

Impact of centralization on aCGH-based genomic profiles for precision medicine in oncology

F. Commo^{1,2,#}, C. Ferte^{1,2,3,#}, JC. Soria^{2,3}, S.H. Friend¹, F. André^{2,3}, J. Guinney¹

1) Sage Bionetworks, Seattle, WA, USA

2) INSERM U981, Gustave Roussy, University Paris XI, Villejuif, France

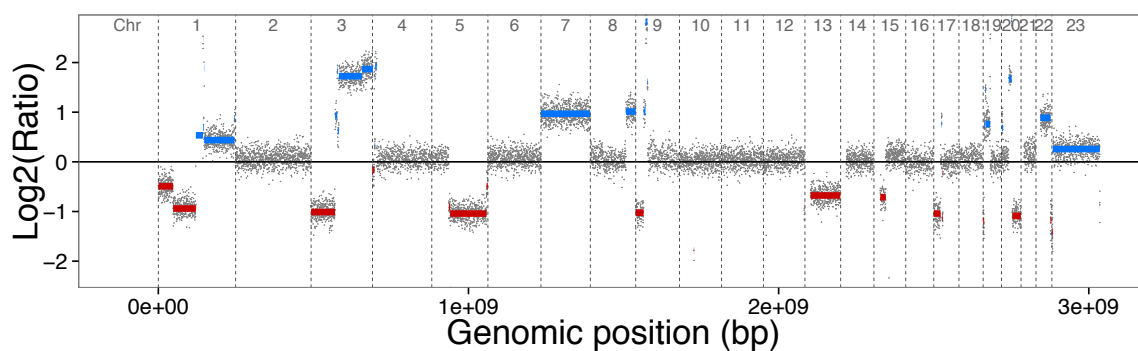
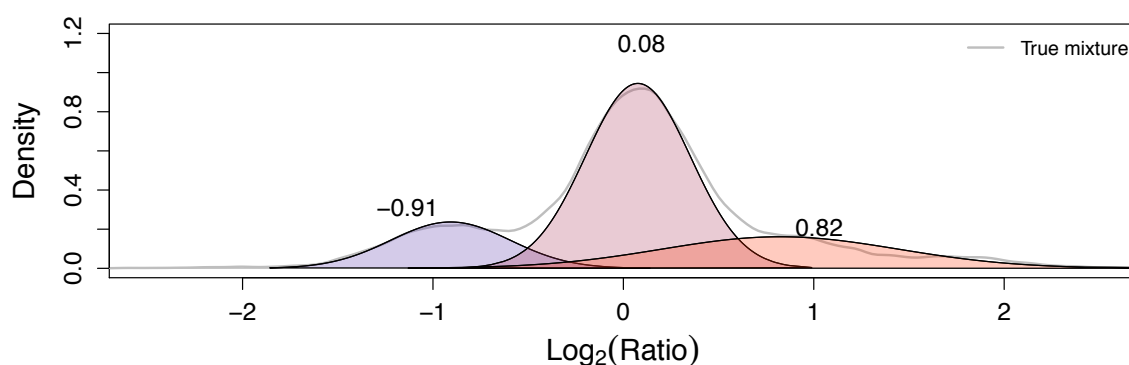
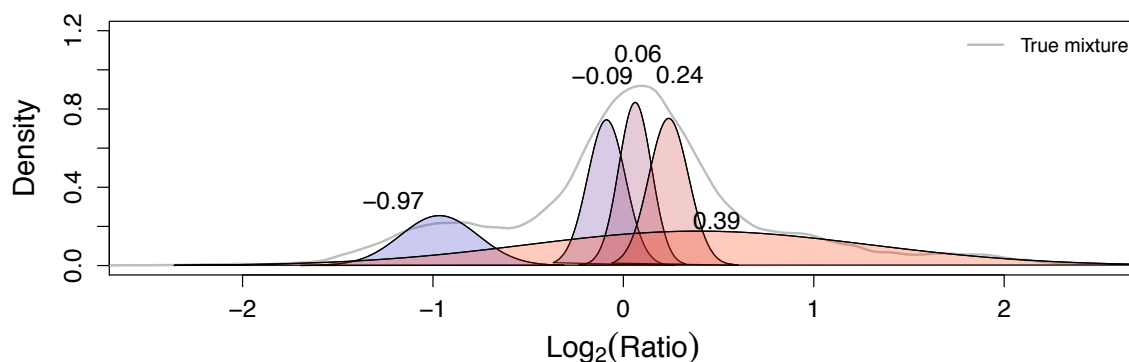
3) Department of Medical Oncology, Gustave Roussy, Villejuif, France

These authors contributed equally to this work.

Supplementary methods and results:

The EM algorithm - EM stands for Expectation-Maximization - is well known and widely used for modeling mixture of normally distributed populations, and extracting their respective parameters. Unfortunately, EM does not perform well on very large and noisy vectors such as CGH $\text{Log}_2(\text{Ratios})$ computed from hundreds of thousands of probes. On such values, this method is time consuming, and not well adapted for detecting centralization peaks, since EM tends to over estimate (or underestimate) the number of sub-populations (figure 1).

Figure 1: The EM algorithm was applied on real data (Safir01 sample). On its original form (top), the method considers the central population as itself a mixture of 3 populations (means: -0.09, 0.062 and 0.23, respectively), which is probably true, but maybe not relevant in that context. Instead, the adapted resampling method (center) identifies one unique peak, and simplifies the decision for centering the entire vector.



To significantly reduce the computation time, as well as to improve the efficacy of the EM algorithm in that particular context, we adapted the algorithm as follow:

A $\text{Log}_2(\text{Ratios})$ is modeled as a mixture of Gaussian distributions, using $1e3$ values randomly picked, instead of using the entire vector of values. The procedure is repeated 100 times, independently. Then, means and variances of each population in the mixture are averaged, considering only the models for which the number of groups corresponds to the median of groups detected over all the models.

