Science Discovery
2019; 7(2): 65-71
http://www.sciencepublishinggroup.com/j/sd
doi: 10.11648/j.sd.20190702.12
ISSN: 2331-0642 (Print); ISSN: 2331-0650 (Online)

Using NGS Technology for Transcriptome Comparison of Normal and Tumor Tissues in Colorectal Cancer Patients

Chen Chumo

Shanghai World Foreign Lanuage Academy, Shanghai, China

Email address:
chenchumo2019@163.com

To cite this article:
Chen Chumo. Using NGS Technology for Transcriptome Comparison of Normal and Tumor Tissues in Colorectal Cancer Patients. *Science Discovery*. Vol. 7, No. 2, 2019, pp. 65-71. doi: 10.11648/j.sd.20190702.12

Received: April 21, 2019; **Accepted**: May 20, 2019; **Published**: May 23, 2019

Abstract: Colorectal cancer is a kind of malignant tumor, which results from life habits and genetic factors. This research project is called "Using NGS Technology for Transcriptome Comparison of Normal and Tumor Tissues in Colorectal Cancer Patients", which belongs to basic research. Next-generation sequencing technology (NGS), also known as high-throughput sequencing technology, can comprehensively obtain whole genome and transcriptome information of cells and tissues through experimental operations including DNA or RNA extraction, purification, library construction, and bioinformatics analysis; in this way, NGS can screen the mutation and abnormal expression of tumor genes and provide relatively accurate therapeutic targets as the guidance of medication. In this study, the transcriptome of five pairs of tumor tissue samples (t: tumor) and normal tissue samples (n: normal) from five patients with colorectal cancer were compared by NGS: the total RNA from those tissues were extracted to construct mRNA library, from which cancer-related transcriptome information was obtained through NGS-based RNA-seq technology and bioinformatics analysis, which may guide clinical research and treatment. The RNA-seq data of transcriptome information from the colorectal cancer patients reflected variably expressed pattern in the comparison of CRCs/normal tissue pairs; two notable targets, COL1A1 and SPP1, were consistent with two other previous researches and matched pathways enrichment with one of them.

Keywords: NGS RNA-Seq Technology, Transcriptome Comparison, Bioinformatics Analysis, Colorectal Cancer

利用NGS技术对结直肠癌患者癌症及正常组织的转录组比对研究

陈楮墨

上海市世界外国语中学，上海，中国

邮箱
chenchumo2019@163.com

摘要：结直肠癌是一种常见的恶性肿瘤，与生活习惯和遗传等因素相关。本研究项目名为“利用NGS技术对结直肠癌患者癌症及正常组织的转录组比对研究”，属于基础领域研究。下一代测序技术(NGS)，又称作高通量测序技术，通过DNA或RNA提取、纯化、建库、生物信息学分析等实验操作，全面获取细胞与组织的总RNA与转录组RNA，检测肿瘤基因的变异与异常表达，提供较为准确的治疗靶点。本实验通过NGS技术比较五位结直肠癌患者的五对肿瘤组织样本中正常组织(n: normal)与肿瘤组织(t: tumor)的转录组：从中提取总RNA用于mRNA建库，通过基于NGS转录组测序技术(RNA-seq)与生物信息学分析收集与导致结直肠癌的转录组信息，作为临床研究与治疗的指导。来自结直肠癌患者的转录组信息的RNA-seq数据反映了结直肠癌/正常组织对的比较中不同的表达模式；两个显著的靶点COL1A1和SPP1，与其他两个以前的研究一致，其中一个研究里富集的通路也与我的研究一致。

关键词：NGS转录组测序技术，转录组比对研究，生物信息学分析，结直肠癌

1．研究背景

结直肠癌(Colorectal cancer, CRC)，是世界上第三大肿瘤。[1]尽管结直肠癌在过去二十年间常见于西方国家，近年来，其发病率在东方国家逐步上升。红肉的摄入，酗酒，长期吸烟与肥胖现已被证明为导致结直肠癌的重要原因。[2]随着中国的发展，中国人的生活习惯发生了许多改变，这成为了中国结直肠癌患者上升的诱因。

从症状上看，结直肠癌早期无症状或无明显症状，仅感不适，消化不良，便中带血等。随着癌症的恶化，其症状趋于明显，表现为腹痛，腹部肿块，肠梗阻，便血，发热与消瘦等。结直肠癌的临床特征因发病部位的不同而不同，可转移并影响其他身体部位。治疗结直肠癌的主要方式为外科手术，放疗和化疗。手术切除针对特定的肿瘤部位，且对于肿瘤再生或癌细胞转移的情况也有一定效果。当癌细胞转移至淋巴结时，建议进行辅助性化疗。

