Counting individual DNA molecules by the stochastic attachment of diverse labels

Glenn K. Fu, Jing Hu, Pei-Hua Wang, and Stephen P. A. Fodor[1]

Affymetrix, Inc., 3420 Central Expressway, Santa Clara, CA 95051

Edited* by Ronald W. Davis, Stanford Genome Technology Center, Palo Alto, CA, and approved March 22, 2011 (received for review November 27, 2010)

We implement a unique strategy for single molecule counting termed stochastic labeling, where random attachment of a diverse set of labels converts a population of identical DNA molecules into a population of distinct DNA molecules suitable for threshold detection. The conceptual framework for stochastic labeling is developed and experimentally demonstrated by determining the absolute and relative number of selected genes after stochastically labeling approximately 360,000 different fragments of the human genome. The approach does not require the physical separation of molecules and takes advantage of highly parallel methods such as microarray and sequencing technologies to simultaneously count absolute numbers of multiple targets. Stochastic labeling should be particularly useful for determining the absolute numbers of RNA or DNA molecules in single cells.

absolute counting | digital PCR | next-generation sequencing | single molecule detection

Determining small numbers of biological molecules and their changes is essential when unraveling mechanisms of cellular response, differentiation or signal transduction, and in performing a wide variety of clinical measurements. Although many analytical methods have been developed to measure the relative abundance of different molecules through sampling (e.g., microarrays and sequencing), the only practical method available to determine the absolute number of molecules in a sample is digital PCR (1–3), a powerful analytical technique typically limited to examining only a few different molecules at a time.

In 2003, a theoretical approach to measure the number of molecules of a single mRNA species in a complex mRNA preparation was proposed (4). To our knowledge no experimental demonstration of this idea has been published. We have generalized this idea and have expanded it to a highly parallel method capable of absolute counting of many different molecules simultaneously. The concept is illustrated in Fig. 1. Each copy of a molecule randomly captures a label by choosing from a large, nondepleting reservoir of diverse labels. The subsequent diversity of the labeled molecules is governed by the statistics of random choice, and depends on the number of copies of identical molecules in the collection compared to the number of kinds of labels. Once the molecules are labeled, they can be amplified so that simple present/absent threshold detection methods can be used for each. Counting the number of distinctly labeled targets reveals the original number of molecules of each species.

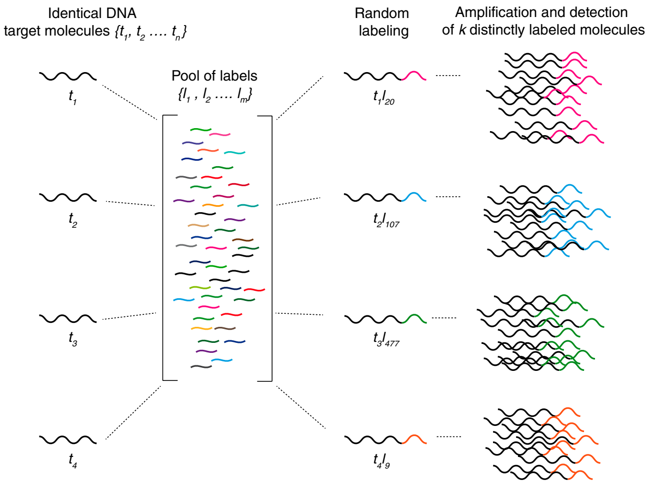

Identical DNA target molecules $\{t_1, t_2 .... t_n\}$
Pool of labels $\{l_1, l_2 .... l_m\}$
Random labeling
Amplification and detection of k distinctly labeled molecules
t_1
t_2
t_3
t_4
t_1l_{20}
t_2l_{107}
t_3l_{477}
t_4l_9

Fig. 1. A schematic representation of the labeling process. An example showing four identical target molecules in solution. Each DNA molecule randomly captures and joins with a label by choosing from a large, nondepleting reservoir of m labels. Each resulting labeled DNA molecule takes on a new identity and is amplified to detect the number of k distinct labels.

We can generalize the stochastic labeling process as follows. Consider a given set of copies of a single target sequence $T = \{t_1,t_2\ldots t_n\}$; where n is the number of copies of T. A set of labels is defined as $L = \{l_1,l_2\ldots l_m\}$; where m is the number of different labels. T reacts stochastically with L, such that each t becomes attached to one l. If the ls are in nondepleting excess, each t will choose one l randomly, and will take on a new identity l_it_j; where l_i is chosen from L and j is the jth copy from the set of n molecules. We identify each new molecule l_it_j by its label subscript and drop the subscript for the copies of T because they are identical. The new collection of molecules becomes $T^* = \{l_1t,l_2t,\ldots l_it\}$; where l_i is the ith choice from the set of m labels. At this point, the subscripts of l refer only to the ith choice and provide no information about the identity of each l. In fact, l_1 and l_2 will have some probability of being identical, depending upon the diversity m of the set of labels. Overall, T^* will contain a set of k unique labels resulting from n targets choosing from the nondepleting reservoir of m labels. Or, $T^*(m,n) = \{l_kt\}$; where k represents the number of unique labels that have been captured. In all cases, k will be smaller than m, approaching m only when n becomes very large. We can define the stochastic attachment of the set of labels on a target using a stochastic operator S with m members, acting upon a target population of n, such that $S(m)T(n) = T^*(m,n)$ generating the set $\{l_kt\}$. Furthermore, because S operates on all molecules independently, it can act on many different targets. Hence, by combining the information of target sequence and label, we can simultaneously count copies of multiple target sequences. The probability of the number of labels generated by the number of trials n, from a diversity of m, can be approximated by the Poisson equation, $P_x = [(n/m)^x/x!]e^{-(n/m)}$. Then P_0 is the probability that a label will not be chosen in n trials, therefore, $1 - P_0$ is the probability that a label will occur at least once. It follows that the expected number of unique labels captured is given by:

Author contributions: G.K.F. and S.P.A.F. designed research; G.K.F. and P.-H.W. performed research; G.K.F., J.H., and S.P.A.F. analyzed data; and G.K.F., J.H., and S.P.A.F. wrote the paper.

Conflict of interest statement: The authors are employees of Affymetrix, Inc. and the subject matter of this article may be a future commercial product.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

[1] To whom correspondence should be addressed. E-mail: steve_fodor@affymetrix.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1017621108/-/DCSupplemental.

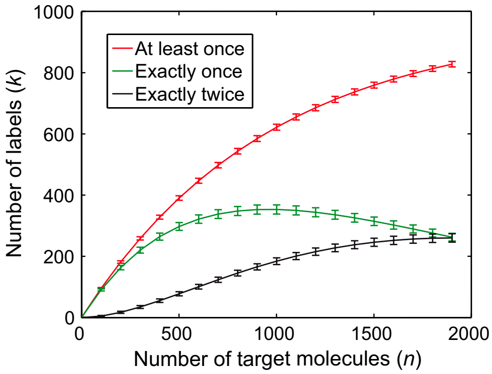


Fig. 2. The number of stochastically captured labels for a given number of target molecules calculated using a nondepleting reservoir of 960 diverse labels. The red curve represents the average number of labels observed at least once (calculated from Eq. **S1**); the green and black curves represent the number of labels observed exactly once and twice (calculated from Eq. **S3**), respectively. Error bars indicate one standard deviation (calculated from Eqs. **S2** and **S4**) away from the corresponding mean values.

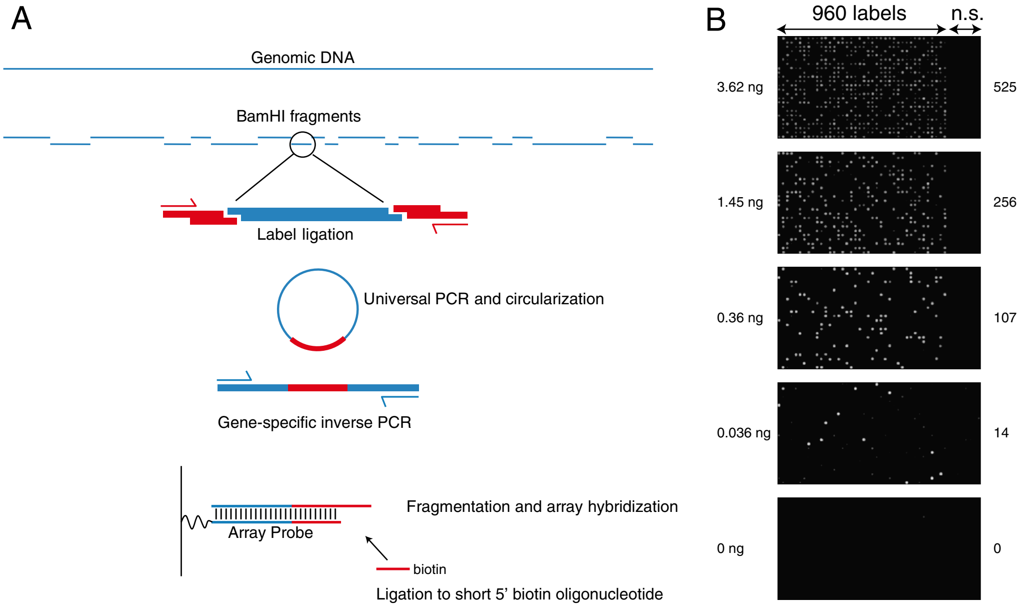
$$k = m(1 - P_0) = m[1 - e^{-(n/m)}]. \qquad [1]$$

Given k, we can calculate n. In addition to using the Poisson approximation, the relationship for k, n, and m can be described using the binomial distribution, or simulated using a random number generator, each yielding similar results (*SI Text*).

