BMC Genomics

METHODOLOGY ARTICLE Open Access

Designing deep sequencing experiments: detecting structural variation and estimating transcript abundance

Ali Bashir*[1], Vikas Bansal[2] and Vineet Bafna*[1]

Abstract

Background: Massively parallel DNA sequencing technologies have enabled the sequencing of several individual human genomes. These technologies are also being used in novel ways for mRNA expression profiling, genome-wide discovery of transcription-factor binding sites, small RNA discovery, etc. The multitude of sequencing platforms, each with their unique characteristics, pose a number of design challenges, regarding the technology to be used and the depth of sequencing required for a particular sequencing application. Here we describe a number of analytical and empirical results to address design questions for two applications: detection of structural variations from paired-end sequencing and estimating mRNA transcript abundance.

Results: For structural variation, our results provide explicit trade-offs between the detection and resolution of rearrangement breakpoints, and the optimal mix of paired-read insert lengths. Specifically, we prove that optimal detection and resolution of breakpoints is achieved using a mix of exactly two insert library lengths. Furthermore, we derive explicit formulae to determine these insert length combinations, enabling a 15% improvement in breakpoint detection at the same experimental cost. On empirical short read data, these predictions show good concordance with Illumina 200 bp and 2 Kbp insert length libraries. For transcriptome sequencing, we determine the sequencing depth needed to detect rare transcripts from a small pilot study. With only 1 Million reads, we derive corrections that enable almost perfect prediction of the underlying expression probability distribution, and use this to predict the sequencing depth required to detect low expressed genes with greater than 95% probability.

Conclusions: Together, our results form a generic framework for many design considerations related to high-throughput sequencing. We provide software tools http://bix.ucsd.edu/projects/NGS-DesignTools to derive platform independent guidelines for designing sequencing experiments (amount of sequencing, choice of insert length, mix of libraries) for novel applications of next generation sequencing.

Background

Massively parallel sequencing technologies provide precise digital readouts of both static (genomic) and dynamic (expression) cellular information. In genetic variation, whole genome sequencing uncovers a complete catalog of all types of variants including SNPs [1] and structural variations [2]. Transcript sequencing [3,4], small RNA sequencing and CHip-Seq [5] allow a measurement of dynamic cellular processes. These technologies provide unprecedented opportunities for genomics research but also pose significant new challenges in terms of making the optimal use of the sequencing throughput. The individual laboratory might not be equipped to provide correct, and cost-effective designs for the new experiments. By 'design', we refer to questions such as "How much sequencing needs to be done in order to reliably detect all structural variations in the sample to a resolution of 400 bp?" Confounding this further is the proliferation of a large number of sequencing technologies, including three widely used platforms, Roche/454 [6], Illumina [1] and ABI SOLiD [7], and others such as Pacific BioSciences [8] and Helicos [9,10]. These technologies offer the end-user a bewildering array of design-parameters, including cost per base, read-length, sequencing error rates, clone/insert lengths, etc. It is not straightforward to make a rea-

* Correspondence: abashir@ucsd.edu, vbafna@cs.ucsd.edu
Dept. of Computer Science and Engineering, UC San Diego, La Jolla, CA, USA
Full list of author information is available at the end of the article

BioMed Central

© 2010 Bashir et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

soned choice of technology and design-parameters in conducting a particular experiment. Likewise, the technology developers are faced with difficult choices on which parameters to improve in future development.

For any particular application, the goal of the researcher is to achieve the desired objective in a cost-effective manner. For example, in genome resequencing, the primary objective is the sensitive and accurate identification of various forms of sequence variants. Accurate SNP detection can be achieved even using short 36 bp Illumina reads [1]. However, for other applications such as de novo assembly of genomes, longer reads are significantly better than short reads [11]. RNA-seq is a novel application of sequencing to determine the expression levels of different mRNA transcripts in the cell [12]. However, the exponential variability in transcript expression levels poses new design questions regarding the required depth of sequencing to sample low abundance transcripts. Resolving such design questions can allow one to expand the scope of next-generation sequencing in novel directions. In this paper, we address and resolve some of the common design questions relating to structural variation and transcript profiling.

Structural variation

Structural variations (SVs) refer to events that rearrange a genome (*query*) relative to a reference genome [13] and include deletions, insertions, inversions and translocations of genomic regions. Paired-end Sequence Mapping (PEM) [14,15] represents a powerful approach to detect such events. In PEM, the ends of a large number of randomly selected inserts (clones) from the genome of an individual (query) are sequenced, and mapped to a reference genome. Inserts which map aberrantly to the reference genome in distance or orientation form an "invalid pair" and suggest an SV [14]. The general approach underlying PEM is illustrated in Fig. 1a-d. A number of recent informatics tools have been developed for the systematic detection of structural variation using the PEM framework [16,17].

Modeling SV detection

As detailed below, and in Figures 1a-d, SVs often involve the creation of *breakpoints*: a pair of coordinates (a, b) in the reference genome, that are brought together to a single location ζ in the query. Consider the *deletion* event in Figure 1a. A reference segment of length $l = b - a + 1$ is absent in the query, relative to the reference. For the breakpoint (a, b) to be detected a paired-end insert must span ζ. Note that the insert-size is not fixed, but distributed tightly around a mean ($L \pm \sigma$). Deletion is confirmed if the breakpoint is spanned and $l >> \sigma$. Typically, $\sigma << L$ so we simply require that $l > L$, which is sufficient but not strictly necessary.

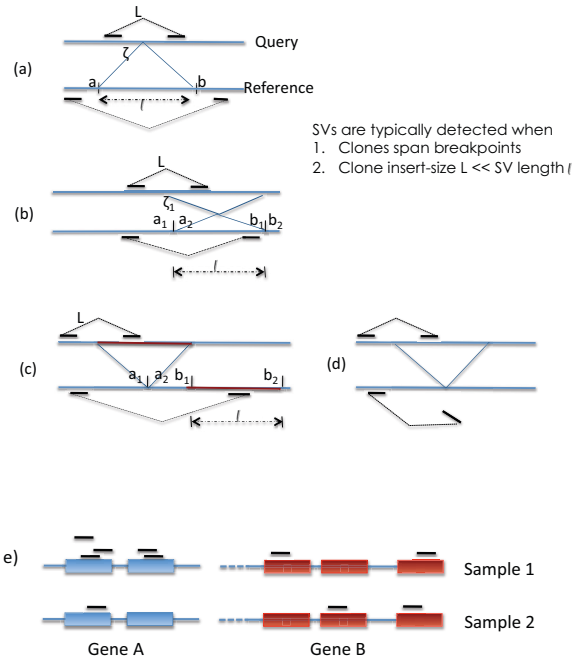

Figure 1 Applications of next-generation sequencing. Figure 1: (a) Structural variant (Deletion) detection using genomic PEM. The deletion brings disparate points (a, b) together to a fusion point in the query. (b) An inversion event is detected by detecting the two breakpoints (a_1, b_1), and (a_2, b_2) (c) An insertion caused by a translocation is detectable and has > 2 breakpoints. (d) Insertion of novel DNA causes dangling ends, and is harder to detect. (e) RNA-seq to detect gene expression. The number of reads mapped back to each gene indicate its relative abundance between and within a sample.

