Supplemental material

Materials and methods

Preparation of BAC library CHORI-507

The CHORI-507 BAC library was prepared as described [Osoegawa et al. 1998]. Briefly, high-molecular-weight DNA from WAV 17 cells was prepared in agarose using DNA Plug Kits (BioRad). The DNA was partially digested with MboI (New England Biolabs) to generate fragments between 100 kb and 400 kb, which were size-fractionated by pulsed-field electrophoresis [Osoegawa et al. 1998]. DNA fractions in the 150 to 250 kb range were sliced from the gel and DNA was electro-eluted from the gel matrix. Vector DNA (pTARBAC1.3) was prepared by treatment with a combination of restrictions enzymes (ApaLI and BamHI) and alkaline phosphatase and then purified by gel-electrophoresis. *E. coli* DH10B T1 phage-resistant cells (Invitrogen) were transformed by electroporation with ligation products from vector and genomic DNA. Prior to full-scale library production, 379 random clones were picked for DNA size analysis. BAC DNA, prepared using an Autogen960 robot, was digested with NotI (New England Biolabs), and analyzed by pulsed-field electrophoresis in a 1% agarose gel. The library clone size distribution is given in Supplemental Fig. 2, and the average insert size was estimated to be 141 kb. There is a small fraction (1 out of 379) of non-recombinant clones. Following this quality assessment, 221 184 BAC clones (approximately 10-fold coverage) were arrayed into 576 384-well microtiter plates. Duplicate sets of twelve high-density colony filters were created for hybridization screening of the library with various human-specific repeat probes.

BAC end sequencing

Positive clones from CHORI-507 were end sequenced using the following protocol. BAC clones were inoculated into 1.5ml of 2xTY containing 12.5μg/ml chloramphenicol in 2ml 96 well growth boxes (Costar) and incubated for 22 hours at 37°C with shaking at 320rpm. Plates were centrifuged at 4000rpm for 3 min to obtain pellets, the supernatant was discarded and the cells were re-suspended in 100μl of GTE+RNaseA, 100μl of NaOH/SDS was added and mixed before adding 100μl of 3M KOAc and further mixing. After filtration through two Millipore plates (MADVN6550/MANUBA50) on a vacuum manifold, the resultant material was dried and re-suspended in 35μl of 10mM Tris pH8.0. The resulting DNA was sequenced with BigDye Terminator Ready Mix v3.0 and BigDye Bufferx5 (Applied Biosystems) using T7 and SP6 primers. The sequenced products were cleaned up by washing with 5μl 3M NaOAc and 125μl 96% ethanol. The DNA was precipitated by centrifugation at 4000rpm for

15 mins, then washed with a further 100μl 70% ethanol and the centrifugation repeated. The products were re-suspended in 10μl of 0.1M EDTA, pH 7.4 and loaded on to ABI3700 capillary sequencers.

BAC clones were fingerprinted using HindIII restriction enzyme fingerprinting [Humphray et al. 2001; Schein et al. 2004]. BAC clones were directly cultured in 170μl 2xTY in 96 well plates. After overnight growth the BAC DNA was extracted by alkaline lysis, and digested with HindIII. Following electrophoresis on 121 lane 1% agarose gels the data are collected using a Typhoon 8600 fluorimager, raw images were entered into the fingerprint database using IMAGE software (http://www.sanger.ac.uk/Software/Image). The output of normalized band values, sizes and gel traces were analyzed in FPC [Soderlund et al. 2000], which bins and orders clones on the basis of shared bands. Sequence tilepaths were identified following inspection of assembled contigs. End sequences were aligned to the human and mouse genomes using BLAST and ssaha2 (www.ensembl.org).

BAC sequencing

For the shotgun phase [Bankier et al. 1987], pUC plasmids with inserts of mostly 1.4-2 kb were sequenced from both ends using the dideoxy chain termination method [Sanger et al. 1977] with different versions of big dye terminator chemistry [Rosenblum et al. 1997]. The resulting sequencing reactions were analyzed on ABI 3730 sequencing machines and the generated data were processed (http://www.sanger.ac.uk/Software/sequencing/) and assembly with PHRED [Ewing and Green 1998; Ewing et al. 1998] and PHRAP (http://www.phrap.org/) algorithms. For the finishing phase, we used GAP4 [Bonfield et al. 1995] to help assess, edit and select reactions to eliminate ambiguities and close sequence gaps. Sequence gaps were closed by a combination of primer walking, PCR, short/long insert sublibraries [McMurray et al. 1998], oligo screens of such sublibraries and transposon sublibraries.

Gene model copy number

We also performed absolute quantification [Hijri and Sanders 2005] using one assay for gene models 25.1, 28.1, 6.2 and two assays for gene model 15.1. For each gene model, a known amount of plasmid DNA containing a 50mer insert was used as a standard. Two-fold serial dilutions (ranging between 1863 and 3.63 pg) of the purified plasmid DNA were included in the experiment to generate the standard curve. Real-time PCR on the plasmid DNA was performed in six technical replicates (10 dilutions per replicate). The Ct values were then plotted against the log of the initial vector concentration containing the amplicon insert. All replicates offered very similar results producing a standard curve with regression coefficients

(R^2) >0.99. This regression was then used for predictions of all gene models copy number for a given Ct value. In the same experiments, two-fold serial dilutions of genomic DNA (ranging between 3200 and 6.25 pg) were included in each experiment to generate the Ct values for nuclear DNA copy number estimation. The number of copies of each gene model could then be calculated for each sample of target DNA, based on the regressions using the plasmid DNA as standards. From this we calculated the copy number of the each gene model per 3.3pg of gDNA (the amount of DNA per each copy of the entire human genome). As a control, we used herring sperm DNA as a carrier to yield 10 ng of total DNA per sample in a total volume of 40 µl, irrespective of the dilution of the target DNA to show that the estimation of copy number is not biased by the amount of non-target DNA added to the reaction.

WAV17

The somatic cell hybrid WAV 17 [Slate et al. 1978] was obtained from Coriell (ccr.coriell.org/nigms; GM08854) and cultured according to the recommended protocols. The monoallelic composition of HSA21 in WAV17 was assessed by genotyping with markers D21S11, D21S1270, D21S1435, D21S1411, D21S226 and *IFNAR*.

Screening BAC library CHORI-507

To identify human clones, CHORI-507 was screened with ^{32}P labeled α-satellite and Alu probes, and total human genomic DNA. This resulted in 2400 (1.09%) positive clones after screening using α-satellite and Alu probes, and 2752 (1.24%) positives with total human genomic DNA. This is consistent with the fact that the HSA21 content of WAV 17 is approximately 1-3% in an aneuploid mouse background. The non-redundant positive clones (3 208) were re-arrayed into 384-well plates and gridded on nylon membrane for screening with 21p-specific probes. To identify specific 21p BACs, overgo probes were designed to STSs corresponding to the YAC ends 2E4R, 2E4L, 4E9L, 4E9R, 1B8L, 2C2R, 2C9R from Wang *et al*. [Wang et al. 1999]. A summary of the library screens is given in Supplemental Table 1 and Supplemental Table 2.

FISH

Cell lines were grown in RPMI 1640 with Glutamax I medium (Invitrogen) supplemented with 10% fetal calf serum and a 1% penicillin/streptomycin mix. Cultures were exposed to colcemid (0.1 μg/mL; Invitrogen) for 1.5 h at 37ºC and harvested according to routine cytogenetic protocols. G-banding was performed following standard cytogenetic methods.

