Specific Aims

Emerging evidence suggest that contrary to previous assumption not all cells of the human body have identical DNA sequence. Besides **single nucleotide variation (SNV)**, dividing cells accumulate larger somatic **structural variants (SVs).** These include copy number variations (CNVs, i.e., duplications and deletions), as well as inversions and translocations, all involving from few hundreds to several millions nucleotides. In addition, a high degree of retrotransposon mobilization is thought to occur in the developing human brain. Larger somatic SVs are likely to intersect genes and/or regulatory elements and may have considerable phenotypic effects. Furthermore, depending on their allele frequency they are likely to confound the relationships between genotype and phenotype, if "genotype" is assumed to be uniform throughout the body. Currently there is no comprehensive estimate of the number and allelic frequency of genomic variation in somatic cells. All studies so far have been done on the population level, which is not sensitive enough to evaluate the full range of somatic variants and precludes a true estimate of their frequency.

Somatic variations are more frequent in proliferating cells and are more likely to occur during the early phases of embryonic development [2-4]. The human cerebral cortex displays a very high degree of mitotic expansion and thus is likely to be particularly susceptible to accumulate somatic variations during ontogenesis. While some somatic variants are likely to be eliminated via negative selection, a portion may persist because they may be initially neutral or confer adaptive properties to the cells. The accumulation of somatic variation in the brain may play a role in adaptation, learning and gene-environment interaction and may play a role in shaping individual susceptibility and resilience to neuropsychiatric disorders. Developmental disorders such as autism have been found to be associated with higher frequency of *de novo* CNVs [6-10]. However, it is currently unclear whether these higher rates of *de novo* CNVs reflect those present in the germline, as commonly assumed, or rather reflect somatic variants in the blood.

In this proposal we analyze the full extent of somatic mosaicism in the embryonic human cerebral cortex and basal ganglia by comparing their genomes with genomes of clonal cell populations from these tissues and by comparing cortex and basal ganglia genomes with another tissue from the same embryo, i.e., the blood. *We will sequence the genomes of tissues and clonal cell populations in order to identify the full range (sizes and types) of somatic variants and estimate their overall allelic frequency in the brain.* Our computational analyses of the sequence around variants, particularly CNV breakpoints, will also give important clues as to their origin.

Aim 1. Construct a map of somatic variation present in progenitor cells of the cerebral cortex and basal ganglia and estimate their frequency in brain tissue as well as in the blood.

1a. Generate clonal cell populations from the cerebral cortex and basal ganglia of postmortem human embryos at 14-15 weeks of gestation and obtain the complete genomic DNA sequence of each clone. Using computational analyses, discover genomic variants manifested in each clone as compared to the original cell population and to another tissue (blood). Validate variant existence and, for large CNVs, determine their precise breakpoints by PCR and qPCR.

1b. Using the dataset of validated genomic variants manifested in progenitor clones, utilize an ultra deep targeted re-sequencing approach, PCR and digitalPCR to determine the presence and allele frequency for specific genomic variants in the original brain tissue(s) and in the blood.

Aim 2. Determine the impact of somatic variants on mRNA transcripts. Analyze the transcriptome of the clonal cell populations by RNA-Seq to assess whether clone-manifested genomic variants correspond to variations in transcripts. Perform gene network analyses to evaluate the effect of somatic genomic variation on modules of functionally related genes. Identify the module(s) whose expression profile differs in clonal cell populations as compared to original brain tissue and determine whether these modules are enriched in somatic variants. A similar enrichment analysis will be performed across individuals in our dataset and in larger datasets of developing human brain, with the aim of identifying those somatic variants that may potentially be causative for expression change in sets of functionally related genes. Our expectation is that these analyses will allow us to assess whether somatic variation has general implications for processes of brain development.

Aim 3. **Investigate the most likely biological origin of somatic variants.** Analyze sequence features at variation sites, correlate variants with recombination hotspots, CpG islands and histone marks to yield a hypothesis about mechanisms responsible their creation (like recombination, double stranded breaks, etc.)

Together, these specific aims will provide the **first comprehensive estimate of the number and allelic frequency of genomic variation in somatic cells of the brain** and will yield hypotheses about mechanisms responsible for their creation as well as their significance for brain development.

Research Strategy

Significance

Genomic variants may either be inherited (i.e., generated in the germline) or caused by *de novo* mutation in somatic tissues. Emerging evidence suggest widespread genomic mosaicism in somatic lineages. Somatic variations are believed to be one of the causes of cancer [11, 12], aging [13, 14] and several diseases [15-17]. While it has been established [15-17] for many decades that cells in the human body carry somatic variations, their extent is still to be determined [13, 14, 18]. For example, in recent studies by exome sequencing, multiple low frequency somatic single nucleotide variant (SNVs) were found to be transmitted into human induced pluripotent stem cell (hiPSC) lines [19, 20]. Similarly, in our ongoing study (see **Preliminary Data**) in which we examined by whole genome sequencing the extent of structural variants (SVs) in hiPSC derived from skin fibroblasts, we discovered that between 50 and 80% of the CNVs found in hiPSC were already present in the fibroblasts of origin. Moreover, assuming that an hiPSC line is an expansion of a single cell, we estimated that ~38% skin fibroblast cells carry somatic CNVs, suggesting somatic variability within cells in human body [21].

Differently patterned instances of somatic mosaicism in CNS regions have been shown to be present in monozygotic twins [22] and throughout different tissues within an individual [23, 24]. Additionally, retrotransposition of ancient virus-like elements (mostly L1, Alu and SVA) has very recently been proposed as a mechanism for the occurrence of somatic mosaicism in the mammalian brain [25, 26]. Emerging evidence suggest that substantial variation in the structure of DNA between different regions of the human brain is mediated by retrotransposon mobilization and re-insertion into area of the genome that are highly expressed. While retrotransposition is now emerging as an important cause of somatic variation in the brain [25, 27, 28], the real extent of this phenomenon is unclear and whether it might result in changes in gene expression and play a role in cellular differentiation and ultimately brain function is unknown. Retrotransposon re-insertion can directly change the DNA structure and can furthermore facilitate additional CNV formation, as retrotransposons have high content of repetitive sequence elements.

The emergence of new somatic variants and retrotransposon mobilization in neural cells could be developmentally adaptive by varying the genetic makeup and epigenetic regulation in innovative ways and allowing selection of favorable mutations during the phase of amplification of neural progenitors cells [27]. However, excessive somatic variation and element mobilization may be detrimental by knocking out genes and destabilizing the genome. Currently, possible effects of somatic variations on gene expression and cell phenotypes are unknown.

The mechanisms by which different types of genomic variants arise during development is not clear, however it is well known that SNVs and indels are created as a result of replication errors and spontaneous mutations [29]. Suggested mechanisms for creating SVs are either replication errors or post-replication chromosome crossing over [5, 30, 31], hence also mostly related to cell division. Furthermore, the results of Shi et al. [32] and Kubo et al. [33] strongly suggest that cell division as a major rate-limiting factor for retrotransposition. In our discovery study of novel retrogenes (i.e., those created by retrotransposition of mRNA) absent from the reference genome (see **Preliminary Data**) we observed that retrogenes are created from genes that have the highest expression at transition from M to G1 phase of the cell cycle. We hypothesize that expression of genes during cell division gives their mRNA the highest chance of being retrotransposed.

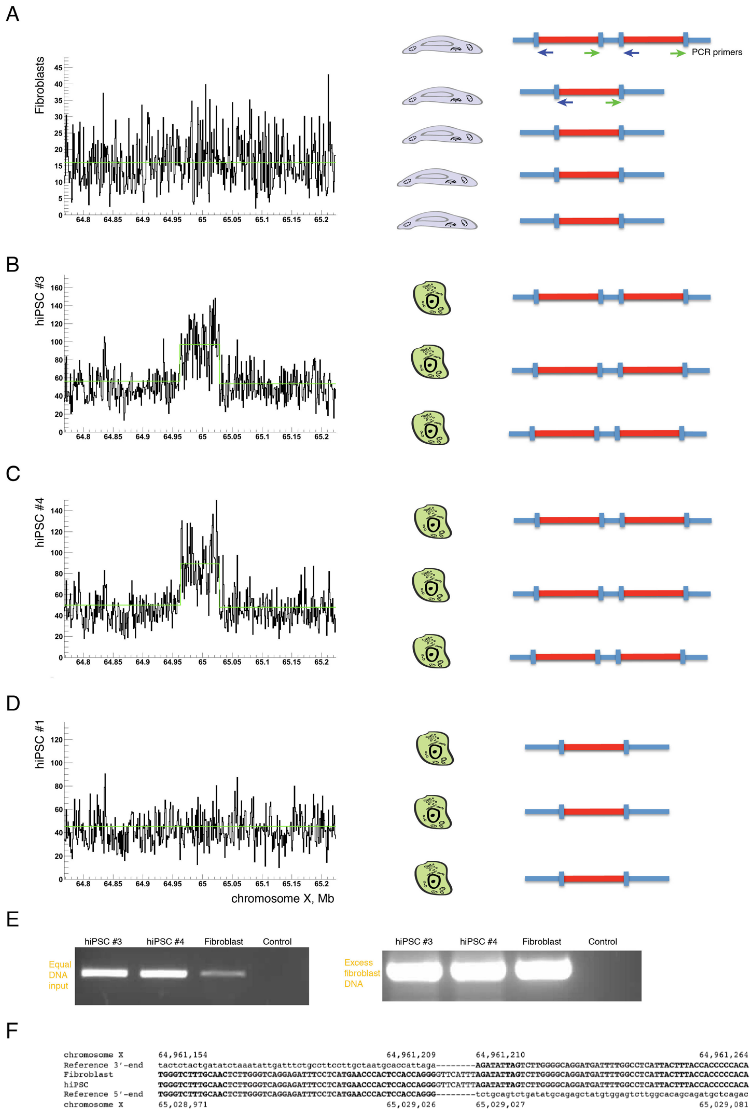
In summary, it is very likely that all types of somatic variations occur more frequently in proliferating cells and, in particular, in early phases of embryonic development during extensive growth. The brain, and in particular the human cerebral cortex, displays a very high degree of mitotic expansion during a restricted period in ontogenesis, and thus **cortical progenitor cells are likely to be particularly susceptible to accumulate genomic variants during development**. Somatic mutations are likely to play a role in cortical development and perhaps in disorders of the cerebral cortex. The accumulation of somatic variation in the brain may play a crucial role in adaptation, learning and gene-environment interactions, and it may be a significant factor in shaping susceptibilities to neuropsychiatric disorders. Furthermore, persons are likely to be different in their catalog of variants, suggesting that genomic mosaicism could contribute to interindividual variability.

