SUPPLEMENTARY DATA

Table S1. Smith-Waterman scoring parameter applied for Illumina-PacBio hybrid alignments

	proovread		SHRiMP2
	default	refined	default
match	5	5	10
mismatch	−11	−10	−15
gap-open on LR	−2	−5	−33
gap-open on SR	−1	−5	−33
gap-extend on LR	−4	−2	−7
gap-extend on SR	−3	−2	−3

SHANNON-ENTROPY

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i) \qquad \text{(S1)}$$

PHRED CONVERSION

$$p = \min(\sqrt{f \cdot 50}, 40); \qquad \text{(S2)}$$

Table S2. Frequency to Phred conversion table

Freq.	Phred	ASCII Phred+64	Freq.	Phred	ASCII Phred+64
0	0	@	20	32	`
1	7	G	21	32	`
2	10	J	22	33	a
3	12	L	23	34	b
4	14	N	24	35	c
5	16	P	25	35	c
6	17	Q	26	36	d
7	19	S	27	37	e
8	20	T	28	37	e
9	21	U	29	38	f
10	22	V	30	39	g
11	23	W	31	39	g
12	24	X	32	40	h
13	25	Y	33	40	h
14	26	Z	34	40	h
15	27	[	35	40	h
16	28	\	36	40	h
17	29	]	37	40	h
18	30	^	38	40	h
19	31	_	39	40	h
20	32	`	40	40	h

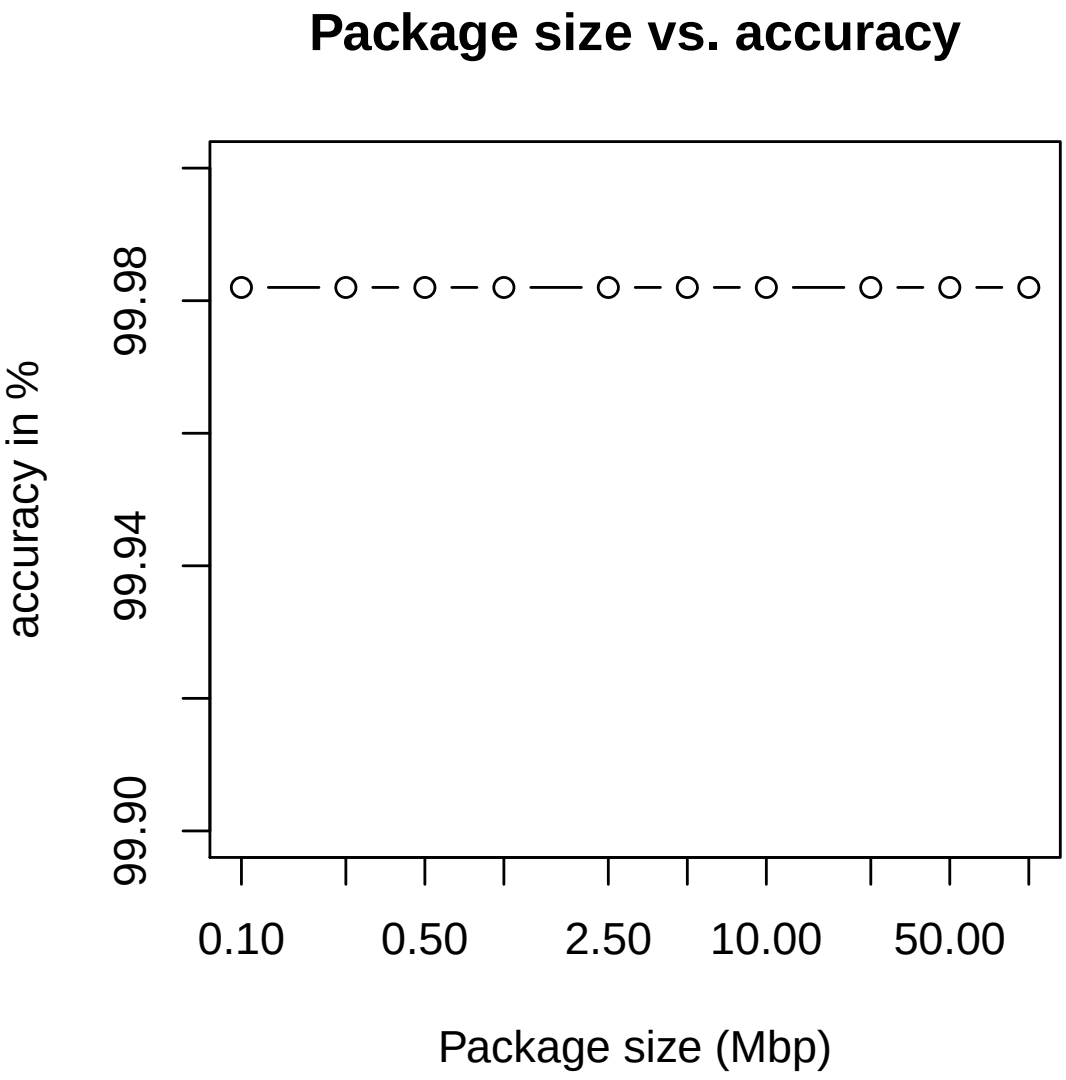
Package size vs. accuracy
accuracy in %
99.90
99.94
99.98
0.10
0.50
2.50
10.00
50.00
Package size (Mbp)

Fig. S1

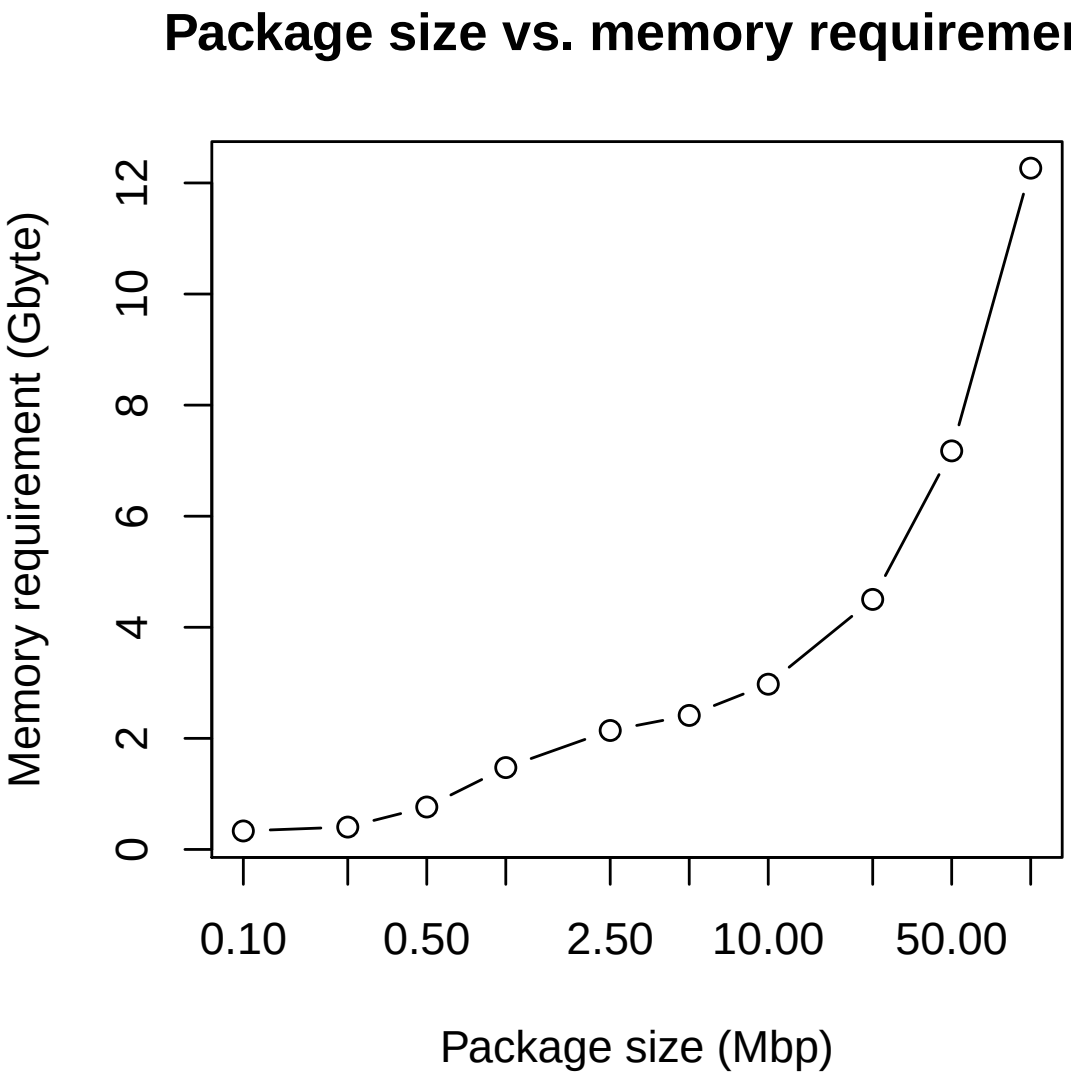
Package size vs. memory requirement
Memory requirement (Gbyte)
0
2
4
6
8
10
12
0.10
0.50
2.50
10.00
50.00
Package size (Mbp)

