NEXT GENERATION SEQUENCING: STATISTICAL CHALLENGES AND OPPORTUNITIES

Susmita Datta

susmita.datta@louisville.edu

www.susmitadatta.org

Department of Bioinformatics and Biostatistics
School of Public Health and Information Sciences
University of Louisville

Jt work with Somnath Datta, Ryan Gill and Seongho Kim

"The evolution of 'omic science through microarray transcriptomics, metabolomics, proteomics, and whole genome SNP-omics has in many ways come full circle with a new focus on genomics and genome sequencing."

- Bateman and Quackenbush, 2009

• Second-generation sequencing generates millions of short reads matures and it has the potential to be applied in a wide range of biological and clinical problems.

- It is absolutely critical to have clear metrics in place for data quality, reliability, reproducibility and biological relevance. Naturally statistical research dealing with this data can potentially help achieving the goal.

- We will discuss the use of statistical methodology used in achieving any of these goals.

Data quality and reproducibility:

- Maroni et al., 2008 *Genome Research*: observed that next generation sequencing data from Illumina are highly reproducible (better than microarray).

- Fu et al., 2009, *BMC Genomics*: arrived at a similar conclusion by comparing the relative accuracy of transcriptome sequencing (RNA-seq) and microarrays with protein expression data from adult human cerebellum using 2D-LC MS/MS.

- Wall et al., 2009, *BMC Genomics*: They concluded that NG sequencing offer great benefit in terms of coverage over capillary-based sequencing (used simulation model).

• Dohm et al., 2008, *Nucleic Acids Research*: found that wrong base calls are often preceded by base G. Base substitution errors were significantly disproportionate with A to C substitution error being 10 times more frequent than the C to G substitution.

• Irizarry and Bravo, 2009, Working Paper: who reported A to T miscall to be the most common error in their calibration study.

• Both studies reported that the error rates vary with the position on the read.

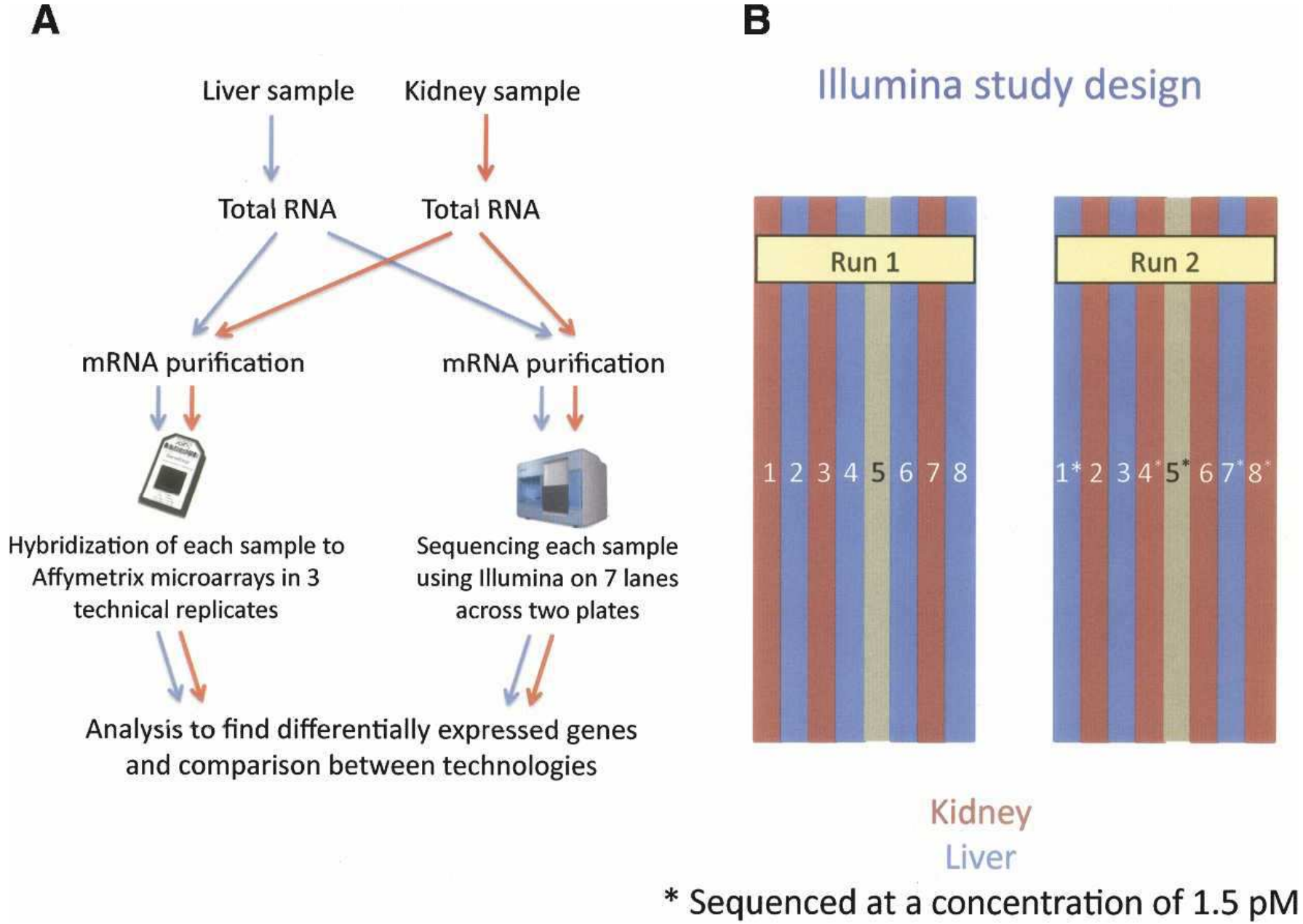
* These problems call for better base-calling procedure with a measure of uncertainty with each of them.

