Sequencing data basics; GLMs

Statistics for Genomic Data Analysis

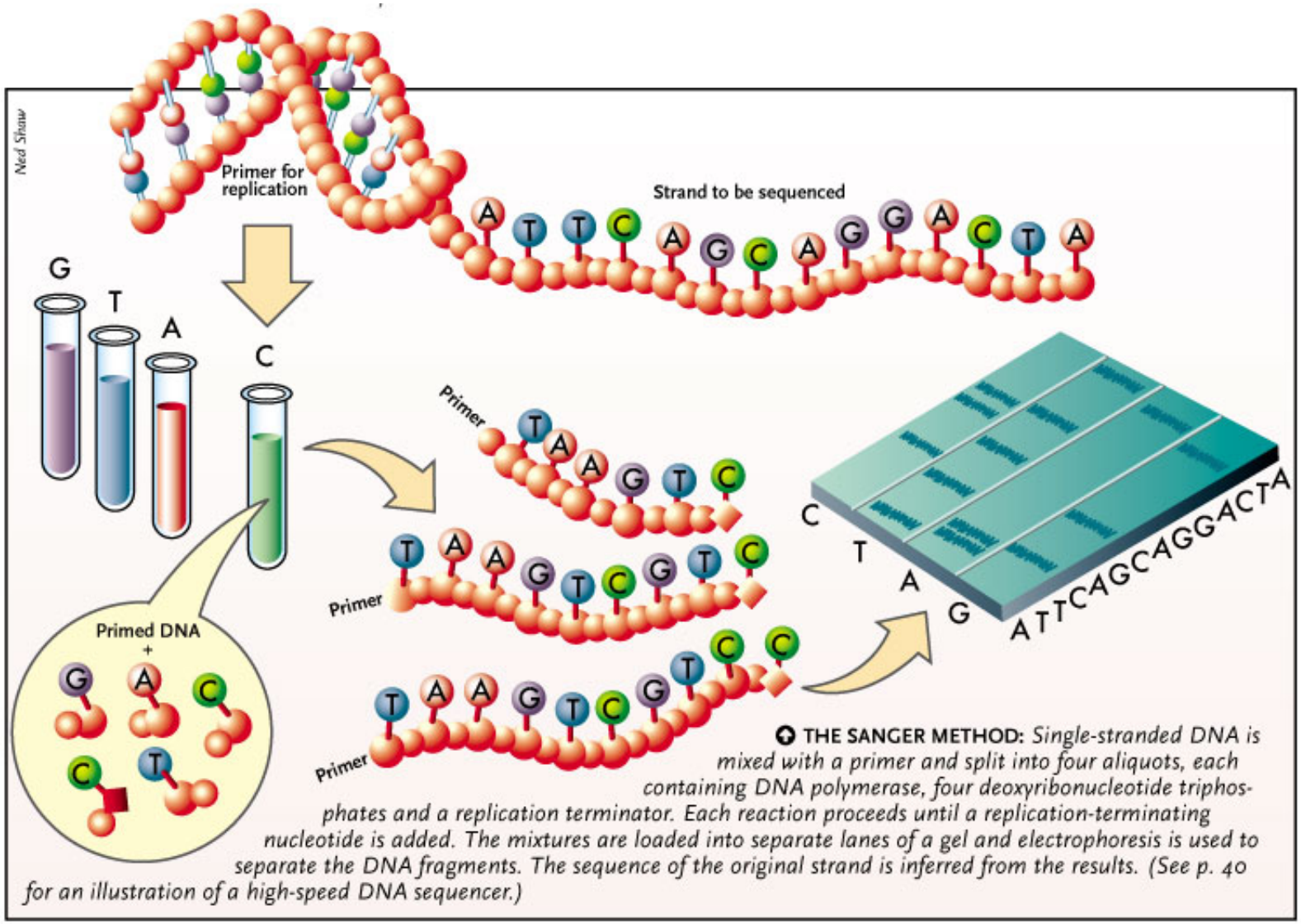
Lecture 9

http://moodle.epfl.ch/course/view.php?id=15271

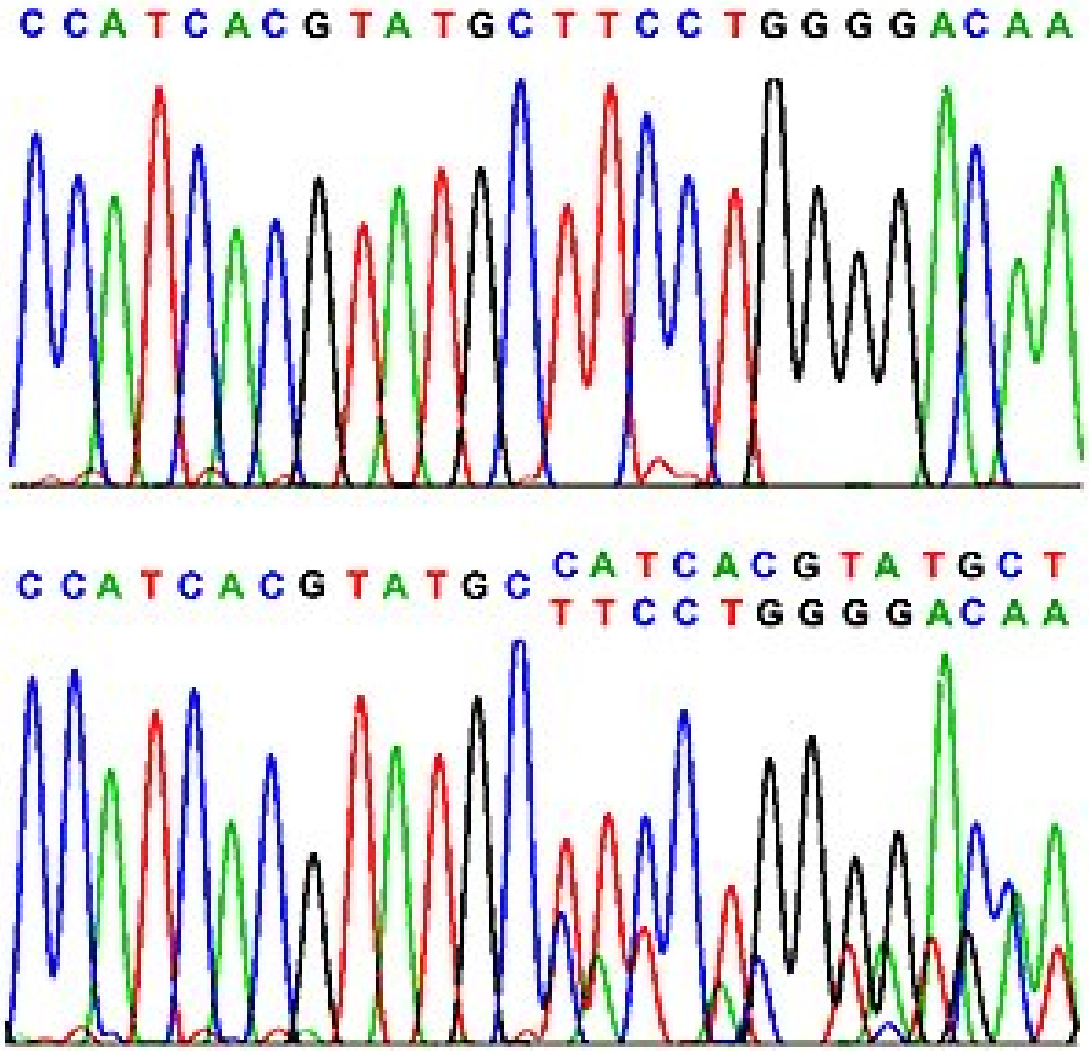
DNA sequencing

- (Automated) Sanger sequencing
  - 'first-generation' technology
  - F. Sanger, 1977
- Process :
  - bacterial cloning or PCR
  - template purification
  - labelling of DNA fragments using the chain termination method with energy transfer, dye-labelled dideoxynucleotides and a DNA polymerase
  - capillary electrophoresis
  - fluorescence detection
- Data : four-colour plots that reveal the DNA sequence

Sanger sequencing



Base-calling



Next-generation sequencing

- Several newer sequencing technologies
  - 'Next-generation sequencing' (NGS data)
  - 'Ultra high-throughput sequencing' (UHTS data)
- These newer technologies use various strategies that rely on a combination of template preparation, sequencing and imaging, and genome alignment and assembly methods
- Data : four-colour plots that reveal the DNA sequence
- Major advance : ability to produce a *large amount* of data relatively *cheaply*
- Expands experimental possibilities beyond just determining the order of bases

