

# Getting NGS data

Aim : Associate any NGS related study publication with the databases. Link the raw data files with the study and download all the relevant data.

Course Wiki: <http://10.152.154.17/wiki/doku.php>

# Last Wednesday

- NGS study types
  - Genome Assembly
  - Resequencing
  - Gene Expression
  - Metagenomics
  - Epigenomics
- Navigating NCBI Taxonomy
- Species Specific information
- Related Species information

# Getting NGS data

- Navigating NGS Data repositories
- Understanding How data is arranged
- Linkage between different datasets
- Downloading NGS

# Data Repositories

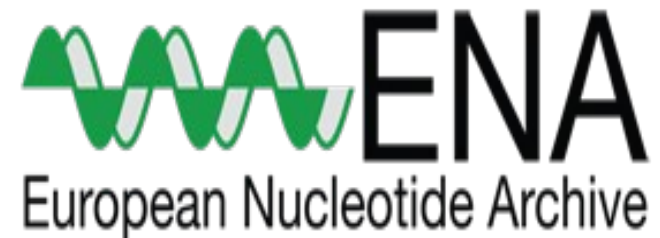
NCBI Resources How To Sign in to NCBI

SRA SRA Search

Advanced Help

**SRA**

Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and



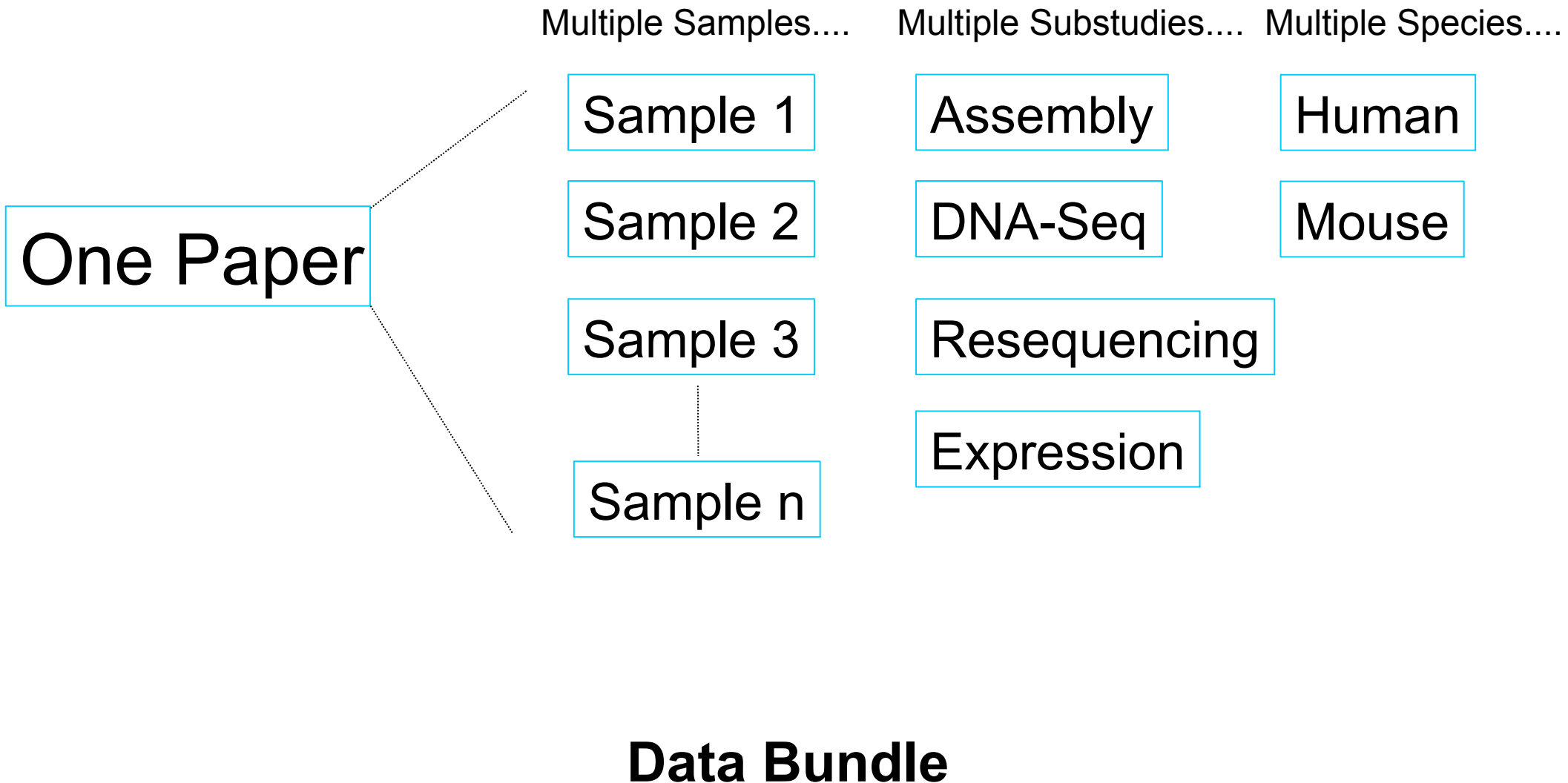
# Data repositories

- NCBI SRA is the main repository
- SRA links out to other NCBI databases (e.g. pubmed, bioproject, taxonomy )
- DDBJ and ENA support SRA
- All 3 are synced regularly
- SRA ids can be used in DDBJ and ENA
- Most Journals ask mandatory data submission to SRA
- Embargos and privacy concerns might make some data unavailable
- SRA has been criticized for its complexity and its future is uncertain

# Main Goal

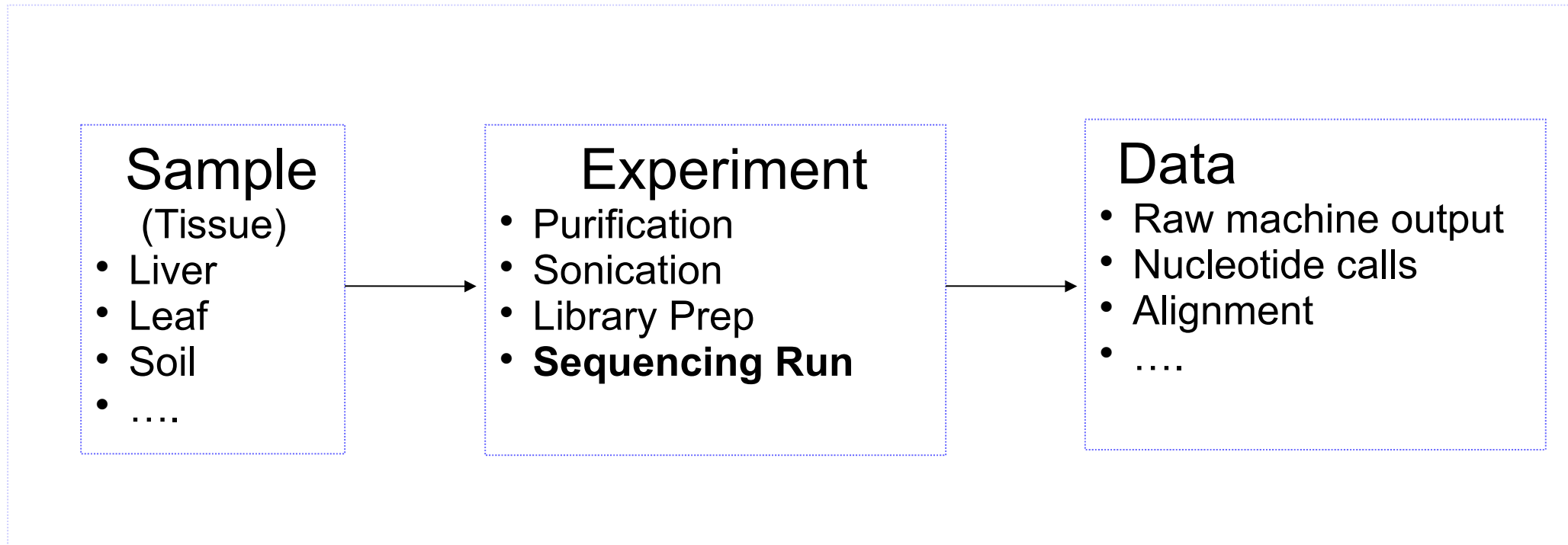
- Take a relevant project (Publication)
- Dissect it in different modules
- Understand each module
- Find each module in database
- Get the data for each experiment
- Access files
  - .sra
  - FastQ
  - Bam
  - Bed