Reproducibility

- Maroni et al., 2008 *Genome Research*: studied the technical variance (between and within lane etc.) associated with Illumina sequencing in the context of finding gene expression difference of liver and kidney RNA samples. They also compare it with the result with affy data using from a human male.

- They used Illumina to sequence each sample on seven lanes across two plates.

- The gene counts were highly correlated across lanes (Spearman correlation average $\approx$ 0.96).


A
Liver sample
Kidney sample
Total RNA
Total RNA
mRNA purification
mRNA purification
Hybridization of each sample to Affymetrix microarrays in 3 technical replicates
Sequencing each sample using Illumina on 7 lanes across two plates
Analysis to find differentially expressed genes and comparison between technologies
B
Illumina study design
Run 1
Run 2
1 2 3 4 5 6 7 8
1* 2 3 4* 5* 6 7* 8*
Kidney
Liver
* Sequenced at a concentration of 1.5 pM

Test for *lane effect* by comparing each pair of lanes:

For each mapped gene: test null hypothesis that gene counts in one lane represent a random sample from the reads in both lanes

$$p-value = 1-2\left|p^{*}-.5\right| \quad \text{where} \quad p^{*} = \sum_{x=0}^{x_{t1}} p_0(x) + U p_0(x_{t1})$$

$$p_0(x) = \frac{\binom{C_1}{x}\binom{C_2}{x_{t1}+x_{t2}-x}}{\binom{C_1+C_2}{x_{t1}+x_{t2}}},$$

x_{tk} = number of counts for gene t in lane k

C_k = total number of reads in lane k

U = randomly generated Uniform(0,1) r.v.

Test for *lane effect* by comparing *L* lanes:

For each sample *i*:

$x_{ijk} = \text{number of reads mapped to gene } j \text{ for lane } k$

$\text{Assume } x_{ijk} \text{ follows indept Poisson}(\mu_{ijk} = c_{ik}\lambda_{ijk}) \text{ where}$

$c_{ik} = \text{total rate that lane } k \text{ produces reads}$

$\lambda_{ijk} = \text{rate of reads to gene } j \text{ in lane } k \text{ relative to other genes } (\sum_{\text{j}} \lambda_{ijk} = 1)$

- To test $\mathrm{H}_0 : \lambda_{ij1} = \lambda_{ij2} = \ldots = \lambda_{ijL}$
- compute goodness of fit statistic $X_{ij} = \sum_k \frac{\left(x_{ijk} - \hat{\mu}_{ijk}\right)^2}{\hat{\mu}_{ijk}}$
- Under H_0, X_{ij} follows a $\chi^2_{L\text{-}1}$ distribution

Experiment: Identifying Differentially Expressed Genes

- Assume x_{ijk} follows indept Poisson($\mu_{ijk} = c_{ik}\,\lambda_{ijk}$)

- For each gene *j*, separate *L* lanes into groups *A* and *B* and test

$$\mathrm{H}_0 : \lambda_{ijk} = \lambda_j \text{ for all } (i,k) \text{ vs.}$$

$$\mathrm{H}_\mathrm{a} : \lambda_{ijk} = \lambda_j^A \text{ for } (i,k) \in A \text{ and } \lambda_{ijk} = \lambda_j^B \text{ for } (i,k) \in B$$

- Likelihood ratio test:

$$\Lambda(x_j) = -2 \ln L(x_j)$$

$$\text{where } L(x_j) = \frac{\prod_{(i,k)} f(x_{ijk} \mid \hat{\lambda}_j)}{\prod_{(i,k)\in A} f(x_{ijk} \mid \hat{\lambda}_j^A) \prod_{(i,k)\in B} f(x_{ijk} \mid \hat{\lambda}_j^B)}$$

$$\text{and } f(\bullet \mid \lambda) \text{ is the Poisson}(\lambda) \text{ mass function}$$

- Under H_0,

$$\Lambda(x_j) \text{ follows a } \chi_1^2 \text{ distribution}$$

- 11,493 genes were differentially expressed in liver-versus-kidney samples

• The list of differentially expressed genes obtained for the Illumina data by the likelihood ratio test were compared with results based on Affymetrix U133 Plus 2 arrays where an Empirical Bayes approach was used to identify differentially expressed genes.

