PUBLISHED IN ASSOCIATION WITH
COLD SPRING HARBOR LABORATORY

Direct genomic selection

Stavros Bashiardes, Rose Veile, Cynthia Helms, Elaine R Mardis, Anne M Bowcock & Michael Lovett

Department of Genetics, Washington University School of Medicine, Box 8232, 4566 Scott Avenue, St. Louis, Missouri 63110, USA. Correspondence should be addressed to M.L. (Lovett@genetics.wustl.edu).

Searching for genetic variants and mutations that underlie human diseases, both simple and complex, presents particular challenges. In the case of complex diseases, these searches generally result in a single nucleotide polymorphism (SNP), or set of SNPs, associated with disease risk. Frequently, these SNPs lie outside the gene coding regions[1,2]. One is thus left in a quandary: do the detected SNPs represent the only genetic variation in the region or are there additional variants that might show even higher associations with disease risk? In the case of cancer, identification of mutations in tumor suppressor genes has also proved to be an arduous and frequently fruitless task[3]. The problem also arises in mouse genetics where mutational screens—for example, using ethylnitrosourea (ENU)—frequently require resequencing of large genomic regions to find a single base change[4]. The problem devolves to one of resequencing a large region of genomic DNA, usually of >100 kilobases (kb), from affected individuals or tissue samples to identify all sequence variants. Here, we describe modifications to direct selection[5,6] that allow for the rapid and efficient discovery of new polymorphisms and mutations in large genomic regions. Biotinylated bacterial artificial chromosome (BAC) DNAs are used in two rounds of hybridization selection with a target of total genomic DNA, and the selected sequences are amplified by the polymerase chain reaction (PCR) (Fig. 1). The procedure results in enrichments of 10,000-fold, in which ~50% of the resulting sequence-ready clones are from the targeted region (Box 1).

MATERIALS

REAGENTS

Genomic DNA, pooled isolates from individual patients or tissue samples
Restriction enzymes *Sau*3AI (Promega), *Hin*fI (Invitrogen) and *Eco*RI (Invitrogen)
DNA polymerases: T4 DNA polymerase (Invitrogen), *PfuUltra* HF DNA polymerase (Stratagene)
5× T4 DNA polymerase reaction mix: 5× T4 DNA polymerase buffer, 1 mM each dNTP
Oligonucleotide 3: 5′-CTCGAGAATTCTGGATCCTC-3′, Oligonucleotide 4: 5′-GAGGATCCAGAATTCTCGAGTT-3′
10× annealing buffer: 100 mM Tris-HCl (pH 7.5), 1 M NaCl, 10 mM EDTA
T4 DNA ligase (Roche)
QIAquick PCR purification kit (Qiagen)
Purified BAC DNA, from clone encoding region of interest
Biotin-16-dUTP (Enzo Biochemicals)
10× nick translation buffer: 500 mM Tris-HCl (pH 7.5), 100 mM $MgCl_2$, 50 mM DTT, 400 μM each dNTP
[α-^{32}P]dCTP (3,000 Ci/mmol) (Amersham)
DNA polymerase/DNase I (Roche)
Cot-1 DNA (Invitrogen)
2× hybridization buffer: 1.5 M NaCl, 40 mM sodium phosphate buffer (pH 7.2), 10 mM EDTA (pH 8), 10× Denhardt's, 0.2% SDS
Streptavidin-coated paramagnetic beads and magnetic bead separator (Dynabeads M-280 Streptavidin)(Dynal)
Streptavidin bead binding buffer: 10 mM Tris-HCl (pH 7.5), 1 mM EDTA (pH 8) and 1 M NaCl
dNTP solution (25 mM each) (Invitrogen)
p-AMP1 plasmid DNA (Invitrogen)
DH10B electrocompetent *Escherichia coli* (Invitrogen)

EQUIPMENT

Thermal cycler programmed with the desired amplification protocol
Sephadex G-50 spin columns (Roche)

Preparation of the genomic DNA and double-stranded linkers

PROCEDURE

1| Set up two reactions for digestion of genomic DNA:

Genomic DNA	10 μg	Genomic DNA	10 μg
*Sau*3AI	200 U	*Hin*fI	200 U
10× enzyme buffer	40 μl	10× enzyme buffer	40 μl
Water	to 400 μl	Water	to 400 μl