This approximation illustrates an important difference between 'algorithm design' for SV detection, and experiment design. Using a clever algorithm based on higher coverage, and variation in insert length (σ), it may be possible to detect smaller deletions ($\sigma < l < L$) as well. however, in deciding how much sequencing is done, we simply focus only on $l > L$. This simplification allows us to handle many different types of SV using identical design criteria. The similarity to other cases is described below.

The case of *inversion* is shown in Figure 1b. Here, two breakpoints, (a_1, b_1), and (a_2, b_2) are fused together in the query. Denote the length of the inversion SV as $l = b_1 - a_1 = b_2 - a_2$. The inversion is detected when both breakpoints are detected. As in the case of deletions, either breakpoint is detected when a shotgun insert of the query spans the corresponding fusion point, and has exactly one end-point inside the inversion. We enforce this by requiring that $L < l$, even though the condition is sufficient but not strictly necessary.

The case of *insertion* into a query sequence, relative to the reference, is slightly more complex, and can be bro-

ken into two sub-cases. In the first case (mediated by transpositions, or chromosomal translocations), a distal region b_1, b_2 of the reference genome is inserted at coordinate a_1, creating at least 2 breakpoints ((a_1, b_1), and (b_2, a_2) in Figure 1c). Let $l = b_2 - b_1$ denote the length of the insertion SV. Again, SV detection depends upon the detection of 2 or more breakpoints. If on the other hand, the inserted sequence is not in the reference genome (Figure 1d), then detection is challenging, often involving a *de novo* assembly of the inserted region (see Bentley, 2009, Figure S21 [1]). We do not consider this case further. The case analysis above reveals the following common thread. An SV is characterized by its length l, and a collection of breakpoints. For an SV to be detected,

1. One or more of the SV breakpoints must be detected.
2. For each breakpoint:
   (a) A shotgun insert must span the corresponding fusion point.
   (b) The reads at the ends of the insert must map unambiguously to the ends.
   (c) The insert-size must be dominated by the SV length ($l > L$).

This abstraction clarifies the design questions considerably. While the algorithmic questions must still deal with each SV separately, the design questions focus on breakpoint detection. We consider 2(b, c) first. For any choice of technology, and insert length, the distribution can be empirically computed by looking at concordantly mapped reads. Using this distribution, we can compute the probability of a randomly picked insert having a specific size.

Consider a typical experiment for SV detection. The researcher would like to detect a large fraction of all SVs of length $\geq l$, with high confidence ($\geq 1 - \varepsilon$). They must choose (a) a specific instrument technology; (b) insert-size(s) from the ones available; (c) read-length, and (d) the amount of sequencing. First, the researcher must choose a technology and insert-size constraint, where

$$\Pr\left(\text{Length of arbitrary insert} > l\right) \leq \varepsilon \qquad (1)$$

The choice of a specific read-length is somewhat less important, but the reads must be long and accurate enough to map unambiguously. We model both points by introducing a parameter f, referring to the fraction of reads that map unambiguously. Therefore if N inserts must map unambiguously to satisfy design constraints, then N/f inserts need to be sequenced, on the average. In the remainder, we limit the discussion to detecting breakpoints, considering only the technologies and insert sizes that satisfy the size-constraint (1); and, we assume a mapping parameter f to scale the answers. The issue now is to choose from available insert-sizes, and second, to determine the amount of sequencing. In this paper, we formulate, and resolve design issue 2(a) as:

• *Given a choice of insert-sizes, and parameter ε, compute the amount of sequencing needed to detect* $1 - \varepsilon$ *of all breakpoints in the query genome.*

We address the questions of breakpoint detection conjunction with the related notion of breakpoint *resolution*. With most technologies, a breakpoint detected as a pair of regions ($[a_1, a_2]$, $[b_1, b_2]$), such that $a \in [a_1, a_2]$, and $b \in [b_1, b_2]$. The resolution, defined by $|a_2 - a_1| + |b_2 - b_1|$ refers to the uncertainty in determining (a, b). Good resolution is critical elucidating the phenotypic impact of the variation In an earlier work, we described the use of tightly re solved breakpoints in detecting gene fusion events cancer [18]. This framework was extended to form general geometric approach for detecting structural variants [16]. We reformulate and resolve the question

• *Given a choice of insert-sizes, and parameters ε, s, compute the amount of sequencing needed to detect* $1 - \varepsilon$ *of all breakpoints in the query genome to a resolution of* $\leq s$ *bp.*

Intuitively, the likelihood of detection would be maximized by choosing the largest available insert-size. However, the longer insert-sizes increase the uncertainty in resolving the breakpoint. One result of our paper is an explicit trade-off between detection and resolution. We also derive a formula that computes the probability of resolving a breakpoint to within 's' base-pairs, given a fixed number of shotgun reads from a specific paired-end sequencing technology. Another result of our paper is that it is advantageous to use a mix of insert-sizes. For example, we can show that only 1.5× mapped sequence coverage of the human genome using Illumina (Solexa) can help resolve almost 90% of the breakpoints to within 200 bp using a mix of inserts. All other parameters being equal, we show that the best resolution of a structural variant comes from using exactly two possible insert-lengths: one that is as close as possible to the desired resolution, and one that is as long as technologically possible (with reasonable quality).

In summary, the researcher can use our formulae in designing his experiment to (a) select appropriate insert-sizes; (b), the optimum amount of sequencing for each insert library. A web-based tool based on the above is available.

Transcript sequencing

Transcript sequencing is a direct approach for measuring abundance, and variations involving splicing, and SV mediated gene disruptions, and fusions [3]. In most transcript sequencing methods, RNA is fragmented, and converted into cDNA, which is subsequently sequenced and mapped back to a reference [12]. This protocol has shown

great promise in detecting aberrant splice forms and SVs that lead to gene disruptions, and fusions [4].

Often, transcript sequencing is used for gene expression profiling. See Figure 1e. The significant difference in sampled reads (5 to 1) between Samples 1 and 2 suggests that gene A's expression level has changed between the two samples. In measuring relative abundance, RNAseq mimics older technologies like microarrays. However, sequencing stands alone in being able to compute relative abundance between two distinct transcripts. In sample 1, the difference in read coverage between genes A and B suggest that A is more than twice as abundant as B (assuming A and B are approximately the same length).

