Clustering pipeline for determining consensus sequences in targeted next-generation sequencing
Raunaq Malhotra1,∗, Daniel Elleder2,∗, Le Bao3, David R Hunter3, Raj Acharya1, Mary Poss4
1 Department of Computer Science and Engineering, Pennsylvania State University, UP, PA, 16801
2 Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Prague, 14220, Czech Republic
3 Department of Statistics, Pennsylvania State University, University Park, PA, 16801
4 Department of Biology, Center for Infectious Disease Dynamics, Pennsylvania State University, University Park, PA, 16801 * E-mail: Corresponding raunaq@psu.edu

Abstract
Analyses of targeted genomic sequencing data from next-generation-sequencing (NGS) technologies typically involves mapping reads to a reference sequence or clustering reads. For a number of species a reference genome is not available so the analyses of targeted sequencing data, for example polymorphic structural variation caused by mobile elements is difficult; clustering methods are preferred for such data analysis. Clustering of reads requires a clustering threshold parameter, which is used to compare and group reads. However, determining the optimal clustering threshold for a read dataset is challenging because of different sequence composition, the number of sequences present, and also the amount of sequencing errors in the dataset. High values of the clustering threshold parameter can falsely inflate the number of recovered genomic regions, while low values of clustering threshold can merge reads from distinct regions into a single cluster. Thus, an algorithm that can empirically determine clustering threshold is needed.

We propose a pipeline for clustering genomic sequences wherein the clustering threshold is empirically determined from the NGS data. The optimal threshold is decided based on two internal clustering measures, namely the Dunn Index and Davies-Bouldin Index, which assess clusters for small intra-cluster diameters and large inter-cluster distances. We evaluate the pipeline on two simulated datasets derived from human genome sequence simulating different genomic regions and sequencing depth. The optimal thresholds as predicted from our pipeline produce compact and well-separated clusters. The total number of clusters obtained is closer to the actual number of reference sequences when compared to single round of clustering. Also, the number of clusters whose consensus sequence matches a corresponding reference sequence is higher in our pipeline. We observe that the presence of repeat regions affects clustering accuracy. Nevertheless, the proposed pipeline clusters unique regions well even in the presence of sequences from repeat regions in the genome.

Introduction
Targeted genomic sequencing using next generation sequencing (NGS) technologies produces a large number of short sequences, called reads, from a small portion of the genome. It allows researchers to study specific regions of interest from a number of samples in a cost-effective manner. Examples of studies include exome sequencing projects [19,21], population targeted sequencing [4,5], and transposon tagging projects [9].

For organisms with high quality reference genome (eg. human, mouse), the reads can be mapped to the organism’s reference sequence to determine the number of unique genomic locations sequenced. There are a number of mapping tools available which can perform this task in a fast and accurate manner [1,2,8,14,16]. Clustering is preferred in non-model organisms where a reference genome is not available. Examples of such studies include identification of unique integration sites of mobile elements
in non-model organisms \cite{9,13}. In clustering, reads with similarity above a certain threshold, known as the clustering threshold, are grouped together into a single cluster. The similarity of two reads is computed using their pairwise alignment scores \cite{24}. Reads originating from identical genomic regions have the same sequence, and thus a high pairwise alignment score, and are grouped together. Thus, the consensus sequence of reads in a single cluster represents the sequence of the target region and the number of clusters determines the number of genomic locations sequenced.

However, with the large number of reads obtained from NGS technologies, computing pairwise alignment scores for all pairs of reads is computationally prohibitive. There are a number of tools available that perform fast clustering of reads \cite{8,10,15}. Instead of computing pairwise alignment score for every pair of read, similar reads are first filtered using heuristics and the pairwise alignment scores are only computed for reads that have similarity above a certain cutoff. This reduces the number of pairwise comparison computations thereby reducing the computational time.

The choice of clustering threshold is influenced by the nature of genomic sequences and the technology used for their sequencing. With the presence of sequencing errors in the reads, a high value of clustering threshold can generate a large number of false clusters. Relaxing the clustering threshold allows reads that are not from the same genomic regions to collapse in a single cluster. This latter problem will affect reliable clustering of sequences derived from repeat regions. Thus, it is important to choose an optimal value of clustering threshold for obtaining an accurate set of consensus sequences representing the sequenced genomic regions.

An ideal clustering algorithm should separate the reads into a compact set of clusters with large pairwise distances. A compact cluster is one where all reads in the cluster are more similar to each other than to reads in a different cluster, leading to a small diameter for the cluster. There are a number ways for evaluating a clustering algorithm that quantify the overall compactness of each of the clusters \cite{6,7,11,12}. In this paper, we develop a pipeline for clustering reads into a collection of compact and well-separated clusters representing the genomic regions from which the reads were sequenced. We use two rounds of clustering and vary the thresholds in each round to optimize two cluster compactness measures (Dunn Index and Davies-Bouldin Index) \cite{6,7}. Our approach enables us to determine the optimal clustering threshold based on the sequence content, depth and error profiles that are specific to a data set.

We use the UCLUST algorithm (available in the software USEARCH) \cite{8} as the base clustering algorithm in our pipeline. USEARCH software has been used for clustering in a number of genomic sequencing projects \cite{3,17,18,23}. We also compare the clustering results obtained from our pipeline to those obtained from the UCLUST algorithm alone. This pipeline is applied to two simulated datasets obtained from human genome representing different genomic sequences and coverage which are typically obtained in current sequencing technologies. We evaluated our pipeline and a single round of clustering using the cluster compactness measures and also performed external validation of individual clusters obtained. The results indicate that the proposed pipeline outperforms a single round of clustering at any clustering threshold. Sequence depth does not affect clustering performance but the sequence content of a dataset, specifically sequences from repeat regions, impacts the overall clustering accuracy.

