Introducing bioinformatics

NextGen sequencing and analysis tools

Dr. Mark Peterson

2014-Jan-20

An overview

An overview

1. Biology I
2. Chemistry & Physics

An overview

1. Biology I
2. Chemistry & Physics
3. Computers

An overview

1. Biology I
2. Chemistry & Physics
3. Computers
4. Stats

An overview

1. Biology I
2. Chemistry & Physics
3. Computers
4. Stats
5. Biology II

An overview

1. Biology I
2. Chemistry & Physics
3. Computers
4. Stats
5. Biology II
6. Facebook and Twitter

An overview

1. Biology I
2. Chemistry & Physics
3. Computers
4. Stats
5. Biology II
6. Facebook and Twitter
7. Conclusions

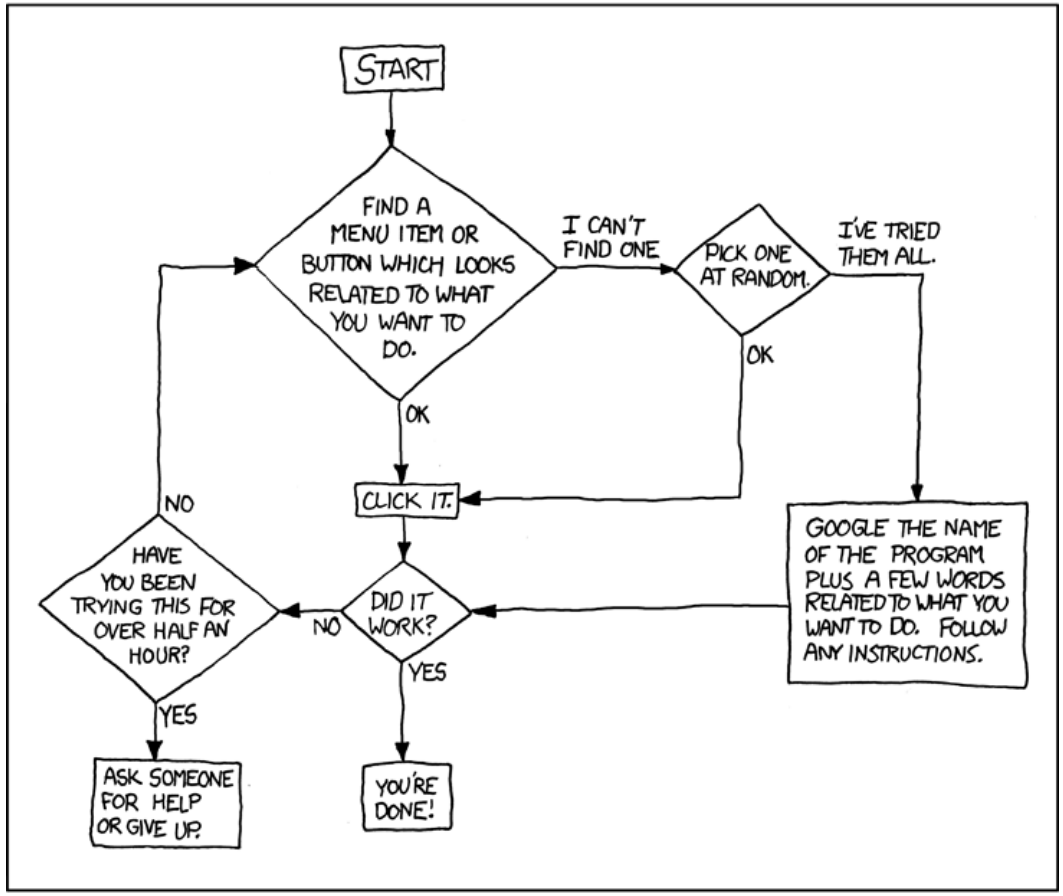
Most of genomics work

Computers

Computers

DEAR VARIOUS PARENTS, GRANDPARENTS, CO-WORKERS, AND OTHER "NOT COMPUTER PEOPLE."

WE DON'T MAGICALLY KNOW HOW TO DO EVERYTHING IN EVERY PROGRAM. WHEN WE HELP YOU, WE'RE USUALLY JUST DOING THIS:



PLEASE PRINT THIS FLOWCHART OUT AND TAPE IT NEAR YOUR SCREEN. CONGRATULATIONS; YOU'RE NOW THE LOCAL COMPUTER EXPERT!

xkcd.com/627

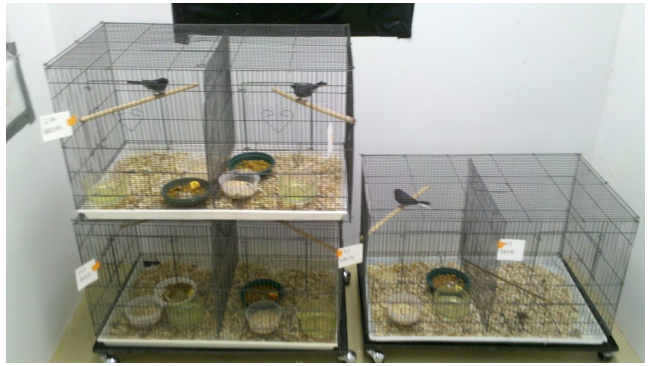
Experimental Design

Experimental Design

- Study question
  - How are males and females different?

Experimental Design

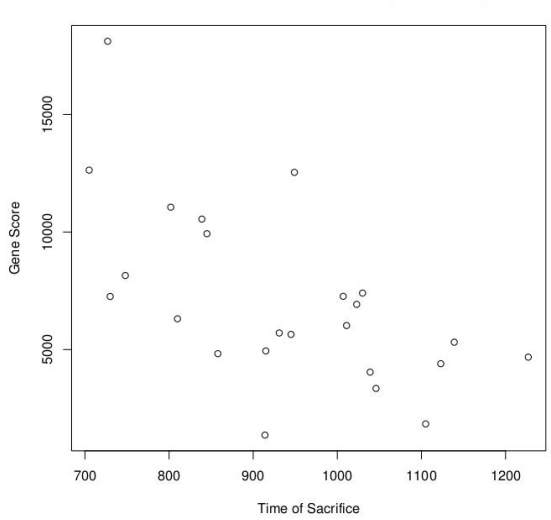
- Study question
  - How are males and females different?
- Sample size
  - How many can I get?



Sample Collection

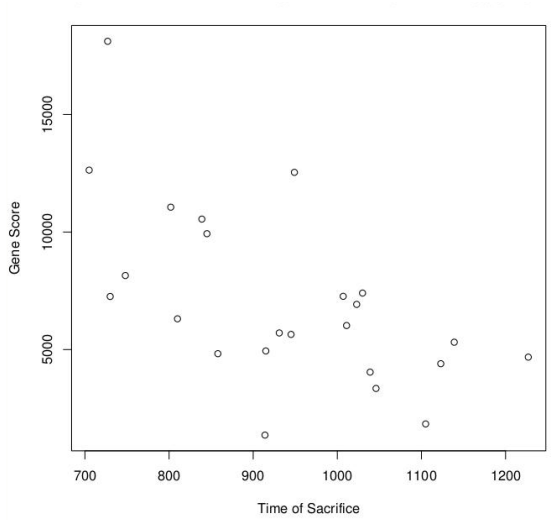
Sample Collection

- Covariates
  - Will time of death affect expression?



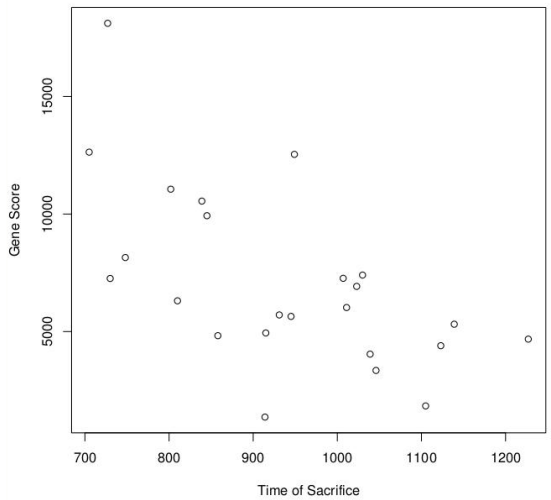
Sample Collection

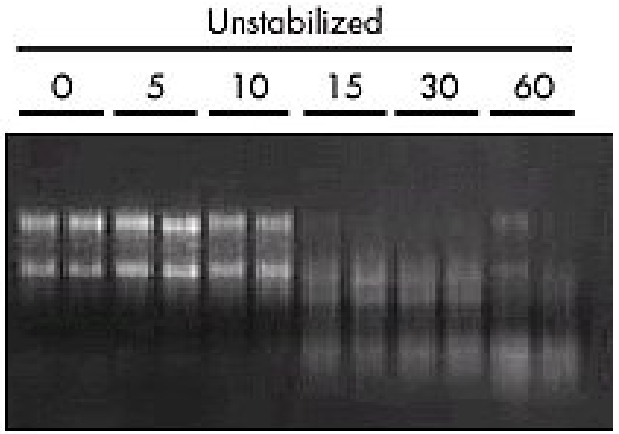
- Covariates
  - Will time of death affect expression?
- RNA degrades rapidly
  - Deep freeze immediately



Sample Collection

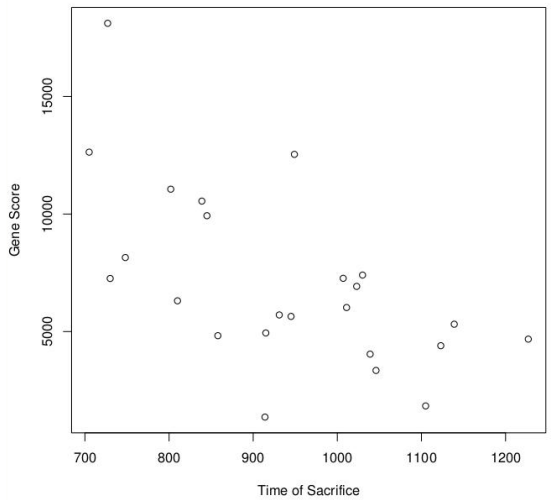
- Covariates
  - Will time of death affect expression?
- RNA degrades rapidly
  - Deep freeze immediately

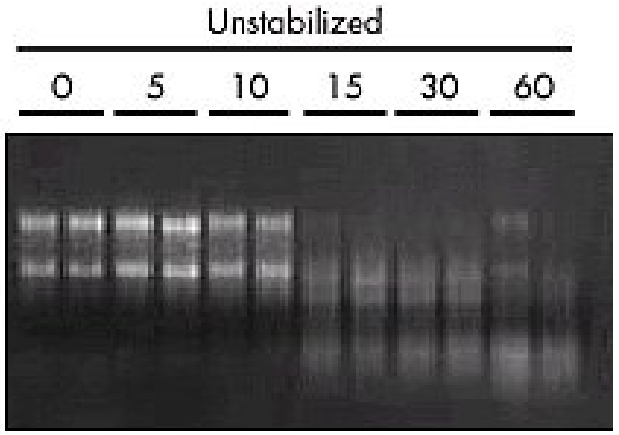




Sample Collection

- Covariates
  - Will time of death affect expression?
- RNA degrades rapidly
  - Deep freeze immediately
- Get enough tissues
  - Balance time and questions

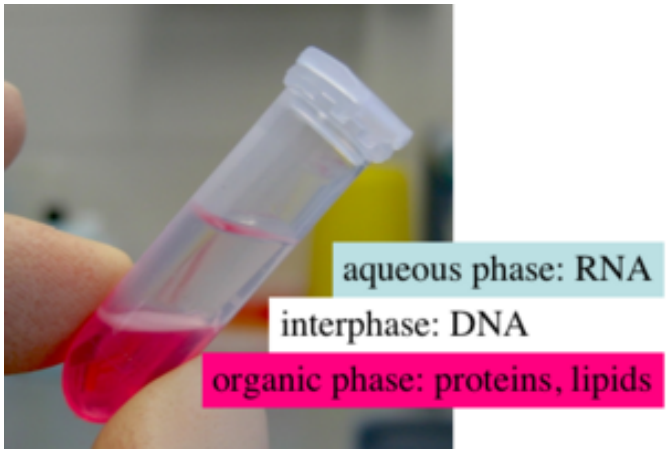




Sample prep

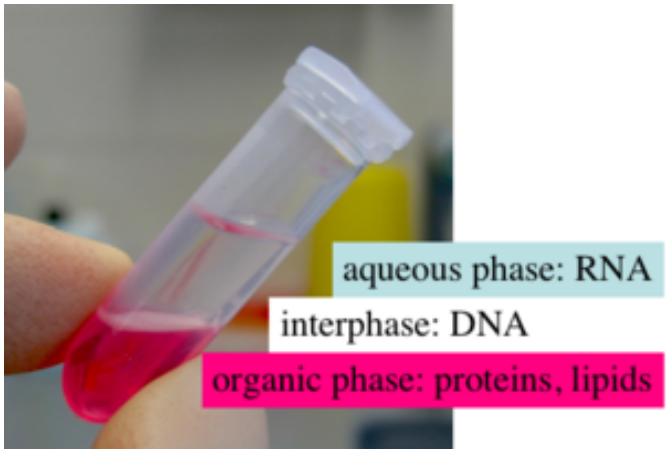
Sample prep

- Extract RNA
  - Phenol-chloroform, multi-step



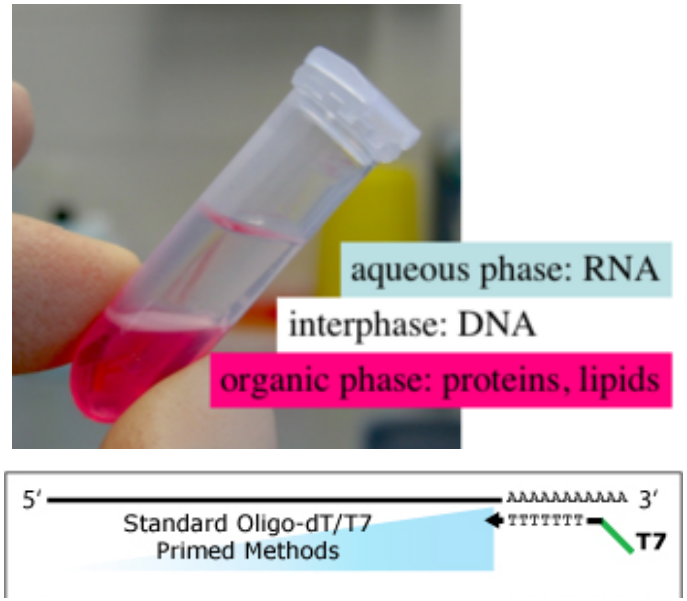
Sample prep

- Extract RNA
  - Phenol-chloroform, multi-step
- Deplete rRNA



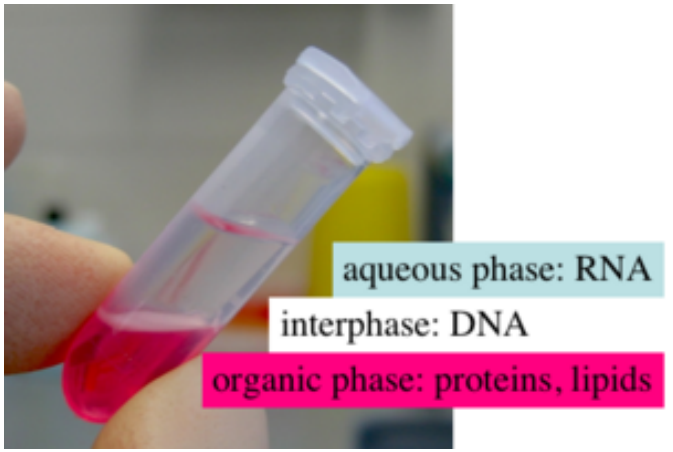
Sample prep

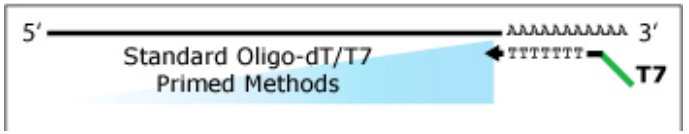
- Extract RNA
  - Phenol-chloroform, multi-step
- Deplete rRNA
  - Select poly-A



