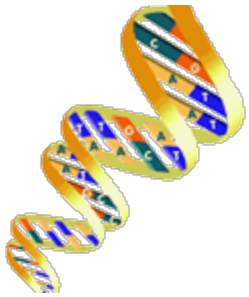


Exceptional Chromosome Regions II

May 24, 2001
Oakland, California

Site Provided by the U.S. Department of Energy

An HGP Program Perspective
Marvin Stodolsky

Agenda

Abstracts

- Exploring TAR cloning for isolation and analysis of exceptional regions of human genome
  Natalay Kouprina, Carl Barrett, and Vladimir Larionov
  - Additional Publications (from PubMed)
    - Abstract: "Closing the gaps on human chromosome 19 revealed genes with a high density of repetitive tandemly arrayed elements," S. H. Leem, N. Kouprina, J. Grimwood, J. H. Kim, M. Mullokandov, Y. H. Yoon, J. Y. Chae, J. Morgan, S. Lucas, P. Richardson, C. Detter, T. Glavina, E. Rubin, J. C. Barrett, and V. Larionov, Genome Res. 2004 Feb;14(2):239-46. Epub 2004 Jan 12.
    - Abstract: "A novel strategy for analysis of gene homologues and segmental genome duplications," V. N. Noskov, S. H. Leem, G. Solomon, M. Mullokandov, J. Y. Chae, Y. H. Yoon, Y. S. Shin, N. Kouprina, and V. Larionov, J Mol Evol. 2003 Jun;56(6):702-10.
    - Abstract: "Segments missing from the draft human genome sequence can be isolated by transformation-associated recombination cloning in yeast," N. Kouprina, S. H. Leem, G. Solomon, A. Ly, M. Koriabine, J. Otstot, E. Pak, A. Dutra, S. Zhao, J. C. Barrett, and V. Larionov, EMBO Rep. 2003 Mar;4(3):257-62.

- Chromosome 19: Sequencing Status and Remaining Problem Regions
  Anne Olsen
- Segmental Duplications: Organization and Impact Within the Current Assembly
  Jeffrey A. Bailey, Amy M. Yavor, Julie E. Horvath, Barbara Trask, and **Evan E. Eichler**
  - Eichler Laboratory (http://eichlerlab.gs.washington.edu/)
- Gap Closing of Chromosome 16 with Unigene Sequences
  Cliff Han and Norman Doggett
- Chromosome 16 Mapping Data Resources
  Robert D. Sutherland, Cliff S. Han and **Norman A. Doggett**
- Status Report on Gap Closure of the Human Chromosome 5 BAC Map
  Steve Lowry, Joel Martin, Duncan Scott, and **Jan-Fang Cheng**
  - Presentation
- Integration of Telomere Sequences with the Draft Human Genome Sequence
  Robert K. Moyzis

Presentation
- Challenges to Human Genome Sequencing: Not so Cryptic Duplications and the Genomic Abyss
  Julie R. Korenberg, Xiao-Ning Chen, Pranay Bhattacharyya
- RPCI-11 BES Pairs to Human Genome
  Shaying Zhao
  - Presentation

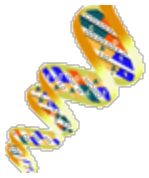
ECR I

Abstracts and References

- Human Telomere Mapping and Sequencing
  R.K.Moyzis, H. Chi, D.L. Grady, and H. Riethman
- One Origin of Man: Primate Evolution Through Genome Duplication
  Julie R. Korenberg, Xiao-Ning Chen, Steve Mitchell, Rajesh Puri, Zheng-Yang Shi and Dean Yimlamai
- Abstract
  Evan E. Eichler
- References for *Arabidopsis* centromer studies
  Greg Copenhaver & Daphne Pruess
- Abstract
  Gary H. Karpen
- Abstract
  David C. Schwartz
- Reconstruction and Annotation of Transcribed Sequences: The TIGR Gene Indices
  John Quackenbush, Ingeborg Holt, Feng Liang, Geo Pertea, Jonathan Upton, and Thomas S. Hansen
- Direct isolation of a centromeric region from the human Y chromosome by TAR cloning for structural and functional studies
  Natalay Kouprina and Vladimir Larionov
- Nucleic Acid Strands Specific for Binding via Non-Denaturing Triple Helix Formation (TISH) to Centromeric DNAs of Each of the Human Chromosomes and to the DNA Common to the Tips of the Achrocentrics
  Jacques R. Fresco
- References on sequencing through difficult chomosomal regions
  John Dunn
- References on duplicated regions
  Barbara Trask
- Finding Genes in Centromeric Regions of the Genome
  Jim Tucker

Please contact Marvin Stodolsky (Marvin.Stodolsky@science.doe.gov) with updates or corrections.

This site produced by the Human Genome Management Information System of Oak Ridge National Laboratory.



