doi:10.1006/geno.2001.6584, available online at http://www.idealibrary.com on IDEAL

A High-Resolution 6.0-Megabase Transcript Map of the Type 2 Diabetes Susceptibility Region on Human Chromosome 20

Sallyanne C. Fossey,[1] Josyf C. Mychaleckyj,[2,3,4] Joanne K. Pendleton,[1] Jonathon R. Snyder,[1] Jeannette T. Bensen,[1] Shohei Hirakawa,[1] Stephen S. Rich,[4] Barry I. Freedman,[3] and Donald W. Bowden[1,3]*

Departments of [1]Biochemistry, [2]Physiology and Pharmacology, [3]Internal Medicine, and [4]Public Health Sciences, Wake Forest University School of Medicine, Winston-Salem, North Carolina 27157, USA

To whom correspondence and reprint requests should be addressed. Fax: 336-716-7200. E-mail: dbowden@wfubmc.edu.

Recent linkage studies and association analyses indicate the presence of at least one type 2 diabetes susceptibility gene in human chromosome region 20q12–q13.1. We have constructed a high-resolution 6.0-megabase (Mb) transcript map of this interval using two parallel, complementary strategies to construct the map. We assembled a series of bacterial artificial chromosome (BAC) contigs from 56 overlapping BAC clones, using STS/marker screening of 42 genes, 43 ESTs, 38 STSs, 22 polymorphic, and 3 BAC end sequence markers. We performed map assembly with GraphMap, a software program that uses a greedy path searching algorithm, supplemented with local heuristics. We anchored the resulting BAC contigs and oriented them within a yeast artificial chromosome (YAC) scaffold by observing the retention patterns of shared markers in a panel of 21 YAC clones. Concurrently, we assembled a sequence-based map from genomic sequence data released by the Human Genome Project, using a seed-and-walk approach. The map currently provides near-continuous coverage between SGC32867 and WI-17676 (~ 6.0 Mb). EST database searches and genomic sequence alignments of ESTs, mRNAs, and UniGene clusters enabled the annotation of the sequence interval with experimentally confirmed and putative transcripts. We have begun to systematically evaluate candidate genes and novel ESTs within the transcript map framework. So far, however, we have found no statistically significant evidence of functional allelic variants associated with type 2 diabetes. The combination of the BAC transcript map, YAC-to-BAC scaffold, and reference Human Genome Project sequence provides a powerful integrated resource for future genomic analysis of this region.

Key Words: diabetes mellitus, physical map, chromosome 20, transcript map, genomics software, diabetes genetics

INTRODUCTION

Type 2 diabetes is one of the most prevalent metabolic diseases, characterized by peripheral insulin resistance, impaired insulin production, and increased hepatic glucose production, all of which contribute to hyperglycemia [1]. Despite intensive investigation, the etiology of the disease remains obscure. Identification of the molecular defects that contribute to type 2 diabetes may provide a better understanding of the biological pathways involved in disease progression and ultimately lead to effective patient intervention and treatment.

Recent linkage studies suggest the presence of at least one type 2 diabetes susceptibility gene on human chromosome 20q12–q13.1 in Caucasian type 2 diabetes families [2–5]. Evidence of linkage disequilibrium with type 2 diabetes has been observed with alleles of two genetic markers within this linked region, adenosine deaminase (*ADA*) and *D20S888*, markers separated by about 6 cM [6]. Although this interval contains hepatocyte nuclear factor-4α (HNF-4α), the gene responsible for maturity onset diabetes of the young type 1 (MODY1), mutations in the coding sequence and proximal promoter regions of HNF-4α do not account for diabetes in the general population of people affected with type 2 diabetes [7–11]. This suggests that mutations in at least one other gene in the interval are

0888-7543/01 $35.00

doi:10.1006/geno.2001.6584, available online at http://www.idealibrary.com on IDEAL

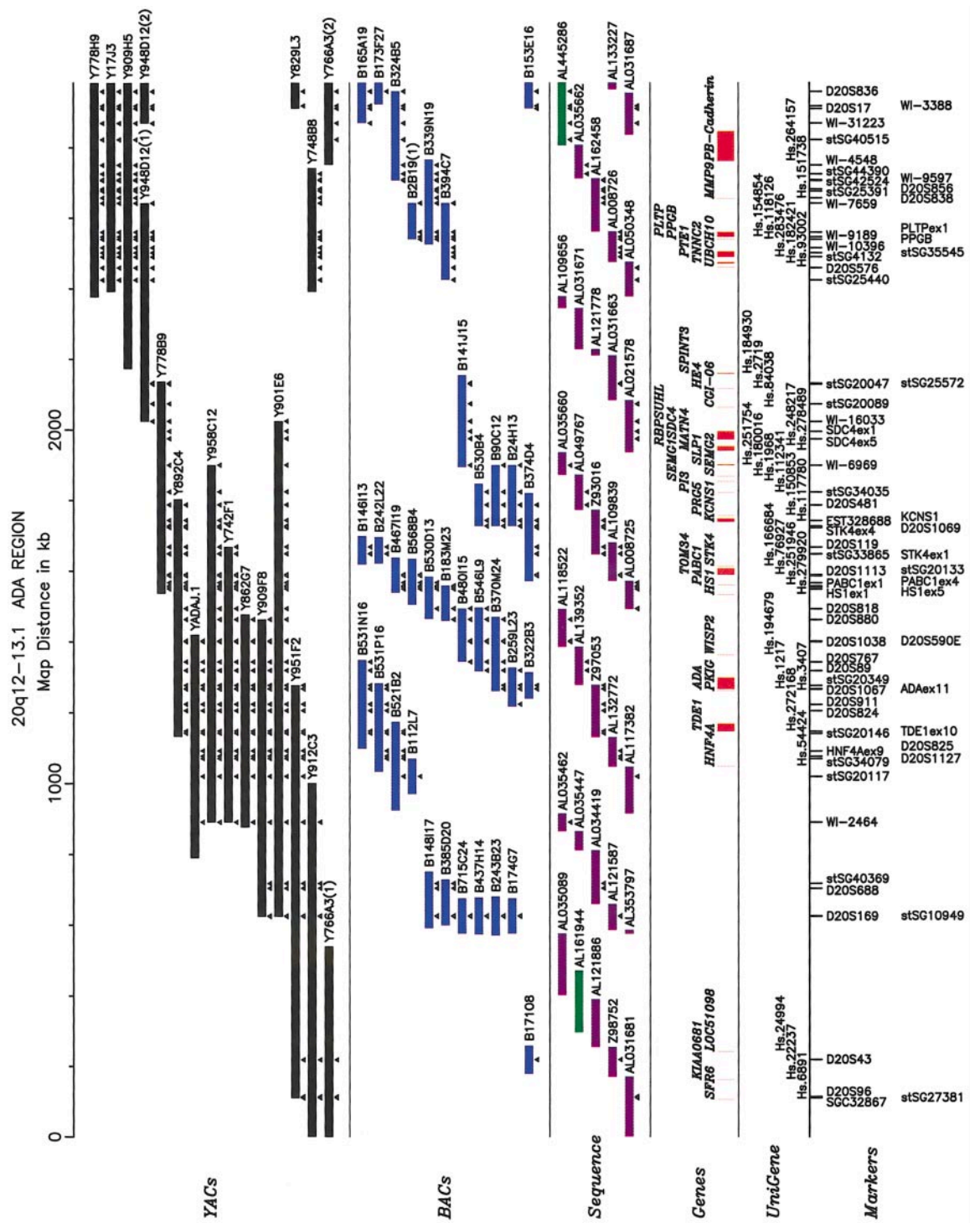

Figure 1 *Continued on next page.*

FIG. 1. Physical and transcript map for the cytogenetic interval 20q12–q13.1, including screened YACs, BACs, genomic sequence, genes, UniGene clusters, and STS/EST markers. The interval is divided into two regions (*ADA* and *D20S888*) but is fully contiguous throughout the 5832-kb interval. Known or putative chimeric YACs and BACs are shown with the parenthetical fragment number after the clone name. In the sequence clones, unfinished clones are light gray and finished clones are dark gray. All map distances are in kilobases and were calculated with a sequence-based map (HGP clones labeled "Sequence"). UniGene cluster identifier labels are left aligned with the centromeric edge of the gene to which they refer. UniGene cluster identifier for *RBPSUHL* (Hs.548217) is omitted for clarity, but it lies in the group of UniGene clusters at 2000 kb. We determined approximate gene positions by BLAST alignment of the mRNA transcript sequence to reference genomic sequence (Table 2).

