DESIGN OF A HIGH-THROUGHPUT ASSAY FOR ALTERNATIVE SPLICING USING POLYMERASE COLONIES

JEREMY BUHLER, RICHARD SOUVENIR, AND WEIXIONG ZHANG
Department of Computer Science and Engineering
Washington University in St. Louis
One Brookings Drive, St. Louis, MO 63130
Email: {jbuhler,rms2,zhang}@cse.wustl.edu

ROBI MITRA
Department of Genetics
Washington University School of Medicine
660 S. Euclid Ave., St. Louis, MO 63110
Email: rmitra@genetics.wustl.edu

We propose an assay to detect and quantify alternative splicing simultaneously for numerous genes in a pool of cellular mRNA. The assay exploits *polymerase colonies*, a recently developed method for sampling and amplifying large numbers of individual transcript molecules into discrete spots on a gel. The proposed assay combines the advantages of microarrays for transcript quantitation with the sensitivity and precision of methods based on counting single transcript molecules. Given a collection of spots s_i, each containing an unknown splice variant of some known gene G_i, we design a series of hybridizations to short oligonucleotide probes to determine in parallel which exons of G_i are present in every spot s_i. We give algorithms to minimize the cost of such designs.

1 Introduction

Alternative splicing of gene transcripts[1,2] is believed to be a major mechanism by which eukaryotes can amplify the number of distinct proteins produced from a limited number of genes. Estimates of the fraction of alternatively spliced genes in the human genome range from 20% to nearly 60%[3,4]. In several cases, different splice variants of a gene have been shown to play distinct or tissue-specific functional roles[5,6,7]. These facts have driven the development of assays to discover and quantify alternative splicing.

Quantitative detection of alternative splicing aims to measure, for one or more genes, the amounts of each splice variant of that gene present in a pool of RNA. In this work, we focus on splicing events that result in insertion or deletion of one or more complete exons from a transcript. A gene is treated as an ordered list of exons $G = \{E_1 \dots E_n\}$, with each splice variant containing a subset of these exons. We seek to determine which subsets of G describe splice variants present in a sample of mRNA, and to estimate how often each

variant occurs. Although this formulation does not consider variation arising from alternative exon starts or ends, it does encompass a wide variety of possible splice variants for a gene.

The amounts of specific splice variants for one or a few genes can be quantified by, e.g., an rtPCR assay. More challenging, however, is the task of devising a high-throughput assay to quantify all variants of numerous genes at once. Historically, high-throughput splicing assays have relied either on counting splicing events in ESTs[3,4] or on microarray methods in which each spot specifically recognizes a sequence arising from a particular splicing event[8,9,10].

EST-based methods directly count transcripts and so allow precise quantitation of splice variants. Moreover, ESTs can span several exons, so they can reveal correlations between splicing events involving different cassettes of exons. However, EST counting requires large-scale DNA sequencing, and its quantitative accuracy is limited by biases in which transcripts survive the process of EST library construction. In contrast, array-based methods are less resource-intensive and require less processing of the sample RNA. However, the oligonucleotides used on arrays typically target a single boundary between two exons, so that these methods cannot easily detect correlations in combinatorial splicing events. Array-based methods also suffer from limited quantitative accuracy, particularly for rare transcripts.

This work proposes a high-throughput assay to quantitate alternative splicing using *polymerase colonies* ("polonies" for short). Polony-based assays combine EST counting's precise quantitation and detection of combinatorial splicing events with microarray-like RNA preparation, hybridization, and imaging. A polony gel is a collection of up to ten million spots, each containing many copies of a single transcript molecule sampled randomly from a pool of RNA. The gene whose transcript gave rise to each spot can rapidly be determined. Given this information, we show how to design short (7–10 base) oligonucleotide probes to determine which exons are present in each spot on the gel, and how to pool probes so as to minimize the number of hybridizations needed for this determination.

The remainder of this work is organized as follows. Section 2 describes polony technology and proposes our assay to quantify alternative splicing. Section 3 poses the problem of designing oligonucleotide probes to detect all splice variants of a set of genes while minimizing the cost of the assay. Although this problem is combinatorially challenging, we derive a spectrum of solutions to trade off the costs of oligo synthesis and hybridization. Section 4 evaluates designs from our methodology, and Section 5 concludes.

