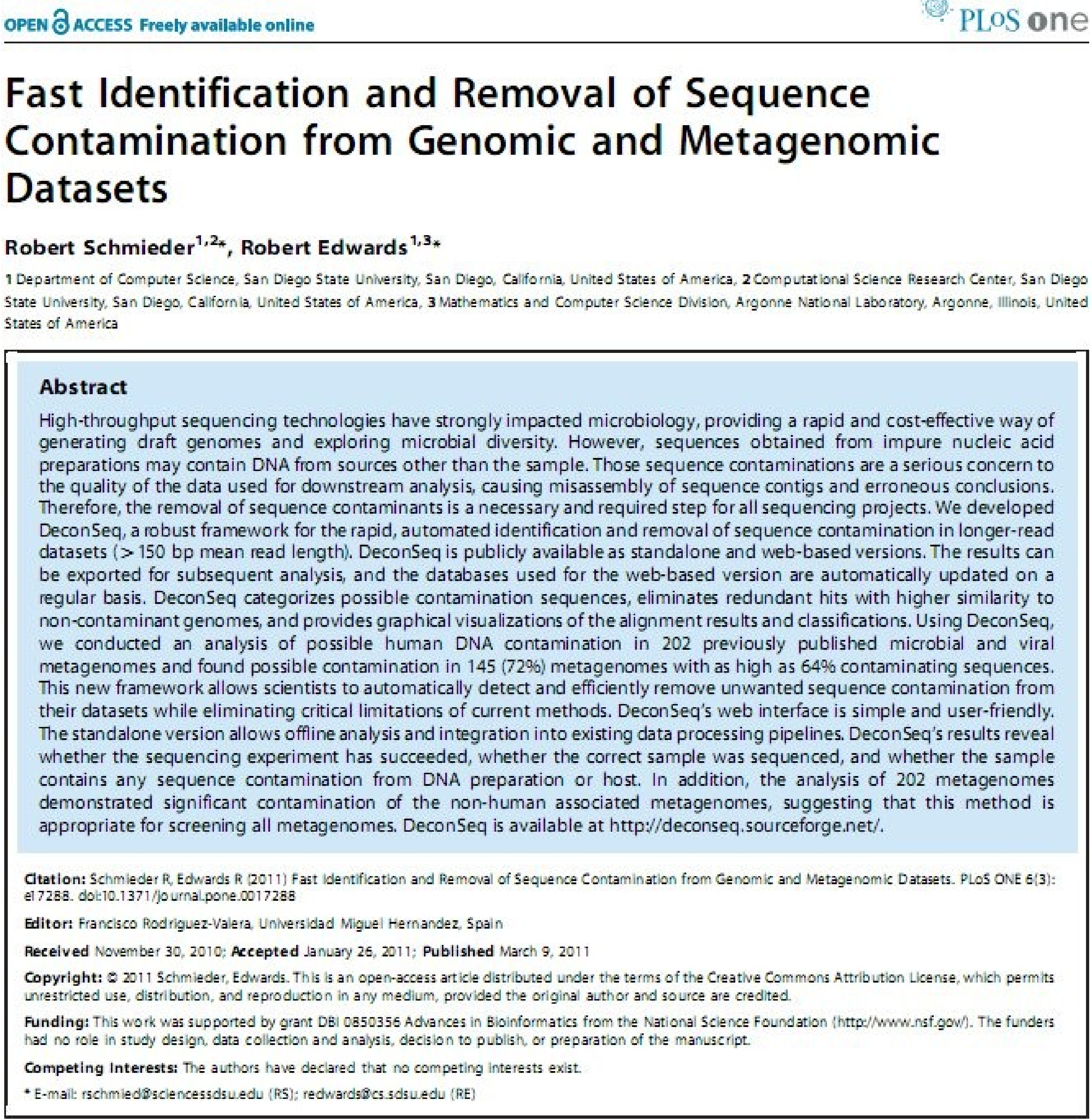
# Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets

*jouranl club* 23.05.11
Aleksander Sudakov



OPEN ACCESS Freely available online

PLoS one

# Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets

**Robert Schmieder**[1,2*], **Robert Edwards**[1,3*]

**1** Department of Computer Science, San Diego State University, San Diego, California, United States of America, **2** Computational Science Research Center, San Diego State University, San Diego, California, United States of America, **3** Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois, United States of America

## Abstract

High-throughput sequencing technologies have strongly impacted microbiology, providing a rapid and cost-effective way of generating draft genomes and exploring microbial diversity. However, sequences obtained from impure nucleic acid preparations may contain DNA from sources other than the sample. Those sequence contaminations are a serious concern to the quality of the data used for downstream analysis, causing misassembly of sequence contigs and erroneous conclusions. Therefore, the removal of sequence contaminants is a necessary and required step for all sequencing projects. We developed DeconSeq, a robust framework for the rapid, automated identification and removal of sequence contamination in longer-read datasets (>150 bp mean read length). DeconSeq is publicly available as standalone and web-based versions. The results can be exported for subsequent analysis, and the databases used for the web-based version are automatically updated on a regular basis. DeconSeq categorizes possible contamination sequences, eliminates redundant hits with higher similarity to non-contaminant genomes, and provides graphical visualizations of the alignment results and classifications. Using DeconSeq, we conducted an analysis of possible human DNA contamination in 202 previously published microbial and viral metagenomes and found possible contamination in 145 (72%) metagenomes with as high as 64% contaminating sequences. This new framework allows scientists to automatically detect and efficiently remove unwanted sequence contamination from their datasets while eliminating critical limitations of current methods. DeconSeq's web interface is simple and user-friendly. The standalone version allows offline analysis and integration into existing data processing pipelines. DeconSeq's results reveal whether the sequencing experiment has succeeded, whether the correct sample was sequenced, and whether the sample contains any sequence contamination from DNA preparation or host. In addition, the analysis of 202 metagenomes demonstrated significant contamination of the non-human associated metagenomes, suggesting that this method is appropriate for screening all metagenomes. DeconSeq is available at http://deconseq.sourceforge.net/.

**Citation:** Schmieder R, Edwards R (2011) Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets. PLoS ONE 6(3): e17288. doi:10.1371/journal.pone.0017288

**Editor:** Francisco Rodriguez-Valera, Universidad Miguel Hernandez, Spain

**Received** November 30, 2010; **Accepted** January 26, 2011; **Published** March 9, 2011

**Funding:** This work was supported by grant DBI 0850356 Advances in Bioinformatics from the National Science Foundation (http://www.nsf.gov/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

* E-mail: rschmied@sciencessdsu.edu (RS); redwards@cs.sdsu.edu (RE)

# Motivation

- **high-throughput sequencing** allows rapid and cost-effective way of generating draft genomes and exploring microbial diversity
- sequences may be obtained from impure samples e.g. **metagenomes**
- contamination affects quality of data for downstream analysis, misassembly of contigs, erroneous conclusions etc
- sequence **cleaning up** required before further processing:
  read duplicates
  low quality reads
  contaminating sequences
  adaptor or barcode sequences

# Problems

- high-throughput sequencing produces longer reads, up to 400 bp with 1000 bp in sight
- short read aligners maximize global alignment
- long read aligners must deal with gaps (indels most frequent sequencing error)
- speed and memory becomes bottleneck for amount of data
- repeats cause problems, but masking them not best option with long reads
- dark matter – regions that do not exist in reference genomes (insertions or structural variants)

