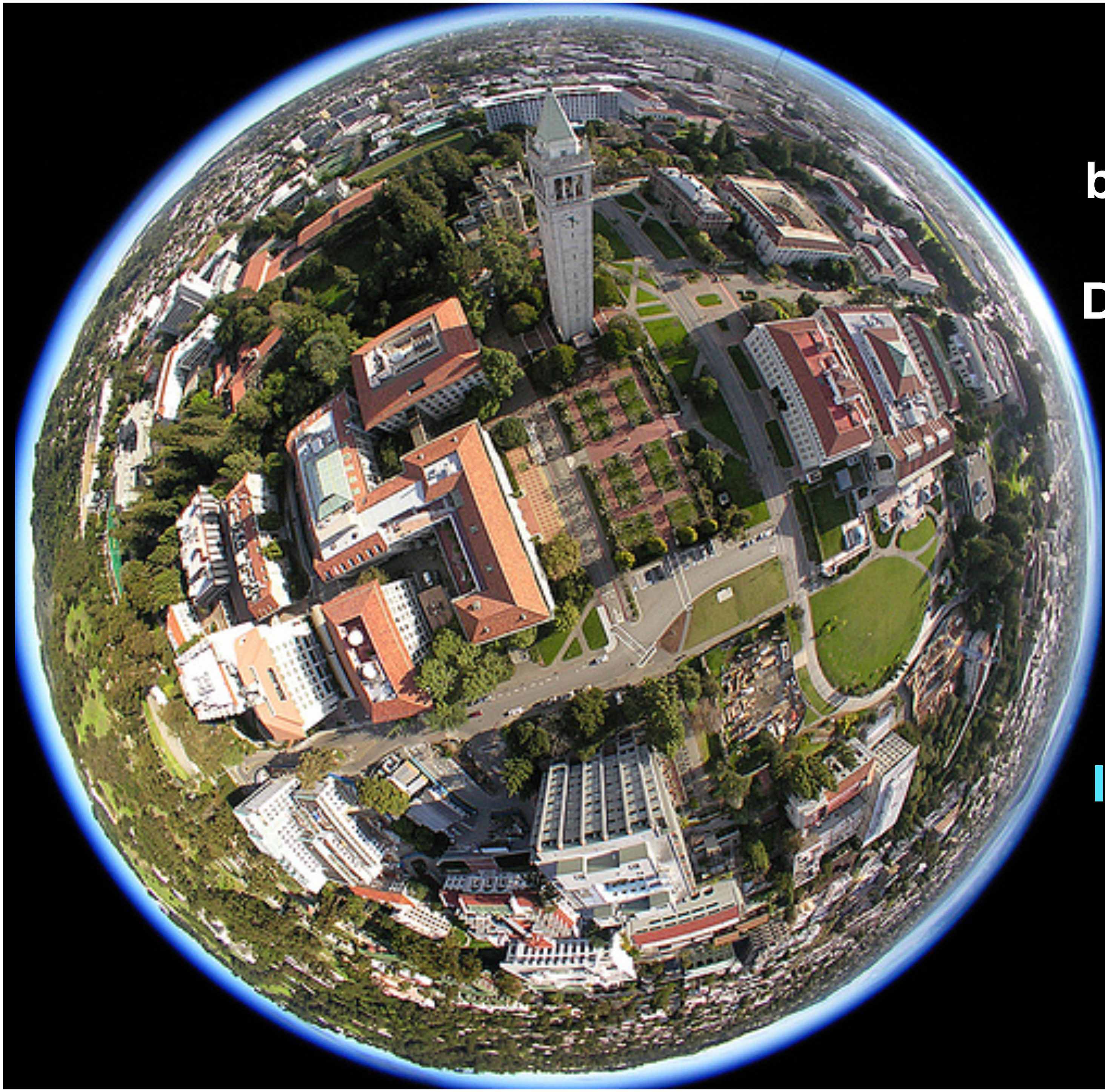


Dealing with GC-content bias in second generation DNA-sequence data

Terry Speed &

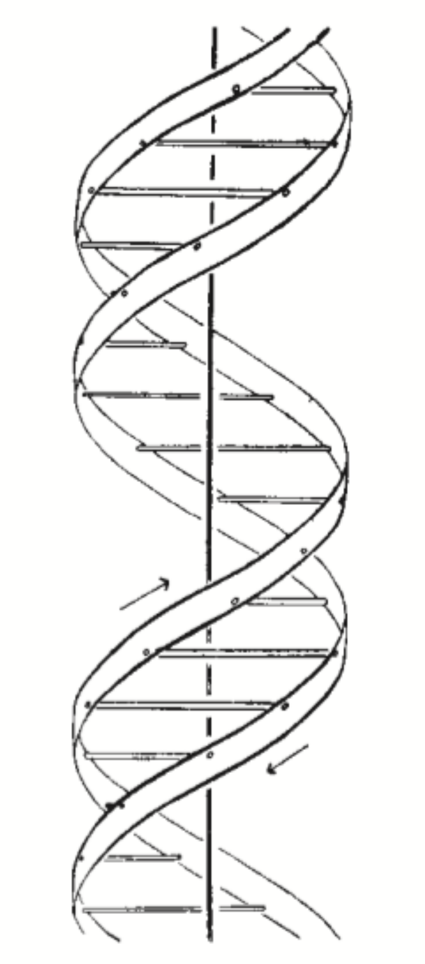
Yuval Benjamini

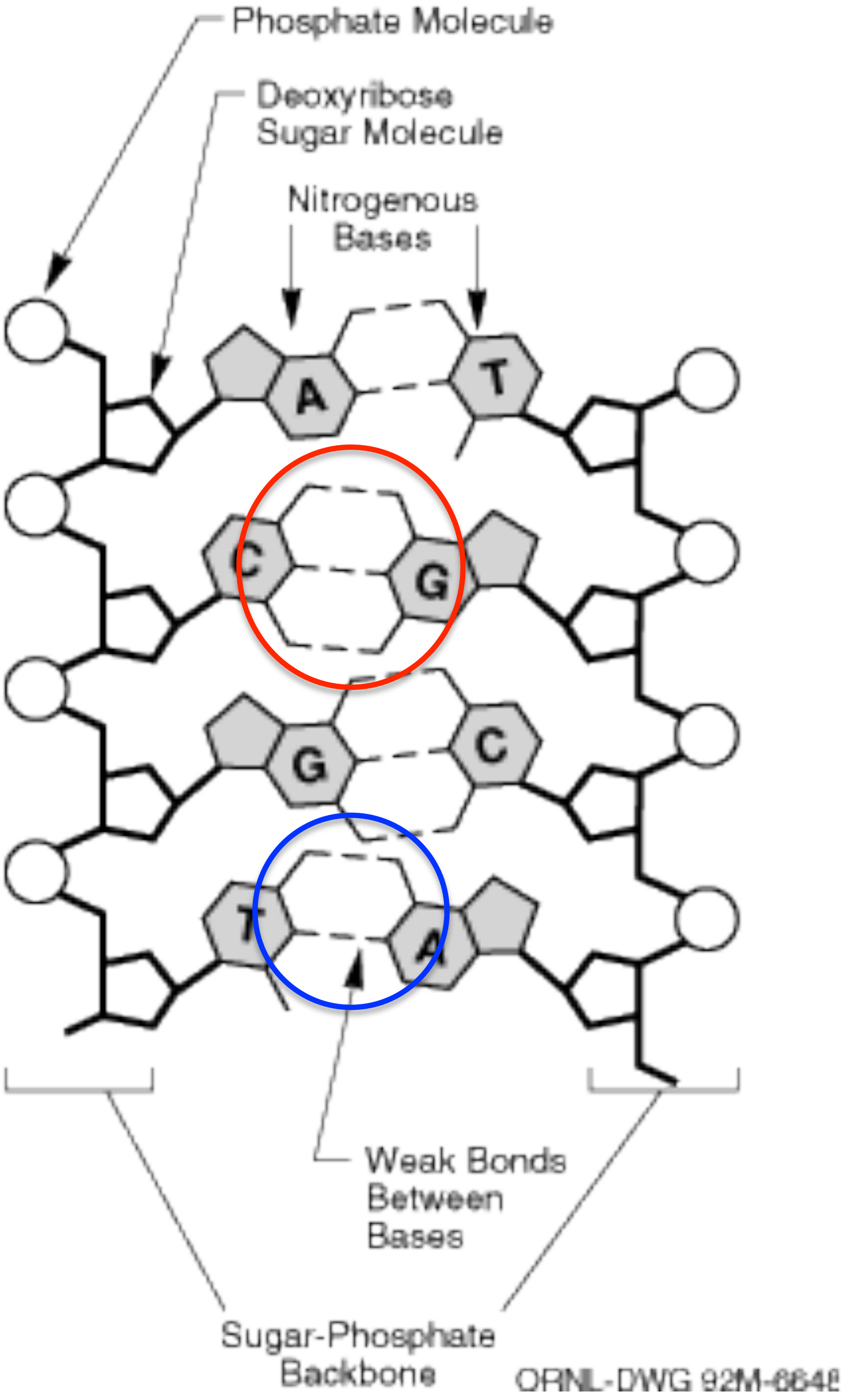
UC Berkeley

IPAM GenMini

11.01.2011

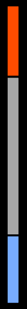
DNA and GC



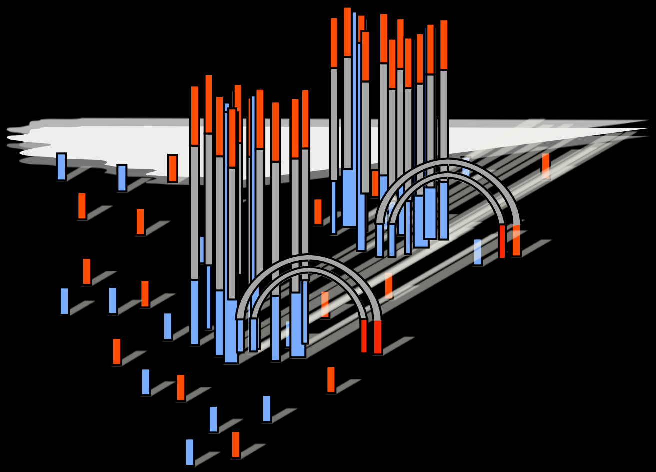

Phosphate Molecule
Deoxyribose Sugar Molecule
Nitrogenous Bases
A
T
C
G
G
C
T
A
Weak Bonds Between Bases
Sugar-Phosphate Backbone
ORNL-DWG 92M-6648

Illumina Sequencing Technology

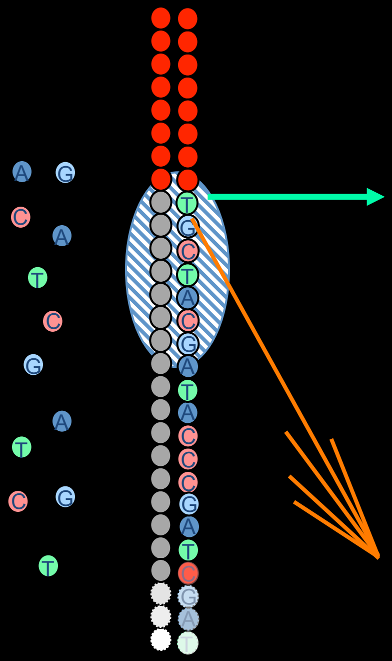
DNA
(0.1-1.0 ug)



Library preparation: fragmentation,
end repair, A-tailing, adaptor ligation,
size selection (melting) and PCR



Cluster growth



Sequencing

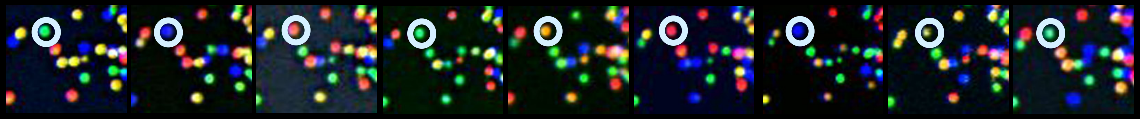


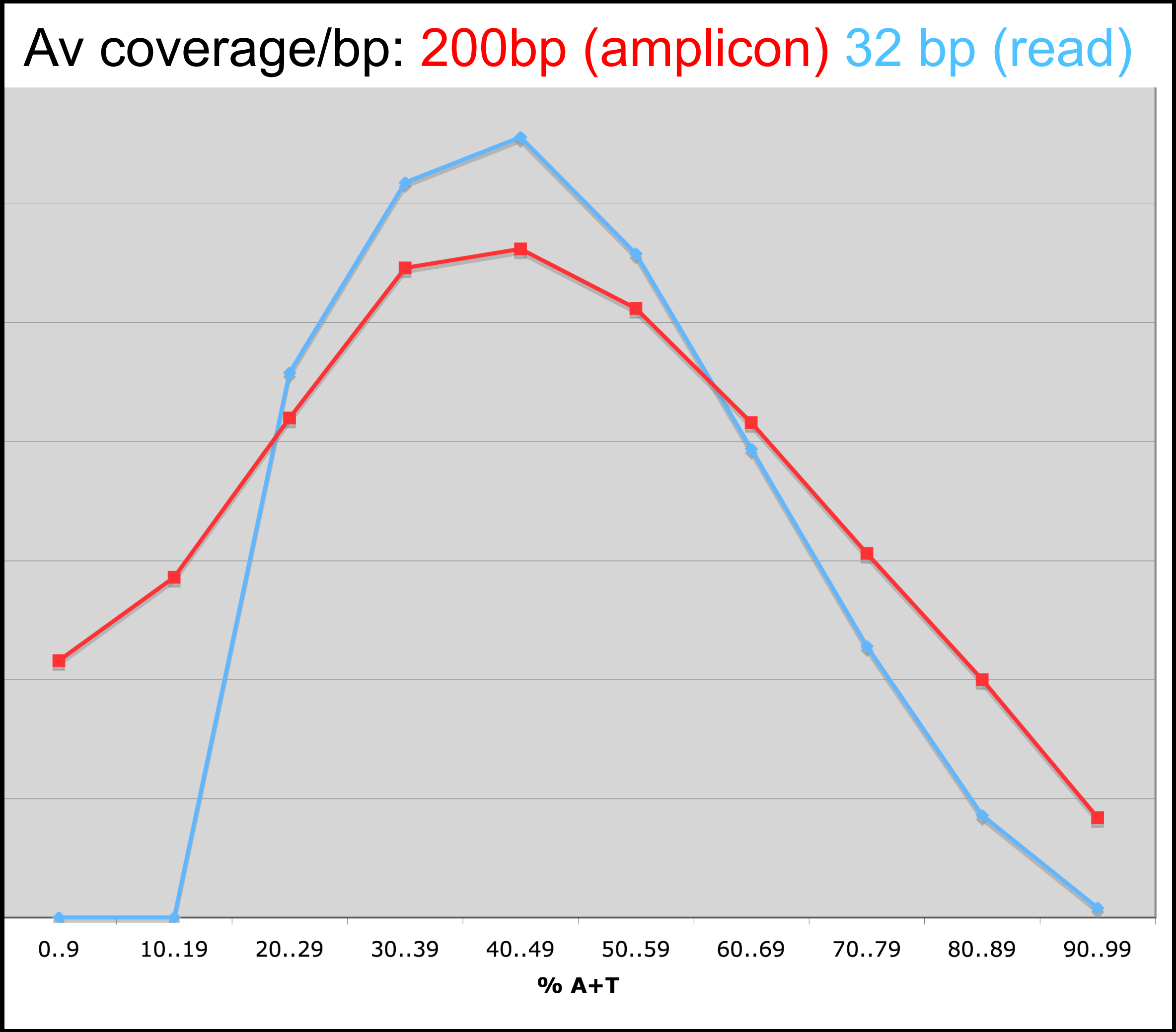
Image acquisition

T G C T A C G A T ...

