Yin et al. BMC Medical Genetics 2013, 14:32
http://www.biomedcentral.com/1471-2350/14/32

BMC Medical Genetics

RESEARCH ARTICLE Open Access

Mutation spectrum in human colorectal cancers and potential functional relevance

Hongzhuan Yin, Yichao Liang, Zhaopeng Yan, Baolin Liu and Qi Su*

Abstract

Background: Somatic variants, which occur in the genome of all cells, are well accepted to play a critical role in cancer development, as their accumulation in genes could affect cell proliferations and cell cycle.

Methods: In order to understand the role of somatic mutations in human colorectal cancers, we characterized the mutation spectrum in two colorectal tumor tissues and their matched normal tissues, by analyzing deep-sequenced transcriptome data.

Results: We found a higher mutation rate of somatic variants in tumor tissues in comparison with normal tissues, but no trend was observed for mutation properties. By applying a series of stringent filters, we identified 418 genes with tumor specific disruptive somatic variants. Of these genes, three genes in mucin protein family (*MUC2*, *MUC4*, and *MU12*) are of particular interests. It has been reported that the expression of mucin proteins was correlated with the progression of colorectal cancer therefore somatic variants within those genes can interrupt their normal expression and thus contribute to the tumorigenesis.

Conclusions: Our findings provide evidence of the utility of RNA-Seq in mutation screening in cancer studies, and suggest a list of candidate genes for future colorectal cancer diagnosis and treatment.

Keywords: Colorectal cancers, Mutation spectrum, RNA-Seq, Transcriptome

Background

As the third most common malignancy and the fourth major cause of cancer mortality [1], colorectal cancer is an important threat to human health which accounts for 1 million new cases worldwide each year. The consistency between incidence rates and economic development reflects a westernized lifestyle and attendant risk factor exposures [1]. As a complex condition, colorectal tumor progression is associated with both genetic and environmental factors. To date, only a few common low-penetrance variants attributing to cancer risk have been identified using genome-wide association studies (GWAS), and it is still largely unknown to us the underlying mechanisms and genes involved in tumor development.

Recently, the importance of somatic mutations in cancer development has been widely accepted. It is thought that cancer evolves through the accumulation of somatic mutations in specific genes, depending on various tumor type [2]. Evidence showed that mutation frequency of candidate cancer genes is much higher than expected, and that the particular combination of mutations could influence the tumor's properties [3-6]. These mutations are caused by a combination of environmental and heritable factors [7]. Since the release of the human genome sequence, great efforts have been taken to identify somatic variants in colorectal cancers. For example, Sanger sequencing technique is applied to 13,023 genes and resulted in 189 genes with unexpected excess of somatic mutations in human breast and colorectal cancers [5]. Another group of scientists have used mismatch repair detection (MRD) approach to screen 93 matched tumor-normal sample pairs and 22 cell lines for somatic mutations in 30 cancer relevant genes, and found a total of 152 somatic mutations in breast and colorectal cancers [8], including previously reported genes, such as *BRAF* and *KRAS*.

The recent development of novel high-throughput sequencing methods has provided an unprecedented opportunity to conduct whole-genome scale studies at an affordable cost, and is extensively applied in transcriptome

* Correspondence: suqi100@hotmail.com
Department of General Surgery, Shengjing Hospital of China Medical University, No. 36 Sanhao Street, Heping District, Shenyang, Liaoning Province 110004, China

profiling. This method, termed RNA-Seq, gives a far more precise measurement of expression levels of transcripts and a far more sophisticated characterization of their isoforms [9,10], and has brought successes including identification of differentially expressed genes [11], fusion genes in tumor tissue [12-14], allele-specific expressed genes [15,16]. Moreover, it can also serve as an efficient and cost-effective approach to systematically screen variants in transcribed regions [17-20]. To gain insight into the variation spectrum in tumor samples, we developed a sophisticated variant discovery pipeline and applied it to deep-sequencing transcriptome data from 2 colorectal cancer tissues and their matched normal tissues. There are more variants found in tumor tissues than in normal tissues. After additional filters, we also identified tumor-specific mutations in unreported genes, which supplement the increasing list of candidate colorectal cancer genes.

Methods

Sequence data

Whole transcriptome sequencing data of paired tumor and normal tissues from 2 stage III colorectal cancer patients were downloaded from NCBI Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo), with the accession number SRP006900. Specifically, 65-bp single-end short reads were generated by Illumina Genome Analyzer, following the standard procedure.

Sequence alignment

All single-end reads were aligned to UCSC human genome reference assembly (hg19), limited to chromosomes 1–22, X and Y. The alignment was carried out using BWA [21] with default parameters, which allows 4% mismatches in each alignment.

