Copy Number Variant calling for CCGD and Profile

Outline

- Background
  - Why is calling CNVs from targeted capture data so hard?
    - Sparse, noisy data
- ReCapSeg & RobustCNV algorithm
- QC for Normalization
  - Metrics and expectations
- Assessment and validation for CCGD and Profile
  - Data sets
  - Methods
  - Results
- Additional Considerations
  - Focal gains/losses, tumor suppressor genes
- Conclusions and Future Directions

Expectation

A change in copy number at a locus → A corresponding change in the probability that reads are sampled from that locus. → A change in relative mapping coverage of sequenced reads at that locus.

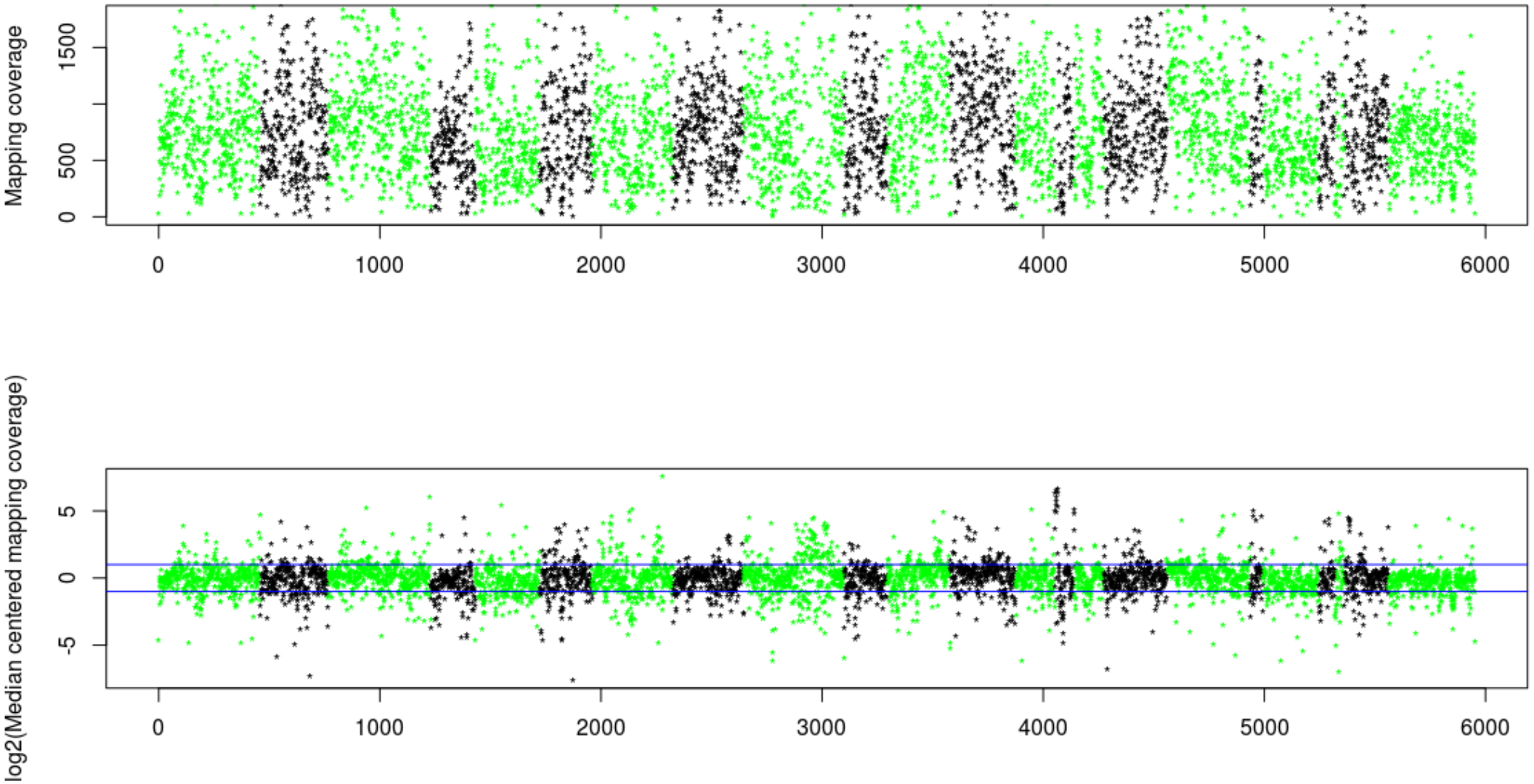
Caveats:

- Unbiased sampling
  - Biased sampling may mask CNVs
- Random noise < CNV signal
- Sufficient sampling density to detect CNVs

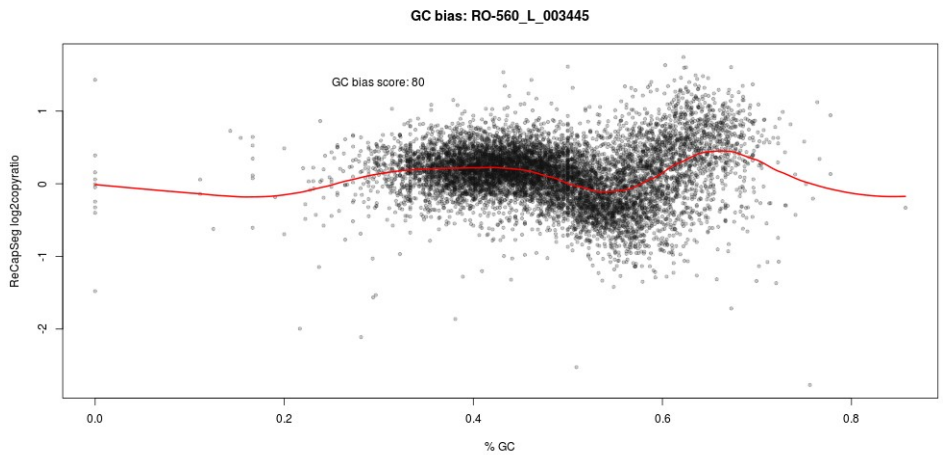
Sparse Noisy Data

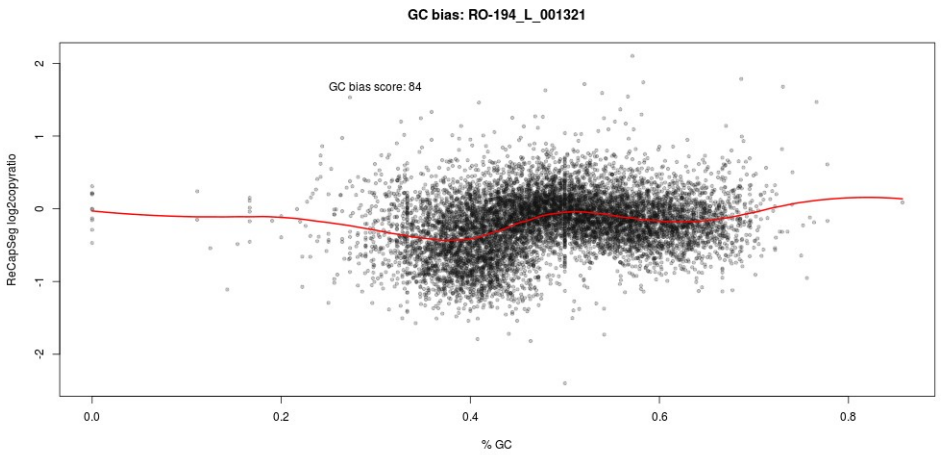
- Sparsity
  - Capture data
    - 3000 – 10,000 total loci baited
    - Some chromosomes have < 100 baited loci
    - Some genes have minimal representation (POPv1 32/283 <= 4 intervals)
- Noise
  - Random noise (sampling)
  - Systematic biases
    - GC content (differs across samples)
    - Capture design (differing capture probe depths)
    - Batch effects (sequencing depth, library prep)
    - Other sources (?)