# Why now Download directly ?



# Basic Module

Study

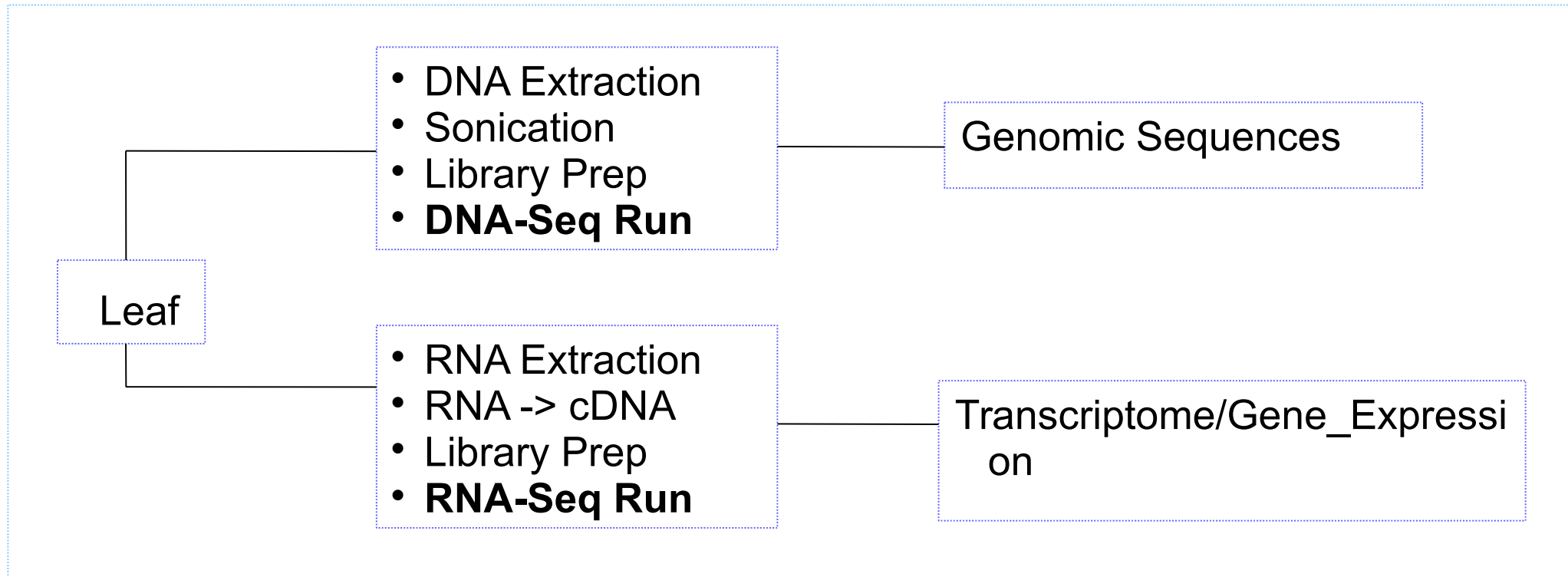


**S → E → D**

**Sample** → **Experiment** → **Data**



# Study



# Study

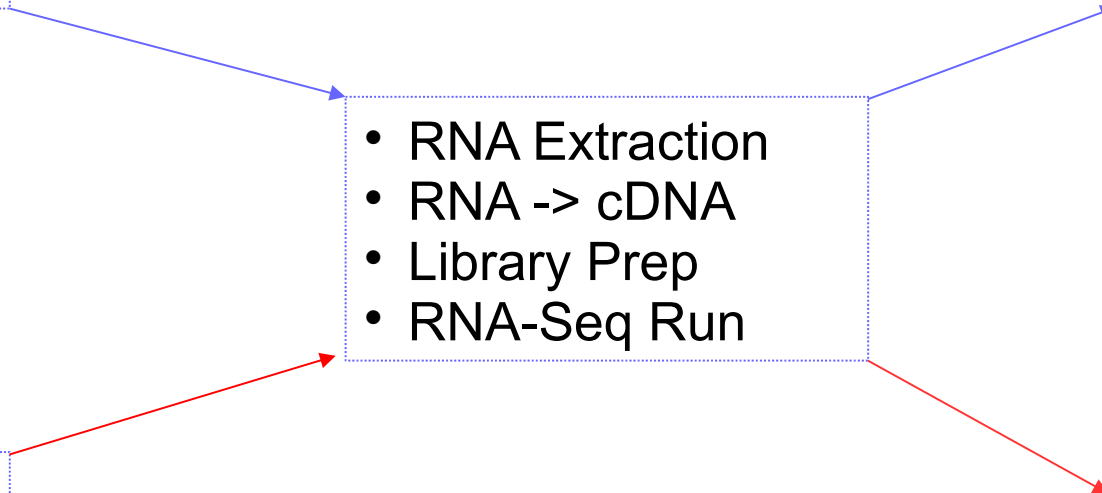
Healthy  
Leaf

Control Transcriptome

- RNA Extraction
- RNA -> cDNA
- Library Prep
- RNA-Seq Run

Infected  
Leaf

Case Transcriptome



# Study Structure

$S \rightarrow E \rightarrow D$

$S \rightarrow E \rightarrow D$

$S \rightarrow E \rightarrow D$

$S \rightarrow E \rightarrow D$

$S \rightarrow E \rightarrow D$

# BioProject: Complicated Studies

- Called as BioProjects
- Big projects (Initiative)
- Organizations/Consortium
- Many studies in one project

