Dealing with the GC-content bias in second-generation DNA sequence data

With Yuval Benjamini, UC Berk<mark>el</mark>ey

School of Mathematical Sciences, University of Adelaide, Fride



- Difeas

GC matters



First, note base pairing: A with T, C with G G with C, T with A



Three hydrogen bonds



Two hydrogen bonds



aaatatattc aatatggaga gaataaaaga actgagagat ctaatgtcgc tcgcgagata ctcactaaga ccactgtgga ccatatggcc ataatcaaaa aggaaggcaa gagaagaacc ccgcactcag aatgaagtgg atgatggcaa aattacagca gacaagagaa taatggacat gattccagag aggaatgaac cctctggagc aaaacaaacg atgctggatc agaccgagtg atggtatcac aacatggtgg aataggaatg gcccaacaac aagtacagtt cattacccta aacttatttc gaaaaggtcg aaaggttgaa acatggtacc ttcggccctg aaatcaagtt aaaataagga ggagagttga tacaaaccct ggccatgcag caaggaggca caggatgtga ttatggaagt tgttttccca aatgaagtgg actgacatca gagtcacagc tggcaataac aaaagagaag aaagaagagc taaaattgct cccttgatgg tggcgtacat gctagaaaga gaattggtcc gtttctccca gtagccggcg gaacaggcag tgtttatatt gaagtgttgc agggacgtgc tgggagcaga tgtacactcc aggaggagaa gtgagaaatg ccaaagtttg attatcgctg ctagaaacat agtaagaaga gcagcagtgt attagcatct ctcttggaaa tgtgccacag cacacagatt ggaggagtaa catcettaga cagaateeaa etgaggaaca ageegtagae atatgeaagg gttgaggatt agctcatctt tcagttttgg tgggttcact ttcaaaagga

atcagtcaag aaagaagaag aagtgctaac gggcaacctc caaacactga acatgaaggg tatgaagaat tcacaatggt tgggagaaga gcaacagcta ggcaaccagg agattgatcc agttgatagt aagcgggaga gacgagcagt

caaggaggca caggatgtga ttatggaagt tottttccca aatgaagtgg

aaatatattc aatatggaga gaataaaaga actgagagat ctaatgtcgc tcgcgagata ctcactaaga ccactgtgga ccatatggcc ataatcaaaa aggaaggcaa gagaagaacc ccgcactcag aatgaagtgg atgatggcaa aattacagca gacaagagaa taatggacat gattccagag aggaatgaac cctctggagc aaaacaaacg atgctggatc agaccgagtg atggtatcac aacatggtgg aataggaatg gcccaacaac aagtacagtt cattacccta aacttatttc gaaaaggtcg aaaggttgaa acatggtacc ttcggccctg aaatcaagtt aaaataagga ggagagttga tacaaaccct ggccatgcag

actgacatca GC content here 445/1000 taaaattgct ______ gaattggtcc gtttctccca gtagccggcg gaacaggcag tgtttatatt gaagtgttgc agggacgtgc tgggagcaga tgtacactcc aggaggagaa gtgagaaatg ccaaagtttg attatcgctg ctagaaacat agtaagaaga gcagcagtgt attagcatct ctcttggaaa tgtgccacag cacacagatt ggaggagtaa catccttaga cagaatccaa ctgaggaaca agccgtagac atatgcaagg gttgaggatt agctcatctt tcagttttgg tgggttcact ttcaaaagga atcagtcaag aaagaagaag aagtgctaac gggcaacctc caaacactga acatgaaggg tatgaagaat tcacaatggt tgggagaaga gcaacagcta ggcaaccagg agattgatcc agttgatagt aagcgggaga gacgagcagt

aaagaagagc

8

An illustration of the spatial distribution of GC content of non-overlapping 1,024-bp windows along a fragment, approximately 1.4 Mb in length, from human chromosome 19



© The Author 2005. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved. For permissions, please e-mail: journals.permissions@oupjournals.org





Illumina short-read DNA sequencing

Illumina Sequencing Technology



Image acquisition

Slide courtesy of G Schroth, Illumina

	chr12:58,145,445-58,145,490 p13.32 p13.2 p12.3 p12.1 p11.21 q12 q13.11 q13.2 q14.2 q15 q21.2 q21.32 q22 q23.2 q24.11 q24.22 q24.32	=
		н90 bp
hg19_DNA_778_merged Q_only.bam Coverage hg19_DNA_778_merged Q_only.bam	$\begin{bmatrix} 0 - 665 \end{bmatrix}$	
35+ I	c p reads G T C C C A T A G G C A C C G A C A C C A A T T T C A G C C A C T G G C T C A T A T C G A T G Y A G V G I E A V P E Y R CDK4	G A O
	Reads mapped to the reference genome	×

Our original motivation: DNA copy number estimation

- Amplifications and losses occur in most genomes, relative to a reference, especially in tumor genomes, which are our focus.
- These can be seen at different scales, from one or a few basepairs to entire chromosomes (in cases of aneuploidy).
- We would like to measure (count) the number of times a region appears in a genome using DNA-seq data on that genome, ideally without using reference or matched normal data.



24 color karyotypes of 6 breast cancer cell lines

Davidson et al BrJCa, 2000



Sensitivity of Noninvasive Prenatal Detection of Fetal Aneuploidy from Maternal Plasma Using Shotgun Sequencing Is Limited Only by Counting Statistics

H. Christina Fan, Stephen R. Quake*

Department of Bioengineering, Stanford University and Howard Hughes Medical Institute, Stanford, California, United States of America

Abstract

We recently demonstrated noninvasive detection of fetal aneuploidy by shotgun sequencing cell-free DNA in maternal plasma using next-generation high throughput sequencer. However, GC bias introduced by the sequencer placed a practical limit on the sensitivity of aneuploidy detection. In this study, we describe a method to computationally remove GC bias in short read sequencing data by applying weight to each sequenced read based on local genomic GC content. We show that sensitivity is limited only by counting statistics and that sensitivity can be increased to arbitrary precision in sample containing arbitrarily small fraction of fetal DNA simply by sequencing more DNA molecules. High throughput shotgun sequencing of maternal plasma DNA should therefore enable noninvasive diagnosis of any type of fetal aneuploidy.

Citation: Fan HC, Quake SR (2010) Sensitivity of Noninvasive Prenatal Detection of Fetal Aneuploidy from Maternal Plasma Using Shotgun Sequencing Is Limited Only by Counting Statistics. PLoS ONE 5(5): e10439. doi:10.1371/journal.pone.0010439

Editor: Joanna Mary Bridger, Brunel University, United Kingdom

Received January 22, 2010; Accepted April 6, 2010; Published May 3, 2010



Our story proper begins here

GC content biases first noted

Published online 26 July 2008

Nucleic Acids Research, 2008, Vol. 36, No. 16 e105 doi:10.1093/nar/gkn425

Substantial biases in ultra-short read data sets from high-throughput DNA sequencing

Juliane C. Dohm¹, Claudio Lottaz², Tatiana Borodina¹ and Heinz Himmelbauer^{1,*}

¹Max Planck Institute for Molecular Genetics, Ihnestr. 63-73, 14195 Berlin and ²Institute for Functional Genomics, Computational Diagnostics, University of Regensburg, Josef-Engert-Str. 9, 93053 Regensburg, Germany

Received December 21, 2007; Revised June 16, 2008; Accepted June 19, 2008

They used Illumina 1G data.

From Dohm et al 2008

Roughly linear GC effect on reads



Another view: a human data set

Reads mapped to Chrom. 2 (both ends mapped)



- Position of reads on forward strand of chr 2
- Binned to 10 kb intervals

The GC bias is non-linear in human data (5 kb bins below, but it looks similar for all bin sizes)



Horizontal axis: fraction GC; lines are loess curves in all cases

Our goals

- To study the nature of the GC content effect
- Find how best to correct for it
- Try to understand relation between the effect and study design, i.e. its causes
- Perhaps identify designs that minimize it.

Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. *Genome Biol.* 2011 Feb 21;12(2):R18.

Systematic bias in high-throughput sequencing data and its correction by BEADS. Cheung MS, Down TA, Latorre I, Ahringer J. *Nucleic Acids Res.* 2011 Jun 6. [Epub ahead of print]

Our goal today

- To study the nature of the GC content effect
- Find how best to correct for it

Digression: mappability

- Some % of reads not mapped due to ambiguity (depends on read length & mapping criteria)
- Mappability = the probability that a read beginning in region can be *successfully* mapped.
- Can take a simple 0-1 approach (as here), and bin.



Our data

Two samples of DNA from an ovarian patient: one sample from the tumor, the other matched normal from their white blood cells.

Each sample was turned into two separate fragment libraries, differing in fragment length distribution.

Fragments were sequenced to 75bp at both ends using the standard Illumina procedure.

Each sequenced read pair was mapped back to the human reference genome using bwa (version0.4.9). [A few more details are omitted here.]

Most of the time we present results for just one chromosome

But it doesn't matter....

GC loess curves for chromosomes 1-5, 10kb bins



Counts for library 1 scaled to match those for library 2

Is the GC-bias specific to a lab, protocol, sample, library preparation, sequencing machine,....?

E.g. can we adjust binned tumor counts by those of a matched normal sample, or, in a Chip-seq experiment, IP-counts by input of other control counts?



Conclusion: the effect seems largely to be run specific.

Is there a right bin size? People have used 100bp, 5 kb, 10 kb, 20 kb, 100 kb.

Variation about loess curve for different bin sizes

Loess bin size (kb)	10	5	2	1	0.5	0.2
Library 1 (MAD)	49.1	47.8	45.1	43.4	43.4	52.2
Library 2 (MAD)	26.0	24.7	22.5	21.7	23.6	41.6

Avoiding binning: single position analyses (also done by Cheung *et al*, 2011)

We work with a *random sample* of ~10M *mappable* locations on the genome locations denoted by *x*. All paired end, and forward strand, unless otherwise stated.

The fragment count at location x may depend on the GC content of the window $W_{a,l} = [x+a, x+a+l)$, which we will denote by gc = GC(x+a,l).

A) Random sample locations B) Partition by GC window C) Count reads and read-rate





Here the window begins at the base of interest. It need not do so. 34

What's interesting about these *read rate* vs *GC-content* curves as we vary window size and location?

Superficially: their shape, that is, their deviation from flatness, which is GC-independence. More interestingly, their ability to help explain variation in read depth. We return to this later. Let's keep it superficial for now, and measure deviation from flatness.

Total variation distance from GC independence.

A surrogate measure of how much is explained by conditioning on GC.

Reads by GC, 32bp Window



TV distance = a weighted average of the brown lengths₆

In symbols,

$$\hat{\lambda}_{gc} = \frac{F_{gc}}{N_{gc}}, \quad \hat{\lambda} = \frac{F}{n}$$
$$TV(W_{a,l}) = \frac{1}{2\hat{\lambda}} \sum_{gc=0}^{l} \frac{N_{gc}}{n} |\hat{\lambda}_{gc} - \hat{\lambda}|,$$

where $W_{a,l}$ is the window [x + a, x + a + l).

Next we look at some *TV* values. We can vary *a* and *I*, and we do so, separately here, for simplicity

Varying the window size from a fixed point (here the 5'-end of the fragment)

TV of models from fragment 5' end



GC window length (I)

Varying the window size from a fixed point (here the 5'-end of the fragment)



GC Curve for best window (a=2, l=176)

Varying the location of a fixed size window (here 50 bp; library 1)



Interim conclusion from many such plots

The "best" interval is in the middle of the fragment, excluding the bits at the very ends (see later).

Next steps: dealing with both strands, and fragment size.

Forward and reverse strands behave similarly



TV Scores of each strand (50 bp windows)

Forward and reverse strands behave similarly



Stratifying by fragment size s

$$\hat{\lambda}_{gc}^{s} = \frac{F_{gc}^{s}}{N_{gc}^{s}}$$

Fragment size matters



Conclusion: GC bias of not simply determined by the ratio GC count/fragment length: there is an interaction.

And now for some predictions

Predicted rates at a given mappable position

$$\hat{\mu}_{x} = c \sum_{s} \hat{\lambda}_{GC(x+a,s-m)}^{s}$$
$$\hat{\mu}_{B} = \sum_{x \in B} \hat{\mu}_{x}$$

Here *c* is a scaling constant to equalize the predicted and the observed median. From now on, our window is the fragment minus 2 bp at each end, i.e. a=2, l=s-2.

Predicted and observed bin counts for bins of different sizes



Lowess lines are based on the observed points.

Conclusion: the predictions seem to be working.

Some other biases/models

Breakpoint effects



Breakpoint model: uses GC(x-2,x+4)

50

End effects



Slight AT preference

D Two ends model: uses *GC(x,I)+GC(x+s-I,I).* We use *s=180/I=30* below. ⁵¹

Some cher prediction (all aggregated to 1kb bins

Counts

52

Counts



Fraction GC in all cases

Conclusion: These predictions don't work too well.

How well does our correction work?

Theory

Spread of observed counts around predictions



Predicted Counts

Conclusion: we don't "explain" everything.54

How well does our correction work?

Practice

Copy number: corrections to normal samples





Slight improvement over loess.

Copy number: crude corrections to tumor samples



Summary

- We seem to have ruled out GC-content of the read parts of the fragment as producing the GC bias.
- Similarly we seem to have ruled out GC content on a scale more "global" than just the fragment.
- Base composition (not just GC-content) around the two fragment break points plays a noticeable role, but not enough to explain everything.
- Speculation over causes is left for another day. There now seems little doubt that PCR amplification bias accounts for the majority, as shown in a beautiful recent paper by D. Aird *et al* (2011) in the Feb 21 issue of *Genome Biology*.

Many thanks to

Yuval Benjamini

- Oleg Mayba, Pierre Neuvial, Henrik Bengtsson, and Su Yeon Kim
- Paul Spellman and Mark Robinson for data
- Leath Tonkin for discussions on the bias
- The whole Berkeley NGS group

And to you...

ChIP-seq data (*A. thaliana***)** Here two initially incompatible *technical* replicates



Problem mainly solved (cf Cheung et al, 2011)

Other examples and phenomena (if time)

Plots for a BrCa tumor



Plots for ChIP-seq sample rep 1, A. thaliana



Plots for one 1,000 genomes sample



Fragments partitioned by quality score

