Published in final edited form as:
Annu Rev Nurs Res. 2011 ; 29: 1–26.

Molecular Genomic Research Designs

Kelley Baumgartel[1], **Jamie Zelazny**[1], **Theresa Timcheck**[1], **Chantel Snyder**[1], **Mandy Bell**[1], and **Yvette Conley**[1,2,*]

[1]University of Pittsburgh School of Nursing

[2]University of Pittsburgh Department of Human Genetics

Abstract

Genetic and genomic research approaches have the capability to expand our understanding of the complex pathophysiology of disease susceptibility, susceptibility to complications related to disease, trajectory of recovery from acquired injuries and infections, patient response to interventions and therapeutics, as well as informing diagnoses and prognoses. Nurse scientists are actively involved in all of these fields of inquiry and the goal of this manuscript is to assist with incorporation of genetic and genomic trajectories into their research and facilitate the design and execution of these studies. New studies that are going to embark on recruitment, phenotyping, and sample collection will benefit from forethought about research design to ensure that it addresses the research questions or hypotheses being tested. Studies that will utilize existing data or samples will also benefit from forethought about research design for the same reason but to also address the fact that some designs may not be feasible with the available data or samples. This manuscript discusses candidate gene association, genome wide association, candidate gene expression, global gene expression, and epigenetic/epigenomic study designs. Information provided includes rationale for selecting an appropriate study design, important methodology considerations for each design, key technologies available to accomplish each type of study, and online resources available to assist in executing each type of study design.

In the last decade we have progressed from a rough draft of the human genome sequence to availability of an abundance of publicly available databases and high throughput data collection technologies to facilitate genetic and genomic study design. Genetic (focus on one gene at a time) and genomic (focus on entire genome as well as gene-gene interactions) research continues to hold great promise for understanding a wealth of human conditions, providing objective data for diagnosis and prognosis, informing therapeutics, and providing the cornerstone for evidence based practice for genomic health care (Green, Guyer, & National Human Genome Research Institute [NHGRI], 2011; Lander, 2011). The research programs of many nurse scientists are ripe for incorporating a genetic/genomic research component or movement of existing genetic or genomic research in a new direction.

The goal of his paper is to bring together key information about designing studies with a molecular genetic or genomic focus coupled with dynamic resources offered to the reader to

*Corresponding Author: Yvette P. Conley, University of Pittsburgh, 3500 Victoria Street, 440 Victoria Building, Pittsburgh, PA 15261, yconley@pitt.edu, 412-383-7641.

Contributor's list:
Kelley Baumgartel, RN, Doctoral Student, University of Pittsburgh, Pittsburgh, PA
Jamie Zelazny, RN, MPH, Doctoral Student, University of Pittsburgh, Pittsburgh, PA
Theresa Timcheck, RN, Doctoral Student, University of Pittsburgh, Pittsburgh, PA
Chantel Snyder, RN, Doctoral Student, University of Pittsburgh, Pittsburgh, PA
Mandy Bell, RN, Doctoral Student, University of Pittsburgh, Pittsburgh, PA
Yvette P. Conley, PhD, Associate Professor of Nursing and Human Genetics, University of Pittsburgh, Pittsburgh, PA

expand their understanding and ensure access to state of the science information. It is not meant to be an exhaustive resource, but one that sets the stage for contemplation of embarking on such research designs and key issues to ponder during study design phase. This paper is written for the researcher who has a basic understanding of genetics and is contemplating adding a genetic or genomic component to their research or designing the next step in their genetic or genomic program of research. Readers are encouraged to visit an extremely useful resource, the National Human Genome Research Institute's talking glossary at http://www.genome.gov/glossary, for clarification of unfamiliar terms and expansion of knowledge about genetic terminology. Technology to collect genetic and genomic data changes rapidly, therefore proper study design, and selection of appropriate methodology to accomplish a study also change rapidly. This paper incorporates a large number of online resources that are continuously updated in an attempt to keep the paper as up to data as possible. Readers are encouraged to visit these online resources when designing their study to ensure that their study design is state of the science.

DNA POLYMORPHISM BASED ASSOCIATION STUDIES

The overall objective of a polymorphism based association study is to examine the relationship between DNA variation and a phenotype (e.g., diabetes, fatigue). A polymorphism is defined as a DNA variation that is present in at least one percent of the population (NHGRI, n.d.). One advantage of this approach compared to other genetic/genomic approaches is the use of DNA. DNA is a very stable template for experiments, allowing for use of previously collected samples. Such a retrospective approach could save time and money that would be needed to prospectively recruit participants and collect samples; however, attention must be given to subject consent to assure that informed consent was obtained for future genetic/genomic evaluation related to the phenotype of interest. Another advantage is that this approach does not require that subjects be related, which is a requirement for linkage analysis, an approach not discussed in this manuscript. It should be noted that while related individuals are not required, newer software has been developed to allow for the analyses of related individuals within the context of an association study. Two very appealing additional advantages of polymorphism based studies are the fact that polymorphisms do not change over time and the DNA template that is utilized can be extracted from any tissue. The sample for DNA extraction and collection of polymorphism data only need to be collected once, yet that polymorphism data can be evaluated within the context of a phenotype that changes over time. While blood and saliva are the most frequently used cell/tissue type for DNA extraction, any cells/tissues that have a nucleus can serve as samples for polymorphism based studies. Because DNA polymorphisms do not change and are not tissue specific, investigators need not worry about collection of DNA samples over time or from what tissue DNA extraction occurs. These advantages are not carried over to other genomic approaches detailed in this manuscript.

Candidate Gene Association Studies

Rationale for taking a candidate gene association approach—Candidate gene association studies investigate polymorphisms representing a specific gene(s) to determine if it is associated with a phenotype of interest. With this hypothesis-driven approach, the investigator pre-selects the candidate gene(s) to be evaluated. This approach is only appropriate when *a priori* assumptions about the gene(s) that may be involved in the phenotype of interest can be justified.

Genome wide association studies (GWAS); discussed in the next section, have large sample size requirements (e.g., 1000 cases/1000 controls), and one relative advantage of the candidate gene approach is that it often requires half that number or less. This reduced sample size requirement compared to a GWAS is due to the focused evaluation of a

candidate gene(s), which reduces multiple testing concerns. The candidate gene association approach is also ideal when studying rarer phenotypes since attainment of a large sample may not be feasible for a condition with a low population frequency.

Subject and sample considerations—Clearly defined inclusion/exclusion criteria, which include a detailed definition of the phenotype, are essential to the candidate gene association approach. Structured inclusion/exclusion criteria help to ensure that individuals with/without the phenotype of interest are similar in all aspects except for the condition being investigated. Moreover, phenotypic assessment of controls should be as comprehensive as the phenotypic assessment of cases. Ultimately, carefully crafted criteria, and thorough phenotypic assessments help reduce the impact of confounding variables.

Population stratification represents another potential source of confounding in candidate gene association studies utilizing a case-control design. The case-control design compares allele, genotype, or haplotype frequencies between the groups. Because these frequencies can be extremely disparate for different ancestries, it is important to control for ancestry to avoid spurious results/conclusions (e.g., concluding that there is an association between a phenotype/allele when in reality the association is fueled by ancestral differences in allelic frequencies). The risk for population stratification can be mitigated. Subgroup analysis represents one option, but it relies on self report to categorically measure race/ethnicity. An option that controls for population stratification statistically is the use of ancestral informative markers (AIM), which are polymorphisms in the DNA that allow one to calculate an admixture proportion for an individual. The application of these proportions are used for analysis rather than the traditionally used, though unreliable, method of self-reported race/ethnicity. In a recent study, only 30 AIMs were needed to estimate European admixture in a group of African American women (Ruiz-Narváez, Rosenberg, Wise, Reich, & Palmer, 2011). Although different AIMs may be needed to estimate other admixture proportions, this example demonstrates that population stratification can be successfully controlled through the analysis of genetic markers.