Incubate the two reactions at 37 °C for 3 h. Extract the reactions with phenol/chloroform (1:1, v/v), recover the DNA by precipitation with ethanol and dissolve the DNA pellets in 50 μl of water. Set aside 3-μl aliquots for use as controls below.
This combination of enzymes is used to compensate for fragment length representational differences that might be introduced by digestion with a single enzyme.
▲ **CRITICAL STEP**

2| Set up two reactions to fill in the overhang ends of the DNA digestion products:

*Sau*3AI-digested DNA	41 μl	*Hin*fI-digested DNA	41 μl
T4 DNA polymerase	20 U	T4 DNA polymerase	20 U
T4 DNA polymerase reaction mix	20 μl	T4 DNA polymerase reaction mix	20 μl
Water	to 100 μl	Water	to 100 μl

Incubate the reactions at 11 °C for 30 min. Combine the two reactions, extract with phenol/chloroform and recover the DNA by ethanol precipitation. Dissolve the pellet in 10 μl water (to give a final concentration of 2 μg/μl).
➨**TROUBLESHOOTING**

3| Anneal oligonucleotides 3 and 4 to create a double-stranded linker, by mixing the following:

Oligonucleotide 3 (1 μg/μl)	22.5 μl
Oligonucleotide 4 (1 μg/μl)	22.5 μl
10× annealing buffer	5 μl
Water	to 50 μl

Heat the reaction at 65 °C for 10 min; then allow it to cool at 15–25 °C for 2 h.
The oligonucleotides in this protocol are designated 3 and 4 to distinguish them from oligonucleotides 1 and 2 used in the related protocol for direct selection of cDNAs[6].

4| Purify the double-stranded linker by column chromatography through a Sephadex G-50 spin column; then concentrate the purified linker solution by lyophilization to a concentration of 2 μg/μl.

Ligation of linkers to genomic DNA fragments

5| Set up the following reaction to ligate the linkers to genomic DNA fragments
Incubate the reaction at 14 °C overnight.

Annealed linkers from Step 4 (20 μg)	10 μl
Genomic DNA from Step 2 (10 μg)	5 μl
T4 DNA ligase	10 U
10× ligation buffer	2 μl
Water	to 20 μl

➨**TROUBLESHOOTING**

6| Adjust the reaction volume to 500 µl with water and purify the ligated genomic DNA using a QIAquick PCR purification kit. Store the purified DNA at a concentration of 1 µg/µl.

7| To incorporate biotinylated residues into the BAC DNA, prepare the following nick translation reaction:

Biotinylation and blocking of the BAC DNA

Purified BAC DNA	100 ng
Biotin-16-dUTP (40 µM stock)	2 µl
10× nick translation buffer	2 µl
[α-^{32}P]dCTP (3,000 Ci/mmol)	1 µl
DNA polymerase/DNase I	5 U
Water	to 20 µl

Incubate the reaction at 4 °C for 2 h.
The isotope is included as a tracer to confirm that the biotinylation reaction has proceeded efficiently and to confirm binding of the BAC DNA to streptavidin-coated magnetic beads.
▲ **CRITICAL STEP**

8| Purify the biotinylated products by column chromatography through a Sephadex G-50 spin column and lyophilize the resulting solution to dryness.

9| Dissolve the biotin-labeled DNA pellet in 5 µl of Cot-1 DNA (2 µg/µl) and transfer to a 200-µl PCR tube. Overlay the solution with mineral oil, denature by incubation at 95 °C for 5 min, and incubate at 65 °C for 15 min. Add 5 µl of 2× hybridization buffer and incubate further at 65 °C for 6 h.
▲ **CRITICAL STEP**

Primary selection and capture of hybrids

10| Deliver 1 µg of linkered genomic DNA from **Step 6** (1 µg/µl) in 5 µl of water into a 200-µl PCR tube and overlay with mineral oil. Denature the DNA at 95 °C for 5 min and incubate at 65 °C for 15 min. Add 5 µl of 2× hybridization buffer and transfer the entire sample to the tube containing the Cot-1–blocked BAC DNA from **Step 9**. Incubate the hybridization reaction at 65 °C for a further 70 h ($C_0t_{1/2}$ of ~25).

