Microarray-based genomic selection for high-throughput resequencing

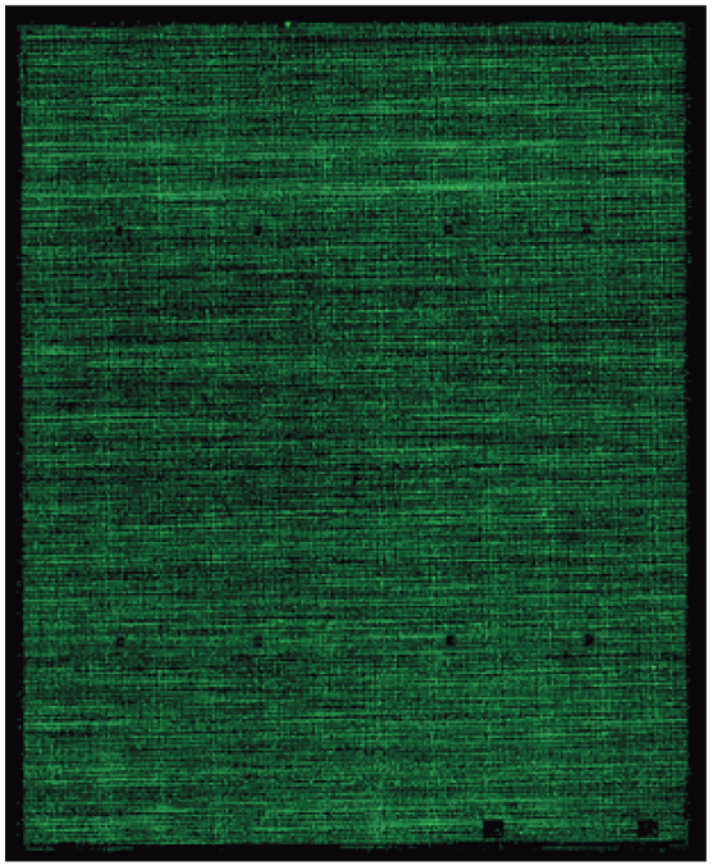
David T Okou, Karyn Meltz Steinberg, Christina Middle, David J Cutler, Thomas J Albert & Michael E Zwick

Supplementary figures and text:

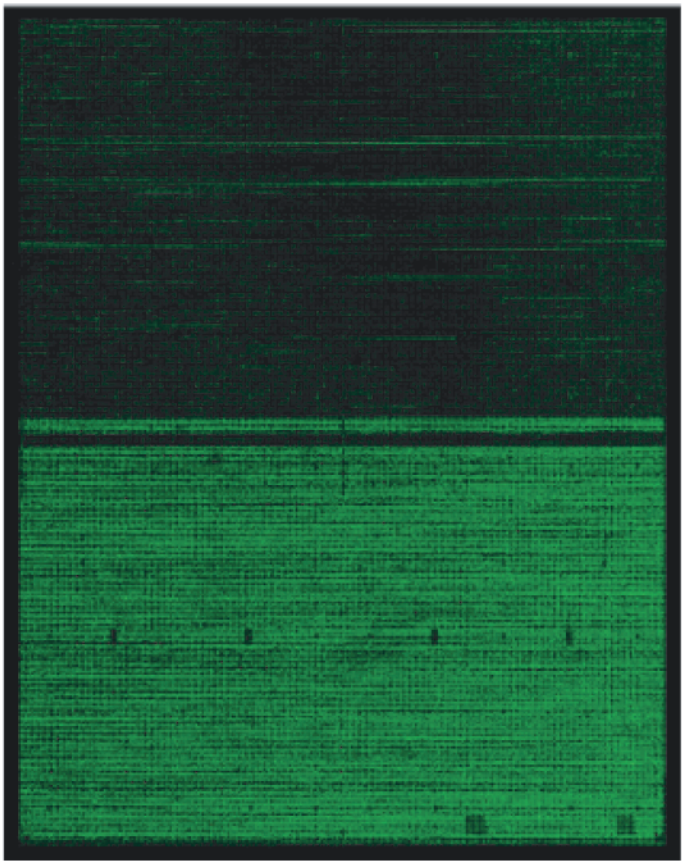
Note: Supplementary Data 1–4 are available on the Nature Methods website.

