

乳腺癌癌旁组织特异性表达基因分析

禹奇超^{1,2}, 宋彬^{1,2}, 邹轩轩^{1,2,3}, 王岭⁴, 刘德权⁵, 李波^{1,2}, 马昆^{1,2}

1. 深圳华大生命科学研究院, 深圳 518083
2. 深圳国家基因库, 深圳 518120
3. 中国科学院大学华大教育中心, 深圳 518083
4. 第四军医大学西京医院血管内分泌外科, 西安 710033
5. 云南省肿瘤医院, 昆明医科大学第三附属医院乳腺外科, 昆明 650118

摘要: 癌症研究中常用癌旁组织(normal tissues adjacent to the tumour, NAT)作对照, 而癌旁组织与无肿瘤的正常组织的基因表达谱是有差异的。癌旁组织特异性表达基因的存在通常会干扰传统的转录图谱研究, 然而目前关于癌旁与无肿瘤组织的基因表达谱差异的研究相对较少。本研究对 14 例乳腺癌患者的癌组织、癌旁组织和对侧正常乳腺组织样本进行高深度 RNA 测序和分析, 发现癌旁组织相对对侧正常乳腺组织有 102 个差异表达基因。基因富集和蛋白-蛋白互作分析揭示这些差异表达基因显著富集在肿瘤坏死因子(tumour necrosis factor, TNF)和上皮间质转化(epithelial-mesenchymal transition, EMT)等癌症相关的基因集中。通过比较癌旁组织与癌组织、癌旁组织与对侧正常乳腺组织的转录图谱, 发现 23 个癌旁组织特异性高表达的基因, 即癌旁特异性激活(tumour-adjacent specific activation, TASA)基因。这些基因显著富集在 TNF 基因集中, 其中 15 个是新发现的基因。结果表明, TASA 基因在乳腺癌癌旁组织中普遍存在, 并且与免疫系统的 TNF 信号有关。癌旁中存在类肿瘤型表达模式的基因, 这些基因可能与肿瘤形成有关, 但是往往在肿瘤-癌旁成对研究中被遗漏。

关键词: 乳腺癌; 癌旁特异激活基因; RNA 测序; 基因表达谱

收稿日期: 2019-04-09; 修回日期: 2019-05-15

基金项目: 深圳市科创委项目(编号: JCYJ20150629114130814)和深圳市工信局项目(编号: 20170731162715261)资助[Supported by Science, Technology and Innovation Commission of Shenzhen Municipality (No. JCYJ20150629114130814) and Shenzhen Municipal Government of China (No. 20170731162715261)]

作者简介: 禹奇超, 硕士, 专业方向: 肿瘤基因组学与生物信息学。E-mail: yuqichao@genomics.cn

宋彬, 硕士, 专业方向: 肿瘤基因组学与生物信息学。E-mail: songbin@genomics.cn

禹奇超和宋彬并列第一作者。

通讯作者: 马昆, 博士, 研究方向: 肿瘤基因组学。E-mail: makun1@genomics.cn

DOI: 10.16288/j.ycz.19-099

网络出版时间: 2019/5/24 7:27:27

URI: <http://kns.cnki.net/kcms/detail/11.1913.R.20190524.0726.001.html>

Analysis of normal tissues adjacent to the tumour-specific expressed genes in breast cancer

Qichao Yu^{1,2}, Bin Song^{1,2}, Xuanxuan Zou^{1,2,3}, Ling Wang⁴, Dequan Liu⁵,
Bo Li^{1,2}, Kun Ma^{1,2}

1. BGI-Shenzhen, Shenzhen 518083, China

2. China National GeneBank, BGI-Shenzhen, Shenzhen 518120, China

3. BGI Education Center, University of Chinese Academy of Sciences, Shenzhen 518083, China

4. Department of Vascular and Endocrine Surgery, Xijing Hospital of Fourth Military Medical University, Xi'an 710033, China

5. Department of Mammary Surgery, Yunnan Tumor Hospital, The Third Affiliated Hospital of Kunming Medical University, Kunming 650118, China

Abstract: Normal tissues adjacent to the tumour (NAT) are widely used as controls in comparative studies to search for cancer-associated genes. However, the gene expression profiles between NAT and non-tumour-bearing tissues are different. The presence of NAT-specific expressed genes often hinders traditional transcriptional profiles studies. Further, studies on the differences in gene expression profiles between NAT and tumour-free tissues are infrequently performed. In this study, we sequenced and analysed the transcriptomes of tumour tissues (T), matched NAT and contralateral breast normal tissues (CBN) of 14 breast cancer patients, and identified 102 differentially expressed genes (DEGs) between CBN and NAT. Gene enrichment and protein-protein interaction (PPI) analyses revealed that these DEGs are significantly enriched in TNF (tumour necrosis factor) signalling and EMT (epithelial-mesenchymal transition) gene sets closely associated with oncogenesis. Comparative analyses of the transcriptomic profiles between NAT and CBN, NAT and T identified 23 NAT-specific highly-expressed genes, namely tumour-adjacent specifically activated (TASA) genes. These genes were significantly enriched in TNF signalling gene set, and 15 of which have not been previously reported. The results indicate that TASA genes are common in adjacent tissues and are related to the TNF signalling in the immune system. The tumour-adjacent tissues harbour tumour-like expressed genes that could contribute to tumour initiation but are often missed in NAT-T pair-wise studies.

