Published in final edited form as:

Methods. 2009 September ; 49(1): 63–69. doi:10.1016/j.ymeth.2009.06.009.

Integration of genetic and genomic methods for identification of genes and gene variants encoding QTLs in the nonhuman primate

Laura A. Cox^{1,2,*}, Jeremy Glenn¹, Simon Ascher³, Shifra Birnbaum¹, and John L. VandeBerg^{1,2}

¹Department of Genetics, Southwest Foundation for Biomedical Research, San Antonio, TX 78227

²Southwest National Primate Research Center, Southwest Foundation for Biomedical Research, San Antonio, TX 78227

³Duke University School of Medicine, Durham, NC 27708

Abstract

We have developed an integrated approach, using genetic and genomic methods, in conjunction with resources from the Southwest National Primate Research Center (SNPRC) baboon colony, for the identification of genes and their functional variants that encode quantitative trait loci (QTL). In addition, we use comparative genomic methods to overcome the paucity of baboon specific reagents and to augment translation of our findings in a nonhuman primate (NHP) to the human population. We are using the baboon as a model to study the genetics of cardiovascular disease (CVD). A key step for understanding gene-environment interactions in cardiovascular disease is the identification of genes and gene variants that influence CVD phenotypes. We have developed a sequential methodology that takes advantage of the SNPRC pedigreed baboon colony, the annotated human genome, and current genomic and bioinformatic tools. The process of functional polymorphism identification for genes encoding QTLs involves comparison of expression profiles for genes and predicted genes in the genomic region of the QTL for individuals discordant for the phenotypic trait mapping to the QTL. After comparison, genes of interest are prioritized, and functional polymorphisms are identified in candidate genes by genotyping and quantitative trait nucleotide analysis. This approach reduces the time and labor necessary to prioritize and identify genes and their polymorphisms influencing variation in a quantitative trait compared with traditional positional cloning methods.

Keywords

Nonhuman primate (NHP); quantitative trait loci (QTL); cardiovascular disease (CVD); functional polymorphism; discordant sib-pairs; gene array; gene networks

© 2009 Elsevier Inc. All rights reserved.

*Corresponding Author: Laura A. Cox, Ph.D., Associate Scientist, Department of Genetics, Core Scientist, Southwest National Primate Research Center, Southwest Foundation for Biomedical Research, 7620 NW Loop 410, San Antonio, TX 78227, lcox@sfbgenetics.org.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1. Introduction

Our laboratory is using the baboon as a model to understand gene-environment interactions that influence the process of atherogenesis. Central to these studies is the mapping and identification of genes underlying variation in cholesterol metabolism. The commonly used methods for positionally cloning novel genes encoding QTLs are labor and time intensive. In order to identify novel genes encoding QTLs, we have developed an efficient strategy to prioritize candidate genes and identify functional polymorphisms in the gene(s) influencing variation in the QTL. This strategy uses information from the baboon linkage map, the human genome sequence, annotated and predicted genes in the human genome, the pedigreed, phenotyped baboon colony at the Southwest National Primate Research Center (SNPRC), quantitative measures for atherosclerosis-related traits, gene expression profiling, and gene network analysis tools. Using this approach we identified endothelial lipase as the gene and variants within endothelial lipase that influence variation in HDL₁-C [1].

2. Overall strategy

To identify novel cardiovascular related genes that contribute to atherogenesis or dyslipidemia, we initially use classical genetic methods to identify chromosomal regions containing loci that influence the trait of interest. The foundation resource for these studies is a baboon genetic linkage map that we constructed using 284 random microsatellite markers from the human linkage map [1]. In addition to constructing the linkage map, scientists in the Department of Genetics at the Southwest Foundation for Biomedical Research have collected quantitative trait data on more than 150 lipid and lipoprotein quantitative traits in the same 2044 pedigreed baboons that were used to construct the linkage map (<http://baboon.sfbgenetics.org/BabPedigreesBL.php>). Genome scans were performed for each quantitative trait to identify quantitative trait loci (QTL) influencing each atherosclerosis related trait (e.g. [2–7]). After QTL identification, QTL regions of interest are fine mapped to reduce the chromosomal region of interest (e.g. [1]). After identifying and refining the QTL region of interest (ROI), we use a modified genomic expression profiling method integrated with bioinformatics analyses to prioritize candidate genes in the QTL region of interest. This method is dependent upon collection of target tissues relevant to the phenotype from baboons fed diets relevant to lipoprotein metabolism. In addition, we use network analysis of transcriptome data to identify networks underlying each phenotype and networks connected to QTL ROI genes to augment gene prioritization. The evaluation of candidate genes in the QTL ROI is all-inclusive with analysis of both annotated and predicted genes. Prioritized candidate genes are then analyzed in detail by identification and genotyping of polymorphisms that may regulate variation in the quantitative trait. Functional polymorphisms are identified by statistical functional analyses and validated by molecular genetic analyses [1]. An outline of this approach is shown in Figure 1.

