Published in final edited form as:
Nat Methods. 2009 July ; 6(7): 507–510. doi:10.1038/nmeth.1343.

Filter-based hybridization capture of subgenomes enables resequencing and copy-number detection

Daniel S Herman[1], **G Kees Hovingh**[1,2], **Oleg Iartchouk**[3], **Heidi L Rehm**[3,4], **Raju Kucherlapati**[1,3], **J G Seidman**[1,3,6], and **Christine E Seidman**[1,3,5,6]

[1] Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA [2] Department of Vascular Medicine, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands [3] Partners Healthcare Center for Personalized Genetic Medicine, Boston, Massachusetts, USA [4] Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts, USA [5] Howard Hughes Medical Institute, Harvard Medical School, Boston, Massachusetts, USA

Abstract

To exploit contemporary sequencing technologies for targeted genetic analyses, we developed a hybridization enrichment strategy for DNA capture that uses PCR products as subgenomic traps. We applied this strategy to 115 kb of the human genome encompassing 47 genes implicated in cardiovascular disease. Parallel sequencing of captured subgenomic libraries interrogated 99.8% of targeted nucleotides ≥20 times (~40,000-fold enrichment), enabling sensitive and specific detection of sequence and copy number variation.

Emerging knowledge about the genetic basis of human disease provides unparalleled opportunity to improve classification of pathologies and to predict disease susceptibility by DNA sequencing. Realizing these opportunities requires robust strategies to study multiple genes potentially involved in clinical phenotypes. Two cardiovascular disorders illustrate the considerable challenges of comprehensive genetic analyses. Cardiac hypertrophy, in the absence of a physiologic explanation, occurs in one in 500 individuals1 and can arise from dominant mutation in genes that encode components of the cardiac sarcomere (defining hypertrophic cardiomyopathy, HCM) or in genes involved in myocyte metabolism2. Plasma high-density lipoprotein (HDL) cholesterol level, a risk factor for coronary artery disease, is strongly influenced by rare3 and common4 variants in a growing list of genes.

Optimal screening of genes implicated in cardiac hypertrophy and HDL metabolism, or any other target subgenome, requires assessment of both nucleic acid sequence and copy

Correspondence should be addressed to C.E.S. (cseidman@genetics.med.harvard.edu).

[6] These authors contributed equally to this work.

AUTHOR CONTRIBUTIONS D.S.H., G.K.H., J.G.S. and C.E.S. conceived and designed the study; D.S.H. and G.K.H. constructed libraries and target concatemers; D.S.H. analyzed sequence data; O.I. and R.K. sequenced the libraries; H.L.R., D.S.H. and R.K. designed and amplified HCM sequences; D.S.H., J.G.S. and C.E.S. wrote the manuscript.

number. Parallel sequencing technologies simultaneously detect sequence and copy number variation5–8 at costs considerably less than that of conventional strategies, and could, in principle, be employed to study target subgenomes.

Traditionally, target DNA has been captured by hybridization enrichment, as in the pairing of a genomic locus with corresponding cDNA9. The utility of hybridization enrichment of genomic DNA libraries for targeted resequencing has been recently demonstrated using bacterial artificial chromosomes10 and long oligonucleotides tethered to microarrays11 or in solution12. Other groups have developed multiplex target amplification methods, such as molecular inversion probes13. These approaches show reasonable specificity, but are relatively expensive and have yet to demonstrate detection of copy number variation (CNV) or sufficient uniformity for applications requiring complete detection of sequence variation.

We describe a practical approach to targeted resequencing: filter-based hybridization capture using PCR amplimers to capture a targeted subgenome followed by massively parallel DNA sequencing. We assess this methodology’s performance in detecting sequence and copy number variation in two complex, discontinuous cardiovascular subgenomes comprising 184 targets (exons ± 10 bp, 5′ and 3′ untranslated regions, conserved non-coding regions, and microRNAs) within 54.7 kb (HCM) and 323 targets within 60.1 kb (HDL).

Selected targets were PCR amplified (Supplementary Tables 1 and 2), pooled into respective subgenome target sets, ligated into large DNA concatemers, and isothermally amplified~4,000-fold with Φ29 DNA Polymerase (HCM only) (Fig. 1 and Supplementary Fig. 1). Subgenomic concatemers were then bound to small nitrocellulose membrane filters, which served as subgenomic traps.

We generated eight genomic DNA libraries from three subjects previously genotyped for the HapMap project (HMapl–3), four subjects with abnormal HDL cholesterol levels (HDL1–4), and one HCM subject with a 215 bp deletion and 9 bp insertion in *MYBPC3* (HCM1). Each genomic DNA (1.5 or 2 μg) was sheared and ligated with an adaptor including a 2 bp identifying barcode (Supplementary Table 3). Genomic libraries were then combined into two pools of four and each pool was separately hybridized to HCM and HDL filters. Following stringent washing, captured subgenomic library pools were eluted from the filters, quantified by real-time PCR (Supplementary Fig. 2 and Supplementary Table 4), PCR amplified, and sequenced on a single lane of an Illumina Genome Analyzer I or II14, yielding 2.4–3.7 million (HDL) and 6.5–6.8 million (HCM) 36 bp high quality reads per pooled library.

