Genomics & Informatics Vol. 4(4) 141-146, December 2006

Chromosomal Information of 1,144 Korean BAC Clones

Mi-Hyun Park[1†], Hee-Jung Lee[1,3†], Kwang-Joong Kim[1], Jae-Pil Jeon[1], Hye-Ja Lee[1], Jun-Woo Kim[1], Hung-Tae Kim[1], Hyo Soung Cha[1], Cheol-Hwan Kim[2], Kang-Yell Choi[3], Chan Park[1], Bermseok Oh[1], Kuchan Kimm[1], Jong-Young Lee[1*] and Bok-Ghee Han[1*]

[1]Center for Genome Science, National Institute of Health, 5 Nokbun-dong, Eunpyung-gu, Seoul 122-701, Korea, [2]Dong-A SeeTech, Co.Ltd, Seoul Building, 81-12 Nonhyeon-2-dong, Gangnam-gu, Seoul 135-818, Korea, [3]Department of Biotechnology, Yonsei University, Seoul 120-752, Korea

Abstract

We sequenced 1,841 BAC clones by terminal sequencing, and 1,830 of these clones were characterized with regard to their human chromosomal location and gene content using Korean BAC library constructed at the Korean Science (KCGS). Sequence analyses of the 1,830 BAC clones was performed for chromosomal assignment: 1,144 clones were assigned to a single chromosome, 190 clones apparently assigned to more than one chromosome, and 496 clones to no chromosome. Evaluating gene content of the 1,144 BAC clones, we found that 706 clones represented 1,069 genes of which 415 genes existed in the BAC clones covering the full sequence of the gene, 180 genes covering a 50%~99%, and 474 genes covering less than 50% of the gene coverage. The estimated covering size of the KBAC clones was 73,379 kilobases (kb), in total corresponding to 2.3% of haploid human genome sequence. The identified BAC clones will be a public genomic resource for mapped clones for diagnostic and functional studies by Korean scientists and investigators worldwide.

Keywords: Korean, BAC clone, genome, cancer

Introduction

The International Human Genome Sequencing Consortium published the first draft of the human genome in *Nature* in February 2001 (Lander *et al.*, 2001), and later published the sequence of the entire genome (International Hunam genome Sequencing Consortium, 2004). These Reports suggested that the number of human genes may be ~34,000, significantly fewer than previous estimates that ranged from 50,000-140,000 (Roest Crollius *et al.*, 2000). Complete analyses of chromosomes 6, 7, 14, and 20-22 have already been published. In October 2003, complete data for chromosome 6 were published online in Nature, and complete data for chromosomes 7, 14, 20, 21 and 22 were published in July 2003 (Hillier *et al.*, 2003), February 2003 (Heilig *et al.*, 2003), December 2001 (Deloukas *et al.*, 2001), May 2000 (Hattori *et al.*, 2000) and December 1999 (Dunham *et al.*, 1999), respectively.

Bacterial artificial chromosomes (BACs) use F-factor-based vectors and are suitable for propagation of DNA segments up to 300 kb in *Escherichia* coli (Shizuya *et al.*, 1992). Their stability over hundreds of generations, capacity to hold genomic inserts of large size, and ease of manipulation have established BACs as invaluable tools for a variety of applications in human genetics, most importantly for the mapping and sequencing of the human genome (McPherson *et al.*, 2001; Lander *et al.*, 2001).As a result of the Human Genome Project, the majority of human genes are available in a characterized and sequenced BAC format. BACs have many advantages for the cloning of large-insert genomic DNAs of a species and have been used for a variety of genetic applications, including physical mapping of chromosomes and large-scale sequencing of genomes.

