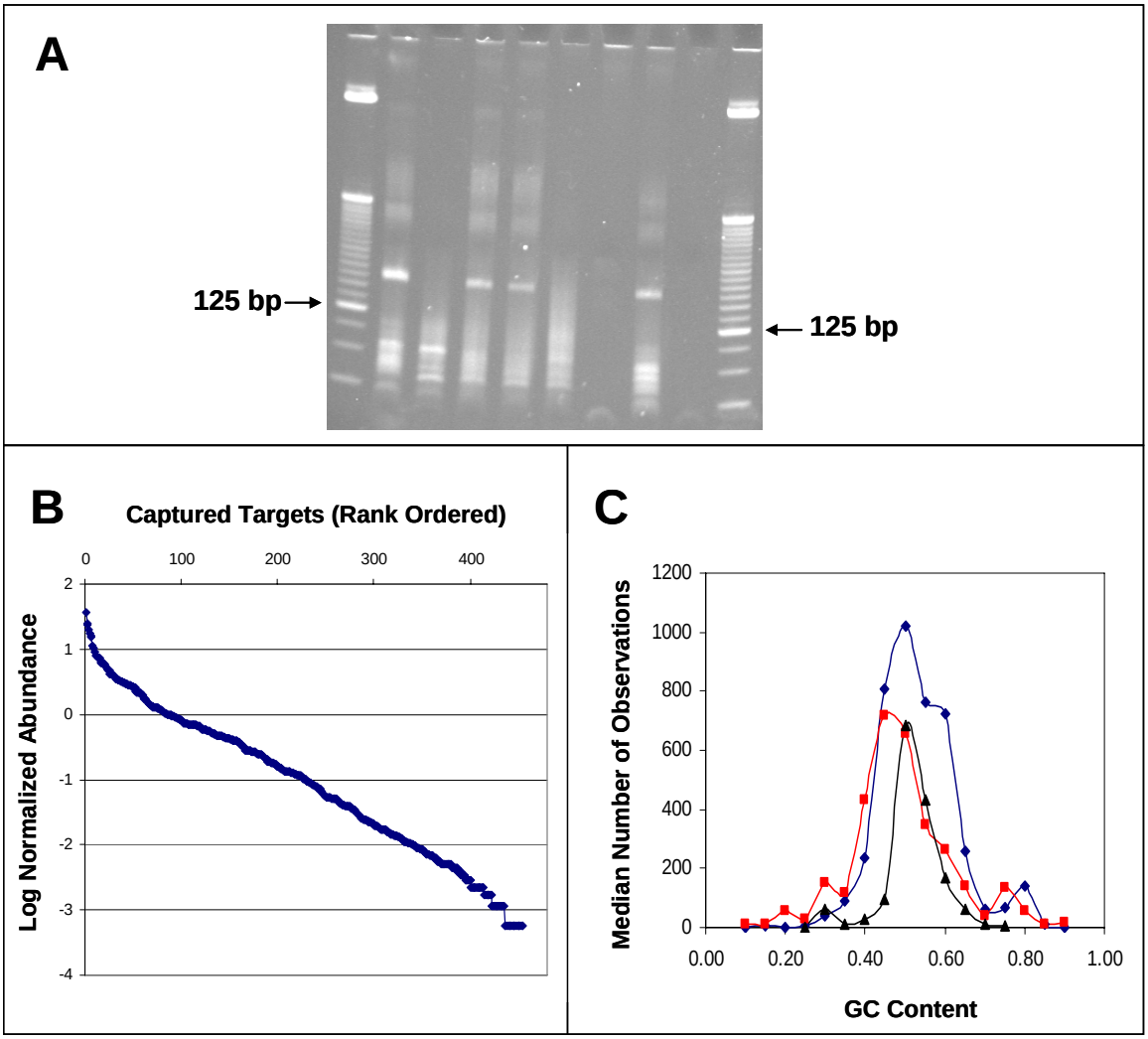
Multiplex amplification of large sets of human exons

Gregory J Porreca, Kun Zhang, Jin Billy Li, Bin Xie, Derek Austin, Sara L Vassallo, Emily M LeProust, Bill J Peck, Christopher J Emig, Fredrik Dahl, Yuan Gao, George M Church & Jay Shendure

Supplementary figures and text:

Note: Supplementary Data 1(list of 55,000 capture oligonucleotides) and 2(list of 480 capture oligonucleotides) are available on the Nature Methods website.