In this proposal, we analyze the extent of **somatic mosaicism in forebrain neural stem cells** of mid-fetal human brain. Somatic variations are difficult to analyze because typically (unless they happen in the early stages of development), they are present in a subset of cells within a tissue and thus are not easily detectable when analyzing large pools of cells. In principle, the ideal approach for the analysis of somatic variation is single cell sequencing and analysis [34]. However, this is still extremely challenging as whole genome amplification (a prerequisite for single cell sequencing) amplifies only about ~6% of a genome. A valuable alternative is single cell clonal expansion (growing a colony of cells from a single cell) followed by analysis.

More specifically, in this project we will **discover rare genomic variants** manifested in clonal population derived from a single neural stem/progenitor cell, amplified using the neurosphere assay. We will them conduct follow up experiments to prove the existence of the variants in the original sample of stem/progenitor cells. Additionally, we will compare genomes of different tissues to find **more frequent somatic variants that are tissue specific**.

Traditionally, new variants have been detected with capillary based sequencing [35-37] and SNP [38] or CGH [39] array applicable for CNV discovery. Recent advances in DNA sequencing technology have further enabled fine-scale identification of all types of variants in an unbiased and comprehensive manner. We will, therefore, apply DNA sequencing to discover a comprehensive catalog of variants manifested in single cells of the embryonic human cerebral cortex and their clonal derivatives.

Besides understanding the full extent of somatic variations, it is important to assess how widespread particular variants are among tissue within an organism. Variants established earlier in ontogenesis will be common between tissues, while those arising during organogenesis will be private to a given tissue or organ. While this project focuses on the brain, we will determine to what extent genomic somatic variants discovered in the brain (which derive from neuroectoderm) are private to this tissue or shared with a tissue of different embryological origin, i.e., blood (which derives from the mesoderm).



```
chromosome X     64,961,154                                             64,961,209   64,961,210                                             64,961,264
Reference 3'-end tactctactgatatctaaatattgatttctgccttccttgctaatgcaccattaga--------AGATATTAGTCTTGGGGCAGGATGATTTTGGCCTCATTACTTTACCACCCCCACA
Fibroblast       TGGGTCTTTGCAACTCTTGGGTCAGGAGATTTCCTCATGAACCCACTCCACCAGGGGTTCATTTAGATATTAGTCTTGGGGCAGGATGATTTTGGCCTCATTACTTTACCACCCCCACA
hiPSC            TGGGTCTTTGCAACTCTTGGGTCAGGAGATTTCCTCATGAACCCACTCCACCAGGGGTTCATTTAGATATTAGTCTTGGGGCAGGATGATTTTGGCCTCATTACTTTACCACCCCCACA
Reference 5'-end TGGGTCTTTGCAACTCTTGGGTCAGGAGATTTCCTCATGAACCCACTCCACCAGGG--------tctgcagtctgatatgcagagctatgtggagtcttggcacagcagatgctcagac
chromosome X     65,028,971                                             65,029,026   65,029,027                                             65,029,081
```

Figure 1. Genomic heterogeneity in fibroblasts. Example of manifested CNV (duplication) in iPSC, which originates in fibroblasts in small allele frequency. (A-D) The duplication allele was present in the fibroblasts at ~20% allele frequency and escaped detection. Two of the three produced hiPSC were made from the cells with the duplication allowing for its confident detection. (E) Subsequent PCR amplification revealed that the duplication was indeed present in fibroblasts. (F) Sequencing of amplified band showed that the duplication breakpoints are exactly the same in fibroblasts and the two hiPSC, further proving that the event was present in fibroblasts and was carried over to hiPSC. Location of primers for PCR amplification is depicted as an example in (A).

Innovation

Our project contains five points of conceptual and technical innovation:

1. The idea of somatic mosaicism is just emerging. We detected somatic CNVs in skin fibroblast (see below), and retrotransposition of ancient virus-like elements has very recently been proposed as a mechanism for the occurrence of somatic mosaicism in the mammalian brain.
2. No one has yet attempted to examine the extent of these phenomena in the embryonic human brain. We approach this question using an innovative experimental design that will involve isolating clonal populations of neural precursor cells from the embryonic brain to obtain their complete DNA sequence.
3. Genomic data will be analyzed by recently developed bioinformatics tools allowing the discovery of all variants (including SVs) by comparing DNA sequence of clonal cell populations with the tissue of origin.

4. We will analyze a new type of genomic variations – novel processed pseudogenes. This type of variation has not been studied so far.
5. The presence of SVs in the tissue of origin will be ascertained by digital PCR, a recently developed technology allowing high-resolution analyses of DNA sequence variants.

Preliminary Data

Discovering somatic CNVs. We have previously performed discovery of somatic CNVs in skin fibroblast cells [21].. Briefly, for each fibroblast sample from 7 individuals we have produced 3 human induced pluripotent stem cell (hiPSC) lines. We then sequenced most of those lines along with the corresponding fibroblast samples in order to compare their genomes. As reprogramming is very inefficient, the resulting hiPSC colonies are very likely to be produced from a single cell and, thus, represent a clonal cell population. Analysis of their genomes can unmask variations present in the founder cell but absent from most of other fibroblast cells. Using CNVnator [40] software that utilizes the read depth approach, we have predicted CNVs that are manifested in hiPSC lines (i.e., present in hiPSC lines and absent in fibroblasts), termed line-manifested CNVs (LM-CNVs) and consisting of deletions, tandem duplications and dispersed duplications. These could be low allele frequency CNVs in fibroblasts that had been transmitted to hiPSCs but could also be CNVs generated *de novo* during the hiPSC de-differentiation process and/or hiPSC proliferation. Using qPCR we have validated 28 of the predictions.

Two manifested tandem duplications with almost the same boundaries were predicted in two different lines originated from the same fibroblast sample (**Fig. 1**). We hypothesized that these CNVs were the same somatic CNV pre-existing in fibroblasts. To prove that, we designed PCR primers (**Fig. 1**) that would give an amplification product only if the duplication was present in tandem in the genome. PCR analysis revealed that the duplication was indeed present in both hiPSC and fibroblasts. Because the PCR amplicon in fibroblast was much weaker than that in hiPSC, we concluded that this duplication was present at a small allele frequency in fibroblasts. Sanger sequencing of the amplicon in hiPSCs and fibroblast revealed the exact same breakpoints [21]. By genotyping this region with CNVnator [40], we estimated that the duplication allele was present in fibroblast with ~20% allele frequency, which explains why it had escaped detection.

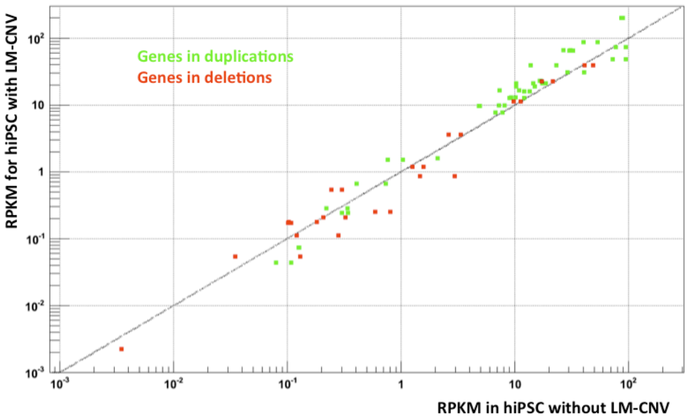


Figure 2. Comparison of gene expression (RPKM) intersecting line-manifested CNVs in iPSCs. Clear tendency (p-value of 0.02 by Fischer's exact test) of increase in expression for genes in duplications and decrease in expression for genes in deletions can be observed.

Using PCR amplification across CNV junctions, we could confidently detected 7 additional somatic CNVs in fibroblasts of 4 individuals. From read depth genotyping by CNVnator and qPCR we estimated that all of the somatic CNVs (except the one discussed above) have much smaller allele frequency than 20%. *We, therefore, demonstrated the applicability of single cell clonal expansion for discovering somatic CNVs in a tissue sample. Moreover we show that we can achieve breakpoint resolution of CNVs, which can be used for analysis of their origin.* For example, the CNV in **Fig. 1** is a tandem duplication, which strongly suggests recombination as a mechanism for its creation. However, it is apparent that sequences around breakpoints are not homologous, and there is an insertion at breakpoints (**Fig. 1F**). Therefore, non-homologous end joining [5] is the likely mechanism responsible for creation of this CNV.

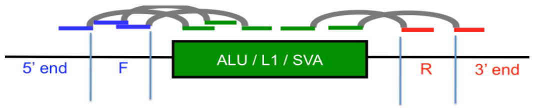


Fig.3A Paired-end TE mapping: 5' (blue) or 3' (red) end fragment maps to unique sequence, and the other fragment (green) to non-unique TE database sequence.

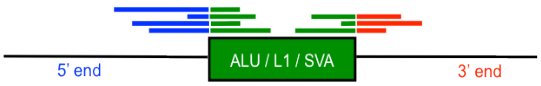


Figure 3B: Split-read TE mapping: part of each informative read can be uniquely mapped (red or blue), and the other part can be mapped to the TE database.

Discovering somatic SNVs. Recent studies have reported discovery of somatic SNVs [19, 20]. Similarly to our study, they studied hiPS cells and the design of their study was conceptually similar to the one we used in our analysis (see above). Using targeted ultra-deep sequencing they have shown that a large fraction of SNVs manifested in hiPS lines was actually present as somatic variants in the original somatic tissue samples. *These analyses demonstrate that not only CNVs/SVs but also single nucleotide somatic variants can be discovered through clonal expansion and sequencing.*

Effect of LM-CNVs on gene expression. To examine whether LM-CNVs affect expression of the genes that they intersect, we compared the expression levels (as obtained from deep coverage RNAseq data) of the genes whose ORFs intersected a CNV in hiPSC lines carrying the LM-CNV vs. hiPSC without the CNV. We limited our comparison to pairs of hiPSC lines derived from the same person to ensure that the differences were not the result of interpersonal genomic variability. As we produced three hiPSC lines for each person while the observed LM-CNVs are typically found in only one hiPSC line, we could make two expression comparisons per gene. Our selection resulted in 40 genes for which we had 80 data points: 52 for duplications and 28 for deletions (**Fig. 2**). For genes in duplicated CNVs the expression typically increased, and this was found in 36 out of 52 cases (69%). For deletions the trend was weaker but still noticeable; in 16 of 28 cases (57%) gene expression decreased. Statistical analysis, using Fischer's exact test, showed that with the p-value of 0.02 there was a direct association of gene expression with its copy number, i.e., duplication increased expression while deletion decreased it. *We, therefore, demonstrated that with RNAseq we can observe an effect of CNVs on gene expression.*

Somatic mobile element insertion in brain. In recent study [25] the authors captured DNA of mobile elements and flanking regions. The DNA was sequenced by ILLUMINA paired-end sequencing and analyzed with the aim of finding somatic mobile elements insertions. While this study provides direct evidence for somatic mobile element insertions the question about its extent is yet to be answered.

