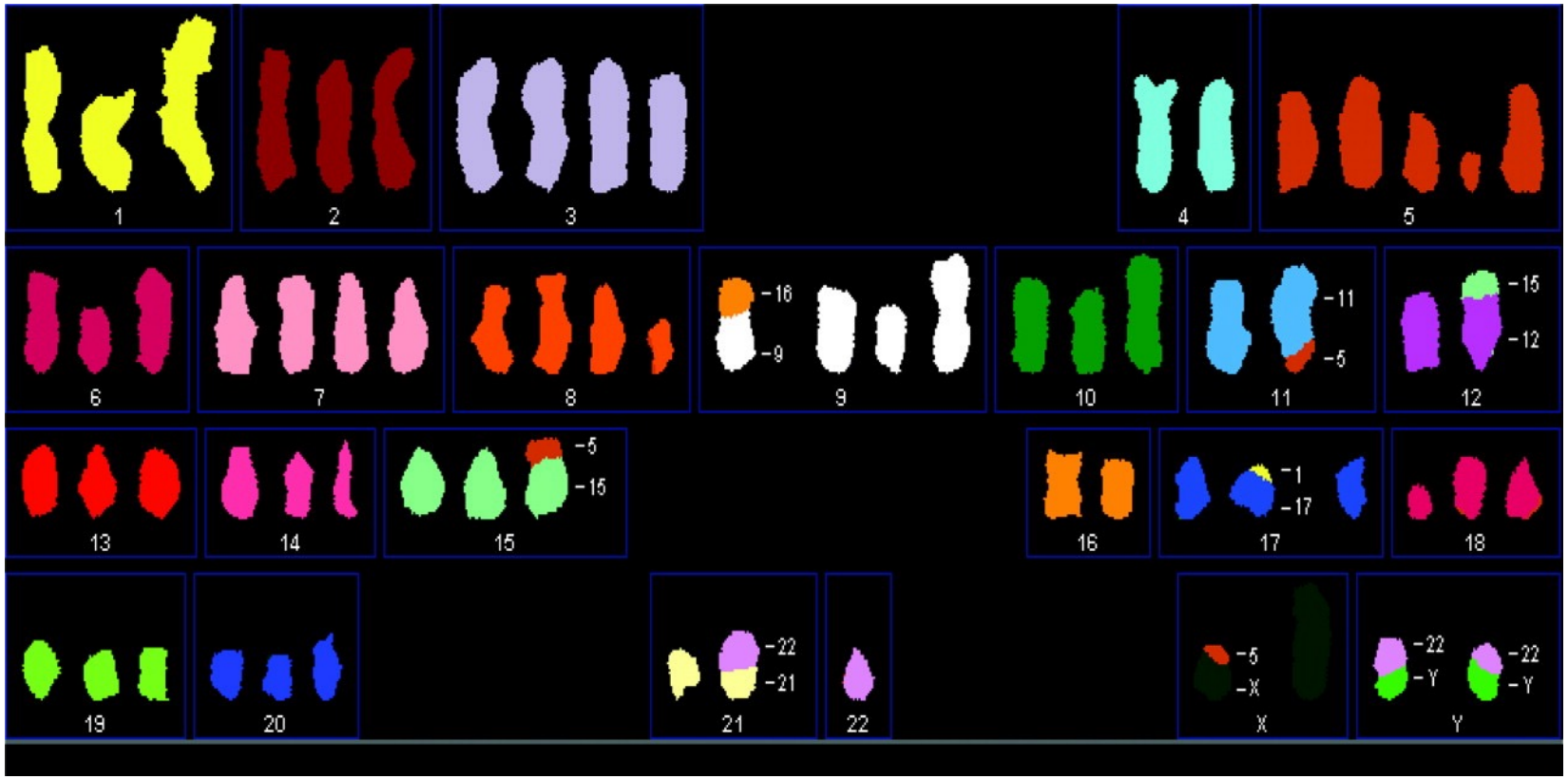
Copy number analysis

Spectral karyotyping data from a representative resistant cell line (UMSCC 81), which is highly aneuploid (chromosome number 69) and shows pronounced numerical and structural chromosomal changes.



Jan Akervall et al. Clin Cancer Res 2004;10:8204-8213

Clinical Cancer Research
AACR American Association for Cancer Research

Copy number analysis

- Amplifications (AR, EGFR, …)
- Homozygous deletions (TP53, PTEN, …)
- Deletions (TSG second hit, translocations, ERG fusion, …)
  Including germline
- Loss of heterozygosity (TSG second hit)
- Duplications (BRAF fusion, …)
- Tumor ploidy
- Tumor cell or DNA fraction
- Other QC

Copy number analysis

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing

Eric Talevich[1,2,3], **A. Hunter Shain**[1,2,3], **Thomas Botton**[1,2,3], **Boris C. Bastian**[1,2,3]*

1 Department of Dermatology, University of California, San Francisco, San Francisco, California, United States of America, **2** Department of Pathology, University of California, San Francisco, San Francisco, California, United States of America, **3** Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, California, United States of America

* boris.bastian@ucsf.edu

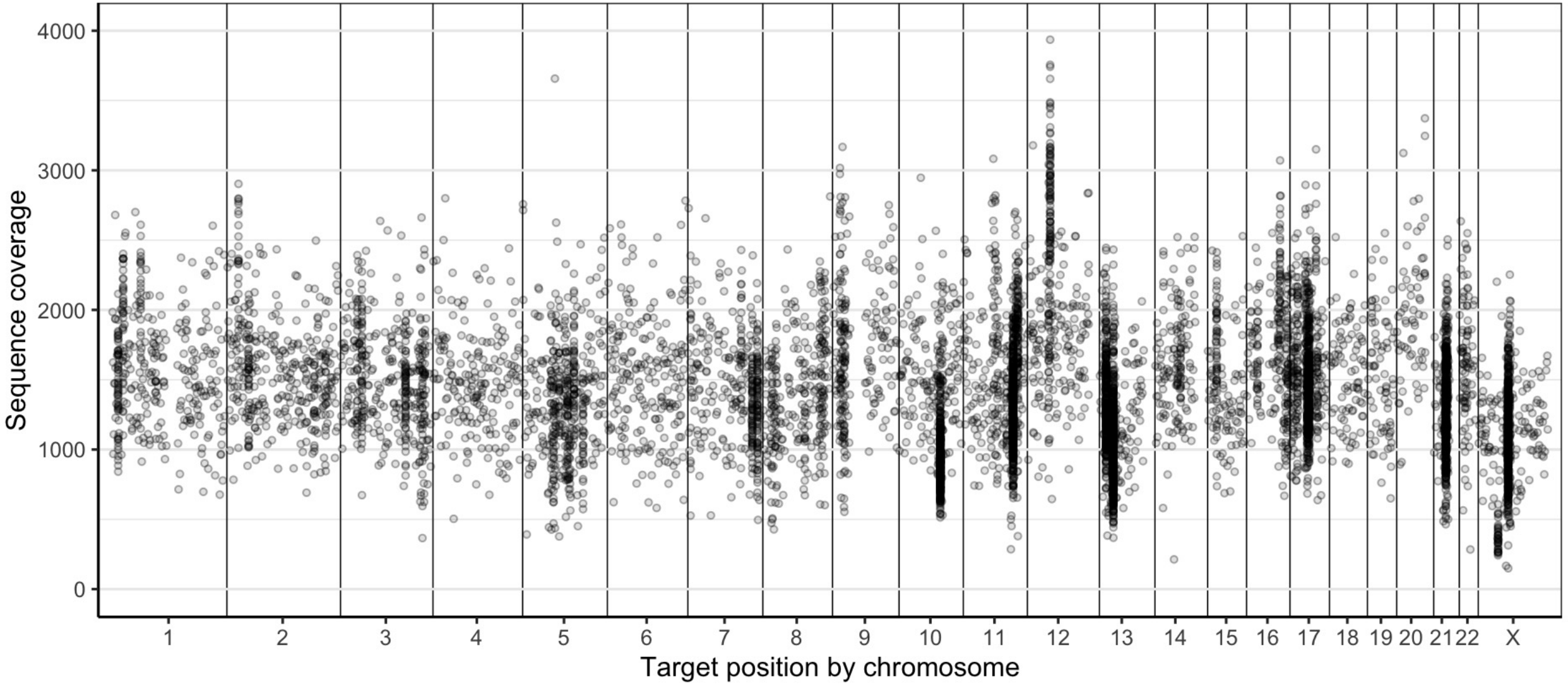
OPEN ACCESS

Citation: Talevich E, Shain AH, Botton T, Bastian BC (2016) CNVkit: Genome-Wide Copy Number

