GPN

Funded by NICHD

Genomic and Proteomic Network for Preterm Birth Research

Scientific Protocol

Version 9.0 (replacing former version 8.0)

Prepared by:

Data Management, Statistics and Informatics Core

for the GPN Network

Collaborative Center for Statistics in Science

Yale University School of Medicine

300 George Street, Suite 523
New Haven, CT 06511
(203) 785-5185

Last updated: October 19, 2009

Table of Contents

1 Acronyms

Advisory Board	AB	Multiple Reaction Monitoring	MRM
Data Management, Statistics, and Informatics	DMSI	Odds Ratios	OR
Deoxyribonucleic Acid	DNA	Optical Character Recognition	OCR
Estimated Date of Delivery	EDD	Preterm Birth	PTB
False Discovery Rate	FDR	Preterm Premature Rupture of Membranes	pPROM
Gestational Age	GA	Prostate Specific Antigen	PSA
Gene Ontology	GO	Protected Health Information	PHI
Genomic and Proteomic Network for Preterm Birth Research	GPN	Quality Control	QC
Good Bioanalytical Practices	GBP	RIBONUCLEIC ACID	RNA
Hardy-Weinberg Equilibrium	HWE	SEXUALLY TRANSMITTED DISEASES	STD
Health Insurance Portability and Accountability Act	HIPAA	Single Nucleotide Polymorphism	SNP
Identification	ID	Spontaneous Preterm Birth	SPTB
Institutional Review Board	IRB	Stable Isotope Labeling by Amino Acids in Cell Culture	SILAC
Laboratory Information Management System	LIMS	Term Delivery	TD
Last Menstrual Period	LMP	Tumor necrosis Factor	TNF
Linkage Disequilibrium	LD	Yale Microarray Database	YMD
Liquid Chromatography/Tandem Mass Spectrometry	LC-MS/MS	Yale Protein Expression Database	YPED

2 Introduction

2.1 Overarching Hypotheses

The ultimate goal of the Genomic and Proteomic Network for Preterm Birth Research (GPN) is to study the genetic and environmental etiologies and mechanisms of spontaneous preterm birth (SPTB). Understanding those mechanisms will help us predict SPTB, and design more effective prevention strategies for SPTB as well as more effective treatment strategies for sequelae of preterm birth.

2.2 Specific Aims

The specific aims of the network are: (a) design and implement hypothesis-driven, mechanistic studies based on wide-scale, high-output genomic, proteomic and metabolomic strategies; and (b) provide the research community with a public, web-based genomic, proteomic, and metabolomic database for data mining and data deposition.

2.3 Primary Hypotheses of the GPN

1. Mothers who deliver SPTBs have a unique maternal genetic polymorphism profile including gene-gene and gene-environment interactions distinct from mothers who deliver at term.
2. Maternal proteomic and metabolomic profiles can prospectively identify women at risk for a spontaneous preterm birth.
3. The expression profile (at mRNA, protein and metabolite levels) in the uterus reflects mechanism and provides insight into processes underlying SPTB.
4. Unique genomic, proteomic, and metabolomic profiles are mechanistically associated with development of preterm premature rupture of membranes (pPROM). Specifically, there are unique genomic and metabolomic profiles associated with:
    a) pPROM, which differentiate pregnancies with and without pPROM among pregnancies delivered preterm and not in labor.
    b) Onset of labor in pPROM pregnancies which differentiate pregnancies with and without labor among pregnancies with pPROM and delivered preterm.

2.4 Secondary Hypotheses of the GPN

Closely related to the primary hypotheses, this network also addresses a secondary hypothesis:

- SPTBs in patients with prior preterm birth(s) have different maternal genetic polymorphism profiles including gene-gene and gene-environment interactions between mothers delivering at term and with recurring preterm deliveries.

2.5 Primary Study Outcome

For the purpose of this network, SPTB, spontaneous onset of labor, term delivery (TD), rupture of membranes (ROM), premature ROM (PROM) and preterm PROM (pPROM) are defined as follows:

2.5.1 Spontaneous early preterm birth

1. Gestational age at delivery – 20 weeks 0 days to 33 weeks 6 days AND
2. Spontaneous onset of labor.

2.5.2 Spontaneous late preterm birth

1. Gestational age at delivery – 34 weeks 0 days to 36 weeks 6 days AND
2. Spontaneous onset of labor.

2.5.3 Spontaneous term delivery

1. Gestational age at delivery 39 weeks 0 days to 41 weeks 6 days.
2. Spontaneous onset of labor.

2.5.4 Rupture of the membranes (ROM)

Membrane rupture is documented by the presence of any two of the following:

1. Pooling of fluid in the vaginal vault.
2. Positive Nitrazine test.
3. Positive Ferning of dried vaginal fluid observed microscopically.
4. Amniosure or α– Microglobin tests positive

or any one of the following:

5. Indigo carmine pooling in the vagina after intra-amniotic installation.
6. Visible leaking of amniotic fluid from the cervix.

2.5.5 Premature rupture of the membranes (PROM)

PROM refers to rupture of the membranes occurring before the onset of labor.

2.5.6 Preterm premature rupture of the membranes (pPROM)

We define pPROM as PROM between 20 weeks and 0 days and 36 weeks 6 days.

2.5.7 Spontaneous Onset of Labor

Labor is defined as follows for the purpose of the study. Early preterm labor occurs between 20 weeks 0 days and 33 weeks 6 days gestation. Term Labor occurs between 39 weeks 0 days and 41 weeks and 6 days.
Spontaneous onset of labor – Operational definition may be adjusted during initial course of the study if yielding insufficient number of cases:

≥ 4 spontaneously occurring contractions (each > 30 sec duration) in 20 min **OR** 10 contractions per hour **AND** absolute dilation of >=2 cm in addition to one of the following: a) increase in dilation of ≥ 1 cm b) effacement of ≥ 75%

2.5.8 Lethal Fetal Anomalies

Lethal fetal anomalies include anencephaly; Alobar holoprosencephaly with cyclopia, ethmocephaly or cebocephaly; Exencephaly; Body Stalk Anomaly/Limb Body Wall Complex; Ectopia Cordis with intracardiac heart defect or other associated anomalies; Pentalogy of Cantrell with complex congenital heart defect, large omphalocele, or other associated anomalies; Absence of a functioning renal system (i.e., renal agenesis or bilateral multicystic dysplastic kidneys with non-visualization of the bladder); Non-immune hydrops with associated cystic hygroma or cardiac defect; Short-limbed, short-rib skeletal dysplasia syndromes (Achondrogenesis, Campomelic dysplasia, Thanatophoric dysplasia, Short-rib Polydactyly syndrome—Type I Saldino-Noonan and Type II Majewski); Severe defects of skeletal mineralization—Hypophophatasia, Type II osteogenesis imperfecta; and Triploidy, Trisomy 13, Trisomy 18.

2.6 Purpose of the Present Study Protocol

This protocol describes the background, design and organization of the study and may be viewed as a written agreement between the study investigators. It is reviewed by the Advisory Board (AB), and is approved by the Steering Committee (SC) and the Institutional Review Board (IRB) of each clinical site before recruitment begins. Any changes to the protocol during the study require the approval of the Steering Committee.

A manual of operations supplements the protocol with detailed specifications of the study procedures.

3 Background

Preterm birth, which is defined as birth prior to 37 weeks of gestation, is a leading cause of infant morbidity and mortality. In the U.S. population, approximately 12 percent of all births are preterm. This accounts for approximately 480,000 infants born annually (Martin, Hamilton et al. 2005). The incidence of preterm birth, however, is not equally distributed among races and ethnic groups. For example, African-Americans have the highest rate of preterm birth, followed by Mexican-Americans, Asians, and Caucasians. Strikingly, a substantial health disparity exists between African-Americans and Caucasians, with African-Americans being 1.6 times more likely to deliver preterm infants than Caucasians (Martin, Hamilton et al. 2005).

Most of the infant mortality and morbidity of preterm birth is associated with the two percent of infants born very preterm (birth at less than 32 weeks of gestation) (Martin, Hamilton et al. 2005). Excluding congenital malformations, preterm birth accounts for approximately 70 percent of all neonatal deaths and nearly 50 percent of long-term neurological problems (Mathews, Menacker et al. 2004; Kramer, Demissie et al. 2000; Wood, Marlow et al. 2000; and Hack and Fanaroff 1993). These long-term neurological problems include serious physical and mental disabilities, such as cerebral palsy, mental retardation, as well as vision and hearing loss.

Despite decades of research, there has been little progress in developing effective interventions to prevent preterm birth. In fact, the rate of preterm birth has increased slightly over the last several decades. Although advances have been made in identifying some of the possible causes of preterm birth, such as intrauterine infection, uterine bleeding, excessive uterine stretch, maternal psychosocial stress, and fetal physiological stress, a much deeper understanding at the molecular level is necessary to aid in formulating effective interventions. Thus, a new initiative is required to address this need as a means to accelerate knowledge in the mechanisms of preterm birth.

Although about 25 % of preterm births are medically indicated because the mother and/or fetus are at high risk, the majority of preterm births occur spontaneously. Spontaneous preterm birth (SPTB) is a complex trait that involves a number of pathologies and numerous risk factors. In the study of such a complex trait, it is unlikely that a single marker will ever be identified that will be both highly sensitive and specific. It is more likely that combinations of genetic, protein and/or metabolite markers, in association with environmental factors, will be required to achieve the goal of a truly sensitive and specific diagnostic profile for spontaneous preterm birth.

The rapidly evolving disciplines of genomics, proteomics and metabolomics are now beginning to be applied to the study of many diseases (Hirschhorn and Daly 2005; Hunter 2005; Carlson, Eberle et al. 2004; Calvo, Liotta et al. 2005; Craig and Stephan 2005; Dunckley, Coon et al. 2005; Dietel and Sers 2006; and Nielsen and Oliver 2005). Using these high-throughput analytical tools, it is possible to screen a vast number of molecules (in the thousands to hundred of thousands range) and to correlate a molecular signature that is highly diagnostic for a particular disease.

In the past, a candidate gene approach has been applied to the problem of SPTB. A number of investigators have shown associations with certain polymorphisms in a number of genes believed to be important in various pathways leading to SPTB (Crider, Whitehead et al. 2005 and Giarratano 2006). These include genes involved with the inflammatory cytokine pathways:

tumor necrosis factor –alpha (TNF), interleukin-1 receptor antagonist, interleukin 6, interleukin 4 and toll-like receptor 4. Others investigators have studied polymorphisms in other pathways that are associated with SPTB. These include matrix metalloproteinase 1 and 9, involved in fetal membrane rupture; beta-2-adrenergic receptor, involved in uterine quiescence; vascular endothelial growth factor, involved in angiogenesis and vascular homeostasis; and Factor 5, involved in the coagulation cascade.

The largest genetic association to date is a study examining 426 SNPs for 55 genes in 300 SPTB (delivered less than 37 weeks) women and 458 term controls in a multi-ethnic population (Hao, Wang et al. 2004). These candidate genes were chosen based on five suspected pathogenic pathways (infection/inflammation, uteroplacental pathology, maternal and fetus stress, premature contraction pathway, environmental toxins) involved with SPTB based on biological plausibility. The results showed that coagulation factor V haplotypes (uteroplacental pathology pathway), was significantly associated with SPTB in all ethnic populations. Coagulation factor V codes for a protein cofactor in the prothrombin-activating complex and accelerates Xa-catalyzed conversion of prothrombin. Interleukin1 receptor 2 (infection/inflammation pathway), nitric oxide synthase 2A (uteroplacental pathology pathway), and opioid receptor mu 1 (environmental toxins pathway) gene haplotypes were also found to be associated with SPTB but only in Black, White and Hispanic groups, respectively. A limitation of this study was that the study population was skewed towards late SPTB, consequently these findings may not be relevant to early SPTB. Moreover, all of these studies were based on relatively small study populations and thus need to be verified in larger cohorts and also, importantly, studies should be designed to examine multiple gene-gene and gene-environmental interactions.

Although studies based on a candidate gene(s) approach studying SPTB are plentiful in the literature, there is a paucity of studies based on high-throughput transcription profiling. Besides several animal studies, only one human study could be found in the literature using a high-throughput approach (Esplin, Peltier et al. 2005). These investigators used microarray analysis of 6000 genes to characterize the labor-selective transcriptome of the human myometrium during labor. One highly up-regulated transcript, monocyte chemotactic protein-1 (MCP-1), a pro-inflammatory cytokine, was further examined with respect to SPTB. The investigators found that MCP-1 levels were elevated in the AF of women experiencing preterm labor who delivered within 7 days of presentation as compared to those delivering beyond that time period, suggesting that the measurement of MCP-1 in the amniotic fluid may be useful as a predictor of time to delivery for SPTB.

Most recently, gene expression analyses that use arrays of smaller numbers of genes selected for their involvement in specific biologic processes, or gene ontologies, offer the opportunity to evaluate human parturition from the perspective of fundamental biologic function that include not only individual gene products at the molecular level but also biologic processes accomplished by ordered assemblies of molecular functions, cellular component functions and sequence ontology (Harris et al, 2005). Haddad and associates (Haddad et al, 2006) have recently applied this approach to human term chorioamnionic membranes from women in labor and not in labor (Haddad et al, 2006). Restricting their analysis to cases with no histologic evidence of inflammation, they demonstrated that human spontaneous labor without histologic chorioamnionitis is characterized by an acute inflammation gene expression signature.

Investigators have also taken a candidate molecule approach based on biological plausible to identify markers associated with SPTB (Goldenberg, Goepfert et al. 2005). Although these markers have a high negative predictive value, they generally suffer from a poor positive predictive value and thus have limited clinical utility. Hence, it is unlikely that the measurement of a single or a small number of molecules will predict the clinical outcome of an individual effectively for an accurate diagnosis.

Recently, investigators in the field of SPTB have taken a high-throughput proteomic approach using mass spectroscopy. Several investigators have identified unique protein profiles in the amniotic fluid that are associated with several risk factors for SPTB. Two studies identified a molecular profile that correlated to women with an intra-amniotic infection (Gravett, Novy et al. 2004 and Park, Yoon et al. 2006). Another study identified a molecular profile correlated with intra-amniotic inflammation (Buhimschi, Christner et al. 2005). In addition to other proteins, all three studies identified proteins from the calgranulin family (A, B and/or C) that were consistently associated with infection and/or inflammation. The calgranulins are expressed by macrophages and epithelial cells in acutely inflamed tissues. Although initially promising, these tests using amniotic fluid will have limited generalized clinical utility because of the invasive procedure required to obtain the specimen. A more promising area of investigation would be to analyze readily available specimens (non-invasively obtained) such as blood, urine, saliva or vaginal fluid. Moreover, as noted earlier in the genomic studies of SPTB, all of these studies were based on small study populations and thus need to be also validated in larger cohorts.

In the near future, because of the rapid development of high-throughput technologies coupled with an expected concomitant reduction in expense, a more complete analysis of the genome, proteome and metabolome will be possible allowing a more comprehensive assessment of potential molecular signatures associated with SPTB, in lieu of the "educated guess" or candidate molecule(s) approach. Furthermore, ascertaining these molecular signatures will likely provide valuable insight into the mechanisms underlying SPTB that would be useful for formulating more effective interventional strategies. Thus, the network intends to take such a high-throughput approach for the study of SPTB and to extend these findings to understand the underlying mechanisms.

Substantial strides have been made in the areas of genomics and proteomics, both of which have made major impacts in accelerating medical advances in many fields of medicine. However, the medical advances brought by these revolutionary technologies are generally due to coordinated, wide-scale, high-output endeavors that are presently beyond the resources of investigators in the field of preterm birth research. In addition, the research community lacks a centralized database for data mining and the deposition of genomic, proteomic, and metabolomic data. Hence, this collaborative network was created with the purpose to help fill these gaps. It is anticipated that this network will hasten a deeper understanding of the pathophysiology of SPTB at the molecular level, discover novel target molecules and diagnostic biomarkers, and ultimately aid in formulating more effective interventional strategies for SPTB prevention.

4 Study Design

4.1 Design Summary

To test our primary hypotheses we propose to conduct the following three observational studies:

I. **Case-Control Study** of 1000 spontaneous early preterm births before 34 weeks (cases) and 1000 spontaneous term deliveries between 39 and less than 42 weeks (controls). The controls must not have history of prior PTB (before 37 weeks of gestational age). Cases will be 1:1-matched with controls according to ethnicity, age, and parity. The criterion for cases of a spontaneous early preterm delivery before 34 weeks was chosen based on the clinical relevance due to perinatal morbidity and mortality and feasibility of recruitment of sufficient number of cases (to achieve adequate power) within a 2-year period. The criterion for controls of no prior PTB before 37 weeks was chosen to maximize the phenotypic (genotypic) differences between cases and controls without affecting the feasibility of the study.

II. **Longitudinal Cohort study** of 500 women with a history of previous SPTB before 37 weeks recruited at or before 18 weeks of pregnancy. The recruitment criterion of a prior preterm delivery was chosen to enrich the cohort for SPTB cases. This cohort is expected to have 30 projected SPTBs prior to 34 weeks and will be 1:2-matched with 60 SPTBs between 34 and less than 37 weeks and 1:2-matched with 60 term deliveries (between 39 and less than 42 weeks) according to ethnicity, age, and parity. The number of patients in each group was based on other studies of similar nature and sample sizes (Eisen, Spellman et al. 1998) that have revealed statistically important findings.

III. **Expression Profiling Study** of six groups of 20 patients. Four groups are without pPROM: in labor or not in labor at term or preterm before 35 weeks. Two groups are all pPROM: one followed by spontaneous labor before 35 weeks and the second delivered electively before 35 weeks. We selected 20 in each group because this is a reasonable number of patients to study RNA/protein/metabolite expression based on the literature (Zhang, Yu et al. 2001) and limited by affordability.

4.2 Gestational Age Determination

For each of the network studies, a 'Project Gestational Age' will be calculated from an established 'Project Estimated Date of Delivery (EDD)'. Once determined for an individual participant, the Project EDD cannot be revised. Because the Project EDD depends on information from the earliest dating ultrasound, if no ultrasound has been performed previously, one must be performed before the patient can be enrolled in a Network protocol. For all Network protocols, the Project EDD (and hence the Project GA) will be determined as follows:

1. The first day of the last menstrual period (LMP) is determined and a judgment made as to whether or not the patient has a 'certain' LMP date.

2. If the LMP date is uncertain, the ultrasound measurements obtained at the patient's first ultrasound examination are used to determine the project gestational age, by the standard method of ultrasound gestational age determination at the specific clinical site.
3. If the date of the LMP is certain, and the ultrasound confirms this gestational age within the number of days specified in Table 1, then the LMP derived gestational age is used to determine the Project Gestational Age.
4. If the ultrasound determined gestational age does not confirm the LMP generated gestational age within the number of days specified then the ultrasound is used to determine the project gestational age.

Also see Table 1 below for the cutoffs of using LMP to determine gestational age.

Table 1. Cutoffs for Using LMP to Determine Gestational Age

Gestational age at first ultrasound by LMP	Ultrasound agreement with LMP
up to 19^6 weeks	$\pm$7 days
20^0 weeks to 29^6 weeks	$\pm$14 days

4.3 Serious Medical Conditions

For the purpose of this protocol, serious medical conditions will include:

1) Renal insufficiency with serum creatinine > 1.5, known proteinurina > 300 mg/24 hrs, or receiving dialysis.
2) Chronic liver disease with transaminases >2 times the upper limit of the normal range.
3) Organ transplant recipients.
4) Women with significant disability due to spinal cord injuries.
5) Severe pulmonary disorders (such as significant obstructive/restrictive disorders, pulmonary hypertension, cystic fibrosis, and chronic oxygen requirement).
6) Severe heart disease (such as cardiomyopathy with a diminished LVEF, obstructive valvular disease, replaced valve, uncorrected aortic coarctation or tetralogy, and severe IHSS).
7) Malignancy not in remission.
8) Established lupus anticoagulant syndrome, antiphospholipid syndrome, or thrombophyllia with a prior thrombotic event.
9) Advanced class diabetes mellitus (class C and greater).
10) Homozygous hemoglobinopathy.
11) Isoimmunized pregnancy, Red Cell Alloimmunization warranting antenatal surveillance or treatment (e.g. serial Doppler, amniocentesis, or transfusion).
12) Serious chronic medical conditions that require more than 1 maintenance medication for control:

a. Chronic Hypertension requiring multiple agents.
b. Asthma requiring medications other than inhaled beta agonists.
c. Psychiatric disorders requiring multiple agents.
d. Seizure disorders requiring multiple agents.
e. Chronic pain on multiple medications or methadone maintenance.

13) Other Medical conditions that exclude participation regardless of medication requirements:
a. SLE
b. Inflammatory Bowel Disease
c. Hyperthyroid Disease currently on medication (Hypothyroid disease is not an exclusion nor is a past history of Graves Disease).
d. Placenta previa that does not deliver vaginally (Case Control study) and central placenta previa diagnosed on a midtrimester scan (Longitudinal study). Placenta previa is not an exclusion for the Expression Profiling Study.
e. Abruptio placenta suspected prior to delivery.
f. Nephrolithiasis with planned surgical treatment postpartum.
g. Cholelithiasis with planned post partum surgical treatment postpartum.
h. Dermatoses of pregnancy requiring maintenance oral steroids.

14) Acquired Immunodeficiency Syndrome (but not mere HIV infection without AIDS).

15) Other severe medical conditions deemed to place the patient at high risk for medically-indicated preterm delivery.

16) For the purpose of Case Control and Expression profiling study exclusions: stillbirth is defined as a fetal demise diagnosed prior to the onset of labor or prior to the presentation to Labor and Delivery (L&D) for delivery. Fetuses that are alive at the time of admission but suffer demise during the hospital course should not be excluded due to the demise. In the Longitudinal study, a fetal demise diagnosed by ultrasound prior to enrollment is an exclusion.

4.4 Informed Consent Criteria

Written informed consent must be obtained from patients before they can be enrolled into the study. Full disclosure of the nature and potential risks of participating in the studies are to be made. Each site will develop its own consent forms according to the requirements of its own Institutional Review Board using the sample consent form provided for each study below. Each site also will develop its own patient authorization documents, as required by the HIPAA Privacy Rule, following the guidelines of its own institution.

Women who are not fluent in English will be provided informed consent documents in their primary language. If this is not available, the consent form must be translated by a person fluent in their primary language.

5 Case Control Study

5.1 Definitions

For the case control study, cases are defined as a preterm delivery between 20 weeks 0 days and 33 weeks 6 days (inclusive) following the spontaneous onset of labor (see definition Section 2.5.7) either in the presence or absence of pPROM (see definition 2.5.6), while controls are defined as a term delivery between 39 weeks 0 days and 41 weeks 6 days following the spontaneous onset of labor either in the presence or absence of PROM (see definition Section 2.5.5).

5.2 Eligibility Criteria

The Steering Committee and Protocol Subcommittee devoted substantial effort and discussions to determine the eligibility (inclusion and exclusion) criteria for this study. The overall rationale is to maintain scientific rigor in order to construct a study cohort in which the disease group will have the highest chance to share a common genomic basis that differs from the controls, and to maximize the feasibility by enhancing the study population.

5.2.1 Inclusion and Exclusion Criteria for Cases

5.2.1.1 Inclusion Criteria

1. Delivery between 20 weeks 0 days and 33 weeks 6 days (inclusive) following the spontaneous onset of labor (see definition 2.5.7).
2. pPROM (see definition 2.5.6) that is followed at any interval by the spontaneous onset of labor at <34 weeks gestation. (Patients with pPROM that experience spontaneous labor at < 34 weeks and 0 days gestation but undergo cesarean section for malpresentation or other obstetric indication are eligible for enrollment).
3. Gestational age at delivery estimated by and meeting criteria of Project gestational age.

5.2.1.2 Exclusion Criteria

1. Indicated delivery (for maternal or fetal indications).
2. Maternal uterine anomalies.
3. Multi-fetal gestations.
4. Known aneuploidy or lethal fetal anomalies (see Section 2.5.8).
5. Polyhydramnios (defined as Amniotic Fluid Index ≥25 cm or Deepest vertical Pocket ≥ 12 cm).
6. pPROM at less than 34 weeks 0 days that is not followed by spontaneous labor (if delivery occurs at any gestational age as a result of induction of labor or cesarean section without spontaneous labor the patient is excluded from enrollment).
7. Cervical Cerclage.

5.2.2 Inclusion and Exclusion Criteria for Controls

5.2.2.1 Inclusion Criteria

1. Delivery between 39 weeks 0 days and 41 weeks 6 days (inclusive).
2. PROM (see definition 2.5.5) that is followed at any interval by the spontaneous onset of labor between 39 weeks 0 days and 41 weeks 6 days are considered eligible for enrollment. (Patients with PROM that experience spontaneous labor between 39 weeks and 0 days and 41 weeks 6 days gestation but undergo cesarean section for malpresentation or other obstetric indication are eligible for enrollment).
3. Gestational age at delivery estimated by and meeting criteria of Project gestational age.

5.2.2.2 Exclusion Criteria

1. Indicated delivery (for medical or obstetrical complications not related to preterm labor).
2. Prior history of any preterm birth (spontaneous or indicated)
3. Maternal uterine anomalies.
4. Multi-fetal gestation.
5. Known aneuploidy or lethal fetal anomalies (see Section 2.5.8).
6. Polyhydramnios (defined as Amniotic Fluid Index >=25 cm or Deepest vertical Pocket ≥12 cm).
7. PROM that is not followed by the spontaneous onset of labor.
8. Cervical Cerclage

5.3 Screening for Eligibility

All patients are eligible for screening who deliver a singleton pregnancy following spontaneous onset of labor between 20 weeks 0 days and 33 weeks 6 days (cases) or between 39 weeks 0 days and 41 weeks 6 days (controls). The inclusion and exclusion criteria will be reviewed with the patient's chart. If an ultrasound examination has not been performed, one must be arranged prior to enrollment. Gestational age determination and assignment of the Project Gestational Age will be then performed based on the results of the ultrasound examination and patient's menstrual history according to criteria defined in Section 4.2 "Gestational Age Determination".

If a patient meets the criteria for enrollment and expresses interest in participating in the study, she will be told about the study design, its potential risks and benefits and asked to sign the informed consent and patient authorization (if applicable) forms. If she accepts, she will then be enrolled in the study. A copy of the signed consent form(s) will be provided to the patient.

5.4 Summary Table of Design and Data Collection

Table 2 provides an overall summary of data and sample collection for the Case Control study.

Table 2. Summary Table for Case Control Study

	Labor & Delivery SPTB (20 to < 34wks)	**Labor & Delivery TD (39 to <42 wks)**
N	1000	1000
Specimen Collection		
Maternal blood	X	X
Maternal Oragene saliva (if no maternal	X	X

blood)		
Cord blood	X	X
Neonatal Oragene saliva (if no cord blood)	X	X
Clinical and Demographic Data		
Demographics	X	X
Enrollment	X	X
Intake Form (I, II, III)	X	X
Current Medications	X	X
Psychosocial Form	X	X
Ultrasound and Labs	X	X
Hospital Admission Form	X	
Labor and Delivery Form	X	X
Neonatal Baseline	X	X
Adverse Event Form	*[1]	*[1]

*[1] Unscheduled forms to be used as needed

5.5 Recruitment and Enrollment Procedures

The following steps are the procedure for recruiting cases and controls:

1. An active surveillance system will be in place to notify (via pager or phone call) the clinical site coordinator (or designee) as soon as possible when a delivery is made at a participating hospital.
2. The site coordinator (or designee) will review the inclusion/exclusion criteria with the individual making the notification, and others if necessary and appropriate.
3. If selected, the site coordinator (or designee) will notify the responsible care provider regarding consenting the patient for GPN enrollment. The mother and legal guardian (if applicable) will be approached by hospital staff seeking to gain permission for contact by a representative of the GPN clinical site. An information brochure will be available for distribution by the surveillance staffs to help with this introduction to the GPN.
4. For patients who agree to be contacted, the site coordinator (or designee) will travel to the facility where the mother delivered or transferred to seek consent or assent for mothers who are minors.

