whole plant. Stem injury is characterized by the oozing of latex from the stem surface about 24–45 cm above the ground level for a week, which leads to the rotting of stem and ultimately death of plants. Symptoms of leaf injury in the form of scorching is visible 10–12 h after the occurrence of frost, which becomes apparent after 20–24 h. Similarly, symptoms of frost injury on growing shoots, flowers and fruits can be seen 6–12 h after frost. Usually, mature fruits are less affected by frost than immature fruits[16,17].

Farmers can protect their standing fruit crops under frosty weather by creation of hot air or smoke screens, covering small fruit plants with straw, dry grass, etc. applying irrigation to maintain soil moisture in the soil profile and foliar spraying of water on the plants twice early in the morning. Wind breaks or shelter belts can also be raised around the existing plantation. This will moderate the abnormal climatic condition to a conducive condition by reducing cold wave damage, reduce wind speed and maintain optimum temperature in and around the orchards. It is also suggested that new orchard growers select low temperature tolerant fruit species and varieties like peach, sweet orange, mandarin, lemon, phalsa, loquat and kagzi lime, with a provision of wind breaks or shelter belts. The Doon Valley is suitable for the cultivation of fruits which are not affected by low temperatures. Citrus species, such as lemon and lime were damaged to an extent of <5% and the overall health of the plants remained unaffected. Loquat is an evergreen, subtropical and underutilized fruit plant, which is being grown on a limited scale by growers. These plants showed no damage and appeared to thrive better under low temperature. Phalsa also a minor, subtropical and winter deciduous fruit plant, remains under dormancy during December–February (chilling period) and did not show any symptoms of damage either on leaves and shoots. The resistance of these hardy plants depends upon species, leaf structure and concentration of solutes, as these reutilize water formed after melting of crystals for initiating metabolic activities in the cells, maintain higher concentration of solutes, cell protoplasm remains more elastic and vacuole splits into a number of small vacuoles to reduce turgor pressure of the cell.

1. Levitt, J., In *Frost Killing and Hardiness of Plants*, Burgers, Minneapolis, Minnesota, USA, 1941.
2. Lyons, J. M., *Annu. Rev. Plant Physiol*., 1973, **24**, 445–466.
3. Terashima, J., Huang, L. R. and Osmond, C. B., *Plant Cell Physiol*., 1989, **30**, 841–850.
4. Sonoike, K., *J. Plant Res*., 1998, **111**, 121–129.
5. Povoles, S. B., Berry, J. A. and Bijorkman, O., *Plant Cell Environ*., 1983, **6**, 117–123.
6. Raison, J. K. and Brown, N. A., *Plant Physiol*., 1989, **91**, 1471–1475.
7. Garber, M. P., *Plant Physiol*., 1977, **59**, 981–985.
8. Geiger, D. R., *Ohio J. Sci*., 1969, **69**, 356–366.
9. Peters, R. A., In *Biochemical Lesions and Lethnal Synthesis*, Macmillan, New York, 1963.
10. Allen, D. J. and Ort, D. R., *Trends Plant Sci*., 2001, **6**, 36–42.
11. Lutze, J. L., Roden, J. S., Holly, C. J., Wolfe, J., Egerton, J. J. G. and Ball, M. C., *Plant Cell Environ*., 2002, **21**, 631–635.
12. Singh, L. B. and Singh, R. N., *Indian J. Hortic*., 1955, **12**, 108.
13. Chattopadhyay, T. K., A *Text Book on Pomology–Tropial Fruits*, 1996, vol. II, p. 128.
14. Pareek, O. P. and Sharma, P. K., *Indian Hortic*., 1992, 35.
15. Sturrock, T. T., A study of the identification of mango varieties in Florida, Mango Studies, Fla Mango Forum, 1951, p. 71.
16. Popenoe, W., *Manual of Tropical and Subtropical Fruits*, Macmillan, New York, 1932.
17. Singh, I. D. and Sirohi, S. C., In Proceedings of the Symposium on Improved Fruit Crops, India, Punjab Agric. Univ., Ludhiana, 9–11 June 1975.