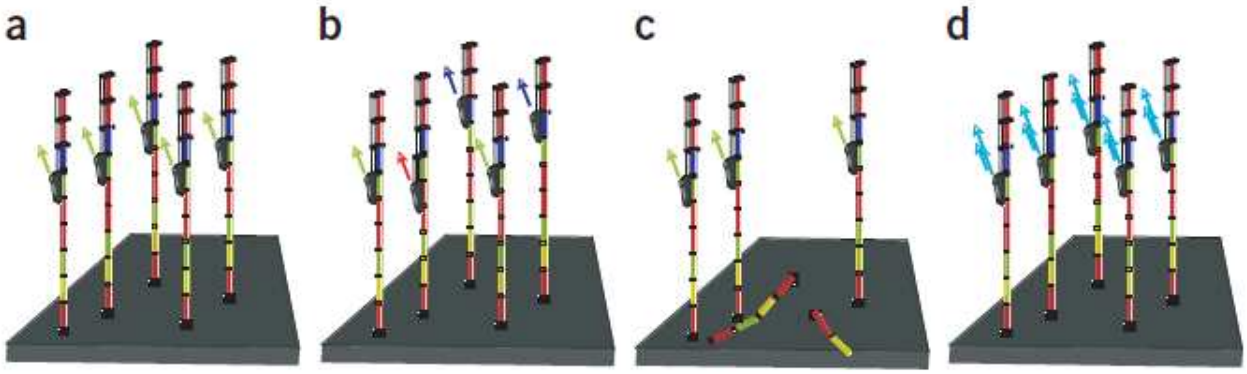
• Of 8113 differentially expressed genes found by the array, 81% were also found to be differentially expressed using Illumina.

• Quantitative PCR (qPCR) was used to examine discrepancies, and overall, the qPCR results agreed more with Illumina than with the microarrays.

Base calling methods for Solexa

- Bustard (Illumina)
  - Developed by Illumina; Cycle-independent
- Alta-Cyclic (Erlich et al., 2008 Nature Methods)
  - SVM
- Rolexa (Rougemont et al., 2008 BMC Bioinformatics)
  - Entropy-based Clustering
- Swift (Whiteford et al., 2009 Bioinformatics)
  - Image analysis
- Ibis (Kircher et al., 2009 Genome Biology)
  - Multiclass-SVM (Cycle-dependent modeling)
- Irizarry and Bravo (2009 Berkeley Electronic Press)
  - read/base- cycle effects
- BayesCall (Kao et al., 2009 Genome Research)
  - Hierarchical (Graphical) Cycle-dependent stochastic modeling

Main Illumina noise factors (b)-(d)



Erlich et al.
Nature Methods
5: 679-682
(2008)

- (a) In the ideal situation, after several cycles the signal (green arrows) is strong, coherent and corresponds to the interrogated position
- (b) Phasing noise introduces lagging (blue arrows) and leading (red arrow) nascent strands, which transmit a mixture of signals
- (c) Fading is attributed to loss of material that reduces the signal intensity
- (d) Changes in the fluorophore cross-talk cause misinterpretation of the received signal (teal arrows).

Image analysis

- Correct for the imperfect repositioning of the CCD camera between cycles and for chromatic aberration of its lens

  align the current image to a reference (initial cycle) image

- Identify clusters from their surrounding background containing identical copies of DNA templates

  thresholding and segmentation image analysis

- Time series data of fluorescence intensities

Base calling

- Convert the fluorescence signals into actual sequence data with quality scores

Summary (Bustard)

Estimate $\mathcal{Z}_k = (\mathcal{Z}_{1,k}, \ldots, \mathcal{Z}_{L,k}) \in \mathbb{R}^{4\times L}$

$\boldsymbol{I}_{t,k} = \boldsymbol{X}\mathcal{Z}_{t,k}$

Intensities: $\boldsymbol{I}_{t,k} = (I^A_{t,k}, I^C_{t,k}, I^G_{t,k}, I^T_{t,k})' \in \mathbb{R}^{4\times 1}$

Concentration: $\mathcal{Z}_{t,k} = (\mathcal{Z}^A_{t,k}, \mathcal{Z}^C_{t,k}, \mathcal{Z}^G_{t,k}, \mathcal{Z}^T_{t,k})' \in \mathbb{R}^{4\times 1}$

Crosstalk matrix: $\boldsymbol{X} = (X_{ij}) \in \mathbb{R}^{4\times 4}$

Renormalize $\mathcal{Z}_k = (\mathcal{Z}_{1,k}, \ldots, \mathcal{Z}_{L,k}) \in \mathbb{R}^{4\times L}$

x $\bar{\mathcal{Z}}_1/\bar{\mathcal{Z}}_t$

$$\bar{\mathcal{Z}}_t := \frac{1}{N}\sum_{k=1}^{N}(\mathcal{Z}^A_{t,k} + \mathcal{Z}^C_{t,k} + \mathcal{Z}^G_{t,k} + \mathcal{Z}^T_{t,k})$$

Estimate $\mathring{\mathcal{Z}}_k = (\mathring{\mathcal{Z}}_{1,k}, \ldots, \mathring{\mathcal{Z}}_{L,k}) \in \mathbb{R}^{4\times L}$

$\mathcal{Z}_k = \mathring{\mathcal{Z}}_k\boldsymbol{Q}$

$\mathcal{Z}_k\boldsymbol{Q}^{-1}$

$\boldsymbol{Q} = (Q_{jt})$ is an L-by-L matrix with $Q_{jt} = [\boldsymbol{P}^t]_{o,j}$

$\mathcal{Z}_{t,k} = \sum_{j=1}^{L} \mathring{\mathcal{Z}}_{j,k}[\boldsymbol{P}^t]_{o,j}$

$$P_{ij} = \begin{cases} p, & \text{if } j = i, \\ 1-p-q, & \text{if } j = i+1, \\ q, & \text{if } j = i+2, \\ 0, & \text{otherwise.} \end{cases}$$