The following probes were used in FISH to check the integrity of HAS21 in the WAV 17 cell line: HSA21 paint (Vysis), cosmid c55A10 [Chen et al. 1999], D21Z1 [Maratou et al. 1999], human and mouse Cot-1 DNA (Invitrogen). Interphase nuclei and metaphase spreads were counterstained with DAPI (Vysis) diluted in Vectashield antifade (Vector Labs). BAC and cosmid DNA was labeled with either a Biotin or Digoxigenin-Nick Translation kit (Roche) and detected with anti-dig fluorescein (green) or avidin (red). Cells were viewed under a Nikon fluorescence microscope equipped with appropriate filter combinations. Monochromatic images were captured and superimposed using the Applied Imaging automated imaging system.

Gene predictions

EST sequences from dbEST [Boguski et al. 1993] were aligned to the HSA21p sequence using the program EXONERATE [Slater and Birney 2005]. We considered only spliced alignments with sequence identity of 90% or higher and at least 88% coverage, and discarded those for which there existed a better spliced alignment elsewhere in the genome. Selected EST alignments were divided into four groups according to the properties of the alignment: best, the best alignment for this EST is in HSA21p sequence; pseudo, the best alignment for this EST is in euchromatin, but this is not spliced, indicating a possible processed pseudogene in the euchromatin; paralogue, the coverage of the EST alignment is 100% but the percentage identity is not as high as the best match in euchromatin, hence a potential paralogue; random, there is a better alignment for this EST in an unassembled euchromatin contig. From these alignments we calculated the set of unique exon-exon pairs (75 in total).

Using RepeatMasker (www.repeatmasker.org) without the low complexity filter (option -nolow) we masked the HSA21p sequences and performed in silico gene prediction using the two *ab initio* gene prediction programs, GeneID [Parra et al. 2000] and GENSCAN [Burge and Karlin 1997], and the comparative gene predictor SGP [Parra et al. 2003], in this case using mouse genome sequence (version mm5) as reference. We thus obtained 3 sets of predictions which we merged together into a single set of predicted exon-exon pairs removing duplicates. We further removed those exon-exon pairs that overlap with any of the EST exon-exon pairs calculated before. We obtained a set of 182 unique exon-exon pairs obtained by *ab initio* and comparative gene prediction and not occurring in the EST-predicted set. We classified the 182 exon-exon pairs into those being predicted by all three gene predictors, by two of them and those that are specific to one gene predictor. This classification, the coordinates and the sequences of these predictions are available through the on-line Supplemental material website at http://genome.imim.es/datasets/hsa21p.

Exon-exon pairs were assembled into 26 gene models (21pGMs) by aligning them with the genomic sequence and combining exon pairs conforming to the splice-site consensus sequence.

Sequence analysis

For general sequence analysis programs within the EMBOSS package were used [Rice et al. 2000]. BLAST searches were carried out at Ensembl (www.ensembl.org). Segmental duplications were detected using a BLAST-based detection scheme (WGAC) 45 to identify all pairwise similarities representing duplicated regions (≥1 kb and ≥90% identity) within the finished BAC sequence of chromosome 21 p arm. The 21p BAC sequences were manually joined in their natural order and compared to all other chromosomes in the NCBI genome assembly (hg17 and 18). Satellite repeats were detected using RepeatMasker (version: 2002/05/15) on sensitive settings (www.repeatmasker.org) [Smit et al. 1996]. The program PARASIGHT (Bailey, unpublished) was used to generate images of pairwise alignments.

RT-PCR

Gene models were tested in 24 human cDNAs (brain, heart, kidney, spleen, liver, colon, small intestine, muscle, lung, stomach, testis, placenta, skin, PBLs, bone marrow, fetal brain, fetal liver, fetal kidney, fetal heart, fetal lung, thymus, pancreas, mammary glands, prostate). All amplicons spanned introns. Each cDNA was mixed with JumpStart REDTaq ReadyMix (Sigma) and 4 ng/ul primers (Sigma-Genosys) using a BioMek 2000 robot (Beckman) as described and modified [Reymond et al. 2002]. The first 10 cycles of PCR amplification were performed with a touchdown annealing temperature decreasing from 60 to 50ºC; annealing temperature of the next 30 cycles was carried out at 50ºC. Amplimers were separated on Ready-to-Run precast gels (Pharmacia) and sequenced.

Gene copy number

Relative DNA copy number was determined as described [Lyle et al. 2006] with some modifications. Genomic DNA was isolated from lymphoblastoid cell lines using DNeasy Tissue kits (Qiagen) according to the manufacturer's protocol. DNAs from a total of 13 individuals were used: 8 unrelated individuals, and mother and son from CEPH families (Coriell), and 3 unrelated caucasian individuals from the University of Geneva DNA bank. Oligonucleotides were designed using PrimerExpress (Applied Biosystems) with default parameters. For each gene prediction a minimum of two assays were designed. Seven assays were used for normalization of input DNA. Control assays were designed in regions that are present in one (*AKAP1*) and two copies (*HBG2*) per haploid genome. Assay sequences are

available in Supplemental Table 10.

Quantitative PCR reactions were performed using PowerSYBR Green PCR master Mix (Applied Biosystems) and each DNA sample were amplified in three replicates. Fluorescence data were collected using the ABI Prism 7900 Sequence Detection System (Applied Biosystems) and data analysis was performed with Qbase (Hellemans et al., unpublished).

Polymorphisms

RNA was obtained from from human brain, testis, heart, spleen, thymus and lung (Ambion). RNA from somatic cell hybrids carrying human chromosome 13, 14, 15, 21, and 22 (GM10898, GM10479, GM11418, GM08854, GM10888; Coriell) was extracted with RNeasy mini kit (Qiagen). cDNA was synthesized using SuperScript II (Invitrogen) and each predicted gene model was tested with the appropriate primer set (Supplemental Table 10) designed to span the gene model. Platinum *Pfx* DNA Polymerase (Invitrogen) was used to perform the PCR. After purification with QIAquick PCR Purification Kit (Qiagen), PCR products were cloned into pCR4Blunt-TOPO vector (Invitrogen). Purification of clones was performed using Plasmid Miniprep[96] kits (Millipore) and sequencing reaction was carried out with ABI Prism BigDye Terminator Cycle Sequencing Ready Reaction Kit (Applied Biosystems). Lasergene v6.0 (DNAStar) was used to quality check and align sequences. Alignments were edited and clustered using BioEdit (http://www.mbio.ncsu.edu/BioEdit/bioedit.html) and GeneDoc (http://www.psc.edu/biomed/genedoc).

References

Bankier, A.T., Weston, K.M., Barrell, B.G. 1987. Random cloning and sequencing by the M13/dideoxynucleotide chain termination method. *Methods Enzymol* **155:** 51-93.

Boguski, M.S., Lowe, T.M.J., Tolstoshev, C.M. 1993. dbEST - database for "expressed sequence tags". *Nature Genetics* **4:** 332-333.

Bonfield, J.K., Smith, K., Staden, R. 1995. A new DNA sequence assembly program. *Nucleic Acids Res* **23:** 4992-4999.

Burge, C., Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268:** 78-94.

Chen, H., Rossier, C., Morris, M.A., Scott, H.S., Gos, A., Bairoch, A., Antonarakis, S.E. 1999. A testis-specific gene, TPTE, encodes a putative transmembrane tyrosine phosphatase and maps to the pericentromeric region of human chromosomes 21 and 13, and to chromosomes 15, 22, and Y. *Hum Genet* **105:** 399-409.

Ewing, B., Green, P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8:** 186-194.

Ewing, B., Hillier, L., Wendl, M.C., Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8:** 175-185.

