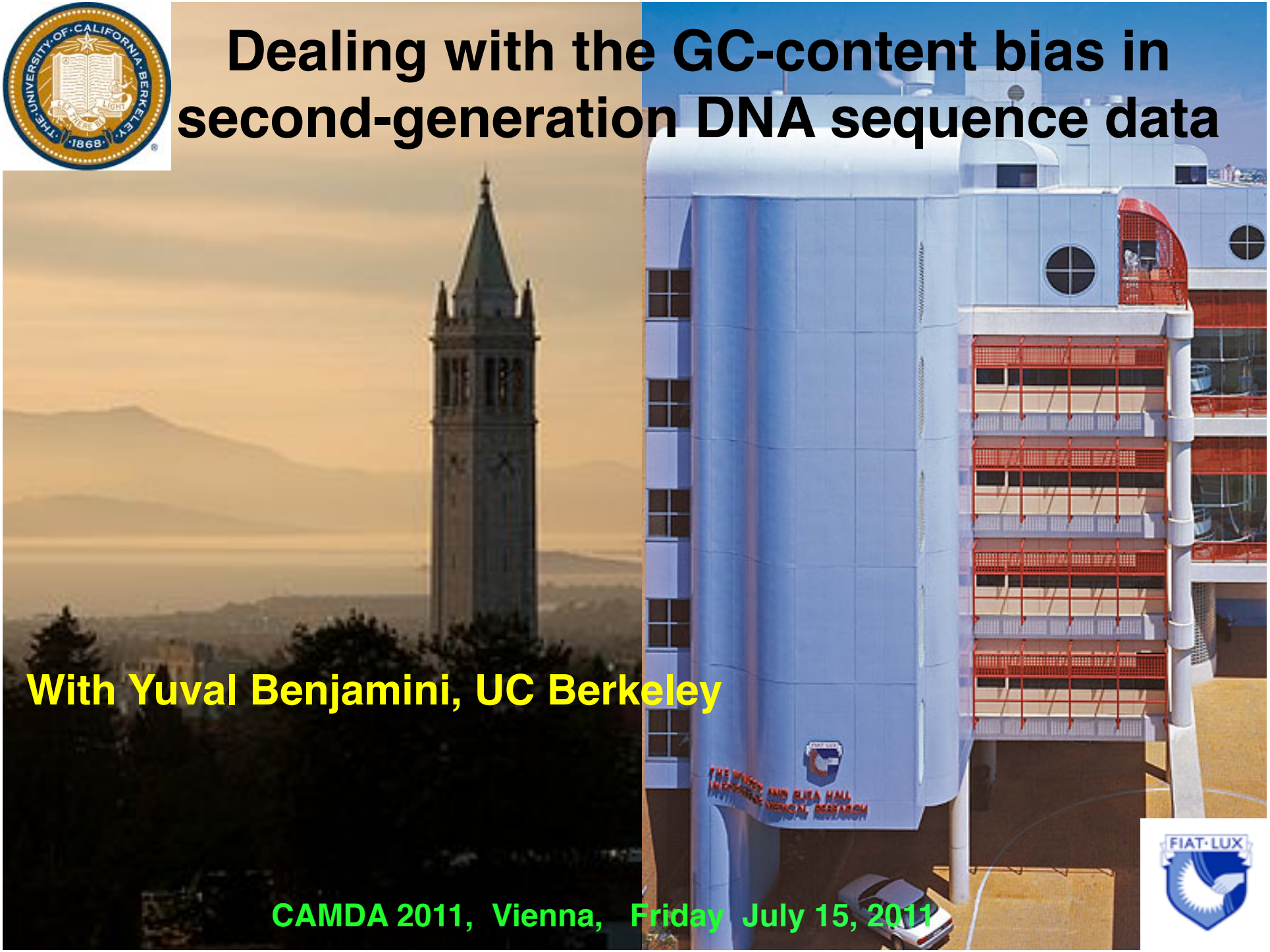
# Dealing with the GC-content bias in second-generation DNA sequence data



With Yuval Benjamini, UC Berkeley

CAMDA 2011, Vienna, Friday July 15, 2011

# 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. Dohm[1], Claudio Lottaz[2], Tatiana Borodina[1] and Heinz Himmelbauer[1,*]**

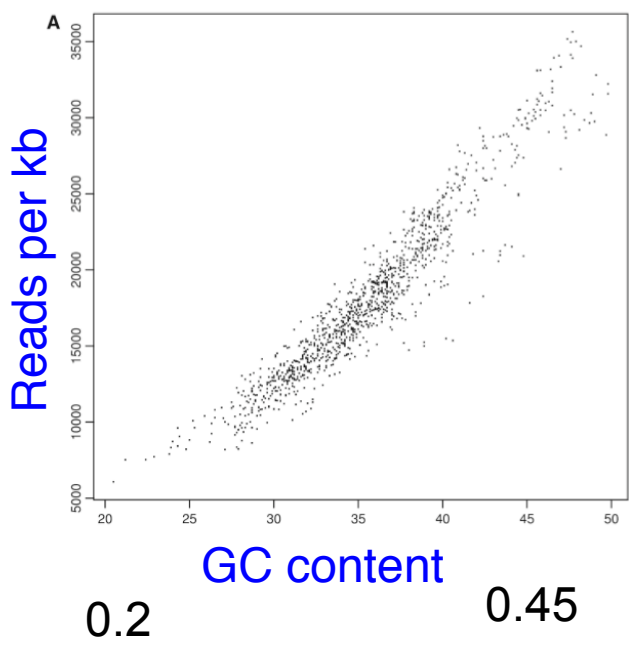
[1]Max Planck Institute for Molecular Genetics, Ihnestr. 63-73, 14195 Berlin and [2]Institute for Functional Genomics, Computational Diagnostics, University of Regensburg, Josef-Engert-Str. 9, 93053 Regensburg, Germany

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

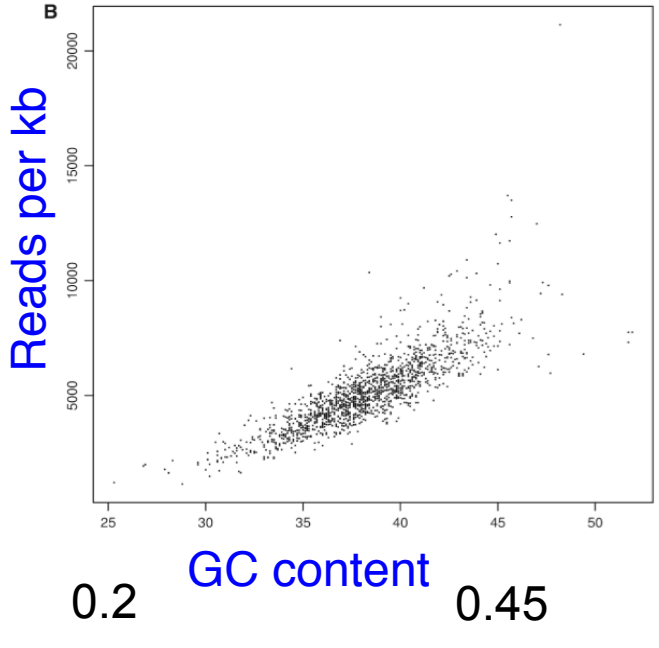
They used Illumina 1G data.

