SeqMap: mapping massive amount of oligonucleotides to the genome

Hui Jiang and Wing Hung Wong

Bioinformatics. 2008 Oct 15;24(20):2395-6.

- high-throughput sequencing
- oligo mapping tools
- mismatches and ins/del

Jclub 14.10.2008

SeqMap: introduction

- http://biogibbs.stanford.edu/~jiangh/SeqMap/

- Command-line tool for mapping large amount of short oligonucleotides to the reference genome

- Data generation:
  - ✓ Illumina-Solexa and ABI-SOLiD = 50-100M reads (30-50 nt)
  - ✓ Roche 454 = 400,000 reads (200-300 nt)

- Alternatives:
  - ✓ ELAND – only for short reads, max 2 substitutions
  - ✓ SOAP – slow with larger genomes, max 2 substitutions or one gap (1-3 nt.)
  - ✓ RMAP – ungapped mapping, accounts read quality

Jiang H, Wong WH. "**SeqMap: mapping massive amount of oligonucleotides to the genome.**" *Bioinformatics*. 2008 Oct 15;24(20):2395-6.

SeqMap: method

- Flexibility in the mapping:
  - ✓Allows 5 mixed substitutions and ins/dels
  - ✓Various command-line options and output formats
  - ✓Sequences can contain ‘N’
  - ✓Sequences can have different lengths
  - ✓Operates with FASTA files

- Alternatives:
  - ✓ELAND – only for short reads, max 2 substitutions
  - ✓SOAP – slow with larger genomes, max 2 substitutions or one gap (1-3 nt.)
  - ✓RMAP – ungapped mapping, accounts read quality

Jiang H, Wong WH. “**SeqMap: mapping massive amount of oligonucleotides to the genome.**” *Bioinformatics*. 2008 Oct 15;24(20):2395-6.

SOAP: algorithm

- Store the genomic sequence in RAM. Two bits for each base, so one byte can store 4 bps
- Split reads into 4 parts - a,b,c,d, two mismatches will be distributed on at most two of the 4 parts at the same time
- Use look up table to judge how many mismatches between reference and read. To have best efficiency, the table used 3 bytes to check a fragment of 12-bp on a time. The table occupied 2^24=16Mb RAM
- Search for identical hits first, if no hits, then 1-mismatch hits will be picked up, then 2-mismatch hits, then gapped hits.

Li R., Li Y., Kristiansen K., Wang J. "**SOAP: short oligonucleotide alignment program.**" *Bioinformatics* 2008 Mar 1;24(5):713-4.

SeqMap: algorithm

- Split query sequence into several fragments
  - ✓ Using 2 mismatches and splitting into 4 fragments, two of them have always perfect matches

1				
2		MM1		MM2
3		MM1		
4				MM2
5		MM1, MM2		
6				MM1, MM2

- Size of the fragment is dynamic

- Creating hash table from the query not the genome

Jiang H, Wong WH. “**SeqMap: mapping massive amount of oligonucleotides to the genome.**” *Bioinformatics*. 2008 Oct 15;24(20):2395-6.