The performances of the procedure were estimated on simulated Gaussian mixtures of various total lengths, N from $1e3$ to $1e6$, with 3 arbitrary components, C_i $i=1$ to 3 , each with different means (μ), standard deviations (σ) and proportions (p):

$$C_1 \sim N(-0.58, 0.25); p_1 = 0.15$$

$$C_2 \sim N(0, 0.5); p_2 = 0.7$$

$$C_3 \sim N(0.58, 0.75); p_3 = 0.15$$

The choice of the different parameters was arbitrary, but consistent with mixtures observed in real situations.

This simulations showed that the EM computation time increases exponentially with the size of vector, while the resampling method increases almost linearly, and becomes advantageous on large data, namely $N > 2e5$. This value roughly corresponds to the length of a vector of $\text{Log}_2(\text{Ratios})$ generated from Agilent 4x180K microarrays (figure2). Moreover, when N is large, $N > 2e5$, the EM algorithm tends to split some of the sub-components itself into separate populations, and is unable to correctly estimate the mixture and its parameters. Under the same conditions, repeated random resamplings lead to better estimation of the mixture, close to the expected results (figure 3 & 4).

Figure 2: The EM computation time increases exponentially with the length of the vector (black line). The resampling approach is more stable, and becomes advantageous on very large data ($N > 2e5$) (red line).

Computation time

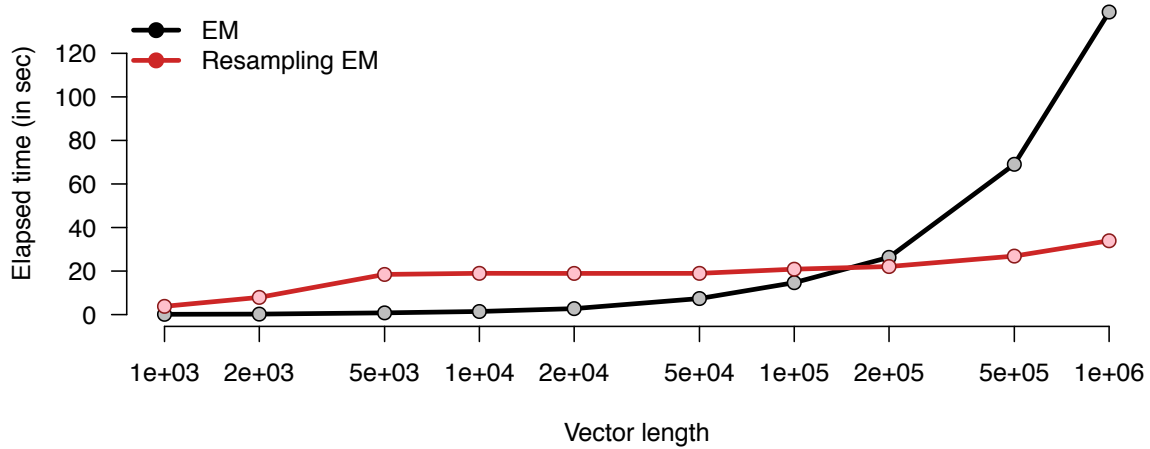


Figure 3: a mixture of 3 normally distributed vectors was simulated. When applying the original EM, the ability of detecting the right number of groups depends on the size of the entire vector (black line), while the resampling approach always detect the correct number of sub-populations, independently of the size of the data (red line).

