Supporting Information

Fu et al. 10.1073/pnas.1017621108

SI Text

Figure S4 is available at the following institutional web site: http://www.affymetrix.com/community/publications/affymetrix/fu_et_al_2011_pnas_fig_s4.pdf.

Table S1 is available at the following institutional web site: http://www.affymetrix.com/community/publications/affymetrix/fu_et_al_2011_pnas_table_s1.pdf.

Stochastic Model. The stochastic labeling process can be defined as follows. Consider n copies of a given target molecule, $T=\{t_j,j=1,2,\ldots,n\}$, and a nondepleting reservoir of m diverse labels, $L=\{l_i,i=1,2,\ldots,m\}$. T reacts with L stochastically, such that each t_j will choose exactly one $l_{i(j)}$, $1\le i(j)\le m$, to take on a new identity $l_{i(j)}t_j$, and may be identified by its label subscript (the subscript j for the copy of target molecule t_j may be dropped). Therefore, the new collection of molecules T^* may be denoted as $T^*=\{l_{i(j)}t,j=1,2,\ldots,n,1\le i(j)\le m\}$. When different copies of the target molecules react with the same label, $i(j)$ for those molecules will assume the same value, therefore, the number of uniquely labeled target molecules k cannot be greater than m. The stochastic reaction of the set of labels on a target may be described by a stochastic operator S with m members, acting upon a target population of n, such that $S(m)T(n)=ST=T^*(m,n)$ generating the set $T^*=\{l_{i(j)}t,j=1,2,\ldots,n,1\le i(j)\le m\}$. For simplicity, we may write $T^*=\{l_kt\}$. Furthermore, because S operates on all molecules independently, it will act on many different target sequences and the method can be expanded to count copies of multiple target sequences, w, simultaneously: $ST^w=ST_1+ST_2+\ldots+ST_w=T_1{}^*+T_2{}^*+\ldots+T_w{}^*=\{l_kt\}_1+\{l_kt\}_2+\ldots+\{l_kt\}_w$, where each $T_i{}^*,i=1,2,\ldots,w$ consists of a set $\{l_kt\}_i$. The net result of S operating on a specific target population is to map the number of molecules, n, of that target, to the number of distinct labels captured, k, which is a random variable.

Because target molecules randomly react with a label with probability $\frac{1}{m}$, the probability of a label being captured by exactly x out of n copies of a target molecule can be modeled as a binomial distribution, $P(x)=\frac{n!}{x!(n-x)!}\left(\frac{1}{m}\right)^x\left(1-\frac{1}{m}\right)^{n-x}$, where $x!$ denotes the factorial of x. The probability that a label will not be captured by any copy of the target molecule is $P(0)=(1-1/m)^n$, and the probability that a label will be captured at least once is $1-P(0)$. When $n\to\infty$ and $1/m\to 0$ in the way that $n/m\to\lambda$, $P(x)$ converges to the Poisson distribution with mean λ, i.e., $P(x)=\frac{\lambda^x}{x!}e^{-\lambda}$.

To compute the number of unique labels captured by n copies of a target molecule, we introduce an index random variable, X_i, which is one if a label has been captured at least once, and zero otherwise. The number of unique labels captured is thus $k=\sum_{i=1}^m X_i$. The mean and variance of k can be derived,

$$E[k]=m\left[1-\left(1-\frac{1}{m}\right)^n\right] \qquad \textbf{[S1]}$$

$$\begin{aligned}\mathrm{Var}[k]=&\,m\left[1-\left(1-\frac{1}{m}\right)^n\right]\left(1-\frac{1}{m}\right)^n\\&+m(m-1)\left[\left(1-\frac{2}{m}\right)^n-\left(1-\frac{1}{m}\right)^{2n}\right]\end{aligned} \qquad \textbf{[S2]}$$

Similarly, to compute the number of labels captured by exactly x copies of a target molecule, we introduce another index random variable, Y_i, which is one if a label has been captured exactly x times, and zero otherwise. The number of labels captured x times is thus $t=\sum_{i=1}^m Y_i$. The mean and variance of t are,

$$E[t]=\frac{m\cdot n!}{x!(n-x)!}\left(\frac{1}{m}\right)^x\left(1-\frac{1}{m}\right)^{n-x} \qquad \textbf{[S3]}$$

$$\begin{aligned}\mathrm{Var}[t]=&\,A(1-A)+(m-1)m\cdot\binom{n}{2x}\\&\cdot\left(\frac{2}{m}\right)^{2x}\left(1-\frac{2}{m}\right)^{n-2x}\binom{2x}{x}\left(\frac{1}{2}\right)^{2x},\end{aligned} \qquad \textbf{[S4]}$$

where $A=m\cdot\binom{n}{x}\left(\frac{1}{m}\right)^x\left(1-\frac{1}{m}\right)^{n-x}$, and the combination $\binom{n}{x}=\frac{n!}{x!(n-x)!}$.

