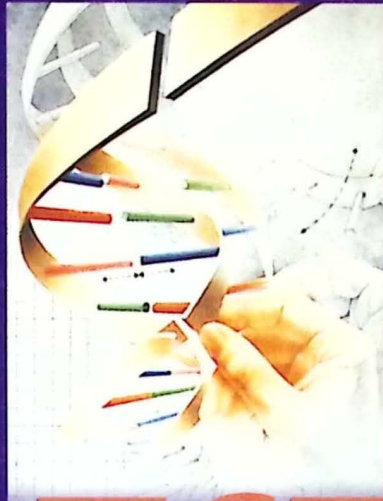




ORESTES

UMA NOVA
TECNOLOGIA
PARA A
DESCOBERTA
DE GENES

ORESTES

UMA NOVA TECNOLOGIA PARA A DESCOBERTA DE GENES

EMMANUEL DIAS NETO e ANDREW J. G. SIMPSON

III Prêmio Alcatel de Tecnologia

OUTUBRO 2000

Instituto Ludwig
de Pesquisas
sobre o Câncer

Fundação de Amparo
à Pesquisa do
Estado de São Paulo

A TECNOLOGIA ORESTES

BREVE APRESENTAÇÃO

Em 26 de março de 1999, foi lançado no auditório da Fundação de Amparo à Pesquisa do Estado de São Paulo-FAPESP, o Projeto Genoma Humano do Câncer, iniciativa resultante de uma parceria estabelecida entre essa fundação, uma das mais importantes agências brasileiras de fomento à pesquisa, e o Instituto Ludwig de Pesquisas sobre o Câncer, respeitada instituição científica internacional, com sede central em Zurique, Suíça, e com uma de suas 10 unidades instalada em São Paulo. A meta do projeto: determinar em dois anos, valendo-se de uma nova e revolucionária tecnologia desenvolvida por dois pesquisadores do Instituto Ludwig de São Paulo, com apoio da FAPESP, 500 mil seqüências de genes expressos em tumores malignos de grande incidência no Brasil. Alcançada essa meta, os pesquisadores envolvidos no projeto teriam, com certeza, dado uma contribuição fundamental para ampliar o conhecimento da parte do genoma humano que realmente tem importância fundamental, ou seja, os cerca de 3% do DNA humano que codificam nossos genes ou que são, propriamente, os genes. Teriam contribuído também para uma maior compreensão sobre as diferenças genéticas entre células normais e células malignas e, portanto, teriam ajudado a alargar o caminho para o desenvolvimento de novos métodos de diagnóstico precoce do câncer e, mais adiante, de novas terapias contra a doença.

Palavras vibrantes marcaram a cerimônia de lançamento do Projeto Genoma Humano do Câncer e, ali, nem de longe elas eram sinônimo de louvação gratuita ou vazia. Em vez disso, refletiam a aguda consciência de que se estava diante de um fato de indiscutível importância para a ciência feita no Brasil e, de modo mais amplo, para a ciência contemporânea, com suas profundas implicações sociais e econômicas, e mesmo repercussões sobre a vida humana, em caráter individual.

Tudo isso porque os responsáveis pelo projeto, aqueles que o formularam e os que decidiram financiá-lo, estavam convencidos de que ele iria permitir ao Brasil entrar efetivamente no grande esforço internacional, liderado pelos Estados Unidos e Inglaterra, para desvendar o genoma humano. E iria permiti-lo exatamente porque o projeto dispunha de um trunfo muito especial em seu poder, a nova tecnologia de

seqüenciamento elaborada em São Paulo e batizada de ORESTES, sigla adaptada para *Open Reading Frames Expressed Sequences Tags*.

Ressalte-se que desvendar o genoma humano, missão do Projeto Genoma Humano, nas suas vertentes pública e privada, é, sem sombra de dúvida, um dos maiores e mais importantes empreendimentos científicos em curso nesta virada do milênio, em razão de seu potencial para alterar radicalmente as bases da biologia e as muitas tecnologias a ela relacionadas. Em razão também de seus imensos efeitos potenciais sobre a saúde e mesmo sobre a duração normal da vida humana.



MIGUEL BOYAN

Festa de abertura: a tecnologia ORESTES viabilizou o Projeto Genoma Humano do Câncer, lançado em março de 1999; na mesa (da direita para a esquerda), o vice-governador de São Paulo, Geraldo Alckmin, o presidente da FAPESP, Carlos Henrique de Brito Cruz, o secretário da Ciência, Tecnologia e Desenvolvimento Econômico do Estado de São Paulo, José Aníbal Pontes, o presidente do Instituto Ludwig, Edward McDermott e o diretor presidente da FAPESP, Francisco Romeu Landi

Essas percepções estavam todas em cena no momento em que o Projeto Genoma Humano do Câncer, movido pela tecnologia ORESTES, estava sendo lançado (como parte do Programa Genoma da FAPESP, iniciado por esta fundação em 1997, sob a liderança de seu diretor científico, José Fernando Perez). E passado, desde então, apenas cerca de um ano e meio, a tecnologia não só tem um indiscutível reconhecimento internacional em meios científicos e em outros mais amplos, como levou o Genoma Humano do Câncer a ter dobradas suas metas quantitativas iniciais de seqüenciamento. Mais: a excelência da tecnologia tornou-se fator de atração de poderosos parceiros interessados em trabalhar conjuntamente com o grupo de pesquisadores brasileiros responsável pelo projeto, e alavanca para a ampliação dos próprios objetivos científicos do empreendimento.

É assim que deverá ser formalizada numa reunião a ser realizada ainda em outubro, em Genebra, Suíça uma parceria proposta pelo Instituto Nacional do Câncer (NCI), uma das mais respeitadas instituições científicas dos Estados Unidos, à FAPESP e ao Instituto Ludwig. A partir disso, os grupos de pesquisa brasileiro e norte-americano passam a seqüenciar, com base nas informações já obtidas pelo projeto, além de fragmentos de genes, também genes inteiros. E passa-se a trabalhar não só com genes expressos em células cancerosas, mas com qualquer gene do genoma humano. Isso leva Robert Strausberg, coordenador da pesquisa do NCI, e Andrew Simpson, coordenador do Projeto Genoma Humano do Câncer, a prever que os dois grupos, com suas técnicas diferentes e complementares, serão os maiores responsáveis pelo seqüenciamento de genes do genoma humano no mundo.