作为本研究中的主要实验手段，NGS技术由Sanger等人开创后[3]，DNA双脱氧核苷测序发生了巨大的变化，对生物信息学与计算机数据库等方面的要求也随之提高。此时，下一代测序技术Next-generation Sequencing (NGS或高通量测序技术)成为可能，并对遗传学产生了重大影响。与传统的测序技术相比，NGS能够处理数百万次的数据读取，然后产生大量的数据以发现关于人类基因组的新颖或重要的信息。NGS技术从全基因组测序(whole-genome sequencing, WGS) 发展至更为高效的组织和解释方法，例如全外显子组测序(whole-exome sequencing, WES)，靶向测序或转录组测序 (transcriptome sequencing, RNA-Seq)。[4]

通过高通量测序，可以满足即时转录组分析，以研究细胞的生理或状态，从而实现靶向基因的预测与验证，不同样本间基因表达差异分析，转录组表达谱分析；同时，在基因组意义上发现和获得miRNA图谱以及新的miRNA分子变为可能。此外，转录组还可用于分析剪接变异、等位基因表达[5]，RNA编辑[6]和3′-UTR多腺苷酸化。[7]

NGS作为一种全面的全基因组工具，目前已经被用于诊断体细胞突变和急性髓细胞白血病（AML）的亚型分类。基于NGS的移植术后可监测AML患者，并区分接受异基因造血细胞移植（HCT）后复发的高危患者，以检查是否存在等位基因负担，它在化疗后的持续存在与更高的复发可能性有关。NGS证实了HCT后突变的逐步清除，并证实了HCT后突变的起源于最初诊断的突变，以及HCT第21天后总变异等位基因频率与更差的总生存率之间的正相关，以及复发的可能性。[8]

NGS技术还被应用于预测或诊断一些重要的癌症。它确认了前列腺癌新的复发改变(例如TMPRSS2-ERG移位，SPOP和CHD1突变)，并验证了此前的通路(如雄激素受体的过表达和突变; PTEN, RB1和TP53的缺失或突变)。[9]通过NGS研究，结合临床病理和放射学数据，可以克服前列腺癌的高度异质性，在诊断、预后和患者特异性药物鉴定方面具有更高的准确性。

利用NGS后的RNA-seq基因表达，结合各种癌症类型和生存率分析，以转录组数据为基础，可以得出与体细胞突变同质性和异质性研究相比存在缺陷的基因表达谱。在33种癌症类型中发现了持续上调的基因和下调的基因，一些基因，如PLP1, MYOM1, NKAPL和USP2，通常在各种癌症中失调，这证实了新的发现。[10]

根据以上的研究，NGS技术可以用来检测结直肠癌的一些有用的全基因组信息。Vogelstein等人总结出不同基因中的体细胞突变(例如APC, KRAS, BRAF, SMAD4, TP53)导致腺瘤性息肉发展为浸润性结直肠癌。[11]以往的研究表明:“典型结直肠癌（CRC）体细胞基因突变发病率较高（>10%），包括TP53, APC, KRAS, SMAD4, FBXW7与PIK3CA.”然而，根据中国关于《中国散发性结直肠癌体细胞突变概况》的研究，APC和Wnt信号通路显示出比TCGA数据显著偏低的突变频率。因此，这种差异干扰了对中国结直肠癌的认识。[12]

在此基础上，与以往的研究相比，中国散发性结直肠癌的体细胞突变在某些通路上似乎是独一无二的。这些突变对目标的实质性影响，以及散发性中国结直肠癌的转录组情况均值得探究。因此，NGS在转录组中的应用有助于验证结直肠癌进展过程中不同表达的基因和疾病机制，有助于了解中国结直肠癌机制和个体化治疗的特异性。

本研究从结直肠患者的5对正常组织和肿瘤组织样本中提取RNA，构建用于RNA测序的mRNA文库。通过测序之后的转录组表达谱分析，发现两组样本间基因表达模式存在差异，预测和验证了特异的肿瘤靶点，对我国结直肠癌的临床治疗有一定的参考和借鉴价值。

2．研究方法

从结直肠癌患者处收集结直肠癌组织与正常组织经同济大学肿瘤研究所处同意，获得五对结直肠癌患者肿瘤与正常组织样本。

从结直肠肿瘤和正常组织样本中提取总RNA：每50－100mg样本组织中加入 1 mL TRIzol™ 试剂并用匀浆器匀浆；室温孵育5分钟使核蛋白复合物充分分离；每1 mL TRIzol™ 试剂裂解的样本加入0.2 mL氯仿，然后小心盖好管盖；剧烈摇晃30秒并孵育2-3分钟；将样本 12,000 × g，4°C离心15分钟；将上层含RNA的水相转移至一个新管；每1 mL TRIzol™ 试剂裂解的样本加入0.5 mL 异丙醇至水相中进行沉淀；孵育10分钟；12,000 × g，4°C离心10分钟，用移液器去除上清；用20－50 μL DEPC水重悬沉淀；用水浴或金属浴55－60°C孵育10－15 分钟；继续下游实验，或将RNA保存于－70°C。