ACKNOWLEDGEMENTS. We thank the Director of Central Soil and Water Conservation Research and Training, Dehradun for encouragement and providing the necessary facilities to carry out the study. We also thank U. C. Tiwari, Fateh Singh, Brijesh Kumar and Raghavendra Singh for help in the field.

Received 30 January 2012; revised accepted 21 February 2012

A. C. RATHORE*
A. RAIZADA
J. JAYA PRAKASH
V. N. SHARDA

Central Soil and Water Conservation Research and Training Institute,
218 Kaulagarh Road,
Dehradun 248 195, India
**For correspondence.*
e-mail: rathoreac@gmail.com

From data repositories to potential biomarkers: application to prostate cancer

The digital nature of RNA-seq technologies offer sensitivities in measurements that were not achievable by hybridization-based technologies; thus fast replacing two decades of genomics technologies, including gene expression microarrays, single nucleotide polymorphism (SNP) arrays, exon arrays and splice variant microarrays. Unlike microarray-based measurements, where one has the limited view of RNA composition through known genes, known splice events, known SNPs, or known exons represented on the microarrays, RNA-seq cracks open the cells exposing diverse RNA phenotypes in its entirety. Thus, RNA-seq technology is becoming a method of choice to measure context-specific translocation[1], gene expression[2], alternative splicing[3], single nucleotide variations (SNVs)[4], non-coding RNAs[5], transcript-induced chimera[6], and micro-RNAs[7] from a single experiment. This, together with the steep drop in sequencing costs, has made data analysis a bottleneck in biology. The NCBI repository is home to a large number of data

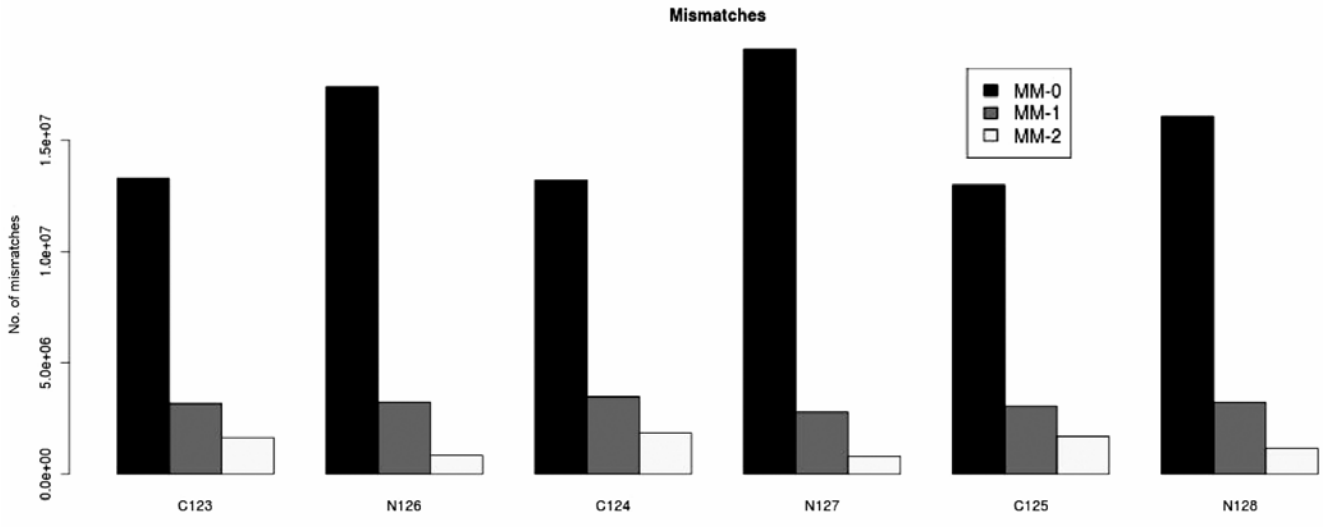

Figure 1. Quality of the six RNA-seq datasets used in the present study.

