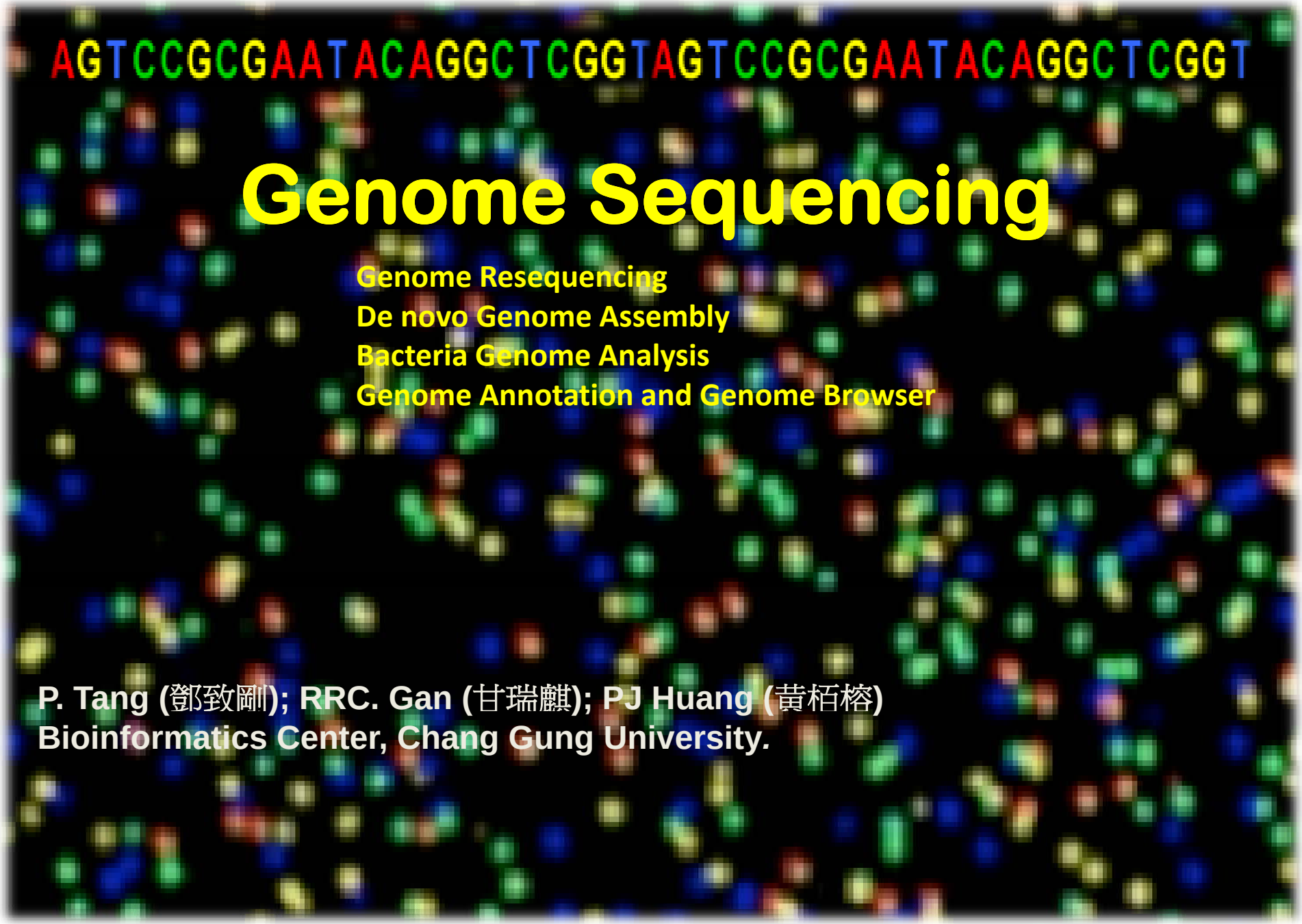
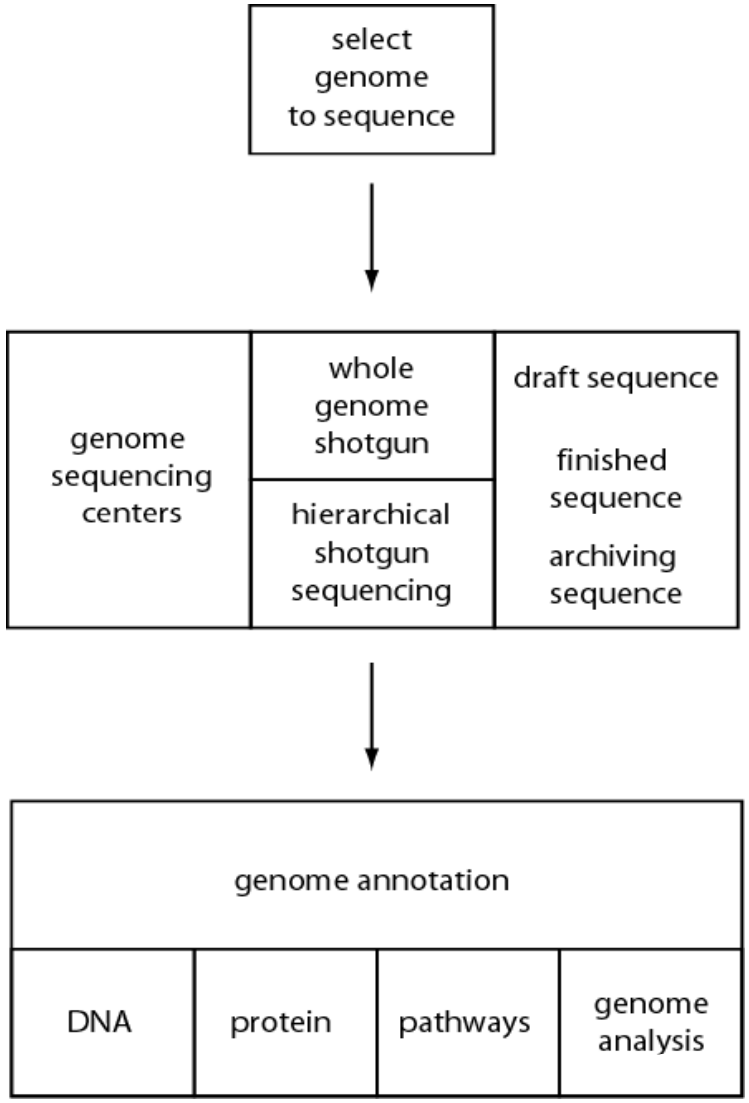