***Analyses of Retrotransposon insertions*:** Whole-genome sequencing can detect all types of transposable elements (TEs) (L1, Alu, SVA, HERV etc.) and at no added cost since we can re-analyze the already existing whole-genome data for this purpose [41] (**Figure 3,** from [41]). TE calling is integrated with the MOSAIK mapping software, but instead of discarding non-unique sequences, they are mapped to the reference genome in relation to the database of TE sequences (as well as locations of known TE polymorphisms). We have participated in the development and testing of methodology to detect TE insertion into the genomic sequence in whole-genome sequence data [41]. The algorithm detected, in 1000 Genomes low-coverage (2-3X) whole-genome PE sequencing data for 150 individuals and in high-coverage (15-40X) data for 2 parent-child trios (DNA from cell lines), a total of 7,380 TE insertion sites (85% ALU, 12% L1, 2.5% SVA), of which 69% were novel, with a ~95% validation rate by PCR, suggesting that computational detection using whole-genome PE sequencing is the current method of choice. TE sites resembled SNPs in evidence for selection (similar distribution of frequencies, including 50x fewer events in coding regions than expected by chance) and in the clustering of allelic frequencies by continental ancestry.

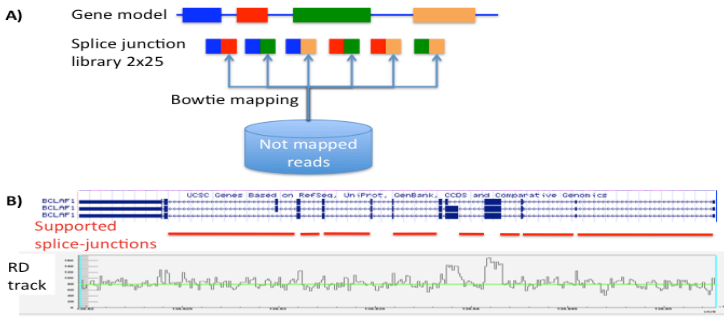

A)
Gene model
Splice junction library 2x25
Bowtie mapping
Not mapped reads
B)
BCLAF1
BCLAF1
BCLAF1
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
Supported splice-junctions
RD track

Figure 4. A) Conceptual pipeline for discovering novel retrogenes. B) Candidate of parent gene, BCLAF1, with novel orphan retrogene. Eight splice-junctions are confirmed by uniquely mapped reads. Novel retrogene is corroborated by increased read depth in exons.

Discovering processed pseudogenes (retrogenes) from sequencing data. We developed a computational pipeline for discovering from sequencing data novel processed pseudogenes, i.e., those that are absent from the reference genome, (referred to below as retrogenes), and have applied it to analyze data from the 1000 Genomes Project (manuscript is under review). *Retrogenes are created by retrotransposition, so, if extensive somatic retrotransposition in brain does exist, we expect to find novel retrogenes.* Thus far, variation in retrogenes has not received any attention. Our pipeline is the first and unique attempt for systematic analysis of novel retrogenes. In brief, we constructed a splice-junction library by joining sequences flanking introns (**Fig. 4**). Namely, for each gene, we exhaustively constructed sequences consisting of 25 bases at the 5'-end of each intron and of 25 bases at the 3'-end of the same and all downstream introns (the number of bases at the junction is increased with read length). We then mapped gDNA reads that previously could not be mapped to the reference genome to this splice-junction library. Reads that mapped uniquely to a splice-junction are retained for the analysis as they are indicative of the junction being present in the studied genome as a continuous (without intron) sequence, suggesting that a corresponding gene, i.e., a parent gene, has an orphan retrogene that is not in the reference genome. We have created additional elaborate analysis procedures (i.e, comparison with null model, detecting insertion points, employing read depth information) to select confident discoveries. An example of a discovered novel retrogene is shown in **Fig. 4B**. When using data from the 1000 Genomes Project we have discovered hundreds of novel retrogenes in almost a thousand people from 14 populations. We also found that parent genes for those retrogenes have higher expression than other genes specifically when cells transition from M to G1 phase of the cell cycle (**Fig. 5**). We, thus, hypothesize that their high expression and nuclear membrane disruption during division are conducive to creation of retrogenes from their mRNA. This hypothesis directly implicates cell division in retrotransposition, which has been suggested by previous studies [32, 33].

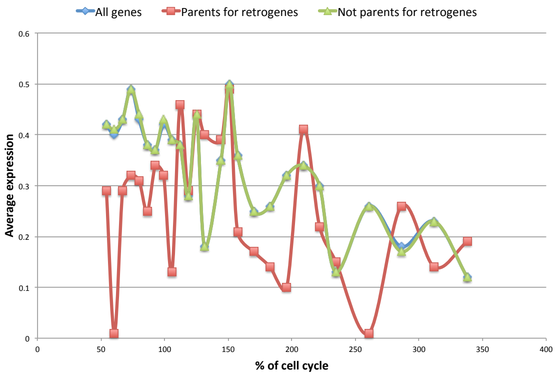


Figure 5. Average gene expression during cell cycle. Data from [1]. Different curves show expression during cell cycles for all genes with data, for genes with variable pseudognees, and for the remaining genes. 100% of cell cycle marks transition from M to G1 phase. Genes with novel pseudogenes have higher expression around that point.

<u>Advantages of clonal sequencing over deep tissue sequencing.</u> Sequencing of clonal populations has a number of advantages over deep sequencing of original tissue. First, it can unmask somatic variants at much smaller frequency. Even at 1000x coverage of original tissue discovering of variants at <0.1% of allele frequency is impossible, while confident discovery is likely to be archived for variants with allele frequency >1%. The same coverage can be split over 30 clones to yield average >30x coverage, which is enough to discover all variant within a clone, even those that are at extremely low frequency in original tissue. Second, clonal sequencing provides information about which variants may co-exist within a single cell. One can observe co-occurrence of CNVs and/or SNVs and/or mobile element insertions, as well as high and low allele frequency variants. Finally, clonal sequencing can also give important information about the timeline of variant generation. For example, in our analysis of hiPSC line we observed a somatic CNV in two lines. For one line it was the only somatic CNV, while for the other one we detected two more CNVs. The most likely explanation is that those two were created after the one shared by both lines.

Approach

Aim 1. Construct a map of somatic variations present in progenitor cells of the cerebral cortex and basal ganglia and estimate their frequency in the original cell population as well as in other embryonic tissues.

Aim 1a. Generate clonal cell populations from the cerebral cortex and basal ganglia of postmortem human embryos at 14-15 weeks of gestation and obtain the complete genomic DNA sequence of each clone. Using computational analyses discover genomic variations manifested in each clone as compared to the whole, original cell population validate this information by PCR, qPCR, digitalPCR and determine the exact breakpoints of SVs.

General design. The following steps are required:

1. Sample preparation
2. Clonal expansion and sequencing

3.1. Bioinformatics predictions of low-allele frequency (i.e., clone-manifested) variants by comparing whole genome sequencing data from clonally expanded lines and original brain tissues (**Fig. 6**)

3.2. Bioinformatics predictions of high allele-frequency variants by comparing data from different tissues and brain regions

4. Validating predictions

<u>1. Sample collection/preparation.</u> The research outlined in this grant application will be conducted using the tissue procurement pipeline at Nenad Sestan’s laboratory in the Department of Neurobiology at Yale University School of Medicine. In the first two years of this project, tissue will be collected after parental or next of kin consent and with approval by the institutional review boards at the Yale University School of Medicine and of any other institution from which tissue specimens may be obtained. Appropriate written informed consent will be obtained and only non-identifying information will recorded for each specimen. The handling of tissue is performed in accordance with ethical guidelines and regulations for the research use of human brain tissue set forth by the NIH (http://bioethics.od.nih.gov/humantissue.html) and the WMA Declaration of Helsinki (www.wma.net/en/30publications/10policies/b3/index.html). To ensure the highest standard for data protection, no personal identifying information will be collected nor will it be accessible by any of the investigators on this project. In addition, we will doubly code in publications all information received to comply with the latest HIPAA regulations. We will use this non-identifying medical history of the subject from which the brain tissue will be obtained, or the mother's medical history in the case of pre- and neonatal specimens, for quality control measures. We will review available ante mortem information, including gender, ethnicity, weight, cause of death, medications, and relevant medical conditions. This information will be only used to exclude some postmortem specimens from the study, like those with known history of drug or alcohol abuse, neurological or psychiatric disorders. Also, available information showing specific agonal conditions, including coma, hypoxia, pyrexia, seizures, severe dehydration, hypoglycemia, multiple organ failure, head injury, and ingestion of neurotoxic substances by mother at time of death will also be grounds for exclusion of the postmortem tissue.

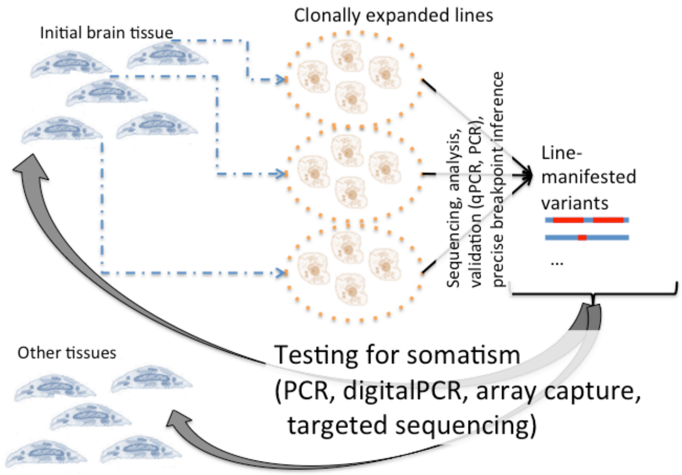

Initial brain tissue
Clonally expanded lines
Sequencing, analysis, validation (qPCR, PCR), precise breakpoint inference
Line-manifested variants
...
Other tissues
Testing for somatism (PCR, digitalPCR, array capture, targeted sequencing)

Fig. 6. Conceptual design for analysis of somatic variants. Clonally expanded neurospheres are produced from single cells of initial brain tissues. Comparison of sequencing data from the clones and initial tissue allows for discovery of line manifested CNVs. Then these CNVs are tested for their presence in the initial tissue at low allele frequency.

2. Clonal expansion and sequencing. The ventricular and subventricular zone (VZ/SVZ) of frontal and occipital poles of the cerebral cortical wall and of the ganglionic eminences of the ventral telencephalon will be microdissected from clinically unremarkable mid-fetal brain specimens, i.e., from 13 to 15 post-conceptional weeks (PCW). A total of five fetuses will be collected in the first two years of this grant. The following fetal tissues will be freshly harvested: **VZ/SVZ from the dorsal and the ventral telencephalon**, the **dorsal cortical plate**, and a sample of **blood**.

