Journal of Proteomics & Bioinformatics - Open Access

Editorial JPB/Vol.1/August 2008

Whether Transcriptome and Proteome Technical Platforms are Intact to Human Genome now?

Fan Zhong[1], Fuchu He[1,2]*

[1]Institutes of Biomedical Sciences, Department of Chemistry, Fudan University, 130 Dong'an Road, Shanghai, 200032, China

[2]State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing Institute of Radiation Medicine, 33 Life Science Park Road, Beijing, 102206, China

*Corresponding author: Fuchu He, E-mail: hefc@nic.bmi.ac.cn

Received July 01, 2008; Accepted July 21, 2008; Published August 13, 2008

Citation: Fan Z, Fuchu H (2008) Whether Transcriptome and Proteome Technical Platforms are Intact to Human Genome now?. J Proteomics Bioinform 1: 237-241. doi:10.4172/jpb.1000030

Transcriptomic and proteomic technologies are vogues in analyzing biological entities. The integrality of their probe sets or searching databases is the prerequisite of full identification, which could be estimated by their coverages over genome.

After the completion of human genome atlas with painstaking effort, expression profiling technology becomes routine analysis in both transcriptome and proteome levels. The mainstream strategies of such profiling, based on expression microarrays and mass spectrum, are enclosed in probe set and sequence database respectively. It is conceivable that their integralities, i.e. their coverages to the whole human genome, determine their capacity in cataloguing transcriptome and proteome.

For the estimate of their integralities, human genome repository in NCBI known as Entrez Gene was used as genome background. Totally 36,545 items including 25,611 **P**rotein **C**oding **G**enes (**PCGs**), and 1,114 **R**NA **C**oding **G**enes (**RCGs**) beyond **PCG** were compiled in version 2007-05-21.

In **M**icroarray-**b**ased **T**ranscriptome **P**rofiling (**MbTP**), two most widely used human "whole genome" expression microarray, Affymetrix **HG-U133 plus 2.0** and Agilent **Human Genome Whole-Four-Plex 44K** were chosen as representatives. The **HG-U133 plus 2.0** contains 54,675 probe sets which can be mapped to 21,083 genes in Entrez Gene, and the 41,000 probe sets in the **Whole-Four-Plex** covers 19,610 genes. For **PCGs**, both of them have close but only medium coverage (76.23% in the **HG-U133 plus 2.0**, and 71.31% in the **Whole-Four-Plex**). Their specific parts include 1,899 and 638 **PCGs** respectively. Their combination can just slightly promote the **PCG** coverage up to 78.72%. The 5,450 **PCGs** lost by **MbTP** could be categorized as "hypothetical" (2,291 genes) or "similar to"/"-like" (2,012 genes), and "others" including 234 ORF-related, 75 zinc finger proteins, 36 immunity products and 802 other functional-known genes. It is notable that 233 members of olfactory receptor gene family and 49 keratin coding genes were lost by **MbTP**. Obviously, current **MbTPs** could not fully identify the proteome-encoding transcriptome. On the other hand, for **RCGs**, the **MbTP** covers only less than 15% coding genes of miscellaneous RNA (**miscRNA**), small nucleolar RNA (**snoRNA**), and small nuclear RNA (**snRNA**); even worse, none of the two microarrays have any ability to detect rRNA and tRNA (see Table 1). Evidently, those non-protein-coding RNAs should be explored only by virtue of other specialized microarrays.

The latest Affymetrix "whole genome" expression microarray **Human Gene 1.0 ST** covers 19,915 genes with 18,570 **PCGs**. Those numbers are even less than those of the **HG-U133 plus 2.0**. The replacement of the **Human Gene 1.0 ST** to the **HG-U133 plus 2.0** will result in the loss of other 2,920 genes including 2,038 **PCGs**. So we finally choose the later as the "whole genome" representative from Affymetrix series. Honestly, addition of the **Human Gene 1.0 ST** to the **MbTP** can slightly raise the total gene coverage (61.15%→64.60%) and **PCG** coverage (78.72%→81.48%). Of the additional 708 **PCGs,** 533 have known functions. Moreover, all the lost 233 olfactory receptor genes and partial keratin genes (31 of 49 lost) have been retrieved in the **Human Gene 1.0 ST**. Even so, the

HG-U133 plus 2.0 is still with more but not too much preponderance coverages of all the genes and **PCGs**.

Probe set design should keep up with genome databases, but is fettered by period of products and somewhat lag behind genome updating. As a remission, some extensional probe sets which corresponding to low quality annotated sequences in current genome database can be designed in microarray in advance. Besides of the "enclosed" transcriptome detection technique, **S**erial **A**nalysis of **G**ene **E**xpression (**SAGE**) and further developed **M**assively **P**arallel **S**ignature **S**equencing (**MPSS**) are two major complementary techniques to transcriptome microarray. They can detect transcripts from unknown (novel) genes.

