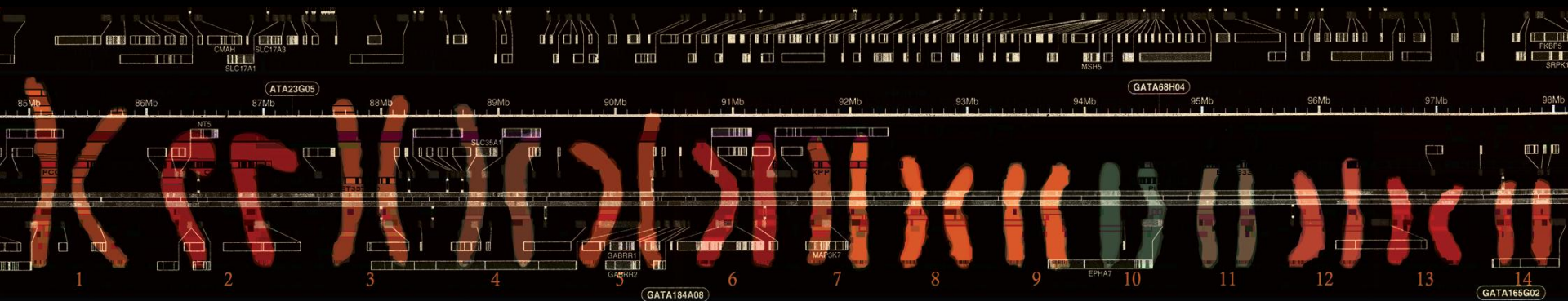
O GENOMA HUMANO



Sequenciamento
Montagem
Anotação

Conceitos

Sequência bruta: sequências de nucleotídeos originadas de cada inserto clonado (*reads*)

Sequências de final pareado: leituras obtidas das duas extremidades de um fragmento de DNA (*mates*)

Cobertura: número médio de vezes que um nucleotídeo é representado em um conjunto de leituras aleatórias

Clone BAC: vetor BAC (Bacterial artificial chromosome) contendo um inserto de DNA genômico, geralmente de 100 a 200 kb.

Conceitos

Assembly: processo de montagem do genoma

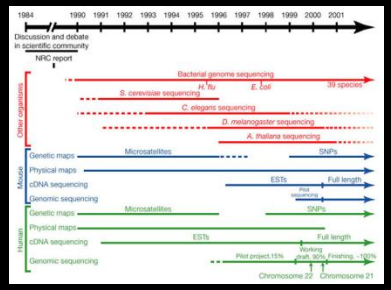
Contig: o resultado, contínuo, da sobreposição de um conjunto de sequências

Scaffold: resultado da conexão dos contigs, considerando ordem e orientação

Mapa Genético: mapa genômico no qual um locus é posicionado em relação a outro com base na frequência de recombinação. A unidade de distância é o centimorgan (cM) denotando 1% de chance de recombinação

A idéia do sequenciamento foi primeiramente proposta em encontros cientificos

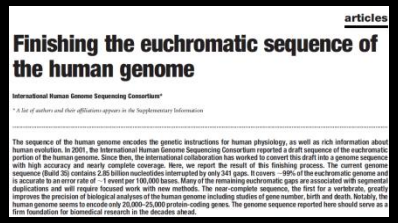
1984-1986



1995

Fim do sequenciamento dos outros organismos

1999


articles

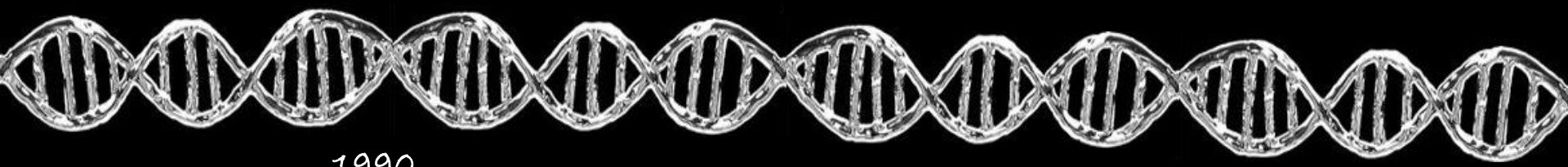
Finishing the euchromatic sequence of the human genome

International Human Genome Sequencing Consortium*

* A list of authors and their affiliations appears in the Supplementary Information

The sequence of the human genome encodes the genetic instructions for human physiology, as well as rich information about human evolution. In 2001, the International Human Genome Sequencing Consortium reported a draft sequence of the euchromatic portion of the human genome. Since then, the international collaboration has worked to convert this draft into a genome sequence with high accuracy and nearly complete coverage. Here, we report the result of this finishing process. The current genome sequence (Build 35) contains 2.85 billion nucleotides interrupted by only 341 gaps. It covers ~99% of the euchromatic genome and is accurate to an error rate of ~ 1 event per 100,000 bases. Many of the remaining euchromatic gaps are associated with segmental duplications and will require focused work with new methods. The near-complete sequence, the first for a vertebrate, greatly improves the precision of biological analyses of the human genome including studies of gene number, birth and death. Notably, the human genome seems to encode only 20,000–25,000 protein-coding genes. The genome sequence reported here should serve as a firm foundation for biomedical research in the decades ahead.

2004



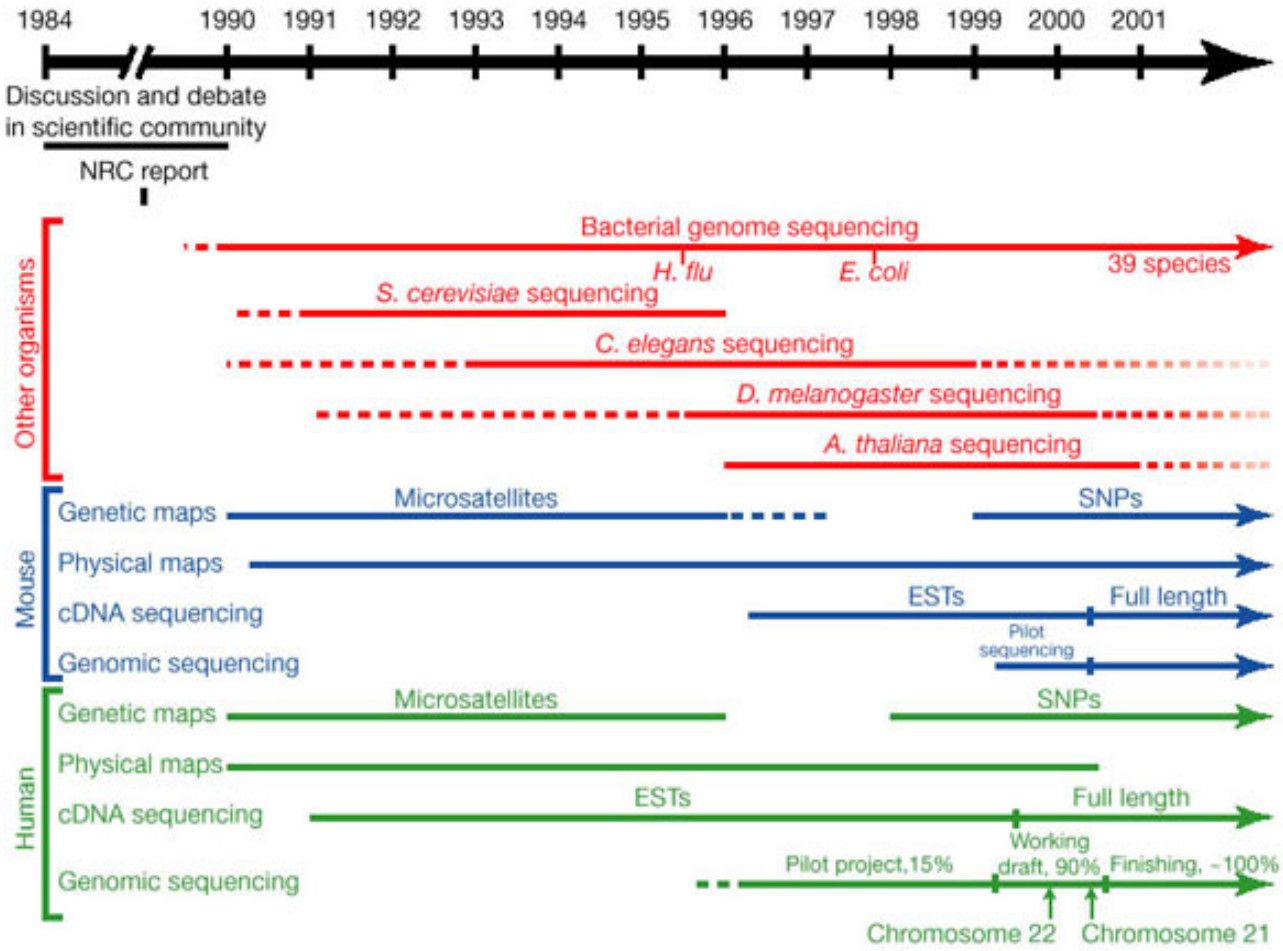
1990

O Projeto Genoma Humano foi lançado com a criação de centros de genoma nos seguintes países: EUA, Reino Unido, Japão, França, Alemanha e China

1998

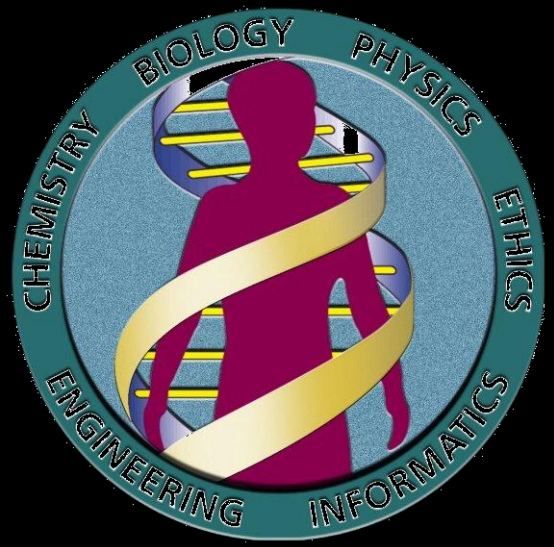
CELERA

2001


1984
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
Discussion and debate in scientific community
NRC report
Other organisms
Bacterial genome sequencing
H. flu
E. coli
39 species
S. cerevisiae sequencing
C. elegans sequencing
D. melanogaster sequencing
A. thaliana sequencing
Mouse
Genetic maps
Microsatellites
SNPs
Physical maps
cDNA sequencing
ESTs
Full length
Pilot sequencing
Genomic sequencing
Human
Genetic maps
Microsatellites
SNPs
Physical maps
cDNA sequencing
ESTs
Full length
Working draft, 90%
Pilot project,15%
Finishing, ~100%
Genomic sequencing
Chromosome 22
Chromosome 21

determinar a sequencia completa de nucleotídeos do genoma humano

Consórcio Internacional do Sequenciamento do Genoma Humano



CANADA
UNITED STATES
RUSSIA
CHINA
BRAZIL
AUSTRALIA
INDIA
Antarctica

Projeto patrocinado pelo Instituto Nacional de Saúde dirigido por Francis Collins e pelo Departamento de Energia - EUA