Results

The outcome of stochastic labeling is illustrated by examining the graph of k (the red curve in Fig. 2) calculated using a label diversity (m) of 960. The expected number of unique labels captured depends on the ratio of molecules to labels, n/m. When n is much smaller than m, each molecule almost always captures a unique label, and counting k is equivalent to counting n. As n increases, k increases more slowly as given by Eq. **1**. For example, when n/m is approximately 0.01, the counting efficiency, which is defined as the ratio of unique labels to molecules k/n is approximately 0.99, and we expect that an increase of 10 molecules will generate 10 new labels. As n/m approaches 0.5 (i.e., 480 molecules reacted with 960 labels), k/n becomes approximately 0.79 and six new labels are expected with an increase of 10 molecules. At high n/m, k increases more slowly as labels in the set are more likely to be captured more than once. The green curve in Fig. 2 shows the number of labels chosen exactly once, and the black curve shows the number of labels chosen exactly twice as n increases. A more complete description of the number of times a label is chosen and of the counting efficiency as a function of n is shown in Figs. S1 and S2.

To demonstrate stochastic labeling, we performed an experiment to count small numbers of nucleic acid molecules in solution. We used genomic DNA from a male individual with Trisomy 21 to determine the absolute and relative number of DNA copies of chromosomes X, 4, and 21, representing one, two, and three target copies of each chromosome, respectively. The DNA concentration in the stock solution was measured by quantitative staining with PicoGreen fluorescent dye, and dilutions containing 3.62, 1.45, 0.36, and 0.036 ng were prepared. In each dilution, the number of copies of target molecules in the sample was calculated from a total DNA mass of 3.5 pg per haploid nucleus (5), and represent approximately 1,000, 400, 100, and 10 haploid genome equivalents. As outlined in Fig. 3*A*, the genomic DNA sample was first digested to completion with the BamHI restriction endonuclease to produce 360,679 DNA fragments. A diverse set of labels consisting of 960 14-nt sequences was synthesized as adaptors harboring BamHI overhangs (Table S1). This set of labels adequately addresses a broad dynamic range and was chosen for favorable thermodynamic properties as described in *Materials and Methods*. For the stochastic labeling reaction, each DNA fragment end randomly attaches to a single label by means



A
Genomic DNA
BamHI fragments
Label ligation
Universal PCR and circularization
Gene-specific inverse PCR
Fragmentation and array hybridization
Array Probe
biotin
Ligation to short 5' biotin oligonucleotide
B
960 labels
n.s.
3.62 ng
525
1.45 ng
256
0.36 ng
107
0.036 ng
14
0 ng
0

Fig. 3. (*A*) A schematic drawing of the method used to attach labels to fragments of DNA in the genome. Red bars represent a pool of synthetic deoxyoligonucleotide adaptors incorporating a collection of 960 labels used as counting sequences. A common primer sequence flanks each unique label adaptor, allowing universal amplification of all fragments with PCR. Circularization of amplified DNA molecules simplifies the selection and amplification of label-ligated DNA fragments through inverse PCR with gene-specific primers. The identity of labels that have been ligated to the genomic DNA fragment is determined using microarray hybridization, or DNA sequencing. (*B*) Microarray scan images of the 960 tiled probes for chromosome 4 corresponding to the labels used, as well as an additional 192 nonspecific (n.s.) probes serving as negative controls. The amount of genomic DNA used in each experiment is given on the left side of each image and the number of labels detected on microarrays is provided on the right side.

of enzymatic ligation of compatible cohesive DNA ends. High coupling efficiency is achieved through incubation with a large molar excess of labels and DNA ligase enzyme (>10^{13} molecules each). At this stage, the stochastic labeling process is complete, and the samples can be amplified as desired for detection. A universal primer is added, and the entire population of labeled DNA fragments is PCR amplified. The PCR reaction preferentially amplifies approximately 80,000 fragments in the 150 bp–2 kb size range. After circularization of the amplified products, three test target fragments were isolated using gene-specific PCR; one on each of chromosomes X, 4, and 21, and prepared for detection.

The three labeled targets were counted using two sampling techniques: DNA microarrays and next-generation sequencing. For the array counting, a custom DNA array detector capable of distinguishing the set of labels bound to the targets was constructed by dedicating one array element for each of the 960 target-label combinations. Each array element consists of a complementary target sequence attached to one of the complements of the 960 label sequences (Fig. 3*A*, Fig. S3). To maximize the specificity of target-label hybridization and scoring, we employed a ligation labeling procedure on the captured sequences (Fig. S3). We set thresholds to best separate the intensity data from the array into two clusters, one of low intensity and one of high intensity (Fig. S4*A*). We scored a label as "present" if its signal intensity exceeded the threshold. The number of labels detected on microarrays is summarized in Table S2. Fig. 3*B* shows examples of microarray scan images where bright spots/features were counted as present. As an alternate form of detection, sequencing adaptors were added (Fig. S5) and the samples were subjected to two independent DNA sequencing runs. Between several hundred thousand to several million high-quality reads were used to score the captured labels (Table S3). Similarly, we set thresholds for the number of sequencing reads observed for each label, and scored a label as present if the number of sequencing reads exceeded the threshold (Fig. S4*B*). The number of attached labels, k, detected for each target in each dilution either by microarray counting or sequence counting is presented in Table S4, and plotted in Fig. 4 *A* and *B*.

