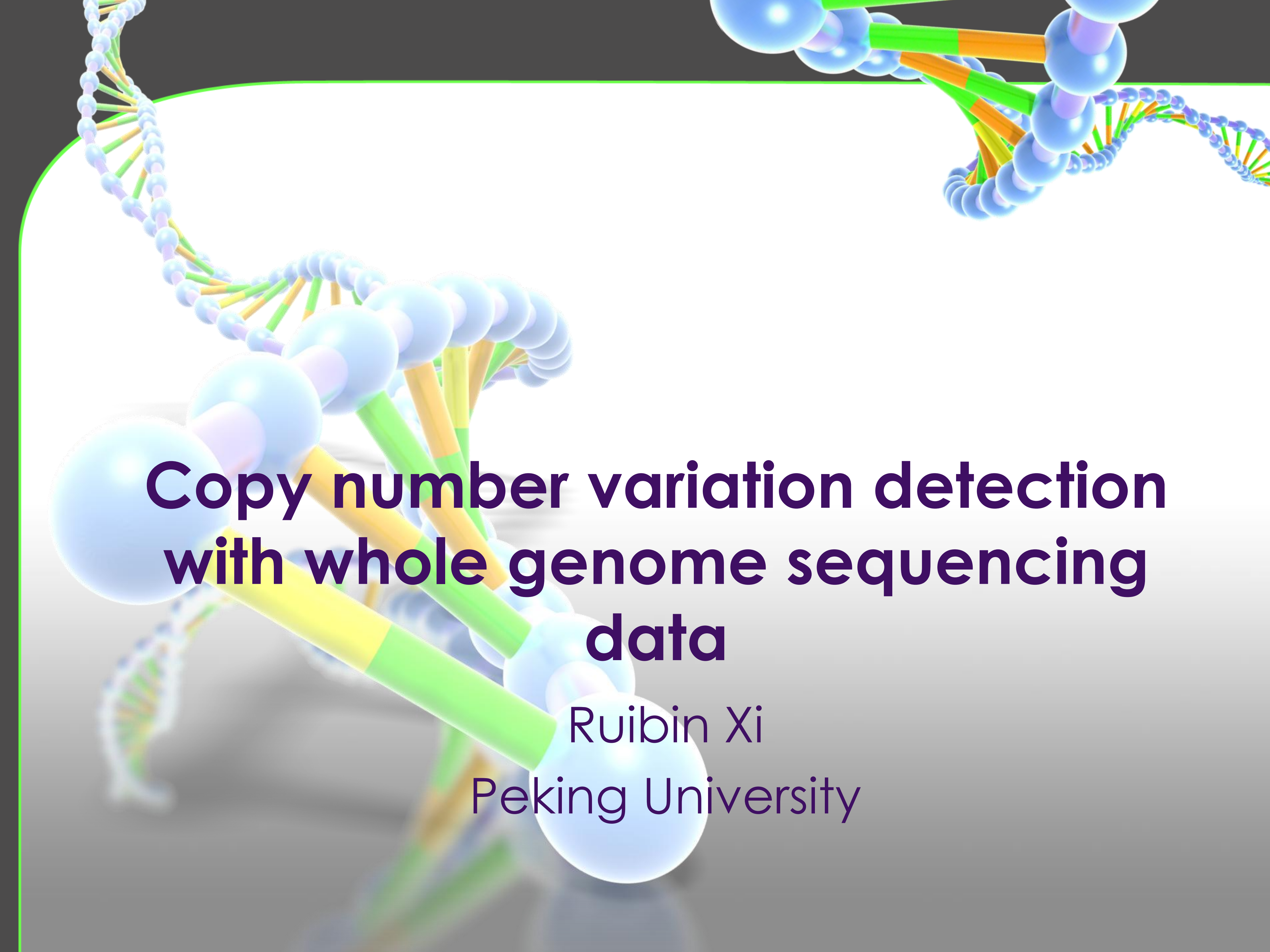


Copy number variation detection with whole genome sequencing data

Ruibin Xi

Peking University

Copy number variation (CNV)

