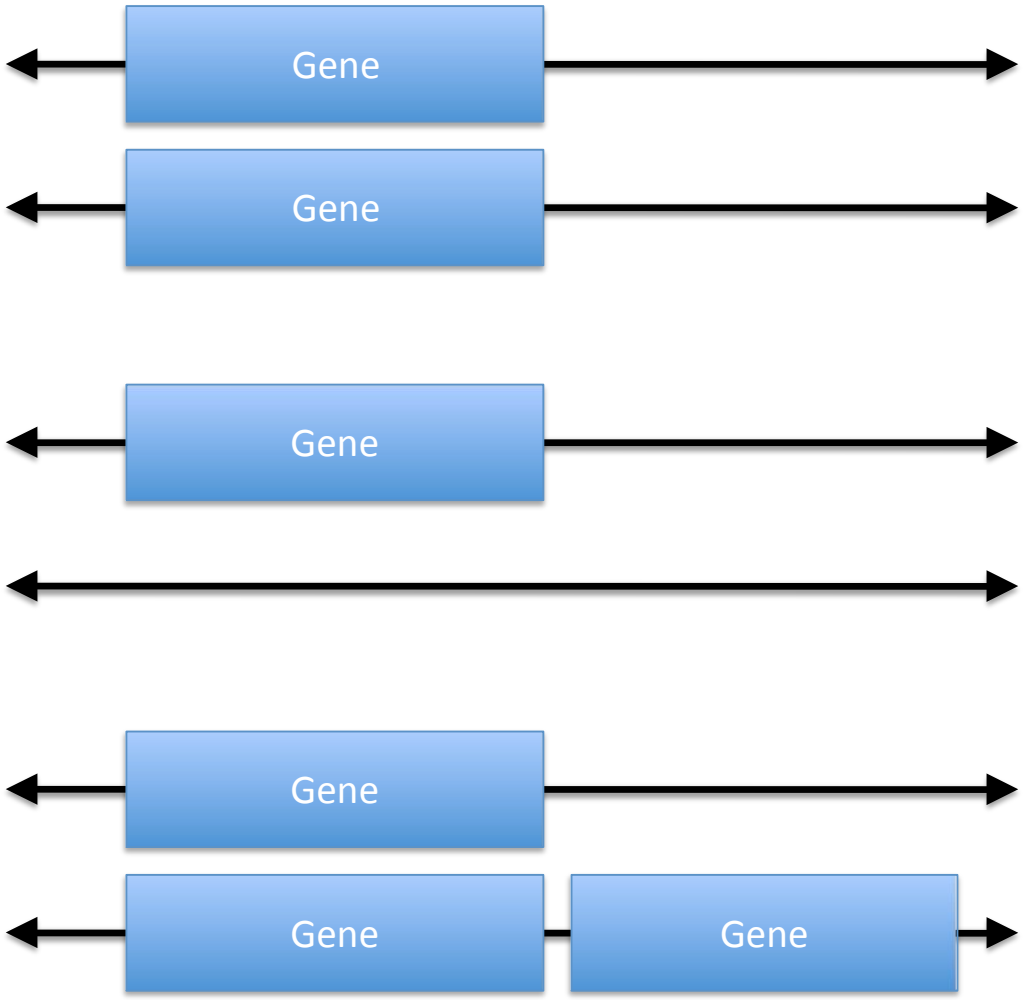
Discovery of copy number variations (CNVs) from exome read depth using XHMM (eXome-Hidden Markov Model)

Dr Reza Rafiee
Newcastle University
2017

Copy number variation (CNV)

- Reference:
- Deletion:
- Duplication:



Aim

- Use exome sequencing to accurately call copy number variation (CNV) at exon-level resolution
  - Based on “depth” of sequencing (number of items a portion of the genome is “read”) = read-depth

"Analysis-ready" exome BAMs using GATK

↓

Mean per-target ***coverage*** using GATK (4-6)

↓

Filter out "extreme" targets and samples, center targets (7-13)

↓

Run ***PCA***, remove top principal components (14-15)

↓

Filter out targets that are "extremely variable" (16)

↓

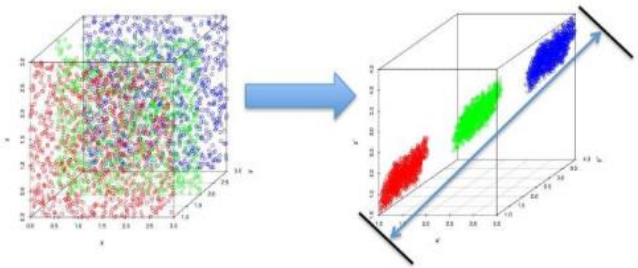
Calculate ***z-score*** for read depths of each sample (16)

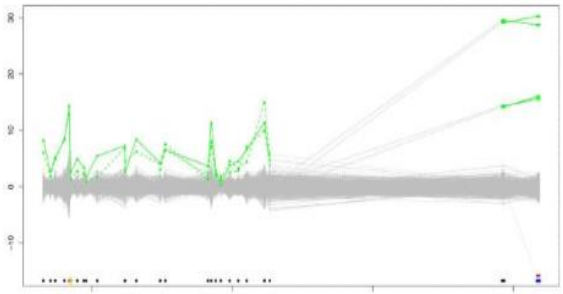
↓

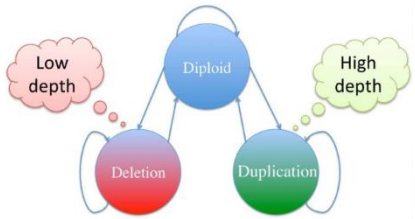
HMM to merge targets and ***discover CNV*** (18)

↓

Genotype CNV across all samples (19)







XHMM Overview

CNV calling pipeline

Mean per-target ***coverage*** using GATK

Filter out "extreme" targets and samples

Mean center each target, run ***PCA***, and remove the top principal components

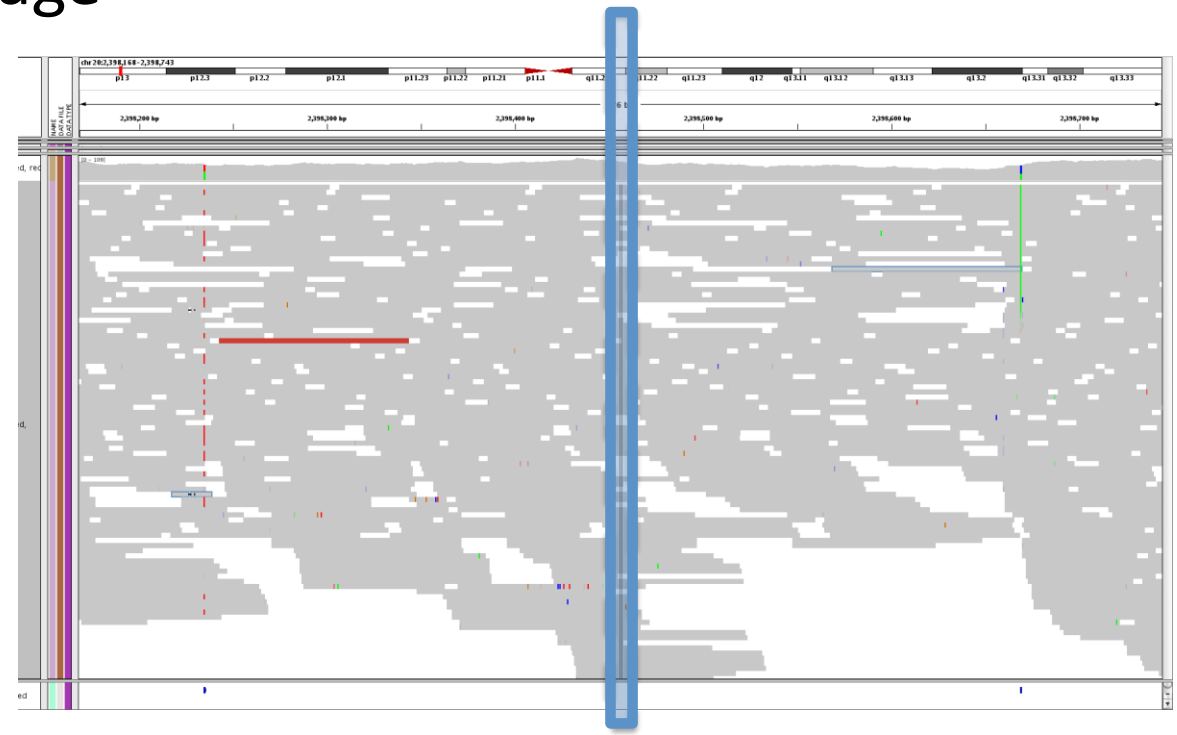
Filter out targets that are extremely variable

