

Article

Hierarchical clustering of DNA k-mer counts in RNAseq Fastq files identifies sample heterogeneities

Wolfgang Kaisers^{1,3*} , Holger Schwender² and Heiner Schaal³ 

¹ Department of Anaesthesiology, HELIOS University Hospital Wuppertal, University of Witten/Herdecke, Heusnerstr. 40, 42283, Wuppertal, Germany

² Mathematisches Institut, Heinrich-Heine-Universität Düsseldorf; schwender@math.uni-duesseldorf.de

³ Institut für Virologie, University Hospital Düsseldorf, Heinrich Heine University, Düsseldorf; schaal@uni-duesseldorf.de

* Correspondence: wolfgang.kaisers@gmail.com; Tel.: +49-202-896-1641

Abstract: We apply hierarchical clustering (HC) of DNA k-mer counts on multiple Fastq files. The tree structures produced by HC may reflect experimental groups and thereby indicate experimental effects, but clustering of preparation groups indicates the presence of batch effects. Hence, HC of DNA k-mer counts may serve as a diagnostic device. In order to provide a simple applicable tool we implemented sequential analysis of Fastq reads with low memory usage in an R package (seqTools) available on Bioconductor. The approach is validated by analysis of Fastq file batches containing RNAseq data. Analysis of three Fastq batches downloaded from ArrayExpress indicated experimental effects. Analysis of RNAseq data from two cell types (dermal fibroblasts and Jurkat cells) sequenced in our facility indicate presence of batch effects. The observed batch effects were also present in reads mapped to the human genome and also in reads filtered for high quality (Phred > 30). We propose, that hierarchical clustering of DNA k-mer counts provides an unspecific diagnostic tool for RNAseq experiments. Further exploration is required once samples are identified as outliers in HC derived trees.

Keywords: DNA-kmer; Fastq;RNAseq

1. Introduction

DNA k-mer's are DNA subsequences of length k. The analysis of DNA k-mer counts constitutes the (word) frequency based methods of *alignment free sequence comparison* which are applied in wide areas of DNA sequence analysis (see [1–3] and references therein).

1.1. Analysis of k-mer counts

Word based analysis of DNA sequencing data is utilised for isoform quantification in RNAseq (Sailfish [4]), genome assembly (Velvet [5] or Celera [6,7]), detection and correction of sequencing errors [8,9] and metagenomics (CLARK [10]). The quality of sequence assembly can be improved by k-mer based filtering of reads [11,12].

Counting of small word sizes ($k \leq 10$) can be done offhand, whilst analysis of longer motifs usually requires elaborate strategies due to computational and storage demands. Standard approaches in this area include hash tables (Jellyfish [2]), bloom filter based methods (BFCounter [13]) and disk based implementations (KMC[14], KMC2 [15] and DSK [16]), but numerous other methods have been described [17].

1.2. Quality control of RNAseq data

Analysis of RNAseq data usually includes quality control of raw reads, which includes counting of DNA k-mer's [18]. FastQC [19], a widely used quality control software, for example indicates

over-represented k-mer's in a FASTQ file by counting 7-mers on 2 % of the contained reads in order to indicate expansive read duplication.

The comparison of DNA k-mer spectra over multiple samples for quality control of RNAseq data was initially described in a GEUVADIS report[20]. Subsequently, a *k-mer Profile Analysis Library* (kPAL), implemented in Python, has been published [21].

1.3. Word sizes

Counting of DNA k-mer's results in vectors of length \mathbb{N}^n ($n = 4^k$) indicating, that 4^k words exist. Small values of k ($2 \leq k \leq 4$) diminish discrimination capabilities of counting results. At intermediate values of k ($9 \leq k \leq 15$), unique k-mer's and *nullomers* (absent k-mer's) begin to appear as consequence of the limited complexity of the human genome[21]. At larger values of k ($k > 30$), the k-mer count vectors become sparse (as only a small fraction of k-mer's are present) [22].

Word sizes (k) used in DNA k-mer analysis depend on application type, for example $k \approx 15 - 30$ in metagenomics [10], $k \approx 20 - 80$ in genome assembly [23], $k \approx 20$ isoform quantification [4].

In comparison of DNA k-mer spectra over multiple RNAseq samples, the word size was $k=9$ in the initial analysis [20] and $k=12$ using smoothed k-mer count profiles [21].

The word size determines the specificity of the sampled sequences. While k-mers from small k presumably will be present in almost all sequencing samples, long k-mers ($k > 20$) become increasingly specific for species or proteins. As HcKmer analysis of all biological samples in this study was done using $k=9$ which can be specific for at most 3 amino acids, diagnostic criteria derived from HcKmer analysis are regarded as unspecific for presence of biological entities.

1.4. Analysis of DNA k-mer counts

The vector representation $v \in \mathbb{N}^n$ ($n = 4^k$) of DNA k-mer counts allows application of analytic procedures provided by euclidean geometry (for example distance measures [1]) and machine learning algorithms, for example principal component analysis (PCA) or hierarchical clustering (HC).

In hierarchical clustering (HC), different entities are located in bi-parting trees according to their pairwise similarity (quantified by a distance measure). Although the trees provide no absolute measure, accumulation of biological or technical related samples in different sub-trees may reveal relevant heterogeneities. Observed sample (dis-)similarities in DNA k-mer counts have been shown to be indicative of problematic samples (for example due to read duplication or presence of rRNA)[21].

1.5. HcKmer analysis algorithm

We implemented a k-mer counting algorithm in C and provide a programming interface allowing to run the complete analysis inside R. The software is available as R package *seqTools* for download from Bioconductor. The Canberra distance is utilised as distance on DNA k-mer counts.

1.6. Analysed samples

Three sample batches were downloaded from ArrayExpress (accession E-MTAB-4842, E-MTAB-4104 and E-MTAB-691), each containing two treatment groups. Second, a batch of 61 samples, sequenced in our facility and containing RNAseq data from two tissues, was analysed. The first moiety consists of 57 dermal fibroblast samples (ArrayExpress accession E-MTAB-4652); the second comprises 4 Jurkat cell samples.