2 Exon Profiling with a Polony Gel

Polony exon profiling is a single-molecule technology for quantifying alternatively spliced mRNAs[11]. We first describe the current form of this technology, which quantifies all isoforms of a single gene. We then suggest an extension to quantify isoforms of multiple genes, up to an entire genome, in a single assay.

Polony exon profiling includes two steps: amplification and hybridization. **Step 1: Amplification.** A dilute cDNA sample is cast into a thin acrylamide gel attached to a microscope slide. Because the sample is dilute, individual molecules are well separated from one another on the slide. Next, PCR is performed in the gel, using primers specific to a gene of interest. Single cDNA molecules are amplified *in situ*; the acrylamide restricts the diffusion of amplification products so that they remain localized near their parent molecules. Each single cDNA molecule produces a discrete polony containing 10^6 to 10^7 identical copies, with each DNA molecule covalently attached[12] to the gel. Over ten million polonies can be amplified on one slide[13].
Step 2: Hybridization. The slide is first denatured and washed so that each polony contains single-stranded DNA. Next, a fluorescently labeled oligonucleotide complementary to the first exon (known or putative) is diffused into the gel. Only polonies amplified from transcripts containing exon 1 will bind the oligonucleotide. The slide is imaged using a confocal laser scanner to identify these polonies. Finally, the gel is prepared for the next round of hybridization by heating the slide to remove the bound probe. The next hybridization is performed with an oligo complementary to exon 2, and so on for all exons. To increase efficiency, hybridizations can be multiplexed using several fluorescent labels.

The outcomes of k successive hybridizations assign each polony a *signature* of k bits. Each 1 bit indicates a successful hybridization to the polony, while each 0 bit indicates absence of hybridization. Each polony's signature specifies the exons in one sampled transcript. For example, in Figure 1, the indicated polony with signature "1100" was amplified from a transcript containing exons 1 and 2 but not exons 3 or 4. The number of polonies with a given signature is proportional to the number of transcripts of the corresponding isoform in the sampled RNA. To quantify (up to sampling error) the abundance of each isoform, we count the number of polonies assigned each signature.

Polony exon profiling has been used to quantify alternative splicing in several genes, including CD44, a gene with 1,024 potential isoforms[11]. The current protocol can realistically be expected to multiplex 10–50 genes, but further multiplexing is unlikely to be feasible for two reasons. First, multiplex PCR does not typically scale beyond 30–50 primer pairs per tube; greater mul-

tiplexing tends to cause primer-dimer artifacts and other mispriming events. Second, the cost of making exon-specific probes for the roughly 30,000 human genes would be prohibitive. At an average of 8.8 exons per gene[14], 264,000 probes would be needed at a cost of roughly $40 per probe.

To address the limits of multiplex PCR and the high cost of exon-specific probes, we propose a modified protocol to quantify splice variants of numerous genes simultaneously using polony technology. Our proposal includes three key changes: (1) Create cDNA so that (as in a typical cDNA library) each transcript is flanked by universal priming sequences. Polony amplification is performed using this single universal primer pair. Hence, every mRNA molecule on the slide now produces a polony. (2) Identify the gene present in each polony by sequencing a few bases from its ends using *fluorescent in situ sequencing*[15]. Sequencing 10–12 bases from each end of a polony's DNA should identify its gene. (3) To reduce the cost of oligo synthesis, use short (7–10 base) oligo probes. Each probe is specific to one exon *within a single gene* but can identify exons in more than one gene, so many fewer probes than exons are needed.

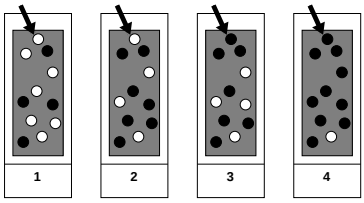

Figure 1. Reading a polony's signature from successive hybridizations against oligos specific to exons 1–4 of a gene. White/black indicate positive/negative outcomes. The indicated polony has signature "1100."