The counting results span a range of approximately 1,500 to 5 molecules, and it is useful to consider the results in two counting regimes, below and above 200 molecules. There is a striking agreement between the experimentally observed number of molecules and that expected from dilution in the first regime where the ratio of molecules to labels $(n/m) < 0.2$ (Table S4). Below 200 molecules the data are in tight agreement, including the data from the lowest number of molecules—5, 10, and 15—where the counting results are all within the expected sampling error for the experiment. (The sampling error for 10 molecules is 10 ± 6.4, where 10 and 6.4 are the mean and two standard deviations from 10,000 independent simulation trials.)

In the second regime above 200 molecules, there is an approximate 10–25% undercounting of molecules, increasing as the number of molecules increases. We attribute this deviation to be

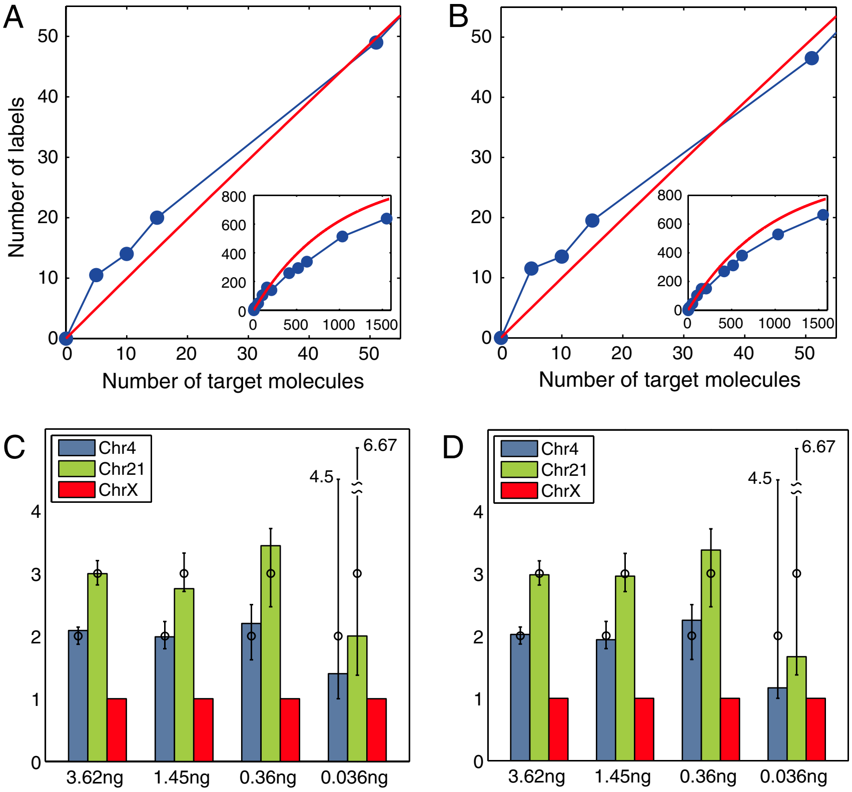


Fig. 4. Absolute counting results for DNA molecules. 3.62, 1.45, 0.36, and 0.036 ng dilutions of DNA isolated from cultured lymphoblasts of a Trisomy 21 male individual were processed for microarray hybridization and DNA sequencing. Three gene targets were tested, one from each of chromosomes X, 4, and 21, and the numbers of detected labels (blue curve) are shown for microarray (*A*) and DNA sequencing (*B*). The number of target molecules for each sample was determined from the amount of DNA used, assuming a single haploid nucleus corresponds to 3.5 pg. For comparison, the calculated number of labels expected from a stochastic model is also plotted in red. Numerical values are provided in Table S2. Copy number ratios of the three gene targets ChrX (red bar), Chr4 (blue bar), and Chr21 (green bar) representing one, two, and three copies per cell, respectively, are shown in (*C*) and (*D*). The calculated number of target molecules was determined from the number of labels detected on microarrays (Table S2, column 9) or from DNA sequencing. For each sample dilution, the copy number ratio of each gene target relative to ChrX is shown for microarray (*C*) and DNA sequencing (*D*). For comparison, copy number ratios obtained from *in silico* sampling simulations are also shown; where circles indicate the median values from 10,000 independent trials and error bars indicate the 10th and 90th percentiles. The 90th percentile values of the ratios at the lowest concentration (0.036 ng) are explicitly labeled in the plots.

due to a distortion in the amplification reaction. PCR-introduced distortion occurs from small amounts of any complex template due to the differences in amplification efficiency between individual templates (6–8). In the present case, stochastic labeling will produce only one (at low n/m ratios), and increasingly several copies (at higher n/m ratios) of each template. Modeling suggests that simple random dropout of sequences (PCR efficiencies under 100%) generates significant distortion in the final numbers of each molecule after amplification. At any labeling ratio, random dropout of sequences because of PCR efficiency will result in an undercount of the original number of molecules. At high n/m ratios, the number of labels residing on multiple targets will increase and have a statistical survival advantage through the PCR reaction causing greater distortion. In support of this argument, we observe a wide range of intensities on the microarray and a wide range in the number of occurrences of specific sequences in the sequencing experiments (Fig. S4 *A* and *B*). This effect can be reduced by carrying out the reaction at n/m ratios near or less than 0.2, increasing the number of labels m, further optimization of the amplification reaction, or by employing a linear amplification method.

