# bsDoctor: a versatile tool to quality diagnosis for single-cell bisulfite-sequencing data

Functionalities

Base, read, and mapping quality

Coverage

Depth

chromosome

strand-specific

uniformness: chr, chr location, replicated sequences

DNA templet lose

Saturation curve

Bisulfite conversion

Lambda DNA

Mt

CHH

mt

DNA copy number

Error rate

methylation level

bias: C and mC, methylation (of different genomic features) vs coverage

methylation variation

pan-gene, pan-CGI

strand-specific

uniformness

error

appliable for WGBS and RRBS

diagnosis for each sample and for distributions of all samples

report

a html report with interactive figures and tables

figures can be saved as .svg or pixel figures

shutil

html template: python jinja2

~~interactive figures: highcharts~~

base64.b64encode

comparison of different single-cell methylome protocols

scBS-seq, scRRBS, 低覆盖高通量技术

高通量高内涵药物/靶点筛选

find the direct associations in microbiome/virome-wide association study with graphical lasso

gene regulation network (direct association between DNAme and RNA/protein expression)

for multi-omics integration

understanding the genome with sequence and annotation

analyses during sampling

read-wise

**Quality**

Base quality (position-free)

read quality (~~mean base quality~~, number of low-quality bases)

mapping quality

cigar proportions

**Coverage**

Whole-genome coverage rate, depending on depth

Coverage rate of each chromosome

Binning coverage of each chromosome

strand-specific coverage: whole genome, whole chr, whole bin

~~uniformness: chr, chr location, replicated sequences~~

**DNA templet lose**

Saturation curve, from depth threshold dependent coverage, DP>=1

**Bisulfite conversion**

Lambda DNA, from reads of lambda DNA

Mt, from reads of MT

CHH, from CHH-context cytosine

**mt**

DNA copy number, from coverage of whole genome and MT

**methylation level**

same results of coverage, all reads of genomic sites

methylation level distribution, 20 bins of [0, 1]

pan-gene, from all reads of specific genomic sites

~~genomic features, from all reads of specific genomic sites~~

**Error**

~~Mismatches, from reads~~

Base error rate, from all reads mapped to MT

~~C/T error rate, from inconsistent reads with methylation status (sc only)~~

**bias**

methylation vs depth

coverage vs kmer

methylation (of different genomic features) vs coverage

**coverage and methylation of kmer, XXCGXX, 4^4=256**

methylation vs **technical variables ?**

~~methylation variation~~

**strandness**

CpGs of double-stranded coverage

stranded coverage of CpGs, difference distribution

stranded methylation of CpGs, difference distribution

**for RRBS sample, coverage vs ideal CCGC sites**

**report of samples**

**distribution of metrics**

sampling over a set of genomic sites

whole genome -> an interval -> all reads mapped onto the interval -> a read -> bases of the read

or all reads in the sample, maybe > 10h