Rede ONSA e o Projeto Genoma Humano do Câncer: Contribuição ao Genoma Humano

atualização

Edna T. Kimura
Gilson S. Baía

RESUMO

A contribuição maior da ciência brasileira ao genoma humano foi trazida pelo Projeto Genoma Humano do Câncer (*Human Genome Cancer Project* – HCGP) uma parceria da FAPESP e do *Ludwig Institute for Cancer Research* e desenvolvido por 29 diferentes laboratórios de seqüenciamento e um centro de bioinformática. Foram seqüenciados mais de 1 milhão de fragmentos gênicos expressos (*expressed sequences tags, ESTs*), provenientes de diferentes tumores humanos. Grande parte destes dados é de acesso público através da website do *Gene Bank* (www.ncbi.nlm.nig.gov), mantido pelo NCBI – *National Center for Biotechnology Information*. Atualmente, diversos projetos estão em desenvolvimento utilizando informações geradas no HCGP e abrangem observar a expressão diferenciada dos genes em diferentes tumores, caracterização completa de genes específicos, assim como o estudo funcional e estrutural dos produtos protéicos. É promissora a perspectiva de que num futuro próximo, diferentes resultados provenientes destas investigações possam trazer benefícios preventivos, prognósticos e clínicos em câncer e outras doenças. **(Arq Bras Endocrinol Metab 2002;46/4:325-329)**

Descritores: Genoma humano; ESTs; ORESTES; GenBank; Transcriptoma; Genoma

Departamento de Histologia e Embriologia, Instituto de Ciências Biomédicas, Universidade de São Paulo, SP.

ABSTRACT

ONSA Network and The Human Cancer Genome Project - Contribution to The Human Genome.

The Human Cancer Genome Project - HCGP - has been the most important contribution of Brazilian science to the understanding of the Human Genome. This project was co-sponsored by FAPESP (São Paulo Research Foundation) and the Ludwig Institute for Cancer Research. Twenty-nine sequencing laboratories integrated by a bioinformatic service and a laboratory carried out the project. This effort has generated more than 1 million of expressed sequence tags (ESTs) in various human cancers. Most of these sequences are available to public accessing at NCBI Gene Bank (www.ncbi.nlm.nih.gov by ORESTES keyword). Currently there are several ongoing projects that utilize the information and genomic library generated by HCGP. These studies are focused in the characterization of a complete physical map of genes, differential gene expression in tumors and also in the structural and functional understanding of the expressed gene. In the future, the generated knowledge would contribute to new insights into the prevention, diagnosis and therapeutic procedures in cancer and other diseases. **(Arq Bras Endocrinol Metab 2002;46/4:325-329)**

Keywords: Human genome; ESTs; ORESTES; GenBank; Transcriptome; Genome

Recebido em 30/06/2002
Aceito em 05/07/2002

EM MEADOS DA DÉCADA DE 80, o Consórcio Genoma Humano foi lançado pelos EUA e congregou, posteriormente, países da Europa e o Japão, inicialmente interessados na construção de um mapa genético humano (1). O objetivo principal do Consórcio Público Internacional visava produzir uma seqüência contínua de cada um dos 24 cromossomos e o delineamento de todos os genes humanos. No final da década de 90, o seqüenciamento em larga escala do genoma teve considerável impulso, devido especialmente ao aperfeiçoamento contínuo e à automação do processo de seqüenciamento de DNA e, também, pelo rápido avanço da tecnologia computacional de análise de seqüências de DNA (2,3).

Em meados de 2000, um esboço do genoma humano foi anunciado após uma acirrada competição entre o Projeto de financiamento público *(Human Genome Project*) e uma empresa privada norte americana, a Celera, e o resultado foi publicado nas revistas *Nature* e *Science*, respectivamente (4,5). O livre acesso *online* ao primeiro "rascunho" do genoma humano gerado pelo consórcio público é mantido pela *University of California at Santa Cruz* (*Human Genome Project Working Draft* - http://genome.cse.ucsc.edu/) e pelo *National Center for Biotechnology Information* - NCBI (http://www.ncbi.nlm.nih.gov). Estes *websites,* em atualização contínua, possibilitam uma análise de predição *in silico* (computacional) para possíveis genes desconhecidos, por meio de análises comparativas de fragmentos gênicos expressos com a seqüência genômica completa, assim como permitem comparações com informações acumuladas em diferentes bancos de dados de ESTs humanos (6).