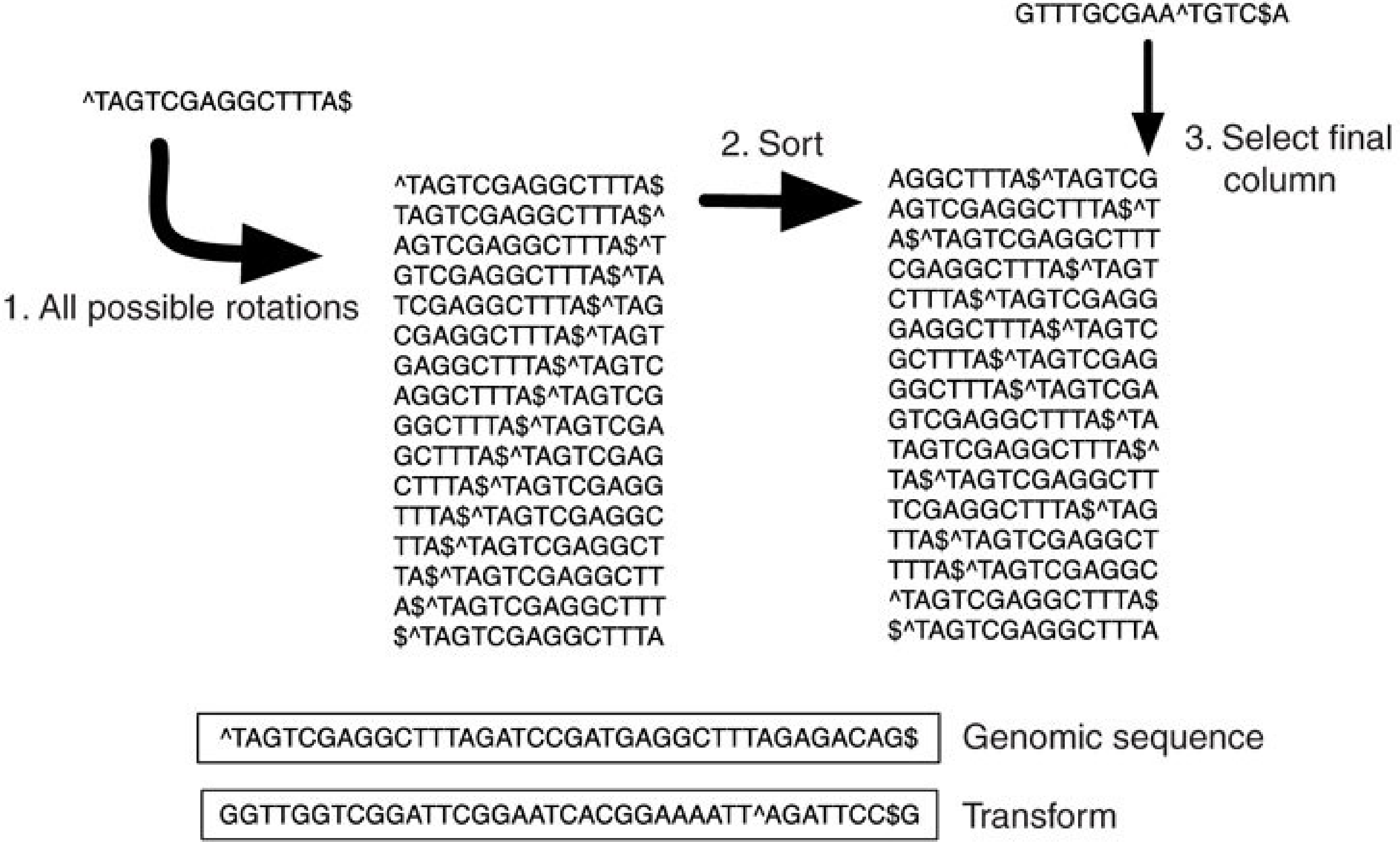
# Available software overview

Three approaches to long read alignment programs:

- **hash table** (BLAST, SSAHA2, BLAT, SOAP, ELAND, MOSAIK)
  target hash database - memory
  query length - time
- **suffix/prefix tree** (MUMmer)
  target suffix tree – memory (10 bytes per nucleotide)
  linear time search
- **Burrows-Wheeler Transform** (BOWTIE, SOAP2, BWA)

  BWA-SW (BWT alligner + Smith-Waterman search heuristics)

  Ferragina-Manzini compressed FM-index: suffix array is much more efficient if it is created from the BWT sequence, rather than from the original sequence (0,5-2 bytes per nucleotide)



^TAGTCGAGGCTTTA$

1. All possible rotations

^TAGTCGAGGCTTTA$
TAGTCGAGGCTTTA$^
AGTCGAGGCTTTA$^T
GTCGAGGCTTTA$^TA
TCGAGGCTTTA$^TAG
CGAGGCTTTA$^TAGT
GAGGCTTTA$^TAGTC
AGGCTTTA$^TAGTCG
GGCTTTA$^TAGTCGA
GCTTTA$^TAGTCGAG
CTTTA$^TAGTCGAGG
TTTA$^TAGTCGAGGC
TTA$^TAGTCGAGGCT
TA$^TAGTCGAGGCTT
A$^TAGTCGAGGCTTT
$^TAGTCGAGGCTTTA

2. Sort

AGGCTTTA$^TAGTCG
AGTCGAGGCTTTA$^T
A$^TAGTCGAGGCTTT
CGAGGCTTTA$^TAGT
CTTTA$^TAGTCGAGG
GAGGCTTTA$^TAGTC
GCTTTA$^TAGTCGAG
GGCTTTA$^TAGTCGA
GTCGAGGCTTTA$^TA
TAGTCGAGGCTTTA$^
TA$^TAGTCGAGGCTT
TCGAGGCTTTA$^TAG
TTA$^TAGTCGAGGCT
TTTA$^TAGTCGAGGC
^TAGTCGAGGCTTTA$
$^TAGTCGAGGCTTTA

3. Select final column

GTTTGCGAA^TGTC$A

^TAGTCGAGGCTTTAGATCCGATGAGGCTTTAGAGACAG$ Genomic sequence

GGTTGGTCGGATTCGGAATCACGGAAAATT^AGATTCC$G Transform

BWT of a 14-mer genomic sequence. Construct all rotations of the given sequence by taking the first character of the sequence and placing it at the end of the sequence (step 1).The characters ^ and $ mark the beginning and end of the sequence, respectively. Once these sequences are created, they are sorted (step 2). From this sorted matrix, the final column is selected as the transformed sequence (step 3). The transformed sequences is exactly the same length and has exactly the same characters as the original sequence, but in a different ordering. The sequence at the bottom is a longer sequence starting with the same 14-mer that demonstrates the effect on the transformed sequence of using a longer input sequence.

# Program performance test

- 16 GB memory
- 50GB HDD space
- time limit 24 hours
- simulated metagenome dataset 1M reads
- 100K human “contamination” sequences from human reference genome build 37 + filtered sequences from Watson, Asian and Yoruban genomes

# Test of software

- Mosaic (memory needed 57GB)
- NUCMer (109 hours running time)
- BLAST (tabulated output generates Gbs of data ,reports all alignments, not done in 24h)
- MegaBLAST (segfault after 4-5 h with human dataset)
- BLAST+ (too much memory on unmasked human dataset)
- BWA-SW (22 minutes, 3,4 GB memory)
- (did not test BLAT and SSAHA2, these programs were compared to BWA-SW in previous work and were slower)

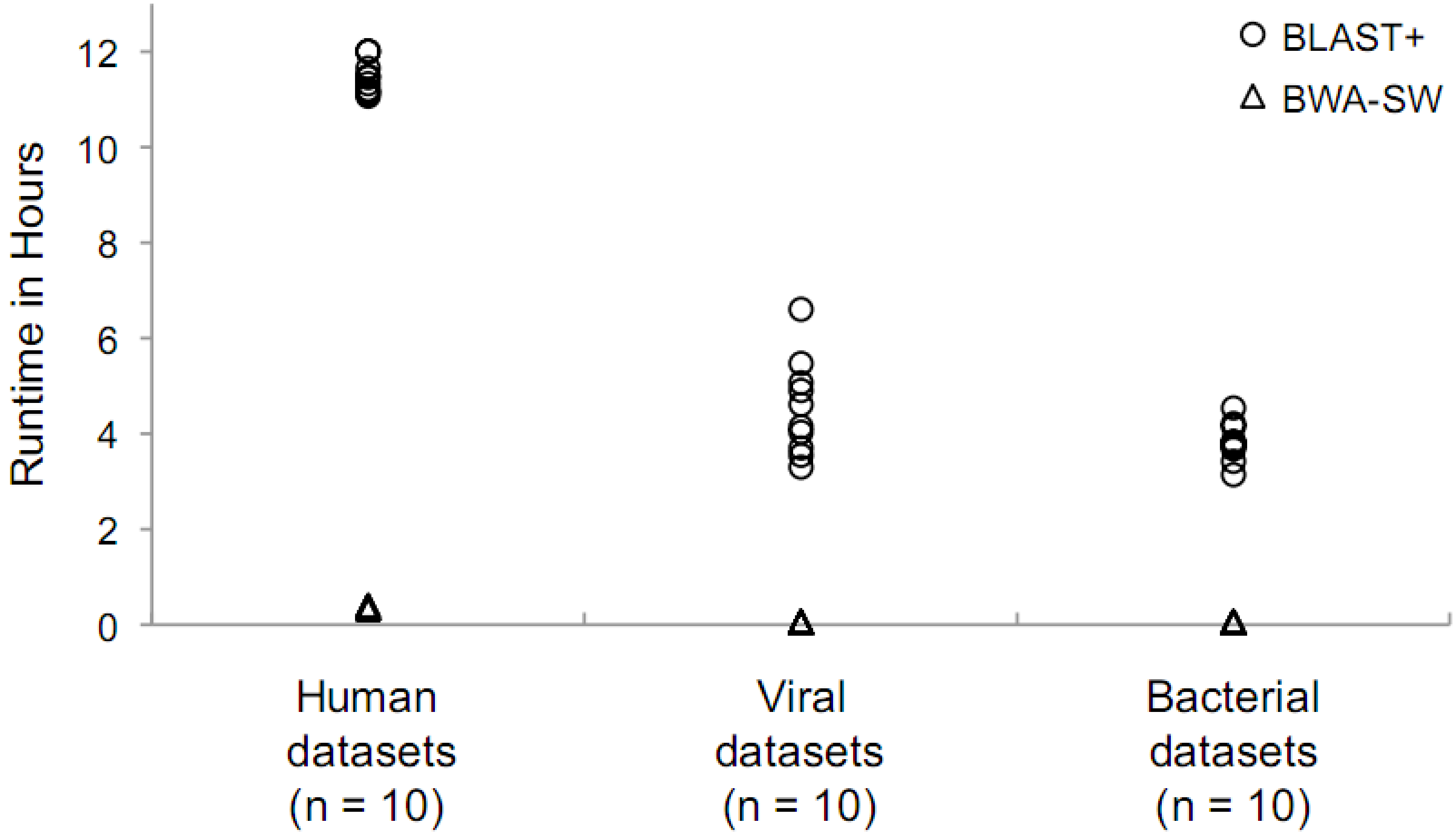


**Figure 1: Comparison of runtimes between BLAST+ and BWA-SW**

Comparison of the runtimes between BLAST+ and BWA-SW for ten human, viral and bacterial simulated metagenomic datasets. BWA-SW performed with the lowest running time of approximately 22 minutes for the human simulated datasets and four minutes for the bacterial and viral simulated datasets.