11| In a 1.5-ml microcentrifuge tube, wash 6.7×10^7 (100 µl) of streptavidin-coated beads twice in 200 µl Streptavidin bead binding buffer. After each wash, remove the beads from the binding buffer with a magnetic separator.

12| Resuspend the beads in 150 µl Streptavidin bead binding buffer and add the hybridization buffer to the bead suspension. Carry out binding at 15–25 °C for 30 min with periodic gentle mixing to ensure the beads remain evenly suspended.

13| Remove the beads from the binding buffer using a magnetic separator and discard the supernatant. Wash the beads once, at 15–25 °C for 15 min, in 1 ml of 1× SSC with 0.1% SDS, and then three times, each at 65 °C for 15 min, in 1 ml 0.1× SSC with 0.1% SDS.

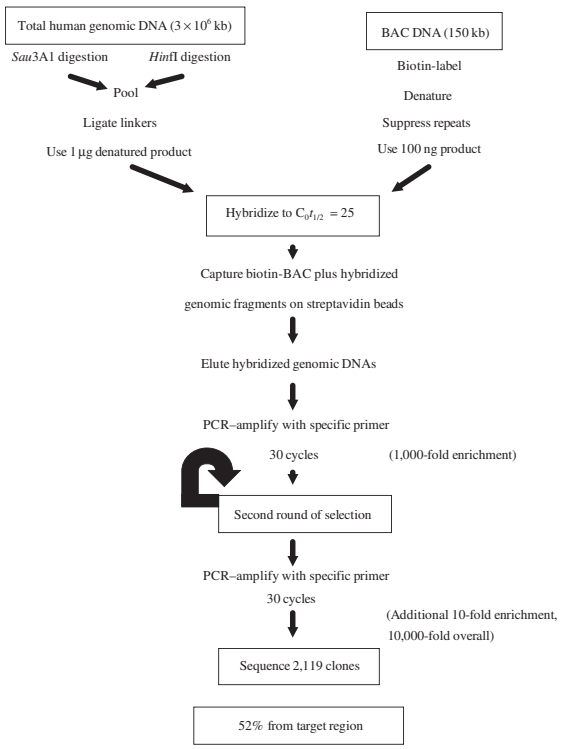

Figure 1 | A flow diagram of the direct genomic selection process. Specific steps are discussed in the text. We illustrate the proof of concept of the method, also described in **Box 1**, using a 150-kb BAC from human chromosome 17.

14| To elute the genomic DNA hybridized to the BAC, add 100 µl of 0.1 M NaOH to the beads and incubate at 15–25 °C for 10 min. Remove the beads from the elution mixture using a magnetic separator.

15| Transfer the supernatant to a fresh 1.5-ml microcentrifuge tube containing 100 µl of 1 M Tris-HCl (pH 7.5) and desalt the solution by spin-column chromatography through Sephadex G-50 resin.

Amplification of the primary selected DNA

16| Set up three separate amplification reactions in 200-µl PCR tubes:

Reagents	Reaction 1	Reaction 2	Reaction 3
Template: primary selection material (Step 15)	1 µl	5 µl	10 µl
Oligonucleotide 3 (200 ng/µl)	1 µl	1 µl	1 µl
dNTPs (25 mM each)	0.4 µl	0.4 µl	0.4 µl
10× *PfuUltra* HF DNA polymerase reaction buffer	5 µl	5 µl	5 µl
PfuUltra HF DNA polymerase	2.5 U	2.5 U	2.5 U
Water	to 50 µl	to 50 µl	to 50 µl

▲CRITICAL STEP

17| Amplify the three reactions according to the following program:

Cycle number	Denaturation	Annealing	Polymerization
1	2 min at 95 °C		
2–31	30 s at 95 °C	30 s at 55 °C	1 min at 72 °C

18| To determine which aliquot of template (1 µl, 5 µl or 10 µl) yields the best range of amplified products, analyze the reaction products by agarose gel electrophoresis, including as control 1 µl of the DNA sample set aside in **Step 1**.
It is advisable to transfer the DNA from the gel by Southern blotting and hybridize the membrane with a single-copy probe to assess enrichment of the template. The primary selection should yield approximately one thousand-fold enrichment.
▲CRITICAL STEP

19| Select the volume of template identified in **Step 18** and use this volume to prepare a series of eight amplification reactions according to the conditions described in **Step 16**. Carry out the amplification as described in **Step 17**.

