Data Mining Empowers the Generation of a Novel Class of Chromosome-specific DNA Probes

H. Zeng\textsuperscript{1,†}, H.U.G. Weier\textsuperscript{1,†}, J. Kwan\textsuperscript{1,5}, M. Wang\textsuperscript{2,6} and B. O’Brien\textsuperscript{1,3,4,*}

\textsuperscript{1} Life Sciences Division, E.O. Lawrence Berkeley National Laboratory, Berkeley, CA, USA
\textsuperscript{2} Human Genome Center, California Institute of Technology (Caltech), Pasadena, CA, USA
\textsuperscript{3} William Harvey Research Institute, Barts and The London School of Medicine, London, UK
\textsuperscript{4} Department of Anesthesiology, Berlin Heart Institute, Berlin, Germany

* Corresponding author:

B. O’Brien, M.D., FRCA
William Harvey Research Institute
Translational Medicine & Therapeutics
Charterhouse Square
London, EC1M 6BQ
United Kingdom
E-mail: benobrien@doctors.org.uk
Telephone: +44 207 882 3402
Fax: +44 207 882 3408

\textsuperscript{5} Present address: Albany Medical College, Albany, NY, USA

\textsuperscript{6} Present address: Department of Diabetes, Beckman Research Institute of the City of Hope, Duarte, CA, USA

† These authors contributed equally to this study.
Abstract

Probes that allow accurate delineation of chromosome-specific DNA sequences in interphase or metaphase cell nuclei have become important clinical tools that deliver life-saving information about the gender or chromosomal make-up of a product of conception or the probability of an embryo to implant, as well as the definition of tumor-specific genetic signatures. Often such highly specific DNA probes are proprietary in nature and have been the result of extensive probe selection and optimization procedures. We describe a novel approach that eliminates costly and time consuming probe selection and testing by applying data mining and common bioinformatics tools. Similar to a rational drug design process in which drug-protein interactions are modeled in the computer, the rational probe design described here uses a set of criteria and publicly available bioinformatics software to select the desired probe molecules from libraries comprised of hundreds of thousands of probe molecules. Examples describe the selection of DNA probes for the human X and Y chromosomes, both with unprecedented performance, but in a similar fashion, this approach can be applied to other chromosomes or species.

Keywords: X-linked diseases; Gender determination; Aneuploidy; Cytogenetics; Fluorescence in situ hybridization; Prenatal diagnosis; DNA probes; Rational probe design; Data mining; Genomics

Abbreviations: BAC – Bacterial Artificial Chromosome, BLAST – Basic Local Alignment Search Tool, FISH – Fluorescence in situ Hybridization, LINE - Long Interspersed Element, NCBI – National Center for Biotechnology Information, UCSC – University of California Santa Cruz
Introduction

Fluorescence in situ hybridization (FISH) is an established laboratory method to delineate specific nucleic acid sequences within interphase cell nuclei or metaphase spreads, which has proven indispensible for the detection of specific features in DNA and a variety of chromosomal rearrangements in cytogenetic and cancer research [1-3]. In the typical procedure, non-isotopically labeled nucleic acid probes are prepared first, and hybridized to denatured, single-stranded DNA or RNA targets, before unbound probe molecules are removed by wash steps. The bound probes can be seen in a fluorescence microscope equipped with filter combinations that match the excitation and emission properties of the fluorescent reporter molecules. Several targets can be tagged or ‘decorated’ simultaneously, and seen as individual hybridization signals, if the probes were labeled with different haptens [4-7].

The studies described below focus on the detection of DNA targets, and more specifically on detection, scoring or enumeration of specific chromosomes in interphase cell nuclei. This type of investigation finds clinical applications in a variety of fields. For example, if one of the parents of an unborn child, embryo or fetus is a known carrier of one of the many known recessive X-linked diseases, the chances of the offspring to be affected by the disease can be better prognosticated by determining its gender [8-10]. This is a relevant issue: according to Online Mendelian Inheritance in Man (OMIM), as of February 22, 2011, the current estimate of sequenced X-linked genes has reached 620 and the total including vaguely defined traits is an estimated 1138 [11].