# BWA-SW limitations

- 2-bit representation of DNA sequence: N converted to random ACGT. Possible false positive hits  
(SSAHA2 converts all N to A)
- BWA faster on short reads and small genomes
- SSAHA2 more accurate for reads with high error rates
- BWA-SW fails to index complete multiple human genome dataset (4 GB max)

Output:

- SAM output contains too much data, huge files
- SAM mapping quality useless
- Cigar output no identity value

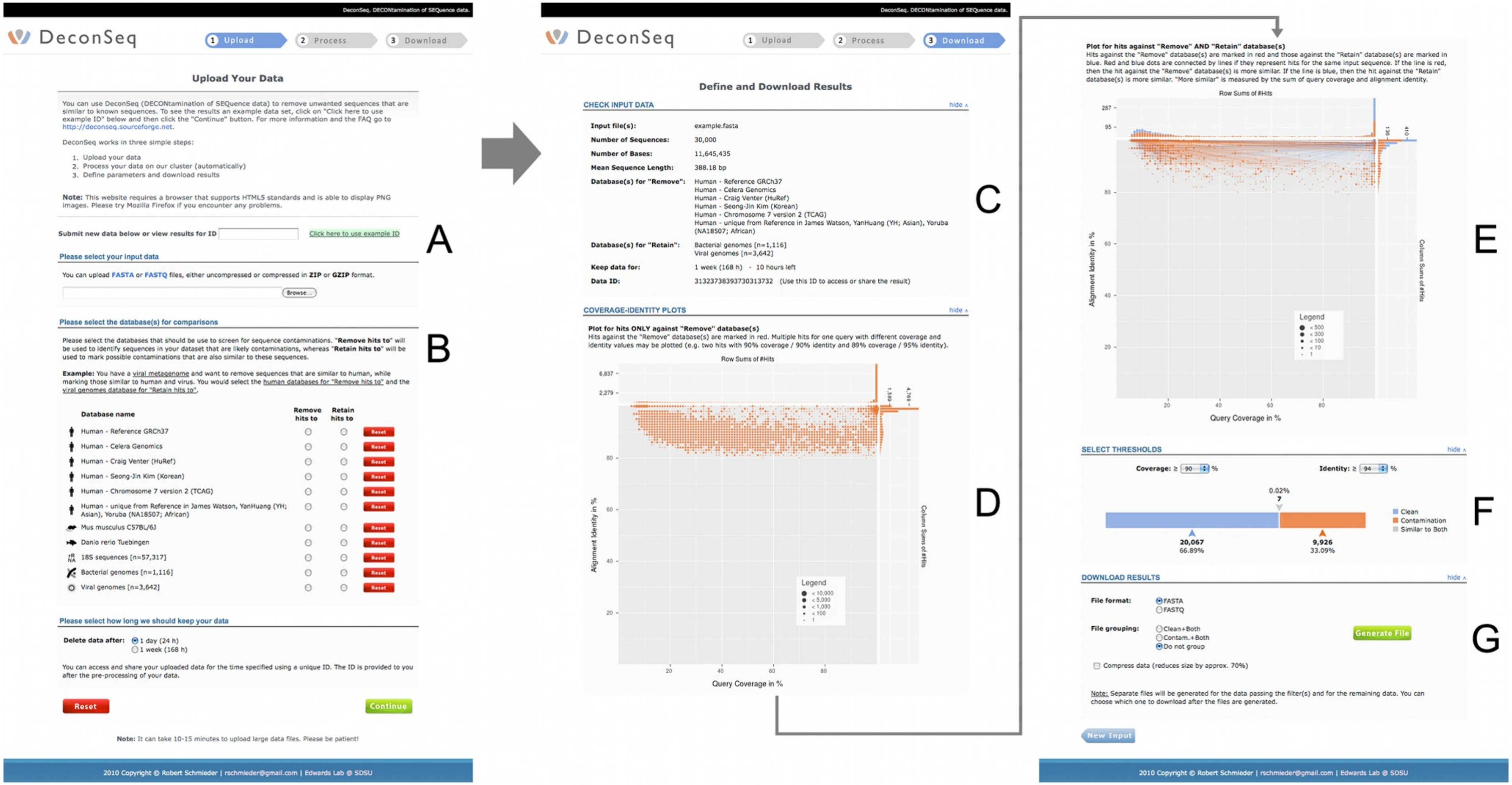
This required modification of BWA-SW and reference data

# DeconSeq

- standalone and web-based
- implemented in Perl, based on BWA-SW in C
- Alignments computed on cluster
  10 nodes, each 8 CPU and 16 GB RAM
- input data automatically split into chunks for distribution over nodes
- input FASTA or FASTQ (also in ZIP or GZIP)
- no limit on number of sequences or size of input file

- User selects coverage and identity thresholds!

# DeconSeq



DeconSeq. DECONtamination of SEQuence data.

DeconSeq

1 Upload | 2 Process | 3 Download

### Upload Your Data

You can use DeconSeq (DECONtamination of SEQuence data) to remove unwanted sequences that are similar to known sequences. To see the results an example data set, click on "Click here to use example ID" below and then click the "Continue" button. For more information and the FAQ go to http://deconseq.sourceforge.net.

DeconSeq works in three simple steps:

1. Upload your data
2. Process your data on our cluster (automatically)
3. Define parameters and download results

**Note:** This website requires a browser that supports HTML5 standards and is able to display PNG images. Please try Mozilla Firefox if you encounter any problems.

**Submit new data below or view results for ID** ________ Click here to use example ID

**Please select your input data**

You can upload **FASTA** or **FASTQ** files, either uncompressed or compressed in **ZIP** or **GZIP** format.

________ Browse...

**Please select the database(s) for comparisons**

Please select the databases that should be use to screen for sequence contaminations. "**Remove hits to**" will be used to identify sequences in your dataset that are likely contaminations, whereas "**Retain hits to**" will be used to mark possible contaminations that are also similar to these sequences.

**Example:** You have a viral metagenome and want to remove sequences that are similar to human, while marking those similar to human and virus. You would select the human databases for "Remove hits to" and the viral genomes database for "Retain hits to".

| Database name | Remove hits to | Retain hits to | |
|---|---|---|---|
| Human - Reference GRCh37 | ○ | ○ | Reset |
| Human - Celera Genomics | ○ | ○ | Reset |
| Human - Craig Venter (HuRef) | ○ | ○ | Reset |
| Human - Seong-Jin Kim (Korean) | ○ | ○ | Reset |
| Human - Chromosome 7 version 2 (TCAG) | ○ | ○ | Reset |
| Human - unique from Reference in James Watson, YanHuang (YH; Asian), Yoruba (NA18507; African) | ○ | ○ | Reset |
| Mus musculus C57BL/6J | ○ | ○ | Reset |
| Danio rerio Tuebingen | ○ | ○ | Reset |
| 18S sequences [n=57,317] | ○ | ○ | Reset |
| Bacterial genomes [n=1,116] | ○ | ○ | Reset |
| Viral genomes [n=3,642] | ○ | ○ | Reset |

**Please select how long we should keep your data**

**Delete data after:** ● 1 day (24 h)
○ 1 week (168 h)

You can access and share your uploaded data for the time specified using a unique ID. The ID is provided to you after the pre-processing of your data.

Reset | Continue

**Note:** It can take 10-15 minutes to upload large data files. Please be patient!

2010 Copyright © Robert Schmieder | rschmieder@gmail.com | Edwards Lab @ SDSU

DeconSeq. DECONtamination of SEQuence data.

DeconSeq

1 Upload | 2 Process | 3 Download

### Define and Download Results

**CHECK INPUT DATA** hide ^

| | |
|---|---|
| **Input file(s):** | example.fasta |
| **Number of Sequences:** | 30,000 |
| **Number of Bases:** | 11,645,435 |
| **Mean Sequence Length:** | 388.18 bp |
| **Database(s) for "Remove":** | Human - Reference GRCh37<br>Human - Celera Genomics<br>Human - Craig Venter (HuRef)<br>Human - Seong-Jin Kim (Korean)<br>Human - Chromosome 7 version 2 (TCAG)<br>Human - unique from Reference in James Watson, YanHuang (YH; Asian), Yoruba (NA18507; African) |
| **Database(s) for "Retain":** | Bacterial genomes [n=1,116]<br>Viral genomes [n=3,642] |
| **Keep data for:** | 1 week (168 h) - 10 hours left |
| **Data ID:** | 31323738393730313732 (Use this ID to access or share the result) |

**COVERAGE-IDENTITY PLOTS** hide ^

**Plot for hits ONLY against "Remove" database(s)**
Hits against the "Remove" database(s) are marked in red. Multiple hits for one query with different coverage and identity values may be plotted (e.g. two hits with 90% coverage / 90% identity and 89% coverage / 95% identity).

**Plot for hits against "Remove" AND "Retain" database(s)**
Hits against the "Remove" database(s) are marked in red and those against the "Retain" database(s) are marked in blue. Red and blue dots are connected by lines if they represent hits for the same input sequence. If the line is red, then the hit against the "Remove" database(s) is more similar. If the line is blue, then the hit against the "Retain" database(s) is more similar. "More similar" is measured by the sum of query coverage and alignment identity.