Exceptional Chromosome Regions II

Home

An HGP Program Perspective

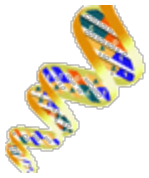
Marvin Stodolsky
U.S. Department of Energy
Marvin.Stodolsky@science.doe.gov

The existence of chromosomal regions with exceptional structures and or posing exceptional analytical problems has been manifest, since the recognition of telomeres and centromeres. Over time, the definition of these ECR have become more precise, with several contributions from within the DOE HGP. Robert Moyzis recognized the repeating sequence of the telomere tips. This led to strategies for their cloning with Bob's follow though into the sequencing of the near telomeric regions. When BAC resources began to mature in Mel Simon's lab, Julie Korenberg with Xiao-Ning Chen used FISH to characterize the resource. I recall the feeling of relief during the next HGP Workshop, from the result that any chimerism of the resource was not worse than 5%, as compared to the common chimerism of YAC resources. I recall feeling of mild terror, as it became progressively apparent that much of the 5% represented real duplications/replications. A 5% of 3.5 megabases was potentially a Lot of Trouble. Barbara Trask's FISHing added the complication, that there was heterogeneity in the human population with respect to the duplication pattern of the near telomeric sequences across the chromosomes. In the course of mapping a fragile X site, David Nelson's team discovered that a copy had been transposed to a near centromeric region of chr22. It has become progressively more apparent, that the centromeric regions are relative hotbeds for acquiring distant genes, reshuffling them locally, and with occasional flings to distant regions of the genome. Detail on these dynamics has been filled in by Evan Eichler and collaborators. While coverage of the genome by BACs and Pieter de Jong's PACs was manifestly good, both teams had recognized a few holes and unstable clones. Natasha Kouprina and Vladimir Larionov are now achieving a rough quantitation of their prevalence. Perhaps it is fortunate that the ECR couldn't be quantitated earlier, as it might have been necessary to acknowledge a very rough Battle in Progress as contrasted to a Victory in Hand.

Last modified: Wednesday, October 22, 2003

Base URL: www.ornl.gov/meetings/ecr2/

Site sponsored by the U.S. Department of Energy Office of Science, Office of Biological and Environmental Research, Human Genome Program



Exceptional Chromosome Regions II

Home

Agenda

Integration of Telomere Sequences with the Draft Human Genome Sequence
Robert K. Moyzis

Status Report on Gap Closure of the Human Chromosome 5 BAC Map
Steve Lowry, Joel Martin, Duncan Scott, and Jan-Fang Cheng

Exploring TAR cloning for isolation and analysis of exceptional regions of human genome
Natalay Kouprina, Carl Barrett, and Vladimir Larionov

Segmental Duplications: Organization and Impact Within the Current Assembly
Jeffrey A. Bailey, Amy M. Yavor, Julie E. Horvath, Barbara Trask, and Evan E. Eichler

Chromosome 16 Mapping Data Resources
Robert D. Sutherland, Cliff S. Han and Norman A. Doggett

Gap Closing of Chromosome 16 with Unigene Sequences
Cliff Han and Norman Doggett

Chromosome 19: Sequencing Status and Remaining Problem Regions
Anne Olsen

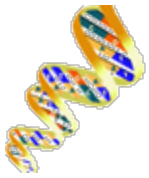
Challenges to Human Genome Sequencing: Not so Cryptic Duplications and the Genomic Abyss
Julie R. Korenberg, Xiao-Ning Chen, Pranay Bhattacharyya

An HGP Program Perspective
Marvin Stodolsky

Last modified: Wednesday, October 22, 2003

Base URL: www.ornl.gov/meetings/ecr2/

Site sponsored by the U.S. Department of Energy Office of Science, Office of Biological and Environmental Research, Human Genome Program



Exceptional Chromosome Regions II

Home

Exploring TAR cloning for isolation and analysis of exceptional regions of human genome

Natalay Kouprina, Carl Barrett, and Vladimir Larionov
Laboratory of Biosystems and Cancer, Center for Cancer Research, National Cancer Institute, NIH, Bethesda MD 20854

The TAR (Transformation-Associated Recombination) cloning technique allows a direct isolation of specific chromosomal regions and genes from mammalian genomes without a time-consuming step of construction of a representative library of random clones. The technique is based on homologous recombination between a vector containing a gene-specific sequence and a genomic DNA fragment during co-transformation into yeast spheroplasts. Using this technique, chromosomal regions up to 400 kb can be rescued in yeast as circular YACs. These YACs can be easily retrofitted into BACs and transferred into *E. coli* cells for further structural and functional studies. A new technology has several potential utilities for a human genome project. Among them: i) verification of the assembled contigs, ii) closing the gaps and iii) isolation of centromeric regions and other exceptional regions that can not be cloned by a routine technique based on *in vitro* ligation.

Isolation of exceptional regions that are not clonable in *E. coli* cells

During last two years a TAR cloning strategy was successfully applied in our lab for isolation of dozen of genes entire copies of those have not been found in BAC libraries. Most of the genes were structurally stable during propagation in yeast cells as YACs and in *E. coli* cells as BACs. However, some genes, including the metastasis-suppressor gene *KAI1* and the mouse mucin gene *MUC2* were stable in yeast cells but can not be stably transferred into *E. coli* cells. During electroporation, the YAC/BACs containing these genes exhibited a low efficiency of *E. coli* transformation (approximately 100 times lower compared to that of other randomly selected BACs with a similar size insert). Moreover, all the transformants obtained contained large deletions eliminating up to 90% of the insert. We demonstrated that such inserts represent a significant fraction of human genome (~6%). Analysis of such regions requires isolation of circular YACs directly from yeast cells.