# Genome Sequencing

Genome Resequencing
De novo Genome Assembly
Bacteria Genome Analysis
Genome Annotation and Genome Browser

P. Tang (鄧致剛); RRC. Gan (甘瑞麒); PJ Huang (黃栢榕)
Bioinformatics Center, Chang Gung University.

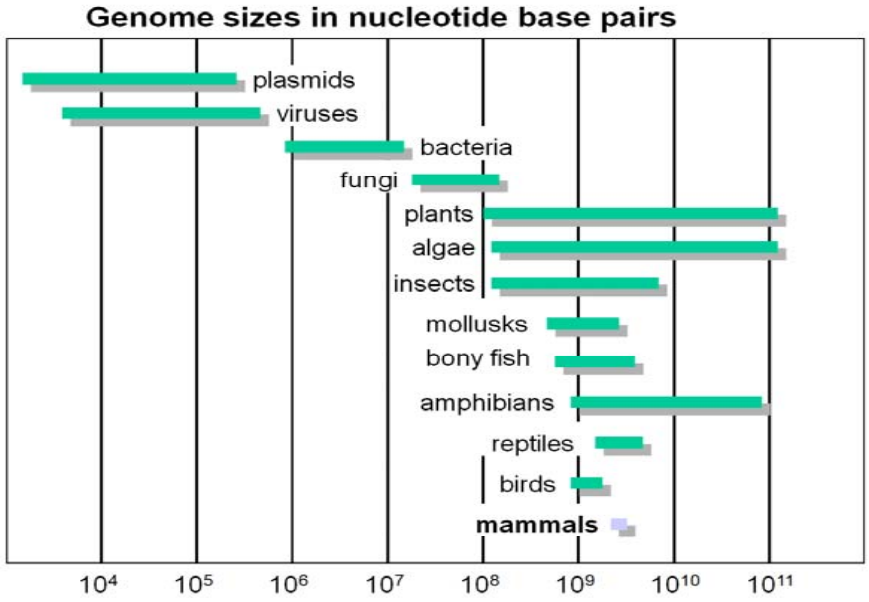
# Overview of Genome Analysis



# Criteria for selecting genomes for sequencing

Criteria include:

- genome size (some plants are >>>human genome)
- cost
- relevance to human disease (or other disease)
- relevance to basic biological questions
- relevance to agriculture



# Criteria for selecting genomes for sequencing

Sequence one individual genome, or several?

Try one…

--Each genome center may study one
chromosome from an organism

--It is necessary to measure polymorphisms
(e.g. SNPs) in large populations

For viruses, thousands of isolates may be sequenced.

For the human genome, cost is the impediment.

# Ancient DNA projects

Special challenges:

- Ancient DNA is degraded by nucleases
- The majority of DNA in samples derives from unrelated organisms such as bacteria that invaded after death
- The majority of DNA in samples is contaminated by human DNA
- Determination of authenticity requires special controls, and analysis of multiple independent extracts

# Metagenomics projects

Two broad areas:

- Environmental (ecological)
  e.g. hot spring, ocean, sludge, soil

- Organismal
  e.g. human gut, feces, lung

http://www.ncbi.nlm.nih.gov/sites/entrez?db=bioproject



# BioProject

A BioProject is a collection of biological data related to a single initiative, originating from a single organization or from a consortium. A BioProject record provides users a single place to find links to the diverse data types generated for that project.

