

Quake: quality-aware detection and correction of sequencing errors

Kelley et al (2010) Genome Biology

Age Tats

Bioinformatics Journal Club

31.01.2011

SOFTWARE

Open Access

Quake: quality-aware detection and correction of sequencing errors

David R Kelley^{1*}, Michael C Schatz², Steven L Salzberg¹

Abstract

We introduce Quake, a program to detect and correct errors in DNA sequencing reads. Using a maximum likelihood approach incorporating quality values and nucleotide specific miscall rates, Quake achieves the highest accuracy on realistically simulated reads. We further demonstrate substantial improvements in *de novo* assembly and SNP detection after using Quake. Quake can be used for any size project, including more than one billion human reads, and is freely available as open source software from <http://www.cbcb.umd.edu/software/quake>.

¹Center for Bioinformatics and Computational Biology, Institute for Advanced Computer Studies, and Department of Computer Science, University of Maryland, College Park, MD 20742, USA. ²Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA.

- Primary errors from Illumina sequencers are substitution errors, at rates of 0.5-2.5%, with errors rising in frequency at the 3' ends of reads.

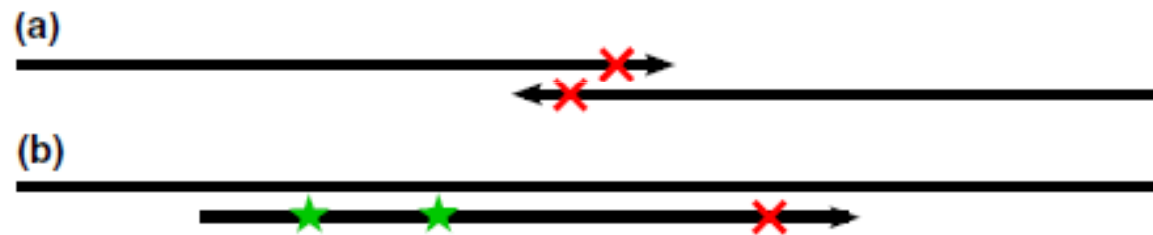


Figure 1 Alignment difficulty. Detecting alignments of short reads is more difficult in the presence of sequencing errors (represented as X's). (a) In the case of genome assembly, we may miss short overlaps between reads containing sequencing errors, particularly because the errors tend to occur at the ends of the reads. (b) To find variations between the sequenced genome and a reference genome, we typically first map the reads to the reference. However, reads containing variants (represented as stars) and sequencing errors will have too many mismatches and not align to their true genomic location.

