PNAS
www.pnas.org

Supplementary Information for

Sub-megabase copy number variations arise during cerebral cortical neurogenesis as revealed by novel single-cell sequencing approach

Suzanne Rohrback, Craig April, Fiona Kaper, Richard R. Rivera, Christine S. Liu, Benjamin Siddoway, Jerold Chun

Jerold Chun
Email: jchun@sbpdiscovery.org

This PDF file includes:

Supplementary text
Captions for datasets S1 to S5
References for SI reference citations
Figs. S1 to S6

Other supplementary materials for this manuscript include the following:

Datasets S1 to S5

Supplementary Information Text

Extended Experimental Methods

Resources. Fluorescence activated cell sorting (FACS) instruments used for FANS was performed at The Scripps Research Institute (TSRI) flow cytometry core. Sequencing was performed at the TSRI next generation sequencing core. Computational pipeline development and analysis was performed on the Triton Shared Computing Cluster (TSCC) at the San Diego Supercomputer Center (SDSC; University of California, San Diego). Sequencing data are available on the NCBI Short Read Archive (BioProject PRJNA415480). Code for CNV identification, including the development and implementation of FUnC, is available on GitHub (https://github.com/suzannerohrback/somaticCNVpipeline).

Single nucleus isolation and staining. All tissue samples were obtained from C57BL/6J mice and flash-frozen in liquid nitrogen after dissection and stored at -80°C until nuclei preparation. Samples were thawed on ice in fresh nuclei extraction buffer (NEB: 0.32 M sucrose, 5 mM $CaCl_2$, 3 mM $Mg(Ac)_2$, 0.1 mM EDTA, 0.1% TritonX-100, 10 mM Tris-HCl, pH 8) for 20 minutes, then homogenized and filtered (50 micron pores) into a collection tube. Nuclei were pelleted by centrifugation 5 minutes x 1,800 rpm, resuspended in PBSE-BSA (1x PBS, 2 mM EGTA, 1% BSA), and incubated with shaking on ice for at least 20 minutes. Propidium iodide (PI, 50 µg/mL, Sigma P4170) and RNase (50 µg/mL, Thermo Fisher) were then added to the suspension and incubated for at least 60 minutes. Adult cortex samples were stained for NeuN using a 1:750 primary antibody dilution for 1.5 hours (Millipore MABN140), washed for 20 minutes, followed by a 1:500 secondary antibody dilution (Invitrogen A21206) for 1.5 hours, and washed for at least 20 minutes with PI included throughout. Embryonic nuclei were gated on the 2n DNA peak from propidium iodide (PI, 50µg/ml, Sigma) staining. Nuclei from adult mice were labeled with an

anti-NeuN antibody (Millipore) and PI and then isolated by gating for both 2n PI intensity and a NeuN-positive signal.

Splenocyte isolation. Splenic samples were collected from adult mice and enriched for lymphocyte cells using Lympholyte-M (Cedarlane Labs) as described by the manufacturer. The spleen was dissected from adult C57BL/6J mice and dissociated by grinding between glass slides in HBSS + 1% FCS. The solution was allowed to settle, then cells in the supernatant pelleted by centrifugation for 5 min at 1,500 rpm. Cells were resuspended in HBSS with 0.2x PBS and 1% FCS, filtered, and diluted to 1 x 10^7 cells/mL. Dead cells, erythrocytes, and debris were removed using Lympholyte-M (Cedarlane Labs) as described by the manufacturer, and washed with HBSS containing 1% FCS. Cells were stained with PI and DRAQ5 (5μM, Cell Signaling) for live-dead and DNA content measurements, respectively. Splenocytes were required to be PI-negative (live-dead), and then gated on the 2n DNA peak from DRAQ5 (5μM, Cell Signaling) staining.

GenomePlex sequencing library preparation. For GenomePlex (WGA4, Sigma) amplifications, cells were sorted into 3 uL sterile PBS with 1% BSA. Manufacturer instructions were followed for GenomePlex (WGA4, Sigma) amplifications apart from using half the volume. Sequencing libraries were then generated using the NEBNext library preparation kit (New England BioLabs). Sequencing for all libraries was performed on a NextSeq™ with a 500/550 High Output Kit v2 (Illumina, Inc) to obtain a minimum of 36 cycle single end reads.

Bioinformatic software implementation and references. Low depth sequencing (~1.5 M unique reads) was obtained for each sample. If necessary, sequencing reads were trimmed to 36 bp by removing bases from the 3' end. Samples amplified by GenomePlex also required the removal of a 5' adapter sequence encompassing 30 bp. Reads were mapped to the mm10 genome using Bowtie v1.1.2 (options –n 2, -e 70, -m 1 –-best, --strata) (1), and PCR duplicates were removed using rmdup from samtools v0.1.19

(2). Methods described by Baslan et. al. (3) were used to generate a genomic reference of 25,000 bins containing equal numbers of uniquely mappable bases for the mm10 reference genome and then to count the number of sequencing reads in each bin from individual samples.