Verification of the assembled contigs

End sequencing of YAC/BAC clones isolated from human genome by TAR allows to verify a quality of contigs assembling. Such analysis revealed several mistakes. Some YAC/BAC clones contained inserts with sizes bigger than that in contigs, suggesting a loss of internal sequences during assembling of the contigs. Other clones contained sequences mapped to different contigs. Thus, the TAR cloning technique can greatly simplify verification of contigs assembling.

Isolation of centromeric and pericentromeric regions

The TAR cloning technique has been also applied for isolation and characterization of human centromeric

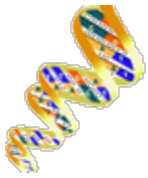
regions. These regions remain the most poorly mapped areas in the human genome and they are underrepresented in the libraries. Using the TAR technique we isolated and physically characterized centromeric and pericentromeric regions from seven human chromosomes (2, 5, 8, 15, 16, 22 and X).

Therefore, we conclude that TAR cloning provides a powerful tool for verification of the genome draft sequence and for development of continuous sequences of human chromosomes. Significance of TAR cloning may be even more important for mouse genome that has not been physically mapped beforehand.

Last modified: Wednesday, October 22, 2003

Base URL: www.ornl.gov/meetings/ecr2/

Site sponsored by the U.S. Department of Energy Office of Science, Office of Biological and Environmental Research, Human Genome Program



Exceptional Chromosome Regions II

Home

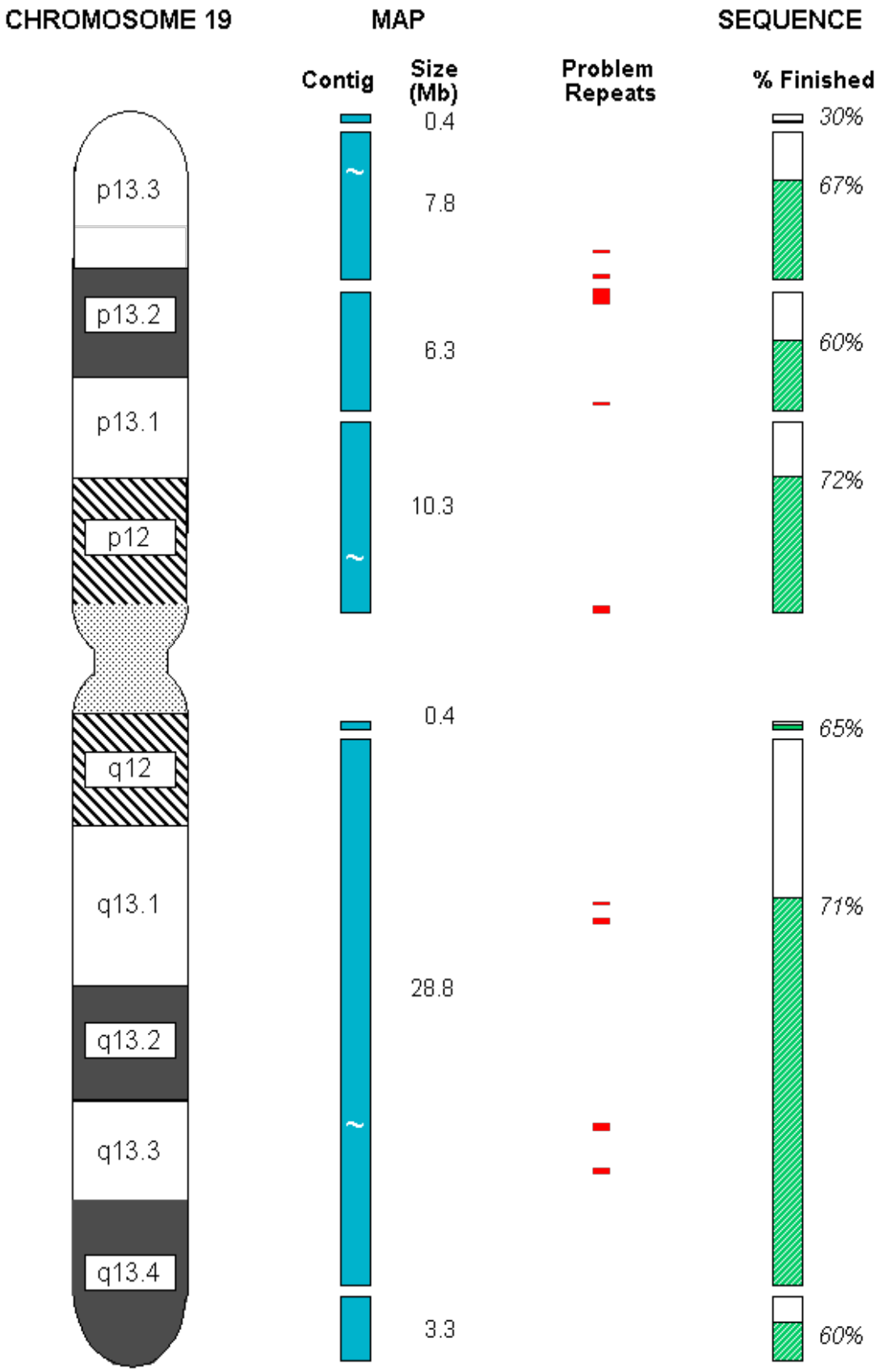
Chromosome 19: Sequencing Status and Remaining Problem Regions

