Hobby Genomics with R

Earl Hubbell

Getting Personal Genomic Data

- I signed up for Seeq
  - "tl;dr: We want to make genomics inexpensive, interactive, and fun; get the app at seeq.io."
  - "The high-level technical description is that we're running an ultra-low-coverage genome sequencing assay."
- They return the raw data
  - "What is "raw data"? For our purposes, these are the sequencing reads generated from your sample during our "ultra-low-coverage" sequencing assay, provided in BAM format."

What is low-coverage sequencing?

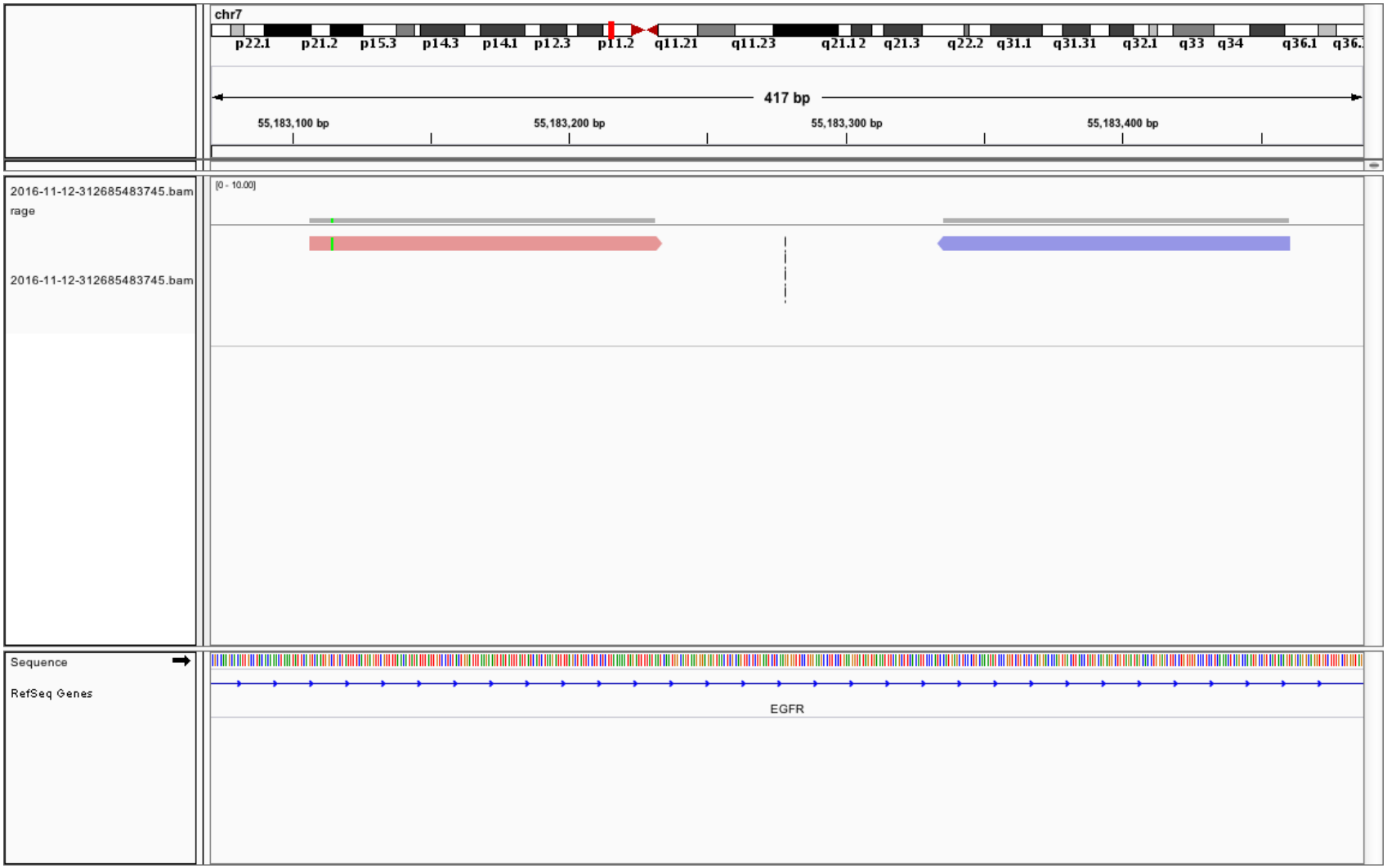


Figure 1: Paired reads: 2x125

Personal Genomics

- Goal: Copy Number Analysis
  - Am I a typical human?
- Tools To Analyze The Data
  - Auxillary data
- Correct Systematic Effects
  - Clean the data

Basic Copy Number Analysis

- (Most) Humans have 23 chromosome pairs
  - 1-22 2 each
  - 2X, 0Y (most) female
  - 1X, 1Y (most) male
  - MT = mitochrondria, lots of copies
- Can miss/gain little bits compared to reference

Read Depth Model

- Probability of seeing a read uniform
  - 2 copies = standard probability
  - 1 copy = 1/2 probability
- Count reads falling in genomic regions
  - Read depth proportional to copies
  - Account for systematic effects

Homebrew: getting some standard tools

/usr/bin/ruby -e "$(curl -fsSL
https://raw.githubusercontent.com/Homebrew/install/
master/install)"

```
brew tap homebrew/science

brew install tabix

brew install samtools

brew install bedtools
```

Setting up the data and references

- which version of the genome was used?

```
samtools view -H 2016-11-12-312685483745.bam | more

@PG ID:bwa PN:bwa VN:0.7.12-r1039
CL:/nfs/sw/bwa/bwa-0.7.12/bwa mem
/gpfs/commons/groups/pickrell_lab/datasets/
genome_data/Genome/hg19/human_g1k_v37.fasta
```