Fig. S2

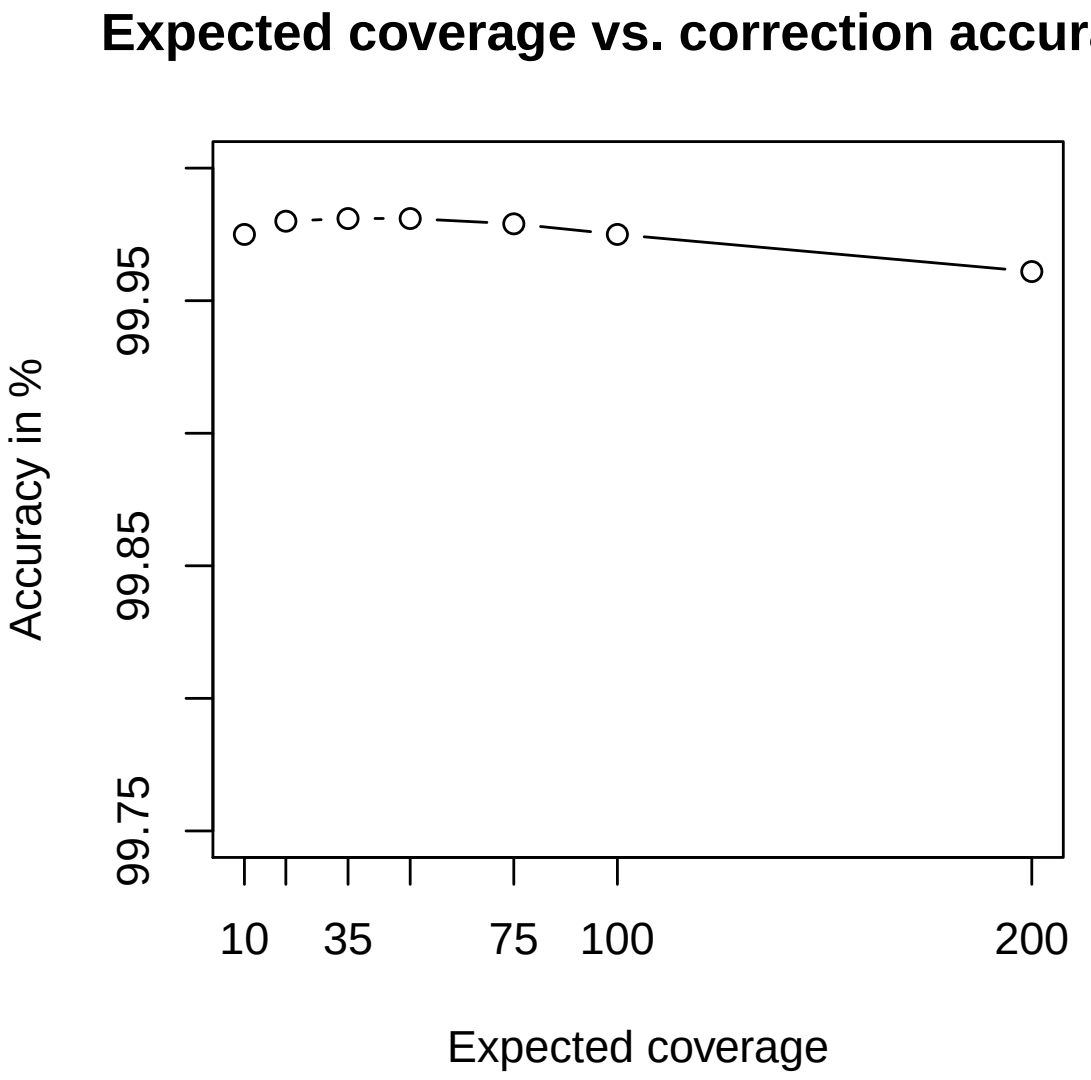
Expected coverage vs. correction accuracy
Accuracy in %
99.75
99.85
99.95
10
35
75
100
200
Expected coverage

Fig. S3

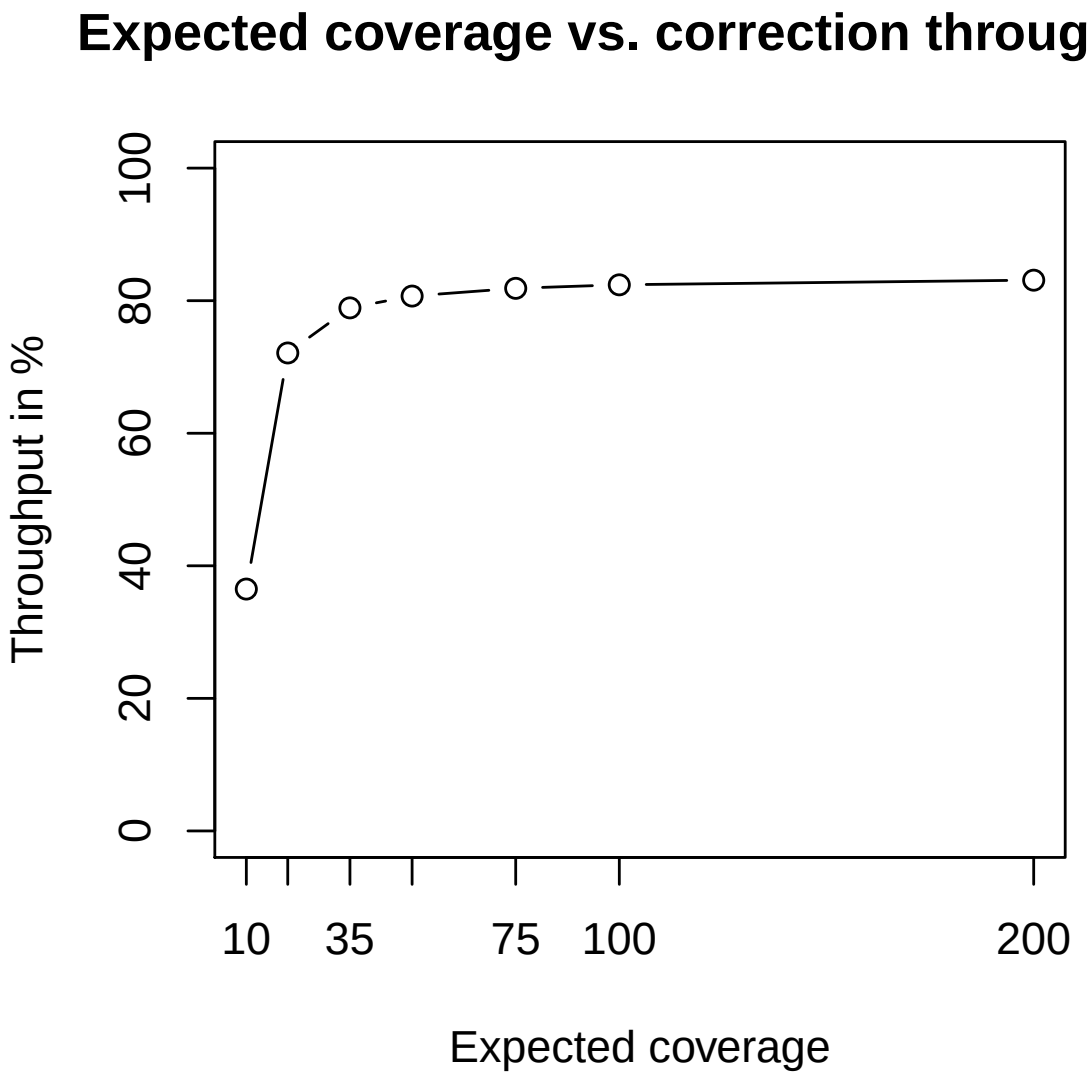
Expected coverage vs. correction throughput
Throughput in %
0
20
40
60
80
100
10
35
75
100
200
Expected coverage

