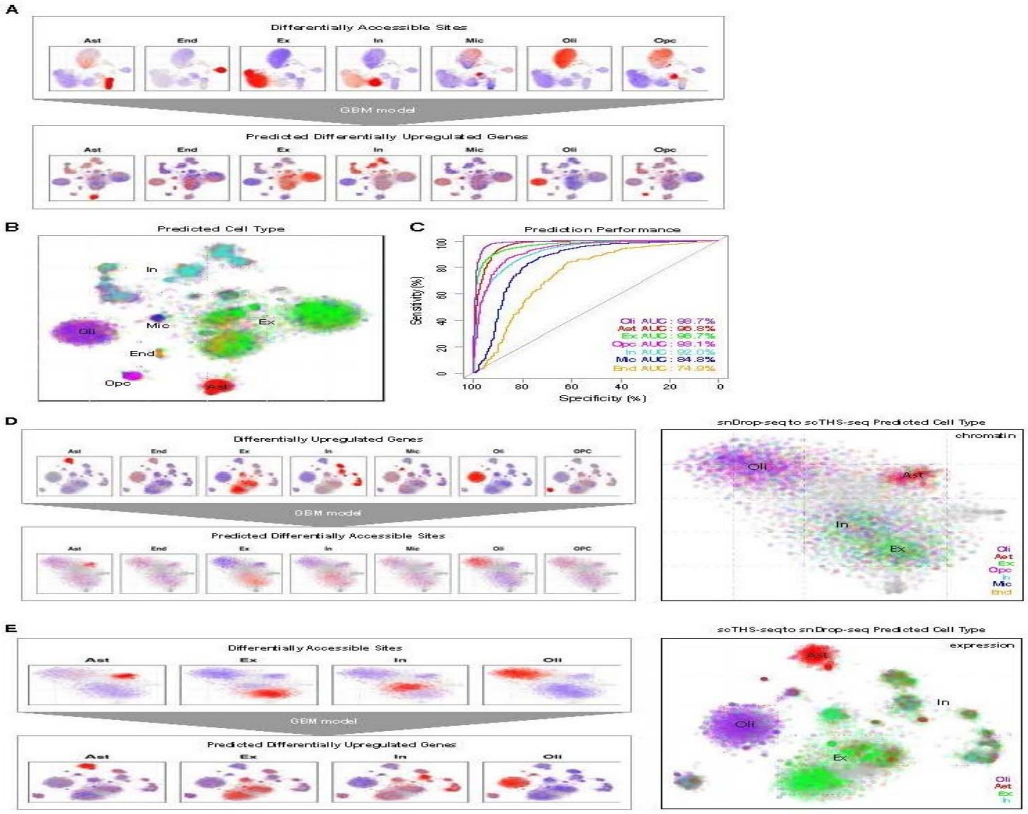


Figure 3.17. Cell type prediction based on probabilistic inference of differential expression and differential accessibility signals.
A. Significantly differentially accessible sites (Z > 1.96) for each major cell type from scTHS-seq for the visual cortex identified and visualized with red corresponding to high accessibility of blue for low accessibility. A GBM model predicts differential expression of genes observed in snDrop-seq based on select features (see Methods). Predicted differentially expressed genes are visualized with red corresponding to high expression and blue for low expression. B. Predicted cell type identities are assigned based on the cell type with the maximal joint score on predicted differentially expressed genes. C. Prediction performance is measured by ROC curves and AUC comparing joint scores and previously inferred cell type identities identified from snDrop-seq analysis alone. D. Cell type prediction based on probabilistic inference of differential accessibility from differential expression for the frontal cortex. E. Cell type prediction based on probabilistic inference of differential expression from differential accessibility for the frontal cortex.

3.2.4 Integrative transcription factor analysis

Having established the cell-type identity of each epigenetically-distinct subpopulation, we sought to identify the transcription factors (TFs) relevant to the regulatory state of each cell type. To do so, we looked for TFs whose predicted binding sites are over-represented within regions of differential chromatin accessibility distinguishing a given cell type. Screening a set of 379 TFs (**Table 14**) with position weight matrices from the JASPAR database[64], we identified 195 TFs factors showing statistically significant (FDR $<10^{-5}$) association with at least one of the cell types (**Fig. 3.18A-B, Table 15**). As expected, TFs associated with the neuronal subpopulations (**ExA**, **ExB**, and **In**) are distinct from TFs with binding relevant to non-neuronal subpopulations (**Oli**, **Ast, OPCs, Mic, and End**) (**Fig. 3.18B**). We also recover TFs with known relevance of cell subtypes. For instance *OLIG1*, an oligodendrocyte progenitor marker[65, 66], showed enriched binding within the **Oli** cluster. As an independent validation, we integrated snDrop-seq data to confirm that the TFs showing significant association with a particular cell type also tend to show higher expression levels within that cell type (**Fig. 3.19**). Some of TFs achieve positive feedback by targeting their endogenous loci. For instance, *PKNOX2* has 8 of such binding sites within its own locus, 5 of which are differentially accessible, mostly in neuronal subpopulations compared to glia (**Fig. 3.18C**), highlighting the potential of this approach to identify positive self-regulation.

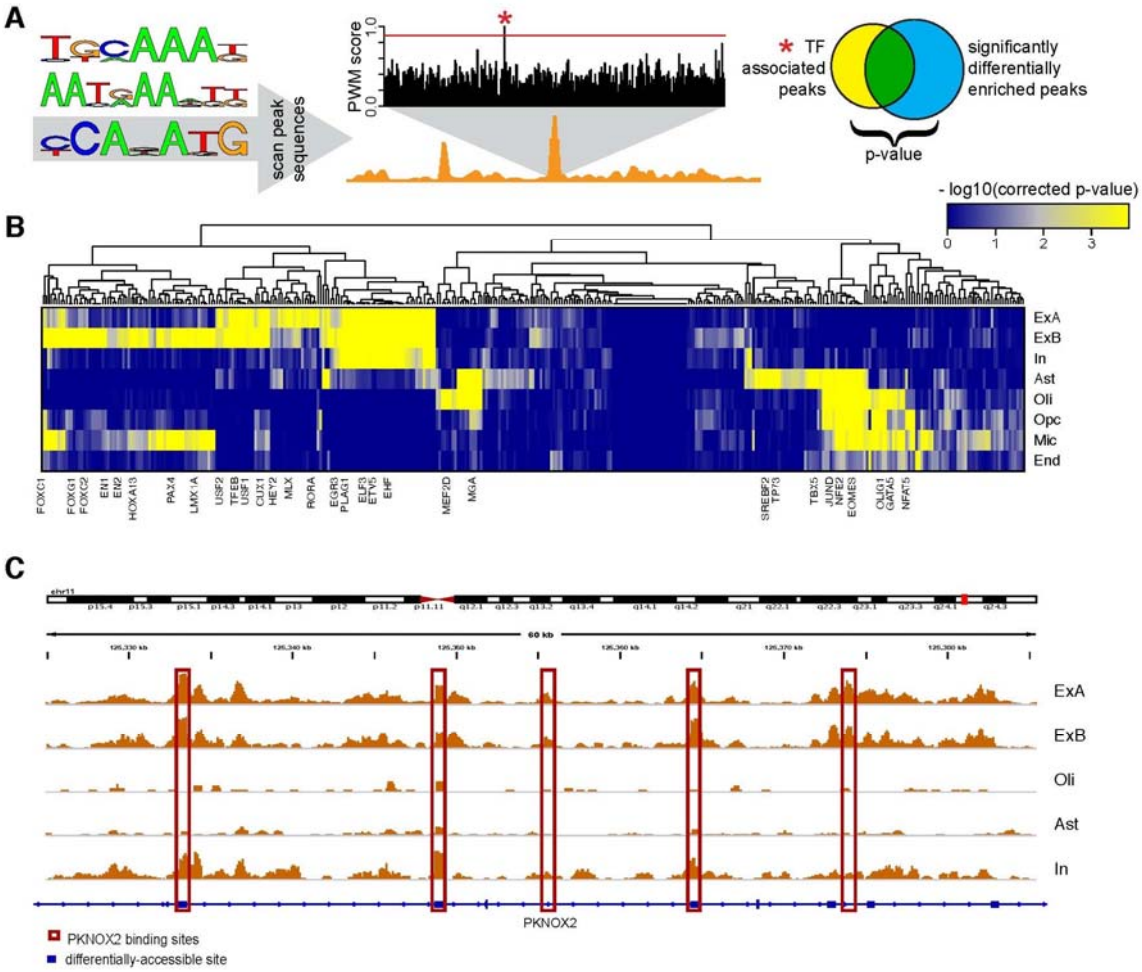


Figure 3.18. Mapping of transcription factor binding status and disease risk variants to specific brain cell types.
A. Schematic of TF analysis. Briefly, putative TF binding sites (TFBS) were identified within all hypersensitive sites based on PWMs matching. To identify relevant factors for a given cell type, sites showing differential accessibility within that cell type were tested for statistical enrichment of different TFBS. B. Heatmap of TF association to epigenetic subpopulations. Each column is a TF. Each row is an epigenetic subpopulation. Select TFs are annotated. C. PKNOX2 transcription factor shows potential for self-regulation. The IGV view of a representative region for tracks corresponding to 5 subpopulations highlights patterns of differential accessibility in the PKNOX2 gene. Differentially-accessible sites within PKNOX2 are noted with blue boxes. Putative binding sites for PKNOX2 based on PWM sequence similarity within PKNOX2 are noted with red boxes.

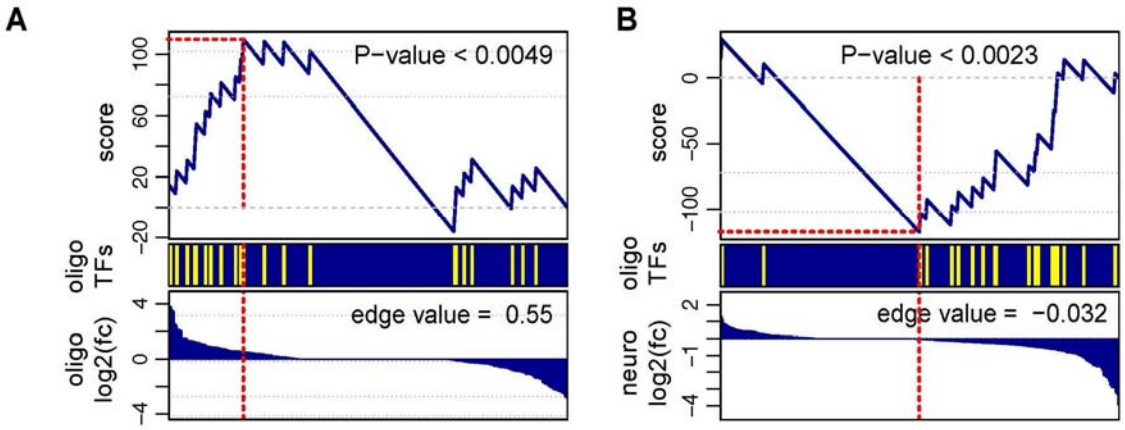


Figure 3.19. Gene set enrichment analysis plots to assess significance of identified TFs with expression of TFs.
A. Oligodendrocyte associated TFs are significantly upregulated based on ranked fold-change comparing TF expression in oligodendrocytes to neurons and astrocytes. B. Oligodendrocyte associated TFs are significantly depleted in upregulated TFs based on fold-change comparing neurons to oligodendrocytes and astrocytes.