Table 1. Public dataset used to harness diverse RNA phenotypes in prostate cancer

Individuals	Accession ID tumour/normal	Referred as	Spots T/N
Patient A	SRX027123/SRX027126	T123/N126	30.8M/33.5M
Patient B	SRX027124/ SRX027127	T124/N127	32M/33.1M
Patient C	SRX027125/ SRX027128	T125/N128	65.5M/31M

from RNA-seq experiments. For lack of mature analysis tools, these datasets are underutilized and remain rich with yet undiscovered disease-specific RNA phenotypes. For example, a RNA-seq experiment performed to decipher tumour-specific translocation remains rich with other diverse disease-specific molecular phenotypes[8]. Here, we identify prostate tumour-specific gene expression, splicing regulation and coding mutations from a single RNA-seq experiment containing data from matched tumour–normal samples of prostate tissues from three individuals (Table 1).

Figure 1 shows the quality of RNA-seq data by counting the number of reads with zero, one and two mismatches according to mapping by bowtie onto hg19 genome assembly. The quality is considered acceptable if the percentage of reads with two mismatches is significantly less than reads with one mismatch, etc.

Measuring gene expression from RNA-seq data is a numbers game. The number of RNA reads from a sample that maps to a gene locus on the reference genome provides a measure of expression of that gene. Although the concept is simple, one has to relax the mapping conditions to account for limited mutations in reads compared to the reference genome. However, allowing for too many mutations will result in false mapping of low-complexity reads and also demand increased computational resources. According to a number of publications, a generally accepted compromise is to allow for two mismatches for reads 30–50 bases long and only select those reads that map to a unique site on the reference genome. For comparing expression levels between tumour and normal samples, read counts are transformed to RPKM (Reads Per Kilobase of exon model per Million mapped reads) values, which normalizes for both gene length and variable depth in sequencing. A comparison of the RPKM values for each gene between tumour and normal samples provides a list of genes whose expression is significantly regulated in cancer.

Bowtie, a publicly available, ultra-fast read mapping tool[9] with default parameters was used allowing for two mismatches to map tens of millions of RNA reads from both normal and tumour samples of all three individuals onto the reference human genome (hg19). Only reads mapped to unique genome loci were selected to compute differential expression. Using a pipeline developed in-house, the RPKM values were compared to identify tumour-specific regulation in genes in all three individuals. Table 2 lists the genes that are upregulated and downregulated in a tumour-specific fashion.

As evident from Table 2, a large number of small nucleolar RNAs are upregulated in all three tumour samples. This is despite the fact that one of the normal samples (patient C) had higher percentage of tumour cells according to both our analysis and as confirmed by the author of the data. For lack of discovery technologies, small nucleolar RNAs (snoRNA) have not been observed and reported until the invent of RNA-seq technology. Since RNA-seq experiments reveal the RNA content in its entirety, strong tumour-specific expression of snoRNAs has been reported recently. In Figure 2, strong differential expression of six snoRNAs in all three tumour samples has been compared with their respective matched normals. This observation is consistent with a recent report that many snoRNAs display strong differential expression in prostate cancer[10]. Also, many other snoRNAs are being tested for their diagnostics value in lung cancer. We have recently confirmed that *SNORA76* is highly expressed in two prostate cell lines, including PC-3 and LNCap. Other genes are currently being validated *in vitro* in cell lines and tissue samples.

Among the downregulated genes shown in Table 2, *NTF3* is the most significant, and is in a translocation point known to be active in mesothelioma[11].