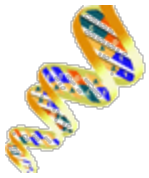
Anne Olsen
DOE Joint Genome Institute, Walnut Creek, CA, 94598
olsen2@llnl.gov

The map of chromosome 19 consists of seven contigs spanning a total of over 57 Mb. It covers >98% of the length of the p and q arms, as estimated by pronuclear FISH and restriction mapping. The five remaining map gaps (excluding the centromere) are estimated by FISH to be small (i.e. less than single BAC size), but we have been unable to close them by hybridization screening of cosmid and BAC libraries, BAC end sequence analysis, or analysis of FPC contigs. The most distal p-arm gap occurs within the SLI gene and is spanned by cDNAs but no genomic clones. Two of the gaps are adjacent to, or within, areas with concentrations of problem repeats that have presented challenges for sequencing and probe development.

Map coverage appears to extend relatively close to the centromere and telomere on each arm. Sequence from the most proximal p-arm and q-arm clones contains alpha satellite repeats, suggesting that the map extends close to the centromeric junction. The most distal p-arm cosmid is highly homologous to a telomeric YAC, yRM2001, and cosmids from this region exhibit FISH signals on multiple telomeres. Sequence from the most distal q-arm cosmid contains small blocks of (TTAGGG)n repeats, suggesting a subtelomeric location. The most distal q-arm cosmids also have FISH signals on multiple telomeres.

A total of 39 Mb of unique sequence, about two-thirds of the chromosome, is now finished. Most of the remainder of the chromosome is currently in finishing, but about 2 Mb, distributed in 7 locations, has been unfinishable due to problem repeats. Cosmids are being sequenced to supplement BAC coverage to assist in finishing these areas, but special sequencing strategies may be required for some repetitive regions. A summary of the current status of the map and finished sequence, indicating the location of problem repeat regions, is illustrated in the accompanying figure (drawn to scale, except gap sizes exaggerated for visibility).


CHROMOSOME 19
MAP
SEQUENCE
Contig
Size (Mb)
Problem Repeats
% Finished
p13.3
p13.2
p13.1
p12
q12
q13.1
q13.2
q13.3
q13.4
0.4
7.8
6.3
10.3
0.4
28.8
3.3
30%
67%
60%
72%
65%
71%
60%



Exceptional Chromosome Regions II

Home

Segmental Duplications: Organization and Impact Within the Current Assembly

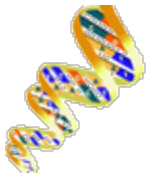
Jeffrey A. Bailey, Amy M. Yavor, Julie E. Horvath, Barbara Trask, and **Evan E. Eichler**
Department of Genetics and Center for Human Genetics, Case Western Reserve School of Medicine and University Hospitals of Cleveland, Cleveland, OH, 44106

Segmental duplications play fundamental roles in both genomic disease and gene evolution. Over the past year, we developed the computational tools and methods necessary to detect identity between long stretches of genomic sequence despite the presence of high copy repeats and large insertion-deletions. We focused our analysis on large recent duplication events that fell well-below levels of draft sequencing error (alignments 90-98% similar and >1 kb in length) revealing an unprecedented amount of duplicated sequence (3.6%) in the human draft assembly (oo15). Here we present a more refined analysis of the most recent genome assembly (oo23) in which we focus on the role duplications play in whole-genome assembly process. Duplications (90-98%; > 1 kb) comprise 3.6% of all sequence in oo23. These duplications show clustering and up to 10-fold enrichment within pericentromeric and subtelomeric regions. Despite this bias, complex regions of duplication have also been identified within gene rich regions. In terms of assembly, duplicated sequences are 6.7-fold over-represented in unordered and unassigned contigs indicating that duplicated sequences are difficult to assign to their proper position. Further, utilizing data from 134 sequence BACs with FISH signals to multiple chromosomes, only 57% (280/571) of chromosomes positive by FISH had a corresponding chromosomal BLAST hit to oo23. We present data that indicates that this is due to misassembly/misassignment and decreased sequencing coverage within duplicated regions. Suprisingly, if we consider putative duplications >98%, we identify 10.3% (286 Mb) of the current assembly as paralogous. At high similarities (>98%) 10.3% of the sequence is involved in pairwise alignments. The majority of these alignments, we believe, represent unmerged overlaps within unique regions. Taken together the above data indicates that segmental duplications represent a significant impediment to accurate human genome assembly, requiring the development of specialized techniques to finish these exceptional regions of the genome. Specific examples from chromosomes 16 and 19 will be presented.