Hijri, M., Sanders, I.R. 2005. Low gene copy number shows that arbuscular mycorrhizal fungi inherit genetically different nuclei. *Nature* **433:** 160-163.

Humphray, S.J., Knaggs, S.J., Ragoussis, I. 2001. Contiguation of bacterial clones. *Methods*

Mol Biol **175:** 69-108.
Lyle, R., Radhakrishna, U., Blouin, J.L., *et al.* 2006. Split-hand/split-foot malformation 3 (SHFM3) at 10q24, development of rapid diagnostic methods and gene expression from the region. *Am J Med Genet A* **140:** 1384-1395.
Maratou, K., Siddique, Y., Kessling, A.M., Davies, G.E. 1999. Novel methodology for the detection of chromosome 21-specific alpha-satellite DNA sequences. *Genomics* **57:** 429-432.
McMurray, A.A., Sulston, J.E., Quail, M.A. 1998. Short-insert libraries as a method of problem solving in genome sequencing. *Genome Res* **8:** 562-566.
Osoegawa, K., Woon, P.Y., Zhao, B., Frengen, E., Tateno, M., Catanese, J.J., de Jong, P.J. 1998. An improved approach for construction of bacterial artificial chromosome libraries. *Genomics* **52:** 1-8.
Parra, G., Agarwal, P., Abril, J.F., Wiehe, T., Fickett, J.W., Guigo, R. 2003. Comparative gene prediction in human and mouse. *Genome Res* **13:** 108-117.
Parra, G., Blanco, E., Guigo, R. 2000. GeneID in Drosophila. *Genome Res* **10:** 511-515.
Reymond, A., Marigo, V., Yayiaoglu, M.B., *et al.* 2002. Human chromosome 21 gene expression atlas in the mouse. *Nature* **420:** 582-586.
Rice, P., Longden, I., Bleasby, A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16:** 276-277.
Rosenblum, B.B., Lee, L.G., Spurgeon, S.L., Khan, S.H., Menchen, S.M., Heiner, C.R., Chen, S.M. 1997. New dye-labeled terminators for improved DNA sequencing patterns. *Nucleic Acids Res* **25:** 4500-4504.
Sanger, F., Nicklen, S., Coulson, A.R. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Science USA* **74:** 5463-5467.
Schein, J., Kucaba, T., Sekhon, M., Smailus, D., Waterston, R., Marra, M. 2004. High-throughput BAC fingerprinting. *Methods Mol Biol* **255:** 143-156.
Slate, D.L., Shulman, L., Lawrence, J.B., Revel, M., Ruddle, F.H. 1978. Presence of human chromosome 21 alone is sufficient for hybrid cell sensitivity to human interferon. *Journal of Virology* **25:** 319-325.
Slater, G.S., Birney, E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6:** 31.
Smit, A.F.A., Hubley, R., Green, P. 1996. RepeatMasker Open-3.0.
Soderlund, C., Humphray, S., Dunham, A., French, L. 2000. Contigs Built with Fingerprints, Markers, and FPC V4.7. *Genome Research* **10:** 1772-11787.
Wang, S.Y., Cruts, M., Del Favero, J., *et al.* 1999. A high-resolution physical map of human chromosome 21p using yeast artificial chromosomes. *Genome Research* **9:** 1059-1073.

Supplemental figures and tables

Supplemental Fig. 1. Confirmation of HSA21 content of WAV 17 cell line. (**a**) G-banding of a single nucleus identifies that this cell contains four chromosomes 21 (red arrowheads). (**b**) FISH paint with HSA21 probe confirms the identity of the human chromosome in the mouse background. Counting of 100 nuclei showed that each cell contains 3-5 chromosomes 21. (**c**) Hybridization with the cosmid c55A10 confirms the presence of the short arm on HSA21 in the cell line. D21Z1 is a chromosome 21 centromere probe [Maratou et al. 1999]. (**d**) Hybridization with human Cot-1 DNA shows that HSA21 is the only human component of WAV 17 and that no mouse sequences could be detected on this chromosome.

Supplemental Fig. 2. Insert size distribution of CHORI-507 clones.

Supplemental Fig. 3. FISH of a 21p BAC clone to a metaphase spread from normal human diploid lymphocytes. Clone 68N6 hybridizes to 21p and 13p, as well as to the pericentromeric regions of 2 and 9. All clones hybridize to 21p, and to various extents to the other acrocentrics. +, positive hybridization signal; ?, chromosome not identified but in this group.

Supplemental Fig. 4. Dotplots of HSA21p BAC sequences.

Supplemental Fig. 5. Major alternative spliced isoforms of 21pGM6. Spliced exons are represented for spleen, thymus and HSA13 somatic cell hybrid. For each RNA source, the number of different alternatively spliced transcripts detected is given, and the group representing the highest percentage (determined as a fraction of the total number of clones sequenced) is shown. For HSA13 all variants show 2 nt substitutions (compared to the original gene model) represented by two white asterisks.

Supplemental Table 1. Results of CHORI-507 screen for HSA clones.

Supplemental Table 2. Results of CHORI-507 screen for 21p clones. Probes are STS probes derived from 21p YAC end sequences [Wang et al. 1999]. BACs were binned according to probe. BACs selected for sequencing are marked in bold.

Supplemental Table 3. Sequenced 21p BAC clones. Non-redundant length refers to contigs constructed from overlap of clones 71C21/54M18 and 201H5/216K13. *This clone is a human/mouse chimera, sequence length is for the human portion only.

Supplemental Table 4. BAC repeat content. Data for 21p BACs was obtained using RepeatMasker; data for the euchromatic proportion of the genome is from Lander et al.

(2001).

Supplemental Table 5. 21p bases involved in segmental duplication and pairwise alignment. The table summarizes the sequence properties of 21p segmental duplications (>90% identity >1kb) compared to the human genome (hg18), including the fraction of duplicated sequence (non-redundant duplication), the number of pairwise alignments, the average length of the pairwise alignments (kb) and their sequence identity. *21p duplication with random chromosomes.

Supplemental Table 6. Expression of 21p gene models. Method: e, based on 100% EST matches; p, *ab initio* prediction only. [a], data from two alternatively spliced transcripts. [b], number of PCR fragments sequenced: gene models with a sequenced fragment, but not marked as positive, indicates that RT-PCR expression was positive in at least one tissue but the fragment did not match 21p exactly.