Base calling model-BayesCall,
Kao et al., 2009 *Genome Research*

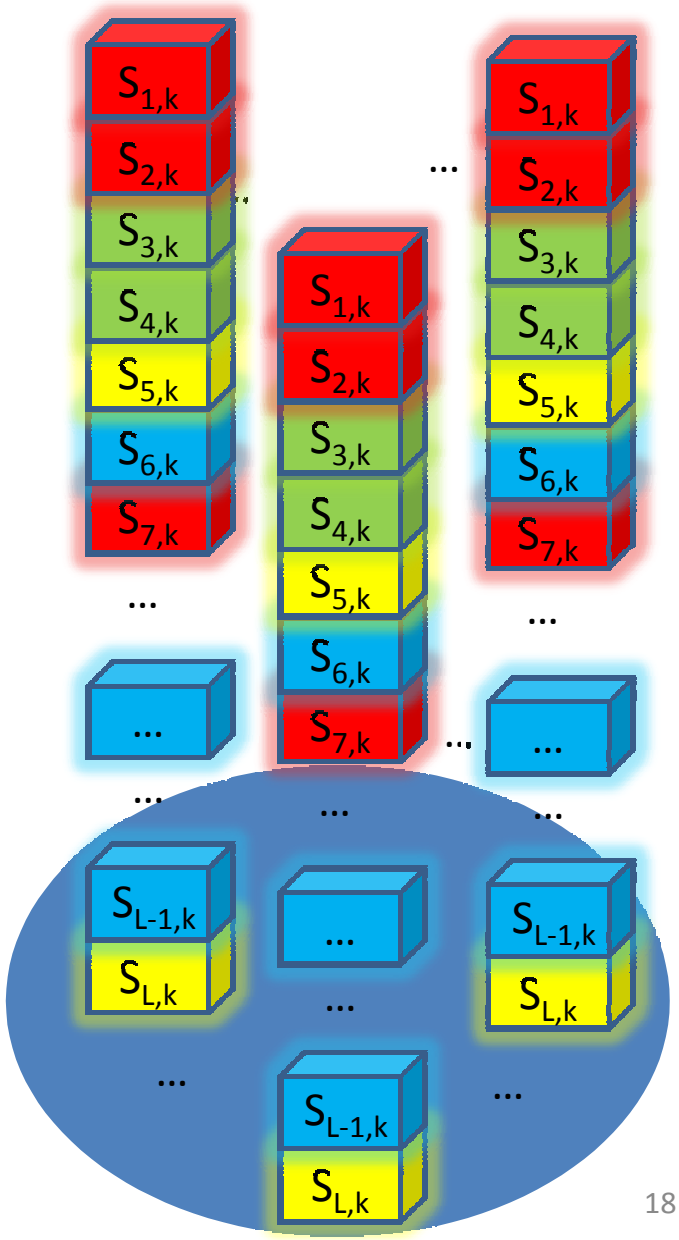
Notations

$s_{1,k}, s_{2,k}, ..., s_{L,k}$: the length - L prefix of the true complement ary DNAsequence in cluster k

$\mathbf{S}_k = (\mathbf{S}_{1,k}, \mathbf{S}_{2,k}, ..., \mathbf{S}_{L,k})$

$e.g.$

$\mathbf{S}_3 = (\mathbf{S}_{1,3}, \mathbf{S}_{2,3}, ..., \mathbf{S}_{60,3})$: length - 60 prefix matrix at 3 - th cluster



Cycle-dependent parameters

$$\Lambda_{t,k} = (1-d)\Lambda_{t-1,k} + (1-d)\Lambda_{t-1,k}\varepsilon$$
where $\varepsilon \sim N(0,\sigma^2)$

$$\Lambda_{t,k} = (1-d_t)\Lambda_{t-1,k} + (1-d_t)\Lambda_{t-1,k}\varepsilon_t$$

1st Layer: $\boldsymbol{I}_{t,k} \mid \boldsymbol{I}_{t-1,k}, \boldsymbol{S}_k, \Lambda_{t,k} \sim \mathcal{N}(\boldsymbol{\mu}_{t,k}, \| \boldsymbol{Z}^{\omega}_{t,k} \|_2^2 \boldsymbol{\Sigma})$

2nd Layer: $\Lambda_{t,k} \mid \Lambda_{t-1,k} \sim \mathcal{N}((1-d)\Lambda_{t-1,k}, (1-d)^2 \Lambda^2_{t-1,k} \sigma^2)$

Hyperparameter (prior): $\boldsymbol{S}_{t,k} \sim \text{Unif}(\{\boldsymbol{e}_A, \boldsymbol{e}_C, \boldsymbol{e}_G, \boldsymbol{e}_T\})$

where $\| \cdot \|_2$ denotes the 2-norm and

- $\boldsymbol{\mu}_{t,k} = \begin{cases} \boldsymbol{X}\boldsymbol{Z}^{\omega}_{t,k}, & \text{if } t=1 \\ \boldsymbol{X}\boldsymbol{Z}^{\omega}_{t,k} + \alpha(1-d)\boldsymbol{I}_{t-1,k}, & \text{if } t>1 \end{cases}$
- $\boldsymbol{Z}^{\omega}_{t,k} := \Lambda_{t,k}\boldsymbol{S}_k\boldsymbol{Q}^{\omega}_t$
- $\boldsymbol{Q} = (Q_{jt})$ is an L-by-L matrix with $Q_{jt} = [\boldsymbol{P}^t]_{o,j}$,
- transition matrix $\boldsymbol{P} = (P_{ij})$, where $0 \le i,j \le L$:

$$P_{ij} = \begin{cases} p, & \text{if } j = i, \\ 1-p-q, & \text{if } j = i+1, \\ q, & \text{if } j = i+2, \\ 0, & \text{otherwise.} \end{cases}$$

