СРАВНЕНИЕ РЕЗУЛЬТАТОВ ПРОТЕОМНОГО И ТРАНСКРИПТОМНОГО АНАЛИЗА И ОПРЕДЕЛЕНИЕ КОРРЕЛЯЦИИ МЕЖДУ НИМИ

Рыбина А.В., Андреевский Т. , Згода В.Г., Курбатов Л.К.

ГУ НИИ Биомедицинской химии им.В.Н.Ореховича РАМН, г.Москва

Haranga@mail.ru

В данной работе были сравнены результаты протеомного и транскриптомного анализа для клеточной линии HepG2, и была выявлена некоторая корреляция, между ними, однако, не было найдено прямой зависимости для каждой отдельной пары ген-белок

В настоящее время данные протеомных и транскриптомных исследований активно используются для создания моделей функционального состояния эукариотических и прокариотических клеток (Perco, 2010… Keasling u.a). Однако, вопрос о разрешении противоречий между экспериментальными данными по экспрессии генов и присутствию белков в клетках остаётся открытым. Эти противоречия имеют в своей основе как биологические, так и "технологические" причины. К биологическим можно отнести

процессы, приводящие к деградации или модификации белков или молекул мРНК, в результате чего теряется прямая связь между наличием или отсутствием генного продукта и соответствующим белком. "Технологические" причины заключаются, как правило, в степени достоверности идентификации белков при протеомном анализе и степени специфичности зондов для мРНК в случае транскриптомного анализа на микрочипах. Также важна и разница в чувствительности методов детекции мРНК и белков в клетке. Наиболее чувствительные методы протеомного анализа способны регистрировать и идентифицировать десятки и сотни тысяч молекул. Благодаря реакции амплификации, транскриптомные методы способны регистрировать 1 молекулу мРНК на клетку исследуемого образца (McClintick и Edenberg, 2006) . Кроме того, если транскриптомный анализ является полуколичественным и позволяет определить относительную разницу в экспрессии по интенсивности флуоресценции меченой кРНК, то наиболее распространенные методы протеомных исследований являются только качественными.

Для оценки качественного и иногда количественного содержания белков и мРНК используются различные методы. Данные по массовому анализу мРНК получают, как правило, с помощью микрочипового метода. Белковый анализ чаще всего выполняется с применением двумерного электрофореза (2D) с последующей идентификацией белков масс-спектрометрическими методами. В работах, предполагающих одновременное использование транскриптомных и протеомных данных показано, что уровень соответствия этих данных сильно зависит от исследуемой системы и может в некоторых случаях превышать 80% (Mijalski и др., 2005). Другие же источники сообщают об очень низком уровне соответствия. (Gygi и др., 1999), (Chen и др., 2002).

В данной работе представлены результаты анализа экспрессии мРНК и протеома клеточной линии HepG2. Протеомное исследование осуществлялось с помощью трехмерной жидкостной хроматографии (3D – LC MS/MS), а транскрипционный анализ проводился на полногеномных микрочипах фирмы Agilent.

МАТЕРИАЛЫ И МЕТОДЫ

Клеточная линия

Клеточная линия Hep G2 была получена от компании ATCC (кат. номер HB-8065). Клетки выращивались при стандартных условиях в CO^2 инкубаторе с использованием культуральных флаконов фирмы Corning (№430639). Среда роста RPMI-1640 (Sigma R8758-500ML) содержала 10% эмбриональной бычьей сыворотки (Sigma F6178-500ML). Пересев клеток осуществляли при достижении клетками конфлуентного монослоя с

разведением 1:5. Отбор клеток для анализа осуществляли при достижении 50% монослоя. Клетки дважды споласкивали раствором Хенкса и обрабатывали раствором Версена при 37ºС, визуально следя за откреплением клеток. Затем клетки дважды промывали центрофугированием в растворе Хенкса и окончательно в фосфатном буфере. Осадок клеток замораживали в жидком азоте и до использования хранили при -80ºС.

Выделение РНК и транскриптомный анализ

Подготовка образцов для транскриптомного анализа осуществлялась согласно протоколу фирмы производителя Agilent. Использовался вариант двухцветной гибридизации на полногеномные чипы G4112A фирмы Agilent (Human Whole Genome Oligo Microarray).