O Consórcio ONSA no Brasil

No ano de 1997, a Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) lançou a rede ONSA (*Organization for Nucleotide Sequencing and Analysis - The Virtual Genomics Institute),* uma iniciativa histórica no Brasil com o objetivo de executar um projeto genoma nacional, numa área de alta competitividade no mundo científico. O consórcio ONSA, uma rede virtual cujo modelo de formação de consórcio de pesquisa privilegiou a criação de redes de cooperação entre centros de pesquisa, integrou neste primeiro projeto diversos laboratórios paulistas para seqüenciamento de DNA em larga escala, com o suporte centralizado para bioinformática no Instituto de Computação da Universidade Estadual de Campinas (Unicamp) (7).

O primeiro projeto desenvolvido foi o seqüenciamento do genoma completo da bactéria *Xylella fastidiosa*, agente causal da Clorose Variegada do Citrus (CVC), doença que nos últimos anos provocou perdas econômicas consideráveis na produção e, conseqüentemente, na exportação de suco de laranja. Em 1996, estimou-se que em 30% da área plantada no estado de São Paulo foram encontradas plantas com sintomas severos de CVC. Na rede ONSA para o Genoma da *Xylella*-CVC (http://aeg.lbi.ic.unicamp.br/xf/), cerca de 30 laboratórios foram equipados com seqüenciadores automáticos de DNA, e foi a primeira das diversas iniciativas que se seguiram em todo o país, com o objetivo de criar competência e de produzir seqüenciamento de DNA em larga escala. Por cerca de dois anos, o projeto de seqüenciamento e análise do genoma da *Xylella* foram desenvolvidos e resultou em artigo publicado na revista *Nature* em Julho de 2000 (8). Em seu editorial, a revista chama a atenção pelo marco histórico sob o ponto de vista científico e também político que a comunidade científica brasileira conquistou, ressaltando que no âmbito mundial o genoma da *Xylella*-CVC foi o primeiro organismo fitopatogênico seqüenciado e anotado (9). Atualmente, cerca de 30 projetos explorando o genoma desta bactéria estão sendo financiados pela FAPESP, desenvolvendo estudos funcionais, visando o entendimento das interações planta-patógeno, e possibilidades de buscar formas mais eficazes para o controle da doença.

A iniciativa de seqüenciamento genômico da rede ONSA na área da agricultura foi ampliada para o seqüenciamento dos genomas de *Xanthomonas ssp* (http://genoma4.iq.usp.br/xanthomonas) (10) e ESTs de cana-de-açúcar (http://sucest.lad.dcc.unicamp.br/en). Mais recentemente, o *"Agronomical & Environmental Project"* (*AEG Project* - http://aeg.lbi.ic.unicamp.br/) vem seqüenciando sucessivamente os genomas das bactérias *Xylella fastidiosa (Pierce's Disease Strain)*, *Leifsonia xyli subsp. xyli*, ESTs de *Eucaliptus grandis* e de *Coffee arabica*. Dentre os parasitas humanos, está em andamento o seqüenciamento de ESTs de *Schistosoma mansoni* (http://verjo18.iq.usp.br/schisto).

Genoma Humano do Câncer

Em 1999, a rede ONSA lançou o Projeto Genoma Humano do Câncer (*Human Cancer Genome Project* - HCGP), área do Genoma Humano de alta competitividade no cenário internacional, com o objetivo inicial de produzir meio milhão de *expressed sequence tags* (ESTs), de tecido tumoral humano. O Projeto HCGP, uma parceria financeira entre a FAPESP e o *Ludwig Institute for Cancer Research* (Instituto Ludwig para Pesquisa sobre o Câncer), envolveu o trabalho conjunto de cerca de 29 laboratórios paulistas, por dois

anos (http://www.ludwig.org.br/ORESTES/grupos.html).

