W.E. Crusio and R.T. Gerlai (Eds.)
Handbook of Molecular-Genetic Techniques for Brain and
Behavior Research (Techniques in the Behavioral and Neural Sciences, Vol. 13)

CHAPTER 2.2.5

Analyzing genomic DNA discordance between monozygotic twins

J. Bouchard,[1] C. Foulon,[2] N. Storm,[1] G.H. Nguyen[1] and C.L. Smith[1]

[1]*Center for Advanced Biotechnology and Departments of Biomedical Engineering, Biology and Pharmacology, Boston University, 36 Cummington Street, Boston, MA 02215, USA*
[2]*Institute des Sciences et Techniques de L'Ingénieur d'Angers, 62 Rue Notre Dame du Lac, 49100 Angers, France*

Introduction

Monozygotic (MZ) or "identical" twins result from the fertilization of one ovum by one sperm. For unknown reasons, the embryo divides into two genetically "identical" embryos during the early stages of development following fertilization. Presently, sex, placentation, cord blood type, HLA antigens, and DNA fingerprinting are used for zygosity testing (Derom et al., 1987; Hill and Jeffreys 1985; Akane et al., 1991). Thus far, no method is considered the "gold standard" and all have the possibility for error.

Intra-uterine environmental differences in the allocation of cells in the placental vascular supply to each twin, as well as stochastic developmental events, may lead to major phenotypic discordance within a MZ twin pair (Hall, 1996). Application of new molecular techniques to MZ twins reveals that they do not possess identical genomes when individual loci are established (Jansen et al., 1994; Kruyer et al., 1994; Machin, 1996; Reyniers et al., 1993). Hence, it is important to reassess the role of genomic differences between MZ twins in producing phenotypic discordances. For example, the study of MZ twins is particularly valuable when only one twin is affected by disease, because a comparison of the almost identical genomes can lead to the identification of genetic abnormalities.

Various methods, including subtractive hybridization (Lisitzyn et al., 1993a), serial analysis of gene expression (SAGE; Velculescu et al., 1995), comparative genomic hybridization (CGH; Kallioniemi et al., 1992), and differential display (DD: Liang and Pardee, 1992) can be used for comparing genomes. However, these methods do not allow direct global comparisons between genomic DNA from complex genomes. Instead, they focus on comparing cDNAs. Working with cDNAs allows comparative studies to focus on important functional units (genes) and reduce sample complexity. For example, when cDNA is analyzed, the human genome complexity is reduced from 3×10^9 base pairs (bp) to an estimated 1×10^8 bp ($\sim$100,000 genes $\times$ $\sim$1000 bp/gene).

Targeted Genomic Differential Display (TGDD) was developed by us to compare complex genomes using genomic DNA rather then cDNA (Broude et al., 1997, 1999). In TGDD, it is essential that genomic complexity be reduced to an analyzable level. Targeting in TGDD is used to reduce genome complexity and to focus analysis on, and nearby, sequences of interest. The target sequence can be a simple repeating sequence

(e.g. $(CAG)_n$ or $(TG)_n$), a sequence coding for a protein motif, a transcriptional regulator element, or any other important genomic region. In TGDD, as in DD, DNAs are fractionated (displayed) by size to produce a DNA fingerprint. Essentially, TGDD detects restriction fragment length polymorphisims (RFLPs), where each difference must be characterized to understand its origin. RFLPs may arise from different causes including DNA sequence variation, insertions, deletions recombination events or methylation changes.

Two TGDD protocols have been developed. Method I (Broude et al., 1997) uses a capture and PCR protocol whereas Method II (Broude et al., 1999) solely uses PCR (Figs 1 and 2, respectively). Here, the TGDD protocols are described and illustrated with results obtained from the genomic analysis of twin pairs.

Method

Isolation of genomic DNA (Method I and II)

TGDD can use genomic DNA isolated from human blood lymphocytes, buccal scrapes, or sperm. Blood samples were collected in 6 ml ACD (Becton-Dickenson) tubes to maximize the allowable time between collection and DNA isolation (i.e. 5 d maximum). The buccal scrape was performed by wiping each side of the oral cavity 10 times with a cytobrush (Medscand #1101). The brush was then shaken in 300 μl cell lysis solution and immediately removed. DNA from buccal scrape samples was collected and stored in the cell lysis solution included in the Puregene DNA Isolation Kit (Gentra Systems, Minneapolis). Freshly ejaculated sperm was mixed with 0.3X ESP (1X ESP = 1 mg/ml ethelene diamine tetraacetic acid (EDTA), 1% sodium lauroylsarcosine (Sigma #L5000), and 1 mg/ml proteinase K; Sigma), and 7% (V:V) β-mercaptoethanol as described by Smith et al., 1993. The high concentration of β-mercaptoethanol is necessary because of the large number of disulfide bonds in the sperm coat proteins. Samples stored in ESP may be shipped and stored at room temperature until further use.

The experimental data shown in this chapter were obtained from blood samples. The MZ twin samples were obtained from Terry Reed of the National Heart, Lung, and Blood Institute (NHLBI) and E. Fuller Torrey of the National Institute of Mental Health. The sibling samples were obtained from Clinton Baldwin of the Boston University Medical Center. The DNA was extracted using the protocol accompanying the Puregene DNA cell line isolation kit (Gentra System, Minneapolis). The red blood cells were lysed to facilitate their separation from white blood cells by adding 18 ml RBC lysis solution to 6 ml whole blood. After 10 min incubation at room temperature, the samples were centrifuged for 10 min at 3000 rpm and the supernatant was removed. To the remaining white pellets, 6 ml of Lysis Solution was added and pipeted up and down to lyse the white blood cells. Proteins were sedimented by centrifugation at 3000 rpm for 10 min after vigorous mixing of 2 ml protein precipitation solution. The supernatant was then transferred to a new tube and the genomic DNA was precipitated with 6 ml isopropanol by inverting the tube until the white threads of DNA formed a visible clump. The white DNA threads were scooped out with a heat-sealed Pasteur pipet and dissolved in a DNA hydration solution contained in a microcentrifuge tube. The tube was then incubated overnight in a 37°C oven. The sample was treated with RNase (1:1 (wt:wt) RNase:DNA;Boehringer Mannheim) at 37°C for 30 min, and then the DNA was stored at 4°C until used.

Genomic DNA concentration (Method I and II)

It is of great importance that the concentrations of DNAs being compared are well matched. The number of false differences between samples increases proportionally with differences in DNA template concentrations used in the PCR reaction. The DNA concentrations should be

Method I

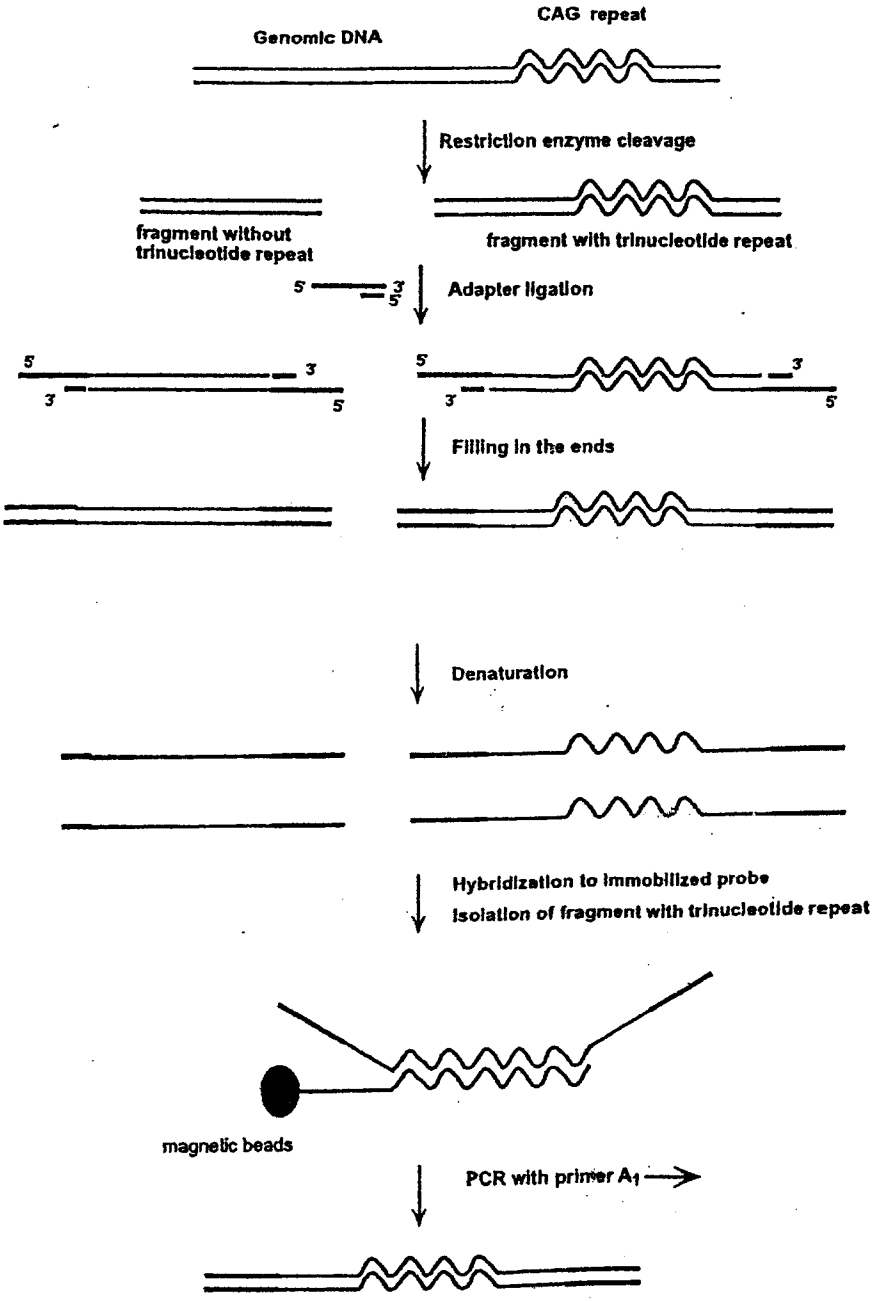


Fig. 1. General protocol for TGDD Method I (see text for description).