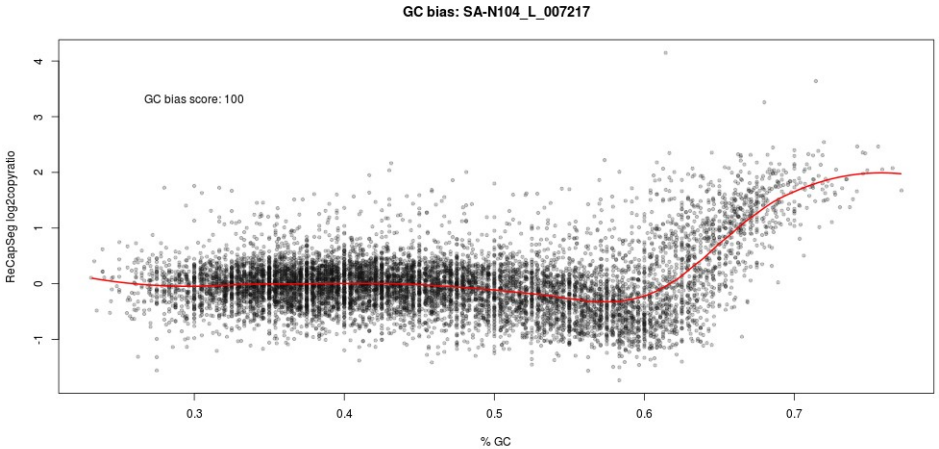
Random noise

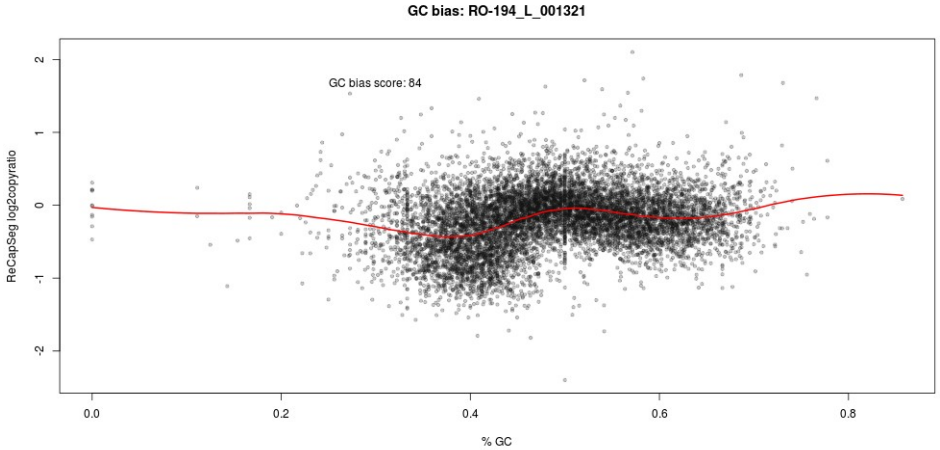


GC bias

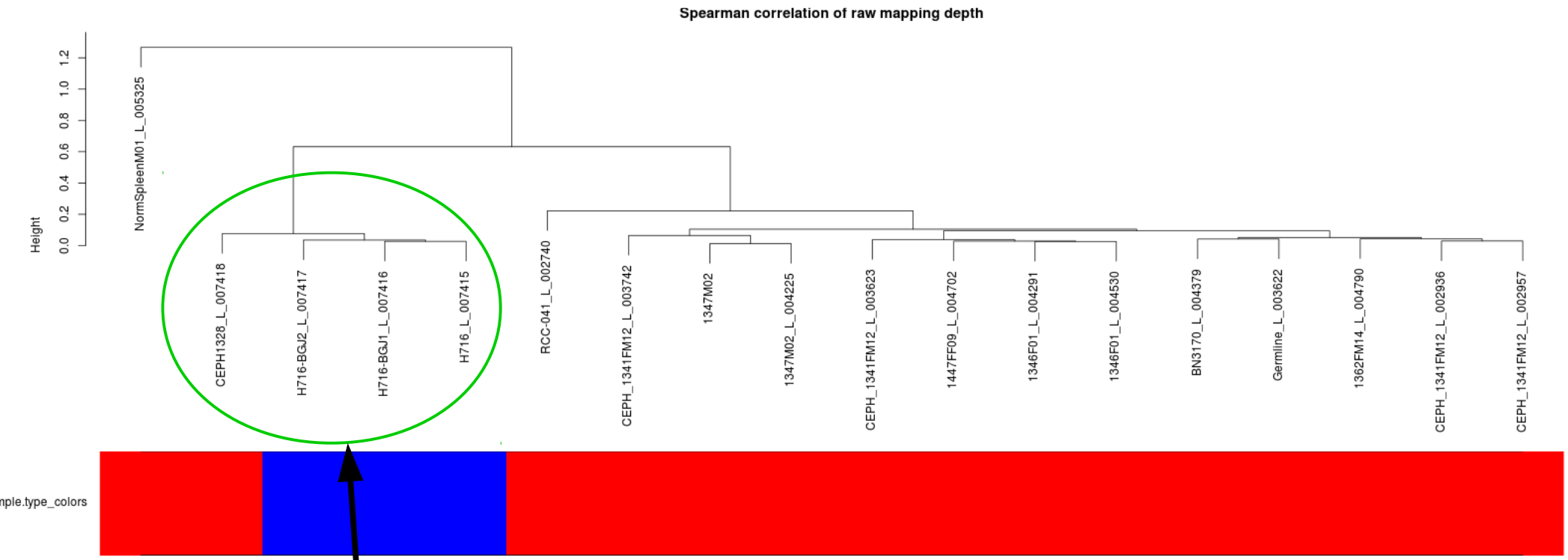

GC bias: RO-560_L_003445
GC bias score: 80
ReCapSeg log2copyratio
% GC


GC bias: RO-194_L_001321
GC bias score: 84
ReCapSeg log2copyratio
% GC


GC bias: SA-N104_L_007217
GC bias score: 100
ReCapSeg log2copyratio
% GC

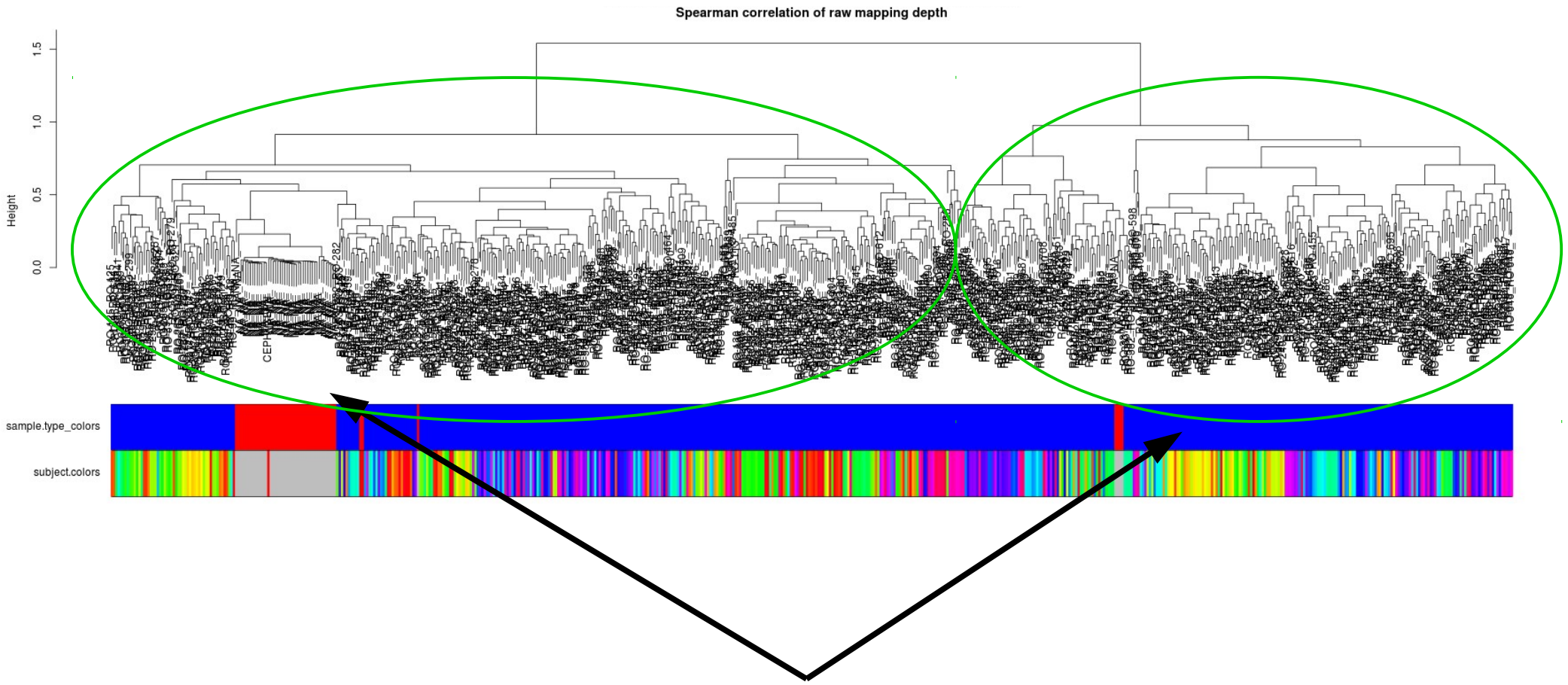

GC bias: RO-194_L_001321
GC bias score: 84
ReCapSeg log2copyratio
% GC

Batch Effects/Clustering



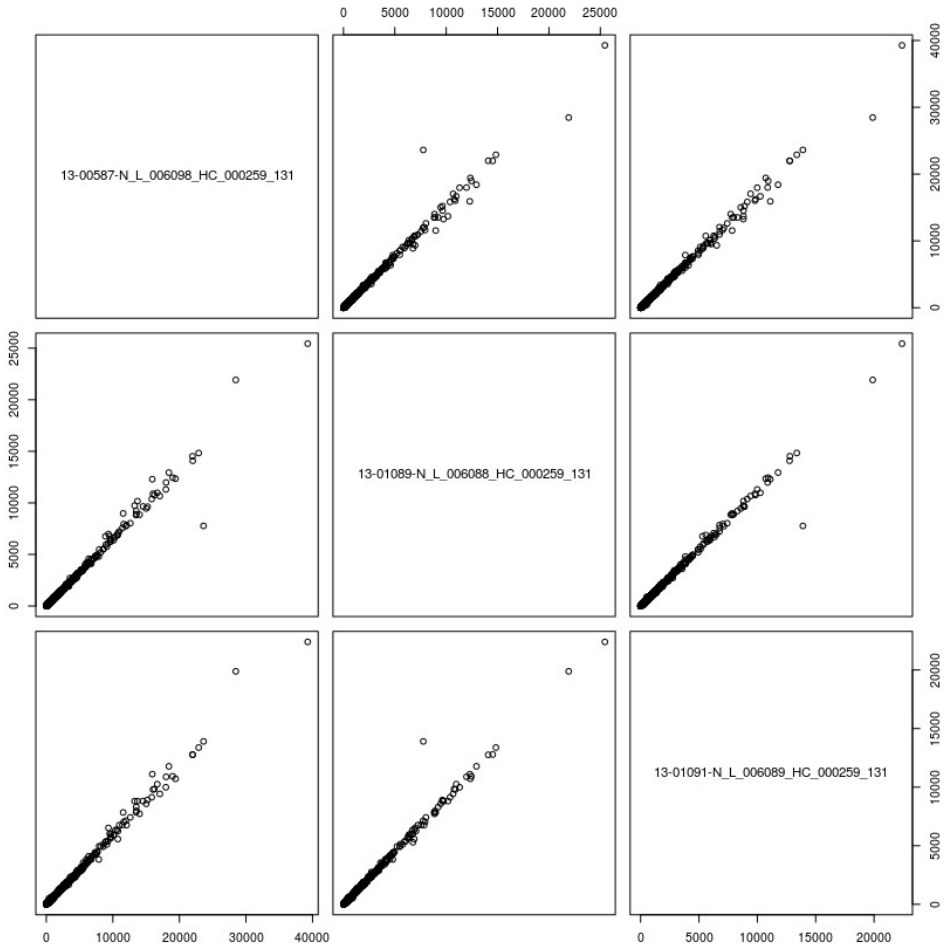
Batch effect is stronger than tissue type effect.

Batch Effects



Two clear groupings between samples

Correlation between samples



Background

- Conclusions
  - Mapping coverage in targeted capture data is:
    - Noisy
    - Sparse
    - Strong correlation patterns provide evidence of systematic biases which change from sample to sample and batch to batch.
  - Many CN events are likely to be obscured by noise.

ReCapSeg

- Normalization and CNV calling strategy
  - Attempts to remove systematic bias through “Tangent normalization”.
  - Attempts to average-out random noise through segmentation. Calls are made on the average value of all intervals within a segment.

ReCapSeg

(slide adapted from Lee Lichtenstein)

Generate Proportional Coverage

- Input: bam file
- Output: target bins
- Output: Proportional Coverage
- Run time: ~1.5h

Replaced with a faster version

Tangent Normalization

- **Input: Proportional Coverage**
- **Input: panel of normals (PoN)**
- **Output: Normalized coverage profile**

Segment

- Input: Normalized coverage profile
- Output: Segments
- DNAcopy – Circular Binary Segmentation Olshen and Venkatraman, 2004.

Call

- Input: Segments
- Input: Normalized coverage profile
- Output: Segments with calls
- Minimum segment size: 3 targets
- Calls based on cutoff derived from variance of sample

Oncotator

- Input: segments with calls
- Output: segments and calls sorted by gene/exon
- Run time: 4 minutes

Replaced with R script that only reports captured genes

Performance with POPv1, (~4803 intervals)

Time	~5 min	< 30s	< 30s per sample	< 30s per sample	Not Run

ReCapSeg

(slide adapted from Lee Lichtenstein)

Generate Proportional Coverage

- Input: bam file
- Output: target bins
- Output: Proportional Coverage
- Run time: ~1.5h

Tangent Normalization

- Input: Proportional Coverage
- Input: panel of normals (PoN)
- Output: Normalized coverage profile

Segment

- **Input: Normalized coverage profile**
- **Output: Segments**
- **DNAcopy – Circular Binary Segmentation Olshen and Venkatraman, 2004.**

Call

- Input: Segments
- Input: Normalized coverage profile
- Output: Segments with calls
- Minimum segment size: 3 targets
- Calls based on cutoff derived from variance of sample

Oncotator

- Input: segments with calls
- Output: segments and calls sorted by gene/exon
- Run time: 4 minutes

Replaced with a faster version

Replaced with R script that only reports captured genes

Performance with POPv1, (~4803 intervals)

Time	~5 min	< 30s	< 30s per sample	< 30s per sample	Not Run

ReCapSeg

(slide adapted from Lee Lichtenstein)

Generate Proportional Coverage

- Input: bam file
- Output: target bins
- Output: Proportional Coverage
- Run time: ~1.5h

Tangent Normalization

- Input: Proportional Coverage
- Input: panel of normals (PoN)
- Output: Normalized coverage profile

Segment

- Input: Normalized coverage profile
- Output: Segments
- DNAcopy – Circular Binary Segmentation Olshen and Venkatraman, 2004.