Error correction methods

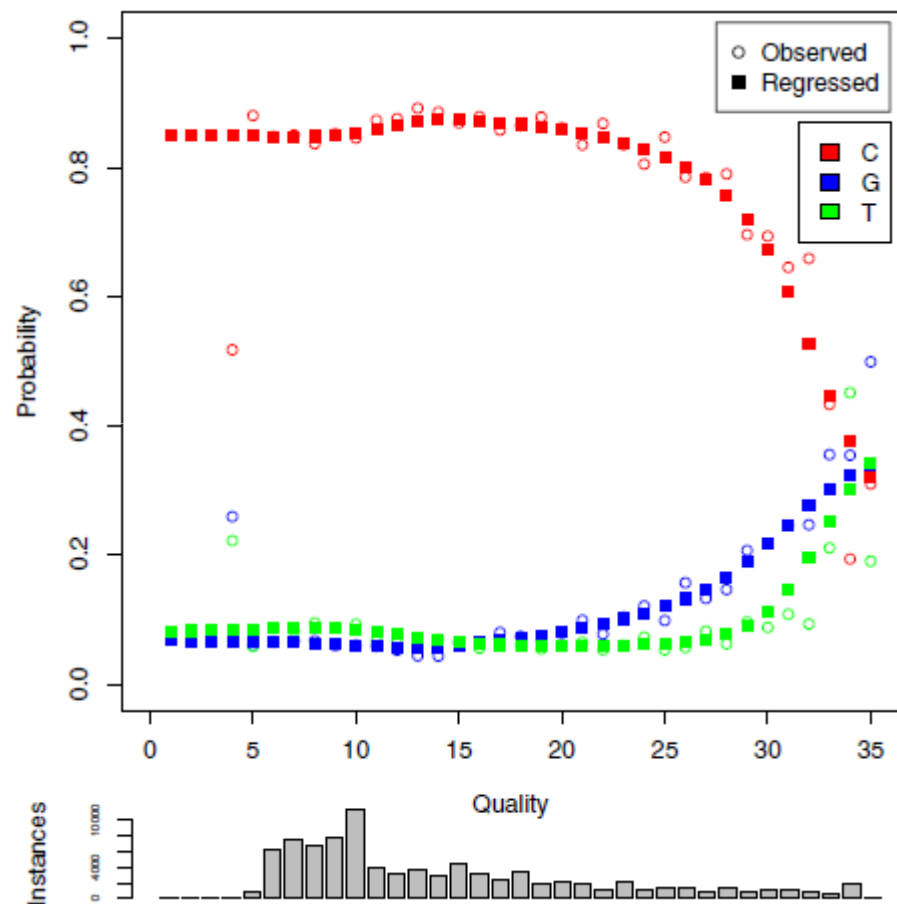
- Aim is to minimize edit distance from untrusted k -mer to trusted k -mer
 - Trusted k -mers = high coverage, highly likely to occur in the genome
 - Untrusted k -mers = low coverage, occurring just once or twice

Error correction methods

- Edit distance methods (EULER assembler)
 - 1) determines a coverage cutoff to separate low and high coverage k -mers
 - 2) corrects reads with low coverage k -mers by making nucleotide edits to the read that reduce the number of low coverage k -mers until all k -mers in the read have high coverage

- Edit distance method treats all bases the same regardless of quality
 - Quality values can be useful even if they only rank one base as more likely to be an error as another
- Edit distance method treats all error substitutions as equally likely
 - Illumina technology cause certain miscalls to be more likely than others: A<->C; G<->T

Figure 2 Adenine error rate. The observed error rate and predicted error rate after nonparametric regression are plotted for adenine by quality value for a single lane of Illumina sequencing of *Megachile rotundata*. The number of training instances at each quality value are drawn as a histogram below the plot. At low and medium quality values, adenine is far more likely to be miscalled as cytosine than thymine or guanine. However, the distribution at high quality is more uniform.



Quake

- Specifically intended for Illumina sequencing reads
- For sequencing projects >15x coverage
- Uses quality values based weighting of k -mer counts to choose cutoff between untrusted and trusted k -mers
- Inappropriate for applications where low coverage does not necessarily implicate a sequencing error (metagenomics, RNA-Seq, ChIP-Seq)

