Das Internationale »H-invitational« Projekt zur funktionellen Annotation humaner cDNAs

Stefan Wiemann · Abteilung Molekulare Genomanalyse, Deutsches Krebsforschungszentrum

Mit dem Beginn des Humangenomprojekts kamen die systematische Isolierung und die Funktionsanalyse der menschlichen Gene ins Blickfeld. Hochdurchsatz-Entwicklungen blieben hierbei weitgehend auf das Gebiet der genomischen Sequenzierung beschränkt, während die Bedeutung der systematischen Analyse von cDNAs längere Zeit unbemerkt blieb. Über die Sequenzierung zunächst von ESTs – Expressed Sequence Tags (Adams *et al.*, 1992) und später auch von Volle-Länge cDNAs – FL-cDNAs (Nomura *et al.*, 1994) wurde schließlich auch die experimentelle Analyse der mRNA auf eine systematische Ebene gestellt. Das Deutsche cDNA-Konsortium wurde 1996 im DHGP als weltweit zweites Projekt seiner Art gegründet, mit dem Ziel, im hohen Durchsatz neue Gene zu identifizieren und zu analysieren. Nicht zuletzt durch dieses Konsortium, das bisher über 12.300 cDNAs mit zusammen 40 Megabasen Sequenz analysiert hat, wurde der wissenschaftlichen Öffentlichkeit die Bedeutung der cDNA-Analyse verdeutlicht (Abbott 2001, Pennisi 2001). Derzeit gibt es drei große cDNA-Projekte weltweit, das Deutsche cDNA-Konsortium (Wiemann *et al.*, 2001), die US-amerikanische Mammalian Gene Collection (Strausberg *et al.*, 2002) und das vereinte Japanische FLJ-Projekt (Ota *et al.*, 2004), die gemeinsam die Identifizierung sämtlicher humaner Gene sowie die Bereitstellung von physikalischen Klon-Ressourcen zum Ziel haben. Diese Ressourcen sind insbesondere für die aufkommenden Hochdurchsatz-Applikationen der »Functional Genomics« und »Proteomics« unentbehrlich (Wiemann *et al.*, 2003).

Seit 1996 treffen sich die Leiter der drei großen cDNA-Projekte in regelmäßigen Abständen zu »Workshops on Complete cDNA Sequencing – WCCS«, um neue Entwicklungen in der Technologie zu erörtern und die Aktivitäten der Projekte zu koordinieren. Neben diesen Workshops hat sich die Serie von »Transcriptome« Konferenzen etabliert, die in den Jahren 2000 (Paris), 2002 (Seattle) und 2003 (Tokyo) Wissenschaftler aus aller Welt zusammenbrachte, um neueste Entwicklungen von cDNA-basierter Hochduchsatz-Forschung bis zur Systembiologie zu kommunizieren und zu diskutieren.

In den WCCS Workshops wurde die Idee entwickelt, in einer gemeinsamen Anstrengung sämtliche in Hochdurchsatz-Projekten sequenzierten humanen cDNAs gemeinsam funktionell zu annotieren und so eine bis dato einmalige internationale Zusammenarbeit zu initiieren.

Die Organisation eines ersten Annotations-Jamborees in diesem internationalen Projekt wurde von den japanischen Kollegen übernommen. 150 Wissenschaftler von 60 Institutionen aus fünf Kontinenten trafen sich im Sommer 2002 für zehn Tage zum 1. »H-invitational« cDNA Annotations-Workshop in Tokyo, Japan (Cyranoski 2002). Das Deutsche cDNA-Konsortium war mit fünf eingeladenen Wissenschaftlern vertreten. Basis der Annotationsarbeit waren 1. die Sequenzen der Hochdurchsatz cDNA-Projekte aus Japan, Deutschland, den USA und China (Abbildung 2), 2. die an mehreren Bioinformatik-Instituten (z.B. DDBJ, NCBI, EBI, MIPS, SANBI) durchgeführte Vorarbeit mit automatischen Analysen und 3. die Expertise der beteiligten Wissenschaftler aus Institutionen, die systematische Analyse und Annotation von Sequenzen betreiben. Speziell für dieses Projekt wurde ein web-basiertes Annotationssystem entwickelt, das unmittelbar an eine Datenbank gekoppelt ist, in der die Annotationen abgelegt und abfragbar zugänglich sind. Im »H-Invitational« Workshop wurden insgesamt 41.118 cDNA-Sequenzen aus den Hochdurchsatz-Projekten geclustert, um letztlich 20.899 Gen-Loci zu definieren, die sämtlich während des Workshops manuell annotiert wurden. Die Ergebnisse des Workshops werden in diesen Tagen veröffentlicht (Imanishi *et al.*, 2004).