Getting the version of the reference

- human genome with index

`http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/human_g1k_v37.fasta.fai`

`http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/human_g1k_v37.fasta.gz`

human_g1k_v37.fasta

human_g1k_v37.fasta.fai

Setting up bins across the genome

- Make bins of size 100kb

bedtools makewindows -g chr.sizes -w 100000 >chr.bed

- know I'm going to look at gc content

bedtools nuc -fi human_g1k_v37.fasta -bed chr.bed >chr.nuc.txt

Let's look at coverage across the genome

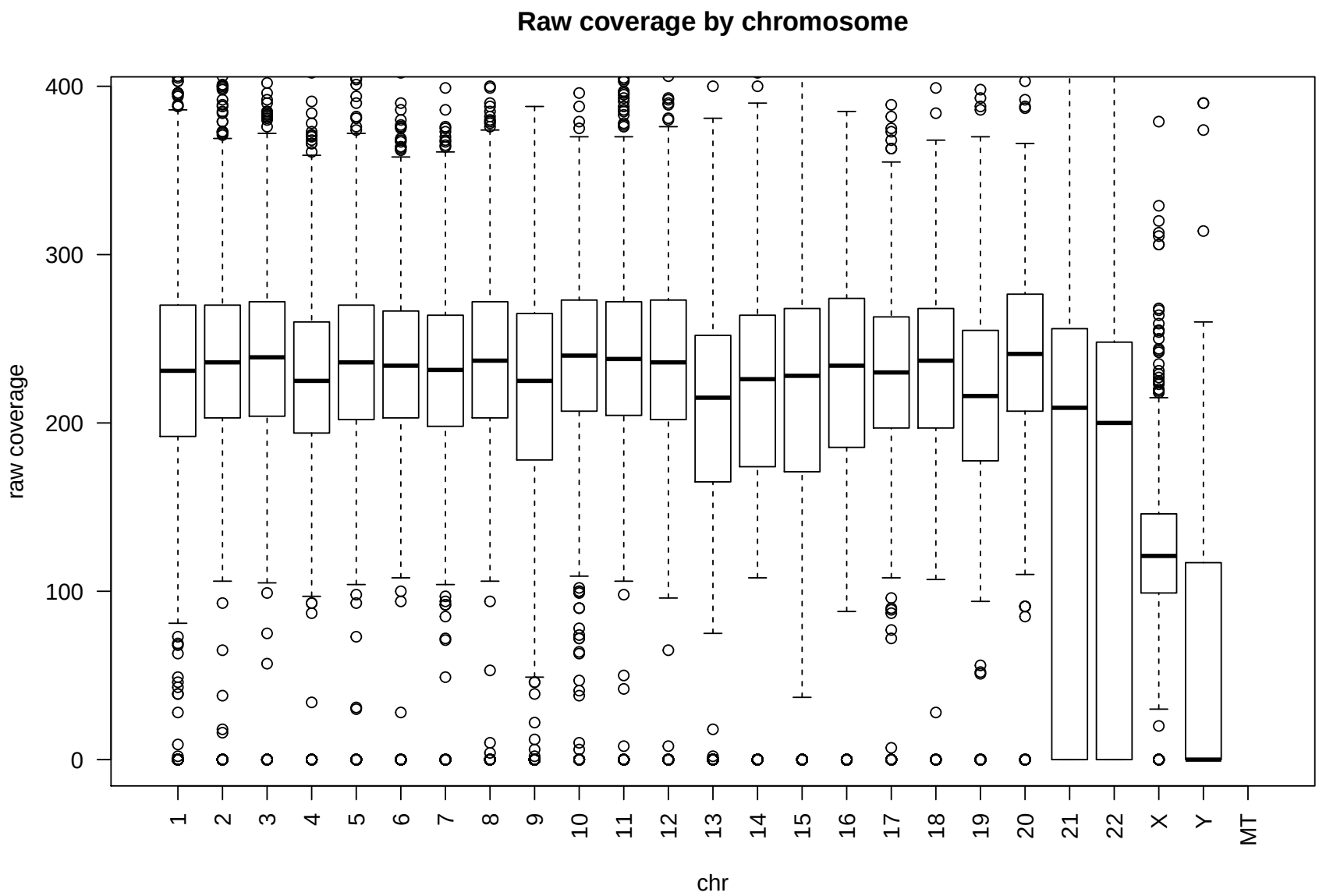
- index the bam file for speed

samtools index 2016-11-12-312685483745.bam

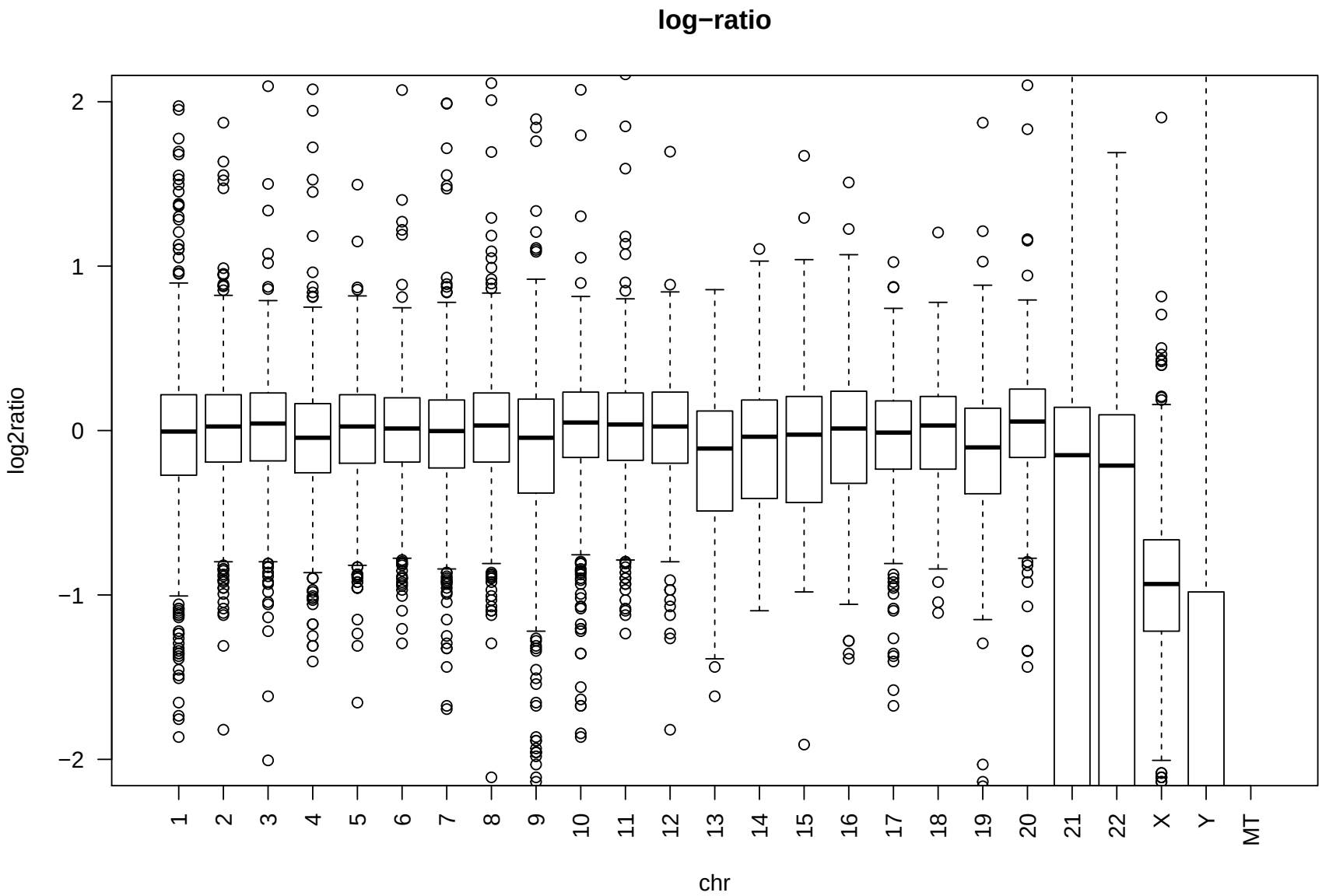
