© 1992 Oxford University Press Nucleic Acids Research, Vol. 20, No. 19 5173–5179

# The selective isolation of novel cDNAs encoded by the regions surrounding the human interleukin 4 and 5 genes

John G.Morgan, Gregory M.Dolganov, Sabrina E.Robbins, Linda M.Hinton and Michael Lovett[1,*]
Department of Molecular Genetics, Genelabs Incorporated, 505 Penobscot Drive, Redwood City, CA 94063 and [1]Department of Biochemistry and McDermott Center, University of Texas Southwestern Medical Center at Dallas, 5323 Harry Hines Boulevard, Dallas, TX 75235, USA

Received April 28, 1992; Revised and Accepted August 29, 1992

## ABSTRACT

**We have developed modifications to direct cDNA selection that allow the rapid and reproducible isolation of low abundance cDNAs encoded by large genomic clones. Biotinylated, cloned genomic DNAs are hybridized in solution with amplifiable cDNAs. The genomic clones and attached cDNAs are captured on streptavidin coated magnetic beads, the cDNAs are eluted and amplified. We have applied this protocol to a 425kb YAC that contains the human IL4 and IL5 genes. After two cycles of enrichment twenty-four cDNAs were evaluated, all of which were homologous to the YAC. DNA sequencing revealed that nine cDNAs were 100% homologous to the interferon regulatory factor 1 (IRF1) gene. Six clones were 70% homologous to the murine P600 gene, which is coexpressed with IL4 and IL5 in mouse Th2 cells. The nine remaining clones were unique within the sequence databases and were non redundant. All of the selected cDNAs were initially present at very low abundance and were enriched by as much as 100,000-fold in two cycles of enrichment. This modified selection technique should be readily applicable to the isolation of many candidate disease loci as well as the derivation of detailed transcription maps across large genomic regions.**

## INTRODUCTION

The isolation of rare human transcripts as cDNA clones can be fraught with problems due to cDNA abundance and sequence complexity considerations. These can be summarized by the following approximations: it has been estimated that about 10,000 genes are expressed in a given human cell type at levels that may be as high as 200,000 mRNA molecules per cell or less than one mRNA molecule per cell, with approximately one third of the genes being expressed at 1–10 molecules per cell (1,2). The complexity of sequences is thus within manageable limits, but the variation in abundance classes dictates that at least several hundred thousand clones must be screened to have a reasonable chance of finding a particular low abundance transcript (3). This problem is further exacerbated when a complex tissue is used as a source for cDNAs. In this case, numerous cell types exist, each containing widely varying transcript abundance classes (4,5). It is therefore unlikely that a conventional cDNA library of approximately one million cDNA clones will adequately represent all the transcripts that are expressed in such a tissue, since it will probably not contain the lower abundance cDNAs. One approach that can be taken to surmount these obstacles is to attempt an approximate normalization of cDNA abundance classes using cDNA reassociation methods, so that fewer clones need be constructed and screened (6,7). An alternative strategy is to devise methods for rapidly screening very large numbers of primary cDNAs. We (8), and others (9), have recently described hybridization selection schemes that have this capability and may also result in some level of abundance normalization (10). These applications are described in greater detail below. However, the immediate application of these techniques is for the detection of coding regions within large regions of genomic DNA.

Identifying coding sequences within large genomic clones adds additional complications to those already mentioned above. On average, only 3% of the genomic DNA will be homologous to a cDNA (1,2) and this homology may be very patchy due to the presence of introns. In addition, the repetitive elements within the genomic DNA are also present in a substantial proportion of the target cDNA population (11). As a result, the use of very large genomic clones in conventional screening schemes usually results in very poor signal to noise ratios and consequently is not very reproducible (8,12,13). Thus, by current methodologies, the task of identifying coding sequences is significantly difficult even when the region of interest is only 20–40kb, and becomes truly daunting when the region is a megabase in length.

Several techniques have been devised that are targeted at the identification of coding regions in human genomic DNA and seek to address the aforementioned problems. These methods can be roughly divided into two types: methods that are based upon the transcription of a genomic region either in a somatic cell hybrid (14,15), or within an artificial construct (16,17), and the subsequent detection of conserved transcriptional or processing sequences; and hybridization based schemes (8,9,18,19). Among the latter group, we have recently reported a method, direct selection (7,20,21), that is based upon hybridizing an entire

[*] To whom correspondence should be addressed

population of cDNAs to a genomic clone or genomic contig. These genomic clones are either in the form of yeast artificial chromosome clones (22) or are within cosmid contigs. The hybridizing cDNAs are eluted and then amplified using the polymerase chain reaction (23). Our first report on this method (8) summarized the results we obtained with a 550kb YAC clone that contained the human erythropoietin (EPO) gene. These data indicated that one round of hybridization selection can enrich a low abundance EPO cDNA (1 clone in $10^6$) by approximately 1,000-fold and that new low abundance cDNAs can be rapidly isolated and identified by this method. We report here modifications that we have made to enhance this protocol in light of our experience with the original methodologies, and the application of this adapted protocol to the detection of novel, low abundance cDNAs that are encoded by a 425kb YAC from human chromosome 5 that contains the interleukin 4 and interleukin 5 genes. The adapted protocol is shown schematically in Figure 1 and is described in detail below.

