Chapter 17

Identification of low allele frequency mosaic mutations in Alzheimer Disease

Carlo Sala Frigerio [1,2,*], Mark Fiers [1,2], Thierry Voet [3,4], and Bart De Strooper [1,2,5,*]

[1] VIB Center for Brain & Disease Research, Leuven, Belgium

[2] Center for Human Genetics, Universitaire ziekenhuizen and LIND, KU Leuven, Leuven, Belgium

[3] Department of Human Genetics, University of Leuven, KU Leuven, Leuven, Belgium

[4] Wellcome Trust Sanger Institute, Hinxton, UK

[5] Dementia Research Institute (DRI-UK), University College London, Queen Square, WC1N 3BG London, UK

* Corresponding authors contact information:

Bart De Strooper

Email: bart.destrooper@cme.vib-kuleuven.be

VIB Center for Brain & Disease Research KU Leuven O&N 4

Herestraat 49, bus 602

3000 Leuven

Belgium

Carlo Sala Frigerio

Email: carlo.salafrigerio@cme.vib-kuleuven.be

VIB Center for Brain & Disease Research KU Leuven O&N 4

Herestraat 49, bus 602

3000 Leuven

Belgium

Running head: Somatic mutations in AD

i. Summary

Germline mutations of *APP*, *PSEN1* and *PSEN2* genes cause autosomal dominant Alzheimer disease (AD). Somatic variants of the same genes may underlie pathogenesis in sporadic AD, which is the most prevalent form of the disease. Importantly, such somatic variants may be present at very low allelic frequency, confined to the brain, and are thus very difficult or impossible to detect in blood-derived DNA. Ever-refined methodologies to identify mutations present in a fraction of the DNA of the original tissue are rapidly transforming our understanding of DNA mutation and their role in complex pathologies such as tumours. These methods stand poised to test to what extend somatic variants may play a role in AD and other neurodegenerative diseases.

ii. Key words

single cell sequencing, mosaicism, somatic variant, Alzheimer's disease, Parkinson's disease

1. Introduction

Many neurodegenerative diseases, such as Alzheimer disease (AD) and Parkinson disease (PD), have an important genetic component, and may present as familial or sporadic forms. In AD, familial forms (FAD) show dominant autosomal inheritance and are associated with point mutations in three genes (*APP*, *PSEN1* and *PSEN2*) [1], although cases of duplication of the *APP* locus are also known [2]. The causes of sporadic AD (SAD) have not yet been identified, however the clinical, histopathological and biochemical similarities between SAD and FAD cases suggest a common pathological cascade.

Given the enormous genomic heterogeneity found in cells comprising the human brain [3–7], it is fair to hypothesize that during brain organogenesis some individuals may acquire somatic mutations in genes known to be causative in FAD forms. These individuals would then have patches of brain tissue bearing pathogenic mutations, which could start the same pathological cascade of events as seen in FAD patients, causing sporadic AD. Indeed, patches of neurons bearing somatic mutations of FAD genes would produce Aβ or tau aggregates, which could then spread in the brain parenchyma seeding further aggregation of amyloid or tau, respectively, in a process known as template-seeded aggregation [8, 9] (Figure 1).

Single nucleotide variants (SNVs) are the most abundant class of mutations responsible for genomic variation between humans at a population level [10], and hence a main cause of cellular genomic heterogeneity within an individual. Indeed, it is estimated that per cell division approximately 10^{-10} errors per base pair accumulate [11], suggesting that a large majority of cells within a body may be genetically dissimilar. Whole genome sequencing of single human neurons suggests the presence of approximately 1500 somatically acquired SNVs per neuron [3]. Interestingly, transcriptionally active genes were enriched for somatic SNVs and the latter were often private to single neurons, indicating that also mechanisms other than DNA replication are causal for somatic DNA mutations [3] (Figure 1).

Analysis of somatic mutations is a rapidly evolving field, that has seen a fast improvement in recent years thanks to the development of high throughput / next generation sequencing (NGS) and more refined data analysis algorithms. There is not yet a unique gold standard method for somatic mutation detection, moreover different methods are required for investigating fine-grain mosaicism (due to private somatic mutations present in very small numbers of cells) when compared to coarse-grain mosaicism (due to somatic variants present in a sizeable percentage of brain cells) (Figure 1). In this chapter we will present methodologies available for investigating somatic variants, with a particular focus on those we have employed in our laboratory. Of importance, the same methodologies could be applied to the study of other neurodegenerative diseases for which a similar somatic mutation-based hypothesis can be envisaged, e.g. Parkinson disease (by targeting *SNCA*, *LRRK2*, *VPS35* and possibly *PINK1*, *DJ-1* and *PARK2* genes) and prion diseases such as Creutzfeldt-Jakob (by targeting the *PRNP* gene) [12].

2. Methodologies

The search for somatic variants present in a sizeable fraction of brain cells can be performed by analysing bulk DNA extracted from a frozen post-mortem brain tissue sample. The technical challenge of this approach lies in the fact that when only a few clonal cells contribute the signal for the mutant allele, thousands or millions of other cells will deliver a wild type signal. Classic Sanger DNA-sequencing is not very well suited to detect somatic mutations with low allele frequency, having a sensitivity threshold that approximates 20% mutant allele frequency (ALT)

[13, 14]. The development of NGS technologies has allowed more profound analysis of bulk DNA samples: by independently sequencing hundreds to thousands of alleles from the starting DNA sample, wild type and mutant signals at a particular genomic locus can be efficiently detected thus allowing to reach sensitivities of 5% ALT and lower. Although powerful, NGS sample preparation often involves several steps of PCR amplification of the original DNA, leading to polymerase errors. Moreover, the sequence-by-synthesis chemistry of Illumina (a widely used sequencing technology today) is also error-prone. Effectively, all *in vitro* polymerase errors limit the sensitivity of detection of genuine low frequency mutations. To overcome such limitations of NGS, several methods for sample preparation and data analysis have been developed.

To achieve the high sequencing depth that is required for the detection of somatic mutations with low ALT cost efficiently, it is recommended to focus the analysis only on specific target regions of the genome that are already suspected to be involved. In the case of AD, it makes sense to selectively sequence the *APP*, *PSEN1* and *PSEN2* genes, as these are the only genes known to cause FAD when mutated, foregoing the analysis of the rest of the genome/exome. This targeted approach can be attained with two different methodologies: 1) targeted enrichment or capture using a custom probeset on a genome sequencing library, or 2) target amplicon generation using a custom multiplexed PCR. Several vendors offer custom-designed sequence-enrichment and/or target-amplicon panels, e.g. Agilent SureSelect Target Enrichment system, Illumina Nextera Rapid Capture Custom Enrichment kit and TruSeq Custom Amplicon kit, Qiagen GeneRead DNAseq custom panels, Raindance ThuderStorm and ThunderBolt systems, and Roche NimbleGen's SeqCap. It is recommended to test the sensitivity and accuracy of the targeting approach chosen at the beginning of a research project. This can be done by setting up a preliminary experiment analysing a series of "synthetic mosaic" samples, which can be obtained by serially diluting a bulk DNA sample containing known heterozygous mutations in the region of interest with bulk *wild type* DNA. The expected allelic frequency of the "synthetic mosaic" variants can then be compared to the observed values and to the overall sequencing noise. This provides information about the lowest allelic frequency reliably detectable and about which parameters in sample preparation / data analysis would need optimization.

