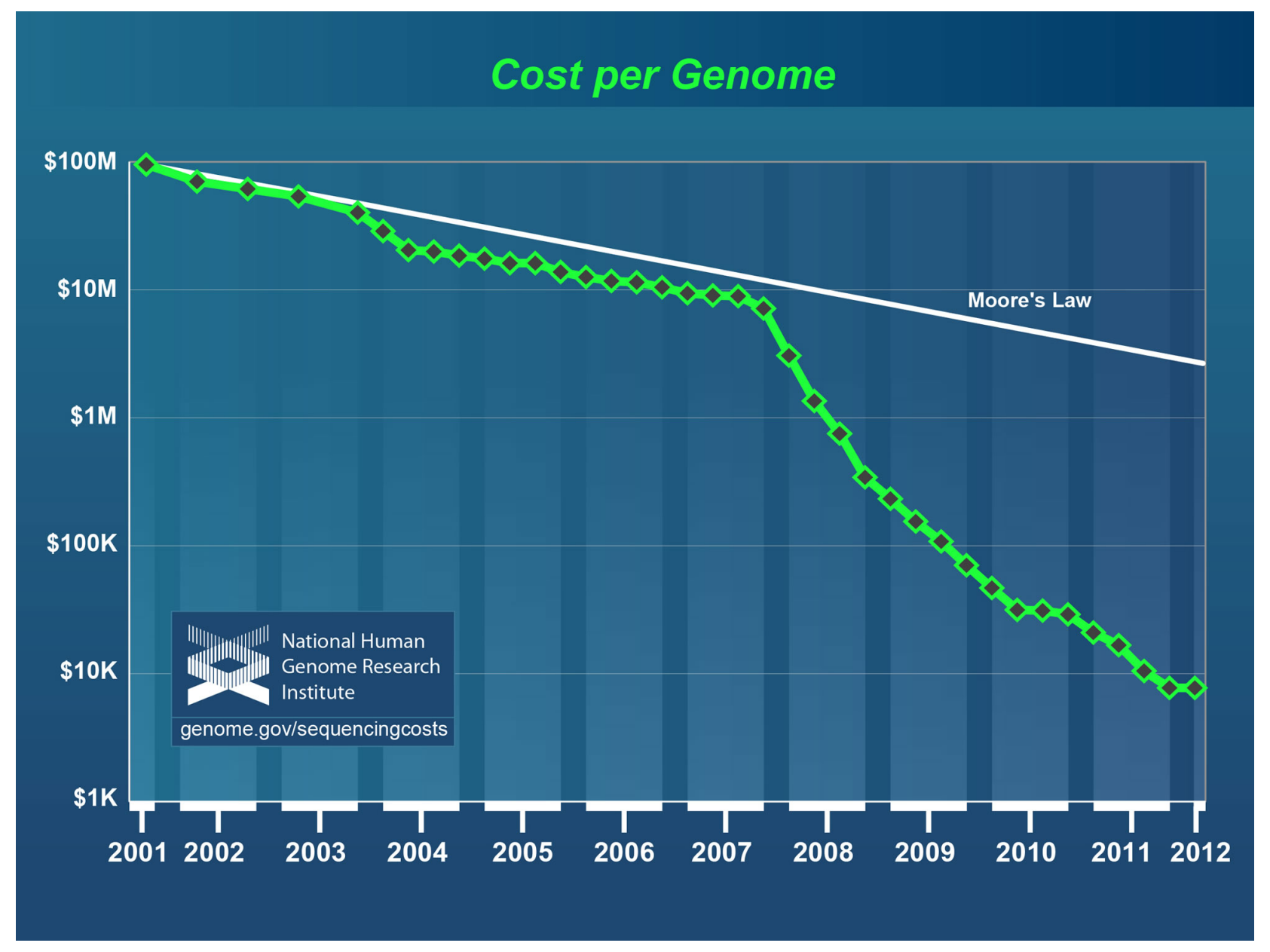
A new lossy compression algorithm for quality scores based on rate distortion theory

Idoia Ochoa, Himanshu Asnani, Dinesh Bharadia, Mainak Chowdhury, Tsachy Weissman and Golan Yona
Electrical Engineering, Stanford University

Introduction

Next Generation Sequencing has reduced the time and cost required for sequencing.



As a result, large amounts of sequencing data are being generated. A typical sequencing data file may occupy tens or even hundreds of gigabytes of disk space, prohibitively large for many users, who have to download, store and analyze the data.