This revised protocol gives rise to two key computational challenges. First, we must choose short exon-specific probes for all genes while realizing the promised savings in synthesis costs. Second, because the probes are too short to guarantee specificity across genes, we must somehow keep probes intended for one gene from producing false positive hybridizations to another. The next section addresses each of these challenges.

3 Distinguishing Splice Variants with Short Oligonucleotides

In this section, we formulate and solve problems arising in the design of short probes for distinguishing splice variants in a polony assay. We first describe formal criteria by which to choose probes for one gene, or for multiple genes simultaneously. We then address the problem of testing all these probes using as few hybridizations as possible. Finally, we identify a tradeoff between the size of the probe set and the number of hybridizations needed and show how to obtain designs that favor one or the other side of this tradeoff.

In what follows, we reduce our problems of interest to two problems known to be NP-hard. In each case, we have also constructed the opposite reduction (omitted for space reasons), showing that we have not set ourselves more difficult tasks than necessary. Also, we select probes directly from the input sequences, recognizing that the real assay must use their reverse complements.

3.1 Assay Designs with Unique Probes

Let G be a gene consisting of exons $E_1, E_2, \ldots, E_n$. A single transcript of G contains a (nonempty) subset of these exons. We wish to construct a collection C of oligo probes with common length ℓ, such that hybridizing each probe in C against a transcript from G unambiguously reveals which exons it contains.

An ℓ-mer probe p is *unique* to exon E_i if it occurs in every splice variant of G that contains E_i and in no variant that does not. Given a set of unique probes $p_1 \ldots p_n$ for each exon of G, we can hybridize each p_i in turn against a transcript of G to determine if it contains exon E_i. This design uses only n probes, the fewest needed to distinguish all $2^n - 1$ splice variants.

Although an exon of G is not guaranteed to contain a unique probe, the following lemma shows how to find such probes when they exist.

Lemma 1. *Let p be an ℓ-mer probe occurring as a substring of exon E_i of gene G. If each exon of G has length at least ℓ, then p is unique to E_i iff (1) p is not a substring of any other exon E_j of G, $j \neq i$; and (2) for any pair of exons E_j and E_k of G, $j < k$, $j, k \neq i$, p is not a substring of the concatenated string $E_j \cdot E_k$.*

Proof. The probe p occurs in any splice variant of G containing E_i. Moreover, in any splice variant lacking E_i, p cannot occur in any of the remaining exons (by Condition 1) or at the boundary between two exons (by Condition 2). A single ℓ-mer cannot span three or more exons if each exon has length $\geq \ell$. Hence, p is unique to E_i.

Conversely, a probe p that occurs uniquely as a substring of E_i cannot appear in a transcript containing only E_j, $j \neq i$ (hence Condition 1). Moreover, p cannot appear across an exon boundary in a transcript containing only exons E_j and E_k, $j, k \neq i$ (hence Condition 2). □

Unless an exon is extremely short, unique probes can generally be obtained by choosing long enough ℓ. In practice, setting ℓ in the range 7–10 yields at least one unique probe (and usually tens of such probes) for well over 90% of predicted coding exons in the human genome.

While the above design produces probe sets for a single gene, we seek to test thousands of genes at once. Naively, we could choose probe sets for each

gene independently; however, such a design is wasteful because a single ℓ-mer can be unique both to (say) exon E_5 within gene G_1 and to exon E_7 within gene G_2. Reusing probes when possible lowers the cost of oligo synthesis. We therefore consider the following optimization problem:

Problem 1. *Let $G_1 \ldots G_m$ be genes for which we want to distinguish all possible splice variants. Each gene G_x has exons $E_{x1} \ldots E_{xn_x}$. For each exon E_{xi}, let $\mathcal{U}_{xi}$ be the set of all ℓ-mer probes unique to E_{xi} within gene G_x. Find a set C of ℓ-mer probes of minimum size such that C contains at least one element of every set $\mathcal{U}_{xi}$.*

A solution to Problem 1 yields a probe set C containing unique probes for every exon of every G_x. Hence, testing each probe in C is sufficient to distinguish all splice variants of these genes. If a solution uses probe p to detect the presence of exon E, we say that p is that solution's *representative* for E. One probe can represent exons of several genes.

