Somatic copy number variant load in neurons of healthy controls and Alzheimer's disease patients

Turan *et al.*
Acta Neuropathologica Communications (2022) 10:175
https://doi.org/10.1186/s40478-022-01452-2

Acta Neuropathologica Communications

RESEARCH Open Access

Somatic copy number variant load in neurons of healthy controls and Alzheimer's disease patients

Zeliha Gözde Turan[1*], Vincent Richter[2], Jana Bochmann[2], Poorya Parvizi[1,3], Etka Yapar[4], Ulas Işıldak[1], Sarah-Kristin Waterholter[2], Sabrina Leclere-Turbant[5], Çağdaş Devrim Son[1], Charles Duyckaerts[5], İdil Yet[6†], Thomas Arendt[2†], Mehmet Somel[1†] and Uwe Ueberham[2*†]

Abstract

The possible role of somatic copy number variations (CNVs) in Alzheimer's disease (AD) aetiology has been controversial. Although cytogenetic studies suggested increased CNV loads in AD brains, a recent single-cell whole-genome sequencing (scWGS) experiment, studying frontal cortex brain samples, found no such evidence. Here we readdressed this issue using low-coverage scWGS on pyramidal neurons dissected via both laser capture microdissection (LCM) and fluorescence activated cell sorting (FACS) across five brain regions: entorhinal cortex, temporal cortex, hippocampal CA1, hippocampal CA3, and the cerebellum. Among reliably detected somatic CNVs identified in 1301 cells obtained from the brains of 13 AD patients and 7 healthy controls, deletions were more frequent compared to duplications. Interestingly, we observed slightly higher frequencies of CNV events in cells from AD compared to similar numbers of cells from controls (4.1% vs. 1.4%, or 0.9% vs. 0.7%, using different filtering approaches), although the differences were not statistically significant. On the technical aspects, we observed that LCM-isolated cells show higher within-cell read depth variation compared to cells isolated with FACS. To reduce within-cell read depth variation, we proposed a principal component analysis-based denoising approach that significantly improves signal-to-noise ratios. Lastly, we showed that LCM-isolated neurons in AD harbour slightly more read depth variability than neurons of controls, which might be related to the reported hyperploid profiles of some AD-affected neurons.

Keywords: Single-cell whole-genome sequencing, Copy number variation, Alzheimer's disease, Brain, Laser capture microdissection, Fluorescence-activated cell sorting, Denoising

Introduction

Alzheimer's disease (AD) is a neurodegenerative disease of multifactorial aetiology, with numerous genetic and environmental factors each explaining a small proportion of variance in disease onset and progression [1]. One of the less-studied potential contributors is somatic copy-number variations (CNVs) in neurons, which can include the gain or loss of whole chromosomes (aneuploidy) or of chromosomal segments. It is generally accepted that mature neurons in healthy brains can carry somatic CNVs, but their frequency is uncertain. Early studies estimated aneuploid neuron frequencies between 4 and 40% in neurotypical brains [2–4], while analyses using single-cell whole-genome sequencing (scWGS) estimated aneuploid neuron frequencies at < 1% [5]. Beyond aneuploidy,

†İdil Yet, Thomas Arendt, Mehmet Somel and Uwe Ueberham have contributed equally to this work

*Correspondence: turan.zelihagozde@gmail.com; uwe.ueberham@medizin.uni-leipzig.de

[1] Department of Biological Sciences, Middle East Technical University, 06800 Ankara, Turkey
[2] Paul Flechsig Institute for Brain Research, Leipzig University, Leipzig, Germany
Full list of author information is available at the end of the article

recent scWGS studies also estimated CNV-carrying neurons at around 30% in young adults and 10% in old adults [6].

Over the last two decades, a number of fluorescence in situ hybridization (FISH) and cytogenetic-based studies investigated CNV frequencies in AD and healthy control brains [2, 7–12]. Several of these reported extra copies of chromosomes in the AD brain [7–12]. This, in turn, implies that the chromosomal imbalance might contribute to AD pathogenesis via altered gene expression levels. An example of such imbalance is seen in individuals with Down's syndrome (DS); carrying an extra copy of chromosome 21 appears to facilitate aggregation of amyloid-β (Aβ) plaques in the brains of DS individuals similar to the AD phenotype [9, 13, 14].

There are various explanations for why post-mitotic neurons in AD brains could carry high frequencies of somatic CNV [15]. According to one view, the high CNV burden in the AD brain originates from neurogenesis in the embryonic period. This excessive somatic mutation may be pathogenic and manifest itself as increased AD risk during ageing [16]. However, Abascal et al. recently showed that somatic mutation (single nucleotide change or indel) accumulation in cells with mitotic capacity and in post-mitotic neurons follow similar trajectories. That is, mutational processes (possibly also including CNVs) appear to occur in a time-dependent manner rather than being division-dependent [17]. Accordingly, CNVs in AD brains may have accumulated during their lifetime. However, this scenario also appears inconsistent with the observation that CNV-bearing neuron frequencies decrease from young to old adulthood [6]. Another view suggests that AD itself might cause dysregulation in neurons, and AD-affected mature neurons might re-enter the cell cycle, resulting in increased CNV load [8, 18], which may then be eliminated at later stages of AD, thus causing neurodegeneration [10].

Over the last decade, advances in next-generation sequencing (NGS) technologies gave fresh impetus to somatic CNV analyses by allowing variants to be determined at the single-cell level [19]. In one such study, van den Bos and colleagues used scWGS to compare the prevalence of aneuploidy in neurons from healthy control and AD patients [5]. Analyzing 1482 neurons from 10 AD patients and 6 control individuals, the authors reported aneuploid prevalence at 0.7% and 0.6% for control and AD neurons, respectively, and concluded that aneuploid cells are not more common in the AD brain.