Sample prep

- Extract RNA
  - Phenol-chloroform, multi-step
- Deplete rRNA
  - Select poly-A
  - Use magnets to target rRNA

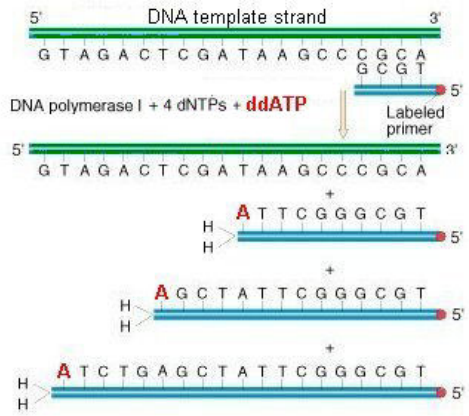




Sanger Sequencing

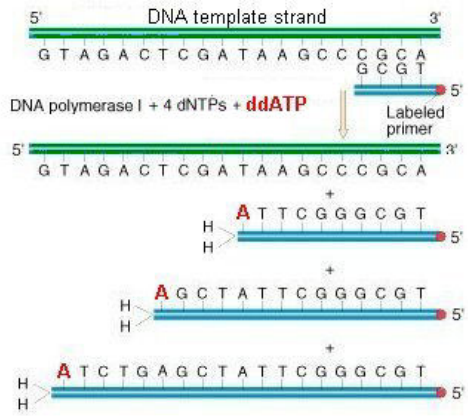
Sanger Sequencing

- Chain terminal sequencing
  - Compare length of fragments



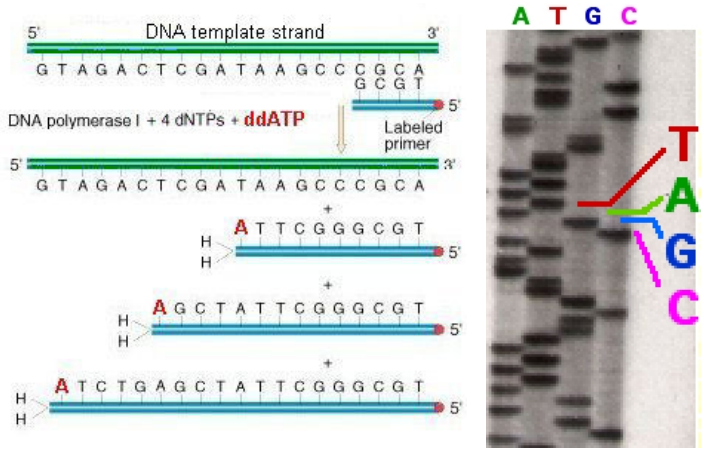
Sanger Sequencing

- Chain terminal sequencing
  - Compare length of fragments
- Measure size of each fragment



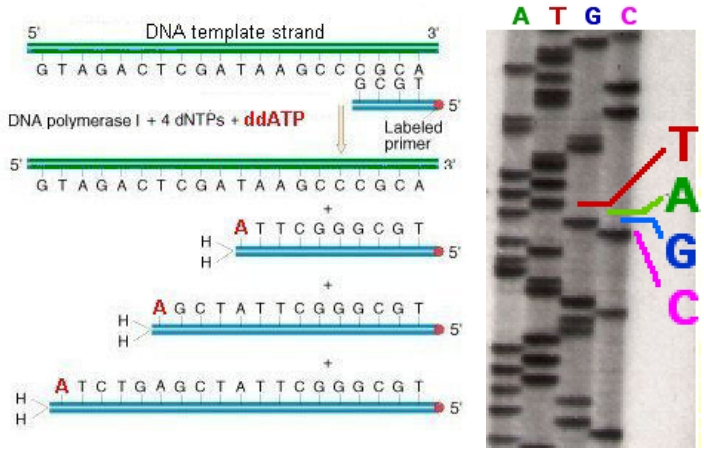
Sanger Sequencing

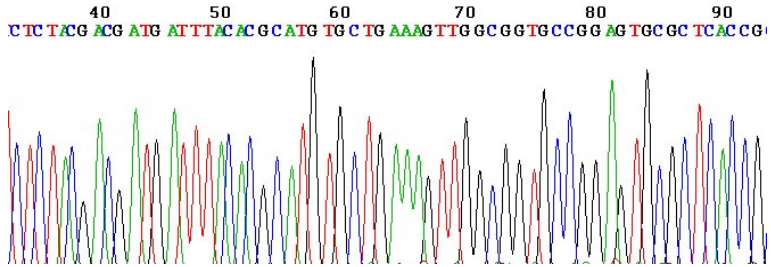
- Chain terminal sequencing
  - Compare length of fragments
- Measure size of each fragment
  - Gel, one reaction per base



Sanger Sequencing

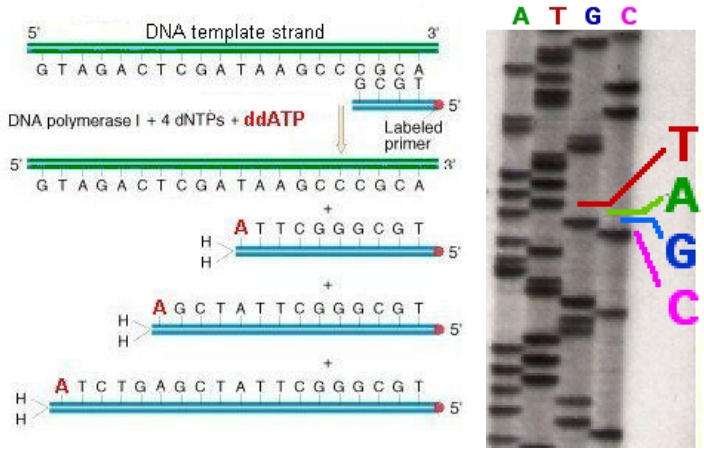
- Chain terminal sequencing
  - Compare length of fragments
- Measure size of each fragment
  - Gel, one reaction per base
  - Capilary, single reaction with color

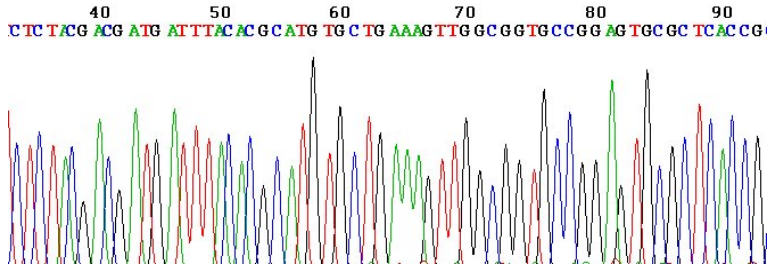




Sanger Sequencing

- Chain terminal sequencing
  - Compare length of fragments
- Measure size of each fragment
  - Gel, one reaction per base
  - Capilary, single reaction with color
- Slow, expensive, requires specifc primers





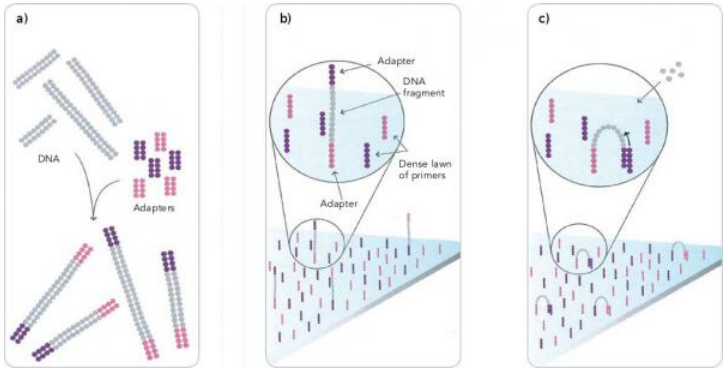
Illumina sequencing

Illumina sequencing

- Parallel sequencing
  - Many simultaneous reactions

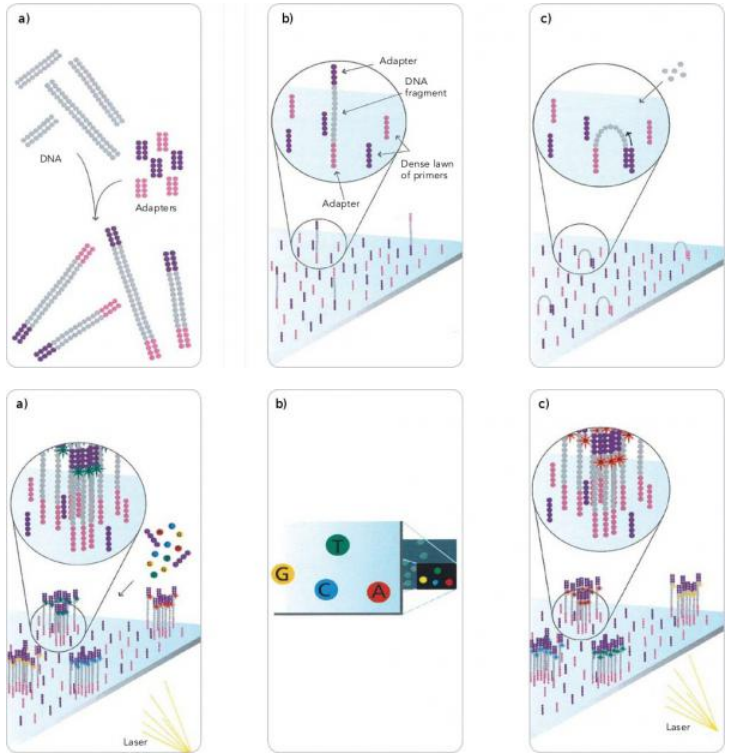
Illumina sequencing

- Parallel sequencing
  - Many simultaneous reactions
- Bind to plate
  - PCR replicate clusters



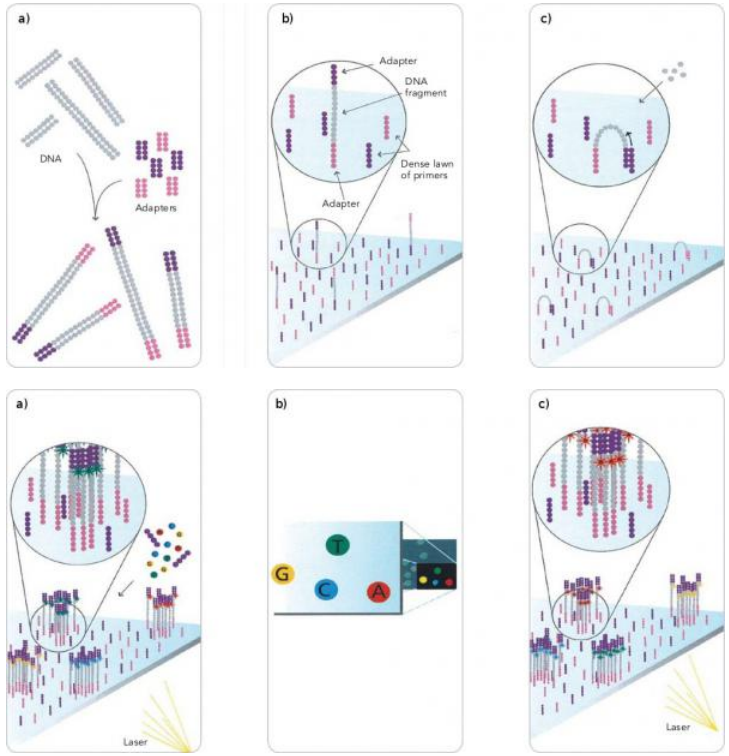
Illumina sequencing

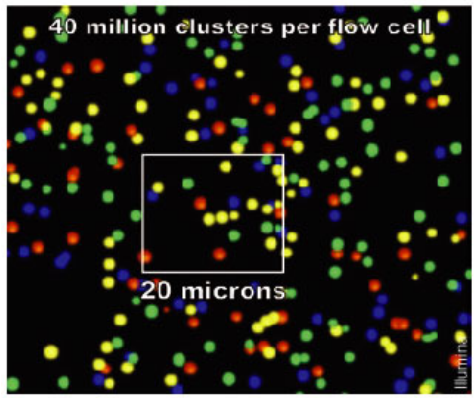
- Parallel sequencing
  - Many simultaneous reactions
- Bind to plate
  - PCR replicate clusters
- Like Sanger, use terminators
  - However, these are reversible



Illumina sequencing

- Parallel sequencing
  - Many simultaneous reactions
- Bind to plate
  - PCR replicate clusters
- Like Sanger, use terminators
  - However, these are reversible
- Read individual spots