SeqMap: speed

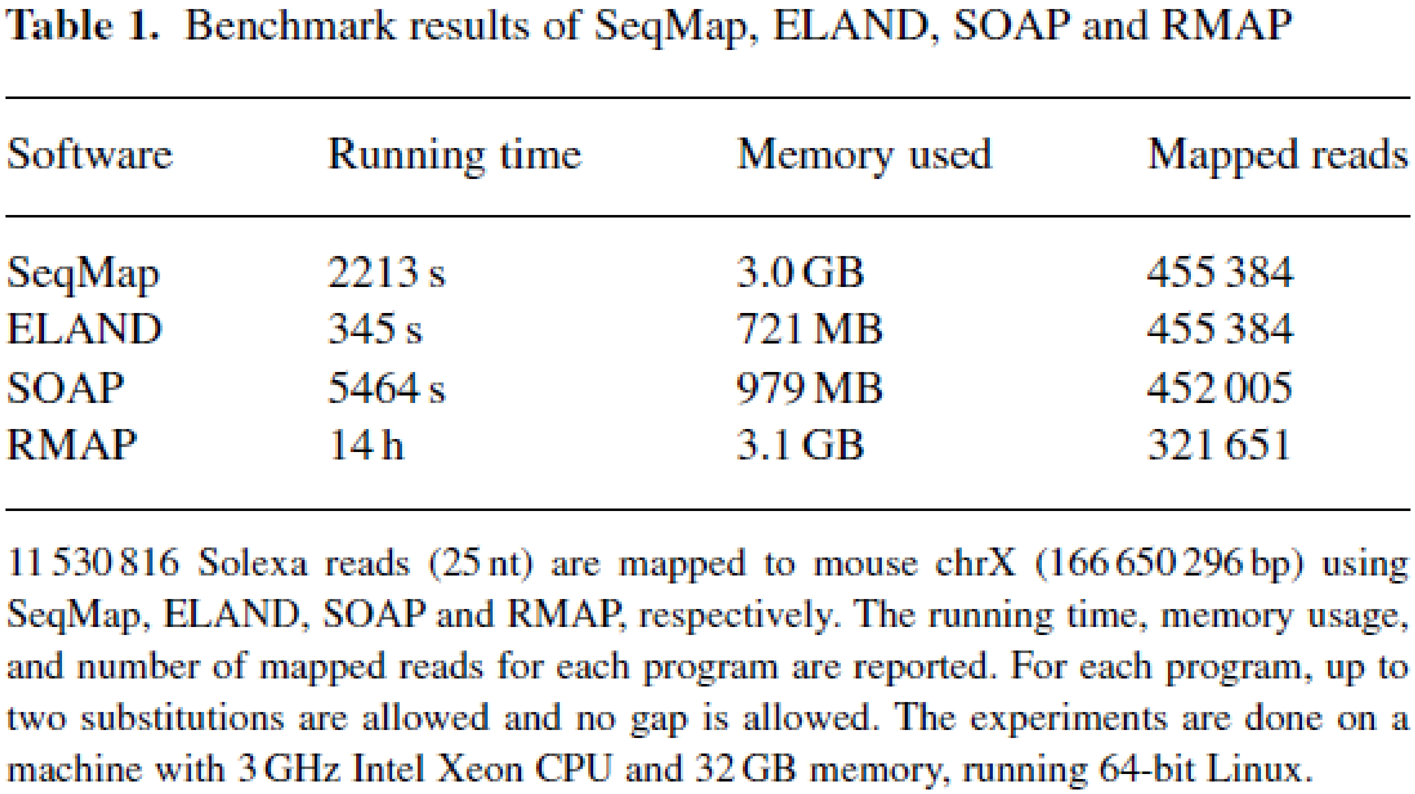


Table 1. Benchmark results of SeqMap, ELAND, SOAP and RMAP

Software	Running time	Memory used	Mapped reads
SeqMap	2213 s	3.0 GB	455 384
ELAND	345 s	721 MB	455 384
SOAP	5464 s	979 MB	452 005
RMAP	14 h	3.1 GB	321 651

11 530 816 Solexa reads (25 nt) are mapped to mouse chrX (166 650 296 bp) using SeqMap, ELAND, SOAP and RMAP, respectively. The running time, memory usage, and number of mapped reads for each program are reported. For each program, up to two substitutions are allowed and no gap is allowed. The experiments are done on a machine with 3 GHz Intel Xeon CPU and 32 GB memory, running 64-bit Linux.

Jiang H, Wong WH. “**SeqMap: mapping massive amount of oligonucleotides to the genome.**” *Bioinformatics*. 2008 Oct 15;24(20):2395-6.

SeqMap: sensitivity

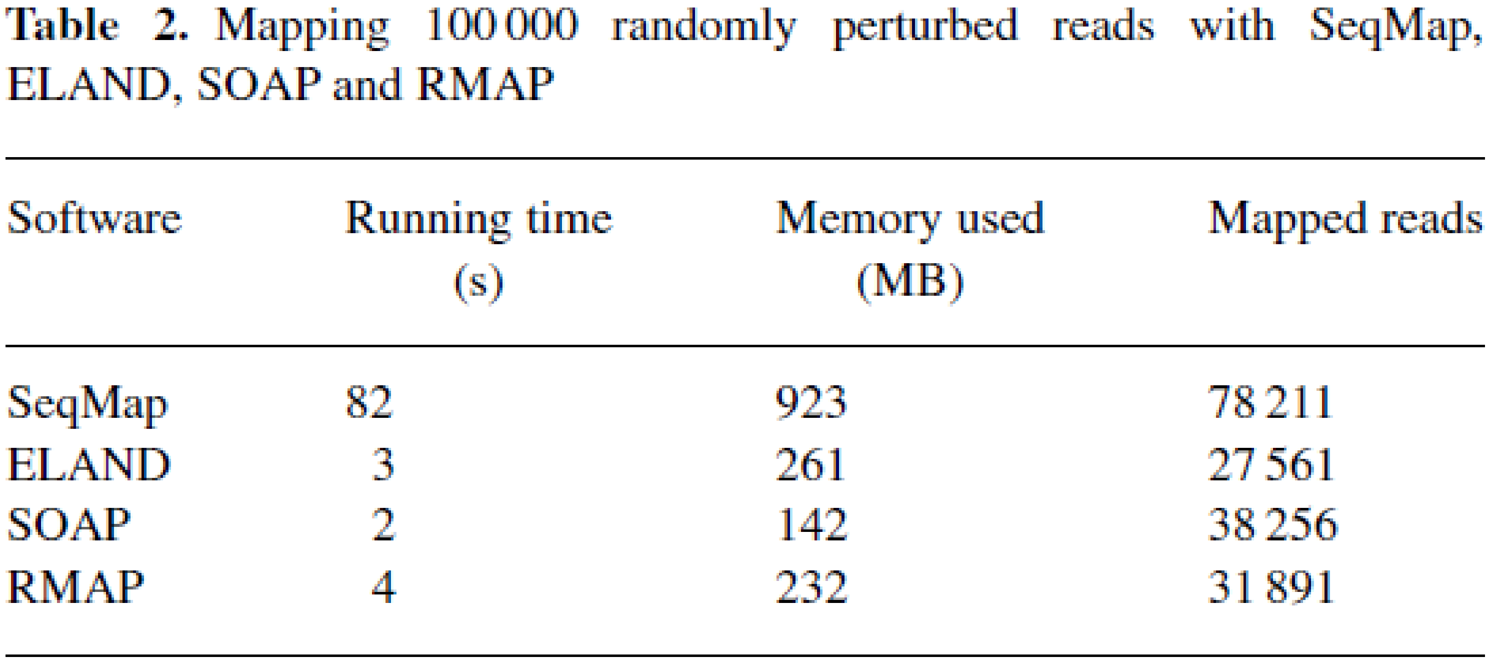


Table 2. Mapping 100 000 randomly perturbed reads with SeqMap, ELAND, SOAP and RMAP

Software	Running time (s)	Memory used (MB)	Mapped reads
SeqMap	82	923	78 211
ELAND	3	261	27 561
SOAP	2	142	38 256
RMAP	4	232	31 891

- 1Mbp random sequence, 100Kbp random substitutions, ‘N’s and ins/dels against 100K random 25-mers

Jiang H, Wong WH. “**SeqMap: mapping massive amount of oligonucleotides to the genome.**” *Bioinformatics*. 2008 Oct 15;24(20):2395-6.