These findings by van den Bos and colleagues implied that CNVs might have no relationship to AD pathogenesis, in contrast with earlier finds from FISH and cytometry. However, the study by van den Bos and colleagues had a number of limitations. One was that the authors only estimated aneuploidy (full chromosome gain or loss), while large CNVs, which could also contribute to pathogenesis, remained uncharacterized. Another limitation was that only one brain region was examined, the frontal cortex, while atrophy of the medial temporal lobe and specifically the hippocampus is generally considered to be a strong predictor of AD [20]. The study did not distinguish among neuron types that may carry sensitivity to AD differentially [21]. Thirdly, the study discarded a large fraction of cells (39%) for showing high within-cell variability in genome coverage, although it was unclear to what extent these represented pure technical error versus cells with complex karyotypes. Fourthly, only NeuN positive neurons were included, which substantially restricts the significance of this study due to different reasons: (1) Recently, up to 30% of cortical neurons have been reported being NeuN-negative following diffuse brain injury, which may be related to certain neurons being particularly vulnerable to membrane disruption [22], a process recently associated with AD [23, 24]. (2) Considerable or even complete loss of NeuN immunoreactivity was also reported for neurons affected by ischemic insults (middle cerebral artery occlusion) without significant cell loss [25] or in neurons that just entered the cell death process [26]. Interestingly, these neuronal populations are of special interest because energy and nutritional deficiency and cell loss are essential characteristics of the AD brain [27]. (3) The intensity of NeuN staining is reported to be lower in AD samples [28], and further (4) due to many NeuN negative cortical neurons in FTLD-TDP (frontotemporal lobar degeneration with TDP-43 inclusions) patients, Yousef et al. suggested NeuN staining as an indicator of healthy neurons [29]. However, if NeuN reflects a neuron's health, any selection of NeuN positive cells would lead to a substantial bias for studying any neurodegenerative disease.

These methodological issues could potentially explain the discrepancies between the findings by van den Bos et al. and those based on FISH and cytogenetic studies [7–12]. Notably, a recent technical comparison between FISH and scWGS using mock aneuploid cells reported a tendency of the latter to severely underestimate aneuploidy [30]. It is thus possible that both neurons with CNV and nuclei thereof display altered physicochemical properties. This may result in selection bias against abnormal nuclei with high CNV loads when using the fluorescence activated cell/nuclei sorting (FACS, FANS) isolation method (exerting mechanical stress [31]) and high hydrodynamic pressure [32], applied by van den Bos and colleagues, and artificially inflate euploidy frequencies. Moreover, besides restriction to NeuN positive cells, usage of only intact nuclei could preclude or

bias AD neurons with nuclear envelope stress or rupture [33].

These observations call for additional data and approaches to tackle this issue. Accordingly, here we generated and analyzed scWGS data to establish the frequency of CNVs (both full chromosome aneuploidies and sub-chromosomal CNVs) in five different brain regions that differ in vulnerability to AD [34]. We employed two different single-cell isolation methods, laser capture microdissection (LCM) and FACS, to isolate neuronal nuclei. LCM, despite being technically challenging, has the advantages of allowing for specific neuron types to be chosen, and being neutral towards normal and abnormal nuclei. We further employed a principal component analysis-based denoising approach to eliminate false positive CNV calls that might result from either systematic experimental biases or repetitive regions in the human genome. Finally, we analyzed published datasets to replicate our main results and check the sensitivity and specificity of our bioinformatics pipeline.

Materials and methods

Tissue sources

Frozen postmortem human brain tissues -temporal cortex, hippocampal subfields cornu ammonis (CA) 1, hippocampal subfields cornu ammonis (CA) 3, cerebellum (CB) and entorhinal cortex (EC)- from a total of 13 AD patients and 7 non-demented age-matched controls were obtained from the GIE NeuroCEB Brain Bank (France) (Additional file 1: Table S1-A). AD cases were diagnosed according to the National Institute of Aging and Reagan Institute Criteria [35] and immunohistochemically processed for tau and amyloid pathologies [36, 37]. Control cases were non-demented individuals who died without known neurological disorders. Post-mortem delays and mean ages of control and AD cases were not significantly different. The average age of death was for control cases ($n=7$) 71 0.57 years (± 5.13 years SEM) and for AD cases ($n=13$) 70.15 years ($\pm 3{,}63$ years SEM) ($p=0.822$). The average post-mortem delays were 31.14 h (± 7.10 h SEM) for control cases and 26.17 h (± 4.08 h SEM) ($p=0.52$). All experiments were conducted at Paul-Flechsig-Institute (Leipzig University, Germany).

Fluorescence-activated cell sorting (FACS)

Neuronal nuclei were extracted following the protocol described in [38]. Briefly, frozen brain samples were thawed in the hypotonic lysis buffer. Neuronal nuclei were stained with propidium iodide and sorted using BD FACSAria II SORP (BD Biosciences). Genomic DNA was then isolated and amplified as described below (see scWGS library preparation and sequencing).

Laser capture microdissection (LCM)

Frozen brain samples at −80 °C were thawed to −20 °C, sliced using CryoCut Freezing Microtome at 30 µm thickness, and mounted on a membrane slide (Carl Zeiss). After staining with cresyl violet, single cells were cut out and placed into an adhesive cap by PALM MicroBeam (Carl Zeiss). Neurons of the individual *5603* were collected using both FACS ($n=12$) and LCM ($n=64$).

scWGS library preparation and sequencing

Genomic DNA was amplified using WGA4 (GenomePlex® Single Cell Whole Genome Amplification Kit) and then purified using the MinElute PCR Purification Kit (Qiagen). The specific adapters were added to the DNA via Phusion® PCR followed by purification with the MinElute PCR Purification Kit (Qiagen). Sample quality was evaluated using agarose gel electrophoresis. Sequencing was performed on the HiSeq2500 platform (Illumina) with paired-end 100 bp (PE100) or 150 bp (PE150) modes.

Read quality control and alignment

The *FastQC* tool (version 0.11.9) was used to check the quality of the raw Illumina reads. The results of *FastQC* were summarized using *MultiQC* (version 1.9) [39]. The mean sequence lengths of the reads (ranging between 101 and 151) were inspected using the output of the *MultiQC* (*general_stats_table*). To avoid biases that would affect the interpretation of the results, all reads were trimmed to a length of 66 (the longest possible length in all reads). Illumina adapter and low-quality bases (the first 35 bp) were removed using *Trimmomatic* [40] with the following parameters: "ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10:8:TRUE HEADCROP:35 MINLEN:66 CROP:66". The quality of the trimmed reads was checked again using both *FastQC* and *MultiQC*. Adapter-trimmed paired-end FASTQ files were mapped to the hg19 human reference genome using Burrows-Wheeler Alignment (BWA v.0.7.17) [41] with *aln* and *sampe* options.

