ClockWork: Genotyping of frameshift mutations induced by CRISPR-Cas9

Bryan van den Brand [1], Job Smink [1] and Juan I. Garaycoechea [1]

[1] Group Garaycoechea, Hubrecht Institute, Utrecht, 3584 CT Utrecht, The Netherlands

Abstract

Motivation: Here we present Clockwork, a high-throughput bioinformatics pipeline for genotyping CRISPR-Cas9 induced frameshift mutations in single neurons. Clockwork is part of a project that aims to elucidate the molecular mechanisms of replication-independent mutations associated with the mutational signature SBS5. In this project key components involved in DNA repair and mutagenesis are mutated by CRISPR-Cas9, introducing frameshift mutations and subsequently studying its effect on the mutational rate, pattern and load. However, the success rate of CRISPR-Cas9 is highly dependent on the local sequence context and guide sequences, with samples frequently having no mutations. Therefore Clockwork addresses a need for genotyping successful frameshifts, saving valuable resources by avoiding whole genome sequencing of unedited samples. Clockwork's efficacy is interrogated with a proof-of-concept dataset consisting of hybrid *Mus musculus* and *Mus spretus* embryonic stem cells as well as human retinal pigment epithelium cells.
Availability: https://github.com/Bryan-vd-Brand/
Contact: bryan.v.d.brand@gmail.com

1 Layman's summary

In the last 10 years, sequencing of DNA has become commonplace. Sequencing is a laboratory process used to learn the exact sequence (order) of the four building blocks (bases) that make up the DNA, A, C, G and T. Initiatives like the cancer genome atlas project have aggregated the sequencing data of thousands of cancers and their patients. Such projects represent a powerful tool for researchers to investigate the cause of cancer, mutations (changes) in the order of the DNA. Researchers collect samples of tissue from both the tumour and healthy cells to discern two classes of mutations, germline and somatic mutations. Germline mutations are mutations present from the moment you were born, present in the germline cells, cells that form the egg and sperm. Somatic mutations arise after your birth and are not present in all the cells of your body. Cancers often accumulate significant amounts of somatic mutations and those mutations are catalogued in databases like The Catalogue of Somatic Mutations in Cancer (COSMIC). Careful analysis of the data available in COSMIC revealed distinct patterns of mutations, generally linked to a specific process or cancerous substance. For example the mutational signature SBS4, found in lung tumours is known to be caused by exposure to cigarette smoke. SBS is short for single base substitution, in this case the chemicals in the cigarette smoke cause specific DNA mutations often swapping C's to A's in the DNA sequence of lung cells.

There are many more distinct signatures that were found and for one of them, SBS5, we hope to find the biochemical or molecular process underlying the pattern of mutations associated with it. To do this, our approach will be to disable genes encoding key components of DNA repair systems. Subsequently we observe the effect of the disabled genes on the location, type, pattern and occurrence of mutations on the DNA. However CRISPR-Cas9, the complex used to disable the genes is not perfect, often failing to edit the DNA.

Therefore within this project there is a need to identify cells that were successfully mutated before sequencing all their DNA, a costly process that would be wasted on cells without disabling mutations. Here we designed Clockwork, a bioinformatics pipeline, in order to genotype (identify the DNA sequence) of cells in a high-throughput manner for disabling mutations. Clockwork relies on the sequencing of a small part of the DNA, specifically in the gene where CRISPR-Cas9 introduces mutations and looks for missing bases in the sequence. The absence of the bases indicates the disabling of the component encoded by the gene and subsequently all these results are compiled into graphs for inspection and ease of use. Since the sequencing of a small part of the DNA for genotyping is cheap compared to the sequencing of all the DNA in a sample, Clockwork saves significant resources sequencing samples without disabling mutations.

2 Introduction

Recent advances in next generation sequencing (NGS) has led to the proliferation of sequencing data across many fields of study. Due to initiatives like the cancer genome atlas (TCGA) project, the molecular characterizations of thousands of cancers and matched normal samples have been aggregated in large databases (Chang *et al.*, 2013). Furthermore the collaboration between the TCGA and the international cancer genome consortium (ICGC) led to the Pan-Cancer Analysis of Whole Genomes (PCAWG) project containing whole genome sequencing of over 2600 primary cancers and their matching normal tissue for 38 distinct tumour types (Campbell *et al.*, 2020). By collecting both primary cancer and matching normal tissue, somatic DNA mutations can be discerned from germline mutations. The Catalogue of Somatic Mutations In Cancer (COSMIC) (Tate *et al.*, 2018), carefully analysed the somatic mutations in the PCAWG dataset aiming to categorise and quantify unique combinations of somatic mutations dubbed COSMIC - Mutational Signatures. Four classes of variants are studied in the COSMIC-MS dataset. Single Base Substitutions (SBS) defined as the replacement of a certain nucleotide base. Double Base Substitutions (DBS) are defined as the replacement of two consecutive nucleotide bases. Small Insertions and Deletions (ID) defined as the incorporation or loss of nucleotides and Copy Number Variations (CN) defined as the gain or loss of large segments of the genome as well as the loss-of-heterozygosity, total copy number and segment length. In general, mutational signatures are linked to mutagenic processes. By estimating the relative contribution of known exogenous or endogenous mutagenic processes to discovered signatures in individual cancer genomes, associations can be found. For example, SBS4 found in lung tumours and SBS7 found in melanoma are caused by exposure to cigarette smoke and UV radiation, respectively. SBS1 is an example of a endogenous mutagenic process, associated with the with the deamination of 5-methylcytosine to thymine; failure to repair the resultant G:T mismatch results in T to C substitutions upon DNA replication(Nik-Zainal *et al.*, 2012). However, the molecular mechanisms driving many of the mutational signatures still remains unknown.