GC-correction was applied to each sample by Lowess regression on $\log_{10}$ median-ratio of the bin counts (statsmodels v0.6.1, nonparametric.smoothers_lowess, frac=0.05) (4). Each dataset was then normalized against a reference generated by combining the corrected bin counts from all single cell samples within the same project that were amplified by the same method and contained at least 500,000 counted reads (5). Circular binary segmentation (CBS) was performed using cghcbs in Matlab (10,000 permutations, stopping rule off, alpha 0.01) (6, 7).

Sex was determined by assessing the copy number states of the X and Y chromosomes. Median absolute pairwise difference (MAPD) (8, 9), median absolute difference (MAD) (10), variability score (VS) (11, 12) and confidence score (CS) (10) were calculated as previously described, with the exception of scaling copy number estimate by segment length and taking the median rather than mean when calculating CS. Datasets were required to have $\geq$ 600,000 unique reads, MAPD $\leq$ 0.40, and CS $\geq$ 0.80 to be included in further analysis. Adjacent chromosomal segments with matching copy number state estimates were merged together prior to CNV filtering. The Y chromosome was excluded from biological CNV analysis due to both its small size and the uncharacterized pseudo-autosomal regions in *Mus musculus* that could not be masked.

Genomic element references were generated using information from the UCSC Table Browser (13), including the gap track (centromere/telomere), RefSeq (genes), segmental duplications, CpG islands, and Repeat Masker, and phastCons60way (vertebrate evolutionary conservation). Replication timing data was obtained from Encode and converted to mm10 coordinates using UCSC liftOver (ENCFF001JVQ). Machine learning procedures including SVM, hierarchical clustering, and dimensionality reduction were

performed in Python with scikit-learn v0.17.1 (14). Additional statistical testing was performed in Python using SciPy.stats v0.14.0 (15). The Kruskal-Wallis test was used for multi-group comparisons, the Mann-Whitney U test for two groups, and the Wilcoxon Signed-Rank for paired data. The Fisher's exact test was applied to 2x2 contingency tables and $\chi 2$ for 2xN contingency tables. Z-scores were used to assess enrichment of genomic elements relative to a bootstrapped null distribution for each biological group.

Extended Bioinformatic Analysis and Results

Manual calculation of sample ploidy. Overall cellular ploidy for a species is discussed in integer values. However, this is not a valid assumption for a single cell with imbalanced CNVs (e.g., an aneuploid cell). To compensate, we calculated the optimal ploidy states of each sample by maximizing the fit to a discrete distribution, similar to the strategy used on the open-source platform Ginkgo (16).

Circular binary segmentation (CBS) output consists of segments in $\log_2$-ratio units, which can be converted to (non-integer) copy number estimates by:

$$Copy\ Number = ploidy\ \times\ 2^{segment\ value}$$

For each dataset, copy number profiles were generated for all potential ploidies from 1.25 to 2.75 in 0.01 unit intervals. The CS generated from each ploidy was then calculated, and the ploidy generating the highest CS score was assigned to the sample. This approach increased the median CS by 5.7% and allowed the inclusion of 11% more samples in analysis (*SI Appendix*, Fig. S6).

Unbiased determination of QC requirements. We anticipated that a proportion of splenocytes of unknown type were actually lymphocytes with V(D)J recombination obscured due to low quality data, meaning robust QC cutoffs should exclude the unidentified cells more frequently than samples classified

as B- or T-cells. Indeed, the distributions of QC measurements were skewed towards lower-quality values for uncategorized lymphocyte samples (*SI Appendix,* Fig. S1*A-C*).

Using the full lymphocyte dataset, natural clustering of samples within high-quality data space was observed (*SI Appendix,* Fig. S2*D*). The boundaries of this cluster were defined using an adaptation of Dijkstra's contraction algorithm (17) as follows:

1. Calculate the Euclidean distance matrix for all pairwise sample comparisons using range-scaled (0 to 1) QC values for read number, MAPD, and CS.
2. Combine the two samples with the smallest distance into one cluster. Remove this pairwise comparison from the distance matrix.
3. Repeat Step 2, removing all pairwise comparisons between samples within the same group from consideration each time, until 95% of samples have been clustered.
4. Select the largest cluster as containing high quality samples.
    a. Select the lowest observed value in the cluster for the reads cutoff (rounding down to the nearest 500,000).
    b. Select the lowest observed value for the CS cutoff (rounding down to the nearest multiple of 0.05).
    c. Select the highest observed value for the MAPD cutoff (rounding up to the nearest multiple of 0.05).
    d. Rounding was used to slightly expand the boundaries so they might be less susceptible to overfitting caused by dataset size limitations.