Another aspect of the candidate gene association study that should be considered is sample size requirements. Quanto (http://hydra.usc.edu/gxe/) is a freely downloadable computer program that can assist with sample size and/or power calculations for candidate gene association studies. User defined criteria can be manipulated according to the polymorphisms that have been selected for evaluation and according to study design specifications.

Candidate gene selection—Candidate gene selection is often based on biologic plausibility. This plausibility can be based on biological pathways implicated in the condition, biomarker data implicating a gene/gene product in the phenotype of interest, pharmacologic treatments for the condition that may indicate a target gene(s), or data from animal models (Hattersley & McCarthy, 2005). Bio-informatics databases, such as the Gene Ontology (http://www.geneontology.org/), may also aid in the identification of genes whose products may impact the phenotype of interest (The Gene Ontology, 1999–2011). Moreover, consideration should be given to number of genes on which to focus, ranging from a single gene to genes within a candidate biological pathway. Because more biologically global conclusions can be drawn, the study of a biologic pathway has the advantage of being more informative than the singular gene approach in most situations (Jorgenson, Ruczinski, Kessing, Smith, Shugart, & Alberg, 2009).

Polymorphism selection—Once selection of the candidate gene(s) is finalized, polymorphisms must be selected to evaluate candidate gene variability, and these are the genetic data used for analyses. The candidate gene association approach includes the

evaluation of single nucleotide polymorphisms (SNPs), repeat polymorphisms, insertion/deletion polymorphisms (INDEL), and copy number variants (CNV).

Resources for polymorphism selection: The SNP is the most common type of polymorphism and is a nucleotide (also known as a base) in the DNA where the nucleotide present (e.g., A, T, C, G) varies in the population (Genetics Home Reference, 2011). The scientific literature and a variety of online databases provide excellent resources for SNP identification and selection. A simple literature search combining the candidate gene(s) with the keyword "functional polymorphism" will help to identify SNPs known to alter the function of the candidate gene(s). Because functional polymorphisms modify the function of a gene regardless of phenotype, the literature search should not be limited to just the phenotype of interest. In addition to the literature, investigators also commonly use the Database of Single Nucleotide Polymorphisms (dbSNP) (http://www.ncbi.nlm.nih.gov/projects/SNP/) and the International HapMap Project (http://hapmap.ncbi.nlm.nih.gov/) to identify/select SNPs and tagging SNPs, respectively.

HapMap is accessed for the selection of tagging SNPs (tSNP), which represent the current gold standard for the evaluation of genetic variation in the candidate gene association study. The goal of HapMap is to develop a haplotype map of the human genome and to describe common patterns of genetic variation in humans (International HapMap Project, 2006). Essentially, HapMap is based on the premise that DNA is inherited in chunks/blocks (haploblock). Within these haploblocks, certain variants are inherited together. If the genotype of one variant within that block of DNA is known the genotype of a second variant within the same block can be determined since they are inherited together. Thus, HapMap assists the user in selecting SNPs that tag a certain haploblock of DNA (tagging SNPs or tSNPs). Ultimately, utilization of tSNPs allows one to fully evaluate the genetic variability of the candidate genes with the least number of SNPs (International HapMap Project, n.d.).

Repeat polymorphisms are characterized by repeating units of DNA bases. The number of times these DNA units repeat is variable in the population (Passarge, 2007). While repeat polymorphisms are less frequent in the genome than SNPs, they are often more informative as they usually have more alleles in the population than SNPs, which typically only have 2. The short tandem repeat (STR) is typically comprised of a repeating unit of two to four DNA bases (e.g., CAG CAG CAG) while the variable number tandem repeat (VNTR) is comprised of a larger repeating unit (Passarge), usually greater than 5 bases. For the evaluation of STRs and VNTRs, the literature continues to be the best source for identification and characterization.

An INDEL polymorphism occurs when a base(s) is added or subtracted from a place in the DNA. It is the presence or absence of the INDEL that is variable in the population (Nussbaum, McInnes, & Willard, 2007). Like SNPs, the dbSNP can be freely accessed to identify small-scale INDELs.

The CNV occurs when the number of copies of a particular genomic sequence/segment is variable in the population (NHGRI, n.d.). CNVs can be identified through scientific literature and online databases. The Database of Genomic Structural Variation (dbVar) (http://www.ncbi.nlm.nih.gov/dbvar) and The Copy Number Variation Project by the Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/humgen/cnv/) are two online resources that may assist in CNV identification.

Genotype data collection technologies—Multiple options are available for SNP genotyping, including the polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP) technique, real-time PCR allelic discrimination (e.g.

TaqMan®), multiplexing via mass spectrometry, and bead chip technology. Selection of the genotyping technique is guided by the number of samples and polymorphisms to be genotyped and available resources. PCR-RFLP, which is used to genotype SNPs based on differences in fragment lengths, is suitable when the number of SNPs and samples to be genotyped is relatively small. Real-time PCR allelic discrimination (http://www.appliedbiosystems.com; http://www.roche-applied-science.com), which genotypes SNPs based on allele-specific fluorescence intensity signals, is suitable for a medium number of SNPs and sample size. Because PCR-RFLP and real-time PCR allelic discrimination can only genotype one SNP at a time, the use of high throughput technologies have become the gold standard for SNP genotype collection when the number of SNPs to be evaluated approaches 24. The iPLEX® Gold-SNP Genotyping assay (http://www.sequenom.com), which genotypes SNPs based on differences in molecular mass, allows for the analysis of up to 36 SNPs per assay (Sequenom, 2010) in larger sample sizes. Not only can an investigator analyze multiple SNPs simultaneously, but time, assay to assay variability, and costs are reduced. The GoldenGate Genotyping Assay (http://www.illumina.com) is another high throughput bead based technology that can be utilized when the number of SNPs and samples to be analyzed is too large for other technologies.

There are several genotyping technologies also available for repeat polymorphisms, INDELs, and CNVs. PCR amplification followed by fragment sizing can be used for genotyping repeat polymorphisms. As with SNPs, real-time PCR allelic discrimination can be used to genotype small INDELs. Finally, TaqMan® Copy Number Assays (http://appliedbiosystems.com) or cytogenetic techniques (e.g., Fluorescence In Situ Hybridization) can be utilized for genotyping candidate CNVs.

Genome Wide Association Studies (GWAS)

Rationale for taking a GWAS approach—A GWAS gentoypes thousands to millions of polymorphisms across the genome for individuals who are phenotypically well-characterized (DiStefano & Taverna, 2011). If genetic variability is significantly different between cases and controls, those variations may be associated with susceptibility to or protection from the phenotype of interest and can provide direction as to which region of the genome these differences might be located. Ongoing efforts of the Human Genome Project and the International HapMap Project have made this approach possible through the generation of large databases that reference and map both sequence and variability.

The major advantage of a GWAS approach is that the biology of the phenotype of interest does not need to be completely understood prior to implementing this approach and the SNPs or genes of interest do not need to be defined *a priori*. Instead of selecting genes and polymorphisms *a priori*, polymorphisms that cover haploblocks across the entire genome are used for genotype data collection and non-parametric based analyses determine what genes/regions of the genome are relevant to the phenotype of interest (Hakonarson & Grant, 2011). The data derived from GWAS will provide direction regarding which areas of the genome warrant additional study.