Method II

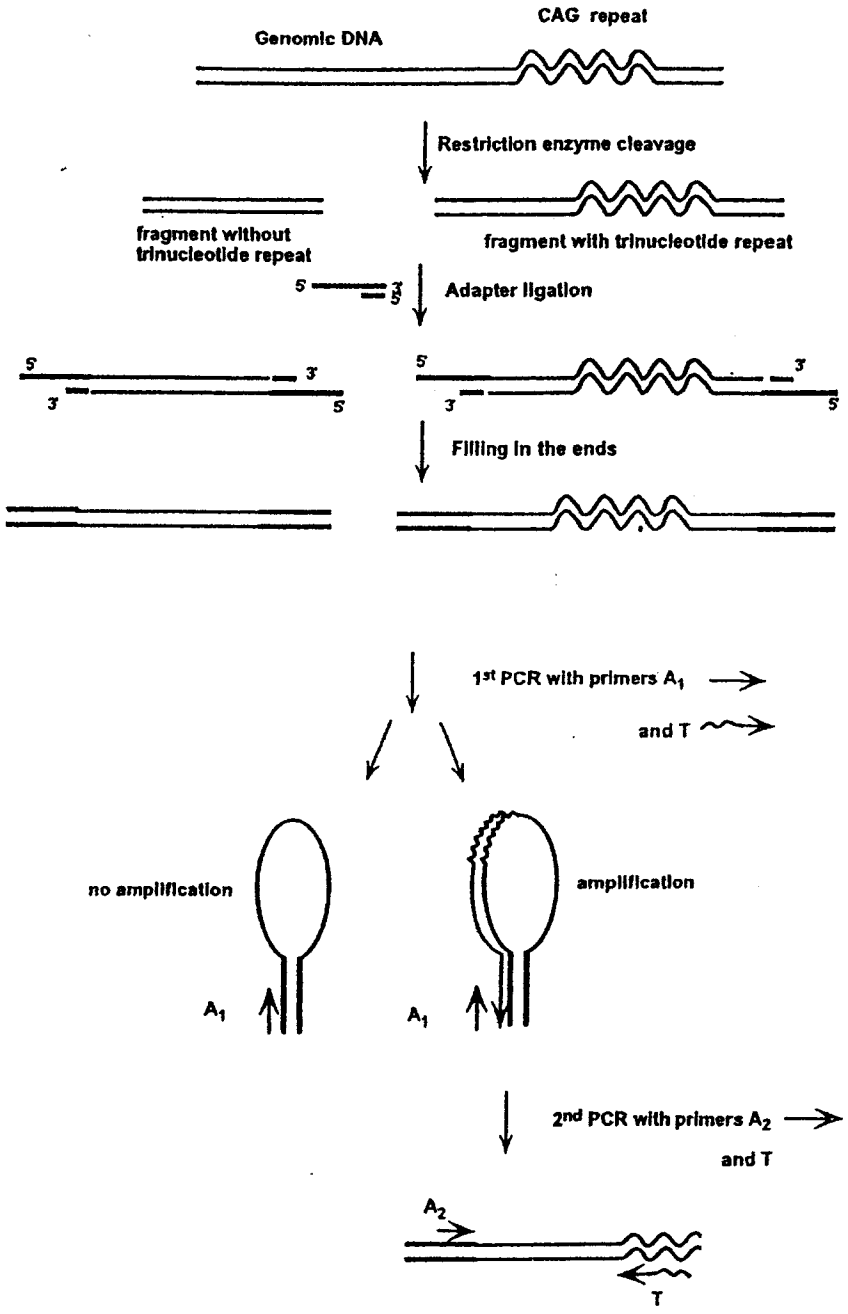


Fig. 2. General Protocol for TGDD Method II (see text for description).

matched throughout the protocol. In some cases, these differences in concentration may only be apparent after the products of TGDD are examined.

There are a number of ways to determine DNA concentrations. All methods have problems. Hence, two goals must be kept in mind. The DNA concentrations used must be well-matched and within an acceptable range. The acceptable range before the first PCR is ~5–20 ng, but some adjustment in the number of PCR cycles may be necessary to insure good signal-to-noise ratios are obtained.

Matching DNA concentrations is best determined on an agarose gel stained with ethidium bromide. The fluorescence of the samples are compared to each other and a standard DNA sample. The samples are briefly electrophoresed to ensure that the RNA component of the sample is separated from the DNA. DNA concentration has also been determined using a DipStick Kit (Invitrogen) as recommended by the manufacturer. Matching fluorescence on agarose gels, however, appears to be the most reliable method for DNA concentration determination and matching. It is important that an individual be very consistent to insure reproducibility.

DNA preparation (Method I and II)

Approximately 1 μg of MZ twin genomic DNA was digested in reactions containing 20 units of *Sau3AI* or *Hae III* (Method I and II respectively; New England Biolabs) and 1X New England Biolabs Buffer *Sau3AI* (100 mM NaCl; 10 mM Bis Tris Propane-HCl; 10 mM $MgCl_2$; 1 mM dithiothreitol, pH 7.0 at 25C) or New England Biolabs Buffer #2 (50 mM NaCl; 10 mM Tris-HCl; 10 mM $MgCl_2$; 1 mM dithiothreitol, pH 7.9 at 25°C) respectively supplemented with 100 μg/ml bovine serum albumin (BSA). Reactions were incubated at 37°C overnight, and the restriction fragments were purified using Phase Lock Gel Tubes (5 Prime-3 Prime, Boulder, CO) and ethanol precipitated.

Ligation (Method I and II)

The digested genomic DNA was then ligated to oligonucleotides of a known sequence (adapters) in a 20 μl reaction. Thus far, the adapter sequences used for Method I and Method II were different. Method I and II used oligonucleotides 1 and 2 or 3 and 4 respectively (Table 1). The reaction consisted of 10 μl of restricted genomic DNA (approximately 1 μg), 2 μl of 10X T4 ligation buffer from New England Biolabs (1X solution: 50 mM Tris-HCl, pH 7.5; 10 mM dithiothreitol; 1 mM ATP; 25 μg/ml BSA), 400 units (1 μl) T4 ligase (New England Biolabs), 2 μl of each 10 μM adapter mixture (equimolar amounts of oligonucleotides 1 and 2 or 3 and 4 in Methods I and II respectively; Table 1), and 3 μl of H_2O. Samples were incubated overnight at 16°C. DNA was phenol-extracted, precipitated with ethanol, washed with 70% ethanol, dried, and dissolved in TE buffer (10 mM Tris HCl, pH 8.0; 1 mM EDTA).

Capture/PCR targeting (Method I)

A biotinylated oligonucleotide (10 pmol) containing a target sequence (i.e. $(CTG)_{12}$; oligonucleotides 5, Table 1) was mixed with 50 ng of ligation products in 50 μl of TE buffer containing 2 μM of the corresponding adapter oligonucleotides to prevent annealing of the fragment ends to each other. After the addition of mineral oil, the sample was heated to 95°C, slowly cooled to room temperature, added to 100 μg of prewashed streptavidin coated magnetic beads M-280 [as directed by Dynal (Oslo)] using a 3-fold molar excess of biotin binding capacity over biotinylated oligonucleotides, and incubated at room temperature for 1 h with gentle rotation. The beads were collected with a magnet, washed twice at 55–60°C for 20 min with 3X standard saline citrate (SSC; 1X SSC = 0.15 M NaCl, 15 mM sodium citrate) and 0.5% SDS and, at room temperature, twice each, with TE containing 1 M NaCl and with

TABLE 1

Synthetic oligonucleotides used in this work

Number	Description	Sequence (5′ → 3′)[a]
1	*Sau3AI* adapter 24	CGGGAATTCTGGCTCTGCGACATG
2	*Sau3AI* adapter 10	GATCCATGTC
3	*Hae III* adapter 43	TGTAGCGTGAAGACGACAGAAAGGGCGTGGTGCGGAGGGCGGT
4	*Hae III* adapter 11	ACCGCCCTCCG
5	CTG-12	b-GATGATCCGACGCAT$(CAG)_{12}$
6	CTG-A T-primer	$(CTG)_6A^b$
7	CTG-G T-primer	$(CTG)_6G^b$
8	CTG-T T-primer	$(CTG)_6T^b$
9	Na21 A-primer	TGTAGCGTGAAGACGACAGGA
10	ST19*HaeIII* A-primer[c]	AGGGCGTGGTGCGGAGGGCGGTCC
11	ST19*HaeIII*G A-primer[c]	AGGGCGTGGTGCGGAGGGCGGTCCG
12	ST19*HaeIII*GG A-primer[c]	AGGGCGTGGTGCGGAGGGCGGTCCGG
13	ST19*HaeIII*TG A-primer[c]	AGGGCGTGGTGCGGAGGGCGGTCCTG
14	ST19*HaeIII*AC A-primer[c]	AGGGCGTGGTGCGGAGGGCGGTCCAC

[a] b = biotin; [b] Cy5 labeled; [c] Note the presence of 3′ CC bases. These bases anneal to genomic GG sequences remaining from the *HaeIII* recognition site; 3′ anchored bases are terminal to the CC dinucleotide.