Calculate ***z-score*** for each sample's read-depths

HMM to merge targets and call per-sample CNVs

Calculation of depth-of-coverage

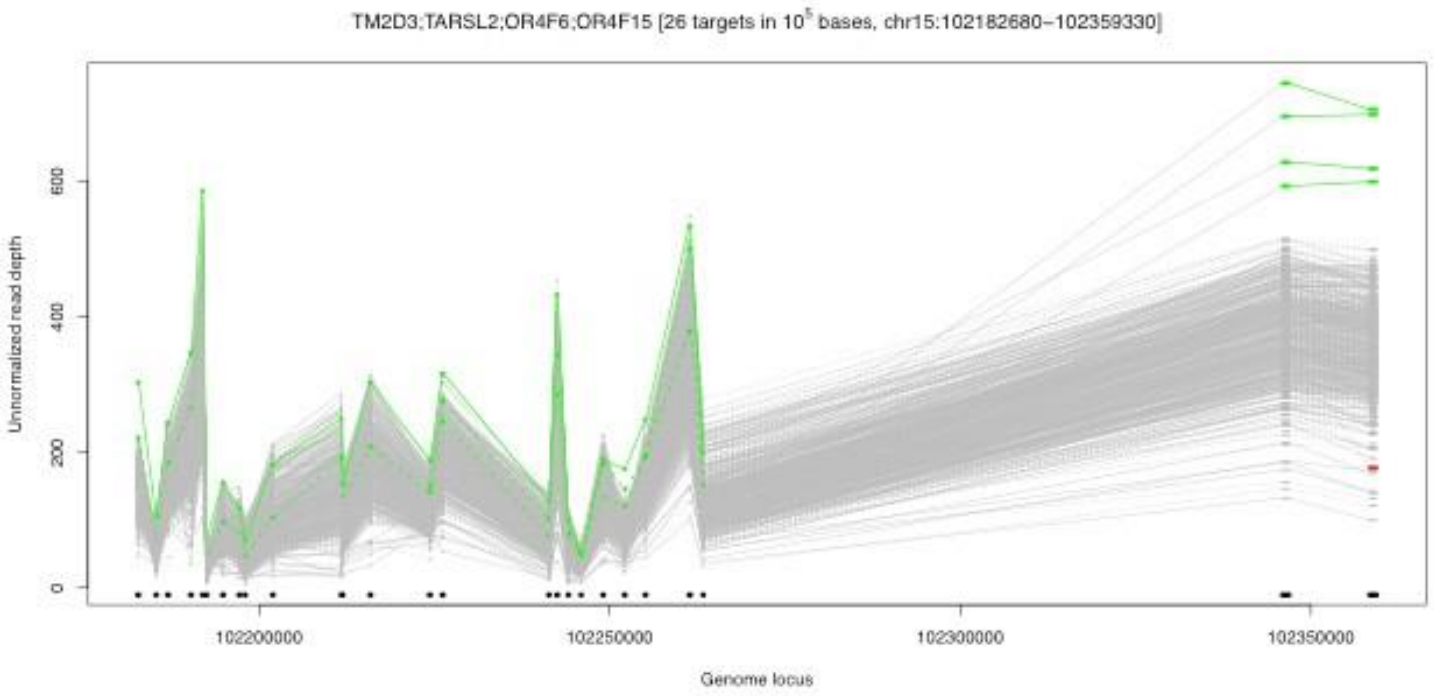
- For each target, calculate the number of reads covering each base in the target, and then average



Original read-depth for region

- 26 targets across 177 kB

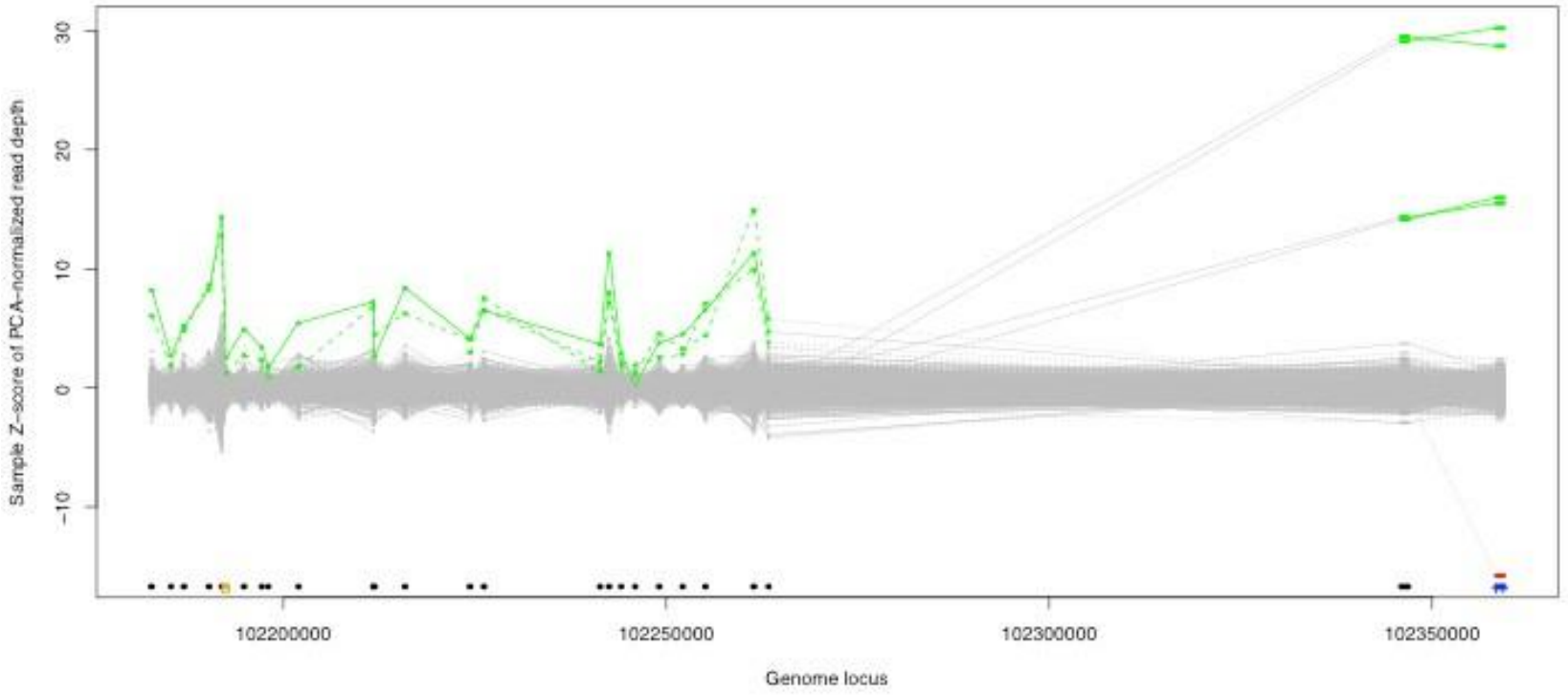
Duplication?



Normalized read-depth for region

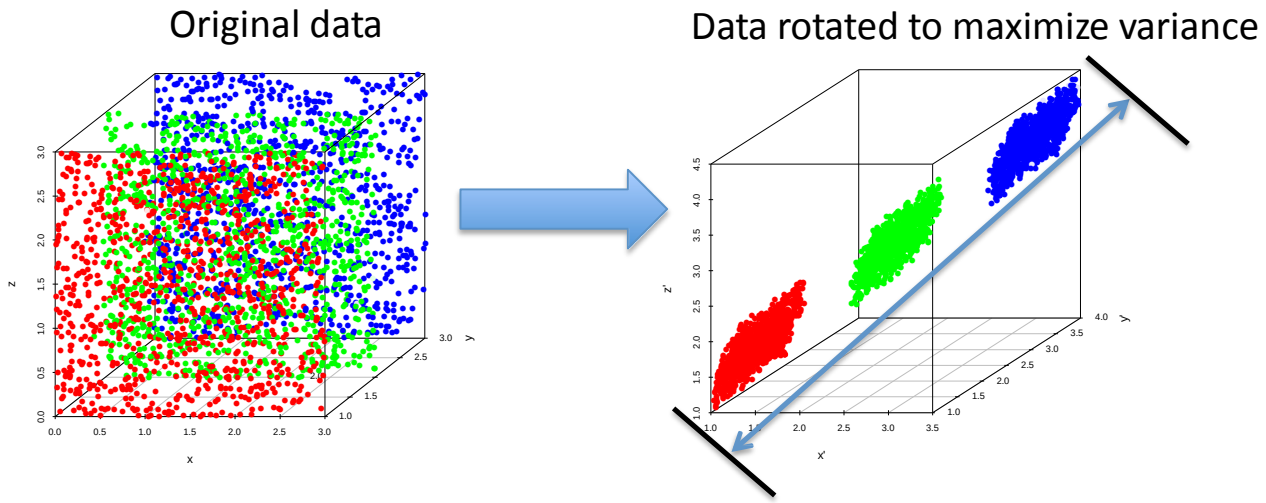
- CNVs can now be detected more easily!