Tecidos tumorais humanos foram analisados para a identificação de genes expressos na caracterização de *ESTs* pelo método ORESTES (*open reading frame expressed sequence tags*). Esta metodologia privilegia a porção central dos genes, pelo uso de oligonucleotídeos (*primers*) arbitrários, não degenerados, em condições de baixa estringência, em reações de PCR antecedidas de uma reação de transcrição reversa (11,12). Aliado a esta metodologia que detecta posições centrais de genes expressos, o sucesso do Projeto deveu-se ao investimento em seqüenciadores de última geração, ainda mais potentes que os utilizados em projetos anteriores no Brasil, os seqüenciadores automáticos multi-capilares, MegaBace 1000‰ (Amersham BioSciences). A utilização deste equipamento, capaz de seqüenciar 96 seqüências em cerca de 3 horas, fez com que o Projeto atingisse em menos de um ano a meta inicial de seqüenciar 500.000 seqüências, sendo então expandida a meta inicial para 1 milhão de seqüências, para o período total do Projeto. Na realidade mais de 1 milhão de ESTs foram seqüenciadas, tendo sido o resultado criteriosamente avaliado pelo Centro de Bioinformática, que além da qualidade técnica do seqüenciamento, baseado no Programa computacional phredPhrap (*Washington University*), checava possíveis contaminações dos *reads* de ESTs com seqüências de vetores plasmidiais e de bactérias utilizados no procedimento metodológico (13). Somente seqüências com padrões de qualidade estabelecida foram computadas para constituir o banco de dados do HCGP (http://www.ludwig.org.br/ORESTES). Até o momento 823.987 seqüências ORESTES do Banco HCGP estão depositados no GenBank – *National Center for Biotechnology Information*, EUA, sendo acessíveis publicamente no *website* http://www.ncbi.nlm.nih.gov, busca em <nucleotide>, com a palavra-chave ORESTES.

O alinhamento ou *assembly* das seqüências (*reads*) para formação de *contigs*, conjuntamente com os dados públicos do GenBank e UniGene, marcam enorme oportunidade para a descoberta de potenciais *open reading frames* (pORFs), os quais podem ser aqueles com funções gênicas já anotadas (genes de função conhecida ou presumida) e suas formas de *splicing*, ou genes novos, ainda desconhecidos (14). Particularmente, as seqüências ORESTES agregaram dados significativos aos bancos de dados públicos de cDNAs, possibilitando a construção de *contigs* mais completos e representativos de todas as seqüências transcritas. Estes dados facilitam o mapeamento de genes humanos e o entendimento de sua funcionalidade, bem como contribuem para a determinação da expressão gênica diferencial e para a distinção de variantes de *splicings*. Para a elucidação do transcriptoma humano é imperativo não só acumular seqüências expressas de DNA, mas sobretudo identificar e analisar os diferentes transcritos de um mesmo gene, e também analisar inteiramente o conjunto de transcritos de células e tecidos, para o início da compreensão da complexa rede de interações, codificada pelos mais diversos transcritos. Cerca de 25% dos segmentos gênicos expressos são ESTs ainda desconhecidos, não apresentando seqüências com similaridade em nenhum banco de dados (humanos e não humanos) disponíveis, indicativo da presença de novos genes ainda não caracterizados (15).

Desta forma, ao término do HCGP, diversos projetos subseqüentes de análise destes bancos de dados estão em desenvolvimento. Dentre as iniciativas relacionadas com o Projeto HCGP, destacamos alguns projetos em andamento:

Identificação de Polimorfismo de Base Única (SNPs-ORESTES):

Este projeto em desenvolvimento no Hemocentro da Faculdade de Medicina da USP de Ribeirão Preto tem o objetivo de identificar *Single Nucleotide Polymorphisms* em regiões codificantes (cSNPs) do genoma humano, a partir de 1,2 milhões de ESTs/ORESTES do HCGP (http://bit.fmrp.usp.br). A importância do estudo e mapeamento de SNPs está relacionada à possibilidade de sua associação com genótipos de suscetibilidade para patogenicidades diversas, sendo promissor no entendimento do mecanismo de resposta de suscetibilidade à doença e tratamento individualizado.