3. Resources for QTL functional polymorphism identification

As mentioned above, data were collected from the SNPRC pedigreed baboons for quantitative traits related to atherosclerosis. Genome scans were performed for these quantitative traits using the baboon linkage map and a number of QTLs were identified. We have identified 19 regions on 13 chromosomes for which there is significant evidence of one or more QTLs influencing lipid and lipoprotein traits (Rainwater et al., under review). QTL identification requires: 1) a pedigreed colony of animals, in this case baboons (3a); 2) a genome linkage map based on a genotyped, pedigreed colony of animals [8] (3b.); 3) the quantitative trait of interest must be quantified in the pedigreed, genotyped colony of animals (3c.) and these data used to perform genome scans to detect QTLs; and 4) the genome scans must reveal a significant QTL (typically LOD > 3.0) for the quantitative trait of interest (3d.). These are well-established

statistical genetic methods (presented in detail in [1]). The resources necessary for QTL identification are then used for identification of the gene and variant(s) that influence variation in the quantitative trait mapping to the QTL.

4. Resources for candidate gene identification

Candidate gene prioritization integrates available resources from the pedigreed baboon colony, the baboon linkage map and the annotated human genome sequence. This includes the defined QTL ROI, baboon sib-pairs discordant for the quantitative trait of interest, and collection of target tissues from the discordant sib-pairs before and after a relevant environmental challenge. The feasibility of identifying sib-pairs discordant for the phenotype and the accessibility of target tissues under specific environmental conditions underline the advantages of using a nonhuman primate for identification of genes underlying human complex diseases.

4a. Narrowing the ROI and identifying genes in the ROI

Genome scans to identify QTLs using the baboon genome linkage map, which has a 7 cM resolution, results in a broad linkage signal. Therefore, prior to identifying genes encoded in the QTL ROI we first fine map the QTL region to reduce the size of the target region. As with many model organisms, no physical map for baboon currently exists. Previously we screened human microsatellite markers to identify new microsatellite markers that were polymorphic in baboon and thus suitable for fine mapping [8,9]. However this approach had a success rate less than a 25%. We have increased the success rate of polymorphic microsatellite marker identification to 67% using comparative genomic methods. When we began our QTL gene identification projects, the rhesus genome had not yet been sequenced; therefore, we performed the comparative genomics using the human genome map as the reference genome. These methods however can be used for any species with a non-sequenced or unassembled genome (target) against a species with a sequenced, assembled genome (reference).

With the availability of baboon genome sequence in the NCBI Trace Archive (<http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?>; [10]), we use baboon trace archive sequence compared with the well-annotated human genome sequence to identify new microsatellite markers in the QTL region. The underlying assumption is that for conserved syntenic regions, repetitive elements, encoded genes, non-coding RNAs, regulatory elements, etc. are conserved between target and reference genomes. Multiple species' genome sequences can be aligned (Vista Genome Browser; <http://pipeline.lbl.gov/cgi-bin/gateway2>; [11,12] for the region of interest to test the extent of element conservation between reference and target syntenic regions. Based on our work using human, rhesus and baboon microsatellite markers in the baboon genome and the human genome, we know that repetitive elements common to two species may be polymorphic in one species but not the other. Therefore, sequence alignment will provide a list of repetitive elements that are good candidates for microsatellite markers based on repeat length; however, variation in a repetitive element length must be tested empirically (e.g. [1, 13]).

We devised a comparative genomics approach to identify and test putative baboon microsatellite markers. First we define the genomic sequence included in our region of interest by identifying the physical map location of microsatellite markers flanking the region of interest using the reference genome. We enter the microsatellite identifiers into the University of California Santa Cruz (UCSC) Genome Browser (<http://genome.ucsc.edu>; [14]) query box and retrieve the genomic locations delimiting the QTL region of interest. We then scan human genomic DNA sequence in the region of interest at 1 million basepair (Mbp) blocks in 5 Mbp intervals for repetitive elements of 12 or more di, tri, or tetra repeats using the UCSC Genome Browser Table Browser function ([http://genome.ucsc.edu/](http://genome.ucsc.edu)) [15] and use the UCSC Genome Browser Table Browser function [15] to list all microsatellite and simple repetitive elements

in the region of interest including 300 bp of flanking sequence 5' and 3' of each repeat. After excluding 1 Mbp regions that already contain microsatellite markers in the baboon linkage map, putative markers are prioritized by proximity to annotated genes, providing another link to the reference genome map. After repetitive element identification, we use the BLAST tool (http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi?PAGE=Nucleotides&PROGRAM=blastn&BLAST_SPEC=TraceArchive&BLAST_PROGRAMS=megaBlast&PAGE_TYPE=BlastSearchrch; [16] with the predicted PCR product sequence from the human genome against the baboon Trace Archive to determine the repetitive element repeat number in baboon and to identify baboon flanking sequence for primer design. Since many species now have genomic sequence data available in the NCBI Trace Archive (<http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?>; [10]), but the genomes have not yet been assembled, this tool is also useful as a secondary reference sequence when the primary reference sequence is more evolutionarily divergent to the target than the secondary reference. Parameters for primer design include PCR product length of 150–300 bp, PCR primer length of 18–24 nucleotides (nt), GC content greater than 55%, and a T_m of 55–68°C. Also, the stability (ΔG) of primer-template duplexes must be less than 10°C difference between the T_m of each primer and primer/dimer pair formation is not allowed. We use the BLAT alignment tool (<http://genome.ucsc.edu/cgi-bin/hgBlat>; [17] with the human genome to increase the likelihood of primer specificity. To optimize the chances of identifying polymorphic baboon microsatellite markers for the pedigreed baboon colony, for marker testing we use a panel of baboons that represent a large portion of the genetic diversity in the pedigreed colony. Markers that amplify and are polymorphic are then tested for heritability using 2–3 baboon nuclear families (i.e. sire, dam, 2–3 offspring). Polymorphic microsatellite markers are genotyped for the phenotyped, pedigreed baboons and the new markers are included in the linkage map and the genome scan for the quantitative trait is repeated.