Projeção de duração de 15 anos
Duração total de 13 anos (1990–2003)

Planejamento de custo de $3 bilhões

Maior genoma sequenciado até agora (2001)

1º genoma de veterbrados sequenciado

Curto período de realização: de 10% a 90% em 15 meses

Genoma Humano é formado por 3 bilhões de pares de bases

Parecem haver de 30.000 a 40.000 genes codificadores de proteínas
somente 2 vezes mais que em vermes ou moscas
genes mais complexos: splicing gera maior numero de proteínas

O conjunto de proteínas codificado pelo genoma humano é mais complexo do que em invertebrados

Mais de 1.4 milhões de SNPs tem sido identificados

Informações do projeto e descrição da geração, montagem e avaliação da sequencia do genoma

Análise da sequencia: aspectos cromossomais, elementos repetidos, as diferenças e similaridades entre genes e proteínas, a história dos segmentos genômicos

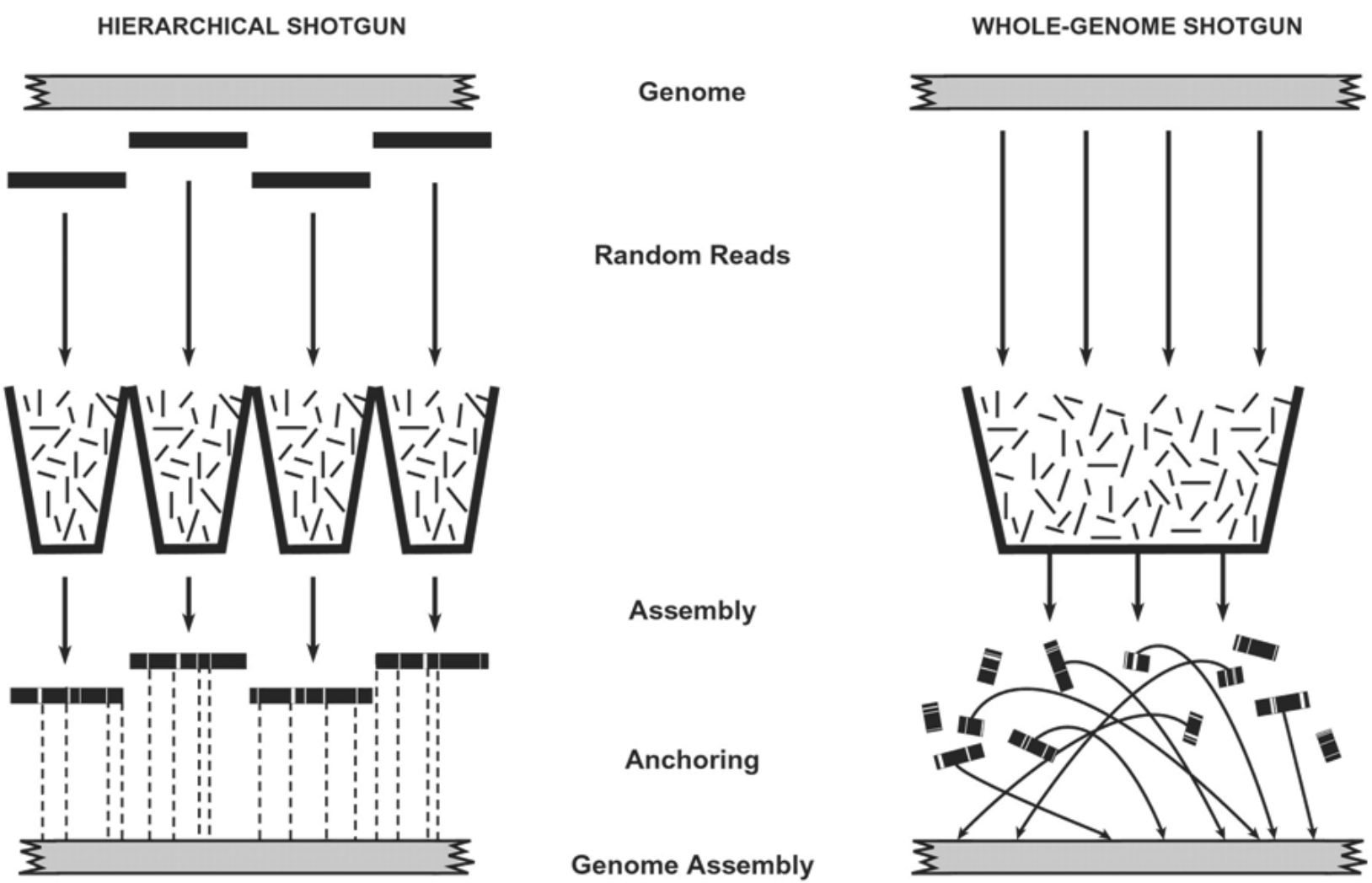
Discussão de aplicações da sequencia para a biologia e medicina e descrição dos próximos passos do projeto

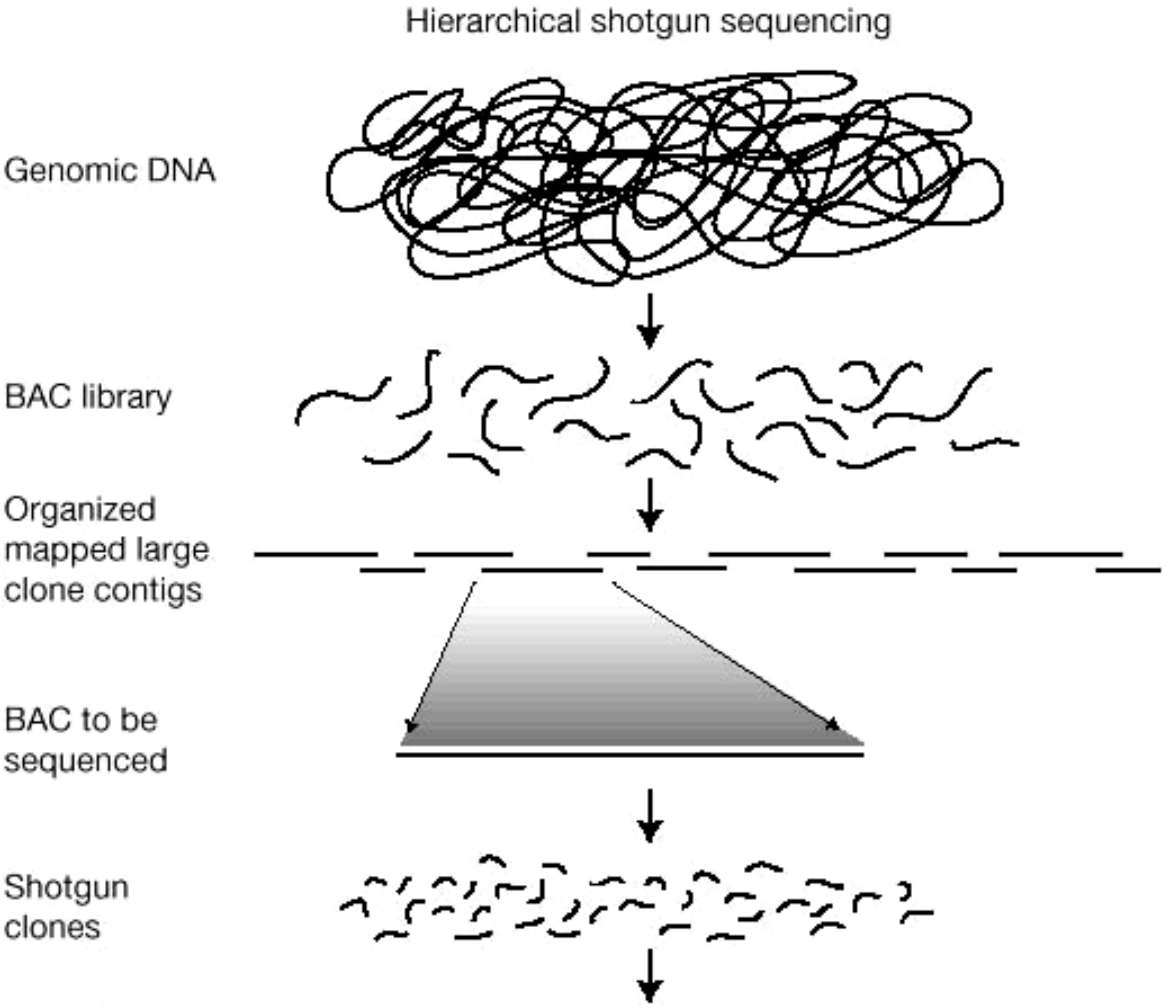
SEQUENCIAMENTO DO GENOMA HUMANO EM 3 PASSOS