   For those consenting to full or limited participation, the site coordinator (or designee) will proceed with in-hospital and other procedures as appropriate. Electronic records will be kept of those deliveries screened for eligibility to the case-control study.
5. If an eligible woman is contacted for study enrollment and refuses to participate, the site coordinator will record the mother's refusal and consider the next eligible patient if available.
6. All cases that fulfill the five steps will be enrolled.Mothers may be consented within 14 days of delivery.
7. Controls will be recruited only when there are cases that need to be matched.

8. Every control should be recruited to match with a case yet to be matched by race (Whites, Hispanics, African Americans, and others), maternal age (<20, 20-29, 30-39, 40+), and parity (0 or greater).
9. If multiple controls meet the matching criteria, the earliest control will be the first one to contact for this study. Then contact will be made to the next control that meets the matching criteria if the earlier one(s) cannot be enrolled.

5.5.1 In-hospital Procedures

Labor and delivery units at participating hospitals will have materials stored to facilitate implementing the case-control study in the site hospitals. These materials will include brochures describing the GPN to be given to potential participants, consent forms for cases and controls, containers for biological specimens, and "instruction sheets" and "checklists" for hospital personnel. The site coordinators will have appropriate data collection instruments on hand as well as pre-printed labels to affix to the hospital record, the consent forms, biological specimen containers, and hard copy data forms.

5.6 Study Procedures

In addition to information collected for determination of eligibility during the screening visit, the following information will be obtained by patient interview, review of the chart, and collection of the following samples.

5.6.1 Patient Data

Samples will be collected as soon after delivery as possible. Maternal data will be collected before discharge from the hospital. Maternal DNA and the psycho-social questionnaire will be obtained within 14 days of delivery. Neonatal DNA can be obtained at any time.

Patient data will be collected through chart reviews and patient interviews. Data will be entered using the Demographics, Enrollment, Intake (I, II, III), Current Medications, Ultrasound and Labs, Labor and Delivery and Neonatal Baseline Forms. Stress, anxiety, and depression during the pregnancy will be assessed by the Psychosocial Questionnaire. All forms and instructions are enclosed in the Manual of Operations. In addition, the Adverse Event Form will be used as needed and is unscheduled. The adverse event form will only be filled out for neonates up to 28 days post delivery (during the neonatal period) who experience an adverse event as defined in the Manual of Operations.

1. Demographic information: parental age, race, ethnicity, etc.
2. Medical history: pre-pregnancy weight, height, medications, STD history, medical conditions, etc.
3. Social history: marital status, years of education, alcohol use and tobacco use.
4. Obstetrical history including outcome of all prior pregnancies and dates of termination.
5. History of preterm labor or premature rupture of membranes symptoms, evaluation or hospitalization for those symptoms in current pregnancy.
6. Medications taken during current pregnancy.
7. History of cervical evaluation (manually or ultrasonographically) or fetal fibronectin measurements.

8. Pregnancy complications.
9. Labor: type, indications, membrane status, and induction method.
10. Delivery type and indications for cesarean section.
11. Neonatal outcome: sex, weight, length, Apgar score, neonatal complications, admission to intensive care unit, etc.
12. Maternal family history of SPTB.

5.6.2 Samples

See Manual of Operations for more details. Samples will be sent from Clinical Cores to the Analytical Core, and data will be sent from the Analytical Core to the DMSI Core.

1. Maternal blood (10 ml) will be collected to obtain DNA for genotyping and measurements of markers of pregnancy exposures (cotinine, heavy metals, folic acid, stress, etc.).
2. Maternal saliva (via Oragene collection) will only be collected if maternal blood could not be collected for DNA.
3. Cord blood (10 ml) will be obtained following delivery for fetal genotyping (DNA) and for serum for future proteomic and metabolomic analyses and measurements of markers of exposures during pregnancy.
4. Neonatal saliva sample (via Oragene collection) will be obtained only if cord blood could not be collected for fetal genotyping.

5.6.3 Receiving Samples from the Clinical Cores to the Analytical Core

1. Barcoding samples will be performed at clinical sites. Samples will be identified (i.e., maternal blood/gestational age, and maternal urine/gestational age) with the following assigned numbers at each clinical site:

- Alabama-Birmingham 1,000 –1,700 series
- Texas-Galveston 7,000 – 7,700 series
- Utah 4,001 – 4,700 series

2. Penn tracking system – caTISSUE is an Oracle-based tissue bank designed by caBIG™. Samples will be linked to TrialDB through an Excel file with barcodes and patient IDs.
3. Shipping instructions (i.e., dry ice, etc.) for individual samples are provided in the Manual of Operations. All samples will be shipped to University of Pennsylvania.

5.6.4 DNA Acquisition

DNA extractions: Automated extractions are performed on the Qiagen M48 robot, and samples are organized in 96-well plates with barcoding and sample tracking systems. Subsequent manipulations are performed on the Biomek FX liquid handling workstation with 96-channel pipetting. DNA quantitation is performed by dye-based assay on a Molecular Devices SpectraMax Plus plate reader.

5.6.5 Analysis

Genome-wide DNA analysis (genotyping of single nucleotide polymorphisms) will be conducted in the Penn Microarray Facility (Donald Baldwin).

5.6.6 Quality Assurance and Control of DNA Data

Samples for DNA extraction are shipped in EDTA blood tubes or Oragene containers on dry ice. Barcoded tubes and storage boxes are logged into the caTissue inventory database and stored in -80C freezers with back-up power and alarm systems. Controls for DNA extractions include genotyping plates that will include no template control wells and 10% randomly duplicated samples to measure concordance. The Illumina Infinium SNP and Affymetrix SNP GeneChip assays incorporate technical controls to monitor the performance of sample preparation steps and of the microarrays.

5.7 Statistical Analyses

5.7.1 Descriptive Statistics and Data Quality Control

On a regular basis, personnel at the DMSI Core will examine descriptive statistics for SPTB, prenatal variables, perinatal variables, and postnatal variables as well as polymorphisms in some selected SNPs. These data will be extracted from the databases and saved as flat line files, and then will be read into SAS for summary statistics. Progress will be reported at the Steering Committee meeting.

As described in Section 5.6.6, quality control steps will be taken by the Analytical Core before the data are sent to the DMSI Core for statistical analysis. For SNP data, we will examine concordance with the Hardy-Weinberg equilibrium and the Mendelian inheritance errors. The significance level of 0.05 will be used to alert inconsistencies or errors. Inconsistencies and errors as a result of SNP genotyping will be resolved and corrected by the programs already in place. After the quality of the SNP data is assured, the SNP data per subject will be stored in the network server for further analysis. These data will not have identifiable information.

We should note that the significance level used to examine genotyping errors or subpopulations discussed below in Section 5.7.3.5 does not involve the same multiple comparison issue as in the hypothesis testing. The reason is that our objective is to identify potential factors that may affect our genetic findings, and any false indication of errors or violations of assumptions will unlikely alter our final conclusions.

5.7.2 Sample Size and Power

Since SPTB is a complex disease and there is little understanding of genetic mechanisms, we have adopted a conservative approach for power estimation, namely the Bonferonni method to correct for genome wide type I error. In addition, we have utilized different analytical strategies (described in depth below) and we have assessed the implication of statistical power in our study.

A Genome Wide Scan with Selected SNPs: The obvious problem with performing a test for each of the 500K SNPs is the large number of tests. The other problem is the potentially limited information in a single SNP. Thus, it is natural and necessary to select SNPs and consider haplotypes.

One approach for selecting SNPs is a two-stage strategy (Satagopan, Verbel et al. 2002). Specifically, we will use 500 cases and 500 controls to scan the entire 500K SNPs in the interim

analysis, followed by the analysis of the entire sample of 1000 cases and controls. However, this strategy is not as powerful as a single stage scan.

Selection of TagSNPs is another approach to reduce the number of SNPs for analysis. Using strong linkage disequilibrium in the dense SNPs may dramatically reduce the number of SNPs and hence the number of significance tests (Gabriel, Schaffner et al. 2002; Daly, Rioux et al. 2001; Patil, Berno et al. 2001; and Zhang, Qin et al. 2005). We will discuss below some of the existing approaches to selecting TagSNPs.

There is a general consensus that, if done appropriately, haplotype based analysis is more effective than single SNP based analysis. As we will also discuss shortly, we will conduct haplotype based association analysis. Since we expect the number of haplotypes examined to be less than the number of SNPs, our sample has much greater power (>80%) to assess the haplotypes if we assume similar relative risks and haplotype frequencies as we assume for the single SNPs. In reality, haplotype relative risks and frequencies are higher than those for a single SNP, which only improves the power, underscoring the rationale for haplotype based association analysis. Furthermore, Morris (2005) reported that a sample of 1000 cases and 1000 controls is adequate to identify associations for moderate genotype relative risk (GRR) of about 1.5 in a variety of realistic settings; see Morris (2005) for more technical details. Here, GRR is the fold increase in risk for having a disease over the general population due to having a disease allele genotype.

Our analysis will proceed in several steps. We will first examine a single SNP based association, next a single gene based association, followed by a haplotype based association analysis based on TagSNPs, and finally gene-gene interactions within a biological pathway.

Figure 1 displays the required sample size as a function of the risk genotype frequency to detect a specific odds ratio (OR). Figure 2 displays the required sample size as a function of the risk genotype frequency to detect a specific level of interaction between two SNP genotypes. A variety of the interaction effects, $\theta = \frac{OR(1,1)}{OR(1,0)OR(0,1)}$, are considered, where the 2nd SNP genotype frequency is set at 0.2. Figure 3 is similar to Figure 2 except that θ is set at 8.

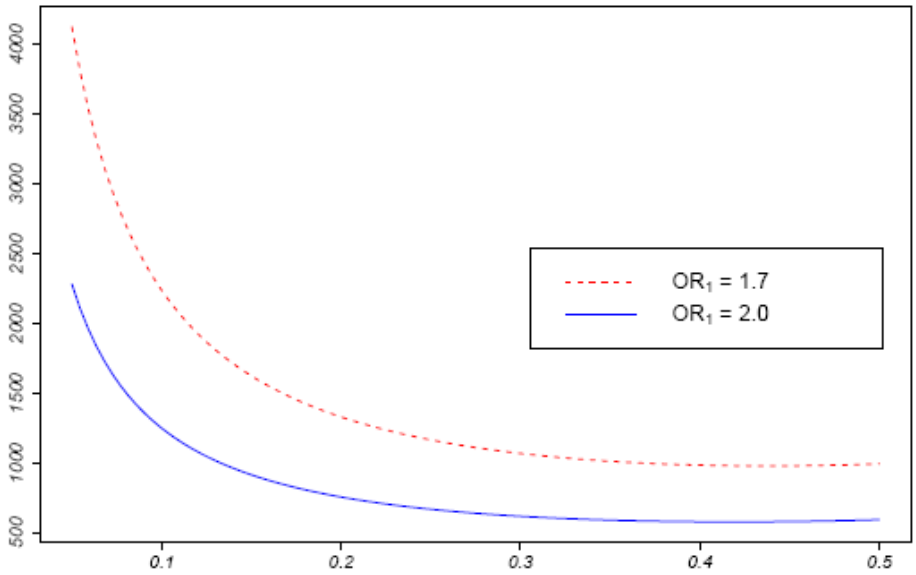

Figure 1. The required sample size is plotted against the genotype frequency of a single SNP in order to detect a specified OR (Odds Ratio) (1.7 or 2). The risk in the population is assumed to be 2%. The type I error is controlled at 0.0001 and power at 80%.

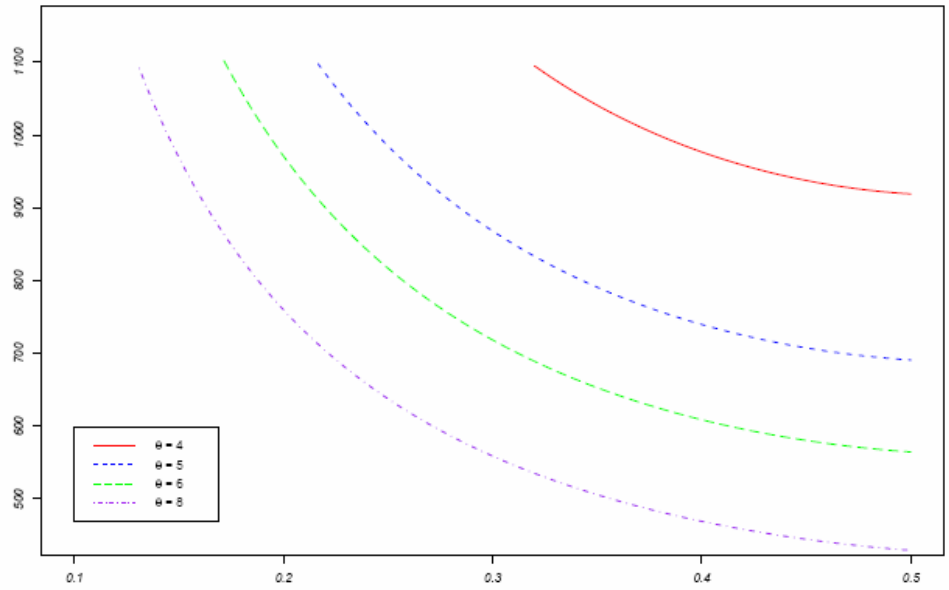

Figure 2. The required sample size is plotted against the genotype frequency of the 1st SNP in order to detect a specified level of interaction between two SNP genotypes. ORs of 1st and 2nd SNP, as the main effect terms, are 1.2 and 1.5, respectively, and the 2nd SNP genotype frequency is set at 0.2. The type I error is controlled at 0.0001 and power at 80%.

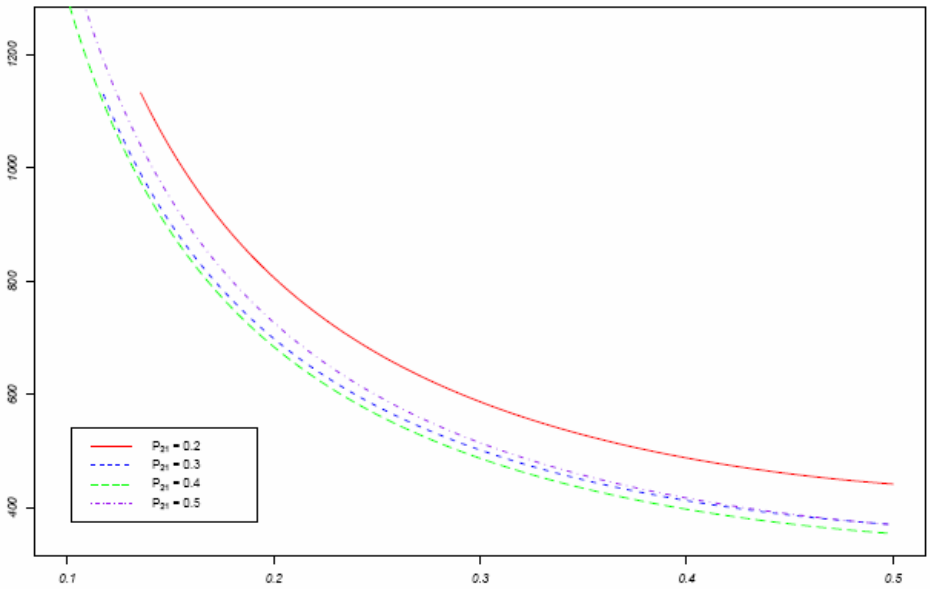

Figure 3. The required sample size is plotted against the genotype frequency of the 1st SNP in order to detect an interaction effect, θ, of 8 between two SNP genotypes. ORs of 1st and 2nd SNP, as the main effect terms, are 1.5, and the 2nd SNP genotype frequency varies from 0.2 to 0.5. The type I error is controlled at 0.0001 and power at 80%.

5.7.3 SNP-based Analysis

5.7.3.1 Testing a single SNP based association.

The analysis can be done through the standard logistic regression model by defining two dummy variables for three possible genotypes that produces the estimates for odds ratios (OR) and their confidence levels. This analysis is applicable regardless of the number of SNPs to be tested, although the overall p-value must reflect the number of tests.

Although we took the most conservative and simple approach - Bonferonni correction – in our power estimation for performing genome wide tests of significance, we will use potentially more powerful strategies in our actual analysis to take advantage of the advanced statistical techniques (permutation). For example, we will determine the statistical significance empirically through permutation testing. The permutation is done by randomly and repeatedly permuting SPTB status in the cases and controls so that the disease status is expected not to be associated with any SNPs, except by chance. In other words, on the basis of our observed data, we will generate the data under the null hypothesis, which allow us to empirically obtain the distribution of the testing statistic under the null hypothesis and hence produce a genome-wide measure of significance. Another analytical approach for dealing with a large number of significance tests is the use of the false discovery rate (FDR). Instead of controlling the genome wide type I error, FDR focuses on the proportion of false positive results in the set of rejected null hypotheses (Benjamini and Hochberg 1995). The same permutation procedure can be used to estimate the p-values for all markers. After sorting all p-values, a cut-off for the p-values is determined to achieve a given level of FDR such as 5%.

Another way to increase the power of our association test is to use and examine haplotype blocks by identifying TagSNPs. This will reduce the number of SNPs to be tested and retain as much information as possible. To this end, we will employ existing methods to identify haplotype blocks and TagSNPs. Since many markers are in strong linkage disequilibrium (Daly, Rioux et al. 2001; Fallin, Cohen et al. 2001; and Patil, Berno et al. 2001), haplotype blocks can be constructed based on marker-marker linkage disequilibrium estimates (Abecasis and Cookson 2000) and TagSNPs can be selected from the haplotype blocks to represent the polymorphisms within a block. The number of tagSNPs will depend on LD patterns.

5.7.3.2 Haplotype based association

Because SNPs are di-allelic, the information in an individual SNP may be limited. The informativeness of the markers, and hence the power of association tests, can be increased by using haplotypes of several SNPs. Furthermore, because each new allele is associated with its own chromosomal history, haplotype-based analyses are warranted to detect unique chromosomal segments that harbor disease-predisposing alleles (Fallin, Cohen et al. 2001). We will examine differences in estimated haplotype frequencies among cases and controls. Before the differences can be compared, we will construct the haplotypes and estimate their frequencies because haplotypes are not directly observable. While this is still an active research area including the development of molecular haplotyping methods, there are many published approaches for determining haplotypes and haplotype blocks as reviewed by Niu (2004). A general approach for choosing haplotpye is to identify tag SNPs that are in strong LD.

For a small number of specified SNPs (4 to 10), a population-based approach can be used to find the haplotypes that maximize the likelihood function through the expectation-maximization (EM) algorithm under the assumption of Hardy-Weinberg equilibrium (HWE) (Excoffier and Slatkin 1995). Despite the HWE assumption, it has been noted that the method is not too sensitive to HWE (Niu 2004). If we use a large number of SNPs to construct haplotypes, the partition-ligation EM algorithm can be used. HWE will be validated in our control population before we apply several approaches as described by Niu (2004). Now, how do we choose a particular set of SNPs for haplotype analysis? One approach is to focus on certain genomic regions or candidate genes and choose, say, 8 tag SNPs from each region or candidate gene. Another approach is to use moving windows. For example, Jawaheer and colleagues used a three-SNP moving window to detect the association between human leukocyte antigens and rheumatoid arthritis (Jawaheer, Li et al. 2002) .

As described above, the haplotypes can be constructed with well-chosen tagSNPs or through moving windows of a few adjacent SNPs. Once the haplotypes and their frequencies are estimated using methods such as PHASE (Stephens, Smith et al. 2001), they can be treated as predictors in a logistic regression model (Zhao, Li et al. 2003 and Schaid, 2004). We will also use the method proposed by Fallin, Cohen et al. (2001) to compare the distribution profiles of haplotypes between cases and controls. Without considering covariates, the data become a 2 x k table, where k is the number of estimated haplotypes. Since we have to determine the tagSNPs and their haplotypes (and consequently the frequencies of the haplotypes) after we collect the data, it is premature to consider specific individual haplotypes. Broadly speaking, however, due to the haplotype uncertainties, the power calculation depends on many unknown quantities. Even though the number of haplotypes to be considered is relatively large, it is generally believed that

the use of haplotypes increases the power over the use of individual SNPs. A common sense approach to gauging the power of the haplotype analysis is by collapsing haplotypes into two major types (a wild type/a risk variant), and then the locus of interest would essentially become a di-allelic locus.

After the SNP-based and haplotype-based associations are performed, it is important to achieve a certain synergy in the two types of analysis. It is generally agreed that no association testing in epidemiological studies alone can distinguish between the true-positive and false-positive signals obtained in a multistage genomewide scan. Approaches that have been suggested include comparative sequence analysis (Sidow 2002 and Bejerano, Pheasant et al. 2004), linkage analysis of expression data (Morley, Molony et al. 2004), or computational approaches to predicting function (Ng and Henikoff 2003; Livingston, von Niederhausern et al. 2004; Xi, Jones et al. 2004; and Zhu, Spitz et al. 2004), before launching into the labor-intensive tests (Thomas, Haile et al. 2005). The other two aims of this network will use expression profiling technologies to further examine our findings.

5.7.3.3 Testing gene-gene interactions

Since we consider a number of genes and each gene has several haplotypes, there are usually a large number of potential gene-gene interactions and pathways. For this reason, this stage of analysis is exploratory and we do not provide specific power analysis. However, the work of Gauderman (2002) assures that our study sample will have adequate power for this type of analysis. According to Table 1 in their work, a sample of 500 cases and controls leads to 80% of power at the significance level of 0.05 to detect a gene-gene interaction of odds ratio of about 2.2 that measures the departure from a purely multiplicative model when they considered two asthma genes, *GSTM1* and *GSTT1*, as examples. After gene-gene interactions are examined, pathways can be identified according to the genes that interact with each other, and then we can use our DNA samples to further understand the functions of those genes that may underlie the cause for preterm.

Identifying gene-gene interactions are important and challenging. Marchini, Donnelly et al. (2005) examined the power of three strategies for analyzing genome-wide association studies: strategy I, locus-by-locus search (requiring at least locus meeting the significant criterion); strategy II, search over all pairs of loci; and strategy III, a two-stage strategy in which all loci meeting some low threshold in a single-locus search are subsequently examined for a significant full model fit. They considered 300K markers, 2000 cases and controls, and three multi-locus disease models. They noted that there are many configurations in which the interaction-based search strategies are more powerful than searching locus-by-locus. We will also use these strategies and will especially examine those interactions revealed by all strategies. Even though we may miss some positive results, given the explorative nature of these analyses, a conservative approach is appropriate.

5.7.3.4 Testing gene-environment interactions

The number of environment covariates to be considered is considerably fewer than the number of SNPs. We can use a modified strategy based on strategy III investigated by Marchini, Donnelly et al. (2005). That is, we begin with all genes meeting some low threshold in a single-gene search, and then pair them with the environment factors to assess their interactions.

5.7.3.5 Population Substructures

It is well-known that population substructure, sometimes referred to as cryptic substructure, can provide spurious results for case-control association tests. For example, association studies of type II diabetes in Pima Indians (who have high rates of diabetes) were flawed because Pima individuals with a high degree of Caucasian ancestry had lower diabetes susceptibility. Thus, any marker loci that were at higher frequency in the Pima than in Caucasians were "associated" with the disease (Gauderman 2002; Lander and Schork 1994; Hanson and Knowler 1998; and Pritchard and Rosenberg 1999). There have been many discussions on this issue. Some researchers believe that this phenomenon might have been over-stated (Spence, Greenberg et al. 2003). Further, Risch (2000) pointed out that population stratification has been invoked numerous times as the cause for an observed high false-positive rate in association studies using candidate genes, yet it has rarely been demonstrated as the culprit. In a relatively recent extensive simulation study, Setakis, Stirnadel et al. (2005) concluded that explicit allowance for cryptic substructure may often be unnecessary provided that good study design principles have been used so that case and control populations are similar. Since our cases and controls are reasonably matched, we expect the two populations to be similar. However, those authors also pointed out that methods that do protect against cryptic substructure typically perform well in limiting the number of false positives, and the cost of this protection, in terms of lost power, is often small. Thus, we will use appropriate methods to consider cryptic substructure. Several methods of assessing population stratification have been proposed using unlinked markers. A classic statistic for detecting cryptic substructure is Wright's FST (Wright 1951), which is estimated as a single value that summarizes the average deviation of a collection of populations away from the mean. While there are a number of methods for adjusting associations for substructure and admixture, unlinked markers are used to adjust associations and the methods may be broadly divided into model based and non-model based approaches (Satten and Epstein 2004). Genomic control (Devlin and Roeder 1999) is a non-model based approach that essentially corrects asymptotic distribution of the classic Armitage trend test statistic by an over-dispersion factor, which is estimated from the empirical distribution of the trend statistic at a given number of null markers.

Alternative methods such as those implemented in the program called STRUCTURE use model-based approaches to determine the underlying population structure (Pinkel, Segraves et al. 1998). According to Setakis, Stirnadel et al. (2005), none of these methods is uniformly superior to the others, nor is any one method uniformly inferior in the presence of the population structure. Nonetheless, if population substructures are evident in our data as suggested by STRUCTURE, we will adjust the population substructure in our association test. The use of genomic control is simple but it can only be applied to SNPs. Generally, we need to use latent variables or mixture models that estimate the number of underlying subpopulations and the probability for an individual marker to be originated from each subpopulation, and then use mixed effects logistic models as described in Satten and Epstein (2004). The advantage of this approach is that it is applicable for both SNPs and haplotypes. The computation for genomic control and other methods to deal with population structures (such as clustering) can be done by Shaun Purcell's PLINK program (http://pngu.mgh.harvard.edu/~purcell/plink).

We should note that our objective is not to test the hypothesis that there are subpopulation structures. Instead, our principle is to examine whether subpopulations may compromise our genetic findings and how we should take them into account. In the worst scenario, if there are

subpopulation structures, and genes are distinct in different subpopulations, we would have to stratify our samples and the power would be reduced.

5.7.3.6 Exploratory Data Analysis

This study is expected to collect rich and important genetic and clinical data. In addition to the hypothesis testing and regression models as described above, we will use other contemporary approaches such as tree-based analysis to take advantage of all information and to simultaneously examine multiple SNPs as well as haplotypes. Dr. Zhang has published extensively in this area and we will perform the respective types of exploratory analyses to strengthen the validity of our findings (Zhang and Singer 1999 and Zhang and Bonney 2000).

5.7.4 DNA Copy Number Changes

In addition to SNP based analysis, we will also examine genomic DNA copy number changes. It has long been known that DNA copy number changes play an important role in the development of human cancers. More recently, it is now recognized that DNA copy number is also a common structural variation or polymorphism in the human genome and that differences between individuals may be involved in disease susceptibility (Feuck, Marshall, et al. 2006). One of the challenges is to design efficient molecular methods that can identify DNA copy number variations locus-by-locus across the whole genome. The basic technique for this objective is Comparative Genomic Hybridization (CGH), that is, arraying a set of probes and hybridizing with genomes from cases ("test") and controls ("reference") cells, and then doing comparisons. The difference of the existing methods in the literature mainly lies in the different choices of probe sets. Pinkel, Segraves et al. (1998) proposed to use arrays of bacterial artificial chromosome (BAC) DNAs or other large insert clones as hybridization probes; another report on use of BACs can be found in Snijders, Nowak et al. (2001). Pollack, Perou et al. (1999) utilized fragment cDNAs as probes. Barrett, Scheffer et al. (2004) recommended using oligonucleotide-array as probes, and they used a set of 60-mer oligonuleotides in their implementation. Lucito, Healy et al. (2003) chose a set of representative 70-mer oligonucleotide probes and developed a methodology called ROMA (representational oligonucleotide microarray analysis). Huang, Wei et al. (2006) used a high-density oligonucleotide array-based SNP genotyping method to find the changes of DNA copy number and 25-mer probes were used in their research. Adam, Qu et al. (2002) developed another algorithm called CARAT (copy number analysis with regression and tree) that used probe intensity information to infer copy number variations, and the probe set were originally designed to genotype 100,000 SNPs. Software such as DChip (Li and Wong 2001 and Zhao, Weir et al. 2005) have been developed to analyze CGH data. Both the technology and the software are still evolving. For example, genomic profiling of loss of heterozygosity (LOH) and DNA copy number is possible with Illumina Infinium® Whole-Genome Genotyping (http://www.illumina.com/General/pdf/CGHApplicationNote.pdf). Affymetrix is developing an optimized 500K SNP array for determining both SNP genotypes and DNA copy number variation. Presently, we favor the use of the SNP approach to determining DNA copy number. This approach will allow both SNP genotyping and copy number variation analyses to be obtained simultaneously for a sample on the same chip. This will result in a significant cost saving, without which DNA copy number analysis would not be feasible given the Network's budget constraints.