Captured subgenomic libraries were evaluated for target specificity and uniformity. Of all sequence reads that uniquely aligned to the human reference genome, ~58% (HCM) or ~67% (HDL) corresponded to the targeted segments and 94.4% (HCM) or 89.0% (HDL) of those reads fell within target amplimers (Fig. 2a, Supplementary Fig. 3 and Supplementary Table 5). These specificities are equivalent to enrichments of at most 42,042-fold (HCM) and 39,603-fold (HDL). For each captured subgenomic library, sequence read depths were calculated as the number of high quality base calls at each position. The captured subgenomic libraries’ median sequence read depths were relatively uniform within each target subgenome, which included the portion of the amplified subgenome ≥ 50 nucleotides within target amplimers and extended to amplimers' edges for overlapping or adjacent amplimers (Fig. 2a). Subjects' target subgenomes had median read depths of 277 (HCM) and 145 (HDL). Across target subgenomes, 99.8% of nucleotides had read depths ≥ 20 and 85.6% and 90.8% of read depths were within three-fold and 98.1% and 98.5% of read depths were within ten-fold for HCM and HDL, respectively (Fig. 2b and Supplementary Fig. 3). The distribution of these read depths correlated with amplimer GC-content ($R^2 = 0.30$; Supplementary Fig. 4). Outside of the target subgenomes, median read depths declined dramatically toward and beyond amplimers' edges (Fig. 2a and Supplementary Fig. 5).

Within the target subgenomes, captured subgenomic libraries were sufficiently complex and relatively unbiased; the distribution of unique sequence read starting positions versus read depth approached random sampling (Supplementary Figs. 6 and 7) and the base frequencies at known heterozygous sites were consistent with random binomial sampling ($P = 0.890$, *chi-square*, $N = 121$; Fig. 2c). Moreover, heterozygous base frequencies appeared unaffected by target concatemers' base distributions ($P = 0.768$, *linear regression*, $n = 72$; Supplementary Fig. 8).

Analysis of sequence data predicted 522 heterozygous and 291 non-reference homozygous genotypes at 289 single nucleotide polymorphisms (SNPs), of which 82% were present in dbSNP. 923 genotypes of 316 SNPs were available in HapMap (Release 22) for subjects HMapl–3. At these sites, we observed 732 AA, 121 AB, and 70 BB genotypes (Fig. 2c). These base calls were concordant with HapMap data at all but six sites; all six discordant base calls were confirmed correct by Sanger dideoxy sequencing (data not shown). 18.9 kb of the HCM subgenome was previously sequenced in subject HCM1. We observed complete concordance with these sequences, which included seven heterozygous and two homozygous sequence variants.

Of the 51 novel sequence variants detected, 22 heterozygous and 3 non-reference homozygous genotypes were discovered in subjects HDL1–4. These included three previously unidentified nonsynonymous variants encoding residues conserved through mammalian evolution within *ABCA1*, *ABCG8,* and *NR1H4*. All novel variant calls in subjects HDL1–4 were confirmed by Sanger dideoxy sequencing (Supplementary Table 6; data not shown).

To assess copy number, we compared the number of aligned reads within 1,789 HCM and 2,214 HDL ~32 bp windows across samples, excluding HCM1 from the control set (Supplementary Fig. 9). Sample windows with deviated read counts (*z-test*) were selected as sites of putative CNVs. Across the X-chromosome (*GLA*, *LAMP2*), 849 of 880 windows in male subjects (96%) were detected as putative CNVs and the overall distribution of read counts was significantly lower in males than in females ($P < 2.2 \times 10^{-16}$; one-sided *t-test*). CNV detection by this approach was reproduced in six additional captured subgenomic libraries from control subjects CTL1–5 and an independent repeat of subject HCM1 (Fig. 3a).

Outside of the X-chromosome, read counts in subject HCM1 were consistent with CNVs in 234 out of 1,491 windows across the HCM subgenome (Fig. 3a). These variations were focused within two intervals: six adjacent windows encompassing the previously identified 215 bp deletion and 9 bp insertion (indel) in *MYBPC3* exon 29 ($P = 1.5 \times 10^{-5}$, *binomial test*) and 94 out of 118 windows spanning *MYBPC3* exon 13 to exon 27 (80%; $P = 2.9 \times 10^{-53}$, *binomial test*) (Fig. 3b). PCR amplification followed by Sanger dideoxy sequencing and gel electrophoresis confirmed the smaller exon 29 indel and revealed an 11 kb tandem insertion (Fig. 3c,d, Supplementary Fig. 9 and Supplementary Table 6). The in-frame tandem insertion is predicted to add 1,815 nucleotides of coding sequence; the indel produces a frame shift that is predicted to truncate the encoded protein. Both variants were detected in the affected brother of HCM1 (data not shown), suggesting that they derive from a common ancestor and account for disease.