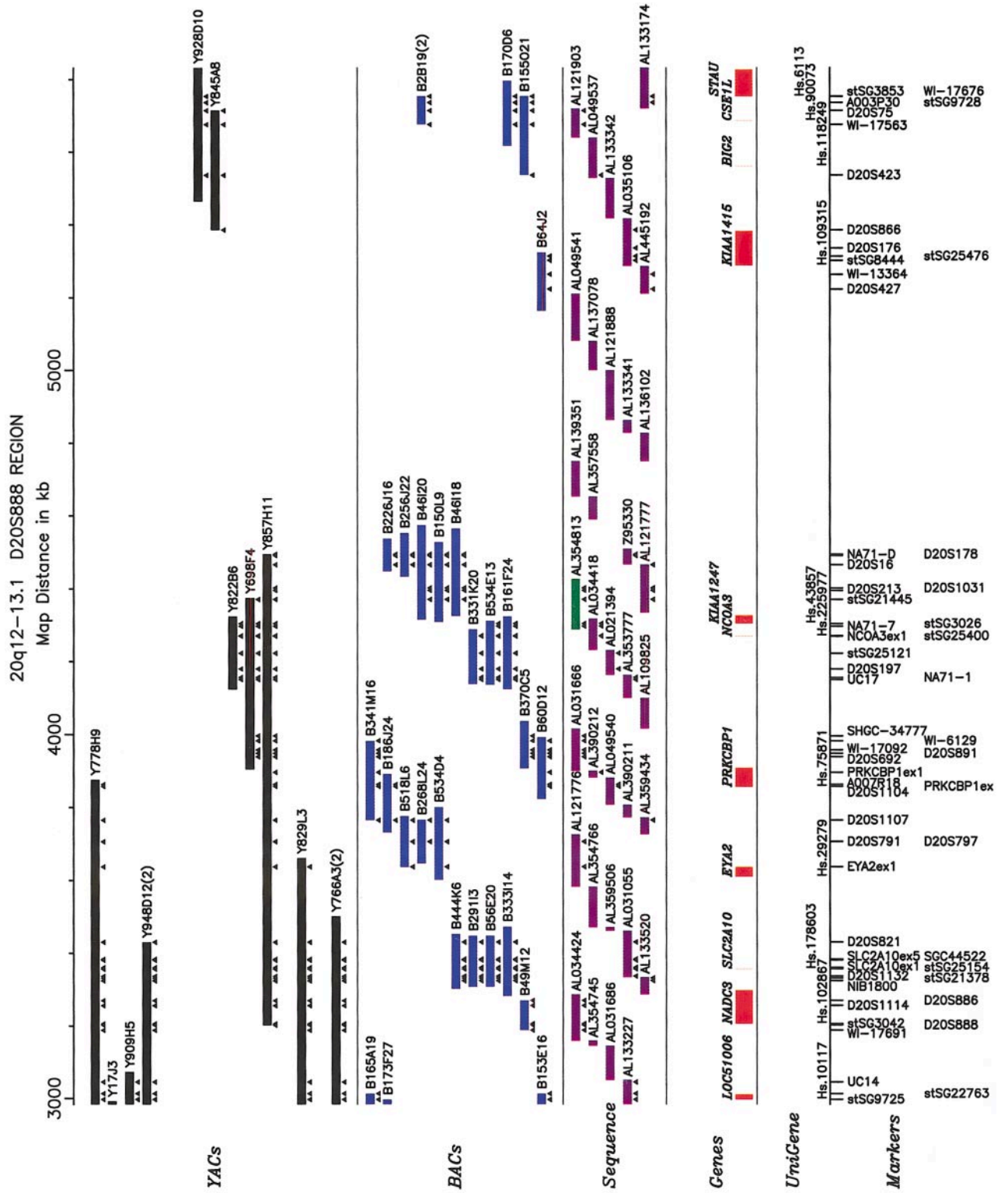

FIG. 1. *Continued from previous page.*

doi:10.1006/geno.2001.6584, available online at http://www.idealibrary.com on IDEAL

responsible for the observed linkage and association with type 2 diabetes.

Several other diseases have been linked to this same region of chromosome 20, including hereditary prostate cancer [12], systemic lupus erythematosus [13], and extreme human obesity [14]. Additionally, the 20q12–q13.1 interval is frequently deleted or amplified in many types of cancer, including malignant myeloid diseases [15,16], pancreatic endocrine tumors [17], and ovarian and breast cancers [18].

To facilitate identification of the type 2 diabetes gene(s) within this clinically important region of chromosome 20q, we previously constructed a framework radiation hybrid and physical map of the linked interval between *PLCG1* and *CEPBP* [19]. In this study, we have constructed a 6.0-megabase (Mb) sequence-based bacterial artificial chromosome (BAC)/yeast artificial chromosome (YAC) transcript map encompassing the linkage disequilibrium peaks detected around *ADA* and *D20S888* [6].

RESULTS

Construction of the Sequence-Based BAC/YAC Transcript Map

Genes, polymorphic markers, STSs, and ESTs were identified. We used selected markers to isolate BACs from the Human CITB BAC Library version 4.0 (Research Genetics). This BAC library provides a 13-fold coverage of the human genome, with an average insert size of 200 kb [20]. We used the GraphMap software program to generate a contiguous assembly of markers and BAC clones. Fig. 1 shows the full cytogenetic interval 20q12–q13.1 subdivided into two contiguous regions that include *ADA* and *D20S888* markers (Table 1). The marker content of 56 overlapping BAC clones was correlated with retention patterns observed in a panel of 21 local YAC clones [19]. The BAC/YAC contig map contains 42 known genes (Table 2), 43 unique ESTs, 38 unique STSs, and 22 polymorphic and 3 BAC end sequence markers. There is a 400-kb gap within the BAC/YAC contig, located about 5 Mb from the centromeric end between the markers *D20S178* and *D20S866*. One BAC clone, 2B19, appears to have a large internal deletion. PCR screening of this BAC with markers localized within the region verified the deletion.

We concurrently generated a sequence-based map from genomic sequence data released by the HGP, providing near-continuous coverage of the region between SGC32867 and WI-17676. The sequences were taken from Release 120.0 (October 15, 2000) of GenBank nonredundant (nr) and high-throughput genomic sequence (htgs) databases including daily updates, up to and including the first week of November 2000. The physical map of the region is estimated to be 5832 kb from the centromeric end of AL031681 to the telomeric end of AL133174 and about 5642 kb between the most centromeric (SGC32867) and telomeric (WI-17676) markers. The map has near-complete sequence clone coverage consisting of 63 finished (nr) accessions (62 unique clones) and 3 draft (htgs) accessions (3 unique clones), with the exception being a single gap in the *D20S888* region not currently captured by a clone within the GenBank databases, at about 4500 kb, between clones Z95330 and AL357558. This gap is captured by a BAC end sequence clone (CIT-HSP-2313F9, Caltech D1 BAC library) whose end sequences are in NCBI dbGSS (AQ027860 and AQ027862). Using BLAST analysis, these end sequences align with Z95330 (about 3.4 kb relative to the centromeric clone end) and AL357558 (about 8.0 kb relative to the centromeric clone end). Sizes for BAC end sequence clones are not generally available in the GenBank database, but using the average insert size for the library (129 kb; http://www.tree.caltech.edu/lib_status.html) and subtracting overlapping sequence segments with flanking clones Z95330 (clone size = 43.9–3.4 kb) and AL357558 (8.0 kb) provide an expected gap size of about 80 kb.

Unfinished clones in the *ADA* region are AL161944 (175.5 kb, 14 contigs), contributing 11.3 kb to map length (net of flanking overlaps); AL445286 (178.4 kb, 2 contigs), contributing two intervals of net 28.9 and 10.4 kb; and, in the *D20S888* region, AL354813 (139.0 kb, three contigs), contributing net 16.3 kb.

The identified genes were positioned in the map using BLAST alignment of the corresponding mRNA transcript sequence/cDNA accession with the local augmented database of 345 genomic sequences spanning the region.

Identification and Screening of Type 2 Diabetes Candidate Genes

The sequence-based transcript map facilitates identification of potential candidate genes for type 2 diabetes. We have localized 42 known genes (Table 2) and 43 unique ESTs. A detailed search of the EST/STS marker and genomic sequence clones identified 68 UniGene clusters (Table 3) within the region. We have begun to systematically evaluate the mapped genes and ESTs for allelic variants that may provide evidence of association with type 2 diabetes. The genes fall into four major categories: genes with an established role in glucose metabolism (*ADA, CEBP*), transcription factors (*EYA2, HNF-4α, RBPSUHL*), genes that may participate in diabetes-related signaling pathways (*HS1, PRKCBP1, SDC4, STK4/KRS2*), and genes whose function does not suggest an obvious role in the biological processes contributing to diabetes (*MMP9, PABC1, PLTP1, KCNS1*). We screened ESTs and the proximal promoter, coding, and 3′ untranslated sequences of genes in 100 unrelated type 2 diabetes patients and 100 unrelated healthy controls by single-stranded conformational polymorphism (SSCP) analysis. The results of the evaluation of 13 candidate genes and 10 ESTs are presented in Table 4. We detected 30 allelic variants (29 single-nucleotide polymorphism (SNPs) and one 7-bp deletion) within the candidate genes. SSCP analysis of the coding sequences of *CEBPB, EYA2, KCNS1*, and *PABC1* did not detect any allelic variants. We identified 16 SNPs identified within coding sequences (cSNPs) and 9 of them resulted in amino acid substitutions. We detected two cSNPs within the gene for *ADA*, an enzyme involved in the purine salvage pathway [21]. We detected a novel G-to-A transition at nucleotide 227 (G227A), which resulted in an Arg-76 to Glu (Arg76Glu) substitution in one