**Methods**

**Definitions**

We denote the collection of reads obtained from a sequencing project containing $P$ different genomic regions as $R = \{R_1, R_2, ..., R_N\}$. A clustering algorithm $C(R, \theta)$ takes as input reads $R$ and a clustering threshold $\theta$ to output a collection of clusters $M(\theta) = \{M_1, M_2, ..., M_K\}$ in which the reads are grouped together. In other words, $C(\theta) : R \rightarrow \{M_1, M_2, ..., M_K\}$ defines a map from the reads to $K$ clusters, wherein reads assigned to a cluster $M_i$ have an alignment score greater than $\theta$ when aligned to a consensus sequence, denoted as $Con(M_i)$, for that cluster. The consensus sequence $Con(M_i)$ of a cluster is computed
by a position-by-position majority vote on the multiple sequence alignment (MSA) of reads assigned to cluster $M_i$, and then removing the gaps, if any, from the MSA.

Typically, when sequencing, the number of unique genomic regions $P$ is unknown, and the goal of a clustering algorithm $C(\theta)$ is to cluster the reads $R$ into $P$ clusters. Thus, an ideal collection of clusters should have $K = P$ unique clusters, with the consensus sequence $\text{Con}(M_i)$ of each cluster in $M(\theta)$ being identical to the genomic sequence from which reads were sampled.

### Clustering Pipeline

We will refer to an invocation of $C(R, \theta)$ as a clustering run. The pipeline consists of four major steps (Figure 1). In the first step, we cluster the collection of reads $R$ at clustering threshold $\theta^{(1)}$ using the clustering algorithm $C(R, \theta^{(1)})$. The clustering algorithm generates a collection of clusters $M^{(1)}(\theta^{(1)}) = \{M_1^{(1)}, M_2^{(1)}, \ldots, M_K^{(1)}\}$ (The superscript 1 denotes that the clusters are obtained in step one of the pipeline). We also compute the consensus sequences $\text{Con}(M^{(1)})$ for each cluster in $M^{(1)}$ based on the reads assigned to it. In the second step, we obtain a subset of clusters $M^{(2)} \subset M^{(1)}$ by ignoring clusters in $M^{(1)}$ containing less than a threshold $T_{err}$ number of reads, namely, $M^{(2)}(\theta^{(1)}, T_{err}) = \{M_i^{(1)} \mid |M_i^{(1)}| > T_{err}\}$. In step three, we re-cluster the consensus sequences of $M^{(2)}$ clusters $\text{Con}(M^{(2)})$ using the clustering algorithm $C(M^{(2)}, \theta^{(3)})$ to obtain a revised cluster set $M^{(3)}(\theta^{(1)}, T_{err}, \theta^{(3)}) = \{M_1^{(2)}, M_2^{(2)}, \ldots, M_K^{(2)}\}$. Next we map all the reads in $R$ to the consensus sequences of $M^{(3)}$ obtained from step three. The clusters are updated to the consensus sequences of their constituent reads. The cluster set $M^{(3)}$ obtained in step four constitutes the final set of clusters obtained from read set $R$ using the pipeline.

The clustering algorithm $C(R, \theta)$ can be any available clustering tool [2, 8, 15, 16]. There are three parameters in our pipeline, namely $(\theta^{(1)}, T_{err}, \theta^{(3)})$. We describe a procedure for determining the optimal values for these three parameters given a read data $R$.

In step two of the pipeline, each cluster $M_i^{(2)} \in M^{(2)}$ has more than $T_{err}$ reads in it. The rationale for removing small clusters in step two is that they likely correspond to some genomic regions represented by clusters in $M^{(2)}$. Thus, there are more than one cluster in $M^{(1)}$ corresponding to the same genomic location. These small clusters arise due to the presence of sequencing errors in the reads and at the clustering threshold $\theta^{(1)}$, reads from the same genomic location are split into more than one cluster. As each cluster corresponds to a unique genomic region, this falsely increases the apparent number of genomic regions present in the dataset.

The clustering in step three $M^{(3)}(\theta^{(3)})$ further combines clusters corresponding to the same genomic region. The clustering threshold in step three $\theta^{(3)}$ can be different from the one used in step one. Step four reassigns every single read to the clusters obtained in step three. This step ensures that the clusters and reads removed in step two are also assigned to the genomic region of their origin.

### Clustering algorithm $C$

We use the clustering algorithm UCLUST [8] based on its favorable comparisons to other clustering softwares on speed and relative accuracy [8]. UCLUST takes the reads and a clustering threshold as input data. It selects the first read as a seed sequence for a cluster, and then iteratively assigns reads to that cluster if their alignment score to the seed sequence is greater than the clustering threshold. A new cluster is created if the alignment score of a read is below the clustering threshold for all existing clusters (seed sequences of existing clusters), with the current read forming the seed sequence of the new cluster. The seed sequence for a cluster remains the same throughout the UCLUST clustering. Thus, the choice of the seed sequence can have an effect on the number and quality of clusters obtained from UCLUST.

The reads assigned to a single cluster can be aligned to generate a consensus sequence for that cluster. The similarity score between two reads is defined in terms of their pairwise alignment score.
The expected computational complexity of a fast clustering algorithm $C$ is linear with the number of reads. The alignment score of a read to seed sequences of clusters is only computed for a fixed number of clusters, before assigning the read to a new cluster. The computational complexity of computing alignment score is quadratic in the read length in the worst case. Thus a clustering run has computational complexity of $O(|R| \cdot |R|^2)$. Our pipeline performs two rounds of clustering, where in step 3, only $|M^{(3)}|$ sequences are clustered. Thus, the overall complexity of the pipeline is of the order of $O((|R| + |M^{(3)}|) \cdot |R|^2)$, which is also linear in the number of reads and similar to the computational complexity of a single run of the clustering algorithm $C$.