# From Dohm *et al* 2008

Roughly linear GC effect on reads

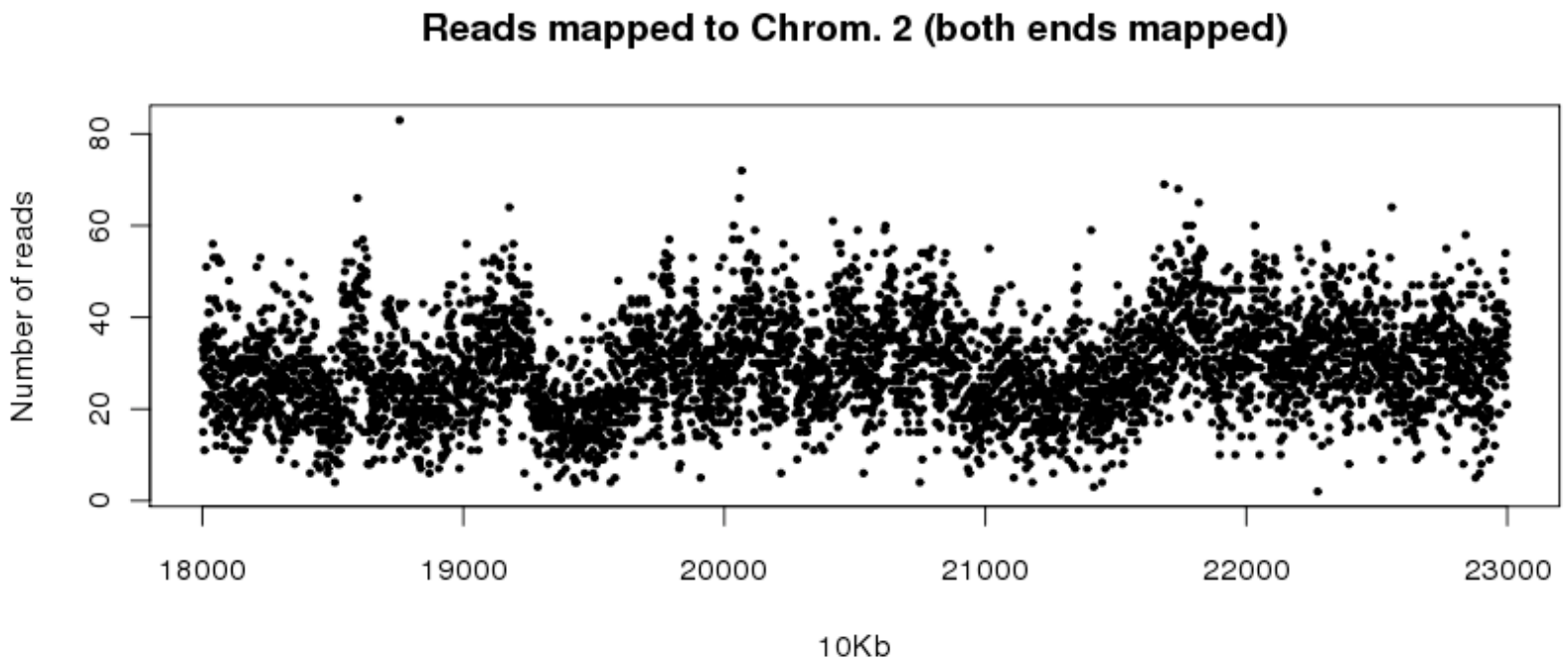



*Beta Vulgaris* 1kb bins




*Helicobacter* 1kb bins

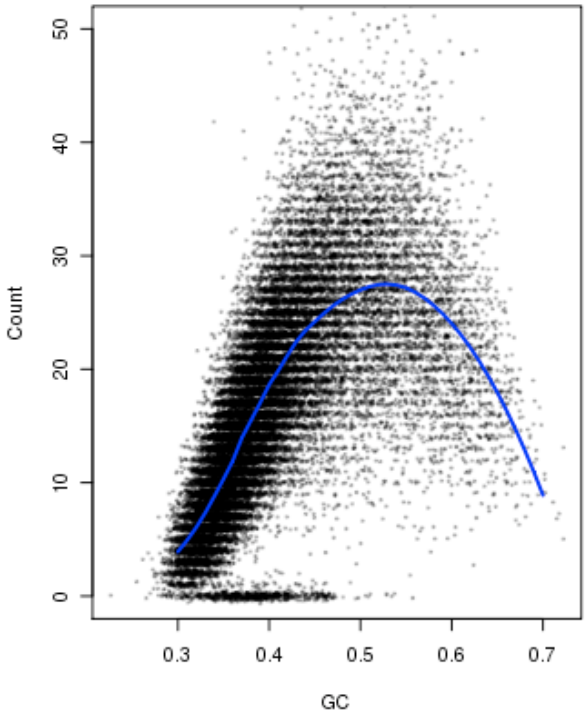
# Another view: a human data set



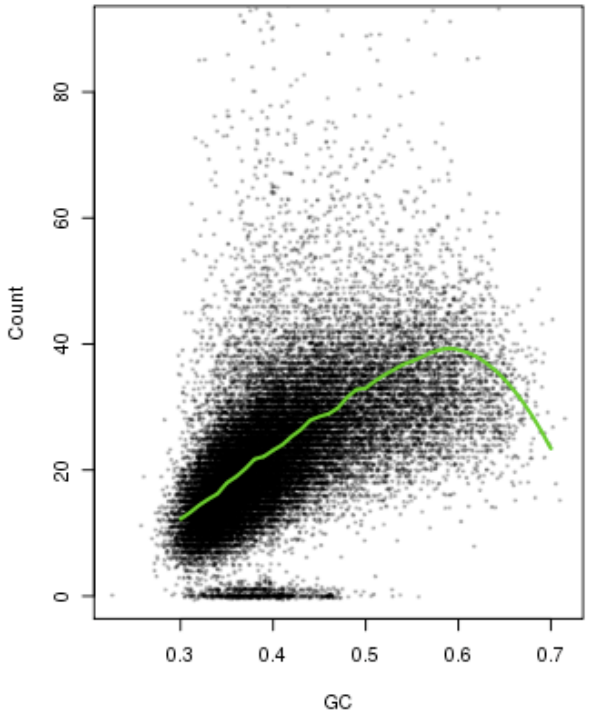

- Position of reads on forward strand of chr 2
- Binned to 10 kb intervals

# The GC bias is non-linear in human data

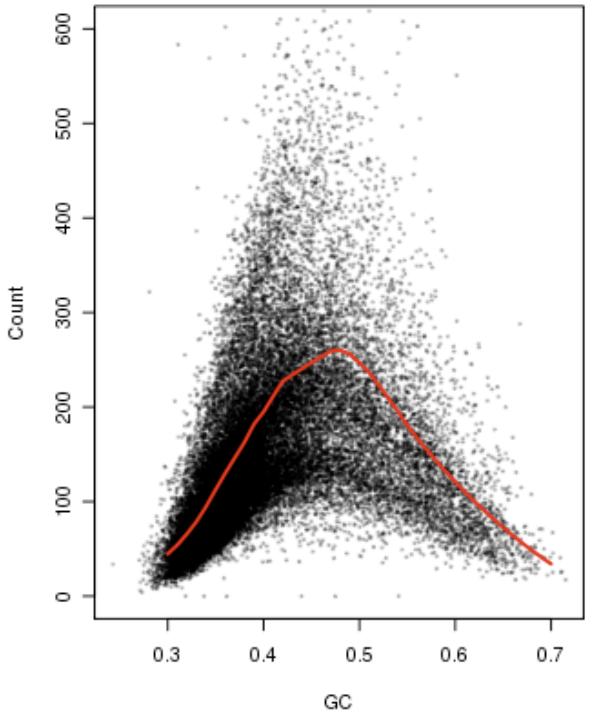
## (5 kb bins below, but it looks similar for all bin sizes)




Data – M. Robinson




Data – D. Chiang




Data – P. Spellman

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

# Our goals

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

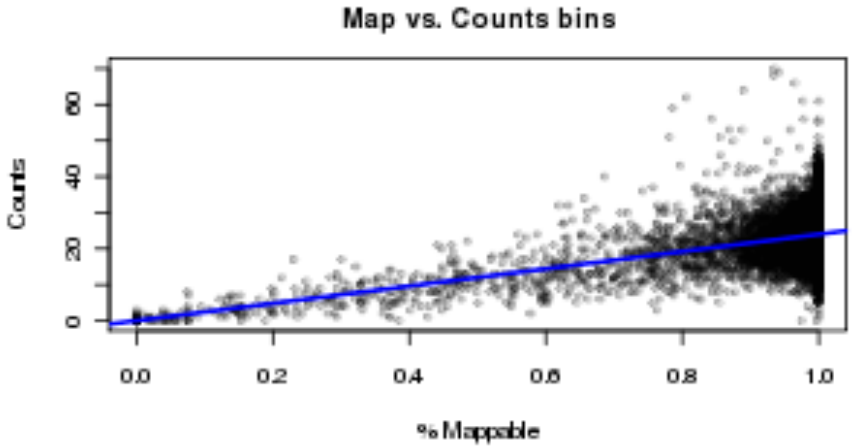
See especially:

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]

# 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.

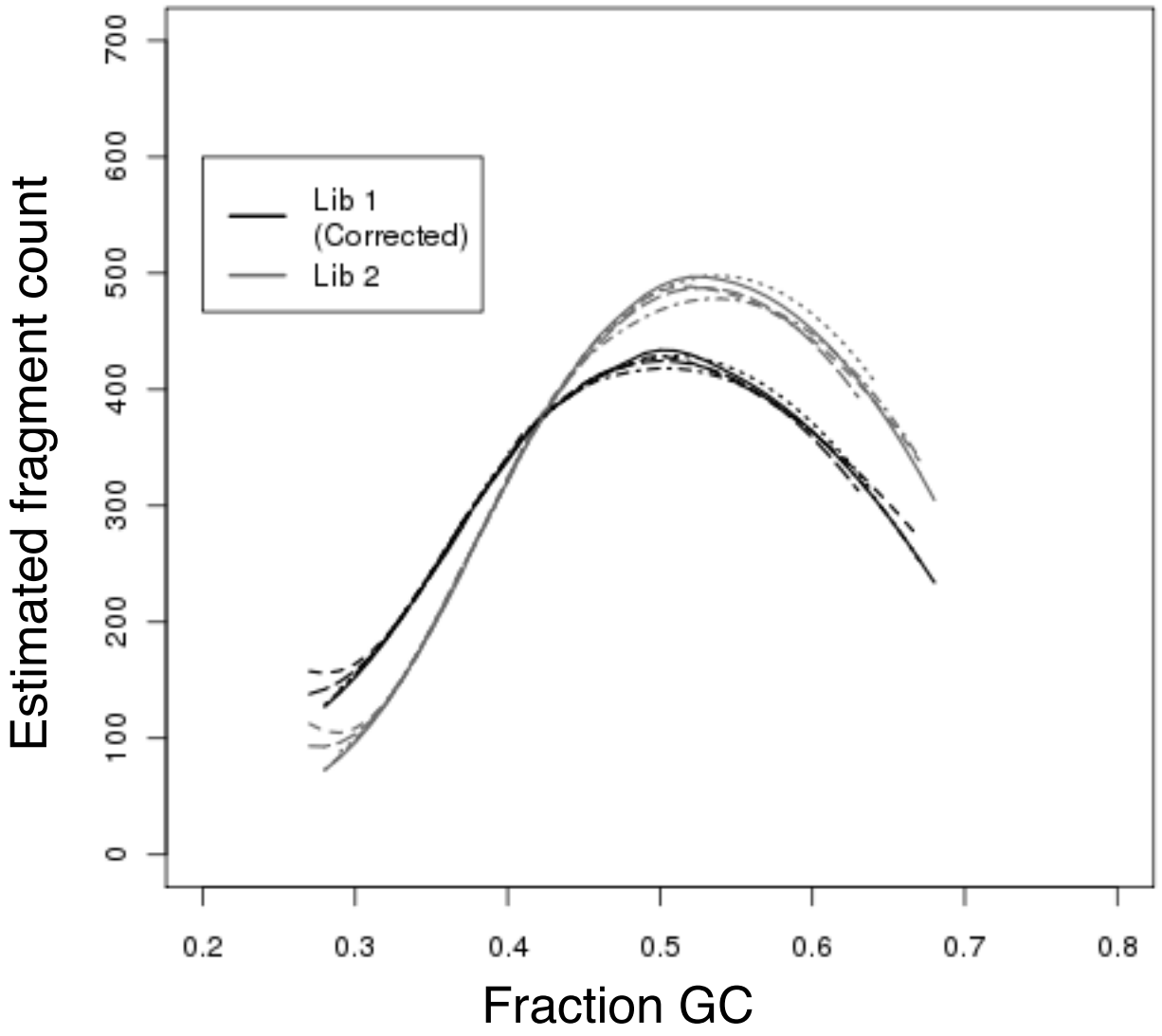
Fragments were sequenced to 75bp at both ends using the standard Illumina procedure.

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….**

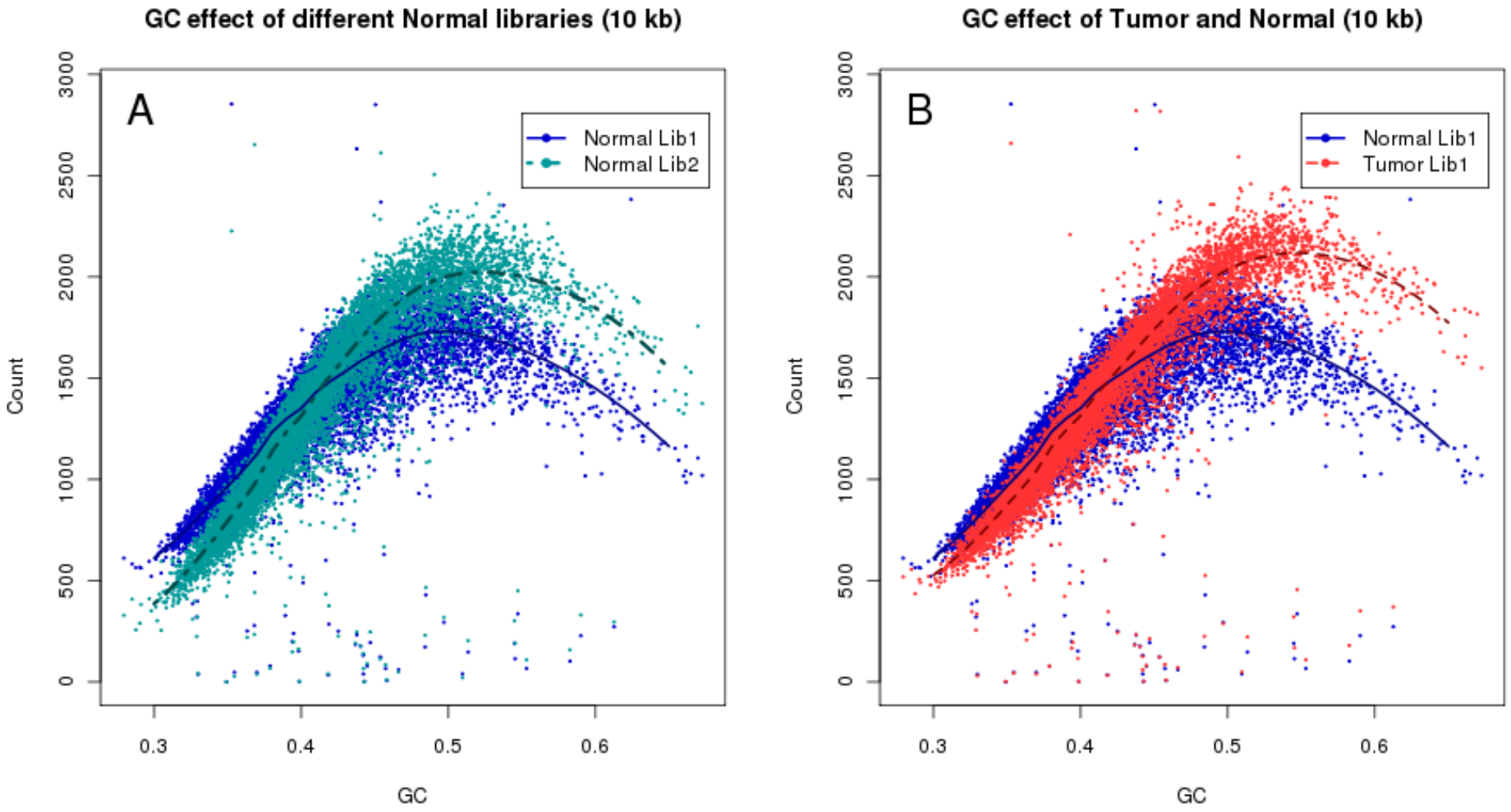
# GC loess curves for chromosomes 1-5, 10kb bins




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 |

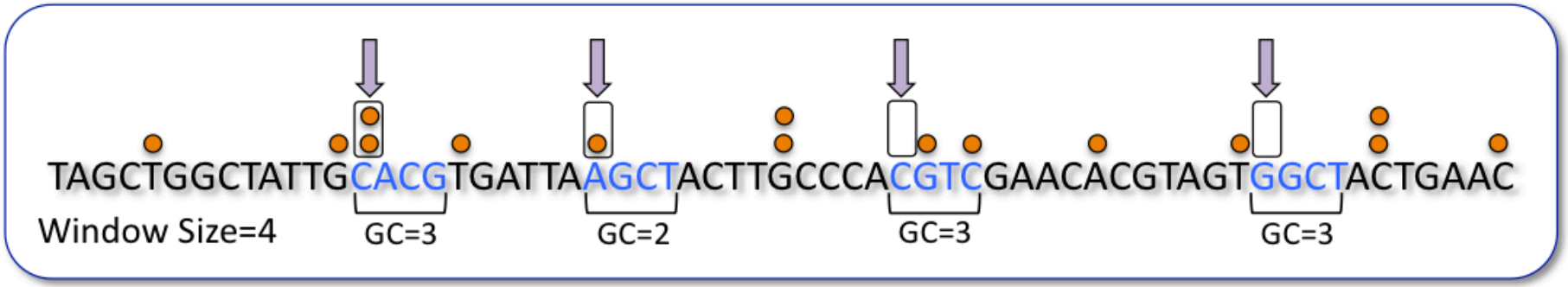
# Avoiding binning: single position analyses

**(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.

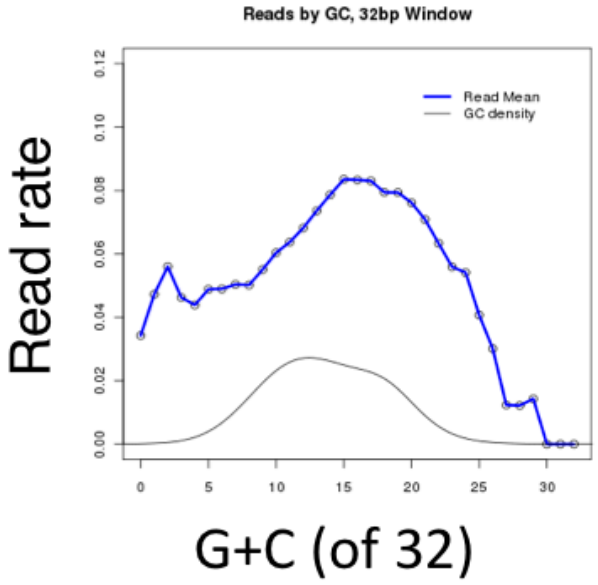
The fragment count at location *x* may depend on the *GC* content of the window $W_{a,l} = [x+a, x+a+l)$, which we will denote by $gc = GC(x+a,l)$.

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




| GC | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Locations | - | - | 1 | 3 | - |
| Reads | - | - | 1 | 2 | - |
| Rate | - | - | 1 | 0.66 | - |

D) Plot GC curve




$$Rate = \frac{\#\,reads}{\#\,locations}$$

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

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

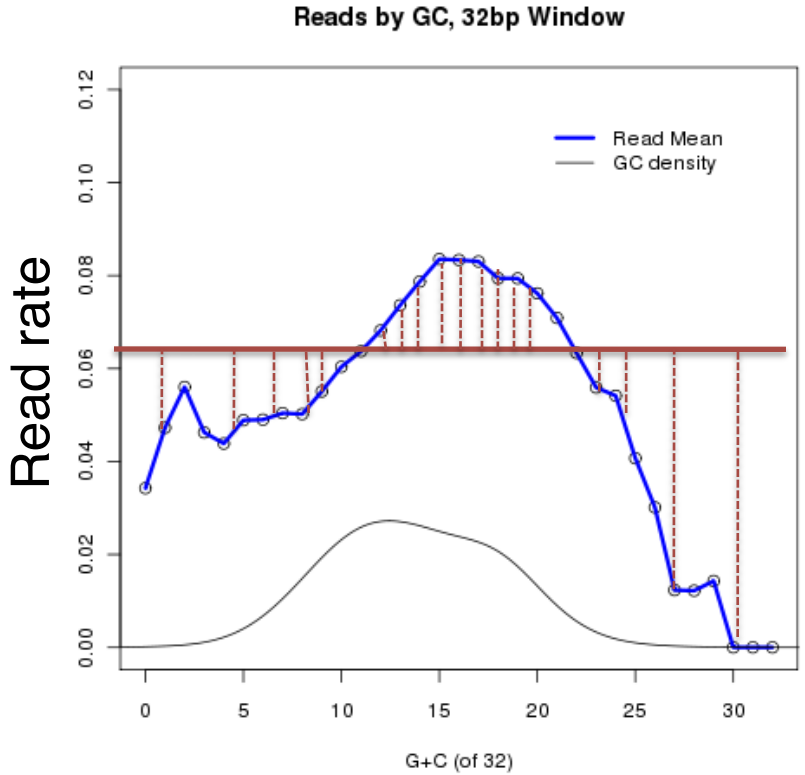
Superficially: their shape, that is, their deviation from flatness, which is GC-independence.

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.




*TV* distance = a weighted average of the brown lengths

## In symbols,

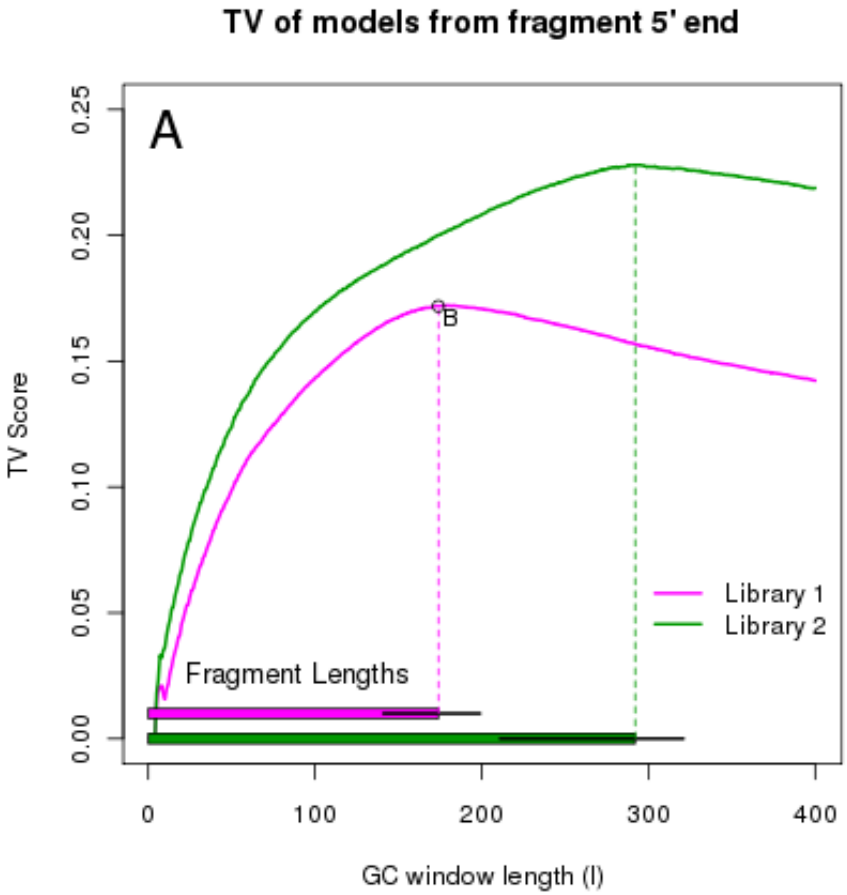
$$\hat{\lambda}_{gc} = \frac{F_{gc}}{N_{gc}}, \quad \hat{\lambda} = \frac{F}{n}$$

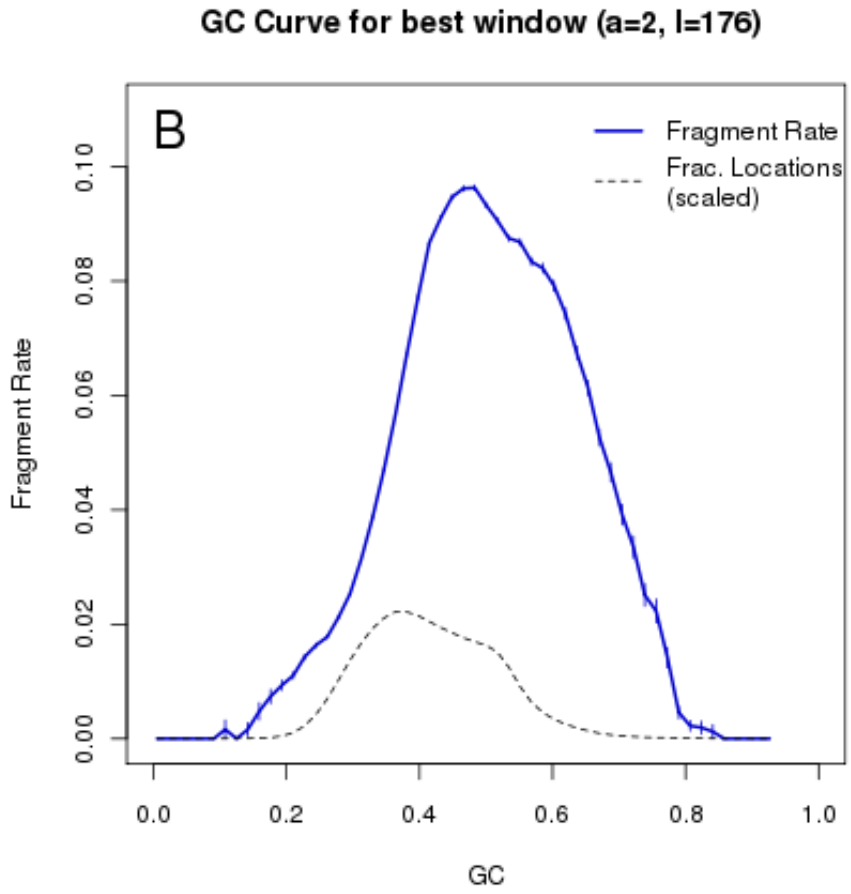
$$TV(W_{a,l}) = \frac{1}{2\hat{\lambda}} \sum_{gc=0}^{l} \frac{N_{gc}}{n} \mid \hat{\lambda}_{gc} - \hat{\lambda} \mid,$$