Keywords: breast cancer; TASA gene; RNA-sequencing; gene expression profile

已有研究表明, 距离肿瘤组织(tumour tissues, T)1 cm 以上的癌旁组织(normal tissues adjacent to the tumour, NAT)与肿瘤组织在 pH^[1]、转录图谱和表观遗传修饰^[2]等方面都存在显著的差异。由于 NAT 可以在肿瘤切除手术中获得, 样本采集相对容易, 长期以来, 绝大多数实体瘤研究都是以 NAT 作为对照, 如癌症基因组图谱计划(The Cancer Genome Atlas, TCGA)^[3]和国际癌症基因组联盟(International Cancer Genome Consortium, ICGC)^[4]。然而, 针对 NAT 的转录组研究却十分有限, NAT 作为对照能否获得真实可靠的肿瘤组织差异表达基因(differen-

tially expressed gene, DEG)仍存在疑问^[5]。

Slaughter 等^[6]认为癌症的形成是一个变异逐渐积累的过程, 将 NAT 视为一种中间状态, 即肿瘤形成前形态学正常但分子水平已经发生变化的一群细胞。根据这个理论, NAT 在分子生物学水平并不是真正的“正常”。因此, 用 NAT 作为对照可能难以获得准确的分子水平上的差异。

人类的多种器官, 包括肺、肾脏、乳房等都是成对存在的, 但是在原发性肿瘤患者中左右两侧同时产生肿瘤的比例却非常低。由此推测, 无肿瘤的对照侧正常组织(contralateral breast normal tissues, CBN)在

分子生物学水平更接近真实的“正常”, 用作癌症研究的对照应更为合适。基于此, 本研究收集了 14 例在身体一侧患有乳腺癌的患者, 对 CBN、NAT 和 T 3 种组织进行 RNA 测序^[7], 并通过对每个患者的 CBN、NAT 和 T 的全转录组各基因进行差异表达分析, 结合基因富集和蛋白-蛋白互作(protein-protein interaction, PPI)分析, 检测到 NAT 相比 CBN 有 102 个差异表达基因。进一步分析发现, 这些差异基因主要富集到肿瘤坏死因子(tumour necrosis factor, TNF)和上皮间质转化(epithelial-mesenchymal transition, EMT)基因集中, 其中有 23 个基因是癌旁特异激活基因, 仅显著富集到 TNF 基因集中。此外, 本研究还发现一些基因在 NAT 中的表达水平与肿瘤组织无显著差异, 但与 CBN 相比有显著差异, 而这些基因往往在以 NAT 为对照的研究中被遗漏。

1 材料与方法

1.1 样本采集

用于本研究的组织材料来自云南省肿瘤医院和第四军医大学西京医院, 从 14 例单侧乳腺癌患者(表 1)获得肿瘤组织、癌旁组织和对侧乳房正常乳腺组织, 这些患者的另一个乳房有非肿瘤良性病

变, 为防止恶化, 良性病变组织被切除。本研究取得了患者的知情同意, 并通过了上述医院和华大基因生命伦理生物安全与遗传资源管理委员会(BGI-IRB)的伦理审查。医院采集了这 14 例患者的肿瘤组织(T), 癌旁组织(NAT, 距离肿瘤组织 5 cm 以上)和对侧正常乳腺组织(CBN), 共计 42 个组织样本。所有样本的 RNA 提取、逆转录、扩增、建库等实验操作由深圳国家基因库完成。cDNA 文库由 BGISEQ-500 测序仪完成单端 50 bp (SE50)和双端 100 bp (PE100)测序。

1.2 数据获取

本研究产生的测序数据存放于国家基因库数据库(CNGBdb)的核酸序列归档系统(CNSA, <https://db.cngb.org/cnsa/>)中, 检索码(accession code)为 CNP-0000385。

1.3 测序数据处理

首先用 SOAPaligner 将原始序列比对到 rRNA 上, 映射到 rRNA 的序列被去除, 剩下的序列用 Bowtie2^[8]比对到人类 RefSeq 参考基因上(hg19)。本研究使用 RSEM^[9]获得样本的基因读段(read)数目及表达水平的结果。

1.4 基因表达谱分析

使用上四分位数(upper quartile)标准化的方法校正了基因的读段数目, 然后使用 edgeR 包^[10]基于标准化后的读段数据进行差异表达基因分析。对每个患者的 3 种组织两两之间都进行了差异表达基因分析。在 2 个或 2 个以下样本中检测到少于 10 条读段的基因被去除。使用错误发现率(false discovery rate, FDR)、倍数变化(fold change, FC)和基因覆盖的读段数目对差异表达基因进行过滤。基因覆盖深度使用 CPM (counts per million), 即每 100 万条读段中映射到某基因上的读段数进行量化。同时满足 $FDR < 0.05$ 、 $\log_2(FC)$ 的绝对值大于 1 且 $\log_2(CPM) > 3$ 的基因被视作差异表达基因。校正后的 CPM 数据用于 Rtsne 降维分析。