3.2.5 Mapping disease risk variants to susceptible cell types

Cell-type specific epigenomics information has been highly valuable for identifying pathogenic cell types and specific regulatory mechanisms underlying many common genetic diseases[67-69], yet brain diseases remained inadequately understood due to the lack of epigenomic maps with any cell-type resolution. To fill this gap, we obtained NIH GRASP database SNPs that were identified from genome-wide association studies (GWAS) as significant (p-values < 10^{-6}) in ten brain related disorders, as well as seven additional diseases unrelated to the brain for controls. Given that causal variants are often located at different positions in linkage disequilibrium with the GWAS SNPs, we searched for enrichment of DNA accessibility regions in 100kb windows centered on all GWAS SNPs of a given disease, and assessed the significance by random permutations (**Fig 3.20A**, see **Methods**). This analysis identified strong enrichments in multiple cell types or sub-types, contrasting with an alternative possibility of uniformity (**Fig. 3.20B-C, Fig. 3.21, Table 18**), including: highly significant enrichment for common risk variants linked to Alzheimer's Disease in **Mic** and **Oli** cells, Schizophrenia in **ExB** and **In** neurons, **Mic** and **Oli** cells; ALS in **In** and **Ast** cells; Parkinson's Disease for the **ExB** and **In** neurons; Bipolar disorder for all **Ex** neurons and **Mic**; Autism for all **Ex** and **In** neurons; ADHD for al **Ex** and **In** neurons, **Oli** and **Opc**; ALS in **In** and **Ast** cells; with significant associations for Epilepsy for all **Ex** and **In** neurons and **Oli**, and for Depression for all **Ex** neurons and **Mic**. Significant enrichment for **Oli** was found in glaucoma, and similarly in the other non-brain diseases significant enrichments were not found in any

neurons. However, enrichments were found in the most closely related cell types to those implicated in the disease. For the autoimmune diseases Crohns, Celiac and Type I diabetes, **Mic** and **End** were enriched, while in lung disease **Mic** was enriched. No enrichment was found for the two non-brain related diseases (Chronic kidney disease and prostate cancer), demonstrating the specificity of these analyses. While further validation is required, our chromatin maps provide a new framework through which new aspects of brain diseases can be understood at the level of specific cell types or subtypes.

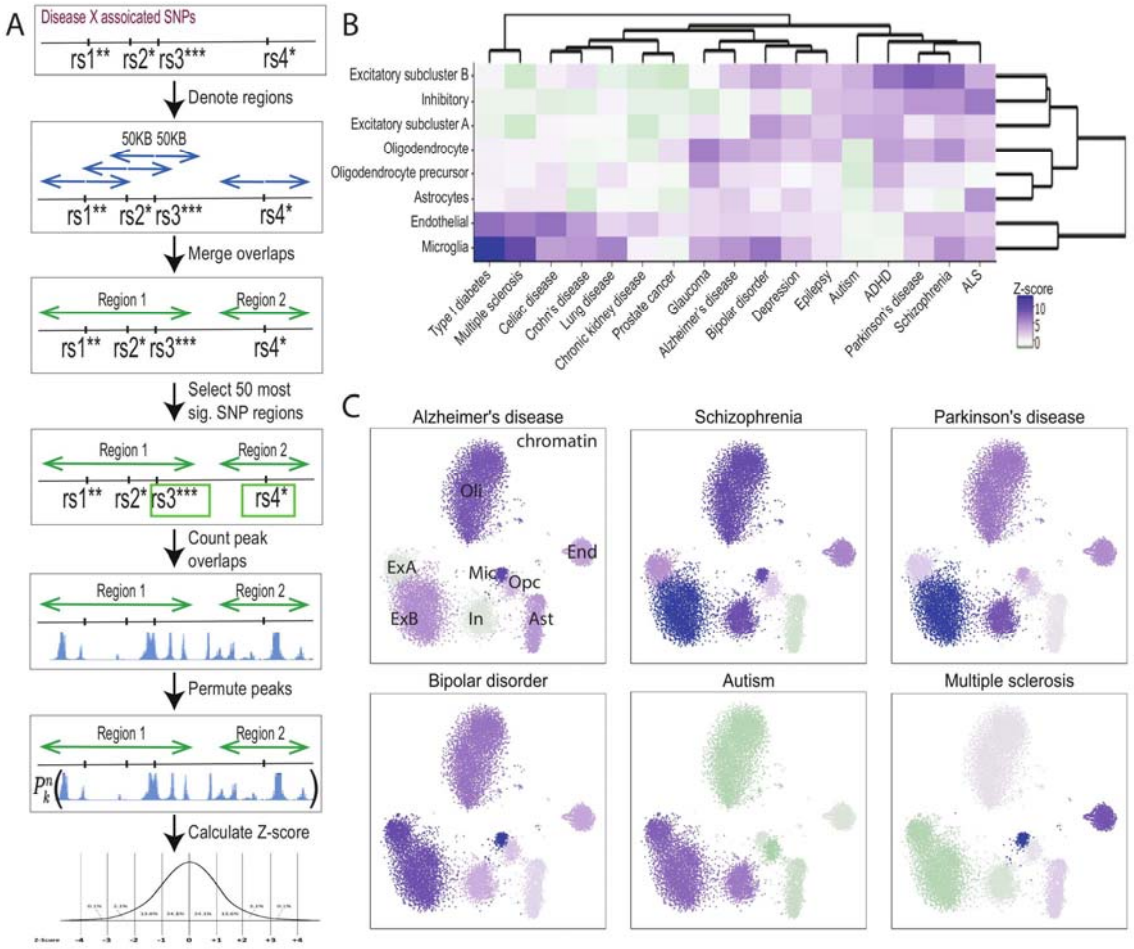


Figure 3.20. Mapping of common disease risk variants to specific brain cell types.

A. Method used to map disease risk variants to specific cell types. Briefly, GWAS SNPs were obtained for each disease, extended to 100KB, merged, the top 50 most significant SNPs selected, number of peaks in overlaps counted, peaks permuted and the number of peaks counted in each region for each permutation, then lastly Z-scores were calculated. B. Heat map representing the enrichment Z-scores across 8 cell clusters (rows) for 10 brain diseases (columns) and 7 unrelated diseases. Dark purple and purple represent a significant Z-score over 1.96, where light purple, gray and light green represent an insignificant Z-score, and green represents a significant negative association with a Z-score less than -1.96. C. Z-scores for the enrichment of GWAS SNPs in the open chromatin of ExB, In, ExA, Oli, Opc, Ast, End, Mic, populations were overlaid onto the cell clusters. Six brain disorders are shown.

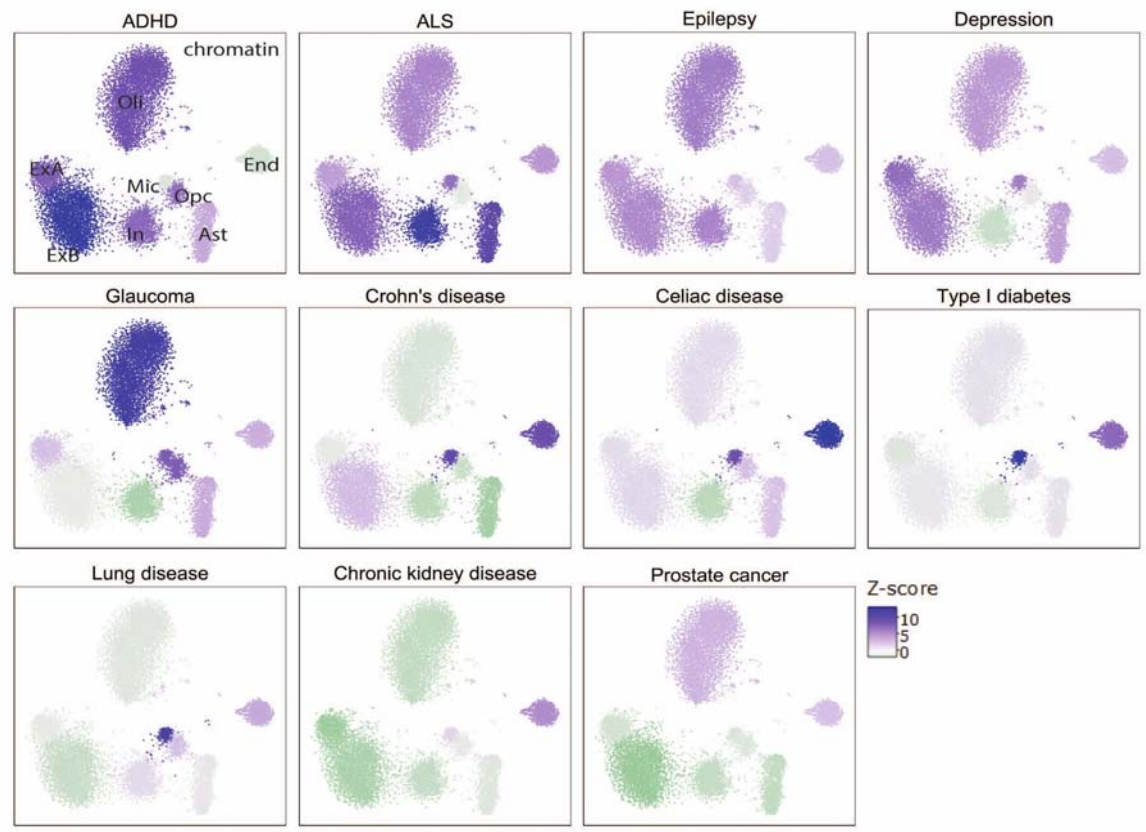


Figure 3.21. Mapping of disease risk variants to specific brain cell types in brain diseases and non-brain diseases.
Disease risk variants were mapped to specific brain cell types in four brain diseases and seven non-brain diseases, including three autoimmune diseases (Crohn's disease, Celiac disease, Type I diabetes). Dark purple and purple represent a significant Z-score over 1.96, where light purple, gray and light green represent an insignificant Z-score, and green represents a significant negative association with a Z-score less than -1.96. C. Z-scores for the enrichment of GWAS SNPs in the open chromatin of ExA, In, ExB, Oli, Opc, Ast, End, Mic, populations were overlaid onto the cell clusters.

3.3 Discussion

Reconstruction of cellular composition is an important aim towards understanding the human brain. Our study provides a first demonstration of an integrative single-cell analysis on the human adult brain, utilizing two highly scalable methods for acquiring transcriptional and epigenetic information from post-mortem tissues: snDrop-Seq and scTHS-seq. Using nuclei isolation to overcome challenges associated with processing of post-mortem tissues, we recovered all known non-neuronal and most neuronal subtypes in human adult cortex. These assays rely on a much shallower sampling of many more cells than our previous efforts, yet we were able to resolve these cell subpopulations and provide informative aggregate profiles for each subtype. Our results underscore the power of sparse sampling of single cells in complexed tissues at a massive scale: as long as the data from the single cells are informative enough for clustering and “virtual sorting” into different groups[25, 70], they can be combined into aggregate profiles that are as rich as bulk sequencing of different cell populations.

