Supporting Information for

Prospective identification of parasitic sequences in phage-display screens

Wadim L. Matochko, S. Cory Li, Sindy K.Y. Tang, Ratmir Derda*

* - corresponding author: ratmir@ualberta.ca

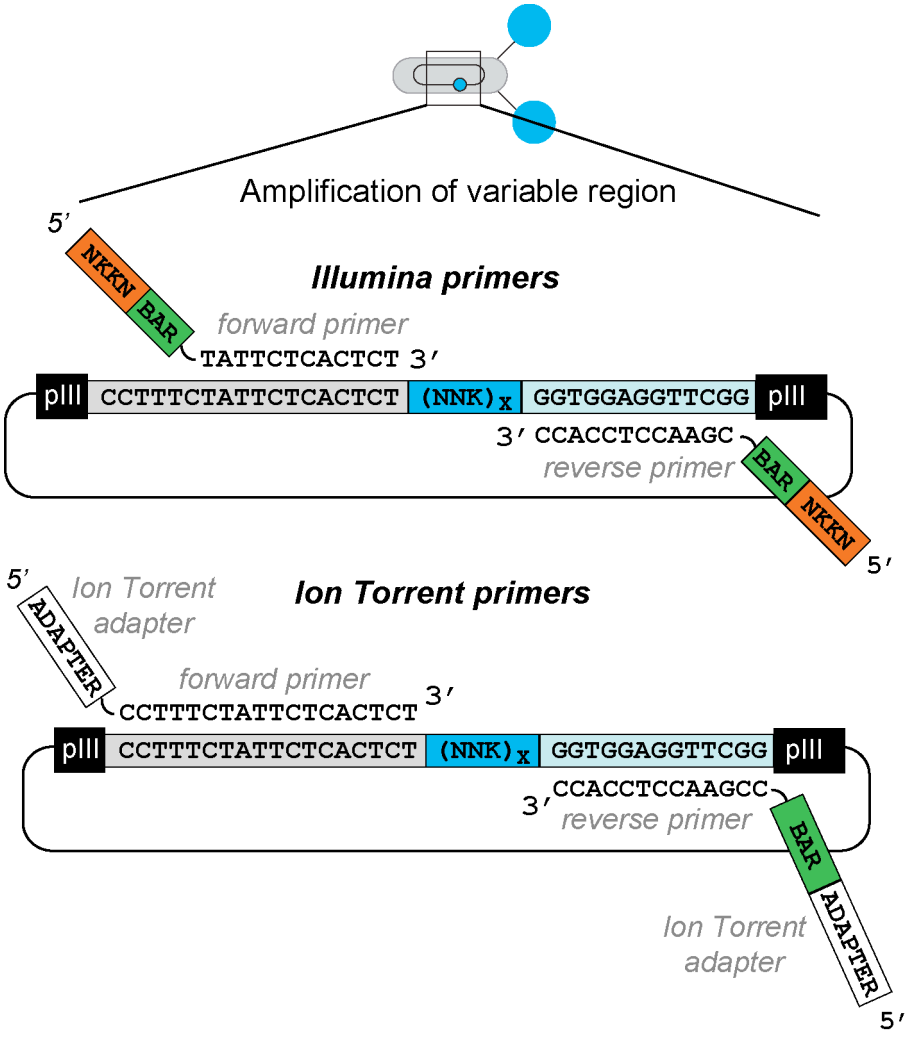
Table of Content

Scheme S1. Illumina and Ion Torrent primer design. Primer sequences are universal for all libraries made by New England Biolabs (Ph.D.-7™, Ph.D.-12™, and Ph.D.-C7C™) because all three libraries contain the same flanking regions. Note that Tyr-Ser-His-Ser is part of the pIII leader sequence, which is removed during the periplasmic export of the phage.

PhD7: ..TAT TCT CAC TCT (NNK)7 GGT GGA GGT TCG GCC..
Tyr Ser His Ser Rnd Gly Gly Gly Ser Ala

PhD12: ..TAT TCT CAC TCT (NNK)12 GGT GGA GGT TCG GCC..
Tyr Ser His Ser Rnd Gly Gly Gly Ser Ala

PhDC7C: ..TAT TCT CAC TCT GCT TGT (NNK)7 TGC GGT GGA GGT TCG GCC..
Tyr Ser His Ser Ala Cys Rnd Cys Gly Gly Gly Ser Ala



Scheme S2. Illumina Analysis workflow.

1. Read FASTAQ file. Parse the reads and create a tagged file containing reads in which the adapters can be identified. Discard non-tagged reads (~2%)
Example of the tagged read:

```
Library
 |  Adapter
 |   |    Adapter
 |   |    quality
 |   |     |        barcode
 |   |     |    NKKN  |  adapter seq.    variable sequences                      quality string parsed in the same way as sequence string
 |   |     |      |   |      |                                  |
Ph7 <F> <PERF> ATGC TTG TATTCTCACTCT ACTACGGTTACGGGGAGTCTGGTGGAGGTTCG.......  CCCF FFF EHHHHHJJJJJJ JJJJJJJIIJJJHIJJGJFHHJJHHHHHFFFFD.......
Ph7 <F> <PERF> CTTC TTG TATTCTCACTCT GTTGATAGGCTTCTGTTGATGGGTGGAGGTTCG.......  CCCF FFF FHHHHHJJJJJJ JJJ+AFHIJJJJJJJIIIFHIJJDGIGIJIIJJ.......
P12 <F> <PERF> GTGC TCA TATTCTCACTCT GCGGAGCCGACGGATTGGCTTACTGTTCGTGAT.......  CBBF FFF FHHHHHJJJJJJ JJJ)BFHJJJJJIIJJJJJJJIJJJJJJJHHHF.......
P12 <F> <PERF> GGGG TTG TATTCTCACTCT TGGTCTCCGGGTCGTGGTCCGATTGAGTCGCAG.......  BCCF DDF DHHHHHJJJJJJ JJJHIIJJJJJGHIHIJHIJJJJIIGHHHIIHH.......
Ph7 <F> <PERF> CTTG CGA TATTCTCACTCT CGGACTGCGCCTACTGTGATGGGTGGAGGTTCG.......  B@@F FFD FHHHHHJJJJJJ JJJJJIIIJIJJHIIJGGDIGIJFHIIJJJJJG.......
C7C <F> <PERF> TGGC TTG TATTCTCACTCT GCTTGTCATTGGTCTCCGGCTTCTCCGTGCGGT.......  ???; 4B1 AD?DDDEEEEEE C@<E9<::9E9<2AEECEADDA)?9?D/??AD#.......
P12 <F> <PERF> GTGG GAC TATTCTCACTCT ACGCCGAATGCTGTTTTGCCGACTGGGCGTACG.......  @@@D FBD FFHHHHHIEIIII IGHIIIIIIHIIDGHFHGIIIIAHGCGFHIHHF.......
Ph7 <F> <PERF> GGGC TTG TATTCTCACTCT AGGCCTGCGTCTCTGCCTCCGGGTGGAGGTTCG.......  CBCD FFF FGHHHHJJJJJJ IJJJIIEGIHGGIIFIJJJJIIJFHJ;FH@EHI.......
Ph7 <F> <PERF> CGGG TTG TATTCTCACTCT CAGCGGCATCATGCGGATGTTGGTGGAGGTTCG.......  B@BF DDF FHHHHHJJJJJJ JGJJJJJIJJJFHIJHJDGHJJEGIIJJJJHH.......
P12 <F> <PERF> AGGG TCA TATTCTCACTCT GTTCCTGCTTGGGCGGTGGAGGTTCGTACGCCT.......  BCCF DDF FHHHHHJJJJJJ JJJJJJJJJJJJHIJJGHGHIJFHIJHIIHHHF.......
Ph7 <F> <PERF> AGGG TCA TATTCTCACTCT ATTGAGTCGACGCGGCATCTGGGTGGAGGTTCG.......  BCCF DDF FHHHHHJJJJJJ JJJ+AFFHIJJJJJJJJJJJJJJ@HIHHHFFFF.......
C7C <F> <PERF> CGGT CGA TATTCTCACTCT GCTTGTCATCAGGCTGGTGCTCCGCATTGCGGT.......  @@CD FFF FHHHHHJJJJJJ JGHGHIIJJGGHHIIIF?DFGGIIJ0DFHHIHF.......
          . . .
C7C <R> <PERF> CGTC GTA CGAACCTCCACC GCAAAAACTCGACGCCTGATTCTTACAAGCAGA.......  CCCF FFF FHHHHHJJJJJJ JJJJJ00:?BG)?FHIJJJJJJJJJJHHHHHFF.......
Ph7 <R> <PERF> CTTG TTG CGAACCTCCACC ATGAAGACGCATACTAAACGCAGAGTGAGAATA.......  CCCF FFF FHGHHHJJJJJJ JIJIJCIJJJJJJJJJJJGJIJJJJJJJJJJG.......
C7C <R> <PERF> AGGT GTA CGAACCTCCACC GCACAGATTCCGCTGCAACGGACTACAAGCAGA.......  BCCD DEF FHHHHHJJJJJJ JJJJJ<GHJHJJJJJJJJIJJJJJJJIJJJHHH.......
P12 <R> <PERF> AGGC GTA CGAACCTCCACC AGAAGTCTCCCTCCTATGCAGCTGCGAATCATG.......  @CCF FDF FFFHHFIJJJJJ GGGHIGIIJIJJGIJJIEIIJJJJJGIJJIJII.......
C7C <R> <PERF> GTGT TTG CGAACCTCCACC GCACTTCTGCACATTATGCCCCCCACAAGCAGA.......  CBBF FFF FHHHHHJJJJJJ JJJJJJJJJJJ9DHIJJJJJJJJJJHHHHFFFF.......
P12 <R> <PERF> TGGG GAC CGAACCTCCACC AACCTTAGAACTAGGCATAGTAAGCGAAATATC.......  CCCF FFF FHHHHHJJJIJJ JJJJJCHG4BHHFIJJJJGIGHIJJJJJJJJJH.......
P12 <R> <PERF> TGGG GAC CGAACCTCCACC CCTAGGCACAACCTCATACGCCTGCCGATACGG.......  CCCF FFF FHHHHHJJJJJJ JJJJJJJJJJJJJJJJJIGI8FHJJJJHHFFDD.......
Ph7 <R> <PERF> CTGG CGA CGAACCTCCACC AGGAAACGAATGCCCAAGCGCAGAGTGAGAATA.......  CCCF FFF FHHHHHJJJJJJ JJJJJ000)0?FGIJJJJFIGIHHHHFFFFFFC.......
C7C <R> <PERF> CGGG GTA CGAACCTCCACC GCACAGCTCAATAGACTTACTCAGACAAGCAGA.......  CCCF FFF FHHHHHJJJJJJ JJJJJIJJJJJJJJJJJJJJJJJJJJJJJJJJJ.......
                                                                  Etc
```

>50,000,000 lines

2. Process tagged file, filter the Phred<30 reads, and reads that contain NNM codons, sort the reads by libraries and experiments (barcodes).
3. Process F and R reads separately. Translate. Save into separate files for inspection.
Example of one of the processed files (F read on the left, R read on the right)