Let x_t denote the *true-expression* of transcript *t*, defined as the number of copies of *t* in the sample. Additionally, the transcript is broken into a number of pieces, roughly proportional to its length, l_t. Therefore, we assume that transcript *t* yields $l_t x_t$ copies in the sample [3]. This contrasts with earlier technologies like EST sequencing, which were biased towards the 3' (or 5') end. Let a_t denote the number of sequences sampled from x_t. We denote the *normalized-expression* for *t* (likelihood of a randomly sampled read coming from *t*) by

$$\nu_t = \frac{x_t}{\sum_u x_u} = \frac{\frac{a_t}{l_t}}{\sum_u \frac{a_u}{l_u}}$$

A typical design question for transcript sequencing is to determine the amount of sequencing required to sample a given fraction (Say, 90%) of the expressed transcripts. The question is particularly difficult to answer because different transcripts have vastly different normalized-expression values. Using empirical and analytical observations, we show that the p.d.f of the normalized-expression can be computed using a small sample. Therefore, a researcher can start with an initial sequencing run (< 500 K reads), and use the mapping data to compute the additional amount of sequencing needed. Formally, we resolve the following:

- *Given transcript mappings from a small sample of sequences, and parameter ε, compute the amount of additional sequencing needed to detect* 1 - *ε of all expressed transcripts*.

Our results are based on novel extrapolation for the low abundance genes that are not accurately represented in the sample. They allow the researchers to efficiently allocate resources for large RNA sequencing studies. This is particularly relevant when many related samples are being sequenced and one needs to assess the trade-offs between sequencing depth and sample coverage.

Results and Discussion

Structural Variation

As discussed in the introduction, we can limit the question of SV detection to detection of SV breakpoints. Let breakpoint (*a, b*) in the reference genome fuse to a single point ζ in the query genome. Let P_ζ, denote the probability that an arbitrary breakpoint is detected. Our goal is to derive an expression for P_ζ, given a certain amount of sequencing.

Direct application of breakpoint formulae requires that one selects from insert-sizes that are smaller than the desired SV length. In the following, we work with available inserts, where the mean insert-size ranges from L = 200 bp to L = 10 Kbp. Therefore, a result that says $P_\zeta = 0.9$ can be interpreted to mean that 90% of all breakpoints from SVs of length significantly larger than L Kbp can be detected. These specific values are chosen for illustration purposes only. Identical results apply for smaller or larger SVs, except that we would be limited to choosing from appropriate insert-sizes. All analytical results are derived assuming a fixed value for L. However, all results on real data use the natural variation in insert-size, and show excellent concordance with the analytical results.

Detection-Resolution trade-off

Consider N inserts with fixed insert-size L sampled at random and end-sequenced. For a genome of length G, the clonal coverage $c = NL/G$, describes the expected number of inserts spanning ζ. A breakpoint is detected exactly when at least one insert spans ζ. Therefore, P_ζ, the probability of detecting an arbitrary breakpoint, is given by the Clarke-Carbon Formula [19,20].

$$P(\zeta) = 1 - e^{-c} \tag{2}$$

Equation 2 demonstrates the effect of L and N. Larger values L (among allowable insert-sizes), or the amount of sequencing N improve the probability of detection. However, the greater insert length also creates a greater uncertainty in the location of ζ. Define *resolution-ambiguity* as the size of the region θ (denoted by $|\theta|$) in which ζ is constrained to lie. Order the inserts spanning ζ by their right endpoint. Let A be the distance of the right end point of the leftmost insert to the right of ζ. Then,

$$\begin{aligned} &Pr(A > s \wedge \zeta \text{ is covered}) \\ &\quad = (1 - s/G)^N \left(1 - \left(1 - \frac{L-s}{G}\right)^N\right) \\ &\quad \approx e^{-sc/L} - e^{-c} \end{aligned}$$

We show (see METHODS) that

$$E(A \mid \zeta \text{ is covered}) = \frac{\sum_s Pr(A>s \wedge \zeta \text{ is covered})}{Pr(\zeta \text{ is covered})} = \frac{L}{c} - \frac{L}{e^c - 1}$$

Using symmetry arguments,

$$E(|\Theta| \mid \zeta \text{ is covered}) = 2 \cdot E(A \mid \zeta \text{ is covered}) = \frac{2L}{c} - \frac{2L}{e^c - 1} = \frac{2G}{N} - \frac{2L}{e^{\frac{NL}{G}} - 1} \quad (3)$$

Equations 2, and 3 provide an SV detection versus resolution trade-off. For a fixed number of sequences N, increasing L increases the probability of detection, but also increases the resolution-ambiguity. The effect decreases for large N. To validate this using experimental data, we used the publicly available Illumina generated human reference sequence from NA18507, a Yoruban male [1]. Using the complete data, we computed a set of "true breakpoints" from SVs of length ≥ 2000 (see METHODS).

Next, we collected all inserts with mean insert-size either 200 bp, or 2000 bp. Choosing the number of mapped reads as a parameter N, we collected random sub-sets of N paired-reads, and computed the fraction of true breakpoints detected as well as the expected resolution (see METHODS). Figure 2 illustrates the trade-off between detection and resolution. The plotted-lines correspond to theoretical predictions which do not use variance in insert-sizes. The dark ovals show the experimentally observed values for detection and resolution, which can be compared against the corresponding theoretical values (squares).

Nevertheless, current sequencing capability allows us to detect and resolve a large fraction of breakpoints. For example, with an Illumina run with 2 Kbp inserts and 25 × 10^6 mappable reads one could detect nearly 100% of breakpoints with an average resolution-ambiguity of less than 500 bp.

Mixing insert lengths

Many of the next generation sequencing technologies offer a variety of insert lengths. For example, the ABI SOLiD technology claims a variety of insert lengths ranging from 600 bp to about 10000 bp [21]. Given the trade-off between detection and resolution, we next asked if using a mix of insert lengths could help with detection *and* resolution. To address this, we first derived bounds on the probability of resolving a breakpoint to a desired

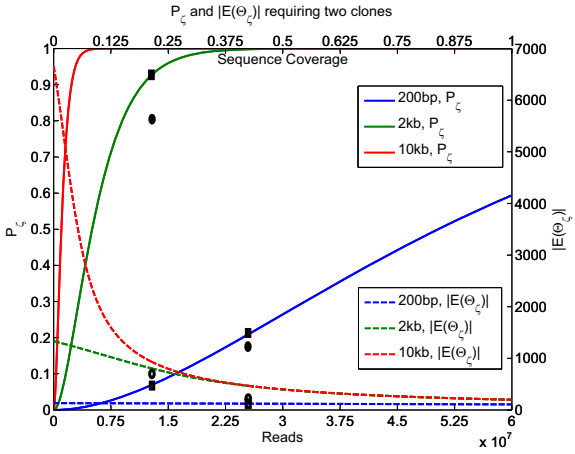

Figure 2 Detection-resolution trade-off. Figure 2: The detection probability, P_ζ (left-axis), increases with increased sequencing (N), as well as insert-length (L). The expected resolution-ambiguity, $|E(\Theta_\zeta)|$ (right-axis), increases with increasing insert-length L. The bottom x-axis shows total number of reads (N) while the top x-axis shows the corresponding sequence coverage (with 50 bp paired reads relative to the human genome). '□' and '●' correspond to the expected and observed values, respectively, for specific values of N and L.