5.7.5 Multiple Comparisons

We have discussed the multiple comparison issue above. We should also note that the multiple testing issue arises in different settings for this Network including multiple markers, multiple phenotypes, selection of environmental exposures, and various choices of models. We will adjust for the multiple testing on a case-by-case basis. For example, although Bonferroni correction is the most conservative approach, we will consider it as the first approach.

5.8 Human Subjects - Template Consent Form

1.0 **You are invited to take part in an important research study about spontaneous preterm birth.**

Spontaneous (unplanned) preterm birth is the spontaneous delivery of a baby more than three weeks before its due date. Researchers from five areas in the United States are trying to learn about the causes of spontaneous preterm birth. The researchers hope that this study will help prevent spontaneous preterm births in the future. This research study is funded by the National Institutes of Health (NIH) / National Institute of Child Health and Human Development (NICHD). It is called the Genomic and Proteomic Network for Preterm Birth Research. Approximately 2,500 pregnant women from across the nation will take part in the study.

You are being asked to take part in this study because you have just delivered within the past day a baby either at full term (between 1 week before your due date and 2 weeks after your due date) or at less than 34 weeks gestation (more than six weeks before your due date). This part of the study will recruit 2000 such women from across the nation. We would like to tell you about a study that hopes to find reasons why spontaneous preterm birth occurs.

This study will use several powerful new research methods to study the problem of spontaneous preterm birth. Genomics is the study of specific genes that are associated with any given medical problem, in this case, spontaneous preterm birth. Genes are passed on from our parents and are the building blocks of our bodies that tell our bodies how to work. Proteomics is the study of proteins (produced by specific genes) that are seen with a given medical problem, in this case, spontaneous preterm birth. Metabolomics is the study of small molecules that are associated with a given medical problem, in this case, spontaneous preterm birth."

What we learn from this study could give us valuable information about this common and serious problem that could lead to more effective treatment and/or prevention.

2.0 **What will happen if I join this study?** For this study of spontaneous preterm birth, we would:

a) ask you some questions
b) collect facts from your medical records and those of your baby
c) obtain your blood or saliva taken within a day after you deliver
d) obtain a sample of blood from the umbilical cord after delivery, and,
e) obtain a cheek swab from your baby after birth

Genomic, proteomic and metabolomic testing would be performed on the samples that are collected and the findings would be compared with the outcomes of the women and babies enrolled in the study. All information will be kept strictly confidential. More details on how we will protect your privacy are given in questions 10.0 and 11.0 below.

2.1 **What kinds of questions will I be asked?** A trained member of our research team will ask you some questions about your pregnancies, your medical history, your family history, and your day-to-day life and work, including smoking, drinking, and drug use. The whole interview can be done while you are in the hospital. The initial interviews should take about 30 to 45 minutes.

2.2 **How will my medical records be used?** We wish to look at the medical records of your pregnancies, including records of your doctor and from your hospital visits. We wish to collect information from your records to help us better understand spontaneous labor, both preterm and full term. This request to access your records (and those of your baby) will comply with HIPAA, which is a law protecting the privacy of your health information.

2.3 **How will my baby's medical records be used?** We would like to collect information from your baby's medical records from birth to the time that your baby leaves the hospital. We will use this information to look at differences between babies that are born pretermly and those that are born full term (within three weeks of the due date).

2.4 **How will my and my baby's blood and saliva samples be used for this study?** In addition to your usual blood tests, an additional 10 milliliters (about two teaspoons) of blood would be drawn from a vein in your arm while you are in the hospital. If possible, this will be done at a time when you are having another blood sample drawn as a part of your care. Your baby's blood will be obtained after delivery from the discarded umbilical cord. We will only save your baby's blood if you agree; otherwise, it will be destroyed. All collected samples will be sent to a specific research laboratory at the University of Pennsylvania (Philadelphia, PA), where specialized laboratory testing will be done that might predict whether or not spontaneous preterm birth will, or will not, occur.

If a blood sample cannot be obtained from you or your baby, we would ask your permission to obtain a sample of saliva from you or your baby. Both blood and saliva are a very good source of DNA (deoxyribonucleic acid), which is the substance that contains our genes. Most genes do not differ among people, but some genes and combinations of genes are unique to each person. If you agree, DNA from you and your baby will be used to see whether the genes of mothers and babies with and without spontaneous preterm birth are different. We will store a sample of your and your baby's DNA. These samples will only be used to study spontaneous preterm birth, and pregnancy or baby-related diseases. These samples would also be sent to the University of Pennsylvania, where DNA would be studied in hopes of finding specific genes that might predict whether or not spontaneous preterm birth will, or will not, occur.

2.5 **How will my samples be stored?** All blood and saliva samples will be stored in a secure facility at the University of Pennsylvania for five years and then in a secured facility (to be determined) until the samples are no longer useable or until destroyed. To protect you and your baby, samples will only be labeled with a study identification number (study ID). We will keep your study ID linked with your name in our local

records (locked and secure) for up to ten years after the end of the study. If you decide that you want the stored samples destroyed while we have this link, you can let us know in writing and we will have the samples that are labeled with your study ID destroyed. The stored samples will only be released for other studies if an independent committee that reviews research on people (called an Institutional Review Board) has approved the research. Research findings on stored samples will not be reported back to you. Stored samples will only be used for the study of spontaneous preterm birth, and pregnancy or baby-related diseases unless you are contacted for your permission to use the samples for other purposes (see question 2.6 below).

2.6 **Could I be contacted for future studies?** A research person authorized by this study may contact you in the future if there are any new research questions or to follow up on the health and well-being of your baby and you. If you ask us not to contact you for future studies, your last contact with us will be at the time that you and your baby are discharged from the hospital.

2.7 **Could information from my baby and me be used in other future studies?** The results of your and your baby's research tests (including DNA tests) and related medical records will be placed in an electronic database that could be used for other research studies in addition to this study. These studies would involve only pregnancy or baby-related diseases. Your privacy will be strictly maintained, since information that would identify you or your baby would not be available. However, in certain situations a researcher might want to contact you for additional information that would be useful only for other research purposes. In these cases, the researcher would make a request to an oversight committee. This committee would review the request and either deny or approve the researchers' request for your contact information. The researchers would have to guarantee that they would not release either your contact information or your database records before permission would be granted. This would only apply if you have given prior permission to be contacted in the future by a researcher. If you have not given your prior permission, the request would be automatically denied.

3.0 Risks and Benefits from Study Participation

3.1 What discomforts or risks could I have from participating in the study?

It is possible that some of the interview questions could make you uncomfortable or upset. You do not have to answer any questions that you do not want to answer. You may stop the interview at any time without affecting your medical care. Occasionally drawing blood may cause pain and bruising. Rarely, drawing blood also causes some people to faint or feel light-headed and causes blood clots or infection. Research staff members are available to help you get medical care if you need it because of the study.

If you agree to be in this study we would be obtaining information from blood or saliva samples about whether you and your baby have been exposed to different kinds of drugs. We will not provide the results of these research tests to any law enforcement or social agency. We have obtained a Certificate of Confidentiality from the federal government so that we cannot be forced to release this information to anyone (see question 11.0 below).

Laboratory tests, including DNA testing, carry some unique risks. The concern is that test results might someday be used against a person. However, the risk for something like that

happening as a result of being in this study is extremely small. First of all, the researchers will be very careful to ensure that only authorized people have access to the samples and your personal information. Secondly, the types of tests we plan to do cannot, at the current time, be directly linked to an increased risk of a disease, for example cancer or diabetes. In addition, the results will not be placed in your medical record. You will not be given any information at all about the results of the tests we are doing, because even if we are successful in identifying protein and metabolites, genes or other types of chemicals, we will still not have enough information to directly benefit you or your family. We will only use the tests from you and your baby to study spontaneous preterm birth and pregnancy or baby-related diseases unless you give us separate permission to study other diseases.

There may be risks while participating in this study that are unknown.

<<LOCAL SITES MAY ADD INFORMATION REGARDING LIABILITY, ETC., PER LOCAL IRB REQUIREMENT HERE. >>

3.2 **Will I benefit from joining the study?** There are likely no direct benefits to you from participating in this study. However, we hope to learn how to reduce the likelihood of spontaneous preterm birth. You could have the satisfaction of knowing that you have given information that may help prevent spontaneous preterm births for other pregnant women in the future. General information from this research will be published in medical journals, but will not be relayed to you or benefit you directly.

4.0 **Do I have to join this study? What are my other choices?** You do not have to be in this study. You may choose to participate in some, but not all, parts of the study. Your medical care will not be affected if you choose not to take part in this study; you will still be able to get all appropriate medical care without being in the study. Your choice about participating in this study will not affect your eligibility for any health plan or any health plan benefits or payments.

5.0 **Withdrawing from the Study**

5.1 **What if I change my mind about how I want to take part in the study?** Even after you agree to be in this study, you can change your mind at any time and pull out of the study. You can refuse to answer a question or refuse to take part in any part of the study. If you change your mind, your medical care and health plan eligibility and benefits will not change. At any time you can also request that your and your baby's samples be destroyed.

5.2 **Can the researchers remove me from the study?** You may withdraw from the study at any time without penalty. Likewise, Dr. XXXXX and HIS/HER associates or the National Institutes of Health can withdraw you without your approval. A possible reason for withdrawal could be the early termination of the study by the National Institutes of Health.

6.0 **If I have expenses because I took part in this study, can the study help pay these costs?** If you agree to take part in this study, usual travel and child care expenses (<<UP TO $??>>) that result from your being in this study will be paid. <<ADDITIONALLY, YOU WILL RECEIVE $?? AS COMPENSATION FOR THE TIME YOU SPEND ANSWERING QUESTIONS. (SENTENCE ONLY NEEDED IF REIMBURSEMENT

FOR TIME IS OFFERED AT THE SITE. SOME SITES MAY SPLIT PAYMENT BETWEEN THE TWO PARTS OF THE INTERVIEW. SUCH DETAILS WILL BE GIVEN HERE. SEE PROTOCOL FOR GUIDELINES.)>>

7.0 **Will joining this study increase my health care costs?** No. Your routine medical care and routine testing will be billed to your insurance company or to you the same as your other medical care. The costs of tests that are performed only for purposes of the research study will not be charged to you or your insurance company.

8.0 **What will happen if I get a bill for research tests by mistake?** If you receive a bill for any study-specific test that is not paid by your insurance company or Medicaid or if you are not sure if the bill is for a study-specific test, call the study team at <<PHONE NUMBER>> and they will ask you to send a copy of the bill to the following address to find out if this test was done for the study. If it was done because of the study, the study will pay for it.

<<NAME AND ADDRESS OF PI>>

9.0 **What do I do if I have questions or need to report a problem?** If you have questions now, feel free to ask them to the study staff. If you have more questions later, please contact

<<NAME AND CONTACT INFORMATION OF APPROPRIATE FIRST CONTACT>>.

If he/she is not available, you may contact <<NAME AND CONTACT INFORMATION OF APPROPRIATE SECOND CONTACT>>. <<PROVIDE INSTRUCTIONS FOR AFTER-HOUR CONTACTS AT YOUR INSTITUTION>>

The committee that reviews research on human subjects (Institutional Review Board) at <<NAME OF INSTITUTION>> will answer any questions about your rights as a research subject. They can be contacted at <IRB CONTACT NUMBER (AND CONTACT PERSON IF AVAILABLE)>>.

10.0 **How will my privacy be protected?** The researchers are taking many steps to protect the privacy of you and your baby. First, we will use a number to identify research records and tissue and blood samples, instead of your name. If you agree to take part in the study, a study ID number will be assigned to you and your baby. This same number will be used to identify medical information kept by the study. Any records linking your name to the study ID will be stored in a locked (password protected) file.

All copies of your medical records and the links between your name and the study ID will be destroyed within ten years of the end of the study. After that time, your blood or tissue samples and information from your interview and medical records collected for the study will be marked with your study ID number only. Even though the information could still be used for research, it will no longer be possible to link it to you. Any results of this study reported in medical journals or at meetings will not identify you or your baby.

Only those who work with this study will be allowed access to your information. However, representatives from the National Institutes of Health may inspect and/or copy the records that identify you.

Results of the study may be published. However, your name and other identifying information will be kept private.

11.0 **What is a Certificate of Confidentiality?** To help us further protect the privacy of you and your baby, the researchers have obtained a Certificate of Confidentiality from the National Institutes of Health (NIH) / National Institute of Child Health and Human Development (NICHD). This Certificate means that the researchers cannot be forced (for example by court order) to disclose any information that might identify you to any federal, state, or local court. A Certificate of Confidentiality does not prevent you from voluntarily giving information to others about yourself or your involvement in this research. You should know that we may provide information to your health care providers if we suspect that you may harm yourself or others. We will not release any information collected as part of the research regarding use of illicit drugs and testing for drugs done on samples collected for the research.

12.0 **If I agree to be in all parts of this study, what does that include?**
Full participation in this study includes all of the following:

- An interview
- Collection, review and storage of medical records
- Collection of your and your baby's blood and saliva samples that were obtained within a day of your delivery
- Laboratory tests on your and your baby's blood and/or saliva, including DNA tests.
- Storage and future use of blood or saliva samples, including DNA.
- Research staff possibly contacting you later for other research questions.

13.0 **Not Participating in a Part of the Study**

13.1 **If I still want to join this study but would like to avoid some parts of it, what should I do?** Agreeing to do all parts of this study will allow the researchers to gather the most useful information to study spontaneous preterm birth. However, if you do not want to do all parts of the study, you can refuse to do any one or more parts of the study.

Parts of this study that you do not wish to do are:
None ☐ Restrictions, if any: _____

13.2 **Follow Up of you and your baby after you deliver and go home**
Please indicate your preference about being contacted after you and your baby have gone home:

☐ Yes, I give my permission for researchers to contact me after my baby and I go home. I understand that this contact will be limited to follow up of our health and well-being.

☐ No, I do not want to be contacted after my baby and I go home.

14.0 **Will my information and samples be used in other studies of pregnancy complications?** The researchers may want to study other pregnancy complications using the information and samples that have been collected in this study. This could only happen if you give the researchers permission. You do not have to give them your permission but we would like you to consider initialing one of the three boxes if you are comfortable doing so. If you do not initial a box giving the researchers this permission, your samples will not be used.

☐ I give my permission for my and my baby's samples to be used in future studies of pregnancy and baby complications. I understand that my baby and I will not be identified in any way.

☐ I may give my permission for my and my baby's samples to be used in a non-identifiable fashion in future studies but wish to be contacted about the future study first.

☐ I do not give permission for my and my baby's samples to be used in the future and request that they be destroyed when the current study is completed.

15.0 **SIGN THIS FORM ONLY IF ALL OF THE FOLLOWING ARE TRUE:**
You have read the above information.

Your questions have been answered to your satisfaction and you believe you understand all of the information given about this study.

You have freely decided to take part in this research study.

You agree to the storage and future use of blood, saliva, and DNA samples or you have indicated restrictions on the storage and future use on the previous page.

Subject:	Other:
_____ Signature of Subject	_____ Signature of Witness
__ __ /__ __ / __ __ __ __ __ __ :__ __ Month Day Year Hour Minute (24-hr clock)	_____ Signature of Person Obtaining Consent
	_____ Printed Name of Person Obtaining Consent
	_____ Printed Title of Person Obtaining Consent
	_____ Signature of Interpreter (if applicable)

YOU WILL BE GIVEN A COPY OF THIS SIGNED CONSENT FORM.

6 Longitudinal Cohort Study

6.1 Definitions

For the longitudinal cohort study, cases (preterm delivery) are defined as a delivery between 20 weeks 0 days and 33 weeks 6 days (inclusive) following the spontaneous onset of labor (see definition Section 2.5.7) either in the presence or absence of pPROM (see definition Section 2.5.6), late preterm controls are defined as a delivery between 34 weeks 0 days and 36 weeks 6 days (inclusive) following the spontaneous onset of labor (see definition Section 2.5.7) either in the presence or absence of pPROM (see definition Section 2.5.6), controls (term delivery) are defined as a delivery between 39 weeks 0 days and 41 weeks 6 days following the spontaneous onset of labor either in the presence or absence of PROM (see definition Section 2.5.5).

6.2 Eligibility Criteria

The Steering Committee and Protocol Subcommittee devoted substantial effort and discussions to determine the eligibility (inclusion and exclusion) criteria for this study. The overall rationale is to maintain scientific rigor in order to construct a study cohort in which the disease group will have the highest chance to share a common basis that differ from the controls, and to maximize the feasibility by enhancing the study population.

6.3 Inclusion Criteria

1. Prior SPTB between 20 weeks 0 days and 36 weeks 6 days (inclusive), including pPROM (see definition 2.5.6) if spontaneous onset of labor occurred, regardless if the delivery was vaginal or cesarean section.
2. Current project gestational age 10 weeks 0 days to 18 weeks 6 days (Inclusive).

6.4 Exclusion Criteria

1. Maternal uterine anomalies.
2. Multi-fetal gestation.
3. Known aneuploidy or lethal fetal anomalies (see Section 2.5.8 for specifications).
4. Polyhydramnios (defined as Amniotic Fluid Index >=25 cm or Deepest vertical Pocket >=12 cm).
5. Planned or probable delivery at a non-Network site.
6. Non-availability for prospective specimen/data collection.
7. Serious maternal medical conditions defined in Section 4.3.
8. Cervical cerclage.

6.5 Screening for Eligibility

All patients with a history of prior SPTB between 20 weeks 0 days and 36 weeks 6 days who present in pregnancy between 10 weeks 0 days to 18 weeks 6 days of gestational age are eligible for screening. The inclusion and exclusion criteria will be reviewed with the patient's chart. If an ultrasound examination has not been performed, one must be arranged prior to enrollment. Gestational age determination and assignment of the Project Gestational Age will be then

performed based on the results of the ultrasound examination and patient's menstrual history according to criteria defined in Section 4.2 "Gestational Age Determination".

If a patient meets the criteria for enrollment and expresses interest to participate in the study, she will be told about the study's design, its potential risks and benefits and asked to sign the informed consent and patient authorization (if applicable) forms. If she accepts, she will then be enrolled in the study. A copy of the signed consent form(s) will be provided to the patient.

6.6 Study Procedures

6.6.1 Initial Visit (10 weeks 0 days – 18 weeks 6 days)

6.6.1.1 Patient data

Patient data will be collected through chart reviews and patient interviews. Data will be entered using the Demographics, Enrollment, Intake Forms (I, II, III), Current Medications, and Ultrasound and Labs forms as enclosed in the Manual of Operations. Stress, anxiety, and depression during the pregnancy will be assessed by the Psychosocial Questionnaire. In addition, the Adverse Event Form will be used as needed and is unscheduled. The adverse event form will only be filled out for neonates up to 28 days post delivery (during the neonatal period) who experience an adverse event as defined in the Manual of Operations.

1. Demographic information: parental age, race, ethnicity, etc.
2. Medical history: pre-pregnancy weight, height, medications, STD history, medical conditions, etc.
3. Social history: marital status, years of education, alcohol use and tobacco use.
4. Obstetrical history including outcome of all prior pregnancies and dates of termination.
5. Maternal family history of SPTB.

6.6.1.2 Samples

Samples will be sent from Clinical Cores to the Analytical Core, and data will be sent from the Analytical Core to the DMSI Core.

1. Maternal blood (10 ml) will be collected to obtain serum for proteomic and metabolomic analyses and measurements of markers of pregnancy exposures (cotinine, heavy metals, folic acid, stress hormones, etc.) and whole blood for DNA for genotyping will also be collected.
2. Maternal saliva (2ml) will be collected for proteomic and metabolomic analyses and measurements of markers of pregnancy exposures (cotinine, heavy metals, folic acid, etc.) and DNA for genotyping as a back up if maternal blood cannot be obtained.
3. Maternal Urine (2 to 4 ml) will be collected for proteomic and metabolomic analyses.
4. Maternal cervicovaginal fluid (4 Dacron swabs) will be collected from vaginal fornices during initial pelvic examination. Samples collected this way will be used for analysis of RNA expression, microbial DNA, and proteomic/metabolomic analyses, and for a slide for diagnosis of Bacterial Vaginosis.

6.6.2 Return Study Visits

The following is the schedule for each visit. Visits may be scheduled anytime within the prescribed windows, but each visit must occur at a minimum of 4 weeks apart:

- First Return Visit (19 weeks 0 days – 23 weeks 6 days).
- Second Return Visit (28 weeks 0 days – 32 weeks 6 days).
- Admissions for Preterm Labor or pPROM that occur between 20 weeks 0 days and 33 weeks 6 days. Participants who have multiple hospital admissions for PTL or pPROM should undergo data and specimen collection for admissions that are two or more weeks apart.

6.6.2.1 Patient data

Patient data will be collected through chart reviews and patient interviews. Data will be entered using the Return Visit Form, Ultrasound and Labs, and the Current Medication Form as enclosed in Manual of Operations. In addition, the Adverse Event Form and Hospital Admission Form will be used as needed. The adverse event form will only be filled out for neonates up to 28 days post delivery (during the neonatal period) who experience an adverse event as defined in the Manual of Operations.

1. History of preterm labor or premature rupture of membranes symptoms, or evaluations or hospitalizations for those symptoms.
2. New medical or obstetrical complications
3. New medications
4. History of cervical evaluation manually or ultrasonographically or fetal fibronectin measurements.

6.6.2.2 Samples

Samples will be sent from Clinical Cores to the Analytical Core, and data will be sent from the Analytical Core to the DMSI Core.

1. Maternal blood (10 ml at first visit, 5 ml at subsequent visits) will be collected to obtain serum for proteomic and metabolomic analyses and future measurements of markers of pregnancy exposures and to obtain a backup source of DNA and for future analyses.
2. Maternal saliva (2 ml) will be collected for proteomic and metabolomic analyses and future measurements of markers of pregnancy exposures (cotinine, heavy metals, folic acid, etc.) and DNA for genotyping as a back up.
3. Maternal urine (2 to 4 ml) for proteomic and metabolomic analyses.
4. Maternal cervicovaginal fluid (4 Dacron swabs) will be collected from vaginal fornices during initial pelvic examination. Samples collected this way will be used for analysis of RNA expression, microbial DNA, and proteomic and metabolomic analyses, and for a slide for diagnosis of Bacterial Vaginosis.
5. Amniotic fluid will be obtained when available at 15-24 weeks during amniocentesis performed for clinical indications such as 2nd trimester genetic screening. Amniotic fluid will not be collected from an amniocentesis performed later in pregnancy, such as the fetal lung maturing testing. 5 mL will be obtained in addition to clinical sample for RNA expression and proteomic and metabolomic analyses.

6.6.3 Labor and Delivery

Samples and Patient's Data will be collected from:

1. Outcome data will be collected from enrolled women, irrespective of subsequent gestational age and whether or not their PTB was spontaneous or indicated. We will only collect specimens in the following three groups.
2. All spontaneous preterm deliveries between 20 and 33 weeks 6 days (we expect 30 of them based on the cohort of 500 patients with a history of preterm deliveries). *
3. Spontaneous preterm deliveries from 34 weeks 0 days to 36 weeks 6 days (Many are expected to be available, and twice as many as group 1 will be used to 1:2-match with group 1).*
4. Spontaneous term deliveries from 39 weeks 0 days to 41 weeks 6 days (Many are expected to be available, and twice as many as group 1 will be used to 1:2-match with group 1).*

* Includes pPROM or PROM if spontaneous labor follows.

6.6.3.1 Patient Data

Patient data will be collected through chart reviews and patient interviews. Data will be entered using the Labor and Delivery Form and Neonatal Baseline Form. In addition, the Adverse Event Form will be used as needed and is unscheduled.

1. Labor type, indications, membrane status, and induction method.
2. Pregnancy complications.
3. Delivery type, indications for cesarean section.
4. Neonatal outcome: sex, weight, length, Apgar score, neonatal complications admission to intensive care unit, etc.

6.6.3.2 Samples

Samples will be sent from Clinical Cores to the Analytical Core, and data will be sent from the Analytical Core to the DMSI Core.

1. Maternal blood (5 ml) will be collected for proteomic and metabolomic analyses, DNA for genotyping, and measurements of markers of pregnancy exposures (cotinine, heavy metals, folic acid, stress, etc).
2. Maternal saliva (2 ml) will be collected for proteomic and metabolomic analyses, and measurements of markers of pregnancy exposures (cotinine, heavy metals, folic acid, stress hormones, etc.) and DNA for genotyping as a back up.
3. Maternal urine (2 to 4 ml) will be collected for proteomic and metabolomic analyses.
4. Cord blood (10 ml) will be obtained following delivery for fetal genotyping (DNA) and proteomic and metabolomic analyses.
5. Neonatal saliva sample (via Oragene collection) will be obtained if cord blood could not be collected for fetal genotyping.
6. Placental samples will be collected for RNA expression, proteomic and metabolomic analyses, epigenetic analysis and histology.
7. Fetal membranes will be collected for RNA expression, proteomic and metabolomic analyses, epigenetic analysis and histology.

6.6.4 Receiving Samples from the Clinical Cores to the Analytical Cores

1. Barcoding samples will be performed at clinical sites. Samples will be identified (i.e., maternal blood/gestational age, maternal urine/gestational age) with the following assigned numbers at each clinical site:

- Alabama-Birmingham 2,001 – 2,200 series
- Texas-Galveston 8,001 – 8,200 series
- Utah 5,001 – 5,200 series

2. Penn tracking system – caTISSUE is an Oracle-based tissue bank designed by caBIG™. Samples will be linked to TrialDB through an Excel file with barcodes and patient IDs.
3. Shipping instructions (i.e., dry ice, etc.) for individual samples are provided in the Manual of Operations (section 8). All samples will be shipped to University of Pennsylvania.

6.6.5 RNA Acquisition

RNA Extractions: Extractions are performed using Trizol and Qiagen protocols. RNA samples only proceed to target labeling if appropriate size and abundance distributions are seen for ribosomal bands using the Agilent Bioanalyzer, and the RNA has an A260/280 ratio of 1.7-2.1 as measured on the Nanodrop spectrophotometer. One to five micrograms of input total RNA usually result in 60 to 100 ug of cRNA after IVT; this yield is a very good indicator of the quality of the initial RNA and the overall hybridization intensity that will be seen on GeneChips. Similar QC determinations are performed on each round of amplification for RNA samples that start with less than 1ug.

6.6.6 Analysis

1. RNA profiling will be conducted in the Penn Microarray Facility (Donald Baldwin). Protein and metabolite profiling will be conducted in the Penn Proteomics Core (Ian Blair).
2. RNA profiling: All protocols will be conducted as described in the Affymetrix GeneChip Expression Analysis Technical Manual. Briefly, 1 ug of total RNA will be converted to first-strand cDNA using Superscript II reverse transcriptase primed by a poly(T) oligomer that incorporates the T7 promoter. Second-strand cDNA synthesis will be followed by in vitro transcription for linear amplification of each transcript and incorporation of biotinylated nucleotide analogs. If a class of samples to be analyzed will typically not yield 1 ug of total RNA, multi-round IVT amplification will be performed by converting the first round cRNA to cDNA with random primed reverse transcription. The cDNA will serve as template for the next round of amplification, and this process can use a minimum of 0.2ng of input total RNA. After final amplification and biotinylation, the cRNA products will be fragmented to 200 nucleotides or less, heated at 99 C for 5 min and hybridized for 16 h at 45 C to Affymetrix microarrays. The microarrays will then be washed at low (6X SSPE) and high (100mM MES, 0.1M NaCl) stringency and stained with streptavidin-phycoerythrin. Fluorescence will be amplified by adding biotinylated anti-streptavidin and an additional aliquot of streptavidin-phycoerythrin stain. An Affymetrix scanner 3000 7G will be used to collect fluorescence signal after excitation at 570 nm.