The lymphoblast cell line used in this study provides an internal control for the relative measurement of copy number for genes residing on chromosomes X, 4, and 21. Fig. 4 *C* and *D* presents the ratio of the absolute number of molecules from all three chromosomes normalized to copy number 1 for the X chromosome. As shown, the measurements above 50 molecules all yield highly precise relative copy number values. At low numbers of molecules (0.036 ng) uncertainty results because the error associated with sampling an aliquot for dilution is significant. Numerical simulations were performed to estimate the sampling error, and summarized medians along with the 10th and 90th percentiles of the copy number ratios are shown in Fig. 4 *C* and *D* as circles and range bars, respectively. At the most extreme dilutions, where approximately 5, 10, and 15 molecules are expected for the chromosome X, 4, and 21 genes, the deviation in copy number ratio is within the expected sampling error.

Overall, the identity of labels detected on the microarrays and in sequencing are in good agreement, with only a small subset of labels unique to each process (Fig. S4*C*). Despite a high sequencing sampling depth (Table S3), a small number of labels with high microarray intensity appear to be missing or underrepresented in the sequencing results. In contrast, labels that appear in high numbers in the sequencing reaction always correlate with high microarray intensities. No trivial explanation could be found for the labels that are missing from any given sequencing experiment. Although underrepresented in some experiments, the same labels appear as present with high sequence counts in other experiments, suggesting that the sequences are compatible with the sequencing reactions. We used PCR as an independent method to investigate isolated cases of disagreement, and demonstrated that the labels were present in the samples used for the sequencing runs (Table S5). Although we can clearly confirm their presence in the sequencing libraries, it is unclear as to why these labels are missing or underrepresented in the sequencing reads.

To test the stochastic behavior of label selection, we pooled the results of multiple reactions at low target concentrations (0.36 and 0.036 ng), where the probability that a label will be chosen more than once is small. Fig. S6 shows that the number of times each label is used closely follows modeling for 1,064 label observations from microarray counting. Furthermore, because each end of a target sequence chooses a label independently, we can compare the likelihood of the same label occurring on both ends of a target at high copy numbers. Table S2, columns 10–11 present the experimentally observed frequency of labels occurring in common across both ends of a target and their expected frequency from numerical simulations. No evidence of nonstochastic behavior is observed in these data.

Discussion

It is interesting to contrast the attributes of stochastic labeling with other quantitative methods. Microarray and sequencing technologies are commonly used to obtain the relative abundance of multiple targets in a sample. In the case of microarray analysis, intensity values reflect the amount of hybridization bound target and can be used to compare to the intensity of other targets in the sample. In the case of sequencing, the number of times a sequence is found is compared to the number of times other sequences are found. Although the techniques differ by using intensity in one case and a digital count in the other, they both provide relative comparisons of the number of molecules in solution. To obtain absolute numbers, quantitative capture of all sequences would need to be assured, and distortions due to amplification biases understood; however, in practice the efficiency of capture and/or distortions due to amplification biases with sequencing or other counting approaches (9–12) are unknown. With stochastic labeling, high-efficiency enzymatic reactions coupled with a large molar excess of labels ensures quantitative labeling, and after amplification, threshold detection diminishes the effects of distortions due to amplification bias.

Digital PCR is an absolute counting method where solutions are stochastically partitioned into multiwell containers, typically until there is an average probability of less than one molecule per two containers, then detected by PCR (3). This condition is satisfied when, $1 - P_0 = (1 - e^{-n/c}) = \frac{1}{2}$; where P_0 is the probability that a container does not contain any molecule, n is the number of molecules and c is the number of containers, or n/c is 0.693. If quantitative partitioning is assumed, the dynamic range is governed by the number of containers available for stochastic separation. Once the molecules are partitioned, high-efficiency PCR detection gives the yes/no answer and absolute counting is enabled. To vary dynamic range, microfabrication (13) or picoliter droplets (14) can be used to substantially increase the number of containers. Similarly, in stochastic labeling, the same statistical conditions are met when $1 - P_0 = (1 - e^{-n/m}) = \frac{1}{2}$; where m is the number of labels, and one half of the labels will be used at least once when $n/m = 0.693$. The dynamic range is governed by the number of labels used, and the number of labels can be easily increased to extend the dynamic range. The number of containers in digital PCR plays the same role as the number of labels in stochastic labeling and by substituting containers for labels we can write identical statistical equations. Using the principles of physical separation, digital PCR stochastically expands identical molecules into *physical space*, whereas the principle governing stochastic labeling is chemically based and expands identical molecules into *chemical space*.