SeqMap: PCR primer mapping

- 1000 primer pairs against human genome (2MM) 3896.77s (> 1 h)

- RAM = 6MB (query hash tables)

```
trans_id  trans_coord  target_seq        probe_id     probe_seq     num_mismatch strand
chr1      1166    CAAGAGGGCCCTGCAGTGCC    90560_L CAAGAGGGCCCTGCACTTCC    2       -
chr1      16663   ATTACAGGCGTGAGCCGCTG    90224_L ATTACAGGCGTGAGCCACCG    2       -
chr1      16670   TGCTGGGATTACAGGCGTGA    90588_L TGCTGGGATTACAGGCGTGA    0       -
chr1      16672   AAAGTGCTGGGATTACAGGCGT  90514_L AAAGTGCTGGGATTACAGGCGG  1       -
chr1      16681   ATCCCAGCACTTTGGGAGGC    90377_L ATCCCAGCAATTTGGGACGC    2       +
```

Jiang H, Wong WH. “**SeqMap: mapping massive amount of oligonucleotides to the genome.**” *Bioinformatics*. 2008 Oct 15;24(20):2395-6.

Substantial biases in ultra-short read data sets from high-throughput DNA sequencing

Juliane C. Dohm, Claudio Lottaz, Tatiana Borodina and Heinz Himmelbauer

Nucl. Acids Res. 2008 36:e105

- Illumina-Solexa technology
- short read error rates
- substitution types
- GC content correlation with the read coverage

Solexa sequencing biases: overview

- Solexa read lengths up to 36 bases
- >40 M reads with 3 days
- Main drawbacks of the high-throughput DNA sequencing using short reads:
  - ✓Wrong base calls
  - ✓Coverage of low-complexity regions

Dohm JC, Lottaz C, Borodina T and Himmelbauer H. "**Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.**" *Nucleic Acids Research* 2008 36:e105.

Solexa sequencing biases: research topics

- Detect biases of error positions, rates and erroneous base calls (neighboring bases and ins/dels)
- Determine the compensation of erroneous base calls by correct ones with higher coverage
- Analyze read start positions, coverage along target sequence and coverage dependencies of the local sequence characteristics
- Assess the reliability of quality values for wrong and correct base calls

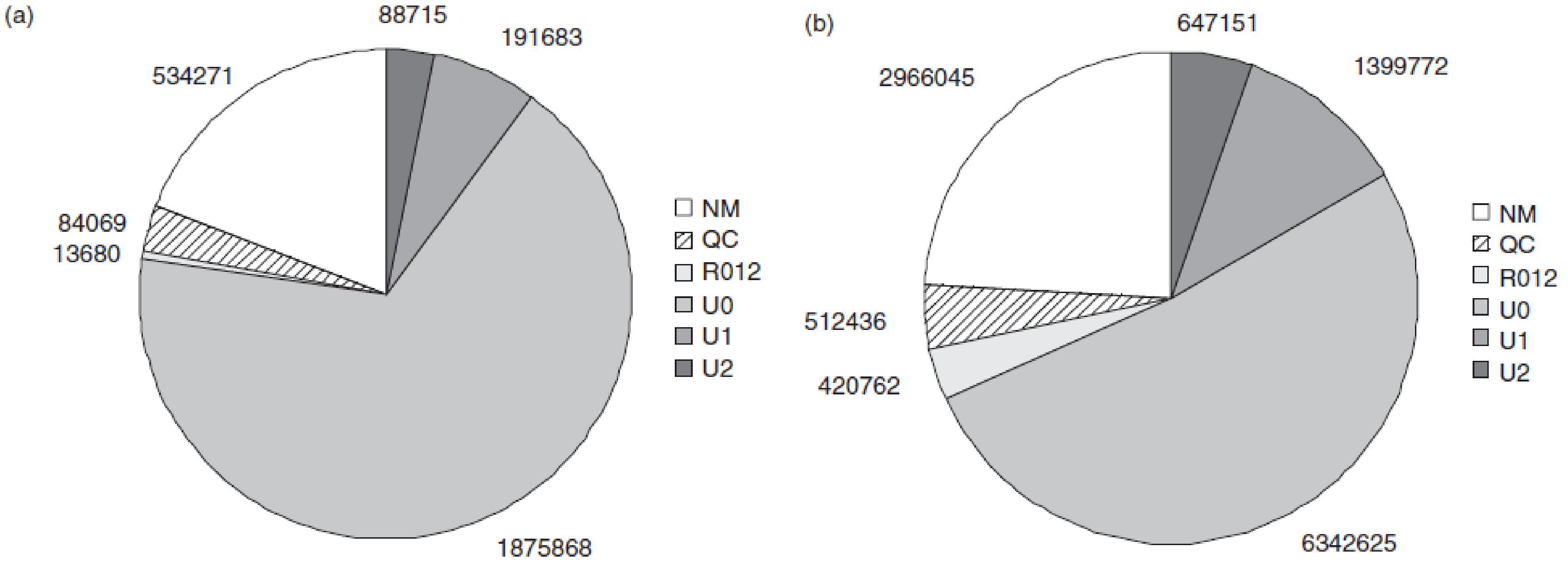
Dohm JC, Lottaz C, Borodina T and Himmelbauer H. “**Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.**” *Nucleic Acids Research* 2008 36:e105.

Solexa sequencing biases: data

- 27mer reads from **Beta vulgaris** clone ZR-47B15 (2 788 286 in total)
- 32mer reads from **Helicobacter acinonychis** (12 288 791 in total)
- ELAND software for mapping reads
- Perl scripts for ins/del detection

Dohm JC, Lottaz C, Borodina T and Himmelbauer H. "**Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.**" *Nucleic Acids Research* 2008 36:e105.