1.5 蛋白-蛋白相互作用分析

将得到的差异表达基因集输入在线工具 STRING

表 1 患者临床信息

Table 1 Clinical information of patients

患者编号	年龄	病理类型	患病部位	TNM
P01	46	乳腺浸润癌	左侧	T2N1aM0
P02	58	乳腺浸润性导管癌	左侧	T1cN1M0
P03	53	乳腺浸润性导管癌	左侧	T2N1aM0
P04	46	乳腺浸润性导管癌	左侧	T1cN0M0
P05	50	乳腺浸润性小叶癌	左侧	T2N0M0
P06	43	乳腺浸润癌	左侧	T2N1aM0
P07	44	乳腺浸润性导管癌	左侧	T1N0M0
P08	44	乳腺浸润性导管癌	左侧	T2N0M0
P09	49	乳腺浸润性导管癌	左侧	T2N3M0
P10	44	乳腺浸润性导管癌	右侧	T1N0M0
P11	60	乳腺浸润性导管癌	右侧	T1cN0M0
P12	50	乳腺浸润性导管癌	左侧	T1cN0M0
P13	60	乳腺浸润性导管癌	左侧	T1cN1aM0
P14	32	乳腺浸润癌	左侧	T1bN2aM0

(<https://string-db.org/>)^[11], 筛选出差异表达基因间的相互作用关系, 获得蛋白-蛋白相互作用网络。同时, 该工具还可对得到的 PPI 网络进行图形可视化展示。

1.6 差异表达基因的富集分析

使用 50 个 Hallmark 基因集在基因富集分析 (gene set enrichment analysis, GSEA)^[12]网站(<http://software.broadinstitute.org/gsea/msigdb/annotate.jsp>) 对各组差异表达基因进行富集分析。通过计算了富集分数(enrichment score, ES), 即 $-\log_{10}(P \text{ value})$, 对不同的富集结果进行排序。当 P 值为 0.05 时, 对应的富集分数为 1.3。富集分数越大, 说明富集的显著性越高。

1.7 统计方法

统计分析由 R 程序实现。双侧 t 检验 $P < 0.05$ 被视为具有统计学显著性。

2 结果与分析

2.1 差异表达基因分析

14 例患者 CBN、NAT 和 T 3 种组织样本平均

测序深度分别为 168.4×、207.5×和 186.4×, 平均测序深度达 186.4× (hg19); 平均检测基因数分别为 10 998.7、10 769.1 和 10 580.8, 平均每个组织能检测到 10 782.9 个基因。

平均每个患者 CBN 与 NAT 的差异表达基因的有 967.6 个, 在 55% 以上患者共享的 DEG 有 102 个。NAT 特异性表达基因(即与 CBN 和 T 相比, 在 NAT 中都上调或下调)有 2045 个, 平均每个患者携带 246 个。

CBN、NAT 和 T 3 种组织 t -SNE 降维分析的结果显示 T 形成了一个独立的簇, 说明肿瘤组织与 CBN 和 NAT 的表达谱有明显差异(图 1A)。CBN 与 NAT 形成一个簇, 说明二者的表达谱有一定的相似性。通过比较 3 个组织两两之间检测到的 DEG 数目, 发现 NAT 与 T 的差异表达基因数目(平均值 2477.6), 显著小于 CBN 与 T 的 DEG 数目(平均值 3069.2, $P < 0.05$, 图 1B)。从 DEG 数目上看, 与 CBN 相比, NAT 与肿瘤组织的基因表达谱一致性更高, 即癌旁组织的基因表达谱是处于对侧正常组织与肿瘤组织中间的一种状态。

2.2 NAT 相比 CBN 的差异表达基因分析

对 14 例患者 CBN 相比 NAT 的 DEG 分别进行

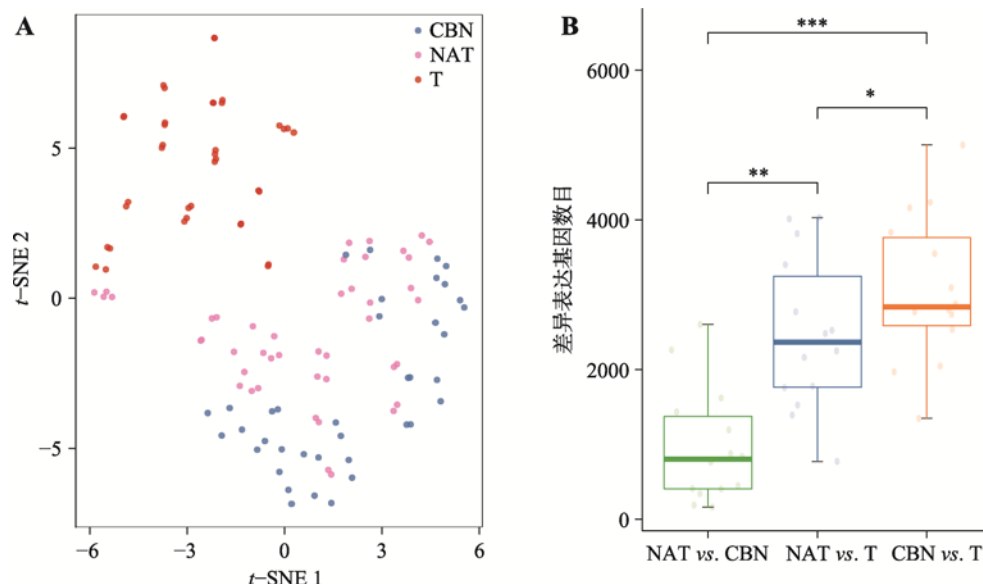


图 1 CBN、NAT 和 T 的转录组特征

Fig. 1 Transcriptomic features of CBN, NAT and T

A: 3 种组织的 t -SNE 图; B: 3 种组织两两比较 DEG 的数目。*: $P < 0.05$, **: $P < 0.01$, ***: $P < 0.001$ 。NAT: 癌旁组织, CBN: 对侧正常乳腺组织, T: 肿瘤组织。

