International Journal of
Molecular Sciences

MDPI

Article

From Forensics to Clinical Research: Expanding the Variant Calling Pipeline for the Precision ID mtDNA Whole Genome Panel

Filipe Cortes-Figueiredo [1,2], Filipa S. Carvalho [1], Ana Catarina Fonseca [3,4], Friedemann Paul [2,5], José M. Ferro [3,4], Sebastian Schönherr [6], Hansi Weissensteiner [6,*] and Vanessa A. Morais [1,*]

[1] VMorais Lab—Mitochondria Biology & Neurodegeneration, Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina, Universidade de Lisboa, 1649-028 Lisbon, Portugal; filipe.figueiredo@medicina.ulisboa.pt (F.C.-F.); filcarvalho.uc@gmail.com (F.S.C.)

[2] NeuroCure Clinical Research Center, Charité—Universitätsmedizin Berlin, 10117 Berlin, Germany; friedemann.paul@charite.de

[3] José Ferro Lab—Clinical Research in Non-communicable Neurological Diseases, Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina, Universidade de Lisboa, 1649-028 Lisbon, Portugal; acfonseca@medicina.ulisboa.pt (A.C.F.); jmferro@medicina.ulisboa.pt (J.M.F.)

[4] Serviço de Neurologia, Hospital de Santa Maria, Centro Hospitalar Universitário Lisboa Norte, 1649-035 Lisbon, Portugal

[5] Experimental and Clinical Research Center, Charité—Universitätsmedizin Berlin and Max Delbrück Center for Molecular Medicine, 13125 Berlin, Germany

[6] Institute of Genetic Epidemiology, Department of Genetics and Pharmacology, Medical University of Innsbruck, 6020 Innsbruck, Austria; sebastian.schoenherr@i-med.ac.at

* Correspondence: hansi.weissensteiner@i-med.ac.at (H.W.); vmorais@medicina.ulisboa.pt (V.A.M.); Tel.: +43-0512-9003-70560 (H.W.); +351-217-999-573 (V.A.M.)

check for updates

Citation: Cortes-Figueiredo, F.; Carvalho, F.S.; Fonseca, A.C.; Paul, F.; Ferro, J.M.; Schönherr, S.; Weissensteiner, H.; Morais, V.A. From Forensics to Clinical Research: Expanding the Variant Calling Pipeline for the Precision ID mtDNA Whole Genome Panel. *Int. J. Mol. Sci.* **2021**, *22*, 12031. https://doi.org/10.3390/ijms222112031

Academic Editor: Andrey V. Kuznetsov

Received: 7 October 2021
Accepted: 29 October 2021
Published: 6 November 2021

Abstract: Despite a multitude of methods for the sample preparation, sequencing, and data analysis of mitochondrial DNA (mtDNA), the demand for innovation remains, particularly in comparison with nuclear DNA (nDNA) research. The Applied Biosystems™ Precision ID mtDNA Whole Genome Panel (Thermo Fisher Scientific, USA) is an innovative library preparation kit suitable for degraded samples and low DNA input. However, its bioinformatic processing occurs in the enterprise Ion Torrent Suite™ Software (TSS), yielding BAM files aligned to an unorthodox version of the revised Cambridge Reference Sequence (rCRS), with a heteroplasmy threshold level of 10%. Here, we present an alternative customizable pipeline, the PrecisionCallerPipeline (PCP), for processing samples with the correct rCRS output after Ion Torrent sequencing with the Precision ID library kit. Using 18 samples (3 original samples and 15 mixtures) derived from the 1000 Genomes Project, we achieved overall improved performance metrics in comparison with the proprietary TSS, with optimal performance at a 2.5% heteroplasmy threshold. We further validated our findings with 50 samples from an ongoing independent cohort of stroke patients, with PCP finding 98.31% of TSS's variants (TSS found 57.92% of PCP's variants), with a significant correlation between the variant levels of variants found with both pipelines.

Keywords: mitochondrial DNA; next-generation sequencing; massively parallel sequencing; whole genome sequencing; Precision ID; Thermo Fisher Scientific; variant calling; mixture; performance metrics

1. Introduction

Mitochondria are the primary source of cellular ATP, while also prominently contributing to cell survival, differentiation, and apoptosis. They contain their own double-stranded circular DNA (mtDNA), with 16,569 base pairs (bps) in humans, responsible for encoding 22 transfer RNAs, two ribosomal RNAs, and 13 essential proteins of the oxidative phosphorylation (OXPHOS) chain. In comparison with nuclear DNA (nDNA), mtDNA has a higher copy number per cell (polyploidy), replicating independently of the nuclear genome, which allows for the existence of multiple genotypes within the same cell (heteroplasmy). Additionally, mtDNA shows a higher mutation rate due to the absence of protective histones, exposure to reactive oxygen and nitrogen species, and more rudimentary repair systems [1,2].

The peculiar nature of mtDNA has hindered the development of specific data analysis guidelines, since the large majority of datasets, guidelines, and bioinformatic pipelines for variant discovery analysis in next-generation sequencing (NGS)/massively parallel sequencing (MPS) are primarily, and sometimes exclusively, focused on nDNA [3,4], with mtDNA NGS analysis lagging years behind [5]. Out of the existing bioinformatic tools for mtDNA, with quality control assessment, variant calling, and haplogroup assignment [6–21], very few incorporate user accessibility, integrative data analysis, and regular updates. A suitable option is mtDNA-Server [17], which allows for FASTQ and BAM files as input and, considering the revised Cambridge Reference Sequence (rCRS) [22] as a reference, identifies heteroplasmic and homoplasmic variants, assigns a haplogroup with HaploGrep 2 [18], based on PhyloTree Build 17 [23] with an updated algorithm [24], and performs a contamination check and coverage analysis.

Recently, a novel approach to mtDNA NGS through whole genome sequencing (WGS) has been developed with the Applied Biosystems™ Precision ID mtDNA Whole Genome Panel (Thermo Fisher Scientific, USA) [25]. This library preparation kit, with 162 amplicons that target the whole mtDNA, is mostly used for forensic samples, achieving reliable results in often degraded samples [26–29] and with low DNA input (usually 0.1 ng of genomic DNA, but 6.25 pg [30] and, very recently, 0.6 pg [31] have been reported). Despite its most common use in forensic sciences, it has also been used in a range of other applications, from worldwide lineage studies [32], to rare mtDNA differences in monozygotic twins [33].

A disadvantage of this approach, however, is its use of a modified mtDNA reference with 16,649 bps, instead of the conventional 16,569 bps. Since the last amplicon corresponds to the last 28 bps and the first 80 bps in the widely used rCRS, the Precision ID mtDNA reference duplicates the first 80 bps of the rCRS at the end of the reference. Thus, most of its bioinformatic processing must be done opaquely in the enterprise software, Ion Torrent Suite™ Software (TSS), with the possible addition of other company-owned software [31,34–36]. In addition to producing an unorthodox reference [37], which impedes further processing in open-source bioinformatic tools, sensitivity analyses have set its heteroplasmy threshold level to 10% [30,38,39].

In this study, we aimed to establish an alternative fully customizable pipeline—PrecisionCallerPipeline (PCP)—for the Precision ID mtDNA Whole Genome Panel, which (**I**) Produced BAMs accurately mapped to the rCRS, allowing complete integration with existing open-source mtDNA pipelines; and (**II**) Revisited the TSS's heteroplasmy threshold of 10%. Our approach yielded improved performance metrics in comparison with TSS, additionally enabling the detection of heteroplasmic variants at the 2.5% threshold.

On the basis of this extensive validation process, we compared our pipeline to TSS by investigating 50 clinical samples from an ongoing stroke cohort, where we further validated our approach.

2. Results

2.1. Pipeline Validation with Samples from the 1000 Genomes Project

We acquired three DNA samples previously sequenced within the 1000 Genomes Project [4] on Illumina HiSeq (Illumina, Inc., USA), with three different sequencing runs (more details in Section 4.1. Sample acquisition/collection): (**I**) Exome sequencing—exome (mean coverage: 420.59 reads/base pair); (**II**) Low-coverage sequencing—lowCov (mean coverage 3863.37 reads/base pair); and (**III**) High-coverage sequencing—highCov (mean coverage: 15079.28 reads/base pair). Samples were then processed and mixed at five different mixtures levels, ranging from 1 to 25% (Figure 1A), and sequenced on Ion Torrent™ Ion S5™ (Thermo Fisher Scientific, USA). The original sequenced samples and their mixtures then underwent bioinformatic analysis with our pipeline, the PrecisionCallerPipeline (PCP) (Figure 1B). For that, we aimed to optimize the protocol for the heteroplasmy threshold and the removal of nuclear insertions of mitochondrial DNA (NUMTs). Correspondingly, we ran the samples with three different NUMT removal approaches and 21 thresholds, ranging from 0.4% to 10% (Figure 1C), through the mutserve variant caller, an offline and command-line version of mtDNA-Server [17,40]. After data analysis and variant classification for primary and mixture samples (Table S1), we determined the performance metrics (Figure S1) considering both Grade A variants, which were homoplasmic in the three independent sequencing runs, and Grade B variants, which were heteroplasmic in the same three sequencing runs (more details in Section 4.4. Data analysis). Our primary outcome was the highest possible F_1 score.

NUMT removal immediately after the first alignment and before merging the BAM files (Figure 1B) was superior to the other NUMT removal approaches (Table S2), while the 2.5% heteroplasmy threshold with NUMT removal before merging yielded the highest mean F_1 score between the four different datasets—Primary Grade A, Mixture Grade A, Primary Grade B, and Mixture Grade B—(Table S3).

Correspondingly, all BAM files were then run through the mutserve variant caller with a 2.5% threshold and NUMT removal before merging (Figure 1B). For simplicity, we denote samples by their simplified haplogroup identifier. Thus, HG00256 as H, HG01626 as T, and HG01757 as U. Similarly, names for mixtures follow an analogous pattern with the minor component being followed by its absolute mixture level and the major component. Therefore, *U0.01H* is sample HG01757 (U) at 1% mixed with sample HG00256 (H) at 99%.

Genome coverage and mappability, defined as the ability to read a base pair (bp) above a certain coverage threshold, were adequate in comparison to other sequencing runs in the primary analysis. Haplogroup assignment was also homogeneous for both PCP and the output from Ion Torrent Suite™ Software (TSS) (Table S4 and Figure 2).

Interestingly, PCP had an increased number of mutations in comparison with TSS (Figure 2B), despite a reduction in the mean coverage in PCP (Figure 2A). To better understand this phenomenon, we considered all samples analyzed on the Ion Torrent (Table S5 and Figure 3).

In PCP, coverage and mappability were uniform, haplogroup assignment was successful for mixtures up to 10%, and sample contamination was correctly detected, with 5% as the lowest detection limit (Table S5 and Figure S2). TSS correctly identified haplogroups up to 10% as well, albeit with errors in the contamination detection for unmixed samples U and T (lower than TSS's own limit of ~10%), and a higher contamination detection limit of 10% (Table S5).

In comparison with TSS, samples processed with PCP have a reduced total number of sequences (Table S5 and Figure 3A)—mean difference of 16.94%, 95% CI [15.98%–17.89%], an adjusted p-value of 1.43×10^{-17}, and a significantly lower mean coverage (Table S5 and Figure 3B)—mean difference of 25.83%, 95% [24.97%–26.70%], an adjusted p-value of 3.87×10^{-21}; paired t-tests adjusted with false discovery rate (FDR). This arises from our read trimming, NUMT removal, and quality control protocol, since, in comparison to the unaltered FASTQ files, the BAM files from TSS have the exact same number of reads and Phred score patterns (Figures 3D and 3F).

In comparison with PCP, these results indicate that TSS achieves a higher coverage at the expense of a lack of read selection from the initial FASTQ files. Despite having significantly reduced coverage and fewer reads, more variants are called with PCP than with TSS (Table S5 and Figure 3C)—mean difference of 5.83, 95% CI [1.95–9.72], an adjusted p-value of 5.60×10^{-3}; paired t-tests with FDR.