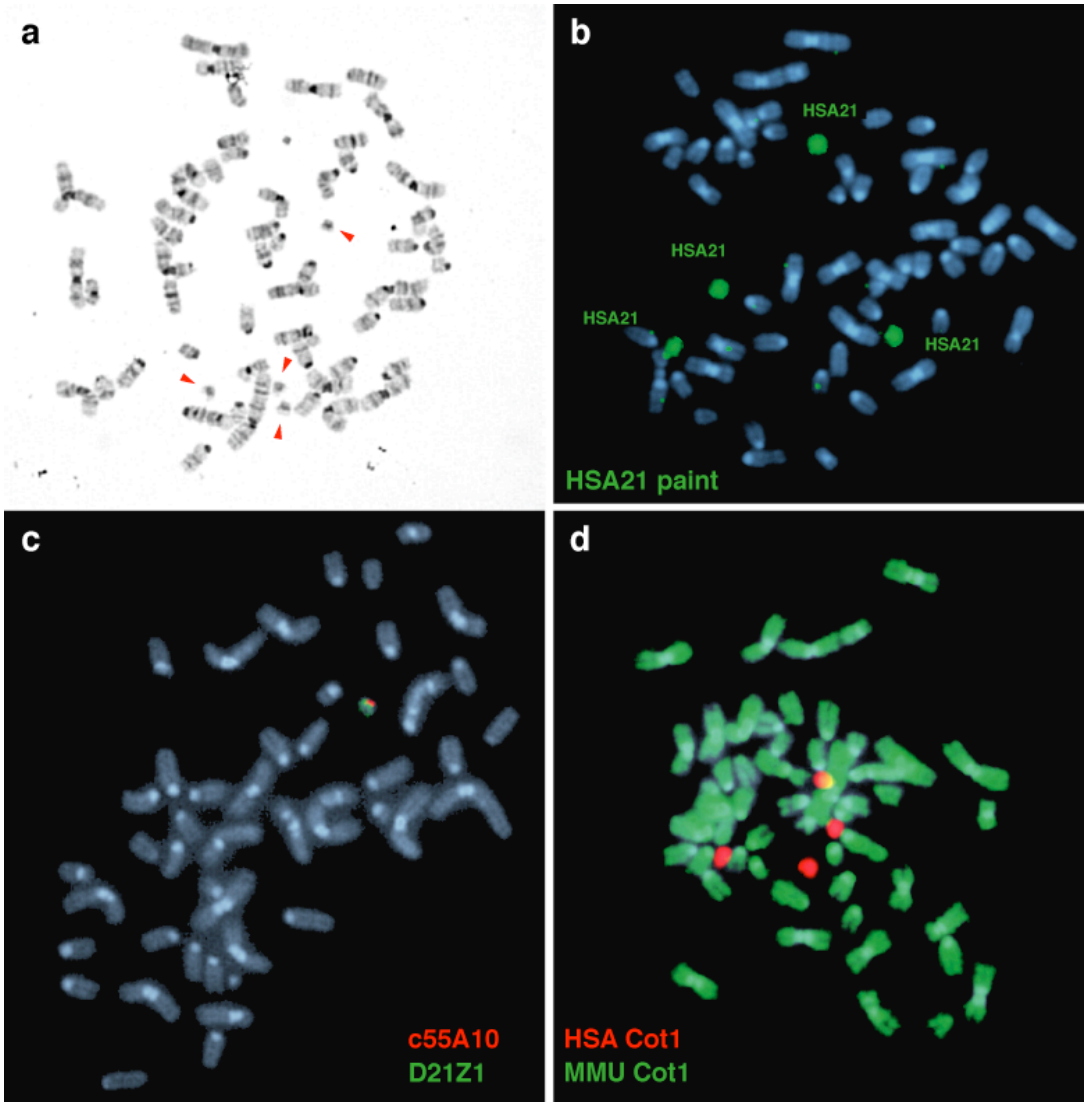
Supplemental Table 7. Homology and conservation of 21p gene models. ORFs of gene models were translated and searched against the databases CCDS Peptides and Peptides using ENSEMBL BLAST. In the case of gene models with no obvious single ORF (i.e. multiple short ORFs), all three forward frames were searched. For human homology, the best match gene is given. For conservation, results are presented as: name of gene with best match, length of match, % ID, e-value. PTR, Pan troglodytes; MMU, Mus musculus; GGA, Gallus gallus; DRE, Danio rerio; DME, Drosophila melanogaster; CEL, Caenorhabditis elegans. §, bp.

Supplemental Table 8. Copy Number results obtained by qRT-PCR. The numbers shown are the copy number of a particular assay for a gene model in each of 13 individuals. The IDs for the 13 individuals used in the analysis are given. [1]Different assays for the same gene model (see text).

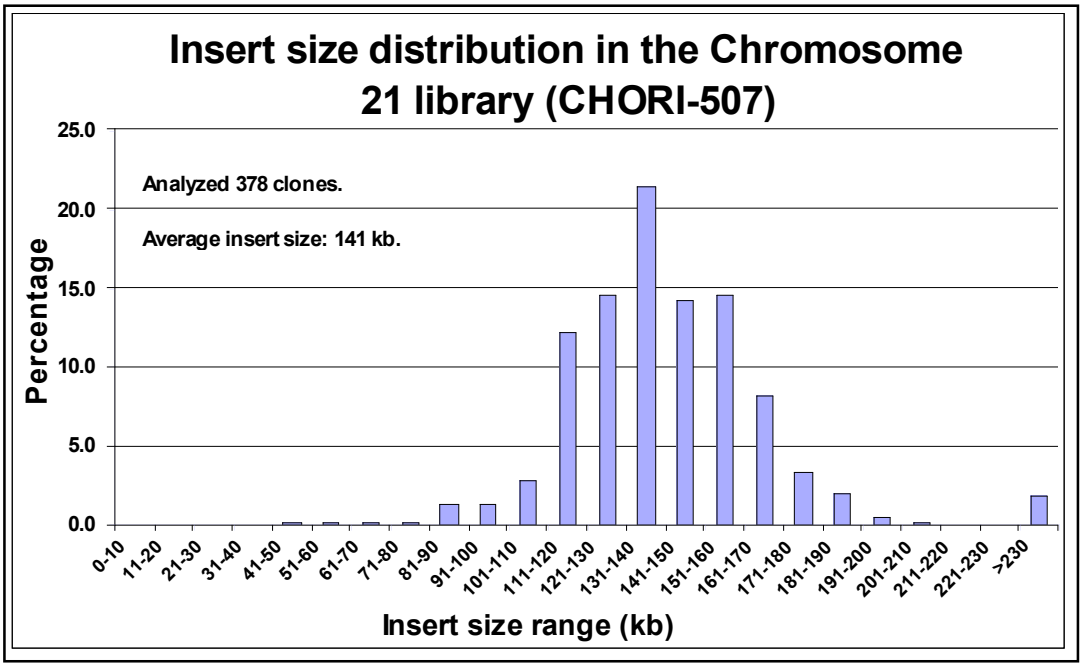
Supplemental Table 9. Polymorphism results. (**a**) The total number of sequence performed per gene model. (**b**) Number of sequence relative to different cDNAs derived from different from somatic cell hybrid (HSA) carrying acrocentric chromosomes and tissues. (**c**) Percentage of sequences that are identical to the original gene model predicted on the HAS21p. (**d**) The total number of groups obtained after the alignment and clustering of identical sequences. (**e**) Number of groups containing sequences with nucleotide and splicing variants. In parenthesis the number of sequences related to this group. (**f**) Number of groups containing splicing isoforms. (**g**) Percentage of groups containing protein truncation codon (PTC).

Supplemental Table 10. Oligonucleotides for expression, copy number and variation analyses.