There are several limitations to GWAS. The variant identified may not be what's accounting for the association, but is rather "tagging along" with the actual causal variant(s). This obstacle is also present for candidate gene association studies, particularly those that utilize a tSNP approach. Therefore, it may be necessary to follow up with more focused genotype data collection, including denser polymorphism evaluations and/or sequencing of that specific region of the genome to identify the exact allele accounting for the association (NHGRI, 2010). A major limitation for the GWAS approach, and perhaps a reason why many investigators are unable to pursue this approach, is the need for thousands of subjects

who are phenotypically well characterized and for which DNA is available. The need for large samples sizes for GWAS is due to the inherent issue of multiple testing that accompanies the evaluation of thousands to millions of different genetic variables. Additionally, the need for very large sample sizes, coupled with the cost of commercial genome-wide scanning techniques makes this approach very costly. GWAS approaches are also not optimal to assess rare polymorphisms as the data collection approaches for the GWAS are more focused on optimizing informativeness of the data (Ku, Loy, Pawitan, & Chia, 2010).

Subject and sample considerations—The cross-sectional case-control study design is the most frequently used approach for a GWAS. Study subjects should be selected based on a well-defined and heritable phenotype. Cases are defined as individuals who meet criteria for a phenotype of interest. Controls are individuals who have never met criteria for the phenotype and ideally have passed through the age or period of risk for the phenotype (Hakonarson & Grant, 2011). Like candidate gene associations studies, ancestry must be considered to avoid issues related to population substructure and this is why some investigators have conducted these types of studies with homogeneous populations (Psychiatric GWAS Consortium Coordinating Committee, 2009). Case and control groups should be matched on ancestry as much as possible to avoid false-positives. Despite this consideration, an advantage of GWAS is that whole genome data can provide adequate data to identify stratification and inflation of test statistics due to population substructure can be addressed (Hakonarson & Grant).

Obtaining a sufficiently large sample size is essential to ensure sufficient statistical power for a GWAS approach. Approximately 1,000 cases and a similar number of controls are required to detect 1–5 variants associated with a given trait. A larger sample is needed to uncover additional variants that may have diminishing contributions to the disease (Hakonarson & Grant, 2011).

Informed consent issues: While informed consent is of paramount importance with any research study, researchers who are considering a GWAS should be cognizant of issues related to conducting such as study and the National Institutes of Health (NIH) policy on data sharing for GWAS. In January 2008, the NIH adjusted its policy mandating the sharing of GWAS data obtained in NIH-funded or conducted studies. The details of this policy can be found at http://gwas.nih.gov/. Most NIH-funded GWAS are required to include language in the consent document that addresses public sharing of de-identified genotype and phenotype data. Researchers who are planning to study existing samples must ensure that the original consent signed by the subjects is consistent with conducting a GWAS.

Genotype data collection technologies—There are currently two commonly used vendors that provide technology for collection of GWAS data, Affymetrix and Illumina. The companies use different technological approaches, which are both widely used in the research community. The Affymetrix^{R} Genome Wide SNP Array 6.0 features 1.8 million genetic markers, including 906,600 SNPs and more than 946,000 probes for the detection of CNVs. This platform also includes a high resolution reference map and a copy number polymorphism (CNP) algorithm (see http://www.affymetrix.com for additional information). The Illumina Omni Microarrays provide a multiple bead chip option which will soon include nearly 5 million markers per sample, including both common and rare variants identified by the 1000 Genomes project. Omni microarrays assess structural variation, including CNVs and copy neutral variants (inversions and translocations) which may also be significant contributors to disease (see http://www.illumina.com for additional information).

Resources of interest for GWAS: The Center for Inherited Disease Research (CIDR) at Johns Hopkins University (http://www.cidr.jhmi.edu/requirements/applications.html) is funded by NIH Institutes and provides genotyping and statistical genetic services to investigators who have received access after a competitive peer review process. Interested investigators are required to submit an application for projects supported by the NIH. In order to maximize access to resources, the application process to CIDR should ideally take place before or at the time of grant application, though this is not a requirement.

The repository for GWAS data is currently the Database of Genotypes and Phenotypes (dbGaP; http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gap). This database was developed to archive the results of studies that have investigated the genotype-phenotype interaction and serves as a useful resource in reviewing the work that has already been completed and aids in planning future research. The dbGaP database provides the opportunity for *in silico* research. Researchers have the option of two levels of access (open and closed) to dbGaP: Open-access data are aggregate data that are publicly available while closed level access requires an application and approval process that includes de-identified subject specific data. The genotype data and their linked phenotype data are invaluable resources and researchers are encouraged to investigate this database as it pertains to their phenotypes of interest prior to designing a study.

GENE EXPRESSION STUDIES

Gene expression studies evaluate the activity of a gene using the level of messenger RNA (mRNA) from a gene(s) and determine if that level is associated with the phenotype of interest. DNA contains a code to generate mRNA through a process called transcription. The amount of mRNA produced from a gene, if at all, depends on many factors including tissue type, local cell environment, and point in the cell cycle.

A gene expression study is different from a polymorphism based study because an expression study evaluates mRNA levels that can change over time, uses less stable mRNA instead of DNA, and mRNA levels can be dramatically different based on what tissue is used for analysis, since gene expression is tissue-specific. Gene expression studies therefore should address whether multiple samples over time are needed for evaluation (similar to other types of biomarkers that change over time), RNA stabilization, and what cell/tissue type is most appropriate to evaluate for the phenotype of interest. For these reasons, many stored samples may not be appropriate for this approach.

Candidate Gene Expression Studies

Rationale for taking a candidate gene expression approach—Candidate gene expression studies investigate mRNA levels for a specific gene(s) to determine if it is associated with a phenotype of interest. Similar to a candidate gene association approach, this is a hypothesis-driven approach where the investigator *a priori* selects the candidate gene(s) to be evaluated. This approach is only appropriate if the investigator has ample justification for investigating a specific gene(s).

Subject and sample considerations—Gene expression studies often involve relative comparisons of mRNA levels between two groups (these groups can be different types of tissues, groups that vary for a particular exposure or groups that vary by the presence or absence of a phenotype of interest). Clearly defined inclusion/exclusion criteria are necessary, due to the relative comparison nature of this approach.

RNA stabilization: Stabilization of RNA is essential to obtain accurate gene expression profiles of biological samples. Immediately after sample collection, RNA degradation and

other transcriptional changes begin to occur. These alterations may result in false up or down regulation of gene expression levels. RNA stabilization preserves a representative gene expression profile for later analysis (e.g. quantitative RT-PCR and microarray analysis). RNA stabilization methods vary based on the type of biological sample. Five of the most common RNA stabilization methods are: (a) PAXgene Blood RNA System (http://www.preanalytix.com), which utilizes a single tube (pre-filled with RNA stabilization reagent) for blood collection, RNA stabilization, sample transport and storage, and purification of total RNA (PreAnalytiX, 2010); (b) LeukoLOCK System (http:www.lifetechnologies.com), which filters and isolates leukocytes from whole blood. RNAlater solution is then used to stabilize the RNA of the leukocytes. A notable advantage to the LeukoLOCK System is the ability to remove a large proportion of reticulocyte-derived globin mRNA. Depletion of the globin mRNA allows for the detection of thousands of additional genes on microarray (Life Technologies Corporation LeukoLOCK, 2010); (c) RNAlater (http://www.ambion.com; http:www.qiagen.com) stabilizes RNA in a variety of fresh samples including animal tissue, tissue culture cells, leukocytes, yeast, and bacteria. After collection, the sample is submerged in the RNAlater stabilization solution. This solution permeates and stabilizes the sample eliminating the need for immediate processing or freezing of samples (Life Technologies Corporation RNAlater, 2010; Qiagen, 2006); (d) Oragene RNA for Expression Analysis Self Collection Kit (http://www.dnagenotek.com): allows for the non-invasive collection of RNA from saliva. Donors are instructed to expectorate into a vial, cap the container, and shake vigorously to release a stabilization solution from the cap. Oragene RNA samples can remain stable for months at room temperature (DNA Genotek, 2011); and (e) Snap Freeze quick freezes solid tissues with liquid nitrogen and dry ice can be used to preserve RNA; however, disruptions during freezing and thawing can lead to RNA degradation. Due to potential RNA degradation and difficulty of obtaining and working with liquid nitrogen and dry ice, RNAlater described above may be a more viable option for solid tissue RNA stabilization.