level of resolution using a mix of two insert lengths. Suppose we generate N_1, N_2 reads, respectively from insert libraries of lengths L_1, L_2. Then, for an arbitrary s (see METHODS)

$$\Pr(|\Theta_\zeta| \le s) = \begin{cases} 1 - e^{-s(N/G)}\left(1 + s\frac{N}{G}\right) \\ (*\text{if } s < L_1 *) \\ 1 - e^{-c_1}e^{-s(N_2/G)}\left(1 + s\frac{N_2}{G}\right) \\ (*\text{ if } L_1 \le s < L_2 *) \\ 1 - e^{-c} \\ (*\text{ if } s > L_2, c = \frac{N_1L_1 + N_2L_2}{G} *) \end{cases} \quad (4)$$

Note that the resolution-ambiguity $|\Theta| \le L_1$, or $|\Theta| = L_2$ can be obtained using single insert libraries, but the likelihood of resolving between L_1 and L_2 is optimized by using an appropriate mix of the two libraries. Analogous equations can be derived when two overlapping inserts or more are required to detect a breakpoint.

Figure 3 illustrates this principle using publicly available Illumina generated human reference sequence from NA18507, a Yoruban male [1], assuming one had chosen to split a single run (flow cell) between 2 insert-sizes. As described earlier, we first used the complete data to com-

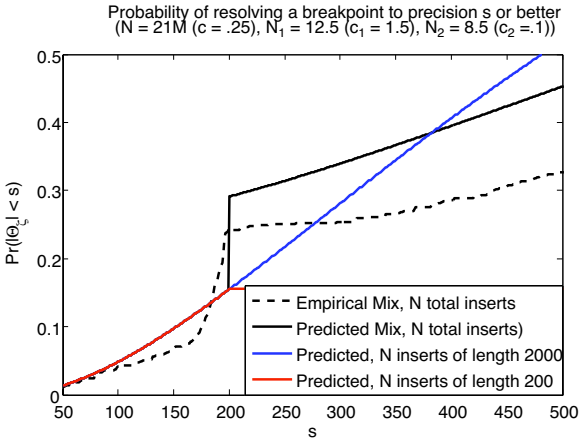

Figure 3 Combination of insert lengths for breakpoint detection. Figure 3: The solid lines indicated the theoretically predicted probabilities as described in METHODS. The dotted line indicates an empirical mix, using 4 lanes each from 200 and 2000 base pair insert runs. *N* (*c*) correspond to the combined number of reads (sequence coverage) uniquely mapped at both ends.

pute a set of "true breakpoints" from SVs of length ≥ 2000 (see METHODS).

Next, we collected all inserts with mean insert-size either 200 bp, or 2000 bp. For a fixed amount of sequencing, we confirmed the theoretically predicted boost in probability of detecting a breakpoint to within a resolution-ambiguity of 200 bp. The results are in Figure 3. The probability is doubled from 0.15 to over 0.29 using a mix of insert libraries. Similar results are obtained for other sequencing studies, such as an ABI SOLiD sequencing with 600 and 2700 length libraries (data not shown). In a further extension of the analysis, we show that to maximize the likelihood of resolving breakpoints to *s*bp, we need only two libraries-one with insert-length *s*, and the other as large as possible (see METHODS). A restatement of these results can be found in Additional file 1. We note that only 1.5× mapped sequence coverage of the human genome using Illumina (Solexa) can help resolve almost 90% of the breakpoints to within 200 bp using a mix of inserts. Similar results were obtained when applied to runs from the ABI SOLiD system [21].

While our analytical results treat the insert sizes as fixed, empirical data very closely approximates the theoretical curve (Figure 3, dotted lines). Though the theoretical model performs better (mostly due to mapping variation resulting from repeat-like genomic regions), the magnitude of the 'boost' at 200 bp is maintained. The concordance between theoretical and experimental results shows the limited effect of insert-length variation.

It is useful to revisit the case of SVs with very small lengths. Mechanisms such as non-homologous end-joining (NHEJ), often gives rise to small insertions and deletions [2], that are valuable as genetic markers. If the event size is smaller than the variance in available insert-size, the event will not be detected by paired end mapping (in the case of deletions and insertions). In these situations, detection is improved by longer reads (such as those available in Roche-454). If single reads are used to detect the fusion point, then there is no ambiguity in resolution. In that case, the design question becomes simple, and the desired number of reads can be computed using the Clark-Carbon formula, and scaled using the mapping parameter *f*.

Transcript sequencing

As transcripts have variable expression, the amount of sequencing needed to detect a transcript is variable. A key design issue is to determine if sufficient sequencing has been performed to sample all transcripts at a certain expression level. For example, in large patient surveys one needs to identify the number of samples that can be sequenced at minimal cost, while ensuring detection of genes at a desired expression level. Similarly, when evaluating a given sample it is important to know whether the required sequencing depth has been reached, or if more sequencing is necessary to detect a given transcript, isoform, or fusion gene. We show here that a relatively low level of transcriptomic sequencing has sufficient information regarding the variability of expression that it can be used to compute the likelihood of a specific transcript being sampled.

While deep sequencing is required to accurately estimate the normalized expression, ν_t, for each transcript, t, a more modest level of sequencing allows us to estimate the distribution of ν values among all transcripts. Formally, define a p.d.f $f(\nu)$ for a randomly sampled transcript to have normalized-expression ν. Consider a transcript sequencing experiment with N reads. If we could estimate ν_t, then

$$\begin{aligned} Pr[t \text{ is sampled} \mid \nu_t = \nu] &= 1-(1-\nu)^H \\ &\simeq 1-e^{-\nu N} \end{aligned}$$

Instead, we propose to use the estimate of f to make predictions about sampling transcripts.

$$\begin{aligned} Pr[t \text{ is sampled}] &= \int_\nu Pr[t \mid \nu_t = \nu] f(\nu) d\nu \\ &= \int_\nu (1-e^{-\nu N}) f(\nu) d\nu \end{aligned} \quad (5)$$

We tested the predictive accuracy of Eq. 5 using data from Marioni et al. [3]. An empirical p.d.f was derived (see METHODS) from the total sequence used in each of two tissue studies (kidney and liver, ~35×10^6 reads each).

Additional file 2a shows the similarity between the empirical distribution of normalized-expression values between the two studies.

