The NTP Biomolecular Screening Branch

Title

The NTP/DIR Mouse Methylome Project

BSB Scientists

John E. (Jef) French, Ph.D.
Scott S. Auerbach, Ph.D.
Alex Merrick, Ph.D.

Associated NTP Scientists

Angela King-Herbert, DVM, Ph.D., CMPB
Dave Malarkey, DVM, Ph.D., CMPB

Associated NIEHS Scientists

Trevor Archer, Ph.D., LMC
Takashi Shimbo, Ph.D., LMC
Keith Shockley, Ph.D., BB
Paul Wade, Ph.D. LMC

Collaborating Scientists

Simon Gregory, Ph.D., Duke University, Durham, NC
Jim Mullikin, NHGRI, NIH Intramural Sequencing Core, Rockville, MD
Ruchir Shah, Ph.D, SRA, RTP, NC

Background and Rationale

An individual's response to exposure related toxicity and concomitant disease is influenced at the genome level by genetic, epigenetic, gene-gene interactions (**intrinsic factors**), and interaction with the environment (**extrinsic factors**). Individual DNA sequence variation does not account for all of the heritability for susceptibility to toxicity and diseases such as asthma, cancer, or diabetes. One intrinsic factor that quantitative and molecular geneticists believe could contribute to the observed "missing heritability" is the **methylome** (Eichler et al., 2010), an individual's genome wide pattern of cytosine methylation. The methylome (a component of the epigenome) may be a major epigenetic modifier of the susceptibility to cancer and other chemical exposure related diseases. Major basic questions about the nature of epigenetic variation within individuals remain understudied. It is not known whether, or to what extent, DNA methylation patterns (or other epigenetic marks) are inherited from parent to offspring. Likewise, variability in patterns of DNA methylation within individuals is not well described at the genome level. Although extrinsic factors are hypothesized to impact the epigenome (and the methylome), the extent to which such interactions influence biological outcomes of exposure, and whether any such effects are heritable, remain unclear.

Presently, there is no mouse reference database for the methylome akin to the NTP/Perlegen DNA sequence database of 15 commonly used inbred strains (plus the C57BL/6 reference strain)(Frazer et al., 2007; Yang et al., 2007). The DNA sequence data has significantly increased our knowledge of the genomic structure of the inbred mouse and has provided the

basis for imputation of the haplotype structure of more than 90 inbred strains used in biological research (Kirby et al., 2010). The absence of a methylome reference database for the mouse significantly handicaps our knowledge and understanding of the mouse model in toxicology and environmentally related diseases and hinders the design and performance of hypothesis based genetic and epigenetic research studies to understand the associated mechanisms and relationships.

Two high content technologies have been recently developed that 1) permit genome-wide determination of cytosine methylation, DNA sequence variation, and RNA sequence at base pair resolution (massively parallel sequencing, bisulfite seq (BIS seq), RNA seq) from a single biological sample; and 2) fractionate DNA sequences using differential restriction and/or affinity capture (MMDE-seq) to enrich for methylated DNA sequences from limited quantities of biological material including formalin fixed, paraffin embedded (FFPE) tissue samples (Bormann Chung et al., 2010; Down et al., 2008; He et al., 2010; Hughes and Jones, 2007; Jacinto et al., 2008; Mill and Petronis, 2009; Mill et al., 2006; Serre et al., 2009). Together, these tools allow targeted interrogation of genomic regions of interest using bioinformatic data mining tools. The proposed study will use these technologies to create a definitive map of the mouse liver methylome from the two parental strains (C57BL/6N and C3H/HeN) and their F1 hybrid (B6C3F1/N) offspring that exhibit dramatically different rates of intrastrain and interstrain as well as sex dependent spontaneous liver cancers. The high, but variable, incidence of liver tumors in the F1 hybrid mouse often confounds interpretation of 2-year toxicology and carcinogenesis studies. Although highly penetrant quantitative trait loci have been identified in the C3H/He strain, the liver cancer incidence varies significantly in untreated control B6C3F1 mice from generation to generation. This variable incidence may be due in part to cytosine methylation in critical tumor suppressor genes, regulatory regions of the genome, and associated pathways. The reference database will aid our understanding of the relationship between variations in sporadic and induced disease incidence associated with individual variations in the methylome, DNA sequence, and exon specific transcript expression critical to understanding the potential functional consequences from generation to generation. The data from this project, will directly address critical knowledge gaps in the nature of the methylome, its variability across individuals, its association with disease, and its heritability across generations. Further, these data will create a reference for future investigations into environment-induced changes in methylome variation and its role in spontaneous and induced disease.