This data is presented in the widely accepted FASTQ format, which consists of millions of entries. Each entry is composed of both the nucleotide sequence and per-base quality scores that indicate the level of confidence in the readout of these sequences.

Example of an entry in the FASTQ format:

```
@SRR001666.1
GATTTGGGGTTCAAAGCAGTGCAAGC
+
IIIHIIHABBBAA=2))!!!(!!!((
```

Motivation

Quality scores account for about half of the required disk space, and therefore the compression of the quality scores can significantly reduce storage requirements and speed up analysis and transmission of sequencing data.

Unlike header lines and nucleotide sequences, quality scores are particularly difficult to compress, due to their higher entropy and larger alphabet.

Quality scores are important and very useful in many downstream applications such as trimming (used to remove untrusted regions), alignment or Single Nucleotide Polymorphism (SNP) detection, among others. However, they significantly increase the size of the files storing raw sequencing data.

With this in mind, we design a lossy compression algorithm for the quality scores, that minimizes the Mean Square Error (MSE) for a given rate (bits per quality score), specified by the user.

Quality Scores

A quality score Q is the integer mapping of P (the probability that the corresponding base call is incorrect). The higher the quality score, the higher the reliability of the corresponding base. It is normally represented in the following standards:

- *Sanger or Phred* scale: $Q = -10\log_{10} P$.
- *Solexa* scale: $Q = -10\log_{10} \frac{P}{1-P}$.

Different NGS technologies use different scales, *Phred + 33*, *Phred + 64* and *Solexa + 64* being the most common ones.
For example, *Phred + 33* corresponds to values of Q in the range $[33:73]$.

The compression Method

Denote the quality score sequences presented in a FASTQ file by $\{\mathbf{Q}_i\}_{i=1}^N$, where $\mathbf{Q}_i = [Q_i(1), \ldots, Q_i(n)]$.

We assume each quality score vector is i.i.d. as $P_{\mathbf{Q}} \sim \mathcal{N}(\mu_{\mathbf{Q}}, \Sigma_{\mathbf{Q}})$.

This is justified by the fact that, given a vector source with a particular covariance structure, the gaussian multivariate source is the least compressible and, further, a code designed under the gaussian assumption will perform at least as well on any other source of the same covariance.

Due to the correlation of quality scores within an entry, $\Sigma_{\mathbf{Q}}$ is not in general a diagonal matrix. We then perform the singular value decomposition ($\Sigma_{\mathbf{Q}} = VSV^T$), and generate a set of new quality vectors $\{\mathbf{Q'}_i\}_{i=1}^N$, where $\mathbf{Q'}_i = V^T(\mathbf{Q}_i - \mu_{\mathbf{Q}})$. Due to the gaussian nature of the vectors, we have $\mathbf{Q'}_i \sim \mathcal{N}(\mathbf{0}, S)$, where S is diagonal.

We can now apply a well known result on rate distortion theory, and optimally allocate nR bits to reduce the MSE by solving the following optimization problem:

$$\min_{\rho=[\rho_1,\cdots,\rho_n]} \frac{1}{n}\sum_{j=1}^{n} \sigma_j^2 2^{-2\rho_j} \qquad (1)$$

$$\text{s.t.} \sum_{j=1}^{n} \rho_j \leq nR, \qquad (2)$$

where ρ_j represents the number of bits allocated to the j^{th} position of $\mathbf{Q'}_i$, for $i \in [1:N]$.

Finally, we normalize each component of $\mathbf{Q'}_i$, for $i \in [1:N]$, to a unit variance Gaussian and map them to decision regions representable in ρ_j bits.

The decision regions that minimize the MSE for different values of ρ and their representative values are found offline from a Lloyd Max procedure on a scalar gaussian distribution with mean zero and variance one.