## MATERIALS AND METHODS

### Preparation of cDNA

Peripheral blood lymphocytes (PBLs) were separated from whole blood by density centrifugation over Ficoll (Sigma, St Louis, MO). Cultures were maintained in complete Iscove's modified Dulbecco's medium (Gibco, Grand Island, NY). Cells (two flasks for each time point) were stimulated separately with Con A (2.5 μg/ml)/PHA (2.5 μg/ml) and PMA (1 ng/ml)/calcium ionophore A23187 (20 μM) for 1, 6, 12, and 24 hours. RNA was isolated (24) from $10^8$ cells for each time point and the two RNA samples for each time point representing both activation pathways were mixed. Polyadenylated RNA was isolated by oligo(dT)cellulose chromatography and double stranded cDNA was synthesized from 5 μg of this RNA using an Amersham cDNA synthesis kit. Double stranded cDNA was digested with *Mbo* I, phenol/chloroform extracted and ethanol precipitated. The cDNA was resuspended and ligated to a linker-adaptor oligonucleotide (Mbo linker I, 5′GATCGAATTCACTCGAGCATCAGG3′ 3′CTTAAGTGAGCTCGTAGTCC5′). Unligated linker-adaptor was removed by passage over a primerase column (Stratagene, La Jolla, CA). Note that the Mbo linker I contains an *Eco* RI site for subsequent cloning steps. For the derivation of ribosomal probes, 5μg of nonpolyadenylated RNA was converted into single stranded cDNA by random priming and this was subsequently radiolabelled for counterscreening selected clones.

### Preparation of genomic DNAs

The 425kb yeast artificial chromosome (YAC) clone (A94G6) that contains both the IL4 and IL5 genes was isolated from the Washington University human YAC library using PCR-based screening (25). This clone was subsequently found to not be detectably chimeric as determined by fluorescence in situ hybridization (FISH) to metaphase chromosome spreads (D. Saltman, personal communication). Yeast chromosomes were prepared and electrophoresed on a contour-clamped homogeneous electric field gel and the 425kb YAC containing the IL-4 and IL-5 genes (YAC 4/5) was excised and purified using Geneclean II (Bio 101, La Jolla, CA). YAC 4/5 DNA was digested with *Mbo* I and ligated to a second oligonucleotide linker/adaptor (Mbo linker II; 5′GATCTCGACGAATTCGTGAGACCA3′ 3′AGCTGCTTAAGCACTCTGGT5′). Excess linkers were separated as described above. Approximately 0.1 to 1 ng of the YAC DNA and cDNA samples were then separately amplified by PCR (see conditions below) using the relevant primer for each linker/adaptor. In each case, the smaller oligonucleotide constituted the primer and each primer was only capable of priming its cognate linker. In the case of the YAC 4/5 DNA, a 5′ biotinylated primer was used in the amplification. Free primer was removed by gel purification of the PCR products. Approximately 1μg of cDNA and 1μg of YAC DNA were prepared by this method with an average length of approximately 0.8kb. The DNAs were separately ethanol precipitated and resuspended in 100μl each of 10mM Tris HCl pH 8.0, 1mM EDTA (TE). Cosmid clones were derived by screening and walking within an arrayed chromosome 5-specific cosmid library using IL4 and IL5 probes and cosmid end probes (G.M.Dolganov unpublished data). The cosmid library was constructed and arrayed by Dr. Larry Deaven at the Los Alamos National Laboratories. Phage genomic clones were isolated from a total human genomic DNA library (Stratagene) by conventional methods (3) using IL4 and IL5 cDNA probes.

### Blocking repeats within the cDNAs

In the experiments described here, we used COT1 DNA (Gibco BRL) as a blocking agent. COT1 DNA and the uncloned starting cDNA were mixed in a one to one (w/w) ratio, denatured and annealed at 60°C to a Cot value of 20 moles nucleotide/liter.seconds in 0.12M sodium phosphate pH7 [80μg/ml for 1 hour is approximately a Cot of 1 (26)].

### Hybridization

The blocked cDNA (1μg) was hybridized at 60°C in 0.12M sodium phosphate pH7, to the biotinylated YAC DNA in solution under paraffin oil, to a calculated Cot of 120 moles nucleotide/liter.seconds, at a cDNA to YAC DNA ratio of 10:1 (w/w). The YAC DNA plus bound cDNAs were captured on streptavidin coated magnetic beads (Dynal, Oslo). Binding was performed using 100μl of beads in TE plus 1M NaCl for 15 minutes at room temperature. Beads were removed from the solution using a magnet for 1 minute. The beads were then subjected to the following washes in a volume of 100μl: 0.1×SSC, 0.1% SDS two washes at room temperature and three washes at 65°C, each wash for 15 minutes. cDNAs were eluted with 50μl of 50mM NaOH for ten minutes at room temperature, followed by neutralization with 50μl of 1M Tris HCl pH 7.0. The eluate was then desalted by chromatography over a Sephadex G-50 (Pharmacia) spun column. For secondary cycles of enrichment, approximately 1μg of amplified eluted cDNA was recycled through the above process, including the repeat blocking step.

### Polymerase chain reactions

PCR amplifications were performed for 30 cycles. Each 100μl reaction contained between 1 and 10μl of eluate (or an appropriate volume of the starting cDNA or YAC) and 5μl of a 20μM primer stock, with all other components being standard Perkin Elmer Cetus stocks. Each cycle consisted of denaturation for 1 minute at 94°C, annealing at 55°C for 1 minute, and extension for 1 min at 72°C. PCR primers for IRF1 were; 5′CTCTAGGCAAGCAGGACCT3′ 5′TCATACCAAGGCGCTCACA3′. PCR primers for IL4 were; 5′GAGCCTGAGATCAACACATG3′ 5′CAGCTCGAACACTTTGAATA3′ and PCR primers for the human P600 homolog were; 5′ATGGCGTTTGTTGACCAC3′ 5′ACAGTACATGCCAGCTGGTCA3′.