Supplementary Figure 1: Resequencing Array Images for 50kb Region

A.



B.

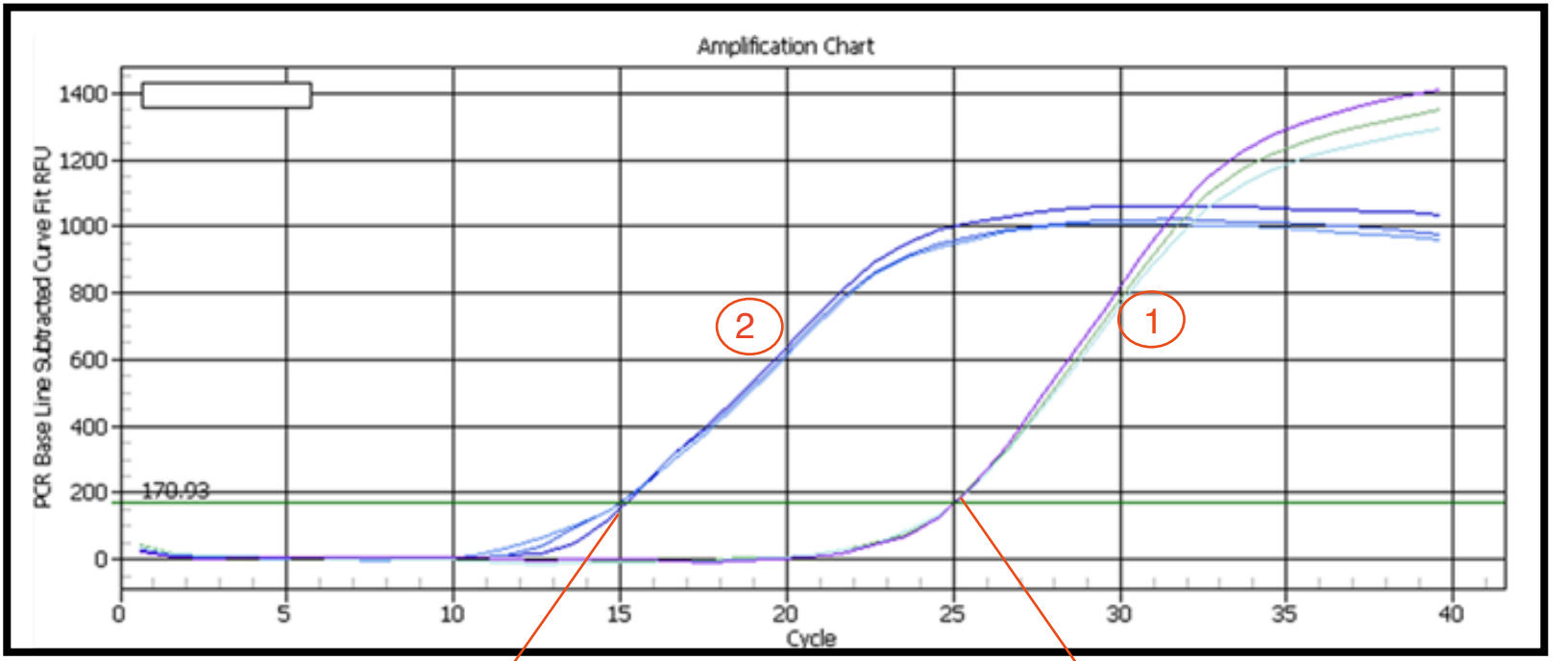


Supplementary Figure 1. RA hybridization results for TR91 (A) and DM316 (B) samples. The large absence of hybridization on the DM316 array is the result of a large deletion of much of the FMR1 locus.