CAGE-Câncer:

O Projeto "Cooperação para a Análise dos Genes e sua Expressão no Câncer", em desenvolvimento no Instituto de Química da USP (http://verjo19.iq.usp.br/cagecancer), tem como objetivo observar as alterações na expressão de genes em tecidos humanos normais e tumorais, com ênfase em tecidos de próstata e pulmão, a partir de seqüências ORESTES geradas no HCGP. Procura ainda identificar novos marcadores genéticos de malignidade e de prognóstico para o tratamento do câncer de próstata, utilizando a técnica de DNA-*chip* ou DNA-*microarray*. Recentemente, na reportagem "*Caçadores de genes*" da Revista Pesquisa FAPESP (edição de no. 74, de Abril de 2002 - http://revistapesquisa.fapesp.br), o grupo CAGE-Câncer anunciou a descoberta de seis genes nunca

antes descritos na literatura médica e que podem estar relacionados ao câncer de próstata. Segundo os autores, os genes poderão se tornar uma ferramenta importante para auxiliar no diagnóstico precoce da enfermidade e/ou da sua evolução clínica, se estudos funcionais comprovarem relação com esse tipo de câncer.

Transcript Finishing Initiative – TFI:

O Projeto *TFI* (http://www.compbionet.org.br/transcript), uma cooperação entre o Instituto Ludwig e a FAPESP, está sendo desenvolvido por uma rede de 29 laboratórios paulistas, e tem como objetivo validar a estrutura gênica completa de 4.000 genes humanos (*full-length genes*). A seqüência genômica do Projeto Genoma Humano Internacional é comparada com bancos de dados de ESTs, incluindo o banco de dados do HCGP. A busca de potenciais fragmentos expressos para genes específicos é realizada pela verificação da expressão gênica em diferentes cDNAs de tecido ou linhagem tumoral humana, supondo-se que estas regiões não sejam totalmente caracterizadas, pela ausência de fragmentos de ESTs que se alinham a estas porções cromossômicas. Ao final do projeto são esperadas informações acumuladas de novos exons, formas alternativas de *splicing* não descritas e respectivas localizações cromossomais, com o mapa físico de 4.000 genes humanos.

Head & Neck Transcriptome:

Este projeto, envolvendo dois centros de Bioinformática (Instituto de Química/USP, SP e Instituto de Biologia, Unicamp) e diversos pesquisadores com interesse em câncer de cabeça e pescoço, está analisando detalhadamente o conjunto de dados ORESTES provenientes destes tecidos, incluindo tumores de tiróide, gerados pelo HCGP (http:// www.lge.ibi.unicamp.br/cancerhn). As seqüências do Banco HCGP provenientes de tumores de cabeça e pescoço foram *clusterizados*, e os *clusters* foram comparados aos dados públicos de genes. A análise objetiva identificar a função gênica (processo de anotação) relacionada a conjuntos de *ESTs,* formas ainda não descritas de *splicing* alternativo, assim como a identificação de novos genes expressos nestes tumores e tecido controle.

Genoma Clínico:

O Projeto Genoma Clínico, uma parceria do Instituto Ludwig e da FAPESP (http://mbl.fmrp.usp.br) congrega centros de estudos para neoplasias osteo-hematopoéticas (leucemia aguda, osteosarcoma, mieloma múltiplo), neoplasias gastro-intestinais (câncer gástrico, câncer de esôfago, câncer colo-retal, carcinomas de cabeça e pescoço), neoplasias neurológicas (astrocitomas). Neste projeto, a expressão gênica das células neoplásicas está sendo relacionada com características clínicas de pacientes para a identificação de genes relevantes no diagnóstico e prognóstico, para definição de subtipos da neoplasia e predição de respostas terapêuticas a diferentes tratamentos.