Number of groups

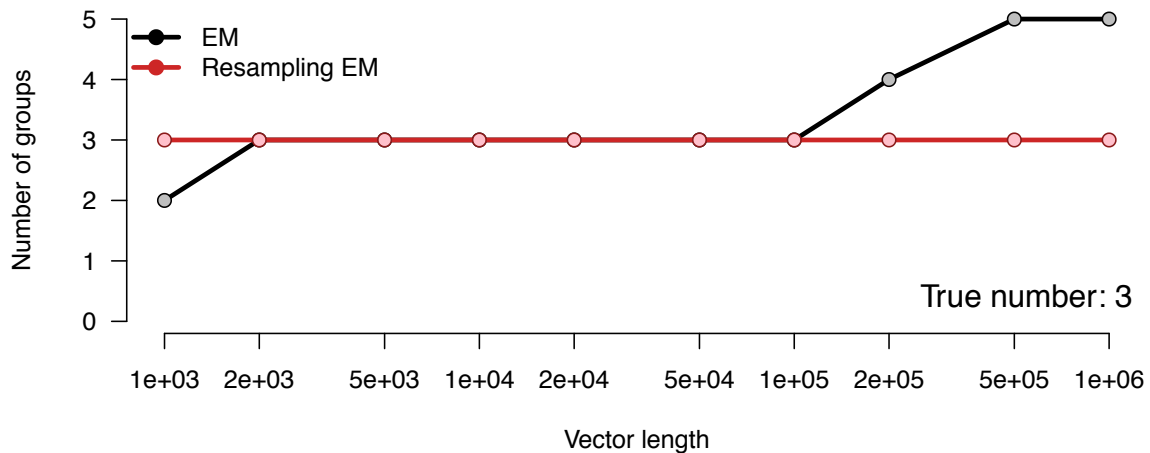
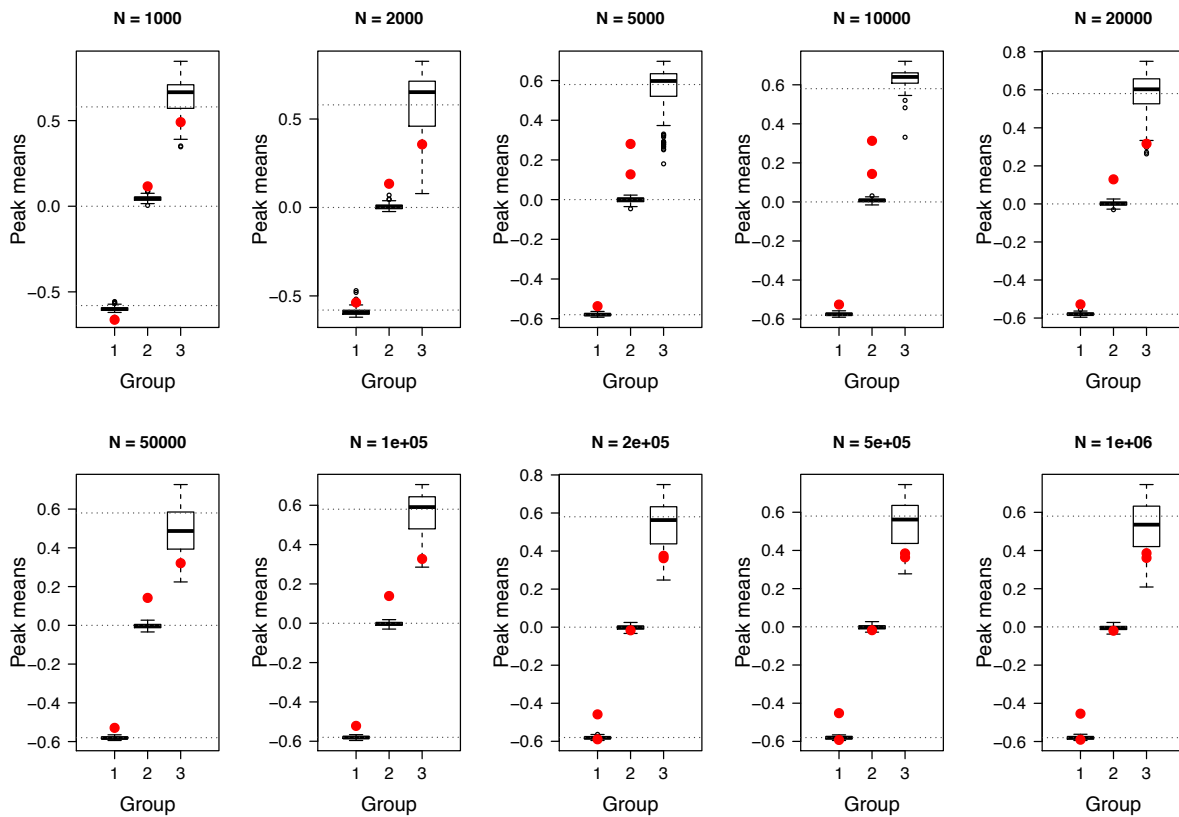


Figure 4: On the same 3-groups mixture simulation, the original EM fails to properly identify the means of each group (grey dash lines at -0.58, 0 and 0.58), while the resampling approach returns values close to the expected ones. Red dots are population means estimated with the original EM, boxes are the distributions of means across the replicated resampling method, bold lines are the medians of the

means for each sub-population. Note that, the EM estimated means were assigned to the group with the closest expected value.



Validation on NCI cell lines

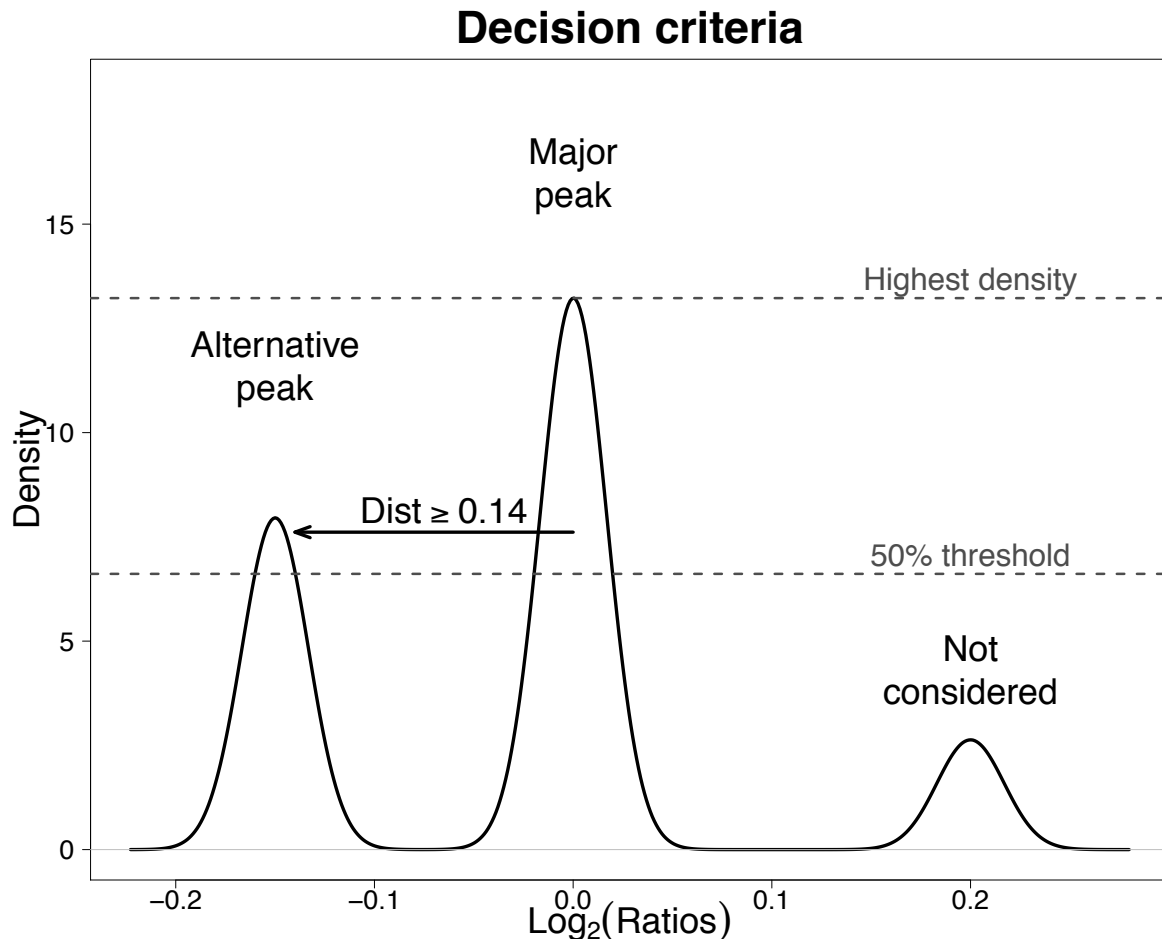
This approach was validated on the NCI data by estimating, for each sample, the reproducibility of the maximum peak detection using this procedure, on 1000 independent tests. The 2-standard deviation interval of the maximum peak values distribution was used to refine our definition of a valid alternative peak.