Supplementary Figure 2: Results of Quantitative PCR Assay

Treatment 1: WGA Genomic DNA --------> Ligation ---------> LMPCR

Treatment 2: WGA Genomic DNA --------> Ligation ---------> GS Chip Hybridization ----------> Elution ----------> LMPCR



Supplementary Figure 2. Results of quantitative PCR assay measuring the extent of enrichment after a single round of microarray-based genomic selection (MGS). Treatment 1 consisted of a whole genome amplified sample that was passed through the entire MSG protocol, but never hybridized to an array. Treatment 2 consisted of a whole genome amplified sample processed through the entire MGS protocol. The DNA from treatment 2 has a cycle threshold of 15 while the cycle threshold for treatment 1 is 25. If we assume that DNA doubles every cycle then enrichment can be calculated by 2^N, with N equaling the difference between the cycle thresholds of the two treatments. In this case, this corresponds to approximately a 1000-fold increase in enrichment.

SUPPLEMENTARY METHODS

1. Array Design

We used the UCSC Table Browser function with repeats masked on the latest human genome build (March 2006) to identify the unique sequences within a selected genomic region[1]. The CGG repeat sequence of FMR1 from the human genome reference sequence was included in the design. Since genetic variants in regulatory elements away from the coding sequences may influence the expression of a gene [2], unique sequence upstream and downstream of the target genes were also be included. We then selected among the unique sequence to obtain ~50 Kb or 304 Kb of unique sequence. We excluded unique sequences 100 bp or less and in some cases, we added short (<100 bp) stretches of previously masked sequence, to avoid breaking up long stretches of genomic regions.

The FASTA format sequences were then provided to chip design engineers at Nimbelgen to select oligonucleotides for the microarray-based genomic selection (MGS) array. Standard bioinformatics filters that check for genomic uniqueness against an indexed human genome (15mers) were used to select capture oligos. The capture oligonucleotides were between 50 and 93 basepairs long and were designed to achieve optimal isothermal hybridization across the microarray. No other optimization of oligos was performed. For the 50 Kb region, there were four pairs of probes for every targeted base, while the 300Kb region has one pair of probes for every 1.5 targeted bases.

Resequencing arrays were designed from the FASTA format sequences provided to design engineers at Affymetrix (FMR1/FMR2) and NimbleGen (FMR1 only). Resequencing Arrays (RAs) query a given base by using overlapping oligonucleotide probes, tiled at a 1-basepair (bp) resolution. The oligonucleotide probes, referred to as features, are typically 25 basepairs long. Both the forward and reverse strands are interrogated, so sequencing a single base requires a total of 8 features. A set of four features contains oligonucleotides identical to the

forward reference strand, except at position 13 (the base to be queried), where there is either A, C, G, or T. The remaining four features are similarly designed for the complementary strand. When a labeled DNA sample, called a target, is hybridized to these eight features on the array, the two features complementary to the reference sequence (forward and reverse complement) will yield the highest signal. If, however, the target DNA contains a variant base at position 13, the two features complementary to that variant base will yield the highest signal. Given eight features for each base, interrogation of an L-length duplex strand would require 8L oligonucleotide probes.

2. Sample Selection

DNA samples were purchased from the Coriell Cell Repository (http://ccr.coriell.org) and included 10 individual genomes represented by two populations of different ancestry: a European descent (ED) population (n=5) selected from the Centre d'Etude du Polymorphism Humain (CEPH) panel with the Coriell Cell Repository numbers: NA07029, NA07048, NA10846, NA10851 and NA10860; and an African descent (AD) population (n=5) selected from the Hapmap with the Coriell Cell Repository numbers: NA18500, NA18503, NA18506, NA18515 and NA18521. MGS was replicated twice for each of the ten samples. Other samples used in this study were extracted from cell lines representing fragile X patients with either disease causing point mutation (A>T) at position 146825745 on the X chromosome (Tr91) or deletion (DM316) in the fragile X mental retardation (FMR1) gene [3,4] (Supplementary Table 1).