## Using BioProject

- Help
- Submission

## Browse BioProject

- By Project attributes
- Download (FTP)

## Large Initiatives

- 1000 Genomes
- ENCODE
- HMP

## NCBI Resources

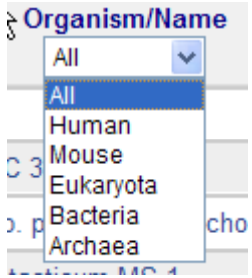
- BioSample
- dbGaP
- Genome

## External Resources

- Genome projects at DOE
- Genome News Network
- GOLD - Genome On Line Database

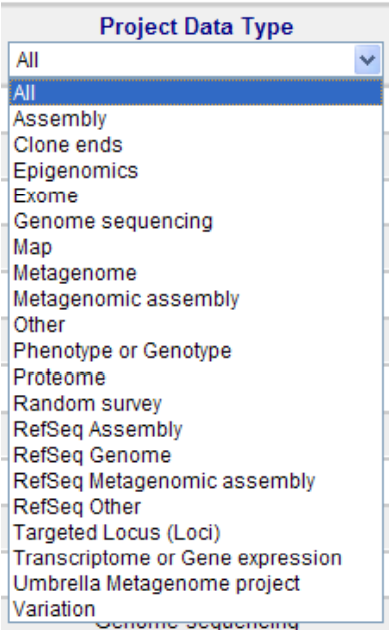
# BioProject

A BioProject is a collection of biological data related to a single initiative, originating from a single organization or from a consortium. A BioProject record provides users a single place to find links to the diverse data types generated for that project.