Call

- **Input: Segments**
- **Input: Normalized coverage profile**
- **Output: Segments with calls**
- **Minimum segment size: 3 targets**
- **Calls based on cutoff derived from variance of sample**

Oncotator

- Input: segments with calls
- Output: segments and calls sorted by gene/exon
- Run time: 4 minutes

Replaced with a faster version

Replaced with R script that only reports captured genes

Performance with POPv1, (~4803 intervals)

Time	~5 min	< 30s	< 30s per sample	< 30s per sample	Not Run

Tangent Normalization
extract “e”

CNV sample
(observed)

$$y = ß_1 * PON\_1 + ß_2 * PON\_2 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ ß_n * PON\_n + e$$

minimize “e” with Least Squares

$$\hat{y} = ß_1 * PON\_1 + ß_2 * PON\_2 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ ß_n * PON\_n$$

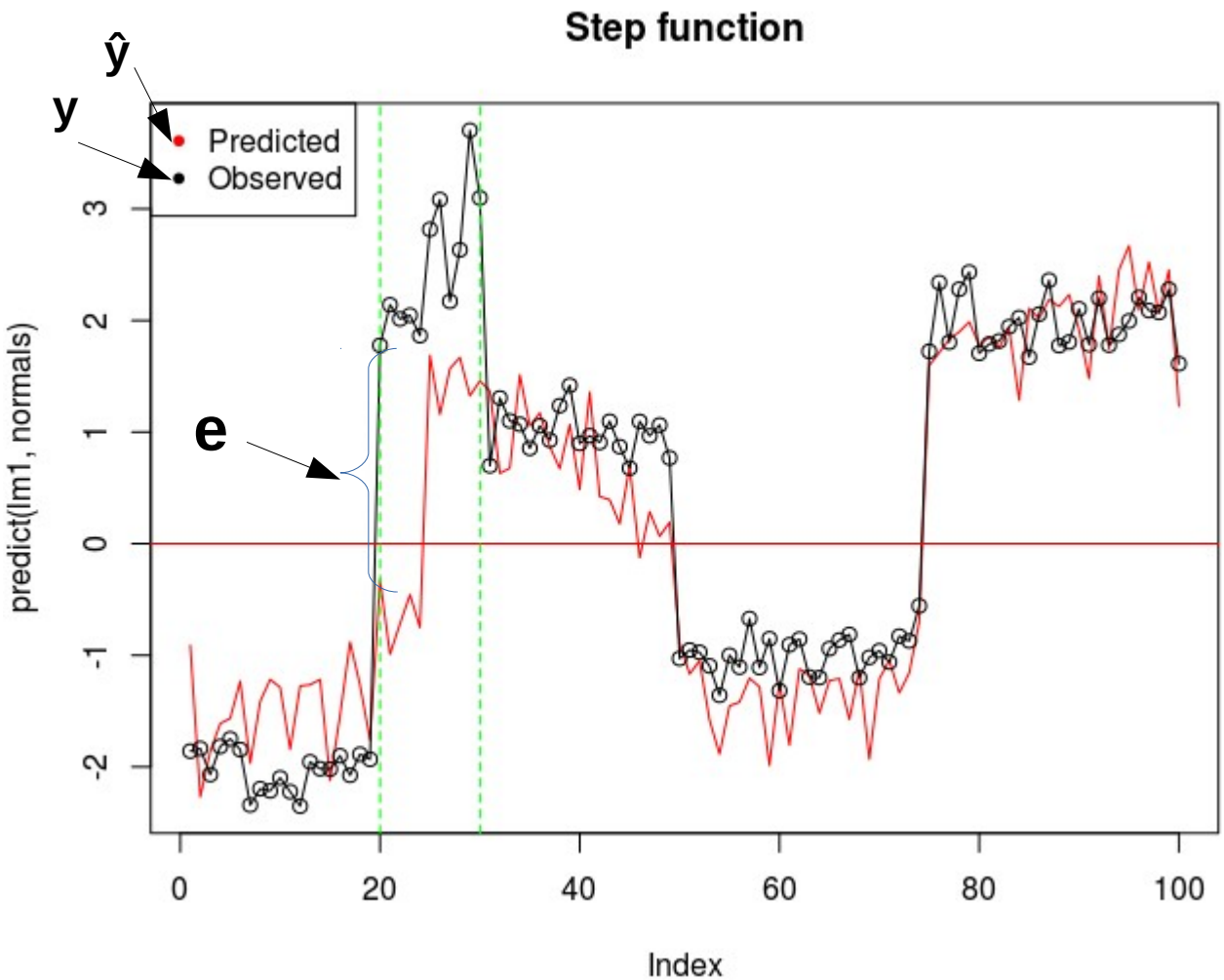
predicted

$$e = y - \hat{y}$$

- **$\hat{y}$ is a linear combination of ALL normals (like a weighted average)**
- **e is the Tangent Normalized value.**

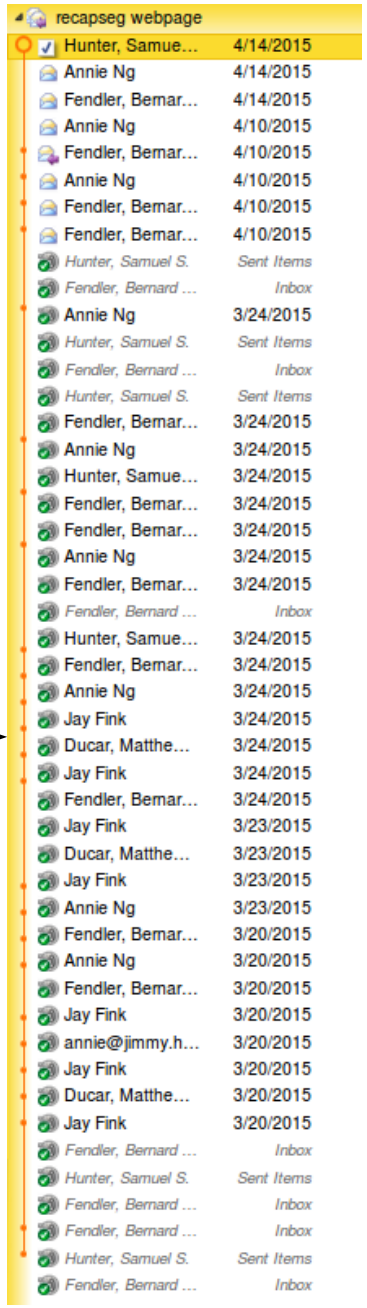
Note:
- A new set of **ß** values are estimated for each tumor sample.
- Tangent normalization assumes that there will be a poor fit at CNV loci (localized large **e** values).
  - A normal sample with a tumor-like profile will defeat this assumption.

CNV: predicted vs observed



Limitations of ReCapSeg

- Relies heavily on PON.
- Large, complicated code base (11,188 lines of Python in 138 files).
  - Relies of R libraries for DNAcopy.
- Unpublished and under development.
- Difficult to install with dozens of dependencies.
- Each step is run separately.
- Limited configuration options
  - No adjustment for sensitivity/specificity



RobustCNV

- Similar strategy to ReCapSeg
  - Iterated re-weighted least squares → more robust to outliers (CNVs) than ordinary least squares
  - Explicit GC normalization step removes most remaining GC bias when PON is poor
  - Pure R code, < 1000 lines
  - ~2x as fast as ReCapSeg
  - Easier to run

QC for normalization

- Quality of normalization dictates quality of CNV calls.
- Optimal:
  - Was sufficient systematic bias was removed?
- Actual:
  - Was systematic bias removed?

QC for normalization

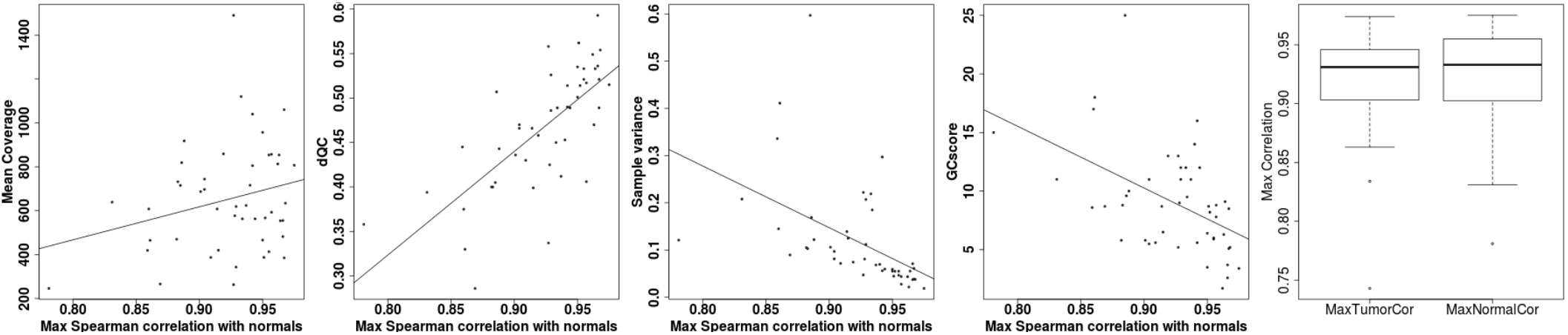
- Metrics
  - MaxNormalCor – maximum Spearman Rank correlation between a tumor sample and any normal sample
  - MaxTumorCor – maximum Spearman Rank correlation between a tumor sample and any other tumor sample
  - MeanCoverage – average mapping coverage
  - dQC – change in average difference between adjacent intervals (before – after → higher is better)
  - Var – variance of post-normalized interval values
  - GCscore – sum of absolute value of loess line fitted to %GC

Assessment of normalization (metrics)

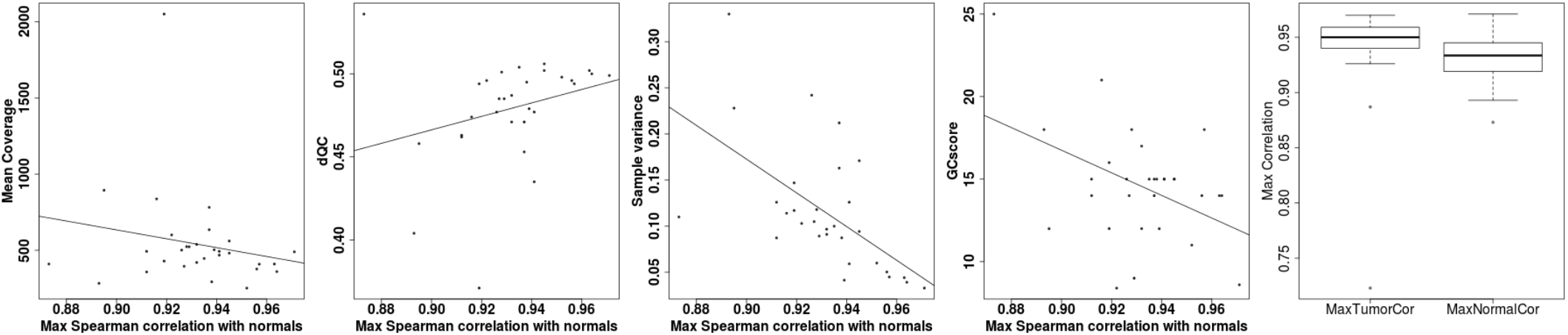
Sample	MaxNormalCor	MaxTumorCor	MeanCoverage	dQC	Var	GCscore
3943-2_L_006659	0.872	0.869	484	0.4	0.354	32
3949-2_L_006660	0.914	0.954	484	0.47	0.115	8
3950-2_L_006661	0.939	0.953	479	0.44	0.299	7.6
3951-2_L_006662	0.929	0.954	419	0.44	0.083	4.8
3959-2_L_006663	0.97	0.963	419	0.45	0.028	9.3
3963_L_006576	0.927	0.936	631	0.49	0.083	6.3
3964_L_006577	0.961	0.97	407	0.45	0.028	13
3966_L_006579	0.854	0.868	290	0.41	0.281	9.3
3968_L_006581	0.931	0.948	342	0.43	0.075	8.3
3969_L_006583	0.95	0.954	375	0.46	0.03	10
3970_L_006584	0.91	0.908	260	0.44	0.187	11
3971_L_006585	0.881	0.872	342	0.44	0.158	11
3972_L_006586	0.829	0.91	352	0.48	0.143	68
3974_L_006588	0.921	0.936	504	0.51	0.095	9
3976_L_006590	0.972	0.974	495	0.49	0.018	9.7
3977_L_006592	0.904	0.911	367	0.51	0.105	19
3980_L_006593	0.97	0.974	495	0.47	0.034	9.7
3981_L_006594	0.941	0.923	488	0.49	0.084	21
3982_L_006595	0.95	0.944	467	0.46	0.053	8.1
3984_L_006596	0.857	0.923	348	0.53	0.183	48
3986_L_006598	0.978	0.971	604	0.5	0.02	6.9