Supplementary Figure 1. Design, results and analysis of a 480-plex pilot experiment.



Our initial evaluation of the key step of our capture strategy (**Fig. 1B**), consisted of an experiment in which we attempted the multiplex amplification of 480 exons. There were several key differences in this experiment relative to the subsequent experiments described in the manuscript. First, the targets consisted of 480 human genomic sequences of a *single length* (121 bp). Second, the design of targeting arms was *not* restricted to those with intermediate, i.e. 30% to 70% GC contents. Third, single-stranded 70-mer targeting oligonucleotides were generated by conventional column-based synthesis, rather than via programmable microarrays. These 480 oligonucleotides were synthesized individually (Invitrogen), combined to generate an equimolar pool, and 5' phosphorylated with T4 polynucleotide kinase. Fourth, the

concentrations of these oligos during the capture reaction was higher (80 nM total concentration; estimated concentration of ~167 pM for each species). Fifth, no linear rolling circle amplification step (as in **Fig. 1C**) was performed. Rather, capture circles were directly subjected to PCR amplification with the common primer pair. (**a**) The products of multiplex exon capture reaction amplifications are shown on a 6% non-denaturing polyacrylamide gel. Lane 1 = 25 bp ladder (Invitrogen); Lane 2 = 96-plex exon capture reaction (a subset of the 480-plex), replicate #1; Lane 3 = negative control of 96-plex reaction (no genomic DNA); Lane 4 = 480-plex exon capture reaction, replicate #1; Lane 5 = 480-plex exon capture reaction, replicate #2; Lane 6 = negative control of 480-plex reaction (no genomic DNA); Lane 7 = negative control with genomic DNA only, no targeting probe; Lane 8 = 96-plex exon capture reaction, replicate #2; Lane 9 = no-template negative control; Lane 10 = 25 bp ladder (Invitrogen). A sharp band with an expected size of ~200 bp (121 bp targets + 40 bp of targeting arm sequence + 40 bp of common primer sequence → 201 bp amplicons) is seen in lanes 2, 4, 5, and 8, corresponding to all replicates of the 96-plex and 480-plex reactions, and in none of the negative control lanes. One of the ~200 bp bands from a 480-plex amplification was gel-purified, sub-cloned and Sanger sequenced. 118 of 118 sequencing reads that aligned to the human genome corresponded to an expected target, consistent with a very high specificity. (**b**) To characterize uniformity, amplicons resulting from a 480-plex capture reaction were end-sequenced to a high depth on by polony tag sequencing[1, 2]. 25 targets were excluded from analysis because the polony tags for those sequences did not define them uniquely. 853,912 sequencing reads were obtained that could be confidently mapped to one of the remaining 455 targets. Counts were normalized relative to the mean abundance for each reaction. The logs (base 10) of the estimated relative abundances were calculated, sorted, and plotted (blue). All of the 455 targets were observed at least once (0% “dropout” rate). However, uniformity was still poor, as the relative abundances of individual targets varied over 4 to 5 logs. Targeting oligo sequences, target sequences, and counts for each target in this experiment are provided in **Supplementary**