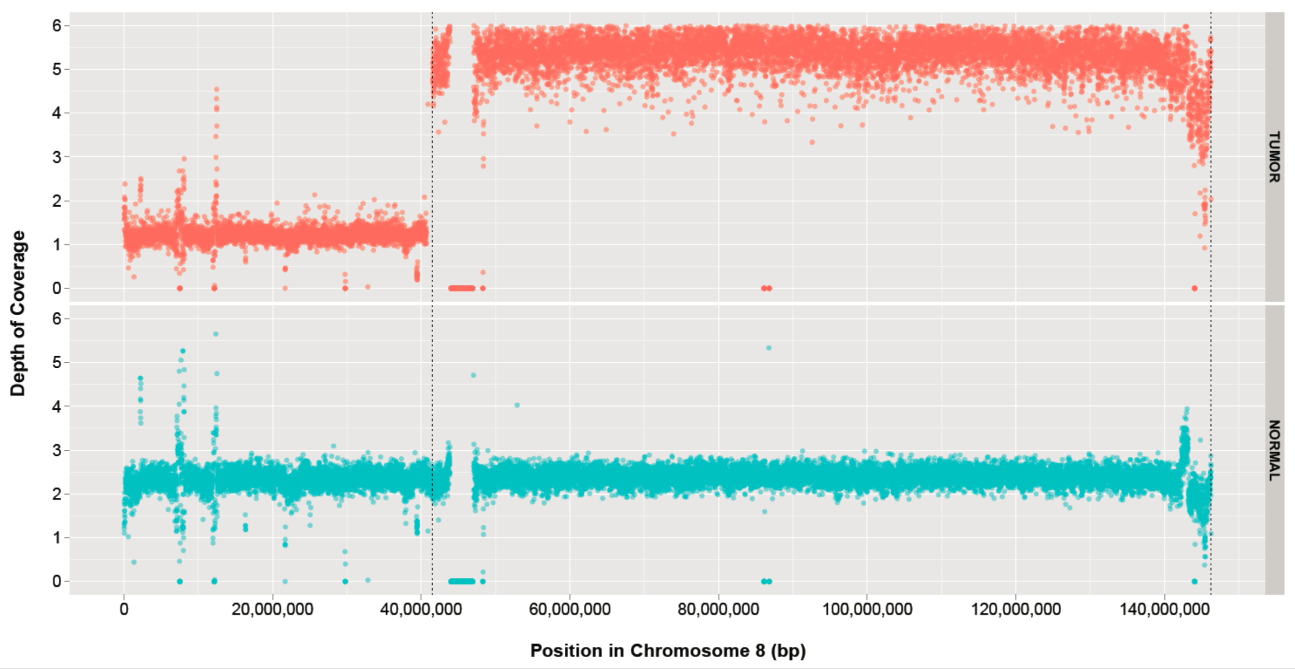
- CNVs: gains or losses of genomic segments
- CNVs accounts for a substantial proportion of human genomic variations
  - In the Database of Genomic Variation (DGV), over 30% of human genome can be influenced by CNVs
- CNVs have been associated with a wide spectrum of diseases
  - Autism, Schizophrenia and Obesity
- In Cancer genomes, we often see very large copy gains of losses

Next-generation sequencing (NGS)