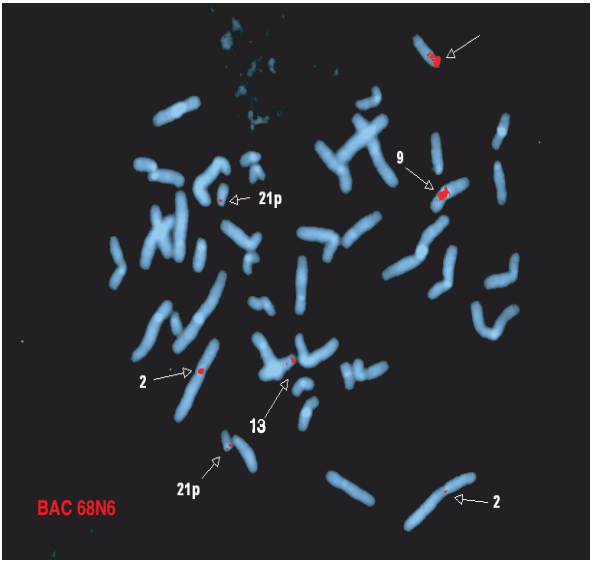
Supplemental Figure 1



Supplemental Figure 2



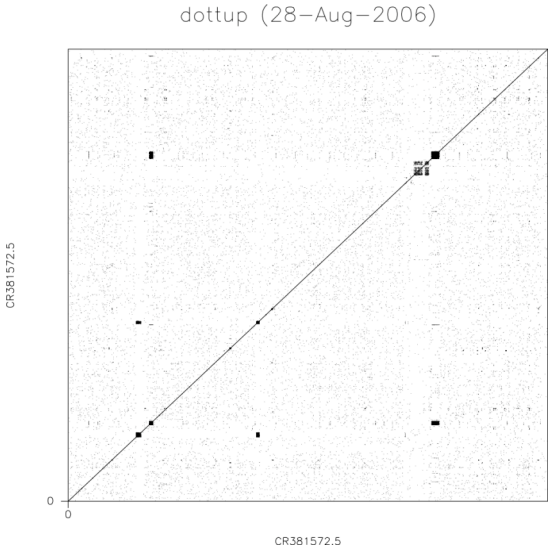
Supplemental Figure 3



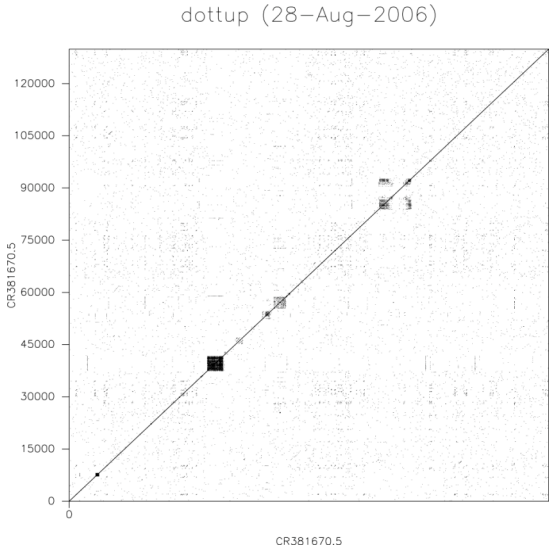
Chr	Group	BAC clone						
		68N6	8909	71C21	54M18	264C1	120F14	201H5
1								
2	A	+						
3			+			+		
4	B		+		+			+
5								
6								
7								
8								
9	C	+			?	?		+
10								
11								
12								
13		**+**	**+**	**+**	**+**	**+**	**+**	**+**
14	D		**+**		**+**	**+**		**+**
15				**+**	**+**	**+**		
16								
17	E							
18								
19	F							
20								
21	G	**+**	**+**	**+**	**+**	**+**	**+**	**+**
22				**+**		**+**		
X								
Y								

Supplemental Figure 4

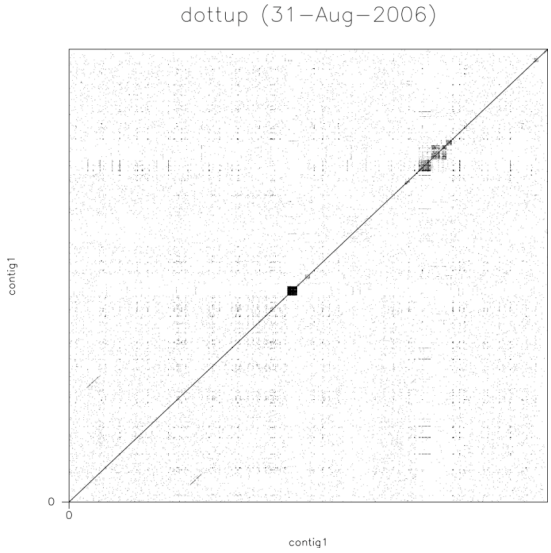
CHORI507-68N6 CR381572



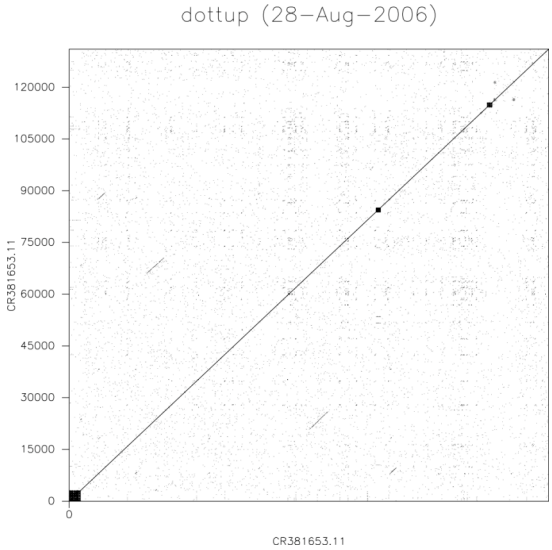
CHORI507-89O9 CR381670



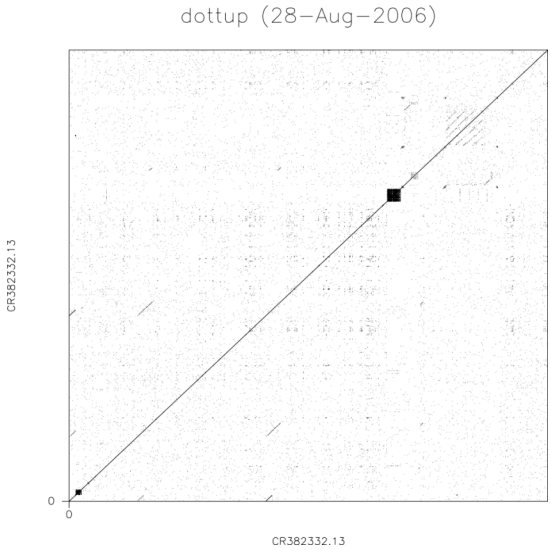
CHORI507-71C21 CR392039
CHORI507-54M18 CR381535



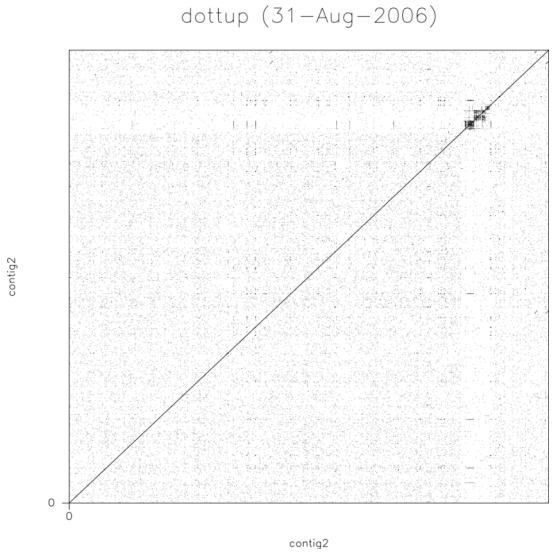
CHORI507-264C1 CR381653



CHORI507-120F14 CR382332



CHORI507-201H5 CR382285
CHORI507-216K13 CR382287



Supplemental Figure 5

Source	No. groups with alternative splicing	% of the total splicing variants	Transcripts
21pGM6	-	-	1 2 3 4 5 6 7 8
Spleen	4	37	1 4 5 6 8
		39	1 5 6 8
Thymus	5	63	1 3 4 5 6 8
HSA13 somatic cell hybrid	5*	80*	1 3 5 6 8