- compute coverage

bedtools coverage -a chr.bed -b 2016-11-12-312685483745.bam > earl.txt

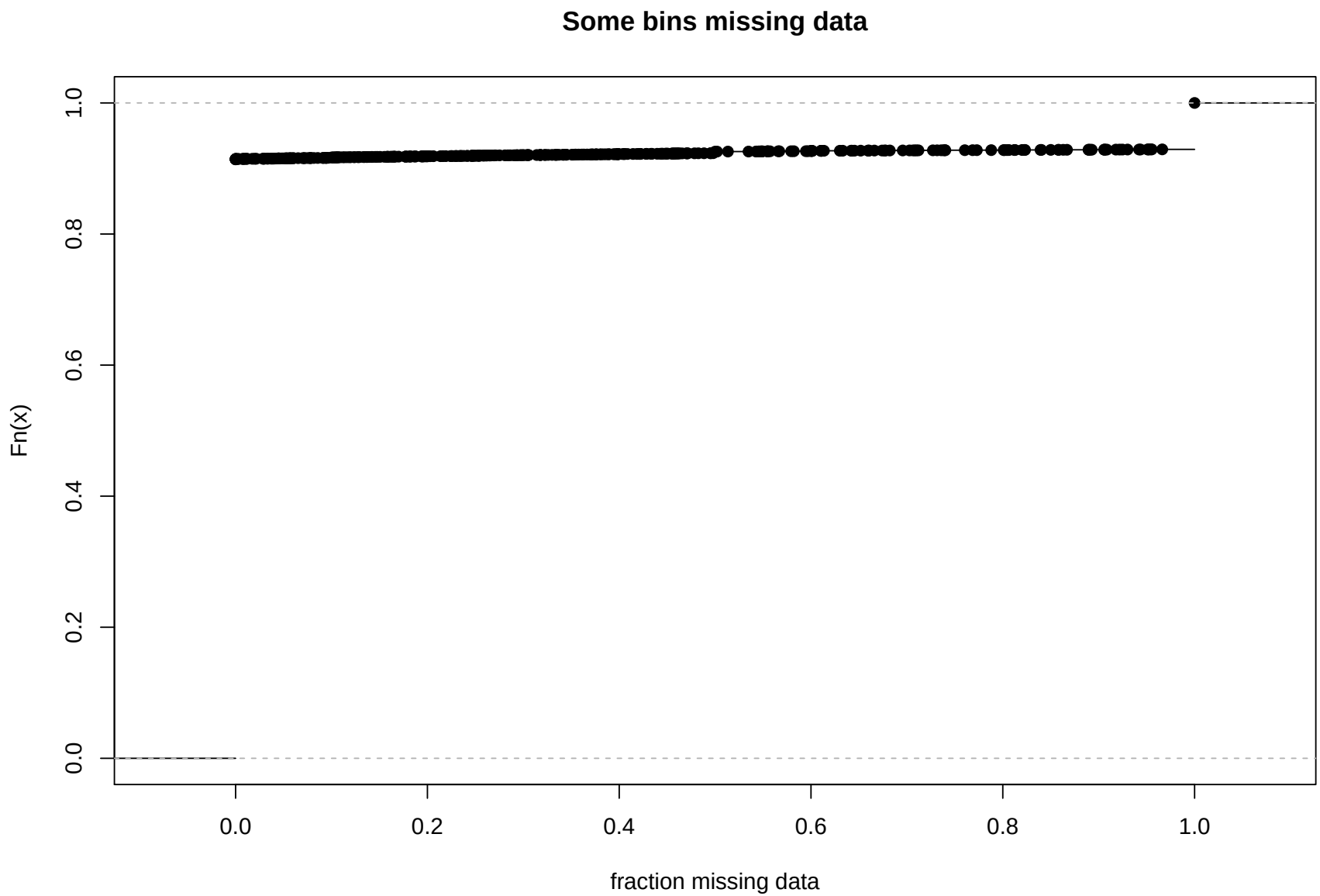
Look at the raw data



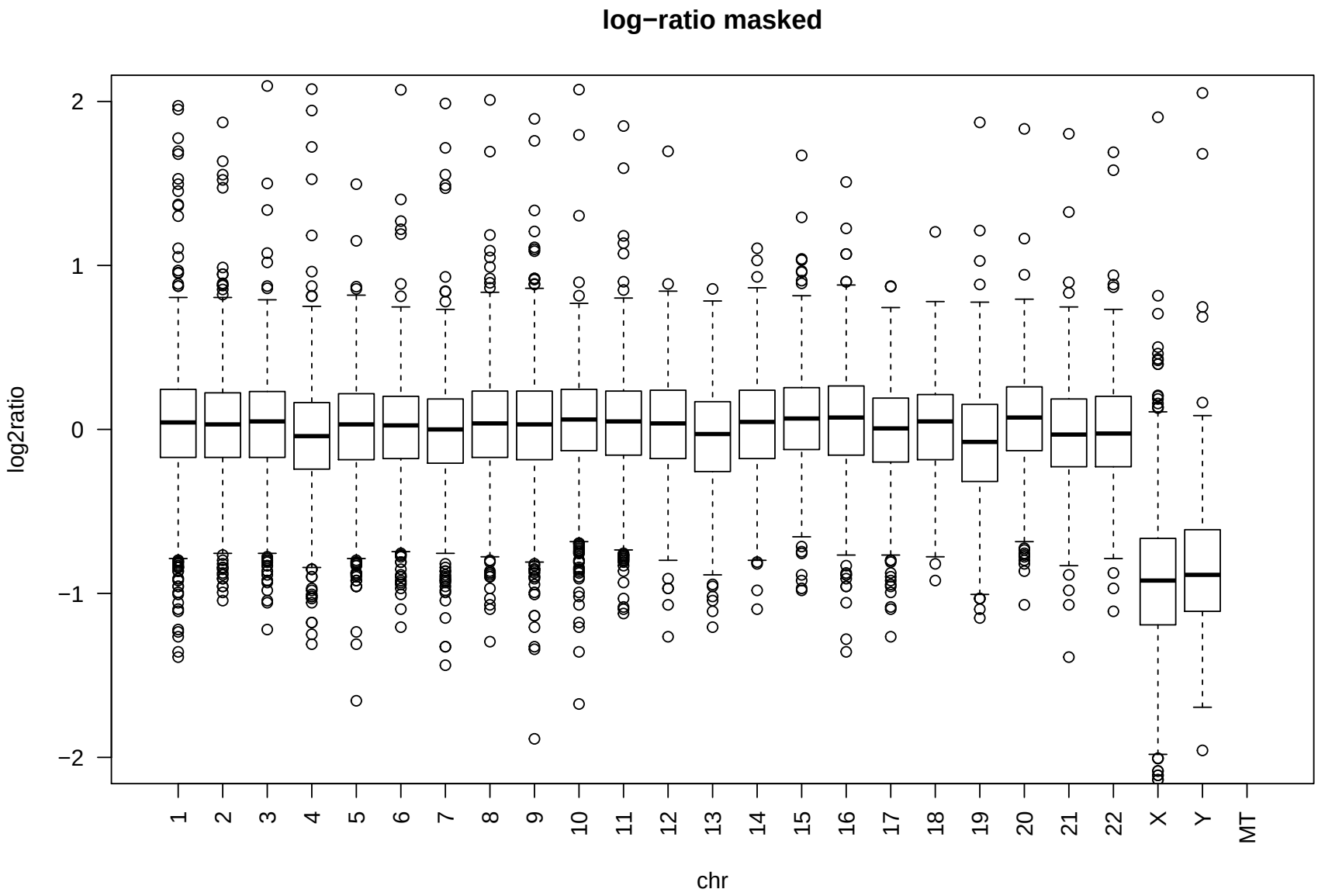
Convert to log2 ratio



