**Comparing mutational signatures from different regions or organisms**

Steve Rozen steverozen@gmail.com

**Background:**

When comparing mutational signatures from different parts of the genome (e.g. an exome or the transcribed region of the genome) or from different organisms, we have to take into account that the trinucleotide composition of the sequenced area can differ. For example, CG dinucleotides are severely depleted in most of the mammalian genome (due to deamination of methylated cytosines). However, the area of the genome sequenced with exome sequencing contains a relatively high number of CpG-islands, which makes that CG dinucleotides are overrepresented in exome data compared to whole-genome data.

We can normalize for this by taking into account the “opportunity” (this is a term we use internally, more correct is to call it trinucleotide abundance, or bi-, tetra-, or penta-nucleotide abundance) of the sequenced area (e.g. the human exome) to the opportunity that we want to plot in (e.g. the human genome).

**Mutational signature representations:**

There are 2 ways to display mutational signatures:

1. The convention used in Alexandrov et al. in 2013 is to plot signatures relative to the opportunity of the human genome. The y-axis of such a plot would either be “counts” (if you have WGS data, no normalization has to be done to display this way, so you directly plot the counts) or “proportion”.
2. Alternatively, we can use the “standard flat” representation, which is plotting mutational signatures as if all trinucleotides are equally abundant in the genome. The y-axis of the plot generally is in mutations/million.

**Examples:**

To give an example of how to normalize for different opportunities:

1. If we have data from a human exome and we want to plot it in “standard flat” representation.

“Standard flat” representation is displaying assuming all trinucleotides are equally abundant in the genome. Therefore, to go from raw counts to “standard flat” representation, we have to divide the counts of mutations in a certain trinucleotide, but the number of times that particular trinucleotide exists in the sequenced area (in this case the human exome). The number we then have is the counts of mutations per 1 trinucleotide, so if we want to display as mutations/million, we multiply by 1\*106.

Putting that in a formula, we get:

# Convert to "Standard flat"

for each mutation type, t {

Standard.flat(t) <- exome.count(t) / opportunity.in.exome(t) \*1000000

}

1. If we have data from a mouse exome, and want to compare it to the PCAWG signatures (which are plotted to human genome opportunity)

In this case, we need to get from mouse-exome to human genome. To do this, we take the raw counts, divide it by the trinucleotide opportunity of the mouse exome (so then we have mutations per trinucleotide), and then multiply by the opportunity in the human genome. NB: we cannot plot this with “counts” on the y-axis, as these are normalized values, so the proper plotting scale now would be “proportions”, so we should divide each mutation class by the sum of mutation classes.

Putting that in formulas, we get:

# Convert to "mutations/trinucleotide"

for each mutation type, t {

mut.per.trinuc(t) <- counts(t) / opportunity.in.mouse.exome(t)

}

# Convert to "human genome opportunity"

for each mutation type, t {

mut.per.human.genome.freq(t) <- mut.per.trinuc(t) \* opportunity.in.human.genome(t)

}

# Convert to proportions

for each mutation type, t {

proportion.of.per.human.genome.freq{t} <-

mut.per.human.genome.freq(t) / sum(mut.per.human.genome.freq)

}