Crescimento surpreendente: em julho de 2000, Mário Covas, governador de São Paulo (ao centro), anunciou a duplicação das metas do projeto; na mesa (a partir da esquerda), estavam o diretor científico e o presidente da FAPESP, José Fernando Perez e Brito Cruz, o ministro-chefe da Secretaria de Comunicação do governo brasileiro, Andrea Matarazzo, o secretário José Aníbal Pontes, o diretor do Instituto do Câncer em São Paulo, Ricardo Brentani e o coordenador do Projeto Genoma Humano do Câncer, Andrew Simpson

Originalidade e outras características singulares de ORESTES

A tecnologia ORESTES, descrita em todos os detalhes no artigo técnico de seus criadores, Emmanuel Dias Neto e Andrew J. G. Simpson, a partir da página 7, é uma tecnologia para a descoberta de genes. E é original, muito diferente da técnica usada, por exemplo, pelo NCI, porque, enquanto esta se dedica ao seqüenciamento nas extremidades dos genes, a tecnologia brasileira permite seqüenciar a região central dos genes, na qual se concentram as informações mais relevantes para que eles desempenhem suas funções. É aí justamente que se concentra a região codificadora dos genes. Esse dado permite que se tenha uma primeira idéia da competência científica e da capacidade criativa que a técnica requereu de seus autores.

Se a isso se soma o alvo pioneiro de aplicação da tecnologia ORESTES, ou seja, tumores malignos, em particular os de alta incidência no Brasil e na América Latina, em cujos tecidos genes normalmente raros do genoma humano se expressam com maior abundância, pode-se formar também uma primeira idéia do impacto positivo que o Projeto Genoma Humano do Câncer prometia sobre as batalhas que a ciência vem movendo contra o câncer. Um impacto que a curto prazo deve se dar mais sobre o diagnóstico precoce do câncer, nos estágios em que a doença tem cura.

Vale a pena resumir alguns aspectos básicos para a apreciação da tecnologia ORESTES nessa breve apresentação, mesmo considerando que eles estão muito melhor descritos no artigo dos autores.

Assim, no que diz respeito ao *caráter inovador* da tecnologia é necessário insistir no fato de que ORESTES permite determinar a informação mais valiosa conti-

da nos genes, auxiliando substancialmente na determinação e caracterização de todos os genes humanos.

Em relação à sua aplicação, vale registrar que a obtenção de fragmentos de genes expressos (chamados de ESTs, do inglês *Expressed Sequence Tags*) é uma ferramenta de fundamental importância para a compreensão dos genomas de organismos complexos como o homem. Depois de se ter decidido o genoma completo, feito notável recentemente realizado pelo Projeto Genoma Humano, é necessário

identificar e determinar a função de todos os genes ativos. A tecnologia ORESTES permite atingir esses objetivos, graças a suas características singulares. Entre elas, está a de privilegiar as seqüências derivadas de genes raros do genoma humano, difíceis de serem obtidas por processos tradicionais, além de favorecer a obtenção da porção central dos genes expressos, que, conforme dissemos, é a região mais rica em informações



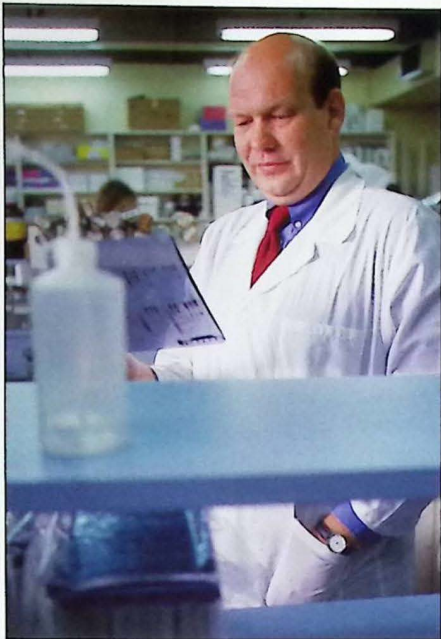
MIGUEL BOYAN

Parceria com o NCI: Victor Jongeneel, coordenador de bioinformática do Ludwig em Lausanne (à esquerda), Robert Strausberg, assistente de direção do Instituto Nacional do Câncer-NCI, dos Estados Unidos, Simpson e Kenneth H. Buetow, da Divisão de Epidemiologia e Genética do Câncer do NCI, acertam os termos da cooperação entre os grupos de pesquisa brasileiro e norte-americano

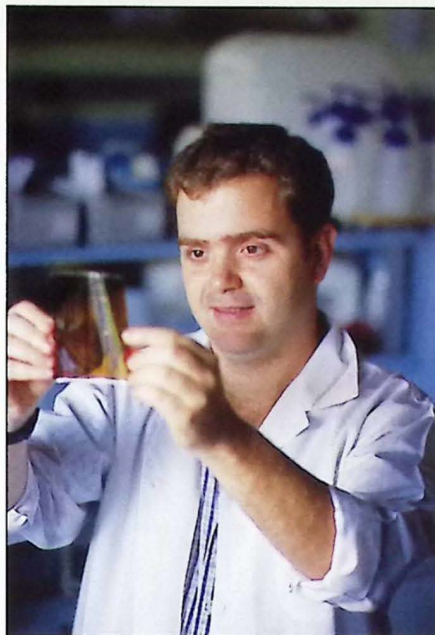
Em relação ao valor e ao conteúdo tecnológico da aplicação, vale registrar que ORESTES oferece um ganho competitivo, porque possibilita atingir-se em larga escala a porção mais informativa dos genes de qualquer organismo, sendo, ao mesmo tempo, uma técnica simples, barata e passível de rápida implantação nos laboratórios de seqüenciamento.

Quanto à escala atual de aplicação da tecnologia e ao potencial futuro da aplicação, observamos que ORESTES é a base do Projeto Genoma Humano do Câncer, que envolve cerca de 150 pesquisadores do Estado de São Paulo, com um investimento a fundo perdido de US\$ 20 milhões. Possivelmente, o Genoma Humano do Câncer é o maior projeto brasileiro em termos de financiamento. Sua produção de dados encontra-se em ritmo extremamente acelerado: mais de 700 mil seqüências de DNA já foram produzidas, o que o transforma no segundo maior projeto do mundo, no campo da pesquisa genômica. O potencial de aplicação de ORESTES, que começa a motivar parcerias internacionais, como a já referida com o NCI, é praticamente ilimitado, porque a técnica pode ser empregada no genoma de qualquer organismo complexo. Ela poderá ser usada como base em projetos como o do genoma do arroz, que está em negociação pelo Ministério da Agricultura do Brasil e Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA), os do genoma do fungo *Paracoccidioides brasiliensis* e do parasita causador da esquistossomose, estes dois com possível financiamento da FAPESP.