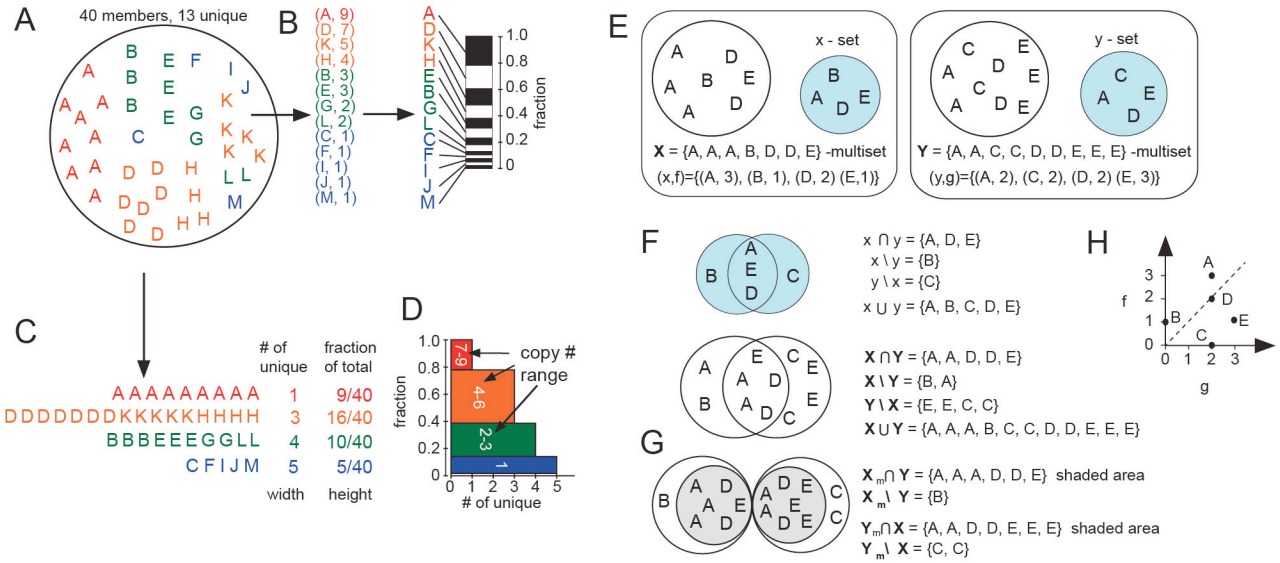
GAGCCGCTGCAGCTGAAGATG	EPLQLKM	74758	GAGCCGCTGCAGCTGAAGATG	EPLQLKM	67210
CATGCTATTTATCCGCGTCAT	HAIYPRH	68194	CATGCTATTTATCCGCGTCAT	HAIYPRH	64004
GGGAAGCCTATGCCTCCGATG	GKPMPPM	60099	GGTCCTATGCTGGCTCGTGGT	GPMLARG	60240
TCGCCGCAGATGACTCTTTCG	SPQMTLS	52692	TCGCCGCAGATGACTCTTTCG	SPQMTLS	57230
CATGCTCTGGGTCCGTCTTCG	HALGPSS	51036	CATGCTCTGGGTCCGTCTTCG	HALGPSS	55316
TGGCCGCAGAAGGCTCAGCCT	WPQKAQP	45254	GGGAAGCCTATGCCTCCGATG	GKPMPPM	54771
TCGACGGCGTCTTATACTCGT	STASYTR	42262	GCGACGACTGTTCCAGCTTCG	ATTVPAS	44445
CAGCCTTGGCCGACGAGTATT	QPWPTSI	31321	TGGCCGCAGAAGGCTCAGCCT	WPQKAQP	37943
TGGCCTACGCCGCCTTATGCG	WPTPPYA	30741	AGTCCGACGCAGCCTAAGTCG	SPTQPKS	34869
AGTCCGACGCAGCCTAAGTCG	SPTQPKS	27246	GGTAAGGTGCAGGCGCAGTCG	GKVQAQS	29403
GCTATGTCGTCTCGTTCGCTT	AMSSRSL	25774	GCGGCTGGTCAGCAGTTTCCT	AAGQQFP	25604
GCGACTCCGCTTTGGCTTAAG	ATPLWLK	25487	CAGCCTTGGCCGACGAGTATT	QPWPTSI	25588
CAGCCTCCTCGTTCGACGTCG	QPPRSTS	22462	GATTCGCATACTCCGCAGAGG	DSHTPQR	24964
CAGGCTACGCATCGTTCGCAT	QATHRSH	20473	CATAGGGCGGATATGCATTTT	HRADMHF	19965
GATTCGCATACTCCGCAGAGG	DSHTPQR	20319	ACGCGGGCTGGTCTGGATTTT	TRAGLDF	19930
GCGACGACTGTTCCAGCTTCG	ATTVPAS	19646	TGGCCTACGCCGCCTTATGCG	WPTPPYA	19884
		Etc >1,000,000 lines			Etc >1,000,000 lines

Combine the F and R reads using multiset union definition.

GAGCCGCTGCAGCTGAAGATG	EPLQLKM	74758
CATGCTATTTATCCGCGTCAT	HAIYPRH	68194
GGTCCTATGCTGGCTCGTGGT	GPMLARG	60240
GGGAAGCCTATGCCTCCGATG	GKPMPPM	60099
TCGCCGCAGATGACTCTTTCG	SPQMTLS	57230
CATGCTCTGGGTCCGTCTTCG	HALGPSS	55316
TGGCCGCAGAAGGCTCAGCCT	WPQKAQP	45254
GCGACGACTGTTCCAGCTTCG	ATTVPAS	44445
TCGACGGCGTCTTATACTCGT	STASYTR	42262
AGTCCGACGCAGCCTAAGTCG	SPTQPKS	34869
CAGCCTTGGCCGACGAGTATT	QPWPTSI	31321
TGGCCTACGCCGCCTTATGCG	WPTPPYA	30741
GGTAAGGTGCAGGCGCAGTCG	GKVQAQS	29403
GCTATGTCGTCTCGTTCGCTT	AMSSRSL	25774
GCGGCTGGTCAGCAGTTTCCT	AAGQQFP	25604
GCGACTCCGCTTTGGCTTAAG	ATPLWLK	25487
GATTCGCATACTCCGCAGAGG	DSHTPQR	24964
CAGCCTCCTCGTTCGACGTCG	QPPRSTS	22462
CAGGCTACGCATCGTTCGCAT	QATHRSH	20473
CATAGGGCGGATATGCATTTT	HRADMHF	19965
ACGCGGGCTGGTCTGGATTTT	TRAGLDF	19930
CAGCGGCTGCCTCAGACGGCG	QRLPQTA	17573
CAGTCTAGTGTTCTGCGGCAT	QSSVLRH	17312
GCTGCTAAGACGCCTACGGAG	AAKTPTE	16774
		Etc >1,000,000 lines

The files were saved as *.txt and used for all graphical processing by the command_center.m script.

Scheme S3. (A) Example of multiset or set with multiple elements. (B) A multiset could be represented as a stacked bar. (C) Isolation of sub-sets with different copy numbers could be used to represent the multiset as 2D stacked-bar (D). (E) Comparison of two multisets **X** and **Y,** which consist of sets *x* and *y* and multiplicity functions *f* and *g*. (F) Venn diagram describing intersection, difference and union in sets and multisets. (G) Multiset-specific m-intersect and m-difference operators. Note that $\mathbf{X}_m \cap \mathbf{Y}$ and $\mathbf{Y}_m \cap \mathbf{X}$ are non-commutative and they are different from intersect $\mathbf{X} \cap \mathbf{Y}$. (H) Scatter plot of f vs. g multiplicity functions describes the abundances of elements in **X** in **Y**.



Data Visualization using Multisets. A result from the selection can be represented mathematically as a *multiset*. There are few standard visualization techniques for multisets. 1D-stacked bars describe both the number of unique sequences and their copy number in the multiset (Figure 2B) (21,22). To describe large multisets, the set elements could be grouped by their copy number and represented in 2D: the width of each segment illustrates the number of the unique sequences in this segment; its height represents the fraction of these sequences in the library (Figure 2D). Multisets can be compared using Venn diagrams. Figure 2E describes examples of two multisets, **X** and **Y,** and the results of intersection ($\mathbf{X} \cap \mathbf{Y}$) and difference ($\mathbf{X} \setminus \mathbf{Y}$) operators (Figure 2F). We define the multiset operations m-intersection $\mathbf{X}_m \cap \mathbf{Y}$, and m-difference $\mathbf{Y}_m \setminus \mathbf{X}$, (Figure 2G) in order to describe weighted contribution of common sequences. M-intersect is defined as $\mathbf{X}_m \cap \mathbf{Y} := \{(s,n) \mid (s,m) \in \mathbf{X} \cap \mathbf{Y} \wedge (s,n) \in \mathbf{X}\}$ (i.e., **X** m-intersect **Y** is equal to the multiset of elements (s,n) such that that s exists in **X** intersect **Y** and n is the count of element s in **X**). That is, m-intersect contains every unique element in the intersection of **X** and **Y** at multiplicities of the element's original count in **X**. We define the m-difference as the remainder: $\mathbf{X}_m \setminus \mathbf{Y} = \mathbf{X} - \mathbf{X}_m \cap \mathbf{Y}$. An intuitive tool for multiset comparison is a scatter plot, which describes pairwise differences in abundances of the individual elements (Figure 2H). These plots could be equipped with color gradients and quantification grids (akin to those used in flow-cytometry software).

Description of all scripts (included in Derda_Scripts.gz.rar)

Folder: Illumina Processing

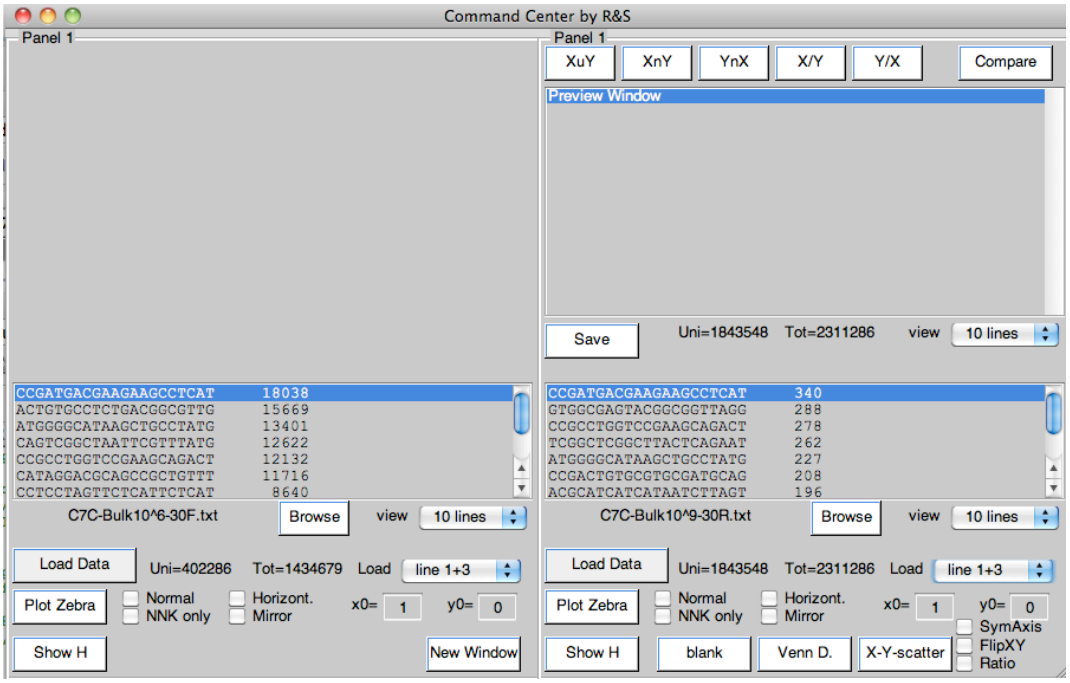
% This folder contains all script for processing FASTQ files according to workflow in **Scheme S1**. Scripts are started by **RunAllScripts.m** script. Name(s) of FASTQ files and directories in which they are stored have to be defined in the script:

line 15: name = {'Derda-sample2_NoIndex_L002_R1_001.fastq'};
Line 28: indir = 'rawfiles';

The file can be downloaded at http://www.chem.ualberta.ca/~derda/zebrapaper/

Once started, the script runs through the remaining 15 scripts and generates text-based status about the processing. Files will be placed in the newly created folders **LIB/** and **RAW/** located within the same folder as the original FASTQ file

Folder: Command Center
% This folder contains all scripts necessary for generating many figures in this manuscript (Figure 2B, 2C, 3A-C, 5C, 6A, 7B-D; Supporting Figures S3, S5A-D, S5E, S6A-D, S6E, S9D, S10A, S10C, S10E, and S11). Scripts are controlled by **command_center.m**. It starts an GUI, which can be used to load files, compare them, plot bars, dot plots, difference plots, generate m-intersects, m-unions and m-differences, and generate Venn diagrams for these plots. Click Browse to load sequence text files generated in **LIB/** folder, select Load option (usually line 1+3, which contains nucleotide sequence and copy number) and load data. Once loaded, data can be plotted or compared. The window will "turn off" the buttons while calculations are performed. Be patient. This script is a pre-beta version and it has not been fully tested or debugged. Screenshot of GUI:



Folder: MimoDB
This folder contains mimoDB.mat structure, which represents entire MimoDB 2.0 database. Files describing libraries can be extracted from it using a **mimodb_convert.m** script. See commentary inside the script. Extracted txt files can be used by **command_center.m** script to run comparison between literature and other data. Examples of libraries **C7C-literature.txt**, **PhD7-literature.txt** and **PhD12-literature.txt** are included.
Folder:MiscScrips
This folder contains MatLab scripts named *makeFigureX.mat*, where X is the number of specific figure. Each script is self-sufficient. Running the script generates the data in specific figure. Note that in most cases, the data has to be downloaded from http://www.chem.ualberta.ca/~derda/zebrapaper/

% All scripts were written by Ratmir Derda; any use of these scripts should be acknowledged by reference to this paper. Thank you!