These copy number results suggest that this strategy can detect (with ≥ 95% sensitivity) deletions ≥ 32 bp (one window) and insertions ≥ 64 bp (two windows). Combining this copy number signal with gapped alignment and local assembly should enable detection of insertions and deletions of any size. No CNVs were detected in 12 subjects across 33,936 autosomal windows. However, unexplained copy number deviations were observed at 10% of autosomal windows in the one subject excluded from the CNV control set, HCM1. Presuming all of these variations to be false, 99.9% specificity for CNV detection can be achieved by combining three CNV windows (~96 bp). We expect this specificity to be improved with larger subject samples sizes.

Here, we present filter-based hybridization capture, a new method for DNA capture with specificity comparable to other hybridization approaches10–12 and sensitivity and uniformity superior to existing methods10–13. Combining filter-based hybridization capture with parallel DNA sequencing allowed efficient detection of both nucleotide sequence and copy number variation. We observed complete sensitivity and specificity for the detection of SNPs and identified a known *MYBPC3* indel and a previously unrecognized 11 kb *MYBPC3* tandem insertion in subject HCM1. Among 210 previously identified HCM mutations in *MYBPC3,* 50% encode a truncated polypeptide (http://cardiogenomics.med.harvard.edu and HLR, JGS, CES unpublished). Yet this 11 kb tandem insertion is the first reported HCM structural variant larger than 33 bp and it could not have been detected by standard dideoxy sequencing of exons.

Filter-based hybridization capture enables efficient and comprehensive study of medium sized subgenomes. Construction of subgenomic traps by PCR and concatemer generation allows for rapid design and synthesis, flexible trap size, and facile modification. The ability to tailor trap size to targeted segments enables uniform coverage from 36 bp single-end reads of captured subgenomic libraries, therein alleviating the requirement for long sequence reads or for additional rounds of concatenation and shearing. The ease of trap modification facilitates normalization, which minimizes the amount of sequence data needed for sensitive variant detection; permits customization and combination of subgenomic traps; and facilitates addition of new targets to existing studies. On the other hand, the generation of subgenomic concatemers by PCR is more labor intensive than oligonucleotide tiling.

Extending subgenomic concatemers to capture very large, discontinuous subgenomes much larger than 1 megabase may require synthesis from synthetic oligonucleotides.

We expect filter-based hybridization capture to scale well to studies of large population cohorts. Here, pools of four or six genomic DNA libraries (250 ng to 1 μg of each) were successfully captured, suggesting that at least 16 libraries (of 250 ng each) may be captured in one reaction. Moreover, these libraries were captured by subgenomic concatemers amplified by Φ29 Polymerase. These properties facilitate cost-effective scaling to larger sample sizes and capture of multiple subgenomes in tandem or one after another (Supplementary Table 7).

Given the simplicity, flexibility, subject scalability, and low cost of this approach, in concert with its demonstrated power to detect both sequence and copy number variants, we suggest that targeted resequencing of subgenomes via filter-based hybridization capture will be broadly useful to research and clinical communities.

Methods

Human Subject

The study population consisted of three individuals from the HapMap project (HMapl-3; NA12717 (CEPH, female), NA19222 (YRI, female), NA19153 (YRI, male)), one subject with HCM (HCM1; male), four subjects with abnormal HDL cholesterol levels (HDL1,3 (male), HDL2,4 (female)), and five control subjects (CTL1 (male), CTL2-5 (female)) (Supplementary Methods). Target concatemers were generated from seven HCM subjects and HapMap individual NA108355 for HCM and five HDL subjects for HDL. All subjects have provided written informed consent Studies were performed in accordance with protocols approved by the Institutional Review Boards of Brigham and Women's Hospital or the Academic Medical Center in Amsterdam. Blood samples were obtained and genomic DNA was extracted according to standard protocols15. Genomic DNA for HapMap individuals was obtained from the Coriell Institute.