Two SBS signatures are of special interest, the aforementioned SBS1 and SBS5. For these signatures, the number of mutations associated with them increases over time and is correlated to the age of the patient across a broad range of cancer types (Alexandrov *et al.*, 2015). This correlation has led to naming SBS1 and SBS5 as “clock-like” mutational signatures. However, the rate of SBS1 does not correlate with SBS5, indicating different biological processes drive their mutation rates (Alexandrov *et al.*, 2015). Recently, single-cell whole genome sequencing of 161 postmitotic single neurons from the prefrontal cortex of 15 individuals showed SBS5 as one of three dominant mutational signatures (Lodato *et al.*, 2018; Bae *et al.*, 2022). Surprisingly, the number of SBS5 mutations continues to increase over time even in post-mitotic cells. During replication, replication coupled repair mechanisms such as interstrand crosslink repair, double-strand break repair and mismatch repair work to address replication stress caused by DNA damage (Chatterjee and Walker, 2017; Cortez, 2019). However these repair mechanisms are not perfect, occasionally transforming DNA damage encountered during replication into DNA mutations. Since SBS5 arises in a postmitotic neuron this challenges the aforementioned mutagenesis dogma, as the mutations have arisen independently of replication.

The overall aim of the project is to elucidate the molecular mechanism of replication-independent SBS5 mutations. To do this, we will study mutagenesis in non-dividing (i.e. post-mitotic) cells to focus the analysis on replication-independent sources of mutation and exclude mutations introduced by replication-coupled mutagenesis (e.g. polymerase errors). Our approach will be to inactivate genes encoding key components involved in DNA repair and mutagenesis, and subsequently study its effect on the mutational burden, rate and pattern of the single neurons. Knockouts will be generated by applying CRISPR-Cas9 to post-mitotic neurons; edited neurons will be subsequently incubated for a period of up to 6 months. Finally, the single cell genomes are amplified using Primary Template-directed Amplification (PTA, Gonzalez *et al.* (2021)) and sequenced for mutational signature analysis. By combining a systematic dissection of DNA repair pathways and single-cell sequencing, we hope to elucidate the aetiology of SBS5 mutations.

Within this project, there is a need for genotyping the single neurons to identify cells with successful frameshifts introduced by CRISPR-Cas9, to avoid whole genome sequencing of samples with in-frame mutations or unedited cells. The success rate of CRISPR-Cas9 is highly dependent on the local sequence context and guide sequence and a significant portion of cells can be unedited (Canver *et al.*, 2017). Following whole-genome amplification by PTA, single neurons will be screened by PCR amplification using primers targeting the CRISPR-Cas9 cut locus and subsequent sequencing of this amplicon. Here we design Clockwork, a bioinformatics pipeline, in order to determine the genotype of single cells in a high-throughput manner identifying cells carrying frame-shift mutations. Alignment of sequencing reads is performed using CRISPResso (Clement *et al.*, 2019), improving alignment of insertions and deletions by preferentially aligning them in the region defined by the CRISPR-Cas9 guide sequence. Clockwork is interrogated with a proof of concept dataset consisting of human retinal pigment epithelium cells (RPE-1) as well as hybrid *Mus musculus* and *Mus spretus* embryonic stem cells (ESCs).

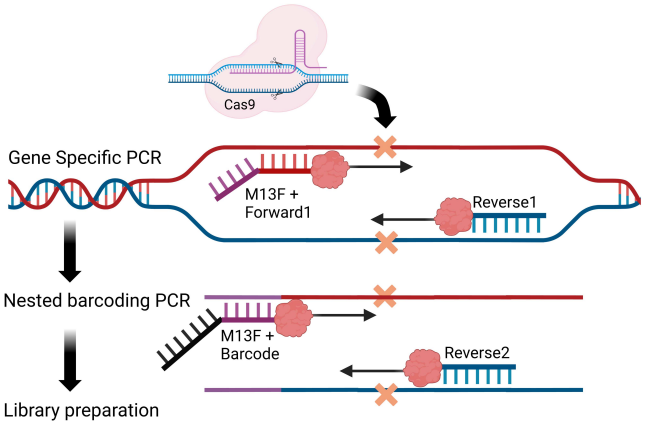

Fig. 1: Graphical representation of the PCR amplification for the human RPE-1 cells and the mouse ES cells. PCR was performed in two steps, the product of the first amplification reaction is used as the template for the second PCR, which is primed by reverse2 that is placed internal to the reverse1.

3 Results

3.1 Proof of concept dataset

Currently, wet lab efforts to transduce and incubate neurons for an extended period of time are ongoing. In order to verify the efficacy of Clockwork, two proof-of-concept datasets are applied. Human retinal pigment epithelium (RPE-1) cells as well as hybrid *Mus musculus* and *Mus spretus* embryonic stem cells (ESCs) were transiently transfected with Cas9 and small guide RNA (sgRNA) sequences targeting two genes of interest. 48 hours later the cells were FACS sorted, isolating GFP+ transfected cells. In RPE-1 cells, *REV7* was targeted, encoding a subunit of polymerase zeta involved in translesion synthesis and microhomology mediated break-repair (Martin and Wood, 2019a). In ESCs *Msh2* was inactivated, which encodes the MSH2 protein, involved in mismatch repair (MMR) and interstrand crosslink repair (ICL) (Edelbrock *et al.*, 2013). Two 96-well plates containing single-cell derived clones of the aforementioned cell lines were PCR amplified. The PCR was performed in two steps, first targeting the loci of interest then a subsequent nested PCR adding 96 barcodes (Fig.1). Subsequent sequencing of the pooled PCR products (Fig.2) provides a representative dataset for interrogating Clockwork's function of detecting frameshifts introduced by CRISPR-Cas9 in 96 samples.