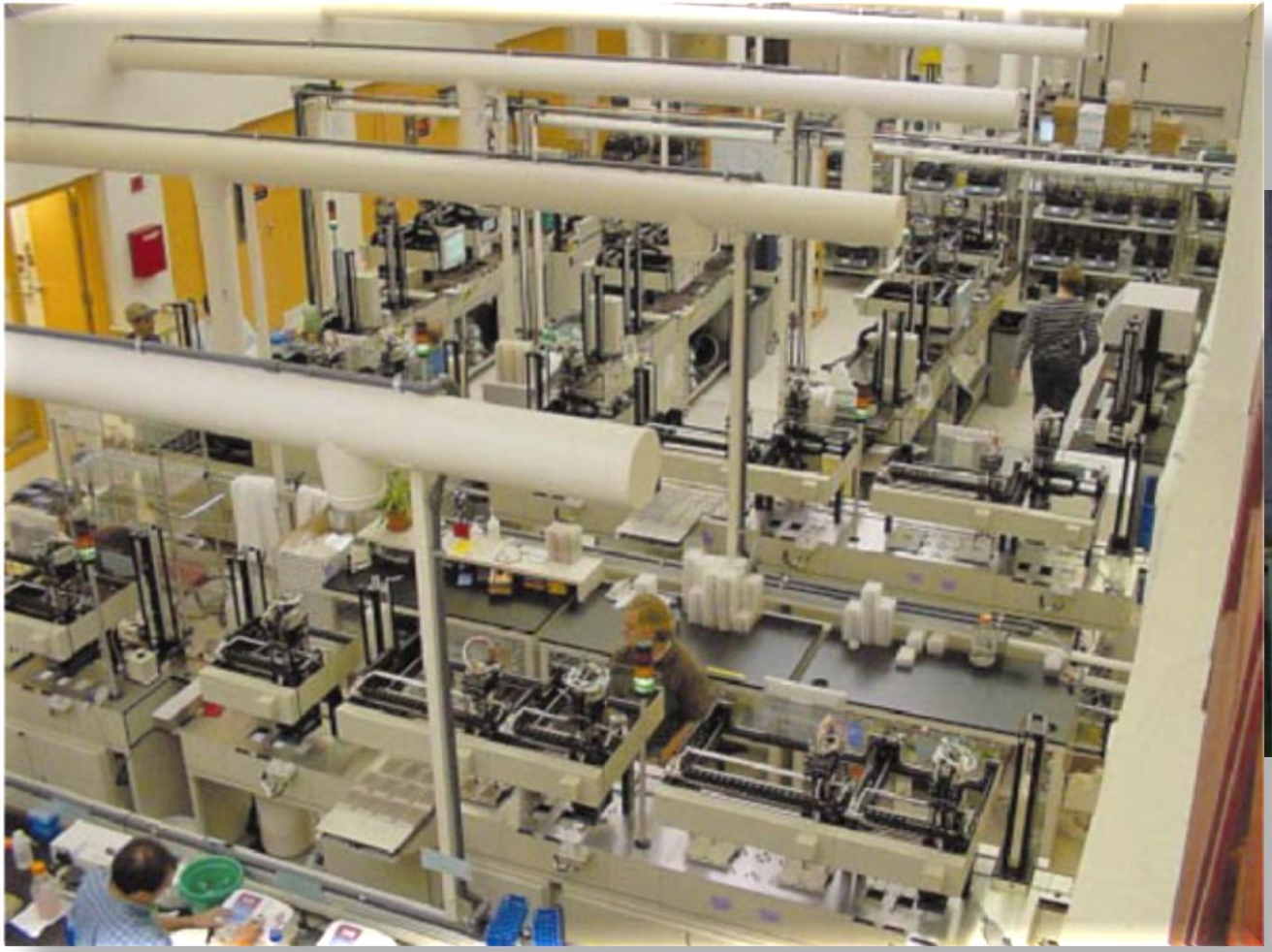
Seleção de clones BAC para serem sequenciados

Sequenciamento dos clones

Montagem dos clones sequenciados em uma sequencia total do genoma

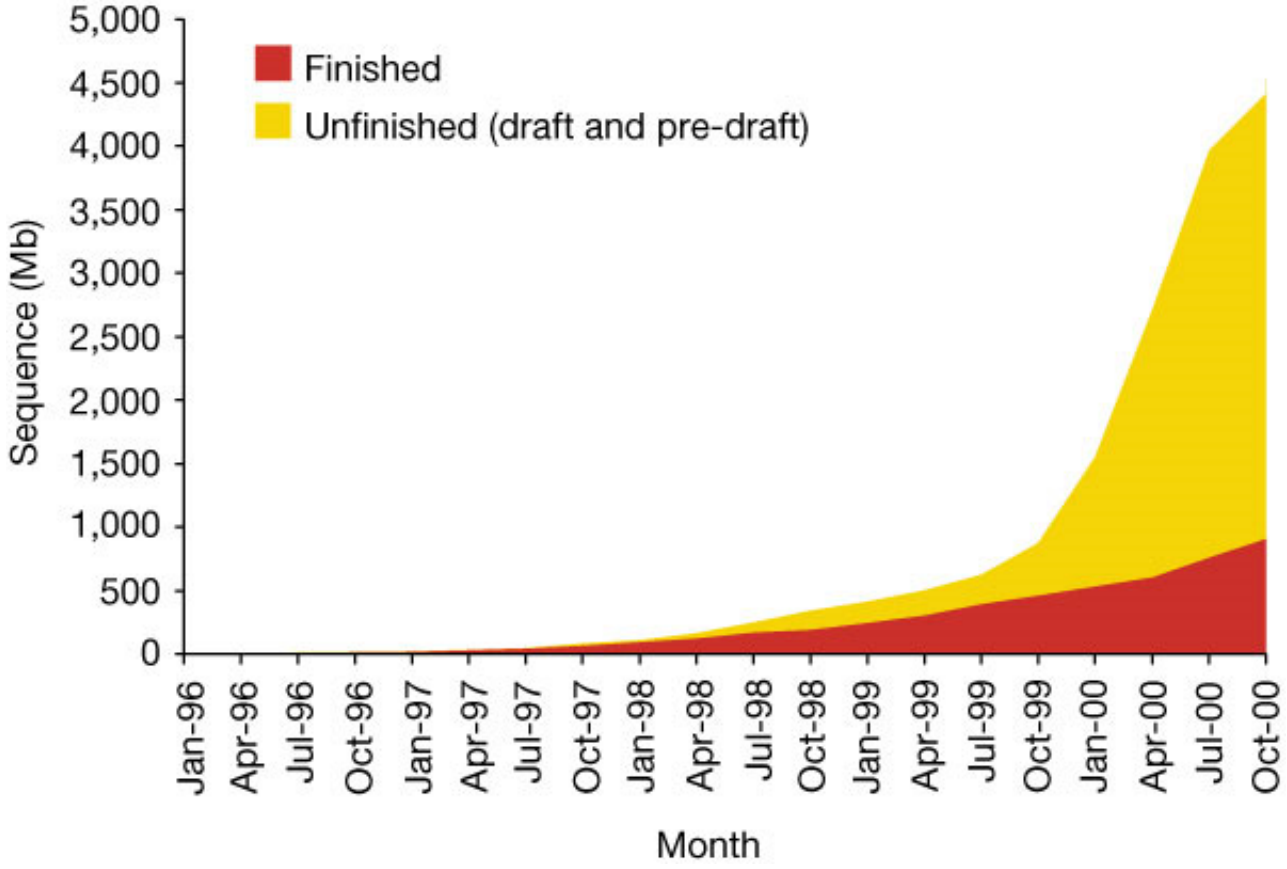

HIERARCHICAL SHOTGUN
WHOLE-GENOME SHOTGUN
Genome
Random Reads
Assembly
Anchoring
Genome Assembly


Hierarchical shotgun sequencing
Genomic DNA
BAC library
Organized mapped large clone contigs
BAC to be sequenced
Shotgun clones
Shotgun sequence
...ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG...
Assembly
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...