Another common use of Y chromosome-specific DNA probes is the identification and tracking of grafted cells in sex-mismatched xenograft studies. For example, when Y
chromosome bearing white blood cells from a male donor are transplanted to a female recipient, minute amounts of grafted cell can be identified following cell harvest and hybridization due to the presence of the Y-specific signal [12-15].

Materials and Methods

All procedures involving human subject have been reviewed and approved by the UC Berkeley Institutional Review Board.

Retrieval of target sequence for the Y probe

The nucleic acid sequence used for data mining was defined in our previous studies on in vitro DNA amplification of Y chromosome-specific DNA repeat sequences. Specifically, we designed pairs of oligonucleotide primers to amplify stretches of 124 bp from the 3.6 kb pentanucleotide DNA repeat described by Nakahori et al. (DYZ1, Genbank accession number X06228) [16,17] (Table 1). Primer annealing sites were chosen that had minimal homology with the human satellite III DNA repeat consensus sequence ‘TTCCA’ [16,18-21]. Blood samples from six normal human volunteers were used to validate the Y chromosome-specific PCR assay [22,23].

Serial cell dilution experiments and artificial mixing of flow-sorted Y chromosome carrying cells in predetermined aliquots of white blood cells from female donors determined that the primer combination WYR 4 (5’-GAACCGTACGATTCCATTCCTTTTGAA-3’) – WYR 6 (5’-TTCCATTCCATTCCATTCCTTTTCTT-3’) amplifying a 248 bp DNA fragment corresponding to position 2965-3212 in Genbank accession number X06228
was sufficiently specific to detect a single male cell in the background of 1 million female cells [17,20,23]. Samples comprised entirely of female cells did not yield this product [20-23].

**Database searches**

We screened the National Center for Biotechnology Information (NCBI) human genome nucleotide DNA database for homologous sequences with one of the most widely used bioinformatics programs, Basic Local Alignment Search Tool (BLAST) [24]. The BLAST approach to rapid sequence comparison directly approximates alignments that optimize a measure of local similarity, the maximal segment pair (MSP) score. The basic algorithm is simple, robust and versatile; it can be implemented in a number of ways and applied in a variety of contexts including straightforward DNA and protein sequence database searches, motif searches, gene identification searches, and in the analysis of multiple regions of similarity in long DNA sequences [24,25].

Execution of the BLAST querying the human genome database with the 27-nucleotide (nt) sequence ‘ATTCCGTACGATTCCATTCCTTTTGAA’.from position 3089-3115 of the human Y-specific 3564 bp repeat (Genbank accession number X06228) performed at [http://ncbi.nih.gov/blast](http://ncbi.nih.gov/blast) retrieved multiple hits. Parameters were set to identify clones with a range of levels of nucleic acid homology (setting: ‘Optimize for somewhat similar sequences’, (BLASTn)).

**Retrieval of probe information for the pericentromeric region of the X chromosome**

We used the University of California Santa Cruz (UCSC) Genome Browser
GRCh37/hg19, built Februar 2009, at the University of California, Santa Cruz, accessible at http://genome.ucsc.edu/ to identify bacterial artificial chromosome (BAC) clones with high satellite DNA content [26]. The graphic user interface was set to display BAC end pairs and repeat DNA elements in the pericentromeric region of the human X chromosome, i.e., from position 58,232,531 bp to position 61,922,800 bp.

**DNA probe preparation**

The BAC DNAs were extracted from overnight cultures following an alkaline lysis protocol [27] or using a ZR BAC DNA Miniprep Kit (Zymo Research; Irvine, CA). The isolation of high molecular weight BAC DNAs was confirmed on 1% agarose gels and quantitated by Hoechst fluorometry using a Hoefer TK 100 instrument (Hoefer; South San Francisco, CA). Probe DNAs were labeled with biotin-14-dCTP (Invitrogen; Carlsbad, CA), digoxigenin-11-dUTP (Roche Molecular Systems; Indianapolis, IN), Spectrum Orange-dUTP (Abbott; Abbott Park, IL) or Spectrum Green-dUTP (Abbott) by random priming using a commercial kit (BioPrime Kit, Invitrogen)[7,28]. When incorporating fluorochrome-labeled deoxynucleoside triphosphates, the dTTP to Spectrum Orange-dUTP or Spectrum Green-dUTP ratio in the reaction was adjusted to 2:1 [28-30].