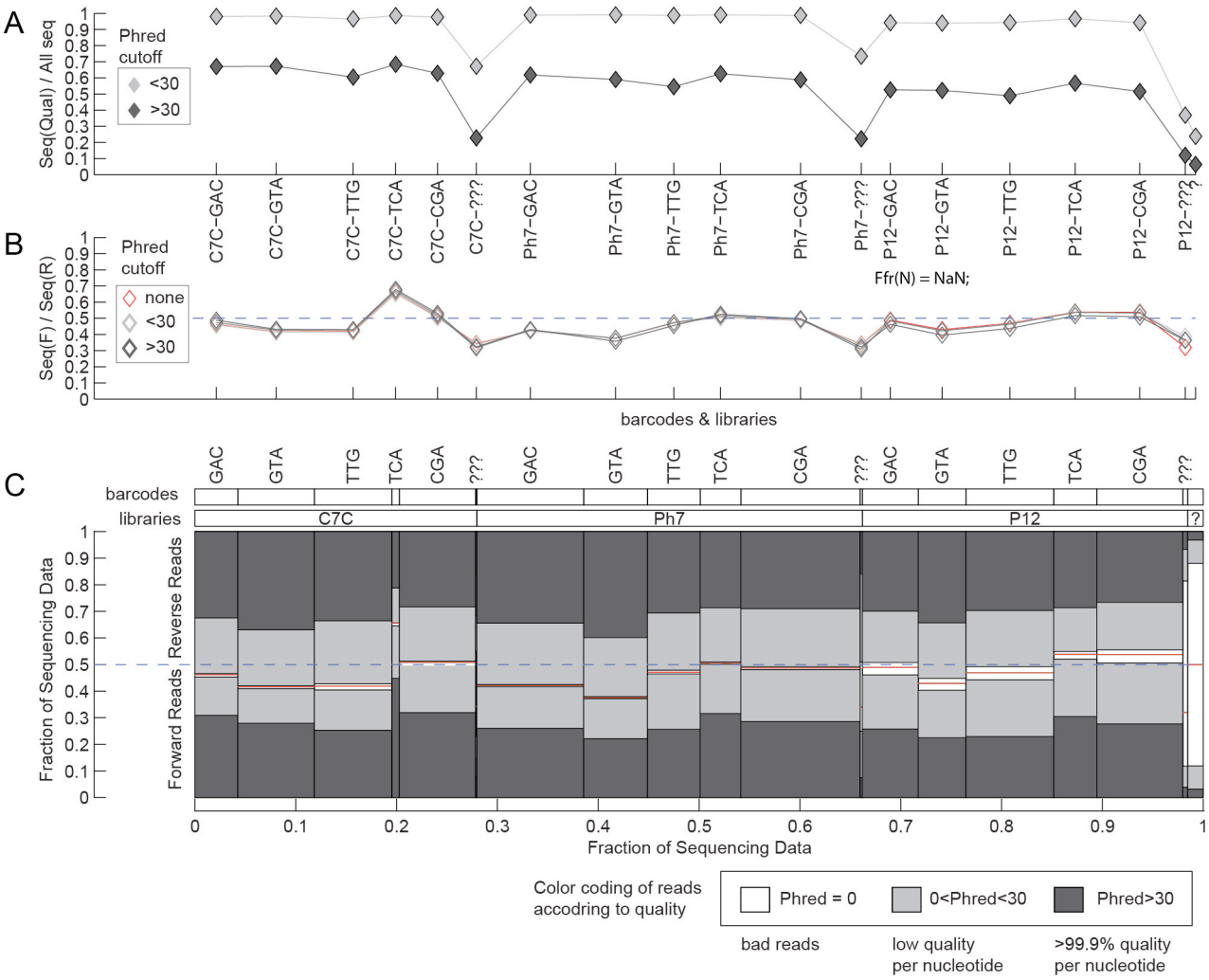


Figure S1. One Illumina sequencing run was used to analyze 3 different libraries and 5 experiments within each library. Experiments were identified using barcodes; libraries were identified by sequence structure. (A) We processed the data using three different cutoffs. Only Phred>30 sequences were used in this paper. The fraction of Phred>0, >13 and >30 sequences in each experiment was consistent. Sequences in which barcodes were damaged (labeled by "???") had significantly lower quality. Damaged barcodes are sequences that do not correspond to any sequences in the original list of barcode sequences. (B) Each experiment contained forward and reverse reads. Their ratio was skewed in reads with damaged barcodes (i.e. low quality). (C) Overall view of the library. Each rectangle represents an experiment. The area of each rectangle is proportional to the fraction of this experiment in the overall pool of sequences. Color represents sequences of certain quality. For example, the first vertical stacked bar represents C7C library tagged by GAC-barcode. Sequences constitute ~4% of the overall sequence space. Ratio of forward sequences is ~45%. Reads with Phred>13 and Phred>30 cutoffs constitute ~80% and ~60% respectively. This graph shows that ~3% of the sequences could not be mapped to any library or any barcode. Majority of those sequences are bad reads (i.e. they contain at least one unknown nucleotide with Phred=0).

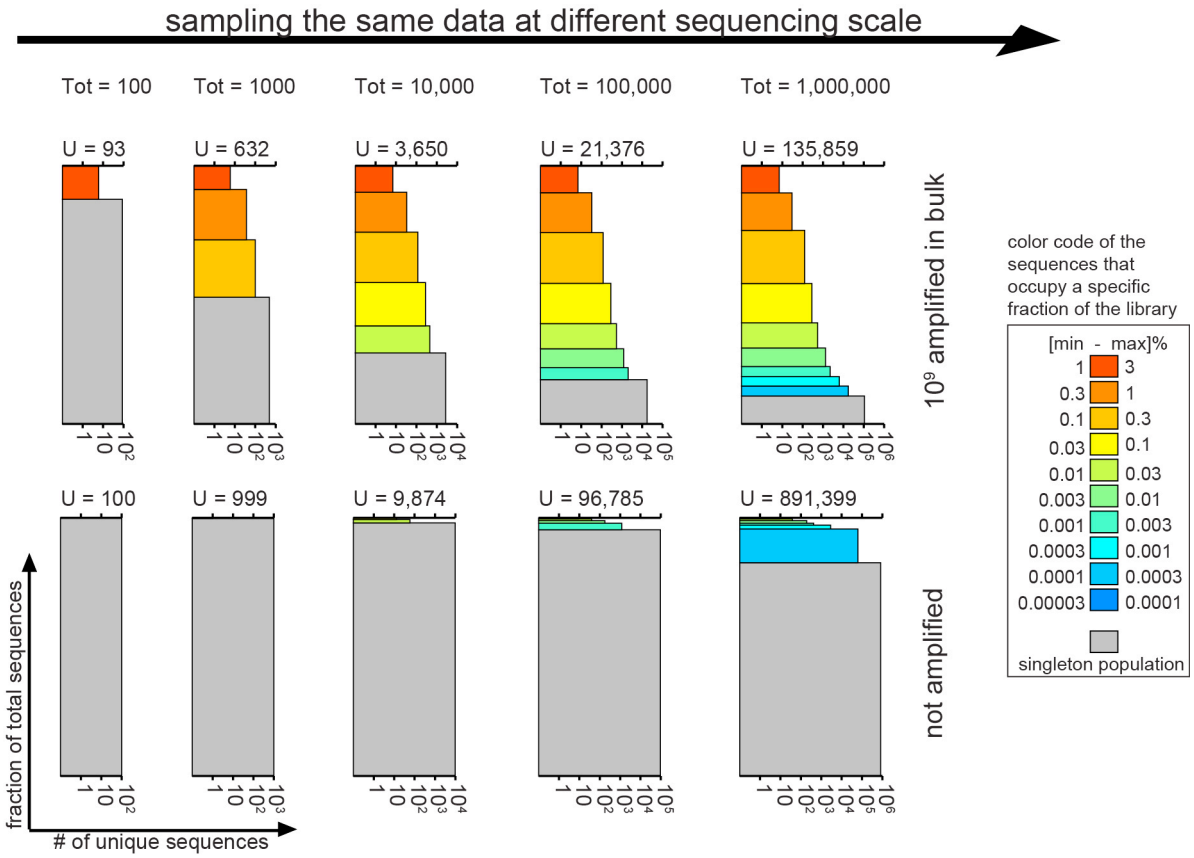


Figure S2. This figure displays how the same sequencing data looks at lower sequencing resolution (tot stands for "total number of reads). The figure is generated by random sampling of Illumina data. Naïve ("not amplified") library is practically "invisible" to sequencing below 10,000 total reads (<100 reads have copy number 2-3, the rest are singleton reads). Amplified library could be reliably analyzed with depth of sequencing of 10,000 to 100,000 reads. In 1000 total reads, one could see a few hundred "parasites" with copy number 2-3 (orange). There are only 10 sequences with copy number above 3-10 (orange-red segments). Observations, thus, do not have high confidence. Sequencing the amplified library at 100-reads scale (typical Sanger sequencing scale) could uncover only a few "parasites" with unreliable confidence (copy number of 2).

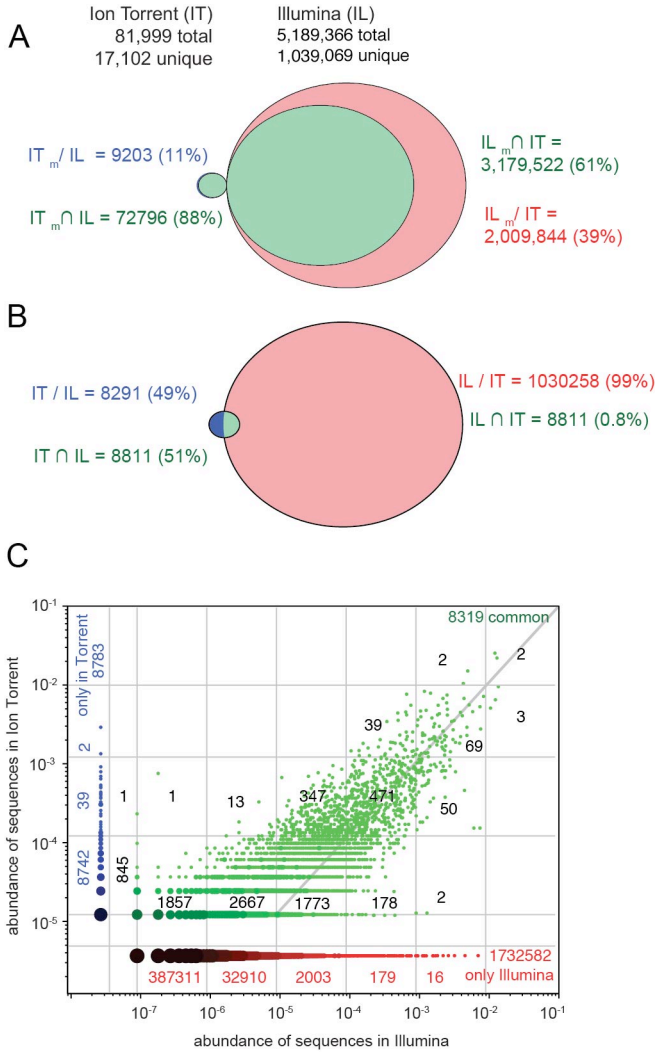


Figure S3. (A) Venn diagram describing multiset-specific m-intersect and m-difference between Ph.D.-7 amplified library sequenced by Ion Torrent (**IT**) and Illumina (**IL**). Over 60% of sequences found in the Illumina multiset are also present in the Ion Torrent multiset. The m-intersect for **IL** $_{m}\cap$ **IT** contains all elements within **IL** that are also found in **IT**. Similarly **IT** $_{m}\cap$ **IL** contains 88% of sequences. (B) The set **IT** ∩ **IL** and **IT** ∩ **IL** as it would be constructed without the consideration of frequency of each sequence found in both **IT** and **IL** sets. (A) and (B) are drawn to scale in relation to the size of each set. (C) Scatter plot describing Ph.D.-7 amplified library sequenced by Ion Torrent (**IT**) and Illumina (**IL**). Each dot is a unique sequence; multiple data at the same (x,y) coordinate are bigger, darker dots (see legend). Green data describe m-intersect, while blue and red describe m-difference population (data unique to **IL** or **IT**). Parasitic sequences are found in both sequencing methods. The **IL** and **IT** sets are in good agreement as sequences concentrate around the identity line (1:1 line drawn through the center of the scatter plot). Numbers represent the number of data points within each cell of the rectangular grid. The images were generated using *command_center.mat*

Re-Analysis of the biological replica using edgeR (version 3.2.3).

Written and complied by Andrea Rau in R (version 3.0.1).

The sections **S1-S5** and Figures S4-S7 on pages S8-S14 presents the full source code used to obtain results for the re-analysis of the phage biological replica (BR) amplified from 10^8 PFU. Specifically, the Bioconductor package edgeR is used to fit a negative binomial model to each sequence and to implement an exact test to identify enriched sequences in the amplified library as compared to the naive library.