2. Results

2.1. Data collection

Fastq files from the three ArrayExpress batches were downloaded and data from each experiment was collected into one separate data-set.

The 61 samples had been sequenced in our local facility on 8 Illumina Flowcells. Data from each Flowcell was collected into one data-set; the processing of 9.8×10^9 contained reads took 8.96 hours (3.04×10^5 reads/second) in a single thread with approximately 1 Gigabyte working memory consumption (see section 1.1 of supplemental material for more details).

2.2. Detection of experimental effects

The separation of experimental groups by HcKmer is exemplified by analysis of three sequencing datasets, downloaded from ArrayExpress [24]. The first dataset has been created by RNAseq of mouse liver endothelial cells from normal (4 samples) and tumor infiltrated tissue (4 samples) [25]. Tumor growth was induced by injection of B16-F10 melanoma cells into the portal vein. Figure 1 shows a dendrogram where all samples derived from normal tissue cluster together in one sub-tree. Differential gene expression analysis revealed up-regulation of glycolytic pathways in tumor derived endothelial cells.

In the second example, RNAseq data of KBM-7 chronic myelogenous leukaemia cells was analysed. Three samples KBM-7 cells and three samples of KBM-7 cells with knockout of NUDT2 (a NUDIX hydrolase [26]) have been analysed for differential gene expression. Figure 2 shows a dendrogram where samples with and without knockout of NUDT2 cluster in different sub-trees. The analysis revealed 6,288 differential expressed genes and 1,685 genes with a fold-change ≥ 2 [27].

The third example shows analysis of Fastq files derived from whole genome sequencing. Two pairs of cell lines derived from high-grade ovarian carcinoma of two patients were analysed. Two cell lines, PEO1 obtained during the first relapse (fr) and platinum sensible (sen), and PEO04 obtained during the second relapse (sr) and platinum resistant (res) were derived from patient 2 (p2). Two cell lines, PEO14 obtained during initial presentation (pr) and platinum sensible (sen), and PEO23 obtained during relapse (re) and platinum resistant (res) were derived from patient 3 (p3). Figure 3 shows a dendrogram where samples from both patients are located in different sub-trees and additionally, platinum sensible and resistant samples also are located in different sub-trees. The analysis revealed structural variants appearing during tumor evolution, for example different tandem duplications in samples PEO14 and PEO23 [28].

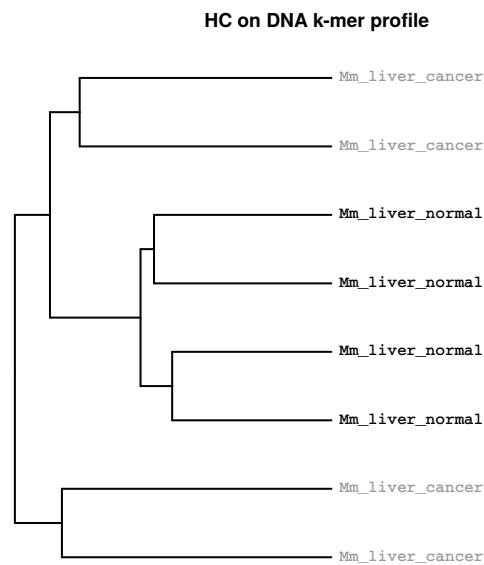


Figure 1. Clustering according to experimental groups. B16-F10 melanoma cells were injected into portal veins in mice resulting in intrahepatic tumor growth. After 14 days, endothelial cells from hepatic tumors and from livers of healthy mice were extracted and analysed with RNAseq (single end). Differential gene expression had been focused on 1,255 metabolic genes expressed in endothelial cells. The sample data was downloaded from ArrayExpress (accession E-MTAB-4842).

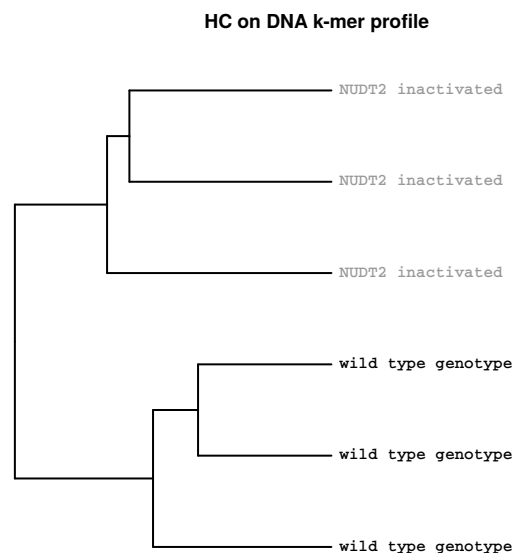


Figure 2. Clustering according to experimental groups. KBM-7 chronic myelogenous leukaemia cells with and without knockout of gene NUDT2 were analysed using RNAseq. The sample data was downloaded from ArrayExpress (accession E-MTAB-4104, only the first of the Fastq files from paired end sequencing (*_1.fastq.gz) are included).

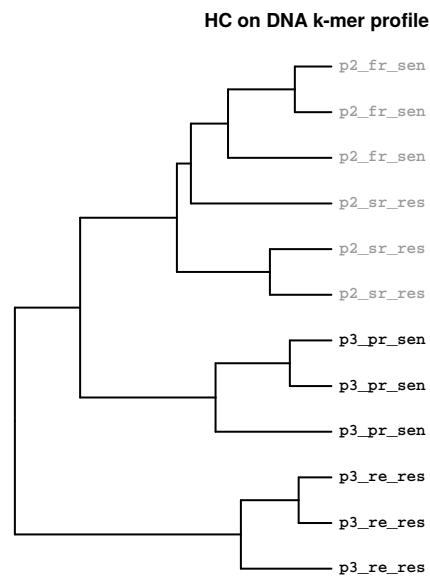


Figure 3. Clustering according to experimental groups. High grade serous ovarian carcinoma cells from two patients were characterised using whole genome sequencing. Leaf labels indicate patient (p2 and p3), clinical status (fr=first relapse, sr=second relapse, pr=primary presentation, re=relapse) and platinum sensibility status (sen=sensible, res=resistant). Samples with identical leaf labels indicate samples obtained from the same cell line. The sample data was downloaded from ArrayExpress (accession E-MTAB-691, only the first of the Fastq files from paired end sequencing (*_1.fastq.gz) are included).