We demonstrate a computational strategy for mapping between transcriptional and corresponding epigenetic states that can be used to reconstruct aggregate epigenetic profiles for fine-grained cell types. Such profiles provide valuable insights into the regulatory processes and elements shaping the identity of different cell types, as well as their relevance to human disease. While previous studies have identified pathogenic cell types for many human common diseases, our analysis enabled an assessment of the relative impacts by

common genetic risk alleles to multiple cell types in an organ. It provides a coherent framework to consolidate previous findings, such as the relative contributions of glia, microglia and neurons to sporadic Alzheimer's disease[71]. Such information is critical for identifying effective therapeutic targets. Generating multiple types of –omics maps from single cells *en mass* also enabled us to leverage the strength of each method to improve the confidence of cell type assignment, greatly enriching cell annotations. This combined approach thus represents a highly scalable strategy for systematic construction of atlases composed of single-cell data for human organs like the brain and eventually, for the full human body.

Chapters 3, in full is an adapted version of the material that has been submitted for publication. Blue B. Lake†, Song Chen†, Brandon C. Sos†, Jean Fan†, Yun Yung, Gwendolyn E. Kaeser, Thu E. Duong, Derek Gao, Jerold Chun*, Peter Kharchenko*, Kun Zhang*. 2017. Integrative single-cell analysis by transcriptional and epigenetic states in human adult brain. In revision. The dissertation author was the co-primary author of all material.

CHAPTER 4: Conclusions

Currently there are two single cell methods to assess chromatin accessibility in single cells, scATAC-seq and scDNase-seq[17, 26]. In this dissertation we improve upon these methods by developing a high throughput and high sensitivity assay, scTHS-seq, to assess chromatin accessibility in single cells and address the shortfalls of current methods. Initially we demonstrate THS-seq is the highest sensitivity chromatin accessibility assay on bulk samples and is able to generate high quality data. We assessed chromatin accessibility in 100 cells, and generated data comparable to data generated in 50,000 cells by ATAC-seq and millions of cells by DNase-seq. The significant finding was that with THS-seq we were able to find more distal elements, peaks, and smaller peaks, thus providing more information about regulatory regions than current methods. We believe this is due to method design, resulting in the preferential capture of open chromatin compared with ATAC-seq. When applied to biological samples this can help provide direction on which regions are important for regulation and should warrant further investigation as well as elucidate functions of previously unannotated regulatory regions. Additionally, we characterized bias in ATAC-seq due to method design that was not found in THS-seq, and should be taken into account during data analysis.

Single cell DNA accessibility assays have been developed are able to provide either high sensitivity or high throughput data, but not both. Single cell DNase-seq is not high throughput, with one reaction per well per cell[17].

Meanwhile, scATAC-seq utilizing combinatorial indexing can generate thousands of cells worth of data in a single experiment, however sensitivity is low[26], which has also been found by a collaborator (unpublished data). Recently Buenrostro et al. has shown scATAC was able to assay ~2,200 cells utilizing the C1 autoprep system, however this is ~5X-7X less cells than what a single scTHS-seq experiment has been able to generate. With higher throughput and higher sensitivity of scTHS-seq, now large single cell datasets can be generated quickly, and with high data quality, opening new avenues of data analysis that were not possible before. Additionally, with higher throughput we can still cluster cell types based on sparse sampling. These large datasets can then be leveraged for generating new biological insights, and integration with other single cell omics datasets to provide a global picture of regulatory states in cells.

Several studies have shown integration of bulk DNA accessibility data with other omics datasets and application to find novel biological insights[29, 30]. In this dissertation we demonstrate integration of data from scTHS-seq and snDrop-seq on human brain nuclei to help refine clustering. With single cell analyses, samples no longer need to be sorted by FACS or other methods into respective cell types, as each single nucleus or cell is assayed individually. However, with the case of the brain nuclei, separation by sorting could not be performed since FACS markers delineating cellular identity are usually on the cellular membrane. Samples can now be grouped computationally using gene markers in accessible regions. Epigenetically similar cellular populations can be further partitioned by integrating RNA expression data, providing a higher resolution epigenetic picture

of the cell types present within the dataset. Our findings demonstrate DNA accessibility may be defining broad cell type classes, while snDrop-seq and RNA expression profiles may define specific cell types within those classes. Development of omics data integration methods provide a framework for clustering and identifying cells to provide a biological structure to samples containing diverse cell types.

In previous GWAS studies on human brain diseases, variants have been linked to disease. However, further studies are needed to implicate specific cell types and brain regions in disease and shed light on the causative mechanisms. Human brain samples are also difficult to obtain so many brain studies are performed on mouse or other model organisms, which though informative may not have direct functional correlation to human brains. Here we demonstrate mapping of disease risk association to specific cell types in brain diseases using chromatin accessibility data, which to our knowledge has not been performed before. For each disease, this provides a global view of the most susceptible cell types. We also examined the specificity of transcription factor motifs to cell types, thus helping us further define the cell subpopulations classified. This analysis framework can be applied to any other human organ scTHS-seq data is generated from and elucidate the most probable cell types involved in disease susceptibility.

In summary, we have demonstrated development of a highly sensitivity chromatin accessibility assay on minute cellular inputs, and on single cells. To demonstrate biological application, we performed scTHS-seq on human brain

samples, performed an integrative data analysis with snDrop-seq transcriptomic data, defined cell type specific transcription factor landscapes, and associated human brain cell types with susceptibility to brain diseases.

APPENDIX A: THS-SEQ MATERIALS AND METHODS

Cell culture

GM12878 lymphoblastoid cells were obtained from Coriell Cell Repositories and maintained in RPMI 1640 with 1% Penstrip and 10% FBS. Cells were kept in suspension at 37°C and were harvested when suspensions reached optimal density. Cell numbers were quantitated with a hemocytometer (Biorad), pelleted at 250 x g for 4 minutes and washed twice with a volume of PBS sufficient to obtain ~1,000,000 cells per milliliter in ~1 mL of PBS, and diluted as needed in PBS for experiments. All biological replicates were performed with material from the same cell culture flask, with independent running of the assays. When cells were diluted to 100,000 cells per milliliter they were quantitated again on the hemocytometer.

Transposome generation

The transposon consisted of two DNA fragments synthesized by IDT, T7tspn-top2 (PAGE purified) and T7tspn-bot (**Supplementary Table 5**). These two fragments were incubated together to form annealed transposon at a concentration of 30 μM per oligo in Qiagen EB buffer for 2 minute at 95°C, and cooled to room temperature at 0.1°C/s. Transposome generation by addition of transposase and annealed transposon, using either EzTn5 or Tn5059 was performed at Illumina. Prepared transposome was stored at -20°C.

Cell lysis, tagmentation and DNA fragment processing

Cells were concentrated and aliquoted so the total number of cells was present in 1 µL optimally. Cells were lysed in 1X lysis buffer in reaction (10X concentration: 100 mM Tris-HCl pH 7.5, 100 mM NaCl, 30 mM MgCl2, 1% NP40), at room temperature for 3 to 5 minutes by adding 1.0 µL of 2X lysis buffer to the cell mixture. Next, the transposition reaction was carried out at 37°C for 30 minutes by adding 1.0 µL tagmentation buffer (5X concentration: 165 mM Tris-OAc pH 7.8, 330 mM K-OAc, 50 mM Mg-OAc, 80% dimethylformamide), 1.0 µL nuclease free water, and 1.0 µL of prepared transposomes (added last) to lysed cells for a total reaction volume of 5 µL. EzTn5 transposomes were added at a final concentration in reaction of 0.5 µM, and Tn5059 transposomes at a final concentration in reaction of 0.7 µM. After transposition, reactions were brought to 15 µL with nuclease free water, and 15 µL of 8M Guanidine HCl was added to degrade chromatin associated proteins and transposase. Next, reactions were purified with 1.8X (54 µL) SPRI beads. After the final wash with 80% ethanol and brief air drying of SPRI beads, 9.6 µL of nuclease free water was added. Importantly, the purified DNA products were not eluted off beads, and DNA products, nuclease free water, and beads remained together in solution. Gap fill-in was performed by addition of 2.4 µL of 5X Taq polymerase solution (NEB), mixed thoroughly, and incubated at 72°C for 3 minutes.

In vitro Transcription

DNA was used directly, in the same tube, and remaining in solution with SPRI beads after gap filling for *in vitro* transcription. Standard protocol with the MAXIscript® T7 Kit (Ambion) was followed. The 10X transcription buffer was kept at room temperature, and if crystals were present, heated slightly at 37°C until crystals dissolved and returned to room temperature. Also as per manufacturer's recommendation, after briefly thawing, ribonucleotides were stored on ice. Reactions were assembled at room temperature, with addition of 2 µL 10X transcription buffer, 1 µL 10 mM ATP, 1 µL 10 mM CTP, 1 µL mM GTP, 1 µL 10 mM UTP and 2 µL T7 enzyme mix to 12 µL of reaction products. Reactions were incubated at 37°C for 16-19 hours. After *in vitro* transcription, RNA was purified with Zymo RNA Clean and Concentrator and eluted with 10 µL nuclease free water. For purification, reactions were brought to 40 µL with nuclease free water before starting the protocol, and whole reactions including SPRI beads in solution were applied to the RNA purification column. Reaction products were visualized and quantitated with a *UV*-transilluminator after running 1 µL of IVT product on a 6% Tris-Borate-Urea (TBU) gel for 20-25 minutes at 250 Volts and staining with SYBR gold. For each sample the region from 250 base pairs to 2000 base pairs was used for RNA quantitation and determination of the volumes of product needed for input into RNA-seq processing, since that portion of the gel is amplified RNA product. 50 nanograms of RNA was used for input into RNA-seq processing.

RNA-seq (custom transposon mediated fragmentation)

For first stranded cDNA synthesis the standard Clontech SMART MMLV reverse transcriptase protocol was followed. 2.5 µL of 20 µM random hexamers were first added to 50 ng RNA for each sample and incubated at 70°C for 3 minutes then cooled immediately on ice. MMLV reverse transcription mastermix was added to make a 20 µL total reaction volume, followed by incubation at room temperature for 10 minutes, and then 42°C for 60 minutes, and inactivated at 70°C for 10 minutes. To remove RNA in DNA/RNA hybrids, samples were incubated for 20 minutes with 0.5 Units/µL RNase H at 37°C. DNA templates were primed with 2.5 µL 20 µM sss_NPA_prmr (**Supplementary Table 5**) and incubated for 2 minutes at 65°C then cooled immediately on ice. For second stranded synthesis 5.88 µL of 5X Taq PCR master mix (NEB) was incubated with samples at 72°C for 8 minutes. Next, 45 µL nuclease free water was added to reactions, and then double stranded DNA was purified with SPRI beads at 1:1.8X ratio (135 µL SPRI beads), eluted with 20 µL nuclease free water and then concentrated to ~4 µL in a Eppendorf™ Vacufuge™ Concentrator, set to 30°C and spun for approximately 15 to 18 minutes. Fragmentation and 3’ end tagging of double stranded cDNA was simultaneously performed by adding 1 µL tagmentation buffer, and 1 µL loaded custom 0.1 µM final in reaction EzTn5 transposomes (ME_BOT, ME_TOP (**Supplementary Table 5**)) to the solution and then incubating at 55°C for 6 minutes, followed by cooling on ice. The same tagmentation buffer that was used in the transposition reaction for tagging open chromatin was used for this reaction. 0.14 µM final in reactionTn5059 can also be used in place of EzTn5 and would provide better read diversity, however for this experiment we did not have

Tn5059 with ME_BOT and ME_TOP available. Also, this transposome complex is prepared the same way as described in transposome generation above, however with different oligos as specified. Next, samples were incubated with 0.1 AU Qiagen protease at 50°C for 10 minutes, then 70°C for 20 minutes. Gap fill-in was performed by addition of 6 µL of 2X Taq PCR master mix (NEB), and incubated at 72°C for 3 minutes. To the sample, 9.0 µL of 2X Taq was added with 2 µL of 10 µM of 5' PCR primer, 2 µL of 10 µM Index barcode primer (for each individual unique sample) (**Supplementary Table 5**), 1.0 µL of 25X SYBR green, 4 µL nuclease free water and 12 µL of DNA template for a 30 µL PCR reaction. PCR cycling consisted of an initial incubation at 72°C for 3 minutes then 95°C for 30 seconds and cycling at 95°C for 10 seconds, 63°C for 30 seconds and 72°C for 3 minutes until curves reach saturation. Generally, reactions were stopped at the end of cycles 9-11. The products of the PCR reactions were visualized and quantitated with a *UV*-transilluminator after running 1 µL of PCR product on a 6% Tris-borate-EDTA (TBE) gel for 20 minutes at 250 Volts and staining with SYBR gold. Barcoded samples were pooled and gel size selected for fragments between ~220-1000 base pairs. Now libraries were ready for sequencing.

