

通过综合分泌蛋白质组和转录组分析对 CHI3L1 和 MASP2 作为肝癌的共同生物标志物进行鉴定

原发性肝细胞癌（HCC）是排名第五位的最常见肿瘤，每年有超过 50 万的确诊新病例。需要发现新的肝癌标志物。通过串联质谱法分析 12 对肝癌和对应癌旁组织的分泌蛋白质组，鉴定了 1528 个蛋白，发现>2 个的特殊肽。假阳性率为 3.4%。利用光谱计数发现 HCC 组中有 87 蛋白，正常组中有 86 个蛋白表现出五倍表达正常组。这些数据中很可能存在候选的生物标志物。我们采用了区分生物标志物组合的新的参数，该参数包括上调的癌症生物标志物和下调器官富集的分子标记，运用转录组和蛋白质组整合分析确定了几丁质酶 3 样蛋白 1 (CHI3L1) 和甘露聚糖结合凝集素丝氨酸蛋白酶-2 (MASP2) 是肝癌诊断的首选生物标志物组合。

采用酶联免疫吸附测定法，进一步在 25 例肝癌患者血清标本和 15 例年龄匹配的正常对照标本中对这种生物标志物进行评估。组合标记（YKL40/MASP2 比值）作为肝癌诊断的表现比任何单独标记好，AUC 为 0.97。在肝癌患者与疾病对照组和独立的群体中对生物标志物对进一步验证是必要的。

原发性肝细胞癌（HCC）在常见肿瘤中排第五，其导致的癌症相关死亡排第三位，每年有超过 50 万确诊为新发病例 [1]。早期发现的肝癌是可以手术治愈的，肝硬化患者通过生物标志物甲胎蛋白（AFP）检测和每 6 个月 B 超检测早期肝癌 [2]。不幸的是，虽然 AFP 被广泛用于诊断肝癌和监测 [3,4]，其假阴性或假阳性率高达 40% [5, 6]。直接将血清样本用于疾病的诊断和预后分析有几个关键的优势，包括低侵袭，最低的成本，方便样品的采集和处理。

然而，由于血清蛋白质组的复杂性和至少为 10^9 - 10^{10} 巨大动态范围为 [7]，蛋白质组学直接分析本质上是具有挑战性的。我们从肝癌和健侧周围组织培养基所采用的方法是 (i) 分析分泌组（分泌蛋白），(ii) 分析肝癌与健侧周围组织的转录组识别差异表达的基因，(iii) 综合转录组和分泌组分析，找出首选的血清标志物，(IV)，使用酶联免疫吸附试验验证候选生物标志物，以及 (v) 结合癌症生物标志物与器官富集生物标志物以提高准确性诊断。使用上面的方法，我们确定了生物标志物 CHI3L1 和 MASP2 组合进行肝癌诊断的 AUC 为 0.97。

2. 实验材料和方法

2.1 临床样本收集

经过 IRB 的批准收集，肝癌和良性相邻配对组织（距离肝癌组织的边缘至

少为 2 厘米) 来自浙江大学附属第一医院接受肝切除或肝移植的 12 肝癌患者。有 12 例的临床和病理资料汇总于表 1。这些患者手术前没有接受抗肿瘤治疗。额外的血清样本来自于浙江大学附属第一医院血清库。

Table 1. Pathologic data from 12 HCC patients used in LC/MS/MS analysis

Patient no.	Gender	Age	Size (cm)	TNM grade	Edmondson grade	Pathologic data	HBV	HCV	Fibrosis stage	Child-pugh grade
P1	M	45	5.1 × 3.5	II	III	Hepatocellular carcinoma (right lobe), grade III, mixed nodular type hepatocirrhosis	+		4	A
P2	M	68	3 × 3	II	III	Hepatocellular carcinoma (right lobe), grade III	+		4	A
P3	F	71	4.9 × 4.0	II	III	Hepatocellular carcinoma (right lobe), grade III, mixed nodular type hepatocirrhosis	+		4	A
P4	M	82	3 × 2	II	II	Hepatocellular carcinoma (left lobe), grade	-		4	A
P5	M	64	5.9 × 7.7	IVa	IV	Carcinoma sarcomatodes (right lobe), grade a, small nodular type hepatocirrhosis	+		4	A
P6	M	50	7.1 × 7.8	IVa	III	Hepatocellular carcinoma (left lobe), grade III, mixed nodular type hepatocirrhosis	+		4	A
P7	M	57	4.5 × 3.7	IVa	III	Hepatocellular carcinoma (right lobe), grade III, mixed nodular type hepatocirrhosis	+		4	B
P8	M	65	4.2 × 4	II	III	Mixed hepatocellular and cholangio-cellular carcinoma (right lobe) grade III, small nodular type hepatocirrhosis		+	4	B
P9	M	45	2.4 × 2.5	II	I	Hepatocellular carcinoma (right lobe), grade I, mixed nodular type hepatocirrhosis	+		4	A
P10	F	49	4 × 3	IIIa	III	hepatocellular carcinoma (right lobe), grade III, mixed nodular type hepatocirrhosis	+		4	A
P11	M	70	2.7 × 2.3	II	III	Hepatocellular carcinoma (right lobe), grade III, mixed nodular type hepatocirrhosis	+		4	A
P12	M	39	3.3 × 3.2	IIIa	IV	Carcinoma sarcomatodes (right lobe), grade, mixed nodular type hepatocirrhosis, mixed nodular type hepatocirrhosis	+		4	A

2.2 组织培养

成对的组织转移到含有 20 毫升 PBS 的培养皿中，并用剪刀剁碎成 2-3-立方毫米片段。此后，组织碎片再悬浮于 50mL PBS 中，并经过不锈钢过滤器（200 毫米直径）以丢弃单一细胞和细胞碎片。所收集的组织用 PBS 洗涤三次，重悬在 20ml 无血清 DMEM（Sigma 公司，圣路易斯，密苏里州）的培养皿中。该组织碎块在细胞培养孵育箱（Thermo Scientific 的，Milford，马萨诸塞州），用 5% CO₂ 中培养在 37C。