Hallmark 基因富集分析，发现“响应 TNF 受 NF-κB 调控的基因”(HALLMARK_TNFA_SIGNALING_VIA_NFKB，以下简称 TNF 基因集)和 EMT 基因集是富集最为显著的两个基因集(图 2A)。

筛选出在 55% 以上患者中复现的 102 个 DEG 进行蛋白-蛋白互作分析，在互作关系的筛选过程中去掉了可靠性较低的(来源于文本挖掘和数据库)蛋白-蛋白互作关系对，得到更可靠的蛋白互作网络(图

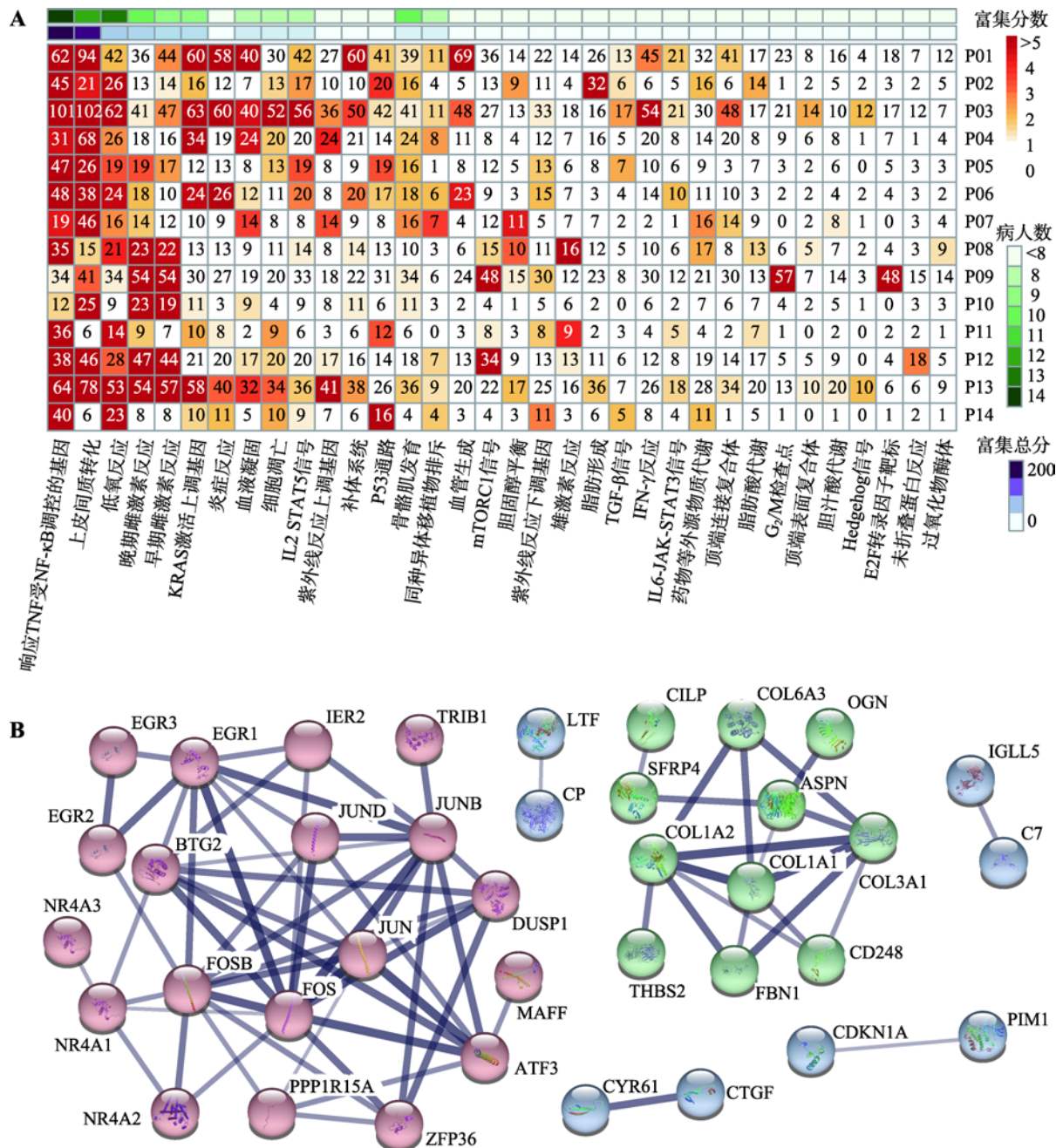


图 2 NAT 相比 CBN 的差异表达基因

Fig. 2 DEGs of NAT vs. CBN

A : 55% 以上患者共有的 102 个 DEG 的 Hallmark 基因富集热图。富集分数 $ES = -\log_{10}(P \text{ value})$ ，富集总分为各样本富集分数之和，热图中的数值表明患者富集到该基因集的 DEG 数目。B : 102 个 DEG 的蛋白-蛋白互作网络。未展示 64 个不在 PPI 网络中的 DEG。包含大于等于 3 个基因的簇(cluster)有 2 个，其中 A 簇(红色)：TNF 基因集；B 簇(绿色)：EMT 基因集。