Filtering

The output of the BWA aligner in Sequence Alignment/Map (SAM) format was further processed by *SAMtools* v1.10 [42] to obtain high-quality uniquely aligned reads. The applied steps are as follows: (1) keep reads mapped in proper pair and discard reads marked with SAM flag 3852 (using the command "samtools view -f 2 -F 3852 -b file.sam > file.bam"), (2) extract uniquely mapped reads from BAM files ("samtools view -h file.bam | egrep -i

"^@|XT:A:U" | samtools view –Shu - > file.bam2") [43], (3) obtain reads having MAPQ scores 60 ("samtools view -h -q 60 file.bam2 > file.bam3"), (4) sort BAM files ("samtools sort file.bam3 > file.sorted.bam"), (5) filter out PCR duplicates ("samtools rmdup -S file.sorted.bam file_rm.sorted.bam"), (6) index BAM files ("samtools index -b file_rm.sorted.bam"), (7) convert BAM file into BED format using the Bedtools *bamToBed* command (Bedtools v2.27.1) [44].

Coverage

Bedtools v2.27.1 algorithm *genomeCoverageBed* was used to obtain coverage of the bases on each BAM file.

CNV prediction and cell elimination

CNV calling was performed using *Ginkgo* [45]. We had three main reasons for using *Ginkgo* over its most commonly used alternative, *HMMcopy* [46]. First, a recent study [47] performed benchmarking on *Ginkgo* and two other widely used methods *HMMcopy* and *CopyNumber*, and found that *Ginkgo* was the most accurate algorithm for inferring the absolute copy number profiles (although *HMMcopy* was superior in identifying breakpoints and running time). Second, *Ginkgo* provided the advantage of outputting data with normalised coverages per cell, which we could use in our PCA-based denoising method, and further in estimating the genome-wide copy number of each cell, which we used to filter cells for high levels of variability in read depth. Third, our tests on the sensitivity and specificity of *Ginkgo* using trisomy-21 in DS and monosomy-X in males in published data [5, 48] revealed 100% and 94% detection rates across the two published datasets. The command-line version of *Ginkgo* was downloaded from https://github.com/robertaboukhalil/ginkgo. The tool was run under the following settings: *(1)* variable size of 500 kb bins [43] based on simulations of 76 bp reads aligned with *BWA*, *(2)* independent segmentation method, *(3) ward* and *euclidean* options for the clustering method and clustering distance metric, respectively. Before the segmentation step, GC correction was performed by *Ginkgo* using the R function *LOWESS* (see [45]). For segmentation, *Ginkgo* uses the CBS algorithm implemented in *DNAcopy* in R [49]. *DNAcopy* runs with the following parameters: alpha = 0.0001, undo.SD = 1, min.width = 5 [50]. We also run *HMMcopy* as described in [43] (using the parameter $e = 0.995$).

The number of reads was divided into the variable size of 500 kb bins that correspond to 5578 genomic windows. Only cells with > 50,000 reads were kept in downstream analyses (approximately nine reads per window), resulting in $n = 1337$ cells.

Published datasets

The van den Bos 2016 dataset: Data was downloaded from EBI ArrayExpress with the accession numbers E-MTAB-4184 and E-MTAB-4185 [5]. Only the cells that were reported as having good quality libraries were included in the analysis (AD: 883; control: 586; Down's syndrome: 34). Adapter sequences were trimmed with the following parameters: "ILLUMINACLIP:adapter.fa:2:30:10:8:TRUE MINLEN:51". Single-end reads were aligned to the hg19 human reference genome using BWA with *aln* and *samse* options. The remaining steps are the same as those described in sections *Filtering*, except that here we used the SAM flag 3844 (because this dataset was single-end sequenced) and used MAPQ scores 20 (because this dataset did not have enough reads which having the MAPQ 60). Note that due to the missing sample information in the database, the number of cells we analyzed does not match what van den Bos and colleagues reported in their original publication.

The McConnell 2013 dataset: FASTQ files of 110 cells were downloaded from the NCBI SRA database with accession number SRP030642 [48]. Adapter sequence was trimmed with the following parameters: "ILLUMINACLIP:adapter.fa:2:30:10:8:TRUE MINLEN:39". Paired-end reads were aligned to the hg19 human reference genome using BWA with *aln* and *sampe* options. The remaining steps are the same as those described in sections *Filtering*.

Statistical modeling of CNV frequencies and index of dispersion (IOD) levels

When modelling CNV frequencies, our null hypothesis was no difference in the frequency of CNVs in the AD brain when compared to healthy controls. The overdispersed and zero-dominated nature of the response variable, i.e. the frequency of CNVs, suggested that the data should be fitted using a zero-inflated negative binomial model. For this reason, we used the "glmmadmb" function (package: glmmADMB) [51] in R 3.6.3 with the following parameters: "zero-inflated = TRUE" and "family = nbinom1". The fixed factors of the model were diagnoses (AD and control), chromosomes (autosomes), sex (male and female), brain regions (temporal cortex, hippocampus CA1, hippocampus CA3, cerebellum, entorhinal cortex), and coverage per cell. The individual effect was added as a random factor. Note that sex could not be used as a fixed factor in the *van den Bos 2016* dataset because cells that remained after filtering only belonged to females. We also compared the difference between AD and control in terms of CNV frequency using *HMMcopy* estimates. The fixed factors of the model were diagnoses

(AD vs. control) and coverage per cell. The individual effect was added as a random factor.

When modelling the index of dispersion (IOD, the ratio between the variance of read coverage and the mean), we used the same approach as above. Levels of the response variable, IOD, was predicted using diagnoses (AD and control), brain regions (temporal cortex, hippocampus CA1, hippocampus CA3, cerebellum, entorhinal cortex) and coverage as explanatory variables using the "glmmadmb" function (package: "glmmADMB") [51] in R 3.6.3. Individual effects were added as a random factor. The distribution of the IOD was right-skewed and the model was run with the "family = gamma" parameter.

To compare the IOD across different brain regions, we used "lme" function (package: "nlme") in R 3.6.3 with diagnoses as fixed effects and the individual as a random effect.

Copy number statistics

After reads were mapped into the bins, read counts in each bin were divided by the mean read counts across bins for each cell. This value corresponds to the normalized read counts as calculated by *Ginkgo* (see [45]).