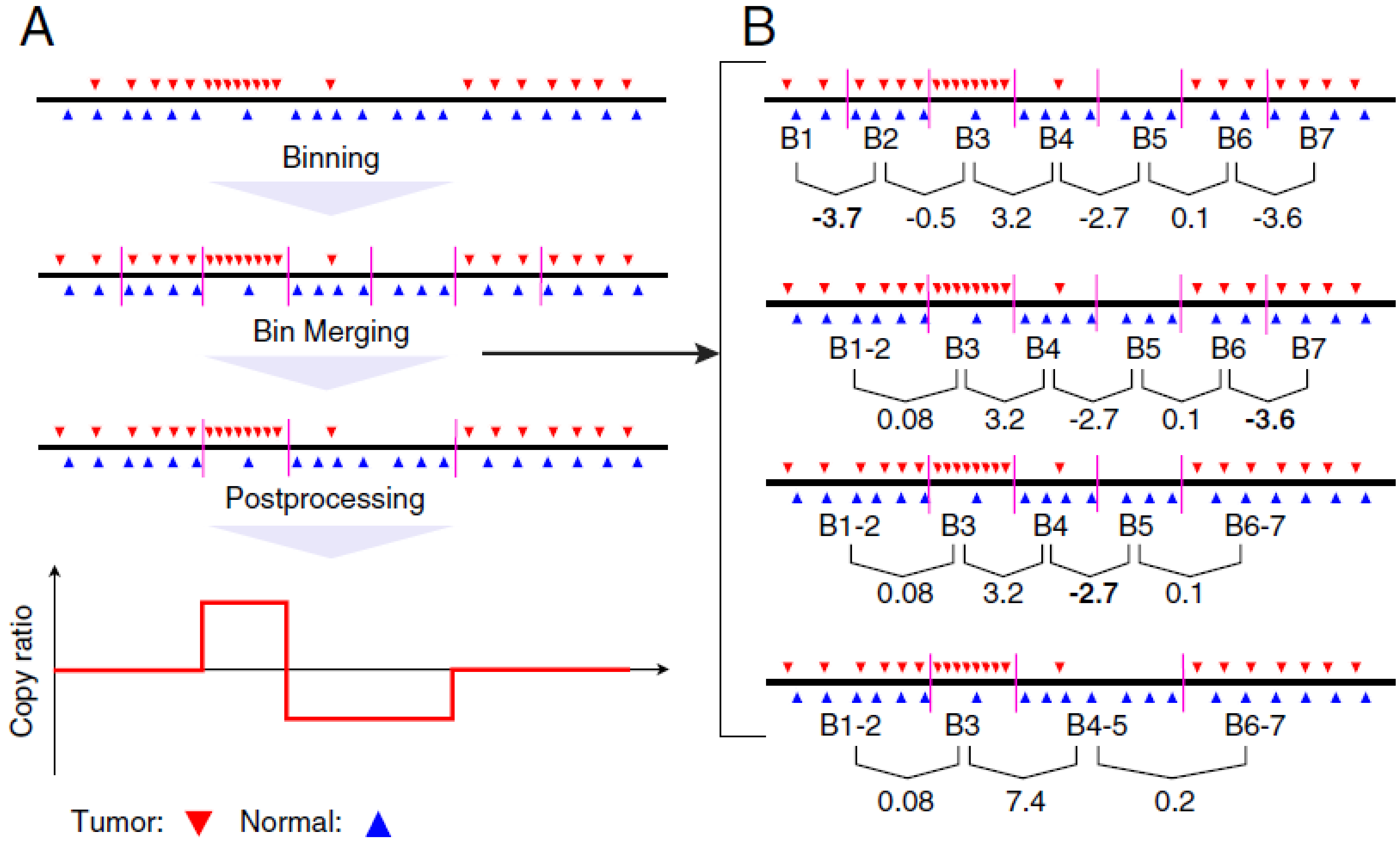
- NGS platforms
  - Roche 454 platforms
  - Illumina/Solexa platforms (most widely used)
  - Applied Biosystem (ABI) SOLiD
  - Helicos HeliSope™ sequencer(single molecular sequencing)
  - Life Technologies platforms
- The throughput is increasing and the price is dropping
- Short read but high throughput

CNV detection using read-depth

- Read-depth: read density in a genomic region
- If there is no bias, the read-depth in a genomic region should be roughly proportional to the copy number
- But there are often biases in the NGS data.