Additionally, to validate these equations, we performed numerical simulations with 5,000 independent runs for each simulated case and observed complete agreement with the analytical solutions.

SI Tables

Table S1 provides a list of DNA sequences of the three chromosome gene fragments, oligonucleotides for PCR primers and ligation adaptors used in the study.

In Table S2, we show the labels detected on microarray experiments. Indicated quantities [column (col.) 2] of genomic DNA derived from a Trisomy 21 male sample were tested on three chromosome targets (col. 1). The estimated number of copies of target molecules (or haploid genome equivalents, col. 3), the number of labels expected by the stochastic model (col. 4), and the actual number of labels detected on microarrays (col. 6–8) are summarized. Because each gene target fragment paired-end consists of random, independent label ligation events at the left (L) and the right (R) termini, the number of identical labels expected (col. 5) can be predicted from computer simulations, and compared to the number actually detected (col. 11). Given the number of labels detected (col. 8), we obtain the corresponding number of copies of target molecules (col. 9) in our stochastic model, and the predicted occurrences of identical labels across paired-ends (col. 10). The numbers in col. 5 and 10 are the means from 5,000 independent simulation runs along with one standard deviation of the corresponding means, given the number of labels at either end (col. 4 and col. 9).

In Table S3 the number of mapped reads from SOLiD DNA sequencing is given, and Table S4 shows label usage summaries from experimental observation (black), or from the stochastic model (red).

Finally, Table S5 gives the PCR detection for the presence of label sequences in the processed DNA sample that was hybridized to microarray, or in the DNA sequencing library. Each PCR contained 0.1 pg of template, which represents approximately 1×10^6 DNA molecules. The number of mapped sequencing reads and the microarray intensity of each of the 16 labels for this selected gene target (chromosome 4, 3.62 ng) are listed. In the last two columns, lettering in red corresponds to reactions where PCR failed to detect the label sequence in the sample.

Other Supporting Information Files

Fig. S1. The expected label usage (y axis) when ligating to a given number of target molecules (x axis). Each copy of a target molecule randomly ligates to only a single copy of one of 1,000 distinct labels equally represented in a library pool in nondepleting quantities. The number of occurrences that each label is used is plotted in the color indicated. The expected label usage was obtained from Eqs. **S1** and **S3** and described in *Materials and Methods.*

Fig. S1 (PDF)

Fig. S2. The target molecule counting efficiency based on a library of 1,000 labels. Counting efficiency is expressed as the ratio of labels observed at least once (red line in Fig. S1) over the number of target molecules present. Low target numbers (inset plot) display a near-linear relationship, but counting efficiency decreases nonlinearly upon increased repetitive usage of the same labels.

Fig. S2 (PDF)

Fig. S3. DNA sequences in the ligation of adaptors harboring label sequence tags, and in the sample preparation for microarray hybridization. The arrangement and position of the adaptors, PCR primers, and the biotinylated array-ligation probe used are shown in detail.

Fig. S3 (PDF)

Fig. S4. (*A*) Labels observed in the microarray experiments. Microarray intensity (y axis, log scale) for each of the 960 labels observed in each side of each gene within each sample is shown. The blue dashed line indicates the counting threshold applied. Microarray intensities for 192 nonspecific labels (negative controls) are shown on the right side of the red dashed line. (*B*) Labels observed in the mapped reads from the first SOLiD sequencing run. The number of reads for each of the 960 labels observed in each side of each gene within each sample is shown. The blue dashed line indicates the counting threshold applied. Reads mapped to any of the 192 nonspecific labels (negative controls) are shown on the right side of the red dashed line. (*C*) Venn diagrams of the observed number of labels in common between the microarray and the two sets of sequencing experiments. The plot size of the Venn diagram circles and the area of overlap are not proportionally scaled to the results displayed. Filled dots within the circles are shown to provide a visual representation of the number of labels in each category.

Fig. S4 (PDF)

Fig. S5. DNA sequences in the construction of SOLiD sequencing libraries. The arrangement and position of the adaptors and PCR primers used to convert the DNA sample hybridized to microarrays into sequencing templates are shown in detail.

Fig. S5 (PDF)

Fig. S6. A comparison between experimentally observed label usage rates (blue bars) with those predicted from stochastic model (red bars). At low target molecule numbers, the chance of multiple target ligations to the same label sequence is low. We therefore only consider data from experiments with low target numbers (0.036 and 0.36 ng of DNA). From these experiments, a total of 1,064 labels were observed, with the total frequency of label usage ranging from zero to six. The theoretically expected label usage frequency for 1,064 target molecules was obtained by performing 5,000 simulation runs, with multiple independent reactions simulated in each run. The error bars indicate one standard deviation from the corresponding means.

Fig. S6 (PDF)

Table S1 (PDF)
Table S2 (DOCX)

Table S3 (DOCX)
Table S4 (DOCX)
Table S5 (DOCX)