After narrowing the QTL ROI, the region is aligned with the human genome using the microsatellite markers and the genomic sequence in the interval is retrieved. The microsatellite sequence including flanking sequence is entered into the human genome BLAT search tool in the UCSC Genome Browser (<http://genome.ucsc.edu/cgi-bin/hgBlat>; [17]. Output from this search includes the sequence alignment and physical map location in the reference (human) genome. Once the reference genome region of interest has been defined, it is possible to identify all known and predicted genes in the interval as well as non-coding RNAs and regulatory elements.

4b. Identification of discordant sib-pairs

To identify baboons for the positional cloning of the gene encoding a QTL, we perform phenotypic and genotypic analysis of the pedigreed baboon population and identify baboon sib-pairs discordant for the quantitative trait. The sib-pairs differ by at least one standard deviation for the quantitative trait. In addition, members of each selected sib-pair do not share IBD (identical-by-descent) alleles in the chromosomal region of interest [2].

4c. Environmental challenge

A major strength of the baboon model for studies of atherosclerosis is the ability to control diet for a specified period of time. For these studies, baboons are exposed to an environmental challenge that is relevant to the quantitative trait of interest and is likely to differ between the discordant sibs. For lipid and lipoprotein-related traits, baboons are fed *ad libitum* commercial monkey chow (basal diet; Teklad) for 7 weeks and then fed a high-cholesterol, high-fat diet (1.7 mg/kcal cholesterol and 40% of calories as fat from lard) [18] for 7 weeks. Blood is collected before and after the challenge diet and the traits measured from serum [2]. For quantitative traits where the QTL peak LOD score differs between the chow diet and high-

cholesterol, high-fat diet, we predict that the gene influencing the trait will be differentially expressed between the two diets.

4d. Target tissues

Another major strength of the baboon model is the feasibility of collecting target tissue biopsies for analysis. This is a key component of a candidate gene identification and prioritization project and it is unlikely that a controlled dietary challenge for selected individuals could be conducted and tissue biopsies collected before and after the challenge in humans. Because of the genetic and physiologic similarity between baboon and human, the findings from these studies are highly likely to be directly applicable to humans. Because liver plays a central role in cholesterol metabolism, for cholesterol metabolism-related traits liver biopsies are collected from the discordant sib-pairs before and after the 7-week high-cholesterol, high-fat diet challenge.

Biopsies are collected by an SNPRC veterinarian from sedated animals with biopsy needle placement guided by anatomical landmarks. Consistency of biopsy collection is ensured due to the skills and the experience of the SNPCR veterinary staff performing biopsies on liver. Three liver punches are collected from each animal for each biopsy procedure. Tissue samples are quick frozen in liquid N₂ and stored at -80°C

5. Prioritization of genes in the QTL ROI

A central hypothesis for our strategy to identify the gene encoding a QTL is that the gene must be expressed in the QTL interval. In this section we integrate the information and resources described above to define all target tissue expressed genes in the region of interest and to define networks relevant to the expressed genes. We developed a Chromosomal Region Expression Array (CREA) strategy that allows us to evaluate all DNA sequences in the region of interest that may encode the gene influencing the QTL. We do not limit our approach to the analysis of known genes; the CREA is inclusive for all genes, ESTs (expressed sequence tags), and predicted genes within the QTL region of interest. To interrogate the arrays, we use heterologous RNA from the tissue most likely to be relevant to the quantitative trait. In addition, we collect tissues from sibling baboons discordant for the quantitative trait in order to minimize genetic variation due to genetic background and to maximize genetic differences for the gene(s) encoding the QTL. Using this approach, we can significantly reduce the number of candidate genes in the QTL region of interest [2].

5a. Design of a chromosomal region expression array

After defining the QTL ROI, we use the well-annotated human genome to identify all known and predicted genes in the region of interest. The microsatellite markers delimiting the QTL ROI are used to retrieve the entire genomic region using the UCSC Genome Browser. The UCSC Table Browser tool is used to provide all annotated genes, predicted genes, and expressed sequence tags (ESTs) in the ROI. The output data for the RefSeq Gene Track and the Gene Scan Prediction Track includes GenBank ID number, exon start site and exon stop site for each exon in the annotated or predicted gene. The Table Function is used to download all exon sequences for each of these genes and predicted genes (<http://genome.ucsc.edu/cgi-bin/hgTables>).

To design gene specific primers for a list of genes for which the cDNA sequence has not yet been determined, we use a comparative genomics approach. With the availability of the baboon genome sequence in the NCBI Trace archives (<http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?>; [10]), each of the gene sequences are compared against the baboon genome to identify baboon coding region sequence using

Sequencher DNA analysis software (Gene Codes, Inc.). The baboon coding region sequence is then used to design oligonucleotides for each gene and predicted gene in the QTL ROI. Oligonucleotide design constraints include: 1) oligonucleotide length ≥ 65 nucleotides; 2) 45–55% GC content; 3) no tetranucleotide repeats; 4) no significant hairpin loops (less than 7 bonds in a hairpin); and 5) optimal probe with highest T_m and the highest negative ΔG value for GC clamp (Oligo Primer Analysis Software; Molecular Biology Insights, Inc). After oligonucleotide design, sequence specificity is confirmed by performing an NCBI-BLAST search and uniqueness of the oligonucleotide is confirmed allowing less than 90% maximum identity with non-target sequences. After oligonucleotide specificity is confirmed, oligonucleotides are synthesized and nylon based arrays printed with oligonucleotides spotted in triplicate.