Als Ziele des Projekts wurden definiert:

1. Die Identifizierung humaner Gene mittels hoch-qualitativer cDNA Ressourcen.
2. Die Durchführung einer manuellen, funktionellen Annotation experimentell validierter Gene und repräsentativer cDNAs.
3. Die Identifizierung von Spleißvarianten, mit denen ein Teil der Komplexität des Transkriptoms und Proteoms erklärt werden kann.
4. Die Identifizierung und Annotation funktioneller RNAs.

Abb. 1: 150 Wissenschaftler aus 60 Instituten trafen sich in Tokyo zur systematischen manuellen Annotation humaner Gene und cDNAs.

Abb. 2: Die Sequenzen aus den internationalen Hochdurchsatz Projekten wurden gemeinsam annotiert und bilden damit die größte Ressource menschlicher FL-cDNAs weltweit.

5. Die Verknüpfung von Krankheits-Informationen mit Genen und repräsentativen cDNAs, nicht zuletzt durch die systematische Analyse von SNPs.

Sämtliche Annotationen wurden in die »H-Invitational«-Datenbank eingegeben und werden mit dem Erscheinen der gemeinsamen Publikation (Imanishi *et al.*, 2004) im Internet über www.h-invitational.jp/ verfügbar gemacht. Damit ist diese in der wissenschaftlichen Gemeinschaft größte humane cDNA-Sequenzdatenbank mit manueller Annotation weltweit frei zugänglich.

Im November 2003 fand, wieder in Tokyo, ein zweiter H-invitational Workshop statt, in dessen Verlauf weitere 15.000 cDNAs annotiert wurden, die im vergangenen Jahr neu sequenziert worden waren. Die Reihe der Annotations-Workshops, wie auch der »Transcriptome« Konferenzen, werden weiter fortgesetzt, um letztlich umfassende Daten der menschlichen Gene und Genprodukte in Händen zu haben und der Öffentlichkeit zu präsentieren. Durch die Verknüpfung der großen systematischen cDNA Projekte weltweit ist eine Ressource von manuell annotierten cDNA-Klonen geschaffen, die für die Funktionsanalyse von Genen und Genprodukten und weiter für die Analyse von Krankheitsassoziationen sowie die Identifizierung von diagnostischen Markern und therapeutischen Targets unverzichtbar ist. H-invitational ist eine bisher einmalige Initiative zur funktionellen Annotation des menschlichen Transkriptoms.

Kontakt

Stefan Wiemann
Abteilung Molekulare Genomanalyse
Deutsches Krebsforschungszentrum
Im Neuenheimer Feld 580 · 69120 Heidelberg
s.wiemann@dkfz.de

Literatur

Abbott, A. (2001). ***Free access to cDNA provides impetus for gene function work.*** *Nature 410, 289-290.*
Adams, M. D., Dubnick, M., Kerlavage et al. (1992). ***Sequence identification of 2,375 human brain genes.*** *Nature 355, 632-634.*
Cyranoski, D. (2002). ***Geneticists lay foundations for human transcriptome database.*** *Nature 419, 3-4.*
Imanishi, T., Itoh, T., Suzuki, Y. et al.(2004). ***Integrative Annotation of 20,899 Human Genes Validated by Full-Length cDNA Clones.*** *PLoS Biol im Druck.*
Nomura, N., Miyajima, N., Sazuka, T. et al. (1994). ***Prediction of the coding sequences of unidentified human genes. I. The coding sequences of 40 new genes (KIAA0001-KIAA0040) deduced by analysis of randomly sampled cDNA clones from human immature myeloid cell line KG-1.*** *DNA Res 1, 47-56.*
Ota, T., Suzuki, Y., Nishikawa, T., et al. (2004). ***Complete sequencing and characterization of 21,243 full-length human cDNAs.*** *Nat Genet 36, 40-45.*
Pennisi, E. (2001). So many choices, so little money. Science 294, 82-85.
Strausberg, R. L., Feingold, E. A., Grouse, L. H. et al. (2002). ***Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences.*** *Proc Natl Acad Sci U S A 99, 16899-16903.*
Wiemann, S., Bechtel, S., Bannasch, D. et al. (2003). ***The German cDNA Network: cDNAs, functional genomics and proteomics.*** *Journal of Structural and Functional Genomics 4, 87-96.*
Wiemann, S., Weil, B., Wellenreuther, R. et al. (2001). ***Toward a Catalog of Human Genes and Proteins: Sequencing and Analysis of 500 Novel Complete Protein Coding Human cDNAs.*** *Genome Res 11, 422-435.*