2.3. Identification of batch effects.

Figure 4 shows a dendrogram where 16 samples from two Flowcells (*d24a* and *c0yr*) and two cell types (dermal fibroblasts and Jurkat cells) are analysed using HcKmer. The partition tree consists of two equally sized sub-trees. Each of them only contains samples sequenced on one Flowcell although on Flowcell *c0yr*, two different cell types are present. Thus, DNA motif dissimilarity is greater between Flowcells *d24a* and *c0yr* than between fibroblasts and Jurkat cells.

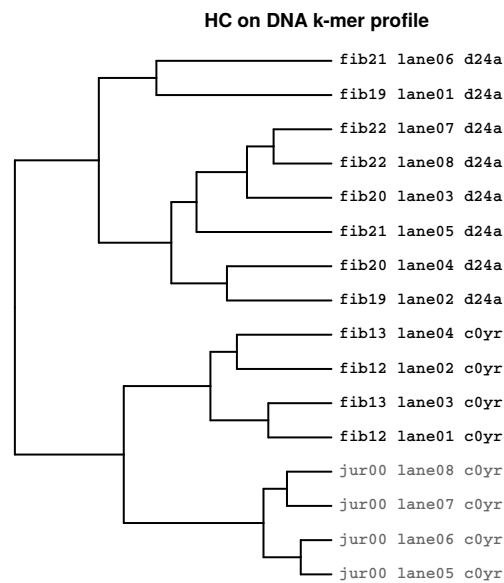


Figure 4. Batch effect indicated by HcKmer on samples sequenced in two Illumina Flowcells: Leaf labels denote cell type (fib, jur) and number of individual, lane number and Flowcell label (*d24a* and *c0yr*). Samples from Jurkat cells are highlighted in grey. The tree clearly separates Flowcell *d24a* and *c0yr* although Flowcell *c0yr* contains two different cell types.

Figure 5, indicates that median Phred scores in sequences from both Flowcells are sufficiently high (see also section 1.4 of supplemental material)

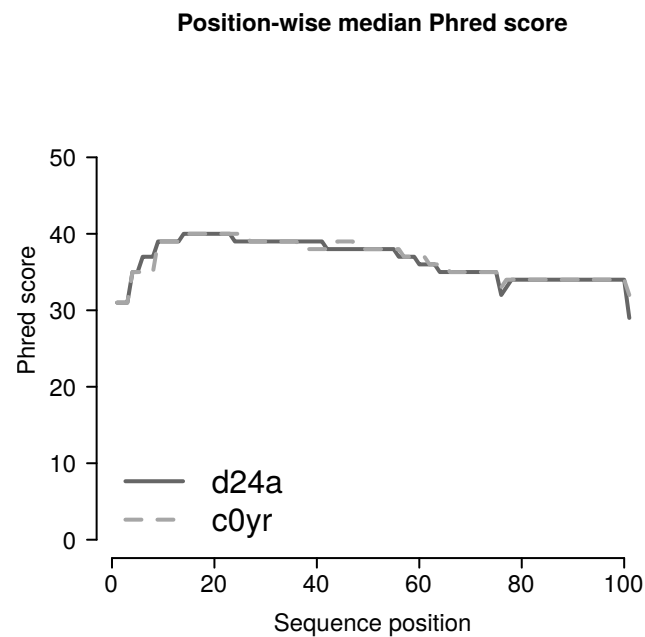


Figure 5. Median Phred score values: For each read position, median Phred scores are shown for samples sequenced on Flowcells *d24a* and *c0yr*. All median Phred scores are > 28 .

2.3.1. Filtering for sufficient Phred scores

In order to explore, whether the observed dissimilarity can be removed, reads containing at least one Phred score < 30 were discarded. The filter was applied to Fastq files sequenced on *d24a* and *c0yr* Flowcells. From Flowcell *d24a*, 14.3 % of reads in Fastq files and from Flowcell *c0yr*, 15.1 % of reads in Fastq files were excluded thereby. Finally, the filtered reads were re-analysed. In the HC-tree, the position of one Flowcell (*fib13 lane03 c0yr*) changed to the opposite sub-tree but still, three fibroblast samples cluster together with the Jurkat samples (Figure 6).

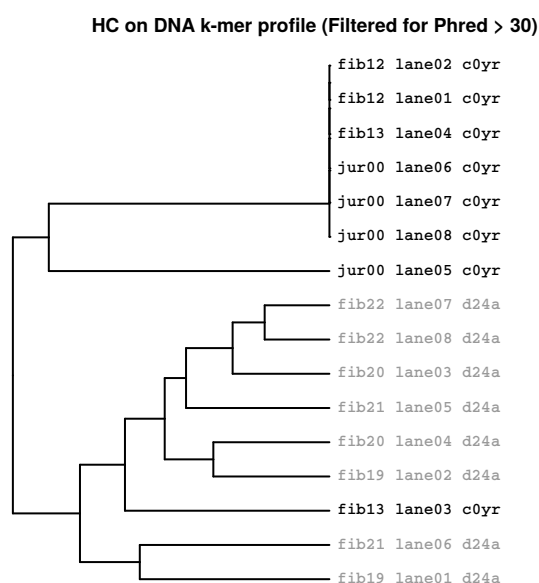


Figure 6. Clustering of Fastq files containing filtered reads: All reads containing Phred scores < 30 had been discarded before HcKmer analysis. On top level clade, the Jurkat cell samples still exclusively cluster together with samples from the same Flowcell.

Thus, filtering based on Phred scores induced only a minor change in cluster formation.

2.3.2. HcKmer on sequences aligned to the human genome

Fastq reads causing HcKmer tree separation may not match to the reference genome and thus would be filtered out by alignment. In order to evaluate this eventuality, clustering of raw Fastq files and mapped reads are compared. Pairs of Flowcells containing 8 fibroblast samples (*d24a*, *d10r*, *c0g9*, *c0yt*, *c2uk*, *d1pd*) are analysed by HcKmer on raw reads and on mapped reads. An example is shown in Figure 7 where raw and mapped reads cluster in similar patterns. In both trees, the size of the largest sub-tree containing samples from only one Flowcell is 7. The mean sizes of these sub-trees in all 15 possible pairs is 7.21 in raw reads and 6.57 in mapped reads. Mapping thus reduces Flowcell cluster sizes by 8.9 % but still, clusters of size ≥ 6 are prevalent. Detailed results are shown in supplemental material (section 2).