QC plots

POPv1 Validation Dataset

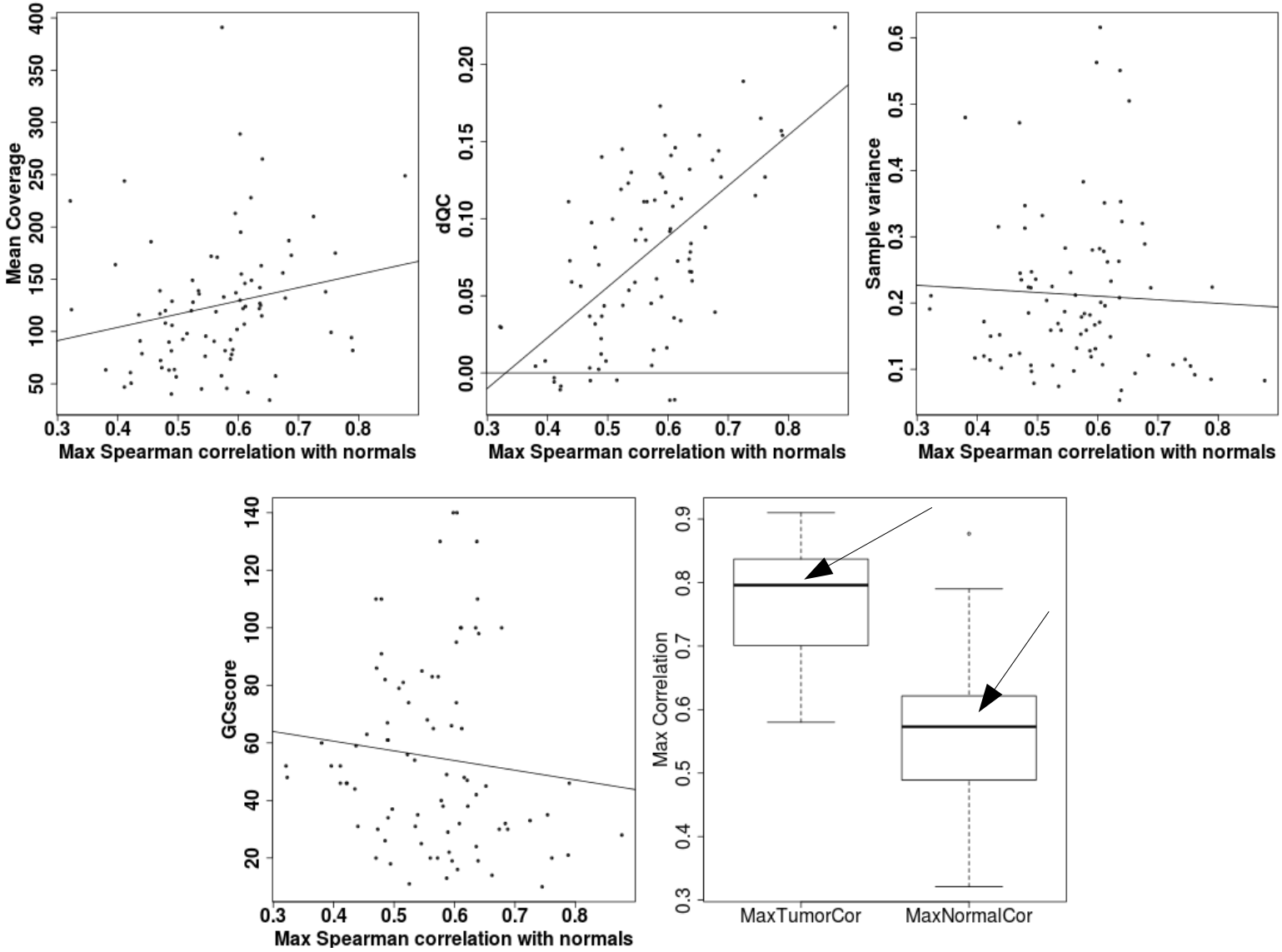


POPv2 validation dataset

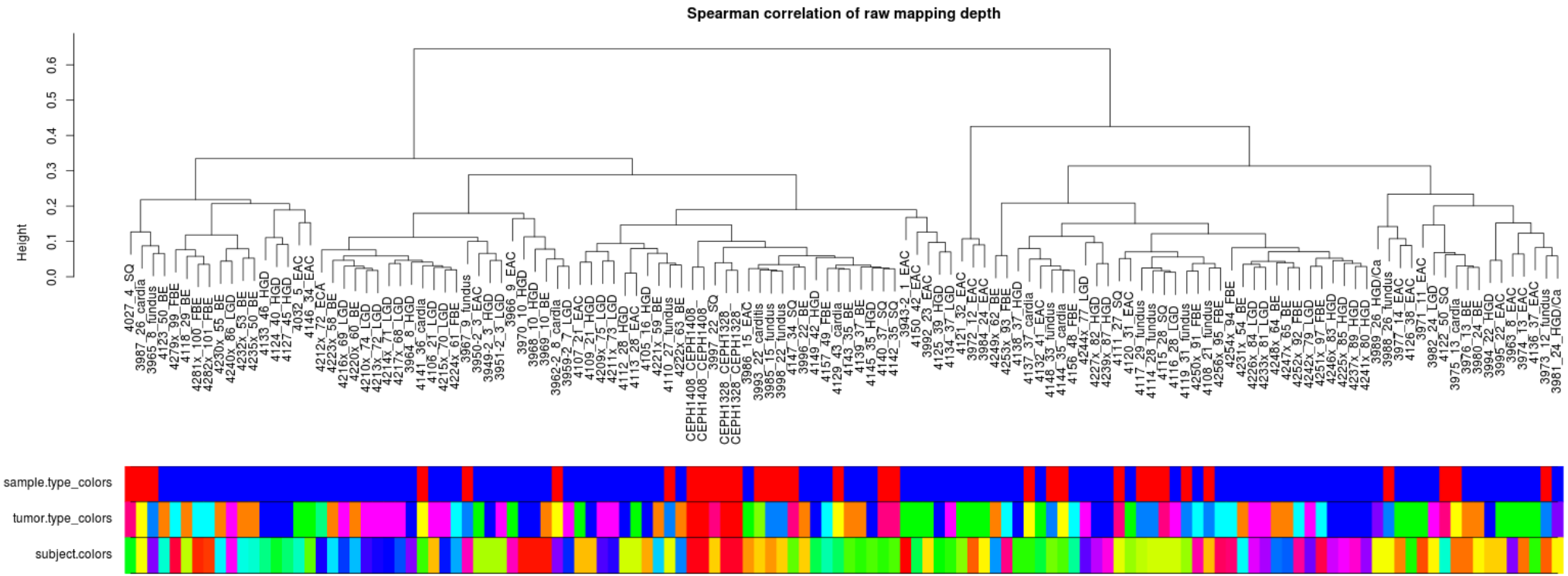


QC plots

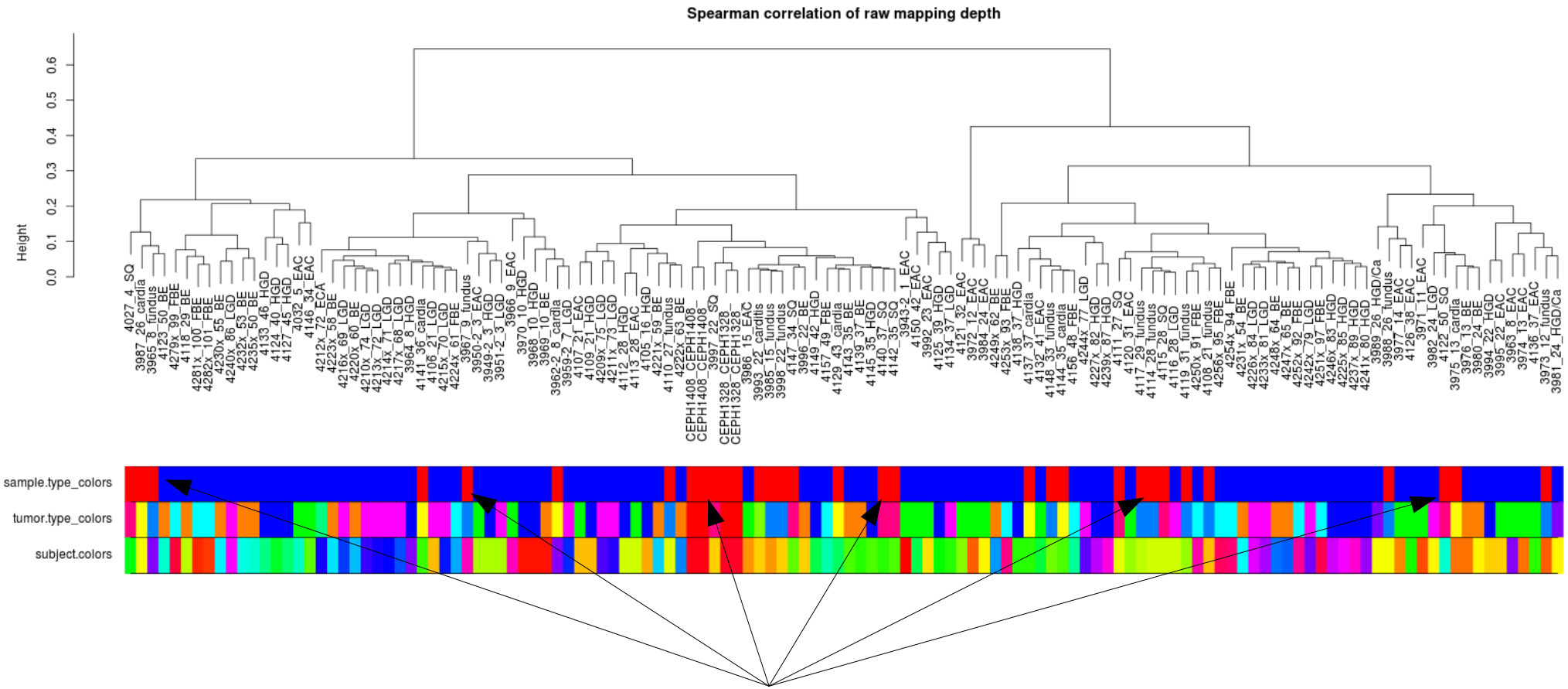
CCGD project w/poor normals



QC plots


Spearman correlation of raw mapping depth
Height
0.0
0.1
0.2
0.3
0.4
0.5
0.6
sample.type_colors
tumor.type_colors
subject.colors