We next asked if *f* could be accurately estimated using a lower sequencing depth. If so, this lower level of sequencing can be used to compute the depth of sequencing required to adequately sample all of the transcripts. To test this, smaller sequence-subsets (100 K, and 1 M) were generated by sampling from the complete set. Expression distributions were computed from each subset as shown in Figure 4a. These were then used to compute the probability of transcript detection. Figure 4(b) plots a *detection-curve*, described as the probability of detecting a transcript from the liver sample as a function of its normalized abundance. While predictions made with smaller samples (blue, red solid lines) roughly track the true detection-curve (black line), there is significant bias as low abundance reads are not accurately sampled (Additional file 3).

Previous work has indicated that gene expression distributions typically follow a power-law [22,23]. Nacher et al. extended this idea, accounting for stochastic noise to provide better fits for low expressed genes [24]. We created a novel regression based strategy (METHODS) to correct for the bias, by fitting a power-law to high-expressed genes and using the simplified variant of models proposed by Nacher et al, to accurately approximate genes with low expression levels. The corrected curves (blue, red dotted lines) track the true estimates closely, even when using a sparse set of 100 *K* reads. With 1 million reads, > 90% of the total observed transcripts were sampled. In this data *f* is well-conserved across samples (as seen in kidney and liver, Additional file 2a). For example, the expression p.d.f. for kidney can be used to roughly predict the probability of detection for liver (Additional file 2b). This implies that *f* may not need to be re-estimated independently for related samples.

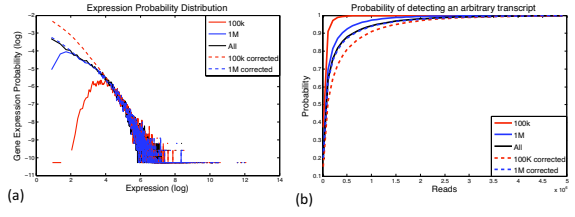

Figure 4 Transcript abundance distribution predicted from sampled reads. Figure 4: (a) Each point on a curve (solid-line) corresponds to the number of transcripts (Y-axis) that had a specific normalized count (X-axis). Note that the number of transcripts small counts relative to coverage drops sharply. For example, the curve for 100K drops at $\log(X) \leq 5$. This undersampling is corrected by the dotted lines. This correction enables the computation of (b), the probability of detecting an arbitrary transcript. The solid lines correspond to predictions made of the empirical (or simulated empirical) distribution. The dotted lines correspond to corrected values from regression (see METHODS). Note the high fit that is obtained after correction, with only 100,000 reads.

Conclusions

We present a number of analytic and empirical results on the design of sequencing experiments for uncovering genetic variation. Our study provides a systematic explanation for empirical observations relating to the amount of sequencing, and the choice of technologies. The theoretical analysis is not without caveats, which are discussed below. Nevertheless, the concordance with empirical data illustrates the applicability of our methods. Some of the results, while not counter-intuitive, provide additional insight. For example, we show that the best design for detecting SV to within 's' bp demands the choice of exactly two insert-lengths, one close to *s*, and the other as large as possible. We explicate the trade-offs between detection and resolution, and provide a method for computing the probability of SV detection as well as the expected resolution-ambiguity for a variety of technology and parameter choices.

Many additional confounding design issues that can be modeled in the context of structural variation. Different technologies have different error rates. This is corrected by introducing a mapping-rate parameter *f*, defined as the fraction of reads that are mapped unambiguously to the reference. Replacing the number of reads *N* by *fN* helps correct somewhat for sequencing errors. New methods have been suggested for dealing with complex scenarios in which it is difficult or impossible to map reads uniquely, such as within recent segmental duplications, using hill climbing [25] or parsimony [26] based approaches which try to minimize the number of observed structural variants. Chimerisms in insert-lengths can be controlled by demanding the use of multiple overlapping inserts. We have extended most analyses to requiring two or more inserts (see METHODS).

An important simplification in our analysis is to treat insert-length as constant. However, choosing a distribution on the insert-length does not influence the expected resolution-ambiguity, only its variance. The variance is important for measuring smaller structural variations. Therefore, experiments that aim to detect small structural variations are constrained to using technologies in which the insert-length variation is significantly smaller than the size of the SV itself. The available technologies are constantly reducing the variance in insert-lengths through better library preparation strategies, which might allow the use of larger insert-lengths in the future.

For transcript sequencing, we address the important question of depth of sequencing, given the large variation in transcript abundance. Our results suggest that estimating the distribution of normalized expression values with modest amounts of sequencing can help address design

questions for transcript sequencing, even when the transcript abundance varies over many orders of magnitude. This approach has a number of caveats, for example, it assumes unbiased sampling of transcripts. Current library preparations have been shown to have biases (such as 3' and 5' depletion) [12] as well as biases towards specific RNAs (specifically small RNAs) within a platform [27]. Additionally, though our results indicate a very good empirical fit on human samples, the assumption of a power-law, or other distribution, may not fit all samples. A number of outstanding questions remain, such as the detection of splicing events, and the resolution of breakpoints. While transcript sequencing is a quick way to detect breakpoints, the location of the breakpoint is confounded by trans-splicing. The issues relating to design can be better resolved only after methods are discovered to resolve breakpoints and predict splicing events based on transcriptome sequencing.

We do not address some important applications of next generation sequencing technologies: the detection of rare (and common) sequence variants in re-sequencing studies. Given the relatively high error rates for some of these technologies, reliable and accurate detection of sequence variants (SNPs) is a challenging problem, and general design principles that would be applicable to all technologies will be addressed in future study. The design of sequencing for 'dark-region' identification (i.e. DNA inserts on the sampled genome that are not in the reference) is not addressed. Lastly, there are practical sample preparation issues which demand consideration. Longer insert-lengths consume more sample for equivalent amount of sequencing. Therefore, if the sample is limited (as in tumors), the best design should also seek to optimize a 'sample-cost' versus detection trade-off.

Technological developments all point to the rapid deployment of personalized genomic sequencing. As large populations of individuals are sequenced, and the sequence is analyzed for a variety of applications, design issues relating to the amount of sequencing, the choice of technology, and the choice of technological parameters become paramount. Our paper helps resolve some of these questions. As current technologies mature and new technologies arise it will be critical to further develop a framework to maximize study efficacy.