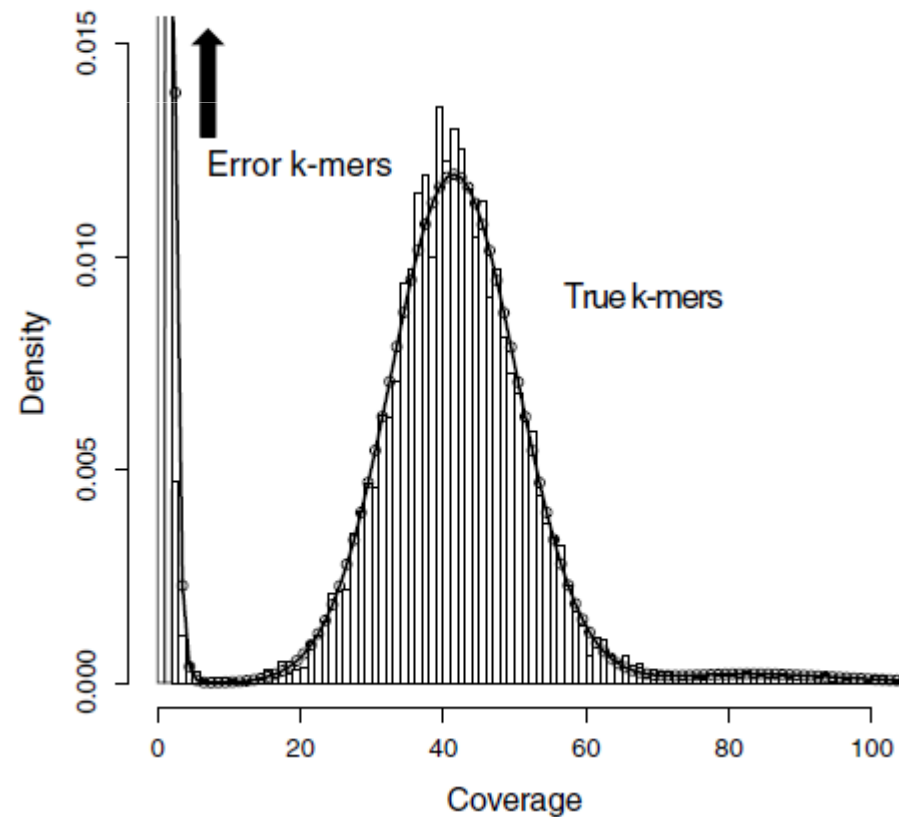
Quake pipeline (1)

- Counting the number of k -mers in the sequencing reads
 - choose k such that $2G/4^k = 0.01$, which simplifies to $k = \log_4 200G$
 - for *E. coli* genome (5 Mbp) $k=15$
 - for human genome (3 Gbp) $k=19$
 - Hadoop cluster: 20 cores, 40 GB RAM, 3.6 TB local disk
- *Q-mer counting*: increment k -mer coverage by the product of the probabilities that the base calls in the k -mer are correct as defined by the quality values

Quake pipeline (2)

- Coverage cutoff

Figure 3 *k*-mer coverage. 15-mer coverage model fit to 76x coverage of 36 bp reads from *E. coli*. Note that the expected coverage of a *k*-mer in the genome using reads of length *L* will be $\frac{L-k+1}{L}$ times the expected coverage of a single nucleotide because the full *k*-mer must be covered by the read. Above, *q*-mer counts are binned at integers in the histogram. The error *k*-mer distribution rises outside the displayed region to 0.032 at coverage two and 0.691 at coverage one. The mixture parameter for the prior probability that a *k*-mer's coverage is from the error distribution is 0.73. The mean and variance for true *k*-mers are 41 and 77 suggesting that a coverage bias exists as the variance is almost twice the theoretical 41 suggested by the Poisson distribution. The likelihood ratio of error to true *k*-mer is one at a coverage of seven, but we may choose a smaller cutoff for some applications.



Quake pipeline (3)