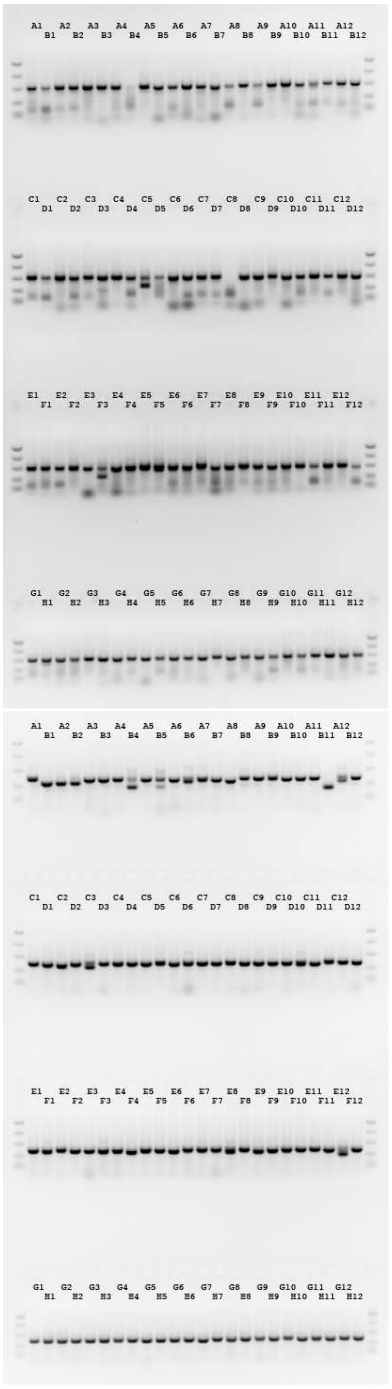

Fig. 2: Top, Gel electrophoresis of the PCR amplification product targeting the gene *REV7* in clonal human RPE-1 cells. Of note samples B4 and C8 do not have a clear band. Bottom, Gel electrophoresis of the PCR amplification product targeting the gene *Msh2* in hybrid mouse ESCs.

3.2 Clockwork's Pipeline

Clockwork's configuration allows for flexibly setting the various definitions and parameters required for analysis. The reference sequence, one for each allele, encompassing the locus of interest for the sample can be set in a fasta formatted file. The guide- and coding- sequences corresponding to that reference can be set in an accompanying file. To allow for multiplexed sequencing of samples, barcodes can be set. Lastly, setting the type of analysis restricts Clockwork to using the paired, forward or reverse reads in the sample. Adapting Clockwork for any locus of interest requires only a change in the aforementioned configuration, allowing for automation of genotyping CRISPR-Cas9 knockouts in wet-lab workflows.

With the configuration set, Clockwork automatically executes the various components of the pipeline. Before processing samples are checked for quality by FastQ and MultiQC (Ewels *et al.*, 2016). Reads are demultiplexed into separate samples. Adapters, barcodes and bad quality bases are trimmed from the reads by cutAdapt (Martin, 2011). CRISPResso2 aligns the reads using an implementation of Gotoh's affine gap alignment algorithm (Gotoh, 1982) that contains an additional parameter, the gap incentive vector. By use of this vector insertions and deletions are preferentially aligned at the CRISPR-Cas9 cut site, improving the accuracy of indel alignments and the genotyping of frameshifts. The various output files and tables of CRISPResso2 are processed by Python's

pandas, quantifying samples as frameshift, in-frame or no modification. Lastly a graphical report for each sample (Fig.3), a visualisation of genotypes in 96-well plate format (Fig.4) and various plots are generated. Fig.3A's barplot gives an indication of alignment efficiency against the reference genome. Fig.3B's piechart shows the percentage of reads modified or unedited for each allele. Fig.3C's Piechart quantifies modified reads as frameshift or in-frame for the *Mus musculus* allele. Fig.3D's Piechart quantifies modified reads as frameshift or in-frame for the *Mus spretus* allele. Fig.3E's plot shows the most frequent alignments of reads for both the *Mus spretus* and *Mus musculus* alleles.

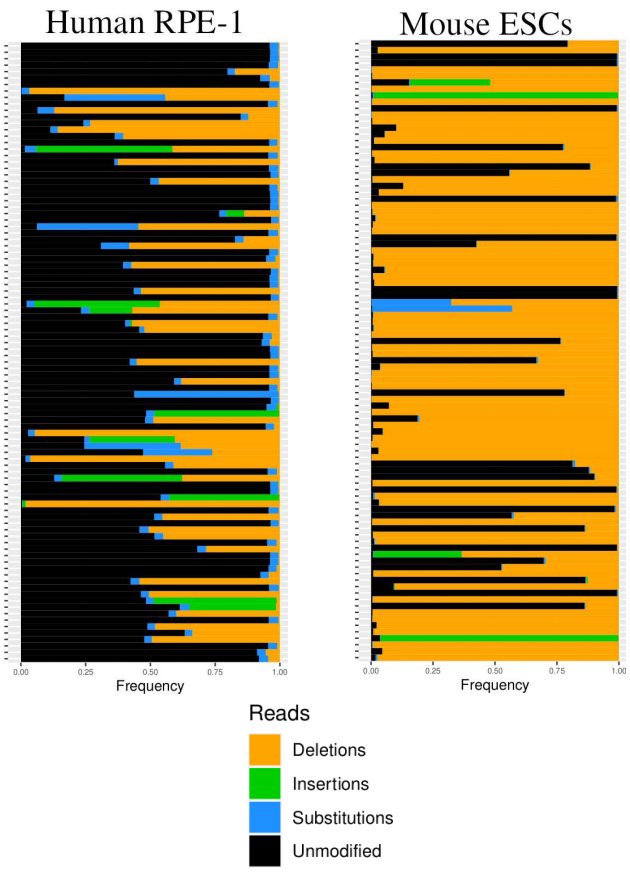

Fig. 5: Left, Frequency of mutation in sequence reads for all 96 samples in the human RPE-1 dataset. Right, Frequency of mutation in sequence reads for all 96 samples in the mouse ESCs dataset.

