A physical map of the human genome – supplemental information and detail.

The International Human Genome Mapping Consortium*

*Complete author list.

Regional approach to large-scale map physical map construction.

Following the positional cloning model, most centers employed sequence tagged site(STS)[1] landmark-content maps to identify regional markers suitable for isolating clones. Once identified, regional clones were characterized by either STS-content alone or in conjunction with restriction enzyme digest fingerprinting methods and assembled into contigs. As an example, the availability of a high-quality, high-resolution yeast artificial chromosome (YAC)[2] STS-content map of chromosome 7[3] made it a logical first target for the Washington University Genome Sequencing Center (WUGSC) human genome sequencing pilot project. In early 1996 the WUGSC began experimenting with an STS-driven clone identification paradigm using collective markers from approximately one megabase intervals simultaneously for clone identification. Initially, clones from a Genome Systems BAC library (Human BAC Release I; http://www.incyte.com/reagents/index.shtml) and from the early California Technical Institute BAC library (http://informa.bio.caltech.edu/Bac_info.html) were identified by the PCR using specific STSs in the laboratory of Eric Green, (NHGRI) and the well locations of STS-positive clones were communicated to the mapping group at the WUGSC. Simultaneously, methods for high-throughput hybridization-based BAC filter screening were developed at the WUGSC (J. McPherson). The general approach involved screening genomic BAC and PAC libraries by hybridization using overgo probes[4] to identify clones corresponding to specific STS markers. Overgo probes are made by filling in the single-stranded overhangs of two overlapping oligonucleotides using radiolabeled nucleotides and Klenow polymerase. Typically, two 24mers overlapping by 8 bp were used to generate a radiolabeled double-stranded 40mer. Overgo probes were arranged in three-dimensional arrays with 6 probes on each axis (6x6x6=216 probes each). A five directional pooling strategy allowed resolution of 80-90% of all markers with only 30 hybridizations. To date, greater than 25,000 human and mouse markers have been associated with BACs using this probe type at the WUGSC (J. McPherson). These efforts targeted the Genome Systems BAC library and the RPCI-4 and RPCI-5 PAC libraries

constructed by P. de Jong (http://www.chori.org/bacpac/). Evolution of the BAC filter screening method resulted in a procedure capable of simultaneous multiplex hybridization of 36 overgo[4] probes at a rate of 450 markers weekly. Once identified, BAC clones were retrieved from 384-well glycerol stocks and colony purified by re-streaking on agar plates. Individual colonies were cultured and DNA purified in 96-well format. DNA was digested with HindIII, the resulting restriction products resolved by agarose gel electrophoresis and analyzed using the FPC fingerprint assembly software package[5,6]. Subsequent efforts utilizing this strategy focused on chromosomes 2 and Y.

Whole Genome BAC map

To generate fingerprint maps of the *Caenorhabditis briggsae* and *Arabidopsis thaliana* genomes, we had evolved the capacity to produce weekly approximately 2,000 BAC fingerprints with a team of six technicians. Each gel analyzed consisted of 40 BACs and 11 standard "marker" DNA lanes. We calculated that to achieve our aim of 300,000 human BAC fingerprints using this gel technology would require 7,500 gels. We devised a tandem 121-lane agarose gel format, allowing the simultaneous electrophoresis of 50 standard "marker" DNA lanes and of 192 BAC restriction digests, thereby reducing the number of gels required. This system, with the increased density of fingerprint information on the gels, would produce the target 300,000 fingerprints in only 1,600 gels. We verified that the new agarose gel format produced accurate fingerprints by comparing fragment sizes between clones prepped and digested in duplicate and by fingerprinting previously sequenced clones and comparing the *in silico* digests of the sequence to the agarose gel fingerprints as shown in Figures 1a and 1b, respectively. With these and other improvements in the fingerprinting technology, and with addition of staff, throughput rose from 2,000 fingerprints weekly to a high of more than 20,000 weekly, the latter number representing approximately one human genome equivalent of fingerprinted DNA produced each week. The majority of the clones fingerprinted were from 3 libraries, RPCI-11, RPCI-13 and CT-C/D1. The fingerprint characteristics of the clones from these libraries are listed in Table 1.

Manipulation of fingerprint data

Due to considerable variability in the mobility of small fragments and that these small fragments were inconsistently identified during subsequent gel image analysis. Hence, fragments smaller than 600 base pairs (corresponding to a an Image mobility value greater than 3590) were removed prior to assembly. In addition, because of the difficulty in determining multiplets (cases where more than one fragment is located at nearly the same position on the gel), all fragments within mobilities of 8 were collapsed to only a single

band in the resulting fingerprint. This "sanitizing" process resulted in assemblies of increased reliability. Assembly parameters were determined empirically by changing various build parameters to obtain assemblies and then evaluating them for consistency using other mapping data associated with clones in the assembly (primarily radiation hybrid (RH) chromosomal localization data from the Stanford Human Genome Center (SHGC; D.R. Cox)). After assembly of the fingerprints with each set of parameters, the number of chimeric assemblies was detected by the presence of conflicting map information affixed to clones in the assembly. Optimal assembly parameters minimizing both the total number of assemblies and the number of chimeric assemblies are given in Table 2. Incremental assemblies of the fingerprints accumulated throughout the project are shown in Figure 2.