Last modified: Wednesday, October 22, 2003

Base URL: www.ornl.gov/meetings/ecr2/

Site sponsored by the U.S. Department of Energy Office of Science, Office of Biological and Environmental Research, Human Genome Program



Exceptional Chromosome Regions II

Home

Gap Closing of Chromosome 16 with Unigene Sequences

Cliff Han and Norman Doggett
Los Alamos National Laboratory

825 Unigenes from chromosome 16 were BLAST against the draft and finished contigs of this chromosome. 41 unigene clusters (5%) did not hit with a similar genomic sequence using a cut-off of >97% identity and >70 bp in length. These 41 sequence were then searched against the whole human genome draft and finished contigs, and 7 of these are located on other chromosomes (1, 14, 5, 17, 19). Of the remaining 34 Unigene clusters without chromosome 16 genomic sequence one is LTR repeat, and two have poor sequence quality leaving 31 clusters for use in BAC library screening. The resulting numbers (784/815 Unigene clusters which match chromosome 16 genomic sequence) suggest that there is 96.2% coverage of the chromosome represented in the draft and finished public databases. We have selected PCR primers for the 31 Unigenes that have potential to close gaps in the 16 map and will use these to screen the RPCI-11 library. This primer list is attached.

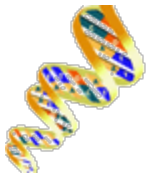
Hs.1137.LEFT	TGTTCAGCCTGCCTTCTTTT
Hs.1137.RIGHT	CTCTTTAAGGAGATGGGGGC
Hs.3076.LEFT	TTGTTTGGGGATGAGAAAGG
Hs.3076.RIGHT	CCTCTTGACACCCAGCTCTC
Hs.3112.LEFT	GCATCCAATTCGTGTCTGTG
Hs.3112.RIGHT	GAGGGGTCAGCTCTGTCTTG
Hs.8040.LEFT	GTCCCTTCCCATCAGTTTGA
Hs.8040.RIGHT	ACACGGTTTTCATTTAGCGG
Hs.9003.LEFT	GTGTTCTGTGCCTGTGTGCT
Hs.9003.RIGHT	GATTAAAGCCTCCCAAAGCC
Hs.9204.LEFT	CACCTTCATCACCCACCTCT
Hs.9204.RIGHT	CAGACCGTCCACTTCCACTT
Hs.9605.LEFT	AATGCACCAGGATATGGACC
Hs.9605.RIGHT	TGTGCTCACAGAGACAAGCG
Hs.74170.LEFT	AGTGACGCGACAACAGACAC
Hs.74170.RIGHT	GGTTGTGAGCTATCCGTGGT
Hs.75124.LEFT	GCCTCCATCCCTGATTCTAA
Hs.75124.RIGHT	GGGAAGGGGTTTGATCTTTC
Hs.75648.LEFT	GCCCTATTCACACTCTGGGA
Hs.75648.RIGHT	GGAACATGAGAGGCCAAAAA
Hs.91910.LEFT	TCCTTCAGGAAGCCACTCTC
Hs.91910.RIGHT	CATCATCACATGAACCCTGC
Hs.96023.LEFT	GGACCCAATCATGAGGAAGA
Hs.96023.RIGHT	TGAATCTTGGGGACTTGAGG
Hs.99880.LEFT	CTTTGGGCCTATAGGGGTGT

Hs.99880.RIGHT	TAAATTAGCGCAAAATGGGG
Hs.118786.LEFT	GACATTACGCGAAGCACTCA
Hs.118786.RIGHT	TTTGTTCTGACTGTCGCCTG
Hs.122823.LEFT	CCCCTCAGACCTCCTCATCT
Hs.122823.RIGHT	CTCCCTCCTTTCCTAGGTGG
Hs.145271.LEFT	CTCAGGAACAAGGGGAAACA
Hs.145271.RIGHT	GTTCCCAAAAACCCCTTCAG
Hs.157113.LEFT	GGGTGGAGTTCAGGATCAGA
Hs.157113.RIGHT	GTCCTGCCAAGTCATCAGGT
Hs.169850.LEFT	CAAACATCAGAGCAATCCGA
Hs.169850.RIGHT	TAGGGGGTCGTATGATCAGG
Hs.183075.LEFT	ACCTCATTTCCTCCAACGTG
Hs.183075.RIGHT	ATGTAGCGGAAGAAGAGCCA
Hs.189762.LEFT	CACAGGTGTGGGTCTCTCAA
Hs.189762.RIGHT	TTTTTCCCACTGTGGTGTCA
Hs.203936.LEFT	GGACCCCAACTGCTCCTG
Hs.203936.RIGHT	AGGAGCAGCAGCTCTTCTTG
Hs.241388.LEFT	TCCAACGCTTATGGGCTATC
Hs.241388.RIGHT	CAGCGACCACTGGAATCATA
Hs.250760.LEFT	CATCATGGCTTTGCCTCTG
Hs.250760.RIGHT	TGTTGGTAGGGACGGAACTC
Hs.275462.LEFT	GTGGCCATTAGCTTGGTGTT
Hs.275462.RIGHT	TGCCTTGCACACTAAGGTCC
Hs.275675.LEFT	GCCAGTCTGGACTGAGGAAA
Hs.275675.RIGHT	GTTCAGGGCTGTAGAGGTGG
Hs.278462.LEFT	CTTCTCCTTGCCTCGAAATG
Hs.278462.RIGHT	AGGAGCAGCAGCTCTTCTTG
Hs.331895.LEFT	AGAAGCACATGGGGGTATTG
Hs.331895.RIGHT	GGACCCATTGTTCTGTGCTT
Hs.332417.LEFT	CAAAGAGGGGCTTAGCACAG
Hs.332417.RIGHT	TTCAACTTTGCAAGAGGGCT
Hs.158336.LEFT	GAAGCCTCGTGCTACCCTG
Hs.158336.RIGHT	TAGATGTGGGAGCAGCGG
Hs.1973.LEFT	CTGTGAGCACACCACTGTCC
Hs.1973.RIGHT	CTGTGGGTCACATACGAACG
Hs.76159.LEFT	ATCCTAGTGCCCATCGTCTG
Hs.76159.RIGHT	GGCGACACACAAGCTAAGGT