# BioProject Structure

$S \rightarrow E \rightarrow D$

$S \rightarrow E \rightarrow D$

$S \rightarrow E \rightarrow D$

$S \rightarrow E \rightarrow D$

$S \rightarrow E \rightarrow D$

$S \rightarrow E \rightarrow D$

$S \rightarrow E \rightarrow D$

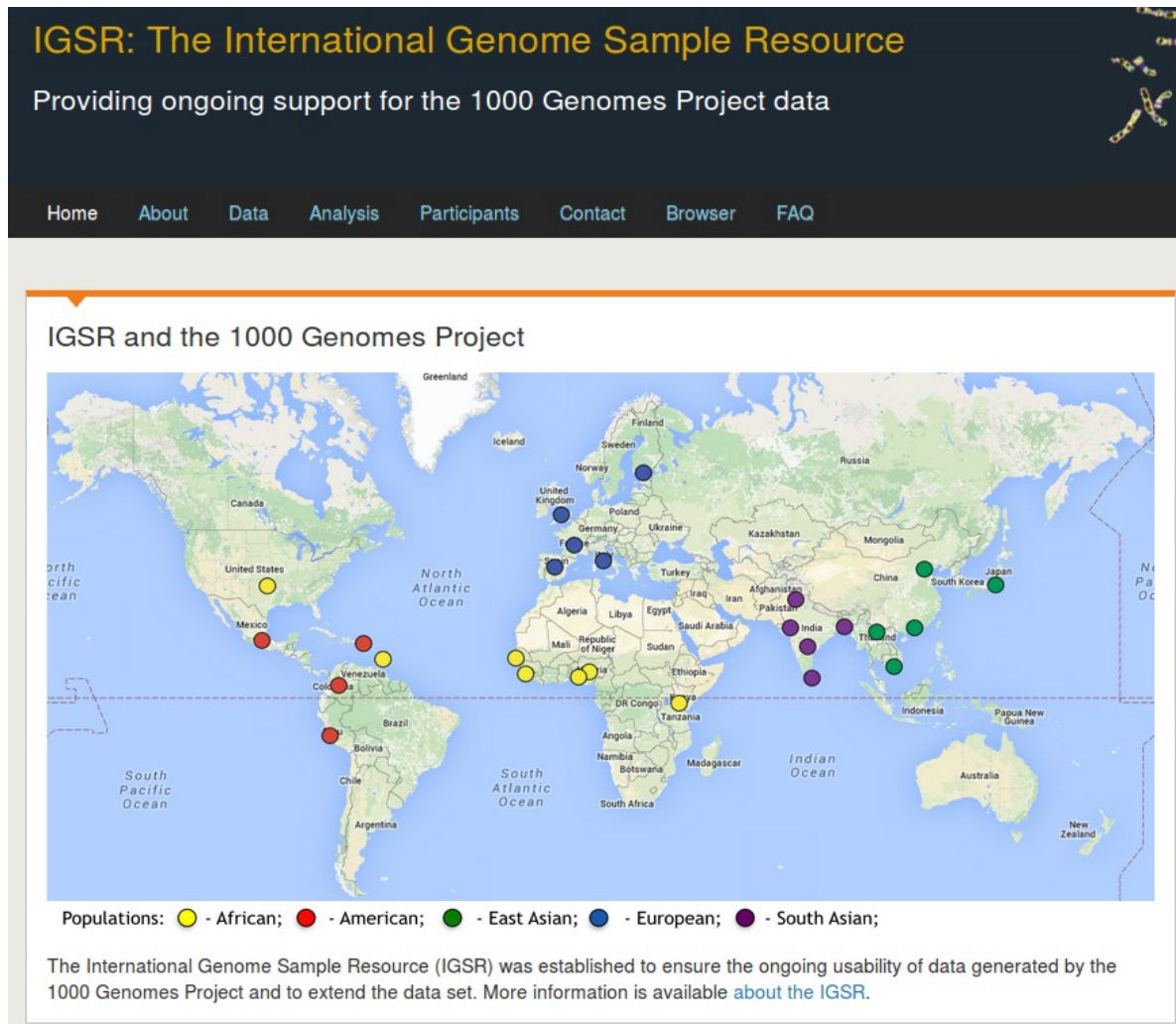
$S \rightarrow E \rightarrow D$

$S \rightarrow E \rightarrow D$

$S \rightarrow E \rightarrow D$

$S \rightarrow E \rightarrow D$

# 1000 genomes bioproject



Accession : PRJNA28889

<http://www.1000genomes.org/>

# ENCODE Project



Encyclopedia of DNA Elements at UCSC 2003 - 2012

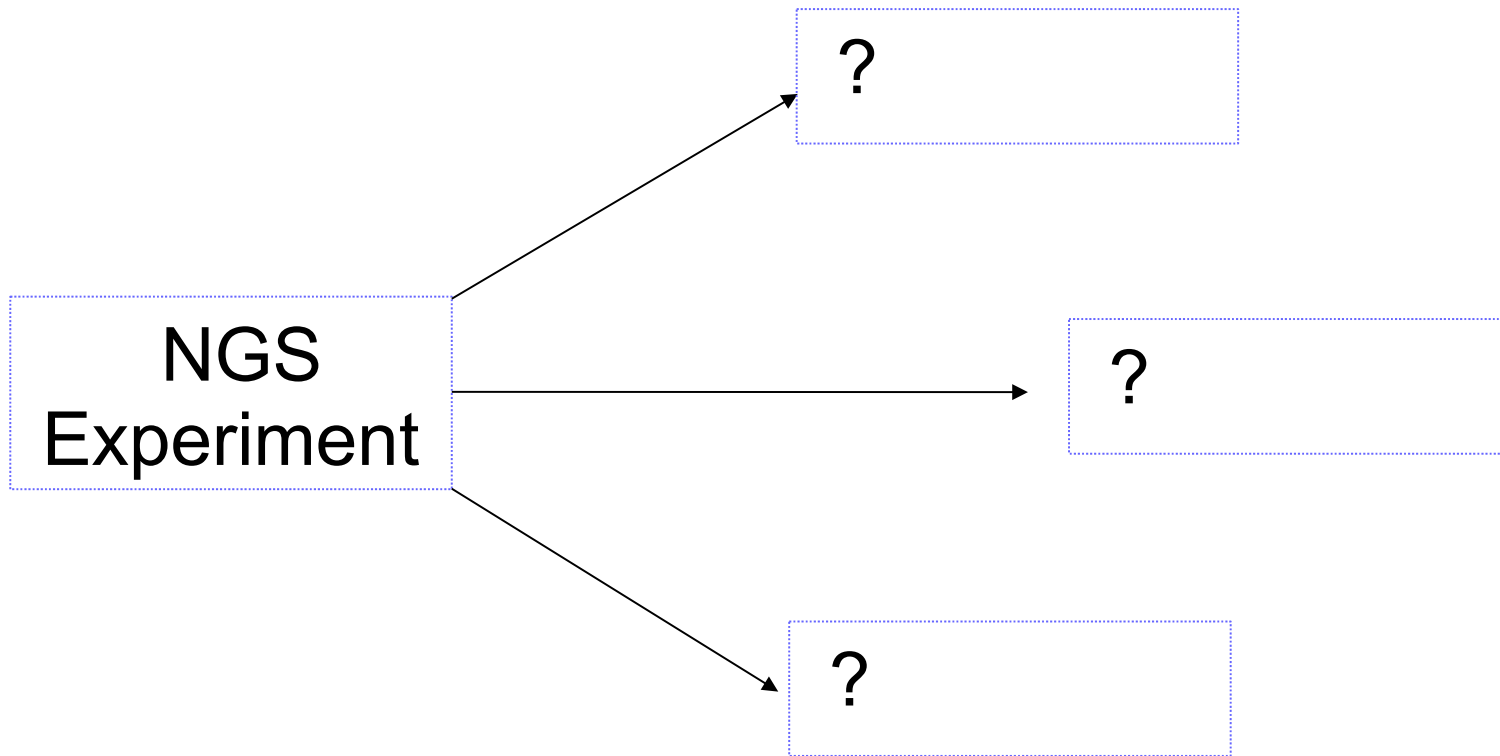
## About

The [Encyclopedia of DNA Elements](#) (ENCODE) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute ([NHGRI](#)). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

**ENCODE results from 2007 and later are available from the ENCODE Project Portal, [encodeproject.org](http://encodeproject.org).** This covers data generated during the two production phases 2007-2012 and 2013-present. The ENCODE Project Portal also hosts additional ENCODE access tools, and ENCODE project pages including up-to-date information about data releases, publications, and upcoming tutorials.

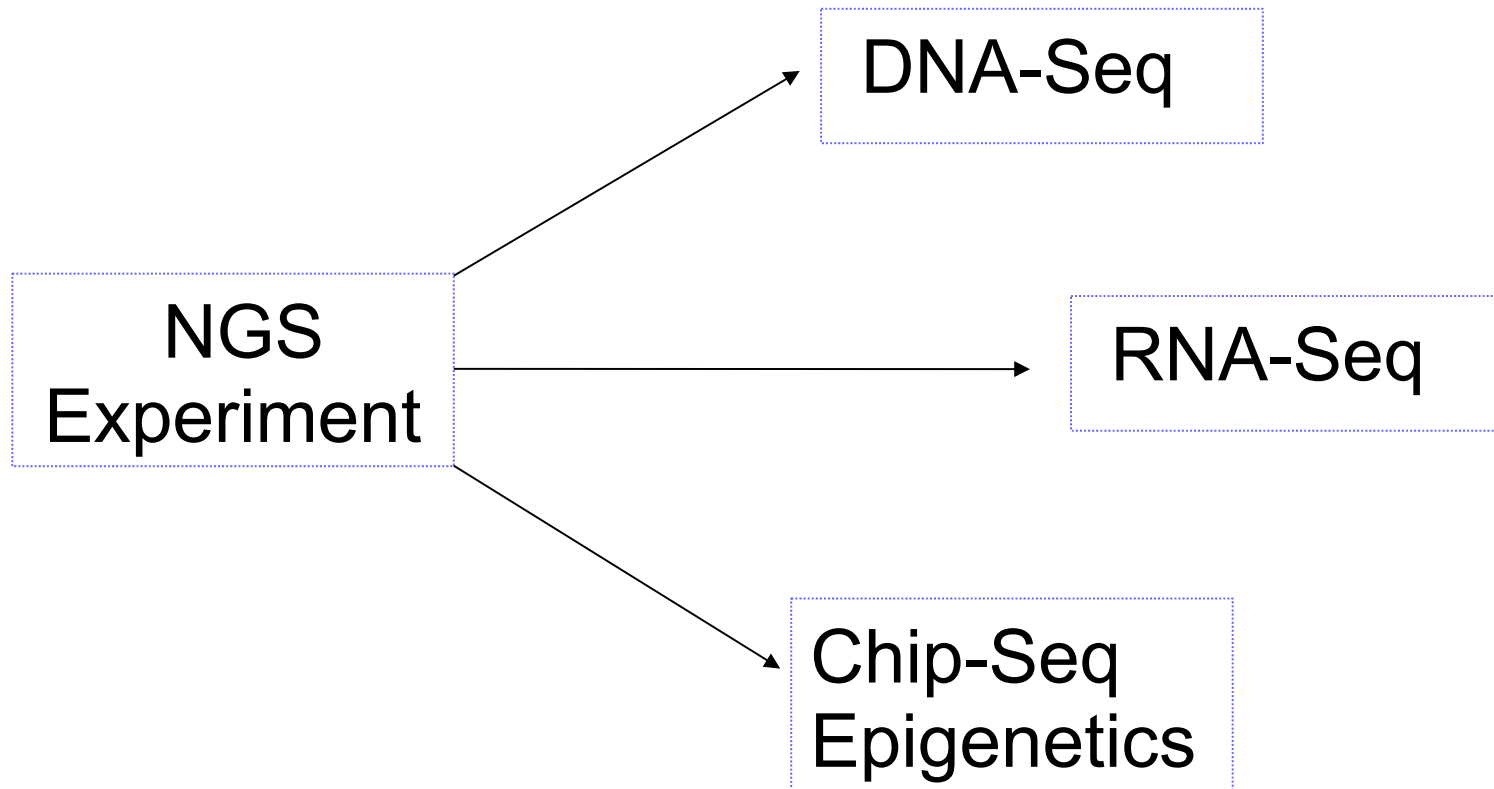
UCSC coordinated data for the ENCODE Consortium from its inception in 2003 (Pilot phase) to the end of the first 5 year phase of whole-genome data production in 2012. All data produced by ENCODE investigators and the results of ENCODE analysis projects from this period are hosted in the UCSC Genome browser and database. Explore ENCODE data using the image links below or via the left menu bar. **All ENCODE data at UCSC are freely available for download and analysis.**

# Linking by experiment type





# Linking by experiment type



# LinkOut

DNA-Seq

404

RNA-Seq

## Gene Expression Omnibus



GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

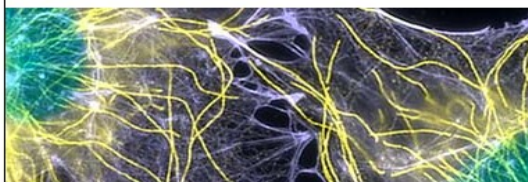
Epigenomics

Epigenomics

Epigenomics

[Advanced](#)