A BAC clone containing an entire specific gene locus was used to study gene regulation in a model tissue culture-based system. Also, a BAC that lacked a putative negatively regulating promoter sequence was constructed and used to transfect cell lines (Bochukova *et al.*, 2003). These experimental systems illustrate the potential of BAC clones in large-scale gene expression studies, new gene therapy strategies, and validation of potential molecular targets for drug discovery (Bochukova *et al.*, 2003). Several very efficient methods have been developed for manipulating BAC clones by retrofitting constructs with marker and selection genes (Mejia *et al.*, 1997; Kim *et al.*, 1998; Wang *et al.*, 2001). Both nonviral (Hart *et al.*, 1998) and infectious (Wade-Martins *et al.*, 2001) novel gene transfer approaches have been developed and used for genomic clone delivery. A BAC library has been constructed to provide reliable and efficient materials for constructing

†Both these anthors contributed equally to the work.
*Corresponding author: E-mail leejy63@nih.go.kr, Tel +82-2-380-2259, Fax +82-2-354-1063
bokghee@nih.go.kr, Tel +82-2-380-2258, Fax +82-2-354-1078
Accepted 13 Sep 2006

sequence-ready maps (Kim *et al.*, 1996). Also, BAC clones have been used to identify SNPs for specific genes or regions in the human genome and for physical mapping studies. Map-based cloning and genome projects have identified many individual genes, and these genes will be useful for other applications in biotechnology, diagnostics and gene therapy. In particular, identified genes can be used in transgenic-based approaches for further functional analysis. Therefore, 1,841 BAC clones were randomly selected from the Korean BAC library, and BAC end sequences were determined. We obtained the information on the chromosomal locations of specific genes should be useful for studies of gene function and genomic analysis.

Materials and Methods

Construction of BAC library

A total of 100,224 clones for the Korean BAC library were prepared from male donor DNA, and a 2.9-fold redundant KBAC library was constructed (Park *et al.*, 2006). 100,224 BAC clones were placed in both of 1,044 96-well plates and 261 384-well plates. And four superpooled KBAC libraries were constructed to quickly isolate desired BAC clones by colony PCR.

Plasmid DNA extraction of BAC clones

A single BAC colony was picked and inoculated into 2 ml 2xYT medium containing chloramphenicol (6.25 μg/ml) and incubated with shaking for 16 h at 37°C. BAC DNA was isolated by the standard alkaline lysis method (Sambrook *et al.* 1989). After incubation, the cells were harvested by centrifugation at 2,500 rpm in swing bucket Qiagen 09366Ffor 10 min, and plasmids were purified with a plasmid purification kit (Qiagen) following the manufacturer's instructions. The supercoiled BAC DNA was isolated and digested with the restriction enzymes *Hind*III or *Not*I to identify DNA inserts from the BAC plasmids.

BAC end sequencing and analysis

BAC ends were sequenced using BigDye Terminators (Perkin-Elmer Applied Biosystems) with an ABI3100 and ABI3730 automated sequencer. The BAC end sequences were determined using M13 oligonucleotide primers: forward (5'-CACGACGTTGTAAAACGAC-3') and reverse (5'-CGATAACAATTTCACACAGG-3'). Reaction mixturess contained 0.5μg of DNA, 2μl of BigDye terminator mix, 50of universal and reverse primers consisting of 13 oligonucleotides in a total volume of 20μl. This mix was denatured at 95°C for 5followed by 50 cycles of 95°C for 45sec, 55°C for 30sec, and 60°C for 4min. Excess BigDye terminators were removed using the Montage SEQ Cleanup kit (Millipore).

As a part of sequence analysis, computational KBAC clone system and a few database systems on analysis information of Korean BAC clones have been constructed to provide information of sequences homology and genes in the BAC clones and their lengths and chromosomal locations, and to manage BLAST search processes and results. This system was consists of process manager and map viewer.

Submitted forward and reverse sequences (the FASTA format) of BAC clones on this system analyzed using BLAST with an E-value better than $1e^{-10}$ to human genomes. Human genomes data was used NCBI data (ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/, build number: 35, version: 1). The BLAST results were parsed by process manager, and the parsed data-set systematically stored in database system. Finally, process manager perform a database search to find the involved complete genes in the KBAC clones.

