

Laura A. Cox

8.1 Introduction

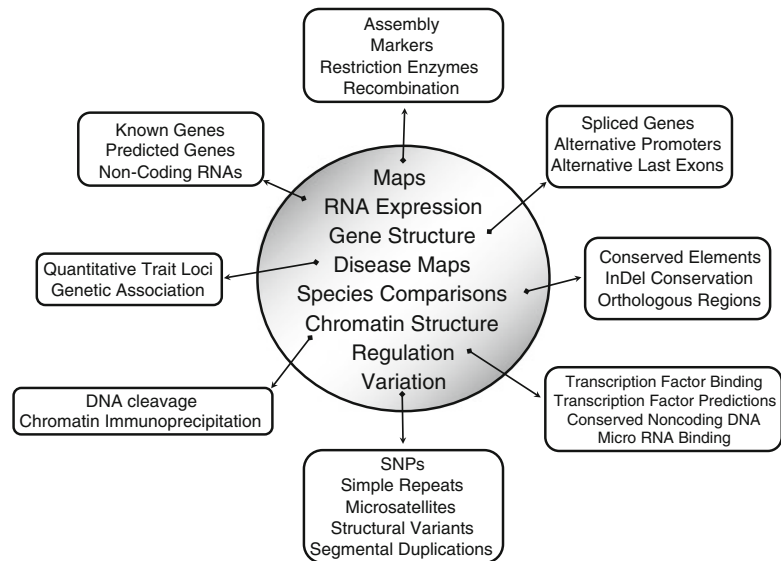
Comparative genomics is a research approach that uses bioinformatics tools to integrate data from multiple genomes for numerous types of information from each genome and is used to address questions ranging from the role played by gene x environment interactions in human health and disease to evolutionary relationships of species using phylogenetic analysis methods. The identification of genomic similarities and/or genomic differences can be used to build models or “systems” to develop a better understanding of specific biological systems. Completion of each successive mammalian genome sequence, each sequence assembly, each whole genome expression study, and each whole genome polymorphism study has exponentially increased the information available and the bioinformatics tools available that support comparative genomic analyses. The databases from which data are drawn for comparative analyses include low-resolution physical maps, high-resolution physical maps, statistical genomic maps, genomic assembly maps, restriction enzyme maps, recombination maps, gene expression profile data, noncoding RNA annotation, alternative

gene structure annotation, conserved coding and noncoding elements, insertion/deletion elements, orthologous chromosomal regions, polymorphisms, and structural variants (Fig. 8.1). Different types of tools have been developed that use assembled genomes as frameworks for mapping disease-related quantitative trait loci and mapping genetic associations with human diseases and model organisms, for predicting expressed genes, noncoding RNAs, protein-DNA binding sites, RNA-DNA binding sites, and for aligning genomic regions from multiple species for phylogenetic analysis such as phylogenetic shadowing. Furthermore, these publicly available databases link to other websites that provide detailed information on each data point (e.g., gene, protein, single nucleotide polymorphism, etc.). Consequently, initial investigation of a genetic element (chromosome, gene, promoter, etc.) using basic bioinformatics tools will typically reveal extensive information about the system of interest before a single laboratory experiment has begun.

In this chapter, I will discuss comparative genomic tools used for the study of gene x environment interactions underlying complex diseases by comparison of a model organism genome with the human genome. In studies of pedigreed baboons, my colleagues and I are using comparative genomic tools combined with classical genetic tools to identify genes that influence variation in complex disease. I will present examples of the use of these new tools for the identification of concordant quantitative

L.A. Cox (✉)
Department of Genetics, Texas Biomedical Research
Institute, 7620 NW Loop 410, San Antonio, TX
78227, USA
e-mail: lcox@txbiomed.org

Fig. 8.1 Overview of comparative genomic resources. Each central element is found in public databases and each connected element links out to one or more specialty databases



trait loci, regulatory elements, conserved gene domains, gene expression profiles, and gene networks relevant to cardiovascular disease. In addition, I will present how these tools can be used for identification of specific polymorphic nucleotides that influence variation in a cardiovascular disease-related quantitative trait in a nonhuman primate. I will also show how these results can be used for identification of polymorphisms that are likely to influence human quantitative trait variation and complex disease. The marked reduction in sequencing costs will soon provide additional data on numerous individuals in multiple species that will again significantly expand the information gained using comparative genomics tools. This information will provide greater power to predict the genes and the polymorphisms in these genes that influence quantitative traits, which will decrease discovery time for the identification of these genes, and their functional polymorphisms.

Our laboratory is using the baboon as a model to determine how genetic variation influences atherogenesis. Central to these studies is the identification of genes underlying variation in cholesterol metabolism. The commonly used methods for positionally cloning novel genes are labor- and time-intensive. In addition, these methods are complicated when localizing and identifying genes regulating multigenic traits. In

order to identify novel genes encoding QTLs, we have developed an efficient strategy to identify candidate genes. This strategy uses information from the baboon linkage map, the human genome sequence, including annotated and predicted genes, the pedigreed baboon colony at the Southwest National Primate Research Center (SNPRC), the quantitative measures for atherosclerosis-related traits, and the human genome database in concert with gene expression array methods. Using this approach we identified the gene and variants within the gene that influence variation in a size fraction of high-density lipoprotein cholesterol (HDL₁-C) (Cox et al. 2007).

To identify novel cardiovascular related genes, that is, genes not previously known to contribute to atherogenesis or dyslipidemia, we initially used 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 (Cox et al. 2007). In addition to the linkage map, scientists in the Department of Genetics at the Texas Biomedical Research Institute have collected quantitative trait data on more than 150 lipid and lipoprotein quantitative traits in the same 951 pedigreed baboons that were used to construct the linkage map. Genome scans were performed for each

quantitative trait to identify quantitative trait loci (QTL) influencing each atherosclerosis-related trait (e.g., Cox et al. 2002; Kammerer et al. 2001, 2003; Rainwater et al. 2003; Vinson et al. 2007; Voruganti et al. 2007). After QTL identification, QTL regions of interest were fine mapped to reduce the chromosomal region of interest (e.g., Cox et al. 2007). After identifying and refining the QTL region of interest, we used a modified genomic expression profiling method integrated with bioinformatics analyses to prioritize candidate genes in the QTL region of interest. The evaluation of candidate genes in the QTL region of interest is all-inclusive, with analysis of both annotated and predicted genes. Prioritized candidate genes were then analyzed in detail by identification and genotyping of polymorphisms that may regulate variation in the quantitative trait. Functional polymorphisms were identified by statistical functional analyses and validated by molecular functional analyses (Cox et al. 2007). Furthermore, we used transcriptome profiling data analyzed using bioinformatics tools to identify genetic pathways and networks underlying each phenotype. In this chapter, I will describe the methods used for each of these steps and provide examples from our work studying genes underlying variation in atherosclerosis-related traits.