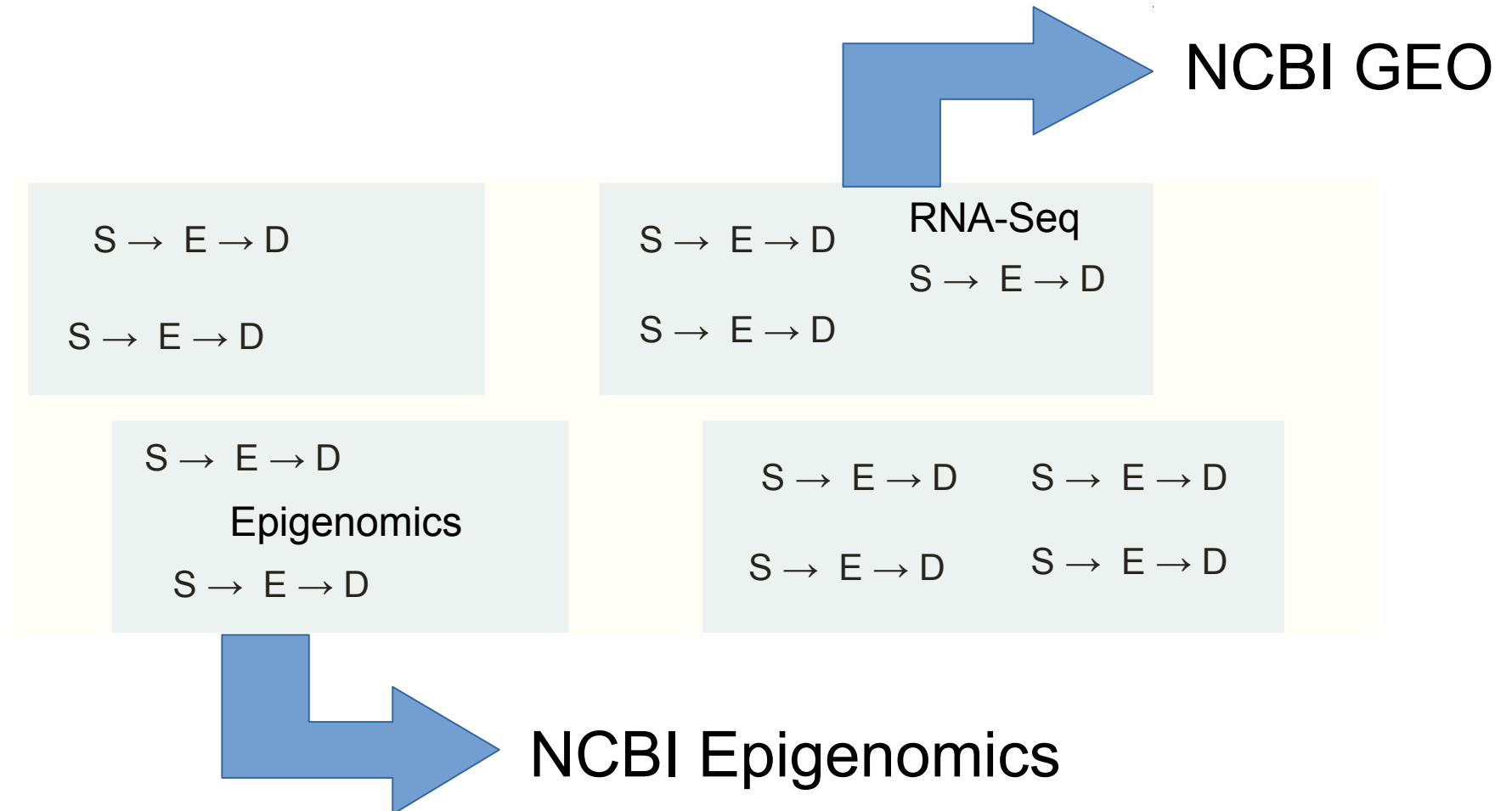
[Help](#)



## Epigenomics

Explore, view, and download genome-wide maps of DNA and histone modifications from our diverse collection of epigenomic data sets

# LinkOut



# Recap

- $S \rightarrow E \rightarrow D$
- Study  $[(S \rightarrow E \rightarrow D), (S \rightarrow E \rightarrow D), (S \rightarrow E \rightarrow D) \dots]$
- BioProject  $[[ (S \rightarrow E \rightarrow D), (S \rightarrow E \rightarrow D) \dots ], [(S \rightarrow E \rightarrow D), (S \rightarrow E \rightarrow D) \dots], [(S \rightarrow E \rightarrow D), (S \rightarrow E \rightarrow D) \dots ], \dots ]$
- NCBI-GEO LinkOut
- NCBI-Epigenomics LinkOut

# BioProject PRJNA\*

## BioSample

- SRS\*/ERS\*
- SAM
- Organism
- Tissue
- ...

## Experiment

- SRX\*/ERX\*
- Machine
- Protocol
- ...

## Study

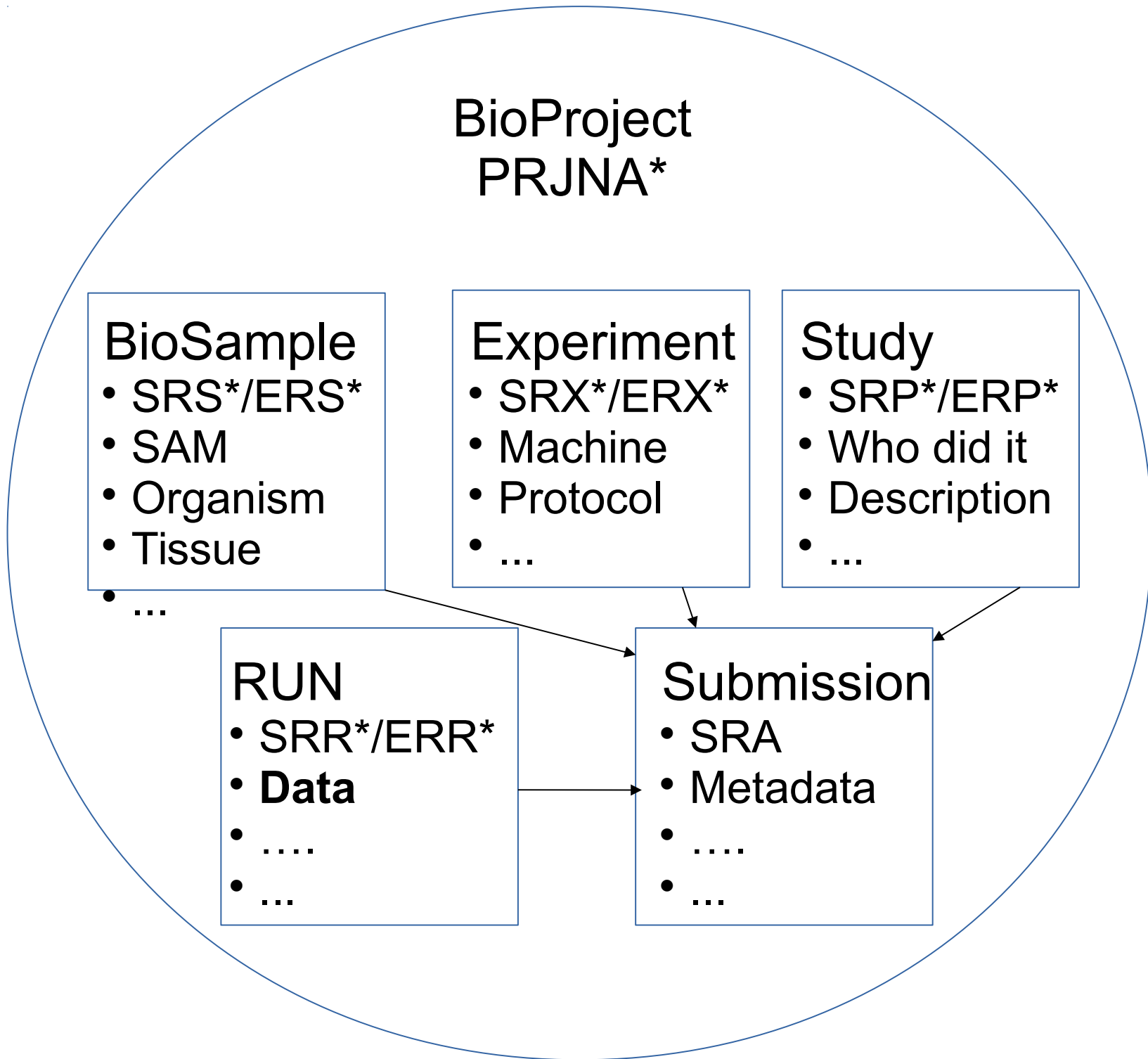
- SRP\*/ERP\*
- Who did it
- Description
- ...

## RUN

- SRR\*/ERR\*
- **Data**
- ....
- ...

## Submission

- SRA
- Metadata
- ....
- ...



# Different SRA ID types

- **Study** (SRP)– A study is a set of experiments and has an overall goal.
- **Experiment** (SRX) – An experiment is a consistent set of laboratory operations on input material with an expected result.
- **Sample** (SRS)– An experiment targets one or more samples. Results are expressed in terms of individual samples or bundles of samples as defined by the experiment.
- **Run** (SRR)– Results are called runs. Runs comprise the data gathered for a sample or sample bundle and refer to a defining experiment.
- **Submission** (SRA) – A submission is a package of metadata and/or data objects and a directive for what to do with those objects.

# PROBLEM

- Explain the type for each id, get the information listed in NCBI
  - SRS372877
  - SRS1165894
  - SRX527715
  - SRP039339
  - SRR1262828

OPEN

EXPAND →

© 2013 American Society of Plant Biologists. All Rights Reserved.

## RNA-Seq of Arabidopsis Pollen Uncovers Novel Transcription and Alternative Splicing<sup>1[C][W][OA]</sup>

Ann E. Loraine<sup>\*</sup>, Sheila McCormick, April Estrada, Ketan Patel<sup>2</sup> and Peng Qin<sup>3</sup>

+ Author Affiliations

← <sup>\*</sup>Corresponding author; e-mail [aloraine@uncc.edu](mailto:aloraine@uncc.edu).