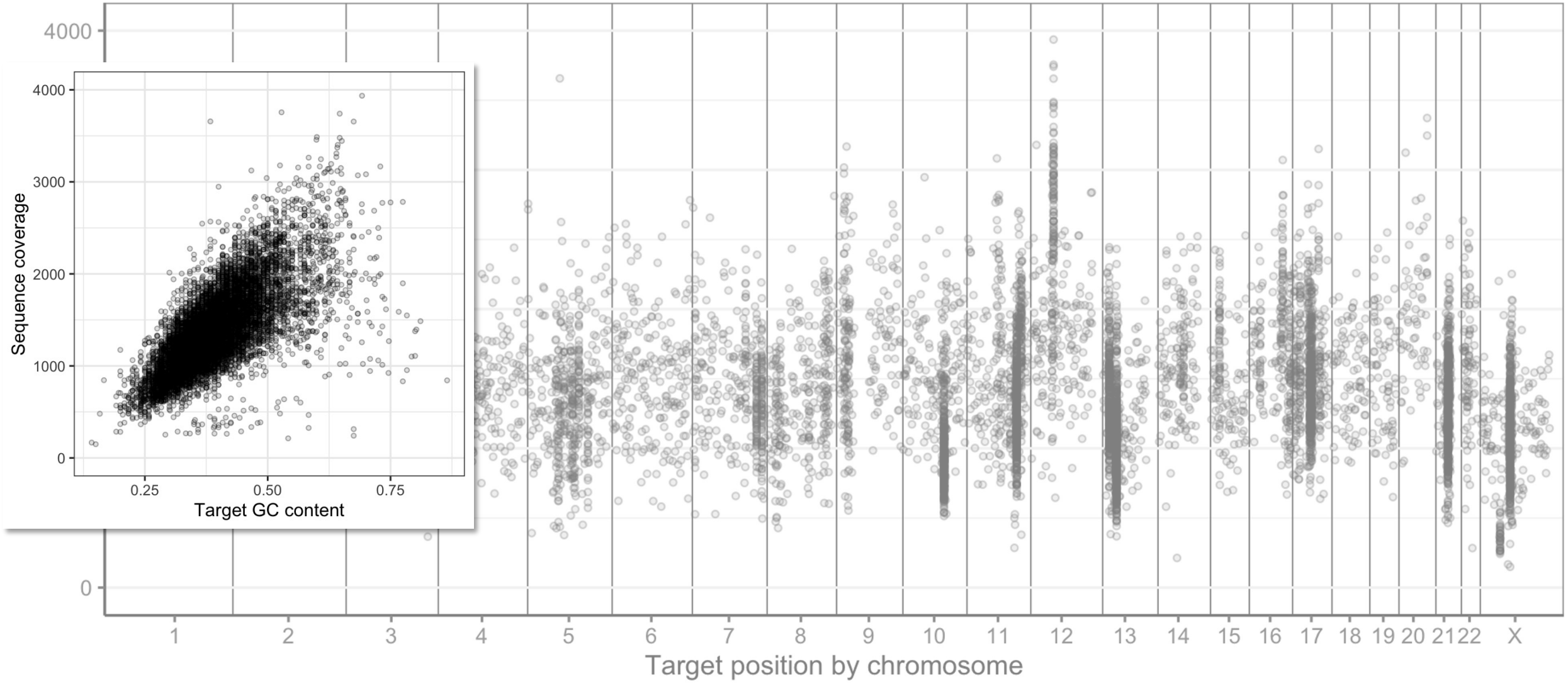
Abstract

Germline copy number variants (CNVs) and somatic copy number alterations (SCNAs) are of significant importance in syndromic conditions and cancer. Massively parallel sequencing is increasingly used to infer copy number information from variations in the read depth in sequencing data. However, this approach has limitations in the case of targeted re-