Fig. S4

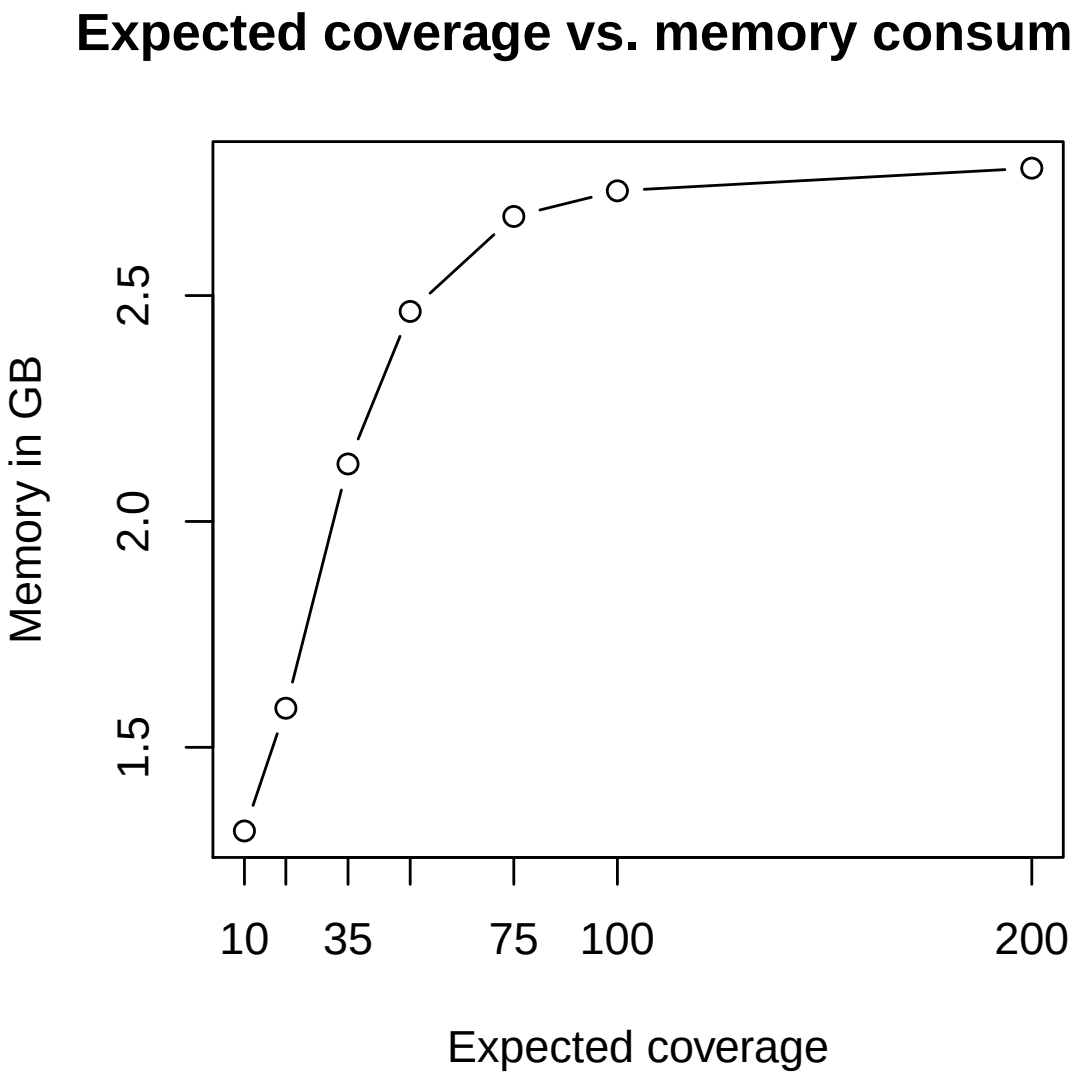
Expected coverage vs. memory consumption
Memory in GB
1.5
2.0
2.5
10
35
75
100
200
Expected coverage

Fig. S5

Table S3. This table comprises the used data sets used for performace estimation of *proovread*.

organism	type	genome reference	genome size	pacbio cell id / accession	SR accession	read length
Escherichia coli	genomic	GCA_000005845.1	4.64 Mbp	c100247042550000001523002504251220	ERX002508	100 bp
Arabidopsis thaliana	genomic	GCA_000001735.1	119.6 Mbp	c100539462550000001823089611241346	SRX158552	110 bp
Homo sapiens	genomic	GCA_000001405.12	3.23 Gbp	c100518541910000001823079209281310	SRX246904	101 bp
					SRX246905	101 bp
					SRX246906	101 bp
					SRX246907	101 bp
					SRX247361	101 bp
					SRX247362	101 bp
Homo sapiens	transcriptomic	GCA_000001405.12	3.23 Gbp	c100202382555500000315044810141103	ERX011200	50 bp
				c100202382555500000315044810141104	ERX011186	75 bp
				c100202382555500000315044810141105		

Table S4. Percentage of ambigous bases obtained from the corrections by *proovread*, PacBioToCA, and LSC

Program	*Escherichia coli*	*Arabidopsis thaliana*	*Homo sapiens*	
			genome	transcriptome
proovread def.	0.54	9.65	3.40	0.27
proovread norm.			1.49	
proovread adapt.				0.31
PacBioToCA	1.37	6.49		0.30
LSC	8.51	16.88		1.65