While 82% of classified lymphocytes passed these thresholds, only 49% of unclassified splenocytes were retained (*SI Appendix,* Fig. S1*E*). Further, the distributions of QC metrics of identified and non-identified lymphocytes had become indistinguishable, suggesting the excessively low-quality

datasets had been removed (*SI Appendix,* Fig. S1*F-H*). The remaining samples lacking recombination are likely endothelial and stromal cells.

Development of CNV cutoffs (FUnC). V(D)J recombination events and euploid regions of lymphocyte samples were used to quantify the appearance of valid copy number assignments. Euploid regions were defined as autosomal segments with a copy number rounding to 2 and sex chromosome segments rounding to copy numbers of 0 (female Y), 1 (male X and Y), or 2 (female X). These regions were used to represent large CNV events. Since we were analyzing samples from non-cancerous tissue gated around a 2n peak, the majority of the genome in the majority of cells was expected to be euploid, meaning the vast majority of regions identified as such most likely had a correct assignment.

intD values were range-scaled, and size (number of bins) was both log and range-scaled to avoid excessive weight given to larger segments. Machine learning was performed using scikit-learn v0.17.1 (svm.OneClassSVM) with kernel rbf, gamma = 10, nu = 0.12587 (14). These values were empirically determined to avoid overfitting (wavy boundaries and gaps) or underfitting (overly spherical, numerous valid samples excluded) the data.

Separate models were built for the V(D)J recombination segments and euploid segments, to avoid overwhelming the smaller number of samples and CNV size range for the V(D)J recombination CNVs. Each model was built following these steps:

1. Select n / 2 segments using random sampling with replacement (n = total number of training data points).
2. Train a one-class support vector machine (SVM) with the bootstrapped dataset from Step 1.
3. Determine which combination of size and intD values are included in the model (size in 1 bin steps, intD in 0.05 unit steps).

4. Repeat steps 1-3 10,000 times.
5. Calculate the frequency with which all combinations of size and intD values are included in a model.
6. Determine the maximal intD included in ≥ 95% of modeling iterations for each CNV size.
7. Perform smoothing of the output boundary from Step 6.
    a. For the V(D)J recombination model, remove variable combinations if intDsize < intDsize-1.
    b. For the euploid segment model, remove variable combinations if intDsize > intDsize-1.
8. Calculate any size and intD variable combinations that were not in the size range between models by linear regression between the nearest 2 points.

CNV simulations to validate FUnC. False positive CNVs were simulated with male samples that contained no CNV calls on the X chromosome. Pairwise sample similarities were calculated by the Euclidean distance of QC metrics, and combinations resulting in the 10 smallest differences for each sample were selected for simulations (582 total iterations). X chromosome bin counts were combined for these pairs and then segmented by CBS. FDR was calculated by:

$$FDR = \frac{\sum_{i=1}^{582} CNV\ count_i \Big/ number\ of\ simulations}{Diploid\ X\ chromosome\ CNVs\ per\ cell}$$

where *i* is one computational simulation, and the denominator was calculated by averaging the X chromosome CNV rate of female datasets.

True positive simulations used male and female samples with no X-chromosome CNV calls. For each female cell the 10 most similar male samples were selected as described for the male pairs, and 10 separate simulations were performed per pair (2,910 total iterations). In each simulation, a random CNV

size, start location, and copy number value (1 or 3) was selected. Deletions were simulated by replacing bins on the female X chromosome with those from the male, and amplifications by adding together the bin counts. After segmentation, false negative rate (FNR) was calculated by:

$$FNR = 1 - \frac{Number\ of\ CNVs\ detected}{Number\ of\ CNVs\ simulated}$$

Additional Dataset Captions

Additional data table S1 (separate file)

Size and integer distance of splenocyte positive control segments used to train the machine learning filtering unreliable CNVs (FUnC) model.

Additional data table S2 (separate file)

Integer distance cutoffs for varying CNV sizes used to implement filtering unreliable CNVs (FUnC).

Additional data table S3 (separate file)

CNVs observed in the developing and adult cerebral cortex.

Additional data table S4 (separate file)

Simulated false positive CNV calls used to validate the filtering unreliable CNVs (FUnC) approach.

Additional data table S5 (separate file)

Simulated true positive CNV calls used to validate the filtering unreliable CNVs (FUnC) approach.

Supplemental References

(insert here!)