Results and Discussion

A total of 1,841 Korean BAC(KBAC) clones were randomly selected from a KBAC library containing 100,224 clones and were sequenced by BAC-end sequencing, which was performed and preserved by the KCGS. Of these, 11 clones have not been sequenced at the first step, but 1,830 clones were characterized and localized to specific chromosomes by BLAST search. BLAST analysis localized 1,144 BAC clones to chromosomes totaling 73.4 Mb, which corresponds to about 2.3% of the human haploid genome (Table 1). A

Table 1. Chromosomal localization and estimated length of BAC clones deduced by BAC-end sequence analysis in this study

Chrmosome	Chromosome size (kb)	Number of BAC clones	Covering size of genome (kb)	Rate
Chr.1	263,000	87	5,707	2.2%
Chr.2	255,000	102	6,759	2.7%
Chr.3	214,000	80	5,194	2.4%
Chr.4	203,000	97	5,641	2.8%
Chr.5	194,000	77	5,188	2.7%
Chr.6	183,000	59	3,976	2.2%
Chr.7	171,000	64	4,051	2.4%
Chr.8	155,000	53	3,343	2.2%
Chr.9	145,000	46	3,087	2.1%
Chr.10	144,000	56	3,605	2.5%
Chr.11	144,000	56	3,503	2.4%
Chr.12	143,000	66	4,412	3.1%
Chr.13	98,000	47	2,671	2.7%
Chr.14	93,000	31	1,862	2.0%
Chr.15	89,000	40	2,589	2.9%
Chr.16	98,000	43	2,869	2.9%
Chr.17	92,000	28	1,688	1.8%
Chr.18	85,000	31	1,706	2.0%
Chr.19	67,000	12	890	1.3%
Chr.20	59,000	23	1,433	2.4%
Chr.21	33,000	9	501	1.5%
Chr.22	34,000	12	993	2.9%
Chr.X	164,000	24	11,599	1.0%
Chr.Y	35,000	1	112	0.3%
Total	3,161,000	1,144	73,379	2.3%

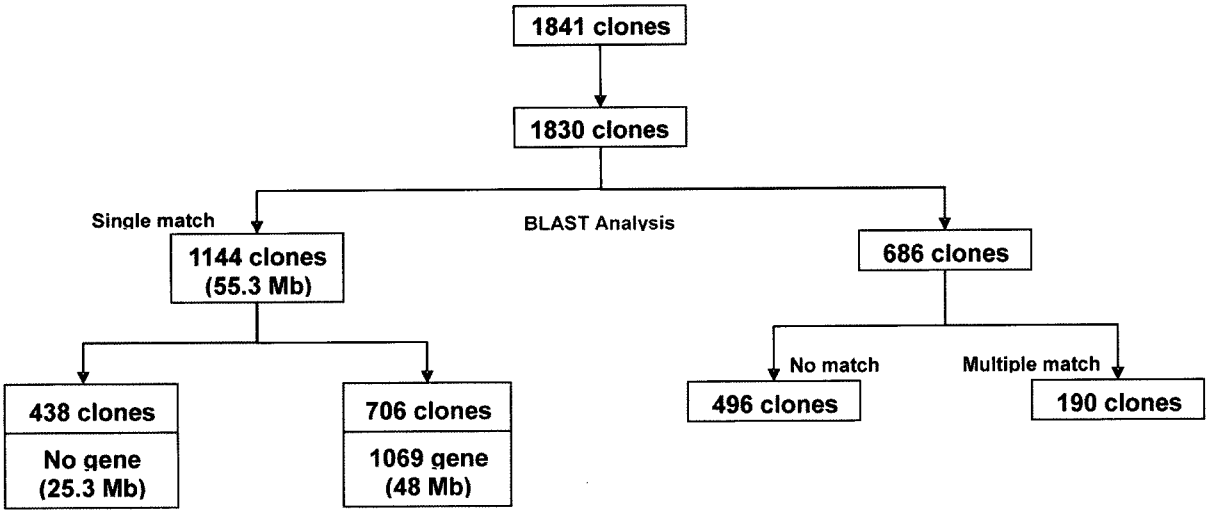

Fig. 1. Summary of 1,841 BAC clones after sequencing and BLAST searching analysis.