Duplication



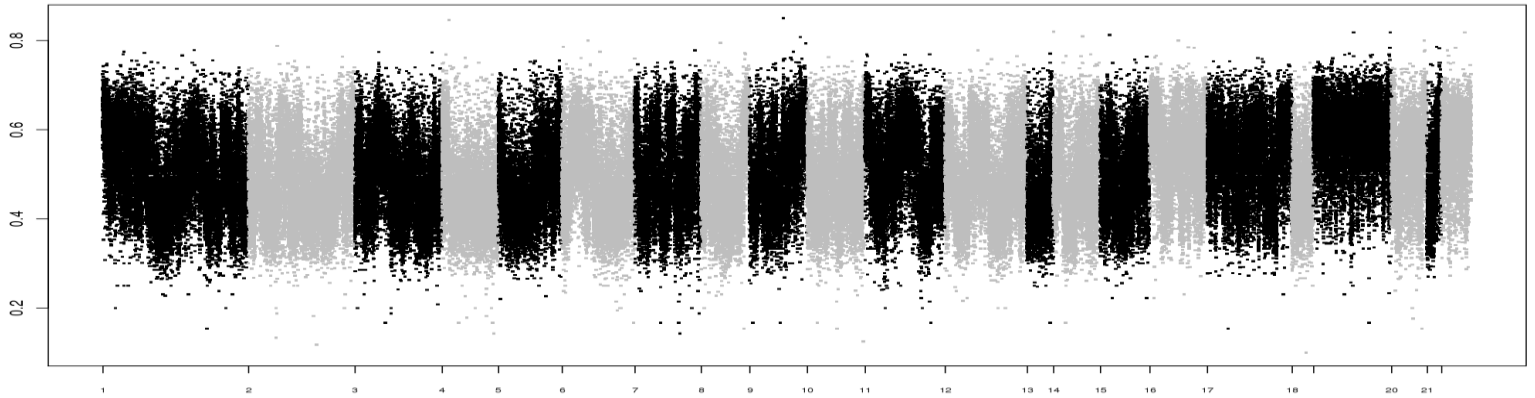
Principal Component Analysis (PCA)

- Rotates high-dimensional data and finds underlying structure
  - Here, we use it to find and remove sample batching effects and target biases (e.g., GC content)
  - Similar to normalisation in CoNIFER (Krumm, et al., Genome Research, 2012)

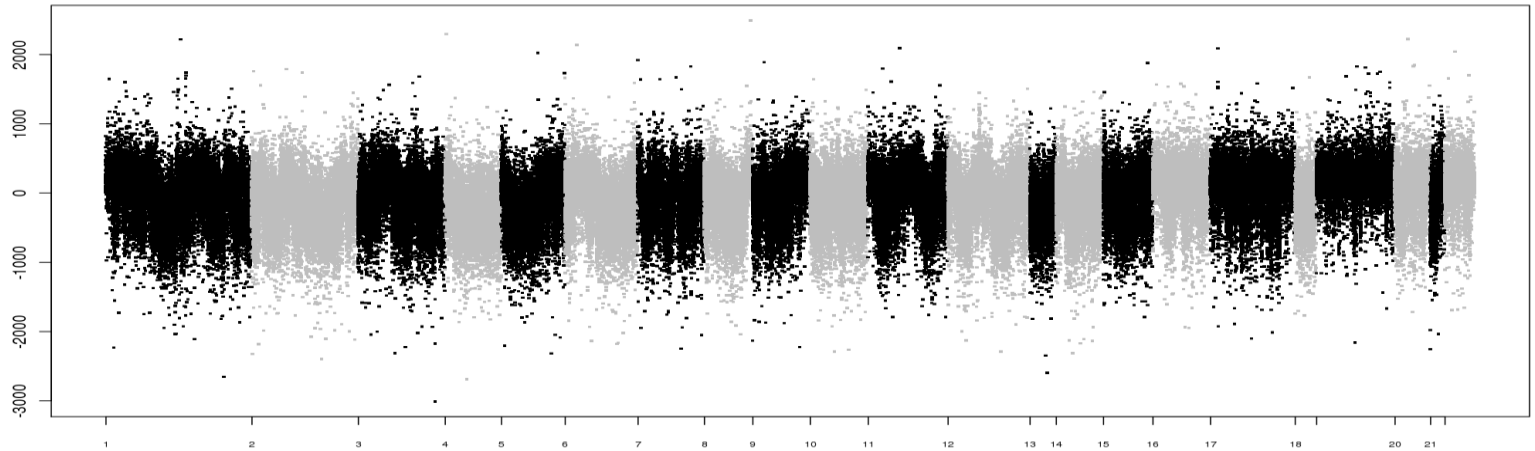


Data-driven ccorrection for GC bias

Target GC content



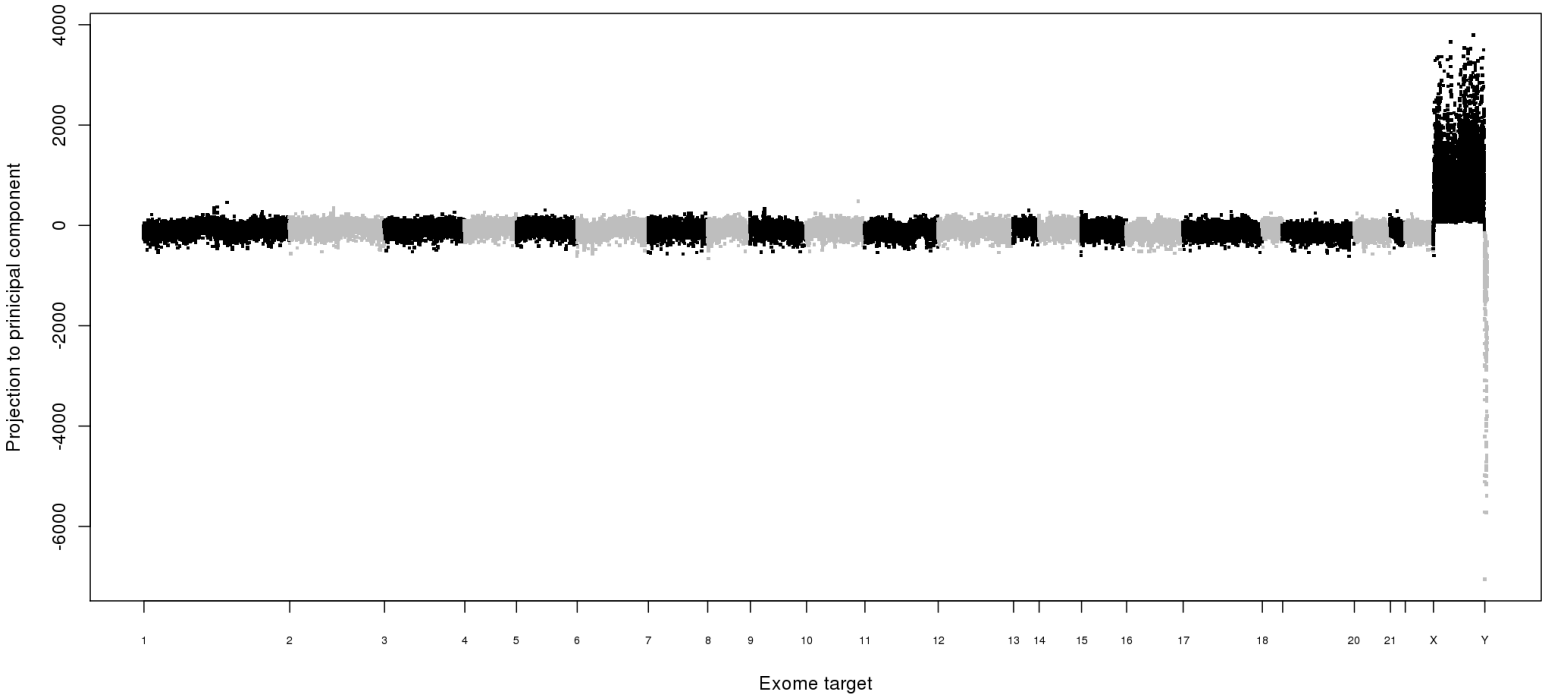
3rd principal component (Pearson correlation = 0.65)