3. Proteomics and Metabolomics: Protein and metabolite expression analyses will be conducted using validated rigorous stable isotope dilution liquid chromatography/tandem mass spectrometry (LC-MS/MS) methodology. Three duplicate quality control (QC) samples in the expected lower quartile, middle range, and upper quartile will be carried through with each analysis. In addition standard curves will be conducted for each analyte in each analytical run. Concentrations of study samples and QC samples will be determined by interpolation from the standard curves.

6.6.7 Quality Assurance and Control of Profiling Data

1. RNA profiling (microarrays) quality control: Samples for RNA extraction are shipped in PAXgene blood tubes or Trizol LS on dry ice. Barcoded tubes and storage boxes are logged into the caTissue inventory database and stored in -80C freezers with back-up power and alarm systems. After completion of RNA extractions, RNA samples only proceed to target labeling if appropriate size and abundance distributions are seen for ribosomal bands using the Agilent Bioanalyzer, and the RNA has an A260/280 ratio of 1.7-2.1 as measured on the Nanodrop spectrophotometer. One to five micrograms of input total RNA usually result in 60 to 100 ug of cRNA after IVT; this yield is a very good indicator of the quality of the initial RNA and the overall hybridization intensity that will be seen on GeneChips. Similar QC determinations are performed on each round of amplification for RNA samples that start with less than 1ug. Controls include standard spiked controls, purchased from Affymetrix, which are included in every hybridization cocktail, and a second set of polyadenylated RNA standards are spiked into primary RNA samples as amplification and labeling controls. Every GeneChip analysis performed in the Facility is accompanied by the QC report page generated in GCOS. This indicates 3'/5' ratios for GAPDH and actin, background values, percent genes detected and scaling factor applied, all of which are used as quality control criteria.
2. Proteomics/metabolomics: Samples for proteomics and metabolomics will be transferred from the Analytical Core to the Penn Proteomics Core. Samples will be logged in through a laboratory information management system (LIMS) and stored in a dedicated -80C. Three duplicate QC samples in the expected lower quartile, middle range, and upper quartile will be carried through with each analysis. In addition standard curves will be conducted for each analyte in each analytical run. Concentrations of study samples and QC samples will be determined by interpolation from the standard curves. Proteomic analyses will be conducted in two separate ways – profiling and quantitative analysis. For profiling studies, protein and metabolite concentrations will be determined by spectrophotometry and analyses will be conducted in the presence of a stable isotope-labeled proteome to provide a check of recovery. If recoveries are < 85 % the analysis will be repeated. For quantitative analyses, protein and metabolites will be analyzed by comparison of the peak area or peak height ratio to a heavy isotope proteome internal standard. A standard QC sample containing 50 protein and metabolites will be analyzed at the same time. If isotope ratios are < 85 % of those expected, the analysis will be repeated. Metabolomics and proteomics data will be transferred from the LIMS to the DMSI Core when QC criteria have been met.

6.7 Summary Table for the Design and Data Collection

Table 3 provides an overall summary of the design, data, and sample collection for the longitudinal study.

Table 3. Summary Table for Longitudinal Study

	Initial Visit	**Return Visit 1**	**Return Visit 2**	**PTL or pPROM**	**Labor & Delivery**	**Labor & Delivery**	**Labor & Delivery**
	10- <18 wks	**≥18- <24wks**	**28- <32 wks**		**20- <34 wks**	**34- <37 wks**	**39-<42 wks**
N	500	500	500		30[1]	60[2]	60[2]
Specimen Collection							
Maternal Blood	X	X	X	X[3]	X	X	X
Maternal Saliva	X	X	X	X[3]	X	X	X
Maternal Urine	X	X	X	X[3]	X	X	X
Maternal Cervicovaginal Fluid	X	X	X	X[3]			
Optional Amniotic Fluid	X[6]	X[6]					
Placenta					X	X	X
Fetal Membranes					X	X	X
Cord Blood (or saliva Oragene)					X	X	X
Forms							
Demographics	X						
Enrollment	X						
Intake Form (I, II, III)	X						
Current Medications	X	X	X				
Hospital Admission Form				X[3]			
Psychosocial Questionnaire	X				X	X	X
Ultrasound and labs	X	X	X				
Return Visit Form		X	X				
Labor and Delivery Form					X[4]	X[4]	X[4]
Neonatal Baseline					X[4]	X[4]	X[4]
Adverse Event Form	[5]	[5]	[5]	[5]	[5]	[5]	[5]

[1] It is anticipated that at least 30 will be collected, but will collect all SPTBs 20- <34 wks.
[2] It is anticipated that at least 60 will be collected, but will collect twice the number of spontaneous deliveries of those obtained at 20-<34 wks.
[3] To be recollected every two weeks until delivery or 34 weeks gestation.
[4] To be collected for all 500 patients in the study.
[5] Unscheduled forms to be used as needed.
[6] To be collected on an optional basis at Initial Visit **or** Return Visit 1

6.8 Statistical Analyses

6.8.1 Descriptive statistics and Data Quality Control

On a regular basis, personnel at the DMSI Core will examine descriptive statistics for SPTB, prenatal variables, perinatal variables, and postnatal variables as well as polymorphisms in some selected SNPs. These data will be extracted from the databases and saved as flat line files, and

then will be read into SAS for summary statistics. Progress will be reported at the Steering Committee meeting.

As described in Section 6.6.7, quality control steps will be taken by the Analytical Core before the data are sent to the DMSI Core for statistical analysis. These data will not have identifiable information.

6.8.2 Sample Size and Power

This study will follow 500 pregnant women who had a history of preterm deliveries. For the purpose of this study, this follow-up will yield a projected 30 SPTBs prior to 34 weeks that will be 1:2-matched by 60 SPTBs between 34 and less than 37 weeks and 1:2-matched by 60 term births between 39 and less than 42 weeks in this same "high risk" cohort. We decided to use the history of preterm deliveries to enhance the cohort after considering other established risk factors including ethnicity, based on the feasibility and clarity of the study design. The sample sizes for this study are determined by empirical experience, in which other studies cited below have used similar sample sizes to achieve similar objectives, and affordability.

6.8.3 Proteomics

Proteomics-based technologies are starting to play a key role in studies of biochemical pathways because they provide a way to define and characterize regulatory and functional networks of proteins (Tyers and Mann 2003). Unlike gene array analysis, proteomics requires the availability of diverse analytical methodologies (Aebersold and Mann 2003). Using such methodology, it is possible to investigate the precise molecular defect(s), and this may help develop specific reagents to better understand different stages of disease pathology. Such studies should ultimately lead to the identification of biomarkers that can be used in monitoring the potential for preterm birth as well as providing insight into the mechanism by which this occurs. Serum plays a central role in clinical diagnosis for many diseases. It is hoped that one day, given sufficient biomarkers, clinical proteomics will provide a network of information allowing for the early diagnosis of disease and individualized treatment of patients. This aspiration lies in the fundamental hypothesis that every cell in the body leaves a record of its physiological state as products shed into the local milieu either as waste or as messages to neighboring or distant cells. It has been estimated that 20-25% of all cellular proteins are secreted or shed. Since all tissues are perfused by blood and lymphatics, protein and protein fragments shed or secreted by cells can enter the circulation. Theoretically, blood contains numerous biomarkers that could reflect the ongoing physiological and pathophysiological states. Serum contains tens of thousands of proteins along with their cleaved or modified forms. These proteins can provide insight into ongoing physiological or pathological events. One current serum biomarker in use for detection and monitoring of prostate cancer is prostate specific antigen (PSA). There is no specific normal or abnormal PSA level, although the higher a man's PSA level, the more likely it is that cancer is present. Unfortunately, various benign factors can cause PSA levels to fluctuate, which can lead to unnecessary, more invasive testing. In addition, due to the known complexity in molecular mechanisms underlying disease, the quest for a single biomarker makes little sense from a biological perspective. Hence there is the need to analyze the entire complex mixture of proteins from serum and plasma, and other available biological fluids such as amniotic fluid and cervicovaginal secretions, and tissue samples.

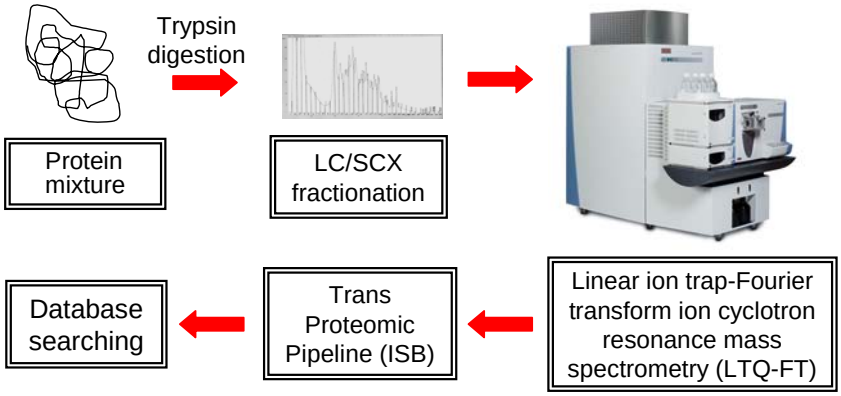

Figure 4: Steps involved in 2D LC-MS/MS proteomics analysis

The search for protein biomarkers has evolved from the original high throughput non-specific methodology based on surface enhanced laser desorption ionization (Ornstein and Petricoin 2004) to modern shotgun techniques (McDonald and Yates 2003; Swanson and Washburn 2005) based upon 2 dimensional liquid chromatography coupled with high resolution nanospray liquid chromatography-tandem mass spectrometry (2D LC-MS/MS) as shown in Figure 4. This technology provides the ultimate sensitivity and specificity when coupled with sophisticated peptide and protein identification algorithms such as PeptideProphet™ and ProteinProphet™, which are available in Transproteomic pipeline software developed by the Institute for Systems Biology. Proteins are subjected to trypsin digestion and then the first dimension of separation is performed using strong cation exchange chromatography. Typically, 15-40 fractions are then analyzed separately by reversed phase nanospray LC-MS/MS using data dependent scanning. Molecular weight information on the tryptic peptides is obtained from the protonated molecules (typically as doubly or triply charge ions). Collision induced dissociation is conducted on the protonated molecules and peptide sequences assignments are made by SEQUEST™ or MASCOT™ using the resulting product ion spectra. The data are then filtered through PeptideProphet™ and ProteinProphet™ using transproteomic pipeline software to eliminate false positive identifications, and grouped into various gene ontology (GO) classes (Figure 5). With this knowledge in hand, quantitative analyses can be conducted to compare expression of proteins from the various samples obtained under different physiological or experimental conditions.

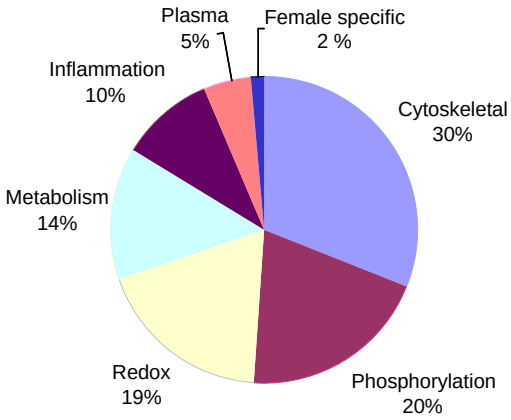

Figure 5: GO chart of 200 proteins identified in cervicovaginal fluids by 2D LC-MS/MS

Quantitative proteomics studies were initially conducted primarily using two-dimensional difference gel electrophoresis (2-D DIGE), which resolves multiple samples labeled with distinct fluorescent dyes on the same 2-D gel (Patton 2002). Differences in fluorophore intensity were translated into relative differences in protein expression in cells (Yan, Weaver et al. 2005) and the secretion of proteins into serum (Yu, Rustgi et al. 2005). 2-D DIGE provided a fast and reproducible way to compare multiple proteomes. However, one of its drawbacks is that when several proteins are present in the same spot, 2-D DIGE does not allow for quantification of the individual protein components (Yan, Weaver et al. 2005). Because of this limitation, various alternative strategies have been developed, all of which have limitations for global protein analysis. The original approach, which was termed isotope coded affinity tag labeling involved covalent addition of a deuterium labeled tag in which a biotin residue was also present to cysteine residues, (Gygi, Rist et al. 1999). Subsequently, a [^{13}C]-derivative replaced the deuterium to prevent separations of internal standard and analyte during LC-MS analyses (Hansen, Schmitt-Ulms et al. 2003). A further modification of the technique (known as isotope coded affinity tag for relative and absolute quantitation or ITRAQ) is still dependent upon a covalent reaction between proteins and the tag (Ross, Huang et al. 2004). Therefore, there is still a potential for differences in reaction rates between experimental and control proteins. Non-covalent quantitative methodology has focused on the use of stable isotope labeling by amino acids in cell culture (SILAC) (Ong, Blagoev et al. 2002). The limitations of this method are that it is confined to cell culture techniques and the experimental cells are grown in different media from the control cells, which could lead to differences in protein expression. More importantly, the amount of endogenous (unlabeled) proteins in the isotopically labeled experimental samples, limits the sensitivity of detection as it does in conventional stable isotope dilution procedures. These issues prompted the development of a modified SILAC approach in which a stable isotope proteome standard is prepared (Yan, Weaver et al. 2005 and Yocum, Busch et al. 2006). When this methodology is coupled with ultrahigh high-resolution 2D nanospray LC-MS it provides a level of specificity that cannot be obtained with other methods.

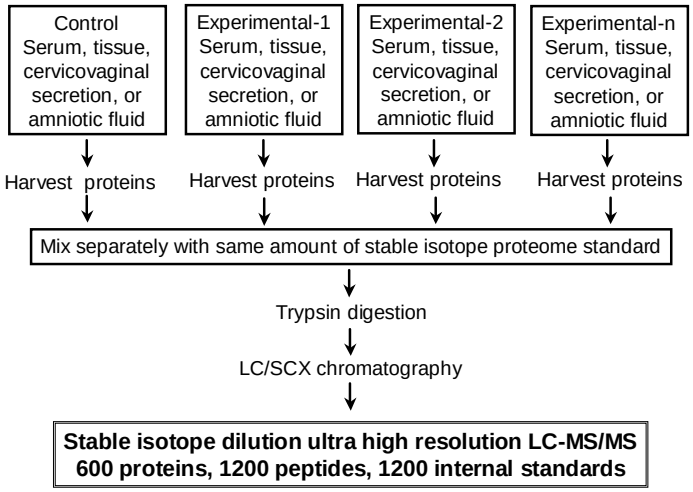

Figure 6: Quantitation of proteins using a stable isotope labeled proteome internal standard.

A stable isotope labeled proteome is first prepared using cell culture techniques (Yan, Weaver et al. 2005 and Yocum, Busch et al. 2006). Cells are grown in lysine- and leucine-deficient media to which [$^{13}C_6{}^{15}N_2$]-lysine and [$^{13}C_6{}^{15}N$]-leucine have been added. After the cells had been

passaged seven times, the labeled proteome is extracted and stored ready for use as an internal standard. Tryptic peptides obtained during the analytical procedure (Figure 4) containing a single [$^{13}C_6{}^{15}N_2$]-lysine are increased in mass by 8 Da and those containing a single lysine and leucine are increased by 15 Da. Peptides with multiple [$^{13}C_6{}^{15}N$]-leucine residues as well as those with missed cleavages at [$^{13}C_6{}^{15}N_2$]-lysine have exactly the same LC retention time as the corresponding unlabeled peptides (unlike deuterium labeled peptide). For quantitative proteomics studies, the same amount of labeled proteome internal standard is added to control and experimental proteins samples from serum, tissue, cervicovaginal fluid, or amniotic fluid (Figure 6). Samples are then processed through the 2D LC-MS/MS procedure and peptides quantified either by selected ion monitoring at ultrahigh resolution or low resolution multiple reaction monitoring (MRM) using specific parent and product ions from selected peptides. The high-resolution method permits up to 600 proteins to be quantified in a single analysis. The low resolution MRM method is useful for focusing on 30-40 higher abundance proteins within a given sample. Ratios between labeled and unlabeled peptides are calculated using the ASAPratio program available in the transproteomic pipeline software and the mean ratio of the two chosen peptides can be compared between control and experimental conditions. Absolute quantitation is conducted simply by constructing standard curves using known protein standards with the same amount of labeled proteome standard that was added to the control and experimental samples. Rigorous validation of the method is conducted using good bioanalytical practices (GBP) as described by Chaudhary, Wickremshinhe et al. (2006).

Highly abundant proteins in serum and plasma significantly interfere with the detection of less abundant proteins (Yocum, Yu et al. 2005). Controversy within the field of proteomics has emerged on the best way to obtain the greatest number of protein identifications from human serum; i.e., to deplete the major components; e.g., albumin, immunoglobulins (Ig), or not. Additionally, the high sequence variability within immunoglobulins complicates protein assignment. As a result, these highly abundant proteins are routinely removed from serum. This reduces the range of protein concentrations that are present allowing lower abundance proteins to be analyzed and quantified. The size of a protein determines how fast it is cleared from the blood. Free flowing low molecular weight proteins and peptides (< 30 kDa) should be rapidly cleared by the kidneys. Abundant high molecular weight proteins that exist above the filtration range of the kidneys will exist in serum until they are proteolytically cleaved into smaller fragments and excreted. Half lives of proteins in serum can vary widely: the half-life of albumin is approximately 20 days as compared with that of complement factors, which are in the range of several minutes. Carrier proteins such as albumin accumulate and amplify lower abundance proteins. Currently, there are a several technologies available to remove the higher abundance carrier proteins from serum, including ultra centrifugal filtration and immunoaffinity depletion. These techniques satisfactorily remove high abundance proteins, although there are selective losses through the non-covalent binding that occurs with selected proteins (Yocum, Yu et al. 2005). Nevertheless, immunodepletion of serum coupled with stable isotope dilution nanospray 2D LC-MS/MS (Figure 4) provides an efficient way to conduct protein biomarker identification and quantitation for preterm birth studies.

6.8.4 Metabolomics

Metabolomics is the systematic study of the unique chemical fingerprints that specific cellular processes leave behind (Weckwerth and Morgenthal 2005 and Fischer 2005). It involves studies

of metabolite target analysis, metabolite profiling, metabolic fingerprinting, and metabolic profiling. The metabolome represents the collection of all metabolites in a biological organism, which are the end products of its gene expression. Thus, while mRNA gene expression data and proteomic analyses do not tell the whole story of what might be happening in a cell, metabolic profiling can give an instantaneous 'snapshot' of the physiology of that cell. Integration of proteomics, transcriptomics, and metabolomics information will provide a more complete picture of potential mechanisms involved in the preterm birth syndrome. Many of the bioanalytical methods used for metabolomics have been adapted from existing biochemical techniques, and thus there are not always clear distinctions between studies that are described as metabolomic and studies that are concerned with metabolism. However, metabolomic research is characterized as being without bias towards a specific metabolite or group of metabolites. Highest specificity is attained by the use of stable isotope dilution LC-MS, which introduces bias towards known metabolites. The gain in specificity more than compensates for this potential problem (Figure 7). Lipidomics is a sub-discipline within the field of metabolomics (Lee, Williams et al. 2005 and Lee, Williams et al. 2003). It is defined as the characterization of lipids and their interacting moieties within a cell, tissue, or organism (Gross, Jenkins et al. 2005 and Han and Gross 2005). The crucial role of lipids in cell signaling and tissue physiology is demonstrated by the many neurological disorders, including bipolar disorders and schizophrenia, and neurodegenerative diseases such as Alzheimer's, Parkinson's and Niemann-Pick diseases, that involve deregulated lipid metabolism. Altered lipid metabolism is also believed to be a key event, which contributes to cerebral ischemic injury. Thus, lipidomics analysis can provide molecular signature to a certain pathways or a disease condition. The application of metabolomics and lipidomics technology for analyses of serum, tissue, cervicovaginal secretions, and amniotic fluid in the field of preterm birth research is in its infancy. This provides a new opportunity to make significant and rapid advances in the field. Rigorous validation of the metabolomics and lipidomics methods are also conducted using GBP as described by Chaudhary, Wickremshinhe et al. (2006).

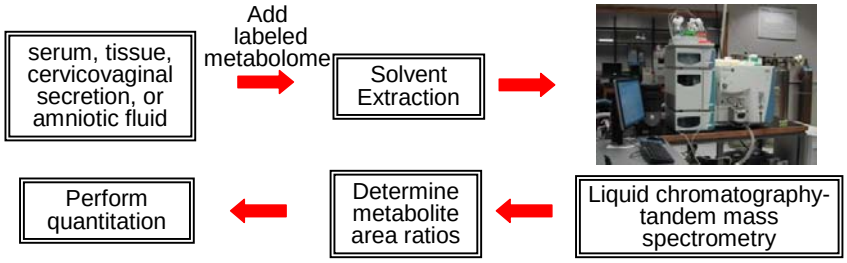

Figure 7: Steps involved in a stable isotope dilution LC-MS/MS metabolomics analysis.

6.8.5 Statistical Analysis of Expression Profiles

For the illustration purpose, we concentrate on the analysis of gene expression profiles using data collected from microarry technology because the proteomic and metabolomic technologies described above will yield data of the same format. The applications of this technology encompass many fields of science from the search for differentially expressed genes (Heller, Schena et al. 1997), to the understanding of regulatory networks (Segal, Shapira et al. 2003), DNA sequencing and mutation study (Hacia, Brody et al. 1998), SNP (single nucleotide

polymorphism) detection (Fan, Chen et al. 2000), cancer diagnosis (Ramaswamy, Tamayo et al. 2001), and drug discovery (Marcotte, Srivastava et al. 2001).

As reviewed by Duan and Zhang (2004), Figure 8 provides a four-level overview of the analytical process. For this study, the first three levels are relevant. The first challenge in dealing with the microarray data is to preprocess the data involving background subtraction, array normalization, and probe-level data summarization. The purpose of this preprocessing is to remove noise and artifacts, in order to enhance and extract hybridization signals. This data pre-processing is also often referred as the low-level analysis (Li and Wong 2001).

The second-level analysis usually contains two steps: one is to filter "unusual" genes whose expression profiles are suspicious due to noise or too extreme, and the other is to identify the differential expressed genes across different samples – among the four groups in the present study. To identify genes that have significantly different expression profiles, the commonly used approaches include the estimation of fold change, student's t-test, Wilcoxon rank sum test, penalized t-test, empirical Bayes (Efron, Tibshirani et al. 2001), and SAM (Significance Analysis of Microarray) (Tusher, Tibshirani et al. 2001). Because we have repeated measures of expression profiles, SAM will be based on MANOVA or random effects models.

The third level of analysis is sometimes referred to as a high-level analysis (Irizarry, Ooi et al. 2003), and it includes clustering, classification and pathway analysis. This is usually conducted on a subset of genes that are selected from the second level analysis. To identify genes that may be correlated to each other, clustering analysis has become particularly popular, and the approaches include hierarchical clustering (Eisen, Spellman et al. 1998), k-means (Soukas, Cohen et al. 2000), self-organization maps (SOM) (Tamayo, Slonim et al. 1999), principle-component analysis (PCA) (Yeung, Haynor et al. 2001), and probabilistic model-based clustering (Yeung, Fraley et al. 2001).

To classify gene expression profiles into different clinical groups, both classic discriminant analysis and contemporary classification methods can be used. The methods include k-nearest neighbors (KNN) (Troyanskaya, Cantor et al. 2001), linear discriminant analysis (LDA) (Zhang and Luo 2003), support vector machine (SVM) (Furey, Cristianini et al. 2000), artificial neural networks (ANN) (Mehrotra, Mohan et al. 1997), classification trees (Zhang, Yu et al. 2001), and random and deterministic forests (Zhang, Yu et al. 2003).

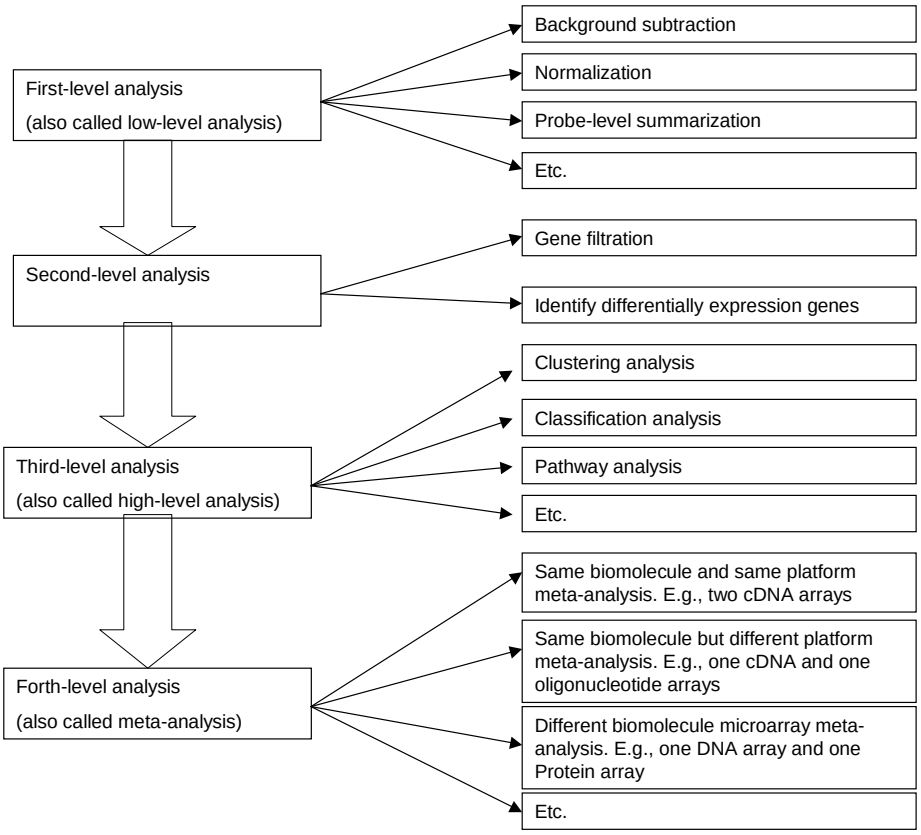

Figure 8. Diagram of the four-level analysis of microarray data (Duan and Zhang 2004).

To identify genes that may be on the same pathway to a particular biological process, relevance networks (Butte and Kohane 2000), linear differential equation (D'Haeseleer, Wen et al. 1999) Boolean networks (Shmulevich, Dougherty et al. 2002), Bayesian networks (Friedman, Linial et al. 2000) and probabilistic rational model (PRM) (Segal, Taskar et al. 2001) can be used.

6.8.6 Example: Clustering Analysis of Gene Expression Data to Assess the Impact of Preterm Birth on Developing Brain

Dr. Ment and her team have developed the murine model of chronic sublethal hypoxia which faithfully resembles the preterm effect on infants. They have collected microarray gene expression data to understand the global molecular mechanisms of cerebral development in the neonatal rodent model of chronic sublethal hypoxia.

Based on the SAM analysis as described above, the expression levels of 1,109 genes were altered at $p < 0.01$ level of significance, and 163 of those 1,109 were increased 1.4-fold or more. Of these 163 genes, 92 were known genes. Figure 9 presents a sample graph from clustering analysis of those genes. When they were analyzed according to function, the mRNAs for many genes involved in cell growth and mitogenesis were significantly increased by hypoxia. Among the components of the cell cycle machinery, cyclin G2 and cdk4 were significantly up-regulated, whereas other cyclins and cyclin inhibitors were not significantly changed.