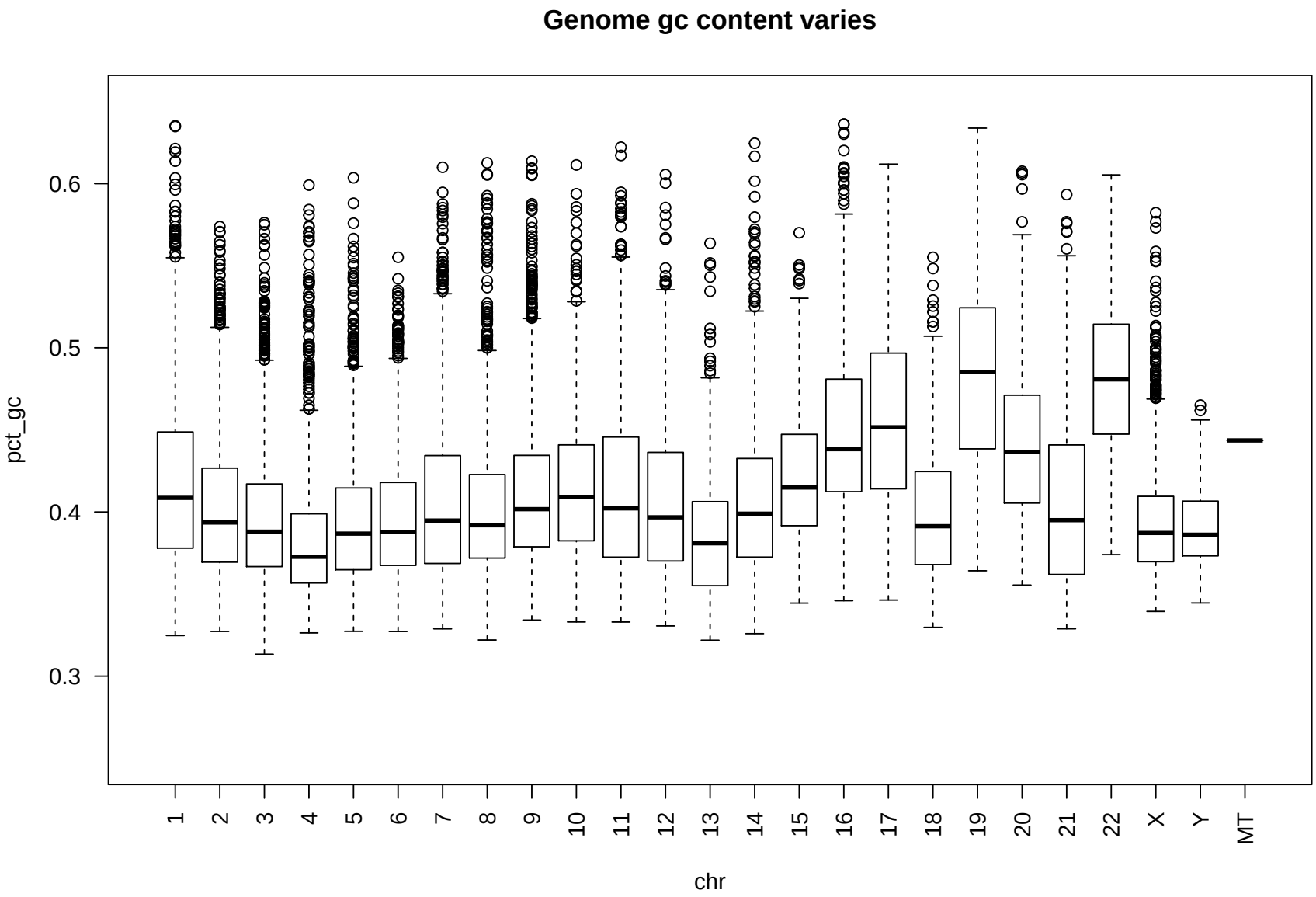
Human genome not complete everywhere



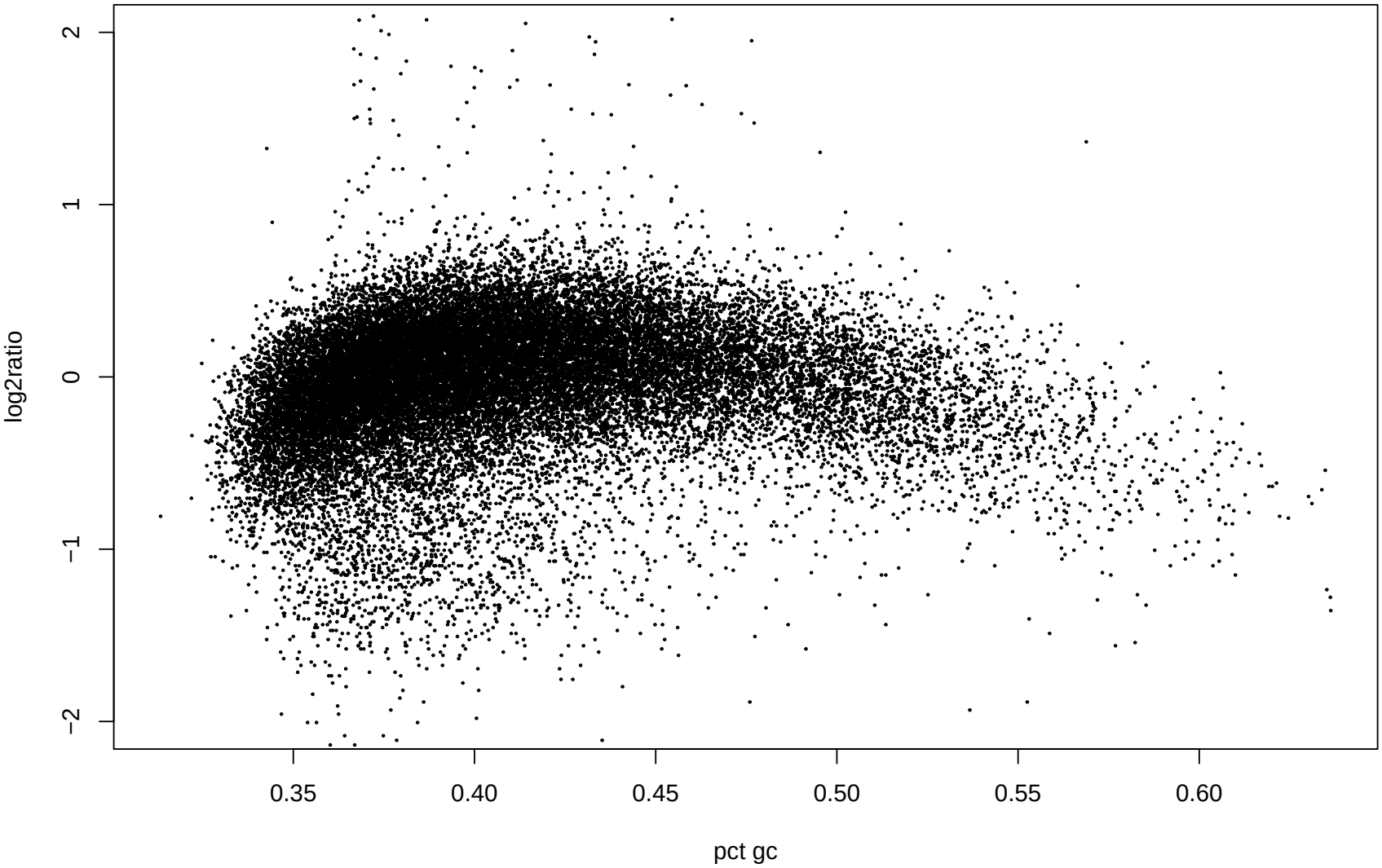
Mask out difficult sections



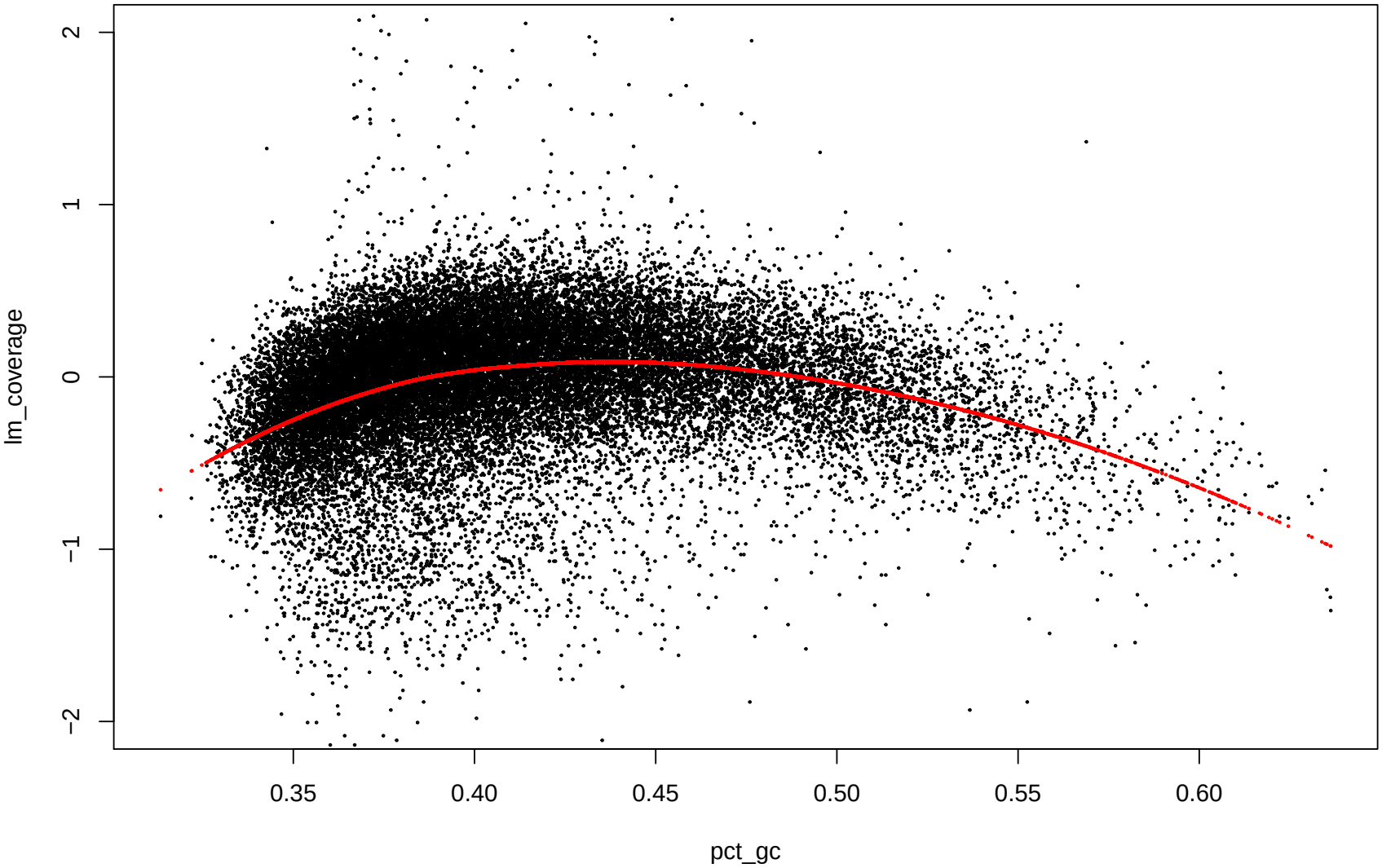
Systematic GC content changes by chromosome