Table 2. List of genes upregulated and downregulated in all three tumour samples with a *p*-value of <0.05 and <0.001 respectively, and sorted by fold change in log to base 2

Gene	Acc_ID	Locus	*p*-value	Fold change
Upregulated				
RPL7A, SNORD36B	NR_000017	chr9	0.035303	5.602432
SNORD36C	NR_000016	chr9	0.037933	5.042827
SNORA76	NR_002995	chr17	0.024127	3.899968
RPSA, SNORA62	NR_002324	chr3	0.017608	3.773465
IGFN1	NM_001164586	chr1	0.015584	2.83712
C10orf53, CHAT	NM_182554	chr10	0.028546	2.329292
SNORA45	NR_002977	chr11	0.007746	2.148374
COL19A1	NM_001858	chr6	0.02781	2.050167
RPLP2, SNORA52	NR_002585	chr11	0.033937	1.80922
EIF4A1, SNORD10	NR_002604	chr17	0.041044	1.77611
GAP43	NM_002045	chr3	0.006045	1.514346
SAMD5	NM_001030060	chr6	0.030295	1.330554
UPF0639	NM_001161498	chr14	0.035588	1.101057
PROZ	NM_003891	chr13	0.009283	0.98888
CPZ,GPR78	NM_080819	chr4	0.01826	0.954344
CABP4	uc009yrw.1	chr11	0.009009	0.851423
ACPT	NM_033068	chr19	0.007261	0.830625
PKDCC	NM_138370	chr2	0.013635	0.809042
DTX1	NM_004416	chr12	0.031843	0.789983
SOD3	NM_003102	chr4	0.036871	0.783709
KIAA1958	NM_133465	chr9	0.003509	0.761201
Downregulated				
NTF3	NM_002527	chr12	0.000657	–2.10823
KLF8	NM_007250	chrX	7.99E-05	–1.95958
KRT86	NM_002284	chr12	0.000986	–1.283
KIAA1671	NM_001145206	chr22	0.000798	–1.16388
RBMS2	NM_002898	chr12	0.000676	–0.96922
SFI1	NM_014775	chr22	0.000198	–0.95898
SENP5	NM_152699	chr3	0.000797	–0.89008
HMGCR	NM_000859	chr5	0.000244	–0.87374
ZNF502	NM_033210	chr3	0.000733	–0.82746
ARMC8	NM_015396	chr3	0.000677	–0.82404
DYNC2LI1	NM_016008	chr2	0.000412	–0.82063
PEX11B	NM_003846	chr1	0.00042	–0.66995
SAP30L	NM_001131062	chr5	0.000176	–0.63946

Considering that translocation events are prevalent in prostate cancer, downregulation of this gene is, perhaps, suggestive of translocation event. *KLF8* is the next most significantly downregulated in all three samples with a *p*-value of < 0.0001. However, upregulation of this gene is implicated in tumour growth and invasion[12]. It will be interesting to see if the *KLF8* gene locus (Xp11.21) also participates in a translocation event in prostate cancer. We are currently involved in identifying all translocation events in this dataset.

Differential regulation of gene expression is only one of the ways by which cells turn cancerous and invasive. With more than 94% of the human genes expressing more than one alternative splice forms, measuring gene expression without addressing splice isoforms only offers a low-resolution molecular profile[13]. Furthermore, many splice isoforms of the same gene have been shown to function like an antagonist[14].

Genome-wide analysis of splicing in cancer has revealed that aberrant and regulated splicing is linked to cancer progression[15]. Differential splicing in cancer cannot be confidently measured by only mapping RNA reads to the reference genome. This is because adjacent exons, which are spliced together in RNA transcripts, are separated by large intronic regions on the chromosomes. Although methods to identify/characterize spliced RNA reads by mapping to the reference genome have been developed[16], they miss many known splice events at the cost of discovering novel ones (data not shown).

