Direct selection of human genomic loci by microarray hybridization

Thomas J Albert, Michael N Molla, Donna M Muzny, Lynne Nazareth, David Wheeler, Xingzhi Song, Todd A Richmond, Chris M Middle, Matthew J Rodesch, Charles J Packard, George M Weinstock & Richard A Gibbs

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

Supplementary Table 1. Exon Capture Data. A series of capture microarrays were employed for direct selection of 6,726 Exonic regions totaling ~5 Mb of genomic sequence from 5 different human genomic DNA samples. These samples were sequenced with one run of the Roche 454 GS-FLX instrument.

DNA SAMPLE	**qPCR Fold Enrichment**	**FLX – Yield (Mb)**	**Percentage of Reads Mapped Uniquely to the Genome**	**Percentage of Total Reads Mapped to Selection Targets**	**Median Fold Coverage for Target Regions**
NA04671	318	63.1	91%	75%	5
NA04671	399	115	89%	65%	7
NA04671	418	93.0	91%	76%	7
HapMap CEPH	217	77.6	88%	74%	7
HapMap JPT	153	96.7	84%	66%	8
HapMap CHB	240	52.8	83%	59%	4
HapMap YRI	363	81.3	53%	38%	4

Supplementary Table 2. SNP Analysis. Coverage of an average of *at least one read* for each base in the target regions

Pop/Indiv	CEPH/ NA11839	CHB/NA18573	JPT/ NA18942	CEPH/ NA11839 CEPH/ NA11839
# Known variant alleles	2235	2257	2206	2334
Stringency of *at least one read* per known variant Hapmap allele				
Positions with ≥ 1 read	2176 (97.3%)	2104 (93.2%)	2168 (98.2%)	2133 (91.3%)
Variant alleles found in ≥ 1 read	2071 (92.6%)	1922 (85.1%)	2080 (94.2%)	1848 (79.1%)
False negative rate	7.4%	14.9%	5.8%	20.9%

Stringency of *at least two reads* per known variant Hapmap allele				
Positions with ≥ 1 read	2176 (97.3%)	2104 (93.2%)	2168 (98.2%)	2133 (91.3%)
Variant alleles found in ≥ 2 reads	1907 (85.3%)	1569 (69.5%)	1939 (87.8%)	1469 (62.9%)
False negative rate	14.7%	30.5%	12.2%	37.1%

The total number of positions in the target regions that were genotyped in the Hapmap project was 8103 (CEU), 8134 (CHB), 8134 (JPT), 8071 (YRI) for each of the four genomes. Of these, most (~6000) sites were homozygous for the reference genome allele. The number of known variant alleles (homozygous or heterozygous) is listed in the second row. These positions were analyzed for coverage and if the allele(s) were found in the captured DNA.

Supplementary Table 3. Locus Capture Data. Capture and sequencing of regions of increasing size containing the human *BRCA1* locus (see text for more details). The individual DNA fragments were captured and sequenced with one run of the Roche 454 GS-FLX instrument.

Tiling Size (kb)	**Average Selection Probe Tiling Density**	**FLX –Yield (Mb)**	**Percentage of Reads Mapped Uniquely to the Genome**	**Percentage of Total Reads That Mapped to Selection Targets**	**Median fold coverage of Unique Portion of Region**
200	1bp	102	55%	14%	79
500	1bp	85.0	61%	36%	93
1,000	2bp	96.7	56%	35%	38
2000	3bp	112.6	81%	60%	37
5,000	7bp	140	81%	64%	18

Supplementary Methods

Loci Selection and Capture Microarray Design: For the initial exonic design 6,726 genome regions of a minimum 500 bp in size were selected from the HG17 build of the human genome (Supplemental Table 1). The specific gene choices were part of an ongoing multicenter collaborative program testing specific loci for mutation in cancer. Overlapping microarray probes (>60 bases) were designed to span each target region, with a probe positioned each 10 bases for the forward strand of the genome. Array probe sequences for these exon regions are provided as supplemental data._For the ‘locus’ capture experiments, five genomic regions of increasing size surrounding the *BRCA1* gene locus were also selected for capture. These regions, from the HG18 build of the genome are listed in the following table:

BRCA1 Region Size	Average Selection Probe Tiling Density	Chromosome 17 coordinates (HG18)
200kb	1bp	38,390,417 – 38,590,417
500kb	1bp	38,240,417 – 38,740,417
1Mb	2bp	37,990,417 – 38,990,417
2Mb	3bp	37,490,417 – 39,490,417
5Mb	7bp	35,990,417 – 40,990,417

The average probe tiling density listed above represents the average distance in bases between the start of one probe and the start of the next probe.