20| Purify the amplification products using a QIAquick PCR purification kit. Pool the eluted samples and determine the concentration of amplified primary selected DNA by spectrophotometry.

21| Remove a volume of DNA equivalent to 1 µg and reduce this volume to 5 µl in a speed vacuum concentrator. Set aside 1 µl (at least 200 ng) of the primary selected material for comparison with the secondary selection products.

Further enrichment by secondary selection

22| Set up a reaction to biotinylate and block 100 ng of purified BAC DNA with Cot-1 DNA, following **Steps 7–9**.

23| Transfer the primary selected DNA (from **Step 21**) to a 200-µl PCR tube and overlay with mineral oil. Denature the sample at 95 °C for 5 min and then incubate at 65 °C for 15 min. Add 5 µl of 2× hybridization buffer and transfer the sample to the tube containing the Cot-1–blocked BAC (from **Step 22**). Incubate the hybridization reaction at 65 °C for 70 h.

24| After hybridization, bind the sample to streptavidin-coated magnetic beads and wash; recover the selected DNA as described for the primary selection (**Steps 11–15**).

Amplification of secondary selected material

25| Set up three amplification reactions as described in ∫. Transfer 1 μl of the secondary selected DNA (**Step 24**) into one tube, 5 μl into the second and 10 μl into the third. Add water to each tube to a final volume of 50 μl and amplify the three reactions according to the program described in **Step 17**.
▲CRITICAL STEP

26| To determine which aliquot of template (1 μl, 5 μl or 10 μl) yields the best range of amplified products, analyze the products by agarose gel electrophoresis. Include as controls 1 μl of the starting genomic DNA (**Step 1**) and 100 ng of primary selected DNA (**Step 21**).
It is advisable to transfer the DNA from the gel by Southern blotting and hybridize the membrane with a single-copy probe to assess enrichment of the template. The secondary selection should yield an approximately tenfold further enrichment.
▲CRITICAL STEP

27| Select the volume of template identified in **Step 26** and prepare a series of eight amplification reactions according to the conditions described in **Step 16**. Amplify the reactions according to the following program:

Cycle number	Denaturation	Annealing	Polymerization
1	2 min at 95 °C		
2–21	30 s at 95 °C	30 s at 55 °C	1 min at 72 °C

If possible, limit the number of PCR cycles to 20 or fewer to minimize the chances of PCR-induced sequence errors.
▲CRITICAL STEP

28| Purify the amplification products using a QIAquick PCR purification kit. Pool the eluted samples and determine the concentration of amplified secondary selected DNA by spectrophotometry.

29| Set up a reaction to digest the secondary selection material:

EcoRI digestion and purification of the amplified products

Purified DNA (**Step 28**)	500 ng–1 μg
*Eco*RI	20 U
10× React 3 buffer (Invitrogen)	3 μl
Water	to 30 μl

Incubate the reaction at 37 °C for 3 h.
The linker (annealed oligonucleotides 3 and 4) is designed with an internal EcoRI site to facilitate cloning of amplified products after selection.

30| Resolve the digested DNA by electrophoresis through a 1% TAE agarose gel. Isolate the selected material ranging from 300 base pairs (bp) to 2 kb by cutting out and removing the appropriate portion of the gel. Purify the sample from the gel using a QIAquick gel extraction kit and quantify the eluted material by spectrophotometry.

31| Transfer a volume of DNA (from **Step 30**) that corresponds to 100 ng into a 200-μl PCR tube and reduce the volume to 3 μl in a speed vacuum concentrator.

32| Clone the DNA fragments into the *Eco*RI site of pAMP1 and transform the recombinant plasmids into *E. coli* strain DH10B.
The transformants may now be analyzed by sequencing to identify the presence of sequence variations within the cloned region.

TROUBLESHOOTING TABLE

PROBLEM	SOLUTION
Steps 2 and 5 The material fails to amplify.	Test 100 ng of the genomic DNA containing linkers in a PCR, as described in **Steps 16–17**. As negative control, run a parallel PCR with 1 μl digested DNA without linkers (from **Step 1**). The DNA with linkers should yield a smear of products on an agarose gel with an average size of ~500 bp, whereas the DNA without linkers should not. If the DNA containing linkers fails to amplify, it is likely that there has been a failure of the fill-in reaction (**Step 2**) or of the ligation (**Step 5**). Repurify the genomic DNA and start from the beginning of the protocol.