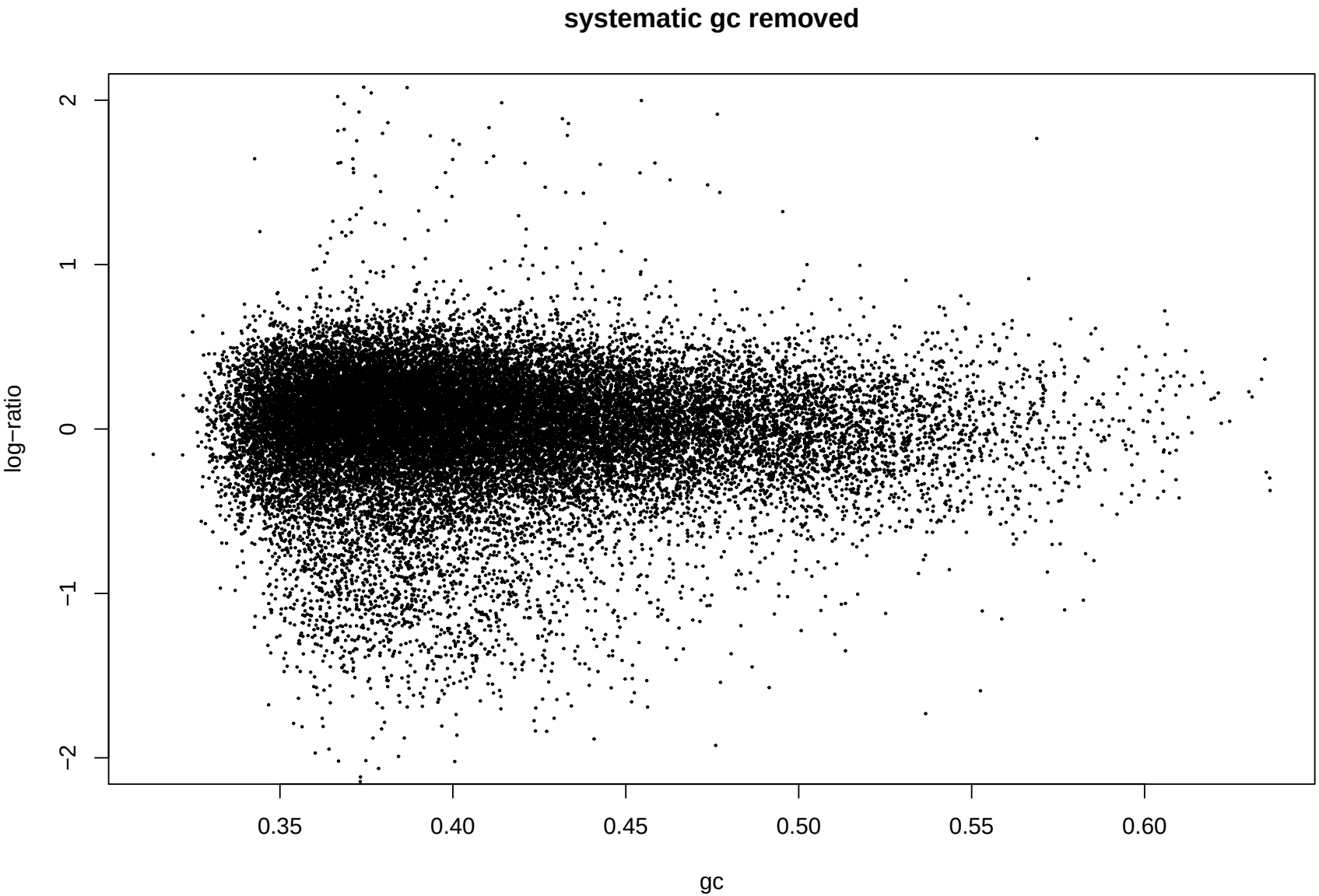
GC content nonlinear effect per bin



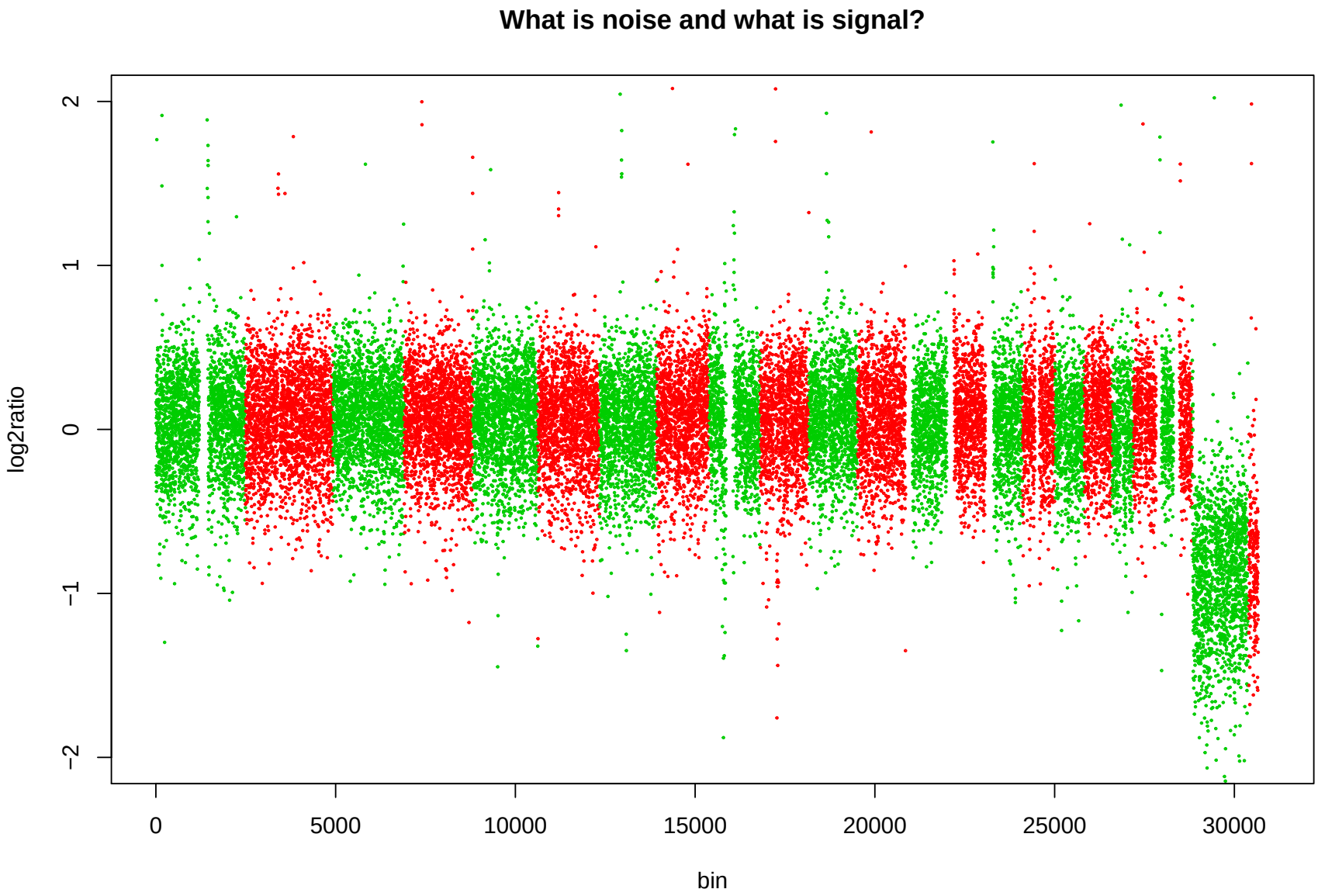
Fit loess model



Remove fitted effect



Corrected coverage



Circular Binary Segmentation

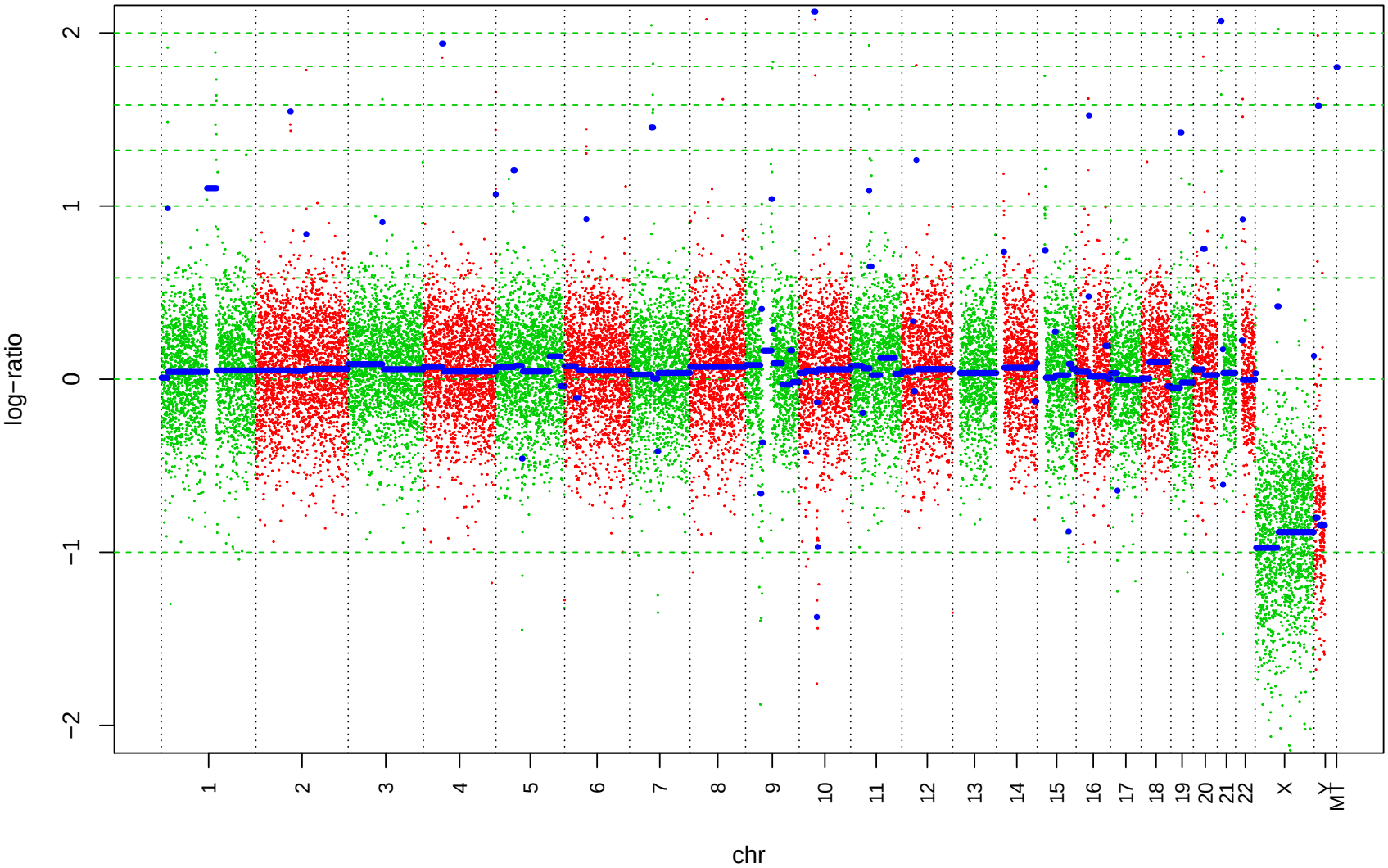
- Copy number should come in discrete segments
  - Mean of segment should be the same
  - Split chromosomes into segments of different means
- Circular Binary Segmentation Algorithm
  - Recursively split segments in two
  - Mean different between sections
  - Larger than by chance variation
  - Treat segments as ‘circles’

Bioconductor

- CBS is a standard algorithm
- Bioconductor toolkit
  - biocLite("DNAcopy")

Segmentation

```
## Analyzing: Sample.1
```



Fun

- All chromosomes present at typical counts
  - Maybe some little bits need further investigation
- Tools
  - Command line
  - R + Bioconductor

Also my day job

- GRAIL
  - Detect Cancer Early When It Can Be Cured
  - Save Lives
- GRAIL is looking for:
  - machine learning
  - reproducible research
  - bioinformatics
- I will pay money for corporate format slide templates
  - beamer, etc.
  - Anyone want a contract job?