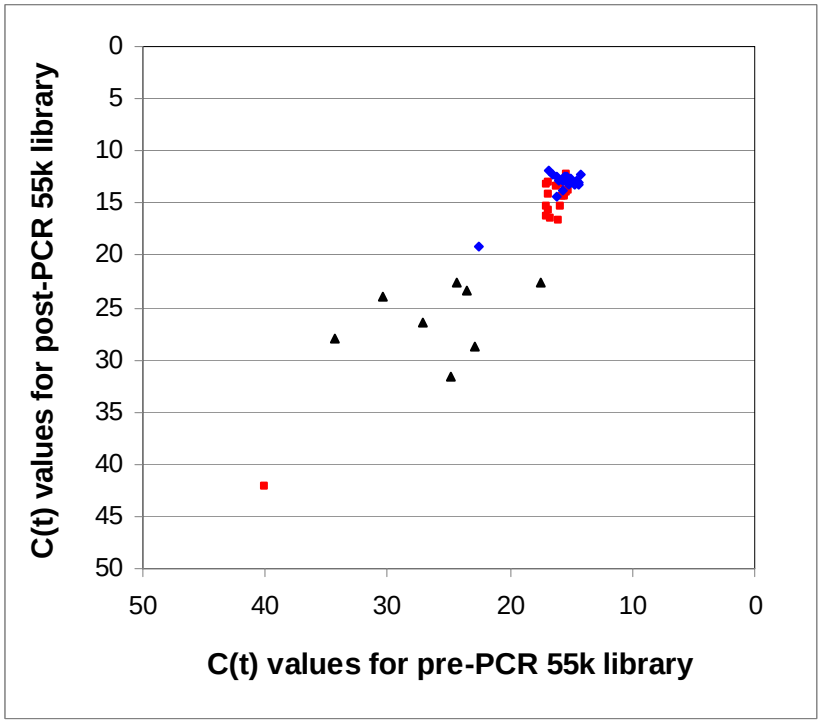
Data 2 online. (**c**) Abundance as a function of GC content. The median abundance is shown as a function of the GC content of the 3' targeting arm (from which polymerase driven gap-fill originates; red squares), the 5' targeting arm (to which ligation occurs to complete a capture event with the formation of a circular DNA molecule; blue diamonds), or the gap-fill sequence itself (black triangles). We clearly observe that intermediate GC contents for all three sets of sub-sequences are associated with higher abundances in products of multiplex amplification.

References

1. Shendure, J. et al. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728-1732 (2005).

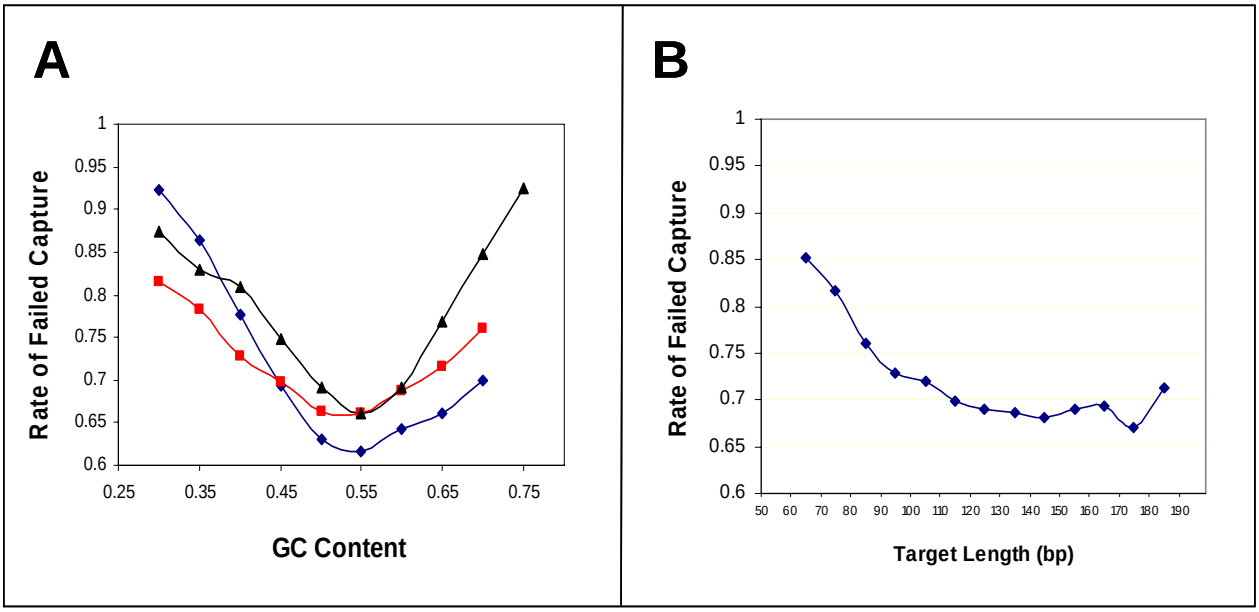
2. Kim, J.B. et al. Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science* **316**, 1481-1484 (2007).