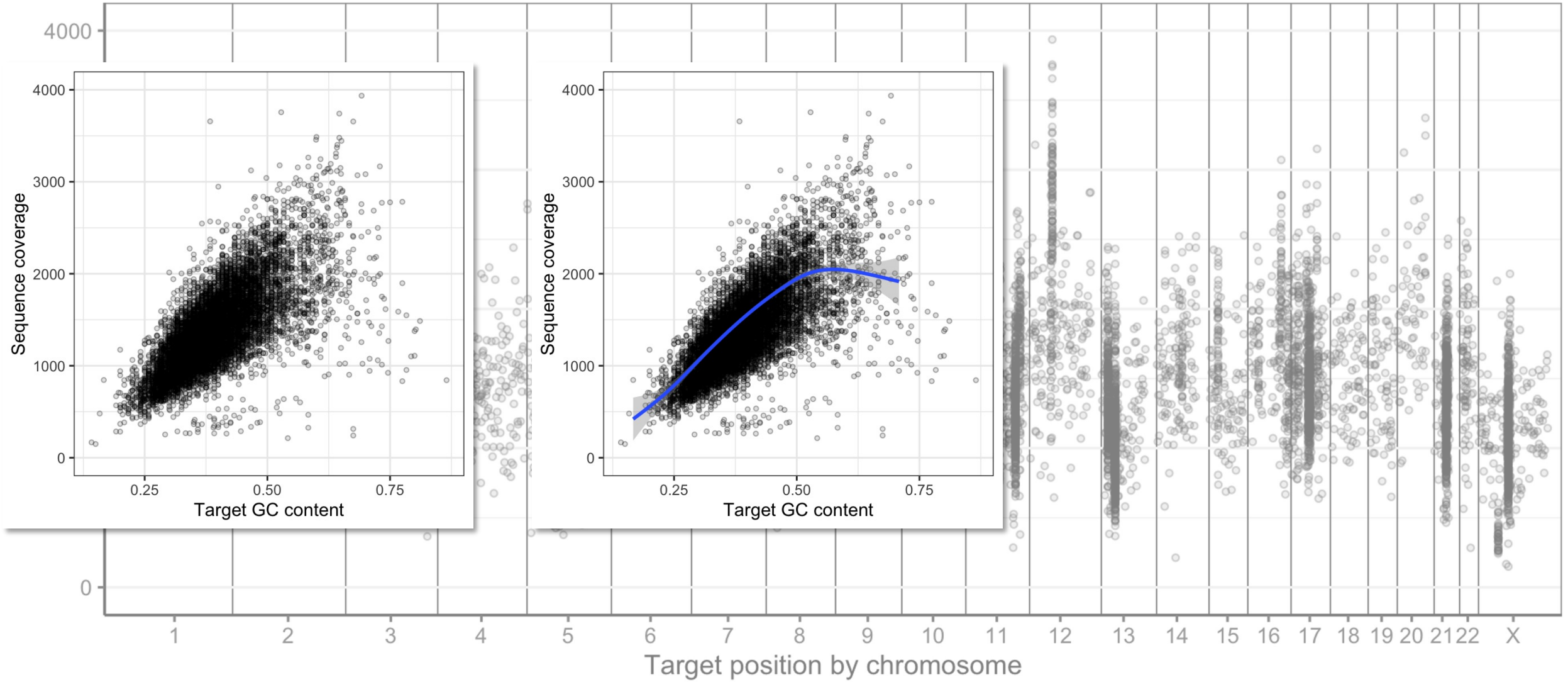
Coverage as DNA abundance



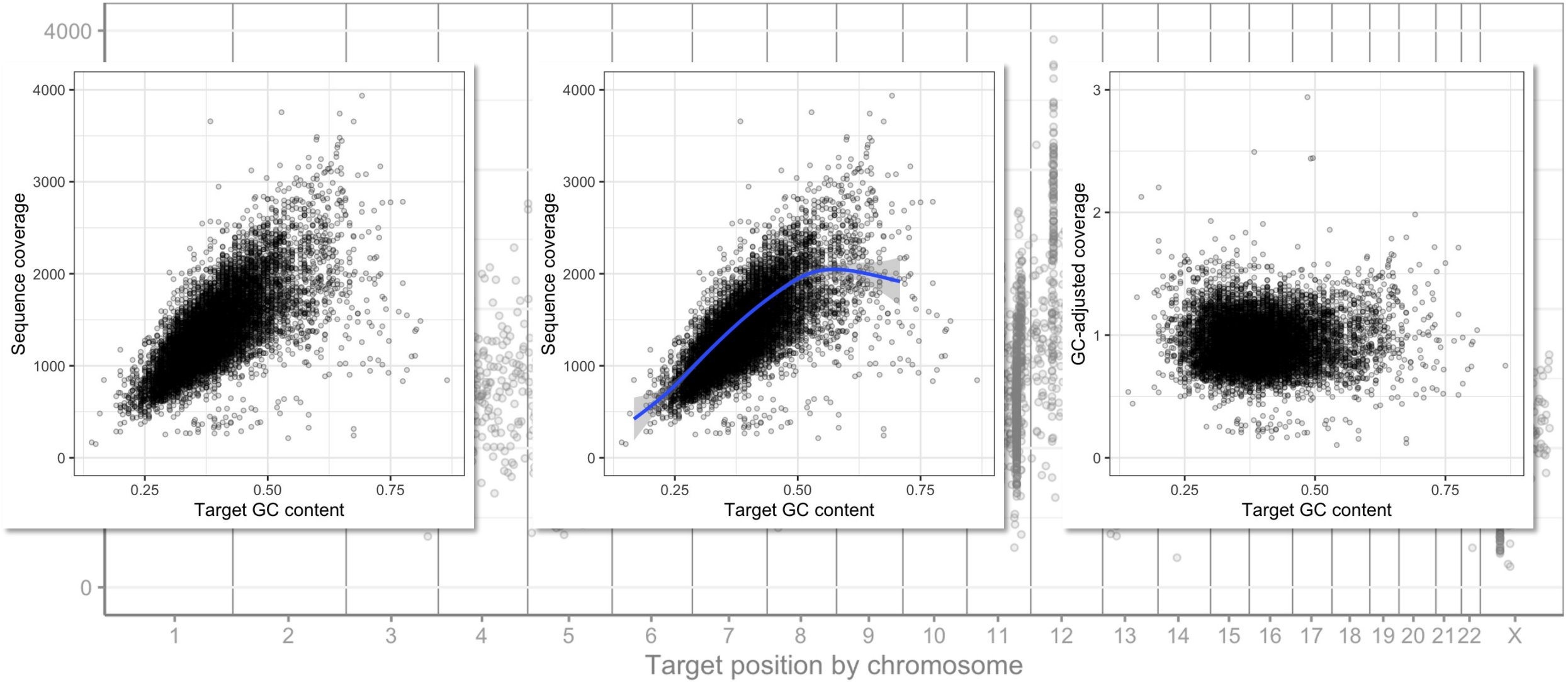
GC content bias correction



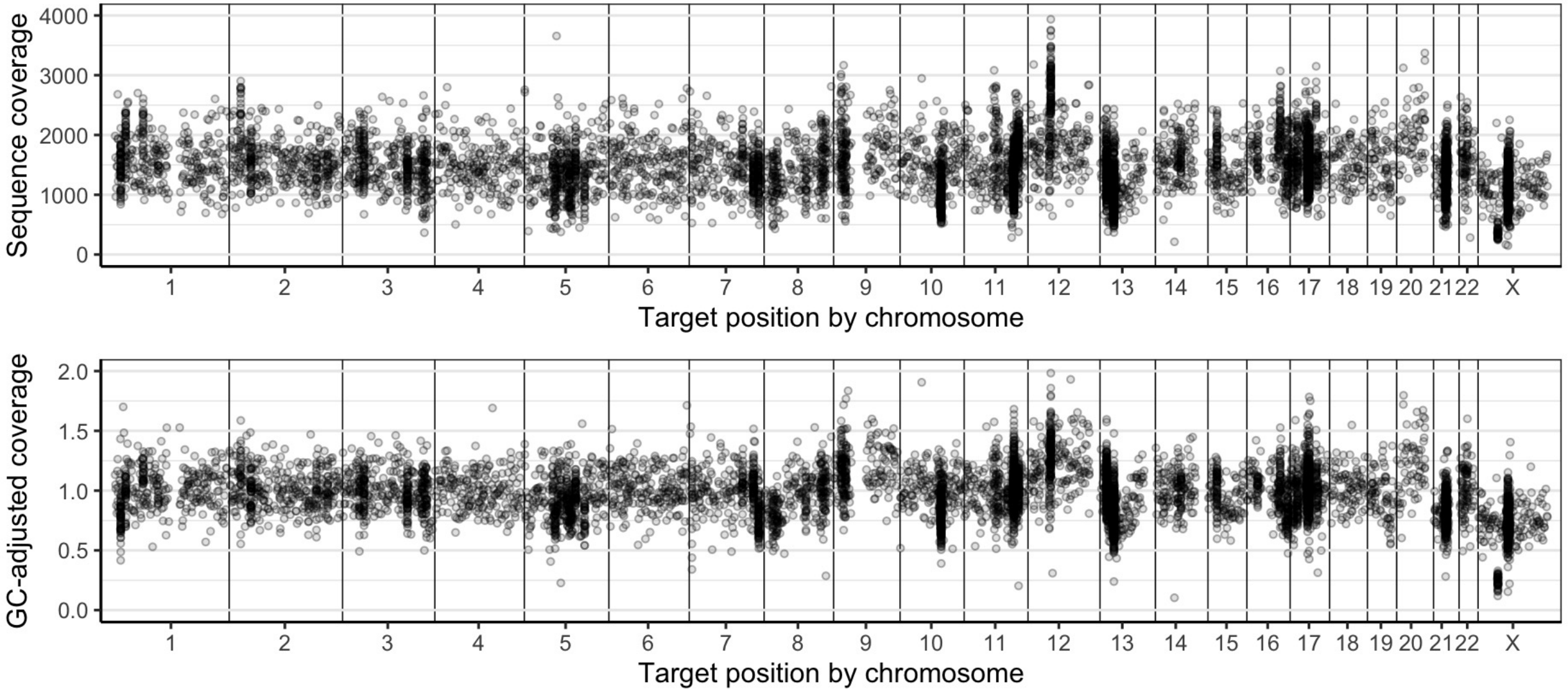
GC content bias correction