From the 60 individual cell lines represented by 72 CGH data, 3 were removed because of too few karyotypic information to generate a genomic-like profile. The remaining 57 cell lines were represented by 69 aCGH data, including replicated experiments.

Over all these remaining 69 experiments, the maximum peak values showed a 2-standard deviation (2SD) from $1.19e-3$ to $1.38e-1$. We then considered the largest observed 2SD interval of 0.14 as an additional constraint. Finally a relevant alternative centralization value was defined as the mean of a peak with a density

height of at least 50% the density height of the major peak, and at least at 0.14, in Log_2 , from this major peak (figure 5).

Figure 5: Given the validation step on the NCI cell lines, a valid alternative peak was defined as a peak with a maximum density higher than 50% of the height of the main peak, and at least at 0.14, in Log_2 scale.



For 6 of the 57 different cell lines, 2 or more replicated aCGH experiments were available. In all the cases, but one, the peak estimations were similar across the replicated experiments. For one of the 4 MCF-7 replicates, an alternative peak was detected, but at a distance lower than 0.14. This profile had also a highest derivative $\text{log}_2(\text{Ratio})$ spread, compared to the 3 others (.256 vs. .252, 0.243 and 0.230, respectively).

Supplementary figure S1: General array-based genomic profiling workflow.

After hybridization, fluorescent signals are digitized and then mapped with the probes locations. LogR are computed against the reference signals (dual-color hybridization) or against an external reference (single-color hybridization). The centralization step adjust the LogR on their median, or on the maximum density value. Then, the segmentation step identifies the breakpoints. Calls rely on the segments magnitude with respect to the base line, considered as the neutral-copy level.

Supplementary figure S2: Frequency of ploidies in the NCI60 cell lines panel, and proportion of alternative peak identification.

The NCI60 cell line panel predominantly includes aneuploid cell lines; more than 50% of the cell lines are, at least, $3n$ (A). Interestingly, centralization alternatives occur in rare cases of $2n$ cell lines, while alternative options are extremely frequent in $3n$ and $4n$ cell lines (B). The unique case of a $5n$ -/+ cell line is specifically described in supplementary figure 2.

Supplementary figure S3: Consistency of genomic profiles according to the centralization methods: the A549 cell line

The A549 genomic profiles has been centered on LogR median (top panel), the maximum density peak (central panel), or the alternative LogR density peak (bottom panel), before being segmented using the CBS algorithm, with the same segmentation parameters. The corresponding centralization values are indicated in bold on each density plot (Centralization). When comparing with the corresponding karyotype, The 2 first methods adjust the entire profile on the main cell line ploidy, namely $3n$, and lead to consider all of the 3-copy chromosomes (1p, 2, 3, 5, 7 to 10, 12, 14 and 16) as in normal count, while most of the 2-copy chromosomes are considered as lost (1p, 4, 6, 13, 19, 21 and 22). Adjusting on the alternative peak dramatically reduce such discrepancies, although errors persist on chr11. Notice that 2 supplementary copies of chr19 are located on chr15, according to the karyotype.

Supplementary figure S4: The $5n$ -/+ SF-295 cell line.

No alternative centralization is suggested when analyzing the LogR density as a mixture of Gaussian populations (left), and only the major density peak seemed to correspond to a sensible value for adjusting the genomic profile (center). However,

this choice led to an erroneous profile, given the karyotype (right); most of the 5-copy chromosomes were considered as in normal counts on the genomic profile, while 3-copy chromosomes (chr10 and 14) were considered as lost.

Supplementary figure S5: Correlation between copy number variation and sensitivity to related inhibitors.

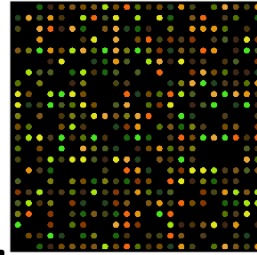
The Spearman correlations of FGFR1 and MET, with their respective inhibitors in the CCLE data (TKI258 and PHA665752, respectively), are not significantly improved when changing the profile centralization strategy: p were at least greater than 0.27 in all comparisons. To note, the relatively weak correlations between these genes copy number variation and responses to their corresponding inhibitor.

Supplementary table S6: Amplification calls in SAFIR01 and MOSCATO-01, according to the centralization method.