Genomic library generation

Genomic DNA libraries were constructed from 1.5–2 μg of genomic DNA. DNA was sonicated in TE (10 mM Tris, pH 8.0; 0.1 mM EDTA) using a Bioruptor (Diagenode) set on HIGH 30 ON/30 OFF for four sessions of 15 minutes (Supplementary Methods). Median DNA fragment sizes, estimated by gel-electrophoresis, were 150–250 bp. Sheared fragments were blunted and phosphorylated using the End-It Repair Kit (Epicentre), purified using the QIAquick PCR Purification Kit (Qiagen), 3′ adenylated with Klenow exo– (NEB), and purified with the Minelute PCR Purification Kit (Qiagen). Products were ligated using the Quick Ligation Kit (NEB) with 5 μl T4 DNA ligase and 10 μl of 100 μM annealed barcoded adaptor in 50 μl at 24 °C for 15 minutes, and purified with a QIAquick column or AMPure beads (2.4× volume of bead mixture; Agencourt). Adaptors consisted of the Illumina genomic DNA adaptor oligonucleotide sequences with the addition of 2 bp barcodes (Supplementary Table 3). Barcodes were designed to contain one G-C and one A-T pair to minimize the differences in melting temperatures and avoid homopolymer runs. Library yield was assessed by real-time PCR using primers PCR_f1 and PCR_r1 (Supplementary Table 4). Approximately half of each library's yield was amplified with the same primers for 6–8 cycles in 12–24 50 μl reactions, purified with a QIAquick column or AMPure beads, quantified by fluorometry, and pooled by equal mass into groups of four or six. The number of amplification cycles was determined by the previous real-time PCR to ensure that the amplifications remained in the linear phase. Libraries were amplified with Modified Gitschier's buffer (67 mM Tris, pH 8.8; 16.7 mM NH_4S0_4; 6.7 mM $MgCl_2$; 10 mM β-mercaptoethanol), 1.5% DMSO, 1 M Betaine inner salt monohydrate (Sigma), 600 μM dNTPs (each; Roche), 600 nM primer (each), 0.5–0.7 μl *Taq*, and 0.25 U Cloned *Pfu* Polymerase (Stratagene) per 25 μl reaction. Thermocycling parameters were 93 °C for 2 min, (93 °C for 20 s, 65 °C for 30 s, 72 °C for 30 s) × # of cycles, 72 °C for 5 min, and 4 °C.

Subgenomic concatemer generation

DNA corresponding to target gene exons and conserved non-coding regions was PCR amplified (Supplementary Methods). PCR products were confirmed by gel electrophoresis, Sanger dideoxy sequencing, and concatemer sequencing. Amplimer concentrations were assessed by picogreen (Invitrogen) and amplimers were pooled in equimolar ratios. To improve the capture uniformity of the HDL subgenome, preliminary HDL amplimers' minimum read depths and HDL concatemers' relative amplimer composition were used to design a separate, supplemental HDL amplified subgenome (Supplementary Methods). For HCM, additional product of two amplimers with 50 bp stretches of GC-content > 85% (*PRKAG2* exon 5 (1× extra) and *TPM1* exon 1 (0.5× extra)) was added to the existing HCM amplified subgenome. Pooled, amplified target exons were depleted of primer dimers using QIAquick PCR Purification and Microcon YM-100 (Millipore) spin columns; blunt ended and phosphorylated using the Quick Blunting Kit (NEB) without heat-inactivation; phenol-chloroform extracted; and ethanol precipitated. Amplimer yield was assessed by the Quant-iT dsDNA BR Assay Kit with a Qubit fluorometer (Invitrogen) or gel electrophoresis. Blunted amplimers (300–600 nM) were ligated to generate concatemers with 400 U μl^{-1} T4 DNA Ligase (NEB) in l× T4 DNA Ligase Buffer at 24 °C for 4 h followed by 4 °C overnight. HDL concatemers and supplemental HDL concatemers were then pooled by mass. As the mass of the HCM concatemers was limited, they were amplified using the REPLI-g Midi Kit (Qiagen) according to the manufacturer's protocol: 10 ng of HCM concatemers was denatured, neutralized, amplified by Φ 29 Polymerase at 30 °C for 16 h, and heat inactivated. Amplified concatemers were analyzed by fluorometry and gel electrophoresis. HDL concatemers were sequenced using the same method as detailed above for genomic library generation, with the addition of an agarose gel size-selection (150–200 bp; Gel Extraction Kit (Qiagen)) following adaptor ligation.

Filter trap generation

Target concatemers (HCM: 20 μg amplified; HDL: 12 μg unamplified (original and supplemental concatemers)) were suspended in 1.5 ml TE, denatured by incubation at 100 °C for 10 min followed by the addition of 1.5 ml 1 M NaOH and incubation at room temperature for 20 min16. Target concatemers were then neutralized with 9 ml Neutralization Solution (10× SSC, 0.25 M Tris-HCl (pH 8), 0.25 N HC1), and applied to a 25 mm (diameter) pre-wet Immobilon nitrocellulose membrane filter (Millipore) at ~1 ml min $^{-1}$ using a multi-filter vacuum manifold (Yeda Scientific, Israel). Filters were washed three times with 4.5 ml 6× SSC, dried, baked at 80 °C under a 10 mmHg vacuum in a vacuum oven (VWR) for 2 h, and stored dry at room temperature17.