The study of fine-grained mosaicism requires to sequence the DNA of single cells: by querying each cell on its own, the contribution of each cell to the overall genomic heterogeneity of the brain can be determined. In addition to the limitations inherent to NGS (see above), single cell DNA sequencing is confounded by the minute amounts of starting material (6.6 pg of DNA for a diploid cell) that has to be amplified prior to sequencing. Such whole-genome amplification (WGA) procedures can lead to false positives –e.g. due to DNA polymerase errors in early rounds of amplification – as well as to false negatives –e.g. due to locus or allelic dropout. Several WGA methods have been developed, the choice for a specific method is primarily guided on the desired classes of genetic variation to be detected genome wide [15, 16].

2.1 Considerations on the tissue samples to be investigated

Depending on the developmental timing of its appearance, a somatic mutation will be spread more or less throughout the body; while early events could lead to a mutation being present in a fraction of both blood and neuronal cells, it is possible to have brain-private somatic mutations if they appeared after gastrulation. Therefore, in order not to miss somatic mutations present only in the brain, we have analysed brain tissue samples from deceased AD patients instead of blood-derived DNA.

Somatic mutation analysis in neurodevelopmental diseases and cancer is facilitated by the fact that diseased tissue can be clearly identified thanks to specific histological features. DNA isolated from the diseased tissue will be enriched for the mutant signal, while parallel sequencing of healthy tissue provides a background reference to exclude germline and de-novo mutations and to control for sequencing errors. Unfortunately, in AD it is not possible to clearly discern a brain area which is more likely to harbour a somatic variant. Reports of a patterning in the perceived spread of tau aggregates suggests that the entorhinal cortex could be one of the earlies areas affected [17], however the mechanistic implications of the "prionoid spread" of amyloid seeds suggests that any brain area could be the source of the first amyloid seeds. Therefore, the search for somatic mutations in sporadic AD should be directed towards several different brain areas.

2.2 Bulk DNA sequencing: sequence-enrichment approach

We have previously employed a custom Roche NimbeGen SeqCap panel to enrich sequencing libraries for the loci of *APP*, *PSEN1*, *PSEN2* and *MAPT* [18]. The *MAPT* gene was included in the analysis even though germline *MAPT* mutations do not cause AD for the reason that tau, the product of the *MAPT* gene, is a primary player in the biochemical pathological cascade of AD; in a somatic mutation scenario it can be thus hypothesized that somatic *MAPT* mutations could lead to AD in a "two-hit" mechanism [19]. A more conservative approach would be of course to only consider the three known FAD-causative genes (*APP*, *PSEN1* and *PSEN2*).

We chose to target the entire loci of our genes of interest, so that we could leverage the sequencing data to simultaneously analyse both somatic SNVs and CNVs, exploiting disturbances in B-allele fractions of germline heterozygous SNPs for the detection of subclonal

CNVs. Both types of somatic mutations could be relevant for the development of AD. We also included 10 kbp pad regions upstream and downstream of each locus to avoid drastic drops in sequencing coverage at both ends of the loci, which would otherwise complicate the analysis. The regions targeted were (based on the Human Genome release hg19): *APP* (chr21:27242859-27553138), *PSEN1* (chr14:73593141-73700399), *PSEN2* (chr1:227048271-227093804) and *MAPT* (chr17:43961646-44115799). As the hg19 release also foresees an alternative assembly for chromosome 17, we also included the *MAPT* regions specific for the alternate assembly (chr17_ctg5_hap1:762280-895830). The actual probes were designed by NimbleGen according to our desired target areas, manufactured and shipped in solution.

The experimental workflow begins with the isolation of high quality gDNA from tissue samples. To isolate bulk gDNA, frozen brain tissue is chopped with a scalpel and incubated overnight with Protease K at 50°C with mild agitation. RNA is then degraded during a 15 minute incubation at 37°C with RNase A (Qiagen, Venlo, The Netherlands). DNA is isolated with phenol:choloroform:isoamyl alcohol, washed twice with choloroform:isoamyl alcohol and precipitated with 100% ethanol. The DNA pellet is further washed with 70% ethanol, dried and finally resuspended in Tris-EDTA buffer. Sample preparation must avoid excessive vortexing or heating, as this would fragment or denature gDNA and render subsequent steps impossible. The concentration of the DNA is determined with a QuBit fluorimeter (Life Technologies, Gent, Belgium), which specifically detects double-stranded DNA, and quality is determined with a NanoDrop spectrophotometer (NanoDrop, Wilmington, DE) to exclude residual contaminants.

High quality gDNA is sheared with a Covaris sonicator (Covaris, Woodingdean Brighton, UK) to produce 300 bp fragments on average and indexed libraries are then prepared with the TrueSeq DNA kit from Illumina (Illumina, San Diego, CA). Individual libraries can then be pooled prior to sequence enrichment, the number of samples that can be pooled is function of the total number of bases captured, the intended sequencing depth for every base and the sequence output of the instrument that will be used to analyse the samples. For this, the projected output of the instrument (according to the manufacturer) can be divided by the size of the target region and by the desired sequencing depth. We have obtained high sequencing depth (>2000X per position, on average) by pooling 10 samples enriched for a ~600 kb target region and sequencing with an Illumina HiSeq 2500 in rapid mode. Demultiplexing the sequencing data amongst the pooled samples is a critical step to avoid the wrong assignment of reads to a particular sample which may result in false signals. Sample-specific indices are preferably at least 3 nucleotides different, allowing maximum one mismatch in the index sequence during demultiplexing. After demultiplexing, sequencing data are usually encoded in FASTQ files.

Sample-individual FASTQ files are aligned to the human reference genome using BWA [20] and converted to a BAM file format for downstream analysis [21]. Next, since indels can cause misalignment of the reads, the alignment should be refined by local realignment around indels using the GATK IndelRealigner tool [22], and base qualities are recalibrated using the GATK BaseRecalibrator tool [22] to correct systematic technical errors in base quality calling by the sequencing instrument. These steps of data pre-processing will yield a BAM file which can then be used to call variants. There is a great variety of variant calling algorithms (see Table) based on different statistical algorithms. We have efficiently used Samtools mPileup function [23] together with VarScan 2.0 [24] to generate a list of candidate somatic variants. A useful approach for variant calling involves the generation of a conservative list of candidate variants called by different algorithms. Indeed, each variant calling software may identify different sets of variants, due to the unique properties of each algorithm. Variants called by multiple algorithms may be considered as high confidence candidates.

Of importance, some somatic variant calling algorithms expect that a test sample is compared with a "matched normal" sample (e.g. a tumour sample compared to a healthy tissue sample) to efficiently rule out false positive calls and germline variants. In the case of AD tissue, it is not possible to perform such comparison, since, differently from tumour studies, it is not obvious which tissue should have a somatic mutation and which should be devoid of it. Hence, we performed variant calling on each sample on its own.

Candidate variants are then annotated, i.e. information on the genomic region, presence in databases and potential functional consequences (synonymous, nonsynonymous, nonsense,…) is retrieved, finally yielding a VCF (variant call format) file. Various tools exist also for variant annotation, we have used SnpEff [25] and Annovar [26]. Further analysis of variants and sequencing data can be carried out efficiently using R (http://www.r-project.org/). Annotation of variants can be useful to prioritize a long list of candidate variants for further validation and is instrumental in gaining insight on the biological consequences of true somatic variants.

It is important to notice that all the above-mentioned software is constantly updated, hence it is recommended to use the latest versions, to consistently use one version for the analysis of all samples in a project, and in any case to correctly report the version number of each software used.