Study, patient Id, and platforms are indicated in columns 1 to 3, respectively. Considering the same genes as in André *et al.*, and using the same decision rules, amplifications are indicated, for each centralization method, in the corresponding column: “none” means no amplification detected, and in bold, genes detected using one method only.

aCGH analysis workflow from experiment to decision

Hybridization



Digitization

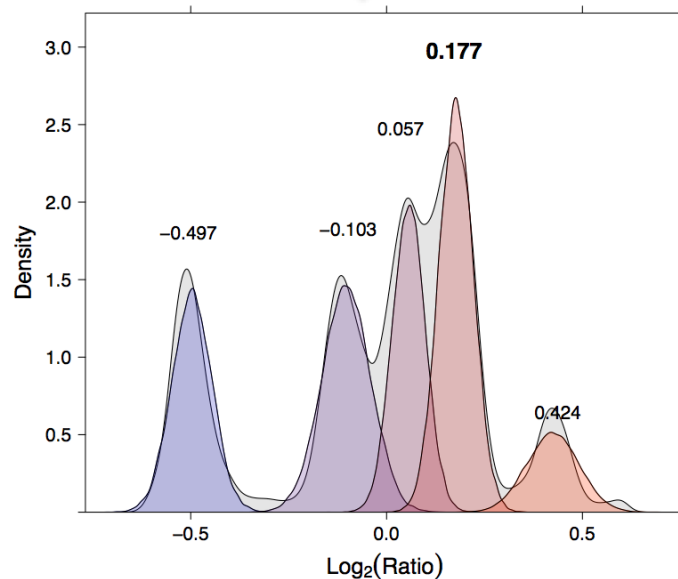
Affymetrix
CEL files

Agilent
FE files