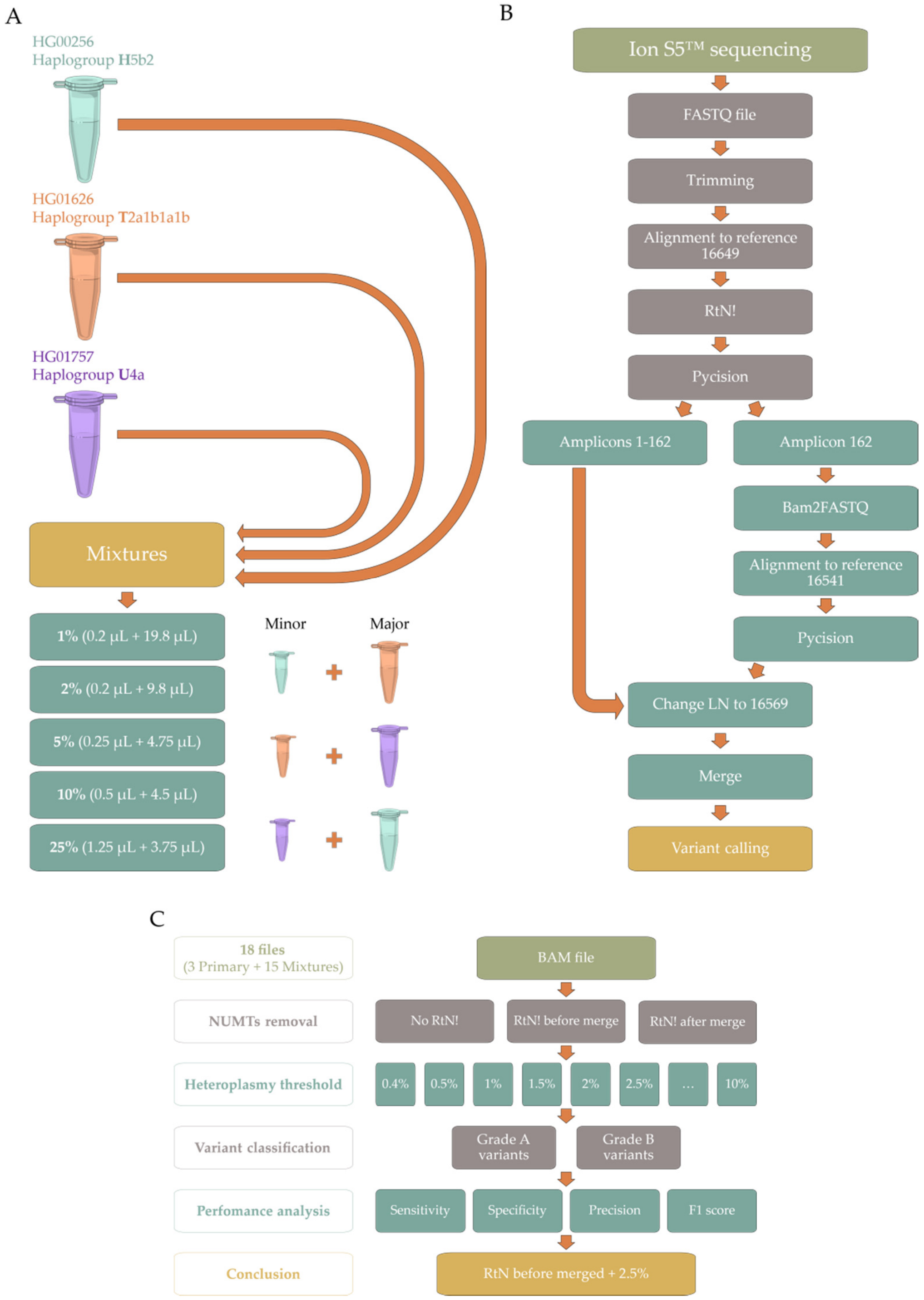

Figure 1. Schematic representation of our workflow. (**A**) Wet lab approach to performing the 15 mixtures from the Coriell Institute for Medical Research's samples, which have been analyzed within the 1000 Genomes Project—templates from Servier Medical Art (CC BY 3.0) were adapted, and are freely available at https://smart.servier.com (accessed on 1 November 2021); (**B**) Visual representation of the PrecisionCallerPipeline (PCP); (**C**) Visual representation of the optimization approach to NUMT removal and heteroplasmy threshold. Abbreviations: NUMTs—nuclear insertions of mitochondrial DNA.

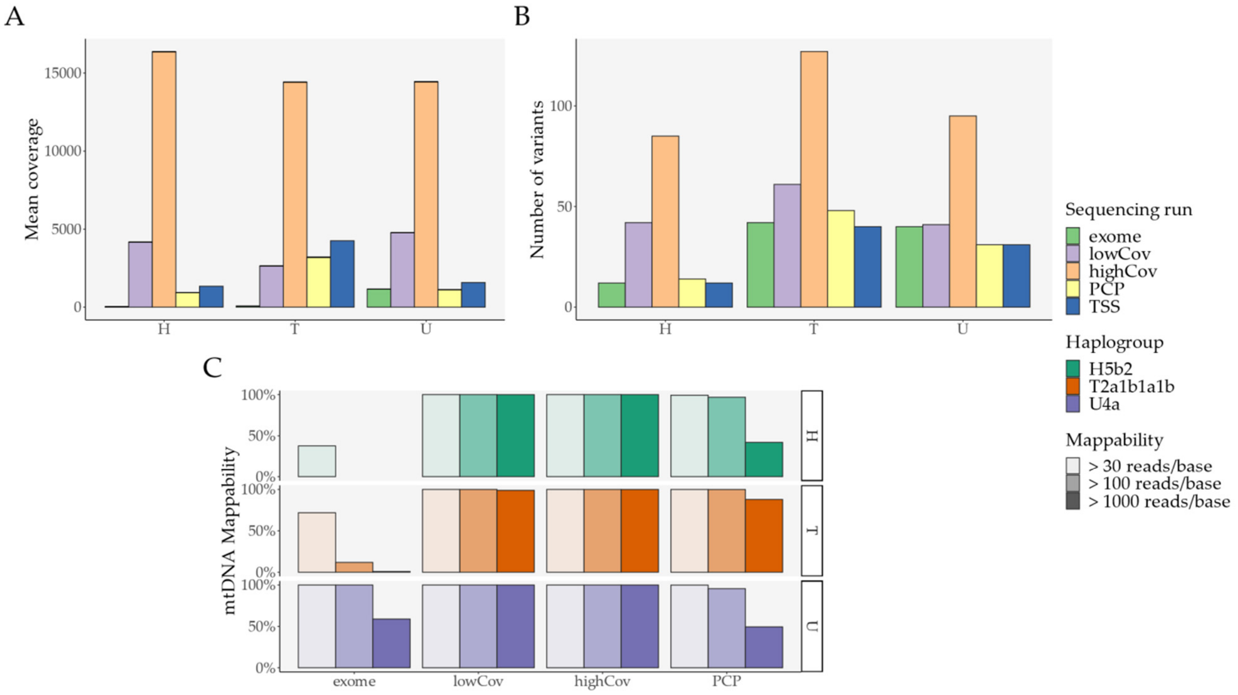

Figure 2. Primary analysis: Overall look; variant threshold was set at 0.4% for exome, lowCov, and highCov, and 2.5% for PCP. (**A**) Mean coverage per sample and per sequencing run/pipeline; (**B**) Number of variants per sequencing run/pipeline; (**C**) Mappability and haplogroup assignment per sequencing run/pipeline (TSS is not shown since it does not output raw data). Error bars denote standard error of the mean (SEM). Abbreviations: PCP—PrecisionCallerPipeline; TSS—Ion Torrent Suite™ Software.

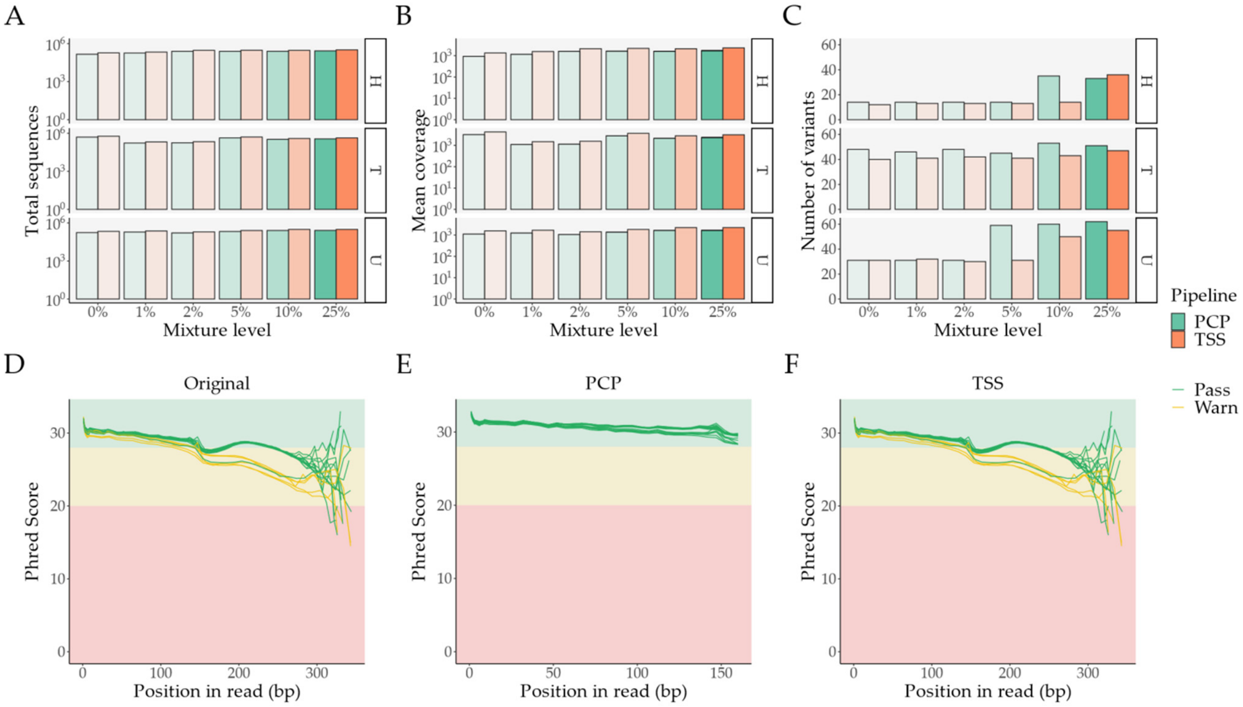

Figure 3. Ion Torrent sequencing: Overall look at differences between PCP and TSS; a 0% mixture level denotes unmixed samples. (**A**) Total number of sequences per pipeline and major sample contributor; (**B**) Mean coverage per pipeline and major sample contributor; (**C**) Number of variants per pipeline and major sample contributor; (**D**–**F**) MultiQC's Phred score per base pair in the unprocessed FASTQ files, PCP's BAM to FASTQ files, and TSS's BAM to FASTQ files, respectively. Abbreviations: bp—base pair; PCP—PrecisionCallerPipeline; TSS—Ion Torrent Suite™ Software.

To further dissect this difference in the number of variants, we first looked at the variants in the primary analysis (Table A1, Figure S3A, and Tables S6–S8).

When comparing variants found with the Ion Torrent sequencing method to the ones previously found on the other three Illumina runs, we observe that: (**I**) PCP significantly increases the proportion of correctly found Grade A/B variants (true positives) in contrast to lost Grade A/B variants (false negatives), while no Grade A/B lost variant with PCP was picked up by TSS (Table S6); (**II**) TSS significantly overestimates the variant level (VL) in found Grade A/B variants in comparison with PCP (Table S7); (**III**) Despite PCP picking up more variants than TSS, the difference in the proportion of "novel" variants (false positives) was not statistically significant between the two pipelines (Table S8).

We then looked at the variants in the mixture analysis (Figure 4, Table A2, Figure S3B–H, Tables S9–S14, Figure S4, Tables S15–S22). Samples with a macro classification of "Mixed status" (more details in Table S1) were excluded because of the impossibility of determining their origin.