BIC-seq: an algorithm for detecting somatic CNVs in tumor genomes



Statistical Model

- Given a short read R that is mapped to the reference genome, it consists of two pieces of information
  - The position s on the reference genome
  - The read type Y: tumor ($Y=1$) or normal ($Y=0$).
- Assume the distribution of $R=(Y,S)$ is $f(y,s)$.
- By Bayes' theorem

$$\begin{aligned} f(y,s) &= \Pr(Y=y \mid S=s)\Pr(S=s) \\ &= \Pr(Y=y \mid S=s) f(s), \end{aligned}$$

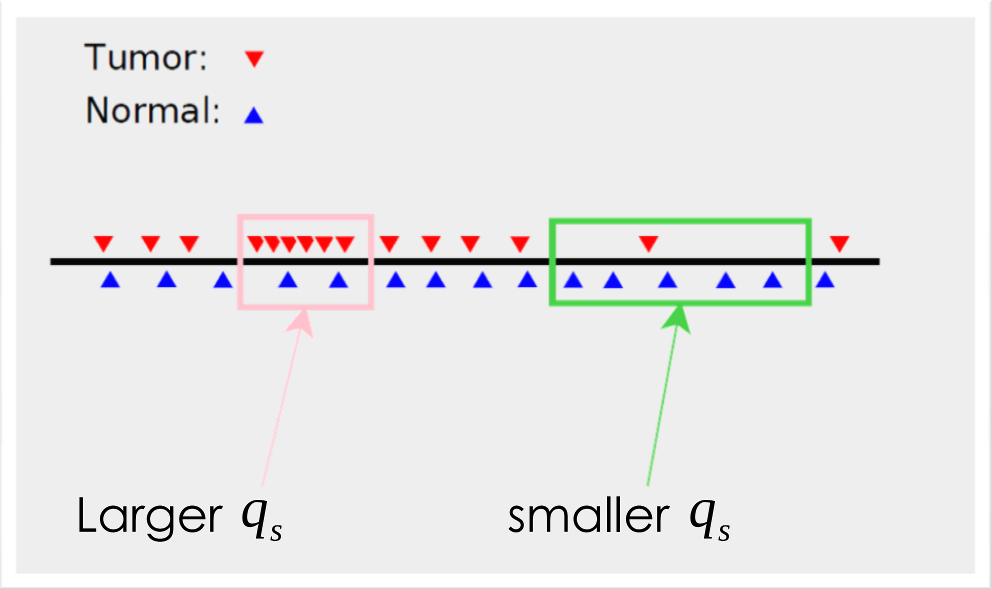
where $f(s)$ is the marginal distribution of S .

Statistical Model (cont. 1)

- Denote q_s be the probability of a read at position s being a tumor read, i.e. $q_s = \Pr(Y=1 \mid S=s)$.
- Given N mapped short reads $R_1=(y_1,s_1),\cdots,R_N=(y_N,s_N)$, the joint likelihood is

$$L_N = \prod_{i=1}^{N} q_{s_i}^{y_i}(1-q_{s_i})^{1-y_i} f(s_i)$$

- To identify CNV regions, it is enough to identify the breakpoints.



Statistical Model (cont. 2)

- Assume that q_s is a constant between any two neighboring breakpoints.
- Given the breakpoints $0 = \tau_0 < \tau_1 < \cdots < \tau_m < \tau_{m+1} = L_c$ on a chromosome c, where L_c is the length of the chromosome c.
- Let p_j be the common probabilities q_s between the breakpoints τ_j and τ_{j+1}. The likelihood can be written as

$$L_N = \prod_{j=0}^{m} \prod_{\tau_j < s_i \le \tau_{j+1}} p_j^{y_i} (1 - p_j)^{1-y_i} f(s_i),$$

- One set of breakpoints corresponds to one model. Then, we could use a model selection criterion such as the Bayesian information criterion (BIC) to select the breakpoints.

Bayesian information criterion (BIC)

- The general definition of the BIC of a model is

$$\mathrm{BIC} = -2\log(L) + k\log(n),$$

  - *L*: the likelihood function evaluated at the MLE
  - *k*: the number of parameters in the model
  - n: the total number of observations

BIC (cont.)