TABLE 1: Markers used to construct the 20q12–q13.1 transcript map

Marker	Position[a]	Type[b]	Forward primer (5′–3′)	Reverse primer (5′–3′)	Size (bp)	GenBank
SGC32867	110.6	E	GATCCCAAGTATTAGGAGCTATTTC	TCTGCCCCTTTAACAAGTTG	150	G25059
stSG27381	111.3	E	TAGGTGAACAAATCGGGAGG	CCCTTTCTTTTACTTGATGGTG	128	T17101
D20S96	114.2	P	CACTGCAACTCTAACCTGGG	CCTGTATGCTGCATTTCCTG	116	Z16449
D20S43	218.74	P	TGCACACCCATGTACACAGACTC	GCCCAGGTCTCCAACTCTCC	200	Z98752
D20S169	624.33	P	GAGTGCTTCTTGAACTCACA	TCTGAATCCTCTAGTGGCTG	197	Z23453
stSG10949	626.23	S	CATGTGATTCAGTCTGACTCTAAT	AGCTCACATTTTTTACAGGAATTAT	94	AL121587
D20S688	703.81	S	AGCTGATTTGATCTTGAGGAGC	TTGCCCTGCATAATTTTAAGTG	262	L44401
stSG40369	717.41	E	TAGACAAAAGTCGTCGCCG	CTGCCAAACCGTTTCGTC	122	R58958
WI-2464	890.80	S	GTCCTTTAACTTCCTGTACTCTCTCC	ATGAGGAACTGGACACCCAC	276	G03998
stSG20117	1020.2	S	TGCTCACACCCTATCTGTTCC	TATGAAATAGAAGATTGCTCGTGC	131	R00866
stSG34079	1070.4	S	GTTCATTCAGGCCACACATG	TCTGCAAAATGGGAAGGATC	131	T71265
D20S1127	1070.5	S	TATTTAATCTCAACCTTCATGGAGG	CATGTGTGGCCTGAATGAAC	146	G15366
HNF4Aex9	1077.2	G	TGGTTGATTGGCCAGCCTG	ATCCTGGTTCTACCTTCTAG	400	NM_000457
D20S825	1092.0	S	GAACAAAGCATTAATGGGATGGAAG	TGGGTAAGATAGATTAGGTAAGGAG	139	G07502
stSG20146	1141.8	E	ATTGTTTCAGACTGAGTCATGCA	CTTTGAGTATAGGGATACATGGTGG	140	R49379
TDE1ex10	1147.5	G	TTAAGCATGGCCTCAAATATCC	CAAGGACACCCACTGGAACT	343	NM_006811
D20S824	1205.3	S	AAAAAGCTGTGGTGGAGTATATGG	TAGTCAACTTTGGAGAAATAGTGAGG	145	G07501
D20S911	1223.4	P	TCCCAAGTGCCTAGAAGAG	GGCCCAATTTGTAGTTCAG	180	Z51767
D20S1067	1266.3	E	TTCACACAAATCATCGGCAT	TCCCAGACCTGCTTGTCC	280	T77999
ADAex11	1267.1	G	AGGATGGACTATCACTACATTG	CAGGGCAACTGCCCAGAAG	240	NM_000022
stSG20349	1276.5	E	GTTCTCAGTTTCCCCATCTGTCCAGTG GGAGCAG	CTGAGGGACAGGCCTGGTCCTAGTCAT AGGGAT	483	M13792
D20S89	1318.7	P	TGAAGTGTAGAGCTTGACA	TGCAGTGAGCCATGTTCAT	120	L29958
D20S767	1343.0	E	ACCTGCTATGTGCCAGGCAT	ACCTCCCATGCAGGATTGCT	79	F01517
D20S1038	1399.9	E	AGAAAAGCCAAAAACTTTAATTTCA	TCAGAATGGAGCCGAACC	279	Z39210
D20S590E	1402.4	S	CATTAAGAGGGTGTTGTTTTCTCC	TCTGTGGAGGTGATATACATTG	187	Z43124
D20S880	1463.1	P	AGCTGCACATACACGTACAC	CAATTGTGATAAGTGCCATAAAA	263	Z51163
D20S818	1493.8	S	GATTAAAACACCAATACCCAGTGC	GGATTCTAGTTTAGGAGGTCCAG	157	G07494
HS1ex1	1549.5	G	CCGAAACCTGACATTGCTC	GCTCAATGCTGGAGATGAC	234	NM_003404
HS1ex5	1554.4	G	TACATACTGGGCCACTTACC	GATATATGTTGAGGGTACAGAG	230	NM_003404
PABC1ex1	1558.1	G	TGGGTGACCCGGCTCCTGCTT	CGCCTGCTTGTCCGTCGG	250	NM_006534
PABC1ex4	1566.9	G	GGCTGATGGCTGGTAGCTG	GAAAAACAGTGATCCCCCCG	240	NM_006534
D20S1113	1586.7	E	GCTGGATGGGTGACCAAC	AAGGAAATCCTCGCTTCCAT	146	G14773
stSG20133	1589.9	E	AAAGCTTTTTAATGCAACAGCT	GTGTCCATGCTTTCTGGGC	150	R39018
stSG33865	1648.5	E	ACCTTCAGAAGAGGCTCTTGG	GGCCTTGCTCAGAAGTTTTG	145	U60207
STK4ex1	1649.2	G	GGCACTGCACCATATAAACTG	TAACAAGCATCAATGTCTGAGG	220	NM_006282
D20S119	1668.2	P	AACTGACACAGTTTCAGTATCTCT	TTTTCCAGATTTAGGGGTGT	110	Z17198
STK4ex4	1723.1	G	TGCCACTGACTTAAGCTTTG	CATAATTCACCCAAGCCTG	340	NM_006282
D20S1069	1727.4	E	CCAAGTTGGTCCCAGTGC	TCCAGTGTCCTTTCTAGCATACC	223	T83726
EST328688	1727.6	E	ACAAACGTGATGAGGTATAG	CCATTGTCTTTGGAATACATG	1200	W45323
KCNS1	1743.2	G	TAACCTGGTCTTCCAGGAAGG	CGCCTCGTCGTAGTCGTCG	280	NM_002251
D20S481	1787.3	P	TGGGTTATGAGTGCACACAG	AACAGCAAAAAGACACACAGC	234	G08051

Table 1 continued on next page

doi:10.1006/geno.2001.6584, available online at http://www.idealibrary.com on IDEAL

