Section - D

DETAILED RESEARCH PROPOSAL

1. Title of the Research Project:

Systems Biology approach to delineate molecular signatures of *Prakriti* in healthy humans.

2. Background:

Prakriti a fundamental concept of Ayurveda categories an individual based on multiple parameters that include physical and clinical analysis and broadly classifies them as Pitta, Vata and Kapha. A sequence-based omics was able to capture the individual group of Prakriti. But it fails to capture the heterogeneous phenomics of Prakriti. To further precisely define and understand the Prakriti, we would propose a novel integrated method that will utilize sequence-based omics; transcriptome and metagenome and also non-sequence based omics; proteomics and metabolites abundance to distil down the information that may precisely correlate with single subject. We will also create a pipeline to integrate above defined omics data with structured and unstructured phenomics data of Prakriti.

3. Objectives:

a) Primary objective:

1) To characterize the proteome profile of healthy male adults of different *Prakriti*.
2) To characterize the transcription profile including lncRNA of healthy male adults of different *Prakriti*.
3) To characterize the gut metagenome of healthy male adults of different *Prakriti*.

b) **Secondary Objective:**

1) To identify potential biological markers for different *Prakriti* through proteome, transcriptome, and metagenome.
2) To identify pathways and predictors for different *Prakriti.*
3)To establish system correlation between proteome, transcriptome and metagenome for different *Prakriti.*

4. Outcome Measures:

Primary Outcome Measures:

1. Measurement and analysis of Proteomics, Transcriptomics and metagenomics (microbiome) data on 588 samples (7 *Prakriti*, 7 CCRAS stations and two *aayans*)
2. Integration of all data using systems biology approach
3. Correlation of omics data with *Prakriti*

Secondary Outcome Measures:

1. New methodologies iteration of different omics data sets according to *Prakriti*.

2. Establishment of Big data analytics methods to convert *Prakriti* phenomics data in to structured data set
3. Base line data set for healthy and correlation with omics data set generated during this project
4. Molecular identifier signature for *Prakriti*

5. Study Procedure:

The research project is a collaborative study by the Central Council for Research in Ayurvedic Sciences, New Delhi in collaboration with the following Institutions:

1. School of life sciences & School of Biotechnology of Jawaharlal Nehru University (JNU) New Delhi;
2. International Centre for Genetic Engineering and Biotechnology (ICGEB) New Delhi
3. Institute of Ayurveda & Integrative Medicine, The University of Trans-Disciplinary Health Sciences and Technology(TDU) Bengaluru.

Name and Address of the Sponsor:
Central Council for Research in Ayurvedic Sciences (CCRAS)
Jawahar Lal Nehru Bhartiya Chikitsa Evam Homoeopathy Anusandhan Bhawan
61-65, Institutional Area, Opposite D-Block, Janakpuri, New Delhi-110058

Particulars of the Investigator(s)/ Co-Investigator(s)

1. Project Coordinator: Prof. Rana Pratap Singh
2. Co-Principal Investigator(s)
Prof. Rupesh Chaturvedi
Prof. Suneel Kateria
Prof. Arun S. Kharat
Dr. Hemant Ritturaj Kushwaha

6. Study Design:

Study Type: Clinical Research-Exploratory
Purpose : To study the molecular signatures of *Prakriti* through systems biology approach.
Masking : Single Blinding (Investigators)
Control : No control
Timing : Prospective
End Point: Characterization & Correlation of proteome, transcriptome and gut metagenome of healthy male adults of different *Prakriti* types. Identification of biological markers, pathways & predictors for different *Prakriti* types.
No. of Groups : Seven groups of Seven *Prakriti* Types
Sample Size : 294 (42 in each Prakriti type group)

7. Study Sites:

For Collection of Samples- Seven CCRAS Institutions:

1. National Ayurveda Research Institute for Panchakarma, Cheruthuruthy
2. Central Ayurveda Research Institute for Hepatobiliary Disorders, Bhubaneswar
3. Central Ayurveda Research Institute for Respiratory Disorders, Patiala
4. Raja Ramdeo Anandilal Podar (RRAP) Central Ayurveda Research Institute for Cancer, Mumbai
5. Regional Ayurveda Research Institute for Eye Diseases, Lucknow
6. Regional Ayurveda Research Institute for Gastro-Intestinal Disorders, Guwahati
7. Regional Ayurveda Research Institute, Tadung, Gangtok, Sikkim

For Sample Processing & Analysis:

1. School of life sciences & School of Biotechnology of Jawaharlal Nehru University (JNU) New Delhi for proteome, transcriptome, gut metagenome and Integromics
2. International Centre for Genetic Engineering and Biotechnology (ICGEB) New Delhi for Metabolomics
3. Institute of Ayurveda & Integrative Medicine, The University of Trans- Disciplinary Health Sciences and Technology (TDU) Bengaluru for Urine Exosome analysis.