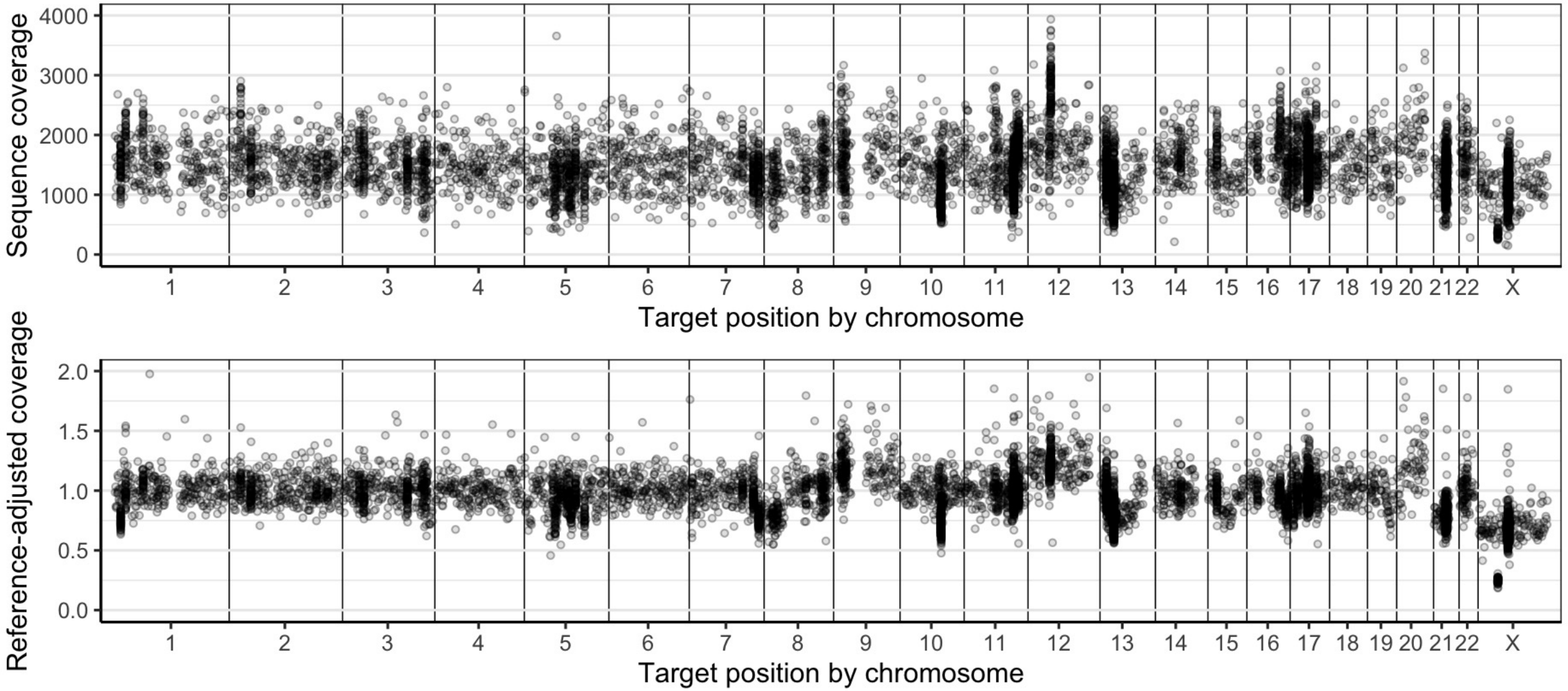
GC content bias correction



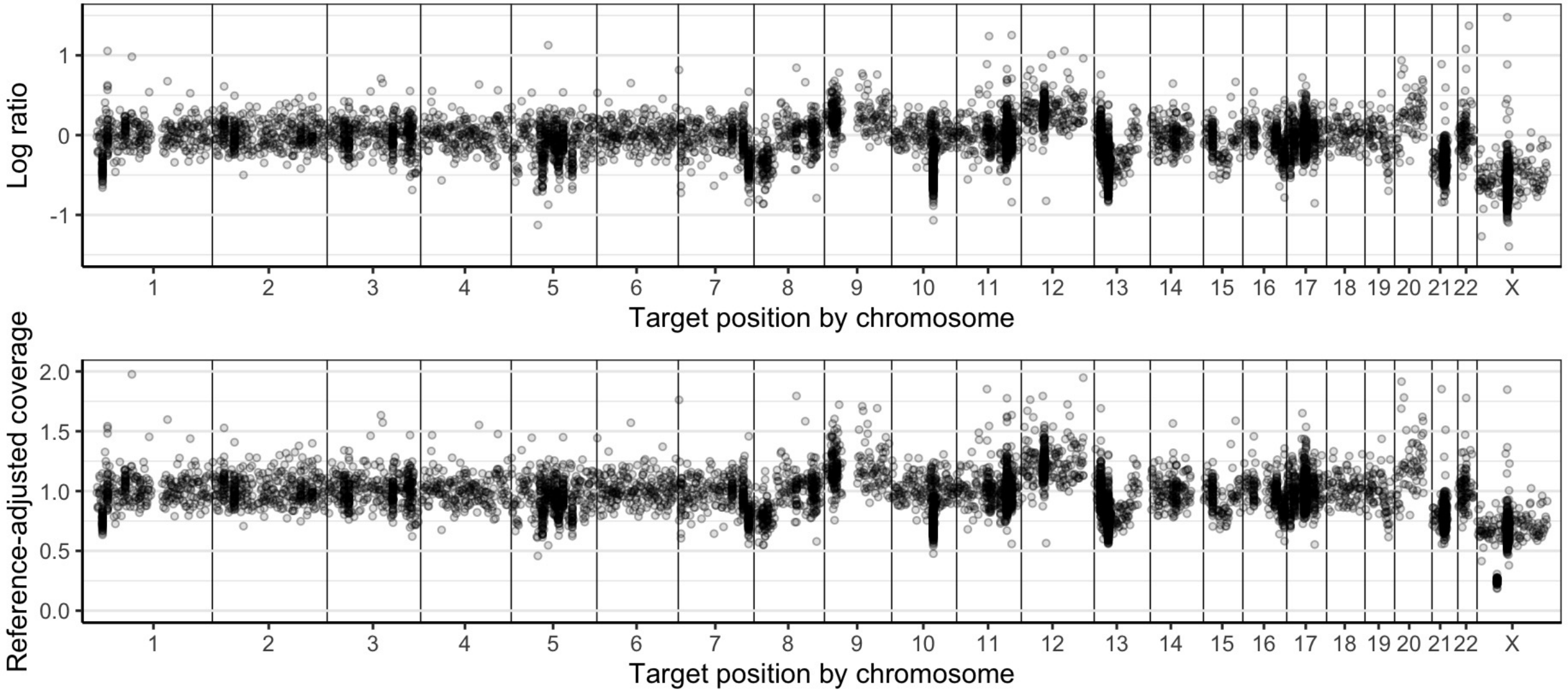
GC content bias correction



Assay-specific bias correction



Log ratio or just ”ratio”