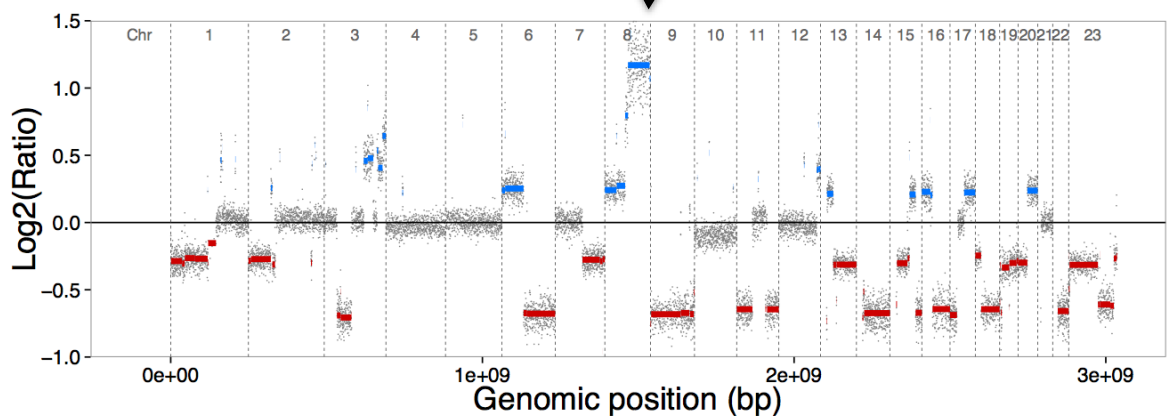
Preprocessing

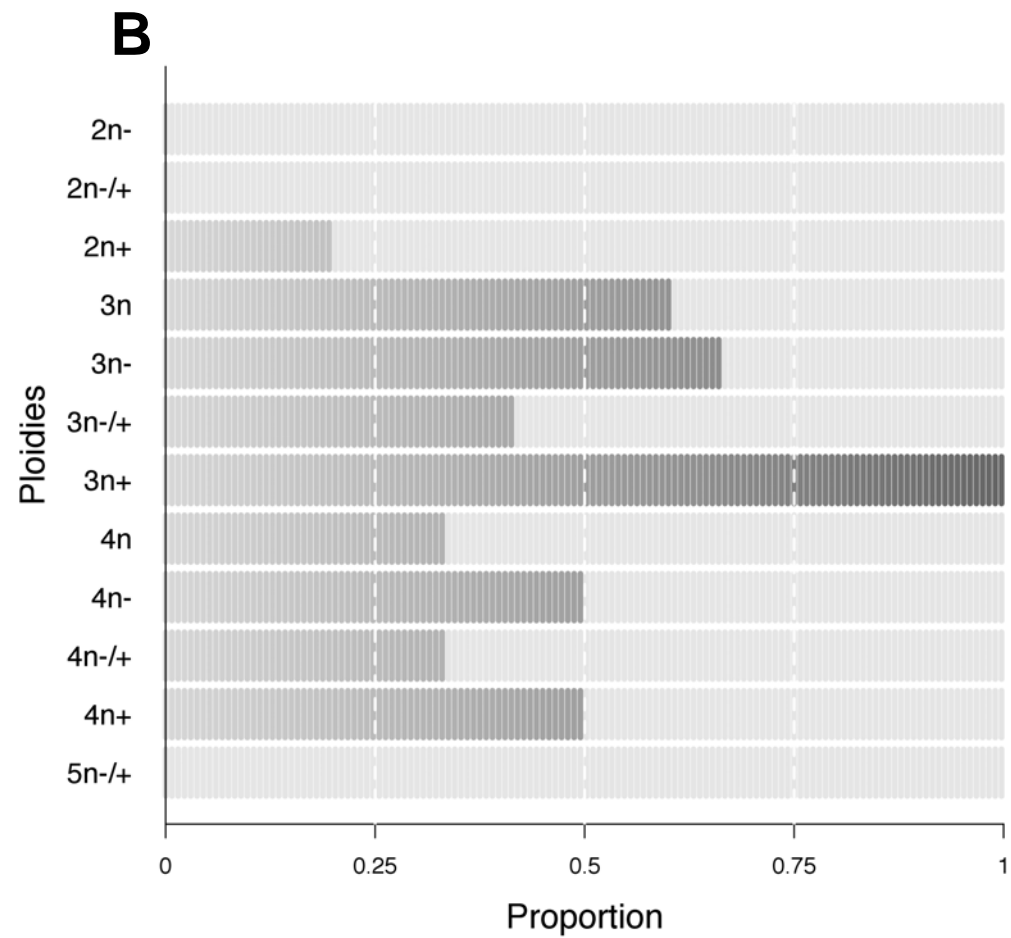
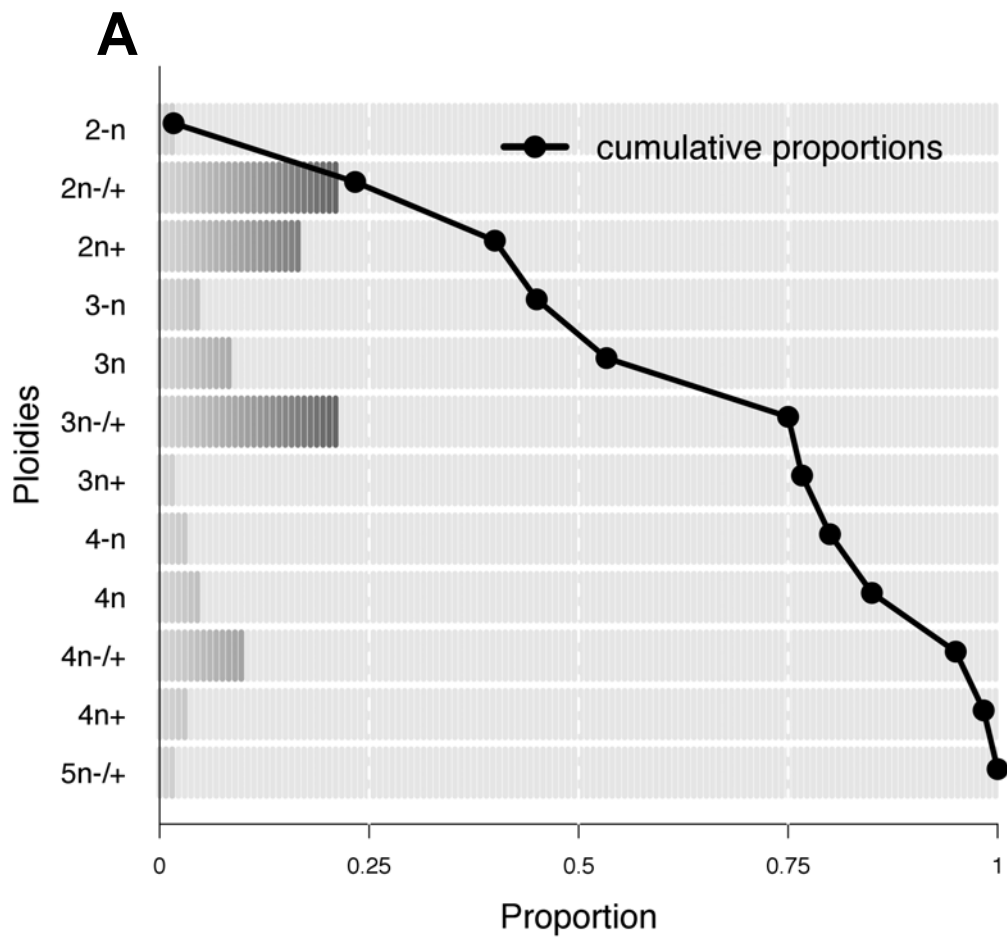
$\text{Log}_2(\text{ratio})$

Centralization



Segmentation



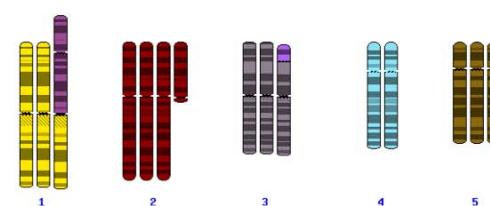
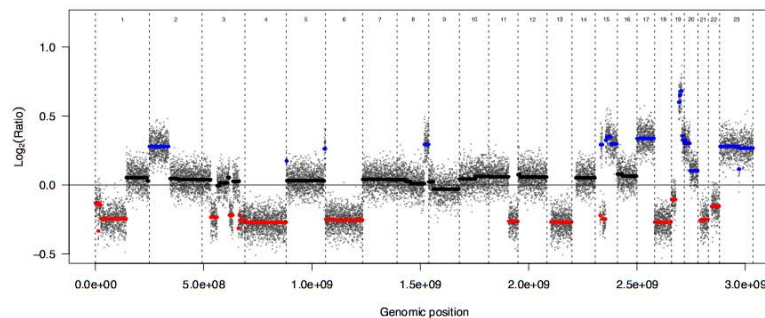
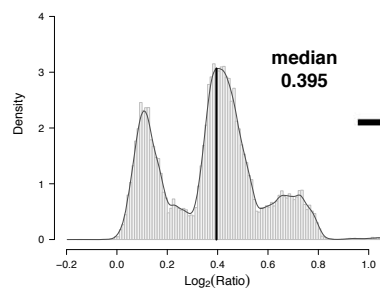


Centralization

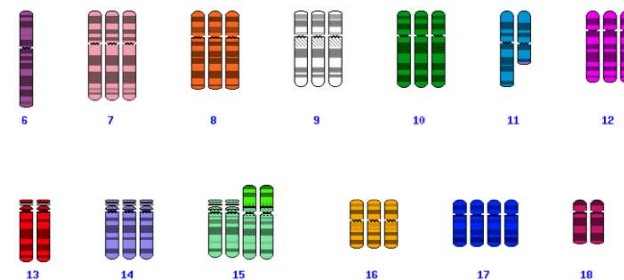
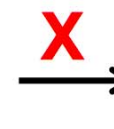
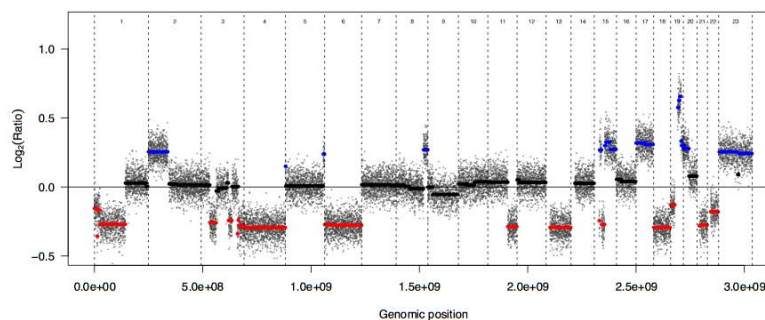
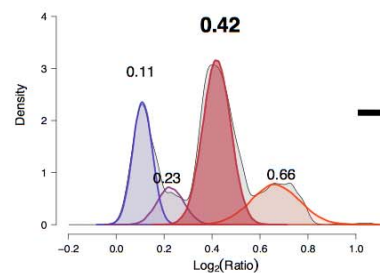
Genomic profile