CRITICAL STEPS

Step 1 Alternatively, random shearing, size selection and repair of genomic DNAs may easily be incorporated into this protocol for regions that lack a good distribution of restriction enzyme sites.

Steps 7 and 9 Check that the BAC DNA has been efficiently biotinylated by removing a small aliquot (1% of the total) and testing its binding to streptavidin beads. Monitor the radioactive tracer on the beads and in the supernatant with a Geiger counter or by Cherenkov counting. The ratio of bound to unbound counts should be at least 8:1. Do not proceed if this ratio is not achieved. If the ratio is lower, either repurify the BAC DNA by extraction with phenol/chloroform and purification on Sephadex spin columns, or increase the ratio of biotin-16-dUTP to unlabeled nucleotide in the labeling reaction (**Step 7**). However, there are upper limits to this latter solution, as too many biotin moieties bound to the BAC DNA can sterically hinder hybridization (**Step 9**).

Steps 16, 25 and 27 It is very important to use the highest-quality and highest-fidelity thermostable polymerase. It is also advisable, if possible, to limit the number of amplification cycles, to reduce the possibility that resulting DNA sequences may be riddled with PCR-induced sequence errors rather than true genomic DNA sequence variants.

Steps 18 and 26 At this point one may either qualitatively or quantitatively assess the degree of enrichment (or one may proceed with blind faith). To gain a qualitative estimate, hybridize the Southern blot of genomic DNA and primary selected material with a radiolabeled single-copy DNA fragment derived from within the genomic region carried in the BAC (for example, a PCR fragment of a few hundred base pairs). The resulting autoradiograph should show a very high degree of enrichment in the primary selected material (~1,000-fold). If this is not the case, do not proceed with the secondary selection; start from the beginning of the protocol and double check DNA purity, concentrations, and bead binding parameters. A quantitative assessment of enrichment can be achieved by deriving a few thousand molecular clones from the primary selected material (**Step 21**) and conducting colony lift hybridizations with the single copy probe.

SOURCE

This protocol is based on new developments to "Direct selection of cDNAs with large genomic DNA clones," in *Molecular Cloning: A Laboratory Manual* (eds. Sambrook, J. & Russell, D.W.) Chapter 11 Protocol 4, pages 11.98–11.106 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA, 2001).

1. Helms, C. *et al.* A putative RUNX1 binding site variant between SLC9A3R1 and RAT9 is associated with susceptibility to psoriasis. *Nat. Genet.* **35,** 349–356 (2003).
2. Pajukanta, P. *et al.* Familial combined hyperlipidemia is associated with upstream transcription factor 1 (USF1). *Nat. Genet.* **36,** 371–376 (2004).
3. Balmain, A., Gray, J. & Ponder, B. The genetics and genomics of cancer. *Nat. Genet.* (Suppl.) **33,** 238–244 (2003).
4. Kile, B.T. *et al.* Functional genetic analysis of mouse chromosome 11. *Nature* **425,** 81–86 (2003).
5. Lovett, M., Kere, J. & Hinton, L.M.. Direct selection: a method for the isolation of cDNAs encoded by extended genomic regions. *Proc. Natl. Acad. Sci. USA* **88,** 9628–9632 (1991).
6. Lovett, M. Direct selection of cDNAs with large genomic DNA clones. In *Molecular Cloning: A Laboratory Manual* Edn. 3 Vol. 2 (eds. Sambrook., J. & Russell, D.W.) 11.98–11.133 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA, 2001).
7. Hacia, J.G. Resequencing and mutational analysis using oligonucleotide microarrays. Nat. Genet. 21, 42-49 (1999)
8. Faham, M., Baharloo, S., Tomitaka, S., DeYoung, J. &

BOX 1 PROOF OF PRINCIPLE

Many methods are available for typing known SNPs but very few for discovering new sequence variants in a cost-effective and rapid manner. Promising new technologies, such as sequencing by hybridization[7] and mismatch repair detection[8], have not as yet been widely adopted. Currently, two well-established but inefficient routes to this end are based on mature and widely available technologies. The first, used in many studies (see, for example, refs. 9–11), consists of designing PCR primer sets at suitable overlapping intervals across the targeted region (not only across exons) and using these to selectively amplify the contig from the genomic DNAs of individuals affected with a condition of interest. In theory, this approach appears direct; in practice it is expensive and difficult to optimize. The second route is to construct an entire genomic DNA library from affected individuals, screen for the particular contig in question and shotgun-sequence all of these clones. This approach is time consuming and expensive.