Обработка транскриптомных данных

После сканирования микрочипов данные экстрагировались при помощи программы Feature Extraction 9.1 фирмы Agilent Technologies. Эта программа осуществляет идентификацию генных продуктов на чипе, вычисление фонового сигнала, а также расчёт интенсивности флуоресценции по красному и зелёному каналу для каждой пробы (точки флуоресценции) и нормализацию сигнала с учетом особенностей двухцветной гибридизации - lowest weighted linear regression – LOWESS (Workman и др., 2002). Дальнейший анализ, нормализация, фильтрация и визуализация данных осуществлялась с помощью программы GeneSpring GX 7.3.1.

Анализ белков

Трехмерную жидкостную хроматографию (3D-LC) проводили, используя устойчивую к нагреванию mRP-колонку как первый шаг разделения. Обращенно-фазовое разделение белков было выполнено на системе LC 1100 Agilent, используя mRP-C18 колонку (4.6 мм × 50 мм, Agilent). Разделения проводили с помощью мультисегментированного линейного градиента при нагревании до 80°C.

Обращенно-фазовую жидкостную хроматографию и тандемную масс-спектрометрию осуществляли на приборе Agilent 1100 nanoflow HPLC-Chip Cube в сочетании с Agilent 1100 XCT Ultra Series MSD ионной ловушкой (Agilent, США). Пептиды разделяли на колонке HPLC Chip (40 нл колонка для обогащения, 75 мкм × 43 мм аналитическая колонка, 5 мкм C-18 SB-ZX, Agilent) с использованием линейного градиента ацетонитрила 5–80% в 0,1% муравьиной кислоте в течение 60 минут при скорости потока 300 нл/мин. Ионы детектировали в ионной ловушке в диапазоне 300–1800 m/z.

Обработку масс-спектров и идентификацию белков проводили с помощью программы Spectrum Mill MS Proteomics Workbench Rev A.03.03.078 (Agilent). Идентификацию белков проводили с использованием базы данных Swiss Prot.

РЕЗУЛЬТАТЫ

Результаты транскриптомного анализа

Анализ экспрессии генов проводился в трёх технических повторах в варианте self-self гибридизации. В результате нормализованная флуоресценция по красному и зеленому каналу в каждой точке флуоресценции, включающей сотни олигонуклеотидных зондов одного типа, должна была иметь равные величины. Таким образом, по интенсивности сигнала можно судить о концентрации данной мРНК в клетке и, соответственно, уровне экспрессии гена. Дальнейшая статистическая обработка и визуализация данных осуществлялась при помощи программного обеспечения GeneSpring GX 7.3.1. После фильтрация данных по «флагам» - показателям качества гибридизации для каждой отдельно взятой точки флуоресценции, получили список из 26007 идентификаторов. Следует отметить, что транскриптомные данные имеют некоторую «избыточность», поскольку включают в себя многочисленные идентификаторы, соответствующие

временным экспериментальным названиям фрагментов неизвестных генов (www.genenames.org).

Результаты протеомного анализа

В данной работе был выбран метод, состоящий из разделения белков методом хроматографии с использованием термостабильного носителя на основе обращенной фазы, с последующим гидролизом полученных белковых фракций трипсином и, наконец, анализом гидролизатов фракций белков методом многомерной технологии идентификации белков MudPit (Multidimensional Protein Identification Technology). Протеомный анализ с использованием данного подхода проводили в трех независимых экспериментах. Идентификацию белков по тандемным масс-спектрам проводили с использованием програмного обеспечения SpectrumMill, которое позволяет проводить идентификацию с различными значениями параметра score, отражающего вероятность правильности идентификации. Для дальнейшего проведения сравнения результатов протеомики и транскриптомики , протеомные данные были сформированы в две группы: S(strong) – максимальная степень достоверности (белки, имеющие score > 7 с последующей автоматической валидацией результатов) и W(weak), включающий в себя все идентифицированные белки. Данные группы включали в себя 1441 и 7709 записей соответственно.

Сравнение результатов протеомного и транскриптомного анализа

Соответствие белковых и транскриптомных идентификаторов оценивалось с помощью биоиформационного ресурса BioMart (http://www.biomart.org/). Поиск осуществлялся при этом по базе данных Ensembl (http://www.ensembl.org/index.html), объединяющей идентификаторы на уровне гена, РНК и конечного продукта в характеристике каждого белка. Сопоставление белковых и генных идентификаторов позволило найти в списке из 7710 белков 7305 транскриптов, имеющих минимально значимый уровень экспрессии.

Само сравнение полученных списков проводилось с помощью программы Microsoft Access 2010 и Microsoft Exel.