8. Study Population: Total 294 participants (6 Participants each of seven *Prakriti* Types from 7 CCRAS centers (including one high altitude) in two aayans = [7*6*7]*2)=588 samples of blood, urine and stool as per the inclusion and exclusion criteria of the Study.

Drop-Outs: An attempt will be made to record the reason for drop outs, if any during the study.

9. Total Duration of the Study: 2 Years

Timelines:

Pre- trial Preparations: 3 months
Recruitment & Sample collection: 12 months
Data Processing: 3 months
Data Analysis: 4 months
Publication: 2 months

10. Study Methodology:

1) **<u>Screening and Enrollment of Participants and Collection of Sample:</u>** Screening and Enrollment of apparently healthy male individuals on the basis of inclusion and

exclusion criteria and collection of Blood, Urine samples & Fecal samples will be done by CCRAS.

2) Sample Transportation:

The samples will be collected from the seven CCRAS participating institutes and will be transported at JNU, New Delhi for Transcriptomic, Proteomic, Metagenomic and Integromics part of study.

3) Execution of Transcriptomic, Proteomic, Metagenomic and Integromics Study for systems biology.

i. Transcriptome Study:

RNA sequencing and mapping of Genes from RNA Seq

Leucocytes will be isolated from whole blood using ficoll method from different type of Prakariti subjects and total RNA will be extracted using trizol. Quality of RNA will be assessed with Agilent Bio analyzer 2100 and library will be prepared using the Illumina TruSeq stranded mRNA Sample Preparation Kit and sequencing will be done on the Illumina HiSeq 2500 using v3 SBS chemistry. Data will be analyzed using Tophat version 2.0.12 with the reference annotation file mm10. The aligned reads will be assembled and transcript expression will be measured by quantifying FPKM (Fragments Per Kilo base of transcripts per Million fragments mapped) in the Cufflinks version 2.2.1. Differential expression between different groups will be detected by using the Cuffdiff.

Transcriptome analysis by Prof. Rupesh Chaturvedi

After mapping of genes, gene IDs will be retrieved using MGI and NCBI database. Using fold change cut-off of >2 and p value <0.05, significant differentially regulated genes will be identified between different Prakariti groups.

Protein-Protein Interaction Network, Module and Gene Ontology Analysis

The up-regulated genes from the datasets of subjects will be used to construct the Protein-Protein Interaction (PPI) networks. All the PPI within the up-regulated proteins will be retrieved from STRING database and will be visualized using Cytoscape. Network analyzer will be used to identify the hub proteins and betweenness centrality. The modules in the Protein-Protein Interaction (PPI) networks of each *Prakriti* will be constructed using MCODE (Molecular Complex Detection) that identified the clusters or modules based on the highly interconnected regions. In order to identify the role of hubs in modules we will consider the top two hub proteins based on their degree from the networks. We will calculate the number of connecting partner for the hubs in each module, which will be identified by MCODE. The modules and hub-modules interactions will be visualized using Cytoscape. GO enrichment analysis for all the modules, up-regulated and down-regulated genes from all the datasets will also be carried out. The gene set enrichment analysis will be

performed using WebGestalt, with the Entrez Gene set as a reference, using a hypergeometric test with Benjamini-Hochberg multiple comparison adjustment corrections.

Validation through long non-coding RNA (lnRNA) sequencing by Prof. Rana P Singh

A subset of serum samples will be selected on the basis of transcriptomics data for validation purpose. From the selected samples RNAs will be purified and next-generation sequencing will be performed. The result obtained will provide information on the long non-coding RNA, which will provide information of gene regulation. This will further help in integrating and increasing resolution and sensitivity of transcriptomics data.

ii. Microbiome/Metagenomics by Prof. Arun S Kharat

Identification of gut microbiome for the particular *Prakriti* will be carried out with metagenomic approach.