Hybridization enrichment

6 mm (diameter) filter circles were cut with a single hole punch from the 25 mm filter loaded with target concatemers. The filter punch was wetted with water, rinsed with 6× SSC, and prehybridized at 60 °C for ≥ 1 h in 4 ml scintillation vials (Wheaton) with 200 μl of prewarmed hybridization solution (6× SSC, 1% SDS, 5× Denhardt's Solution), 500 ng μl^{-1} denatured herring sperm DNA (Invitrogen) and 50 ng μl^{-1} denatured human Cot-1 DNA (Invitrogen). Pools of 250 ng × 4 (HCM hybridization enrichment), 670 ng × 6 (HCM repeat hybridization enrichment), or 1 μg × 4 (HDL hybridization enrichment) of amplified genomic libraries were combined with 1.5 nmoles per μg of library of each Block_f and PCR_r2 oligonucleotides (Supplementary Table 4), 100 μg herring sperm DNA, and 10–20 μg human Cot-1 DNA. This library hybridization mix was speed vacuumed to 15 μl, denatured at 100 °C for 5 min and immediately transferred to an ice water slurry. Following the exchange of the prehybridization solution for fresh prewarmed hybridization solution, the library hybridization mix was added and incubated with the filters at 65 °C for ≥ 17 h. Filters were then rinsed at room temperature with 10 ml 2× SSC and 0.1% SDS 1-2×, 10 ml 0.2× SSC and 0.1% SDS 1-2×, 10 ml 0.1× SSC and 0.1% SDS 0-1×, and 5 ml of 5.2 M Betaine18, 0.1× SSC, and 0.1% SDS and washed at 48 °C in 5 or 20 ml of 5.2 M Betaine, 0.1× SSC, and 0.1% SDS 2× for ≥ 20 min and 0.1× SSC and 0.1% SDS 1× for ≥ 5 min. Filters were rinsed twice in 0.1× SSC and 0.1% SDS at room temperature, transferred to 300 μl of 0.1% SDS, and incubated at 100 °C for 5 min. The eluate was immediately removed, ethanol precipitated, and resuspended in 20 μl TE. Captured subgenomic library yield was assessed by realtime PCR using primers PCR_f2 and PCR_r2 (Supplementary Table 4) and half of the product was amplified with the same primers in 2–10 50 μl reactions for 16–21 cycles (depending on capture yield). Amplified libraries were purified with a QIAquick column and quantified with the Quant-iT dsDNA HS Assay Kit (Invitrogen).

DNA sequencing

Captured subgenomic libraries were single-end sequenced 36 cycles (HCM1, HDL1-4, and HMap1-3) or 26 cycles (HCM1 (repeat) and CTL1-5) using an Illumina Genome Analyzer14. Sequencing images were processed by the Intel FFTW-compiled Illumina GA Pipeline (v0.3.0 or v1.0.0) with automatic generation of cross-talk matrices and phasing/pre-phasing and default read exclusions by base intensity. Generated sequence reads were aligned using Eland with a 25 bp (or 23 bp for 26 cycle sequences) seed to human reference assembled chromosomes, downloaded from NCBI (build 36.3). Read alignments were then expanded to full read length, excluding the 2 bp barcode and the 'T' required for ligation to 3′ adenylated inserts. If after expansion sequence reads had a single best (fewest mismatches) alignment they were considered uniquely aligned. Sequence reads were assigned to subjects according to their 2 bp barcodes by exact matching if both bases had unadjusted quality scores ≥ 10. Subsequent statistical analyses were performed in R19.

Capture specificity and enrichment

Sequence reads (using only the 25 bp alignment seed) were classified as specific to the targeted segments if they aligned to within 200 bp of the edge of a target amplimer. Where p = fraction of reads aligned to the subgenome, t = size of the subgenome, and $\mathbf{g}$ = size of the alignable reference genome, capture enrichment was calculated as: $p/(1 - p) \times (\mathbf{g} - t)/t$. The human reference genome used for alignment included 3.1×10^9 nucleotides; approximately 77% of 25-mers in this reference genome are unique. For each subgenome, the median values across the eight captured subgenomic libraries are reported. For analyses requiring isolated amplimers, only those amplimers separated from all others by at least 300 bp were used.

Capture uniformity

Across the captured subgenome, sequence read depths were calculated as the number of base calls with raw quality scores ≥ 20. Non-uniquely aligned sequence reads were evenly distributed among their best hits if there were 25 or fewer and excluded if there were more than 25. Median read depth and read depth distributions for each captured subgenomic library were calculated and the medians among the eight captured subgenomic libraries were reported. The effect of GC-content and amplimer pooling on capture was assessed by linear regression of the non-supplemented amplimers' mean read depths for the HDL captured libraries over the amplimers' mean read depths in the sequencing of the unadjusted HDL concatemers and a measure of the amplimers' GC-content ($1/\mathrm{L} \times 1/50 \times \Sigma(GC\% - 50)2$, where $GC\%$ is GC-content % calculated in windows of 50 bp at each position in an amplimer of length L).