A modification to this approach is to use a commercially available gene array and supplement the expression profile results with a chromosome region expression array. The rationale being that the CREA will contain predicted genes and ESTs either not included on the array or that do not have a quality signal on the human commercial array due to undetectable gene expression or differences between baboon and human in the array oligonucleotide sequence.

5b. Identify genes expressed in the region of interest

The CREA is interrogated with RNA generated probes from discordant sib-pair samples. Complementary RNA probes are synthesized from total RNA by synthesizing cDNA and using the cDNA to synthesize radioactively labeled cRNA by including $\alpha^{32}\text{P}$ -UTP in the *in vitro* transcription reaction according to manufacturer's instructions. For first strand cDNA synthesis, primers are annealed to the mRNA templates by incubating 1 μl total RNA (1 μg) with 1 μl 5 μM T7-Oligo (dT) (Ambion) for 6 min. at 70°C and then cooling the sample to 4°C for 2 min. The mRNA is then reverse transcribed by adding 1 μl 5X first strand buffer (Invitrogen), 0.5 μl 100 mM DTT, 0.375 μl 10 mM dNTP mix, 0.25 μl RNase inhibitor (40 Units/ μl , Invitrogen), 0.5 μl SuperScript II (200 Units/ μl , Invitrogen), and 0.375 μl DEPC-treated water to the RNA-primer mixture. The reaction is incubated for 1 hr at 42°C and the SuperScript II heat-inactivated for 10 min. at 70°C. The cDNA second strand is synthesized by adding 7.5 μl 5X second strand buffer (Invitrogen), 0.75 μl 10 mM dNTP mix, 0.25 μl DNA Ligase I (10 Units/ μl , Invitrogen), 1 μl DNA Polymerase I (10 Units/ μl , Invitrogen), and 0.25 μl RNase H (2 Units/ μl , Invitrogen) to the first strand reaction and incubated for 2 hrs. at 16°C. One microliter of T4 DNA polymerase (5 Units/ μl , Invitrogen) is then added to the reaction and incubated for 10 min. at 16°C. The cDNA is precipitated by first adding 2 μl Glycogen (5 mg/ml) as a carrier and then 80 μl DEPC-treated water followed by 0.6 volumes 5 M ammonium acetate and 2.5 volumes cold absolute ethanol. After precipitating the cDNA overnight at –20°C, samples are centrifuged at 17,000 $\times g$ for 30 min. at 4°C. DNA pellets are washed in 70% ethanol (with DEPC-treated dH₂O) and air-dried.

Complementary RNA synthesis by *in vitro* transcription is performed using the MAXIscript Kit (Ambion) by adding 7 μl nuclease free water to the cDNA pellet and then adding: 2 μl 10 $\times$ Transcription Buffer, 1 μl 10 mM ATP, 1 μl 10 mM GTP, 1 μl 10 mM CTP, 1 μl 250 μM UTP (12.5 μM), 5 μl $\alpha^{32}\text{P}$ UTP (3000 Ci/mM), and T7 Enzyme Mix. The reaction is incubated for 1 hour at 37°C and the cDNA template removed by adding 1 μl DNase I, incubating 15 min. at 37°C, and the DNase I inactivated with 1 μl 0.5 M EDTA. Complementary RNA (cRNA) is cleaned using a Sephadex G-50 column (Roche mini Quick Spin RNA Columns) according to manufacturer's instructions. An aliquot of cRNA is counted in a scintillation counter to determine synthesis efficiency. Purified cRNA is then fragmented using fragmentation buffer (Ambion) according to manufacturer's instructions.

For hybridization of cRNA probes with nylon membrane oligonucleotide arrays, arrays are prehybridized for 2 hours at 42°C with UltraHyb Buffer (Ambion). Denatured cRNA probe is

added to the membrane in prehybridization buffer and hybridized for 42 hours at 42°C. The nylon membrane arrays are washed: once in 2X SSC / 0.1% SDS for 5 min., twice in 2X SSC / 0.1% SDS for 5min. at 35°C, and twice in 0.2X SSC / 0.1% SDS for 10min. at 35°C. After washing, nylon membrane arrays are air-dried and placed in phosphorimager cassettes for image capture. Each gene array image is acquired by exposing nylon filters to phosphorimager cassettes and capturing the image with a Phosphorimager (Storm 840, Amersham Biosciences) using ImageQuantTL Image Analysis Software (Amersham Biosciences). Each image is loaded into ImaGene 5.6 Microarray Image Analysis software (Biodiscovery, Inc.) and the template containing the annotated grid is applied to the image. Pooled targets are used as reference points to properly align the grid over the image. Data are then quantified by compiling numerical intensity values, quality measurements, and spot location for each spot. Background for each spot is measured in a rectangular region around the spot and a circular buffer region. The median of background values within 5×5 spots is subtracted from the signal value of the center spot for background correction (Local Group Median option). A manual quality control check is then performed on the data to remove miscalled spots (due to background or slight membrane defects) and to flag quality discrepancies for limited manual evaluation and editing. Each dataset is refined by verifying positive and negative controls. Data for empty, poor quality and absent spots (including spots that do not have an acceptable signal from its duplicate) are removed. Intensities and quality values are averaged for replicate spots.