2.3 蛋白质组学样本的制备

在组织培养 24 小时后收集上清液。将上清液离心 20000g 10 分钟以除去可能包含在上清液中的任何细胞或细胞碎片。SpeedVac 中（LABCONCO Centrivap 集中，堪萨斯城，密苏里州）对样品浓缩约 20 倍，并重新悬浮于 25 mM 的碳酸氢铵（NH₄ HCO₃，Sigma 公司，圣路易斯，密苏里州）中。取 60 毫克每个样

品，在 12% SDS 聚丙烯酰胺凝胶上进行分离。凝胶用胶体 CBB 染色。凝胶中的蛋白质使用 Pierce 凝胶内胰蛋白酶消化试剂盒（皮尔斯生物技术，罗克福德，IL）进行胰蛋白酶消化。

2.4 质谱分析

肽段混合物通过配备一个捕获柱（Dionex 公司/ LC 包装间柱前盒 P/ N160454 C18 PepMap100, 5 毫米分离 100 Å, 300 毫米内径×5 毫米，桑尼维尔，CA）和纳米柱（戴安/ LC 包装 P/ N16032115060.075-MMID, C18 PepMap3 毫米, 100Å）的 Ettan MDLC 纳流/毛细管液相色谱系统（GE Healthcare, 匹兹堡, 宾夕法尼亚州）分离，然后经过纳流配置（nanospray configuration）的 LTQ-轨道阱（Thermo Finnigan, 不来梅, 德国）。母离子扫描质谱（M / Z300-1600）所用的 Orbitrap 的分辨率 $R=60000$, m/z 为 400, 累计获得离子是 16106。五个峰值最高的离子被分离和打碎（数累积离子: 36104）。由此产生的碎片离子，记录范围是分辨率 $R=15000$ m/z 为 400。

2.5 质谱数据分析

该 BioWorks 程序 V3.2（Thermo Electron 公司，沃尔瑟姆，马萨诸塞州）的 extract_msn 被用于生成 MS 峰列表，参数为默认。Xcalibur（Thermo Electron 公司）的 ICIS 峰检测算法被用于质谱的自动检测。SEQUEST 算法（Thermo Fisher）被用于 SEQUEST 数据库搜索，搜索使用了 ipi.HUMAN.v3.29.fasta 蛋白质数据库（有 70,757 个条目）（<http://www.ebi.ac.uk/IPI/IPIhuman.html>）。在 TurboSEQUEST 搜索参数设定中，DTA 阈值为 10 000，母离子 DTA 容忍度。为 1.4。SEQUEST 搜索中，肽段偏差设置为 3Da，碎片离子偏差设定在 0.01Da。PeptideProphet™ [8] 用来评估对 MS / MS 谱的质量，接受单张 MS / MS 谱图的 p 值被设定为 0.9，对应于 0.5% 的错误率。允许一个漏切的胰蛋白酶裂解。羧酰胺甲基半胱氨酸（Cys_CAM）（157）是碘乙酰胺还原和烷基化后的固定修饰。由于蛋白质制备用 PAGE，半胱氨酸可能会反应生成丙半胱氨酸（Cys_PAM）。

考虑到潜在丙半胱氨酸，我们增加了 14Da 的可选修饰（CYS- PAM 和 Cys-CAM 之间的质量差是 14）。甲硫氨酸氧化（116 Da）被选为另一种可选修饰。ProteinProphet P 值大于 0.9 并且有两个以上的唯一肽段的蛋白质被视为真实命中。ipi.HUMAN.v3.29.fasta 生成的随机数据库被用作诱饵的数据库来计算蛋白质鉴定的假阳性率。用于随机化的 perl 脚本是从 www.matrixscience.com/download/decoy.pl.gz 下载的。配到随机数据库与比配上 IPI.HUMAN.v3.29.fasta 数据库的比值计算假阳性率（FDR）。

2.6 谱图计数和基因本体论分析

我们将 HCC 组（12 个样品）和对照组（12 个样本）中频谱数量相加。通过发现的所有蛋白质的 *pectra* 的总数对谱图数量进行归一化。计算谱图在肝癌和对照组的比率。oMiner[9]被用来寻找统计上差异显著的 ontology（GO）的类别。在 1528 个至少两次 hits 的蛋白质用作总的输入对有差异的基因使用证据级别 3 进行了分析。

2.7 分析

来自 HCC 和周围的组织上分离蛋白质通过 12% 聚丙烯酰胺凝胶分离，并转移到 PVDF 膜（Amersham Pharmacia 公司。生物技术，ppsala，瑞典）。这些印迹在室温在 TBST 缓冲液温育 2 小时（20mM Tris-CL，40 mMNaCl，pH 值 7.5，0.05 % 吐温 20），含有 5 % 脱脂牛奶，然后用第一抗体抗 anti-AAT Alpha-1-antitrypsin, IPI00553177）（Santa Cruz 公司，加利福尼亚州），4°C 过夜。在 TBST 洗三次之后，室温与 HRP 标记的二抗杂交（1 小时 1:10000，Santa Cruz 公司）。ECL 试剂被用于显影（Pierce Biotechnology）。

2.8 酶联免疫吸附测定法（ELISA）分析

用于鉴定 CHI3L1（YKL40）和 MASP2 的 ELISA 试剂盒购自 Quidel 的公司（圣地亚哥，CA）和 Hycult 生物技术 BV（Uden 的，荷兰）。操作根据制造商的说明书进行。分析之前，血清样本用 PBS 缓冲液稀释三次。ROC 曲线用 GB STAT V10.0 分析（Dynamic Micro-systems, Silver Spring, MD）。