Solexa sequencing biases: ELAND results

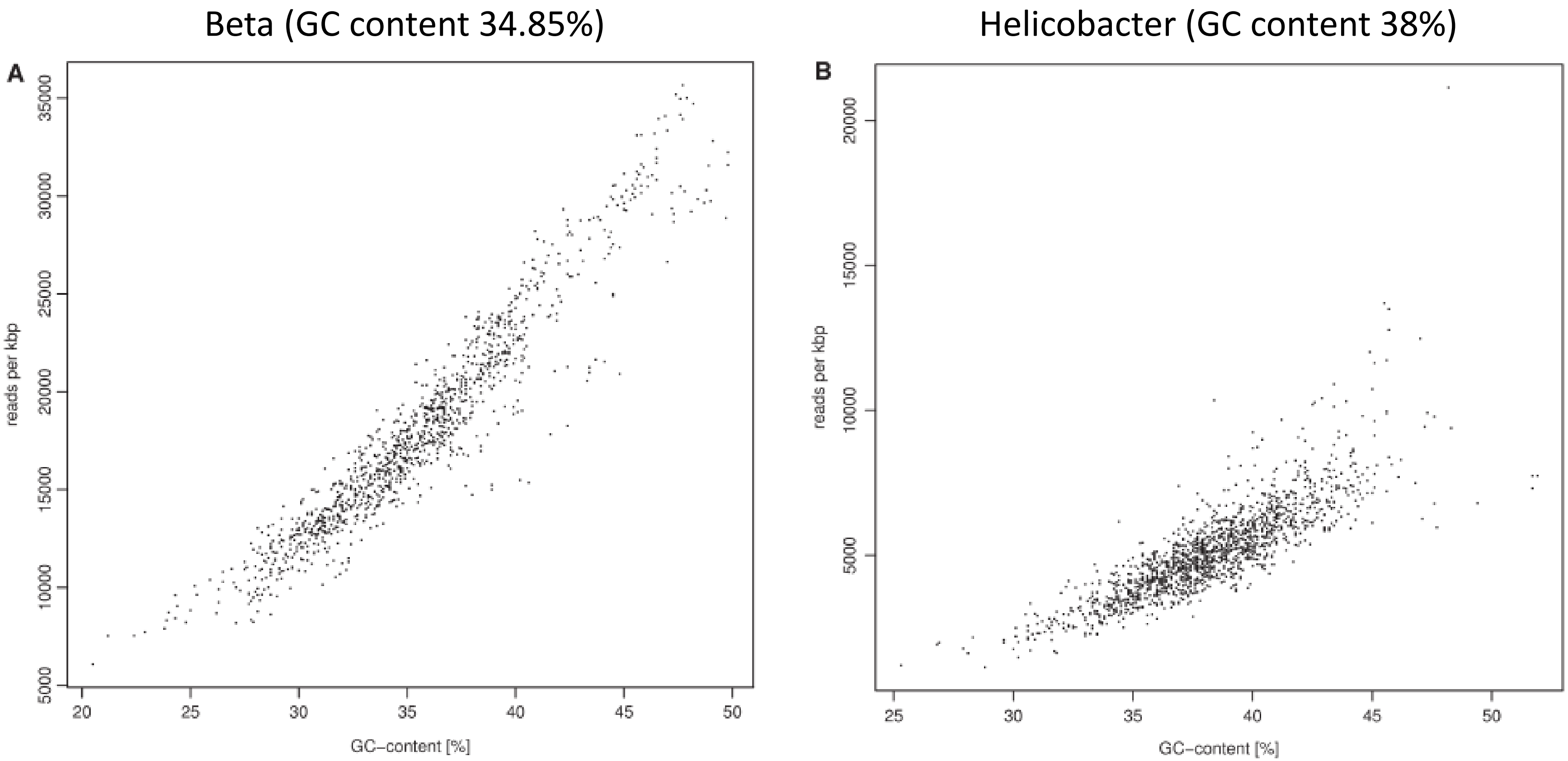

Beta

Helicobacter

ELAND categories are:
QC - no matching done because of low quality of the read (more than two positions with quality score=5),
NM - no match found;
U0 - unique exact match found;
U1 - unique match with one error;
U2 - unique match with two errors;
R0 - multiple exact matches found;
R1 - multiple matches with one error;
R2 - multiple matches with two errors.

Dohm JC, Lottaz C, Borodina T and Himmelbauer H. “**Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.**” *Nucleic Acids Research* 2008 36:e105.

Solexa sequencing biases: GC content



Beta (GC content 34.85%)

Helicobacter (GC content 38%)

Each data point corresponds to the number of reads recorded for a 1-kbp window (shift of 100 bp in Beta and 1 kbp in Helicobacter)

Dohm JC, Lottaz C, Borodina T and Himmelbauer H. “**Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.**” *Nucleic Acids Research* 2008 36:e105.