Custom ATAC-seq

Transposome complexes were generated at Illumina and provided at 2.5 µM. To perform the assay cells were first washed in PBS and then spun down at 500Xg for 5 minutes. Cells were then counted with Bio-rad TC-20 cell counter diluted to

500,000 cells/mL, and then 1 µL was aliquoted to sample tubes. 1 µL of 2X lysis buffer (same as THS-seq) was added to each reaction, followed by 1 µL 5X custom tagmentation buffer (same as THS-seq), 1 µL EzTn5 Transposomes and 1 µL nuclease free water. Reaction was incubated at 37°C for 30 minutes. Next, 1 µL of 0.05 AU/mL Qiagen protease is added to the reaction, and incubated at 50°C for 10 minutes, then 70°C for 20 minutes. Samples are now ready for PCR, and 1.50 µL 10 µM barcoded i5 adaptor, 1.50 µL 10 µM barcoded i7 adaptor, 15.0 µL KAPA SYBR FAST, and 6.0 µL nuclease free water are added to the whole reaction (6.0 µL) from the previous step. PCR cycling consisted of an initial incubation at 72°C for 3 minutes then 98°C for 30 seconds and cycling at 98°C for 10 seconds, 63°C for 30 seconds and 72°C for 1 minute until curves reach saturation. Generally, reactions were stopped at the end of cycles 9-11. Samples were then cleaned with AMPURE SPRI beads at 1.8X ratio. Samples were visualized and quantitated with a *UV*-transilluminator after running 1 µL of each reaction on a 6% Tris-borate-EDTA (TBE) gel for 20 minutes at 250 Volts and staining with SYBR gold. Barcoded samples were pooled and gel size selected for fragments between ~220-1000 base pairs. Now libraries were ready for sequencing.

Primary data Processing

Sequencing was performed on HiSeq with 50 x 6 single end reads with single six base pair indexing for THS-seq data, and 50 x 8 x 8 single end reads with dual eight base pair indexing for ATAC-seq data. Technical replicates were combined

before analysis, while biological replicates were kept separate. Any down sampling of mapped alignments was performed with in house scripts randomly sampling reads from the original datasets. Sequencing reads were first mapped to the hg19 reference genome with BWA 0.7.5a-r405 using default parameters. After conversion to a bam file with samtools, clonal read removal was performed. For identifying significantly enriched regions, Dfilter was used with default parameters, unless stated. Overlaps were calculated using the bedtools function coverageBed. Individual loci read density data were extracted from BAM files, and converted to wig files and examined as a custom tract on the UCSC genome browser. Correlations were performed and graphed in R, and peak distribution graphs among other graphs were graphed with graphpad prism version 5.0. Proportional Venn diagrams were generated with Venn Diagram Plotter, (http://omics.pnl.gov/software/venn-diagram-plotter). Gene ontology analyses on significantly called peaks were performed with GREAT[43].

Counting the number of alignments in peaks

Files with significantly called peaks generated by Dfilter were used to obtain peaks regions called for each sample. To count all alignments in peaks, the function “samtools view –Lh bam_files bed_region >> output“ was used to output and concatenate alignments overlapping the input BED peaks file to a new file, followed by counting the number of reads total in the outputted file. To count alignments in each individual peak, the function “samtools view –Lh bam_files bed_region > output“ was used for each individual peak, and then the number of

alignments in each peak region was counted. To get the normalized number of reads per 100 base pairs, first any peaks that had zero alignments were removed, and then the total number of reads for each peak was divided by 100. Each dataset was sorted first by the number of reads per 100 base pairs, and then by peak length before plotting in excel or graphpad prism version 5.0. For calculating the mean ± SEM, datasets with the normalized number of reads per 100 base pairs were saved as CSV files with peak length in the first column, and reads per 100 base pairs in the second column, and then run through an in house python script calculating the mean ± SEM for each peak length, with each peak length being in increments of 100 base pairs. Results were then graphed in Graphpad Prism version 5.0.

Published ENCODE data analysis

Duke and UW ENCODE GM12878 lymphoblastoid datasets were downloaded from the UCSC genome browser (data accessible at NCBI GEO database[72], accessions GSM816665, GSM736496, GSM736620) and were processed using the same pipeline used to process THS-seq and ATAC-seq data, except bowtie version 1.0.0 with stringent parameters (bowtie -n1 -k1 --best --chunkmbs 10240 --strata -l32 -m1 -p4 --nomaqround –sam) was used for mapping. To get the most comprehensive accessibility data from Duke and UW, all replicates from each respective dataset were combined before mapping. In total, 144,738,228 and 47,417,059 total unique alignments were used for analysis for Duke and UW respectively.

Published ATAC-seq data analysis

All raw paired end ATAC-seq data for 50,000 cells and 500 cells was downloaded from GEO gene expression omnibus (data accessible at NCBI GEO accession GSE47753). Only read 1 of the paired end data was used for analysis to provide a fair comparison to single end THS-seq data. ATAC-seq read 1 datasets were then aligned with BWA using default parameters. Down sampling of mapped alignments was performed with in house scripts randomly sampling reads from the original datasets, and then run through Dfilter for calling of significant peaks.

APPENDIX B: scTHS-SEQ MATERIALS AND METHODS

Sample Origin and Nuclei Preparation

All human tissue protocols were approved by the Scripps Office for the Protection of Research Subjects (SOPRS) at The Scripps Research Institute (TSRI) and conform to National Institutes of Health guidelines. Nuclei were prepared using nuclear extraction buffer (NEB) as described previously (*7*). Briefly, fresh frozen post-mortem brain tissue was sectioned at 50um using a cryostat and placed in 1 ml of ice-cold NEB for 10 minutes. Nuclei were extracted using a glass dounce homogenizer with Teflon pestle using 10-12 up-and-down strokes in 1 ml of NEB. Samples were passed through a 50um filter (Sysmex Partec), incubated on ice for 10 minutes. Samples were spun for 5 minutes at 250-300 x g, washed in PBS + 2 mM EGTA (PBSE), and resuspended in PBSE supplemented with 1% fatty-acid free bovine serum albumin (FAF-BSA, Gemini) containing 4',6-diamidino-2-phenylindole (DAPI). DAPI+ single nuclei were purified by flow cytometry using MoFlo Astrios (Beckman Coulter), concentrated at 900g for 10 minutes and used directly for droplet encapsulation.

Nuclei encapsulation, mRNA-Seq Library Preparation and Sequencing

Drop-seq was performed as described previously[32], but with modifications optimized for processing nuclei, now termed snDrop-seq. Before droplet generation, connecting tubing and syringes were coated with 1% bovine serum albumin (BSA) to prevent non-specific binding of nuclei to the surface, and then

rinsed with PBS. To reduce nuclei settling, Ficoll PM-400 was added to nuclei suspension buffer, rather than the lysis buffer. Nuclei were loaded at the concentration of 100 nuclei/ml, and co-encapsulated in droplets with barcoded beads purchased from ChemGenes Corporation, Wilmington MA (Cat. # Macosko201110). When encapsulation was complete, the droplet-collecting falcon tubes were added with a layer of mineral oil, and then transferred to 72°C water bath. After 5 minutes of incubation, the tubes were removed from the water bath to ice and droplets were broken by perfluorooctanol, following which beads were harvested, and hybridized RNA was reverse transcribed. cDNA was then PCR amplified for 16 cycles with primer, buffer and cycle conditions identical to those described previously (*2*). A total of 32 libraries were prepared from 14 experiments, and cDNA from each replicate was prepared and tagmented by Nextera XT and indexed with different Nextera index 1 primers. cDNA libraries were pooled and sequenced on the Illumina Hiseq 2500 using Read1CustSeqB (*2*) for priming of read 1 (read 1 was 30 bp; read 2 (paired end) was 120 bp).

snDrop-seq data processing

Paired-end sequencing reads were processed largely as described (http://mccarrolllab.com/wp- content/uploads/2016/03/Drop-seqAlignmentCookbookv1.2Jan2016.pdf) with additional correction steps. First, paired-end reads were filtered out if read 1 had more than 4 non-T bases in the last 10 bases (to remove all non poly T-captured contaminated reads), or had one or more bases with poor quality score (less than 10). And cell barcode and

UMI information were then inferred from the first 12 bases and the next 8 bases of read 1 respectively. The right mate of each read pair was trimmed to remove any portion of the SMART adaptor sequence or large stretches of polyA tails (6 consecutive bp or greater). The trimmed reads were then aligned to the human genome (GENCODE GRCH38) using STAR v2.5 with the default parameter settings. Reads mapping to intronic or exonic regions of genes as per the GENCODE gene annotation were both recorded. One further correction step to fix barcode synthesis errors was performed by inserting N at last base of the cell barcode for reads which had identical first 11 bases of the cell barcode and same last T base of UMI. Digital expression matrix was then generated with genes as rows and cells as columns. UMI counts were assigned for each gene of each cell by collapsing UMI reads which had only 1 edit distance.

snDrop-seq Data Clustering and Analyses

UMI matrix cell barcodes were tagged by their associated sequencing library batch ID (**Table 6)** and combined across independent experiments. Mitochondrial genes not expressed in nuclei were excluded and only UMI counts associated with protein-coding genes were used for clustering analyses. Nuclei with fewer than 300 molecules or more than 5,000 molecules (outliers) were omitted. Molecular counts were normalized using the total number of reads, as the estimate of library size for each cell. The expression values were rescaled so that mean expression of a gene within each measurement batch was equal to the dataset-wide average. Winsorization procedure was used to cap the magnitude

of 10 most extreme values for each gene. To estimate the residual variance for each gene, variance dependency on the expression magnitude (log scale) was modeled as a smoothed generalized additive model with smoothing term k=10 (using mgcv package in R). The observed to expected variance ratio for each gene was modeled using F distribution using the degrees of freedom corresponding to the number of successful gene observations. To normalize the contribution of each gene in the subsequent PCA analysis, the variance of each gene was rescaled to match the tail probability obtained from the F distribution under a standard normal sampling process. For both the visual cortex dataset, cell clusters were determined using approximate k-nearest neighbor graph (k=50) based on a cosine distance of the top 150 principal components (PCs) derived from the top 1500 variable genes from the variance-adjusted expression matrix, using the infomap community detection algorithm (as implemented in the igraph R package). Cell clusters with fewer than 30 cells were omitted from further analysis. Similarly, for the smaller frontal cortex dataset, cell clusters were determined using a smaller k=30 based on the consine distance of the top 100 PCs derived from the top 2000 variable genes. Cluster stability was assessed using 85% subsampling. For both datasets, Cluster1 grouped low-depth cells that could not be confidently assigned to other clusters, and was omitted. Resulting cells were reclustered and visualized using t-Distributed Stochastic Neighbor Embedding (t-SNE) on the 150 PCs using a perplexity of 100 for the visual cortex dataset and on the 100 PCs using a perplexity of 300 for the frontal cortex dataset.