Variant calling

In each tissue sample, we called variants from the read alignment using SAMtools package [22]. To avoid potential PCR duplicate fragments, we set –D as 100 when invoking vcfutils.pl script, although it varied little when this option is set to 1000 (~3% increase in the total number of variants). Next, we applied several filters to reduce possible false positive calls.

Filter 1.1 We first removed variants that were mistakenly called with a probability greater than 0.01. This was done by requiring a value ≥20 for the 'QUAL' column in vcf files generated by SAMtools.

Filter 1.2 We eliminated false positives that were caused by extremely high sequence coverage. To obtain the optimal upper bound for sequence coverage, we searched for variants after filter 1.1 which were also showed in the dbSNP build 135, and assign them as known set. Then, we decided a cut-off value as 97.5% of known variants have lower coverage than that and applied it to the remaining variants. This step was performed independently for each sample.

Identification of somatic variants

Somatic variants were called by comparing paired normal and tumor tissues. We used custom tools to parse variants after initial filters with following additional filters:

Filter 2.1 Variants in genomic regions of low quality were first excluded for further analysis. Poor quality regions were defined as regions with read coverage in only one sample of a pair, which could be caused by random bias.

Filter 2.2 We next removed variants that were presented in dbSNP135 [23], leaving novel variants.

Filter 2.3 This filter removes variants that are found in both of the matched normal and tumor tissues.

Filter 2.4 To reduce false positives caused by alignment difficulties around indels, we calculated the local mismatch rate as the percentage of mismatches within 10-bp downstream and upstream of a variant. Variants with high local mismatch rate (≥0.1, or ≥2 mismatches) were discarded.

Gene ontology analysis

The gene ontology (GO) [24] information for genes was assigned using bioconductor (http://www.bioconductor.org) package "org.Hs.eg.db". The enrichment tests were performed using "topGO" package [25].

Result

Read alignment and mutation spectrum

The whole transcriptome data of paired normal and tumor tissues from 2 patients contains ~40 million short reads produced by Illumina Genome Analyzer (9.6 million reads per sample), each 65-bp long. Using BWA aligner [21], we mapped short reads to the human reference genome (hg19), and ~30 million (~76%) short reads were mapped to a unique location (Table 1). Next, we made variant calls using SAMtools package. Since massively parallel sequencing technique has higher error rate, extra care must be taken when we used RNA-Seq to identify variants. Therefore we applied a series of stringent filters to minimize false positive rate. First, we removed variants mistakenly called with a probability greater than 0.01, and obtained 89,129 variants. Since PCR duplicates can cause false positives, we next filtered variants with high sequence coverage. To decide the optimal upper boundary, we denoted known variants as found in dbSNP 135 and novel variants as not, and compared sequence coverage between these two sets. We found that the sequence coverage of known variants is significantly higher than that of novel variants (Figure 1, $P < 2.2 \times 10^{-16}$, Wilcoxon rank sum test), then we used the 97.5% percentile in known variants (47 reads) as the cutoff to filter potential false positives. After this step, 85,863 variants were remained, and we

Table 1 Sample and alignment summary

Sample	# reads	# unique reads	Read length	Total throughput	Aligned %
Normal1	9037384	7022993	65 bp	456494545	77.71
Tumor1	8542144	6524738	65 bp	424107970	76.38
Normal2	11308009	8428484	65 bp	547851460	74.54
Tumor2	11461875	8459429	65 bp	549862885	73.80

found that there are more variants in tumor samples when compared to normal samples (23,549 versus 19,383 per sample, ratio = 1.22), with a higher proportion of novel variants in tumor samples (42% versus 39%). Among these variants, a majority are transitions (Figure 2), and the transition/transversion ratio is 2.64 and 2.67 in tumor and normal samples, respectively. These ratios are slightly higher than 2.1, the expected human genome transition/transversion ratio obtained from whole genome resequencing data [26], and it is not unexpected because during transcription, RNA editing specifically changes