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

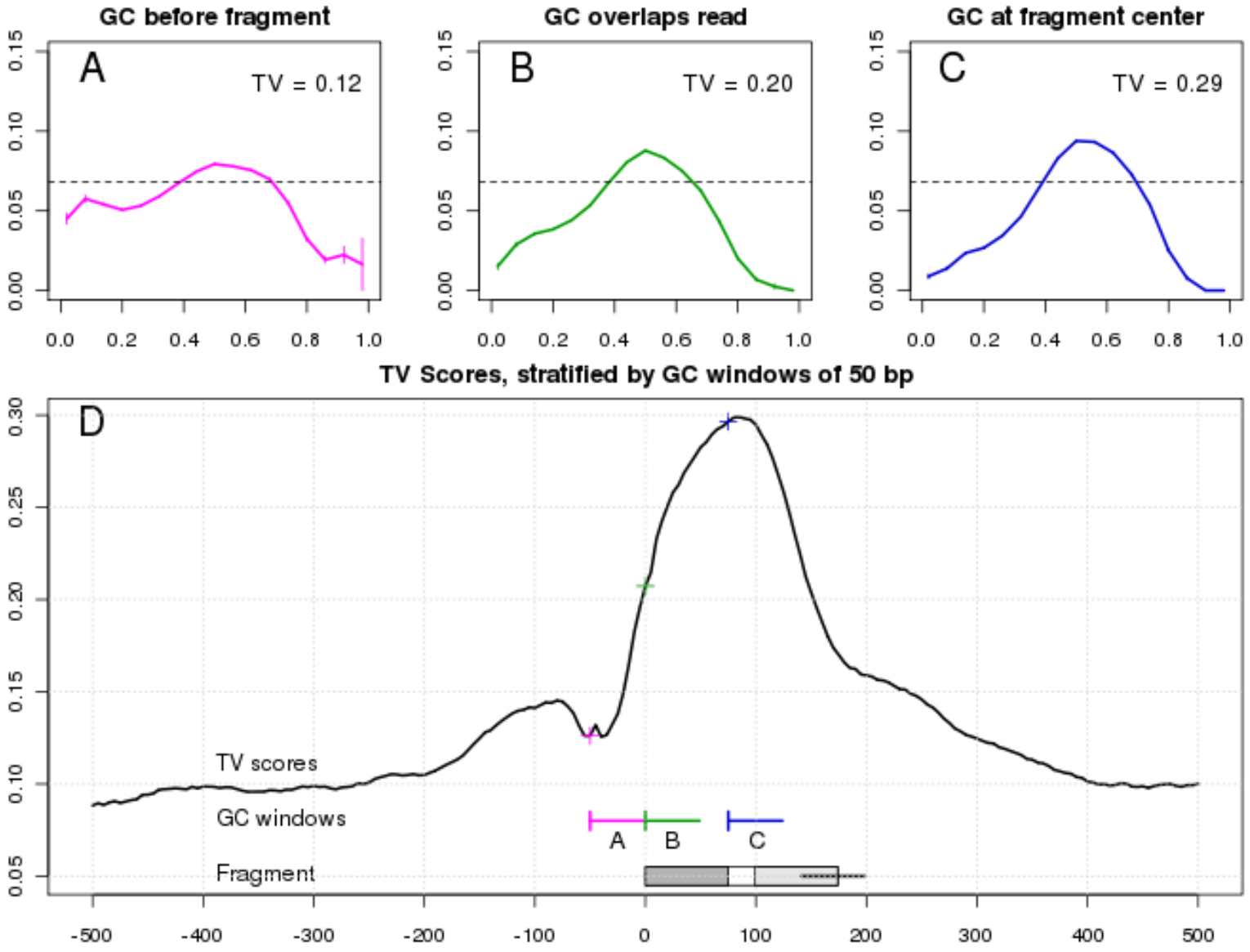
Next we look at some *TV* values. We can vary *a* and *l*, and we do so, separately here, for simplicity

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



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

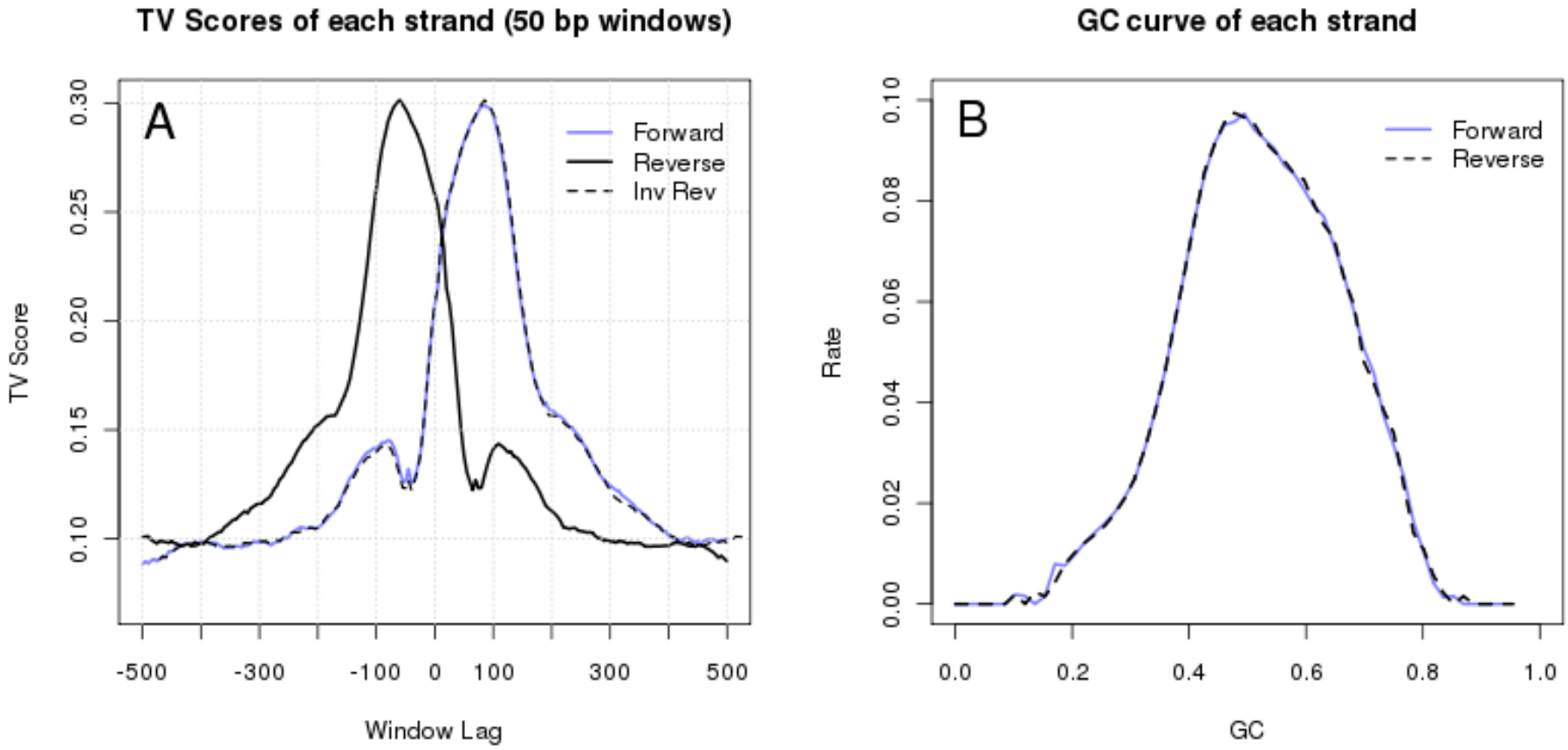


# Interim conclusion from many such plots

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.

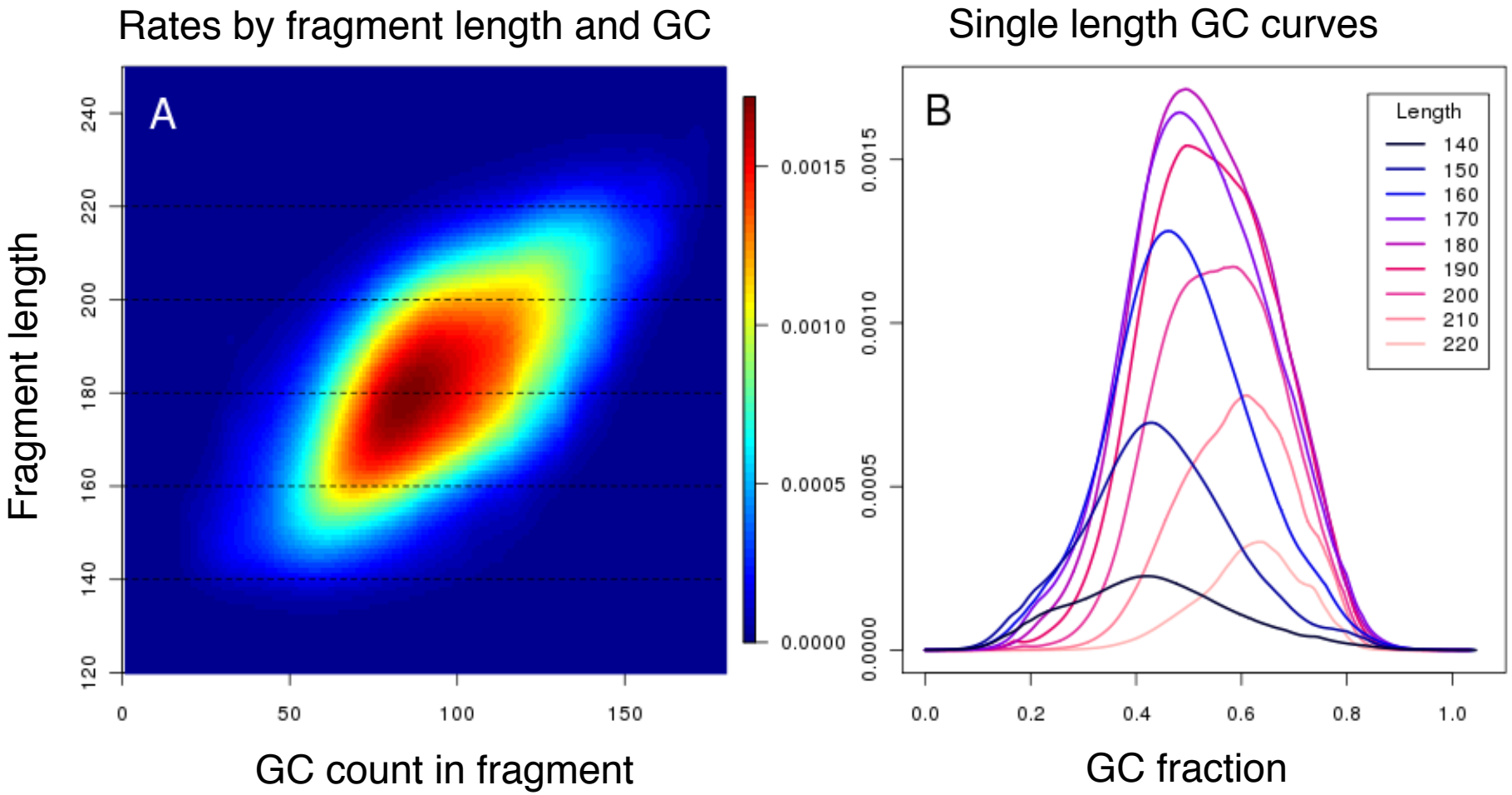
# Forward and reverse strands behave similarly



# Stratifying by fragment size *s*

$$\hat{\lambda}^{s}_{gc} = \frac{F^{s}_{gc}}{N^{s}_{gc}}$$

# 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

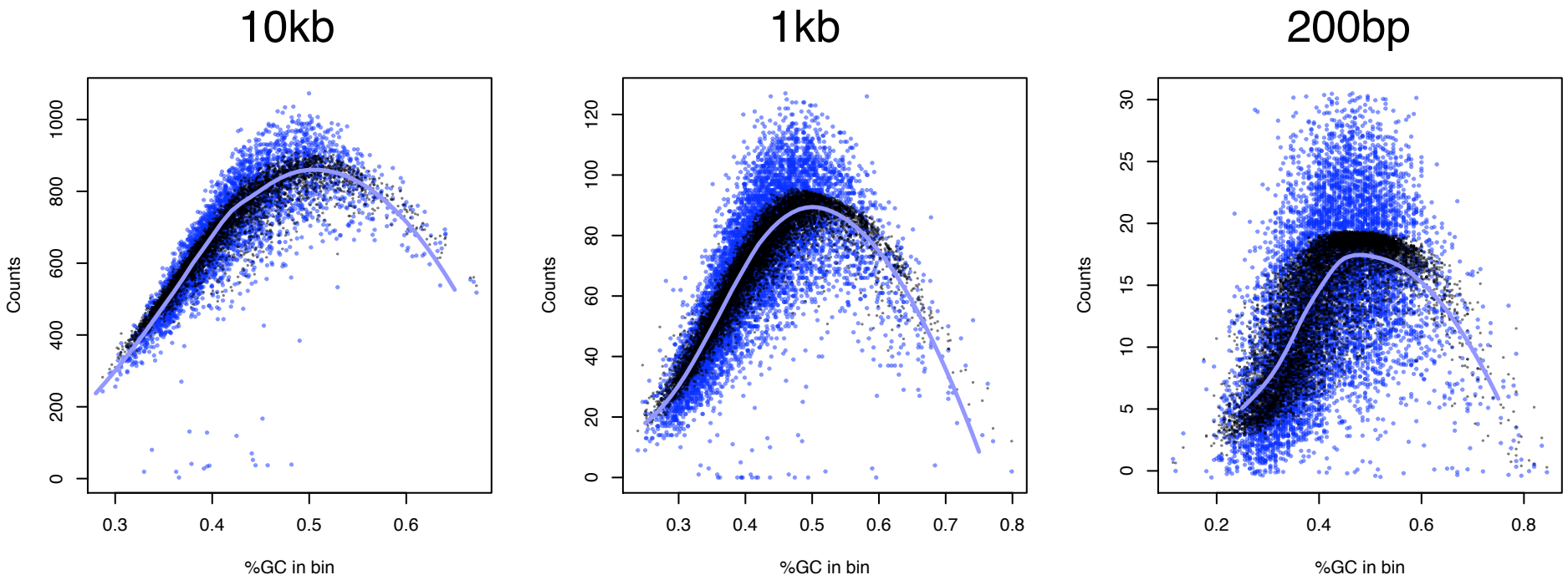
# Predicted rates at a given mappable position