We have shown that a population of indistinguishable molecules can be stochastically expanded to a population of uniquely identifiable and countable molecules. High-sensitivity threshold detection of single molecules is demonstrated, and the process can be used to count both the absolute and relative number of molecules in a sample. The method should be well-suited for determining the absolute number of multiple target molecules in a specified container, such as high-sensitivity clinical assays, or for determining the number of transcripts in single cells. For example, counting on the order of 300,000 molecules of the approximately 30,000 gene transcripts in the human genome in any given cell could be achieved with high efficiency using several thousand labels. We estimate that this experiment should require about 10–30 million sequencing reads, falling within the capacity of modern sequencing devices (the number of reads required using sequencing technology depends on the number of molecules, not the diversity of labels). The number of array elements required depends on the number of different types of molecules times the diversity of labels, or $\sim10^7$ array elements in this example, also within range of current technology. The approach should also be compatible with other molecular assay systems. For example, antibodies could be stochastically labeled with DNA fragments and those that bind antigen harvested. After amplification, the number of labels detected will reveal the original number of antigens in solutions. In the examples shown here, DNA is used as a chemical label because of the great diversity of sequences available, it can be amplified, and because it is easily detectable. In principle, any stochastic chemical change could be used as long as it can be easily detected and generates sufficient diversity for the desired application.

Materials and Methods

DNA Samples. Genomic DNA isolated from cultured B-Lymphocytes of a male Caucasian with Trisomy 21 was purchased from Coriell Institute for Medical Research (Catalog no. GM01921). The DNA quantity was determined by PicoGreen (Invitrogen) measurements using the lambda phage DNA provided in the kit as reference standard. DNA quality was assessed by agarose gel electrophoresis.

BamHI Digestion and Ligation to Labels. Genomic DNA was digested to completion with BamHI [New England BioLabs (NEB)] and ligated to a pool of adaptors consisting of an equal concentration of 960 distinct labels (Fig. 3*A*). Each adaptor consists of a universal PCR priming site, a 14-nt long label sequence, and a BamHI overhang (Fig. S3). The sequence of the labels (Table S1) was selected from an all-possible 4^{14} nucleotide combination to be of similar melting temperature, minimal self-complementation, and maximal differences between one another. Homopolymer runs and the sequence of the BamHI restriction site were avoided. Oligonucleotides were synthesized (Integrated DNA Technologies) and annealed to form double-stranded adaptors prior to pooling. For ligation, the digested DNA was diluted to the desired quantity and added to 100 pmol (equivalent to 6×10^{13} molecules) of pooled label adaptors, and 2×10^{3} units (equivalent to 1×10^{16} molecules) of T4 DNA ligase (NEB) in a 30 μL reaction. The reaction was incubated at 20 °C for 3 h until inactivation at 65 °C for 20 min.

Adaptor PCR. Adaptor-ligated fragments were amplified in a 50 μL reaction containing 1X TITANIUM Taq PCR buffer (Clontech), 1M betaine (Sigma-Aldrich), 0.3 mM dNTPs, 4 μM PCR004StuA primer (Fig. S3), 2.5 units Taq DNA Polymerase (Affymetrix), and 1X TITANIUM Taq DNA polymerase (Clontech). An initial PCR extension was performed at 72 °C for 5 min, 94 °C for 3 min, followed by 5 cycles of 94 °C for 30 s, 45 °C for 45 s, and 68 °C for 15 s. This step was followed by 25 cycles of 94 °C for 30 s, 60 °C for 45 s, and 68 °C for 15 s and a final extension step of 68 °C for 7 min. PCR products were assessed with agarose gel electrophoresis (Fig. S4) and purified using the QIAquick PCR purification kit (Qiagen).

Circularization. The purified PCR product was denatured at 95 °C for 3 min prior to phosphorylation with T4 polynucleotide kinase (NEB). The phosphorylated DNA was ethanol precipitated and circularized using the CircLigase™ II ssDNA Ligase Kit (Epicentre). Circularization was performed at 60 °C for 2 h followed by 80 °C inactivation for 10 min in a 40 μL reaction consisting of 1X CircLigase™ II reaction buffer, 2.5 mM $MnCl_2$, 1M betaine, and 200U CircLigase™ II ssDNA ligase. Noncircularized DNAs were removed by treatment with 20U Exonuclease I (Epicentre) at 37 °C for 30 min. Remaining DNA was purified with ethanol precipitation and quantified with OD_{260} measurement.

Amplification of Gene Targets. Three assay regions were tested, one on each of chromosomes 4, 21, and X. Table S1 lists the genomic location, length, and sequences of these selected fragments. The circularized DNA was amplified with gene-specific primers in a multiplex inverse PCR reaction. PCR primers were picked using Primer3 (http://frodo.wi.mit.edu/primer3) to yield amplicons ranging between 121 and 168 bp. PCR was carried out with 1X TITANIUM Taq PCR buffer (Clontech), 0.3 mM dNTPs, 0.4 μM each primer, 1X TITANIUM Taq DNA Polymerase (Clontech), and approximately 200 ng of the circularized DNA. After denaturation at 94 °C for 2 min, reactions were cycled 30 times as follows: 94 °C for 20 s, 60 °C for 20 s, 68 °C for 20 s, and a 68 °C final hold for 4 min. PCR products were assessed on a 4–20% gradient polyacrylamide gel (Invitrogen) and precipitated with ethanol.