In this study, clustering was performed using the UCLUST algorithm in USEARCH with options set to: global matching, nofastalign, maxrejects , maxaccepts as 0. We also used the identity definition 1 in the clustering options. For mapping of reads in step 4, we used query mapping to database option in USEARCH. The parameters for query mapping were set to: global matching, nofastalign, maxrejects and maxaccepts to 80, identity definition was set to 1.

Determining the clustering thresholds

The collection of clusters $M^{(3)}$ is a function of the clustering thresholds ($\theta^{(1)}, \theta^{(3)}$) and the error threshold $T_{err}$. In order to determine the optimal thresholds for clustering of read set $R$, we use two internal clustering compactness measures, Dunn Index and Davies-Bouldin (DB) Index [6, 7] to determine the optimal thresholds. Dunn Index is defined as the ratio of the minimum inter-cluster distance to the maximum diameter amongst all the clusters.

$$DI(C(\theta) : R \to \{M_1, M_2, \ldots, M_K\}) = \frac{\min_{i,j \in M} d(M_i, M_j)}{\max_{k \in M} \text{dia}(M_k)}$$

The inter-cluster distance is defined as the distance between the consensus sequences of two clusters.

$$d(M_i, M_j) = d(\text{Con}(M_i), \text{Con}(M_j))$$

The distance between two sequences is defined as one minus the pairwise alignment scores of the two sequences. Thus,

$$d(M_i, M_j) = 1 - P(\text{Con}(M_i), \text{Con}(M_j))$$

Here we define the diameter of a cluster as the average distance of all reads assigned to the cluster by $C(\theta)$ to the consensus sequence of the cluster.

$$\text{dia}(M_k) = \frac{1}{|M_k|} \cdot \sum_{R_i \in M_k} d(R_i, \text{Con}(M_k))$$

An optimal clustering should have well separated and compact clusters. Well separated clusters indicates that the minimum value of inter-cluster distances is high. Compact clusters corresponds to small values of the maximum diameter for each cluster. Thus, overall a large value of $DI$ indicates an optimal clustering of the reads. $DI < 1$ implies that minimum value of inter-cluster distance is smaller than the maximum diameter of any cluster. In such a scenario, there are two clusters closer to each other than the diameter of a largest cluster. Thus, for obtaining optimal clustering, we focus on clustering thresholds for which $DI > 1$.

DB-Index is the ratio of the intra-cluster distance to inter-cluster distances, which is averaged over all the clusters [8]. For each cluster $M_i, i \in \{M_1, \ldots, M_K\}$, first a measure of closeness of all other cluster is defined.

$$\text{Cl}(i, j) = \frac{\text{dia}(M_i) + \text{dia}(M_j)}{d(M_i, M_j)}$$
The maximum closeness of each cluster is defined as the maximum value for the closeness defined above.

\[ Cl_{\text{max}}(i) = \max_{j \neq i} Cl(i, j) \]

Subsequently, the DB-Index of a clustering run is defined as the average value of the maximum closeness of each cluster.

\[ DB(C(\theta) : \mathbb{R} \to \{M_1, M_2, \ldots, M_K\}) = \frac{1}{K} \sum_{k \in \{1, \ldots, K\}} Cl_{\text{max}}(k) \]

Note that a compact clustering run should have low values of DB-Index as it corresponds to well separated and compact clusters, wherein each cluster is far from other clusters. If a clustering run produces two clusters corresponding to the same genomic region, then their inter-cluster distance will be small, leading to a large value for the maximum closeness for these two clusters. This will generate a large value for the DB-Index. We thus look for clustering thresholds that correspond to small values of DB-Index.

We combine the above two measures into a single measure for compactness, henceforth mentioned as the \( I - index \), by taking the harmonic mean of Dunn Index and inverse of DB-Index.

\[ I - index = \frac{2 \cdot DI}{DI \cdot DB + 1} \]

We use the \( I - index \) to evaluate our clustering pipeline and to determine the optimal values for the clustering thresholds. A high value for \( I - index \) indicates a compact and well-separated set of clusters. As the measure evaluates the results obtained from a clustering run solely based on the reads, it is known as an internal clustering measure. The clustering thresholds \( \{\theta^{(1)}, \theta^{(3)}\} \) (for step 1 and step 3 in the pipeline) which maximize the clustering compactness measure \( (I - index) \) and have Dunn Index greater than 1 \( (DI > 1) \) are chosen as the optimal clustering parameters for the collection of reads \( R \). We search for the optimal parameters by performing clustering at different values of \( \{\theta^{(1)}, \theta^{(3)}\} \), and retaining the collection of clusters \( M^{(3)} \) which satisfy the optimization criterion.

**Simulated data**

In order to evaluate the proposed clustering pipeline, we simulate targeted sequences from human genome build hg19. We chose 100 bp fragments from 1993 random segments of chromosomes 15 and 22 (dataset S1), and 2201 random segments of the entire human genome (dataset S2). The two datasets represent the different genomic contents of typical targeted sequencing projects (for example, exome sequencing [20]). We analyze the repeat content of two datasets using RepeatMasker [http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker](http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker). S1 contains 19.83% of interspersed repeats while S2 contains 18.60% of interspersed repeats sequences. We simulate Ion Torrent sequencing reads for these two datasets at an average coverage of 149 and 1318, respectively (Table I). Different read depths are typical of those produced by different sequencing technologies and allows us to investigate the effect of sequence depth on correct cluster recovery by our pipeline. We use the simulation software dwgsim [http://github.com/nh13/DWGSIM](http://github.com/nh13/DWGSIM) for simulating the Ion Torrent Sequencing reads. The software introduces an average sequencing error of 2% in the reads. The reads from datasets S1 and S2 were trimmed to 40 bps to avoid sequencing errors at the 3’ end and for faster analysis.

