Table 1 | Evaluation of MoDIL on various datasets

			Overlap with known indels[7]			Simulation	
Size	Type	MoDIL	Total	Found	FNR	Recall	Precision
≥20 bp	Insertion	1,336	78	75	0.04	0.85	0.90
	Deletion	3,799	196	187	0.05	0.91	0.89
15–19 bp	Insertion	1,601	119	84	0.29	0.61	0.65
	Deletion	5,333	178	126	0.29	0.78	0.45
10–14 bp	Insertion	936	370	130	0.65	0.44	0.37
	Deletion	3,682	593	227	0.62	0.54	0.27

Number of insertions and deletions of each size identified by MoDIL from Illumina data[6] and the number of previously known indels[7] (total) overlapped by MoDIL predictions (found). We considered indels discovered in ref. 7 but not by us to be false negatives, and the ratio of these as a function of all indels in ref. 7 the false negative rate (FNR). Using a simulated dataset (simulation), we computed the fraction of true indels discovered by MoDIL (recall) and the fraction of predicted indels that were real (precision).

Gaussian distribution with mean μ and standard deviation σ where

$$\mu = \mu_{p(Y)} - \mu_{D_k} \text{ and } \sigma = \sigma_{p(Y)} / \sqrt{n}$$

with n being the number of mate pairs in the distribution, regardless of the shape of $p(Y)$. MoDIL thus uses higher clone coverage to locate progressively shorter indel variation. For proof and thorough description of the algorithms, see **Supplementary Note**.

To evaluate our method, we conducted simulation experiments by implanting known human indels[5] into chromosome 1 and simulating mate-pair data. We used this simulated dataset (51×10^6 mate pairs) to predict indels in the chromosome. MoDIL achieved both precision and recall ≥ 0.85 for indels that were ≥20 base pairs (**Table 1**). We compared MoDIL to tools[1,2] for structural and indel variation discovery using this simulated data (**Supplementary Note**), and no other tool we evaluated identified 15–40-bp indels.

We also applied MoDIL to Illumina whole-genome shotgun-sequencing data[6]. The 3.5×10^9 reads provided 40-fold read and 120-fold clone coverage of the National Center for Biotechnology Information (NCBI) reference human genome. The reads had been mapped[2], with observed insert size $\mu = 208$ bp and $\sigma = 13$ bp. We required each cluster to have at least 20 mate pairs, and used MoDIL to identify 3,981 insertions and 13,147 deletions in the sequenced individual genome relative to the NCBI reference genome. The sizes were 6–118 nucleotides for insertions and 6–66,361 nucleotides for deletions (a full list of predicted indels is available at http://compbio.cs.toronto.edu/modil/). The genome of the same individual was previously sequenced to 0.3-fold coverage using Sanger sequencing[7], allowing for discovery of a small fraction of the short indels in the genome. We estimated the false negative rate of our approach by computing the fraction of these known indels that were missed by our method, but had 20 overlapping clones in our dataset. The sensitivity of our approach varied widely depending on the indel size, but was >95% for indels ≥20 base pairs (**Table 1**).

Because MoDIL does not observe the indels directly the predicted indel size is an approximation of the true size. To verify the accuracy of MoDIL indel size estimates, we compared them to the sizes of overlapping indels from the Mills dataset[5]. The sizes were extremely highly correlated with a large number of indels of ~300–350 bp owing to *Alu* mobile elements (**Fig. 1c**). As expected, the difference between the true size of an indel and our predicted size followed a Gaussian distribution (**Fig. 1d**) with a mean of zero and variance inversely proportional to the number of mate pairs in the cluster. Together, these results indicate that MoDIL accurately recovered smaller variants than was previously possible using high clone coverage of short-read sequencing technologies.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We acknowledge Canadian Institutes of Health Research Catalyst grant to M.B. for funding, and Q. Morris and A. Valouev for useful discussions.