Neurospheres, which are free-floating clusters of progenies of a single neural stem/progenitor cells, will be generated by suspending the cells dissociated from the **dorsal and the ventral telencephalon VZ/SVZ** using a modification [42] of the original protocol developed by Reynolds and Weiss [43]. Dissociated single primary cells will be cultured at low (clonal) density in a medium lacking adherent substrates and containing necessary growth factors such as epidermal growth factor and fibroblast growth factor. A minimum of **twenty primary neurospheres** will be dissociated and separately expanded by dissociation and growth as secondary and tertiary neurosphere preps. **Genomic DNA** will be extracted from each **clonally expanded neurosphere** prep and from the **original brain tissue** used to generate the neurospheres, as well as from a sample of **blood** from the same fetus. DNA will then be subjected to whole genome, high throughput PE DNA sequencing at 30X coverage to identify all types of genomic variation (CNV/SVs, SNVs, retrotransposon/pseudogene insertion) by comprehensive computational analyses as described below (**Fig. 6**).

First, we will discover genomic variants in the **20 clonally expanded neurospehere preparations** and the **original dorsal and the ventral telencephalon VZ/SVZ cell population** with respect to the reference genome. Then we will genotype each clonal population with respect to the VZ/SVZ cell population of origin to find structural DNA variants manifested only in each clone. Currently, we can prepare genomic DNA libraries suitable for sequencing on the Illumina HiSeq instrument using as little as 50 ng total DNA, which should be easily obtainable by a clonal population of 50,000 cells. The reason for comparing the degree of mosaicism of the dorsal and ventral telencephalic **VZ/SVZ** is that they share the same general location (the telencephalic vesicle) but the ventral VZ/SVZ represents a region with a lower degree of clonal expansion than the dorsal cortex.

3.1 Bioinformatics predictions of low-allele frequency somatic variants. We have extensive experience when working with next-generation sequencing data [40, 44-51]. For our analysis we will adopt a pipeline used by the 1000 Genomes Project for aligning data to the reference genome [46, 48]. In brief, sequencing reads will be aligned to the latest version of the reference genome, filtered to remove duplicates, realigned around known indel locations, followed by recalibration of base quality scores. Then, we'll use a variety of methods based on different approaches for comprehensive discovery of variants in each sample. Namely, for SNV discovery we will use GATK [52] and output of standard ILLUMINA analysis pipeline, for indel discovery we will use GATK [52], PINDEL [53] and DINDEL [54], for mobile element discovery we will use SPANNER [41, 46] and T-Lex [55], for CNV/SV discovery we will use CNVnator [40, 46], PEMer [45], Pindel [53], SRiC [47] and AGE [44], and for pseudogene discovery we will use our existing pipeline (manuscript is under review). We have previously developed or participated in development of a number of those software: CNVnator [40, 46], SPANNER [46], PEMer [45], SRiC [47] and AGE [44]. With almost all others we had previous experience.

For a clonal cell population, the list of **clone-manifested variants** will be a subset of the variants calls, which don't overlap and have no evidence of existence in the original brain sample. For SNVs, that would mean than no reads in sequencing data from the original sample supporting a given SNV are found. For CNVs, that would mean that no discordantly mapped RP and/or no split-read and/or no RD deviation from normal level suggests a CNV in the original brain sample. Similarly, for a tissue its specific variants will be those having no evidence of existence in the other tissue. The variant can be clone-manifested because of two reasons: (i) the variant newly developed during clonal expansion and culturing; (i) the variant was present at low allelic frequency in the original tissue and thus not detectable.

<u>3.2 Bioinformatics predictions of high-allele frequency somatic variants.</u> High-allele frequency somatic variants can be found by comparing sequencing data from two tissue samples. We will compare **dorsal VZ/SVZ, ventral VZ/SVZ and the blood**, a tissue with different embryonic lineage. All the methodologies, comparisons, approaches, experiments, and validations described for step 3.1 above and step 4 below can be applied for such discovery. Namely, the list of sample-manifested variants will be a subset of the variants calls, which don't overlap and have no evidence of existence in other tissues.

Table 1. Validation strategies of predicted clone-manifested and tissues specific variants.

Variant type	Validation experiment	Confirmatory evidence
SNV	PCR amplification of locus & capillary sequencing of band	Base mismatch in the sequence
Indel	PCR amplification of locus & capillary sequencing of band	Indel in the sequence
CNV/SV	PCR amplification across breakpoints & capillary sequencing of band	Split sequence alignment corresponding to the variant (e.g., deletion)
	qPCR in the middle of a CNV	Deviation from 1 of ratio for copy number estimates in two compared samples
Retrotransposon insertion	PCR amplification across breakpoints & capillary sequencing of band	Fraction of the sequence maps to the target location and the rest is repeat
Processed pseudogene insertion	PCR amplification across exon-exon junction(s) & capillary sequencing	Split sequence alignment with gaps at introns
	PCR amplification across breakpoints & capillary sequencing	Fraction of the sequence maps to the target location and the rest maps to parent gene

<u>4. Validation and refinement of predicted variants</u>. Candidate clone-manifested CNVs/SVs, SNVs, indels and retrotransposon/pseudogene insertions discovered by sequencing will be validated by an orthogonal method, PCR and/or qPCR, in the original DNA from the neurosphere clones and the original cell population from the cortical VZ/SVZ (**Fig. 6**) to yield a list of genomic regions that are truly different between clonal neurospheres and brain sample. The purpose of validation is to arrive at a confident set of clone-manifested variants. For SNVs and indels (the expected majority of variants) we will validate 100 random events per sample. Validation results can be used to optimize parameters/cut offs for calling clone-manifested SNVs and indels. For SVs, MEI and pseudogenes (with prior expectation of just a few events per clone) we will validate each event. Theoretically, PCR and subsequent gel band sequencing can be applied for validation of any type of variants: SNVs, indels, CNVs/SVs, mobile element and pseudogene insertions (**Table 1**). However, the success of PCR for validation of CNVs/SVs can be compromised by off-target primer placements due to uncertainty in breakpoint location. We will design multiple pair of primers for large variants to address this issue. Additionally, we will validate CNVs with qPCR. The results will also provide us with an estimate of how often CNVs/SVs could not be validated with regular PCR due to primer mis-placement. In our analysis of somatic CNVs in fibroblast using hiPS lines we were able to design workable PCR primers for 68% of CNVs validated with qPCR. We anticipate that in the proposed research this fraction will be higher as we will have deeper sequencing of each sample, which allows determining CNV boundaries more precisely. We, therefore, anticipate that for most real clone-manifested variants we will be able to conduct a successful PCR. For each successful PCR we will also sequence the resulting band with Sanger sequencing. This will be an ultimate proof of a variant existence and, for large CNVs, will allow us to determine their <u>precise breakpoints</u> [44, 47].

***Expected results*:** *The results of completing this sub-aim will be a set of tissues for 5 embryos, which can be used for analysis of somatic variants in the whole body (not just brain). Another result is the set confident/validated clone-manifested and tissue specific variants in brain with a majority of them resolved at a single basepair level.*

Aim 1b. Using the dataset of validated genomic variants manifested in progenitor clones, utilize target ultra deep sequencing approach, PCR and digitalPCR to determine the presence and allele frequency for specific somatic variations in the original brain cell population and in at least another embryonic tissue lineage such as the blood.

***General design*.** It is crucial to determine the somatic nature and allele frequency of all discovered SVs. Somatic variation in cultured cell populations may not necessarily be present in endogenous cells, as it may be merely a product of *in vitro* expansion and the associated mitoses. To discriminate between genomic variation generated *in vitro* and variation due to mosaicism of the cell population from the VZ/SVZ within the brain, we will **genotype the genomic variants in the original tissue** used to prepare the cell suspension for clonal analyses **using the most sensitive techniques available**.

In addition to testing the presence of SVs in the original sample of cortical and basal ganglia VZ/VZ used to prepare the cell suspension, we will assess their presence in the superficial part of the cortex (cortical plate), in which early-generated neurons, arising from neural stem cells in the cortical and basal ganglia VZ/SVZ, respectively, migrate to form the cortical laminar structure. A sample of microdissected cortical plate of frontal and occipital cortices and basal ganglia will be retained at the time of the original embryo dissection. We will confirm that the genomic mosaicism found in cells generated *in vitro* is also found in postmitotic cortical and striatal neurons and thus is derived by *in vivo* mitoses. Furthermore, to determine whether the same genomic events are present in other tissues, we will assess whether validated genomic variants found in the brain are also present in the blood obtained from the same fetus. **In total, we will assess the presence of clone-manifested variants in 4 tissues: VZ/SVZ of the cerebral cortex, VZ/SVZ of the basal ganglia, cortical plate and blood**.

***Establishing the somatic nature of clone-manifested variants*.** For validated clone-manifested variants we will determine their presence and frequency in the brain samples with **targeted ultra deep sequencing, PCR** and **digitalPCR** approaches (**Fig. 3**). We intend to apply all three techniques, as none of them along is universal. Targeted sequencing can be applied for analysis of all variants but it can miss very low allele frequency variants. For example, for variants at 0.1% allele frequency at least 1000x coverage of targeted regions is required. Also, estimate of variant frequency can be biased due to variable capture efficiency. Control experiments would allow correcting for the bias but it will also increase the cost. PCR and digitalPCR are not suited for assessing low frequency SNVs and indels as the non-variant allele would also be amplified (except in rare cases when the primer would include the variant). However, PCR and digitalPCR could be more sensitive than targeted sequencing to assess low frequency CNVs/SVs, processed pseudogenes and retrotransposons i.e., when amplifying across breakpoints (**Fig. 1E**). Our primary choice between PCR and digitalPCR will be the latter (as it allows not only observing but also estimating the frequency for low frequency variants). However, there is a subtle difference in primer design for successful application of each technique. DigitalPCR is best performed for short amplicons (<100bp) with high affinity hybridization probes, while PCR results are best observed for larger amplicons (>100 bp) and require no hybridization probe. Therefore, depending on nucleotide content, PCR can be more sensitive for detection of low allele frequency variants.