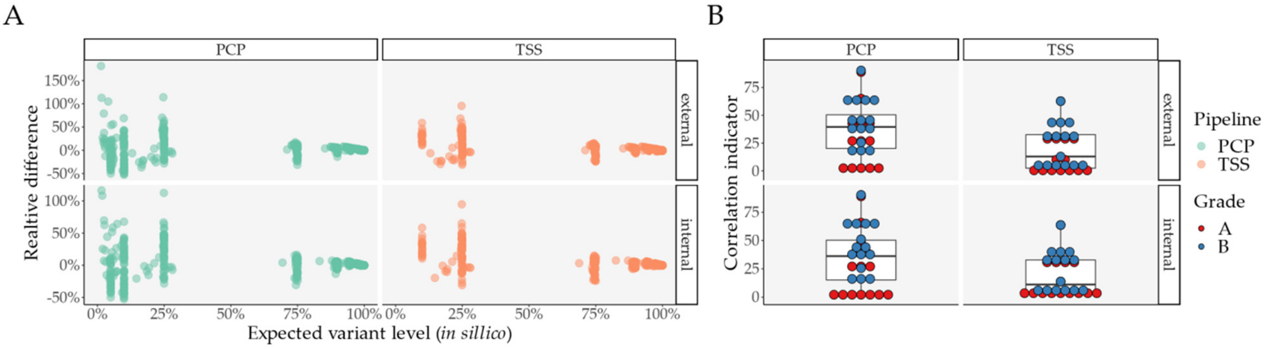

Figure 4. Mixture analysis: similarity in Grade A/B variants between expected and observed variant levels, for internal and external comparisons, per pipeline. (**A**) Relative difference between expected and observed variant level, per pipeline, and per internal and external comparison; (**B**) Correlation indicator for internal and external comparisons, per pipeline, and per variant grade. Abbreviations: PCP—PrecisionCallerPipeline; TSS—Ion Torrent Suite™ Software.

Similar to the primary analysis, Grade A/B variants were best retrieved with PCP, in comparison with TSS, with a single variant found by TSS and lost with PCP (Table S9). Moreover, true positive variants below 11% are completely absent in TSS, while PCP correctly picks up variants until its threshold of 2.5% (Figure 4 and Table A2).

Regarding the difference between the expected and the observed VL, we defined two different approaches:

- External comparison—the difference to the other Illumina sequencing methods (exome, lowCov, and highCov);
- Internal comparison—the difference within Ion Torrent, taking the primary sequencing as a reference.

Consequently, we also considered two different relative differences, where we divided each difference by its expected level in order to observe a percent variance per VL (Figure 4A, Figure S3C–H). When performing a paired comparison regarding differences for Grade A/B variants (Table S10), we observed no significant differences between TSS and PCP for raw VL differences taking found variants into consideration. However, when we considered lost variants, TSS performed worse in the external comparison (Table S10). Regarding relative differences (Table S11), we found no statistically significant differences between PCP and TSS.

In order to assess the similarity between the expected and the observed VLs (internal and external), we also performed linear models per mixture, considering Grade A/B variants for each pipeline, for each comparison (internal and external), and for each grade variant classification (A or B), as they might have varied between them. After extracting the adjusted R^2 and the adjusted p-value with FDR, we created an indicator for correlation strength determined by the log10 of the ratio between the adjusted R2 and the adjusted *p*-value. Thus, a greater value would indicate a stronger correlation between both variables. PCP showed a significantly increased correlation indicator in both comparisons (internal and external) for Grade B variants, while no significant differences for Grade A variants were found (Figure 4B and Table S12).

Since PCP appeared to have increased sensitivity in comparison with TSS, we then looked at the prevalence of novel variants (false positives). Interestingly, no differences were found (Table S13), albeit with TSS showing novel variants at a higher VL, namely, novel variants already present in the primary analysis (Table S14).

When we looked at the location of these novel variants in both analyses (primary and mixture), PCP had one mutation in 10/18 samples (4318T), one mutation in 8/18 samples (8649C), while the remaining four were isolated (2463G, 5752G, 6698G, 8152A). TSS had two mutations shared in 10/18 samples (310C, 10958C), one shared in 4/18 samples (14777C), one shared in 3/18 samples (8249C), and one shared in 2/18 samples (318C).

Looking into the different variant classifications (found, lost, novel, and primary novel variants), we noticed that: (**I**) PCP showed an increased rate (adjusted to the size of each region) of found variants throughout the genome, while maintaining a lower rate of lost variants (Figure S4A–B and Tables S15 and S16); (**II**) Novel variants (primary and mixture) showed the lowest normalized coverage (normalized for the mean coverage of each sample in Tables S4–S5) for both PCP and TSS, while the highest strand bias, calculated as a coverage ratio (forward vs. reverse), was observed in mixture novel variants for PCP, and in primary novel variants for TSS (Figure S4C–E and Tables S17 and S18); (**III**) Since false variants might arise from NUMTs, we calculated the mean value of reported NUMTs based on two different databases [17,41] for each position—the highest rate of NUMTs was observed in primary novel variants for PCP, and in found variants for TSS (Figure S4F and Table S19); (**IV**) Regarding the proportion of variants in positions classified as a low-complexity region (LCR) [17], lost variants showed the highest proportion of LCR variants for PCP, while the same was true in primary novel variants for TSS (Figure S4G and Table S20); (**IV**) Finally, we tested the hypothesis that the distance to the "callable" extremity of each amplicon might be significantly different, depending on each variant class. The lowest distance was observed in primary and mixture novel variants for PCP, while the same was true for primary novel variants alone for TSS, although less consistently (Figure S4H and Table S21). Table 1 offers a summary for novel variants (false positives).

Table 1. Novel variants (false positives): Summary characteristics in the mutserve variant calling method (VCM).

Variable	PCP		TSS	
	Primary Novel Variants	**Mixture Novel Variants**	**Primary Novel Variants**	**Mixture Novel Variants**
Normalized coverage	↓	↓↓	↓↓↓	↓↓
Coverage ratio	-	↑↑↑	↑↑↑	↑
Number of NUMTs	↑↑↑	↓	-	↓↓
LCR prevalence	↓↓↓	↓↓↓	↑↑↑	-
Distance to amplicon edge	↓↓↓	↓↓	↓	-

Using found variants (true positives) as a reference: "-" denotes nonsignificant changes; "↑" or "↓" denote significant changes in the 25–50% range; "↑↑" or "↓↓" denote significant changes in the 51–75% range; and "↑↑↑" or "↓↓↓" denote significant changes > 75% (more details in Tables S17–S21). Abbreviations: PCP—PrecisionCallerPipeline; TSS—Ion Torrent Suite™ Software; NUMTs—nuclear insertions of mitochondrial DNA; LCR—low-complexity region.

Ultimately, we performed a performance analysis for four different datasets: Primary Grade A, Mixture Grade A, Primary Grade B, and Mixture Grade B, with separate paired statistical tests for each dataset (Figure A1 and Table S22). PCP achieved a higher sensitivity and F_1 score in both the primary and mixture analyses for Grade B variants, without performing worse than TSS in the other indicators or in the other datasets (Table S22).

In order to further validate our findings, we performed the same analyses with two other variant callers (freebayes [42] and varscan 2 [43]), which yielded overall overlapping results (Tables S6–S24).

2.2. Pipeline Comparison with An Independent Set of 50 Clinical Samples

As a way of demonstrating the added value of our pipeline, we selected 50 independent samples from a prospective study in patients with ischemic stroke (more details in Section 4.1. Sample acquisition/collection) and analyzed them in parallel with TSS and PCP.

In PCP, genome coverage followed a similar pattern in all samples (Figure 5A), with an overall uniform mappability (Table S25). The genome coverage was also in accordance with what is typical in Ion Torrent™ sequencing with the Precision ID mtDNA Whole Genome Panel [35]. The haplogroup was identical for both pipelines in 45/50 samples (Table S25). In PCP, 3 samples were flagged for contamination in contrast to TSS, which had 11 samples, 10/11 were well below its ~10% threshold (Table S25). In PCP, the number of variants varied extremely, from 12 to 503, while in TSS, they ranged from 10 to 80 (Table S25). Thus, we eliminated Samples #21, #37, and #44 because of their contamination status, and Samples #2 and #3 because of their extremely high number of variants in PCP, from our subsequent analysis.

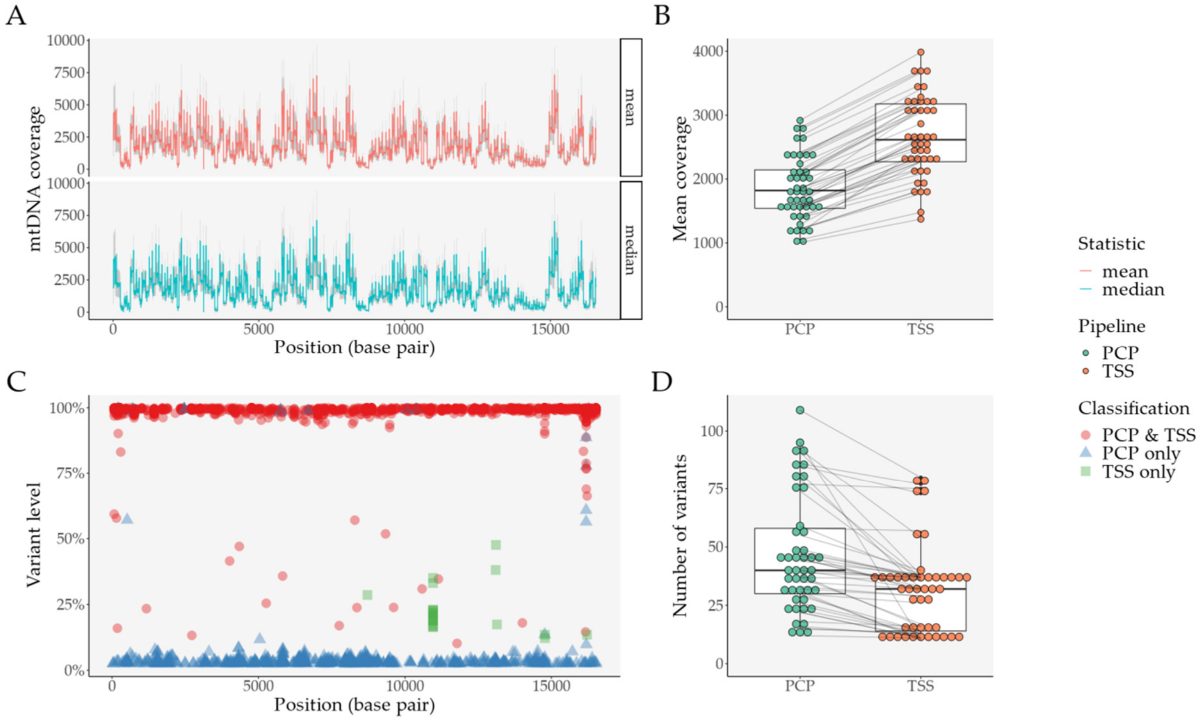

Figure 5. Comparison between PCP and TSS in an independent set of samples (after filtering, N = 45) derived from clinical practice (an ischemic stroke prospective cohort). (**A**) Coverage per base pair with PCP data, per statistic (mean and median)—the grey area represents the standard deviation of each statistic per base pair; (**B**) Mean coverage per sample and per pipeline—lines connect the same sample; (**C**) Distribution of variants per base pair and per variant classification—for variants found in PCP and TSS, mean variant level was plotted; (**D**) Mean number of variants per sample and per pipeline—lines connect the same sample. Abbreviations: PCP—PrecisionCallerPipeline; TSS—Ion Torrent Suite™ Software.

Consistent with what we described previously, while mean coverage was significantly decreased in PCP (Figure 5B)—mean difference of 30.87%, 95% CI [29.17%–32.58%], an adjusted *p*-value of 3.01×10^{-34}, we still observed a higher number of variants with our pipeline (Figure 5D)—mean difference of 15.11%, 95% [10.07–20.16], an adjusted *p*-value of 2.97×10^{-7}; paired t-tests adjusted with FDR.

