

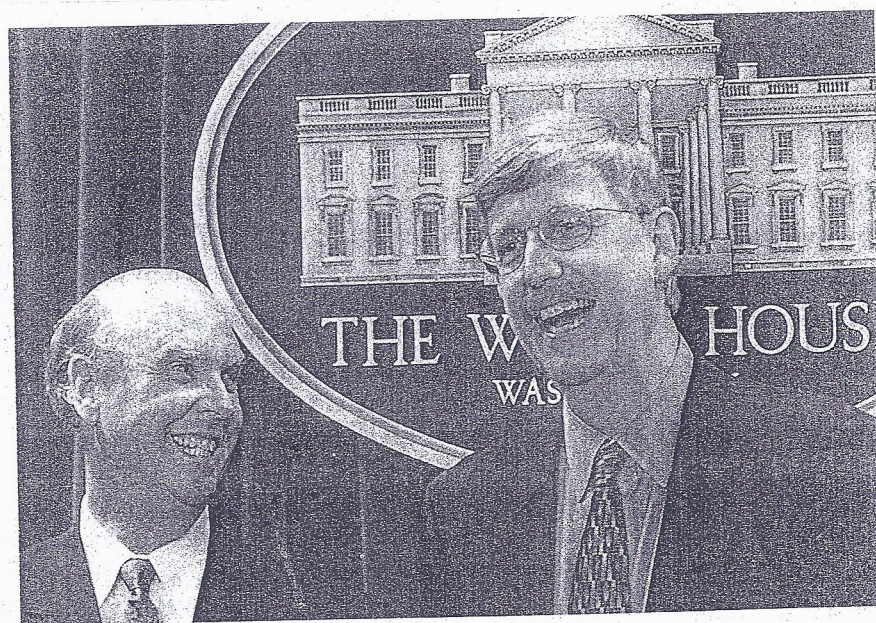
Micklos, Freyer & Crotty (2005) 2a Ed

CAPÍTULO 6

Métodos Modernos de Análise de Genomas Completos

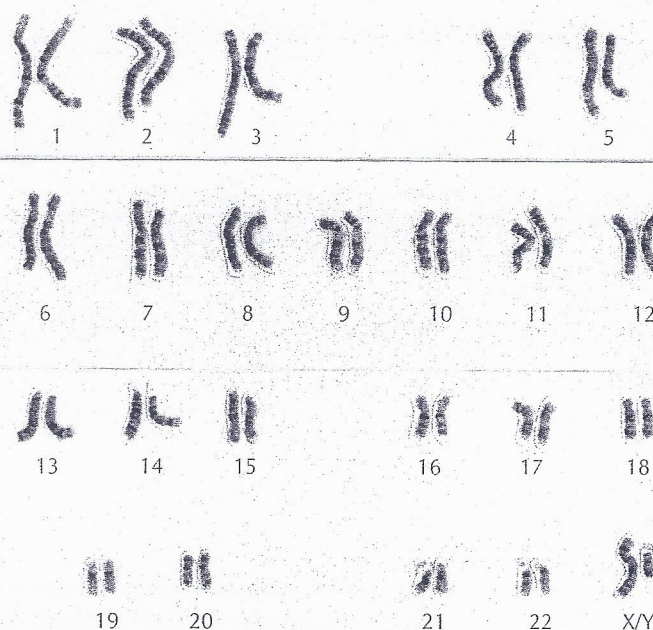
A publicação da primeira versão da sequência de DNA do genoma humano, em fevereiro de 2001, caracterizou-se como o auge da revolução biológica, que teve início meio século atrás com a descoberta da estrutura do DNA por Watson e Crick. O Projeto Genoma Humano é o experimento mais ambicioso da história das ciências biológicas. A grandiosidade dessa tarefa pode ser vislumbrada quando se considera a quantidade de informações necessárias para codificar e perpetuar a espécie humana.

O genoma é o conjunto completo das qualidades genéticas de um organismo – a sequência completa de nucleotídeos do DNA compreendida em todos os seus cromossomos. O genoma humano está contido nos 23 pares de grandes cromossomos presentes no núcleo de cada célula humana – 22 pares de autossomos e um par de cromossomos sexuais (X e Y) – *mais* um pequeno cromossomo presente na mitocôndria. Esse conjunto característico de cromossomos que define as características de cada indivíduo humano é reunido no momento da concepção. Um membro de cada par de cromossomos nucleares (além do cromossomo mitocondrial) é fornecido pelo óvulo materno, enquanto o espermatozóide paterno fornece o outro membro de cada par.



J. Craig Venter e Francis Collins

Coordenadores dos projetos de seqüenciamento das iniciativas privada e pública anunciando o término da seqüência inicial do genoma humano aos jornalistas. (Foto cortesia da A/P WIDE WORLD PHOTOS, Rick Browmer.)



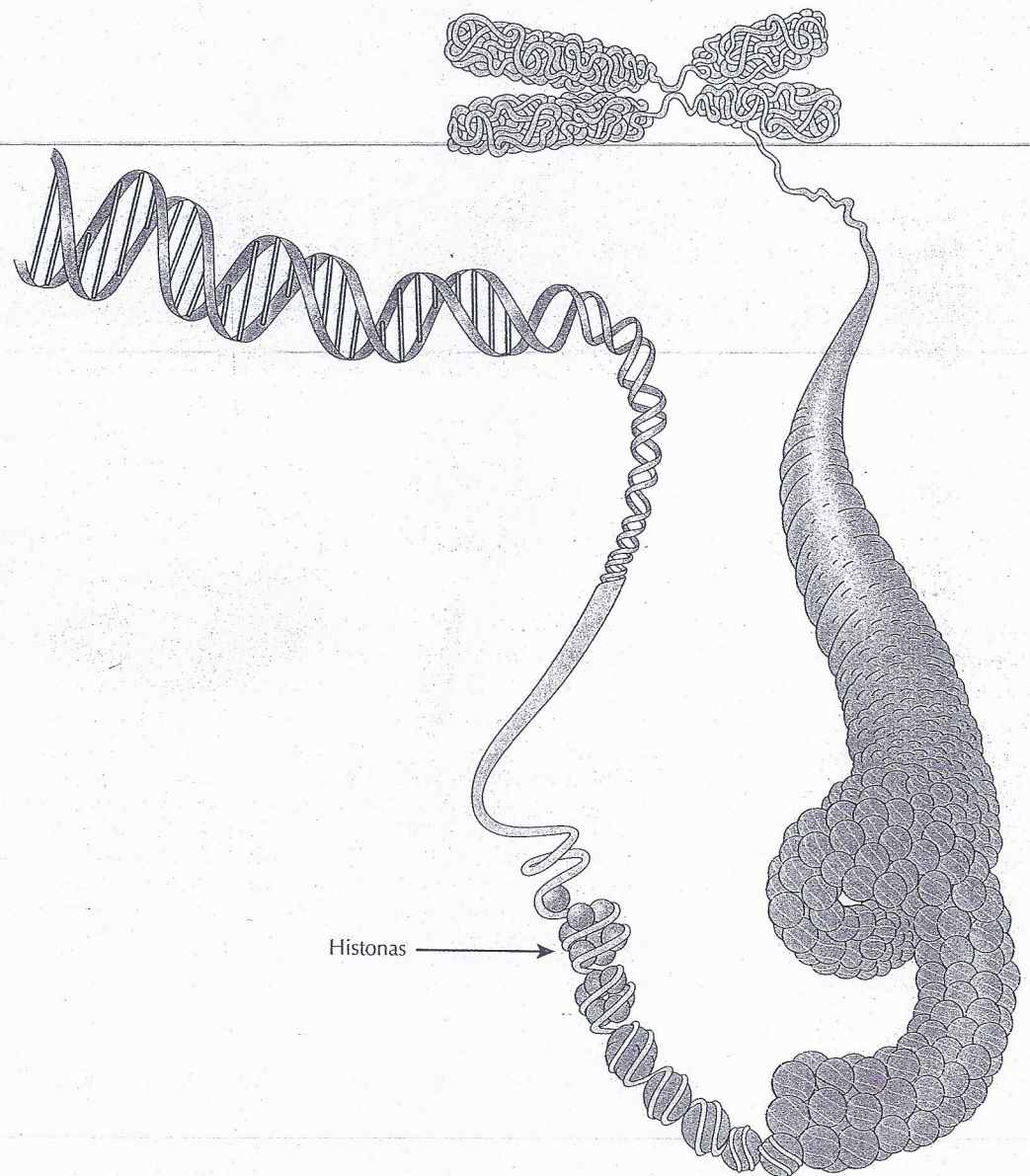
Cariótipo humano

A sequência do genoma humano consiste de 22 cromossomos autossômicos, dos cromossomos X e Y e do DNA mitocondrial. O DNA mitocondrial é muito pequeno comparativamente aos cromossomos, não sendo possível mostrá-lo nesta figura. (Reimpressa, com permissão, de Witherly J.L., Perry G.P. e Leja D.L. 2001. *An A to Z of DNA science*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.)

Excetuando-se o par X/Y, que ocorre em machos, cada par de cromossomos apresenta o mesmo arranjo de genes. Assim, com o objetivo de simplificação, os biólogos preocupam-se apenas com o genoma haplóide constituído de um único grupo de 25 cromossomos: os 22 autossomos, os cromossomos X e Y e o cromossomo mitocondrial. Cada cromossomo é um pacote complexo que contém uma fita única e contínua de DNA. A estrutura geral do cromossomo só é possível devido a moléculas protéicas (histonas) que formam uma armação em torno da qual a fita de DNA se enrola, como uma linha em um carretel. Dessa forma, pode-se pensar no genoma humano como sendo constituído por um conjunto de moléculas de DNA bastante longas, cada uma correspondendo a um dos cromossomos. As 25 moléculas de DNA do genoma humano haplóide contêm aproximadamente 3,1 bilhões ($3,1 \times 10^9$) nucleotídeos.

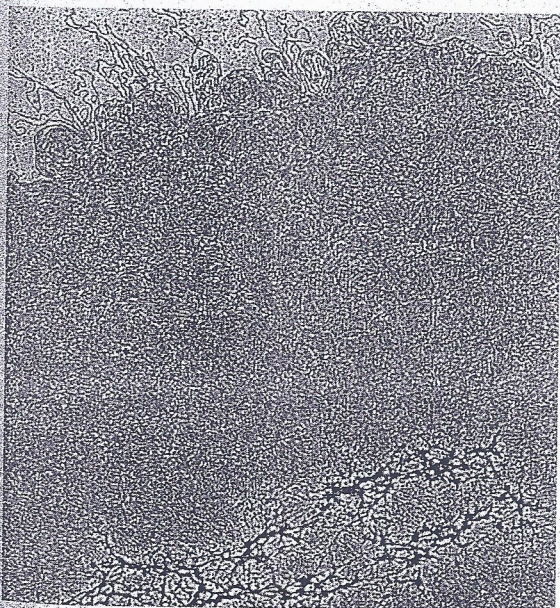
A magnitude do Projeto Genoma Humano transparece quando se imagina um "livro genômico da vida", no qual cada nucleotídeo é representado por uma única letra (A, T, C ou G). Usando a tipografia desse texto como exemplo – sem espaços, pontuação ou figuras –, cada página poderia conter 3.675 nucleotídeos. Sob esse padrão tipográfico, a sequência completa do genoma humano preencheria 843.500 páginas. Se tais páginas fossem organizadas em livros como o que você está lendo (de aproximadamente 600 páginas), o livro genômico da vida estaria composto por um conjunto de 1.406 volumes. Considerando-se 3,5 cm por volume, esse conjunto alcançaria 49,21 metros de comprimento, ou seja, a altura de um prédio de 15 andares.

Apesar de o Projeto Genoma Humano ter sido inicialmente concebido sob a óptica de um esforço puramente acadêmico, rapidamente o setor de negócios viu seu potencial latente. O projeto colocou um consórcio com financiamento público dirigido por cinco dos mais importantes centros de pesquisa acadêmica dos Estados Unidos e da Inglaterra contra uma nova companhia, a Celera, empresa criada a partir da principal produtora de aparelhos de sequenciamento automático. Os diretores das empreitadas pública e privada abriram fogo contra seus rivais – inclusive com acusações de que muitos dos participantes desses projetos seriam representantes das forças "do bem" ou "do mal". A competição para o sequenciamento do genoma hu-



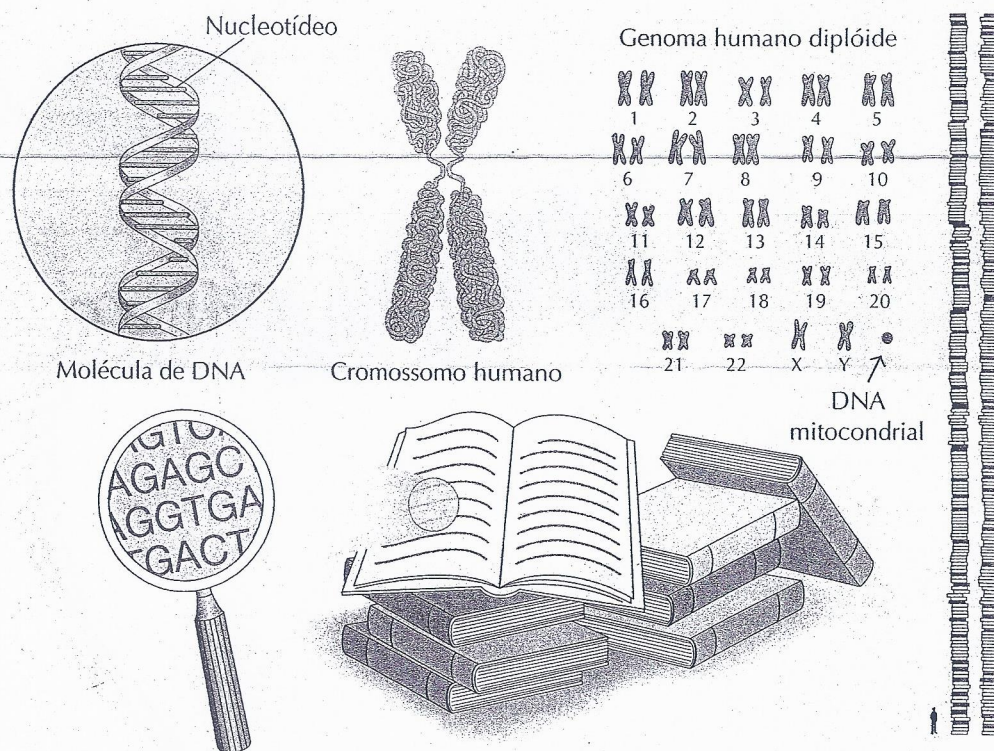
Cromatina

Um enrolamento complexo em torno das proteínas histonas permite que os aproximadamente 2 metros de DNA existentes em cada célula humana sejam compactados sob a forma dos 46 cromossomos nucleares.



Microfotografia eletrônica de um cromossomo depletado de histonas

A armação protéica completa toma a forma de um cromossomo com DNA escapando para todas as direções. (Reimpressa, com permissão, de Moran L.A., Scrimgeour K.G., Horton H.R., Ochs R.S. e Rawl J.D. 1994. *Biochemistry*. Neil Patterson Publishers/Prentice Hall, Englewood Cliffs, Nova Jérsei. Cortesia de U.K. Laemmli.)