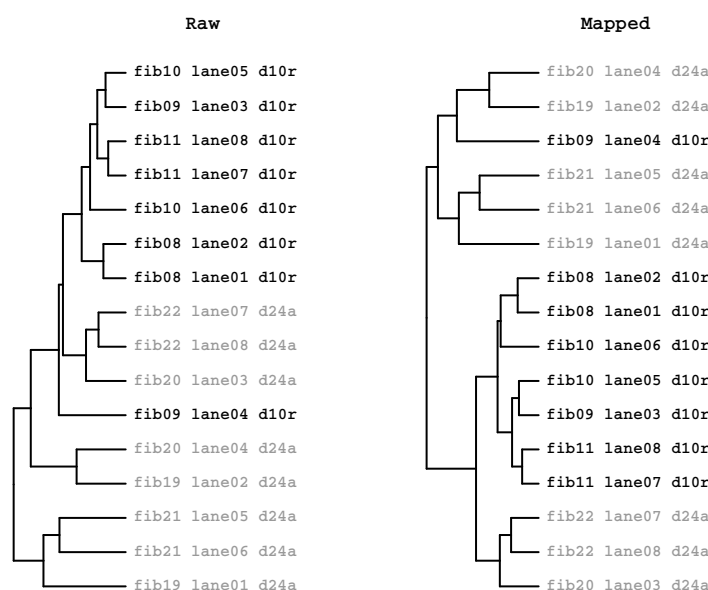


Figure 7. HcKmer on raw and mapped reads: Fastq files containing mapped reads were constructed from raw reads by alignment using TopHat followed by extraction of reads from BAM files using *reader2fastq* (rbamtools) function. HcKmer on raw and mapped reads was calculated on Fastq files from the same Flowcell pairs. Raw reads consist of unmapped and mapped reads. **Left:** HcKmer on raw reads. **Right:** Mapped reads. Leaf labels denote cell type (fib), lane number, and Flowcell label (*d24a* and *d10r*). All but one samples from Flowcell *d10r* (dark grey) cluster within a separate sub-tree. Raw and mapped reads show similar clustering characteristics.

2.3.3. Prevalence of batch effects in RNAseq data

The prevalence of detectable batch effects was analysed on the whole set of 61 samples sequenced on 8 Flowcells. Pairwise comparison of 8 Flowcells results in 28 pairs. All Flowcell pairs are analysed for presence of batch effects using HcKmer and a semi-quantitative score ranging from strong batch effect (b1a = top-level separation of Flowcells; shown in Figure 4) and detectable batch effect (b1b, b2a or b2b) to absence of batch effects (es). From the analysed Flowcell pairs, 6 (21.4 %) show strong batch effects and 17 (60.7 %) show detectable batch effects. Thus, batch effects are present in a considerable fraction of Flowcell pairs. The definitions and details on analysis are shown in supplemental material (section 4).

2.3.4. Influence of batch effects on false discovery rate

Clustering according to sample preparation batches potentially influences differentially Expressed Gene (DEG) analysis in RNAseq data. To address this question, results from DEG analysis are compared between Flowcells with different HcKmer dissimilarity using a two-way ANOVA. Classification as batch effect (b1 or b2 vs. es) is significantly associated with increased number of differentially expressed (DE) genes (see section 5 of supplemental material for details). The ANOVA predicts 3.848 DE genes for b1a dissimilarity and 695 DE genes for idf/es (no batch effect detectable). Thus, increased number of false positives in DEG analysis are found when batch effects are identified by HcKmer.

2.4. Separation sensitivity on simulated data

The degree of sequence dissimilarity required to produce HcKmer trees with top level separation of two sample groups (e.g. Flowcells) is determined in simulation where Fastq files with random DNA are analysed. A pure random group is compared to a group in which a variable percentage of random DNA reads is contaminated with (one or multiple) fixed DNA 6-mers.

Clustering of groups is quantified using a score (Contralaterality Score, CS). A decrease of CS from 41.1% (mean value for pure random sequences) to 0% is regarded as indicative for clustering of groups and in the setting of the simulation, a CS < 10 % is considered to be statistically significant ($p < 0.05$).

Figure 8 shows separation capabilities of HcKmer for contamination of 0 - 6 % of Fastq reads with one fixed DNA 6-mer. The results indicate that with a contamination of 4 % of Fastq reads, significant sample separation reaches a power of 80 %. Details of simulation, definition of CS and results are shown in supplemental data (section 3).

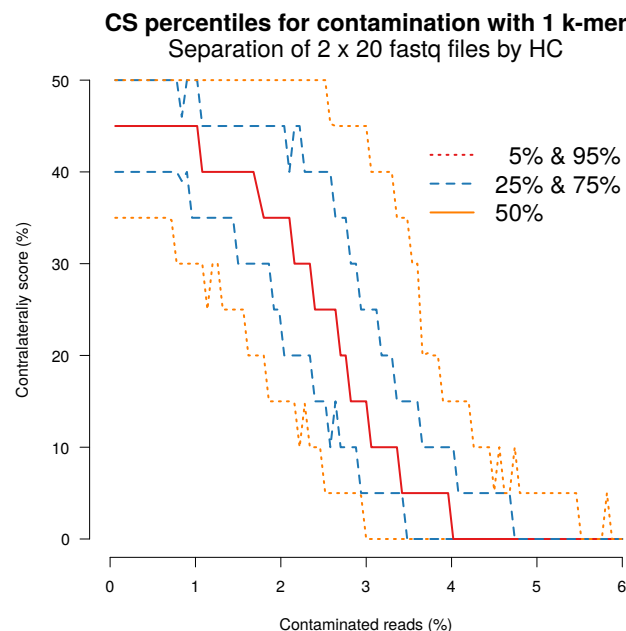


Figure 8. Separation sensitivity on simulated data: HcKmer was performed on Fastq files containing simulated DNA sequence ($k=6$). Percentiles for Contralaterality score (CS) are given for variable percentages of contamination with a single fixed DNA 6-mer. CS quantifies the presence of the minor present group in the first half of the HC-derived group labels. CS values <10 % are considered to be statistically significant ($p < 0.05$).