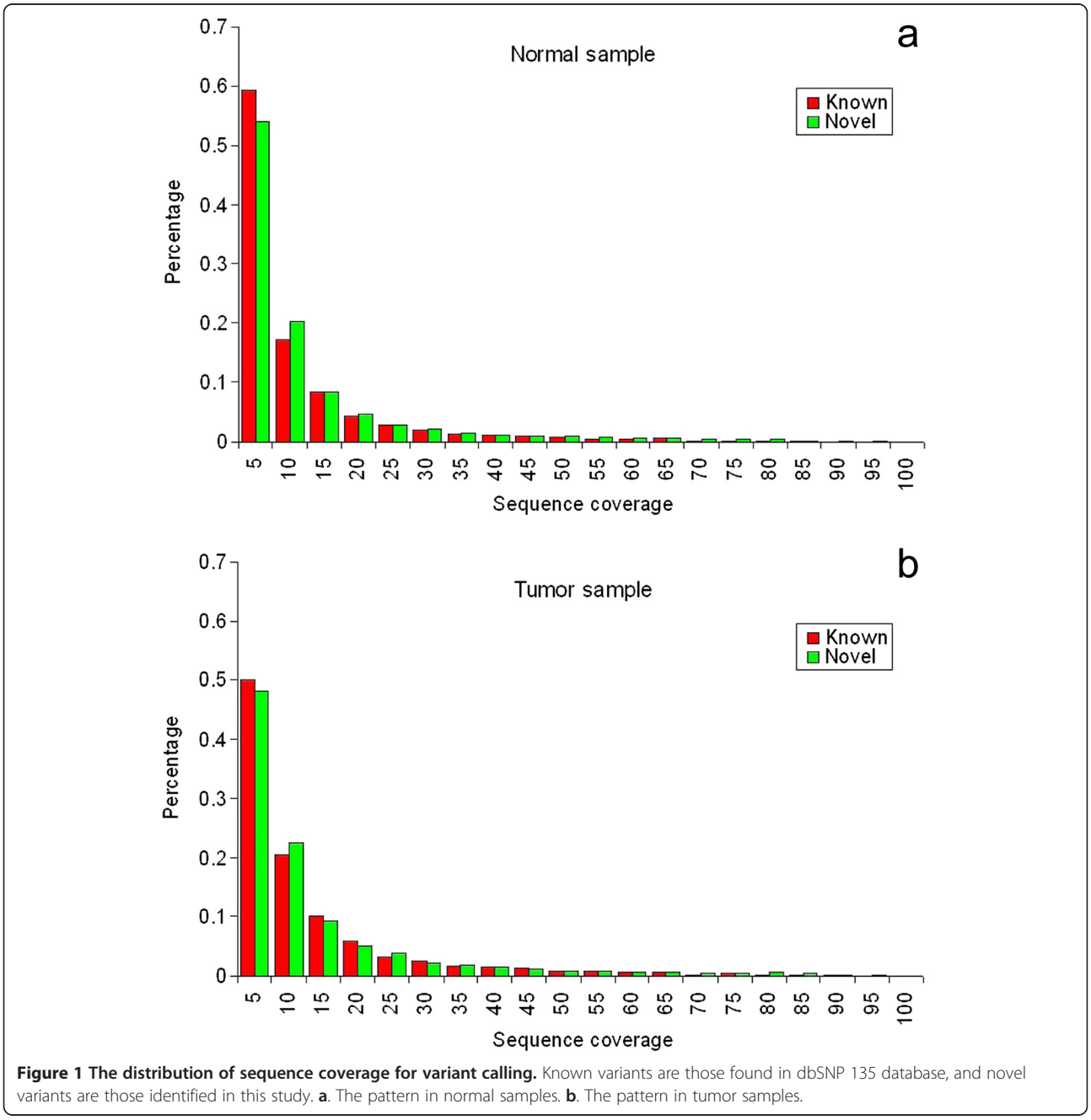

Figure 1 The distribution of sequence coverage for variant calling. Known variants are those found in dbSNP 135 database, and novel variants are those identified in this study. **a**. The pattern in normal samples. **b**. The pattern in tumor samples.