After data cleaning, array data are uploaded into GeneSifter (GeneSifter.net, VizX Labs), all-median normalized and log₂ transformed. Box plots are inspected to ensure that the median for each group is zero and the variance among groups is similar. Data are filtered by spot quality. All genes that pass the quality filter are subjected to pair-wise analysis by t-test and for group analysis by ANOVA assuming unequal variance. $p < 0.05$ is considered statistically significant.

5c. Transcriptome profiling and network analysis

Whole genome expression profiling of liver RNA is performed using Human BeadChip (Illumina Inc., San Diego, CA). Complementary RNA probes are synthesized from baboon total RNA by first synthesizing cDNA and then using the cDNA to synthesize fluorescently labeled cRNA by including biotinylated UTP in the *in vitro* transcription reaction. For first strand cDNA synthesis, primers are annealed to the mRNA templates by incubating 1 µl total RNA (1 µg) with 1 µl 5 µM T7-Oligo (dT) for 6 min. at 70°C and then cooling the sample to 4°C for 2 min. The mRNA will then be reverse transcribed by adding 1 µl 5X first strand buffer, 0.5 µl 100 mM DTT, 0.375 µl 10 mM dNTP mix, 0.25 µl RNase inhibitor (40 Units/µl), 0.5 µl SuperScript III (200 Units/µl) and 0.375 µl DEPC water to the RNA-primer mixture. The reaction is incubated for 1 hr at 42°C and the Super Script III heat-inactivated for 10 min. at 70°C. The cDNA second strand is synthesized by adding 7.5 µl 5X second strand buffer, 0.75 µl 10 mM dNTP mix, 0.25 µl DNA ligase I (10 Units/µl), 1 µl DNA Polymerase I (10 Units /µl) and 0.25 µl RNase H (2 Units/µl) to the first strand reaction and incubated for 2 hrs at 16°C. One microliter T4 DNA polymerase (5 Units/µl) will then be added to the reaction and incubated for 10 min. at 16°C. The cDNA is precipitated by first adding 2 µl glycogen (5 mg/ml) as a carrier and then 80 µl DEPC-treated water followed by 0.6 volumes 5M ammonium acetate and 2.5 volumes cold absolute ethanol. After precipitating the cDNA overnight at -20°C, samples are centrifuged at 17,000 × g for 30 min. at 4°C. DNA pellets are washed in 70% ethanol (with DEPC-treated dH₂O) and air-dried. Complementary RNA synthesis by *in vitro* transcription is performed using the TotalPrep™ RNA Labeling Kit (Ambion, Austin, TX) by adding 7 µl nuclease free water to the cDNA pellet and then adding: 2 µl 10× Transcription Buffer, 1 µl 10 mM NTP mix containing biotinylated UTP, 1 µl 10 mM GTP, 1 µl 10 mM CTP, 1 µl 250 µM UTP (12.5 µM) and T7 Enzyme Mix. The reaction is incubated for 1 hour at 37°C and the cDNA template removed by adding 1 µl DNase I, incubating 15 min. at

37°C and inactivating with 1µl 0.5 M EDTA. Complementary RNA (cRNA) is cleaned using a Sephadex G-50 column (Roche mini Quick Spin RNA Columns) according to manufacturer's instructions.

Gene expression data are acquired using BeadScan software (Illumina Inc., San Diego, CA) and basic data cleaning performed using BeadStudio software (Illumina Inc., San Diego, CA). Array data are all-mean normalized and $\log_2$ transformed using GeneSifter software (GeneSifter.Net, VizX Labs, Seattle, WA). Statistical analyses of array data is performed by t-test using GeneSifter software for pair wise comparisons. We perform a repeated measures ANOVA using the data on each gene across the groups of test and control animals where appropriate [84].

5d. Identify ROI genes contained in networks that are responsive to the environmental challenge

Network analysis of genes reveals QTL ROI expressed genes that are directly connected to networks of genes which are differentially expressed between the groups of discordant sibs and between the diets for each sib group. QTL ROI expressed genes contained in these diet responsive networks are considered higher priority candidate genes than genes not connected to networks. In addition, ontological pathway (<http://www.geneontology.org/>) [19] and KEGG pathway (www.genome.jp/kegg/) [20] analysis of the whole genome expression data provides detailed data on individual genes in the context of that gene's role in described biological/biochemical pathways which may reveal insights into molecular mechanisms by which the gene could influence the QTL.

Networks analysis of whole genome expression data is performed using Ingenuity Pathway Knowledge Base. Each data set containing gene identifiers and corresponding expression values is uploaded into the Ingenuity Pathways Analysis application (Ingenuity® Systems, www.ingenuity.com). Each gene identifier is mapped to its corresponding gene object in the Ingenuity Pathways Knowledge Base. These genes are overlaid onto a global molecular network developed from information contained in the Ingenuity Pathways Knowledge Base. Networks of focus genes will be generated based on their connectivity using algorithms developed and implemented by Ingenuity® Systems.