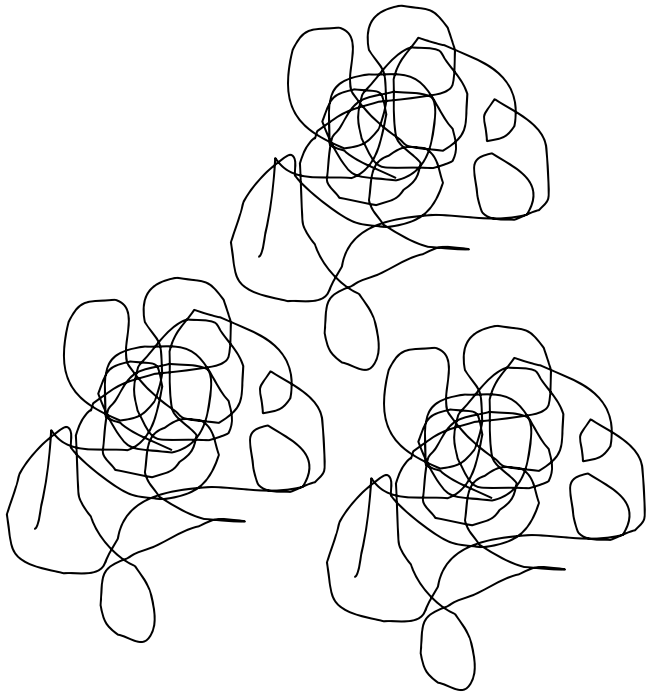

First Previous **Shown: 1 - 100 out of 49433 items** Next Last

| Project Accession | Organism/Name (All) | TaxID | Project Type (All) | Project Data Type (All) | Date |
|---|---|---|---|---|---|
| PRJNA3 | Borrelia burgdorferi B31 | 224326 | Primary submission | Genome sequencing | 2003/02/23 |
| PRJNA4 | Treponema denticola ATCC 35405 | 243275 | Primary submission | Genome sequencing | 2004/04/06 |
| PRJNA5 | Treponema pallidum subsp. pallidum str. Nichols | 243276 | Primary submission | Genome sequencing | 2003/02/25 |
| PRJNA6 | Magnetospirillum magnetotacticum MS-1 | 272627 | Primary submission | Genome sequencing | 2003/02/25 |
| PRJNA7 | Campylobacter fetus subsp. venerealis str. Azul-94 | 593452 | Primary submission | Genome sequencing | 2009/04/22 |
| PRJNA8 | Campylobacter jejuni subsp. jejuni NCTC 11168 = ATCC 700819 | 192222 | Primary submission | Genome sequencing | 2003/02/25 |
| PRJNA9 | Francisella tularensis subsp. tularensis SCHU S4 | 177416 | Primary submission | Genome sequencing | 2004/12/17 |
| PRJNA12 | Pseudomonas fluorescens Pf0-1 | 205922 | Primary submission | Genome sequencing | 2005/10/07 |
| PRJNA13 | Ralstonia solanacearum GMI1000 | 267608 | Primary submission | Genome sequencing | 2003/02/25 |
| PRJNA15 | Xanthomonas campestris pv. campestris str. 8004 | 314565 | Primary submission | Genome sequencing | 2005/05/25 |
| PRJNA16 | Azotobacter vinelandii DJ | 322710 | Primary submission | Genome sequencing | 2003/02/25 |
| PRJNA17 | Bradyrhizobium japonicum USDA 110 | 224911 | Primary submission | Genome sequencing | 2003/03/14 |
| PRJNA18 | Mesorhizobium loti MAFF303099 | 266835 | Primary submission | Genome sequencing | 2003/02/25 |
| PRJNA19 | Sinorhizobium meliloti 1021 | 266834 | Primary submission | Genome sequencing | 2003/02/25 |
| PRJNA20 | Methylobacterium extorquens AM1 | 272630 | Primary submission | Genome sequencing | 2009/06/17 |
| PRJNA21 | Methylococcus capsulatus str. Bath | 243233 | Primary submission | Genome sequencing | 2004/09/24 |
| PRJNA22 | Legionella pneumophila subsp. pneumophila str. Philadelphia 1 | 272624 | Primary submission | Genome sequencing | 2004/09/28 |
| PRJNA23 | Neisseria gonorrhoeae FA 1090 | 242231 | Primary submission | Genome sequencing | 2005/02/14 |
| PRJNA24 | Bordetella bronchiseptica RB50 | 257310 | Primary submission | Genome sequencing | 2003/08/12 |
| PRJNA25 | Bordetella parapertussis 12822 | 257311 | Primary submission | Genome sequencing | 2003/08/12 |
| PRJNA26 | Bordetella pertussis Tohama I | 257313 | Primary submission | Genome sequencing | 2003/08/12 |
| PRJNA27 | Bordetella avium 197N | 360910 | Primary submission | Genome sequencing | 2006/09/29 |
| PRJNA29 | Zymomonas mobilis subsp. mobilis ZM4 | 264203 | Primary submission | Genome sequencing | 2003/03/20 |

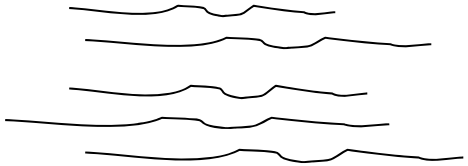


# Whole Genome Sequencing (WGS)

Multiple copies of DNA



**Fragments** of 200 - 200,000 bases



No information is retained on which part of the DNA the fragments came from.

# WGS sequencing: fragments

- We start with millions of pairs of reads, 100 - 1000 bases each
- Multiple copies of DNA provide multiple coverage by reads
- The problem of genome assembly is to recover the original sequence of bases of the genome (as much as possible…).

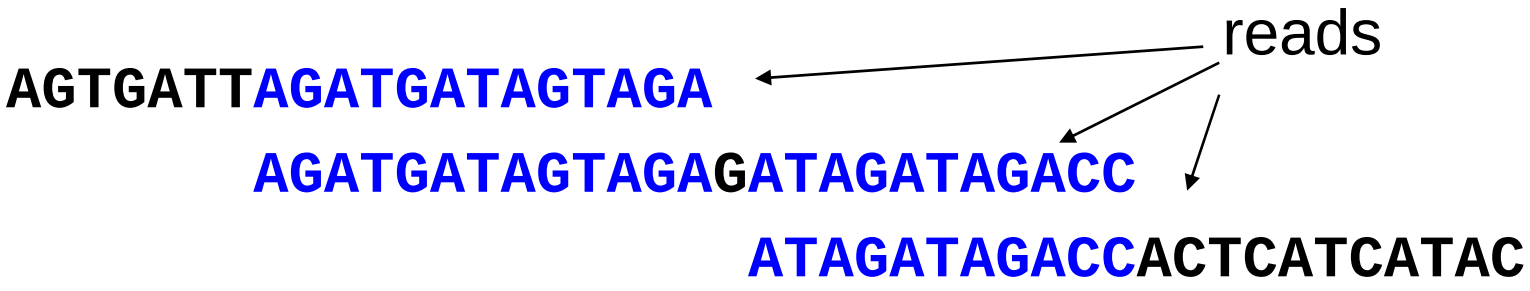
# Assembling a jigsaw puzzle 1

- The task of the assembly becomes the task of assembling a giant jigsaw puzzle
- We look for reads whose sequences suggest that they came from the same place in the genome:

```
AGTGATTAGATGATAGTAGA
         |||||||||
         GATGATAGTAGAGGATAGATTTA
```