Adapted from a slide of G Schroth, Illumina

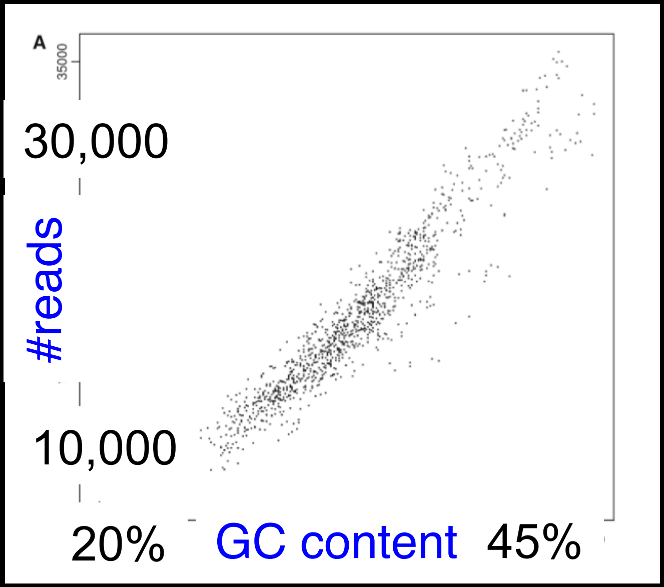
GC matters: Hiller *et al*, Nat Meth 2008

C. elegans Illumina (then Solexa) data

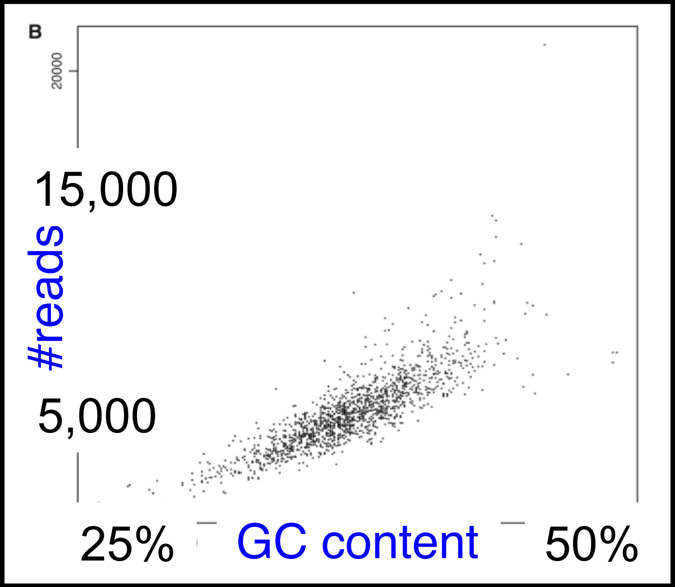


GC bias: Dohm *et al* NAR 2008

1kb bins



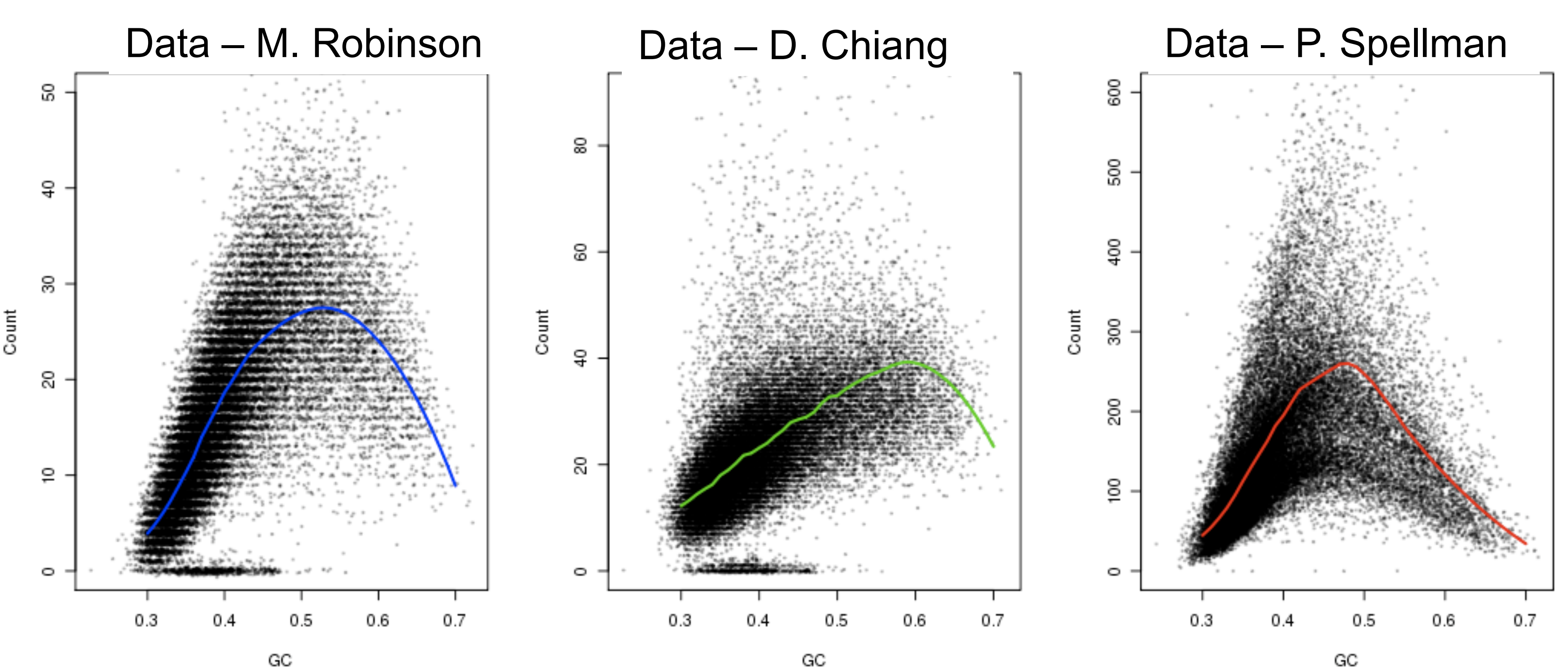
Beta vulgaris



Helicobacter acinonychis

The GC bias is non-linear in human data

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



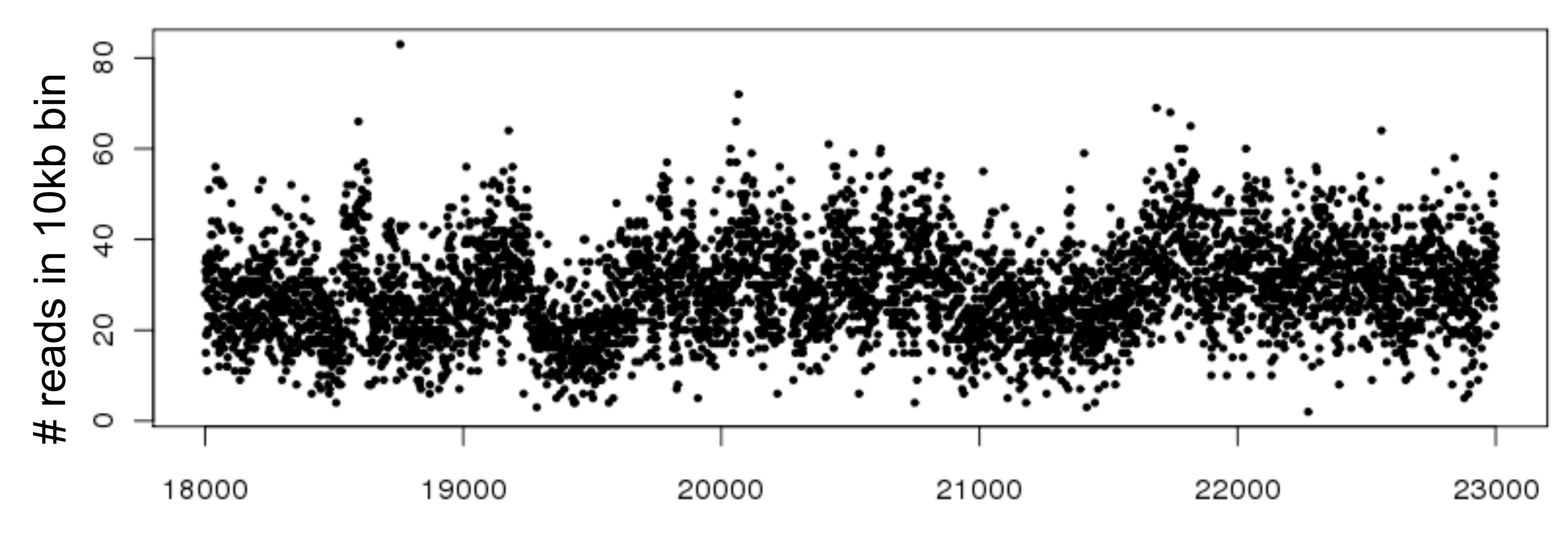
Data – M. Robinson

Data – D. Chiang

Data – P. Spellman

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

Another view: part of a human chr 2



Position of 10 kb bin on forward strand of q-arm of chr 2

The ups and downs reflect changes in GC content (trust me: data not shown, but see later)

Let’s begin with the question: whose GC?

- isochores (>300 kb in size, see slide 52)
- the local region around the read (how much?)
- the fragment itself
- the read itself
- near the read ends (how much?)
- the fragment breakpoints
- ….
- *none*, *some* or *all* of the above
- Your views?

Our main data for today

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

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

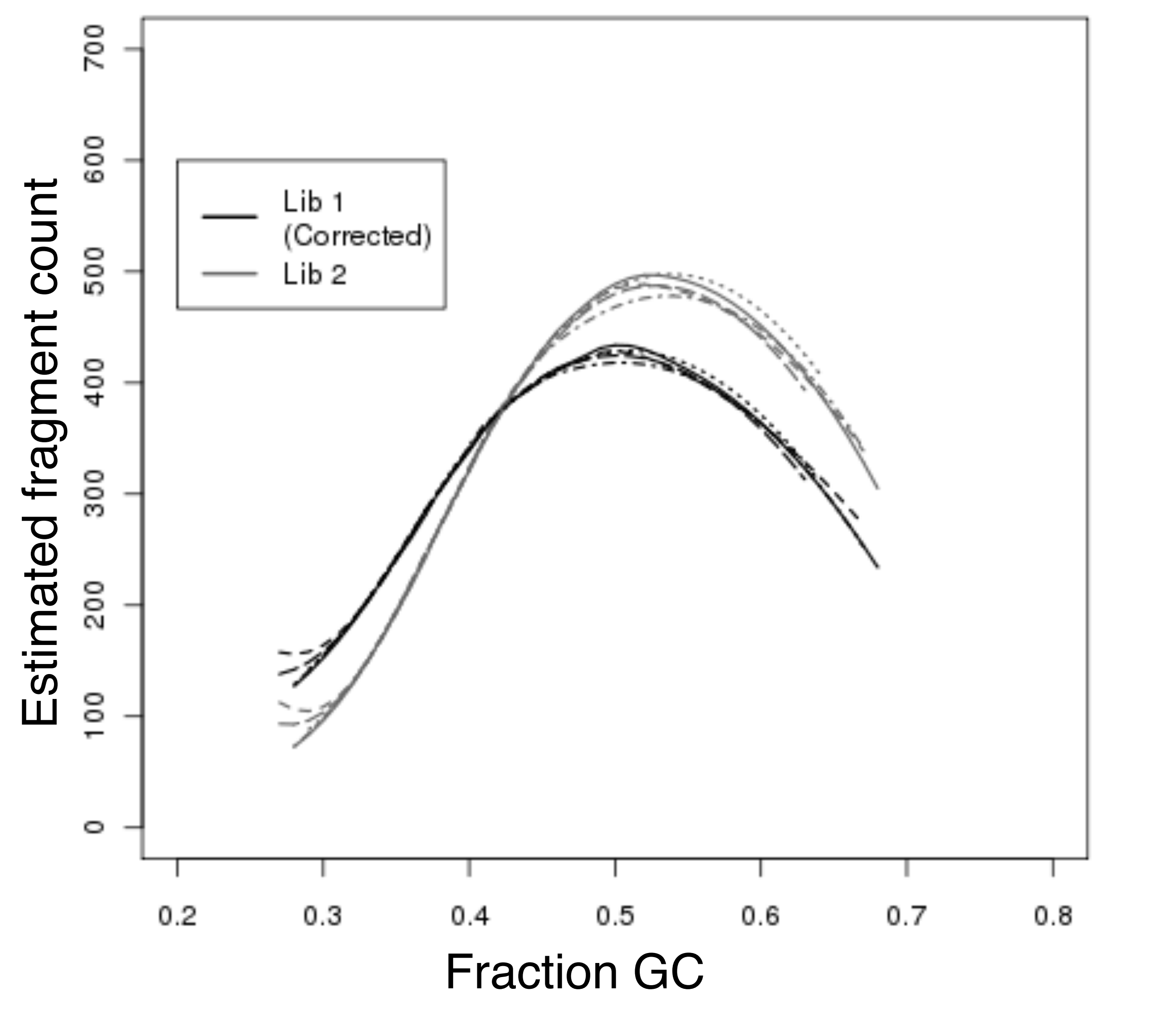
Fragments were sequenced to 75bp at both ends using the standard Illumina procedure.

Each sequenced read pair was mapped back to the human reference genome using bwa (version0.4.9

Most of the time we present results for just one chromosome, for

it doesn’t matter

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

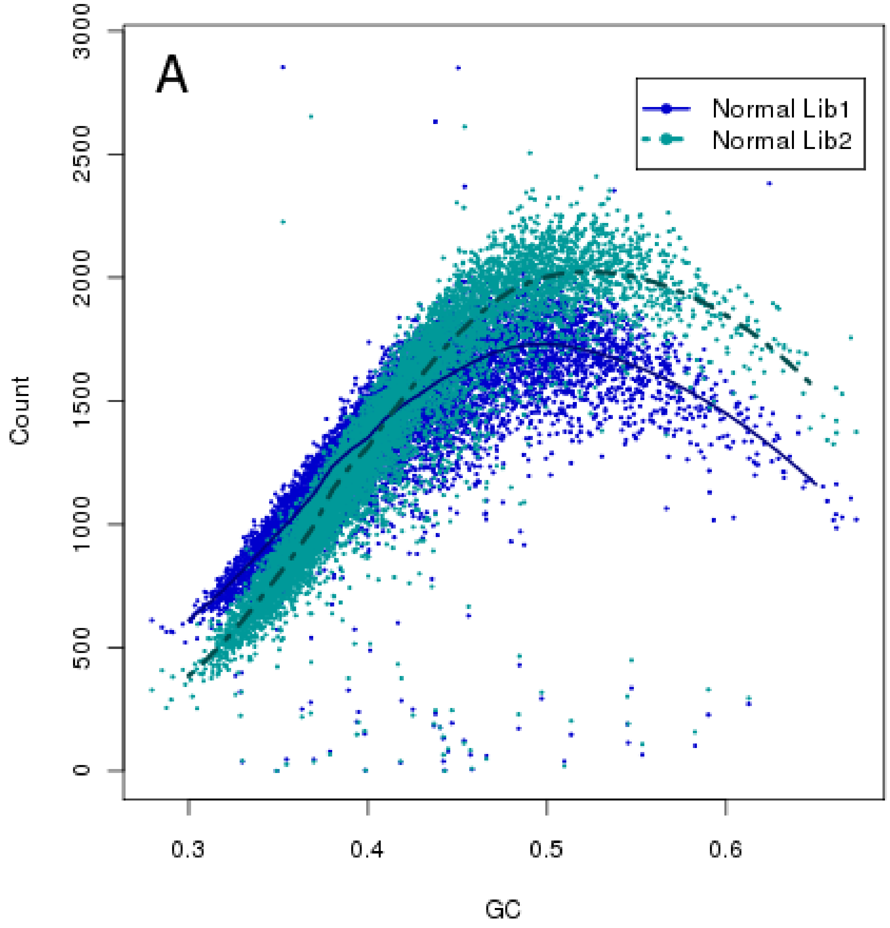


Counts for library 1 scaled to match those for library 2

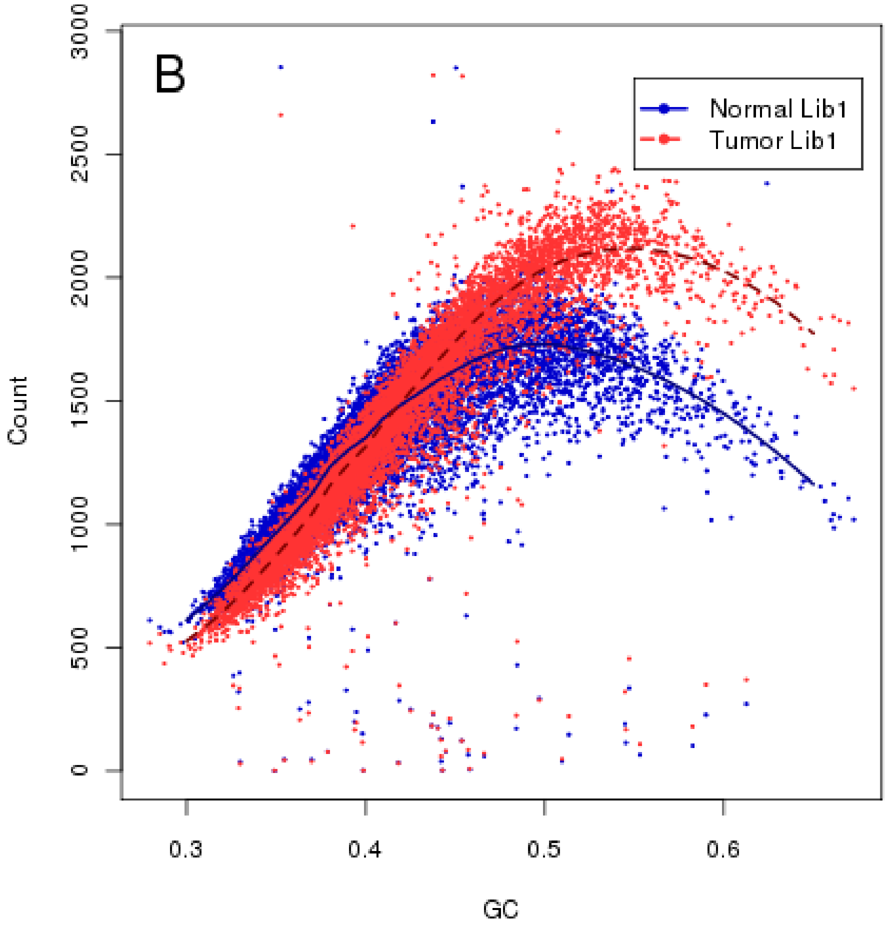
Is the GC-bias specific to a lab, protocol, sample, library preparation, sequencing machine,….?