For pathway analysis, genes that exhibit significant differences in expression are overlaid onto Ontological Pathways (<http://www.geneontology.org/>) [85] and KEGG Pathways (www.genome.jp/kegg/) [86] using GeneSifter software. The ontological and KEGG pathway analyses provide detailed data on individual genes in the context of that gene's role in described biological/biochemical pathways. Pathways are considered significantly altered from the control gene expression profiles if the z-score for that pathway is less than -2 or greater than +2. z-scores are calculated in GeneSifter using the following formula: $z\text{-score} = [r - n(R/N)] / [\sqrt{(n(R/N))(1 - R/N)(1 - ((n-1)/(N-1)))}]$; where R = total number of genes meeting selection criteria, N = total number of genes measured, r = number of genes meeting selection criteria with the specified GO term and n = total number of genes measured with the specific GO term [87].

6. Prioritizing ROI genes

Genes are prioritized based on expression profiles, proximity to the peak LOD score, biological relevance to the trait of interest, and association with cardiovascular disease QTLs from other studies. A positional table is generated using the UCSC table browser that includes annotated genes, expressed genes, and QTLs. The QTL track includes human, mouse and rat QTL data annotated as a component of the rat genome database project [21]. The table is then filtered to retain all CREA expressed genes. Mean values for both sib-pair groups from chow and high-

cholesterol, high-fat diets are added to the table for each CREA expressed gene. In addition, GeneCards (<http://genome-www.stanford.edu/genecards/index.shtml>; [22] and OMIM (<http://www.ncbi.nlm.nih.gov/omim>; [23] databases are accessed for known function(s) of each annotated gene.

Genes are ranked first by consistency of each gene expression profile with the QTL signal. For example, if the QTL signal was observed for the high-cholesterol, high-fat diet but not the chow diet, keeping in mind that the discordant sib-pairs were selected based on their contribution to the QTL signal, we predict that the gene influencing the quantitative trait will be differentially expressed between low and high sib-pairs on the high-cholesterol, high-fat diet, but not the chow diet. Therefore, in this example the highest priority genes are differentially expressed between low and high sib-pairs on the high-cholesterol, high-fat diet. In addition, the low and high sib-pairs show either no differences on the chow diet or no differences in expression for the low sib-pairs comparing chow and high-cholesterol, high-fat diets. Genes included in this group are further prioritized based on biological relevance to the genes' known function with the quantitative trait and proximity to the peak LOD score. Predicted genes can't be prioritized based on known function and are therefore prioritized by expression profiles and location relevant to related QTLs mapped to the QTL region of interest. Using this approach for the chromosome 18 QTL influencing HDL1-C, we began with 354 genes and predicted genes in the region of interest and reduced the number of candidates down to 3 genes [24].

It is possible that the gene encoding a QTL is not differentially expressed between groups of animals discordant for the quantitative trait. For example, if a functional polymorphism encoded a nonsynonymous polymorphism it is possible that this nucleotide change would not correlate with expression levels of the mRNA. Therefore, our prioritization scheme contains a second phase where no high priority candidate genes are identified from differential gene expression profiles or where further interrogation of high priority differential genes does not result in identification of functional polymorphisms.

In the second phase of candidate gene prioritization, all genes that are not differentially expressed between discordant groups and by diet are ranked by proximity to the peak LOD score. In addition, network analysis is performed on all expressed genes, annotating expressed and differentially expressed genes in each network and comparing networks between discordant groups and by diet. In our experience, genes relevant to the quantitative trait that are not differentially expressed are included in networks that are differentially activated between discordant groups. Peak LOD score proximity, network data and biological information are used to prioritize the candidate genes. Proteins encoded by top priority candidate genes are evaluated for expression between discordant groups and by diet (Cox et al., manuscript in preparation). Nackley et al. [25] have shown that both synonymous and nonsynonymous SNPs that influence a quantitative trait can influence gene product expression levels through alterations in mRNA secondary structure. They have shown that synonymous polymorphisms in addition to nonsynonymous polymorphisms can have a pronounced effect on the level of protein expression. Therefore, proteins differentially responsive to the dietary challenge between discordant groups are ranked highest. Using this approach we have identified genes influencing LDL response to dietary fat (Cox et al., manuscript in preparation).

7. Identification and genotyping of polymorphisms in the discordant sibs for the top priority candidate genes

Sequence data from the baboon Trace Archive (<http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?>; [10]) is used to design sequencing primers for resequencing the top priority candidate gene(s). Sequence polymorphisms are identified by

sequencing the candidate gene(s) in a panel of animals discordant for the quantitative trait of interest. To ensure that polymorphisms are identified resequencing is performed on single alleles; all genomic DNA fragments that are sequenced from the panel of discordant animals are subcloned and 8 clones for each animal in the panel are sequenced. To limit the number of polymorphisms identified and genotyped, we focus on the panel of discordant baboons for resequencing. Because these baboons differ by at least one standard deviation for the quantitative trait of interest and each selected sib-pair in the panel does not share IBD (identical-by-descent) alleles in the chromosomal region of interest, then polymorphisms that may influence variation in the gene encoding the QTL will be present in this group of animals.