Seunghak Lee[1], Fereydoun Hormozdiari[2], Can Alkan[3] & Michael Brudno[1,4]

[1]Department of Computer Science, University of Toronto, Toronto, Canada. [2]School of Computing Science, Simon Fraser University, Burnaby, Canada. [3]Department of Genome Sciences, University of Washington and the Howard Hughes Medical Institute, Seattle, Washington, USA. [4]Banting and Best Department of Medical Research, University of Toronto, Toronto, Canada. e-mail: brudno@cs.toronto.edu

PUBLISHED ONLINE 31 MAY 2009; DOI:10.1038/NMETH.F.256

1. Li, H., Ruan, J. & Durbin, R. *Genome Res.* **18**, 1851–1858 (2008)
2. Hormozdiari, F. *et al. Genome Res.* (in the press).
3. Korbel, J.O. *et al. Genome Biol.* **10**, R23 (2009).
4. Lee, S., Cheran, E. & Brudno, M. *Bioinformatics* **24**, i59–i67 (2008).
5. Mills, R.E. *et al. Genome Res.* **16**, 1182–1190 (2006).
6. Bentley, D.R. *et al. Nature* **456**, 53–59 (2008).
7. Kidd, J.M. *et al. Nature* **453**, 56–64 (2008).

Limitations and possibilities of small RNA digital gene expression profiling

To the Editor: High-throughput sequencing (HTS) has proven to be an invaluable tool for the discovery of thousands of microRNA genes across multiple species[1,2]. At present, the throughput of HTS platforms is sufficient to combine discovery with quantitative expression analysis allowing for digital gene expression (DGE) profiling[3]. We observed that methods for small RNA DGE profiling are strongly biased toward certain small RNAs, preventing the accurate determination of absolute numbers of small RNAs. The observed bias is largely independent of the sequencing platform but strongly determined by the method used for small RNA library preparation. However, as the biases are systematic and highly reproducible, DGE profiling is suited for determining relative expression differences between samples.

We generated duplicate small RNA libraries using three library-preparation methods (poly(A) tailing[4], modban adaptor (IDT) ligation[5] and Small RNA Expression kit (SREK; Ambion)) from a single sample (rat brain) and sequenced these on Roche 454, AB SOLiD and traditional capillary dideoxy sequencing platforms (**Supplementary Fig. 1**, **Supplementary Note** and **Supplementary Methods**). To assess the impact of the library-preparation method and sequencing platform, we focused on the distribution of known rat 5′ and 3′ microRNA sequences (miRBase v11.0; ref. 6).