3 结果

3.1 在肝癌的分泌组与正常周围组织之间鉴定差异表达的蛋白

我们在无血清培养条件下比较了 12 对的肝癌及健侧组织培养基（分泌组）中的蛋白组。我们分别鉴定出 1107 和 977 种 ProteinProphet P 值大于 0.9 的蛋白，其中每个蛋白都有两个以上特征肽段分别符合肝癌和正常分泌组（支持信息表 1 和表 2）。另有 70 种蛋白有一个特征肽段在肝癌或正常分泌组中出现，并且该肽段在肝癌分泌组中与在正常组织分泌组中不同。当合并来自肝癌分泌组和正常组织分泌组数据后，这些蛋白质有两个特征肽段。因此，最后有超过 2 个肽段符合的特征肽段总数目为 1528。使用 ipi.HUMAN.v3.29.fasta 的随机数据库作为诱饵数据库，我们计算出这 1528 个蛋白的错误发现率（FDR）为 3.4%。

我们采用质谱计数法[10]来对肝癌和正常组织分泌组进行半定量比较分析。我们比较了 12 个肝癌样本与 12 例正常对照。我们在肝癌组里确定了 87 种过度

表达的蛋白（大于等于 5 倍）（支持信息表 3）在正常组织中发现了 86 种过度表达的蛋白（大于等于 5 倍）（支持信息表 4）。通过质谱计数，比较肝癌与正常组织，AAT（ α -1-抗胰蛋白酶前体， IPI00553177）显示 3.8 倍（1249 在肝癌组中，325 在正常组中）过表达。通过 Western-blot 分析证实，相比于正常组，AAT 确实在肝癌组中高表达（支持资料图 1）。

采用 GO 分类的聚类分析差异表达的基因是由 GoMiner 进行分析的[9]。由 GO 细胞成分分类，在肝癌分泌组中的过表达蛋白中，存在丰富的胞外区蛋白（GO: 0005576）。在 73 个可被映射到胞外区域 GO 项的蛋白，其中有 10 个属于肝癌高表达组，但没有属于对照过表达组的，表明肝癌可能会增加分泌活动。通过 GO 生物学过程项，我们发现，富集在 HCC 中过表达的蛋白质中的 GO 项包括肝素结合（GO: 0008201），钙调蛋白结合（GO: 0005516），糖胺聚糖结合（GO: 0005539），和 I- κ B 的蛋白激酶的 NF- κ B 的级联反应（GO: 0007249）（图 1）。有趣的肝癌聚类低表达蛋白包括雄激素代谢过程（GO : 0008209）（图 1）。AK1C4 和 AK1D1（醛酮还原酶家庭 1 成员 C4 和成员 D1;也被命名为 3 α -羟基类固醇脱氢酶和类固醇 5- β 还原酶，分别）是两种蛋白质在 GO 项 0008209，它们在肝癌组里低表达。这两种蛋白质都参与了雄激素和雌激素代谢（KEGG 通路 hsa00150）。AK1D1 (E.C. 1.3.99.6) 是参与睾酮转化到的更有效的 5- β -二氢睾酮（KEGG 通路 has00150）。

3.2 整合蛋白质组学数据与转录数据

质谱计数方法仅是半定量的。我们试图整合我们的半定量蛋白质组学数据与公共数据和我们的转录组数据来帮助我们识别验证和优先候选基因。采用 Illumina 的新一代测序技术，我们生成了一个肝癌及其健侧周围组织的数字转录组数据。在数字表达图谱上，转录本的丰度被定义为每百万（tpm）中的转录本数目[11]。我们整合在转录组数据和蛋白质组学数据。为了优化我们的基因列表，我们设置了一个过滤器，质谱计数用于蛋白质组学数据将是大于 2 和在转录组数据的比例是大于 3 和癌与正常之间 tpm 绝对差异大于 30（来增加比较低表达基因与低 TPM 值的可信度）。表 2 列出了通过这种综合分析得出的 103 个优先候选基因。当选择最优先的基因，以便进一步验证，我们还比较了我们的数据与 Oncomine 数据库中（www.oncomine.org）（数据未示出）肝癌的基因芯片数据[12 - 14]。

HCC UP GO Terms	HCC Down GO Terms	Go Terms Biological Process
		GO:0009892_negative_regulation_of_metabolic_process
		GO:0051248_negative_regulation_of_protein_metabolic_process
		GO:0051129_negative_regulation_of_cell_organization_and_biogenesis
		GO:0051128_regulation_of_cellular_component_organization_and_biogenesis
		GO:0031324_negative_regulation_of_cellular_metabolic_process
		GO:0048523_negative_regulation_of_cellular_process
		GO:0030042_actin_filament_depolymerization
		GO:0030834_regulation_of_actin_filament_depolymerization
		GO:0030835_negative_regulation_of_actin_filament_depolymerization
		GO:0051016_barbed-end_actin_filament_capping
		GO:0051053_negative_regulation_of_DNA_metabolic_process
		GO:0051693_actin_filament_capping
		GO:0048731_system_development
		GO:0048519_negative_regulation_of_biological_process
		GO:0045934_negative_regulation_of_nucleobase__nucleoside__nucleotide_and_nucleic_acid_metabolic_process
		GO:0007275_multicellular_organismal_development
		GO:0032501_multicellular_organismal_process
		GO:0048513_organ_development
		GO:0050789_regulation_of_biological_process
		GO:0019222_regulation_of_metabolic_process
		GO:0050794_regulation_of_cellular_process
		GO:0008064_regulation_of_actin_polymerization_and_or_depolymerization
		GO:0030832_regulation_of_actin_filament_length
		GO:0051261_protein_depolymerization
		GO:0051246_regulation_of_protein_metabolic_process
		GO:0065007_biological_regulation
		GO:0048856_anatomical_structure_development
		GO:0006996_organelle_organization_and_biogenesis
		GO:0006323_DNA_packaging
		GO:0006325_establishment_and_or_maintenance_of_chromatin_architecture
		GO:0007517_muscle_development
		GO:0051052_regulation_of_DNA_metabolic_process
		GO:0007001_chromosome_organization_and_biogenesis_(sensu_Eukaryota)
		GO:0007249_I-kappaB_kinase_NF-kappaB_cascade
		GO:0051276_chromosome_organization_and_biogenesis
		GO:0032502_developmental_process
		GO:0006333_chromatin_assembly_or_disassembly
		GO:0006334_nucleosome_assembly
		GO:0008154_actin_polymerization_and_or_depolymerization
		GO:0031497_chromatin_assembly
		GO:0050767_regulation_of_neurogenesis
		GO:0042127_regulation_of_cell_proliferation
		GO:0019219_regulation_of_nucleobase__nucleoside__nucleotide_and_nucleic_acid_metabolic_process
		GO:0007010_cytoskeleton_organization_and_biogenesis
		GO:0001501_skeletal_development
		GO:0048518_positive_regulation_of_biological_process
		GO:0008152_metabolic_process
		GO:0006066_alcohol_metabolic_process
		GO:0044237_cellular_metabolic_process
		GO:0019752_carboxylic_acid_metabolic_process
		GO:0006082_organic_acid_metabolic_process
		GO:0032787_monocarboxylic_acid_metabolic_process
		GO:0042445_hormone_metabolic_process
		GO:0006760_folic_acid_and_derivative_metabolic_process
		GO:0008209_androgen_metabolic_process
		GO:0006725_aromatic_compound_metabolic_process
		GO:0006807_nitrogen_compound_metabolic_process
		GO:0009308_amine_metabolic_process
		GO:0006067_ethanol_metabolic_process
		GO:0006069_ethanol_oxidation
		GO:0006544_glycine_metabolic_process
		GO:0006752_group_transfer_coenzyme_metabolic_process
		GO:0008206_bile_acid_metabolic_process
		GO:0008202_steroid_metabolic_process
		GO:0006520_amino_acid_metabolic_process
		GO:0006629_lipid_metabolic_process
		GO:0009069_serine_family_amino_acid_metabolic_process
		GO:0044255_cellular_lipid_metabolic_process
		GO:0006519_amino_acid_and_derivative_metabolic_process
		GO:0006118_electron_transport
		GO:0007586_digestion
		GO:0006732_coenzyme_metabolic_process