Capture complexity and composition

At each nucleotide within target subgenomes, the number of unique starting positions among overlapping sequence reads was calculated. To compare this distribution to that of read depths, we adjusted for library sampling by restricting analysis to a fixed number of reads aligned per kb of targeted subgenome. The expected relationship between sequence read depth and unique read starts for libraries of varying complexities ($C_{effective}$) was determined by random sampling without replacement (5,000 iterations) for each read depth (R). The distribution of the 66 (33 bp read length × 2) possible unique starting positions (D) was calculated by randomly sampling unique starting positions from a uniform population of $C_{effective}$ reads. The expected number of unique read starts (U_e) for read depth R was then calculated by sampling R reads from distribution D. Library complexity and bias was also assessed by comparing observed base counts to binomial expectations at known heterozygous sites using the *chi-squared goodness of fit test*. To test whether concatemers' sequences biased DNA capture, we compared base frequencies between captured libraries and corresponding subgenomic concatemers. The base frequencies (*AB1, AB2*) of known heterozygous SNPs, within the unadjusted HDL amplimers, were corrected for additional minor bases ($adjusted\ AB2 = AB2/(AB1 + AB2)$) and linearly regressed over the minor base frequency observed in the sequencing of the unadjusted HDL concatemers and the distance from the furthest amplimer edge.

SNP detection

RefSeq transcript definitions were downloaded from the UCSC Genome Browser (http://genome.ucsc.edu) and SNP positions were downloaded from NCBI (dbSNP build 129). Sequence data were interrogated for single nucleotide variation within target subgenomes (54.7 kb for HCM and 60.1 kb for HDL). Sequence reads were aligned using Novocraft (v1.05.01; http://www.novocraft.com/) and genotype consensus calls were made using a Bayesian model implemented by MAQ20 component glfProgs (v0.l; http://maq.sourceforge.net), with a prior probability of heterozygosity of 0.001. Variant calls were then filtered by MAQ using default thresholds, except that the minimum consensus quality threshold for a given base and its six flanking bases was lowered to 4. Called novel variants were confirmed by PCR amplification of the corresponding amplimers from subject genomic DNA and Sanger dideoxy sequencing (Agencourt).

Copy number detection

We assessed copy number across each subgenome within non-overlapping ~32 bp windows that were centered at multiples of 32 bp from amplimers' edges. There were 1,693 windows assessed across the HCM subgenome for the HCM repeat hybridization enrichment. For each subject, the cross-correlation was calculated between the number of sequence reads aligned across each subgenome by Eland to the forward and reverse strands. Each set of read counts was then shifted by half of the offset with the peak correlation21. The resulting shifted read counts were tallied within each window for each strand and normalized against the total number of reads aligned to that strand for each subject Window count means and standard deviations were calculated, excluding subject HCM1 and excluding all males (HCM) or females (HCM repeat) from the X-chromosome control set. To avoid common variation, windows with a maximum relative standard deviation (standard deviation/mean) > 0.25 were removed if they fell outside of the pre-defined targets and optimized if overlapping with targets. Optimization consisted of a greedy series of extensions or contractions of 5 bp in the direction of the minimum relative standard deviation until a stable minimum relative standard deviation was reached, the relative standard deviation fell below 0.25, or the window was smaller than 20 bases. Windows were not permitted to expand outside the amplified subgenome or to overlap with another window; windows smaller than 20 bp or with median read counts < 10 were removed. Copy ratio for each window was estimated as the mean for the two strands of the ratio of the sample read count to the mean read count across all control samples. Putative copy number variations were identified as sample windows with counts deviated from the control count distribution, as defined by the product of the p-values for each strand's read count (z-test) $< 1 \times 10^{-3}$. The significance of stretches of putative copy number variation were assessed by calculating their likelihood under binomial sampling from all window copy number calls within each subgenome, excluding the X-chromosome. Variants were confirmed by PCR amplification using primers listed in Supplementary Table 6 and the Expand Long Range dNTPack (Roche), Sanger dideoxy sequencing, and gel electrophoresis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank J. Aach, J. Gorham, J. Kim, J.B. Li, and L. Wang for helpful discussions and advice; J. Kastelein and J.A. Kuivenhoven for HDL subject genomic DNA and for assistance in selecting HDL target genes; S. Depalma for analysis of Affymetrix custom sequencing array data; L. Merrill and D. Sonkin at Partners Center for Personalized Genetic Medicine of Harvard Medical School for Illumina sequencing and support; B. Richter, D. Gurgul, S. Roylance, J. Xu, and A. Zschau at Partners Healthcare for access to high performance computing cluster and data storage; T. Levi and B. McDonough for subject recruitment and genomic DNA isolation; S. Cox and M. Iaculli for generating HCM amplimers; and the Drosophila RNAi Screening Center for access to their Molecular Devices Analyst GT plate reader. G.K.H. was supported by grants from the Dutch Heart Association (NHS 2007R001) and Saal van Zwanenbergstichting. This work was supported by grants from NCI, NHLBI, HHMI, and Leducq Foundation.