TE alone. Beads with captured DNA were stored in TE buffer at 4°C.

One-fifth of the captured DNA was amplified by PCR in a PTC-100 thermal cycler (MJ Research, Cambridge, MA). The 50 μl reaction contained 67 mM Tris HCl, (pH 8.8); 4 mM $MgCl_2$; 16 mM $(NH_4)_2SO_4$; 10 mM 2-mercaptoethanol; 300 μM of each dNTP; 2 units of AmpliTaq DNA polymerase; and 5 μM fluorescent labeled adapter primer (A-primer). After a hot start (see below), samples were subjected to 20–23 PCR cycles, each consisting of 1 min at 94°C and 3 min at 72°C, and a final incubation at 72°C for 5 min (see results for discussion of PCR primer choices).

PCR/targeting (Method II)

Two consecutive hot start PCRs (see below) were performed. The first and second PCRs targeted a specific sequence of interest, whereas the second uses a semi-nested PCR to further reduce the complexity of the amplified fragments. The products of the first PCR were diluted 1/100 and used as templates for a second PCR. Here, conditions for the second PCR were the same as in the first PCR, except that a different oligonucleotide (10–14; Table 1) is used in place of oligonucleotide 9 (Table 1) and 20–25 cycles were performed. Usually the 25 μl PCR contained 10–12 ng DNA, 5 mM $MgCl_2$, 1 mM of each dNTP, 5 pM of each primer and 1.25 Units of enzyme.

Hot start PCR (Method I and II)

Usually, TaqStart (Clontech, Palo Alto, CA) antibody was used to insure hot start PCR. Hot start PCR insures that primer annealing and elongation does not occur until the complementary sequence is found. The hot start mix consisted of 0.25 μl TaqStart Antibody (1.1 mg/μl; 7 μM), 0.25 μl AmpliTaq DNA polymerase (5 units/μl; 0.25 μM) and 1 μl dilution buffer, 50 mM KCl, 10 mM Tris-HCl (pH 7.0) was incubated at room temperature (20–22°C) for 5 min and then added to a 25 μl PCR reaction. The PCR was centrifuged and subjected to the PCR conditions listed above.

Hot start PCR was also performed by withholding one of the reaction components until the sample had gone through it first long denaturating step. It is best to withhold the enzyme to maximize its activity. For instance, the half-life of Amplitaq is 40 min at 95°C. For hot start PCR, the enzyme was diluted into 10 μl of 1X PCR buffer and the other components were made up in a total volume of 15 μl 1X PCR buffer. The 15 μl was overlaid with mineral oil, denatured and brought to 80°C. The diluted enzyme dropped on top of the mineral oil sedimented to the aqueous bottom layer to complete the reaction components. The PCR was then performed as described above.

TGDD fractionation (Method I and II)

Most times the amplification products were analyzed on a 2% agarose gel containing ethidium bromide. The approximate lengths of the amplified fragments were determined by comparison with a 50 bp ladder (Boehringer Mannheim). Then, the DNA fragments were analyzed on a high-resolution DNA sequencing gel. Usually, two and a half microliters of the second PCR were mixed in 3.5 μl of a stop solution (6 mg/μl of dextran blue and 0.1% Sodium dodecyl sulfate in deionized formamide; Pharmacia Biotech, Upsala, Sweden;) and denatured for 5 min at 94°C. After denaturation, the sample was immediately quenched on ice and loaded on a 6% denaturing polyacrylamide gel (PAGE) in 0.6X TBE (54 mM Tris-borate, 1.2 mM EDTA). The samples were fractionated using the ALFexpress Sequencer (Pharmacia Biotech). A Cy-5 labeled 50 bp ladder (ALFexpress sizer 50-500, 27-4539-01, Pharmacia Biotech) was used as a standard to determine the fragment length.

Data analysis

The results were visualized and compared using the Fragment Manager Software provided with the ALFexpress. This software displays band intensity as a function of elution time, which is approximately equal to bp length. The data can be viewed in two different formats: fullscale and autoscale. Full-scale data is unprocessed intensity measurements, where band intensity is shown relative to background intensity. Most of the data presented here is autoscaled with the highest intensity peak in each lane set to 100%. The lower intensity peaks are scaled relative to the largest peak. All data was examined in both formats. Full-scale analysis allows signal-to-noise ratios to be examined, while autoscaled analysis amplifies peak height and is especially useful with low signal-to-noise data. Differences found in twin pairs due to varying signal-to-noise ratios must be ignored in both cases. Figs. 4, 8 and 11 show autoscaled data, while Figs. 6, 9, 10 and 12 show fullscale data.

Isolation of polymorphic fragments (Method I and II)

Specific fragments were isolated from a 3% 3:1 (wt:wt) Nusieve: LE agarose (FMC Bioproducts) by fractionation of 20 μl of sample for 5 h at 130 V in a 1X TAE (40 mM Tris-acetate, 1.0 mM EDTA) buffer. Specific bands of interest were picked from the gel with a pipet tip and reamplified. These products were then analyzed on a 6% denaturing polyacrylamide gel as described above. The gel purification may be repeated for direct genomic sequencing or the fragments may be cloned and sequenced.

Results

Two methods for TGDD are described. These methods are very similar; however, there are important technical details that are described in the Materials and Methods that differentiate Method I from II. The goal of both methods is to create a library composed of restriction fragments that share a target sequence. This means that each member of the library has a common

sequence and a unique sequence. The experiments described here focused on $(CAG)_n$ repeats containing restriction fragment libraries. These libraries can be used to create a complex DNA fingerprint as described in this work or may be used in other applications, such as creating clone libraries containing sequences useful for genetic mapping experiments (Oliveria et al., 1998).

The target libraries are created by a capture and PCR protocol (Method I) or a PCR protocol alone (Method II). Currently, most of our experiments use Method II because it is easier to perform. However, both methods are more than 90% effective in specific sequence targeting (Broude et al., 1997, 1999). It should be noted that the products of the methods are not the same. Specifically, the fragment products of Method I have a target sequence surrounded by unique sequence and end-tagged with the adapter sequence. The ends of the fragment products of Method II are the target and adapter sequences. Hence, additional experiments must be done to isolate the single copy sequence flanking the other side of the target sequence (see below and Figs. 1 and 2).

DNA preparation

Both methods can use the same preparation of DNA. The first decision to make in applying this procedure is the choice of restriction enzyme to fragment the genome. Our restriction enzyme choices were determined empirically. However, as more of the human genome sequence becomes available, it will be possible to choose the appropriate restriction enzyme a priori based not only on the frequency of occurrence of a specific recognition site, but also on the distance of restriction enzyme cleavage sites from specific target sequences. In fact, ongoing TGDD modeling experiments using the sequences of the *Saccharomyces cerevisiae* and *Caenorhabditis elegans* genomes are doing just this (Bouchard et al., manuscript in preparation).

Cleaved genomic restriction fragments are ligated to oligonucleotides (adapters) of a known sequence. A-primers complementary to the adapter sequences allow PCR amplification of the restriction fragments. Thus far, the adapters used in Methods I and II have been of different lengths. Method I uses adapters that are 20-mer or shorter in order to avoid the annealing of the complementary ends, since fragment end-annealing may interfere with A-primer annealing and extension. Method II uses adapter lengths of 40-mer to promote end annealing and inhibition of A-primer annealing and extension. In Method II, the first round of PCR amplification must occur from the target primer (T-primer) complementary to the target sequence (see below).

Method I

In Method I, the tagged restriction fragments are hybridized to an immobilized single-stranded DNA complementary to the target sequence (Fig. 1). The captured genomic fragments are eluted and used as template DNA for PCR. The single-stranded template DNA has the target sequence surrounded by unique sequences and is end-tagged with the known adapter sequence. PCR amplification may be done using different primer combinations (Fig. 3). The entire template DNA is amplified when a single A-primer is used. Alternatively, an A-primer may be used in combination with a T-primer. In the latter instance, the region between the target sequence and the adapter sequence on one strand is amplified along with the intervening unique sequence. Use of the target sequence complement and an A-primer amplifies the region between the target sequence and the adapter on the other strand. The use of a T-primer also increases the robustness of the targeting process. Most of our experiments used a labeled T-primer, but any primer in the experiment may be labeled. A TGDD comparison of an alleged MZ twin pair, generated by Method I, is shown in Fig. 4.