Using the FPC parameter values listed in Table 2, an automated assembly of the 283,287 clones resulted in 7,133 assemblies that contained 93% of all fingerprints in the database (December 1999). The remaining unincorporated clones (*i.e.* singletons) were excluded as they contained fewer than the 3 fragments specified as the minimum number required for assembly (min band = 3) or contained fewer than the total of 6 fragments that were empirically determined necessary for automated assembly under these conditions. Although the initial automated assemblies produced when using the FPC parameter values in Table 2 were generally reliable with respect to the "binning" of related clones, determination of the correct relative order of the clones within the assemblies remained a manual task.

Achieving map continuity by manual editing of assemblies

The initial goals of the manual assembly editing were to refine the relative ordering of the clones within contigs, to identify joins between contigs and to disassemble larger chimeric contigs. This process involved first editing the fingerprint assemblies, using the tools encapsulated in FPC, to ensure that every clone within a contig was properly situated with respect to its most highly related neighbors. A detailed discussion of the manual editing process can be found in Marra *et al.*, 1999.[7] Briefly, this was determined by minimizing the "Sulston" score[8] (a statistical measure of the of coincidental overlap) between adjacent clones and inspecting the fingerprint data extracted from the original gel image. Redundant clones contributing no unique fragment data to the assembly were also "buried" under their parent clones to simplify the contig view if this had not already been done in the automated process. During this manual inspection, additional clones were incorporated into the contig from the remaining pool of previously unassembled clones and newly fingerprinted clones that were not yet part of any assembly. Once the clone order had been established within each contig, clones at the extreme ends of each contig were used to query the FPC database at a reduced required fingerprint overlap stringency from that used to perform the initial assembly. In this way, potential joins between contigs were identified. Fingerprints of clones involved in potential joins were visually inspected to confirm that all restriction fragments were logically consistent and the joins were made if appropriate. The results of this editing activity are illustrated in Table 3. The most notable effect of the intensive editing effort was the nearly 5-fold reduction of total assemblies. The large increase in the number of singletons was due to the continued addition of new clones to the database without a total assembly being performed. Clones from the singleton pool were only manually incorporated into assemblies as needed to extend the ends of the contigs. As a rule, clones from the singleton pool were not incorporated into the interior of contigs where additional redundancy was not needed.

Integration of map data with the BAC contig database

The main purpose of generating a whole genome BAC contig map was to coordinate activities of the larger sequencing groups by making the FPC database available and to specifically select clones to feed the human sequencing capacities of the WUGSC, the Whitehead Institute for Biomedical Research (WIBR) and the Stanford Genome and Technology Center (SGTC). An essential component of the working draft strategy was that the selected clones for these centers be restricted as much as possible to chromosomes 2, 4, 7, 8, 11, 15,

17,18 and Y. Early on, mapped BACs were identified primarily from the hybridization of 13,695 overgo probes generated from sequences mapped to these chromosomes. These consisted predominantly of selected markers from the CEPH Généthon genetic map[9], the GeneMap'99 genome-wide RH map (http://www.ncbi.nlm.nih.gov/genemap/).[10-12] and from plasmid library sequences prepared from flow-sorted chromosomes (Sanger Centre, unpublished data). The 13,695 markers identified 96,283 unique clones, providing many anchor points for the assembled contigs. As the BAC assemblies enlarged and the fingerprint resource gained popularity as a tool for coordinating large-scale sequencing efforts, markers across the genome were sought to anchor all contigs. The remainder of the genome was to be sequenced by other members of the International Sequencing Consortium;[13] however, it was realized that the WG BAC map would also be a useful resource for other centers. To enhance the utility of the map to all centers, as many markers as possible were integrated into the FPC database, resulting in the chromosomal localization of most BAC contigs. Because the RPCI-11 library was being used for many genome initiatives there was a wealth of marker information available. A significant source of such information came from the inclusion of chromosomal assignment data for 9,018 STS derived from BAC end sequences (D.R. Cox, unpublished data). Although greater than 15% of the available BAC end sequences of the RPCI-11 library are apparently mislabeled with respect to the correct BAC name[14], the number of accurately mapped BAC ends facilitated the correct chromosomal assignment of contigs by considering the consistent trend of these data for a given contig. As the working draft sequence accumulated, known markers were readily identified utilizing electronic PCR (ePCR; a program that searches sequence for STS by identifying the associated primer sequences in the correct orientation and with correct spacing),[15] which identifies primers and these data provided for inclusion in the FPC database (G. Schuler, NCBI). The combined ePCR and hybridized data sets contained 69,507 markers, including 1,659 polymorphic markers from the Généthon genetic map. In addition, chromosomal assignment and integration of cytogenetic map positions were achieved by identification of 3,412 BACs mapped by fluorescence *in situ* hybridization (FISH) data as described in an accompanying manuscript.

