

1 **Incorporation of unique molecular identifiers in TruSeq adapters improves the accuracy of quantitative**  
2 **sequencing**

3  
4 Jungeui Hong and David Gresham  
5

6 **SUPPLEMENTARY INFORMATION**

7  
8 **MATERIALS AND METHODS**

9  
10 **Preparing TrUMIseq adapters**

11 We designed two modified oligonucleotides based on the Illumina P5 and P7 oligonucleotides, which were  
12 commercially synthesized ([www.idtdna.com](http://www.idtdna.com)). The modified P5 oligonucleotide contains a phosphorothioate  
13 bond between the 3' T and the adjacent 6 base sample index. A 'T' was added 5' to the sample index to  
14 ensure complementarity with the Illumina forward read sequencing primer. The modified P7 oligonucleotide  
15 contains a 5' phosphate group, a 6-base random sequence that serves as the UMI and the complementary  
16 sample index sequence (**Figure 1A and Supplementary Table 1**). In principle, each pool of adapters should  
17 have 4096 (i.e. 4<sup>6</sup>) unique UMIs. To ensure true randomization of UMIs, reagents were hand mixed by the  
18 oligonucleotide synthesizer. The two partially complementary oligonucleotides were annealed to form the Y-  
19 shaped adapter with a 'T' overhang as follows:

- 20 a) Each individual oligonucleotide was re-suspended at the same molar concentration (20μM) in  
21 annealing buffer (10mM Tris, pH 7.5–8.0, 50mM NaCl, 1mM EDTA).
- 22 b) Equal volumes of partially complementary oligonucleotides were mixed, placed in a standard  
23 heatblock at 95°C for 5 minutes, and then cooled to room temperature on a workbench for 1 hour.
- 24 c) The annealed adapters were checked on a non-denaturing 5-6% PAGE gel. Successful annealing was  
25 determined by ~90 % of the band running at a molecular weight of 300-400bp due to the gel migration  
26 properties of the Y-shaped partially dsDNA molecule.
- 27 d) Annealed adapters were kept at -20°C for long-term storage.

28

29

## 30 **Library sequencing protocol**

31 We used the TrUMIseq adapters for whole genome population DNA-seq (3 libraries), targeted amplicon  
32 sequencing (12 libraries), and strand specific RNA-seq (9 libraries) using samples from *Saccharomyces*  
33 *cerevisiae*. While most library preparation steps were identical to the standard TruSeq protocol, some  
34 variations were introduced as follows:

- 35 a) All reaction cleanup and DNA insert size selections were performed using AMPure® beads (Beckman  
36 Coulter, Pasadena, CA, USA).
- 37 b) For amplicon sequencing, each amplicon was fragmented using sonication prior to adapter ligation.
- 38 c) For RNA-seq libraries in which the amount of starting material was limited, only 0.5µM of the adapter  
39 was used for ligation. Otherwise, all ligation protocols used adaptor concentrations of 20µM, which  
40 we found results in sufficient ligated molecules and minimizes adapter dimer formation.
- 41 d) The number of PCR cycles varied from 8 to 15 depending on the amount of starting materials.
- 42 e) The final concentration of libraries loaded onto a flow-cell was slightly higher than the standard  
43 requirement of 2nM.
- 44 f) PhiX DNA (Illumina San Diego, CA, USA), was added in each sequencing lane in order to minimize  
45 the negative effect of low base diversity in the first 7 sequencing cycles that determine the 6-mer  
46 sample index plus the 'T' overhang in the 7<sup>th</sup> position. As multiple different sample indices were  
47 multiplexed in each sequencing run, only the 7<sup>th</sup> position is identical in all sequence reads. This results  
48 in low quality base calls at the 7<sup>th</sup> position, but does not negatively impact base calls at other read  
49 positions. Multiplexed libraries were sequenced using either 2x50 bp paired end sequencing for DNA-  
50 seq on an Illumina HiSeq 2500 (San Diego, CA, USA) or 2x250 bp paired end sequencing for RNA-  
51 seq and targeted AMP-seq on an Illumina MiSeq (San Diego, CA, USA) for this study and have  
52 subsequently been used in our laboratory with an Illumina NextSeq 500 (San Diego, CA, USA).