Section S1: Load packages, data, and annotation

We begin by loading the necessary packages for the data analysis, as well as the data themselves, which are appropriately formatted for the analysis, and the sequences found to be significantly enriched from the t-test based on normal distribution (Figure 5D). Note that the main package to be used for the differential analysis is edgeR; the remaining packages are tools to assist in visualization.

```
> ## Set working directory, load libraries
> rm(list = ls())
> setwd("C:/Users/arau/Desktop/NAR")
> library(edgeR)
> library(VennDiagram)
> library(gplots)
> library(RColorBrewer)
> library(statmod)
> library(SweaveListingUtils)
> ## Read full data and format
> full.dat <- read.table("10^8-all.txt", fill=TRUE)
> colnames(full.dat) <- c("sequence", "ID",
+ as.matrix(full.dat[1,which(is.na(full.dat[1,]) != TRUE)]))
> full.dat <- full.dat[-1,]
> ## Read previous significant results
> prev.sig <- read.table("volcano.txt")[,-3]
> colnames(prev.sig) <- c("sequence", "ID")
> prev.sig[,3] <- paste(prev.sig$ID, prev.sig$sequence, sep=".")
> colnames(prev.sig)[3] <- "name"
> ## Table of counts alone: N1-N10 vs B1.5-B5
> dat <- full.dat[,-c(1,2)]
> dat <- matrix(as.numeric(as.matrix(dat)), ncol=19)
> rownames(dat) <- paste(full.dat[,2], full.dat[,1], sep=".")
> colnames(dat) <- colnames(full.dat)[-c(1:2)]
> dat <- dat[,c(1:10, 15:19)]
> conds <- factor(c(rep("N", 10), rep("B", 5)), levels = c("N", "B"))
```

For the remaining analysis, we remove sequences that have zero counts or only one count ("single-tons"). This removes a large portion of the sequences to be tested (from 176,158 to 10,159), which will make a significant difference for the necessary correction to be made to p-values to account for multiple testing.

```
> ## Remove sequences with 0 or 1 counts in all samples
> dim(dat) ## 176,158 x 15
```

```
[1] 176158 15
```

```
> dat <- dat[-which(rowSums(dat) <= 1),]
> dim(dat) ## 10,159 x 15
```

```
[1] 10159 15
```

Section S2: Quick exploratory analysis

In this section, we perform some quick exploratory analysis of the data (see Figure 1).

```
> ## Quick exploratory analysis
> hist(log(as.numeric(dat)+1), col="grey", breaks=20, main = "",
+ xlab="Log(count+1)")
> quantile(dat)
0% 25% 50% 75% 100%
0 0 0 0 281
> postscript("exploratory_analysis.ps", horizontal=FALSE, width=10, height=4)
> par(mfrow = c(1,2), mar = c(4,4,2,2))
> boxplot(log(dat+1), ylab = "Log(count + 1)")
> barplot(colSums(dat), col = c(rep("grey50", 10), rep("grey90", 5)),
+ ylab = "Library sizes", main="")
> barplot(apply(dat, 2, max), add=TRUE, col="black")
> dev.off()

windows
2
```

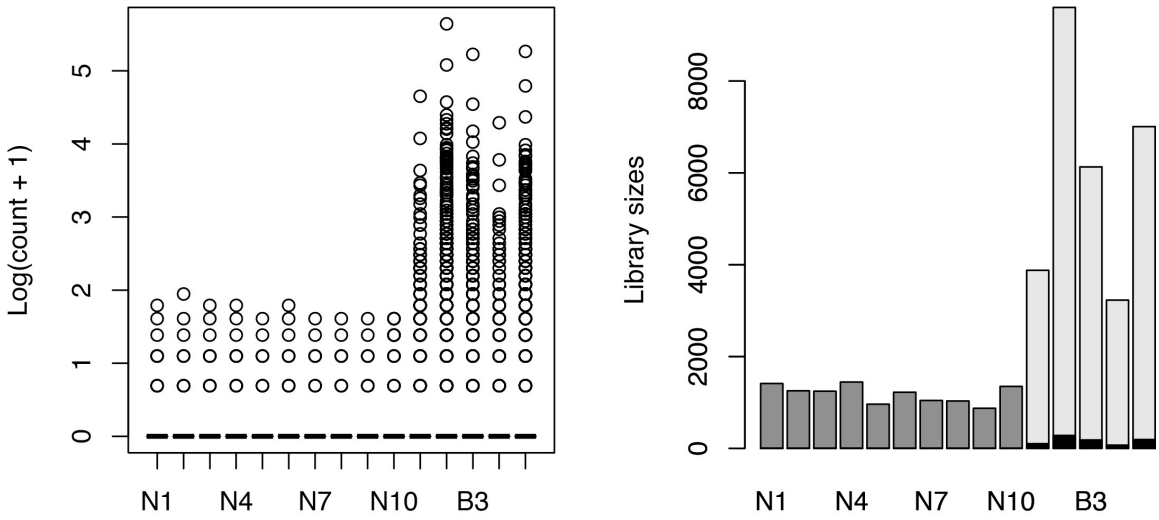


Figure S4: (left) Boxplots of $\log_2$(counts + 1) for each sample. (right) Barplots of library sizes for each sample, with largest single sequence overlaid in black. N1-N10 are re-sequencing replica of the naïve library. B1-B5 are biological replica (re-amplification).

Section S3: EdgeR analysis (with TMM normalization)

In this section, we make use of the Bioconductor package edgeR [2, 3] to fit a negative binomial model to each sequence and to implement an exact test for differential expression. The novelty of the edgeR package is that it implements an empirical Bayes method to estimate sequence-specific biological variation (i.e., overdispersion). In particular, rather than using an overall, common estimate for dispersion parameters or a per-sequence dispersion parameter estimate, edgeR calculates moderated dispersion parameter estimates that are reliable even for small samples. These moderated dispersion parameters allow information to be shared between sequences while still maintaining sequence-specific dispersion estimates by squeezing tagwise dispersions toward the common dispersion. Within the edgeR framework, we make use of the Trimmed Mean of M-values (TMM) normalization [4] to account for differences in library composition among the different libraries. This technique find a set of scaling factors for library sizes (yielding so-called effective library sizes) that minimize log-fold changes between the samples for most sequences.

After running edgeR, we find a total of 606 enriched sequences at signficance level _ = 0.05 if no correction for

multiple testing is included (incorrect!), or 218 enriched sequences after a Benjamini- Hochberg [1] correction to control the false discovery rate (FDR) at $\alpha = 0.05$.

```
> design <- model.matrix(~0 + conds)
> colnames(design) <- c("N", "B")
> DGE <- DGEList(counts=dat, group=conds)
> DGE <- calcNormFactors(DGE)
> ## Square-root of the common dispersion gives the biological coefficient of variation
> ## BCV = coefficient of variation with with abundance of sequence varies between replicates
> DGE <- estimateCommonDisp(DGE, verbose=TRUE)
Disp = 0.15144 , BCV = 0.3891

> DGE <- estimateTagwiseDisp(DGE)
> postscript("BCV.ps", horizontal=FALSE)
> plotBCV(DGE, cex=0.4)
> dev.off()
windows
2
> et <- exactTest(DGE)
> de.none <- decideTestsDGE(et, adjust="none", p.value=0.05)
> de.BH <- decideTestsDGE(et, adjust="BH", p.value=0.05)
> de.none.names <- rownames(dat)[which(de.none == 1)] ## 606
> de.BH.names <- rownames(dat)[which(de.BH == 1)] ## 219
```

We next compare the list of enriched sequences (using both uncorrected p-values and adjusted p-values) to the previously identified list using one-sided t-tests; see Figure S5.

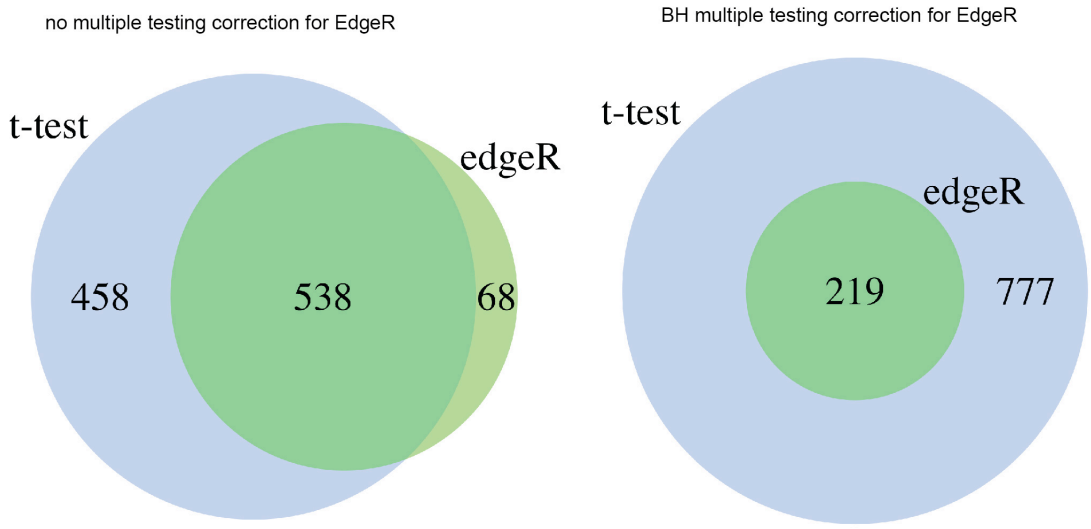


Figure S5: (left) Venn diagram of previously identified significant sequences (t-test) and sequences identified by edgeR without multiple testing correction. (right) Venn diagram of previously identified significant sequences (t-test) and sequences identified by edgeR after BH multiple testing correction.

```
> ## Venn diagrams with edgeR (no multiple testing correction) and t-test results
> venn.plot <- venn.diagram(
+ x = list(
+ "t-test" = prev.sig[,3],
+ "edgeR" = de.none.names
+ ),
```

```
+ filename = "Venn_t-test_edgeR_none.tiff",
+ scaled = TRUE, fontfamily="serif",
+ cex = 2.5, margin = 0.05,
+ fill = c("cornflowerblue", "green"),
+ col = "transparent", cat.dist = c(0.04, 0.04),
+ cat.cex = 2.5, height = 3000, width = 3000,
+ sub = "(no multiple testing correction for edgeR)"
+ );
> ## Venn diagrams with edgeR (BH correction) and t-test results
> venn.plot <- venn.diagram(
+ x = list(
+ "t-test" = prev.sig[,3],
+ "edgeR" = de.BH.names
+ ),
+ filename = "Venn_t-test_edgeR_BH.tiff",
+ scaled = TRUE, fontfamily="serif",
+ cex = 2.5, margin = 0.05,
+ fill = c("cornflowerblue", "green"),
+ col = "transparent", cat.dist = c(0.04, 0.04),
+ cat.cex = 2.5, height = 3000, width = 3000,
+ sub = "(BH multiple testing correction for edgeR)"
+ );
```

Finally, we can visualize the dispersion parameter estimates provided by edgeR in **Figure S6**

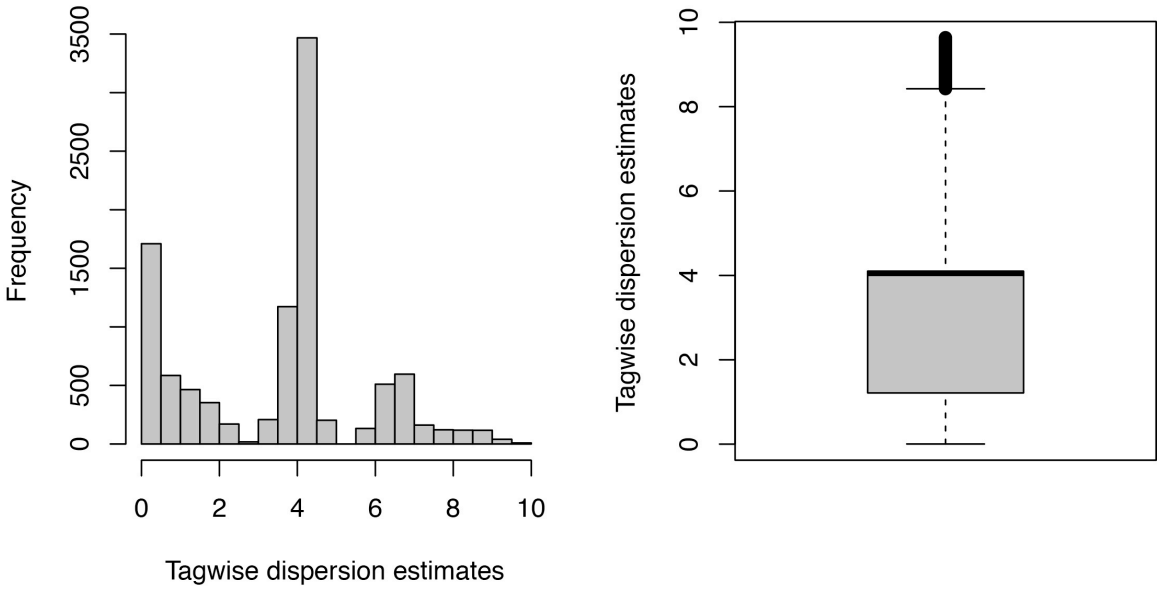


Figure S6: (left) Histogram and (right) boxplot of tagwise dispersion parameter estimates provided by edgeR.

```
> ## Dispersion parameter estimates
> postscript("dispersion.ps", horizontal=FALSE, width=10, height=4)
> par(mfrow=c(1,2), mar=c(4,4,2,2))
> hist(DGE$tagwise.dispersion, xlab = "Tagwise dispersion estimates",
+ main="", col="grey", breaks=25)
> boxplot(DGE$tagwise.dispersion, col="grey", ylab = "Tagwise dispersion estimates")
> dev.off()
windows
2
> write.table(DGE$tagwise.dispersion, "dispersion_estimates.txt", col.names=FALSE,
```