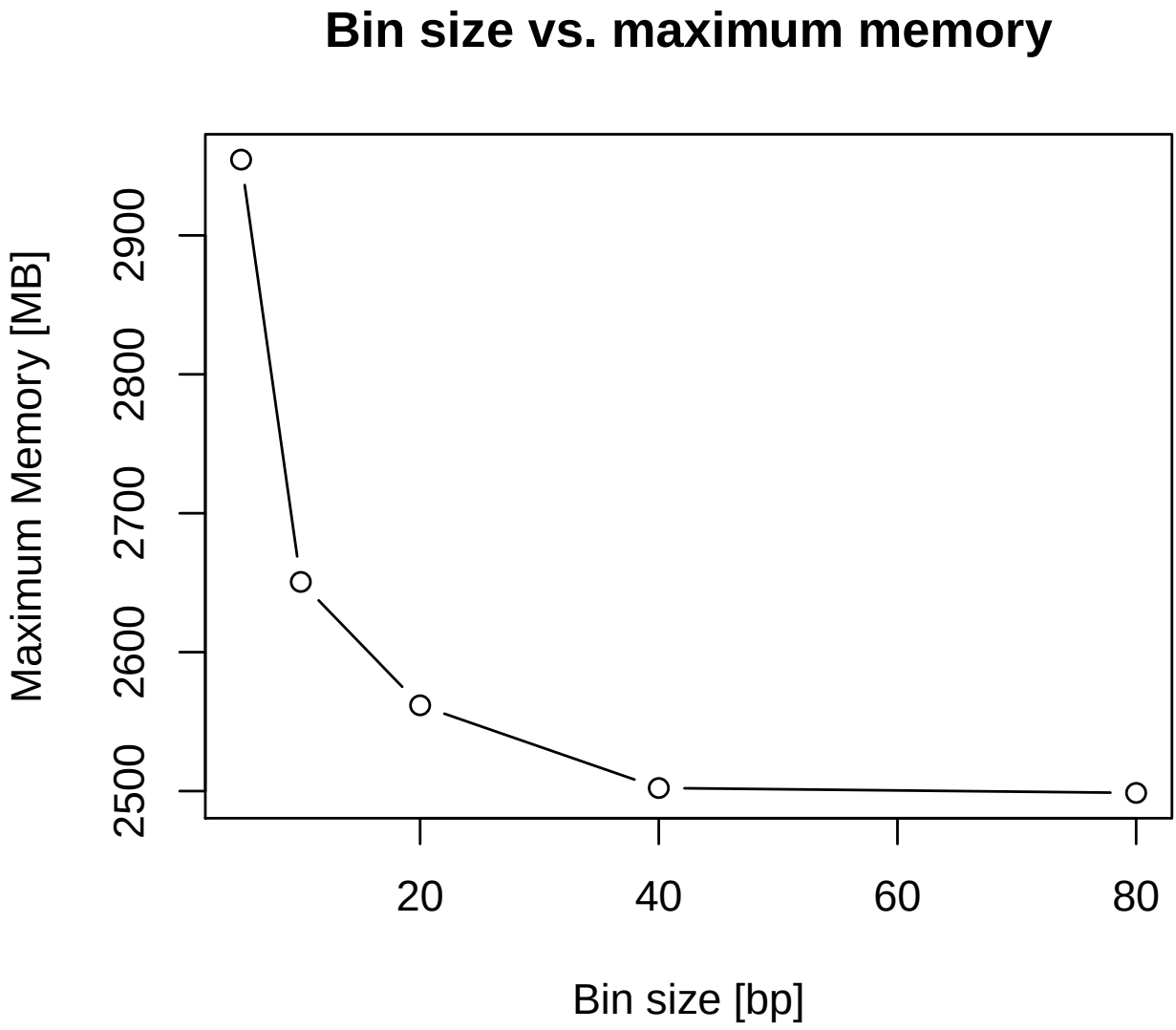

Fig. S6

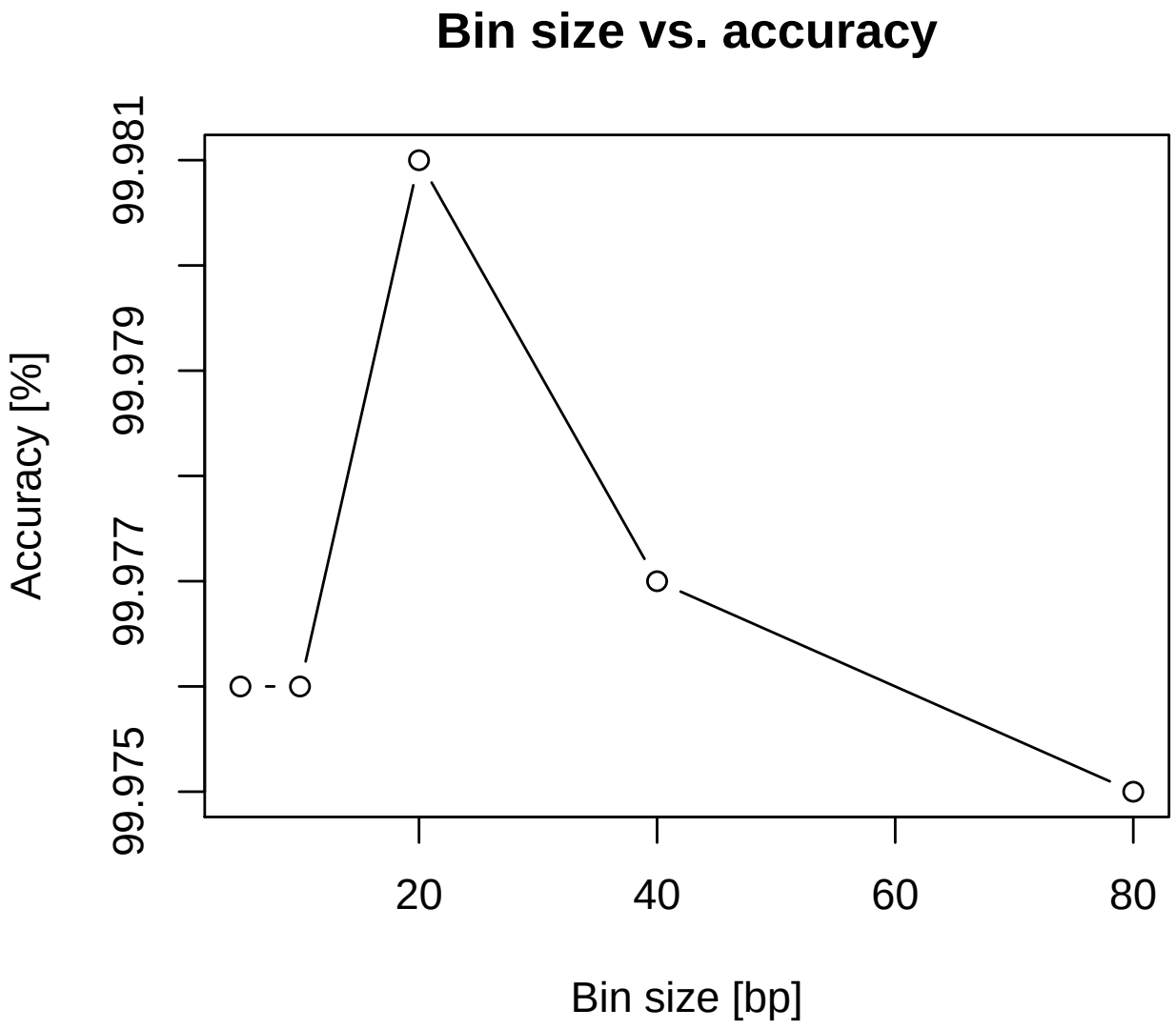
Bin size vs. accuracy
Accuracy [%]
99.975
99.977
99.979
99.981
20
40
60
80
Bin size [bp]