Violin plots and differential gene expression analyses were performed using Seurat software (V1.4.0.5) in R (https://github.com/satijalab/seurat). For normalization, UMI counts for all annotated nuclei were scaled by total UMI counts (excluding mitochondrial genes), multiplied by 10,000 and transformed to log space. Technical effects associated with UMI coverage and batch identity were regressed from scaled data using a linear model (RegressOut function in Seurat). Cell type and subtype enriched genes were identified (Seurat software) using a likelihood-ratio test on all protein-coding genes to identify 0.25 fold (log scale) enriched genes detected in at least 25% of cells in the cluster. Gene enrichment analyses were performed for all clusters and for excitatory or inhibitory neuron subtypes separately.

Comparison of snDrop-seq Data with Published Data

Control bulk RNA-seq data (FPKM values) from the mouse cerebral cortex and human temporal lobe were obtained from: http://web.stanford.edu/group/barres_lab/brain_rnaseq.html and http://web.stanford.edu/group/barres_lab/brainseqMariko/brainseq2.html, respectively. Top 50 cell type enriched genes were derived from comparison of averaged expression values of each cell type against an average of the remaining cell types (with the exception of oligodendrocyte sub-populations which were compared only against non-oligodendrocyte lineages). Type enriched genes from bulk data sets were used for correlation of log averaged

FPKM values of the associated bulk RNA-seq data with log transformed average expression values from snDrop-seq data.

For comparison with single-cell RNA-seq data from the human temporal lobe[52], gene count data was obtained from GEO (GSE67835), normalized using Seurat as mentioned above using a minimum cutoff of 1000 genes detected. Highly variable genes were identified from a mean variability plot (average expression versus dispersion (Variance/mean) assigned to 20 bins based on average expression) using a log(Variance/mean) cutoff of 1 to identify 2235 genes. Principle component analysis (PCA) was performed on these highly variable genes then projected to the entire dataset. Statistically significant PCs (p value < 0.05) were identified using a jack straw approach. Cell identities from the original publication were maintained and the top 50 genes from the statistically significant principal components differentiating these cell types, as well as the top 10 differentially expressed genes associated with each cell type, were identified using Seurat and used for correlation of log transformed averaged expression values from scRNA-seq and snDrop-seq data. For comparison with single cell RNA-seq data from the mouse visual cortex[51], gene RPKM data was obtained (GSE71585), log transformed and loaded into Seurat using published cluster annotations. Neuronal subtypes were combined into a single group and average cluster expression values were obtained across cell types using previously described marker genes present in each cluster[51] and a correlation heatmap of log transformed averaged expression values was generated. SNS data generated on the Fluidigm C1 platform[53] (dbGaP accession

phs000833.v3.p1) were used for correlation of log transformed subtype-averaged expression values for differentially expressed genes (greater than 2-fold) underlying previous subtype clustering and classification[53] combined with the top 10 cluster-specific markers for the visual or frontal cortices (Table 8). For comparison of subtypes between visual and frontal cortex data, correlation of log transformed subtype-averaged expression values for the top 10 marker genes for each region were used.

For comparison of UMI counts and genes detected with scDrop-seq data from the mouse retina (*2*), the full UMI count table for 44808 annotated samples was obtained from GEO (GSE63472). For comparison with 9k brain cell data sets from an E18 mouse generated on the 10X platform (Cell Ranger 1.3., v2 Chemistry), filtered gene matrices were downloaded from the company website. For comparison with human embryonic midbrain single cell data sets generated using the Fluidigm C1 platform[55], annotated UMI count matrices were obtained from GEO (GSE76381). Each data set was analyzed using Seurat for t-SNE visualization of clusters.

RNA In Situ Hybridization (ISH)

For RNA ISH from the visual cortex (Fig. 2), representative images were obtained from the Allen Human Brain Atlas (http://human.brain-map.org) and corresponding links are provided in Table 11.

scTHS-seq Sample Origin and Nuclei Preparation

The same human tissue samples use in used in snDrop-seq were used for scTHS-seq. However, after flow cytometry nuclei were processed differently and were always kept on ice. Nuclei were spun down at 500xg for 5 minutes at 4ºC, then supernatant was removed, and the pellet was resuspended in 1X lysis buffer (1X concentration: 10 mM Tris-Hcl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$ 0.1% NP-40, 2% bovine serum albumin, one roche protease inhibitor tablet per 10 mL, in PBS) and chilled at 4ºC for 5 minutes without shaking. Then nuclei were spun down at 500xg for 5 minutes, supernatant was removed, and the pellet resuspended in 1.5X tagmentation buffer (1.0X concentration: 33 mM Tris-OAc, pH 7.8, 66 mM K-OAc, 10 mM Mg-OAc, 16% Dimethylformamide). Now the nuclei sample was ready for nuclei counting and species/species sample mixing. For running scTHS-seq, a mouse nuclei sample and a human nuclei sample was always mixed. The mouse nuclei sample was prepared the same way after flow cytometry. For species/species mixing, both the human nuclei and mouse nuclei samples were counted on a Bio-Rad TC20 cell counter, and diluted with 1.5X tagmentation buffer, or further concentrated by spinning down nuclei at 500xg for 5 minutes at 4ºC and resuspending in a lower volume with 1.5X tagmentation buffer, so the samples cell counts were within 10% of each other. A cell concentration of ~$2.4X10^5$ nuclei/mL was obtained for each sample with the optimal range being $2.0X10^5$ – $5.0X10^5$ nuclei/mL with ~1 million total nuclei for each sample. Next, equal volumes mouse and human samples were combined, and mixed gently. The sample was now ready for transposition and combinatorial indexing.

scTHS-seq transposon generation

Each transposon consisted of two oligos synthesized by IDT and kept at 100 µM stock solutions in TE buffer, the 74 base pair barcoded transposon and 19 base pair universal 5' phosphorylated mosaic end. In total there were 384 barcoded r5 transposons each with a unique 6 base pair barcode and all barcodes in the barcode set had a minimum edit distance of 2 (**Table 17**). In order to generate the annealed transposons, 10 µL of each 100 µM oligo was aliquoted into one reaction for a total of 384 uniquely barcoded r5 transposon reactions in a 384 well plate with one uniquely barcoded r5 transposon per well. The 384 well plate was incubated at 95ºC for two minutes, and cooled to 14ºC at 0.1ºC/second. Each uniquely barcoded annealed r5 transposon was now at 50 µM in 20 µL, and was then diluted to 8.4 µM and a final concentration of 50% glycerol in a new 384 well plate. Dilution was with a solution of TE and 75% glycerol. Transposons were stored at -20ºC.

scTHS-seq barcoded r5 transposome complex generation

Tn5059 was generated at Illumina, and normalized for activity before being received. It is important to note complexed Tn5059 and transposon slowly degrades transposon over time, with noticeable loss in data quality after a few weeks, however uncomplexed components have a long shelf life. Because of this, r5 transposome complexes were generated the morning scTHS-seq was going to be run. All steps were performed with the E1-Cliptip programmable 12

channel electronic pipettor. First, Tn5059 was diluted to 4.2 μM in standard storage buffer provided by Illumina. In a 384 well plate, 1 μL of 4.2 μM Tn5059 was added to each well. Next, 1 μL of 8.4 μM annealed barcoded r5 transposon was added to each well and mixed, with one uniquely barcoded r5 transposon per well. Activity normalization performed by Illumina found twice the transposon concentration to Tn5059 concentration was needed for comparable activity, due to a longer transposon being used. The 384 well plate was then incubated at room temperature for 30 minutes, and was then ready for tagmentation and combinatorial indexing.

scTHS-seq custom i7 transposome complex generation

For custom nXTv2_i7 Tn5059 transposome generation, first annealed transposon at 50 μM was generated as described above with the nXTv2_i7 transposon (**Table 17)**. To generate a complexed transposome solution of 1.75 μM Tn5059 and 2.5 μM annealed transposon, 3.5 μM Tn5059 was incubated with 5 μM annealed transposon for 30 minutes at room temperature. Then the complexed transposome solution was diluted to 0.7 μM with standard storage buffer provided by Illumina, and was now ready for tagmentation. Transposome complexes were stored at -20ºC and used within a week.

scTHS-seq nuclei tagmentation and barcoding

To the 384 well plate of freshly generated uniquely barcoded Tn5059 r5 transposome complexes, 4 μL of human/mouse mixed cell sample was added for

a total of ~960 nuclei/well and final concentration of 0.7 μM Tn5059 r5 transposome complex. Each sample was mixed gently 5X by the electronic pipettor. Optimally ~2000 nuclei/well are wanted, however nuclei yields were low. The 384 well plate was then incubated at 37ºC for 30 minutes. To stop the reaction, 4.0 μL of 50 mM EDTA was added to each well and mixed gently 5X by the electronic pipettor for a final concentration of 20 mM EDTA in 10 uL, followed by incubating the 384 well plate at 37ºC for 15 minutes. The 384 well plate of samples was then frozen at -20ºC overnight. The next day, the 384 well plate of samples was thawed, and one volume of cold 2X FACS buffer (1X FACS buffer: 2 mM EDTA, 1% BSA in PBS) was added to each well. Samples are then mixed gently 3X by the electronic pipettor and pooled into one tube on ice, which was spun down at 500xg for 5 minutes at 4ºC. Supernatant was removed, and tagmented nuclei resuspended in 1.5 mL cold 1X FACS buffer. Next, 75 μL eBioscience Propidium Iodide (PI) was added to the tube, and gently mixed. Nuclei were then sorted by flow cytometry into 96 well plates containing 10 μL PBS/well at 100 nuclei/well, and put on ice. Doublets were removed based on forward and side scatter plots, and PI stained events selected. Nuclei were now ready for RNA amplification and processing.