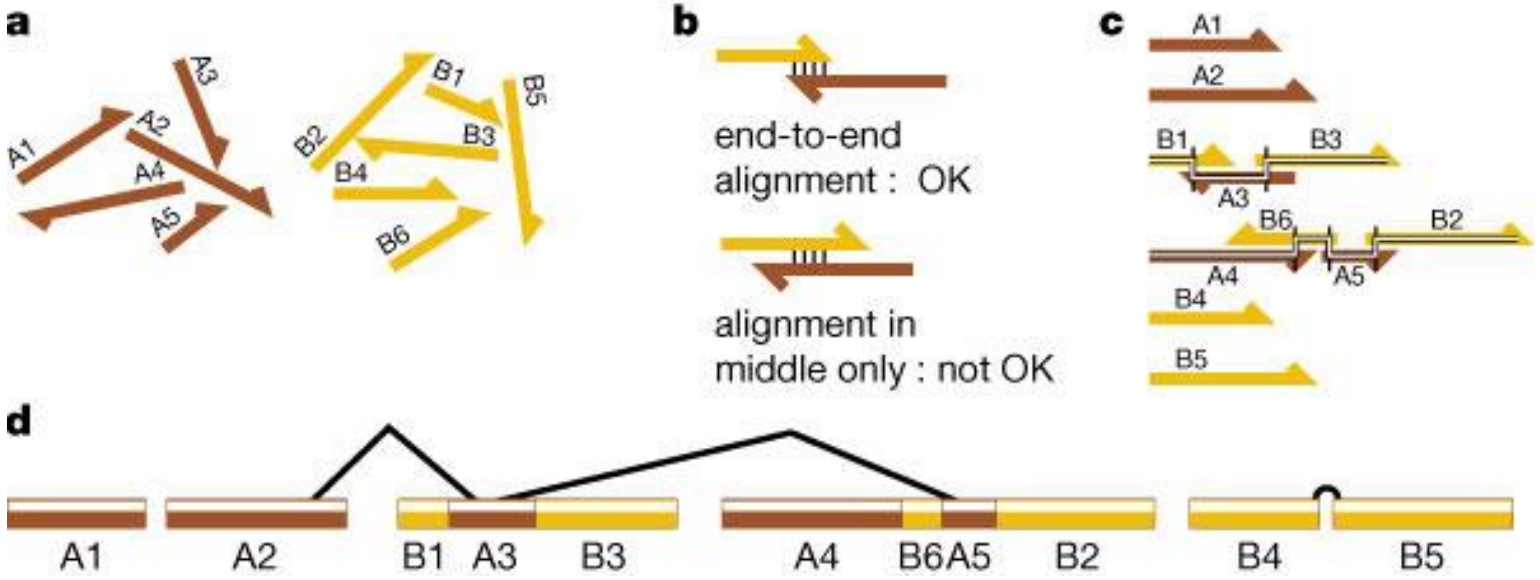


Cambridge, MA - EUA

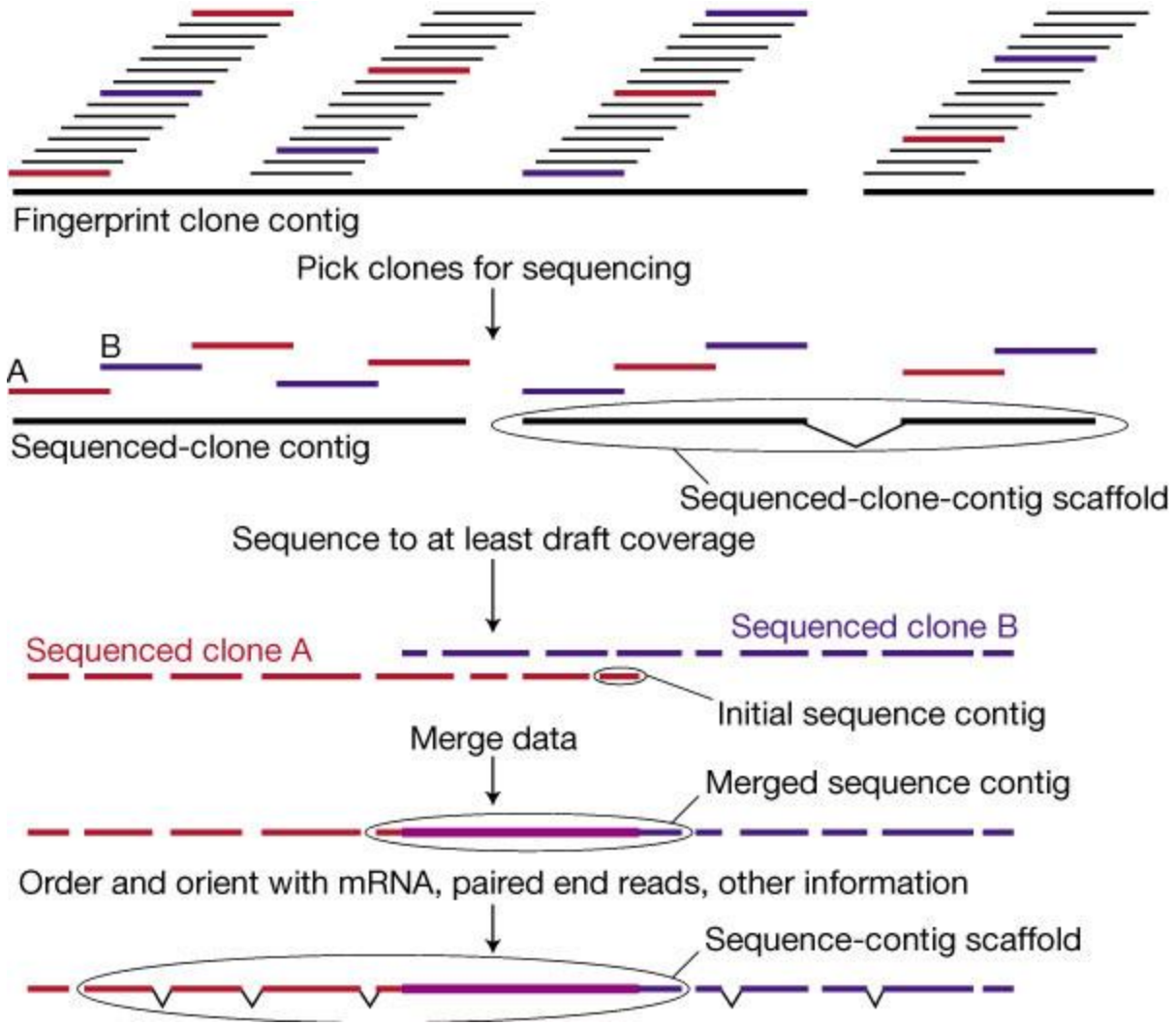
O grau de automação dos centros de pesquisa do Projeto Genoma Humano variou muito. A automação mais agressiva foi capaz de processar mais de 100.000 reações de sequenciamento em 12 horas.



Em Junho de 2000, os centros estavam produzindo sequencias em uma taxa equivalente a cobertura de uma vez todo o genoma em menos de 6 semanas. Isso corresponde a 1000 nucleotídeos por segundo, 24 horas por dia, 7 dias por semana.



Montagem das sequencias dos clones com programa GigAssembler.



contigs

scaffolds

Table 8 Chromosome size estimates

Chromosome*	Sequenced bases† (Mb)	FCC gaps‡		SCC gaps‖		Sequence gaps#		Heterochromatin and short arm adjustments**(Mb)	Total estimated chromosome size (including artefactual duplication in draft genome sequence)†† (Mb)	Previously estimated chromosome size‡‡ (Mb)
		Number	Total bases in gaps§ (Mb)	Number	Total bases in gaps¶ (Mb)	Number	Total bases in gaps[ref] (Mb)			
All	2,692.9	897	152.0	4,076	142.7	145,514	80.6	212	3,289	3,286
1	212.2	104	17.7	347	12.1	[illegible]	[illegible]	30	279	263
2	221.6	50	8.5	296	10.4	12,880	7.1	3	251	255
3	186.2	71	12.1	336	11.8	14,689	8.1	3	221	214
4	168.1	39	6.6	343	12.0	12,768	7.1	3	197	203
5	169.7	46	7.8	337	11.8	10,304	5.7	3	198	194
6	158.1	15	2.6	275	9.6	5,225	2.9	3	176	183
7	146.2	27	4.6	195	6.8	4,338	2.4	3	163	171
8	124.3	41	7.0	249	8.7	8,692	4.8	3	148	155
9	106.9	19	3.2	122	4.3	6,083	3.4	22	140	145
10	127.1	14	2.4	163	5.7	8,947	5.0	3	143	144
11	128.6	29	4.9	193	6.8	8,279	4.6	3	148	144
12	124.5	26	4.4	168	5.9	8,226	4.6	3	142	143
13	92.9	12	2.0	115	4.0	5,065	2.8	16	118	114
14	86.9	13	2.2	40	1.4	775	0.4	16	107	109
15	73.4	18	3.1	104	3.6	5,717	3.2	17	100	106
16	73.1	55	9.4	102	3.6	4,757	2.6	15	104	98
17	72.8	41	7.0	95	3.3	4,261	2.4	3	88	92
18	72.9	22	3.7	113	4.0	4,324	2.4	3	86	85
19	55.4	49	8.3	108	3.8	2,344	1.3	3	72	67
20	60.5	7	1.2	33	1.2	469	0.3	3	66	72
21	33.8	4	0.1	0	0.0	0	0.0	11	45	50
22	33.8	10	1.0	0	0.0	0	0.0	13	48	56
X	127.7	141	24.0	182	6.4	4,282	2.4	3	163	164
Y	21.8	6	1.0	19	0.7	113	0.1	27	51	59
NA	5.1	0	0	134	0.0	577	0.3	0	0	0
UL	9.3	38	0	7	0.0	566	0.3	0	0	0

* NA, sequenced clones that could not be associated with fingerprint clone contigs. UL, clone contigs that could not be reliably placed on a chromosome.
† Total number of bases in the draft genome sequence, excluding gaps. Total length of scaffold (including gaps contained within clones) is 2.916 Gb.
‡ Gaps between those fingerprint clone contigs that contain sequenced clones excluding gaps for centromeres.
§ For unfinished chromosomes, we estimate an average size of 0.17 Mb per FCC gap, based on retrospective estimates of the clone coverage of chromosomes 21 and 22. Gap estimates for chromosomes 21 and 22 are taken from refs 93, 94.
‖ Gaps between sequenced-clone contigs within a fingerprint clone contig.
¶ For unfinished chromosomes, we estimate sequenced clone gaps at 0.035 Mb each, based on evaluation of a sample of these gaps.
# Gaps between two sequence contigs within a sequenced-clone contig.
[ref] We estimate the average number of bases in sequence gaps from alignments of the initial sequence contigs of unfinished clones (see text) and extrapolation to the whole chromosome.
** Including adjustments for estimates of the sizes of the short arms of the acrocentric chromosomes 13, 14, 15, 21 and 22 (ref. 105), estimates for the centromere and heterochromatic regions of chromosomes 1, 9 and 16 (refs 106, 107) and estimates of 3 Mb for the centromere and 24 Mb for telomeric heterochromatin for the Y chromosome[108].
†† The sum of the five lengths in the preceding columns. This is an overestimate, because the draft genome sequence contains some artefactual sequence owing to inability to correctly to merge all underlying sequence contigs. The total amount of artefactual duplication varies among chromosomes; the overall amount is estimated by computational analysis to be about 100 Mb, or about 3% of the total length given, yielding a total estimated size of about 3,200 Mb for the human genome.
‡‡ Including heterochromatic regions and acrocentric short arm(s)[108].

O rascunho do genoma

foi concluído

com uma taxa de erro

de aproximadamente

1 em 10.000 bases

Conteúdo Gênico

As sequências codificantes compreendem apenas uma pequena porcentagem do genoma e uma média de cerca de 5% de cada gene

Genes parecem apresentar mais splicing alternativos

ncRNAs

Estimativa de 70,000-80,000 genes pelo número de ilhas CpG e a frequência da associação com genes conhecidos

Estudos utilizam vários tipos de dados, como ESTs, mRNAs de genes conhecidos, comparações entre genomas e análise de cromossomos sequenciados. Estimativas baseadas em ESTs tem variado muito, de 35.000 a 120.000 genes

Análises mais rigorosas excluem ESTs que aparecem uma única vez em bancos de dados, e calibram sensibilidade e especificidade produzindo estimativas de aproximadamente 35.000 genes.

Atualmente os métodos de predição utilizam combinções de três abordagens: evidência direta de transcrição fornecida por ESTs ou mRNAs, evidência indireta baseada na similaridade com genes ou proteínas previamente identificados e reconhecimento de exons por modelos computacionais e estatísticos.

Melhor estimativa concluída em ~ 24.500 genes

Finishing the euchromatic sequence of the human genome

International Human Genome Sequencing Consortium*

* *A list of authors and their affiliations appears in the* Supplementary Information

The sequence of the human genome encodes the genetic instructions for human physiology, as well as rich information about human evolution. In 2001, the International Human Genome Sequencing Consortium reported a draft sequence of the euchromatic portion of the human genome. Since then, the international collaboration has worked to convert this draft into a genome sequence with high accuracy and nearly complete coverage. Here, we report the result of this finishing process. The current genome sequence (Build 35) contains 2.85 billion nucleotides interrupted by only 341 gaps. It covers ~99% of the euchromatic genome and is accurate to an error rate of ~1 event per 100,000 bases. Many of the remaining euchromatic gaps are associated with segmental duplications and will require focused work with new methods. The near-complete sequence, the first for a vertebrate, greatly improves the precision of biological analyses of the human genome including studies of gene number, birth and death. Notably, the human genome seems to encode only 20,000–25,000 protein-coding genes. The genome sequence reported here should serve as a firm foundation for biomedical research in the decades ahead.