Problem 1 is an instance of the hitting set problem[16]. Hitting set is known to be NP-hard but can be approximated to within a factor $\log(\max_{x,i} |\mathcal{U}_{xi}|)$ by a greedy algorithm[17]. A similar combinatorial formulation was used by Nicodème and Steyaert[18] to design multiplex PCR assays. The hitting set problem generalizes (with comparable approximability) to a variant in which each $\mathcal{U}_{xi}$ must be hit at least $r > 1$ times[19]. We use this extension to design probe sets with at least r representatives per exon. Using only one representative per exon provides no way to recover from failed hybridizations that cause false negative outcomes. In contrast, redundancy ensures multiple chances to detect an exon if it is present.

3.2 Pooling to Minimize Hybridization Costs

We can naively test each probe in set C by hybridizing it sequentially. Polony gels can maintain integrity through tens of washings, so sequential hybridization steps are possible. However, the assay cost increases with the number of steps, as does the probability that the gel will tear or detach from the slide. We therefore ask how few steps are needed to test all probes in C.

The danger of testing two or more probes in one hybridization is that one probe may prevent another from unambiguously detecting its intended exon. Figure 2 illustrates this danger in a gene G with exons E_1 and E_2. Probe p represents exon E_1, while probe q occurs in exon E_2 (though it may not represent E_2). If p and q are mixed with the same fluorescent label, the mixture yields a positive result for variants of G that contain E_2 but lack E_1, making p useless for its intended purpose.

We say that a probe p *forbids* a probe q if (1) p represents exon E of some

gene G; and (2) q occurs in any splice variant of G that lacks E. If p forbids q or vice versa, we say that p and q *conflict*. The above example shows that conflicting probes cannot be pooled in one hybridization. Conversely, if p and q do not conflict, then for any exon E represented by one probe, the other probe either does not bind to any variant of E's gene or binds only to variants that contain E. Hence, p and q can safely be pooled.

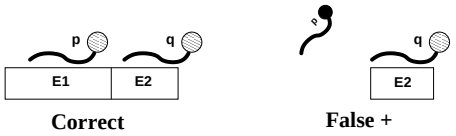

Figure 2. Effect of pooling probe p with probe q when p forbids q. Probe p represents exon E_1 and so yields a positive result iff it is present (left). However, q, which binds to E_2, can cause a false positive result even if E_1 is absent and p does not bind (right).

Any number of non-conflicting probes can be hybridized in a single experiment. Hence, finding large non-conflicting sets of probes reduces the number of hybridizations needed without compromising correctness. We therefore formulate the following problem.

Problem 2. *Let C be a set of unique probes. Divide C into the fewest possible disjoint subsets $C_1 \ldots C_z$ so that no two probes in a subset conflict.*

This problem reduces easily to vertex coloring[20]. Let H be a *conflict graph* whose vertices are the probes of C, such that two vertices are connected iff their probes conflict. In any valid coloring of H, all vertices (probes) of one color are pairwise non-conflicting and so can safely be pooled. Graph coloring, like hitting set, is NP-hard. While approximation algorithms results exist for coloring[21], we use less compute-intensive heuristics to color the conflict graph.

3.3 A Tradeoff in Assay Design

We have formulated a two-step process to design high-throughput alternative splicing assays for polonies: first, select representative probes for all exons of interest; and second, divide these probes into non-conflicting pools. A fundamental tradeoff between these two steps arises because assay designs with fewer probes typically demand more hybridizations.

Consider two probes p and q. As the number of genes with an exon represented by p increases, so too does the chance that q will appear in one of these genes, possibly inducing a conflict. Probe selection seeks to cover all exons with as few probes as possible and so tends to increase the number of exons represented by each probe. As a result, the number of edges in the conflict graph H increases, inducing a likely increase in H's chromatic number and hence in the number of hybridizations needed.

Ideally, our assay design would optimize a joint cost function $f(\pi, \eta)$,

where π is the number of probes and η the number of hybridizations. While we cannot yet directly optimize such a joint cost, we instead seek a spectrum of designs that trade off between π and η, then choose the least-cost design.