# Assembling a jigsaw puzzle 2

- Then we put “overlapping” reads together

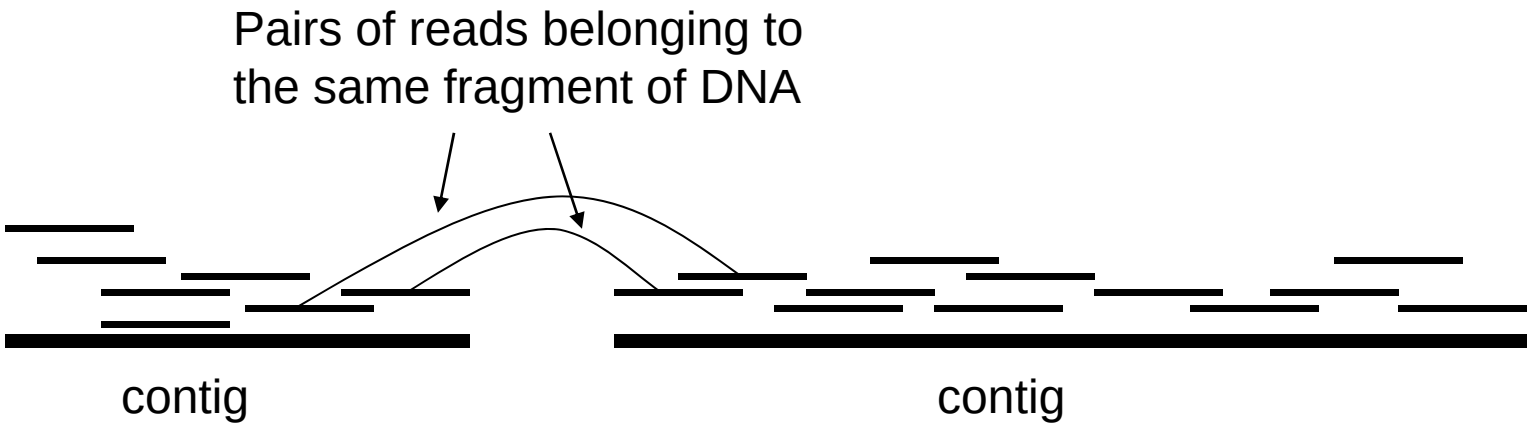



**AGTGATTAGATGATAGTAGAGATAGATAGACCACTCATCATAC**

This yields a “**contig**”

# Assembling a jigsaw puzzle 3

- We use read pairing information to order and orient contigs to produce **scaffolds** – the final product of assembly



# Difficulties in NGS assembly

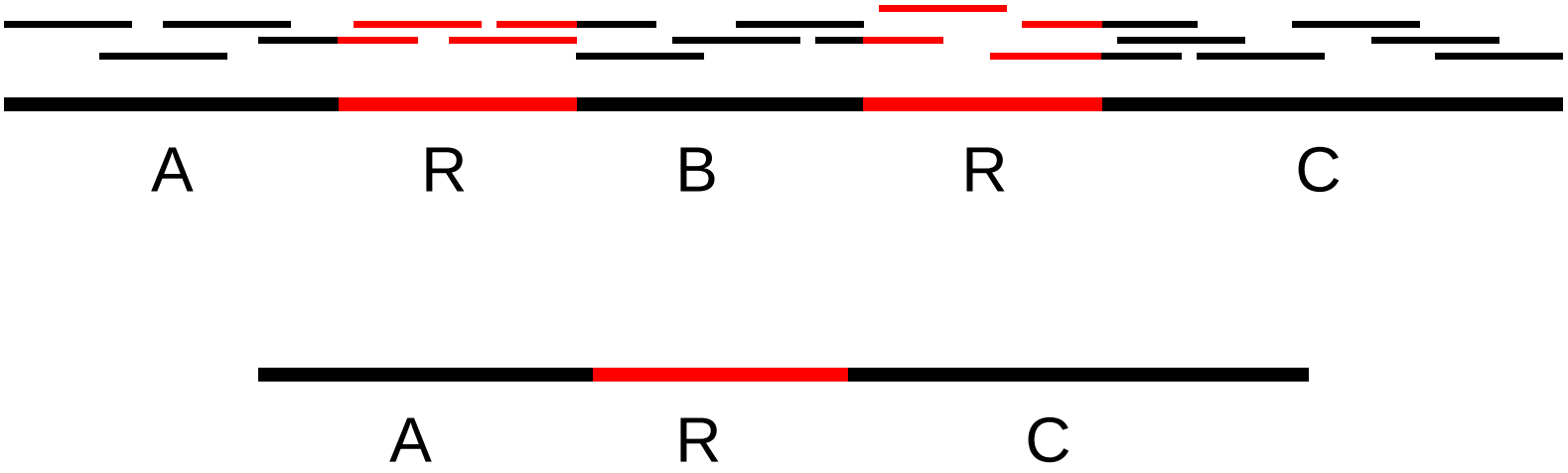
- Sequencing errors: two reads that came from the same place in the genome often have mismatching sequences

```
AGTGATTAGATCATAGTAGAG
          || |||||||||
          ATGATAGTAGAGGATAGAT
```