MIGUEL BOYVAN



MIGUEL BOYVAN



Criadores bem sucedidos: os autores da tecnologia ORESTES, Andrew Simpson e Emmanuel Dias Neto, trabalhando em seus laboratórios no Ludwig em São Paulo

Em relação aos *impactos da inovação tendo em vista a modernização, eficácia e aumento da produtividade para o empreendimento em que está sendo aplicada*, a existência de ORESTES permitiu que o Brasil entrasse no seleto grupo dos países que contribuem em larga escala para a definição dos genes humanos. O Genoma Humano do Câncer já foi capaz de atrair investimentos substanciais que permitiram equipar dezenas de laboratórios de pesquisas em várias universidades. Sendo

um projeto brasileiro, tem um foco importante na geração de dados centrados em tumores de incidência significativa em nível nacional, que não vêm sendo estudados por países mais desenvolvidos, como os tumores de estômago e cabeça-pescoço. A assimilação desses achados científicos deverá se traduzir na modernização dos métodos de diagnóstico precoce e tratamento de neoplasias humanas. Espera-se também métodos diagnósticos mais eficazes em termos de sensibilidade, carga informativa e especificidade em relação aos vários tipos de câncer. A médio prazo, esses benefícios deverão ser estendidos para o desenvolvimento de métodos mais adequados de tratamento, com menor custo, mais eficácia e menos efeitos colaterais. Os dados gerados pelo Projeto Genoma Humano do Câncer são continuamente disponibilizados para toda a comunidade científica internacional, ultrapassando os limites dos laboratórios envolvidos no trabalho, do Estado de São Paulo e do Brasil.

Vale um registro também em relação à *utilização de tecnologias da informação* movida pela tecnologia ORESTES. Na verdade, a troca de informações tem papel fundamental em todo o trabalho de pesquisa gerado pela técnica. O método ORESTES permite a obtenção de seqüências de DNA extremamente ricas em informação. Todos os dados gerados pelos grupos de seqüenciamento, espalhados entre 33 laboratórios de pesquisa em cidades do Estado de São Paulo, são transmitidas via Internet para a coordenação de bioinformática. Esses dados são comparados com os bancos de dados internacionais para a caracterização das seqüências. A correta e acurada análise desses dados permite que eles se transformem em informação extremamente valiosa que possibilita sua aplicação em curto prazo. Tecnologias de informação permitem a interconexão de todos os grupos associados e, mais importante, permite que qualquer pesquisador, de qualquer país do mundo, tenha livre acesso às informações geradas no Projeto Genoma Humano do Câncer. A liberdade de troca de informações é uma garantia de rápida utilização do conhecimen-

to, o que viabiliza o retorno dos benefícios para a comunidade da maneira mais rápida possível.

No que diz respeito à *exemplaridade e efeitos emulador e indutor da aplicação da técnica ORESTES*, sua implantação tem surtido efeitos que ultrapassaram todas as expectativas de seus autores. O Projeto Genoma Humano do Câncer tornou possível, ao Brasil, ter produzido até agora 13,2% de todas as ESTs humanas disponíveis em bancos de dados públicos. Como foi dito, ele é o segundo maior projeto mundial, perdendo em número de dados apenas para o CGAP (*Cancer Genome Anatomy Project*), iniciado pelo menos três anos antes. Quando tumores humanos são avaliados em detalhe, verifica-se que o projeto gerou 64 mil dados derivados de tumores de mama (o dobro das informações geradas pelo resto do mundo junto). Também para tumores de estômago foi gerado o dobro de dados que todo o mundo gerou e, em relação a tumores de cabeça e pescoço, o projeto brasileiro gerou 4 vezes mais dados. A disponibilidade de uma tecnologia alternativa atraiu um investimento de US\$ 20 milhões para o Genoma do Câncer, tornando-o, além do maior projeto de pesquisa nacional, um dos que dispõem de mais recursos originários da iniciativa privada.

Em relação a *equipamentos utilizados no processo de criação e aplicação da tecnologia ORESTES, e ao grau de cooperação de instituições e agentes externos*, é necessário destacar, na criação, o uso de termocicladores, sistemas de eletroforese e sistemas de análise de dados do tipo WorkStations SUN e IBM, além de sistemas de comunicação por linhas dedicadas e de alta velocidade. Na aplicação da tecnologia, foram usados seqüenciadores de DNA de última geração, baseados na tecnologia capilar. Quanto à cooperação externa, obteve-se apoio financeiro de US\$ 10 milhões do Instituto Ludwig de Pesquisas sobre o Câncer. E no âmbito científico propriamente, o projeto teve cooperação externa na forma de valiosa participação de um comitê avaliador externo, que contou com os seguintes membros: Dr. Phillip Sharp (prêmio Nobel de Medicina pela descoberta dos Introns), Dr. John Sgouros (ex-coordenador de Bioinformática do Projeto Genoma de Levedura, Inglaterra), Dr. Marcelo Bento Soares (Universidade de Iowa e Projeto Genoma Humano, USA), Dr. Victor Jongeneel (Office for Information and Technology, Suíça).

Quanto à consolidação da tecnologia Orestes por meio de *patente da inovação e seu registro relativo à propriedade industrial*, foi requerido pedido de patente internacional junto à *World Intellectual Property Organization* #WO 00/31299 (02.06.2000). International Application Number: PCT/US99/27430



MIGUEL BOYAVAN



EDUARDO CESAR



EDUARDO CESAR

Avaliação externa: os cientistas (em sentido horário) Phillip Sharp, prêmio Nobel de Medicina pela descoberta dos introns, Marcelo Bento Soares, da Universidade de Iowa e membro do Projeto Genoma Humano dos Estados Unidos e John Sgouros, ex-coordenador de bioinformática do Projeto Genoma da Levedura, na Inglaterra, examinaram e deram seu aval à tecnologia Orestes e ao Projeto Genoma Humano do Câncer

ORESTES

UMA TECNOLOGIA BRASILEIRA PARA A DESCOBERTA DE GENES

EMMANUEL DIAS NETO e ANDREW J. G. SIMPSON

DNA, diversidade e evolução

Toda a rica diversidade dos 6 bilhões de habitantes do nosso planeta, com sua carga de contrastes e diferenças de corpos e mentes, tem como base as informações contidas no núcleo de cada uma de nossas células. A molécula-chave de toda essa informação é o DNA. Em última instância, toda a sutileza de nossa espécie, nossa arte, nossas conquistas científicas, nossas doenças e um grande número de fatores que influenciam nosso comportamento são determinados por um grupo relativamente pequeno de instruções contidas no nosso DNA, em nossos cerca de 80 mil genes. Com certeza não é exagero dizer que a descoberta dos genes humanos irá iluminar muitos dos mistérios de nossa vida e de nossa biologia.

O genoma humano é representado pela totalidade do material genético contido em uma célula. A fonte de nossa identidade, o nosso genoma completo, é preservada em cada uma dos nossos trilhões de células. Em um nível mais básico, nosso genoma é o nosso DNA, o ácido desoxirribonucleico, um polímero natural composto por apenas quatro tipos de nucleotídeos ou bases (A, C, G e T), repetidos bilhões de vezes. Na famosa dupla hélice de DNA, essas bases formam pares de uma maneira fixa e complementar. Os pares formados são A/T, T/A, C/G e G/C. A sequência e ordenação dessas bases formam a estrutura primária do DNA. Nos cromossomos, as hélices de DNA são enoveladas e empacotadas em estruturas mais complexas com alto nível de organização.