8.2 QTL Identification and Fine Mapping

As with many model organisms, no physical map for baboon exists and the genome sequence map is currently in draft form with numerous gaps. In the absence of a well-annotated baboon reference genome, we use comparative genomic methods to: (1) decrease the QTL region of interest by fine mapping (Sect. 8.2.3) and (2) determine known and predicted genes in the reduced region of interest for each QTL (Sect. 8.3.1). When we began our QTL gene identification projects, the rhesus genome had not yet been sequenced; therefore, we performed the comparative genomic examples presented here using the human genome map as the reference genome. These methods, however, can be used for any species

with a nonsequenced genome (target) against a species with a sequenced genome (reference). To identify gene(s) encoding QTLs, the closer the two species are related evolutionarily the more informative the comparison.

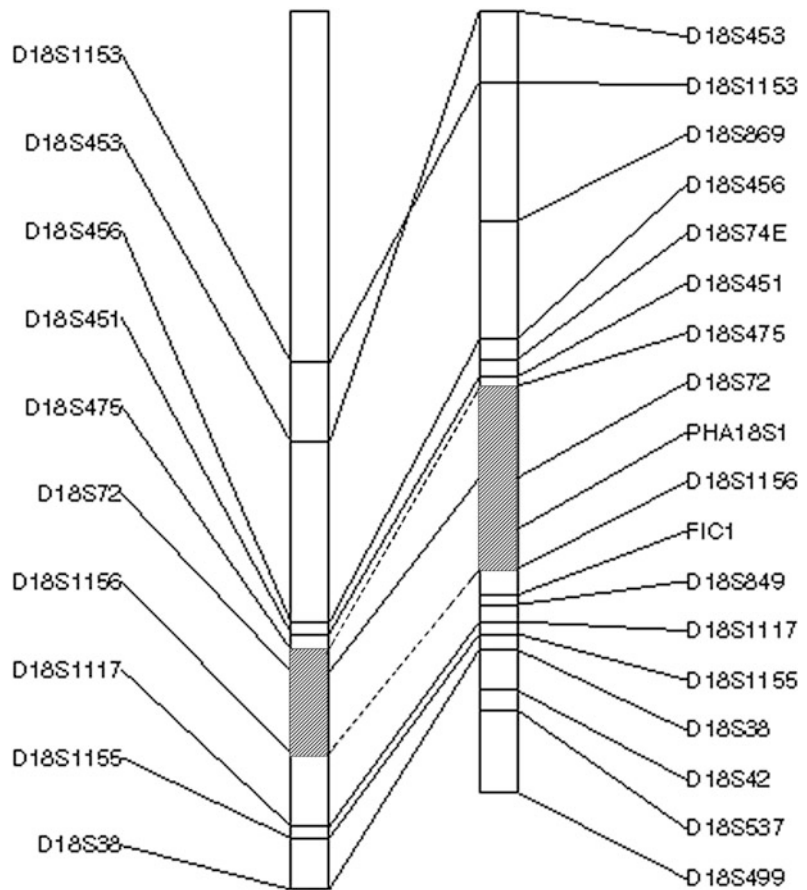
8.2.1 QTL Identification

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. In this chapter, we use results from our work on two QTLs identified from the genome scans as examples, one QTL influencing HDL₁-C (a size fraction of HDL-C) (Cheng et al. 1988) on baboon chromosome 18 (Mahaney et al. 1998) and one QTL influencing the hypertension trait sodium lithium counter transport (SLC) activity on baboon chromosome 5 (Kammerer et al. 2001). For the HDL₁-C QTL on baboon chromosome 18, the two-point linkage analysis showed a peak LOD score of 7.32 at marker D18S72. We defined the region of interest for the QTL, which is the chromosomal region most likely to include the gene(s) influencing HDL₁-C variation, as the two LOD support interval, i.e., the region included in the area under the QTL curve from the peak out to the two LOD drop in the curve (Cox et al. 2002). For the SLC QTL on baboon chromosome 5, we obtained evidence for an SLC QTL with a peak LOD score of 9.3 located near marker D4S1645 (human chromosome 4 is the orthologue of baboon chromosome 5). This QTL accounts for approximately two thirds of the total additive genetic variation in SLC activity in baboons.

8.2.2 Sequence Alignment for Fine Mapping Chromosomal Regions

DNA sequence alignment of the target and reference genomes is necessary for the identification of repetitive elements that can be used to fine map the region of interest and reduce the number of

Fig. 8.2 Alignment of baboon chromosome 18 (left) HDL QTL region of interest (hashed lines) with human chromosome 18 (right) using genotyped microsatellite markers for baboon (modified from http://baboon.txbiomedgenetics.org/Bab_Results/GraphicMaps/chrom18.php)



candidate genes that must be analyzed. By genotyping microsatellite markers and repetitive elements common to both the target and reference genomes it is possible to align the target and reference linkage maps (e.g., Fig 8.2). Because the reference genome has both a linkage map and whole genome sequence, the alignment of the reference genome syntenic block with the target species' QTL region of interest. The underlying assumption is that for conserved syntenic regions, repetitive elements, encoded genes, noncoding 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>) (Frazer et al. 2004; Shah et al. 2004) 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., Cox 2002; Cox et al. 2007).

8.2.3 Fine Mapping a QTL Region of Interest

To fine map a QTL region of interest, we must identify microsatellite markers that are amplifiable and polymorphic in our target (baboon)

species. When we first began fine mapping baboon QTLs, we screened human microsatellite markers in baboon that were included in the human genome linkage maps (Cox et al. 2006a; Rogers et al. 2000). Although this strategy was successful identifying new markers for the baboon linkage map, it was extremely inefficient with less than a 25 % success rate for marker identification. In addition, some of these markers did not yield clean PCR products making genotyping difficult and some markers were not very polymorphic. Therefore, we devised a comparative genomics approach to identify and test putative baboon microsatellite markers. First we defined 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 entered the microsatellite identifiers into the University of California Santa Cruz (UCSC) Genome Bioinformatics browser (<http://genome.ucsc.edu/>); (Kent et al. 2002) query box and retrieved the genomic locations delimiting the QTL region of interest. We then scanned 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 Bioinformatics, Table Browser function (<http://genome.ucsc.edu/cgi-bin/hgTables>) (Karolchik et al. 2004) 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 contained microsatellite markers in the baboon linkage map, putative markers were prioritized by proximity to annotated genes, providing another link to the reference genome map. Because we know there are sequence differences between human and baboon but we don't know what nucleotides differ, we designed two pairs of PCR primers for amplification of each repetitive element (Oligo v6.89, Molecular Biology Insights, Inc.). Parameters for primer design included PCR product length from 150 to 300 bp, PCR primer length of 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 no primer/dimer pair formation is allowed. We used the BLAT alignment tool (<http://genome.ucsc.edu/cgi-bin/hgBlat>; Kent 2002) with the human genome to ensure primer specificity (Cox et al. 2009).