Array Design. For each gene target assayed, the array probes consist of all possible combinations of the 960 label sequences connected to the two BamHI genomic fragment ends (Fig. S3). An additional 192 label sequences that were not included in the adaptor pool were also included to serve as nonspecific controls. This strategy enables label detection separately at each paired end, because each target fragment is ligated to two independent labels (one on either end).

Array Synthesis. Arrays were synthesized following standard Affymetrix GeneChip manufacturing methods utilizing contact lithography and phosphoramidite nucleoside monomers bearing photolabile 5′-protecting groups. Array probes were synthesized with 5′ phosphate ends to allow for ligation. Fused silica wafer substrates were prepared by standard methods with trialkoxy aminosilane, as previously described (15). After the final lithographic exposure step, the wafer was deprotected in an ethanolic amine solution for a total of 8 h prior to dicing and packaging.

Hybridization to Arrays. PCR products were digested with Stu I (NEB), and treated with lambda exonuclease (Affymetrix). Five micrograms of the digested DNA was hybridized to a GeneChip array in 112.5 μL of hybridization solution containing 80 μg denatured Herring sperm DNA (Promega), 25% formamide, 2.5 pM biotin-labeled gridding oligo, and 70 μL hybridization buffer (4.8M TMACl, 15 mM Tris pH 8, and 0.015% Triton X-100). Hybridizations were carried out in ovens at 50 °C for 16 h with rotation at 30 rpm. Following hybridization, arrays were washed in 30 mM NaCl, 2 mM NaH_2PO_4, 0.2 mM EDTA, pH 7.4 containing 0.005% Trition X-100 at 37 °C for 30 min, and with 10 mM Tris/1 mM EDTA, pH 8 (TE) at 37 °C for 15 min. A short biotin-labeled oligonucleotide (Fig. S3) was annealed to the hybridized DNAs, and ligated to the array probes with *Escherichia coli* DNA ligase (Affymetrix). Excess unligated oligonucleotides were removed with TE wash at 50 °C for 10 min. The arrays were stained with streptavidin, R-phycoerythrin conjugate (Invitrogen), and scanned on the GCS3000 instrument (Affymetrix).

Counting Labels. We set thresholds for the array intensity, or the number of sequencing reads to classify labels as either being used or not (Fig. S4 *A* and *B*). Appropriate thresholds were straightforward to determine when used and unused labels fall into two distinct clusters separated by a significant gap. In situations where a gap was not obvious, the function normalmixEM in the R package mixtools was used to classify labels. This function uses the expectation maximization (EM) algorithm to fit the data by mixtures of two normal distributions iteratively. The two normal distributions correspond to the two clusters to be identified. The cluster of labels with a high value is counted as used, and the other as not used. The average of the minimum and maximum of the two clusters, $(I_{\min}+I_{\max})/2$, was applied as the threshold for separating the two clusters.

Sampling Error Calculation. A sampling error can be introduced when preparing dilutions of the stock DNA solution. This error is a direct consequence of random fluctuations in the number of molecules in the volume of solution sampled. For example, when 10 μL of a 100 μL solution containing 100 molecules is measured, the actual number of molecules in the sampled aliquot may not be exactly 10. The lower the concentration of the molecules in the entire solution, the higher the sampling error, and the more likely the actual abundance in the sampled aliquot will deviate from the expected abundance (e.g., 10 molecules in this example). To calculate sampling errors, we determined the number of molecules for each chromosome target in our stock DNA solution and performed numerical simulations to follow our dilution steps in preparing the test samples (3.62, 1.45, 0.36, and 0.036 ng). To illustrate, if the dilution step is sampling 1 μL of a 25 μL solution containing 250 molecules, we create 25 bins and randomly assign each of the 250 molecules into one of the bins. We randomly choose one bin and count the number of molecules assigned to that bin to simulate the process of sampling 1/25th of the entire solution. If a serial dilution was performed, we would repeat the simulation process accordingly. For each dilution, the observed copy number ratios of chromosome 4:X or 21:X for 10,000 independent trials are summarized as observed medians, along with the 10th and 90th percentiles and shown in Fig. 4 *C* and *D*.