Candidate gene selection—Candidate gene selection must be justified and rationale for selection is similar to selection of candidate genes in the candidate gene association section. The same bio-informatics databases mentioned in that section are also applicable to aiding in the selection of candidate genes for an expression study and as with a candidate gene association study, a candidate gene expression study should consider focusing on a group of genes in a biological pathway versus the value of focusing on a single gene.

Expression data collection technologies—Selection of a data collection technology for a candidate gene expression study should take into account the number of genes/loci and the number of samples to be evaluated. The most frequently used technologies for a candidate gene approach include Northern blotting, quantitative real-time PCR (qRT-PCR), and multiplex platforms that support 3–36 genes/loci per reaction.

Northern blotting requires electrophoresis of RNA, transfer to a membrane and hybridizing the membrane with a probe specific for detection of the mRNA of interest. The advantages of blotting are that most laboratories will have the equipment to conduct this type of data collection and assessment of RNA size is possible. Disadvantages of blotting are that RNA degradation is common, it requires more RNA as a template for the experiment compared to other methods, it is laborious, and is not optimal for quantification of mRNA levels.

Currently, one of the most popular techniques for assessing the level of mRNA for a gene/locus is qRT-PCR. qRT-PCR requires conversion of RNA to a more stable template called cDNA (complementary DNA), PCR amplification and probe hybridization for the gene/locus of interest. The probe is fluorescently labeled and liberation of this fluorescent label is quantified, reflecting the amount of starting mRNA template in the sample. One crucial step

in conducting qRT-PCR is normalization of the data generated. Normalization of the data allows for sample to sample comparisons that have been corrected for noise such as what's introduced when sample dispensing between samples isn't uniform. This is often done using qRT-PCR data collected simultaneously for an endogenous control, which usually represents a stably expressed gene (often referred to as a "housekeeping gene") and allows for normalization of data across samples (Guenin et al., 2009). Thought needs to be given to selection of an appropriate endogenous control given that different tissues will have different stably expressed genes (Guenin et al.). If in doubt, endogenous control panels are available for assessment prior to conducting qRT-PCR. Advantages of qRT-PCR include high sensitivity and reduced RNA template requirements, high throughput capabilities, quantification of starting mRNA template is possible with use of proper exogenous reference controls, and for many genes/loci/pathways off the shelf optimized assays are available (http://www.appliedbiosystems.com; http;//www.roche-applied-science.com).

Multiplex gene expression assays are available when the number of genes/loci to be evaluated is in the range of ~3–36. One example is the QuantiGene® Plex 2.0 assay (for more information see http://www.panomics.com/index.php?id=product_6) that uses Luminex technology to collect the data and the assay can be customized.

Global (Genome Wide) Gene Expression Studies

Rationale for taking a global gene expression approach—Whole genome expression (also known as global gene expression or gene expression profiling) offers a comprehensive view of gene activity within a biological sample by examining mRNA levels for all known genes across the genome. In this way, whole genome expression provides functional information regarding "when and where a protein is expressed, when it is degraded, and with which other proteins it may interact" (Altman & Raychaudhuri, 2001, p. 340). Due to the dynamic nature of expression, gene expression profiles are often relatively compared under multiple conditions (such as comparing different tissue types, comparing normal versus abnormal tissues, comparing tissues before and after an exposure) or over a period of time (Altman & Raychaudhuri, 2001; Arcellana-Panlilio & Robbins, 2002). The use of global gene expression profiling is extremely advantageous when little to nothing is known about the genes influencing a condition, a similar advantage held by the GWAS approach. Thus, whole genome expression can identify novel candidate hypotheses through a non-parametric analysis of genome wide expression data.

Subject and sample considerations

Sample selection: Although this is an approach similar to GWAS, with evaluation of thousands of genes in a nonparametric manner, sample size requirements for global gene expression are usually smaller, requiring approximately 10 subjects per variable. Matching of subjects for key variables known to influence the phenotype under investigation can reduce the number of variables that need to be accounted for in the analyses. A sample size calculator for global gene expression experiments can be found at http://bioinformatics.mdanderson.org/MicroarraySampleSize/. Additionally, as with candidate gene expression studies, mRNA stabilization of the collected samples is crucial.

Gene expression data collection technologies

Microarrays: Microarrays are used to examine the expression profile of a single sample (often referred to as single dye array) or to compare expression levels between two different samples/conditions (often referred to as two dye array). The microarray itself is a solid surface covered with an "ordered arrangement of unique nucleic acid fragments derived from individual genes" (Arcellana-Panilio & Robbins, 2002, p. G397). Fluorescently labeled template hybridizes to these nucleic acid fragments (referred to as probes) on the solid

surface through complementary pairing. The intensity of the florescence at each spot on the microarray corresponds to the amount of sample binding to a particular nucleic acid fragment and thus, the gene expression level. If the microarray reveals any interesting findings, q-RT-PCR should be carried out for validation purposes. For a visual representation of microarray methodology visit this web address: http://www.bio.davidson.edu/courses/genomics/chip/chip.html.

Microarrays have revolutionized gene expression analyses, as this technology is able to simultaneously survey thousands of genes in a short period of time. However, the ability to detect novel genes is limited to the hybridization probes represented on the microarray. Off-the-shelf probe sets that contain reference sequences can be used, or custom probe sets are designed based on specific genes of interest or pathways. Additionally, microarrays require specialized lab equipment and are very useful when analyzing a small sample size but become costly as sample size increases. Two popular microarrays platforms include Affymetrix's GeneChip and Illumina's BeadChip.

Affymetrix's GeneChip platform (for more information see http://www.affymetrix.com) utilizes traditional solid support microarray technology. Affymetrix's latest product, the GeneChip Human Gene 1.0 ST Array, is able to interrogate 28,869 genes and covers over 700,000 distinct probes. A greater number of samples can be processed simultaneously (with this same probe set) using the Human Gene 1.1 Array Strip (4 samples/strip) and the Human Gene 1.1 Array Plate (16, 24, or 96 samples/plate). Affymetrix also provides whole transcript expression analysis technology for mice and rats.

Instead of using a solid support platform, the Illumina BeadChip platform (for more information see http://www.illumina.com) employs silica beads (each covered with thousands of copies of a specific oligonucleotide) self-assembled in microwells of fiber optic bundles or planar silica slides. Illumina's most recent whole genome expression array, the HumanHT-12 v4 BeadChip, provides high throughput processing of twelve samples and covers over 47,000 probes. Illumina also offers whole genome expression BeadChip technology for mice and rats.

Normalization of gene expression data is also important with microarray data collection. Unlike qRT-PCR where an appropriate endogenous control needs to be selected and included in the data collection, microarrays already include a range of endogenous controls for which data is simultaneously collected and from which the investigator can select to use for normalization of the data.

Sequence based technologies that utilize next-generation sequencing (NGS; high throughput sequencing) are also available for collection of genome-wide gene expression data. An example of such a technology is the RNA-Seq method (for more information see http://www.illumina.com). This method requires conversion to cDNA, ligation of the cDNA fragments, creation of a library, sequencing of the template, and collection of frequency data for a transcript. An advantage of this approach over microarrays is that it does not require primers or probes therefore novel transcripts that would not be detectable with a microarray can be identified.