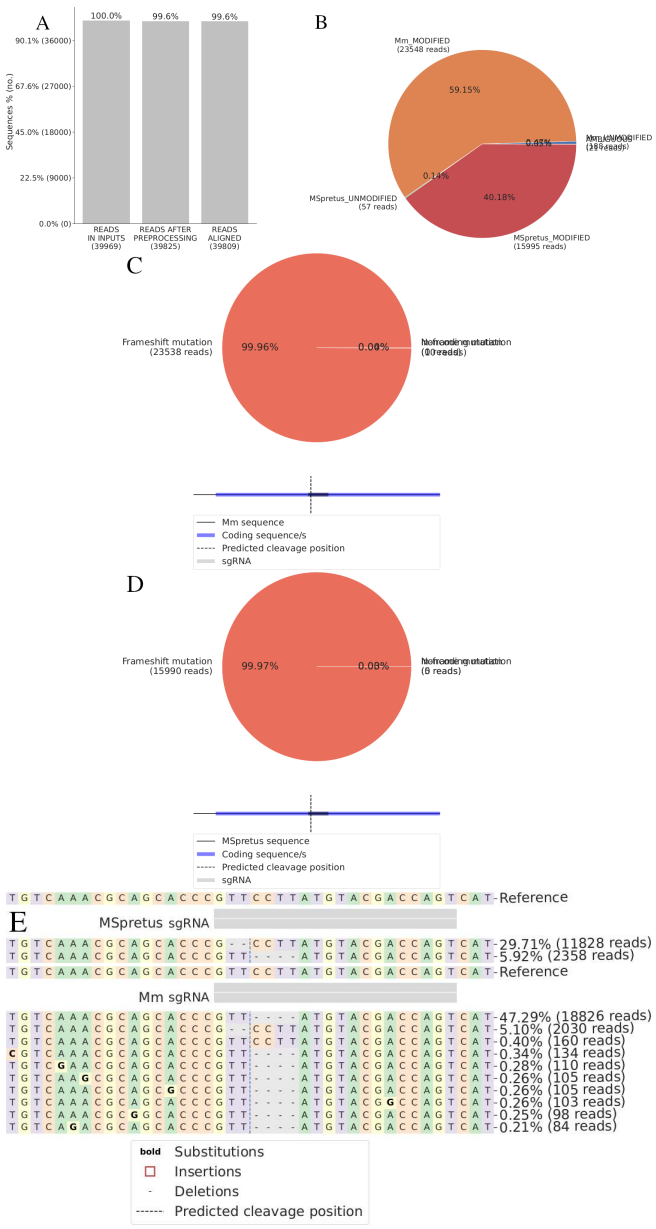

Fig. 3: The Graphical report for sample E1 of the hybrid *Mus musculus* (Mm) and *Mus spretus* (MSpretus) ESCs. A.) Barplot showing the percentage of sequences (reads) aligned to the reference sequence. B.) Piechart quantifying read modifications for both alleles. C,D.) Piechart quantifies the modified reads as frameshift, in-frame or non-coding for each allele. E.) Chart showing the most common sequences for each allele and their modification pattern. Of note are the likely chimeric PCR reads present in the second entry for each allele.

3.3 Genotypes by Clockwork

The alleles are discerned by a single A/G polymorphism 38 bp upstream of the CRISPR-Cas9 cut site and represented as a double circle in Fig.4. Surprisingly, a significant amount of the alleles in the mouse ESC dataset are not genotyped. These untyped alleles fall under the "Insufficient Data" category due to less than 1000 reads aligning against the allele's reference genome. In the case of samples like A2, B1, C4, D7, E6, F4, G3 and H7 either the *Mus spretus* or the *Mus musculous* allele has a significant amount of aligned reads containing a frameshift mutation whilst the other allele's coverage is under the acceptable threshold. Even if the covered allele is consistent with the wild type sequence, allelic dropout still occurs as in samples B2 and C1. In the case of a large deletion at the CRISPR-Cas9 cut site it is possible that the single nucleotide polymorphism is deleted, subsequently distinct mutation patterns can still be discerned but not assigned to an allele. For samples A2, B1 and G3 the polymorphism is missing and a single mutation pattern is present, likely due to the deletion of the primer binding site of the other allele. Interestingly, there appears to be a significant difference in the efficacy of the CRISPR-Cas9 induced mutation between the two datasets (Fig.4, Fig.5). The difference is likely explained by the differing guide sequences (van Overbeek *et al.* (2016), Allen *et al.* (2019)). Furthermore the local sequence context could favour certain repair pathways, with local microhomology known to bias towards microhomology-mediated end joining repair (Allen *et al.*, 2019). The

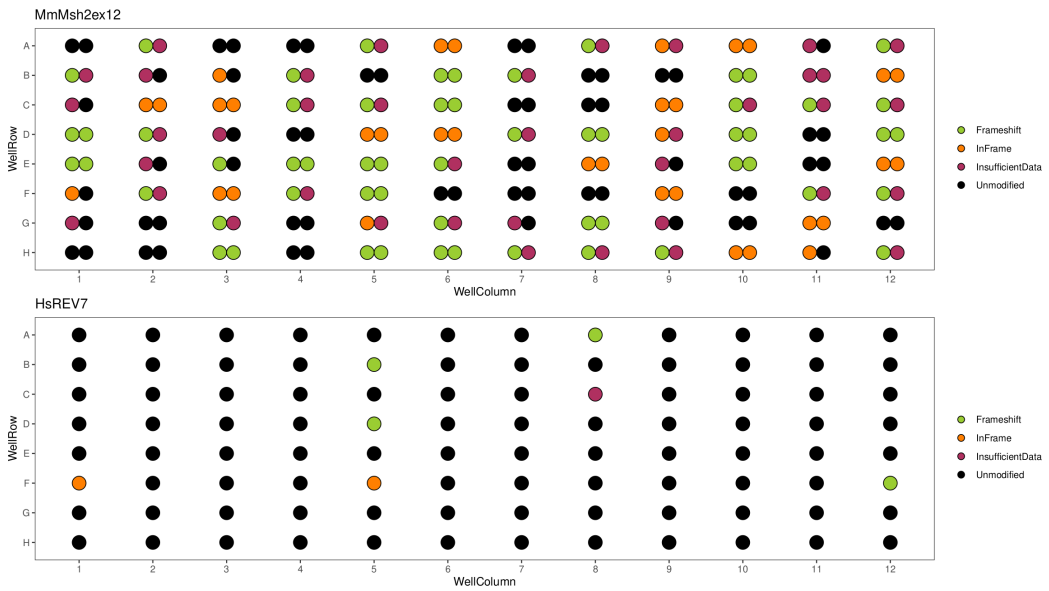

Fig. 4: Top, Visualisation of genotypes in the 96-well plate format for the hybrid mouse ESC dataset. Each sphere represents an allele in the sample and is coloured by the expected genotype of the most common sequence for that allele. Bottom, Visualisation of genotypes in the 96-well plate for the human RPE-1 dataset.

relative activity of the DNA repair pathways in the cell lines can also bias repair outcomes. The comparison of 3777 repair outcomes in *Allen* et al. *(2019)* shows a bias toward microhomology mediated deletions in ESC lines and 1bp insertions in RPE lines. Whilst these factors all influence the resultant genotypes, the efficiency of the guide sequence is the dominant influence (van Overbeek *et al.*, 2016).