This difference mostly derives from an increase in variants below TSS's heteroplasmy threshold (Figure 5C): variants only present in PCP have a mean VL of 7.45%, while variants only present in TSS have a mean VL of 22.5%. Interestingly, PCP finds 98.31% of TSS's variants, while TSS only finds 57.92% of PCP's variants. Although not exclusively, when we consider discordant variants (variants not present in both pipelines) with more than one occurrence (Table S26), we observe similar patterns as false positives in the previous analyses (Figure S4C–H and Tables S17–21). In PCP, for example, we encounter the same variants seen in the previous analysis: 8649C (present in 25/45 samples) and 4318T (present in 20/45 samples). These two samples have a mean VL <5%, a low mean normalized coverage, and a small distance to the amplicon's "callable" edge. Similarly, in TSS, we observe variants that were also previously found: 10958C (present in 17/45 samples) and 14777C (present in 2/45 samples). These two variants have a low mean normalized coverage, without other discernible features.

Regardless of the discordant variants, in the variants found in both pipelines, there is a statistically significant correlation between the VL with TSS, and with PCP (Figure S5)—adjusted R^2 of 0.97 and a *p*-value $< 2.2 \times 10^{-16}$; linear regression model.

3. Discussion and Conclusions

In this study, we have developed a fully customizable pipeline—PCP—for the Precision ID mtDNA Whole Genome Panel, which produced BAMs accurately mapped to the rCRS and that has revisited the heteroplasmy threshold of the current gold standard method—TSS.

On the one hand, despite arising from the same set of samples and sequencing runs, the output from PCP and TSS was sufficiently different to yield significant changes between the two pipelines. In comparison with TSS, our method accomplished better performance metrics in previously sequenced samples, namely, higher sensitivity and F_1 scores without a decrease in specificity and precision. This was mostly due to the additional detection of heteroplasmic variants with a VL above 2.5%, but below TSS's ~10% threshold. Interestingly, this increase in the number of correct variants and the improved performance was achieved despite PCP having lower coverage and fewer sequences. Moreover, the same pattern of a higher number of variants at a lower threshold, notwithstanding a lower mean coverage and fewer sequences, was also observed in an independent set of clinical samples.

On the other hand, mtDNA differences between the same sample are not unheard of, as DNA polymerases, amplification protocols, sequencing runs, and variant callers are frequent sources of disparity in genomics [35,44]. Discerning sequencing errors/false positives from true heteroplasmic variants is a challenging task, usually achieved through post-sequencing curation, looking for signs of poor amplification, strand bias, mutations in LCR, and the presence of NUMTs [41,45–49].

Although sometimes ignored [50], heteroplasmic variants are ubiquitous [51–54] and show significant tissue-specificity [53,54]. Interestingly, low-level heteroplasmic variants (VL <10%) have shown matrilineal inheritance [55,56] with increased relevance in aging [57], and in clinical settings, particularly cancer [58,59]. Nonetheless, the reliability of these low-level variants has been a matter of debate [60–62] and, thus, current guidelines/workflows suggest a heteroplasmy threshold of 10%, based on the limitations of Sanger sequencing and electrophoresis technology [35,37,63–66].

False positives in both PCP and TSS showed a decreased normalized coverage, which indicates poor amplification in those regions, and were mostly transitions, which is in accordance with the literature [66,67]. However, the remaining putative mechanisms of error were very different between the two pipelines, and even showed differences within the same pipeline, depending on where the novel variants were found (unmixed vs. mixed samples). Correspondingly, in the analysis of the 1000 Genomes Project's samples, no false positives were shared.

In parallel, discordant variants in the stroke cohort showed most of the variants previously flagged as false positives, albeit with many more variants only present in PCP at a low level, which had not been flagged before. This different mutational pattern from the samples sequenced within the 1000 Genomes Project might arise from the tissue-specificity of heteroplasmic variants, since blood, from where we extracted the DNA in our stroke cohort, has a lower level of heteroplasmy [53,54] and a very diverse mixture of mtDNA content [68], as it encompasses multiple cell subtypes.

Our pipeline offers multiple advantages in comparison with TSS. Firstly, it is in line with the Findable, Accessible, Interoperable, Reusable (FAIR) principles [69], increasingly important in the field of genomics [70]. Thus, our open-access approach makes it easier to implement validated and uniform bioinformatic protocols in genomics [3,71,72]. Secondly, it outputs a BAM file with the correct rCRS reference, thus allowing for integration into other NGS workflows and greatly facilitating variant annotation. This is particularly important when dealing with high-throughput datasets, where manual curation and inspection are not feasible nor efficient, thus favoring computational approaches to DNA variant analysis [73]. Finally, it is fully customizable. In a world of limited resources (personnel, time, samples, funding), our workflow is easily adaptable, depending on the focus of each research project and tissue mutational pattern. On the one hand, if one wishes to prioritize sample diversity in favor of replicates, variants might be called with PCP and filtered if not present in TSS, since PCP captures ~98% of all TSS variants with a very high VL correlation, and TSS discordant variants are likely to be false positives, in accordance with our analysis. On the other hand, if one prefers to sacrifice sample diversity in favor of VCM performance, the protocol for NUMT removal and heteroplasmy threshold optimization in unmixed and mixed samples yields an optimal workflow specific to that set of samples. Furthermore, the two approaches might be combined, or even expanded, into FASTQ quality control, alignment, or further variant annotation, with the addition of hypervariable segments [74], poly C-stretches [75], which are particularly difficult in Ion Torrrent™ sequencing [76], and the presence of homoplasmic or heteroplasmic variants at the HelixMTdb [67] (see Table S26 for an example). In summary, future clinical research will benefit from using our open-source bioinformatic processing since it keeps the advantages of the Precision ID library kit, particularly its low DNA input, while circumventing the limitations of TSS, namely, its modified reference sequence, proprietary nature, and 10% heteroplasmy threshold. For separating true variants from false variants, we also present a range of options that can be used in combination or not, depending on the samples and research focus: (**I**) A dedicated mixture optimization protocol; (**II**) Variant filtering based on normalized coverage, strand bias, the presence of NUMTs, among other parameters; (**III**) Filtering the output from PCP with TSS.

Nonetheless, our study also has a few limitations. Firstly, we used a limited set of samples for the primary and mixture analysis, which are not representative of the range of possible haplogroup combinations. Secondly, samples from our stroke cohort were not sequenced in duplicate nor mixed, which did not allow for a replication of the performance metrics we did previously. Finally, we did not consider insertions nor deletions and, thus, we are unable to provide any recommendations for the analysis of those variants with our pipeline.

Overall, we have developed the first open-source alternative to the enterprise software, Ion Torrent Suite™ Software (TSS), for the Precision ID mtDNA Whole Genome Panel, with improved performance metrics, and with an output in the correct rCRS reference. Since the majority of existing mtDNA bioinformatic tools [6–21] are only compatible with the correct rCRS format, and the few tools that perform variant calling in samples sequenced with the Precision ID library kit [31,34–36] keep TSS's modified reference sequence, PCP is currently the sole option that bridges this bioinformatic gap, allowing for the universal variant calling of samples sequenced with the aforementioned library.

The herein presented pipeline, PCP, is available in the GitHub repository, https://github.com/filcfig/PCP (accessed on 1 November 2021). Additionally, we provide the generated data for validation (see Data Availability Statement).

4. Materials and Methods

4.1. Sample Acquisition/Collection

In order to validate our pipeline—PrecisionCallerPipeline *(PCP)*, we acquired three cell-line samples from the Coriell Institute for Medical Research (Camden, NJ, USA), i.e., HG00256, HG01626, and HG01757. These samples have been sequenced within the 1000 Genomes Project [4] on Illumina HiSeq (Illumina, Inc., San Diego, CA, USA), with three different sequencing runs, mostly targeted at nuclear DNA: (**I**) Exome sequencing—exome; (**II**) Low-coverage sequencing—lowCov; and (**III**) High coverage sequencing—highCov.

With the aim of evaluating the sensitivity, specificity, precision, and F_1 score (harmonic mean of precision and sensitivity) [17,44] with different variant detection thresholds, we performed three different mixtures at five different mixture levels (1, 2, 5, 10, and 25%):

- Mixture HT—Minor component: HG00256 (haplogroup H5b2, short identifier H) + Major component: HG01626 (haplogroup T2a1b1a1b, short identifier T);
- Mixture TU—Minor component: HG01626 (haplogroup T2a1b1a1b, short identifier T) + Major component: HG01757 (haplogroup U4a, short identifier U);
- Mixture UH—Minor component: HG01757 (haplogroup U4a, short identifier U) + Major component: HG00256 (haplogroup H5b2, short identifier H).

DNA concentration measured prior to shipping ranged from 307 to 331 ng/µL. Thus, DNA mixtures were volume-based (Figure 1A).

As an independent set of samples, we analyzed 50 samples derived from a prospective stroke cohort of patients at the Hospital de Santa Maria, Centro Hospitalar Universitário Lisboa Norte. Blood samples were collected within 72 h of hospital admission and all cases were reviewed and confirmed by trained neurologists. Inclusion criteria were: (**I**) Ischemic stroke; (**II**) Age $\geq$ 18 years old; (**III**) Blood samples collected up to 72h after symptom onset. The exclusion criteria were: (**I**) Active cancer diagnosis; (**II**) Previous cerebral revascularization surgeries; (**III**) Modified Rankin score [77] $\geq$ 5. The approval of the institutional review board (IRB) was conceded by the Comissão de Ética do Centro Académico de Medicina de Lisboa (reference 435/16, approved on 14 December 2016), informed consents were given by every subject, and the study followed the standards of the Declaration of Helsinki. DNA extraction was performed after PBMC isolation with the QIAamp® DNA Blood Midi Kit (QIAGEN GmbH, Hilden, Germany), according to the manufacturer's instructions.

4.2. DNA Sequencing

Samples were sequenced without prior long-range PCR (LR-PCR) with the Applied Biosystems™ Precision ID mtDNA Whole Genome Panel (Thermo Fisher Scientific, Waltham, MA, USA), in conjunction with the Ion Torrent™ Ion S5™ (Thermo Fisher Scientific, Waltham, MA, USA), in accordance with the manufacturer's instructions. Briefly, DNA was quantified with a Qubit® 3.0 fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) and samples were diluted to 0.0067 ng/µL for an input of 0.1 ng of genomic DNA in 15 µL. Libraries were prepared using the Ion Chef™ automated protocol, and samples were then run on 530™ chips with the Ion Torrent™ Ion S5™ at Ipatimup—Instituto de Patologia e Imunologia Molecular da Universidade do Porto (Porto, Portugal).

4.3. Bioinformatic Processing

4.3.1. PrecisionCallerPipeline (PCP)

Our PCP pipeline automatically takes the FASTQ files from the sequencing facility and outputs fully aligned BAM files mapped to the commonly used reference sequence, rCRS [22]. We use a workflow based on Snakemake [78] that uses: (**I**) Awk, for SAM file editing [79]; (**II**) BEDTools, for BAM to FASTQ conversion [80]; (**III**) BWA-MEM, for read alignment [81]; (**IV**) Pycision, for amplicon delimitation and selection [34]; (**V**) SAMtools for BAM conversion, sorting, indexing, and merging [82]; and (**VI**) Trimmomatic for read quality control and trimming [83] (Figure 1B). Removal of NUMTs was tested before and after final BAM merging with RtN! [47] (Figure 1C).

Samples were processed in a Linux-based system. For simplicity, a predetermined file structure and the necessary files (except for the files from external software) may be downloaded from https://github.com/filcfig/PCP.git (accessed on 1 November 2021). A separate Snakefile (Snakefile_no RtN) is also provided to run the samples without the removal of NUMTs.

Read quality analysis was performed with FastQC [84] and MultiQC [85]. For read quality control, we opted to crop reads at 160 base pairs (bps), taking into account: (**I**) The visual inspection of Phred score patterns per read bp (Figure 3D); (**II**) The length range of 73–137 bps for the known "callable" sections of each amplicon [34], also considering the company-reported amplicon average length of 163 bps, as the exact coordinates of the amplicon themselves are unknown.