For the targeted ultra deep sequencing approach, we will first design a capture array with probes targeting the regions around validated clone-manifested variants and then sequence captured DNA material with Illumina MiSeq 2x250 bp reads to extremely high (~1000x) coverage. For this analysis we will prepare short insert libraries of ~400 bp, so that 3'-ends of 250 bps paired reads overlap allowing constructing a single longer read of ~400bp in length. Alignment of these reads will be used to show the presence of variants in the original brain sample, in the same way we intended to use the result of capillary sequencing after PCR (**Table 1**). Clone-manifested CNVs/SVs and pseudogene/retrotransposon insertions showing no evidence of being somatic in brain samples after targeted sequencing will be tested with digitalPCR and PCR. We will also apply these techniques to a subset of SNVs and indels by use restriction enzyme cutting at reference allele site, thus after amplification we should be able to see two short bands for cut reference allele and a long band for non-reference, i.e., somatic, allele. Results of digitalPCR and PCR analysis will be also used to estimate sensitivity and precision of array capture experiment in predicting allele frequency of somatic variants.

For the targeted ultra deep sequencing approach, we have performed array capture experiment for CNV validation in the pilot phase of the 1000 Genomes Project [46], where we targeted ~14 Mbp of genome with 2.1M oligo Nimblegen custom array. For the proposed project, most of the tested variant will be known with basepair resolution (except of some CNVs/SVs for which breakpoints can be off by up to 1 Kbp), so we estimate that on average we will need to target with probes ~500 bp around each variant. Previously, only a few [19] clone-manifested SNVs in coding genomic regions were observed. Scaling up to genome size gives <100 and account for possible negative selection in coding regions yields **<200 somatic SNVs per cell**. All other types of variants are typically significantly less frequent. For example, in our previous analysis we saw only a few line-manifested CNVs per hiPSC lines [21]. So, we conservatively estimate that there will be <200 clone-manifested variants in each clonal population. Forty clones will be sequenced for each individual (20 from cerebral cortex and 20 from basal ganglia) which projects to 6 Mbp of sequence to be captured in total, Thus, through this estimate, we predict that a single array would be enough to capture sequence around all clone-manifested variants from clonal neurospheres preparations from each brain. Thus, only one capture array design per individual is needed, where it will be applied to four tissues. From our previous experience only half of sequencing reads are from targeted regions. Taking this into account one can estimate that targeting 6 Mbp on array will enrich targeted regions by 250 folds. The next generation of MiSeq sequenator (http://www.illumina.com/Documents/products/datasheets/datasheet_miseq.pdf) will produce ~25 million 2x250 bps reads per run, which would allow to archive ~1000x coverage of targeted regions. The machines will be available for commercial use at the end of 2012 while upgrade at the Yale Center for Genomic Analysis would follow thereafter.

Our estimation of the number of somatic variant per clone is based on observed mosaics in adult skin. It can well be (and we think this is the more likely scenario) that there are significantly fewer somatic variants in embryos, as its cells did not divide that many times and were not exposed to the environment. So, in case the number of putative somatic variants is on the order of 10 or less, we will attempt digitalPCR and PCR for all of them in four tissues for each individual.

Potential challenges and solutions. If the number of candidate clone-manifested variants will turn out to be still too large to conduct a PCR validation for each one of them, then we will validate variants of interest, i.e., those overlapping known genomics functional/annotated elements, overlapping differentially expressed genes in clones vs. tissues, and all SVs. We will also apply a "population filter" to select more confident variants. The idea of the filter is to look at occurrence of putative clone-manifested variants in other clones. If a variant is present in many other clones then it is likely a false positive, i.e., a variant present in the original tissue but not discovered there for some reason. Also, we will select a random and manageable subset of variants for validation to estimate the FDR of our analysis. Using validation results we will optimize the computational analysis pipeline to yield more stringent and more reliable results. Capture array and targeted ultra deep sequencing approach will be our first line of testing somatism, and we will employ digital PCR as a second choice, for those SVs that are not validated by the first approach (estimated to be about 10-15%).

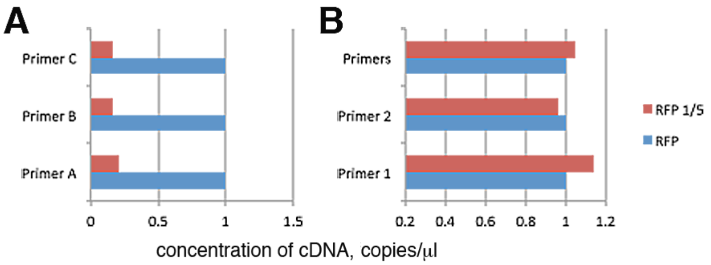


Fig. 7. Linear distribution of gene expression for diluted sample with ddPCR machine QX100 (a) and qPCR from CFX 384 (b). Raw data without normalization by the internal control, showing the absolute amount of cDNAs in each sample. Note that blue bar indicated the amount of undiluted cDNA and red bar indicated the amount of 1/5 diluted cDNA following PCR amplification.

Commercial use of next generation MiSeq sequenator can be delayed. In such case we will use the current generation that we already have access to. It can generate ~14 million 2x150 reads per run, which allow constructing reads (from overlapping 3'-ends) of up to 250 bps. With current generation we would have to do two runs, which would slightly increase the cost of establishing the somatic nature of variants. It should be also noticed that we are not going to use MiSeq until the second year of the project. Thus, there is a significant lag of time from the beginning of the project until the time we need MiSeq, which increases the likelihood of next generation MiSeq being available for use. A possible alternative is using IonTorrent machines and Amplicon Sequencing in particular (i.e., sequencing selectively amplified regions: http://www.iontorrent.com/lib/images/PDFs/amplicon_application_note_040411.pdf).

PCR. Specific primers will be designed to reveal bands that are diagnostic of the events to be genotyped, i.e., to target both sides of adjacent region of the deleted or to target the 5' and 3' end of the duplicated region. In this way specific products will be amplified only when deletions or duplications are present. Genomic DNA from the HapMap cell line GM12878 will be used as negative control. When necessary, a second round of PCR with 30 cycles was performed to further increase the signals for low allele frequency events. All specifically amplified PCR bands will be run on 2 % E-gel (Invitrogen, CA), gel extracted by MinElute gel purification Kit (Qiagen, MD), and sequenced using both forward and reverse primers.

Digital PCR. Digital PCR is conceptually similar to regular PCR. In essence it is parallel PCR in many (tens of thousands) single cells, the number of cells with the observed amplification giving a sense about how frequent the particular locus is. In addition to a pair of amplification primers it requires a primer for a fluorophore, which can be designed with standard tools like Primer3. DNA from tissue will be the signal. DNA from clones where a variant was discovered will be used for normalization. Genomic DNA from the HapMap cell line GM12878 will be used as negative control. We will use the BioradQX100 Droplet Digital PCR (ddPCR) system, which provides a near absolute measure of target DNA molecules for quantitative PCR amplifications. The droplet generator partitions samples into 20,000 nanoliter-sized droplets. After PCR on a thermal cycler, droplets from every sample are streamed in single file on the QX droplet reader. The PCR-positive and PCR-negative droplets are counted to provide absolute quantification of target DNA in digital form (**Fig. 7**).

***Expected results**: The results of completing this sub-aim will be the estimation, for a given set of somatic genomic variants, the number and allele frequency of these variants in the original tissue. Furthermore we will estimate whether this frequency is higher in dorsal as compared to ventral germinal layers. Additionally we will estimate of how many variants discovered and validated in the brain are also present in the blood, which will allow to indirectly estimate when they arose in ontogenesis.*

Aim 2. Determine the impact of somatic variants on mRNA transcripts. Due to complex hierarchy of transcriptional regulation it is challenging to decipher the effect of each individual variant on the expression of a particular gene or set of genes. We, therefore, in a first instance, will aim at determining the overall effect of the variants on gene expression. The simplest cases for such analysis are CNVs. A prior expectation is that expression for genes that intersect duplications will increase while expression for genes that intersect deletions will decrease. In relation to analysis of hiPSC, we showed that indeed there is such statistically significant correlation (preliminary data and **Fig. 2**). We will perform a similar analysis for somatic CNVs discovered in the proposed study.

Determining gene expression from RNAseq data. Our existing RNAseq workflow is described briefly as follows. Tophat [56] is used to align the data against the human genome (hGRC37/hg19), the Gencode v11 annotation [57, 58], and dynamically constructed exons and splice libraries. The output in BAM format [59] is processed by BedTools [60] to compute the raw read counts for each gene. We then quantify expression by Reads Per Kilobase of transcript per million mapped fragments (RPKM) using RSEQtools[61]. The raw counts are used for determining differentially expressed genes by DESeq[62], an R package that uses negative binomial distribution to model read counts per gene and the Benjamini-Hochberg method for adjusting for multiple comparisons [63].

Overall, our hypothesis is that somatic variants lead to differential expression of genes in their vicinity. More specifically, we will compare expression of genes in **clonal** and in the **original cell populations**. To test for this, we will first assign potentially influential variants to each gene. Here we will simply extend annotated genes coordinates by 2 kbp downstream and upstream, and every variant within such regions will be "potentially influential" for the gene. We will then perform an enrichment analysis of whether genes with "potentially influential" variants are more often differentially expressed then genes without such variants with respect to the original population. We also expect that genes with frame shifting indels/MEIs and SNVs leading to stop codons (disrupting splice-site or introducing stop codon) will have decreased expression due to nonsense-mediated decay (NMD). We will, thus, test for an enrichment of down regulated genes among those having such SNVs and indels/MEIs.

We will also attempt to predict an effect of "potentially influential" variants that are likely to reside in the promoter region of an assigned gene, i.e., upstream from the transcription start site. The analysis will be focused on the identification of transcription factor (TF) binding motifs that are created/lost/changed as a result of introducing a variant. In particular, we will look at whether a variants (e.g. SNVs) leads to a predicted increase/decrease in TF binding affinity from Position Weighted Matrix (PWM) [64]. We will focus on analysis of motifs for TF uniformly active across heterogeneous tissues in the human body (such that Pol2, SMAD1, ETS1, JUN, CREB, ATF3) and TFs that are enriched in the brain (such as FOXG1, FEZF2, LHX2, NEUROG2). Given the predicted effect of variants on the gene expression, we will correlate it with the measured expression the way we do for CNVs, i.e., perform an enrichment analysis on many genes.

Observing such enrichments/correlations will potentially allow us to suggest how and through which mechanism (e.g., gene duplication/deletion, TF binding site creation/deletion, etc.) somatic variants affect mRNA expression. We are confident in our abilities to perform such an analysis as we did it previously in various studies where we: correlated gene expression with CNV (ref [21] and preliminary data), studied allele specific expression and binding within ENCODE project [51, 65], surveyed loss-of-function variants within 1000 Genomes Project [48, 66], and analyzed the effect of variants on coding and non-coding genomic elements [67, 68]. Some technical details of our analysis are given below.