Key Issue, Hypothesis Tested or Problem Addressed

Scientific evidence indicates that individual differences in susceptibility to toxicity, sporadic disease, and exposure related disease within a population of individuals is based upon individual differences in genomic structure (e.g., individual differences in single nucleotide polymorphisms and copy number variation). However, this variation explains only part of the heritability of susceptibility to disease. The proposed research will facilitate our understanding of the role of individual differences in epigenomic structure within and between individuals, heritability of DNA methylation patterns, and the relationship of these events to disease susceptibility. By determining an individual's genomic DNA sequence, methylated sequences, and the exon-specific transcript expression, we can correlate genome wide methylation patterns with tissue specific transcript expression and create a reference database under controlled conditions. The mouse has been shown to be an excellent model organism for many human diseases. Using

syngeneic individuals from a population of inbred mice analogous to studies of human monozygotic twins, we can work toward separating "signal from noise" by understanding individual variations in SNPs, CNV, and cytosine methylation patterns that may affect micro RNA, long non-coding RNA and messenger RNA expression. These efforts are expected to facilitate identification of target sequences for hypothesis-based research in environmental health.

Approach (Research Plan)

AIMS

1. Sequence and catalog the genomic DNA sequence, genomic cytosine methylation pattern, and exon specific transcripts of both sexes of the C57BL/6N (B6) strain and C3H/HeN (C3) strain and the female B6 x male C3 and the female C3 x male B6 outcrosses that produce the female and males B6C3F1/N or C3B6F1/N hybrid progeny, respectively, to create a reference mouse DNA genome, methylome, and transcriptome database for each sex of each inbred strain under standard NTP specifications and controlled study and environmental conditions.
2. Determine the individual intra- and inter-strain differences in cytosine methylation (B6 and C3, and the F1 hybrid female and male) that may potentially explain differential toxicity and tissue specific disease outcomes within and between these inbred strains and their F1 hybrid.
3. Determine, correlate, and catalog each strain's individual methylome with its exon specific transcriptome (microRNA, long non-coding RNA, and messenger RNA transcripts) at both the quantitative (expression level) and qualitative (splicing) level.
4. Determine, correlate, and catalog heritable regions of the methylome (DNA sequence specific cytosine methylated sequences) and the heritability of the transcriptome of each strain.

Study Design

Animals: C57BL/6N (B6) female and male, C3H/HeN (C3) female and male, and their B6C3F1/N and C3B6F1/N female and male hybrid progeny will be used (Source: NTP Colony). All breeders and offspring mice will be clearly identified by tattoo and lineage. All mice will have specified tissues sampled at the same age and after being raised under the same environmental conditions as described.

Breeding Scheme: A total of 10 female and 10 male B6 mice and 10 male and 10 female C3 mice will be randomly selected from the NTP strain maintenance colony. One pair of female and male siblings from the same litter will be randomly selected from each breeding unit until 10 of each sex-strain pair have been isolated and uniquely identified after weaning and sex has been confirmed. From this population of randomly selected mice of each strain, 7 breeding pairs of B6 females and C3 males and 3 breeding pairs of C3 females and B6 males will be randomly paired for outcross to produce B6C3F1 or C3B6F1 hybrid female and male progeny. All mice will be uniquely identified and their lineage tracked and confirmed.