- Repetitive DNA (~ 5-20% of human DNA is repetitive):

```
TTAGGGTTAGGGTTAGGGTTAGGGTTAGGG
```

# Repeat regions may cause omissions

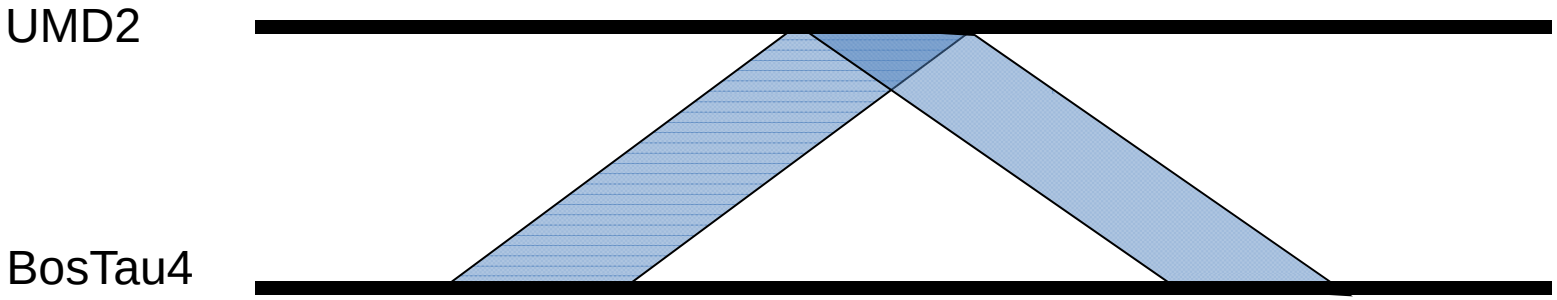



(1) Long insert library :10kb
(2) Mate-paired librared
(3) Long read : 3-4 Kb from 3rd Generation sequencer.

# Erroneous duplications

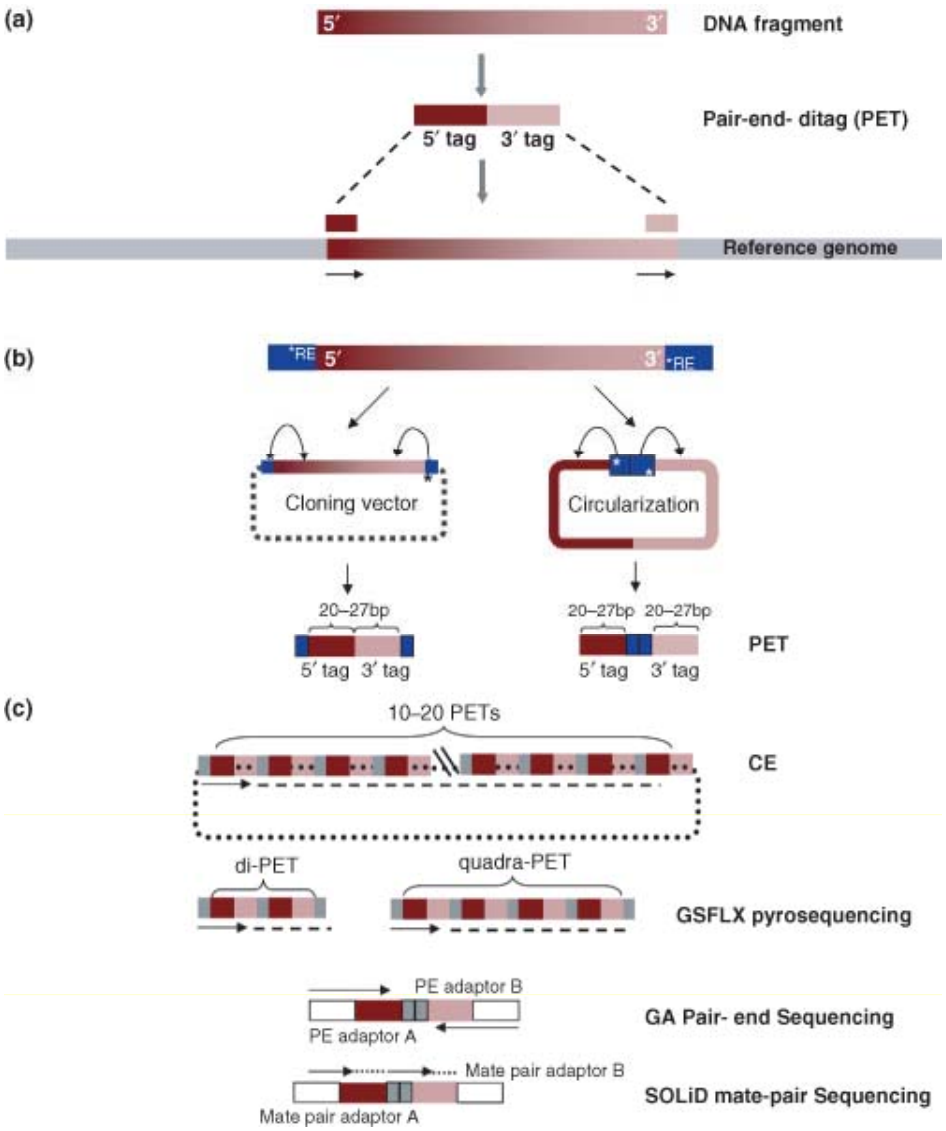
- Two recent published assemblies of the cow genome: UMD2 and BosTau4
- Segmental duplications were a central theme in BosTau4 genome paper
- UMD2 assembly had many fewer duplications