Solexa sequencing biases: overall coverage

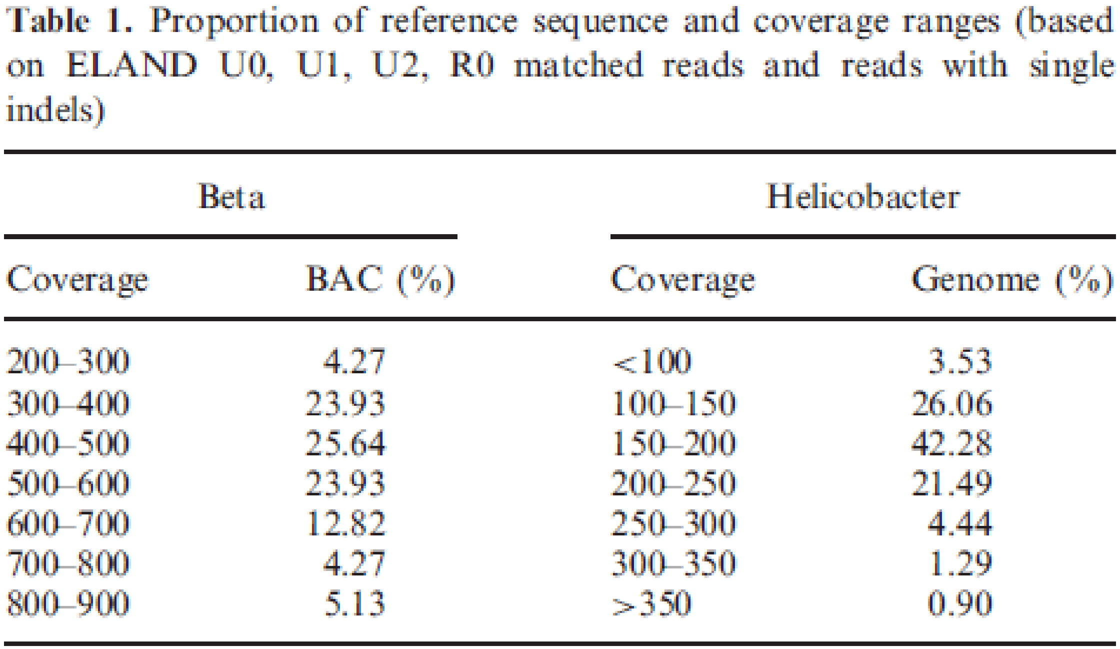


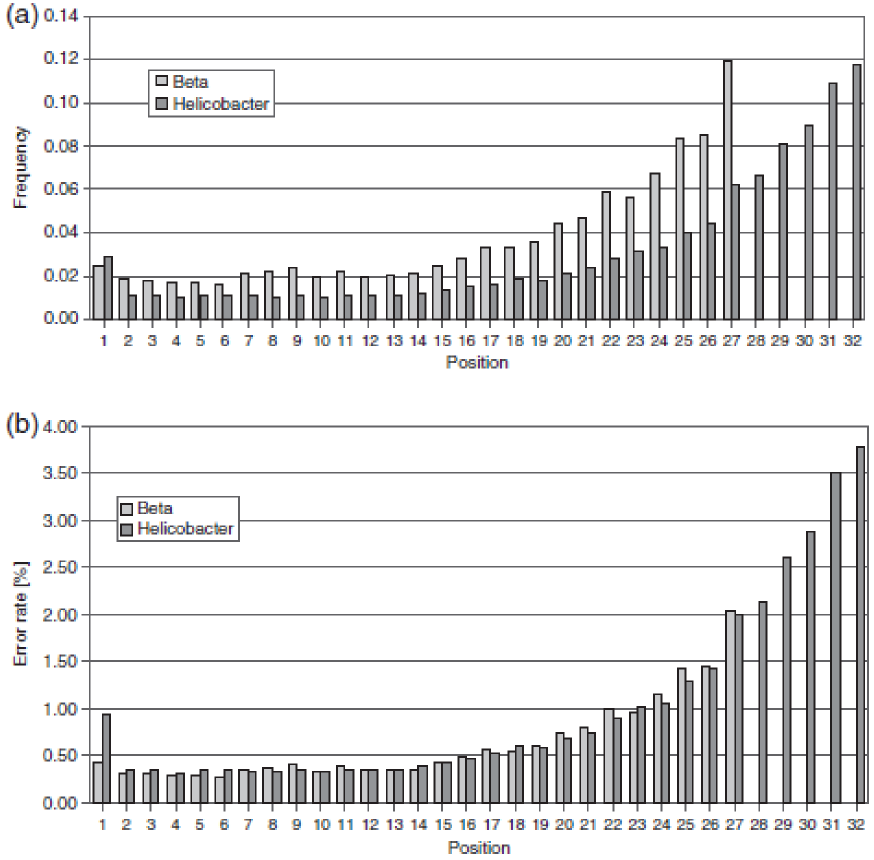
Table 1. Proportion of reference sequence and coverage ranges (based on ELAND U0, U1, U2, R0 matched reads and reads with single indels)

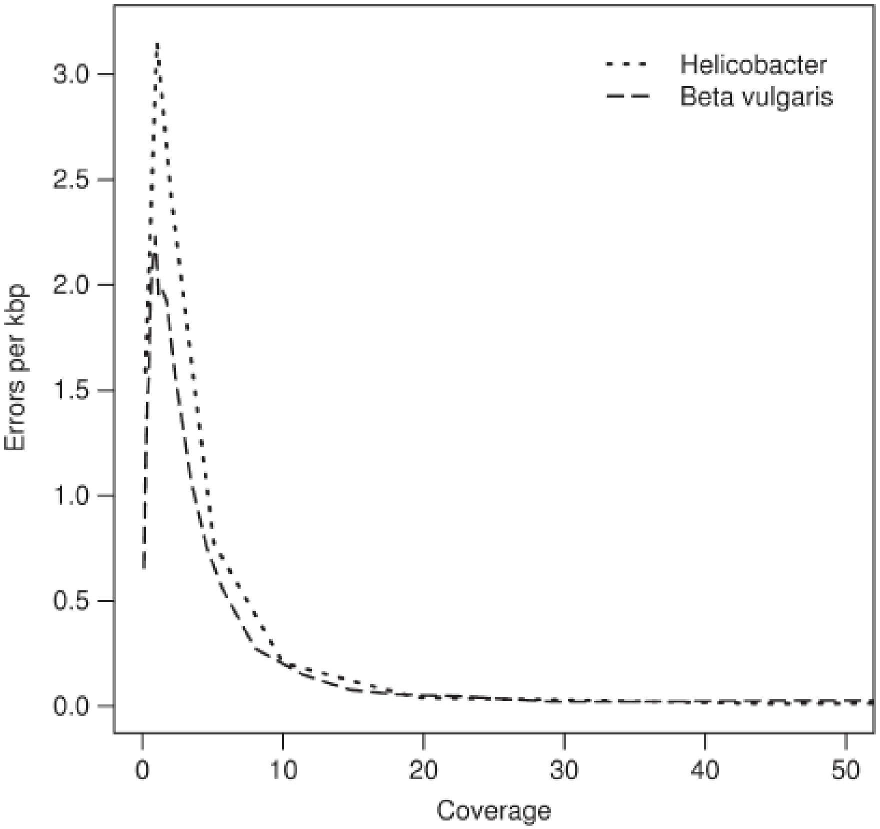
Beta		Helicobacter	
Coverage	BAC (%)	Coverage	Genome (%)
200–300	4.27	<100	3.53
300–400	23.93	100–150	26.06
400–500	25.64	150–200	42.28
500–600	23.93	200–250	21.49
600–700	12.82	250–300	4.44
700–800	4.27	300–350	1.29
800–900	5.13	>350	0.90