A database of splice events in human, SPEventH, containing probe sequences representing less than a million observed splice events from a database of non-redundant human splice events has been created with 28 bases on each side of the junction[17]. Restricting the probes to 28 bases on each side is to avoid exonic reads longer than 32mer from fully mapping to a junction and maximizing mapping of the reads that are asymmetric to the junction[13]. Bowtie with default parameters is used to align millions of RNA reads from all six samples from three individuals onto SPEventH. The bowtie output is loaded into the MySQL database to obtain the number of reads mapped uniquely to each splice event. Using R-package, *p*-value and fold change were computed for all three tumour samples compared to their respective normals. We found 47 events upregulated and 28 events downregulated with < 0.001 and fold change >1.5. We are currently validating these events in prostate cell lines using PCR. Figure 3 lists tumour-specific splice events validated by another publicly available RNA-seq dataset of matched tumour-normal prostate samples generated by a different group (SRP002628). Table 3 shows events upregulated and downregulated in both datasets with a *p*-value less than 0.01 and fold change >1.5.

A splice event in *UBE2L3*, a gene involved in p53 ubiquitination and degradation, is upregulated in the tumour samples from all three individuals, and the tumour samples of nine out of ten individuals were used for validation. The *c20orf3* gene, with a upregulated splice event, showed highest correlation between RNA expression and DNA copy implicated in liver metastasis[18]. Again, it has been reported that the *PABPC1* gene showed higher number of copies in prostate cancer using CGH arrays[19]. Structural characterization of these splice events and biological validation using rt-PCR will shed more light on the functional role of these events in cancer progression.

More often, tumour-specific molecular phenotypes, such as gene expression and splicing, are a result of accumulating single nucleotide variations/mutations that offer advantage to tumour cells[20]. Short reads from RNA-seq experiments also carry exonic SNVs from the respective genome location. In this aspect RNA-seq experiments are like exome sequencing, with added burden from RNA editing[21]. The disadvantage of RNA-

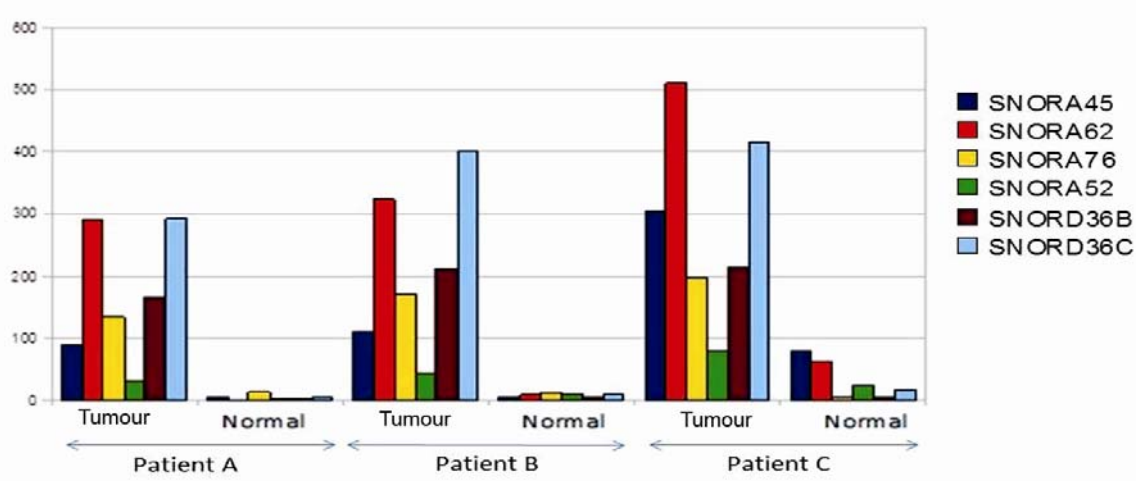

Figure 2. Raw read counts mapped to the six snoRNAs that are significantly upregulated in tumour samples with no expression in the respective normal in all three individuals.