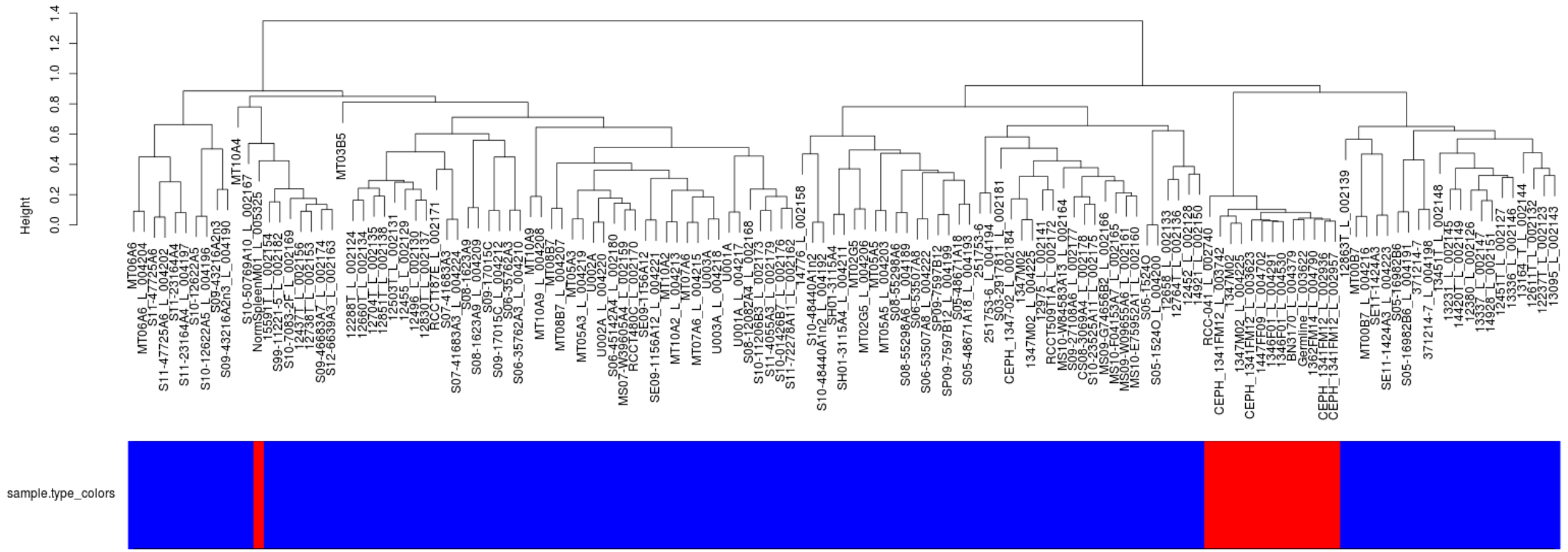
QC plots



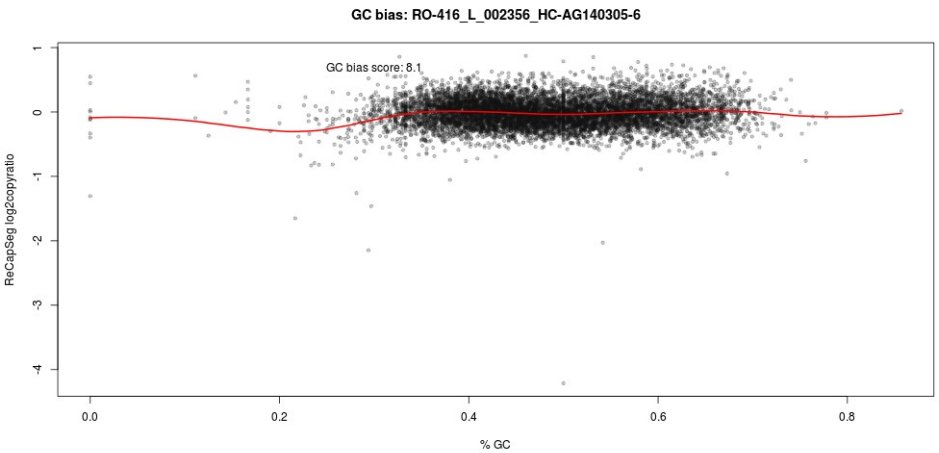
For good normalization, normal samples should be mixed with tumor samples.

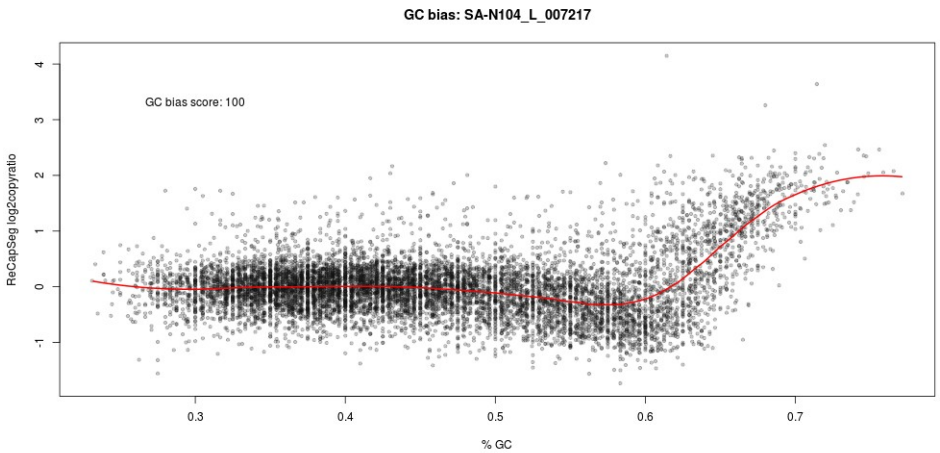
QC plots

CCGD project w/poor normals



QC plots

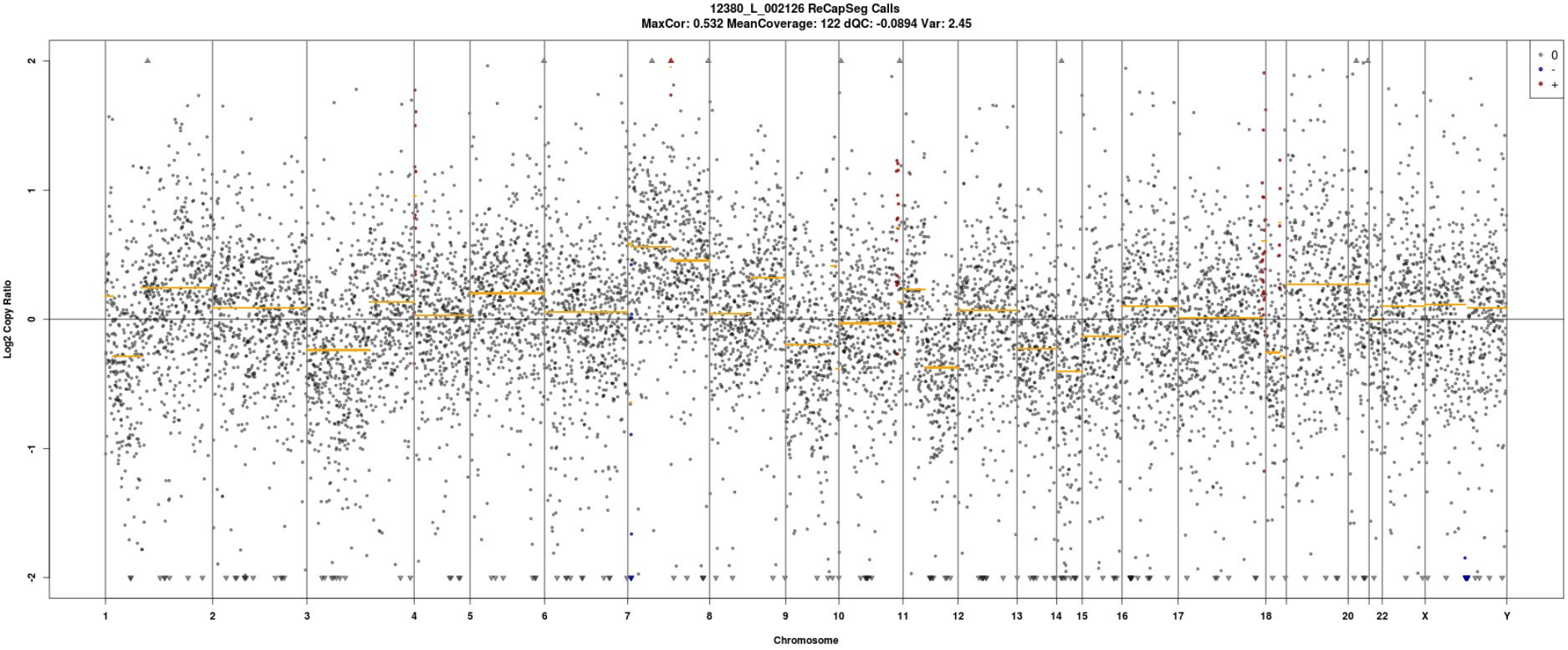




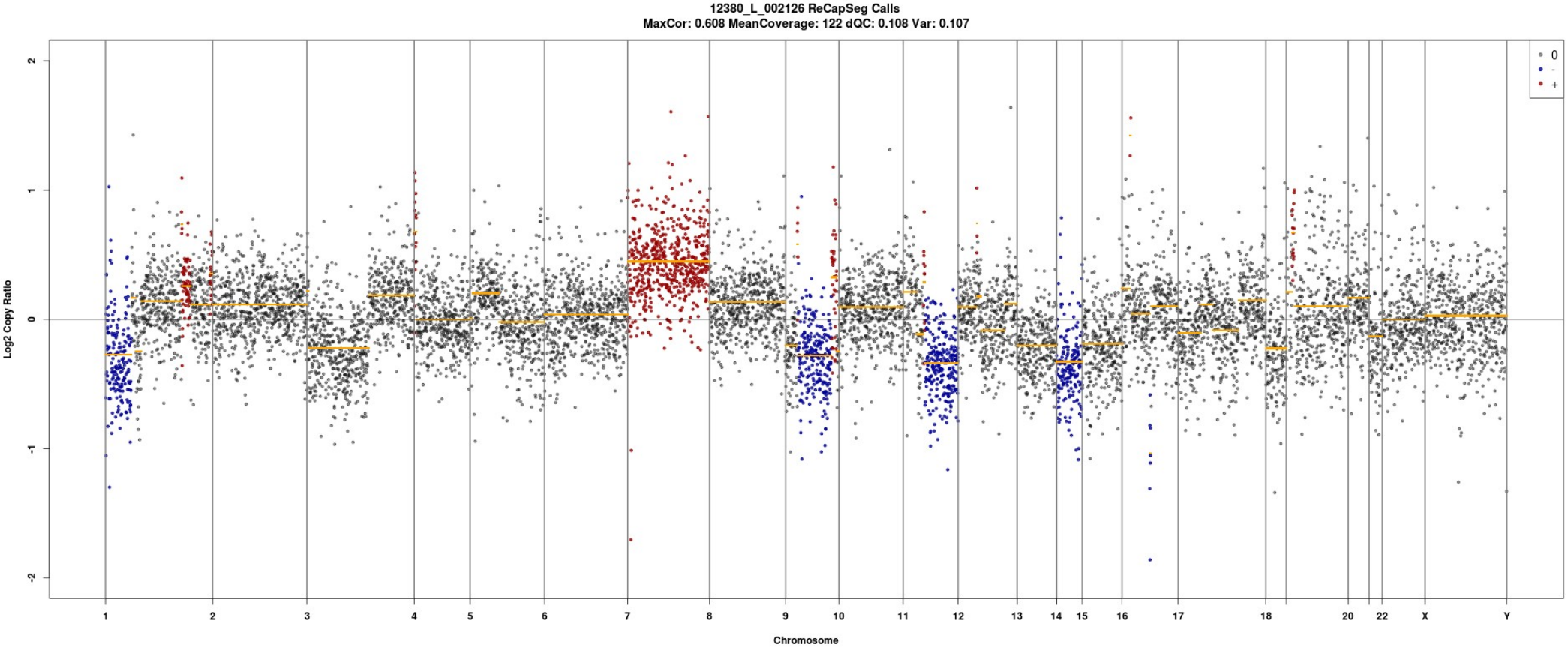
Well normalized samples have ~0 GC bias and a low GCscore.

Poorly normalized samples have remaining GC bias and a high GCscore.

Normals Matter



Normals Matter



QC conclusions

- We have developed a variety of metrics to assess the quality of the normalization.
- Currently CCGD looks for a MaxNormalCor >= 0.8 as a minimum cutoff.
- Good normals matter.

Assessment and validation for CCGD and Profile

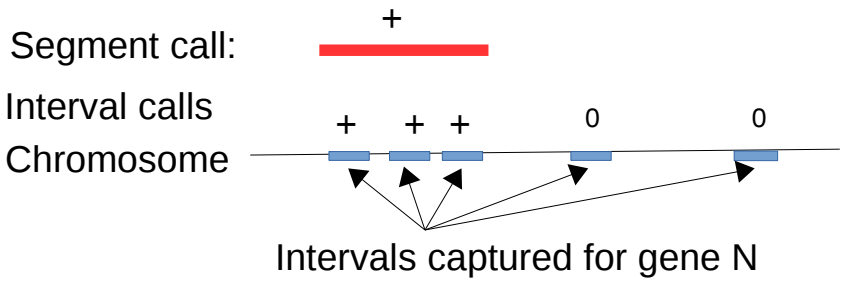
- Objectives
  - Determine whether new algorithms perform as well, or better than the existing approach (VisCap).
  - Measure performance against array-Competitive Genomic Hybridization (aCGH) calls.
  - Identify and implement strategies to further improve performance of these algorithms.