+ row.names=FALSE)

Section S4: Additional visualizations: volcano plots and smear plots

In this section we provide some additional visualizations for the differential analysis results, including volcano plots, smear plots, and heatmaps. For these visualizations, we make use of the results from the differential analysis above after correction for multiple testing using the Benjamini-Hochberg control of FDR. Volcano and smear plots are shown in **Figure S7.**

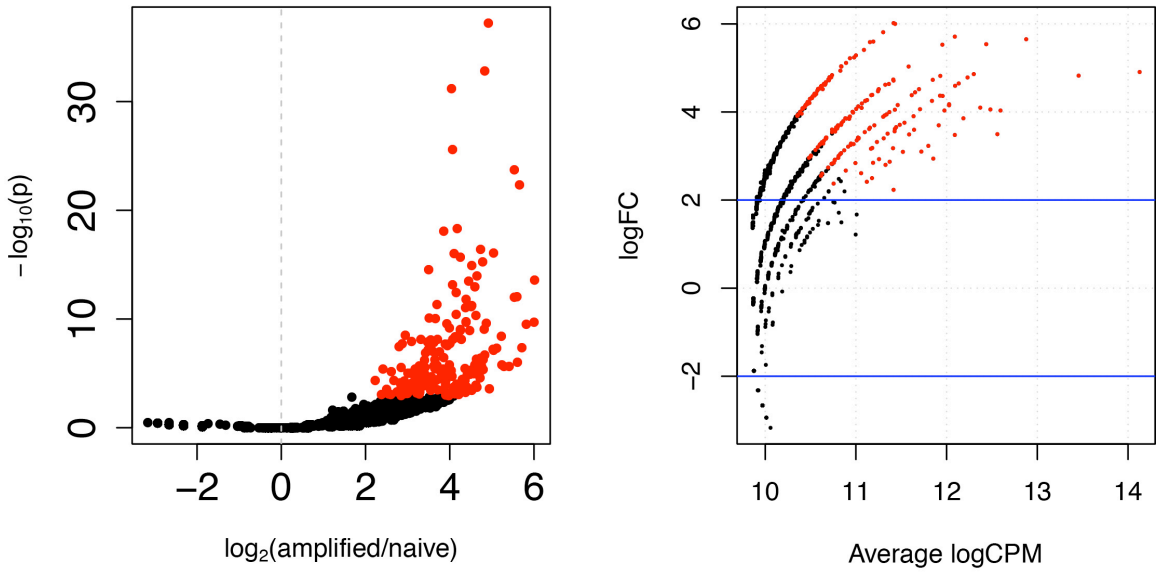


Figure S7: (left) Volcano plot and (right) smear plot (analogous to an MA plot for microarrays) following results from the differential analysis. Points in red are those that have been identified as enriched after correction for multiple testing. Blue lines in the smear plot indicate 4-fold changes.

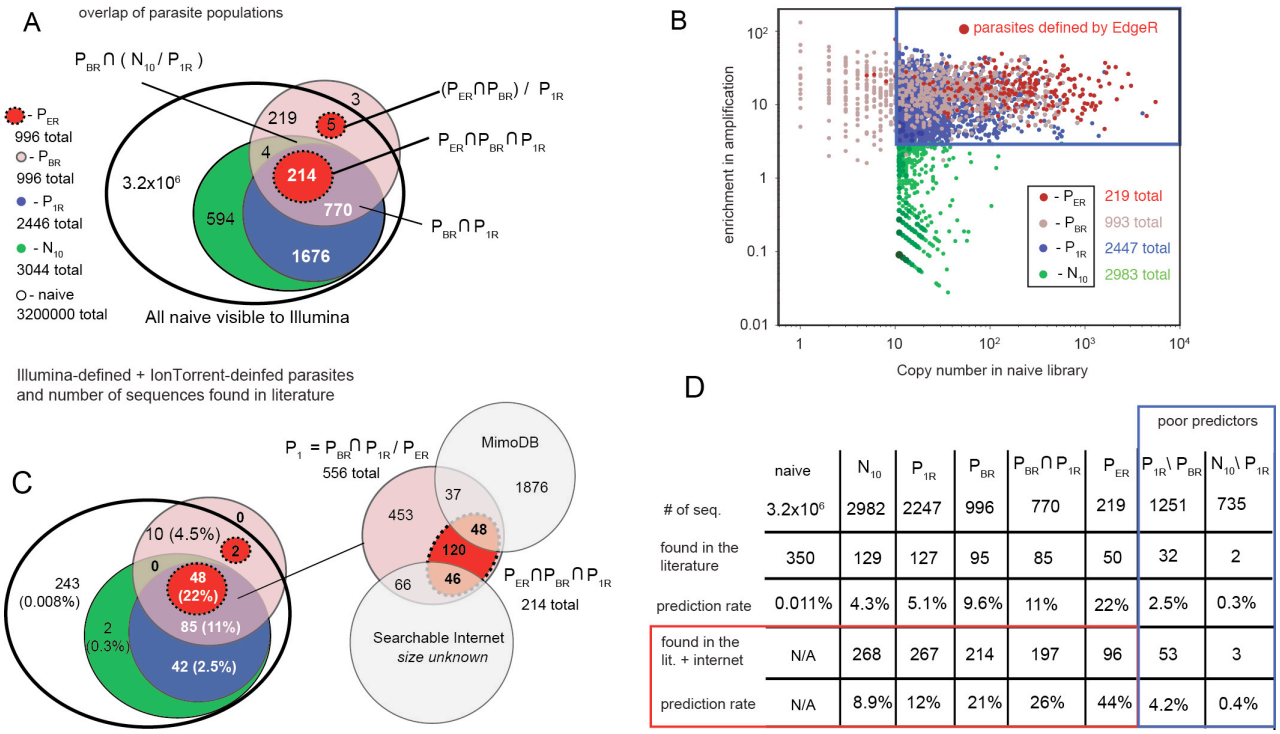
```
> DE <- which(de.BH==1)
> NDE <- which(de.BH==0)
> ## Volcano plots
> postscript("Volcano_MA.ps", width=10, height=4, horizontal=FALSE)
> par(mfrow = c(1,2), mar=c(4,4,2,2))
> plot(0,0, col="white", cex.main=2.5, cex.axis=1.5,
+ ylim=c(min(-log10(et$table$PValue)),max(-log10(et$table$PValue))),
+ xlim=c(min(et$table$logFC), max(et$table$logFC)),
+ xlab=expression(paste(log[2], "(amplified/naive)")),
+ ylab=expression(paste(-log[10], "(p)")))
> points(et$table$logFC[NDE], -log10(et$table$PValue[NDE]),
+ col="black", pch=20)
> points(et$table$logFC[DE], -log10(et$table$PValue[DE]),
+ col="red", pch=20)
> abline(v=0, lty=2, col="grey")
> mtext(side = 1, outer = TRUE, "Log fold change", line = 1, cex=1.5)
6
> mtext(side = 2, outer = TRUE, "Log10 p-value", line = 1, cex=1.5)
> ## Tagwise log-fold changes against log-cpm (analagous to MA plot)
> ## Blue lines indicate 4-fold changes
> plotSmear(et, de.tags=rownames(dat)[DE])
> abline(h = c(-2,2), col="blue")
> dev.off()
```

windows
2
>

Section S5: Session Info

> ## Save sessionInfo()
> sessionInfo()
R version 3.0.1 (2013-05-16)
Platform: x86_64-w64-mingw32/x64 (64-bit)
locale:
[1] LC_COLLATE=French_France.1252
[2] LC_CTYPE=French_France.1252
[3] LC_MONETARY=French_France.1252
[4] LC_NUMERIC=C
[5] LC_TIME=French_France.1252

attached base packages:
[1] grid parallel stats graphics grDevices
[6] utils datasets methods base
other attached packages:
[1] SweaveListingUtils_0.6.1 startupmsg_0.8
[3] statmod_1.4.17 RColorBrewer_1.0-5
[5] gplots_2.11.0.1 MASS_7.3-26
[7] KernSmooth_2.23-10 caTools_1.14
[9] gdata_2.12.0.2 gtools_2.7.1
[11] VennDiagram_1.6.0 DESeq_1.12.0
[13] lattice_0.20-15 locfit_1.5-9.1
[15] Biobase_2.20.0 BiocGenerics_0.6.0
[17] edgeR_3.2.3 limma_3.16.5
loaded via a namespace (and not attached):
[1] annotate_1.38.0 AnnotationDbi_1.22.6
[3] bitops_1.0-5 DBI_0.2-7
[5] genefilter_1.42.0 geneplotter_1.38.0
[7] IRanges_1.18.1 RSQLite_0.11.4
[9] splines_3.0.1 stats4_3.0.1
[11] survival_2.37-4 tools_3.0.1
[13] XML_3.98-1.1 xtable_1.7-1

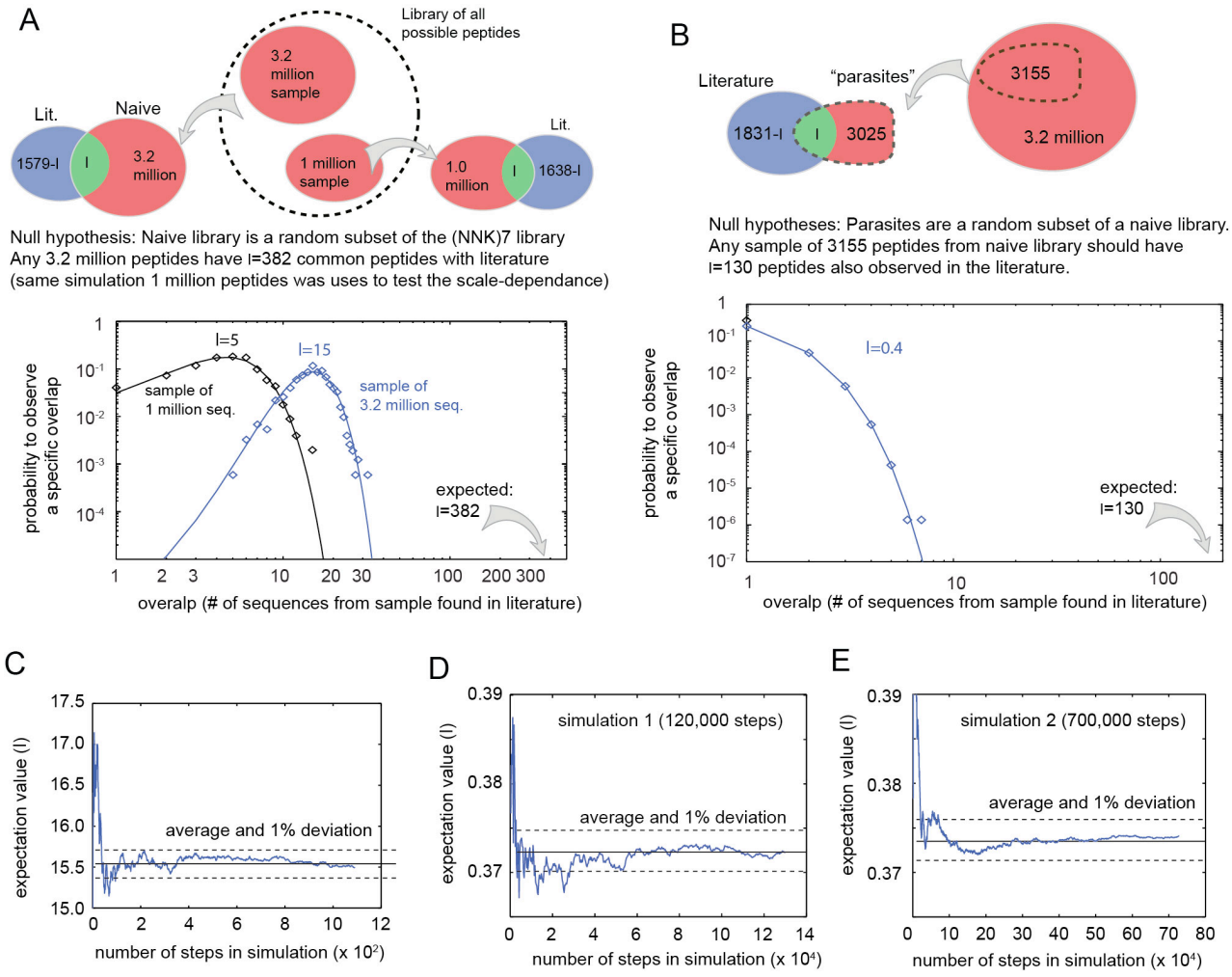


D

	naive	N_{10}	P_{1R}	P_{BR}	$P_{BR} \cap P_{1R}$	P_{ER}	poor predictors: $P_{1R} \setminus P_{BR}$	poor predictors: $N_{10} \setminus P_{1R}$
# of seq.	3.2×10^6	2982	2247	996	770	219	1251	735
found in the literature	350	129	127	95	85	50	32	2
prediction rate	0.011%	4.3%	5.1%	9.6%	11%	22%	2.5%	0.3%
found in the lit. + internet	N/A	268	267	214	197	96	53	3
prediction rate	N/A	8.9%	12%	21%	26%	44%	4.2%	0.4%

Figure S8. Comparison of parasites obtained through one-sided t-statistics (P_{BR} population, pink) and negative binomial model with exact test and correction for multiple comparison (EdgeR, P_{ER} population, red). (A) All 219 P_{ER} hits reside inside P_{BR} population and 214 of them reside in the intersect of all previously defined populations (compare to Figure 5). (B) Scatter plot shows that P_{ER} population overlaps significantly with P_{BR} population at high copy numbers but the overlap diminishes at low copy numbers. (C) Comparison to the literature demonstrates that P_{ER} population has higher rate of identification of the literature sequences (48/214 = 22%) when compared to $P_{BR} \cap P_{1R}$ (85/770=11%). On the other hand 85-48 = 37 literature hits are not accounted for in the P_{ER} population. Overlap with another literature source (searchable internet) reveals the same trend. The P_{ER} population has higher prediction rate (46/214 = 20%) but it also fails to account for ~66 hits, which were found in the broader $P_{BR} \cap P_{1R}$ population. (D) Summary of the parasite populations and their overlap with MimoDB or Internet-search.