Supplementary Figure 2. QC analysis of microarray-derived targeting oligonucleotides.



The presence of individual species within the pool of 55,000 microarray-derived oligonucleotides was assayed by real-time PCR, using primer pairs directed at the unique targeting arm sequences. Primer pairs were designed (Primer3) for: 20 probes that target exons observed in *both* replicates (blue diamonds); 20 probes that target exons observed in *neither* replicate (red squares); and, as negative controls, 8 probes that were not expected to be present (black triangles). The X-axis depicts C(t) values for amplification directly from the probe pool itself. The Y-axis depicts C(t) values for the probe pool after it had undergone PCR amplification as described in the Methods. For all negative controls (black triangles), gel electrophoresis of products showed non-specific amplification. Both pre-PCR and post-PCR, the C(t) values of the two subsets of targeting oligos are not significantly different ($p>0.05$).

Supplementary Figure 3. Failed capture as a function of GC content and target length.



The rate of failed capture, i.e. the "dropout" rate, is calculated as the fraction of the 55,000 intended targets that were observed in *neither* of the replicate reactions. (**a**) The rate of failed capture is shown as a function of the GC content of the 3' targeting arm (from which polymerase driven gap-fill originates; red squares), the 5' targeting arm (to which ligation occurs to complete a capture event with the formation of a circular DNA molecule; blue diamonds), or the gap-fill sequence itself (black triangles). All bins in which there were at least 100 targets are shown. In the design of this experiment, both targeting arms were restricted in GC content to the 30% to 70% range. For all subsequences, we observe that an intermediate GC content is correlated with a lower rate of "dropout". (**b**) The rate of failed capture is shown as a function of the length of the gap-fill sequence itself (range = 60 – 191 bp; ~10 bp bins). We observe that shorter gap-fill lengths are associated with a modestly higher rate of "dropout".

Supplemental Table 1. Cost analysis for synthesis of targeting oligonucleotide pools.

	Column	Array
Synthesis cost	$1,732,500	$12,000
Probe pool yield (fmol)	3.9E+11	8.3E+04
Reactions per synthesis	9.8E+08	2.1E+02
Probe pool cost per fmol (dollars)	4.4E-06	1.6E-01
Probe pool cost per reaction (dollars)	1.8E-03	6.6E+01

We estimated the costs of generating a pool of 55,000 targeting oligonucleotides by two strategies – individual, column-based synthesis vs. parallel synthesis on a microarray. The cost of column-based synthesis is estimated assuming $0.45 per base X 70 nucleotides per probe X 55,000 probes at 100 nmol scale (Integrated DNA Technologies). The cost of array-based synthesis was $12,000 for 55,000 probes, each 100 nucleotides in length (Agilent, academic use). Final yield of column-based synthesis is >= 7.1 nmol per probe, while array-based synthesis is 1.5 fmol per probe (0.3 fmol per probe before amplification). Each selection reaction consumes 400 fmol of probe pool. The estimated cost of array-based synthesis includes amortized PCR amplification and restriction digestion reagent costs of $19.18 per nmol, which increases yield per array 5-fold. The estimated costs for column-based synthesis do not include the costs of normalizing and pooling individually synthesized oligos. We note that: (a) the *upfront* cost of the array-based strategy is ~2 orders of magnitude below the estimated upfront cost of column-based synthesis; (b) In principle, the *amortized* cost of the column-based strategy is dramatically less than that of the array-based synthesis. However, this calculation (as performed above) assumes that the cost is amortized over ~1e9 capture reactions.