Variant calling was performed with freebayes v1.3.5 [42], mutserve v2, the command line interface and successor of the mtDNA-Server pipeline [17,40], and VarScan 2 v2.3.7 [43]. Initially, we used 21 different heteroplasmy thresholds, ranging from 0.4% to 10.0% (Figure 1C and Figure S1). After optimization, the maximum F_1 score was achieved with a heteroplasmy threshold of 2.5% and with NUMT removal before the final BAM merge (Figure 1B, Tables S2 and S3); this processing was then used in all variant calling methods (VCMs). For the Illumina sequencing runs (exome, lowCov, and highCov), a threshold of 0.4% was maintained as a reference. We used very similar parameters as [44], with the exception of a base quality score of 20 for all VCMs, as well as, for freebayes, where we used *"–ploidy 1 –pooled-continuous"*. Similar to [44], variants in positions 302–315 (position 310 was blacklisted in our analysis), 523–524, and 3104–3110 were excluded. Only single nucleotide substitutions were considered, and variants below the established threshold were filtered.

Haplogroup calling was carried out through HaploGrep v2.4.0 [18], and a contamination check was done with Haplocheck v1.3.3 [40], based on the output from mutserve.

4.3.2. Ion Torrent Suite™ Software (TSS)

For the current gold standard, data from each run was processed using the Ion Torrent™ specific pipeline software, Ion Torrent Suite™ Software (TSS), using the reference sequence PrecisionID_mtDNA_rCRS, and target regions PrecisionID_mtDNA_WG_targets with the plugins CoverageAnalysis and VariantCaller. FASTQ and BAM files were generated using the plugin FileExporter. The software versions ranged from v5.8, v5.10, and v5.12, according to the date of each run.

We compiled the VCF files arising from the sequencing runs and corrected all positions > 16,569 to the first 80 bps in the rCRS. When we observed equal variants with different coverages and variant levels (VLs), particularly in the first 80 bps, we calculated the mean coverage and VL per mutation. Only single nucleotide substitutions were considered.

After exporting the corrected variants in a VCF format, where GT 1/0 was defined for VL $\geq$ 90%, we ran the samples through HaploGrep and Haplocheck, similar to PCP.

4.4. Data Analysis

Data analysis was performed with R version 4.1.1 [86] in RStudio [87] with the packages extrafont [88], infer [89], magick [90], patchwork [91], readxl [92], remotes [93], scales [94], svglite [95], and tidyverse [96], as well as Excel 2016 (Microsoft Corporation, Redmond, WA, USA).

Analyses were performed in parallel for PCP's output and TSS's output. For the primary analysis, variants from the resequenced unmixed samples were compared to the ones identified in Illumina for the exome, lowCov, and highCov sequencing runs, and classified according to their reliability:

1. Grade A variants: homoplasmic variants (mean variant level $\geq$ 95%) found in both highCov and lowCov, regardless of exome;
2. Grade B variants: heteroplasmic variants (mean variant level $\geq$ 0.4% and $\leq$ 95%) found in both highCov and lowCov, regardless of exome, or found in highCov plus exome, or lowCov plus exome;
3. Grade C variants: found in a single sequencing run;
4. Novel variants: found in the Ion Torrent runs only.

For the mixture analysis, variants from the 15 mixtures were classified according to the interaction between both VL and primary variant classifications (explained previously) in both components (minor and major). We began with a manual curation of all 512 theoretical combinations, which gave rise to 128 possible scenarios, grouped in 47 micro classes, 7 meso classes, and 5 macro classes (Table S1). In this case, due to the possibility of shared variants—variants in the same position in two different samples—having two different classifications, we assumed that lower grade variants would prevail in combinations. Hence, a shared variant with both Grade A and Grade C variant classifications would receive a Grade C classification.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/ijms222112031/s1.

Author Contributions: Conceptualization, F.C.-F., H.W. and V.A.M.; formal analysis, F.C.-F., H.W., and V.A.M.; funding acquisition, F.P. and V.A.M.; investigation, F.C.-F., F.S.C. and A.C.F.; methodology, F.C.-F., S.S., H.W. and V.A.M.; resources, A.C.F., J.M.F. and V.A.M.; software, F.C.-F., S.S. and H.W.; supervision, H.W. and V.A.M.; visualization, F.C.-F. and H.W.; writing—original draft preparation, F.C.-F., H.W. and V.A.M.; writing—review and editing, F.C.-F., F.S.C., A.C.F., F.P., J.M.F., S.S., H.W. and V.A.M.. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Fundação para a Ciência e Tecnologia (FCT) (FCT/PTDC/MED-NEU/7976/2020), and a project cofunded by FEDER (POR Lisboa 2020—Programa Operacional Regional de Lisboa PORTUGAL 2020), and FCT (PAC-PRECISE LISBOA-01-0145-FEDER-016394), and the National Multiple Sclerosis Society (NMSS), NMSS Pilot Research Grant (PP-1712-29466). F.C.F.'s stipend was supported by FCT (PD/BD/114122/2015), and by Merck Germany (restricted research grant). V.A.M. is an iFCT researcher (IF/01693/2014; IMM/CT/27-2020).

Institutional Review Board Statement: Sample collection within the 1000 Genomes Project was performed with an informed consent by every subject, which explicitly outlined the creation of cell lines from the patients' samples. The distribution of genetic material was carried out by the Coriell Institute. Regarding the independent set of samples from stroke patients, the study was conducted according to the guidelines of the Declaration of Helsinki, and it was approved by the Ethics Committee of the Centro Académico de Medicina de Lisboa (reference 435/16, approved on December 14, 2016).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study, namely, the data generated with the samples previously sequenced within the 1000 Genomes Project, are openly available in Zenodo [97] at doi:10.5281/zenodo.5524539. Data analyzed within the stroke cohort are available upon request from the corresponding author; the data are not publicly available because of ethical and privacy restrictions.

Acknowledgments: We would like to thank the staff at Biobanco-iMM CAML for training and sample handling, namely, Ângela Afonso, nurse Gislaine Castilho de Oliveira for sample collection and transportation; and the staff at Ipatimup—Instituto de Patologia e Imunologia Molecular da Universidade do Porto (Porto, Portugal) for DNA sequencing, namely, Ana Mafalda Rocha. Finally, we would like to thank all the subjects who have altruistically donated their samples for this research project.

Conflicts of Interest: F.P. is a member of the Novartis OCTIMS study steering committee and the MedImmune/Viela Bio steering committee; has received speaker honoraria/travel grants from Bayer, Novartis, Biogen Idec, Teva, Sanofi-Aventis/Genzyme, Merck Serono, Alexion, Chugai, MedImmune, Shire, Roche, Actelion, and Celgene and; is an Academic Editor of PLoS ONE, and an Associate Editor of *Neurology: Neuroimmunology & Neuroinflammation*; has provided consultancies for SanofiGenzyme, BiogenIdec, MedImmune, Shire, and Alexion; and has received research support from Bayer, Novartis, Biogen Idec, Teva, Sanofi-Aventis/Genzyme, Alexion, Merck Serono, the German Research Council (DFG Exc 257), Werth Stiftung of the City of Cologne, the German Ministry of Education and Research (BMBF Competence Network Multiple Sclerosis), Arthur Arnstein Stiftung Berlin, the EU FP7 Framework Program (combims.eu), the Arthur Arnstein Foundation Berlin, the Guthy Jackson Charitable Foundation, and the National Multiple Sclerosis Society of the USA. The remaining authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A

Table A1. Primary analysis: Distribution of variants with the mutserve variant calling method (VCM).

Variant Description		**PCP**				**TSS**			
Broad Classification	**Detailed Classification**	*N*	**Mean Observed VL**	**Mean Primary VL**	**Mean Absolute Difference in VLs**	*N*	**Mean Observed VL**	**Mean Primary VL**	**Mean Absolute Difference in VLs**
			[Min–Max]	**[Min–Max]**	**[Min–Max]**		**[Min–Max]**	**[Min–Max]**	**[Min–Max]**
Grade A variant found	including exome	*79*	99.73% [98.50%–100.00%]	99.61% [97.03%–100.00%]	0.44% [0.00%–2.73%]	*77*	99.43% [98.20%–100.00%]	99.60% [97.03%–100.00%]	0.65% [0.00%–2.77%]
Grade B variant found	including exome	*5*	26.06% [3.50%–92.50%]	28.59% [2.57%–94.77%]	2.91% [0.93%–7.47%]	*2*	55.65% [19.60%–91.70%]	58.57% [22.37%–94.77%]	2.92% [2.77%–3.07%]
	excluding exome	*8*	7.07% [2.50%–28.40%]	6.36% [1.10%–28.35%]	0.79% [0.05%–2.10%]	*1*	28.30%	28.35%	0.05%
Grade C variant found	from highCov	-	-	-	-	*1*	21.70%	0.40%	21.30%
Grade A variant lost	including exome	-	-	-	-	*2*	-	100.00% [100.00%–100.00%]	-
Grade B variant lost	including exome	*4*	-	0.88% [0.50%–1.27%]	-	*7*	-	4.19% [0.50%–17.37%]	-
	excluding exome	*34*	-	1.10% [0.40%–5.75%]	-	*41*	-	1.46% [0.40%–6.20%]	-
	from lowCov and exome	*2*	-	0.48% [0.45%–0.50%]	-	*2*	-	0.48% [0.45%–0.50%]	-
Grade C variant lost	from highCov	*177*	-	0.82% [0.40%–5.80%]	-	*176*	-	0.83% [0.40%–5.80%]	-
	from lowCov	*12*	-	0.93% [0.40%–3.80%]	-	*12*	-	0.93% [0.40%–3.80%]	-
	from exome	*4*	-	0.70% [0.40%–1.00%]	-	*4*	-	0.70% [0.40%–1.00%]	-
Novel variant	Only present in Ion Torrent	*1*	2.70%	-	-	*2*	25.15% [19.30%–31.00%]	-	-

Abbreviations: PCP—PrecisionCallerPipeline; TSS—Ion Torrent Suite™ Software; N—number; VL—variant level; Min—minimum; Max—maximum.

Table A2. Mixture analysis: Distribution of variants with the mutserve VCM.

Variant Description				PCP				TSS		
					Within Platform	Other Platforms			Within Platform	Other Platforms
Macro	**Meso**	**Micro**	*N*	**Mean Observed VL**	**Mean Absolute Difference in VLs**	**Mean Absolute Difference in VLs**	*N*	**Mean Observed VL**	**Mean Absolute Difference in VLs**	**Mean Absolute Difference in VLs**
				[Min–Max]	**[Min–Max]**	**[Min–Max]**		**[Min–Max]**	**[Min–Max]**	**[Min–Max]**
Found variants	Found variants	Major Grade A	*270*	92.51% [51.50%–100.00%]	3.51% [0.00%–22.90%]	3.73% [0.00%–23.42%]	*260*	92.10% [56.80%–100.00%]	3.40% [0.10%–18.20%]	3.55% [0.00%–18.15%]
		Major Grade B	*39*	20.82% [2.60%–94.30%]	1.73% [0.16%–6.32%]	2.09% [0.29%–6.10%]	*15*	43.18% [11.80%–94.20%]	2.26% [0.14%–6.29%]	2.25% [0.08%–4.98%]
		Major Grade C	-	-	-	-	*5*	28.48% [22.30%–34.10%]	8.65% [1.03%–13.49%]	28.11% [21.91%–33.72%]
		Minor Grade A	*133*	16.95% [2.50%–52.90%]	4.38% [0.02%–28.00%]	4.41% [0.06%–28.20%]	*73*	25.31% [11.00%–48.30%]	5.73% [0.00%–23.50%]	5.75% [0.02%–23.60%]
		Minor Grade B	*8*	7.54% [3.30%–26.00%]	1.90% [0.32%–3.95%]	1.72% [0.14%–4.18%]	*1*	26.30%	3.38%	2.61%
		Shared Grade A	*125*	99.89% [99.20%–100.00%]	0.09% [0.00%–0.40%]	0.20% [0.00%–2.41%]	*125*	99.64% [98.60%–100.00%]	0.20% [0.00%–1.30%]	0.45% [0.00%–2.71%]
	Mixture found variants	Major Grade B	*9*	4.08% [2.60%–5.70%]	-	0.44% [0.03%–1.69%]	*4*	12.60% [11.90%–13.20%]	-	3.12% [0.83%–3.99%]
Lost variants	Mixture lost variants	Major Grade B	*26*	-	2.77% [1.88%–3.56%]	2.00% [0.83%–2.72%]	-	-	-	-
		Minor Grade A	*137*	-	2.44% [0.98%–24.80%]	2.44% [0.97%–24.98%]	*187*	-	3.86% [0.98%–10.00%]	3.87% [0.97%–10.00%]
		Minor Grade B	*57*	-	0.48% [0.03%–4.63%]	0.48% [0.01%–4.74%]	*14*	-	2.65% [0.20%–9.17%]	2.78% [0.22%–9.48%]
	Old lost variants	Major Grade A	-	-	-	-	*10*	-	-	91.40% [75.00%–99.00%]
		Major Grade B	*186*	-	-	0.81% [0.30%–2.70%]	*241*	-	-	1.44% [0.30%–16.50%]
		Major Grade C	*430*	-	-	0.72% [0.30%–5.74%]	*425*	-	-	0.73% [0.30%–5.74%]
		Minor Grade A	-	-	-	-	*10*	-	-	8.60% [1.00%–25.00%]

Abbreviations: PCP—PrecisionCallerPipeline; TSS—Ion Torrent Suite™ Software; N—number; VL—variant level; Min—minimum; Max—maximum.