We examined the duplications, > 99.5% identity, >5000bp, one copy in the UMD2 assembly and two copies in the BosTau4




Each base in the genome is covered by 6 reads, on average. A way to judge which assembly is correct is to compute the average read coverage for these regions.

# Next Gen vs. Sanger Sequencing

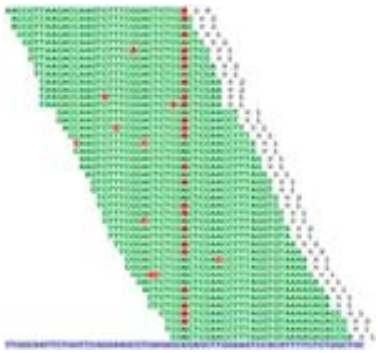


# De novo Sequencing vs Re-sequencing

**Assembly**



**Mapping**

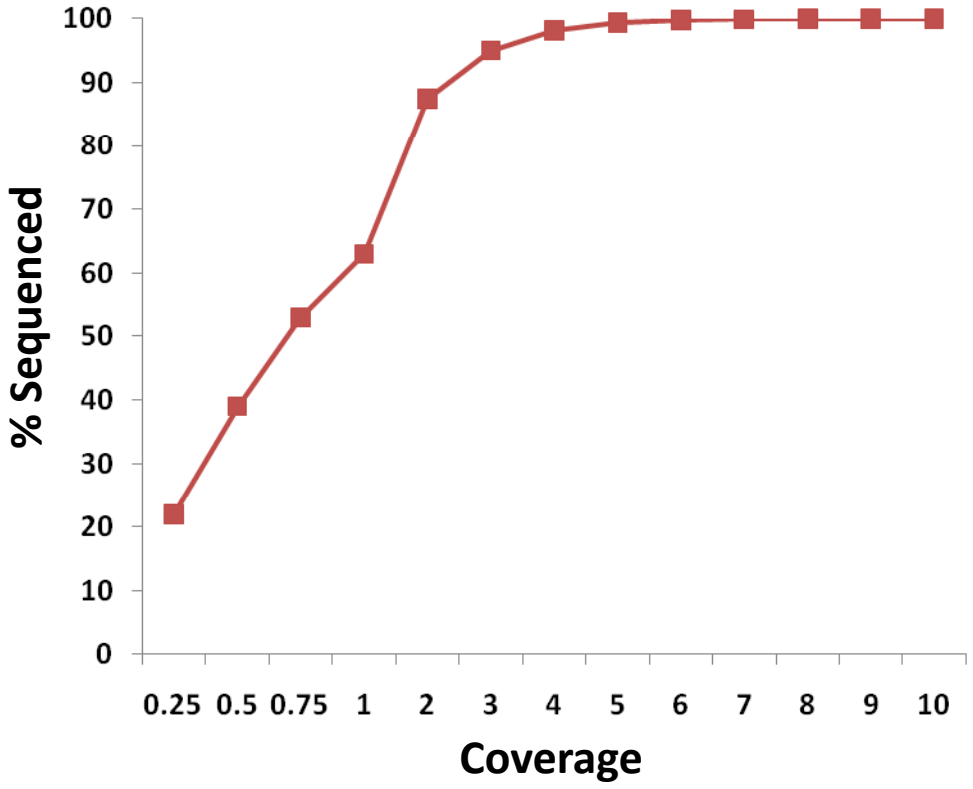


Assembly Tools
ABySS
ALLPATHS
Edena
Euler-SR
SHARCGS
SHRAP
SSAKE
Velvet

**CLC Genomics**

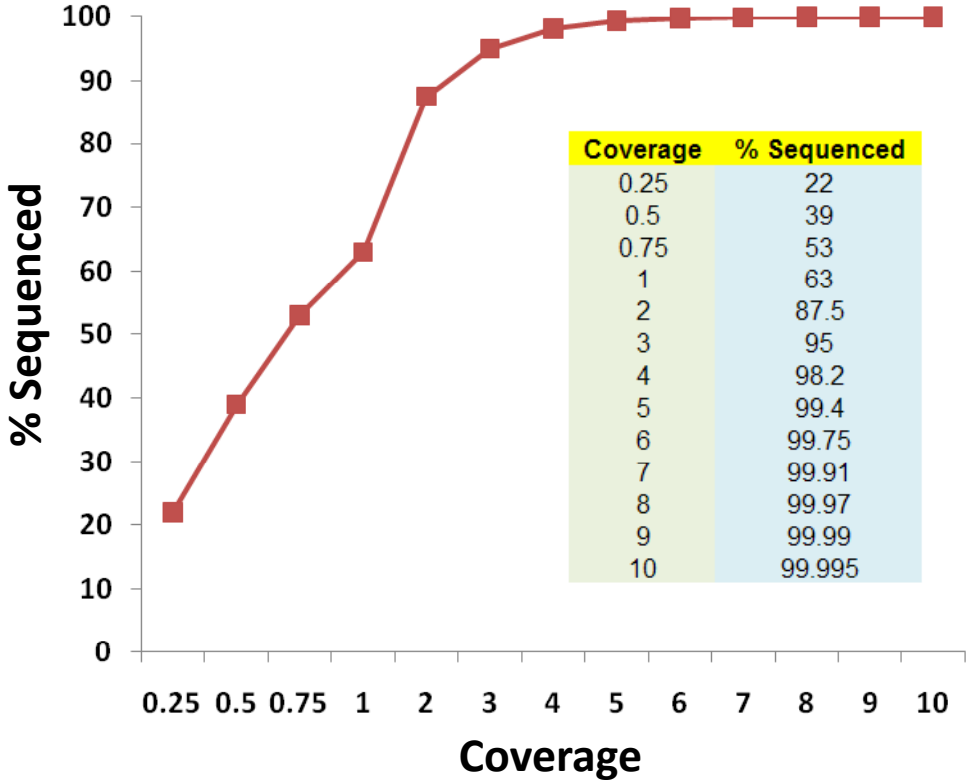
Alignment Tools
Cross_match
ELAND
Exonerate
MAQ
Mosaik
SHRiMP
SOAP
Zoom

# When has a genome been fully sequenced?




| Coverage | % Sequenced |
|---|---|
| 0.25 | 22 |
| 0.5 | 39 |
| 0.75 | 53 |
| 1 | 63 |
| 2 | 87.5 |
| 3 | 95 |
| 4 | 98.2 |
| 5 | 99.4 |
| 6 | 99.75 |
| 7 | 99.91 |
| 8 | 99.97 |
| 9 | 99.99 |
| 10 | 99.995 |

# Read coverage




| Coverage | % Sequenced |
|---|---|
| 0.25 | 22 |
| 0.5 | 39 |
| 0.75 | 53 |
| 1 | 63 |
| 2 | 87.5 |
| 3 | 95 |
| 4 | 98.2 |
| 5 | 99.4 |
| 6 | 99.75 |
| 7 | 99.91 |
| 8 | 99.97 |
| 9 | 99.99 |
| 10 | 99.995 |