Methods

- Datasets
  - Profile Oncopanel V1 (POPv1)
    - 47 aCGH matched samples
  - Profile Oncopanel V2 (POPv2)
    - 30 aCGH matched samples
- Analysis
  - aCGH calls made with Nexus Copy Number software from BioDiscovery.
  - VisCap calls were generated according to settings currently used by Profile group.
  - ReCapSeg calls were generated using default settings except that the top 5% most noisy intervals were discarded (as calculated from normals, default 25%).
  - RobustCNV calls were generated using default setting with no intervals discarded.
  - Calls for X and Y chromosomes were excluded from analysis.
- Comparisons
  - Using coverage calculated from Targeted intervals definition.
    - Other interval strategies were also considered.
  - Per-interval and Per-gene.

Methods
Segments → Interval → Gene

- All callers produce segment calls
  - e.g. chr1:5000-20000 +
- To facility interval-level comparisons, intervals are assigned calls based on their intersection with segments



Gene Call: 3:+, 2:0 → Gain

Methods
Rules for Gene-level calls

Rule	Gene-level Call
'-' call > 2 times and and '+' > 50%	'gain+loss'
'-' call > 2 or is 100%	'loss'
'+' call > 50%	'gain'
'+' and '-' calls in the same gene (below threshold)	'mixed'
'+' calls but below threshold	'Normal+'
'-' calls but below threshold	'Normal-'
No + or - calls	'Normal'

Methods
Sensitivity & Specificity Calculation

- 0 → normal copy, + → gain, - → loss
- Sensitivity, Specificity framework:
  - positive = all non-0 calls
  - negative = all 0 calls
  - TP = (-, +) call when condition is (-, +) *****
  - FP = (-, +) call when condition is (+, -) ***** or 0
  - FN = 0 call when condition is - or +
  - TN = 0 call when condition is 0

Note: condition = ACGH
***** respectively

Gene Level: Sensitivity & Specificity Calculation

- Sensitivity, Specificity framework:
  - *negative* = {NormalCopy, NormalCopy+, NormalCopy-mixed}
  - *positive* = {gain, gain+loss, loss}
  - TP when Caller_call == aCGH_call and aCGH_call is not in *negative*
  - FP when Caller_call in *positive* and aCGH_call in *negative*
  - FN when Caller_call in *negative* and aCGH_calls in *positive*
  - TN = Call_calls in *negative* and aCGH_calls in *negative*

POPv1: ACGH comparison (summarized across 47 samples)

ReCap calls

ReCapCalls	aCGHCall −	aCGHCall +	aCGHCall 0
−	14371	294	8599
+	14	17452	6673
0	2057	8099	157513

VisCap Calls

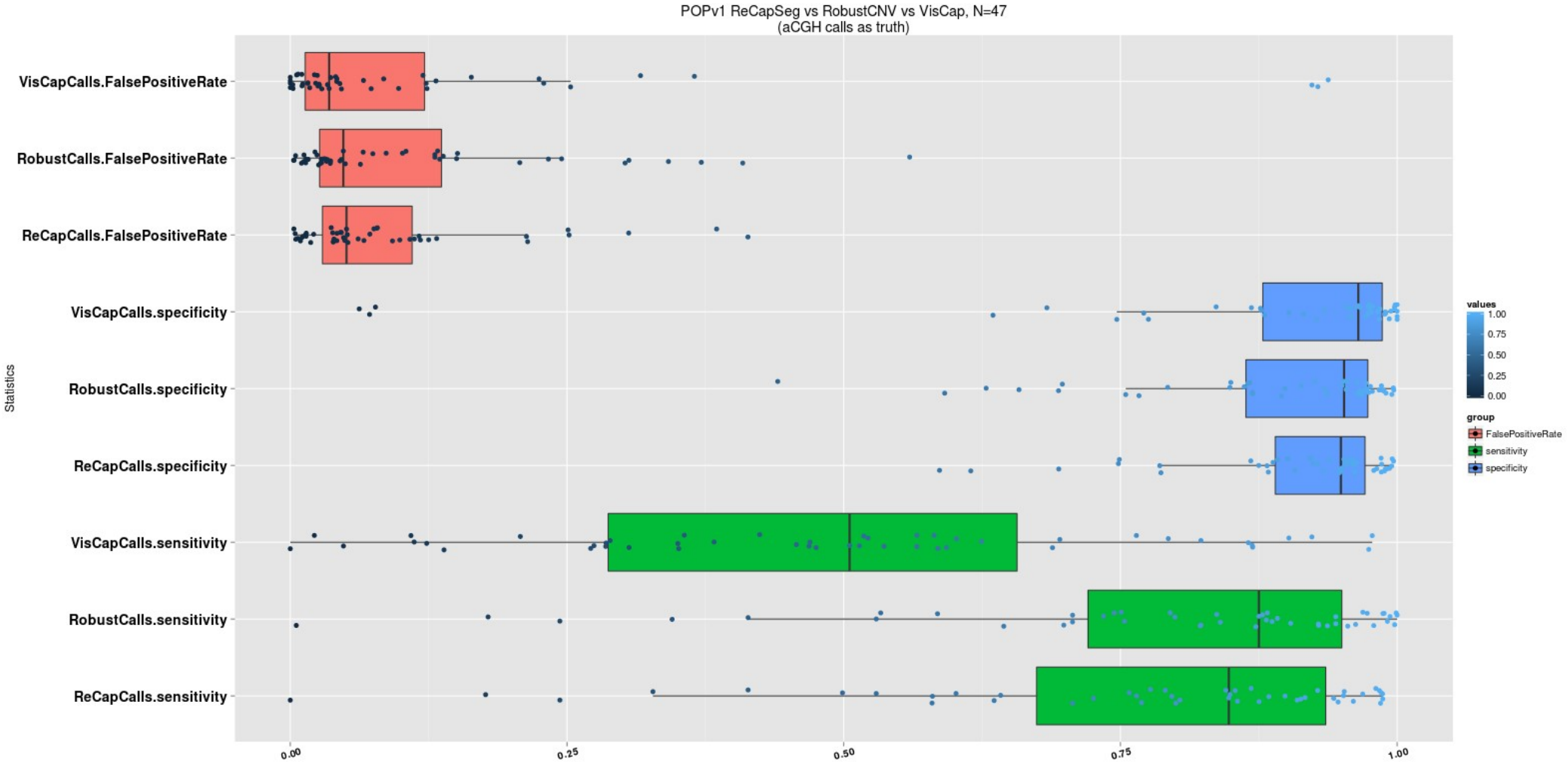
VisCapCalls	aCGHCall −	aCGHCall +	aCGHCall 0
−	8101	60	13063
+	32	11566	8401
0	8309	14219	151321

RobustCNV calls

RobustCalls	aCGHCall −	aCGHCall +	aCGHCall 0
−	14738	303	10593
+	12	18633	7528
0	1692	6909	154664

*Statistics represent calls per-sample summarized to the interval level.

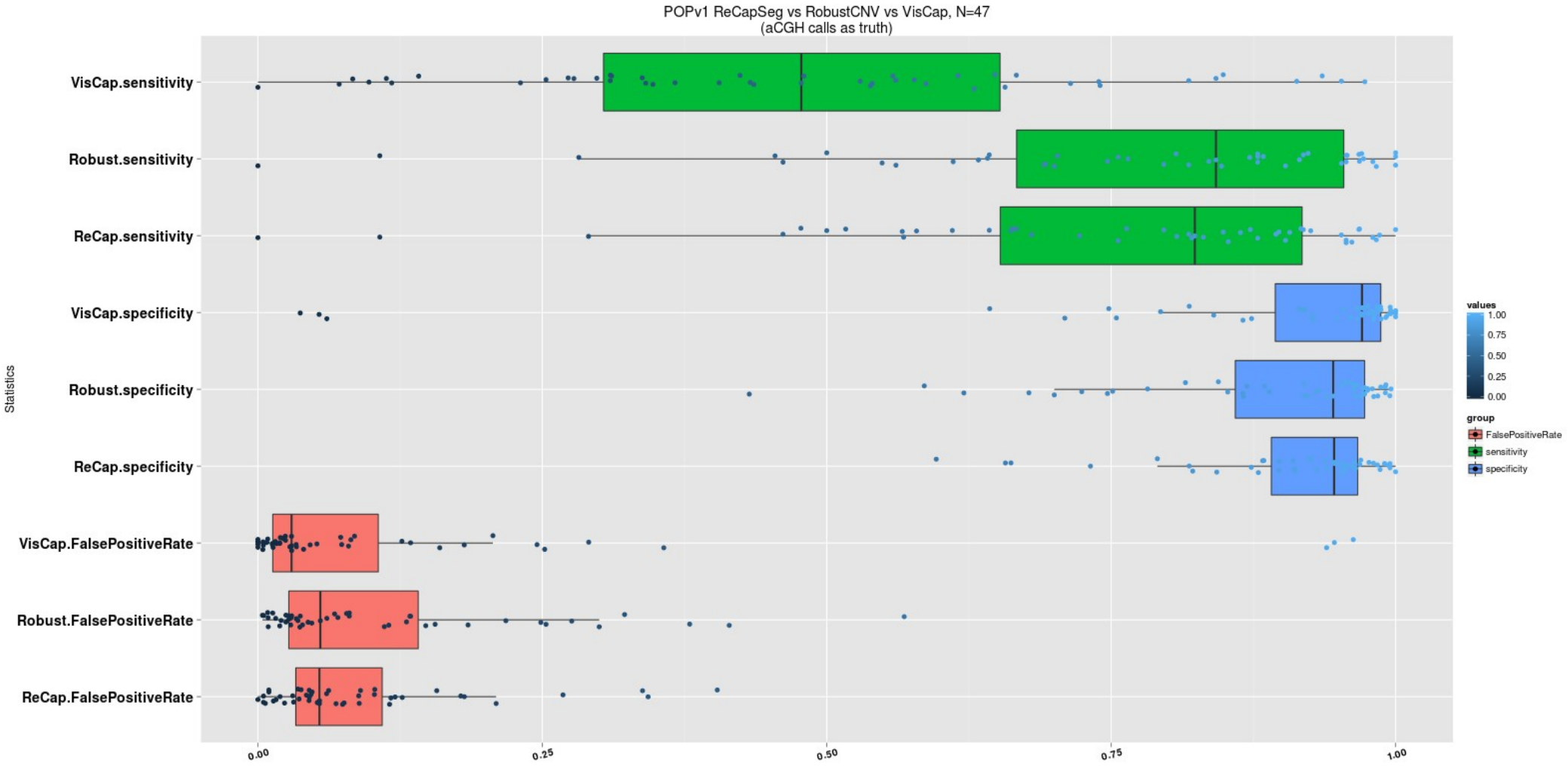
Results: POPv1: intervals



Sensitivity = TP / (TP + FN)	specificity = TN / (FP + TN)	FPR = FP / (FP + TN)

*Statistics represent calls per-sample summarized to the interval level.