6.8.7 Example: Classification Analysis of Gene Expression Data

In the classification setting, we are given a training set of observations that contain a number of features (e.g., gene or protein expressions) as well as the groups(e.g., preterm versus term). These observations are used to induce a classification model. This model can then be applied to predict the class label (preterm versus term) for a set of previously unseen instances (new tissue samples). Sample re-use methods are generally used to assess the prediction precision of classifications rules. Because the number of genes or proteins is far larger than the number of available samples, it has become an interesting and important issue as to which method provides a more faithful measure of prediction precision (Zhang et al. 2001; Ambroise and McLachlan 2002; Dudoit, Fridlyand et al. 2002; Zhang and Yu 2002).

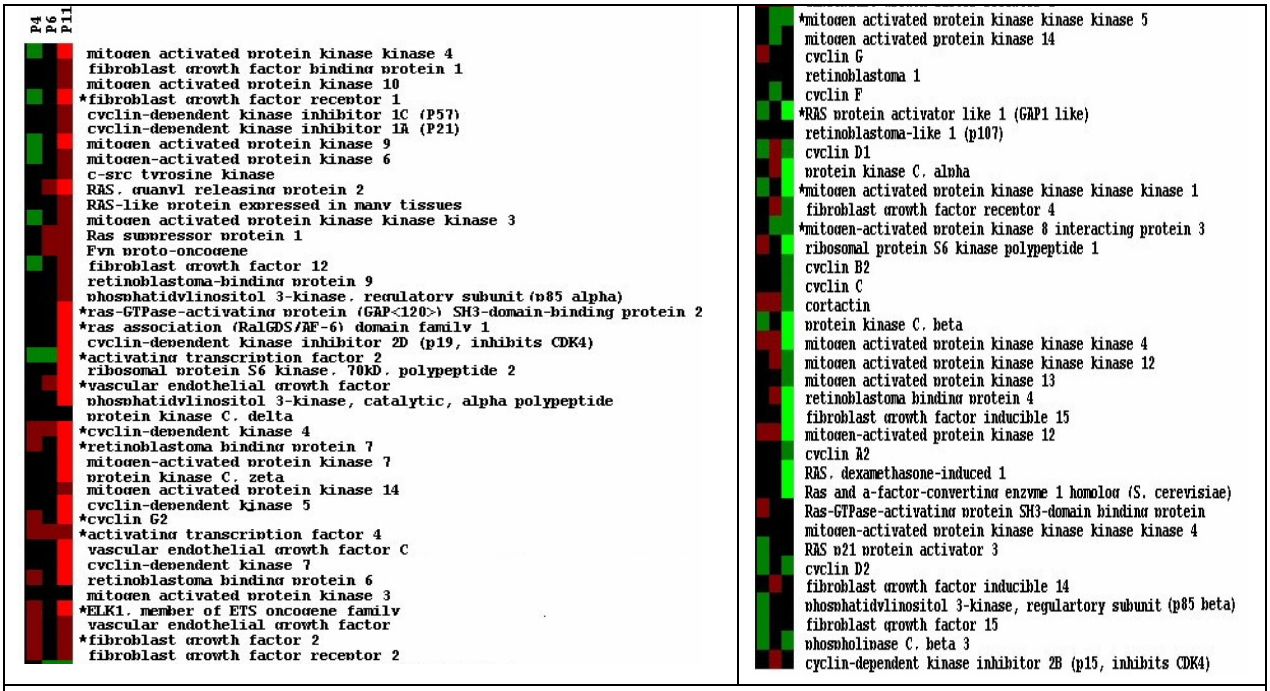

Figure 9. Cluster analysis of cDNA microarray data showing the effect of hypoxia on a set of genes involved in cell growth and mitogenesis. The ratio of the hypoxic intensity to the normal intensity at each time point was carried out for each gene and hierarchical clustering analyses was performed. For each time point, the red color denotes increase in mRNA relative to normoxia; the green color indicates a decrease in mRNA; black indicates no change between normoxia and hypoxia. Star indicates significance at $p<0.01$ level when the data are analyzed across injury and time.

The recursive partitioning technique can deal with feature selection and feature extraction simultaneously. While it can be conveniently used to automatically select genes whose intensity of expression or non-expression can distinguish different groups, it is also flexible enough to allow us to accommodate prior knowledge or expert opinion at any intermediate stage in the procedure, bypassing the statistical selection criteria, if needed. This is critically important because it is very important to incorporate our biological knowledge to the discovery of neurological discovery. As shown in Zhang et al. (2001), the classification rules resulting from recursive partitioning can be remarkably precise compared to those derived from more routine

methods. In addition, competing classification trees can often be obtained, which could be suggestive of co-regulation (by different genes) of genes in a pathway.

We refer to Breiman et al. (1984) and Zhang and Singer (1999) for the general methodology. Here, we examine one published data set on lymphoma data (Alizadeh et al. 2000). Data are available on the three most prevalent adult lymphoma malignancies: B-cell chronic lymphocytic leukemia (B), follicular lymphoma (F), and diffuse large B-cell lymphoma (D). There are a total of 84 samples (29 B, 9 F, and 46 D) with expressions from 4,026 genes. Figure 10 presents a classification using the Gini index (Breiman et al. 1984) as node impurity.

This example illustrates how to grow a simple and precise classification tree for tumor cells. For this longitudinal study, this means the identification of genes that change more dramatically than others under SPTB (<34), SPTB (34 - <37), or term conditions (39- <42). Because we have so many genes and few samples, it is not unusual to have a simple, precise classification tree. What is challenging is the fact that there are potentially many alternative simple and precise trees. Instead of betting on one tree, it is useful to examine forests rather than a single tree. The forest of these competing trees can be extremely useful in this Network. On one hand, it is likely that the information contained in a large number of measures can be captured by a smaller number of them without significant loss of information due to the underlying correlations. Clinically, this could be a direct result of the fact that clusters of measures (e.g., genes) are similarly regulated. On the other hand, variables selected in different trees are not necessarily highly correlated. Those trees may suggest different pathways or mechanisms among biological correlates.

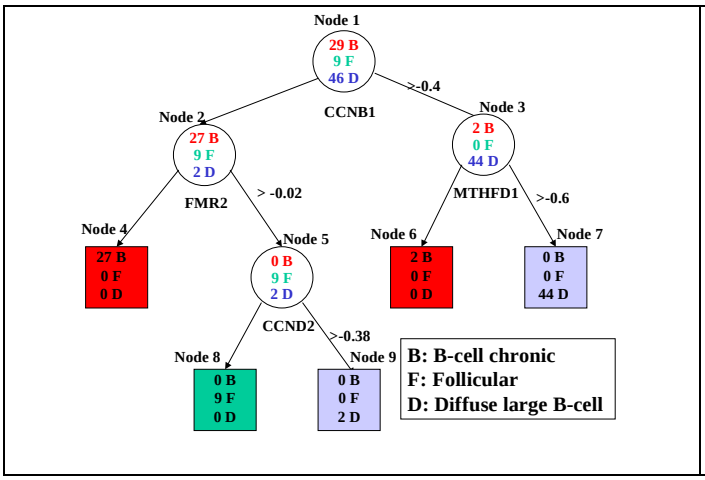

Figure 10. Tree Structure. Circles represent internal nodes that are subsequently divided into daughter nodes. The boxes are terminal nodes that do not have further partition and determine the tissue class membership. Beneath each internal node is the gene whose expression level is used to split the node, and the cutoff is displayed on the arrow next to it, on the right.

Reducing the misclassification rate is one objective, and another objective is to understand the genes that participate in the formation of the forests. To the second end, we can assess the frequencies of different genes appearing in the deterministic forest, because a high frequency is indicative of the importance of a gene in the respective experiment. For the lymphoma data set, 54 different genes appeared in the forest of 100 trees. The left panel in Figure 11 presents the frequencies of the genes that are used relatively frequently. The right panel in Figure 11 displays the expression levels of the three frequent genes with cell type labeled in different colors. Some of these genes are used in Figure 10. Many trees in the forest use these genes with different cut-off values.

One important use of forests in this Network is to rank the pool of genes and proteins that manifest great differential expression levels between preterm and term conditions. For example, from the existing literature, we found evidence supporting most genes in Figure 11 in relation to

lymphoma (Zhang et al. 2003). This is actually not by accident because most of the genes that have been independently identified by the tree-based methods are known to be associated with the disease endpoint of interest in the existing literature (Zhang et al. 2001, 2003).

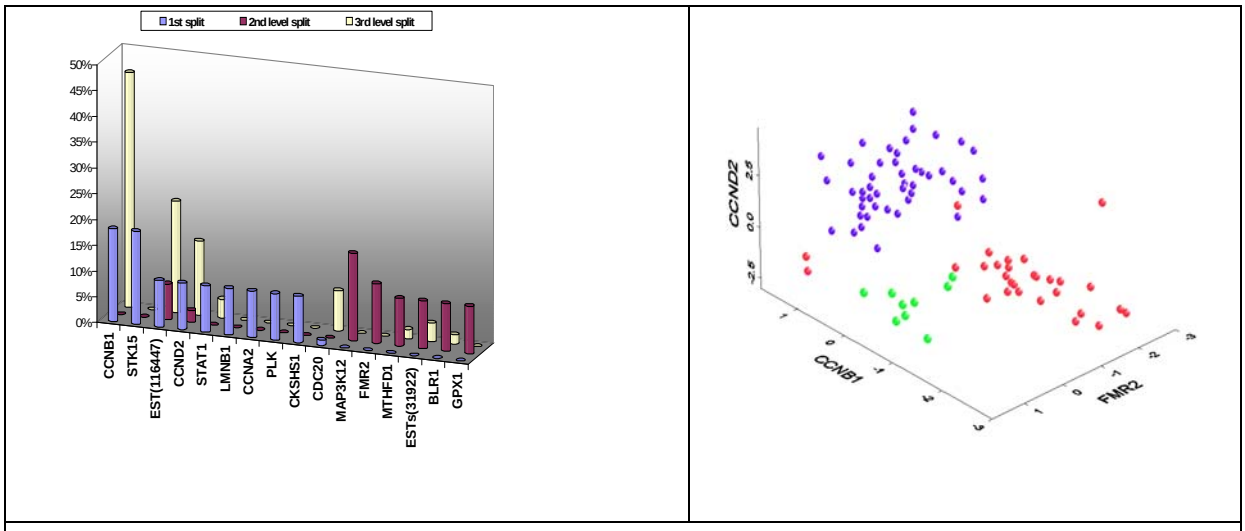

Figure 11. Analysis of Lymphoma Data Set. Left: Frequencies of the genes appearing in the 100-tree deterministic forest. Right: A 3-d representation of a tree based on three most frequent genes. Three colors show a separation of B (red), F (green), and D (purple) lymphoma. Source: Zhang et al. (2003).

6.8.8 Missing Data

Missing data are common in medical research and there is extensive literature concerning how to conduct statistical analysis in the presence of missing data. For our Network, we will make a diligent effort to prevent missing data. However, missing data are expected to occur in covariates and SNP typing. If covariates and SNPs are missing at random, Little and Rubin (1987) is one of the early references that systematically addresses these issues. More recent, excellent reviews include Little (1992), Horton and Laird (1999) and Kenward and Molenberghs (1999). For example, for the regression analysis, Little (1992) assumes a parametric form for the joint distribution of the response and regressors and Robins et al. (1994) propose a class of semiparametric methods based on weighted GEE. Extensions to generalized linear models are also available from Ibrahim (1990). An extensive literature for dealing with missing data is established for both continuous and categorical longitudinal data (Kenward and Molenberghs 1999). We will consider these approaches cautiously in our data analyses when missing data are of concern. We also note that all of the approaches require certain assumptions that cannot be validated from the data under analysis (Kenward and Molenberghs 1999) and that an appropriate use of imputation or other methods to deal with missing data could cause more problems than it addresses.

6.9 Human Subjects –Template Consent Form

1.0 **You are invited to take part in an important research study about spontaneous premature birth.** Spontaneous (unplanned) premature birth happens when a woman goes into labor and delivers a baby more than three weeks before its due date.

Researchers from five areas in the United States are studying the factors that might be related to spontaneous premature birth. The researchers hope that this study will help prevent spontaneous premature births in the future. This research study is funded by the National Institutes of Health (NIH) / National Institute of Child Health and Human Development (NICHD). It is called the Genomic and Proteomic Network for Premature Birth Research. Approximately 2,500 pregnant women from across the nation will take part in the study.

The best single predictor of who will have a spontaneous premature birth is a history of a spontaneous premature birth in a previous pregnancy. Thus, to learn more about spontaneous premature birth, it would be particularly helpful to follow a group of pregnant women who have experienced a previous spontaneous premature birth. You are being asked to take part in this study because you are currently pregnant and have previously had a spontaneous premature birth that was at least six weeks early. This part of the study will sign up 500 such women from across the nation. We would like to tell you about a study that hopes to:

(a) identify those pregnancies that are – and are not – going to experience another spontaneous premature birth, and
(b) identify the ways by which spontaneous premature birth occurs.

This study will use several powerful new research methods to study the problem of spontaneous premature birth. Genomics is the study of specific genes that are associated with any given medical problem, in this case, spontaneous premature birth. Genes are passed on from our parents and are the building blocks of our bodies that tell our bodies how to work. Proteomics is the study of proteins (produced by specific genes) that are seen with a given medical problem, in this case, spontaneous premature birth. Metabolomics is the study of small molecules that are associated with a given medical condition, in this case, spontaneous premature birth."

What we learn from this study could give us valuable information about this common and serious problem that could lead to more effective treatment and/or prevention.

2.0 **What will happen if I join this study?** For this study of spontaneous premature birth, we would:

a) ask you some questions,
b) collect facts from your medical records and those of your baby,
c) obtain up to four samples of your blood, urine, vaginal secretions and saliva taken at different times during your pregnancy,
d) examine and obtain samples of your placenta (afterbirth), umbilical cord and membranes, and
e) obtain a sample of blood from the umbilical cord or a cheek swab from your baby after birth.

If you deliver more than 6 weeks prior to your due date, we would collect all of these samples. Otherwise, you may – or may not – be selected to have your delivery samples collected. The delivery samples include your placenta, umbilical cord and membranes and cord blood or cheek swab from your baby. The decision to collect – or not collect –

your delivery samples will be determined by random selection (like drawing numbers from a hat). If you are selected, we would collect all of these samples.

If you were to have an amniocentesis (withdrawing some fluid from around the baby for clinically indicated testing, such as prenatal diagnosis or lung maturity testing) we would also ask your permission to draw an extra teaspoon of fluid for research testing.

Genomic, proteomic and metabolomic testing would be performed on the samples that are collected, and the findings would be compared with the outcomes of the women and babies enrolled in the study. All information will be strictly confidential and you will not be identified in any way. More details on how we will protect your privacy are given in questions 10.0 and 11.0 below.

2.1 **What kinds of questions will I be asked?** A trained member of our research team will ask you some questions about your pregnancies, your medical history, your family history, and your day-to-day life and work, including smoking, drinking, and drug use. Most of this interview will be done early in your pregnancy. This interview could be done either at the time of a scheduled clinic visit or at some other time that might be more convenient for you. We would update some of the information at other study visits during the pregnancy (up to three total) and after you deliver your baby. The initial interviews should take about 30 to 45 minutes.

2.2 **How will my medical records be used?** We wish to look at the medical records of your pregnancies, including records of your doctor and from your hospital visits. We wish to collect information from your records to help us better understand spontaneous labor, both premature and full term. This request to access your records (and those of your baby) will comply with HIPAA, which is a law protecting the privacy of your health information.

2.3 **How will my baby's medical records be used?** We would like to collect information from your baby's medical records from birth to the time that your baby leaves the hospital. We will use this information to look at differences between babies that are born prematurely and those that are born full term (within 1-2 weeks of the due date).

2.4 **How will my blood, urine, vaginal secretions and saliva samples be used for this study?** In addition to your usual blood tests, an additional 10 milliliters (about two teaspoons) of blood would be drawn from a vein in your arm at up to four different times during your pregnancy and delivery. These times are:

a) 10-18 weeks (2 ½ - 4 months)
b) 18-24 weeks (4 - 5 months)
c) 28-32 weeks (6-7 months), and,
d) around the time of delivery.

These time points have been selected in part because they are times in pregnancy when you would be having blood drawn for routinely recommended clinical testing. These blood samples will be sent to a specific research laboratory at the University of Pennsylvania (Philadelphia, PA), where specialized testing will be done that will look for genes, proteins and metabolites during the course of pregnancy that might tell us whether or not spontaneous premature birth will, or will not, occur.

We would also ask for a sample of your urine and vaginal secretions at each of these times during pregnancy. The urine samples would be studied to see if any protein and metabolites could be identified that could predict whether or not spontaneous premature birth might occur.

At the same time that blood samples are obtained we would also ask you to provide a sample of saliva.

If you were to require a long hospitalization later in your pregnancy because of premature labor or premature ruptured membranes, we would also request your permission to obtain additional blood, saliva, urine and vaginal swab specimens every two weeks from the time of your admission until you deliver. We would talk to you about this again should this complication arise and would not otherwise request any such additional samples.

Your baby's blood will be obtained after delivery from the discarded umbilical cord. We will only save your baby's blood if you agree. Otherwise it will be thrown away. If a blood sample cannot be obtained from your baby's umbilical cord, we would ask your permission to obtain a cheek swab sample from your baby in order to provide a saliva sample that would be used to study your baby's DNA.

Most genes do not differ among people, but some genes and combinations of genes are unique to each person. If you agree, DNA from you, your placenta and your baby will be used to see whether the genes of mothers and babies with and without spontaneous premature birth are different. We will store a sample of your and your baby's DNA. These samples will only be used to study spontaneous premature labor and pregnancy-related diseases. These samples would also be sent to the University of Pennsylvania, where DNA would be studied in hopes of finding specific genes that might predict whether or not spontaneous premature birth will, or will not, occur.

2.5 **What will happen when the placenta (afterbirth), umbilical cord and membranes are examined?** The placenta will be examined in both normal and additional ways. We want to examine the placentas of all women in the study so that we can compare placentas of premature babies to those of babies who are born full term. We will take pictures and measurements and also examine the placenta under the microscope. We will also perform tests on your placenta that look for signs of infection, or biochemical differences (genes, proteins, metabolites) that might occur more often in spontaneous premature births.

2.6 **How will my samples be stored?** All blood and saliva samples will be stored in a secure facility at the University of Pennsylvania for five years and then in a secured facility (to be determined) until the samples are no longer useable or until destroyed. To protect you and your baby, samples will only be labeled with a study identification number (study ID). We will keep your study ID linked with your name in our local records (locked and secure) for up to ten years after the end of the study. If you decide that you want the stored samples destroyed while we have this link, you can let us know in writing and we will have the samples that are labeled with your study ID destroyed. The stored samples will only be released for other studies if an independent committee that reviews research on people (called an Institutional Review Board) has approved the research. Research findings on stored samples will not be reported back to you. Stored

samples will only be used for the study of spontaneous premature birth, and pregnancy or baby-related diseases unless you are contacted for your permission to use the samples for other purposes (see question 2.7 below).

2.7 **Could I be contacted for future studies?** A research person authorized by this study may contact you in the future if there are any new research questions or to follow up on the health and well-being of your baby and you. If you ask us not to contact you for future studies, your last contact with us will be at the time that you and your baby are discharged from the hospital.

2.8 **Could information from my baby and me be used in other future studies?** The results of your and your baby's research tests (including DNA tests) and related medical records will be placed in an electronic database that could be used for other research studies in addition to this study. These studies would involve only pregnancy or baby-related diseases. Your privacy will be strictly maintained, since information that would identify you or your baby would not be available. However, in certain situations a researcher might want to contact you for additional information that would be useful only for other research purposes. In these cases, the researcher would make a formal request to an oversight committee. This committee would review the request and either deny or approve the researchers request for your contact information. The researchers would have to guarantee that they would not release either your contact information or your database records before permission would be granted. This would only apply if you have given prior permission to be contacted in the future by a researcher. If you have not given your prior permission the request would be automatically denied.

3.0 **Risks and Benefits from Study Participation**

3.1 **What discomforts or risks could I have from participating in the study?** Some of the interview questions may make you feel uncomfortable or upset. You do not have to answer any questions that you do not want to answer. You may stop the interview at any time without affecting your medical care. Occasionally drawing blood may cause pain and bruising. Rarely, drawing blood also causes some people to faint or feel light-headed and causes blood clots or infection. Research staff members are available to help you get medical care if you need it because of the study.

If you agree to be in this study we will be obtaining information from your blood, saliva, urine or placental samples about whether you and your baby have been exposed to different kinds of drugs. We will not provide the results of these research tests to any law enforcement or social agency. We have obtained a Certificate of Confidentiality from the federal government so that we cannot be forced to release this information to anyone (see question 11.0 below).

Laboratory tests, including DNA testing, carry some unique risks. The concern is that test results might someday be used against a person. However, the risk for something like that happening as a result of being in this study is extremely small. First of all, the researchers will be very careful to ensure that only authorized people have access to the samples and your personal information. Secondly, the types of tests we plan to do cannot, at the current time, be directly linked to an increased risk of a disease, for example cancer or diabetes. In addition, the results will not be placed in your medical record. You will not be given any information at all about the results of the tests we are doing, because even if

we are successful in identifying protein and metabolites, genes or other types of chemicals, we will still not have enough information to directly benefit you or your family. We will only use the tests from you and your baby to study spontaneous premature birth and pregnancy or baby-related diseases unless you give us separate permission to study other diseases.

There may be risks while participating in this study that are unknown. <<LOCAL SITES MAY ADD INFORMATION REGARDING LIABILITY, ETC., PER LOCAL IRB REQUIREMENT HERE.>>

3.2 **Will I benefit from joining the study?** There are likely no direct benefits to you from participating in this study. However, we hope to learn how to reduce the likelihood of spontaneous premature birth. You could have the satisfaction of knowing that you have given information that may help prevent spontaneous premature births for other pregnant women in the future. General information from this research will be published in medical journals, but will not be relayed to you or benefit you directly.

4.0 **Do I have to join this study? What are my other choices?** You do not have to be in this study. You may choose to participate in some, but not all, parts of the study. Your medical care will not be affected if you choose not to take part in this study; you will still be able to get all appropriate medical care without being in the study. Your choice about participating in this study will not affect your eligibility for any health plan or any health plan benefits or payments.

5.0 **Withdrawing from the Study**

5.1 **What if I change my mind about how I want to take part in the study?** Even after you agree to be in this study, you can change your mind at any time and pull out of the study. You can refuse to answer a question or refuse to take part in any part of the study. If you change your mind, your medical care and health plan eligibility and benefits will not change. At any time you can also request that your samples be destroyed.

5.2 **Can the researchers remove me from the study?** You may withdraw from the study at any time without penalty. Likewise, Dr. XXXXX and HIS/HER associates or the National Institutes of Health can withdraw you without your approval. A possible reason for withdrawal could be the early termination of the study by the National Institutes of Health.

6.0 **If I have expenses because I took part in this study, can the study help pay these costs?** If you agree to take part in this study, usual travel and child care expenses (<<UP TO $??>>) that result from your being in this study will be paid. <<ADDITIONALLY, YOU WILL RECEIVE $?? AS COMPENSATION FOR THE TIME YOU SPEND ANSWERING QUESTIONS. (SENTENCE ONLY NEEDED IF REIMBURSEMENT FOR TIME IS OFFERED AT THE SITE. SOME SITES MAY SPLIT PAYMENT BETWEEN THE TWO PARTS OF THE INTERVIEW. SUCH DETAILS WILL BE GIVEN HERE. SEE PROTOCOL FOR GUIDELINES.

7.0 **Will joining this study increase my health care costs?** No. Your routine medical care and routine testing will be billed to your insurance company or to you the same as your other medical care. The costs of tests that are performed only for purposes of the research study will not be charged to you or your insurance company.

8.0 **What will happen if I get a bill for research tests by mistake?** If you receive a bill for any study-specific test that is not paid by your insurance company or Medicaid or if you are not sure if the bill is for a study-specific test, call the study team at <<PHONE NUMBER>> and they will ask you to send a copy of the bill to the following address to find out if this test was done for the study. If it was done because of the study, the study will pay for it.

<<NAME AND ADDRESS OF PI>>

9.0 **What do I do if I have questions or need to report a problem?** If you have questions now, feel free to ask them to the study staff. If you have more questions later, please contact <<NAME AND CONTACT INFORMATION OF APPROPRIATE FIRST CONTACT>>.

If he/she is not available, you may contact
<<NAME AND CONTACT INFORMATION OF APPROPRIATE SECOND CONTACT>>. <<PROVIDE INSTRUCTIONS FOR AFTER-HOUR CONTACTS AT YOUR INSTITUTION>>

The committee that reviews research on human subjects (Institutional Review Board) at

<<NAME OF INSTITUTION>>
will answer any questions about your rights as a research subject. They can be contacted at

<<IRB CONTACT NUMBER (AND CONTACT PERSON IF AVAILABLE)>>.

10.0 **How will my privacy be protected?** The researchers are taking many steps to protect the privacy of you and your baby. First, we will use a number to identify research records and tissue and blood samples, instead of your name. If you agree to take part in the study, a study ID number will be assigned to you and your baby. This same number will be used to identify medical information kept by the study. Any records linking your name to the study ID will be stored in a locked (password protected) file.

All copies of your medical records and the links between your name and the study ID will be destroyed within seven years of the end of the study. After that time your blood or tissue samples and information from your interview and medical records collected for the study will be marked with your study ID number only. Even though the information could still be used for research, it will no longer be possible to link it to you. Any results of this study reported in medical journals or at meetings will not identify you or your baby.

Only those who work with this study will be allowed access to your information. However, representatives from the National Institutes of Health may inspect and/or copy the records that identify you.

Results of the study may be published. However, your name and other identifying information will be kept private.

11.0 **What is a Certificate of Confidentiality?** To help us further protect the privacy of you and your baby, the researchers have obtained a Certificate of Confidentiality from the National Institutes of Health (NIH) / National Institute of Child Health and Human Development (NICHD). This Certificate means that the researchers cannot be forced (for example by court order) to disclose any information that might identify you to any federal, state, or local court. A Certificate of Confidentiality does not prevent you from voluntarily giving information to others about yourself or your involvement in this research. You should know that we may provide information to your health care providers if we suspect that you may harm yourself or others. We will not release any information collected as part of the research regarding use of illicit drugs and testing for drugs done on samples collected for the research.

12.0 **If I agree to be in all parts of this study, what does that include?**

Full participation in this study includes all of the following:

- An interview
- Collection, review and storage of medical records
- Obtain up to four samples of your blood, urine, vaginal secretions and saliva samples that were obtained at different times during your pregnancy
- Examine and obtain samples of your placenta, umbilical cord and membranes
- Obtain your baby's blood or saliva sample within a day of your delivery
- Genetic, proteomic and metabolomic studies on you and your baby
- Storage and future use of blood, urine, vaginal secretions, saliva, placenta, umbilical cord and membranes, and DNA samples
- Research staff possibly contacting you later for other research questions.

13.0 **Not Participating in a Part of the Study**

13.1 **If I still want to join this study but would like to avoid some parts of it, what should I do?** Agreeing to do all parts of this study will allow the researchers to gather the most useful information to study spontaneous premature birth. However, if you do not want to do all parts of the study, you can refuse to do any one or more parts of the study.

Parts of this study that you do not wish to do are:

None ☐ Restrictions, if any: _____

13.2 **Follow-Up of you and your baby after you deliver and go home**
Please indicate your preference about being contacted after you and your baby have gone home:

☐ Yes, I give my permission for researchers to contact me after my baby and I go home. I understand that this contact will be limited to follow up of our health and well-being.

☐ No, I do not want to be contacted after my baby and I go home.

14.0 **Will my information and samples be used in other studies of pregnancy complications?** The researchers may want to study other pregnancy complications using the information and samples that have been collected in this study. This could only happen if you give the researchers permission. You do not have to give them your permission but we would like you to consider initialing one of the three boxes if you are comfortable doing so. If you do not initial a box giving the researchers this permission, your samples will not be used.

☐ I give my permission for my and my baby's samples to be used in future studies of pregnancy complications. I understand that my baby and I will not be identified in any way.

☐ I may give my permission for my and my baby's samples to be used in a non-identifiable fashion in future studies but wish to be contacted about the future study first.

☐ I do not give permission for my and my baby's samples to be used in the future and request that they are destroyed when the current study is completed.