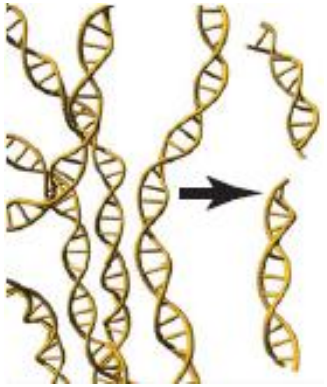
454 sequencing

454 sequencing

- A different approach

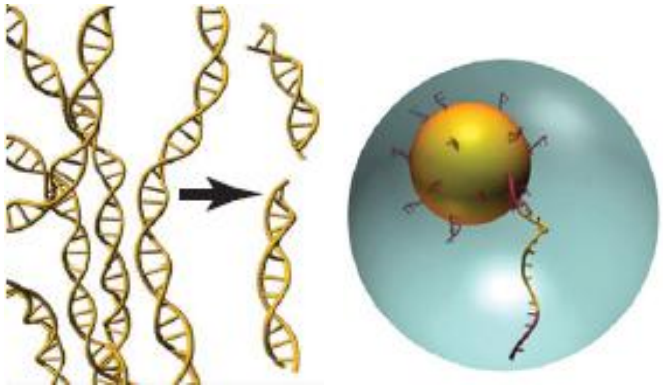
454 sequencing

- A different approach
- Random cleaving
  - Add tags



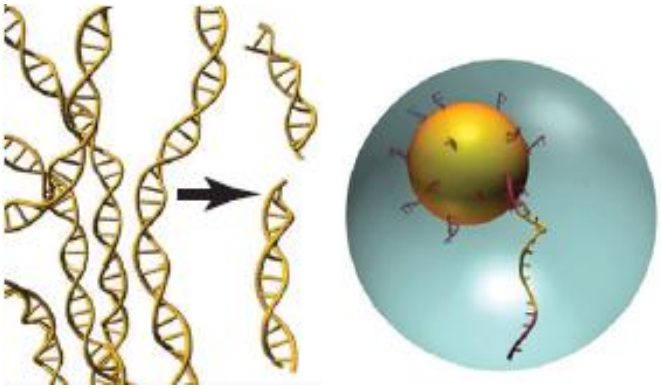
454 sequencing

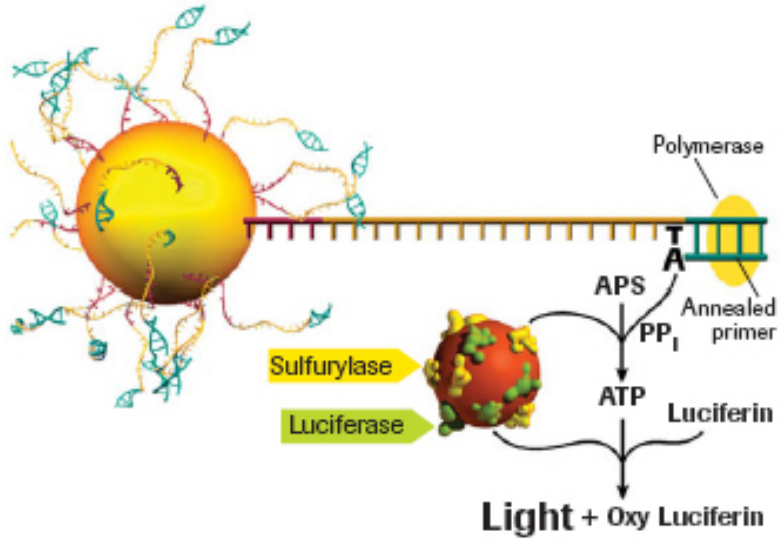
- A different approach
- Random cleaving
  - Add tags
- Emulsion PCR
  - Clonally amplify for reading



454 sequencing

- A different approach
- Random cleaving
  - Add tags
- Emulsion PCR
  - Clonally amplify for reading
- Pyrosequencing
  - Read flashes





PacBio sequencing

PacBio sequencing

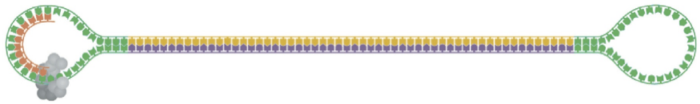
- Single Molecule Real Time (SMRT)

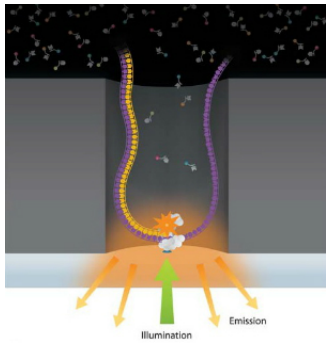
PacBio sequencing

- Single Molecule Real Time (SMRT)
- Circularizes, reads single molecule
  - High (~10%) error
  - Uses consensus from multiple reads

PacBio sequencing

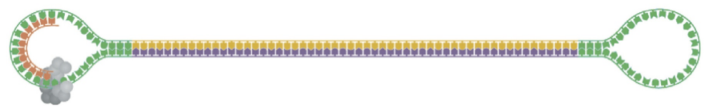
- Single Molecule Real Time (SMRT)
- Circularizes, reads single molecule
  - High (~10%) error
  - Uses consensus from multiple reads
- Gets *incredibly* long reads, starting 2013-Oct-03
  - Mean length 8,500 bases
  - N50 of 10k bases
  - Max reads of 30k bases

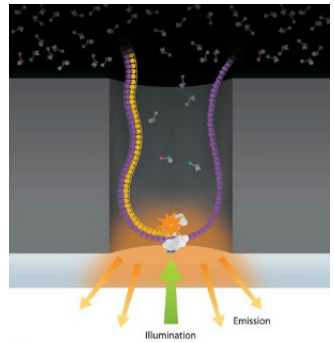


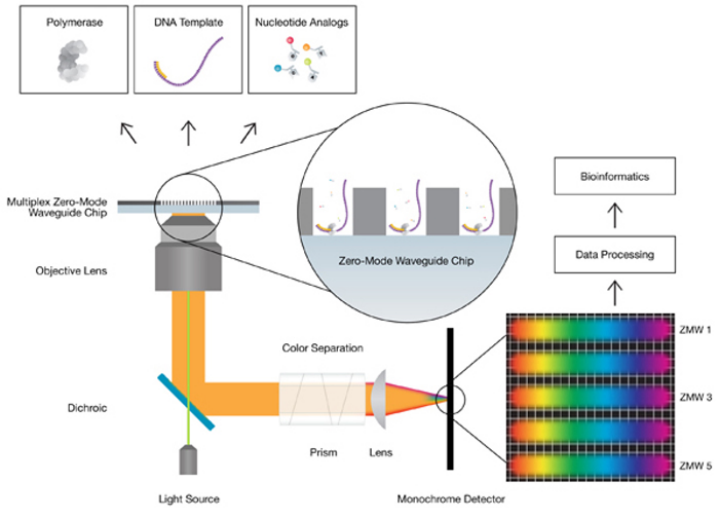


PacBio sequencing

- Single Molecule Real Time (SMRT)
- Circularizes, reads single molecule
  - High (~10%) error
  - Uses consensus from multiple reads
- Gets *incredibly* long reads, starting 2013-Oct-03
  - Mean length 8,500 bases
  - N50 of 10k bases
  - Max reads of 30k bases

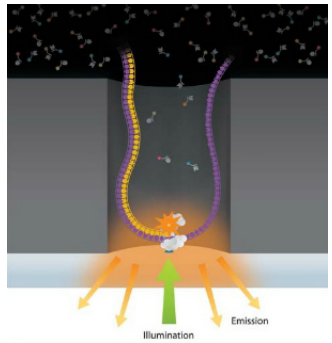


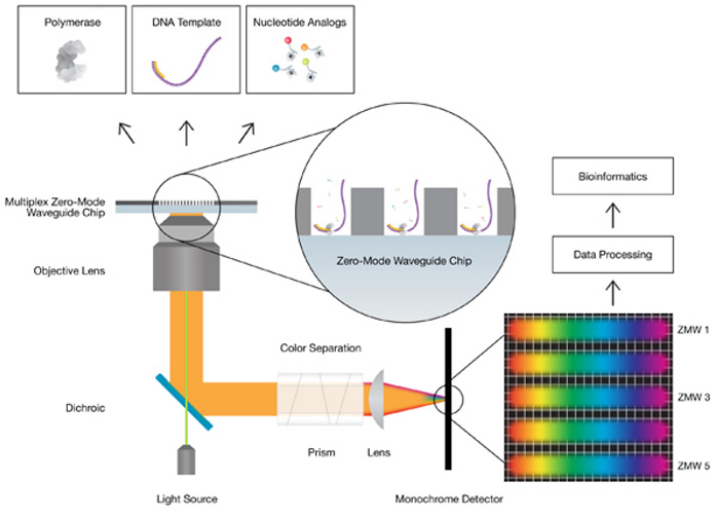




PacBio sequencing

- Single Molecule Real Time (SMRT)
- Circularizes, reads single molecule
  - High (~10%) error
  - Uses consensus from multiple reads
- Gets *incredibly* long reads, starting 2013-Oct-03
  - Mean length 8,500 bases
  - N50 of 10k bases
  - Max reads of 30k bases
- Fewer (50k) reads





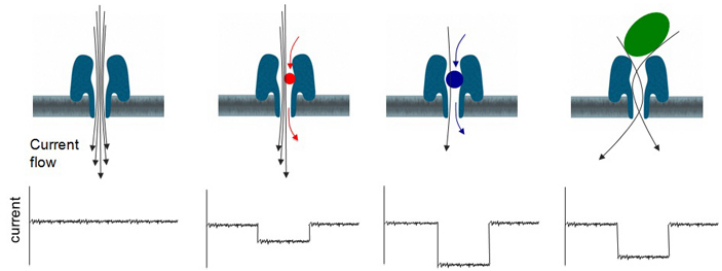
Nanopore Sequencing

Nanopore Sequencing

- Create a *tiny* pore with electrical potential

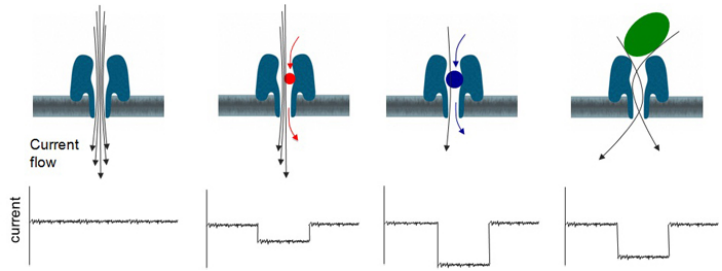
Nanopore Sequencing

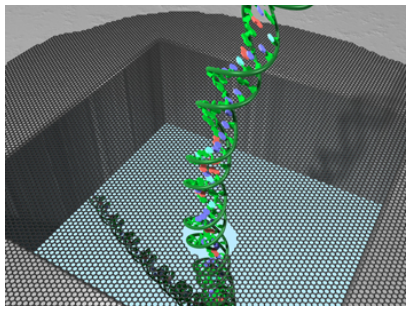
- Create a *tiny* pore with electrical potential
- Current changes as it is blocked



Nanopore Sequencing

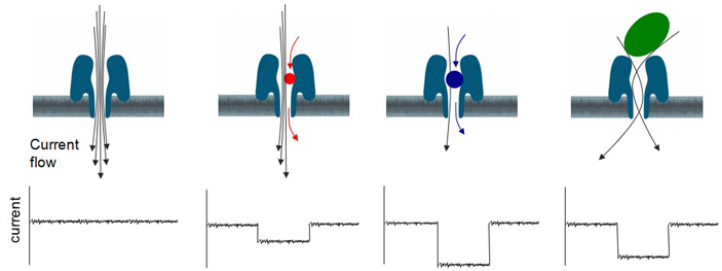
- Create a *tiny* pore with electrical potential
- Current changes as it is blocked
- Pass DNA through

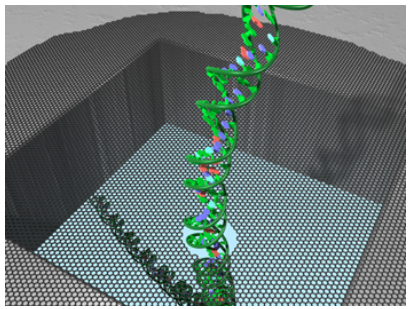


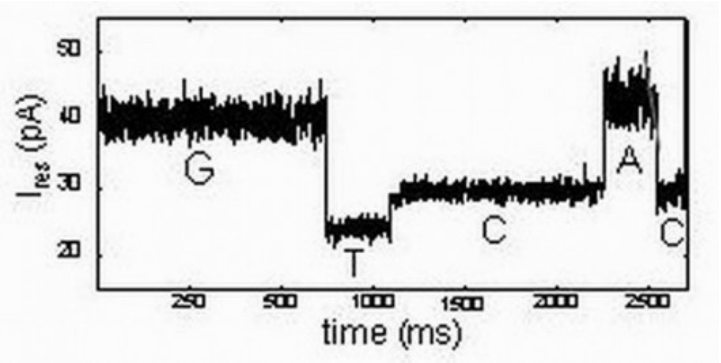


Nanopore Sequencing

- Create a *tiny* pore with electrical potential
- Current changes as it is blocked
- Pass DNA through
- Read current changes

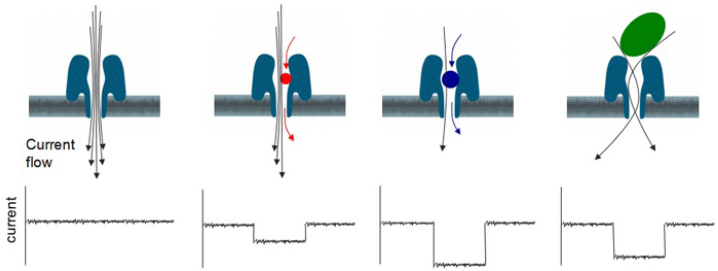


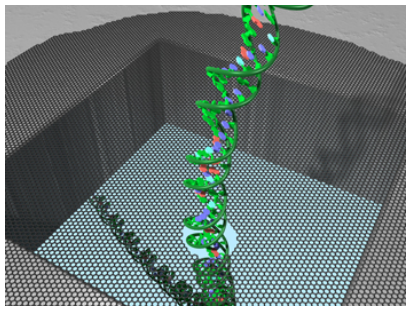


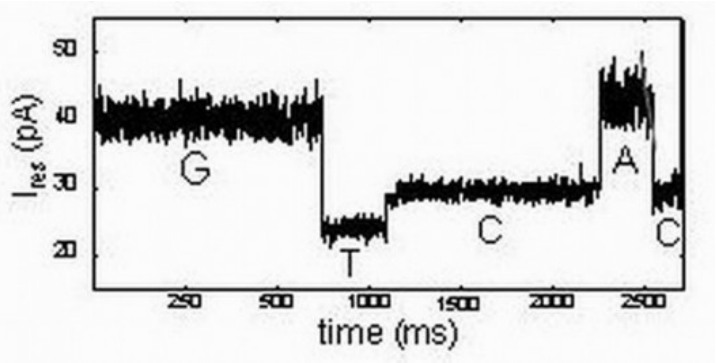


Nanopore Sequencing

- Create a *tiny* pore with electrical potential
- Current changes as it is blocked
- Pass DNA through
- Read current changes
- Arbitrary length reads, but many challenges:
  - DNA moves too fast
  - How are homopolymers addressed?
  - Making good pores







Handling the data

Handling the data

- Data files are *huge*

Handling the data

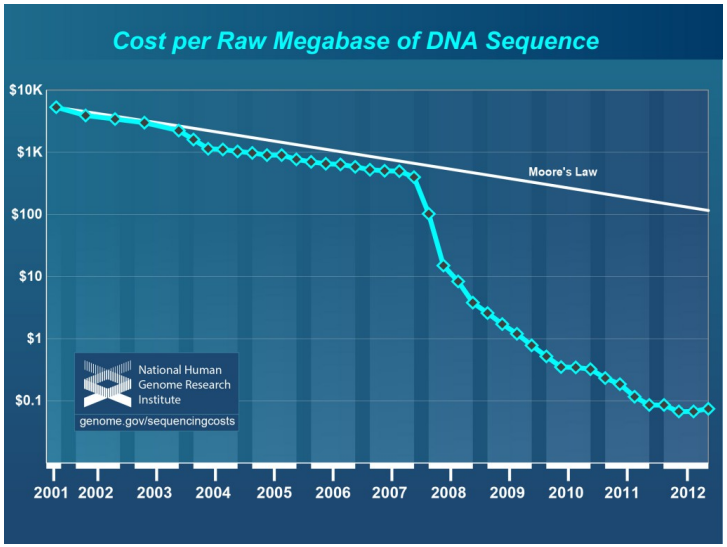
- Data files are *huge*
  - Need *nix computers