Impact of somatic variants on functional gene co-expression modules. Cells have complex regulatory networks, and the effect of genomic variants on gene expression is not always straightforward. In particular, a variant can introduce only a minor change to gene expression while this change can be amplified in a cascade of downstream regulatory events. To partially overcome these problems we will adopt an additional strategy, i.e., we will evaluate the effects of somatic genomic variants on gene co-expression networks, a powerful tool for network inference and data reduction [69]. Specifically, in each individual, we will use the WGCNA package [69] to process gene expression data from all the brain regions, in clonally extended cell and the original cell population. WGCNA relies on the correlation strength between co-expressed genes to reduce the whole set of gene transcripts to a much smaller set of "modules". These "modules" are defined by sets of genes whose expression profiles are correlated across samples. The overall expression level of each module will be described by its **first principal component (the module eigengene**). Modules' eigengene will be tested for differential expression between clonally extended cell populations against the original cell population, to identify the **module(s) whose expression profile is likely affected by the genetic variants**.
The genes in each module will be enriched for gene ontology (GO) classes and biological pathways to assess the biological significance of the module(s). We will use Ingenuity (http://www.ingenuity.com), DAVID (http://david.abcc.ncifcrf.gov) and MetaCore™ (https://portal.genego.com) software and databases for this purpose. These analyses will allow us to derive the most significant biological pathways for list of genes that participate in a network module.

We will then select those modules that are differentially expressed between clonally extended cell populations and the original brain tissues within each brain, and further determine whether transcripts belonging to each module are **significantly enriched with genes harboring potentially influential genomic variants**, similarly to what we will do for genes (see above). Collectively, these analyses will allow us to derive the most significant gene networks and biological pathways whose differential expression correlate with specific somatic variants in each individual. These **genomic variants that are significantly associated with transcript modules will then be selected for further analyses**. The modules derived from the various individuals may underline completely independent or, more likely, overlapping functions. We will therefore test for consensus modules across individuals to identify possible common biological functions affected by the individual's SVs.

In addition to minimizing the issue of multiple comparisons, modular analysis will likely increase our ability to determine the biological significance of the detected genomic somatic variants. The prediction from this analysis is that we will identify sets of co-expressed transcripts that are significantly associated with genomic variants in clonal cell populations where these genomic variant(s) were discovered. Although enrichment does not necessarily mean causation, these analyses will allow us to **focus on those somatic variants that may be associated, or even be causative for expression change in modules of functionally related genes**.

We will further compare differentially expressed modules enriched in somatic variants with modules of co-expressed genes in developing brain using larger datasets available from the Sestan lab. The Sestan laboratory has generated and analyzed RNA-Seq and DNA methylation data for the BrainSpan project (www.BrainSpan.org) [70, 71]. Since the modules of co-expressed genes from these larger datasets are likely to contain a wider representation of transcripts expressed in developing brain across a larger number of individuals, we will further determine which modules in the BrainSpan project datasets are enriched with genes containing somatic variants, as determined by the prior step in our analyses. Our expectation is that these analyses will allow us to draw correlations across individuals, i.e., investigate whether somatic variation has general implications for processes of brain development. In the future, we will validate the modules/pathways arising from the present analysis, by first identifying key regulators in each relevant module/pathway, for instance by classifying genes by their degree of connectivity ("hub" genes), then by selectively perturbing/blocking their activity (e.g. RNAi, etc.) and assessing the impact of such operation on the cellular phenotype.

Statistical analysis. For each gene/module and its "potentially influential" somatic variant(s), we will compare expression of the gene/module in the pair clone – original population, where a variant was discovered. Given a set of variants (e.g., CNVs, mRNA disrupting SNVs/indels, etc.) each such comparison will be a data point. These data points will be used for statistical analysis. For enrichment analysis (e.g. for differential expression of genes with "potentially influential" variants) we will apply test for comparison of two proportions. We will first assign potentially influential variants to each gene as described above and then we will then perform an enrichment analysis of whether genes with "potentially influential" variants are more often differentially expressed then genes without such variants within the differentially expressed module(s). We will also test for an enrichment of down regulated modules among those having frame shifting indels/MEIs and SNVs leading to stop codons. When there are two possible outcomes of our predictions (e.g., increase in expression of duplicated genes and decrease in expression in deleted genes) we will use Fischer's exact test to access statistical significance. The resulting p-values will be correct for multiple-hypothesis testing (i.e., number of correlations analyzed).

Expected results: *Collectively, these analyses will provide another layer of analysis to try to estimate the potential biological significance of the variant, based upon the most significant gene networks and biological pathways that correlate with it within and across individuals. This will greatly increase the predictive power of our analysis and assess the biological significance of the genomic variants.*

Aim 3. Investigate the most likely biological origin of somatic variants

Here we will focus on analysis of CNVs and SVs. Obviously insertion of pseudogenes and mobile elements is the result of retrotransposition. Other SVs, however, can be generated by multiple mechanisms [5, 30, 31]. We have previously developed an algorithm for optimal alignment of sequences with SVs [44]. The strength of the method is that it guarantees an optimal alignment of sequencing around SV breakpoints without prior assumption about sequence features around breakpoints. In particular, the algorithm naturally detects micro-insertions around SV breakpoints (**Fig. 1F**) and can resolve breakpoints masked by long homologous sequences [44]. When performing validation experiments (**Fig. 6**) with PCR and/or targeted resequencing with long reads we will align reads with our algorithm to derive SV breakpoints. We will conduct similar analysis when establishing somatic nature of CNVs. Matching breakpoints from the two analyses will provide additional confirmatory evidence for our findings.

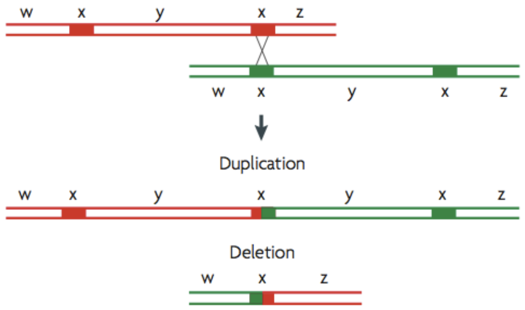


Figure 8. Creation of deletion and tandem duplication alleles as a result of chromosome cross-over. The figure is adopted from [5].

Having found precise CNV/SV breakpoints we will use our existing pipeline to classify mechanisms generating the CNVs [49]. In brief, the pipeline looks at sequence features around SV breakpoints to predict the most likely mechanism generating the SV: Non-allelic Homologous Recombination (NAHR), mobile element insertion (MEI), Variable Tandem Repeats (VTR) and Non-Homologous Rearrangements (NHR). Additionally, just looking at the frequency of each type of CNV (deletion vs. duplication) gives valuable information about cellular processes underling CNV/SV creation. Tandem duplications strongly imply chromosomal cross-over, i.e., homologous/non-homologous recombination, as a CNV creating event. Dispersed duplications were suggested to arise during replication errors or complex recombination event [5, 30, 31]. Deletions could be the result of double stranded DNA break or chromosomal cross-over (**Fig. 8**). In the latter case a tandem duplication is also created. Note, that both alleles with duplication and deletion can be present in original brain cell population but only one of them will be present in each given clone. We will perform an experiment to find out whether at least some of detected duplication can be the result of cross-over in brain cells. Namely, for each deletion (inversely tandem duplication) we will assume that there exists a tandem duplication (inversely deletion) in the brain sample. We will then add these "hypothetical" variants to our test list for establishing somatic nature (see **Aim 1**), and analyze them with described experiments (PCR, digitalPCR, and ultra-deep targeted resequencing).

It is more challenging to determine the mechanisms creating somatic SNVs and indels, as they (mechanisms) have no distinct signatures in sequence. We therefore will analyze whether statistical characteristics of somatic SNVs/indels are different from the characteristics of germline SNPs/indels. In particular, we will compare their distribution across genome, distribution within and outside coding regions, aggregation around recombination hot-spots, and transition to transversion ratios. It may well be that somatic SNVs/indels closely resemble germline SNVs/indels, then this will be the end point of our analysis. In case they are different will further characterize somatic SNVs/indels with the aim of better understanding their nature. In particular, we will attempt identifying clusters of somatic SNVs/indels, bioinformatically correlate their (variants') distribution with chromatin structure and histone marks (using existing datasets, e.g., from ENCODE project), and correlate their distribution with distribution of CNVs/SVs and MEIs. Clustering of somatic SNVs and/or indels and/or CNVs and/or MEIs may identify genomic regions susceptible to somatic variants or even imply common cellular process(es) leading to their generation.

Expected results: *The result of analysis in this aim will be an improved understanding of which mechanism are responsible for generating somatic CNVs/SVs and whether somatic SNVs/indels are similar to germline SNPs/indels.*