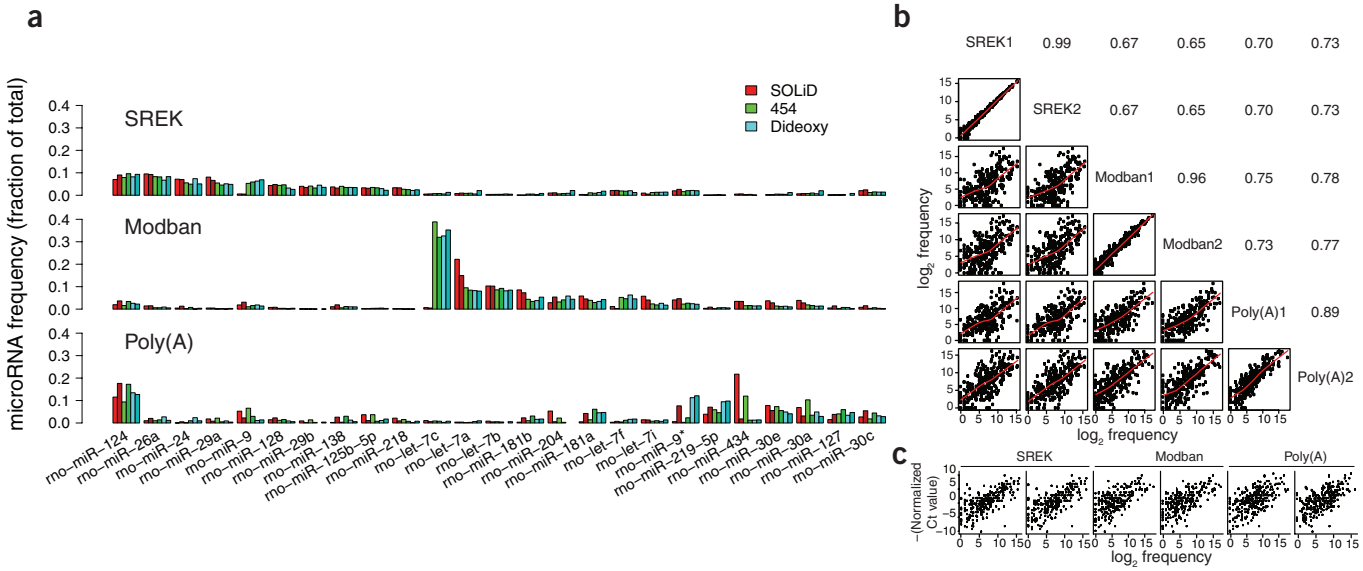

Figure 1 | microRNA expression profiles from different sequencing platforms. (**a**) Relative frequency (specific number of microRNA reads/total number of microRNA reads) of the 10 most frequently cloned microRNAs for each library-preparation method and sequencing platform. Light and dark variants of the colors represent the technical replicates. (**b**) Pairwise comparison of small RNA libraries sequenced on the SOLiD platform. The library identities are indicated, with 1 and 2 being replicates. The $\log_2$ frequencies of individual microRNA sequences are plotted and the corresponding correlation coefficients (Spearman's ρ) are shown. (**c**) Comparison of qPCR analysis with results from the SOLiD platform. Negative median normalized cycle threshold (Ct) values (**Supplementary Methods**) are plotted against normalized microRNA read counts from the DGE profiling experiments.

The 10 most frequently sequenced microRNAs for each library-preparation method showed highly similar relative frequencies on each platform (**Fig. 1a**). The overall correlation between 454 data and SOLiD data was high (Spearman's ρ between 0.79 and 0.95; **Supplementary Fig. 2**).

However, comparing different library-preparation methods revealed large differences in microRNA frequencies. The 10 most frequently sequenced microRNAs of each library-preparation method showed that each approach preferentially captured a distinct set of microRNAs (**Fig. 1a**). These platform-independent biases affected microRNA frequencies over the entire read frequency distribution (**Fig. 1b** and **Supplementary Fig. 3**; ρ, 0.64–0.80 for HTS results). Technical replicates were highly similar (ρ, 0.84–0.99 for HTS results), indicating that these biases have a systematic character (**Fig. 1b**). Furthermore, quantitative PCR (qPCR) (Taqman qPCR; AB), a sequencing-independent approach, showed that DGE profiling frequencies were positively correlated to qPCR results (ρ, ~0.7 for all libraries; **Fig. 1c**). However, none of the library-preparation methods approximated the qPCR results more accurately than any other method.

To better understand the impact of the library-preparation method on microRNA read distribution, we performed both DGE profiling (SREK-SOLiD and modban-Solexa) and qPCR analysis on a sample containing 473 synthetic human microRNA sequences at equal molarity. Both DGE