Exome targets

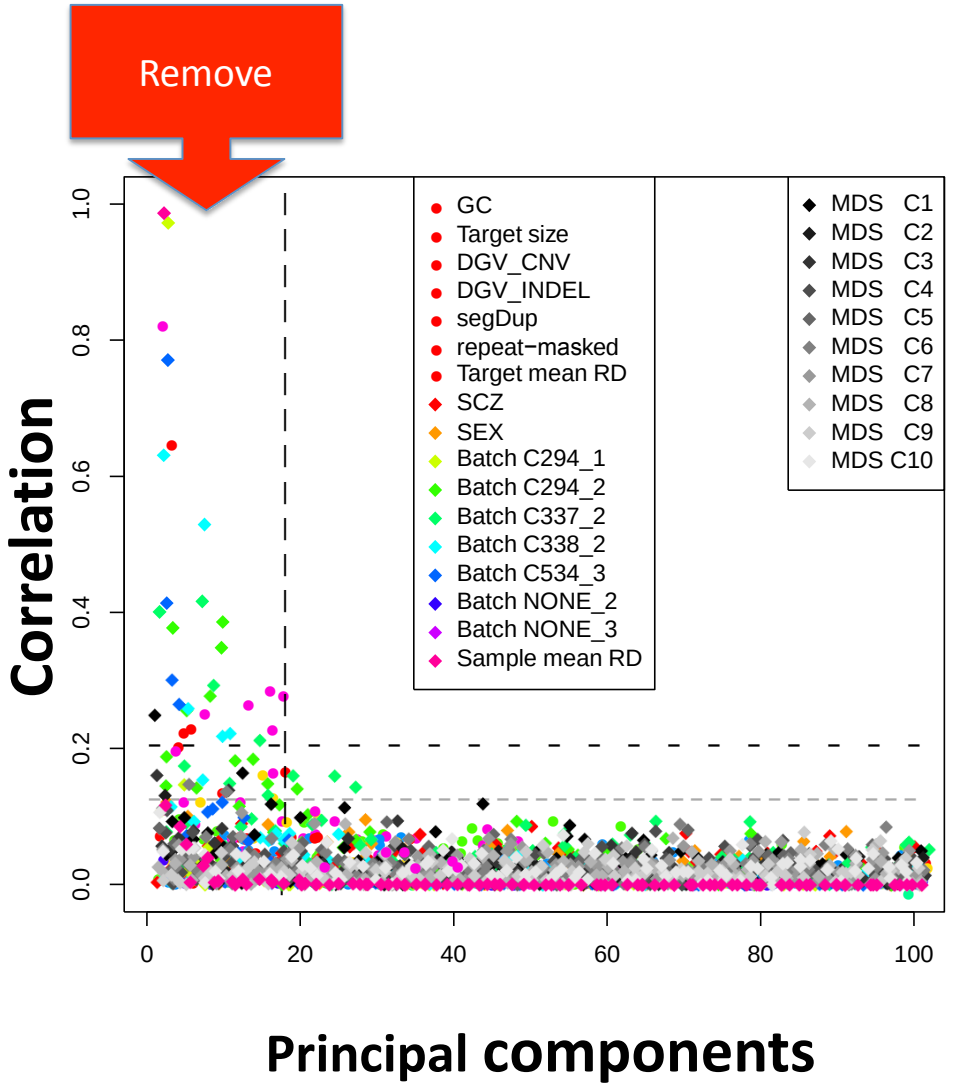
PC correlated with sample sex

- A “female” pattern of read-depth variation:



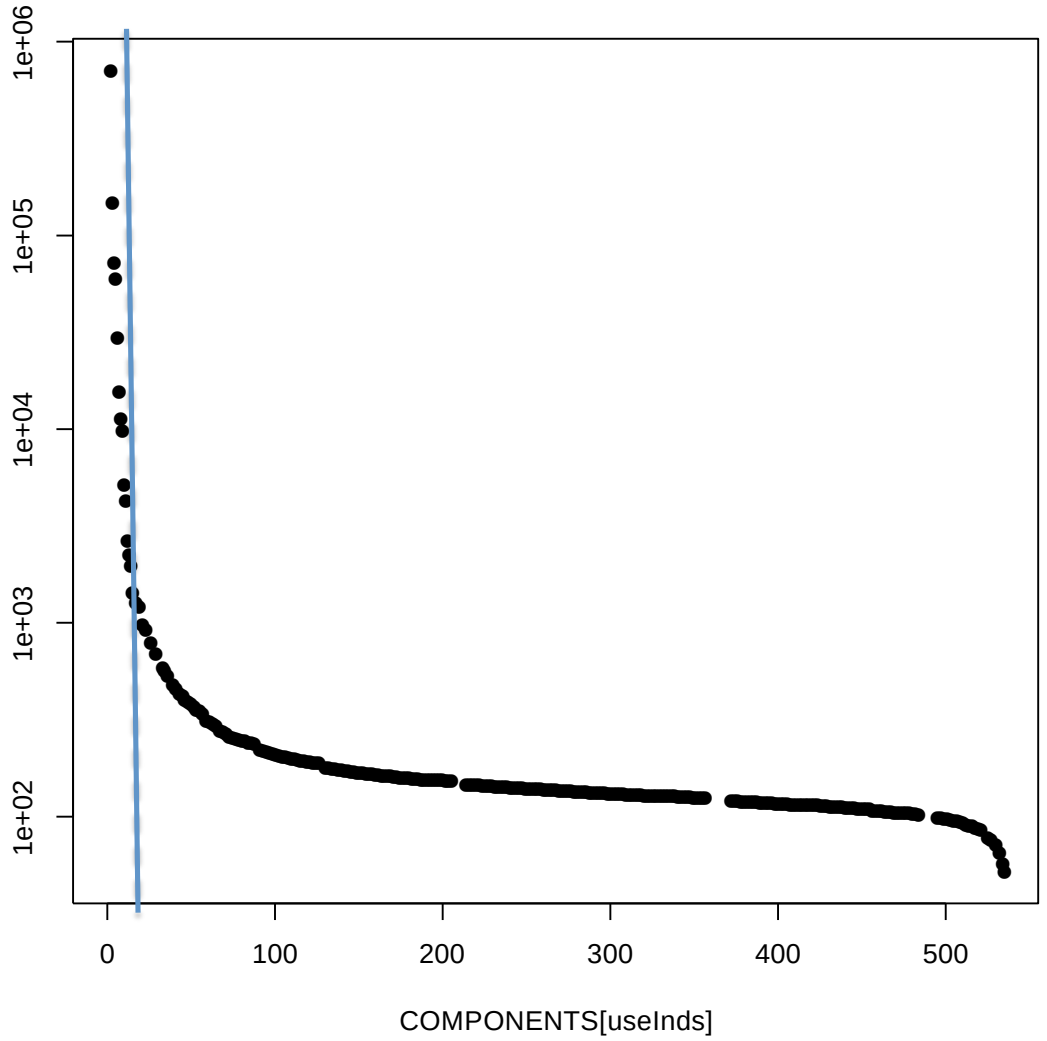
Analysis of PCA components

- Remove first 18 components of read-depth variation, which correlate with:
  - Sample properties
    - Batch, population, mean read depth
  - Target properties
    - GC, size, mean read depth



Variance in PCs also drops a i er ~20 components

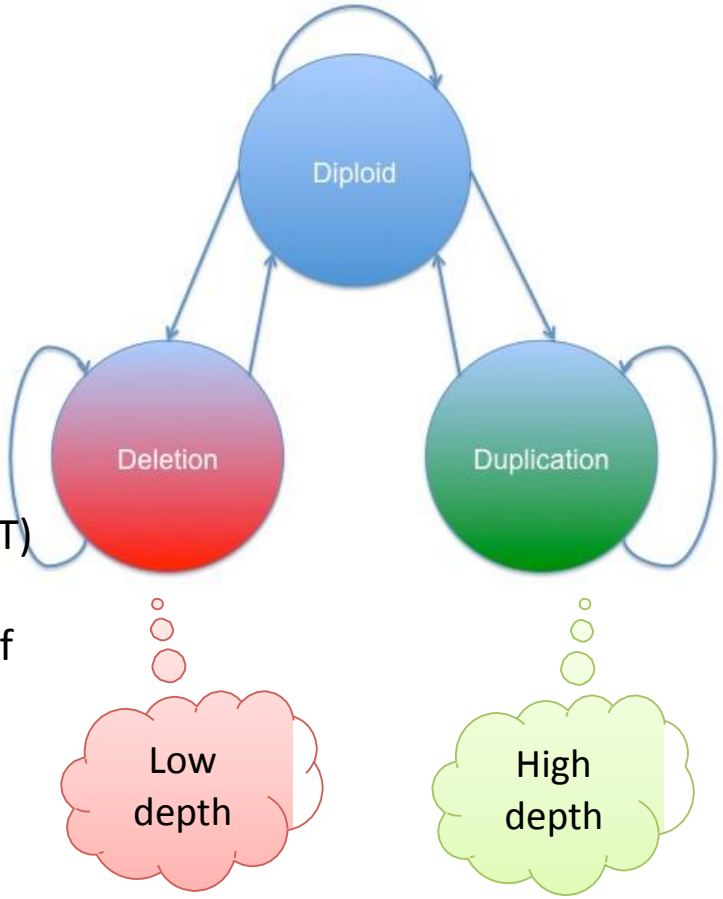
- No a priori knowledge required, except assumption that **largest** effects are **not** CNV signal