mRNA文库构建：按照NEBNext Ultra II RNA 建库试剂盒（适配Illumina）说明书操作。

NGS测序与生物信息学分析：本研究的RNA测序是在Illumina Hiseq4000 平台上进行，生物信息学分析是由美吉公司的I-sanger平台来完成的。

3．数据与讨论

3.1．正常组织与肿瘤组织样本之间基因表达的Venn分析

Venn分析显示了各组的基因/转录本数量以及各组之间的基因/转录本的重合。经Venn分析，正常组有17773个基因/转录本，肿瘤组有17076个基因/转录本，它们共有15895个相同的基因/转录本。因此，正常组有1878个特异性基因/转录本，肿瘤组有1181个特异性基因/转录本。

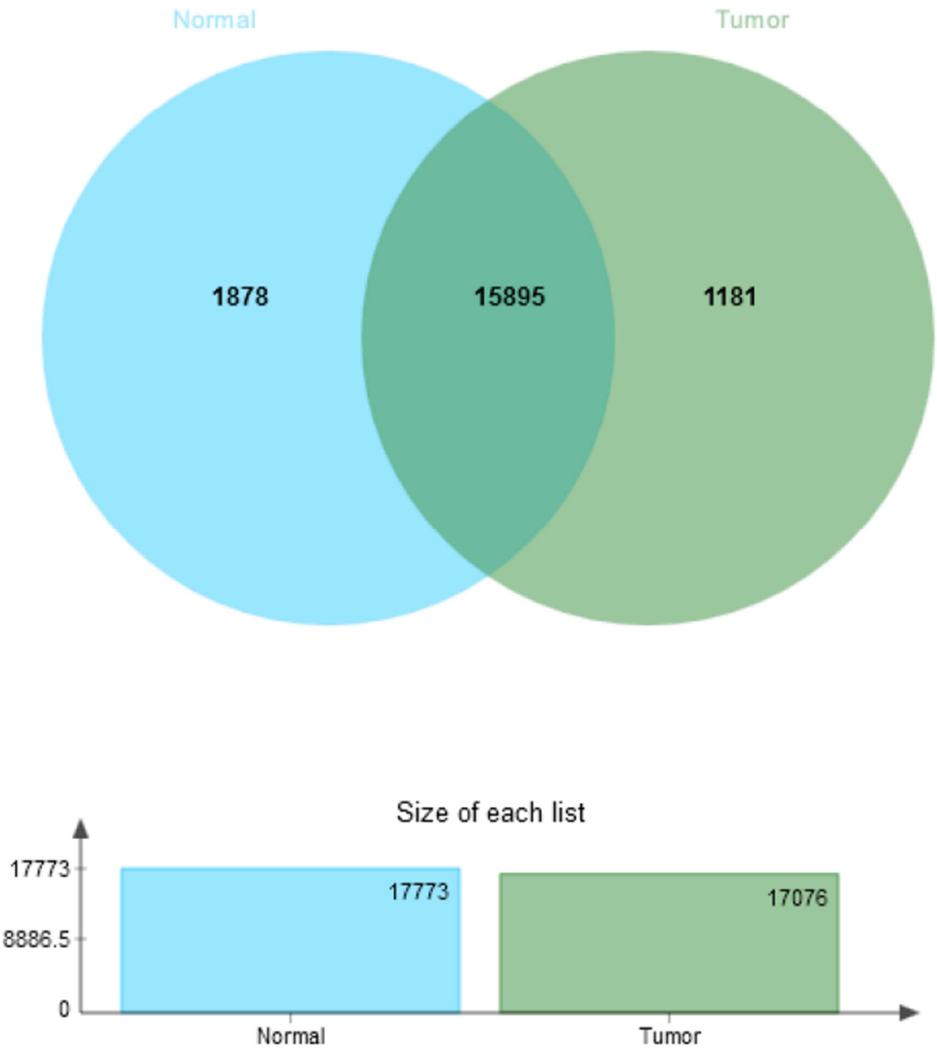

图1 正常组织和肿瘤组织样本之间的Venn分析，不同颜色的圆圈代表每组样本中表达的基因数量。

3.2．正常组织与肿瘤组织样本之间的PCA分析

主成分分析（PCA）利用数学方法从原始变量中重新组合出一组新的自变量（主成分）。PCA能降低数据的复杂度，更深入地挖掘样本之间的关系和变异，并对各分量的重要性进行排序。