With the recent availability of baboon genomic sequence in the NCBI Trace Archive (<http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?>), we have added an additional step to this procedure. 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=BlastSearch; Altschul et al. 1990) 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, but the genomes have not yet been assembled, this tool is also useful as a second reference sequence when the first reference sequence is not as evolutionarily close to the target as the second “Trace Archive” reference.

To optimize the chances of identifying polymorphic baboon microsatellite markers for the pedigreed baboon colony, we used a panel of 12 baboons that represent a large portion of the genetic diversity in the pedigreed colony. Genomic DNA was amplified by PCR for each target region using a fluorescently labeled forward primer and unlabeled reverse primer. PCR products were size-fractionated in an automated sequencer Applied Biosystems, Inc. (ABI) and genotyped using Genotyper software. Heritability was tested for each polymorphic marker by genotyping 2–3 baboon nuclear families (i.e., sire, dam, 2–3 offspring). If multiple polymorphic, heritable markers were identified for a chromosomal interval, the most polymorphic marker was selected for genotyping. Selected microsatellite markers were genotyped for the phenotyped, pedigreed baboons. The new markers were then included in the linkage map and the genome scan for the quantitative trait repeated.

8.3 Characterizing the Refined Region of Interest

8.3.1 Sequence Alignment

After refining the QTL region of interest, the chromosomes must be aligned and the genomic sequence in the region of interest must be retrieved for the reference sequence. Although microsatellite markers from the linkage map were used to align the chromosomes before fine mapping the region of interest, the same analysis must be performed including the new markers. It is possible that small chromosomal rearrangements not apparent with the original alignment are apparent with the new markers. As described in Sect. 8.2.2, microsatellite markers common to both target and reference genomes were used to align the target and reference genomic regions. If repetitive elements were used for genotyping the target genome, these contain physical map “addresses” in the reference genome that can also be used to tie the target genome to the framework of the reference genome. To do so, the microsatellite sequence including flanking region sequence was entered into the reference genome BLAT search tool in the UCSC Genome Bioinformatics browser (<http://genome.ucsc.edu/cgi-bin/hgBlat>; Kent 2002). Output from this search will include the sequence alignment and physical map location in the reference 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 noncoding RNAs and regulatory elements. In addition, it is possible to determine if other scientists have mapped QTLs (Rapp 2000) or genetic disease associations (Becker et al. 2004) with that chromosomal interval.

8.3.2 Sequence Alignment for Rearranged Chromosomal Regions

It is not unusual to find chromosomal rearrangements such as inversions when comparing target and reference linkage maps. In addition, it is not unusual to find QTL regions of interest that

include areas of rearrangement between target and reference chromosomes. A central element in the identification of genes encoding QTLs is to include all possible candidate genes in the initial screening process. Therefore, chromosomal rearrangements and the portions of chromosome that overlap with QTL regions of interest must be defined as clearly as possible for inclusion and exclusion of candidate genes. That is, all possible genes that can be excluded should be excluded in order to reduce the number of genes that must be interrogated; however, because one does not want to exclude a candidate gene that lays in the region of interest the region must be defined as clearly as possible.

With this in mind, we often see rearranged regions with inadequate linkage map data (i.e., number of microsatellite markers in the linkage map) to precisely identify the chromosomal regions of interest in the reference chromosome. An example of this is shown in Fig. 8.3a, where the QTL region of interest includes baboon chromosome 5 from D4S414 to D4S2365. This QTL region of interest spans a chromosomal inversion when comparing baboon against human. Using the mapped microsatellite markers, the region of the orthologous human chromosome (chromosome 4) for the region from D4S414 to D4S1645 is clear; all the markers included between these markers are the same for baboon and human with the order from p to q reversed. Whereas, the DNA that should be included in the region from D4S1645 to D4S2365 is not as clear. D4S2365 borders the baboon region of interest and D4S413 is outside the region of interest for baboon and this is consistent in human. So, the conserved chromosomal region should be p-ter to D4S2365; however, D4S414 is outside the region of interest in baboon but flanks the region likely to include QTL region of interest DNA. In this case, the investigator has 2 choices: (1) fine map additional markers between D4S2365 and D4S414 or (2) include all genes and expressed genes as candidates for the D4S2365–D4S414 region knowing that some genes are likely to be outside the region of interest. Due to the required time and resources for candidate gene interrogation

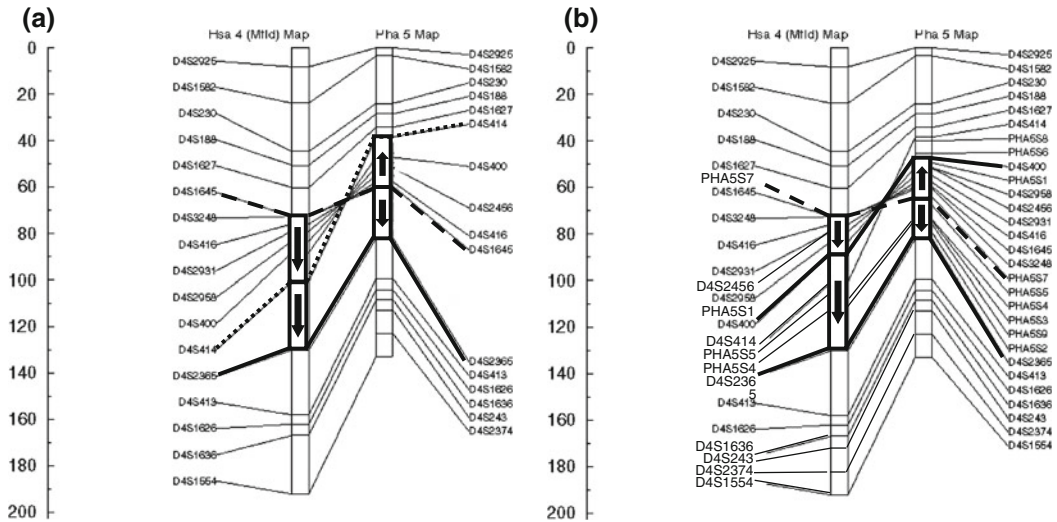


Fig. 8.3 Identification of target DNA in region of interest with rearranged reference chromosome. Human chromosome 4 (Hsa 4) is on the *left*, baboon chromosome 5 (Pha), the ortholog to Hsa 4 is on the *right*. *Lines* between the chromosomes show marker order. *Boxes* show chromosome segment conservation. *Arrows* indicate relative

directions. **a** shows chromosome comparison by mapped microsatellite markers with region of interest in *bold boxes* and chromosome direction indicated by *arrows*. **b** shows reduced region of interest with inclusion of additional microsatellite markers in the linkage map

and prioritization, the first option is usually worth the time invested.