### Cloning PCR products

PCR products were digested with *Eco* RI and were purified on a primerase column. The cDNAs (200 ng) were ligated into *Eco* RI digested, phosphatased λgt10 arms (2 μg), in a final volume of 6μl. Phage were packaged using a commercial packaging kit (Stratagene). Separate libraries of cDNAs were built from the starting cDNAs, the cDNAs obtained from the primary elution and cDNAs from the secondary cycle of enrichment.

### Evaluating enrichments and assigning cDNAs to YAC4/5

DNAs (purified YAC DNAs, genomic DNAs, cosmid DNAs, phage DNAs or cDNAs) were blotted, as recommended by the manufacturer, onto Hybond$^{TM}$-N$^+$ nucleic acid transfer membrane (Amersham) using a Minifold II slot blot system (Schleicher and Schuell, Keene, N.H.). In some cases, individual cDNAs were also hybridized to pulsed field gel blots of YACs to confirm their assignment to YAC4/5, and/or were substrates for the design of PCR primers. These primers were used to confirm the presence of the sequences within YAC4/5 by PCR (see results section for details). Screening of cDNA libraries was performed using duplicate nitrocellulose plaque lifts (3). All DNA probes were radiolabeled with $^{32}$P using a random-priming kit (Boehringer Mannheim). Prehybridization, hybridization and washing conditions were conducted under standard conditions (3).

### DNA Sequencing of PCR products

Purified PCR products were directly sequenced using the Promega fmol$^{TM}$ DNA Sequencing System.

## RESULTS

### The modified selection protocol

The direct selection scheme, as we originally reported it, involved using purified DNA from a genomic clone immobilized on a filter support, and hybridization with an entire library of cDNA inserts (8). Approximately 3–10% of cDNAs contain repetitive elements (11) and these must therefore be either blocked or eliminated from the hybridization prior to the selection step or they result in insurmountable background problems. cDNAs in our original protocol were blocked with total genomic DNA and were then hybridized to the genomic DNA. Specific cDNAs were eluted after post hybridization washing. The eluted cDNAs were then amplified and could be either cloned using restriction sites in the end linker, or used for further cycles of hybridization selection.

Figure 1 illustrates our recent modifications to this procedure. The kinetics of filter hybridizations are difficult to accurately predict in comparison to solution hybridizations (26). We therefore opted to incorporate a biotin/streptavidin capture system and a solution hybridization to better control the hybridization parameters. The cloned genomic DNA can either be labelled with biotin by conventional methods or as in this report, be amplified in the PCR using a biotinylated primer. The addition of linkers to the YAC DNA has the substantial advantage that only small quantities of the YAC need be initially purified. The YAC DNA plus bound cDNAs is then captured after the hybridization using streptavidin coated beads. The beads are extensively washed and bound cDNAs are eluted, amplified, and recycled. Repeats within the cDNA are in this case, blocked with COT1 DNA (Gibco BRL). This DNA is highly enriched in intermediate repeats and has the advantage over total genomic DNA that it does not drastically increase the overall sequence complexity of the mixture in the hybridization. Using cDNA inserts that have been amplified from the vector with vector primers results in a bias against large cDNAs if the original cDNA was full length rather than random primed. We therefore chose to fragment our full length cDNAs into smaller pieces and ligate them to a linker for subsequent amplification (27). As is noted below, this also has some disadvantages. One of the most important considerations in assessing the clones derived by this protocol was the knowledge gained from an extensive evaluation of several hundred clones from our EPO selections (8 and Hinton, Robbins, Morgan and Lovett in preparation). These screens revealed that over 70% of the insert-containing clones in our previous selections were ribosomal in origin. Ribosomal clones represent a common contaminant in most cDNA libraries. We believe these cDNAs are probably selected because of small quantities of contaminating yeast ribosomal DNA in the gel purified YAC DNA preparations. However, these contaminating clones can be largely removed merely by preblocking with ribosomal DNA or can be completely eliminated by counterscreening with a ribosomal probe after the selection. We employed the latter strategy in the experiments described here.

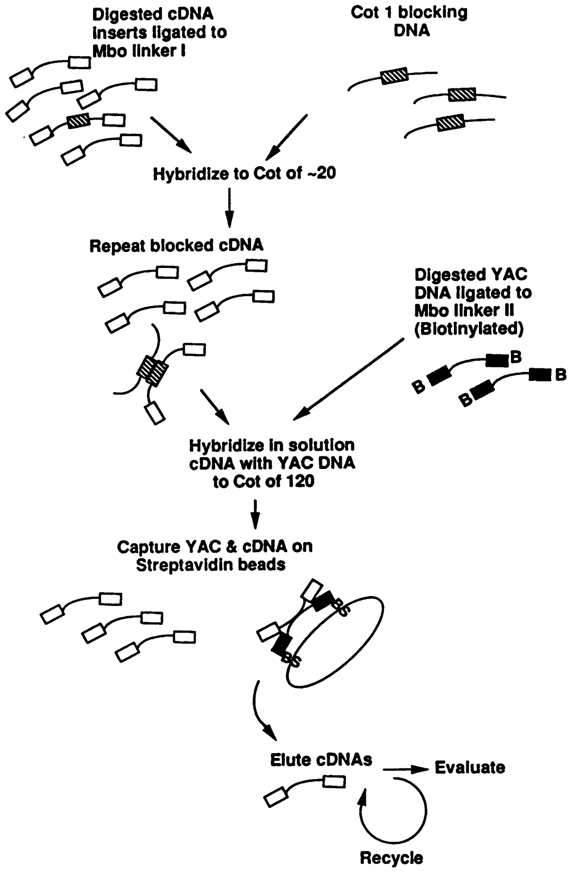
Digested cDNA inserts ligated to Mbo linker I
Cot 1 blocking DNA
Hybridize to Cot of ~20
Repeat blocked cDNA
Digested YAC DNA ligated to Mbo linker II (Biotinylated)
B
B
B
B
Hybridize in solution cDNA with YAC DNA to Cot of 120
Capture YAC & cDNA on Streptavidin beads
S
S
Elute cDNAs
Evaluate
Recycle