Essa estrutura organizacional é mantida com pequenas alterações em todas as espécies vivas e nosso genoma hoje é fruto de milhões de anos de evolução. Durante esses milhões de anos, o material genético foi sofrendo contínuas modificações selecionadas pelo ambiente, que permitiram o surgimento de todas as espécies, incluindo o *Homo sapiens*. Essas alterações foram sendo incorporadas ao nosso DNA através de diversos mecanismos. Trechos de DNA foram inseridos por meio de vírus, mutações ocorreram após contato com substâncias químicas ou raios ultravioleta, erros ocorridos durante processos de divisão celulares foram incorporados, genes foram desligados ou perdidos, e todas as alterações que resultaram em

algum ganho evolutivo para um indivíduo foram fixadas no seu genoma. Como resultado, nossa espécie tem hoje um genoma de 3,12 bilhões de bases. Neste vasto conteúdo do genoma estão espalhados os cerca de 50 a 80 mil genes humanos. Os genes encontram-se distribuídos ao longo de todo o genoma e seu conteúdo informacional está distribuído de uma maneira descontínua, como palavras separadas em sílabas. Na Biologia Molecular, cada “sílabas” é chamada de exon, espaçada de uma outra “sílabas” por uma região denominada íntron. Essa organização está representada na Figura 1. Se somarmos cada uma das “sílabas” de todos os genes iremos totalizar cerca de apenas 3% do conteúdo do genoma. Apenas esse pequeno percentual está diretamente comprometido com o funcionamento da maquinaria celular para a construção, manutenção funcional de um ser humano. As demais regiões do genoma (97%) são constituídas por DNA não informativo.

Projeto Genoma Humano

Para que possamos compreender ampla e totalmente a estrutura completa de funcionamento de nosso organismo, com todas as suas características fisiológicas, estruturais e funcionais, é absolutamente fundamental que tenhamos um conhecimento total das informações contidas em nossos genes. Esse conhecimento é adquirido pela leitura das bases nucleotídicas que compõem os genes (seqüenciamento) e determinação das regiões cromossômicas nas quais estes genes se encontram (mapeamento). O seqüenciamento e o mapeamento do DNA constituem as mais importantes atividades do Projeto Genoma Humano. Decifrar toda a informação genética de nosso genoma foi um desafio que começou a ser planejado ainda nos anos 80. Limitações técnicas foram sendo gradativamente superadas e a entrada de empresas privadas no setor levou a um aumento de competitividade e a uma tremenda redução no período inicialmente estimado para o término dessa etapa de decodificação.

Dessa maneira, no dia 26 de junho de 2000, a empresa privada Celera Genomics e a iniciativa pública do Projeto Genoma Humano (financiado principalmente pelo Department of Energy, National Institutes of Health e The Wellcome Trust) anunciaram em conjunto o término da primeira montagem do genoma humano, feito que foi comunicado pelo presidente norte-americano Bill Clinton e prontamente difundido em toda a imprensa mundial. Certamente nossa geração será referida pelas gerações futuras como sendo a responsável pela decodificação do genoma humano. Uma conquista ímpar que influenciará todos os aspectos da vida humana nas próximas gerações.



Figura 1

Estrutura básica da organização de um gene de organismos complexos.

As regiões espaçadoras e íntrons são representados pela cor azul, enquanto que os exons, ou regiões transcritas de um gene, são representados pela cor vermelha. A junção desses exons, transformados em moléculas de mRNA, forma os genes propriamente ditos, que equivalem a apenas 3% do nosso genoma.

A porção interna desses mRNAs contém a informação central para a síntese das proteínas. Nos projetos genoma os mRNAs são seqüenciados parcialmente, gerando as ESTs. A distribuição posicional dessas ESTs geradas com ORESTES e com técnicas tradicionais é apresentada.

Desafios da era pós-genoma

Após a primeira montagem do genoma humano, o projeto avança para a sua segunda fase, a interpretação do genoma codificado. A montagem e a leitura de toda a sequência de bases do genoma humano equivalem a fazer um grande “caça-palavras” com 3,12 bilhões de letras. Nessa próxima fase, o primeiro desafio será o de encontrar todos os genes humanos nesse longo “caça-palavras”. O desafio não é simples, pois os genes normalmente se encontram divididos em pedaços (exons) equivalentes às sílabas de uma palavra. Análises por computador mostraram-se limitadas no processo de identificar essas “sílabas”, sendo assim incapazes de identificar corretamente todos os genes. Atualmente apenas pouco mais de 10 mil genes humanos são completamente conhecidos e as estimativas do número total de genes são extremamente variadas, indo de 35 mil a 120 mil (Liang *et al.*, 2000; Ewing & Green, 2000; Crollius *et al.*, 2000).

Dessa maneira, para sermos capazes de determinar quais são e identificar corretamente cada um dos genes humanos, devemos buscar alternativas que ofereçam uma possibilidade real e segura. Para isso, nada melhor do que nos utilizarmos dos processos celulares que naturalmente refletem os genes expressos em cada uma de suas formas alternativas. Os 3% do nosso genoma, que contém todos os nossos genes, são responsáveis pela síntese das proteínas e durante esse processo são sintetizadas moléculas intermediárias denominadas mRNAs. Dessa forma, a obtenção e caracterização da sequência desses mRNAs representam um atalho seguro na determinação das sequências dos genes, permitindo economizar 97% do trabalho e dos investimentos. Uma varredura rápida dos genes expressos, usando o sequenciamento parcial dessas moléculas de mRNA, foi descrita por Adams *et al.*, 1991, e utilizada com sucesso por vários grupos mundiais trabalhando com diversos organismos. A técnica denominada *EST sequencing* (sequenciamento de “etiquetas de genes expressos”) permitiu um aumento dramático no número de genes humanos conhecidos (ainda que parcialmente) e a identificação de uma grande série de genes importantes em diversas patologias e processos celulares humanos. A técnica consiste em obter uma sequência parcial, normalmente derivada de uma das extremidades de um gene. Essa sequência, ainda que parcial, é suficiente para determinar se um gene é novo ou não, montando o perfil de genes ativos em um tecido. Atualmente, ESTs são usadas como base de uma série de projetos no mundo, inclusive por grandes grupos, tais como o CGAP (Cancer Genome Anatomy Project – www.ncbi.nlm.nih.gov/CGAP/).