In the example of a QTL mapping to baboon chromosome 5, we chose to fine map the region of interest and more clearly define the region of rearrangement as described in Sect. 8.2.3. Figure 8.3b shows the QTL region of interest after including additional markers in the linkage map. Markers that were identified specifically from baboon genomic sequence as described in Sect. 8.2.3 are indicated by the “PHA” identifier; all of these sequences can be assigned locations relative to the mapped human microsatellite markers using the human genome sequence for the human orthologous chromosome using the UCSC Genome Browser BLAT alignment tool. Addition of the new markers more narrowly defines the chromosomal breakpoint between the human and baboon orthologous chromosomes, shown by the dashed line PHA557 and narrows the QTL region of interest more than 11 Mbp by moving the p-ter border of the QTL from D4S414 to D4S400 and based on the March 2006 human genome assembly (<http://genome.ucsc.edu/staff.html>) reduces the number of

candidate annotated genes by 78 and predicted genes by 38.

8.3.2.1 Identifying Known and Predicted Genes in Region of Interest

The UCSC Genome Browser (Kent et al. 2002) tool is used to identify the physical map location of genes and predicted genes within a QTL region of interest. To do so, the microsatellite identifiers for the two markers delimiting the borders of the QTL region of interest are entered into the UCSC Genome Browser query box and the genomic locations retrieved. These two locations define physical map location of the QTL region of interest on the reference genome. When both of these locations are entered into the query box together, the UCSC Genome Browser window will display the entire genomic region of interest.

To list all genes in the defined region, select the “Tables” link in the top bar in Fig. 8.4 to load the Table Browser tool (Fig. 8.5). This new window defaults to selecting a positional table

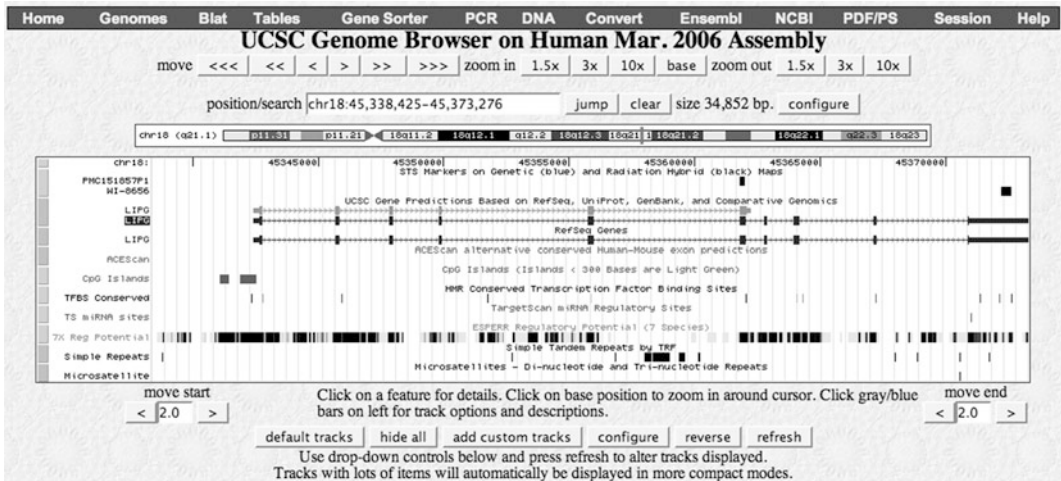


Fig. 8.4 UCSC Genome browser showing annotated gene and predicted gene tracks for the HDL₁-C QTL region of interest. From *top to bottom*, genome browser function links, the genome assembly version, navigational tools, chromosome position numerically, and graphically

on the chromosome diagram. The browser window shows the base number, the track name, and the contents of the track for annotated genes (UCSC based on RefSeq, UniProt, and GenBank) and predicted gene (N scan)

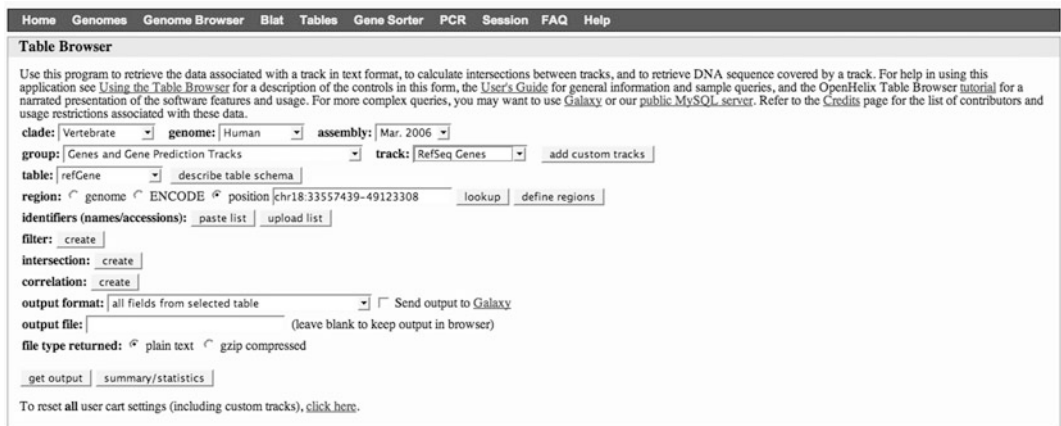


Fig. 8.5 UCSC Table browser function. This tool can be used to provide tabular data on any tracks included in the UCSC genome browser and can be used for defined chromosomal regions or for the entire genome. In this

figure, the table browser has been selected to generate lists of genes and predicted genes for the QTL chromosomal region of interest

for the region of interest viewed in the Genome Browser (Fig. 8.4). The Table Browser tool can be used to provide tabular data on any tracks included in the UCSC genome browser and can be used for defined chromosomal regions as positional data, can be used for non-positional data, or can be used to retrieve data for the entire genome. In this example, the Table Browser is

used to download in tabular form all annotated genes for the region of interest. In addition, using the gene predicted tracks combined with Expressed Sequence Tag (EST) and spliced EST tracks, all high confidence predicted genes can be listed. The gene array tracks, indicating expressed genes detected by whole genome expression profiling, may also be included to gain additional

information regarding expression levels and tissue type expression of genes in the region of interest. 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 and gene sequence. The UCSC Table Browser webpage includes links to descriptions of all table functions and links to tutorials for use of the Table Browser features including merging, filtering, and intersecting data from multiple tracks for output of data in tabular form (<http://genome.ucsc.edu/cgi-bin/hgTables>).