Handling the data

- Data files are *huge*
  - Need *nix computers
- Moving them is non-trivial

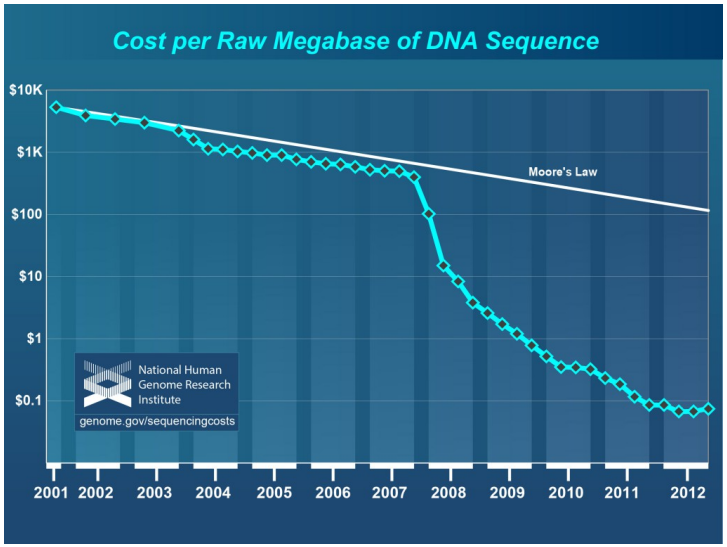
Handling the data

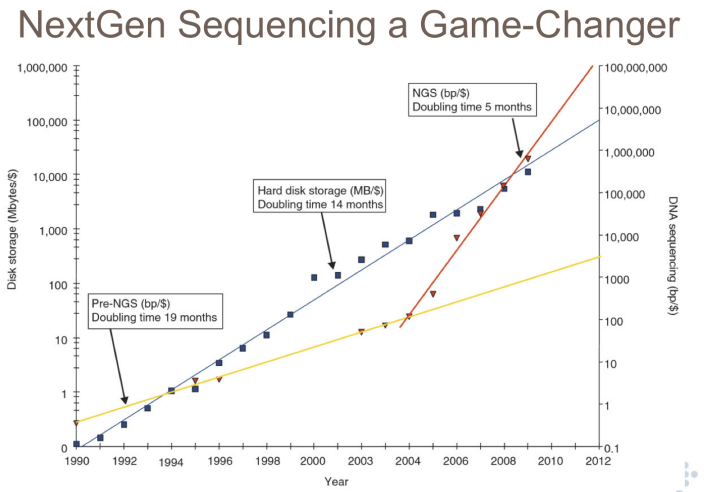
- Data files are *huge*
  - Need *nix computers
- Moving them is non-trivial
- Moore's law exceeded



Handling the data

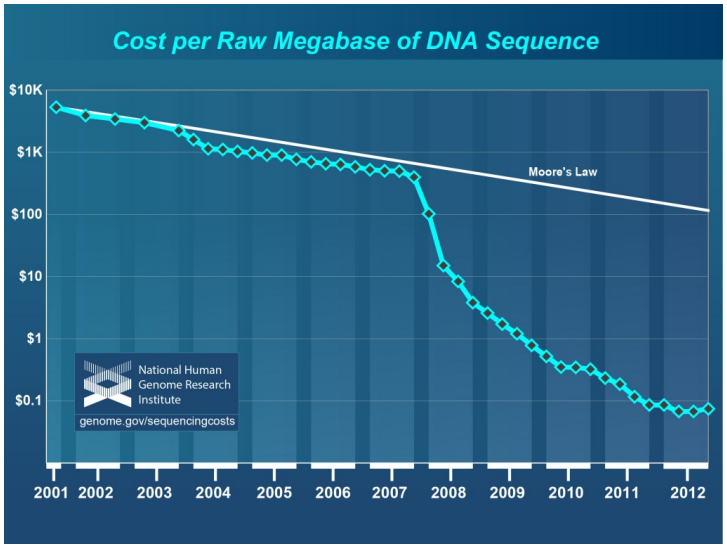
- Data files are *huge*
  - Need *nix computers
- Moving them is non-trivial
- Moore's law exceeded
  - Data storage issues

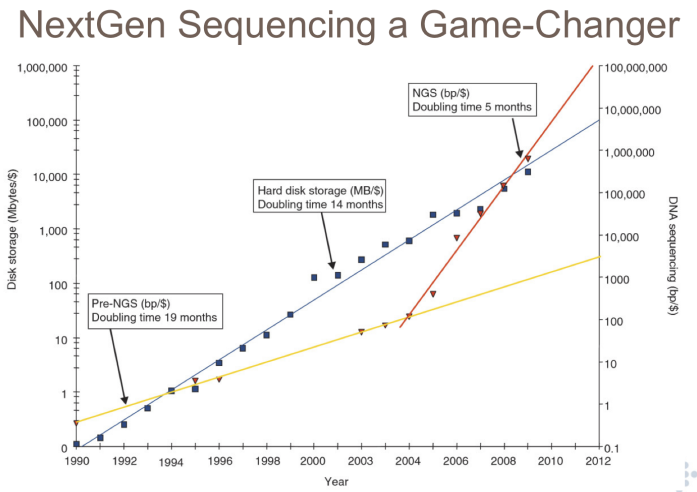




Handling the data

- Data files are *huge*
  - Need *nix computers
- Moving them is non-trivial
- Moore's law exceeded
  - Data storage issues
- Most Data files are shared
  - GEO; SRA; GenBank; EMBL; etc.





Splitting and trimming

Splitting and trimming

- Need to separate multi-plexed samples
  - TCACTTCGTA...
  - TCATCTCGTA...

- `split_libraries.py`
  `-i seq.fastq`
  `-o out`
  `-b s2.fastq`
  `-m map.txt`

Splitting and trimming

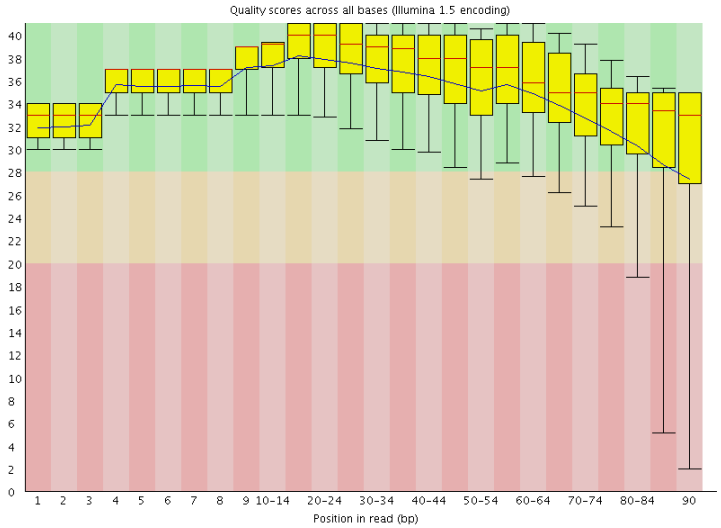
- Need to separate multi-plexed samples
  - TCACTTCGTA...
  - TCATCTCGTA...
- Remove errors in barcodes

- `split_libraries.py`
  `-i seq.fastq`
  `-o out`
  `-b s2.fastq`
  `-m map.txt`

Splitting and trimming

- Need to separate multi-plexed samples
  - TCACTTCGTA...
  - TCATCTCGTA...
- Remove errors in barcodes
- Quality filter and trim
  - Decisions depend on the application

- `split_libraries.py`
  `-i seq.fastq`
  `-o out`
  `-b s2.fastq`
  `-m map.txt`



Aligning data to a reference

Aligning data to a reference

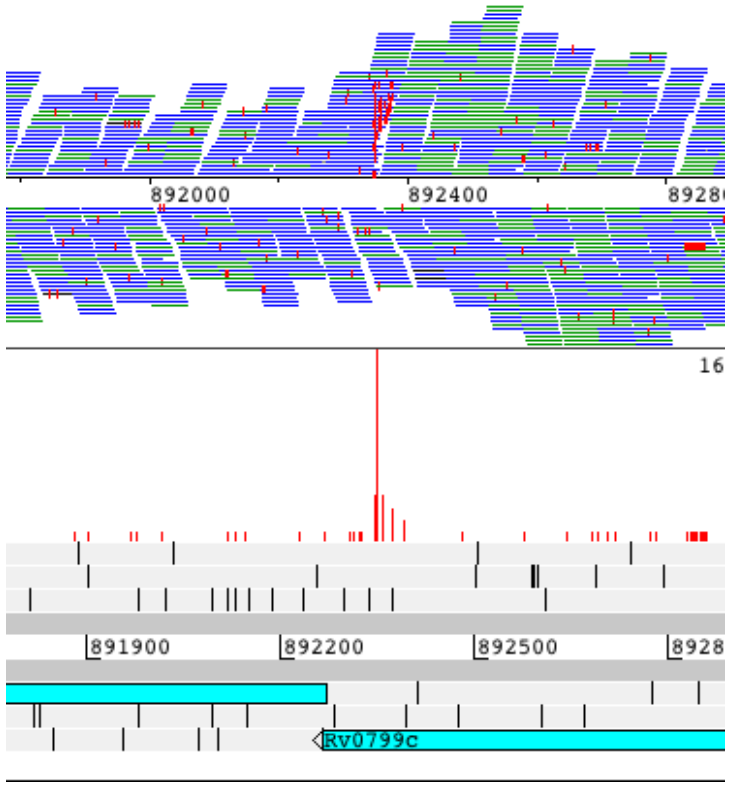
- Now have lots of short sequence fragments

Aligning data to a reference

- Now have lots of short sequence fragments
- Need to identify the source of each

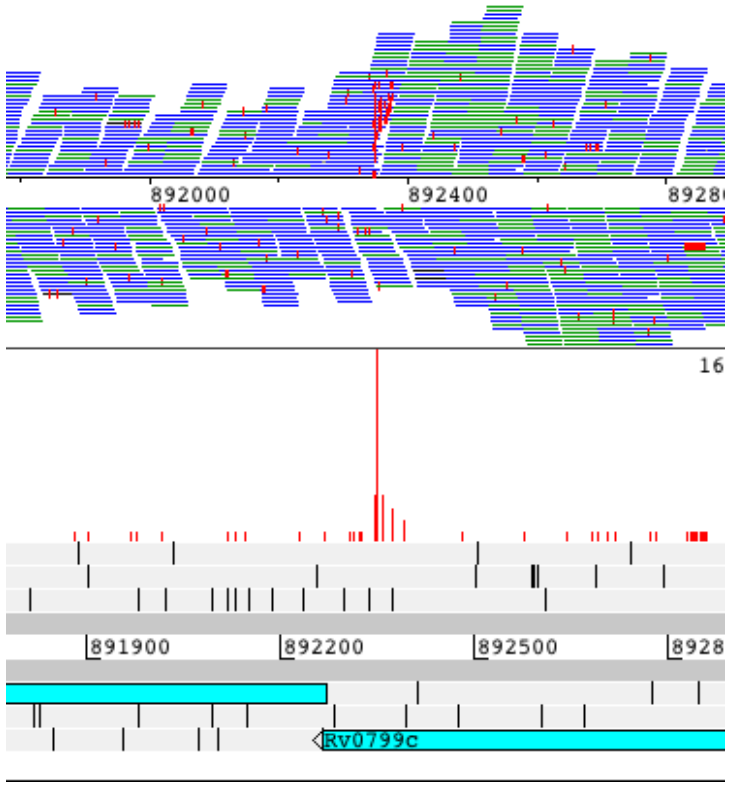
Aligning data to a reference

- Now have lots of short sequence fragments
- Need to identify the source of each
- Align against a reference



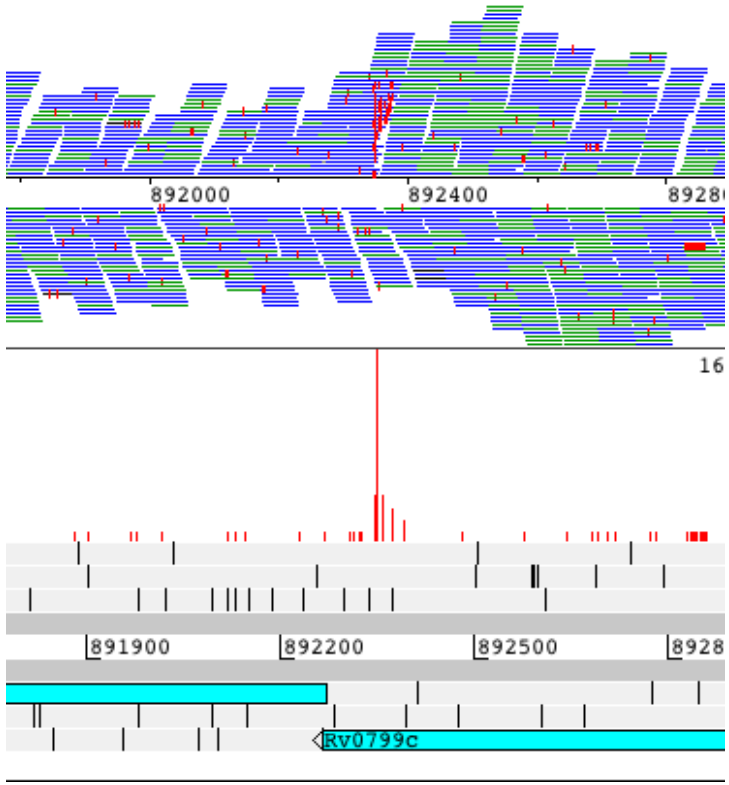
Aligning data to a reference

- Now have lots of short sequence fragments
- Need to identify the source of each
- Align against a reference
  - May need to be created



Aligning data to a reference

- Now have lots of short sequence fragments
- Need to identify the source of each
- Align against a reference
  - May need to be created
- Over 50 possible programs



Assign and normalize scores

Assign and normalize scores

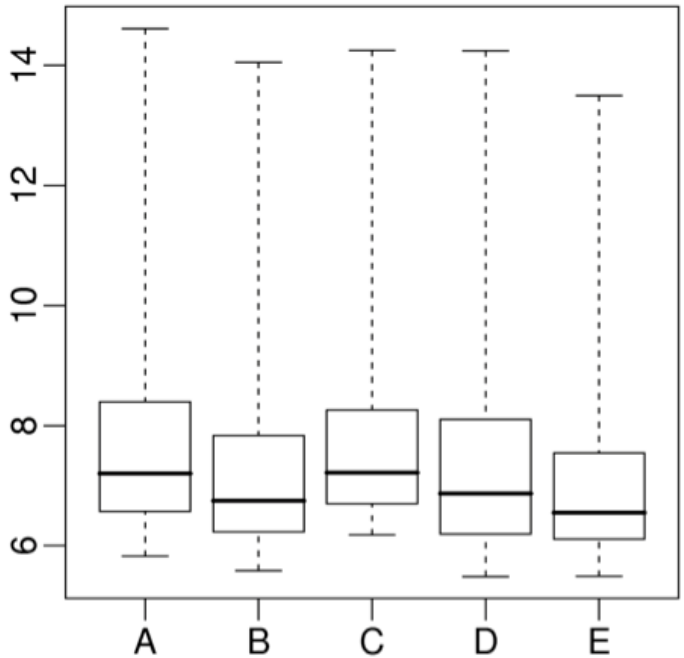
- Count the number of fragments that align to each sequence

Assign and normalize scores

- Count the number of fragments that align to each sequence
- But, how to normalize?