Obstáculos das técnicas usuais de geração de ESTs

A utilização de ESTs na descoberta dos genes demonstrou ser de grande valor. No entanto, dois grandes obstáculos técnicos existem e fizeram com que vários projetos fossem descontinuados. O primeiro se refere a um fator ligado à distribuição dos diversos genes dentro das células. Enquanto alguns poucos genes são muito ativos, com milhares de mRNAs dentro de uma célula, a grande maioria dos genes tem um baixo nível de atividade, ou seja, temos poucas unidades funcionais representadas em cada célula (Tabela 1).

Para gerar as ESTs são feitas as chamadas bibliotecas de cDNA, que refletem a população dos genes ativos em uma célula. Sendo assim, os genes raros são pou-

co representados, e a sua descoberta torna-se cara, pois requer a geração de milhares de ESTs para que os genes raros tenham chances estatísticas de serem representados. Por outro lado, a alta representatividade dos genes abundantes leva a um seqüenciamento redundante dos mesmos, aumentando os gastos de um projeto e impedindo a obtenção de dados informativos referentes a novas seqüências.

Um segundo problema encontrado nessa abordagem se refere à limitação técnica de seqüenciamento de DNA ainda existente nos dias de hoje. Os genes humanos têm em média cerca de 4.000 bases a serem seqüenciadas. Os métodos atuais permitem, em um projeto de larga escala, seqüenciar apenas cerca de 500 bases. Sendo assim, apenas as extremidades dos genes são seqüenciadas quando essa estratégia é utilizada. A porção central normalmente não é representada. Com isso, a seqüência completa dos genes não é atingida, e a porção interna dos genes, aquela que contém a informação necessária para fazer as proteínas, não é determinada. Esse fato pode ser facilmente percebido por análises nos bancos de dados, onde é notada a pequena frequência de ESTs derivadas da porção central dos genes (Figura 2b).

Sendo assim, para que os frutos do Projeto Genoma Humano sejam totalmente aproveitados, tornam-se necessárias alternativas tecnológicas que permitam:

- 1) Ter acesso direto às moléculas correspondentes apenas aos genes, evitando os 97% não informativos do genoma.
- 2) Ter acesso à população de genes raros, permitindo um avanço na direção de descobrir os genes pouco expressos, evitando os genes abundantes já amplamente seqüenciados.
- 3) Gerar dados referentes à porção central e mais informativa desses genes, de modo a facilitar análises comparativas que permitem inferir sua função biológica, além de possibilitar montar a seqüência completa de todos os genes.

ORESTES: Alternativa tecnológica para a descoberta da porção central de genes humanos, com ênfase nos genes raros

De maneira geral, podemos dizer que uma biblioteca rica em informação deve conter um grande número de diferentes títulos. Isso também pode ser transposto para nossos genes. A diversidade de informações é proporcional ao número de genes. Se avaliarmos com cuidado a tabela 1, poderemos observar que a maior diversidade de informações é encontrada na categoria correspondente aos genes raros, onde estão mais de 90% dos genes humanos. Dessa maneira, para favorecermos os genes raros, uma alternativa seria dispormos de uma tecnologia que favorecesse a diversidade e riqueza de informações em vez da abundância de seqüências de baixo conteúdo informacional. Essa é exatamente uma das idéias centrais por detrás de ORESTES.

ORESTES leva em consideração dois fatores:

- 1) A descoberta de genes raros trata da descoberta de uma população de genes bastante complexa em termos de diversidade de seqüência de DNA. Essa população com alta diversidade ou grande quantidade de informação pode ser preferencialmente determinada se esse fator for levado em consideração no desenho experimental. Sendo assim, o fator nível de expressão (importante para os genes

TABELA 1

Classe	Número de diferentes genes/célula	Representatividade (cópias de cada gene por célula)
Abundantes	<10	12.000
Intermediários	500	300
Raros	11.000	10

Fonte: Huang *et al.*, 1999

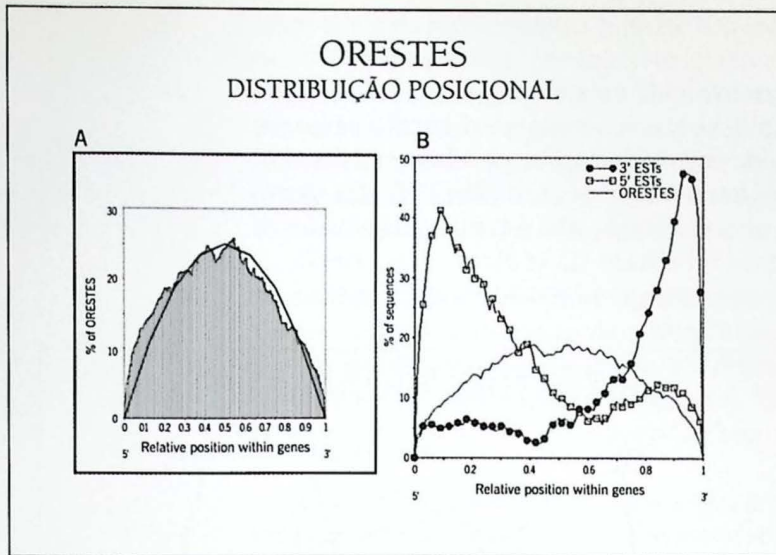


Figura 2

Distribuição posicional de ORESTES. 2a – A linha representa a distribuição teórica de ORESTES prevista de acordo com a fórmula discutida no texto. As barras representam a distribuição real observada após 10 mil ORESTES derivadas de tumores de mama. 2b – Os dados observados em ORESTES a partir da distribuição posicional são comparados com os dados observados para sequências derivadas de outros projetos genoma a partir de cada uma das duas extremidades. Podemos ver a complementaridade das duas abordagens.

abundantes) não deve ser determinante, e a complexidade informacional das moléculas desse grupo deveria ser favorecida.

2) Para atingir a parte central das moléculas, não poderíamos iniciar a reação, como os demais projetos mundiais, utilizando moléculas iniciadoras ancoradas nas extremidades dos genes. Deveríamos utilizar uma grande série de moléculas distintas que se liguem em porções mais internas dos diversos genes.

Sendo assim, toda a idéia por trás da técnica ORESTES se baseia em complementaridade de moléculas. Uma complementaridade que permite a efetivação das reações de uma maneira alternativa. Essa complementaridade entre moléculas se dá sempre onde a diversidade de informações é maior: sendo assim, o elevado nível de diversidade na população dos genes mais raros faz com que estes, seri-

am teoricamente os alvos principais dessa técnica, sejam favorecidos. O fato de não usarmos moléculas complementares à extremidade dos genes, o que resultaria em sequências das extremidades, faz com que tenhamos ESTs altamente informativas vindas da porção central e geralmente pouco representada dos genes. No jargão da Biologia Molecular, esta região é denominada de Moldura Aberta de Leitura, ou *Open Reading Frame*. Surgiu então a técnica ORESTES (*Open Reading Frame ESTs*).