**Fluorescence in situ hybridization (FISH) and detection of bound probes**

For in situ hybridization experiments, 1 µl of labeled DNA probe, 1 µl of salmon sperm DNA (10 mg/ml, 5 Prime – 3 Prime Inc.; Boulder, CO), 1 µl water, and 7 µl of the hybridization master mix (78.6% formamide (FA) (Invitrogen), 14.3% dextran sulfate (Sigma; St. Louis, MO) in 1.43x SSC, pH 7.0) (20x SSC is 3 M sodium chloride and 300 mM tri-sodium citrate, the pH was adjusted to 7.0) were combined, thoroughly mixed
and denatured at 76°C for 10 min. The final concentration in the hybridization mixture is 55% FA, 10% dextran sulfate and 1xSSC, pH 7.0 [29, 30]. Acetic acid:methanol (1:3, vol.:vol.) fixed metaphase spreads were prepared on slides from phytohaemagglutinin-stimulated short-term cultures of peripheral blood lymphocytes from a karyotypically normal male following published procedures [31, 32]. Just prior to hybridization, slides were denatured for 4 min in 70% FA/2x SSC, pH 7.0 at 76°C, dehydrated in a 70%, 85%, 100% ethanol series, two minutes each step, and air dried. About 2.5 μl of hybridization mixture were placed on the denatured samples, covered with a 12 mm diameter circular coverslip, sealed with rubber cement and incubated for 16-20 hours at 37°C in a humidified chamber.

After the rubber cement was carefully removed from the slides, the coverslip was allowed to slip off in 2x SSC at 20°C. Post-hybridization washes were performed as described [7,28 ] and included two washes in 50% FA/2x SSC buffer at 43°C for 15 min each followed by two washes in 2x SSC at 20°C [33,34]. If non-fluorescently-labeled DNA probes were hybridized, i.e., the biotinylated or digoxigenin-labeled probes, unspecific binding sites were blocked with PNM buffer (5% nonfat dry milk, 0.1% NaN₃ in PN buffer (0.1M sodium phosphate buffer, pH 8.0, plus 0.05% NP-40)) [7,28] for 10 min, before probes were detected with either fluorescein-conjugated avidin DCS (Vector; Burlingame, CA) or anti-digoxigenin rhodamine (Roche). An additional two washes in 2x SSC were applied to remove unbounded avidin or antibodies. Finally, the slides were mounted with 4,6-diamino-2-phenylindole (DAPI) (0.1μg/ml) in antifade solution (0.1% p-phenylenediamine dihydrochloride (Sigma), 0.1x phosphate buffered saline (Invitrogen), 45 mM NaHCO₃, 82% glycerol (Sigma), pH 8.0) and coverslipped.
Image acquisition and analysis

Fluorescence microscopy was performed on a Zeiss Axioskop microscope (Zeiss, New York, NY) equipped with a filter sets for observation of Cy5/Cy5.5, Texas red/rhodamine, FITC or DAPI (84000v2 Quad, ChromaTechnology; Brattleboro, VT). Images were collected using a CCD camera (VHS Vosskuehler; Osnabrueck, Germany) and processed using Adobe Photoshop® software (Adobe Inc.; Mountain View, CA).
Results and Discussion

Chromosome Y BAC clone selection and validation

The bacterial artificial chromosome clone with the highest homology score, RP11-242E13 (Genbank accession number AC068123) contains 28 identical copies of 23 of the 27 nt query sequence in its 98295 bp insert (Fig.1). According to the NCBI database, this clone has been mapped to the long arm of the Y chromosome, band q12 (Table 1). With one exception (hit #27), the 23 bp sites (Fig.1) are spaced in more-or-less regular intervals of slightly more than 3500 bp (Table 2). The site #27 is spaced 1994 bp from the previous site of homology with the query sequence, thus indicating a single incomplete 3.5kb DNA repeat in BAC RP11-242E13. Without including site #27, the sites are spaced an average of 3566 bp, which is in excellent agreement with the published sequence for a single repeat of 3564 bp described by Nakahori et al. [16].