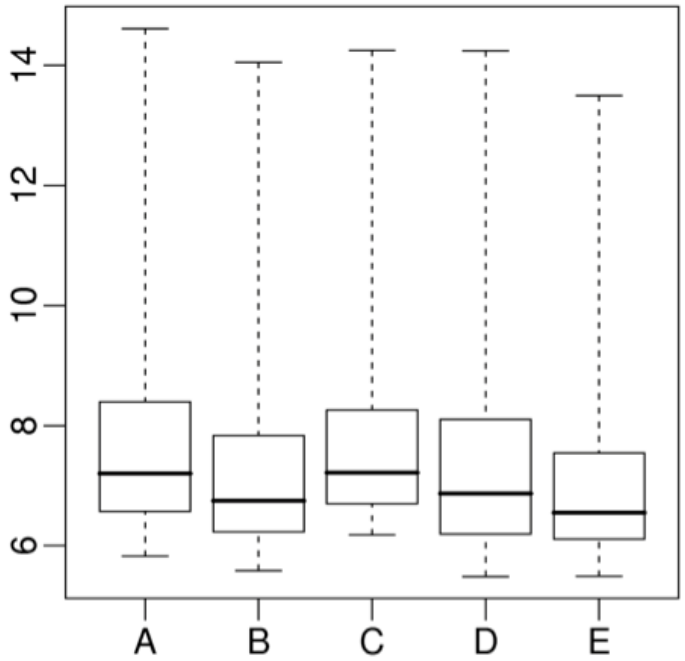
Assign and normalize scores

- Count the number of fragments that align to each sequence
- But, how to normalize?
  - Total reads can vary widely



Assign and normalize scores

- Count the number of fragments that align to each sequence
- But, how to normalize?
  - Total reads can vary widely
- Fragments per kilobase per million fragments - FPKM
  - Controls for total sequencing and gene length



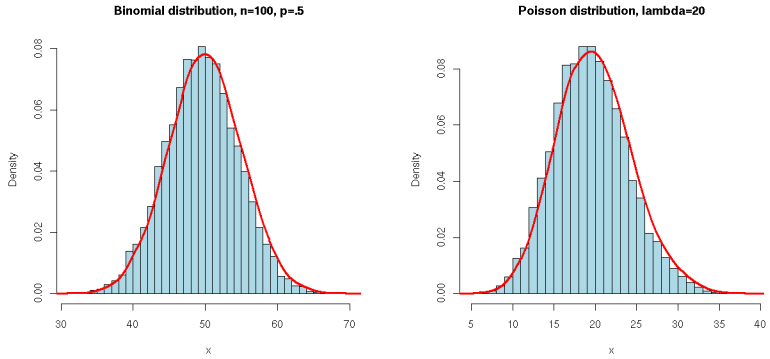
Basics of differential expression

Basics of differential expression

- Many of the goals are to identify differences in expression

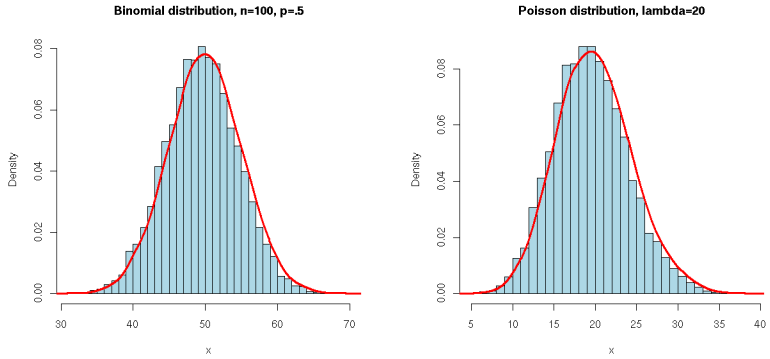
Basics of differential expression

- Many of the goals are to identify differences in expression
- Several different statistical models



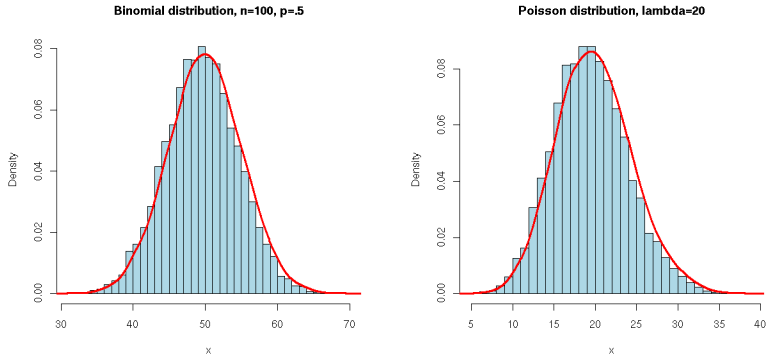
Basics of differential expression

- Many of the goals are to identify differences in expression
- Several different statistical models
- Over 35 different tools



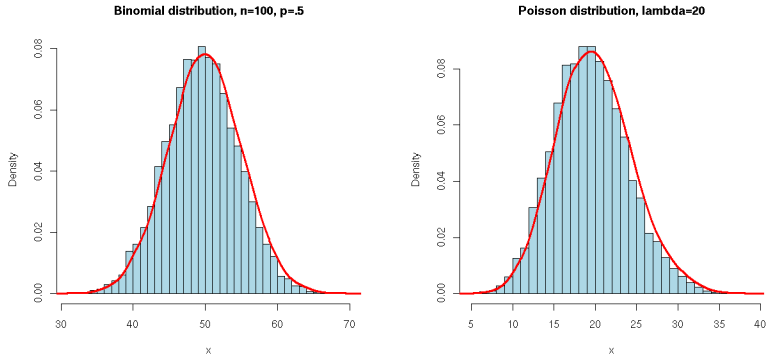
Basics of differential expression

- Many of the goals are to identify differences in expression
- Several different statistical models
- Over 35 different tools
  - Most written for R or Linux command line



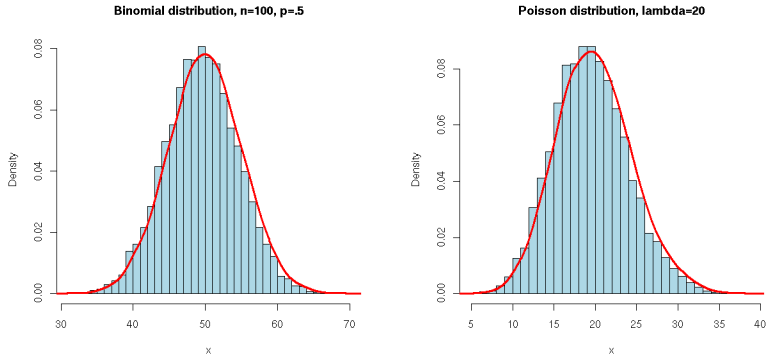
Basics of differential expression

- Many of the goals are to identify differences in expression
- Several different statistical models
- Over 35 different tools
  - Most written for R or Linux command line
- Sample size is key



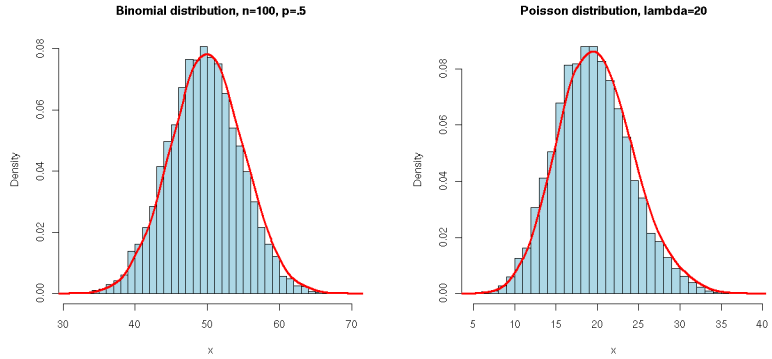
Basics of differential expression

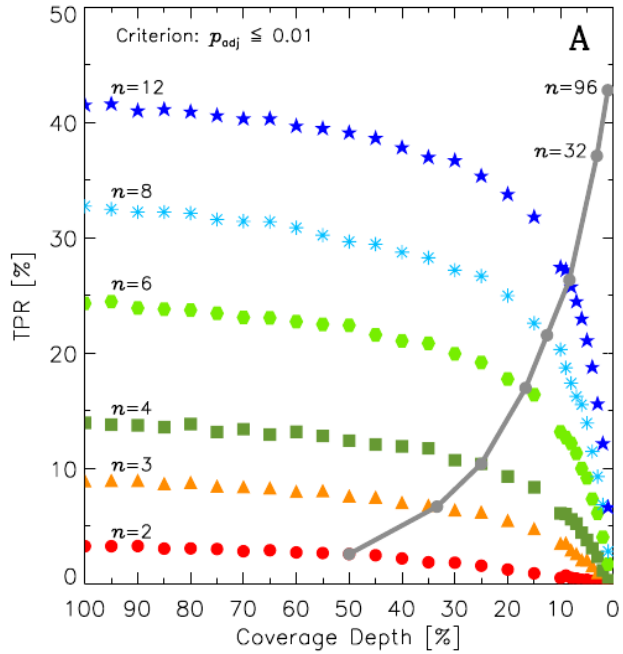
- Many of the goals are to identify differences in expression
- Several different statistical models
- Over 35 different tools
  - Most written for R or Linux command line
- Sample size is key
  - Sequence Coverage



Basics of differential expression

- Many of the goals are to identify differences in expression
- Several different statistical models
- Over 35 different tools
  - Most written for R or Linux command line
- Sample size is key
  - Sequence Coverage
  - Biological replicates

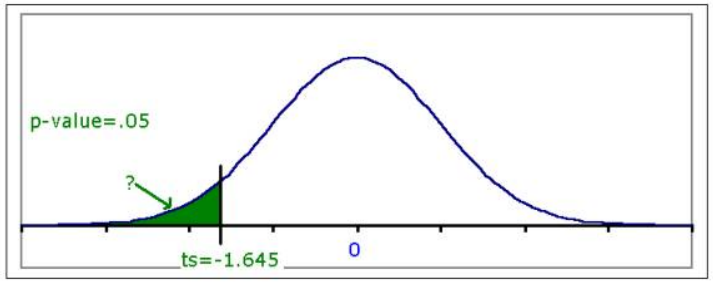




Multiple Testing

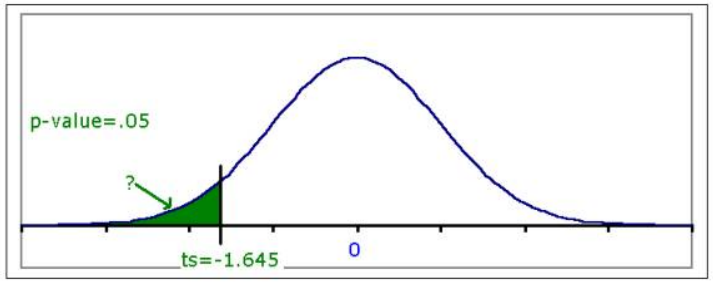
Multiple Testing

- Test 20 genes, how many $p < 0.05$?



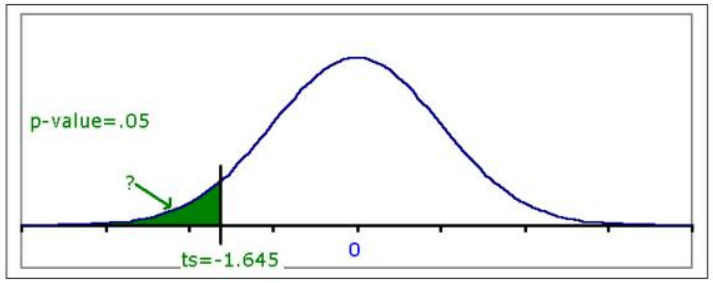
Multiple Testing

- Test 20 genes, how many $p < 0.05$?
  - What about 20k genes?



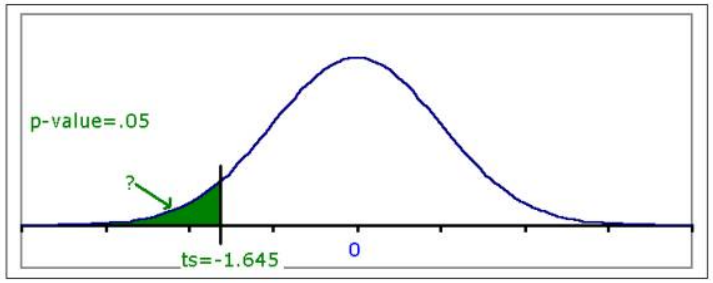
Multiple Testing

- Test 20 genes, how many $p < 0.05$?
  - What about 20k genes?
- Bonferroni was an ass



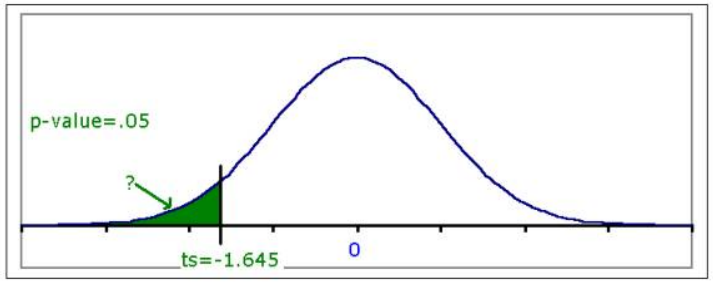
Multiple Testing

- Test 20 genes, how many $p < 0.05$?
  - What about 20k genes?
- Bonferroni was an ass
  - $p < 2.5 * 10^{-6}$



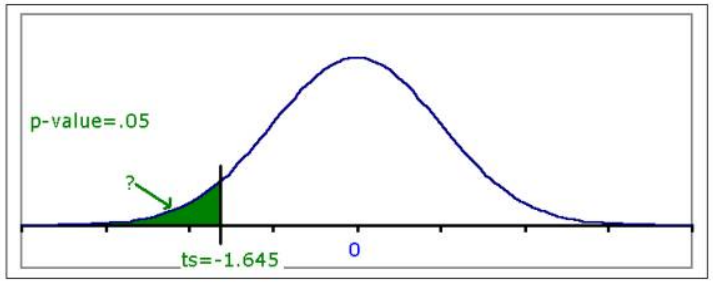
Multiple Testing

- Test 20 genes, how many $p < 0.05$?
  - What about 20k genes?
- Bonferroni was an ass
  - $p < 2.5 * 10^{-6}$
  - But he wasn't wrong