The bioinformatics analysis can be computationally intense, in particular the alignment step, the GATK-based steps, and the steps in R if dealing with a big target region or a high number of candidate variants. Analysis can be efficiently tackled by computing clusters or by the use of dedicated servers.

2.3 Bulk DNA sequencing: targeted approach with tagged reads

Preparation of sequencing libraries involves several PCR steps, which introduce errors which may complicate somatic variant analysis. The main errors due to PCR are: 1) the incorporation of wrong nucleotides and 2) the skew in amplification of mutated alleles and wild type alleles (see Figure 2) which could result in false negative (in case the mutant allele is under-represented) or false positive (in case a PCR error is over-represented) calls. The skewed amplification behaviour and the introduction of incorrect nucleotides are inherent to PCR and DNA polymerases [27, 28], however several smart workarounds have been developed to prevent and counter them.

Two approaches (Duplex Sequencing [29] and UMI-TSCA [30]) share the general principle to add barcodes to the original DNA molecules in order to track the daughter molecules produced during PCR amplification. This allows 1) to correctly remove PCR duplicates (deduplication) by counting unique barcodes instead of raw reads, and 2) to generate consensus reads by pooling all reads sharing the same barcodes (Figure 2).

In the Duplex Sequencing approach [29, 31], gDNA is sheared and fragments are ligated to duplex sequencing adapters, resulting in DNA fragments bearing 12-nucleotide long barcode sequences at both ends. Fragments barcoded at each end are PCR amplified and sequenced, then reads are grouped by the barcodes and consensus sequences are derived. This approach also allows to double-check a variant by identifying its presence in both families of sequencing reads that derive from the complementary Watson and Crick strands of the original gDNA molecule. However, this method still requires the development of a sequence enrichment panel to prepare a targeted sequencing library.

An alternative approach [30] applies a modification of the TruSeq Custom Amplicon (TSCA) kit from Illumina to introduce a Unique Molecule Identifer (UMI) in place of the P5 sample index. The TSCA kit is a custom-designed panel of probes recognizing a list of user-defined genomic targets; for each target, two probes are designed, one upstream and one downstream.

After hybridization to the gDNA targets, the upstream probe is extended by a DNA polymerase onto the downstream probe, producing a copy of the original gDNA region. Next, a unique barcode is added to each copy and the product is PCR amplified to produce the sequencing library: during this step the barcodes are copied together with the genomic target, thus keeping the original labelling. This approach is more straightforward than the former, as it directly generates a targeted amplicon library, however it lacks the possibility of copying both strands of the original gDNA target.

A third method, CirSeq [32], dispenses from using barcodes altogether: instead of PCR amplifying the intended target, the gDNA is sheared, fragments are circularized and using random primers and a high processivity DNA polymerase concatenated copies of the original template are made. The strength of the method lies in the fact that mutations are not propagated by PCR, as each new copy comes from the original DNA sequence. However, it also requires an extra step to select the regions of interest when dealing with a targeted approach, which would require additional PCR reactions.

We have applied the UMI-modified TSCA approach to analyse somatic mutations in AD, our choice was based on the fact that this approach is the most straightforward for a targeted sequencing analysis, and built around a commercially available kit already aimed at the analysis of large genomes (such as the human genome). In order to maximize the sequencing coverage of areas of interest we targeted the exons of *APP*, *PSEN1*, *PSEN2* and *MAPT* known to harbour pathogenic mutations. The TSCA panel was designed by Illumina, to accommodate 250 bp-long amplicons, the total area covered is ~7 kb long thus allowing a very high coverage when sequencing 12 pooled samples on an Illumina MiSeq 2x300 sequencing run.

To prepare sequencing libraries with the modified TSCA method, first the gDNA is denatured and slowly cooled in presence of the probeset, allowing specific annealing of the probes with their cognate sites on the gDNA. Next, a DNA polymerase extends the upstream probe and a

DNA ligase joins the newly synthesized copy of the gDNA to the downstream probe. The ligation-extension products are purified using a filter plate (provided with the kit, per manufacturer's recommendations). Next, in place of the canonical direct PCR amplification of the extension-ligation products, we performed one cycle of PCR with the Illumina P7 primers (bearing a sample-specific index) and a modified P5 primer (P5') which contains a random 12-nucleotide sequence (which constitutes the UMI) in place of the second sample-specific index. Given the high number of possible UMIs (4^{12}=16,777,216), it is highly likely that each copy of a specific genomic target in the original sample will get a different UMI. Next, PCR products are purified with Ampure XP magnetic beads and a second round of PCR is carried out using the same sample-specific P7 primer as before and a P5'' primer that anneals downstream of the UMI and carries the Illumina-specific P5 sequence handle for correct loading on an Illumina flowcell. The final PCR product is again purified using Ampure XP magnetic beads, quantified using the KAPA library quantification kit for Illumina libraries (Kapa Biosystems, Wilmington, MA), and equimolar sample-specific libraries are pooled. Since only the P7 sample index can be used to label different biological samples, the maximum amount of samples that can be pooled on a single sequencing run is 12, as Illumina provides only 12 different P7 indices, restricting high sample throughput.

The Illumina MiSeq sample sheet (which instructs the sequencing instrument on the run parameters) has to be modified to account for a longer (12 nucleotides instead of 8) i5 index read (which covers the UMI). In order to keep the UMI tied to each specific R1 and R2 read pair (which cover the actual amplicon), we have appended the UMI sequence to the header of each R1 and R2 read. To analyse the data, we first aligned the reads to the reference genome using BWA-MEM, and performed local realignment around indels and base quality recalibration with the dedicated GATK software.

Next, UMIs are leveraged to correct for PCR artefacts and to generate consensus nucleotide calls at each position assessed. The overall approach is to group reads aligned to one genomic locus that share the same UMI (called "UMI family"). Subsequently, analysis can focus on a read-by-read basis or on a position-by-position basis. Following the latter approach, we have developed an algorithm (called Scotoplanes, Figure 3) which generates a pileup of the nucleotides aligned at each position of our target region, while taking the UMI into account. Deduplication of reads nucleotides with the same UMI depends on the level of UMI duplication and whether or not all the reads with the same UMI support the same nucleotide: 1) a UMI appears only once: the associated nucleotide is counted (e.g. UMI1, Figure 3); 2) a UMI is observed more than once, but all instances support the same nucleotide: the nucleotide is retained but counted only once (e.g. UMI2, Figure 3); 3) a UMI is observed more than once, and associates with a number of different nucleotides: the most abundantly present nucleotide is retained (and counted once), but only if it is at least four times more abundant than the second most frequent nucleotide associated with the same UMI at that position (e.g. UMI3, Figure 3). After such position-by-position UMI deconvolution, allele frequencies are calculated at each position, and major and minor alleles are called. Candidate variants can then be annotated using SnpEff and Annovar as seen above to prioritize variants for further validation.

2.4 Single cell DNA sequencing

Single cell DNA sequencing has the capacity to detect somatic variants down to the biological unit of organs. Although powerful, the method is still technically highly challenging, which has to be taken into account when designing a project [16]. In particular, a cell's gDNA has to be amplified to obtain enough material for sequencing analysis; and methods for WGA are still in need of optimization to improve uniformity of coverage – whereby unevenness in amplification can lead to locus or allelic drop outs and thus false negative SNVs as well as false positive CNVs – and to mitigate the introduction of polymerase errors which produce false positive SNVs [16].