**SELECT THRESHOLDS** hide ^

**Coverage:** ≥ 90 %  **Identity:** ≥ 94 %

**DOWNLOAD RESULTS** hide ^

**File format:** ● FASTA
○ FASTQ

**File grouping:** ○ Clean+Both
○ Contam.+Both
● Do not group

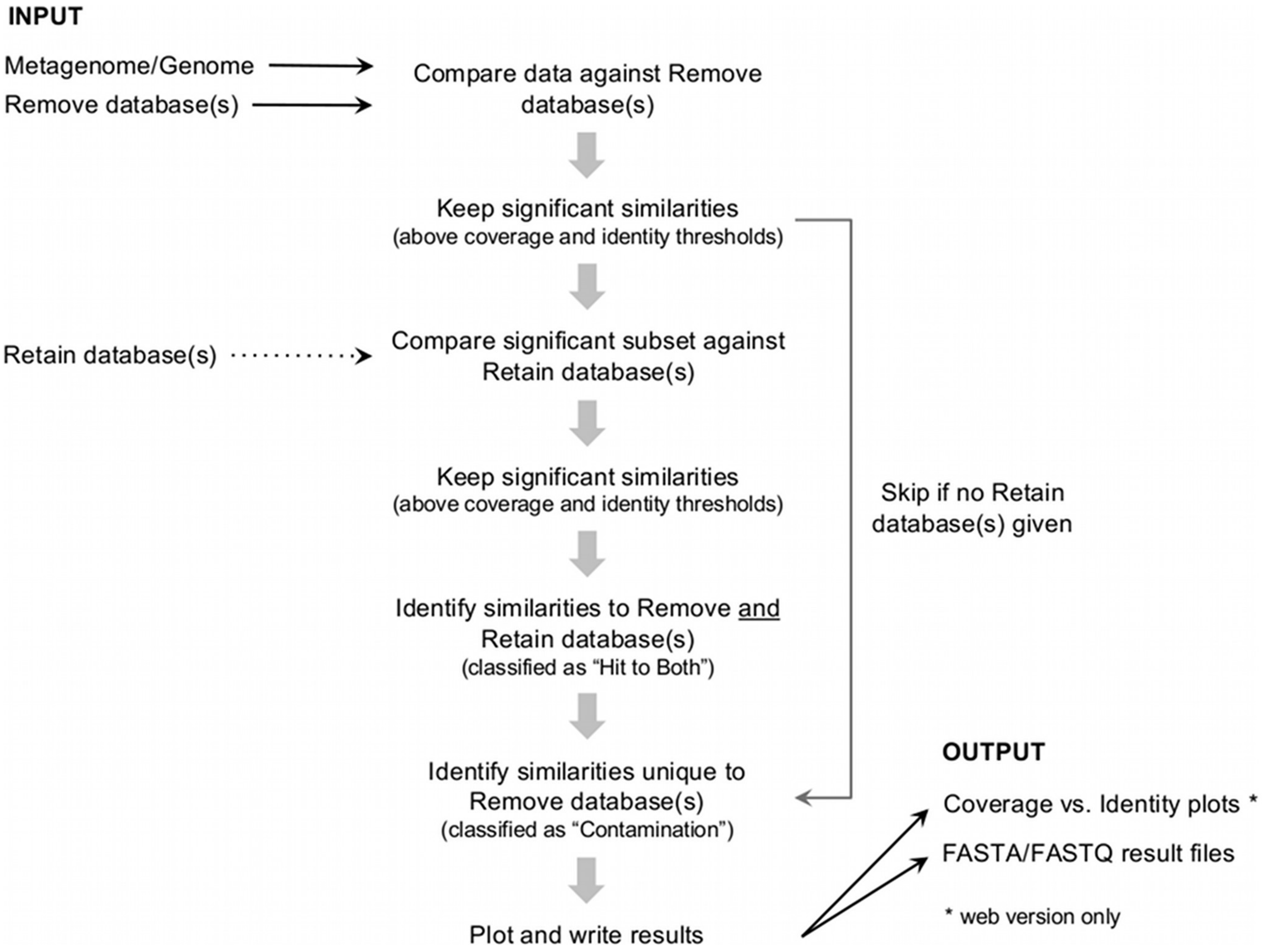
Generate File

□ Compress data (reduces size by approx. 70%)

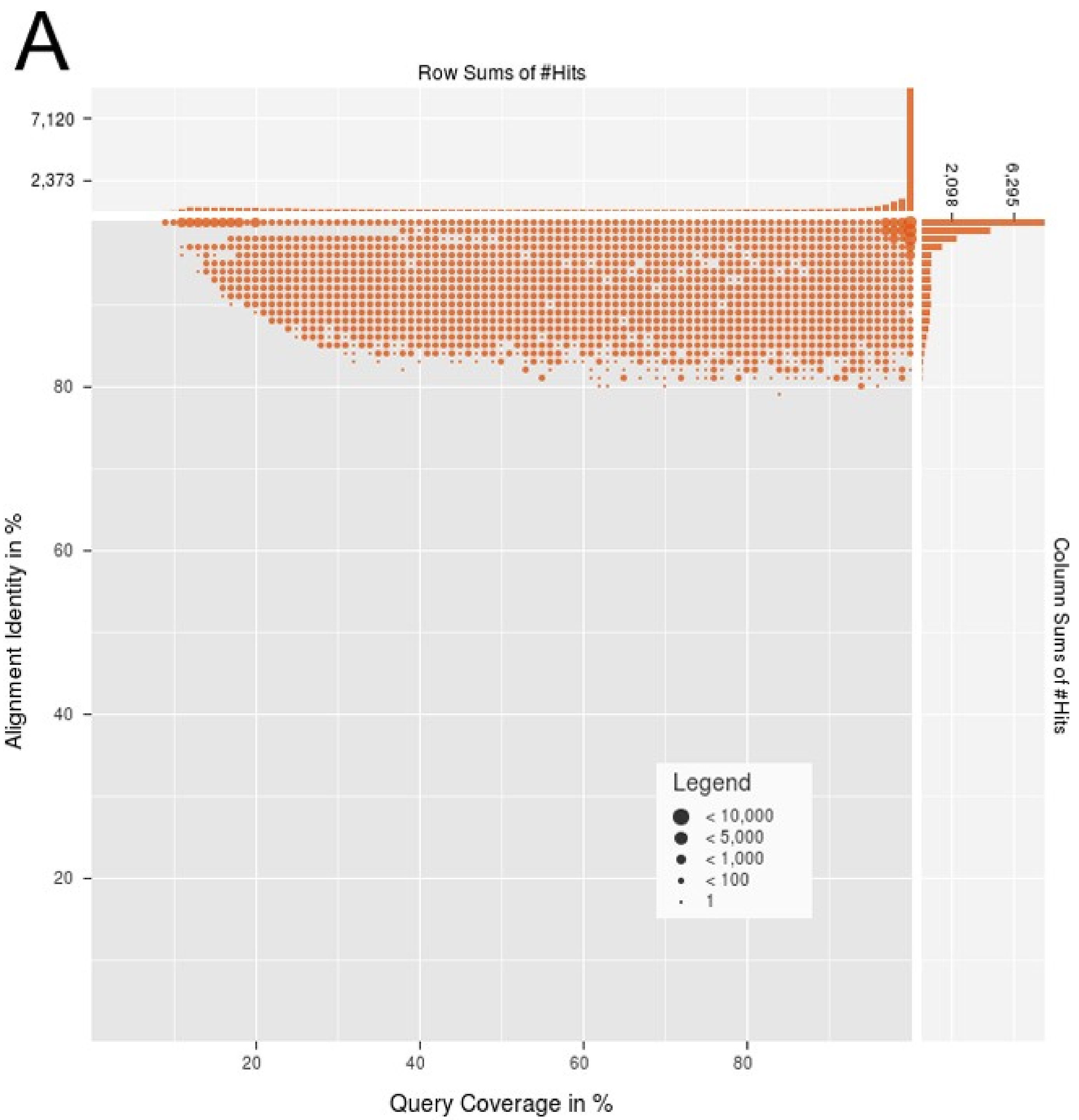
Note: Separate files will be generated for the data passing the filter(s) and for the remaining data. You can choose which one to download after the files are generated.