Works cited

1. Maron BJ, et al. Circulation. 1995; 92:785–789. [PubMed: 7641357]
2. Alcalai R, Seidman JG, Seidman CEJ. Cardiovasc Electrophysiol. 2008; 19:104–110.
3. Hovingh GK, et al. Curr Opin Lipidol. 2005; 16:139–145. [PubMed: 15767853]
4. Wilier CJ, et al. Nat Genet. 2008; 40:161–169. [PubMed: 18193043]
5. Kim JB, et al. Science. 2007; 316:1481–1484. [PubMed: 17556586]
6. Hillier LW, et al. Nat Methods. 2008; 5:183–188. [PubMed: 18204455]
7. Chiang DY, et al. Nat Methods. 2009; 6:99–103. [PubMed: 19043412]
8. Goossens D, et al. Hum Mutat. 2009; 30:472–476. [PubMed: 19058222]
9. Lovett M. Curr Protoc Hum Genet. 2001; Chapter 6(Unit 6):3. [PubMed: 18428299]
10. Bashiardes S, et al. Nat Methods. 2005; 2:63–69. [PubMed: 16152676]
11. Albert TJ, et al. Nat Methods. 2007; 4:903–905. [PubMed: 17934467]
12. Gnirke A, et al. Nat Biotechnol. 2009; 27:182–189. [PubMed: 19182786]
13. Porreca GJ, et al. Nat Methods. 2007; 4:931–936. [PubMed: 17934468]
14. Bentley DR, et al. Nature. 2008; 456:53–59. [PubMed: 18987734]
15. Moore D, Dowhan D. Curr Protoc Mol Biol. 2002; Chapter 2(Unit 2):1A.
16. Parnes JR, et al. Proc Natl Acad Sci U S A. 1981; 78:2253–2257. [PubMed: 6166005]
17. Brown T. Curr Protoc Mol Biol. 2001; Chapter 2(Unit 2):9B.
18. Rees WA, Yager TD, Korte J, von Hippel PH. Biochemistry. 1993; 32:137–144. [PubMed: 8418834]
19. Team RDC. R: A language and environment for statistical computing. 2008
20. Li H, Ruan J, Durbin R. Genome Res. 2008; 18:1851–1858. [PubMed: 18714091]
21. Kharchenko PV, Tolstorukov MY, Park PJ. Nat Biotechnol. 2008; 26:1351–1359. [PubMed: 19029915]

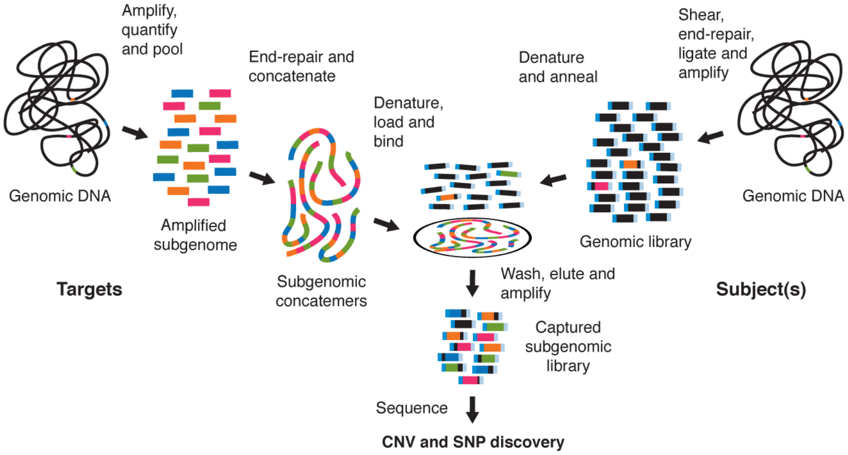

Figure 1.

Filter-based hybridization capture schematic.