Supplemental Figure Legends

Fig. S1. Determination and validation of quality control (QC) cutoffs. (*A*) Natural clustering of high quality lymphocytes was used to define QC requirements (gray planes). (*B-D*) Cumulative histograms show the distributions of QC metrics for lymphocytes that were identified (as B or T cells) versus those that did not have an assigned cell type. Upper panels show distributions of all sequenced samples ("All"; N = 62, 43 for identified and unknown, respectively) and lower panels show distributions of high quality samples ("Good"; N=51, 21 for identified and unknown, respectively). (*E*) Cell type classification rates after removing low quality samples. Note the clear reduction in the number of cortical cells misclassified as B or T lymphocytes after excluding such samples, in contrast to Fig. 4C (Lym, N=71; Ctx, N=345). * $p < 10^{-30}$.

Fig. S2. The impact of filtering unreliable CNVs (FUnC) on true and false positive CNV calls. (*A-B*) Segment size distributions (*A*) are minimally changed after removing low quality segments by FUnC, while integer distance values (*B*) are much more similar for the true positives and putative CNVs. Euploid, N=2,833; immune, N=42; putative, N=665; * $p < 0.01$ compared to euploid regions; ‡ $p < 0.01$ compared to immune CNVs. (*C*) After removing unreliable CNV calls, the majority of splenocytes maintain an identifiable cell type. Note also the further reduction in the number of cortical cells misclassified as B or T cells after removing such samples (in contrast to Fig. 4C and Fig. S1E). Lym, N=71; Ctx, N=345; * $p < 10^{-25}$. (*D*) Positive control CNVs – V(D)J recombination and monosomic sex chromosomes (N=74) – from splenocytes prepared by an independent researcher after the creation of FUnC are maintained after applying CNV cutoffs. (*E-F*) Processing of control samples generated in an independent study

(SRP041670) support the appropriateness of FUnC since the majority of true positive CNV calls (*E*; N=201) are retained, while a minority of CNV calls enriched for false positives (*F*; N=164) are within FUnC thresholds.

Fig. S3. Hierarchical clustering confirms the stochastic distribution of somatic cortical CNVs. (*A*) Clustered copy number profiles of all high quality samples prepared by transposase-based amplification (TbA) (distance metric = correlation, clustering method = single). Only two clear clusters (labeled 1 and 2) are observed in the dendrogram. (*B*) Close up of cluster 1 shows it contains primarily B cells. (*C*) Close up of cluster 2 shows it contains exclusively T cells.

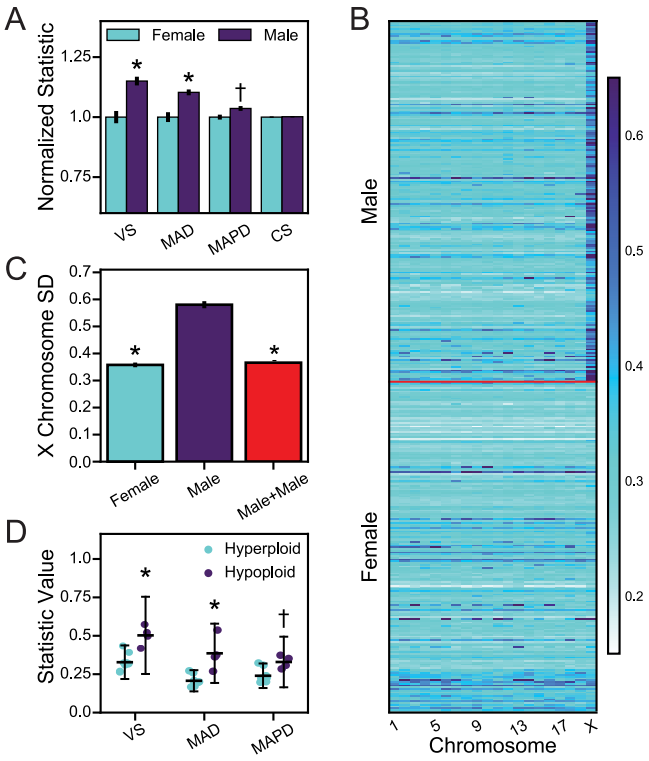
Fig. S4. Confirmation that neurodevelopmental CNVs are randomly distributed. (*A*) Kernel PCA (Gaussian radial basis function (rbf) kernel) of CNV genomic element composition shows no sub-grouping of alterations. (*B*) This lack of clustering is the same as observed after applying kernel PCA to randomly selected genomic intervals. (*C-D*) The genomic location of the CNVs in each cell was randomly shuffled 1,000 times to obtain a null distribution of genomic element composition for each group. The Z-score of the observed genomic element abundance against this distribution was then calculated for amplification (*C*) and deletion (*D*) events. Values > 0 indicate over-enrichment and values < 0 indicate under-enrichment. Statistical Z-tests were used to determine significant differences from the overall genome's composition (*). Centromeres and telomeres are the only consistently enriched component, although at early ages, amplification events may also be enriched for genomic elements associated with early replicating S-phase loci.