O livro genômico da vida

Se cada nucleotídeo de um genoma humano fosse considerado como uma única letra deste livro e se cada sequência fosse escrita sem espaços, pontuação, ilustrações ou tabelas, 3.675 nucleotídeos completariam uma página. Dessa forma, a sequência de DNA do genoma haplóide humano (os 22 autossomos, os cromossomos X e Y e o cromossomo mitocondrial) preencheria 843.500 páginas. Se organizadas em volumes de tamanho aproximado igual a este livro, o livro genômico da vida formaria uma pilha de 49,21 metros de altura. Aproximadamente o dobro dessa quantidade seria necessário para representar a informação genética de um genoma diplóide completo.

mano tomou, então, o aspecto de uma nova copa americana de iatismo – times influenciados por diferentes mentalidades, usando diferentes metodologias, cada um apostando que seria o primeiro a “cantar vitória” sobre a natureza.

No período anterior à disponibilização do primeiro rascunho do genoma humano, a maioria dos cientistas estimava que o mesmo deveria conter 100.000 genes, enquanto algumas estimativas alcançavam a casa dos 150.000 genes. Assim, foi uma surpresa quase que geral o anúncio de dois grupos independentes que declararam que o genoma continha aproximadamente 35.000 genes. Tal fato levou à seguinte questão: como seres humanos conseguem lidar com sua enorme complexidade possuindo menos que o dobro do número de genes de um nemátódeo microscópico (o qual possui 19.099 genes). Desde esse anúncio inicial, alguns estudos têm modificado a estimativa de genes do genoma humano para ~50.000. As estimativas da massa total de sequências codificadoras de proteínas têm sido também reduzidas de 5% para menos de 1,5% do conteúdo do genoma. Apesar do conhecimento da extensão da porção não-codificadora do genoma humano ter aumentado durante os últimos 30 anos, o fato que ~99% do mesmo não-codificam nenhum gene permanece, para muitos, uma questão indecifrável.

O público em geral e grande parte dos cientistas estão interessados nos genes revelados pelo Projeto Genoma Humano. A identificação e o conhecimento de todos os genes do homem será um enorme passo em direção à compreensão de como nossas células funcionam e do que realmente está ocorrendo de errado em diversas doenças.

Até o momento, só foi possível identificar o gene causal de uma pequena fração das vários milhares de doenças genéticas conhecidas. O mapeamento e o sequenciamento desses genes de doenças facilitará o diagnóstico e os estudos bioquímicos de-

Tamanho dos genomas seqüenciados

Espécies	Pares de bases	Número de genes
Genomas completamente seqüenciados		
<i>Mycoplasma genitalium</i> (bactéria)	578.000	400
<i>Haemophilus influenzae</i> (bactéria)	1.830.138	1.738
<i>Escherichia coli</i> (bactéria)	4.639.221	4.377
<i>Saccharomyces cerevisiae</i> (levedura)	12×10^6	5.885
<i>Caenorhabditis elegans</i> (nematódeo)	$95,5 \times 10^6$	19.820
<i>Drosophila melanogaster</i> (mosca-das-frutas)	$1,8 \times 10^8$	13.601
<i>Arabidopsis thaliana</i> (planta)	$1,17 \times 10^8$	25.498
Genomas seqüenciados mas número total de genes ainda desconhecido*		
<i>Mus musculus</i> (camundongo)	3×10^9	30.000
<i>Homo sapiens</i> (ser humano)	$3,3 \times 10^9$	30.000-50.000
<i>Oryza sativa</i> (arroz)	$4,3 \times 10^8$	30.000-63.000

*O número total de genes está estimado.

talhados de seus produtos protéicos poderão levar ao desenvolvimento de melhores terapias. Sendo assim, tornar a vida humana mais saudável e, presumivelmente, mais agradável é o objetivo prático do Projeto Genoma Humano. No entanto, a extensão do DNA não-codante sugere que, assim como algo tangível encontra-se nas entrelinhas de um poema, algo de significativo pode se esconder em nosso genoma sob a forma do "lixo" não codante que existe entre nossos genes.

Este capítulo retrata o desenvolvimento do Projeto Genoma Humano, desde uma idéia controversa até o desenvolvimento de um esforço colaborativo mundial envolvendo milhares de cientistas. Será feita, então, uma revisão dos métodos laboratoriais usados na determinação e na organização das seqüências de DNA, dos métodos computacionais desenvolvidos para a identificação e a comparação de genes e das novas tecnologias de microarranjos que possibilitam a análise de milhares de genes simultaneamente, concluindo com uma discussão sobre a estrutura básica do genoma humano, como se lida com relativamente tão poucos genes, enfocando-se as razões de apenas uma fração tão pequena de nosso DNA codificar proteínas.

A GÊNESE DO PROJETO

A possibilidade de desenvolvimento de um esforço coordenado para o seqüenciamento do genoma humano foi levantada pela primeira vez em um encontro realizado na Universidade da Califórnia, em Santa Cruz, em 1985. Apesar de a idéia ter sido considerada por muitos bastante instigante, poucos pesquisadores a apoiaram na realidade. Os opositores dessa idéia continuaram a levantar o mesmo conjunto inicial de objeções mesmo um bom tempo depois do início formal do projeto. A maioria dos argumentos contrários ao projeto de seqüenciamento em larga escala baseava-se nas seguintes questões-chave.

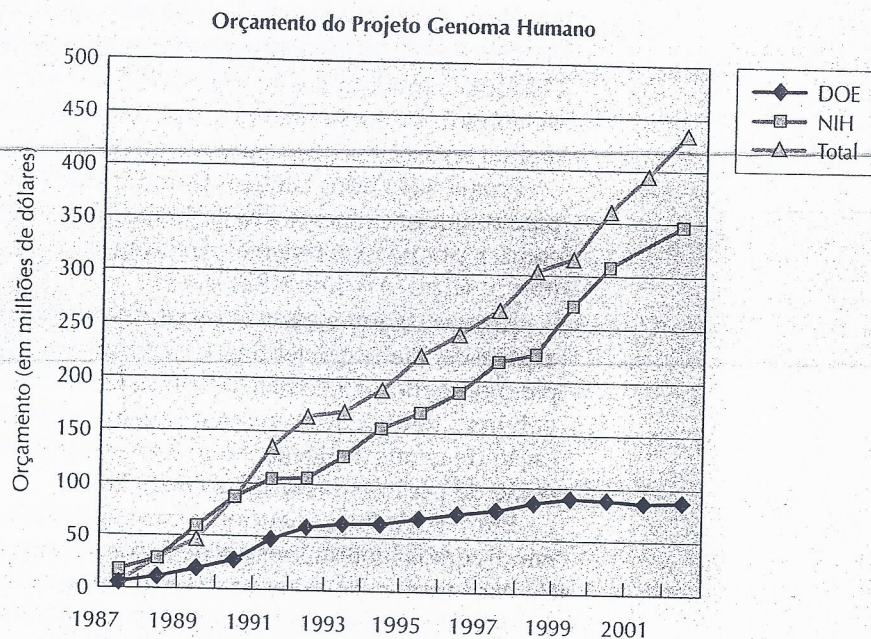
- **O que o projeto traria para a biologia?** Ao longo da maior parte de sua história, a biologia tem-se centrado em investigadores individuais e em seus pequenos grupos de estudantes, pesquisadores assistentes e técnicos de laboratório. Um grupo característico de biologia molecular, por volta de 1980, era composto, em média, por uma dúzia de indivíduos, apenas os cientistas na verdade mais bem estabelecidos possuíam grupos mais numerosos. Um projeto genoma descentralizado transferiria os esforços de numerosos investigadores individuais para um grande consórcio. Naquele tempo, quando as fronteiras entre a pesquisa biológica acadêmica e comercial estavam ainda

bastante definidas, um esforço massivo cheirava à industrialização da ciência. Em resumo, muitos imaginavam que o Projeto Genoma poderia mudar a "alma" da pesquisa biológica.

- Como esse Projeto seria financiado? Expresso em números, o custo e a duração do Projeto Genoma Humano alcançavam uma estimativa comum de 3 bilhões de dólares (1 dólar por nucleotídeo) ao longo de um período de mais de 15 anos. A competição por financiamentos de pesquisa já era árdua o suficiente na metade dos anos de 1980. Assim, muitos biólogos estavam preocupados que um projeto genoma desse porte pudesse redirecionar e absorver os financiamentos de programas já estabelecidos, tornando a obtenção de recursos inacessível a investigadores individuais.
- O seqüenciamento é realmente ciência? Apesar das informações obtidas a partir de um projeto genoma com certeza apresentarem interesse, o seqüenciamento propriamente dito foi visto de início como uma técnica laboriosa e monótona. Para muitos, esse projeto era visto como estando abaixo das aspirações de uma ciência rigorosa baseada em hipóteses. A afirmativa de que o seqüenciamento do DNA em grande escala era um trabalho mais adequado a prisioneiros, feita por Sydney Brenner, ficou famosa. Na verdade, o Centro Sanger, que se tornaria a maior instituição acadêmica de seqüenciamento, usou, no começo, como alternativa de recrutamento para seu corpo de pessoal técnico os empregados de padarias e de pequenas lojas da vizinhança de Cambridge, Inglaterra.
- Por que seqüenciar todo o genoma? Na década de 1980, todos os principais tipos de DNA não-codificador já haviam sido identificados, e estava bem estabelecido que eles eram responsáveis pela maior parte do genoma humano. Muitos cientistas não conseguiam vislumbrar uma justificativa para o seqüenciamento completo do genoma. Por que não seqüenciar apenas a partir dos genes conhecidos para estabelecer ligações entre eles e outros localizados dentro da mesma região de um cromossomo? O debate entre o seqüenciamento de "cabeça-à-cauda" (de telômero-a-telômero) ou de "retirada da nata" (centrado nos genes) nunca foi completamente resolvido por completo. Esse debate reapareceu em dois momentos-chave, dividindo a comunidade científica e fornecendo mais combustível para a "corrida" do seqüenciamento do genoma humano.

Nos Estados Unidos, conforme a comunidade biológica debatia os prós e contras do projeto genoma financiado pelo National Institute of Health (NIH), outra agência financiadora federal havia terminado sua própria análise e decidido dar andamento ao processo. O anúncio, em 1986, do Department of Energy (DOE) em dar início à sua própria Iniciativa Genoma Humano, com um financiamento de 5,3 milhões de dólares previsto para o ano fiscal de 1987, intrigou muitas pessoas. Por que o DOE estaria interessado em seqüenciar o genoma humano? Como primeiro objetivo, o DOE está sobretudo envolvido com os usos civis e militares da energia atômica. Ele foi fundado para controlar o sistema nacional de laboratórios envolvidos na produção de armas nucleares. A partir dessa idéia, o DOE havia penetrado na pesquisa biológica, direcionando-se, por necessidade, para os efeitos e respostas à radiação ionizante. Os estudos de mutações induzidas por radiação e de seu reparo levaram a uma maior atenção da genética molecular de humanos. Assim, existia uma lógica racional por trás do interesse do DOE no seqüenciamento do genoma humano.

O DOE tinha pouco interesse nos pontos de debate que alimentavam os biólogos. Essa instituição tinha experiência na administração de uma grande rede com projetos colaborativos e na administração de grandes somas de financiamento visando a resultados a longo termo, bilhões de dólares e décadas poderiam ser devotadas para o planejamento e a construção de um único acelerador de partículas que seria compartilhado por milhares de pesquisadores. Além disso, o enorme orçamento do DOE



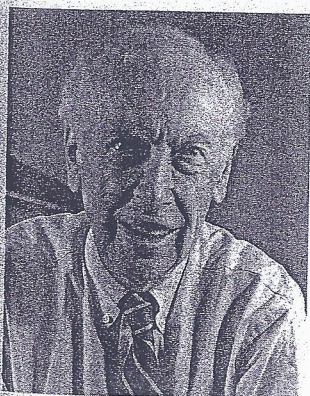
Financiamento do governo dos Estados Unidos para o Projeto Genoma Humano

poderia acomodar com facilidade um redirecionamento para a genômica e o trabalho poderia ser realizado em muitos laboratórios nacionais que desenvolviam fortes programas biológicos. Tendo em vista que muitos de seus projetos apoiavam-se na coleta de conjuntos de dados em larga escala, o seqüenciamento completo do genoma, telômero-a-telômero, não se apresentava sobremaneira assustador ou entediante. Além do mais, o DOE dispunha de uma infra-estrutura computacional adequada para estoque e compartilhamento da enorme quantidade de dados de seqüências que seriam gerados.

O surpreendente posicionamento do DOE deixou aos biólogos um motivo de ponderações sobre as consequências de permitir que uma agência controlada por físicos fosse responsável pelo que prometia ser o projeto biológico do século. Após a publicação em 1988 do relatório do National Research Council apoiando o seqüenciamento do genoma, os biólogos deixaram de lado suas diferenças e o NIH anunciou, em março, seu próprio projeto genoma. Nesse mesmo ano, o NIH e o DOE assinaram um memorando para cooperação dos esforços de seqüenciamento.

Em 1988, James Watson foi selecionado para liderar o esforço de seqüenciamento do NIH. Ele passou os anos seguintes em um contínuo trânsito entre Washington e Long Island, Nova York, onde permanecia como diretor dos Laboratórios de Cold Spring Harbor. Sendo um dos co-descobridores da estrutura do DNA, descoberta que deu início à revolução do DNA, Watson trouxe status científico, intuição e visibilidade para o projeto. Antevendo as muitas implicações da obtenção do roteiro da vida humana, Watson definiu que pelo menos 3% do valor do financiamento para a pesquisa seriam destinados às implicações éticas, legais e sociais da pesquisa genômica (ELSI, do inglês *Ethical, Legal, Social Implications*). O programa ELSI foi um experimento pioneiro na exploração da interação entre ciência e sociedade.

Anteriormente, havia sido determinado que o seqüenciamento do genoma humano poderia ser realizado em paralelo ao seqüenciamento do genoma de diversos organismos experimentais, tais como a bactéria *Escherichia coli*, a levedura *Saccharomyces cerevisiae*, o nematódeo *Caenorhabditis elegans*, a mosca-das-frutas *Drosophila melanogaster* e o camundongo *Mus musculus*. A análise comparativa entre as seqüências derivadas de humanos e aquelas derivadas de outros organismos de pesquisa poderia favorecer avanços em vários níveis. De um lado, genes pré-identificados em organismos experimentais poderiam ser utilizados para a identificação dos genes corres-



James D. Watson
(Cortesia da Coleção de James D. Watson, Cold Spring Harbor Laboratory Archives.)

pondentes em humanos. Por outro lado, as funções dos genes humanos e de seus produtos protéicos poderiam ser facilmente testadas nesses sistemas experimentais. O seqüenciamento dos genomas de *S. cerevisiae* e de *C. elegans*, além de regiões-alvo de genomas de mamíferos, forneceram resultados bastante promissores como primeiros testes de seqüenciamento em larga escala.