Serial analysis of gene expression: Serial analysis of gene expression (SAGE) provides comprehensive quantitative gene expression data. SAGE technology is based on three main principles: (1) a short sequence tag (9–17 bases) contains sufficient information to distinctively identify a transcript, (2) sequence tags can be linked together to form one long molecule that can be cloned and sequenced to allow efficient analysis of transcripts, and (3) the number of times a particular tag is observed corresponds to the expression level of the

transcript (Sagenet, 2005; Velculescu, Zhang, Vogelstein, & Kinzler, 1995). One of the main advantages of SAGE, similar to RNA-Seq, is the ability to detect novel genes as it does not require prior sequence information or hybridization probes for each transcript like microarrays (Velculescu et. al, 1997). Another advantage of SAGE is that is utilizes common laboratory equipment and techniques. Any laboratory that performs PCR and manual sequencing could also execute SAGE. Nonetheless, due to cloning and sequencing, SAGE can be expensive, time consuming, and labor intensive.

EPIGENETIC STUDIES

An epigenetic mechanism is a biochemical alteration to the DNA molecule that does not change the sequence of the DNA but does influence gene expression. Epigenetics is often defined as the "study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence" (Russo, Martienssen, & Riggs, 1996, p. 1).

The epigenetic/epigenomic approach shares many advantages and disadvantages with DNA polymorphism based approaches and gene expression based approaches. Like DNA polymorphism based approaches, the epigenetic/epigenomic approach uses DNA as its template for data collection. Since both DNA sequence and its chemical modifications are stable, stored samples are more likely to be appropriate for this approach than gene expression approaches. Similar to a gene expression based approach; epigenetic/epigenomic alterations can change over time and can differ dramatically between cell/tissue types. Although template stability is not an issue, the investigator should give great consideration to whether multiple samples over time are needed for evaluation and what cell/tissue type is most appropriate to evaluate for the given phenotype of interest. For these reasons, similar to gene expression studies, many stored samples may not be appropriate for this approach.

Chromatin remodeling, non-coding RNAs, histone modifications, and DNA methylation are all epigenetic/epigenomic alterations that impact gene expression. Chromatin remodeling is an enzymatic process that results in altered chromatin and nucleosome composition. This transformed structure provides regulatory proteins access to the DNA molecule. Non-coding RNAs are not translated into protein but have considerable involvement in gene expression through interactions with DNA/mRNA. While chromatin remodeling and non-coding RNAs are important to gene regulation, this paper will focus primarily on the commonly examined epigenetic mechanisms for which the most technology for data collection is available: histone modifications and methylation.

Rationale for Taking an Epigenetic/Epigenomic Approach

The decision to take an epigenetic (candidate gene) or an epigenomic (genome wide) approach is based upon wanting to evaluate the mechanism for gene regulation. There are many environmental factors that impact the severity and frequency of epigenetic/epigenomic alterations and subsequent gene expression; therefore, this approach is often used to examine multifactorial diseases that have an environmental component associated with it. Epigenetic approaches to examine transcriptional regulation have contributed to a more comprehensive understanding of complex conditions that demonstrate aberrant gene expression, including: cancer (Wilop et al., 2011), mental health (Read, Bentall, & Fosse, 2009), and cardiovascular disorders (Ordovás & Smith, 2010). Furthermore, the investigation of diseases for which DNA mutations have not been revealed may benefit from an epigenetic approach.

Subject and Sample Considerations

The epigenome is subject to frequent alterations; therefore, longitudinal sample collection is recommended if evaluating time-sensitive trends. Subject size recommendations for an epigenetic study follow similar guidelines to a gene expression study, and vary on whether the investigator will examine the entire genome (hypothesis generating/larger sample size) or a candidate gene profile (hypothesis driven/smaller sample size). Like the other approaches described, an epigenetic study does not require that subjects be related. The advantages and disadvantages of conducting a genome-wide versus candidate gene epigenetic study are similar to those described in previous sections.

The epigenome is largely determined by cell type, and this is especially true for methylation patterns; therefore, tissue source is extremely important to consider for this type of approach. For example, the methylation profile of a skin cell is very different than the methylation profile of a liver cell, since different genes are expressed in each cell type, and methylation is a driving force behind tissue specific gene expression. Similar to a gene expression study, an epigenetic design requires the samples for epigenetic analyses be from a tissue that appropriately addresses the phenotype of interest. Tissue specific sample collection will capture epigenetic patterns that impact gene expression which are potentially contributing to the disease. Unlike a gene expression study which examines RNA, this design requires DNA, which is advantageous for the investigator who has access to previously collected samples, assuming they were collected from an appropriate tissue for the phenotype under investigation.

Epigenetic and Epigenomic Data Collection Technologies

This section will focus on the two epigenetic mechanisms most frequently studied: (a) histone modification and (b) methylation. Post-translational histone modifications include alteration of the histone tail through biochemical changes that ultimately impact gene activity. Genome-wide histone modifications can be captured with chromatin immunoprecipitation technology (ChIP), and quantified with a microarray (ChIP-chip). Methylation refers to the addition of a methyl group to a cytosine, often at CpG islands, which are regions of the genome that are rich in CG base sequences. Hypermethylation of a gene typically leads to gene suppression, while hypomethylation results in gene expression. Genome-wide methylation intensities can also be measured with affinity-based immunoprecipitation (MeDIP), and quantified with a microarray (MeDIP-chip, Infinium platform). Methylation of candidate genes can also be measured with restriction enzymes that recognize only demethylated CpG regions (HELP assay), or pyrosequencing. Next Generation Sequencing approaches are also becoming increasingly popular, more cost effective, and provide global sequencing for histone modification (ChIP-seq) and methylation (MeDIP-seq), often integrating these with other epigenetic mechanisms. This section will describe each method and provide the reader with technologies and recommendations to aide in the design and implementation of an epigenetic study.

Histone modification analysis

Histone modification signals can be captured with chromatin immunoprecipitation (ChIP), which provides modification position approximation on the genome (Collas, 2010). The ChIP-chip technique combines this ChIP technology with a microarray (chip) to quantify the sum of binding sites on the genome (Aparacio, Geisberg, & Struhl, 2004). The ChIP-seq technique (see Next Generation Sequencing) has become a popular technique compared to ChIP-chip. Unlike ChIP-seq, ChIP-chip requires more amplification, multiplexing is not possible (Park, 2009), and the results have a lower resolution that are limited to the coverage provided by the selected microarray (Evertts, Zee, & Garcia, 2010). Nimble Gen offers a whole-genome ChIP-chip tiling array that allows the investigator to choose between

ordering the entire genome set or individual arrays within a set (http://www.nimblegen.com/products/chip/wgt/index.html). Single gene ChIP technologies are available that target antibodies against specific histone modifications. Mass spectrometry also allows the measurement of mass-to-charge ratio of peptides (Evertts et al.) and allows for changes in modification to be quantified during chromatin assembly (Deal & Henikoff, 2010).

When performing any microarray experiment, it is important to address concerns that may compromise the integrity of the experiment, including: image acquisition, background subtraction, standard normalization and the need to control for biases from dye (Buck & Lieb, 2004). Additionally, the reproducibility of the histone-modification results depends on the quality and specificity of antibodies used. Antibodies may exhibit appropriate specificity, but are ineffective when subjected to ChIP reagents (Egelhoffer et al., 2010). The Center for Biomedical Informatics at Harvard Medical School has developed an online repository that allows investigators to search for antibodies subjected to validation tests (http://compbio.med.harvard.edu/antibodies/about). It is important to note that this validation data should be used as a guide and investigators are encouraged to validate their own findings.