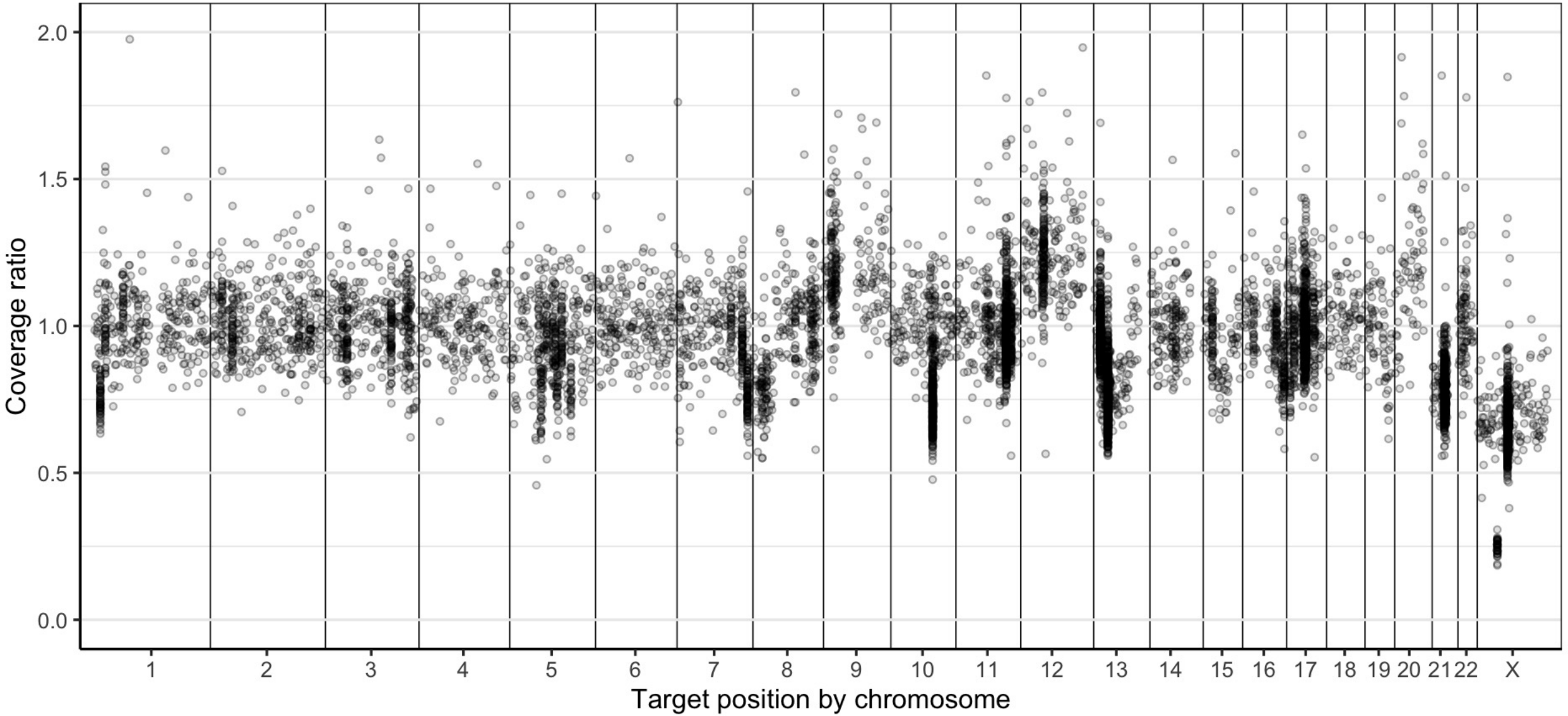


Measuring DNA abundance

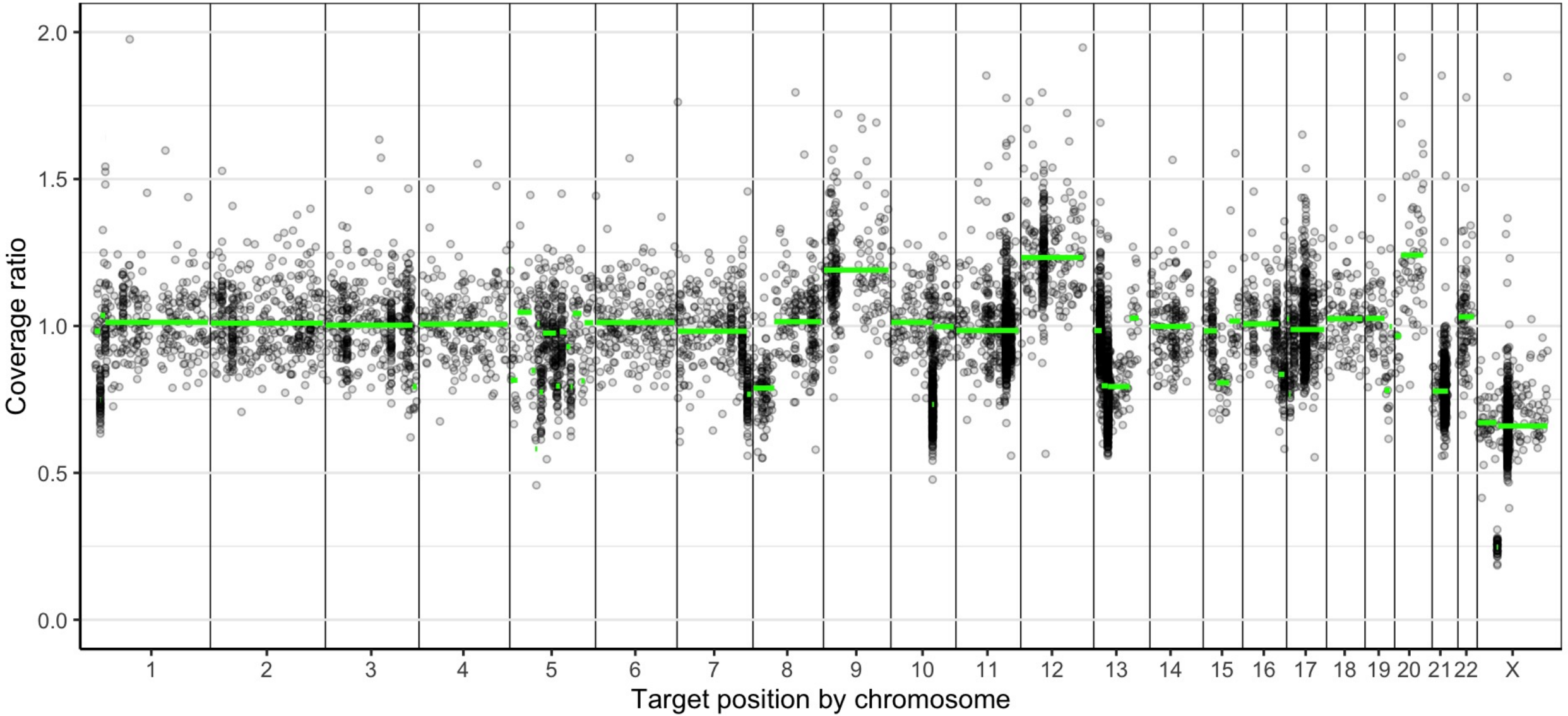
- Quantify coverage in bins along genome
- Correct for sample-specific bias e.g. GC content
    * linear or log space: log ratio = log2(coverage ratio)
- Correct for assay-specific bias
    * pool of normals usually better than matched normal

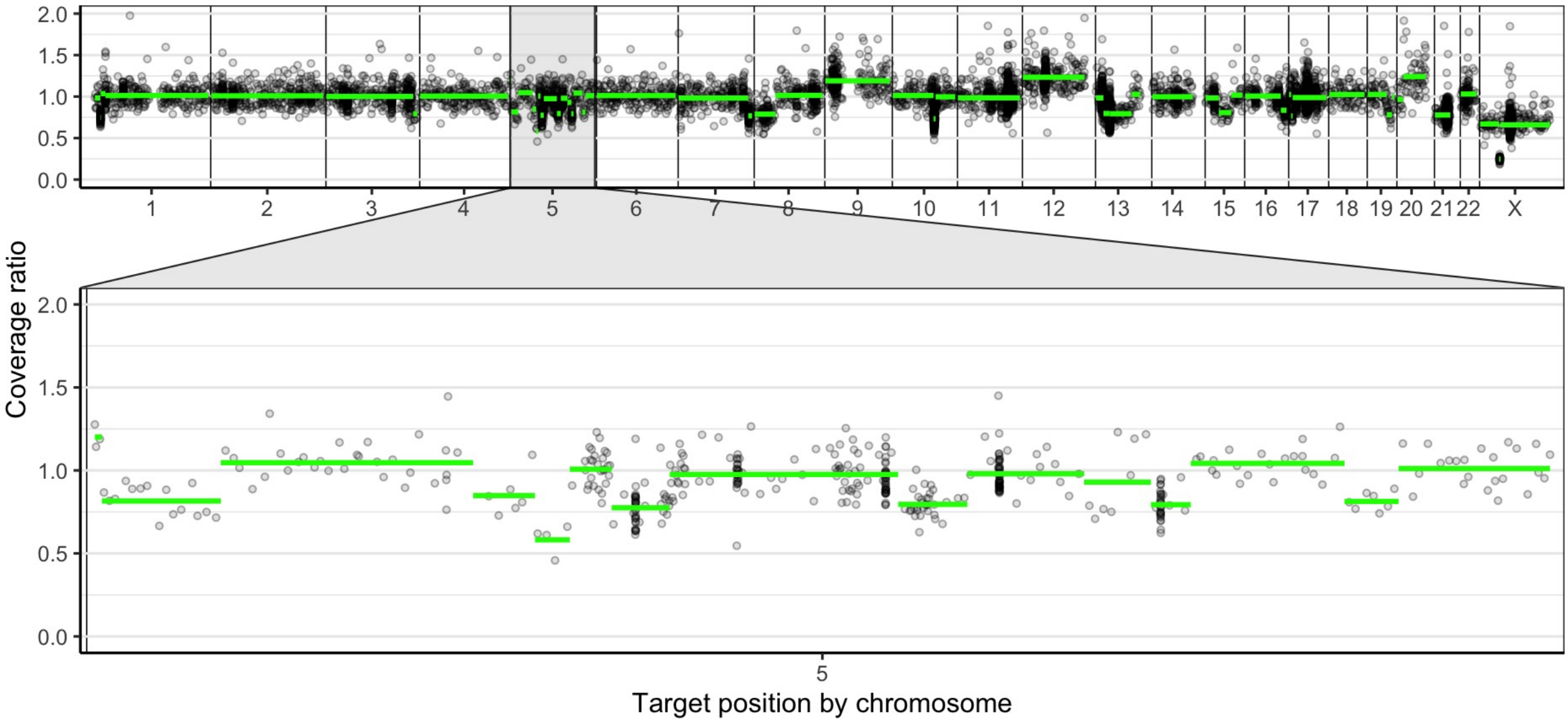
    Run and analyze matched normal separately!