我得到了8个主成分，在表1中列出了它们的变异比例，并用PC1和PC2绘制了图2。通过PC1和PC2降维分析，我两组样本分离良好，即正常组和肿瘤组的转录组差异很大，显示了结直肠癌的转录组特异性。

表1 主成分变异比例。

	Proportion of Variance
PC1	0.327969819
PC2	0.211976673
PC3	0.092682845
PC4	0.086374113
PC5	0.075328948
PC6	0.063086227
PC7	0.057845171
PC8	0.047422316

注：第一列为各主成分，第二列为主成分所占比例，数值越大代表该主成分越能区分样本。

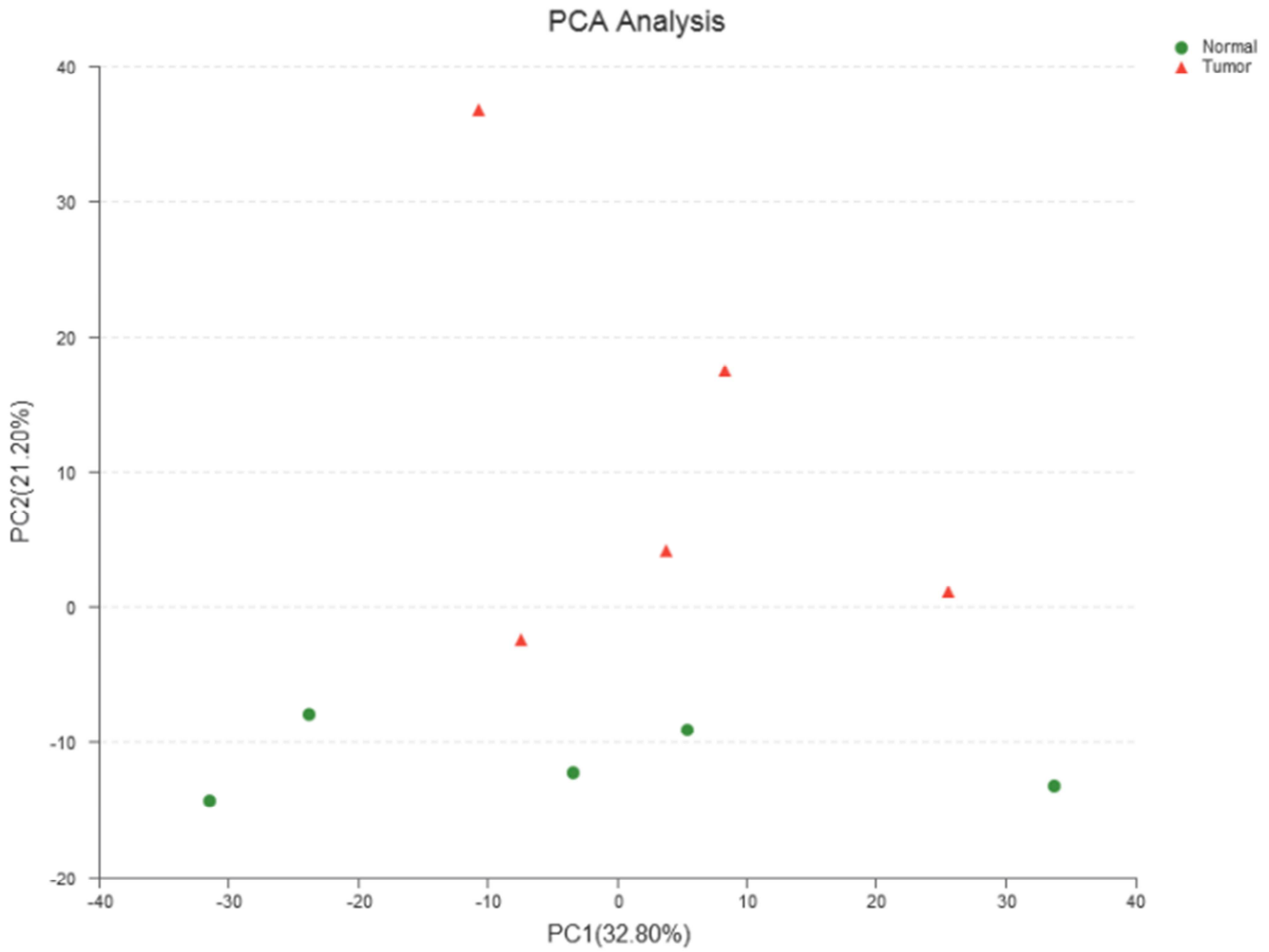

图2 两组正常组织和肿瘤组织样本之间的PCA分析。

样本通过降维分析后，在主成分上有相对坐标点，各个样本点的距离代表了样本的距离，距离越近表明样本间相似性越高。横轴表示二维图中主成分1（PC1)对区分样本的贡献度，纵轴表示二维图中主成分2（PC2)对区分样本的贡献度。