Autoren der Publikation: ***Integrative Annotation of 20,899 Human Genes Validated by Full-Length cDNA Clones.*** *PLoS Biol im Druck.(2004) Imanishi, T., Itoh, T., Suzuki, Y., O'Donovan, C., Fukuchi, S., Koyanagi, K. O., Barrero, R. A., Tamura, T., Yamaguchi-Kabata, Y., Tanino, M., Yura, K., Miyazaki, S., Ikeo, K., Homma, K., Kasprzyk, A., Nishikawa, T., Hirakawa, M., Thierry-Mieg, J., Thierry-Mieg, D., Ashurst, J., Jia, L., Nakao, M., Thomas, M. M., Mulder, N., Karavidopoulou, Y., Jin, L., Kim, S., Yasuda, T., Lenhard, B., Eveno, E., Suzuki, Y., Yamasaki, C., Takeda, J., Gough, C.,* ***Amid, C.****, Bellgard, M., de Fatima Bonaldo, M., Bono, H., Bromberg, S. K., Brookes, A., Bruford, E., Carninci, P., Chelala, C., Couillault, C., de Souza, S., Debily, M., Devignes, M., Dubchak, I., Endo, T., Estreicher, A., Eyras, E., Fukami-Kobayashi, K., Gopinathrao, G., Graudens, E., Hahn, Y.,* ***Han, M.****, Han, Z., Hanada, K., Hashimoto, K., Hinz, U., Hirai, M., Hishiki, T., Hopkinson, I., Imbeaud, S., Inoko, H., Kanapin, A., Kasukawa, T., Kelso, J., Kersey, P., Kikuno, R., Kimura, K., Korn, B.,* ***Kuryshev, V.****, Makalowska, I., Makalowski, W., Makino, T., Mano, S., Mariage-Samson, R., Mashima, J., Matsuda, H.,* ***Mewes, H. W.****, Minoshima, S., Nagai, K., Nagasaki, H., Nigam, R., Ogasawara, O., Ohara, O., Ohtsubo, M., Okada, N., Okido, T., OOta, S., Ota, M., Ota, T., Otsuki, T., Piatier-Tonneau, D.,* ***Poustka, A.****, Ren, S., Saitou, N., Sakai, K., Sakamoto, S., Sakate, R.,* ***Schupp, I.****, Servant, F., Sherry, S., Shimizu, N., Shimoyama, M., Simpson, A. J., Soares, B., Steward, C., Suwa, M., Suzuki, M., Takahashi, A., Tamiya, G., Tanaka, H., Taylor, T., Terwilliger, J. D., Unneberg, P., Watanabe, S., Wilming, L., Yasuda, N., Yoo, H., Veeramachaneni, V., Stodolsky, M., Go, M., Nakai, K., Takagi, T., Kanehisa, M., Sakaki, Y., Quackenbush, J., Okazaki, Y., Hayashizaki, Y., Hide, W., Chakraborty, R., Nishikawa, K., Sugawara, H., Tateno, Y., Chen, Z., Oishi, M., Tonellato, P., Apweiler, R., Okubo, K., Wagner, L.,* ***Wiemann, S.****, Strausberg, R. L., Isogai, T., Auffray, C., Nomura, N., Gojobori, T. and Sugano, S.(****Wissenschaftler des Deutschen cDNA Konsortiums sind hervorgehoben****)*