Table 3. Upregulated and downregulated *p*-values common between discovery and validation with *p*-value <0.01

Gene	Exon_Splice_Id	*p*-value	Fold change	Validation *p*-value	Validation fold change	Is AltSplicing?
Upregulated						
EIF4B	NM_001417.J5-6	0.009355	0.590192	0.00547	0.80229	TRUE
HDLBP	NM_005336.J13-14	0.005846	1.29948	0.00131	0.65318	FALSE
C20orf3	NM_020531.J6-7	0.001436	1.595423	0.00029	0.89069	TRUE
UBE2L3	NR_028436.J1-2	0.003611	1.374891	0.00562	0.65944	TRUE
DDX17	NM_001098504.J4-5	0.00641	1.162784	0.00040	1.41089	TRUE
PPP3CA	NM_001130691.J12-13	3.00E-05	0.590618	0.00027	1.44217	TRUE
PABPC1	BG390274.J2-3	0.009984	1.34529	0.00373	1.01303	TRUE
Downregulated						
PKM2	NM_002654.J8-9	0.003508	–1.13807	0.0007	–0.84242	TRUE
TRPC4AP	NM_015638.J17-18	0.007407	–1.56466	0.00148	–0.83771	TRUE
CTSB	NM_147780.J5-6	0.001649	–2.44132	0.00578	–1.15916	TRUE
CTSB	BF674837.J5-6.	0.009404	–0.94708	0.00758	–0.89826	TRUE
CLU	NM_203339.J7-8	0.005587	–2.46905	0.00205	–1.34296	TRUE
CLU	NM_203339.J8-9.	0.005647	–1.22126	0.00040	–1.10388	TRUE

seq-based SNV discovery is that only SNVs in exons significantly expressed in a given sample can be detected. However, the advantage over exome sequencing is that RNA-seq data are not limited to detecting SNVs in known exons, thus enabling novel coding regions.

Mapping of reads using bowtie allowing for two mismatches reports mutation positions and mutation types within each read relative to the read coordinates. Since bowtie also reports the start position on the chromosome where a read is aligned, the chromosomal position of mismatched base and the variant types can be deciphered. Here, for each sample the mismatched base coordinates and types were extracted and loaded onto a MySQL database. To filter false positives stemming from sequencing errors, only those chromosomal positions with evidence of mismatch from more than three reads at the same position with the same base-type were considered potential SNVs in a sample. The odds of finding such an event occurring from sequencing error alone is 0.0004222. The SNV positions between tumour and normal samples from the same individuals were compared using SQL query to identify tumour-specific SNVs in all the three individuals. Tumour-specific SNVs of the three individuals were compared to collate a list of SNVs common to all three tumours, which are not present in their respective normal samples.

The total number of SNVs unique to tumour is 15,034. This is five times more than the number of SNVs (3026) unique only to the three normal samples. All the tumour-specific SNVs were assigned gene names based on their chromosomal locations. Table 4 lists all the genes with more than 100 SNVs. According to the literature, some of the eight genes that have accumulated more than 100 mutations have been implicated in cancer. Also, characterization of SNVs into coding, non-coding, synonymous and non-synonymous mutations suggests that many SNVs have the potential to alter protein functions. Genes *KLK2* and *KLK3* known to be implicated in prostate cancer are among those with the most accumulated SNVs[22]. *KLK2* polymorphism is also associated with prostate cancer[23]. It should be mentioned that

Table 4. Genes with more than 100 accumulated single nucleotide variations unique to tumour samples

Gene	Accession ID	Total mutations	Coding mutations	Non-coding	Synonymous	Non-synonymous
EEF2	NM_001961	267	190	77	54	136
MALAT1	NR_002819	154	0	154	NULL	NULL
BCAM	NM_005581	115	64	51	15	49
GOLM1	NM_016548	115	57	58	13	44
ANKRD30BL	NR_027019	113	0	113	NULL	NULL
KLK2	NM_005551	108	14	94	0	14
KLK3	NM_001648	104	3	101	0	3
CALR	NM_004343	103	62	41	10	52