3.3．表达差异分析

通过对表达差异的分析，我发现正常组和肿瘤组组织样本中有1504个表达不同的基因。表2展示了部分基因的名称和描述，详表参见附件S1。

表2 两组正常和肿瘤组织样本中不同表达基因的统计分析（附件S1）。

Gene id	Gene name	Gene description	Normal vs Tumor	Sum
ENSG00000247077	PGAM5	PGAM family member 5, mitochondrial serine/threonine protein phosphatase [Source:HGNC Symbol;Acc:HGNC:28763]	yes	1
ENSG00000106003	LFNG	LFNG O-fucosylpeptide 3-beta-N-acetylglucosaminyltransferase [Source:HGNC Symbol;Acc:HGNC:6560]	yes	1

（1）Gene/Transcripts_id：基因/转录本编号；（2）Gene name：基因名称的名称；（3）Gene description：基因信息的描述；（4）不同组中相同的基因数目；（5）特异表达的基因数目。

3.4．基因集聚类分析

基因集是从特定筛选条件（如功能、表达水平和表达差异）中获得的一组基因/转录本集。聚类分析用于确定不同实验条件下基因/转录本的表达模式；具有相似表达模式的基因/转录本可能具有相似的功能，或可能参与相同的代谢过程或细胞途径。因此，通过对具有相同或相似表达模式的基因进行聚类，可以推断出未知基因、基因/转录本的功能或参考基因/转录本的新功能。

从两组正常组织和肿瘤组织样本的聚类分析热图（图3），所有5个肿瘤组织样本均聚集在左侧，所有其他5个正常组织样本均聚集在右侧，表明同一组具有相似的表达模式，并可能参与相同的代谢过程或细胞途径。表3列出了所有基因表达值的详细数据。