Limitations of microarrays

- Microarrays provide powerful technology to generate high-throughput data in several domains : :
  - gene expression, genotyping, transcription factor binding (ChIP-chip)
- However, they also have limitations :
  - require *a priori* knowledge of the genome
  - cross-hybridization between similar sequences restricts microarray analysis to *non-repetitive fraction of genome*
  - cross-hyb complicates analysis of related genes, alternatively spliced transcripts, etc.
  - relatively noisy, limited dynamic range : challenges in detection of low-abundance sequences, resolution of changes in high-abundance sequences
  - may need relatively large amounts of material, relying on amplification by PCR (can introduce bias)
  - reproducibility can be difficult to establish due to the variety of platforms

Some advantages of NGS

- Knowledge of genome annotation not required
- Material *directly sequenced* rather than interrogated by hybridization to user-defined sequences $\Longrightarrow$ removes (some) experimental bias and cross-hybridization issues
- Quantification based on counting sequence tags rather than relative measures between samples, thereby increasing the dynamic range of signal
- Require less starting material, reducing (or eliminating) relianc on PCR amplification
- Can simultaneously monitor RNAs from known and undefined genomic features (promoters, exons, non-coding RNAs (ncRNA) and enhancers)
- Hope is that reproducibility/combinability of results will be improved (since all data of same primary type – sequence counts)

Applications of NGS

- Sequence assembly (original application)
- Resequencing : The sequencing of part of an individual's genome in order to detect sequence differences between the individual and the standard genome of the species
- Gene expression : RNA-Seq
- SNP discovery and genotyping
- Variant discovery and quantification
- Transcription factor binding sites : ChIP-Seq
- Measuring DNA methylation

Death of microarrays ?

- Over the past few years, there have been several articles announcing the death of microarray technology
- Are sequencing technologies displacing microarrays ?
- To some degree yes, BUT : the technologies are rather complementary
- Cost differences (microarrays still cheaper)
- For a simple gene expression study experiment, microarrays are generally chosen (quick, low cost)
- For a study where a *large dynamic range of expression*, sequencing technologies would be preferred (increased sensitivity)
- Rule of thumb :
  - when sensitivity isn't limiting : microarrays
  - when sensitivity is important : (short read) sequencing technologies

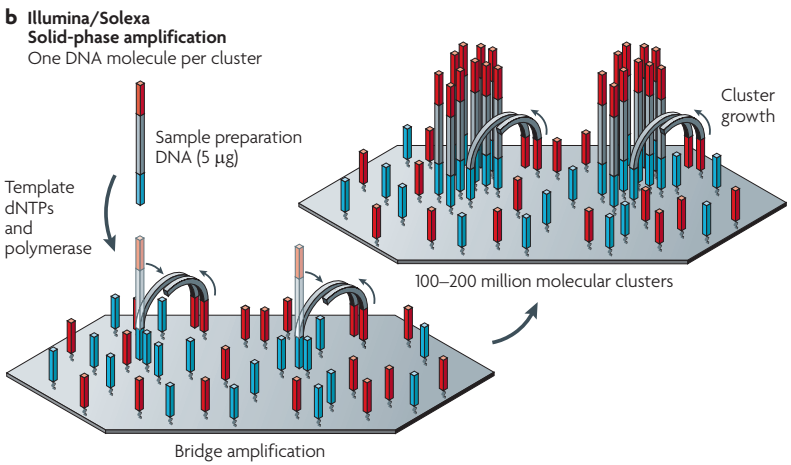
NGS data generation

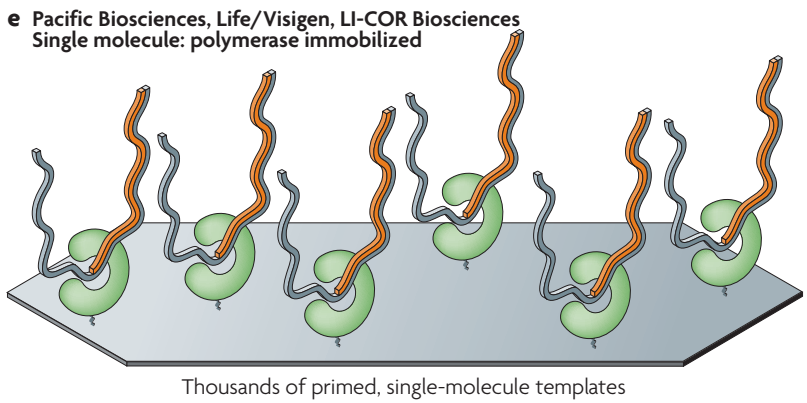
- Sequencing technologies incorporate methods that we can class as
  - template preparation
  - sequencing and imaging
  - data analysis
- Combination of specific protocols distinguishes different technologies
- Major technologies :
  - * Illumina HiSeq (older : Solexa)
  - 454 (Roche)
  - Applied Biosciences SOLiD
  - * Pacific Biosciences SMRT (single molecule real-time)