3. Adaptor and Primer Design

All oligonucleotides used in this project were obtained from Invitrogen Corp. The adaptor was prepared by annealing the forward (21 bp) and reverse (22 bp) oligonucleotides to generate a 21 bp dsDNA fragment with single and double base "T" overhangs at the 3 prime and 5 prime end respectively. Adaptor sequences used were CTCGAGAATTCTGGATCCTCT and

TTGAGCTCTTAAGACCTAGGAG. Annealing of the oligos was performed by mixing both oligonucleotide to a final concentration of 1.5 μg/μl each oligo, heating to 95'C for 10 minutes in a heating block, turning off the heating block and allowing the mixture to slowly cool back to room temperature. The primers used for the enrichment were made by preparing a 20 μM of each oligonucleotide used for the adaptor.

4. Genomic DNA preparation

Whole genome amplification was performed on 100 ng of genomic DNA using the RepliG Kit (Qiagen Inc.). Following amplification, the unpurified samples were quantified using a spectrophotometer (NanoDrop). Twenty-five micrograms of each sample was aliquoted into sterile Eppendorf tubes for a final concentration of 100 ng / μl (250 μl).

5. Target DNA isolation

Samples were sonicated (Misonix sonicator 3000) in eppendorf tubes with a microtip probe using the following parameters: 3 pulses of 30 seconds each with 2 minutes of rest and a power output level of two. After fragmentation, approximately 750 ng of each sample was run on a 1.5% TAE agarose gel against 750 ng of a 1 Kb plus ladder to verify that fragments average 300 bp in size. The samples were then dried down in a SpeedVac at medium heat to 47 μl (75° C).

6. Repairing Ends of Sheared DNA

To the 47 μl fragmented DNA we added 8 μl of dNTPs (2.5 mM, TaKaRa), 8 μl of 10X T4 DNA Polymerase Buffer (NEB), 1 μl of 100X BSA (NEB), 1 μl 100mM ATP, 14 μl of T4 DNA Polymerase (3U/μl, NEB), and 1 μl of T4 Polynucleotide Kinase (10U/μl, NEB). We then incubated in a thermocycler at 12°C for 20 minutes followed by 37°C for 30 minutes and 70°C for 5 minutes. After incubation we directly added 2 μl of 10X T4 DNA Polymerase Buffer

(NEB), 2 μl 100mM dATP (Sigma), 3 μl of 50mM MgCl2, 8 μl of VWR H2O, and 5 μl of Taq DNA Polymerase (5U/μl, NEB). Samples were incubated in a thermocycler at 72°C for 45 minutes. After incubation we used the Promega Wizard® SV Gel and PCR Clean-Up System following the manufacturer protocol. Each column was eluted with 70 μl of water, the volume adjusted to 71 μl and 1 μl removed to perform Nanodrop quantification.

7. Ligation of Adapters

The following reaction(s) were performed in a 0.2 ml PCR tube. To the 70 μl repaired reaction 10 μl of 10X T4 DNA Ligase Buffer (NEB), 15 μl of Adapters (1.5 μg/μl) and 5 μl of T4 DNA Ligase (2000U/μl, NEB) was added. This was incubated at room temperature for 2 hours. The insert to vector ratio was calculated in terms of insert ends to vector ends. The number of ends available for ligation in pmoles can be calculated as follows:

pmol ends/μg of DNA = (2 x 10^6) / (number of base pairs x 660)

The ratio of adapter to DNA should be at least ~12:1. While this increases the chance of getting some adapter concatamer (which should not hybridize to the array), all of the fragments will likely get adapters, which is very important. When the ligation was complete, the sample was transferred to a 1.5 ml tube and 100 μl of VWR water was added. The Promega Wizard® SV Gel and PCR Clean-Up System was used following the manufacturer protocol. Each column was eluted with 50 μl of water and 1 μl was removed to perform Nanodrop quantification.