Methods

Breakpoint Resolution

The insert coverage is given by $c = N\,L/G$ where N is the number of inserts. A breakpoint (a, b) in the reference genome corresponds to a fusion point ζ in the query genome where the coordinates a, b come together. Let ζ be covered by at least one insert, and let A be the distance of the right end point of the leftmost insert from ζ.

$$\begin{aligned} &Pr(A > s \wedge \zeta \text{ is covered}) \\ &= (1 - s/G)^N \left(1 - (1 - \frac{L-s}{G})^N\right) \\ &\approx e^{-sc/L} - e^{-c} \end{aligned}$$

Therefore,

$$\begin{aligned} &E(A \mid \zeta \text{ is covered}) \\ &= \frac{\Sigma_s Pr(A>s\wedge\zeta \text{ is covered})}{Pr(\zeta \text{ is covered})} \\ &= \frac{\int_0^L (e^{-sc/L} - e^{-c})ds.}{1-e^{-c}} \\ &= \frac{L/c(1-e^{-c}) - Le^{-c}}{1-e^{-c}} = \frac{L}{c} - \frac{L}{e^c - 1} \end{aligned}$$

Using symmetry arguments,

$$\begin{aligned} &E(|\,\Theta\,||\,\zeta \text{ is covered}) \\ &= 2 \cdot E(A \mid \zeta \text{ is covered}) \\ &= \frac{2L}{c} - \frac{2L}{e^c - 1} \end{aligned}$$

Requiring coverage by multiple (≥ 2) inserts,

$$\begin{aligned} &Pr(A > s \wedge \zeta \text{ is covered by} \geq 2 \text{ insert}) \\ &= e^{-cs/L} \\ &.(1 - Pr(\text{no inserts}) - Pr(\text{exactly one insert})) \\ &= e^{-cs/L}\left(1 - e^{-(L-s)c/L} - \frac{c(L-s)}{L} e^{-(L-s)c/L}\right) \end{aligned}$$

$$\begin{aligned} &E(A \mid \zeta \text{ is covered by} \geq 2 \text{ inserts}) \\ &= \frac{\int_0^L (e^{-sc/L} - e^{-c}(c+1) + cs/Le^{-c})ds}{1-e^{-c} - ce^{-c}} \\ &= \frac{L/c(1-e^{-c}) - (c+1)Le^{-c} + cL/2e^{-c}}{1-e^{-c} - ce^{-c}} \\ &= \frac{L}{c} - \frac{cL}{2(e^c - (c+1))} \end{aligned}$$

$$\begin{aligned} &E(|\,\Theta\,||\,\zeta \text{ is covered} \geq 2 \text{ inserts}) \\ &= 2 \cdot E(A \mid \zeta \text{ is covered} \geq 2 \text{ inserts}) \\ &= \frac{2L}{c} - \frac{cL}{e^c - c - 1} \end{aligned}$$

Simulation

A set of "true" breakpoints were chosen by mapping Illumina reads for individual NA18507 (obtained from the NCBI short read trace archive) to build 36.1 of the human genome. ELAND alignment tool, where each end mapped separately to detect SVs. Insert libraries were mapped until > 100× insert coverage was reached, in order to obtain a candidate set. To avoid systematic errors within a library (and over-fitting of the test data) at least three distinct libraries were required to span a breakpoint for it to be considered a "true breakpoint". All SV events greater than 2 Kbp were selected to be the final set is considered to be the TRUEBREAKPOINTSET.

To test the theoretical predictions, a 200 bp and a 2 Kbp library were selected at random. For parameter N, we randomly picked N paired-reads in which both ends mapped uniquely to the genome. A true breakpoint was considered to be detected if at least two inserts spanned it. Thus, the fraction of true breakpoints detected was empirically computed. These numbers were compared against theoretical predictions, obtained using Eq. 2,3 respectively. The resolution $|\Theta_\zeta|$ for each detected breakpoint was computed as follows: for each paired-read that spanned a breakpoint, let x_l denote the distance of its left endpoint from the left end of the right-most clone; let x_r denote the distance of its right end-point from the left most clone. Then, the resolution is given by $L - (x_l + x_r)$. $|\Theta_\zeta|$ was obtained by taking the mean (Θ_ζ) of all overlapping paired-reads. The fraction of "true" breakpoints detected (at least 2 inserts spanning the event) and resolved by these libraries is shown in Figure 3, as a function of N.

Mixing insert lengths

Consider the case where we have two different insert-lengths L_1 and L_2 where $L_2 > L_1$ w.l.o.g. Denote the coverages of the insert libraries as c_1 and c_2. Let $c = c_1 + c_2 L_1/L_2$

$$\begin{aligned} &Pr(A > s \wedge s \le L_1 \wedge \zeta \text{ is covered}) \\ &\quad = (1 - s/G)^{N_1+N_2} \cdot \left(1 - \left(1 - \frac{L_1 - s}{G}\right)^{N_1+N_2}\right) \\ &\quad \approx e^{-sc/L_1} - e^{-c} \\ &\quad = e^{-c_1}\left(e^{-sc_2/L_2} - e^{-c_2}\right) \end{aligned}$$

$$\begin{aligned} &E(A \mid \zeta \text{ is covered}) \\ &\quad = \frac{\Sigma_s \, Pr(A > s \wedge s \le L_1 \wedge \zeta \text{ is covered})}{Pr(\zeta \text{ is covered})} \\ &\quad + \frac{\Sigma_s \, Pr(A > s \wedge s > L_1 \wedge \zeta \text{ is covered})}{Pr(\zeta \text{ is covered})} \end{aligned}$$

$$E(A \mid \zeta \text{ is covered}) = \frac{1}{1 - e^{-(c_1+c_2)}} \cdot \begin{pmatrix} \frac{L_1}{e}(1 - e^{-c}) - L_1 e^{-c} \\ + \frac{L_2}{c_2}(e^{-c} - e^{-c(c_1+c_2)}) \\ - e^{-(c_1+c_2)}(L_2 - L_1) \end{pmatrix}$$

$$E(|\Theta| \,|\, \zeta \text{ is covered}) = 2 \cdot E(A \mid \zeta \text{ is covered})$$

Next, we compute the probability of resolving a breakpoint to within 's' bp. We have 3 cases: i) $s < L_1$; ii) $L_1 \le s < L_2$; and, iii) $s >= L_2$. For $s < L_1$, we extend the analysis of [18], where we showed that $Pr(|\Theta_\zeta| = s) = se^{-\frac{sN}{G}}(1 - e^{-\frac{N}{G}})^2$. Denoting $N = N_1 + N_2$,

$$\begin{aligned} Pr(|\Theta_\zeta| \le s \mid s < L_1) &= (1 - e^{-\frac{N}{G}})^2 \sum_{x=0}^{s} x e^{-\frac{xN}{G}} \\ &= (1 - e^{-\frac{N}{G}})^2 \int_0^s x e^{-\frac{xN}{G}} dx \\ &= (1 - e^{-N/G})^2 \\ &\quad \left(\frac{-G}{N}\left(e^{-sN/G}\left(\frac{G}{N} + s\right) - \frac{G}{N}\right)\right) \\ &= 1 - e^{-sN/G} - \frac{sN}{G} e^{-sN/G} \end{aligned}$$

Note that the results are independent of insert-lengths (or, in fact, whether or not a mix of inserts is being used). However, for the case $L_1 \le s < L_2$, we have to consider the event of an L_1 insert spanning the breakpoint $(1 - e^{-c_1})$ *or* the event of two L_2 inserts spanning ζ with no L_1 inserts spanning ζ Therefore,

$$\begin{aligned} &Pr(|\Theta_\zeta| \le s \mid L_1 \le s \le L_2) \\ &\quad = (1 - e^{-c_1}) \\ &\quad + e^{-c_1}(1 - e^{-sN/G} - \frac{sN}{G} e^{-sN/G}) \end{aligned}$$