- Read distribution along the Beta vulgaris BAC sequence (with cloning vector pBeloBACII). 2 166 892 27mer reads were matched against the finished sequence (enclosed by the cloning vector,117 kbp in total). The read coverage was calculated in 200 consecutive 0.58 kbp windows.
- Read distribution along the 1.55Mbp Helicobacter genome, based on 8 700 113 32mer reads. The local coverage is shown in 200 consecutive windows of 7.77 kbp.

Dohm JC, Lottaz C, Borodina T and Himmelbauer H. “**Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.**” *Nucleic Acids Research* 2008 36:e105.

Solexa sequencing biases: wrong base calls (1)

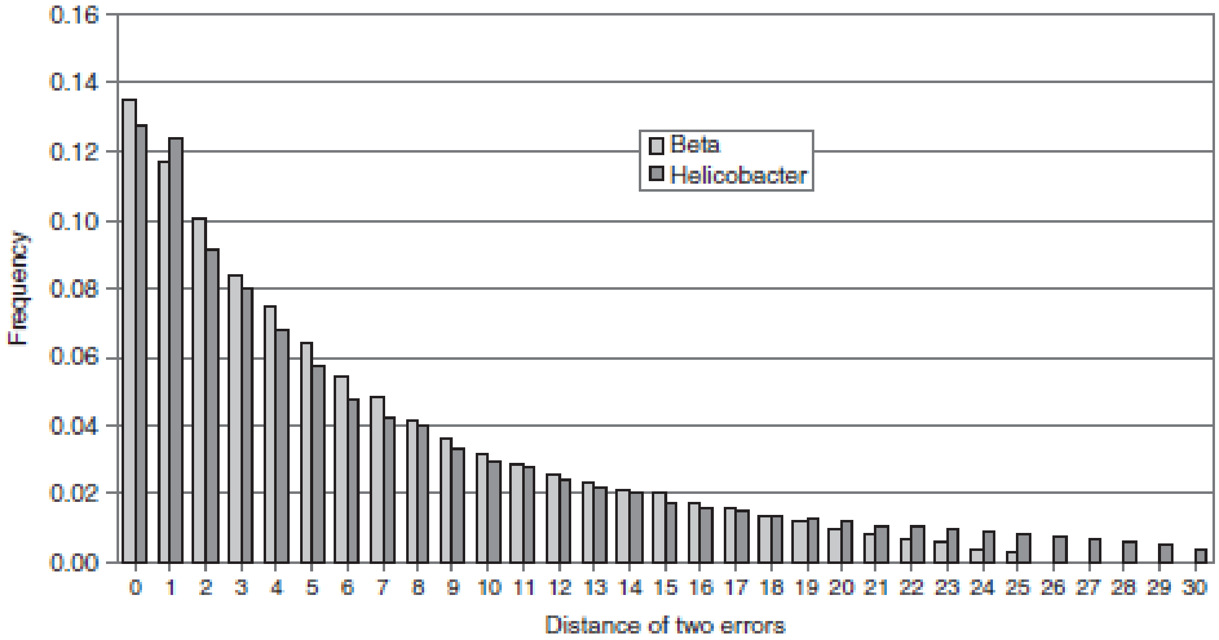




(a) Error frequency per position calculated from considering wrong base calls only. The highest error frequency is observed at the read 30 end. (b) Per-base error rates (overall error frequency per position considering all base calls).

Dohm JC, Lottaz C, Borodina T and Himmelbauer H. "**Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.**" *Nucleic Acids Research* 2008 36:e105.