- Localizing errors – decreases the runtime of the algorithm

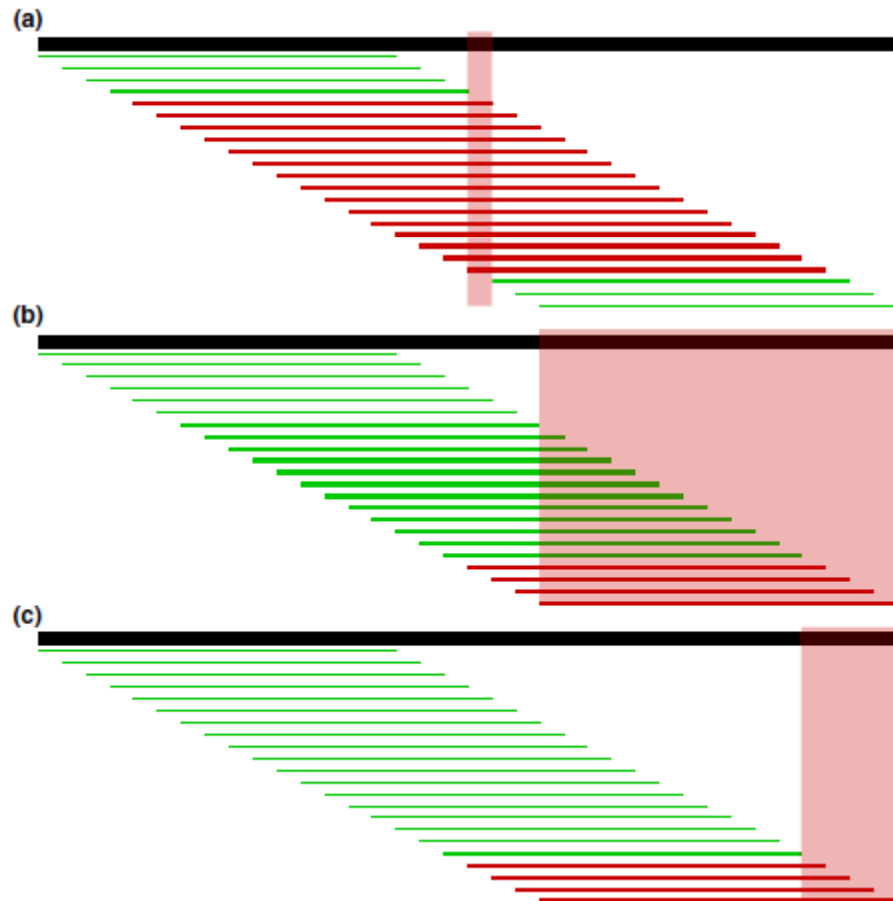


Figure 4 Localize errors. Trusted (green) and untrusted (red) 15-mers are drawn against a 36 bp read. In (a), the intersection of the untrusted k -mers localizes the sequencing error to the highlighted column. In (b), the untrusted k -mers reach the edge of the read, so we must consider the bases at the edge in addition to the intersection of the untrusted k -mers. However, in most cases, we can further localize the error by considering all bases covered by the right-most trusted k -mer to be correct and removing them from the error region as shown in (c).

Quake pipeline (4)

- Search for the maximum likelihood set of corrections that makes all k -mers overlapping the region trusted
- The likelihood of a set of corrections to a read is defined by a probabilistic model of sequencing errors incorporating the read's quality values as well as the rates at which nucleotides are miscalled as different nucleotides.
- Correction proceeds by examining changes to the read in order of decreasing likelihood until a set of changes making all k -mers trusted is discovered.

$O = O_1, O_2, \dots, O_N$

$A = A_1, A_2, \dots, A_N$

observed nucleotides of the read

actual nucleotides of the sequenced fragment

$$p_i = 1 - 10^{-\frac{q_i}{10}}$$

The probability that the nucleotide at position i is accurate, q_i is the corresponding quality value

$E_q(x, y)$

The probability that the base call y is made for the nucleotide x at quality value q given that there has been a sequencing error

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

$$E_q(x, y) = \frac{\sum_i C_{q_i}(x, y)N(q_i; q, 2)}{\sum_i C_{q_i}(x)N(q_i; q, 2)}$$

$$E_q(x, y) = \frac{\sum_i C_{q_i}(x, y) N(q_i; q, 2)}{\sum_i C_{q_i}(x) N(q_i; q, 2)}$$

$C_q(x, y)$ the number of times actual nucleotide x was observed as error nucleotide y at quality value q

$C_q(x)$ the number of times actual nucleotide x was observed as an error at quality value q

$N(q; u, s)$ the probability of q from a Gaussian distribution with mean u and standard deviation s