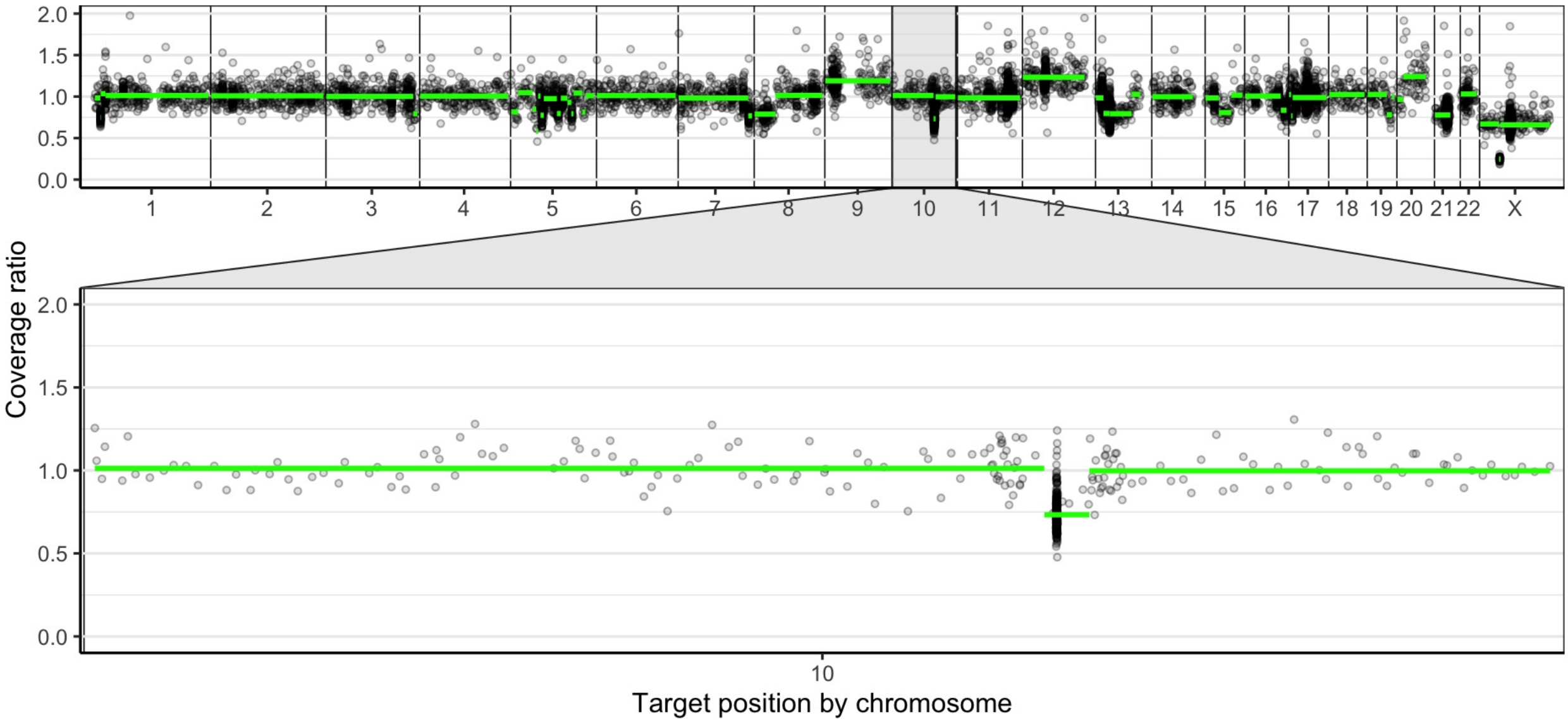
Segmentation

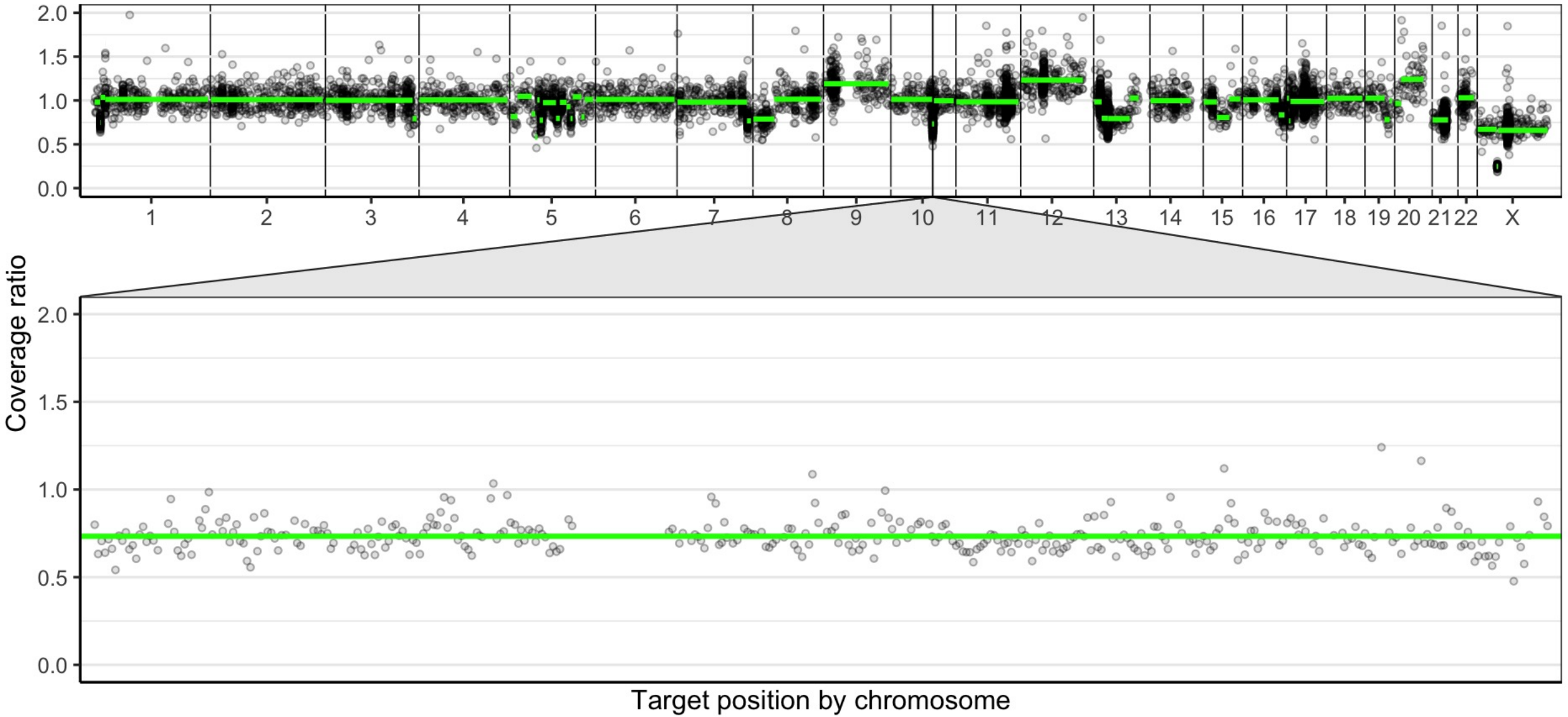


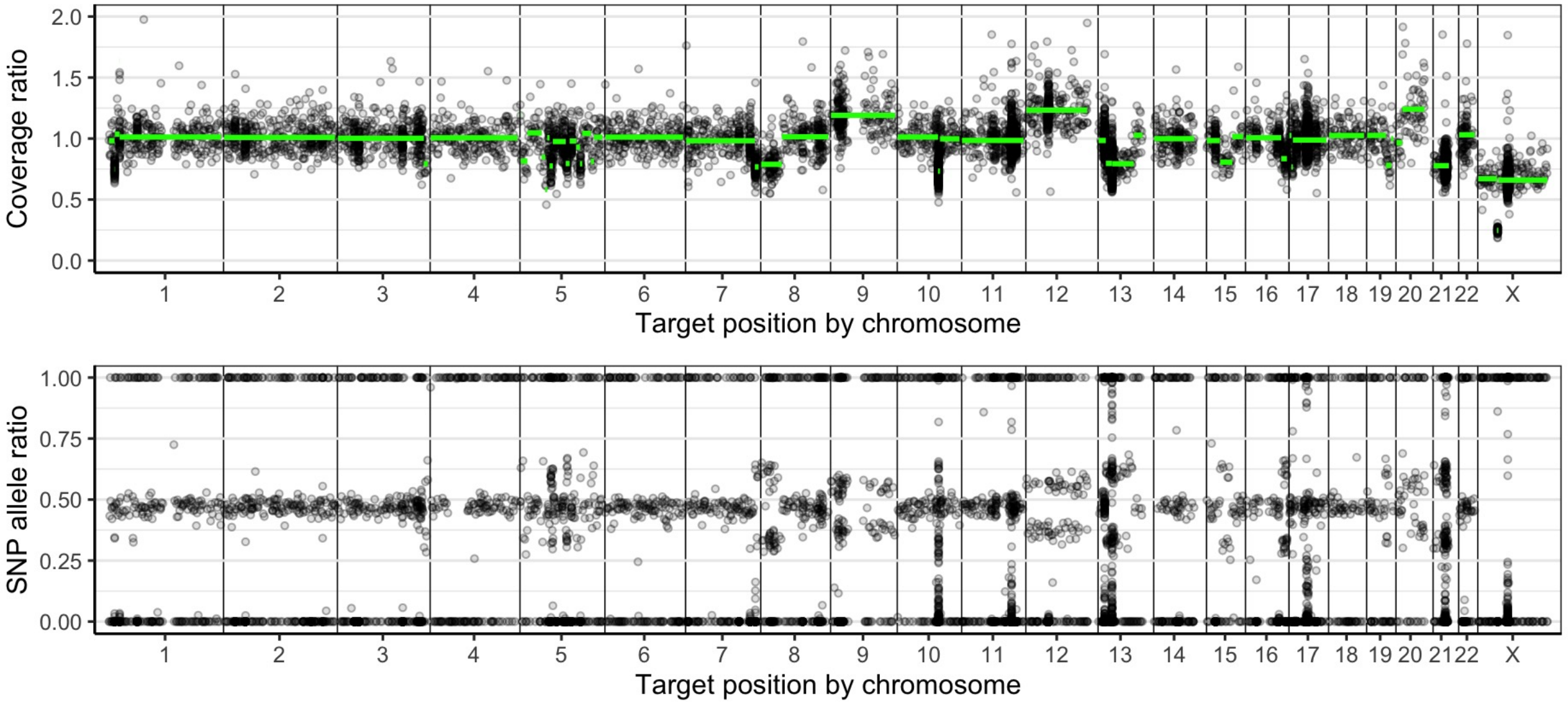
Segmentation