**External measures for cluster evaluation**

As the DNA segments \( P \) for the simulated dataset are known, we can assess the quality of clusters obtained by our pipeline externally as well as internally. Each read in \( R \) is obtained from one of the DNA segments, denoted \( \{C_1, \ldots, C_P\} \). Thus, the true map of reads \( T : \mathbb{R} \to \{C_1, \ldots, C_P\} \) is known.
We define the following external measures for evaluating a clustering run \( C(\theta) : \mathbb{R} \to \{ M_1, \ldots, M_K \} \). A cluster \( M_k \) is called a merged cluster if it contains reads from two different DNA segments.

\[
Merged(M_k) = \begin{cases} 
1, & \text{if } \exists i,j; C(R_i, \theta) = C(R_j, \theta) = M_k, \text{ but } T(R_i) \neq T(R_j) \\
0, & \text{otherwise}
\end{cases}
\]

We define homogeneity of a clustering as a fraction of the clusters that are not merged in the clustering.

\[
h = 1 - \frac{1}{K} \sum_{k=1}^{K} Merged(M_k)
\]

The measure \( h \) indicates the fraction of clusters that contain all reads from a single DNA reference segment. Note that a trivial assignment of every read to a new cluster would have a perfect homogeneity in a clustering and a single cluster containing all the reads will have a zero homogeneity.

We also compute the DNA reference segments \( C_r \), from which the reads are assigned to two or more clusters, denoted as split references.

\[
Split(C_r) = \begin{cases} 
1, & \text{if } \exists i,j; T(R_i) = T(R_j) = C_r, \text{ but } C(R_i) \neq C(R_j) \\
0, & \text{otherwise}
\end{cases}
\]

We define completeness of a clustering as the fraction of DNA reference segments that are not split in the clustering.

\[
c = 1 - \frac{1}{P} \sum_{r=1}^{P} Split(C_r)
\]

Similar to homogeneity, the completeness \( c \) of a clustering indicates the fraction of DNA segments that have all their reads in a single cluster. A clustering where every read is assigned to a new cluster will have a zero value for completeness measure, while the clustering with all reads in a single cluster will have a perfect completeness measure. Thus, completeness and homogeneity are complementary measures.

We also compute for a cluster \( M_k \) whether all the reads from a single DNA segment \( C_a \) and none from a different segment are present in it. We denote such a cluster as a “correct cluster”.

\[
Correct(M_k) = \begin{cases} 
1, & \text{if } \forall R_i \in \mathbb{R}, \text{ such that } T(R_i) = C_a \Leftrightarrow C(R_i, \theta) = M_k \\
0, & \text{otherwise}
\end{cases}
\]

Similarly, the true fraction of a clustering is defined as the ratio “correct clusters” to the total number of DNA reference fragments.

\[
t = \frac{1}{P} \sum_{k=1}^{K} Correct(M_k)
\]

The external measures, as defined above, for an optimized clustering should maximize the number of “correct clusters” \( \sum Correct(M_k) \) and minimize the number of merged clusters \( \sum Merged(M_k) \) and split DNA segments \( \sum Split(C_r) \). In other words, large values for homogeneity, completeness and true fraction denotes an optimized clustering. We compute a single clustering evaluation metric (similar to V-Measure [22]) from the above three measures by computing their harmonic mean.

\[
E-index = \frac{3 \cdot c \cdot h \cdot t}{c \cdot h + h \cdot t + t \cdot c}
\]

The \( E-index \), as defined above, has a value in the range of \([0,1]\), where higher values indicate good clustering.
Apart from the compactness of a clustering, the consensus sequence of a cluster should also be close to the original DNA reference segment used to generate the reads. Thus, we also measure the percent identity of the consensus sequence of a cluster $Con(M_k)$ to the reference DNA segment used to generate it.

**Results and Discussion**

**Clustering results from a single run of UCLUST**

We cluster reads in datasets S1 and S2 at various clustering thresholds starting from 70% to 95% in steps of 5. Using a single clustering run of UCLUST, none of the clustering thresholds recovered the actual number of DNA segments (Table 2). High clustering thresholds yielded inflated number of clusters and the number of clusters decreased below the number of reference DNA segments as the clustering threshold was decreased.

The Dunn index values for all clustering thresholds in the single run are zero or less than one (Table 2) in dataset S1. This indicates that none of the clustering thresholds in the single round of clustering generates well separated or compact collection of clusters. The DB-Index increases with increasing clustering thresholds, also indicating decrease in the clustering quality with increasing clustering thresholds.

The internal clustering measures are sensitive to clusters corresponding to outlier reads in the dataset, as such clusters have a small inter-cluster distance to another cluster in the clustering. Such outlier clusters usually have one or two reads in them. To account for this, we recomputed Dunn and DB indices for the above clustering thresholds after removing clusters that contain two or less reads (Table 2). The DB index is much smaller in magnitude than those computed earlier. However, the Dunn index is still less than one for all clustering thresholds in both datasets, suggesting that the single round of clustering fails to obtain compact and well-separated clusters at any clustering threshold.

The overall internal measure, $I-index$, is optimized for clustering threshold of $\theta^{(1)} = 75$, which is the same as optimal clustering threshold from external clustering measures (Table 2). For S2 dataset, we observe optimal internal compactness measures for clustering threshold of $\theta^{(1)} = 75$. The internal clustering measures across the two datasets are consistent at 5% significance level, suggesting that the clustering algorithm is independent of coverage depth.

The optimal clustering threshold for S1 based on the $E-index$ is ($\theta^{(1)} = 75$), while for dataset S2 is ($\theta^{(1)} = 70$). The $E-index$ decreases from 0.93 to zero with increasing clustering thresholds for both S1 and S2 datasets (Figure 2). For S1, the true fraction is 90% at clustering threshold of ($\theta^{(1)} = 75$), while for S2, the true fraction at clustering threshold of ($\theta^{(1)} = 70$) is 89%.