Bisulfite-conversion based methylation analyses

Bisulfite-conversion of unmethylated cytosines to uracils remains the gold-standard to evaluate methylation (Huang, Huang, & Feng, 2010). Bisulfite-conversion based microarrays use probes that hybridize targets to methylated and unmethylated regions, and release a fluorescent intensity that denotes methylation status (Huang et al.). Recent research indicates that tissue-specific methylation occurs in CpG island shores rather than previously targeted CpG islands (Irizarry et al., 2009); therefore, CpG islands alone are not sufficient to reveal differentially methylated regions and methylome evaluation should also include CpG shores (Gupta, Nagarajan, & Wajapeyee, 2010). Like other non-sequencing-based methods, the results of this platform are “susceptible to certain polymorphisms that were not known or considered at the time the array was designed” (Rakyan, Down, & Balding, 2011, p. 532). Illumina offers the Infinium HumanMethylation450K which provides a whole-genome analysis of methylation intensities of more than 450,000 sites, including CpG islands, shores and other CpG sites outside of islands (for more information see http://www.illumina.com/products/methylation_450_beadchip_kits.ilmn). Candidate gene methylation assessment can be accomplished through technologies such as the EpiTYPER (for more information see http://www.sequenom.com) that uses bisulfite converted DNA as a template for PCR and after modification and cleavage of the PCR product, mass spectrometry is performed to quantify methylated and non-methylated DNA.

Bisulphite-based sequencing (BS-seq) uses bisulphite converted DNA as a template, PCR amplification occurs, and sequencing of the resulting fragments provide a global view of methylation with minimal bias toward CpG dense regions. This approach provides the highest level of coverage and resolution, but is not capable of distinguishing between methylated and hydroxymethylated cytosine bases. BS-seq can be used for both a genome-wide or candidate gene approach. Pyrosequencing examines the methylation intensity of specific sites or genes of interest. Illumina offers a single-site resolution methylation assay that uses bisulfite conversion and pyrosequencing to produce high resolution results (http://www.illumina.com/technology/veracode_methylation_assay.ilmn).

Affinity-based methylation analyses

Genome-wide affinity-based microarrays use enzyme recognition sites within CpG sites that enrich the methylated fraction of the genome. The MeDIP-chip technique (Methylated DNA

Immunprecipitation-chromatin Immunoprecipitation) immunoprecipitates the methylated portion of genomic DNA with an antibody, and is followed by quantification of methylation with a microarray. This technique yields a restricted resolution that is limited by the type of array used. MeDIP-chip should be validated with quantitative PCR, though referencing is not required since bisulfite conversion does not occur. ArrayStar offers MeDIP-chip services that include quality assessments for both methods (http://www.arraystar.com/Microarray/service_main.asp?id=181).

Restriction endonuclease-based methylation analysis

Restriction endonucleases have been adapted to discriminate methylated from unmethylated regions in the DNA (Edwards et al., 2010). This approach uses restriction enzymes that recognize only unmethylated sites, and are therefore unable to cut methylated portions of DNA. This method, combined with high throughput sequencing is limited by the availability of restriction enzyme sites in the target DNA (Gupta et al., 2010). Additionally, this technique requires large amounts of DNA (Biotage, 2007). Advantages for this approach include: a simplified data analysis, straightforward protocol, and it does not require bisulfite-conversion. The use of restriction enzymes to analyze methylation can be used for either candidate-gene or genome-wide studies (Gupta et al.) and has been used as a method of methylation mapping analysis (Edwards et al., 2010).

Data Quality assessments are important to incorporate into an epigenetic study. Quantile and LOESS normalization is recommended, which assumes a similar total strength (source). Additionally, bisulfite-based experiments, especially pyrosequencing since PCR is highly variable, should include verification in independent samples to distinguish methylation from incomplete bisulfite conversion (Laird, 2010). Incomplete conversion of methylated cytosines remains a major weakness of bisulfite-conversion based analysis techniques. Fully methylated and fully unmethylated controls should be provided by commercial vendors which allow the investigator to evaluate bisulphite-conversion efficiency.

Next generation sequencing (NGS) for histone modification analysis

DNA sequencing from epigenetic events may provide a first step toward quantification of epigenetic mechanisms. Similar to ChIP-chip, ChIP-seq uses antibodies to enrich for histone modifications, but is instead followed by high-throughput sequencing that measures gene expression levels (Evertts et al., 2010). This technique determines the genome-wide patterns of modified chromatin, including: histone methylation, acetylation status and binding regions for proteins (Werner, 2010). Unlike ChIP-chip, ChIP-Seq offers higher resolution with fewer artifacts, greater coverage, and requires less DNA. Illumina offers a ChIP-seq assay that provides a wide range of binding sites with varying strength (http://www.illumina.com/technology/chip_seq_assay.ilmn).

Next generation sequencing (NGS) for methylation analysis

MeDIP-seq (Methylated DNA Immunoprecipitation-Sequencing) is a high throughput sequencing technique of methylated DNA fragments that is aligned to a referenced genome. This technique is comparatively easier to analyze and interpret (Gupta et al., 2010); however, this method is best used to study hypermethylation of CpG-rich areas, since methylated CpG-rich sequences are more efficiently enriched than methylated CpG-poor sequences (Bibkova & Fan, 2009).

CONCLUSIONS

Nurse scientists should give much thought to how a genetic or genomic study could positively impact and move forward their program of research. When designing a genetic or

genomic research study it is paramount that one decides if they will take a polymorphism based, gene expression based or epigenetic based approach and then within the context of that study whether they will take a genetic or a genomic approach. This paper, while not providing an exhaustive review of available technologies, demonstrates the variety of technologies available for commonly used approaches, each with advantages and disadvantages. Availability of databases housing information to facilitate study design, data collection, interpretation of findings, and dissemination of data have greatly improved over the past decade. Investigators are encouraged to visit and utilize *in silico* resources when designing a research study to ensure they are conducting novel investigations and using up to date information.

Acknowledgments

The authors would like to acknowledge support available through a National Institutes of Health, National Institute of Nursing Research award “Targeted Research and Academic Training Program for Nurses in Genomics” (T32 NR009759).

References

Altman RB, Raychaudhuri S. Whole-genome expression analysis: challenges beyond clustering. Current Opinion in Structural Biology. 2001; 11(3):340–347.10.1016/S0959-440X(00)00212-8 [PubMed: 11406385]

Aparicio, O.; Geisberg, J.; Struhl, K. Current Protocols in Cell Biology. Vol. Chapter 17. Los Angeles, CA, USA: John Wiley & Sons, Inc; 2004. Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo.

Arcellana-Panlilio M, Robbins SM. Global gene expression profiling using DNA microarrays. American Journal of Physiology – Gastrointestinal and Liver Physiology. 2002; 282(3):G397–G402.10.1152/ajpgi.00519.2001 [PubMed: 11841989]

Bibkova M, Fan J. Genome-wide DNA methylation profiling. WIRE Systems Biology and Medicine. 2009; 2(2):210–223.10.1002/wsbm.35

Biotage. CpG methylation analysis by pyrosequencing: benchmarks and application. 2007 Mar. Retrieved July 11, 2011, from http://www.pyrosequencing.com/graphics/7424.pdf

Buck MJ, Lieb JD. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics. 2004; 83(3):349–360.10.1016/j.ygeno.2003.11.004 [PubMed: 14986705]

Campbell, AM. Molecular movies: DNA microarray methodology. 2001. Retrieved July 13 2011 , from http://www.bio.davidson.edu/courses/genomics/chip/chip.html

Collas P. The current state of chromatin immunoprecipitation. Molecular Biotechnology. 2010; 45(1):87–100.10.1007/s12033-009-9239-8 [PubMed: 20077036]

Corvin A, Craddock N, Sullivan PF. Genome-wide association studies: A primer. Psychological Medicine. 2010; 40(7):1063–1077.10.1017/S0033291709991723 [PubMed: 19895722]

Deal R, Henikoff S. Capturing the dynamic epigenome. Genome Biology. 2010; 11(10):218.10.1186/gb-2010-11-10-218 [PubMed: 20959022]

DiStefana, JK.; Taverna, DM. Technical issues and experimental design of gene association studies. In: DiStefano, Joanna K., editor. Disease Gene Identification: Methods and Protocols, Methods in Molecular Biology. 2011. p. 3-16.

