Introduction to the Analysis of RNAseq Data

4/6/2011

These slides are adapted from slides provided by Peng Liu.

What is RNAseq?

- RNAseq refers to the method of using Next-Generation Sequencing (NGS) technology to measure RNA levels.
- NGS technology is an ultra-high-throughput technology to measure DNA sequences.

Some References for an Introduction to NGS

- Metzker, M.L. (2010). Sequencing technologies – the next generation. *Nature Reviews Genetics* 11:31.
- Current Topics in Genome Analysis 2010, Lecture: Next-Generation Sequencing Technologies Elliott Margulies, NHGRI, lecture on web at youtube.com, GenomeTV (http://www.genome.gov/12514288)
- Bullard, et al. (2010). Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments. *BMC Bioinformatics* 11: 94

Some Advantages of RNAseq over Microarrays

- Microarrays measure only genes corresponding to predetermined probes on a microarray while RNAseq measures any transcripts in a sample.
- With RNAseq, there is no need to identify probes prior to measurement or to build a microarray.
- RNAseq provides count data which may be closer, at least in principle, to the amount of mRNA produced by a gene than the fluorescence measures produced with microarray technology.

Some Advantages of RNAseq over Microarrays

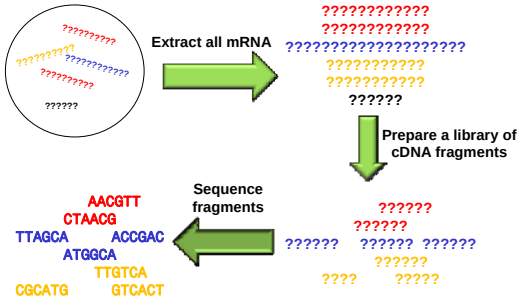
- RNAseq provides information about transcript sequence in addition to information about transcript abundance.
- Thus, with RNAseq, it is possible to separately measure the expression of different transcripts that would be difficult to separately measure with microarray technology due to cross hybridization.
- Sequence information also permits the identification of alternative splicing, allele specific expression, single nucleotide polymorphisms (SNPs), and other forms of sequence variation.

Examples of NGS Instrumentation

1. Roche 454 sequencer
2. Illumina Genome Analyzer (Solexa sequencing)
3. Applied Biosystems SOLiD sequencer

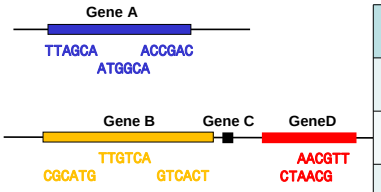
Technologies are similar but vary in speed, cost per base, and length of sequence reads.

Overview of RNAseq Procedure



7

Align Sequences to Genome



Gene ID	Sample1
A	3
B	3
C	0
D	2

For a given gene, the number of reads aligned to the gene measures its expression level.

8

Experiment Protocol

- Prepare library (collection of DNA fragments to be sequenced)
  - For analyzing gene expression, extract mRNA from samples and convert to DNA.
- Generate physical clusters of sequences.
- Sequence clusters.
- Perform bioinformatic analysis to determine genomic origin of sequences.

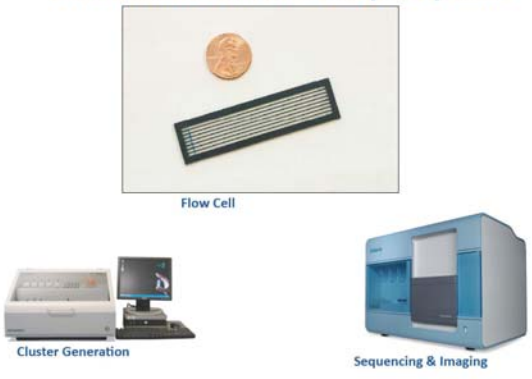
9

Illumina (Solexa) Sequencing by Synthesis Procedure

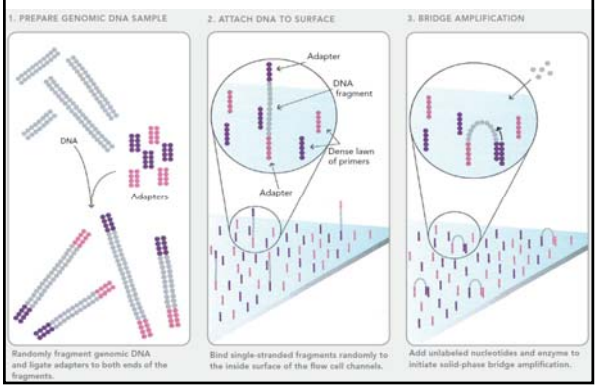
- http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf
- Handout from Illumina©