Выявление зависимости между транскриптомным и протеомным анализом

Сравнение проводилось по уровням score и флуоресцентного сигнала(усреднённого по 3-м экспериментам). Не было обнаружено прямой зависимости score от уровня флуоресцентного сигнала, но было обнаружено, что при сравнении списков генов и белков процент совпадения повышается при одновременном увеличении минимального сигнала флуоресценции и величины score.

Для выявления корреляции, списки флуоресцентных сигналов и score были выстроены в порядке убывания флуоресцентного сигнала и белковый список W, после чего было взято по 100 верхних записей из каждого списка и сравнены друг с другом. После этого было сделано сравнение 200, 300 и так далее с шагом 100 до конца обоих списков. В результате получилась линейная зависимость между взятым и совпавшим количеством записей (Рис1). Также для 406 белков не было найдено транскриптомных сигналов, однако, поскольку при запросе в BioMart для них не было найдено идентификаторов Agilent, что можно объяснить “технологическими” ошибками. (Несоответствием баз Agilent и Uniprot или отсутствием соответствующих данных в Ensemble.)

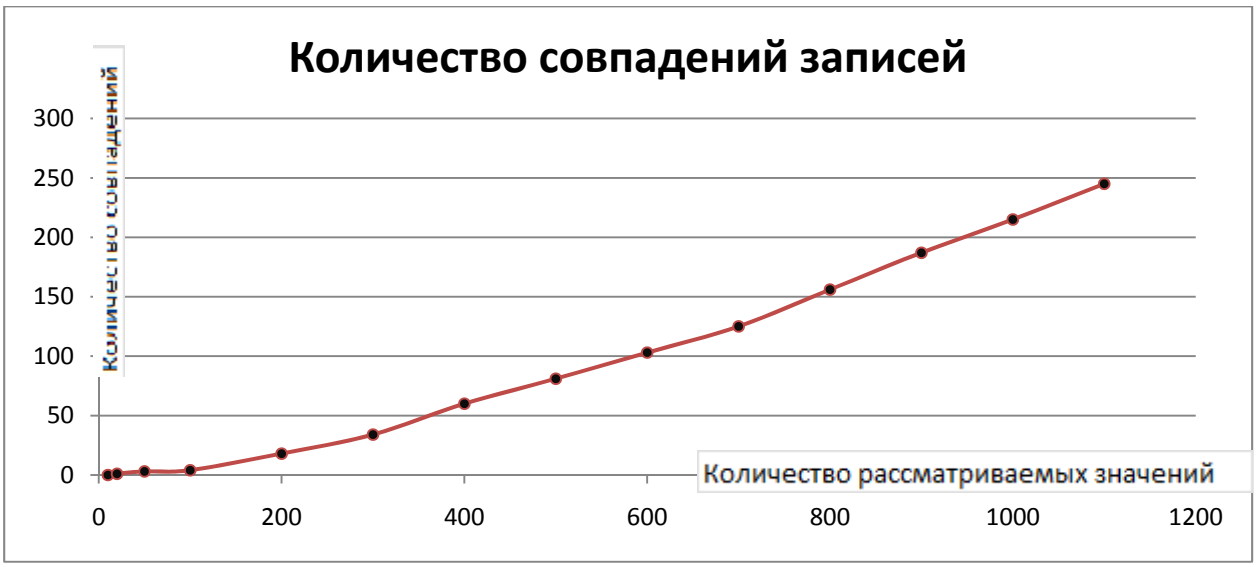

Рис 1.

Записи, присутствующие и в протеомном и в транскриптомном массивах, были выделены в 2 отдельных списка и выстроены в порядке уменьшения score и флуоресцентного сигнала соответственно. Затем для определения есть ли тенденция связи между высоким score и высоким сигналом и отсечения влияния на результат данных с низким score и сигналом, оба списка были разделены на 4 части первый – по уровню сигнала, второй – по score. После этого четверти первого списка были сравнены с соответствующими (часть с максимальным score с частью с максимальным сигналом и т.д.) четвертями второго списка. Аналогичные действия были проделаны для списка S (с параметрами, соответствующим очень надёжной идентификации). В обоих случаях была выявлена закономерность – чем выше score и сигнал, тем больше корреляция между ними, причём для списка со score, соответствующим очень надёжной идентификации, эта зависимость проявилась гораздо заметнее(см Рис2 и Рис3). Для списка W процент совпадения “верхней” четверти составил 48,69%, в то время как для списка S , этот процент составил 76,1%.

Рис2. Сравнение четвертей списков W и ранжированного флуоресцентного сигнала.

Рис3. Сравнение четвертей списков S и ранжированного флуоресцентного сигнала.