Overall, we observed that no single definition of parasite is optimal. The larger populations $P_{B,}$ P_{1R}, N_{10} or their combinations, which were defined using relaxed statistical approaches, capture a larger number of putative parasites in the literature (less false negatives). These populations also contained a large number of false positives. Focused P_{BR} population has lower false positive rate, but also higher false negative rate (see main text for examples of false negatives). Unfortunately, the identity of the "true parasite population" is not known; hence, the false-positive and false negative rates can be defined only qualitatively.



F General sampling problem (see illustration on the right): Given a random sample of N peptides from a random peptide library encoded by (NNK)7 codons, how many peptides are expected to be common between this N-sample and a literature?

N	3,200,000	1,000,000	500,000	100,000	3,000	770
exp. value (I)	15	5	2.5	0.5	0.015	0.004
maximum in in X trials	35 X = 1000	18 X = 5000	10 X = 5000	5 X = 5000	1 X = 10000	1 X = 20000

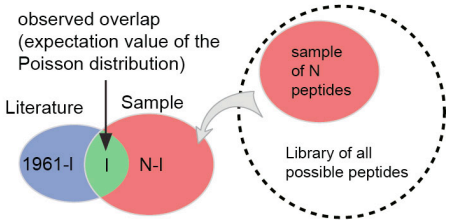


Figure S9. (A) Overlap between literature and naïve library (382 common sequences) cannot occur at random. To prove it, we formulated null hypothesis and sampled 3.2 million peptides from a random peptide library based on NNK codons. In 5000 trials, the average number of common sequences was 15. The extrapolated probability to observe 382 common sequences was $p<<e^{-382}$. The same hypothesis for non-singleton population of the naïve library (ca. 10^6 seq.; ~220 literature hits): a random sample has 5 hits. The probability to observe 220 hits was $p << e^{-200}$ (B) Testing the significance of the overlap between parasite population P_{10} and literature (130 hits). In 10 million random trials, we selected 3155 random sub-sets from naïve library. Average overlap was 0.4 sequences. The probability to observe 130 common sequences was $p<<10^{-130}$. (C) Convergence of the bootstrapping simulation used to generate blue curve in (A). (D-E) convergence of simulations used to generate blue curve in (B). Running two separate simulations with 120,000 or 700,000 steps yields similar results. (E) The null hypothesis tested in (A) could be formulated more broadly for sequencing results that contain <1 million reads. For algorithm see *makeFigureS4.mat*

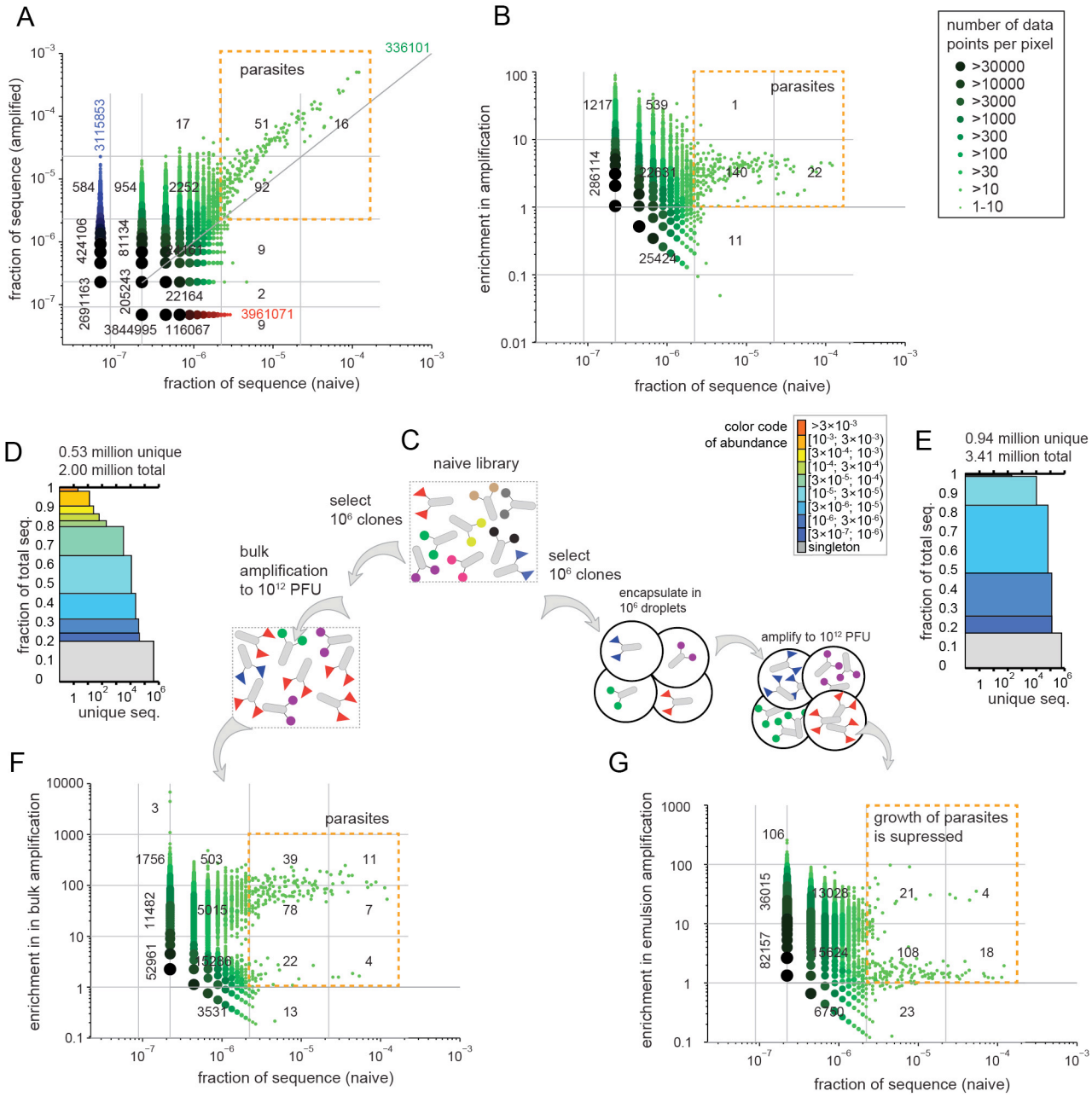


Figure S10. (A) Scatter plot describing naïve (**N**) and amplified (**A**) Ph.D.-C7C library. Each dot is a unique sequence; multiple data at the same (x,y) coordinate are bigger, darker dots (see legend). Numbers represent the number of data points within each cell of the rectangular grid. Green data describe m-intersect, while blue and red describe m-difference population (data unique to **N** or **A**). (B) We calculated sequence enrichment as f_{amp}/f_{naive} and plotted it vs. f_{naive}, where f is a fraction of sequences (copy number normalized by total number of reads). (c) Schematic for the amplification of 10^6 clones taken from Ph.D.-C7C naïve library. Amplification was performed either in bulk or emulsion (as described in Conditions 1 and 3 in the Methods section). (D-F) Amplification in bulk shows significant enrichment of parasitic sequences compared to amplification in emulsion sequences (>100 fold increase respectively from the original fraction in the naïve library). (E-F) Amplification in emulsion yields uniform library without high-copy-number reads. The enrichments of parasitic sequences is suppressed in emulsion amplification.

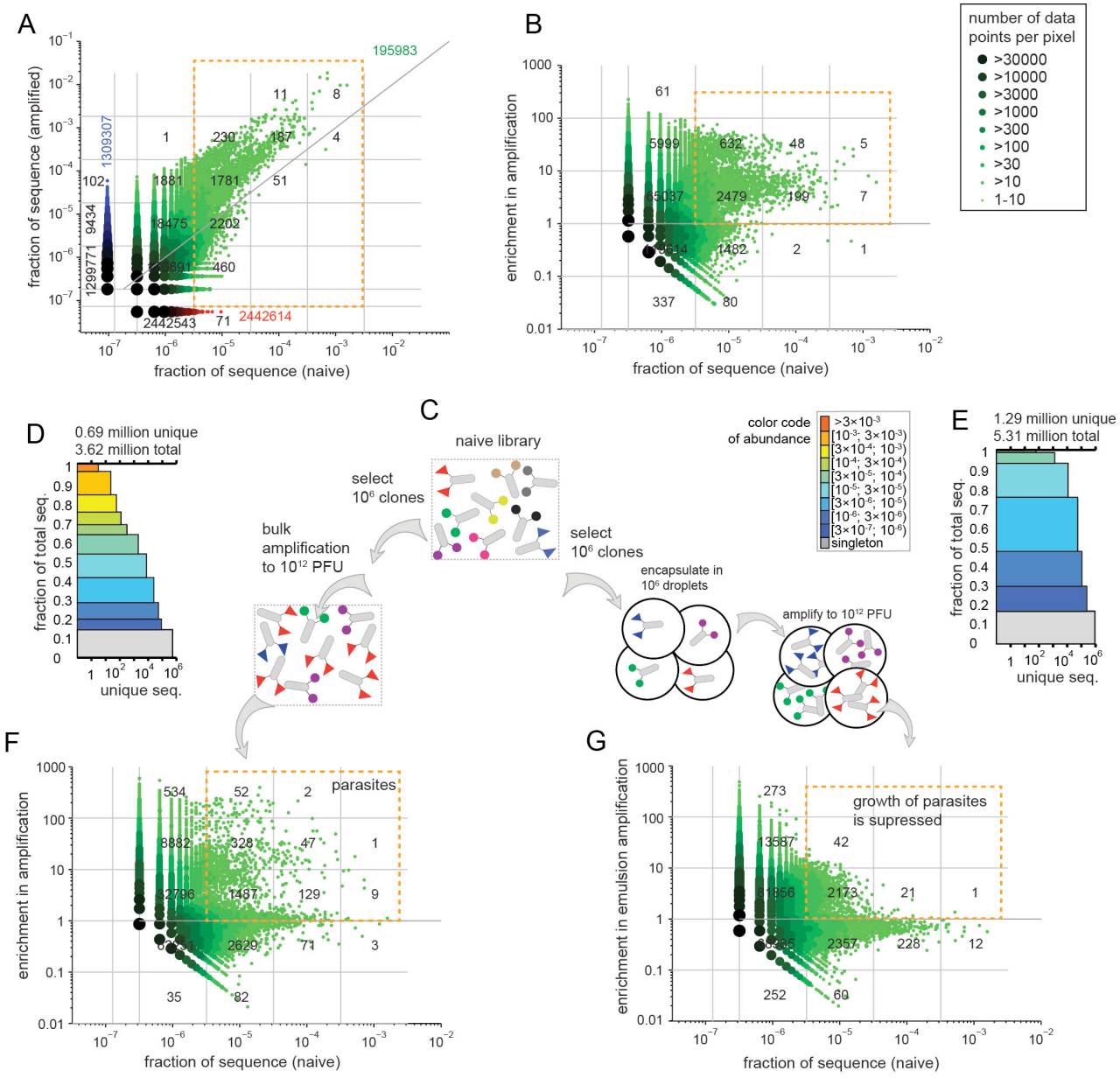


Figure S11. (A) Scatter plot describing naïve (**N**) and amplified (**A**) Ph.D.-12 library. Each dot is a unique sequence; multiple data at the same (x,y) coordinate are bigger, darker dots (see legend). Numbers represent the number of data points within each cell of the rectangular grid. Green data describe m-intersect, while blue and red describe m-difference population (data unique to **N** or **A**). (B) We calculated sequence enrichment as f_{amp}/f_{naive} and plotted it vs. f_{naive}, where f is a fraction of sequences (copy number normalized by total number of reads). (C) Schematic for the amplification of 10^6 clones taken from Ph.D.-12 naïve library. Amplification was performed either in bulk or emulsion (as described in Conditions 1 and 3 in the Methods section). (D-F) Amplification in bulk shows significant enrichment of parasitic sequences compared to amplification in emulsion sequences (>100 fold increase respectively from the original fraction in the naïve library). (E-F) Amplification in emulsion yields uniform library without high-copy-number reads. The enrichment of parasitic sequences is suppressed in emulsion amplification.