10

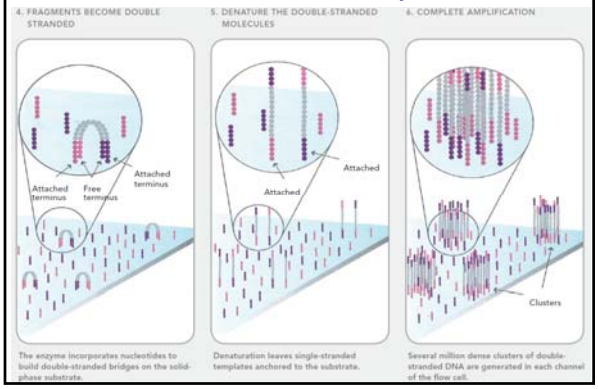
The Illumina Genome Analysis System



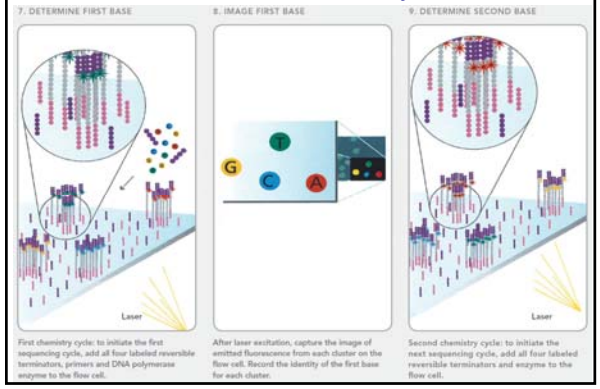
Solexa Genome Analyzer



Solexa Genome Analyzer



Solexa Genome Analyzer



Solexa Genome Analyzer

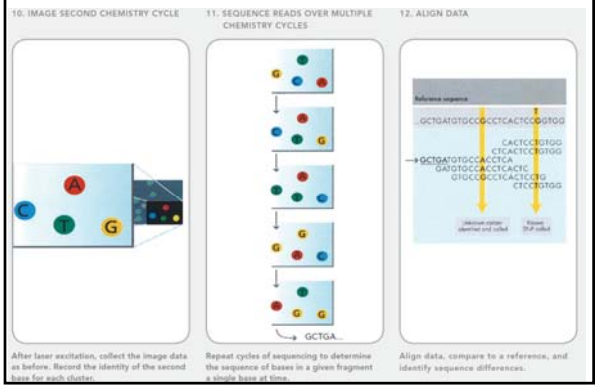
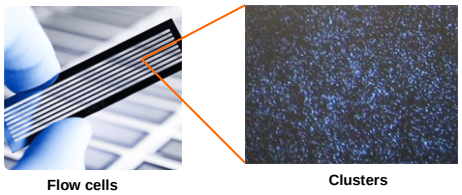


Image analysis



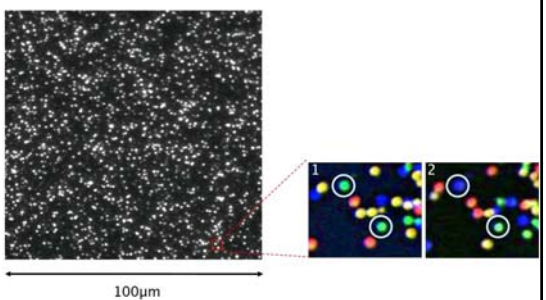
Flow cells **Clusters**

Images copied from http://www.vmsr.net/Sequencing_slides-Intro_Seminar.pdf

16

Solexa

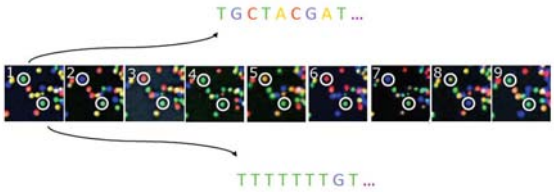
Sequencing by synthesis (SBS)



colony of ≈1000 single-stranded DNA templates

Copied from http://www.vmsr.net/Sequencing_slides-Intro_Seminar.pdf

Base Calling from Raw Data

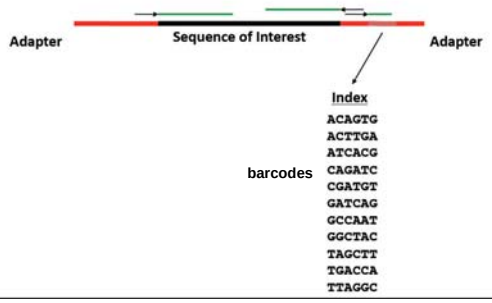


The identity of each base of a cluster is read off from sequential images.

Slide (though modified) Courtesy of Illumina Inc.

Multiplexing Samples

Paired End Indexing Strategy



Illumina Throughput

- **Each lane can sequence 20-30 million molecules**
  - 8 lanes = up to 240 million reads
- **36 bp reads suitable for counting based experiments**
- **Capable of up to 100bp paired end reads**
  - 50 Gigabases of sequence per run

Alignment

- Trapnell, C., Salzberg, SL, How to map billions of short reads onto genomes, Nature Biotechnology, 27(5), 2009.
- Abstract: "Mapping the vast quantities of short sequence fragments produced by next-generation sequencing platforms is a challenge. What programs are available and how do they work?"

21

Comments about Alignment

- A reference genome is used when available. Otherwise, there are methods for *de novo* sequencing.
- Typically, a maximum number of mismatches (1 or 2) are allowed when aligning reads.
- There are many challenges, such as dealing with alternative splicing and reads that match multiple places in the genome.
- Data in its rawest form is huge and requires substantial computing power to manage.

22

Example Dataset after Aligning Reads

Gene	Treatment 1			Treatment 2		
1	14	18	10	47	13	24
2	10	3	15	1	11	5
3	1	0	10	80	21	34
4	0	0	0	0	2	0
5	4	3	3	5	33	29
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
53256	47	29	11	71	278	339
Total	22910173	30701031	18897029	20546299	28491272	27082148

23

Analysis

- To determine if gene 1 is DE, we would like to know whether the proportion of reads aligning to gene 1 tends to be different for experimental units that received treatment 1 than for experimental units that received treatment 2.

14 out of 22910173		47 out of 20546299
18 out of 30701031	vs.	13 out of 28491272
10 out of 18897029		24 out of 27082148

24

Poisson Approximation to the Binomial Distribution

- Recall that if $Y \sim \text{Binomial}(n,\theta)$, then

  $Y \dot{\sim} \text{Poisson}(\lambda = n\theta)$ when n is large.

- Thus, we may choose to model the count for treatment i, gene j, and experimental unit k as

  $Y_{ijk} \sim \text{Poisson}(n_{ik}\theta_{ij})$,

  where n_{ik} is the total number of reads for treatment i experimental unit k.

25

Consider a Generalized Linear Model with Poisson Response and a Log Link

- Suppose $Y_{ijk} \sim \text{Poisson}(n_{ik}\theta_{ij})$ for all i, j, k and assume independence of counts within each gene j.
- Suppose $\log(n_{ik}\theta_{ij}) = \log(n_{ik}) + \log(\theta_{ij})$

  $= o_{ik} + \mu_j + \tau_{ij}$

  where o_{ik} is a known offset for all i and k and μ_j and τ_{ij} are unknown parameters for all i and j.

26

Alternative Offsets and Parameter Interpretation

- Although the offset $o_{ik} = \log(n_{ik})$ seems to be a natural choice, Bullard et al. (2010) argue that log of the 0.75 quantile of the count distribution for treatment i and experimental unit k is a more robust choice than $\log(n_{ik})$.
- Note that $\log(\theta_{ij}) = \mu_j + \tau_{ij}$ is a natural modeling choice for a CRD with an arbitrary number of treatments.
- The test for gene j differential expression between treatments i and i' has null hypothesis

$$H_{0j} : \tau_{ij} = \tau_{i'j}$$

27

Extending the Generalized Linear Model to Other Designs

- It is easy to extend this modeling strategy to, for example, a randomized complete block design.
- The only change is that we would use

  $\log(\theta_{ijk}) = \mu_j + \tau_{ij} + \beta_{jk}$ in place of $\log(\theta_{ij}) = \mu_j + \tau_{ij}$

- The null hypothesis for the test of gene j differential expression between treatments i and i' would remain

$$H_{0j} : \tau_{ij} = \tau_{i'j}$$

28

Overdispersion

- Recall that $Y \sim \text{Poisson}(\lambda)$ implies

  $E(Y)=\lambda$ and $\text{Var}(Y)=\lambda$.

- From the fit of our generalized linear model, we can estimate count means and variances and assess whether the Poisson mean-variance relationship holds.
- When the actual counts are more variable than we would expect based on the Poisson assumption, the data are said to be overdispersed.

29

Estimating Overdispersion

- If $E(Y)=\lambda$ and $\text{Var}(Y)=\varphi\lambda$, φ is said to be the dispersion parameter.
- When $\varphi>1$, the data are overdispersed.
- φ can be estimated by $-2 \log \Lambda / (n-p)$, where n is the number of observations, p is the number of free parameters (rank of the design matrix), and Λ is the likelihood ratio comparing our previous Poisson model with the saturated Poisson model that estimates $E(Y_{ijk})$ by Y_{ijk}.