8. Hybridization

To the ligated sample we added a 5-fold amount (in μg) of human Cot-1 DNA (Invitrogen). The sample was dried in the Speed-Vac at medium heat (75°C) for 45 minutes. The sample was vortexed for 3 minutes and drying continued to the pellet. The following reactions were performed in a 1.5 ml tube. To the pellet from dried sample 7.2 μl of VWR water, 8.25 μl of

2X Hybe Buffer (NimbleGen) and 1.43 μl Hybe Component A (NimbleGen) was added. The samples were vortexed 3 minutes and then heated at 95°C for 10 minutes. The samples were quickly spun down and placed in the MAUI heat block at 42°C until ready to use. Once the samples were applied to the chip surface, we began the mixer on program B and hybridized for 60 hours.

9. Elution

After hybridization, the MGS arrays were first prewashed at 42°C in NimbleGen Buffer 1 followed by two 5 min washes at 47.5°C with Nimblegen Stringent Buffer. The arrays were then washed at room temperature for 2 min with NimbleGen Buffer 1, 1 min with NimbleGen Buffer 2 and 30 seconds with NimbleGen Buffer 3. We placed the washed chip on the Hybriwheel (NimbleGen) at 100°C and secured with a Hybe Puck (NimbleGen). We added 400 μl of 95°C VWR water and incubated 5 minutes. After the 5 minute incubation we removed as much water as possible and pipetted it into a labeled 1.5 centrifuge tube (placed on ice). We repeated this process one more time beginning with the addition of 400 μl of 95°C VWR water to the puck. When this was complete, we added 350-400 μl of 95°C VWR water and removed it immediately and pipetted it into the 1.5 ml tube.

After elution, the sample was placed in the Speed-Vac at medium heat (75°C) for 45 minutes. The sample was vortexed for 3 minutes and drying continued until the sample was to the pellet. We then hydrated the pellet in 33 μl of VWR water and vortexed for 3 minutes. We performed Nanodrop quantification of single strand DNA (DNA -33) to determine the concentration of the sample (picogreen and ethidium bromide quantification are inefficient for single stranded DNA). Upon eluting the selected target from the capture MGS chip, we obtained yields of between 700ng and 1.2 μg.

10. Amplification by Ligation Mediated PCR (LMPCR)

Each eluted sample was amplified using a single primer pair represented by the adaptors oligos and a high fidelity polymerase. To maintain an optimal concentration of 3ng/µl of template for each 50 µl PCR reaction, between 5 and 10 PCR reaction was done to amplify each entire eluate. One 50 µl reaction included 5 µl of 10X LA PCR buffer (TaKaRa), 5 µl of 2.5 mM dNTPs mix (TaKaRa), 2 µl of 20 µM FWD LMPCR primer, 2 µl of 20 µM REV LMPCR primer, and 2 µl of LA Taq (5U/µl, TaKaRa), and VWR water to 50 µl volume. The reactions were incubated in a thermocycler at (1) 95°C for 2 minutes, (2) 95°C for 60 seconds, (3) 58°C for 60 seconds, (4) 72°C for 60 seconds, (5) Repeat step 2 30 times (35 cycles), then at 72°C for 5 minutes and finally hold at 4°C.

All PCR reactions were pooled by sample and transferred into a 1.5 ml tube. We used the Promega Wizard® SV Gel and PCR Clean-Up System following the manufacturer centrifugation protocol to purify the sample. For spin steps we used 13000 g, and for the elution spin we used 16000 g and 1.5 minutes. Each column was eluted with 50 µl of water.

Three to 5 µl were used to verify size distribution on 1.5 % TAE agarose gel against 500 – 750 ng of 1 Kb plus ladder and positive control (6 X xylene cyanol loading dye for samples). Then the samples were quantified using Nanodrop and sonicated.

11. Resequencing of selected DNA

NimbleGen's Comparative Genomic Sequencing protocol was used for the 50K RA. Briefly, 1 µg of sample was denatured at 98^{o}C for 10 min in random primer buffer and labeled in the dark with Cy3-9mer primers (TriLink BioTechnologies) in the presence of dNTP mix and 100 units of Klenow (50U/µl, NEB) for 2 hours. To guarantee at least 20 µg of label sample for resequencing, 2 labeling reactions were done per sample (2 µg total). Labeled samples were purified using ethanol precipitation method and dried down to the pellet in the dark to avoid

bleaching of the Cy3 dye. After rehydrating the pellets with 20 µl total of VWR H_2O, ten to thirty micrograms of labeled DNA was mixed with NimbleGen's Hybridization cocktail (2X hybe buffer and hybe component A) and denatured at 95^{o}C for 5 min. The arrays were loaded and incubated overnight at 42°C on MAUI Hybridization System (BioMicro). The signal was detected by measuring Cy3-chrome fluorescence using Genepix 4000B (Molecular Devices Corp.).