**Figure 1.** A biotin/streptavidin capture system for direct selection (see text for details).

### The YAC4/5 selection

Deriving a physical map of the regions surrounding the IL4 and IL5 genes within human chromosome 5q23–31 has been of considerable interest to many groups (28–34) because of the close proximity of these loci to each other (29,32–34), to the IL3 and GMCSF genes (29,33,34) and to the chromosomal breakpoints that delineate the region involved in the 5qabnormalities (30,31). We have been interested in building YAC and cosmid contigs throughout this region (35) and in the course of this work isolated a YAC that contains both IL4 and IL5. This YAC, designated YAC4/5, is approximately 425kb in length and prior to this study, no other genes had been localized to it. However, in the course of this work the interferon regulatory factor 1 (IRF1) gene was localized to the 5q23−q31 region (36) and then further localized by radiation hybrid mapping close to the IL4 and IL5 genes (34). Our subsequent PCR-based assessment of YAC4/5 indicated that the IRF1 gene was contained within it and, as is described below, we discovered simultaneous with this observation, that a substantial proportion of our selected cDNAs were derived from the IRF1 gene.

The starting cDNA for these studies was constructed from activated peripheral blood mononuclear cells, a source that is known to express the two reporter genes, IL4 and IL5 (37). Pooled samples from various timepoints and using different activating mitogens were used (see Materials and Methods). The cDNA was intentionally not cloned so as to retain the large sequence complexity of the starting source. The double stranded cDNA was digested and ligated to a linker to render it amplifiable. This was then hybridized in solution with COT1 DNA to block the repeats within the cDNA (shown as cross hatched boxes in Figure 1). The purified YAC DNA was amplified with a biotinylated primer. The repeat blocked cDNA was then hybridized to the biotinylated YAC DNA to an intermediate Cot value. The YAC DNA plus attached cDNAs was captured on streptavidin coated magnetic beads (Dynabeads), washed extensively and an aliquot of the eluted cDNAs was amplified and recycled through a second cycle of hybridization selection. The selection of the positive control cDNAs (IL4 and IL5) was monitored at the end of each selection by hybridization of the two reporter genes to slot blots of the starting, primary selection and secondary selection cDNAs (data not shown). Surprisingly, the IL4 cDNA was enriched but the IL5 gene was depleted. Analysis of the IL5 cDNA sequence revealed that it only contains one *Mbo*I site and was thus not capable of being ligated to two linker adaptors (38), and hence could not be amplified. This trivial technical problem has been resolved in more recent selections (see discussion). An aliquot of the original starting cDNA was cloned and the amplified selected cDNAs were also separately cloned after the primary and the secondary selections. Plaque lifts from these libraries were screened with ribosomal probes and with an IL4 cDNA (data not shown). Approximately 60% of the clones were ribosomal in origin and IL4 had been enriched by approximately 2000-fold (see Table 1).

**Table 1.** Summarized enrichment and localization data of selected clones

| cDNA | Starting % | Primary % | Secondary | Location |
|---|---|---|---|---|
| IL4 | 0.0025 | 3.5 | 4.0 | IL4 coding |
| #20 (IRF1) | 0.00022 | 3.0 | 33.0 | >100kb outside |
| #50 (P600) | <0.00022 | 1.8 | 25.0 | upstream IL4 |
| #8 | 0.00067 | 1.0 | 4.0 | upstream IL5 |
| #32 | 0.00022 | 0.3 | 4.0 | upstream IL4 |
| #39 | 0.00045 | 0.25 | 4.0 | downstream L4 |
| #40 | 0.00045 | 2.2 | 4.0 | >100kb outside |

cDNA refers to the selected clones from the secondary library. Starting refers to frequency per 450,000 clones in the starting cDNA library, <0.00022% indicates that no positives were detected in a screen of 450,000 clones from the starting library. However, the presence of all of the selected cDNAs within this source was separately verified by Southern blotting total starting cDNA from the library and hybridization with each cDNA (data not shown). Primary refers to frequency per 5,000 clones in the primary selected library and secondary to final frequency in the secondary library. All numbers are corrected for the frequency of ribosomal clones which comprised less than 1% of the starting library and constituted 60% of the primary and secondary selected libraries. Numbers for IRF1 constitute the sum of the frequencies for each of the three separate cDNA fragments from this locus. Numbers for P600 constitute the sum for the two identifiable fragments from that locus. Location refers to the approximate location within a partial contig of IL4 and IL5 (see Figure 2 and text for details).