Run HMM to call CNVs across neighboring targets

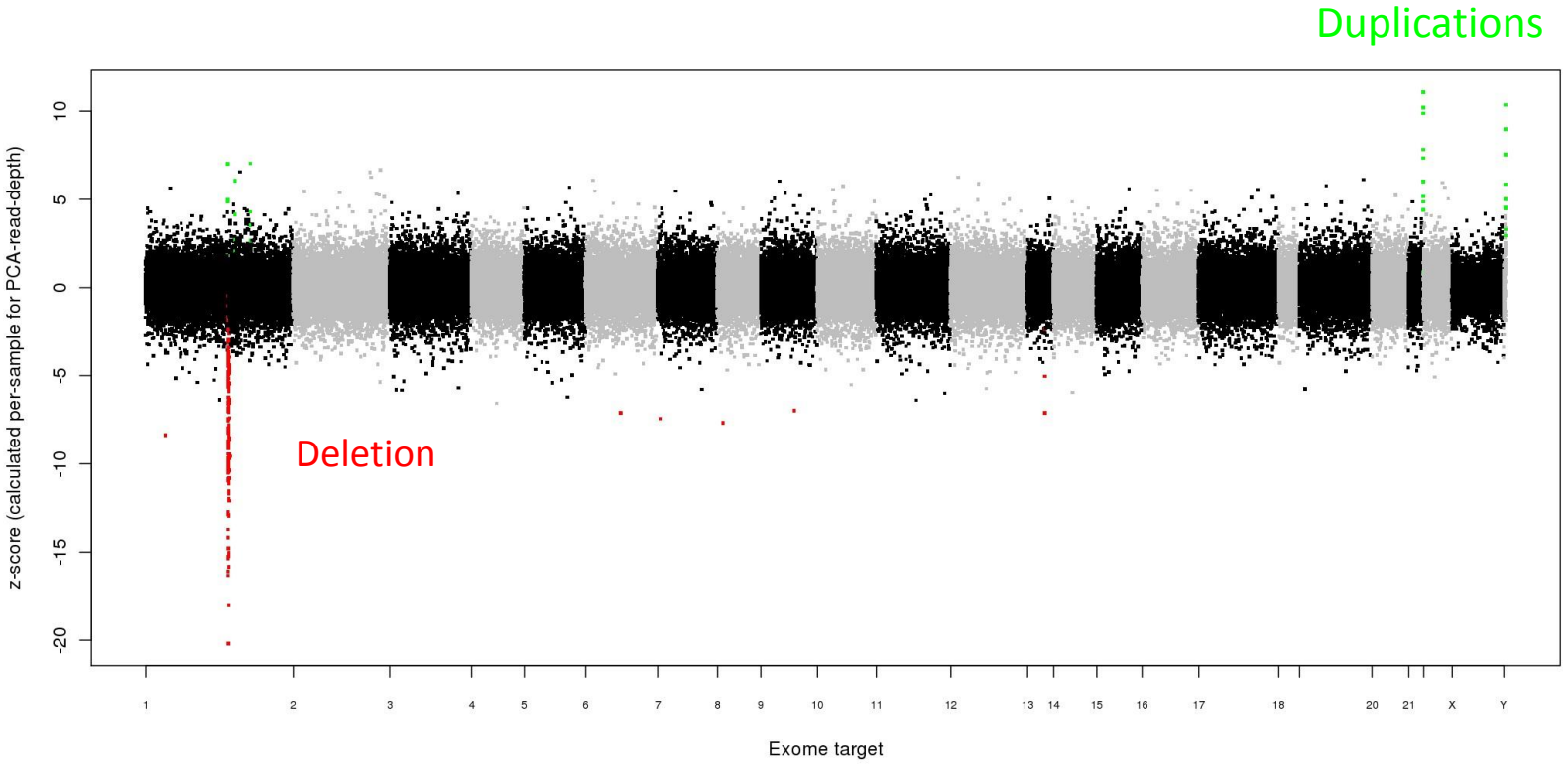
- Hidden Markov model (HMM) to "connect the dots" between adjacent targets
  - Takes into account genome-wide CNV rate, length, and distance between exome targets

- Related methods using NGS
  - Yoon, et al., 2009: event-wise testing (EWT) of intervals of read-depth data
  - Nord, et al., 2011: set-normalized depth of coverage is corrected for GC bias, and supplemented by a scan for partially-mapped reads at CNV edges



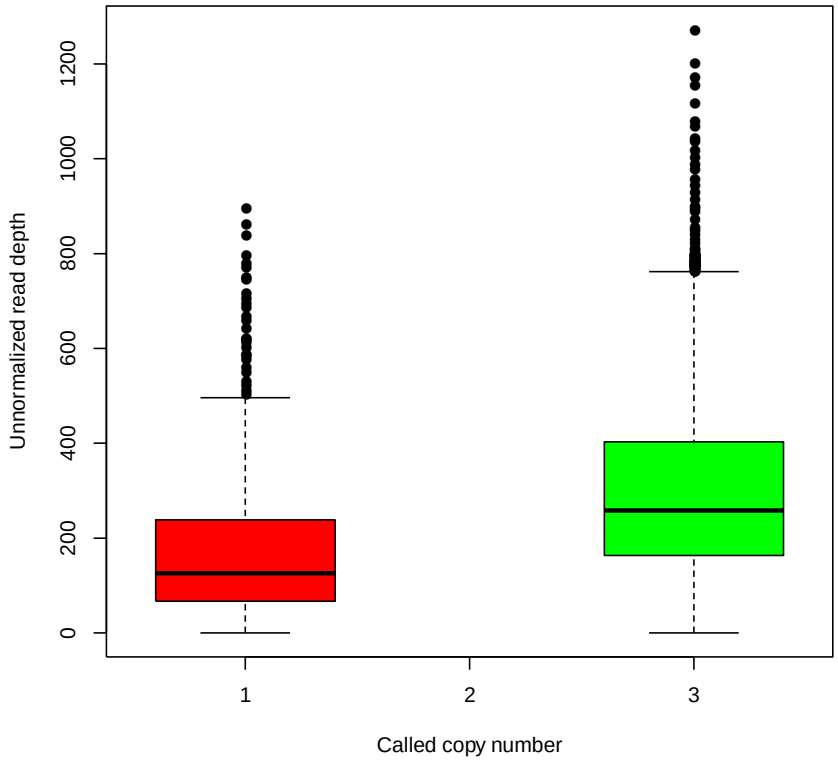
HMM finds regions of consistent deviation in read-depth

- Run hidden Markov model for each sample



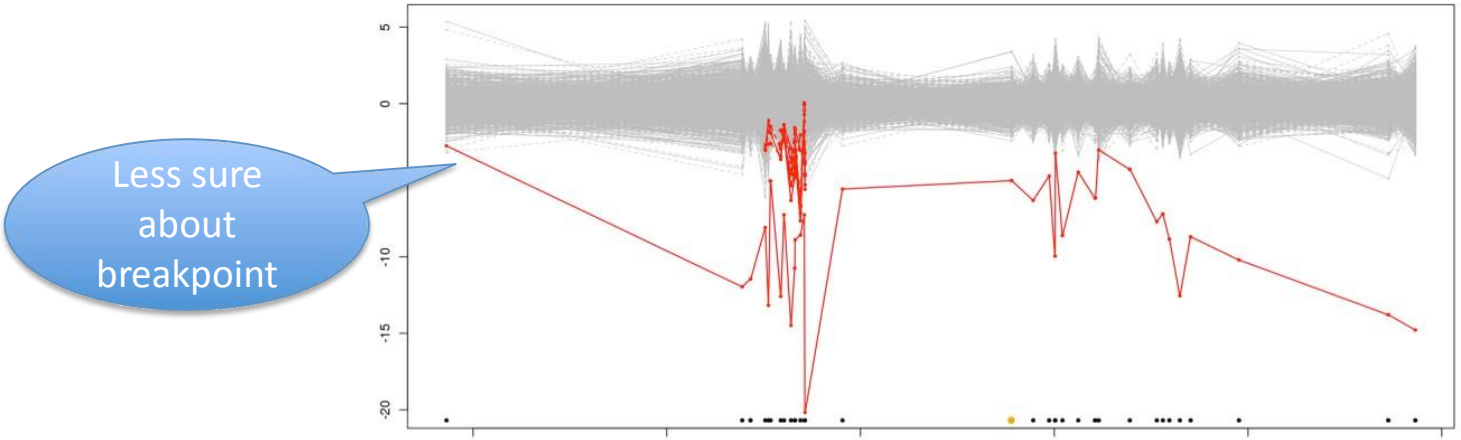
Uncovered duplications and deletions make statistical sense

- Duplication read-depth > Deletion read-depth



Quality filtering using HMM

- CNV call quality (Phred-scaled):
  - $10 * \log_{10}[1 - P(\text{Some CNV in called region})]$
  - Higher quality means event is likely to be real
- For example, for a called deletion, we calculate the probability that at least one of the targets is deleted
  - Calculated efficiently using HMM chain structure



CNV by XHMM
(eXome-Hidden Markov Model)

- Key steps:
  - Running depth of coverage calculations using GATK
  - Coverage normalization
  - CNV calling using Hidden Markov Model
  - Statistical genotyping

- Input:
  - list of exome targets
  - exome sequencing reads
  - Markov model parameters
- Output:
  - CNVs of all target regions

Input files

- List of exome targets (exons)
  - EXOME.interval_list
    - e.g., 22:17072368-17072566
- Exome sequencing reads
  - BAM files (and .bai BAM index files)
- XHMM model parameters
  - params.txt

XHMM parameters: params.txt

1. Exome-wide CNV rate
2. Mean number of targets in CNV
3. Mean distance between targets within CNV (in KB)
4. Mean of DELETION z-score distribution
5. Standard deviation of DELETION z-score distribution
6. Mean of DIPLOID z-score distribution
7. Standard deviation of DIPLOID z-score distribution
8. Mean of DUPLICATION z-score distribution
9. Standard deviation of DUPLICATION z-score distribution

Goes into transition probability

```
As an example, the file with parameters:

**********************************************************************
Input CNV parameters file:
**********************************************************************
1e-08   6       70      -3      1       0       1       3       1
**********************************************************************

translates into XHMM parameters of:

**********************************************************************
Pr(start DEL) = Pr(start DUP) = 1e-08
Mean number of targets in CNV [geometric distribution] = 6
Mean distance between targets within CNV [exponential decay] = 70 KB

DEL read depth distribution ~ N(mean=-3, var=1)
DIP read depth distribution ~ N(mean=0, var=1)
DUP read depth distribution ~ N(mean=3, var=1)
**********************************************************************
```

XHMM parameters: params.txt

1. Exome-wide CNV rate
2. Mean number of targets in CNV
3. Mean distance between targets within CNV (in KB)
4. Mean of DELETION z-score distribution
5. Standard deviation of DELETION z-score distribution
6. Mean of DIPLOID z-score distribution
7. Standard deviation of DIPLOID z-score distribution
8. Mean of DUPLICATION z-score distribution
9. Standard deviation of DUPLICATION z-score distribution

