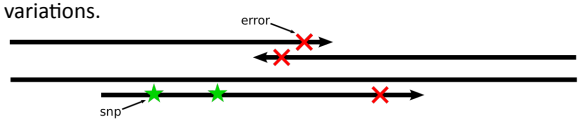
Quality-aware error correction of sequencing reads

David Kelley, Michael Schatz, and Steven Salzberg
Center for Bioinformatics and Computational Biology, Computer Science, U. Maryland College Park

Introduction

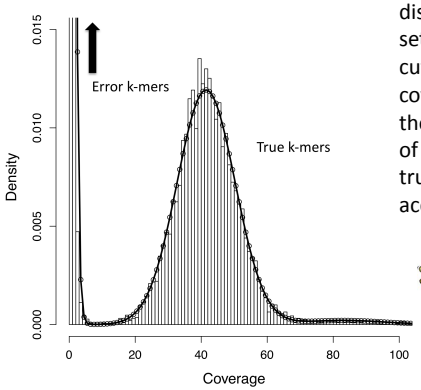
Massively parallel DNA sequencing technologies have permeated nearly all areas of biological research. Illumina reads, typically 35-125 bp, may contain errors at rates of 1-2%, particularly at the 3' ends of reads. These characteristics complicate a variety of important bioinformatics tasks; for example finding overlaps between sequencing reads before assembly and aligning reads to a reference genome to find variations.



Quake is a program to correct substitution errors in deep coverage of sequencing reads that uses the k-mer coverage framework of previous methods, but incorporates quality values and rates of specific nt to nt miscalls learned from the data, which greatly improves correction accuracy.
Quake is available open source from *http://www.cbcb.umd.edu/software/quake*.

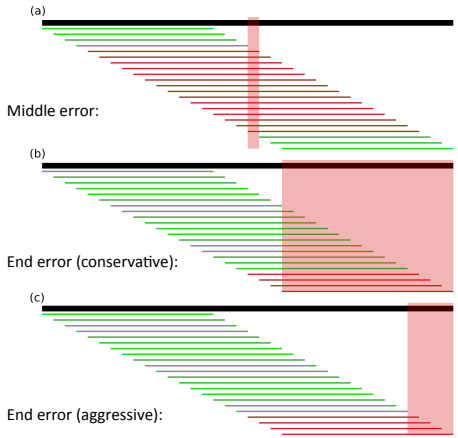
Q-mer coverage

Quake's first step is to count k-mers in the reads using the quality values to devalue k-mers with low quality bases, a process we refer to as q-mer counting. A histogram of q-mer counts shows a mixture of two distributions – true k-mers and error k-mers – that must be separated. We model these two distributions and set the separating cutoff to the coverage at which the likelihood ratio of error k-mer to true k-mer is at an acceptable level.

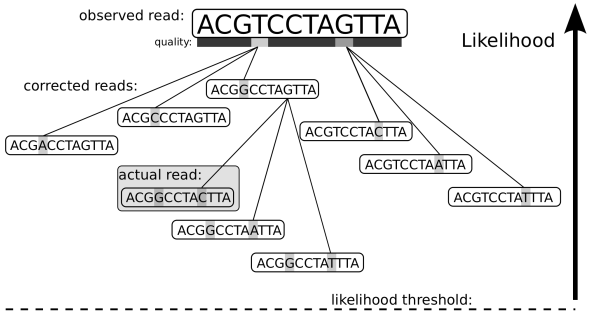


Localizing errors

Reads containing untrusted error k-mers become candidates for correction. The pattern of untrusted k-mers generally localizes the sequencing error(s) to a small region.



Correction search



Every set of corrections can be assigned a likelihood. We search this space in order of decreasing likelihood for a set of corrections that makes all k-mers in the read trusted. If no trusted corrections are found, we try to trim the read. If two trusted sets of corrections within a likelihood factor of 10 are found, we abandon the read. If the read is of very poor quality, we abandon the read before searching.

Error model

Our goal is to assign a likelihood to a set of corrections. Let *A* be the actual sequence covered by a read and *O* be the observed sequence.

1. Assume independence of sequencing errors.
2. Use the quality values to find the probability p_i that a base is correct.
3. Learn the rate at which nt's are miscalled as other nt's from the data at different quality values $E_q(a_i, o_i)$.

$$P(O_i = o_i | A_i = a_i) = \begin{cases} p_i & \text{if } o_i = a_i \\ (1 - p_i) E_q(a_i, o_i) & \text{otherwise} \end{cases}$$

Results

We simulated 40x coverage of *E. coli* using real 124 bp read quality values to simulate errors, providing 296577 error reads.

	Corrections	Trim corrections	Mis-corrections	Error reads kept
Quake	283687	6592	243	460
SOAP	276770	2942	7019	5490
EULER	228316	16577	3763	414
Shrec	165943	0	33140	96626

We assembled 152x coverage of real 36 bp *E. coli* reads using Velvet, which does not have a stand-alone error correction module, but instead performs extensive correction on the de Bruijn graph.

	Contigs	N50	N90	Scaffolds	N50	N90
Uncor	398	94827	17503	380	95365	23869
Cor	345	94831	25757	332	95369	26561

We sampled 35x coverage of real 36 bp *E. coli K12* reads, aligned them with Bowtie to strain *536*, and called SNPs using the SAMtools pileup program.

	SNPs	Mean cov
Uncor	77794	24.58
Cor	78865	26.54

Acknowledgement: This work was supported in part by NIH grant R01-LM006845.