ОБСУЖДЕНИЕ

В настоящее время данные протеомных и транскриптомных исследований активно используются для создания моделей функционального состояния эукариотических и прокариотических клеток.

Как правило, комплексный протеомный и транскриптомный анализ проводится при исследовании ответных реакций клеток на различные воздействия или индуцировании дифференцировки (Zheng с сотр., 2005; Conrads с сотр., 2005). При этом даже при использовании довольно ограниченного по возможностям 2D-электрофореза идентифицируются белки с неподтвержденной экспрессией соответствующей мРНК (Zheng с сотр., 2005). В связи с этим делается вывод, что данные протеомики и транскриптомики являются дополнением друг для друга и позволяют дать более полное представление о функциональных преобразованиях в клетках.

При анализе мРНК и белков клеточной линии HepG2 видно, что не существует абсолютной зависимости между уровнем флуоресцентного сигнала генного продукта на микрочипе и score, однако было обнаружено, что при сравнении списков генов и белков процент совпадения повышается при одновременном увеличении минимального сигнала флуоресценции и величины score. Эта корреляция связана с тем, что данные транскриптомики представляют собой сигнал, зависящий от количества мРНК, а белковые данные - достоверность, которая в какой-то степени зависит от количества белка в пробе.

Деление ранжированных списков протеомных и транскриптомных данных на четверти было проведено для уменьшения влияния низких показателей. Было показано, что для списка W процент совпадения “верхней” четверти составил 48,69%, в то время как для списка S , этот процент составил 76,1%, и чем ближе четверть к минимальным показателям, тем этот процент ниже (см Рис2 и Рис3). В то же время прямое сопоставление флуоресцентного сигнала и score массивов даже “верхних” четвертей не показывает линейной зависимости (см Рис4). Таким образом, высокий уровень флуоресцентного сигнала на чипе, отражающий количество мРНК определённого гена, может не совпадать с высоким score соответствующего белка и наоборот.

Тем не менее возможно, что существует зависимость сигнала от score, позволяющая определить нижние пределы достоверности. Под таким пределами подразумевается уровень флуоресцентного сигнала в транскриптомном анализе, при котором велика вероятность обнаружения соответствующего белка высокопроизводительным протеомным методом.