Fig. S7

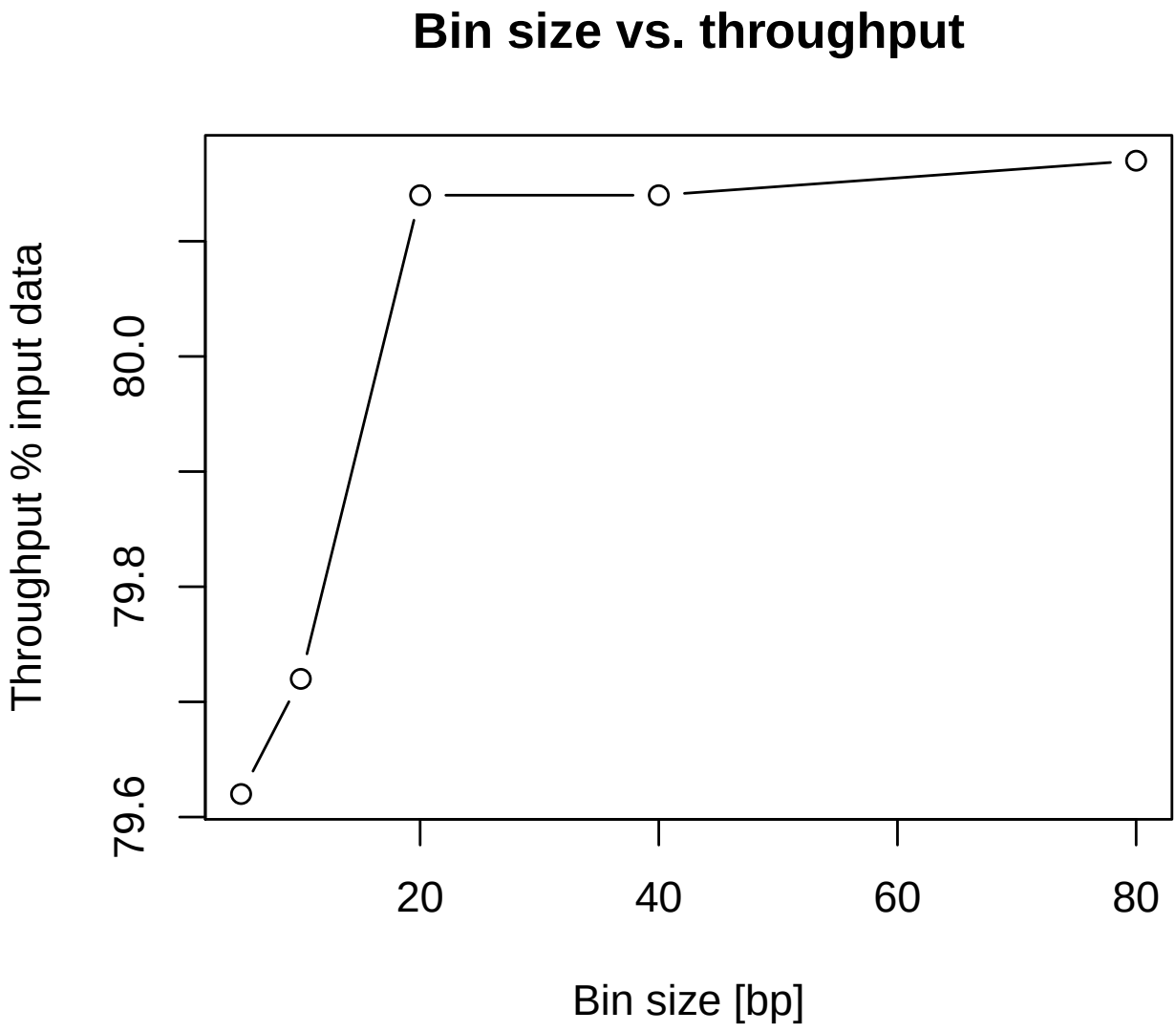
Bin size vs. throughput
Throughput % input data
79.6
79.8
80.0
20
40
60
80
Bin size [bp]

Fig. S8

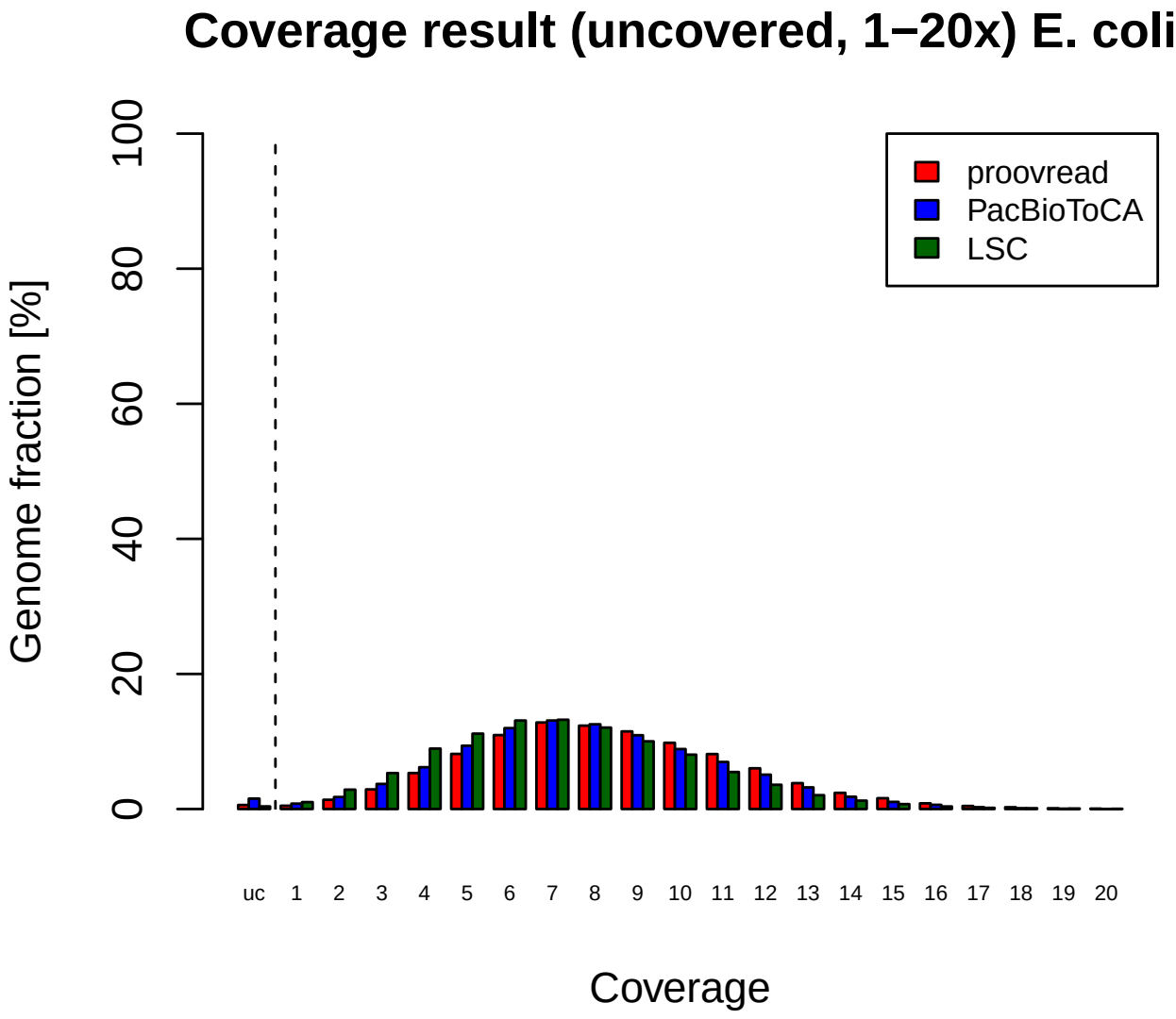
Coverage result (uncovered, 1−20x) E. coli
Genome fraction [%]
0
20
40
60
80
100
proovread
PacBioToCA
LSC
uc
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Coverage