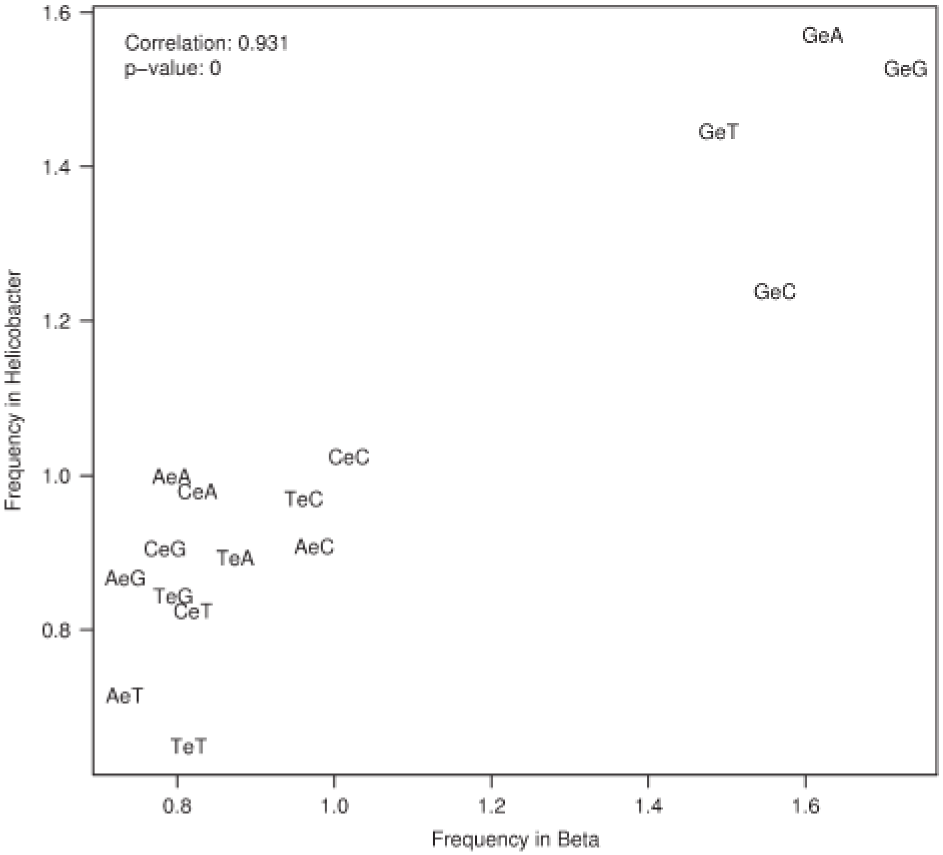
Solexa sequencing biases: wrong base calls (2)

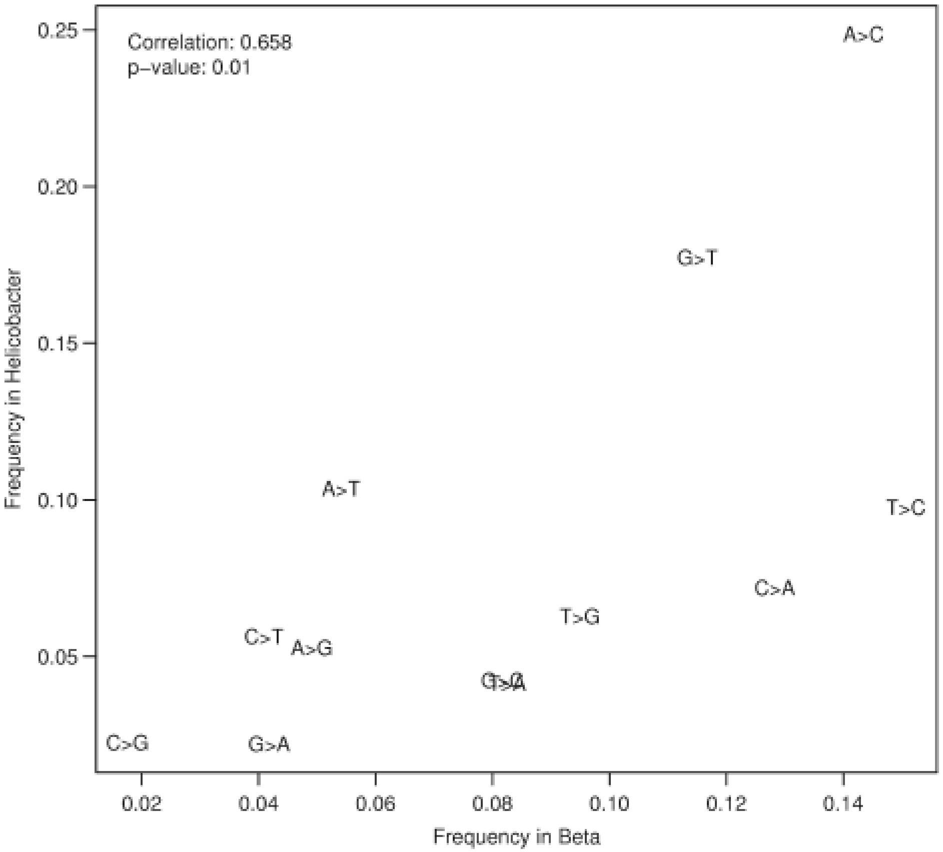


Distance between two errors on a read in the Helicobacter and Beta vulgaris data sets. ‘0’ indicates that the erroneous base calls are next to each other.

Dohm JC, Lottaz C, Borodina T and Himmelbauer H. “**Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.**” *Nucleic Acids Research* 2008 36:e105.

Solexa sequencing biases: error context





Dohm JC, Lottaz C, Borodina T and Himmelbauer H. “**Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.**” *Nucleic Acids Research* 2008 36:e105.

Solexa sequencing biases: score quality (1)