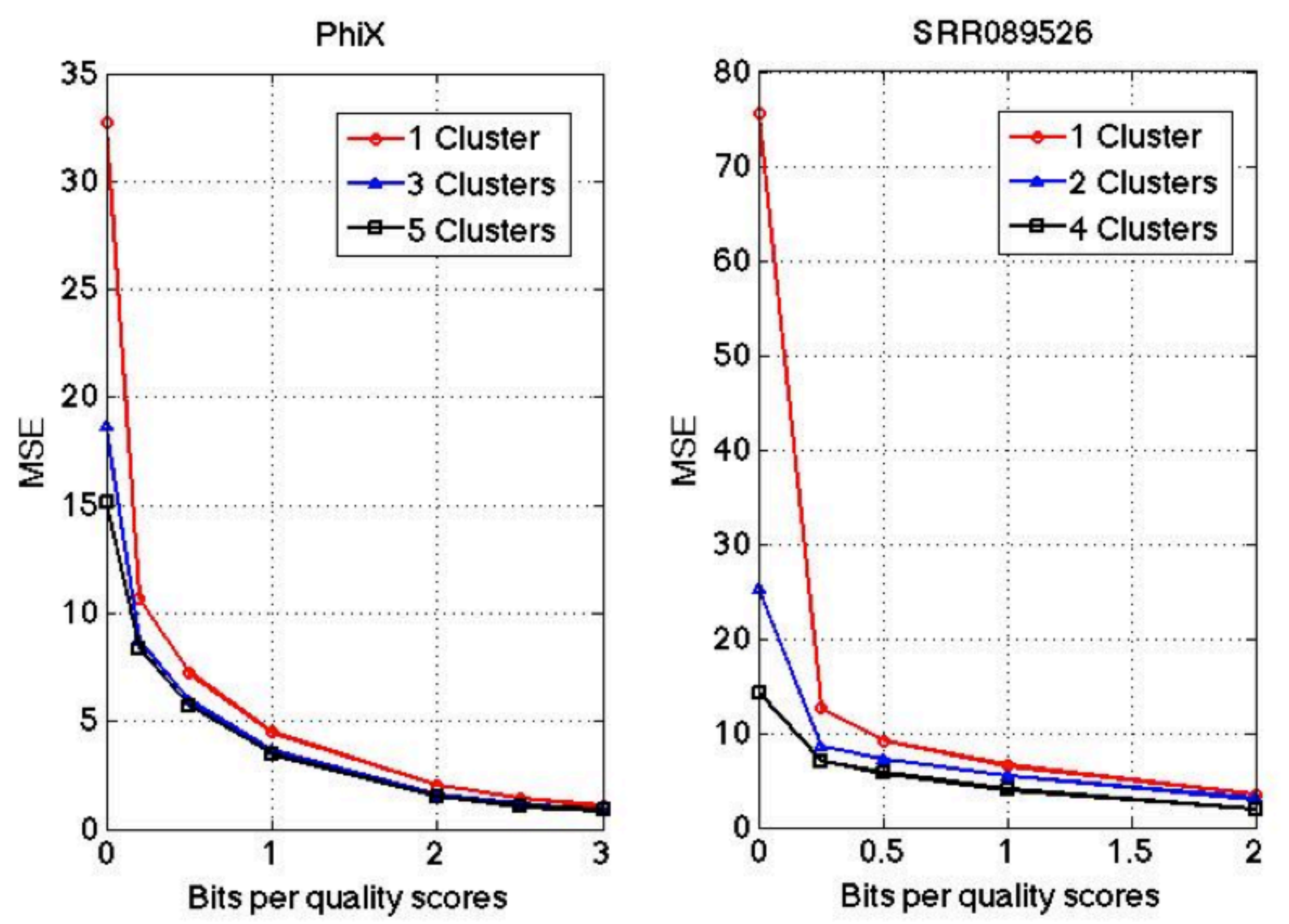
To improve the performance, we allow clustering prior to compression.

Simulation Results

- Data used for simulations:

	Numer of Reads	Length of each read	Quality Scores
PhiX (Virus)	13,310,768	100	66:98
SRR089526 (H. Sapiens)	23,892,841	48	33:73

- Rate vs MSE:



- Alignment with the Burrows Wheeler Aligner (BWA) followed by SNP calling with Samtools for data *SRR089526*:

T.P., F.P. and F.N. stand for true positives (detected both with the original FASTQ file and the reconstructed one), false positive (detected only with the reconstrued FASTQ file) and false negative (detected only with the original FASTQ file), respectively.

The selectivity parameter is computed as T.P./(T.P. + F.P.), and sensitivity as T.P./(T.P. + F.N.).

Four clusters						
R	MSE	T.P.	F.P	F.N.	Selectivity (%)	Sensitivity (%)
0	14.25	54708	4868	5719	91.82	90.53
0.25	6.99	58963	4722	1464	92.58	97.57
0.50	5.76	59110	4480	1317	92.95	97.82
1.00	3.99	59318	4028	1109	93.64	98.16
2.00	1.91	59592	3229	835	94.85	98.61

Software Availability

The software is freely available for download at:

- http://www.stanford.edu/~iochoa/QualComp.html

Acknowledgement

- This work is supported in part by La Caixa Fellowship and by the NSF Center for Science of Information.