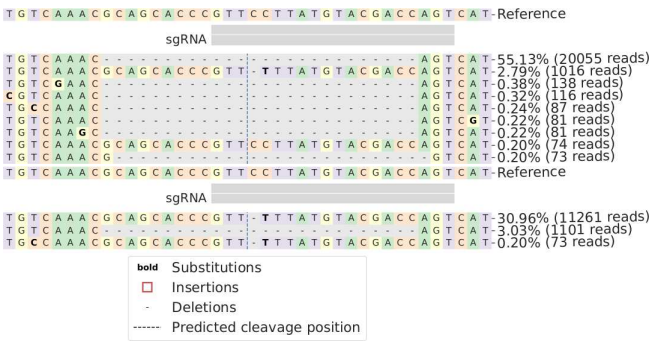

Fig. 6: Allele sequence plot for sample B6 of the hybrid mouse ESC dataset. The top plot represents the *Mus musculus* allele and the bottom plot represents the *Mus spretus* allele. The second entry in both plots represents a chimeric PCR reads with sequence belonging to both alleles.

3.4 Chimeric reads

Several samples of the mouse ESC dataset show clear signs of chimeric PCR products, with reads belonging to *Mus spretus* and the *Mus musculus* alleles having the same deletion (Fig. 3E, 6, 7). On average between 2-5% of the reads of each allele show evidence of chimerism. Chimeric PCR products are formed by the incomplete elongation or premature termination of synthesis. Subsequently the incomplete strand re-anneals to a heterologous sequence like the other allele and continues elongation leading to the formation of a chimeric PCR product (Omelina *et al.*, 2019). Chimeric read formation during PCR can be controlled by adjusting the temperature in the thermal cycles, adding less input template and increasing the extension time (Liu *et al.*, 2014). Chimeric PCR reads can complicate and mislead the upstream analysis, introducing false positive signals. Thresholds applied in Clockwork's pipeline remove the chimeric PCR signal.

Table 1. Genotype per sample

A8	1bp deletion
A9	10bp indel and WT
B5	1bp indel and C insertion
D3	1bp deletion and WT
WT	Parental WT

3.5 Growth inhibition assay

In order to verify Clockwork's prediction of frameshift and in-frame mutations in the human RPE-1 sampleset, we performed a growth inhibition assay on a representative set of samples (Table.1). Samples with a frameshift mutation have a knockout for the REV7 protein, a subunit of polymerase zeta involved in trans-lesion synthesis (Martin and Wood, 2019a). 4NQO has mutagenic properties, introducing bulky lesions analogous to UV damage (Bailleul *et al.*, 1989). Cells deficient in polymerase zeta are unable to bypass these lesions through trans-lesion synthesis, causing replication fork collapse, double stranded breaks and toxicity (Martin and Wood, 2019b). Therefore samples with frameshift mutations are expected to survive at lower rates then their wild-type equivalent.

Surprisingly, Sample A9 called unedited by Clockwork (Fig. 4), shows similar performance to A8 constituting a false negative (Fig. 8). Sample B5, called frameshift by Clockwork, shows similar performance to the WT control constituting a false positive. Samples A8 and D3 perform as expected and were called as a frameshift and unedited genotype respectively.

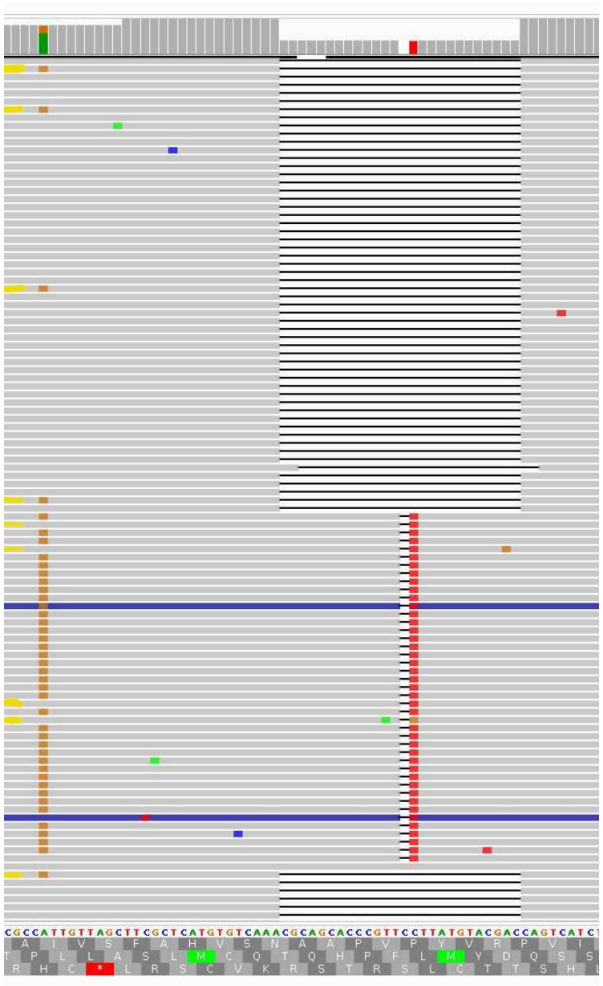

Fig. 7: Visualisation of the read alignments by Integrated Genome Viewer (IGV) for sample B6 (Fig.6). The single nucleotide polymorphism discerning the alleles is present in the 5th nucleotide and is coloured brown for *Mus spretus*. Reads that likely originate from chimeric PCR templates are marked by a yellow highlight.