## Isolation of new cDNAs and their approximate localization

Twenty-four clones that were neither IL4 or ribosomal were randomly picked from the secondary selected library and were assessed to determine whether they were encoded by the YAC. The average length of these inserts was 0.8kb with a range from 0.2kb to 1.8kb. Each cDNA insert was radiolabelled and hybridized to a slot blot that contained purified YAC4/5 DNA, an unrelated 550kb purified YAC DNA from chromosome 7 (YACEPO), total yeast genomic DNA, a selection of cosmid and phage genomic DNA clones from the regions surrounding IL4 and IL5, and controls. Figure 2 shows a representative set of slot blots from these experiments and indicates that all of the selected clones are homologous to YAC4/5 and that at least some of these cDNAs also hybridize to the partial IL4/IL5 cosmid contig that is included on the slot blots. A control cDNA from chromosome 7 (GNB2) that is encoded by YACEPO (8) only detects that slot, verifying the selectivity of the asssay. In addition, the localization of individual cDNAs was confirmed by hybridization to pulsed field gel blots of YAC4/5 (data not shown) and by PCR-based assays (see below). Because the slot blots included cosmid and phage clones that cover part but not all of the DNA contained within the YAC, the data from Figure 2 also enables some conclusions to be drawn about the approximate location of the regions of cDNA homology relative to IL4 and IL5. Thus, clone #50 is homologous to a region upstream from IL4 (relative to the direction of transcription), and clone #20 is located outside of the regions immediately flanking the IL4 and IL5 genes. Two of the clones used in this experiment yield lower intensity signals; clone #8 hybridizes weakly to all of the slots but most strongly to clones upstream form IL5; clone #39 hybridizes strongly to cosmid and phage clones surrounding IL4 but only weakly to the YAC4/5 slot. These two cDNA are nevertheless both single copy and specifically homologous to the YAC as determined by the confirmatory assays discussed below. This data is summarized as part of Table 1.

## Identifying known and novel cDNAs

Approximately 200bp of DNA sequence from each of the twenty-four cDNAs was next determined and compared with the accessible sequence databases. This analysis revealed that nine of the clones were 100% homologous to parts of the interferon regulatory factor 1 gene [IRF1, (39)]. The nine clones comprised three separate *Mbo* I fragments from the IRF1 cDNA. One of these clones is #20 shown in Figure 2 and this places the IRF1 locus outside of the immediate vicinity of the IL4 and IL5 genes. Six of the clones, including #50 mentioned above and shown in Figure 2, were 70% homologous to the murine P600 gene, a gene that is coexpressed with IL4 and IL5 in a murine T helper cell subtype known as Th2 cells (40). It would appear that the clones we have isolated are part of the previously undescribed human P600 locus and that this gene is located very close to the

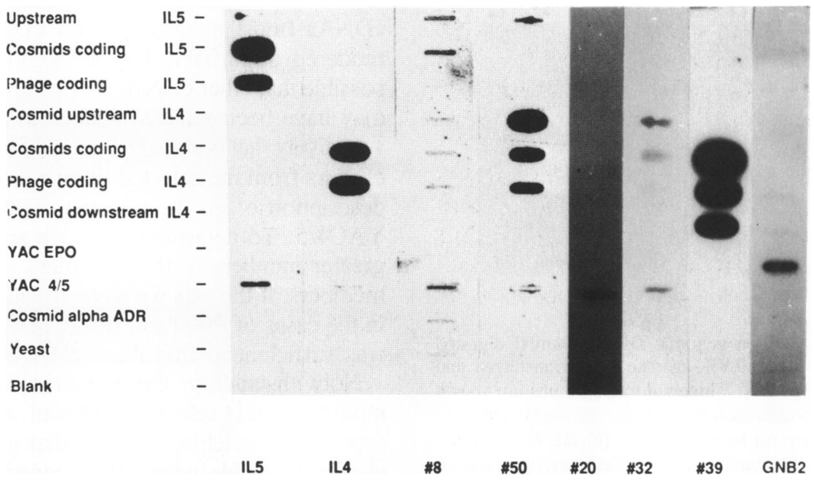

**Figure 2.** Hybridization of selected cDNA clones to YAC4/5, cosmids and phages covering the IL4 and IL5 loci. A series of slot blots are shown, each comprised of twelve slots. At the left side the DNA contained in each slot is shown and its location within a partial IL4/IL5 genomic contig. YAC EPO is a control YAC from chromosome 7, yeast refers to total AB1380 genomic DNA, cosmid alpha ADR is a cosmid clone that contains a human alpha adrenergic gene (located distal to the IL4/IL5 loci on human chromosome 5), and blank is a buffer alone control. The probes used for hybridization to each series of slots are shown at the bottom. The GNB2 probe is located on the YACEPO control DNA (8).

IL4 locus (Figure 2 and Table 1). The six P600 homologous clones comprised two distinct *Mbo* I fragments and a comparison of the conceptual translation products of this sequence with the mouse P600 protein is shown in Figure 3. The two amino acid sequences are approximately 64% homologous and the DNA sequences are approximately 70% homologous. However, this number should be interpreted with caution since the DNA sequence was derived from a cDNA that had undergone a total of 60 cycles of PCR in two selection cycles, and may therefore have a substantial error rate. The nine additional cDNAs are unique within the accessible DNA sequence databases (Genbank, EMBL), and are non-redundant. The aforementioned cosmid and phage hybridization data indicates that at least four of the unique nine cDNAs are encoded by a region upstream of the IL4 gene (clone #32 in Figure 2 and Table 1 is an example of one of these). These four may be separate parts of one larger cDNA that is encoded by this region and experiments are underway to isolate full length clones for all of the selected cDNAs to determine if this is the case.