Fig. S9

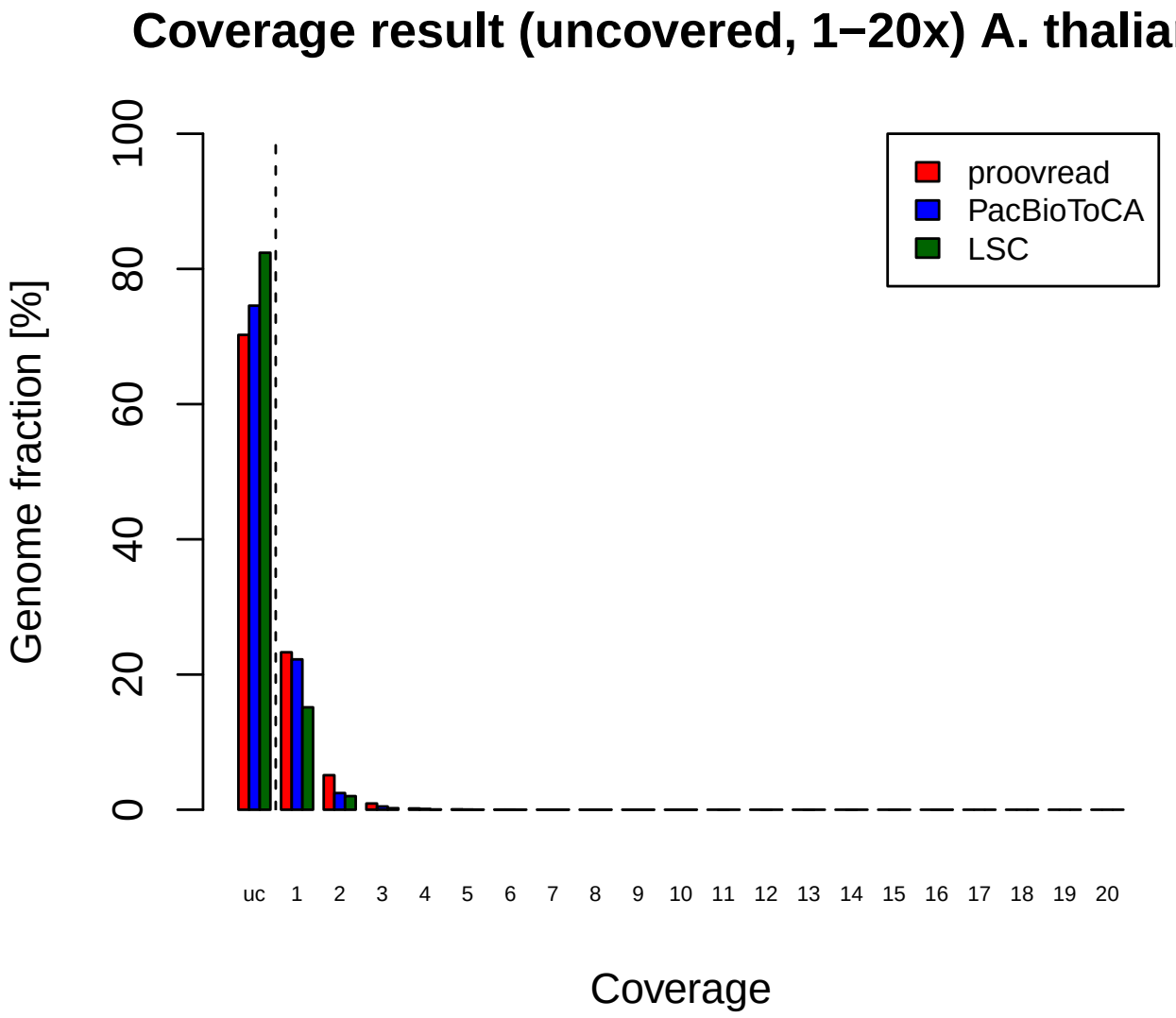
Coverage result (uncovered, 1−20x) A. thaliana
Genome fraction [%]
0
20
40
60
80
100
proovread
PacBioToCA
LSC
uc
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Coverage

Fig. S10

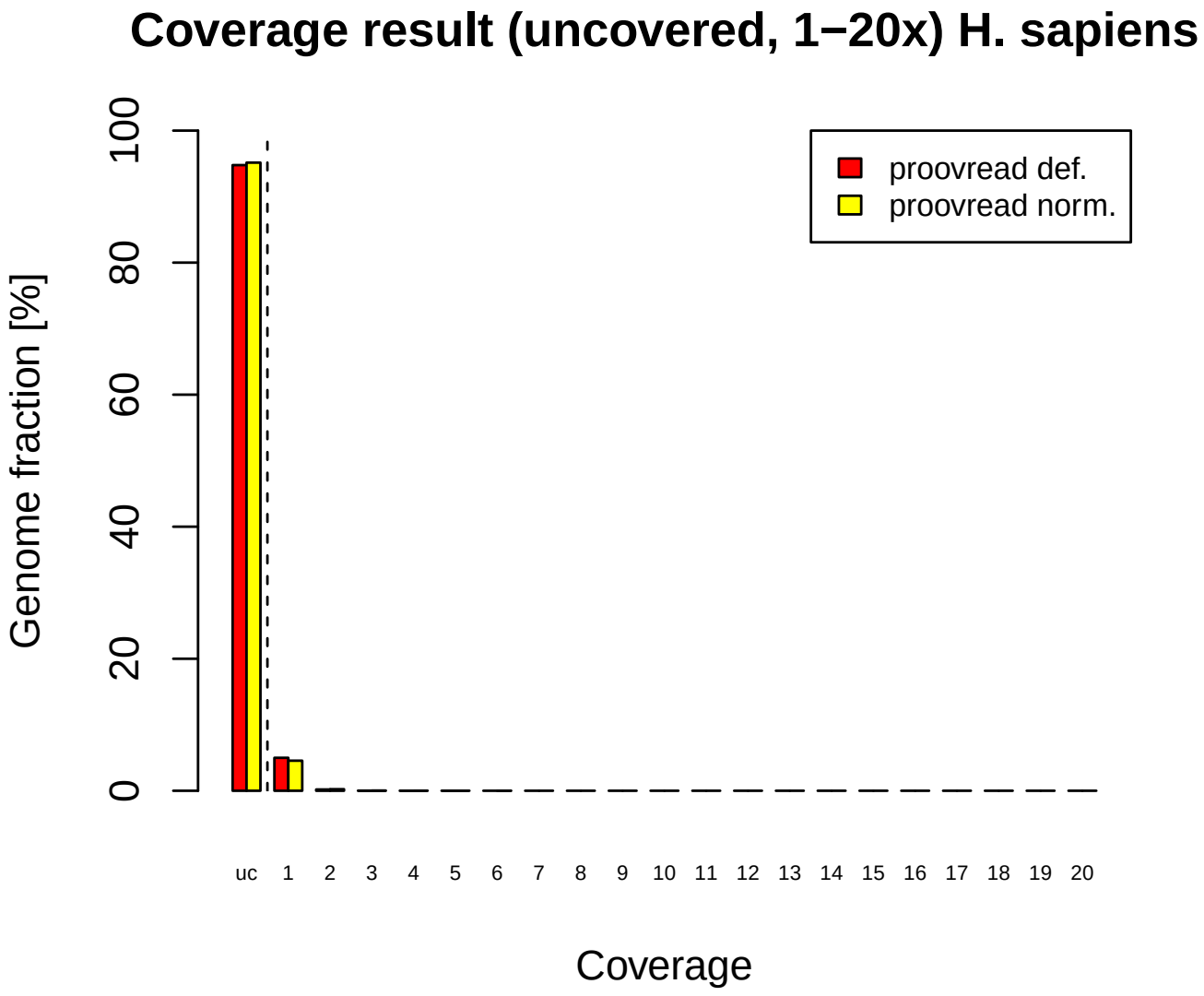
Coverage result (uncovered, 1−20x) H. sapiens
Genome fraction [%]
0
20
40
60
80
100
proovread def.
proovread norm.
uc
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Coverage