图 1 对于肝癌以及正常组织中发现的 5 倍高表达的基因进行 GO 生物学过程聚类项分析。聚类 p 值小于 0.05 的用深灰色表示，p 值大于 0.05 的用浅灰色

3.3 鉴定出肝癌血清标志物几丁质酶 3 类似蛋白 (CHI3L1)

我们搜索了具有商业 ELISA 试剂盒或者好的抗体的差异表达蛋白以便于确认分析。几丁质酶 3 类似蛋白 1 (CHI3L1)，这是其中一个在表 2 中排名很靠前的基因，具备一个 ELISA 试剂盒。CHI3L1 是一个 39 kDa 的分泌糖蛋白属于几丁质酶蛋白家族，它在巨噬细胞分化和组织重塑中发挥作用[15, 16]。我们发现 CHI3L1 只在肝癌分泌组中被鉴定出，但是没有在正常分泌组中出现过(表 2)。在转录组中分析，它是最佳的差异表达基因之一（在肝癌中 3335 tpm 在健侧周围组织中 15 tpm）。通过 Chen 等[12]发表的肝癌芯片数据，CHI3L1 被证明是表达在癌组织中比正常组织高表达（t-检验：3.742；p 值：2.5E -4；数据来自 www.oncomine.org）。我们购买 CHI3L1（别名 YKL-40）ELISA 试剂盒并直接对 25 例肝癌病人的血清标本和 15 例年龄匹配的正常对照组进行验证来评估这个分子标记物。均值和血清表达的范围水平显示为箱须图（图 2，上图）。详细临床资料在支持材料表 5 中提供。ROC 曲线表明 CHI3L1 的 AUC 为 0.92，截断值为 57.3，敏感度为 0.92，特异性 0.87，准确度 0.9（图 3A）。

3.4 结合肝富集表达的生物标志物 MASP2 与 CHI3L1 来进行肝癌诊断

因为 YKL-40 似乎是一个对多种类型的癌症，包括神经胶质瘤[17]，子宫内膜癌[18]和卵巢癌[19]通用的标记物，我们试图增加一个组织富集表达的基因，在这种情况下，肝富集表达基因，并且分泌和表达改变与疾病相关。我们先前提出鉴定组织丰富的表达基因作为生物标志物[20]。我们发现，MASP2（甘露聚糖结合凝集素丝氨酸蛋白酶 2）是一种肝组织富集表达的基因，它被预测为分泌型蛋白质，它包含一个信号肽（信号肽概率：1.000 通过 SignalP3.0 软件预测服务器 <http://www.cbs.dtu.dk/service/SignalP/>），另外有一个奇怪的得分为 2.52 的由 SecretomeP 程序预测的分泌蛋白（<http://www.cbs.dtu.dk/services/SecretomeP/>）。MASP2 是第二个确定的血清蛋白酶成员在激活甘露聚糖结合凝集补体系统过程中发挥作用[21]。我们的 MPSS 数据（数据未所示），以及 Chen 的肝癌芯片[12]的数据表明此基因在正常肝组织中相比肝癌组织是下调的。我们推测，肝癌可能影响正常肝脏的功能，因此下调了肝富集表达基因的表达。我们购买 MASP2 酶联免疫试剂盒，并直接对 25 例肝癌病人的血清标本和 15 例年龄匹配的正常对照组进行验证来评估这个分子标记物(支持信息表 5)。血清表达水平的平均值和范围显示为盒须图(图 2，下图)。ROC 曲线分析显示 MASP2 具有 0.83 的 AUC。使用 292.9 作为截断值，灵敏度为 0.87，特异度是 0.77，精度是 0.8（图 3B）。然后，我们计算 YKL40 和 MASP2

作为组合标志物在肝癌诊断中的可行概率。我们发现，使用的截止值 14.36 增加了单独使用任一标志物检测灵敏度（0.96 为合并后的灵敏度 0.92 和 0.87 为单个标记物的灵敏度），剩余的准确性在 0.9（图 3C），特异性为 0.8（高于单独使用 MASP2 但比单独使用 CHI3L1 与先前描述的截止值的情况下低）。以 YKL40/MASP2 比率作为肝癌诊断的 AUC 为 0.97，其中大于 MASP2 的 AUC（0.83）和 YKL-40 的 AUC 0.92。