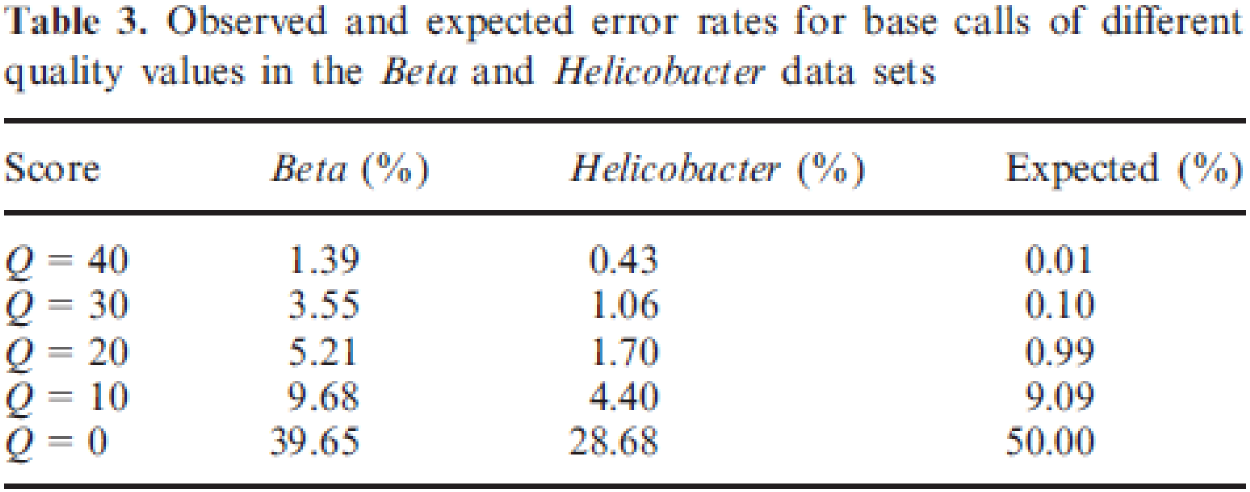


Table 3. Observed and expected error rates for base calls of different quality values in the *Beta* and *Helicobacter* data sets

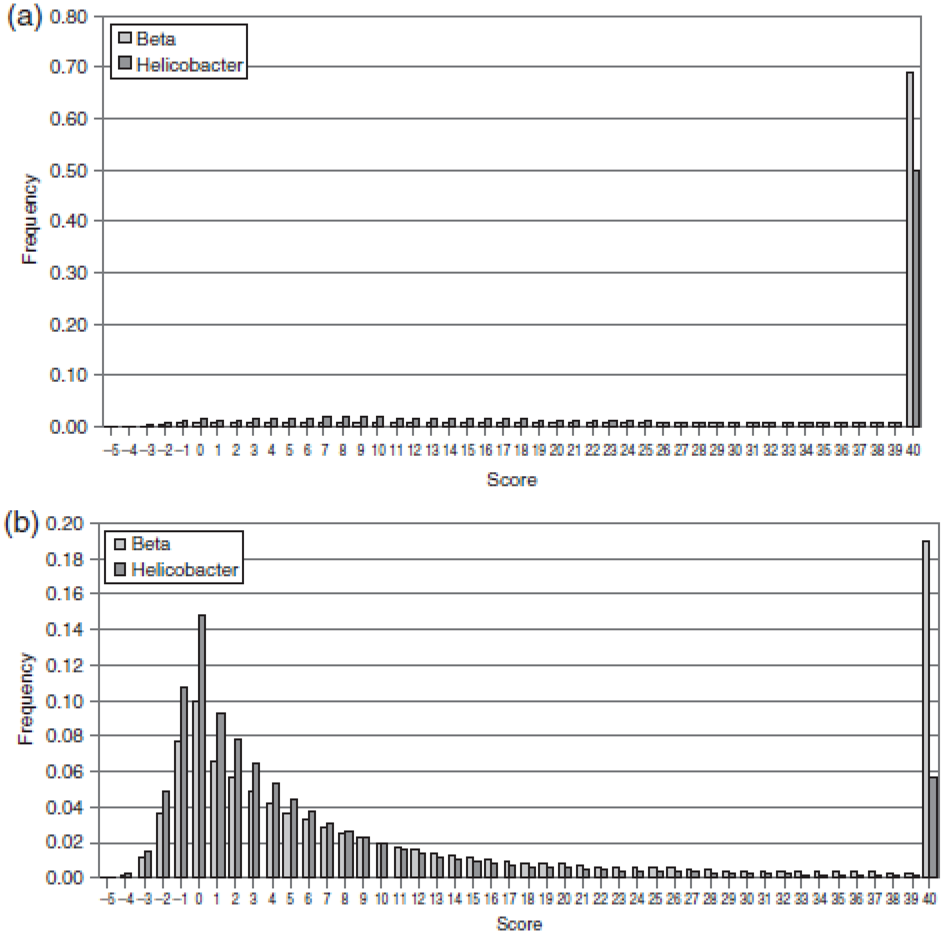
Score	*Beta* (%)	*Helicobacter* (%)	Expected (%)
$Q = 40$	1.39	0.43	0.01
$Q = 30$	3.55	1.06	0.10
$Q = 20$	5.21	1.70	0.99
$Q = 10$	9.68	4.40	9.09
$Q = 0$	39.65	28.68	50.00

The Solexa base caller Bustard reports the quality of each base call by estimating a quality score similar to the phred score based on the image output without considering the reference sequence.

Q=40 -> expected error probability of P=0.01%
Q=0 -> expected error probability of P=50%

Dohm JC, Lottaz C, Borodina T and Himmelbauer H. "**Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.**" *Nucleic Acids Research* 2008 36:e105.

Solexa sequencing biases: score quality (2)



Histograms of base quality values for all correct base calls (a) and all wrong base calls (b) in the Beta and Helicobacter data sets.

Six percent of all wrong base calls in Helicobacter and 19% of all wrong base calls in Beta have Solexa quality scores Q=40.

Dohm JC, Lottaz C, Borodina T and Himmelbauer H. “**Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.**” *Nucleic Acids Research* 2008 36:e105.

Solexa sequencing biases: conclusions

- General sequence bias around read starting positions were not detected
- Strong correlation between GC richness and read coverage
- Base call errors occur preferentially at the 3’ end of the reads
- Sequence tuples before error position are mainly G-rich
- Solexa software underestimates true-error rate up to 100-fold for high quality values and overestimates for low quality values

Eddy J, Maizels N. “**Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes.”** *Nucl. Acids Res*. 2008 36: 1321-1333.