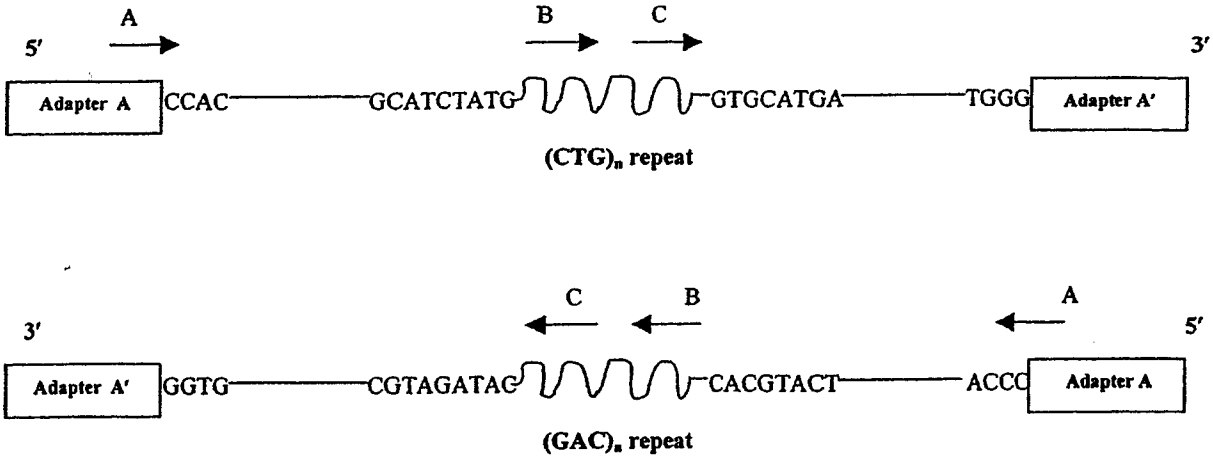


Fig. 3. Possible TGDD primer combinations. A-primers are indicated by A and T-primers are indicated by B or C. Note that the possible primer combinations are A alone, A and B, and A and C for Method I and A and B or A and C for Method II.

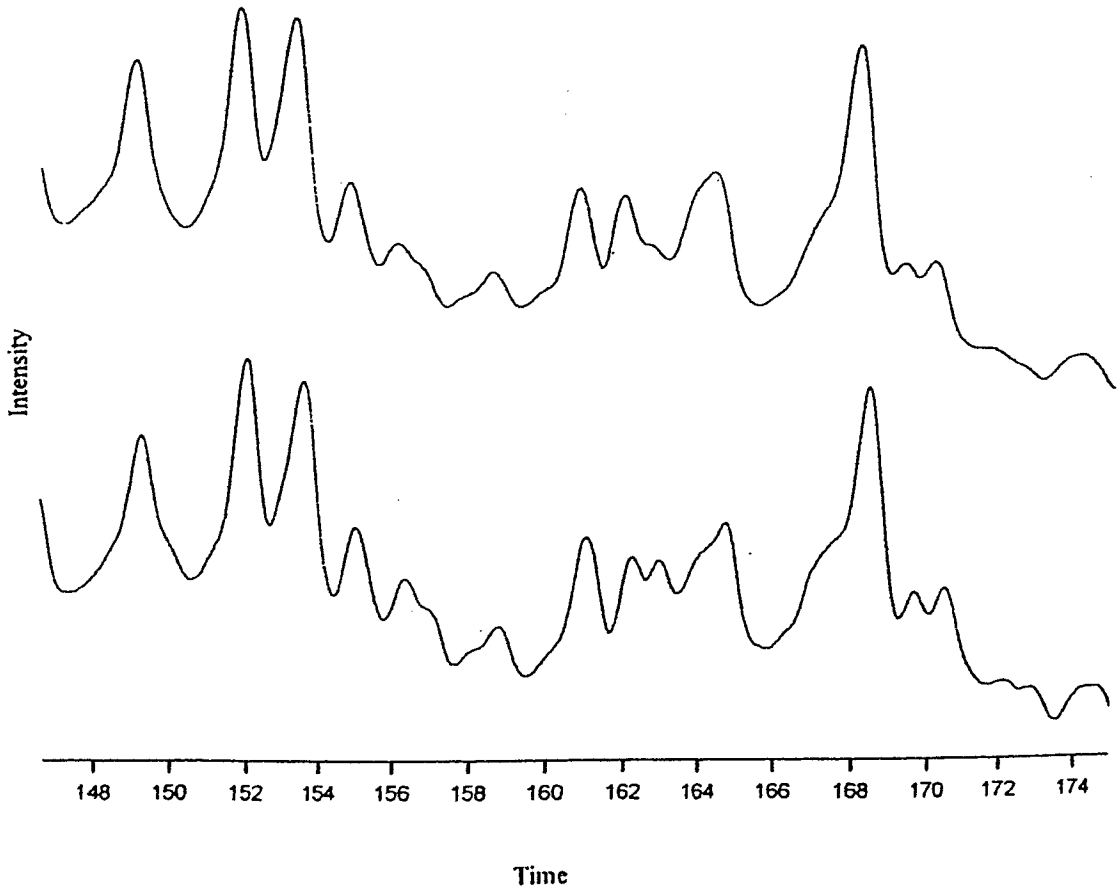


Fig. 4. A Method I comparison of MZ twins. The immobilized capture probes and T-primer were both a $(CTG)_n$ repeating sequences (oligonucleotide 5 and 6 respectively, Table 1).

Method II

Method II uses PCR alone for targeting (Fig. 2). This means that the biotin capture step is eliminated. Targeting is performed using a T-primer under PCR suppression conditions (PS: Siebert et al., 1995). PS takes advantage of the fact that the adapters ligated to the ends of restriction fragments are self-complementary, which enhances targeting. Annealing of the ends of a fragment to each other can interfere with the annealing and extension of a complementary A-primer. The interference is enhanced by using long (e.g. 40 bp) GC-rich adapters. Efficient amplification of such templates occurs only when a T-primer target is located on the internal single-stranded region of the template that is not annealed. Elongation of T-primer will produce a product with non-complementary ends (i.e. one end will contain the target sequence and the other end will contain the adapter sequence). This is the same structure that is obtained with Method I when a T-primer is used in the PCR amplification step. All of the remaining TGDD data presented in this paper used Method II (Figs 6, 8–12).

The T-primer

The T-primer can be composed of any sequence of interest in the genome. Thus far, our focus has centered on $(CAG)_n$ and $(CA)_n$ repeating sequences (Broude et al., 1997, 1999; Oliveria et al., 1998), long terminal repeats (LTR) of human endogenous retroviruses (Lavrentieva et al., 1999), and one Zn-finger gene family (Foulon et al., manuscript in preparation).

The design of the T-primer is a fundamental aspect of TGDD. A major question is whether or not a further reduction in sample complexity is necessary beyond that afforded by the selected T-primer. A sample that is too complex will be difficult to analyze. Sample complexity is reduced by adding unique bases (anchors) to the 3′ or 5′ end of the T-primer and/or A-primer (Fig. 5). It is important to remember that 3′ anchors added to A-primers must be terminal to bases needed to anneal to any remaining restriction enzyme recognition sequences.

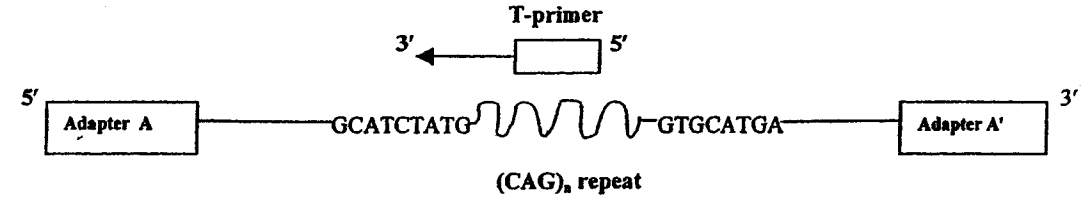
The unique bases also anchor the T-primer to either the 3′ or 5′ end of a T-primer for a simple repeating sequence. Anchor bases anneal to the bases adjacent to the target sequence in genomic DNA. PCR amplification selects those target-containing restriction fragment pool that contain the bases complementary to anchors. Anchor T-primers must be used when the target is a simple repeating sequence to ensure amplification from the same location. A 3′ anchored T-primers amplifies genomic DNA adjacent to simple repeat sequence, and in general, do not provide information about repeat length. This is due to the fact that the PCR products all have the same 5′ end made from the T-primer sequence. A 5′ anchored T-primer should amplify the entire repeat sequence. The effectiveness of 5′ anchoring may be improved by increasing the number of bases (anchors) added to the 5′ end.

It should be noted that sequences such as $(CAG)_n$ repeats are capable of forming alternative structures which may interfere with T-primer annealing and/or extension (see Broude et al., 1997, 1999 for discussion). In fact, long $(CAG)_n$ repeat lengths form hairpin structures which appear to inhibit PCR with a $(CAG)_n$ containing T-primer. Hence, long versus short $(CAG)_n$ repeating sequences within the Huntington's disease locus were distinguished by TGDD (Broude et al., 1997).