Template preparation

- Need robust methods capable of producing a *representative, non-biased* source of nucleic acid material from the genome under investigation
- Clonally amplified vs. single-molecule templates
- Current methods generally involve *randomly breaking* genomic DNA into smaller sizes
- *Templates* created from the pieces
- Typically, the template is *attached* or *immobilized* to a solid surface or support
- The *immobilization* of spatially separated template sites allows thousands to billions of sequencing reactions to be performed simultaneously

Template preparation

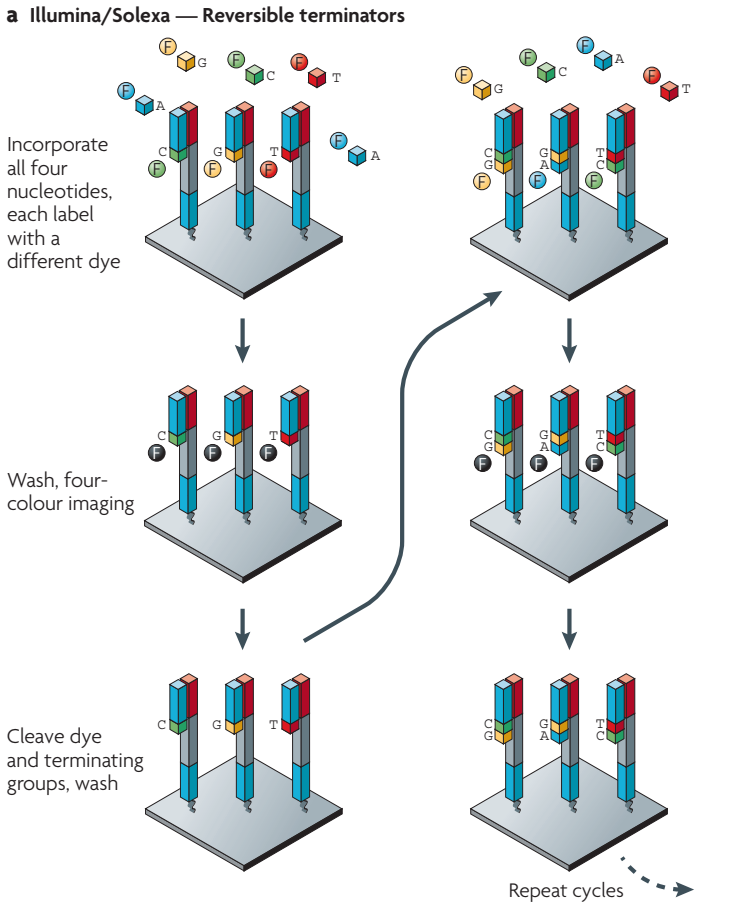




Sequencing and imaging

- Fundamental differences in sequencing different template types
- Clonal amplification results in a population of identical templates the undergo the sequencing reaction
- Once imaged, the observed signal is a *consensus* of the nucleotides or probes added to the identical templates for a given cycle
- Signal dephasing, which occurs with step-wise addition methods, increases fluorescence noise, causing base-calling errors and shorter sequence reads $\Longrightarrow$ need cycle efficiency
- (not an issue with single-molecule templates)
- Single molecules subject to different types of *errors* (e.g. deletion errors due to quenching effects between adjacent dye molecules)