DNA Genotek. Oragene RNA. 2011. Retrieved July 13, 2001, from http://www.dnagenotek.com/DNA_Genotek_Product_RNA_Overview.html

Edwards JR, O’Donnell AH, Rollins RA, Peckham HE, Lee C, Milekic MH, Bestor TH. Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. Genome Research. 2010; 20(7):972–980.10.1101/gr.101535.109 [PubMed: 20488932]

Egelhoffer TA, Minoda A, Klugman S, Lee K, Kolasinska-Zwierz P, Alekseyenko AA, Lieb JD. An assessment of histone-modification antibody quality. Nature Structural & Molecular Biology. 2010; 18(1):91–93.10.1038/nsmb.1972

Everts A, Zee B, Garcia B. Modern approaches for investigating epigenetic signaling pathways. Journal of Applied Physiology. 2010; 109(3):927–933.10.1152/japplphysiol.00007.2010 [PubMed: 20110548]

Genetics Home Reference. What are single nucleotide polymorphisms (SNPs)?. 2011. Retrieved July 19, 2011, from http://ghr.nlm.nih.gov/handbook/genomicresearch/snp

Grant S, Hakonarson H. Microarray technology and applications in the arena of genome-wide association. Clinical Chemistry. 2008; 54(7):1116–1124.10.1373/clinchem.2008.105395 [PubMed: 18499899]

Green ED, Guyer MS. National Human Genome Research Institute. Charting a course for genomic medicine from base pairs to bedside. Nature. 2011; 470(7333):204–213.10.1038/nature09764 [PubMed: 21307933]

Guenin S, Mauriat M, Pelloux J, Van Wuytswinkel O, Bellini C, Gutierrez L. Normalization of qRT-PCR data: the necessity of adopting a systematic, experimental conditions-specific, validation of references. Journal of Experimental Botany. 2009; 60(2):487–493.10.1093/jxb/ern305 [PubMed: 19264760]

Gupta R, Nagarajan A, Wajapeyee N. Advances in genome-wide DNA methylation analysis. Biotechniques. 2010; 49(4):iii–xi.10.2144/000113493 [PubMed: 20964631]

Hakonarson H, Grant S. Planning a genome-wide association study: Points to consider. Annals of Medicine. 2011; 43(6):451–460.10.3109/07853890.2011.573803 [PubMed: 21595511]

Hattersly AT, McCarthy MI. Genetic Epidemiology 5: What makes a good genetic association study? The Lancet. 2005; 366(9493):1315–1323.10.1016/S0140-6736(05)67531-9

Huang Y, Huang T, Wang L. Profiling DNA methylomes from microarray to genome-scale sequencing. Technology in Cancer Research and Treatment. 2010; 9(2):139–147. Retrieved from PubMed database. [PubMed: 20218736]

International HapMap Project. About the International HapMap Project. 2006. Retrieved July 18, 2011, from http://hapmap.ncbi.nlm.nih.gov/abouthapmap.html

International HapMap Project. What is the HapMap?. n.d. Retrieved July 18, 2011, from http://hapmap.ncbi.nlm.nih.gov/whatishapmap.html.en

Irizarry RA, Ladd-Acosta B, Wen Z, Wu C, Montano P, Onyango H, Feinberg AP. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nature Genetics. 2009; 41(2):178–186.10.1038/ng.298 [PubMed: 19151715]

Jorgensen TJ, Ruczinski I, Kessing B, Smith MW, Shugart YY, Alberg AJ. Hypothesis-drive candidate gene association studies: Practical design and analytical considerations. American Journal of Epidemiology. 2009; 170(8):986–993.10.1093/aje/kwp242 [PubMed: 19762372]

Ku CK, Loy EY, Pawitan Y, Chia KS. The pursuit of genome-wide association studies: Where are we now? Journal of Human Genetics. 2010; 55(4):195–206.10.1038/jhg.2010.19 [PubMed: 20300123]

Laird P. Principles and challenges of genome-wide DNA methylation analysis. Nature Reviews Genetics. 2010; 11:191–203.10.1038/nrg2732

Lander ES. Initial impact of the sequencing of the human genome. Nature. 2011; 470(7333):187–197.10.1038/nature09792 [PubMed: 21307931]

Life Technologies Corporation. LeukoLOCK total RNA isolation system. 2010. Retrieved July 13, 2011, from http://www.ambion.com/techlib/prot/fm_1923.pdf

Life Technologies Corporation. RNAlater tissue collection: RNA stabilization solution. 2010. Retrieved July 13, 2011, from http://www.ambion.com/techlib/prot/bp_7020.pdf

National Center for Biotechnology Information. Microarrays: Chipping away at the mysteries of science and medicine. 2007. Retrieved June 6, 2011, from http://www.ncbi.nlm.nih.gov/About/primer/microarrays.html

National Human Genome Research Institute [NHGRI]. Talking Glossary of Genetic Terms: Copy Number Variation (CNV). n.d. Retrieved July 19, 2011, from http://www.genome.gov/glossary/index.cfm?id=40

National Human Genome Research Institute [NHGRI]. Talking Glossary of Genetic Terms: Polymorphism. n.d. Retrieved July 26, 2011, from http://www.genome.gov/glossary/index.cfm?id=160

Nussbaum, RL.; McInnes, RR.; Huntington, FW. Thompson & Thompson genetics in medicine. 7. Philadelphia, PA: Saunders Elsevier; 2007.

Ordovás JM, Smith C. Epigenetics and cardiovascular disease. Nature Reviews Cardiology. 2010; 7(9):510–519.10.1038/nrcardio.2010.104

Park P. ChIP-seq: advantages and challenges of a maturing technology. Nature Reviews. 2009; 10(10): 669–680.10.1038/nrg2641

Passarge, E. Color Atlas of Genetics. New York: Thieme; 2007.

PreAnalytiX. PAXgene blood RNA: The better the source, the more to explore. 2010. Retrieved July 13, 2011, from http://www.qiagen.com/literature/render.aspx?id=200337

Psychiatric GWAS Consortium Coordinating Committee. Genomewide Association Studies: History, rationale, and prospects for psychiatric disorders. American Journal of Psychiatry. 2009; 166(5): 540–556.10.1176/appi.ajp.2008.08091354 [PubMed: 19339359]

Qiagen. RNAlater handbook. 2006. Retrieved July 13, 2011, from www.qiagen.com/literature/render.aspx?id=403

Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nature Reviews Genetics. 2011; 12(8):529–541.10.1038/nrg3000

Read J, Bentall R, Fosse R. Time to abandon the bio-bio-bio model of psychosis: Exploring the epigenetic and psychological mechanisms by which adverse life events lead to psychotic symptoms. Epidemiologia e Psichiatria Sociale. 2009; 18(4):299–310.10.1017/S1121189X00000257 [PubMed: 20170043]

Ruiz-Naváez EA, Rosenberg L, Wise LA, Reich D, Palmer JR. Validation of a small set of ancestral informative markers for control of population admixture in African Americans. American Journal of Epidemiology. 2011; 173(5):587–592.10.1093/aje/kwq401 [PubMed: 21262910]

Russo, VA.; Martienssen, RA.; Riggs, AD. Epigenetic Mechanisms of Gene Regulation. Vol. 32. Plainview, NY: Cold Spring Harbor Laboratory Press; 1996.