Table 2. Number of genes and rate of coverages included in the BAC clones.

Ranges	Genes	Percentage
100%	415	39%
90~99%	32	3%
80~89%	34	3%
70~79%	25	2%
60~69%	42	4%
50~59%	47	4%
40~49%	58	5%
30~39%	74	7%
20~29%	116	11%
10~19%	110	10%
0~9%	116	11%
Total	1,069	100%

total of 686 BAC clones were not located by BLAST, including 190 BAC clones were too much matched on chromosomes and 496 clones that did not localize to any chromosome. Of the 1,144 BAC clones, 438 clones did not encode a gene; 706 clones, however, were found to encode partial or full sequences of 1,069 genes that were eliminated duplicated genes in the genome (Fig. 1). The 706 clones included 415 genes covering 100% of the whole gene sequence, 180 genes covering >50%, and 474 genes covering <50% (Table 2). The 415 fully covered genes may be used directly or indirectly to study gene function. The KBAC sequence information also will be useful for the unambiguous identification of genomic regions, for obtaining information on full-length genes, and for

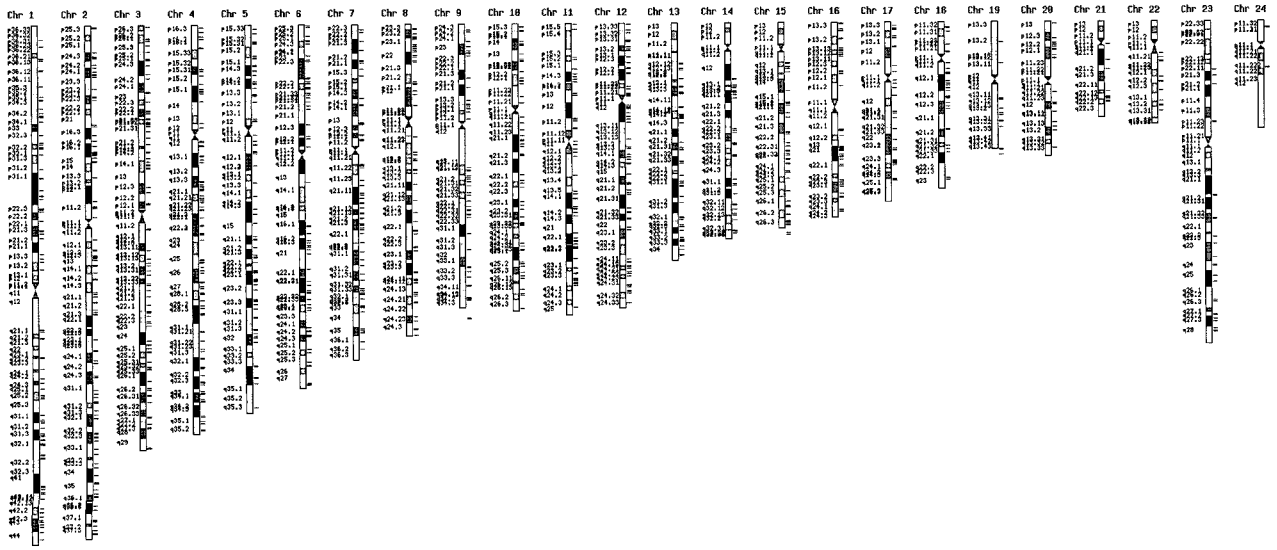

Fig. 2. Distribution of KBAC clones among the human chromosomes using sequence based maps.