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?

Normal libraries 1 and 2 (10 kb bins)



Tumor and normal library 1 (10kb bins)



Conclusion from more of the same: anything can matter.

Is there a right bin size?

People have used 100bp, 5 kb, 10 kb, 20 kb, 100 kb.

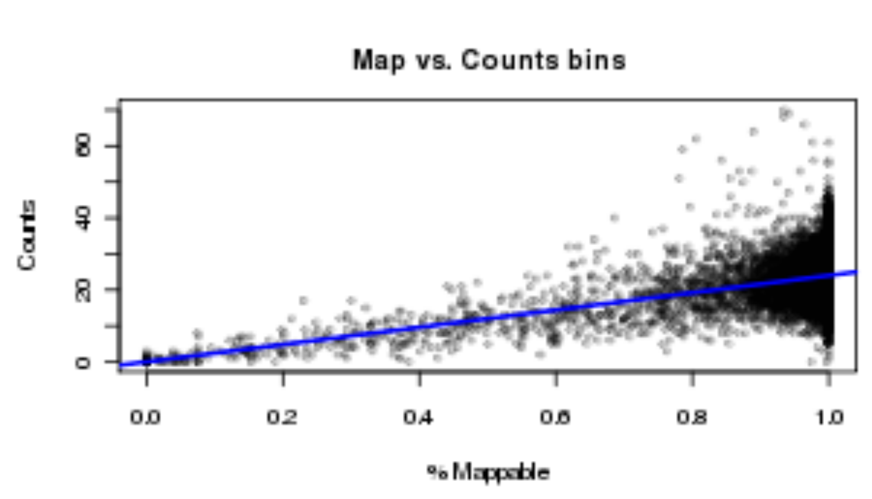
Variation about the smooth curve for different bin sizes

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

Answer: the smaller the better, until we lose 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.



Avoiding binning: single position analyses

We work with 10M mappable genome locations denoted by *x.*

We assign fragments to the 5’ end on the + strand.

The fragment count at location *x* may depend on the *GC* content of the window length *l*, offset *a* relative to x:

$$W_{a,l} = [x+a,\ x+a+l),$$

whose GC we will denote by $gc = GC(x+a,l)$.

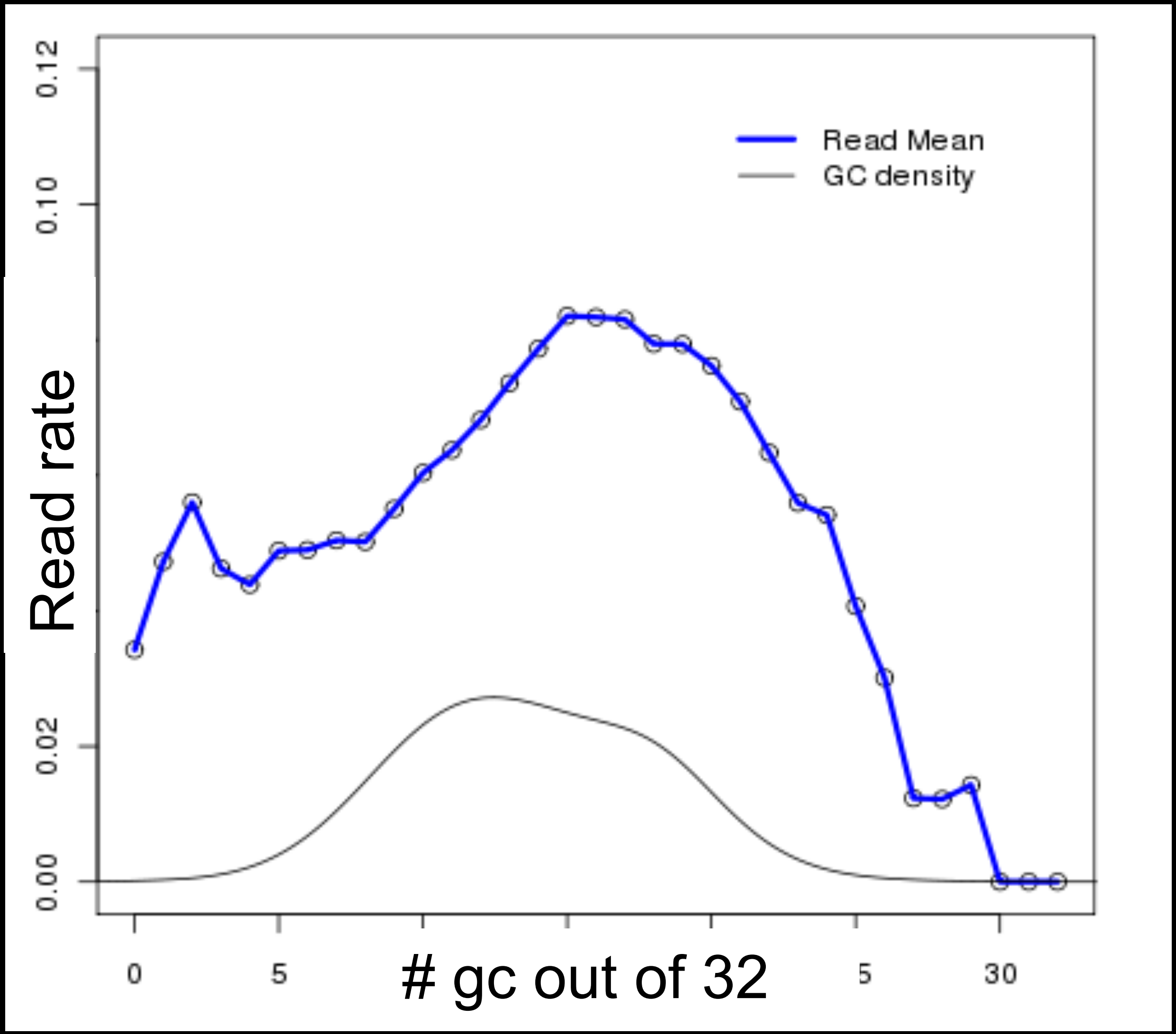
In symbols,

Let N_{gc} be the total number of x's whose window $W_{a,l}$ has $GC{=}gc$, and let F_{gc} be the total number of fragments mapping to such x's. The GC-stratified rate λ_{gc} and the overall rate λ of fragments mapping to such x's are estimated by

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

Apologies for omitting this slide and messing up this explanation in the lecture.

Rate vs GC curve: *a=0, l=32*



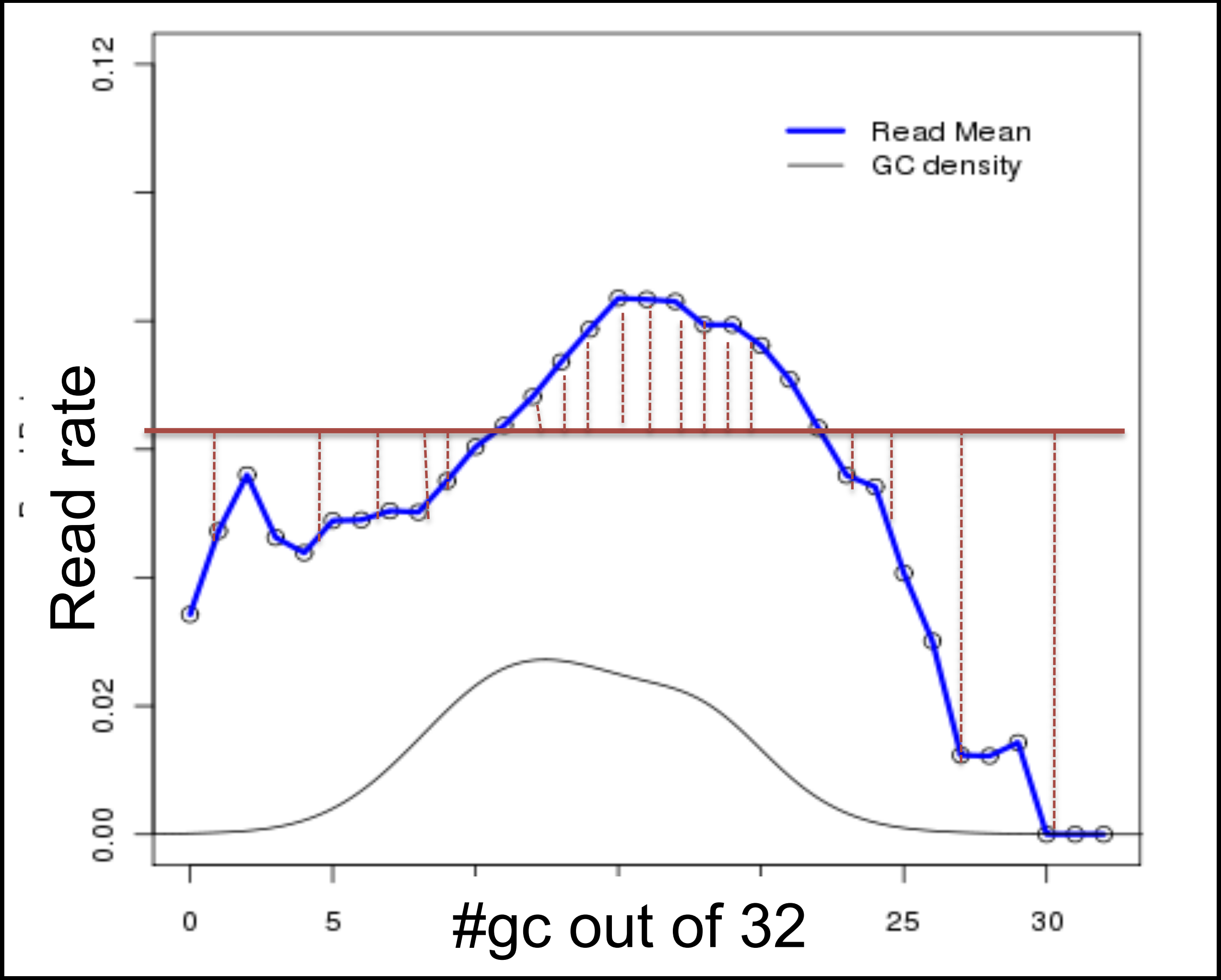
What's interesting about these *read rate* vs *GC-content* curves as we vary window size and location?

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.

TV distance from GC independence.



TV distance = a weighted average of the brown lengths

In symbols,

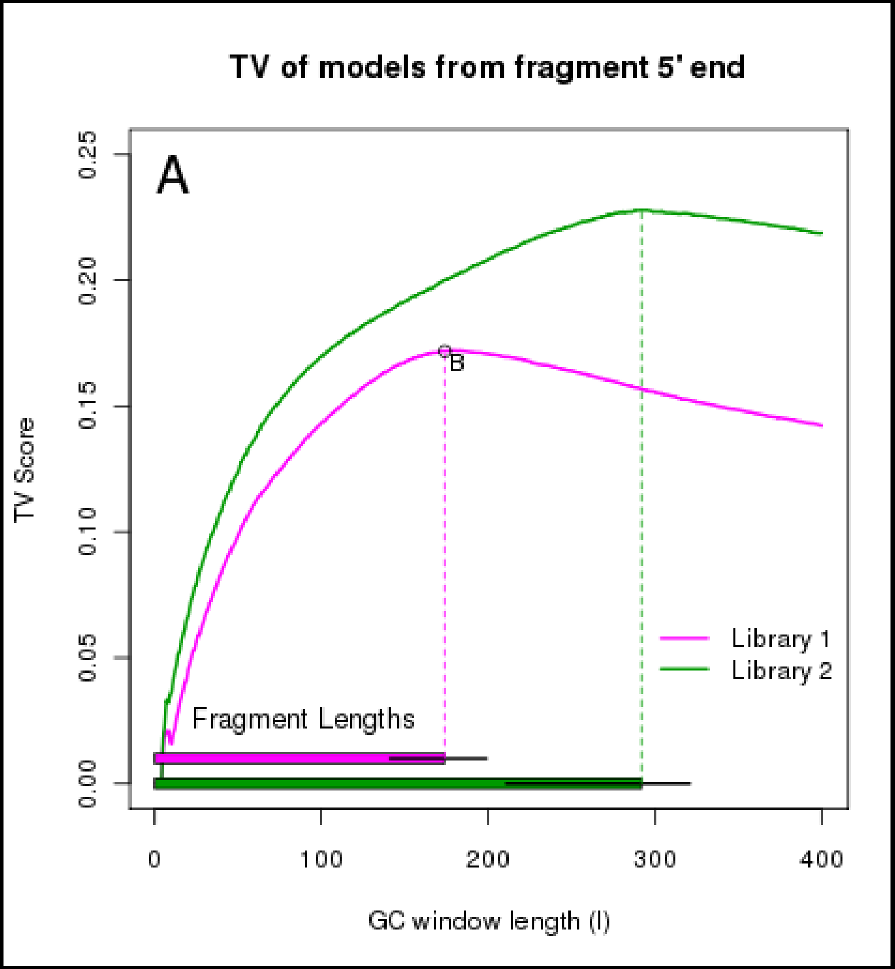
$$TV(W_{a,l}) = \frac{1}{2\hat{\lambda}} \sum_{gc=0}^{l} \frac{N_{gc}}{n} \mid \hat{\lambda}_{gc} - \hat{\lambda} \mid,$$

where $W_{a,l}$ *is the window* $[x + a, x + a + l)$,
and n *is the total number of* $x's$.