Goes into transition probability

```
As an example, the file with parameters:

**********************************************************************
Input CNV parameters file:
**********************************************************************
1e-08    6      70      -3      1       0       1       3       1
**********************************************************************

translates into XHMM parameters of:

**********************************************************************
Pr(start DEL) = Pr(start DUP) = 1e-08
Mean number of targets in CNV [geometric distribution] = 6
Mean distance between targets within CNV [exponential decay] = 70 KB

DEL read depth distribution ~ N(mean=-3, var=1)
DIP read depth distribution ~ N(mean=0, var=1)
DUP read depth distribution ~ N(mean=3, var=1)
**********************************************************************
```

XHMM parameters: params.txt

1. Exome-wide CNV rate
2. Mean number of targets in CNV
3. Mean distance between targets within CNV (in KB)
4. Mean of DELETION z-score distribution
5. Standard deviation of DELETION z-score distribution
6. Mean of DIPLOID z-score distribution
7. Standard deviation of DIPLOID z-score distribution
8. Mean of DUPLICATION z-score distribution
9. Standard deviation of DUPLICATION z-score distribution

Goes into transition probability

As an example, the file with parameters:

```
**********************************************************************
Input CNV parameters file:
**********************************************************************
1e-08   6       70      -3      1       0       1       3       1
**********************************************************************

translates into XHMM parameters of:

**********************************************************************
Pr(start DEL) = Pr(start DUP) = 1e-08
Mean number of targets in CNV [geometric distribution] = 6
Mean distance between targets within CNV [exponential decay] = 70 KB

DEL read depth distribution ~ N(mean=-3, var=1)
DIP read depth distribution ~ N(mean=0, var=1)
DUP read depth distribution ~ N(mean=3, var=1)
**********************************************************************
```

XHMM parameters: params.txt

1. Exome-wide CNV rate
2. Mean number of targets in CNV
3. Mean distance between targets within CNV (in KB)
4. Mean of DELETION z-score distribution
5. Standard deviation of DELETION z-score distribution
6. Mean of DIPLOID z-score distribution
7. Standard deviation of DIPLOID z-score distribution
8. Mean of DUPLICATION z-score distribution
9. Standard deviation of DUPLICATION z-score distribution

Emission probability

As an example, the file with parameters:

```
**********************************************************************
Input CNV parameters file:
**********************************************************************
1e-08   6       70      -3      1       0       1       3       1
**********************************************************************

translates into XHMM parameters of:

**********************************************************************
Pr(start DEL) = Pr(start DUP) = 1e-08
Mean number of targets in CNV [geometric distribution] = 6
Mean distance between targets within CNV [exponential decay] = 70 KB

DEL read depth distribution ~ N(mean=-3, var=1)
DIP read depth distribution ~ N(mean=0, var=1)
DUP read depth distribution ~ N(mean=3, var=1)
**********************************************************************
```

XHMM parameters: params.txt

1. Exome-wide CNV rate
2. Mean number of targets in CNV
3. Mean distance between targets within CNV (in KB)
4. Mean of DELETION z-score distribution
5. Standard deviation of DELETION z-score distribution
6. Mean of DIPLOID z-score distribution
7. Standard deviation of DIPLOID z-score distribution
8. Mean of DUPLICATION z-score distribution
9. Standard deviation of DUPLICATION z-score distribution

Emission probability

As an example, the file with parameters:

```
**********************************************************************
Input CNV parameters file:
**********************************************************************
1e-08   6       70      -3      1       0       1       3       1
**********************************************************************

translates into XHMM parameters of:

**********************************************************************
Pr(start DEL) = Pr(start DUP) = 1e-08
Mean number of targets in CNV [geometric distribution] = 6
Mean distance between targets within CNV [exponential decay] = 70 KB

DEL read depth distribution ~ N(mean=-3, var=1)
DIP read depth distribution ~ N(mean=0, var=1)
DUP read depth distribution ~ N(mean=3, var=1)
**********************************************************************
```

XHMM parameters: params.txt

1. Exome-wide CNV rate
2. Mean number of targets in CNV
3. Mean distance between targets within CNV (in KB)
4. Mean of DELETION z-score distribution
5. Standard deviation of DELETION z-score distribution
6. Mean of DIPLOID z-score distribution
7. Standard deviation of DIPLOID z-score distribution
8. Mean of DUPLICATION z-score distribution
9. Standard deviation of DUPLICATION z-score distribution

Emission probability

As an example, the file with parameters:

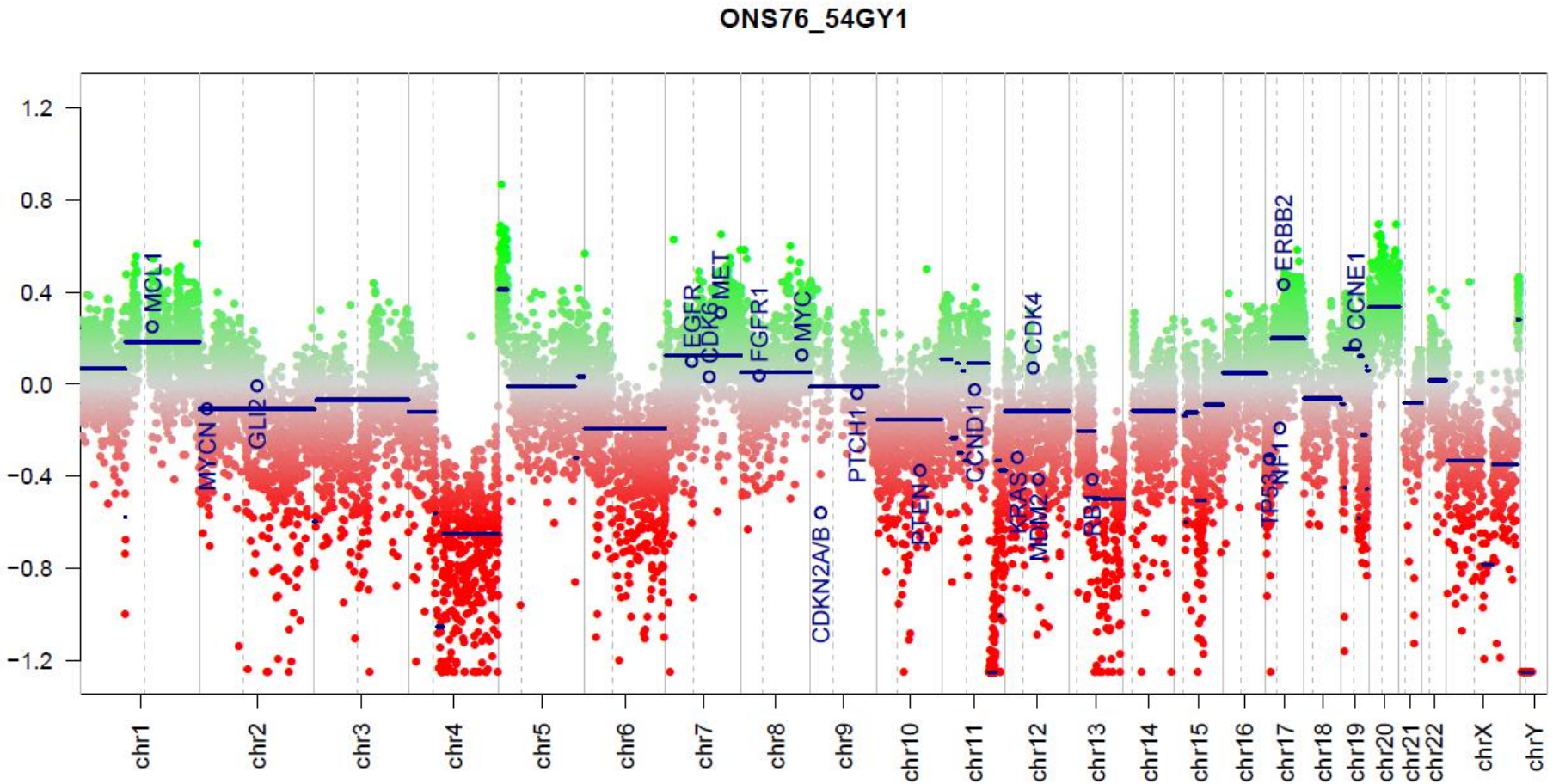
```
**********************************************************************
Input CNV parameters file:
**********************************************************************
1e-08   6       70      -3      1       0       1       3       1
**********************************************************************

translates into XHMM parameters of:

**********************************************************************
Pr(start DEL) = Pr(start DUP) = 1e-08
Mean number of targets in CNV [geometric distribution] = 6
Mean distance between targets within CNV [exponential decay] = 70 KB

DEL read depth distribution ~ N(mean=-3, var=1)
DUP read depth distribution ~ N(mean=0, var=1)
DUP read depth distribution ~ N(mean=3, var=1)
**********************************************************************
```