Número de gaps reduzido de ~150.000 para 341

Taxa de erro diminuída de 1 a cada 10.000 bases para 1 a cada 100.000 bases

Total de pares de bases de 3 bilhões para 2,851,330,913

Número de genes entre 20.000 a 25.000

A versão atual do catálogo de genes consta de 19.438 genes conhecidos e 2.188 genes preditos

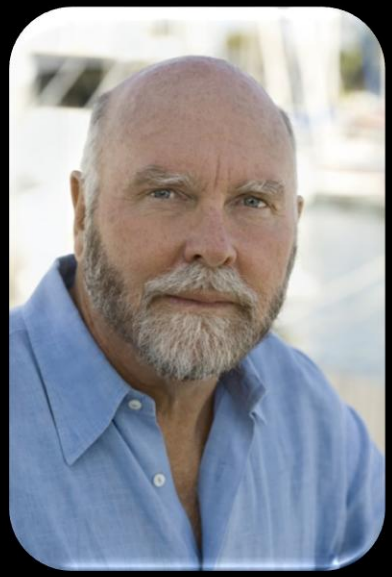
Cobertura da cromatina de 90% para 99%

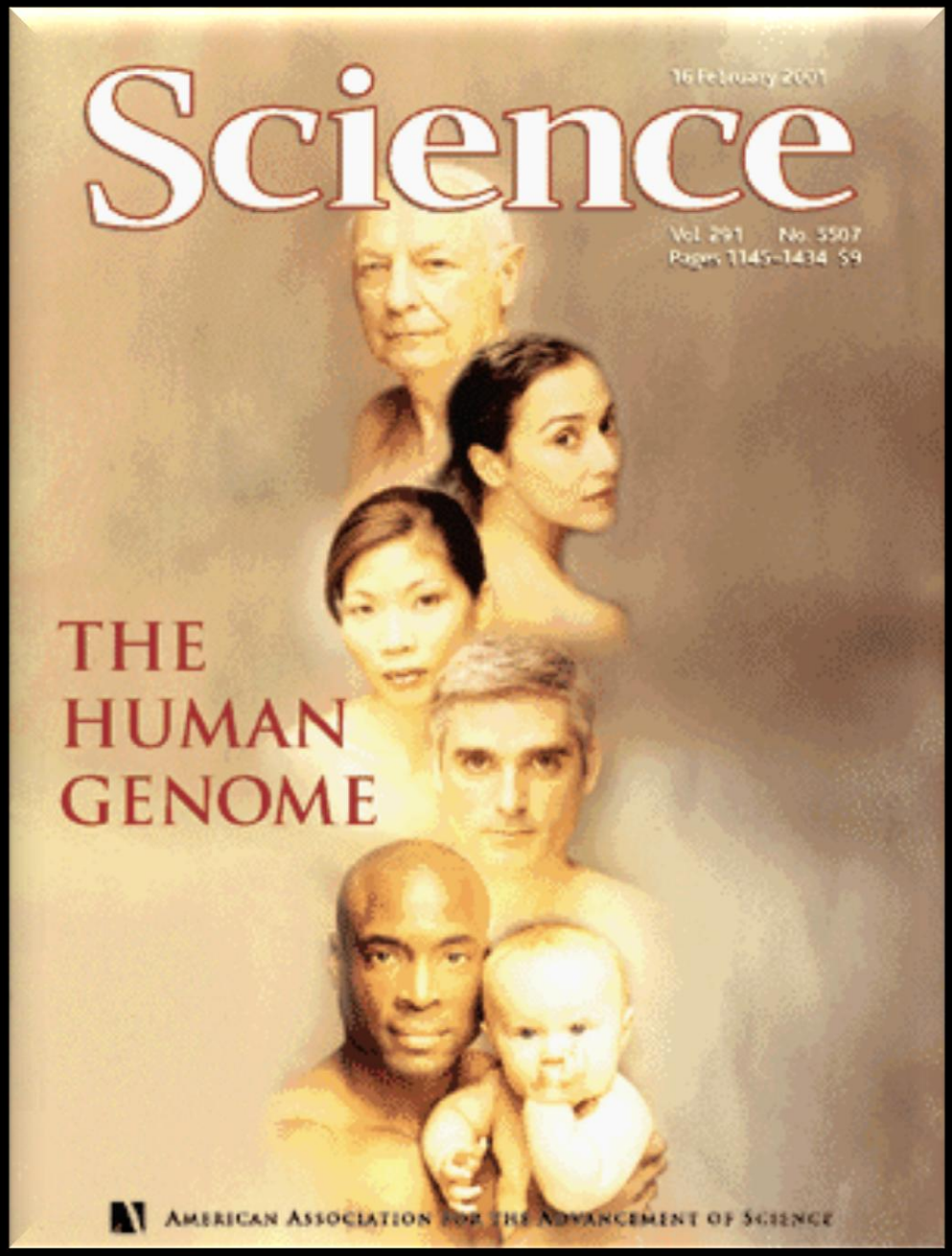
3.7 milhões de SNPs mapeados

Area	Goal	Achieved	Date
Genetic Map	2- to 5-cMresolution map (600 - 1,500 markers)	1-cM resolution map(3,000 markers)	September 1994
Physical Map	30,000 STSs	52,000 STSs	October 1998
DNA Sequence	95% of gene-containing part of human sequence finished to 99.99% accuracy	99% of gene-containing part of human sequence finished to 99.99% accuracy	April 2003
Capacity and Cost of Finished Sequence	Sequence 500 Mb/year at < $0.25 per finished base	Sequence >1,400Mb/year at <$0.09 per finished base	November 2002
Human Sequence Variation	100,000 mapped human SNPs	3.7 million mapped human SNPs	February 2003
Gene Identification	Full-length human cDNAs	15,000 full-lengthhuman cDNAs	March 2003
Model Organisms	Complete genome sequences of *E. coli, S .cerevisiae, C. elegans, D. melanogaster*	Finished genome sequences of *E. coli, S. cerevisiae, C. elegans, D. melanogaster,* plus whole-genome drafts of several others, including *C. briggsae, D. pseudoobscura,* mouse and rat	April 2003
Functional Analysis	Develop genomic-scale technologies	High-throughput oligonucleotide synthesis DNA microarrays Eukaryotic, whole-genome knockouts (yeast) Scale-up of two-hybrid system for protein-protein interaction	1994 1996 1999 2002

The Sequence of the Human Genome

CELERA





Um total de 21 doadores foram cadastrados. Destes, foram selecionados 5 doadores (por um conjunto de fatores): 1 africano, 1 asiático, 1 hispânico e 2 caucasóides
Aproximadamente 130ml de sangue foram coletados

Construção de uma Biblioteca de plasmídeos e BACs com inserto de vários tamanhos seqüenciados a partir das duas extremidades formando leituras pares de cada inserto (mates)

Método de sequenciamento
Whole Genome Random Shotgun

Seqüenciamento pelo método de Sanger

Sequenciador ABI PRISM 3700 DNA Analyzer

Laboratório ocupando 30.000 square feets (~ 2787 m^2)

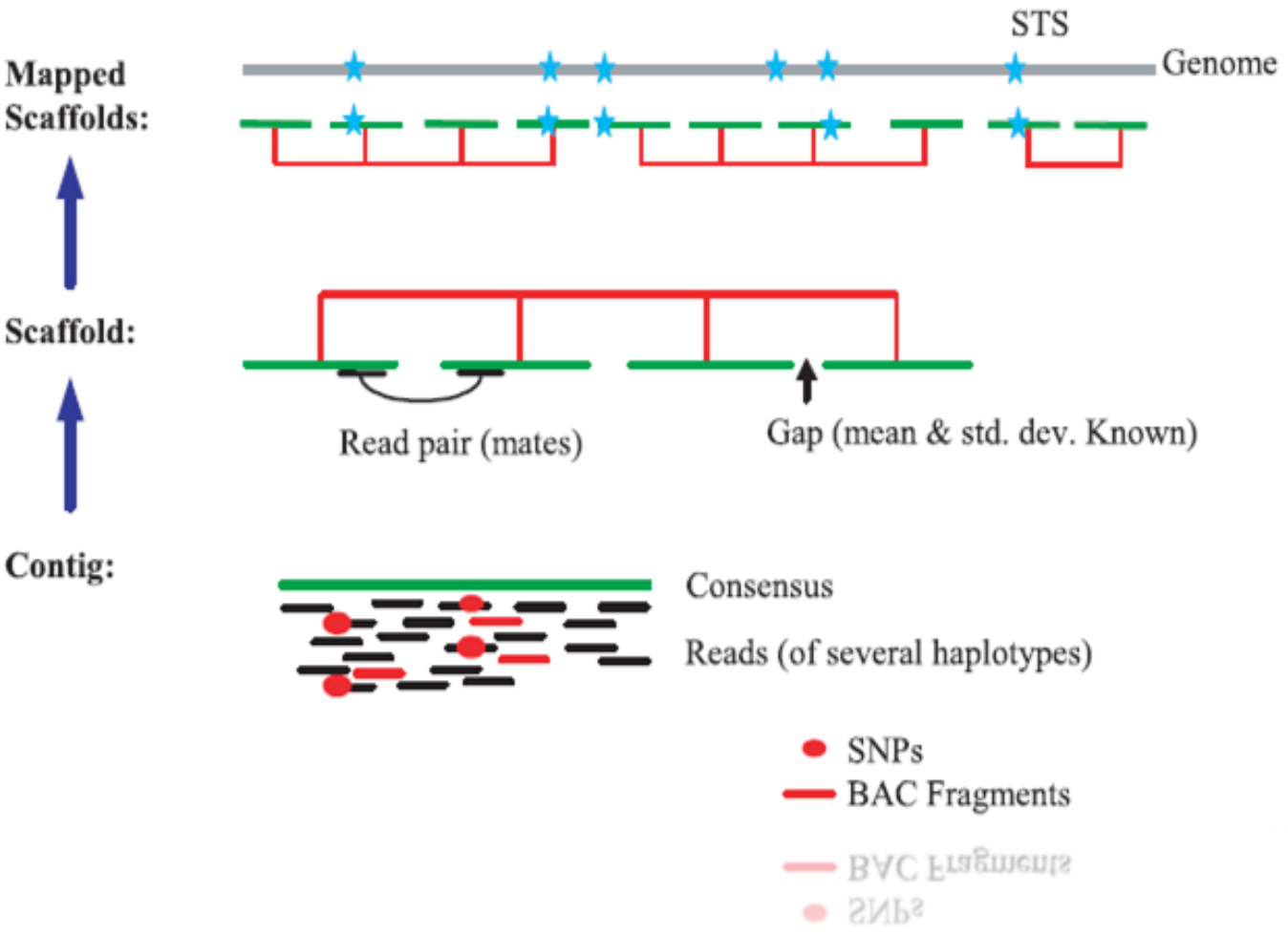
175.000 reads por dia - 27.27 milhões de reads no final cada uma com média de 543pb.