TABLE 1: Continued

Marker	Position[a]	Type[b]	Forward primer (5′–3′)	Reverse primer (5′–3′)	Size (bp)	GenBank
stSG34035	1824.1	E	CAGTCTCCACTAAGCCTGGC	ACAGGTGCAGCAAGGACC	179	Z18538
WI-6969	1899.7	E	TCCTGCCATATGGAGGAGG	GTGCAAAGAGAAATAGGCTCG	175	G06121
SDC4ex5	1974.8	G	CCTTAGGTCCTGATGAGGAG	CCACCCTTCAAAATCCCCTG	290	NM_002999
SDC4ex1	1995.8	G	CGCCTATAAGATGGGTGGCG	ACGCTCCGACGAACAAAGGAG	210	NM_002999
WI-16033	2024.2	E	TTTTTTTTACCATGCCTCCG	CTGTAGGATGGTACTTAGCAGGC	108	H09823
stSG20089	2073.7	E	TTTATTCCACAAACAGTAAACTCCA	ACAGCACAGAAAAAGATTTCCA	125	H25231
stSG20047	2129.4	E	GCTGAAAGTGGTTACTTTATTGG	ACCAGGCTGAGCAGTGAGG	150	F02471
stSG25572	2132.3	S	ACGCTTTAAGATCATGAACTGC	AGCAGGTGTGGAGCCTAAGA	178	Z94648
stSG25440	2424.5	S	CCCCAAATGAAATCTTAGTTGG	CTAGCCTGGGGTACAGACCA	113	Z94564
D20S576	2459.1	E	CTACAAACTCGCTCAGGCAC	CTCCAAAGCCCTCTTCCTC	96	Z41789
stSG4132	2489.9	E	GCCAAAAGGGACTGTAACTCC	CAGGATGGAGTCCTAGCTGTG	164	H29206
stSG35545	2501.6	E	AGTTCAAGTAATGCCCCCA	GCCAGGGTCTTATCTTTATGC	153	T16332
WI-10396	2515.8	S	GCCTTGGAGTATATCTAAACTGTGG	ACAAAGTGTTTACAAATGGTGGC	231	AL008726
WI-9189	2546.8	E	AGCCTGGGGGCAAGTTAG	TGGAATCAGTGCATCATAAAGG	102	M22960
PLTPex9	2559.7	G	CAAGAGAAGGTCTGTGACAGA	CCTTGGTCTCACTGGTGTG	150	NM_006227
WI-7659	2641.7	E	CCCGTCCTGCTTTGCAGT	ATCCAAGTTTATTAGAAACACTCCA	174	J05070
D20S838	2656.9	S	CTCATGCTGGTGCTGC	GAGGCGCTCCTGTGAC	119	Z52250
stSG25391	2676.2	S	TGATGCCACTTCCTGAGATG	CTTGATCATCCTGGGACCC	80	Z94533
D20S856	2681.7	P	CCCTTCAACGTGCTGG	GGAATGCTGTGTGCTGTG	161	Z52662
stSG42524	2705.9	E	TACAGACCCTGTGCCCGT	GAAAAAGGGCAGCGAAGAC	151	H84419
WI-9597	2708.2	E	TTATTGCATTTTGTGCAGACG	ATCCTGCCTCAGTATTGATCTTG	149	Z239181
stSG44390	2724.9	E	AGGGAAATATCACCAAGGGG	AAATGAGGTCATTTGGTGGTG	126	R16784
WI-4548	2749.9	S	CAGCATAGGCAGGTCCCAG	GTCTTGGTGGAAGTCCTTGG	127	G03661
stSG40515	2822.0	S	CCATTTGAAAACAAAAATTTATTGC	TGACCTATGTCACAAGAGGTGG	233	G14610
WI-31223	2868.8	E	TGGGCTTACTCACGCATCTG	GCTATGTAAGGAAGCCAGCC	150	H86779
D20S17	2910.6	P	GGCTAAGTATGCAGCAGTTAG	GTACTTTCTCTTTGCAGACCTTG	197	L12127
WI-3388	2918.0	S	CCTAAGTCTGTGTTCATCTTTACCC	CTCCTGCCTTAAATATCATCAGC	132	NT_002193
D20S836	2959.1	P	TAAGAGCAGCCTCCCCATC	GTCTGAACGCCCTAACAGC	144	X97925
stSG9725	2996.9	E	CTTAAGACCAGACATTTGAA	AGAGAGCAGGTTTCTTTATAG	156	H52166
stSG22763	3013.3	E	TTTTCCCCAAGGTTCCATC	GCATTCAGTATTGGTGGCCT	146	T90783
UC14	3045.2	S	GGCTACCTGTACGTTACTAC	TGGGTTTGCTCTTACCCACT	104	G01528
WI-17691	3188.1	S	GATCAGGCTCATCTCATCTGC	TTGAGGCATATAAATAAGTTCCAGG	102	D20888
D20S888	3202.2	P	GGACTTGCTAAGCCTCCAC	GTCAGGGCTCCCTAGAGAA	167	Z53429
stSG3042	3205.5	E	CTCCTTTCCAAGCGAGCAC	CCTCTTGCTTCCAGTCCC	91	Z39412
D20S1114	3255.8	E	CACTGAGGTAAATGATCCAAAGC	AACTTCCGGGTTCTGTCCTT	130	T56713
D20S886	3269.6	P	TGACCTTAGAGTGTCCCTAGC	CTCAACAGGAATTGGTGTG	149	Z53418
NIB1800	3324.3	E	GCCTCACAGCTTTTTATTGA	AGGGGCAGTGATTTGGAG	179	T16751
stSG21378	3331.6	E	GCCGAACAAGAGGGAAAAG	TCTCATCTTGAATTGTTCCCG	162	N35484
D20S1132	3336.6	S	TTACTGAGCGTTGCCGAAC	TCCGGCCTGTAAGTGGTTAC	133	G15621
SLC2A10ex1	3356.5	G	—	—	—[c]	
stSG25154	3360.3	S	ATCCCCTTCTCGGTGAGG	CCTTGAGGCTGACTTCAGG	179	Z94364
SLC2A10ex5	3380.7	G	—	—	—[c]	
SGC44522	3383.2	E	GGGTATGTTTGTTGCTCACAA	GGGTATGTTTGTTGCTCACAA	261	4999952
D20S821	3430.1	P	ACAGGAAATAAACTAGGCATGAGG	CAACTCGATGAAACTAAGATTTCAAC	166	G07497

Table 1 continued on next page

TABLE 1: Continued

Marker	Position[a]	Type[b]	Forward primer (5′–3′)	Reverse primer (5′–3′)	Size (bp)	GenBank
EYA2ex1	3637.0	G	GAACTAGTGATCTCACCCAGCC	GCACTGTACTGCGTCTG	220	NM_005244
D20S791	3705.9	S	TAATGCTGGGAGAAAACAGA	GCCTGTTGCATATTCTCTTG	126	G07635
D20S797	3706.2	S	GAGTGAGACTCCATGTCAACA	CACAGTTCCTGGGACATAAA	162	G07641
D20S1107	3764.9	S	TTCCATCCTGTGACAAACACA	CTCCGTTCCAGAGCTACAGG	197	G14650
D20S1104	3857.5	S	AGTCTGCAGTCAAAGCCGAT	GCAGGAGAACAGCCACTTTC	115	G14595
PRKCBP1ex9	3857.7	G	TTGTGCTCAGGCGCAGGTGG	CTTTTCCTGGCGTCTGGGT	245	AF233453
A007R18	3862.7	E	CTCACTCTGTTGCCCAGC	GTTAGTTAGATGTCTCTTC	162	T65756
PRKCBP1ex1	3896.1	G	CTCCTCAACTGTCAGCTCCT	CTGAGGCTGCTGCTGCTGG	277	AF233453
D20S692	3938.5	S	TTCCTTCCTGCTGCTCTCAT	TTTTTACTTGCAAAAGCCATAGC	104	L44405
D20S891	3948.2	P	GCAAGCATCTACAAGGCTCTTCAT	CTACAGGTGAGCGCCACCAT	212	Z53706
WI-17092	3957.3	E	AATGGGGCTTGTTGATGTGT	ACTGCTTTAAAACCTGTGGTCC	137	R06466
WI-6129	3982.5	S	GTCATTCCTTCATTGTTCTCTATTG	AAAGGAGAATTGTATGGTATGTGAA	253	G05022
SHGC-34777	3995.6	E	ATTCCAGCAATCAGACTGGG	TTCCTCATTCCACTTTTGTGG	385	H57758
UC17	4150.3	S	CAGCATTCGCTGAAGACCTA	GAGTAAGATTGCTGCATCTC	150	G01531
NA71-1	4156.1	B	GGTGCTCTGTATGGTTTTCCA	GTCAGCACTGAAAGCTGCAT	250	—[d]
D20S197	4179.2	P	TCCTGGTGTCCTTGTTTAAGTATCA	CATGTGTTGCCTATTCCTTAGATGT	144	Z24408
stSG25121	4222.4	S	TTCTTTTTACAAAATGGCCTCA	GCTTCTTGCTAACATGCTGAC	81	Z94341
NCOA3ex1	4269.4	G	GTCGACGTGGCGGCCGGCGG	GAATATACATCAGCAACTGG	180	NM_006534
stSG25400	4270.4	S	CCAAAGTCAGTGTCAATTGAAG	TGGGCAGAGTTAGCCTCAGT	190	Z94538
NA71-7	4296.8	B	GCTCAGCTTCCACAAGGATG	ATGAGGCAGATGACCACACA	150	—[d]
stSG3026	4303.7	E	CCTTGAGGTTTTGCTCCTAG	CGTAAACAGAATGGATTGCC	110	Z25162
stSG21445	4370.2	S	CCTACCAAAGACACTAGGCCC	GGAGACACATGGTTAGAAAGCC	145	N47785
D20S213	4395.1	P	GGCAACAGAGTGAGACTTCG	AGCAATGGCTGTTGATAAGG	139	250449[e]
D20S1031	4401.7	S	AGAACAAGAGGACACAAATCTTCC	GTCCTCTCTCCCCTCAACCT	177	G11908
D20S16	4465.5	P	—	—	—	166926[e]
NA71-D	4490.6	B	GCTTCAGGAGAGCTGAGACC	GCCATGCACAACTGTGAGTC	205	—[d]
D20S178	4494.0	P	GCCATGTCCATACAGAAC	GGATTCCTGAAAAGTGAAG	252	Z23757
D20S427	5222.7	S	AATGACAAAAAAAGGAAGGCAG	CATTCACTTCAGACCTAGCAG	363	L17847
WI-13364	5263.5	E	TTTTTCTTTTGTGCTCTTTTTTTT	TGACATTTTACATTTCAAAGAAAGG	110	T90531
stSG8444	5301.3	S	AGTTTTCCATTGGAGGCCTT	GAAGGACTTTGGTGAGATGAGG	175	H65114
stSG25476	5313.0	S	ATCCCGAAGGTGTTTGCTC	AGGATAACACACAATGCCTGC	233	Z94857
D20S176	5335.4	P	CTTGGGACTTGTCAGCCTC	TCTTAACTTCTGCCCCTTG	183	Z23738
D20S866	5385.1	P	TAAACAGGAGGTGCTCAGCC	AGGACTTGCTCAAGGTCTCTGC	173	Z53097
D20S423	5535.4	S	GAGACAGAAAATAGATTAGAGG	TGACAGAGTGAGACTCCGTG	270	L29969
WI-17563	5674.9	E	TCACCTGCCAGAGGAAAATG	AAACAGTGGATGGGTAATTTTTAT	150	R77775
D20S75	5713.3	P	GCTGAAATGGGAGGATCG	GCTGCAGTGAGACATGATCA	173	M87717
A003P30	5735.5	E	ACAAGTTCAAAAGGAGAACT	AGTATCTCCAAGGGTACC	150	T91320
stSG9728	5735.8	E	AGGATTTAGTAACAGAAAGTCTC	GGACAAGTATTATAGTTCAGT	112	D60376
stSG3853	5752.7	S	AAGCAGTACTTAACTCGCAGGG	TGATTTTCCAGAGTGCTCGA	187	H11397
WI-17676	5752.8	E	CTCCATGTAGTCCATATTAACCTGC	ACACAGCTGTATACAGAAACGTAGG	113	D20243