First Published on April 16, 2013, doi: <http://dx.doi.org/10.1104/pp.112.211441>

Plant Physiology June 2013 vol. 162 no. 2 1092-1109

Free via OPEN

Go  
View this article with LENS

### ↑↓ Navigate This Article

Top

Abstract

RESULTS

DISCUSSION

MATERIALS AND METHODS

Acknowledgments

Footnotes

REFERENCES

Cited By ?



BioSample = ?

Experiment = ?



OPEN

EXPAND ⇨

© 2013 American Society of Plant Biologists. All Rights Reserved.

## RNA-Seq of Arabidopsis Pollen Uncovers Novel Transcription and Alternative Splicing<sup>1[C][W][OA]</sup>

Ann E. Loraine<sup>\*</sup>, Sheila McCormick, April Estrada, Ketan Patel<sup>2</sup> and Peng Qin<sup>3</sup>

+ Author Affiliations

←<sup>\*</sup>Corresponding author; e-mail [aloraine@uncc.edu](mailto:aloraine@uncc.edu).

First Published on April 16, 2013, doi: <http://dx.doi.org/10.1104/pp.112.211441>

Plant Physiology June 2013 vol. 162 no. 2 1092-1109

Free via OPEN



### ↑↓ Navigate This Article

Top

Abstract

RESULTS

DISCUSSION

MATERIALS AND METHODS

Acknowledgments

Footnotes

REFERENCES

Cited By ?



Study

BioSample = Pollen

Experiment = RNA-Seq

S → E → D

Pollen → RNA-Seq → Data

# RNA-Seq of Arabidopsis Pollen Uncovers Novel Transcription and Alternative Splicing<sup>1[Cl][W][OA]</sup>

Ann E. Loraine\*, Sheila McCormick, April Estrada, Ketan Patel<sup>2</sup>, and Peng Qin<sup>3</sup>

Department of Bioinformatics and Genomics, University of North Carolina, Kannapolis, North Carolina 28081 (A.E.L., A.E., K.P.); and Plant Gene Expression Center, United States Department of Agriculture-Agricultural Research Service/University of California-Berkeley, Albany, California 94710 (S.M., P.Q.)

Pollen grains of Arabidopsis (*Arabidopsis thaliana*) contain two haploid sperm cells enclosed in a haploid vegetative cell. Upon germination, the vegetative cell extrudes a pollen tube that carries the sperm to an ovule for fertilization. Knowing the identity, relative abundance, and splicing patterns of pollen transcripts will improve our understanding of pollen and allow investigation of tissue-specific splicing in plants. Most Arabidopsis pollen transcriptome studies have used the ATH1 microarray, which does not assay splice variants and lacks specific probe sets for many genes. To investigate the pollen transcriptome, we performed high-throughput sequencing (RNA-Seq) of Arabidopsis pollen and seedlings for comparison. Gene expression was more diverse in seedling, and genes involved in cell wall biogenesis were highly expressed in pollen. RNA-Seq detected at least 4,172 protein-coding genes expressed in pollen, including 289 assayed only by nonspecific probe sets. Additional exons and previously unannotated 5' and 3' untranslated regions for pollen-expressed genes were revealed. We detected regions in the genome not previously annotated as expressed; 14 were tested and 12 were confirmed by polymerase chain reaction. Gapped read alignments revealed 1,908 high-confidence new splicing events supported by 10 or more spliced read alignments. Alternative splicing patterns in pollen and seedling were highly correlated. For most alternatively spliced genes, the ratio of variants in pollen and seedling was similar, except for some encoding proteins involved in RNA splicing. This study highlights the robustness of splicing patterns in plants and the importance of ongoing annotation and visualization of RNA-Seq data using interactive tools such as Integrated Genome Browser.

BioSample = Pollen, Seedlings

Seedlings → RNA-Seq → Data

Pollen → RNA-Seq → Data

# MATERIALS AND METHODS

## Sample Collection for RNA-Seq

*Arabidopsis* (*Arabidopsis thaliana*) Col-0 seeds were sown on soil, kept at 4°C for 3 d, and then transferred to a temperature-controlled growth room set to 22°C. Seedlings were grown in a 16-h/8-h light/dark cycle under 100 to 120  $\mu\text{mol m}^{-2} \text{s}^{-1}$ . After 21 d of growth, the aerial portions of the seedlings were collected twice per day for 5 d and pooled for RNA extraction. The experiment was then repeated following the same collection scheme, thus providing two distinct biological replicates. Mature, dry pollen was harvested from Col-0 plants using a vacuum collection device as described (Johnson-Brousseau and McCormick, 2004).

BioSample = Pollen, Seedlings (REP1), Seedlings (REP2)

# MATERIALS AND METHODS

## Sample Collection for RNA-Seq

*Arabidopsis* (*Arabidopsis thaliana*) Col-0 seeds were sown on soil, kept at 4°C for 3 d, and then transferred to a temperature-controlled growth room set to 22°C. Seedlings were grown in a 16-h/8-h light/dark cycle under 100 to 120  $\mu\text{mol m}^{-2} \text{s}^{-1}$ . After 21 d of growth, the aerial portions of the seedlings were collected twice per day for 5 d and pooled for RNA extraction. The experiment was then repeated following the same collection scheme, thus providing two distinct biological replicates. Mature, dry pollen was harvested from Col-0 plants using a vacuum collection device as described (Johnson-Brousseau and McCormick, 2004).

Seedlings(R1) → RNA-Seq → Data

Seedlings(R2) → RNA-Seq → Data

Pollen → RNA-Seq → Data



## RNA Extraction and Illumina Library Preparation

Seedlings or pollen were ground into a fine powder with a mortar and pestle, and RNA was isolated via TRI-reagent extraction with cleanup on Qiagen Plant RNeasy (catalog no. 74104) columns. All RNAs were treated with DNaseI using the Plant RNeasy Kit. A starting amount of 15  $\mu\text{g}$  of total RNA from each sample (seedling and pollen) was used in the library preparation, using the Illumina mRNA-seq Sample Preparation Kit (catalog no. RS-930-1001, part no. 1004898). mRNA was isolated and purified via Sera-Mag Magnetic Oligo(dT) Beads, washed, and then fragmented using divalent cations under elevated temperature. First- and second-strand cDNAs were synthesized, and an end-repair step was performed to convert overhangs into blunt ends. Next, the 3' ends were adenylated for ligation of the Illumina adapters. cDNA templates were then purified by gel isolation for a size selection of approximately 250 bp and amplified via PCR so that an adequate amount of sample was available for sequencing (10  $\mu\text{L}$  of a 25 ng  $\mu\text{L}^{-1}$  sample). Products were isolated with the Qiagen QIAquick PCR Purification Kit (catalog no. 28104) and added to the flow cell for sequencing.

## Data Availability

Illumina sequence data are available from NCBI under Short Read Archive accession SRP022162. Processed alignment, junction, and coverage graph files are available for visualization in Integrated Genome Browser via IGB QuickLoad data source <http://www.igbquickload.org/pollen>.

## Testing Our Model!

Seedlings(R1) → RNA-Seq → Data

Seedlings(R2) → RNA-Seq → Data

Pollen → RNA-Seq → Data

# Final Model

Seedlings(R1) → RNA-Seq → Data

Seedlings(R1) → RNA-Seq → Data

Seedlings(R2) → RNA-Seq → Data

Seedlings(R2) → RNA-Seq → Data

Pollen → RNA-Seq → Data

Pollen → RNA-Seq → Data

RESEARCH

Open Access



# Kiwi genome provides insights into evolution of a nocturnal lifestyle

Diana Le Duc<sup>1,2\*</sup>, Gabriel Renaud<sup>2</sup>, Arunkumar Krishnan<sup>3</sup>, Markus Sällman Almén<sup>3</sup>, Leon Huynen<sup>4</sup>, Sonja J. Prohaska<sup>5</sup>, Matthias Ongyerth<sup>2</sup>, Bárbara D. Bitarello<sup>6</sup>, Helgi B. Schiöth<sup>3</sup>, Michael Hofreiter<sup>7</sup>, Peter F. Stadler<sup>5</sup>, Kay Prüfer<sup>2</sup>, David Lambert<sup>4</sup>, Janet Kelso<sup>2</sup> and Torsten Schöneberg<sup>1\*</sup>

## Abstract

**Background:** Kiwi, comprising five species from the genus *Apteryx*, are endangered, ground-dwelling bird species endemic to New Zealand. They are the smallest and only nocturnal representatives of the ratites. The timing of kiwi adaptation to a nocturnal niche and the genomic innovations, which shaped sensory systems and morphology to allow this adaptation, are not yet fully understood.

**Results:** We sequenced and assembled the brown kiwi genome to 150-fold coverage and annotated the genome using kiwi transcript data and non-redundant protein information from multiple bird species. We identified evolutionary sequence changes that underlie adaptation to nocturnality and estimated the onset time of these adaptations. Several opsin genes involved in color vision are inactivated in the kiwi. We date this inactivation to the Oligocene epoch, likely after the arrival of the ancestor of modern kiwi in New Zealand. Genome comparisons between kiwi and representatives of ratites, *Galloanserae*, and *Neoaves*, including nocturnal and song birds, show diversification of kiwi's odorant receptors repertoire, which may reflect an increased reliance on olfaction rather than sight during foraging. Further, there is an enrichment of genes influencing mitochondrial function and energy expenditure among genes that are rapidly evolving specifically on the kiwi branch, which may also be linked to its nocturnal lifestyle.