We produce a spectrum of designs for probe length ℓ by generalizing probe selection to *weighted hitting set*. This problem variant assigns each probe a weigh and seeks to minimize the total weight of probes chosen. Weighted hitting set can be approximated within the same bound as the unweighted version by a modified greedy algorithm[22]. We will use probe weighting as a heuristic to select probes that induce fewer conflicts and hence are less likely to increase the conflict graph's chromatic number.

For each probe p, we define a *conflict weight* $w_c(p)$. A probe's conflict weight estimates how many other probes would forbid p were it chosen as part of a probe set. We will define $w_c(p, G)$, the conflict weight of p versus a single gene G, and set $w_c(p) = \sum_G w_c(p, G)$. Suppose G has n exons. Then

$$w_c(p, G) = \begin{cases} \text{n} & \text{if } p \text{ occurs non-uniquely in } G \\ \text{n-1} & \text{if } p \text{ is unique to one exon of } G \\ 0 & \text{otherwise.} \end{cases}$$

The rationale for this weighting is as follows. Each exon of G must be represented in the probe set. If p occurs non-uniquely in G, then for each exon E of G, p occurs in a splice variant of G that lacks E. Hence, as described in Section 3.2, p cannot be mixed with E's representative. All n representatives of G's exons (whatever they may be) will therefore forbid p. If p is unique to exon E_i of G, a similar argument shows that p cannot be mixed with any representative *except* that for E_i. Hence, p is forbidden by all but one exon representative for G. Finally, if p never occurs[a] in G, none of the representatives for G's exons forbid p.

To vary the extent to which conflict weighting affects the design, we compute $\overline{w}$, the average conflict weight of all candidate probes, and set $w(p) = \alpha\overline{w} + (1 - \alpha)w_c(p)$ in the hitting set problem. Setting α closer to 0 favors solutions that minimize conflict, while setting it closer to 1 favors solutions with fewer probes. Of course, our weighting scheme is only heuristic, since (1) it overcounts the number of potential conflicts when probes can represent more than one exon, and (2) the conflict count is not a perfect predictor of the conflict graph's chromatic number. However, the results of the next section show that conflict weighting is effective in producing a spectrum of designs that trade off between probe count and number of hybridizations.

[a] We have refined $w_c(p, G)$ to handle cases when p occurs only at boundaries between exons.

4 Empirical Results

In this section, we describe the empirical properties of our assay designs on a comprehensive set of genes predicted by the Twinscan program[23] on NCBI release 31 of the human genome. The test set includes 21,845 multi-exon genes, with an average of 8.5 and a maximum of 80 exons per gene. Although Twinscan's predictions are among the most accurate available, they do not include the difficult-to-predict UTR regions, so the gene sizes and exon counts in our experiments are slightly reduced compared to the real human genome.

We implemented our own software for greedy probe selection and conflict graph generation. For coloring, we used an existing implementation of the `DSATUR` heuristic[24,25]. We avoided probes within six bases of an exon boundary to accommodate small inaccuracies in Twinscan's exon predictions. When enumerating ℓ-mers that cross exon boundaries, we considered splice variants that could link exon E_i with any exon in the range $[E_{i-5}, E_{i+5}]$; however, the exact set of boundary ℓ-mers considered minimally impacted our designs.

An important first test of design methods using small ℓ is whether most exons have unique probes. For ℓ between 7 and 10, we were assigned 97.5% of exons at least one unique probe and 97.4% at least two probes. Discounting initial and terminal exons (which were artificially truncated by Twinscan at the first and last codons), we assigned two probes to over 99% of exons.

Figure 3A illustrates the range of tradeoffs achieved on the test genes between probe set size and number of pools, assuming one unique probe per exon, ℓ from 7 to 10, and α from 0 to 1 in steps of 0.1. All designs use many fewer probes than either the number of exons or 4^ℓ. Longer probes are less likely to cause conflicts, so the number of hybridization pools required decreases as the probe length ℓ increases.

For each $\ell \geq 8$, varying the probe weighting permits tradeoffs as described in Section 3.3. For smaller ℓ, weighting has little effect on the cost of the solution, perhaps because *any* solution that hits every exon has close to 4^ℓ probes and hence unavoidably has a high density of conflicts.