Supplementary Methods

Structure & Sequence of Array-Synthesized 100-mers. 100-mer targeting oligo precursors were defined for each potential target. The sequence of each 100-mer was: 5'-AGGACCGGATCAACTxxxxxxxxxxxxxxxxxxxxCTTCAGCTTCCCGATAT CCGACGGTAGTGTyyyyyyyyyyyyyyyyyyyyCATTGCGTGAACCGA-3' (x's and y's indicate variable 20 nt sequences that correspond to targeting arms). A set of 55,000 targets was selected that: (a) were 60-191 bp in length; (b) had targeting arms that did not overlap with sequences annotated as repetitive; (c) had targeting arms with individual GC contents of between 0.30 and 0.70, inclusive; (d) did not contain Nt.AlwI or Nb.BsrDI recognition sites within its targeting arms.

Shotgun sequencing of multiplex capture products. To generate a shotgun library of captured material, gel-purified PCR amplicons were "recircularized" with a common adaptor as follows: 20 ul, containing 2 ul of template (gel-purified capture amplicons), 1 mM ATP, 1 mM DTT, 250 nM CP_2_CIRC_A splint oligo, 250 nM CP_2_CIRC_NO_A splint oligo, 5 units Optikinase (USB), and 5 units Ampligase in 1x Ampligase Buffer, was incubated at 37°C for 30 min, (95°C for 30 s, 60°C for 10 min) x 10. To degrade uncircularized material, 4 ul of exonuclease mix was added (containing 40 units of exonuclease I and 200 units of exonuclease III (New England Biolabs)), followed by incubation at 37°C for 1 hr and 80°C for 20 min. Recircularized capture products were then subjected to randomly primed hyperbranched rolling circle amplification (hRCA) as follows: 50 ul, containing 10 ul of template (recircularized capture products), 1 mM dNTPs, 0.4x SybrGreen, 50 uM phosphorthiolated random hexamers, and 250 units of phi29 in 1x phi29 amplification buffer (Epicentre), was prepared on ice and incubated at 30°C for 5 hr, 65°C for 10 min. Products of hyperbranched RCA reaction were sheared and prepared for sequencing with the Solexa Genomic Library Kit (Illumina), and sequencing was carried out on the Illumina Genome Analyzer as per manufacturer's instructions.

Analysis of Solexa Sequencing Data. For end-sequencing and shotgun sequencing, reads were ~35 bases in length. Sequence reads were subjected to a simple algorithm that compared each read to all possible expected alignments, and assigned to the alignment with the minimum number of mismatches (reads with 'ties' were discarded). Indels were not considered by this algorithm. For end-sequencing, a maximum of 8 mismatches, and for shotgun sequencing, a maximum of 4 mismatches, were allowed. These thresholds were chosen such that when simulations were performed in which reads were aligned to databases of random sequences of equivalent complexity, a negligible number of assignments were made. For end-sequencing, counts from each replicate are provided in **Supplementary Data 1** online. HapMap genotyping data for GM12248 was filtered for "validated" positions. SNP positions were considered validated if the minor allele was observed on at least two chromosomes in the full HapMap dataset. Targets at which there was strong potential for "paralog" capture, in that there were multiple instances in the human genome of 'exact matches' for the targeting arms within 1000 bp of one another, were excluded from resequencing data analysis.

Oligonucleotide Sequences. Sequences of oligonucleotides (5' → 3') used here are as follows:

eMIP_CA1_F: TGCCTAGGACCGGATCAACT

eMIP_CA1_R: GAGCTTCGGTTCACGCAATG

RCA_2_RA: ACCGTCGGATATCGGGAAGC

CP-2-FA: GCACGATCCGACGGTAGTGT

CP-2-RA: CCGTAATCGGGAAGCTGAAG

CP_2_CIRC_A: ACACTACCGTCGGATCGTGCACCGTAATCGGGAAGCTGAAG

CP_2_CIRC_NO_A: ACACTACCGTCGGATCGTGCCCGTAATCGGGAAGCTGAAG;

Random hexamers: NNNN*N*N, with asterisk representing phosphorthiolated positions.

Note. All column purifications were performed with the Qiaquick PCR Purification Kit (Qiagen). All PCR and hRCA reactions were performed on a real-time PCR machine.