53

54

55 **Data processing and analysis**

56 Demultiplexing was performed using a custom perl script (demultiplex\_TrUMIseq.pl) using NYU's high  
57 performance computing facility. For sample demultiplexing, we allowed only one mismatch in the six  
58 nucleotide sample index. The first 7 nucleotides comprising the sample index and 'T' overhang in every read  
59 were trimmed prior to downstream analysis. For read alignment we used BWA -mem (1) and Tophat2 (2) to  
60 align against the *Saccharomyces cerevisiae* S288C reference genome, obtained from the SGD database on Feb  
61 03, 2011. PCR duplicate rates were calculated based on analysis of SAM files using a custom perl script  
62 (check\_dup\_TrUMIseq\_v2.pl): all alignments reporting identical coordinate information were selected and for  
63 each set of reads mapping to the same coordinates, only those with a unique 6 base UMI were considered to be  
64 non-PCR duplicates (**Figure 1b**). We considered all UMIs to be unique regardless of their edit distance from  
65 other UMIs. We confirmed random incorporation of bases during oligonucleotide synthesis by assessing the  
66 nucleotide frequency at each position in UMIs from all sequencing reads, which is close the expected  
67 frequencies of 0.25 (**Supplementary Figure 1**). All poorly mapped (mapping quality less than 10) and  
68 misaligned paired reads were removed in this analysis. We used Picard (<http://picard.sourceforge.net>) to  
69 identify PCR duplicates on the basis of coordinate information only using the MarkDuplicates tool. For SNP  
70 detection, and allele quantification, in population samples based on the DNA-seq or AMP-seq, we used SNVer  
71 (3) with a minimum detection limit of 1%. We used EdgeR (4) to identify differently expressed genes  
72 between control and treated samples from RNA-seq data. All statistical analyses were conducted using R.  
73 Custom perl scripts for demultiplexing and analyzing TrUMIseq data are available at the github repo:  
74 <https://github.com/GreshamLab/trumiseq>  
75  
76 Nucleotide and homopolymeric sequence frequencies for all UMIs identified in an RNAseq sample sequenced  
77 using a Nextseq 500 were determined using the grep command in unix. Expected counts were determined for  
78 each class of homopolymeric sequence using either the probability of the homopolymeric sequence multiplied  
79 by the number of possible locations within the UMI (for  $N_6-N_2$ ) or the binomial distribution (to compute the  
80 probability of at least one N or of zero N occurring in a UMI).

## 81 Commands used in this analysis

```
#####  
# PBS script for de-multiplexing #  
#####  
NUM= # Number of temporary split file  
FQ1= # FASTQ R1 FILE NAME  
FQ2= # FASTQ R2 FILE NAME  
LIB= # LIBRARY FILE NAME  
# → A tab-delimited text file with Col1 = library name / Col 2=sample barcode  
perl demultiplex_TrUMIseq.pl -l $LIB -f1 ${FQ1}_${NUM} -f2 ${FQ2}_${NUM} -bq 10 -m 1
```