4 Discussion

During development of Clockwork and its subsequent interrogation by the two datasets several possible improvements and challenges were discovered. Chimeric reads, primarily caused by the two step PCR library preparation, can be avoided by changing the protocols annealing time and number of cycles (Kanagawa, 2003). Indels larger then 100bp causing the deletion of the primer binding site used during library preparation can lead to the absence of either or both alleles in the sample. Whilst both chimeric reads and absence of alleles is recognizable from the sequencing data they reduce the effectiveness of Clockwork as a genotyping tool, requiring manual investigation of the data to resolve the ambiguities. Clockwork relies on the overlapping forward and reverse reads to construct a consensus sequence. Whilst effective at removing sequencing artefacts, 3' read quality trimming can lead to gaps in the coverage of the consensus sequence. Overall these challenges can be addressed by increasing the read-length to 250 bp allowing for more options for primer design, increased overlap of the reads and avoiding the deletion of the primer binding site. Whilst this sequencing chemistry is more expensive than the 150bp variant, multiplexing and barcoding allows for the massively

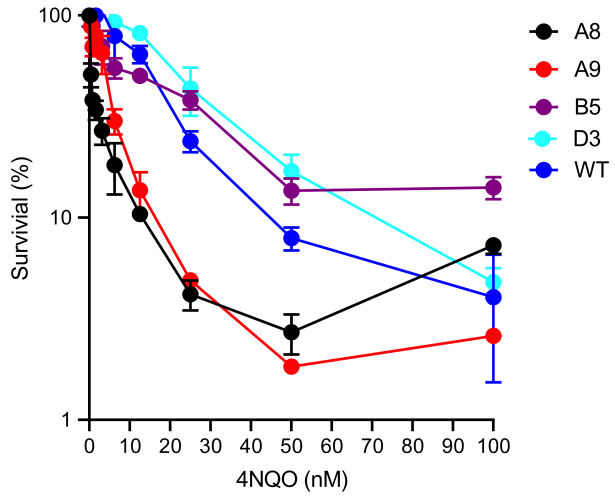

Fig. 8: Growth inhibition assay plots for a set of samples from the human RPE-1 dataset representing several types of frameshift and in-frame deletions. The assay was performed on a range of concentrations of 4NQO, a carcinogenic inducing bulky adducts. The experiment was performed with 3 replicates per sample.

parallel sequencing of many samples reducing the costs per sample. Alternatively, repeating the interrogation with long-read sequencing could be an interesting follow up. The relatively higher error rate of long-read sequencing compounds with the challenging alignment of indels in short amplicons, possibly invalidating the approach if the constructed consensus sequence cannot address sequencing artifacts sufficiently.

Verifying Clockwork's predictions through growth inhibition assays turned out to be challenging. As a bioinformatician this was my first introduction to working in a wet-lab and performing experiments. Unfortunately my three attempts at performing a growth inhibition assay failed due to difficulties with measuring cell concentrations accurately. Luckily the second author had performed a growth inhibition assay in parallel. Therein two false predictions were found, in sample A9 and B5. In the first case, closer inspection of the read alignment shows three signals of possible allele's, a 10bp indel at 50% of reads, a wild-type sequence at 24% of reads and a 1bp indel at 9% of reads. Since A9's survival rate is worse then the wildtype control, the wildtype sequence could have arisen by contamination of the two step PCR amplification. In the case of B5, there are two clear signals, a 1bp deletion and a 1bp insertion. Further inspection of the data does not indicate a clear cause for the false positive, leaving mutations outside the amplicon or contamination of the assay as a possible explanation.

Overall, Clockwork addresses a need for high-throughput genotyping in experiments applying CRISPR-Cas9 to induce frameshift mutations. We have seen the efficiency of CRISPR-Cas9 can vary wildly between cell types and guide sequences causing samples to be frequently unedited. Genotyping of samples is relatively cheap compared to the costs of whole genome sequencing, saving significant resources sequencing unedited samples.

Table 2. Primer Sequences for human RPE-1

Gene Specific Forward	GTAAACGACGGCCAGTTAAAAGTCCACCCTGTACCAC
Barcoding Forward	CGATNNNNNNNNGTAAACGACGGCCAGT
Reverse 1	TTCAACTCCAGAACAGCACACT
Reverse 2	TCCAGGTCGGAGGGATGGA

Table 3. Primer Sequences for mouse ESCs

Gene Specific Forward	GTAAAACGACGGCCAGAGGCTACGTAGAGCCAATGC
Barcoding Forward	CGATNNNNNNNNGTAAAACGACGCCAGT
Reverse 1	AACCAGATGTAAGTCTAGGACT
Reverse 2	AGGTTTACTGCACGTGAAAC

5 Methods

5.1 Two step PCR amplification

Human retinal pigment epithelium (RPE-1) cells as well as hybrid *Mus musculus* and *Mus spretus* embryonic stem cells (ESCs) were transiently transfected with Cas9 and small guide RNA (sgRNA) sequences targeting two genes of interest. 48 hours later the cells were FACS sorted, isolating GFP+ transfected cells. Two 96-well plates containing single-cell derived clones of the aforementioned cell lines were PCR amplified. The PCR was performed in two steps, first amplifying the locus of interest with the gene specific forward primer and the reverse 1 primer. Then the PCR product was re-amplified using the barcoding forward primer and the reverse 2 primer. The gene specific forward primer and barcoding forward primer share the M13 universal sequence, extending the amplicon with a unique barcode sequence represented as 8N in Tables 2 and 3.

5.2 Trimming of adapters and bad quality basepairs

Raw reads from the sequencer are trimmed for adapter sequences and bases with unacceptable quality. Often these sequences occur at the 5' and 3' side of the read respectively. Cutadapt (Martin, 2011), a commonly used library for read trimming, was supplied with Illumina adapter sequences and a quality value of 5. With larger quality values the quality trimming is more aggressive, 5 was chosen to conservatively trim 3' read bases. The effectiveness of adapter and quality trimming was assessed by FastQC and MultiQC after processing by Cutadapt (Ewels *et al.*, 2016).

5.3 Barcoding and demultiplexing

Samples in the experiment were barcoded to allow for simultaneous sequencing during a single run on the Illumina sequencer. Barcoding was applied using in-house primers containing barcode sequences during PCR amplification of the target locus. After sequencing the reads were processed by Cutadapt which after aforementioned trimming and quality control separated reads by barcode sequence (Martin, 2011). A maximum error rate of 15% of the barcode sequence was set to allow for sequencing errors in the barcode sequence. Reads for which no barcode sequence was present are removed before upstream analysis.