Karyotype

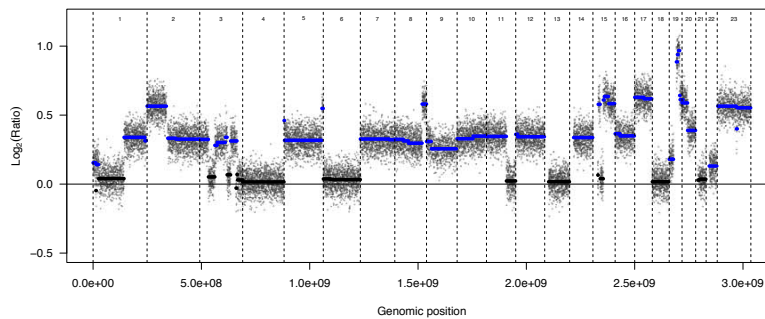
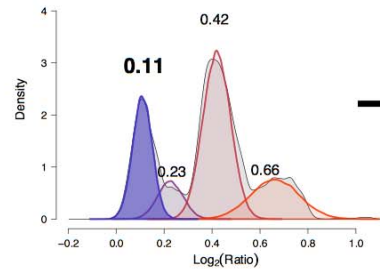
Median



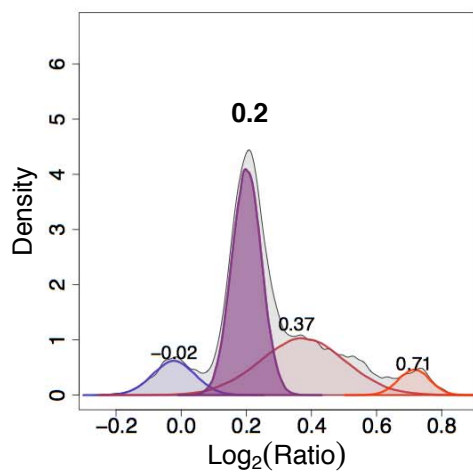
Maximum density



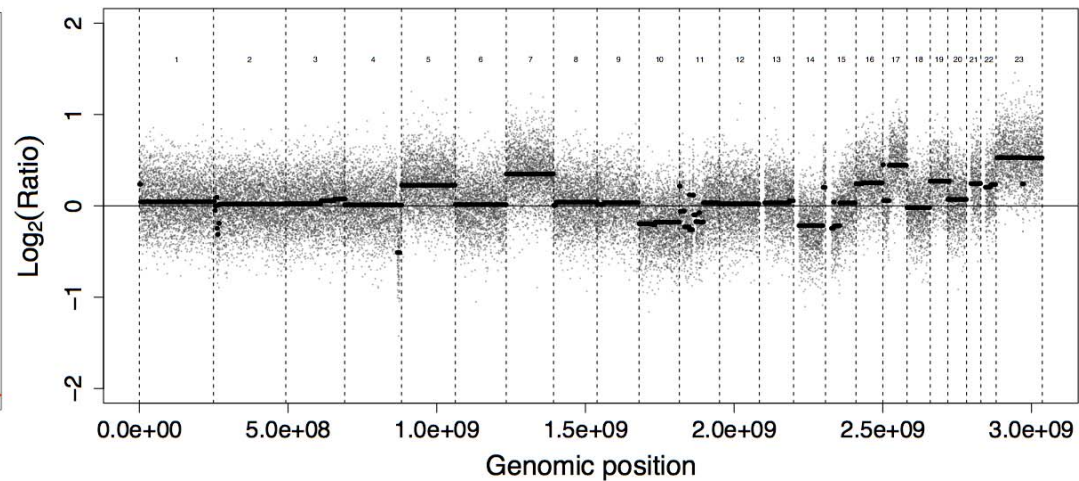
Alternative peak



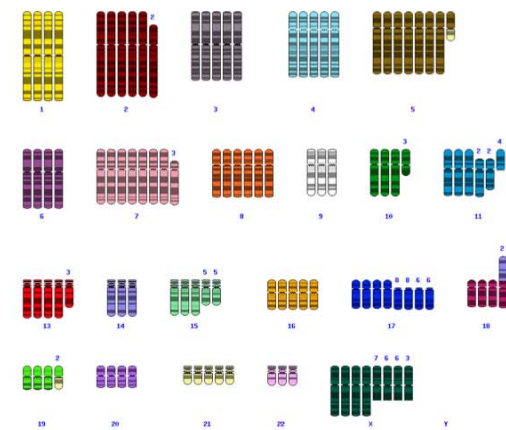
Density plot



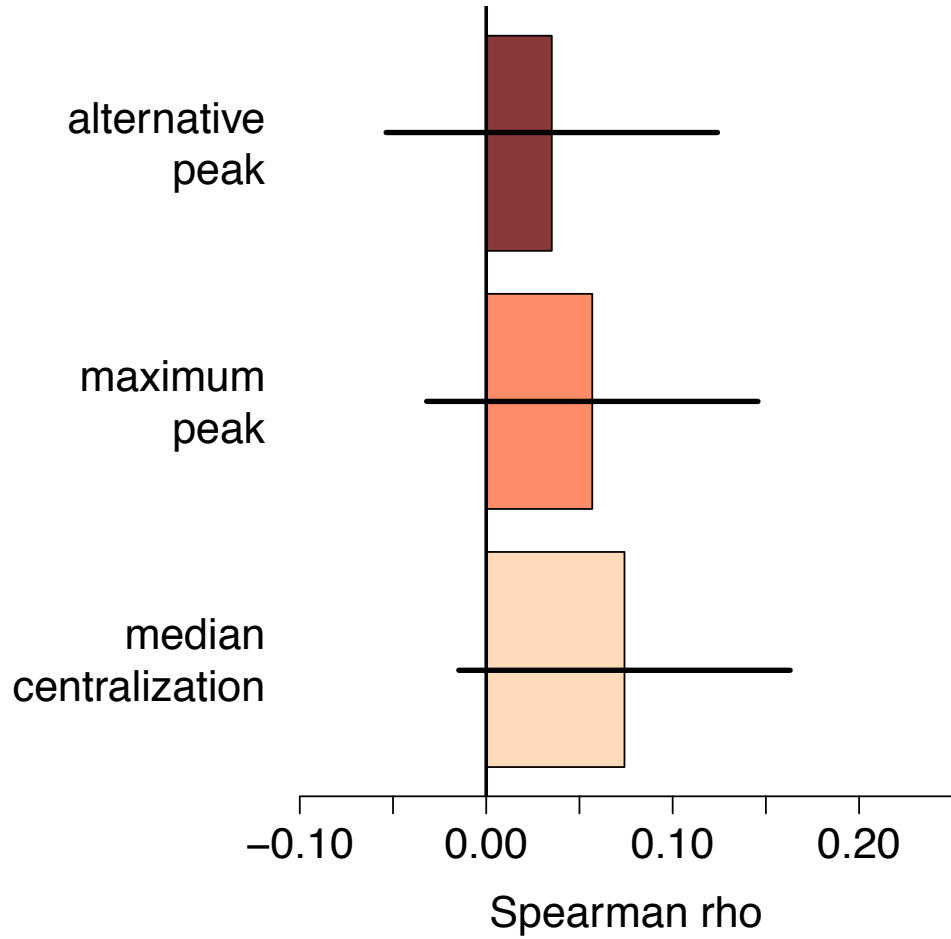
Genomic profile



Karyotype