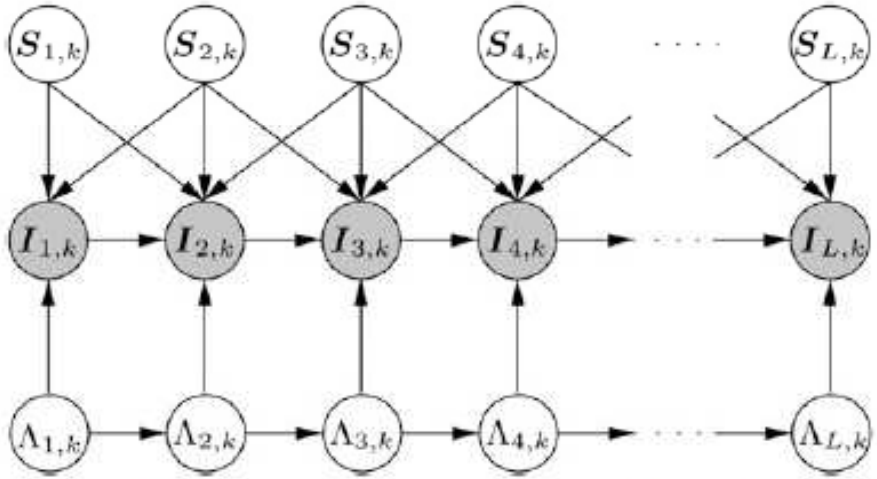


Figure 1. The graphical model corresponding to our base-calling algorithm for cluster k. The observed random variables are the intensities $\boldsymbol{I}_{t,k}$. Base-calling is done by finding the MAP estimates of $\boldsymbol{S}_{t,k}$. In this example, the window size is 3, with $l = r = 1$. See Methods for a detailed description.

$$\Theta = \{p, q, d, \alpha, \sigma^2, \boldsymbol{X}, \Sigma\}$$

$$d_t, \alpha_t, \sigma^2_t, \boldsymbol{X}_t, \Sigma_t$$

Estimation methods

- EM algorithm
  - E-step: Latent Variables (Λ_k, S_k)
    - Metroplis-Hastings algorithm (Random-Walk Metropolis) ≈ MCEM (Wei and Tanner, 1990)
  - M-step: α, σ^2, X, Σ
    - Steepest ascent method (updating one parameter at a time) ≈ ECM (Meng and Rubin, 1993)

➔ MCECM aglorithm

- d: assume d is independent of $\mu_{t,k}$
- Interior point method: p, q
- Simulated Annealing : $P(\Lambda_k, S_k | I_K, \Theta)$

Phred quality scores

The most common representation of uncertainty in base calling is quality score. Quality scores have been an important feature of modern sequencing platforms since the early days of the human genome project. The classic definition of a quality score was laid out in the implementation of the widely used traditional sequencing base-caller *phred* (Ewing and Green 1998). It defines a quality score $Q_{phred}(b)$ for each called base b as a scaled log of the error probability:

$$Q_{phred}(b) = -10 \log_{10} \mathbb{P}(B_t \neq b).$$

B_t: the base at position t

Statistical methods for using sequence reads:

- Li, Ruan, Durbin, 2008, *Genome Research*.

mapping software (MAQ) assembling short sequence reads to match a reference genome and making a final genotypic call using Bayesian calculations.

- Dalevi, Ivanova, Mavromatis, Hooper, Szeto et al., 2008, *Bioinformatics*

matching individual short reads sampled from the collective genome of a microbial community to protein families.

- Boyle, Guinney, Crawford,Furey, 2008, *Bioinformatics*

software employs kernel smoothing in converting sequencing reads into continuous signals along a chromosome. Useful to find transcription factor binding sites (ChIP-seq) or regions of open chromatin (DNase-seq).

- Zhang, Liu, Meyer, Jérôme, Johnson et al., 2008, *Genome Biology*

MACS (Model-based Analysis of ChIP-seq) that utilizes Poisson modeling and to capture local biases in the genome resulting in for more robust predictions of binding sites.

- Jiang and Wong (2009), *Bioinformatics*

Statistical inferences for isoform expression in RNA-Seq

With this Next Gen. sequencing technology one reads millions of short reads from the transcript population of interest and by mapping these reads to the genome, RNA-Seq produces digital (counts) rather than analog signals and offers the chance to detect novel transcripts. Obviously there are several protocols for transcript quantification.

Advantage: Measures transcription with a high precision and modeling transcription abundance.

Problem: The counts of reads fall into a locus of genome annotated with multiple isoforms.

Solution: May be one needs to determine the expression for isoforms.

Background: Mortazavi et al. (2008), *Nat. Methods* introduced transcript quantification as (RPKM)

RPKM: reads per killobase of the transcript per million mapped reads to the transcriptome.

Normalizing the counts of reads mapped (all exons belonging to) a gene against the transcript length and the sequencing depth this RPKM can compare the expression measures across different genes and different experiments.