5.4 Alignment by CRISPResso2

Reads are aligned against the supplied reference sequence by CRISPResso2[4]. CRISPResso2 specialises in aligning sequencing of genome editing experiments by cleaving nucleases like CRISPR-Cas9. CRISPResso2 contains a modified implementation of Gotoh's algorithm (Gotoh, 1982), where the affine gap penalty is extended by a gap incentive vector. The gap incentive vector promotes aligning an indel at the cut site of the CRISPR-Cas9 complex. Cut sites are defined by supplying the guide sequence for the sample and creating a 'window' of base pairs around the cut site for which the gap incentive vector will be positive. Window sizes are set to 2bp for paired analysis and 10bp for forward/reverse read analysis. The middle of the window corresponds to the cut site defined by the guide sequence. After alignment CRISPResso2 uses the supplied reference, guide and exon sequences to quantify insertions, deletions and substitutions. These quantifications are stored in table format as well as visualised.

5.5 Quantify frameshifts with Python's Pandas

Python scripts read CRISPResso2's various output files into tables using Pandas (pandas devteam, 2020). For each sample these files are processed applying a set of thresholds to ensure confidence in frameshift calls. Firstly any sample must contain at most 5% of reads with a wildtype sequence, avoiding calling frameshifts in samples with a significant wildtype signal. Secondly the sample must have at least 1000 reads to ensure sufficient confidence and coverage in the alignment. Lastly if more than 10% of the reads for each allele contains an in-frame variant the sample is assumed to be fit. In the case where the change in exon sequence is not a multiple of 3, the read contains a frameshift mutation. Otherwise, where the change in exon sequence is a multiple of 3, the read contains an in-frame mutation. The quantification of the frameshift, the length and type of mutation for each allele and overall alignment parameters are saved for upstream visualisation.

5.6 Snakemake

Snakemake (Mölder *et al.*, 2021), a library for automating scientific workflows, automates the pipeline. By defining rules that encode the workflow steps the pipeline is capable of reproducing and scaling data analysis to any new query or data set size. The rules are split into separate sets, with each set representing a discrete step in the pipeline. By changing the configuration and supplying a new reference sequence any locus targeted by CRISPR-Cas9 can be analysed, detecting frameshifts and visually reporting the results. The configuration allows for custom barcode sequences, defining alleles for each sample, preferentially analysing forward, reverse or paired reads as well as simultaneously processing multiple reference genomes and/or samples. Furthermore data analysis is supported by snakemake's logging, automatically deleting partial files upon fault and the capability of restarting a partially executed analysis.

5.7 Integrated Genome Viewer

To further investigate the benefits of CRISPResso's gap incentive vector all samples have been aligned against their respective references using bwa mem on default settings (Li and Durbin, 2009). The resultant alignment was automatically visualised and saved using the Integrated Genome Viewer (IGV) bash script option.

5.8 PyPDF and GGplot2

A combination of python and R scripts and libraries read and visualise CRISPResso2's quantification. Pandas reads the data into a dataframe table, then mutates into the proper format for visualisation by R's GGplot2 library (Wickham, 2016). For each sample a set of graphs is created. The set of graphs are then merged into a summary pdf file specifically for that sample by PyPDF (Fenniak *et al.*, 2022). All samples that confidently contain a frameshift mutation are aggregated into a super summary pdf for printing and ease of use in the wet lab. A 96-well plate visualisation showing coloured spheres representing alleles and their respective mutations allows for easy adaption into existing wet lab protocols.

5.9 Anaconda

The pipeline is stored alongside an anaconda environment file describing the various libraries and their versions used within the pipeline. Upon cloning the git storing ClockWork the file can be used to reconstruct the environment allowing for reproducibility and consistency. Anaconda is set to recent versions of libraries from both conda-forge and bioconda.

5.10 Growth inhibition assay

5 samples were chosen for a growth inhibition assay from the human RPE-1 plate selecting for various patterns of mutation (Table. 1). Cells were grown on medium, then diluted to a concentration of 15.000 cells per 1 mL. 96 well plates were prepared with a range of concentrations of 4NQO (Bailleul *et al.*, 1989), a carcinogenic inducing bulky adducts. Wells were seeded with 100 µl of the cell suspension, with three technical replicates for each sample. After 72 hours of incubation, the medium was removed and replaced with CellTiter-Glo (Hannah *et al.*, 2001). Subsequently the plate was mixed on an orbital shaker and imaged on a plate reader. Lastly the values were normalised from blank measurements and plotted as a percentage of the untreated well.

6 Acknowledgements

I would like to thank the second author Job Smink for his invaluable assistance in learning proper wet lab protocol for cleanliness, cell culturing, survival assays and working in a ML-lab and the team at the Garaycoechea group for their welcoming attitude and support in understanding the biochemical processes behind mutational signatures.

7 Code Availability

A snakemake pipeline was written for this paper. The pipeline is published on github at https://github.com/Bryan-vd-Brand/. The supplementary files used are also available alongside the pipeline.

8 Supplementaries

Large files like the barcode, reference and exon sequences, the anaconda environment specification file and the snakemake file can be found on github. See the code availability section.