A Z_1-score for each CNV was calculated using the normalized read counts. It was calculated as the cell mean (mean normalized read counts across autosomes) minus the CNV mean (mean read counts between CNV boundaries) divided by the standard deviation (sd) of CNV:

$$Z_1\text{-score} = (mean_{cell} - mean_{CNV})/\, sd_{CNV}$$

The Z_2-score of each CNV was calculated by calculating the difference between the *Ginkgo*-estimated integer copy number state (1 or 3) and the observed normalized read count, dividing by the standard deviation (*sd*) of the normalized read counts:

$$Z_2\text{-score} = [a - b]/sd_{CNV}$$
$$\text{with} \quad a = ESTIMATED\_STATE_{CNV}$$
$$b = mean(OBSERVED\_READCOUNT_{CNV}).$$

CNVs with two standard deviations below or above the cell's mean and CNVs with Z_2-score smaller than or equal to 0.5 were kept in the analysis. Using these combinations, monosomy X (≥ 90% of the chromosome's length) was correctly predicted in 58.1% (217 of 373) of males in the uncorrected data.

Principal component analysis (PCA)

To remove experimental noise from the data, the following steps were applied for every cell: (1) one cell (x) at a time was discarded from the analysis. For the remaining cells, PCA was applied on the normalized read counts using the"prcomp" function with the parameter"scale = TRUE" in R 3.6.3. (2) *n* PCs that explained at least 90% of the variance in total were chosen. (3) To remove the effect of the chosen PCs from the focal cell x, a linear regression model with normalized read counts from cell x as a response, and the *n* PCs as explanatory variables was constructed using the R"lm" function. (4) Residuals from this model were calculated. (5) To prevent errors during a lowess fit of GC content (log transformation of negative residuals produces NaNs), we added the constant 1 to the residuals. If there still remained values ≤ 0, these were replaced with the smallest positive number for the focal cell x. (6) The resulting value was set as a new value of the focal cell x, and *Ginkgo* was run with the new values.

PCA of the normalized read counts across different datasets was performed in R 3.6.3 using the "prcomp" function with the parameter "scale = FALSE".

Results

Summary of the dataset

We used scWGS to determine the frequency of CNVs in the temporal cortex, hippocampal subfields cornu ammonis (CA) 1, hippocampal subfields cornu ammonis (CA) 3, cerebellum (CB) and entorhinal cortex (EC) of 13 AD patients and 7 age-matched healthy controls (Figs. 1A, 2A,B, Additional file 1: Table S1). The Braak stages of AD patients ranged between III and VI (Fig. 2C). Neuronal nuclei were isolated using either FACS (sorted with propidium iodide, $n = 12$) or LCM (sorted with cresyl violet, $n = 1552$), the latter performed on frozen brain slices (Fig. 1B, see Methods). LCM-isolated non-neuronal "blank" regions were used as negative control ($n = 10$). The LCM method, although more difficult to implement than FACS, was chosen to ensure the selection of nuclei of pyramidal neurons for sequencing, known to be particularly sensitive to AD [21]. For technical comparison, neurons of a single individual were collected both using FACS ($n = 12$) and LCM ($n = 64$) (see Methods). scWGS libraries were prepared using GenomePlex whole-genome amplification and specific adapters were inserted using Phusion® PCR. Paired-end reads were mapped to the human reference genome, followed by stringent filtering to obtain uniquely mapped reads (see Methods). This resulted in a median of 276,446 reads, corresponding to a coverage of 0.006X per LCM-isolated cell (range [133–1,909,016] reads and [0.000003X–0.04X] coverage) (Fig. 3A).

CNVs were predicted using *Ginkgo*, which uses circular binary segmentation (CBS) to estimate deletion or duplication events [45]. Negative controls ($n = 10$) and FACS-isolated neurons ($n = 12$) were analyzed separately and are not included in the main results. *Ginkgo* was run