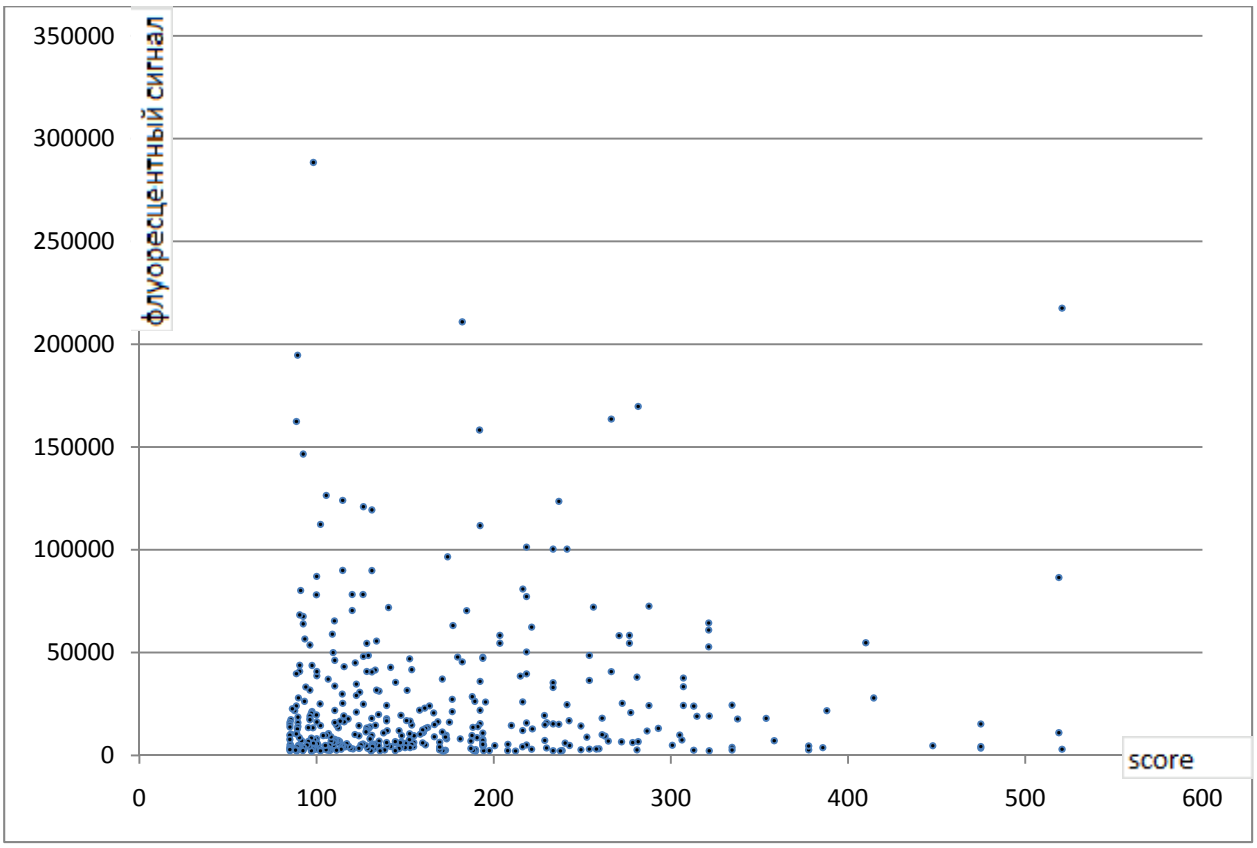

Рис 4. Корреляция между score и флуоресцентным сигналом для "верхних" четвертей списков S и ранжированного флуоресцентного сигнала.

ЗАКЛЮЧЕНИЕ

Поскольку наше исследование показало, что увеличение сигнала для транскриптомики и критериев достоверности для протеомики значительно повышает процент совпадения массивов данных, можно сделать вывод, что есть некоторая корреляция между этими показателями. Соответственно, следующим шагом данного исследования будет попытка определения минимально достоверных показателей (Trash Hold) score и флуоресцентного сигнала , отражающих реальное наличие белка и соответствующей мРНК в клетке.

Литература

1. *Zheng, P. Z., Wang, K. K., Zhang, Q. Y., Huang, Q. H., et al.*, // Systems analysis of transcriptome and proteome in retinoic acid/arsenic trioxide-induced cell differentiation apoptosis of promyelocytic - leukemia. Proc. Natl. Acad. Sci. USA 2005, 102, 7653–7658.
2. *Mijalski, T., Harder, A., Halder, T., Kersten, M., et al.*, // Identification of coexpressed gene clusters in a comparative analysis of transcriptome and proteome in mouse tissues. - Proc. Natl. Acad. Sci. USA 2005, 102, 8621–8626.
3. *Steven P. Gygi, Beate Rist, Scott A. Gerber, Frantisek Turecek, Michael H. Gelb, and Ruedi Aebersold.* // Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. - Nature Biotechnology 1999, 17, 994-1000
4. *Gygi SP, Rochon Y, Franza BR, Aebersold R.* // Correlation between protein and mRNA abundance in yeast. - Mol Cell Biol 19: 1720–1730, 1999.
5. *Zheng PZ, Wang KK, Zhang QY, Huang QH, Du YZ, Zhang QH, Xiao DK, Shen SH, Imbeaud S, Eveno E, Zhao CJ, Chen YL, Fan HY, Waxman S, Auffray C, Jin G, Chen*

SJ, Chen Z, Zhang J. // Systems analysis of transcriptome and proteome in retinoic acid/arsenic trioxide-induced cell differentiation/apoptosis of promyelocytic - leukemia. Proc Natl Acad Sci U S A. 2005 May 24;102(21):7653-8. Epub 2005 May 13.

6. *Kelly A. Conrads1†, Ming Yi2†, Kerri A. Simpson1†, David A. Lucas3, Corinne E. Camalier1, Li-Rong Yu3, Timothy D. Veenstra3, Robert M. Stephens2, Thomas P. Conrads3 and George R. Beck Jr.* // A combined proteome and microarray investigation of inorganic phosphate-induced pre-osteoblast cells. - MCP Papers in Press. Published on June 14, 2005 as Manuscript M500082-MCP200
7. *Paul Perco, Irmgard Mu¨ hlberger, Gert Mayer, Rainer Oberbauer, Arno Lukas, Bernd Mayer* // Linking transcriptomic and proteomic data on the level of protein interaction networks. - Electrophoresis 2010, 31, 1780–1789
8. *McClintick JN, Edenberg HJ:*// Effects of filtering by Present call on analysis of microarray experiments.
9. *BMC Bioinformatics* 2006, **7:**49. PubMed Abstract | BioMed Central Full Text | PubMed Central Full Text