BioSample=?

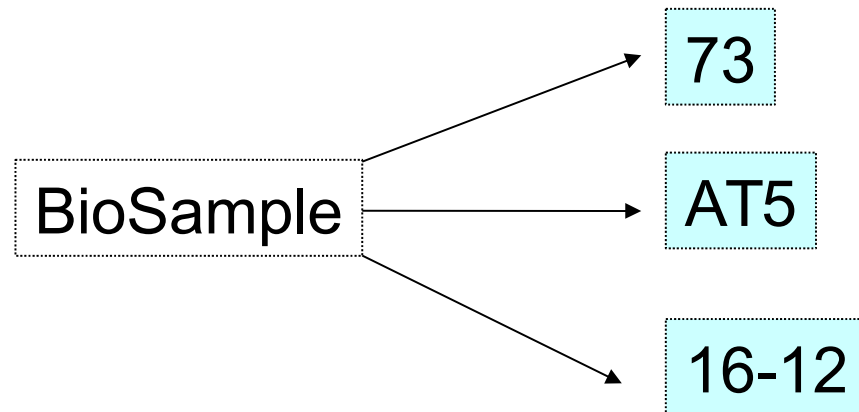
Experiment=DNA-Seq,RNA-Seq ?



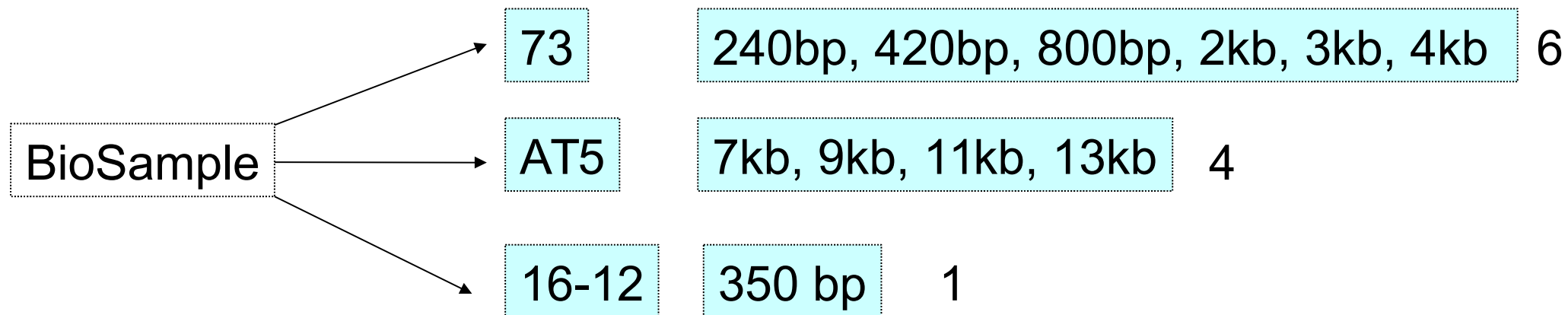
## Methods and materials

### Genome sequence assembly and annotation

We sequenced *Apteryx mantelli* female individuals, which originate from the far North (kiwi code 73) and central part – Lake Waikaremoana (kiwi code AT5 and kiwi code 16–12) of North Island (Additional file 1: Figure S10). They were sampled in 1986 (kiwi code 73) and 1997 (kiwi code AT5 and 16–12) in ‘operation nest egg’ carried out by Rainbow and Fairy Springs, Rotorua. No animals were killed or captured as a result of this study and genome assembly was performed with iwi approval from the Te Parawhau and Waikaremoana Māori Elders Trust.



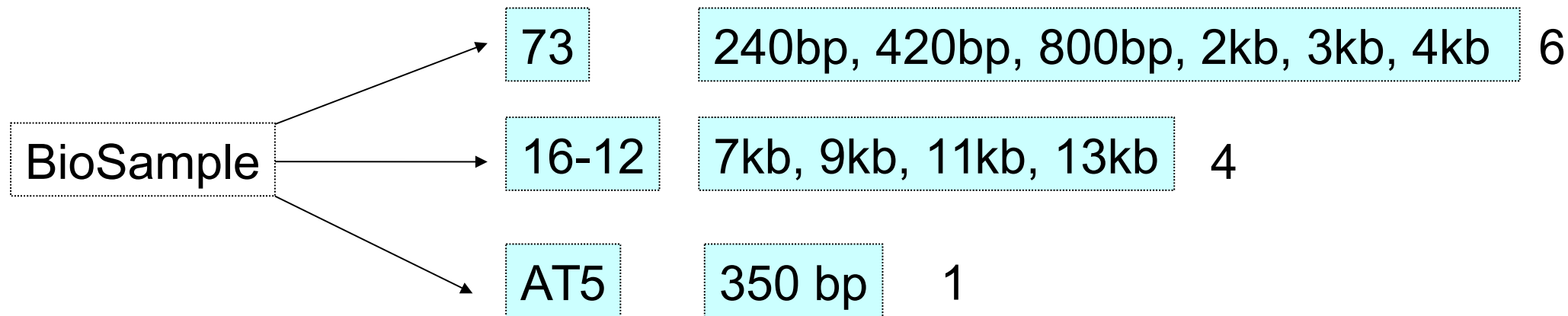
We extracted genomic DNA from *Apteryx mantelli* embryos. Libraries with insert sizes of 240 bp, 420 bp, 800 bp, 2 kb, 3 kb, and 4 kb were obtained from individual kiwi code 73, and mate-paired-end libraries 7 kb, 9 kb, 11 kb, and 13 kb, from individual kiwi code 16–12. DNA from individual AT5 was used to build a 350 bp insert-size library with the purpose of confirming kiwi-specific sequence polymorphisms and was not included in the genome assembly (Additional file 1: Note: Sampling, DNA library preparation and



sequencing; Additional file 1: Table S1). Paired-end sequencing was performed on HiScanSQ and HiSeq platforms with read lengths of 101 bp and 96 bp, respectively.

#### Data availability

Assembly, raw DNA, and RNA sequencing reads have been deposited in the European Nucleotide Archive under the BioProject with accession number: PRJEB6383.



# Model

DNA-Seq

73

S → E → D	S → E → D
S → E → D	S → E → D
S → E → D	S → E → D

DNA-Seq

AT5

S → E → D

DNA-Seq

16-12

S → E → D	
S → E → D	S → E → D
S → E → D	

RNA-Seq

???

S → E → D

12 entries in SRA ??

# Validation

DNA-Seq

73

$S \rightarrow E \rightarrow D$

$S \rightarrow E \rightarrow D$

$S \rightarrow E \rightarrow D$

$S \rightarrow E \rightarrow D$

$S \rightarrow E \rightarrow D$

$S \rightarrow E \rightarrow D$

DNA-Seq

AT5

$S \rightarrow E \rightarrow D$

DNA-Seq

16-12

$S \rightarrow E \rightarrow D$

$S \rightarrow E \rightarrow D$

$S \rightarrow E \rightarrow D$

$S \rightarrow E \rightarrow D$

RNA-Seq

???

$S \rightarrow E \rightarrow D$

# Problem

- Prepare a conceptual model for the following paper.
  - Get all sample, run and study identifiers.
  - Add these identifiers to the model

Xing et al. *Journal of Animal Science and Biotechnology* 2014, **5**:32  
<http://www.jasbsci.com/content/5/1/32>



JOURNAL OF ANIMAL SCIENCE  
AND BIOTECHNOLOGY

## RESEARCH

## Open Access

### The liver transcriptome of two full-sibling Songliao black pigs with extreme differences in backfat thickness

Kai Xing<sup>1</sup>, Feng Zhu<sup>1</sup>, Liwei Zhai<sup>1</sup>, Huijie Liu<sup>1</sup>, Zhijun Wang<sup>2</sup>, Zhuocheng Hou<sup>1\*</sup> and Chuduan Wang<sup>1\*</sup>

#### Abstract

**Background:** Fatness traits in animals are important for their growth, meat quality, reproductive performance, and immunity. The liver is the principal organ of the regulation of lipid metabolism, and this study used massive parallelized high-throughput sequencing technologies to determine the porcine liver tissue transcriptome architecture of two full-sibling Songliao black pigs harboring extremely different phenotypes of backfat thickness.

**Results:** The total number of reads produced for each sample was in the region of 53 million, and 8,226 novel transcripts were detected. Approximately 92 genes were differentially regulated in the liver tissue, while 31 spliced transcripts and 33 primary transcripts showed significantly differential expression between pigs with higher and lower backfat thickness. Genes that were differentially expressed were involved in the metabolism of various substances, small molecule biochemistry, and molecular transport.

**Conclusions:** Genes involved in the regulation of lipids could play an important role in lipid and fatty acid metabolism in the liver. These results could help us understand how liver metabolism affects the backfat thickness of pigs.

