A-statistic as copy number variation estimator.

A-statistic was proposed by Myers and it is calculated as log of ratio of probabilities that segment is unique versus segment has multiplicity at least two when counting only non-duplicate non-contained reads. As sequencing coverage increases it is become harder to distinguish between different values of multiplicity (because expected spacing difference between reads becomes smaller). Counting duplicated or contained reads leads to another problem - method becomes vulnerable to sequencing artefacts.
So I checked the performance of A-statistic test on one of known E.coli assemblies.
The pipeline:
1) Assemble primary contigs from paired reads with SGA.
2) Use sga-astat.py included to sga to calculate A-statistic. (Authors of SGA claims that value of A-statistic above 20 means that contig is to be unique)
3) Use blast to align contigs and filter contigs with align length near equal to contig length and extremely high level of similarity
4) Compare results and make conclusions.

https://github.com/1dayac/summerScience - link to all the scripts

Results:
ok - unique or missing - 150
not ok - unique A-stat but multiple alignments - 293
not ok - multiple A-stat and unique or missing alignments - 0
ok - multiple A-stat and multiple alignments - 121

When considering 293 false results the stats are next:
2: 24, 3: 48, 4: 31, 5: 14, 6: 1, 7: 93, 10: 10, 11: 17, 12: 13, 14: 14, 15: 20
where first number is multiplicity of contig and the second is number of time contigs with such multiplicity appears at genome.

Conclusion: use A-statistic as copy number estimator is not a good idea. It has a lot of misses counting segments as unique when they appeared 2-15 times. Though considering structure of scaffolding step it is possible to say, that A-statistic works well, as it counts really high abundant segments as multiple. Usually low abundant segments stays at completely different locations of genome and it could not lead to ambiguities at scaffolding.
So, we just need to find some more powerful estimator for CNV.