2.5. K-mer spectrum responsible for tree separation

Standard quality control tools (for example FastQC) only report a small number of over-represented k-mers which may not represent a sufficient sample for explanation of batch effects diagnosed by HcKmer. We therefore estimate how many k-mers are needed in order to evoke the observed batch effects using an example.

From the simulation data on separation sensitivity, it is assumed that $\approx 3\%$ contamination (where the median CS falls below 12.5 % in Figure 8) is required in order to produce a strong batch effect (b1a). Two pairs of Flowcells are selected, one with complete tree separation due to batch effect (Flowcells *d24a* and *c0yr*) and one pair of Flowcells without tree separation due to batch effect (Flowcells *d24a* and *d1pd*). For each k-mer, the logarithmised and normalised sum of k-mer counts on whole Flowcells are

calculated and compared for both pairs and are shown in Figure 9). In the example shown in left panel (*d24a/c0yr* comparison), differences from the 3,678 k-mers with the largest difference in k-mer count need to be accumulated in order to attain a 3 % contamination rate.

Thus, the observed tree separations are caused by the sum of widespread small differences instead of large disparities on a small number of k-mers. Also, the read content of problematic samples seemingly cannot be easily corrected (for example by removing or clipping a small fraction of reads).

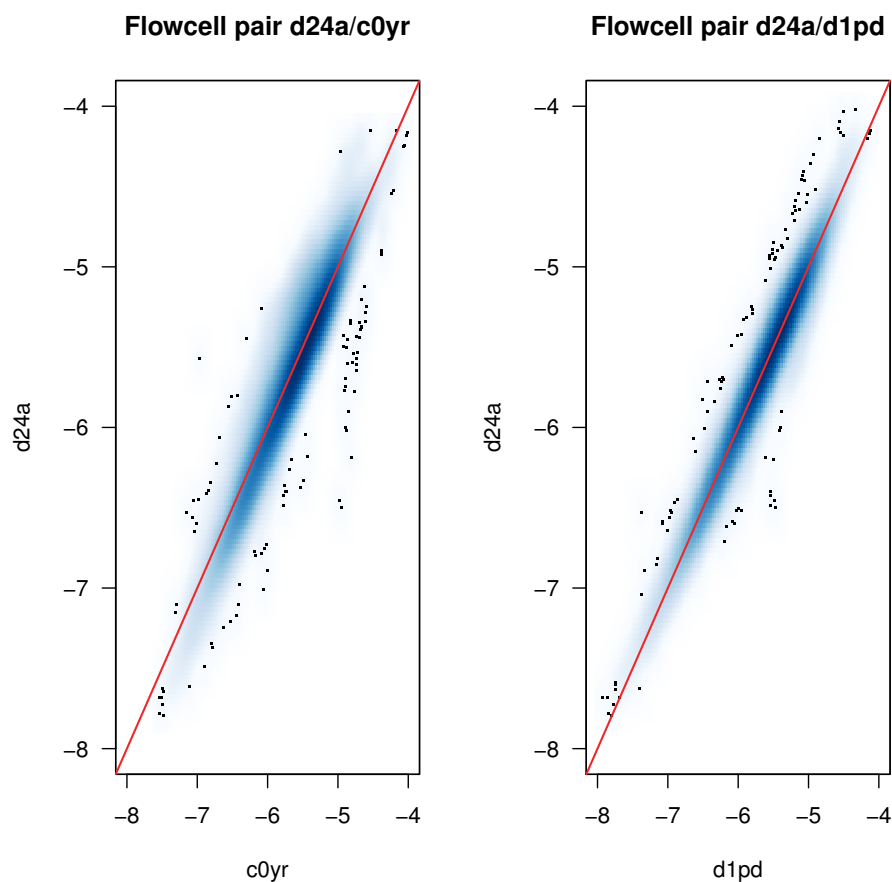


Figure 9. K-mer spectrum responsible for tree separation: Comparison of K-mer counts in two Flowcell pairs using density scatter plots: All axes represent normalised and \log_{10} transformed k-mer counts on a whole Flowcell (each 8 Fastq files). The diagonal (red line) indicates equal normalised counted numbers for k-mers in both Flowcells. **Left panel:** Comparison of k-mer counts on Flowcells *d24a/c0yr* (strong batch effect (=b1a) identified by HcKmer). **Right panel:** Comparison of k-mer counts on Flowcells *d24a/d1pd* (no batch effect (=es) identified by HcKmer). **Result:** The k-mer count differences are larger for *d24a/c0yr* (mean= 0.30, sd= 0.12) than for *d24a/d1pd* (mean=0.01, sd=0.07). The HcKmer diagnosed difference of sample similarity between the Flowcell pairs is due to larger deviation of k-mer counts from the diagonal for a broad variety of k-mers (thus not generated by a small group k-mers with large deviation).

3. Discussion

The underlying principle of HcKmer is the comparison of multiple samples using a high dimensional criterion which is closely linked to the central information content of DNA sequencing data. As such, this principle differs from standard quality control algorithms which analyse single samples, but also imposes restrictions, as a minimum of 3-4 samples per group are required for

assessment.

Clustering of samples according to biological or experimental entities is an unspecific criterion as different k-mer spectra may be caused by differential gene expression or differing splicing patterns (as biologically unspecific indicator, see 1.3) as well as the mentioned disturbing factors (read duplication or presence of rRNA [21]). Thus, etiologic factors may not be apparent.

3.1. Indication of batch effects

Transcriptome sequencing data is prone to disturbing effects resulting from experimental design for example ozone levels [29], random hexamer priming [30], GC-content [31], and transcript length [32]. Batch effects, variation due to experimental design, is a prevalent phenomenon and clustering according to surrogate values (processing date or batch) is a common approach for recognition of underlying sources of variation [33].