scTHS-seq library preparation

Each 96 well plate of nuclei was processed individually. First, 11 μL guanidine hydrochloride was added to each well and mixed by lightly vortexing. Reactions were purified by addition of 40 μL (1.8X) AMPure SPRI beads and lightly

vortexed, then bead pelleting and 80% ethanol washes were performed with the “flick and blot” method and magnetic plate from V&P Scientific. After 80% ethanol washes were complete the plate was quickly spun down at 500xg and leftover 80% ethanol removed by pipetting. 10 µL 1X NEB Taq polymerase was added to each reaction and the plate was lightly vortexed to resuspend the beads (SPRI beads left in reaction), followed by running the reactions at 72ºC for 3 minutes for end fill in and put on ice. For *in vitro* transcription (IVT) amplification the NEB HiScribe T7 high yield synthesis kit was used. To each reaction a mastermix of 2 µL of 10X transcription buffer, 2 µL ATP, 2 µL CTP, 2 µL GTP, 2 µL UTP was directly added to the end fill in reactions, lightly vortexed, and incubated at 37ºC for 19 hours. To check IVT amplification, two full reactions were purified with SPRI binding buffer, run on a 6% Tris-Borate-Urea (TBU) gel for 20 minutes at 250 volts followed by SYBR gold staining and visualization on a UV-transilluminator. Next, reactions were purified by addition of 44 µL (2.0X) SPRI binding buffer (20% PEG 8000, 2.5 M NaCl, 10 mM Tris–HCl, 1 mM EDTA) to each reaction and the plate vortexed thoroughly. 80% ethanol washes and leftover 80% ethanol removal were performed as describe above, and SPRI beads were resuspended in 9 µL nuclease free water. For reverse transcription, first 2.5 µL of 20 µM random hexamers was added to each reaction, the plate vortexed lightly and then heated to 70ºC for 3 minutes, and immediately cooled on ice. Then Clontech SMART® MMLV Reverse Transcriptase kit was used with addition of 4 µL 5X first strand synthesis, 2 µL dNTP mix, 2 µL 100 mM DTT, 0.5 µL SMART MMLV RT in a mastermix to each reaction and the plate vortexed

lightly. Reactions were incubated at 22ºC for 10 minutes, then 42ºC for 60 minutes, and terminated at 70ºC for 10 minutes. To degrade RNA in cDNA-RNA hybrids 1 µL of 0.5 units Enzymatics RNase H was added to each reaction, the plate was vortexed lightly and incubated at 37ºC for 20 minutes. For second strand synthesis first 2.5 µL of 20 µM sss_scnXTv2 (Table 17) was added to each reaction and lightly vortexed, then incubated for 2 minutes at 65ºC and immediately cooled on ice. Then 5.9 µL of NEB taq5X was added to each reaction and incubated at 72ºC for 8 minutes. After cooling on ice, 60 µL (2.0X) SPRI binding buffer was added to each sample. The plate was vortexed thoroughly and 80% ethanol washes and leftover 80% ethanol removal performed as described previously, and SPRI beads resuspended by light vortexing in 7 µL of nuclease free water. Double stranded cDNA fragments then undergo simultaneous fragmentation and 3' adaptor addition with a custom nXTv2_i7 Tn5059 transposome (Table 17). To 7 µL of sample, 2 µL of 5X tagmentation buffer was added to each sample, followed by addition of 2 µL of prepared 0.7 µM custom nXTv2_i7 Tn5059 transposomes (final transposome concentration of 0.14 µM), vortexed lightly, then incubated at 55ºC for 6 minutes, and cooled briefly on ice. Immediately after cooling, 19 µL of 6.32M guanidine hydrochloride, for a final guanidine hydrochloride concentration of 4M was added to each reaction and briefly vortexed. Then 60 µL (2.0X) SPRI binding buffer was added to each sample. The plate was vortexed thoroughly and 80% ethanol washes and leftover 80% ethanol removal was performed as described previously. However, for this purification SPRI beads were resuspended in 16 µL

nuclease free water and the plate was placed back onto the magnetic plate. Sample was eluted off SPRI beads held by the magnetic plate, and transferred to a qPCR plate. Standard Illumina Nextera XT v2 barcoding in an 8x12 (i5xi7) format was performed with qPCR, using custom scTHS-seq i5 indexes and standard Illumina i7 indexes (Table 17). In total, 20 µL of KAPA SYBR Fast, 2 µL of 10 µM scT7_S5XX index primer and 2 µL of 10 µM nXTv2_i7XX index primer were added to each reaction for a total volume of 40 µL, and mixed well. qPCR was run at 72ºC for 3 minutes, 95ºC for 30 seconds, followed by cycling for (95ºC for 10 seconds, 63ºC for 30 seconds, 72ºC for 1 minute) until curves reach saturation, typically 9-12 cycles. Plates were stored at -20ºC.

scTHS-seq library validation, pooling and sequencing

To validate libraries, 1 µL of each qPCR reaction was run on 6% Tris-borate-edta (TBE) gels at 250 volts for 20 minutes, staining with SYBR Gold, and visualizing smears with a UV-transilluminator. For pooling, 2 µL of each of each uniquely barcode qPCR reaction was added to one tube. If smears on TBE gels were visibly significantly lighter, then 4 µL or 6 µL of the reaction was added to the pooled sample. After all samples had been added, the pooling tube was mixed thoroughly. For gel size selection a TBE gel of the pooled sample was run at 250 volts for 20 minutes and size selected from 220-1000 base pairs with a Dark Reader blue light transilluminator, and scalpel. Selected bands were sheared by poking a hole in the bottom of a 0.5 mL tube with a 22-gauge needle, placing the gel piece inside then placing the 0.5 mL tube into a 2.0 mL tube, and centrifuging

in a bench top centrifuge at 21,000xg for 1 minute. Next the 0.5 mL tube was removed, 450 µL 1X TE added then shaken vigorously at 37ºC for 45 minutes. The supernatant and all possible retrievable liquid was pipetted into Nanosep 0.2 µM filter tubes and centrifuged at 21,000xg for 1 minute. Collected sample in the nanosep tube then underwent ethanol precipitation, with 40 µL NaOAc, 1.3 µL Ambion GlycoBlue, and 1 mL 100% Ethanol added to each tube. Samples were mixed gently and put at -80ºC for 15 minutes, followed by centrifuging at 12,000xg at 4ºC for 20 minutes. Supernatent was discarded and 700 µL ice cold 75% ethanol added. Sample was inverted for mixing, then centrifuged at 18,000xg for 5 minutes at room temperature. Supernatent was discarded, leftover ethanol pipeted out, pellet dried, and pellet resuspended in 5 µL nuclease free water. Resultant size selected libraries were quantified with Qubit, sequenced on Illumina MiSeq (50x32x32) single end reads for validation, then Illumina HiSeq 2500 high throughput (50x8x32) single end reads for data generation.

scTHS-seq data processing

Raw BCL files were demultiplexed to fastq files Read1, Index1, and Index2 files using bcl2fastq v2.17.1.14, then used as input to deindexder (https://github.com/ws6/deindexer) with 0 mismatch barcode demultiplexing. Barcode combinations associated with each read were appended to each reads header with in house perl scripts and all fastq's were combined and mapped to a hg38 no alternative loci plus decoy reference genome

(GCA_000001405.15_GRCh38_no_alt_plus_hs38d1_analysis_set) and mm10 no alternative loci reference genome (GCA_000001635.5_GRCm38.p3_no_alt_analysis_set) using BWA 0.7.12-r1039. Mapped sam outputs were redemultiplexed by barcodes, converted to bam files and clonal reads removed with samtools 1.3.1, while gathering read statistics for each barcode combination. To determine which uniquely barcoded nuclei were suitable for downstream analysis, several dataset filtering steps were implemented. First, all multimapping reads MAPQ<5 were removed from each single cell dataset to remove reads in repeat regions, and to remove biases before clustering. Second, nuclei were filtered requiring log10(total reads + 1) > 3. To better focus on potential microglial cells, an additional analysis without multimapping reads removed was also conducted with the same nuclei filtering criteria.

Joint peak calling was performed on pooled bam files using SPP (v1.13; https://github.com/hms-dbmi/spp). In total, 14781 cells were pooled. Reads mapping within 100bps of known repeat regions according to annotations from Repeat Masker (http://www.repeatmasker.org) were removed. A smoothed density of pooled reads was generated using window tag counts with a window size of 500bps and a window step of 100bps. DNA accessible regions (peaks) were called based on the smoothed density with a minimum threshold of 5 reads and minimum span of 5 steps between each peak. Peaks were filtered using a permutation based FDR of 10e-8 and filtered for presence in at least 40 cells from the visual cortex, resulting in 52694 final peaks called (**Table 13**). Called

peaks were then assessed for reads in each individual cell to generate a matrix of peaks versus cells for downstream clustering and analysis.

scTHS-seq Data Clustering

The molecular count matrix was binarized for further analysis. 52,694 sites that were observed in 30 or more cells were selected. The data was modeled as a right-censored Poisson process (observing at most 1 molecule per site). To determine cell depth and batch-specific site observation probabilities, an EM algorithm was used, with each iteration fitting MLE values for library size and batch-specific site probabilities sequentially. To evaluate over-dispersion of each site, total deviance was calculated across all observations was calculated for a given site under the censored Poisson process. The relationship between the total deviance and mean site occurrence frequencies was modeled using generalized additive model (mgcv R package, smoothing term k=10). The observed / expected deviance difference was scored using variance gamma distribution.

To cluster and visualize the cells in the visual cortex, top 30 PCs were determined on the censored Poisson deviance residual matrix. The negative deviance residuals associated with non-observed sites were ignored. Cell clusters were determined on a k-nearest neighbor graph (k=50) using multilevel community detection method (igraph R package). Cell clusters with fewer than 30 cells were omitted from further analysis. Upon inspection, three smallest clusters appeared to represent poorly-resolved cells or doublets mixing signals from two

or more subpopulations and were also not examined in further analysis. Two dimensional visualization was achieved by applying t-Distributed Stochastic Neighbor Embedding (t-SNE) on the 30 PCs using a perplexity of 50.

The 30 PCs derived from the visual cortex were also used to project cells from the frontal cortex. Due to there being fewer cells in the frontal cortex dataset than the visual cortex dataset, a smaller k=30 and perplexity of 30 was used for the k-nearest neighbor graph and t-SNE embedding respectively.

scTHS-seq collision rate determination

For each unique barcode combination, the proportion of unique reads that map to either mouse or human genome was calculated. A unique barcode combination was determined to belong to one species if 89% of the reads mapped to one genome, otherwise the barcode combination was determined to be a detectable collision. For visualization, results were then graphed in R using a X,Y scatter plot and density plot.

scTHS-Seq and snDrop-Seq joint analysis

To map between transcriptional and epigenetic space we trained a generalized boosted regression model (GBM) to predict the probability of differential expression from patterns of nearby accessibility differences, and a separate GBM in reverse, predict probability of differential accessibility given the differential expression observations. GBM was implemented using the caret (V6.0-72) package in R. The differential expression prediction GBM utilized the

following features: 1) maximum differential accessibility Z score observed within 5kb of the transcription start site; 2) maximum differential accessibility Z score of any site located within the gene or within 3kb of the gene boundaries; 3) number of differentially accessible (Z>5) sites within 3kb of the gene; 4) number of non-differentially accessible sites within the 3kb of the gene (Z<1). The differential site prediction GBM was based on the following features: 1) distance to closest transcription start site; 2) expression difference Z score of the genes with TSS within 5kb of the site; 3) max log2 fold expression difference of the gene with TSS within 5kb of the site; 4) minimum distance to a gene (0 if the site is overlapped by a gene); 4) maximum Z score of differential expression of any gene within 3kb of the site. Models were trained on Astrocyte and Oligodendrocyte data from the visual cortex only to learn relevance of features. Models were fit using 10x cross-validation. Joint scores (across multiple genes or sites) were calculated as probability means of individual elements (sites or genes). In the first round of mapping, differentially accessible sites for each major cell type including Ast, Oli, Ex, In, as well as Mic, Oli, and End combined with Per (when available) were identified using a Fisher's exact test. Joint scores across resulting predicted differentially expressed genes were then derived for each major cell type for each cell. ROC curves and AUC were computed by comparing the resulting joint scores with original annotations. In additional rounds of mapping, we restricted to cells predicted to be either Ex or In in order to identify additional excitatory and inhibitory subtypes.

scTHS-seq transcription factor analysis

To infer relevant transcription factors (TFs) and transcription factor binding sites (TFBSs), we obtained DNA sequences corresponding to scTHS-seq peaks and position weight matrices (PWMs) 379 TFs from the JASPAR database. A sliding window was used to identify the maximum PWM score for each peak, taking into consideration both the plus and minus strands. PWM scores within each peak were normalized by subtracting the theoretical minimum and dividing by the maximum score possible for each PWM using the PWMscoreStartingAt() function from the matchPWM package in R assuming a uniform prior distribution on all nucleotides. Scores for each peak as well as TF were then standardized to Z-scores by subtracting the mean and dividing by the standard deviation of scores for each TF to control for background rates of binding and non-specificity. TFBSs were inferred as corresponding to peaks with Z-score > 1.96 for each TF. We assess the overlap of inferred TFBSs with previously identified cell type specific peaks using a Fisher's exact test. TFs with TFBSs significantly overlapping with cell type specific peaks (Bonferroni corrected P-value < 0.2) were inferred to be relevant to the cell type. We integrate snDrop-Seq data to assess the expression fold change of these TFs in each cell type, assessing significance by using rank-based gene set enrichment analysis (GSEA). Specifically, TF expression was averaged across cells for each cell type. A log2 fold change comparing the average expression in oligodendrocyte versus neuronal cell types was used to assess for enrichment of expression for predicted oligodendrocyte-related TFs.