So for different isoform expression, one can think it is the counts of reads mapped to a specific isoform normalized against the isoform length and sequence depth.

Difficulty: Most reads that are mapped to a gene shared by more than one isoform.

Solution: Hui and Wong (2009) paper provides a statistical model to describe how the mapped to the exons of the genes are related to the isoform specific expression and estimate isoform specific expression index and quantifies the uncertainty score.

- Let G be the set of genes;

- For $g \in G$, $F = \{f_{g,i} | i \in [1, n_g]\}$ is the set of isoforms for all possible isoforms for all genes, which stands for all different possible transcripts in the sample being sequenced.

- For any isoform $f \in F$, l_f = length of f

  k_f = number of copies of the transcripts in the form of isoform f

- The total length of the transcripts in the sample is $\sum_{f \in F} k_f l_f$

- If we assume every read is independently and uniformly sampled from all possible nucleotides in the sample then probability that a read comes from a isoform f is:

$$p_f = \frac{k_f l_f}{\sum_{f \in F} k_f l_f} = \theta_f k_f$$, say, where $\sum_{f \in F} \theta_f l_f$=1.

- Let w be the total number of mapped reads. The reads are sampled independently and uniformly. The number of reads coming from a region of length l in f is denoted by a R.V. $\boldsymbol{X \sim Bin(w, p = l\theta_f)}$

As $\boldsymbol{w}$ is large and $\boldsymbol{p}$ is small the binomial can be approximated by Poisson with $\boldsymbol{\lambda = lw\theta_f}$.

• Suppose there are m exons with lengths $L = [l_1, l_2, ..., l_m]$ and n isoforms with expressions $\Theta = [\theta_1, \theta_2, ...\theta_n]$

These isoforms share an exon j as a whole. Set of observations $X = \{X_s | s \in S\}$ falling into a region g can be modeled with a Poisson RV with parameter $\lambda = l_j w \sum_{i=1}^{n} c_{ij}\theta_i$, where $c_{ij} = 1$ if isoform i contains exon j and 0 otherwise.

For exon-exon junctions $\lambda = lw \sum_{i=1}^{n} c_{ij}c_{ik}\theta_i$, where l is the length of the isoform and w is the total no. of reads and j and k are the indices of the exons falling in the junction being investigated.

Estimate Θ :

MLE of Θ using the Poisson likelihood

1) For the single-isoform case it is

$$\widehat{\Theta} = x/a$$

where, x is the no. of reads falling into some region of length l.

$a = lw$, w is the total no. of reads.

2) Multiple Isoform case

Simple closed form solution does not exist. Numerical method (e.g. hill climbing) is applied to solve for the maximum likelihood estimation.

In this case authors show that the joint log likelihood of multiple isoforms is concave and so any local minimum is guaranteed to be global maximum.

Applications of NGS data involving interesting statistical modeling

- Choi, Nesvizhskii, Ghosh, Qin, 2009, *Bioinformatics*,

used NG ChIP-seq data together with array hybridization data towards enhancing the detection of transcription factor binding sites.

- Zang, Schones, Zeng, Cui, Zhao, Peng, 2009, *Bioinformatics*,

looked for spatial clusters of signals, for identification of ChIP enriched signals for histone modification profiles.

- Chu, Bunce, Hogge, Peters, 2009, *Bioinformatics*,

applied whole genome sequencing to diagnose the fetal genetic disease using cell-free DNA from maternal plasma samples in the first trimester of pregnancy.

- Cokus, Feng, Zhang, Chen et al.,2008, *Nature*,

used NG sequencing to identify novel components of the Arabidopsis for methylation.

- Quon and Morris, 2009, *Bioinformatics*

developed a statistical method to identify the primary origin of a cancer sample via next generation sequencing. This utilizes a detail profile of tissues of each primary origin and not a data based classifier.

Choi *et al.*, 2009 Joint analysis of ChIP-chip and ChIP-seq data

- ChIP-seq offers genome-wide coverage in a single base pair resolution at low cost