Analysis of our samples shows, that clustering of preparation batches may be prevalent (21 % - 61 %), and that resulting disturbances may not be removed by Phred based filters or by alignment to a genome.

HcKmer clustering according to sample preparation groups had also been demonstrated for whole genome sequencing (WGS) data as well as for whole exome sequencing (WES) data (see Figure 3 and 4 in [21]). Although unproblematic according to FastQC measures, the WGS samples exhibited significant inconsistencies in paired end read mapping suggesting, that the indicated batch effects might affect DEG analysis, which is in line with the results shown in section 2.3.4. Thus, differences identified by HcKmer should initiate exploration of underlying effects due to potentially influential effects on downstream analysis.

Due to analysis results of differential gene expression on our fibroblast samples [34], we assume that there are no consistent differences between different groups and thus, GC-content and transcript length are unlikely to evoke the observed batch effects. Also, observed sample dissimilarities seem to result from a larger number of small differences in k-mer counts which can not be diagnosed by identification of few over- or underrepresented k-mers. Based on the shown results, our group decided to discard data from samples sequenced on two other Flowcells (data not shown here) exhibiting strong batch effects when compared with all Flowcells in the shown analysis.

3.2. Application of HcKmer

Standard quality control procedures address Phred-scores, GC content, over-represented K-mers and sequence duplication [18] and are implemented in ubiquitously used software products like FastQC. The shown results indicate, that HcKmer represents a quality differing from qualities examined by standard procedures which essentially can be explained by the fact that between-sample heterogeneity is examined at high resolution.

A limitation is, that HcKmer provides no quantitative measure for similarity and that subjective judgement is required. But, as HcKmer can easily be applied (using *seqTools*), it appears advisable to include HcKmer into analysis standards.

4. Materials and Methods

4.1. Algorithmic framework

For a DNA sequence of length k , 4^k different sequence motifs (k-mers) exist each defining a category for which occurrences can be counted. The value of k is usually chosen in the range of 5 to 9 (resulting in 1,024 to 262,144 different k-mers) as shorter motifs impair separation capabilities and longer motives increase computational demands.

The algorithms for performing HcKmer are available in R package *seqTools* on Bioconductor [35]. DNA

k-mer counts are collected with other quality measures (for example Phred scores and GC content) from Fastq files. Data from multiple samples (for example from whole Illumina Flowcells) can be collected at once into merge-able objects. Merged objects then serve as input for hierarchical clustering. The Canberra distance, defined as

$$f_{cb}(x_i, y_i) = \frac{|x_i - y_i|}{|x_i| + |y_i|} \quad (1)$$

$$d(x, y) = \sum_{i=1}^k f_{cb}(x_i, y_i), \quad (x_i)_{i=1, \dots, k}, (y_i)_{i=1, \dots, k} \in \mathbb{R}^k, \quad (2)$$

is used as distance measure between samples. The Canberra distance (initially described in 1966 [36]), is a weighted version of the L_1 (Manhattan) metric which is sensitive for differences when both components (x_i, y_i) are small. Additionally, total read numbers ($\sum_{i=1}^m x_i$) are scaled before calculation of Canberra distance to a common value in order to compensate a systematic offset caused by different sequencing depth's (thereby increasing $|x_i - y_i|$ values). Subsequent cluster-analysis was calculated using the *hclust* function (CRAN *stats* package) [37]. Detailed source code used for the analysis is shown in the supplemental material (section 1.2).

Sample preparation, sequencing and alignment

In this study, transcriptome data from 61 samples is analysed. Therefrom, 57 samples of human short term cultured human dermal fibroblasts were obtained from 27 human healthy individuals and sequenced for a study on human ageing [34]. Collection and processing of dermal samples from donors was approved by the Ethical Committee of the Medical Faculty of the University of Düsseldorf (# 3361) in 2011. The Fastq files from these samples are available under ArrayExpress accession E-MTAB-4652 (ENA study ERP015294). Also, 4 samples from cultured human Jurkat cell lines were sequenced for a study on transcriptome alterations caused by HIV infection.

Sample preparation and sequencing has been described elsewhere [34]. In short, cellular mRNA was amplified on 8 Illumina Flowcells (v1.5) and sequenced on a Illumina HiSeq 2000 sequencer. From each lane, the resulting 101-nucleotide sequence reads were converted to Fastq by CASAVA 1.8.2. A Fastq file contained in average 162.2×10^6 reads. In total the 61 Fastq files contained 9.8×10^9 reads. Subsequent alignments were calculated on unprocessed Fastq files with *TopHat* (v 2.0.14) using human GRCh38 assembly.

In order to compare batch effects between raw Fastq files and mapped reads, BAM file content was transformed back into Fastq (using *bam2fastq* from CRAN package *rbamtools* [38]). All DNA k-mer counts were collected using $k=9$ ($4^9 = 262,144$ DNA motifs), except for the simulation study ($k=6$).

Differential gene expression analysis was performed using Quasi-likelihood F-Tests from the edgeR (3.12.0) framework [39]. Genes with a reported FDR < 0.1 were considered to be significantly differential expressed. The same analytic procedure performed on the whole dataset resulted in *no* significantly differential expressed gene [34].

5. Conclusions

HcKmer applies an unsupervised learning algorithm onto the raw high throughput sequencing data with the capability of detecting potentially prevalent and influential unwanted variation. Experimental designs allowing HcKmer analysis (for example defined library preparation on a defined set of Flowcells) thus are favourable. When multiple samples are sequenced on the same Flowcell, the reads should be demultiplexed before HcKmer analysis. Based on contrasts identified by HcKmer, further exploration of results from sequencing experiments as well as exclusion of contaminating samples may be reasonable.

Author Contributions: Conceptualisation, W.K.; Methodology, W.K. and H.Schw.; Software, W.K.; Validation, W.K.; Formal Analysis, W.K. and H.Schw.; Investigation, W.K. and H.Scha.; Resources, H.Scha. and H.Schw.; Data Curation, W.K.; Writing—Original Draft Preparation, W.K.; Writing—Review & Editing, W.K., H.Scha. and H.Schw.; Visualisation, W.K.; Supervision, H.Scha. and H.Schw.; Project Administration, W.K.; Funding Acquisition, H.Scha. and H. Schw.