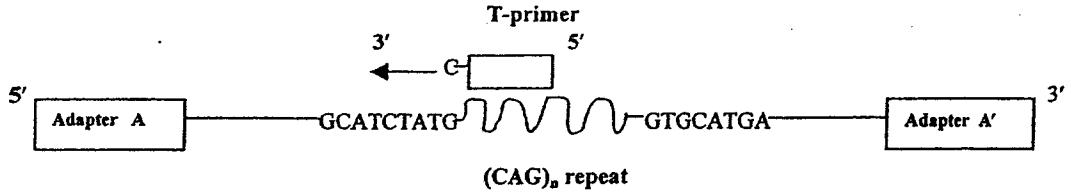
In a random sequence, the addition of a single anchor would reduce the complexity of the sample by four-fold, whereas the addition of two bases would reduce the complexity by sixteen-fold. An example of complexity reduction using anchors is shown in Fig. 6.

Designing a T-primer to target a gene family is more complex. Conventionally, gene families have been identified by amino acid homologies. An example of amino acid alignments in different gene families is shown in Table 2. Some families (e.g.

Unanchored T-primer:



3′ C anchored T-primer:



5′ C anchored T-primer:

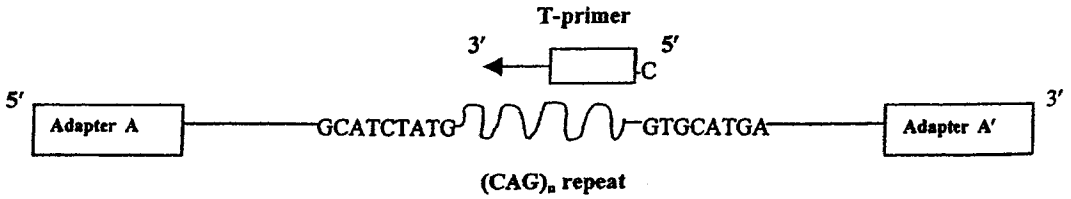


Fig. 5. Principles of anchoring (see text for description).

heat shock 70 proteins) have a well-defined amino acid consensus region while others do not (e.g. protein kinase C family terminal domain). Some gene families may also be subdivided into groups with different consensus sequences. Note that only 6 to 7 amino acids are needed for the PCR primer design.

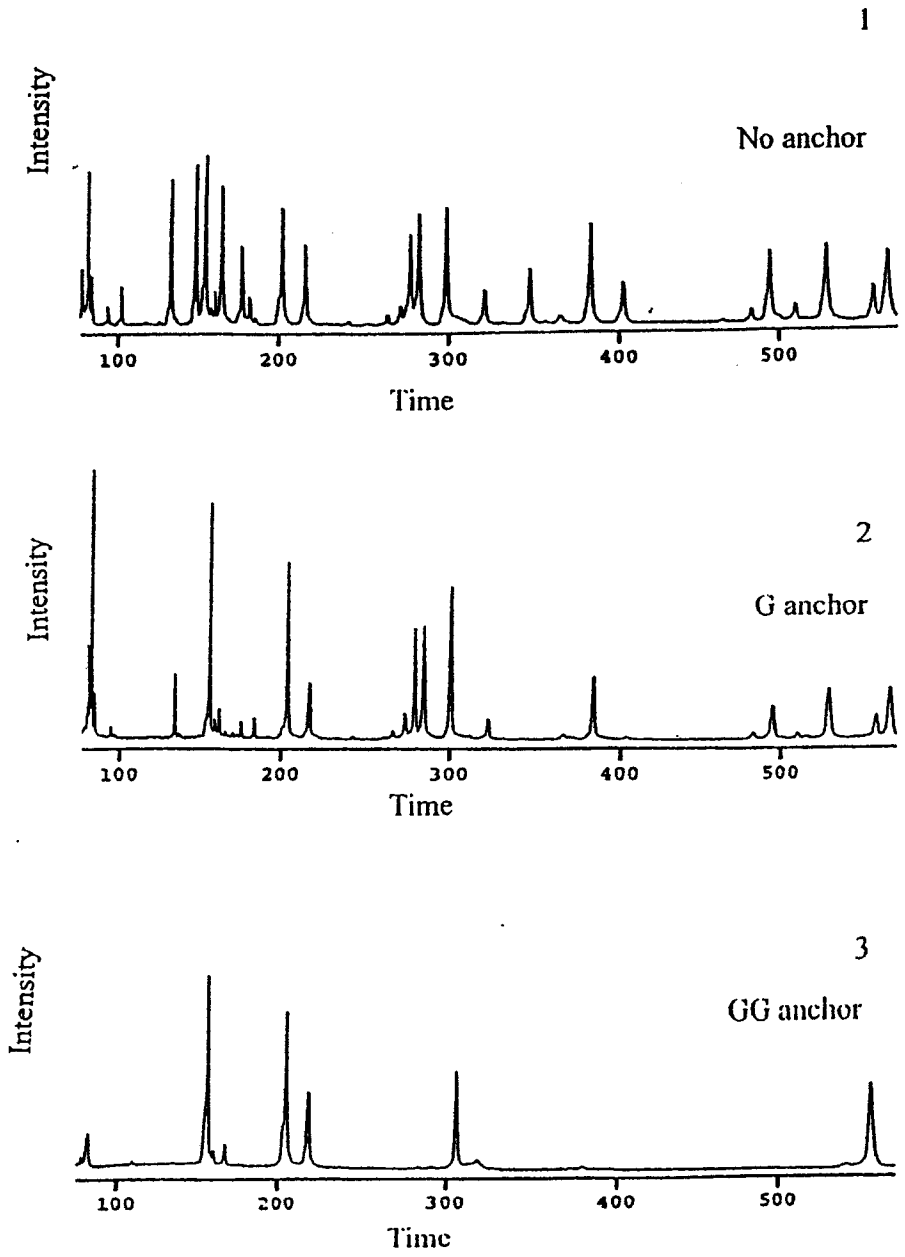


Fig. 6. The effects of 3′ anchoring on TGDD (Method II). The $(CAG)_n$ containing fragments were targeted using oligonucleotide 7 (Table 1). The unanchored A-primers were ST19*HaeIII* (lane 1) and the anchored A-primer was ST19*HaeIII*G (lane 2) or ST19*HaeIII*GG (lane 3); (oligonucleotides 10, 11, and 12 respectively; Table 1).

TABLE 2

Examples of consensus amino acid sequences of protein families. This table was adapted from information obtained from the Sanger Center web site: (http://www.sanger.ac.uk/Pfam/)

Family	Identification number	Selected alignments	Consensus % identity[a]	Total % identity[b]
Heat shock Hsp 70 proteins	DNAK_BACME/4-574	IIGIDLGTTNSCVAVLEGGEPKV	100	92
	DNAK_METMA/4-578	ILGIDLGTTNSCVAVMEFFEAVV	100	88
	DNAK_BORBU/4-599	IIGIDLGTTNSCVAIMEHGKPVV	100	87
	DNAK_CLOAB/4-576	VIGIDLGTTNSCVAVMEGGDPAV	100	85
	DNAK_CHLPN/10-603	IIGIDLGTTNSCVSVMEFFQAKV	100	82
	consensus:	IIGIDLGTTNSCVAVMEGGEPVV		
Protein Kinase C terminal domain	KPC1_YEAST/1084-1149	RNINFDDILNLRVKPPYIPEIKSP	100	79
	PCK2_SCHPO/943-1008	SNINWDDIYHKRTQPPYIPSLNSP	83	77
	KPC1_CANAL/1030-1095	HDVNFDDVLNCRIPAPYIPEVQSE	83	54
	KPC1_HUMAN/601-667	RYIDWEKLERKEIQPPYKPKARDK	66	52
	PCK1_SCHPO/924-988	ASIVWDDLYNKLYEPSYKPLINDP	50	48
	consensus:	RNINWDDLLNRR8QPPYIPEINSPIY		

[a] % identity to underlined consensus sequence
[b] % identity of total consensus

The consensus amino acid sequence is only a part of the puzzle. When deciding on which part of the consensus amino acid sequence to use for primer synthesis, it is best to look for amino acids with as few codons as possible (Table 3). For instance, tryptophan and Methiomne are coded for by only one codon. Most, but not all of the codon variability lies in the third position of the codon. Thus, variable bases used in the primer can compensate for these ambiguities and allow entire gene families or subgroups to be targeted. However, the introduction of ambiguity can have adverse effects on the PCR reactions. Each variable base included in the primer decreases the primer concentration by half. Hence, many times it is not effective to have a broad T-primer. An example is shown in Figure 7.

Instead, it may be useful to use a chimeric primer consisting of a variable and a constant region. This is called Tagged PCR (T-PCR; Grothues et al., 1993). In this approach, several initial PCR cycles are performed with the chimeric primer. The remaining PCRs are performed using a constant region primer, which usually has a higher T_m than the chimeric primer.