- Given the breakpoints $0 = \tau_0 < \tau_1 < \cdots < \tau_m < \tau_{m+1} = L_c$, the BIC is

$$\begin{aligned}\mathrm{BIC}(\lambda) &= -2\sum_{j=0}^{m}\left[k_j \log(\hat{p}_j) + (n_j - k_j)\log(1-\hat{p}_j)\right] \\ &\quad -2\sum_{i=1}^{N} f(s_i) + (m+1)\lambda\log(N),\end{aligned}$$

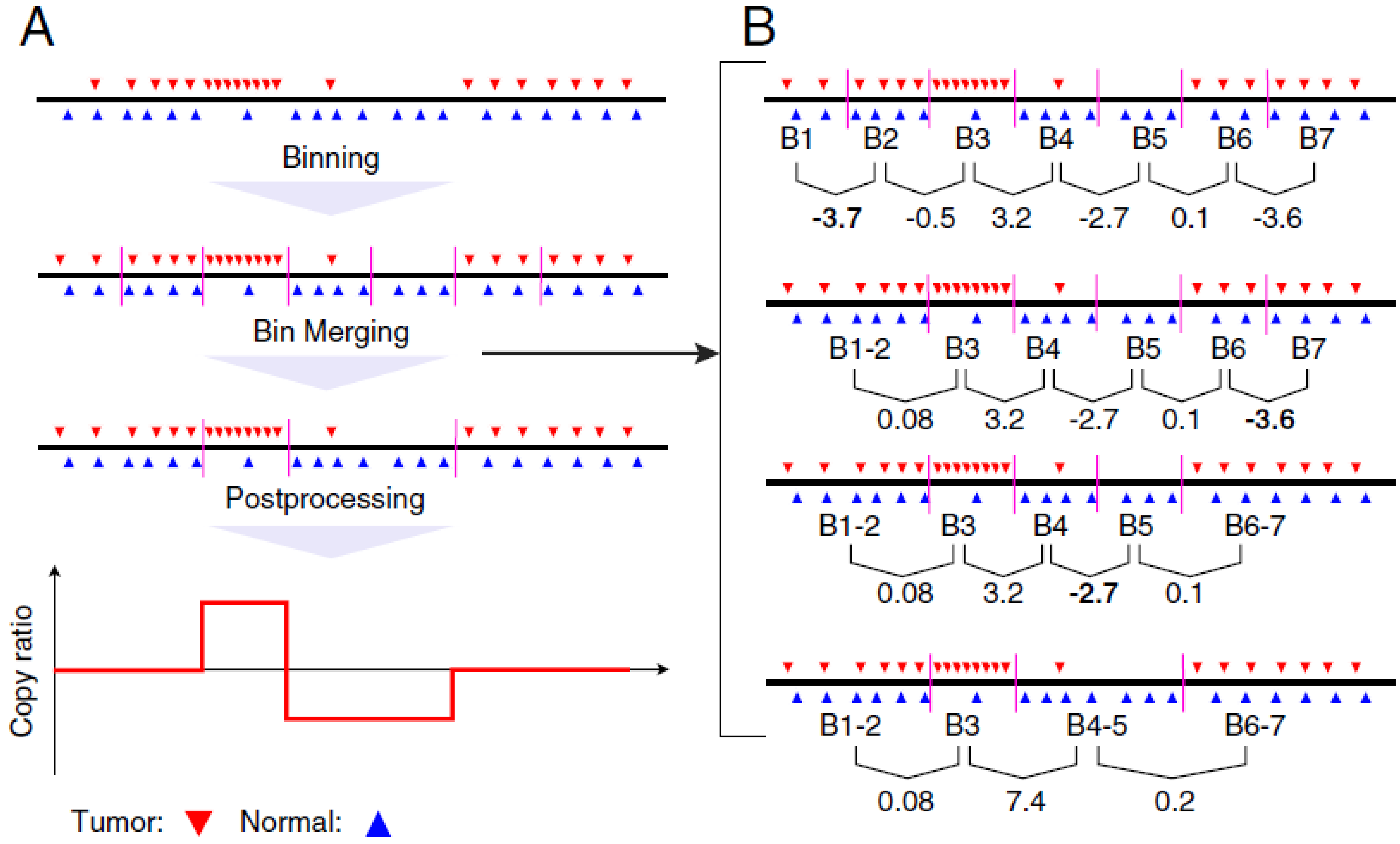
  - k_j : the number of tumor reads between τ_j and τ_{j+1} .
  - n_j : the total number of reads between τ_j and τ_{j+1}
  - $\hat{p}_j = k_j / n_j$: the MLE of the parameter p_j
  - λ>0 : tuning parameter

- Note that the term $-2\sum_{i=1}^{N} f(s_i)$ is common for all different models. Therefore, we can drop it when comparing different models.