New Input

2010 Copyright © Robert Schmieder | rschmieder@gmail.com | Edwards Lab @ SDSU

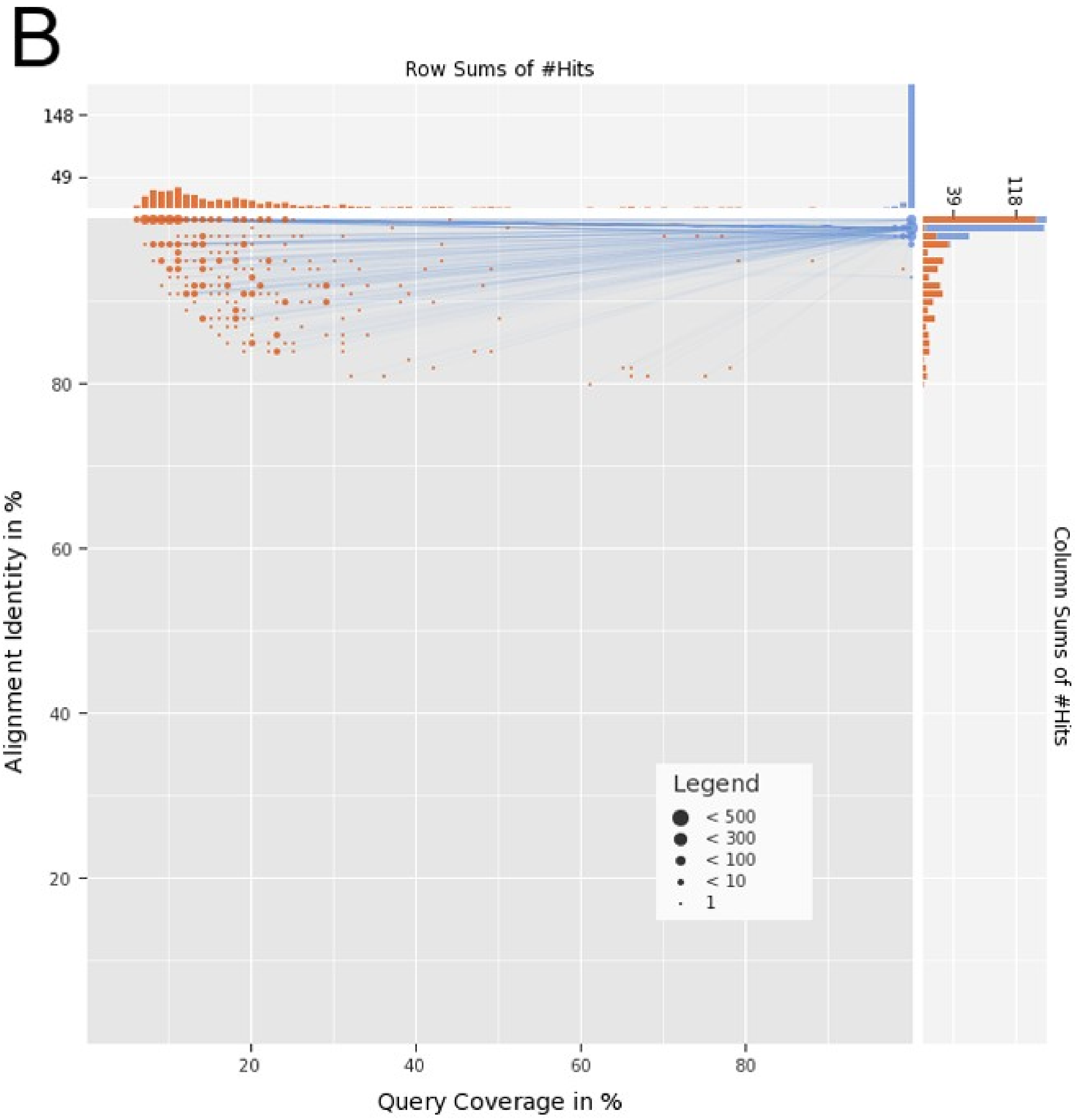

INPUT
Metagenome/Genome
Remove database(s)
Compare data against Remove database(s)
Keep significant similarities
(above coverage and identity thresholds)
Retain database(s)
Compare significant subset against Retain database(s)
Keep significant similarities
(above coverage and identity thresholds)
Skip if no Retain database(s) given
Identify similarities to Remove and Retain database(s)
(classified as "Hit to Both")
Identify similarities unique to Remove database(s)
(classified as "Contamination")
Plot and write results
OUTPUT
Coverage vs. Identity plots *
FASTA/FASTQ result files
* web version only

Matching reads against "remove" database



Matching reads against both “remove” (red) and “retain” (blue) databases. Majority is more similar to “retain”

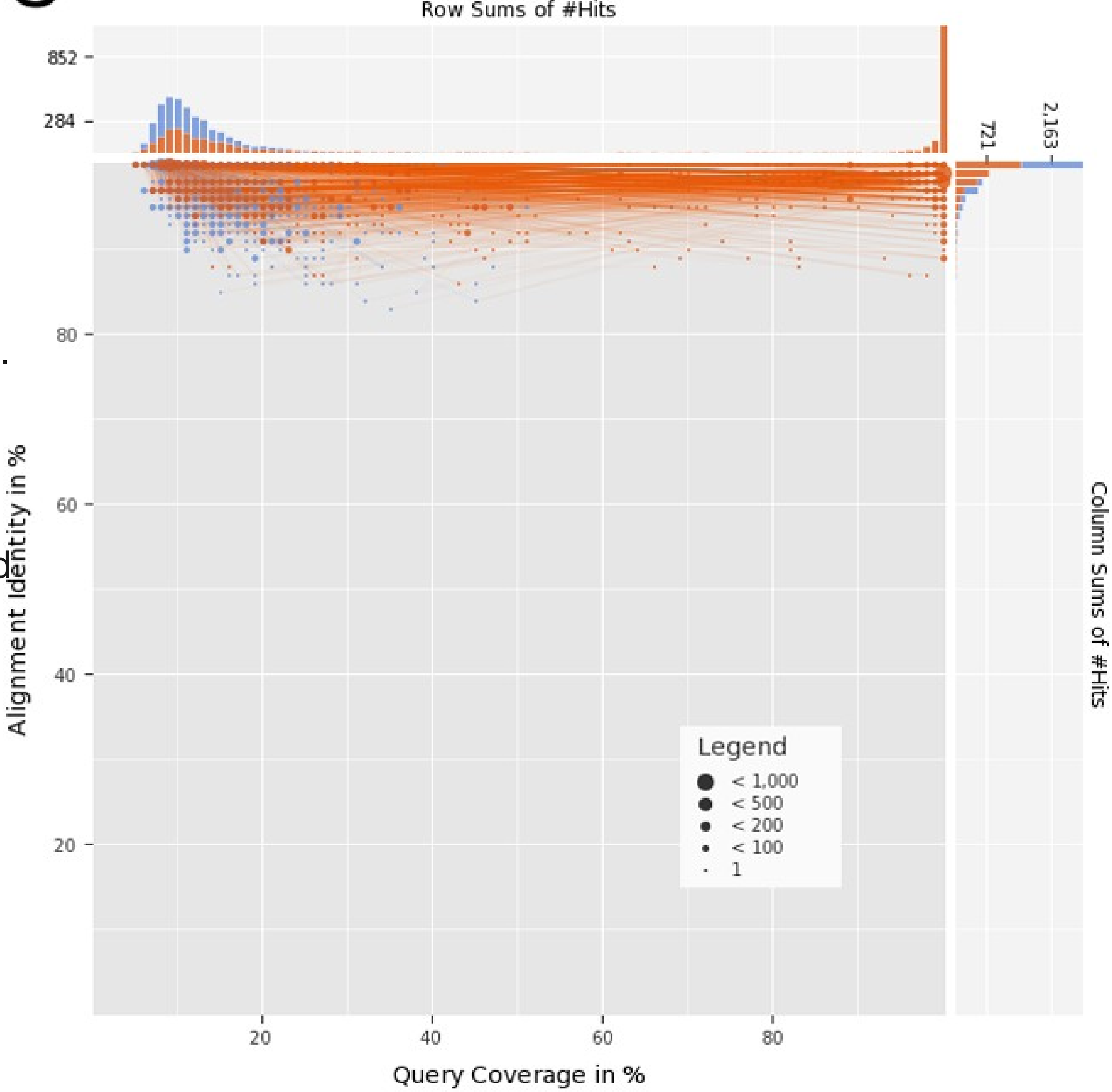
Reads matching both databases are connected by lines



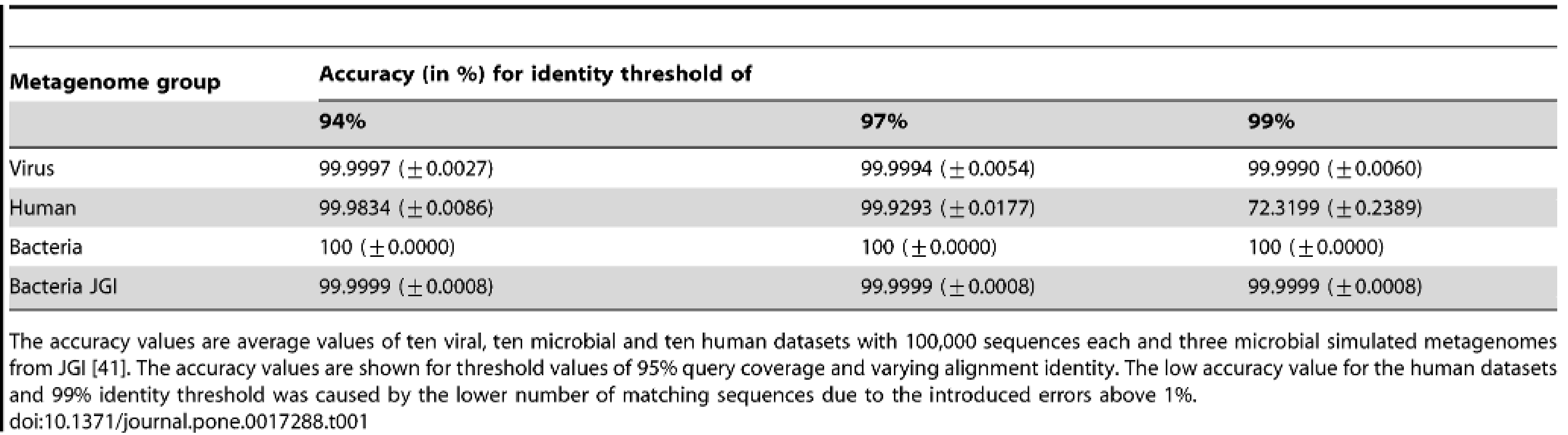
Matching reads against both “remove” (red) and “retain” (blue) databases. Majority is more similar to “remove”

Reads matching both databases are connected by lines

C



# Evaluation of DeconSeq accuracy



| Metagenome group | Accuracy (in %) for identity threshold of | | |
|---|---|---|---|
| | **94%** | **97%** | **99%** |
| Virus | 99.9997 (±0.0027) | 99.9994 (±0.0054) | 99.9990 (±0.0060) |
| Human | 99.9834 (±0.0086) | 99.9293 (±0.0177) | 72.3199 (±0.2389) |
| Bacteria | 100 (±0.0000) | 100 (±0.0000) | 100 (±0.0000) |
| Bacteria JGI | 99.9999 (±0.0008) | 99.9999 (±0.0008) | 99.9999 (±0.0008) |