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

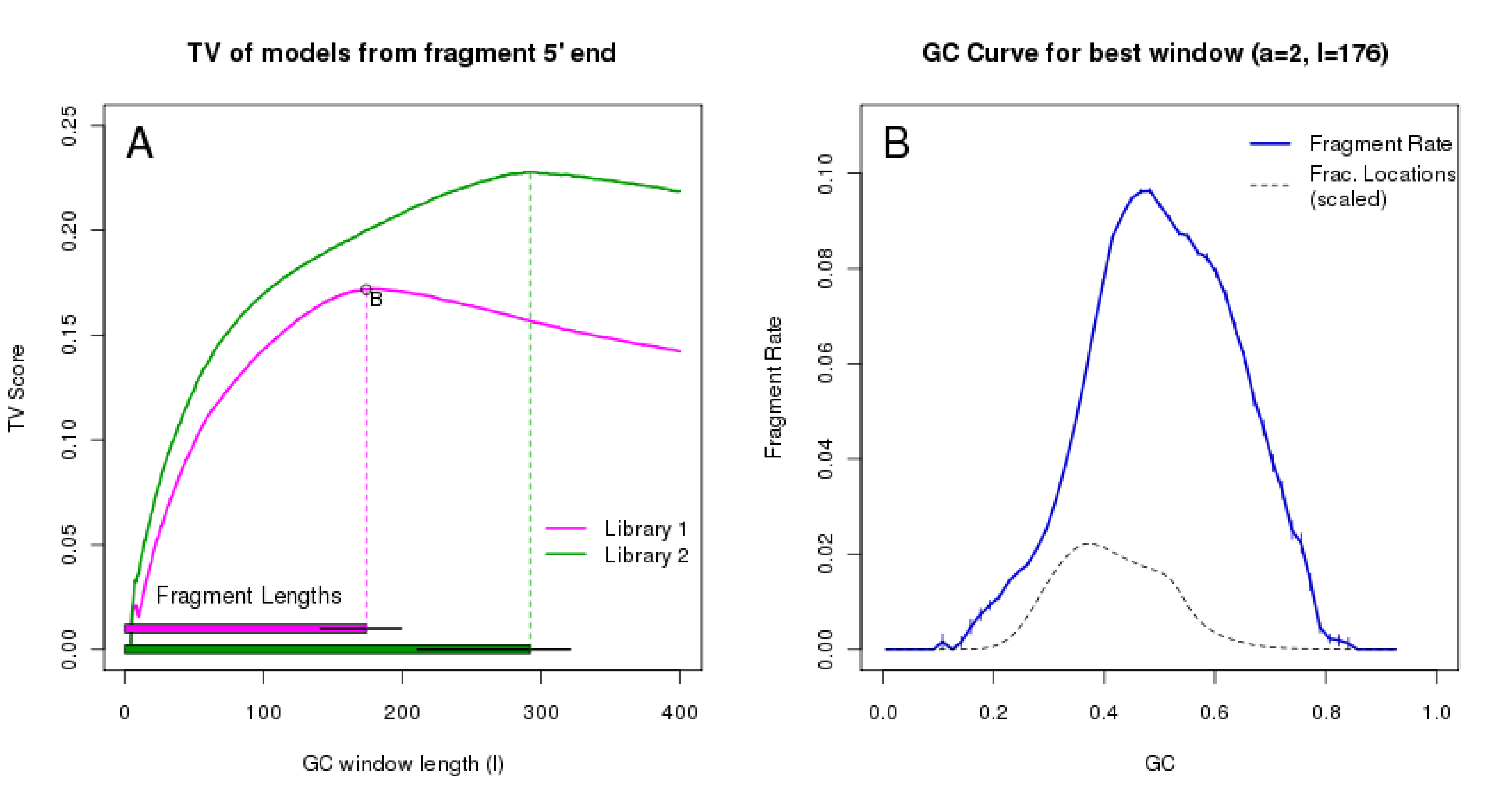
Varying the window *size* from a fixed point

(here the 5’-end of the fragment)



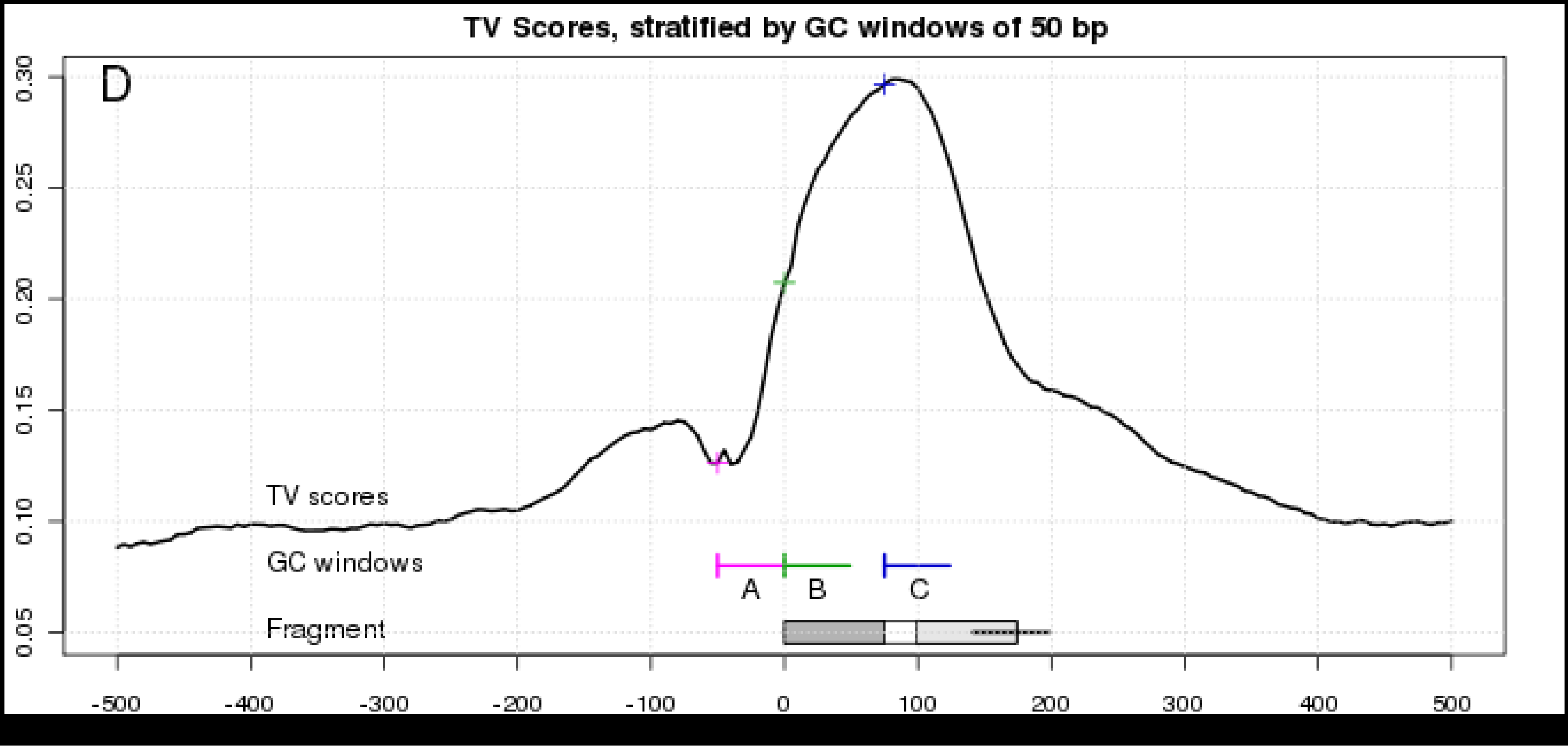
Varying the window *size* from a fixed point

(here the 5'-end of the fragment)



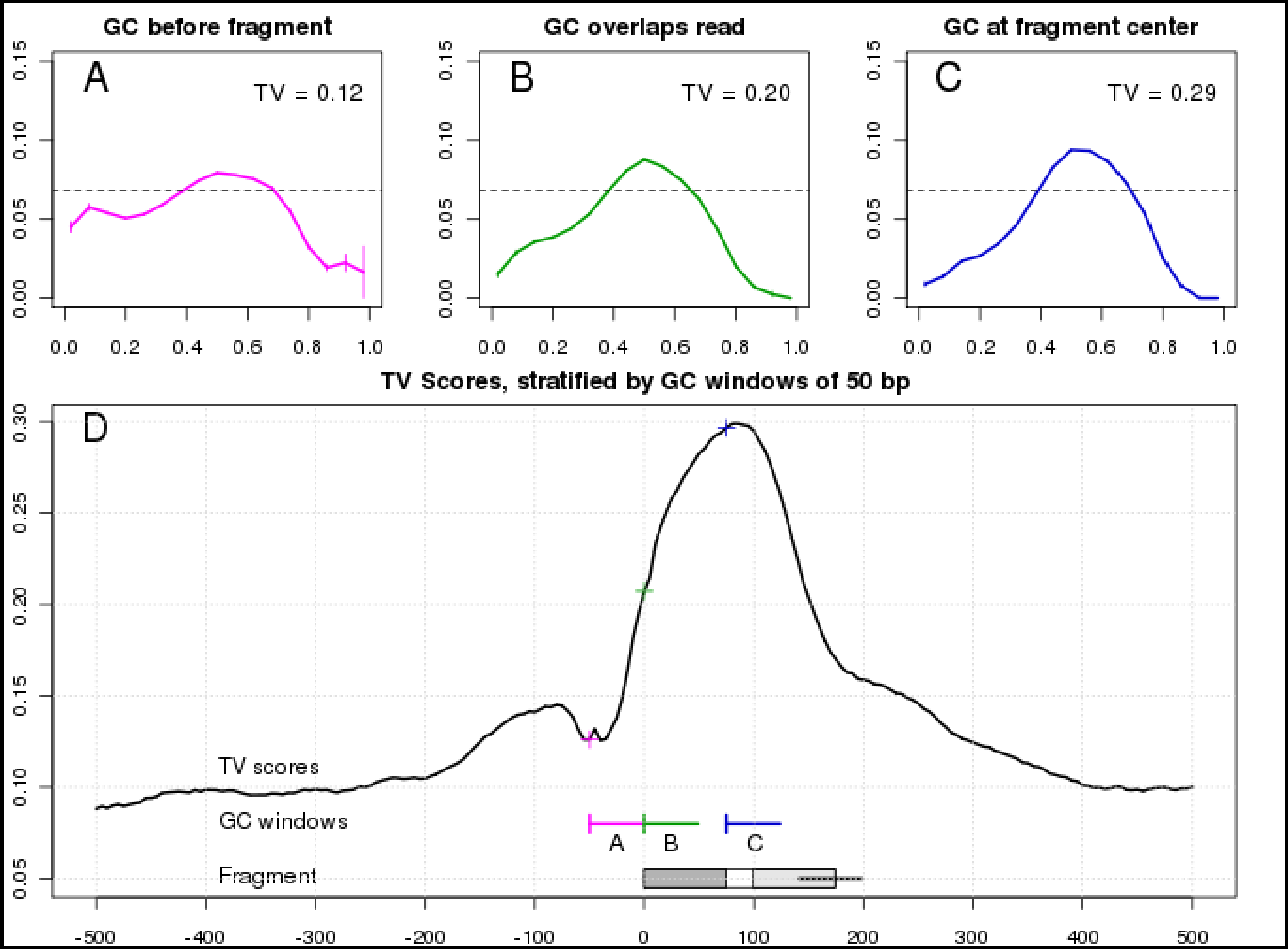
Varying the *location* of a fixed size window

(here 50 bp; library 1)



Varying the *location* of a fixed size window

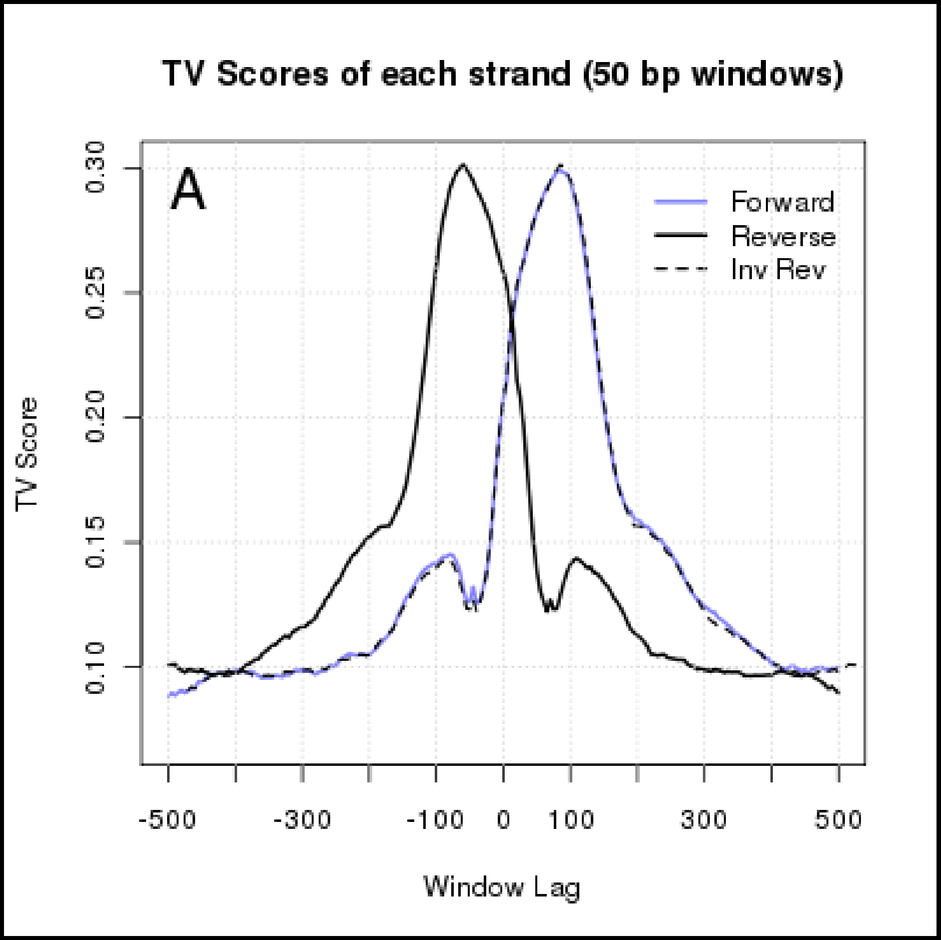
(here 50 bp; library 1)



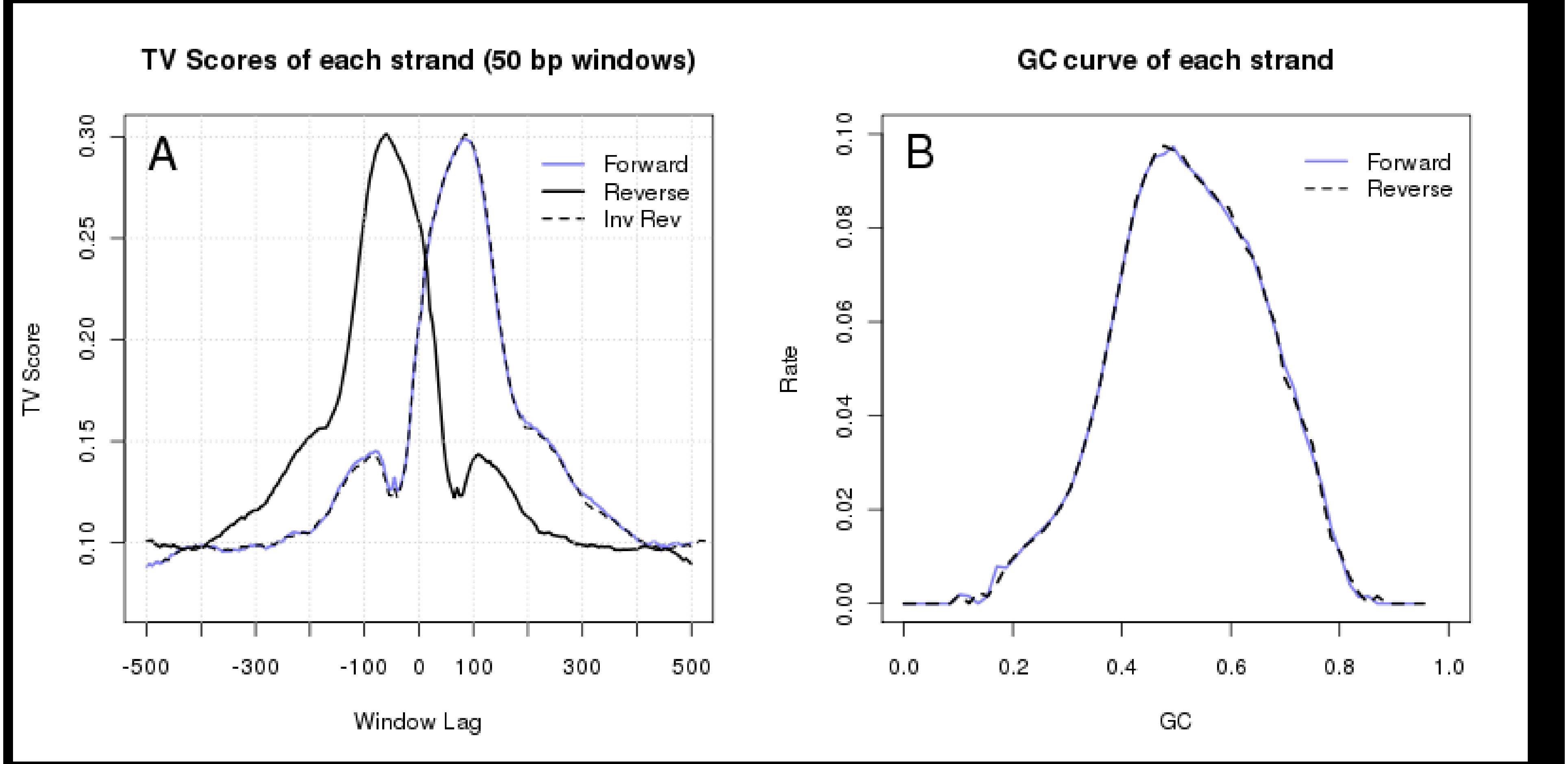
Interim conclusion from many such plots

The *best* interval is in the middle of the fragment, excluding the bits at the very ends.

