REPORT

Report on participation of the ICMR International Fellow (ICMR-IF) in Training/Research abroad.

1.	Name and designation of ICMR- IF	: Sharmila Bapat, Ph.D Scientist G
2.	Address	: National Centre for Cell Science, NCCS Complex, SPPU University Campus, Ganeshkhind, Pune 411 007, Maharashtra, India
3.	Frontline area of research in which Training / research was carried out	: Proteogenomics
4.	Name & address of Professor and host institute	: Professor David Fenyö Institute for Systems Genetics NYU School of Medicine Science Building, 435 East 30th St, New York, NY 10016
5.	Duration of fellowship with exact date	: 2 months 13 days w.e.f. 2nd December, 2019 to 14h February,2020.
6.	Highlights of work conducted	:
	i) Technique/expertise acquired	: Development of a full-length protein database for chimeric proteins
	ii) Research results, including any papers, Prepared / submitted for publication	: This data will definitely be published once the entire project is completed
	iii) Proposed utilization of the experience in India	: This database will be used to screen for expression of these novel peptides in Indian patient tumor samples

Overall Work Achieved –

1. **Collation of a backend, targeted peptide database for detection of full-length chimeric peptides using Mass spectrometry (MS) datsets** – From the previous -based chimeric transcript detection pipeline developed by me earlier on the Seven Bridges Genomics Cloud, we extracted Chimerascan-outputs and collated these to derive around 5000 chimeric transcripts. This is an addition to the sanctioned proposal of generating a

database from the 120 recurrent chimeric transcripts identified earlier. Collation of these outputs was performed and scripts coded and tested for extraction of the full-length transcript sequences of not only all these 5000 transcripts but also all the known variants of each chimeric transcript in the Ensemble Genome Browser. This vastly extended the proposed work to 32,000 full length transcripts all of which were extracted.

2. ***In silico* translation** - Another script was developed to generate all 6 putative reading frame isoforms of the proteins predicted from these 32000 transcripts. Further, these 1.96,000 protein sequences were collated to generate a full-length backend database. Further testing of the database using MS datasets for its validation was carried out towards debugging of the database.
3. **Development of an efficient pipeline for further analyses** - The database was docked onto the Seven Bridges Genomics Cloud that also provides access to the CPTAC (a publically available proteomics resource for samples banked in the TCGA). Prof. Fenyo has developed a pipeline for quantitation of proteins in mass spectra, which I adapted for detection of the novel chimeric as well as known proteins.
4. **Downstream analyses to reconstruct the entire protein and quantify isoforms** – The entire efforts so far were effectively channelized to identify the novel proteins that are possibly being expressed in ovarian cancer. This effort has led to the generation of a much larger quantum of data than was initially proposed in the sanctioned project. I have been mining this data at present to identify the isoforms of a few proteins. However, extracting the data for all the expressed proteins is a huge task that will be soon completed at NCCS.

Sabapathy 20/2/2020

Signature of ICMR-IF

ICMR Sanction No. INDO/FRC/452/S-60/2019-20-IHD