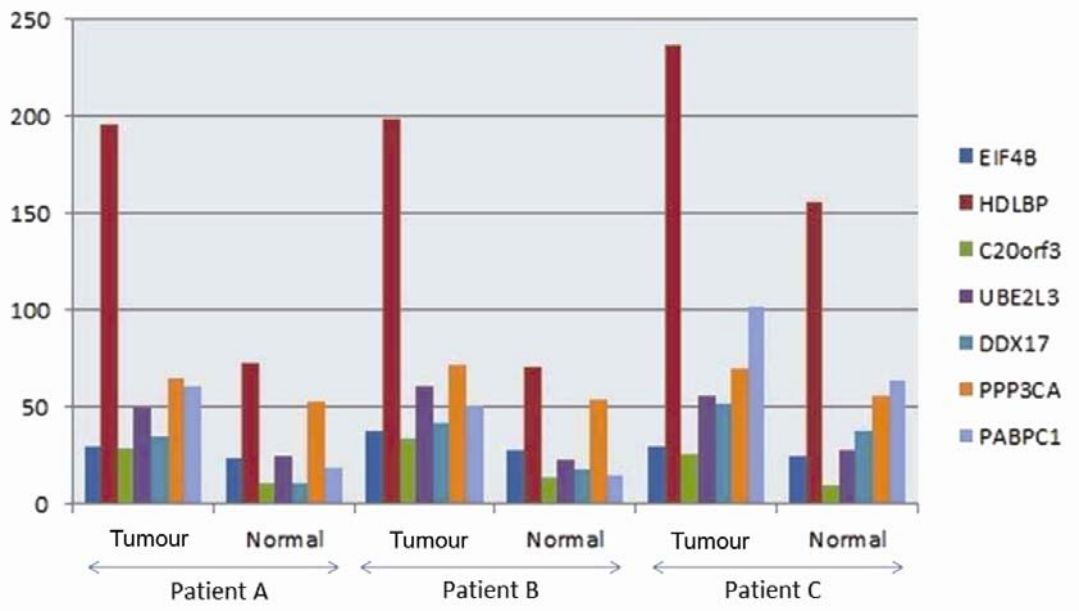

Figure 3. Raw read counts of the spliced reads mapped to the seven splice events in SPEventH.

very high expression levels of some genes may have increased the false positive rate in our findings. However, while more stringent filters may reduce false positives, it will do so at the cost of real SNVs. More work is required to optimize filtering criteria in discovering SNVs from RNA-seq.

RNA-seq experiments are becoming an important part of genomics research. However, unlike other technologies in the past, data from RNA-seq experiments are home to diverse context-specific phenotypes. The speed with which scientists have embraced this technology leaves little room for the equally diverse analysis tools to mature. Because of this disparity between data generation and analysis, public data resources for RNA-seq experiments remain rich with context-specific RNA phenotypes. Here, we have demonstrated that RNA-seq data from public repositories can be harnessed to discover thousands of disease-specific RNA phenotypes, including differential gene expression, differential coding mutation and differential splicing common to the three individuals.

We have demonstrated the usefulness of public databases in biomarker discovery and validation. To our knowledge, there have been no attempts earlier to use RNA-seq data from a single experiment for the identification of three diverse RNA-phenotypes, including gene expression, splice events and coding mutations. We have also attempted to perform computational validation of findings in RNA-seq datasets from the public resources generated by independent groups for the same biological context. As data from RNA-seq experiment grow in repositories, computational validation will become routine. However, for certain phenotypes such as differential splicing, experimental validation may be more rewarding. We are currently validating our findings in prostate tumour sample by PCR methods in cell lines and tumour samples.

Based on the pattern of differential gene and splice expression between tumour and matched normal samples, as shown in Figures 1 and 2, we were able to predict that the normal sample from patient C was perhaps contaminated. This finding was confirmed by communication with the author of the data. However, the impact of this contamination is minimal in this work because the sequencing quality was good. At the most, this contamination may have reduced the odds of finding many more low-expressing, tumour-specific phenotypes.