Table A2. *Cont.*

Variant Description			**PCP**				**TSS**			
					Within Platform	**Other Platforms**			**Within Platform**	**Other Platforms**
Macro	**Meso**	**Micro**	***N***	**Mean Observed VL**	**Mean Absolute Difference in VLs**	**Mean Absolute Difference in VLs**	***N***	**Mean Observed VL**	**Mean Absolute Difference in VLs**	**Mean Absolute Difference in VLs**
				[Min–Max]	**[Min–Max]**	**[Min–Max]**		**[Min–Max]**	**[Min–Max]**	**[Min–Max]**
		Minor Grade B	*195*	-	-	0.09% [0.00%–1.44%]	*245*	-	-	0.16% [0.00%–4.34%]
		Minor Grade C	*430*	-	-	0.06% [0.00%–0.95%]	*425*	-	-	0.06% [0.00%–0.95%]
		Shared Grade C	*540*	-	-	0.86% [0.40%–2.69%]	*540*	-	-	0.86% [0.40%–2.69%]
Novel variants	Found variants	Minor novel	-	-	-	-	*5*	21.24% [16.80%–26.00%]	18.07% [11.98%–24.07%]	-
		Shared novel	-	-	-	-	*4*	31.65% [28.00%–35.10%]	12.10% [8.10%–15.78%]	-
	Novel variants	Novel variant: mixture	*12*	27.26% [2.50%–99.20%]	-	-	*9*	13.92% [10.60%–21.10%]	-	-
Primary novel variants	Mixture lost variants	Major novel	*5*	-	2.47% [2.02%–2.67%]	-	*5*	-	28.33% [23.25%–30.69%]	-
		Minor novel	*5*	-	0.23% [0.03%–0.68%]	-	*5*	-	1.15% [0.19%–3.10%]	-
		Shared novel	-	-	-	-	*1*	-	19.35%	-

Abbreviations: PCP—PrecisionCallerPipeline; TSS—Ion Torrent Suite™ Software; N—number; VL—variant level; Min—minimum; Max—maximum.

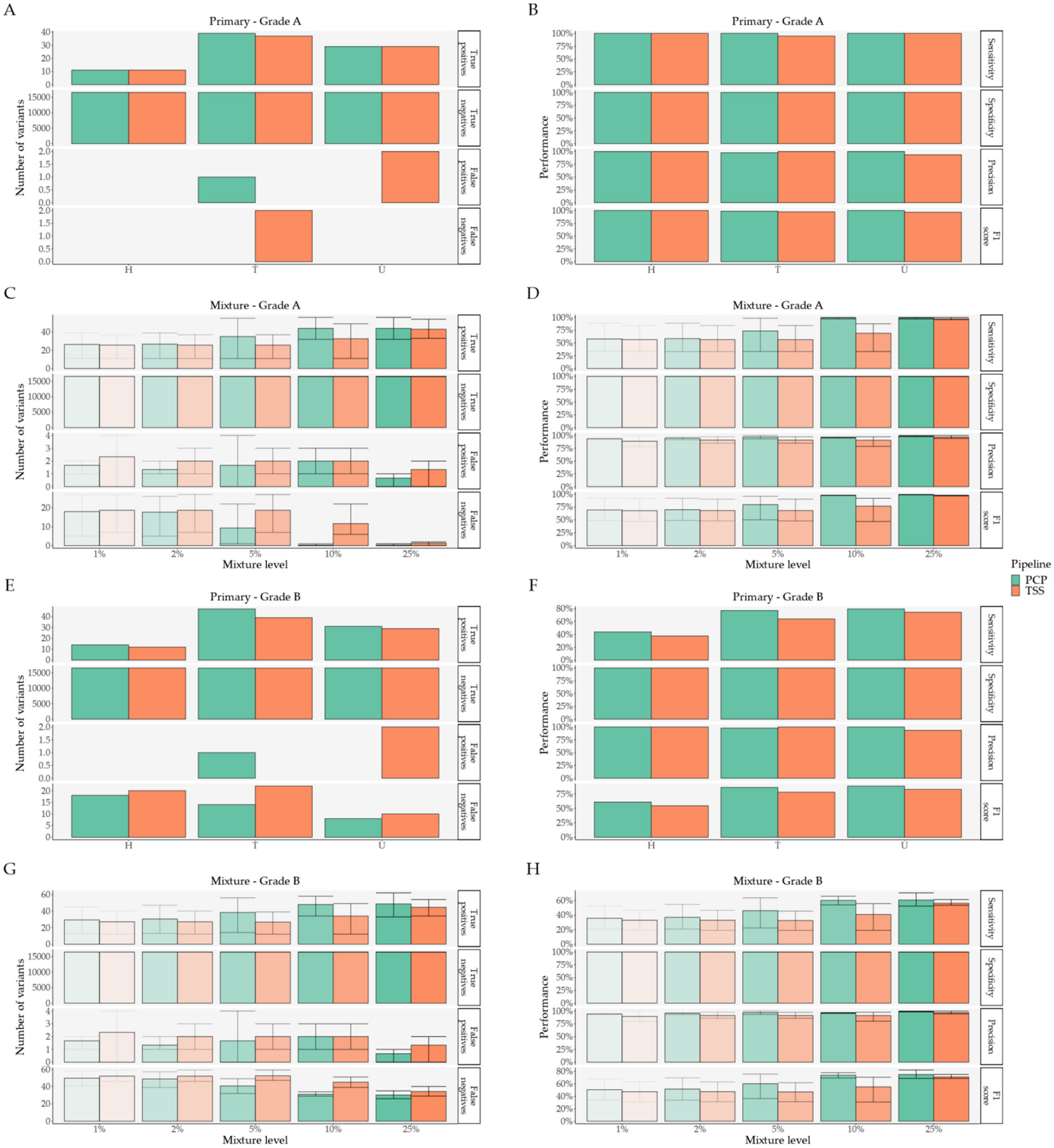

Figure A1. Performance analysis in PCP and TSS. (**A**,**C**,**E**,**G**) Performance metric variant classification per pipeline for Primary Grade A, Mixture Grade A, Primary Grade B, and Mixture Grade B datasets, respectively; (**B**,**D**,**F**,**H**) Performance metrics per pipeline for Primary Grade A, Mixture Grade A, Primary Grade B, and Mixture Grade B datasets, respectively. Error bars denote minimum and maximum values. Abbreviations: PCP—PrecisionCallerPipeline; TSS—Ion Torrent Suite™ Software.

References

1. Taylor, R.W.; Turnbull, D.M. Mitochondrial DNA Mutations in Human Disease. *Nat. Rev. Genet.* **2005**, *6*, 389–402. [CrossRef]
2. Tuppen, H.A.L.; Blakely, E.L.; Turnbull, D.M.; Taylor, R.W. Mitochondrial DNA Mutations and Human Disease. *Biochim. Biophys. Acta* **2010**, *1797*, 113–128. [CrossRef]
3. Van der Auwera, G.A.; Carneiro, M.O.; Hartl, C.; Poplin, R.; Del Angel, G.; Levy-Moonshine, A.; Jordan, T.; Shakir, K.; Roazen, D.; Thibault, J.; et al. From FastQ Data to High Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr. Protoc. Bioinform.* **2013**, *43*, 11.10.1–11.10.33. [CrossRef]
4. 1000 Genomes Project Consortium; Auton, A.; Brooks, L.D.; Durbin, R.M.; Garrison, E.P.; Kang, H.M.; Korbel, J.O.; Marchini, J.L.; McCarthy, S.; McVean, G.A.; et al. A Global Reference for Human Genetic Variation. *Nature* **2015**, *526*, 68–74. [CrossRef]
5. Gatk4-Mitochondria-Pipeline. Available online: https://github.com/gatk-workflows/gatk4-mitochondria-pipeline (accessed on 18 August 2021).
6. Lee, H.Y.; Song, I.; Ha, E.; Cho, S.-B.; Yang, W.I.; Shin, K.-J. MtDNAmanager: A Web-Based Tool for the Management and Quality Analysis of Mitochondrial DNA Control-Region Sequences. *BMC Bioinform.* **2008**, *9*, 483. [CrossRef] [PubMed]
7. Fan, L.; Yao, Y.-G. MitoTool: A Web Server for the Analysis and Retrieval of Human Mitochondrial DNA Sequence Variations. *Mitochondrion* **2011**, *11*, 351–356. [CrossRef]
8. Zhidkov, I.; Nagar, T.; Mishmar, D.; Rubin, E. MitoBamAnnotator: A Web-Based Tool for Detecting and Annotating Heteroplasmy in Human Mitochondrial DNA Sequences. *Mitochondrion* **2011**, *11*, 924–928. [CrossRef]
9. Guo, Y.; Li, J.; Li, C.-I.; Shyr, Y.; Samuels, D.C. MitoSeek: Extracting Mitochondria Information and Performing High-Throughput Mitochondria Sequencing Analysis. *Bioinformatics* **2013**, *29*, 1210–1211. [CrossRef] [PubMed]
10. Vianello, D.; Sevini, F.; Castellani, G.; Lomartire, L.; Capri, M.; Franceschi, C. HAPLOFIND: A New Method for High-Throughput MtDNA Haplogroup Assignment. *Hum. Mutat.* **2013**, *34*, 1189–1194. [CrossRef] [PubMed]
11. Yang, I.S.; Lee, H.Y.; Yang, W.I.; Shin, K.-J. MtDNAprofiler: A Web Application for the Nomenclature and Comparison of Human Mitochondrial DNA Sequences. *J. Forensic Sci.* **2013**, *58*, 972–980. [CrossRef] [PubMed]
12. Lott, M.T.; Leipzig, J.N.; Derbeneva, O.; Xie, H.M.; Chalkia, D.; Sarmady, M.; Procaccio, V.; Wallace, D.C. MtDNA Variation and Analysis Using Mitomap and Mitomaster. *Curr. Protoc. Bioinform.* **2013**, *44*, 1.23.1–1.23.26. [CrossRef]
13. Calabrese, C.; Simone, D.; Diroma, M.A.; Santorsola, M.; Guttà, C.; Gasparre, G.; Picardi, E.; Pesole, G.; Attimonelli, M. MToolBox: A Highly Automated Pipeline for Heteroplasmy Annotation and Prioritization Analysis of Human Mitochondrial Variants in High-Throughput Sequencing. *Bioinformatics* **2014**, *30*, 3115–3117. [CrossRef]
14. Navarro-Gomez, D.; Leipzig, J.; Shen, L.; Lott, M.; Stassen, A.P.M.; Wallace, D.C.; Wiggs, J.L.; Falk, M.J.; van Oven, M.; Gai, X. Phy-Mer: A Novel Alignment-Free and Reference-Independent Mitochondrial Haplogroup Classifier. *Bioinformatics* **2015**, *31*, 1310–1312. [CrossRef] [PubMed]
15. Falk, M.J.; Shen, L.; Gonzalez, M.; Leipzig, J.; Lott, M.T.; Stassen, A.P.M.; Diroma, M.A.; Navarro-Gomez, D.; Yeske, P.; Bai, R.; et al. Mitochondrial Disease Sequence Data Resource (MSeqDR): A Global Grass-Roots Consortium to Facilitate Deposition, Curation, Annotation, and Integrated Analysis of Genomic Data for the Mitochondrial Disease Clinical and Research Communities. *Mol. Genet. Metab.* **2015**, *114*, 388–396. [CrossRef] [PubMed]
16. Vellarikkal, S.K.; Dhiman, H.; Joshi, K.; Hasija, Y.; Sivasubbu, S.; Scaria, V. Mit-o-Matic: A Comprehensive Computational Pipeline for Clinical Evaluation of Mitochondrial Variations from next-Generation Sequencing Datasets. *Hum. Mutat.* **2015**, *36*, 419–424. [CrossRef] [PubMed]
17. Weissensteiner, H.; Forer, L.; Fuchsberger, C.; Schöpf, B.; Kloss-Brandstätter, A.; Specht, G.; Kronenberg, F.; Schönherr, S. MtDNA-Server: Next-Generation Sequencing Data Analysis of Human Mitochondrial DNA in the Cloud. *Nucleic Acids Res.* **2016**, *44*, W64–W69. [CrossRef]
18. Weissensteiner, H.; Pacher, D.; Kloss-Brandstätter, A.; Forer, L.; Specht, G.; Bandelt, H.-J.; Kronenberg, F.; Salas, A.; Schönherr, S. HaploGrep 2: Mitochondrial Haplogroup Classification in the Era of High-Throughput Sequencing. *Nucleic Acids Res.* **2016**, *44*, W58–W63. [CrossRef]
19. Ishiya, K.; Ueda, S. MitoSuite: A Graphical Tool for Human Mitochondrial Genome Profiling in Massive Parallel Sequencing. *PeerJ* **2017**, *5*, e3406. [CrossRef]
20. Rueda, M.; Torkamani, A. SG-ADVISER MtDNA: A Web Server for Mitochondrial DNA Annotation with Data from 200 Samples of a Healthy Aging Cohort. *BMC Bioinform.* **2017**, *18*, 373. [CrossRef]
21. Preste, R.; Vitale, O.; Clima, R.; Gasparre, G.; Attimonelli, M. HmtVar: A New Resource for Human Mitochondrial Variations and Pathogenicity Data. *Nucleic Acids Res.* **2019**, *47*, D1202–D1210. [CrossRef]
22. Andrews, R.M.; Kubacka, I.; Chinnery, P.F.; Lightowlers, R.N.; Turnbull, D.M.; Howell, N. Reanalysis and Revision of the Cambridge Reference Sequence for Human Mitochondrial DNA. *Nat. Genet.* **1999**, *23*, 147. [CrossRef]
23. Van Oven, M. PhyloTree Build 17: Growing the Human Mitochondrial DNA Tree. *Forensic Sci. Int. Genet. Suppl. Ser.* **2015**, *5*, e392–e394. [CrossRef]
24. Dür, A.; Huber, N.; Parson, W. Fine-Tuning Phylogenetic Alignment and Haplogrouping of MtDNA Sequences. *Int. J. Mol. Sci.* **2021**, *22*, 5747. [CrossRef]
25. Chaitanya, L.; Ralf, A.; van Oven, M.; Kupiec, T.; Chang, J.; Lagacé, R.; Kayser, M. Simultaneous Whole Mitochondrial Genome Sequencing with Short Overlapping Amplicons Suitable for Degraded DNA Using the Ion Torrent Personal Genome Machine. *Hum. Mutat.* **2015**, *36*, 1236–1247. [CrossRef] [PubMed]