Sagenet. Description of SAGE. 2003–2005. Retrieved June 6, 2011, from http://www.sagenet.org/findings/index.html

Sequenom. MassARRAY® iPLEX® gold-SNP genotyping: From target discovery to HTP validating (version 2) [Brochure]. 2010.

The Gene Ontology. An Introduction to the Gene Ontology. 1999–2011. Retrieved July 19, 2011, from http://www.geneontology.org/GO.doc.shtml

The University of Texas MD Anderson Cancer Center: Department of Bioinformatics and Computational Biology. Sample size for microarray experiments. 2003–2010. Retrieved July 13, 2011, from http://bioinformatics.mdanderson.org/MicroarraySampleSize/

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science. 1995; 270(5235):484–487.10.1017/S1121189X00000257 [PubMed: 7570003]

Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE, Kinzler KW. Characterization of the yeast transcriptome. Cell. 1997; 88:243–251.10.1016/S0092-8674(00)81845-0 [PubMed: 9008165]

Werner T. Next generation sequencing in functional genomics. Briefings in Bioinformatics. 2010; 2(5):499–511.10.1093/bib/bbq018 [PubMed: 20501549]

Wilop S, Fernandez AF, Jost E, Herman JG, Brummendorf TH, Esteller M, Galm O. Array-based DNA methylation profiling in acute myeloid leukaemia. British Journal of Haematology. 2011; 155(1):65–72.10.1111/j.1365-2141.2011.08801.x [PubMed: 21790528]

Table 1

Online Genome Databases and Resources

Name and Address	Description
Database of Short Genetic Variations (aka SNP database) http://www.ncbi.nlm.nih.gov/snp/?term=	This database houses documented SNPs, microsatellites, and small-scale INDELs. It provides population specific allele frequencies; genotype data, genome location, and information on function (e.g., change in an amino acid).
International HapMap Project http://hapmap.ncbi.nlm.nih.gov/	This database is used to identify and select tagging SNPs. User defined criteria under the configure tab include population selection, R^2 cutoff values, and mean allele frequency cutoff. SNPs identified in the literature or dbSNP can also be included in the tagger SNP configuration.
Database of Genomic Structural Variation http://www.ncbi.nlm.nih.gov/dbvar	This database houses information on documented structural variants, including CNVs. User defined limits include criteria such as study design, method type (e.g., SNP genotyping, FISH), project ID, and variant type
Copy Number Variation (CNV) Project http://www.sanger.ac.uk/humgen/cnv/	This database provides CNV data from two projects (Global CNV assessment; High-resolution CNV discovery)
Genetics Home Reference http://ghr.nlm.nih.gov/	This website by the National Library of Medicine contains information concerning genetic conditions, genes, and chromosomes.
Talking Glossary of Genetic Terms http://www.genome.gov/glossary/index.cfm	This glossary provides definitions, illustrations, and animations of commonly used genetic/genomic terms.
The Gene Ontology Project http://www.geneontology.org/	This database can be used to identify genes whose products may impact a phenotype of interest. The domains covered include cellular component, molecular function, and biological process.
Catalog of Published Genome-Wide Association Studies http://www.genome.gov/gwastudies/	Database containing all published GWA studies attempting to genotype at least 100,000 SNPs in the initial stage
Genome-Wide Association Studies Data Repository http://was.nih.gov/	Website for the NIH Genome Wide Association Study Portal
The Genes, Environment, and Health Initiative http://www.genesandenvironment.nih.gov	Website for Genes, Environment and Health Initiative (GEI)
Database of Genotypes and Phenotypes http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gap	Database containing results of studies investigating genotype-phenotype interaction. Currently houses NIH GWAS repository.
Center for Inherited Disease Research http://www.cidr.jhmi.edu	Provides genotyping and statistical genetic services to investigators approved for access through competitive peer review process
Understanding the Basics of Microarrays http://www.ncbi.nlm.nih.gov/About/primer/microarrays.html	This publication from the National Center for Biotechnology Information (NCBI) provides an overview of DNA microarrays explaining gene expression, the technology underlying microarrays, the purpose and importance of microarrays, and the basics of microarray experiments.
Gene Expression Omnibus http://www.ncbi.nlm.nih.gov/geo	GEO: the Gene Expression Omnibus. GEO serves as public repository and online resource for storage and retrieval of gene expression data. GEO currently maintains microarray and serial analysis of gene expression (SAGE) data on over 100
European Bioinformatics Institute http://www.ebi.ac.uk/	The European Bioinformatics Institute (EBI) is a nonprofit organization that focuses on research and services in bioinformatics. EBI's website enables access to gene expression databases (Array Express Archive and Gene Expression Atlas) and microarray analysis tools (Expression Profiler, Next Generation and Bioconductor).
Serial Analysis of Gene Expression Portal http://www.sagenet.org	Sagenet provides a detailed description of serial analysis of gene expression (SAGE). This website also provides SAGE applications, publications, and resources.
Histone Database http://www.research.nhgri.nih.gov/histones	NHGRI histone database Histone sequence information, including posttranslational modifications
Antibody Validation Database http://compbio.med.harvard.edu/antibodies/about	Collect and to share experimental results on antibodies that would otherwise remain in individual laboratories, thus aiding researchers in selection and validation of antibodies.
Chromatin Structure and Function http://www.chromatin.us	Information on chromatin biology, histones and epigenetics (hosted by Jim Bone)

Name and Address	Description
Database for DNA Methylation and Environmental Epigenetic Effects http://www.methdb.de/	Human DNA methylation Database DNA methylation data readily available to public Future develop includes environmental impact on methylation
CpG Island Searcher http://www.uscnorris.com/cpgislands2/cpg.aspx	CpG island searcher CpG Island sequence search algorithm Allows for selection of % methylation and length of (ISLAND?) and gaps between islands
Catalogue of Parent of Origin Effects http://igc.otago.ac.nz/home.html	Imprinted Gene Catalogue Catalogue of parent of origin effects Can search by taxon, chromosome, gene name or key word
Database of Noncoding RNAs http://www.noncode.org	Knowledge database dedicated to ncRNA Information on: class, name, location, related publications, mechanism through which it exerts its function Includes all traditional ncRNAs, but excludes tRNAs and rRNAs
MicroRNA Database http://www.mirbase.org	MicroRNA data resource Searchable database of >16,000 published miRNA sequences and annotation – includes location and sequence of mature miRNA Can search by name, keyword, reference and/or annotation
Epigenome Network of Excellence http://www.epigenome-noe.net	Epigenome Network of Excellence Web site of European interdisciplinary epigenetics research network Includes protocols, an antibody database and reference information on epigenetics
Human Epigenome Project http://www.epigenome.org	The Human Epigenome Project Research Consortium Collaborative effort to catalogue and interpret genome-wide methylation patterns of all human genes and major tissues

Table 2

Online Commercial Resources Used in Manuscript

Name	Address
Applied Biosystems Incorporated	http://www.appliedbiosystems.com
Roche Applied Science	http://www.roche-applied-science.com
Illumina Incorporated	http://www.illumina.com
Affymetrix Incorporated	http://www.affymetrix.com
Millipore	http://www.millipore.com
Sequenom Incorporated	http://www.sequenom.com
Preanalytix	http://www.preanalytix.com
Life Technologies Corporation	http://www.lifetechnologies.com
Ambion	http://www.ambion.com
Qiagen Incorporated	http://www.qiagen.com
DNAGenoteck Incorporated	http://www.dnagenotek.com
Panomics	http://www.panomics.com
Roche Niblegen Incorporated	http://www.nimblegen.com
Arraystar Incorporated	http://www.arraystar.com