Para obtermos a representação de um determinado ponto dentro de um gene, é necessário termos moléculas complementares que se liguem dos dois lados desse ponto. A chance de esses dois eventos de complementaridade ocorrerem é proporcional ao comprimento da sequência de cada um dos lados desse ponto. Sendo assim, se o tamanho de um gene é 1, e a distância do ponto determinado até uma das extremidades é tida como S, a probabilidade de termos uma interação para cada um dos eventos será S e 1-S. Como consequência, a probabilidade de representação desse ponto será $S(1-S)$. Em termos práticos, a chance de uma ORESTES conter exatamente a porção central de um dado gene é igual a $0,5 \times 0,5 = 0,25$, enquanto que a chance de representar um ponto relativo aos 10% finais ou iniciais é igual a $0,1 \times 0,9 = 0,09$, etc. Ao plotarmos os valores dessa fórmula simples, para as diversas regiões de um gene, teremos a linha sólida observada na Figura 2a. As barras nesta mesma figura representam os dados reais resultantes das primeiras 10 mil ORESTES derivadas de tumores de mama. A similaridade das duas representações permite concluir que ORESTES funciona como previsto, sendo a única técnica disponível para gerar dados em larga escala referentes à porção central dos genes.

ORESTES

Esta técnica começou a ser desenvolvida no ano de 1993, com genes derivados do parasita humano *Schistosoma mansoni*, o agente causador da esquistossomose. Cerca de mil ESTs foram geradas e analisadas, permitindo avaliar algumas características de ORESTES. Os genes abundantes, já conhecidos nesse parasita, não foram representados, e um alto percentual de seqüências inéditas foi obtido (Dias Neto, *et al.*, 1997).

Posteriormente, passamos a avaliar com maior minúcia as características de ORESTES. Para isso se fazia necessária a sua avaliação em um modelo já bastante estudado, onde pudéssemos ter grande número de genes conhecidos – verificando a posição exata de nossas ESTs – e avaliar em maior escala a capacidade de obter genes raros. Partimos então para gerar um número similar de dados derivados de um tumor cólon-retal humano, tecido bastante estudado e que oferecia uma real avaliação de ORESTES. Para nossa satisfação, mesmo trabalhando com o genoma mais estudado, continuamos com um alto percentual de dados novos referentes a seqüências humanas jamais descobertas por outros grupos mundiais. Diante desses dados, o Instituto Ludwig decidiu financiar em US\$ 250 mil um projeto piloto para gerar 10 mil ORESTES derivadas de tumores de mama. Para produzir essa quantidade de dados em um curto espaço de tempo, fomos auxiliados por alguns grupos do Estado de São Paulo que fizeram o seqüenciamento dos clones que geramos. Foi também montado um Laboratório de Bioinformática, capaz de lidar com a grande quantidade de dados gerados, permitindo uma análise minuciosa. Mais uma vez os dados se mostraram surpreendentes. As 10 mil ORESTES confirmaram as predições de distribuição central e mostraram uma capacidade de refletir genes raros pelo menos equivalente à das melhores tecnologias do mundo. Esses dados foram reunidos na forma de um artigo científico publicado pela revista da Academia Nacional de Ciências dos Estados Unidos (PNAS), após ser recomendado pelo Prêmio Nobel Dr. Phillip Sharp (Massachusetts Institute of Technology – MIT). As figuras desse trabalho (apresentadas também nas figuras 2a e 2b) demonstram que ORESTES são realmente derivadas da porção codificadora dos genes. As análises também demonstram com clareza que os genes raros são favorecidos e mais facilmente representados.

A alta qualidade dos dados fez com que mais uma vez recebêssemos um apoio inestimável do Instituto Ludwig, que desta vez foi somado ao apoio da FAPESP. O total financiado, a “fundo perdido”, foi de US\$ 10 milhões, compartilhado igualmente entre o Ludwig e a FAPESP, surgindo então o Projeto Genoma Humano do Câncer Ludwig/FAPESP (Human Cancer Genome Project - HCGP). No dia 26 de março de 1999 o HCGP foi lançado pelas diretorias do Instituto

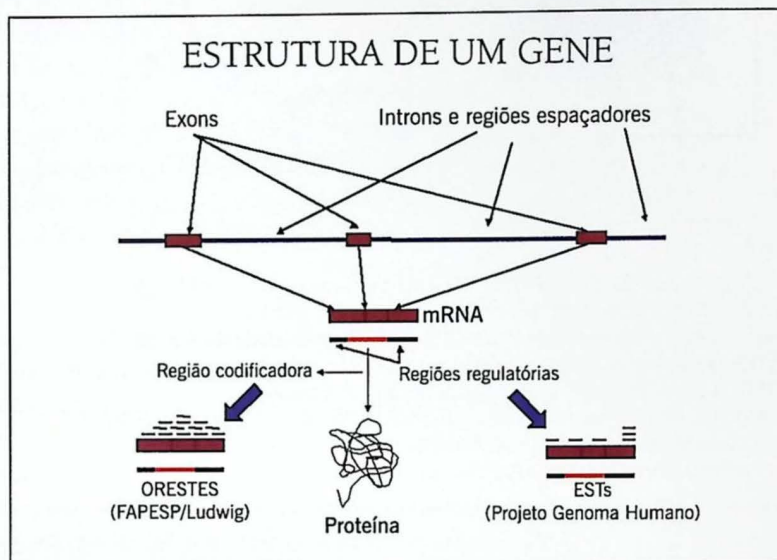


Figura 3

Ilustração representando as fases iniciais de organização do Projeto Genoma do Câncer Humano – Ludwig/FAPESP.

Ludwig e pela FAPESP. Uma rede de 33 laboratórios foi formada, envolvendo cerca de 200 pesquisadores. Os objetivos então delimitados foram extremamente ousados – descobrir os genes envolvidos na formação de tumores de grande importância em nosso país. A organização do projeto é baseada em uma coordenação geral, a cargo de Andrew J. G. Simpson. Os tecidos a serem usados no estudo são geralmente microdissecados pelo Dr. Fernando Soares, chefe do Depto de Patologia do Hospital do Câncer. Esse material é tratado pelas coordenações específicas, a saber:

Coordenador de RNA – Luis Fernando Lima Reis

Coordenador de Bibliotecas de cDNA – Emmanuel Dias Neto

Coordenadores de Sequenciamento – Fernando Costa (UNICAMP), Marco Antônio Zago (USP-Ribeirão Preto), Maria A. Nagai (Fac. Medicina da USP), Marcelo Briones (Unifesp) e Sérgio Verjovski (Inst. Química – USP).

Ao redor desses centros temos os laboratórios de sequenciamento espalhados pelas cidades anteriores e também por Botucatu, Araraquara e São José dos Campos. A organização, desde a obtenção dos tecidos até a geração de ORESTES, é mostrada na Figura 3.