2B)。这些 DEG 中,有 38 个基因出现在 PPI 网络中(75 条边, $P < 1.0 \times 10^{-16}$),形成 2 个相互独立的簇(cluster)。通过对每个簇中的基因进行富集分析发现, A 簇(红色)显著富集于 TNF 基因集(包含 17 个基因, $P = 6.6 \times 10^{-40}$), B 簇(绿色)显著富集于 EMT 基因集(包含 6 个基因, $P = 5.2 \times 10^{-13}$)。其中 A 簇富集的基因集与以往的研究结果一致^[5]。

2.3 对侧正常乳腺组织、癌旁组织和肿瘤组织的基因表达模式分析

对 CBN、NAT 与 T 进行差异表达分析,根据基因从 CBN 到 NAT 再到 T 的表达水平变化,可将基因的表达模式划分成 4 种类型,即 NAT 特异性表达型(NAT-specific, A 型),梯度型(gradient, G 类),类肿瘤型(tumour-like, T 型)和类正常型(normal-like, N 型)^[5]。同时用“1”和“2”分别表示基因上调和下调,可以将基因的表达模式细分为 8 个小类。对每个病人的 DEG 划分类型后统计每种类型的平均基因数目和高频复现基因数目(定义为 55% 以上患者共有的基因),并分别做了富集分析,计算了平均富集分数(表 2)。

NAT 相比 CBN 和 T 都上调的基因(A1 型)被称

为癌旁特异激活基因(tumour-adjacent specific activation, TASA)。本研究共计检测到 1403 个 TASA 基因,平均每个患者有 190 个。高频复现的 TASA 基因有 23 个,其中包括 8 个已报道的基因: *ATF3*、*CSRNP1*、*EGR1*、*EGR2*、*EGR3*、*FOS*、*FOSB* 和 *NR4A3*, 15 个新的 TASA 基因: *ADAMTS1*、*APOLD1*、*CYR61*、*DUSP1*、*JCHAIN*、*JUN*、*KLF2*、*KLF4*、*MAFF*、*NR4A3*、*PPP1R15A*、*RASD1*、*TNXB*、*VEGFD* 和 *ZFP36*。对上述 23 个 TASA 基因进行 PPI 分析,得到一个包含 19 个节点、27 条边的蛋白互作网络($P < 1.0 \times 10^{-16}$)。富集分析发现,这些基因显著富集在 TNF 基因集中(包含 16 个基因, $P = 1.8 \times 10^{-34}$, 图 3C),与以往的研究结果相似^[5]。值得注意的是,新发现的基因中, *DUSP1*、*JUN*、*KLF2*、*KLF4*、*MAFF*、*NR4A1*、*PPP1R15A* 和 *ZFP36* 也包含在 TNF 基因集中,其中 *KLF4* 和 *JUN* 是 COSMIC 数据库(<https://cancer.sanger.ac.uk/census>)收录的癌基因(图 3, A 和 B)。NAT 相比 CBN 和 T 都下调的基因(A2 型)基因有 642 个,平均每个患者 56 个,没有发现高频复现基因和显著富集结果。此外, A2 型表达模式的基因数目显著低于 A1 型($P < 0.05$)。

平均每个患者检测到 37 个 G1 型(从 CBN、NAT

表 2 基因在对侧正常乳腺组织、癌旁组织和肿瘤组织中的表达模式

Table 2 Expression patterns in CBN, NAT and T

类型	模式	MGC	RGC	富集结果 1(MS)	富集结果 2(MS)
A	UD/DU	246	23	响应 TNF 受 NF- κ B 调控的基因(14.17)	低氧反应(3.43)
G	UU/DD	91	0	上皮间质转化(6.39)	—
T	US/DS	463	1	上皮间质转化(5.70)	响应 TNF 受 NF- κ B 调控的基因(2.69)
N	SU/SD	1672	335	上皮间质转化(10.19)	IFN- γ 反应(5.38)
A1	UD	190	23	响应 TNF 受 NF- κ B 调控的基因(15.32)	低氧反应(3.72)
A2	DU	56	0	—	—
G1	UU	37	0	上皮间质转化(7.34)	—
G2	DD	54	0	—	—
T1	US	154	0	上皮间质转化(7.52)	响应 TNF 受 NF- κ B 调控的基因(4.24)
T2	DS	309	1	—	—
N1	SU	828	111	上皮间质转化(10.29)	IFN- γ 反应(10.12)
N2	SD	844	224	骨骼肌发育(4.26)	早期雌激素反应(3.88)

模式:两个字母分别表示从 CBN 到 NAT, NAT 到 T 的变化趋势,其中 U:上调(up-regulated); D:下调(down-regulated); S:稳定(stable); MGC:平均基因数(mean gene count); RGC:在 55% 以上患者复现的基因数目(recurrent gene count); MS:富集分数的平均值(mean score), $MS = \Sigma[-\log_{10}(P \text{ value})]/n$, n 为患者数目。富集结果 1 和 2 分别展示了在 55% 以上患者中富集且平均富集分数最高的前两个基因集。