8.4 Prioritizing Candidate Genes Using Expression Profiling

Central to our strategy for the identification of genes encoding QTLs is based on the premise that the gene regulating the QTL must be encoded within the region of the QTL signal and the gene must be expressed in the relevant tissue. Consequently, 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 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 (Cox et al. 2002).

Since developing the custom arrays we have moved to using current sequencing methods to analyze gene expression in QTL intervals. The advantage of RNA Seq methods (Illumina GA IIx platform) is that we are able to identify and quantify all transcripts expressed in a sample. In

addition, quantification of each transcript is not dependent on knowing the precise gene or non-coding RNA sequence beforehand, as is the case when designing primers for array-based methods. For candidate gene prioritization, only those genes in the QTL interval are used from the RNA Seq data. However, with the use of network analysis, it is possible to strengthen candidate gene priority by inclusion or exclusion of the candidate genes in networks constructed from the entire transcriptome.

8.4.1 Discordant Sibs CREA Analysis

To identify baboons for the positional cloning of the gene encoding the HDL₁-C QTL study, we performed phenotypic and genotypic analysis of the pedigreed baboon population and identified baboon sib-pairs discordant for HDL₁-C serum concentrations. The sib-pairs differed by at least one standard deviation for HDL₁-C values. In addition, members of each selected sib-pair did not share IBD (identical-by-descent) alleles, or for some markers shared only one IBD allele, in the chromosomal region of interest. For details of sib-pair HDL₁-C phenotype data see Cox et al. (2002). Because the QTL peak LOD score is greater for the high cholesterol high fat diet than the chow diet, we predicted that the gene influencing HDL₁-C would be differentially expressed between the two diets. Therefore, we collected liver biopsies from baboons before and after a 7-week, high cholesterol, high fat (HCHF) diet challenge. RNA was extracted from the liver biopsies and used to measure expression of all known and predicted genes in the QTL region of interest. In addition, gene expression was compared between the chow and the high cholesterol, high fat diets (Cox et al. 2002).

8.4.2 Designing a CREA

The CREA approach can be achieved by either constructing a custom array or by analyzing RNA Seq data for the chromosomal region of interest. The CREA method is less expensive but

may miss a gene or noncoding RNA due to probe mismatches or may miss novel genes or noncoding RNAs not predicted in the reference genome. Both of these methods are consistent with a conservative approach to positionally cloning the gene encoding a QTL where one evaluates all genes, noncoding RNAs and predicted genes in a QTL region of interest. The analysis relies on an annotated reference genome such as the human genome.

For the CREA method, after defining the physical map locations of the markers delimiting the QTL region of interest, we use the UCSC Table Browser to identify all genes and predicted genes in the QTL region of interest and use the Table Function to download all exon sequences for each of these genes and predicted genes. The exon sequence of each gene is then used to design 65-mer oligonucleotides specific for each gene. The oligonucleotides are then arrayed and used for chromosome region specific expression profiling. 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. First we align the human cDNA sequence with the rat and or mouse cDNA sequence (USCS Genome Browser, NCBI-Search Nucleotide, GeneLynx and Rat Genome Database). We assume that nucleotides conserved between human and rodent will be conserved between human and baboon. We then import both sequences into Sequencer (Gene Codes, Inc.), align the cDNAs, and design oligonucleotides for the gene based on conserved coding regions using Oligo Primer Analysis Software (Molecular Biology Insights, Inc). Oligonucleotide design constraints include: (1) oligonucleotide ≥ 65 nucleotides long; (2) less than 8 mismatches between species; (3) 45–55 % GC content; (4) no tetranucleotide repeats; (5) no significant hairpin loops (less than 7 bonds in a hairpin); and (6) optimal probe with highest T_m and the highest negative ΔG value for GC clamp. 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 nontarget sequences. After gene orientation is confirmed, oligonucleotides are synthesized and nylon-based arrays printed with oligonucleotides spotted in triplicate (Northcott et al. 2012).

Some investigators use a modified CREA approach where they perform whole genome expression profiling using a commercial gene array and analyze genes in the QTL region of interest. For species that do not have a commercial array available, this presents a problem for QTL candidate gene prioritization. If the investigator uses an array from a different species, such as a human gene array for baboon gene expression, there are likely to be sequence differences between human and baboon for some genes resulting in some array probes that do not cross react with baboon sequence. In these instances, the lack of signal for a specific gene may be because the gene is not expressed but it may also be because the gene probe does not cross react. Another limitation of some commercial arrays is that they include only annotated genes and not predicted genes. We know from previous experience that some “predicted genes” in one assembly of the human genome can become annotated genes in later assemblies. Therefore, if a commercial array is used, an investigator should supplement those data with a custom array that includes all predicted genes in the QTL interval and includes all genes that did not give a signal using the commercial array.

The Next Gen sequencing platforms provide a means to sequence and determine abundance of all transcripts (cDNAs) expressed in a tissue. Using the RNA Seq method, genomic DNA in the QTL region of interest is used to map all expressed transcripts. Genome annotations are used to annotate known and predicted genes and noncoding RNAs. In addition, because transcript abundance is measured using this method it is possible to identify differentially expressed transcripts in response to a challenge or that differ among groups with variation in the phenotype of interest. Using this method, we have identified novel baboon transcripts that were not known or predicted in human (Cox et al. unpublished data).

8.4.3 Prioritizing Genes in Region of Interest

Regardless of the method used in Sect. 8.4.2, to quantify expression of genes in the QTL region of interest, 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 (Rapp 2000). The table is then filtered to retain all CREA expressed genes. Mean values for both groups (e.g., low and high HDL₁-C from chow and HCHF diets) are added to the table for each CREA expressed gene. In addition, GeneCards (<http://www.genecards.org/>; Rebhan and Prilusky 1997) and Online Mendelian Inheritance in Man (OMIM) (<http://www.ncbi.nlm.nih.gov/omim>; OMIM 2008) databases are accessed for known function(s) of each annotated gene.