The GeneMap'99 map was chosen for the purpose of anchoring and ordering contigs as it has a substantial marker set (> 50,000), is well-integrated with the Généthon genetic map and provides local ordering at <1 Mb resolution. It was also the most widely known genome map at that time. To ensure that markers assigned to BACs provided consistent localizations, single markers associated with clones in multiple contigs were not used for contig placement. For each contig, all chromosome assignment data associated with its clones were tallied and the contig was assigned to the chromosome with the majority of supportive data. Each chromosomal assignment datum was given equal weight with the exception of FISH data, which was considered the strongest evidence for placement. A clone could have more than one piece of supportive datum. For example, a clone with one chromosome 7 STS marker, chromosome 7 FISH data and BAC end sequence RH data for chromosome 2 would contribute two supportive data points for contig assignment to chromosome 7 and one for chromosome 2. In the case of a tie in the cumulative data for the contig, no initial chromosomal assignment was made.

Once a chromosomal assignment was made, the majority of contigs could be further localized to the median RH map position of all markers associated with the contig and assigned to that chromosome. To determine the median map position of a contig: 1. Each clone was allowed only one map position. This was the median position of all markers associated with that clone; 2. Markers that are likely outliers were

removed. A contig may have outliers if the map distance covered by the contig is greater than expected for a contig with the given number of clones. The outlier markers likely represent local mapping errors; 3. The median position was taken from the markers that remain.

The orientation of the contig with respect to the map was determined next. The marker content of accessioned sequences associated with the clones from the contig was determined by ePCR. Those markers found on one of five maps were examined (Généthon genetic map, Marshfield genetic map[16], WIBR YAC STS-content (http://carbon.wi.mit.edu:8000/cgi-bin/contig/phys_map), GeneMap'99, SHGC G3 RH map (http://www-shgc.stanford.edu/Mapping/rh/index.html), NCBI framework map (http://www.ncbi.nlm.nih.gov/genome/guide/)) and all mapped markers were then compared with their adjacent neighboring markers to determine if the order of markers within a contig for any given map were ascending or descending with respect to that map. Contigs were flipped as needed to orient them with respect to the majority consensus of all maps examined.

Lastly, finished map data provided by other groups for several chromosomes was examined. Although the contigs ordered by the methods above were largely consistent with these data, a few manual adjustments were made to more accurately reflect these well-characterized maps. The maps considered at this stage were for chromosomes 14 (Genoscope, J. Weissenbach)[17], 19 (Lawrence Livermore National Laboratory, http://www-bio.llnl.gov/bbrp/genome/html/chrom_map.html), X (Sanger Centre, D. Bentley) and a 20 Mb segment of chromosome 15 (University of Washington, L. Rowen). Telomeric contigs were identified and positioned where possible. The data for the telomeric contigs is discussed in a separate manuscript in this issue of *Nature*.[18]

Automated clone selection for sequencing

To maximize the amount of sequence produced while still ensuring the integrity of the clones selected, seed clones were chosen from contigs consisting of at least eight clones by finding the largest clone in each contig that had no more than 2 bands that were not confirmed by neighboring clones. Two fragments in each clone were assumed to be the insert-vector junction fragments as HindIII was used to generate fingerprints from this BAC library constructed using a partial EcoRI restriction digest. Once selected, each seed clone was compared to *in silico* digests of all of the human clones in GenBank to identify any

clones that may overlap previously generated sequence. If clones were found to excessively overlap an existing sequence by a Sulston score of $<3x10^{-6}$, or if the overlap suggested localization to a chromosome other than those specifically targeted, then the seed clone was discarded. Based on the estimated sizes of the chromosomes and the number of seed clones selected from contigs localized to each chromosome, it was determined whether each chromosome had been seeded to a point of saturation at approximately one contig per 500 kb. When this point was reached, seed clones were no longer chosen from that chromosome and these nucleation points were extended to avoid excessive gaps when completing the sequence map. Since the fingerprint database was growing weekly and all data were not yet incorporated into existing contigs, a set of tools was developed that allowed extension from seed clones by identifying overlapping fingerprints in the rest of the database independent of contig assembly. Once manually ordered contigs became available, the automated clone selection tools attempted to walk in a minimal tiling path, except in cases where a clone chosen for a walk overlapped heavily with another sequenced clone. In this case, a bridging clone with significant overlap to both neighbors was chosen to complete the coverage. For choosing gap-spanning clones and for choosing minimal tiling paths, BAC end data, when available and consistent, were used in conjunction with fingerprint data. To complete the working draft, the end of each contig was extended where possible using fingerprint analysis, even if the degree of overlap was higher than our allowed degree of overlap for automated clone selection (25%). Clones were automatically selected from the ends of contigs when at least 30 kb of new sequence data would likely be obtained.

