List of parameters

Wavelength:

Wavelength is the wavelength of emitted light by the molecules in the image in nanometer

NA:

NA is the numerical aperture of the observation objective and defines the resolution of the observation instrument

FragmentSize:

The fragment size is the size of the optical maps to be simulated (in pixels).

PixelSize:

This is the calibrated pixel size of the pixels in the image (in nanometer)

Enzyme:

The methyl-transferase enzyme, or restriction enzyme employed in the OM pipeline

NumTransformations:

This is the amount of times that a single genome is sampled. Sampling happens by sliding across the entire genome and sampling fragments of certain size. Therefore, the number of transformations refers to the amount of times that a single genome is sampled.

StretchingFactor:

The factor of overstretching of DNA, typically ranges from 1.7-1.8

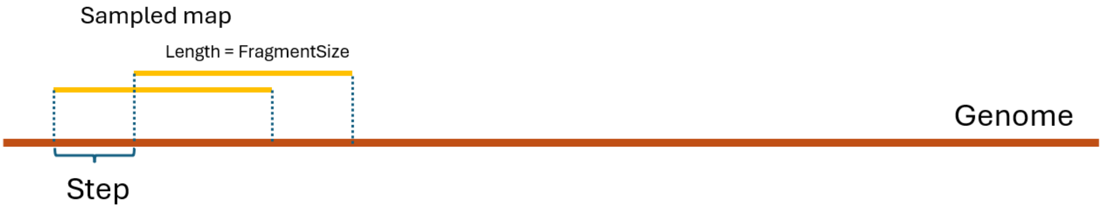
LowerBoundEffLabelingRate:

The lower bound on the labelling rate. We suggest to keep it at 75%, as going lower may make it significantly harder for networks to converge.

UpperBoundEffLabelingRate:

The upper bound on the labelling rate. We suggest to keep it at 100%, so that the network may be able to observe fragments that were not modified by the labelling rate.

Step:



As mentioned above, the genome is sampled by sliding a fragment across the genome and sampling from every location. 'Step' determines the distance in pixels between two 'slides', and can be seen as a stride between two sampling regions.

PixelShift:

During sampling, each dye is allowed to shift +- PixelShift from its original position as to introduce data augmentation.

NoiseAmp:

The relative noise amplitude compared to the signal of 1 dye. 0.2 therefore corresponds to 20% noise, or SNR of 5.

LocalNormWindow:

Local normalization window size. Local normalization aids in normalizing the dye intensity locally, allowing to correct for illumination artifacts. We suggest to keep that at 10000 bp.

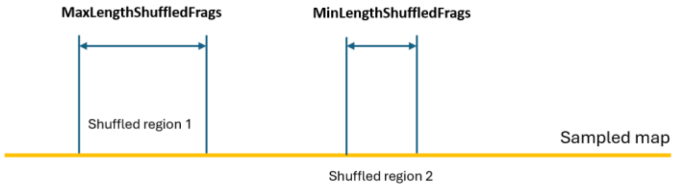
Min#ShuffledFrags:

Minimal number of shuffled regions. In an attempt to diversify the sequences present in the training dataset, and to increase the network generalization capability in presence of SV's, we introduce regions in the fragment where the dyes are randomly shuffled. The minimal number of these regions is determined by this parameter.

Max#ShuffledFrags:

Maximal number of shuffled regions. In an attempt to diversify the sequences present in the training dataset, and to increase the network generalization capability in presence of SV's, we introduce regions in the fragment where the dyes are randomly shuffled. The maximal number of these regions is determined by this parameter.

MinLengthShuffledFrags:



Each of the shuffled regions has a certain length. This is the minimal shuffled length for shuffled region. The total percentage of shuffled regions within a given fragment can therefore be estimated as: $P = \frac{\sum_i^N L_i}{L_g}$, with L_i being the length of each shuffled region, N, the number of shuffled regions. We suggest to choose the number and length of shuffled regions as to keep this ratio below 50%.

MinLengthShuffledFrags:

Each of the shuffled regions has a certain length. This is the maximal shuffled length for shuffled region. The total percentage of shuffled regions within a given fragment can therefore be estimated as: $P = \frac{\sum_i^N L_i}{L_g}$, with L_i being the length of each shuffled region, N, the number of shuffled regions. We suggest to choose the number and length of shuffled regions as to keep this ratio below 50%.

Random:

Whether to simulate random fragments as well.

RandomLength:

The number of random fragments. We suggest to keep the number of random genome fragments approximately equal to the number of fragments of the genome of interest, as to obtain a 50/50 ratio.

FPR:

The false positive labelling rate. The choice of false positive labelling rate was based on the random dye deposition in the background, which influenced the FP rate the most. The measured FPR was around 0.7 labels/kb.

FPR2:

The double false positive labelling rate. This parameter was introduced to allow for variation in the dye background intensity and was kept at 0.2 labels/kb, inferred from the background of experimental images.

Random-min:

The minimal number of dyes on a random fragment.

Random-max:

The maximal number of dyes on a random fragment. We do not suggest to go beyond 6-7/kb as there are not many bacterial species that have such density of TCGA sites.