Last modified: Wednesday, October 22, 2003

Base URL: www.ornl.gov/meetings/ecr2/

Site sponsored by the U.S. Department of Energy Office of Science, Office of Biological and Environmental Research, Human Genome Program



Exceptional Chromosome Regions II

Home

Chromosome 16 Mapping Data Resources

Robert D. Sutherland, Cliff S. Han and **Norman A. Doggett**
Los Alamos National Laboratory, Joint Genome Institute, MS M888, Los Alamos, NM 87545

We are coordinating the complete sequencing of human chromosome 16 by continued mapping efforts and improvement of the fully integrated map. This includes both restriction fragment and sequence based data for the BAC map of this chromosome. The data is maintained both in the SIGMA application and in multiple MS Excel spreadsheets. The Excel spreadsheets are divided into seven major regions--three p-arm and four q-arm maps. The three p-arm sections are the telomere to the FMF region (16p13.3), FMF region to the beginning of the TIGR effort (16p13.12), and the inner twenty Mb's initially sequenced by TIGR (upto 16p11.2). The q-arm is divided roughly into four equal parts. The size and distribution of the maps has a great deal with the spreadsheet size limitation in MS Excel. As map size increases, they are further subdivided to worksheets to facilitate data handling.

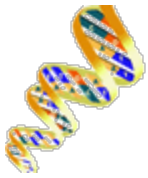
The chromosome 16 map has been years in the making starting with a flow-sorted cosmid library for which fingerprinted cosmids were assembled into cosmid contigs. STS's from these cosmids became the anchor points for assembly of a whole chromosome low-resolution YAC map and then a high-resolution BAC map. Sequencing commenced on this integrated framework. Five of the seven Excel map sections contain overgo and STS versus BAC-hit data.

There are gaps in the high-resolution BAC map resulting primarily from uneven STS or overgo distribution and to a lesser degree to uneven coverage of BAC libraries. These gaps have been identified and flanking BACs are analyzed with TIGR's BAC-end database search to find the optimum new BAC to extend into the gap. Using this technique, we walk across or meet in the middle of the gap. Another strategy is to find unique sequences that have no representation in the already sequenced BACs and screen for new ones.

This full data set, is included in four Excel files, that are available on the Web at www.jgi.doe.gov. Two different files represent the chromosome: 16parm and 16qarm. Both files have multiple worksheets denoting the different sections of the map. The two additional spreadsheets are for reference purposes. The first file is the chromosome 16 tiling set which contains all available genomic sequenced clones with accession numbers in chromosome order. This file also shows the minimal tiling set and size summaries at the bottom. It also contains entries for clones labeled as chromosome 16 but are either bad or reside on other chromosomes. The last file is the chromosome 16 reference list that will help locate on which map a BAC or STS can be found.

Any questions, input, or corrections to the chromosome 16 map may be directed to Robert Sutherland at rds@lanl.gov.

Last modified: Wednesday, October 22, 2003



Exceptional Chromosome Regions II

Home

Status Report on Gap Closure of the Human Chromosome 5 BAC Map

Steve Lowry, Joel Martin, Duncan Scott, and **Jan-Fang Cheng**
Lawrence Berkeley National Laboratory, Joint Genome Institute