As it is difficult to isolate intact single cells from a complex tissue such as the human brain, it is preferable to isolate single nuclei instead. For nuclear isolation, tissue samples are homogenized in the presence of very low levels of detergent in order to avoid compromising the nuclear envelope. Nuclei can be recovered by centrifugation on an Optiprep density gradient. The nuclei are then labelled with a DNA stain (e.g. DAPI or DRAQ5) and can be additionally marked for cell-type specific nuclear antigens. For example, to specifically recover neuronal nuclei one can use a fluorescently-conjugated antibody for the neuronal-specific splicing regulator Rbfox3 (NeuN). Fluorescently labelled nuclei can subsequently be single-sorted into 96 well plates containing lysis buffer using fluorescent activated cell sorting (FACS) platforms. Following isolation, for the purpose of SNV analysis the gDNA of the cell is preferably amplified by isothermal amplification using random primers and the Φ29 polymerase, a DNA polymerase with high processivity and low error rates [15, 33]. The resulting multiple displacement amplification (MDA) product can be converted in a sequencing library using conventional methods, or can be used in conjunction with a sequence-enrichment approach. As for bulk DNA approaches, it is best to include a sequence-enrichment step to focus on genomic regions which, if mutated, could drive AD pathogenesis.

Although variant calling algorithms developed for bulk DNA have also been used for calling variants in single cell DNA sequencing data, this latter kind of data has peculiar aspects which warrants the use of bespoke algorithms. Specifically, the technical artefacts introduced during WGA need to be taken into account, to prevent calling false variants. Due to the error rate of WGA polymerases –e.g. the per-cycle per-base error rate of MDA has been estimated to approximate 3.2 x 10^{-6} [34] – it is currently impossible to call SNVs that are private to a cell with absolute confidence [15]. Instead, the candidate SNV has to be reported by at least a few cells to increase reliability. Algorithms specifically designed for analysing SNVs in single cell DNA sequencing data, as Monovar [35] and Single Cell Genotyper [36], are emerging.

Single cell DNA sequencing has been used widely to study tumour evolution, and recently it has been used for the identification of somatic mutational signatures in the human brain [3]. A possible future application to investigate somatic genetic variation underlying the cause of AD would be to couple single nucleus sorting and DNA library preparation with a sequence enrichment panel targeted for the *APP*, *PSEN1* and *PSEN2* loci. Alternatively, a direct amplicon based assay targeting the most pathogenic mutant sites for AD may be directly applied on single cell DNA at scale without upfront WGA [37]. Although single-cell sequencing allows SNV analysis at relatively low coverage, many different cells may need be to be pooled and sequenced for the discovery of low frequency somatic SNVs in the diseased tissue.

3. Confirmation of candidate somatic variants

Although bioinformatics analysis of sequencing data is based on continuously improving algorithms, the stochastic nature of errors introduced at any step of sample preparation and sequencing means that the analysis can at best give a list of candidate variants with differing degrees of confidence. Confirmation of called somatic variants by an alternative approach remains therefore important to validate the results and to develop increasingly reliable somatic variant callers. There are several ways to confirm candidate variants, depending on the approach used at first pass (bulk tissue or single cell analyses), the allelic frequency of the candidate variant and the number of candidates to be tested (for budgeting reasons).

3.1 Sanger sequencing

The resolution of Sanger sequencing for detecting somatic variants is limited to variants with an allelic frequency of 20% or more [13, 14]. In this procedure, the candidate mutant locus is amplified from bulk DNA with specific PCR primers and then directly sequenced. However, it requires that the sequencing runs have very low levels of noise to be able to reliably identify the somatic variant. Parameters that have to be carefully controlled are in particular the clean-up and resuspension of the sample after the PCR with fluorescent dideoxy nucleotides. In practice, it is more efficient to use alternative methods even for high allele frequency somatic variants.

3.2 TA-subcloning and Sanger sequencing

The principle of this approach is to exploit bacteria to partition the bulk DNA signal, so that Sanger sequencing can be efficiently used to validate a candidate somatic variant.

The candidate mutation is amplified from the original bulk gDNA sample using a pair of specific PCR primers. The PCR amplicons are then cloned into a plasmid vector, ideally one which allows quick and efficient subcloning without the need for restriction digest of the amplicon such as those provided by the TOPO TA cloning kit from ThermoFisher (ThermoFisher, Waltham, Massachusetts). After bacterial transformation and plating, single independent colonies are isolated and sequenced. Each colony will bear only one allele, hence Sanger sequencing will yield a yes/no answer for the presence of the candidate mutation. By counting the ratio of mutant colonies, we can infer the allelic frequency of the target mutation in the original DNA sample.

This approach works best with mutations having a relatively high allelic frequency, to avoid having to sequence many colonies, but could in theory be up-scaled to detect low frequency variants.

3.3 Digital droplet PCR

Following the same principle of partitioning the original bulk gDNA signal, digital droplet PCR (ddPCR) allows to achieve very good sensitivity, down to 0.1% ALT or lower (Figure 4) [38]. In a ddPCR assay, a PCR reaction is mixed with oil in a microfluidics chip, leading to emulsification of the PCR reaction into thousands of nanoliter droplets containing either zero, one or more than one template DNA molecules. After completion of the PCR reaction, droplets are read one by one, measuring the presence of fluorescently labelled allele-specific probes. Statistical analysis of the count data, based on Poisson statistics, allows calculating confidence levels for the allelic frequency determined for each sample. The extremely high partition of the template DNA and the discrete nature of droplet counting allows to detect very low levels of mutant allele molecules.

For each candidate variant a specific PCR primer set is designed together with two hydrolysis probes (e.g. Taqman probes), one recognizing the wild type allele and one recognizing the mutant allele, labelled with fluorophores with different spectral wavelengths (e.g. FAM and HEX). The use of two differentially labelled probes allows the simultaneous detection of the two alleles in the same reaction, thus offering a better quantification than running mutant and wild type assays in different tubes. For each newly developed assay, several parameters need to be optimized, including correct annealing temperature, starting gDNA quantity and the numbers of wells/droplets required.

Each assay should be run with appropriate negative and positive control samples. As negative control samples, assays should be run without template gDNA to assess the levels of inherent background fluorescence of the probes, and should be run with gDNA devoid of the target somatic variant (wild type control) to assess the signal coming from non-specific PCR products. As positive controls, a dilution series of a construct containing the target somatic variant spiked to wild type DNA could be analysed. For this purpose, a gBLOCK synthetic double stranded DNA molecule mimicking the genomic region containing the somatic variant can be ordered from IDT (Integrated DNA Technologies, Leuven, Belgium). Such dilution series will allow to readily assess the sensitivity of the assay against the noise signal derived from the analysis of pure wild type samples. BioRad (Hercules, CA) provides a complete workflow for ddPCR analysis. Assays can be designed using their online software (https://www.bio-rad.com/digital-assays/#/assays-create/mutation). The BioRad QX200 Droplet Digital PCR system foresees an instrument for the preparation of the emulsion PCR reaction and a droplet counter. Moreover, BioRad provides a software (QuantaSoft) to analyse droplet counts (Figure 4).

3.4 Resequencing of candidate variants

Another approach to validate candidate variants is to re-sequence at very high depth target amplicons: as most PCR and sequencing errors are random, it would be unlikely to retrieve a false positive variant in two separate sample preparations and sequencing runs.