Another important consideration is codon usage. Not all codons are used equally within a particular genome. Table 3 shows general codon

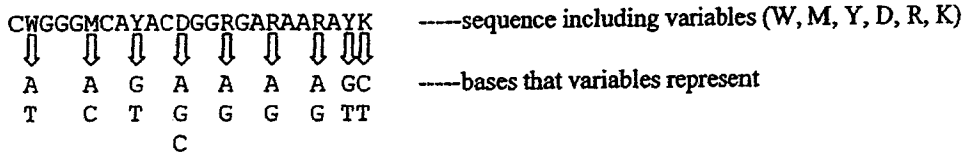


Fig. 7. Variable base options in a T-primer to a protein motif. Eight positions have 2 variables bases and 1 position has three variable bases. If all variability is built into the primer the concentration of any single sequence will be 768 fold ($2^8 \times 3^1$) less than the total concentration. For example, a 1.5 μM primer solution would only have the effectiveness of a 1 nM primer concentration. The best way to compensate for this problem is not obvious.

TABLE 3

The universal genetic code and human codon usage frequency. Information was adapted from www.nih.go.jp/~jun/research

Amino acid	Codons	Frequency	Amino acid	Codons	Frequency
Phe	TTT	0.43		GCC	0.40
	TTC	0.57		GCA	0.22
Leu	TTA	0.06		GCG	0.10
	TTG	0.12	Tyr	TAT	0.42
	CTT	0.12		TAC	0.58
	CTC	0.20	His	CAT	0.41
	CTA	0.07		CAC	0.59
	CTG	0.43	Gln	CAA	0.27
Ile	ATT	0.35		CAG	0.73
	ATC	0.52	Asn	AAT	0.44
	ATA	0.14		AAC	0.56
Met	ATG	1.00	Lys	AAA	0.40
Val	GTT	0.17		AAC	0.56
	GTC	0.25	Asp	GAT	0.44
	GTA	0.10		GAC	0.56
	GTG	0.48	Glu	GAA	0.41
Ser	TCT	0.18		GAG	0.59
	TCC	0.23	Cys	TGT	0.42
	TCA	0.15		TGC	0.58
	TCG	0.06	Trp	TGG	1.00
	AGT	0.14	Arg	CGT	0.09
	AGC	0.25		CGC	0.19
Pro	CCT	0.29		CGA	0.10
	CCC	0.33		CGG	0.19
	CCA	0.27		AGA	0.21
	CCG	0.11		AGG	0.22
Thr	ACT	0.23	Gly	GGT	0.18
	ACC	0.38		GGC	0.33
	ACA	0.27		GGA	0.26
	ACG	0.12		GGG	0.23
Ala	GCT	0.28			

usage in humans. However, different classes of genes use different codons. As the human genome project provides more sequence data, considerations such as codon usage will become clearer. A more complicating fact is that some gene families share structural motifs rather than protein sequence motifs. Presently, computational methods are being developed to address the identification of such families. It is not yet clear how the structural families relate to DNA sequence.

Decreasing the complexity of a targeted gene family fragment pool can be accomplished using anchored bases as described for targeting a repeated sequence. Alternatively, the T-primer can be designed to anneal only to a subset of the gene family sequence.

TGDD analyzes data rich Gaussian patterns

It should be noted that our TGDD products are viewed as data-rich Gaussian distributions, rather

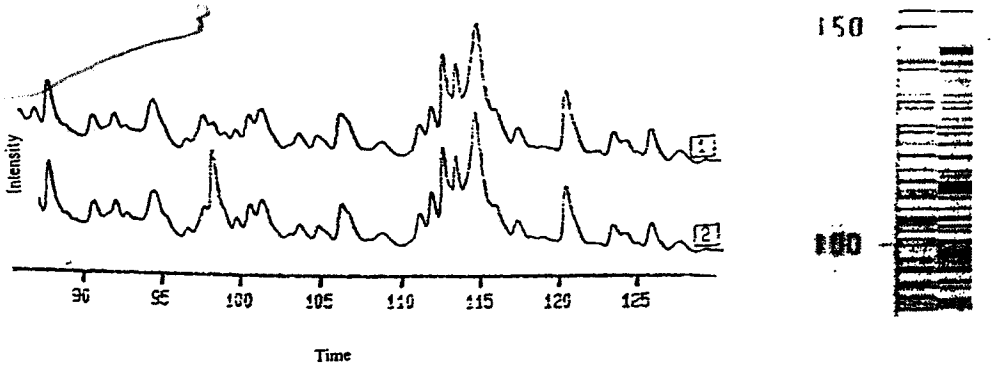


Fig. 8. Comparison of a high resolution TGDD fingerprint with low resolution TGDD fingerprint of an alleged MZ twin pair (Method II). Most implementations of DD by others use the low-resolution method of analysis. The $(CAG)_n$ containing fragments were targeted using oligonucleotide 8 (Table 1).

than low-resolution banding pattern. The high-resolution Gaussian analysis eliminates many false differences between samples. In Fig. 8, low and high-resolution analysis of the same data is shown. It is quite clear that many differences suggested in the low-resolution analysis do not survive high-resolution scrutiny.

Genomic DNA discordance within MZ twin pairs

The reliable de novo detection of genomic differences between MZ twin pairs depends strongly upon adherence to technical details described in materials and methods of this paper and elsewhere (Broude et al., 1997, 1999; Foulon et al., manuscript in preparation; Bouchard et al., manuscript in preparation). A variety of conditions were tested in order to optimize TGDD including: number of PCR cycles, primer concentrations, sample concentrations, and DNA extraction techniques. It was found that one of the most important factors is the matching of the DNA concentrations (see methods for details).

The existence of genomic differences between MZ twins is well established. Here, we describe a method that allows for the de novo detection of genomic differences in the absence of any knowledge of causation or location in the genome. TGDD also allows a quantitative assessment of genomic differences between individuals. An example of genomic discordances within a monozygotic twin pair is shown in Fig. 9. Here, the focus was directed on and nearby $(CAG)_n$ repeating sequences. Changing anchors on the T-primer and A-primers (oligonucleotides 6–8 and 10–14 respectively, Table 1) allowed a search through the pool of $(CAG)_n$ containing restriction fragments. This search was done until differences were identified. The fragments of interest were then picked from a high percentage agarose gel and reamplified. Figure 10 shows the isolated 250 and 390 bp fragments that will be sequenced.

Zygosity testing

As previously stated (see above), there is no "gold standard" zygosity test. All conventional methods are error prone. The zygosity of the twin samples studied here was determined to be MZ through conventional methods by other researchers. For example, the zygosity of the twin pair presented in Fig. 8 was determined by serology (22 antigens) in the following systems: ABO, Rh, P, Kell, Duffy, Kidd, and Lewis (Feinleib et al., 1977).

Ongoing studies are examining TGDD's application in zygosity testing. TGDD can effectively present a comparison of over 100 loci from a twin pair simultaneously. Our studies have shown that

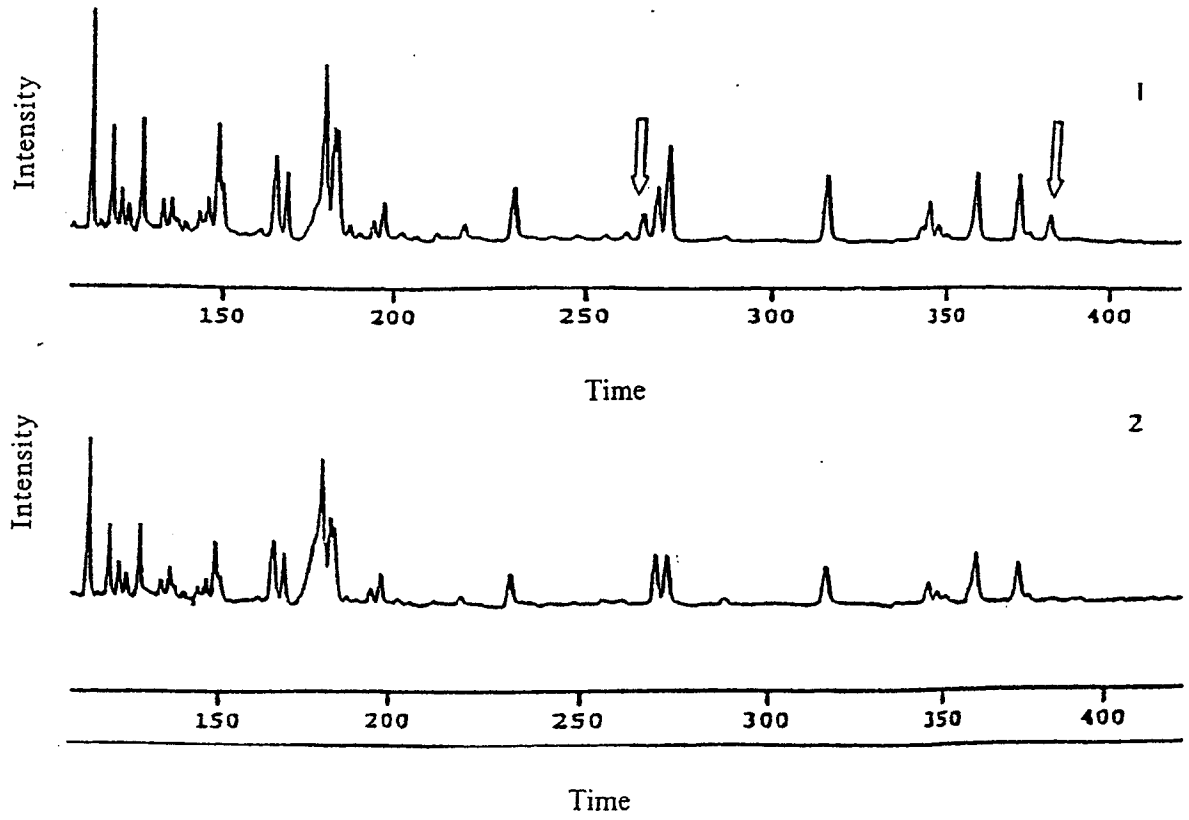


Fig. 9. TGDD (Method II) comparison between an alleged monozygotic twin pair. The A-primer was ST19*HaeIII*AC and the T-primer was $(CTG)_6A$ (oligonucleotides 14 and 6 respectively; Table 1). Arrows indicate discordant fragments.

at (CA6)n targets siblings and DZ twins have a higher level of variation than MZ twins. An example of TGDD targeting of an alleged MZ twin pair and a sibling pair is shown in Fig. 11(a) and (b) respectively. The exact variations in the level of genomic DNA discordance between MZ twin pairs and siblings is not yet fully understood. However, a preliminary assessment of our results indicates the presence of at least one incorrect zygosity determination.