Genes are then ranked first by consistency of each gene expression profiles with the QTL signal. In this example, the QTL signal was observed for the HCHF diet but not the chow diet. Because these baboons were selected based on their contribution to the QTL signal and because these baboons are discordant for HDL₁-C, we predicted that the gene influencing HDL₁-C would be differentially expressed between low and high responders on the HCHF diet, but not the chow diet. Therefore, in this case the highest priority genes were differentially expressed between low and high responders on the HCHF diet and showed either no differences between low and high responders on the chow diet or no differences in expression for the low responders comparing chow and HCHF diets. Genes included in this group were further prioritized based on biological relevance to the genes' known function with the quantitative trait and proximity to the peak LOD score. Predicted genes cannot 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 to 3 genes (Cox et al. 2005).

8.5 Functional Polymorphism Identification

After prioritization of candidate genes, functional polymorphism(s) in the gene, that is, the polymorphisms that influence variation in the quantitative trait must be identified. To date, there are no good prediction tools for the identification of functional polymorphisms. In our baboon HDL₁-C QTL candidate gene study of endothelial lipase (*LIPG*), we evaluated the orthologous human gene for conserved noncoding 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 to the 5' untranslated region and the other region was located -2,446 bp from the transcription start site. No polymorphisms were identified in the conserved region proximal to the 5' untranslated region and none of the polymorphisms located in the upstream conserved region influenced *LIPG* expression of HDL₁-C variation. Furthermore, our study of *LIPG* revealed two functional single nucleotide polymorphisms (SNPs) and one deletion-insertion polymorphism (DIP). SiteSeer (Boardman et al. 2003) 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). Therefore, traditional methods must still be used to identify polymorphisms in each candidate gene, all polymorphisms 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 (Curran et al. 2005). 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.

To limit the number of polymorphisms that must be genotyped, we used 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 identical-by-descent (IBD) 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.

8.5.1 Sequencing Candidate Genes

In our baboon HDL₁-C QTL example, the baboon candidate genes in the QTL region of interest had not yet been sequenced. Therefore, we used gene and genome sequence information from the human reference genome to isolate and sequence the baboon gene. To sequence each candidate gene for which no gene sequence exists, we first isolated Bacterial artificial Chromosome (BAC) clones containing the gene from a baboon BAC library (BACPAC Resources; BACPACorders@chori.org). We used the human DNA sequence to design primers for amplification of a fragment from each candidate gene using Oligo software (Molecular Biology Insights, Inc). The gene fragment was amplified using these primers and the fragment was then used as a probe to isolate a baboon BAC clone containing the gene. The baboon gene of interest was then sequenced from the BAC clone using sequencing primers based on the reference sequence gene data. To download the human gene sequence, we used the Genome Browser

“get DNA” feature (<http://genome.ucsc.edu/cgi-bin/hgc?hgsid=107910572&o=33557438&g=getDna&i=mixed&c=chr18&l=33557438&r=49123308&db=hg18&hgsid=107910572>; Kent et al. 2002). To do so, we first entered the gene name or GenBank ID number into the “position/search” box in the Browser window (Fig. 8.4). The Browser displays a link for that gene. After activating the link, the Browser displays the gene from the transcription start site to the end of the 3'UTR and shows intron–exon structure of the gene. The genomic location was indicated in the “position/search” box. The lower number in the “position/search” box could be changed to 4,000 bp less than the number displayed and the “jump” button used to display the gene including 4,000 bp of promoter (Fig. 8.4).

Clicking on the “DNA” link along the top of the page loaded a new page asking for display preferences; the gene position was auto filled into the “position” box (Fig. 8.6). If there was a preference for DNA display such as lower case for noncoding and upper case for coding sequences, this can be selected using the “extended case/color options” feature. Selecting “get DNA” will prompt the browser to display the DNA sequence for the gene or region of interest (Fig. 8.7). The DNA sequence was copied from the display and pasted into the Oligo software program for design of sequencing primers. In addition, the DNA sequence could be pasted into Sequencher software (GeneCodes, Inc., Ann Arbor, MI) for alignment and distribution analysis of sequencing primers. Each exon for the reference gene was acquired in the same manner and included in the Sequencher alignment as landmarks of coding regions in the candidate gene. For the top priority candidate genes, we sequenced the introns, exons, untranslated regions, and ~4,000 bp of the promoter. We chose to sequence beyond the traditional 1,000 bp of promoter sequence because strong enhancer elements have frequently been found in gene promoters between –4,000 and –1,000 bp from the transcription start site.

If RNA Seq methods are used to sequence and quantify gene expression in a relevant panel of animals, then resequencing will only need to be done for the introns and promoter regions. In

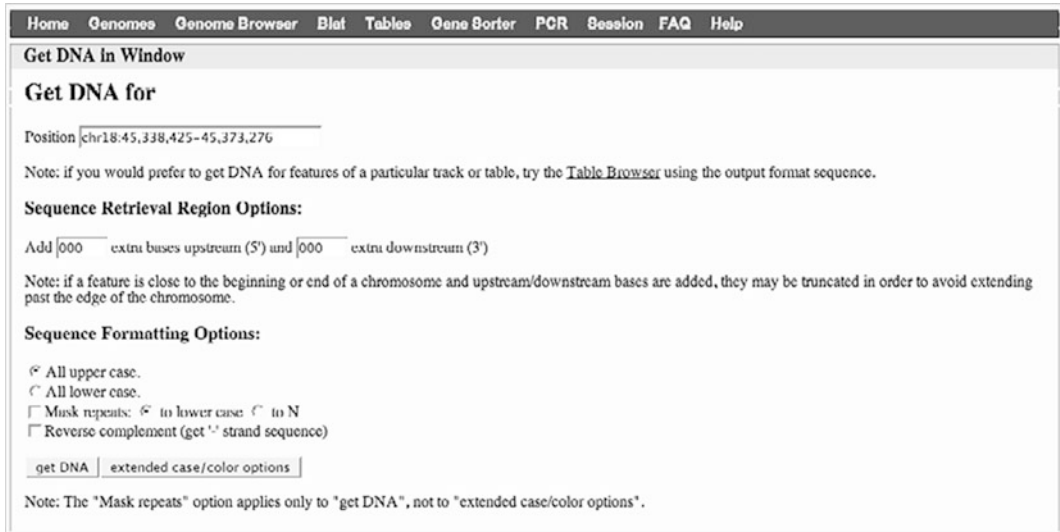


Fig. 8.6 The “Get DNA” feature in the UCSC genome browser can be used to download sequences from any track. This tool allows the user to define the chromosomal