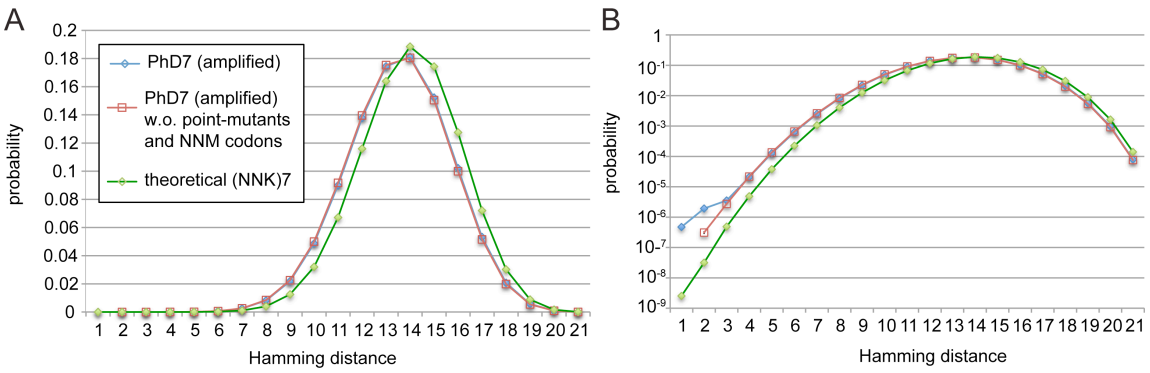


Figure S12. (A) Distribution of Hamming distances between nucleotides in the library (B) is the log scale plot of (A). The data was based on library amplified using condition 1 (10^6 PFU amplified to 10^{12} PFU in bulk). Theoretical library of 10^6 $(NNK)_7$ nucleotides was generated *in silico*. For each library, we selected 1000 random members and calculated distances to the members in the rest of the library (10^9 distances total). Calculation with 100 or 10000 randomly selected members yielded the same distribution. See *makefigureS7.m* for specific algorithm and http://www.chem.ualberta.ca/~derda/parasitepaper/ for raw files.

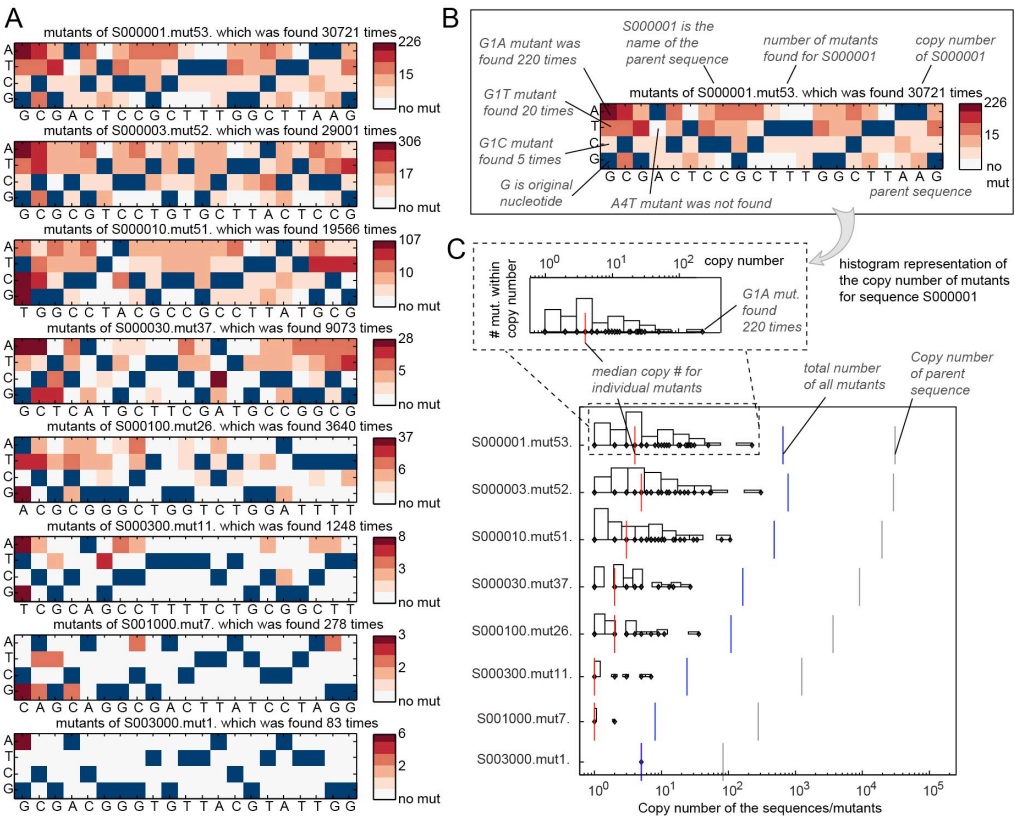
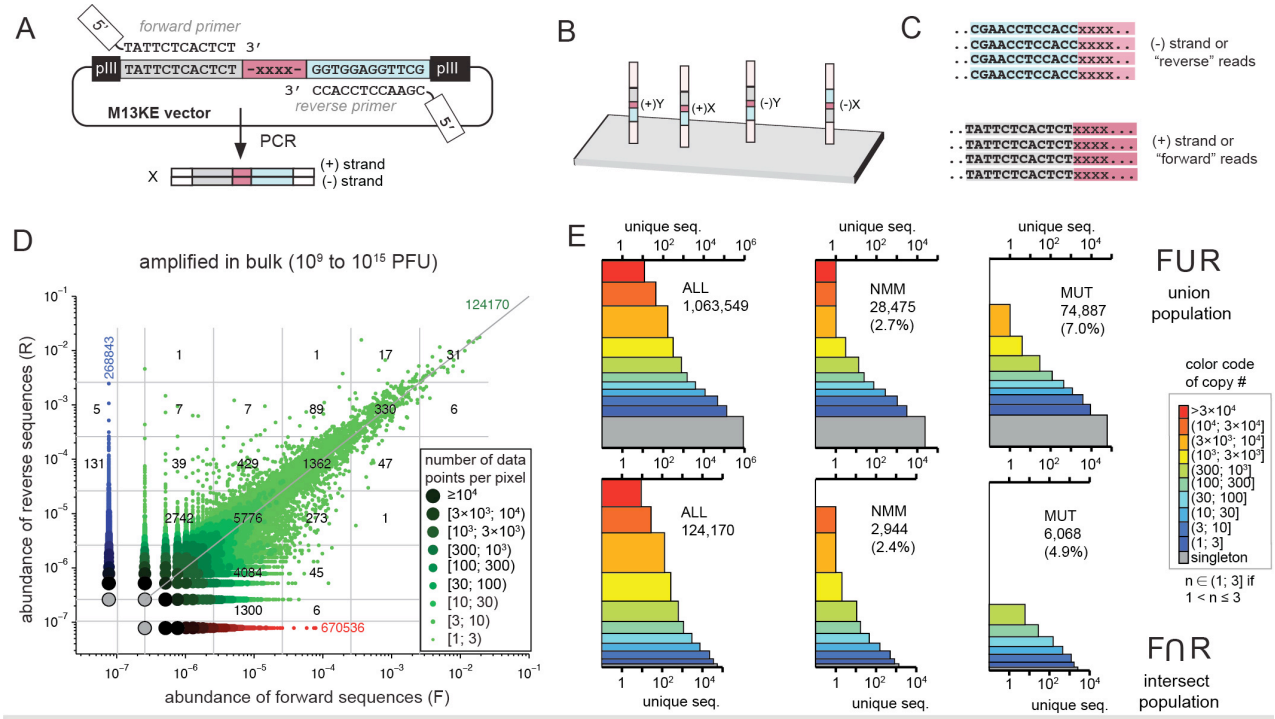


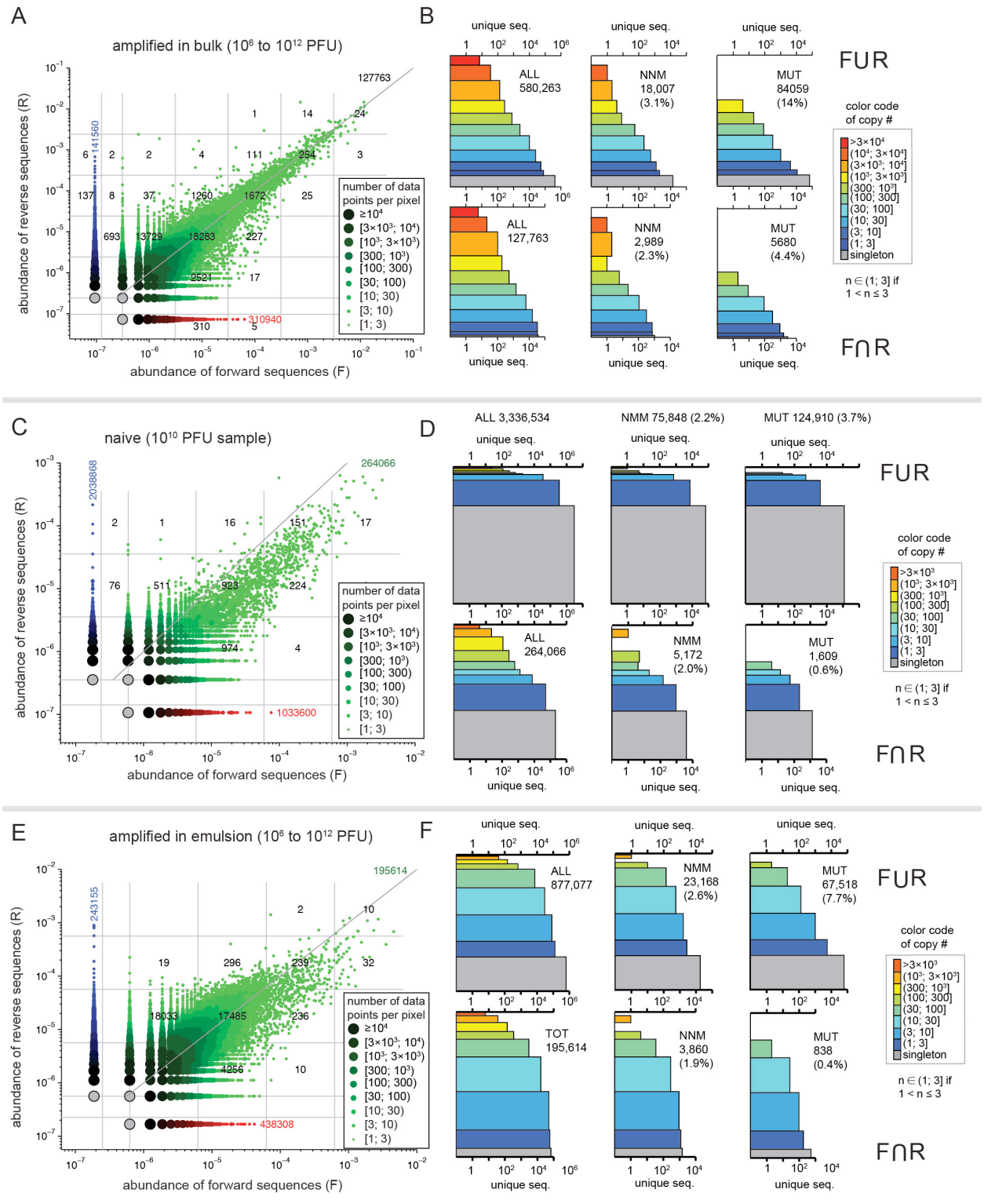
Figure S13. Analysis of point mutants in deep sequencing of PhD-7 library. (A) In heatmap, blue denotes the original nucleotide and shades of red denote the mutants. The intensity of red color is proportional to the number of times each mutant was found.
(B) The results show that point mutations are ubiquitous throughout (NNK)7 seqeunce. They occur both in N and in K positions with similar frequencies. (C) The number of unique mutants scales with the abundance of the parent sequence. Abundance of mutants: most of them are clustered around a median frequency value (red line). Total number of mutants (blue line) was on average 20 times less than the abundance of parent sequence (grey line). See *makefigureS8.m* for specific algorithm and http://www.chem.ualberta.ca/~derda/parasitepaper/ for raw files.