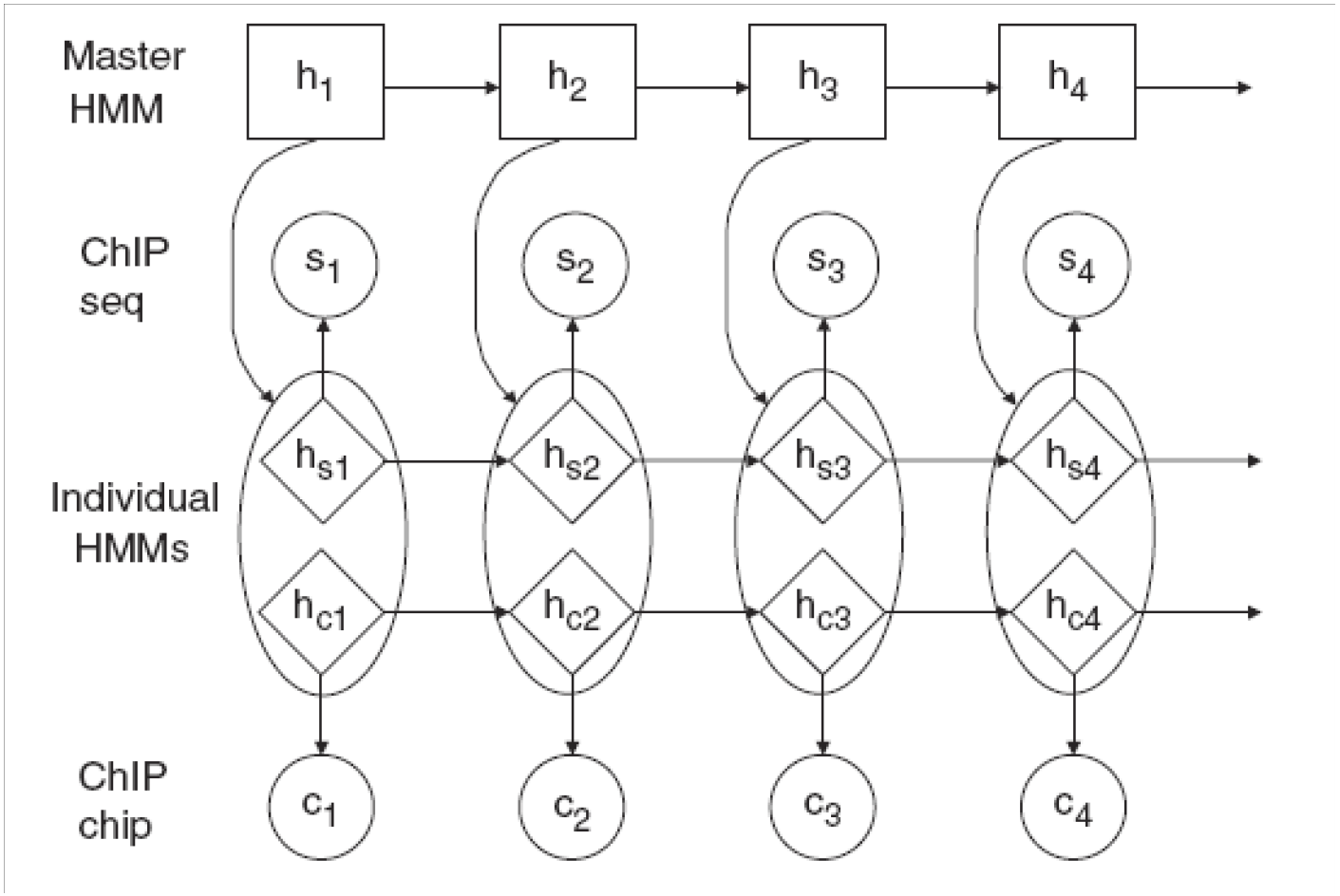
- With ChIP-seq, different mapping strategies may identify mutually exclusive peak regions as candidate binding sites.

- Massively parallel sequencing may not work well for all DNA fragments uniformly. Other mapping methods not relying on direct sequencing, e.g. ChIP-chip, can be a valuable source to complement the weakness of the sequencing technology.

- The peaks identified by ChIPseq are expected to form regions that are much sharper than those in ChIP-chip due to its superior resolution, whereas ChIP-chip tends to report broader regions with moderate significance including potential false positives.

- The signals from the two data sources have to be appropriately weighted in order to keep the overall false positive rates low and good sensitivity in the joint analysis. This is done through a mostly Bayesian strategy.

- Individual HMMs are fit to both ChIP-seq and ChIP-chip data which in turn are controlled by a master or hierarchical HMM consisting of either ChIP enriched or background states.


Master HMM
h1
h2
h3
h4
ChIP seq
s1
s2
s3
s4
Individual HMMs
hs1
hs2
hs3
hs4
hc1
hc2
hc3
hc4
ChIP chip
c1
c2
c3
c4

HMM in ChIP-chip: $C_t|h_{ct} = 1 \sim Unif$; $C_t|h_{ct} = 0 \sim Normal$

HMM in ChIP-seq: $s_t|h_{st} = 1 \sim Generalized\, Poisson$;

$$s_t|h_{st} = 0 \sim Zero - inflated\, Poisson$$

Master HMM: $(h_{st}, h_{ct})|h_t \sim Multinomial$

Posterior probabilities of the master states are computed and a state is declared to be ChIP enriched if this probability exceeds a given threshold, say, 90%.

• In a simulation experiment, the HHMM approach yielded the best ROC curves compared to either ChIP-chip or ChIP-seq alone or by naive attempts such as taking their intersection or union.

R and Bioconductor packages

- *rtracklayer* (Lawrence et al., 2009, Bioinformatics)

provides an interface between R and genome browsers. This package includes functions that import/export track data and control/query external genome browser sessions/views.

- *chipseq* (Kharchenko et al., 2008, Nature Biotechnology)

provides useful tools for design and analysis of ChIP-seq experiments and detection of protein-binding positions with high accuracy. These tools include functions that improve tag alignment and correct for background signals.

- *Biostrings 2* (Pages, 2009, Bioconductor)

allows users to manipulate big strings easily and fast by introducing new implementations and new interfaces into *Biostrings 1.*

- *ShortRead* package (Morgan et al., 2009, Bioinformatics)

provides useful tools for analyzing high-throughput data produced by Solexa, Roche 454, and other sequencing technologies. These tools include input and output, quality assessment, and downstream analysis functions.

- *IRanges* (Pages et al., 2009, R)

includes functions for representation, manipulation, and analysis of large sequences and subsequences of data as well as tools for attaching information to subsequences and segments.