For Affymetrix RAs, 30 µg of enriched samples were digested to 20 to 100 bp for 3 min in a 42µl reaction comprised of 10X Phor-All_Buffer (Amersham Biosciences), 10X Acetylated BSA and 3 units of DNAse1 (Promega). Reactions were heated at 75° C for 10 minutes to inactivate the DNAse then to 95° C for 15 minutes to separate the strands. The reactions were then cooled at 4° C for 45 minutes. The fragmented DNA was labeled using 17.13 nmol of a biotinylated proprietary labeling reagent (Affymetrix), 4.5 units of terminal deoxynucleotidyl transferase (Affymetrix) and terminal deoxynucleotidyl transferase buffer (Affymetrix) at a final concentration of 1X. The reactions were brought to a volume of 60µl with nuclease free water (VWR). Each reaction was incubated at 37°C for 4 hours followed by heat- inactivation for 15 minutes at 95°C and stored at 4° C until ready to use.

The labeled DNA samples were combined with 160 µl Hybridization buffer comprised of 1M Tris HCl pH 7.8 (Sigma), 5M TMACL (Sigma), 0.10% Tween 20 (Pierce Biotechnology), 100 µg/µl of Herring Sperm DNA (Promega), 500ug/ml Acetylated BSA (Invitrogen), and 200pM biotinylated SNPHy948B (Invitrogen). The hybridization mix was then heated to 95°C for 5 minutes, equilibrated at 49°C and hybridized to the high-density oligonucleotide array at 49°C for 16 hours. All signal detection steps were performed using an Affymetrix fluidics. The arrays were washed in 6X SSPE, 0.01% Tween 20 solution (wash A) 6 times at 25°C then in .6X SSPE, 0.01% Tween 20 solution (wash B) 6 times at 45°C. For signal detection, the arrays were incubated with stain 1 (6X SSPE, 0.01% Tween 20, 1X Denhardt's solution (Sigma), and

10ug/ml SAPE (Invitrogen), final concentration) for 10 minutes at 25°C, followed by 6 washes with wash A at 25°C. Incubation with stain 2 (6X SSPE, 0.01% Tween 20, 1X Denhardt's solution (Sigma), and 10ug/ml anti-streptavidin antibody (Vector), final concentration) was done for 10 minutes at 25°C. A second incubation with stain 1 was done for 10 minutes at 25°. The arrays were rewashed 10 times in wash A at 30°C and filled with a holding buffer (5M NaCl, 10% Tween 20, MES hydrate and MES sodium salt). They were stored at 25°C until they were ready to be scanned. The signal was detected by measuring Cy-chrome fluorescence using a G7 Genechip scanner (Affymetrix). For both the Nimblegen and Affymetrix resequencing arrays, all bases calls were made with the RATools program RA_PopGenCaller (http://www.dpgp.org/).

12. Validation Sequencing

Discrepancies between RA data and HapMap data were evaluated using independent sequencing (Agencourt). PCR primers were designed using Primer 3 (http://frodo.wi.mit.edu/). PCR Reactions were composed of 400 ng of sample DNA was mixed with 8 µl of dNTP mix (TaKaRa), 5 µl of 10X LA Taq buffer (TaKaRa), 1.5 µl LA Taq (TaKaRa), 0.8 µl of each forward and reverse primers and VWR water to 50 µl total reaction volume. DNA was amplified using the following parameters: 94°C for 4 min, 30 cycles of 94°C for 20 sec, 58°C for 1 min, and 72°C followed by 72°C for 5 minutes. This method was also used to validate discrepancies in the Tr91 RA data. The primers that amplified the SNP discrepancies are listed in Supplementary Table 2. PCR products were run on a 1% TAE agarose gel, excised from the gel and purified using the Promega Wizard® SV Gel and PCR Clean-Up System.