Processing and extraction of DNA will be carried out from stool samples to obtain high quality microbial DNA. Fractionation and selective lysis will be carried out to ensure the minimize host DNA contamination. Along with this centrifugation, selective filtration and flow cytometry will be applied to enrich target fraction.

The stool samples will be collected and portion of which will be used for analysis while the remaining will be stored for future use. High quality DNA will be extracted from the stool sample, and the V2 and V3 regions of the bacterial 16S rRNA will be amplified using universal 16S primers 8F (5′-AGAGTTTGATCCTGGCTCAG-3′) and 541R (5′-WTTACCGCGGCTGCTGG-3′). Individual sample will be barcoded and pooled to construct the amplicon based library. The vast diversity among the library both at sequence level and relative abundance of sequence within the library will be estimated. Nucleotide sequencing of amplicons with an Illumina MiSeq instrument will be carried out to generate pair-ended 2 x 300 reads. Microbiome profiles will be evaluated using Qiime, Mothur, Ribosomal Database Project (RDP). In silico analysis will be carried out to identify identity of nucleotide sequences. After establishing identity for the microbes within gut-microbiome (qualitative microbiome) their relative abundance will be estimated by studying differential expression of amplicon within the sample (quantitative microbiome).

iii. Label free untargeted Proteomics by Prof. Suneel Katariya

25 µg of serum sample will be reduced with 5 mM of Tris (2-carboxyethyl) phosphine (TCEP), followed by alkylation using 50 mM iodoacetamide and digested with trypsin (1:50, trypsin: lysate ratio) at 37 °C for 16 hours. These digests will be subsequently cleaned with C18 silica cartridge before drying with speed vac (Concentrator plus/ Vacufuge® plus - Eppendorf). Resulting dried pellet will be resuspended in Buffer A containing 5% acetonitrile and 0.1% formic acid. All the

experiments will be performed with the Thermo Fisher Scientific TM EASY-nLCTM 1000 liquid chromatograph coupled to Thermo Orbitrap Fusion mass spectrometer (Thermo Fisher Scientific) equipped with nano-electrospray ion source. One μg of the peptide mixture will be loaded and resolved with 15 cm EASY-Spray LC column. Peptides were loaded with Buffer A and eluted with a 0-40% gradient of Buffer-B (95% acetonitrile/0.1% Formic acid) at a flow rate of 300 nL/min for 60 minutes.

Survey MS scan will be acquired in profile mode using Orbitrap mass analyzer at 350–1700 *m*/*z* mass range MS1 resolution of 60,000, $5x10^5$ AGC target at 50 ms injection time. Uppermost precursor ions will be screened with quadrupole mass filter at isolation width of 1.2 Da and fragmented with High-energy C-trap dissociation using 30% normalized collision energy. MS2 spectra will be acquired with Ion trap in rapid mode using 10,000 AGC target and 35 ms injection time. Lock mass option will be enabled for polydimethylcyclosiloxane (PCM) ions (m/z = 445.120025) for internal recalibration at every run. MS data for this study will be acquired in a data-dependent top ten method dynamically selecting the most abundant precursor ions from the survey scan.

All MS raw files will be analyzed with Proteome Discoverer 2.2 against the UniProt Human reference proteome database. For Sequest HT and MS Amanda 2.0 search, fragment mass tolerances and the precursor were set at 0.5 Da and 10.0 ppm respectively. The protease will be used in generating peptides like enzyme specificity would be set for trypsin/P (cleavage at the C terminus of "K/R: unless followed by "P") and the maximum missed cleavages value of two. Carbamidomethyl on cysteine as fixed modification, oxidation of methionine and N-terminal acetylation will be termed variable modifications for database search. Both peptide-spectrum match and protein identification false discovery rate (FDR as determined using percolator node) will be set to 0.01.

Relative protein quantification will be performed using label-free quantification method based on the Minora feature detector node of Proteome Discoverer 2.2 with default settings and considering only high PSM (peptide-spectrum matches) confidence. Based on UniProt accession number Pfam, KEGG pathways and GO annotations will be assigned for the list of identified proteins. Only proteins with at least two unique peptides will be included in further analyses.