Running XHMM on cell lines

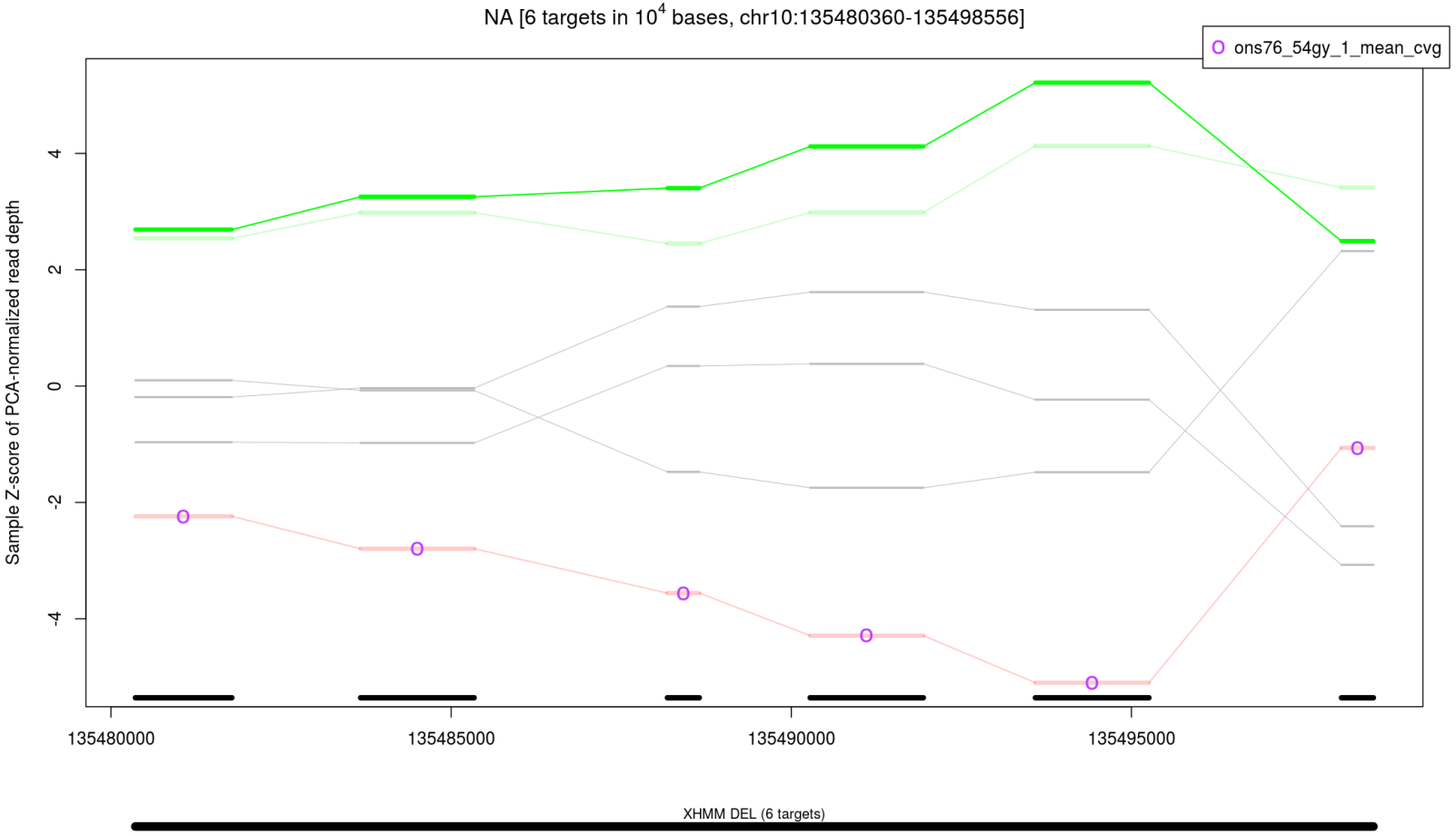
only Chr10 as an example

	SAMPLE	CNV	FULL_INTERVAL	TARGET_IND	TARGET	CHR	MID_BP	POSTERIO	RD	ORIG_RD
2	ons76_neg	DUP	10:135480360-135498556	U-2	10:135439254-13543992	10	135439587	0	0.18	48.79
3	ons76_neg	DUP	10:135480360-135498556	U-1	10:135439968-13544036	10	135440167	0	-2.23	42.88
4	ons76_neg	DUP	10:135480360-135498556	1	10:135480360-13548177	10	135481068	0.97	2.69	222.29
5	ons76_neg	DUP	10:135480360-135498556	2	10:135483670-13548533	10	135484504	1	3.26	220.03
6	ons76_neg	DUP	10:135480360-135498556	3	10:135488179-13548864	10	135488413	1	3.4	185.21
7	ons76_neg	DUP	10:135480360-135498556	4	10:135490278-13549194	10	135491109	1	4.12	230.51
8	ons76_neg	DUP	10:135480360-135498556	5	10:135493588-13549525	10	135494422	1	5.22	206.04
9	ons76_neg	DUP	10:135480360-135498556	6	10:135498087-13549855	10	135498321	0.99	2.49	179.56
10	ons76_54gy_1	DEL	10:135480360-135498556	U-2	10:135439254-13543992	10	135439587	0	-0.24	49.52
11	ons76_54gy_1	DEL	10:135480360-135498556	U-1	10:135439968-13544036	10	135440167	0	1.45	61.04
12	ons76_54gy_1	DEL	10:135480360-135498556	1	10:135480360-13548177	10	135481068	0.88	-2.24	216.32
13	ons76_54gy_1	DEL	10:135480360-135498556	2	10:135483670-13548533	10	135484504	1	-2.8	205.76
14	ons76_54gy_1	DEL	10:135480360-135498556	3	10:135488179-13548864	10	135488413	1	-3.56	177
15	ons76_54gy_1	DEL	10:135480360-135498556	4	10:135490278-13549194	10	135491109	1	-4.29	215.03
16	ons76_54gy_1	DEL	10:135480360-135498556	5	10:135493588-13549525	10	135494422	1	-5.1	194.87
17	ons76_54gy_1	DEL	10:135480360-135498556	6	10:135498087-13549855	10	135498321	0.52	-1.06	174.99
18	ons76_54gy_2	DUP	10:135480360-135498556	U-2	10:135439254-13543992	10	135439587	0	0.07	53.51
19	ons76_54gy_2	DUP	10:135480360-135498556	U-1	10:135439968-13544036	10	135440167	0	-2.53	55.73
20	ons76_54gy_2	DUP	10:135480360-135498556	1	10:135480360-13548177	10	135481068	0.95	2.54	237.03
21	ons76_54gy_2	DUP	10:135480360-135498556	2	10:135483670-13548533	10	135484504	1	2.98	223.87
22	ons76_54gy_2	DUP	10:135480360-135498556	3	10:135488179-13548864	10	135488413	1	2.45	201.47
23	ons76_54gy_2	DUP	10:135480360-135498556	4	10:135490278-13549194	10	135491109	1	2.99	236.63
24	ons76_54gy_2	DUP	10:135480360-135498556	5	10:135493588-13549525	10	135494422	1	4.13	230.34
25	ons76_54gy_2	DUP	10:135480360-135498556	6	10:135498087-13549855	10	135498321	1	3.41	191.1
26										

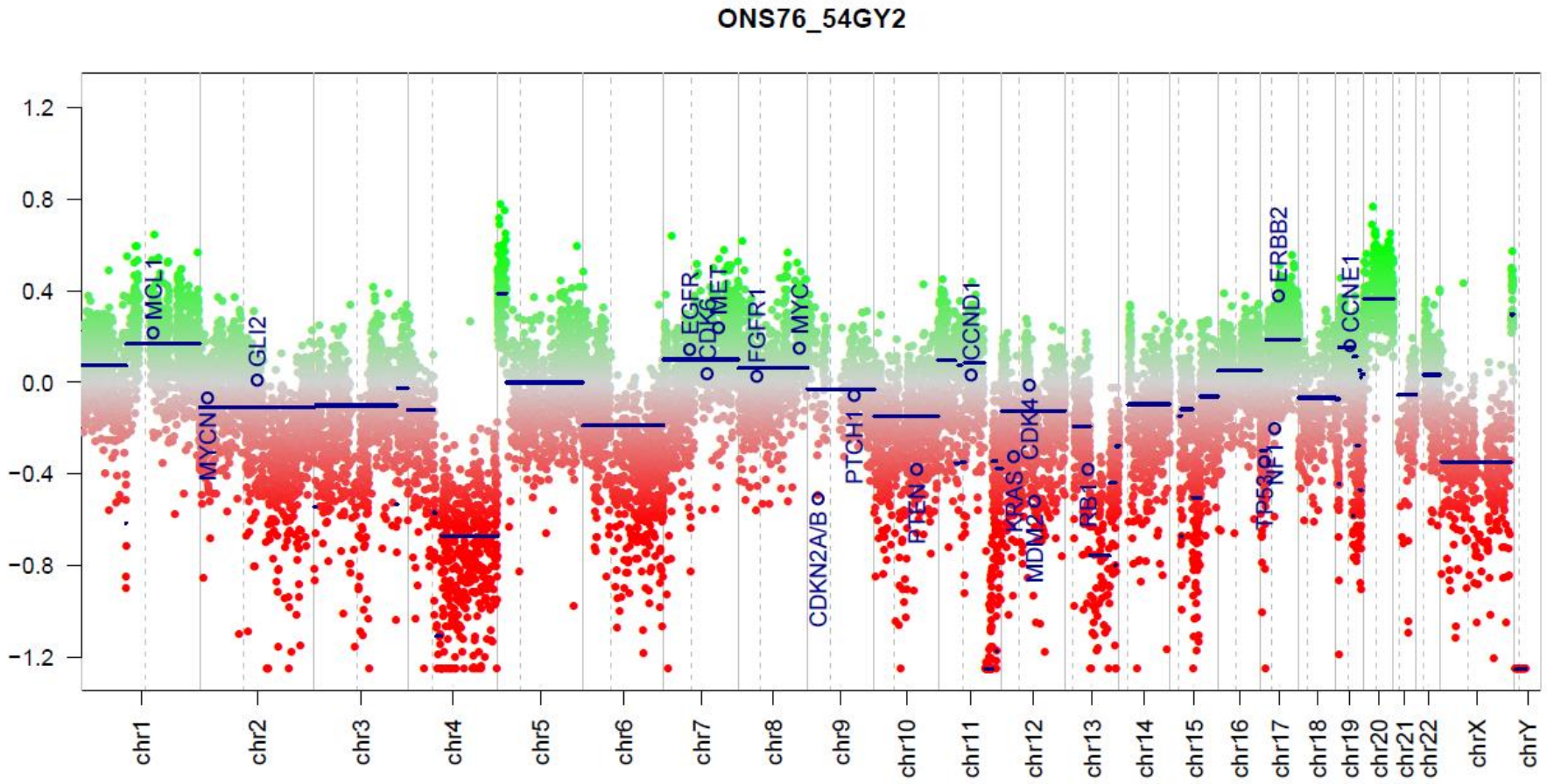
CNV by conumee: ONS76_54GY_1 (cell lines)