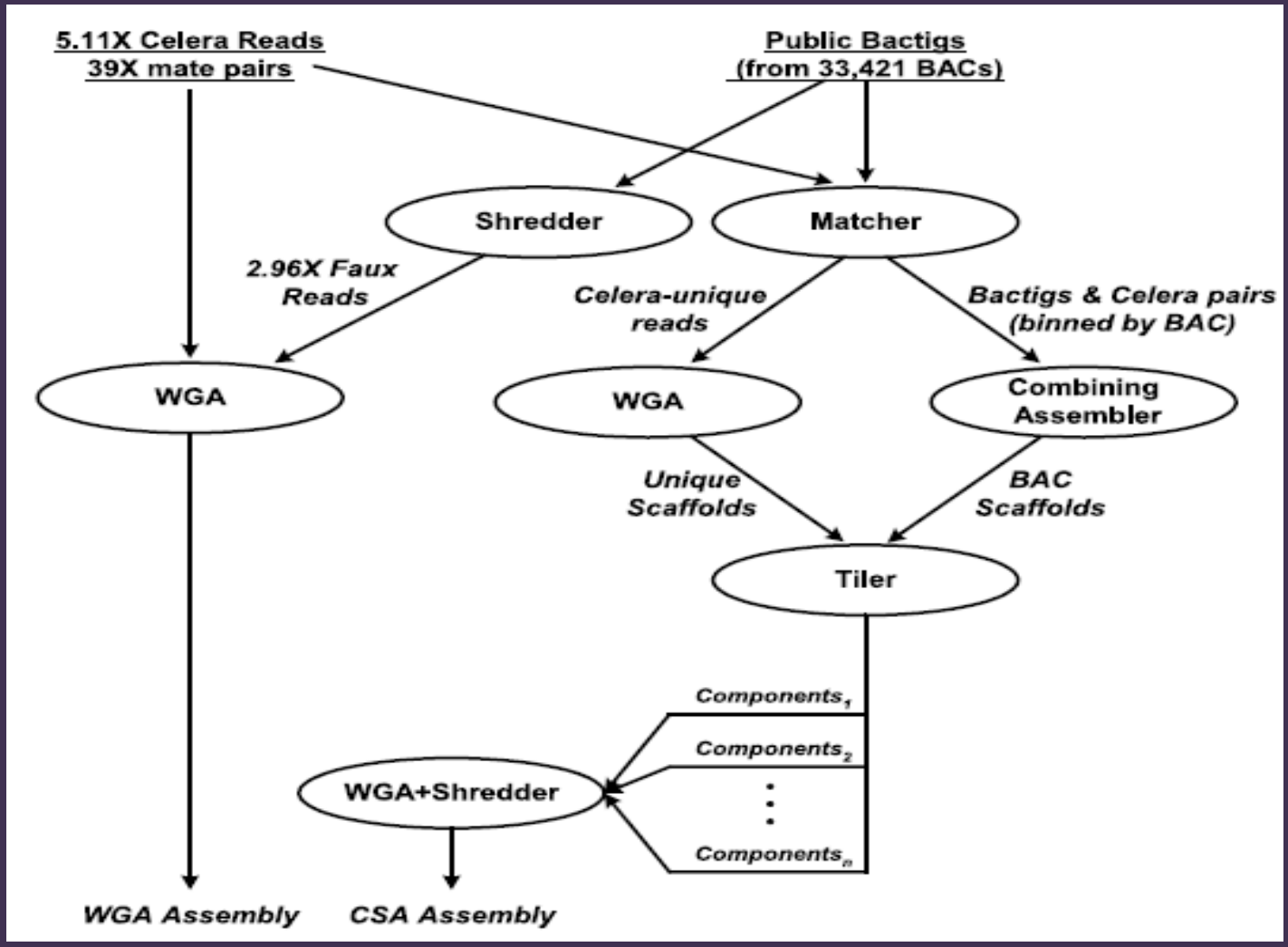
Estratégia de montagem em duas abordagens

Combinação computacional de todas as reads com dados do GenBank

Reunião de todos os fragmentos com dados de mapeamento
(abordagem utilizada na fase de análise por ter menos gaps e mais cobertura)

Primeira fonte de informação: reads
Segunda fonte de informação: dados do consorcio público.

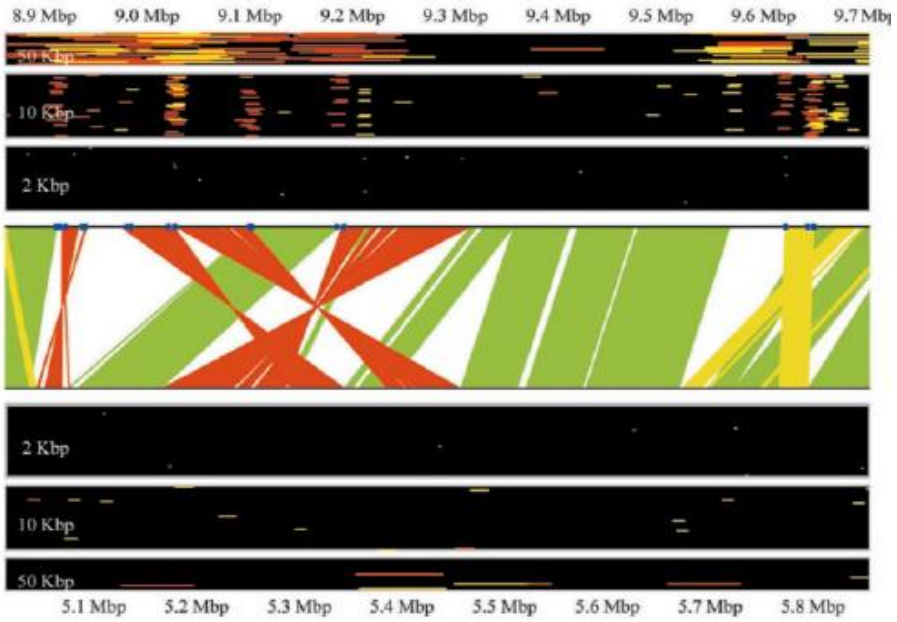

STS
Genome
Mapped
Scaffolds:
Scaffold:
Read pair (mates)
Gap (mean & std. dev. Known)
Contig:
Consensus
Reads (of several haplotypes)
SNPs
BAC Fragments

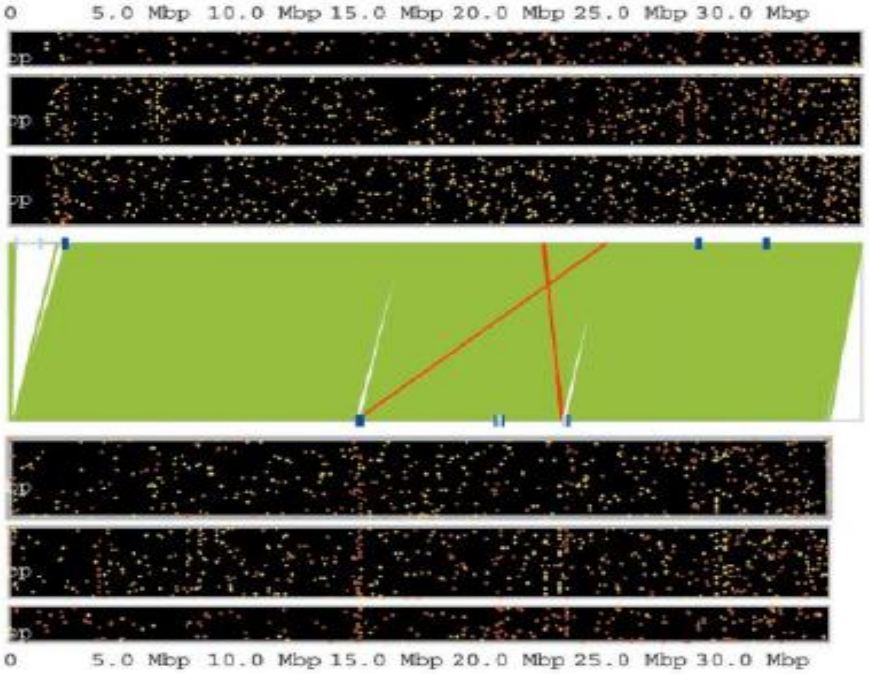

5.11X Celera Reads
39X mate pairs
Public Bactigs
(from 33,421 BACs)
Shredder
Matcher
2.96X Faux Reads
Celera-unique reads
Bactigs & Celera pairs (binned by BAC)
WGA
WGA
Combining Assembler
Unique Scaffolds
BAC Scaffolds
Tiler
Components1
Components2
Componentsn
WGA+Shredder
WGA Assembly
CSA Assembly

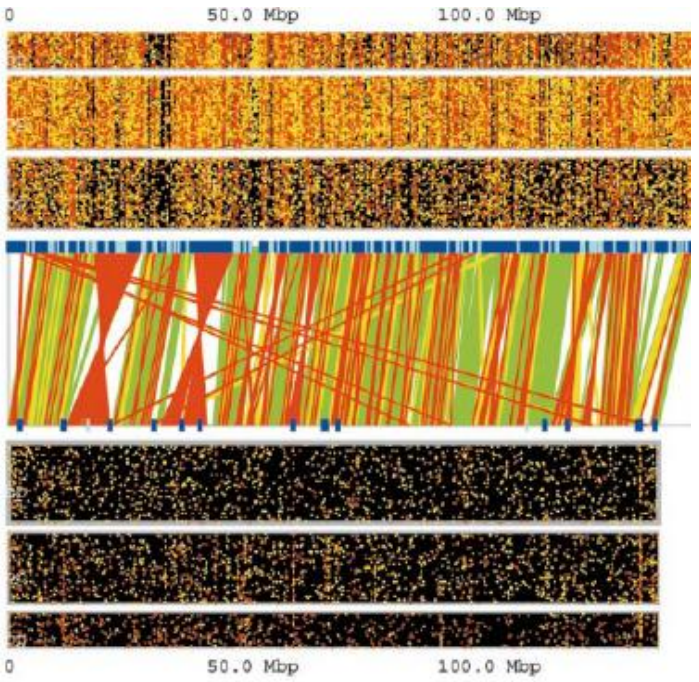
Total de horas de CPU para a montagem: 20.000

Passo final da montagem: ordenar e orientar os scaffolds nos cromossomos com base em mapas físicos

Validação de acordo com a cobertura e a precisão da montagem do genoma.







