Supplementary Information for Initial Sequencing and Analysis of the Human Genome.

International Human Genome Sequencing Consortium.

Methods and additional notes

Section: Generating the draft genome sequence (p. 864)

Subsection: Clone selection (p. 865)

Page 866 col. 2, para.3 "Fingerprint data were reviewed ….bias against rearranged clones).

Seed clones were picked from the growing contigs as follows: We began by identifying fingerprint clone contigs that had been localized to targeted locations and that did not contain any clones that had previously been selected for sequencing. Contigs were localized using mapping data from a variety of sources that could be attached to the fingerprinted clones, including STS/hybridization data from McPherson and colleagues[86], FISH data from several sources (C. McPherson et al., ref. 103), STS/PCR mapping data from several sources[92,95,103], electronic PCR data (http://www.ncbi.nlm.nih.gov/STS/) matching the BAC end sequences with mapped STSs and others. Beginning with the largest available clone in a valid contig (clones >250 kb were excluded to avoid artifacts), the FPC program[451] evaluated the fingerprints of all of the clones in the contig to determine largest clone for which all (but 2) of the individual bands in the restriction fragment pattern were common to or shared with (confirmed; having a band of equivalent size ±3%) with bands in the patterns of flanking clones (again, ignoring >250 kb flanking clones >250 kb). (Since the restriction enzyme used to produce the clone inserts is different than the enzyme used to produce the fingerprints, two bands may arise from the insert-vector junction, which are not found in the genome or in flanking clones.) Selected clones were then checked for excessive overlap with previously selected or sequenced clones and with each other. The allowable overlap at this stage was varied to suit the demands of the project.

Clones (walking clones) extending from seed or other selected clones were selected as follows: In the early phases of the effort, clones were not necessarily correctly ordered within a fingerprint clone contig and indeed not all of the available clones had necessarily been incorporated into the contig. Starting with a previously selected (seed) clone, the FPC program compared the restriction fragment pattern of that clone with the patterns of all of the clones in the fingerprint database that overlapped with the seed clone. It then iteratively analyzed the clones identified in the first round of analysis to identify the additional clones that overlapped with those. In this way, a set of overlapping clones was identified and the clones in the set were ordered based on their overlap statistics. After ordering, all of the valid clones were identified (valid clones were defined as those with all but three of their bands confirmed by clones within 4 clones on either side). Any clone that also had outside evidence of overlap, e.g. through BAC end sequence matches or shared

> STS/hybridization data was selected for further evaluation. In cases with more than one clone with such outside evidence, the clone with the lowest overlap statistic (i.e., the one that was least redundant) was selected (in the case of ties, the largest clone was favored). Where there was no outside evidence, a clone was picked based on evaluation of the overlaps. The candidate clone was the first one that was found to have the minimal overlap with the seed clone (initially <20% overlap, rising to 30% in later phases of the mapping effort; the percentage overlap was estimated by dividing the sum of the sizes of the common bands by the size of the smaller of the two clones). To be picked, the clone also had to be bridged to the seed clone by a third, intermediate clone that confidently (<1e^{-4}) overlapped both the seed clone and the candidate clone. The candidate clone was then further evaluated for fingerprint overlap with previously selected or sequenced clones.
>
> Once clones were ordered within fingerprint clone contigs, a similar algorithm that exploited the known clone order was used to pick the walking clones. This algorithm was also adapted to pick a spanning/walking clone for complex contigs with 2 or more clones in the sequencing pipeline, using the fingerprint map as a guide.

Subsection: Sequencing (p. 867)

Page 868, left-hand column, line 20: “By examining … 500 bp.”

> The sizes of the gaps between adjacent initial sequence contigs in draft clones were measured using alignments of the initial sequence contigs from individual draft clones to contigs of size ≥ 40 kb from overlapping clones, usually finished clones. 10,999 gaps were examined. 1,726 gaps larger than 6,000 bp were discarded as probable artefacts due to misassemblies or incorrect alignments. The mean size of the gaps between the initial sequence contigs in draft clones was 554 bases. When the cutoff for discarding gaps was lowered to 3000 bp or raised to 12,000 bp, the mean gap size decreased to about 400 bp (estimated from 9,801 gaps) and increased to about 800 bp (estimated from 11,972 gaps) accordingly, indicating that there is still considerable uncertainty in the mean value. The 554 bp estimate for the mean gap size was used, along with the number of initial sequence contigs (Table 7) and the total number of bases in the initial sequence contigs (data not shown) to estimate the percentage of the draft clones that were covered by the initial sequence contigs. It was thus determined that, on average, about 96% of the draft clones was covered; assuming a mean gap size between 400 and 800 bp, the range in coverage is about 94-97%.