For resequencing, genomic DNA (50 ng) is amplified using species-specific gene primers, PCR buffer and Taq DNA Polymerase. PCR products are subcloned into pTOPO (Invitrogen) and transfected into competent cells (Invitrogen). Plasmid DNA is purified (Qiagen) and sequenced (Applied Biosystems, Inc.). Sequencing products are purified using Exonuclease I (USB) and Shrimp Alkaline Phosphatase (USB) and size fractionated (Applied Biosystems Inc). Sequence data are imported into Sequencher (Gene Codes, Inc.) for alignment and identification of polymorphisms. Nucleotides and insertion/deletions are considered polymorphic if they are validated by their presence in either 1) two or more baboons in the sib-pair panel and data are consistent using primers from both directions, or 2) one baboon and the data were consistent for sequence data from multiple clones, i.e. 4 clones with one variant and 4 clones for a second variant.

8. Functional polymorphism identification and validation

After prioritization of candidate genes, the functional polymorphism(s) in the gene, that is the polymorphisms that influence variation in the quantitative trait, must be identified. To date, robust predictive tools for the identification of functional polymorphisms are not available. In our baboon HDL₁-C QTL candidate gene study of endothelial lipase (*LIPG*), we evaluated the orthologous human gene for conserved non-coding sequences (Vista Genome Browser, <http://pipeline.lbl.gov/cgi-bin/gateway2>). These analyses showed conservation from mouse to human for two regions in the 5' flanking region of *LIPG*. One region was immediately upstream of the 5' untranslated region (UTR) and one region was located -2,446 bp from the transcription start site. No polymorphisms were identified in the conserved region proximal to the 5' UTR and none of the polymorphisms located in the upstream conserved region influenced *LIPG* expression of HDL₁-C variation. Furthermore, our study of *LIPG* revealed 2 functional single nucleotide polymorphisms (SNPs) and one deletion-insertion polymorphism (DIP). SiteSeer [26] was used to determine predicted transcription factor binding to the *LIPG* promoter binding for the functional DIP and SNPs in the 5' flanking region. One SNP was located in a predicted transcription factor binding site and the insertion for the DIP included a predicted transcription binding site; however the second SNP was not located in any predicted or annotated regulatory element (Cox et al., 2007). From this study and others we know that functional polymorphisms are not necessarily in linkage disequilibrium with neighboring polymorphisms. Therefore, all polymorphisms in each candidate gene must be genotyped in the population from which the QTL was detected and quantitative trait nucleotide analyses must be performed on each polymorphism to identify functional polymorphisms. In cases where candidate genes are predicted to be differentially expressed and the variation in gene expression influences variation in the quantitative trait of interest, polymorphisms in potential regulatory regions as well as the coding regions must be identified. In cases where candidate genes are not differentially expressed, polymorphisms in coding sequence and untranslated regions must be identified. In addition, resequencing is most likely to reveal informative polymorphisms if animals representative of variation in the quantitative trait of interest are resequenced for polymorphism identification.

For identification of functional polymorphisms, all polymorphisms in the gene of interest must be genotyped. We have found direct sequencing of genomic DNA to be the most effective method for genotyping candidate gene polymorphisms identified in the discordant sib-pair. All pedigreed baboons phenotyped for the quantitative trait are genotyped. Resequencing primers from polymorphism identification are used and sequencing is performed as described above. Genotype data cleaning, genotype analysis, and quantitative trait nucleotide analysis are described in detail in Cox et al., 2007 [27].

9. Translating findings from nonhuman primates to humans

The goal of our studies is to translate baboon disease gene identification and functional variant identification to humans. For genes we have sequenced previously, we have not found the same SNPs and DIPs and microsatellite markers in humans as baboons. However, we have found the same “class” of polymorphism that has the same effect in the orthologous genes. For example, we identified a splice site mutation in baboon apolipoprotein(a), a gene that influenced LDL-C, that results in a transcript positive null mRNA [28]. The same polymorphism was not found in humans; however, a polymorphic splice site was found that resulted in transcript positive null alleles in humans [29].

In another example, we identified functional polymorphisms in the promoter of endothelial lipase that play roles in transcriptional activation of the gene and influence HDL₁-C [27]. Although these same polymorphisms are not found in human, the same class of polymorphisms is found in the human endothelial lipase gene promoter (Cox et al., in preparation). To identify conserved classes of functional polymorphisms, we align the region of the gene containing the baboon functional polymorphism with the human orthologous gene region. The functional polymorphism including flanking regions are queried using the UCSC BLAT alignment tool [17] to identify similar regions in the gene. Using this approach we have identified endothelial lipase gene promoter regions in humans that are conserved with baboons and contain polymorphisms. Experiments are underway to determine if these polymorphisms influence gene expression and HDL₁-C.

10. Conclusions

Integration of resources generated from the SNPRC pedigreed baboon colony, including a baboon linkage map, quantitative trait data, genotype data, target tissue accessibility from animals under controlled environmental conditions, combined with the well-annotated human genome and genomic methods such as transcriptome profiling and network analysis, provide a wealth of data on positional candidate genes encoding QTLs. Analysis of these data provides a mechanism to prioritize all expressed candidate genes in a QTL interval and dramatically reduce the number of genes that must be resequenced and genotyped for functional polymorphism identification. Because the baboon is genetically and physiologically very similar to humans, the identification of genetic polymorphisms influencing variation in disease-related quantitative traits is directly applicable to the identification of disease-related polymorphisms and the mechanisms by which they influence disease risk in humans.