[a]Estimated marker map distance (kb) relative to the centromeric end of AL031681.
[b]Indicates marker type: B, primers derived from BAC end sequencing; E, EST marker; G, gene; P, polymorphic marker; S, STS marker.
[c]Fossey *et al.*, unpublished data.
[d]BAC end sequence markers described by Price *et al.* [19].
[e]Indicates GDB database accession number.
[f]Complex restriction fragment length polymorphism derived from genomic clone CRI-I1214 [45].

doi:10.1006/geno.2001.6584, available online at http://www.idealibrary.com on IDEAL

TABLE 2: Genes within the 20q12–q13.1 transcript map

Gene symbol	Description	Accession number[a]	UniGene cluster
ADA	adenosine deaminase	NM_000022	Hs.1217
BIG2	brefeldin A-inhibited guanine nucleotide-exchange protein 2	NM_006420	Hs.118249
CGI-06	*Homo sapiens* CGI-06 protein	NM_015937	Hs.84038
CSE1L	cellular apoptosis susceptibility protein	NM_001316	Hs.90073
EYA2	Eyes absent homologue 2	NM_005244	Hs.29279
HE4	human epididymis specific protein E4 precursor	NM_006103	Hs.2719
HS1	14-3-3/KCIP protein kinase inhibitor	NM_003404	Hs.279920
KCNS1	potassium channel protein, α-subunit	NM_002251	Hs.117780
KIAA0681	lethal (3) malignant brain tumor l(3)mbt protein (*Drosophila*) homologue	U89358	Hs.22237
KIAA1247	similar to glucosamine 6-sulfatases	AB03373[b]	Hs.43857
KIAA1415	KIAA1415 protein	AB037836	Hs.109315
LOC51006	CGI-15 protein	NM_015945[b]	Hs.10117
LOC51098	CGI-53 protein	NM_016004	Hs.24994
MATN4	matrilin 4	NM_003833	Hs.278489
MMP9	matrix metalloprotease, member 9	AX011001[c]	Hs.151738
NADC3	sodium-dependent high-affinity dicarboxylate transporter	AF154121	Hs.102867
NCOA3	nuclear receptor coactivator 3	NM_006534	Hs.225977
PABC1	polyadenylate-binding protein cytoplasmic 1	AL008725.2	Hs.251946
PB-Cadherin	similar to long type PB-cadherin	AF035300[b]	Hs.264157
PI3	SKALP/elafin, elastase-specific inhibitor	NM_002638	Hs.112341
PKIG	protein kinase (cAMP-dependent, catalytic) inhibitor-γ	NM_007066	Hs.3407
PLTP	phospholipid transfer protein	NM_006227	Hs.154854
PPGB	protective protein for β-galactosidase	NM_000308	Hs.118126
PRG5	p53-responsive gene 5 (putative)	AF147078[b]	Hs.150853
PRKCBP1	protein kinase C binding protein 1, RACK-like protein	AF233453	Hs.75871
PTE1	peroxisomal acyl CoA thioesterase homologue	NM_005469	Hs.283476
RBPSUHL	recombining binding protein suppressor of hairless-like (*Drosophila*)	NM_014276	Hs.248217
SDC4	syndecan-4, integral membrane heparan sulfate proteoglycan	NM_002999	Hs.252189
SEMG1	semenogelin I	NM_003007	Hs.1968
SEMG2	semenogelin II	NM_003008	Hs.180016
SLC2A10	GLUT10, facilitative glucose transporter	AF248053[d]	Hs.178603
SLP1	secretory antileukoproteinase	NM_003064	Hs.251754
SFRS6	splicing factor, arginine/serine-rich 6	U30883	Hs.6891
SPINT3	serine protease inhibitor, Kunitz type 3	X77166	Hs.184930
STAU	RNA-binding protein staufen	NM_004602	Hs.6113
STK4/KRS2	stress responsive serine/threonine kinase	NM_006282	Hs.166684
TCF/HNF4a	hepatocyte nuclear factor 4α, MODY1 gene	NM_000457	Hs.54424
TDE1	tumor differentially expressed/Diff33	NM_006811	Hs.272168
TNNC2	troponin C2, fast	NM_003279	Hs.182421
TOM34	putative outer mitochondrial membrane 34-kD translocase	NM_006809	Hs.76927
UBCH10	ubiquitin carrier protein E2-C	NM_007019	Hs.93002
WISP2	WNT1 inducible signaling pathway protein 2	NM_003881	Hs.194679

[a]All genes include full codons, unless otherwise indicated.
[b]Partial codons only.
[c]Patent sequence.
[d]Fossey *et al.*, unpublished data.

type 2 diabetes patient (DB1). We detected a previously reported SNP, G239A, which resulted in a Lys80Arg substitution [22] in 12% of type 2 diabetes and in 14% of control chromosomes. We detected six sequence variants within the gene *MMP9*, a member of a group of secreted zinc metalloproteases that degrade the collagens of the extracellular matrix [23]. We detected five previously reported SNPs [24], two of which involved amino acid substitutions. A C60T transition resulted in an Ala20Val substitution, and SNP A836G led to an Arg279Glu substitution. We also detected two SNPs in the promoter region and one in the 3′ UTR of *MMP9*. We evaluated one sequence variant identified in dbSNP, rs13969 (A1821C, (Gly607Gly)), and identified a novel allele, C1722G (Pro574Arg), but the frequencies of the SNP alleles did not differ between type 2 diabetes patients and control populations examined. We identified four novel SNPs (one cSNP) within the phospholipid transfer protein *PLTP*. A458G produced a Met458Val substitution in 15% of type 2 diabetes patients and in 16% of control chromosomes analyzed. Analysis of the protein kinase C binding protein (*PRKCBP1*) revealed two sequence variants, C1413T and C198T, neither of which produced amino acid changes [25]. We detected the C1413T SNP detected in 25% of type 2 diabetes and in 24% of control chromosomes evaluated. We detected the C198T SNP in one individual with type 2 diabetes. We detected four novel SNPs (1 cSNP) in the evaluation of retinol binding-like protein *RBPSUHL*. The cSNP C27611A occurs within the wobble base of the codon and has no functional consequence. The allelic frequencies observed did not differ between the two population groups. Our analysis of HNF-4α, the MODY1 gene [26], detected five previously reported sequence variations [8,10]. We also identified two novel alleles, a 7-bp deletion (5′-GGAGGGC-3′) in the proximal promoter region in one type 2 diabetes patient and an Arg324His mutation in exon 8 of a type 2 diabetes patient [11].

Evaluation of the 10 ESTs identified five novel SNPs in five ESTs. We did not detect any allelic variants in EST328688, stSG9728, stSG34035, WI-4548, or WI-9189. We detected one SNP in each of the following ESTs: SGC30446, stSG25154, stSG4132, WI-6969, and WI-8404. The frequencies of the identified alleles did not differ between the two populations evaluated.

Discussion

To facilitate the identification of a type 2 diabetes gene within the 20q12–q13.1 interval, we have constructed a 6.0-Mb sequence-based BAC/YAC transcript map encompassing the linkage disequilibrium peaks detected around *ADA* and *D20S888* [6]. The transcript map we have generated from a combination of traditional BAC/YAC contig mapping techniques and genomic sequence alignment provides increased resolution to confidently determine the precise order for markers and sequences whose locations were previously ambiguous, such as *D20S176* [19].

There is one gap (about 400 kb) in the BAC/YAC contig between YAC clones 857H11 and 845A8. No genes or EST/STS markers mapped to this region of the contig, although genomic sequence clones span the interval. This region is adjacent to the *D20S16* locus. We previously reported the *D20S16* locus structure is highly polymorphic, consisting of a complex pattern of interspersed repeats of a tandemly reiterated sequence [27]. So far, we have been unable to isolate a stable YAC clone spanning this region of chromosome 20.

Our YAC/BAC map was useful for resolving ambiguities in the sequence map during the early stages of map construction when less HGP sequence was available for the region and a greater percentage of the clone sequence was assembled to only draft phase. By inspection of the map orders, putative false positives and negatives, chimeric clones, and nonunique STS markers can be identified for follow-up lab analysis, followed by iteration of the process with the amended marker data.