Forward and reverse strands behave similarly



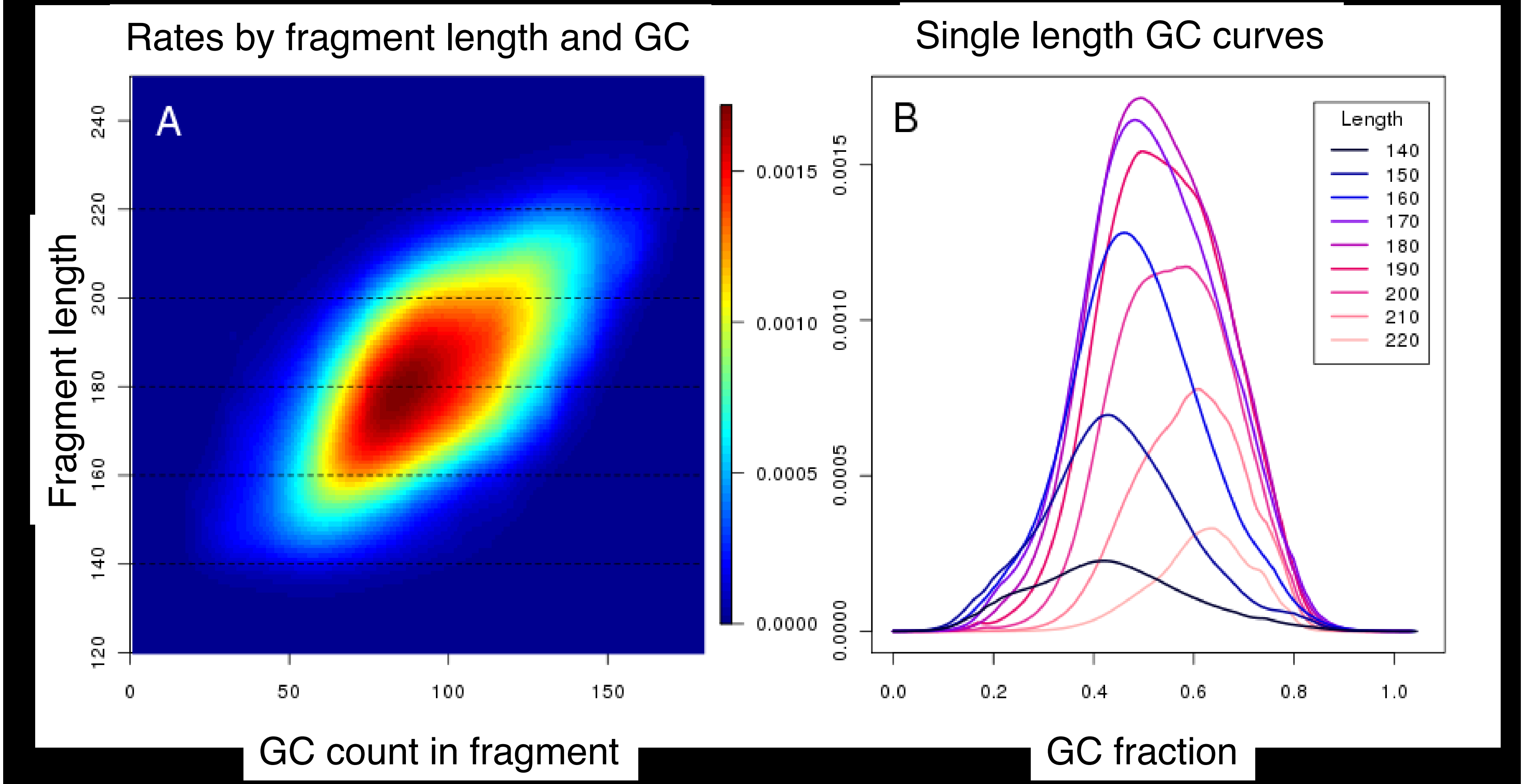
Forward and reverse strands behave similarly



Stratifying by fragment size *s*

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

Fragment size matters



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

And now for some predictions

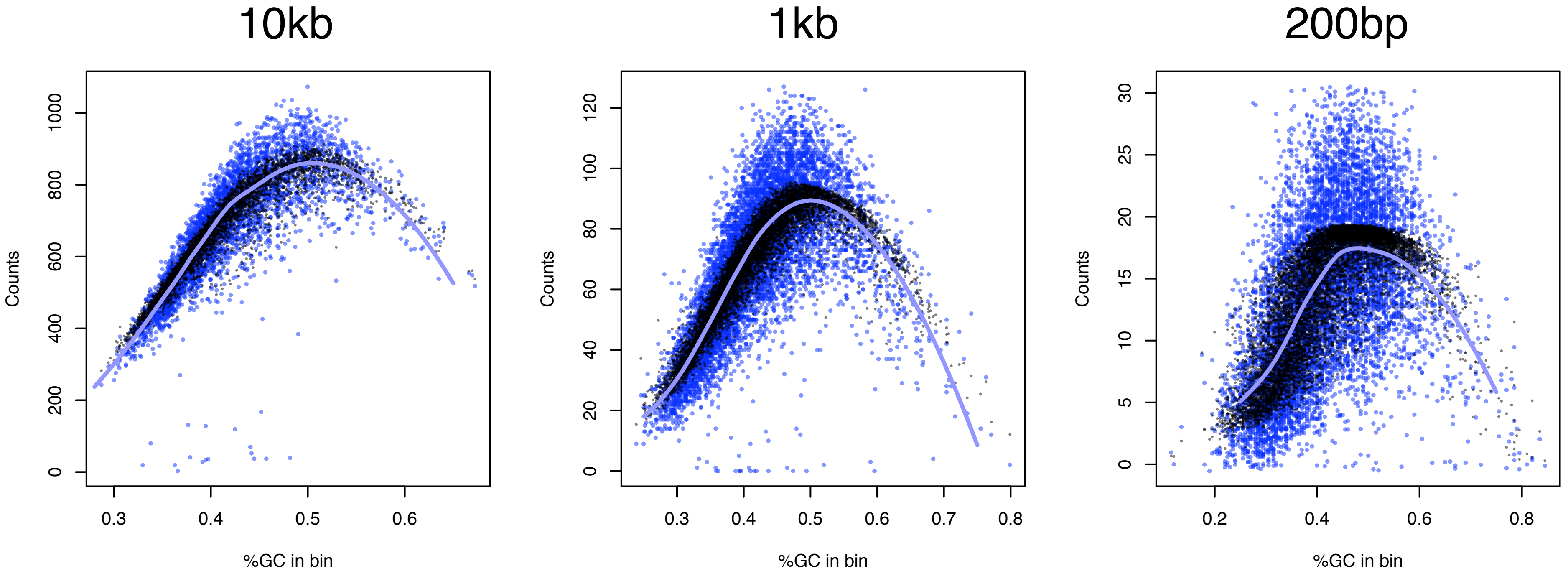
Predicted rates at a given mappable position

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

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

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

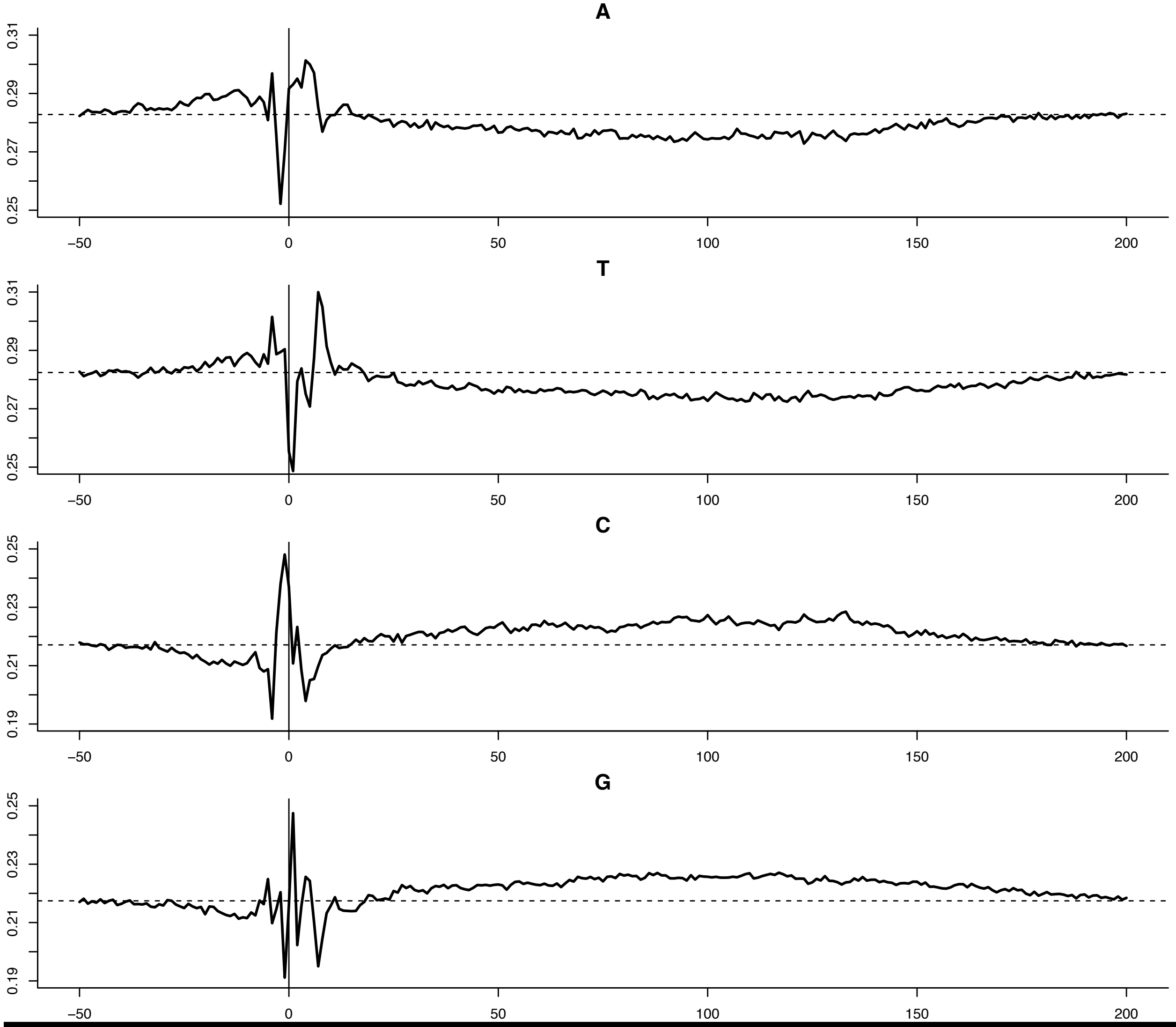


black = predicted, blue= observed
lowess lines are based on the observed points.

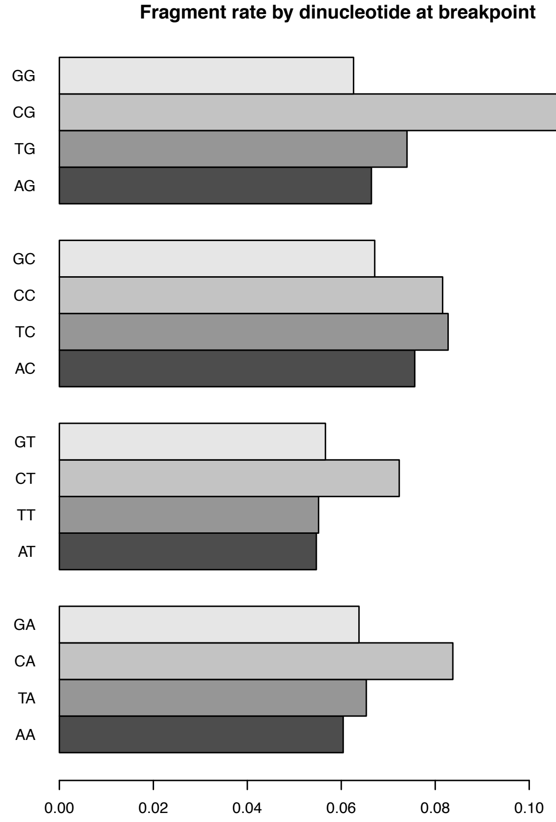
Conclusion: the predictions seem to be working.