A principal instituição mundial de pesquisas sobre o câncer é o NCI (National Cancer Institute), um dos principais Institutos do NIH (National Health Institute, USA). A nosso pedido, o seu diretor, Dr. Richard Klausner, emitiu seu parecer sobre o nosso projeto. Este parecer foi uma importante fonte de reconhecimento de nosso trabalho, servindo como estímulo para continuarmos. Seguem as palavras do Dr. Klausner:

"I am impressed and pleased with the enormous effort of the Brazilian cancer genome project. The production of this amount of valuable information and the generous and open sharing with the international scientific community is a model of the spirit of the international fight against cancer. This fight can only be waged with the full participation of the talents of people throughout the world and the free and open exchange of ideas and information".

Projeto Genoma Humano do Câncer Ludwig/FAPESP - Investir em vez de esperar

No momento do lançamento do HCGP, havia apenas um grande projeto mundial centrado na geração de ESTs derivadas de tumores, o CGAP (Cancer Genome Anatomy Project). Este projeto, lançado em 1997 com financiamento do Instituto Nacional do Câncer nos Estados Unidos, tinha como objetivos principais a determinação dos genes envolvidos nos tumores de maior importância epidemiológica nos Estados Unidos, tais como mama, pulmão, próstata, cólon. Questões políticas fazem com que o Instituto Nacional de Saúde dos Estados Unidos priorize a geração de dados relevantes aos tumores mais importantes em seu próprio país.

No entanto, sabemos que os tipos, a prevalência e a mortalidade por câncer são diferentes entre países de Primeiro e de Terceiro Mundo. Questões raciais, geográficas, culturais e socioeconômicas atuam de maneira importante, gerando índices bastante diversificados entre diferentes países. Dados recentes sobre incidência e mortalidade de câncer em todo o mundo foram publicados por Pisani *et al.*, 1999 e por Parkin *et al.*, 1999). De maneira geral, vemos que a incidência e a mortalidade da maioria dos tipos de câncer vêm crescendo de maneira brutal nos últimos anos. Estima-se para o ano de 2010 um aumento de 52,4% nas mortes por câncer, totalizando cerca de 7.900.000 mortes neste ano. O qua-

dro é ruim em todo o mundo, e ainda pior nos países em desenvolvimento, onde o crescimento percentual de mortes por câncer tende a dobrar em relação ao crescimento em países desenvolvidos. Os dados podem ser mais bem visualizados na tabela 2:

No trabalho de Parkin *et al.*, 1999, vemos que tumores de colo de útero ainda apresentam uma alta incidência em países de Terceiro Mundo, especialmente América Central (31,2 casos/100.000 habitantes) e América do Sul (27 casos/100.000 habitantes). Em países desenvolvidos essa taxa é inferior a 14/100.000. Os índices são ainda mais preo-

cupantes para tumores de nasofaringe (agrupados em nossos estudos como tumores de cabeça/pescoço), onde a incidência comparada, normalizada para 100.000 habitantes do sexo masculino, entre países desenvolvidos e em desenvolvimento é de 35,7 contra 4,1 respectivamente. O quadro não melhora quando tratamos da mortalidade causada por câncer (Pisani *et al.*, 1999). Se considerarmos o mundo todo, temos anualmente 197 mil mortes apenas por câncer de cavidade oral. Setenta e quatro por cento dessas ocorrem em países em desenvolvimento. Das 190 mil mortes anuais por câncer de colo de útero, 78% ocorrem em países em desenvolvimento. Dessa maneira, para descobrirmos os genes ligados a tumores importantes em nosso país, e que ficaram de fora das prioridades do programa americano, tínhamos duas opções: esperar ou investir.

O HCGP foi fruto da coragem e ousadia de investir. Investir na descoberta de genes importantes para tumores que atingem países de Terceiro Mundo, com a coragem de apostar em uma tecnologia alternativa totalmente desenvolvida no Brasil e com a ousadia de participar, sem convite, do seleto clube de países desenvolvidos ligados ao seqüenciamento de genes humanos.

Não temos dúvidas de que o grande sucesso desse projeto se deve ao comprometimento do grupo de cerca de 200 pesquisadores brasileiros de altíssima capacidade, distribuídos em 33 laboratórios ao redor do Estado de São Paulo. Além disso, o espírito humanitário de pacientes que permitiram o estudo de material tumoral e a colaboração de diversos grupos de cirurgia foram de um valor inestimável.

Resultados atuais do HCGP

Temos a satisfação de termos alcançado, cerca de oito meses antes do prazo, a meta de meio milhão de seqüências humanas, todas produzidas pelo método ORESTES. Desse total, cerca de 30% representam fragmentos novos, jamais obtidos em nenhum outro projeto. Além desses 30%, outras ORESTES permitiram a descoberta de novas formas de expressão de genes humanos já conhecidos, ou mostraram pela primeira vez a determinação de tecidos onde alguns genes são ativos. Nossos dados permitiram descobrir genes mesmo em fragmentos do genoma humano exaustivamente estudados por outros grupos mundiais. Um exemplo pode ser dado pela análise do cromossomo 22 humano, seqüenciado pelo Sanger Centre na Inglaterra. Quando os dados de ORESTES foram usados no mapeamento de genes nesse cromossomo, 144 novos genes foram descobertos (Souza *et al.*, no prelo).

Os números obtidos falam por si mesmos. O HCGP já é o segundo maior projeto de geração de ESTs humanas, ficando atrás apenas do CGAP/NIH. Para di-

TABELA 2

	Ano 2000	A	Ano 2010	B
Mundo	6.366.000	22,8%	7.899.000	52,4%
Países desenvolvidos	2.669.000	14,5%	3.019.000	29,5%
Países em desenvolvimento	3.697.000	29,6%	4.880.000	71,1%

A: Crescimento percentual no ano de 2000 em relação a 1990.

B: Crescimento percentual no ano de 2010 em relação a 1990.

Fonte: Pisani *et al.*, 1999. Estimates of the worldwide mortality from 25 cancers in 1990. International Journal of Cancer 83:18-29.