Deep amplicon sequencing is a viable validation option for candidates identified by a sequence-enrichment approach. Primer sets are designed to amplify each target candidate, and each PCR product is labelled with a sample-specific index to allow for pooling. Amplicons can then be sequenced at very high coverage (>100,000X) on an Illumina MiSeq.

For candidates identified in an amplicon-based approach a more relevant re-sequencing approach would entail change of sequencing chemistry altogether. Pacific Biosciences (PacBio) sequencing offers a good alternative to Illumina sequencing for this kind of validation, as the sequencing chemistry, and therefore the error pattern, are radically different between the two. For PacBio sequencing, after preparing indexed amplicons, library preparation foresees the ligation of bubble adaptors on both ends of each DNA molecule, enabling recursively sequencing of the Watson and Crick DNA strands of that molecule on an RSII or Sequel platform (Pacific Biosciences, Menlo Park, CA). Following sequencing, all subreads from a DNA molecule are piled and a consensus sequence is derived. As PacBio polymerase errors are stochastic, the consensus sequence generation is able to correct most sequencing errors, thus providing a high confidence validation.

4. Conclusions

Somatic variant analysis is a rapidly developing field, with continuously improved sample preparation and sequencing methodologies and continuously refined statistical algorithms for variant detection. Interest is that somatic variant detection goes across multiple fields of biology, e.g. developmental biology, tumour biology, neurobiology, toxicology and has even forensics applications.

In the field of AD research, where the primary cause of sporadic AD is still unknown, the possibility of somatic mutations being pathogenic drivers is a long standing question [39, 40]. We foresee that continuous optimization of the methodologies will finally clarify the role of somatic mutations in AD. Moreover, the methodologies illustrated in this chapter can be extended to other neurodegenerative diseases, in particular for those in which a template-seeded aggregation mechanism is involved, e.g. Parkinson disease [41, 42] and prion diseases.

Bibliography

1. Cruts M, Theuns J, Van Broeckhoven C (2012) Locus-specific mutation databases for neurodegenerative brain diseases. Hum Mutat 33:1340–1344. doi: 10.1002/humu.22117

2. Rovelet-Lecrux A, Hannequin D, Raux G, et al (2006) APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. Nat Genet 38:24–26. doi: ng1718 [pii]10.1038/ng1718

3. Lodato MA, Woodworth MB, Lee S, et al (2015) Somatic mutation in single human neurons tracks developmental and transcriptional history. Science (80-) 350:94–98. doi: 350/6256/94 [pii]10.1126/science.aab1785

4. Upton KR, Gerhardt DJ, Jesuadian JS, et al (2015) Ubiquitous L1 mosaicism in hippocampal neurons. Cell 161:228–239. doi: 10.1016/j.cell.2015.03.026

5. Evrony GD, Lee E, Park PJ, Walsh CA (2016) Resolving rates of mutation in the brain using single-neuron genomics. Elife. doi: 10.7554/eLife.12966

6. McConnell MJ, Lindberg MR, Brennand KJ, et al (2013) Mosaic copy number variation in human neurons. Science 342:632–637. doi: 10.1126/science.1243472

7. Evrony GD, Cai X, Lee E, et al (2012) Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. Cell 151:483–496. doi: 10.1016/j.cell.2012.09.035

8. Aguzzi A, Lakkaraju AK (2015) Cell Biology of Prions and Prionoids: A Status Report. Trends Cell Biol 26 (1):40-51. doi:S0962-8924(15)00158-0 [pii]10.1016/j.tcb.2015.08.007.

9. Brettschneider J, Del Tredici K, Lee VM, Trojanowski JQ (2015) Spreading of pathology in neurodegenerative diseases: a focus on human studies. Nat Rev Neurosci 16:109–120. doi: nrn3887 [pii]10.1038/nrn3887

10. Auton A, Brooks LD, Durbin RM, et al (2015) A global reference for human genetic variation. Nature 526:68–74. doi: nature15393 [pii]10.1038/nature15393

11. Nussbaum R, McInnes RR, Willard HF (2007) Thompson & Thompson Genetics in Medicine, 7th edition. Saunders

12. Alzualde A, Moreno F, Martinez-Lage P, et al (2010) Somatic mosaicism in a case of apparently sporadic Creutzfeldt-Jakob disease carrying a de novo D178N mutation in the PRNP gene. Am J Med Genet B Neuropsychiatr Genet 153B:1283–1291. doi: 10.1002/ajmg.b.31099

13. Tsiatis AC, Norris-Kirby A, Rich RG, et al (2010) Comparison of Sanger sequencing, pyrosequencing, and melting curve analysis for the detection of KRAS mutations: diagnostic and clinical implications. J Mol Diagn 12:425–432. doi: S1525-1578(10)60082-7 [pii]10.2353/jmoldx.2010.090188

14. Jamuar SS, Lam AT, Kircher M, et al (2014) Somatic mutations in cerebral cortical malformations. N Engl J Med 371:733–743. doi: 10.1056/NEJMoa1314432

15. Macaulay IC, Voet T (2014) Single Cell Genomics : Advances and Future Perspectives. doi: 10.1371/journal.pgen.1004126

16. Gawad C, Koh W, Quake SR (2016) Single-cell genome sequencing: current state of the science. Nat Rev Genet 17:175–188. doi: nrg.2015.16 [pii]10.1038/nrg.2015.16

17. Braak H, Braak E (1991) Neuropathological stageing of Alzheimer-related changes. Acta Neuropathol 82:239–259.

18. Sala Frigerio C, Lau P, Troakes C, et al (2015) On the identification of low allele frequency mosaic mutations in the brains of Alzheimer’s disease patients. Alzheimers Dement 11:1265–1276. doi: S1552-5260(15)00120-X [pii]10.1016/j.jalz.2015.02.007

19. Small SA, Duff K (2008) Linking Abeta and tau in late-onset Alzheimer’s disease: a dual pathway hypothesis. Neuron 60:534–542. doi: S0896-6273(08)00956-2 [pii]10.1016/j.neuron.2008.11.007

20. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589–595. doi: btp698 [pii]10.1093/bioinformatics/btp698

21. Li H, Handsaker B, Wysoker A, et al (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. doi: btp352 [pii]10.1093/bioinformatics/btp352

22. McKenna A, Hanna M, Banks E, et al (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303. doi: gr.107524.110 [pii]10.1101/gr.107524.110

23. Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987–2993. doi: btr509 [pii]10.1093/bioinformatics/btr509

24. Koboldt DC, Zhang Q, Larson DE, et al (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22:568–576. doi: gr.129684.111 [pii]10.1101/gr.129684.111

25. Cingolani P, Platts A, Wang le L, et al (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly 6:80–92. doi: 19695

[pii]10.4161/fly.19695

26. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38:e164. doi: gkq603 [pii]10.1093/nar/gkq603

27. Kanagawa T (2003) Bias and artifacts in multitemplate polymerase chain reactions (PCR). J Biosci Bioeng 96:317–323. doi: S1389-1723(03)90130-7 [pii]10.1016/S1389-1723(03)90130-7

28. Gundry M, Vijg J (2011) Direct mutation analysis by high-throughput sequencing: from germline to low-abundant, somatic variants. Mutat Res 729:1–15. doi: S0027-5107(11)00266-1 [pii]10.1016/j.mrfmmm.2011.10.001

29. Schmitt MW, Kennedy SR, Salk JJ, et al (2012) Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci U S A 109:14508–14513. doi: 1208715109 [pii]10.1073/pnas.1208715109