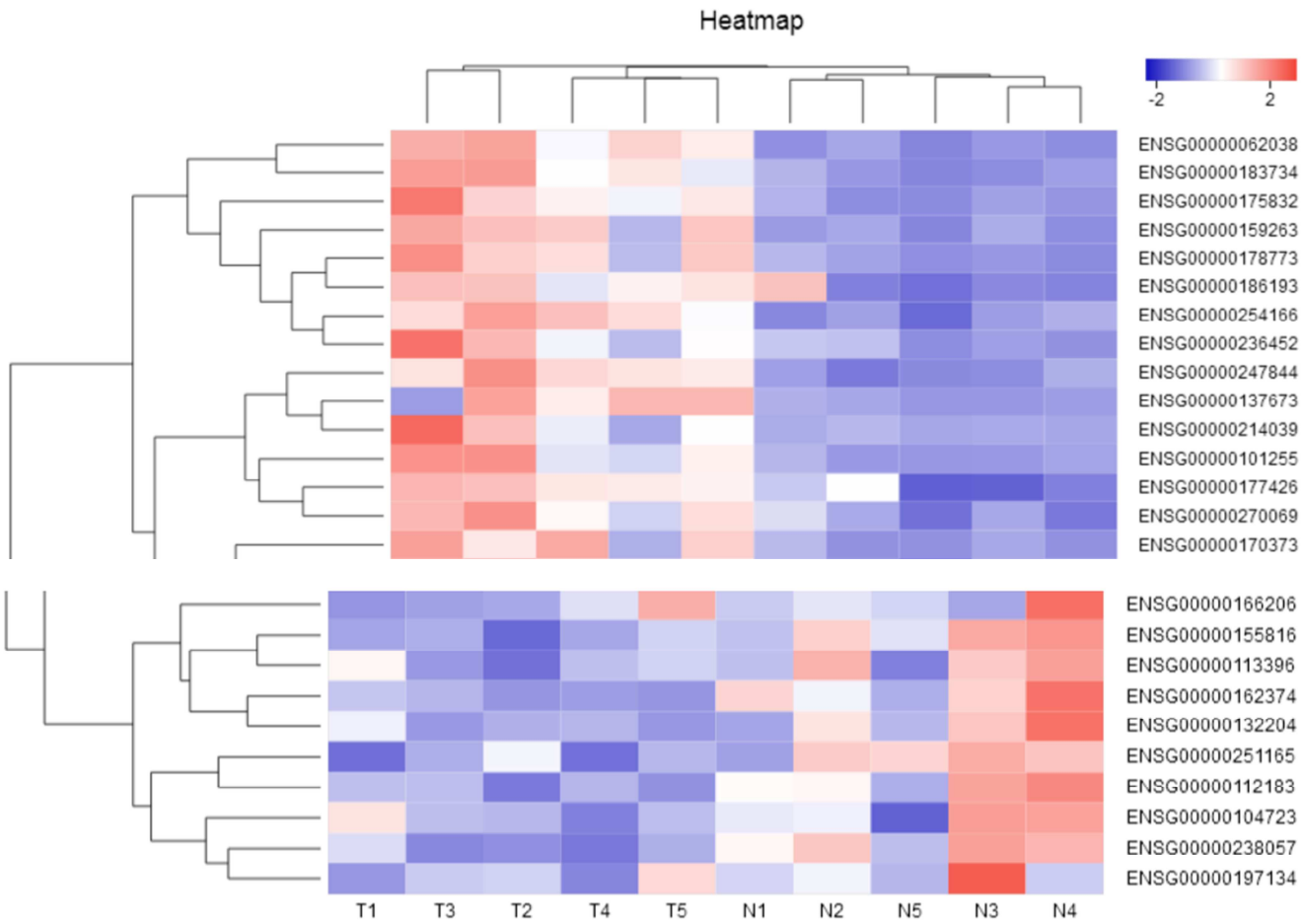

图3 两组正常和肿瘤组织样本的聚类分析热图（附件S2）。

图中的每个列表都显示了一个样本，每行有一个基因。图中的颜色显示了样本中基因表达的大小[log10（TPM+1）]。红色代表样品中基因的高表达，蓝色代表低表达。表达大小的趋势显示在右上角的颜色栏中。基因聚类树型图在左边，基因名在右边。两个基因分支越近，它们的表达就越近。样本聚类的树形图在顶部，下面是样品的名称。两个样本的分支越近，两个样本中所有基因的表达模式越接近，即基因表达的趋势越近。

表3中（附件S3）分别列举出了五个肿瘤组织样本与五个正常组织样本1504个基因表达值的详细数据，表格从左至右依次列举出了基因编号(Gene_id)、基因名称(Gene_name)、基因介绍(Gene Description)、五个肿瘤组织样本基因表达值数据(T1, T3, T2, T4, T5)、五个正常组织样本基因表达值数据(N1, N2, N5, N3, N4)

3.5. 基因集比较中的KEGG功能分析

图4显示了KEGG数据库中两组正常和肿瘤组织样本不同表达基因集的功能注释。基因集分为6个第一类和42个第二类KEGG通路。统计数据也显示在表4中：前三种通路是“环境信息处理中的信号转导”（134个基因），“癌症：人类疾病综述”（96个基因）和“组织系统中的内分泌系统”（77个基因）。

纵坐标是KEGG通路的名称；横坐标是注释到通路中的基因或转录本的数量。KEGG通路分为7类：代谢、遗传信息处理、环境信息处理、细胞过程、生物体系统、人类疾病、药物开发。根据我的样本，它只显示了6个类别。

通过二代测序（NGS）技术，从结直肠癌患者的5对正常组织和肿瘤组织样本中获取转录组信息。在生物信息学分析之后，我总结了最重要的几点提示：1）正常组有1878个特异性基因/转录本，肿瘤组有1181个特异性基因/转录本，两组共有15895个相同的基因/转录本；2）两组样品能很好地采用PC1和PC2降维分析进行分离；3）正常组与肿瘤组有1504个不同表达的基因；4）基因集聚类分析表明，每组5个样本的表达模式相似，每组参与了相似的代谢过程或细胞途径；5）两组基因集的比较KEGG功能分析表明，结直肠癌前三个功能特异性通路是“环境信息处理中的信号转导”（134个基因）、“癌症：人类疾病综述”（96个基因）和“组织系统中的内分泌系统”（77个基因）。