Table 3. Information on 45 candidate cancer-related genes included in the BAC clones.

Gene	Start position of gene	End posiotn of gene	Length	Gene-matching start position in clone	Gene-matching end position in clone	Gene-matching Clone lengths	Coverages
AAA1	34163274	34684434	521,160	34163274	34214250	50,976	9.8%
ALK	29327293	30056083	728,790	29327293	29340595	13,302	1.8%
ANGPT2	6347601	6408172	60,571	6387727	6408172	20,445	33.8%
ARAF1	47176761	47187553	10,792	47176761	47187553	10,792	100.0%
ARVCF	18331973	18378863	46,890	18338147	18378863	40,716	86.8%
AXIN1	277441	342465	65,024	314536	342465	27,929	43.0%
AXL	46416663	46459511	42,848	46416663	46459511	42,848	100.0%
BCAR1	73820430	73843004	22,574	73832439	73843004	10,565	46.8%
BID	16591460	16631812	40,352	16591460	16598812	7,352	18.2%
CDH1	67328696	67426945	98,249	67400598	67418037	25425	25.9%
CDH13	81218079	82387702	1,169,623	81919466	82006929	87,463	7.5%
CXADR	17807201	17861147	53,946	17839262	17861147	21,885	40.6%
CXCL10	77299452	77301829	2,377	77299452	77301829	2,377	100.0%
CXCL9	77279802	77285820	6,018	77279802	77285820	6,018	100.0%
DPYD	97255323	98098600	843,277	97481400	97516484	35,084	4.2%
ESR1	152220800	152516520	295,720	152220800	152228623	7,823	2.6%
EWSR1	27988825	28021059	32,234	27988825	28021059	32,234	100.0%
FER	108111422	108551272	439,850	108434010	108538626	104,616	23.8%
HFE	26195427	26205038	9,611	26195427	26205038	9,611	100.0%
HSD17B4	118816122	118905923	89,801	118853382	118905923	52,541	58.5%
IL12A	161189331	161196508	7,177	161194981	161196508	1,527	21.3%
INPP4B	143307499	143710137	402,638	143307499	143340712	72011	17.9%
KIAA0153	41887126	41907573	20,447	41889038	41907573	18,535	90.6%
LASP1	34279894	34331541	51,647	34279894	34284870	4,976	9.6%
LRP6	12164958	12311013	146,055	12164958	12203979	39,021	26.7%
MAPK8	49279693	49313189	33,496	49297059	49313189	16,130	48.2%
MBD2	49934573	50005156	70,583	49946119	50005156	59,037	83.6%
MMP12	102238686	102250889	12,203	102243133	102250889	7,756	63.6%
MMP7	101896449	101906688	10,239	101896449	101897796	1,347	13.2%
MSH3	79986295	80208390	222,095	80115752	80152945	37,193	16.7%
MTR	233284759	233390003	105,244	233362914	233390003	27,089	25.7%
MYH11	15704495	15858369	153,874	15704495	15771968	67,473	43.8%
NRG1	32525295	32741615	216,320	32549287	32633411	84,124	38.9%
PGR	100414313	100506465	92,152	100460598	100506465	45,867	49.8%
PLA2G6	36832003	36902263	70,260	36863430	36897254	33,824	48.1%
RERG	15151985	15265571	113,586	15202055	15262047	59,992	52.8%
SLC2A4	7125833	7131125	5,292	7125833	7131125	5,292	100.0%
TERC	170965100	170965550	450	170965100	170965550	450	100.0%
TGFB3	75494195	75517242	23,047	75494195	75516125	21,930	95.2%
TGM4	44891131	44931097	39,966	44891131	44931097	39,966	100.0%
THBS1	37660572	37676960	16,388	37660572	37676960	16,388	100.0%
THRB	24139236	24511317	372,081	24486015	24511317	25,302	6.8%
TIMP1	47199061	47202441	3,380	47199061	47202441	3,380	100.0%
TRIM29	119487205	119514073	26,868	119487205	119514073	26,868	100.0%
WRN	31010320	31150819	140,499	31010320	31036531	26,211	18.7%

producing clones for functional studies. The 1,069 genes were evenly distributed among chromosomes using sequence-based maps (Fig. 2).