The alignment of genomic sequence clones provides near-continuous coverage between SGC32867 and WI-17676, with the exception of an estimated 80-kb gap between Z95330 and AL357558, which is included in the current estimated map size of 5832 kb. As several genomic sequence clones are unfinished, the precise map intervals and clone sizes will change as these clones approach a finished status and therefore may lead to map contraction or expansion. The finished assembly of clone AL161944, currently consisting of 14 unordered, unoriented contigs, could increase the map size by up to 150 kb. Clones with smaller numbers of contigs are likely to contain contigs that are correctly ordered and oriented, but each could still expand the map by 50–100 kb. Because the sequence gap size is bounded by the size of a single Caltech D1 BAC (less about 48 kb of overlapping sequence), the gap could be as large as 150 kb (assuming an upper BAC limit of 200 kb). Therefore, we roughly estimate the upper limit of the map size as 5832 kb+ 150 kb + 2 × 100 kb) + (150 kb – 80 kb) = 6252 kb.

Several maps have been published that include parts of the type 2 diabetes linked interval [28,29]. In general, our contig map agrees with these maps. However, the order of markers and genes between PABC1 and PLTP in our transcript map differs significantly from the recently published Wang map [29]. From the combination of BAC/YAC contig assemblies and genomic sequence alignment, we ordered the genes from centromere to telomere, *PABC1-SLPI-SDC4-RBPSUHL-HE4-PPGB-PLTP*. In contrast, the Wang map [29] orders the same genes *PABC1-HE4-SDC4-RBPSUHL-SLPI-PLTP-PPGB*. We believe the combination of traditional BAC/YAC contig mapping techniques and genomic sequence alignment within this region provides increased resolution to precisely order markers and therefore supports our arrangement.

The sequence-based transcript map we have generated provides us with the genomic resources necessary for detailed molecular analysis of candidate genes within the chromosome 20 type 2 diabetes susceptibility region. We have localized 42 genes, 43 unique ESTs, and 38 unique STSs within our sequence-based map, and we identified 68 UniGene clusters.

doi:10.1006/geno.2001.6584, available online at http://www.idealibrary.com on IDEAL

TABLE 3: Identified UniGene clusters within the 20q12–q13.1 transcript map

UniGene cluster	ESTs	HGP sequence	UniGene cluster	ESTs	HGP sequence
Hs.10087	D20S1114	AL034424	Hs.252189	—	AL021578
Hs.10117	stSG9725	AL133227	Hs.26213	D20S576	AL050348
Hs.102867	stSG3042	AL034424	Hs.264157	WI-21844	AL035662
Hs.106233	D20S767	AL139352	Hs.26608	stSG40369	AL034419
Hs.109315	WI-13364	AL035106	Hs.267458	stSG42524	AL162458
Hs.112341	stSG34035	AL049767	Hs.270001	WI-31223	AL031687
Hs.117780	—	Z93016	Hs.2719	SGC30446	AL031663
Hs.118126	WI-9189	AL008726	Hs.272168	WI-8404	Z97053
Hs.118249	WI-17092	AL121903	Hs.272285	stSG25440381	AL050348
	WI-17563		Hs.272520	—	AL031663
Hs.121031	stSG21378	AL133520	Hs.278489	—	AL021578
Hs.121084	—	AL031663	Hs.279920	—	AL008725
Hs.1217	stSG20349	M13792	Hs.282990	stSG34084	AL031055
Hs.134594	—	AL035447	Hs.283007	stSG34025	AL008726
Hs.150853	—	Z93016	Hs.283869	—	AL132772
Hs.151738	WI-7659	AL162458	Hs.283476	stSG4132	AL008726
Hs.178603	SGC44522	AL031055	Hs.288058	A007R18	AL049450
Hs.180016	—	AL049767	Hs.29279	WI-14748	AL049540
Hs.182351	WI-16033	AL021578	Hs.30793	WI-17691	AL034224
Hs.182421	—	AL050348	Hs.3407	D20S1067	Z97053, M13792
Hs.184930	—	AL031663	Hs.35140	stSG33865	Z93016
Hs.186571	D20S1038	AL118522	Hs.37372	SGC34777	AL031666
Hs.190075	stSG44390	AL035662	Hs.43857	stSG3026	AL034418
Hs.194679	—	AL139352	Hs.54424	D20S1127	AL132772
Hs.1968	—	AL049767	Hs.6113	WI-17676	AL133174
Hs.21413	stSG2530	AL162458	Hs.6511	stSG35545	AL008726
	WI-9597		Hs.6777	stSG3045	AL133520
Hs.22237	—	AL110279		NIB-1800	
Hs.225977	AL034418		Hs.6891	stSG27381	AL031861
Hs.226666	stSG20146	M13792		SGC32867	
Hs.247855	—	AL031686	Hs.75871	stSG33870	AL031666, AL049540
Hs.248217	—	AL021578	Hs.76927	stSG20133	AL109839
Hs.24994	—	AL121886	Hs.84038	stSG20089	AL021578
Hs.250824	D20S1069	Z93016	Hs.90073	stSG9728	AL133174
	EST328688			A003P30	
Hs.251754	WI-6969	AL035660	Hs.93002	—	AL050348
Hs.251946	D20S1113	AL109839	Hs.96560	stSG22763	AL133227

TABLE 4: Allelic variations detected within candidate genes and ESTs

Gene/EST	Exons	Polymorphism	SNP ID	Patient frequency	Control frequency	Fisher's *P* exact
ADA	12	G227A (Arg76Glu)	Novel[a]	0.005	–	–
		G239A (Lys80Arg)	102700.0001[b]	0.12	0.14	0.39
CEBPB	1	–	–	–	–	–
EYA2	3	–	–	–	–	–
HS1	5	C366T (Thr123Thr)	Novel[a]	0.23	0.18	0.08
KCNS1	4	–	–	–	–	–
MMP9	13	C[–2118]T[c]	–	0.01	0.03	0.15
		C[–1562]T[c]	–	0.14	0.14	–
		C60T (Ala20Val)[c]	–	0.01	0.02	0.61
		A836G (Arg279Gln)[c]	–	0.30	0.35	0.21
		C1722G (Pro574Arg)	Novel[a]	0.04	0.06	0.24
		A1821C (Gly607Gly)	rs13969[d]	0.36	0.42	0.16
		3′UTR C[+6bp]T[c]	–	0.38	0.35	0.37
PABC1	6	–	–	–	–	–
PLTP1	14	Intron2 T→G	rs435306[d]	0.35	0.38	0.30
		intron4 G→A	Novel[a]	0.2	0.17	0.26
		A544G (Met153Val)	Novel[a]	0.15	0.16	0.45
		intron11 T→C	Novel[a]	0.28	0.25	0.29
PRKCBP1	9	C1413T (Thr477Thr)[e]	Novel[a]	0.25	0.24	0.46
		C198T (Ser66Ser)[e]	Novel[a]	0.005	–	–
RBPSUHL	11	G19561A[f]	Novel[a]	0.35	0.33	0.40
		C20211T[f]	Novel[a]	0.05	0.07	0.30
		C27611A[f]	Novel[a]	0.03	0.03	–
		G28205A[f]	Novel[a]	0.22	0.25	0.31
SDC4	5	C138T (Ser46Ser)	Novel[a]	0.3	0.32	0.40
STK4/KRS2	5	C129T (Ser43Ser)	Novel[a]	0.34	0.36	0.28
		intron 3 C→T	Novel[a]	0.21	0.19	0.36
TC14/HNF4A[g]	10	1066–1071 delGGAGGGC	Novel[a]	0.002	–	–
		C114T (T38I)[h]	–	0.03	0.01	0.16
		G718A (R324H)	Novel[a]	0.005	–	–
		G1288A[h]	–	0.05	0.04	0.45
		G1563A (V521M)[i]	–	0.11	0.08	0.25
		T3142C[h]	–	0.46	0.47	0.48
		C3175T[h]	–	0.21	0.26	0.16
EST328688	–	–	–	–	–	
SGC30446	G101A	rs9880[d]	0.12	0.14	0.37	
stSG25154	G294A	rs321476[d]	0.05	0.07	0.29	
stSG34035	–	–	–	–	–	
stSG4132	C121T	rs197665[d]	0.16	0.14	0.31	
stSG9728	–	–	–	–	–	
WI-4548	–	–	–	–	–	
WI-6969	G101A	rs8282[d]	0.12	0.14	0.37	
WI-8404	G121A	rs13786[d]	0.33	0.29	0.26	
WI-9189	–	–	–	–	–	

[a]Novel allele identified; submitted to dbSNP database.
[b]OMIM allelic variant.
[c]Alleles described by Zhang, *et al.* [24].
[d]NCBI dbSNP ID.
[e]Alleles described by Fossey, *et al.* [25].
[f]Nucleotide numbers refer to position within AL021578.
[g]Nucleotide numbers refer to GenBank accession no. HSHNF4AS01.
[h]Alleles described by Malecki, *et al.* [10].
[i]Alleles described by Moller, *et al.* [8].

doi:10.1006/geno.2001.6584, available online at http://www.idealibrary.com on IDEAL

The distribution of these transcripts is nonrandom, with 34 of the 68 UniGene clusters localized in the 1.5-Mb interval between *D20S824* and WI-31223, which indicates a gene-rich region. An 800-kb interval between LOC51098 and HNF-4α contains no mapped genes or UniGene clusters. The 1-Mb interval between LOC1247 and KIAA1415 is also devoid of identified transcripts. This is the same region for which no BAC or stable YAC clones have been identified.