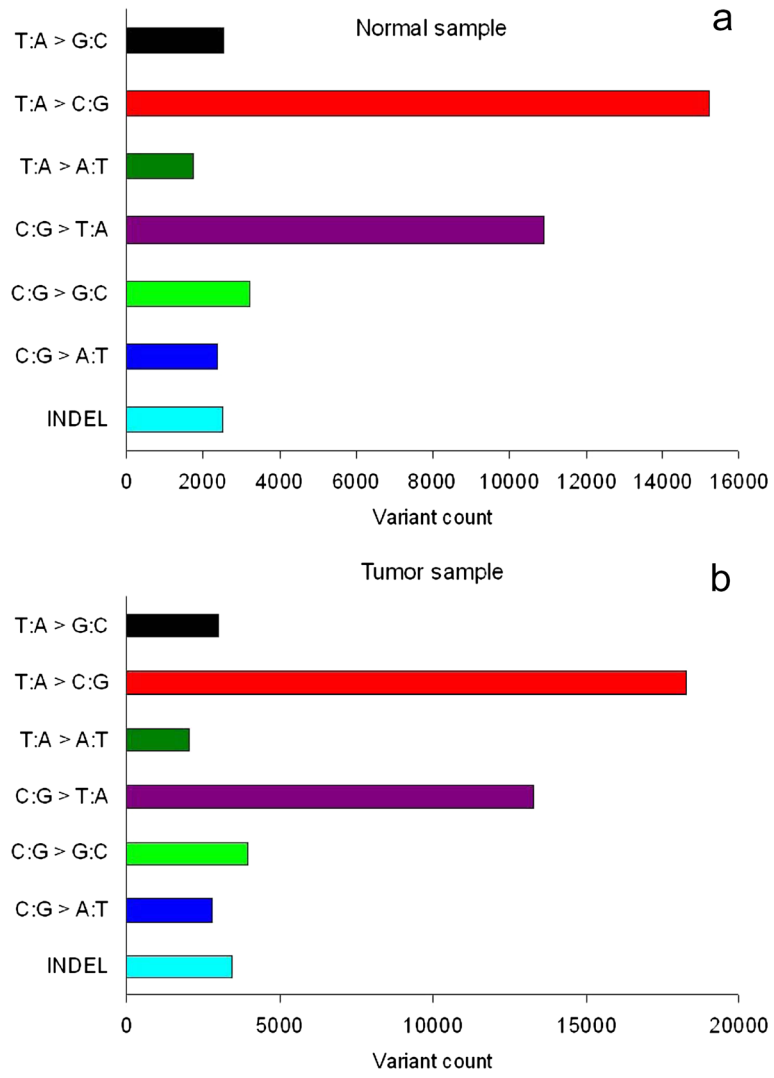

Figure 2 Mutation spectra of normal and tumor tissues. The numbers of each of the six classes of base substitution and insertion/deletions are shown. **a**. The pattern in normal samples. **b**. The pattern in tumor samples.

adenosine (A) to inosine (I), which, in turn, is called as guanosine (G) by sequencers [27].

Identification of somatic variants

To investigate the potential effect of variants on oncogenesis, we next compared somatic variants between paired normal and tumor samples. Several additional filters were applied to call high confident somatic variants. First, if variant positions were only covered in one sample, we removed them to avoid false positives that are probably caused by sequence bias, resulting in 18,970 and 16,409 tumor and normal variants per sample. Next, we filtered known variants found in dbSNP135 [23], which leads to 11,749 and 9,857 novel variants in each tumor and normal sample, respectively. We also removed variants found in both tumor samples and matched normal samples, as well as variants with a high local mutation rate (2 mismatches in the flanking 20-bp region), which might be a result of local misalignment. In total, we obtained 3,382 tumor-specific novel variants and 1,812 variants per sample, across all autosomes and sex chromosomes.

Of note, the ratio of tumor versus normal samples is significantly higher for novel variants when compared to all variants (3,382/1,812 versus 23,549/19,383, $P < 2.2 \times 10^{-16}$, Fisher's exact test), but no bias is observed for transition/transversion ratio between tumor and normal samples (2,054/719 versus 3,929/1,466, $P = 0.235$, Fisher's exact test), so it is less likely that the excess of somatic variants in tumor samples are due to high false positive rate.

Furthermore, we mapped these somatic variants to protein coding genes to screen for potential important genes for tumor progression. In summary, 1,104 tumor-specific

variants and 627 normal-specific variants were found in coding regions. Of them, 671 (60.8%) and 413 (65.9%) variants were disruptive variants (which either change encoding amino acids or reading frames), belonging to 418 and 245 genes, respectively. Additionally, there were 33 genes found to embed somatic mutations in both tumor samples (Table 2).

Functional characterization of genes with somatic variants

It is of great interest to understand functions and putative contributions of genes bearing tumor- and normal-specific variants, thus we extracted gene ontology (GO) [24] annotation for these genes and performed gene enrichment analysis. 5 biological processes were found enriched in tumor samples (Table 3), compared to none in matched normal samples. Among these processes is protein localization (GO:0008104) related to tumor development. Researches found that aberrantly localized proteins have been linked to human diseases, including cancers [28-30], suggesting that variants we identified here may promote tumor progression through this process. We also found that tumor-specific variants were enriched in several molecular functions including nucleotide binding (GO: 0000166), which is not unexpected, as several nucleotide binding genes, such as *GNB2L1*, are found to be involved in cancers [31].

Table 2 List of genes that contain somatic disruptive variants in both tumor samples in this study