We describe here modifications to a direct selection method originally developed and widely used for cDNA isolation[5,6]. The strategy (summarized in **Fig. 1**) consists of using selected BAC clones from the reference human genomic DNA library as a hybridization 'hook' to isolate corresponding sequences from a pool of genomic DNAs. As proof of this concept, we embarked upon the identification of previously undescribed sequence variations within the *RAPTOR* gene on human chromosome 17q. SNPs adjacent to this gene have been associated with an increased risk for psoriasis[1]. The peak area of association covers ~100 kb; however, it was unclear whether additional (perhaps more highly associated) SNPs or disease-causing variants exist elsewhere in this gene.

We applied our procedure to a pool of genomic DNAs from 14 patients with psoriasis. A BAC clone (RP11-582A10) of 150 kb covering the region of peak association was labeled with biotin-dUTP, then denatured and suppressed for highly repetitive elements with Cot-1 DNA. Genomic DNA from 14 individual psoriatic patients was pooled in equal proportions, digested and hybridized to the BAC clone as described in the protocol. To validate this primary round of selection the clones were screened by hybridization with a 300-bp single-copy DNA fragment from the BAC, indicating that the targeted sequence was present in the primary selection at a frequency of 1 in 10,000 clones (corresponding to an enrichment of ~1,000-fold). This observation was confirmed by sequencing ~100 randomly picked clones, and 1 in 20 clones were shown to derive from the original target region. To increase the enrichment, the (uncloned) primary selection was subjected to another round of hybridization selection. Screening of the resulting clones with the same 300-bp fragment detected 10 in 10,000 clones, implying a total enrichment of ~10,000-fold. To directly test the levels of enrichment (and background clones), we sequenced 2,119 clones. Of these, 52% aligned with the BAC DNA sequence, 16% had poor alignments that could reflect the quality of the sequence and 34% either had no unequivocal genomic alignments (short or repetitive sequences) or mapped elsewhere in the genome. The average *Eco*RI fragment length was 511 bp. The average coverage of the BAC sequence was threefold, with alignments throughout the BAC (although not across highly repetitive tracts).

With this sequence analysis we detected 69 previously described SNPs and more than 100 putative new SNPs across this 150-kb interval, which we are currently validating. We also detected a three-base deletion corresponding to nucleotides 79,419,439–79,419,441 within intron 4 of *RAPTOR* and a two-base insertion at position 79,351,708. The deletion and insertion were validated by sequencing the corresponding regions in all 14 patient DNAs within the psoriasis patient DNA pool. The deletion was present in the DNA of one patient and the insertion was present in the DNA of seven patients. Neither variant was present in the starting BAC DNA.

This modified application of direct selection results in the production of clones in a sequence-ready format (in plasmid vectors with inserts of ~500 bp). It results in frequencies of 'correct' clones of ~50% by random picking (a number we confirmed in additional selections using other BACs covering regions elsewhere in the genome). This frequency is price-compatible with a random picking and sequencing strategy. The detection of previously described SNPs, a new microdeletion and a new microinsertion indicates that this method should prove useful for the rapid enrichment and sequence-based detection of new sequence variants, mutations and genomic rearrangements.

Freimer, N.B. Mismatch repair detection (MRD): high-throughput scanning for DNA variations. *Hum. Mol. Genet.* **10,** 1657–1664 (2001).

9. Stengard, J.H. *et al.* Contributions of 18 additional DNA sequence variations in the gene encoding apolipoprotein E to explaining variation in quantitative measures of lipid metabolism. *Am. J. Hum. Genet.* **71,** 501–517 (2002).
10. Crawford, D.C. *et al.* Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *Am. J. Hum. Genet.* **74,** 610–622 (2004).
11. Fullerton, S.M. *et al.* The effects of scale: variation in the APOA1/C3/A4/A5 gene cluster. *Hum. Genet.* **115,** 36–56 (2004).