The multiple molecular interactions that contribute to the diabetes phenotype remain elusive, which increases the difficulty in selecting genes to evaluate as diabetogenic candidates. Defects in insulin-stimulated glucose uptake and intracellular glucose metabolism seem to be responsible for the peripheral insulin resistance observed in type 2 diabetes [1,30–33]. However, no common causative point mutations in glucose transporters, the insulin gene, or the insulin receptor have been identified, although risk-incurring extended haplotypes or more complex epistatic effects cannot be ruled out. To begin a systematic evaluation of the transcripts localized within our map, we organized the 42 identified genes into four categories: genes with an established role in glucose metabolism, transcription factors, genes that may participate in signaling pathways, and genes whose function does not suggest an obvious role in the biological processes contributing to diabetes. We used SSCP techniques to screen the coding, proximal promoter, and 3′ sequences of 13 genes for allelic variants that may be associated with diabetes (Table 4). We identified 16 coding SNPs (9 of which resulted in an amino acid substitution) and a 7-bp deletion within the 13 candidate genes evaluated. Concurrently, we scanned 10 ESTs for allelic variants and identified five SNPs. We deposited novel SNPs in the dbSNP database. From analysis of the candidate genes and ESTs, we did not find evidence of a common coding mutation associated with type 2 diabetes. It may be appropriate to repeat these analyses on an expanded set of patients and controls in the future. It should be noted that the method we used to survey for allelic variants, SSCP, is at best only 80% efficient at detecting sequence variants. It may be fruitful to survey some of these genes for allelic variants with a more sensitive method. We are systematically evaluating mapped genes and transcripts in an effort to identify diabetogenic alleles.

There has been considerable interest recently in the use of SNPs to evaluate the genetic component of complex diseases [34,35]. The SNPs we have identified are valuable loci, which we can use to construct a dense SNP map of this clinically important region.

We have constructed a 6.0-Mb sequence-based BAC/YAC transcript map of the type 2 diabetes-linked interval on chromosome 20q12–q13.1. The combination of reference HGP sequence, BAC transcript map, YAC scaffold, and SNP localization has generated one of the most comprehensive maps available for this clinically important region. We have begun to systematically evaluate the localized genes and expressed sequences in an effort to find alleles that may contribute to type 2 diabetes.

Materials and Methods

Mapping ESTs, STSs, genes, and novel transcripts. The genetic markers, ESTs, STSs, genes, and novel transcripts analyzed in this study were identified from the framework physical map of this region described by Price *et al.* [19], the Sanger Centre chromosome 20 sequencing project (http://www.sanger.ac.uk/HGP/Chr20), and previously published maps (GeneBridge 4 RH map at WI/MIT [36], G3 radiation hybrid map at Stanford Human Genome Center [37], and an independent YAC map [38]). We determined STS retention patterns in individual BAC and YAC clones by PCR screenings conducted in triplicate, using primer pairs unique for each marker (Table 1).

BAC and YAC clone isolation. We identified novel BAC clones from the Human CITB BAC Library version 4.0 [20] (Research Genetics, Huntsville, AL) by sequential PCR screenings using markers in the region. We isolated DNA from colony-purified BAC clones with a Qiagen midi kit (Qiagen, Chatworth, CA). BAC clone DNA was digested with *Not*I restriction endonuclease (Promega, Madison, WI) and sized by pulsed-field gel electrophoresis with a CHEF II Mapper (Bio-Rad, Hercules, CA.). The YAC clones used to construct the genomic scaffold were identified and isolated as described [19].

Construction of the transcript map. We used two concurrent strategies to construct the transcript map. To help automate the physical map assembly from our BAC STS/marker screen data, we designed and wrote GraphMap, a Java-based computer program. GraphMap uses a local, greedy search algorithm for the best Hamiltonian path through the markers, similar to approaches that find a maximal spanning tree [39,40] but with extra local decision-making heuristics that use information about clones hybridized to a particular marker and its immediate neighbors. A double breadth-first traversal of the marker-clone links identifies contigs that are at least doubly linked [41] and forms a structure graph to enable identification of the markers in the extremity layers [42,43]. Each contig is recast as a weighted marker intersection graph, with vertices corresponding to markers and weighted edges corresponding to the number of clones linking each marker pair. The best Hamiltonian path through the graph is computed for each contig, initiated with an extremity layer marker. At each vertex where the path has multiple next choices, the graph edge with the maximum weight is selected. In cases of ambiguity, local clone identity heuristics are used to refine the selection of the probable path from one marker to the next in the map. This is repeated sequentially, initiated with all candidate extremity layer markers. An overall map quality score is accorded to each marker permutation in the contig and used to select interesting marker map permutations that minimize the contig map score. BAC contigs assembled via this process were visually aligned and oriented within the YAC scaffold by the observed retention patterns in the markers shared with the BACs. This straightforward approach works well here because the genetic distance in the map is a small segment of a single chromosome, BACs have a low chimerism rate compared with YACs, and the approach is applied in a semiautomated fashion with manual verification.

Concurrently, a sequence-based map was assembled. Seed marker sequences downloaded from NCBI dbSTS and dbEST databases (http://www.ncbi.nlm.nih.gov/Entrez) were used to retrieve HGP clone sequences from the corresponding NCBI htgs and nr databases and from the Sanger Centre chromosome 20 project (http://www.sanger.ac.uk/HGP/chr20). BAC end sequences of 100–500 bp automatically generated from the clones and after repeat masking (RepeatMasker software, A. F. Smit and P. Green, unpublished data) were used to iteratively probe NCBI htgs and nr databases for overlapping clones using BLAST [44]. With a local augmented BLAST database of the region containing 345 sequences and 10,820,697 letters, 100 bases of clone end query sequence were perfectly, unambiguously aligned with Expect = $1e^{-51}$.

We calculated all physical map distances and the total interval size using the known sizes of HGP clone sequences, with allowances for clone overlaps and estimated gap sizes. We located markers in sequence clones with BLAST alignments to verify YAC/BAC marker order generated by the GraphMap program and used to calculate intermarker distances. We calculated approximate map positions for the marker-screened YAC and BAC clones by centering the clones at the mean map position of the outermost pair of screened markers that lie within the clones, thereby equally apportioning the excess lab-measured clone distance (compared with the maximum intermarker distance) between centromeric and telomeric ends. The two assembled maps were regularly compared and used for mutual refinement.

Transcript annotation. We used EST/STS marker and human genome project sequences to search the UniGene database at NCBI (http://www.ncbi.nlm.nih.gov/UniGene/) and to identify and verify gene and EST clusters (Tables 2 and 3).

Identification of novel allelic variants in identified genes and ESTs and association with type 2 diabetes. We used SSCP analysis to screen expressed sequences including identified coding regions, proximal 5′ promoter, and 3′ untranslated sequences of candidate genes (Table 4) in 100 unrelated Caucasian type 2 diabetes patients and 100 unrelated Caucasian healthy controls. Ascertainment and other characteristics of the patients have been described in detail [11]. PCRs were performed with primer pairs end labeled with [γ-^{32}P]dATP (ICN Radiochemicals, Irvine, CA). The resultant products were denatured and analyzed by electrophoresis on native 0.5% MDE/5% glycerol (FMC Products, Rockland, ME) gels in 0.6 $\times$ tris-borate EDTA at 15 W for 15 h. Gels were exposed overnight to X-ray film (Fuji, Stamford, CT) between intensifying screens.

We calculated SNP allele frequency differences between type 2 diabetes patients and controls by Fisher's exact procedure in 2 $\times$ 2 contingency tables. We determined the power to detect differences in SNP allele frequency between the patients and controls based on the frequency of the SNP in the patient group and the SNP frequency in the control group. For these calculations, the type 1 error rate (α) is 5%. For example, if a SNP allele has a frequency of 0.40 in the controls, we can detect (with 81% power) a SNP with 0.60 frequency in patients (equivalent to an odds ratio of 2.25). Similarly, with 100 patients and 100 controls, we have greater than 80% power for SNP alleles with 0.25 frequency in controls and greater than 0.45 frequency in patients (odds ratio of 2.45). Thus, this study is adequately powered to detect SNPs whose risk allele increases overall risk by about 2.5-fold (an odds ratio of 2.5 with a 95% confidence interval of 0.56–10.70). It should be noted, however, that should the control SNP allele frequencies be higher than expected, the power would be reduced.

Patient populations. Patients with type 2 diabetes and controls have been described [11].

ACKNOWLEDGMENTS

This work was supported by NIH grants R01-DK56289 and R01-DK53591 (D.W.B.) and R01-HL56266 (B.I.F.).