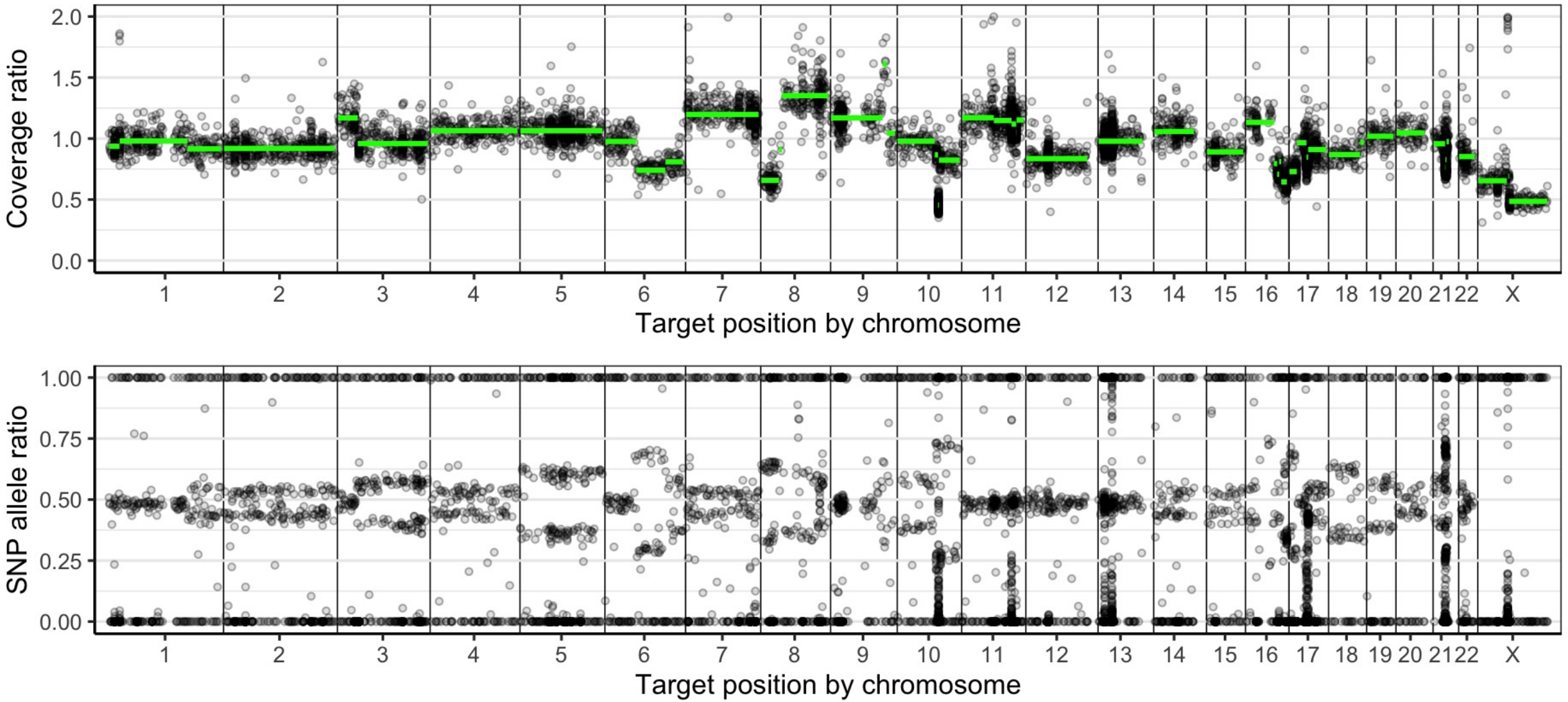



2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Coverage ratio
5
Target position by chromosome