CONCLUSÃO

A participação do Brasil na pesquisa genômica mundial nos últimos anos é um marco histórico na ciência brasileira. O passo decisivo nesta área se deve, especialmente, à visão pioneira de organizar os grupos numa rede virtual e criar competência qualificada para o seqüenciamento em larga escala, análise de genomas e bioinformática em diversos centros de pesquisa. O banco de genes humanos mantido pela FAPESP em centros de pesquisa e pelo Instituto Ludwig – Hospital do Câncer e os dados acumulados no Projeto HCGP, estão contribuindo para a caracterização dos genes humanos e suas funções, para a expansão nos estudos de expressão diferenciada de genes pela utilização em técnicas de *microarrays* e com potencial utilização para estudos de proteoma. A informação do seqüenciamento e análise completa do genoma huma-no serão imprescindíveis para a medicina futura, am-pliando as possibilidades de diagnósticos preventivos e prognósticos assim como novas terapêuticas.

Ao assumir um Projeto tão competitivo como o HCGP, e outros programas na área genômica, o Brasil assegurou um potencial inestimável para a ciência brasileira, não somente pelos bancos genômicos e os da-dos gerados, mas também pela formação de profissionais altamente qualificados. O resultado do investimento desta iniciativa brasileira certamente se traduzirá em importantes contribuições científicas futuras.

AGRADECIMENTOS

Agradecemos o auxílio financeiro da Fundação de Amparo à Pesquisa no Estado de São Paulo em projetos de nosso laboratório relacionados a este artigo (FAPESP 99/03661-9, 00/10165-7, 00/12488-8) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

REFERÊNCIAS

1. Olso MV. Time to sequence. **Science 1995**;270:394-6.
2. Wolfsberg TG, McEntyre J, Schuler GD. Guide to the draft human genome. **Nature 2001**;409:824-6.
3. Baltimore D. Our genome unveiled. **Nature 2001**;409:814-6.
4. International Human Genome Consortium. **Nature 2001**;409:860-921.
5. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton, et al. The sequence of the human genome. **Science 2001**;291:1304-51.
6. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. **Genome Res 2002**;12:996-1006.
7. Simpson AJG, Perez JF. ONSA, the São Paulo Virtual Genomics Institute. Organization for Nucleotide Sequencing and Analysis. **Nat Biotechnol 1998**;I16:795-6.
8. Simpson AJ, Reinach FC, Arruda P, Abreu FA, Acencio M, Alvarenga R, et al. The genome sequence of the plant pathogen *Xylella fastidiosa.* - The Xylella fastidiosa Consortium of the Organization for Nucleotide Sequencing and Analysis. **Nature 2000**;406:151-9.
9. Nature editorial. Genome sequencing for all. **Nature 2000**;406:109.
10. da Silva AC, Ferro JA, Reinach FC, Farah CS, Furlan LR, Quaggio RB, et al. Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities. **Nature 2002**;417:459-63.
11. Dias Neto E, Garcia Correa R, Verjovski-Almeida S, Briones MR, Nagai MA, da Silva W Jr, et al. Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. **Proc Natl Acad Sci USA 2000**;97:3491-6.
12. Camargo AA, Samaia HP, Dias-Neto E, Simao DF, Migotto IA, Briones MR, et al. The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. **Proc Natl Acad Sci USA 2001**;98:12103-8.
13. Ewing B, Hiller LD, Wendl MC, Green P. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. **Genome Res 1998**;8:175-85.
14. Strauberg RL, Riggins GJ. Navigating the human transcriptome. **Proc Natl Acad Sci 2001**;98:11837-8.
15. de Souza S, Camargo AA, Briones MRS, Costa FF, Nagai MA, Verjovski-Almeida S, et al. Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. **Proc Natl Acad Sci USA 2000**;97:12690-3.

Endereço para correspondência:

Edna T. Kimura
Departamento de Histologia e Embriologia, ICB, USP
Av. Prof. Lineu Prestes 1524, sala 414
05508-000 São Paulo, SP
e.mail: etkimura@usp.br