Projetos isolados também foram iniciados (apesar de não-financiados pelo NIH) para o seqüenciamento dos genomas de diversas espécies vegetais importantes, tais como a planta experimental *Arabidopsis thaliana* (da família das mostardas), o arroz (*Oryza sativa*) e o milho (*Zea mays*).

Apesar da intercessão dos críticos para que fosse dada atenção especial a genes expressos, Watson determinou de forma clara como objetivo o seqüenciamento completo do genoma telômero-a-telômero. No início, o projeto teve como foco o desenvolvimento de mapas cromossômicos e dos marcadores necessários para a identificação de genes de doenças que fornecessem pontos de ancoramento para a especificação da posição das seqüências de DNA.

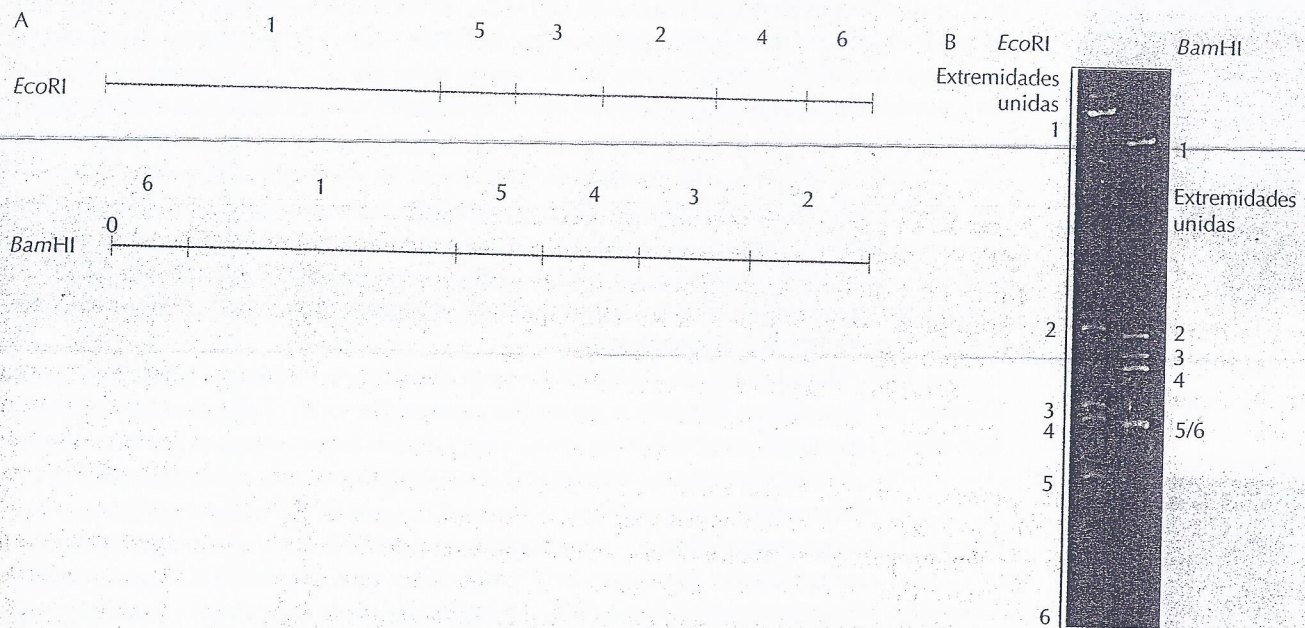
Em 1995, o NIH financiou centros-piloto de seqüenciamento do genoma humano em diversos laboratórios acadêmicos, tendo-se focado finalmente na Universidade de Washington (St. Louis), no Baylor College of Medicinal (Houston) e no Whitehead Institut Massachusetts Institut of Technology (Cambridge). O DOE agregou os esforços de seqüenciamento de vários laboratórios nacionais no Joint Genome Institute em Walnut Hills, Califórnia. Na metade dos anos 90, a Wellcome Trust, maior financiadora médica filantrópica do mundo, estabeleceu sua própria estrutura de seqüenciamento próximo a Cambridge, na Inglaterra. Esse instituto recebeu o nome do vencedor do prêmio Nobel Fred Sanger e tornou-se o maior centro acadêmico de seqüenciamento. Esses cinco centros de seqüenciamento do Projeto Genoma Público ficaram conhecidos como "G5" e contribuíram com 80% dos dados usados na montagem da seqüência inicial do genoma. Quinze outros centros importantes de seqüenciamento nos Estados Unidos, Japão, França, Alemanha e China cooperaram no seqüenciamento de cromossomos humanos específicos.

O trabalho desenvolvido com modelos experimentais e nos principais centros levou ao desenvolvimento de tecnologias-chave que aceleraram de maneira significativa a velocidade na qual o DNA pôde ser seqüenciado e montado, resultando em diminuição dos custos de seqüenciamento de aproximadamente US\$ 1,00 por par de base concluída para menos de US\$ 0,10. Nessas novas técnicas incluem-se (1) métodos automatizados de processamento de DNA clonado, (2) sistemas analíticos e bioquímicos para o rápido seqüenciamento de DNA, (3) sistemas computacionais de organização e distribuição de dados e (4) algoritmos para acesso da fidelidade da seqüência, montagem das seqüências de DNA e identificação de genes.

MAPAS CROMOSSÔMICOS E MARCADORES

A construção de mapas de alta resolução para cada cromossomo foi um importante objetivo na abordagem inicial do Projeto Genoma. Assim como mapas de estradas, os mapas de cromossomos humanos podem ser construídos em uma ampla gama de diferentes resoluções. Os mapas genéticos descrevem a localização dos genes e seus padrões de herança. Os mapas de ligação genética, como os construídos no início para a *Drosophila* por Thomas Hunt Morgan e colaboradores, no início da década de 1900 (ver Capítulo 1), têm sido úteis para a montagem de esquemas de baixa resolução de genomas totais, os quais deram origem à medida clássica de distância genética, o centiMorgan (cM).

O cM é a medida da frequência de recombinação durante a formação dos gametas. Lembre que, durante a meiose, os cromossomos pareados se alinham e se "recombinam" para a troca de segmentos de DNA. Um cM é definido como a distância tal entre dois genes ou marcadores que eles serão separados durante a meiose em 1%



Mapa de restrição do DNA do bacteriófago λ

(A) Mapas mostrando os sítios de restrição das enzimas *EcoRI* e *BamHI* no DNA de λ . As barras verticais representam sítios nos quais a enzima corta.
 (B) Uma separação eletroforética em gel de DNA de λ digerido pelas enzimas *EcoRI* e *BamHI*. Os números indicam o fragmento correspondente mostrado no mapa. O DNA não foi suficientemente corrido e separado por eletroforese para permitir a resolução das bandas 5 e 6 da digestão com *BamHI*. As bandas indicadas por "extremidades unidas" representam os dois fragmentos terminais, unidos por extremidades coesivas (sítios cos) do DNA de λ . (Reimpressa, com permissão, de Snyder L.A., Freifelder D. e Hartl D.L. 1985. *General Genetics*. Jones and Bartlett, Boston.)

das vezes. No início do Projeto Genoma, um cM em distância genética era estimado como igual a cerca de 1 milhão de nucleotídeos em distância física. A sequência inicial do genoma humano produzida pela empresa Celera mostrou uma média de 1,22 milhões de nucleotídeos (megabases, Mb) por cM.

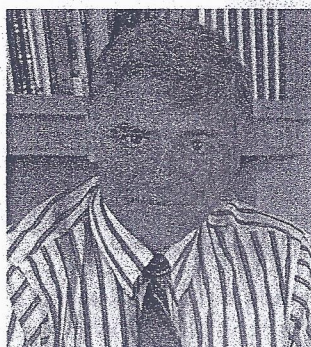
Os experimentos iniciais realizados na década de 1970 originaram mapas dos sítios de clivagem de enzimas de restrição ao longo dos cromossomos virais, os quais foram os primeiros *mapas físicos*, nos quais cada sítio marcava a ocorrência de uma característica estrutural da própria molécula de DNA. Apesar de enzimas de restrição terem-se mostrado bastante úteis no mapeamento de pequenos cromossomos, seus sítios de reconhecimento ocorrem com demasiada frequência para serem usados na construção de mapas de larga escala de cromossomos humanos. No entanto, tal problema foi superado pelo uso de enzimas de restrição que reconhecem sítios pouco frequentes. A eletroforese em gel de campo-pulsado, um método de separação de grandes pedaços de DNA, foi usada para a análise desses grandes fragmentos.

A identificação de diversidade populacional nos padrões de corte de enzimas de restrição forneceu um método para a produção dos primeiros mapas físicos de cromossomos humanos completos. Na metade dos anos 80, observou-se que mutações pontuais podiam às vezes criar ou deletar sítios de reconhecimento de enzimas de restrição ao longo de um cromossomo. Também a adição de pequenas repetições em tandem pode aumentar a distância entre os sítios de corte das enzimas. Cada mutação ou variação da repetição em tandem é um alelo diferente de uma molécula de DNA. Distintos alelos produzem diferentes tamanhos de fragmentos de restrição ou polimorfismos do tamanho do fragmento de restrição (RFLPs), os quais podem ser medidos em nucleotídeos.

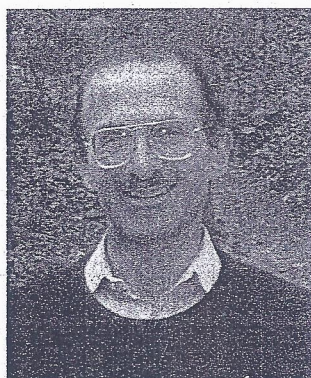
A genética clássica havia usado os genes como marcadores para definir posições discretas (loci) sobre os cromossomos. Um alelo era definido como uma forma alternativa de um gene, no que diz respeito a um fenótipo específico detectável. Os sítios de reconhecimento das enzimas de restrição forneceram uma nova definição para os



Maynard Olson
(Cortesia de CSHL Archives.)



Leroy Hood



Charles Cantor



David Botstein
(Cortesia de CSHL Archives.)

loci, como seqüências específicas de DNA. Os RFLPs estenderam a definição de alelo para significar uma forma alternativa de DNA – um comprimento diferente da molécula de DNA. Apesar de os RFLPs terem sido, no início, utilizados para testes de identidade e diagnóstico de DNA (ver Capítulo 8), eles se tornaram extremamente úteis para o mapeamento de cromossomos, mesmo na ausência de qualquer relação fenotípica. Por meio da medida da frequência de recombinação entre marcadores RFLP pode ser fornecida uma primeira medida para relacionar a distância genética (calculada em cMs) com a distância física (calculada em nucleotídeos). O primeiro mapa de ligação do genoma humano, que incorporava 403 marcadores RFLP, foi publicado em 1987 por Helen Donis-Keller e colaboradores, do Collaborative Research, Inc. de Boston, Massachusetts.

Em 1989, Maynard Olson (Universidade de Washington), Leroy Hood (California Institut of Tecnology), Charles Cantor (Lawrence Berkeley Laboratories) e David Botstein (Genentech, Inc.) estabeleceram uma nomenclatura comum para os marcadores moleculares pela qual os cientistas podem facilmente trocar informações de seqüências e obter sondas para regiões conhecidas do genoma. Em retrospectiva, o que eles propuseram parece óbvio – uma seqüência de DNA *per se* pode agir como um marcador cromossômico. Qualquer seqüência cópia única de 200 a 500 nucleotídeos – um sítio marcador de seqüência (STS, do inglês *sequence-tagged site*) – pode ser usada como uma marca física sobre um cromossomo.

Cada marcador RFLP ou STS deve ter sido identificado em sua posição específica do cromossomo pelo uso de fragmentos clonados de DNA. Os mapas *contíguos* (*contig*) alinham as extremidades de fragmentos de DNA adjacentes, dessa maneira enfileirando entre si seqüências contínuas. Os mapas contig orientam marcadores físicos em fragmentos adjacentes e, sendo assim, fornecem pontos de referência espaçados a intervalos regulares ao longo do genoma. No final da década de 1980, Ray White e seus colaboradores, da Universidade de Utah, haviam estabelecido um conjunto de marcadores RFLF que abrangia os cromossomos em intervalos de aproximadamente 10 cM. Um dos objetivos iniciais do Projeto Genoma era fornecer mais marcadores aos mapas cromossômicos e diminuir o tamanho dos intervalos existentes entre tais marcadores. Quando da publicação da seqüência geral inicial, a resolução do mapa físico havia sido aumentada em um fator de 100 – existindo então 30.000 distintos marcadores em intervalos de cerca de 0,1 cM (100.000 pb).

Conforme a qualidade desses mapas genômicos aumentava, eles se tornavam ferramentas valiosas para a identificação de genes envolvidos em doenças humanas. Anteriormente, a descoberta de um gene associado a uma doença era uma tarefa árdua que envolvia anos de trabalho intenso. Por exemplo, Mark Skolnick, da Universidade de Utah, demorou quatro anos (1990 a 1994) para identificar e posicionar o gene *BRCA1*, gene relacionado ao câncer de mama. Mapas melhorados propiciaram que Richard Wooster, do Haddow Laboratories Institute of Cancer Research, clonasse o gene *BRCA2* em um período de poucos meses.

Os STSs representam a etapa final da evolução dos marcadores de DNA. Genes, sítios de reconhecimento de enzimas de restrição e RFLPs representam propriedades funcionais associadas ao DNA. Os STSs, por outro lado, são “marcadores puros”, considerando-se que podem ser mapeados sem o conhecimento de genes ou de enzimas de restrição. No entanto, a viabilidade e a disseminação do uso dos marcadores STS necessitaram de duas tecnologias que surgiram nos anos 80: a reação em cadeia de polimerase e o seqüenciamento automatizado de DNA.