根据这些数据，正常组和肿瘤组的组织样本表现出不同的表达模式，特别是在癌症相关通路中。在详细研究了转录组谱的特征后，我将结直肠癌的基因靶点与此前的研究进行了比较。

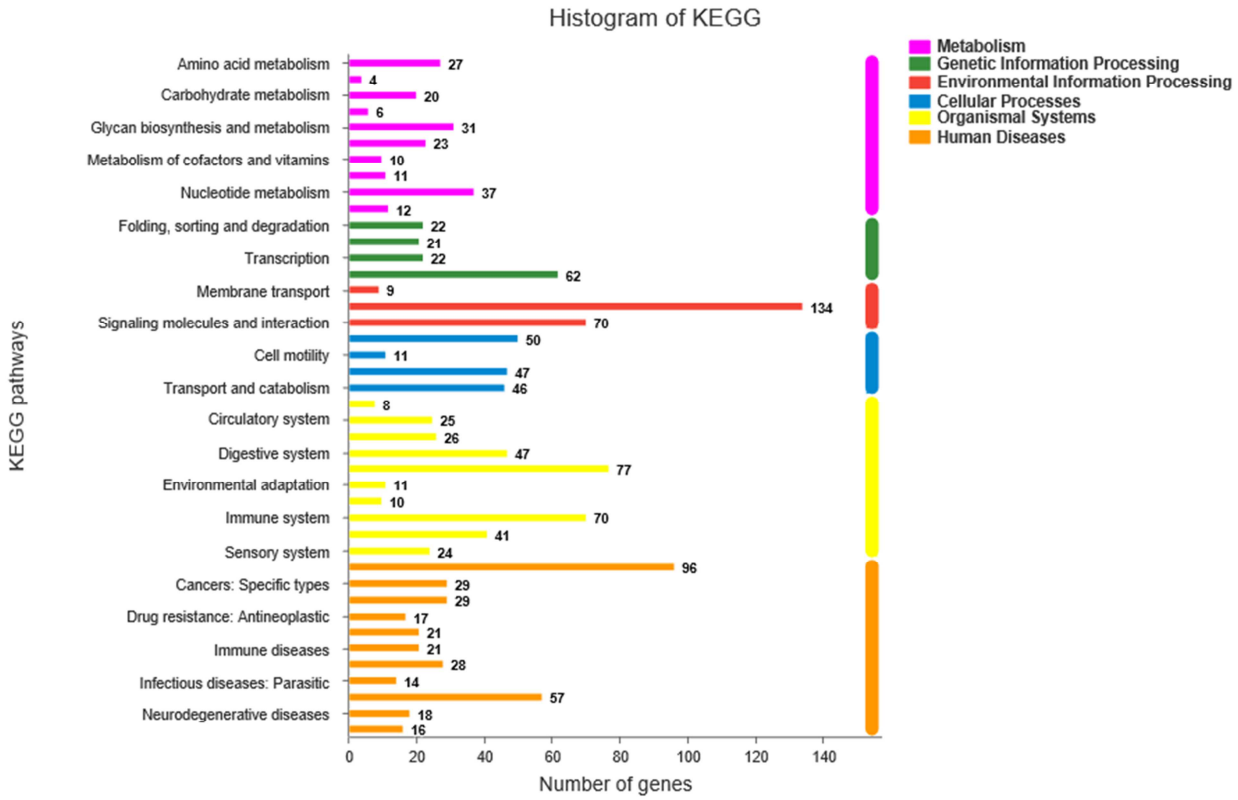

图4 KEGG通路对两组正常和肿瘤组织样本不同表达的基因列表进行的分类统计。

表3 KEGG通路统计分析两组正常和肿瘤组织样本中不同表达的基因集。

First category	Second category	Normal vs Tumor G Number	First category	Second category	Normal vs Tumor G Number
Metabolism	Amino acid metabolism	27	Organismal Systems	Aging	8
Metabolism	Biosynthesis of other secondary metabolites	4	Organismal Systems	Circulatory system	25
Metabolism	Carbohydrate metabolism	20	Organismal Systems	Development	26
Metabolism	Energy metabolism	6	Organismal Systems	Digestive system	47
Metabolism	Glycan biosynthesis and metabolism	31	Organismal Systems	Endocrine system	77
Metabolism	Lipid metabolism	23	Organismal Systems	Environmental adaptation	11
Metabolism	Metabolism of cofactors and vitamins	10	Organismal Systems	Excretory system	10
Metabolism	Metabolism of other amino acids	11	Organismal Systems	Immune system	70
Metabolism	Nucleotide metabolism	37	Organismal Systems	Nervous system	41
Metabolism	Xenobiotics biodegradation and metabolism	12	Organismal Systems	Sensory system	24
Genetic Information Processing	Folding, sorting and degradation	22	Human Diseases	Cancers: Overview	96
Genetic Information Processing	Replication and repair	21	Human Diseases	Cancers: Specific types	29
Genetic Information Processing	Transcription	22	Human Diseases	Cardiovascular diseases	29
Genetic Information Processing	Translation	62	Human Diseases	Drug resistance: Antineoplastic	17
Environmental Information Processing	Membrane transport	9	Human Diseases	Endocrine and metabolic diseases	21
Environmental Information Processing	Signal transduction	134	Human Diseases	Immune diseases	21
Environmental Information Processing	Signaling molecules and interaction	70	Human Diseases	Infectious diseases: Bacterial	28
Cellular Processes	Cell growth and death	50	Human Diseases	Infectious diseases: Parasitic	14
Cellular Processes	Cell motility	11	Human Diseases	Infectious diseases: Viral	57
Cellular Processes	Cellular community -eukaryotes	47	Human Diseases	Neurodegenerative diseases	18
Cellular Processes	Transport and catabolism	46	Human Diseases	Substance dependence	16