interval from which the sequence will be retrieved and annotation of the retrieved sequence

```
>hg18_dna range=chr18:45338425-45373276 5'pad=0 3'pad=0 strand=+ repeatMasking=none
TATTTATTTGGGTAGAGATGAGGTCTCCTTATGTTGCCCTGGCTGGTCTC
AAATGCCCTAGCCTCAAGCCATCCTTCCACTTTGGCCCTCCCAAAGTGCCAG
GATTACAGGCGTGAGCCACCACACCCAGCCACTTAATTTTAATTTTCATGT
GTTTCTTTTTACCTTTATAATAGGACCACTAGGAAACATAAAATTTATACA
TGTGCCGCTATGGACTGAATTGGGACCCCTCAAATTCCTATGTTGAAGC
CCTAACTCCCTATGATGATTTTGGAGATGGGGCCTGTGGGAGATCATTG
GTTTAGATGAGGTATGAGGTGAGGCACCATGATGGGATTAGAGTTGTTA
TTGGAAGAGACATCAGCGTGTCTTTCTCTCTCTCTCTCTCTCTCTCTCT
CTCTCTCTCTCTCTCTCTCTGCCATGTGAGGACACAGTGAGAAGGCAGC
CATCTATAAGCCAGAAAGAGGGCCCTACCAGAAACCGACTATGCTGGTA
CCCTGATCTTGGACTTCTAGTCTCCAGAATATGAGAAAATAAATTTCTG
TTTTAAAGCCACTAAGTCTATGGTATACTGTTCTGGCAGCCCAAAC TGACC
AAGACATGCGGTTTGTATTATATATTTCTGTTGGACAGCATTTGGTCCAGA
TATCTGGGAACCTCCTACATACCAGCCAGCCTTCTGGCACTTGTAACCTTC
TGTATTGTCTGTGAGAGCACAGGCATGGTCTCCAGGCCAGTGTTCCT
CCCTAGCAGCTGCTCAATAAGCTTCTGGCAATTAAGTCATTTCTTGGTT
GTAAGAATATAAACAGTCTCTGGATAATGTATGTAAAAGAGGACCTCATT
AAGAGAATATTGGGTAACAACCTGGAATCTGCAGGGCAGAGAACCCAGGCT
TGGCCGAAATGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGG
```

Fig. 8.7 Output from genomic DNA sequence retrieval from the “get DNA” feature in the UCSC Genome Browser. The *top line* describes the chosen parameters.

Sequence can be copied and pasted into any sequence analysis program or word processing file

addition, the exon sequences from the RNA Seq can be used to design primers for intron sequencing.

8.5.2 Resequencing Candidate Genes

Sequence data from Sect. 8.5.1 is used to design sequencing primers for resequencing the gene. Sequence polymorphisms are identified by

sequencing the candidate gene 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 will be sequenced from the panel of discordant animals are subcloned and 8 clones for each animal in the panel are sequenced. Briefly, 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. (ABI)). Sequencing products are purified using Exonuclease I (USB) and Shrimp Alkaline Phosphatase (USB) and size fractionated. Sequence data are imported into Sequencher, Gene Codes, Inc. (GCI) 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.

8.6 Cross Species Use of Whole Genome Expression Arrays

We use whole genome expression profiling (from gene arrays or RNA Seq) to provide additional data for QTL candidate genes. Ontological pathway (<http://www.geneontology.org/>) (Ashburner et al. 2000) and KEGG Pathway (www.genome.jp/kegg/) (Kanehisa et al. 2004) analysis of whole genome expression data provide detailed data on individual genes in the context of that gene's role in described biological/biochemical pathways and may reveal insights into molecular mechanisms by which a gene influences a QTL. Cross species use of whole genome expression arrays provides a list of genes that provide quality signal for the RNA samples of interest. These experiments provide extensive information about expression of many genes regardless of the species specificity of the array. One caveat of the cross species use of gene arrays is that the lack of signal for a gene could be due to either low gene expression or lack of cross species hybridization for that gene. From the perspective of simply studying expressed genes, this is not a problem. However, if the investigator wishes to perform pathway analyses for the dataset then the issue of "no-signal" genes becomes an issue. Z-score calculations defining significant gene categories and pathways are

based on the total number of genes on the array that could give a signal (Doniger et al. 2003). Thus, to accurately calculate z-scores, the array of baboon genes for which expression was detected on the human gene chip must be defined. Therefore, for our baboon gene expression studies, we evaluated both human Affymetrix (Affymetrix U133A 2.0) and human Illumina (Illumina Human WG-6 v2) gene arrays for whole genome expression profiling of baboon RNA samples.

To evaluate each human whole genome expression array, we used baboon RNA from 12 baboons for 13 different tissues including liver, kidney, lymphocytes, fat, placenta, 0.5 gestation (G) and 0.9G fetal liver, 0.5G and 0.9G fetal frontal cortex, 0.5G and 0.9G fetal kidney, and 0.5G and 0.9G fetal adrenal. Whole genome expression profiling was performed for each sample and the samples were quality filtered based on 0.5 for Affymetrix (for details see Cox et al. 2006b) and 0.95 for Illumina gene arrays. Because Affymetrix and Illumina use different types of probes and different measures to assess signal quality, the quality filter setting differs for these two platforms. The lists of quality genes from each tissue were merged for each array to generate a list of genes providing a quality signal for baboon RNA on that array platform. Using this method, 16,186 of the 22,227 genes on the Affymetrix GeneChip and 17,231 of 25,538 annotated genes and 4,916 of 20,658 predicted genes on the Illumina BeadChip were detected with quality signal. The merged list for each array platform is the virtual "custom" baboon array for that platform. After determining the genes included in each custom array, the list is uploaded into Genesifter (VizX Labs, Seattle, WA) as the "custom" baboon array and used to perform pathway analyses on the whole genome expression profiling datasets.

8.7 Conclusion

Comparative genomic methods provide a wealth of data for many genetic questions before the first laboratory experiment begins. A basic knowledge

of the central data repositories, available databases, and basic analytical tools will help determine what is known about a system, what can be inferred using data from multiple species, and generate specific hypotheses and questions to address the hypotheses. The UCSC Genome Browser has become a central repository for annotated genome data. In addition, the Genome Browser links out to more detailed information for all included data types. The database is continually updated and new tools are continually developed and added to the Genome Browser. With that said, information in the UCSC Genome Browser depends on data that are provided by other investigators. For example, baboon genome sequence is routinely downloaded to the NCBI trace archive (<http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?>) from the Genome Sequencing Center (Baylor Genome Sequencing Center for baboon) as the data are generated. However, these data will not be downloaded to the UCSC Genome Browser until the baboon genome has been assembled. Consequently, species-specific genome sequence data may be found prior to release to the UCSC Genome Browser. This is the case for bottlenose dolphin, kangaroo rat, and echinoid genome sequences to name just a few. A list of ongoing genome sequencing projects can be found at the NCBI Entrez Genome Project website (<http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi>). In addition, early genome sequence data may be found at different websites (different databases) than cDNA databases with the two datasets generated by laboratories independent from each other. Scientists involved in sequencing a species' genome may not be part of the community of scientists who routinely use that species as a model system. For this reason, often scientists who use a particular model organism may not be aware that the genome sequencing for that organism is underway. The search for sequences specific to your species of interest, even if they are unassembled and unannotated, will add confidence to your comparative genomic analyses and are worth the time spent searching to see if they exist.

