Auto-Karyotyping Report for "DN18249"

If you have any queries relating to these data, please contact robert.goldstone@crick.ac.uk who will be able to put you in touch with best person(s) to help you.

Sample_Name	LIMS_ID	Genome_Tag	Genome_Build	Raw_Yield	Tot_Aln	Percent_Aln	FLAGS
AA004_C12	NIC216A687	Homo sapiens	GRCh38-r89	87210865	86927946	99.68	NA
AA004_C7	NIC216A688	Homo sapiens	GRCh38-r89	84340059	84155444	99.78	NA
AA004_C8	NIC216A689	Homo sapiens	GRCh38-r89	78029175	77821910	99.73	NA
AA004_ER10	NIC216A690	Homo sapiens	GRCh38-r89	101288153	101024186	99.74	NA
AA002_CL3G8G11	NIC216A691	Homo sapiens	GRCh38-r89	64209663	64058053	99.76	NA

Description

Following alignment to the reference genome, copy number estimation was performed using the QDNASeq package:

- Scheinin et al "DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and eclusion of problematic regions in the genome assembly."

In brief, the genome is subdivided into bins of fixed width (1000kb by default) and the number of reads mapping within each bin is calculated.
These "raw" counts are then corrected for local GC content and mapability by estimating the median count across bins of the same GC content and mapability.
"Smoothing" is then performed by fitting a LOESS surface through the medians.
Finally, the raw counts are corrected by dividing the count for each bin by the LOESS fit corresponding to its GC and mapability.
The plots in the associated PDF files display the log2 transformed ratios.

One might think of this process as using each sample to estimate its own "expected" number of reads for a given GC-content and mapability conbination, with the smoothing
mitigating outliers based on the assumption that regions of identical GC-content and similar mapability should have similar counts, and simultaneously,
regions of similar GC-content and identical mapability should also have similar counts.
Presentation on the log2 ratios of raw counts to "expected" then allows for the standard interpretation: a log2 ratio of 1 indicates that that specific bin has twice as many reads as expected given the other bins for *that sample*
It should be noted then that these plots do not necessarily indicate "absolute" copy number estimates -- for example, in the event that the entire genome is doubled, all regions will be equally over-represented, leading to log2 fold changes of zero rather than 1. Further, the absolute values of the log fold change estimates are ***not*** directly comparable between samples.

Finally: the sex chromosomes are explicitly excluded from the median and smoothing process described above.

Software Version Info:

Rscript	R scripting front-end version 3.5.1 (2018-07-02)
FastQC	FastQC v0.11.8
Trimgalore	Quality-/Adapter-/RRBS-/Speciality-Trimming [powered by Cutadapt] version 0.6.0 Last update: 01 03 2019
bwa	Version: 0.7.15-r1140
samtools	Version: 0.1.19-44428cd