Throughout the clone selection process for the working draft, a central registry was maintained at the NCBI to track clones being sequenced by all groups in the consortium. Incorrect clone names and the depth of clone coverage for any given region limited the usefulness of this registry but it did provide a means for examining maps for conflicts. In addition, the first set of 96 sequence reads for any clone was used to check for overlap with known sequences. Initially, a clone was discarded at this stage if greater than 25% of sequences matched another clone. This overlap cutoff was increased as the working draft accumulated to allow for gap-spanning clones to enter the pipeline. Even with all these efforts, redundancy within the working draft sequence is higher than the <20% that was targeted; however, this redundancy in fact may have benefits for analysis of sequence variation within the genome.

Mapping of accessioned sequences.

For the placement of each accessioned sequence, *in silico* digests of each sequence were used to query the FPC human BAC database. If the FPC database fingerprint with the highest restriction fragment similarity to the *in silico* digest corresponded to the clone referenced in the associated GenBank record, then the accessioned sequence was mapped to this position; however, if it corresponded to a different clone than that referenced in the GenBank record, additional information was used to position this sequence as accurately as possible. In many cases, the highest matching fingerprint was associated with a clone overlapping with, or completely redundant to, the referenced clone. This could occur when working draft sequence was used as all restriction fragments may not be adequately represented in the assembled sequence and thus would be absent from the *in* silico digest. In these cases, the accessioned sequence was positioned with respect to the GenBank referenced clone. In a third scenario, the highest matching fingerprint was associated with a clone mapping independently of that referenced in the GenBank record. If the *in silico* digest match to the fingerprint was sufficiently stringent (Sulston score $<1x10^{-9}$) then the accessioned sequence was mapped to the matching position. If the match was less stringent, then two pieces of additional information were required to confirm this placement. Additional information considered was the alignment of BAC end sequences to the accessioned sequence. If the majority of these BAC end alignments represented clones overlapping with the matching clone, then the accessioned sequence was placed in that position on the map. At the time of last analysis, there were 302,351 BAC end sequences that could be aligned to 32,235 HTGS human sequences. Of these alignments, 83,219 of them were from clones contained in the WG BAC map and confirmed or contributed to accessioned sequence placement. Additional evidence for placement of the sequence came from examining believed sequence overlaps with neighboring sequenced clones. The believed overlaps were generated by an all-against-all sequence comparison of all accessioned sequences (G. Schuler and J. Kent). The believed overlaps were ranked as either strong (includes additional supporting data such as paired BAC end linkages) or weak (sequence overlap alone) and were weighted accordingly in this consideration.

1. Olson, M., Hood, L., Cantor, C. & Botstein, D. A common language for physical mapping of the human genome. *Science* **245**, 1434-5 (1989).

2. Burke, D. T., Carle, G. F. & Olson, M. V. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* **236**, 806-12 (1987).

3. Bouffard, G. G. et al. A physical map of human chromosome 7: an integrated YAC contig map with average STS spacing of 79 kb. *Genome Res* **7**, 673-92 (1997).

4. Ross, M. T., LaBrie, S., McPherson, J. & Stanton, V. P. in *Current Protocols in Human Genetics* (eds. Dracopoli, N. C. et al.) 5.6.1-5.6.5 (John Wiley and Sons, New York, 1999).

5. Soderlund, C., Longden, I. & Mott, R. FPC: a system for building contigs from restriction fingerprinted clones. *Comput Appl Biosci* **13**, 523-35 (1997).

6. Soderlund, C., Humphray, S., Dunham, A. & French, L. Contigs Built with Fingerprints, Markers, and FPC V4.7. *Genome Res* **10**, 1772-1787 (2000).

7. Marra, M. et al. A map for sequence analysis of the Arabidopsis thaliana genome. *Nat Genet* **22**, 265-70 (1999).

8. Sulston, J. et al. Software for genome mapping by fingerprinting techniques. *Comput Appl Biosci* **4**, 125-32 (1988).

9. Dib, C. et al. A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**, 152-4 (1996).

10. Schuler, G. D. et al. A gene map of the human genome. *Science* **274**, 540-6 (1996).

11. Deloukas, P. et al. A physical map of 30,000 human genes. *Science* **282**, 744-6 (1998).

12. The International Radiation Hybrid Mapping Consortium, A new gene map of the human genome: GeneMap'99. (cited October 2000) *http://www.ncbi.nlm.nih.gov/genemap/* (1999).