Figure 3B extends the assay design to pick at least two unique probes for every exon. Protection against false negatives increases the number of required probes by a factor of 2–3. At small probe lengths ℓ, this increase brings the number of probes much closer to 4^ℓ, dramatically increasing the number of pools required; however, for larger ℓ, the effects of redundancy on hybridization count are less pronounced.

We now consider the practical utility of our designs. The controlling variable for practicality is likely to be the number of hybridizations, each of which increases the chance that the gel will tear or detach, ruining the experiment.

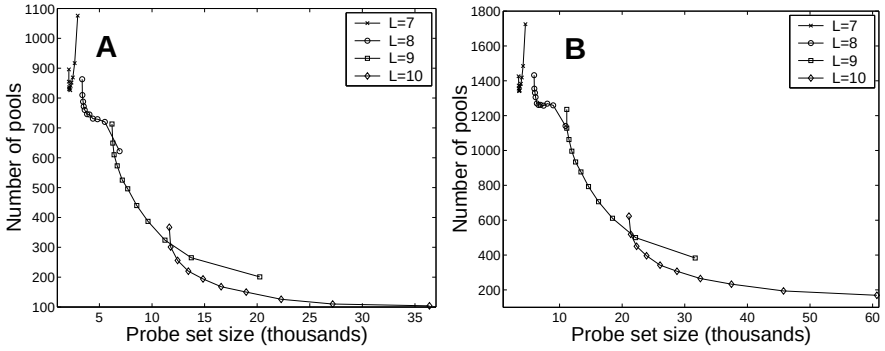

Figure 3. Tradeoffs between number of probes and number of pools. (A) Designs assigning at least one unique probe per exon. (B) Designs assigning at least *two* unique probes per exon to reduce false negatives.

We estimate that realistic assays must use fewer than 100 (preferably fewer than 50) hybridizations. Assuming four-color probe labeling, designs should therefore use at most 400 (preferably at most 200) pools. We achieve such designs for $\ell \geq 9$ while still using many fewer probes than exons.

To make the abstract cost function of Section 3 concrete, we conclude by estimating the actual cost of our assay, assuming it can be fully ramped up to a high-throughput genome-wide survey of alternative splicing. We assume that \$40 will purchase enough of one labeled oligo to test 10,000 gels, and that a high-throughput survey amortizes this cost over the full 10,000 gels. The cost per gel is assumed to be the cost of oligos consumed (0.4 cents per probe, since each probe is used only once per gel) plus roughly \$35 for materials, labor, and machine costs associated with 50 hybridizations. Assuming 45,000 10-mer probes to achieve two unique probes per exon (as in Figure 3B), the cost per gel is \$225. The cost of large-scale polony exon profiling is therefore competitive with microarray and EST-based methods.

5 Discussion

Polony gel technology provides a cell-free method to probe and count individual transcripts from a sample of cellular RNA. It avoids most biases introduced by EST library construction while still yielding a digital readout that can probe every exon in a transcript. Previous work[11] has shown the feasibility of polony gels for assaying splice variants of one or a few genes, but we seek to scale the technology to thousands of genes for high-throughput use.

This work computes two key aspects in the design of high-throughput assays for polony exon profiling: the set of oligonucleotide probes to use, and their pooling into hybridization experiments. Our methods permit systematic selection of redundant probes to limit the rate of false negative outcomes. Our cost estimates show sufficient promise to pursue development of the new assay, which will entail empirically optimizing both the specificity of the oligos and the ability of polony gels to withstand large numbers of washings.

Two issues demanding further exploration are the need for full-length gene predictions and, more generally, the problem of false positives. Full-length genes are necessary to accurately design probe sets, since unexpected sequences in a transcript could cause false positive matches to probes. Accurate prediction of exon structure in UTRs is still an open problem, which means we are unlikely to be able to design probes for many UTR exons. However, our designs can tolerate some degree of overprediction if the UTRs are treated as "forbidden" sequences that, while not themselves probed, restrict the sets of unique probes for the coding exons. We plan to use computational (over)prediction, especially of 5' UTRs, combined with EST evidence from e.g. the NCBI Refseq[26] project to estimate UTR boundaries.