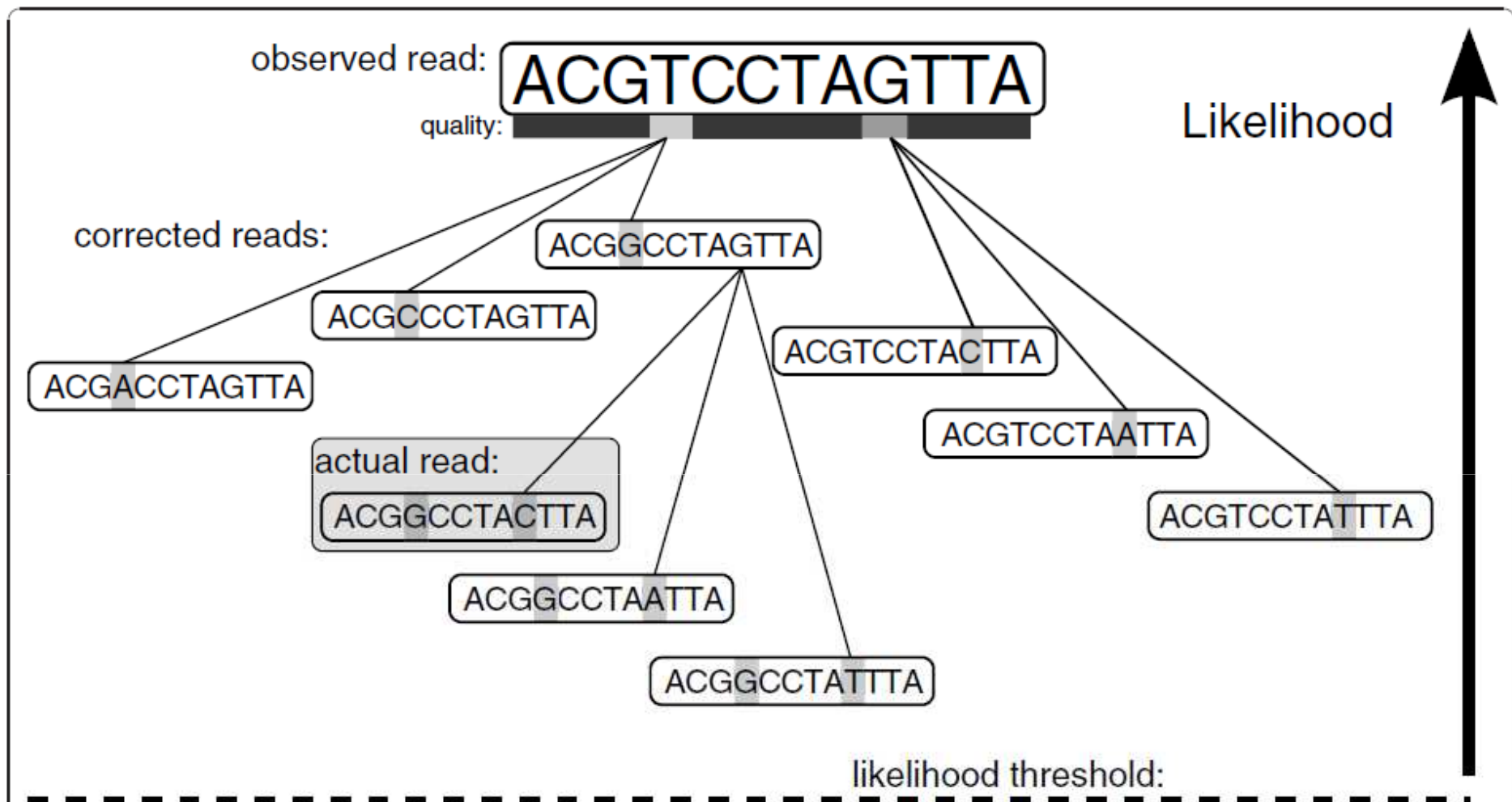


Figure 6 Correction search. The search for the proper set of corrections that change an observed read with errors into the actual sequence from the genome can be viewed as exploring a tree. Nodes in the tree represent possible corrected reads (and implicitly sets of corrections to the observed read). Branches in the tree represent corrections. Each node can be assigned a likelihood by our model for sequencing errors as described in the text. Quake's algorithm visits the nodes in order of decreasing likelihood until a valid read is found or the threshold is passed.