A reação em cadeia de polimerase

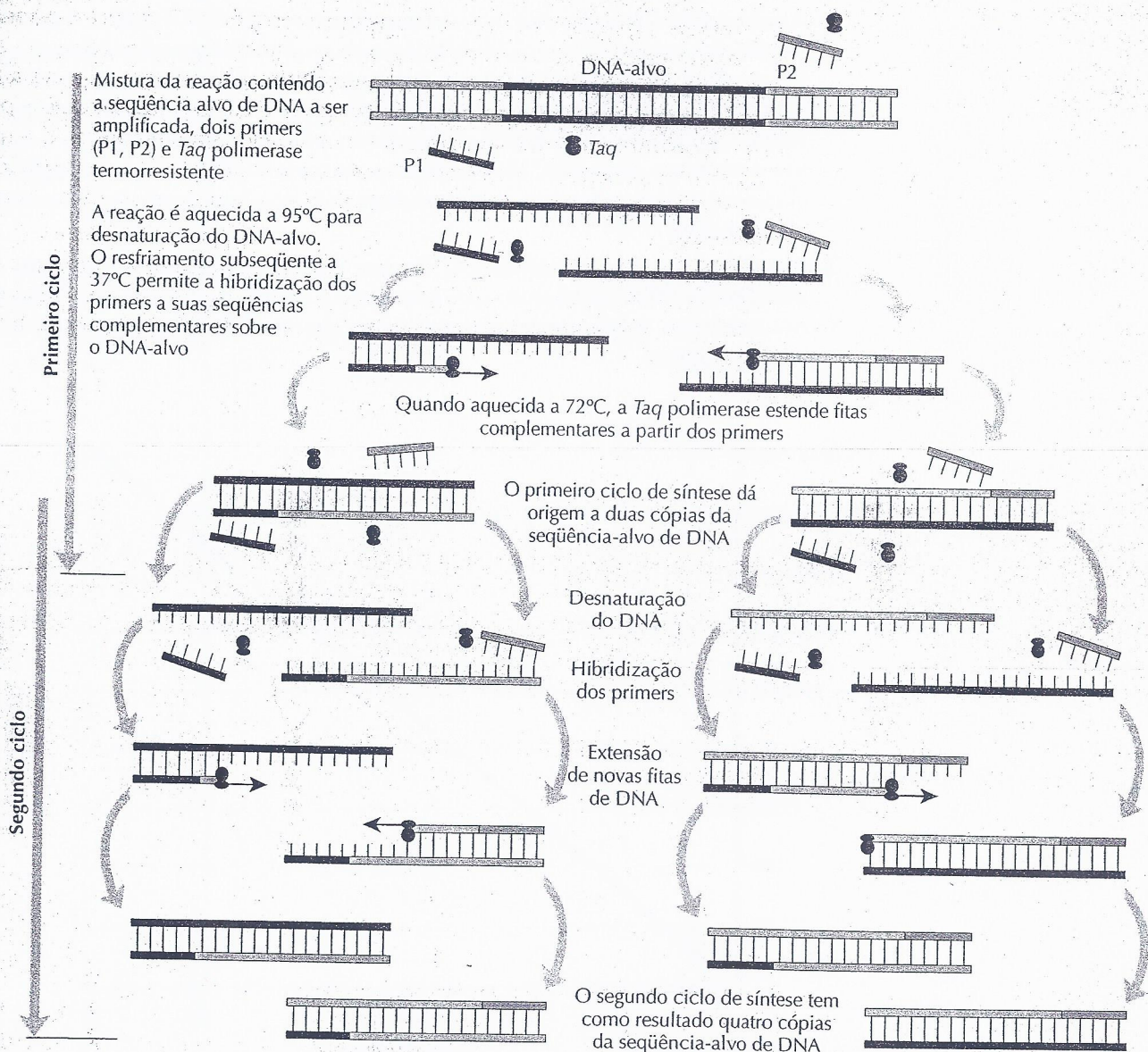
A reação em cadeia de polimerase (PCR), descoberta em 1985 por Kary Mullis, da Cetus Corporation, mostrou-se uma tecnologia-chave para o Projeto Genoma em muitos aspectos. Primeiro, ela ofereceu uma maneira simples para a obtenção de marcadores de DNA e para a genotipagem de fragmentos de cromossomos, sem a necessi-



Kary Mullis
(Cortesia de CSHL Archives.)

dade de obtenção de clones. Segundo, ela ofereceu uma nova abordagem química que permitiu a automatização do sequenciamento de DNA.

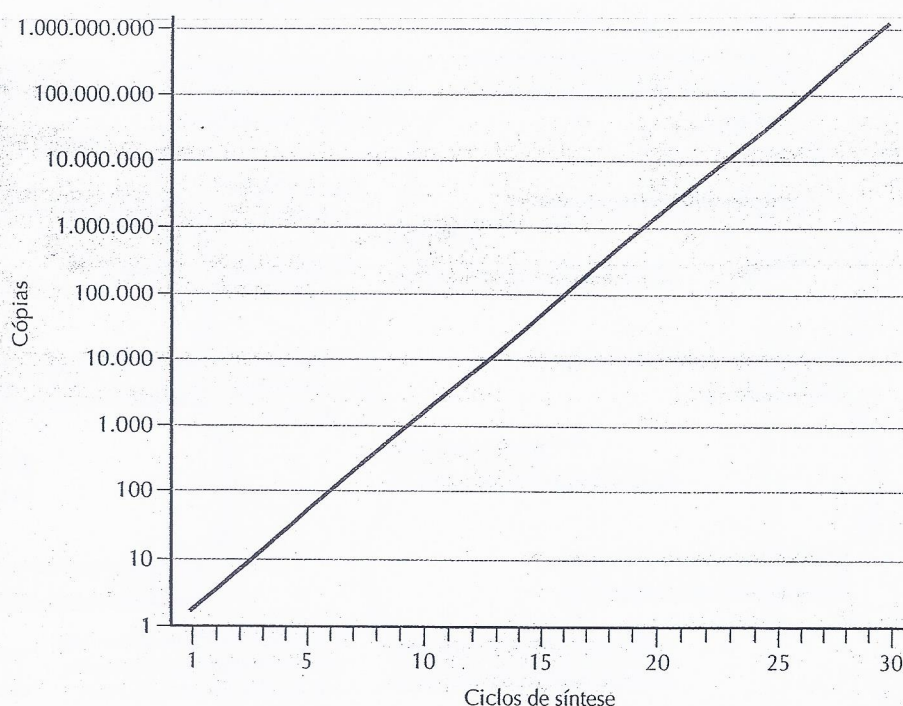
Até o advento da PCR, a única maneira de purificar uma sequência de DNA era por intermédio da amplificação biológica em células em cultura. Para a obtenção de uma amostra de uma dada sequência de DNA suficientemente pura, era necessário cloná-la em um vetor e fazer crescer esse vetor clonado em uma cultura de bactérias ou de leveduras. Assim, no início do Projeto Genoma não existiam alternativas além da manutenção de culturas-padrão de bactérias e leveduras, cada uma transformada com um fragmento de um cromossomo humano identificado por STS específico ou por um marcador RFLP. A perspectiva de manutenção de um gigantesco banco central de clones e da distribuição desses clones entre os grupos de pesquisa associados no mundo era assustadora. A PCR forneceu uma alternativa simples para o isolamento de quantidades puras de sequências específicas de DNA. A PCR faz uso da amplificação enzimática para aumentar o número de cópias de um dado fragmento de DNA com tamanho de até 6000pb. Por meio da utilização de dados provenientes de bancos de dados STS, um pesquisador localizado em qualquer ponto do globo pode usar a PCR para testar o fragmento de DNA em relação ao STS determinado em poucas horas.



Amplificação de DNA por meio da reação em cadeia de polimerase

A PCR está baseada no fenômeno e na extensão de um primer pela DNA polimerase, descoberto nos anos 60 (ver Capítulo 2). A síntese de cada fita de DNA é " iniciada " por um pequeno pedaço de dupla fita. Trabalhando a partir desse primer, a polimerase adiciona nucleotídeos complementares à fita de DNA-molde, estendendo a região de dupla fita. No início, um par de primers de oligonucleotídeos de DNA (*oligo* significa poucos) de comprimento igual a aproximadamente 20 nucleotídeos é sintetizado de tal forma que estes se ligarão delimitando a região " alvo " a ser amplificada. Os primers são desenhados de forma a anelar seqüências complementares de DNA que se encontram a 5' sobre cada uma das moléculas de DNA. Um excesso de cada um dos dois primers é adicionado à amostra de DNA que contém a seqüência-alvo, em conjunto com uma DNA polimerase. Também serão fornecidos o co-fator magnésio (Mg^{++}) e os quatro desoxirribonucleotídeos trifosfato (dNTPs). A mistura de reação será então submetida a múltiplos ciclos de síntese de acordo com o seguinte protocolo.

1. **Desnaturação:** o aquecimento à temperatura próxima da ebulição ($94^{\circ}C$) desnatura o DNA-alvo e cria um conjunto de moldes fita simples. O aquecimento aumenta a energia cinética da molécula de DNA a um ponto mais alto que a energia necessária para a manutenção das pontes de hidrogênio entre os pares de bases. O DNA dupla fita é separado em fitas simples.
2. **Anelamento (pareamento):** o resfriamento a cerca de $65^{\circ}C$ favorece o anelamento dos primers oligonucleotídeos a suas seqüências complementares sobre as fitas simples molde. A temperatura ótima de anelamento varia de acordo com a proporção de pares de bases AT e GC na seqüência dos primers. Considerando que os primers foram adicionados em excesso e que são estruturas pequenas, eles anelarão com mais facilidade a suas seqüências-alvo complementares do que as duas fitas originais conseguirão sofrer renaturação.
3. **Extensão:** o aquecimento a $72^{\circ}C$ fornece uma temperatura ótima para a atuação da DNA polimerase que procederá à extensão do primer de oligonucleotídeos. A polimerase sintetiza uma segunda fita complementar à fita molde original.



Amplificação teórica de uma reação em cadeia de polimerase

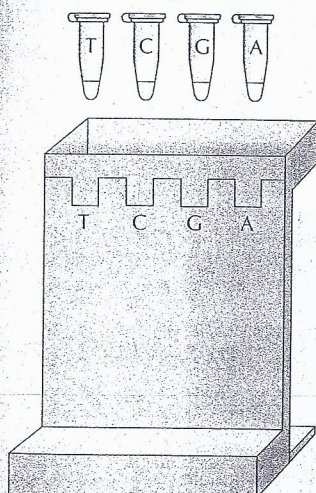
Ao longo de cada ciclo de síntese, que dura cerca de 2 minutos, o número de cópias da molécula de DNA-alvo é duplicado. *Vinte e cinco* ciclos de síntese produzem, teoricamente, uma amplificação de 1.000.000 de vezes da sequência-alvo em um período de aproximadamente uma hora.

As altas temperaturas usadas na PCR requerem o uso de uma DNA polimerase termo-estável ou termo-resistente. No final da década de 1980, a primeira DNA polimerase termoestável foi isolada a partir de uma cultura de *Thermus aquaticus*, originalmente coletado nos anos 60, no Mushroom Pool do Parque Nacional de Yellowstone. Essa bactéria está adaptada à vida em águas termais, que atingem 80°C, e sua DNA polimerase (denominado *Taq* polimerase) se mantém ativa mesmo depois de repetidas exposições a temperaturas próximas da ebulição, o que a torna ideal para a PCR. As chaminés vulcânicas oceânicas tornaram-se outra fonte importante de organismos adaptados a condições extremas (extremófilos) que possuem polimerases termo-resistentes.

A sequência de mudanças de temperatura era, no início, realizada por meio da transferência dos tubos de reação de um banho-maria para outro. Essa operação tediosa foi substituída pelo advento dos termocicladores de DNA automatizados, nos quais as temperaturas do ciclo e o tempo de permanência dos tubos a cada temperatura são inscritos em um programa de computador que aumenta ou diminui a temperatura de blocos de aquecimento. Esse processo automatiza a reação e os tubos permanecem em um mesmo local, no bloco, enquanto sofrem a contínua ciclagem por meio das etapas da PCR. A combinação de polimerases termo-resistentes e de termocicladores automatizados permite que praticamente qualquer cientista ou estudante possa realizar uma rápida análise de DNA.

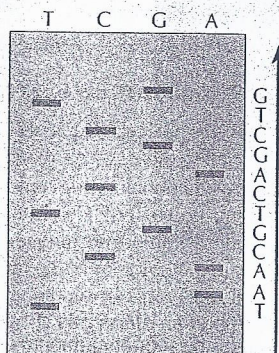
Eletroforese em gel de poliacrilamida

Reações de sequenciamento são aplicadas em um gel desnaturante



Auto-radiografia do Gel

As canaletas são lidas em sequência, de baixo para cima



Método de sequenciamento de DNA por terminação dideoxi

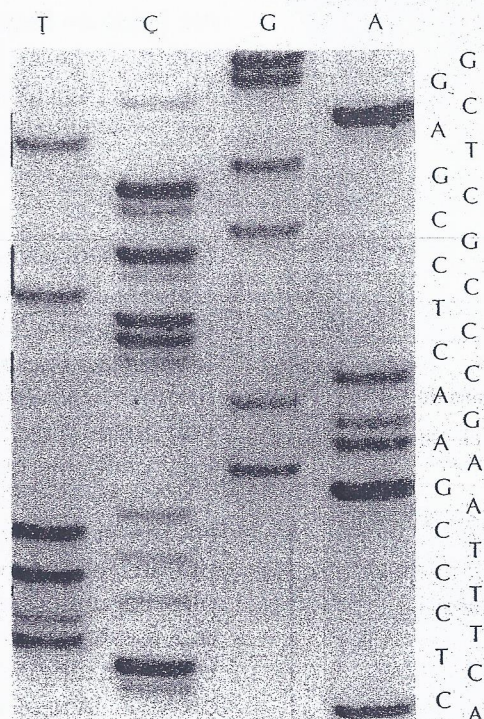
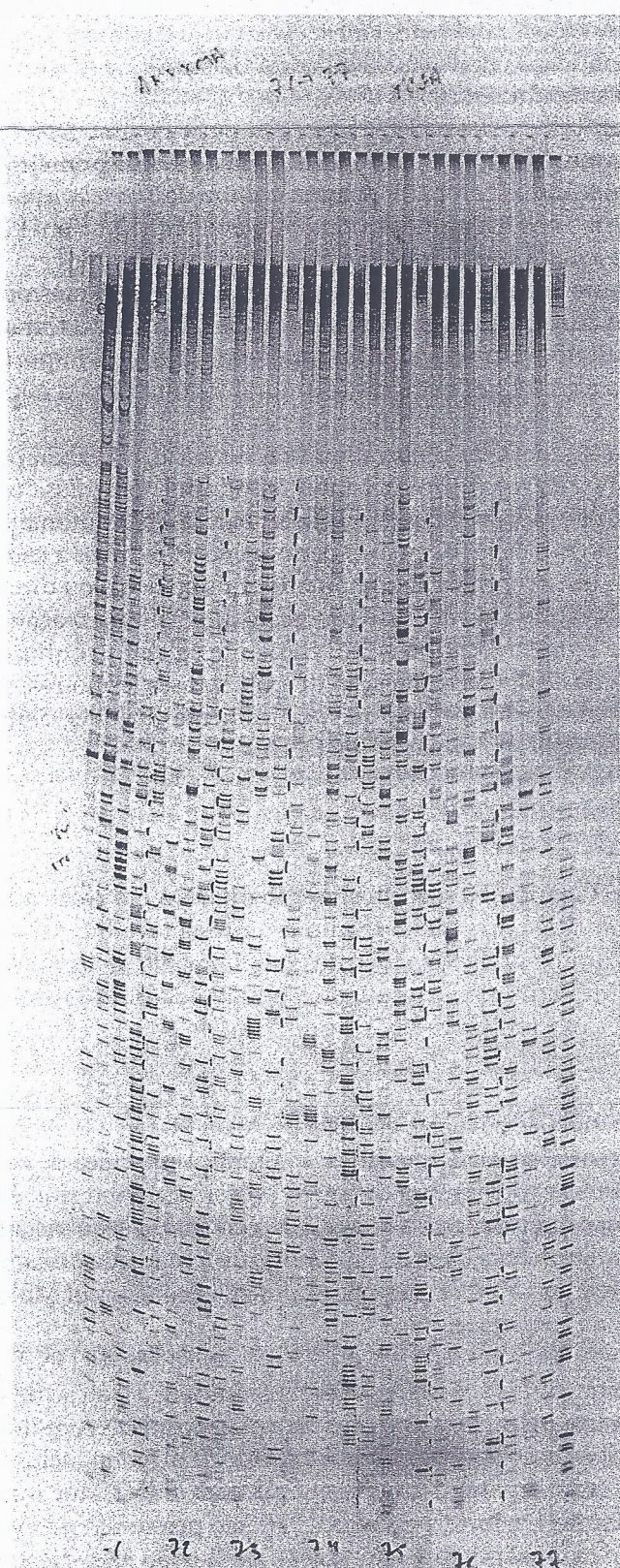
Sequenciamento de DNA

No início do Projeto Genoma, o sequenciamento de DNA era realizado manualmente, de forma que consumia muito tempo e representava intensivo trabalho. Um dos principais trunfos do Projeto foi o desenvolvimento de tecnologias de sequenciamento que trouxeram como resultado a possibilidade de leitura de 500.000 ou mais nucleotídeos em um único dia e a um custo de poucos centavos por nucleotídeo.