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410
- 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 (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25:25–29
- Becker KG, Barnes KC, Bright TJ, Wang SA (2004) The genetic association database. *Nat Genet* 36:431–432
- Boardman PE, Oliver SG, Hubbard SJ (2003) SiteSeer: visualisation and analysis of transcription factor binding sites in nucleotide sequences. *Nucleic Acids Res* 31:3572–3575
- Cheng ML, Kammerer CM, Lowe WF, Dyke B, VandeBerg JL (1988) Method for quantitating cholesterol in subfractions of serum lipoproteins separated by gradient gel electrophoresis. *Biochem Genet* 26:657–681
- Cox LA (2002) The FIC1 gene: structure and polymorphisms in baboon. *J Med Prim* 31:1–12
- Cox LA, Birnbaum S, VandeBerg JL (2002) Identification of candidate genes regulating HDL cholesterol using a chromosomal region expression array. *Genome Res* 12:1693–1702
- Cox L, Birnbaum S, Mahaney M, VandeBerg J (2005) Characterization of candidate genes regulating HDL-C using expression profiling. In: Proceedings of the XIII International Congress on Genes, Gene Families, and Isozymes. Medimond, Bologna Italy, pp. 177–180
- Cox LA, Mahaney MC, VandeBerg JL, Rogers J (2006a) A second-generation genetic linkage map of the baboon (*Papio hamadryas*) genome. *Genomics* 88:274–281
- Cox LA, Nijland MJ, Gilbert JS, Schlabritz-Loutsevitch NE, Hubbard GB, McDonald TJ, Shade RE, Nathanielsz PW (2006b) Effect of 30 per cent maternal nutrient restriction from 0.16 to 0.5 gestation on fetal baboon kidney gene expression. *J Physiol* 572:67–85
- Cox LA, Birnbaum S, Mahaney MC, Rainwater DL, Williams JT, VandeBerg JL (2007) Identification of promoter variants in baboon endothelial lipase that regulate HDL-cholesterol levels. *Circulation* 116:1185–1195
- Cox LA, Glenn J, Ascher S, Birnbaum S, VandeBerg JL (2009) Integration of genetic and genomic methods for identification of genes and gene variants encoding QTLs in the nonhuman primate. *Methods* 49:63–69
- Curran JE, Jowett JB, Elliott KS, Gao Y, Gluschenko K, Wang J, Abel Azim DM, Cai G, Mahaney MC, Comuzzie AG, Dyer TD, Walder KR, Zimmet P, MacCluer JW, Collier GR, Kissebah AH, Blangero J (2005) Genetic variation in selenoprotein S influences inflammatory response. *Nat Genet* 37:1234–1241

- Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, Conklin BR (2003) MAPPFinder: using gene ontology and genMAPP to create a global gene-expression profile from microarray data. *Genome Biol* 4(1):R7
- Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I (2004) VISTA: computational tools for comparative genomics. *Nucleic Acids Res* 32:W273–W279
- Kammerer CM, Cox L, Mahaney MC, Rogers J, Shade R (2001) Sodium lithium counter transport activity is linked to chromosome 5 in baboons. *Hypertension* 37:398–402
- Kammerer CM, Rainwater DL, Schneider JL, Cox LA, Mahaney MC, Rogers J, VandeBerg JL (2003) Two loci affect angiotensin I-converting enzyme activity in baboons. *Hypertension* 41:854–859
- Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res* 32 (Database Issue): D277–D280. DOI: [10.1093/nar/gkh063](https://doi.org/10.1093/nar/gkh063)
- Karolchik D, Hinrichs AS, Furey TS, Roskin K, Sugnet CW, Haussler D, Kent WJ (2004) The UCSC table browser data retrieval tool. *Nucl Acids Res* 32:D493–D496
- Kent WJ (2002) BLAT-The BLAST-like alignment tool. *Genome Res* 12:656–664
- Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. *Genome Res* 12:994–1006
- Mahaney MC, Rainwater DL, Rogers J, Cox LA, Blangero J, Almasy L, VandeBerg JL, Hixson JE (1998) A genome search in pedigreed baboons detects a locus mapping to human chromosome 18q that influences variation in serum levels of HDL and its subfractions. *Circulation* 98:15 (Abstract)
- Northcott CA, Glenn JP, Shade RE, Kammerer CM, Hinojosa-Laborde C, Fink GD, Haywood JR, Cox LA (2012) A custom rat and baboon hypertension gene array to compare experimental models. *Exp Biol Med* 237:99–110
- OMIM (2008) Online Mendelian Inheritance in Man. Johns Hopkins University, Baltimore, MD Retrieved MIM Number: {606945}, <http://www.ncbi.nlm.nih.gov/omim/>. Accessed June 12, 2007
- Rainwater DL, Kammerer CM, Mahaney MC, Rogers J, Cox LA, Schneider JL, VandeBerg JL (2003) Localization of genes that control LDL size fractions in baboons. *Atherosclerosis* 168:15–22
- Rapp JP (2000) Genetic analysis of inherited hypertension in the rat. *Physiol Rev* 80:135–172
- Rebhan M, Prilusky J (1997) Rapid access to biomedical knowledge with GeneCards and HotMolecBase: implications for the electrophoretic analysis of large sets of gene products. *Electrophoresis* 18:2774–2780
- 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 (2000) A genetic linkage map of the baboon (*Papio hamadryas*) genome based on human microsatellite polymorphisms. *Genomics* 67:237–247
- Shah N, Couronne O, Pennacchio LA, Brudno M, Batzoglou S, Bethel EW, Rubin EM, Hamann B, Dubchak I (2004) Phylo-VISTA: interactive visualization of multiple DNA sequence alignments. *Bioinformatics* 20:636–643
- Vinson A, Mahaney MC, Cox LA, Rogers J, VandeBerg JL, Rainwater DL (2007) A pleiotropic QTL on 2p influences serum Lp-PLA(2) activity and LDL cholesterol concentration in a baboon model for the genetics of atherosclerosis risk factors. *Atherosclerosis* 196:667–673
- Voruganti VS, Tejero ME, Proffitt JM, Cole SA, Freeland-Graves JH, Comuzzie AG (2007) Genome-wide scan of plasma cholecystokinin in baboons shows linkage to human chromosome 17. *Obesity* 15:2043–2050