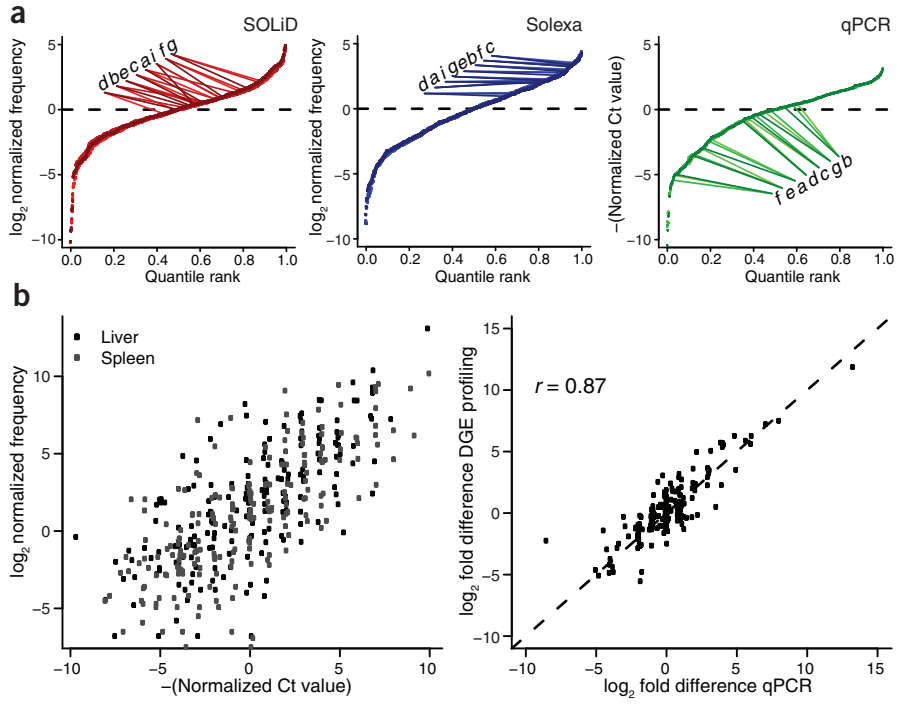

Figure 2 | DGE profiling analysis of a synthetic RNA pool. (**a**) Synthetic microRNAs were measured using SREK library preparation followed by SOLiD sequencing, using modban library preparation followed by Solexa sequencing and using qPCR analysis. The lines highlight the contribution of hsa-let-7a to hsa-let-7g and hsa-let-7i isoforms (represented by the labels a–g and i). Dashed lines indicate median microRNA frequencies or Ct values. (**b**) Differential expression analysis of microRNAs by DGE profiling and qPCR. Small RNAs were amplified by SREK from the indicated tissues followed by SOLiD sequencing and compared to qPCR results. The dashed line represents perfect correlation.

profiling and qPCR recovered a nonuniform distribution of microRNAs (**Fig. 2a**); we observed up to four orders of magnitude difference between the most and least frequently detected microRNAs. Only 61% (SREK-SOLiD) and 52% (modban-Solexa) of the microRNAs varied within a single order of magnitude (**Fig. 2a**). These results showed the inherent quantification biases of both DGE profiling and qPCR based on microRNA sequence, complicating comparison of microRNA amounts in a sample.

Correction of the biological dataset with the frequency bias obtained using the synthetic RNA pool did not improve the correlation between the library-preparation methods (data not shown). We therefore used the synthetic small RNA dataset to explore the potential basis of systematic biases. Although we found clear effects of certain terminal mono- and dinucleotides (**Supplementary Fig. 4**), we could not identify a satisfactory correction model based on primary (RNA sequence) and secondary (for example, folding characteristics) parameters (**Supplementary Fig. 5** and **Supplementary Note**). This might be explained by our observation that even single nucleotide differences influenced the read frequencies (**Supplementary Fig. 6**). RNA ligase preferences[7] may contribute to the observed different terminal nucleotides over the read frequency spectrum. In addition, the reverse-transcriptase reaction as well as the PCR could be a contributor to the bias[8].

To determine whether DGE profiling allows for differential expression analysis, we sequenced small RNA libraries from rat spleen and liver (SREK-SOLiD). In parallel, we analyzed the input RNA by qPCR. Similar to our previous results, qPCR data differed substantially from the read frequencies within a sample (**Fig. 2b**). However, differential expression results between samples obtained by qPCR and DGE profiling were strongly correlated (**Fig. 2b**), showing that the systematic biases do not prohibit the comparison of relative microRNA amounts between samples.

Despite the limitations described here, small RNA profiling by DGE is the method of choice for studying small RNA expression. In contrast to most other existing methods, DGE profiling is hybridization-independent, accurate in discriminating microRNA family members that differ by only a single nucleotide, capable of detecting 5′ and 3′ end variability (for example, isoMirs), and as the approach does not require a priori information, it can be used to simultaneously detect known and discover new biomolecules.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

This work was supported by the Cancer Genomics Center and Netherlands Bioinformatic Center through Netherlands Genomics Initiative (to E.C.), Fred Hutchinson Cancer Research Center and Canary Foundation funding (New Development funds to M.T.), Core Center of Excellence in Hematology Pilot Grant P30 DK56465 (to M.T.), Pacific Northwest Prostate Cancer Specialized Program of Research Excellence Grant P50 CA97186 (to M.T.) and a Rosetta Inpharmatics Fellowship in Molecular Profiling (to S.K.W.). We thank P. Toonen for animal care.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