Literature Cited

1. Whitfield, M.L., Sherlock, G., Saldanha, A.J., Murray, J.I., Ball, C.A., Alexander, K.E., Matese, J.C., Perou, C.M., Hurt, M.M., Brown, P.O. & Botstein, D. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular biology of the cell* **13**, 1977-2000 (2002).
2. Frank, S.A. Evolution in health and medicine Sackler colloquium: Somatic evolutionary genomics: mutations during development cause highly variable genetic mosaicism with risk of cancer and neurodegeneration. *Proceedings of the National Academy of Sciences of the United States of America* **107 Suppl 1**, 1725-1730 (2010).
3. Frank, S.A. & Nowak, M.A. Problems of somatic mutation and cancer. *BioEssays : news and reviews in molecular, cellular and developmental biology* **26**, 291-299 (2004).
4. Frank, S.A. & Nowak, M.A. Cell biology: Developmental predisposition to cancer. *Nature* **422**, 494 (2003).
5. Hastings, P.J., Lupski, J.R., Rosenberg, S.M. & Ira, G. Mechanisms of change in gene copy number. *Nature reviews. Genetics* **10**, 551-564 (2009).
6. Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., Leotta, A., Pai, D., Zhang, R., Lee, Y.H., Hicks, J., Spence, S.J., Lee, A.T., Puura, K., Lehtimaki, T., Ledbetter, D., Gregersen, P.K., Bregman, J., Sutcliffe, J.S., Jobanputra, V., Chung, W., Warburton, D., King, M.C., Skuse, D., Geschwind, D.H., Gilliam, T.C., Ye, K. & Wigler, M. Strong association of de novo copy number mutations with autism. *Science* **316**, 445-449 (2007).
7. Sanders, S.J., Ercan-Sencicek, A.G., Hus, V., Luo, R., Murtha, M.T., Moreno-De-Luca, D., Chu, S.H., Moreau, M.P., Gupta, A.R., Thomson, S.A., Mason, C.E., Bilguvar, K., Celestino-Soper, P.B., Choi, M., Crawford, E.L., Davis, L., Wright, N.R., Dhodapkar, R.M., DiCola, M., DiLullo, N.M., Fernandez, T.V., Fielding-Singh, V., Fishman, D.O., Frahm, S., Garagaloyan, R., Goh, G.S., Kammela, S., Klei, L., Lowe, J.K., Lund, S.C., McGrew, A.D., Meyer, K.A., Moffat, W.J., Murdoch, J.D., O'Roak, B.J., Ober, G.T., Pottenger, R.S., Raubeson, M.J., Song, Y., Wang, Q., Yaspan, B.L., Yu, T.W., Yurkiewicz, I.R., Beaudet, A.L., Cantor, R.M., Curland, M., Grice, D.E., Gunel, M., Lifton, R.P., Mane, S.M., Martin, D.M., Shaw, C.A., Sheldon, M., Tischfield, J.A., Walsh, C.A., Morrow, E.M., Ledbetter, D.H., Fombonne, E., Lord, C., Martin, C.L., Brooks, A.I., Sutcliffe, J.S., Cook, E.H., Jr., Geschwind, D., Roeder, K., Devlin, B. & State, M.W. Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863-885 (2011).
8. Sanders, S.J., Murtha, M.T., Gupta, A.R., Murdoch, J.D., Raubeson, M.J., Willsey, A.J., Ercan-Sencicek, A.G., Dilullo, N.M., Parikshak, N.N., Stein, J.L., Walker, M.F., Ober, G.T., Teran, N.A., Song, Y., El-Fishawy, P., Murtha, R.C., Choi, M., Overton, J.D., Bjornson, R.D., Carriero, N.J., Meyer, K.A., Bilguvar, K., Mane, S.M., Sestan, N., Lifton, R.P., Gunel, M., Roeder, K., Geschwind, D.H., Devlin, B. & State, M.W. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* (2012).
9. Gilman, S.R., Iossifov, I., Levy, D., Ronemus, M., Wigler, M. & Vitkup, D. Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron* **70**, 898-907 (2011).
10. Levy, D., Ronemus, M., Yamrom, B., Lee, Y.H., Leotta, A., Kendall, J., Marks, S., Lakshmi, B., Pai, D., Ye, K., Buja, A., Krieger, A., Yoon, S., Troge, J., Rodgers, L., Iossifov, I. & Wigler, M. Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron* **70**, 886-897 (2011).
11. Failla, G. The aging process and cancerogenesis. *Annals of the New York Academy of Sciences* **71**, 1124-1140 (1958).
12. Vijg, J. Somatic mutations and aging: a re-evaluation. *Mutation research* **447**, 117-135 (2000).
13. Youssoufian, H. & Pyeritz, R.E. Mechanisms and consequences of somatic mosaicism in humans. *Nature reviews. Genetics* **3**, 748-758 (2002).
14. McClellan, J. & King, M.C. Genetic heterogeneity in human disease. *Cell* **141**, 210-217 (2010).
15. Carlson, E.A. & Southin, J.L. Chemically Induced Somatic and Gonadal Mosaicism in Drosophila. I. Sex-Linked Lethals. *Genetics* **48**, 663-675 (1963).
16. Sastry, G.R., Cooper, H.B., Jr. & Brink, R.A. Paramutation and somatic mosaicism in maize. *Genetics* **52**, 407-424 (1965).
17. Cotterman, C.W. Somatic mosaicism for antigen A2. *Acta genetica et statistica medica* **6**, 520-521 (1956).
18. De, S. Somatic mosaicism in healthy human tissues. *Trends in genetics : TIG* **27**, 217-223 (2011).
19. Gore, A., Li, Z., Fung, H.L., Young, J.E., Agarwal, S., Antosiewicz-Bourget, J., Canto, I., Giorgetti, A., Israel, M.A., Kiskinis, E., Lee, J.H., Loh, Y.H., Manos, P.D., Montserrat, N., Panopoulos, A.D., Ruiz, S., Wilbert, M.L., Yu, J., Kirkness, E.F., Izpisua Belmonte, J.C., Rossi, D.J., Thomson, J.A., Eggan, K., Daley, G.Q.,

Goldstein, L.S. & Zhang, K. Somatic coding mutations in human induced pluripotent stem cells. *Nature* **471**, 63-67 (2011).

20. Ji, J., Ng, S.H., Sharma, V., Neculai, D., Hussein, S., Sam, M., Trinh, Q., Church, G.M., McPherson, J.D., Nagy, A. & Batada, N.N. Elevated coding mutation rate during the reprogramming of human somatic cells into induced pluripotent stem cells. *Stem Cells* **30**, 435-440 (2012).
21. Abyzov, A., Gerstein, M. & Vaccarino, F. Origin of genomic CNVs in human induced pluripotent stem cells. **In revision**.
22. Maiti, S., Kumar, K.H., Castellani, C.A., O'Reilly, R. & Singh, S.M. Ontogenetic de novo copy number variations (CNVs) as a source of genetic individuality: studies on two families with MZD twins for schizophrenia. *PloS one* **6**, e17125 (2011).
23. Piotrowski, A., Bruder, C.E., Andersson, R., Diaz de Stahl, T., Menzel, U., Sandgren, J., Poplawski, A., von Tell, D., Crasto, C., Bogdan, A., Bartoszewski, R., Bebok, Z., Krzyzanowski, M., Jankowski, Z., Partridge, E.C., Komorowski, J. & Dumanski, J.P. Somatic mosaicism for copy number variation in differentiated human tissues. *Human mutation* **29**, 1118-1124 (2008).
24. Mkrtchyan, H., Gross, M., Hinreiner, S., Polytiko, A., Manvelyan, M., Mrasek, K., Kosyakova, N., Ewers, E., Nelle, H., Liehr, T., Volleth, M. & Weise, A. Early embryonic chromosome instability results in stable mosaic pattern in human tissues. *PloS one* **5**, e9591 (2010).
25. Baillie, J.K., Barnett, M.W., Upton, K.R., Gerhardt, D.J., Richmond, T.A., De Sapio, F., Brennan, P., Rizzu, P., Smith, S., Fell, M., Talbot, R.T., Gustincich, S., Freeman, T.C., Mattick, J.S., Hume, D.A., Heutink, P., Carninci, P., Jeddeloh, J.A. & Faulkner, G.J. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* (2011).
26. Coufal, N.G., Garcia-Perez, J.L., Peng, G.E., Yeo, G.W., Mu, Y., Lovci, M.T., Morell, M., O'Shea, K.S., Moran, J.V. & Gage, F.H. L1 retrotransposition in human neural progenitor cells. *Nature* **460**, 1127-1131 (2009).
27. Muotri, A.R., Marchetto, M.C., Coufal, N.G. & Gage, F.H. The necessary junk: new functions for transposable elements. *Hum Mol Genet* **16 Spec No. 2**, R159-167 (2007).
28. Muotri, A.R. & Gage, F.H. Generation of neuronal variability and complexity. *Nature* **441**, 1087-1093 (2006).
29. Pray, L.A. DNA replication and causes of mutation. *Nature Education* (2008).
30. Liu, P., Carvalho, C.M., Hastings, P. & Lupski, J.R. Mechanisms for recurrent and complex human genomic rearrangements. *Current opinion in genetics & development* (2012).
31. Sharp, A.J., Cheng, Z. & Eichler, E.E. Structural variation of the human genome. *Annual review of genomics and human genetics* **7**, 407-442 (2006).
32. Shi, X., Seluanov, A. & Gorbunova, V. Cell divisions are required for L1 retrotransposition. *Molecular and cellular biology* **27**, 1264-1270 (2007).
33. Kubo, S., Seleme, M.C., Soifer, H.S., Perez, J.L., Moran, J.V., Kazazian, H.H., Jr. & Kasahara, N. L1 retrotransposition in nondividing and primary human somatic cells. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 8036-8041 (2006).
34. Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., Muthuswamy, L., Krasnitz, A., McCombie, W.R., Hicks, J. & Wigler, M. Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90-94 (2011).
35. Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., Dermitzakis, E., Bonnen, P.E., Altshuler, D.M., Gibbs, R.A., de Bakker, P.I., Deloukas, P., Gabriel, S.B., Gwilliam, R., Hunt, S., Inouye, M., Jia, X., Palotie, A., Parkin, M., Whittaker, P., Yu, F., Chang, K., Hawes, A., Lewis, L.R., Ren, Y., Wheeler, D., Gibbs, R.A., Muzny, D.M., Barnes, C., Darvishi, K., Hurles, M., Korn, J.M., Kristiansson, K., Lee, C., McCarrol, S.A., Nemesh, J., Dermitzakis, E., Keinan, A., Montgomery, S.B., Pollack, S., Price, A.L., Soranzo, N., Bonnen, P.E., Gibbs, R.A., Gonzaga-Jauregui, C., Keinan, A., Price, A.L., Yu, F., Anttila, V., Brodeur, W., Daly, M.J., Leslie, S., McVean, G., Moutsianas, L., Nguyen, H., Schaffner, S.F., Zhang, Q., Ghori, M.J., McGinnis, R., McLaren, W., Pollack, S., Price, A.L., Schaffner, S.F., Takeuchi, F., Grossman, S.R., Shlyakhter, I., Hostetter, E.B., Sabeti, P.C., Adebamowo, C.A., Foster, M.W., Gordon, D.R., Licinio, J., Manca, M.C., Marshall, P.A., Matsuda, I., Ngare, D., Wang, V.O., Reddy, D., Rotimi, C.N., Royal, C.D., Sharp, R.R., Zeng, C., Brooks, L.D. & McEwen, J.E. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-58 (2010).
36. Mills, R.E., Luttig, C.T., Larkins, C.E., Beauchamp, A., Tsui, C., Pittard, W.S. & Devine, S.E. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome research* **16**, 1182-1190 (2006).