At 10 weeks of age, female mice and male mice will be randomly selected and pair mated to carry out the conventional and the reciprocal backcross as described. After confirmation of

pregnancy, the male mice will be removed and housed separately. At weaning, the B6C3F1/N and C3B6F1/N hybrid pups will be uniquely identified, sexed, and the females and males housed separately and their lineage tracked. At 17 weeks of age, their diet will be changed from NIH31 to NTP 2000 until each has reached 20 weeks of age.

Total mice: **Parental lines**: 10 mice/sex x 2 sexes x 2 strains (B6 and C3) = 40 mice to be used for tissue sampling as specified. **Progeny**: 1 mouse/sex x 2 sexes x 10 breeding pairs = 20 mice plus 2 additional female and male siblings from one randomly selected paired mating for both B6C3F1/N and C3B6F1/N (2 sexes x 2 siblings x 2 outcrosses = 8 mice) to be used for tissue sampling as specified. TOTAL: 68 mice (10 male and 10 females/strain plus 3 siblings/sex from 2 litters (B6C3F1/N and C3B6F1/N) for sampling and tissue prepared for sequencing and archival.

Environment: Conducted under standard NTP Specifications without test agent exposure.

Diet: All mice will be placed on NIH31 (NTP breeding diet) for the first 17 weeks of life and then switched to NTP2000 (study diet) for the final 3-week prior to euthanasia and tissue collection. This scheme is the best approximation of diet and dietary exposures in the production of B6C3F1/N hybrids for NTP studies.

Tissue samples: The primary tissue of initial interest is the liver. Liver was selected for the primary analysis because of its relative homogeneity (80-90% hepatocytes) and its relative importance to NTP carcinogenicity studies in the B6C3F1/N mouse (as noted above). The left lateral lobe will be rapidly removed and dissected in a dish on ice into 3 – 4 mm cubes and flash frozen in 1 mL Eppendorf screw cap tubes in liquid nitrogen and stored at -80°C prior to sample preparation for DNA isolation and library production for bisulfite sequencing and exon specific transcript expression analysis of individual mouse genomes.

Other tissues (adipose, brain, cardiac muscle, and skeletal muscle) will also be collected as described by multiple prosectors as rapidly as possible, flash frozen, and archived for future studies as warranted.

Molecular studies: Up to 3 liver samples for each sex-strain pair will be initially investigated to determine the within and between strain variation. More samples may be analyzed as required after statistical evaluation of the results

Phase 1 – C57BL/6N female, C3H/HeN male, B6C6F1 male and female

Phase 2 – C3H/HeN female, C57BL/6N male, C3B6F1 male and female

Phase 3 – additional replicates and targeted resequencing as necessary

We will incorporate into our study:

1. Bisulfite sequencing (BIS-Seq) with DNA sequence genomic controls
2. Targeted re-sequencing of specific sites (MMDE-seq)
3. Whole exon-specific transcriptome expression profiles (microRNA, long non-coding RNA, and messenger RNA transcripts; RNA-Seq)

Data Analysis:

1. Determine and catalog the genome wide cytosine methylation patterns: Bioinformatic methods will be employed to align and map reads to the genome of the 3 strains to create a high resolution map of the methylome

2. Identify heritable regions of the methylome: Compare the methylome of each of the parental B6 females to their B6C3F1/N female offspring and the C3 males to their B6C3F1/N male's offspring. For the reciprocal outcross, compare the methylome of each of the parental C3 females to their C3B6F1/N female offspring and the B6 males to their C3B6F1/N male offspring to determine effects of germline transmission of imprinted genes (cytosine methylation variation in known imprinted genes).
3. Determine intra-strain variation by sex: Identify differentially methylated sites within females and males of both parental strains (B6 and C3) and their F1 hybrid
4. Determine inter-strain variation by sex: Identify differentially methylated sites between the parental strains (B6 and C3) and their F1 hybrid
5. Determine whether sexual dimorphisms in DNA methylation patterns exist, and their extent across the genome.
6. Identify association between local methylation patterns in the genome and both quantitative and qualitative variation in the transcriptome. Correlative analysis will be performed to identify these relationships.