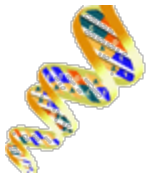
- Presentation

The human chromosome 5 BAC contig map is being assembled by restriction fragment overlaps and sequence overlaps. Other information used to confirm the map include BAC end sequence matches, FISH data, UniSTS matches, and FPC data. The current chr5 map consists of 95 gaps with bridging clones (from the FPC map) and 33 gaps with no bridging clones. We compared the flanking sequences of the 33 unbridged gaps with Celera assembled sequence and found that 11 are located within a scaffold, suggesting that these gaps are clonable by E coli vectors. The reminding 22 gaps include the 5p telomeric region and a gap surrounding the centromere. For the 5q telomeric region, the work of Dr. Harold Riethman's group has identified a half-YAC telomeric clone spanning the 120kb gap between the telomeric end of RP11-324K20 (the most distal BAC on 5q) and the molecular telomere. A cosmid contig is being constructed for sequencing. The most distal BAC on 5p, RP11-811I15, whose physical distance from the molecular telomere is not known, contains characteristic subtelomeric repeats. For the pericentromeric regions, 6 YACs were identified to span 1.9 Mb and 2.7 Mb of the p and q arms, respectively. The alphoid repeat subsets were found in the YACs flanking the centromere. Restriction maps from digests with 5 rare cutters have been determined for the 6 YACs and can be used to estimate the size of the gaps between BAC contigs surrounding the centromere. We are in the process of screening additional human BAC libraries to fill the remaining physical gaps.

Other regions requiring careful scrutiny are those containing conflicting map and sequence data. For instance, we have found 24 BAC sequences containing UniSTSs from more than one chromosome even though they all mapped to chr5 by FISH. In 7 BACs, the sequences containing multi-chromosomal UniSTSs were assembled into a single sequence contig. Possible explanations are chimeric clones, interchromosomal duplications, or misassembly. The other 17 dubious sequences may contain mixed projects since the UniSTSs hit different sequence contigs. We also have 23 BACs that were mapped to chr5 by FISH but contain UniSTSs from other chromosomes. This apparent discrepancy may be the result of interchromosomal duplications or clone/sequence mistracking. There are three large intrachromosomal duplicated segments on chr5, one on the p arm and two on the q arm. The BACs spanning these regions can be identified in FISH hybridization.

Last modified: Tuesday, December 18, 2018

Base URL: www.ornl.gov/meetings/ecr2/



Exceptional Chromosome Regions II

Home

Integration of Telomere Sequences with the Draft Human Genome Sequence

Robert K. Moyzis
University of California, Irvine
rmoyzis@uci.edu

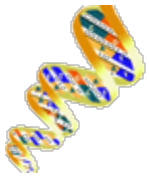
Presentation

Human telomeres end with a stretch of the conserved simple repeat sequence (TTAGGG)n. To capture single-copy human DNA regions linked to telomeres, large telomere-terminal fragments of human chromosomes were cloned using specialized yeast artificial chromosome (YAC) vectors. By contrast, bacterial artificial chromosome (BAC) libraries are not expected to contain sequences extending to the telomere, owing to the absence of restriction sites in (TTAGGG)n, the effects of length associated with the construction of size-selected DNA recombinant clones, and the genomic instability of these regions. By DNA end-sequencing of cosmid subclones derived from telomere YACs, connection to the working draft human sequence has now been accomplished (Riethman et. al., Nature 409, 948-951, 2001; www.genome.uci.edu). Integration with the working draft sequence was confirmed for 32 telomeres (out of the 46 distinct ends), with framework sequence extending to within 250kb-50kb of the physical end of these chromosomes. The remaining 14 telomere ends have not been connected, due to 1) inability to obtain YAC clones (5p and 20q), 2) the lack of extension of the working draft sequence into telomeric regions, and/or ambiguous identification due to repetitive sequences (2q,7p,17p,17q,19p,19q,and Xp/Yp), and 3) cloning instability/repeats in the ribosomal gene containing acrocentric chromosomes (13p,14p,15p,21p, and 22p). Subtelomeric sequence structure appears to vary widely, mainly as a result of large differences in subtelomeric repeat sequence abundance and organization at individual telomeres. Many subtelomeric regions appear to be gene-rich, matching both known and unknown expressed genes.

The successes and problems encountered with human telomeric regions suggest a number of directions that future research on ECRs might take:

a. A focused effort to confirm current assemblies and close remaining gaps will be critical. These are difficult regions, and must be attacked by a variety of methods. RARE cleavage analysis (to determine gap length/assembly consistency) and TAR cloning (to target gaps), have great potential.
b. The great variability in subtelomeric (and other ECR) regions between individuals has potential biological significance. Finishing a "single" sequence in these regions has little meaning, without extensive population/species sampling. A number of ECRs should be targeted for "high-depth" reanalysis/resequencing/haplotyping.

Last modified: Tuesday, December 18, 2018



Exceptional Chromosome Regions II

Home

Challenges to Human Genome Sequencing: Not so Cryptic Duplications and the Genomic Abyss

Julie R. Korenberg, Xiao-Ning Chen, Pranay Bhattacharyya
Departments of Human Genetics and Pediatrics, UCLA and Cedars Sinai, Medical Genetics Birth Defects Center, Los Angeles, CA, 90048
And Collaborators*

Like doughnuts, the holes in the genome are not found in the final product. Most of the gaps in the finished sequence are due to the instability of genomic regions containing repeated sequences. This results in underrepresentation in recombinant libraries, misaligned clone maps and difficulties in sequencing. The underlying sequence structure of such regions ranges from clustered (centromeres, pericentromeres, telomeres,) or interspersed reiterated simple sequences to megabase sized highly conserved regions of genomic DNA. Such regions are important to sequence in order to elucidate their roles in cellular function, human genomic variability, germ line and somatic instability, and gene regulation. Therefore, in recognition of the unavoidable biases that were introduced in sequence sets and their descendants by such unstable regions, a set of BACs were defined at random and integrated with draft sequence to provide anchor points for sequencing centromeres, pericentromeres and duplications in chromosome arms. Clones likely to fill unsequenced gaps in chromosomes 5, 16 and 19 are defined below.