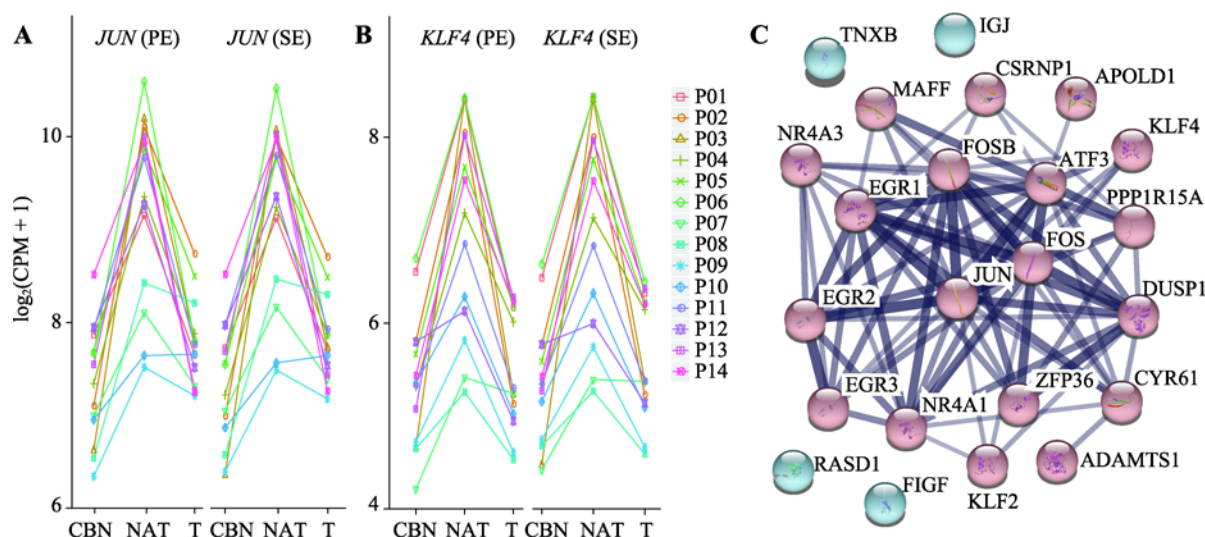


图 3 多个患者中复现的癌旁特异激活基因

Fig. 3 TASA genes in multiple patients

A: *JUN* 基因在 3 种组织中的表达水平; B: *KLF4* 基因在 3 种组织中的表达水平; C: 高频复现的 TASA 基因 PPI 网络。SE: 单端测序数据; PE: 双端测序数据。

到 T 持续上调)基因和 54 个 G2 型(从 CBN、NAT 到 T 持续下调)基因,未发现高频复现基因, G1 类型基因同样显著富集到了 EMT 基因集。平均每个患者检测到 154 个 T1 型基因和 309 个 T2 型基因,而这些基因在常规的肿瘤-癌旁成对转录组研究的 DEG 分析中可能会被遗漏。其中 T1 型显著富集到 EMT 和 TNF 等基因集中, T2 型没有显著富集结果,仅有一个高频复现基因 *MIRLET7BHG*。这些基因中有 85 个的复现率大于 35%, 包含 COSMIC 数据库中 4 个已知的癌基因: *BRIP1*、*CDKN1A*、*COL3A1* 和 *POU5F1*。平均每个患者检测到 828 个 N1 型基因和 844 个 N2 型基因, 高频复现基因数目分别为 111 和 224 个。N1 型基因显著富集到 EMT 基因集和 IFN- γ 反应等基因集中, N2 型基因显著富集到骨骼肌发育和早期雌激素反应等基因集中(表 2)。

A2、G2 和 T2 型基因几乎没有高频复现基因, 且没有显著富集结果。放宽过滤条件发现它们都能富集到早期雌激素反应基因集中, 与 N2 型基因的富集结果类似, 但富集分数远低于 A1、G1 和 T1 型基因, 说明在肿瘤组织中下调的基因的共表达模式与上调基因相比随机性升高。此外, 本研究还分别对 A、G、T、N 4 种模式的基因进行富集分析, 富集结果与 A1、G1、T1、N1 高度一致(表 2)。

3 讨论

通过对 14 例乳腺癌患者 3 种组织(CBN、NAT 和 T)高深度 RNA 测序数据分析, 本研究在比较 NAT 和 CBN 的转录图谱分析中得到了 102 个在大于 55% 患者中复现的 DEG, 表明 NAT 与 CBN 的基因表达谱存在明显差异。通过对 DEG 进行 Hallmark 基因富集分析, 发现这些 DEG 显著富集到 TNF 基因集和 EMT 基因集中。蛋白-蛋白互作分析将这些 DEG 分为 2 个显著的簇, 而这两个簇分别对应了 TNF 基因集(A 簇)和 EMT 基因集(B 簇)。富集到 TNF 基因集的基因包括 *JUN*、*FOS* 和 *FOSB* 等, 富集到 EMT 基因集的基因包括 *COL1A1*、*COL3A1* 和 *COL1A2* 等(图 2B)。