13. The International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, (this issue).

14. Zhao, S. et al. Human BAC ends quality assessment and sequence analyses. *Genomics* **63**, 321-32 (2000).

15. Schuler, G. D. Electronic PCR: bridging the gap between genome mapping and genome sequencing. *Trends Biotechnol* **16**, 456-9 (1998).

16. Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. & Weber, J. L. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am J Hum Genet* **63**, 861-9 (1998).

17. Bruls, T. et al. A physical map of human chromosome 14. Nature (this issue).

18. Reithman, H. C. et al. Integration of telomeric DNA sequences with the human working draft sequence. *Nature* (this issue).

The International Human Genome Mapping Consortium

Full author list

Washington University School of Medicine, Genome Sequencing Center:

http://genome.wustl.edu

John D. McPherson[1], Marco Marra1[*], LaDeana Hillier[1], Robert H. Waterston[1], Asif Chinwalla[1], John Wallis[1], Mandeep Sekhon[1], Kristine Wylie[1], Elaine R. Mardis[1], Richard K. Wilson[1], Robert Fulton[1], Tamara A. Kucaba[1], Caryn Wagner-McPherson[1], William B. Barbazuk[1] Amanda Aabbott[1], Johar Ali[1], Stephanie Andrews[1], Larisa Belaygorod[1], Gary Bemis[1], Amy Berghoff[1], Merry Brook[1], Marco Cardenas[1], Jason Carter[1], Jim Cloud[1], Marc Cotton[1], Jye'mon Crockett[1], Kristy Drone[1], Feiyu Du[1], Jen Edwards[1], Jackie Fedele[1], Dan Fischer[1], Nat Florence[1], Catherine Franklin[1], Tony Gaige[1], Diane Gaige[1], Marilyn Gibbons[1], Neenu Grewal[1], Heather Grover[1], Chris Gund[1], Gwen Harmon[1], Njata Harvey[1], Shunfang Hou[1], Sara Jaeger[1], Corrie Joshu[1], Sara Kohlberg[1], Colin Kremitzki[1], Dan Layman[1], Shawn Leonard[1], Jason Maas[1], Ken MacDonald[1], Catherine Marquis-Homeyer[1], Rachel Maupin[1], Ryan Mcadow[1], Cindy McCabe[1], Rebecca McGrane[1], Kelly Mead[1], Richard Morales[1], Nancy Mudd[1], Christine Nguyen[1], Phil Ozersky[1], Carrie Ragan[1], Amy Reiley[1], Kerry Robinson[1], Ellen Ryan[1], Samuel Sasso[1], Debra Scheer[1], Kelsi Scott[1], Kelsi Scott[1], Ryan Seim[1], Karina Shapiro[1], Proteon Shelby[1], Aimee Smith[1], Tamberlyn Stoneking[1], Hui Sun[1], Carrie Sutterer[1], Elizabeth Sweet[1], Brenda Theising[1], Jane Threideh[1], Rebecca Walker[1], J. Patrick Woolley[1], & Martin Yoakum[1]

Wellcome Trust Genome Campus:

http://www.sanger.ac.uk/

Simon G. Gregory[2], Sean J. Humphray[2], Lisa French[2], Richard S. Evans[2], Graeme Bethel[2], Adam Whittaker[2], Jane L. Holden[2], Owen T. McCann[2], Andrew Dunham[2], Carol Soderlund[2*], Carol E. Scott[2] & David R. Bentley[2]

National Center for Biotechnology Information:

http://www.ncbi.nlm.nih.gov

Gregory Schuler[3], Hsiu-Chuan Chen[3] & Wonhee Jang[3]

National Human Genome Research Insititute:

http://genome.nhgri.nih.gov/chr7/

Eric D. Green[4], Jacquelyn R. Idol[4] & Valerie V. Braden Maduro[4]

Albert Einstein College of Medicine:

http://sequence.aecom.yu.edu/chr12/

Kate T. Montgomery[5], Eunice Lee[5], Ashley Miller[5], Suzanne Emerling[5] & Raju Kucherlapati[5]

Baylor College of Medicine, Human Genome Sequencing Center:

http://www.hgsc.bcm.tmc.edu/

Richard Gibbs[6], Steve Scherer[6], J. Harley Gorrell[6], Erica Sodergren[6], Kerstin Clerc-Blankenburg[6], Paul Tabor[6], Susan Naylor[7] & Dawn Garcia[7]

Roswell Park Cancer Institute:

Pieter J. de Jong[8*], Joseph J. Catanese[8*], Norma Nowak[8] & Kazutoyo Osoegawa[8*]

Multimegabase Sequencing Center:

Shizhen Qin[9], Lee Rowen[9], Anuradha Madan[9], Monica Dors[9] & Leroy Hood[9]

Fred Hutchinson Cancer Research Institute:

Barbara Trask[10], Cynthia Friedman[10]& Hillary Massa[10]The Children's Hospital of Philadelphia: Vivian G. Cheung[11], Ilan R. Kirsch[12], Thomas Reid[12] & Raluca Yonescu[12]