---

---

```
#####  
# PBS script for DNA-seq and AMP-seq data analysis #  
#####  
TAG=TEST # Sample name  
REF=Ref.SGD020311.fasta # reference fasta sequence  
READ1=${TAG}_R1.fastq # input original fastq file 1  
READ2=${TAG}_R2.fastq # input original fastq file 2  
  
#####  
# (1) indexing the reference file #  
#####  
bwa index -a bwtsv $REF  
  
#####  
# (2) Align reads to the reference sequence to generate SAM file. #  
#####  
bwa mem -t 12 -C $REF $READ1 $READ2 > ${TAG}.sam  
  
#####  
# (3) check pcr duplicates based on the aligned SAM file #  
# and output non-PCR duplicates reads in fastq format #  
# -m : number of mismatches allowed in the UMI #  
#####  
check_dup_TrUMIseq_v2.pl -in ${TAG}.sam -m 1
```

```
#####  
# PBS script for RNA-seq data analysis #  
#####  
TAG=TEST # Sample name  
REF=Ref.SGD020311.fasta # reference fasta sequence  
REF_DB=Ref.SGD020311 # reference DB name for bowtie2 alignment  
GFF=saccharomyces_cerevisiae_R64-1-1_20110208_only_orf_converted.gff # gene info file  
READ1=${TAG}_R1.fastq # input original fastq file 1  
READ2=${TAG}_R2.fastq # input original fastq file 2  
  
#####  
# (1) indexing the reference file using bowtie2-build #  
#####  
bowtie2-build $REF $REF_DB  
  
#####  
# (2) align reads to the reference sequence to generate SAM file #  
#####  
# - p : multithreading / -G : input GFF file  
# --no-convert-bam : output should be SAM format  
# --library type : first strand / -o : output folder
```

```
tophat -p 12 -G $GFF --no-convert-bam --library-type fr-firststrand -o tmp $REF_DB $READ1 $READ2
```

```
#####  
# (3) check pcr duplicates based on the aligned SAM file #  
# and output non-PCR duplicates reads in fastq format #  
# -m : number of mismatches allowed in the UMI #  
#####  
perl check_dup_TrUMIseq_v2.pl -in ${TAG}.sam -m 1
```

82

83

84 **Accession number for sequencing data**

85 All sequencing data (in fastq format) are available from the NCBI Sequence Read Archive with accession

86 number SRP101366 for DNA-seq, SRP101367 for AMP-seq and SRP101370 for RNA-seq.

87 **REFERENCES**

88

- 89 1. **Li, H., and R. Durbin.** 2009. Fast and accurate short read alignment with Burrows-Wheeler transform.  
90 *Bioinformatics* 25:1754–1760.
- 91 2. **Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S.L. Salzberg.** 2013. TopHat2: accurate  
92 alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*  
93 *14*:R36.
- 94 3. **Wei, Z., W. Wang, P. Hu, G.J. Lyon, and H. Hakonarson.** 2011. SNVer: a statistical tool for variant  
95 calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Res* 39:e132–  
96 e132.
- 97 4. **Robinson, M.D., D.J. McCarthy, and G.K. Smyth.** 2010. edgeR: a Bioconductor package for  
98 differential expression analysis of digital gene expression data. *Bioinformatics* 26:139–140.

99

100 **SUPPLEMENTARY TABLES**  
101

102 **Supplementary Table 1. Sequences of oligonucleotides used for generating TrUMIseq adapters used in**  
103 **this study.** Additional adapters can be designed by modification of the 6-base sample index sequence  
104 (underlined).

105

106 **Supplementary Table 2. Sequencing quality metrics for experiments using TrUMIseq adapters.**

107 Sequencing quality metrics for seven different sequencing experiments using TrUMIseq adapters on a MiSeq,  
108 HiSeq and NextSeq machine.

109

110 **Supplementary Table 3. Analysis of homopolymeric sequences in TrUMIseq adapters.** The distribution  
111 of homopolymeric sequences was assessed in a single RNAseq library analyzed on a NextSeq 500. A total of  
112 4026 UMIs were present in 675,876 unique reads following deduplication. Nucleotide frequencies across  
113 UMIs were determined for each base, the frequency of UMIs containing the expected homopolymeric  
114 sequences computed, and compared to the observed frequency of each UMI in the sample. Note that the  
115 counts for each class of homopolymeric sequence is cumulative as no constraints were imposed on bases that  
116 are not within the homopolymeric sequence.