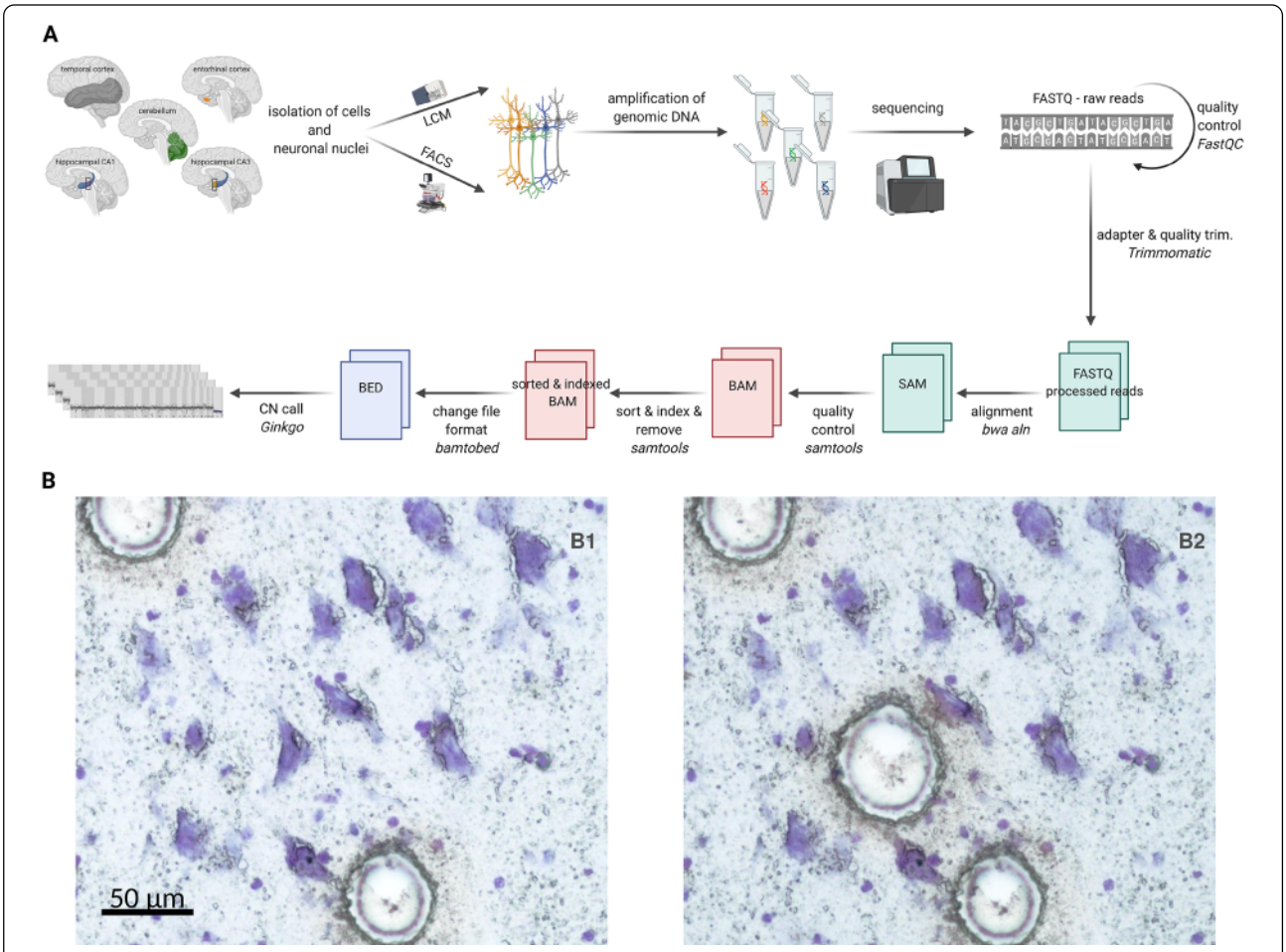


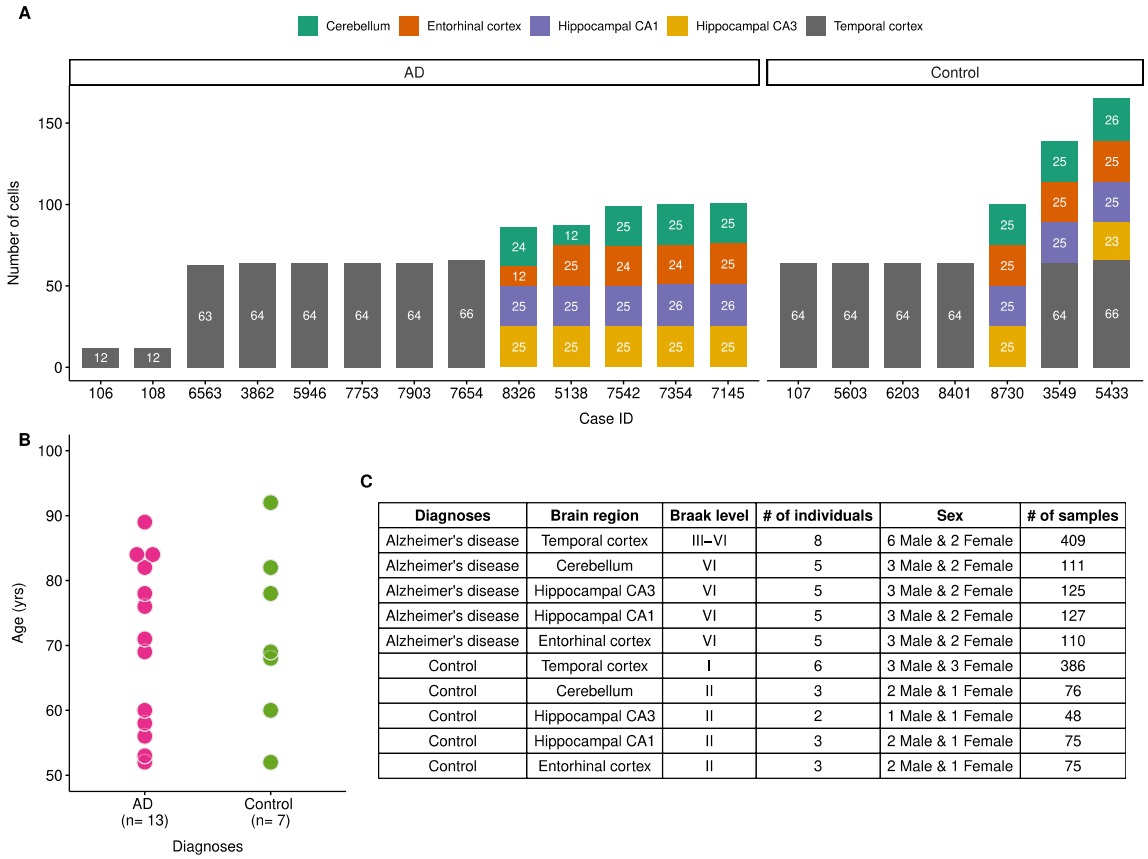
Fig. 1 Schematic of the workflow and information about the samples. **A** The pipeline of NGS data analysis and CNV detection. **B** Images from a frozen hippocampal brain slice stained with cresyl-violet showing a pyramidal cell before *(B1)* and after *(B2)* laser capture microdissection-based isolation process using the PALM device. Circles in *B1* indicate positions where two pyramidal cells have already been isolated just prior to the picture being taken. Scale bar, 50 µm

on our dataset with $n=1542$ cells, while in parallel, we also analyzed two published scWGS datasets: one by van den Bos and colleagues ("van den Bos 2016"), comprising $n=1469$ cells from healthy and AD brains (median coverage 0.005X) and another by McConnell and colleagues ("McConnell 2013"), comprising $n=110$ cells from healthy brains (median coverage 0.047X) (Fig. 3D) [5, 48]. Note that the van den Bos 2016 dataset includes only 61% of cells produced in that study, because data from cells filtered for high noise levels were not published and thus could not be included here.

LCM-isolated cells show a high frequency of depth variability

We first evaluated the sensitivity and specificity of our bioinformatics pipeline on scWGS data using trisomy-21 in DS and monosomy-X in males in published data. Analyzing $n=34$ neuronal nuclei from DS individuals [5], trisomy-21 was correctly predicted across all samples without any false positive or false negative calls. In addition, monosomy-X was accurately predicted in 94.2% (338 of 359) of cells from males across the two published datasets [5, 48].

Ginkgo includes an algorithm that uses the distribution of read depth across the genome to infer the average DNA copy number of each cell, which is estimated within a range of 1.5 to 6. It would be expected that the majority of human neurons would carry on average two copies of each autosome, although high frequencies (10–35%) of hyperploid neurons have also been reported, especially in AD brains [10].