The homogeneity of a clustering increases to one as the clustering threshold increases for both datasets. This suggests that clusters containing reads only from a single reference DNA segment increases with the increasing clustering thresholds (Supplementary Table 1). However, the completeness of the clustering decreases to zero as the clustering threshold increases (Figure 2), indicating that more reference DNA segments are split into two or more clusters. Also, the true fraction for all clustering thresholds is less than 90%, showing that none of the clustering thresholds recovered more than 90% of the reference DNA segments correctly (Figure 2 and Supplementary Table 1).

**Clustering using the proposed pipeline**

We applied the four-step clustering pipeline (Figure 1) on the S1 and S2 datasets. In step one, the reads are clustered at a clustering threshold ($\theta^{(1)}$) varying from 75% to 95% in steps of 5. In step two, for each clustering, we remove the clusters that contain two reads or less assigned to them ($T_{err}$ is set to two in our pipeline). In step three, we re-cluster the remaining clusters from step two at clustering threshold ($\theta^{(3)}$) varying again from 75% to 95% in steps of 5 (Here the superscript on the clustering threshold indicates
the step of our pipeline). The consensus sequence of the clusters thus obtained in step three represent the reference DNA sequences, and we map all the reads to them in step four.

We evaluate the clustering performed at each clustering threshold pairs \((\theta^{(1)}; \theta^{(3)})\) based on internal and external measures to determine an optimal clustering threshold. For S1, the combined internal measure, \(I - index\), is maximized for clustering threshold pairs \((\theta^{(1)}; \theta^{(3)}) = ((75; 75), (75; 80))\) (Table 3). The external clustering measure \(E - index\) in S1 dataset is maximized for clustering threshold pairs \((\theta^{(1)}; \theta^{(3)}) = ((80; 80), (80; 75))\). There are two sets of clustering parameters for which both the I-index and E-index are near maximum \(((75; 75) and (80; 80))\). However, the number of clusters obtained using these thresholds differs (1894 vs 1982) (Table 3). In all cases the Dunn Index is greater than one. As this is simulated data, we can also evaluate the external measures of cluster quality. Threshold pair \((75; 75)\) returns fewer correct clusters than does threshold pair \((80; 80)\) (1804 vs 1850, respectively. See Supplementary table 2, and there are more merged clusters but fewer split references in the \((75; 75)\) clustering runs (Supplementary table 2).

We observe that there are multiple clustering threshold pairs which appear as optimal under the different measures. For experimental data sets, the only parameters that can be obtained are the internal measures and the total number of clusters recovered. Thus we determined if there were differences in the cluster quality for each of the optimal thresholds. We select the optimal clustering threshold pairs in S1 suggested by the internal clustering measures, namely \((\theta^{(1)}; \theta^{(3)}) = ((75; 75), (75; 80), (80; 80))\), and compare the consensus sequences of the clusters to their original DNA reference segments. The distribution of the sequence differences between consensus sequences of clusters and the DNA reference segments is similar for all the clustering threshold pairs (Figure 3). This suggests that any one of the optimal clustering threshold pairs will return suitable clusters. As the total number of reference DNA segments will be unknown for a real dataset, choosing the clustering threshold pairs that generates the maximum number of clusters could be imposed as a user-specified criterion.

**Effect of Read Depth:** The dataset S2 is used to evaluate the impact of increased sequencing depth on our pipeline. Our results for clustering S2 dataset with our pipeline are similar to those obtained for S1 (Supplementary Text, Supplementary Table 3). The clustering threshold pairs \((\theta^{(1)}; \theta^{(3)}) = ((75; 75), (70; 70))\) maximize the combined internal measure \(I - index\), which contains the clustering threshold pairs \((\theta^{(1)}; \theta^{(3)}) = (75; 75)\) for which the external clustering index \(E - index\) is also optimized. The number of clusters generated at optimal clustering threshold pairs are 94%-98% of the actual number of reference DNA segments.

The values for overall internal index \(I - index\) are comparable amongst the two datasets, indicating read depth has little effect on the overall clustering accuracy (Tables 3, 4). We did note that the Dunn indices were smaller for dataset S2 compared to those for S1, which is likely due to increased cluster diameters with the increased read number. The true fraction at optimal clustering thresholds is more than 90% in S2 (Supplementary table S3), even when the coverage of DNA segments was ten-fold that of the S1 dataset.

**Genome Content:** The presence of reference segments from repeat regions in the genome can affect the accuracy of a clustering as two different reference segments may collapse into a single cluster if their pairwise distance is less than the clustering threshold. We investigate the effect of DNA segments from repeat regions of the human genome on the clustering pipeline. In the S1 and S2 datasets, 464 of 1993 (23.3%) and 492 of 2201 (22.4%) the DNA segments, respectively, are from repeat regions in the human genome (Table 1). At the optimal clustering threshold pairs for S1, \((\theta^{(1)}; \theta^{(3)}) = (80; 80)\), less than 55% of the repeat region clusters are recovered that differ by only 1 bp from their corresponding DNA segment (Figure 4). On the other hand, 84% of the unique region clusters are recovered to within 1 bp of their corresponding DNA segment. The repeat region clusters might account for the difference in the total number of clusters obtained for the optimal clustering threshold pairs. The results for repeat region reference segment recovery to those from unique regions for S2 dataset are similar to S1 data (Supplementary Figure 1).
Conclusions