Liver samples from all mice at 20 weeks of age will be collected and flash frozen as described above in "Tissue Sample". From each of the 10 breeding pairs, the female C57BL/6N dam along with a male sibling and the male C3H/HeN sire along with a sibling female will have liver samples collected (10 pairs x 2 sexes x 2 strains = 40 samples). From the F1 progeny, 1 randomly selected male and 1 randomly selected female B6C3F1/N from each set of the 10 mating pairs will have liver samples collected and frozen as described. In addition, one set of B6C3F1 sibling females and males will be randomly identified and 3 individuals of each sex will also have liver samples collected [(10 pairs progeny x 1 strain x 2 sexes) + (2 progeny x 2 sex x 2 strains)] = 28 samples. Thus, there will be a total of 68 liver sample sets will be collected and available for processing and for analysis.

In the first phase of sequencing, only the liver samples from one B6 female (dam), one C3 males (sire), one male and one female B6C3F1/N hybrid (2 genomes, 4 BIS-Seq and RNA-Seq total) of the conventional outcross will be sequenced and analyzed. In the second phase, only the liver samples from one C3 female (dam), one B6 male (sire), and one female and one male C3B6F1/N hybrid (2 genomes, 4 BIS-Seq, and 4 RNA-Seq total) of the reciprocal outcross will be sequenced and analyzed. This will allow sufficient sequencing data to be analyzed and processed to determine the strategy for the third phase of sequencing to answer the question of within and between strain and within litter variation relative to the nature of the outcross and germline effects relevant to transgenerational outcomes. Further analysis (increasing the number of observations per strain/sex) samples will allow examination of the within and between strain variation of the males (C3 and B6C3F1) and the females (B6 and B6C3F1) and the reciprocal cross (if warranted). Further sequencing of samples will be carried out as warranted to complete the study aims by permitting comparison of sexual differences within and across the two inbred strains (B6 and C3) and by providing sufficient replicates to discriminate "signal from noise" (i.e., random vs. non-random cytosine methylation). Estimation of the strain and sex variation in stochastic methylation is deemed critical to identification of candidate genomic regions for hypothesis-based research on the epigenetic basis of strain and sexual variation in spontaneous liver cancer incidence.

Tiered Approach: Number of mice to be sequenced and analyzed

The initial critical question being asked in the first phase of this project is: What is the extent and genomic location of cytosine methylation variation within the B6 female and male and the C3 female and male inbred strains and within female and male siblings of a heterozygous hybrid strain of the conventional and reciprocal outcross. By defining this variation in methylated cytosine sequences and the corresponding DNA sequence context genome-wide, targeted sequencing strategies can be developed for hypothesis-based research, including the basis for trans-generational effects. To control efficiency and cost, three individuals (n = 2 or 3) of each sex and genotype will be considered the minimum number necessary to discriminate between random and non-random variation in methylated CpG sequences and the maximum number in regard to sequencing and development of high content data set costs. After examining the initial sequence results, sequencing of additional samples, including female and male B6 and female and male C3 to estimate within sex and strain variation and to increase the number of observation to more than 3 samples each sex/strain may be required in order to increase the power of detection.

Results/Progress

Breeding, sample collection, and archival of tissues will be undertaken by NTP and is in progress. Libraries for whole genome and bisulfite sequencing will prepared by the laboratory of Paul Wade, LMC, DIR, who have significant experience in the preparation of libraries. Bioinformatic analysis will be carried out by NIEHS and NIH bioinformaticians. The analysis tools for massively parallel sequencing or NexGen sequencing are still being developed and a major bioinformatic effort will be required. This effort will include sequence alignment and computational analysis to correlate DNA sequence, methylated sequence, and exon-specific transcripts.

