Introduction:

The infinite sites model of molecular evolution requires that every base in the genome is mutated at most once (Fig. 1). It is a cornerstone of (tumour) phylogenetic analysis, and is often implied when calling, phasing and interpreting variants3-5 or studying the mutational landscape as a whole. It is unclear however, whether this assumption holds in practice for bulk tumour samples.

Here we provide frameworks to model (Fig. 2) and detect (Fig. 3, 4) infinite sites violations, identifying 23,826 in total, including 6 candidate biallelic driver events, in 680 bulk tumour samples (25.6%) from the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes project (Fig. 5). Violations generally occur at mutational hotspots and their frequency and type can accurately be predicted from the overall mutation spectrum (Fig. 6). In melanoma, their local sequence context evidences how not only ETS, but also NFAT-family transcription factor binding creates hotspots for UV-induced cyclobutane pyrimidine dimer formation (Fig. 7). In colorectal adenocarcinoma, violations reveal hypermutable special cases of the trinucleotide mutational contexts identified in POLE-mutant tumours.

Results:

3. Parallel violations show an increased Variant Allele Frequency and aberrant phasing

4. Divergent violations are reliably called by Mutect2 but filtered out by default

7. Violations are enriched in specific contexts, revealing local mutational determinants

Conclusions

We reveal the infinite sites model breaks down at the bulk level for a considerable fraction of tumours. Violations in melanoma and POLE-mutant tumours often occur in specific hypermutable contexts. We provide frameworks to model and detect these violations, as well as recommendations for variant calling. These results also highlight the need for careful evaluation of current methods for reconstructing tumour genotypes, especially when scaling up to larger sets of mutations and lineages.