Missing values for the protein abundance will be computed using random numbers from a normal distribution and measure of central tendency (the mean and standard deviation) of which will be selected to best simulate low abundance values close to the noise level. Protein abundance will be log2-transformed, and the dataset will be subsequently filtered by 50% minimum valid values in each group. Significance will be assessed using both ANOVA and two-tailed student's *t*-test to understand the differentially expressed proteins in the study groups of different *Prakriti*, while

correction for multiple testing will be based on Benjamini-Hochberg, 10% FDR; P value < 0.05. Fold change and volcano plots will also be generated to support two-tailed student's *t*-test. Finally, a richer and straightforward feature identification strategy called Significance Analysis of Microarrays (SAM), which assigns a score to each protein on the basis of changes in their expression relative to the standard deviation of repeated measurements will be used to confirmed differential proteome. SAM has been reported to provide much robust information about the protein differential expression profiles and is usually germane in the evaluation of FDR and miss rates.

For data visualization, Principal Component Analysis (PCA), Partial Least Squares-Discriminant Analysis (PLS-DA), Orthogonal Projections to Latent Structures Discriminant Analysis (OPLS-DA) and Hierarchical Clustering Analysis (HCA) heatmaps with LFQ normalized abundance values and volcano plots will be built using the open-source statistical online package Metaboanalyst and programming language R called MetaboAnalystR. Protein ANalysis THrough Evolutionary Relationships (PANTHER) analysis will be used to gain a better understanding of the functions of all the statistically perturbed proteome profile in this study. All differentially expressed proteins will be categorized into different *Prakriti* groups based on their biological process, cellular localization, molecular function and biological pathway they influenced.
Both ROC and the area under the ROC curve (AUROC), 95% confidence interval (CI) and probability values will be calculated to appraise the power and robustness of protein group in discriminating different *Prakriti*.

iv. **Integromics for systems biology by Dr. Hemant Rituraj Kushwaha**

The integration of the multi-omics data is performed by combining the data as predictor variables to allow varied type of comprehensive modeling of the complex traits, which results in elaborating the interplay among the biological variations at various levels of regulation. There are two main approaches of the data integration: multi-staged analysis, which involves integrating information using a stepwise or hierarchical analysis approach; and meta-dimensional analysis, which refers to the concept of integrating multiple different data types to build a multivariate model associated with a given outcome. In order to integrate multiple data types meta-dimensional analysis is used which are based on concatenation based integration, transformation-based integration and model-based integration. In the proposed analyses, transformation-based integration and model-based integration will be performed. In the transformation-based integration multiple data sets is combined after transforming each data type into an intermediate form, such as a graph or a kernel matrix. Multiple graphs or kernels will then be merged for producing the predictive models. The transformation-based integration approach has the advantage of preserving data-type-specific properties from each data set when each type of data is transformed into an appropriate intermediate representation. Model-based

integration will be used to generate multiple models using the different types of data as training sets, and a final model is then generated from the multiple models created during the training phase, preserving data-specific properties. This approach can combine predictive models from different types of data.

4) Data Analysis:

The Data analysis will be carried out after execution of the processing of the Samples to characterizing and correlating Transcriptomic, Proteomic, Metagenomic, and Systems biology Integromics data according to different *Prakriti* types and identifying potential biomarkers, pathways & predictors for different *Prakriti* types.

5) Inclusion Criteria:

1. Healthy male volunteers of age between 25-50 years of different *Prakriti* types.
2. Willing to provide written informed consent and participate in the study.

6) Exclusion Criteria:

1. Males < 25 years or > 50 years and female gender.
2. Anyone who is using any form of narcotics like alcohol, tobacco and smoking etc. currently or in last 10 years.
3. Suffering from any acute disease condition or chronic diseases such as diabetes, hypertension or any other disease that may affect participation in the study.
4. Suffering from diseases like dengue, tuberculosis, malaria, chikungunya currently or in the last 5 years
5. Currently suffering any infectious disease like influenza, typhoid etc.
6. Persons on any type of medications or health supplements.
7. Mentally challenged persons.
8. Has changed his city in last five years
9. Not willing to participate in the study.

7) Withdrawal Criteria:

The participant may be withdrawn from the study if

- The participant does not come up/ not willing for giving sample second time.
- The participant himself wants to withdraw from the study.
- Participant develops any major illness/ hospitalized/ any accident during the study.

If the decision to withdraw a participant from the study will be taken by the Principal Investigator, then it will be justified in terms of the actual reason. The same will be informed to the Sponsor and the Ethics Committee within two working days.

8) Laboratory Investigations: