A normalization technique for next generation sequencing experiments

Günter Klambauer[1], Karin Schwarzbauer[1], Andreas Mayr[1], Sepp Hochreiter[1]
[1] Institute of Bioinformatics, Johannes Kepler University Linz, 4040 Linz, Austria

Abstract

Motivation: Next generation sequencing (NGS) are these days one of the key technologies in biology. NGS' cost effectiveness and capability of finding the smallest variations in the genome makes them increasingly popular. For studies aiming at genome assembly, differences in read count statistics do not affect the outcome. However, these differences bias the outcome if the goal is to identify structural DNA characteristics like copy number variations (CNVs). Thus a normalization step must removed such random read count variations subsequently read counts from different experiments are comparable. Especially after normalization the commonly used assumption of Poisson read count distribution in windows on the chromosomes is more justified. Strong deviations of read counts from the estimated mean Poisson distribution indicate CNVs.

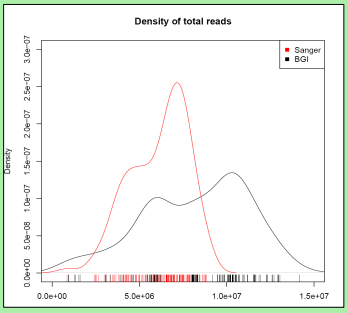
Results: We test our normalization technique on sequencing results from three different sequencing centers with a wide range of quality levels. After normalization, regions that deviate from the estimated Poisson distribution are have been identified as sex specific or previously identified CNV regions.

Motivation of Normalization

Without normalization: assumption of Poisson read count distribution not justified
- different number of total reads
- Reads with multiple mapping positions:
  - Excluding: underestimated read counts → wrongly detected deletions
  - All possible matches: overestimated read counts → wrongly detected amplifications

Normalization is essential for NGS quantitative data analysis

Biases in Sequencing data due to lane quality and other effects



Density of the total number of reads

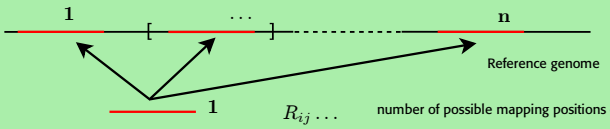
Genomic position	BGI_FC20CA5AAXX2	BGI_FC_20CA6AAXX1	BGI_FC20CA6AAXX2
Chr1:200,000-250,000	7317	9025	9165
Chr1:250,000-300,000	321	442	450
Chr1:300,000-350,000	5504	6303	6453
Chr1:350,000-400,000	22949	26791	28536
Chr1:400,000-450,000	13954	15583	16259
Total mapped reads	6,012,023	7,495,175	7,830,023

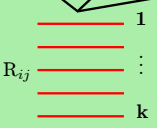
Data matrix of unnormalized read counts.

Suggested Normalizaton Procedure

Normalization Step 1: Reads with multiple mapping positions

Multiplicity of Reads





R_{ij} ... number of possible mapping positions of read

$\mathbf{n}(R_{ij})$ ... the i-th read of sample j aligning to a certain window

$\mathbf{k}(R_{ij})$ ... number of reads of sample j with identical sequence as R_{ij}

Reads with multiple mapping positions

Normalized read count:

$$\hat{x}_j = \sum_{i=1}^{l} \frac{\mathbf{k}(R_{ij})}{\mathbf{n}(R_{ij})}$$

Usually only the number of reads in the windows are just counted

Normalization Step 2: Lane effect

Each lane shows different characteristics
- → different number of reads mapped back
- → normalization of the read counts per lane (by the number of reads which were mapped back)
- → quality and the number of reads implicitly considered

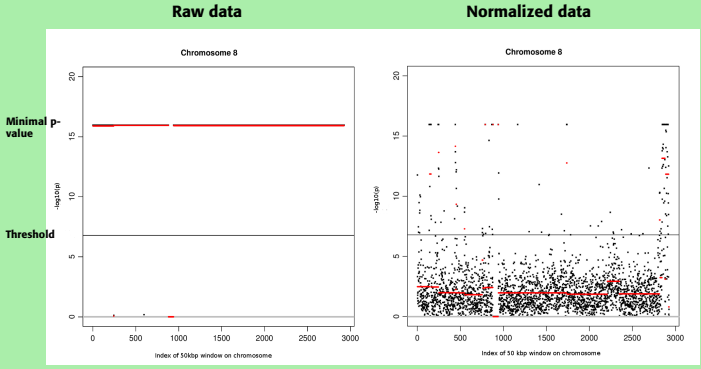
T_j ... number of total aligned reads of sample j

Normalized read count of sample j in window i: $\overline{x}_{ij} = x_{ij} \cdot \frac{\min_j T_j}{T_j}$

Experiments on HapMap data

Normalization justifies the Poisson distribution Assumption

P-values of a Poisson test before and after normalization



-log10 p-values of a test for Poisson distribution on 45 HapMap samples. Left: Raw read counts. Right : Normalized read counts.

Testing the Poisson Assumption

Poisson test: Brown and Zhao then Bonferroni correction of the p-values.

Rejection rate of the Poisson Assumption

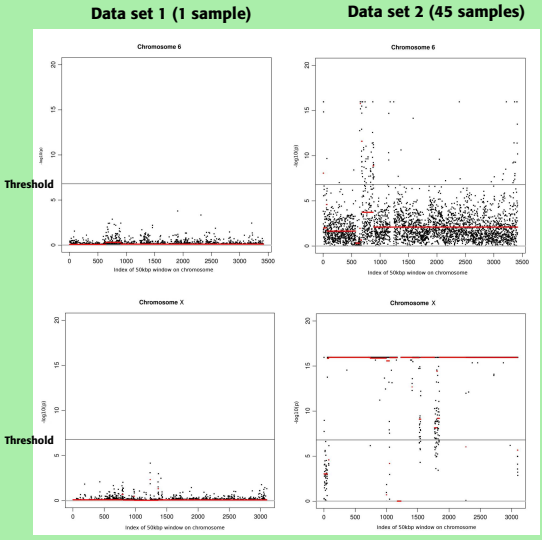
	Data set 1 (46 lanes of one sample)	Data set 2 (45 lanes of 45 samples)
Raw data	93,1 %	93,1 %
Multiple reads normalization	92,9%	73,6%
Lane effect normalization	84,9%	18.4%
Both normalizations	34,0%	**0.1%**

Rejection rate of the Poisson hypothesis by a Poisson test.

Data sets

From 1000 Genomes on HapMap samples sequenced by the Solexa Genome Analyzer
- Data set 1: lanes from single sample NA19328
- Data set 2: lanes from 45 different samples

P-values of the Poisson test



-log10 p-values of a test for Poisson distribution on the two data sets. Left: data set 1(single sample). Right : data set 2 (45 samples)