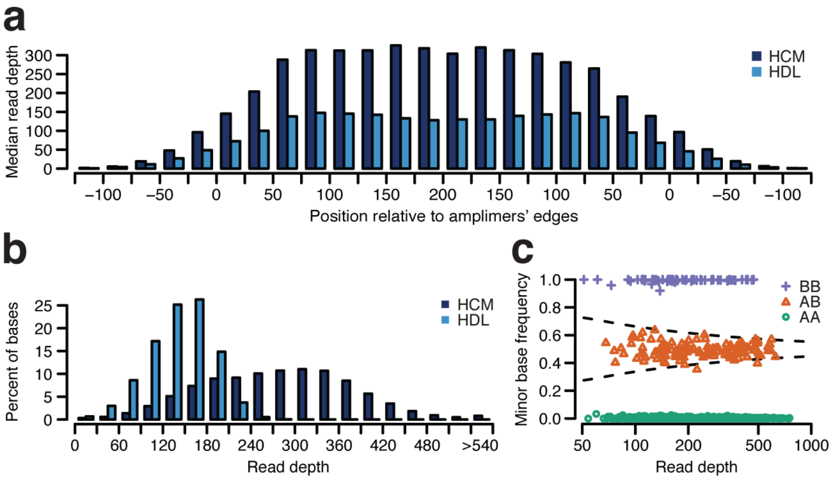

Figure 2.

Captured subgenomic library uniformity and single nucleotide polymorphism detection for HCM and HDL subgenomes. (a) Median read depths within isolated amplimers at different locations relative to amplimers' edges, (b) Distribution of read depths within the target subgenomes. (**c**) Plot of minor base frequencies versus read depth for all HapMap genotyped SNPs in subjects HMapl–3. HapMap and Sanger dideoxy genotype calls: AA (○), AB (△), BB (+), N (□); sequencing genotype calls: AA (green), AB (orange), BB (purple). The 0.05 and 99.95 percentiles of the expected base frequencies at heterozygous bases are indicated (dashed lines).

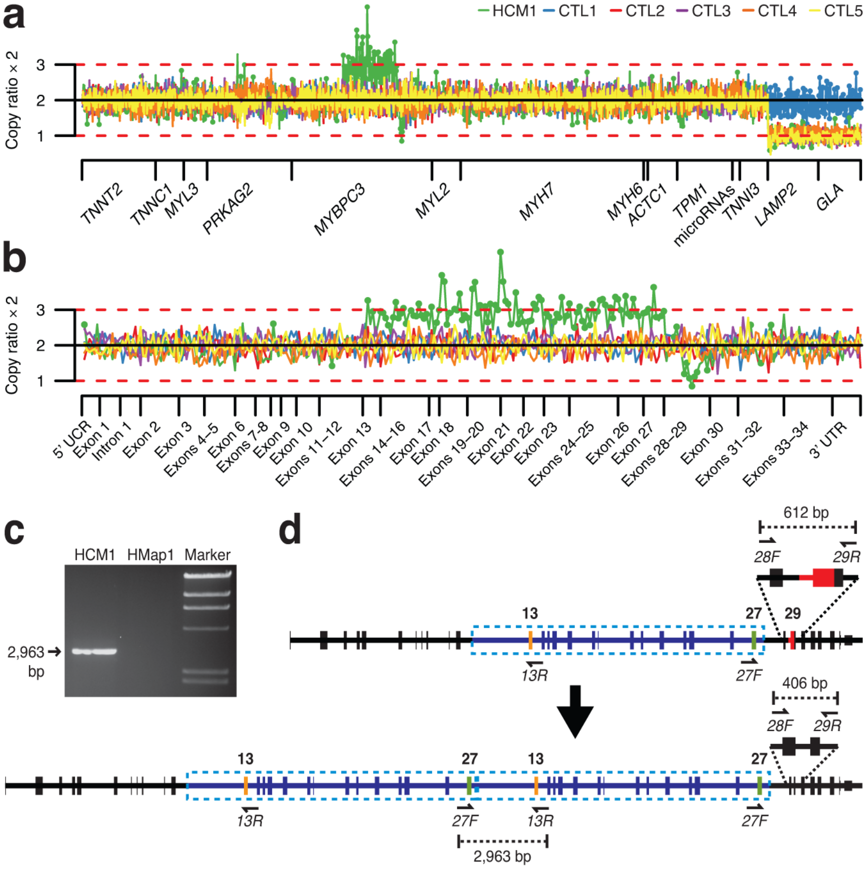

Figure 3.

Detection of copy number variation. Plot of twice the copy ratio (ratio of sample read counts to the control mean read count) for each subject's captured subgenomic library (CTL1 (female); HCM1 and CTL2–5 (male)) within ~32 bp windows across (a) the entire HCM subgenome or (b) *MYBPC3*. Solid circles indicate sample windows with deviated read counts. UCR = upstream conserved region, (c) 1% agarose gel showing PCR product using primers 13R and 27F for subjects HCM1 and HMapl. DNA marker is ΦX174 DNA *Hae*III digested (Roche) and indicated size is in basepairs. (d) Complex rearrangement model including an 11 kb tandem insertion spanning exon 13 to exon 27 (blue and boxed) and a 215 bp deletion and 9 bp insertion including parts of exon 29 and intron 28 (red). Annealing sites for primers 13R, 27F, 28F, and 29R and the size of their predicted PCR products are indicated.