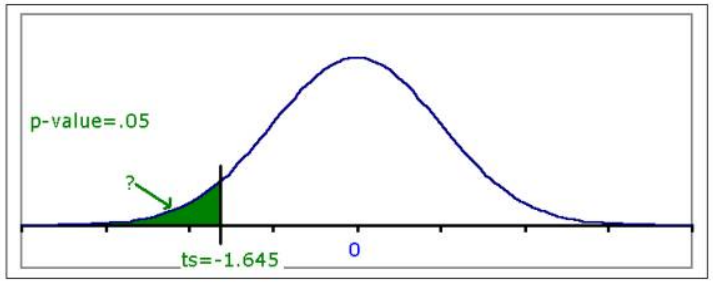
Multiple Testing

- Test 20 genes, how many $p < 0.05$?
  - What about 20k genes?
- Bonferroni was an ass
  - $p < 2.5 * 10^{-6}$
  - But he wasn't wrong
- Less restrictive alternative
  - False Discovery Rate



Multiple Testing

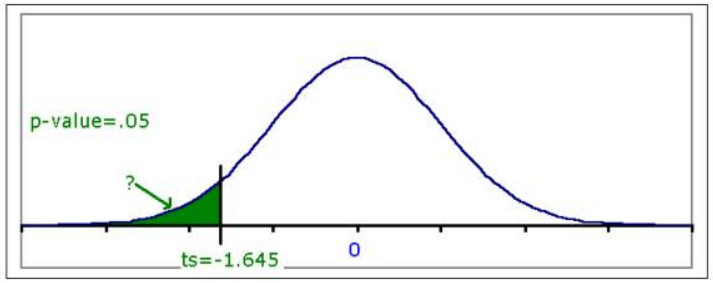
- Test 20 genes, how many $p < 0.05$?
  - What about 20k genes?
- Bonferroni was an ass
  - $p < 2.5 * 10^{-6}$
  - But he wasn't wrong
- Less restrictive alternative
  - False Discovery Rate



	H_0 is True	H_1 is True	Total
Call sig	V	S	R
Call non-sig	U	T	m - R
Total	m_0	m - m_0	m

Multiple Testing

- Test 20 genes, how many $p < 0.05$?
  - What about 20k genes?
- Bonferroni was an ass
  - $p < 2.5 * 10^{-6}$
  - But he wasn't wrong
- Less restrictive alternative
  - False Discovery Rate
  - Bayes



	H_0 is True	H_1 is True	Total
Call sig	V	S	R
Call non-sig	U	T	m - R
Total	m_0	m - m_0	m

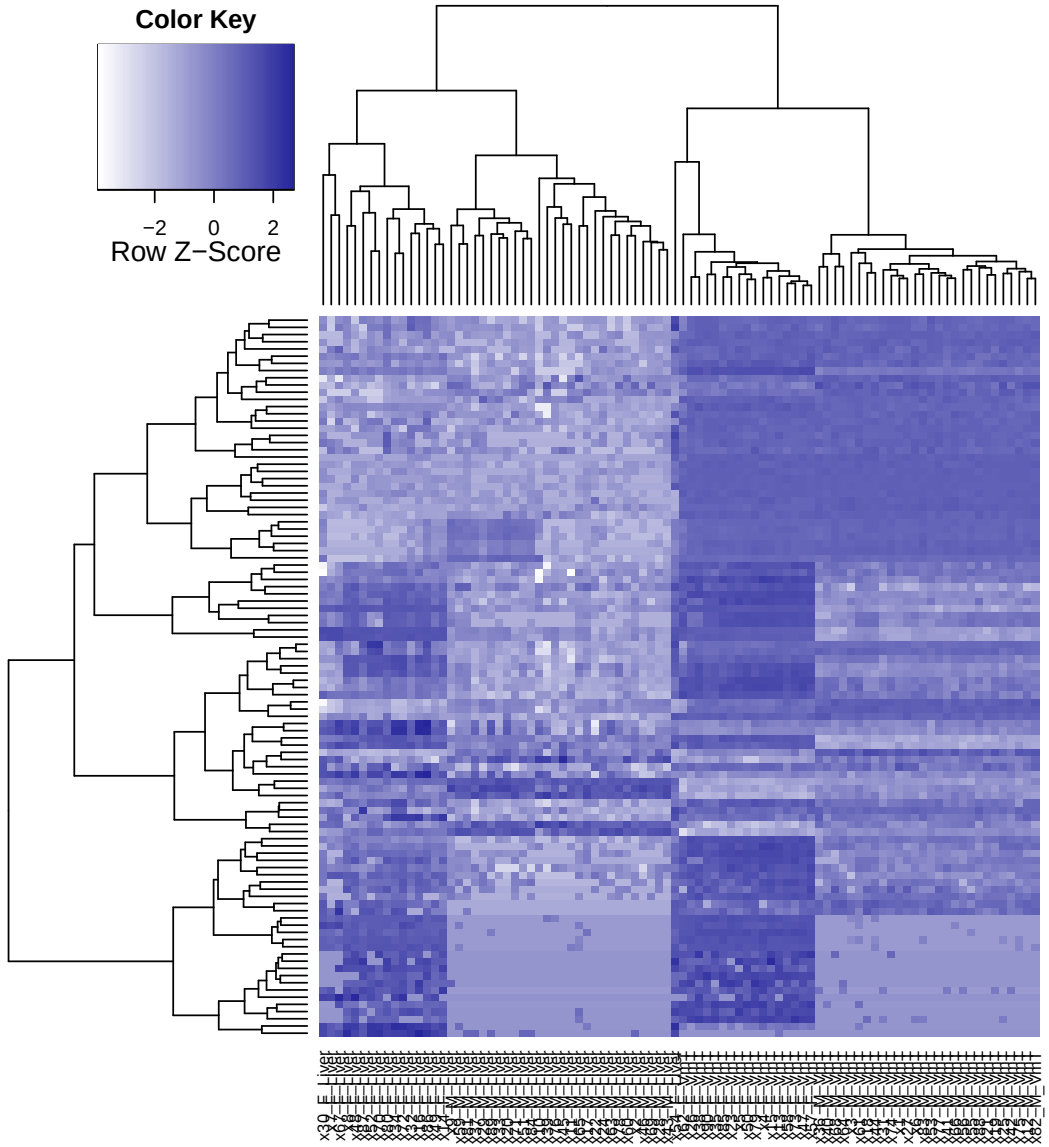
A big list of genes

A big list of genes

657 Genes: isogroup19373; isogroup19515; isogroup13068; isogroup10667; isogroup11511; isogroup07184; isogroup17762; isogroup00499; isogroup02469; isogroup22206; isogroup16440; isogroup05470; isogroup22638; isogroup16100; isogroup19892; isogroup07895; isogroup03190; isogroup21725; isogroup16366; isogroup03130; isogroup21894; isogroup08311; isogroup12192; isogroup01746; isogroup09440; isogroup18430; isogroup15653; isogroup06151; isogroup00660; isogroup14952; isogroup01183; isogroup22319; isogroup11507; isogroup06246; isogroup07270; isogroup06737; isogroup11168; isogroup08955; isogroup07610; isogroup06101; isogroup21172; isogroup09396; isogroup20679; isogroup02199; isogroup00176; isogroup00147; isogroup13098; isogroup03280; isogroup12307; isogroup04792; isogroup19416; isogroup13712; isogroup10572; isogroup02056; isogroup21399; isogroup13499; isogroup04474; isogroup20441; isogroup04909; isogroup00097; isogroup11302; isogroup07447; isogroup12033; isogroup13464; isogroup20646; isogroup13511; isogroup15410; isogroup21533; isogroup00006; isogroup02618; isogroup11706; isogroup20691; isogroup03917; isogroup03398; isogroup02863; isogroup14748; isogroup08046; isogroup06354; isogroup01902; isogroup05820; isogroup02653; isogroup14064; isogroup21159; isogroup12317; isogroup16504; isogroup00260; isogroup04114; isogroup15569; isogroup16200; isogroup11299; isogroup16326; isogroup19367; isogroup21611; isogroup06580; isogroup14905; isogroup03379; isogroup03149; isogroup00907; isogroup17296; isogroup20211; isogroup11774;

Visualization

Visualization



Sex chromosome analysis sanity check

Sex chromosome analysis sanity check

- Diverged sex chromosomes are common

Sex chromosome analysis sanity check

- Diverged sex chromosomes are common
- However; they are often under-annotated

Sex chromosome analysis sanity check

- Diverged sex chromosomes are common
- However; they are often under-annotated
  - Many bird genomes lack the W chromosome

Sex chromosome analysis sanity check

- Diverged sex chromosomes are common
- However; they are often under-annotated
  - Many bird genomes lack the W chromosome
  - In fish, rapid evolution of sex determination

Sex chromosome analysis sanity check

- Diverged sex chromosomes are common
- However; they are often under-annotated
  - Many bird genomes lack the W chromosome
  - In fish, rapid evolution of sex determination
- I have lots of gene expression data in birds

Sex chromosome analysis sanity check

- Diverged sex chromosomes are common
- However; they are often under-annotated
  - Many bird genomes lack the W chromosome
  - In fish, rapid evolution of sex determination
- I have lots of gene expression data in birds
  - Commonality may reveal Z and W genes

Sex chromosome analysis sanity check

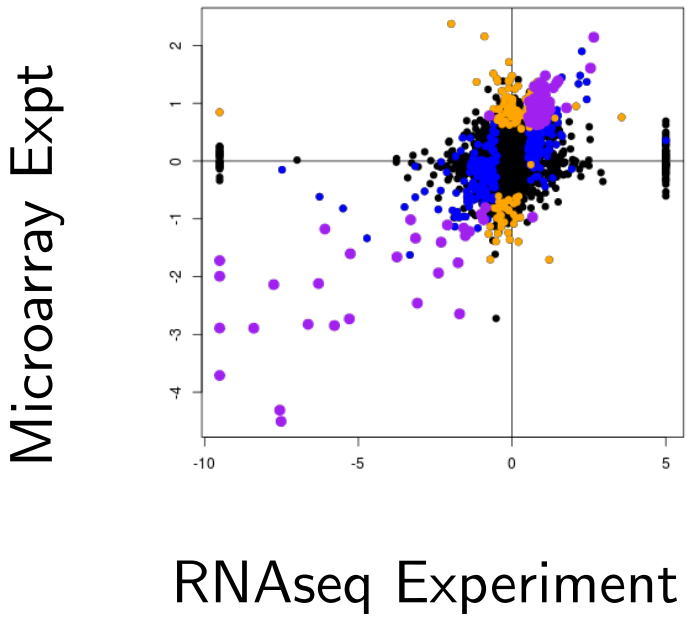
- Diverged sex chromosomes are common
- However; they are often under-annotated
  - Many bird genomes lack the W chromosome
  - In fish, rapid evolution of sex determination
- I have lots of gene expression data in birds
  - Commonality may reveal Z and W genes

Microarray Expt

RNAseq Experiment

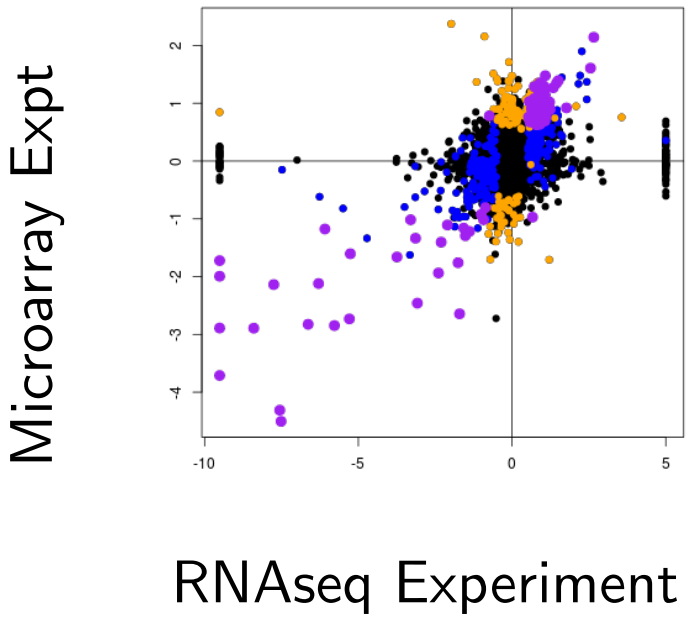
Sex chromosome analysis sanity check

- Diverged sex chromosomes are common
- However; they are often under-annotated
  - Many bird genomes lack the W chromosome
  - In fish, rapid evolution of sex determination
- I have lots of gene expression data in birds
  - Commonality may reveal Z and W genes



Sex chromosome analysis sanity check

- Diverged sex chromosomes are common
- However; they are often under-annotated
  - Many bird genomes lack the W chromosome
  - In fish, rapid evolution of sex determination
- I have lots of gene expression data in birds
  - Commonality may reveal Z and W genes
- Population genomic project may reveal fish sex chromosomes



GO analysis

GO analysis

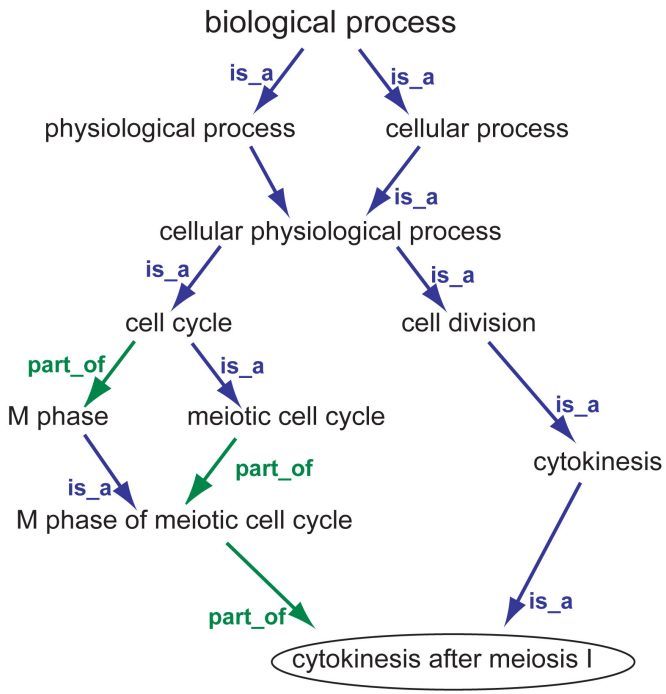
- Look for over-represented functions

GO analysis

- Look for over-represented functions
- Need heirarchical, conistent annotation

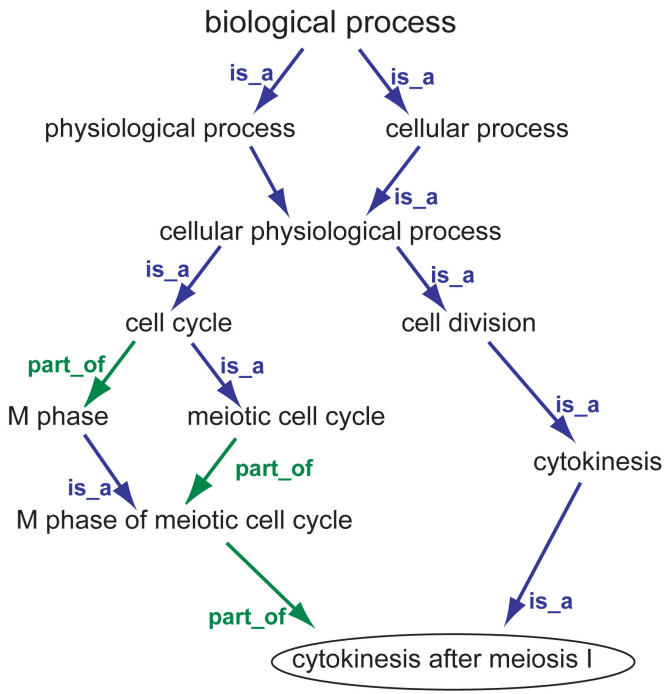
GO analysis

- Look for over-represented functions
- Need heirarchical, conistent annotation
  - Gene Ontology
  - 25k terms