30

Accounting for Overdispersion

- Suppose T is a test statistic that has an approximate chi-squared distribution with df_T degrees of freedom under the null hypothesis when there is no overdispersion (e.g., T could be the likelihood ratio test statistic for comparing full and reduced models).
- When there is overdispersion, use $F=T/(\hat{\phi}df_T)$ as a test statistic and assume that F has an approximate F distribution with df_T and n-p degrees of freedom under the null.
- Simulation should be used to check how well this works in practice.

31

```
#Data for an example gene.

cbind(trt,y,log.75q)

     trt   y  log.75q
[1,]   1  47 4.691348
[2,]   1  29 4.844187
[3,]   2  11 4.779123
[4,]   2  71 5.129899
[5,]   3 278 5.283204
[6,]   3 339 5.384495

#Fit the generalized linear model
#with Poisson response and log link.
#Use the log of the .75 quantile of
#the count distribution of each
#experimental unit as the offset.

o=glm(y~trt,family=poisson(link=log),offset=log.75q)
```

32

```
#Examine the results and test for evidence
#of differential expression across treatments.
summary(o)
Call:
glm(formula = y ~ trt, family = poisson(link = log), offset = log.75q)

Deviance Residuals:
      1       2       3       4       5       6
 1.9084  -1.9638  -4.5841   3.0785  -0.8775   0.8208

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.1331     0.1147  -9.878   <2e-16 ***
trt2         -0.1231     0.1592  -0.773    0.439
trt3          1.5297     0.1216  12.583   <2e-16 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 448.098  on 5  degrees of freedom
Residual deviance:  39.434  on 3  degrees of freedom
AIC: 81.82

Number of Fisher Scoring iterations: 5
```

33

```
a=anova(o,test="Chisq")
a

Analysis of Deviance Table

Model: poisson, link: log

Response: y

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                     5     448.10
trt   2   408.66         3      39.43 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

34

```
#Test for overdispersion.

1-pchisq(deviance(o),df.residual(o))

[1] 1.404665e-08

#Estimate the dispersion parameter.

phihat=deviance(o)/df.residual(o)
phihat

[1] 13.14457
```

35

```
#Test for evidence of differential expression
#across treatments while accounting for overdispersion.

Fstat=a[2,2]/a[2,1]/phihat
Fstat

[1] 15.54499

pvalue=1-pf(Fstat,a[2,1],a[2,3])
pvalue

[1] 0.02610609
```

36

```
#Test for evidence of differential expression
#between treatments 1 and 2
#while accounting for overdispersion.

full=o
reduced=glm(y~factor(c(1,1,1,1,2,2)),
            family=poisson(link=log),offset=log.75q)
a=anova(reduced,full)
a
Analysis of Deviance Table
Model 1: y ~ factor(c(1, 1, 1, 1, 2, 2))
Model 2: y ~ trt
  Resid. Df Resid. Dev Df Deviance
1         4     40.031
2         3     39.434  1  0.59681

Fstat=a[2,4]/a[2,3]/phihat
Fstat 0.04540376

pvalue=1-pf(Fstat,a[2,3],a[2,1])
pvalue 0.8449218
```

37

```
#Test for evidence of differential expression
#between treatments 1 and 3
#while accounting for overdispersion.

full=o
reduced=glm(y~factor(c(1,1,2,2,1,1)),
            family=poisson(link=log),offset=log.75q)
a=anova(reduced,full)
a
Analysis of Deviance Table
Model 1: y ~ factor(c(1, 1, 2, 2, 1, 1))
Model 2: y ~ trt
  Resid. Df Resid. Dev Df Deviance
1         4    269.942
2         3     39.434  1   230.51

Fstat=a[2,4]/a[2,3]/phihat
Fstat 17.5364

pvalue=1-pf(Fstat,a[2,3],a[2,1])
pvalue 0.02482486
```

38

```
#Test for evidence of differential expression
#between treatments 2 and 3
#while accounting for overdispersion.

full=o
reduced=glm(y~factor(c(1,1,2,2,2,2)),
            family=poisson(link=log),offset=log.75q)
a=anova(reduced,full)
a
Analysis of Deviance Table
Model 1: y ~ factor(c(1, 1, 2, 2, 2, 2))
Model 2: y ~ trt
  Resid. Df Resid. Dev Df Deviance
1         4     330.78
2         3      39.43  1   291.35

Fstat=a[2,4]/a[2,3]/phihat
Fstat 22.16476

pvalue=1-pf(Fstat,a[2,3],a[2,1])
pvalue 0.01813760
```

39