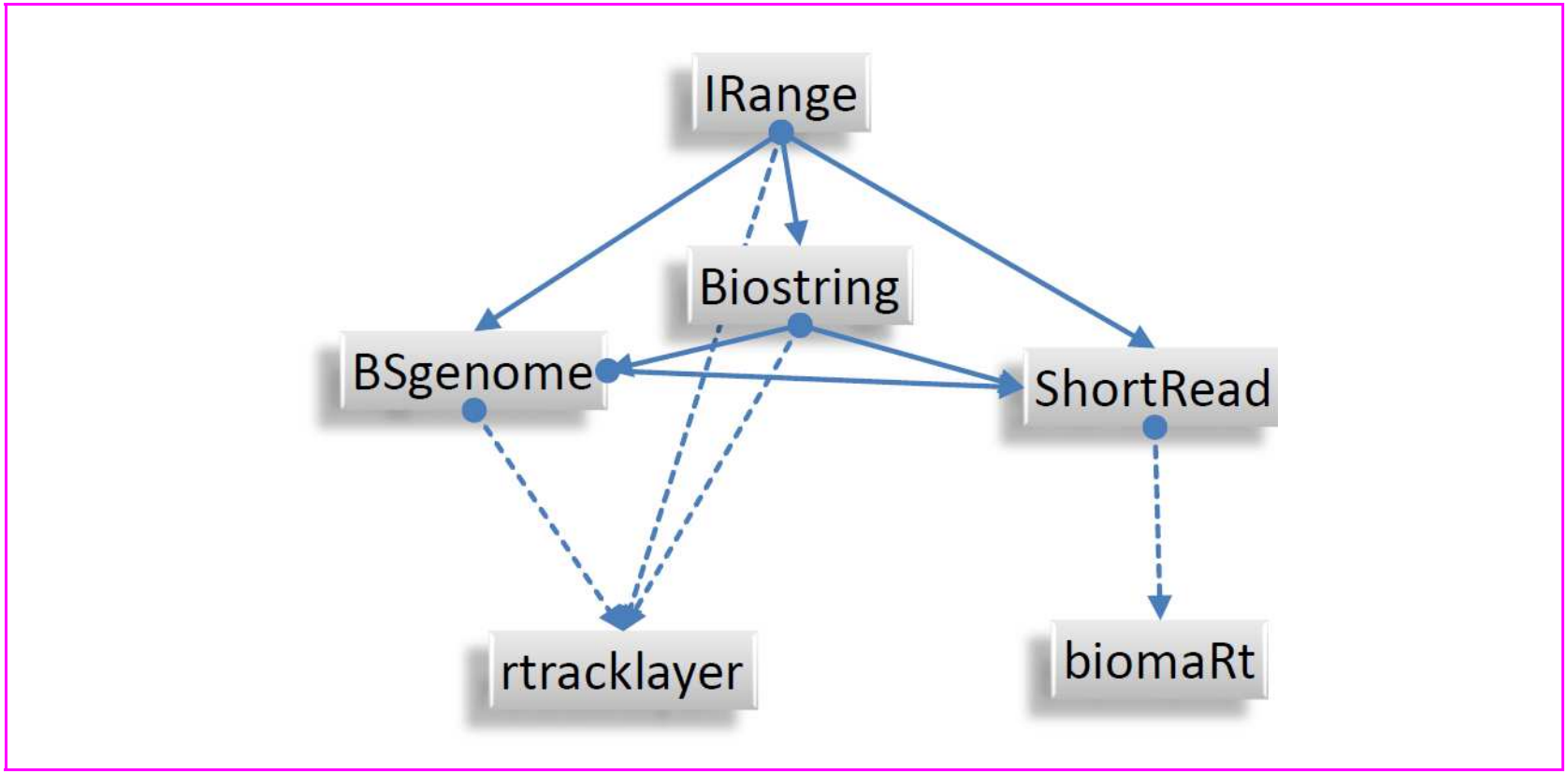
- *BSgenome* (Pages, 2009, R)

provides infrastructure for accessing, analyzing, creating, or modifying data packages containing full genome sequences of a given organism.

- *biomaRt* (Durinck et al., 2006, Bioinformatics)

allows users to connect to and search BioMart databases and integrates them with software in Bioconductor. This package includes functions that annotate identifiers with genetic information and that allow retrieval of data on genome sequences and single nucleotide polymorphisms.

The dependency among the released R/Bioconductor packages.



Challenges

1. Technical replicates versus biological replicates

2. Complex statistical methods (mostly Bayesian) - somewhat ad hoc ways of piecing together various statistical techniques. The overall statistical properties are difficult to assess.

3. For Bayesian methods, care must be taken to ensure the property of the posterior and the effect of prior parameter on the final answer.

4. High dimension; correlation; global error rate control (sensitivity, specificity, FDR, FNR)

5. Statistical designs

Opportunities

1. Availability of large amount of publicly available data with interesting scientific and clinical implications
2. Scope of development of novel statistical methods and software (including adaptation of existing methods)
3. Interdisciplinary research
4. Establishing statistical standards

Acknowledgements

Research Time Supported by:

NSF-DMS -0706965 (So Datta)

NSF DMS- 0805559 (Su Datta)

NCI R15-CA133844 (Su Datta)

NIH/NIEHS P30ES014443 (Su Datta)

Thank you very much for your attention!