The accuracy values are average values of ten viral, ten microbial and ten human datasets with 100,000 sequences each and three microbial simulated metagenomes from JGI [41]. The accuracy values are shown for threshold values of 95% query coverage and varying alignment identity. The low accuracy value for the human datasets and 99% identity threshold was caused by the lower number of matching sequences due to the introduced errors above 1%.

doi:10.1371/journal.pone.0017288.t001

# Identification of human contamination in metagenomes

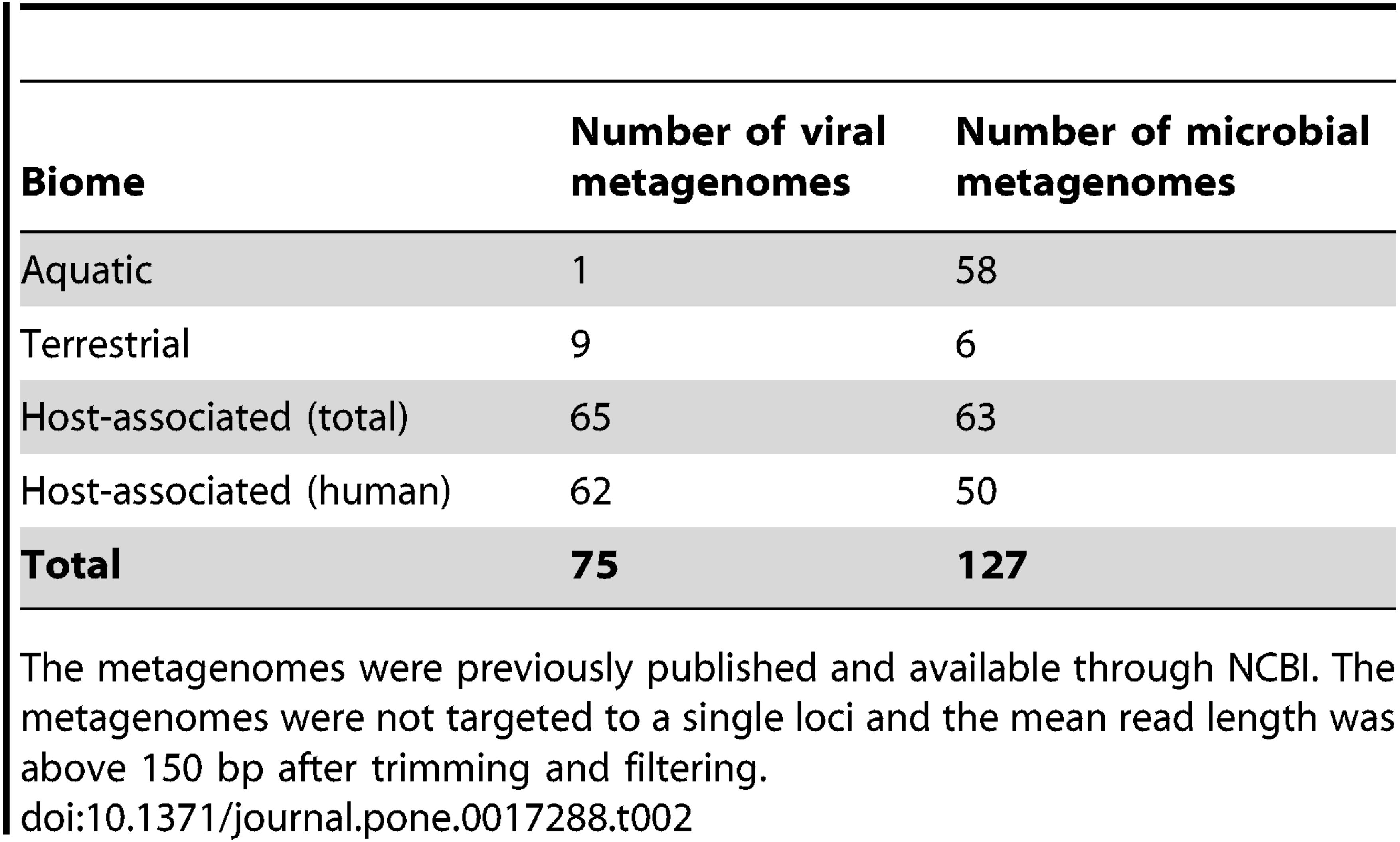
- 202 longer read metagenomes from NCBI (150 bp mean read length)
- viral and bacterial communities from different biomes

Query pre-processed:

- screened for vector contamination with UniVec and cross_match (phrap)
- TagCleaner was used to trim adapter and tag sequences
- PRINSEQ was used to remove exact read duplicates, sequences shorter than 50 or longer than 10000
  sequences containing more than 5% N
  sequences containing non IUPAC characters for DNA

DeconSeq was run on all human databases for “remove” and corresponding type (microbial or virus) database for “retain”

# Identification of human contamination in 202 metagenomes

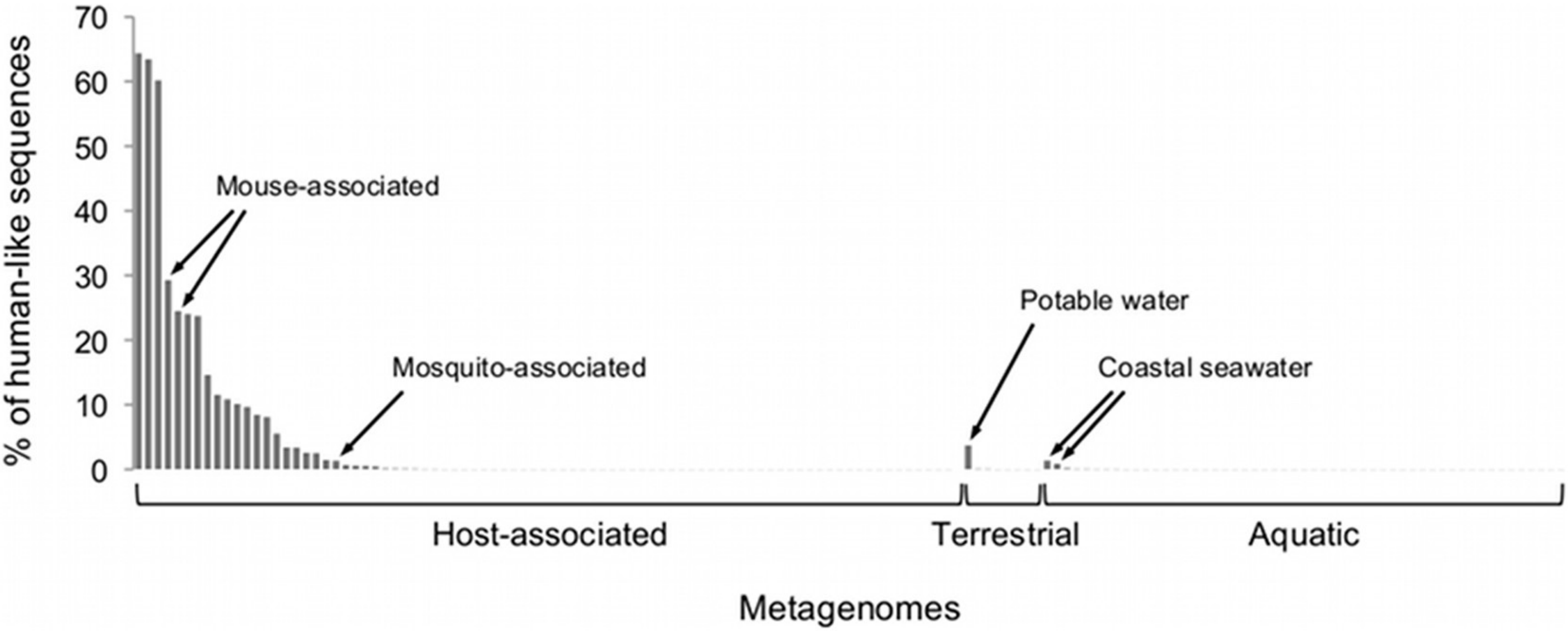


| Biome | Number of viral metagenomes | Number of microbial metagenomes |
|---|---|---|
| Aquatic | 1 | 58 |
| Terrestrial | 9 | 6 |
| Host-associated (total) | 65 | 63 |
| Host-associated (human) | 62 | 50 |
| **Total** | **75** | **127** |

The metagenomes were previously published and available through NCBI. The metagenomes were not targeted to a single loci and the mean read length was above 150 bp after trimming and filtering.