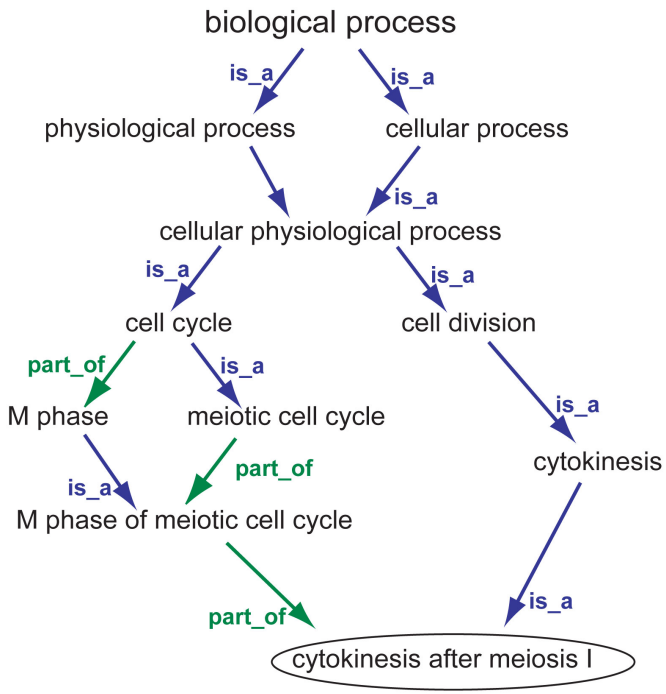
GO analysis

- Look for over-represented functions
- Need heirarchical, conistent annotation
  - Gene Ontology
  - 25k terms
- Use Fisher's Exact test



GO analysis

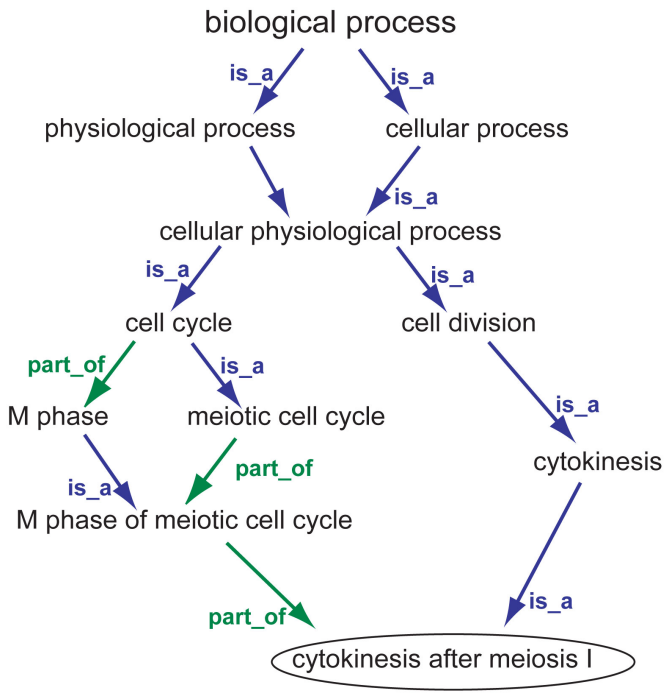
- Look for over-represented functions
- Need heirarchical, conistent annotation
  - Gene Ontology
  - 25k terms
- Use Fisher's Exact test



	is Sig DE	not Sig DE
Has GO term	A	C
Not this GO	B	D

GO analysis

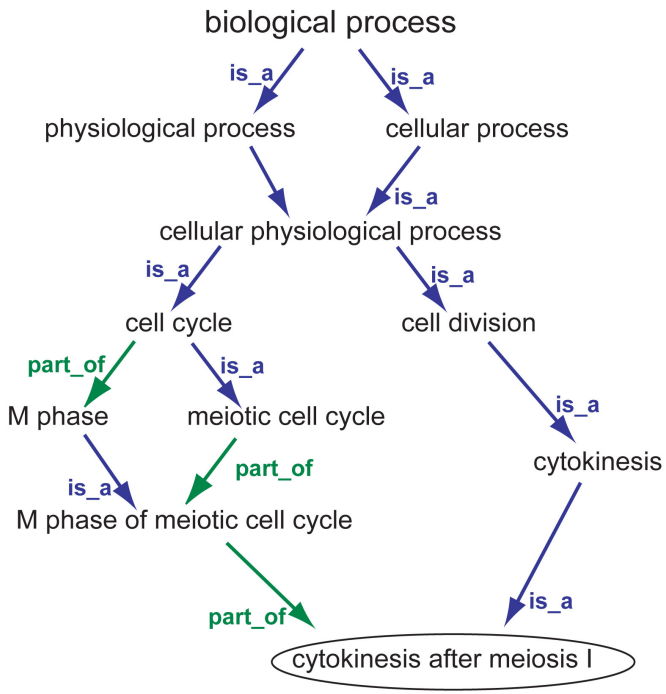
- Look for over-represented functions
- Need heirarchical, conistent annotation
  - Gene Ontology
  - 25k terms
- Use Fisher's Exact test
- Several alternatives



	is Sig DE	not Sig DE
Has GO term	A	C
Not this GO	B	D

GO analysis

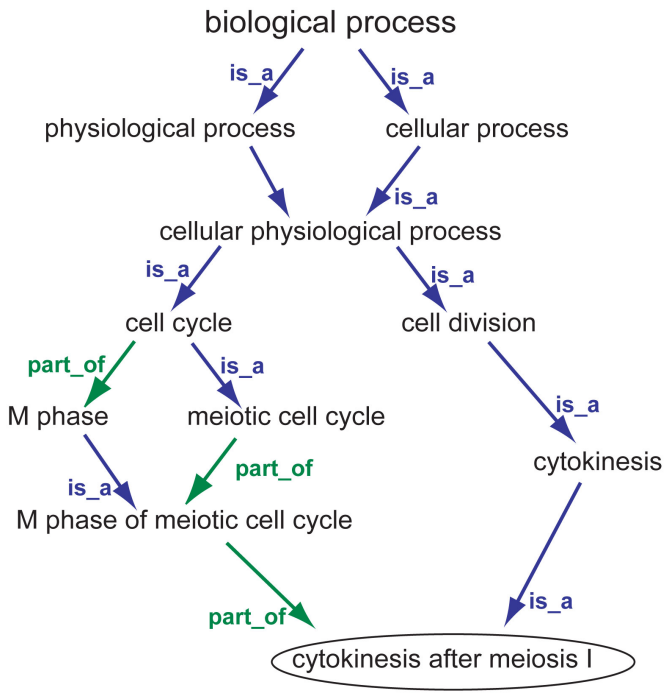
- Look for over-represented functions
- Need heirarchical, conistent annotation
  - Gene Ontology
  - 25k terms
- Use Fisher's Exact test
- Several alternatives
  - Chi-square



	is Sig DE	not Sig DE
Has GO term	A	C
Not this GO	B	D

GO analysis

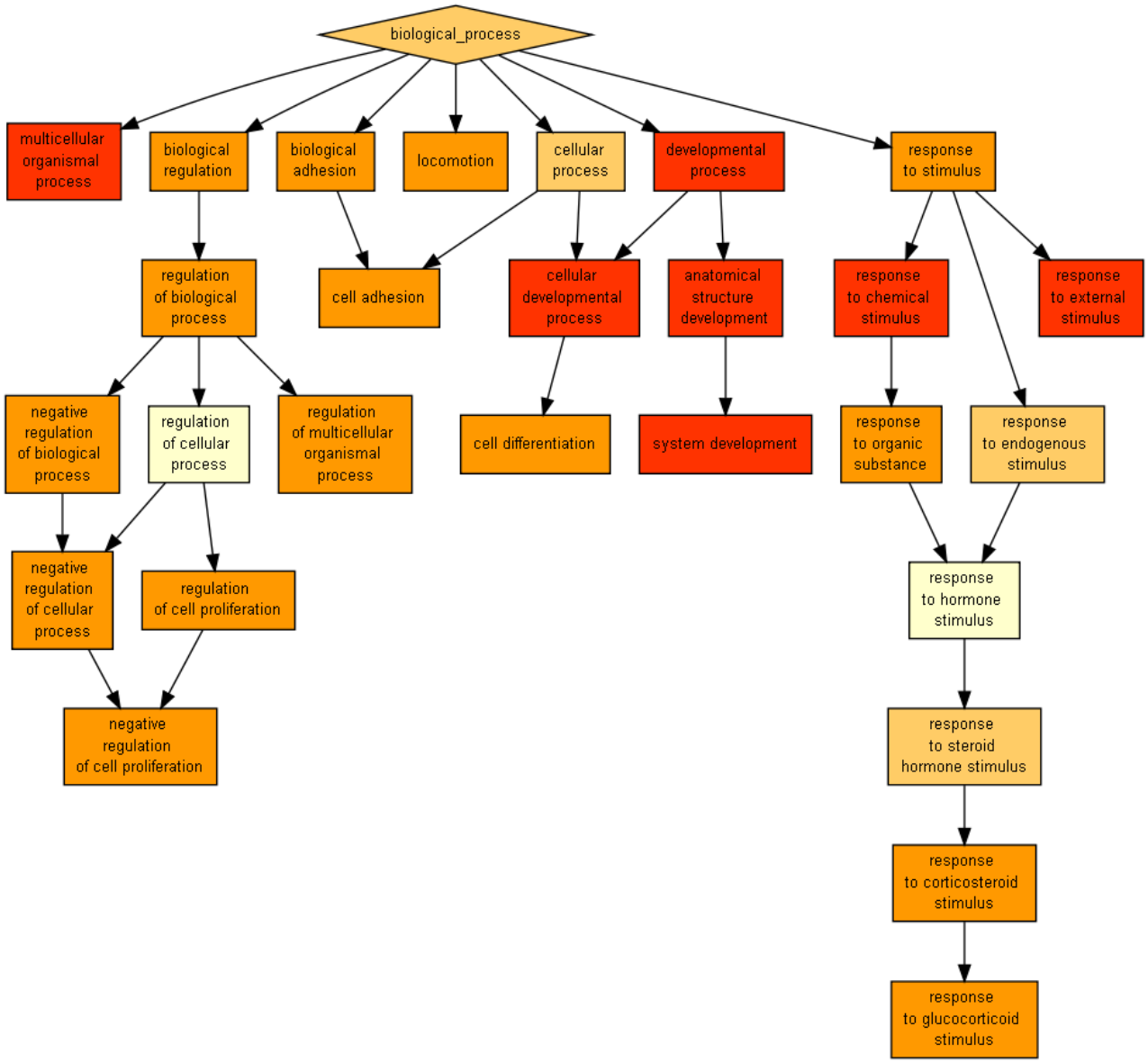
- Look for over-represented functions
- Need heirarchical, conistent annotation
  - Gene Ontology
  - 25k terms
- Use Fisher's Exact test
- Several alternatives
  - Chi-square
  - Use heirarchy



