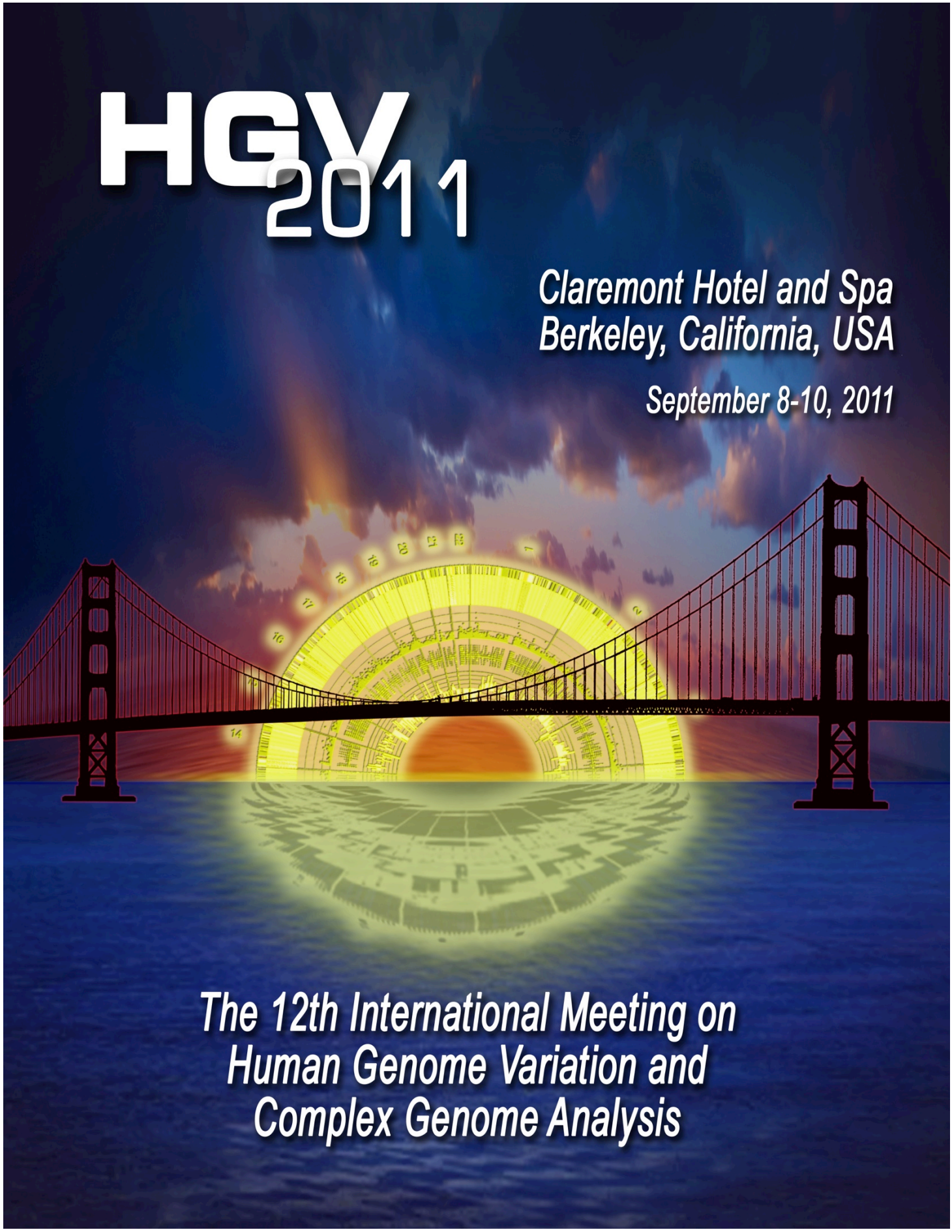
HGV
2011
Claremont Hotel and Spa
Berkeley, California, USA
September 8-10, 2011
The 12th International Meeting on
Human Genome Variation and
Complex Genome Analysis

cn.MOPS: mixture of Poissons for discovering copy number variations in next generation sequencing data

Guenter Klambauer, Djork-Arné Clevert, Sepp Hochreiter

The quantitative analysis of next generation sequencing (NGS) data like the detection of copy number variations (CNVs) is still challenging. Current methods detect CNVs as changes of read densities along chromosomes, therefore they are prone to a high false discovery rate (FDR) because of technological or genomic read count variations, even after GC correction. A high FDR means many wrongly detected CNVs that are not associated with the disease considered in a study, though correction for multiple testing must take them into account and thereby decreases the study's discovery power.

We propose "Copy Number estimation by a Mixture Of PoissonS" (cn.MOPS) for CNV detection from NGS data, which constructs a model across samples at each genomic position, therefore it is not affected by read count variations along chromosomes. In a Bayesian framework, cn.MOPS decomposes read variations across samples into integer copy numbers and noise by its mixture components and Poisson distributions, respectively. The more the data drives the posterior away from a Dirichlet prior corresponding to copy number two, the more likely the data is caused by a CNV, and, the larger is the informative/non-informative (I/NI) call. cn.MOPS detects a CNV in the DNA of an individual by a region with large I/NI calls. I/NI call based CNV detection gurantees a low FDR because wrong detections are less likely for large I/NI calls.

We compare cn.MOPS with the five most popular CNV detection methods for NGS data at three benchmark data sets: (1) artificial, (2) NGS data from a male HapMap individual with implanted CNVs from the X chromosome, and (3) the HapMap phase 2 individuals with known CNVs. At all benchmark data sets cn.MOPS outperformed its five competitors with respect to precision (1-FDR) and recall both at gains and losses.

Günter Klambauer
Institute of Bioinformatics, Linz
Altenberger Straße 69
Linz, Upper Austria A-4040
Austria
Tel: +43-732-2468-1566
Email: klambauer@bioinf.jku.at