15.0 **SIGN THIS FORM ONLY IF ALL OF THE FOLLOWING ARE TRUE:**

You have read the above information.

Your questions have been answered to your satisfaction and you believe you understand all of the information given about this study.

You have freely decided to take part in this research study.

You agree to the storage and future use of blood, urine, vaginal secretions, saliva, amniotic fluid, placenta, umbilical cord and membrane tissues, and DNA samples or you have indicated restrictions on the storage and future use on the previous page.

Subject: _____ Signature of Subject	Other: _____ Signature of Witness

__ __/__ __/ __ __ __ __ __ __:__ __ Month Day Year Hour Minute (24-hr clock)	_____ Signature of Person Obtaining Consent _____ Printed Name of Person Obtaining Consent _____ Printed Title of Person Obtaining Consent _____ Signature of Interpreter (if applicable)

YOU WILL BE GIVEN A COPY OF THIS SIGNED CONSENT FORM.

7 Expression Profiling Study

7.1 Definitions

For the expression profiling study, cases are defined as a caesarean preterm delivery between 20 weeks 0 days and 34 weeks 6 days (inclusive) following the presence or absence of the spontaneous onset of labor (see definition Section 2.5.7) either in the presence or absence of pPROM (see definition Section 2.5.6), while controls are defined as a term caesarean delivery between 39 weeks 0 days and 41 weeks 6 days following the presence or absence of the spontaneous onset of labor either in the presence or absence of PROM (see definition Section 2.5.5).

7.2 Eligibility Criteria

The Steering Committee and Protocol Subcommittee devoted substantial effort and discussions to determine the eligibility (inclusion and exclusion) criteria for this study. The overall rationale is to maintain scientific rigor in order to construct a study cohort in which the disease group will have the highest chance to share a common basis that differs from the controls, and to maximize the feasibility by enhancing the study population.

7.2.1 Group 1: Preterm delivery <u>without</u> labor (n = 20)

7.2.1.1 Inclusion Criteria

1. Cesarean delivery between 24 weeks 0 days and 34 weeks 6 days (Inclusive).
2. Delivery for maternal or fetal indications (preferably delivery for fetal distress).
3. Planned tubal ligation at cesarean if required by clinical site.

7.2.1.2 Exclusion Criteria

1. Spontaneous labor (see Section 2.5.7 for definition).
2. Induction of labor.
3. pPROM (see Section 2.5.6 for definition).
4. Maternal uterine anomalies.
5. Multi-fetal gestation.
6. Known aneuploidy or lethal fetal anomalies (see Section 2.5.8 for specifications).
7. Polyhydramnios (defined as Amniotic Fluid Index >=25 cm or Deepest vertical Pocket >=12 cm).
8. Serious maternal medical conditions (defined in Section 4.3).
9. Cervical Cerclage.

7.2.2 Group 2: Preterm delivery with labor (n = 20)

7.2.2.1 Inclusion Criteria

1. Cesarean delivery between 24 weeks 0 days and 34 weeks 6 days (inclusive).

2. Spontaneous labor (see Section 2.5.7 for definition).
3. Planned tubal ligation at cesarean if required by clinical site.

7.2.2.2 Exclusion Criteria

1. pPROM (see Section 2.5.6 for definition).
2. Maternal uterine anomalies.
3. Multi-fetal gestation.
4. Known aneuploidy or lethal fetal anomalies (see Section 2.5.8 for specifications).
5. Polyhydramnios (defined as Amniotic Fluid Index >=25 cm or Deepest vertical Pocket >=12 cm).
6. Serious maternal medical conditions (defined in Section 4.3).
7. Cervical Cerclage.

7.2.3 Group 3: Term Delivery <u>with</u> Labor (n = 20)

7.2.3.1 Inclusion Criteria

1. Cesarean delivery between 39 weeks 0 days and 41 weeks 6 days (Inclusive).
2. Spontaneous labor (see Section 2.5.7 for definition).
3. Planned tubal ligation at cesarean (this may vary by site).

7.2.3.2 Exclusion Criteria

1. PROM (see Section 2.5.5 for definition).
2. Maternal uterine anomalies.
3. Multi-fetal gestation.
4. Known aneuploidy or lethal fetal anomalies (see Section 2.5.8 for specifications).
5. Polyhydramnios (defined as Amniotic Fluid Index >=25 cm or Deepest vertical Pocket >=12 cm).
6. Other known fetal complications (IUGR, etc.).
7. Serious maternal medical conditions (defined in Section 4.3).
8. Cervical Cerclage.

7.2.4 Group 4: Term Delivery <u>without</u> Labor (n = 20)

7.2.4.1 Inclusion Criteria

1. Cesarean delivery between 39 weeks 0 days and 41 weeks 6 days (Inclusive).
2. Delivery indications for prior cesarean section or abnormal fetal lie.
3. Planned tubal ligation at cesarean (this may vary by site).

7.2.4.2 Exclusion Criteria

1. Labor (see Section 2.5.7 for definition).
2. PROM (see Section 2.5.5 for definition).
3. Maternal uterine anomalies.
4. Multi-fetal gestation.
5. Known aneuploidy or lethal fetal anomalies (see Section 2.5.8 for specifications).

6. Polyhydramnios (defined as Amniotic Fluid Index >=25 cm or Deepest vertical Pocket >=12 cm).
7. Other known fetal complications (IUGR, etc.).
8. Serious maternal medical conditions (defined in Section 4.3).
9. Cervical Cerclage.

7.2.5 Group 5: Preterm Premature of the Rupture of the Membranes with Labor (n = 20)

7.2.5.1 Inclusion Criteria

1. Cesarean delivery between 24 weeks 0 days and 34 weeks 6 days (Inclusive).
2. Preterm Premature Rupture of the membranes (see Section 2.5.4 for definition)
3. Labor (see Section 2.5.7 for definition).
4. Planned tubal ligation if required by clinical site.

7.2.5.2 Exclusion Criteria

1. Serious maternal medical complications (defined in Section 4.3).
2. Maternal uterine anomalies.
3. Multi-fetal gestation.
4. Known aneuploidy or lethal fetal anomalies (see Section 2.5.8 for specifications).
5. Polyhydramnios (defined as Amniotic Fluid Index >=25 cm or Deepest vertical Pocket >=12 cm).
6. Cervical Cerclage.

7.2.6 Group 6: pPROM without Labor (n = 20)

7.2.6.1 Inclusion Criteria

1. Cesarean delivery between 24 weeks 0 days and 34 weeks 6 days.
2. Preterm Premature Rupture of the membranes (see Section 2.5.4 for definition) .
3. Delivery indications for prior cesarean section or abnormal fetal lie.
4. Planned tubal ligation at cesarean (this may vary by site).

7.2.6.2 Exclusion Criteria

1. Labor (see Section 2.5.7 for definition).
2. Serious maternal medical complications (defined in Section 4.3)
3. Maternal uterine anomalies.
4. Multi-fetal gestation.
5. Known aneuploidy or lethal fetal anomalies (see Section 2.5.8 for specifications).
6. Polyhydramnios (defined as Amniotic Fluid Index >=25 cm or Deepest vertical Pocket >=12 cm).
7. Cervical Cerclage.
8. Induction of labor.

7.3 Screening for Eligibility

Eligible for screening are all patients who deliver singleton pregnancy via cesarean section with bilateral tubal ligation (the need for bilateral tubal ligation may vary by site). Eligible patients will be divided into the following groups:

1. Indicated Preterm deliveries between 24 weeks 0 days and 34 weeks 6 days due to fetal indications for delivery (Group 1).
2. Spontaneous Preterm deliveries between 24 weeks 0 days and 34 weeks 6 days due to spontaneous onset of preterm labor (Group 2).
3. Deliveries between 39 weeks 0 days and 41 weeks 6 days following the spontaneous onset of labor (Group 3).
4. Deliveries between 39 weeks 0 days and 41 weeks 6 days without evidence of labor. Most will be elective cesarean section or cesarean for abnormal fetal lie (Group 4).
5. Spontaneous preterm deliveries between 24 weeks and 34 weeks 6 days due to pPROM (Group 5).
6. Elective deliveries at 24 weeks 0 days to 34 weeks 6 days due to pPROM (Group 6).

The inclusion and exclusion criteria will be reviewed with the patient's chart. If an ultrasound examination has not been performed, one must be arranged prior to enrollment. Gestational age determination and assignment of the Project Gestational Age will be then performed based on the results of the ultrasound examination and patient's menstrual history according to criteria defined in Section 4.2 "Gestational Age Determination".

If a patient meets the criteria for enrollment and expresses interest to participate in the study, she will be told about the study design, its potential risks and benefits and asked to sign the informed consent and patient authorization (if applicable) forms. If she accepts, she will then be enrolled in the Study. A copy of the signed consent form(s) will be provided to the patient.

7.4 Study Procedures

7.4.1 Patient Data

Patient data will be collected through chart reviews and patient interviews. Data will be entered using the Demographics, Enrollment, Intake Forms (I, II, III), Current Medications, Ultrasound and Labs, Labor and Delivery Data Form, and the Neonatal Baseline as enclosed in the Manual of Operations. Stress, anxiety, and depression during the pregnancy will be assessed by the Psychosocial Questionnaire. In addition, the Adverse Event Form will be used as needed and is unscheduled. The adverse event form will only be filled out for neonates up to 28 days post delivery (during the neonatal period) who experience an adverse event as defined in the Manual of Operations.

1. Demographic information: parental age, race, ethnicity, etc.
2. Medical history: pre-pregnancy weight, height, medications, STD history, medical conditions, etc.
3. Social history: marital status, years of education, alcohol use and tobacco use.
4. Obstetrical history including outcome of all prior pregnancies and dates of termination.

5. History of preterm labor or premature rupture of membranes symptoms, evaluation or hospitalization for those symptoms in current pregnancy.
6. Medications in current pregnancy.
7. History of cervical evaluation (manually or ultrasonographically) or fetal fibronectin measurements.
8. Pregnancy complications.
9. Labor: type, indications, membrane status, induction method.
10. Delivery type and indications for cesarean section.
11. Neonatal outcome: sex, weight, length, Apgar score, neonatal complications, admission to intensive care unit, etc.
12. Family history of SPTB (father of child and his family members).

7.4.2 Samples

Samples will be sent from the Clinical Cores to the Analytical Core, and data will be sent from the Analytical Core to the DMSI Core (Manual of Operations Section 11).

1. Maternal blood (10 ml) will be collected for RNA expression profiling, proteomic/ metabolomic analyses, DNA, and measurements of markers of pregnancy exposures (cotinine, heavy metals, folic acid, stress, etc.).
2. Maternal saliva (2ml) will be collected for proteomic/metabolomic analyses. It will also serve as an alternate source of DNA and pregnancy exposures if maternal blood cannot be collected.
3. Maternal urine (5 to 10 ml) will be collected for proteomic and metabolomic analyses.
4. Amniotic fluid (10 ml) will be obtained in patients with intact membranes after uterine incision to study RNA expression and proteomic/ metabolomic analyses.
5. Cord blood (15 ml) will be obtained following delivery for DNA and RNA expression profiling, and proteomic/metabolomic analyses.
6. Neonatal saliva sample (via Oragene collection) will be obtained if cord blood could not be collected for fetal genotyping.
7. Placenta samples ($1cm^3$) will be collected for RNA expression, proteomic/metabolomic analyses, epigenetics, and histology.
8. Fetal membranes will be collected for RNA expression, proteomic/metabolomic analyses, epigenetics, and histology.
9. Decidual sample ($0.5\ cm^3$) will be collected for RNA expression, proteomic/ metabolomic analyses, epigenetics, and histology.
10. Fundal myometrial sample, lower segment myometrial sample, and cervical sample ($0.5\ cm^3$) will be obtained for RNA expression and proteomic/ metabolomic analyses, and epigenetics. Uterine biopsies must be obtained under direct visualization. If optimum visualization is not possible, then the biopsies will not be obtained. For fundal myometrium samples, the uterus must first be exteriorized.

7.4.3 Receiving Samples from the Clinical Cores to the Analytical Core

Barcoding samples will be performed at clinical sites. Samples will be identified (i.e., maternal blood/gestational age, maternal urine/gestational age) with the following assigned numbers at each clinical site:

- Alabama-Birmingham 2,501 – 2,550 series
- Texas-Galveston 8,501 – 8,550 series

- Utah 5,501 – 5,550 series

2. Penn tracking system – caTISSUE is an Oracle-based tissue bank designed by caBIG™. Samples will be linked to TrialDB through an Excel file with barcodes and patient IDs.
3. Shipping instructions (i.e., dry ice, etc.) for individual samples are provided in the Manual of Operations (section 8). All samples will be shipped to University of Pennsylvania.

7.4.4 RNA Acquisition

See Section 6.6.5.

7.4.5 Analysis

See Section 6.6.6.

7.4.6 Quality Assurance and Control of Profiling Data

See Section 6.6.7.

7.5 Summary Table for the Design and Data Collection

Table 4 provides an overall summary of the design, data, and sample collection for the expression profiling study.

Table 4. Summary Table for Expression Profiling Study

	Indicated PTD not in labor <35 wks with CS and BTL	**Spontaneous PTD in labor <35 wks with CS and BTL**	**Term deliveries in labor >39 wks with CS and BTL**	**Term deliveries not in labor >39 wks with CS and BTL**	**pPROM <u>with labor</u> <35 wks with CS and BTL**	**pPROM <u>without labor</u> 35 wks with CS and BTL**
N	20	20	20	20	20	20
Specimen Collection						
Maternal Blood	X	X	X	X		
Amniotic Fluid	X	X	X	X	X	X
Maternal Urine	X	X	X	X	X	X
Maternal Saliva	X	X	X	X	X	X
Cord Blood	X	X	X	X	X	X
Neonatal Oragene Saliva (if no cord blood)	X	X	X	X	X	X
Placenta	X	X	X	X	X	X
Fetal Membranes	X	X	X	X	X	X
Decidua	X	X	X	X	X	X
Uterus (fundal & lower segments)	X	X	X	X	X	X
Cervix	X	X	X	X	X	X
Clinical and Demographic Data						
Demographics	X	X	X	X	X	X
Enrollment	X	X	X	X	X	X
Intake Form (I,II III)	X	X	X	X	X	X
Current Medications	X	X	X	X	X	X
Psychosocial Questionnaire	X	X	X	X	X	X
Ultrasound and Labs	X	X	X	X	X	X
Labor and Delivery Data Form	X	X	X	X	X	X
Neonatal Baseline	X	X	X	X	X	X
Adverse Event Form	*[1]	*[1]	*[1]	*[1]	*[1]	*[1]

*[1] Unscheduled forms to be used as needed

7.6 Statistical Analyses

The objective is to identify differential expression profiles (RNA, protein and metabolites) in six groups of 20 patients: in labor and not in labor at term and preterm before 34 weeks, pPROM and spontaneous preterm labor, and pPROM and indicated preterm delivery. The sample sizes for this study are determined by empirical experience, in which other studies cited above have used similar sample sizes to achieve similar objectives, and affordability.

The expression profiling technologies and statistical analysis are analogous to the longitudinal study as presented in Section 6.8.

7.7 Human Subjects - Template Consent Form

1.0 **You are invited to take part in an important research study about spontaneous premature birth.** Spontaneous preterm birth is the spontaneous delivery of a baby more than three weeks before its due date. Researchers from five areas in the United States are trying to learn about the causes of spontaneous premature birth. The researchers hope that this study will help prevent spontaneous premature births in the future. This research study is funded by the National Institutes of Health (NIH) / National Institute of Child Health and Human Development (NICHD). It is called the Genomic and Proteomic Network for Preterm Birth Research. Approximately 2,500 pregnant women from across the nation will take part in the study.

You are being asked to take part in this study because your doctors have recommended that you deliver your baby by cesarean section and are either:

a) at full term (between 3 weeks before your due date and 1 week after your due date) and in labor,
b) at full term and NOT in labor,
c) at less than 35 weeks gestation (more than six weeks before your due date) and in labor, or,
d) at less than 35 weeks gestation and NOT in labor.
e) at less than 35 weeks gestation due to preterm premature rupture of the membranes
f) elective deliveries at 35 weeks 0 days due to preterm premature rupture of the membranes

This part of the study will recruit 80 such women from across the nation, with 20 in each of these six groups. We would like to tell you about this study. This study is trying to learn more about the causes of spontaneous preterm birth.

This study will use several powerful new research methods to study the problem of spontaneous premature birth. Genomics is the study of specific genes that are associated with any given medical problem, in this case, spontaneous premature birth. Genes are passed on from our parents and are the building blocks of our bodies that tell our bodies how to work. Proteomics is the study of proteins (produced by specific genes) that are seen with a given medical problem, in this case, spontaneous premature birth.

Metabolomics is the study of small molecules that are associated with a given medical problem, in this case, spontaneous premature birth.

What we learn from this study could give us valuable information about this common and serious problem that could lead to more effective treatment and/or prevention.

2.0 **What will happen if I join this study?** For this study of spontaneous premature birth, we would:

a) ask you some questions,
b) collect facts from your medical records and those of your baby,
c) obtain your blood, urine and saliva taken within a day after you deliver,
d) obtain a sample of amniotic fluid at the time of your delivery,
e) obtain a small piece of tissue from the upper part of your uterus, the area of your uterus where the cesarean is performed (called the lower uterine segment) and the cervix,
f) examine and obtain samples of your placenta (afterbirth), umbilical cord and membranes,
g) obtain a sample of blood from the umbilical cord after delivery, and,
h) obtain a cheek swab from your baby after birth.

Genomic, proteomic, and metabolomic testing would be performed on the samples that are collected and the findings would be compared with the outcomes of the women and babies enrolled in the study. All information will be strictly confidential and you would not be identified in any way. More details on how we will protect your privacy are given in questions 10.0 and 11.0 below.

2.1 **What kinds of questions will I be asked?** A trained member of our research team will ask you some questions about your pregnancies, your medical history, your family history, and your day-to-day life and work, including smoking, drinking, and drug use. Most of this interview will be done while you are in the hospital. However, the whole interview could also be done after you have left the hospital. The initial interviews should take about 30 to 45 minutes.

2.2 **How will my medical records be used?** We wish to look at the medical records of your pregnancies, including records of your doctor and from your hospital visits. We wish to collect information from your records to help us better understand spontaneous labor, both premature and full term. This request to access your records (and those of your baby) will comply with HIPAA, which is a law protecting the privacy of your health information.

2.3 **How will my baby's medical records be used?** We would like to collect information from your baby's medical records from birth to the time that your baby leaves the hospital. We will use this information to look at differences between babies that are born prematurely and those that are born full term (within three weeks of the due date).

2.4 **How will my samples be used for this study?** In addition to your usual blood tests, an additional 10 milliliters (about two teaspoons) of blood would be drawn from a vein in your arm while you are in the hospital. If possible this will be done at a time when you are having another blood sample drawn as a part of your care. This blood sample will be sent to a specific research laboratory at the University of Pennsylvania (Philadelphia,

PA), where specialized testing will be done that will look for genes, protein and metabolites that might tell us whether or not spontaneous premature birth will, or will not, occur.

We would obtain a urine sample while you have a catheter in your bladder for your cesarean. The urine sample would also be studied to see if any protein and metabolites could be identified that could predict whether or not spontaneous premature birth might occur.

At some time during your hospital stay we would also ask you to provide a sample of saliva. The saliva sample would be studied to see if any proteins or metabolites could be identified that could predict whether or not spontaneous premature birth might occur. Saliva is also a very good source of DNA (deoxyribonucleic acid), which is the substance that contains our genes. Most genes do not differ among people, but some genes and combinations of genes are unique to each person. If you agree, DNA from you, your placenta and your baby will be used to see whether the genes of mothers and babies with and without spontaneous premature birth are different.

Your baby's blood will be obtained after delivery from the discarded umbilical cord. We will only save your baby's blood if you agree. Otherwise it will be thrown away. If a blood sample cannot be obtained from your baby's umbilical cord, we would ask your permission to obtain a cheek swab sample from your baby in order to provide a saliva sample that would be used to study your baby's DNA.

We will store a sample of your and your baby's DNA. These samples will only be used to study spontaneous premature labor and pregnancy-related diseases. These samples would also be sent to the University of Pennsylvania, where DNA would be studied in hopes of finding specific genes that might predict whether or not spontaneous premature birth will, or will not, occur.

The amniotic fluid and tissue samples obtained from your uterus and cervix will be used to look for genes or other markers (proteins, metabolites) that can tell us how the uterus contracts and the cervix dilates in the course of labor.

2.5 **What will happen when the placenta (afterbirth), umbilical cord and membranes are examined?** The placenta will be examined in both normal and additional ways. We want to examine the placentas of all women in the study so that we can compare placentas of premature babies to those of babies who are born full term. We will take pictures and measurements and also examine the placenta under the microscope. We will also perform tests on your placenta that look for signs of infection or biochemical differences (i.e., genes, proteins, metabolites) that might occur more often in spontaneous premature births.

2.6 **How will my samples be stored?** All samples collected during this study will be stored in a secure facility at the University of Pennsylvania for five years and then in a secured facility (to be determined) until the samples are no longer useable or until destroyed. To protect you and your baby, samples will only be labeled with a study identification number (study ID). We will keep your study ID linked with your name in our local records (locked and secure) for up to ten years after the end of the study. If you decide that you want the stored samples destroyed while we have this link, you can let us know in writing and we will have the samples that are labeled with your study ID destroyed.

The stored samples will only be released for other studies if an independent committee that reviews research on people (called an Institutional Review Board) has approved the research. Research findings on stored samples will not be reported back to you. Stored samples will only be used for the study of spontaneous premature birth, and pregnancy or baby-related diseases unless you are contacted for your permission to use the samples for other purposes (see question 2.7 below).

2.7 **Could I be contacted for future studies?** A research person authorized by this study may contact you in the future if there are any new research questions. If you ask us not to contact you for future studies, your last contact with us will be at the time that you and your baby are discharged from the hospital.

2.8 **Could information from my baby and me be used in other future studies?** The results of your and your baby's research tests (including DNA tests) and related medical records will be placed in an electronic database that could be used for other research studies in addition to this study. These studies would involve only pregnancy or baby-related diseases. Your privacy will be strictly maintained, since information that would identify you or your baby would not be available. However, in certain situations a researcher might want to contact you for additional information that would be useful only for other research purposes. In these cases, the researcher would make a request to an oversight committee. This committee would review the request and either deny or approve the researchers request for your contact information. The researchers would have to guarantee that they would not release either your contact information or your database records before permission would be granted. This would only apply if you have given prior permission to be contacted in the future by a researcher. If you have not given your prior permission the request would be automatically denied.

3.0 **Risks and Benefits from Study Participation**

3.1 **What discomforts or risks could I have from participating in the study?**
Occasionally drawing blood may cause pain and bruising. Rarely, drawing blood also causes some people to faint or feel light-headed and causes blood clots or infection. It is possible that you might have bleeding or infection at the places where the uterus and cervix samples are obtained. Research staff members are available to help you get medical care if you need it because of the study.

It is possible that some of the interview questions could make you feel uncomfortable or upset. You do not have to answer any questions that you do not want to answer. You may stop the interview at any time without affecting your medical care.

If you agree to be in this study we will be obtaining information from blood and placental samples about whether you and your baby have been exposed to different kinds of drugs. We will not provide the results of these research tests to any law enforcement or social agency. We have obtained a Certificate of Confidentiality from the federal government so that we cannot be forced to release this information to anyone (see question 11.0 below).

Laboratory tests, including DNA testing, carry some unique risks. The concern is that test results might someday be used against a person. However, the risk for something like that happening as a result of being in this study is extremely small. First of all, the researchers will be very careful to ensure that only authorized people have access to the samples and

your personal information. Secondly, the types of tests we plan to do cannot, at the current time, be directly linked to an increased risk of a disease, for example cancer or diabetes. In addition, the results will not be placed in your medical record. You will not be given any information at all about the results of the tests we are doing, because even if we are successful in identifying protein and metabolites, genes or other types of chemicals, we will still not have enough information to directly benefit you or your family. We will only use the tests from you and your baby to study spontaneous premature birth and pregnancy or baby-related diseases unless you give us separate permission to study other diseases.

There may be risks while participating in this study that are unknown.

<<LOCAL SITES MAY ADD INFORMATION REGARDING LIABILITY, ETC., PER LOCAL IRB REQUIREMENT HERE.>>

3.2 **Will I benefit from joining the study?** There are likely no direct benefits to you from participating in this study. However, we hope to learn how to reduce the likelihood of spontaneous premature birth. You could have the satisfaction of knowing that you have given information that may help prevent spontaneous premature births for other pregnant women in the future. General information from this research will be published in medical journals, but will not be relayed to you or benefit you directly.

4.0 **Do I have to join this study? What are my other choices?** You do not have to be in this study. You may choose to participate in some, but not all, parts of the study. Your medical care will not be affected if you choose not to take part in this study; you will still be able to get all appropriate medical care without being in the study. Your choice about participating in this study will not affect your eligibility for any health plan or any health plan benefits or payments.

5.0 **Withdrawing from the Study**

5.1 **What if I change my mind about how I want to take part in the study?** Even after you agree to be in this study, you can change your mind at any time and pull out of the study. You can refuse to answer a question or refuse to take part in any part of the study. If you change your mind, your medical care and health plan eligibility and benefits will not change. At any time you can also request that your samples be destroyed.

5.2 **Can the researchers remove me from the study?** You may withdraw from the study at any time without penalty. Likewise, Dr. XXXXX and HIS/HER associates or the National Institutes of Health can withdraw you without your approval. A possible reason for withdrawal could be the early termination of the study by the National Institutes of Health.

6.0 **If I have expenses because I took part in this study, can the study help pay these costs?** If you agree to take part in this study, usual travel and child care expenses (<<UP TO $??>>) that result from your being in this study will be paid. <<ADDITIONALLY, YOU WILL RECEIVE $?? AS COMPENSATION FOR THE TIME YOU SPEND ANSWERING QUESTIONS. (SENTENCE ONLY NEEDED IF REIMBURSEMENT FOR TIME IS OFFERED AT THE SITE. SOME SITES MAY SPLIT PAYMENT BETWEEN THE TWO PARTS OF THE INTERVIEW. SUCH DETAILS WILL BE

GIVEN HERE. SEE PROTOCOL FOR GUIDELINES. COMPENSATION FOR TIME PLUS REIMBURSEMENT FOR EXPENSES SHOULD NOT EXCEED $150.)>>

7.0 **Will joining this study increase my health care costs?** No. Your routine medical care and routine testing will be billed to your insurance company or to you the same as your other medical care. The costs of tests that are performed only for purposes of the research study will not be charged to you or your insurance company.

8.0 **What will happen if I get a bill for research tests by mistake?** If you receive a bill for any study-specific test that is not paid by your insurance company or Medicaid or if you are not sure if the bill is for a study-specific test, call the study team at <<PHONE NUMBER>> and they will ask you to send a copy of the bill to the following address to find out if this test was done for the study. If it was done because of the study, the study will pay for it.

<<NAME AND ADDRESS OF PI>>

9.0 **What do I do if I have questions or need to report a problem?** If you have questions now, feel free to ask them to the study staff. If you have more questions later, please contact

<<NAME AND CONTACT INFORMATION OF APPROPRIATE FIRST CONTACT>>.

If he/she is not available, you may contact

<<NAME AND CONTACT INFORMATION OF APPROPRIATE SECOND CONTACT>>.

<<PROVIDE INSTRUCTIONS FOR AFTER-HOUR CONTACTS AT YOUR INSTITUTION>>

The committee that reviews research on human subjects (Institutional Review Board) at <<NAME OF INSTITUTION>> will answer any questions about your rights as a research subject. They can be contacted at <<IRB CONTACT NUMBER (AND CONTACT PERSON IF AVAILABLE)>>.

10.0 **How will my privacy be protected?** The researchers are taking many steps to protect the privacy of you and your baby. First, we will use a number to identify research records and tissue and blood samples, instead of your name. If you agree to take part in the study, a study ID number will be assigned to you and your baby. This same number will be used to identify medical information kept by the study. Any records linking your name to the study ID will be stored in a locked (password protected) file.