We present a pipeline for empirically determining threshold values for de novo clustering of reads obtained from a targeted sequencing run. Our results demonstrate improved performance for the proposed pipeline compared to a single round of clustering in a number of ways. First, in contrast to a single round of clustering, the total number of clusters recovered approximates the actual number of DNA segments used in the simulations at more combinations of clustering threshold pairs. The second round of clustering (in step three) merges similar clusters which were erroneously separated in the first round of clustering. Second, we were able to achieve higher values for Dunn indices and lower DB indices for several threshold combinations compared to single round of clustering indicating well-separated and compact clusters. Third, by using the pipeline, the true fraction of reference DNA segments recovered is higher for both datasets compared to the single round of clustering. Although there a multiple optimal clustering threshold pairs, the consensus sequences of the clusters are consistent across these optimal clustering threshold pairs. The proposed pipeline is insensitive to the sequence depth of the dataset and recovers compact and well-separated clusters. However, the clusters corresponding to repeat region sequences have less accurate consensus sequences as compared to clusters corresponding to unique regions in the genome. We observe that the optimal clustering threshold pairs obtained from internal clustering measures occur at the same pairs as those from external clustering measures for both datasets. Thus, we can use the internal clustering measures (specially the combined internal measure $I - index$) in our pipeline for determining optimal clustering thresholds for real datasets.

References

1. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, October 1990.
2. Christof Angermüller, Andreas Biegert, and Johannes Sding. Discriminative modelling of context-specific amino acid substitution probabilities. *Bioinformatics*, 28(24):3240–3247, 2012.
3. Petr Baldrian, Miroslav Kolářík, Martina Štursová, Jan Kopecký, Vendula Valášková, Tomáš Větrovský, et al. Active and total microbial communities in forest soil are largely different and highly stratified during decomposition. *The ISME journal*, 6(2):248–258, 2011.
4. 1000 Genomes Project Consortium et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.
5. John W Davey, Paul A Hohenlohe, Paul D Etter, Jason Q Boone, Julian M Catchen, and Mark L Blaxter. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, 12(7):499–510, 2011.
6. David L Davies and Donald W Bouldin. A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (2):224–227, 1979.
7. Joseph C Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. 1973.
8. Robert C. Edgar. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461, October 2010.
9. Daniel Elleder, Oekyung Kim, Abinash Padhi, Jason G Bankert, Ivan Simeonov, Stephan C Schuster, Nicola E Wittekindt, Susanne Motameny, and Mary Poss. Polymorphic integrations of an endogenous gammaretrovirus in the mule deer genome. *Journal of virology*, 86(5):2787–2796, 2012.
10. Mohammadreza Ghodsi, Bo Liu, and Mihai Pop. Dnaclust: accurate and efficient clustering of phylogenetic marker genes. *BMC bioinformatics*, 12(1):271, 2011.

11. Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2-3):107–145, 2001.

12. Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. Clustering validity checking methods: part ii. *ACM Sigmod Record*, 31(3):19–27, 2002.

13. Pauline L Kamath, Daniel Elleder, Le Bao, Paul C Cross, John H Powell, and Mary Poss. The population history of endogenous retroviruses in mule deer (odocoileus hemionus). *Journal of Heredity*, 105(2):173–187, 2014.

14. Heng Li and Nils Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in bioinformatics*, 11(5):473–483, 2010.

15. Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)*, 22(13):1658–1659, July 2006.

16. Yongchao Liu, Bertil Schmidt, and Douglas Maskell. Cudasw++2.0: enhanced smith-waterman protein database search on cuda-enabled gpus based on sint and virtualized simd abstractions. *BMC Research Notes*, 3(1):93, 2010.

17. Rachel Mackelprang, Mark P Waldrop, Kristen M DeAngelis, Maude M David, Krystle L Chavarria, Steven J Blazewicz, Edward M Rubin, and Janet K Jansson. Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature*, 480(7377):368–371, 2011.

18. Barbara A Methé, Karen E Nelson, Mihai Pop, Heather H Creasy, Michelle G Giglio, Curtis Huttenhower, Dirk Gevers, Joseph F Petrosino, Sahar Abubucker, Jonathan H Badger, et al. A framework for human microbiome research. *Nature*, 486(7402):215–221, 2012.

19. Sarah B Ng, Kati J Buckingham, Choli Lee, Abigail W Bigham, Holly K Tabor, Karin M Dent, Chad D Huff, Paul T Shannon, Ethylin Wang Jabs, Deborah A Nickerson, et al. Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics*, 42(1):30–35, 2010.

20. Sarah B Ng, Kati J Buckingham, Choli Lee, Abigail W Bigham, Holly K Tabor, Karin M Dent, Chad D Huff, Paul T Shannon, Ethylin Wang Jabs, Deborah A Nickerson, et al. Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics*, 42(1):30–35, 2010.

21. Sarah B Ng, Emily H Turner, Peggy D Robertson, Steven D Flygare, Abigail W Bigham, Choli Lee, Tristan Shaffer, Michelle Wong, Arindam Bhattacharjee, Evan E Eichler, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261):272–276, 2009.

22. Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *EMNLP-CoNLL*, volume 7, pages 410–420. Citeseer, 2007.

23. Patrick D Schloss and Sarah L Westcott. Assessing and improving methods used in operational taxonomic unit-based approaches for 16s rrna gene sequence analysis. *Applied and environmental microbiology*, 77(10):3219–3226, 2011.