Diagnoses	Brain region	Braak level	# of individuals	Sex	# of samples
Alzheimer's disease	Temporal cortex	III–VI	8	6 Male & 2 Female	409
Alzheimer's disease	Cerebellum	VI	5	3 Male & 2 Female	111
Alzheimer's disease	Hippocampal CA3	VI	5	3 Male & 2 Female	125
Alzheimer's disease	Hippocampal CA1	VI	5	3 Male & 2 Female	127
Alzheimer's disease	Entorhinal cortex	VI	5	3 Male & 2 Female	110
Control	Temporal cortex	I	6	3 Male & 3 Female	386
Control	Cerebellum	II	3	2 Male & 1 Female	76
Control	Hippocampal CA3	II	2	1 Male & 1 Female	48
Control	Hippocampal CA1	II	3	2 Male & 1 Female	75
Control	Entorhinal cortex	II	3	2 Male & 1 Female	75

Fig. 2 Sample information. **A** Bar plot showing the number of cells that have been sequenced for each individual. Brain regions are illustrated in different colors (see the colour key on the top of the figure). **B** Dot plot showing age of AD (pink) and control (green) individuals. **C** The table summarises sex, diagnoses and Braak level of the individuals. Detailed information on the individuals is available at Additional file 1: Table S1-A

Applying *Ginkgo* on the two published datasets, we found that for 99.9% (1577 of 1579) of cells the estimated average copy number lies within [1.9–2]. Using the same algorithm on our dataset, however, only 45% (687 of 1542) of the cells had average copy numbers estimated within the [1.9–2] range; i.e. 55% were non-euploid. Although hyperploid neurons have been described in control brains at ~ 10% frequency using FISH [10], the observed non-euploidy estimates suggest that our dataset carries particularly high levels of variability in read depth. These differences, in turn, could be related to the LCM protocol used, as the published scWGS experiments had used FACS.

To investigate this possibility, we compared the quality metrics of cells we had collected using FACS or LCM for this study. These metrics were mapping proportion (the number of mapped reads/ the total number of reads), coverage, and index of dispersion (IOD, the ratio between the variance of read coverage and the mean). FACS-isolated cells had higher sequencing coverage and mapping proportions than the LCM-isolated ones (Wilcoxon two-sided rank-sum test, $p < 0.0001$ and $p < 0.001$ for coverage and mapping proportion, respectively) (Fig. 3A, B). Note that the difference in coverage variability between FACS and LCM has not been reported elsewhere. In addition, FACS-isolated cells had low IOD values, indicating less variation in sequence depth than the rest of the samples (Kruskal–Wallis test, $p = 1.5e{-}07$) (Fig. 3C). Because our LCM and FACS samples originated from different brain regions with different cell type proportions, we also asked whether such differences could explain the observed LCM vs. FACS differences. To rule out this possibility, we compared the index of dispersion value of the cells that were taken from the temporal cortex of the same individual using FACS ($n = 12$) and using LCM ($n = 64$). We found a significant difference in the direction of higher variability in LCM (Wilcoxon rank-sum test $p < 0.001$), indicating that the observed variability between LCM and FACS can not be simply explained by differences in cell type proportion among brain regions. We note that the higher noise observed in LCM data was not solely due to higher