Em 1977, dois métodos de sequenciamento de DNA foram independentemente desenvolvidos. O método de clivagem química de Allan Maxam e Walter Gilbert, da Universidade de Harvard, envolve a modificação química específica de cada nucleotídeo diferente. Tal alteração permite a clivagem posterior em um sítio específico e identifica o nucleotídeo presente nessa posição. No entanto, o sequenciamento moderno de DNA está inteiramente baseado no método de terminação de cadeia, desenvolvido por Fred Sanger, no Medical Research Council Laboratory of Molecular Biology, em Cambridge, na Inglaterra.

O método de Sanger está baseado em dois fatos da síntese do DNA. Primeiro, em presença dos quatro dNTPs, a DNA polimerase iniciará a síntese de uma nova fita de DNA quando um pequeno primer for hibridizado ao DNA molde fita simples. Segundo, se dideoxynucleotídeos trifosfato (ddNTPs) forem incluídos na reação, a extensão do DNA será interrompida quando um ddNTP for incorporado. Essa interrupção da síntese ocorre, pois o ddNTP não possui um grupamento hidroxil 3' livre (-OH), necessário para a formação da ligação fosfodiéster que une nucleotídeos adjacentes.

No protocolo de sequenciamento por dideoxi, quatro tubos de reação (A, T, G e C) são preparados. Cada um desses tubos contém DNA-molde, sequência primer, DNA polimerase e quatro deoxynucleotídeos trifosfato (dATP, dTTP, dCTP e dGTP, um dos quais se encontra marcado radioativamente). Um único tipo de ddNTP é adicionado a cada uma das quatro reações – ddATP (ao tubo A), ddTTP (ao tubo T), ddCTP (ao tubo C) ou ddGTP (ao tubo G).



Auto-radiografia de um gel de seqüenciamento por terminação dideóxi

Sete clones separados contendo fragmentos de restrição foram seqüenciados em reações aplicadas lado a lado e separadas por eletroforese em um gel de seqüenciamento. Cada conjunto de quatro colunas (reações T, C, G e A), nesta auto-radiografia, representa os dados de seqüenciamento de um clone diferente. Um segmento dos dados de seqüenciamento do clone 5 está apresentado em detalhe. O referido experimento foi realizado no laboratório de Fred Sanger, o cientista que desenvolveu o método de seqüenciamento por dideóxi. (Cortesia de W. Herr, Cold Spring Harbor Laboratory.)

Começando a partir do primer, a polimerase adiciona aleatoriamente dNTPs ou didNTPs complementares ao DNA molde. O balanço entre dNTPs e didNTPs na reação está ajustado de modo a um didNTP ser incorporado na cadeia de DNA que está sendo sintetizada aproximadamente a cada 100 nucleotídeos. Cada vez que um didNTP é incorporado, a síntese é interrompida, sendo, em consequência, gerada uma fita de DNA de tamanho correspondente. Após a replicação, existem milhões de cópias da sequência de DNA-molde cuja síntese foi interrompida em cada possível posição nucleotídica.

Na reação A, por exemplo, a cada 100 resíduos de timina encontrados sobre a fita de DNA-molde, a polimerase incorporará 99 unidades do nucleotídeo complementar deoxiadenosina (lembre-se que, na síntese de DNA, o A complementa o T da fita oposta). Em um desses 100, no entanto, a polimerase irá incorporar uma di-deoxiadenosina, provocando a interrupção da síntese da fita neste ponto. Em outras palavras, a cadeia de DNA será terminada na posição de timina em 1% do total de moléculas molde. Tendo em vista que milhões de moléculas de DNA estão presentes, a reação A resultará em uma coleção de fragmentos de DNA nos quais a síntese foi interrompida de modo a retratar as posições de cada resíduo de timina na fita de DNA-molde.

Quando as reações estão completas, adiciona-se formamida para desnaturar as novas fitas sintetizadas a partir do DNA molde. Cada reação é aplicada em uma canaleta diferente de um gel de poliacrilamida contendo uréia, o que evita que as fitas de DNA formem estruturas secundárias durante a eletroforese. Os fragmentos migram pelo gel de acordo com seu tamanho e, dessa maneira, formam uma espécie de "escada" de bandas, nas quais cada degrau será composto por moléculas de DNA que diferem dos degraus adjacentes em um único nucleotídeo de comprimento.

Após a eletroforese, o gel é colocado em contato com um filme de raio X. Os nucleotídeos marcados radioativamente, que foram incorporados durante a síntese, sensibilizam o filme de raio X, produzindo uma série de bandas e, como resultado, indicando o tamanho dos fragmentos gerados nas reações A, T, C e G. O gel é então "lido" de baixo para cima, iniciando-se com o menor dos fragmentos de DNA e seguindo-se ao longo das canaletas para a identificação sucessiva dos fragmentos maiores. Na verdade, *scanners* ópticos foram desenvolvidos para a leitura automatizada de géis, pois o seqüenciamento manual era um processo muito lento para que se pudesse realizar o seqüenciamento do genoma humano de forma rápida e acurada.

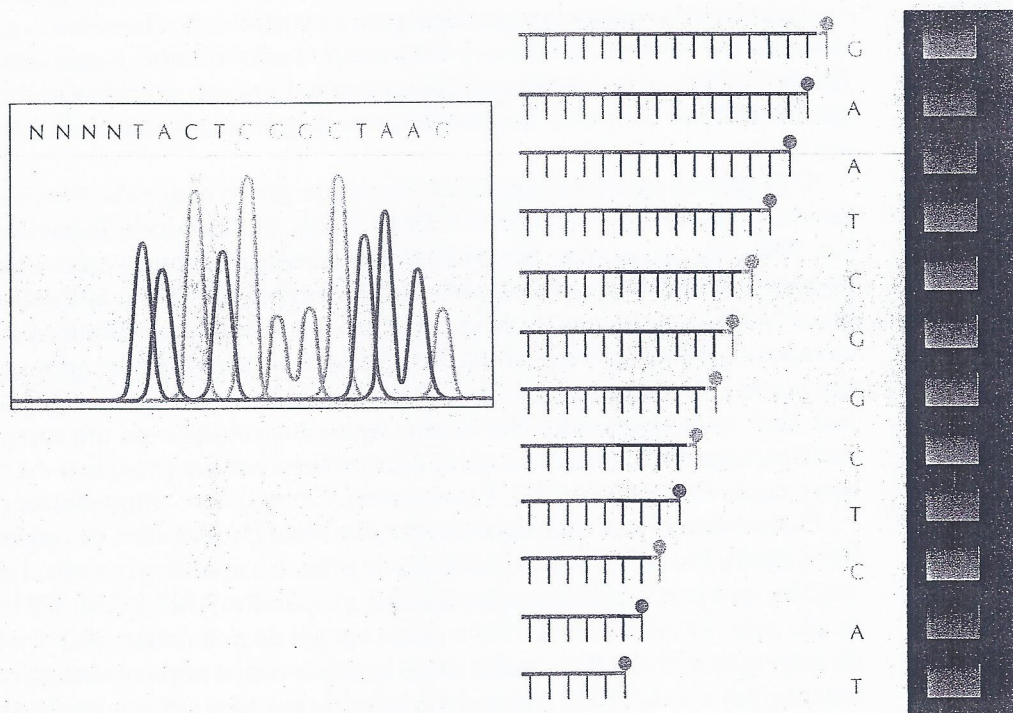
O seqüenciamento automatizado tornou-se possível devido a colorações químicas desenvolvidas por Lee Hood e Lloyd Smith, do California Technology Institute. Em 1986, eles conjugaram um corante fluorescente diferente a cada uma das quatro reações didNTP. As quatro reações de seqüenciamento foram aplicadas sobre uma única canaleta em um gel de poliacrilamida e os marcadores fluorescentes foram detectados conforme os fragmentos de terminação migravam pelo gel e passavam em um ponto determinado sobre o qual incidia um *laser* de argônio. Quando atingido pelo *laser*, cada terminador fluorescente emitia luz colorida em um comprimento de onda característico, o qual era então interpretado por um programa de computador como equivalente a A (verde), T (vermelho), C (azul) ou G (amarelo) na posição.

Hood desenvolveu uma colaboração com Mike Hunkapiller, da empresa Applied Biosystems, Inc. (ABI), para a produção do primeiro aparelho comercial de leitura de seqüências a partir de fragmentos corados. O aparelho ABI Modelo 370 foi comercializado pela primeira vez em 1987 e usava um gel de poliacrilamida para a realização da migração eletroforética, assim como acontece com o seqüenciamento manual. No entanto, em vez de leitura manual ou óptica do gel, esse sistema empregava um *laser* para a detecção dos nucleotídeos marcados com fluorescência. O seqüenciador incorporava um programa de computador que construía uma imagem simulada do gel sob a forma de bandas de DNA coloridas conforme elas passavam pelo laser detector durante a eletroforese. A aparência final dos resultados era um eletroferograma,

mostrando picos coloridos correspondentes a cada posição nucleotídica. O Seqüenciador de DNA ABI 370, equipado com um gel de poliacrilamida de 16 canaletas, tinha capacidade de seqüenciar até 20.000 nucleotídeos por dia. O aumento do número de canaletas para 32, 48 e, finalmente, para 96 aumentou a capacidade de seqüenciamento para 80.000 nucleotídeos diários ou mais.

Permitindo que todos os quatro nucleotídeos fossem analisados em uma única canaleta, a química fluorescente desenvolvida por Hood quadruplicou a capacidade dos géis de seqüenciamento. Outros avanços em paralelo, com relação ao preparo de DNA, aumentaram ainda mais a eficiência do sistema. No final dos anos de 1980, polimerases termor-resistentes (como a *Taq*, por exemplo) e termocicladores de DNA foram colocados a serviço da marcação automatizada, uma metodologia híbrida que ficou conhecida como seqüenciamento em ciclos. Em 1987, a Dupont introduziu os "terminadores marcados", nos quais um corante fluorescente diferente estava conjugado a cada um dos quatro nucleotídeos terminadores (didATP, didTTP, didCTP ou didGTP). Isso permitiu que todos os quatro nucleotídeos fossem marcados ao mesmo tempo em uma única reação, em uma corrente contínua de preparação de amostras e, na realidade, quadruplicou a capacidade de trabalho diário de um técnico de laboratório.

A fase final do desenvolvimento do seqüenciamento em larga escala foi a substituição dos géis de poliacrilamida por múltiplos capilares preenchidos por gel, cujo potencial havia sido demonstrado em 1990. Os novos aparelhos de seqüenciamento automático incluem conjuntos de 96 capilares conectados a um mecanismo capaz de reaplicar amostras de forma automática mais de 12 vezes por dia. Os aparelhos com capilares eliminaram elementos altamente trabalhosos que envolviam um grande dispêndio de tempo, como o preparo de géis e a aplicação manual das amostras, restringindo as intervenções humanas apenas à manutenção dos níveis de reagentes e à



Dados amostrais obtidos a partir de um seqüenciador automático de DNA

(à esquerda) Eletroferograma mostrando os picos correspondentes a cada posição nucleotídica. (ao centro) Representação dos fragmentos de DNA, gerados na reação de seqüenciamento e que correspondem a cada um dos picos. (à direita) Representação do gel de seqüenciamento desses fragmentos.

colocação de placas de amostras no sistema de aplicação automático. Assim, enquanto a corrida visando o primeiro quadro geral do genoma humana completo "animava-se", cada um dos centros de seqüenciamento havia se tornado uma fábrica automatizada, equipada com mais de 30 desses seqüenciadores, cada um gerando seqüências de até 400.000 nucleotídeos por dia. A rapidez do processo de seqüenciamento havia aumentado mais de 400 vezes desde o início do Projeto.

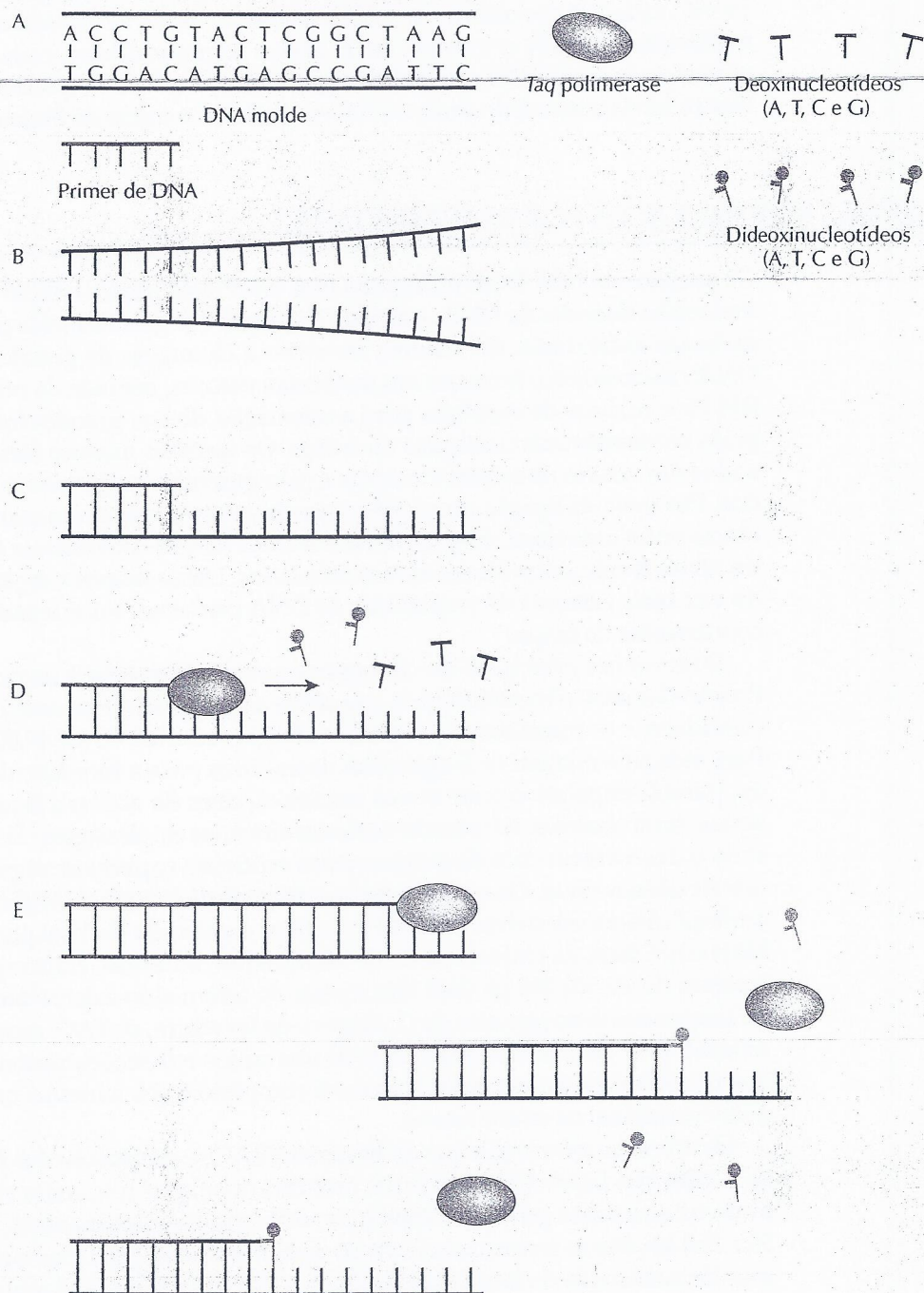
BANCOS DE CLONES HIERÁRQUICOS POR FRAGMENTAÇÃO

Os projetos do NIH-DOE utilizaram uma estratégia para a criação de um banco de moléculas clonadas de DNA, partindo de mapas em grande escala para chegar às seqüências individuais. Tal sistema envolveu a clonagem de grandes fragmentos de DNA cromossômico humano em moléculas vetores, seguida da utilização de STSs, RFLPs e enzimas de restrição para a orientação desses grandes fragmentos em regiões cromossômicas contíguas ou *contigs*. Os insertos maiores foram então subclonados em vetores de seqüenciamento e os segmentos seqüenciados foram remontados. Por meio da ligação de seqüências adjacentes e mantendo marcadores de referência entre cada nível, os pesquisadores puderam se movimentar entre os diferentes níveis hierárquicos desses clones ordenados. Dessa maneira, regiões contíguas cada vez mais extensas de seqüências de DNA puderam ser montadas pelo uso das coordenadas do mapa.