Appendix

Custom oligonucleotide arrays for CREA construction were synthesized and spotted onto nylon membranes by Sigma Aldrich (www.sigmaaldrich.com/life-science/custom-oligos.html). CREA images were captured using a Storm Phosphorimager (Molecular Dynamics.). Human whole genome expression profiling is performed using Illumina BeadChips with the BeadXpress the BeadStation with BeadStudio software (Illumina Inc.).

Acknowledgements

This work was supported by National Institutes of Health grants P01 HL028972, P51 RR013986. This investigation was conducted in part in facilities constructed with support from Research Facilities Improvement Program Grant Number C06 RR013556 and C06 RR015456 from the National Center for Research Resources, National Institutes of Health.

References

1. Cox LA, Birnbaum S, Mahaney MC, Rainwater DL, Williams JT, Vandeberg JL. *Circulation* 2007;116:1185–1195. [PubMed: 17709635]
2. Cox LA, Birnbaum S, Vandeberg JL. *Genome Res* 2002;12:1693–1702. [PubMed: 12421756]
3. Kammerer C, Cox L, Mahaney M, Rogers J, Shade R. *Hypertension* 2001;37:398–402. [PubMed: 11230307]
4. Kammerer CM, Rainwater DL, Schneider JL, Cox LA, Mahaney MC, Rogers J, Vandeberg JL. *Hypertension* 2003;41:854–859. [PubMed: 12624008]
5. Rainwater DL, Kammerer CM, Mahaney MC, Rogers J, Cox LA, Schneider JL, Vandeberg JL. *Atherosclerosis* 2003;168:15–22. [PubMed: 12732382]
6. Vinson A, Mahaney MC, Cox LA, Rogers J, Vandeberg JL, Rainwater DL. *Atherosclerosis*. 2007
7. Voruganti VS, Tejero ME, Proffitt JM, Cole SA, Freeland-Graves JH, Comuzzie AG. *Obesity* 2007;15:2043–2050. [PubMed: 17712122]
8. Cox LA, Mahaney MC, Vandeberg JL, Rogers J. *Genomics*. 2006
9. Rogers J, Mahaney MC, Witte SM, Nair S, Newman D, Wedel S, Rodriguez LA, Rice KS, Slifer SH, Perelygin A, Slifer M, Palladino-Negro P, Newman T, Chambers K, Joslyn G, Parry P, Morin PA. *Genomics* 2000;67:237–247. [PubMed: 10936045]
10. Shumway, M.; Alexeyev, V.; Church, D.; Salzberg, S. Series. 2005. 1.1:[Available from: <http://www.ncbi.nlm.nih.gov/Traces>
11. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. *Nucleic Acids Res* 2004;32:W273–W279. [PubMed: 15215394]
12. Shah N, Couronne O, Pennacchio LA, Brudno M, Batzoglou S, Bethel EW, Rubin EM, Hamann B, Dubchak I. *Bioinformatics* 2004;20:636–643. [PubMed: 15033870]
13. Cox LA. *J. Med. Prim* 2002;31:1–12.
14. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. *Genome Res* 2002;12:994–1006.
15. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. *Nucl. Acids Res* 2004;32:D493–D496. [PubMed: 14681465]
16. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. *J. Mol. Biol* 1990;215:403–410. [PubMed: 2231712]
17. Kent WJ. *Genome Res* 2002;12:656–664. [PubMed: 11932250]
18. McGill HC Jr, McMahan CA, Kruski AW, Kelley JL, Mott GE. *Arteriosclerosis* 1981;1:337–344. [PubMed: 6956267]
19. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. *Nat Genet* 2000;25:25–29. [PubMed: 10802651]
20. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. *Nucleic Acids Res* 2004;32:D277–D280. [PubMed: 14681412] Database issue
21. Rapp JP. *Physiol Rev* 2000;80:135–172. [PubMed: 10617767]
22. Rebhan M, Prilusky J. *Electrophoresis* 1997;18:2774–2780. [PubMed: 9504809]
23. Omim, Series, 2008. 2007 Dec 06. Available from: <http://www.ncbi.nlm.nih.gov/omim/>
24. Cox, L.; Birnbaum, S.; Mahaney, M.; Vandeberg, J. *Proceedings of the XIII International Congress on Genes, Gene Families, and Isozymes Medimond*; 2005.
25. Nackley AG, Shabalina SA, Tchivileva IE, Satterfield K, Korchynskiy O, Makarov SS, Maixner W, Diatchenko L. *Science* 2006;314:1930–1933. [PubMed: 17185601]

26. Boardman PE, Oliver SG, Hubbard SJ. Nucleic Acids Res 2003;31:3572–3575. [PubMed: 12824368]
27. Cox LA, Birnbaum S, Mahaney MC, Rainwater DL, Williams JT, Vandeberg JL. Circulation 2007;116:1185–1195. [PubMed: 17709635]
28. Cox LA, Jett C, Hixson JE. J Lipid Res 1998;39:1319–1326. [PubMed: 9684734]
29. Ogorelkova M, Gruber A, Utermann G. Hum Mol Genet 1999;8:2087–2096. [PubMed: 10484779]

Figure 1. Strategy for QTL functional polymorphism identification. The resources needed and the integration of these resources are shown. Numbers on the left hand side of the figure correspond to section numbers in the text.