Genoscope:

Jean Weissenbach[13], Thomas Bruls[13] & Roland Heilig[13]

US DOE Joint Genome Institute:

http://www.jgi.doe.gov/JGI_home.html

Elbert Branscomb[14], Anne Olsen[14], Norman Doggett[14], Jan-Fang Cheng[14] & Trevor Hawkins[14]

Stanford Human Genome Center and Department of Genetics:

http://www-shgc.stanford.edu/

Richard M. Myers[15], Jin Shang[15], Lucia Ramirez[15], Jeremy Schmutz[15], Olivia Velasquez[15], Kami Dixon[15], Nancy E. Stone[15] & David R. Cox[15]

University of California, Santa Cruz:

http://genome.ucsc.edu

David Haussler[16,17], W. James Kent[18] , Terrence Furey[17] , Sanja Rogic[17] & Scot Kennedy[19]

Department of Genome Analysis, Institute of Molecular Biotechnology:

http://genome.imb-jena.de/printHum.html

André Rosenthal[21], Gaiping Wen[21], Markus Schilhabel[21], Gernot Gloeckner[21], Gerald Nyakatura[21]*, Reiner Siebert[22] & Brigitte Schlegelberger[22]

Departments of Human Genetics and Pediatrics, University of California:

Julie Korenberg[23] & Xiao-Ning Chen[23]

RIKEN Genomic Sciences Center:

http://hgp.gsc.riken.go.jp/

Asao Fujiyama[24], Masahira Hattori[24] , Atsushi Toyoda[24] , Tetsushi Yad[24] ,

Hong-Seok Park[24] & Yoshiyuki Sakaki[24]

Department of Molecular Biology, Keio University School of Medicine:

http://www.dmb.med.keio.ac.jp

Nobuyoshi Shimizu[25], Shuichi Asakawa[25], Kazuhiko Kawasaki[25], Takashi Sasaki[25], Ai Shintani[25], Atsushi Shimizu[25], Kazunori Shibuya[25], Jun Kudoh[25] & Shinsei Minoshima[25]

Max-Plank-Institute for Molecular Genetics:

http://seq.molgen.mpg.de/hgs/

Juliane Ramser[26] , Peter Seranski[26,27] , Celine Hoff[26,27] , Uwe Radelof[26], Ralf Sudbrak[26],Annemarie Poustka[26,27], Richard Reinhardt[26] & Hans Lehrach[26]

1. Washington University School of Medicine, Genome Sequencing Center, Department of Genetics, 4444 Forest Park Boulevard, St. Louis, Missouri 63108, USA

2. Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

3. National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland 20894, USA

4. National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

5. Department of Molecular Genetics, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, New York 10461, USA

6. Baylor College of Medicine, Human Genome Sequencing Center, Houston, Texas, USA
http://www.hgsc.bcm.tmc.edu/

7. University of Texas, San Antonio, Texas, USA

8. Roswell Park Cancer Institute, Buffalo, New York 14263, USA

9. Multimegabase Sequencing Center, Institute for Systems Biology, Seattle, Washington 98105, USA

10. Division of Human Biology, Fred Hutchinson Cancer Research Institute, Seattle, Washington 98109, USA

11. Department of Pediatrics, The Children's Hospital of Philadelphia, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

12. Genetics Department, Medicine Branch, National Cancer Institute, Washington DC, USA

13. Genoscope, Centre National de Séquencage, 2 Rue Gaston Crémieux, CP 5706, 91057 Evry, France

14. US DOE Joint Genome Institute, Walnut Creek, California, USA

15. Stanford Human Genome Center and Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA

16. Howard Hughes Medical Institute, University of California, Santa Cruz, Santa Cruz, California 95064, USA

17. Department of Computer Science, University of California, Santa Cruz, Santa Cruz, California 95064, USA

18. Department of Biology, University of California, Santa Cruz, Santa Cruz, California 95064, USA

19. Department of Mathematics, University of California, Santa Cruz, Santa Cruz, California 95064, USA

20. British Columbia Cancer Research Centre, 600 West 10th Avenue, Room 3427, Vancouver, British Columbia V5Z 4E6, Canada

21. Dept. of Genome Analysis, Institute of Molecular Biotechnology, Beutenbergstrasse 11, D-07745 JENA, Germany

22. Institute of Human Genetics, University of Kiel, Germany

23. Departments of Human Genetics and Pediatrics, University of California, Los Angeles, California, USA

24. RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan

25. Department of Molecular Biology, Keio University School of Medicine, 35 Shinanomachi Shinjuku-ku, Tokyo 160-8582, Japan

26. Max-Plank-Institute for Molecular Genetics, Ihnestrasse 73, D-14195, Berlin, Germany

27. Abteilung Molekulare Genomanalyse, Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

*Present addresses: British Columbia Cancer Research Centre, 600 West 10th Avenue, Room 3427, Vancouver, British Columbia V5Z 4E6, Canada (M.M.); Clemson University Genome Institute, 100 Jordan Hall, Clemson University, Clemson, South Carolina 29634-5727, USA (C.S.); Children's Hospital Oakland Research Institute, BACPAC Resources, Oakland, California 94609, USA and Pfizzer Global Research & Development, Alameda Laboratories, Alameda, California 94502 USA (P.J.d.J., J.J.C., K.O.); MWG-Biotech AG, Ebersberg, Germany (G.N.).