This comment also pertains to page 874, left-hand column, line 57: “Assuming that the sequence gaps … gaps within the draft sequenced clones”

Subsection: Assembly of the draft genome (p. 868)

Page 868, right-hand column, l. 47, "To eliminate such problems, sequenced clones were associated with the fingerprint clone contigs in the physical map…"

> An FPC match statistic better than 1e^{-7} for the sequenced clone against the fpc fingerprint database was considered significant, based on empirical evidence. This match level was the weakest value used for placement when there was other confirmatory evidence to support the placement. In the absence of additional supportive data, a match score of better than 1e^{-9} was required for placement. In general, only the best match was used. Other confirmatory evidence included BAC end matches; the BAC end sequences were obtained from NCBI (dbGSS; http://www.ncbi.nlm.nih.gov/dbGSS/index.html). Only BAC end sequences with 15 or fewer matches to the genomic sequence were used to eliminate repetitive sequences. Additional information used to place clones included BAC paired-end sequence matches, shared STS matches, and "believed" sequence overlap relationships determined by investigators at the NCBI and at UC-Santa Cruz. In instances in which the data led to conflicting placements, the data were weighted based on estimates of reliability. In some cases, if there was conflicting placement data or only weak data for placement and, according to GigAssembler, the sequenced clone failed to overlap any clones in the assembly at their original placement positions, a placement was attempted at secondary sites suggested by the placement data.

Page 869, left-hand column, line 48 "Of these 942 contigs with sequenced clones… "

> In general, merges between fingerprint clone contigs were based primarily on evaluation of the fingerprint data. Information about the STS map location of the fingerprint contigs was used to prevent spurious merges, to break spurious contigs and to suggest possible merges that had not been previously recognized. In addition, 62 contigs were merged on the basis of sequence overlap information, supported by STS map positions.

Subsection: Quality assessment (p. 871)

Sub-subsection: Alignment of the fingerprint clone contigs (p. 873)

Page 873, right-hand column, line 28: "The positions of most of the STSs… about 1.7% differed from one or more of them."

> We localized the STS markers from seven different physical maps (the Genethon[101] and Marshfield (http://research.marshfieldclinic.org/genetics/) genetic maps, the GeneMap99[100], the G3 and Stanford TNG radiation hybrid maps (http://www-shgc.stanford.edu/Mapping/Marker/STSindex.html), and the Whitehead YAC and radiation hybrid map[29]) on the draft genome sequence using e-PCR, allowing one mismatch per primer and the default distance constraints between primers (50 bp deviation from expected size of product). Only those markers that were uniquely placed on the draft sequence were considered. There were 62,239 such markers. Of these, 1,095, or 1.7%, were mapped by ePCR to a chromosome of the draft sequence that was different from the chromosome indicated by the information from a genetic or radiation hybrid map.

Subsection: representation of random raw sequences (p. 874)

Page 875, left-hand column, line 9: "We compared the raw sequences … using the BLAST computer program."

> We processed whole genome shotgun reads from four independently constructed libraries as follows. All reads with fewer than 300 bases of PHRED quality 20 or greater were removed. The remaining reads were then trimmed for vector and for quality, looking at the 5' end for the first window with at least 15 continuous non-vector bases of >PHRED20 and at the 3' end, starting from the left cutoff, for 12 contiguous non-vector bases with <PHRED20 scores. Only trimmed reads that had >95% of their trimmed bases with PHRED>20 and a length of >250 bases were kept. The reads after trimming were composed of 40% GC base pairs. Reads were masked for repeats using the RepeatMasker program (A.F.A. Smit & P. Green, http://repeatmasker.genome.washington.edu/cgi-bin/RM2_req.pl) and for low entropy data using the nseg option of BLAST (W. Gish, unpublished; http://blast.wustl.edu)Reads were retained and used only if there were at least 100 consecutive bases of PHRED quality 20 or greater and 100 consecutive unmasked bases.
>
> Based on a test data set of random reads from finished projects, the following BLAST parameters were found to match 100% of the reads without false matches: -filter seg S=170 S2=150 W=13 gapW=4 gapS2=150 M=5 N=-11 Q=11 R=11. The set of masked trimmed reads was compared to the 7 October 7 2000 freeze of the HTGS data set, to all of Genbank and to the TSC SNP database using BLASTN 2.0MP (W. Gish, unpublished; http://blast.wustl.edu). The highest scoring match was aligned against the read using CROSSMATCH, demanding alignment of the full trimmed read at ≥97% identity for genomic sequence and with appropriate topological constraints for the SNP reads. Typically 1-2% of the matches were eliminated by this step.