Validation of the Y-specific BAC clone RP11-242E13 by FISH

The result of an in situ hybridization experiment using a biotinylated DNA probe prepared from BAC clone RP11-242E13 is shown in Figure 2A. As anticipated, the 98 kb DNA probe decorated the distal long arm of the human Y chromosome suggesting that there are hundreds or even thousands of copies of the complementary DNA target sequence [35] (Figure 2A). Interestingly, even in this simple hybridization experiment that did not include any type of blocking DNA, the probe lit up just the chromosome Y, band q12. No cross-hybridization with other chromosomes that are known to carry satellite III-type heterochromatin such as human chromosome 9 [19] was found. In summary, our findings suggest that BAC clone RP11-242E13 carries multiple copies of the 3.6kb DNA repeat which hybridize exclusively to the human Y chromosome.
BAC clones from the pericentromeric region of the X chromosome

The data mining effort searching for a large insert, recombinant DNA clone that hybridizes specifically to the heterochromatic DNA tandem repeats in the pericentromeric region of the human X chromosome was guided by our prior observation that chromosome-specific alpha satellite DNA repeats can be identified in BAC clones through database searches [36]. We had noticed that alpha satellite DNA repeats which typically appear as chromosome-specific, high order tandem repeats in the pericentromeric region of human autosomes [26, 37] can also be found dispersed in low copy numbers in some significant distance from the centromeric regions and in vitro DNA amplification can be used to enrich the alpha satellite sequences [36].

The graphical user interface of the UC Santa Cruz Genome Browser suggested that several clones from the RP11 BAC library map into the X chromosome-specific pericentromere and contain alpha satellite DNA repeats (Figure 3). We decided to test two BAC clones, one that is essentially free of interspersed DNA repeats such as short or long interspersed element (LINE) sequences (RP11-294C12) mapping to chromosome X, band q11.2 (Table1) and a second clone (RP11-348G24), which contains a few interspersed DNA repeats and maps to chromosome X, band p11.1 (Figure 3). After retrieval from the freezer, both clones grew in Luria Bertoni broth at approximately the same rate. The DNA yields from 5 ml overnight cultures were similar, but the labeled DNAs showed drastically different hybridization patterns. The results are shown in Figures 2B and 2C.

The digoxigenin-labeled probe prepared from BAC clone RP11-348G24 containing the interspersed repeats gave a strong signal on the X chromosome, but also
unacceptably high levels of cross-hybridization with other autosomes (Figure 2B). Many interphase nuclei exhibited such high levels of cross-hybridization that it was not possible to delineate the X chromosome target. In contrast, hybridization of a biotinylated DNA probe prepared from BAC clone RP11-294C12 resulted in signals that localized exclusively to the pericentromeric region of the human X chromosome (Figure 2C).

To demonstrate that the highly specific DNA probes for chromosomes X and Y can be combined in dual-color multiplex hybridization experiments [38-40], we labeled the Y chromosome-specific DNA from BAC clone RP11-242E13 with Spectrum Green-dUTP (green fluorescence) and the X chromosome-specific DNA from BAC clone RP11-294C12 with Spectrum Orange-dUTP (red fluorescence). Hybridization of this probe mixture gave strong, specific signals in metaphase as well as interphase cells (Figure 2D).

**Probe generation and testing rates**

Turnaround times for generation and validation of DNA probes prepared from BAC’s are at least one order of magnitude shorter than those observed with yeast artificial chromosome clones as large insert human genomic DNA probe template [41-43].
Conclusions

This study was undertaken to specifically seek answers to three questions:

1. Can researchers with limited access to computational capabilities data mine and take advantage of the huge resources generated by the International Human Genome Project?