Fig. S11

Table S5. Overall horizontal coverage for the correction of genomic data sets of *Escherichia coli*, *Arabidopsis thaliana*, and *Homo sapiens*

Method	*E. coli*	*A. thaliana*	*H. sapiens*
proovread default	99.46 %	29.76 %	5.23 %
proovread norm.			4.85 %
PacBioToCA	98.46 %	25.43 %	
LSC	99.62 %	17.60 %	

proovread default config

```
## proovread Config File
##-- command line parameter ------------------------------------------------##
## LIST of Pacbio read files to correct. FASTA or FASTQ format.
'long-reads' => [],
## LIST of high confidence short read files used for correction in FASTQ or
## FASTA format.
'short-reads' => [],
## Prefix to output files. Defaults to 'proovread'
'prefix' => undef,
## Coverage cutoff for highest scoring mappings at each location.
'coverage' => 40,
## Number of threads to use for mapping. Defaults to 8 (or maximum available
##  number of processors, if kess than 8 availabe).
'threads' => 4,
## to auto-detect max processors use:
##  'threads' => qx(grep '^processor' /proc/cpuinfo | wc -l) =~ s/\n//r,
'mode' =>  undef, #'pacbio-pre', 'pacbio-iterative', 'pacbio-ccs', 'external-sam'
# for custom modes see mode_passes in the "advanced options" section
## Use already created mapping in SAM format to create corrected consensus
##  sequences from. Use this as alternative input to --long-reads/short-reads
##  If --sam is specified, --mode is automatically set to "external-sam".
'sam' => undef,
## By default, while mapping, a temporary SAM file is created and an index to
##  this file is kept in memory. This limits the memory requirement for one SMRT
##  cell (>100Mbp, 50X coverage) to less than 10Gb.
## Specify '--ram-sam' to hold the SAM presentation entirely in memory. This is
##  faster and saves disk space, but might require up to 100GB per SMRT cell.
'ram-sam' => undef,
## Sort the filtered SAM files by coordinates in addition to the
##  sorting of references. This has no effect on the pipeline, and is just
##  a convenience if you need the files for something else.
'sort-sam-by-coordinates' => undef,
## Specify '1' to keep temporary file of each pass, '2' to also keep the
##  individual temporary file of each thread.
'keep-temporary-files' => 0,  # 0,1,2
## overwrite exiting output folder
'overwrite' => 0,
##-- advanced parameter ----------------------------------------------------##
## don't mess with these unless you know what you are doing
## Short read quality offset, usually 64 or 33, use 0 for FASTA. Defaults to
##  guessing, Specify value if guessing fails. Needs to be the same for all
##  short read files provided.
'sr-qv-offset' => undef,
## Short read length. Defaults to guessing, sampling 1000 reads from input
##  file. Specify value if guessing fails.
'sr-length' => undef,
## Number of short reads provided, used for ETA calculation. Defaults to
##  guessing based on 1000 randomly sampled reads. Specify value if guessing
##  fails.
'sr-count' => undef,
## Toggle short reads head/tail trimming including leading/trailing indels
##  sr-indel-taboo-length
'sr-trim' => 1,
## Fraction of short reads used in pre masking steps
'sr-pre-fraction' => {
        'shrimp-pre-1' => 0.2,
        'shrimp-pre-2' => 0.3,
        'shrimp-pre-3' => 0.5,
        'shrimp-pre-4' => 0.3,
},
## expand lcr-min-length (min-sr-length + max-sr-dev) to this ratio.
'lcr-min-ratio' => 1.3,
##
'lcr-end-ratio' => {
        'shrimp-pre-1' => 0.8,
        'shrimp-pre-2' => 0.3,
        'shrimp-pre-3' => 0.5,
        'shrimp-pre-4' => 0.5,
},
## Trim reads to prevent insertions/deletions within the first
## 'sr-indel-taboo-length' fraction of the read. N=0 deactivates the feature.
'sr-indel-taboo' => 0.1,
## Long read qv-offset, required if --sam and --long are used together, and
##  it cannot be detected automatically from --long file.
'lr-qv-offset' => undef,
## Detect and identify chimera like looking reads
'detect-chimera' => 1,
## Number of bases at the end of a hcr from the previous iteration not be
##  masked w/r/t sr-length
'hcr-sticky-ratio' => .3,
## Minimum length of a hcr region to be considered as such w/r/t sr-length,
##  0 deactivates hcr detection.
'hcr-min-ratio' => .8,
## Number of reads to check out at once for individual consensus correction
##  process. Memory intensive step, be cautios with great values
'chunk-size' => 100,
## Size in base pairs of bins for local score comparisons
'bin-size' => 20,
##-- task settings ---------------------------------------------------------##
'mode-tasks' => {
        # illumina, evenly covered, e.g. genome
        'pacbio-pre' => ['read-long', 'shrimp-pre-1', 'shrimp-pre-2',
                         'shrimp-pre-3', 'shrimp-finish'],
        # use an externally created SAM file
        'external-sam' => ['read-long', 'read-sam'],
        # illumina, unevenly covered, e.g. quant. RNA-seq, slower than pacbio-pre
        'pacbio-iterative' => ['read-long', 'shrimp-iter-1', 'shrimp-iter-2',
                               'shrimp-finish'],
        # Experimental: illumina, bowtie2 + shrimp
        'bowtie2' => ['read-long', 'bowtie2-iter-1', 'shrimp-pre-2',
                      'bowtie2-finish'],
        # Experimental: Circualar Consensus, 454...
        'pacbio-ccs' => ['read-long', 'shrimp-ccs', 'shrimp-finish'],
        # custom => ['my-pass-settings, finish],  #...
},
## shrimp-pre-1
'shrimp-pre-1' => {
        '-h' => "55%",
        '--report' => 200,
        '-s' => "1"x11,
        '-w' => "130%",
        '--no-mapping-qualities' => '',
        '--match' => 5,
        '--mismatch' => -11,
        '--open-r' => -2,
        '--open-q' => -1,
        '--ext-r' => -4,
        '--ext-q' => -3,
},
## shrimp-pre-2
'shrimp-pre-2' => {
        '-h' => "55%",
        '--report' => 200,
        '-s' => "1"x10,
        '-w' => "140%",
        '-r' => "45%",
        '--no-mapping-qualities' => '',
        '--match' => 5,
        '--mismatch' => -11,
        '--open-r' => -2,
        '--open-q' => -1,
        '--ext-r' => -4,
        '--ext-q' => -3,
},
## shrimp-pre-3
'shrimp-pre-3' => {
        '-h' => "50%",
        '--report' => 200,
        '-s' => "11111111,1111110000111111",
        '-w' => "140%",
        '-r' => "35%",
        '--no-mapping-qualities' => '',
        '--match' => 5,
        '--mismatch' => -11,
        '--open-r' => -2,
        '--open-q' => -1,
        '--ext-r' => -4,
        '--ext-q' => -3,
},
## shrimp finish
'shrimp-finish' => {
        '-h' => "90%",
        '--report' => 200,
        '-s' => "1"x20,
        '--hash-spaced-kmers' => '',
        '--match' => 5,
        '--mismatch' => -10,
        '--open-r' => -5,
        '--open-q' => -5,
        '--ext-r' => -2,
        '--ext-q' => -2,
},
##-- Chimera filter --------------------------------------------------------##
'chimera-filter' => {
        '--min-score' => 0.01,
        '--trim-length' => 20,
        '--verbose' => 2
},
##-- SeqFilter settings ----------------------------------------------------##
'seq-filter' => {
        '--trim-win' =>  15,
        '--trim-lcs' =>  '3,50,100',
        '--min-length' => 100,
},
##-- SeqChunker settings ---------------------------------------------------##
'seq-chunker' => {
        '--chunk-number' => 100,
},
```