To evaluate the importance of individual BAC clones identified in this study, we performed sequence comparisons using the CancerGenetics and Cancer500 databases, which contain 1,584 candidate cancer genes. Of these cancer genes, 45 genes were partially included in our BAC clones while12 genes, 2,970 bp, AXL(42,848 bp), CXCL9 (6,018 bp), CXCL10(2,377 bp), EWSR1(32,234 bp), HFE(9,611 bp), SLC2A4(5,292 bp), TERC(450 bp), TGM4(39,966 bp), THBS1(16,388 bp), TIMP1(3,380 bp), and TRIM29(26,868 bp) were completely included in the KBAC clones (Table 3). Those clones could provide

genome sequences to isolate the whole gene and other applications (e.g., promoter analysis).

Our BAC library of Korean is constructed for several purposes, for example to compare nucleotide diversity among ethnic groups. This KBAC library will be a useful reference, especially for comparing the levels of nucleotide diversity in genome-wide regions. In addition, the KBAC clones were also intended to be used for BAC-based comparative genomic hybridization (CGH) analyses. Comparative genomic studies with vertebrate genomes will lead to a better understanding of the molecular mechanisms responsible for vertebrate novelties and provide insight into the origins of modern species. Our KBAC library will benefit both biomedical and biological communitiesfor these genomic applications. Recently, BAC-based CGH showed a significant impact in the field of cancer cytogenetics as a powerful tool for the detection of aberrations in chromosome copy numbers, even in epithelial solid tumors, in which tumor-specific genomic alterations are difficult to detect using conventional cytogenetics (Veltman *et al.*, 2002; Albertson and Pinkel, 2003; Mantripragada *et al.*, 2004). Finally, BAC end sequencing was introduced as a tool for virtual library screening to search for BACs overlapping with large sequence contigs and to establish a baseline from which to expand contiguous sequences. KBAC sequence information also will be useful for the unambiguous identification of genomic regions, for obtaining information on full-length genes, and for producing clones for functional studies.

The immediate goal of this study was to obtain sequences of the KBAC clones. These sequences will be useful for a variety of genetic applications, including physical mapping of chromosomes and sequencing of large parts of the genome. The mapped BAC clones will serve as a valuable resource for reference sequences in highly variable regions, for promoter analysis and for comparative genomics studies.

References

Albertson, D.G. and Pinkel, D. (2003). Genomic microarrays in human genetic diseases and cancer. *Hum Mol. Genet.* 12, 145-152.

Bochukova, E.G., Jefferson, A., Francis, M.J., and Monaco, A.P. (2003). Genomic studies of gene expression: regulation of the Wilson disease gene. *Genomics* 81, 531-542.

Deloukas, P., Matthews, L.H., Ashurst, J., Burton, J., Gilbert, J.G., Jones, M., Stavrides, G., Almeida, J.P., Babbage, A.K., Bagguley, C.L. *et al.* (2001). The DNA sequence and comparative analysis of human chromosome 20. *Nature* 414, 865-871.

Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J. *et al.* (1999). The DNA sequence of human chromosome 22, *Nature* 402, 489-495.

Hart, S.L., Arancibia-Carcamo, C.V., Wolfert, M.A., Mailhos, C., O'Reilly, N.J., Ali, R.R., Coutelle, C., George, A.J., Harbottle, R.P., Knight, A.M. *et al.* (1998). Lipid-mediated enhancement of transfection by a nonviral integrin-targeting vector. *Hum. Gene Ther.* 9, 575-585.