Funding: This study was supported in part by the Deutsche Forschungsgemeinschaft (DFG, SCHW 1508/3-1) to H. Schw. and (DFG, SCHA 909/4-1) and the German Ministry of Research and Education (Network Gerontosys, Stromal Aging WP3, part C) to H. Scha.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

RNAseq	Whole transcriptome sequencing
HC	Hierarchical clustering
HcKmer	Hierarchical clustering of DNA k-mer counts
Phred score	Quality score for identification of nucleotides in sequencing data
DEG	Differently expressed gene
QC	Quality control
CS	Contralaterality Score

- Zielezinski, A.; Vinga, S.; Almeida, J.; Karlowski, W.M. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biol.* **2017**, *18*, 186.
- Marcais, G.; Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **2011**, *27*, 764–770.
- Ren, J.; Bai, X.; Lu, Y.Y.; Tang, K.; Wang, Y.; Reinert, G.; Sun, F. Alignment-Free Sequence Analysis and Applications. *Annual Review of Biomedical Data Science* **2018**, *1*, 93–114. doi:10.1146/annurev-biodatasci-080917-013431.
- Patro, R.; Mount, S.M.; Kingsford, C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.* **2014**, *32*, 462–464.
- Zerbino, D.R.; Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **2008**, *18*, 821–829.
- Myers, E.W.; Sutton, G.G.; Delcher, A.L.; Dew, I.M.; Fasulo, D.P.; Flanigan, M.J.; Kravitz, S.A.; Mobarry, C.M.; Reinert, K.H.; Remington, K.A.; Anson, E.L.; Bolanos, R.A.; Chou, H.H.; Jordan, C.M.; Halpern, A.L.; Lonardi, S.; Beasley, E.M.; Brandon, R.C.; Chen, L.; Dunn, P.J.; Lai, Z.; Liang, Y.; Nusskern, D.R.; Zhan, M.; Zhang, Q.; Zheng, X.; Rubin, G.M.; Adams, M.D.; Venter, J.C. A whole-genome assembly of *Drosophila*. *Science* **2000**, *287*, 2196–2204.
- Denisov, G.; Walenz, B.; Halpern, A.L.; Miller, J.; Axelrod, N.; Levy, S.; Sutton, G. Consensus generation and variant detection by Celera Assembler. *Bioinformatics* **2008**, *24*, 1035–1040.
- Medvedev, P.; Scott, E.; Kakaradov, B.; Pevzner, P. Error correction of high-throughput sequencing datasets with non-uniform coverage. *Bioinformatics* **2011**, *27*, i137–141.
- Kelley, D.R.; Schatz, M.C.; Salzberg, S.L. Quake: quality-aware detection and correction of sequencing errors. *Genome Biol.* **2010**, *11*, R116.
- Ounit, R.; Wanamaker, S.; Close, T.J.; Lonardi, S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics* **2015**, *16*, 236.
- Carvalho, A.B.; Dupim, E.G.; Goldstein, G. Improved assembly of noisy long reads by k-mer validation. *Genome Res.* **2016**, *26*, 1710–1720.
- Mapleson, D.; Garcia Accinelli, G.; Kettleborough, G.; Wright, J.; Clavijo, B.J. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* **2017**, *33*, 574–576.
- Melsted, P.; Pritchard, J.K. Efficient counting of k-mers in DNA sequences using a bloom filter. *BMC Bioinformatics* **2011**, *12*, 333.