Accuracy (1)

Table 1 Simulated 36 bp E. Coli

	Corrections	Trim corrections	Mis-corrections	Error reads kept	Time (min)
Quake	1035709.4	26337.0	1744.0	5537.0	14.2
SOAPdenovo	969666.4	120529.0	3912.8	9288.4	12.4
Shrec	964431.8	0.0	165422.0	41733.6	87.6

Table 2 Simulated 124 bp E.coli

	Corrections	Trim corrections	Mis-corrections	Error reads kept	Time (min)
Quake	283769.4	6581.2	243.0	393.6	11.8
SOAPdenovo	276770.4	2942.6	7019.4	5490.2	16.9
Shrec	165942.7	0.0	33140.3	96626.7	97.1
EULER	228316.4	16577.4	3763.0	414.8	6.9

Accuracy (2)

- Human genome – simulated 325M 124 bp reads from chr 1 (34x coverage)
- 18-mers
- Corrected 89.6% of error reads (11% more than SOAPdenovo)
- 64% less mis-corrections than SOAPdenovo
- Kept 15% less error reads than SOAPdenovo

Accuracy (3)

- More uncorrected reads in human data than in *E. coli* data is caused by repetitive elements
- 13.8% of all single base 18-mer mutations in chr1 create another 18-mer that also exists in human genome (11.1% for 19-mers)

Genome assembly

Table 3 Velvet E.coli assembly

	Contigs	N50	N90	Scaffolds	N50	N90	Breaks	Miscalls	Cov
Uncorrected	398	94827	17503	380	95365	23869	5	456	0.9990
Corrected	345	94831	25757	332	95369	26561	4	4	0.9992

Table 4 SOAPdenovo bee assembly

Assembly	Trimmed Only	Corrected	Removed
Uncorrected Corrected	146.0 M	-	12.9 M
SOAPdenovo Corrected	134.4 M	15.7 M	15.6 M
Quake	146.9 M	16.5 M	13.0 M

Table 4 continues

Assembly	Contigs	N50	N90	Scaffolds	N50	N90	Reads
Uncorrected Corrected	312414	2383	198	90201	37138	9960	167.3 M
SOAPdenovo Corrected	188480	4051	515	36525	36525	9162	164.8 M
Quake	189621	4076	514	37279	37014	9255	167.3 M

SNP detection

1)

Table 5 E. coli SNP calling

Method	Reads mapped	SNPs	Recall	Precision
Two mismatch uncorrected	3.39 M	79,748	0.746	0.987
Two mismatch corrected	3.51 M	80,796	0.755	0.987
Quality-aware uncorrected	3.56 M	85,071	0.793	0.984
Quality-aware corrected	3.55 M	85,589	0.798	0.984

We called SNPs in 35× coverage of 36 bp reads from *E. coli* K12 by aligning the reads to a close relative genome *E. coli* 536 with Bowtie using both a two mismatch and quality-aware alignment policy and calling SNPs with SAMtools pileup. SNPs were validated by comparing the *E. coli* K12 and *E. coli* 536 reference genomes directly. Under both alignment policies, correcting the reads with Quake helps find more true SNPs.

2) Korean individual (1.7 billion reads)

2% more SNPs were called. Read coverage on SNP locations increased 4.8% compared to uncorrected reads

Conclusions

- Quake corrects more reads more accurately than previous methods
- q -mer counting, which uses quality values as a means of weighting each k -mer, separates trusted and untrusted k -mers distributions better
- Improves genome assembly using Velvet and SOAPdenovo
- Improves SNP calling
- Provides statistics to compare and analyze different experiments/lines/runs