The case when $s > L_2$ can be modeled by a single library with $c = (N_1 L_1 + N_2 L_2)/G$.

$$Pr(|\Theta_\zeta| \le s \mid s \ge L_2) = 1 - e^{-c}$$

The equations can be modified to require that at least 2 inserts overlap a breakpoint. Case (i) is unchanged, as it requires 2 inserts. Likewise for the second term in case (ii). We constrain case (ii) to require 2 or more inserts for the first term.

$$\begin{aligned} Pr(|\Theta_\zeta| \le s \mid L_1 \le s \le L_2) &= (1 - e^{-c_1} - c_1 e^{-c_1} e^{-c_2}) \\ &+ e^{-c_1}(1 - e^{-sN/G} - \tfrac{sN}{G} e^{-sN/G}) \\ &= (1 - e^{-c_1} - c_1 e^{-c}) \\ &+ e^{-c_1}(1 - e^{-sN/G} - \tfrac{sN}{G} e^{-sN/G}) \end{aligned}$$

For case (iii), we can extend the generic cluster coverage case.

$$Pr(|\Theta_\zeta| \le s \mid s \ge L_2) = 1 - e^{-c} - ce^{-c}$$

Proof of Optimality of Two Insert Design

We show that it is sufficient to consider exactly two insert lengths for resolving a breakpoint to within 's' bp. We show first that for a given s and N, and a collection of insert-lengths, $Pr(|\Theta_\zeta| = s)$, is maximized using a mixture of ≤ 2 insert lengths.

Assume to the contrary that an optimal mix requires ≥ 3 distinct insert-lengths. This implies that for some insert length L', $L \neq s$, and $L \neq L_M$, where L_M is the maximum available insert-length. In other words, either a) $L' < s$, or, b) $s < L' < L_M$. We consider each case in turn.

$L' < s$: From earlier discussion, the contribution of the inserts with length L' to $Pr(|\Theta| \le s)$ is proportional to coverage (c_1). Replacing inserts of length L' with inserts of length s will increase coverage without changing N, contradicting optimality.

$s < L' < L_M$: Once again, for inserts larger than the desired resolution-ambiguity s, their contribution to $Pr(|\Theta| \le s)$ is completely dependent on coverage. Replacing by a insert of length L_M improves the resolution probability, a contradiction.

An immediate corollary is that the optimal design consists of a mix of two insert lengths, s and L_M. The mix of the two libraries (the ratio N_1/N_2 s.t. $N_1 + N_2 = N$ is fixed) only needs to be optimized for Case (ii).

$$\begin{aligned} Pr(|\Theta_\zeta| \le s \mid L_1 \le s \le L_2) &= (1 - e^{-c_1} - c_1 e^{-c}) + e^{-c_1}(1 - e^{-N_2/G})^2 \\ &\cdot \left(\frac{-G}{N_2} \left(e^{-sN_2/G} \left(\frac{G}{N_2} + s \right) - \frac{G}{N_2} \right) \right) \end{aligned}$$

We compute the optimal mix empirically by iterating over $N_1 \in [0, N]$.

Simulation for mix of inserts

The set of breakpoints, and method for computing mean size of Θ_ζ, followed that of the previous simulation. A single 2 Kbp and 200 Kbp library were analyzed, using 4 lanes from each corresponding flow cell. Clusters of invalid pairs were generated by combining the two reduced libraries.

Transcript Sequencing

Mapped RNA-seq data, generated by Marioni, et al. [3], was obtained from http://giladlab.uchicago.edu/data.html. The genomic mappings were converted to a list of overlapping exons in Refseq. For each transcript, a count of the number of reads sampling it was generated. This enabled the estimation of ν_t which was calculated as described earlier. To obtain smaller data sets, random sampling of the reads was performed and ν_t was re-calculated.

The sample is used to estimate the p.d.f of normalized abundance values, and is shown in Additional file 2. It can be observed that each sample of r reads is accurate for highly expressed genes (normalized expression > $1/r$). Below $1/r$, the chance of sampling a gene is low, and so the p.d.f cannot be estimated accurately. It has been shown empirically that most tissues follow a power law distribution [22,23].

$$\begin{aligned} f(\nu) &= \beta \nu^{\alpha} \\ \log(f(\nu)) &= \alpha \log(\nu) + \log(\beta) \end{aligned}$$

Figure 4a shows a plot of $\log f(\nu)$ vs. $\log(\nu)$. Performing a regression analysis on the line reveals the slope α, and the intercept $\log(\beta)$.

Nacher et al. suggested a stochastic model of gene expression which, in practice, provides a better fit to gene expression data. They provide the equation:

$$P(\nu) = \frac{e^{(-4\delta_D - 1)\ln(\nu+2) - (8\delta_D + 1.2)/(\nu+2)}}{N(\frac{1}{4}\nu^2 + \nu + 1)} \quad (6)$$

Where δ_D is a noise parameter (relating to decay of RNA molecules) and N is a normalization constant [24]. Note, that this equation is approximated by the power law $\nu^{-(4\delta_D+1)}$ at high values of ν.

Generating the fit requires two important steps: fitting a power law at high gene expression and identification of a "reliable point". Note that "high gene expression" can be

maintained for all samples (in our simulations we used $\frac{1}{10{,}000}$th of overall expression). Performing a regression on gene expression values above this threshold provides δ_D. Intuitively, the reliable point can be identified independently for each distribution by determining the point of inflection of the graph $\log(f(v))$ vs $log(v)$; the set of points immediately downstream of the inflection are used to fit Equation 6. One can accurately determine a "reliable point", v_r, by computing the gene expression value at which there is a 95% probability of detecting a transcript, $v_r = \frac{\ln(1-.95)}{-r}$, where r is the number of reads. The corrected p.d.f. utilizes the empirically generated p.d.f after this reliable point, and the theoretical p.d.f before then. It is important to note that the empirical p.d.f. derived using all reads implies that there is a drop off in abundance for very low abundance genes, which the fitting procedure would over-predict. However, this could be an artifact of incomplete sampling and a regression of the full data may provide a better estimate.

Additional material

Additional file 1 The statistics obtained in this research have been implemented in tools, available via web-service and download at http://bix.ucsd.edu/projects/NGS-DesignTools. Figure 3: An optimal mix of inserts (200 bp and 2 kb) improves the probability of resolving breakpoints to a given precision (200 bp). The probability was computed for an experimentally computed set of "true breakpoints" for SVs of length greater than 2000 bp. The bottom x-axis shows total number of reads while the top x-axis shows the corresponding sequence coverage (with 50 bp paired reads relative to the human genome). Different ratios consistently outperform/underperform one another - a single insert size consistently underperforms any ratio of mixed insert sizes.