4 讨论

从组织间隙液蛋白质的鉴定或条件化的细胞培养基作为候选生物标志物和治疗目标之前已经提出[22 ， 23]。在这里，我们将这个类似的方法在鉴定潜在的体外肝癌组织培养分泌蛋白作为生物标志物的候选。在 GO 分析，在体外组织培养基蛋白中还包括许多细胞内蛋白。

Table 2. Top candidates identified by integrative proteomics and transcriptomics analysis

ID	GeneID	Symbol	Description	Transcriptomics HCC (tpm)	Transcriptomics normal (tpm)	Proteomics Spectral counting (HCC) ¹	Proteomics Spectral counting (normal)
IP100002147	1116.0	CHI3L1	Chitinase-3-like protein 1 precursor	3335	15	4.3	0
IP100220827	9168.0	TMSB10	Thymosin beta-10	2038	217	7.2	0
IP100029737	2182.0	ACSL4	Isoform Long of Long-chain-fatty-acid-CoA ligase 4	1337	14	5.8	0
IP100008527	6176.0	RPLP1	60S acidic ribosomal protein P1	1691	551	15.9	5
IP100032292	7076.0	TIMP1	Metalloproteinase inhibitor 1 precursor	1277	194	53.6	0
IP100221093	6218.0	RPS17	40S ribosomal protein S17	1587	506	8.7	0
IP100012011	1072.0	CFL1	Cofilin-1	1185	275	173.9	45
IP100007221	5104.0	SERPINAS	Plasma serine protease inhibitor precursor	1015	257	7.2	0
IP100418169	302.0	ANXA2	Annexin A2 isoform 1	871	122	68.1	10
IP100221224	290.0	ANPEP	Aminopeptidase N	980	287	142.0	14
IP100006114	5176.0	SERPINF1	Pigment epithelium-derived factor precursor	826	229	137.7	5
IP100219038	3020.0	H3F3A	Histone H3.3	699	140	97.1	20
IP100465439	226.0	ALDOA	Fructose-bisphosphate aldolase A	640	83	323.2	64
IP100014572	6678.0	SPARC	SPARC precursor	573	97	4.3	0
IP100024320	5835.0	RBM3	Putative RNA-binding protein 3	462	49	7.2	0
IP100221222	10823.0	SUB1	Activated RNA polymerase II transcriptional coactivator p15	452	81	85.5	27
IP100215914	375.0	ARF1	ADP-ribosylation factor 1	495	129	30.4	11
IP100216308	7416.0	VDAC1	Voltage-dependent anion-selective channel protein 1	517	153	4.3	0
IP100020956	3068.0	HDGF	Hepatoma-derived growth factor	476	149	144.9	32
IP100514127	3068.0	HDGF	Hepatoma-derived growth factor	476	149	124.6	29
IP100022443	174.0	AFP	Alpha-fetoprotein precursor	329	2	2.9	0
IP100014587	1211.0	CLTA	Isoform Brain of Clathrin light chain A	387	73	58.0	15
IP100023673	3959.0	LGALS3BP	Galectin-3-binding protein precursor	274	40	17.4	7
IP100302592	2316.0	FLNA	Filamin A, alpha	251	32	1153.6	214
IP100029266	6635.0	SNRPE	Small nuclear ribonucleoprotein E	247	47	40.6	9
IP100010790	633.0	BGN	Biglycan precursor	259	71	58.0	5
IP100304612	23521.0	RPL13A	60S ribosomal protein L13a	244	55	4.3	0
IP100027285	6628.0	SNRPB	Isoform SM-B' of Small nuclear ribonucleoprotein-associated proteins B and B'	200	17	37.7	17
IP100009750	3960.0	LGALS4	Galectin-4	196	21	21.7	10
IP100304962	1278.0	COL1A2	Collagen alpha-2(I) chain precursor	240	65	2.9	0
IP100643041	5901.0	RAN	GTP-binding nuclear protein Ran	238	66	15.9	0
IP100025491	1973.0	EIF4A1	Eukaryotic initiation factor 4A-1	239	68	68.1	5
IP100010896	1192.0	CLIC1	Chloride intracellular channel protein 1	203	34	40.6	2
IP100296099	7057.0	THBS1	Thrombospondin-1 precursor	177	22	333.3	7
IP100030910	4076.0	CAPRIN1	GPI-anchored membrane protein 1	212	62	15.9	3
IP100550363	8407.0	TAGLN2	Transgelin-2	197	51	194.2	68
IP100413778	2280.0	FKBP1A	FKBP1A protein	200	57	50.7	17
IP100292452	80780.0	ITIH5	Inter-alpha (Globulin) inhibitor H5	141	2	4.3	0
IP100298971	7448.0	VTN	Vitronectin precursor	158	21	55.1	11
IP100296922	3913.0	LAMB2	Laminin subunit beta-2 precursor	177	40	24.6	0
IP100297646	1277.0	COL1A1	Collagen alpha-1(I) chain precursor	152	17	5.8	0
IP100021263	7534.0	YWHAZ	14-3-3 protein zeta/delta	173	52	344.9	160
IP100000760	23564.0	DDAH2	NG,NG-dimethylarginine dimethylaminohydrolase 2	151	33	23.2	3
IP100026844	4811.0	NID1	Isoform 1 of Nidogen-1 precursor	140	23	37.7	0
IP100554737	5518.0	PPP2R1A	Serine/threonine-protein phosphatase 2A 85 kDa regulatory subunit A alpha isoform	145	31	26.1	0
IP100007119	5054.0	SERPINE1	Plasminogen activator inhibitor 1 precursor	115	5	30.4	0
IP100217296	5524.0	PPP2R4	Isoform 3 of Serine/threonine-protein phosphatase 2A regulatory subunit B'	121	14	7.2	3
IP100021700	5111.0	PCNA	Proliferating cell nuclear antigen	139	33	24.6	0
IP100034319	51586.0	CUTA	Isoform A of Protein CutA precursor	154	51	36.2	16
IP100306322	1284.0	COL4A2	Collagen alpha-2(IV) chain precursor	126	25	5.8	0
IP100003966	768.0	CA9	Carbonic anhydrase 9 precursor	100	1	10.1	0