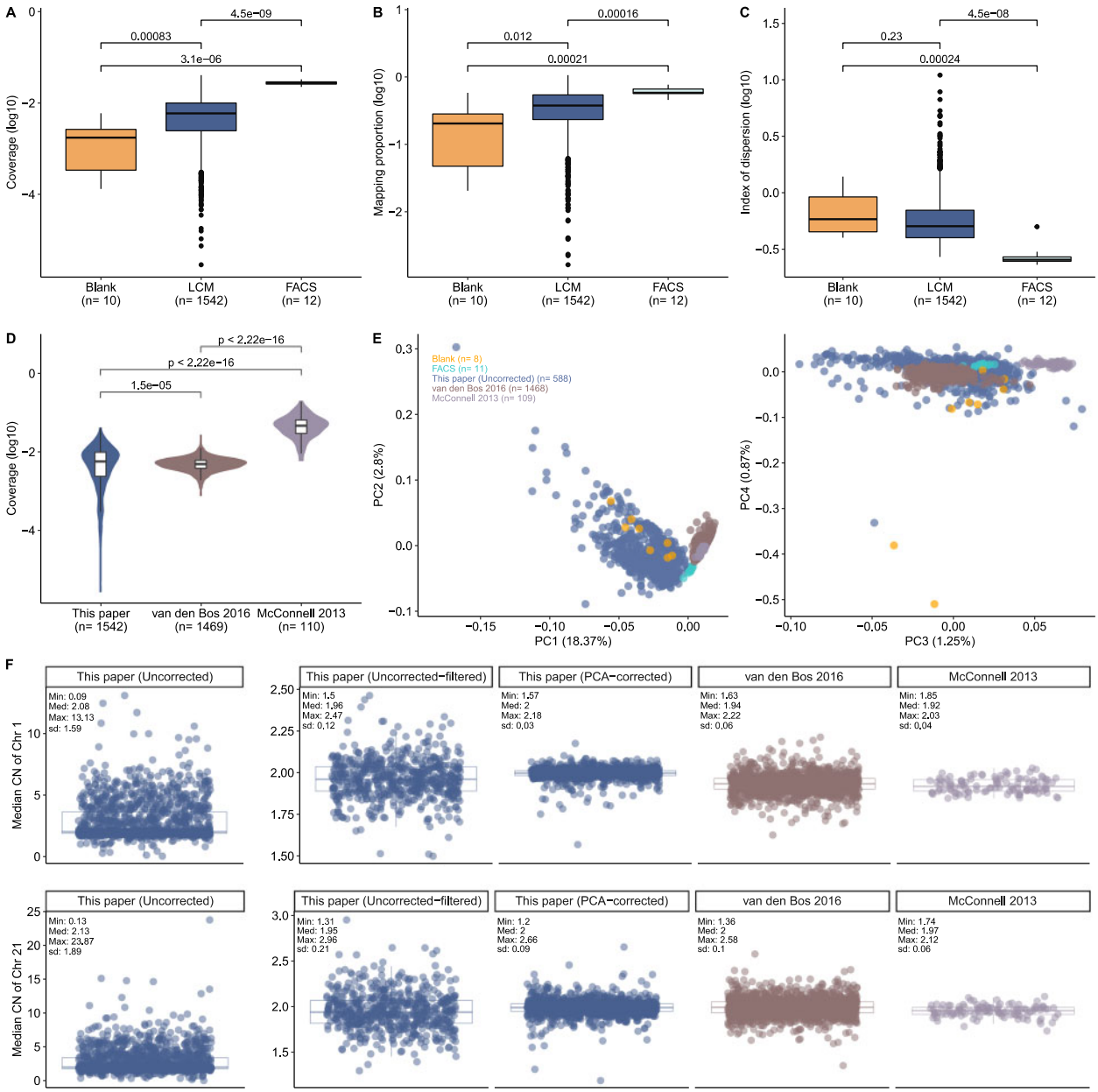


Fig. 3 Comparison between different cell isolation methods and published datasets. Box plots showing the distribution of coverage (**A**), mapping proportion (**B**), index of dispersion (**C**) among FACS-isolated, LCM-isolated and LCM-isolated blank samples. *P*-values were calculated using Kruskal–Wallis test among groups and Wilcoxon rank-sum test between groups. (**D**) Violin plot showing the distribution of coverage among different datasets. P-values were calculated using Wilcoxon rank-sum test between groups. This study, including only LCM: blue; van den Bos 2016: brown; McConnell 2013: purple. (**E**) A principal components analysis (PCA) was performed using the normalized read counts across autosomal bins ($n = 5243$) in published datasets and this study. Because they dominated the PCs, cells deviating from the [1.9–2] range were not included in the analyses. The number of cells for each dataset are indicated on the plot. X-axes illustrate PC1 and PC3 that explain 18.4% and 1.3% of the total variance, respectively. Y-axes show PC2 and PC4 that explain 2.8% and 0.9% of the total variance, respectively. (**F**) Boxplots showing the distribution of median CN of chromosome 1 (chr1, upper part of the figure) and chromosome 21 (chr21, lower part of the figure) across bins ($n = 440$ and $n = 68$ for chr1 and chr21, respectively). Each point corresponds to the median CN of each cell. Minimum ("Min"), median ("Med"), maximum ("Max") and standard deviation ("sd") of each distribution were shown on the boxplot. Cells that deviated from the [1.9–2] range were excluded from the analyses to be consistent with our filtering criteria (except for the uncorrected datasets). This study [Uncorrected ($n = 1337$), Uncorrected-filtered ($n = 588$), PCA-corrected ($n = 1301$)]: blue; van den Bos 2016 ($n = 1468$): brown; McConnell 2013 ($n = 109$): purple

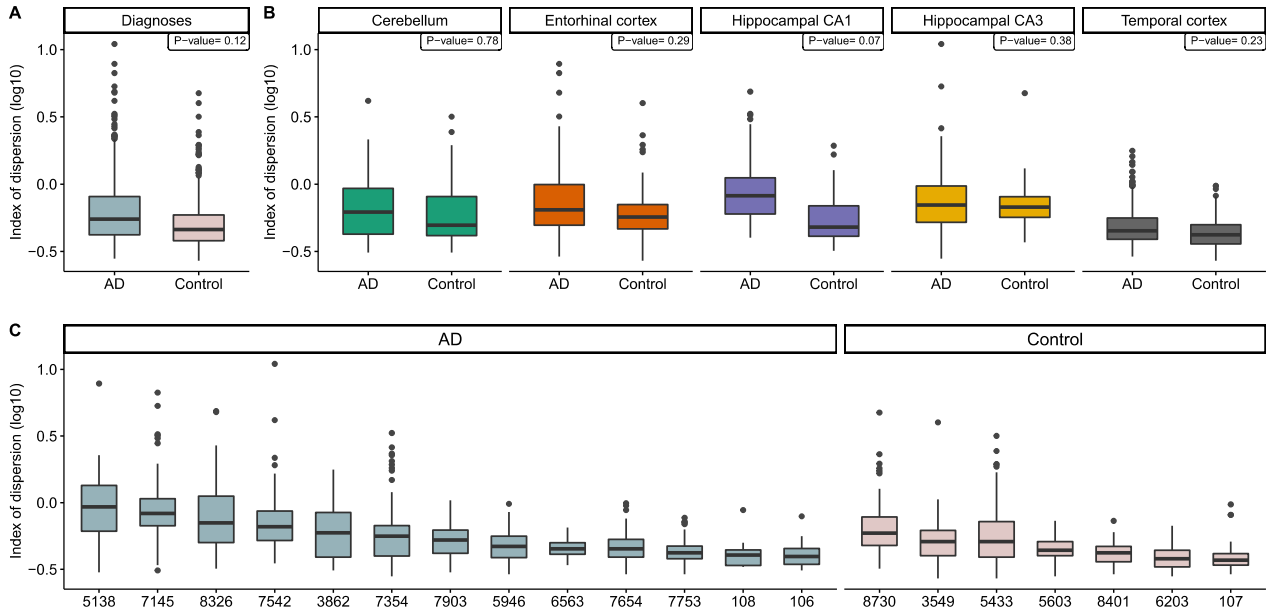


Fig. 4 The distribution of index of dispersion ("IOD") for LCM-isolated cells ($n = 1542$) according to **A** diagnoses (Alzheimer's disease ("AD"), control), **B** brain regions (hippocampal subfields CA1 ("Hippocampal CA1"), hippocampal subfields CA3 ("Hippocampal CA3"), Entorhinal cortex, Cerebellum, Temporal cortex). For each brain region, we tested whether AD diagnosis was predictive of IOD using a linear mixed-effects (lme) model (see Methods). Individuals were added as a random factor. Across all tested brain regions, differences were only marginally significant ($p = 0.069$). **C** The distribution of IOD across individuals ($n = 20$). Box plots were ordered by the median. Y-axes illustrate the IOD values on the log10 scale

genome coverage, as the FACS-based data from the van den Bos 2016 dataset had a median coverage comparable to ours (0.005X vs. 0.006X), but did not show comparable variability as in our LCM data (Fig. 3D). These differences in IOD between LCM and FACS could be potentially explained by the higher sensitivity of the LCM procedure to experimental noise, compared to FACS. Alternatively, they could partly represent abnormal nuclei selected out in FACS but captured by LCM.