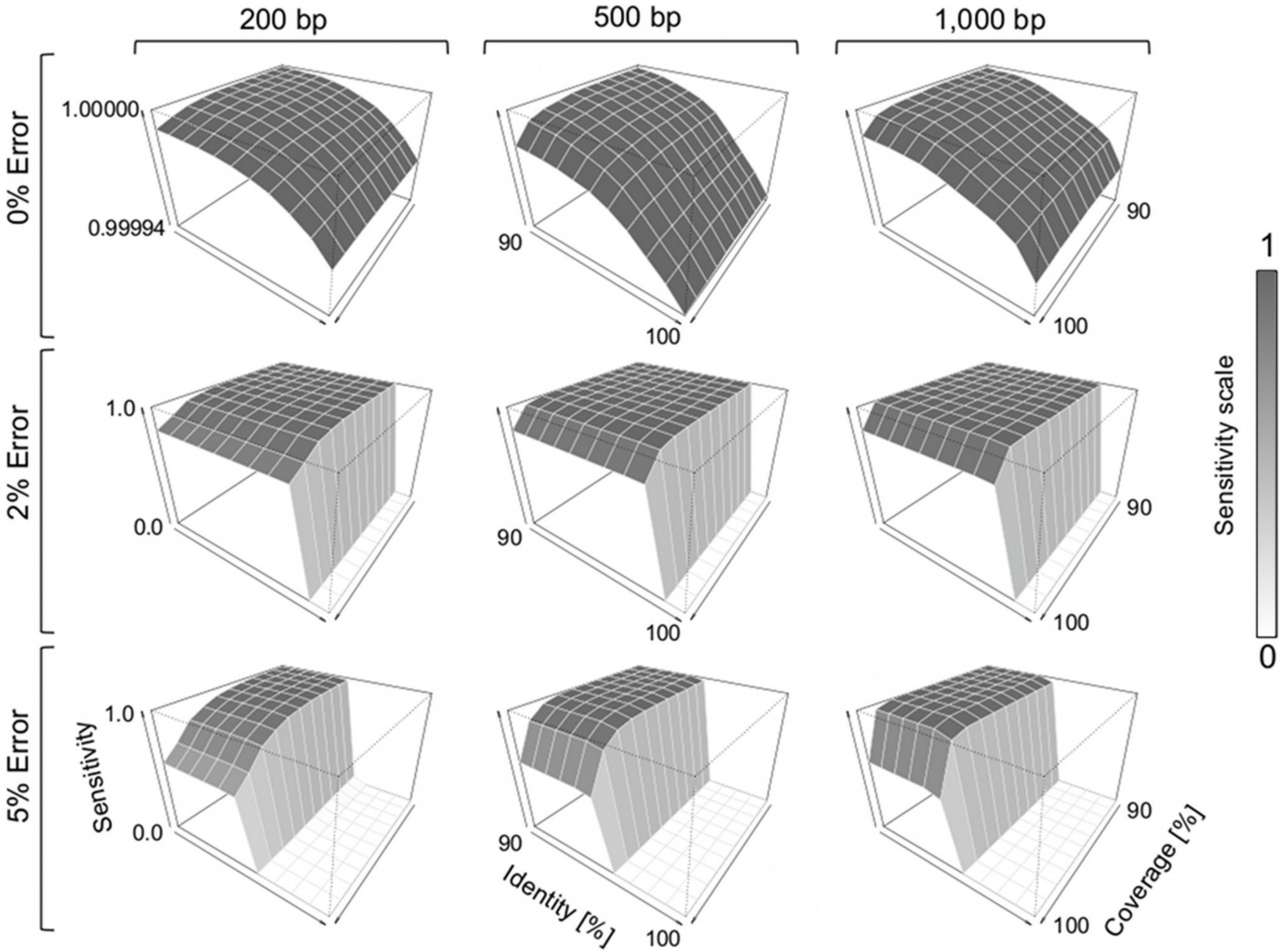
doi:10.1371/journal.pone.0017288.t002

# Identification of human contamination in 202 metagenomes



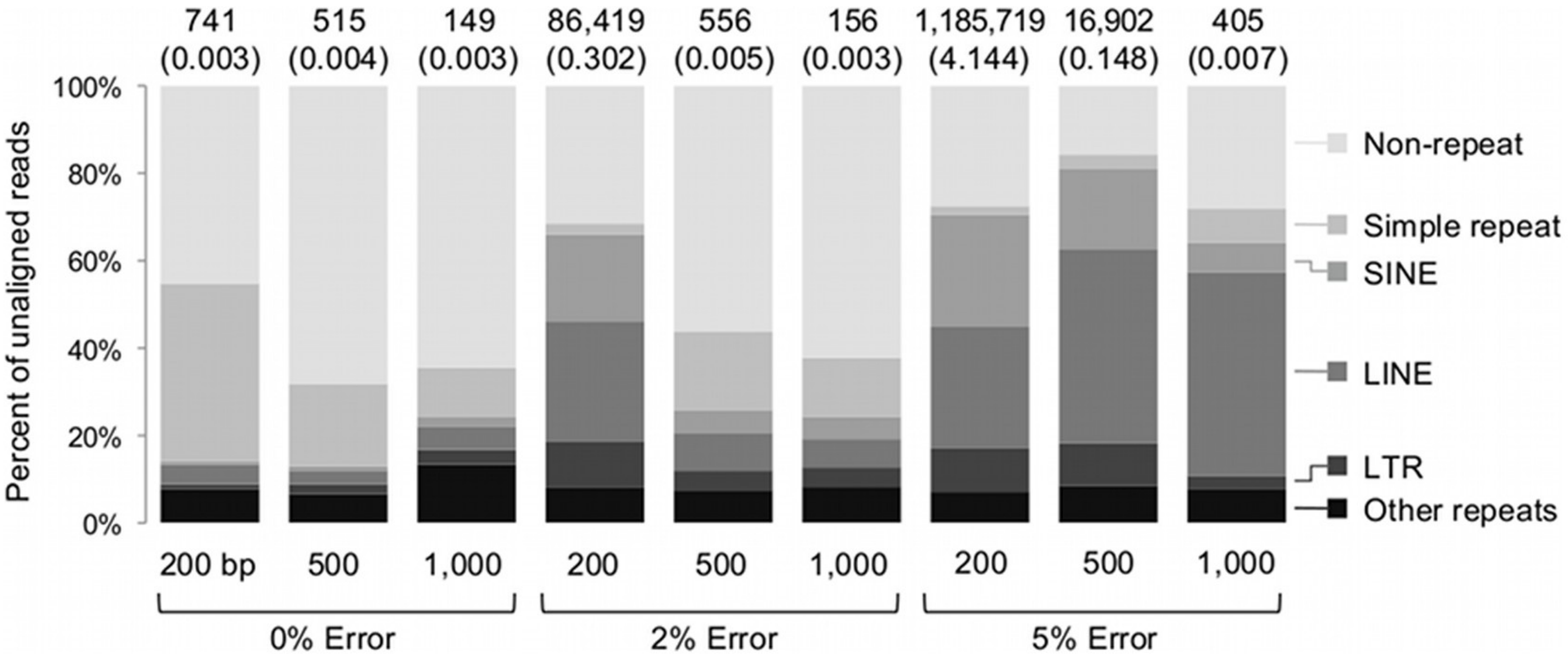
Human contamination up to 64% of metagenome
145 (72%) metagenomes contained at least one possible contamination sequence
Two mouse associated metagenomes reported 24% and 29% possible human contamination. Origin probably host-related (56% and 57% mouse-like sequences)

# BWA-SW alignment sensitivity



**Alignment sensitivity of BWA-SW for human sequences.**
Query coverage and alignment identity values ranged from 90% to 100%. The sensitivity shows how many sequences could be aligned back to the reference. The simulated datasets contained 28,612,955 reads for 200 bp, 11,444,886 reads for 500 bp, and 5,722,210 reads for 1,000 bp.

# Unaligned reads



**Repeats causing alignment problems for BWA-SW.**
The query coverage was set to 95%, with identity set to 99%, 97% and 94% for error rates of 0%, 2% and 5%, respectively. The numbers above the bars show the number of unaligned sequences of each category for the given thresholds. The values shown in parenthesis represent the percentage of unaligned sequences. The simulated datasets contained 28,612,955 reads for 200 bp, 11,444,886 reads for 500 bp, and 5,722,210 reads for 1,000 bp.