Validation by DNA Sequencing (First SOLiD Run). DNA targets that were used for hybridization to arrays were converted to libraries for sequencing on the SOLiD instrument (Applied Biosystems). P1 and P2 SOLiD amplification primers were added to the DNA ends through adaptor ligation and strand extension from gene-specific primers flanked by P1 or P2 sequences (Fig. S5). Each sample received a unique ligation adaptor harboring a four-base encoder (Table S1)) that unambiguously identifies the originating sample of any resulting read. Individual libraries were prepared for each sample, and quantified with PicoGreen before equal amounts of each sample was combined into a single pooled library. DNA sequencing was performed on SOLiD v3 by Cofactor Genomics. A total of approximately 46 million 50 base reads were generated. Each read is composed of three segments, including corresponding to the sample encoder, label sequence, and gene fragment (Fig. S5). We removed reads if uncalled color bases were present, the average quality value (AQV) of the whole read <10, the AQV of the sample encoder <20, or the AQV of the label sequence <18. Forty percent of the raw reads were removed. Filtered reads were mapped to reference sequences using the program Short Oligonucleotide Color Space (http://solidsoftwaretools.com/gf/project/socs/) with a maximum tolerance of four color mismatches between the first 45 color bases in each read and reference sequences (the last five color bases on 3′end of each read were trimmed in alignment). About 64.3% reads were uniquely mapped to reference sequences, of which 89.5% (16 million) have high mapping quality, i.e., with no mismatch in the sample encoder and at most one mismatch in the label sequence. These high-quality reads, accounting for approximately 35% of the total reads generated, were used in subsequent counting analysis.

Sequencing Replication (Second SOLiD Run). An aliquot of the exact same DNA library originally sequenced by Cofactor Genomics was subsequently resequenced by Beckman Coulter Genomics. Approximately 50 million 35 base reads were generated and processed following the same rules. Approximately 61% of the raw reads passed quality filters, of which 81% uniquely mapped to a reference sequence with a maximum tolerance of three color mismatches. (An adjusted mismatch tolerance was applied in the alignment step to account for the shorter length of these reads.) Of the mapped reads, 91% (22.5 million) are of high mapping quality, i.e., with perfect match in the sample encoder and at most one mismatch in the label sequence. These high-quality reads (45% of the total raw reads generated) were used for counting analysis. Table S3 lists the number of high-quality reads from the two SOLiD sequencing runs.

PCR Validation. PCR was used to detect the presence of 16 label sequences (Table S5), which were either observed as high or low hybridization intensity on microarrays, and observed with either high or low numbers of mapped reads in SOLiD sequencing. We PCR-amplified the Chr4 gene target with three dilutions (0.1, 1, and 10 pg) of the 3.62 ng NA01921 sample, using the DNA target that was hybridized to microarrays, or the prepared SOLiD library template. PCR products were resolved on 4% agarose gels and fluorescent DNA bands were detected after ethidium bromide staining.

ACKNOWLEDGMENTS. We thank Professor Lubert Stryer for illuminating discussions on statistics and chemical diversity, and Professor Stephen Quake for critical comments on the manuscript.

1. Kalinina O, Lebedeva I, Brown J, Silver J (1997) Nanoliter scale PCR with TaqMan detection. *Nucleic Acids Res* 25:1999–2004.
2. Sykes PJ, et al. (1992) Quantitation of targets for PCR by use of limiting dilution. *BioTechniques* 13:444–449.
3. Vogelstein B, Kinzler KW (1999) Digital PCR. *Proc Natl Acad Sci USA* 96:9236–9241.
4. Hug H, Schuler R (2003) Measurement of the number of molecules of a single mRNA species in a complex mRNA preparation. *J Theor Biol* 221:615–624.
5. Gregory TR, et al. (2007) Eukaryotic genome size databases. *Nucleic Acids Res* 35: D332–338.
6. Karrer EE, et al. (1995) In situ isolation of mRNA from individual plant cells: Creation of cell-specific cDNA libraries. *Proc Natl Acad Sci USA* 92:3814–3818.
7. Wagner A, et al. (1994) Surveys of gene families using polymerase chain reaction: PCR Selection and PCR Drift. *Syst Biol* 43:250–261.
8. Makrigiorgos GM, Chakrabarti S, Zhang Y, Kaur M, Price BD (2002) A PCR-based amplification method retaining the quantitative difference between two complex genomes. *Nat Biotechnol* 20:936–939.
9. Kurimoto K, Yabuta Y, Ohinata Y, Saitou M (2007) Global single-cell cDNA amplification to provide a template for representative high-density oligonucleotide microarray analysis. *Nat Protoc* 2:739–752.
10. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5:621–628.
11. Fortina P, Surrey S (2008) Digital mRNA profiling. *Nat Biotechnol* 26:293–294.
12. Harris TD, et al. (2008) Single-molecule DNA sequencing of a viral genome. *Science* 320:106–109.
13. Fan HC, Blumenfeld YJ, El-Sayed YY, Chueh J, Quake SR (2009) Microfluidic digital PCR enables rapid prenatal diagnosis of fetal aneuploidy. *Am J Obstet Gynecol* 200:543.e1–543.e7.
14. Beer NR, et al. (2007) On-chip, real-time, single-copy polymerase chain reaction in picoliter droplets. *Anal Chem* 79:8471–8475.
15. Fodor SPA, et al. (1991) Light-directed, spatially addressable parallel chemical synthesis. *Science* 251:767–773.