Test/retest reliability of TGDD

Nine of the twelve MZ twin pairs analyzed by TGDD were found to have distinct and reproducible differences. The reproducible differences were seen when the samples were run in quadruplicate and also when different sets of experiments using the same samples were compared. The final conformation of the differences is done after obtaining the sequence of the restriction fragments. Once the sequence is obtained, a new primer that is specific for the difference is designed and used in a PCR. Such a conformation of a TGDD identified difference in a MZ twin pair is shown in Fig. 12. After the conformation of the difference, the two sequences are compared in order to understand how the difference arose and its impact on phenotype. For instance, the samples used in ongoing studies are discordant for schizophrenia or hypertension. Hence, a genomic discordance may impact an expression of these diseases.

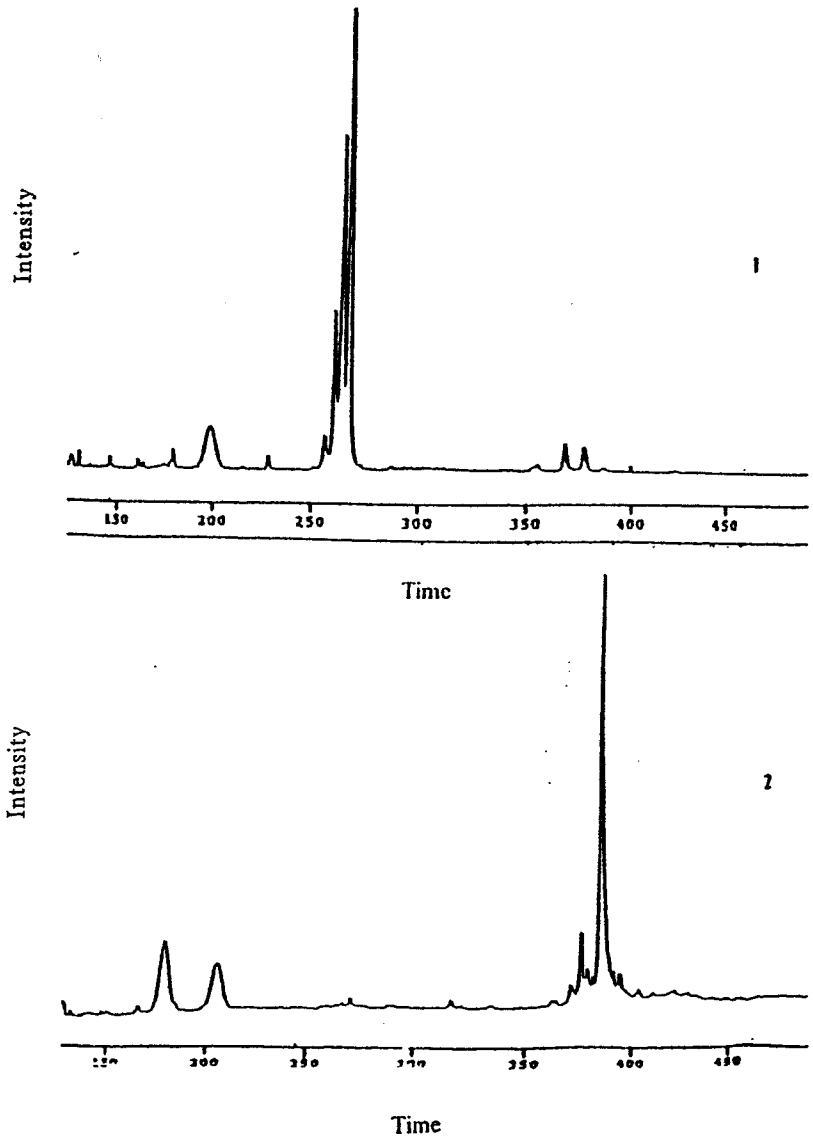


Fig. 10. The isolation of the discordant fragments shown in Fig. 9. The fragments were isolated and reamplified before fractionation on the Alfexpress as discussed in the Materials and Methods.

Prospectus

Currently, TGDD is being used to study monozygotic twins with schizophrenia or hypertension. Other studies are establishing the level of genomic identity of MZ twins and (re)assessing zygosity determination of others. The results shown here demonstrate that zygosity testing should be done with a large number of DNA sequences since the genomes of MZ twin pairs are very similar but not identical. Our preliminary work has detected an unexpected amount of genomic discordance in monozygotic twins (Nguyen et al., manuscript in preparation).

Understanding how and when the genomic DNA differences arise is the subject of ongoing

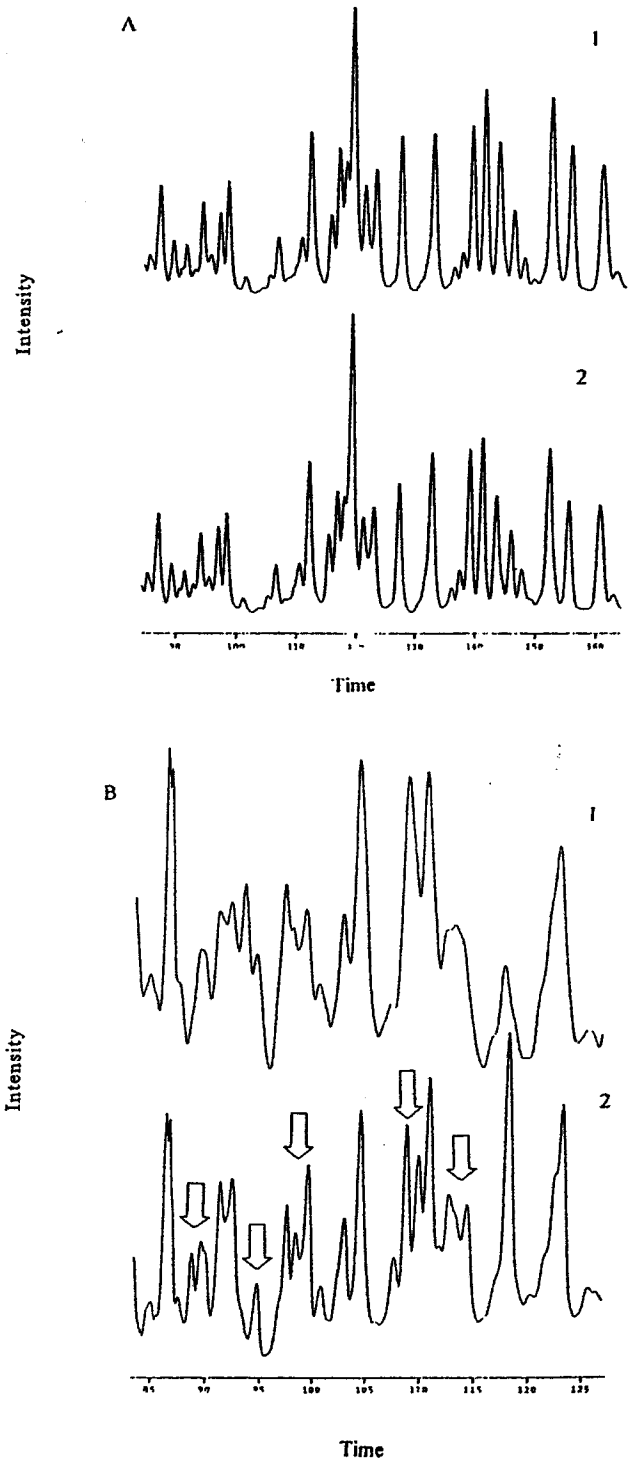


Fig. 11. A TGDD (Method II) comparison targeting $(CAG)_n$ repeat containing fragments using oligonucleotide 7 (Table 1). Samples being compared were (A) an alleged MZ twin pair and (B) a sibling pair.

