Controlling for different tri-nucleotide (or other) abundances in viewing and analyzing mutational spectra and signatures

Steve Rozen steverozen@gmail.com

We will start thinking about single nucleotide substitutions in trinucleotide context, but the concepts here generalize to other classifications of mutations.

To plot (or think about) a mutational spectrum, there are two broad approaches:

1. One can simply count up the number of mutations in each category (and plot that), and call it the mutational spectrum. The issue with this is that in the portion of a genome under consideration there
may be many more "opportunities" for one kind of mutation than for another.
Notably in many mammalian genomes, CG sequences are rare, so that
mutations from ACG, CCG, GCG, and TCG will be rare,
not because mutations from any give CG are rare, but because CG sequences are rare. If we are basing a spectrum on counts in the human genome, we would call this a "human-genome count" mutational spectrum.

2. It is also straightforward to conceptualize a mutational spectrum as the proportion of each trinucleotide that are mutated, so(can be deleted?) e.g. the proportion of(can be deleted?)
of ACGs that are mutated to AAG, to AGG, to ATG, etc. Each trinucleotide is an "opportunity" for
a mutation from that trinucleotide.
We refer to this as a "per-trinucleotide frequency" mutational spectrum.
(Since the proportion of any
given trinucleotide that is mutated is low, we usually plot per-trinucleotide frequencly(frequency?)
spectra as mutations per million trinucleotides.)
An advantage of this concept of a spectrum is that
spectra from exomes and genomes or from different organisms can be compared
directly, even though trinucleotide abundances differ between exome(exomes?) and genomes and between different
species. Also, this concept of spectrum reveals the sequence propensities of the
underlying mutational processes, since these propensities are not composed with the sequence-based
opportunities for processes to operate.

Now suppose that you want to compare count spectra from exomes and(to?)genomes, or from human and(to?) mouse genomes.
The solutions is(are?)to covert(convert?) to per-trinucleotide mutation frequencies and then back to the appropriate counts based on the target trinucleotide frequencies (opportunities).

Example: Convert from genome counts to counts as they would be in the exome

```
# Convert to "per trinucleotide frequency"
for each mutation type, t {
     per.trinuc.freq(t) <- genome.count(t) / opportunity.in.genome(t)
     # The opportunity of a mutation type e.g. ACT > ATT is the opportunity of ACT.
}

for each mutation type, t {
     inferred.exome.count(t) <- opportunity.in.exome * per.trinuc.freq(t)
}
```

This can be simplified to

```
for each mutation type, t {
    per.trinuc.freq(t) <- genome.count(t) / opportunity.in.genome(t)
     inferred.exome.count(t) <- opportunity.in.exome * per.trinuc.freq(t)
}
```

or equivalently, by algebraic simplification:

```
for each mutation type, t {
     inferred.exome.count(t)    <-    genome.count(t)    *    opportunity.in.exome(t)    /
opportunity.in.genome(t)
}
```

Now, to move on to signatures as opposed to spectra, let us take the
(correct) perspective that spectra due to a single exposure can be thought
of as a signature, provided we ignore the intensity of the exposure.
So to go from a spectrum due to one mutational process
to a signature based on *counts*, we divide the count of each mutation type by
the total number of mutations. To go from a spectrum based on
per-trinucleotide frequencies to a signature based on
per-trinucleotide frequencies we divide the
per-trinucleotide frequency of each mutation type
by the total of all per-trinucleotide
frequencies.

Converting between a signature based on genome counts to one based on per-trinucleotide frequencies:

```
for each mutation type, t {
     tmp(t) <- genome.count.proportion(t) / opportunity.in.genome(t)
}
for each mutation type, t {
  proportion.of.per-trinucleotide-frequency <- tmp(t) / sum_over_all_t(tmp(t))
}
```

You can probably fill in the missing pieces to get the
procedure for going from proportions based on
genome counts (i.e., a genome-count-based signature)
to proportions based on exome counts (i.e. an exome-count-based
signature):

```
for each mutation type, t {
     tmp(t) <- genome.count.proportion(t) * opportunity.in.exome / opportunity.in.genome(t)
}
for each mutation type, t {
  inferred.per-trinucleotide-proportion-in-exome <- tmp(t)  / sum_over_all_t(tmp(t))
}
```