UC San Diego
MOORES CANCER CENTER

**2019 MCC CCSG Project Grant**

**Computational Infrastructure for Analysis of Precancer Data**

Ludmil Alexandrov, PhD

**Project Description:**

An essential step of cancer prevention is understanding the early events leading to cancer initiation and developing strategies to reduce or even inhibit such events. All cancers are caused by somatic mutations[1]. These mutations may be due to the intrinsic mutational processes, for example, infidelity of the DNA replication machinery, or to extrinsically triggered mutational processes, for example, exposures to environmental carcinogens. In principle, spontaneous somatic mutagenesis starts as early as the first division of the fertilized egg and mutations accumulate in every single cell of the body throughout a person's lifetime[1]. The majority of mutations found in somatic cells are believed to be "passengers" that have very little effect on cellular fitness. However, just by chance, a mutation may disrupt the function of a gene or enable a novel function of a gene, which may lead to an increased cellular fitness and subsequent cellular proliferation. Such mutations are termed "driver mutations". Large-scale genomic initiatives examining thousands of cancer patients across the spectrum of human neoplasia have identified more than 300 genes harboring driver mutations[2,3]. In most cases, the number of driver mutations in an individual cancer patient is limited to a handful, usually between two and ten mutations[3]. In contrast, passenger mutations are much more prevalent with most cancers harboring thousands, and, in some cases, even millions, of passenger mutations[4]. While the exact numbers of passenger and driver mutations necessary to accumulate prior to the development of a morphologically recognizable clonal expansion remain unknown, a recent analysis indicates that around half of the mutational burden found in a cancer is generated during the normal lineage of the cell, i.e., before the cell has undergone malignant transformation[5].

Prior studies have estimated that at least 40% and, possibly, as much as 90% of all human cancers are attributable to epidemiologically identified risk factors such as tobacco smoking, environmental pollutants, alcohol intake, exposure to ultraviolet light, viral or bacterial infections, and others[6,7]. Although the molecular mechanisms governing these associations have not been fully understood, genomic analyses of advanced cancers have identified distinct mutational patterns, termed "mutational signatures." Mutational signatures have been causally linked to many of the known cancer risk factors and shown to have a higher propensity to cause specific driver mutations[4,8]. Our previous pan-cancer analyses[4,8-11], encompassing whole-exome and whole- genome sequences of more than 20,000 primary cancers, revealed the activity of almost eighty distinct mutational processes. Some are present in many cancer types, notably a signature attributed to the APOBEC family of cytidine deaminases, whereas others are confined to a single cancer class. Certain signatures are associated with age of the patient at cancer diagnosis, known mutagenic exposures or defects in DNA maintenance, but many are of cryptic origin. However, for most mutational signatures and for most driver events, it is unclear whether they occur early or late in the lineage of a cancer cell. Examination of the

genomic landscape of human precancers will allow elucidating the early genomic events responsible for cancer initiation, which, in turn, may provide opportunities for practical intervention leading to cancer prevention.

The term precancer reflects a mass of cells that is not cancerous but that has an opportunity of becoming a cancer. In principle, any normal cell in the human body can be considered as a precancer since: *(i)* normal cells are not cancerous; *(ii)* each cell exhibits a certain, albeit minuscule, potential to undergo a malignant transformation. Considering that the average adult has ~40 trillion cells[12] and that one in every two people will get cancer in their lifetime[13], one rough estimate for the average probability for any cell in the body to become a cancer is $10^{-13}$. In practice, considering every cell in the human body as a precancer is impractical as this probability is extremely low. Further, considering benign lesions or the recently described macroscopic clones[14,15] in normal tissue as precancerous is also not practical as, currently, there is no evidence that these have an increased chance for neoplastic expansions; in fact, there is evidence that some of these clonal expansions are negatively selected and likely have a lower opportunity for becoming malignant[15]. A much more practical definition of precancer is to consider only lesions that have been shown to have a realistic opportunity of transforming into cancer. In this proposal, we consider a precancer to be a mass of cells that has not invaded neighboring tissues and that either transforms into cancer in at least 10% of cases or increases the relative risk for developing cancer at least two times. As such, we plan to examine lesions that are non-cancerous but that have a reasonable probability for transforming into advanced cancers.

While large-scale cancer atlas projects have brought unprecedented insights of the genomic events associated with advanced stages of human cancer, few studies have comprehensively profiled the genomic alterations in human precancers. In this project, **our goal is to accumulate genomics data from previously generated human precancers by curating more than 2,500 precancers from 38 tissue types.** To achieve this goal, we have already performed a detailed examination of the literature and contacted collaborators working on different precancer datasets. Overall, we have identified a set of **2,591 available precancer matched-normal cases** with an approximate size of 1,000 terabytes (~1 petabytes). Each sample has been subjected either to a whole- exome or to a whole-genome sequencing. In this pilot grant, we request funding for storage in order to be able to download these data. More specifically, we plan to purchase 1.1 petabytes of usable storage (corresponding to 1.4 petabytes of raw storage) at the Triton Shared Computing Cluster (TSCC) maintained by the San Diego Supercomputer Center (SDSC). After the storage is purchased and installed, we anticipate downloading all identified precancer cases within six months. The downloaded data will form the backbone of all our future precancer analysis. It will be used for generating preliminary data for grant applications as well as for performing genomics analysis for peer-reviewed publications.