All copies of your medical records and the links between your name and the study ID will be destroyed within ten years of the end of the study. After that time, your blood or tissue samples and information from your interview and medical records collected for the study will be marked with your study ID number only. Even though the information could still be used for research, it will no longer be possible to link it to you. Any results of this study reported in medical journals or at meetings will not identify you or your baby.

Only those who work with this study will be allowed access to your information. However, representatives from the National Institutes of Health may inspect and/or copy the records that identify you.

Results of this study may be published. However, your name and other identifying information will be kept private.

11.0 **What is a Certificate of Confidentiality?** To help us further protect the privacy of you and your baby, the researchers have obtained a Certificate of Confidentiality from the National Institutes of Health (NIH) / National Institute of Child Health and Human Development (NICHD). This Certificate means that the researchers cannot be forced (for example by court order) to disclose any information that might identify you to any federal, state, or local court. A Certificate of Confidentiality does not prevent you from voluntarily giving information to others about yourself or your involvement in this research. You should know that we may provide information to your health care providers if we suspect that you may harm yourself or others. We will not release any information collected as part of the research regarding use of illicit drugs and testing for drugs done on samples collected for the research.

12.0 **If I agree to be in all parts of this study, what does that include?**

Full participation in this study includes all of the following:

- An interview
- Collection, review and storage of medical records
- Perform tests on your blood and saliva samples that were obtained within a day of your delivery
- Obtain an amniotic fluid sample as well as samples from the upper and lower part of the uterus and the cervix at the time of your cesarean section
- Examine and obtain a sample of your placenta, umbilical cord and membranes
- Obtain your baby's blood or saliva sample within a day of your delivery
- Genetic, proteomic and metabolomic studies on you and your baby
- Storage and future use of blood, urine, saliva, tissue from the uterus, cervix, placenta, umbilical cord and membranes, and DNA samples
- Research staff possibly contacting you later for other research questions.

13.0 **If I still want to join this study but would like to avoid some parts of it, what should I do?** Agreeing to do all parts of this study will allow the researchers to gather the most useful information to study spontaneous premature birth. However, if you do not want to do all parts of the study, you can refuse to do any one or more parts of the study.

Parts of this study that you do not wish to do are:

None ☐ Restrictions, if any: _____

14.0 **Will my information and samples be used in other studies of pregnancy complications?** The researchers may want to study other pregnancy complications using

the information and samples that have been collected in this study. This could only happen if you give the researchers permission. You do not have to give them your permission but we would like you to consider initialing one of the three boxes if you are comfortable doing so. If you do not initial a box giving the researchers this permission, your samples will not be used.

☐ I give my permission for my samples to be used in future studies of pregnancy and baby complications. I understand that my baby and I will not be identified in any way.

☐ I may give my permission for my and my baby's samples to be used in a non-identifiable fashion in future studies but wish to be contacted about the future study first.

☐ I do not give permission for my and my baby's samples to be used in the future and request that they be destroyed when the current study is completed.

15.0 **SIGN THIS FORM ONLY IF ALL OF THE FOLLOWING ARE TRUE:**

You have read the above information.

Your questions have been answered to your satisfaction and you believe you understand all of the information given about this study.

You have freely decided to take part in this research study.

You agree to the storage and future use of blood, saliva, placenta, urine, amniotic fluid, umbilical cord and membrane tissues, and DNA samples or you have indicated restrictions on the storage and future use on the previous page.

Subject:	Other:
_____ Signature of Subject __ __ /__ __ / __ __ __ __ __ __:__ __ Month Day Year Hour Minute (24-hr clock)	_____ Signature of Witness _____ Signature of Person Obtaining Consent _____ Printed Name of Person Obtaining Consent _____ Printed Title of Person Obtaining Consent _____ Signature of Interpreter (if applicable)

YOU WILL BE GIVEN A COPY OF THIS SIGNED CONSENT FORM

8 Data Collection and Management

8.1 Assignment of Study Identification Number

All study patients will be assigned an identification number and identified by study number only, and the list of study patient names and study numbers will be maintained in a locked file cabinet in the PI's offices at the clinical cores. Sequential, site-specific pre-numbered sealed packets will be available at each site from the Analytic Core. At the time of study number assignment, the site PI will open the sealed packet pre-labeled with the study number to which the patient has been assigned. This packet will include data sheets/booklets and other necessary materials such as tubes for blood or buccal swabs pre-labeled with a barcode corresponding to the patient's study number. These will be prepared and sent by the Analytical Core. All biological analyses will be identified only by barcode corresponding to study number. Upon arrival of the sample at the Analytical Core for genotyping and profiling, each sample will be subsequently identified by barcode stickers.

8.2 Data Collection Forms

Data will be collected on standardized forms on which nearly all responses have been precoded. Each form is briefly described below:

- Demographics Form
- Enrollment Form
- Intake Forms I, II, III.
- Current Medications Form
- Ultrasound and Labs Form
- Return Form
- Psychosocial Questionnaire
- Hospital Admission Form
- Neonatal Baseline
- Labor and Delivery Form
- Adverse Event Form

8.3 Centralized Data Management System

Data will be hosted on a secure HIPAA-compliant database server located in the Yale Information Technology Services. The front end will be maintained on a Web server hosted at the Yale Center for Medical Informatics and will utilize the TrialDB software.

All created and updated forms are transmitted from the clinical sites to the DMSI Core through the network web server. Biological samples will be sent from the Clinical Sites to the Analytical Core, and data will be uploaded by the Analytical Core to the network database server after quality control assurance.

8.4 Data Entry

Data will be entered into the system in two ways:

1. Data can be entered directly on laptops or desktop PCs from the individual clinical sites. PCs will require Internet Explorer with a minimum of 512 MB RAM. Apple computers are currently not supported by TrialDB for data entry purposes, although this support will be available when TrialDB is fully ported to the Microsoft .NET platform within the next year and a half. Forms can be printed out by the individual sites, if needed, for subsequent data entry.
2. Bulk import of legacy data can be retrieved from the local systems. Individual sites will need to have their systems tailored to their needs while conforming to the network data entry system since each site will have its own data formats and methods for variable naming.
    - TrialDB is very capable of mapping variables in local data sets into the master data set. It can also handle automated recoding of coded/enumerated data that is of the same granularity as that in the master. However, the clinical sites will need to ensure that the variables have the same semantics and granularity as defined in the master set and recode if necessary. When the data entry forms are agreed upon by the network and fully developed, we will have a standard set of questions and answers. The individual sites must ensure that their answers are consistently coded as the standard ones. The DMSI core will provide necessary training on this issue.
    - If necessary, each clinical site will be responsible for the curation-intensive task of transforming narrative text (clinical notes, for example) into individual coded variables.
    - There are two methods for performing bulk import: incremental updates and re-import after purging old data. Incremental updates are more sophisticated, but require that all local sites maintain sophisticated databases that time-stamp every record with the dates of creation and last change. Software that enables the user to time-stamp data rows is generally found in rare cases of sophisticated programming. A more practical tactic for local data management will utilize a re-import option.
    - Prior to importing bulk data, checks will be performed to verify the consistency in the interactive mode. A list of errors will be reported and given to each site. It is the responsibility of the clinical sites to review the list and correct errors from the original data. This will prevent a reoccurrence of errors during a data re-import.

- At any time, each site will have the ability to download the complete data on their own patients in the form of tab-delimited files accompanied by a data dictionary (“extracts”). These can be imported into microcomputer, high-end database systems, spreadsheets or statistical packages. Currently, SAS and SPSS dictionaries are supported if the statistical-package export option is chosen.

TrialDB also generates numerous tracking reports, such as Patient Accrual, Missing forms, problem data, and audit trail. In addition, it is capable of facilitating data monitoring by displaying the data summary. The DMSI PI, in collaboration with the clinical sites, will be instrumental in defining the need for additional reports. We will particularly monitor the

distributions of race, maternal age groups, and parity within each site between cases and controls.

8.5 Data Flow and Sample Tracking

Data Flow will proceed as follows:

Clinical Data

Data Collection forms include: Enrollment Form, Intake Forms I, II, III, Current Medications Form, Ultrasound and Labs Form, Hospital Admission Form, Symptomatic Admission Form, Return Form, Psychosocial Questionnaire, Labor and Delivery Form, and Neonatal Baseline. The Adverse Event Form will also be used as needed. The information extracted from these forms will be entered directly over the internet into TrialDB from the research sites. Any site will be able to download, at any time, the complete data on its own patients (all CRFs + demographics) as a TrialDB data extract. In addition, the research sites as well as NICHD personnel will be able to run a variety of standard monitoring reports (e.g., patient accrual) on their data.

Sample data

With the network investigators, the DMSI Core can design a standard system for patient identification and send it to all sites to use. Basic sample identification data will be sent from the clinical sites to the Analytical Core at the University of Pennsylvania, along with the samples themselves. Currently, we assume that the Analytical Core has the capability to provide access to sample tracking data via the internet for use by the clinical sites. TrialDB has the ability to track sample shipments across institutions. This feature can be provided as part of the present GPN protocol. This requires that information regarding samples associated with a patient is entered into TrialDB. At that point, the Analytical Core can access this information over the internet through a standard report. Data generated at the Analytical Core will be sent to the DMSI Core using standard formats for Microarray and Proteomics data. The DMSI Core will make the data available to the clinical sites via the internet through the network web server or other Yale servers such as the Yale Microarray Database (YMD) and the Yale Protein Expression Database (YPED). The DMSI Core will implement cross-links between TrialDB and other genomic and proteomic databases such as YMD/YPED, based on the Biospecimen/Sample IDs.

Figure 12 shows the proposed sample flow and tracking scheme. In our Steering Committee meeting that was held on July 17-18, 2006, we decided to extend the TrialDB sample tracking module to meet the sample tracking needs of our network, while the Analytical Core will continue to use caTissue to perform their own sample tracking for the genomic/proteomic experiments. Also in consideration are other forms of sample tracking such as paper forms through fax and electronic forms through email communications. Prior to sending off patients' DNA samples to the Analytical Core for experimental processing, the staff at each clinical site will enter basic sample description/information upon receipt of the patient samples. We will establish a standard interface to allow synchronization of basic sample data between the Analytical Core's caTissue system and TrialDB sample tracking system. The staff at the Analytical Core will enter more detailed description of the samples (received from the clinical

sites) into caTissue during their experimental process. Figure 13 gives the schematic diagram of caTissue. The figure shows the ability of caTissue to capture information corresponding to different types of samples (e.g., blood serum and tissue samples), which are used for different types of genomic/proteomic experiments (e.g., DNA microarrays and 2-D gel). Although the current needs of the network do not require detailed sample description, the potential advantage of using caTissue is the ability of integration with other caBIG compliant software systems. If such needs arise in the future, our current approach can still allow us to switch to use caTissue, since our proposal includes synchronization of sample data between TrialDB and caTissue. The data generated from the genomic/proteomic experiments will be stored in the network server, and/or on YMD and YPED servers, all of which will be linked to the sample information stored in TrialDB.

We expect that the DMSI Core will store the processed/derived data generated by the Analytical Core rather than the voluminous raw signal files. It is expected that raw data will be archived off-line at the Analytical Core. However, if the network prefers, raw data can be archived at the DMSI Core at an additional cost.

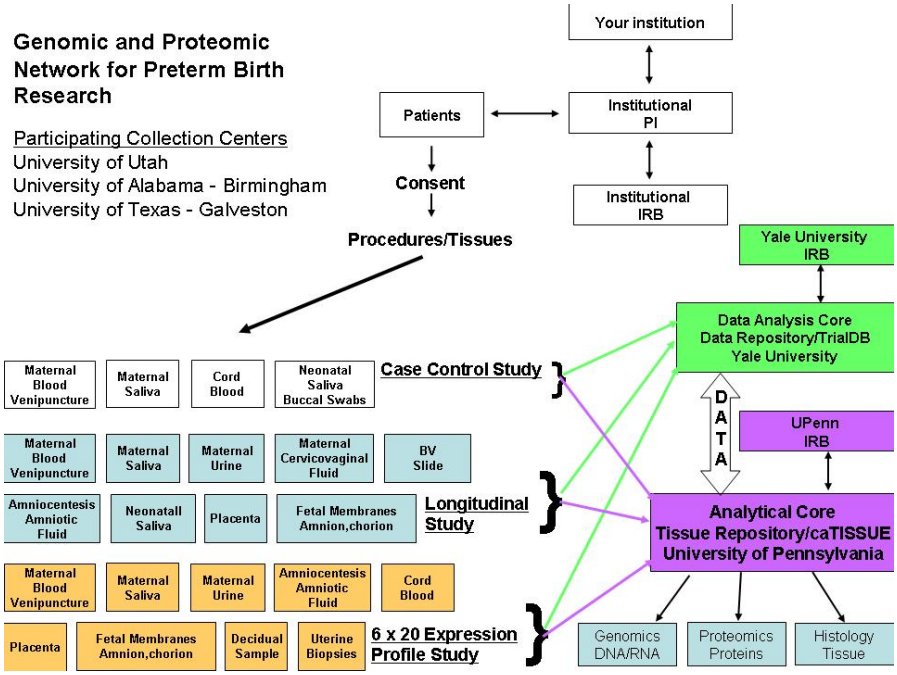

Figure 12. Data Flow Chart

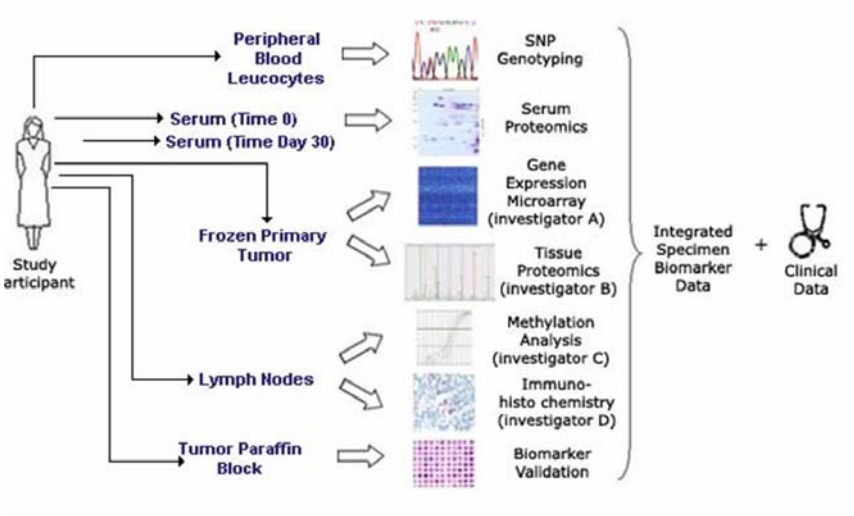

Figure 13. CaBIG Tracking System

8.6 Reporting and Performance Monitoring

Generating administrative and statistical reports are an important role of the DMSI Core. The DMSI Core will present regular reports to the Steering Committee and the Advisory Board. These include:

- Monthly recruitment reports: reports of the number of women screened and enrolled by month and by clinical site will be provided monthly to the members of the Steering Committee.
- Steering Committee reports (three times a year): a report detailing recruitment, baseline patient characteristics, data quality, incidence of missing data and adherence to study protocol by clinical site, will be provided to the members of the Steering Committee prior to the Steering Committee meeting.
- Data quality status reports: every four weeks, each clinical site will receive a data quality status report specific to that site.
- Advisory Board reports: for every meeting of the Advisory Board, a report is prepared which includes site performance information with respect to data quality, timeliness of data submission and protocol adherence.
- Statistical reports include reports to the steering committee and advisory board from the data analysis, and special reports for scientific manuscripts.

Our database system has a reporting module that automatically publishes a variety of administrative reports to our website for the internal Network access, which can be generated by the user on demand, in addition to periodically scheduled reports. The latter reports will include a summary of enrollment by clinical cores, ethnicity, and gender, reasons for study exclusion, protocol compliance percentages, and data quality. It is assumed that each clinical site would indicate how many subjects it expects to enroll per protocol per year and how many births would be minorities. We would propose that the Network develop practical procedures to ensure that the clinical sites accrue patients on schedule with representative gender and minorities and collect high quality clinical data.

Statistical Reports will be generated in SAS. Reports are provided for data safety and monitoring reviews, and for final analysis of study results in preparation for scientific publications. The content of the interim reports will be very complete, and will serve as the template for the final

report of each study, which in turn will form the basis of the publication of the results. Our proposed reports to the Advisory Board would include the following: a protocol description and history; accrual rates; site performance in terms of accrual; eligibility; protocol violations; data accuracy and minority representation; patient characteristics by treatment and site; and the rate of adverse experiences.

8.7 Quality Control and Assurance

The data are automatically edited on an intra-form basis for missing, out of range and inconsistent values. After review at the DMSI Core, edit printouts are returned to the appropriate clinical and the Analytical Cores for correction or clarification on a weekly basis. In regular intervals, the DMSI Core will run audits on the entire database or on a specific subset of data to compare data across forms (inter-form). These reports are also submitted to the other sites for corrections.

Furthermore, quality control of data will be handled on three different levels. The first level takes place at the clinical site where the data entry personnel ensure the data accuracy with nurse coordinators and the site PI. The second level is the real-time logical and range checking built into the web-based data entry system. The third level is the remote data monitoring and validation that is the primary responsibility of the data manager and programmer at the DMSI Core. Paper forms from a randomly selected 1% of the recruited patients will be sent to the DMSI Core to validate the consistency between the paper forms and the data in the databases. The data manager will conduct monthly comprehensive data checks. The data manager will query a site until each irregularity or identified error is resolved.

See Manuals of Operation for more details.

8.8 Data Entry Training, Certification and Support

Since nurse coordinators will be the primary users of TrialDB, the user interface will be adjusted to their needs.

The training will proceed as follows: a designated person from each institution will be trained in the operation of the user interface for data entry, browsing and reporting. Intensive training will take place primarily by internet video conferences if feasible to reduce the cost. However, the DMSI Core can coordinate in-person training. After the training, certification will be issued to qualified personnel. They will be responsible for training others within their institution and will be the primary point of contact with the TrialDB support team. All personnel will be certified using the same process designed by the data management team.

All requests from an institution will be directed through a designated contact. Requests may include reports that are designed to facilitate local workflow. After the network approves the budget, a DMSI specialist will provide support by phone to the site designated person and/or by e-mail between the hours of 8:00 am - 5:00 pm EST Mondays through Fridays. In all cases, the TrialDB team, consisting of Drs. Nadkarni and Cheung, our programmer, and the specialist, will respond within a maximum of 24 hours, though corrective actions may take longer if the questions and problems require a substantial correction of the TrialDB system. See the Manual of Operation for additional details.

9 Study Administration

9.1 Organization and Funding

This study is funded by the National Institute of Child Health and Human Development (NICHD), and conducted by the Genomic and Proteomic Network for Preterm Birth Research. The Network consists of the following: 1) a Clinical Core, comprising three clinical sites, responsible for subject recruitment and specimen collection; 2) an Analytical Core responsible for genomic and proteomic analyses; and 3) a Data Management, Statistics, and Informatics (DMSI) Core responsible for central data collection, analysis, and management; information technology; and coordination of the administrative activities of the Network.

The Clinical Core consists of three clinical research sites that will work cooperatively with one another. This core is responsible for recruitment of human subjects, collection of relevant clinical information, and collection and temporary storage of biological specimens. The Clinical Core is also responsible for data entry into the DMSI Core databases. The Analytical Core is responsible for "state-of-the-art", high-throughput DNA, RNA, and protein and metabolite analyses, as well as entry of data into the DMSI Core databases. This core is also responsible for the receipt and storage of the biological specimens collected from the various clinical sites. In addition, the Analytical Core is responsible for the computational analysis of genomic, proteomic and metabolomic data and synergistically interacting with the DMSI Core in performing genomic, proteomic and metabolomic statistical analyses. The DMSI Core is responsible for central data collection, analysis, and management. It will integrate clinical data with genomic, proteomic and metabolomic data. It is responsible for the statistical analysis of data aided, in part, by the Analytical Core. It is also responsible for creating and maintaining an internal and a public web-based database. In addition, it acts as the administrative coordinating center for the Network.

Each of the funded institutions is represented by a Principal Investigator.

9.1.1 Participating Clinical Cores

The Principal Investigators of the clinical cores have agreed to abide by the study protocol, to have the necessary staff, facilities and equipment and to ensure the proper conduct of the study at each of their sites including: recruitment of patients as specified in the protocol, accurate data collection, data entry to the network web-based database server, and the transmission of information to the Steering Committee.

9.1.2 Analytical Core

The Principal Investigator of the Analytical core has agreed to abide by the study protocol, to have comparable staff, facilities and equipment and to ensure the proper conduct of the study at the Analytical Core including: accurate sample analysis and data collection, uploading of data to the network web-based database server, and the transmission of information to the Steering Committee.

9.1.3 Data Management, Statistics, and Informatics Core (DMSI)

The DMSI Core is responsible for all aspects of biostatistical design, analysis and data management of the study, in addition to the interim and final statistical analyses and preparation of publications based on the study results.

9.1.4 NICHD

In addition to its role as funding agency, the NICHD participates in the activities of the Network, including the development of protocols, administration and conduct of the studies, and preparation of publications.

9.2 Management and Guidance Structure

The management and guidance structure consists of a Steering Committee and Advisory Board (AB). The Steering Committee consists of the Principal Investigators, participating NICHD staff, and other appropriate members. This committee is responsible for the formulation of new projects, review of proposed projects, prioritization of workload, consideration of long-term strategic issues, policy and budgetary matters, and development of guidelines for operation of the Network. The AB consists of individuals not affiliated with the Network and is responsible for providing technical and operational advice.

9.2.1 Steering Committee

The Steering Committee is a voting body that is solely responsible for protocol development. It is advised by the AB and, at times, consultants to the Steering Committee. A majority vote is required for acceptance of a proposed item. The Steering Committee's primary responsibility is to formulate and conduct protocols, as well as to create manuscripts for publication of the results of completed protocols. The Steering Committee is comprised of the Principal Investigators (voting members), two designated NICHD Project Scientists from the Pregnancy and Perinatology Branch (who will share one vote), and other non-voting and non-NIH members who are deemed necessary for the conduct of the Network. In addition, an individual not affiliated with the Network acts as chair of the Steering Committee. The chairperson will be a voting member only in case of a tie vote. The specific responsibilities of the Steering Committee are as follows:

- Development of guidelines and procedures for the operation of the Network
- Formulation of new projects
- Review of proposed projects
- Prioritization of workload
- Consideration of long-term strategic issues
- Policy and budgetary matters

9.2.2 Advisory Board

The Advisory Board (AB) is responsible for providing technical and operational advice to the Steering Committee. This board is strictly advisory and consequently their opinion is not

binding on the activities of the Network or the NICHD. Nonetheless, advice given by the board will be heavily weighted by the Steering Committee and the NICHD in the decision making process. The board was chosen by the NICHD, and is composed of individuals not associated with the Network or its members. The board includes members with expertise in premature birth, genomics, proteomics, molecular epidemiology, bioinformatics, and bioethics. The Chairperson of the Steering Committee, the principal investigator of the DMSI core, and the NICHD Project Scientists attend AB meetings to provide information as needed. Additional members participate based on the need for specific expertise. The specific responsibilities of the AB include evaluating:

- Scientific merit of protocols
- Overall Network operations
- Safeguards for subject/data confidentiality
- Quality control and assurance

9.2.3 NICHD Project Scientists

There are two NICHD Project Scientists on the Steering Committee. The specific responsibilities of the Project Scientists are as follows:

- Assistance in the identification of important areas of study
- Assistance in the development of study protocols
- Assistance in the development and review of capitation-based budgets including the identification of study costs and special institutional needs
- Assistance in the review and evaluation of each stage of the program before subsequent stages are started, in conjunction with the Steering Committee and the Advisory Board
- Assistance in the efficient conduct of the study, including ongoing review of progress; possible redirection of activities to improve performance and cooperation; and frequent communication with other members of the Steering Committee
- Participation on the Steering Committee and all active subcommittees
- Recommend consultants for appointment to the Steering Committee on an as-needed basis

9.2.4 NICHD Project Officer

The NICHD has appointed a Project Officer, apart from the Project Scientists, who will:

- Carry out continuous review of all activities to ensure that the objectives are being met and that all regulatory, fiscal, and administrative matters are handled according to NIH guidelines
- Have the option to withhold support to a participating institution if technical performance requirements are not met
- Perform other duties required for normal program stewardship of grants

9.2.5 Protocol Subcommittee

For each study undertaken by the Network, a subcommittee is appointed which is responsible for the preparation and conduct of the study. The subcommittee reports the progress of the study to the Steering Committee. The subcommittee includes a Chairman (who is a Principal Investigator or an alternate investigator from one of the clinical cores), Principal Investigators or alternate investigators from each site, and the NICHD Program Scientists.

9.2.6 Publications Subcommittee

The Publications Committee is a standing committee of the Steering Committee. The functions of this committee are to develop publication policies and to review all manuscripts and abstracts prior to submission. The goals of this committee are to ensure fair and appropriate authorship, credit and high quality publications.

9.3 Study Violation Management

The Protocol Deviation Form will be completed for every episode in which procedures specified in the protocol are not carried out or are carried out incorrectly. Those deviations that are significant will be labeled as protocol violations.

Once a woman is enrolled, protocol violations that could occur include the following:

1. Failure to obtain and/or document informed consent or assent if applicable;
2. Failure to obtain and/or document HIPAA Authorization of Disclosure;
3. Breach of participant confidentiality;
4. Failure to keep Institutional Review Board (IRB) approval current;
5. Enrollment of an ineligible participant, for example:
    a. Clinical gestational age outside range for eligibility;
    b. Sampling strategy violation in the recruitment of a control participant;
    c. Documented conditions making participant ineligible.
6. Collection and/or examination of a specimen without the consent of the study participant;

The DMSI Core will provide protocol violation reports to NICHD. The NICHD will propose solutions. If a problem persists, it may require immediate action by NICHD.

10 Data and Resource Sharing

A unique aspect of this Network is its open policy on data and resource accessibility to the outside scientific community. The NICHD expects that the Network investigators benefit from the first and continuing use, but not from prolonged exclusive use of data, resources or biological specimens. Furthermore, one of the aims in the creation of the Network was to provide the scientific community access to the Network's data, resources, and specimens in order to stimulate research in preterm birth and its sequelae. Thus, provisions are defined consistent with these principles (also see Manual of Operations).

11 Human Subjects

11.1 Waivers of Informed Consent to View Protected Health Information

All participating clinical sites and cores will obtain the appropriate waivers from their IRB to view and record protected health information (PHI) about potential study participants for purposes of determining their eligibility for inclusion in a study. Women will be made aware of the case-control study by someone directly involved in their care (i.e., their physician or nurse) and permission will be obtained through the care providers for study personnel to approach each woman to further discuss her participation.

11.2 Informed Consents

Written informed consent will be obtained from all study participants or their legal guardians. Template consent forms are provided for each study. Consent will be obtained by study-related personnel who are knowledgeable of the study and who have knowledge and training in the consent process and in protection of human subjects. No study-related procedures (interview, sample collection, chart abstraction) will be undertaken before a signed consent form has been completed by the subject or her legal guardian. During the process of obtaining informed consent, potential participants will have the nature of the study, specimen collection, data collection procedures, the importance of compliance to study procedures, and the potential risks and benefits explained to them. Potential participants will be told that there is no obligation to participate, that there will be no penalty for declining to participate, and that their treatment will not be compromised if they choose not to participate or cease participation at any time.

Ample time will be provided for each potential participant to read and understand the consent form and to ask questions. If a potential participant and/or legal guardian (if applicable) cannot read, the consent form will be read aloud or an audio-tape of the consent form and a tape player will be provided. Written consent forms and formal interpreters will be available to conduct informed consent in English and Spanish. Written translations and formal interpreters for conducting informed consent in other languages will be available with variance by clinical site, consistent with local IRB requirements and approvals. A woman who consents to the study will be given a copy of the signed consent form for her personal records. A copy will be placed in her medical record (if required). The original signed copy will be kept in a locked file at the clinical site with other confidential information on the participant.