To avoid non-specific binding of genomic elements to capture arrays, highly repetitive elements of the genome were excluded from selection microarray designs, using a new method that utilizes a strategy similar to the WindowMasker program developed by Morgolis (2006) to identify these regions and exclude them from probe selection.[1] The process compared the set of probes against a pre-computed frequency histogram of all possible 15-mer probes in the human genome. For each probe, the frequencies of the 15-mers comprising the probe are then used to calculate the average 15-mer frequency of the probe. The higher the average 15-mer frequency, the more likely the probe is to lie within a repetitive region of the genome. Only probes with an average 15-mer frequency less than 100 were used. This method results in better coverage of the genome, as compared to the conventionally used RepeatMasker, while still avoiding highly repetitive regions[2]. DNA microarrays were synthesized according to standard NimbleGen microarray manufacturing protocols[3].

DNA Sources: Purified genomic DNA was purchased from Coriell for the following cell lines: Burkett lymphoma cell line: ATCC#NA04671, HapMap samples: (CEPH/ NA11839, CHB/NA18573, JPT/NA18942, YRI/NA18861). Burkitt’s Lymphoma DNA was whole genome amplified using Qiagen service (Hilden, Germany).

Sample Preparation and Microarray Capture: 20 ug of DNA from each sample (either total genomic DNA or whole genome amplified DNA) were sonicated to an average size of 500bp and treated with the Klenow fragment of DNA polymerase I (NEB, Beverly MA) to generate blunt-ends, and then 5′ phosphorylated with polynucleotide kinase (NEB). The following oligonucleotides were annealed and ligated to the fragment ends: 5′-Pi-GAGGATCCAGAATTCTCGAGTT-3′, 5′-CTCGAGAATTCTGGATCCTC-3′. Ligated samples were hybridized to capture arrays in the presence of 1x NimbleGen hybridization buffer (NimbleGen, Madison WI) for approximately 65 hours at 42°C with active mixing using a MAUI hybridization station (NimbleGen). After hybridization, arrays were stringently washed 3 x 5 minutes with Stringent Wash Buffer (NimbleGen) and rinsed with Wash Buffers 1, 2, and 3 (NimbleGen). Captured DNA fragments were immediately eluted with 2 x 250 µl of water at 95°C. Samples were dried down and resuspended for amplification using a primer complementary to the linker ligated earlier.

To quantify the fold enrichment of the exonic regions, eight loci were selected at random for qPCR analysis. These regions were amplified using the following primers:

region 1	F: 5′-CTACCACGGCCCTTTCATAAAG-3′
	R: 5′-AGGGAGCATTCCAGGAGAGAA-3′
region 2	F: 5′-GGCCAGGGCTGTGTACAGTT-3′
	R: 5′-CCGTATAGAAGAGAAGACTCAATGGA-3′
region 3	F: 5′-TGCCCCACGGTAACAGATG-3′
	R: 5′-CCACGCTGGTGATGAAGATG-3′
region 4	F: 5′-TGCAGGGCCTGGGTTCT-3′
	R: 5′-GCGGAGGGAGAGCTCCTT-3′
region 5	F: 5′-GTCTCTTTCTCTCTCTTGTCCAGTTTT-3′
	R: 5′-CACTGTCTTCTCCCGGACATG-3′
region 6	F: 5′-AGCCAGAAGATGGAGGAAGCT-3′
	R: 5′-TTAAAGCGCTTGGCTTGGA-3′
region 7	F: 5′-TCTTTTGAGAAGGTATAGGTGTGGAA-3′
	R: 5′-CAGGCCCAGGCCACACT-3′
region 8	F: 5′-CGAGGCCTGCACAGTATGC-3′
	R: 5′-GCGGGCTCAGCTTCTTAGTG-3′

Samples were analyzed using an ABI 7300 real time PCR system (Applied Biosystems, Foster City, CA) according to manufacturers protocols, measuring SYBR green fluorescence to compare enriched, amplified samples with genomic DNA that was ligated to linkers and LM-PCR amplified, but not hybridized to a capture array.

After elution from arrays, DNA fragments were ligated to linkers compatible with 454 sequencing. These samples were amplified on beads using emulsion PCR and sequenced using the 454 sequencing instrument (454, Branford CT). Because each sequenced fragment contained the 20bp linker for the LM-PCR used immediately after microarray elution, the majority of 454 sequencing reads contained this linker sequence.

Following *in-silico* removal of the linker sequence, we used BLAST[4] to compare each of the sequencing reads to the entire appropriate version of the Human Genome. We use a cutoff score of e = 10^-48, tuned to maximize the number of unique hits. The reads that did not uniquely map back to the genome (between 10 and 20%) were discarded. The rest were considered “captured sequences”. The captured sequences that, according to the original BLAST comparison, map uniquely back to regions within the target regions are considered “sequencing hits”. These are then used to calculate the % of reads that hit target regions, and the fold sequencing coverage for the entire target region. Data was visualized using SignalMap software (NimbleGen).

[1] Morgulis, A., Gertz, E.M., Schäffer, A.A., Agarwala, R. *Bioinformatics* **15**, 134-41 (2006).

[2] www.reapeatmasker.org.

[3] Nuwaysir, E.F., et al.. *Genome Res.* **12**:1749-1755 (2002).

[4] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. *J. Mol. Biol.* **215**:403-410 (1990).