Fig. S5. Developmental CNV trends are reproducible across biological preparations. Overall CNV frequency (*A*) and separated amplification and deletion rates (*B*) for cells from replicate sample collections. Error bars show SEM; * $p < 0.05$ vs the first splenocyte (Lym) preparation; ‡ $p < 0.05$ vs the second splenocyte preparation; N=13, 17, 30 for E13.5 litters; N=28, 28 for E14.5 litters; N=69, 19 for Lym samples.

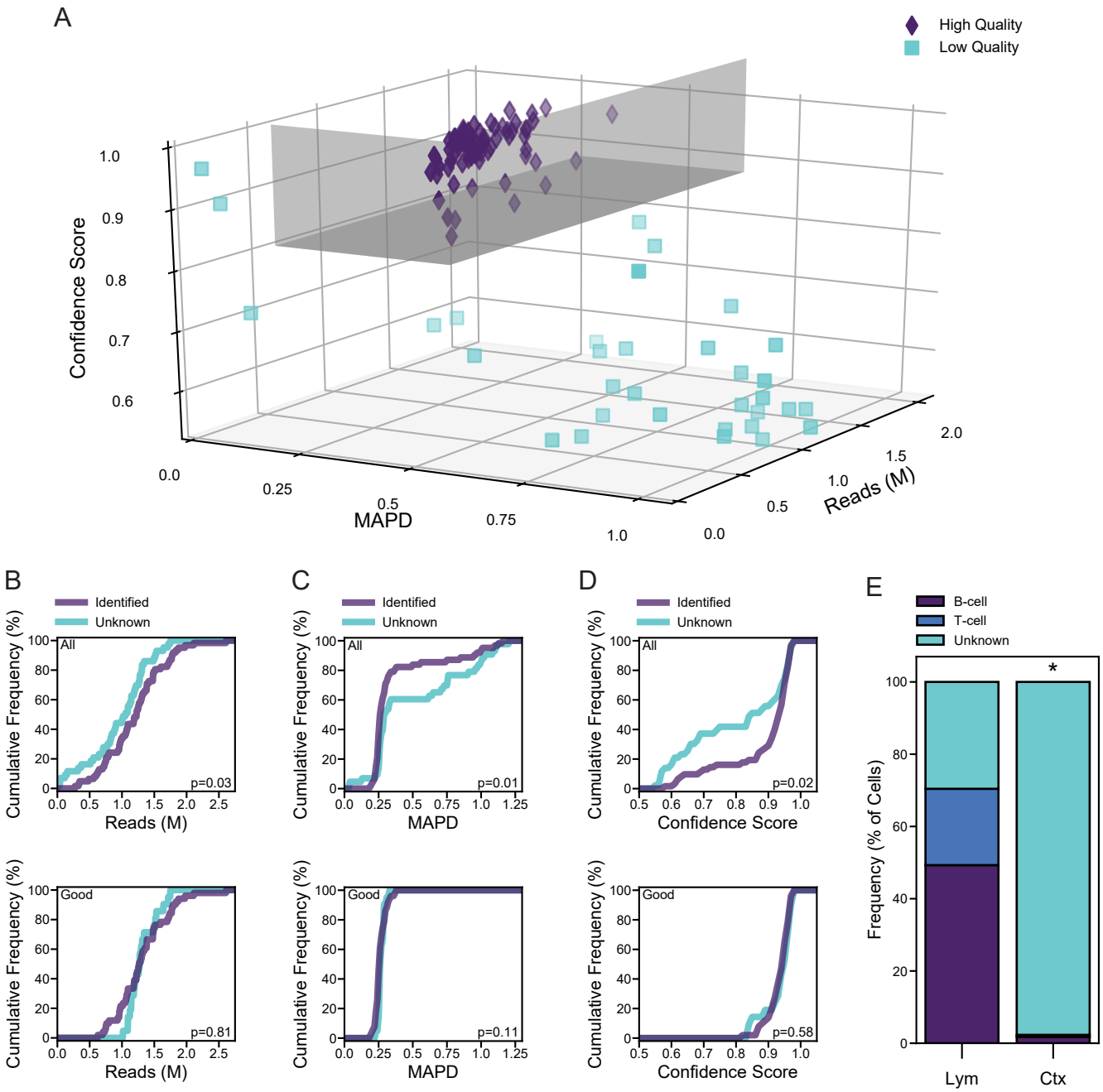
Fig. S6. Single-cell ploidy calculation improves the retention of cells with highly altered genomes that would have been eliminated by quality control requirements using standard protocols. (*A*) A euploid female cell conforms to the 2n diploid assumption (point of maximal confidence score, CS). (*B*) A euploid male cell has a true ploidy value of ~1.9 because of monosomy of the sex chromosomes. (*C*) A cell with many large deletions is fit to a ploidy state much lower than 2. (*D*) Sample CS values when assuming diploidy (X-axis) are significantly lower than CS values after calculating the optimal ploidy (Y-axis). The gray line shows the boundary of a 1:1 relationship, and the red box contains samples that would have been erroneously excluded without empirical determination of ploidy. N=653.