13. Long PCR Control

To minimize the number of amplifications, we used long PCR to amplify genomic regions that contain one or more unique sequence blocks tiled onto the variant resequencing

array. A total of 14 primer pairs spanning 48 Kb (including the 39 kb FMR1 genome region) were used. Except for one primer close to the CGG repeat (20 bp) Long PCR primers were 31 to 34 base pairs long and were selected by using Amplify 3.1.4 [6] to ensure that they bound uniquely within a 48 kb region and had a primer stability value between 70 and 80. Primers had GC content between 45% and 60%.

Amplification of genomic DNA was accomplished in 50 µl reactions carried out in thin-walled polypropylene tubes using LA Taq (TaKaRa). The manufacturer's recommendation was followed. LPCR amplification of the human samples employed either a standard or a modified mixture where 5% DMSO (or manufacturer GC Buffer) was added to aid the amplification of GC rich regions. The standard conditions for the LPCR were: (1) 94˚C for 2 minutes, (2) 94˚C for 10 seconds, (3) 68˚C for 1 minute per kb fragment size, (4) repeat to step 2, 30 times, and (5) final extension time equal to step 3 plus five minutes. Each LPCR required a minimum of 200ng of human genomic DNA and most fragments were between 3.4 and 11 kb long. To obtain optimal performance across the microarray, we pooled equal molar concentration of PCR product, to ensure that an equal number of targets existed for each probe on the array. The primer sequences that amplified each fragment are listed in Supplementary Table 2.

14. Quantitative PCR

We performed quantitative PCR on sample DNA with two treatments: (1) whole genome amplified, ligated and then amplified using LMPCR protocol but never hybridized to a genomic selection array (Treatment 1) and a (2) whole genome amplified, ligated, hybridized to a genomic selection array, eluted from the array, and then amplified using LMPCR. Reagents used included iQ SYBR® Green Supermix (Bio-Rad) and the following primer pair:

FW: ACAGTAGGGCTGTGCTTACTGC

REV: CTCATTTTCAGCCTCAATCCTC

The primers amplify 156 bases from exon 10 in the FMR1 gene. Reactions contained 12.5 µl of

1X iQ SYBR® Green Supermix, 1 µl of FW Primer (10mM), 1µl of REV Primer (10mM), 9.5 µl of VWR water and 1 µl of DNA template (30 ng/µl) for a total volume of 25 µl. The standard curve was created using whole genome amplified DNA at concentrations ranging from 7.8 ng/µl to 500 ng/µl. The reactions were performed in triplicate. The reactions were incubated in a Bio-Rad iQ5 Multicolor Real Time PCR Detection Light Cycler using the following parameters: (1) 94˚C for 3 minutes, (2) 94˚C for 10 seconds, (3) 58˚C for 30 seconds, (4) 72˚C for 30 seconds, and (5) Repeat steps 2-4 for 40 cycles. From our quantitative PCR result we conservatively estimate at least 1000X enrichment of DNA used for resequencing (treatment 2) when compared to whole genome amplified DNA that underwent LMPCR amplification (treatment 1). The DNA from treatment 2 had a cycle threshold of 15 while the cycle threshold for treatment 1 was 25. If we assume that DNA concentration doubles every cycle then enrichment can be calculated by 2^N, with N equaling the difference between the cycle thresholds of the two treatments (Supplementary Figure 2).

REFERENCES

1. Karolchik, D. et al. The UCSC Genome Browser Database. *Nucleic Acids Res* **31**, 51-4 (2003).
2. Kleinjan, D.A. & van Heyningen, V. Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am J Hum Genet* **76**, 8-32 (2005).
3. De Boulle, K. et al. A point mutation in the FMR-1 gene associated with fragile X mental retardation. *Nat Genet* **3**, 31-5 (1993).
4. Gu, Y., Lugenbeel, K.A., Vockley, J.G., Grody, W.W. & Nelson, D.L. A de novo deletion in FMR1 in a patient with developmental delay. *Hum Mol Genet* **3**, 1705-6 (1994).
5. Cutler, D.J. et al. High-throughput variation detection and genotyping using microarrays. *Genome Res* **11**, 1913-25 (2001).
6. Engels, B. Amplify 3. *http://engels.genetics.wisc.edu/amplify/* (2005).