TABELA 3

TUMOR	América do Sul (clima temperado)		América do Sul (clima tropical)		ESTs geradas (x1000)	
	Incidência	Mortalidade*	Incidência	Mortalidade*	CGAP	HCGP1
Boca/Faringe	8,4	1,1 (11)	10,9	7,9 (8)	3,5	147,6
Estômago	19,6	6,2 (2)	23,3	30,8 (1)	22,0	38,8
Mama	76,1	5,6 (4)	28,6	15,1 (4)	15,5	70,9
Próstata	22,9	3,5 (6)	16,8	11,9 (6)	14,7	19,2
Cólon	26,8	5,9 (3)	9,6	14,6 (5)	53,5	62,5
Pulmão	53,3	10,2 (1)	15,1	22,4 (2)	61,8	5,4
Colo de útero	28,2	2,6 (8)	24,5	15,6 (3)	0,9	21,4
TOTAL	235,3	35,1	128,8	105,8	171,9	365,8

Taxa de Incidência é apresentada para grupos de 100.000 indivíduos. Mortalidade: número total de mortes anuais causadas por determinado tipo de câncer (em milhares). (Fonte: Pisani et al., 1999). Estimates of the worldwide incidence from 25 cancers in 1990. *International Journal of Cancer* 83:18-29; Parkin et al., 1999. Estimates of the worldwide mortality from 25 cancers in 1990. *International Journal of Cancer* 80:827-841.

* Os números entre parênteses indicam a posição de um tumor específico em termos de mortalidade;

(1) Atualizado em 20 de setembro de 2000.

versos tumores humanos já somos o primeiro grupo mundial em termos de seqüências de genes. Quando avaliamos a nossa contribuição ao estudo de tumores específicos, constatamos que nosso grupo lidera em número de ESTs geradas a partir dos tumores de mama, cabeça e pescoço, estômago, próstata, cólon intestinal, testículo e próstata. No total, temos hoje (dia 02 de outubro de 2000) 701.471 ORESTES geradas. Dentre estas, cerca de 30% representam seqüências humanas inéditas.

A produtividade de nosso projeto em relação a al-

guns dos principais tipos de tumores é apresentada junto à incidência e mortalidade dos mesmos em diferentes regiões sul-americanas (Tabela 3).

O impacto desses tumores é dramático, causando apenas na América do Sul cerca de 141 mil das 764,7 mil mortes anuais por câncer.

As ORESTES produzidas, após passarem por nosso controle de qualidade, são depositadas em bancos de dados públicos, de acesso gratuito, permitindo o seu uso por toda a comunidade científica internacional (www.ncbi.nlm.nih.gov).

Esse grande número de dados de alta qualidade, proporcionados pelo uso de ORESTES, fez com que o financiamento inicial fosse dobrado de US\$ 10 milhões para US\$ 20 milhões, totalmente voltados para a pesquisa. Com isso deveremos dispor de um parque instalado de 11 seqüenciadores automáticos de tecnologia capilar, o que permitirá uma significativa aceleração do projeto. As metas foram reavaliadas e até o final de 2000 temos o compromisso de gerar 1 milhão de ORESTES.

Perspectivas finais

Ao seu final, o Projeto Genoma Humano deverá gerar um imenso e detalhado atlas da biologia humana. O HCGP, através da técnica de ORESTES, deverá permitir o mapeamento de um alto número de genes humanos dentro desse atlas. O atlas completo do genoma humano fornecerá informações em nível de seqüências de DNA que permitirão diagnosticar, estudar, tratar, curar e prevenir a maioria das doenças humanas. A pesquisa biomédica ficará dividida entre o que foi feito antes e o que será feito após o PGH. Novas drogas poderão ser desenhadas, de modo a agir apenas em vias específicas de agentes infecciosos, outras drogas poderão ser desenhadas de modo a suprir a função de genes defeituosos, podendo ser usadas no tratamento e prevenção de doenças como o câncer, diabetes e doenças cardíacas. Não apenas os médicos deverão estar prontos para lidar com essa nova tecnologia, mas toda a sociedade deverá estar preparada para lidar com essa nova perspectiva.

O PGH vem repleto de expectativas, mas certamente com uma carga ainda maior de implicações sociais. Os dados gerados correm o risco de serem mal utilizados e a possibilidade de invasão de privacidade individual deve ser seriamente considerada. O PGH não promete curar o câncer, corrigir a esquizofrenia ou o mal de Alzheimer; não promete também definir as mudanças moleculares precisas que ocorrem na transformação de um embrião até um adulto, ou explicar comportamentos agressivos anti-sociais. No entanto, temos a certeza de que se algum dia conseguirmos descobrir os mistérios da carcinogênese, se algum dia compreendermos a bioquímica dos distúrbios neurológicos e se realmente tivermos a pretensão de compreender os processos de crescimento e desenvolvimento, é necessário que tenhamos antes de tudo um quadro completo do genoma humano. Nesse momento o "código dos códigos" estará realmente decifrado. O grande desafio será o de utilizarmos essas informações com consciência e responsabilidade.

Bibliografia:

- Crollius, HR, Jaillon, O, Bernot, J, DaSilva, C, Bouneau, L, Fischer, C, Fizames, C, Wincker, P, Brottier, P, Quetier, F, Saurin, W & Weissenbach, J. 2000. Estimate of human gene number provided by genome-wide analysis using Tetradon nigroviridis DNA sequence. *Nature Genetics* 25: 235-238.
- Dias Neto, E, Harrop, R, Corrêa-Oliveira, R, Wilson, RA, Pena, SDJ & Simpson AJG. 1997. Minilibraries constructed from cDNA generated by arbitrarily primed RT-PCR: an alternative to normalized libraries for the efficient generation of ESTs from nanogram quantities of mRNA. *Gene* 186: 135-142.
- Dias-Neto, E, Corrêa, RG, Verjovski-Almeida, S, Briones, MRS, Nagai, MA, Silva Jr, W, Zago, MA, Bordin, S, Costa, FF, Goldman, GH, Carvalho, AF, Matsukuma, A, Baia, GS, Simpson, DH, Brunstein, A, Oliveira, PSL, Bucher, P, Jongeneel, CV, O'Hare, M, Soares, F, Brentani, RR, Reis, LFL, Souza, SJ & Simpson, AJG. 2000. Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proceedings of the National Academies of Sciences (USA)* 97: 3491-3496.
- Ewing, B & Green, P. 2000. Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genetics* 25: 232-234.
- Huang, GM, Ng, W, Farkas, J, He, L, Liang, HA, Gordon, D, Yu, J & Hood, L. Prostate Cancer expression profiling by cDNA sequencing analysis. 1999. *Genomics* 59: 178-186.
- Liang, F, Holt, I, Perte, G, Karamycheva, S, Salzberg, SL & Quackenbush, J. 2000. Gene Index analysis of the human genome estimates approximately 120,000 genes. *Nature Genetics* 25: 239-240.
- Parkin, DM, Pisani, P & Ferlay, J. 1999. Estimates of the worldwide mortality from 25 cancers in 1990. *International Journal of Cancer* 80:827-841.
- Pisani, P, Parkin, DM, Bray, F & Ferlay, J. 1999. Estimates of the worldwide incidence from 25 cancers in 1990. *International Journal of Cancer* 83:18-29.
- Souza, SJ *et al*; 2000. Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. *Proceedings of the National Academies of Sciences* (no prelo).