26. Wai, K.T.; Barash, M.; Gunn, P. Performance of the Early Access AmpliSeq^{TM} Mitochondrial Panel with Degraded DNA Samples Using the Ion Torrent^{TM} Platform. *Electrophoresis* **2018**, *39*, 2776–2784. [CrossRef] [PubMed]
27. Yao, L.; Xu, Z.; Zhao, H.; Tu, Z.; Liu, Z.; Li, W.; Hu, L.; Wan, L. Concordance of Mitochondrial DNA Sequencing Methods on Bloodstains Using Ion PGM^{TM}. *Leg. Med.* **2018**, *32*, 27–30. [CrossRef]
28. Strobl, C.; Eduardoff, M.; Bus, M.M.; Allen, M.; Parson, W. Evaluation of the Precision ID Whole MtDNA Genome Panel for Forensic Analyses. *Forensic. Sci. Int. Genet.* **2018**, *35*, 21–25. [CrossRef]
29. Cuenca, D.; Battaglia, J.; Halsing, M.; Sheehan, S. Mitochondrial Sequencing of Missing Persons DNA Casework by Implementing Thermo Fisher's Precision ID MtDNA Whole Genome Assay. *Genes* **2020**, *11*, 1303. [CrossRef] [PubMed]
30. Pereira, V.; Longobardi, A.; Børsting, C. Sequencing of Mitochondrial Genomes Using the Precision ID MtDNA Whole Genome Panel. *Electrophoresis* **2018**, *39*, 2766–2775. [CrossRef]
31. Faccinetto, C.; Sabbatini, D.; Serventi, P.; Rigato, M.; Salvoro, C.; Casamassima, G.; Margiotta, G.; De Fanti, S.; Sarno, S.; Staiti, N.; et al. Internal Validation and Improvement of Mitochondrial Genome Sequencing Using the Precision ID MtDNA Whole Genome Panel. *Int. J. Legal Med.* **2021**, *135*, 2295–2306. [CrossRef]
32. Strobl, C.; Churchill Cihlar, J.; Lagacé, R.; Wootton, S.; Roth, C.; Huber, N.; Schnaller, L.; Zimmermann, B.; Huber, G.; Lay Hong, S.; et al. Evaluation of Mitogenome Sequence Concordance, Heteroplasmy Detection, and Haplogrouping in a Worldwide Lineage Study Using the Precision ID MtDNA Whole Genome Panel. *Forensic Sci. Int. Genet.* **2019**, *42*, 244–251. [CrossRef] [PubMed]
33. Ming, T.; Wang, M.; Zheng, M.; Zhou, Y.; Hou, Y.; Wang, Z. Exploring of Rare Differences in MtGenomes between MZ Twins Using Massively Parallel Sequencing. *Forensic Sci. Int. Genet. Suppl. Ser.* **2019**, *7*, 70–72. [CrossRef]
34. Woerner, A.E.; Ambers, A.; Wendt, F.R.; King, J.L.; Moura-Neto, R.S.; Silva, R.; Budowle, B. Evaluation of the Precision ID MtDNA Whole Genome Panel on Two Massively Parallel Sequencing Systems. *Forensic Sci. Int. Genet.* **2018**, *36*, 213–224. [CrossRef] [PubMed]
35. Cihlar, J.C.; Amory, C.; Lagacé, R.; Roth, C.; Parson, W.; Budowle, B. Developmental Validation of a MPS Workflow with a PCR-Based Short Amplicon Whole Mitochondrial Genome Panel. *Genes* **2020**, *11*, 1345. [CrossRef]
36. Roth, C.; Parson, W.; Strobl, C.; Lagacé, R.; Short, M. MVC: An Integrated Mitochondrial Variant Caller for Forensics. *Aust. J. Forensic Sci.* **2019**, *51*, S52–S55. [CrossRef]
37. Parson, W.; Gusmão, L.; Hares, D.R.; Irwin, J.A.; Mayr, W.R.; Morling, N.; Pokorak, E.; Prinz, M.; Salas, A.; Schneider, P.M.; et al. DNA Commission of the International Society for Forensic Genetics: Revised and Extended Guidelines for Mitochondrial DNA Typing. *Forensic Sci. Int. Genet.* **2014**, *13*, 134–142. [CrossRef]
38. Cho, S.; Kim, M.Y.; Lee, J.H.; Lee, S.D. Assessment of Mitochondrial DNA Heteroplasmy Detected on Commercial Panel Using MPS System with Artificial Mixture Samples. *Int. J. Legal Med.* **2018**, *132*, 1049–1056. [CrossRef]
39. Churchill, J.D.; Stoljarova, M.; King, J.L.; Budowle, B. Massively Parallel Sequencing-Enabled Mixture Analysis of Mitochondrial DNA Samples. *Int. J. Legal Med.* **2018**, *132*, 1263–1272. [CrossRef]
40. Weissensteiner, H.; Forer, L.; Fendt, L.; Kheirkhah, A.; Salas, A.; Kronenberg, F.; Schoenherr, S. Contamination Detection in Sequencing Studies Using the Mitochondrial Phylogeny. *Genome Res.* **2021**, *31*, 309–316. [CrossRef]
41. Smart, U.; Budowle, B.; Ambers, A.; Soares Moura-Neto, R.; Silva, R.; Woerner, A.E. A Novel Phylogenetic Approach for de Novo Discovery of Putative Nuclear Mitochondrial (PNumt) Haplotypes. *Forensic Sci. Int. Genet.* **2019**, *43*, 102146. [CrossRef]
42. Garrison, E.; Marth, G. Haplotype-Based Variant Detection from Short-Read Sequencing. *arXiv* **2012**, arXiv:1207.3907.
43. Koboldt, D.C.; Zhang, Q.; Larson, D.E.; Shen, D.; McLellan, M.D.; Lin, L.; Miller, C.A.; Mardis, E.R.; Ding, L.; Wilson, R.K. VarScan 2: Somatic Mutation and Copy Number Alteration Discovery in Cancer by Exome Sequencing. *Genome Res.* **2012**, *22*, 568–576. [CrossRef]
44. Fazzini, F.; Fendt, L.; Schönherr, S.; Forer, L.; Schöpf, B.; Streiter, G.; Losso, J.L.; Kloss-Brandstätter, A.; Kronenberg, F.; Weissensteiner, H. Analyzing Low-Level MtDNA Heteroplasmy—Pitfalls and Challenges from Bench to Benchmarking. *Int. J. Mol. Sci.* **2021**, *22*, 935. [CrossRef]
45. Ring, J.D.; Sturk-Andreaggi, K.; Alyse Peck, M.; Marshall, C. Bioinformatic Removal of NUMT-Associated Variants in Mitotiling next-Generation Sequencing Data from Whole Blood Samples. *Electrophoresis* **2018**, *39*, 2785–2797. [CrossRef]
46. Genomics England Research Consortium; NIHR BioResource; Wei, W.; Pagnamenta, A.T.; Gleadall, N.; Sanchis-Juan, A.; Stephens, J.; Broxholme, J.; Tuna, S.; Odhams, C.A.; et al. Nuclear-Mitochondrial DNA Segments Resemble Paternally Inherited Mitochondrial DNA in Humans. *Nat. Commun.* **2020**, *11*, 1740. [CrossRef] [PubMed]
47. Woerner, A.E.; Cihlar, J.C.; Smart, U.; Budowle, B. Numt Identification and Removal with RtN! *Bioinformatics* **2020**, *36*, 5115–5116. [CrossRef] [PubMed]
48. Cihlar, J.C.; Strobl, C.; Lagacé, R.; Muenzler, M.; Parson, W.; Budowle, B. Distinguishing Mitochondrial DNA and NUMT Sequences Amplified with the Precision ID MtDNA Whole Genome Panel. *Mitochondrion* **2020**, *55*, 122–133. [CrossRef] [PubMed]
49. Marshall, C.; Parson, W. Interpreting NUMTs in Forensic Genetics: Seeing the Forest for the Trees. *Forensic Sci. Int. Genet.* **2021**, *53*, 102497. [CrossRef]
50. Yonova-Doing, E.; Calabrese, C.; Gomez-Duran, A.; Schon, K.; Wei, W.; Karthikeyan, S.; Chinnery, P.F.; Howson, J.M.M. An Atlas of Mitochondrial DNA Genotype-Phenotype Associations in the UK Biobank. *Nat. Genet.* **2021**, *53*, 982–993. [CrossRef]
51. Sosa, M.X.; Sivakumar, I.K.A.; Maragh, S.; Veeramachaneni, V.; Hariharan, R.; Parulekar, M.; Fredrikson, K.M.; Harkins, T.T.; Lin, J.; Feldman, A.B.; et al. Next-Generation Sequencing of Human Mitochondrial Reference Genomes Uncovers High Heteroplasmy Frequency. *PLoS Comput. Biol.* **2012**, *8*, e1002737. [CrossRef]