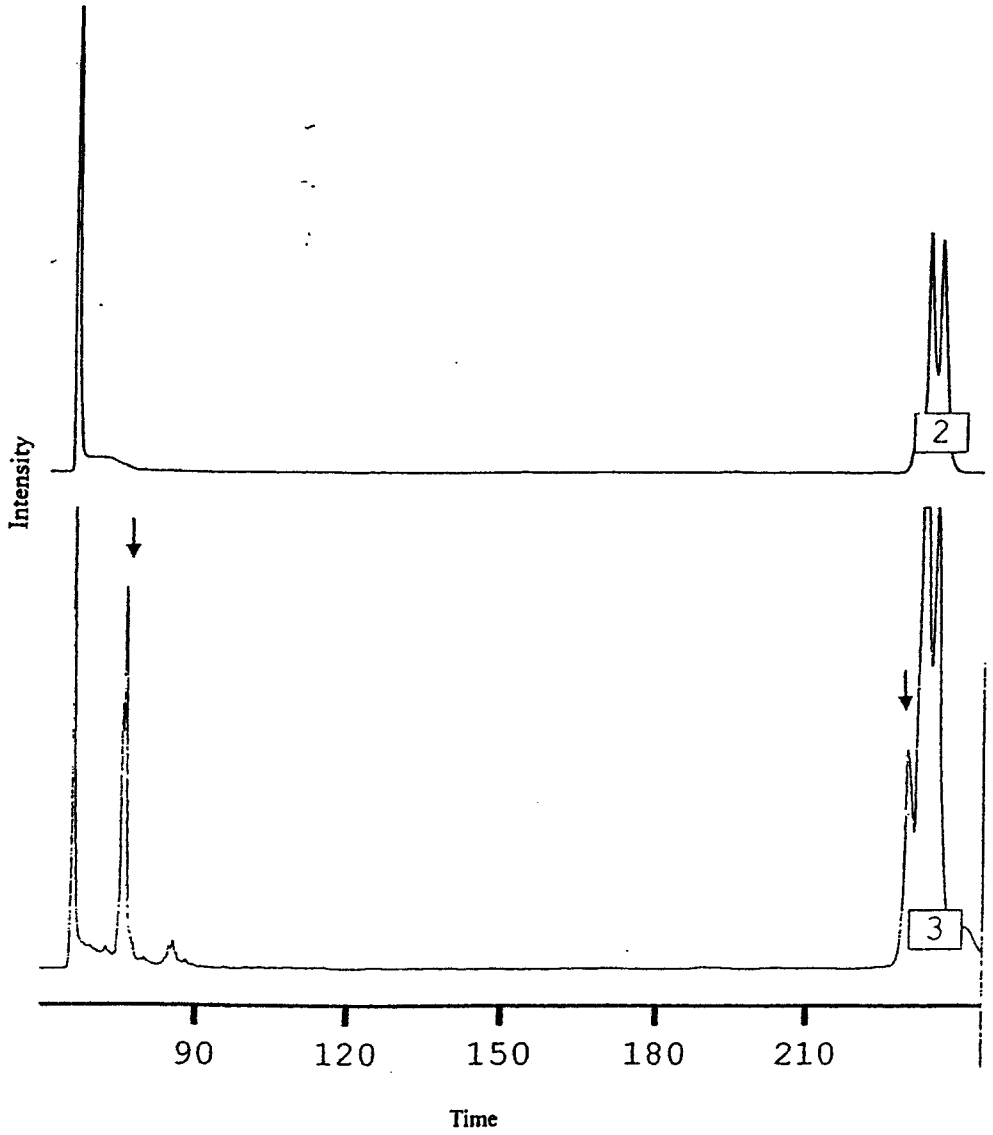


Fig. 12. Confirmation that TGDD identifies discordant genomic restriction fragments. A discordant fragment was isolated, cloned and sequenced (details in Nguyen et al., manuscript in preparation. Specific PCR primers were used to amplify homologous sequences from the genomic DNA of the MZ twin pair. Note comparison between the twin pair confirms differences, which are indicated by arrows.

experiments. It will be of special interest to determine the impact of these differences on phenotype, especially on disease discordance. The results also question the interpretation of disease discordance due to environmental factors. A great deal of the previous and ongoing research is focused on developing a quantitative method with a minimum of false positive and false negatives. This has led to the development of a model TGDD system using genomes whose entire sequence is known (Bouchard et al., manuscript in preparation). The model systems allowed experimental TGDD

results to be compared to theoretical results. Thus, errors can be identified and eliminated.

Also underway is work focused on developing automated methods of analysis. This will allow comparisons between a highly complex mixture of fragments efficiently. These methods are also being tested in the model system.

Our long term goal is to analyze TGDD with DNA array microchip technology. Here, pools of fragments can be sorted by unique sequence adjacent to the target sequence by hybridization to DNA arrays. The use of targeting should greatly enhance this technology.

Acknowledgements

The authors would like to thank Haralambos Gavras, E. Fuller Torrey, Louis Keith, Terry Reed, David Shepro, and Karen Pimpis and the NHLBI twin study for their support. This work was supported by grants DOA (DAMD17-94-J-414) and NIH (1P50 HL55001) to CLS.

References

Akane, A., Matsubara, K., Shiono, H., Yamada, M. and Nakagome, Y. (1991) Diagnosis of twins zygosity by hypervariable RPLP markers. *Am. J. Med. Genet.*, 41, 96–98.

Broude, N.E., Chandra, A. and Smith, C. L. (1997) Differential display of genome subsets containing specific interspersed repeats. *Proc. Natl. Acad. Sci. USA*, 94: 4548–4553.

Broude, N.E., Storm, N., Malpel, S. and Smith, C.L. (1999) A PCR-based Targeted genomic and cDNA differential display method. *Genetic Analysis (Biomolecular Engineering)*, 15, 51–63.

Derom, C., Vlietinck, R., Derom, R., Van Den Berghe, H. and Thiery, M. (1987) Increased monozygotic twinning rate after ovulation induction. *Lancet*, 1: 1236–1238.

Feinleib, M., Garrison, R.J., Fabsitz, R., Christian, J.C., Hrubec, Z., Borhani, N.O., Kannel, W.B., Rosenman, R., Schwartz, J.T. and Wagner, J.O. (1977) The NHLBI twin study of cardiovascular disease risk factors: methodology and summary of results. *Am. J. Epidemiol.*, 106(4): 284–285.

Grothues, D., Canter C.R. and Smith, C. L. (1993) PCR amplification of megabase DNA tagged random primers (T-PCR). *Nucleic Acid Research*, 21: 1321–1322.

Hall, J.G. (1996) Twinning: mechanisms and genetic implications. *Curr. Opin. Genet. Dev.*, 6: 343–347.

Hill, A.V.S. and Jeffreys, A. (1985) Use of minisatellite DNA probes for determination of twin zygosity at birth. *Lancet*, 2: 1394–1395.

Jansen, G., Willems, P., Coerwinkel, M., Nillesen, W., Smeets, H., Vits, I., Howeler, C., Brunner, H. and Wieringa, B. (1994) Gonosomal mosaicism in myotonic dystrophy patients: involvement of mitotic event in $(CTG)_n$ repeat variations and selections against extreme expansion in sperm. *Am. J. Hum. Genet.*, 54: 575–585.

Kruyer, H., Mila, M., Glover, G., Carbonell P., Ballesta, F. and Estivill, X. (1994) Fragile X syndrome and the $(CGG)_n$ mutation: two families with discordant monozygotic twins. *Am. J. Hum. Genet.*, 54: 437–442.

Liang, P. and Pardee, A.B. (1992) Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*, 257: 967–971.

Lisitzyn, N. and Wigler, M. (1993) Cloning the differences between two complex genomes. *Science*, 259: 946–951.

Kallioniemi, A., Kallioniemi, O.P., Sudar, D., Rutovitz, D., Gray, J. W. and Waldman, F. (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, 258: 818–821.

Lavrentieva, I., Broude, N.E., Lebedev, Y., Gottesman, I.I., Lukyanov, S.A., Smith, C.L. and Sverdlov, E.D. (1999) High polymorphism level of genomic sequences flanking insertion sites of human endogenous retroviral long terminal repeats. *FEBS Lett.*, 443: 341–347.

Machin G.A. (1996) Some causes of genotypic and phenotypic discordance in monozyotic twin pairs. *Am. J. Med. Genetics*, 61: 216–228.

Oliveira, R.P., Broude, N.E., Macedo, A.M., Cantor, C.R., Smith, C.L. and Pena, S.D. (1998) Probing the genetic population structure of *Trypanosoma cruzi* with polymorphic microsatellites. *Proc. Nat. Acad. Sci.*, 95: 3776–3780.

Reyniers, E., Vits, L., De Boulle, K.,Van Roy, B., Van Velzen, D., De Graaff, E., Verkerk. A.J. Jorens, H.Z., Darby, J.K., Oostra, B. and Willems., P.J. (1993) The full mutation in the FMR-1 gene of male fragile X patients is absent in their sperm. *Nat. Genetics*, 4: 143146

Siebert, P.D., Chenchik, A., Kellogg, D.E., Lukyanov, K.A. and Lukyanov, S.A. (1995) An improved PCR method for walking in uncloned genomic DNA. *Nucleic Acids Res.*, 23: 1087–1088.

Smith, C., Klco, S., Zhang, T., Fang, H., Rafael, O., Wang, D., Bremer, M. and Lawrance, S. (1993) Analysis of megabase DNA using pulsefield gel electrophoresis techniques. *Methods in Molecular Genetics*, 2: 155–175.

Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995). Serial Analysis of gene expression. *Science*, 270: 484–487.