# Abstract version on Thursday night:

Identifying highly mutated regions from population scale sequencing is a key way to discover cancer drivers.  Nevertheless, it is challenging to identify burdened regions because of severe mutation rate heterogeneity across the genome and across individuals gives rise to highly over-dispersed mutation counts. Moreover, it is known that part of this heterogeneity relates to confounding genomic features, such as replication timing and chromatin organization.
Results
Here, we address these issues with a Negative binomial regression based Integrative Method for mutation Burden analysis (NIMBus). Our approach (1) uses a Gamma-Poisson mixture model to capture the mutation rate heterogeneity across different individuals and (2) regresses the regional mutation counts across the genome against many features (381) extracted from REMC and ENCODE to estimate a local background mutation rate. Because of these two aspects it accurately models the over-dispersed mutation counts as a negative binomial distribution. As a demonstaration of NIMBus, we applied it to 649 whole-genome cancer sequences. It successfully identified well-known coding and noncoding drivers, such as TP53 and the TERT promoter. In addition, it also found known cancer related pathways, such as TP53 signaling and apoptosis pathways, to be significantly mutated.
Conclusion
NIMBus is a powerful tool to identify mutational hotspots. We make NIMBus available at [nimbus.gersteinlab.org](http://nimbus.gersteinlab.org/%22%20%5Ct%20%22_blank) and release our results as an online resource. Finally, we explain how the our approach to somatic mutations can readily be extended to examing the burdening of rare germline mutations in diseassed individuals.

# My edited abstract

Background

Identifying highly mutated regions from population scale sequencing is a key way to discover cancer drivers. Nevertheless, it is challenging to identify burdened regions because of severe mutation rate heterogeneity across the genome and across individuals, which gives rise to highly over-dispersed mutation counts. Moreover, it is known that part of this heterogeneity relates to confounding genomic features, such as replication timing and chromatin organization.

Results

Here, we address these issues with a Negative binomial regression based Integrative Method for mutation Burden analysis (NIMBus). Our approach (1) uses a Gamma-Poisson mixture model to capture the mutation rate heterogeneity across different individuals and (2) regresses the regional mutation counts across the genome against many features (381) extracted from REMC and ENCODE to estimate a local background mutation rate. Because of these two aspects it accurately models the over-dispersed mutation counts as a negative binomial distribution. As a demonstration of NIMBus, we applied it to 649 whole-genome cancer sequences. It successfully identified well-known coding and noncoding drivers, such as TP53 and the TERT promoter. In addition, it also found known cancer related pathways, such as TP53 signaling and apoptosis pathways, to be significantly mutated.

Conclusion

NIMBus is a powerful tool to identify mutational hotspots. We make NIMBus available at nimbus.gersteinlab.org and release our results as an online resource. Finally, we explain how our approach to somatic mutations can readily be extended to examine the burdening of rare germline mutations in diseased individuals.

# Abstract version of last week

Background

Identifying highly mutated regions is a key way that scientists can use sequencing on a population scale to discover key genomic regions associated with complex diseases such as cancer. Nevertheless, it is challenging to identify such regions because severe mutation rate heterogeneity across different genome regions gives rise to highly over-dispersed mutation counts. Moreover, it is known that part of this heterogeneity relates to confounding genomic features, such as replication timing and chromatin organization.

Results

Here, we address these issues with a Negative binomial regression based Integrative Method for mutation Burden analysis (NIMBus). This approach uses a Gamma-Poisson mixture model to capture the mutation rate heterogeneity across different individuals, and consequently models the over-dispersed mutation counts as a negative binomial distribution. Furthermore, the model regresses the mutation counts against 381 genomic features extracted from REMC and ENCODE to accurately estimate the local background mutation rate, which can be readily extended to accommodate additional genomic features. NIMBus was used to analyze 649 whole-genome cancer sequences. It successfully identified well-known coding and noncoding drivers, such as TP53 and the TERT promoter. In addition, NIMBus was also used for network based mutation burden analysis and successfully found known cancer related pathways, such as TP53 signaling and apoptosis pathways, to be significantly mutated.

Conclusion

NIMBus is a powerful tool to identify mutational hotspots as driver candidates in complex diseases so as to better allow biologists and clinicians to understand the underlying biological mechanisms of these diseaeses. We make NIMBus available and release our results as an online resource (nimbus.gersteinlab.org).