F

sequence number	nucleotide	peptide	copy#	# of mutants found	copy # of mut.	NMM-key	MUT-key
1	GGGAAGCCTATGCCTCCGATG	GKPMPPM	5548	38	136	.....	.
2	CAGCCTTGGCCGACGAGTATT	QPWPTSI	4478	31	167	.....	.
3	TGGCCTACGCCGCCTTATGCG	WPTPPYA	4044	25	168	.....	.
4	CATGCTCTGGGTCCGTCTTCG	HALGPSS	3066	33	132	.....	.
...							
14	GCGACGACTGTTCCAGCTTCG	ATTVPAS	1602	25	100	....A..	.
...							
185	CTACCTTCATATCATGTGCCT	LPSYHVP	248	8	18	A.A....	.
...							
701	GGGCCTACGCCGCCTTATGCG	GPTPPYA	55	7	56	.....	.S000003:T01G.
...							
1890	AATGCTCTGGGTCCGTCTTCG	NALGPSS	16	5	14	.....	.S000004:C01A
...							
49447	AGTGCTCTGGGTCCGTCTTCG	SALGPSS	3	2	4	.....	.S001890:A02G.S011426:C02G.
...							
100482	TGACCTACGCCGCCTTATGCG	*PTPPYA	3	1	2	A.....	.S000003:G03A.
...							
379467	TGACCTGAGCCGGTGGGTCTG	*PEPVGL	2	0	0	A.....	.
382592	TGGCCTACGCCACCTTATGCG	WPTPPYA	2	0	0	...A...	.S000003:G12A.

SI Figure S14. (A) Scheme describing the origin of F and R reads from (+) and (-) copy generated during PCR amplification. (B) F and R strands are read separately and (C) they can be identified during processing. (D) Normalized abundance of reads is F and R populations. Most reads are found in F and R reads and their copy number follow noisy Poisson distribution, some reads are found only in F or R population (red or green data). (E) One can consider union or intersect of F and R for analysis. In both populations, there are reads with NNM structure (they should not be present in the library) and point-mutations. Their distribution in the library is represented by stacked bar plots. Abundance of mutants is lower in the intersect population. (F) Description of specific mutants and NNM sequences (sequences that have A or C in every 3rd nucleotide). Scatter plot was generated using *command_center.mat*, stacked plots of erroneous populations were made by *makeFigureS10.mat* using sequence files listed in Table S1 below.

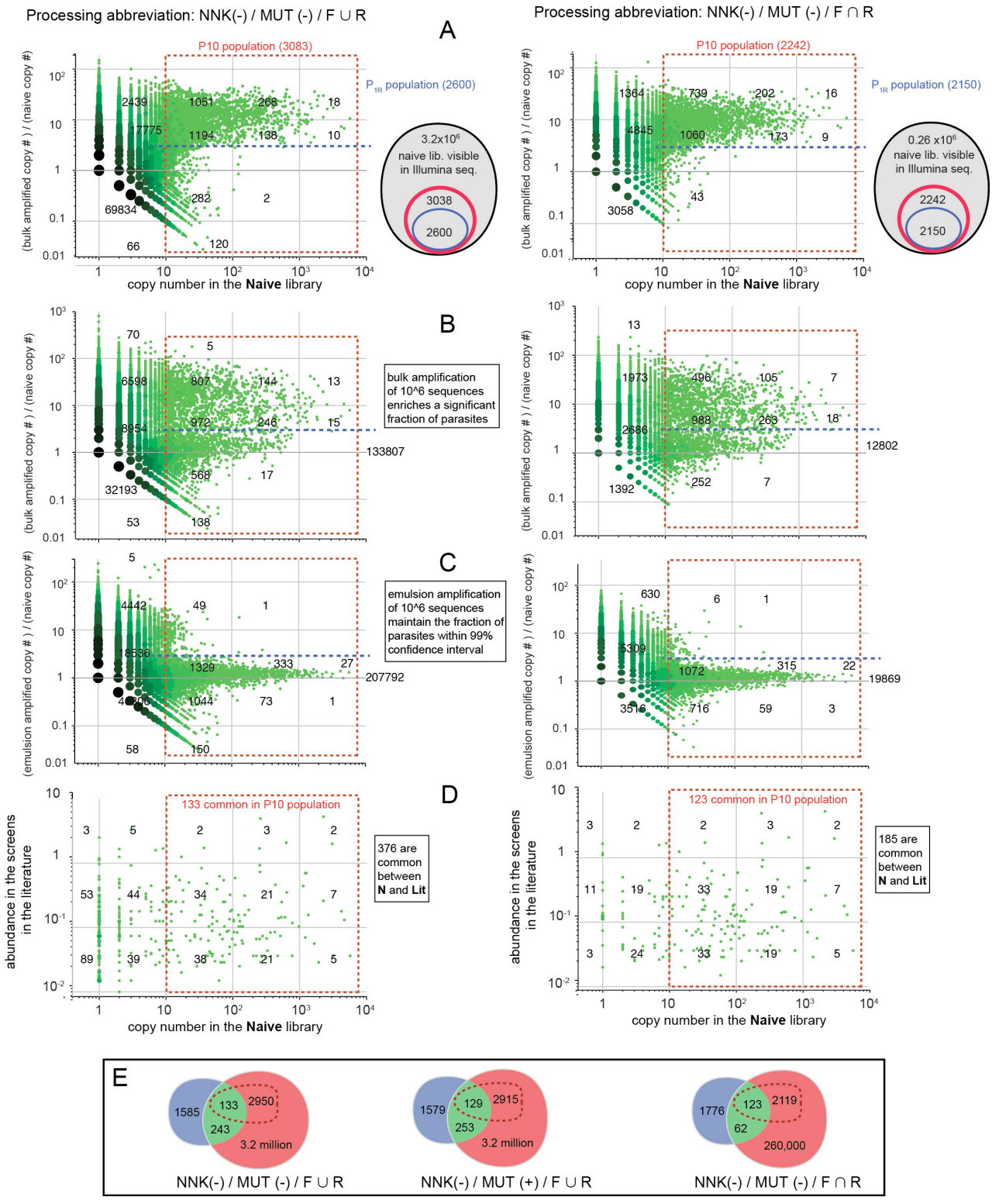


SI Figure S15. Description of **F** vs. **R** reads (A, C, E) and MUT and NNM errors (B, D, F) in three different sequencing experiments. Scatter plot was generated using *command_center.mat*, stacked plots were made by *makeFigureS10.mat* using the files described in Table S1 (next page). Every library made of union of F and R read contains ~2% of NNM error and 4-14% of MUT errors. The intersect library (F ∩ R), contains similar number of NNM errors and significantly less MUT errors.

Table S1. Files used to generate images in Figure S9 (top) and S10 (bottom). Files can be found at: http://www.chem.ualberta.ca/~derda/parasitepaper/

9	File name	Library type and processing type
D	PhD7-CGA-30F.txt	10^9-10^{15} bulk amplification, F-population, Phred>30, NNM and mutants were not removed.
D	PhD7-CGA-30R.txt	10^9-10^{15} bulk amplification, R-population, Phred>30, NNM and mutants were not removed.
E	PhD7-CGA-30FuR.txt	10^9-10^{15} bulk amplification, F-R-union, NNM and mutants were not removed.
E	MUTPhD7-CGA-30FuR.txt	10^9-10^{15} bulk amplification, F-R-union, file with all mutants and NNM errors tagged (generated in situ).
E	PhD7-CGA-30FnR.txt	10^9-10^{15} bulk amplification, F-R-intersect, NNM and mutants were not removed.
E	MUTPhD7-CGA-30FnR.txt	10^9-10^{15} bulk amplification, F-R-intersect, file with all mutants and NNM errors tagged (generated in situ).

10	File name	Library type and processing type
A	PhD7-GAC-30F.txt	10^6-10^{12} bulk amplification, F-population, Phred>30, NNM and mutants were not removed.
A	PhD7-GAC-30R.txt	10^6-10^{12} bulk amplification, R-population, Phred>30, NNM and mutants were not removed.
B	PhD7-GAC-30FuR.txt	10^6-10^{12} bulk amplification, F-R-union, NNM and mutants were not removed.
B	MUTPhD7-GAC-30FuR.txt	10^6-10^{12} bulk amplification, F-R-union, file with all mutants and NNM errors tagged (generated in situ).
B	PhD7-GAC-30FnR.txt	10^6-10^{12} bulk amplification, F-R-intersect, NNM and mutants were not removed.
B	MUTPhD7-GAC-30FnR.txt	10^6-10^{12} bulk amplification, F-R-intersect, file with all mutants and NNM errors tagged (generated in situ).
C	PhD7-GTA-30F.txt	Naïve library, F-population, Phred>30, NNM and mutants were not removed.
C	PhD7-GTA-30R.txt	Naïve library, R-population, Phred>30, NNM and mutants were not removed.
D	PhD7-GTA-30FuR.txt	Naïve library, F-R-union, NNM and mutants were not removed.
D	MUTPhD7-GTA-30FuR.txt	Naïve library, F-R-union, file with all mutants and NNM errors tagged (generated in situ).
D	PhD7-GTA-30FnR.txt	Naïve library, F-R-intersect, NNM and MUT were not removed.
D	MUTPhD7-GTA-30FnR.txt	Naïve library, F-R-intersect, file with all mutants and NNM errors tagged (generated in situ).
E	PhD7-TTG-30F.txt	10^6-10^{12} emulsion amplification, F-population, Phred>30, NNM and mutants were not removed.
E	PhD7-TTG-30R.txt	10^6-10^{12} emulsion amplification, R-population, Phred>30, NNM and mutants were not removed.
F	PhD7-TTG-30FuR.txt	10^6-10^{12} emulsion amplification, F-R-union, NNM and mutants were not removed.
F	MUTPhD7-TTG-30FuR.txt	10^6-10^{12} emulsion amplification, F-R-union, file with all mutants and NNM errors tagged (generated in situ).
F	PhD7-TTG-30FnR.txt	10^6-10^{12} emulsion amplification, F-R-intersect, NNM and mutants were not removed.
F	MUTPhD7-TTG-30FnR.txt	10^6-10^{12} emulsion amplification, F-R-intersect, file with all mutants and NNM errors tagged (generated in situ).



SI Figure S16. We removed point mutations from the libraries (left column) or processed libraries using intersect of sequences instead of union (right column). Although the size of the library changed, the conclusions remained unchanged. Two different processing methods, removal of mutations from union (left column) and intersect (right column) of **F** and **R** reads, is used to

compare the Ph.D.-7 library. Ratio plot comparing the normalized ratio of each sequence between naïve and amplified library and copy number in the naïve library for 10^9 PFU (A) (Similar to Figure 3B), 10^6 PFU (B) (Similar to Figure 7E), and 10^6 PFU emulsion amplified (C) (Similar to Figure 7F). Scatter plot comparing the abundance of sequences found in the literature (MimoDB database) and the naïve library sequenced by Illumina (Similar to Figure 6A). All four panels down a column share the same X-axis. (E) Venn diagrams show overlap of the naïve library (red), N10 parasite population (dotted red line) and literature (blue). Three diagrams describe results from three separate processing methods (outlined under the Venn diagram). The data was generated using *command_center.mat* and the following files (**Table S2**):

Table S2. Files used to generate images in Figure S11. Files can be found at:
Files can be found at: http://www.chem.ualberta.ca/~derda/parasitepaper/

	File name	Library type and processing type
1	NoMuPhD7-CGA-30FnR.txt	10^9-amplified population (bulk), F-R-intersect, mutants removed
2	NoMuPhD7-GTA-30FnR.txt	Naïve population, F-R-intersect, mutants removed
3	NoMuPhD7-GAC-30FnR.txt	10^6-amplified population (bulk), F-R-intersect, mutants removed
4	NoMuPhD7-TTG-30FnR.txt	10^6-amp. population (emulsion), F-R-intersect, mutants removed
5	NoMuPhD7-CGA-30FuR.txt	10^9-amplified population (bulk), F-R-union, mutants removed
6	NoMuPhD7-GTA-30FuR.txt	Naïve population, F-R-union, mutants removed
7	NoMuPhD7-GAC-30FuR.txt	10^6-amplified population (bulk), F-R-union, mutants removed
8	NoMuPhD7-TTG-30FuR.txt	10^6-amp. population (emulsion), F-R-union, mutants removed
9	PhD7-literature.txt	List of all sequences from MimoDB2.0

References

[1] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. JRSS: B 57, 289-300.
[2] Robinson, M.D. and Smythn G.K. (2007). Moderated statistical tests for assessing differences in tag abundance. Bioinformatics 23, 2881-2887.
[3] Robinson, M.D. and Smyth, G.K. (2008). Small sample estimation of negative binomial dispersion, with applciations to SAGE Data. Biostatistics 9, 321-332.
[4] Robinson, M.D. and Oshlack, A. (2010). A scaling normalizaiton method for differnetial expression analysis of RNA-seq data. Genome Biology 11, R25.