Clones from Exceptional Regions were defined as follows: Spanning the human genome, a total of 6,000 BACs were mapped by in situ hybridization, 3500 defined at random, 184 from screens using concensus alpha satellite, 346 using telomeric oligonucleotides and about 2000 from other screens of the Caltech BAC libraries A and B. STS linkage was established for about 957, end sequences for 272 and fingerprints for 976 clones.

Centromeric Regions: A total of 373 mapped to centromeric regions (~40Mb), of which 192 mapped to single centromeres (~20Mb), 150 mapped to multiples and 31 (~3.4Mb) to all human centromeres (defined as universals). For Chromosome 5, 43 BACs are centromeric, of which one is specific, five map to both 5 and 19, seven map to multiple chromosomes and 30 are universals. For chromosome 16, in addition to the universals, one BAC is specific, two map to three sites and six map to 6 or more sites. For chromosome 19, there are no specific BACs, five map to chromosomes 5 and 19, and five to 4 or more sites.

Chromosome Arm FISH Multisite Clones: A Majority were defined at random.

Fingerprint Data:
Of the total mapped BACs, 990 revealed multiple sites, of which 350 (of 3500) had been defined at random suggesting that a minimum of 10% of the genome was duplicated in addition to the known regions. Of the 990 multisite, 489 were fingerprinted, of which thirty-three with 5-29 bands had no database match, and 58 had too few bands, suggesting a minimum of 8 % of duplications (non centromeric) were not represented in the

fingerprint database. For chromosome 5, eight of 56 BACs (14%) having at least one FISH site on chromosome 5 were not represented in the FP database and 3 were on orphan contigs. For chromosome 16, 13 of 51 BACs (25%) were not represented in the FP database, four were located on orphan and three at the ends of contigs. For chromosome 19, six of 22 (27%) were not represented in the FP database and one mapped to an orphan contig. For chromosomes 5, 16 and 19, a minimum of 35 BACs from the current dataset, or 50 from the complete multisite set may contribute to filling gaps in the draft sequence.

End Sequence Data based on 272 of 718 multisite BACs (2/2001)
Two hundred and seventy-two multisite BACs had at least one end sequence available (446 further remain to be analyzed), of which 160 had at least one database hit of greater than 80% homology, of which 95 (60%) were above 98% homology and 65 (40%) had hits of 80-98%. Three were located on orphan contigs. This suggested that at least 40% of the multisite clones detected repeated regions which were not included in the draft sequence. One hundred and twelve of the Multisite BACs or 41% had no match in the draft sequence.

Analyses 5.23.01 for Chromosome 5, 16, 19.
For chromosome 5, of 25 end sequenced multisite BACs two had no hits (8%) and 16 had hits of 98% (two on chromosome 5 and 4 of which had multiple sites) and 7 had hits below 98%, with two mapping on orphan contigs. This suggested that 13 of the original 25 represented repeated families for which all clones had not been sequenced. For chromosome 16, three of 26 multisite BACs had no hits in draft sequence. Of the 23 hits, 14 were to multiple sites (Less than 98%) and 9 were to single sites. This suggested that 12% (3/26) defined unsequenced repeats and 61% (14/23) defined repeats for which not all members had been sequenced, some on 16. For chromosome 19, 2/8 multisite BACs had perfect hits and the remaining 6 hits were less than 98%, or on multiple chromosomes, suggesting that 75% represented repeat families for which the original BAC had not been sequenced. This results in a minimum of 36 clones in the current dataset and 70 predicted for the complete set of multisite BACs, that may yield sequence information for 5, 16 and 19.

In summary, we have defined a subset of BAC reagents for duplicated regions in the genome, a number of which are neither mapped nor sequenced. End sequencing the remainder of the 446 multisite clones may provide defined reagents that, together may help to cover a total of ~11 Mb of regions related to or duplicated on chromosomes 5, 16 and 19. The clones defined in the current report as not present in the draft sequence may derive both from the random approach to clone selection and the use of an alternative library. Such end sequenced BAC clones that are not included in the genome draft sequence may provide one cost efficient source of BACs for filling gaps, for defining hotspots of genomic instability and for sequencing centromeric regions containing genes.

*SY Zhao (TIGR, Rockville, MD)
* M. Sekhon and J. McPherson (Washington Univ Genome Sequencing Center, St Louis, Missouri
*H. Shizuya and M. Simon (Celtech, Pasedena, CA)

Last modified: Wednesday, October 22, 2003

Base URL: www.ornl.gov/meetings/ecr2/

Site sponsored by the U.S. Department of Energy Office of Science, Office of Biological and Environmental Research, Human Genome Program