1. Langmead B, Trapnell C, Pop M, & Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10(3):R25.
2. Li H*, et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078-2079.
3. Baslan T*, et al.* (2012) Genome-wide copy number analysis of single cells. *Nature protocols* 7(6):1024-1041.
4. Seabold S & Perktold J (2010) Statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference*).
5. Evrony GD*, et al.* (2015) Cell lineage analysis in human brain using endogenous retroelements. *Neuron* 85(1):49-59.
6. Olshen AB, Venkatraman ES, Lucito R, & Wigler M (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5(4):557-572.
7. Venkatraman ES & Olshen AB (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* 23(6):657-663.
8. Affymetrix (2008) Median of the absolute values of all pairwise differences and quality control on affymetrix genome-wide human SNP array 6.0.
9. Cai X*, et al.* (2014) Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. *Cell Rep* 8(5):1280-1289.
10. McConnell MJ*, et al.* (2013) Mosaic copy number variation in human neurons. *Science* 342(6158):632-637.
11. Knouse KA, Wu J, & Amon A (2016) Assessment of megabase-scale somatic copy number variation using single-cell sequencing. *Genome Res* 26(3):376-384.
12. Knouse KA, Wu J, Whittaker CA, & Amon A (2014) Single cell sequencing reveals low levels of aneuploidy across mammalian tissues. *Proc Natl Acad Sci U S A* 111(37):13409-13414.
13. Karolchik D*, et al.* (2004) The UCSC Table Browser data retreival tool. *Nucleic Acids Research* 1(32):D493-D496.
14. Pedregosa F*, et al.* (2011) Scikit-learn: Machine Learning in Python. *J Mach Learn Res* 12:2825-2830.
15. Jones E, Oliphant T, & Peterson P (2001) SciPy: Open Source Scientific Tools for Python.
16. Garvin T*, et al.* (2015) Interactive analysis and assessment of single-cell copy-number variations. *Nat Methods* 12(11):1058-1060.
17. Knuth DE (1977) A generalization of Dijkstra's algorithm. *Information Processing Letters* 6(1):1-5.