Table 2. Continued

ID	GeneID	Symbol	Description	Trans-criptomics HCC (tpm)	Trans-criptomics normal (tpm)	Proteomics Spectral counting (HCC) ^a	Proteomics Spectral counting (normal)
IPI00215911	328.0	APEX1	DNA-(apurinic or apyrimidinic site) lyase	138	42	42.0	11
IPI00399319	11316.0	COPE	Epsilon subunit of coatomer protein complex isoform c	115	20	5.8	0
IPI00217465	3006.0	HIST1H1C	Histone H1.2	133	42	92.7	0
IPI00010133	11151.0	CORO1A	Coronin-1A	109	17	10.1	0
IPI00396321	55379.0	LRRC59	Leucine-rich repeat-containing protein 59	123	35	33.3	13
IPI00028946	10313.0	RTN3	Isoform 3 of Reticulon-3	101	14	2.9	0
IPI00168479	128240.0	APOA1BP	Apolipoprotein A-1 binding protein precursor	112	26	84.1	31
IPI00743696	1282.0	COL4A1	161 kDa protein	103	20	11.6	0
IPI00418262	230.0	ALDOC	Fructose-bisphosphate aldolase C	96	14	30.4	14
IPI00297160	960.0	CD44	Isoform 12 of CD44 antigen precursor	108	26	5.8	0
IPI00032957	7329.0	UBE2I	SUMO-conjugating enzyme UBC9	110	32	11.6	0
IPI00443909	10330.0	CNPY2	Isoform 1 of MIR-interacting saposin-like protein precursor	115	38	21.7	5
IPI00059366	9555.0	H2AFY	H2A histone family, member Y isoform 2	100	23	87.0	0
IPI00018931	55737.0	VPS35	Vacuolar protein sorting-associated protein 35	104	27	8.7	0
IPI00293464	1642.0	DDB1	DNA damage-binding protein 1	112	37	21.7	7
IPI00479997	3925.0	STMN1	Stathmin	91	17	33.3	0
IPI00374563	375790.0	AGRN	Agrin precursor	81	10	5.8	0
IPI00290770	7203.0	CCT3	Chaperonin containing TCP1, subunit 3 isoform b	92	21	15.9	0
IPI00438229	10155.0	TRIM28	Isoform 1 of Transcription intermediary factor 1-beta	94	24	14.5	0
IPI00032325	1475.0	CSTA	Cystatin-A	81	11	2.9	0
IPI00029468	10121.0	ACTR1A	Alpha-centractin	102	32	4.3	0
IPI00216153	6209.0	RPS15	40S ribosomal protein S15	83	18	5.8	0
IPI00026833	159.0	ADSS	Adenylosuccinate synthetase isozyme 2	76	13	2.9	0
IPI00014361	7264.0	TSTA3	GDP-L-fucose synthetase	81	20	15.9	7
IPI00018769	7058.0	THBS2	Thrombospondin-2 precursor	75	14	113.0	0
IPI00012535	3301.0	DNAJA1	DnaJ homolog subfamily A member 1	80	18	26.1	0
IPI00012382	6626.0	SNRPA	U1 small nuclear ribonucleoprotein A	70	16	18.8	8
IPI00005198	3608.0	ILF2	Interleukin enhancer-binding factor 2	67	14	95.6	38
IPI00016862	2936.0	GSR	Isoform Mitochondrial of Glutathione reductase, mitochondrial precursor	77	25	63.8	27
IPI00018768	7247.0	TSN	Translin	78	26	46.4	20
IPI00013894	10963.0	STIP1	Stress-induced-phosphoprotein 1	66	16	133.3	7
IPI00105407	57016.0	AKR1B10	Aldo-keto reductase family 1 member B10	61	12	466.6	30
IPI00455033	10801.0	39334.0	Isoform 3 of Septin-9	58	9	4.3	0
IPI00018236	2760.0	GM2A	Ganglioside GM2 activator precursor	65	15	2.9	0
IPI00006052	5202.0	PFN2	Prefoldin subunit 2	67	21	18.8	0
IPI00014589	1212.0	CLTB	Isoform Brain of Clathrin light chain B	65	21	14.5	3
IPI00306825	7165.0	TPD52L2	Isoform 1 of Tumor protein D54	55	12	8.7	0
IPI00007138	22839.0	DLGAP4	Isoform 1 of Disks large-associated protein 4	50	8	4.3	0
IPI00032516	8907.0	AP1M1	AP-1 complex subunit mu-1	54	12	4.3	0
IPI00030968	286257.0	C9orf142	Uncharacterized protein C9orf142	44	5	2.9	0
IPI00478390	1029.0	CDKN2A	Cyclin-dependent kinase inhibitor 2A, isoform 4	40	1	2.9	0
IPI00001560	1029.0	CDKN2A	Isoform 1 of Cyclin-dependent kinase inhibitor 2A, isoforms 1/2/3	40	1	2.9	0
IPI00014537	813.0	CALU	Isoform 1 of Calumenin precursor	53	15	11.6	0
IPI00456750	64855.0	FAM129B	Niban-like protein	45	8	5.8	0
IPI00010154	2664.0	GDI1	Rab GDP dissociation inhibitor alpha	51	17	4.3	2
IPI00033025	989.0	39332.0	Isoform 1 of Septin-7	46	13	27.5	4
IPI00032460	57819.0	LSM2	U6 snRNA-associated Sm-like protein LSM2	40	7	8.7	3
IPI00411426	9559.0	VPS26A	Vacuolar protein sorting-associated protein 26A	47	15	4.3	0

Table 2. Continued

ID	GeneID	Symbol	Description	Trans-criptomics HCC (tpm)	Trans-criptomics normal (tpm)	Proteomics Spectral counting (HCC) ^a	Proteomics Spectral counting (normal)
IPI00294158	644.0	BLVRA	Biliverdin reductase A precursor	45	12	2.9	0
IPI00022078	10397.0	NDRG1	Protein NDRG1	45	14	7.2	0
IPI00020672	10072.0	DPP3	Isoform 1 of Dipeptidyl-peptidase 3	34	4	20.3	8
IPI00013454	6001.0	RGS10	Isoform 1 of Regulator of G-protein signaling 10	40	10	2.9	0

a) Normalized by total spectral numbers.