Supplemental Table 1

	Positives		Percentage	
Filter	HSA genomic	Alu/alpha	HSA genomic	Alu/alpha
1	316	194	1.71%	1.05%
2	215	232	1.17%	1.26%
3	241	190	1.31%	1.03%
4	254	128	1.38%	0.69%
5	179	155	0.97%	0.84%
6	278	215	1.51%	1.17%
7	212	219	1.15%	1.19%
8	247	201	1.34%	1.09%
9	189	207	1.03%	1.12%
10	219	262	1.19%	1.42%
11	210	200	1.14%	1.09%
12	192	197	1.04%	1.07%
Total	2752	2400		
Combined unique	3208			
Average	229	200	1.24%	1.09%

Supplemental Table 2

	2E4R	**4E9L**	**2E4L**	**2C2L**	**4E9R**	**1B8L**	**2C9R**
Positive BACs	26P1	45H15	1M11	45H15	47O22	**120F14**	115J4
	308O4	20F14	47O22	31O20	**54M18**	190J10	106P6
	68N6	9I18	**54M18**	20F14	**120F14**	**264C1**	129N2
	123D1	207C5	120F14	9I18	190J10		173P2
	153A9	309H6	92J24	97E12	**264C1**		193E19
	144M20	281N23	153M15	83C3	504L7		193E20
	266I21	378N19	**264C1**	**71C21**			**201H5**
	222M23	382G21	269D3	94F10			280D19
	470N15	426P9	285C4	175P13			303M9
	512F17	465D15	382F17	228O24			546C3
	563P24	468A5	462H17	309H6			555O7
	513E4	441J5		281N23			**216K13**
		517M20		378N19			308O4
		503E24		382G21			26P21
		505K17		426P9			123D1
		546A15		465D15			
		218I16		468A5			
		468G13		441J5			
		89O9		471M5			
				473M4			
				497J20			
				516A14			
				517M20			
				503E24			
				505K17			
				546A15			
				546B17			
				218I16			
				468G13			
				89O9			
BACs selected for sequencing	**68N6**						
		89O9		**89O9**			
				71C21			
			54M18		**54M18**		
			264C1		**264C1**	**264C1**	
			120F14		**120F14**	**120F14**	
							201H5
							216K13

Supplemental Table 3

BAC	Acc number	Length	Status	Non-redundant length	% g/c
CHORI507-68N6	CR381572	184355	finished	184355	37.3
CHORI507-89O9	CR381670	129889	finished	129889	41.3
CHORI507-71C21	CR392039	158069	finished	209483	43.3
CHORI507-54M18	CR381535	152296	finished		
CHORI507-264C1	CR381653	131056	finished	131056	39.3
CHORI507-120F14	CR382332	166871	2 unordered pieces	166871	40.8
CHORI507-201H5	CR382285	178865	finished	281920	36.1
CHORI507-216K13*	CR382287	150002	finished		
	Total	**1251403**		**1103574**	**39.7**

Supplemental Table 4

	21p BACs			Euchromatic genome			
	Number of elements	**length occupied**	**% sequence**	**Number of elements**	**length occupied**	**% sequence**	**χ2 p value**
SINEs	371	93798	8.5	1558000	359600000	13.1	0.00000
LINEs	309	252694	22.9	868000	558800000	20.4	0.10040
LTR elements	253	156953	14.2	443000	227000000	8.3	0.00000
DNA elements	80	29906	2.7	294000	77600000	2.8	0.83780
total length		1103312			2736681887		
bases masked		608555	55.2		1532256900	56.0	

Supplemental Table 5

	Non-redundant duplication (Mb)*			Number of pairwise alignment			Average length of pairwise alignment (kb)			Total aligned base pairs (Mb)			Average percent identity of pairwise alignment (%)		
	Inter	Intra	Both	Inter	Intra	Both	Inter	Intra	Both	Inter	Intra	Both	Inter	Intra	Both
21p	1.042	0.703	1.05	249	29	278	26.8	56.77	29.93	6.673	1.646	8.32	96.41	98.05	96.73
Random*	0.581	0.703	0.892	24	25	49	41.01	64.26	52.87	0.984	1.607	2.591	97.38	98.16	97.86

Supplemental Table 6

BAC clones	Models	Method	No. exons	Length, bp	ORF, aa	Brain	Heart	Kidney	Spleen	Liver	Colon	Sm. Intestine	Muscle	Lung	Stomach	Testis	Placenta	Skin	PBLs	Bone Marrow	Fetal Brain	Fetal Liver	Fetal Kidney	Fetal Heart	Fetal Lung	Thymus	Pancreas	Mammary Gl.	Prostate	No. +ve tissues	Sequenced[b]
68N6	21pGM1	p	6	615	204																										1
	21pGM2	p	4	2532	843																										
	21pGM3	p	2	120	28																										1
	21pGM4	e	2	461	54																										1
	21pGM5	e	2	374	46																										
89O9	21pGM6	e	8	343	303		+	+	+	+		+	+	+	+	+	+				+	+	+		+	+	+		+	17	2
	21pGM7	p	5	778	164											+														1	1
	21pGM8	p	7	1209	402																										
71C21/54M18	21pGM9	e	5	782	129		+						+		+	+		+	+				+	+			+			9[a]	3
	21pGM10	e	2	196	35																										
	21pGM11	e	3	515	70				+																					1	1
	21pGM12	p	3	266	88																										1
	21pGM28	e	2	237	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	24	1
264C1	21pGM13	e	2	242	35			+	+	+					+	+														5	1
	21pGM14	p	5	686	214																										1
	21pGM15	e	2	214	60	+	+				+		+		+	+	+		+											8	1
	21pGM16	p	9	1923	640				+				+			+	+				+						+		+	7	1
201H5/216K13	21pGM17	e	3	317	44																										
	21pGM18	e	2	535	29																										1
	21pGM19	e	2	329	71																										
	21pGM20	e	2	301	22																										1
	21pGM21	p	8	682	221																										6
120F14	21pGM24	e	3	473	43	+	+		+				+			+		+				+	+					+		9	2
	21pGM25	e	2	710	60	+								+						+	+		+	+			+			7	1
	21pGM26	e	2	382	34																										
	21pGM27	p	6	888	295																										