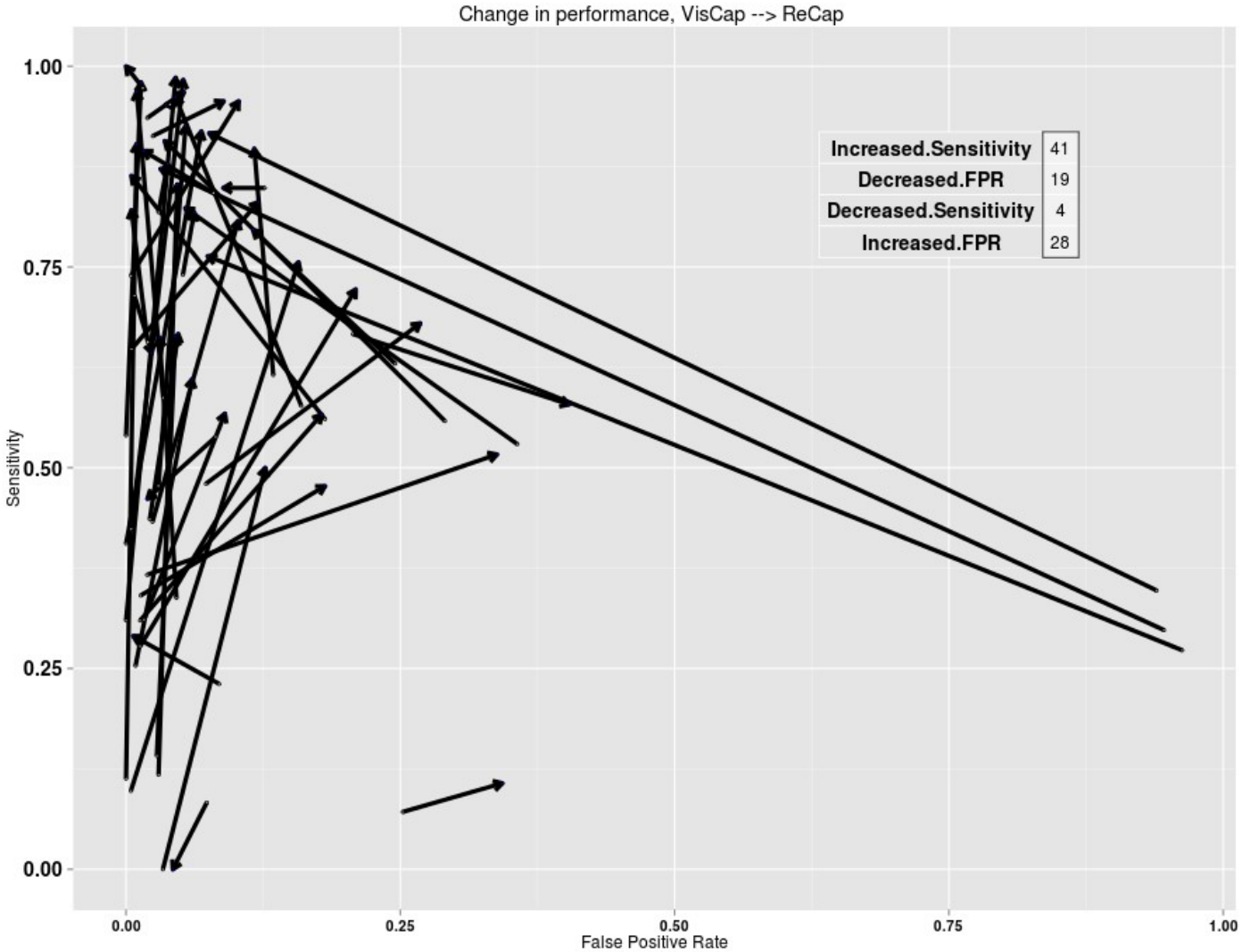
Results: POPv1: genes



Sensitivity = TP / (TP + FN)	specificity = TN / (FP + TN)	FPR = FP / (FP + TN)

*Statistics represent calls per-sample summarized to the interval level.

Results: POPv1: genes



Sensitivity = TP / (TP + FN)	specificity = TN / (FP + TN)	FPR = FP / (FP + TN)

*Statistics represent calls per-sample summarized to the gene level.

POPv2: ACGH comparison (summarized across 30 samples)

ReCap calls

ReCapCalls	aCGHCall -	aCGHCall +	aCGHCall 0
-	9853	26	3840
+	14	10983	4019
0	1825	4520	131750

VisCap Calls

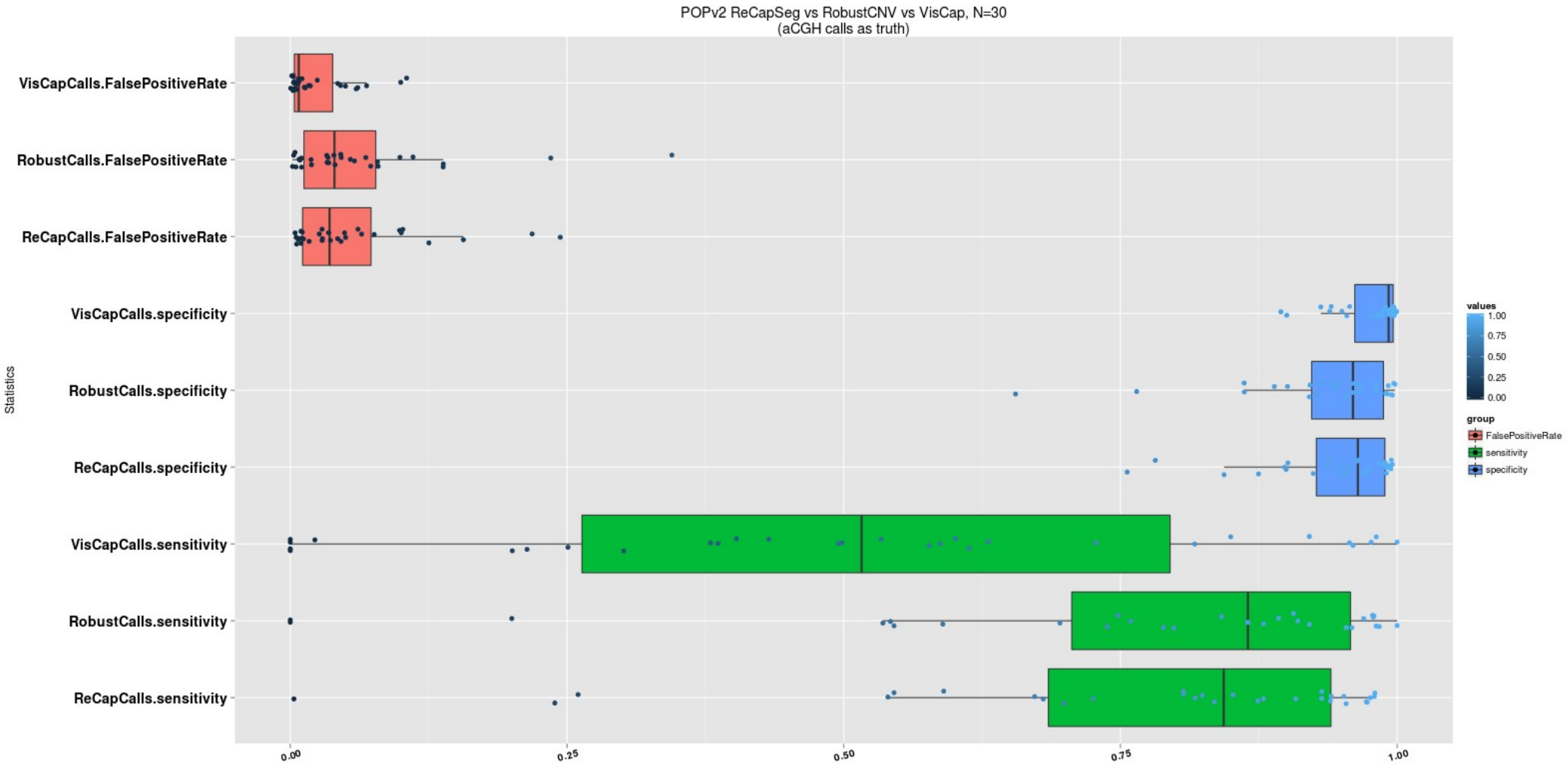
VisCapCalls	aCGHCall -	aCGHCall +	aCGHCall 0
-	5429	13	562
+	12	7953	2656
0	6251	7563	136391

RobustCNV calls

RobustCalls	aCGHCall -	aCGHCall +	aCGHCall 0
-	9727	36	4082
+	8	11378	4507
0	1957	4115	131020

*Statistics represent calls per-sample summarized to the interval level.

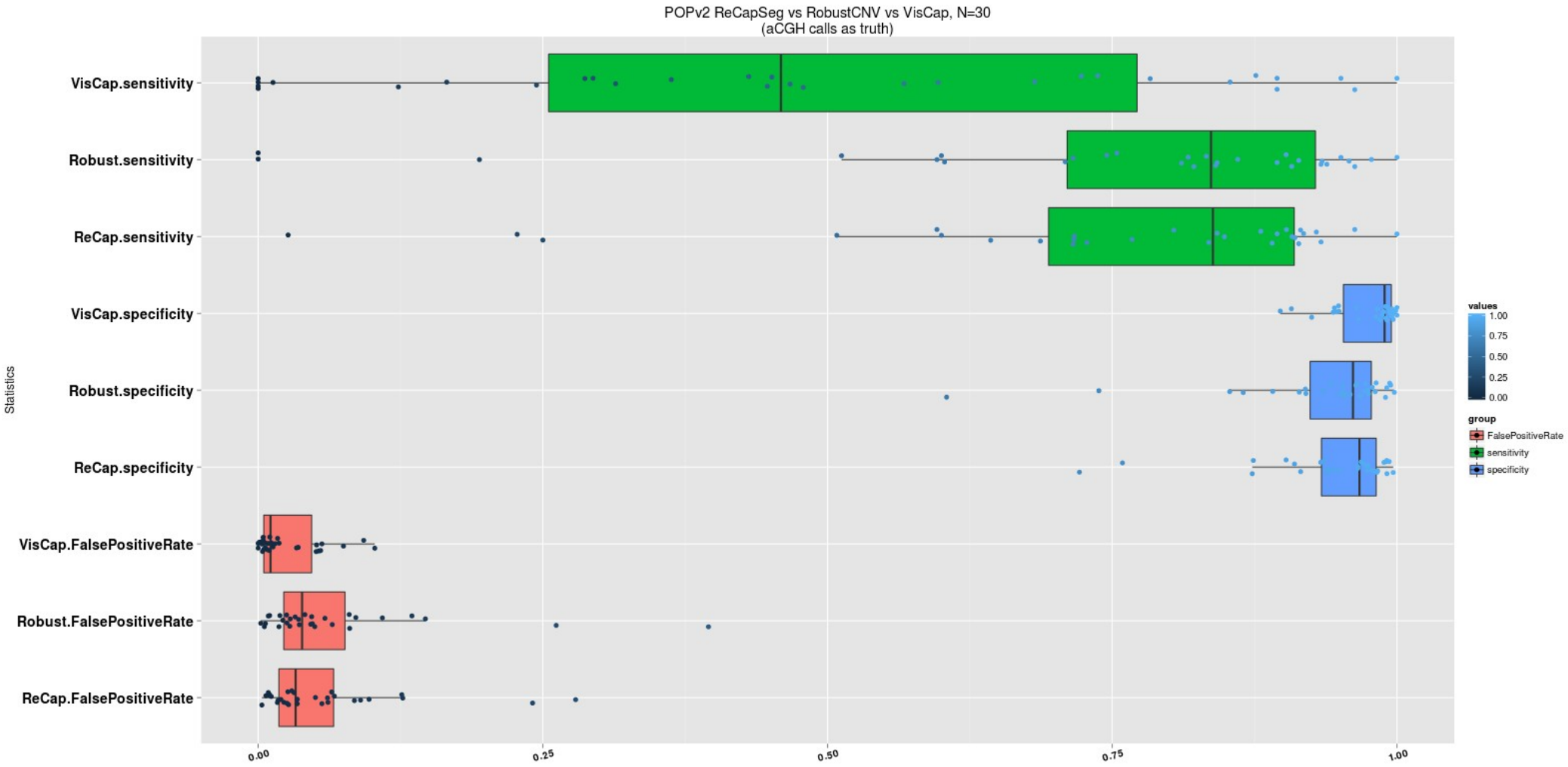
Results: POPv2: intervals



Sensitivity = TP / (TP + FN)	specificity = TN / (FP + TN)	FPR = FP / (FP + TN)

*Statistics represent calls per-sample summarized to the interval level.