近年来发表的几篇论文集中讨论了NGS对CRC/正常组织的转录组学分析。[13-16]其中两个被用来与本研究中数据比较不同表达的基因。在Wu的论文中，研究了肿瘤、邻近非肿瘤和远处正常组织的转录组特征。在由RNA测序和qRT-PCR确认的5个基因中，我的数据显示了两个共同的靶点：COL1A1（I型胶原α1链）和SPP1（分泌型磷蛋白1）。而在Slattery的论文中，除了COL1A1和SPP1外，还发现了更常见的靶点，如E2F1, MYC, CDKN2B, CDK4, CDK2和CCND1，这些靶点主要与癌症通路有关。在他们的论文中，诸如“染色体复制的细胞周期控制”、“雌激素介导的S期进入”和“细胞周期：G1/S检验点调节”等基因富集通路也与本研究中的分析相匹配。

总之，本研究项目收集了一些关于结直肠癌转录组信息的初步数据。它与正常样本不同，验证了以往研究的一些重要指标，为我国结直肠癌的发病机制研究提供了更多的线索。

4．结论

从5例结直肠癌患者的RNA-seq数据中得到的转录组信息显示，在结直肠癌/正常组中有不同的表达模式，有两个靶点COL1A1和SPP1与之前的两篇相似论文一致，并与其中一篇富集基因的通路相匹配，为我国结直肠癌的发病机制提供了更多的研究基础。该项目通过比对得出大肠癌所对应的转录组表达谱特异性，由此可以进一步开展靶向试验和治疗。筛查肿瘤基因的异常表达，为临床提供较为准确的治疗靶点，以及检测、预防肿瘤复发与评估药效。

参考文献

[1] Ferlay, J., Soerjomataram, I. *et al*. (2013). GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase. No. 11. Retrieved 2nd September, 2018 from https: //doi.org/10.1016/j.ucl.2013.01.011.

[2] Van Blarigan EL., Meyerhardt JA. *“Role of physical activity and diet after colorectal cancer diagnosis.”* J Clin Oncol. 2015; 16: 1825–34.

[3] Sanger F. *et al*. *“DNA sequencing with chain-terminating inhibitors.”* Proc. Natl. Acad. Sci. U. S. A. 1977; 74: 5463-5467.

[4] Guan YF., Li GR. *et al*. *“Application of next generations sequencing in clinical oncology to advance personalized treatment of cancer.”* Chin J Cancer. 2012; 31 (10): 463–70.

[5] Sultan M., Schulz MH. *et al*. *“A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome.”* Science. 2008; 321: 956–60.

[6] Ju YS., Kim JI. *et al*. *“Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals.”* Nat Genet.2011; 43: 745–52.

[7] Fu Y., Sun Y. *et al*. *“Differential genome-wide profiling of tandem 3'UTRs among human breast cancer and normal cells by high-throughput sequencing.”* Genome Res. 2011; 21: 741–7.

[8] Kim T., Moon JH. *et al*. “*Next-generation sequencing based post-transplant monitoring of acute myeloid leukemia.”* Blood. 2018 Aug 14. pii: blood-2018-04-848028.

[9] Yadav S. S., Li J. *et al*. *“Next-generation sequencing technology in prostate cancer diagnosis, prognosis, and personalized treatment.”* Urologic Oncology: Seminars and Original Investigations, (2015) 33 (6), pp. 267.e1-267.e13.

[10] Li MY., Sun QR. *et al*.*“Transcriptional Landscape of Human Cancers.”* Oncotarget 8.21 (2017): 34534–34551. PMC. Web. 12 Sept. 2018.

[11] Vogelstein B., Fearon ER. *et al*. *“Genetic alterations during colorectal tumor development.”* N Engl J Med. 1988; 319:525–32

[12] Liu Z. *et al*. *“The Landscape of Somatic Mutation in Sporadic Chinese Colorectal Cancer.”* Oncotarget 9.44 (2018): 27412–27422. *PMC*. Web. 12 Sept. 2018.

[13] Seshagiri S., Stawiski EW. *et al*. *“Recurrent R-spondin fusions in colon cancer.”* Nature. 2012; 488 (7413):660–4.

[14] Wu Y., Wang X. *et al*. *“Transcriptome profiling of the cancer, adjacent non-tumor and distant normal tissues from a colorectal cancer patient by deep sequencing.”* PLoS ONE. 2012; 7(8):e41001. Retrieved from 2nd September, 2018 from *https: //doi.org/10.1371/journal.pone.0041001.*

[15] Slattery ML., Herrick JS. *et al*. *“Improved survival among colon cancer patients with increased differentially expressed pathways.”* BMC Med. 2015; 13: 75.

[16] Li M., Zhao LM. *et al*. *“Differentially expressed lncRNAs and mRNAs identified by NGS analysis in colorectal cancer patients.”* Cancer Med. 2018 Jul 23. doi: 10.1002/cam4.1696.