Figure 1: Assays of fingerprint reproducibility and fingerprint accuracy.

Figure 1a: Verification of fingerprint reproducibility.

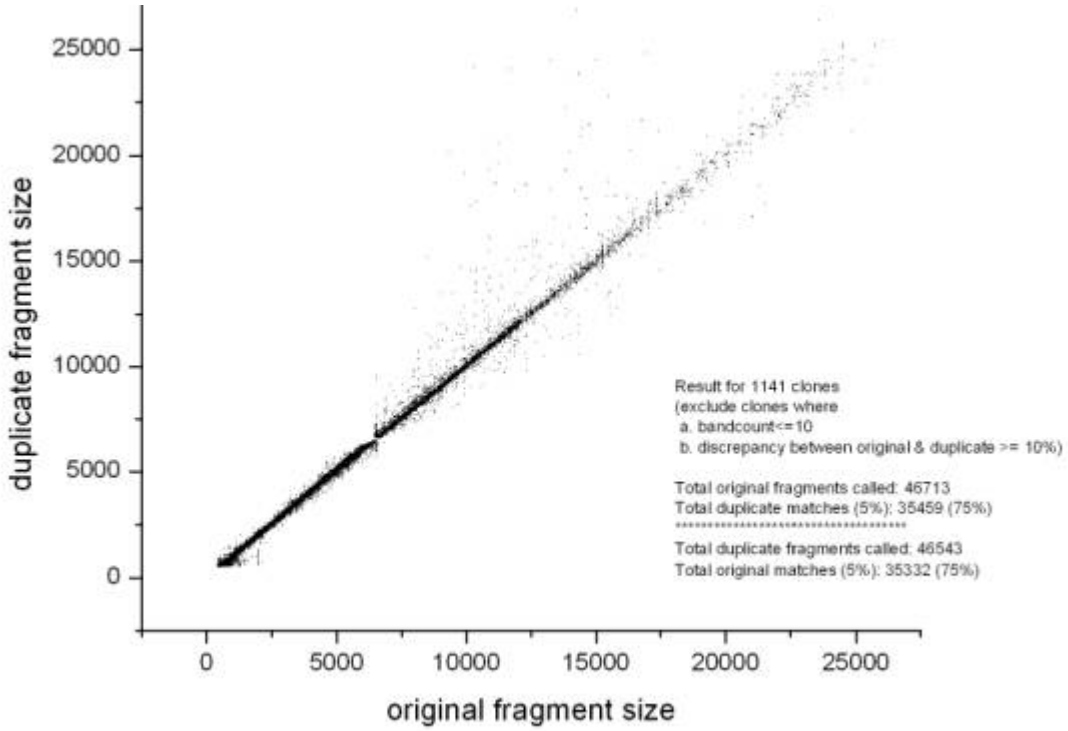
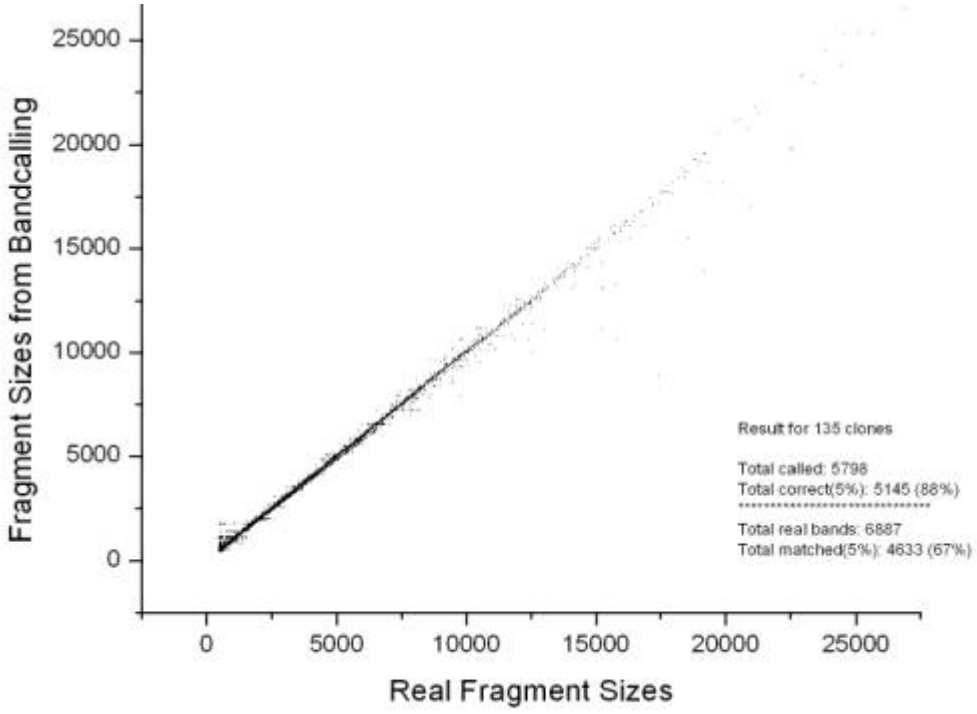