Supplemental Table 7

	HSA homology			Conservation						
Models	**Homology**	**% ID**	**e-value**	**PTR**	**MMU**	**GGA**	**DRE**	**DME**	**CEL**	**Note**
21pGM1	-									
21pGM2	-									
21pGM3	-									
21pGM4	-									
21pGM5	-									
21pGM6	ankyrin-containing protein (*ANKRD20A*)	71	8.00E-60	ANKRD26, 135, 27, 5e-09	-	-	-	-	-	
21pGM7	-									
21pGM8	sorting nexin associated golgi protein 1 (*SNAG1*)	44	5.40E-16	SNAG1, 140, 35, 2.2e-10	SNAG1, 112, 44, 3.3e-16	SNAG1, 111, 43, 8.7e-16	XP_684661 (similar to SNAG1), 175, 30, 2.3e-13	CG40368, 246, 23, 5.5e-06	-	SNAG1 related
21pGM9	tektins (*TEKT4*)	96	1.40E-63	TEKT1, 110,44, 3.3e-20	Tekt4, 111, 68, 8.1e-42	TEKT4, 117, 58, 2.2e-39	ENSDARG00000028899 (similar to TEKT4), 129, 40, 6.4e-23	CG32819 (Tekt3-related), 108, 32, 7.2e-14	-	Tektin family
21pGM10	-									
21pGM11	asparagine synthetase (*ASNS*)	87	2.70E-36	ASNS, 70, 87, 2.2e-36	Asns, 70, 81, 1.5e-34	ASNS, 70, 73, 5.0e-30	asns, 71, 70, 3.3e-30	asparagine-synthetase, 70, 36, 4.0e-07	MO2D8.4 (asparagine-synthetase), 55, 44, 3.6e-08	asparagine-synthetase
21pGM12	-									
21pGM28	SSU_rRNA_5	99	5.10E-46	SSU_rRNA_5, 107, 2.5e-42	SSU_rRNA_5, 110, 89, 1.9e-39	SSU_rRNA_5, 102, 84, 9.5e-36	SSU_rRNA_5, 92, 93, 4.3e-29		-	SSU_rRNA_5
21pGM13	-									
21pGM14	-									
21pGM15	-									
21pGM16	sorting nexin associated golgi protein 1 (*SNAG1*)	40	3.00E-14	ENSPTRG00000005666, 101, 73, 5.3e-34	Snag1, 89, 54, 5.3e-20	SNAG1, 146, 39, 6.8e-19	SNAG1, 78, 47, 1.1e-15	-	-	
21pGM17	similar to TBC1 domain family member 3C isoform 23	68	1.40E-11	ENSPTRG00000009501 (USP6NL family, RabGAP_TBC domain), 36, 64,	Usp6nl (RabGAP_TBC domain), 33, 45, 6.1e-5	ENSGALG00000006897 (RabGAP_TBC domain), 33, 52, 7.2e-07	TBC-domain (ENSDARP00000068022), 33, 48, 8.2e-07	RN-tre (RabGAP_TBC domain), 36, 39, 1.7e-06	TBC-domain (Y46G5A.1), 33, 48, 2.4e-06	Similarity to TBC domain Pfam PF00566
21pGM18	-									
21pGM19	-									
21pGM20	-									
21pGM21	60S ribosomal protein L7a (*SNORD36A*)	77	1.10E-24	ENSPTRG00000024092 (RPL7A family), 57, 77, 2.4e-24	Rpl7a, 57, 77, 2.2e-24	ENSGALG00000013415 (Rpl7a family), 57, 74, 3.0e-24	rpl7a, 57, 77, 1.5e-24	rpl7a, 84, 46, 1.4e-17	Y24D9A.4 (60S RIBOSOMAL L7A), 165, 35, 2.1e-22	60S RIBOSOMAL L7A
21pGM24	ENSG00000206169	90	2.10E-28	ENSPTRG00000020863, 58, 64, 2.2e-18	-	-	-	-	-	ANKYRIN REPEAT DOMAIN CONTAINING
21pGM25	sorting nexin associated golgi protein 1 (*SNAG1*)	31	6.10E-08	SNAG1, 147, 44, 6.4e-25	Snag1, 148, 51, 7.9e-35	SNAG1, 151, 46, 1.4e-3	ENSDARG00000052829 (Snag1-like), 43, 148, 6.9e-28	-	Y37A1B.2 (Lateral Signaling Target family member lst-4), 95, 26, 8.6e-07	SNAG1 related
21pGM26	-									
21pGM27	-									

Supplemental Table 8

			Individual												
		Assay[1]	**0-154**	**0-397**	**GM07006**	**GM11993**	**GM12003**	**GM12045**	**GM12139**	**GM12156**	**GM12750**	**GM12872**	**GM6991**	**GM7002**	**GP1973**
21p gene model	21pGM15	b	16	16	15	17	16	17	18	17	19	16	15	15	18
		c	9	10	9	10	10	10	10	10	8	9	10	10	11
		f	6	5	5	7	6	5	6	7	6	7	4	7	5
	21pGM25	d	7	8	7	7	8	7	8	8	8	7	7	9	9
		c	9	10	9	10	11	11	11	10	12	8	11	14	12
		e	23	24	22	20	21	24	23	22	23	19	22	23	27
		h	8	8	7	8	8	10	10	8	8	8	8	8	10
	21pGM28	f	7	15	20	17	14	29	14	30	28	14	29	37	18
		a	6	14	19	16	13	26	14	27	25	13	26	30	15
		h	10	23	34	25	24	45	24	54	45	24	42	54	24
		n	2	3	3	2	2	3	2	4	3	3	3	5	2
	21pGM6	a	11	12	11	13	12	12	11	11	12	11	12	12	11
		b	6	7	6	6	8	7	5	6	8	7	6	7	6
	21pGM9	i	4	4	5	4	3	4	3	3	4	4	4	3	4
		e	0	1	2	0	0	1	2	0	1	0	2	0	2
		f	1	1	1	1	1	0	1	0	0	0	1	2	0
Control genes	TBL2		1	1	1	1	1	1	1	1	1	1	1	1	1
	WBSCR14		1	1	1	1	1	1	1	1	1	1	1	1	1
	RFC2		1	1	1	1	1	1	1	1	1	1	1	1	1
	CYLN2		1	1	1	1	1	1	1	1	1	1	1	1	1
	WBSCR16		1	1	1	1	1	1	1	1	1	1	1	1	1
	AKAP1_1		1	1	1	1	1	1	1	1	1	1	1	1	1
	AKAP1_2		1	1	1	1	1	1	1	1	1	1	1	1	1
	HBG2_1		2	2	2	2	2	2	2	2	2	2	2	2	3
Normalization genes	SAMSN1		1	1	1	1	1	1	1	1	1	1	1	1	1
	NCAM2		1	1	1	1	1	1	1	1	1	1	1	1	0
	cgh4_A		1	1	1	1	1	1	1	1	1	1	1	1	1
	cgh6_A		1	1	1	1	1	1	1	1	1	1	1	1	1
	C21orf1		1	1	1	1	1	1	1	1	1	1	1	1	1
	ELN		1	1	1	1	1	1	1	1	1	1	1	1	1
	POR		1	1	1	1	1	1	1	1	1	1	1	1	1

Supplemental Table 9

	Gene model				
	21pGM6	**21pGM9**	**21pGM15**	**21pGM25**	**21pGM28**
number of sequenced clones[a]	121	48	185	69	31
somatic cell hybrid/tissues (n of sequences)b	HSA13 (25) spleen (57) thymus (39)	heart (35) testis (13)	HSA13 (32) HSA15 (24) HSA21 (23) HSA22 (28) heart (33) brain (45)	HSA21 (16) HSA22 (18) lung (35)	brain (31)
sequences identical to HSA21p[c]	0	0	38%	16%	84%
number of groups[d]	25	26	34	35	5
nucleotide variation groups[e]	19 (36%)	23 (85%)	29 (60%)	34 (84%)	4 (16%)
splicing isoform groups[f]	6 (64%)	3 (15%)	4 (2%)	0	0
groups with PTC[g]	25	3	15	1	-