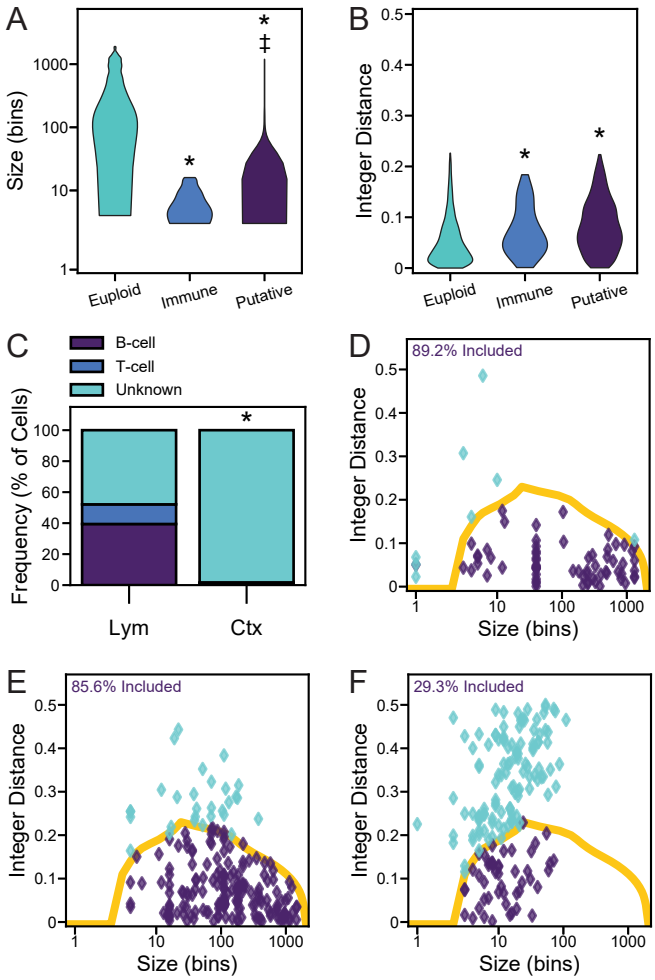
Supplementary Figure 1



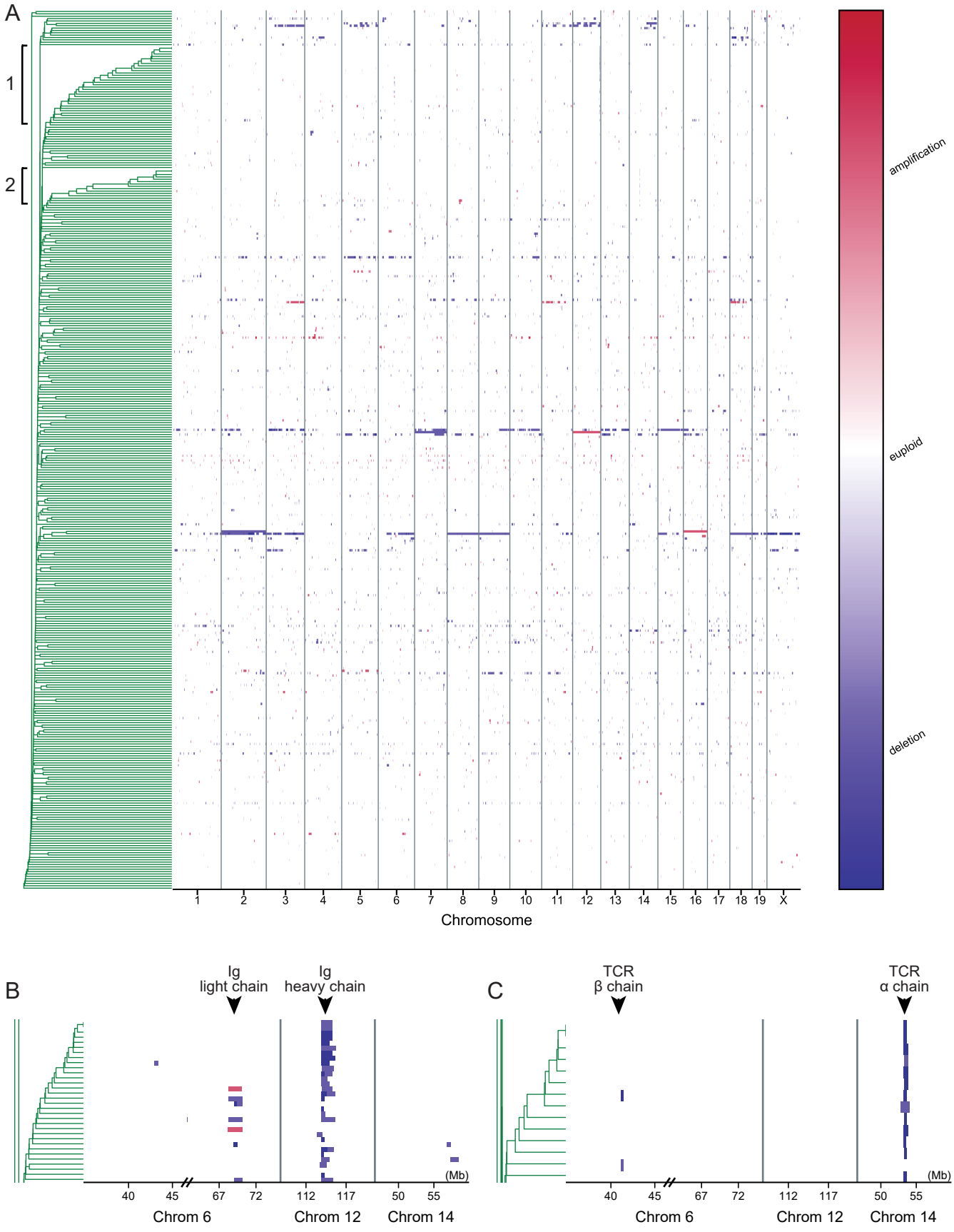
Supplementary Figure 2



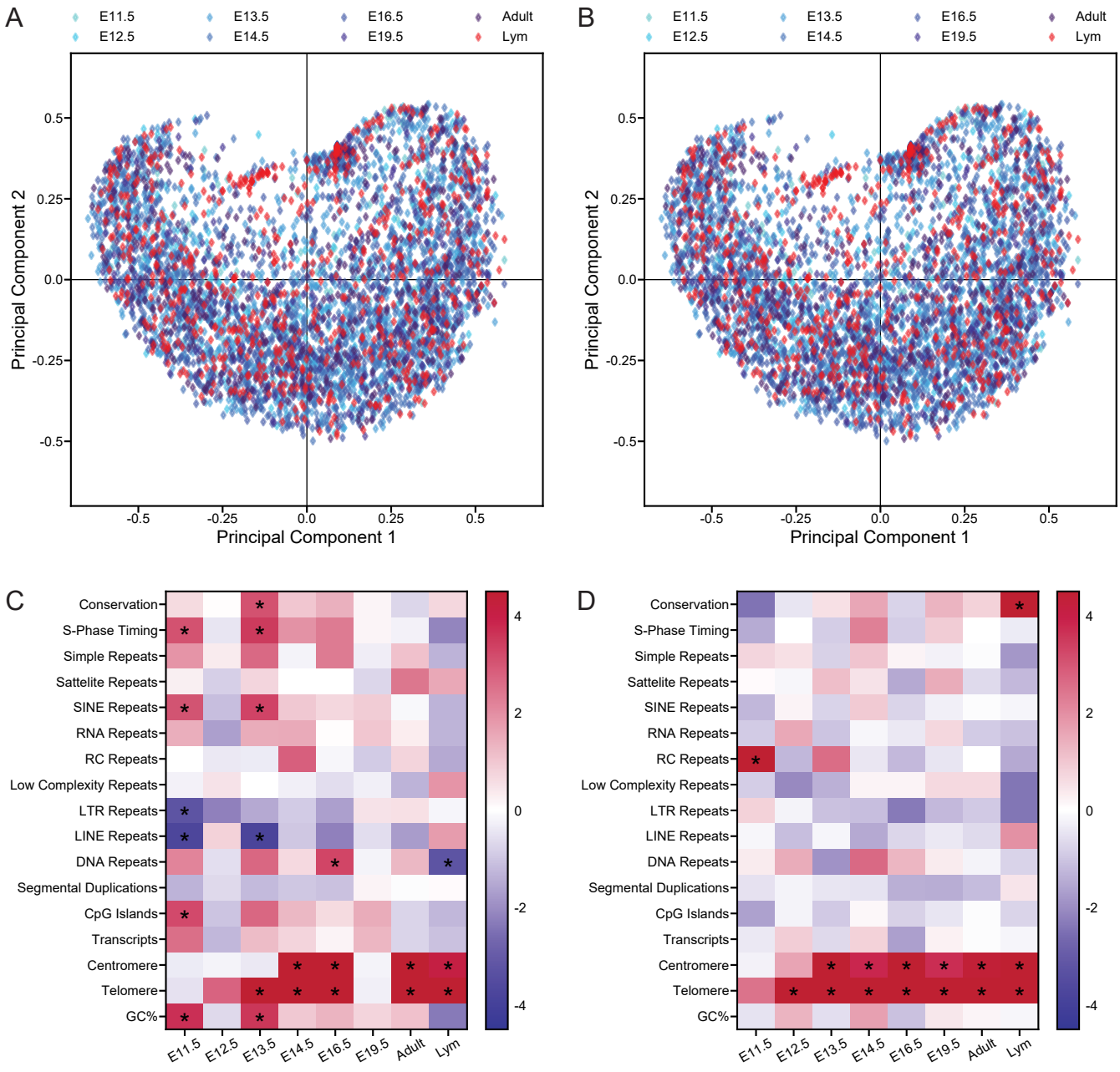