Dois vetores principais de clonagem estavam disponíveis no período inicial do Projeto Genoma: o bacteriófago λ , que aceita a clonagem de insertos de até 20.000 nucleotídeos; e os cosmídeos, que podem abrigar insertos de até 40.000 nucleotídeos. Para reduzir o número de fragmentos necessários para a formação de um mapa *contig*, parecia imperativo criar novos vetores capazes de aceitar a incorporação de insertos muito maiores. Afastando-se desses sistemas de clonagem clássicos, foi necessário o desenvolvimento de cromossomos artificiais capazes de se propagar no interior de células vivas. Os cromossomos artificiais de levedura (YACs, do inglês *yeast artificial chromosomes*) desenvolvidos no final da década de 1980 por David Burke e Maynard Olson, da Universidade de Washington, St. Louis, eram capazes de aceitar insertos de até 500.000 pb (500 kb). Apesar de terem sido extensivamente utilizados no mapeamento do genoma de *C. elegans* e de levedura, os YACs mostraram-se hereditariamente instáveis. A ocorrência de rearranjos e deleções tornou difícil garantir que o inserto em seqüenciamento estava completo e sob a mesma ordenação da seqüência original no cromossomo.

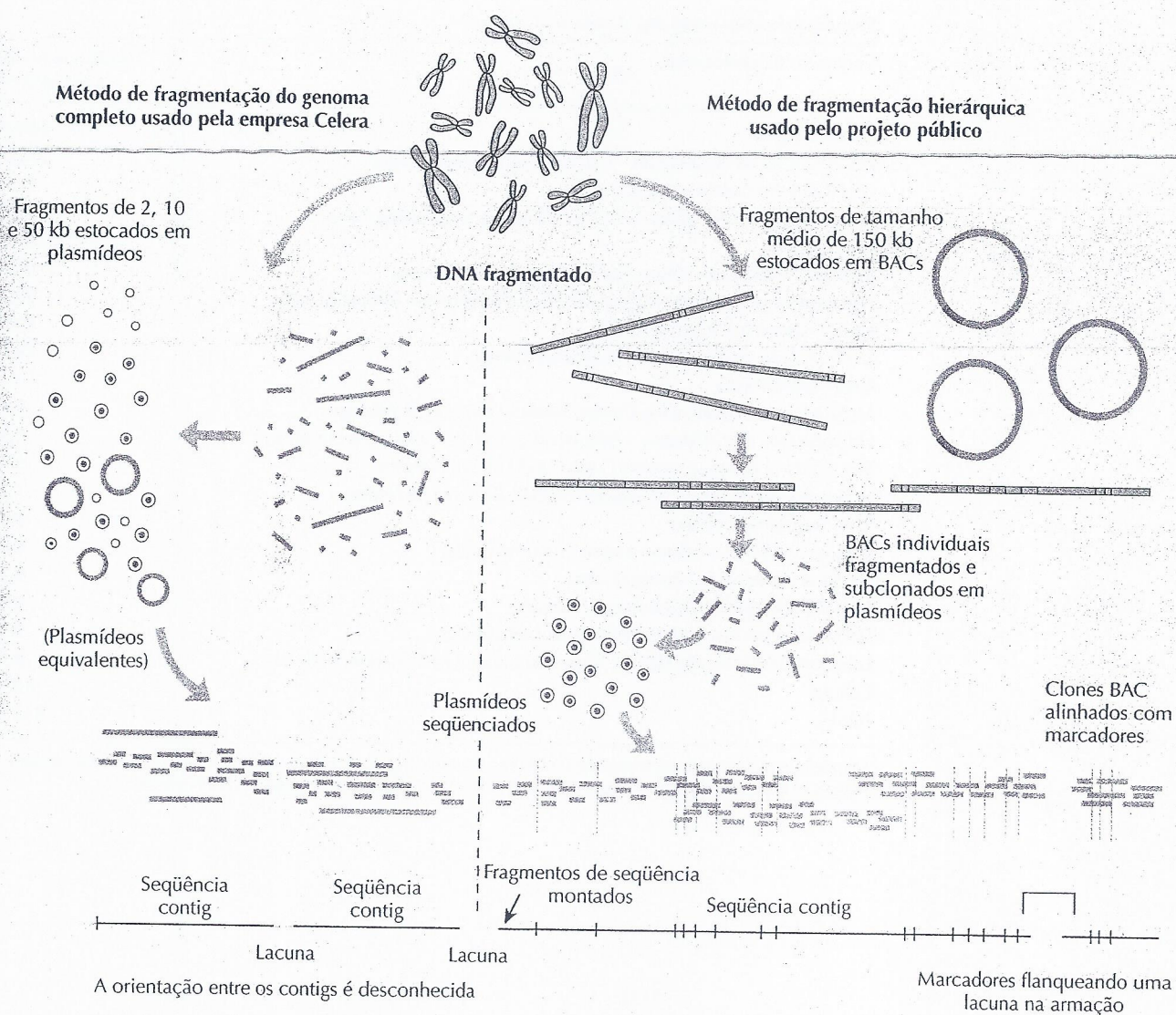
Os cromossomos artificiais de bactérias (BACs), desenvolvidos por Mel Simon, no California Technology Institute, mostraram-se, por fim, como as grandes locomotivas do projeto genoma. Apesar de aceitarem insertos menores, que variam de 50 a 250 kb, eles se mostraram estáveis ao extremo e de fácil manipulação em sistemas de automação de larga escala. Simon e Pieter de Jong, do Roswell Park Cancer Institute, em Buffalo, Nova York, construíram extensivas bibliotecas de BAC que foram utilizadas como base de grande parte do trabalho de seqüenciamento do projeto público.

Para construir as bibliotecas de BAC, o DNA genômico humano era clivado de maneira aleatória com enzimas de restrição gerando fragmentos que possuíam em média 150.000 nucleotídeos. Essas clivagens podiam ser realizadas com enzimas tais como *MluI*, *NruI* e *PvuI*, que reconhecem seqüências raras nos genomas de mamíferos. De modo alternativo, uma digestão parcial de DNA genômico pode ser realizada por meio da exposição do DNA a uma mistura de *EcoRI* e *EcoRI* metilase. Sob concentrações adequadas, a metilase protege um número suficiente de sítios de restrição fazendo com que as clivagens pela endonuclease sejam infreqüentes, o que, conseqüentemente, produz fragmentos de restrição bastante grandes (ver Capítulo 4). Ambos os métodos caracterizam-se por fragmentar o genoma em pedaços, mais ou me-



Seqüenciamento em ciclos

(A) Uma dupla fita de DNA-molde é adicionada a um tubo de ensaio contendo um único primer de DNA fita simples, *Taq* polimerase, deoxinucleotídeos (A, T, C e G) e dideoxinucleotídeos marcados com fluorescência (A, T, C e G, cada um dos quais marcado com um corante diferente). O primer foi desenhado de forma a anelar por complementaridade uma das extremidades da região a ser seqüenciada. (B) A reação é aquecida e a dupla fita do DNA-molde se separa. (C) A reação é resfriada, permitindo que o primer ligue-se por complementaridade ao DNA que se encontra sob a forma de fita simples. (D) A *Taq* polimerase liga-se à extremidade do primer e dá origem a uma nova fita complementar, por meio da incorporação de nucleotídeos que se encontram em solução. (E) Quando um dideoxinucleotídeo é incorporado, ocorre o término da extensão da nova fita. Tal ciclo é então repetido várias vezes.



Sequenciamento do genoma humano

nos como um disparo de pistola fragmenta o material atingido à queima roupa. Assim, as bibliotecas montadas por meio desses métodos para o sequenciamento ficaram conhecidas como sequenciamento por fragmentação.

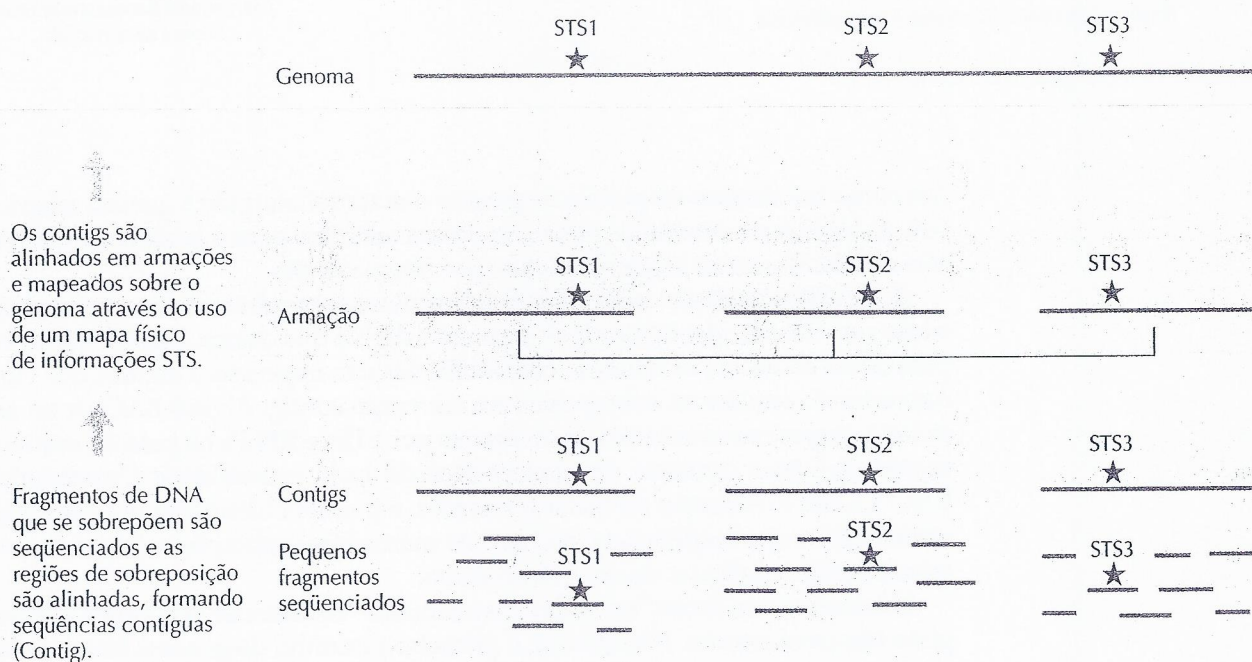
Os BACs individuais eram identificados pelos seus padrões característicos de restrição por *HindIII*, identificando a impressão digital dos clones. Um algoritmo computacional analisava as bandas compartilhadas das impressões digitais dos clones, convertendo sequências sobrepostas em um mapa contig. Alguns BACs eram ancorados a mapas cromossômicos já existentes por STSs e RFLPs ou pelo do sequenciamento de várias centenas de nucleotídeos de cada extremidade ("extremidades BAC"). Com a utilização de tais informações, é possível identificar um conjunto de clones BAC cujas sequências e padrões de marcadores sobrepõem-se, cobrindo extensas regiões contíguas de um cromossomo.

Idealmente, um único "caminho pavimentado" estender-se-ia cobrindo por completo um cromossomo. No entanto, o primeiro rascunho do genoma humano apresentava várias lacunas ou quebras. Em muitos casos, contigs adjacentes podem ser orientados uns em relação aos outros por meio das extremidades de BACs, das extremidades de plasmídeos, dos mRNAs conhecidos e dos mapas de informações para que se possa atingir a mais completa ordem de integração em grande escala, denominada de "andaimes" ou "armações". Como o nome indica, uma armação fornece

Seqüência humana total depositada no GenBank

Centro de seqüenciamento	Seqüência humana concluída (kb)
Center Sanger	284.353
Universidade de Washington – Genome/Sequencing/Center	175.279
US DOE Joint Genome Institute	78.486
Baylor College of Medicine – Human Genome Sequencing Center	53.418
Genoscope	48.808
Instituto Whitehead – Centro para a Pesquisa do Genoma	46.560
Department of Genome Analysis, Institute of Molecular Biotechnology	17.788
RIKEN Genomics Sciences Center	16.971
Universidade de Washington – Genome Center	14.692
Universidade Keio	13.058
Multimegabase Sequencing Center; Institute for Systems Biology	9.676
Universidade de Oklahoma – Advanced Center for Genome Technology	9.155
The Stanford Human Genome Center and Dep. of Genetics	9.121
Universidade do Texas, Southwestern Medical Center de Dallas	7.028
GTC Sequencing Center	7.014
Beijing Genomics Institute/Human Genome Center	6.297
Stanford Genome Technology Center	3.530
Max Planck Institute for Molecular Genetics	2.940
GBF – Geiman Research Center for Biotechnology	2.338
Laboratory Cold Spring Harbor, Genome Center Lita Annenberg Hazen	2.104
Outros	35.911
Total	842.027

Adaptada, com permissão, de International Human Genome Sequencing Consortium (2001. Initial sequencing and analysis of the human genome. *Nature* 409: 860-921). HTGS, seqüência de genoma de alta eficiência.



Anatomia da montagem de um genoma completo

(Adaptada, com permissão, de Venter J.C., Adams M.D., Myers E.W. et al. 2001. The sequence of the human genome. *Science* 291: 1304-1351.)