Illumina sequencing



Comparison of sequencing technologies

Platform	Library/ template preparation	NGS chemistry	Read length (bases)	Run time (days)	Gb per run	Machine cost (US$)	Pros	Cons	Biological applications
Roche/454's GS FLX Titanium	Frag, MP/ emPCR	PS	330*	0.35	0.45	500,000	Longer reads improve mapping in repetitive regions; fast run times	High reagent cost; high error rates in homo-polymer repeats	Bacterial and insect genome *de novo* assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics
Illumina/ Solexa's GA_{II}	Frag, MP/ solid-phase	RTs	75 or 100	4‡, 9§	18‡, 35§	540,000	Currently the most widely used platform in the field	Low multiplexing capability of samples	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics
Life/APG's SOLiD 3	Frag, MP/ emPCR	Cleavable probe SBL	50	7‡, 14§	30‡, 50§	595,000	Two-base encoding provides inherent error correction	Long run times	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics
Polonator G.007	MP only/ emPCR	Non-cleavable probe SBL	26	5§	12§	170,000	Least expensive platform; open source to adapt alternative NGS chemistries	Users are required to maintain and quality control reagents; shortest NGS read lengths	Bacterial genome resequencing for variant discovery
Helicos BioSciences HeliScope	Frag, MP/ single molecule	RTs	32*	8‡	37‡	999,000	Non-bias representation of templates for genome and seq-based applications	High error rates compared with other reversible terminator chemistries	Seq-based methods
Pacific Biosciences (target release: 2010)	Frag only/ single molecule	Real-time	964*	N/A	N/A	N/A	Has the greatest potential for reads exceeding 1 kb	Highest error rates compared with other NGS chemistries	Full-length transcriptome sequencing; complements other resequencing efforts in discovering large structural variants and haplotype blocks

Data analysis pipeline

- Data are counts of short sequences (called 'reads')
- Quality control of data
- Match to reference sequence, read mapping
- Count/summarize number of reads per feature
- Statistical analysis (depends on the specific application)
- Next week, we will consider the problem of identifying differential gene expression from RNA-seq data

BioConductor sequencing resources

- `IRanges`, `GenomicRanges`, `genomeIntervals` : for range-based (e.g., chromosomal regions) calculation, data manipulation, and general-purpose data representation
- `Biostrings` : for alignment, pattern matching (e.g., primer removal), and data manipulation of large biological sequences or sets of sequences
- `ShortRead`, `Rsamtools` : for file I/O, quality assessment, and high-level, general purpose data summary
- `rtracklayer` : for import and export of tracks on the UCSC genome browser
- `edgeR`, `DESeq`, `baySeq`, `DEGseq`, `Genominator` : differential expression
- Use biocViews hierarchy to discover other packages : Software : AssayTechnologies : HighThroughputSequencing

BREAK

Modeling overview

- Want to capture important features of the *relationship between* a (set of) *variable(s)* and one or more *response(s)*
- Many models are of the form

$$g(Y) = f(\mathbf{x}) + \text{error}$$

- *Differences* in the form of g, f and distributional assumptions about the error term

Examples of models

- Linear : $Y = \beta_0 + \beta_1 x + \epsilon$
- Linear : $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$
- (Intrinsically) Nonlinear : $Y = \alpha x_1^{\beta} x_2^{\gamma} x_3^{\delta} + \epsilon$
- Generalized Linear Model (e.g. Binomial) : $\log \frac{p}{1-p} = \beta_0 + \beta_1 x + \beta_2 x_2$
- Proportional Hazards (in Survival Analysis) : $h(t) = h_0(t) \exp(\beta x)$

Linear modeling

- A simple linear model : $E(Y) = \beta_0 + \beta_1 x$
- Gaussian measurement model : $Y = \beta_0 + \beta_1 x + \epsilon, \epsilon\ N(0, \sigma^2)$
- More generally : $Y = X\beta + \epsilon$, where Y is $n \times 1$, X is $n \times p$, β is $p \times 1$, ϵ is $n \times 1$, often assumed $N(0, \sigma^2 I_{n \times n})$

Analysis of designed experiments

- An important use of linear models
- Define a (design) matrix X so that for response variable Y :

$$E(Y) = X\beta,$$

  where β is a vector of *parameters* (or contrasts)
- Many ways to define design matrix/contrasts

Model fitting

- For the standard (*fixed effects*) linear model, estimation is usually by *least squares*
- Can be more complicated with *random effects* or when x-variables are subject to measurement error as well

Model checking

- Examination of *residuals*
  - Normality
  - Time effects
  - Nonconstant variance
  - Curvature
- Detection of *influential observations*

Linear regression model (again)

- Linear model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

- Another way to write this :

$$Y \sim N(\mu, \sigma^2), \quad \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- Suitable for a *continuous* response
- **NOT** suitable for a *binary* response
- **NOT** suitable for a *count* data

Modified model

- Instead of modeling the response directly, could instead model some function of the response (here, a count)
- i.e., Instead of modeling the expected response *directly* as a linear function of the predictors, model a *suitable transformation*
- For count data, this is often taken to be the *log* transformation

Count response in a linear model

- In a standard linear model, the *response variable* is modeled as a *normally distributed*
- However, if the response variable is a *count*, it does not make sense to model the outcome as normal
- Generalized linear models (GLMs) are an extension of linear models to model non-normal response variables
- A GLM consists of three components :
  - A *random component*, specifying the conditional distribution of the response variable, Y_i, given the values of the explanatory variables in the model
  - A *linear predictor*
  - A smooth and invertible linearizing *link function*
- We might consider *Poisson regression* for a count response

Generalized linear models : some theory

- Allows unified treatment of statistical methods for several important classes of models
- Response Y assumed to have *exponential family distribution* :

$$f(y) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$$

- For a standard linear model

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon, \text{ with } \epsilon \sim N(0, \sigma^2)$$

- The *expected response* is $E[Y \mid x] = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$
- Let η denote the *linear predictor* $\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$
- For a standard linear model, $E[Y \mid x] = \eta$
- In a *generalized linear model*, there is a *link function* g between η and the expected response :

$$g(E[Y \mid x]) = \eta$$

- For a standard linear model, $g(y) = y$ (*identity link*)

Link function for count data

- We can model the count data $Y_i \sim Pois(\mu_i),\ i = 1, \ldots, n$
- Want to relate the mean μ_i to one or more *covariates* (for example, treatment/control status)
- A convenient link function in this case is the log :

$$\log \mu_i = \eta = x_1^T \beta$$

- Using a log link ensures that the fitted values of μ_i will remain in the parameter space $[0, \infty)$
- A Poisson model with a log link is sometimes called a *log-linear model*

Link function : examples

	Family Name				
Link	binomial	Gamma	gaussian	inverse.gaussian	poisson
logit	D				
probit	•				
cloglog	•				
identity		•	D		•
inverse		D			
log		•			D
1/mu^2				D	
sqrt					•

Variance function

- The Poisson distributions are a discrete family with probability function indexed by the rate parameter $\mu > 0$:

$$p(y) = \frac{e^{-\mu}\mu^{y}}{y!}, \quad y = 0, 1, 2, \ldots$$

- Under the Poisson model :

$$E[Y_i] = Var(Y_i) = \mu_i$$

- Real data are often *overdispersed*, exhibiting more variation than allowed by the Poisson model

Variance function

Overdispersion usually handled with an alternative model :

- *Quasi-Poisson Model* : Assume $Var(Y_i) = \phi \mu_i$ and estimating the *scale parameter* ϕ
- *Zero-Inflated Poisson Model* : for modeling the case when there are too many '0' values
- **Negative Binomial Model** : Can arise from a two-stage model :

$$Y_i \sim Pois(\mu_i^*) \qquad \mu_i* \sim \Gamma(\mu_i/\omega, \omega)$$

Then $Y_i \sim NegBin$, with $E[Y_i] = \mu_i$ and $Var(Y_i) = \mu_i + \mu_i^2/\omega$

Analogous to linear regression

- The link function g has many of the desirable properties of a linear regression model :
  - Mathematically convenient and flexible
  - Can meaningfully interpret parameters
  - Linear in the parameters
  - A difference : Error distribution not normal

Fitting the model

- For linear regression, typically use *least squares*
- For count data, the 'nice' statistical properties of least squares estimators no longer hold
- The general estimation method that leads to least squares (for normally distributed errors) is *maximum likelihood*
- Write out the likelihood, take the derivative, set equal to zero and solve
- Estimating equations typically nonlinear functions of the regression parameters so must be solved numerically (IRLS)

Assessing model fit

- In linear regression, an anova table partitions SST, the total sum of squared deviations of observations about their mean, into two parts :
  - SSE, or residual (observed - predicted) sum of squares
  - SSR, or regression sum of squares
- Large SSR suggests the explanatory variable(s) is(are) important
- Use same guiding principle in Poisson regression : compare observed response to predicted response obtained from models with/without the variable(s)
- Comparison based on log likelihood function

Differential gene expression for NGS data

- Several BioConductor packages for identifying differential expression from NGS data
- These mostly use the negative binomial model, since the counts are typically over-dispersed compared to the Poisson model
- Next week, we will look more closely at the approach used in the `edgeR` package
- `edgeR` uses an overdispersed Poisson model to account for both biological and technical variability, and uses empirical Bayes methods to moderate the degree of overdispersion across transcripts
- This is the same approach that we have already seen in `limma` (same developer group)