GSEA was performed using the LIGER package in R (https://github.com/JEFworks/liger).

scTHS-seq GWAS data analysis

GWAS SNPs were downloaded from the GRASP database, using categories with any trait for selection with p-values $<1x10^{-6}$. Categories that were selected Alzheimer's disease, Schizophrenia, Parkinson's disease, Bipolar disorder, Autism, Multiple sclerosis, ADHD, ALS, Epilepsy, Depression, Glaucoma, Crohn's disease, Celiac disease, Type I diabetes, Lung disease, Chronic kidney disease, Prostate cancer. For the rest of the analysis in house python and shells scripts were used. For each category all SNPs were extended at each end to encompass a 100 KB region. Any SNP regions overlapping each other were merged with bedtools to generate a larger SNP region containing both SNPs. Next, for each SNP region, the most significant p-value SNP was selected. This removed any multiple instances of linked variants for the same trait, and ensured there were no variants in linkage disequilibrium. Next, the top 50 most significant p-value SNPs and their gene regions were selected for further analysis. To determine overlap of accessible regions in each cluster defined during cell clustering and identification, first peak calling with SPP v1.2 was performed on merged data of each cluster to generate list of peak regions, and then lists of differential peaks (peaks present in one or more cell types and not others) were generated for all the cell types. Peaks with Z-scores <400 were removed to generate a final peak list for each cluster. Next, those peaks were overlapped

with the top 50 SNP regions for each disease category and the number of overlaps counted. To determine if enrichment was significant, Z-scores were calculated. First, 20,000 permutations of the peak regions over the hg38 reference genome using only autosomes was performed, and for each permutation overlaps of peaks for each cluster with SNP regions was counted. From the permutations, averages and standard deviation were calculated, and in conjunction with previously calculated total overlaps, the Z-score for each cluster was calculated. For visualization, R was used to overlap Z-scores onto the clusters, and generate a heat map of similarity between cell types and diseases.

REFERENCES

1. Sabo, P.J., Humbert, R., Hawrylycz, M., Wallace, J.C., Dorschner, M.O., McArthur, M. & Stamatoyannopoulos, J.A. Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 4537-4542 (2004).

2. Sabo, P.J., Kuehn, M.S., Thurman, R., Johnson, B.E., Johnson, E.M., Cao, H., Yu, M., Rosenzweig, E., Goldy, J., Haydock, A., Weaver, M., Shafer, A., Lee, K., Neri, F., Humbert, R., Singer, M.A., Richmond, T.A., Dorschner, M.O., McArthur, M., Hawrylycz, M., Green, R.D., Navas, P.A., Noble, W.S. & Stamatoyannopoulos, J.A. Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nature methods* **3**, 511-518 (2006).

3. Song, L. & Crawford, G.E. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor protocols* **2010**, pdb prot5384 (2010).

4. Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R. & Lieb, J.D. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome research* **17**, 877-885 (2007).

5. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213-1218 (2013).

6. Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T.K., Diegel, M., Dunn, D., Ebersol, A.K., Frum, T., Giste, E., Johnson, A.K., Johnson, E.M., Kutyavin, T., Lajoie, B., Lee, B.K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E.D., Qu, H., Reynolds, A.P., Roach, V., Safi, A., Sanchez, M.E., Sanyal, A., Shafer, A., Simon, J.M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M.O., Hansen, R.S., Navas, P.A., Stamatoyannopoulos, G., Iyer, V.R., Lieb, J.D., Sunyaev, S.R., Akey, J.M., Sabo, P.J., Kaul, R., Furey, T.S., Dekker, J., Crawford, G.E. & Stamatoyannopoulos, J.A. The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82 (2012).

7. John, S., Sabo, P.J., Thurman, R.E., Sung, M.H., Biddie, S.C., Johnson, T.A., Hager, G.L. & Stamatoyannopoulos, J.A. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature genetics* **43**, 264-268 (2011).

8. Neph, S., Stergachis, A.B., Reynolds, A., Sandstrom, R., Borenstein, E. & Stamatoyannopoulos, J.A. Circuitry and dynamics of human transcription factor regulatory networks. *Cell* **150**, 1274-1286 (2012).

9. Valouev, A., Johnson, S.M., Boyd, S.D., Smith, C.L., Fire, A.Z. & Sidow, A. Determinants of nucleosome organization in primary human cells. *Nature* **474**, 516-520 (2011).

10. Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman, R.E., John, S., Sandstrom, R., Johnson, A.K., Maurano, M.T., Humbert, R., Rynes, E., Wang, H., Vong, S., Lee, K., Bates, D., Diegel, M., Roach, V., Dunn, D., Neri, J., Schafer, A., Hansen, R.S., Kutyavin, T., Giste, E., Weaver, M., Canfield, T., Sabo, P., Zhang, M., Balasundaram, G., Byron, R., MacCoss, M.J., Akey, J.M., Bender, M.A., Groudine, M., Kaul, R. & Stamatoyannopoulos, J.A. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83-90 (2012).

11. Segal, E. & Widom, J. Poly(dA:dT) tracts: major determinants of nucleosome organization. *Current opinion in structural biology* **19**, 65-71 (2009).

12. Natarajan, A., Yardimci, G.G., Sheffield, N.C., Crawford, G.E. & Ohler, U. Predicting cell-type-specific gene expression from regions of open chromatin. *Genome research* **22**, 1711-1722 (2012).

13. Furey, T.S. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nature reviews. Genetics* **13**, 840-852 (2012).

14. Song, L., Zhang, Z., Grasfeder, L.L., Boyle, A.P., Giresi, P.G., Lee, B.K., Sheffield, N.C., Graf, S., Huss, M., Keefe, D., Liu, Z., London, D., McDaniell, R.M., Shibata, Y., Showers, K.A., Simon, J.M., Vales, T., Wang, T., Winter, D., Zhang, Z., Clarke, N.D., Birney, E., Iyer, V.R., Crawford, G.E., Lieb, J.D. & Furey, T.S. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome research* **21**, 1757-1767 (2011).

15. Cusanovich, D.A., Daza, R., Adey, A., Pliner, H.A., Christiansen, L., Gunderson, K.L., Steemers, F.J., Trapnell, C. & Shendure, J. Epigenetics. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910-914 (2015).

16. Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y. & Greenleaf, W.J. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486-490 (2015).

17. Jin, W., Tang, Q., Wan, M., Cui, K., Zhang, Y., Ren, G., Ni, B., Sklar, J., Przytycka, T.M., Childs, R., Levens, D. & Zhao, K. Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature* **528**, 142-146 (2015).

18. Adey, A., Morrison, H.G., Asan, Xun, X., Kitzman, J.O., Turner, E.H., Stackhouse, B., MacKenzie, A.P., Caruccio, N.C., Zhang, X. & Shendure, J. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome biology* **11**, R119 (2010).

19. Caruccio, N. Preparation of next-generation sequencing libraries using Nextera technology: simultaneous DNA fragmentation and adaptor tagging by in vitro transposition. *Methods in molecular biology* **733**, 241-255 (2011).

20. Picelli, S., Bjorklund, A.K., Reinius, B., Sagasser, S., Winberg, G. & Sandberg, R. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome research* **24**, 2033-2040 (2014).

21. Naumann, T.A. & Reznikoff, W.S. Tn5 transposase active site mutants. *The Journal of biological chemistry* **277**, 17623-17629 (2002).

22. Goryshin, I.Y. & Reznikoff, W.S. Tn5 in vitro transposition. *The Journal of biological chemistry* **273**, 7367-7374 (1998).

23. Amini, S., Pushkarev, D., Christiansen, L., Kostem, E., Royce, T., Turk, C., Pignatelli, N., Adey, A., Kitzman, J.O., Vijayan, K., Ronaghi, M., Shendure, J., Gunderson, K.L. & Steemers, F.J. Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nature genetics* **46**, 1343-1349 (2014).

24. Vitak, S.A., Torkenczy, K.A., Rosenkrantz, J.L., Fields, A.J., Christiansen, L., Wong, M.H., Carbone, L., Steemers, F.J. & Adey, A. Sequencing thousands of single-cell genomes with combinatorial indexing. *Nature methods* **14**, 302-308 (2017).

25. Ramani, V., Deng, X., Qiu, R., Gunderson, K.L., Steemers, F.J., Disteche, C.M., Noble, W.S., Duan, Z. & Shendure, J. Massively multiplex single-cell Hi-C. *Nature methods* **14**, 263-266 (2017).

26. Cusanovich, D.A., Daza, R., Adey, A., Pliner, H.A., Christiansen, L., Gunderson, K.L., Steemers, F.J., Trapnell, C. & Shendure, J. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910-914 (2015).

27. Corces, M.R., Buenrostro, J.D., Wu, B., Greenside, P.G., Chan, S.M., Koenig, J.L., Snyder, M.P., Pritchard, J.K., Kundaje, A., Greenleaf, W.J., Majeti, R. & Chang, H.Y. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature genetics* **48**, 1193-1203 (2016).

28. Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., Amin, V., Whitaker, J.W., Schultz, M.D., Ward, L.D., Sarkar, A., Quon, G., Sandstrom, R.S., Eaton, M.L., Wu, Y.C., Pfenning, A.R., Wang, X., Claussnitzer,

M., Liu, Y., Coarfa, C., Harris, R.A., Shoresh, N., Epstein, C.B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R.D., Lister, R., Hong, C., Gascard, P., Mungall, A.J., Moore, R., Chuah, E., Tam, A., Canfield, T.K., Hansen, R.S., Kaul, R., Sabo, P.J., Bansal, M.S., Carles, A., Dixon, J.R., Farh, K.H., Feizi, S., Karlic, R., Kim, A.R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T.R., Neph, S.J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R.C., Siebenthall, K.T., Sinnott-Armstrong, N.A., Stevens, M., Thurman, R.E., Wu, J., Zhang, B., Zhou, X., Beaudet, A.E., Boyer, L.A., De Jager, P.L., Farnham, P.J., Fisher, S.J., Haussler, D., Jones, S.J., Li, W., Marra, M.A., McManus, M.T., Sunyaev, S., Thomson, J.A., Tlsty, T.D., Tsai, L.H., Wang, W., Waterland, R.A., Zhang, M.Q., Chadwick, L.H., Bernstein, B.E., Costello, J.F., Ecker, J.R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J.A., Wang, T. & Kellis, M. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330 (2015).

29. Ackermann, A.M., Wang, Z., Schug, J., Naji, A. & Kaestner, K.H. Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. *Molecular metabolism* **5**, 233-244 (2016).