**Keywords:** Backfat thickness, Liver, Pig, RNA-Seq

## Methods

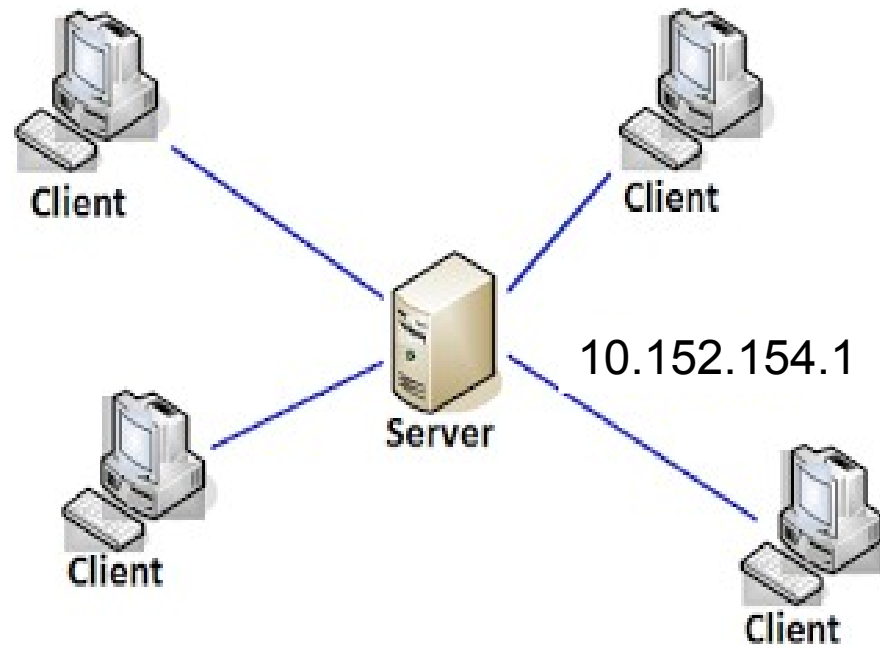
### Experimental design, animals, and phenotypes

The Songliao black female pig population (average age, 217 (range, 216–218) days; average live weight, 100 kg (range, 92.5–116.4 kg) was housed in consistent and standard environmental conditions with natural, uncontrolled room temperature and light. Animals were fed three times a day and had access to water *ad libitum*. Pedigree information was available for all animals. Live backfat thickness was measured on the last 3/4 rib using B-mode real-time ultrasound (HS1500, Honda, Japan). We analyzed a total of 53 individuals with full/half-sibs for backfat thickness to identify pairs with two divergent phenotypes. To minimize the noise of different genetic back grounds, full-sibs were selected as a priority.

We set out to compare transcriptome changes between two groups with a high variation in backfat thickness: pigs with higher backfat thickness (BH) and those with lower backfat thickness (BL) which had a backfat thickness 2–3 times lower than that of BH pigs. The chosen animals also had to have a similar backfat thickness within the same group (BH/BL) after adjustment for live body weight. Based on our criteria, experimental samples were made up of two pairs of pigs with extreme backfat thickness differences, both of which were full-sibs.



# Client Server Architecture



- All storage and computation happens on Server
- Client connects securely (Encryption, Passwords)
- Commands are issued by Client on terminal
- These are sent to the server and executed
- Screen Output of these commands is sent back to terminal
- No computation happens on the client



# Command Line Utilities

command arguments input output

- Filename
- Multiple
- Pattern

- Print to Screen
- Do Nothing

- Print to a file
- Precede by > or >>

- Pipe to another command
- Precede by |

- Essential
- Only one

- Optional
- Change command behavior
- Multiple
- **Space separated**
- Typically preceded by -
- Can be key=value pairs

# Example Commands

- ls
- head/tail
- grep
- wc
- pwd
- cat
- mkdir
- Touch
- locate

- cp
- mv
- cd
- less
- rm
- df
- cal
- file
- man

# Example patterns

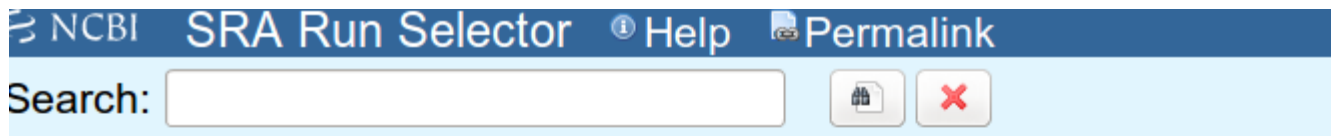
- ~ (Your home directory)
- . (Current directory)
- ../ (Directory above current directory)
- \* (All files)
- \*.txt (All files with extension .txt)
- \*abc\* (All files with character abc in filename)

# Accessing Data

- SRR/ERR entry is available for download
- Navigate to ERR/SRR for download
- Still library insert size is not mentioned in most cases
- Aggregated data for a single BioProject

# Getting a list of all Downloadable data for a Project

- SRA Run Selector
- Awesome Tool!

A screenshot of the NCBI SRA Run Selector web interface. The top navigation bar is dark blue with the NCBI logo, 'SRA Run Selector', a 'Help' link, and a 'Permalink' button. Below this is a light blue search area with the label 'Search:' followed by a text input field. To the right of the input field are two buttons: one with a document icon and another with a red 'X' icon.

Please type accession(s) of the studies, samples or experiments

- Studies:
  - SRA Study accessions (prefixes SRP, DRP, ERP)  
Examples: [SRP000002](#), [DRP000617](#), [ERP002000](#)
  - BioProject accessions (prefixes PRJNA, PRJDB, PRJEB)  
Examples: [PRJNA111397](#), [PRJDB90](#), [PRJEB1976](#)
  - dbGaP study accessions (prefix phs)  
Example: [phs000159](#)
  - GEO Study (prefix GSE)  
Example: [GSE12578](#)

# Direct Download

- Direct Download (ftp, http, Aspera) (Browser or commandline)
- Direct Download from EBI/DDBJ
- sra-toolkit software has a command fastq-dump

Command	Argument	Input	Output
fastq-dump	-h		Print help
fastq-dump		SRR_ID	Download entire file
fastq-dump	-X <number>	SRR_ID	Download N spots
fastq-dump	--skip-technical	SRR_ID	Do not include technical reads
fastq-dump	-Z	SRR_ID	Print to terminal
fastq-dump	-F	SRR_ID	Get original id
fastq-dump	--split-files	SRR_ID	Print read pairs in separate files

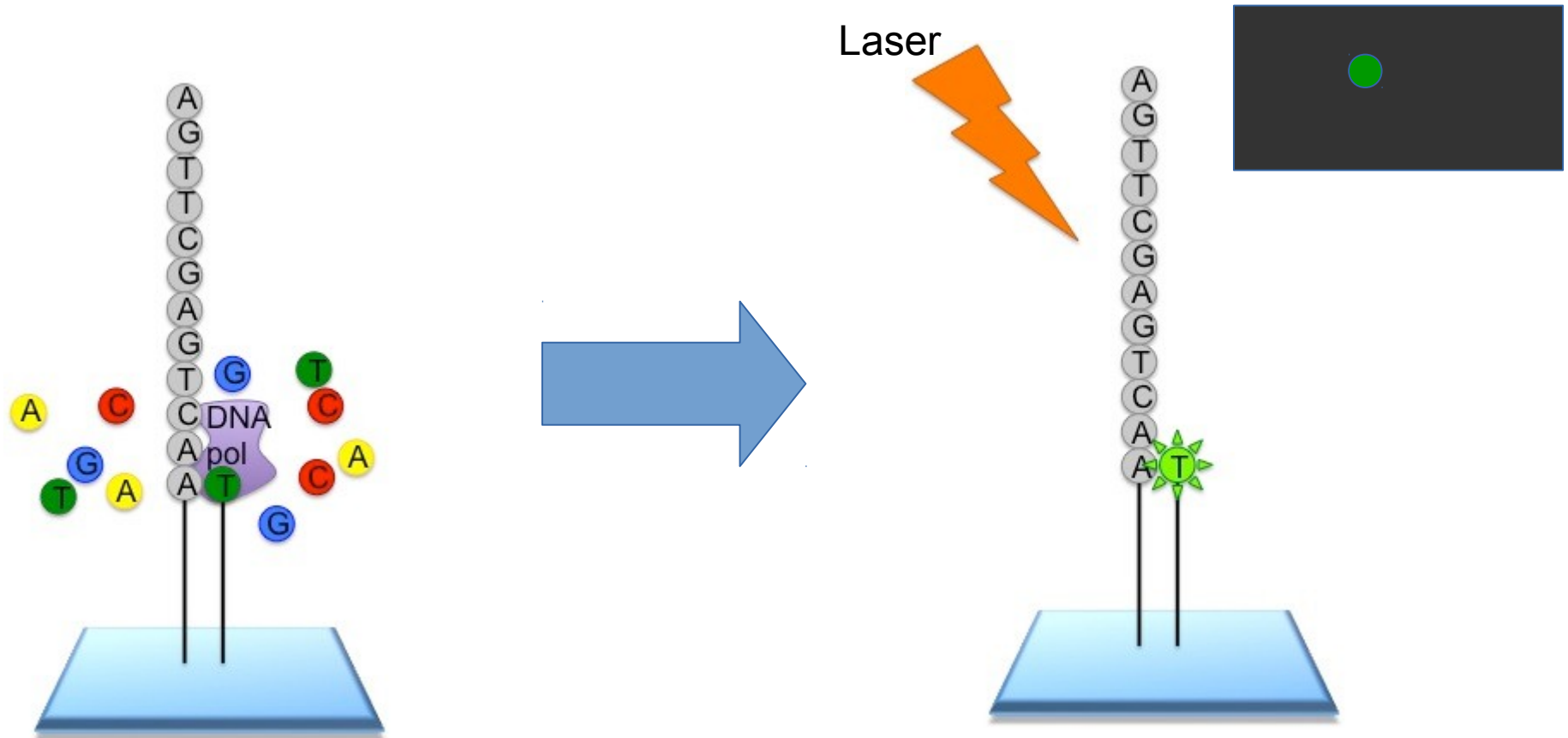
# FASTQ FORMAT

# Simplified Sequencing Run

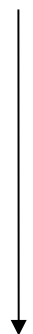
- Single Stranded DNA Template primed at one end
- Fluorescently labeled nucleotides are passed over the template
- Each nucleotide type is labeled with different fluorophore A T G C
- The complementary nucleotide at the position in the strand sequenced pairs
- The fluorophore is excited using a light (lazer)
- The corresponding color is registered by a detector
- **Software analyzes the color and intensity to make a nucleotide call**