Table 4. small RNA guide sequences

human	ACGTGCGCGAGGTCTACCCCG
mouse	ACTGGTCGTACATAAGGAAC

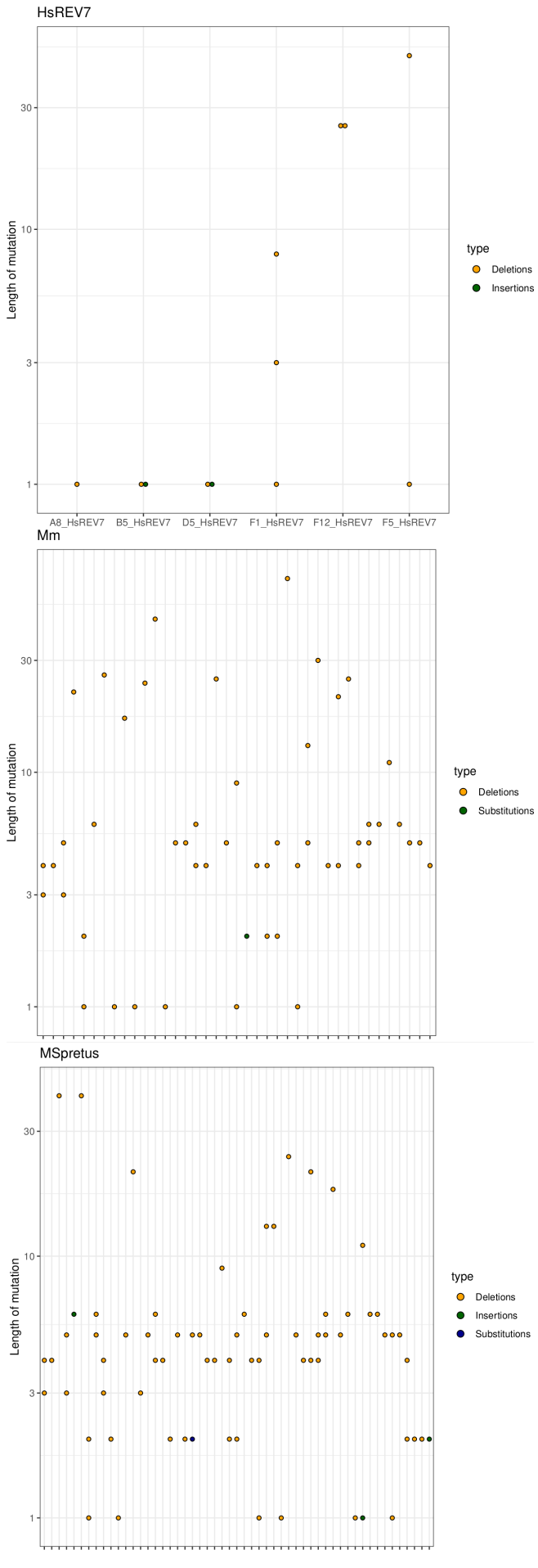

Fig. 9: Length of mutation determined by clockwork, coloured by mutation type, plotted for each allele

References

Alexandrov, L. *et al.* (2015). Clock-like mutational processes in human somatic cells. *Nature genetics*, **47**.

Allen, F. *et al.* (2019). Predicting the mutations generated by repair of cas9-induced double-strand breaks. *Nat Biotechnol*, **37**, 64–72.

Bae, T. *et al.* (2022). Analysis of somatic mutations in 131 human brains reveals aging-associated hypermutability. *Science*, **377**(6605), 511–517.

Bailleul, B. *et al.* (1989). Molecular basis of 4-nitroquinoline 1-oxide carcinogenesis. *Japanese Journal of Cancer Research*, **80**(8), 691–697.

Campbell, P. *et al.* (2020). Pan-cancer analysis of whole genomes. *Nature*, **578**, 82–93.

Canver, M. *et al.* (2017). Characterization of genomic deletion efficiency mediated by clustered regularly interspaced short palindromic repeats (crispr)/cas9 nuclease system in mammalian cells. *Journal of Biological Chemistry*, **292**, 2556–2556.

Chang, K. *et al.* (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, **45**, 1113–20.

Chatterjee, N. and Walker, G. C. (2017). Mechanisms of dna damage, repair, and mutagenesis. *Environmental and Molecular Mutagenesis*, **58**(5), 235–263.

Clement, K. *et al.* (2019). Crispresso2 provides accurate and rapid genome editing sequence analysis. *Nature Biotechnology*, **37**, 224 – 226.

Cortez, D. (2019). Replication-coupled dna repair. *Molecular Cell*, **74**(5), 866–876.

Edelbrock, M. A. *et al.* (2013). Structural, molecular and cellular functions of msh2 and msh6 during dna mismatch repair, damage signaling and other noncanonical activities. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, **743-744**, 53–66. DNA Repair and Genetic Instability.

Ewels, P. *et al.* (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, **32**(19), 3047–3048.

Fenniak, M. *et al.* (2022). The PyPDF2 library.

Gonzalez, V. *et al.* (2021). Accurate genomic variant detection in single cells with primary template-directed amplification. *Proceedings of the National Academy of Sciences*, **118**.

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, **162**(3), 705–708.

Hannah, R. *et al.* (2001). Celltiter-glo™ luminescent cell viability assay: a sensitive and rapid method for determining cell viability. *Promega Cell Notes*, **2**, 11–13.

Kanagawa, T. (2003). Bias and artifacts in multitemplate polymerase chain reactions (pcr). *Journal of bioscience and bioengineering*, **96**, 317–23.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**(14), 1754–1760.

Liu, J. *et al.* (2014). Extensive recombination due to heteroduplexes generates large amounts of artificial gene fragments during pcr. *PloS one*, **9**, e106658.

Lodato, M. A. *et al.* (2018). Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science*, **359**(6375), 555–559.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**(1), 10–12.

Martin, S. K. and Wood, R. D. (2019a). DNA polymerase ζ in DNA replication and repair. *Nucleic Acids Research*, **47**(16), 8348–8361.

Martin, S. K. and Wood, R. D. (2019b). DNA polymerase ζ in DNA replication and repair. *Nucleic Acids Research*, **47**(16), 8348–8361.

Mölder, F. *et al.* (2021). Sustainable data analysis with snakemake [version 2; peer review: 2 approved]. *F1000Research*, **10**(33).

Nik-Zainal, S. *et al.* (2012). Mutational processes molding the genomes of 21 breast cancers. *Cell*, **149**(5), 979–993.

Omelina, E. *et al.* (2019). Optimized pcr conditions minimizing the formation of chimeric dna molecules from mpra plasmid libraries. *BMC Genomics*, **20**, 536.

pandas devteam, T. (2020). pandas-dev/pandas: Pandas.

Tate, J. G. *et al.* (2018). COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, **47**(D1), D941–D947.

van Overbeek, M. *et al.* (2016). Dna repair profiling reveals nonrandom outcomes at cas9-mediated breaks. *Molecular Cell*, **63**(4), 633–646.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.