Hattori, M. (2000). Human genome sequencing. *Tanpakushitsu Kakusan Koso* 45, 1978-1985.

Heilig, R., Eckenberg, R., Petit, J.L., Fonknechten, N., Da Silva, C., Cattolico, L., Levy, M., Barbe, V., de Berardinis, V., Ureta-Vidal, A. *et al.* (2003). The DNA sequence and analysis of human chromosome 14. *Nature* 421, 601-601.

Hillier, L.W., Fulton, R.S., Fulton, L.A., Graves, T.A., Pepin, KH, Wagner-McPherson C, Layman, D., Maas, J., Jaeger, S., Walker, R. *et al.* (2003). The DNA sequence of human chromosome 7. *Nature* 424, 157-164.

International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature* 431, 931-945.

Kim, S.Y., Horrigan, S.K., Altenhofen, J.L., Arbieva, Z.H., Hoffman, R., Westbrook, C.A.(1998). Modification of bacterial artificial chromosome clones using Cre recombinase: introduction of selectable markers for expression in eukaryotic cells. *Genome Res* 8, 404-412.

Kim, U.J., Birren, B.W., Slepak, T., Mancino, V., Boysen, C., Kang, H.L., Simon, M.I., Shizuya, H.(1996). Construction and characterization of a human bacterial artificial chromosome library. *Genomics* 34, 213-218.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860-921.

Mantripragada, K.K., Buckley, P.G., de Stahl, T.D., Dumanski, J.P. (2004). Genomic microarrays in the spotlight. *Trends Genet* 20, 87-94.

McPherson, J.D., Marra, M., Hillier, L., Waterston, R.H., Chinwalla, A., Wallis, J., Sekhon, M., Wylie, K., Mardis, E.R., Wilson, R.K. *et al.*(2001). A physical map of the human genome. *Nature* 409, 934-941.

Mejia, J.E., and Monaco, A.P. (1997). "Retrofitting vectors for Escherichia coli-based artificial chromosomes (PACs and BACs) with markers for transfection studies", *Genome Res. 7,* 179-186.

Park, M.H., Lee, H.J, Bok, J., Kim, C.H., Hong, S.T., Park, C., Kimm, K., Oh B., and Lee, J.Y. (2006). Korean BAC library construction and characterization of HLA-DRA, HLA-DRB3. *J Biochem Mol Biol.* 39(4):418-425.

Roest, Crollius, H., Jaillon, O., Bernot, A., Dasilva, C., Bouneau, L., Fischer, C., Fizames, C., Wincker, P., Brottier,

P., Quetier, F., Saurin, W., and Weissenbach, J. (2000). Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence. *Nat. Genet.* 25, 235-238.

Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989). "Molecular Cloning: A Laboratory Manual", 2nd ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Shizuya, H., Birren, B., Kim, U.J., Mancino, V., Slepak, T., Tachiiri, Y., and Simon, M. (1992). Cloning and stable maintenance of 300-kilobasepair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proc. Natl. Acad. Sci. USA* 89, 8794-8797.

Veltman, J.A., Schoenmakers, E.F., Eussen, B.H., Janssen, I., Merkx, G., van Cleef, B., van Ravenswaaij, C.M., Brunner, H.G., Smeets, D., and Van Kessel. A.G. (2002). High-throughput analysis of subtelomeric chromosome rearrangements by use of array-based comparative genomic hybridization. *Am. J. Hum. Genet.* 70, 1269-1276.

Wade-Martins, R., Smith, E.R., Tyminski, E., Chiocca, E.A., and Saeki, Y. (2001). An infectious transfer and expression system for genomic DNA loci in human and mouse cells. *Nat Biotechnol. 19*, 1067-1070.

Wang, Z., Engle,r P., Longacre, A., and Storb, U. (2001). An efficient method for high-fidelity BAC/PAC retrofitting with a selectable marker for mammalian cell transfection. *Genome Res.* 11, 137-142.