30. Wu, J., Huang, B., Chen, H., Yin, Q., Liu, Y., Xiang, Y., Zhang, B., Liu, B., Wang, Q., Xia, W., Li, W., Li, Y., Ma, J., Peng, X., Zheng, H., Ming, J., Zhang, W., Zhang, J., Tian, G., Xu, F., Chang, Z., Na, J., Yang, X. & Xie, W. The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature* **534**, 652-657 (2016).

31. Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A. & Kirschner, M.W. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187-1201 (2015).

32. Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., Trombetta, J.J., Weitz, D.A., Sanes, J.R., Shalek, A.K., Regev, A. & McCarroll, S.A. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).

33. Svensson, V., Natarajan, K.N., Ly, L.H., Miragaia, R.J., Labalette, C., Macaulay, I.C., Cvejic, A. & Teichmann, S.A. Power analysis of single-cell RNA-sequencing experiments. *Nature methods* **14**, 381-387 (2017).

34. Leslie, R., O'Donnell, C.J. & Johnson, A.D. GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics* **30**, i185-194 (2014).

35. Beck, T., Hastings, R.K., Gollapudi, S., Free, R.C. & Brookes, A.J. GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. *European journal of human genetics : EJHG* **22**, 949-952 (2014).

36. Price, A.L., Spencer, C.C. & Donnelly, P. Progress and promise in understanding the genetic basis of common diseases. *Proceedings. Biological sciences* **282**, 20151684 (2015).

37. Tabor, S. & Richardson, C.C. A bacteriophage T7 RNA polymerase/promoter system for controlled exclusive expression of specific genes. 1985. *Biotechnology* **24**, 280-284 (1992).

38. Hoeijmakers, W.A., Bartfai, R., Francoijs, K.J. & Stunnenberg, H.G. Linear amplification for deep sequencing. *Nature protocols* **6**, 1026-1036 (2011).

39. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep* **2**, 666-673 (2012).

40. Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A. & Amit, I. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776-779 (2014).

41. Kumar, V., Muratani, M., Rayan, N.A., Kraus, P., Lufkin, T., Ng, H.H. & Prabhakar, S. Uniform, optimal signal processing of mapped deep-sequencing data. *Nature biotechnology* **31**, 615-622 (2013).

42. Rosenbloom, K.R., Sloan, C.A., Malladi, V.S., Dreszer, T.R., Learned, K., Kirkup, V.M., Wong, M.C., Maddren, M., Fang, R., Heitner, S.G., Lee, B.T., Barber, G.P., Harte, R.A., Diekhans, M., Long, J.C., Wilder, S.P., Zweig, A.S., Karolchik, D., Kuhn, R.M., Haussler, D. & Kent, W.J. ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic acids research* **41**, D56-63 (2013).

43. McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M. & Bejerano, G. GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology* **28**, 495-501 (2010).

44. Carter, K.L., Cahir-McFarland, E. & Kieff, E. Epstein-barr virus-induced changes in B-lymphocyte gene expression. *J Virol* **76**, 10427-10436 (2002).

45. Ersing, I., Bernhardt, K. & Gewurz, B.E. NF-kappaB and IRF7 pathway activation by Epstein-Barr virus Latent Membrane Protein 1. *Viruses* **5**, 1587-1606 (2013).

46. Iskra, S., Kalla, M., Delecluse, H.J., Hammerschmidt, W. & Moosmann, A. Toll-like receptor agonists synergistically increase proliferation and activation of B cells by epstein-barr virus. *J Virol* **84**, 3612-3623 (2010).

47. Vierstra, J., Wang, H., John, S., Sandstrom, R. & Stamatoyannopoulos, J.A. Coupling transcription factor occupancy to nucleosome architecture with DNase-FLASH. *Nature methods* **11**, 66-72 (2014).

48. Ameur, A., Zaghlool, A., Halvardson, J., Wetterbom, A., Gyllensten, U., Cavelier, L. & Feuk, L. Total RNA sequencing reveals nascent transcription and

widespread co-transcriptional splicing in the human brain. *Nat Struct Mol Biol* **18**, 1435-1440 (2011).

49. Zhang, Y., Chen, K., Sloan, S.A., Bennett, M.L., Scholze, A.R., O'Keeffe, S., Phatnani, H.P., Guarnieri, P., Caneda, C., Ruderisch, N., Deng, S., Liddelow, S.A., Zhang, C., Daneman, R., Maniatis, T., Barres, B.A. & Wu, J.Q. An RNA-sequencing transcriptome and splicing database of glia, neurons, and vascular cells of the cerebral cortex. *J Neurosci* **34**, 11929-11947 (2014).

50. Zhang, Y., Sloan, S.A., Clarke, L.E., Caneda, C., Plaza, C.A., Blumenthal, P.D., Vogel, H., Steinberg, G.K., Edwards, M.S., Li, G., Duncan, J.A., 3rd, Cheshier, S.H., Shuer, L.M., Chang, E.F., Grant, G.A., Gephart, M.G. & Barres, B.A. Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. *Neuron* **89**, 37-53 (2016).

51. Tasic, B., Menon, V., Nguyen, T.N., Kim, T.K., Jarsky, T., Yao, Z., Levi, B., Gray, L.T., Sorensen, S.A., Dolbeare, T., Bertagnolli, D., Goldy, J., Shapovalova, N., Parry, S., Lee, C., Smith, K., Bernard, A., Madisen, L., Sunkin, S.M., Hawrylycz, M., Koch, C. & Zeng, H. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat Neurosci* **19**, 335-346 (2016).

52. Darmanis, S., Sloan, S.A., Zhang, Y., Enge, M., Caneda, C., Shuer, L.M., Hayden Gephart, M.G., Barres, B.A. & Quake, S.R. A survey of human brain transcriptome diversity at the single cell level. *Proc Natl Acad Sci U S A* **112**, 7285-7290 (2015).

53. Lake, B.B., Ai, R., Kaeser, G.E., Salathia, N.S., Yung, Y.C., Liu, R., Wildberg, A., Gao, D., Fung, H.L., Chen, S., Vijayaraghavan, R., Wong, J., Chen, A., Sheng, X., Kaper, F., Shen, R., Ronaghi, M., Fan, J.B., Wang, W., Chun, J. & Zhang, K. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* **352**, 1586-1590 (2016).

54. Shekhar, K., Lapan, S.W., Whitney, I.E., Tran, N.M., Macosko, E.Z., Kowalczyk, M., Adiconis, X., Levin, J.Z., Nemesh, J., Goldman, M., McCarroll, S.A., Cepko, C.L., Regev, A. & Sanes, J.R. Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell* **166**, 1308-1323 e1330 (2016).

55. La Manno, G., Gyllborg, D., Codeluppi, S., Nishimura, K., Salto, C., Zeisel, A., Borm, L.E., Stott, S.R., Toledo, E.M., Villaescusa, J.C., Lonnerberg, P., Ryge, J., Barker, R.A., Arenas, E. & Linnarsson, S. Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell* **167**, 566-580 e519 (2016).

56. Zeng, H., Shen, E.H., Hohmann, J.G., Oh, S.W., Bernard, A., Royall, J.J., Glattfelder, K.J., Sunkin, S.M., Morris, J.A., Guillozet-Bongaarts, A.L., Smith, K.A., Ebbert, A.J., Swanson, B., Kuan, L., Page, D.T., Overly, C.C., Lein, E.S., Hawrylycz, M.J., Hof, P.R., Hyde, T.M., Kleinman, J.E. & Jones, A.R. Large-scale cellular-resolution gene profiling in human neocortex reveals species-specific molecular signatures. *Cell* **149**, 483-496 (2012).

57. Mariani, J., Coppola, G., Zhang, P., Abyzov, A., Provini, L., Tomasini, L., Amenduni, M., Szekely, A., Palejev, D., Wilson, M., Gerstein, M., Grigorenko, E.L., Chawarska, K., Pelphrey, K.A., Howe, J.R. & Vaccarino, F.M. FOXG1-Dependent Dysregulation of GABA/Glutamate Neuron Differentiation in Autism Spectrum Disorders. *Cell* **162**, 375-390 (2015).

58. Liu, S.J., Nowakowski, T.J., Pollen, A.A., Lui, J.H., Horlbeck, M.A., Attenello, F.J., He, D., Weissman, J.S., Kriegstein, A.R., Diaz, A.A. & Lim, D.A. Single-cell analysis of long non-coding RNAs in the developing human neocortex. *Genome Biol* **17**, 67 (2016).

59. Collins, C.E., Airey, D.C., Young, N.A., Leitch, D.B. & Kaas, J.H. Neuron densities vary across and within cortical areas in primates. *Proc Natl Acad Sci U S A* **107**, 15927-15932 (2010).

60. Sos, B.C., Fung, H.L., Gao, D.R., Osothprarop, T.F., Kia, A., He, M.M. & Zhang, K. Characterization of chromatin accessibility with a transposome hypersensitive sites sequencing (THS-seq) assay. *Genome Biol* **17**, 20 (2016).

61. Kia, A., Gloeckner, C., Osothprarop, T., Gormley, N., Bomati, E., Stephenson, M., Goryshin, I. & He, M.M. Improved genome sequencing using an engineered transposase. *BMC Biotechnol* **17**, 6 (2017).

62. Hansen, D.V., Lui, J.H., Flandin, P., Yoshikawa, K., Rubenstein, J.L., Alvarez-Buylla, A. & Kriegstein, A.R. Non-epithelial stem cells and cortical interneuron production in the human ganglionic eminences. *Nat Neurosci* **16**, 1576-1587 (2013).

63. Ma, T., Wang, C., Wang, L., Zhou, X., Tian, M., Zhang, Q., Zhang, Y., Li, J., Liu, Z., Cai, Y., Liu, F., You, Y., Chen, C., Campbell, K., Song, H., Ma, L., Rubenstein, J.L. & Yang, Z. Subcortical origins of human and monkey neocortical interneurons. *Nat Neurosci* **16**, 1588-1597 (2013).

64. Mathelier, A., Fornes, O., Arenillas, D.J., Chen, C.Y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R., Zhang, A.W., Parcy, F., Lenhard, B., Sandelin, A. & Wasserman, W.W. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **44**, D110-115 (2016).

65. Lu, Q.R., Sun, T., Zhu, Z., Ma, N., Garcia, M., Stiles, C.D. & Rowitch, D.H. Common developmental requirement for Olig function indicates a motor neuron/oligodendrocyte connection. *Cell* **109**, 75-86 (2002).

66. Zhou, Q. & Anderson, D.J. The bHLH transcription factors OLIG2 and OLIG1 couple neuronal and glial subtype specification. *Cell* **109**, 61-73 (2002).

67. Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B.E., Liu, X.S. & Raychaudhuri, S. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet* **45**, 124-130 (2013).

68. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M. & Bernstein, B.E. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43-49 (2011).

69. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A.K., Canfield, T.K., Giste, E., Diegel, M., Bates, D., Hansen, R.S., Neph, S., Sabo, P.J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S.R., Kaul, R. & Stamatoyannopoulos, J.A. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-1195 (2012).

70. Fan, H.C., Fu, G.K. & Fodor, S.P. Expression profiling. Combinatorial labeling of single cells for gene expression cytometry. *Science* **347**, 1258367 (2015).

71. Mattson, M.P. Pathways towards and away from Alzheimer's disease. *Nature* **430**, 631-639 (2004).

72. Edgar, R., Domrachev, M. & Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research* **30**, 207-210 (2002).