2. How can one identify recombinant, large insert DNA clones with features that make them ideal hybridization probes? and

3. What constitutes an inexpensive and rapid approach to validate the selected clones?

We believe that all three questions have been answered in this study. First, we have been able to demonstrate that simple, publicly available bioinformatics tools such as BLAST searches running on a simple desktop computer allow the operator to data mine and extract the desired information from publicly available archives. Secondly, we began to define rules to predict DNA probe properties based on DNA sequence analysis. Lastly, we were able to show that in a matter of days, chromosome-specific DNA probes can be defined, prepared and validated by FISH.

The importance of probe specificity can not be overemphasized. While recombinant DNA clones carrying large chunks of the human genome such as the BAC or P1 clones (39, 40,43-44) are easy to propagate, the presence of non-chromosome-specific interspersed DNA repeats can lead to major impediments in FISH-based interphase cell analysis [36]. As shown above, careful probe selection based on mining of the publicly accessible databases may circumvent some of these problems.
The simple, yet efficient data mining approach presented in this paper is our first step towards creation of an ‘in silico’ DNA probe set that may find extensive use in the analysis of NexGen, deep sequencing data [45,46]. Thus, this new class of ‘in silico probes’ has little or nothing in common with the conventional recombinant DNA probes [1,2,4,19,47], many of which have not even been sequenced, and it is more akin to the well known ‘in silico PCR’ [48]. Here, we describe a software-based interactive approach in which the user checks the results of the data mining operation and validates the BAC clones via FISH. In our lab setting, the approach allowed us to select and validate a few clones (<10) per week.

In closing, it’s important to mention that polymorphisms or centromeric heteromorphisms might present serious problems in the FISH-based analysis of cells from affected individuals using DNA repeat probes [49] regardless whether such an analysis is performed by convention FISH, Spectral Imaging or multicolor FISH [6,34,50]. The before-mentioned deep sequencing in combination with ‘in silico’ copy number analysis might prove to be a reasonable alternative approach.
Acknowledgements

This work was supported in parts by a grant from the Director, Office of Energy Research, Office of Health and Environmental Research, U.S. Department of Energy, under contract DE-AC02-05CH11231, the Leonard Rosenman Fund and NIH grants HD45736, CA123370, CA132815, CA132815-02S1 and CA136685 (to HUG).

Disclaimer

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof, or The Regents of the University of California.
### Table 1. BAC clones for human chromosomes X and Y

<table>
<thead>
<tr>
<th>Clone</th>
<th>Band</th>
<th>Start (bp)</th>
<th>End (bp)</th>
<th>Insert size (bp)</th>
<th>Database</th>
<th>Accession numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>RP11-242E13</td>
<td>Yq12</td>
<td>58819362</td>
<td>58917656</td>
<td>98295</td>
<td>NCBI</td>
<td>AC068123</td>
</tr>
<tr>
<td>RP11-294C12</td>
<td>Xq11.1</td>
<td>61696702</td>
<td>61722561</td>
<td>25860</td>
<td>UCSC</td>
<td>AQ508215,AQ508213</td>
</tr>
<tr>
<td>RP11-348G24</td>
<td>Xp11.1</td>
<td>58356061</td>
<td>58564667</td>
<td>208607</td>
<td>UCSC</td>
<td>AQ528470,AQ528473</td>
</tr>
</tbody>
</table>