Ensembl ID	HNCN symbol	Mutation count
ENSG00000100345	*MYH9*	3
ENSG00000100353	*EIF3D*	2
ENSG00000100461	*RBM23*	2
ENSG00000101182	*PSMA7*	2
ENSG00000108821	*COL1A1*	2
ENSG00000110080	*ST3GAL4*	4
ENSG00000113161	*HMGCR*	2
ENSG00000115457	*IGFBP2*	3
ENSG00000119888	*EPCAM*	3
ENSG00000125124	*BBS2*	2
ENSG00000125970	*RALY*	4
ENSG00000125991	*ERGIC3*	2
ENSG00000128298	*BAIAP2L2*	2
ENSG00000130429	*ARPC1B*	2
ENSG00000134398	*ERN2*	2
ENSG00000144659	*SLC25A38*	2
ENSG00000145113	*MUC4*	10
ENSG00000151846	*PABPC3*	2
ENSG00000163399	*ATP1A1*	2
ENSG00000166794	*PPIB*	2
ENSG00000166888	*STAT6*	3
ENSG00000168542	*COL3A1*	4
ENSG00000173988	*LRRC63*	3
ENSG00000180138	*CSNK1A1L*	2
ENSG00000182944	*EWSR1*	2
ENSG00000184840	*TMED9*	3
ENSG00000188846	*RPL14*	2
ENSG00000197324	*LRP10*	2
ENSG00000198788	*MUC2*	2
ENSG00000204628	*GNB2L1*	4
ENSG00000205277	*MUC12*	2
ENSG00000215570	—	4

Characterization of potential colorectal cancer genes

As is well-known, accumulation of somatic variants is the basic mechanism leading to the development of malignancy. Due to the development of massively parallel sequencing, which makes large-scale sequencing affordable and available, we witnessed a rapid accumulation of somatic variants found in colorectal cancer, such as *MLH3, BRAF, GALNT12,* and *TP53* [32-36]. In the present analysis, we have identified 418 genes with somatic disruptive variants in two tumor samples. Among these genes, we found previously identified genes, such as *TP53*, and tumor-related or oncogenes, such as *RAB5C*, *PIM-3*, *TPT1*, *ST14*. Here we only present several high confident candidate genes that were found in both tumor samples and were good target for diagnosis marker and drug development. Guanine nucleotide binding protein (G protein), beta polypeptide 2-like 1 (*GNB2L1)*, which is also known as *RACK1*, encodes a ubiquitously expressed scaffolding protein and plays a crucial regulatory role in tumor growth [37]. We have detected a 1-bp insertion in both tumor samples, and another 2-bp insertion and a C->T point mutation in one tumor sample. These changes could impact the normal function of *GNB2L1* and thus tumor progression. We also found several members of the mucin protein family that have somatic variants in both tumor samples. Mucin proteins are the major constituents of mucus, which is the viscous secretion that covers epithelial surfaces. There were 2 indels in *MUC2*, 10 indels and point variants in *MUC4*, as well as 1 indel and 1 point variant in *MUC12*. Since the expression of mucin proteins has been correlated with aggressiveness of colorectal cancer [38], the excess of disruptive variants in mucin genes further confirmed their importance in colorectal carcinogenesis.

Discussion

Recent advances in sequencing technologies continuously reduce sequencing costs and increase sequence output at an unprecedented rate, making RNA-Seq an appropriate method to characterize transcriptome profiles, such as gene expression differences or splicing variations. Wang et al. also used RNA-Seq data to derive sample-specific protein databases [39]. By applying this method to two colorectal

Table 3 Enriched molecular function categories in GO analysis

GO.ID	Term	Annotated	Significant	Expected	*P*-value	Corrected *P*
Biological process						
GO:0044419	interspecies interaction between organisms	397	28	9.22	1.80E-07	0.001729
GO:0033036	macromolecule localization	1443	61	33.5	2.70E-06	0.010247
GO:0051704	multi-organism process	943	45	21.89	3.20E-06	0.010247
GO:0008104	protein localization	1190	52	27.63	6.60E-06	0.015852
GO:0030030	cell projection organization	712	35	16.53	2.30E-05	0.044192
Molecular function						
GO:0005515	protein binding	7367	235	171.36	6.30E-12	2.26E-08
GO:0000166	nucleotide binding	2307	84	53.66	1.30E-05	0.01791
GO:0005488	binding	12172	314	283.12	1.50E-05	0.01791
GO.ID	Term	Annotated	Significant	Expected	*P*-value	Corrected *P*
Biological process						
GO:0044419	interspecies interaction between organisms	397	28	9.22	1.80E-07	0.001729
GO:0033036	macromolecule localization	1443	61	33.5	2.70E-06	0.010247
GO:0051704	multi-organism process	943	45	21.89	3.20E-06	0.010247
GO:0008104	protein localization	1190	52	27.63	6.60E-06	0.015852
GO:0030030	cell projection organization	712	35	16.53	2.30E-05	0.044192
Molecular function						
GO:0005515	protein binding	7367	235	171.36	6.30E-12	2.26E-08
GO:0000166	nucleotide binding	2307	84	53.66	1.30E-05	0.01791
GO:0005488	binding	12172	314	283.12	1.50E-05	0.01791

cancer cell lines SW480 and RKO, they found a significant improvement in protein identification. In addition, RNA-Seq can also be used for variant detection in transcribed regions, which is suitable for identification of somatic mutations [17-20,40,41]. However, it has been concerned that variant-calling by RNA-Seq is prone to error [18] and could generate a high false discovery rate. To minimize that, we implemented a series of stringent filters in our bioinformatic discovery pipeline. First, we required each variant should have a quality score no less than 20, removing poorly called variants. Next, we used variants that were found in dbSNP135 dataset to train our pipeline and filtered variants with extremely high read coverage. We also applied additional stringent filters to call high confident tissue-specific novel variants, including removing variants with high local mismatch rate. In our final list, we identified more somatic variants in tumor samples than in normal samples, and some variants were in tumor-related genes. Due to our strict filters, we argued that there should be more genes containing tumor-specific somatic variants.