In Mass **S**pectrum-**b**ased **P**roteome **P**rofiling (**MSbPP**), protein sequence databases can follow more tightly with updating of genomic sequences. The **I**nternational **P**rotein **I**ndex (**IPI**) in EBI and the "**N**on-**R**edundant" protein database (**NR**) in NCBI are two dominant databases for mass spectrum data searching. **IPI** updates monthly and **NR** updates in real time through web interface. **IPI** Human v.3.32 released in 2007-08-06 contains 67,524 items covers 21,768 **PCGs**; **NR-Human** from FTP download (release 2007-08-08) and web download (release 2007-08-10) contains 175,947 and 410,546 items respectively, totally covering 24,155 **PCGs**. The both proteomic databases reach significantly higher **PCG** coverage (**IPI** 84.99%, **NR** 94.31%), especially the later. **IPI**-mapped **PCGs** are almost included in those of **NR**, only with 185 specifically mapped **PCGs** (see Table 1). We can construct an **MSbPP** searching database by combining **IPI** and **NR**. It will achieve a little bit higher **PCG** coverage at 95.04%. Even so, there are also 1,271 **PCGs** lost by **MSbPP,** including 482 hypothetical proteins, 269 "similar to"/"-like" type proteins, 165 ORFs related, 55 zinc finger proteins, 29 immunity products and 268 genes with other functions. Evidently, although **MSbPP** searching databases are much better in integrality than the probe sets of transcriptome profiling, but remain insufficient for full identification of proteome.

When we talk about the completion of transcriptome and proteome, two kinds of "completion" should be differentiated: Other than genome blue print, transcriptome is with temporal (expression status) variable; proteome is with both spatial (subcellular) and temporal variables. The completion of all infinite statuses is of course infeasible. If the "completion" refers to grasp all elements (mRNAs or proteins) in a measurement, such as low abundance issues, we think that is possible to achieve the goal by technique improving.

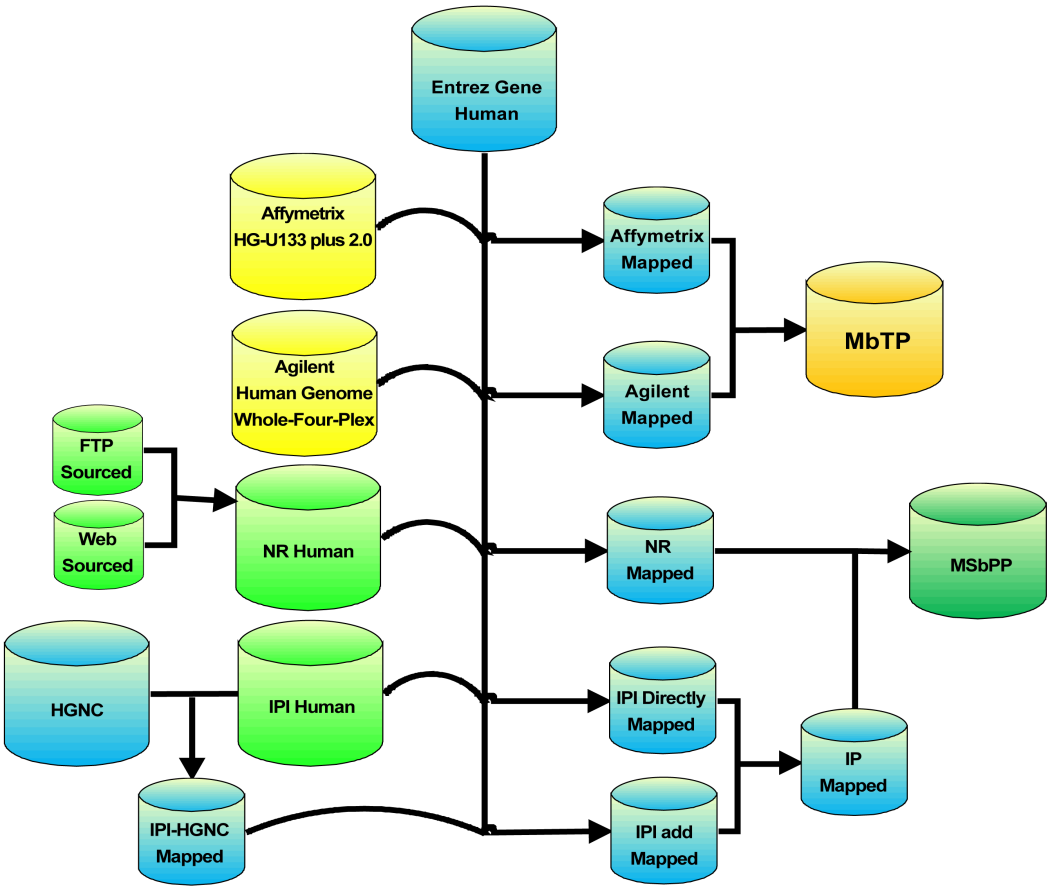

Figure 1: Flowchart of **MbTP** and **MSbPP** datasets construction.