## Confirmatory assays to determine localization to YAC4/5

To independently confirm the localization of the selected cDNAs to YAC4/5, additional assay systems were employed. Figure 4 shows an example of a PCR-based assay in which primers designed to the IRF1 cDNA, the IL4 cDNA and the human P600 cDNA were used in PCR reactions on YAC4/5 DNA or on yeast DNA alone. In the case of the IL4 and IRF1 PCR products, the expected genomic PCR products were observed (1.4kb and 0.6kb respectively). The P600 cDNA sequence predicted an extension product of 0.19kb and the observed product has a length of 1.3kb (Figure 4 track 2) most probably reflecting the presence of an intron within the coding sequence over which we designed our primers. The second confirmatory assay consisted of hybridizing individual cDNAs to a panel of genomic DNAs to verify copy number and absence of any repetitive elements. Figure 5 shows representative data from such an analysis for clones #8, #50 and #32. The panel of *Bam*HI digested genomic DNAs consisted of DNA from a somatic cell hybrid that contains one copy of human chromosome 5 in a hamster background (HHW105, track a), DNA from the hamster parental cell line, human genomic DNA and yeast genomic DNA. In all but one case the cDNAs hybridized to single, chromosome 5-specific genomic DNA fragments (tracks a and c). The exception is clone #32 shown in the right hand panel of Figure 5. In this case, two *Bam*HI fragments are detected in human genomic DNA and two in the

```
Mouse P600  MALWVTAVLALACLGGLAAPGPVPRSVSLPLTLLELIEELSNITQDQ-TPLCN
            :::  .:. .:::::::..::::::.:..:.    :::: ::: :::: . : .::::
  Clone 50  MALLLTTVIALACLGGFDSPGPVPPSTALR----ELIEELSNITQTQKAPLCN
```

**Figure 3.** The conceptual translation products from the putative human P600 cDNA compared to the published mouse amino acid sequence. The one letter amino acid code is shown and the sequences are shown aligned at the first methionine residue in the mouse P600 sequence (40).

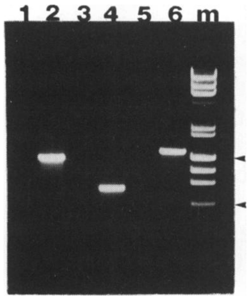

**Figure 4.** Confirmatory PCRs on YAC4/5. An ethidium bromide stained 1% agarose gel of electrophoresed PCR products is shown. Tracks are (1) yeast DNA with human P600 primers, (2) YAC4/5 with P600 primers, (3) yeast with IRF1 primers, (4) YAC 4/5 with IRF1 primers, (5) yeast with IL4 primers, and (6) YAC4/5 with IL4 primers. The marker track on the right is a mixture of lambda *Hind*III and phiX174 *Hae*III DNAs, the positions of the 1,353bp and 603bp marker fragments are shown by the arrows at the right side.

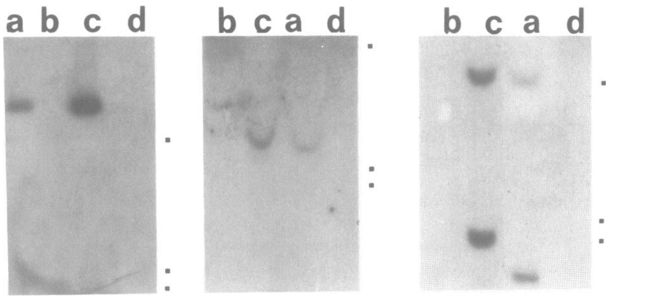

**Figure 5.** Confirmatory Southern blots on genomic DNAs. BamHI digested genomic DNAs were eletrophoresed on 0.8% agarose gels, transferred and hyridized with radiolabelled cDNA probes. Three autoradiogramns are shown representing from left to right, hybridizations with clones #8, #50 and #32 respectively. The genomic DNAs are (a) hamster DNA, (b) HHW105 DNA, a hamster X human hybrid that contains human chromosome 5 as its only human chromosome, (c) human genomic DNA, and (d) yeast genomic DNA. At the left of each panel are shown from top to bottom, the positions of the lambda *Hind*III 4.3kb, 2.3kb and 2.0kb markers.

hybrid, but the length of the shorter fragment differs between the two tracks. This is due either to the presence of a *Bam*HI polymorphism or to partial digestion of the human genomic DNA. In any case, the data indicate that clone #32 is also low in copy number and chromosome 5-specific.

### Assessing enrichment

To determine the sensitivity of the selection and the abundance levels of the various cDNAs throughout the selection, duplicate plaque lifts from the starting, primary, and secondary selected libraries were separately hybridized with IL4, IRF1, P600 and four of the nine novel cDNAs. Approximately 450,000 clones from the starting cDNA library, 5,000 clones from the primary selected library, and 100 clones from the secondary selected library were screened with each probe. Duplicate positive signals were counted and the results are presented in Table 1. These data indicate that with the exception of IL4 which was moderately abundant in the starting library, all of the selected cDNAs were in the low abundance classes (<0.001%) of the library prior to selection and comprised at least 4% of the secondary selected material. Overall, enrichments varied from 2,000-fold for IL4, up to greater than 100,000-fold for the P600 and IRF1 cDNAs.

## DISCUSSION

In this report, we have described the isolation of nine novel low abundance cDNAs that are homologous to a 425kb YAC clone from chromosome 5q23–31. As was mentioned above, because of the manner in which our starting cDNA was constructed (complete digestion with *Mbo* I) it is not possible to immediately determine how many individual transcription units these cDNAs were derived from, and an additional step to convert each fragment to a full length cDNA is necessary. This problem is surmountable by either returning to the full length starting cDNA as a source for full length clones, or by using random primed cDNAs as a starting source for selections and thus presumably retaining some overlap between clones after selection. If accurate DNA sequence data is required, then the eventual isolation of a full length cDNA by conventional methods is advisable to avoid PCR-based sequence alterations. The future use of blunt-end linker/adaptors is also an important consideration since some cDNAs (most notably the IL5 cDNA in this study) were not rendered amplifiable by the addition of *Mbo* I linkers. It is possible that other cDNAs that were present in the starting source may have been selected against in a similar manner. It is also very likely that our analysis of only twenty-four randomly picked cDNAs from the selected library does not constitute a complete description of all the selected cDNAs that are homologous to YAC4/5. To determine this, we are currently analyzing much greater numbers of the secondary selection clones that are not members of the sets we have already picked. However, at least in the cases of P600 and IRF1, picking only twenty-four clones was sufficient to find these cDNAs multiple times.