Some other biases/models

Breakpoint effects

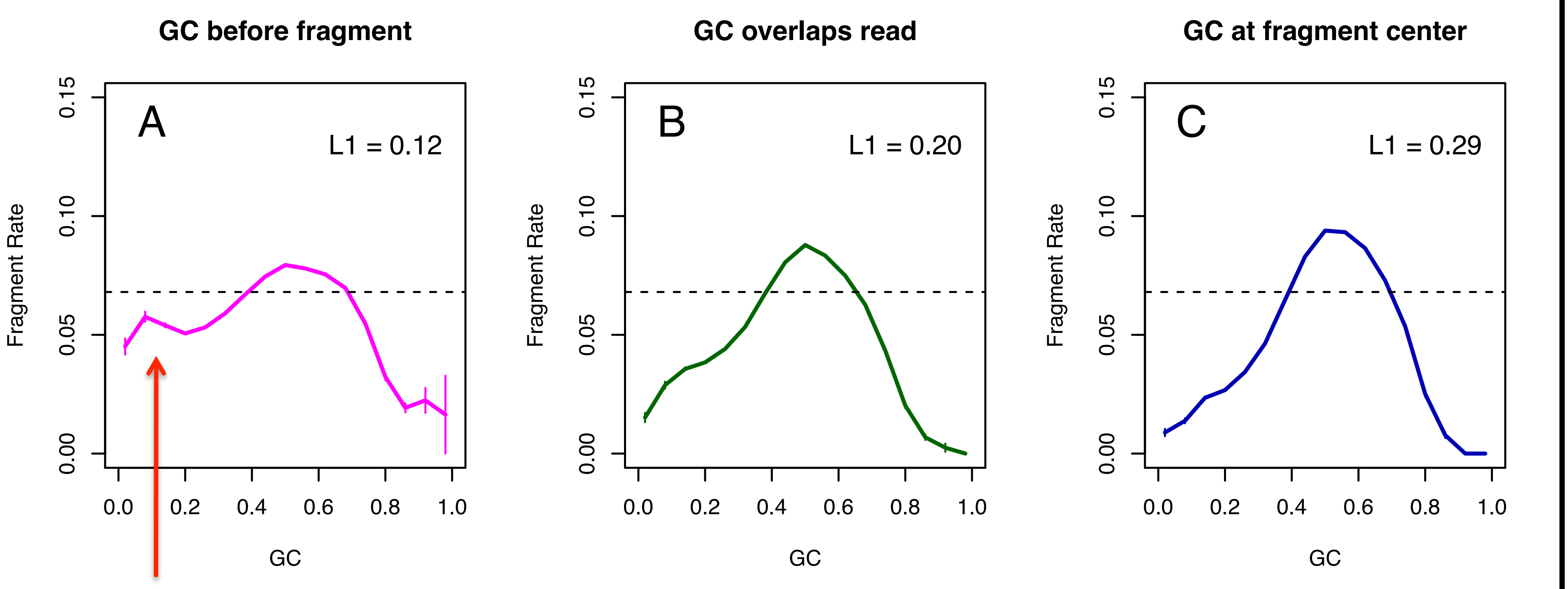


Fragment rate by dinucleotide at breakpoint



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

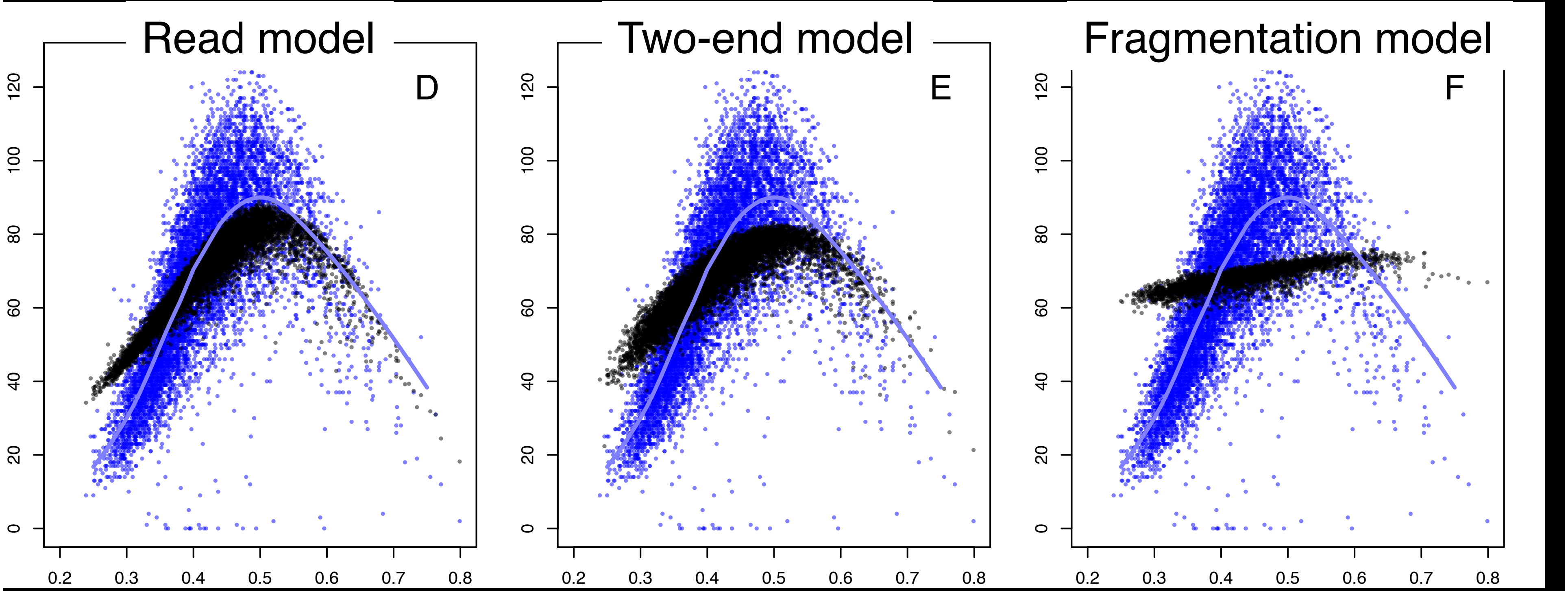
End effects



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)

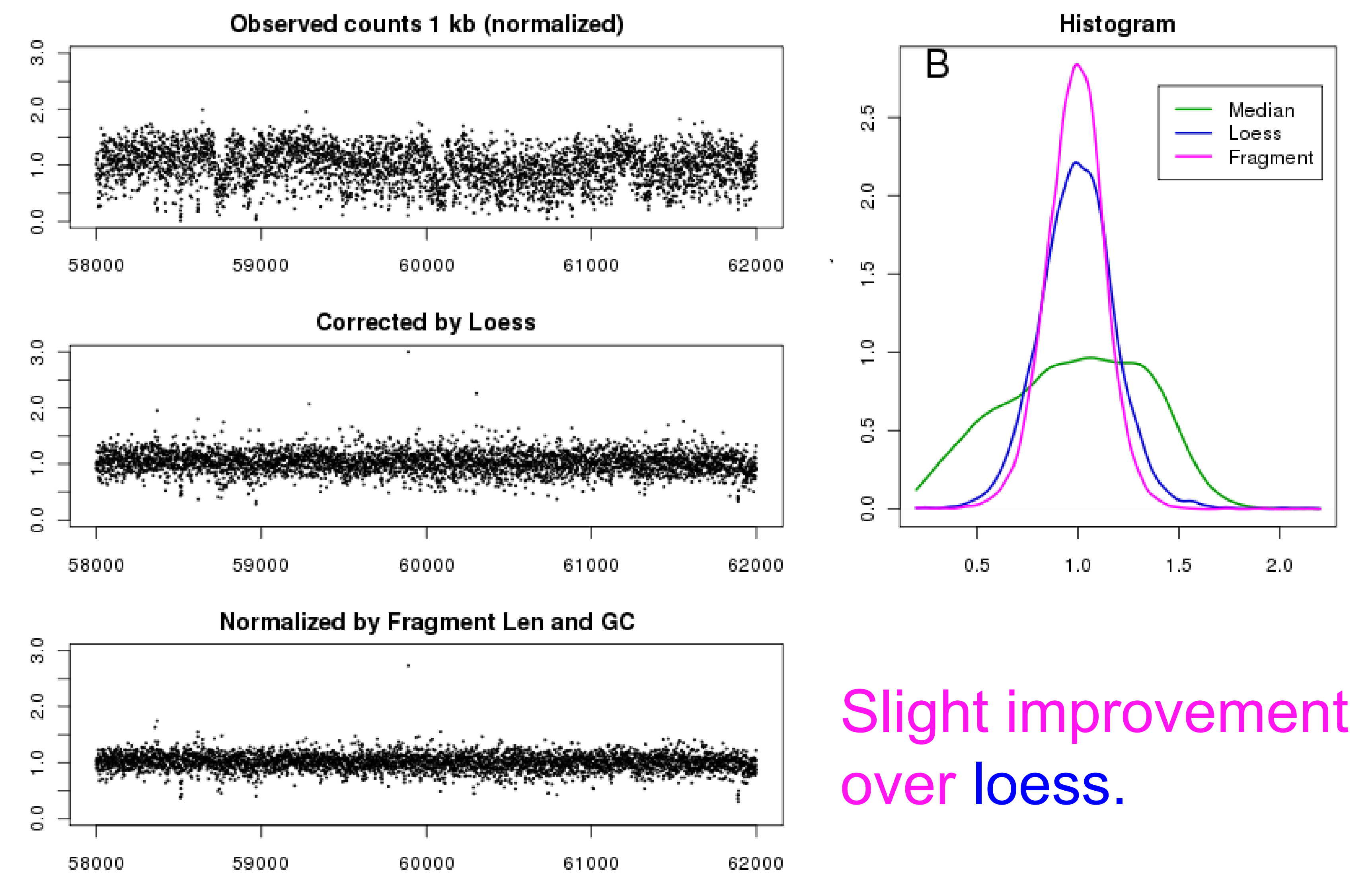


Conclusion: These predictions don’t work too well.

How well does our correction work?

Practice

Copy number: corrections to normal samples

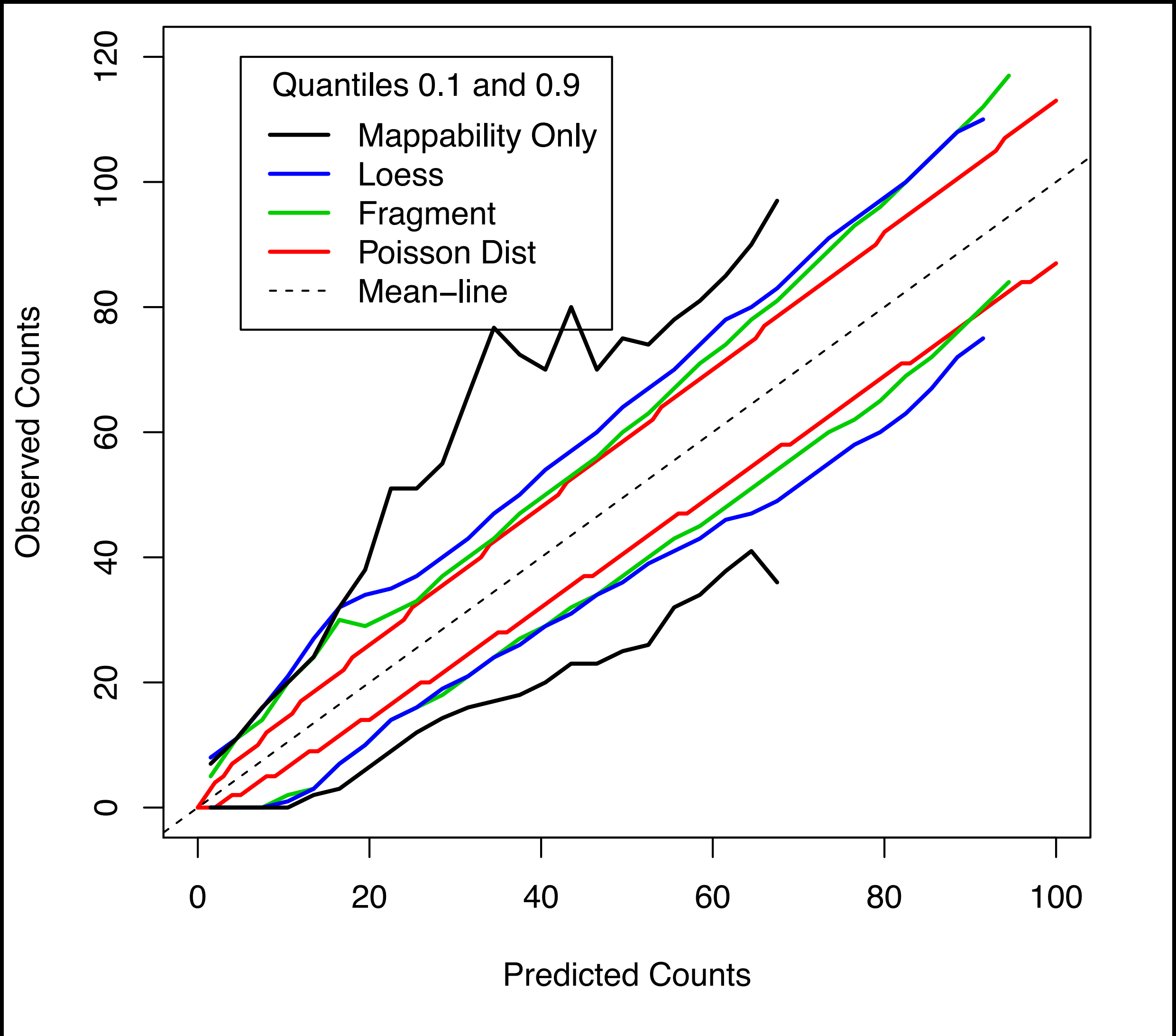


Slight improvement over loess.

How well does our correction work?

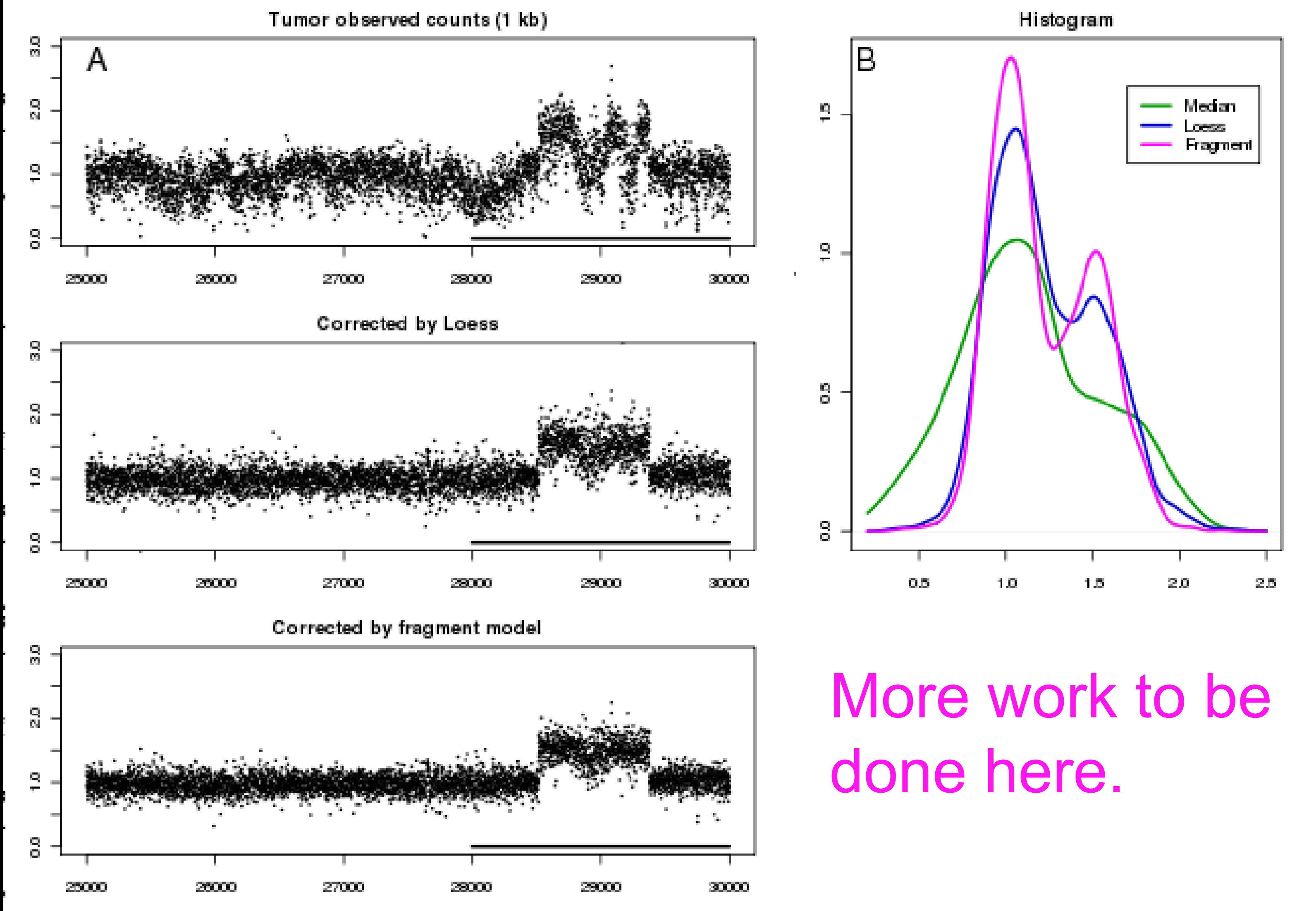
Theory

Spread of observed counts around predictions



Conclusion: we don't "explain" everything.

Crude corrections to tumor samples

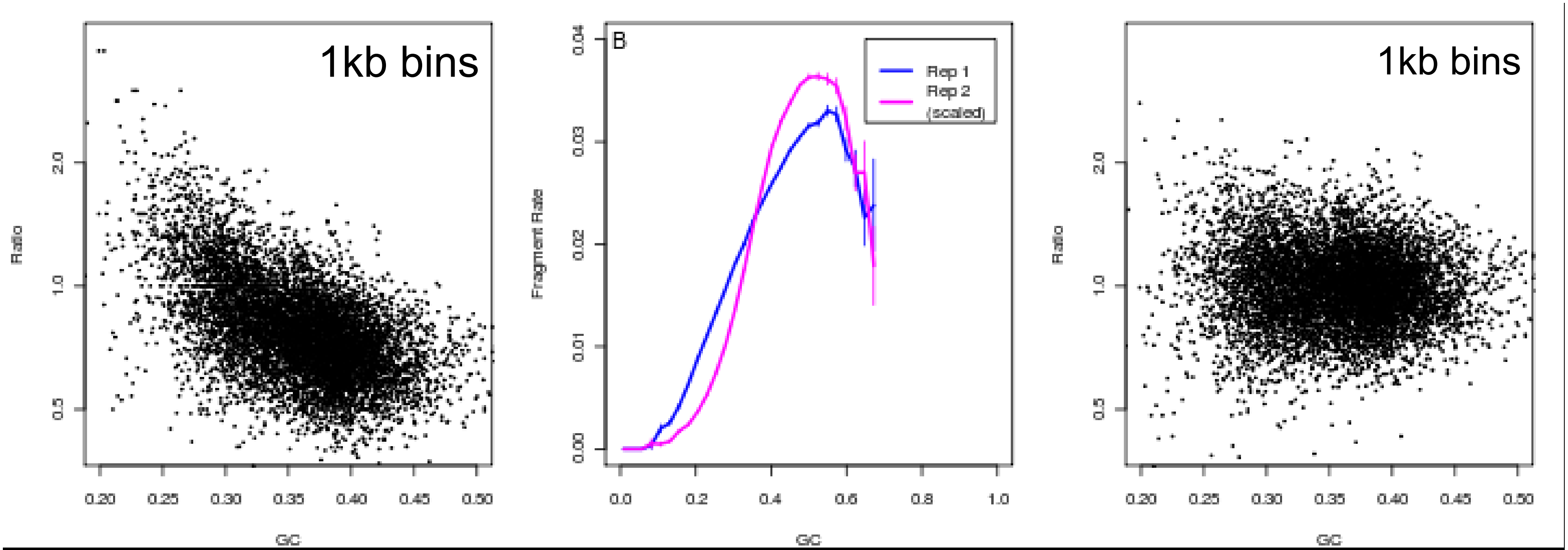


More work to be done here.