Gene Type	All Entrez	Transcriptomic Microarray Probe sets						Proteomic Searching Databases					
		Affy	Agil	Affy∩Agil	Affy∪Agil	Affy Only	Agil Only	IPI	NR	IPI∩NR	IP∪NR	IPI Only	NR Only
PCG	25,611	19,523	18,262	17,624	20,161	1,899	638	21,768	24,155	21,583	24,340	185	2,572
pseudogene	6,543	733	802	244	1,291	489	558	464	557	377	644	87	180
miscRNA	638	88	62	55	95	33	7	49	5	4	50	45	1
snoRNA	340	18	3	0	21	18	3	2	0	0	2	2	0
snRNA	51	0	1	0	1	0	1	0	0	0	0	0	0
rRNA	10	0	0	0	0	0	0	0	0	0	0	0	0
tRNA	75	0	0	0	0	0	0	0	0	0	0	0	0
other	710	59	26	23	62	36	3	136	272	135	273	1	137
unknown	2,567	662	454	399	717	263	55	712	912	663	961	49	249
sum	36,545	21,083	19,610	18,345	22,348	2,738	1,265	23,131	25,901	22,762	26,270	369	3,139

Affy:Affymetrix HG-U133 plus 2.0; Agil:Agilent Human Genome Whole-Four-Plex 44K; ∩:Overlapped; ∪:Combined;
PCG: protein coding gene; miscRNA: miscellaneous RNA; snoRNA: small nucleolar RNA; snRNA: small nuclear RNA;
other: when the type is known, but there is no specific enumeration for it; unknown: when the type of gene is uncertain.

Table 1: Coverage of most widely used transcriptomic probe sets and proteomic searching databases to human genome.

Reference

1. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, et al. (2003) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 18: 630-634. » CrossRef » Pubmed » Google Scholar

2. Henzel WJ, Watanabe C, Stults JT (2003) Protein identification: the origins of peptide mass fingerprinting. J Am Soc Mass Spectrom 14: 931-942. » CrossRef » Pubmed » Google Scholar

3. Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, et al. (2004) The International Protein Index: An integrated database for proteomics experiments. Proteomics 4: 1985-1988. » CrossRef » Pubmed » Google Scholar

4. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921. » CrossRef » Pubmed » Google Scholar

5. Lipshutz RJ, Morris D, Chee M, Hubbell E, Kozal MJ, et al. (1995) Using oligonucleotide probe arrays to access genetic diversity. Biotechniques 19: 442-447. » CrossRef » Pubmed » Google Scholar

6. Maglott D, Ostell J, Pruitt KD, Tatusova T (2007) Entrez Gene: gene-centered information at NCBI. Nucleic acids Res 35: D26-31. » CrossRef » Pubmed » Google Scholar

7. Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270: 467-470. » CrossRef » Pubmed » Google Scholar

8. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270: 484-487. » CrossRef » Pubmed » Google Scholar

9. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304-1351. » CrossRef » Pubmed » Google Scholar

10. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, et al. (2008) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 36 : D13-21. » CrossRef » Pubmed » Google Scholar

OMICS Publishing Group

List of Journals

1. Journal of Proteomics & Bioinformatics (JPB) **Open Access**
2. Journal of Bioequivalence & Bioavailability (JBB) **Open Access**
3. Journal of Computer Science & Systems Biology (JCSB) **Open Access**
4. Journal of Antivirals & Antiretrovirals (JAA) **Open Access**
5. Journal of Data Mining in Genomics & Proteomics (JDMGP) **Open Access**
6. Journal of Glycomics & Lipidomics (JGL) **Open Access**
7. Journal of Pharmacogenomics & Pharmacoproteomics(JPP) **Open Access**
8. Journal of Postgenomics: Drug & Biomarker Development (JPDBD) **Open Access**
9. Journal of Biochips and Tissue Chips (JBTC) **Open Access**
10. Journal of Molecular Biomarkers & Diagnosis **(JMBD) Open Access**
11. Journal of Bioanalysis & Biomedicine (JBABM) **Open Access**
12. Journal of Nanomedicine & Biotherapeutic Discovery **(JNBD) Open Access**
13. Journal of Cancer Science & Therapy (JCST) **Open Access**
14. Journal of Biotechnology & Biomaterials (JBTBM) **Open Access**
15. Journal of Petroleum & Environmental Biotechnology **(JPEB) Open Access**
16. Journal of Microbial & Biochemical Technology (JMBT) **Open Access**
17. Journal of Tissue Sciences & Engineering (JTSE) **Open Access**
18. Briefings in Intellectual Property Rights (BIPR) **Open Access**

ISSN:0974-276X JPB, an open access journal