Supplementary Figure 3



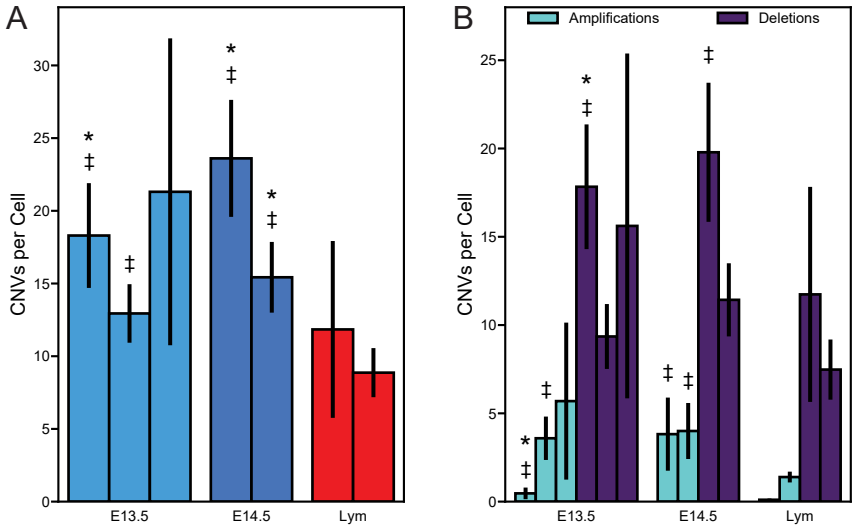
Supplementary Figure 4



Supplementary Figure 5



Supplementary Figure 6



Supplementary Figure 7