这可能是由于一些蛋白质具有多个细胞定位。此外，它很可能是在组织培养

过程中，一些细胞死亡，从而导致释放细胞内蛋白质进入培养基中。此现象已在以前的细胞培养基蛋白质组学分析中被提出过[23, 24]。我们试图通过限制组织培养的时间长短来尽量减少这种效应。然而，我们发现，约 42%经鉴定的蛋白质被预测为分泌蛋白，高于从细胞裂解物中蛋白质组数据来预测的分泌蛋白数以及远高于从随机选择的基因列表中来预测的分泌蛋白数（数据未显示）。

我们采用了一种简单的频谱计算方法[10]以确定的相对丰度，并成功地确定了在肝癌分泌组相比正常分泌组至少五倍过度或低表达的大约 190 个差异表达的蛋白（参考信息表 3 和表 4）。质谱计数被证明是定量蛋白质组学分析的有效方法 [25]。正如我们分析的 12 对样品，质谱计数方法显然更容易实现比较与稳定同位素标记和定量分析。在质谱计数中经过比较得到计数值越高，则比较的准确性越好。质谱数目在单独癌与正常对照中通常很小（小于 10）除了那些高丰度蛋白，它们之间通过比较质谱计数方法是不可靠的。为了提高我们差异检测的敏感性和可信度，我们首先总结了肝癌组与对照组的总质谱数，然后进行比较。

许多在肝癌分泌组里我们鉴定为过度表达的蛋白质在此前已报道为肝癌生物标志物[4, 26-29]。我们发现许多热休克蛋白家族成员呈过表达（表 2）包括热休克 70- kDa 蛋白（HSPA5）（质谱计数：1023/102; 肝癌/正常），热休克 70- kDa 蛋白 4（HSPA4）（质谱计数：29 /6, 肝癌/正常），热休克 71- kDa 蛋白亚型 1（质谱计数：848/445）和热休克蛋白 HSP 90- α -2（质谱计数：355 /161）。Sun 等人[26]通过 2D- DIGE 分析肝癌组织并表明几个 HSP 在肝癌组织中表达上调。我们发现，载脂蛋白 E（APOE）过度表达在肝癌组中（359）相比正常组（2）。它已经由 2-DE 和 Western-blot 方法证实在肝癌中上调[29]。我们还确定了其他先前报道的肝癌相关蛋白质，如肝癌衍生生长因子（ HDGF ） [30, 31]，谷胱甘肽 S 转移酶（GST） [32~34]，以及醛缩酶 A [35, 36]。

一些我们确定的分泌组蛋白被证明是良好的肝癌血清生物标志物。我们发现钙网蛋白和蛋白质二硫键异构酶 A3（PDIA3）的在肝癌分泌组相比对照组中过表达（质谱计数：钙网蛋白在肝癌中 758, 223 在对照中；对 PDIA3 在肝癌中 699 在对照中 227）。Chignard 等 [37]显示在钙网蛋白和 PDIA3 片段在肝癌患者和健康人的血清水平之间存在有统计性意义的显著差异。有趣的是，我们发现，除了 PDIA3 外，PDIA2（在肝癌中 115, 在对照中 0）， PDIA4（在肝癌中 383, 在对照中 113），PDIA6（在肝癌中 226, 在对照中 76）均在肝癌中相比于对照显著高表达（参考信息表 1 和表 2）。

从我们的分泌组的分析结果，我们确定 CHI3L1（几丁质酶 3 类似蛋白 1，

又名 YKL-40)是一个很好的诊断肝癌标记物, AUC 为(ROC 曲线下面积)0.92。(图 3A)。CHI3L1 是几丁质酶蛋白家族的成员[15]。CHI3L1 的功能是未知的。有可能在组织重塑中起作用[16]。此外, YKL-40 被证明是一种预测结直肠癌及乳腺癌生存时间 [38, 39]的预后标记物, 以及卵巢癌和子宫内膜癌的诊断标志物[18, 19]。

Johansen 等人[40]的研究表明, 血清 YKL-40 (CHI3L1) 随着患者的肝纤维化的出现而增高, 包括酒精肝硬化, 后肝炎肝硬化和非肝硬化纤维化。最近, YKL-40 的表达水平显示出与丙肝相关的纤维化[41, 42]有关。由于我们的病人群体有严重纤维化(纤维化阶段 4, 表 1 和支持信息表 5), 我们病人群体里增加的 YKL-40 水平可能与肝癌的肝纤维化或乙肝感染相关。肝纤维化可能由病毒感染引起(例如乙肝病毒或丙肝), 化学品或肿瘤细胞的生长[43]。进一步的研究比较纤维化但没有癌症的患者, 那些纤维化同时有肝癌的患者, 那些有肝癌但没有纤维化的患者将是必要的, 以确定 YKL-40 在肝癌诊断中的作用。

为了提高我们的鉴别能力, YKL-40 的增加是由于肝癌而不是其他癌症, 我们增加了肝富集表达基因 MASP2。结合这两个标记将增加肝癌诊断的 AUC。YKL-40/MASP2 比例的 AUC 对于肝癌的诊断是 0.97, 它是大于单独 MASP2 (0.83) 的 AUC 和单独 YKL-40 (0.92) 的。然而, 如我们比较肝癌与正常对照, 而不是肝癌患者与那些有肝纤维化但没有癌症的患者, AUC 的肝癌诊断可能被高估。另外, 在本研究的样本量不大, 进一步评估使用别的病人群体和更大的样本量将是必要的, 以确定该生物标志物对的真正的实用价值。此外, 后续研究肝癌患者与患相同的类型和严重程度肝病的疾病对照(如纤维化)来确定是否 YKL40 和 MASP2 标记物的组合对可以帮助从其他肝脏疾病中来区别鉴别肝癌。