24. Temple F Smith and Michael S Waterman. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.
Figure 1: Proposed clustering pipeline: The four main steps of our clustering pipeline include 1). Cluster the reads at a clustering threshold $\theta$ to obtain a set of clusters and their consensus sequence. 2). Remove clusters containing small reads. 3). Re-cluster the remaining clusters using clustering threshold $\theta'$ and obtain consensus sequence. 4). Map all reads to the consensus sequences obtained in step three.

| Step 1 | Step 2 | Step 3 | Step 4 |
|--------|--------|--------|--------|
| - Cluster reads $R$ using clustering algorithm $C(R, \theta^{(1)})$ to obtain set of clusters $M^{(2)}(\theta^{(1)})$<br>- Compute the consensus sequences $Con(M_i)$, $\forall M_i \in \mathcal{M}^{(1)}(\theta^{(1)})$ | - Remove clusters with less than $T_{err}$ reads to obtain new cluster set $\mathcal{M}^{(2)}(\theta^{(1)}, T_{err}) \subset \mathcal{M}^{(1)}(\theta^{(1)})$ | - Re-cluster the remaining clusters using $C(M^{(2)}, \theta^{(2)})$ to obtain $\mathcal{M}^{(3)}(\theta^{(1)}, T_{err}, \theta^{(2)})$ | - Map all reads $R$ to cluster set $\mathcal{M}^{(3)}$ |

Figure 1. Proposed clustering pipeline
Figure 2: Single round of clustering using UCLUST: Clustering analysis at various percentage identities for S1 (a) and S2 (b). Homogeneity is the fraction of clusters that have all reads from a single reference DNA segment. Completeness measures the fraction of reference DNA segments that have all reads assigned to a single cluster. True fraction measures the fraction of reference DNA segments that are complete and their corresponding clusters being homogeneous (See methods for details). The external measure, $E-index$, is an harmonic mean of homogeneity, completeness and the true fraction. As homogeneity increases, the completeness decreases. None of the clustering thresholds has true fraction more than 90%.

(a) Dataset S1 (Total number of DNA segments 1993)

(b) Dataset S2 (Total number of DNA segments 2201)
Figure 3: Comparing cluster consensus to reference DNA segments in S1 The figure displays the number of nucleotide differences between the cluster consensus sequence and the original reference. The maximum number of differences was 12. The error distributions are the same at the three optimal clustering threshold pairs \{(75; 75), (75; 80), (80; 80)\}. More than 90% of the clusters are within 4 bp of the corresponding DNA reference segment.

Figure 3. Comparing cluster consensus to reference DNA segments in S1
Figure 4: Effect of repeat region sequences on correct clusters for S1  
The figure displays the percent of cluster consensus sequences derived from repeat regions and unique regions that differ from the reference sequence by less than (left histograms) or more than (right histograms) 2 nucleotides. The data are shown for the three optimal clustering pair thresholds.

![S1 unique versus repeat regions](image)

**Figure 4.** Effect of repeat region sequences on correct clusters for S1

**Supporting Information Legends**

**Supplementary figure 1: Comparing cluster consensus sequences to the DNA segments for S2** The figure displays the number of nucleotide differences between the cluster consensus sequence and the original reference for S2. The maximum number of differences was 12. The distributions are similar for the optimal clustering thresholds.

**Supplementary figure 2: Effect of repeat region sequences on correct clusters for S2** The figure displays the percent of DNA segments from repeat regions and unique regions that differ from the reference sequence by less than (left histogram) or more than (right histogram) 2 nucleotides. The data are shown for all the optimal clustering thresholds.

**Supplementary figure 3: Cluster analysis for S1 after step four of pipeline** The number of merged clusters, correct clusters, and DNA segments obtained after mapping step of the pipeline, at different clustering thresholds for S1 dataset are shown. At higher percentage identities, the number of correct clusters is highest, while the number of merged clusters is also minimum.

**Supplementary Table 1: Single run clustering from UCLUST**

**Supplementary Table 2: Results for step three of clustering pipeline for dataset S1**

**Supplementary Table 3: Results for step three of clustering pipeline for dataset S2**
Table 1. Statistics on the simulated data

| Dataset name | Chromosomes used | # of DNA segments | # of reads | % of Interspersed repeats | Average Coverage |
|--------------|------------------|-------------------|------------|--------------------------|-----------------|
| S1           | 15 & 22          | 1993              | 295,985    | 23.3                     | 149             |
| S2           | 1-22             | 2201              | 2,901,970  | 22.4                     | 1318            |

\[a\] Repeat percentage was computed using RepeatMasker (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker)

\[b\] Ion Torrent Sequencing simulation used WGS software (dwgsim https://github.com/nh13/DWGSIM)

Table 2. Single run clustering from UCLUST

| % Clustering Threshold \(\theta^{(1)}\) | No. of clusters \(a\) | E-Index \(b\) | Dunn Index | DB Index | Dunn Index \(c\) | DB Index \(c\) | I-Index \(c,d\) |
|-----------------------------------------|------------------------|---------------|------------|----------|-----------------|-----------------|-----------------|
| S1                                      |                        |               |            |          |                 |                 |                 |
| 70                                      | 1960                   | 0.919         | 0.0189     | 1.251    | 0.995           | 1.22            | 0.897           |
| 75                                      | 2014                   | 0.936         | 0          | 192365   | 0.979           | 1.14            | 0.927           |
| 80                                      | 2251                   | 0.895         | 0          | 344222   | 0.976           | 1.18            | 0.909           |
| 85                                      | 3543                   | 0.583         | 0          | 1968269  | 0.899           | 1.33            | 0.818           |
| 90                                      | 10155                  | 0.0969        | 0          | 3166512  | 0.883           | 1.61            | 0.729           |
| 95                                      | 39360                  | 0             | -          |          | 0.942           | 1.75            | 0.712           |
| S2                                      |                        |               |            |          |                 |                 |                 |
| 70                                      | 2241                   | 0.926         | 0          | 518636   | 0.896           | 1.26            | 0.841           |
| 75                                      | 2429                   | 0.897         | 0          | 637993   | 0.911           | 1.20            | 0.870           |
| 80                                      | 3849                   | 0.601         | 0          | 10367454 | 0.909           | 1.46            | 0.781           |
| 85                                      | 10652                  | 0.120         | 0          | 40080499 | 0.795           | 1.89            | 0.635           |
| 90                                      | 43625                  | 0.00272       | 0          |          |                 |                 |                 |
| 95                                      | 190733                 | 0             | -          |          |                 |                 |                 |

\[a\] Number of clusters obtained at the clustering threshold \(\theta^{(1)}\), Actual number of clusters S1 = 1993, S2 = 2201

\[b\] Combined External measure for the clustering consisting of completeness, homogeneity, and true fraction.