- mesma ordem e orientação
- mesma orientação mas fora da ordem
- divergentes

Breakpoints
Large gaps

Predição e anotação dos genes

Utilização de um programa chamado Otto o qual utilizava informações como regiões conservadas entre homens e ratos, similaridade com ESTs ou outros dados de mRNAs e cDNAs além de similaridade com proteínas, para predizer porções que poderiam ser genes

Dependendo do tipo de análise e informações utilizadas (linhas de evidências) foram encontradas diferentes totais de genes

26.383 gene anotados com boa confiaça – ao menos 2 linhas de evidência

Table 10. Features of the chromosomes. De novo/any refers to the union of de novo predictions that do not overlap Otto predictions and have at least one other type of supporting evidence; de novo/2× refers to the union of de novo predictions that do not overlap Otto predictions and have at least two types of evidence. Deserts are regions of sequence with no annotated genes.

	Sequence coverage (CS assembly)						Base composition			Gene prediction*					Gene density (genes/Mbp)						
Chr.	Size (Mbp)	No. of scaffolds	Largest scaffold (Mbp)	No. of scaffolds >500 kbp	Sequence covered by scaffolds >500 kbp	% of total sequence in scaffolds >500 kbp	% repeat	% GC	No of CpG islands	Otto	De novo/ any	De novo/ 2×	Total (Otto + de novo/ any)	Total (Otto + de novo/ any)	Sequence in deserts >500/ kbp	Sequence in deserts >1 Mbp	Otto	De novo/ any	De novo/ 2×	Otto + de novo/ any	Otto + de novo/ 2×
1	220	2,549	11	82	192	88	37	42	2,335	1,743	1,710	710	3,453	2,453	29	6	8	8	3	16	11
2	240	3,263	13	78	217	91	36	40	1,703	1,183	1,771	633	2,954	1,816	55	19	5	7	2	12	7
3	200	3,532	7	78	173	87	37	40	1,271	1,013	1,414	598	2,427	1,611	50	12	5	7	3	12	8
4	186	2,180	10	70	169	91	37	38	1,081	696	1,165	449	1,861	1,145	55	18	4	6	2	10	6
5	182	3,231	11	63	163	89	37	40	1,302	892	1,244	474	2,136	1,366	46	15	5	7	2	11	7
6	172	1,713	13	58	160	93	37	40	1,384	943	1,314	524	2,257	1,467	38	9	6	7	3	13	8
7	146	1,326	14	53	130	89	38	40	1,406	759	1,072	460	1,831	1,219	26	12	5	7	3	12	8
8	146	1,772	11	54	135	92	36	40	948	583	977	357	1,560	940	33	6	4	7	2	11	6
9	113	1,616	8	40	101	89	38	41	1,315	689	848	329	1,537	1,018	22	9	6	7	3	13	8
10	130	2,005	9	55	116	89	36	42	1,087	685	968	342	1,653	1,027	21	8	5	7	2	12	7
11	132	2,814	9	44	116	88	39	42	1,461	1,051	1,134	535	2,185	1,586	27	9	8	8	4	16	12
12	134	2,614	8	51	117	87	38	41	1,131	925	936	417	1,861	1,342	24	9	7	7	3	14	10
13	99	1,038	13	34	91	91	36	38	644	341	691	241	1,032	582	31	16	4	7	2	10	5
14	87	576	11	16	83	95	40	41	913	583	700	290	1,283	873	34	20	7	8	3	14	10
15	80	1,747	8	31	70	87	37	42	722	558	640	246	1,198	804	8	1	7	8	3	15	10
16	75	1,520	8	27	62	82	40	44	1,533	748	673	247	1,421	995	13	3	10	9	3	19	12
17	78	1,683	6	40	61	78	39	45	1,489	897	648	313	1,545	1,210	15	6	12	8	4	19	15
18	79	1,333	13	18	72	92	36	40	510	283	543	189	826	472	21	10	4	7	2	10	6
19	58	2,282	3	31	38	67	57	49	2,804	1,141	534	268	1,675	1,409	3	0	20	9	4	29	23
20	61	580	14	17	58	94	41	44	997	517	469	180	986	697	7	1	8	7	3	16	11
21	33	358	10	6	32	96	38	41	519	184	265	102	449	286	15	9	6	8	3	13	8
22	36	333	11	12	32	88	44	48	1,173	494	341	147	835	641	3	0	14	9	4	23	17
X	128	1,346	4	91	93	73	46	39	726	605	860	387	1,465	992	29	8	5	6	3	11	7
Y	19	638	2	10	12	65	50	39	65	55	155	49	210	104	4	2	3	8	2	11	5
U*	75	11,542	1						479	196	278	132	474	328							
Total	2907	53,591		1,059	2,490				28,519	17,764	21,350	8,619	39,114	26,383	606	208					
Avg.	116	2,144	9	44	104	87	40	41	1,160	714	812	333	1,526	1,047	25	9	7	7	3	14	9

*Chromosomal assignment unknown.

Table 11. Genome overview.

Size of the genome (including gaps)	2.91 Gbp
Size of the genome (excluding gaps)	2.66 Gbp
Longest contig	1.99 Mbp
Longest scaffold	14.4 Mbp
Percent of A+T in the genome	54
Percent of G+C in the genome	38
Percent of undetermined bases in the genome	9
Most GC-rich 50 kb	Chr. 2 (66%)
Least GC-rich 50 kb	Chr. X (25%)
Percent of genome classified as repeats	35
Number of annotated genes	26,383
Percent of annotated genes with unknown function	42
Number of genes (hypothetical and annotated)	39,114
Percent of hypothetical and annotated genes with unknown function	59
Gene with the most exons	Titin (234 exons)
Average gene size	27 kbp
Most gene-rich chromosome	Chr. 19 (23 genes/Mb)
Least gene-rich chromosomes	Chr. 13 (5 genes/Mb), Chr. Y (5 genes/Mb)
Total size of gene deserts (>500 kb with no annotated genes)	605 Mbp
Percent of base pairs spanned by genes	25.5 to 37.8*
Percent of base pairs spanned by exons	1.1 to 1.4*
Percent of base pairs spanned by introns	24.4 to 36.4*
Percent of base pairs in intergenic DNA	74.5 to 63.6*
Chromosome with highest proportion of DNA in annotated exons	Chr. 19 (9.33)
Chromosome with lowest proportion of DNA in annotated exons	Chr. Y (0.36)
Longest intergenic region (between annotated + hypothetical genes)	Chr. 13 (3,038,416 bp)
Rate of SNP variation	1/1250 bp

*In these ranges, the percentages correspond to the annotated gene set (26, 383 genes) and the hypothetical + annotated gene set (39,114 genes), respectively.

Análise das variações da sequencia

Comparação com bases de dados de SNPs e com a sequencia do consórcio público.

SNPs demonstram distribuição não aleatória

Pequena porção (<1%) com potencial impacto sobre proteínas

Encontrados 2.104.820 SNPs putativos de um total de 2.778.474 substituições diferentes na comparação entre os dois consórcios

Table 17. Distribution of SNPs in classes of genomic regions.

Genomic region class	Size of region examined (Mb)	Celera-PFP SNP density (SNP/Mb)
Intergenic	2185	707
Gene (intron + exon)	646	917
Intron	615	921
First intron	164	808
Exon	31	529
First exon	10	592

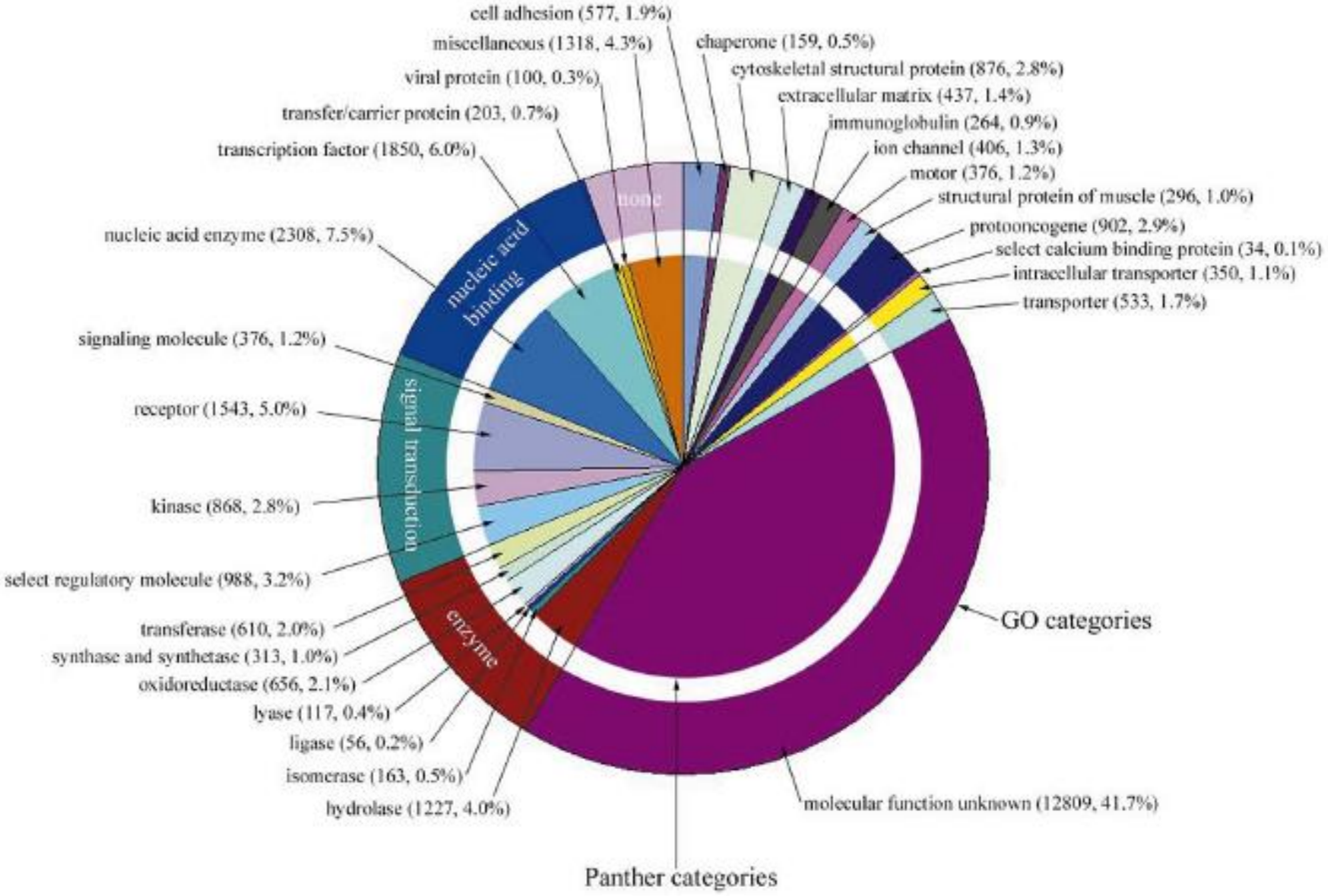
Visão geral das funções dos genes preditos no genoma humano