综上所述, 我们提出了一种合并生物标志物的新范式, 包括 (i) 一种癌症的生物标志物是相比正常个体是增加的, 及 (ii) 一个器官富集表达的标记, 在癌症由于器官功能丧失而向下调节可能帮助识别癌症的组织/器官的起源, 哪怕第一个癌生物标志物不是癌症类型特异性的。我们已经采用这种方法来确定肝癌的生物标志物对 CHI3L1 和 MASP2。我们发现的其他的候选生物标志物(支持信息表 3 和表 4)可以利用稳定同位素标记进行标记, 并使用质谱进一步进行定量评估如果 ELISA 检测不可用或开发起来过于昂贵。我们的方法可能是发现癌症生物标志物的有用的一般方法, 最终可能有助于利用多种标志物来评估器官的正常功能或疾病状态。

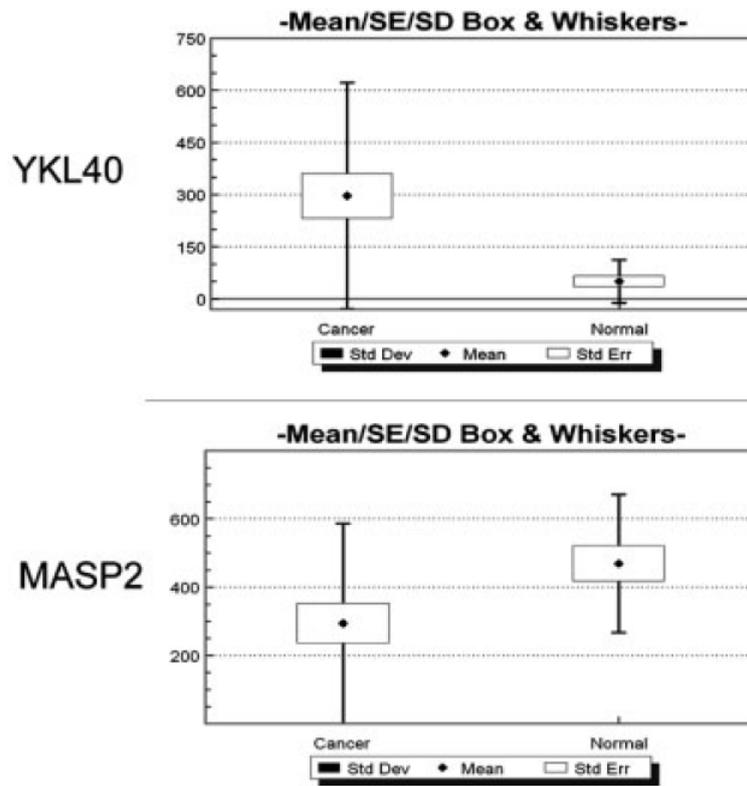


图 2. 盒须图显示平均值, 标准偏差 (标准偏差) 和标准错误 (标准错误) CHI3L1 和 MASP2 在 25 个肝癌患者的血清样品和 15 个年龄对应的正常对照。

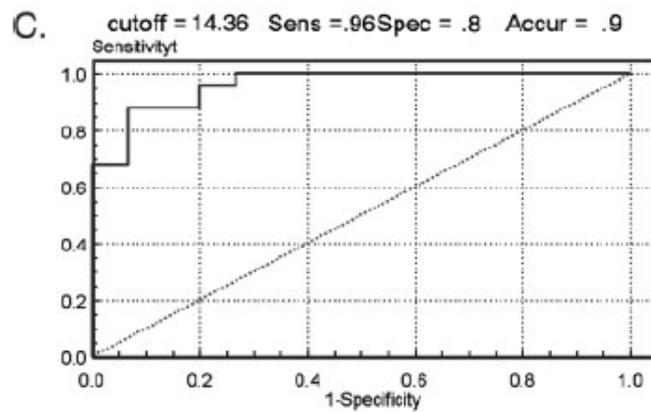
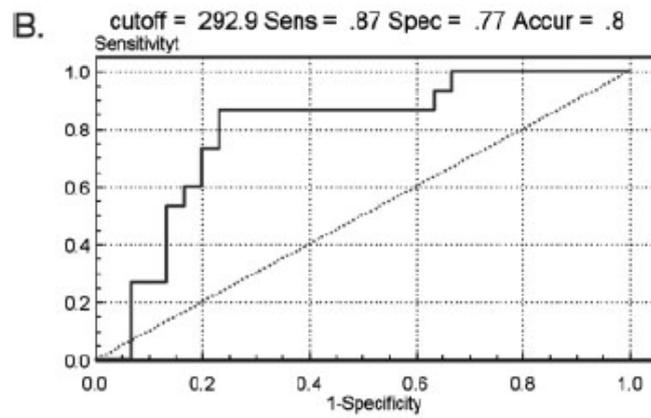
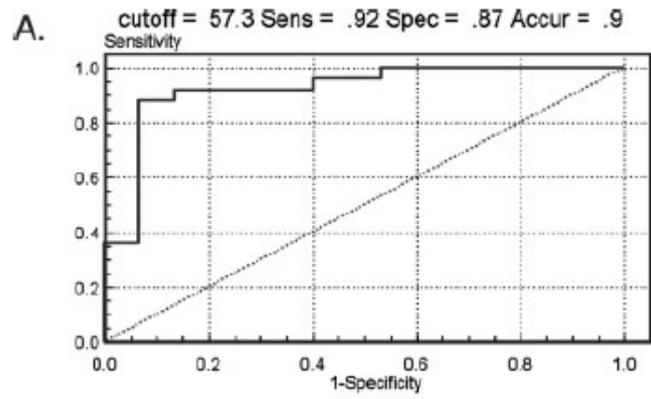


图 3. CHI3L1 (A) 和 MASP2 (B) 以及 CHI3L1/MASP2 (C) 的 ROC 曲线，表明它们在检测肝癌的 AUC，灵敏度以及特异性。