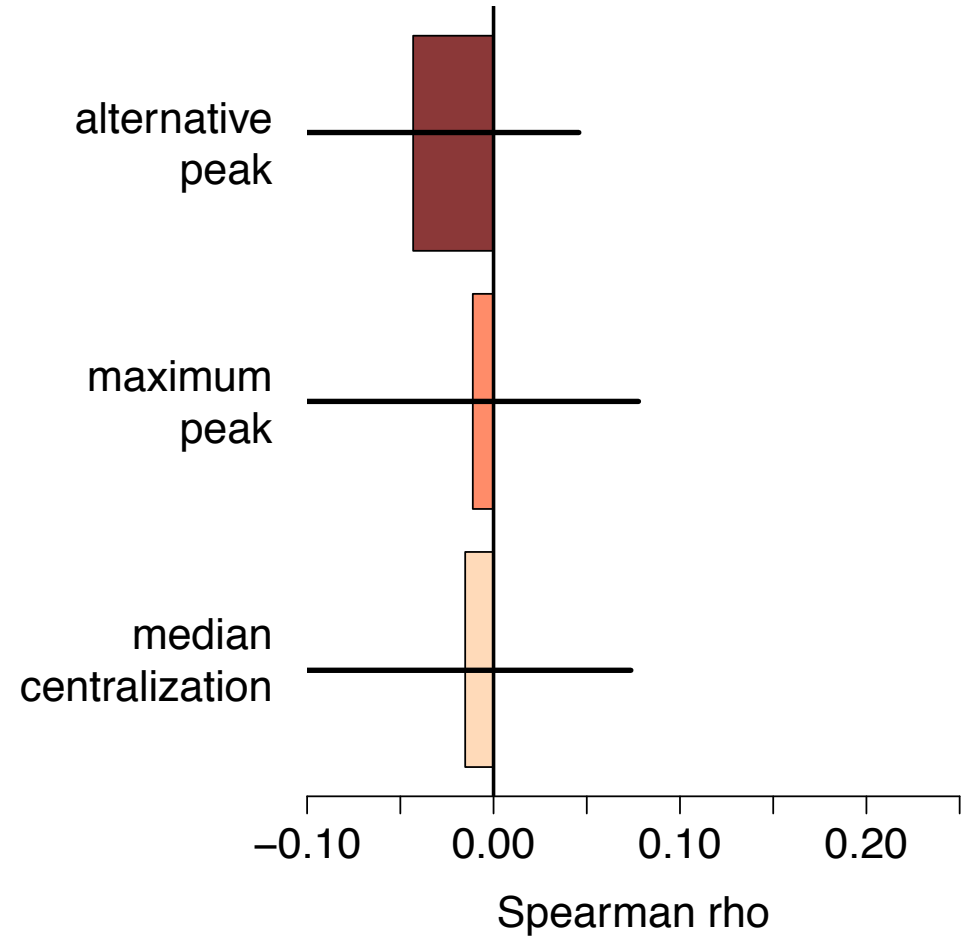
FGFR1/TKI258



FGFR1 Vs. TKI258

	Median	Maximum	Alternative
Rho	0.074	0.057	0.035
p-values		Maximum	Alternative
Median		0.394	0.271
Maximum			0.367

MET/PHA665752



MET Vs. PHA665752

	Median	Maximum	Alternative
Rho	-0.015	-0.011	-0.043
p-values		Maximum	Alternative
Median		0.475	0.332
Maximum			0.309