Asymptotic result

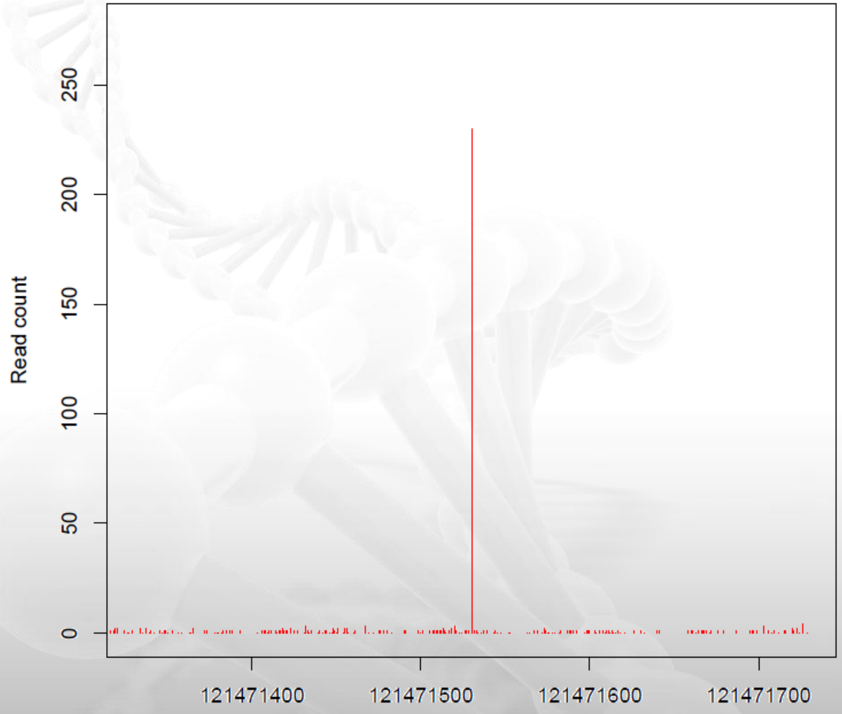
- Assume $f(s) > 0$ for all s. Then, the breakpoint set that minimizes the BIC is a consistent estimator of the true breakpoint set, i.e. it will converge to the true breakpoint set in probability as N, the number of observations, goes to infinity.

The BIC-seq algorithm (revisit)



Some remarks

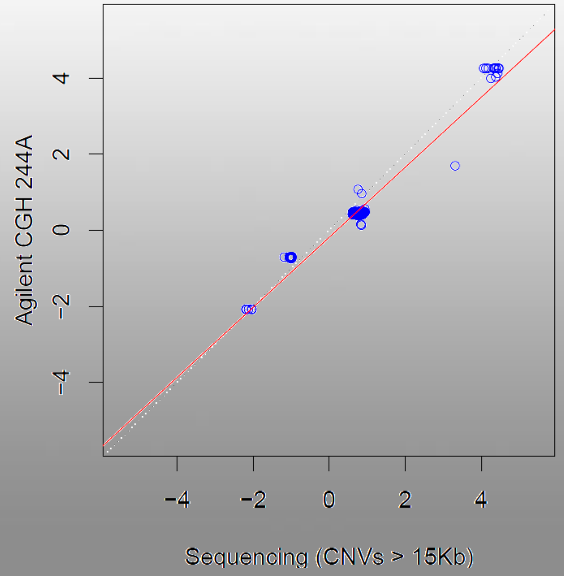
- Outlier removal:
  - Look at local genomic window to determine if the read count at a nucleotide position is an outlier
- Assign credible interval to a breakpoint
  - Gibbs sampling
- Assign false discovery rate
  - Permutation based