Supplementary Table 1. Primer sequences used in independent sequencing validation of HapMap and Tr91 discrepancies.

HapMap Samples	
rs16994908_FW2	**CTTCACCATTTTTGCATGTACC**
rs16994908_REV	**TTGCAACCACATTTGAAGTGAC**
rs12688573_FW	**AAAGTCGCACAGATACCCTCTC**
rs12688573_REV	**CTTTTCTGTCTTGCCATTAGCC**
rs11117557_3_FW	**ACTGCATCTGCAGAGAAACAAC**
rs11117557_3_REV	**AACAGTTGTGAAACTACGTCAGG**
rs7052829_FW	**TTATGGGAAGAATCCACTCCAG**
rs7052829_REV_2	**AGTAGCAGCAACAGCAACAAAG**
rs7052654_rpt_FW	**CAGGGCAGGGATGATTAGAG**
rs7052654_rpt_REV	**AGAAAGGAAGAGATGCATGGAC**
rs6626955_6_FW	**TCCCTTGTGTTCATGGAGTATG**
rs6626955_6_REV	**AACAGGAGCTTCTTCCTGATTG**
rs2761622_2_FW	**AAATGAAATGCACCTTCCAGAG**
rs2761622_2_REV	**GCACTTGTTTCACAGGTACAGC**
rs1805422_FW	**GTAGCAGTAGTGCGTTTGTTGG**
rs1805422_REV	**TTTCCTATAGCCAAACGTGTCC**
rs1265401_FW	**GGGTATGGGTTTAACATAGGACAG**
rs1265401_REV	**GACTTACGGGCTGCTTCTCAC**
rs1265397_FW	**GCATGCGTGTCTTACTCCATAG**
rs1265397_REV	**AAGCTCTGTCAGTGTGATGTGG**
rs25699_FWD	**GCCAGAGGCTATTTCCCTAACTTAC**
rs25699_REV	**TGATGACGAACTCTGGAATTTGAC**
rs4949_FWD	**AGAGTGCTTTTGTTGGGATGTAC**
rs4949_REV_2	**attacacacataGGTGGCACTA**
rs1442280_FWD	**AGACATTGCAAACATCCAGAAC**
rs1442280_REV	**ATGCAGTCAGCCAGGTAATAGA**
rs16994869_FWD	**tgAACAGTCACTTGACATCCAAAG**
rs16994869_REV	**GATTGGAGGAGGCAGAGAAATAGT**
Tr91	
rs29284_int9_FW	**CTCTGGTACCTGACCAAAGGAG**
rs29284_int9_REV	**AAAGCAGTAAGCACAGCCCTAC**
rs29288_int13_FW	**CATGCCATTCATTCTTATGGTG**
rs29288_int13_REV	**AATCCTAACTCTCCAGGCCTTC**
rs25707_ex5_FW	**CCTGCCACAAAAGATACTTTCC**
rs25707_ex5_REV	**TTCTCCATTGCTCTTGCAAAC**
I304N_ex10_FW	**ACAGTAGGGCTGTGCTTACTGC**
I304N_ex10_REV	**CTCATTTTCAGCCTCAATCCTC**
rs29286_int12_FW	**GTGGCTTCATCAGTTGTAGCAG**
rs29286_int12_REV	**CACATACCCACAAACACTCCTC**
rs5904816_int14_FW	**GCACATCAAGGTTTGAACTTAGG**
rs5904816_int14_REV	**CAGAGACGTTTCAGGGGTAATC**
rs25704_ex17_FW	**GGAAGGTCATTTCCATGTATGC**
rs25704_ex17_REV	**AAAACCAAACCCCAACACTTC**