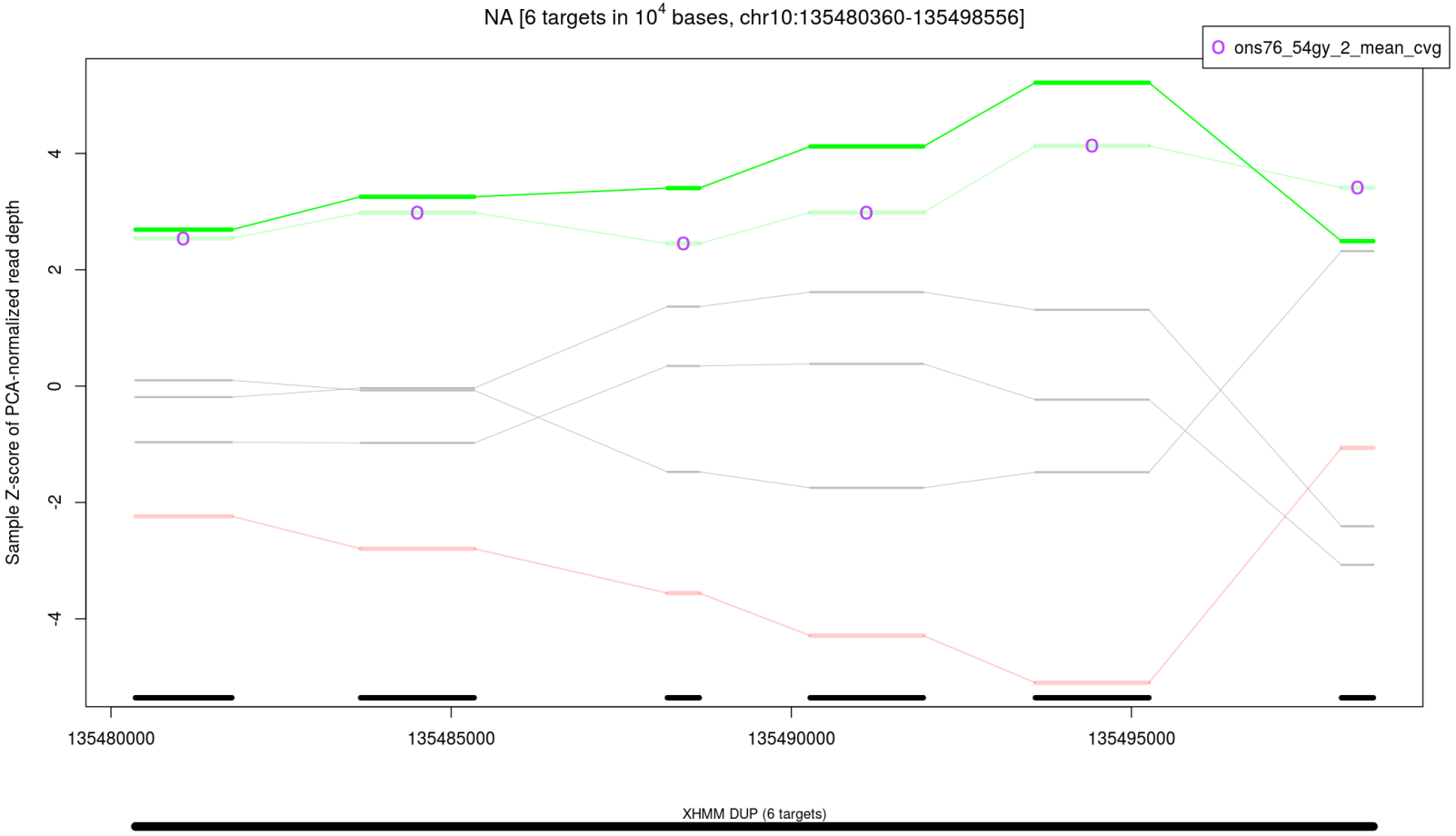
CNV by XHMM: Ons76_54gy_1



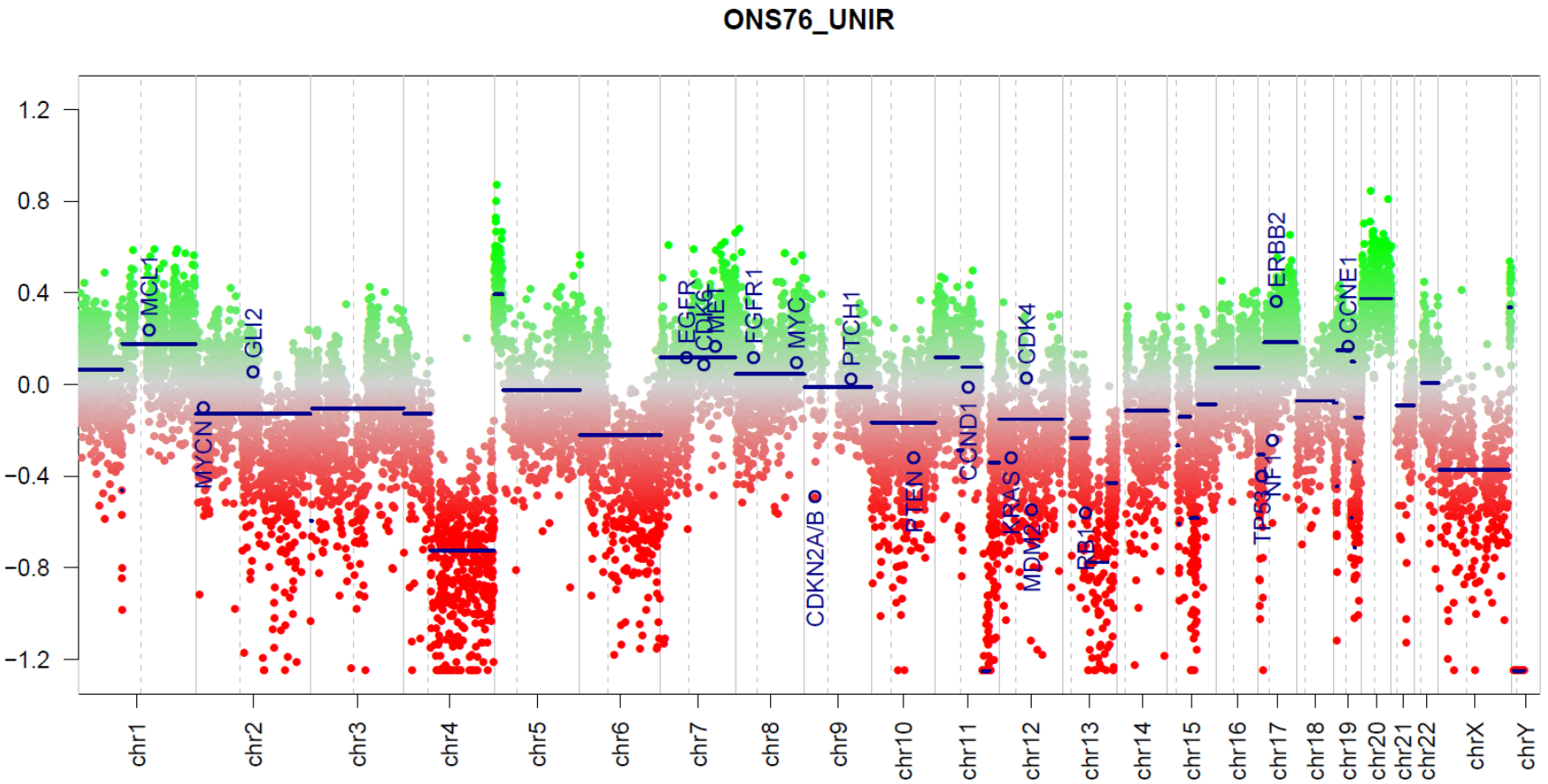
CNV by conumee: ONS76_54GY2



CNV by XHMM: Ons76_54gy_2



CNV by conumee: ONS76_neg



CNV by XHMM: Ons76_neg