	is Sig DE	not Sig DE
Has GO term	A	C
Not this GO	B	D

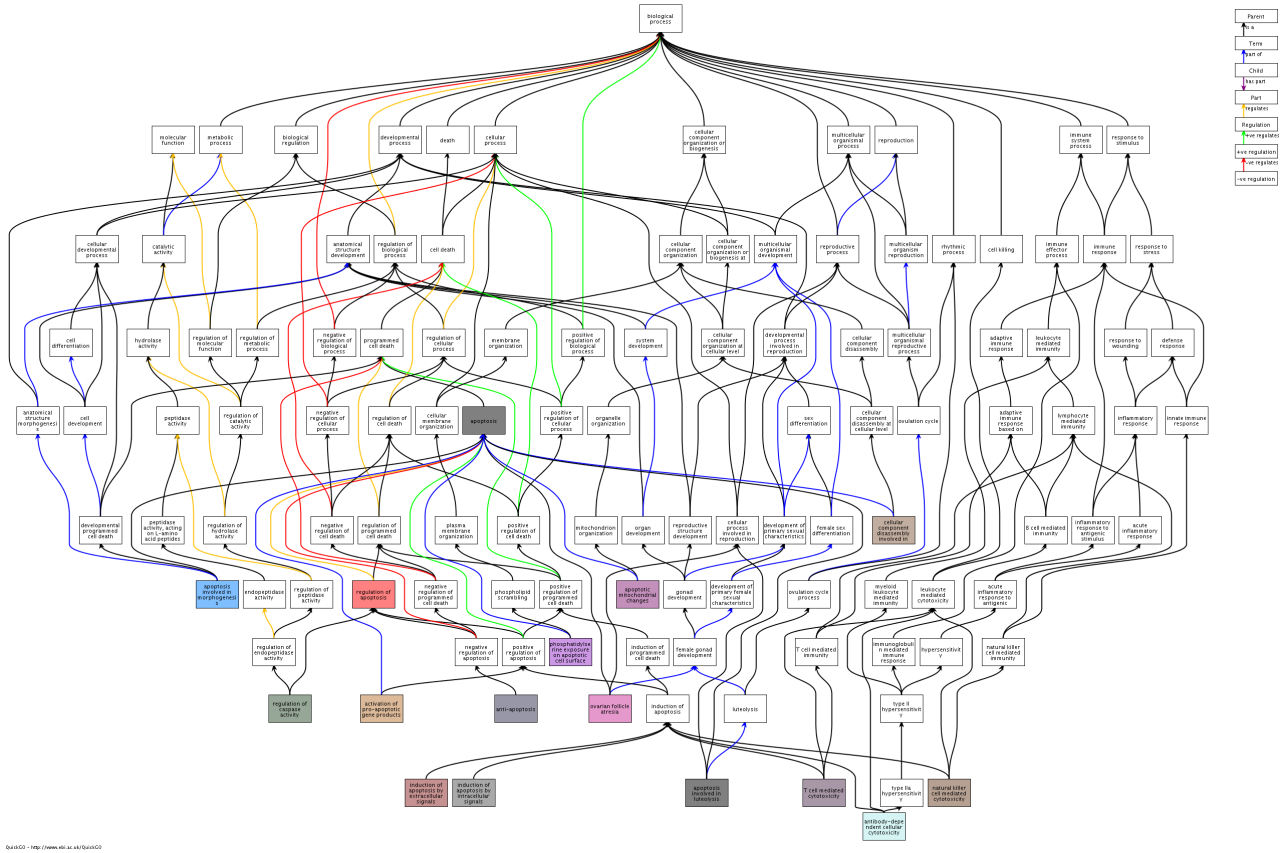
A simple example

A simple example



A bigger example

A bigger example



Facebook and Twitter analysis

Facebook and Twitter analysis

- Facebook networks are similar to gene networks

Facebook and Twitter analysis

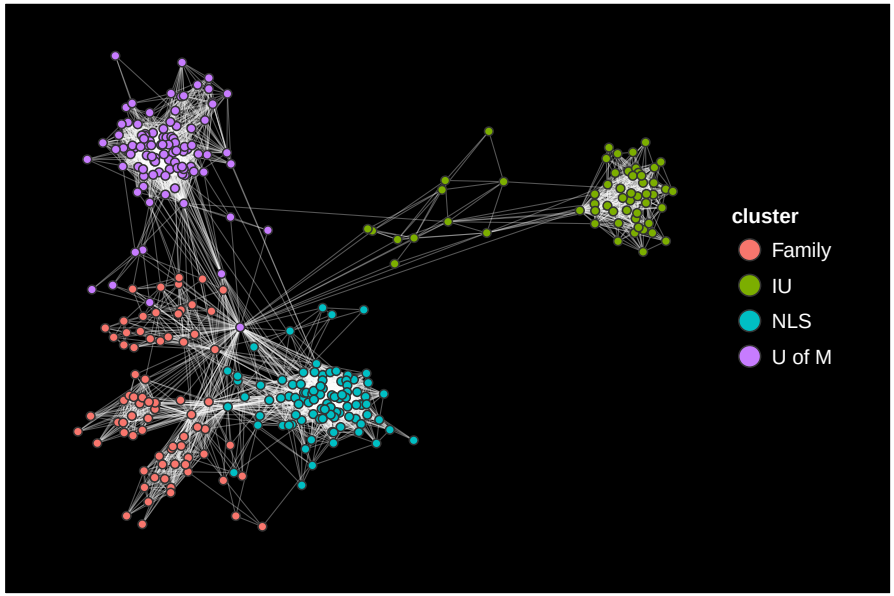
- Facebook networks are similar to gene networks
- Similar visualizations and analyses possible

Facebook and Twitter analysis

- Facebook networks are similar to gene networks
- Similar visualizations and analyses possible
  - My friends network

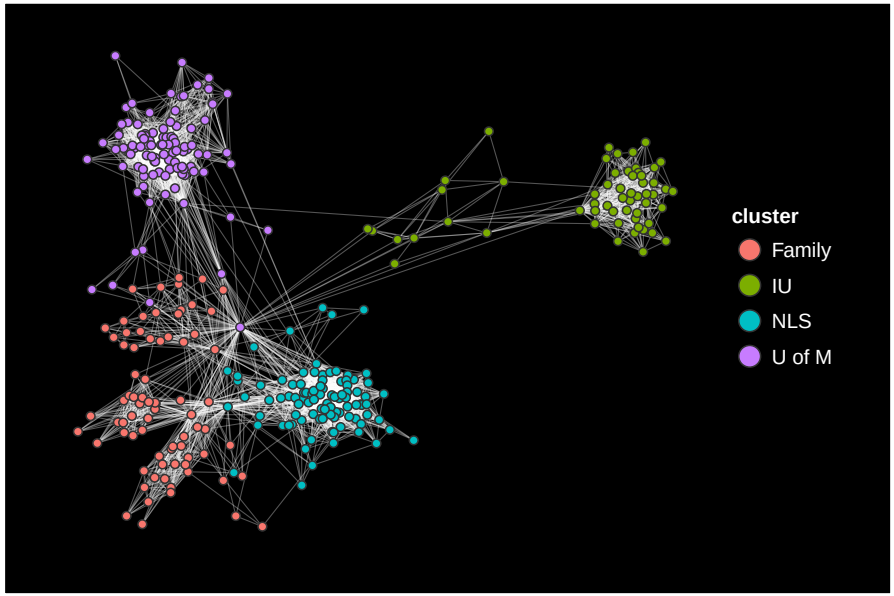
Facebook and Twitter analysis

- Facebook networks are similar to gene networks
- Similar visualizations and analyses possible
  - My friends network



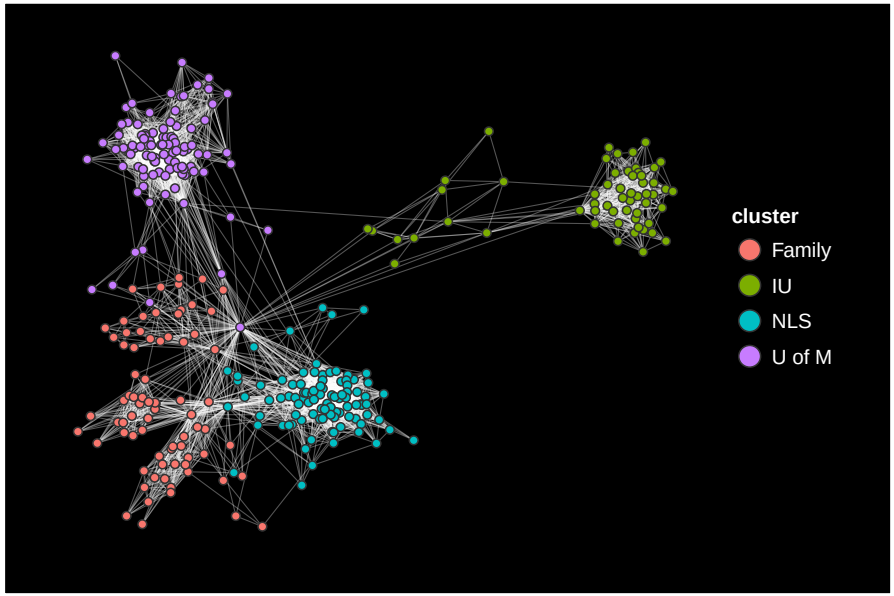
Facebook and Twitter analysis

- Facebook networks are similar to gene networks
- Similar visualizations and analyses possible
  - My friends network
- Both Facebook and Twitter offer ability to access data



Facebook and Twitter analysis

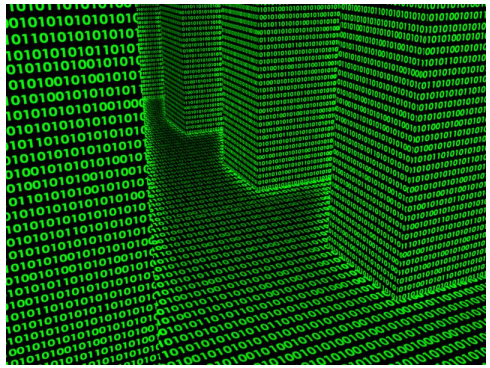
- Facebook networks are similar to gene networks
- Similar visualizations and analyses possible
  - My friends network
- Both Facebook and Twitter offer ability to access data
  - Planning text analysis



Conclusions

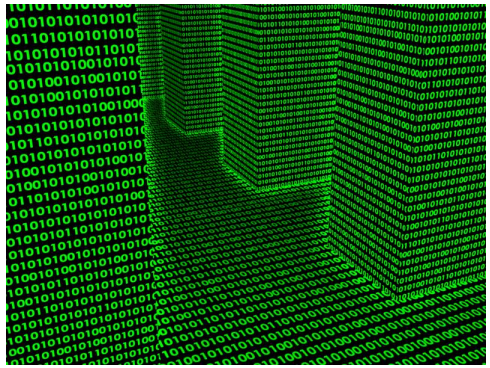
Conclusions

- Very little *Biology*



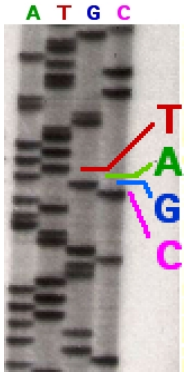
Conclusions

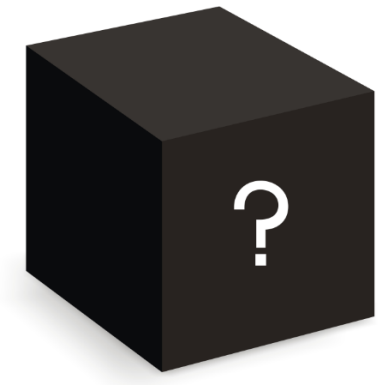
- Very little *Biology*
- Most inovation comes from other fields

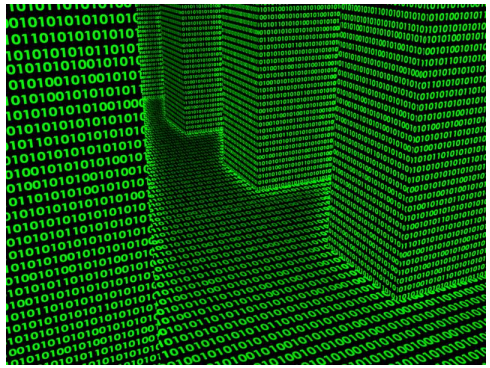


Conclusions

- Very little *Biology*
- Most inovation comes from other fields

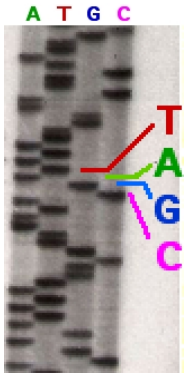


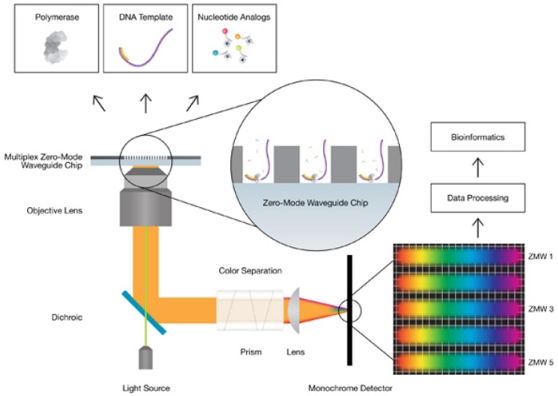


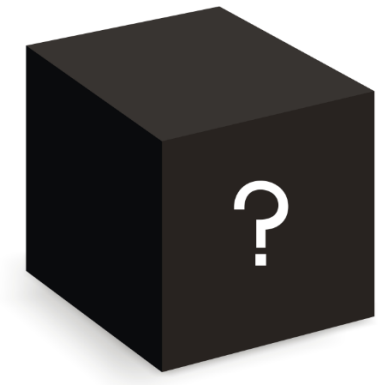


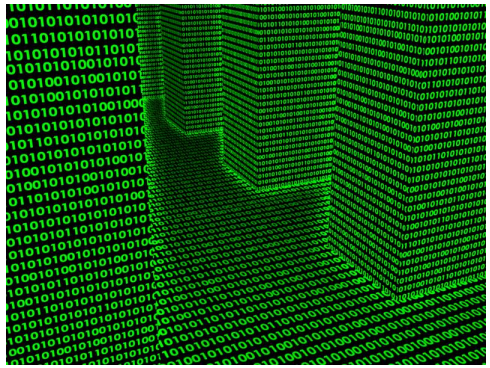
Conclusions

- Very little *Biology*
- Most inovation comes from other fields



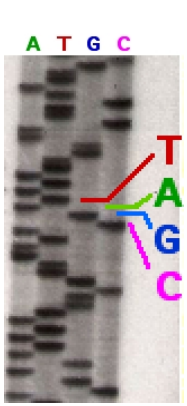


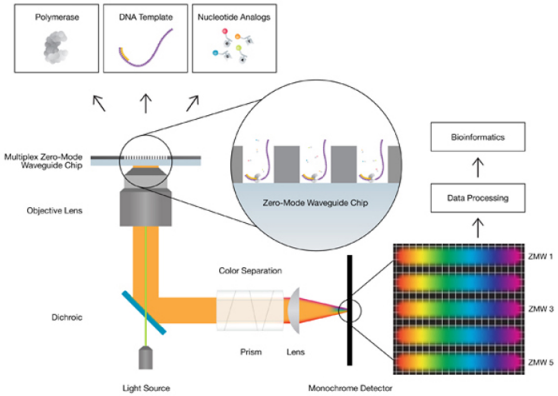


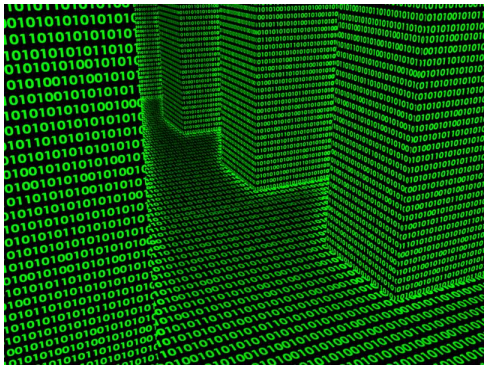


Conclusions

- Very little *Biology*
- Most inovation comes from other fields
- A *lot* of open-source tools

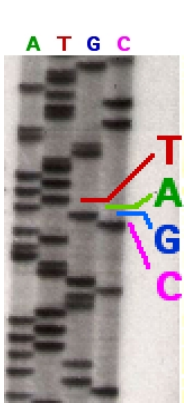


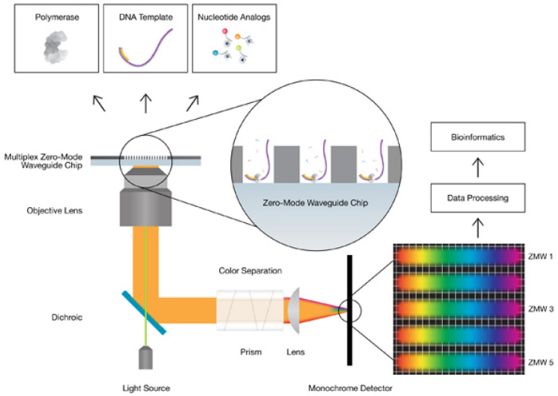


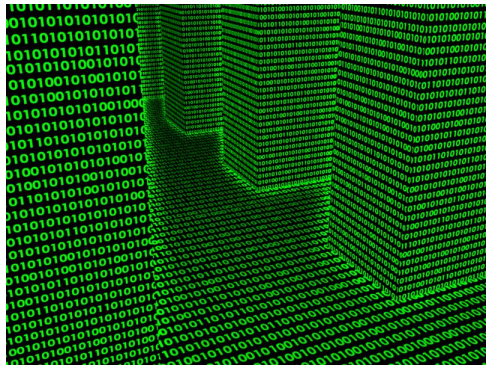


Conclusions

- Very little *Biology*
- Most inovation comes from other fields
- A *lot* of open-source tools
- Data is the bottle neck







Conclusions

- Very little *Biology*
- Most inovation comes from other fields
- A *lot* of open-source tools
- Data is the bottle neck
- Genomics is *integrative*