$$\hat{\mu}_x = c \sum_s \hat{\lambda}^s_{GC(x+a,\, s-m)}$$

$$\hat{\mu}_B = \sum_{x \in B} \hat{\mu}_x$$

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



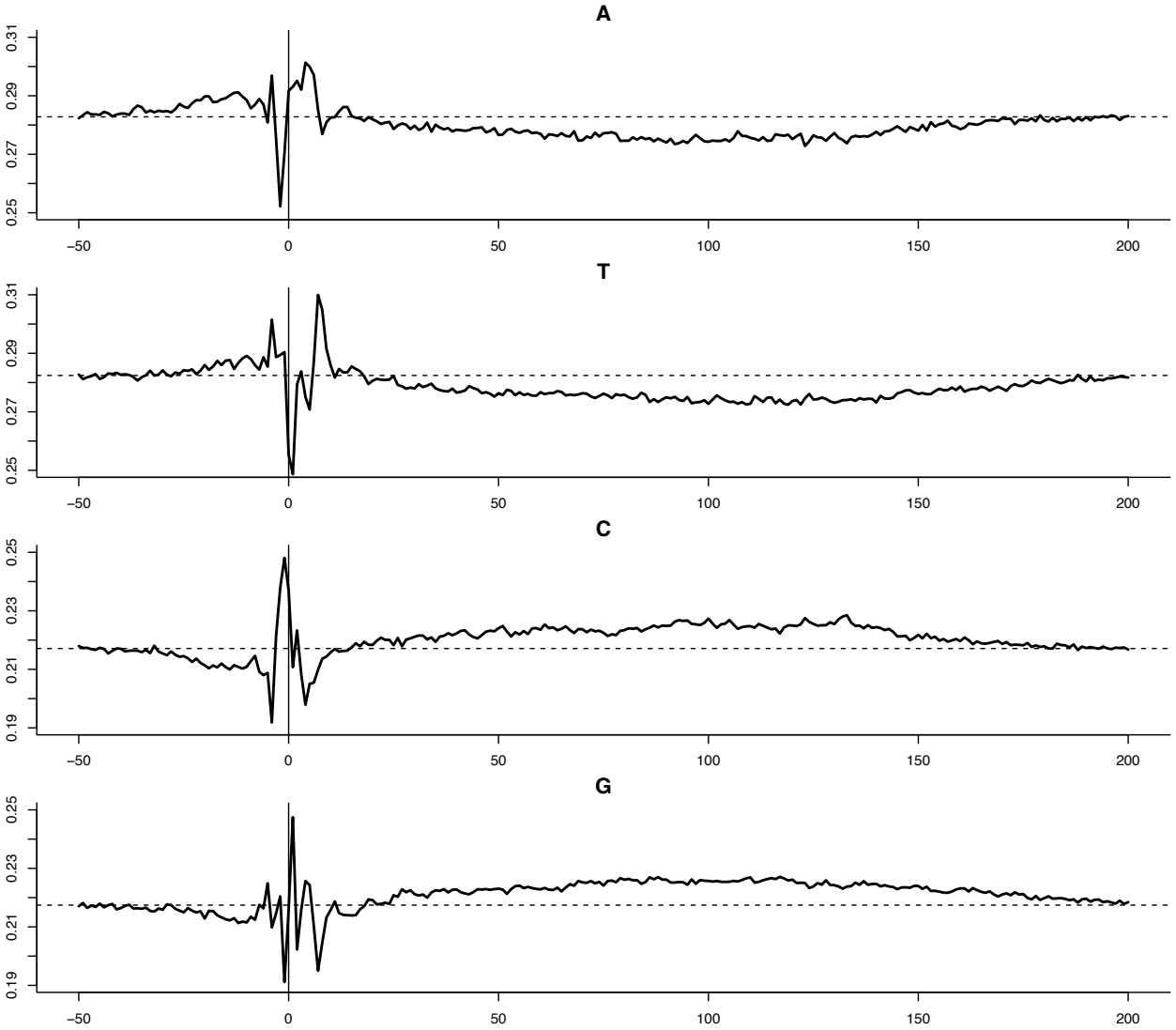

Lowess lines are based on the observed points.

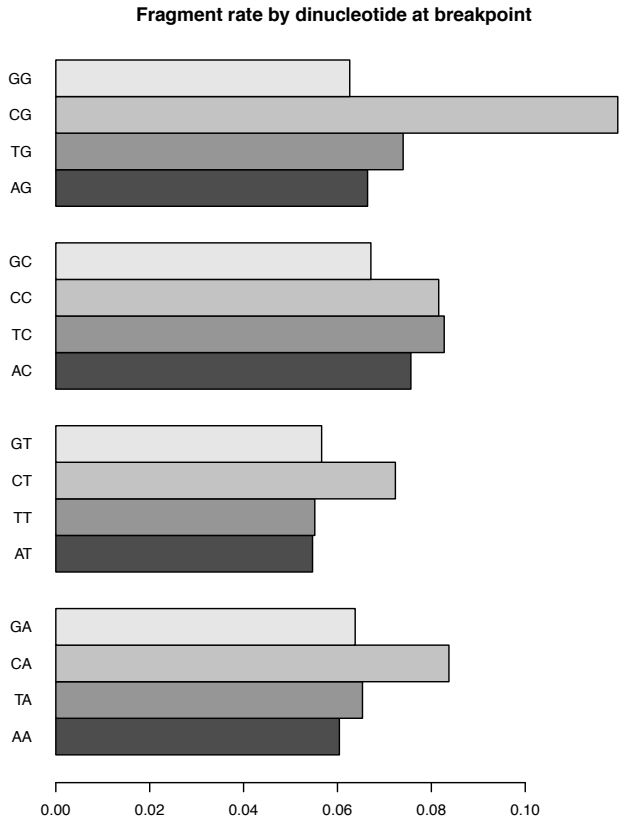
**Conclusion: the predictions seem to be working.**

# Some other biases/models

# Breakpoint effects

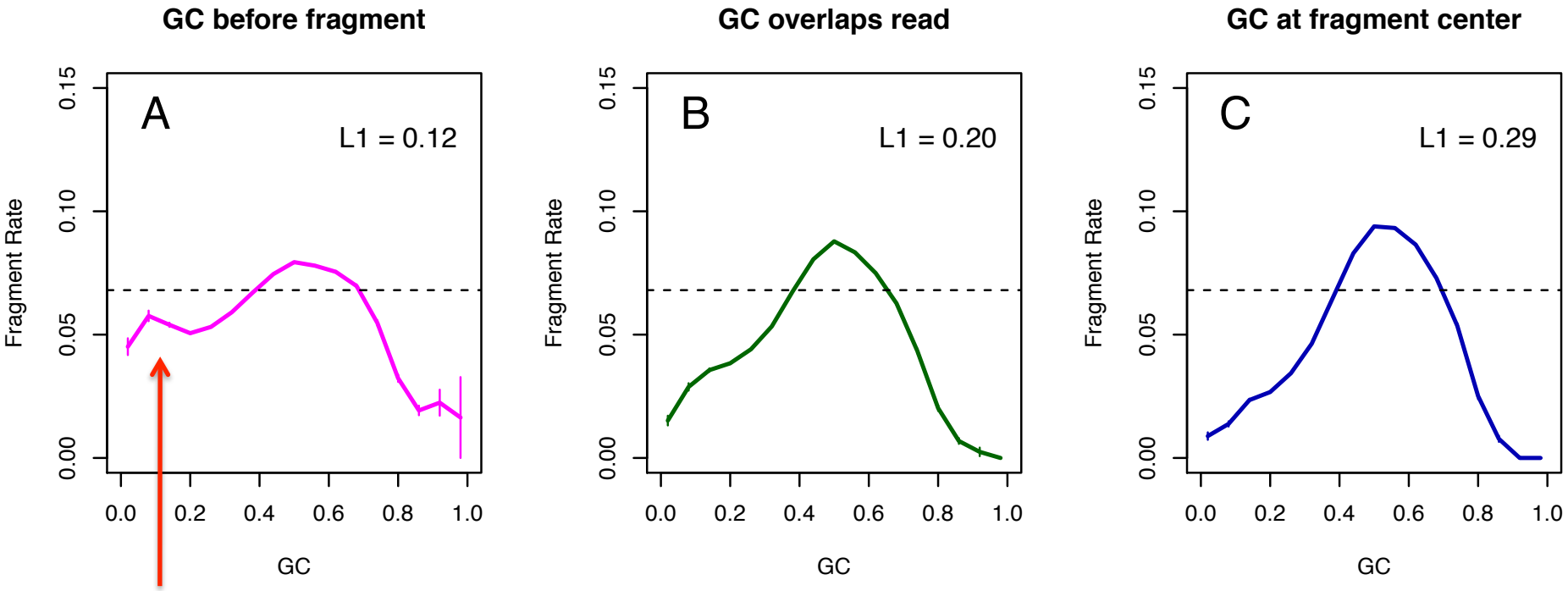
## (*cf* Hansen *et al*, Nucleic Acids Research, 2010)