We next investigated the possibility that underlying variation may be caused by technical and/or biological factors. For this, we used a generalized linear mixed model (GLMM) to explain IOD (the response variable) per LCM-isolated cell ($n = 1542$) as a function of diagnosis (AD vs. control), genome coverage, and brain region as fixed factors, and individual as a random factor (see Methods). Note that p-values of the pairwise differences between AD and control (Fig. 4A–C) was calculated using a linear mixed-effects model (see Fig. 4 legend). We found that coverage has a significant negative effect on IOD, as may be expected ($z = -21.06$, $p < 0.0001$). Compared to the cerebellum, the region least affected by neurodegenerative diseases [34], we found a significantly high IOD for the entorhinal cortex ($z = 2.61$, $p < 0.05$), hippocampal CA1 ($z = 3.34$, $p < 0.001$) and hippocampal CA3 ($z = 3.75$, $p < 0.001$), but not for the temporal cortex ($z = -0.28$, $p = 0.78$) (Fig. 4B). Finally, neurons from control individuals have slightly less IOD than AD patients ($z = -1.93$, $p = 0.054$) (Fig. 4A–C). This result might suggest a tendency for neurons of AD patients to carry more variable DNA content and is consistent with cytometry analyses reporting a high occurrence of hyperploid neurons in the AD brain [10]. Although these findings imply a role of biological factors in read count variation within cells, it still remains possible that confounding technical factors influence our data. Given this uncertainty about the source of variability, we continued the analyses by filtering our dataset to remove the most variable cells.

No significant difference in CNV frequency between AD and control in the "uncorrected-filtered" dataset

We then used *Ginkgo* to call CNV events from the "uncorrected-filtered" dataset ($n = 882$ cells from 13 AD patients, and $n = 660$ cells from 7 healthy controls). We found 19,608 events in 882 cells from AD patients (22.2 per cell), and 14,844 events in 660 cells from healthy controls (22.5 per cell). We then tested the observed

frequency difference between AD and control using a GLMM with a negative binomial error distribution (see Methods). The response variable (the frequency of CNVs) was predicted using a combination of fixed factors, including diagnoses, chromosomes, brain regions, sex and coverage (Fig. 6D). The individual effect was added as a random factor. We found no statistically significant difference between AD and control across all tested combinations (GLMM, $p \geq 0.17$; Additional file 3: Table S3).

CNV estimation from low coverage scWGS data is known to be highly sensitive to technical noise, and a large proportion of the called CNV events likely represent false positives. We thus decided to filter both cells and CNV events in our dataset to obtain a more reliable dataset [6, 52, 53]. We started by removing the most highly variable cells among the LCM-isolated ones ($n = 1542$) using the following criteria. First, 13% (205 of 1542) of the cells with a low number of reads (< 50,000) were discarded from the analysis (see Methods). Second, as most cells are expected to be diploid, and also given that the *Ginkgo*-estimated copy number (CN) profiles of 99% of cells in the McConnell 2013 and van den Bos 2016 datasets were observed to lie between [1.9–2], we excluded those cells with CN values beyond this range (54% excluded, 726 of 1337). Third, we filtered out 23 of the remaining 611 cells (4%) that showed extreme CNV intensity, which we defined as three or more chromosomes of a cell carrying predicted CNVs that cover > 70% of their length (Fig. 6A). Information about the remaining cells ($n = 588$) is provided in Additional file 2: Table S2 and Additional file 4: Fig. S1.

From these 588 cells, we called 3521 CNVs (~ 5.9 events per cell) in the uncorrected data, which we call the "uncorrected-filtered" dataset. We further applied a number of conservative filtering criteria to remove potential false positives: (1) We only included megabase scale CNVs (≥ 10 Mb), considering that detection of small events with low coverage data will be unreliable. (2) We limited the analyses to 1-somy and 3-somy events, assuming that most somatic CNVs involving chromosomes or chromosome segments would involve loss or duplication of a single copy. (3) We only included CNVs with unique boundaries across all analysed cells, assuming that somatic CNV breakpoint boundaries should be generally randomly distributed across the human genome. (4) We removed CNVs on the proximal portion of the chr19 p-arm, where frequently observed duplications were previously reported as low coverage sequencing artifacts [43]. (5) To ensure the reliability of the CNV signal, we calculated a standard *Z*-score for each CNV that reflects the deviation in read count distribution in that region compared to the rest of the cell (which we call Z_1, see Methods), and only accepted CNVs with absolute values of Z_1-scores ≥ 2. (6) We reasoned that read counts in a real CNV should be closely clustered around expected integer values (e.g. 1 or 3). To assess this, we calculated a *Z*-score for the deviation from the expectation (called Z_2), and only accepted events with absolute values of Z_2-scores ≤ 0.5 (see Methods, Fig. 6A, Additional file 4: Fig. S3).

After CNV filtering, we found 12 CNV events across 295 cells in 13 AD individuals and 4 CNV events across 293 cells in 7 controls. Among the 295 pyramidal neurons analyzed from the 13 AD patients, we found 10 deletions (3.39% per cell) and 2 duplications (0.68% per cell) (Fig. 6B). These events ranged in size from about 10.14 to 77.01 Mb (median: 19.31 Mb) and were observed in the temporal cortex and the entorhinal cortex. Of the 293 neurons from 7 control brains, 1 deletion (0.34% per cell) and 3 duplications (1.02% per cell) were detected in the temporal cortex with a size range of 10.81 to 54.67 Mb (median: 14.51 Mb) (Fig. 6B). Again testing the CNV frequency differences between AD and control brains using a GLMM, we found no statistically significant effect (GLMM, $p \geq 0.88$) (Additional file 2: Table S2, Additional file 3: Table S3).

We also implemented an alternative algorithm, *HMMcopy* [46], to predict CNVs (see Methods). Overall, 75% (12/16) of the HMMcopy predictions overlapped with the CNV events that we found after filtering the uncorrected Ginkgo predictions. Comparing predicted CNV event frequencies between AD and control we again found no significant difference ($z = -1.34$, $p = 0.18$).

A PCA-based denoising approach minimizes within-cell depth variability

To gain further insight into within-cell variability in our dataset (the uncorrected-filtered version) compared to the two published scWGS datasets, we calculated the median CN of chr1 and chr21 (the largest and smallest chromosomes) across all three. We still found conspicuously higher within-cell variation in our dataset, despite having discarded highly variable cells (Fig. 3F). We then used the autosomal normalized read counts to perform a PCA on the uncorrected-filtered data and published datasets. We also included blank (negative control) samples and FACS-isolated cells to illustrate how reads counts from these two groups relate to others. According to the PCA, LCM-isolated uncorrected-filtered data and blank samples were separated from the published datasets and FACS-isolated cells (Fig. 3E). This result might also highlight distinct profiles of LCM-isolated cells.

We then sought an approach that could reduce this elevated within-cell variability in read depth, assuming it is of technical origin and possibly related to the LCM procedure. Experimental biases could involve