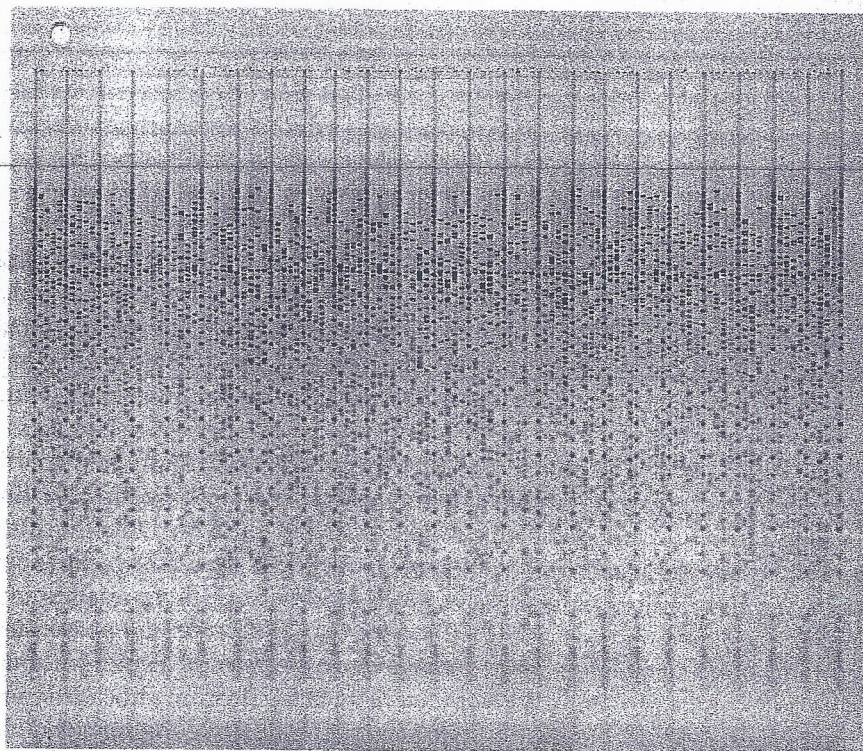


Foto de um gel mostrando um painel de impressões digitais de clones BAC
(Cortesia de James McPherson, School of Medicine, Washington University, St. Louis.)

uma moldura para a ligação de contigs adjacentes por meio de regiões cromossômicas não-seqüenciadas.

Para a obtenção da seqüência final, BACs foram selecionados a partir desse caminho pavimentado. Cada BAC sofreu digestão com enzimas de restrição para a produção de fragmentos de cerca de 1.500 nucleotídeos, sendo que tais fragmentos foram clonados em vetores de seqüenciamento. Considerando o tamanho do clone BAC original, um número suficiente de clones de seqüenciamento foi selecionado de tal forma que cada nucleotídeo do clone BAC estivesse representado de quatro a oito vezes nos novos clones. Tendo em vista a amplitude significativa de sobreposição, o seqüenciamento foi realizado a partir de uma ou de ambas as extremidades de cada clone, mas a porção central geralmente não foi seqüenciada por completo.

Os maiores centros de seqüenciamento funcionavam como verdadeiras fábricas, nas quais robôs isolavam automaticamente clones, purificavam amostras de DNA, preparavam reações de seqüenciamento e passavam as amostras em seqüenciadores automáticos de capilares.

Trabalhando em turno contínuo, esses centros podiam processar mais de 200.000 reações de seqüenciamento por dia. As seqüências cruas eram processadas e montadas usando os programas PHRED e PHRAP desenvolvidos em 1995 por Phil Green e Brent Ewing, da Universidade de Washington. De acordo com as normas estabelecidas no encontro de seqüenciamento do genoma público, nas Bermudas, em 1996, todas as seqüências estabelecidas, maiores de 1.000 pb (1 kb), eram depositadas em bancos de dados públicos dentro de um prazo de 24 horas. Pelo uso dos mapas de informação fornecidos a partir de impressões digitais de BAC, um programa chamado GigAssembler unia as seqüências desde arranjos locais até aproximadamente 5.000 contigs que formam uma seqüência genômica praticamente contínua, apelidada de "caminho dourado". O código do GigAssembler foi escrito em quatro semanas por Jim Kent, um aluno de graduação da Universidade da Califórnia, em Santa Cruz.

SEQÜÊNCIAS-ALVO EXPRESSAS

Os esforços iniciais de seqüenciamento de DNA mostraram com rapidez que cDNAs correspondentes a mRNAs transcritos (ver Capítulo 5) seriam muito importantes para a verificação de genes preditos por meio da seqüência genômica. A metodologia de seqüências-alvo expressas (EST, do inglês *expressed sequence tag*), desenvolvida em 1991 por J. Craig Venter, do National Institut of Mental Hearch, forneceu um sistema altamente eficiente para a identificação das regiões de DNA codificadoras de milhares de genes transcritos. Esse método começa com uma biblioteca de cDNA humano, criada pela transcrição reversa de mRNA proveniente de tecido cerebral, e prossegue por meio da clonagem dos cDNAs resultantes em partículas de fago. Pelo uso de um primer não-específico, que se liga a uma grande diversidade de seqüências, ele utiliza PCR para amplificar aleatoriamente fragmentos de 150 a 400 pb.

Cada fragmento é seqüenciado para criar um "alvo" característico para um gene expresso. A partir desse método, Venter identificou com rapidez mais de 3.000 ESTs característicos de cérebro humano. Uma busca em bancos de dados disponíveis mostrou que apenas 17% dos ESTs de cérebro apresentavam concordância com genes pré-identificados, ou seja, 83% dos ESTs representavam, potencialmente, genes desconhecidos.

O anúncio que o NIH pretendia patentear os genes identificados pelo método de seqüências-alvos expressas de Venter causou grande comoção entre os pesquisadores acadêmicos. Muitos ficaram chocados com as possíveis implicações do patenteamento de fragmentos de um gene cuja identidade e função biológica ainda eram desconhecidas. Uma atitude dessas seria mais condizente a uma companhia farmacêutica do que ao governo federal. O responsável pelo Projeto Genoma no NIH, James Watson, chegou a qualificar tal possibilidade de "insanidade absoluta". Seu desagrado em relação a esse ponto o levou a uma ruptura com o diretor do NIH, Bernardine Healy e, considerando outros pontos de desacordo, Watson retirou-se do Projeto Genoma. Sua vaga foi ocupada por Francis Collins, que havia construído sua reputação na Universidade de Michigan ao utilizar a clonagem posicional para a identificação do gene responsável pela fibrose cística.

O valor da abordagem de ESTs na separação "do joio do trigo" no genoma não passou despercebido pelos setores comerciais da biotecnologia. Várias companhias deram início a um processo de "mineração de genes" por meio do rápido estabelecimento de bibliotecas de ESTs de dúzias de tecidos normais e cancerosos. No final dos anos 90, tais companhias possuíam coleções de 50.000 a 100.000 ESTs específicos que, teoricamente, representavam a maioria dos genes do genoma humano. Por sua vez, essas empresas solicitaram patentes para centenas de ESTs de uso potencial no diagnóstico e na indústria farmacêutica.

Apesar de o NIH haver retirado sua aplicação de patentes para os ESTs de Venter, a questão levantou certas considerações que ainda figuram no centro da discussão a respeito do patenteamento de genes. Para receber uma patente, uma invenção deve ser nova e útil. Um EST pode estar de acordo com o critério de novidade, porém, sem que se tenha uma completa informação de sua seqüência, é virtualmente impossível demonstrar sua utilidade. Sendo assim, até o momento, o Registro de Patentes do Estados Unidos não permitiu o patenteamento de porções de genes com funções biológicas desconhecidas. Sob algumas circunstâncias, foram permitidas patentes de ESTs de genes humanos cuja função pôde ser inferida por homologia com um gene já descrito. Até o momento, nenhuma patente de EST foi levada ao âmbito jurídico, portanto ainda permanece difícil definir o que ocorrerá nesse campo.

SEQÜENCIAMENTO DO GENOMA COMPLETO POR FRAGMENTAÇÃO

O método de Venter a respeito dos ESTs foi importante sob uma óptica filosófica. Considerando que era possível a utilização de um método simples, em grande escala, para a identificação de genes expressos – e exclusão de DNA não-codificador –, es-

se método colocou em dúvida o sistema "telômero-a-telômero" utilizado pelo Projeto Genoma. Vinculado a outra instituição, *The Institute for Genomic Research* (TIGR), Venter dedicou-se a mostrar que o algoritmo usado na análise de ESTs poderia ser utilizado para a montagem de um genoma completo a partir de clones de fragmentação, sem a necessidade da construção inicial de mapas cromossômicos como os que haviam consumido a primeira fase do Projeto Genoma Público.

Tendo sido retirado do sistema de financiamento do Projeto Genoma do NIH em 1995, o grupo de Venter utilizou o seqüenciamento do genoma completo por fragmentação para determinar o genoma completo de 1,8 milhões de pares de bases (megapares de base Mbp) da bactéria *Haemophilus influenzae*. O grupo TIGR partiu então para o uso dessa metodologia no seqüenciamento de vários outros genomas microbianos.

Em 1997, apenas cerca de 5% do genoma humano havia sido completamente seqüenciado. James Weber, da Mansfield Medical Research Foundation, e Edward Myers, da Universidade do Arizona, propuseram o uso dessa metodologia visando a aceleração do processo do seqüenciamento humano. A estratégia estava baseada no seqüenciamento de ambas extremidades de insertos curtos (0,4 a 1,2 kb) e longos (5 a 20 kb) clonados em *E.coli*. "Extremidades pareadas lidas" de pequenos insertos que se sobrepõem forneceriam informação contínua de seqüência, ao passo que as extremidades lidas dos longos insertos forneceriam informação sobre espaçamento e orientação, favorecendo uma montagem de seqüências mais efetiva, sobretudo em regiões que contêm elementos repetitivos interespaçados. No entanto, a maioria dos cientistas tinha dúvidas sobre a possibilidade de desenvolver tal metodologia na escala necessária, ou seja, 150 vezes maior do que o seqüenciamento de um genoma microbiano. Além disso, os genomas microbianos praticamente não possuem DNA repetitivo e muitos estavam céticos em relação à capacidade dos algoritmos de montagem em lidar com a grande quantidade de DNA repetitivo existente em um genoma eucariote.

Assim, os pesquisadores genômicos ficaram impactados quando, em maio de 1998, Venter anunciou que estava deixando o grupo TIGR para dirigir uma nova companhia denominada Celera. Uma subsidiária da Applied Biosystems, a Celera usaria os mais recentes seqüenciadores por capilaridade para seqüenciar o genoma humano em três anos a um custo de 300 milhões de dólares. O projeto público respondeu com um cronograma acelerado para produzir um rascunho básico do genoma humano em 2001. O Wellcome Trust duplicou o financiamento do projeto público para 330 milhões de dólares, assumindo a responsabilidade de seqüenciar um terço do genoma no Center. No final de 1998, John Sulston, do Sanger, e Robert Waterston, da Universidade de Washington, noticiaram o primeiro seqüenciamento de um organismo eucariote multicelular, o nematódeo *C.elegans*.

Para testar se era factível a montagem completa do genoma de um organismo eucariote, a Celera, em conjunto com Gerald Rubin, do Projeto *Drosophila*, de Berkeley, desenvolveu o seqüenciamento de 120 Mbp da porção eucromatina do genoma da mosca-das-frutas no período de um ano. Apesar da seqüência não cobrir o genoma completo, o seqüenciamento foi considerado *finalizado* – as principais regiões que abrigavam genes haviam sido seqüenciadas. As regiões repetitivas não foram seqüenciadas. Várias lacunas grandes permaneceram e diversas questões de arranjos anômalos não foram resolvidas. A seqüência de *Drosophila* foi completada, com todas as lacunas preenchidas e todas as seqüências confirmadas, sob uma baixa taxa de erros, no final de 2002.

A questão do que se considera uma seqüência *finalizada* é de extrema importância e depende daquilo que o cientista pretende realizar com os dados de seqüenciamento. Por exemplo, para identificar um gene de doença, um mapa de alta qualidade com a localização do gene, bem como sua ordenação já é suficiente. Em contraste, estudos de evolução molecular (quantas alterações as seqüências de DNA sofrem, entre as espécies, ao longo do tempo) necessitam de seqüências muito mais acuradas e, nesse caso, as seqüências de regiões não-codificadoras são também importantes.

Desde o começo do Projeto Genoma, os cientistas perceberam a importância de obtenção de seqüências acuradas. Mesmo um erro em um único nucleotídeo pode ter um significativo impacto na interpretação de uma seqüência de DNA. Por exemplo, a inserção ou deleção de um nucleotídeo dentro de uma seqüência codificadora (éxon) altera a fase de leitura e induz a predição de uma seqüência de aminoácidos errada desde esse ponto até o final da região codificadora. Lembre-se também que a seqüência do genômica de cada indivíduo (com exceção de gêmeos monozigóticos idênticos) é única e específica. Uma alta fidelidade de seqüenciamento é necessária para acessar as verdadeiras diferenças de seqüências entre indivíduos.

No momento da implantação do Projeto Genoma, a acurácia média do seqüenciamento era de aproximadamente 99%, cerca de um erro a cada 100 nucleotídeos seqüenciados. A maior parte dos especialistas acreditava que seria necessário atingir uma acurácia consistente de seqüenciamento da ordem de 99,9%, que se traduz em um erro a cada 1.000 nucleotídeos. Se assumir que o tamanho médio de um éxon é igual a 200 nucleotídeos, mesmo sob essa taxa, haveria a chance de que ocorresse um erro a cada cinco éxons. Felizmente, o aumento exponencial na velocidade e a diminuição do custo de seqüenciamento foram acompanhados pelo aumento na acurácia. A lição é a mesma que tem sido observada em linhas de montagem de produção de bens de consumo em massa – quanto maior a mecanização, menor a chance de erros humanos.

COMPLETANDO A SEQÜÊNCIA DO GENOMA HUMANO

Armada com a experiência do seqüenciamento e montagem do genoma de *Drosophila*, a empresa Celera iniciou os trabalhos sobre o genoma humano. A maior parte do seqüenciamento do genoma completo por fragmentação foi concluída em de 10 meses, de setembro de 1999 a junho de 2000. Da mesma forma que o público, o projeto Celera partiu de DNA obtido com consentimento informado de múltiplos doadores anônimos de ambos sexos e de várias origens étnicas. Da mesma forma que o projeto público, a referida empresa ainda anunciou que o seqüenciamento fora realizado a partir de uma mistura de DNAs provenientes de um conjunto anônimo. No entanto, mais tarde, Craig Venter revelou que a maior parte do DNA seqüenciado pela Celera era, de fato, seu próprio material genético.