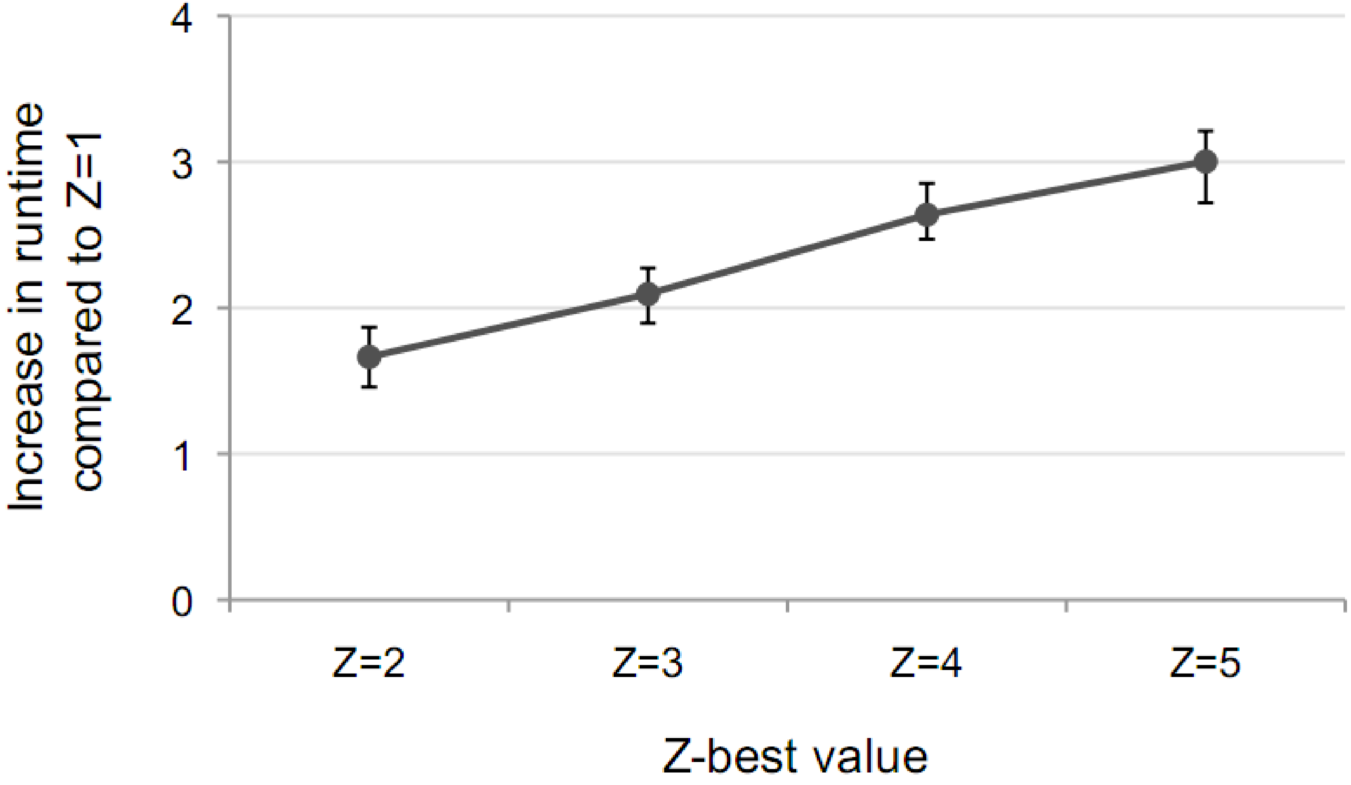


**Figure 3: Runtime of BWA-SW for different *Z*-best values**

The change in runtime was measured for *Z*-best values ranging from one to five using the 30 simulated metagenomic datasets with 100,000 sequences each. The sequences were compared to the human reference genome and the change in runtime compared to the $Z = 1$ runtime is shown.

To identify contamination it is sufficient to find a single match above given threshold without calculation all possible matches

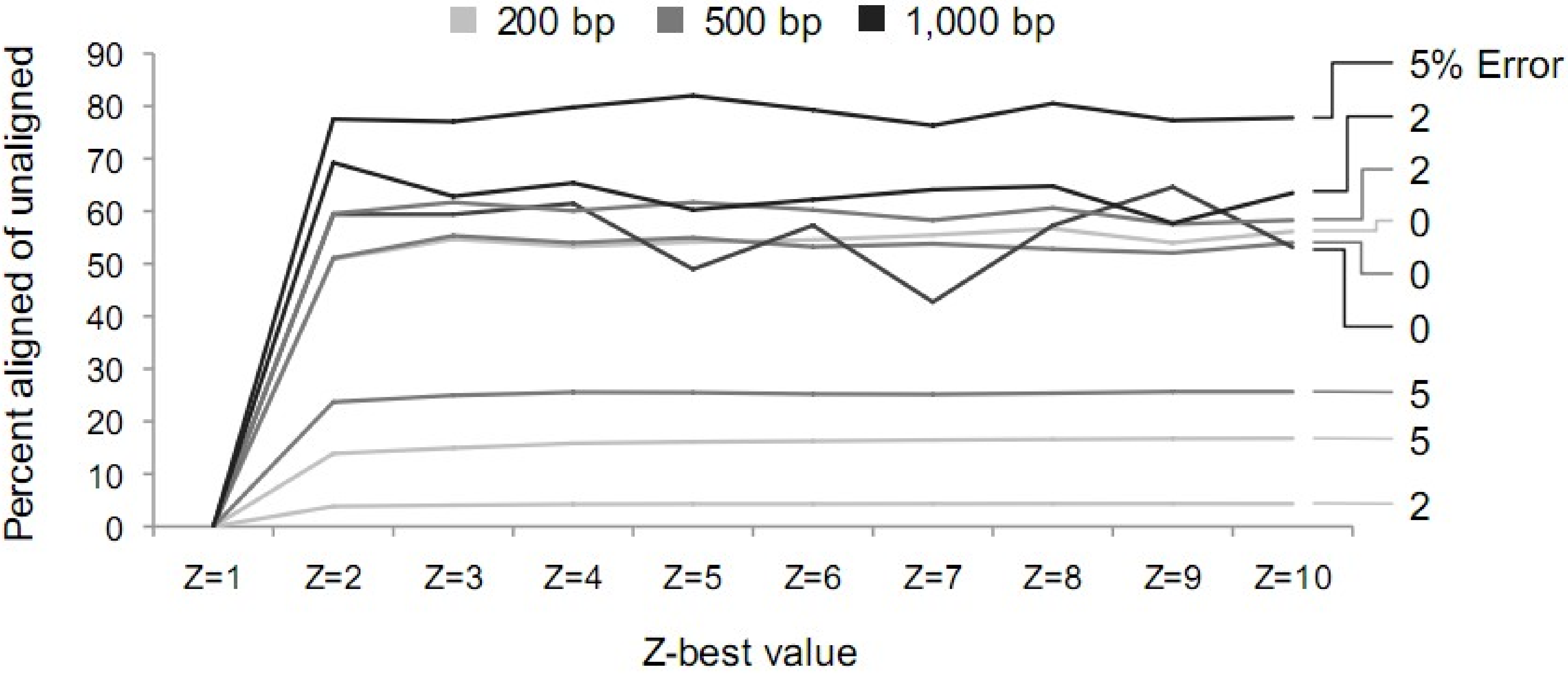


**Figure 2: Percentage of unaligned sequences that could be aligned using higher *Z*-best values**

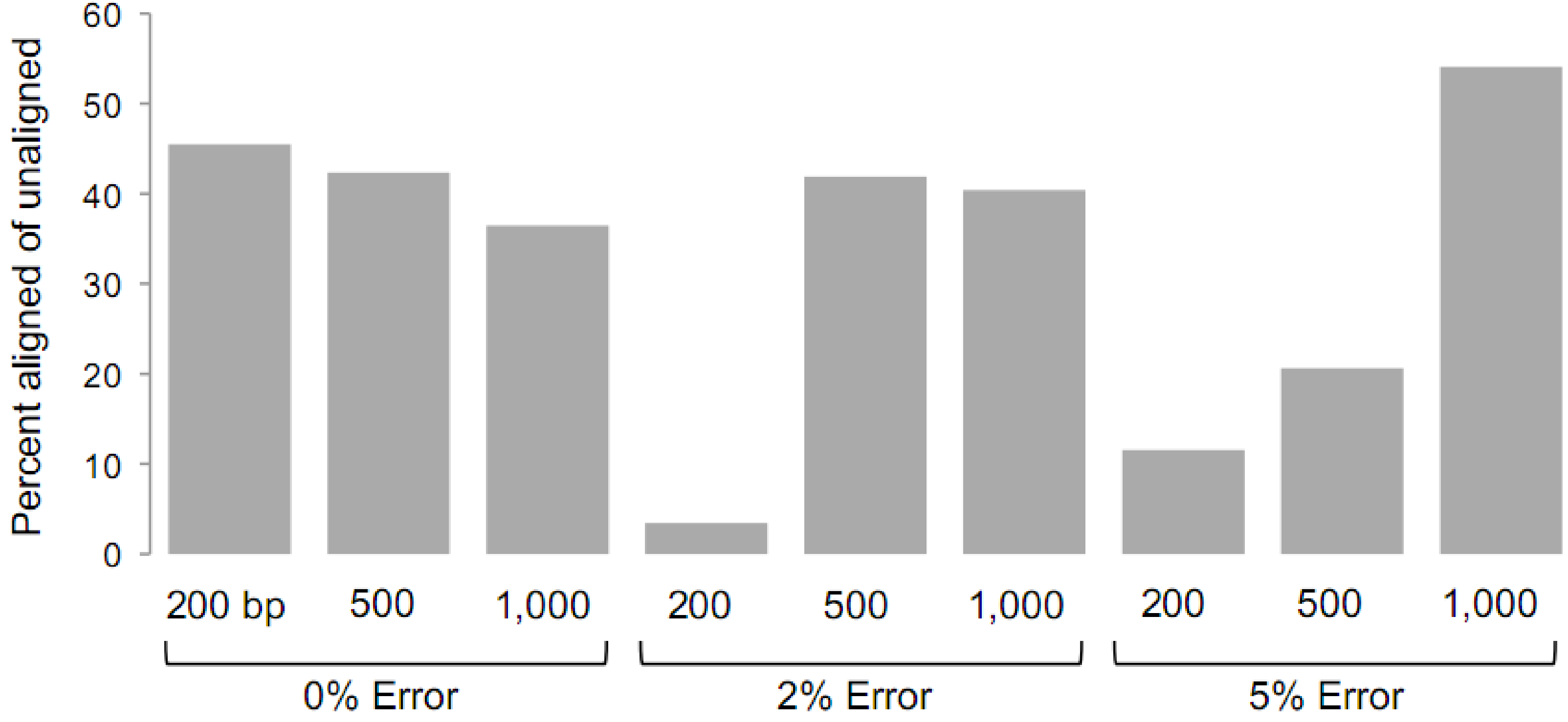


**Figure 4: Percentage of unaligned sequences that could be aligned using additional human genome data**

# Conclusions

- sequence contamination a serious concern to quality of data
- Burrows-Wheeler algorithm was adopted as optimal in speed/memory/accuracy
- DeconSeq allows rapid and robust identification and removal of sequence contaminants
- contamination was detected in 145 of 202 previously published metagenomes
- contamination was also detected in non human-associated metagenomes, suggesting that this method is appropriate for screening all metagenomes