### Table 2. Location of 28 BLASTn hits with the 27 nt query sequence in BAC RP11-242E13

<table>
<thead>
<tr>
<th>Hit</th>
<th>Position in RP11-242E13</th>
<th>Distance to previous hit (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3112</td>
<td>Not applicable.</td>
</tr>
<tr>
<td>2</td>
<td>6670</td>
<td>3558</td>
</tr>
<tr>
<td>3</td>
<td>10234</td>
<td>3564</td>
</tr>
<tr>
<td>4</td>
<td>13823</td>
<td>3589</td>
</tr>
<tr>
<td>5</td>
<td>17357</td>
<td>3534</td>
</tr>
<tr>
<td>6</td>
<td>20831</td>
<td>3474</td>
</tr>
<tr>
<td>7</td>
<td>24410</td>
<td>3579</td>
</tr>
<tr>
<td>8</td>
<td>27985</td>
<td>3575</td>
</tr>
<tr>
<td>9</td>
<td>31579</td>
<td>3594</td>
</tr>
<tr>
<td>10</td>
<td>35143</td>
<td>3564</td>
</tr>
<tr>
<td>11</td>
<td>38702</td>
<td>3559</td>
</tr>
<tr>
<td>12</td>
<td>42276</td>
<td>3574</td>
</tr>
<tr>
<td>13</td>
<td>45858</td>
<td>3582</td>
</tr>
<tr>
<td>14</td>
<td>49452</td>
<td>3594</td>
</tr>
<tr>
<td>15</td>
<td>53026</td>
<td>3574</td>
</tr>
<tr>
<td>16</td>
<td>56570</td>
<td>3544</td>
</tr>
<tr>
<td>17</td>
<td>60159</td>
<td>3589</td>
</tr>
<tr>
<td>18</td>
<td>63713</td>
<td>3554</td>
</tr>
<tr>
<td>19</td>
<td>67242</td>
<td>3529</td>
</tr>
<tr>
<td>20</td>
<td>70826</td>
<td>3584</td>
</tr>
<tr>
<td>21</td>
<td>74400</td>
<td>3574</td>
</tr>
<tr>
<td>22</td>
<td>77984</td>
<td>3584</td>
</tr>
<tr>
<td>23</td>
<td>81543</td>
<td>3559</td>
</tr>
<tr>
<td>24</td>
<td>85107</td>
<td>3564</td>
</tr>
<tr>
<td>25</td>
<td>88680</td>
<td>3573</td>
</tr>
<tr>
<td>26</td>
<td>92264</td>
<td>3584</td>
</tr>
<tr>
<td>27</td>
<td>94258</td>
<td>1994</td>
</tr>
<tr>
<td>28</td>
<td>97817</td>
<td>3559</td>
</tr>
</tbody>
</table>
Figure Legends

Figure 1: The results of a BLASTn query using a human Y chromosome-specific 27nt DNA sequence. Shown are the first seven hits of the alignment search, which in total resulted in 28 perfect matches for 23/27 nt query sequence.

Figure 2: Clone validation by FISH on to normal male human metaphase spreads and interphase cells. (A) In situ hybridization of a biotinylated probe prepared from BAC RP11-242E13 exhibited strong signals on the long arm the human Y chromosome (arrow). The insert shows an enlarged picture of the Y chromosome with the DAPI picture to the left, the FITC picture in the center and a superposition of both on the right. (B) A digoxigenin-labeled DNA probe prepared from BAC RP11-348G24 resulted in specific signals on the X chromosome (arrow) in the presence of a significant amount of cross-hybridization to other chromosomes. Please note the high level of cross-hybridization in the interphase nucleus shown in the lower right corner. (C) In contrast to the result shown in (B), a probe prepared from BAC RP11-294C12 bound virtually exclusively to the centromeric heterochromatin of the human X chromosome (arrow). (D) Multiplex analysis of human gonosomes was achieved by combining differently labeled DNA probes in a single hybridization experiment. In this experiment, the Y chromosomal target appears in green (arrow), while the X-chromosomal centromeric repeat DNA is shown decorated in red (arrow).

Figure 3: Selection of an X chromosome-specific BAC clone using the UCSC Genome Browser. On top, the figure depicts a X chromosome ideogram with the selected pericentromeric region 58,232,531-61,922,800 bp highlighted by a red square.
The lower part of the picture shows BAC end alignments within this interval as well as the locations of DNA repeat sequences (bottom). Clones and DNA repeat sequences of interest are indicated by colored squares and circles, respectively. Please note the extensive area of central, unchartered DNA representing the centromere.
References


(D10S170/H4) and deletion mapping in the thyroid cancer cell line TPC-1. Cytogenet Cell Genet 75: 254-257.