Notwithstanding these minor technical adjustments, the modified direct selection protocol, as described here, is clearly capable of enriching cDNAs that are in the lowest abundance classes to levels where they constitute several percent of the ending library and can be randomly picked. These cDNAs would be entirely missed by conventional screening protocols. Indeed, in the case of the P600 cDNA a screen of the starting cDNA library with the pure cDNA probe yielded no positives in a screen of 450,000 clones (Table 1). Thus, a conventional cDNA screen was negative, but the cDNA is nevertheless present in the library and is encoded by the YAC4/5. The function of the mouse P600 gene has not been described, however, its amino acid sequence suggests that it might be membrane anchored (40) and it is known to be expressed by a mouse T helper cell subtype known as Th2, the functional equivalent of which has not been described in humans. It is noteworthy that the Th2 cell type also expresses both IL4 and IL5 (41) and this raises the interesting possibility that the close physical linkage of these three genes may reflect an evolutionary conversation in the arrangement of cell type specific genes. One prediction from this work is that the arrangement of these three genes will be conserved across species and that the mouse P600 gene will be found to reside close to the IL4 and IL5 genes on mouse chromosome 11 (42–44). The IRF1 gene encodes a nuclear transcription factor that binds to the virus inducible elements of the IFN-alpha and IFN-beta gene promoters as well as to IFN-inducible promoters (39). Unlike the P600 gene, the IRF1 gene appears from our cosmid contig data to be located at some distance (>100kb) from IL4 and IL5.

The enrichment data summarized in Table 1 indicate that a moderately abundant cDNA such as IL4 has essentially plateaued in its enrichment after one cycle of selection. In contrast, most of the low abundance cDNAs show a ten-fold improvement in abundance with a secondary cycle of selection. A notable exception to this is clone #40 which only shows a two-fold improvement. Our preliminary data indicate that clones containing repetitive elements are relatively depleted in the secondary library compared to the primary and this may account for some of the additional enrichment, but it does not account for all of it. No individual cDNA species (IRF1 cDNAs constitute three species), accounts for greater than approximately 12% of the selected clones in the secondary selection. In addition, the relative abundance ratio of IL4 to IRF1 changes from 10:1 in the starting library to 3:1 in the primary and 1:3 in the selected library. These observations suggest that some level of abundance normalization may be occurring in these selections (10). Theoretically, this should be possible using direct selection methods. If the genomic DNA is limiting in the hybridization it will dictate the ending level of each cDNA. Thus, the abundant cDNAs will quickly saturate their respective genomic targets, leaving the majority of higher abundance cDNAs in solution rather than hybridized to the genomic target. Low abundance cDNAs will also be selected, but because there are less of these to start with, they will not be present in such a vast excess over their genomic sites. The net result when both sets are eluted from the genomic DNA should be an approximate normalization of the abundant cDNAs downwards and the low abundance classes upwards. This conceptually simple postulation obviously ignores the problems that long genes with multiple exons, pseudo genes, gene families, hybridization networking of cDNAs, or repeats would introduce. However, some of these anticipated problems should be addressable. For example, repeats can be blocked with reasonable efficiencies or could be depleted from cDNA libraries or genomic DNAs by modifications of the biotin/streptavidin capture scheme shown in Figure 1.

Approximate abundance normalization coupled with the speed, selectivity, and accuracy of the direct selection method should provide a powerful tool for the isolation of cDNAs encoded by much larger genomic regions, and large sets of cDNAs from multiple tissues could conceivably be used. These methods should also be readily applicable to the isolation of cDNAs that are encoded by candidate disease gene regions and to the derivation of temporal and spatial transcription maps across extensive regions of the human genome.

## ACKNOWLEDGEMENTS

This work was supported by grants RO1 HG00368 and R44 HG00508 to M.L. by the National Center for Human Genome Research. The authors are grateful to Annette Chapman and Rick Cuevas for their assistance in preparing this manuscript.