52. Ye, K.; Lu, J.; Ma, F.; Keinan, A.; Gu, Z. Extensive Pathogenicity of Mitochondrial Heteroplasmy in Healthy Human Individuals. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 10654–10659. [CrossRef]
53. Li, M.; Schröder, R.; Ni, S.; Madea, B.; Stoneking, M. Extensive Tissue-Related and Allele-Related MtDNA Heteroplasmy Suggests Positive Selection for Somatic Mutations. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 2491–2496. [CrossRef]
54. Naue, J.; Hörer, S.; Sänger, T.; Strobl, C.; Hatzer-Grubwieser, P.; Parson, W.; Lutz-Bonengel, S. Evidence for Frequent and Tissue-Specific Sequence Heteroplasmy in Human Mitochondrial DNA. *Mitochondrion* **2015**, *20*, 82–94. [CrossRef]
55. Guo, Y.; Li, C.-I.; Sheng, Q.; Winther, J.F.; Cai, Q.; Boice, J.D.; Shyr, Y. Very Low-Level Heteroplasmy MtDNA Variations Are Inherited in Humans. *J. Genet. Genom.* **2013**, *40*, 607–615. [CrossRef] [PubMed]
56. Zaidi, A.A.; Wilton, P.R.; Su, M.S.-W.; Paul, I.M.; Arbeithuber, B.; Anthony, K.; Nekrutenko, A.; Nielsen, R.; Makova, K.D. Bottleneck and Selection in the Germline and Maternal Age Influence Transmission of Mitochondrial DNA in Human Pedigrees. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 25172–25178. [CrossRef] [PubMed]
57. Wachsmuth, M.; Hübner, A.; Li, M.; Madea, B.; Stoneking, M. Age-Related and Heteroplasmy-Related Variation in Human MtDNA Copy Number. *PLoS Genet.* **2016**, *12*, e1005939. [CrossRef] [PubMed]
58. Yuan, Y.; Ju, Y.S.; Kim, Y.; Li, J.; Wang, Y.; Yoon, C.J.; Yang, Y.; Martincorena, I.; Creighton, C.J.; Weinstein, J.N.; et al. Comprehensive Molecular Characterization of Mitochondrial Genomes in Human Cancers. *Nat. Genet.* **2020**, *52*, 342–352. [CrossRef] [PubMed]
59. Fendt, L.; Fazzini, F.; Weissensteiner, H.; Bruckmoser, E.; Schönherr, S.; Schäfer, G.; Losso, J.L.; Streiter, G.A.; Lamina, C.; Rasse, M.; et al. Profiling of Mitochondrial DNA Heteroplasmy in a Prospective Oral Squamous Cell Carcinoma Study. *Cancers* **2020**, *12*, 1933. [CrossRef]
60. Just, R.S.; Irwin, J.A.; Parson, W. Questioning the Prevalence and Reliability of Human Mitochondrial DNA Heteroplasmy from Massively Parallel Sequencing Data. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 4546–4547. [CrossRef]
61. Ye, K.; Lu, J.; Ma, F.; Keinan, A.; Gu, Z. Reply to Just et al. Mitochondrial DNA Heteroplasmy Could Be Reliably Detected with Massively Parallel Sequencing Technologies. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, E4548–E4550. [CrossRef] [PubMed]
62. Just, R.S.; Irwin, J.A.; Parson, W. Mitochondrial DNA Heteroplasmy in the Emerging Field of Massively Parallel Sequencing. *Forensic Sci. Int. Genet.* **2015**, *18*, 131–139. [CrossRef]
63. Bris, C.; Goudenege, D.; Desquiret-Dumas, V.; Charif, M.; Colin, E.; Bonneau, D.; Amati-Bonneau, P.; Lenaers, G.; Reynier, P.; Procaccio, V. Bioinformatics Tools and Databases to Assess the Pathogenicity of Mitochondrial DNA Variants in the Field of Next Generation Sequencing. *Front. Genet.* **2018**, *9*, 632. [CrossRef] [PubMed]
64. Brandhagen, M.D.; Just, R.S.; Irwin, J.A. Validation of NGS for Mitochondrial DNA Casework at the FBI Laboratory. *Forensic Sci. Int. Genet.* **2020**, *44*, 102151. [CrossRef] [PubMed]
65. Poole, O.V.; Pizzamiglio, C.; Murphy, D.; Falabella, M.; Macken, W.L.; Bugiardini, E.; Woodward, C.E.; Labrum, R.; Efthymiou, S.; Salpietro, V.; et al. Mitochondrial DNA Analysis from Exome Sequencing Data Improves Diagnostic Yield in Neurological Diseases. *Ann. Neurol.* **2021**, *89*, 1240–1247. [CrossRef]
66. Laricchia, K.M.; Lake, N.J.; Watts, N.A.; Shand, M.; Haessly, A.; Gauthier, L.; Benjamin, D.; Banks, E.; Soto, J.; Garimella, K.; et al. Mitochondrial DNA Variation across 56,434 Individuals in GnomAD. *bioRxiv* **2021**. bioRxiv:2021.07.23.453510. [CrossRef]
67. Bolze, A.; Mendez, F.; White, S.; Tanudjaja, F.; Isaksson, M.; Jiang, R.; Rossi, A.D.; Cirulli, E.T.; Rashkin, M.; Metcalf, W.J.; et al. A Catalog of Homoplasmic and Heteroplasmic Mitochondrial DNA Variants in Humans. *bioRxiv* **2020**. bioRxiv:798264. [CrossRef]
68. Rausser, S.; Trumpff, C.; McGill, M.A.; Junker, A.; Wang, W.; Ho, S.; Mitchell, A.; Karan, K.R.; Monk, C.; Segerstrom, S.C.; et al. Mitochondrial Phenotypes in Purified Human Immune Cell Subtypes and Cell Mixtures. *bioRxiv* **2021**. bioRxiv:2020.10.16.342923. [CrossRef]
69. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef]
70. Chervova, O.; Conde, L.; Guerra-Assunção, J.A.; Moghul, I.; Webster, A.P.; Berner, A.; Larose Cadieux, E.; Tian, Y.; Voloshin, V.; Jesus, T.F.; et al. The Personal Genome Project-UK, an Open Access Resource of Human Multi-Omics Data. *Sci. Data* **2019**, *6*, 257. [CrossRef]
71. Jennings, L.J.; Arcila, M.E.; Corless, C.; Kamel-Reid, S.; Lubin, I.M.; Pfeifer, J.; Temple-Smolkin, R.L.; Voelkerding, K.V.; Nikiforova, M.N. Guidelines for Validation of Next-Generation Sequencing-Based Oncology Panels: A Joint Consensus Recommendation of the Association for Molecular Pathology and College of American Pathologists. *J. Mol. Diagn.* **2017**, *19*, 341–365. [CrossRef] [PubMed]
72. Marshall, C.R.; Chowdhury, S.; Taft, R.J.; Lebo, M.S.; Buchan, J.G.; Harrison, S.M.; Rowsey, R.; Klee, E.W.; Liu, P.; Worthey, E.A.; et al. Best Practices for the Analytical Validation of Clinical Whole-Genome Sequencing Intended for the Diagnosis of Germline Disease. *NPJ Genom. Med.* **2020**, *5*, 47. [CrossRef]
73. Wong, A.K.; Sealfon, R.S.G.; Theesfeld, C.L.; Troyanskaya, O.G. Decoding Disease: From Genomes to Networks to Phenotypes. *Nat. Rev. Genet.* **2021**. [CrossRef]
74. Amorim, A.; Fernandes, T.; Taveira, N. Mitochondrial DNA in Human Identification: A Review. *PeerJ* **2019**, *7*, e7314. [CrossRef] [PubMed]
75. Zhao, H.; Shen, J.; Medico, L.; Platek, M.; Ambrosone, C.B. Length Heteroplasmies in Human Mitochondrial DNA Control Regions and Breast Cancer Risk. *Int. J. Mol. Epidemiol. Genet.* **2010**, *1*, 184–192. [PubMed]

76. Sturk-Andreaggi, K.; Parson, W.; Allen, M.; Marshall, C. Impact of the Sequencing Method on the Detection and Interpretation of Mitochondrial DNA Length Heteroplasmy. *Forensic Sci. Int. Genet.* **2020**, *44*, 102205. [CrossRef]
77. Bamford, J.M.; Sandercock, P.A.; Warlow, C.P.; Slattery, J. Interobserver Agreement for the Assessment of Handicap in Stroke Patients. *Stroke* **1989**, *20*, 828. [CrossRef] [PubMed]
78. Köster, J.; Rahmann, S. Snakemake—A Scalable Bioinformatics Workflow Engine. *Bioinformatics* **2012**, *28*, 2520–2522. [CrossRef] [PubMed]
79. Aho, A.V.; Kernighan, B.W.; Weinberger, P.J. Awk—A Pattern Scanning and Processing Language. *Softw. Pract. Exper.* **1979**, *9*, 267–279. [CrossRef]
80. Quinlan, A.R.; Hall, I.M. BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features. *Bioinformatics* **2010**, *26*, 841–842. [CrossRef]
81. Li, H. Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM. *arXiv* **2013**, arXiv:1303.3997.
82. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef] [PubMed]
83. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A Flexible Trimmer for Illumina Sequence Data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef] [PubMed]
84. Andrews, S.; Krueger, F.; Segonds-Pichon, A.; Biggins, L.; Krueger, C.; Wingett, S. FastQC: A Quality Control Tool for High Throughput Sequence Data. Available online: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed on 18 August 2021).
85. Ewels, P.; Magnusson, M.; Lundin, S.; Käller, M. MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report. *Bioinformatics* **2016**, *32*, 3047–3048. [CrossRef] [PubMed]
86. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2021.
87. RStudio Team. *RStudio: Integrated Development Environment for R*; RStudio, PBC: Boston, MA, USA, 2021.
88. Chang, W. *Extrafont: Tools for Using Fonts*; 2014.
89. Bray, A.; Ismay, C.; Chasnovski, E.; Baume, B.; Cetinkaya-Rundel, M. *Infer: Tidy Statistical Inference*; 2021.
90. Ooms, J. *Magick: Advanced Graphics and Image-Processing in R*; 2021.
91. Pedersen, T.L. *Patchwork: The Composer of Plots*; 2020.
92. Wickham, H.; Bryan, J. *Readxl: Read Excel Files*; 2019.
93. Hester, J.; Csárdi, G.; Wickham, H.; Chang, W.; Morgan, M.; Tenenbaum, D. *Remotes: R Package Installation from Remote Repositories, Including "GitHub"*; 2021.
94. Wickham, H.; Seidel, D. *Scales: Scale Functions for Visualization*; 2020.
95. Wickham, H.; Henry, L.; Pedersen, T.L.; Luciani, T.J.; Decorde, M.; Lise, V. *Svglite: An "SVG" Graphics Device*; 2021.
96. Wickham, H.; Averick, M.; Bryan, J.; Chang, W.; McGowan, L.; François, R.; Grolemund, G.; Hayes, A.; Henry, L.; Hester, J.; et al. Welcome to the Tidyverse. *JOSS* **2019**, *4*, 1686. [CrossRef]
97. European Organization for Nuclear Research. *OpenAIRE Zenodo: Research. Shared*; 2013. [CrossRef]