Additional file 2 The statistics obtained in this research have been implemented in tools, available via web-service and download at http://bix.ucsd.edu/projects/NGS-DesignTools. Figure 1: (a) Distribution of normalized expression from two transcript sequencing experiments. (a) Histogram of *v* in two separate samples. For clarity, the bins are log distributed and the y-axis represents the fraction of total reads. (b) Fraction of detected transcripts from kidney RNA-seq at different sequencing depths. The expected distribution from a different tissue (liver) tracks, but typically over predicts probability of detection at higher sequencing depths.

Additional file 3 The statistics obtained in this research have been implemented in tools, available via web-service and download at http://bix.ucsd.edu/projects/NGS-DesignTools. Figure 2: Estimating the p.d.f of normalized gene expression values. Note that all samples agree except at low levels of detection, where there are insufficient reads. Thus, the 100 K (10^5) reads sample can only estimate the p.d.f accurately after a normalized expression value of 10^{-5}.

Authors' contributions

A.B., V.Bansal, and V. Bafna all conceived the study, designed the experiments, and wrote the manuscript. A.B. and V. Bansal implemented all of the software related to the study.

Acknowledgements

The research was supported by a grant from the NIH RO1-HG004962-01.

Author Details

[1]Dept. of Computer Science and Engineering, UC San Diego, La Jolla, CA, USA and [2]Scripps Genomic Medicine, Scripps Translational Science Institute, La Jolla CA 92037, USA

Received: 23 December 2009 Accepted: 18 June 2010
Published: 18 June 2010

References

1. Bentley DR, *et al.*: **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature* 2008, **456(7218):**53-59.
2. Korbel J, Urban A, Affourtit J, Godwin B, Grubert F, Simons J, Kim P, Palejev D, Carriero N, Du L, *et al.*: **Paired-end mapping reveals extensive structural variation in the human genome.** *Science* 2007, **318(5849):**420.
3. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: **RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.** *Genome Res* 2008, **18:**1509-1517.
4. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM: **Transcriptome sequencing to detect gene fusions in cancer.** *Nature* 2009, **458:**97-101.
5. Johnson D, Mortazavi A, Myers R, Wold B: **Genome-wide mapping of in vivo protein-DNA interactions.** *Science* 2007, **316(5830):**1497.
6. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM: **The complete genome of an individual by massively parallel DNA sequencing.** *Nature* 2008, **452(7189):**872-876.
7. McKernan K, Blanchard A, Kotler L, Costa G: **Reagents, Methods, and Libraries for Bead-based Sequencing.** 2006. [U.S. Patent 084132]
8. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjorn-son K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S: **Real-time DNA sequencing from single polymerase molecules.** *Science* 2009, **323(5910):**133-138.
9. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X, Jovanovich SB, Krstic PS, Lindsay S, Ling XS, Mastrangelo CH, Meller A, Oliver JS, Pershin YV, Ramsey JM, Riehn R, Soni GV, Tabard-Cossa V, Wanunu M, Wiggin M, Schloss JA: **The potential and challenges of nanopore sequencing.** *Nat Biotechnol* 2008, **26(10):**1146-1153.
10. Pihlak A, Baurén G, Hersoug E, Lönnerberg P, Metsis A, Linnarsson S: **Rapid genome sequencing with short universal tiling probes.** *Nat Biotechnol* 2008, **26(6):**676-684.
11. Chaisson MJ, Brinza D, Pevzner PA: **De novo fragment assembly with short mate-paired reads: Does the read length matter?** *Genome Res* 2009, **19:**336-346.
12. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet* 2009, **10:**57-63.
13. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, Oseroff VV, Albertson DG, Pinkel D, Eichler EE: **Segmental duplications and copy-number variation in the human genome.** *Am J Hum Genet* 2005, **77:**78-88.
14. Volik S, Zhao S, Chin K, Brebner J, Herndon D, Tao Q, Kowbel D, Huang G, Lapuk A, Kuo W, *et al.*: **End-sequence profiling: Sequence-based analysis of aberrant genomes.** *Proc Natl Acad Sci USA* 2003, **100(13):**7696-7701.
15. Tuzun E, Sharp A, Bailey J, Kaul R, Morrison V, Pertz L, Haugen E, Hayden H, Albertson D, Pinkel D, Olson M, Eichler E: **Fine-scale structural variation of the human genome.** *Nat Genet* 2005, **37(7):**727-732.
16. Sindi S, Helman E, Bashir A, Raphael B: **A geometric approach for classification and comparison of structural variants.** *Bioinformatics* 2009, **25(12):**i222.
17. Korbel J, Abyzov A, Mu X, Carriero N, Cayting P, Zhang Z, Snyder M, Gerstein M: **PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data.** *Genome Biology* 2009, **10(2):**R23.

18. Bashir A, Volik S, Collins C, Bafna V, Raphael B: **Evaluation of Paired-End Sequencing Strategies for Detection of Genome Rearrangements in Cancer.** *PLoS Computational Biology* 2008, **4(4):**.
19. Clarke L, Carbon J: **A colony bank containing synthetic Col El hybrid plasmids representative of the entire E. coli genome.** *Cell* 1976, **9:**91-99.
20. Lander ES, Waterman MS: **Genomic mapping by fingerprinting random clones: a mathematical analysis.** *Genomics* 1988, **2(3):**231-239.
21. McKernan KJ, *et al.*: **Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding.** *Genome Res* 2009, **19:**1527-1541.
22. Furusawa C, Kaneko K: **Zipf's law in gene expression.** *Cell Phys Rev Lett* 2003, **90(243):**088102.
23. Ueda H, Hayashi S, Matsuyama S, Yomo T, Hashimoto S, Kay S, Hogenesch J, Iino M: **Universality and flexibility in gene expression from bacteria to human.** *Proceedings of the National Academy of Sciences* 2004, **101(11):**3765-3769.
24. Nacher J, Akutsu T: **Sensitivity of the power-law exponent in gene expression distribution to mRNA decay rate.** *Physics Letters A* 2006, **360:**174-178.
25. Lee S, Cheran E, Brudno M: **A robust framework for detecting structural variations in a genome.** *Bioinformatics* 2008, **24(13):**i59.
26. Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC: **Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes.** *Genome Res* 2009, **19:**1270-1278.
27. Linsen S, de Wit E, Janssens G, Heater S, Chapman L, Parkin R, Fritz B, Wyman S, de Bruijn E, Voest E, *et al.*: **Limitations and possibilities of small RNA digital gene expression profiling.** *Nature Methods* 2009, **6(7):**474-476.

doi: 10.1186/1471-2164-11-385
Cite this article as: Bashir *et al.*, Designing deep sequencing experiments: detecting structural variation and estimating transcript abundance *BMC Genomics* 2010, **11**:385

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central