30. Smith EN, Jepsen K, Khosroheidari M, et al (2014) Biased estimates of clonal evolution and subclonal heterogeneity can arise from PCR duplicates in deep sequencing experiments. Genome Biol 15:420. doi: s13059-014-0420-4 [pii]10.1186/s13059-014-0420-4

31. Kennedy SR, Schmitt MW, Fox EJ, et al (2014) Detecting ultralow-frequency mutations by Duplex Sequencing. Nat Protoc 9:2586–2606. doi: nprot.2014.170 [pii]10.1038/nprot.2014.170

32. Lou DI, Hussmann JA, McBee RM, et al (2013) High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. Proc Natl Acad Sci U S A 110:19872–19877. doi: 1319590110 [pii]10.1073/pnas.1319590110

33. Dean FB, Hosono S, Fang L, et al (2002) Comprehensive human genome amplification using multiple displacement amplification. Proc Natl Acad Sci U S A 99:5261–5266. doi: 10.1073/pnas.08208949999/8/5261 [pii]

34. de Bourcy CF, De Vlaminck I, Kanbar JN, et al (2014) A quantitative comparison of single-cell whole genome amplification methods. PLoS One 9:e105585. doi: 10.1371/journal.pone.0105585PONE-D-14-24544 [pii]

35. Zafar H, Wang Y, Nakhleh L, et al (2016) Monovar : single-nucleotide variant detection in single cells. doi: 10.1038/pj.2016.37

36. Roth A, McPherson A, Laks E, et al (2016) Clonal genotype and population structure inference from single-cell tumor sequencing. Nat Methods 13:573–576. doi: nmeth.3867 [pii]10.1038/nmeth.3867

37. Eirew P, Steif A, Khattra J, et al (2014) Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. Nature 518:422–426. doi: nature13952 [pii]10.1038/nature13952

38. Hindson BJ, Ness KD, Masquelier DA, et al (2011) High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. Anal Chem 83:8604–8610. doi: 10.1021/ac202028g

39. Geller LN, Potter H (1999) Chromosome missegregation and trisomy 21 mosaicism in Alzheimer’s disease. Neurobiol Dis 6:167–179. doi: S0969-9961(99)90236-X [pii]10.1006/nbdi.1999.0236

40. Beck JA, Poulter M, Campbell TA, et al (2004) Somatic and germline mosaicism in sporadic early-onset Alzheimer’s disease. Hum Mol Genet 13:1219–1224. doi: 10.1093/hmg/ddh134ddh134 [pii]

41. Proukakis C, Houlden H, Schapira AH (2013) Somatic alpha-synuclein mutations in Parkinson’s disease: hypothesis and preliminary data. Mov Disord 28:705–712. doi: 10.1002/mds.25502

42. Proukakis C, Shoaee M, Morris J, et al (2014) Analysis of Parkinson’s disease brain-derived DNA for alpha-synuclein coding somatic mutations. Mov Disord 29:1060–1064. doi: 10.1002/mds.25883

Figure captions

Figure 1. Different degrees of mosaicism and recommended workflows.

Depending on the developmental timing of appearance of a mutation, it can appear in multiple cells (exemplified by the red patch in the "coarse-grain mosaicism") or it can be private to a single cell (cell-private mutations are exemplified by the coloured dots in "fine-grain mosaicism", where each dot is a different mutation). Depending on the prevalence of a mutation in a tissue, different sequencing approaches can be undertaken in the discovery phase. Methodologies for the validation phase are mainly dictated by the allelic frequency of the candidate variants.

Figure 2. Principles of amplicon tagging.

A low-frequency mutation (indicated by a star) is present only in a fraction of the reads (represented by a blue line ending with a circle) aligning on a genomic target of interest. When no read tagging is used (e.g. after sequence-enrichment of sheared gDNA), the allelic frequency of a somatic mutation cannot be corrected for PCR duplication artefacts (a). When reads are tagged prior to PCR amplification, duplicates can be easily spotted and the allelic frequency of the mutation can be corrected (b, tags indicated by the coloured circles at the end of each read). Deduplication of PCR artefacts leads to correct assessment of the allelic frequency of somatic variant also in the case that the somatic variant is not present in the duplicated reads (c).

Figure 3. Tag-deduplication algorithm.

The algorithm represented has been implemented in a software developed in the lab to process UMI-labelled reads. The green box represents different possible scenarios of duplicated UMIs and connected reads. After alignment to the genome, reads and their associated UMIs are analysed position-by-position. Reads are grouped based on their UMIs into UMI families (e.g. UMI1, UMI2, UMI3). Depending on the duplication levels and on the supporting nucleotides found for different duplicate molecules, a read is either retained (accept) or discarded (reject). In the example (green box) UMI1 is unique and thus retained; UMI2 is duplicated but all nucleotides are the same and thus UMI2 is retained; UMI3 is duplicated and the most represented nucleotide is less than four-fold more abundant than the second, thus UMI3 is rejected.

Figure 4. Titration of digital droplet PCR assay for a candidate variant.

A ddPCR assay for a candidate variant is first tested for sensitivity by running appropriate negative and positive controls. Negative controls include a non-template control (NTC) sample and a wild type DNA sample, to check background probe fluorescence and non-specific hybridization, respectively. As positive controls we used a series of wild type DNA samples containing a spiked-in gBLOCK construct containing the candidate variant. We spiked in 3000, 600, 120, 24 and 4.8 molecules of mutant gBLOCK construct per reaction, which corresponds to 10%, 2%, 0.4%, 0.08% and 0.016% of the number of alleles present in the template DNA in each reaction (100 ng of DNA, ~30,000 haploid genome copies). Data were analysed with BioRad QuantaSoft, and we report the calculated number of copies per µL of mutant (a) and wild type allele (b); error bars are the 95% Poisson confidence intervals. The red line highlights the lowest sensitivity attainable, i.e. the upper boundary of the mutant allele count in the wild type sample. Hence, for this specific assay a variant can be called if its lower Poisson confidence interval is above 0.42 copies/µL.