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

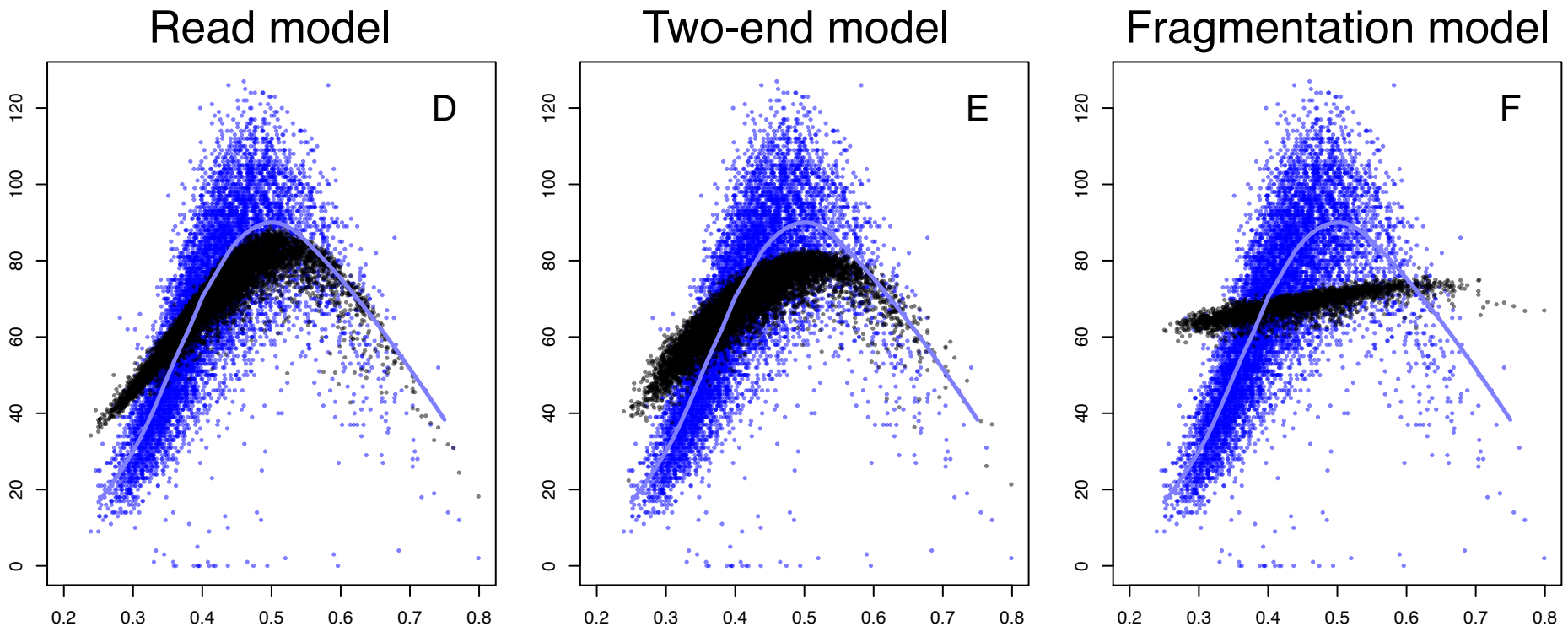
# End effects




Slight AT preference

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

# Some other predictions
## (all aggregated to 1kb bins)




Fraction GC in all cases

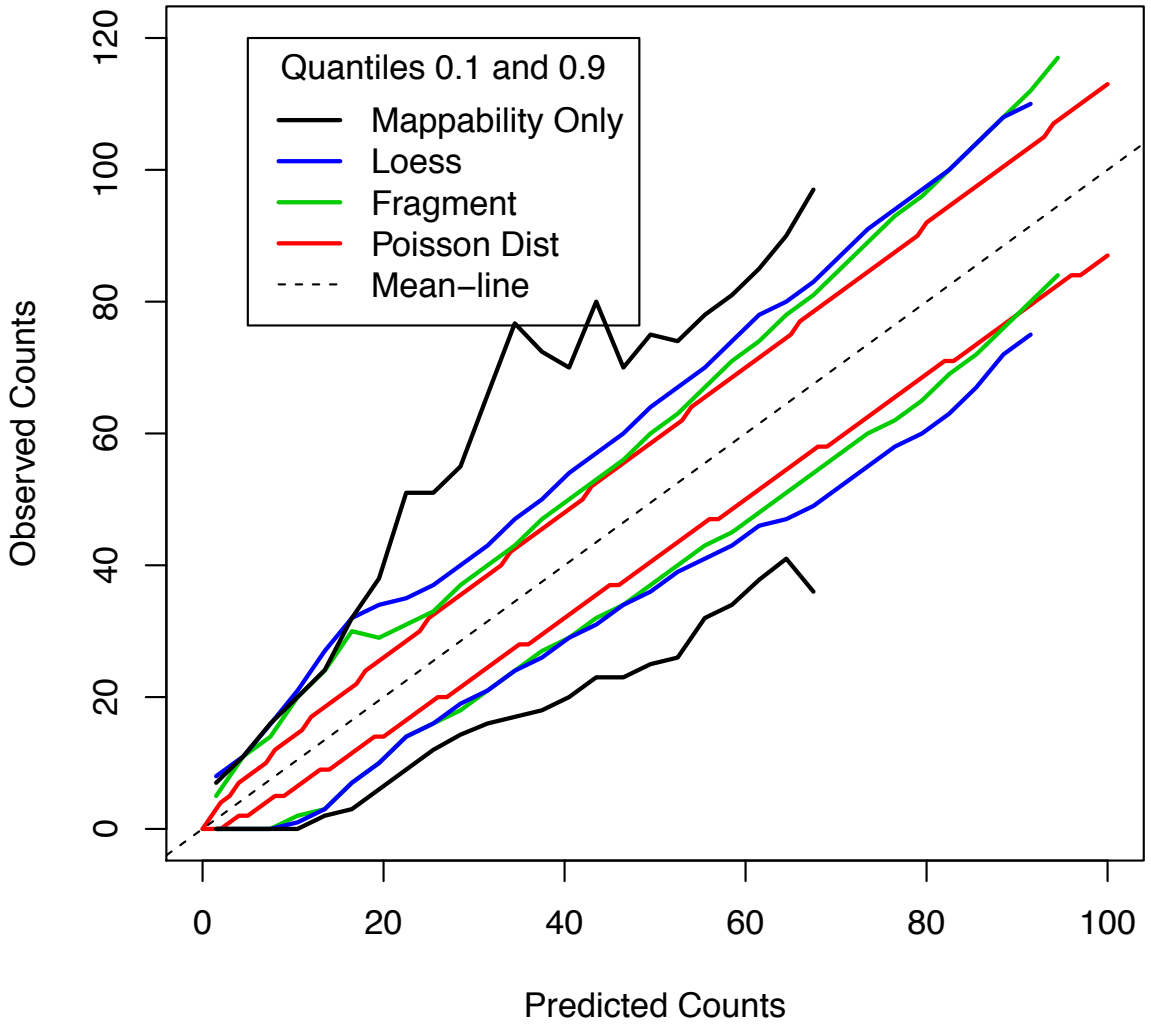


**Conclusion: These predictions don't work too well.**

# How well does our correction work?

**Theory**

# Spread of observed counts around predictions



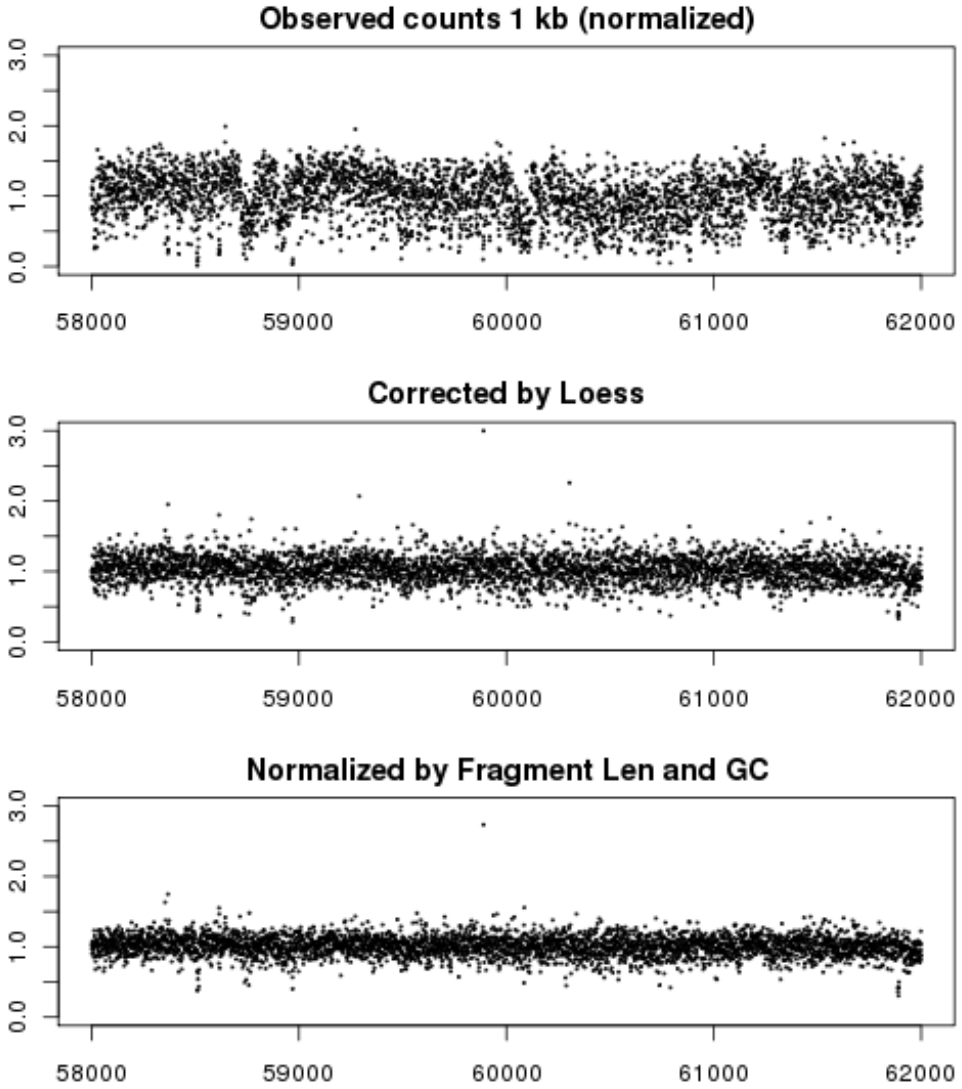

1kb bins

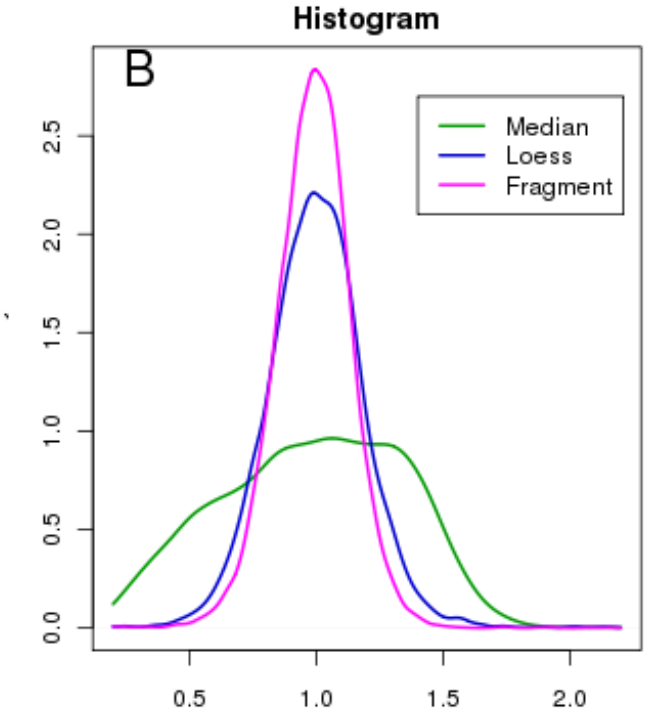
**Conclusion: we don't "explain" everything.**

# How well does our correction work?

**Practice**

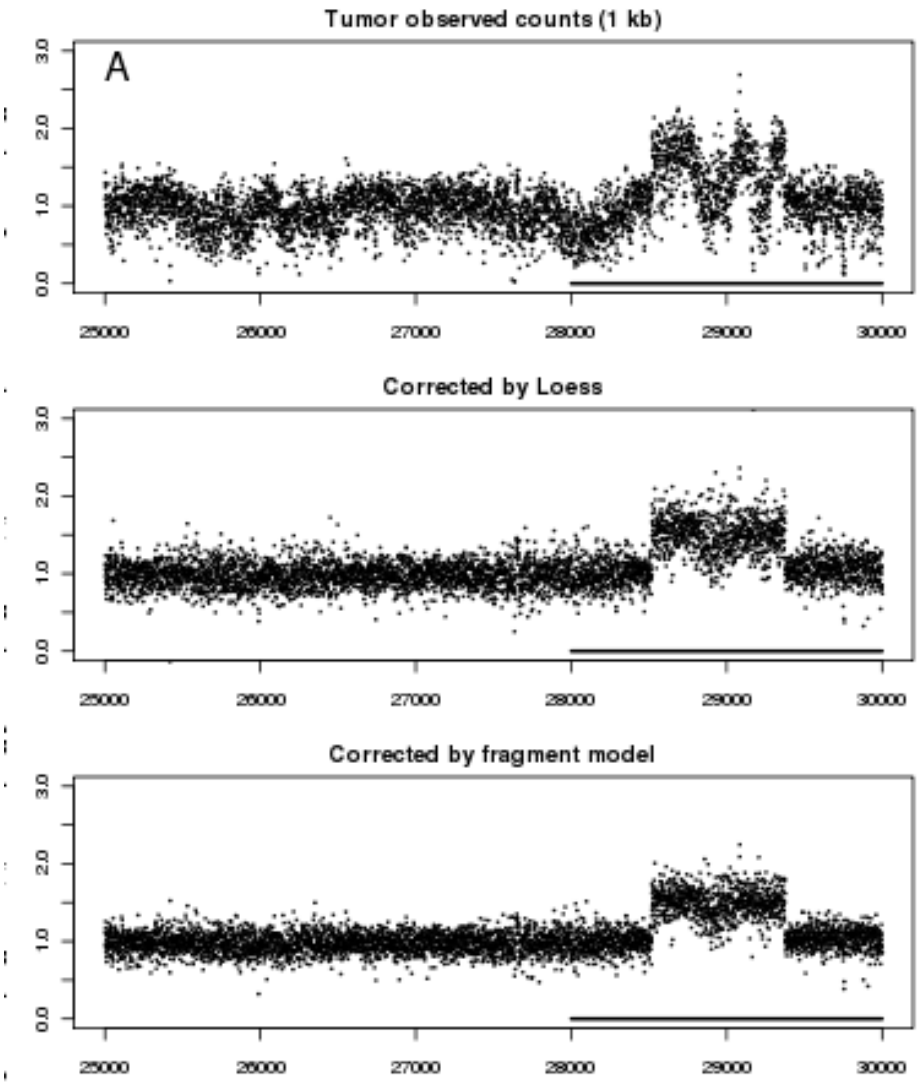
# Copy number: corrections to normal samples