It is widely acknowledged that accumulations of mutations in oncogenes and tumor suppressor genes are the main cause of human cancer [2]. Mutations occurred only in tumor tissues provide important information to understand the potential biological processes underlying carcinogenesis, as well as to facilitate the development of diagnostic and therapeutic markers. As the development of sequencing techniques and the decrease of corresponding costs, large-scale studies begin to accumulate to identify somatic mutations in colorectal cancers. In one study, Sjöblom et al. used polymerase chain reaction (PCR) approach to analyze 13,023 genes in 11 breast and 11 colorectal cancers [5], and found an average of ~90 mutated genes per tumor sample. Using stringent criteria, they identified 189 significantly mutated genes, which affect a wide range of cellular functions, including transcription, adhesion, and invasion. In another study, Timmerman et al. applied next-generation sequencing to sequence the whole exome of primary colon tumors as well as adjacent not affected normal colonic tissue [32]. More than 50,000 small nucleotide variations were identified for each tissue, and there are 359 and 45 most significant mutations in microsatellite stable (MSS) and microsatellite instable (MSI) colon cancers. Somatic mutations were found in the intracellular kinase domain of bone morphogenetic protein receptor 1A, *BMPR1A*, of which germline mutations are associated with juvenile polyposis syndrome. In this present study, we analyzed RNA-Seq data from 2 colorectal tumors and their matched normal tissues to compare their mutation spectra. In general, tumor tissues were enriched in somatic variants compared with normal tissues. By mapping short reads to 54,665 annotated human genes, we have detected 418 genes with somatic variants in tumor

tissues, including 3 mucin genes found in both tumor samples. Mucins are complex glycoproteins and play important roles in protecting epithelial surfaces [38], alterations in mucin expression and the extent of their glycosylation have been reported to be associated with neoplastic progression and metastasis in several human cancers [42-44]. Since disruptive variants may radically change protein functions instead of gene expression, we further used SIFT tool [45] to assess their effects on protein functions. 10 of 12 variants were classified as tolerated variants, which have a limited impact on the protein function. Thus it is more likely that these disruptive mutations in mucin genes regulate gene expression and thus lead to tumorigenesis. Additionally, mucins can form insoluble mucous to protect gut lumen, therefore amino acid changes in these genes could result in the modification of the micro-environment. This change may in turn lead to the proliferation of some bacteria such as *Fusobacterium nucleatum* and *Coriobacteria*, which have been reported to be significantly over-represented in colorectal tumor specimens [46,47]. Somatic disruptive mutations in these genes found here suggest the abnormality of their expression is related to colorectal tumorigenesis.

Conclusions

RNA-Seq is a powerful tool to identify somatic mutations in protein-coding regions after sophisticated filters. The list of genes we found in this study only represents a minimal set of candidate genes, due to the stringent criteria we applied. However, the identification of several oncogenes and tumorigenesis genes, as well as signal pathway genes, provides meaningful candidates to understand the molecular mechanism of colorectal cancer and for future drug target development. Although additional validations and functional examination are helpful, RNA-Seq, with well developed bioinformatic pipeline, can serve as the first step for somatic variant screening in human cancers.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

YZ and SQ carried out the prime studies and drafted the manuscript. All people have participated in the design of the study and the experiments. In addition, YP and LB coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Received: 13 August 2012 Accepted: 10 January 2013
Published: 8 March 2013