37. Mills, R.E., Pittard, W.S., Mullaney, J.M., Farooq, U., Creasy, T.H., Mahurkar, A.A., Kemeza, D.M., Strassler, D.S., Ponting, C.P., Webber, C. & Devine, S.E. Natural genetic variation caused by small insertions and deletions in the human genome. *Genome research* **21**, 830-839 (2011).
38. McCarroll, S.A., Kuruvilla, F.G., Korn, J.M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M.H., de Bakker, P.I., Maller, J.B., Kirby, A., Elliott, A.L., Parkin, M., Hubbell, E., Webster, T., Mei, R., Veitch, J., Collins, P.J., Handsaker, R., Lincoln, S., Nizzari, M., Blume, J., Jones, K.W., Rava, R., Daly, M.J., Gabriel, S.B. & Altshuler, D. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature genetics* **40**, 1166-1174 (2008).
39. Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C.H., Kristiansson, K., Macarthur, D.G., Macdonald, J.R., Onyiah, I., Pang, A.W., Robson, S., Stirrups, K., Valsesia, A., Walter, K., Wei, J., Tyler-Smith, C., Carter, N.P., Lee, C., Scherer, S.W. & Hurles, M.E. Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704-712 (2010).
40. Abyzov, A., Urban, A.E., Snyder, M. & Gerstein, M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome research* (2011).
41. Stewart, C., Kural, D., Stromberg, M.P., Walker, J.A., Konkel, M.K., Stutz, A.M., Urban, A.E., Grubert, F., Lam, H.Y., Lee, W.P., Busby, M., Indap, A.R., Garrison, E., Huff, C., Xing, J., Snyder, M.P., Jorde, L.B., Batzer, M.A., Korbel, J.O. & Marth, G.T. A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS genetics* **7**, e1002236 (2011).
42. Rasin, M.R., Gazula, V.R., Breunig, J.J., Kwan, K.Y., Johnson, M.B., Liu-Chen, S., Li, H.S., Jan, L.Y., Jan, Y.N., Rakic, P. & Sestan, N. Numb and Numbl are required for maintenance of cadherin-based adhesion and polarity of neural progenitors. *Nat Neurosci* **10**, 819-827 (2007).
43. Reynolds, B.A. & Weiss, S. Generation of neurons and astrocytes from isolated cells of the adult mammalian central nervous system. *Science* **255**, 1707-1710 (1992).
44. Abyzov, A. & Gerstein, M. AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. *Bioinformatics* (2011).
45. Korbel, J.O., Abyzov, A., Mu, X.J., Carriero, N., Cayting, P., Zhang, Z., Snyder, M. & Gerstein, M.B. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol* **10**, R23 (2009).
46. Mills, R.E., Walter, K., Stewart, C., Handsaker, R.E., Chen, K., Alkan, C., Abyzov, A., Yoon, S.C., Ye, K., Cheetham, R.K., Chinwalla, A., Conrad, D.F., Fu, Y., Grubert, F., Hajirasouliha, I., Hormozdiari, F., Iakoucheva, L.M., Iqbal, Z., Kang, S., Kidd, J.M., Konkel, M.K., Korn, J., Khurana, E., Kural, D., Lam, H.Y., Leng, J., Li, R., Li, Y., Lin, C.Y., Luo, R., Mu, X.J., Nemesh, J., Peckham, H.E., Rausch, T., Scally, A., Shi, X., Stromberg, M.P., Stutz, A.M., Urban, A.E., Walker, J.A., Wu, J., Zhang, Y., Zhang, Z.D., Batzer, M.A., Ding, L., Marth, G.T., McVean, G., Sebat, J., Snyder, M., Wang, J., Eichler, E.E., Gerstein, M.B., Hurles, M.E., Lee, C., McCarroll, S.A., Korbel, J.O. & Genomes, P. Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59-65 (2011).
47. Zhang, Z.D., Du, J., Lam, H., Abyzov, A., Urban, A.E., Snyder, M. & Gerstein, M. Identification of genomic indels and structural variations using split reads. *BMC genomics* **12**, 375 (2011).
48. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073 (2010).
49. Lam, H.Y., Mu, X.J., Stutz, A.M., Tanzer, A., Cayting, P.D., Snyder, M., Kim, P.M., Korbel, J.O. & Gerstein, M.B. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nature biotechnology* **28**, 47-55 (2010).
50. Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., Taillon, B.E., Chen, Z., Tanzer, A., Saunders, A.C., Chi, J., Yang, F., Carter, N.P., Hurles, M.E., Weissman, S.M., Harkins, T.T., Gerstein, M.B., Egholm, M. & Snyder, M. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420-426 (2007).
51. Rozowsky, J., Abyzov, A., Wang, J., Alves, P., Raha, D., Harmanci, A., Leng, J., Bjornson, R., Kong, Y., Kitabayashi, N., Bhardwaj, N., Rubin, M., Snyder, M. & Gerstein, M. AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Molecular systems biology* **7**, 522 (2011).
52. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D. & Daly, M.J. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491-498 (2011).

53. Ye, K., Schulz, M.H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865-2871 (2009).
54. Albers, C.A., Lunter, G., MacArthur, D.G., McVean, G., Ouwehand, W.H. & Durbin, R. Dindel: accurate indel calls from short-read data. *Genome research* **21**, 961-973 (2011).
55. Fiston-Lavier, A.S., Carrigan, M., Petrov, D.A. & Gonzalez, J. T-lex: a program for fast and accurate assessment of transposable element presence using next-generation sequencing data. *Nucleic acids research* **39**, e36 (2011).
56. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009).
57. Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.K., Chrast, J., Lagarde, J., Gilbert, J.G., Storey, R., Swarbreck, D., Rossier, C., Ucla, C., Hubbard, T., Antonarakis, S.E. & Guigo, R. GENCODE: producing a reference annotation for ENCODE. *Genome Biol* **7 Suppl 1**, S4 1-9 (2006).
58. Coffey, A.J., Kokocinski, F., Calafato, M.S., Scott, C.E., Palta, P., Drury, E., Joyce, C.J., Leproust, E.M., Harrow, J., Hunt, S., Lehesjoki, A.E., Turner, D.J., Hubbard, T.J. & Palotie, A. The GENCODE exome: sequencing the complete human exome. *European journal of human genetics : EJHG* **19**, 827-831 (2011).
59. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & Genome Project Data Processing, S. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
60. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
61. Habegger, L., Sboner, A., Gianoulis, T.A., Rozowsky, J., Agarwal, A., Snyder, M. & Gerstein, M. RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics* **27**, 281-283 (2010).
62. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106 (2010).
63. Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N. & Golani, I. Controlling the false discovery rate in behavior genetics research. *Behavioural brain research* **125**, 279-284 (2001).
64. Portales-Casamar, E., Thongjuea, S., Kwon, A.T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W.W. & Sandelin, A. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic acids research* **38**, D105-110 (2010).
65. The_ENCODE_Project_Consortium. An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature* **accepted** (2012).
66. MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., Albers, C.A., Zhang, Z.D., Conrad, D.F., Lunter, G., Zheng, H., Ayub, Q., DePristo, M.A., Banks, E., Hu, M., Handsaker, R.E., Rosenfeld, J.A., Fromer, M., Jin, M., Mu, X.J., Khurana, E., Ye, K., Kay, M., Saunders, G.I., Suner, M.M., Hunt, T., Barnes, I.H., Amid, C., Carvalho-Silva, D.R., Bignell, A.H., Snow, C., Yngvadottir, B., Bumpstead, S., Cooper, D.N., Xue, Y., Romero, I.G., Wang, J., Li, Y., Gibbs, R.A., McCarroll, S.A., Dermitzakis, E.T., Pritchard, J.K., Barrett, J.C., Harrow, J., Hurles, M.E., Gerstein, M.B. & Tyler-Smith, C. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-828 (2012).
67. Mu, X.J., Lu, Z.J., Kong, Y., Lam, H.Y. & Gerstein, M.B. Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. *Nucleic acids research* **39**, 7058-7076 (2011).
68. Balasubramanian, S., Habegger, L., Frankish, A., MacArthur, D.G., Harte, R., Tyler-Smith, C., Harrow, J. & Gerstein, M. Gene inactivation and its implications for annotation in the era of personal genomics. *Genes & development* **25**, 1-10 (2011).
69. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
70. Kang, H.J., Kawasawa, Y.I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A.M., Pletikos, M., Meyer, K.A., Sedmak, G., Guennel, T., Shin, Y., Johnson, M.B., Krsnik, Z., Mayer, S., Fertuzinhos, S., Umlauf, S., Lisgo, S.N., Vortmeyer, A., Weinberger, D.R., Mane, S., Hyde, T.M., Huttner, A., Reimers, M., Kleinman, J.E. & Sestan, N. Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483-489 (2011).
71. Johnson, M.B., Kawasawa, Y.I., Mason, C.E., Krsnik, Z., Coppola, G., Bogdanovic, D., Geschwind, D.H., Mane, S.M., State, M.W. & Sestan, N. Functional and evolutionary insights into human brain development through global transcriptome analysis. *Neuron* **62**, 494-509 (2009).

Resource Sharing Plan

Data that will be collected in this project are: whole genome sequencing data from DNA and RNA extracted from brain tissue and blood of postmortem human embryos.

Biomaterials are DNA and RNA extracts prepared from clonal cultures derived from primary brain tissue and directly from fetal tissues.

We plan submitting the data and available biomaterials to a public repository to enable rapid sharing of data and biospecimens in the scientific community. For example, all sequencing data will be deposited to NCBI GEO (rnaseq reafs) and dbGAP (dnaseq) as soon as the paper describing the data is submitted

The timetable for deposition of the data and/or biomaterials will be after publication of the research or six months after the end of the award period, whichever is shorter.

Data will be distributed from the Vaccarino, Sestan and Gerstein laboratories. These laboratories will certify the quality of all data generated prior to submission to the repository and will review the data for accuracy after submission. Biospecimens will be distributed by the Vaccarino laboratory at Yale University. We have budgeted for shipping costs and for personnel (Ms. Tomasini) that that will organize the receipt, storage, cataloguing and shipment of samples.

We will ensure that any applicable biomaterials will be stored at Yale University in the Vaccarino laboratory facilities and will be tracked via a web-accessible database where samples are listed without personal identifiers.

To receive the biomaterials, recipient investigators and their Institutions will be required to sign material transfer agreements (MTA) in behalf of Yale University. The recipient will be also required to provide written assurance and evidence that (1) the biomaterials will be used solely in accord with their Institutional review; (2) that biomaterials will not be further distributed by the recipient without consent of our Program; (3) that biomaterials will not be used for commercial purposes.

Requests from for-profit corporations to use the cells commercially will be negotiated by our institution's technology transfer office. All licensing shall be subject to distribution pursuant to our institution's policies and procedures on royalty income. The technology transfer office will report any invention disclosure submitted to them to the appropriate Federal Agency.

Our institution and we will adhere to the NIH Grants Policy on Sharing of Unique Research Resources including the " Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies (GWAS)" issued in August 28, 2007 (http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html) and the “Sharing Human Data via the National Database for Autism Research” to serve the autism research community as a common platform for exchanging data, tools, and research-related information (http://ndar.nih.gov). Should any intellectual property arise which requires a patent, we would ensure that the technology remains widely available to the research community in accordance with the NIH Principles and Guidelines document.

Research resources generated with funds from this grant will be freely distributed, as available, to qualified academic investigators for non-commercial research. My institution and I will adhere to the NIH Grants Policy on Sharing of Unique Research Resources including the "Sharing of Biomedical Research Resources: Principles and Guidelines for Recipients of NIH Grants and Contracts" issued in December, 1999. http://ott.od.nih.gov/policy/rt_guide_final.html Specifically material transfers would be made with no more restrictive terms than in the Simple Letter Agreement or the UBMTA and without reach through requirements. Should any intellectual property arise which requires a patent, we would ensure that the technology remains widely available to the research community in accordance with the NIH Principles and Guidelines document.