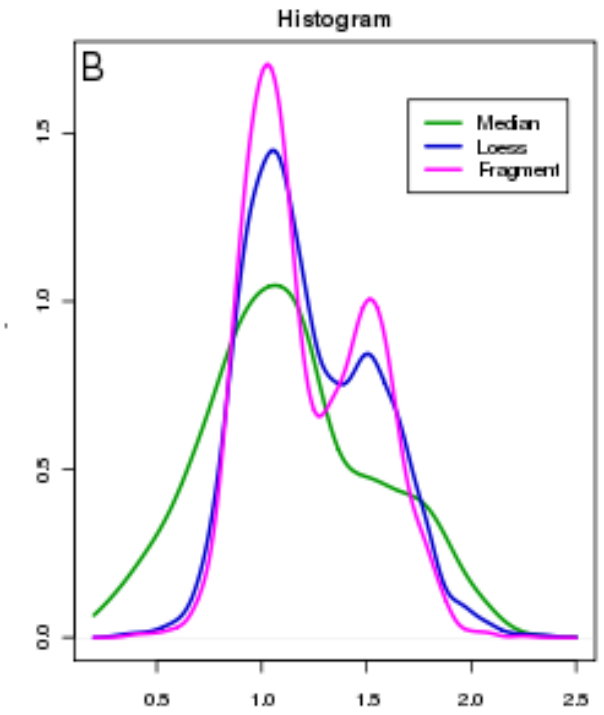



Slight improvement over loess.

# Copy number: crude corrections to tumor samples



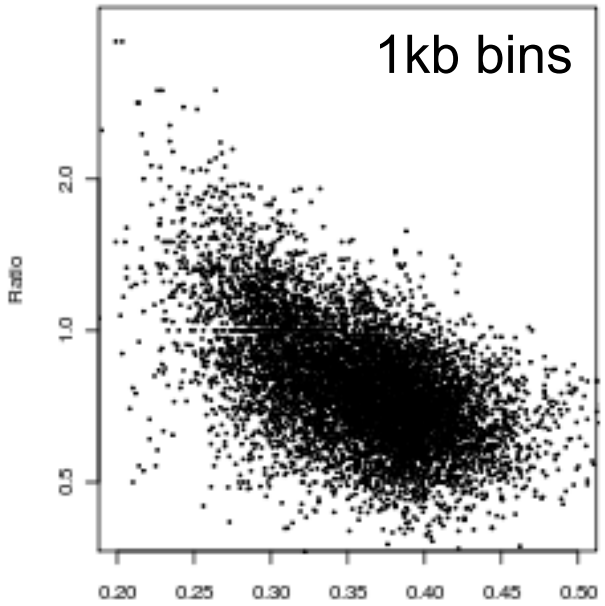




More work to be done here.

# ChIP-seq data (*A. thaliana*)

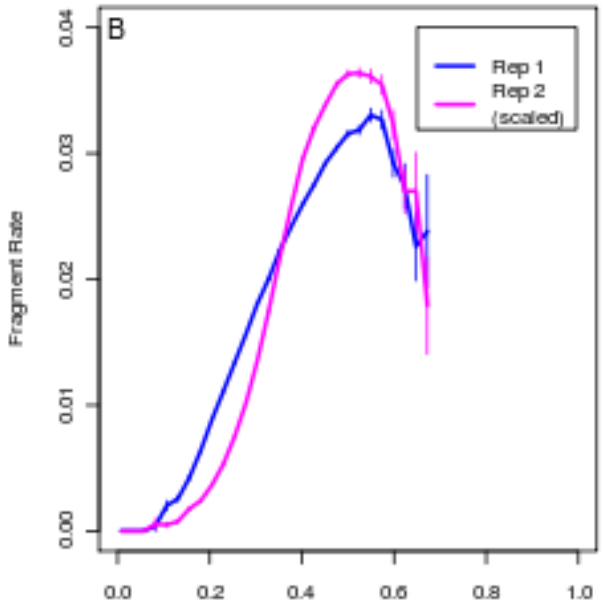
## Here two initially incompatible *technical* replicates

Uncorrected ratios



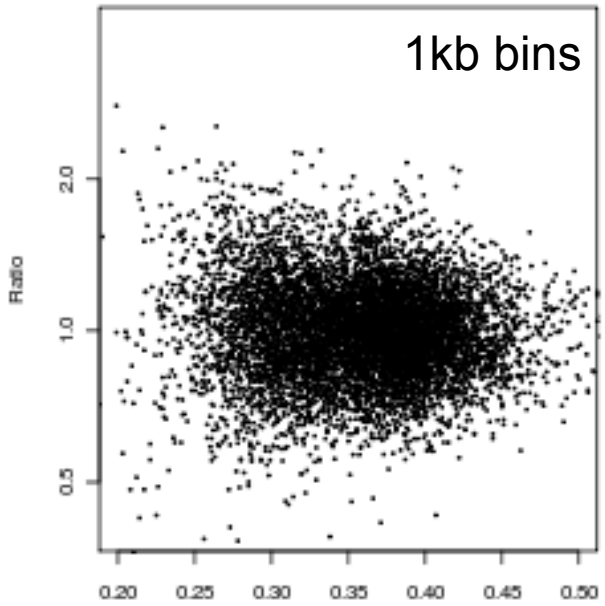

GC

GC curves (a=2, l=122)