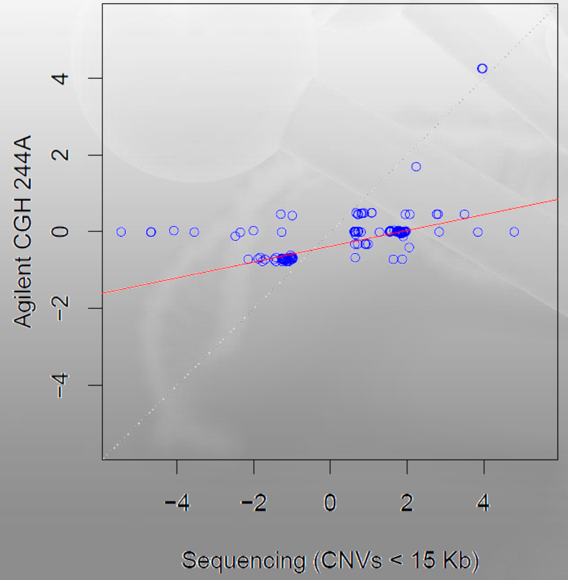


An outlier example

Application of BIC-seq on a GBM tumor genome

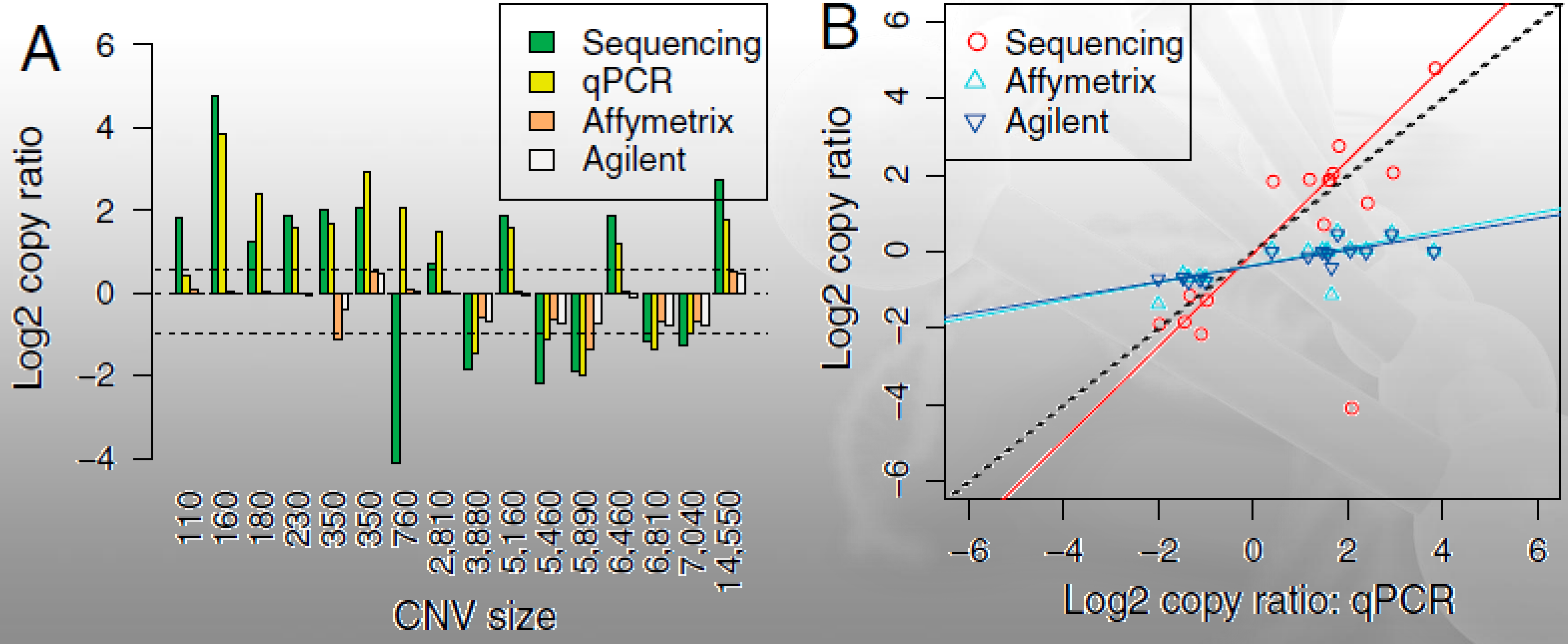
- Applied BIC-seq on a GBM tumor genome
  - Tumor: 10X
  - Normal: 7x
- Detected 291 putative CNVs ranging from 40bp to 5.7 Mb
- Compare the copy ratio estimate given by BIC-seq and an array-based platform