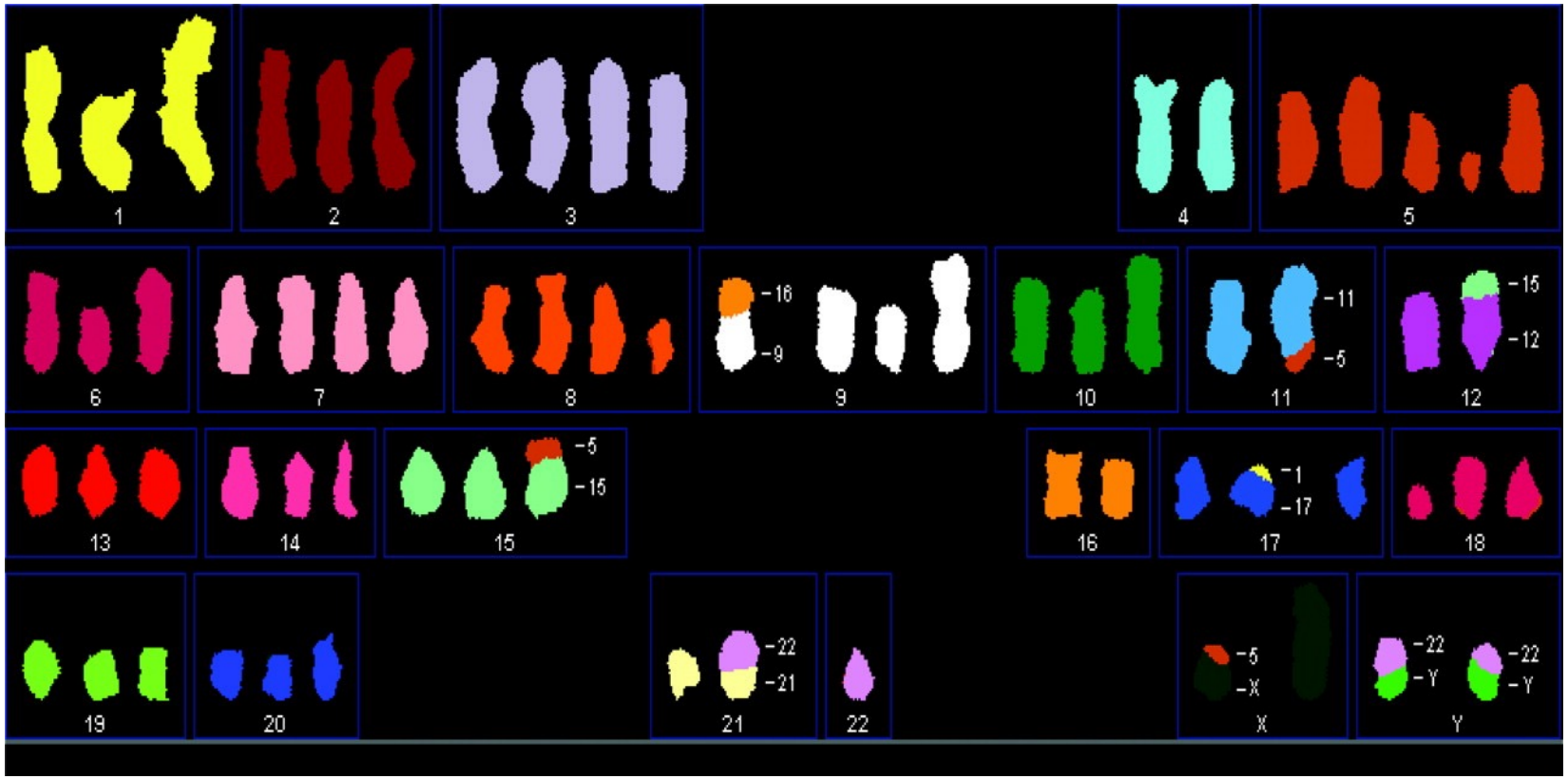

2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Coverage ratio
2.0
1.5
1.0
0.5
0.0
10
Target position by chromosome


2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
Coverage ratio
Target position by chromosome


Coverage ratio
2.0
1.5
1.0
0.5
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Target position by chromosome
SNP allele ratio
1.00
0.75
0.50
0.25
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Target position by chromosome


Coverage ratio
2.0
1.5
1.0
0.5
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Target position by chromosome
SNP allele ratio
1.00
0.75
0.50
0.25
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
Target position by chromosome

Spectral karyotyping data from a representative resistant cell line (UMSCC 81), which is highly aneuploid (chromosome number 69) and shows pronounced numerical and structural chromosomal changes.



Jan Akervall et al. Clin Cancer Res 2004;10:8204-8213

Clinical Cancer Research
AACR American Association for Cancer Research

Estimate ploidy and purity

- Cancer genomes are frequently not (near) diploid.
- Aberrant ploidy may be an important observation in itself
- High ploidy may lead to false positive calls of loss
- DNA from tumor (or ctDNA) contain non-tumor DNA
- Low purity makes it more difficult to distinguish variants

PureCN

Riester *et al. Source Code for Biology and Medicine* (2016) 11:13
DOI 10.1186/s13029-016-0060-z

Source Code for Biology and Medicine

SOFTWARE Open Access

PureCN: copy number calling and SNV classification using targeted short read sequencing

Markus Riester*, Angad P. Singh, A. Rose Brannon, Kun Yu, Catarina D. Campbell, Derek Y. Chiang and Michael P. Morrissey

Abstract

Background: Matched sequencing of both tumor and normal tissue is routinely used to classify variants of uncertain significance (VUS) into somatic vs. germline. However, assays used in molecular diagnostics focus on known somatic alterations in cancer genes and often only sequence tumors. Therefore, an algorithm that reliably classifies variants would be helpful for retrospective exploratory analyses. Contamination of tumor samples with normal cells results in differences in expected allelic fractions of germline and somatic variants, which can be exploited to accurately infer genotypes after adjusting for local copy number. However, existing algorithms for determining tumor purity, ploidy and copy number are not designed for unmatched short read sequencing data.

Results: We describe a methodology and corresponding open source software for estimating tumor purity, copy number, loss of heterozygosity (LOH), and contamination, and for classification of single nucleotide variants (SNVs) by somatic status and clonality. This R package, PureCN, is optimized for targeted short read sequencing data, integrates well with standard somatic variant detection pipelines, and has support for matched and unmatched tumor samples. Accuracy is demonstrated on simulated data and on real whole exome sequencing data.

Conclusions: Our algorithm provides accurate estimates of tumor purity and ploidy, even if matched normal samples are not available. This in turn allows accurate classification of SNVs. The software is provided as open source (Artistic License 2.0) R/Bioconductor package PureCN (http://bioconductor.org/packages/PureCN/).

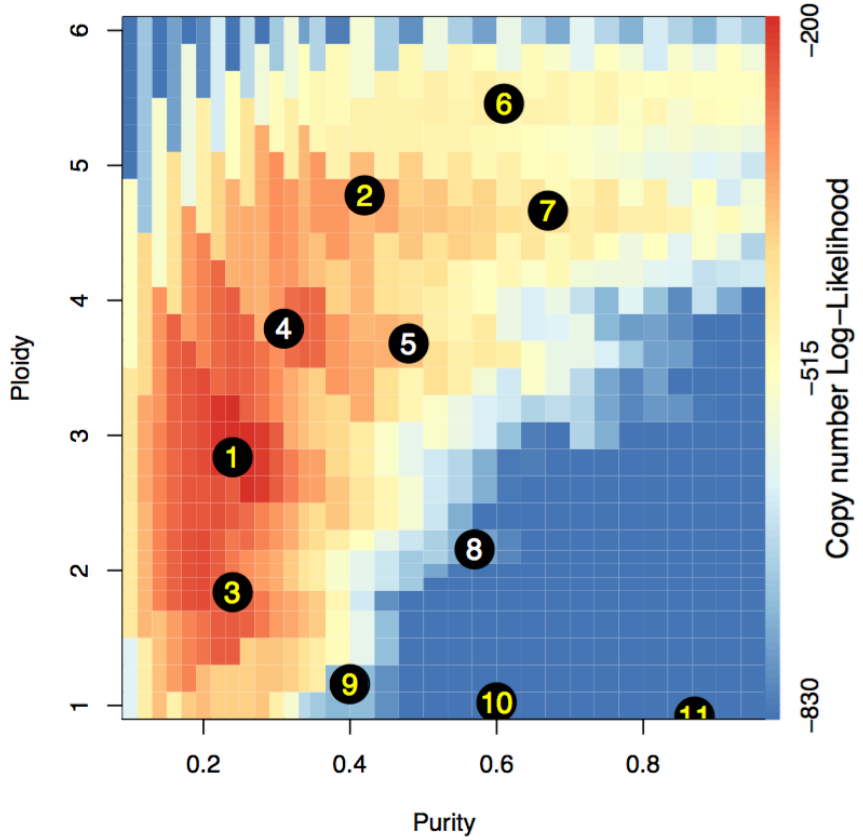
Keywords: Purity, Ploidy, Heterogeneity, Whole exome sequencing, Hybrid capture, Copy number, Loss of heterozygosity, Cell lines

PureCN

- PureCN suggests multiple (ranked) solutions for purity and ploidy

PureCN

- PureCN suggests multiple (ranked) solutions for purity and ploidy



PureCN

- PureCN suggests multiple (ranked) solutions for purity and ploidy
- The ”best” solution is not always the correct one ☹
- Usually successful at roughly ≥25% purity
- A better purity/ploidy fit is sometimes available to choose/curate
- Not always necessary to fix!

Copy number analysis

- Quantification of sequence read depth throughout the reference genome as a measure of DNA abundance in the sample
- Removal of sample-specific systematic noise using features such as GC content and mappability
- Removal of assay-specific systematic noise using normal (non-cancer) reference samples
- Segmentation - partitioning of the reference genome so that each segment can be assigned one copy number
- Copy number calling, assigning each segment some estimate of the number of copies per cell
- Combine normalized coverage with SNP allele frequencies to support copy number estimates