Table captions

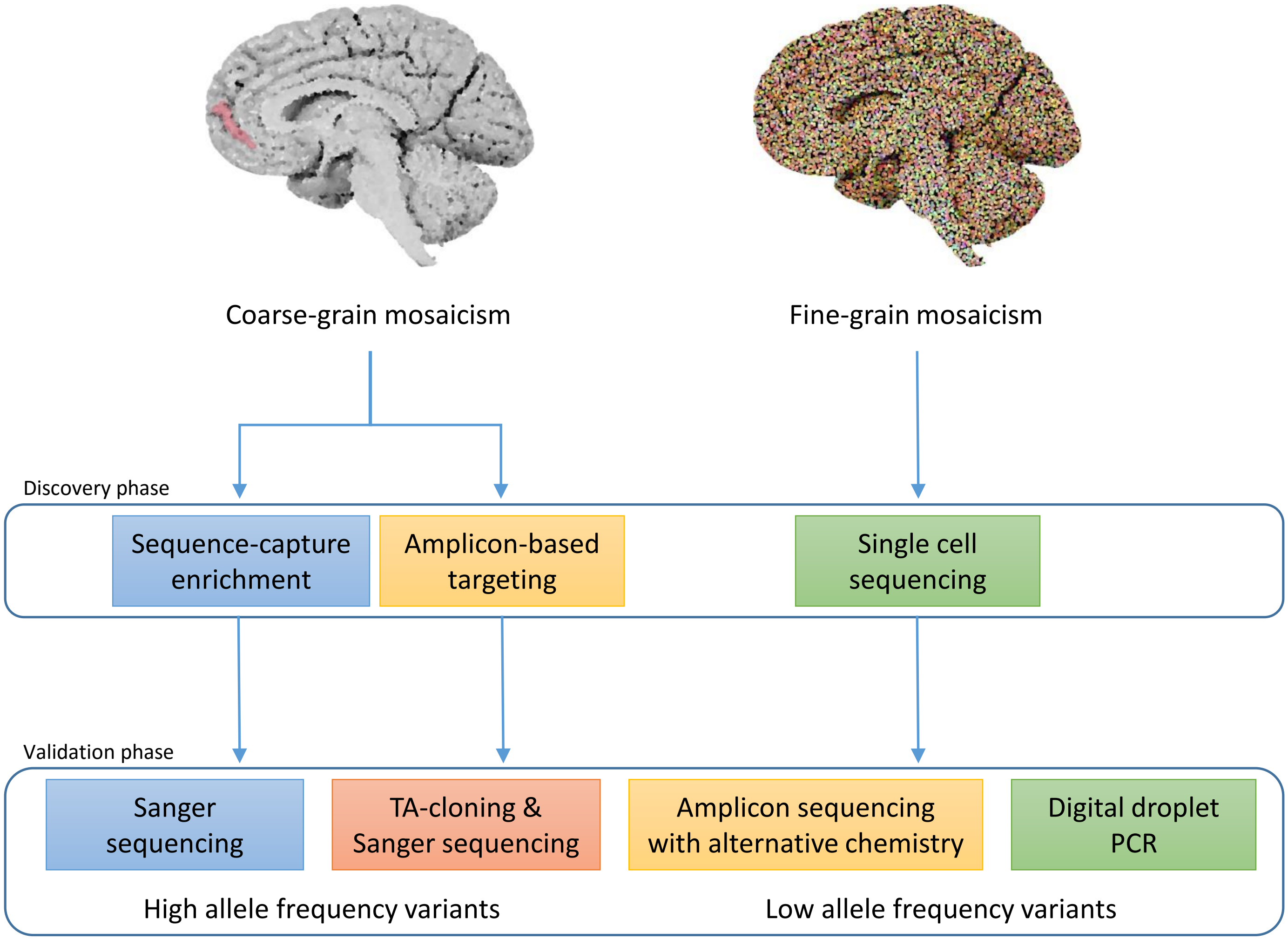
Table. Bioinformatics analysis software for DNA sequencing and variant calling analysis.

In the table we provide a description and links for software mentioned in the methodologies, along with similar software that can be used for DNA sequencing analysis and somatic variant discovery. Links are updated and valid as of December 2016.

Table

Name	Purpose	Link
BWA	alignment of raw reads to reference genome	bio-bwa.sourceforge.net/
SAMtools	handling of aligned reads, pileup of aligned reads	samtools.sourceforge.net/
GATK	handling of aligned reads, error correction, variant calling	https://software.broadinstitute.org/gatk/
VarScan	variant calling	http://dkoboldt.github.io/varscan/
SNVer	variant calling	snver.sourceforge.net/
LoFreq	variant calling	http://csb5.github.io/lofreq/
UMI-tools	handling of unique molecular identifiers	https://github.com/CGATOxford/UMI-tools
SnpEff	variant annotation	snpeff.sourceforge.net/
Annovar	variant annotation	www.openbioinformatics.org/annovar/
Monovar	Variant calling in single cell sequencing data	https://bitbucket.org/hamimzafar/monovar
Single Cell Genotyper	Variant calling in single cell sequencing data	https://bitbucket.org/aroth85/scg/wiki/Home
R	statistical analysis	http://www.r-project.org/

Figure 1



Coarse-grain mosaicism

Fine-grain mosaicism

Discovery phase

Sequence-capture enrichment

Amplicon-based targeting

Single cell sequencing

Validation phase

Sanger sequencing

TA-cloning & Sanger sequencing

Amplicon sequencing with alternative chemistry

Digital droplet PCR

High allele frequency variants

Low allele frequency variants

Figure 2

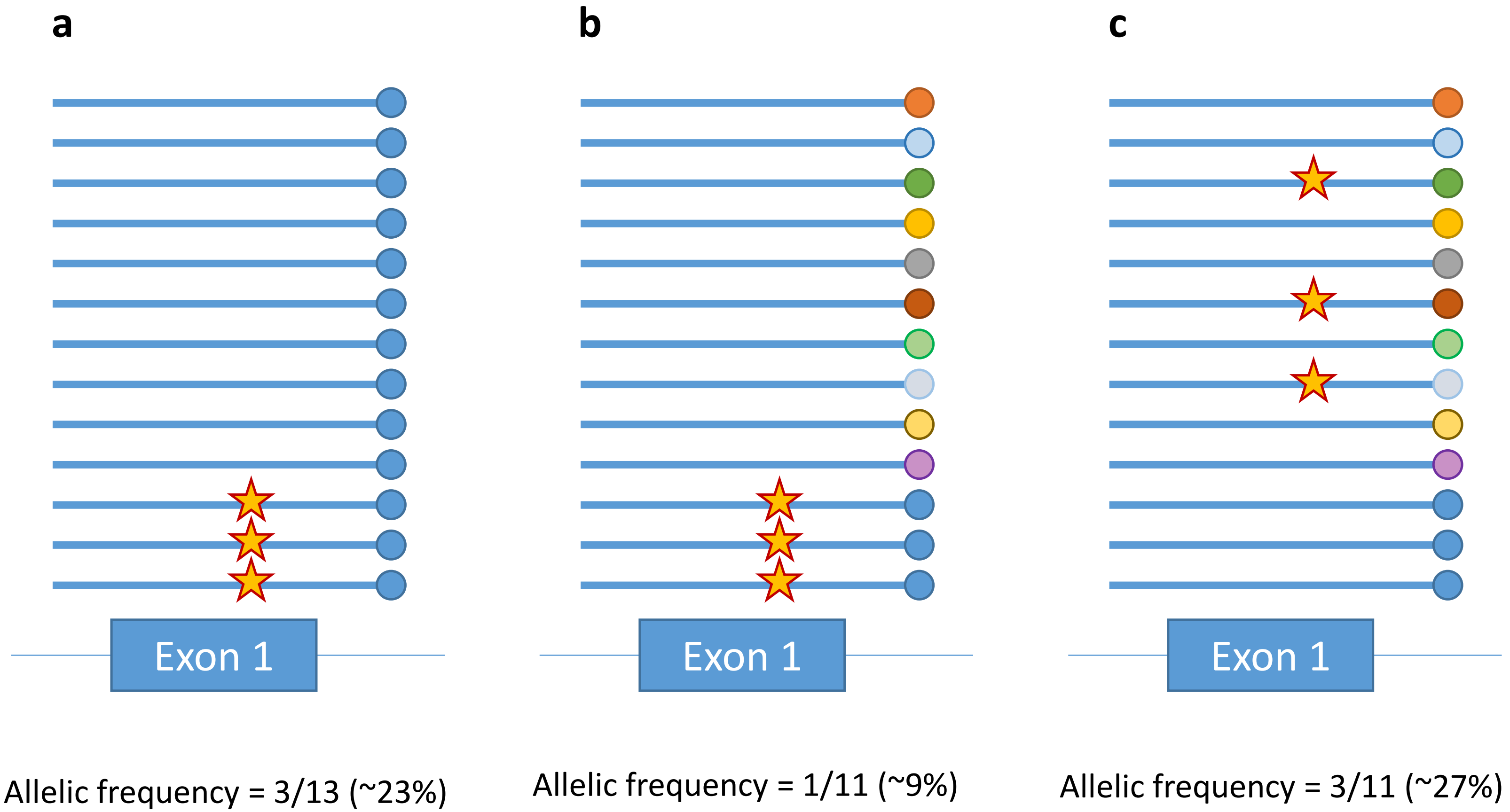

a
b
c
Exon 1
Exon 1
Exon 1
Allelic frequency = 3/13 (~23%)
Allelic frequency = 1/11 (~9%)
Allelic frequency = 3/11 (~27%)