GC

Corrected ratios

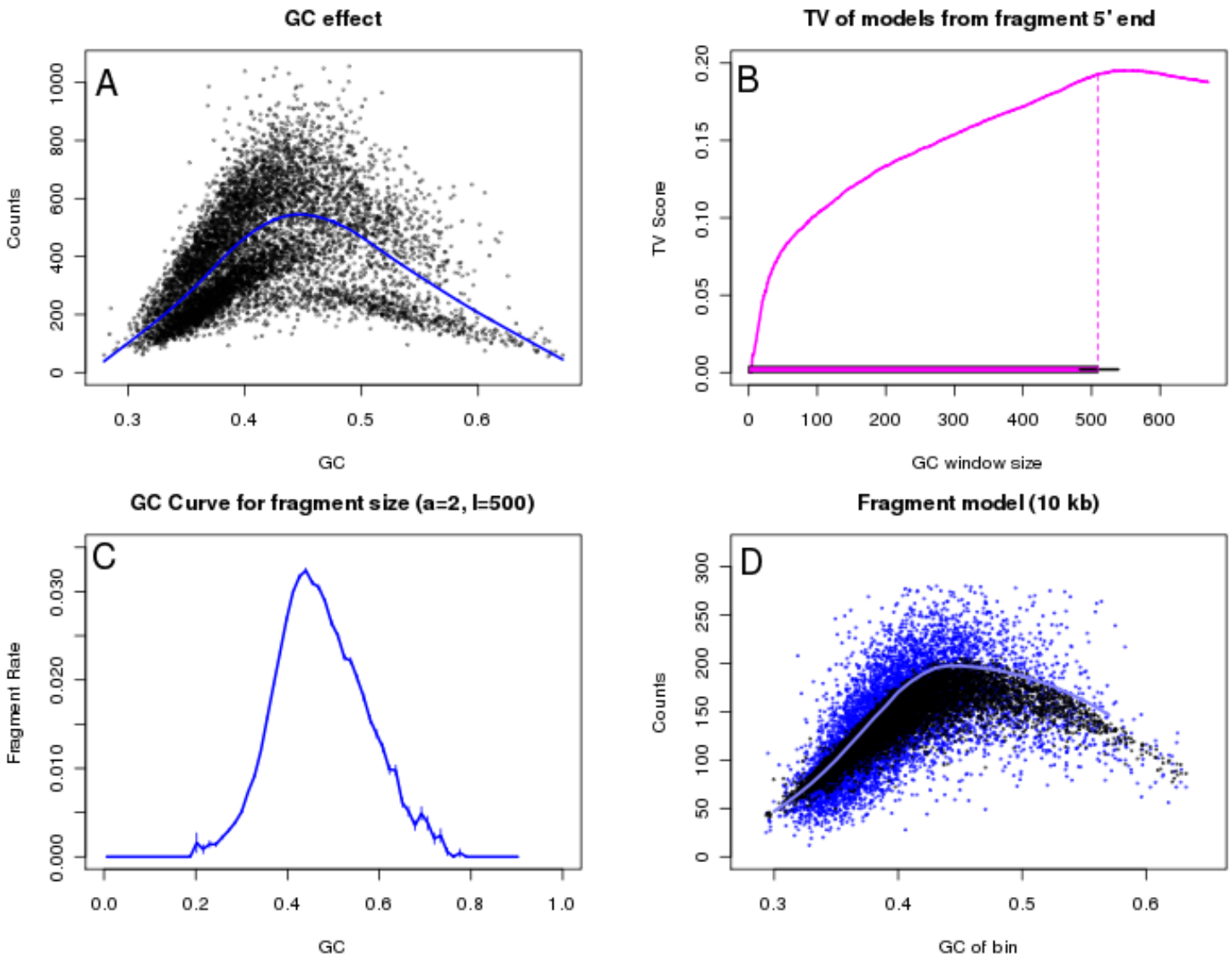



GC

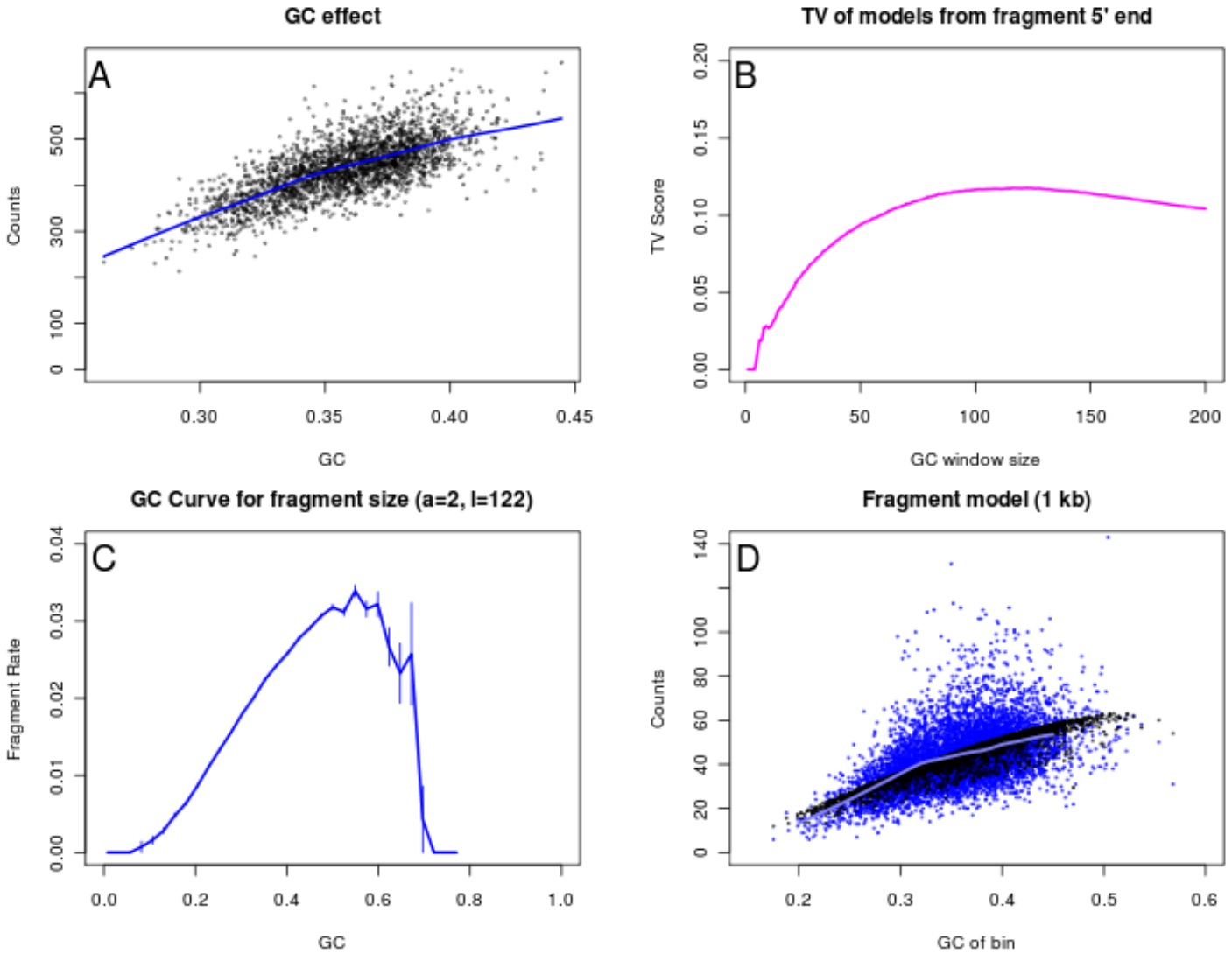
Problem mainly solved (*cf* Cheung *et al*, 2011)

# Other examples and phenomena (if time)

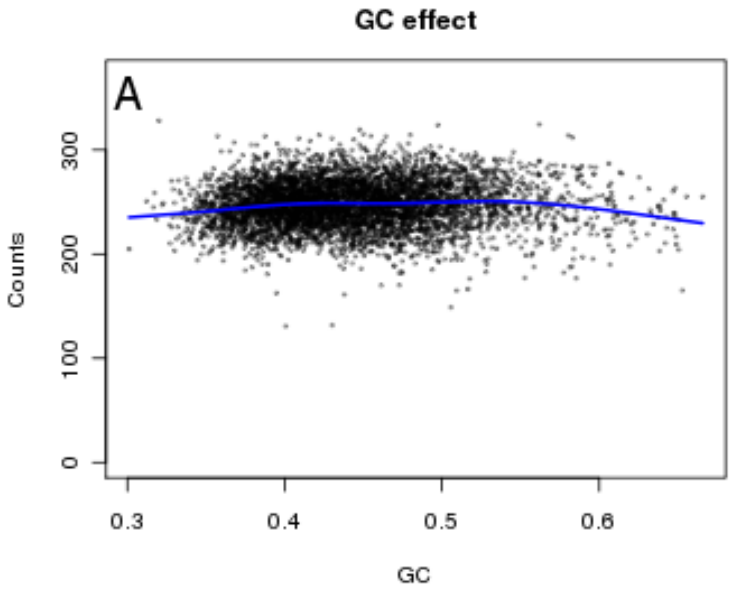
# Plots for a BrCa tumor

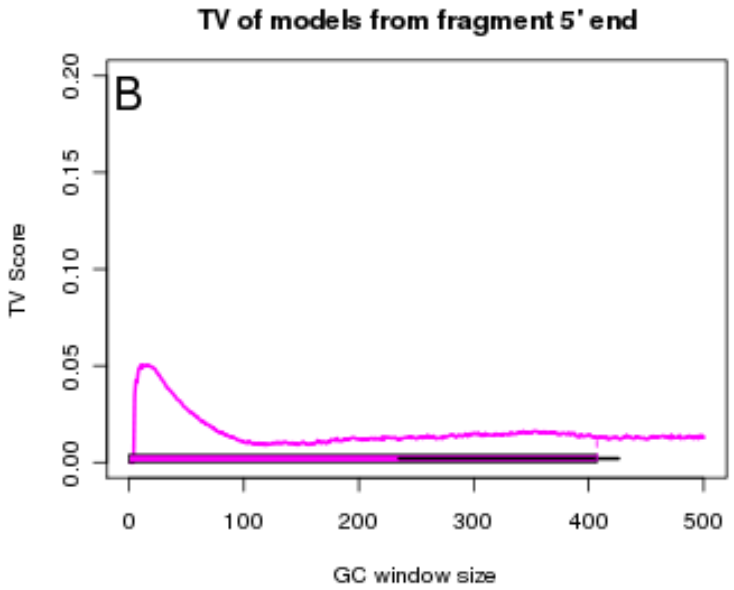


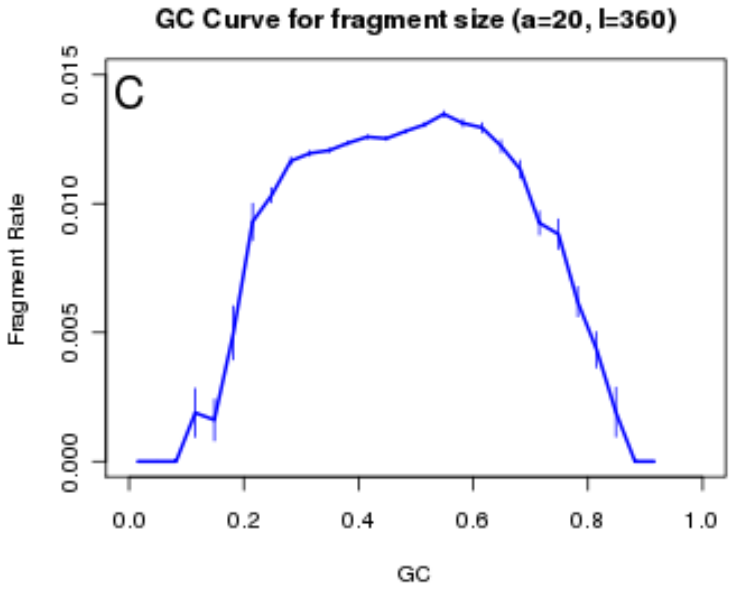
# Plots for ChIP-seq sample rep 1, *A. thaliana*

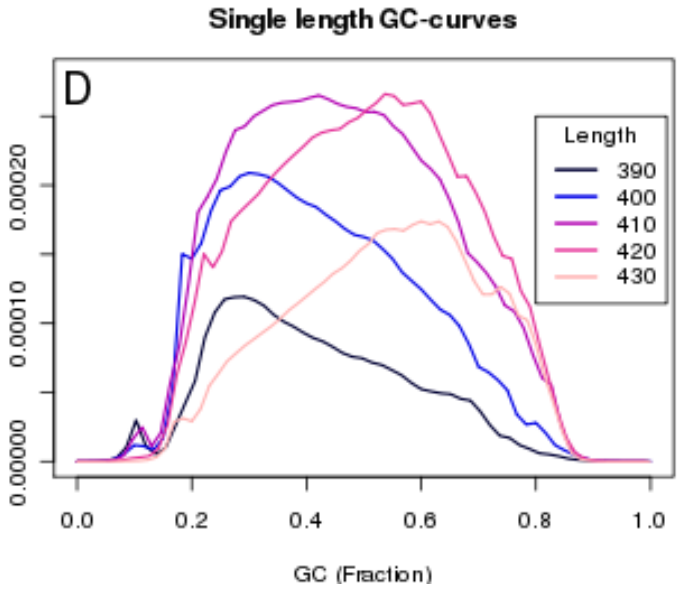


# Plots for one 1,000 genomes sample

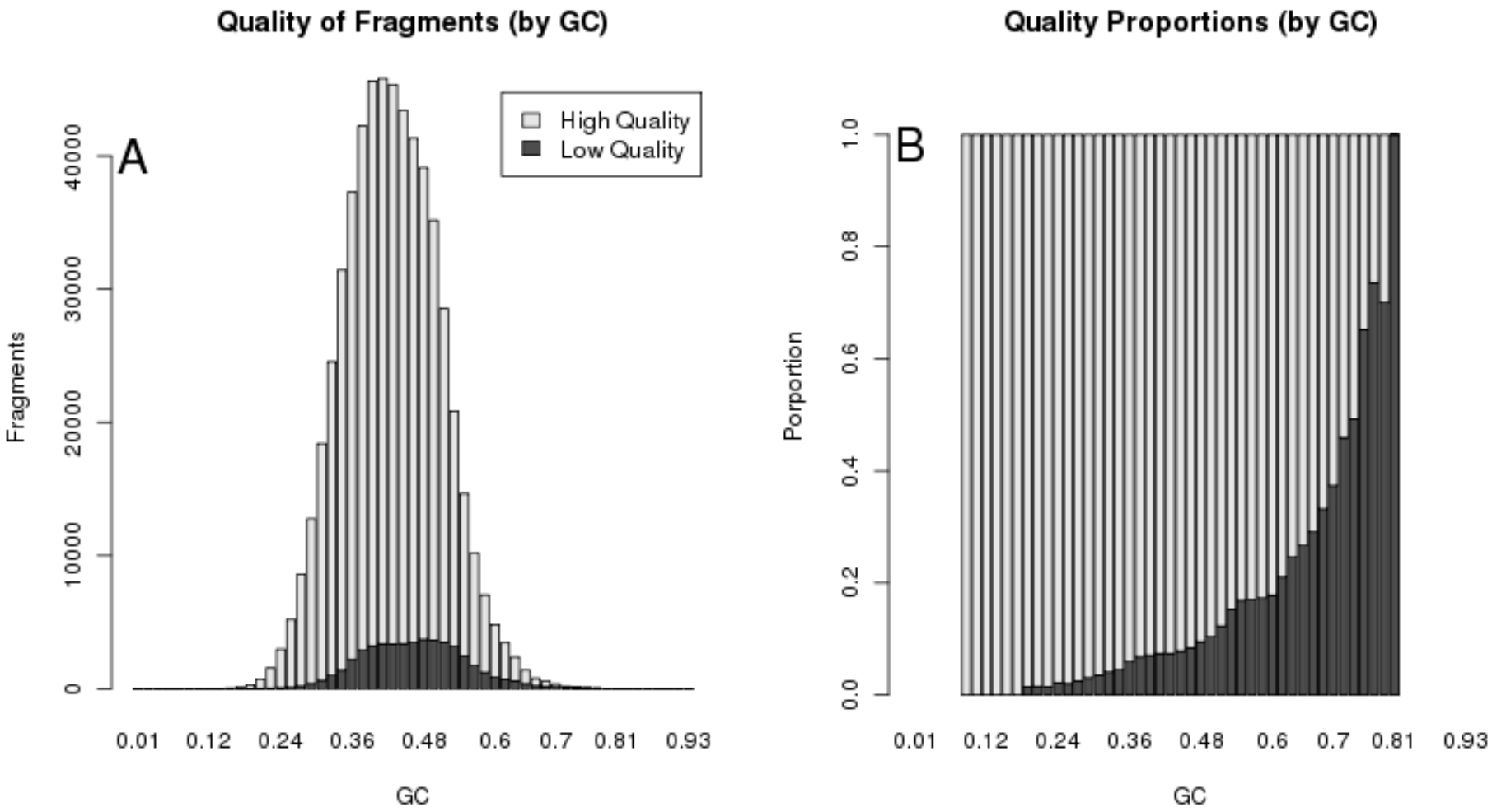


# Fragments partitioned by quality score



# 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…

Reference: http:/www.stat.berkeley.edu/25 Technical Report 804