Sanger sequencing ~1000bp

NGS sequencing

Solexa: ~100bp

SOLiD: ~70bp

For 99.75% - 99.99% Accuracy

**NEED 60X - 100X COVERAGE**

# Prokaryotic genomes sequencing

| | Sanger | unpaired + paired 454 + illumina GAI 36bp | Unpaired 454 + illumina GAIIx MP 4,5Kb 76bp | Illumina GAIIx MP 4,5Kb 76bp |
|---|---|---|---|---|
| Coverage | 7.4X | 25X and 50X | 25X and 40X | 22X |
| Assembler | Arachne (Broad Institute) | Newbler (454 / Roche) | Newbler (454 / Roche) | Soap (BGI) |
| # of contigs | 173 | 119 | 44 | 495 |
| Contigs N50 (Kb) | 39.0 | 58.2 | 197 | 12.2 |
| # of scaffolds | 2 | 10 | 9 | 84 |
| Scaffolds N50 (Kb) | 2,200 | 1,000 | 1,009 | 818 |
| Assembly size (% of reference) | 3.417Mb (95%) | 3.544 Mb (98%) | 3.567 (99%) | 3.674 Mb (102%) |
| Mis-assemblies | 0 | 0 | 0 | 0 |
| # of errors | 3,442 | 163 1 error / 22Kb | 63 1 error / 55Kb | 35 1 error / 100Kb |
| Substitutions | 2,494 | 71 | 60 | 33 |
| Insertions / Deletions | 948 | 92 | 4 | 2 |

# Eukaryotic Genome Sequencing Evolution

✓ Extend prokaryotic strategy to eukaryotic genomes

✓ Sanger sequencing is still used to sequence long DNA fragments :
>20Kb, BAC ends, …