\[c\] Indices when clusters containing two or less reads are removed

\[d\] Combined Internal index consisting of Dunn and DB indices
Table 3. Clustering S1 using the proposed pipeline

| % Clustering Threshold ($\theta^{(1)}, \theta^{(3)}$) | Total no. of clusters $^a$ | E-index $^b$ | I-index $^c$ |
|--------------------------------------------------|-----------------------------|-------------|-------------|
| 70:70                                            | 1835                        | 0.961       | 1.04        |
| 70:75                                            | 1877                        | 0.956       | 0.887       |
| 70:80                                            | 1910                        | 0.95        | 0.94        |
| 70:85                                            | 1929                        | 0.946       | 0.928       |
| 70:90                                            | 1946                        | 0.941       | 0.93        |
| 70:95                                            | 1959                        | 0.938       | 0.893       |
| 75:70                                            | 1841                        | 0.97        | 0.867       |
| 75:75                                            | 1894                        | **0.972**   | **1.07**    |
| 75:80                                            | 1928                        | 0.969       | **1.07**    |
| 75:85                                            | 1950                        | 0.965       | 1.02        |
| 75:90                                            | 1995                        | 0.954       | 0.945       |
| 75:95                                            | 2011                        | 0.952       | 0.927       |
| 80:70                                            | 1849                        | 0.972       | 0.95        |
| 80:75                                            | 1911                        | **0.974**   | 0.957       |
| 80:80                                            | 1982                        | **0.974**   | **1.05**    |
| 80:85                                            | 2038                        | 0.966       | 0.99        |
| 80:90                                            | 2206                        | 0.928       | 0.915       |
| 80:95                                            | 2247                        | 0.921       | 0.873       |
| 85:70                                            | 1879                        | 0.963       | 0.811       |
| 85:75                                            | 1925                        | 0.972       | 1           |
| 85:80                                            | 2077                        | 0.955       | 0.962       |
| 85:85                                            | 2398                        | 0.89        | 0.95        |
| 85:90                                            | 3099                        | 0.715       | 0.871       |
| 85:95                                            | 3490                        | 0.654       | 0.825       |
| 90:70                                            | 1927                        | 0.946       | 0.923       |
| 90:75                                            | 1970                        | 0.962       | 0.906       |
| 90:80                                            | 2154                        | 0.938       | 0.925       |
| 90:85                                            | 3130                        | 0.731       | 0.834       |
| 90:90                                            | 5672                        | 0.323       | 0.76        |
| 90:95                                            | 9703                        | 0.126       | 0.764       |
| 95:70                                            | 1953                        | 0.939       | 0.908       |
| 95:75                                            | 2000                        | 0.954       | 0.874       |
| 95:80                                            | 2223                        | 0.923       | 0.885       |
| 95:85                                            | 3472                        | 0.665       | 0.811       |
| 95:90                                            | 9423                        | 0.136       | 0.781       |

$^a$ Total number of clusters obtained at a clustering threshold pair

$^b$ External clustering measure for the clustering threshold pair, bold highlights the top two values in the column

$^c$ Internal clustering measure for the clustering threshold pair, bold highlights the top two values in the column
Table 4. Clustering S2 using the pipeline

| % Clustering Threshold \((\theta^{(1)}, \theta^{(3)})\) | Total no. of clusters \(^a\) | E-index \(^b\) | I-index \(^c\) |
|--------------------------|---------------------|----------------|----------------|
| 70;70                    | 2073                | 0.969          | 1.05           |
| 70;75                    | 2116                | 0.962          | 0.888          |
| 70;80                    | 2141                | 0.958          | 0.936          |
| 70;85                    | 2165                | 0.953          | 0.924          |
| 70;90                    | 2197                | 0.948          | 0.87           |
| 70;95                    | 2228                | 0.943          | 0.881          |
| 75;70                    | 2080                | **0.973**      | 0.878          |
| 75;75                    | 2126                | **0.973**      | **1.05**       |
| 75;80                    | 2159                | 0.97           | 1.03           |
| 75;85                    | 2201                | 0.963          | 0.908          |
| 75;90                    | 2367                | 0.929          | 0.869          |
| 75;95                    | 2409                | 0.922          | 0.841          |
| 80;70                    | 2098                | 0.966          | 0.989          |
| 80;75                    | 2178                | 0.966          | 0.987          |
| 80;80                    | 2291                | 0.958          | 0.898          |
| 80;85                    | 2395                | 0.942          | 0.926          |
| 80;90                    | 3307                | 0.701          | 0.813          |
| 80;95                    | 3619                | 0.67           | 0.797          |
| 85;70                    | 2137                | 0.957          | 0.862          |
| 85;75                    | 2234                | 0.955          | 0.929          |
| 85;80                    | 2966                | 0.818          | 0.831          |
| 85;85                    | 3602                | 0.713          | 0.819          |
| 85;90                    | 5749                | 0.257          | 0.714          |
| 85;95                    | 8192                | 0.155          | 0.703          |
| 90;70                    | 2188                | 0.95           | 0.796          |
| 90;80                    | 3458                | 0.737          | 0.804          |
| 90;85                    | 7820                | 0.27           | 0.714          |
| 90;90                    | 14934               | 0.0769         | 0.762          |

\(^a\) Total number of clusters obtained at a clustering threshold pair

\(^b\) External clustering measure for the clustering threshold pair, bold highlights the top two values in the column

\(^c\) Internal clustering measure for the clustering threshold pair, bold highlights the top two values in the column