References

1. Tenesa A, Dunlop MG: **New insights into the aetiology of colorectal cancer from genome-wide association studies.** *Nat Rev Genet* 2009, **10**(6):353–358.
2. Vogelstein B, Kinzler KW: **Cancer genes and the pathways they control.** *Nat Med* 2004, **10**(8):789–799.
3. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, *et al*: **Patterns of somatic mutation in human cancer genomes.** *Nature* 2007, **446**(7132):153–158.
4. Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, *et al*: **Core signaling pathways in human pancreatic cancers revealed by global genomic analyses.** *Science* 2008, **321**(5897):1801–1806.
5. Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, *et al*: **The consensus coding sequences of human breast and colorectal cancers.** *Science* 2006, **314**(5797):268–274.
6. Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, *et al*: **The genomic landscapes of human breast and colorectal cancers.** *Science* 2007, **318**(5853):1108–1113.
7. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A, Hemminki K: **Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from Sweden, Denmark, and Finland.** *N Engl J Med* 2000, **343**(2):78–85.
8. Bentivegna S, Zheng J, Namsaraev E, Carlton VE, Pavlicek A, Moorhead M, Siddiqui F, Wang Z, Lee L, Ireland JS, *et al*: **Rapid identification of somatic mutations in colorectal and breast cancer tissues using mismatch repair detection (MRD).** *Hum Mutat* 2008, **29**(3):441–450.
9. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics.** *Nat Rev Genet* 2009, **10**(1):57–63.
10. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5**(7):621–628.
11. Zhang LQ, Cheranova D, Gibson M, Ding S, Heruth DP, Fang D, Ye SQ: **RNA-seq Reveals Novel Transcriptome of Genes and Their Isoforms in Human Pulmonary Microvascular Endothelial Cells Treated with Thrombin.** *PLoS One* 2012, **7**(2):e31229.
12. Ju YS, Lee WC, Shin JY, Lee S, Bleazard T, Won JK, Kim YT, Kim JI, Kang JH, Seo JS: **A transforming KIF5B and RET gene fusion in lung adenocarcinoma revealed from whole-genome and transcriptome sequencing.** *Genome Res* 2012, **22**(30):436–445.
13. Kohno T, Ichikawa H, Totoki Y, Yasuda K, Hiramoto M, Nammo T, Sakamoto H, Tsuta K, Furuta K, Shimada Y, *et al*: **KIF5B-RET fusions in lung adenocarcinoma.** *Nat Med* 2012, **18**(3):375–377.
14. Lee CH, Ou WB, Marino-Enriquez A, Zhu M, Mayeda M, Wang Y, Guo X, Brunner AL, Amant F, French CA, *et al*: **14-3-3 fusion oncogenes in high-grade endometrial stromal sarcoma.** *Proc Natl Acad Sci USA* 2012, **109**(3):929–934.
15. Gregg C, Zhang J, Butler JE, Haig D, Dulac C: **Sex-specific parent-of-origin allelic expression in the mouse brain.** *Science* 2010, **329**(5992):682–685.
16. Gregg C, Zhang J, Weissbourd B, Luo S, Schroth GP, Haig D, Dulac C: **High-resolution analysis of parent-of-origin allelic expression in the mouse brain.** *Science* 2010, **329**(5992):643–648.
17. Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, *et al*: **Stem cell transcriptome profiling via massive-scale mRNA sequencing.** *Nat Methods* 2008, **5**(7):613–619.
18. Cirulli ET, Singh A, Shianna KV, Ge D, Smith JP, Maia JM, Heinzen EL, Goedert JJ, Goldstein DB: **Screening the human exome: a comparison of whole genome and whole transcriptome sequencing.** *Genome Biol* 2010, **11**(5):R57.
19. Kridel R, Meissner B, Rogic S, Boyle M, Telenius A, Woolcock B, Gunawardana J, Jenkins C, Cochrane C, Ben-Neriah S, *et al*: **Whole transcriptome sequencing reveals recurrent NOTCH1 mutations in mantle cell lymphoma.** *Blood* 2012, **119**(9):1963–1971.
20. Canovas A, Rincon G, Islas-Trejo A, Wickramasinghe S, Medrano JF: **SNP discovery in the bovine milk transcriptome using RNA-Seq technology.** *Mamm Genome* 2010, **21**(11–12):592–598.
21. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**(14):1754–1760.
22. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25**(16):2078–2079.
23. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res* 2001, **29**(1):308–311.
24. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**(1):25–29.
25. Alexa A, Rahnenfuhrer J, Lengauer T: **Improved scoring of functional groups from gene expression data by decorrelating GO graph structure.** *Bioinformatics* 2006, **22**(13):1600–1607.
26. Lam HY, Pan C, Clark MJ, Lacroute P, Chen R, Haraksingh R, O'Huallachain M, Gerstein MB, Kidd JM, Bustamante CD, *et al*: **Detecting and annotating genetic variations using the HugeSeq pipeline.** *Nat Biotechnol* 2012, **30**(3):226–229.