Currently, a new batch of students is working on extracting other prostate tumour-specific RNA phenotypes such as

copy number variations, transcript-induced chimera and non-coding RNAs from the same dataset. Emphasis is also to discover novel, tumour-specific exons and splice events, which demand entirely different bioinformatics approaches.

The work reported here is the result of a training-cum-research effort launched at our institute, to both accelerate training in NGS data analysis and to instill interest in students to pursue basic research in genomics. All the work presented here was done by students during their six-month internship, which is a requirement to secure a postgraduate diploma from the institute.

1. Berger, M. F. *et al.*, *Genome Res*., 2010, **20**, 413–427.
2. Sultan, M. *et al.*, *Science*, 2008, **321**, 956–960.
3. Twine, N. A., Janitz, K., Wilkins, M. R. and Janitz, M., *PLoS One*, 2011, **6**, e16266.
4. Chepelev, I., Wei, G., Tang, Q. and Zhao, K., *Nucleic Acids Res*., 2009, **37**, e106 (1–8).
5. Derrien, T., Guigó, R. and Johnson, R., *Front. Genet.*, 2011, **2**, 107 (1–6).
6. Nacu, S. *et al.*, *BMC Med. Genomics*, 2011, **4**, 11 (1–22).
7. Wu, X. *et al.*, *J. Transl. Med*., 2012, **10**, 42 (1–7).
8. Kannan, K. *et al.*, *Proc. Natl. Acad. Sci. USA*, 2011, **108**, 9172–9177.
9. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L., *Genome Biol*., 2009, **10**, R25 (1–10).
10. Martens-Uzunova, E. S. *et al.*, *Oncogene*, 2012, **31**, 978–991.
11. Aerssens, J. *et al.*, *Cytogenet. Cell Genet.*, 1995, **71**, 268–275.
12. Schnell, O. *et al.*, *PLoS One*, 2012, **7**, e30429 (1–12).
13. Wang, E. T. *et al.*, *Nature*, 2008, **456**, 470–476.
14. Srinivasan, S., *Curr. Sci*., 2011, **100**, 813–814.
15. Germann, S., Gratadou, L., Dutertre, M. and Auboeuf, D., *J. Nucleic Acids*, 2012, **2012**, 269570 (1–9).
16. Trapnell, C., Pachter, L. and Salzberg, S. L., *Bioinformatics*, 2009, **25**, 1105–1111.
17. Bingham, J. L., Carrigan, P. E., Miller, L. J. and Srinivasan, S., *OMICS*, 2008, **12**, 83–92.
18. Mekenkamp, L. J. M. *et al.*, *Gut*, 2012; doi: 10.1136/gutjnl-2011-301587
19. van Duin, M. *et al.*, *Genes Chromosomes Cancer*, 2005, **44**, 438–449.
20. Teng, M. *et al.*, *BMC Proc. Suppl. 9*, 2011, **5**, S40 (1–7).
21. Kumar, A. *et al.*, *Proc. Natl. Acad. Sci. USA*, 2011, **108**, 17087–17092.
22. Ciampa, J. *et al.*, *Cancer Res.*, 2011, **71**, 3287–3295.
23. Kohli, M. *et al.*, *Cancer Biomark*, 2010, **7**, 101–108.

ACKNOWLEDGEMENTS. We thank DBT, New Delhi for funds through the Ramalingaswamy Fellowship to S.S. We also thank DIT for support under the program 'Centre of Excellence for research and training in bioinformatics' to IBAB.

Received 26 March 2012; revised accepted 9 April 2012

ARUN H. PATIL
MANJARI DESHMUKH
NEERAJ K. SINGH
ROLI SRIVASTAVA
K. SATYA SWATI
MOHIT VERMA
SAURABH GUPTA
S. VEERESH
R. SRIVATSAN
SUBHASHINI SRINIVASAN*

Institute of Bioinformatics and Applied Biotechnology,
Biotech Park,
Electronic City Phase I,
Bangalore 560 100, India
**For correspondence.*
e-mail: ssubha@ibab.ac.in