Aran 等^[5]根据 GTEx (the Genotype-Tissue Expression)^[13]的非肿瘤组织样本和 TCGA 的多种癌症组织和癌旁组织样本的基因表达谱发现了 262 个 NAT 特异性表达基因, 通过 PPI 分析将蛋白互作网络分成 4 个簇, 包括细胞分裂(I 簇)、免疫反应(II 簇)、细胞刺激(III 簇)和 ATP(IV 簇)。本研究发现 A 簇中的基因与 Aran 等^[5]报道的 III 簇基因基本一致, 这些基因能够显著富集到 TNF 基因集。其中 *FOS* 和 *JUN* 被认为是促炎性即刻早期基因(immediate-early gene, IEG), 可能参与免疫系统的早期反应。此外, 本研

究的结果与 Aran 等^[5]有所不同:(1)免疫反应(II簇)在本研究的结果中不是富集最显著的基因集,其富集总分(32.9)远小于 TNF(208.8)和 EMT(191.9)基因集;(2)ATP(IV簇)没有出现在本研究的显著富集基因集中。

本研究还发现了 23 个高频复现的 TASA 基因显著富集到 TNF 基因集中。与已报导的 18 个 TASA 基因相比^[5],发现了 15 个新基因,其中 8 个包含在 TNF 基因集中。TNF 信号通路能够影响细胞增殖、分化和免疫反应,同时又可激活促进凋亡或抑制凋亡通路:TNF- α 活化 caspase-8 和 caspase-10 诱导凋亡,又能通过 NF- κ B 抑制凋亡^[14]。本研究发现的 TASA 基因富集在“响应 TNF 受 NF- κ B 调控的基因”集合,说明在癌旁组织中这些 TASA 基因可能激活了抑制凋亡通路,癌旁组织很可能受到了癌细胞的影响或者部分癌旁组织的细胞已经开始癌变。此外, *FOS*、*ATF3*、*JUN*、*PPP1R15A* 和 *KLF4* 同时富集在 TP53 通路中,表明 TP53 通路在癌旁组织中可能被激活。除此之外,新发现的 *DUSP1*、*NR4A1* 和 *VEGFD* 都包含在 MAPK 信号通路中。以往研究表明,异常活化的 MAPK 信号通路在细胞恶性转化及演进中发挥重要作用,并且与乳腺癌、卵巢癌、胃癌和肝癌等癌症的发生与进展密切相关^[15,16]。新发现的 *KLF4* 和 *JUN* 都是 COSMIC 数据库收录的重要的癌基因。*KLF4* 在细胞中扮演了双重角色,它既能够促进细胞存活,在一定条件下又能够促进细胞死亡。有研究报道在原发乳腺导管癌中,*KLF4* 的活化起到促进癌症进展的作用^[17]。然而,也有研究报道在乳腺癌 SK-BR-3 细胞系中激活 *KLF4* 能够促进细胞凋亡,从而抑制肿瘤发生过程^[18]。在乳腺癌 MCF-7 细胞系中,*JUN* 基因的过表达显著增加了癌细胞的致瘤性和侵入性^[19],在乳腺癌肝转移过程中也发挥了关键作用^[20]。本研究基于 14 例散发乳腺癌患者,证实了 TASA 基因在乳腺癌癌旁组织中普遍存在,并且与免疫系统的 TNF 信号有关。

Aran 等^[5]报导的 TASA 基因有 10 个在本研究的数据集中复现率较低,除了 *JUND* 基因的复现率(50%)接近阈值之外,其他基因复现率都小于 40%。据此推测,造成这些差异的原因可能有以下 3 点:(1)本研究的每个患者都包含 CBN、NAT 和 T3 种

组织,来源统一,而 Aran 等^[5]的研究显然使用了不同患者;(2)本研究的样本都是来源于同一种癌症患者,而 Aran 等^[5]使用的对照来自 GTEx,而这些样本不是肿瘤患者;(3)本研究的数据来源于相同的测序平台(BGISEQ-500),而 Aran 等^[5]使用的数据来源于 GTEx 和 TCGA,产出数据的平台不同。来源于同一个患者、同一类癌症、同一种测序平台的数据则更为可靠。

通过分析 A、G、T、N 4 种表达模式的基因,本研究发现 A1、G1、T1、N1 和 N2 型的基因能够显著富集在与癌症相关的基因集中,如 TNF、EMT 和 IFN- γ 反应,而 A2、G2、T2 型的基因未能富集到类似的基因集中,推测这 3 种类型的基因可能与癌症发生和进展的关系较小。

由于 NAT 相对容易获取,包括 TCGA 和 ICGC 在内的绝大多数实体肿瘤转录组研究都使用 NAT 作对照。但是,类肿瘤型(T1 和 T2 型)表达模式的基因在癌旁组织和肿瘤组织中的表达水平无显著差异,而与对侧正常组织相比是有显著差异的。在这种情况下,癌旁-肿瘤配对的转录组分析几乎不可避免地会遗漏这种类型的 DEG。另一方面,癌旁特异性表达基因(A1 和 A2 型)的存在会导致对 DEG 结果的误判,即一些在正常组织和肿瘤组织中表达水平没有显著差异的基因会被错误的认为是 DEG。这会对分析结果造成很大的影响,进而干扰研究的结论。