14. Deorowicz, S.; Debudaj-Grabysz, A.; Grabowski, S. Disk-based k-mer counting on a PC. *BMC Bioinformatics* **2013**, *14*, 160.
15. Deorowicz, S.; Kokot, M.; Grabowski, S.; Debudaj-Grabysz, A. KMC 2: fast and resource-frugal k-mer counting. *Bioinformatics* **2015**, *31*, 1569–1576.
16. Rizk, G.; Lavenier, D.; Chikhi, R. DSK: k-mer counting with very low memory usage. *Bioinformatics* **2013**, *29*, 652–653.
17. Manekar, S.C.; Sathe, S.R. A benchmark study of k-mer counting methods for high-throughput sequencing. *Gigascience* **2018**.
18. Conesa, A.; Madrigal, P.; Tarazona, S.; Gomez-Cabrero, D.; Cervera, A.; McPherson, A.; Szczesniak, M.W.; Gaffney, D.J.; Elo, L.L.; Zhang, X.; Mortazavi, A. A survey of best practices for RNA-seq data analysis. *Genome Biol.* **2016**, *17*, 13.
19. Andrews, S. *FastQC: a quality control tool for high throughput sequence data*. Babraham Bioinformatics, 2010.
20. 't Hoen, P.A.; Friedlander, M.R.; Almlof, J.; Sammeth, M.; Pulyakhina, I.; Anvar, S.Y.; Laros, J.F.; Buermans, H.P.; Karlberg, O.; Brannvall, M.; den Dunnen, J.T.; van Ommen, G.J.; Gut, I.G.; Guigo, R.; Estivill, X.; Syvanen, A.C.; Dermitzakis, E.T.; Lappalainen, T.; van Ommen, G.J.; Estivill, X.; Guigo, R.; Syvanen, A.C.; Gut, I.G.; Dermitzakis, E.T.; Antonarakis, S.E.; Brazma, A.; Flicek, P.; Schreiber, S.; Rosenstiel, P.; Meitinger, T.; Strom, T.M.; Lehrach, H.; Sudbrak, R.; Carracedo, A.; 't Hoen, P.A.; Pulyakhina, I.; Anvar, S.Y.; Laros, J.F.; Buermans, H.P.; van Iterson, M.; Friedlander, M.R.; Monlong, J.; Lizano, E.; Bertier, G.; Ferreira, P.G.; Sammeth, M.; Almlof, J.; Karlberg, O.; Brannvall, M.; Ribeca, P.; Griebel, T.; Beltran, S.; Gut, M.; Kahlem, K.; Lappalainen, T.; Giger, T.; Ongen, H.; Padioleau, I.; Kilpinen, H.; Gonzalez-Porta, M.; Kurbatova, N.; Tikhonov, A.; Greger, L.; Barann, M.; Esser, D.; Hasler, R.; Wieland, T.; Schwarzmayr, T.; Sultan, M.; Amstislavskiy, V. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat. Biotechnol.* **2013**, *31*, 1015–1022.
21. Anvar, S.Y.; Khachatryan, L.; Vermaat, M.; van Galen, M.; Pulyakhina, I.; Ariyurek, Y.; Kraaijeveld, K.; den Dunnen, J.T.; de Knijff, P.; 't Hoen, P.A.; Laros, J.F. Determining the quality and complexity of next-generation sequencing data without a reference genome. *Genome Biol.* **2014**, *15*, 555.
22. Zhang, Q.; Pell, J.; Canino-Koning, R.; Howe, A.C.; Brown, C.T. These are not the k-mers you are looking for: efficient online k-mer counting using a probabilistic data structure. *PLoS ONE* **2014**, *9*, e101271.
23. Chikhi, R.; Medvedev, P. Informed and automated k-mer size selection for genome assembly. *Bioinformatics* **2014**, *30*, 31–37.
24. Kolesnikov, N.; Hastings, E.; Keays, M.; Melnichuk, O.; Tang, Y.A.; Williams, E.; Dylag, M.; Kurbatova, N.; Brandizi, M.; Burdett, T.; Megy, K.; Pilicheva, E.; Rustici, G.; Tikhonov, A.; Parkinson, H.; Petryszak, R.; Sarkans, U.; Brazma, A. ArrayExpress update—simplifying data submissions. *Nucleic Acids Res.* **2015**, *43*, D1113–1116.
25. Cantelmo, A.R.; Conradi, L.C.; Brajic, A.; Goveia, J.; Kalucka, J.; Pircher, A.; Chaturvedi, P.; Hol, J.; Thienpont, B.; Teuwen, L.A.; Schoors, S.; Boeckx, B.; Vriens, J.; Kuchnio, A.; Veys, K.; Cruys, B.; Finotto, L.; Treps, L.; Stav-Noraas, T.E.; Bifari, F.; Stapor, P.; Decimo, I.; Kampen, K.; De Bock, K.; Haraldsen, G.; Schoonjans, L.; Rabelink, T.; Eelen, G.; Ghesquiere, B.; Rehman, J.; Lambrechts, D.; Malik, A.B.; Dewerchin, M.; Carmeliet, P. Inhibition of the Glycolytic Activator PFKFB3 in Endothelium Induces Tumor Vessel Normalization, Impairs Metastasis, and Improves Chemotherapy. *Cancer Cell* **2016**, *30*, 968–985.
26. Bessman, M.J.; Frick, D.N.; O'Handley, S.F. The MutT proteins or "Nudix" hydrolases, a family of versatile, widely distributed, "housecleaning" enzymes. *J. Biol. Chem.* **1996**, *271*, 25059–25062.
27. Marriott, A.S.; Vasieva, O.; Fang, Y.; Copeland, N.A.; McLennan, A.G.; Jones, N.J. NUDT2 Disruption Elevates Diadenosine Tetrphosphate (Ap4A) and Down-Regulates Immune Response and Cancer Promotion Genes. *PLoS ONE* **2016**, *11*, e0154674.
28. Ng, C.K.; Cooke, S.L.; Howe, K.; Newman, S.; Xian, J.; Temple, J.; Batty, E.M.; Pole, J.C.; Langdon, S.P.; Edwards, P.A.; Brenton, J.D. The role of tandem duplicator phenotype in tumour evolution in high-grade serous ovarian cancer. *J. Pathol.* **2012**, *226*, 703–712.
29. Fare, T.L.; Coffey, E.M.; Dai, H.; He, Y.D.; Kessler, D.A.; Kilian, K.A.; Koch, J.E.; LeProust, E.; Marton, M.J.; Meyer, M.R.; Stoughton, R.B.; Tokiwa, G.Y.; Wang, Y. Effects of atmospheric ozone on microarray data quality. *Anal. Chem.* **2003**, *75*, 4672–4675.
30. Hansen, K.; S.E., B.; Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Research* **2010**, *38* (12), e131. doi:10.1093/nar/gkq224.

31. Risso, D.; Schwartz, K.; Sherlock, G.; Dudoit, S. GC-content normalization for RNA-Seq data. *BMC Bioinformatics* **2011**, *12*, 480.
32. Gao, L.; Fang, Z.; Zhang, K.; Zhi, D.; Cui, X. Length bias correction for RNA-seq data in gene set analyses. *Bioinformatics* **2011**, *27*, 662–669.
33. Leek, J.; Scharpf, R.; Bravo, H.; Simcha, D.; Langmead, B.; Johnson, W.; Geman, D.; Baggerly, K.; Irizarry, R. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics* **2010**, *11*, 733–739.
34. Kaisers, W.; Boukamp, P.; Stark, H.J.; Schwender, H.; Tigges, J.; Krutmann, J.; Schaal, H. Age, gender and UV-exposition related effects on gene expression in in vivo aged short term cultivated human dermal fibroblasts. *PLoS ONE* **2017**, *12*, e0175657.
35. Kaisers, W. *seqTools: Analysis of nucleotide, sequence and quality content on fastq files.*, 2014. R package version 1.0.0.
36. Lance, G.N.; Williams, W.T. Computer Programs for Hierarchical Polythetic Classification (“Similarity Analyses”). *The Computer Journal* **1966**, *9*, 60–64. doi:10.1093/comjnl/9.1.60.
37. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.
38. Kaisers, W.; Schaal, H.; Schwender, H. rbamtools: an R interface to samtools enabling fast accumulative tabulation of splicing events over multiple RNA-seq samples. *Bioinformatics* **2015**, *31*, 1663–1664.
39. Robinson, M.D.; McCarthy, D.J.; Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **2010**, *26*, 139–140.

Sample Availability: The raw FASTQ files are available under ArrayExpress accession E-MTAB-4652 (ENA study ERP015294). The software is available in R packages: *seqTools* on Bioconductor.