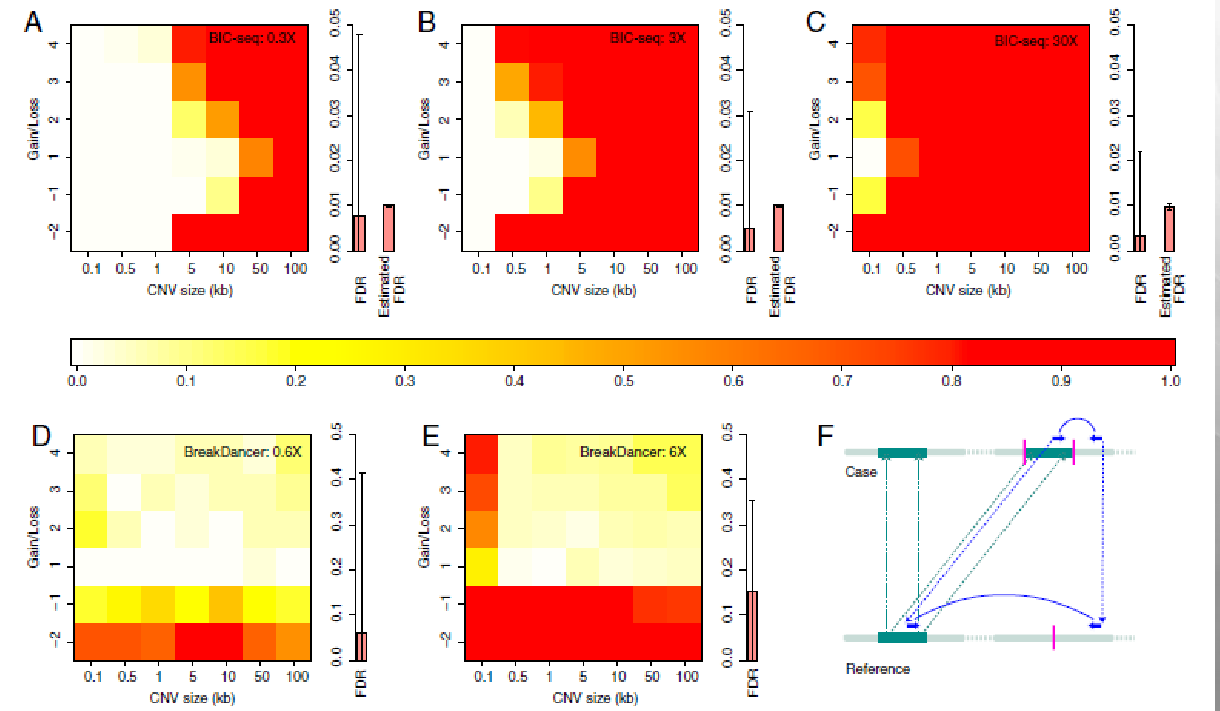
Application of BIC-seq on a GBM tumor genome (Cont.)

- Selected 16 small CNVs ranging from 110 bp to 14 kb for qPCR validation
- 14 out of 16 (87.5%) were validated



Comparison with a PEM method BreakDancer

- Use simulation for comparison
- Created 100 “tumor” chromosomes using chromosome 22 (hg18) as the template
  - Each tumor chromosome contains 42 CNV regions
  - 7 sizes and 6 copy numbers



Summary

- BIC-seq : a read-depth CNV detection algorithm
  - Nonparametric
  - Computationally efficient
  - High accuracy
  - Asymptotically consistent
  - High validation rate

Thank you for your attention!