11.3 HIPAA Requirements

The Health Insurance Portability and Accountability Act (HIPAA) requires that all research collecting identifiable health information on an individual be in compliance with HIPAA standards and regulations. HIPAA regulations specifically apply to research studies collecting PHI.

PHI is defined by HIPPA as health information transmitted or maintained in any form or medium that:

1. Identifies or could be used to identify an individual, and
2. Is created or received by a healthcare provider, health plan or employer, and
3. Relates to past, present or future physical or mental health or condition of an individual.

Given that the present case-control study will utilize participants' PHI; all clinical sites will comply with the HIPAA regulations as they relate to research. Compliance for each clinical site requires that each subject read and sign a hospital specific form "HIPAA Authorization to Use and Disclose Individual Health Information for Research Purposes". This may be combined with the consent form. Case-control study participants will receive a copy of the signed authorization. Importantly, the "Certificate of Confidentiality" issued for the protocol will help assure protection of the PHI data.

In addition, all study personnel who have contact with potential participants or data will have completed a course on human subjects' protection. Each site PI and directors of various laboratories will be responsible for ensuring that their personnel have completed an approved training program.

11.4 Protection of Confidentiality

Participation in this research study exposes subjects to the risk of loss of confidentiality. In order to minimize this risk to the greatest extent possible, certain safeguards will be implemented.

11.4.1 Protecting Subject Confidentiality during the Maternal Interview

Every effort will be made to conduct the maternal interview in a private location that minimizes the risk that answers to interview questions will be overheard by others not directly involved in the research. Information obtained through maternal interview will not be disclosed to anyone who is not directly involved in the study, except in cases where potentially life-threatening conditions are revealed (e.g. suicidal ideation).

11.4.2 Protecting Subject Confidentiality – Chart Abstraction / Record Keeping / Data Transmission

All study participants will be assigned a study ID number by the DMSI core. This number will appear on all data collection instruments. The study ID will be linked to the participant's name and other identifiers during the screening process. The individual identifiers will be removed from the analytical data and the link will be stored separately. Furthermore, this link at the DCAC will be destroyed once the analysis is completed.

Information collected for study purposes may be entered directly onto computers containing the data collection instruments. These computers will be dedicated to Preterm Network activities and will have appropriate password protections in place to prevent unauthorized disclosure of study data.

Data will be transmitted electronically to the DMSI core, using encryption to safeguard the information as it is transmitted. Copies of medical records will be kept at the clinical sites for research purposes. All copies of medical records and the linkage between a participant name and study ID will be destroyed after ten years of study completion. Thereafter, blood or tissue samples and information from the interview and medical records collected for the study will be marked with the study ID number only. Even though the information will still be used for research, it will no longer be possible to be traced to an individual. Any results of this study reported in medical journals or at meetings will not identify individuals.

11.4.3 Protecting Subject Confidentiality – Biological Specimens

Any tissue or fluid samples collected for study purposes will be marked with the participant's study ID number only. The local PI will maintain a record that can link these collected specimens with the individual they came from; this record will be maintained in a secure location (e.g. locked file cabinet or electronic file with adequate protections established) and destroyed within seven years of study completion.

11.4.4 Protecting Subject Confidentiality – Certificate of Confidentiality

To help us further protect the privacy of the study participant, the researchers have obtained a Certificate of Confidentiality from the National Institutes of Health (NIH) / National Institute of Child Health and Human Development (NICHD). This Certificate means that the researchers cannot be forced (for example by court order) to disclose any information that might identify a study participant to any federal, state, or local court.

11.5 Protection of Subjects

Participant safety always takes priority above all else. Toward this end, the Preterm Network will adhere to the International Conference on Harmonization's general principles of Good Clinical Practice. These principles dictate, first and foremost, that we have weighed and continue to weigh the anticipated benefits versus the foreseeable risks to all participants. Participants will be fully informed of these risks before they are enrolled. Close monitoring of participants' adverse events will occur throughout the study.

The Preterm Network Steering Committee will continually review the obstetric literature to determine if new findings substantially affect the study's rationale or justification, or if new or previously unforeseen risks must be conveyed to participants. The study's leadership will share any such findings with the Scientific Advisory and the NICHD Project Officers in a timely manner so informed decisions can be made about continuing the study. Assurances of participant safety are made through informed consent procedures, responsibilities for participant confidentiality, adherence to HIPAA regulations, and close monitoring and reporting of adverse events. These procedures are described in Manuals of Procedures.

11.6 Risks and Discomforts

During the maternal interview, some questions may make subjects feel uncomfortable, or questions may upset them as they recall their pregnancy and their baby. Risks from having blood drawn include pain and bruising where the needle goes in. Very rarely, the site can become infected. Information will be obtained from tests done on banked specimens (umbilical cord homogenate and meconium) about whether a subject or her infant (live born or stillborn) have been exposed to different kinds of drugs, such as marijuana and cocaine. The results of these tests will not be provided to any law enforcement or social agency.

Laboratory tests including genetic testing carries some unique risks. The concern is that test results might someday be used against a person. However, the chances of this happening as a result of this study are small. Blood and tissue samples will only be labeled with a study ID and none of the Preterm Network results from genetic testing will be released or placed in medical records.

11.7 Potential Benefits

By participating in this study, women who had a SPTB may benefit from an increased chance of identifying a cause for why the baby was preterm. The majority of the research findings will not be relayed to the participant. However, the participant will be informed if malignancy is found during the placental examination. Whether or not they choose to participate in this project, they will be offered ongoing counseling and grief support services. The information derived from this study may help prevent SPTBs in the future.

11.8 Compensation for Participation

Some compensation for travel expenses will be made for inconvenience and loss of time.

11.9 Subject Population

Together the three clinical sites (University of Alabama , University of Texas Medical Branch at Galveston and University of Utah) comprise approximately 19,000 spontaneous births per year of which a high percentage (15%) are preterm births (<37 weeks of gestation). Especially noteworthy is the high percentage (3-6%) of early spontaneous preterm births (<34 weeks of gestation), which is associated with the highest incidence of infant mortality and morbidity, and is the focus of the network. In addition, the population is enriched in the major U.S. minority and ethnic group of African Americans and Hispanics, respectively. Overall, based on cumulative spontaneous preterm birth data, approximately 47% of the population is Caucasian (non-Hispanic), 17 % African American and 34% Hispanic. (The national average is approximately 68% Caucasian (non-Hispanic), 13.7 % African American and 14% Hispanic). Table 5 shows the anticipated cumulative racial/ethnic percentages based on site specific enrollment goals and site specific race/ethnic distribution. Although it is recognized that race and ethnicity, besides many other variables, are associated with a spontaneous preterm birth, these studies may not be powered adequately to examine the effects of race/ethnicity. Nonetheless, the data will be analyzed according to race and ethnicity to determine if any significant differences or trends are detected.

Table 5. Anticipated Racial/Ethnic Percentages

Caucasian	African	Hispanic
49%	30%	20%

11.10 Subject Recruitment and Network Timeline

The Network is funded for only 5 years and is not expected to be renewed. The following timeline is anticipated: Recruitment is anticipated to start in year 2 and the studies are to finish near the end of year 3 or in the beginning of year 4. Proportional to the size of the appropriate case catchment population for each of the clinical sites and so that the studies can be concluded within the 2 year time interval, the minimum enrollment numbers per year for each clinical site is shown in Table 6. Specimen analysis is anticipated to occur in year 4 and final statistical analyses of the data is anticipated to be finished by the end of year 5.

Table 6. GPN Minimum Enrollment Goals Per Year

	CASE-CONTROL	LONGITUDINAL	EXPRESSION PROFILING
UAB	186+186=372	47	2 x 6 groups= 12
UTMB	67+67=134	84	3 x 6 groups = 18
UU	247 + 247 = 494	119	6 x 6 groups = 36

12 References

Abecasis, G. R. and W. O. Cookson (2000). "GOLD--graphical overview of linkage disequilibrium." Bioinformatics **16**(2): 182-3.

Adam, B. L., Y. Qu, et al. (2002). "Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men." Cancer Res **62**(13): 3609-14.

Aebersold, R. and M. Mann (2003). "Mass spectrometry-based proteomics." Nature **422**(6928): 198-207.

Alizadeh, A. A., M. B. Eisen, et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. [see comments.]. Nature **403**(6769): 503-11.

Ambroise, C. and G. J. McLachlan (2002). "Selection bias in gene extraction on the basis of microarray gene-expression data." Proc Natl Acad Sci USA **99**: 6562-6566.

Barrett, M. T., A. Scheffer, et al. (2004). "Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA." Proc Natl Acad Sci U S A **101**(51): 17765-70.

Bejerano, G., M. Pheasant, et al. (2004). "Ultraconserved elements in the human genome." Science **304**(5675): 1321-5.

Benjamini, Y. and Y. Hochberg (1995). "Controlling the false discovery rate: A practical and powerful approach to utiple testing." Journal of the Royal Statistical Society, Series B **57**: 289-300.

Breiman, L., J. Friedman, et al. (1984). Classification and Regression Trees. California, Wadsworth.

Buhimschi, I. A., R. Christner, et al. (2005). "Proteomic biomarker analysis of amniotic fluid for identification of intra-amniotic inflammation." Bjog **112**(2): 173-81.

Butte, A. J. and I. S. Kohane (2000). "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements." Pac Symp Biocomput: 418-29.

Calvo, K. R., L. A. Liotta, et al. (2005). "Clinical proteomics: from biomarker discovery and cell signaling profiles to individualized personal therapy." Biosci Rep **25**(1-2): 107-25.

Carlson, C. S., M. A. Eberle, et al. (2004). "Mapping complex disease loci in whole-genome association studies." Nature **429**(6990): 446-52.

Chaudhary, A. J., E. R. Wickremshinhe, et al. (2006). "A risk-based approach to bioanalytical methods validations and sample analysis during drug discovery and development." Amer Drug Discover **1**: 34-42.

Craig, D. W. and D. A. Stephan (2005). "Applications of whole-genome high-density SNP genotyping." Expert Rev Mol Diagn **5**(2): 159-70.

Crider, K. S., N. Whitehead, et al. (2005). "Genetic variation associated with preterm birth: a HuGE review." Genet Med **7**(9): 593-604.

Daly, M. J., J. D. Rioux, et al. (2001). "High-resolution haplotype structure in the human genome." Nat Genet **29**(2): 229-32.

Devlin, B., Roeder, K. (1999). “Genomic control for association studies.” Biomerics, **55**, 997-1004.

D'Haeseleer, P., X. Wen, et al. (1999). "Linear modeling of mRNA expression levels during CNS development and injury." Pac Symp Biocomput: 41-52.
Dietel, M. and C. Sers (2006). "Personalized medicine and development of targeted therapies: the upcoming challenge for diagnostic molecular pathology. A review." Virchows Arch **448**(6): 744-55.
Duan, F. and H. Zhang (2004). "Correcting the loss of cell-cycle synchrony in clustering analysis of microarray data using weights." Bioinformatics **20**(11): 1766-71.
Dudoit, S., J. Fridlyand, et al. (2002). "Comparison of discrimination methods for the classification of tumors using gene expression data." JASA **97**: 77-87.
Dunckley, T., K. D. Coon, et al. (2005). "Discovery and development of biomarkers of neurological disease." Drug Discov Today **10**(5): 326-34.
Efron, B., R. Tibshirani, et al. (2001). "Empirical Bayes Analysis of a Microarray Experiment." Journal of the American Statistical Association **96**(456): 1151-1160.
Eisen, M. B., P. T. Spellman, et al. (1998). "Cluster analysis and display of genome-wide expression patterns." Proc Natl Acad Sci U S A **95**(25): 14863-8.
Esplin, M. S., M. R. Peltier, et al. (2005). "Monocyte chemotactic protein-1 expression is increased in human gestational tissues during term and preterm labor." Placenta **26**(8-9): 661-71.
Excoffier, L. and M. Slatkin (1995). "Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population." Mol Biol Evol **12**(5): 921-7.
Fallin, D., A. Cohen, et al. (2001). "Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease." Genome Res **11**(1): 143-51.
Fan, J. B., X. Chen, et al. (2000). "Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays." Genome Res **10**(6): 853-60.
Feuk L, Marshall CR, et al. (2006) "Structural variants: changing the landscape of chromosomes and design of disease studies." Hum Mol Genet.**15** Spec No 1:R57-66.
Fischer, H. P. (2005). "Towards quantitative biology: integration of biological information to elucidate disease pathways and to guide drug discovery." Biotechnol Annu Rev **11**: 1-68.
Friedman, N., M. Linial, et al. (2000). "Using Bayesian networks to analyze expression data." J Comput Biol **7**(3-4): 601-20.
Furey, T. S., N. Cristianini, et al. (2000). "Support vector machine classification and validation of cancer tissue samples using microarray expression data." Bioinformatics **16**(10): 906-14.
Gabriel, S. B., S. F. Schaffner, et al. (2002). "The structure of haplotype blocks in the human genome." Science **296**(5576): 2225-9.
Gauderman, W. J. (2002). "Sample size requirements for association studies of gene-gene interaction." Am J Epidemiol **155**(5): 478-84.
Giarratano, G. (2006). "Genetic influences on preterm birth." MCN Am J Matern Child Nurs **31**(3): 169-75; quiz 176-7.
Goldenberg, R. L., A. R. Goepfert, et al. (2005). "Biochemical markers for the prediction of preterm birth." Am J Obstet Gynecol **192**(5 Suppl): S36-46.
Gravett, M. G., M. J. Novy, et al. (2004). "Diagnosis of intra-amniotic infection by proteomic profiling and identification of novel biomarkers." Jama **292**(4): 462-9.

Gross, R. W., C. M. Jenkins, et al. (2005). "Functional lipidomics: the roles of specialized lipids and lipid-protein interactions in modulating neuronal function." Prostaglandins Other Lipid Mediat **77**(1-4): 52-64.

Gygi, S. P., B. Rist, et al. (1999). "Quantitative analysis of complex protein mixtures using isotope-coded affinity tags." Nat Biotechnol **17**(10): 994-9.

Hacia, J. G., L. C. Brody, et al. (1998). "Applications of DNA chips for genomic analysis." Mol Psychiatry **3**(6): 483-92.

Hack, M. and A. A. Fanaroff (1993). "Outcomes of extremely immature infants--a perinatal dilemma." N Engl J Med **329**(22): 1649-50.

Haddad R, Tromp G, Kuivaniemi H, Chaiworapongsa T, Kim YM, Mazor M, Romero R. Human spontaneous labor without histologic chorioamnionitis is characterized by an acute inflammation gene expression signature. Am J Obstet Gynecol 2006;195:394-405.

Han, X. and R. W. Gross (2005). "Shotgun lipidomics: electrospray ionization mass spectrometric analysis and quantitation of cellular lipidomes directly from crude extracts of biological samples." Mass Spectrom Rev **24**(3): 367-412.

Hansen, K. C., G. Schmitt-Ulms, et al. (2003). "Mass spectrometric analysis of protein mixtures at low levels using cleavable 13C-isotope-coded affinity tag and multidimensional chromatography." Mol Cell Proteomics **2**(5): 299-314.

Hanson, R. L. and W. C. Knowler (1998). "Analytic strategies to detect linkage to a common disorder with genetically determined age of onset: diabetes mellitus in Pima Indians." Genet Epidemiol **15**(3): 299-315.

Harris MA, Lomax J, Ireland A, Clark JI. 2005. The Gene Ontology project. In: Subjramaniam S (ed.), Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics, Part 4, Bioinformatics. Wiley and Sons, Inc., New York.

Hao, K., X. Wang, et al. (2004). "A candidate gene association study on preterm delivery: application of high-throughput genotyping technology and advanced statistical methods." Hum Mol Genet **13**(7): 683-91.

Heller, R. A., M. Schena, et al. (1997). "Discovery and Analysis of Inflammatory Disease-Related Genes using cDNA Microarrays." Proceedings of the National Academy of Sciences of the United States of America(6): 2150-2155.

Hirschhorn, J. N. and M. J. Daly (2005). "Genome-wide association studies for common diseases and complex traits." Nat Rev Genet **6**(2): 95-108.

Horton, N. J. and N. M. Laird (1999). "Maximum likelihood analysis of generalized linear models with missing covariates." Stat Methods Med Res **8**(1): 37-50.

Huang, J., W. Wei, et al. (2006). "CARAT: a novel method for allelic detection of DNA copy number changes using high density oligonucleotide arrays." BMC Bioinformatics **7**: 83.

Hunter, D. J. (2005). "Gene-environment interactions in human diseases." Nat Rev Genet **6**(4): 287-98.

Ibrahim, J. G. (1990). "Incomplete data in generalized linear models." Journal of the American Statistical Association **85**: 765-69.

Irizarry, R. A., S. L. Ooi, et al. (2003). "Use of mixture models in a microarray-based screening procedure for detecting differentially represented yeast mutants." Stat Appl Genet Mol Biol **2**(1): Article1.

Jawaheer, D., W. Li, et al. (2002). "Dissecting the genetic complexity of the association between human leukocyte antigens and rheumatoid arthritis." Am J Hum Genet **71**(3): 585-94.

Kenward, M. G. and G. Molenberghs (1999). "Parametric models for incomplete continuous and categorical longitudinal data." Stat Methods Med Res **8**(1): 51-83.
Kramer, M. S., K. Demissie, et al. (2000). "The contribution of mild and moderate preterm birth to infant mortality. Fetal and Infant Health Study Group of the Canadian Perinatal Surveillance System." Jama **284**(7): 843-9.
Lander, E. S. and N. J. Schork (1994). "Genetic dissection of complex traits." Science **265**(5181): 2037-48.
LaVallee, T.M., Tarantini, F., Gamble, S., et al. (1998)Synaptotagmin-1 is required for fibroblast growth factor-1 release. Journal of Biological Chemistry **273**: 22217-22223.
Lee, S. H., M. V. Williams, et al. (2005). "Targeted chiral lipidomics analysis." Prostaglandins Other Lipid Mediat **77**(1-4): 141-57.
Lee, S. H., M. V. Williams, et al. (2003). "Targeted lipidomics using electron capture atmospheric pressure chemical ionization mass spectrometry." Rapid Commun Mass Spectrom **17**(19): 2168-76.
Li, C. and W. H. Wong (2001). "Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection." Proc Natl Acad Sci U S A **98**(1): 31-6.
Little, R. J. A. (1992). "Regression with missing X's: a review." Journal of the American Statistical Association **87**: 1227-37.
Little, R. J. A. and D. B. Rubin (1987). Statistical analysis with missing data. New York.
Livingston, R. J., A. von Niederhausern, et al. (2004). "Pattern of sequence variation across 213 environmental response genes." Genome Res **14**(10A): 1821-31.
Lucito, R., J. Healy, et al. (2003). "Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation." Genome Res **13**(10): 2291-305.
Marchini, J., P. Donnelly, et al. (2005). "Genome-wide strategies for detecting multiple loci that influence complex diseases." Nat Genet **37**(4): 413-7.
Marcotte, E. R., L. K. Srivastava, et al. (2001). "DNA microarrays in neuropsychopharmacology." Trends Pharmacol Sci **22**(8): 426-36.
Martin, J. A., B. E. Hamilton, et al. (2005). "Births: final data for 2003." Natl Vital Stat Rep **54**(2): 1-116.
Mathews, T. J., F. Menacker, et al. (2004). "Infant mortality statistics from the 2002 period: linked birth/infant death data set." Natl Vital Stat Rep **53**(10): 1-29.
McDonald, W. H. and J. R. Yates, 3rd (2003). "Shotgun proteomics: integrating technologies to answer biological questions." Curr Opin Mol Ther **5**(3): 302-9.
Mehrotra, K., C. Mohan, et al. (1997). Elements of Artificial Neural Networks. M. Press. Cambridge, Penram International Press.
Morley, M., C. M. Molony, et al. (2004). "Genetic analysis of genome-wide variation in human gene expression." Nature **430**(7001): 743-7.
Morris, A. P. (2005). "Direct analysis of unphased SNP genotype data in population-based association studies via Bayesian partition modelling of haplotypes." Genet Epidemiol **29**(2): 91-107.
Ng, P. C. and S. Henikoff (2003). "SIFT: Predicting amino acid changes that affect protein function." Nucleic Acids Res **31**(13): 3812-4.
Nielsen, J. and S. Oliver (2005). "The next wave in metabolome analysis." Trends Biotechnol **23**(11): 544-6.

Niu, T. (2004). "Algorithms for inferring haplotypes." Genet Epidemiol **27**(4): 334-47.
Ong, S. E., B. Blagoev, et al. (2002). "Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics." Mol Cell Proteomics **1**(5): 376-86.
Ornstein, D. K. and E. F. Petricoin, 3rd (2004). "Proteomics to diagnose human tumors and provide prognostic information." Oncology (Williston Park) **18**(4): 521-9; discussion 529-32.
Park, S. J., W. G. Yoon, et al. (2006). "Proteome analysis of human amnion and amniotic fluid by two-dimensional electrophoresis and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry." Proteomics **6**(1): 349-63.
Patil, N., A. J. Berno, et al. (2001). "Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21." Science **294**(5547): 1719-23.
Patton, W. F. (2002). "Detection technologies in proteome analysis." J Chromatogr B Analyt Technol Biomed Life Sci **771**(1-2): 3-31.
Pinkel, D., R. Segraves, et al. (1998). "High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays." Nat Genet **20**(2): 207-11.
Pollack, J. R., C. M. Perou, et al. (1999). "Genome-wide analysis of DNA copy-number changes using cDNA microarrays." Nat Genet **23**(1): 41-6.
Pritchard, J. K. and N. A. Rosenberg (1999). "Use of unlinked genetic markers to detect population stratification in association studies." Am J Hum Genet **65**(1): 220-8.
Ramaswamy, S., P. Tamayo, et al. (2001). "Multiclass cancer diagnosis using tumor gene expression signatures." Proc Natl Acad Sci U S A **98**(26): 15149-54.
Risch, N. (2000). "Searching for genes in complex diseases: lessons from systemic lupus erythematosus." J Clin Invest **105**(11): 1503-6.
Robins, J. M., R. A., et al. (1994). "Estimation of regression-coefficients when some regressors are not always observed." Journal of the American Statistical Association **89**: 846-866.
Ross, P. L., Y. N. Huang, et al. (2004). "Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents." Mol Cell Proteomics **3**(12): 1154-69.
Satagopan, J. M., D. A. Verbel, et al. (2002). "Two-stage designs for gene-disease association studies." Biometrics **58**(1): 163-70.
Satten, G. A. and M. P. Epstein (2004). "Comparison of prospective and retrospective methods for haplotype inference in case-control studies." Genet Epidemiol **27**(3): 192-201.
Schaid, D. J. (2004). "Evaluating associations of haplotypes with traits." Genet Epidemiol **27**(4): 348-64.
Segal, E., M. Shapira, et al. (2003). "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data." Nat Genet **34**(2): 166-76.
Segal, E., B. Taskar, et al. (2001). "Rich probabilistic models for gene expression." Bioinformatics **17 Suppl 1**: S243-52.
Setakis, E., H. Stirnadel, et al. (2005). "Logistic regression protects against population structure in genetic association studies." Genome Res **16**(2): 290-6.
Shmulevich, I., E. R. Dougherty, et al. (2002). "Gene perturbation and intervention in probabilistic Boolean networks." Bioinformatics **18**(10): 1319-31.
Sidow, A. (2002). "Sequence first. Ask questions later." Cell **111**(1): 13-6.

Snijders, A. M., N. Nowak, et al. (2001). "Assembly of microarrays for genome-wide measurement of DNA copy number." Nat Genet **29**(3): 263-4.
Soukas, A., P. Cohen, et al. (2000). "Leptin-specific patterns of gene expression in white adipose tissue." Genes Dev **14**(8): 963-80.
Spence, M. A., D. A. Greenberg, et al. (2003). "The emperor's new methods." Am J Hum Genet **72**(5): 1084-7.
Stephens, M., N. J. Smith, et al. (2001). "A new statistical method for haplotype reconstruction from population data." Am J Hum Genet **68**(4): 978-89.
Swanson, S. K. and M. P. Washburn (2005). "The continuing evolution of shotgun proteomics." Drug Discov Today **10**(10): 719-25.
Tamayo, P., D. Slonim, et al. (1999). "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation." Proc Natl Acad Sci U S A **96**(6): 2907-12.
Thomas, D. C., R. W. Haile, et al. (2005). "Recent developments in genomewide association scans: a workshop summary and review." Am J Hum Genet **77**(3): 337-45.
Troyanskaya, O., M. Cantor, et al. (2001). "Missing value estimation methods for DNA microarrays." Bioinformatics **17**(6): 520-5.
Tusher, V. G., R. Tibshirani, et al. (2001). "Significance analysis of microarrays applied to the ionizing radiation response." Proc Natl Acad Sci U S A **98**(9): 5116-21.
Tyers, M. and M. Mann (2003). "From genomics to proteomics." Nature **422**(6928): 193-7.
Weckwerth, W. and K. Morgenthal (2005). "Metabolomics: from pattern recognition to biological interpretation." Drug Discov Today **10**(22): 1551-8.
Wood, N. S., N. Marlow, et al. (2000). "Neurologic and developmental disability after extremely preterm birth. EPICure Study Group." N Engl J Med **343**(6): 378-84.
Wright, S. (1951). "The Genetic Structure of Populations." Ann Eugen **15**: 323-54.
Xi, T., I. M. Jones, et al. (2004). "Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function." Genomics **83**(6): 970-9.
Yan, Y., V. M. Weaver, et al. (2005). "Analysis of protein expression during oxidative stress in breast epithelial cells using a stable isotope labeled proteome internal standard." J Proteome Res **4**(6): 2007-14.
Yeung, K. Y., C. Fraley, et al. (2001). "Model-based clustering and data transformations for gene expression data." Bioinformatics **17**(10): 977-87.
Yeung, K. Y., D. R. Haynor, et al. (2001). "Validating clustering for gene expression data." Bioinformatics **17**(4): 309-18.
Yocum, A. K., C. M. Busch, et al. (2006). "Proteomics-based strategy to identify biomarkers and pharmacological targets in luekemias with t(4;11) translocations." J Proteome Res(DOI): 1000-1021.
Yocum, A. K., K. Yu, et al. (2005). "Effect of immunoaffinity depletion of human serum during proteomic investigations." J Proteome Res **4**(5): 1722-31.
Yu, K. H., A. K. Rustgi, et al. (2005). "Characterization of proteins in human pancreatic cancer serum using differential gel electrophoresis and tandem mass spectrometry." J Proteome Res **4**(5): 1742-51.
Zhang, H. and G. Bonney (2000). "Use of classification trees for association studies." Genet Epidemiol **19**(4): 323-32.

Zhang, H. and B. Singer (1999). Recursive Partitioning in the Health Sciences. New York, Springer.
Zhang, H., C. Yu, et al. (2001). "Recursive partitioning for tumor classification with gene expression microarray data." Proc Natl Acad Sci USA **98**: 6730-6735.
Zhang, H. and C. Y. Yu (2002). "Tree-based analysis of microarray data for classifying breast cancer." Frontiers in Bioscience **7**: c63-7.
Zhang, H., C. Y. Yu, et al. (2003). "Cell and tumor classification using gene expression data: construction of forests." Proc Natl Acad Sci U S A **100**(7): 4168-72.
Zhang, K., Z. Qin, et al. (2005). "HapBlock: haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms." Bioinformatics **21**(1): 131-4.
Zhang, L. and L. Luo (2003). "Splice site prediction with quadratic discriminant analysis using diversity measure." Nucleic Acids Res **31**(21): 6214-20.
Zhao, L. P., S. S. Li, et al. (2003). "A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies." Am J Hum Genet **72**(5): 1231-50.
Zhao, X., B. A. Weir, et al. (2005). "Homozygous deletions and chromosome amplifications in human lung carcinomas revealed by single nucleotide polymorphism array analysis." Cancer Res **65**(13): 5561-70.
Zhu, Y., M. R. Spitz, et al. (2004). "An evolutionary perspective on single-nucleotide polymorphism screening in molecular cancer epidemiology." Cancer Res **64**(6): 2251-7.