Sam E V Linsen[1,7], Elzo de Wit[1,7], Georges Janssens[1], Sheila Heater[2], Laura Chapman[2], Rachael K Parkin[3], Brian Fritz[3,6], Stacia K Wyman[3], Ewart de Bruijn[1], Emile E Voest[4], Scott Kuersten[2], Muneesh Tewari[3,5] & Edwin Cuppen[1]

[1]Hubrecht Institute and University Medical Center Utrecht, Cancer Genomics Center, Utrecht, The Netherlands. [2]Life Technologies, Austin, Texas, USA. [3]Human Biology Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. [4]Department of Medical Oncology, University Medical Center Utrecht, Utrecht, The Netherlands. [5]Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. [6]Present address: Illumina, Inc., San Diego, California, USA. [7]These authors contributed equally to this work.
e-mail: e.cuppen@niob.knaw.nl

1. Berezikov, E., *et al. Genome Res.* **16**, 1289–1298 (2006).
2. Ruby, J.G. *et al. Cell* **127**, 1193–1207 (2006).
3. Kuchenbauer, F. *et al. Genome Res.* **18**, 1787–1797 (2008).
4. Berezikov, E. *et al. Nat. Genet.* **38**, 1375–1377 (2006).
5. Lau, N.C., Lim, L.P., Weinstein, E.G. & Bartel, D.P. *Science* **294**, 858–862 (2001).
6. Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A. & Enright, A.J. *Nucleic Acids Res.* **34**, D140–D144 (2006).
7. Romaniuk, E., McLaughlin, L.W., Neilson, T. & Romaniuk, P.J. *Eur. J. Biochem.* **125**, 639–643 (1982).
8. Taube, R., Loya, S., Avidan, O., Perach, M. & Hizi, A. *Biochem. J.* **329**, 579–587 (1998).

RNAiCut: automated detection of significant genes from functional genomic screens

To the Editor: RNA interference (RNAi) is a popular functional genomic technology for identifying genes involved in a biological process. Although higher scores for genes in an RNAi screen suggest more central roles in the pathway, estimating the score threshold separating pathway- or process-relevant hits from noise remains difficult (**Supplementary Table 1**) and is typically done manually.

To overcome this subjective approach, we built a fully automated system, RNAiCut, that objectively and robustly identifies score thresholds from functional genomic data by introducing the use of the connectivity of subgraphs of protein-protein interaction (PPI) networks[1,2]. Unlike some previous work[3], our method does not overlap RNAi and PPI data to find interacting regulators. Instead, its guiding hypothesis is that true positive hits in an RNAi experiment are densely interconnected in the PPI network. For the k highest-scoring genes ($k = 1, 2, 3\ldots$), RNAiCut computes the edge count of the induced subgraph and estimates the P-value of finding a PPI subgraph of at least this size that is induced by k randomly chosen nodes that have the same degrees as these genes (**Supplementary Methods** and **Supplementary Results**). The plot of these P-values as a function of k is typically V-shaped, and we take the global minimum as the score threshold (**Fig. 1**). We used RNAiCut to compute thresholds for several *Drosophila melanogaster* RNAi screens[4] (**Supplementary Figs. 1–10** and **Supplementary Tables 2–3**).

RNAiCut chose successful thresholds, as measured by Gene Ontology (GO)[5] enrichment: the gene lists with above-threshold scores were enriched for functions relevant to the screen, compared to the rest (**Supplementary Table 4**). When the manual screener's threshold was later in the ranked list of hits than the RNAiCut threshold, choosing RNAiCut's threshold may reduce the potentially high number of false positives. When RNAiCut's threshold was later, the GO enrichment for RNAiCut's cutoff was at least as good as for the manually determined cutoff, revealing additional pathway-relevant genes (**Supplementary Results**). Although some of the additional hits identified by RNAiCut may be false positives, analyzing them may be useful given their apparent connectivity to