Page 875, left-hand column, line 30: "We found that 88% of the bases of these cDNAs could be aligned ..."

> We aligned the RefSeq cDNA sequences to the draft genome using the psLayout program[104] and gathered statistics on the percentage of cDNA bases that aligned at various percent identity thresholds.
>
> The distal 200 bases of each cDNA were not included in the computation of the percentage of aligning bases because alignments in these regions are less reliable. If any cDNA aligned in more than one way, each cDNA base involved in any alignment was counted only once. At a threshold of 98% identity for the alignments, we found that 87.9% of the cDNA bases aligned somewhere in the draft genome. When the threshold was increased to 99% identity, the percentage of aligning bases fell to 85.83%, and when the threshold was decreased to 97% identity, it rose to 88.5%. Further decreases in the threshold all the way down to 90% identity only

increased the percentage of aligning bases one more percentage point, so the value of approximately 88% aligning bases, achieved by requiring 98% identity, represents a knee in the curve.

Section: Broad genomic landscape (p. 875)

page 876, right-hand column, line 9: “In addition, the human cytogenetic map ...”

The locations of the cytogenetically mapped clones on the draft genome sequence can be viewed at http://genome.ucsc.edu/goldenPath/mapPlots . Further information about the individual clones can be obtained at http://www.ncbi.nlm.nih.gov/genome/cyto/ and http://www.ncbi.nlm.nih.gov/genome/guide. Here, as well as on the browser at http://genome.ucsc.edu and http://www.ensembl.org/ , they can be viewed in the context of other genome annotation.

Subsection: Long-range variation in GC content (p. 876)

Page 877, left-hand column, line 30 “About three-quarters of the genome-wide variance… consistent with a homogeneous distribution”

All 3,312 windows of length 300 kb that had at least eight gap-free 20 kb subwindows and did not contain more than 50% simple repeats were extracted from the draft genome sequence. The average sample variance of the GC content of the subwindows of a window was 7.3%. The sample variance of all subwindows genome-wide (N = 36,562) was 27.4%. Hence, the variance of GC content within the 20 kb subwindows of a 300 kb window accounts for approximately one quarter of the overall variance of the GC content among all 20 kb subwindows in this sample. The average sample standard deviation of the GC content of the subwindows of a window was 2.4%.

Page 877, left-hand column, line 34: “In fact, the hypothesis … draft genome sequence.”

For each of the 3,312 windows of length 300 kb, we tested the hypothesis that its 20 kb subwindows were sampled from a homogeneous GC distribution. The distribution was defined to have mean m equal to the GC-content in the combined subwindows of the 300 kb window, and the bases were taken as independent. Under this distribution, the GC-content of a 20 kb subwindow would have mean m and variance $s^2 = m(100-m)/20000$. For m = 41%, the typical value, this gives $s^2 = 0.121\%$, which is about 0.017 times the average sample variance of 7.3%. For each window, the variance s^2 and the sample variance $\hat{s}^2$ were determined, along with the value $c^2 = (n-1)\,\hat{s}^2/s^2$, where n is the number of subwindows of the window. Under the hypothesis of homogeneity, the statistic c^2 should have an approximately chi-square distribution with n-1 degrees of freedom. However, for every one of the 3,312 windows, $c^2 > 31.5$, which rejects the hypothesis of homogeneity with p-value >> 0.995.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS

Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS

Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS

Sync your system to PACER to automate legal marketing.

WHAT WILL YOU BUILD? | sales@docketalarm.com | 1-866-77-FASTCASE

fastcase®
Smarter legal research.