## REFERENCES

1. Galau, G.A., Klein, W.H., Britten, R.J. and Davidson, E.H. (1977) *Arch. Biochem. Biophys.* **179**, 584–599.
2. Bishop, J.O., Morton, J.G., Rosebach, M. and Richardson, M. (1974) *Nature (London)* 250, 199–204.
3. Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual, 2nd Edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
4. Milner, R.J., Lai, C., Lenoir, D., Nave. K., Bakhit. C. and Malfroy, B. (1986) *Biochem. Soc. Symp.* **52**, 107–17.
5. Ohlsson, R. (1989) *Cell. Differ. Dev.* **28**, 1–15.
6. Ko, M.S.H. (1990) *Nucleic Acids Res.* **18**, 5705–5711.
7. Patanjali, S.R., Parimoo, S. and Weissman, S.M. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 1943–19477.
8. Lovett M., Kere J. and Hinton, L.M. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 9628–9632.
9. Parimoo S., Patanjali S.R., Shukla H., Cahplin D.D. and Weismann, S.M. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 9623–9627.
10. Weissman, S.M. (1987) *Mol. Biol. Med.* **4**, 133–143.
11. Crampton J.M., Davies K.E. and Knapp T.F. (1981) *Nucleic Acids Res.* **9**, 3821–3834A.
12. Wallace M.R., Marchuk D.A., Anderson L.B., Letcher R., Odeh H.M., Saulino A.M., Fountain J.W., Brereton A., Nicholson J., Mitchell A.L., Brownstein B.H. and Collins F.S. (1990) *Science* **249**, 181–186.
13. Elvin, P., Slynn, G., Black, D., Graham, A., Butler, R., Riley, J., Anand, R. and Markham, A.F. (1990) *Nucleic Acids Res.* **18**, 3913–3917.
14. Liu, P., Legerski, R. and Siciliano, J. (1989) *Science* **246**, 813–815.
15. Corbo, L., Maley, J.A., Nelson, D.L. and Caskey, C.T. (1990) *Science* **249**, 652–655.
16. Duyk, G.M., Kim, S.W., Myers, R.M. and Cox, D.R. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 8995–8999.
17. Buckler, A.J., Chang, D.D., Graw, S.L., Brook, D., Haber, D.A., Sharp, P.A. and Housman D.E. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 4005–4009.
18. Hochgeschwender, U., Sutcliffe, J.G. and Brennan, M.B. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 8482–8486.
19. Lennon G. and Lehrach H. (1991) *Trends in Genetics* **7**, 314–317.
20. Reyes, G.R., Bradley, D.W. and Lovett, M. (1992) *Seminars in Liver Disease.* (in press)
21. Kere, J., Taillon-Miller, P., Hinton, L.M. and Lovett, M. (submitted)
22. Burke, D.T., Carle, G.F. and Olson, M.V. (1987) *Science* **236**, 806–812.
23. Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharfe, S.J., Higuchi, R., Horn, G.T., Mullis, K.B. and Ehrlich, H.A. (1988) *Science* **239**, 487–494.
24. Chirgwin, J.M., Przybyla, A.E., MacDonald, R.J. and Rutter W.J. (1979) *Biochemistry* **18**, 5294–5299.
25. Green, E.D. and Olson, M.V. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 1213–1217.
26. Young, B.D. and Anderson, M.L.M. (1985) In Hames, B.D., & Higgins, S.J. (eds.) Nucleic Acid Hybridization–A Practical Approach. IRL Press Oxford, pp 47–71.
27. Reyes G.R. and Kim J.P. (1991) *Molecular and Cellular Probes* **5**, 473–480.
28. Wasmuth, J.J., Park, C. and Ferrell, R.E. (1989). *Cell. Genet.* **51**, 137–148.
29. van Leeuwen, B.H., Martinson, M.E., Webb, G.C. and Young, I.G. (1989) *Blood* **73**, 1142–1148.
30. Le Beau, M.M., Lemons, R.S., Espinosa, R., Larson, R.A., Arai, N. and Rowley, J.D. (1989) *Blood* **73**, 647–650.
31. Huebner, K., Nagarajan, L., Besa, E., Angert, E., Lange, B.J., Cannizzaro, L.A., van den Berghe, H., Santoli, D., Finan, J., Croce, C.M. and Nowell, P.C. (1990) *Am. J. Hum. Genet.* **46**, 26–36.
32. Chandrasekharappa, S.C., Rebelsky, M.S., Firak, T.A., Le Beau, M.M. and Westbrook, C.A. (1990) *Genomics* **6**, 94–99.
33. Warrington, J.A., Hall, L.V., Hinton, L.M., Miller, J.N., Wasmuth, J.J. and Lovett, M. (1991). *Genomics* **11**, 701–708.
34. Warrington, J.A., Bailey, S.K., Armstrong, E., Aprelikova, O., Alitalo, K., Saltman, D., Wilcox, A., Sikela, J., Lovett, M. and Wasmuth, J.J. (1992)*Genomics* **13**, 803–808.
35. Saltman, D.L., Dolganov, D.M., Warrington, J.A., Wasmuth, J.J. and Lovett, M. (submitted)
36. Itoh, S., Harada, H., Nakamura, Y., White, R. and Taniguchi, T. (1991) *Genomics* **10**, 1097–1099.
37. Yokota, T., Arai, N., de Vries, J., Spits, H., Banchereau, J., Zlotnik, A., Rennick, D., Howard, M., Takebe, Y. and Miyatake, S. (1988) *Immunol. Rev.* **102**, 137–187.
38. Azuma, C., Tanabe, T., Konishi, M., Kinashi, T., Noma, T., Matsuda, F., Yaoita, Y., Takatsu, K., Hammarstrom, L., Smith, C.I., et al. (1986) *Nucleic Acids Res.* **14**, 9149–9158.
39. Maruyama, M., Fujita, T. and Taniguchi, T. (1989) *Nucleic Acids Res.* **17**, 3292.
40. Brown, K.D., Zurawski, S.M., Mosmann, T.R. and Zurawskik, G. (1989) *J. Immunol.* **142**, 679–687.
41. Fong, T.A.T. and Mosmann, T.R. (1990) *J. Immunol.* **144**, 1744–1752.
42. D'Eustachio, P., Brown, M., Watson, C. and Paul, W.E. (1988) *J Immunol.* **141**, 3067–3071.
43. Lee, J.S., Campbell, H.D., Kozak, C.A. and Young, I.G. (1989) *Somat. Cell. Mol. Genet.* **15**, 143–152.
44. Takahashi, M., Yoshida, M.C., Satoh, H., Hilgers, J., Yaoita, Y. and Honjo, T. (1989) *Genomics* **4**, 47–52.