More generally, our designs do not yet address the question of false positive outcomes due to imperfect hybridization. To control this false positive rate, we plan to more accurately predict binding affinity for probes using sequence-specific estimates of their melting temperatures T_m. Our definitions of uniqueness and conflict extend straightforwardly to forbid probes that bind with too high an affinity as well as those that match a sequence exactly. These extensions, combined with our existing provisions to reduce false negative outcomes, will greatly increase our assay's robustness to real-world variations in hybridization.

Acknowledgments

The authors wish to thank Gary Stormo for invaluable discussion and suggestions. Grant support included: JB, NSF DBI-0237903; RS, NIH GM08802 and NSF ITR/EIA-0113618; WZ, NSF IIS-0196057 and ITR/EIA-0113618; and RM, an award from the Whitaker Foundation, St. Louis, MO.

References

1. A. J. Lopez. *Annual Reviews of Genetics*, 32:279–305, 1998.
2. B. Modrek and C. Lee. *Nature Genetics*, 30:13–19, 2002.

3. A. A. Mironov, J. W. Fickett, and M. S. Gelfand. *Genome Research*, 9:1288–93, 1999.
4. Human Genome Sequencing Consortium. *Nature*, 409:860–921, 2001.
5. S. Seino and G. I. Bell. *Biochemical and Biophysical Research Communications*, 159:312–6, 1989.
6. W. C. Horne, S. C. Huang, P. S. Becker, T. K. Tang, and E. Z. Benz, Jr. *Blood*, 82:2558–63, 1993.
7. L. Rowen, J. Young, B. Birditt, A. Kaur, A. Madan, D. L. Phipps, S. Qin, P. Minx, R. K. Wilson, L. Hood, and B. R. Graveley. *Genomics*, 79:587–97, 2002.
8. T. A. Clark, C. W. Sugnet, and M. Ares, Jr. *Science*, 296:907–10, 2002.
9. J. M. Yeakley, J.-B. Fan, D. Doucet, L. Luo, E. Wickham, Z. Ye, M. S. Chee, and X.-D. Fu. *Nature Biotechnology*, 20:353–8, 2002.
10. H. Wang, E. Hubbell, J.-S. Hu, G. Mei, et al. *Bioinformatics*, 19 Suppl 1:315–322, 2003.
11. J. Zhu, J. Shendure, R. D. Mitra, and G. M. Church. *Science*, 2003. To appear.
12. F. N. Rehman et al. *Nucleic Acids Research*, 27:649–55, 1999.
13. R. D. Mitra and G. M. Church. *Nucleic Acids Research*, 27:1–6, 1999.
14. B. Alberts et al. *Molecular Biology of the Cell*. Garland Publishing, 4 edition, 2002.
15. R. D. Mitra, J. Shendure, J. Olejnik, E.-K. Olejnik, and G. M. Church. *Analytical Biochemistry*, 2003. To appear.
16. R. M. Karp. In R. E. Miller and J. W. Thatcher, editors, *Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972.
17. D. S. Johnson. *J. Computer and Systems Science*, 9:256–78, 1974.
18. P. Nicodème and J. M. Steyaert. In *Proc. ISMB '97*, pages 210–3, 1997.
19. D. Hochbaum. *Approximation algorithms for NP-hard Problems*, chapter 3. PWS Publishing, 1997.
20. M. R. Garey, D. S. Johnson, and L. J. Stockmeyer. *Theoretical Computer Science*, 1:237–67, 1976.
21. D. R. Karger, R. Motwani, and M. Sudan. *JACM*, 45:245–65, 1998.
22. V. Chvatal. *Mathmematics of Operations Research*, 4:233–5, 1979.
23. I. Korf, P. Flicek, D. Duan, and M. R. Brent. *Bioinformatics*, 17 Suppl 1:140–8, 2001.
24. D. Brelaz. *CACM*, 22:251–6, 1979.
25. J. C. Culberson. Graph coloring programs, 1997. http://www.cs.ualberta.ca/∼joe/Coloring/Colorsrc/index.html.
26. K. D. Pruitt, K. S. Katz, H. Sciotte, and D. R. Maglott. *Trends in Genetics*, 16:44–7, 2000.