Ma^[21]提出“两侧对称动物胚胎发育左-右分隔原理”:人的左、右体侧是分别由胚胎 8-细胞期左侧四个细胞衍生的后代和右侧 4 个细胞所衍生的后代构成的。分布在人体左、右体侧的成对器官在发育过程中由于细胞分裂 DNA 复制而产生的体细胞变异是独立的和随机的。如果这些变异会导致某些基因表达的变化,那么建造左、右体侧成对器官细胞的转录组是可能有明显差异的。然而,除非不同的人在同种成对器官上患同类疾病,在构成他们左、右体侧同种成对器官的细胞间复现共同的转录组差异的概率是很小的。根据该原理,建议在癌生物学研究,尤其是成对器官一侧患癌,另一侧未患癌的癌症研究中,尽量获得对侧正常组织样本作为对照,以减少癌旁组织特异性表达基因带来的干扰,提高组学数据分析结果的准确性。

参考文献(References):

- [1] Gerweck LE, Seetharaman K. Cellular pH gradient in tumor versus normal tissue: potential exploitation for the treatment of cancer. *Cancer Res*, 1996, 56(6): 1194–1198. [DOI]
- [2] Heaphy CM, Griffith JK, Bisoffi M. Mammary field cancerization: molecular evidence and clinical importance. *Breast Cancer Res Treat*, 2009, 118(2): 229–239. [DOI]
- [3] Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The cancer genome atlas Pan-Cancer analysis project. *Nat Genet*, 2013, 45(10): 1113–1120. [DOI]
- [4] Hu XD, Yang HM, He J, Lv YY. The cancer genomics and global cancer genome collaboration. *Chin Sci Bull*, 2015, 60(9): 792–804.
胡学达, 杨焕明, 赫捷, 吕有勇. 肿瘤基因组学与全球肿瘤基因组计划. *科学通报*, 2015, 60(9): 792–804. [DOI]
- [5] Aran D, Camarda R, Odegaard J, Paik H, Oskotsky B, Krings G, Goga A, Sirota M, Butte AJ. Comprehensive analysis of normal adjacent to tumor transcriptomes. *Nat Commun*, 2017, 8(1): 1077. [DOI]
- [6] Slaughter DP, Southwick HW, Smejkal W. Field cancerization in oral stratified squamous epithelium; clinical implications of multicentric origin. *Cancer*, 1953, 6(5): 963–968. [DOI]
- [7] Qi YX, Liu YB, Rong WH. RNA-Seq and its applications: a new technology for transcriptomics. *Hereditas(Beijing)*, 2011, 33(11): 1191–1202.
祁云霞, 刘永斌, 荣威恒. 转录组研究新技术:RNA-Seq 及其应用. *遗传*, 2011, 33(11): 1191–1202. [DOI]
- [8] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 2012, 9(4): 357–359. [DOI]
- [9] Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform*, 2011, 12: 323. [DOI]
- [10] Nikolayeva O, Robinson MD. EdgeR for differential RNA-seq and ChIP-seq analysis: an application to stem cell biology. *Methods Mol Biol*, 2014, 1150: 45–79. [DOI]
- [11] Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*, 2014, 43(Database issue): 447–452. [DOI]
- [12] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*, 2005, 102(43): 15545–15550. [DOI]
- [13] Consortium GT. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*, 2013, 45(6): 580–585. [DOI]
- [14] van Antwerp DJ, Martin SJ, Kafri T, Green DR, Verma IM. Suppression of TNF- α -induced apoptosis by NF- κ B. *Science*, 1996, 274(5288): 787–789. [DOI]
- [15] Jia Y, Zhou J, Luo X, Chen M, Chen Y, Wang J, Xiong H, Ying X, Hu W, Zhao W, Deng W, Wang L. KLF4 overcomes tamoxifen resistance by suppressing MAPK signaling pathway and predicts good prognosis in breast cancer. *Cell Signal*, 2018, 42: 165–175. [DOI]
- [16] Delire B, Stärkel P. The Ras/MAPK pathway and hepatocarcinoma: pathogenesis and therapeutic implications. *Eur J Clin Invest*, 2015, 45(6): 609–623. [DOI]
- [17] Foster KW, Frost AR, Mckie-Bell P, Lin CY, Engler JA, Grizzle WE, Ruppert JM. Increase of GSK3 β messenger RNA and protein expression during progression of breast cancer. *Cancer Res*, 2000, 60(22): 6488–6495. [DOI]
- [18] Wang B, Zhao MZ, Cui NP, Lin DD, Zhang AY, Qin Y, Liu CY, Yan WT, Shi JH, Chen BP. Kruppel-like factor 4 induces apoptosis and inhibits tumorigenic progression in SK-BR-3 breast cancer cells. *FEBS Open Bio*, 2015, 5: 147–154. [DOI]
- [19] Smith LM, Wise SC, Hendricks DT, Sabichi AL, Bos T, Reddy P, Brown PH, Birrer MJ. c-Jun overexpression in MCF-7 breast cancer cells produces a tumorigenic, invasive and hormone resistant phenotype. *Oncogene*, 1999, 18(44): 6063–6070. [DOI]
- [20] Zhang Y, Pu X, Shi M, Chen L, Song Y, Qian L, Yuan G, Zhang H, Yu M, Hu M, Shen B, Guo N. Critical role of c-Jun overexpression in liver metastasis of human breast cancer xenograft model. *BMC Cancer*, 2007, 7: 145. [DOI]
- [21] Ma K. Embryonic left-right separation mechanism allows confinement of mutation-induced phenotypes to one lateral body half of bilaterians. *Am J Med Genet A*, 2013, 161A(12): 3095–3114. [DOI]