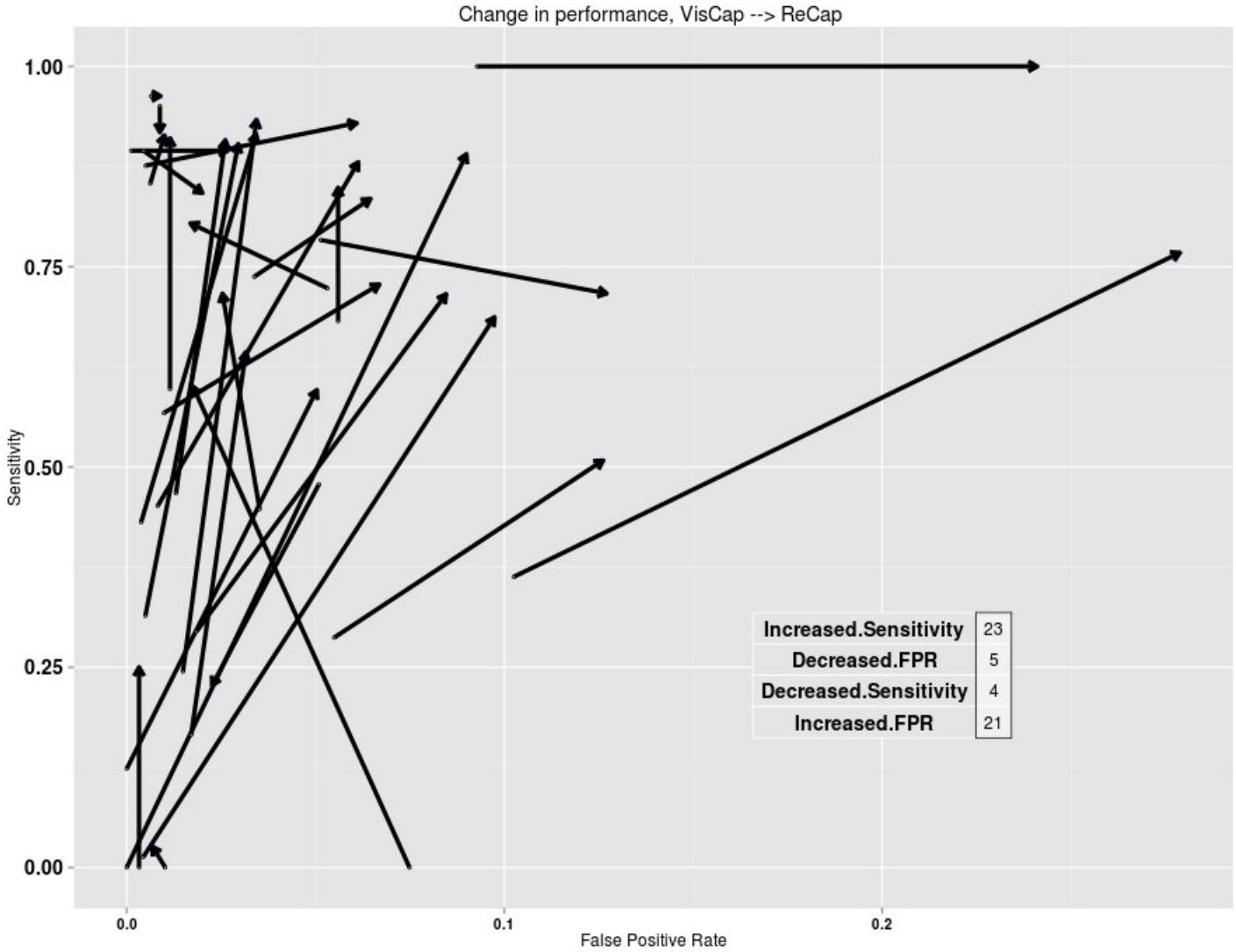
Results: POPv2: genes



Sensitivity = TP / (TP + FN)	specificity = TN / (FP + TN)	FPR = FP / (FP + TN)

*Statistics represent calls per-sample summarized to the interval level.

Results: POPv2: genes



Sensitivity = TP / (TP + FN)	specificity = TN / (FP + TN)	FPR = FP / (FP + TN)

*Statistics represent calls per-sample summarized to the gene level.

Additional Considerations

- Focal gains/losses
  - 1-2 aCGH called intervals flanked by normal calls
- Tumor suppressor genes
  - SMAD2, SMAD4, ATM, RB1, TP53, PTEN, CDKN2A, CDKN2B

POPv1: Focal event ACGH comparison (summarized across 47 samples)

ReCap calls

```
                    aCGHCall
ReCapCalls    -    +
         -   18    0
         +    0   16
         0   22  126

18.7% called
```

VisCap Calls

```
             aCGHCall
VisCapCalls    -    +
          -    2    6
          +    1   13
          0   37  123

8.2 % called
```

RobustCNV calls

```
                     aCGHCall
RobustCalls    -    +
           -   21    3
           +    0   14
           0   19  125

19.2% called
```

*Statistics represent calls per-sample summarized to the interval level.

POPv2: Focal event ACGH comparison (summarized across 30 samples)

ReCap calls

	aCGHCall	
ReCapCalls	-	+
-	60	6
+	6	77
0	134	313

23% called

VisCap Calls

	aCGHCall	
VisCapCalls	-	+
-	5	4
+	5	21
0	190	371

4.4 % called

RobustCNV calls

	aCGHCall	
RobustCalls	-	+
-	61	2
+	4	81
0	135	313

23.8% called

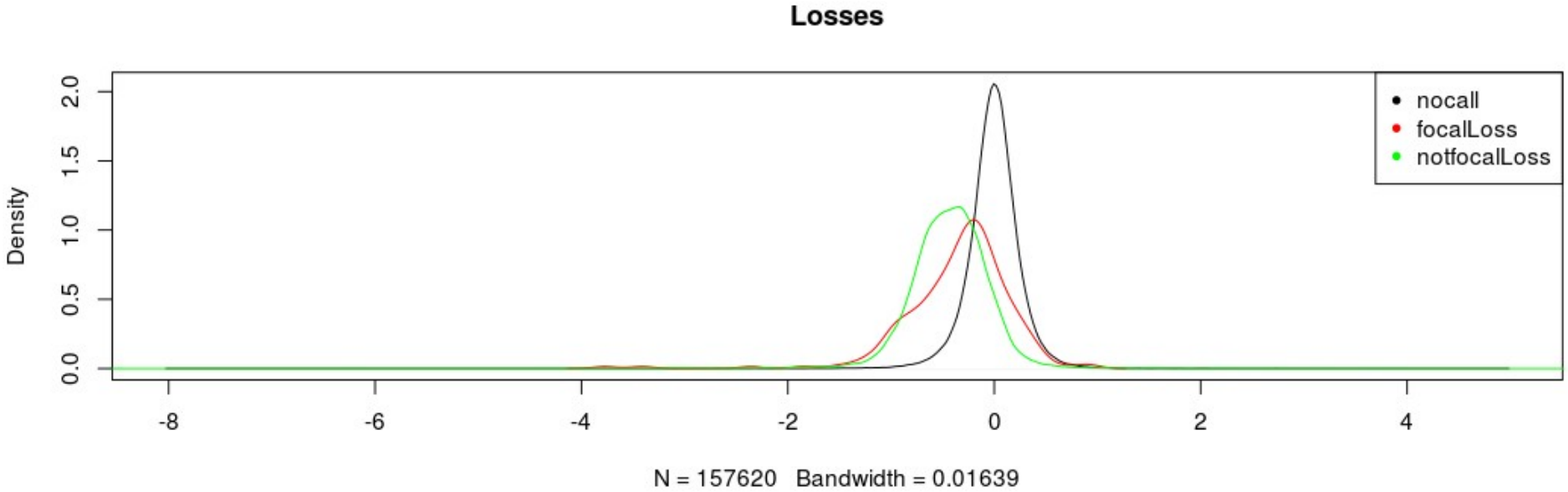
RobustCNV calls (customized intervals)

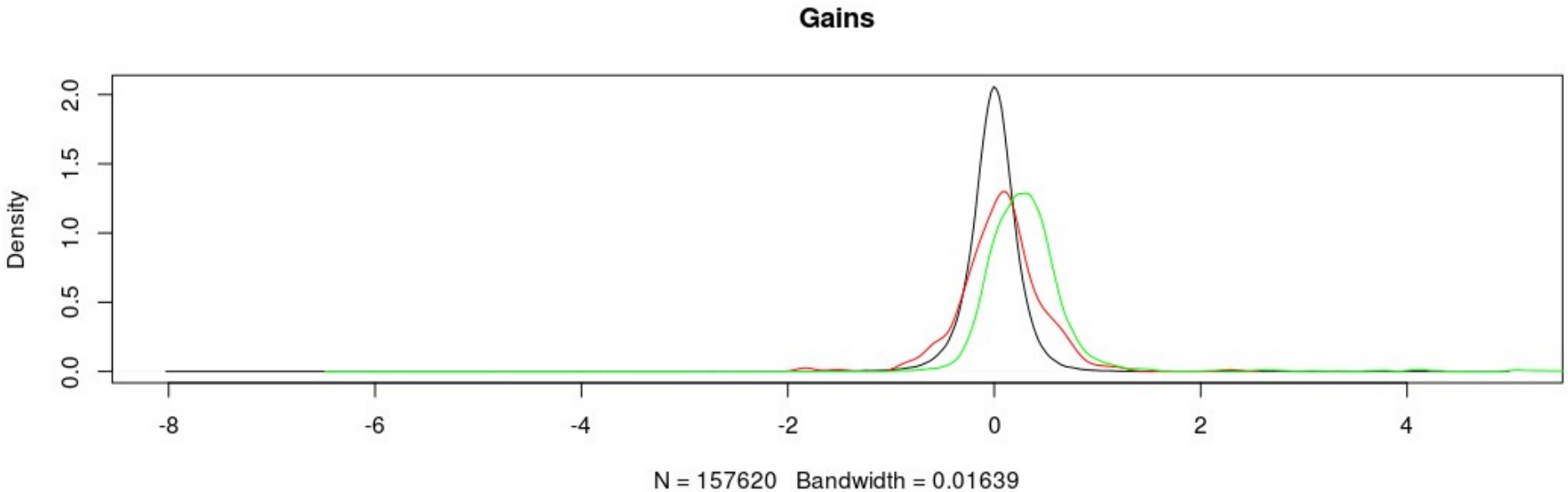
	aCGHCall	
RobustCalls	-	+
-	70	3
+	4	86
0	126	307

26.2% called

*Statistics represent calls per-sample summarized to the interval level.

Focal events are hard to call





POPv1: Tumor suppressor genes**
(summarized across 47 samples)

ReCap calls

ReCapSeg	aCGH gain	loss	NormalCopy	NormalCopy-	NormalCopy+
gain	**18**	0	8	0	0
loss	0	**87**	19	1	0
NormalCopy	15	6	**210**	0	10
NormalCopy+	0	1	1	0	0

RobustCNV calls

RobustCNV	aCGH gain	loss	NormalCopy	NormalCopy-	NormalCopy+
gain	**18**	0	11	0	0
gain+loss	0	0	1	0	0
loss	0	**85**	18	1	0
NormalCopy	15	7	**205**	0	10
NormalCopy-	0	2	0	0	0
NormalCopy+	0	0	3	0	0

VisCap Calls

VisCap	aCGH gain	loss	NormalCopy	NormalCopy-	NormalCopy+
gain	**15**	0	11	0	0
loss	0	**58**	19	0	0
NormalCopy	18	34	**207**	1	10
NormalCopy-	0	2	0	0	0
NormalCopy+	0	0	1	0	0

*Statistics represent calls per-sample summarized to the gene level.
** SMAD2, SMAD4, ATM, RB1, TP53, PTEN, CDKN2A, CDKN2B

POPv2: Tumor suppressor genes**
(summarized across 30 samples)

ReCap calls

	aCGH		
ReCapSeg	gain	loss	NormalCopy
gain	**5**	0	4
loss	0	**48**	5
NormalCopy	5	9	**163**
NormalCopy-	0	0	1

VisCap Calls

	aCGH		
VisCap	gain	loss	NormalCopy
gain	**3**	0	2
loss	0	**34**	3
NormalCopy	7	23	**166**
NormalCopy+	0	0	2

RobustCNV calls

	aCGH		
RobustCNV	gain	loss	NormalCopy
gain	**6**	0	6
loss	0	**46**	7
NormalCopy	4	11	**160**

*Statistics represent calls per-sample summarized to the gene level.
** SMAD2, SMAD4, ATM, RB1, TP53, PTEN, CDKN2A, CDKN2B

Assessment and validation for CCGD and Profile

- Conclusions
  - Determine whether ReCapSeg performs as well, or better than the existing approach (VisCap).
    - ReCapSeg is much more sensitive than VisCap, but has a slightly higher FDR.
  - Measure performance against array-Competitive Genomic Hybridization (aCGH) calls.
    - ReCapSeg has a median sensitivity of ~80% on both test datasets (VisCap ~50%)
  - Identify and implement strategies to further improve performance of these algorithms.
    - Modifications to intervals show marginal improvements.

Questions