Figure 3

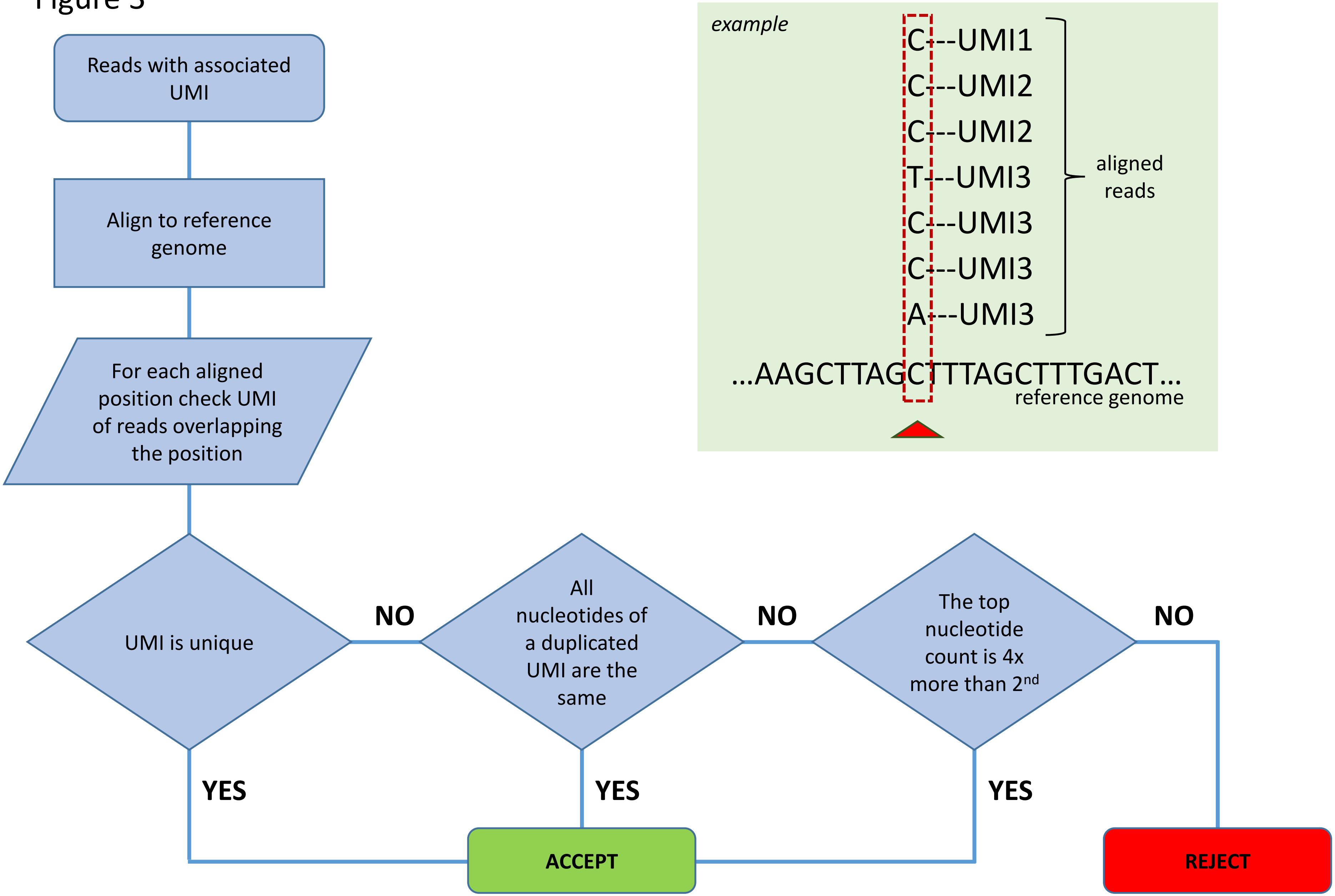
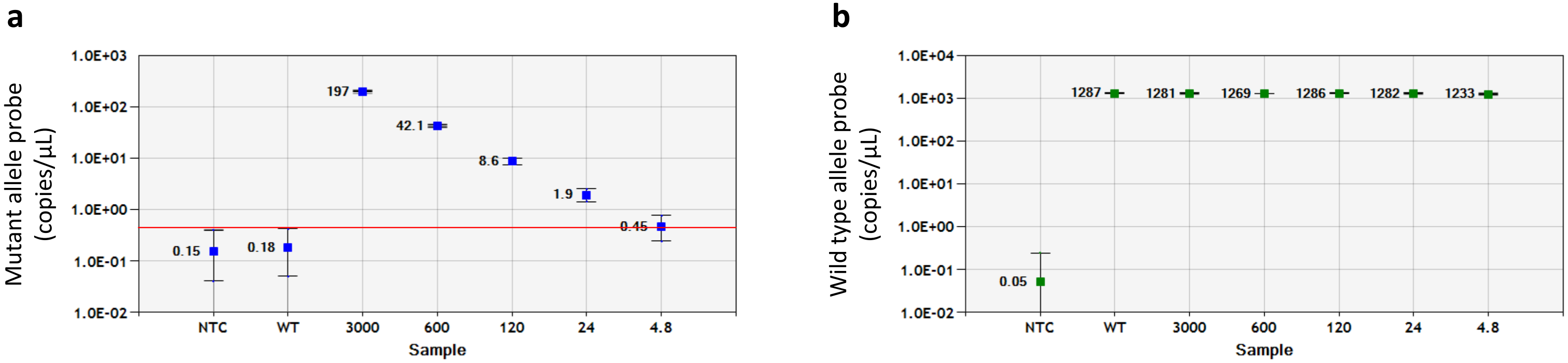

Reads with associated UMI
Align to reference genome
For each aligned position check UMI of reads overlapping the position
UMI is unique
NO
All nucleotides of a duplicated UMI are the same
NO
The top nucleotide count is 4x more than 2nd
NO
YES
YES
YES
ACCEPT
REJECT
example
C---UMI1
C---UMI2
C---UMI2
T---UMI3
C---UMI3
C---UMI3
A---UMI3
aligned reads
...AAGCTTAGCTTTAGCTTTGACT...
reference genome

Figure 4


a
Mutant allele probe (copies/µL)
1.0E+03
1.0E+02
1.0E+01
1.0E+00
1.0E-01
1.0E-02
197
42.1
8.6
1.9
0.45
0.15
0.18
NTC
WT
3000
600
120
24
4.8
Sample
b
Wild type allele probe (copies/µL)
1.0E+04
1.0E+03
1.0E+02
1.0E+01
1.0E+00
1.0E-01
1.0E-02
1287
1281
1269
1286
1282
1233
0.05
NTC
WT
3000
600
120
24
4.8
Sample