ChIP-seq data (*A. thaliana*)

Here two initially incompatible *technical* replicates

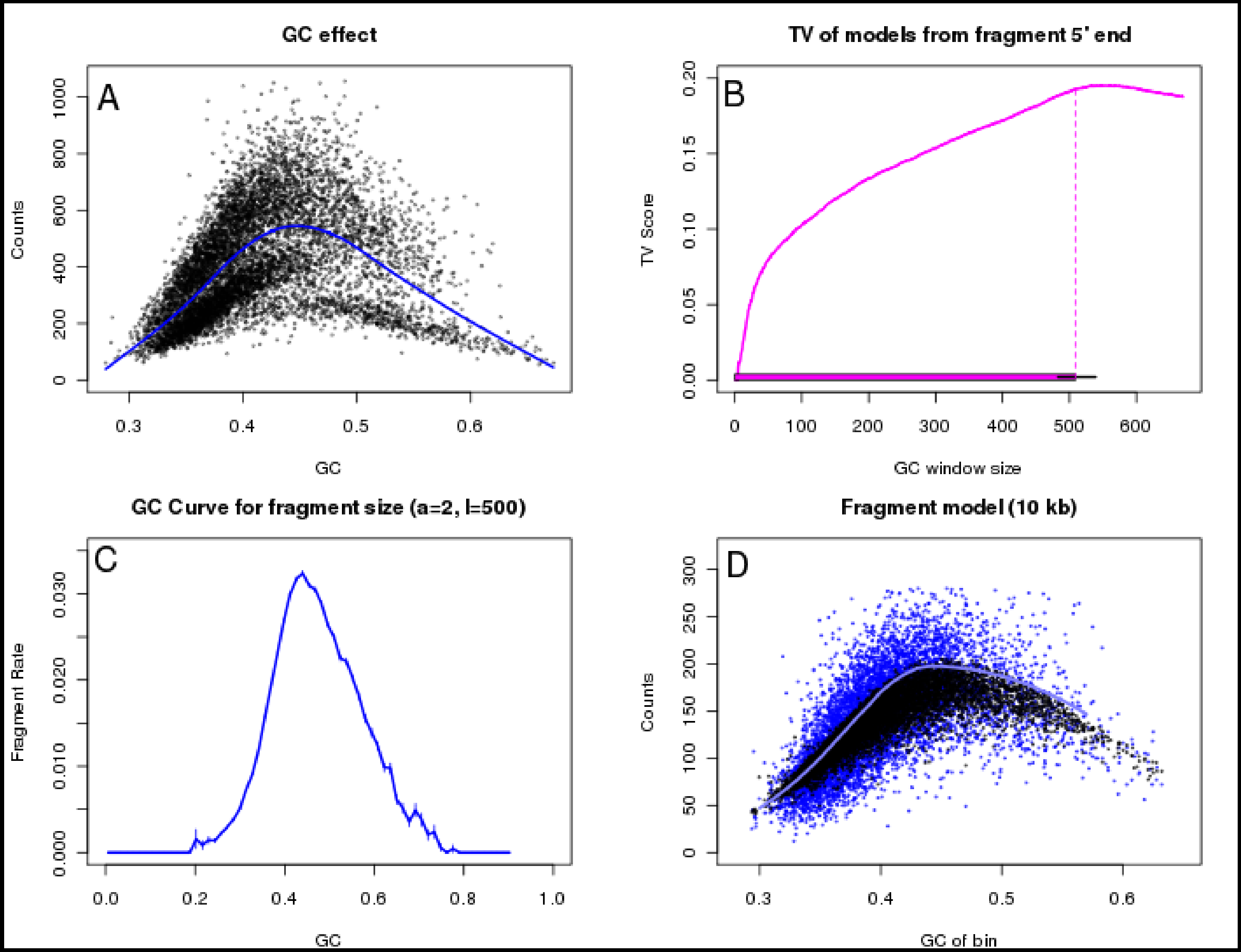
Uncorrected ratios — GC curves (a=2, l=122) — Corrected ratios



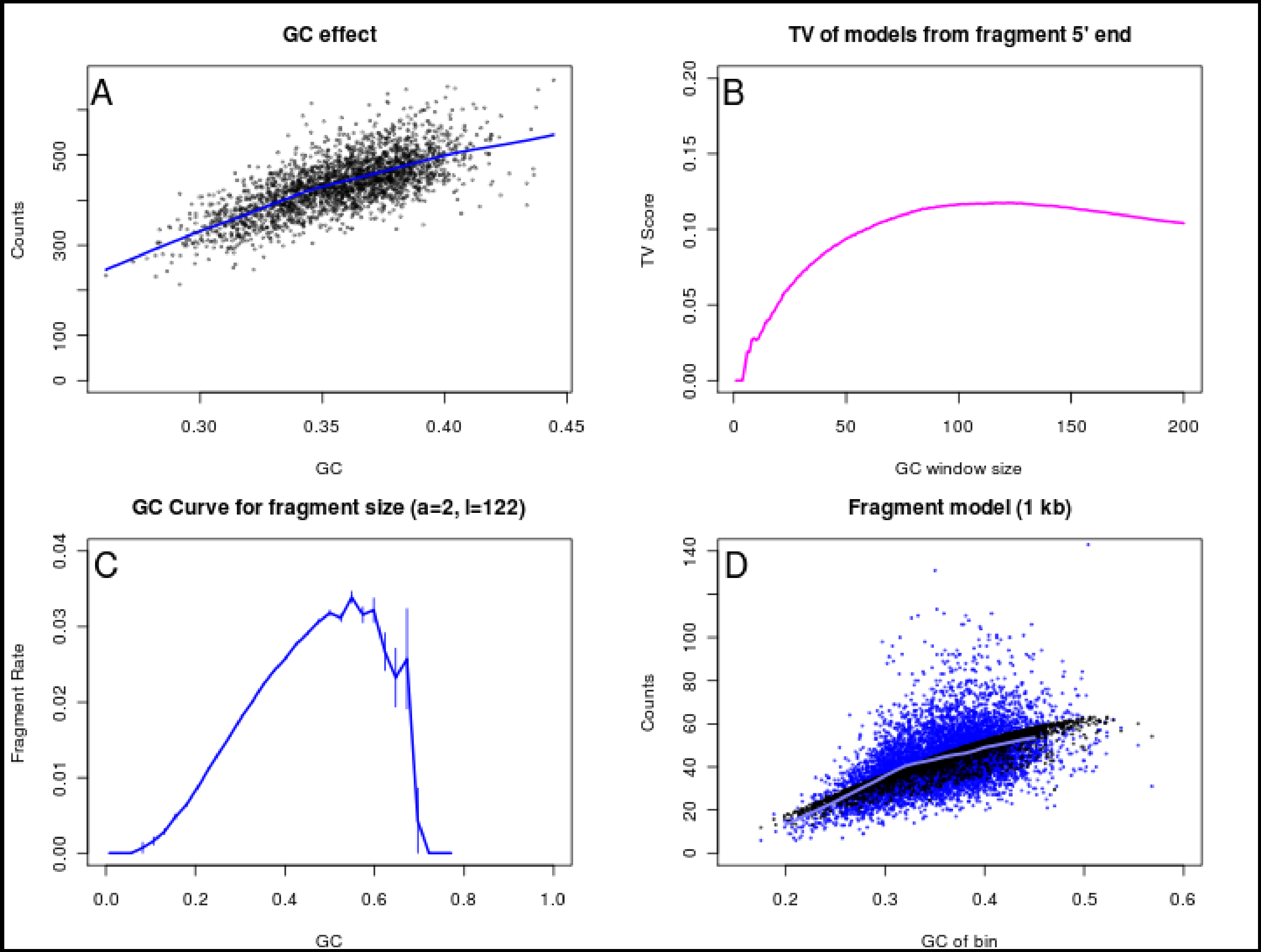
Problem mainly solved (*cf* Cheung *et al*, 2011)

Other examples and phenomena

Plots for a BrCa tumor

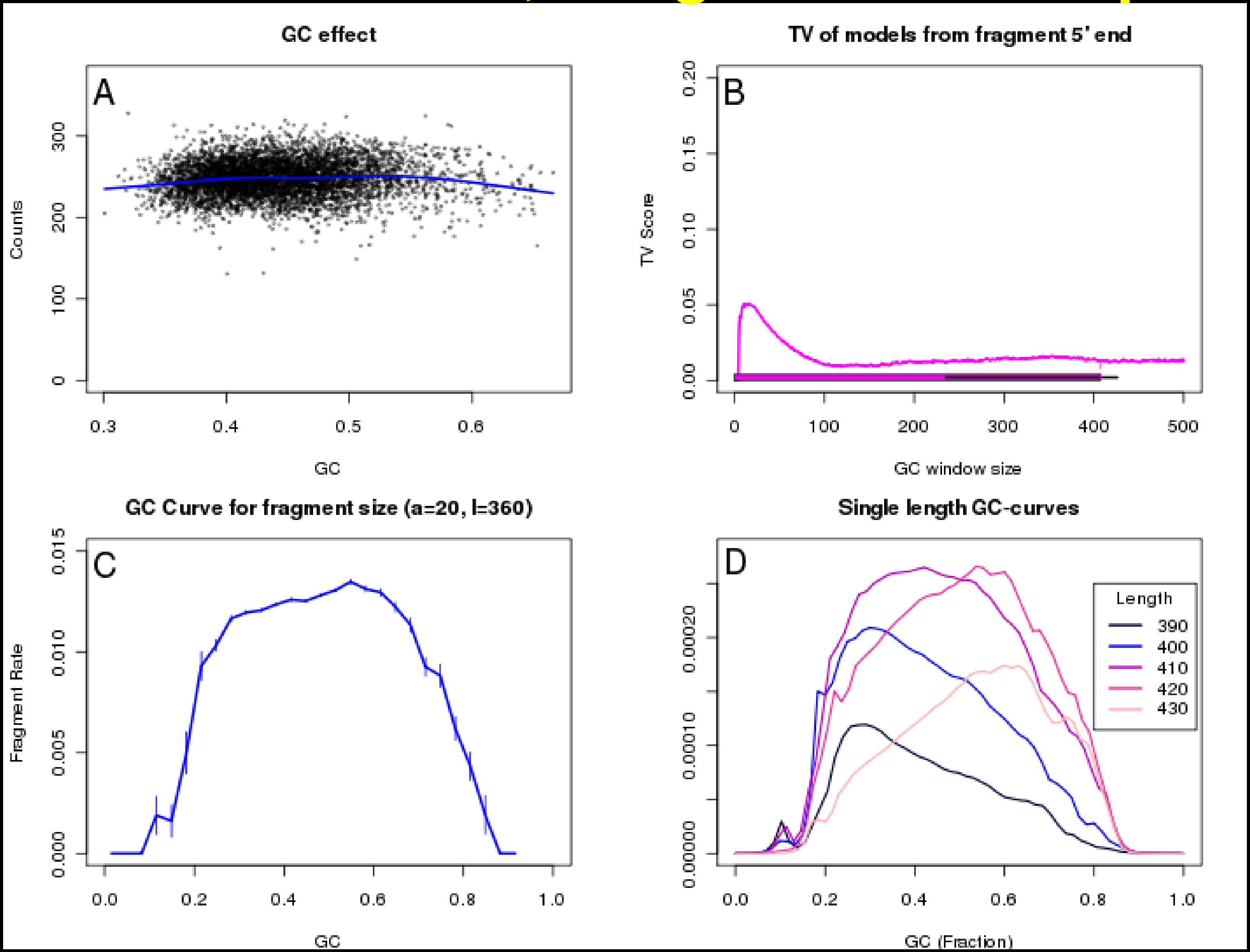


Plots for ChIP-seq sample rep 1, *A. thaliana*



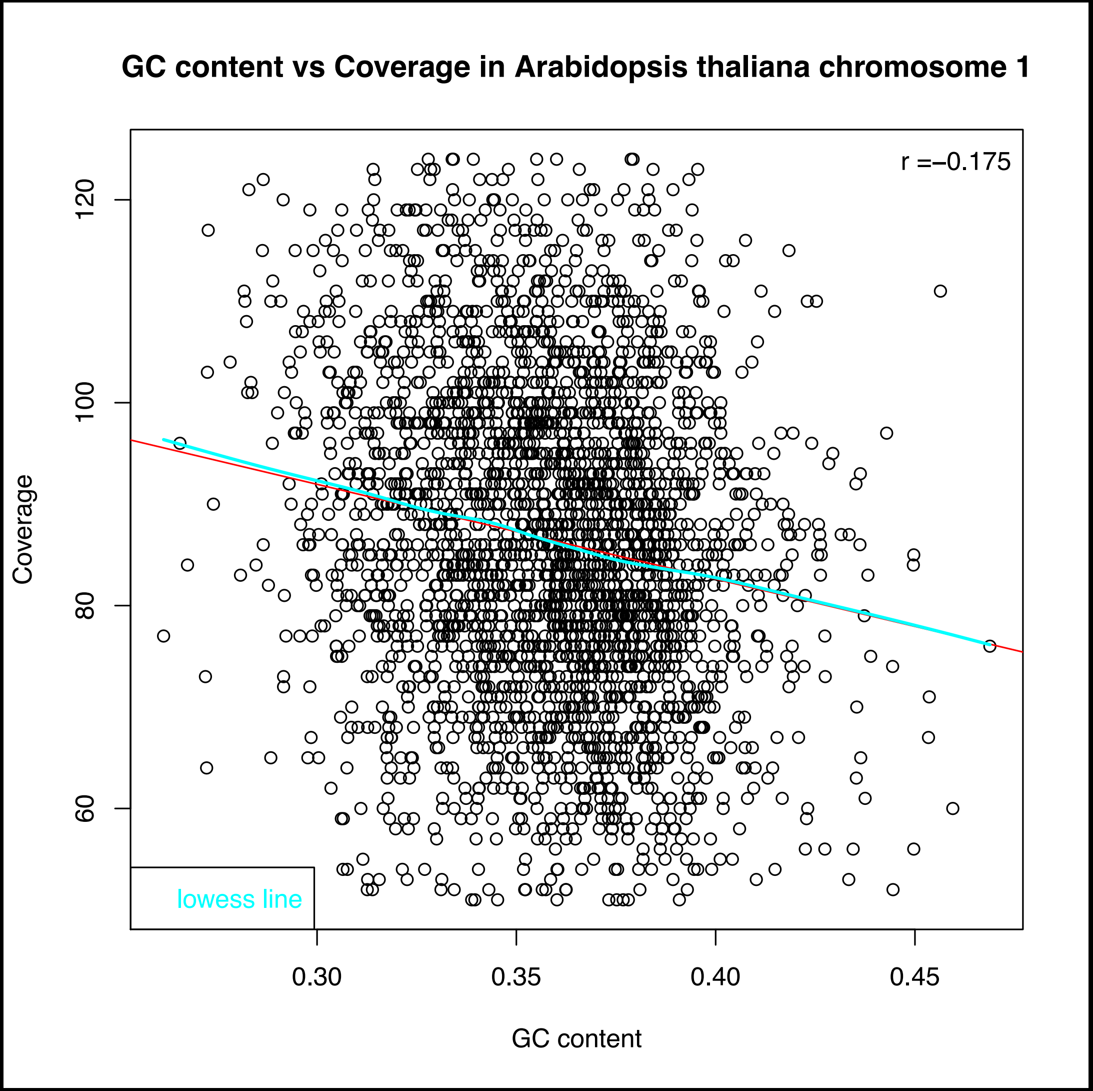
Plots for one 1,000 genomes sample

Note scale here



Typical (?) Pacific Biosciences result

Reads filtered to be > 1kb, >85% accurate
Bin size: 10kb
Bottom and top 2.5% omitted



Thanks to Malinka Jansson & Jim Bullard, PacBio

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.