RECEIVED FOR PUBLICATION MARCH 3; ACCEPTED APRIL 23, 2001.

REFERENCES

1. Kahn, C. R., Vicent, N. D., and Doria, A. (1996). Genetics of non-insulin-dependent (type II) diabetes mellitus. *Annu. Rev. Med.* **47:** 509–531.
2. Bowden, D. W., *et al.* (1997). Linkage of genetic markers on human chromosomes 20 and 12 to NIDDM in Caucasian sib pairs with a history of diabetic nephropathy. *Diabetes* **46:** 882–886.
3. Ji, L., *et al.* (1997). New susceptibility locus for NIDDM is localized to human chromosome 20q. *Diabetes* **46:** 876–881.
4. Zouali, H., *et al.* (1997). A susceptibility locus for early-onset non-insulin dependent (type 2) diabetes mellitus maps to chromosome 20q, proximal to the phosphenolpyruvate carboxykinase gene. *Hum. Mol. Genet.* **6:** 1401–1408.
5. Ghosh S., *et al.* (1999) Type 2 diabetes: evidence for linkage on chromosome 20 in 716 Finnish affected sib pairs. *Proc. Natl. Acad. Sci. USA* **96:** 2198–2203.
6. Price, J. A., *et al.* (1997). Construction of a physical map of chromosome 20q12–13.1 and linkage disequilibrium analysis in diabetic nephropathy patients. *Am. J. Hum. Genet.* **61** (suppl.): A241.
7. Furuta, H., *et al.* (1997). Organization and partial sequence of the hepatocyte nuclear factor-4 α/MODY1 gene and identification of a missense mutation, R127W, in a Japanese family with MODY. *Diabetes* **46:** 1652–1657.
8. Moller, A. M., *et al.* (1997). Studies of the genetic variability of the coding region of the hepatocyte nuclear factor-4α in Caucasians with maturity onset NIDDM. *Diabetologia* **40:** 980–983.
9. Hani, E. H., *et al.* (1998). A missense mutation in hepatocyte nuclear factor-4α, resulting in a reduced transactivation activity, in human late-onset non-insulin-dependent diabetes mellitus. *J. Clin. Invest.* **101:** 521–526.
10. Malecki, M. T., *et al.* (1998). Exclusion of the hepatocyte nuclear factor 4α as a candidate gene for late-onset NIDDM linked with chromosome 20q. *Diabetes* **47:** 970–972.
11. Price, J. A., *et al.* (2000). Analysis of the HNF4α gene in Caucasian type II diabetic-nephropathic patients. *Diabetologia* **43:** 364–372.
12. Berry, R., *et al.* (2000). Evidence for a prostate cancer-susceptibility locus on chromosome 20. *Am. J. Hum. Genet.* **67:** 82–91.
13. Moser, K. L., *et al.* (1998). Genome scan of human systemic lupus erythematosus: evidence for linkage on chromosome 1q in African-American pedigrees. *Proc. Natl. Acad. Sci. USA* **95:** 14869–14874.
14. Lee, J. H., *et al.* (1999). Genome scan for human obesity and linkage to markers in 20q13. *Am. J. Hum. Genet.* **64:** 196–209.
15. Heim, S., and Mitelman, F. (1995). *Cancer Cytogenetics*. Wiley Liss, New York.
16. Asimakopoulos, F. A., and Green, A. R. (1996). Deletions of 20q and the pathogenesis of myeloproliferative disorders. *Br. J. Haematol.* **95:** 219–226.
17. Stumpf, E., *et al.* (2000). Chromosomal alterations in human pancreatic endocrine tumors. *Genes Chromosom. Cancer* **29:** 83–87.
18. Larramendy, M. L., *et al.* (2000). Comparative genomic hybridization reveals complex genetic changes in primary breast cancer tumors and their cell lines. *Cancer Genet. Cytogenet.* **119:** 132–138.
19. Price, J. A., *et al.* (1999). A physical map of the 20q12–13.1 region associated with type2 diabetes. *Genomics* **62:** 208–215, doi:10.1006/geno.
20. Shizuya, H., *et al.* (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl. Acad. Sci. USA* **89:** 8794–8797.
21. Daddona, P. E., *et al.* (1994). Human adenosine deaminase: cDNA and complete primary amino acid sequence. *J. Biol. Chem.* **259:** 12101–12106.
22. Valerio, D., *et al.* (1986). One adenosine deaminase allele in a patient with severe combined immunodeficiency contains a point mutation abolishing enzyme activity. *EMBO J.* **5:** 113–119.
23. Nagase, H., Barrett, A. J., and Woessner, J. F., Jr. (1992). Nomenclature and glossary of the matrix metalloproteinases. *Matrix Suppl.* **1:** 421–424.
24. Zhang, B., *et al.* (1999). Genetic variation at the matrix metalloproteinase-9 locus on chromosome 20q12.2–13.1. *Hum. Genet.* **105:** 418–423.
25. Fossey, S. C., *et al.* (2000). Identification and characterization of PRKCBP1, a candidate RACK-like protein. *Mamm. Genome* **11:** 919–925, doi:s003350010174.
26. Yamagata, K., *et al.* (1996). Mutations in the hepatocyte nuclear factor-4a gene in maturity-onset diabetes of the young (MODY1). *Science* **384:** 458–460.
27. Bowden, D. W., *et al.* (1995). D20S16 is a complex interspersed repeated sequence: genetic and physical analysis of the locus. *Genomics* **25:** 394–403, doi:10.1006/geno.
28. Bench, A. J., *et al.* (1998). A detailed physical and transcriptional map of the region of chromosome 20 that is deleted in myeloproliferative disorders and refinement of the commonly deleted segment. *Genomics* **49:** 351–362, doi:10.1006/geno.
29. Wang, P. W., *et al.* (2000). Refinement of the smallest commonly deleted segment of chromosome 20 in malignant myeloid diseases and development of a PAC-based physical and transcription map. *Genomics* **67:** 28–39, doi: 10.1006/geno.2000.6215.
30. Butler, P. C., Kryshak, E. J., Marsh, M., and Rizza, R. A. (1990). Effect of insulin on oxidation of intracellularly and extracellularly derived glucose in patients with NIDDM. Evidence for primary defect in glucose transport and/or phosphorylation but not oxidation. *Diabetes* **39:** 1373–1380.
31. Cline, G. W., *et al.* (1999). Impaired glucose transport as a cause of decreased insulin-stimulated muscle glycogen synthesis in type 2 diabetes. *N. Engl. J. Med.* **341:** 240–246.
32. Shepherd, P. R., and Kahn, B. B. (1999). Glucose transporters and insulin action: implications for insulin resistance and diabetes mellitus. *N. Engl. J. Med.* **341:** 248–257.
33. Tirosh, A., Rudich, A., and Bashan, N. J. (2000). Regulation of glucose transporters—implications for insulin resistance states. *Pediatr. Endocrinol. Metab.* **13:** 115–133.
34. Kruglyak, L. (1997). The use of a genetic map of biallelic markers in linkage studies. *Nature Genet.* **17:** 21–24.
35. Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* **22:** 139–144.
36. Hudson, T. J., *et al.* (1995). An STS-based map of the human genome. *Science* **270:** 1945–1954, with supplementary data from the Whitehead Institute/MIT Center for Genome Research, Human Genetic Mapping Project, Data Release 11 (October 1996).
37. Stewart, E. A., *et al.* (1997). An STS-based radiation hybrid map of the human genome. *Genome Res.* **7:** 422–433.
38. Stoffel, M., *et al.* (1996). A yeast artificial chromosome-based map of the region of chromosome 20 containing the diabetes-susceptibility gene, MODY1, and a myeloid leukemia related gene. *Proc. Natl. Acad. Sci. USA* **93:** 3937–3941.
39. Mott, R., Grigoriev, A., Maier, E., Hoheisel, J., and Lehrach, H. (1993). Algorithms and software tools for ordering clone libraries: application to the mapping of the genome of *Schizosaccharomyces pombe*. *Nucleic Acids Res.* **8:** 1965–1974.
40. Nadkarni, P. M., *et al.* (1996). CONTIG EXPLORER: interactive marker content map assembly. *Genomics* **31:** 301–310, doi:10.1006/geno.
41. Arratia, R., Lander, E. S., Tavare, S., and Waterman, M. S. (1991). Genomic mapping by anchoring random clones: a mathematical analysis. *Genomics* **11:** 806–827, doi:10.1006/geno.
42. Harley, E., Bonner, A. J., and Goodman, N. (1996). Good maps are straight. Proc. 4th International Conference on Intelligent Systems for Molecular Biology (ISMB-96), pp. 161–169, American Association of Artificial Intelligence Press, Menlo Park, CA.
43. Harley, E., Bonner, A. J., and Goodman, N. (1999). Revealing hidden interval-graph structure in STS-content data. *Bioinformatics* **15:** 278–285.
44. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410, doi:10.1006/jmbi.
45. Donis-Keller, H., *et al.* (1987). A genetic linkage map of the human genome. *Cell* **51:** 319–337.