Supplemental Table 10

	Gene/gene model	Assay	F	R
Expression	21pGM1	18481p4.1.a	TCCAAATGGCATCTCCTACC	CCTCCAGCTTTCAAAGTGCT
	21pGM2	18477p3.1.a	CACATGGGGACACAGTCAAG	GCCTCCTCGGGTTTATGC
	21pGM3	22703p7.5.a	GAAAGTGAGTAATGGCCCAGA	CATTTTATGCTAAGCTTCTTAGTTGC
	21pGM4	18463e6.16.a	TGGCTAGACAACCCCTTCTG	AGGCCCTTCAGGGATAACAC
	21pGM5	18455e5.3.a	GGGTGTCATCATGGACCTTT	GCGGCTTTTTGCCCA
	21pGM6	30204e3.1.d	GGAAGAGCAAGAAAGGTGTGA	TTCTTGGGAAATTGCTGAGG
	21pGM7	22705p7.14.c	GATCATGGCACAAGTTGCAG	GTTAATCGCAAAACTGTACGAGAA
	21pGM8	18485p4.3.a	AAACCGGGGTTCACATCAA	CTCTCCCCTCCCAGAAACTC
	21pGM9	18409e2.1.a	CGGCTGTAGTTGCTCTCACA	GACTGTTGGTCATGGCAGTG
	21pGM9	18419e2.6.a	ACCTGGGACTCCAGTGTGAC	GCTCATGTGGATGTCCTTGA
	21pGM10	18459e6.14.a	TCAGGCCACAAGTGAACATC	CACTGGGGGCTTGATTTTT
	21pGM11	30200e6.1.b	TGCTTTGGATTTCACCCATT	CAAATGCAAACACACCAACC
	21pGM12	18473p1.4.a	TTAATGCAGCTGGGCAACTTT	GCTGATCCAGTGCCCAGAA
	21pGM13	30202e6.12.a	TCAGACTTCAAATAACTGCCAGA	CAATTAAAATGTGTAAATAGCCCAGT
	21pGM14	22715p7.9.c	CAAGAACAGCCATGGAAACA	TTTTCTTTCCAGTCCCATGC
	21pGM15	18393e1.1.a	CTTGGGGCTTGGTTCTATGT	TTCATTTCTTGTTGCTCTTGGA
	21pGM16	18483p4.2.a	TGATCTTATGGCGGAGAAGG	CTGCTGCAAAGGCTTCCTTA
	21pGM17	18457e6.11.b	ATCTGGAGATTTTCTCAAGCTTCT	TTTGTTAATTGACATAATTCTTCAGCA
	21pGM18	18395e1.2.a	GCTTAAATTGGGTTGTGTGCT	CAATGCCTTCCTAATGACCTT
	21pGM19	18447e3.2.a	TTGCTGGGTCCATCTCTCTT	TGAATTTCAACCAGCAAAGAA
	21pGM20	18403e1.6.a	GAGGCAATACACTGGCATCA	GCTTCCCCTTTTTCGTCTCT
	21pGM21	22717p7.17.c	CCAAGGTGCGAAAAAGAAAG	CCATTGAGCAGCTGTTTGGT
	21pGM24	30917C21p-e6.4a	CAAGGTGAGGTCGAGAGTCA	TGGCTTACCCGTGGATTCTA
	21pGM25	18453e5.2.a	GCCATGGACCTGTTAGTGCT	GCACCAGCCCTTTGATACTG
	21pGM26	18439e2.19.a	GGCTTGGTTCTATGTCCCTGT	TCGGCATCAGAAAGATCAGA
	21pGM27	22725p6.15.a	TGCCCACCAGAAGATTTAGG	GCCTTTCCTTCCCTGATTTC
	21pGM28	18397e1.3.a	GCTCGCTCCTCTCCTACTTG	GGCGACTACCATCGAAAGTT
Copy number	21pGM6	Hs21pGM6.2-a	CCTTTACGTGGACCTTCTCATGA	TTTTCCATTGACTATTCTGTTTCCTTT
		Hs21pGM6.2-b	TGCTGCTGACAGATTAACCAAAG	CTGAGGATACGTTTTCCCAATCTT
	21pGM9	Hs21pGM9.1-i	CCTCCACCCCAGAGACCC	GCGTGGCACAGATTGTCCT
		Hs21pGM9.1-e	AGGAACCTCTGCACATAGCCC	GCCGGTGTGAGAGCAACTACA
		Hs21pGM9.1-f	CTGCGCAACCTCAAGGACAT	GGCAGTGACGTCCTTCTCCA
	21pGM15	Hs21pGM15.1-b	GGCTTGGTTCTATGTCCCTGC	TTCCTCTTCGCCCTTGCA
		Hs21pGM15.1-c	CGGACCAGTGTTCCCTAGTTG	CAATGTGCCCACGCGTC
		Hs21pGM15.1-f	TTCCCTAGTTGTGGGAGCAGA	AACATAGCTGCACTTCAGGCC
	21pGM25	Hs21pGM25.1-e	TCACCACTTCCATCCAATTCC	TGCTGCATCTGTGAGTCCATG
		Hs21pGM25.1-h	CCCAGGAATTGATAGTGTTGAATG	GAAACCGTCCACGATCGAGT
		Hs21pGM25.1-d	CACCCCCAGTGTGACACTCA	CCGTCAATCTTGCTCTCCCA
		Hs21pGM25.1-c	GAAGGACCTGGACCTAGCCAT	TGGTGTCCCTGAGACAGCAC
	21pGM28	Hs21pGM28.1-a	GCGCTGACCCCCTTCG	TGGTTTTGATCTGATAAATGCACG
		Hs21pGM28.1-f	GGGCGCTGACCCCCT	TGGTTTTGATCTGATAAATGCACG
		Hs21pGM28.1-h	GCTGACCCCCTTCGCG	CGGGTTGGTTTTGATCTGATAAAT
		Hs21pGM28.1-n	TAGATAACCTCGGGCCGATCG	AAAGTTGATAGGGCAGACGTTCGAA
	TBL2	15_TBL2	CTGTCTGTCATCTAGCCTCTGCA	CCAGGCCCTTCGGCA
	MLXIPL	14_WBSCR14	CACGCATTTCTGATTCCCATT	GCAGGCGGTCTGCGC
	RFC2	6_RFC2	TGACAACCCCCTTGCCAC	TGCCATATTTGGTAATGCGC
	CYLN2	5_CYLN2	CTGCTCTTCCTCCTGTGGCT	AGCTGTTTGCCCAGGTCCT
	WBSCR16	4_WBSCR16	CATGCTTTGTGGGAACTGCTT	AGAGGACAGCAGTGTGAATCCA
	AKAP1	HsAkap1_1	AAAGCACTTAGTCGGTCGGC	AAACTCACATAGCGCCCCTG
	AKAP1	HsAkap1_2	GATACGTGGACTACGGCGGA	GCCGGAGCACGTCTACTTTC
	HBG2	HsHBG2-1	CGGCTGGCTAGGGATGAAG	TGTGGAACTGCTGAAGGGTG
	SAMSN1	hsacgh-21-1A	TCTCGATTCCTTTAGCGTCAACT	GCTCAGCTCAGGATGCAGAAA
	NCAM2	hsacgh-21-3A	CCACCTCAAGTATACCGAAGTCCTA	GGAAGCGCTGCATGTGAA
	-	hsacgh-21-4A	CAATTCAGGTCAGGTGATAACTCAGTA	GCCAGGTTTAGAATGTTTGTCTAAGTC
	-	hsacgh-21-6A	CGTCACATCACATCTCACTCATTG	ACCACGTGAAAGGAGGTTTCC
	-	HsHUG1	GCCAGGAGCCCCACTGTAA	AGATTTCTGCAGCCACCTCAC
	ELN	9_ELN	GGCATGAGACGCTCCACAT	TTCTGAGCCAGGGCAACAC
	POR	0.1_400KboutWBcentro	CACCAAAGACTGGCGTTTCC	GCGGACCCAGATTTCTTGAC
Variation	21pGM6	21pGM6.2cDNA	CAGAGCCTTTACGTGGACCT	TGCGTATTTCCTGTCGGAGT
	21pGM9	21pGM9.1cDNA	GGCGGTCGTGTAGCTGAG	CTGTTTATTTGGGGGATTGG
	21pGM15	21pGM15.1cDNA	GGCTTGGGGCTTGGTTCT	CCAGAGATTTATTCATTTCTTGTTG
	21pGM25	21pGM25.1cDNA	ACAGGGCACTCAGCTCAACT	TTGAACTGGAAACCGTCCAC
	21pGM28	21pGM28.1cDNA	TGTGGTAATTCTAGAGCTAATACATGC	GTGGCGACTACCATCGAAAG