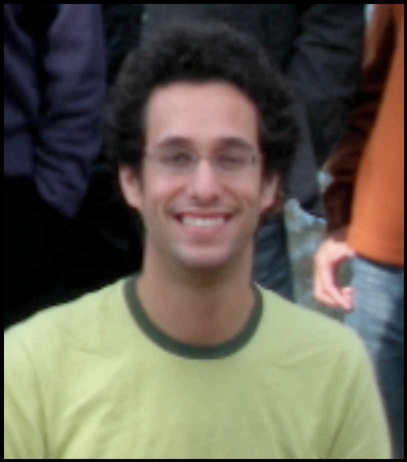
Base composition (not just GC-content) around the two fragment break points plays a noticeable role, but not enough to explain everything.

Speculation over causes is left for another day. There now seems little doubt that PCR amplification bias of the fragment 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*.

All of the above and more can be found in Tech Report #804
http://www.stat.berkeley.edu/25
With luck it will appear in NAR soon. An R package GCcorrect is almost ready (11/3/11)

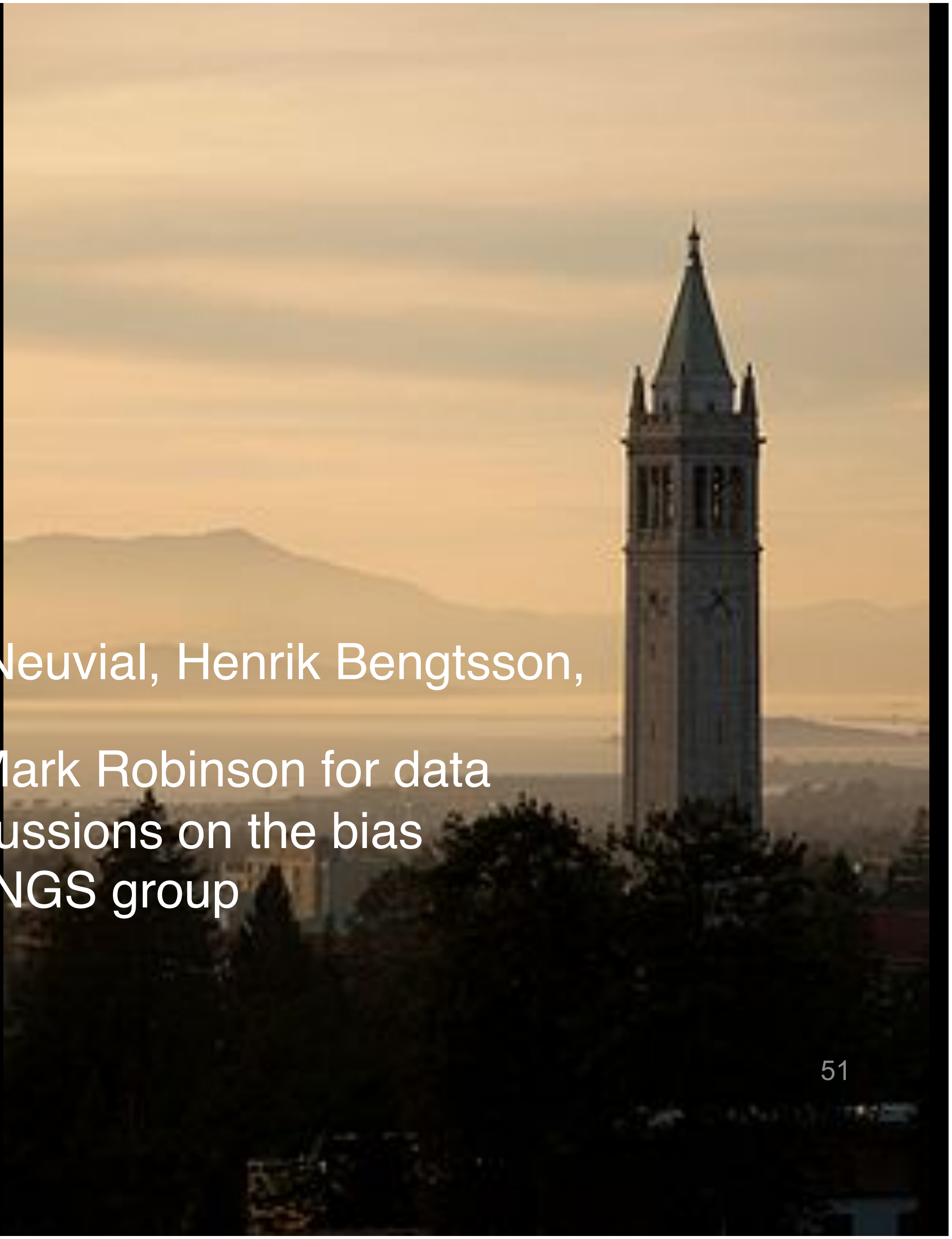
Many thanks to

- **Yuval Benjamini**

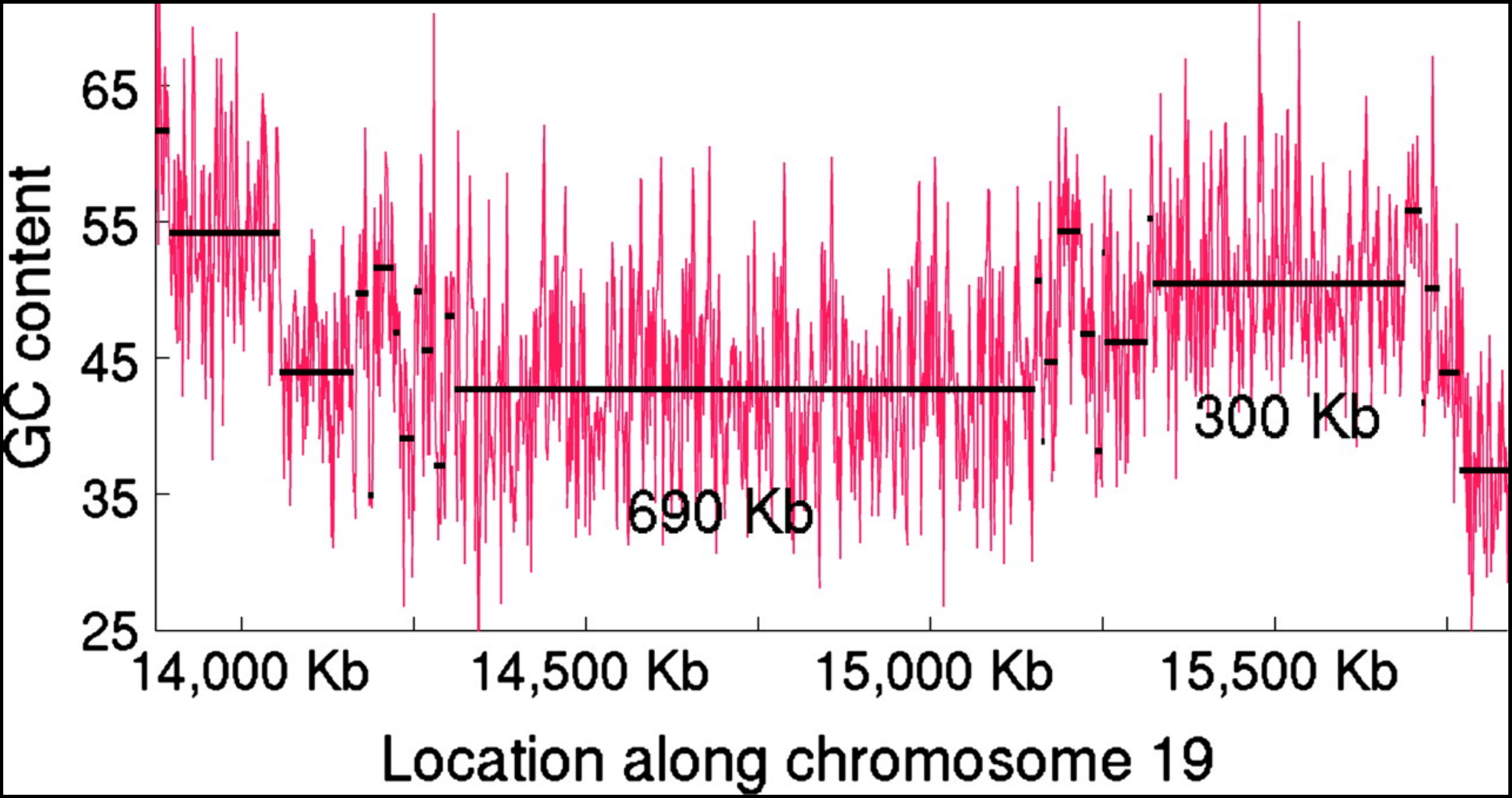


- Oleg Mayba, Pierre Neuvial, Henrik Bengtsson, & Su Yeon Kim
- Paul Spellman and Mark Robinson for data
- Leath Tonkin for discussions on the bias
- The whole Berkeley NGS group
- NIH NCI TCGA

And to you…



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

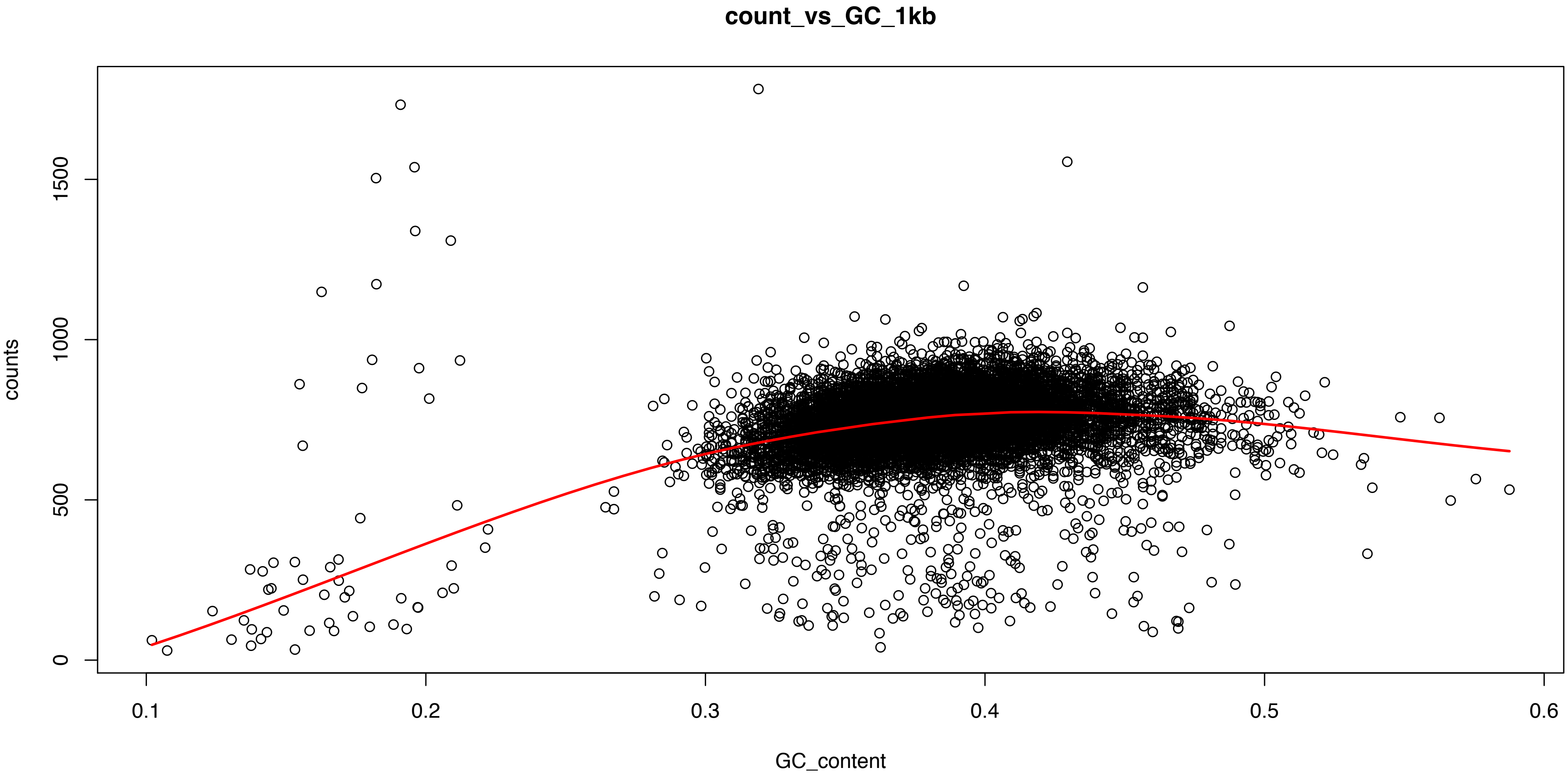


Cohen N et al. Mol Biol Evol 2005;22:1260-1272

The isochores probably don't exist paper.

MBE MOLECULAR BIOLOGY AND EVOLUTION

S. cerevisae GC curve (1kb bins)



Our library prep is a bit different from the Illumina protocol, for one of the steps, we used a heat inactivation step to the stop the enzyme (after the polyadenylation step) instead of using the column or beads to purify the library prep again. (Lin Gen)

Procedure to get a Rate vs GC curve

- Random sample *10M* uniquely mappable locations *x*
- Stratify by the GC-value of the window

$$W_{a,l} = [x+a,\ x+a+l),$$

- Count # reads in each GC-stratified window
- Compute *Rate = # reads / # locations*
- Plot and smooth the Rate vs GC curve

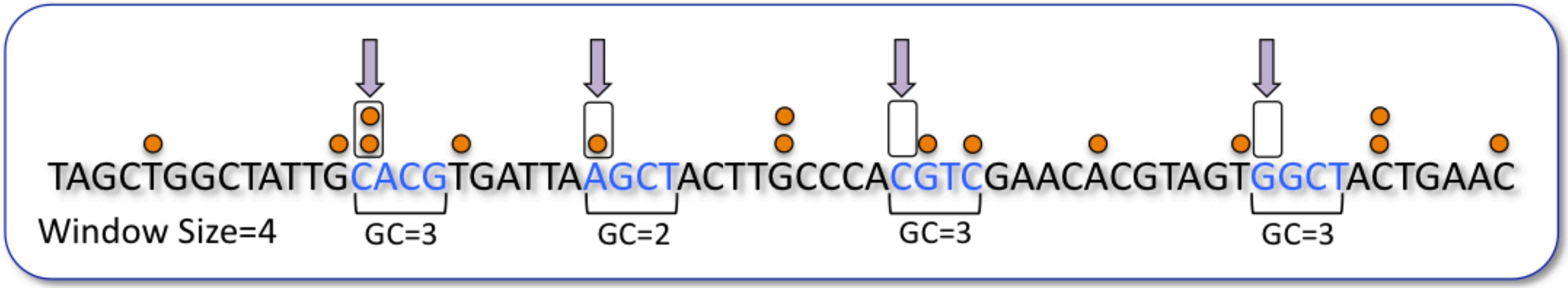
Our overall goals

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

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]

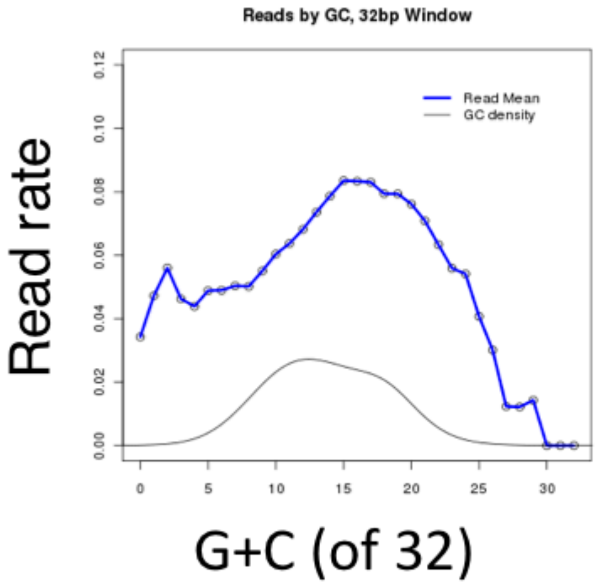
A) Random sample locations B) Partition by GC window C) Count reads and read-rate


TAGCTGGCTATTGCACGTGATTAAGCTACTTGCCCACGTCGAACACGTAGTGGCTACTGAAC
Window Size=4
GC=3
GC=2
GC=3
GC=3

GC	0	1	2	3	4
⇩ Locations	-	-	1	3	-
● Reads	-	-	1	2	-
Rate	-	-	1	0.66	-

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

D) Plot GC curve


Reads by GC, 32bp Window
Read Mean
GC density
Read rate
G+C (of 32)