2 métodos para avaliar proteínas correspondentes aos 26.383 genes preditos

Famílias de proteínas

Domínios das proteínas

Avaliação computacional: resultados limitados


cell adhesion (577, 1.9%)
miscellaneous (1318, 4.3%)
chaperone (159, 0.5%)
viral protein (100, 0.3%)
cytoskeletal structural protein (876, 2.8%)
transfer/carrier protein (203, 0.7%)
extracellular matrix (437, 1.4%)
transcription factor (1850, 6.0%)
immunoglobulin (264, 0.9%)
ion channel (406, 1.3%)
motor (376, 1.2%)
none
structural protein of muscle (296, 1.0%)
nucleic acid enzyme (2308, 7.5%)
protooncogene (902, 2.9%)
nucleic acid binding
select calcium binding protein (34, 0.1%)
intracellular transporter (350, 1.1%)
transporter (533, 1.7%)
signaling molecule (376, 1.2%)
signal transduction
receptor (1543, 5.0%)
kinase (868, 2.8%)
select regulatory molecule (988, 3.2%)
GO categories
transferase (610, 2.0%)
enzyme
synthase and synthetase (313, 1.0%)
oxidoreductase (656, 2.1%)
lyase (117, 0.4%)
ligase (56, 0.2%)
isomerase (163, 0.5%)
hydrolase (1227, 4.0%)
molecular function unknown (12809, 41.7%)
Panther categories

Conclusões Projeto Privado

- Eficiência do método
- 95% da sequência de eucromatina
- Menor conteúdo gênico que o esperado
- Grande número de variações na sequência (SNPs)
- Complexidade genômica

PLoS Biology | www.plosbiology.org October 2007 | Volume 5 | Issue 10 | e254

The Diploid Genome Sequence of an Individual Human

Samuel Levy[1*], Granger Sutton[1], Pauline C. Ng[1], Lars Feuk[2], Aaron L. Halpern[1], Brian P. Walenz[1], Nelson Axelrod[1], Jiaqi Huang[1], Ewen F. Kirkness[1], Gennady Denisov[1], Yuan Lin[1], Jeffrey R. MacDonald[2], Andy Wing Chun Pang[2], Mary Shago[2], Timothy B. Stockwell[1], Alexia Tsiamouri[1], Vineet Bafna[3], Vikas Bansal[3], Saul A. Kravitz[1], Dana A. Busam[1], Karen Y. Beeson[1], Tina C. McIntosh[1], Karin A. Remington[1], Josep F. Abril[4], John Gill[1], Jon Borman[1], Yu-Hui Rogers[1], Marvin E. Frazier[1], Stephen W. Scherer[2], Robert L. Strausberg[1], J. Craig Venter[1]

Sequência do genoma de um único indivíduo

Identificação e comparação de alelos

Identificação de 4.1 milhões de variações, englobando 12.3 Mb, muitas delas descritas pela primeira vez.

3,213,401 SNPs,
53,823 substituições em bloco (2–206 bp),
292,102 inserções/ deleções em heterozigose
559,473 inserções/ deleções em homozigose
90 inversões
Diversas duplicações e variações em numeros de cópias

copyright © 2007 Bill Frymire

O sistema de alinhamento permitindo o agrupamento das reads por alelo.

Identificação de variações dentro do genoma (heterozigose)

Identificação de variações em comparação com a sequência referência

O sistema de montagem era capaz de agrupar as leituras em alelos isolados quando encontrava variações (heterozigose)

Um alelo abrangia leituras que compartilhavam sequências idênticas na região de variação, e era considerado confirmado se representado por um número maior de leituras (scores atribuídos)

Como esperado, normalmente haviam 2 alelos confirmados em cada região. Regiões com mais de 2 alelos aparentes representavam regiões repetitivas ou sequências com suspeita de erro de sequênciamento, e não variação genética verdadeira.

nature Vol 452|17 April 2008|doi:10.1038/nature06884

The complete genome of an individual by massively parallel DNA sequencing

David A. Wheeler[1]*, Maithreyan Srinivasan[2]*, Michael Egholm[2]*, Yufeng Shen[1]*, Lei Chen[1], Amy McGuire[3], Wen He[2], Yi-Ju Chen[2], Vinod Makhijani[2], G. Thomas Roth[2], Xavier Gomes[2], Karrie Tartaro[2]†, Faheem Niazi[2], Cynthia L. Turcotte[2], Gerard P. Irzyk[2], James R. Lupski[4,5,6], Craig Chinault[4], Xing-zhi Song[1], Yue Liu[1], Ye Yuan[1], Lynne Nazareth[1], Xiang Qin[1], Donna M. Muzny[1], Marcel Margulies[2], George M. Weinstock[1,4], Richard A. Gibbs[1,4] & Jonathan M. Rothberg[2]†

Comparação de dois genomas

Comparação com genoma de referência

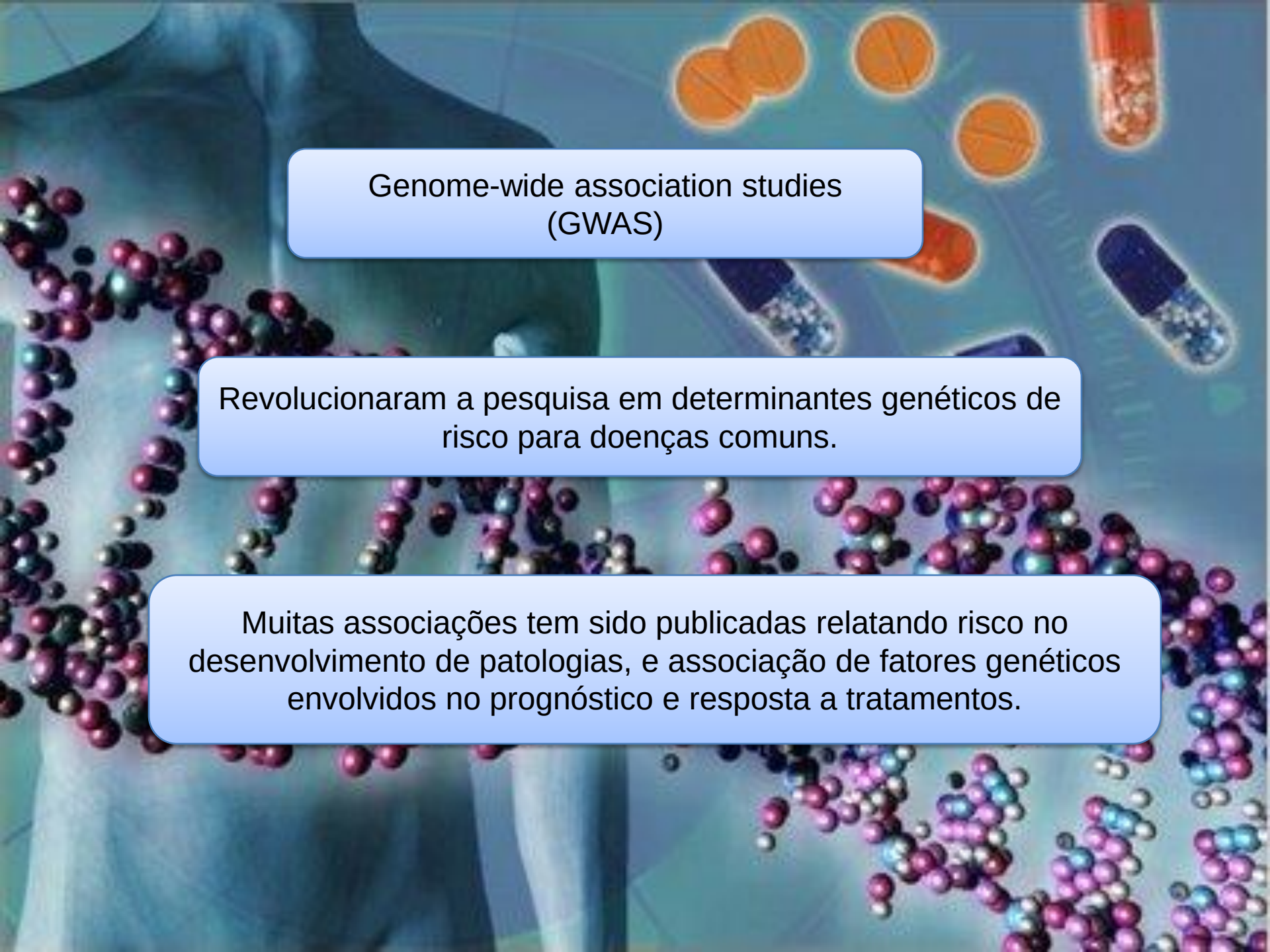
Encontrados 3.3 milhões de SNPs, e ainda inserções, deleções e variações em numeros de cópias resultando em ganhos ou perdas cromossomais variando entre 26000 a 1.5 milhões de pares de base.

Os dois genomas compartilhavam 1.68 milhões de SNPs dos quais 5230 não sinônimos.

Sequênciamentos e comparações individuais levam a busca de correlações genótipo-fenótipo de valor preditivo.

DESAFIOS

- Identificação de polimorfismos na população humana e associação com doenças
- Identificação de alvos para terapias
- Identificação de elementos funcionais do genoma: genes, proteínas, regiões regulatórias e elementos estruturais
- Identificação de como os genes e as proteínas funcionam em conjunto



Genome-wide association studies (GWAS)

Revolucionaram a pesquisa em determinantes genéticos de risco para doenças comuns.

Muitas associações tem sido publicadas relatando risco no desenvolvimento de patologias, e associação de fatores genéticos envolvidos no prognóstico e resposta a tratamentos.

O uso da genética para predizer risco e resposta ao tratamento é um passo em direção a medicina personalizada

Após a identificação de um locus com forte correlação estatística com risco, é necessário caracterizar a relação genótipo-fenótipo para fornecer a base biológica para a prevenção e tratamento

Prevenção e tratamento do câncer

Muitos estudos tem identificado loci que previamente não se suspeitava estarem relacionados com a carcinogênse, apontando novos mecanismos.

O risco conferido por alelos de susceptibilidade geralmente é baixo, porém efeitos combinados podem ser suficientes para predição de risco, para determinação de alvos para screening e para prevenção, especialmente se mais de um local é identificado.

"The real challenge of human biology, beyond the task of finding out how genes orchestrate the contruction and maintenance of the miraculous mechanism of our bodies, will lie ahead as we seek to explain how our minds have come to organize thoughts suffiently well to investigate our own existence"