Figure 1b: Verification of fingerprint accuracy.



Shown in Figure 1a is a comparison of clone fingerprints, generated in duplicate, in which each point represents a restriction fragment compared between duplicates. Shown in Figure 1b are fingerprint-derived restriction fragment sizes compared to computer-derived "*in silico*" virtual digests predicted from DNA sequence data. Restriction fragments were considered identical if their sizes (base pairs) were within 2% of each other.

Table 1: Fingerprint characterization of the major clone classes used.*

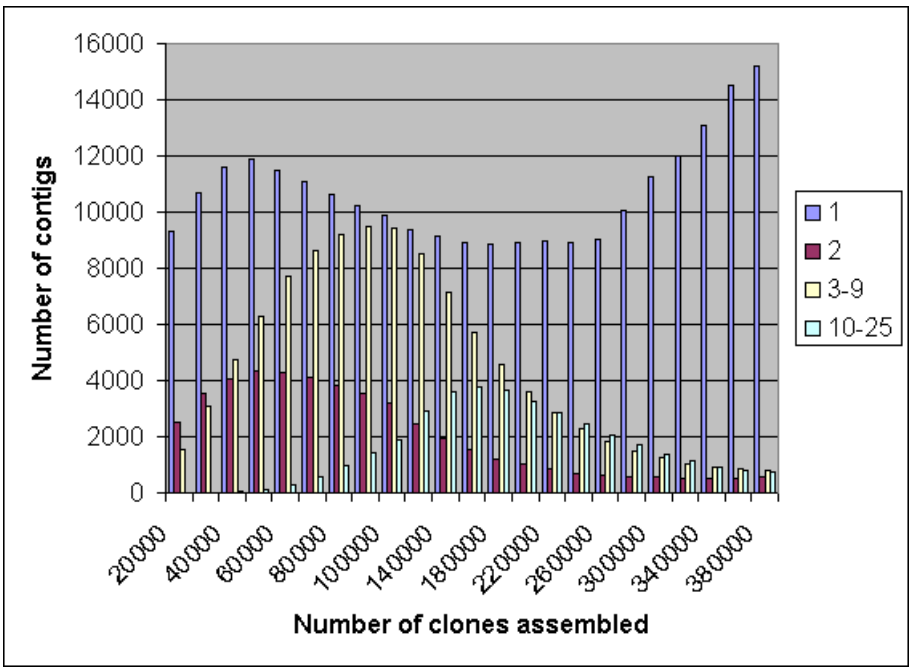
Library	Clones	Ave. insert (kb) (std. dev.; median)	Ave. number of bands (std. dev.; median)
RPCI-11	265,619	167.7 (21.1; 167.4)	41.8 (7.9; 42)
RPCI-13	50,202	151.5 (27.9; 156.6)	38.6 (9.5; 40)
CTC-C/D1	49,349	123.4(29.1; 120.5)	30.9 (8.6; 30)

*Only clones with an estimated insert size greater than 50 kb and less than 300 kb were included to minimize the effect of fingerprinting artifacts.

Table 2: Optimal FPC parameters used for fingerprint assembly.

FPC Parameter	Value
tolerance	7
cutoff	3x10-12
automated bury	0.10
use CpM	no
min bands	3
best of	10

Figure 2: Incremental assemblies of the BAC fingerprints.



At the various points indicated, the database of BAC fingerprints was assembled into contigs using the FPC software suite. Assembly parameters are as in Table 2. The figure key indicates the range of the number of clones per contigs in each group considered (contigs containing 10-15 clones, 3-9 clones, two clones and remaining clones with no match to another clone or "singletons".

Table 3: Status of FPC database after automated assembly and manual editing.

	Automated assembly	Manually edited database
Date	December, 1999	September, 2000
BAC clones in FPC	283,287	372,264
Number of assemblies	7,133	1,447
Clones in assemblies	264,555	295,828
Number of singletons*	11,022	76,436
Assemblies containing:		
>25 clones	2,652	912
10-25 clones	1,720	260
3 – 9 clones	1,886	204
2 clones	875	71

* clones not incorporated into any assembly; see text.