proovread transcriptome config

```
## proovread Config File
##-- command line parameter ----------------------------------------------##
## LIST of Pacbio read files to correct. FASTA or FASTQ format.
'long-reads' => [],
## LIST of high confidence short read files used for correction in FASTQ or
## FASTA format.
'short-reads' => [],
## Prefix to output files. Defaults to 'proovread'
'prefix' => undef,
## Coverage cutoff for highest scoring mappings at each location.
'coverage' => 40,
## Number of threads to use for mapping. Defaults to 8 (or maximum available
##  number of processors, if kess than 8 availabe).
'threads' => 4,
## to auto-detect max processors use:
##  'threads' => qx(grep '^processor' /proc/cpuinfo | wc -l) =~ s/\n//r,
'mode' =>  undef, #'pacbio-pre', 'pacbio-iterative', 'pacbio-ccs', 'external-sam'
# for custom modes see mode_passes in the "advanced options" section
## Use already created mapping in SAM format to create corrected consensus
##  sequences from. Use this as alternative input to --long-reads/short-reads
##  If --sam is specified, --mode is automatically set to "external-sam".
'sam' => undef,
## By default, while mapping, a temporary SAM file is created and an index to
##  this file is kept in memory. This limits the memory requirement for one SMRT
##  cell (>100Mbp, 50X coverage) to less than 10Gb.
## Specify '--ram-sam' to hold the SAM presentation entirely in memory. This is
##  faster and saves disk space, but might require up to 100GB per SMRT cell.
'ram-sam' => undef,
## Sort the filtered SAM files by coordinates in addition to the
##  sorting of references. This has no effect on the pipeline, and is just
##  a convenience if you need the files for something else.
'sort-sam-by-coordinates' => undef,
## Specify '1' to keep temporary file of each pass, '2' to also keep the
##  individual temporary file of each thread.
'keep-temporary-files' => 0,  # 0,1,2
## overwrite exiting output folder
'overwrite' => 0,
##-- advanced parameter --------------------------------------------------##
## don't mess with these unless you know what you are doing
## Short read quality offset, usually 64 or 33, use 0 for FASTA. Defaults to
##  guessing, Specify value if guessing fails. Needs to be the same for all
##  short read files provided.
'sr-qv-offset' => undef,
## Short read length. Defaults to guessing, sampling 1000 reads from input
##  file. Specify value if guessing fails.
'sr-length' => undef,
## Number of short reads provided, used for ETA calculation. Defaults to
##  guessing based on 1000 randomly sampled reads. Specify value if guessing
##  fails.
'sr-count' => undef,
## Toggle short reads head/tail trimming including leading/trailing indels
##  sr-indel-taboo-length
'sr-trim' => 1,
## Fraction of short reads used in pre masking steps
'sr-pre-fraction' => {
        'shrimp-pre-1' => 0.5,
        'shrimp-pre-2' => 1.0,
        'shrimp-pre-3' => 1.0,
        'shrimp-pre-4' => 1.0,
},
## expand lcr-min-length (min-sr-length + max-sr-dev) to this ratio.
'lcr-min-ratio' => 1.3,
##
'lcr-end-ratio' => {
        'shrimp-pre-1' => 0.8,
        'shrimp-pre-2' => 0.3,
        'shrimp-pre-3' => 0.5,
        'shrimp-pre-4' => 0.5,
},
## Trim reads to prevent insertions/deletions within the first
## 'sr-indel-taboo-length' fraction of the read. N=0 deactivates the feature.
'sr-indel-taboo' => 0.1,
## Long read qv-offset, required if --sam and --long are used together, and
##  it cannot be detected automatically from --long file.
'lr-qv-offset' => undef,
## Detect and identify chimera like looking reads
'detect-chimera' => 1,
## Number of bases at the end of a hcr from the previous iteration not be
##  masked w/r/t sr-length
'hcr-sticky-ratio' => .3,
## Minimum length of a hcr region to be considered as such w/r/t sr-length,
##  0 deactivates hcr detection.
'hcr-min-ratio' => .8,
## Number of reads to check out at once for individual consensus correction
##  process. Memory intensive step, be cautios with great values
'chunk-size' => 100,
## Size in base pairs of bins for local score comparisons
'bin-size' => 20,
##-- task settings -------------------------------------------------------##
'mode-tasks' => {

        # illumina, evenly covered, e.g. genome
        'pacbio-pre' => ['read-long', 'shrimp-pre-1', 'shrimp-pre-2',
                         'shrimp-pre-3', 'shrimp-finish'],
        # use an externally created SAM file
        'external-sam' => ['read-long', 'read-sam'],
        # illumina, unevenly covered, e.g. quant. RNA-seq, slower than pacbio-pre
        'pacbio-iterative' => ['read-long', 'shrimp-iter-1', 'shrimp-iter-2',
                               'shrimp-finish'],
        # Experimental: illumina, bowtie2 + shrimp
        'bowtie2' => ['read-long', 'bowtie2-iter-1', 'shrimp-pre-2',
                      'bowtie2-finish'],
        # Experimental: Circualar Consensus, 454...
        'pacbio-ccs' => ['read-long', 'shrimp-ccs', 'shrimp-finish'],
        # custom => ['my-pass-settings, finish],  #...
},
## shrimp-pre-1
'shrimp-pre-1' => {
        '-h' => "55%",
        '--report' => 200,
        '-s' => "1"x11,
        '-w' => "130%",
        '--no-mapping-qualities' => '',
        '--match' => 5,
        '--mismatch' => -11,
        '--open-r' => -2,
        '--open-q' => -1,
        '--ext-r' => -4,
        '--ext-q' => -3,
},
## shrimp-pre-2
'shrimp-pre-2' => {
        '-h' => "55%",
        '--report' => 200,
        '-s' => "1"x10,
        '-w' => "140%",
        '-r' => "45%",
        '--no-mapping-qualities' => '',
        '--match' => 5,
        '--mismatch' => -11,
        '--open-r' => -2,
        '--open-q' => -1,
        '--ext-r' => -4,
        '--ext-q' => -3,
},
## shrimp-pre-3
'shrimp-pre-3' => {
        '-h' => "50%",
        '--report' => 200,
        '-s' => "11111111,1111110000111111",
        '-w' => "140%",
        '-r' => "35%",
        '--no-mapping-qualities' => '',
        '--match' => 5,
        '--mismatch' => -11,
        '--open-r' => -2,
        '--open-q' => -1,
        '--ext-r' => -4,
        '--ext-q' => -3,
},
## shrimp finish
'shrimp-finish' => {
        '-h' => "90%",
        '--report' => 200,
        '-s' => "1"x20,
        '--hash-spaced-kmers' => '',
        '--match' => 5,
        '--mismatch' => -10,
        '--open-r' => -5,
        '--open-q' => -5,
        '--ext-r' => -2,
        '--ext-q' => -2,
},
##-- Chimera filter ------------------------------------------------------##
'chimera-filter' => {
        '--min-score' => 0.01,
        '--trim-length' => 20,
        '--verbose' => 2
},
##-- SeqFilter settings --------------------------------------------------##
'seq-filter' => {
        '--trim-win' =>  20,
        '--trim-lcs' =>  '3,50,100',
        '--min-length' => 100,
},
##-- SeqChunker settings -------------------------------------------------##
'seq-chunker' => {
        '--chunk-number' => 100,
},
```