27. Bass BL: **RNA editing by adenosine deaminases that act on RNA.** *Annu Rev Biochem* 2002, **71**:817–846.
28. Hung MC, Link W: **Protein localization in disease and therapy.** *J Cell Sci* 2011, **124**(Pt 20):3381–3392.
29. Fabbro M, Henderson BR: **Regulation of tumor suppressors by nuclear-cytoplasmic shuttling.** *Exp Cell Res* 2003, **282**(2):59–69.
30. Dansen TB, Burgering BM: **Unravelling the tumor-suppressive functions of FOXO proteins.** *Trends Cell Biol* 2008, **18**(9):421–429.
31. Nymark P, Wikman H, Ruosaari S, Hollmen J, Vanhala E, Karjalainen A, Anttila S, Knuutila S: **Identification of specific gene copy number changes in asbestos-related lung cancer.** *Cancer Res* 2006, **66**(11):5737–5743.
32. Timmermann B, Kerick M, Roehr C, Fischer A, Isau M, Boerno ST, Wunderlich A, Barmeyer C, Seemann P, Koenig J, *et al*: **Somatic mutation profiles of MSI and MSS colorectal cancer identified by whole exome next generation sequencing and bioinformatics analysis.** *PLoS One* 2010, **5**(12):e15661.
33. Li WQ, Kawakami K, Ruszkiewicz A, Bennett G, Moore J, Iacopetta B: **BRAF mutations are associated with distinctive clinical, pathological and molecular features of colorectal cancer independently of microsatellite instability status.** *Mol Cancer* 2006, **5**:2.
34. Guda K, Moinova H, He J, Jamison O, Ravi L, Natale L, Lutterbaugh J, Lawrence E, Lewis S, Willson JK, *et al*: **Inactivating germ-line and somatic mutations in polypeptide N-acetylgalactosaminyltransferase 12 in human colon cancers.** *Proc Natl Acad Sci USA* 2009, **106**(31):12921–12925.
35. Godai TI, Suda T, Sugano N, Tsuchida K, Shiozawa M, Sekiguchi H, Sekiyama A, Yoshihara M, Matsukuma S, Sakuma Y, *et al*: **Identification of colorectal cancer patients with tumors carrying the TP53 mutation on the codon 72 proline allele that benefited most from 5-fluorouracil (5-FU) based postoperative chemotherapy.** *BMC Cancer* 2009, **9**:420.
36. Iacopetta B: **TP53 mutation in colorectal cancer.** *Hum Mutat* 2003, **21**(3):271–276.
37. Wang F, Osawa T, Tsuchida R, Yuasa Y, Shibuya M: **Downregulation of receptor for activated C-kinase 1 (RACK1) suppresses tumor growth by inhibiting tumor cell proliferation and tumor-associated angiogenesis.** *Cancer Sci* 2011, **102**(11):2007–2013.
38. Manne U, Weiss HL, Grizzle WE: **Racial differences in the prognostic usefulness of MUC1 and MUC2 in colorectal adenocarcinomas.** *Clin Cancer Res* 2000, **6**(10):4017–4025.
39. Wang X, Slebos RJ, Wang D, Halvey PJ, Tabb DL, Liebler DC, Zhang B: **Protein identification using customized protein sequence databases derived from RNA-Seq data.** *J Proteome Res* 2012, **11**(2):1009–1017.
40. Chepelev I, Wei G, Tang Q, Zhao K: **Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq.** *Nucleic Acids Res* 2009, **37**(16):e106.
41. Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh T, McDonald H, Varhol R, Jones S, Marra M: **Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing.** *Biotechniques* 2008, **45**(1):81–94.
42. Ho SB, Niehans GA, Lyftogt C, Yan PS, Cherwitz DL, Gum ET, Dahiya R, Kim YS: **Heterogeneity of mucin gene expression in normal and neoplastic tissues.** *Cancer Res* 1993, **53**(3):641–651.
43. Byrd JC, Bresalier RS: **Mucins and mucin binding proteins in colorectal cancer.** *Cancer Metastasis Rev* 2004, **23**(1–2):77–99.
44. Biemer-Huttmann AE, Walsh MD, McGuckin MA, Ajioka Y, Watanabe H, Leggett BA, Jass JR: **Immunohistochemical staining patterns of MUC1, MUC2, MUC4, and MUC5AC mucins in hyperplastic polyps, serrated adenomas, and traditional adenomas of the colorectum.** *J Histochem Cytochem* 1999, **47**(8):1039–1048.
45. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function.** *Nucleic Acids Res* 2003, **31**(13):3812–3814.
46. Castellarin M, Warren RL, Freeman JD, Dreolini L, Krzywinski M, Strauss J, Barnes R, Watson P, Allen-Vercoe E, Moore RA, *et al*: **Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma.** *Genome Res* 2012, **22**(2):299–306.
47. Kostic AD, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM, Ojesina AI, Jung J, Bass AJ, Tabernero J, *et al*: **Genomic analysis identifies association of Fusobacterium with colorectal carcinoma.** *Genome Res* 2012, **22**(2):292–298.

doi:10.1186/1471-2350-14-32
Cite this article as: Yin *et al.*: Mutation spectrum in human colorectal cancers and potential functional relevance. *BMC Medical Genetics* 2013 **14**:32.