Significance

The genetic basis (SNPs, CNV, somatic mutations, etc.) for susceptibility explains only part of the role of individual variation in heritable and derived phenotypic traits, including sporadic and environmental related diseases such as asthma, cancer, diabetes, obesity, etc. Inbred mouse models share significant features in genetic and genomic structures and susceptibility to disease with humans. The range of genetic variation in laboratory and wild-derived strains of mice is similar in magnitude to the variation in SNPs and CNV observed in human populations. This project will address fundamental questions in regard to DNA sequence and the methylome, its variability, heritability, relationship with gene expression, and with disease. Further, this project will propel construction of a reference database for the methylome of inbred strains as a research tool available to the NTP, DIR and the broader scientific community. It is anticipated that such information may spur further studies investigating the mechanistic basis by which the epigenome intersects with environmental exposure in disease incidence, susceptibility and severity.

Future Directions/Plans and Justifications

1. If successful and warranted by the results, we plan to extend the approach to other tissues collected from these animals and to additional inbred strains sequenced previously in the NTP-Perlegen research project. The intent is to develop a cohort of genetically and epigenetically diverse set of inbred strains, characterized across several tissues of interest for quantitative analysis of toxicity and disease phenotypes through haplotype and/or meiotic association mapping.
2. Targeted differential restriction and high throughput bisulfite sequence along with the DNA sequence can be used to develop genetic and epigenetic marker reference data set for high- resolution genome wide haplotype association mapping. Those highly significant markers within or near reference methylated cytosine sequences that are associated with quantitative traits of interest, targeted genome wide HTS bisulfite sequencing can be used for high resolution mapping and the data used to confirm SNP and CpG markers for haplotype association mapping of phenotypic traits. Highly significant candidate sequences may be further examined for functional validation using *in vitro* or *in vivo* models for single and multiple (trans-) generation studies.
3. Develop the tools to determine the value of using the FFPE archived tissues and targeted HTS bisulfite sequencing to investigate mechanisms of toxicity and disease association within NTP studies, and, where possible, compare NTP liver cancer specimens with DIR human liver specimens to address the relevance to human disease.

References

Bormann Chung CA, Boyd VL, McKernan KJ, Fu Y, Monighetti C, Peckham HE, et al. 2010. Whole methylome analysis by ultra-deep sequencing using two-base encoding. PLoS One 5(2): e9320.

Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, et al. 2008. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Nat Biotechnol 26(7): 779-785.

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. 2010. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11(6): 446-450.

Frazer KA, Eskin E, Kang HM, Bogue MA, Hinds DA, Beilharz EJ, et al. 2007. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. Nature 448(7157): 1050-1053.

He D, Choi A, Pipatsrisawat K, Darwiche A, Eskin E. 2010. Optimal algorithms for haplotype assembly from whole-genome sequence data. Bioinformatics 26(12): i183-190.

Hughes S, Jones JL. 2007. The use of multiple displacement amplified DNA as a control for methylation specific PCR, pyrosequencing, bisulfite sequencing and methylation-sensitive restriction enzyme PCR. BMC Mol Biol 8: 91.

Jacinto FV, Ballestar E, Esteller M. 2008. Methyl-DNA immunoprecipitation (MeDIP): hunting down the DNA methylome. Biotechniques 44(1): 35, 37, 39 passim.

Kirby A, Kang HM, Wade CM, Cotsapas C, Kostem E, Han B, et al. 2010. Fine mapping in 94 inbred mouse strains using a high-density haplotype resource. Genetics 185(3): 1081-1095.

Mill J, Petronis A. 2009. Profiling DNA methylation from small amounts of genomic DNA starting material: efficient sodium bisulfite conversion and subsequent whole-genome amplification. Methods Mol Biol 507: 371-381.

Mill J, Yazdanpanah S, Guckel E, Ziegler S, Kaminsky Z, Petronis A. 2006. Whole genome amplification of sodium bisulfite-treated DNA allows the accurate estimate of methylated cytosine density in limited DNA resources. Biotechniques 41(5): 603-607.

Serre D, Lee BH, Ting AH. 2009. MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. Nucleic Acids Res 38(2): 391-399.

Yang H, Bell TA, Churchill GA, Pardo-Manuel de Villena F. 2007. On the subspecific origin of the laboratory mouse. Nat Genet 39(9): 1100-1107.