As bibliotecas plasmidiais foram criadas por meio de um processo que emprega a passagem forçada de uma solução de DNA através de uma seringa e a clonagem dos fragmentos que sofreram quebra mecânica em três classes: 2 kpb, 10 kpb e 50 kpb. O seqüenciamento das extremidades de cada inserto plasmidial deu origem a 27,27 milhões de seqüências lidas compostas, em média, de 543 nucleotídeos, totalizando 14,8 bilhões de pares de bases. Um total de 4,4 bilhões de pares de bases de dados de seqüências BAC provenientes do Projeto Genoma Público foram “fragmentadas” por computador para produzir leituras “falsas” de 550 pb cada. Esse conjunto combinado de dados foi analisado e montado por meio de dois métodos: (1) uma montagem do genoma completo por fragmentação para gerar de novo *contigs* e armações pelo pareamento por sobreposição entre leituras individuais de 550 pb e (2) uma montagem “compartimentalizada” usando o mapa do genoma público para ancorar montagens locais dos dados combinados. Dessa forma, pelo acesso às seqüências e aos mapas cromossômicos gerados pelo projeto público, o trabalho da Celera foi muito facilitado.

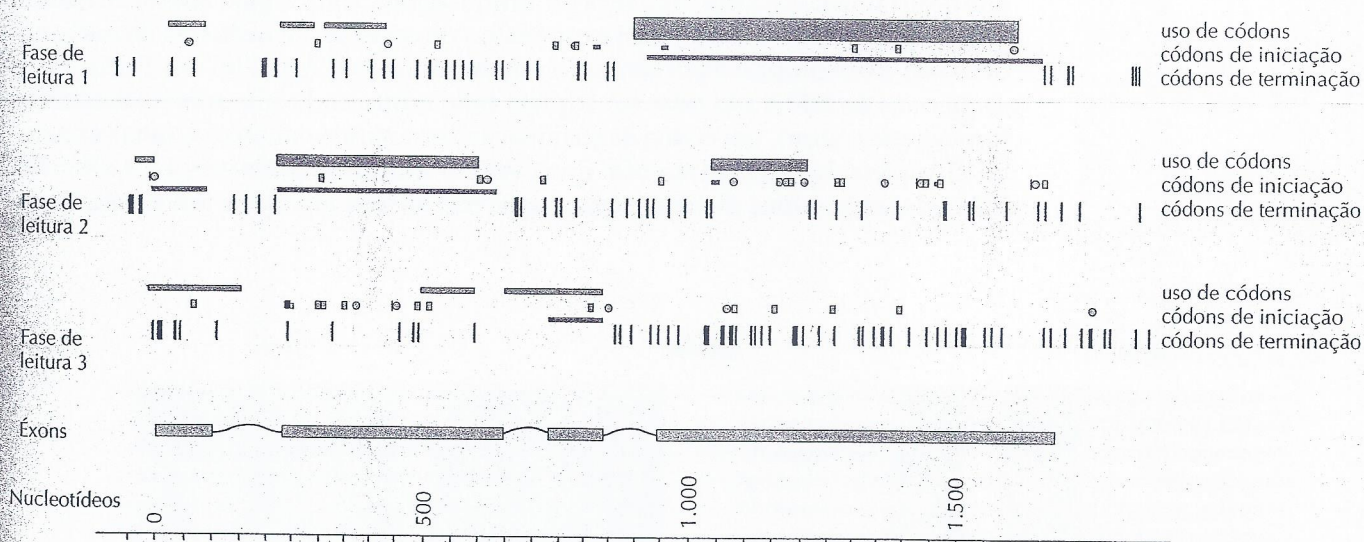
Durante a etapa final do seqüenciamento, em 2000, que levou à publicação do genoma humano, a Celera, com 65 funcionários, chegou a processar 175.000 amostras e seqüenciar 90 milhões de pares de bases por dia. Por um sistema de trabalho praticamente contínuo, cada aparelho de seqüenciamento automático necessitava de apenas aproximadamente 15 minutos de intervenção humana por dia. Os centros que possuíam financiamento público tinham, em conjunto, uma capacidade produtiva equi-

valente. Juntos, os dois projetos genoma competidores produziam mais de 2.000 nucleotídeos de seqüências brutas por segundo, completando o equivalente a um genoma humano inteiro em menos de três semanas. No final, mais de 95% das seqüências foram obtidas com uma acurácia de pelo menos 99,9%.

BIOINFORMÁTICA: DESVENDANDO A INFORMAÇÃO DE NOSSOS GENES

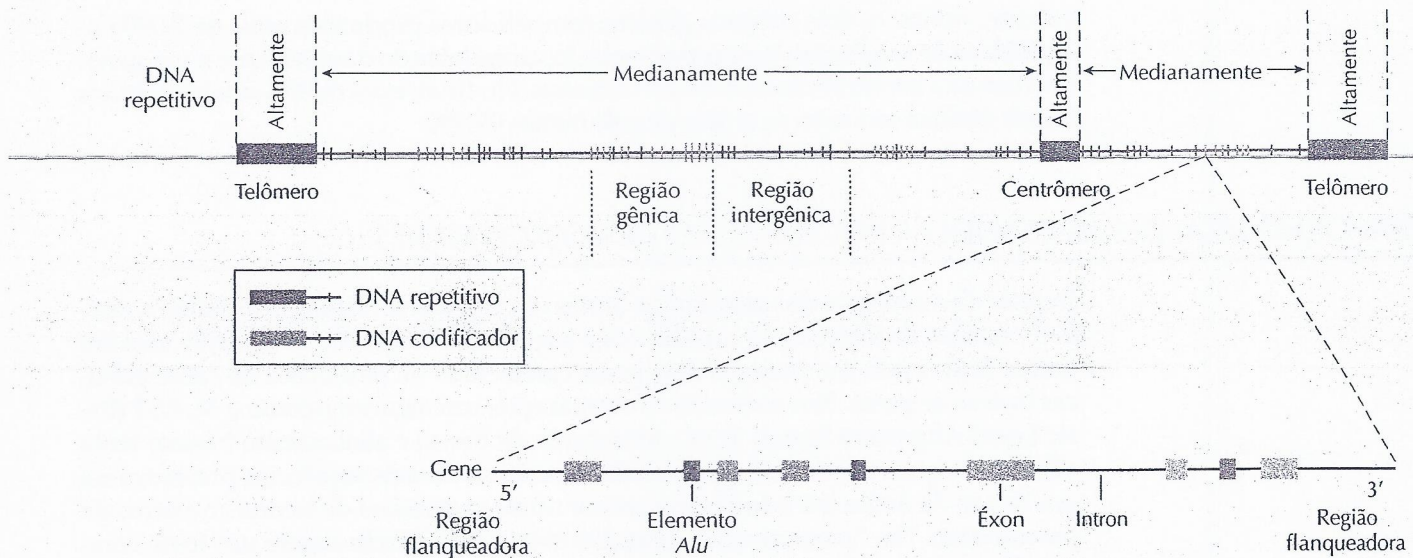
Depois da montagem das seqüências brutas em regiões contíguas, tem início a tarefa de análise da informação contida nessa seqüência. Os genes são identificados em dados brutos pela comparação com genes conhecidos ou pela busca de características típicas de genes. Nos métodos de comparação, um algoritmo como o BLAST (Basic Local Alignment Search Tool – ferramenta de busca e alinhamento básico) tenta alinhar a seqüência que está sendo analisada com as demais seqüências presentes em um banco de dados introduzindo o menor número possível de lacunas ou erros de pareamento. Um “pareamento” estatisticamente significado sugere um forte compartilhamento de propriedades químicas ou estruturais. Uma evidência direta da transcrição da seqüência de DNA (e, como resultado, que essa é uma seqüência codificadora) é fornecida pela identificação de similaridade e pareamento com um EST conhecido ou com uma seqüência de mRNA. A evidência indireta de que a seqüência pertence a um gene é o fato de a mesma apresentar homologia com outros genes ou proteínas de humanos ou de outras espécies. Para esse tipo de comparação, a quantidade crescente de dados provenientes do seqüenciamento paralelo de espécies-modelo – em especial de *E. coli*, *S. cerevisiae*, *C. elegans*, *S. pombe*, *Drosophila*, camundongo e *Arabidopsis* – originou importantes bancos de dados de genes previamente identificados.

Métodos de comparação, como os agora descritos, conduziram à identificação de cerca de 60% dos genes encontrados na primeira versão do genoma humano. Os 40%



Predição de um gene, *ab initio*, em três fases de leitura

Algoritmos de busca e identificação de genes procuram sinais de éxons codificadores nas três fases de leitura possíveis em cada fita de uma molécula de DNA. O uso de códons identifica padrões de seqüências consistentes com o uso não-aleatório de códons específicos em diferentes organismos. A ausência de códons de terminação define uma fase aberta de leitura – e um éxon em potencial. Um códon de iniciação seguido de uma fase aberta de leitura define o primeiro éxon. O mapa genético (na parte inferior) mostra que os dois primeiros éxons encontram-se na fase de leitura 2, seguidos de um éxon na fase de leitura 3 e de um éxon na fase de leitura 1.



Estrutura de um cromossomo humano

O DNA altamente repetitivo inclui os microssatélites (2-5 pb) e os minissatélites (6-50 pb). A repetição telomérica é um exemplo de minissatélite. O DNA medianamente repetitivo inclui SINEs (75-500 pb), LINEs (até 6,5 kb) e genes de RNA ribossomal. A sequência *Alu* é um exemplo de SINE.

remanescentes foram identificados *ab initio* – desde o começo – por meio da busca de uma estrutura na sequência genômica. Em alguns casos, algoritmos computacionais eram usados para procurar, em “janelas” relativamente grandes de sequências, um padrão associado com frequência a regiões que contêm genes, tais como “ilhas CpG”, com uma proporção relativamente alta de dinucleotídeos CG. Tais ilhas têm sido associadas a sequências reguladoras transcricionais. Em outros casos, a busca centrou-se em sequências específicas que definem os vários elementos funcionais dos genes.

O método clássico de identificação de um gene potencial é a busca por códonos que determinem uma fita específica de amino-ácidos. Uma região aberta de leitura (orf, do inglês *open reading frame*) é um trecho de centenas ou milhares de nucleotídeos não-interrompidos por um códon de terminação (TAA, TAG ou TGA). Se as sequências de DNA são lidas em triplets (não esquecendo que três nucleotídeos formam um códon), um códon de terminação deve surgir, em média, uma vez a cada 21 triplets. Lembre-se também que existem 64 triplets possíveis $4 \times 4 \times 4 = 64$. Três deles são códonos de terminação, o que nos dá uma média de um a cada 21 có-

Éxon	Íntron	Éxon
~cgatgccaccgagggcatcag	GTttgttagacgaggcctaac	~ctctttccattgtgtttcag
~atpctggaagcagtatgttg	GTaagctgcataattgaaat	~tgctgtctcgggtctctcag
~gccacgcgtaccccgagag	GTgagtgccagctggggtc	~ctcactgcctttatttccag
~tgccggaaataaaacataag	GTaatcacttgctcgggtgt	~ctcctttgtctgccaatgcag
~tgataggcgtgtttttaaag	GTaagtagagcctcaaattt	~ctccgtctctcaccctgcag
~gtcaccaccccgccgctga	GTgagccgtgccgtaactg	~ctggctgcccctctattctag
~aggaaagaactgtgaattag	GTaagaggttcagagtggct	~taactcttccatatctccag
~tgaagaagaacatgaggag	GTgagtgccagctggggtc	~ctcactgcctttatttccag
		gtatgagacggaggtgtccc

Sítio de *splicing*

Sítio de *splicing*

Sequências consenso das junções de *splicing* íntron/éxon

Os limites entre éxons e íntrons são definidos por sequências consenso contendo quatro nucleotídeos invariáveis (GT e AG) que determinam cada uma das extremidades de um íntron. Uma sequência consenso no interior do íntron é também utilizada para a predição do mesmo. Mesmo o melhor dos algoritmos faz muitas predições falso-positivas de junções de *splicing*.

dons (ou 63 nucleotídeos). Esse método funciona muito bem em genes bacterianos, os quais quase sempre iniciam por um códon de iniciação (ATG) e seguem continuamente sobre uma mesma fase de leitura até sua extremidade final.

Pode ocorrer uma complicação adicional em procariotes e vírus, pois algumas vezes tais organismos apresentam sobreposição de genes, em que são usadas ao mesmo tempo duas ou mais das seis possíveis fases de leitura de uma determinada região de dupla fita do DNA. Isso ocorre porque os vírus sobretudo possuem genomas compactos.

Os genes eucariotes raramente apresentam sobreposição ou ocupam posições em fitas opostas da molécula de DNA. No entanto, a presença de íntrons faz com que as fases abertas de leitura sejam muito menos críveis como preditoras de genes. Se considerarmos um gene "médio" humano identificado pelo Projeto Genoma, ter-se-á uma estrutura com oito éxons de cerca de 135 pb de comprimento cada um, além de sete íntrons com aproximadamente 2.200 pb. No interior dos íntrons, a presença de códons de terminação não apresentam importância funcional e os mesmos são identificados na proximidade na taxa esperada de ocorrência ao acaso. Entretanto, a fase aberta de leitura de um éxon pode se estender para o interior de um íntron, tornando praticamente impossível a determinação do local exato de início ou fim de um éxon. Assim, um gene eucariote característico está composto de várias seqüências codificadoras relativamente pequenas separadas por numerosos íntrons.

Um certo número de elementos gênicos apresenta "seqüências-consenso" características. Apesar de não serem idênticas, as seqüências consenso compartilham uma série de nucleotídeos invariáveis inseridos entre nucleotídeos variáveis. Por exemplo, nucleotídeos GT e AT flanqueiam cada uma das extremidades de um íntron, mas os limites entre íntron/éxon são ainda especificados por uma seqüência consenso maior no sítio de *splicing* e por um sinal mais difuso que ocorre no interior do íntron. O "TATA box", de seqüência consenso TATAAAA, marca o sítio de ligação da RNA polimerase na maioria dos promotores eucariotes. As extremidades terminais de genes são marcadas por um sinal de poliadenilação, AATAAA, que orquestra a adição da cauda de poli(A) encontrada na maioria dos mRNAs eucariotes. Usadas em combinação com as fases abertas de leitura, tais seqüências-sinais-consenso podem auxiliar a identificação de muitos genes.

Em conjunto, os vários elementos funcionais de um gene diminuem a aleatoriedade da seqüência de DNA e criam um direcionamento no sentido da utilização de determinadas combinações de nucleotídeos. Apesar dessas características gerais de seqüências não poderem ser distinguidas com facilidade segundo uma série de regras simples definidas, modelos "hidden Markov" (HMM) incorporam informações estatísticas a respeito de sítios de *splicing*, uso não-aleatório do código genético, bem como de tamanhos de éxons e íntrons de genes conhecidos para a identificação de novos genes.

No nematódeo e na mosca-das-frutas, os HMMs podem corretamente identificar aproximadamente 90% de éxons individualmente, podendo identificar todos os éxons de um dado gene em aproximadamente 40% dos genes. No entanto, no DNA humano tais índices se reduzem para respectivamente 70 e 20%. Os programas de predição de genes são muito eficientes na identificação de éxons internos, que possuem duas junções de *splicing* (à direita e à esquerda) adjacentes a dois íntrons. O primeiro e o último éxon freqüentemente não são identificados, pois apresentam apenas metade das informações de seqüências usadas para a inferência. Além disso, a inferência de éxons gera um grande número de falso-positivos. As seqüências de sítios de *splicing*-consenso são bastante comuns em íntrons, o que torna a existência de "pseudo-éxons" muito mais comum que a existência dos éxons genuínos.