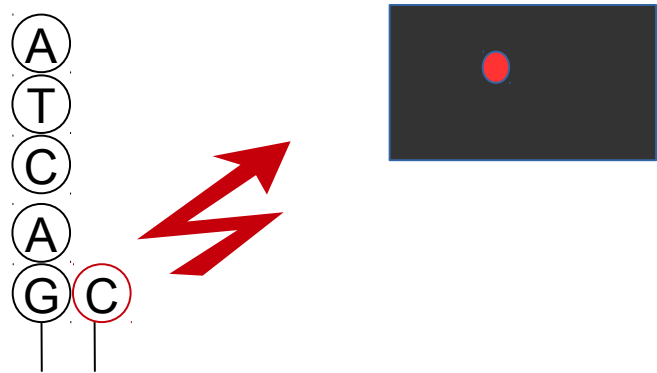
# FASTQ Format



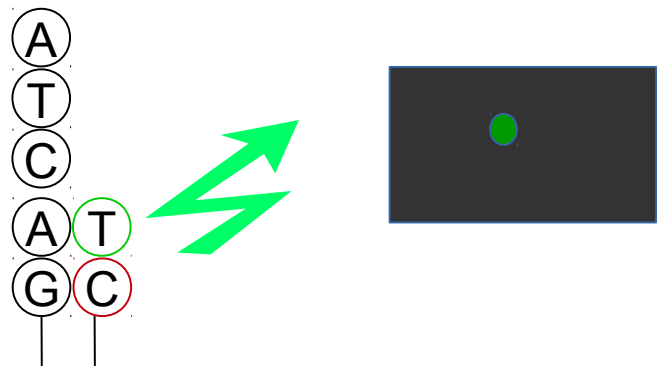
Source: <http://www.ebi.ac.uk>



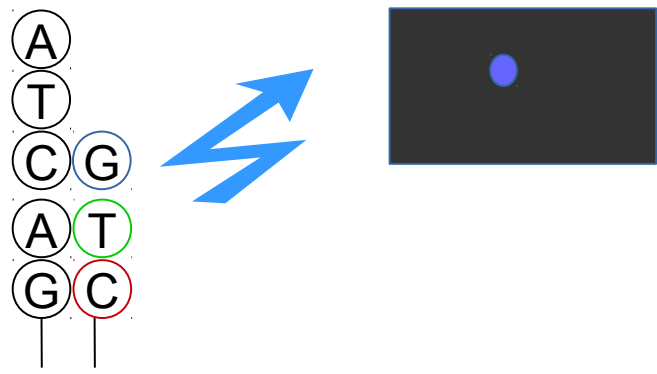
Cycle 1



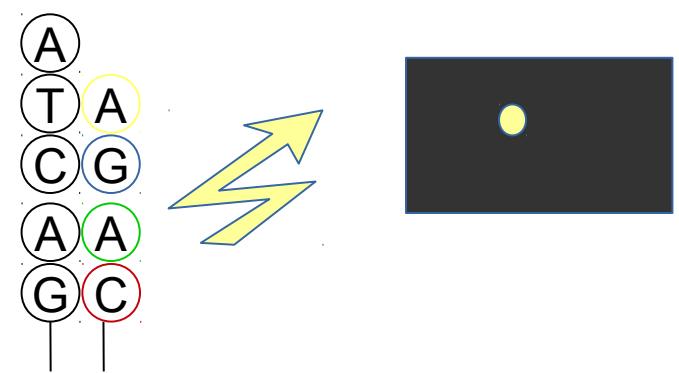
Cycle 2



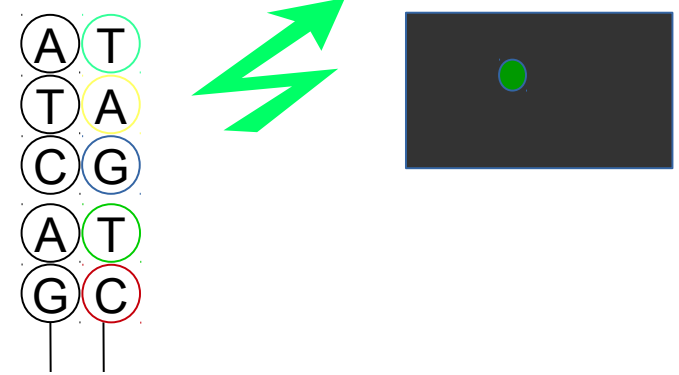
Cycle 3

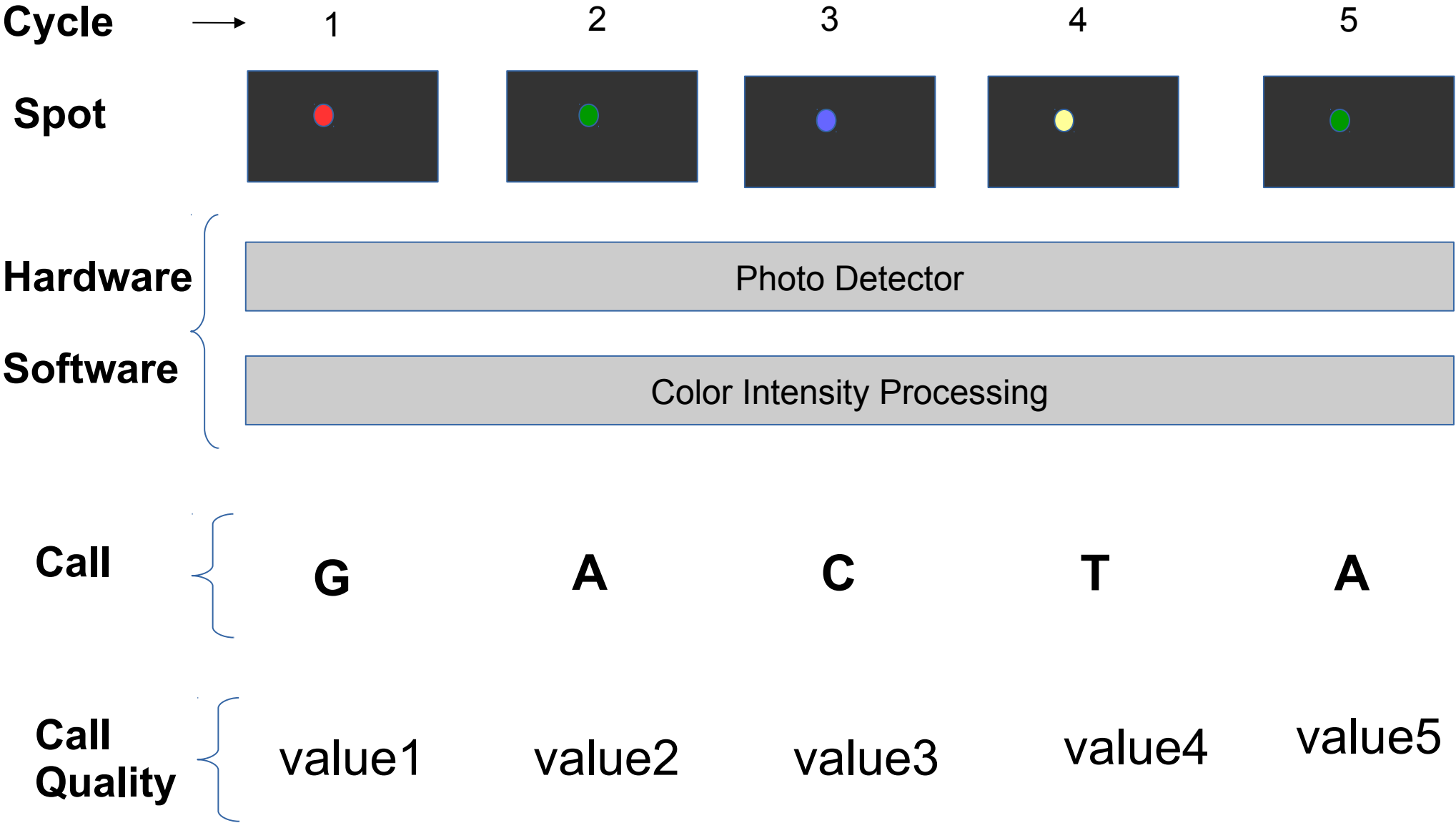


Cycle 4



Cycle 5

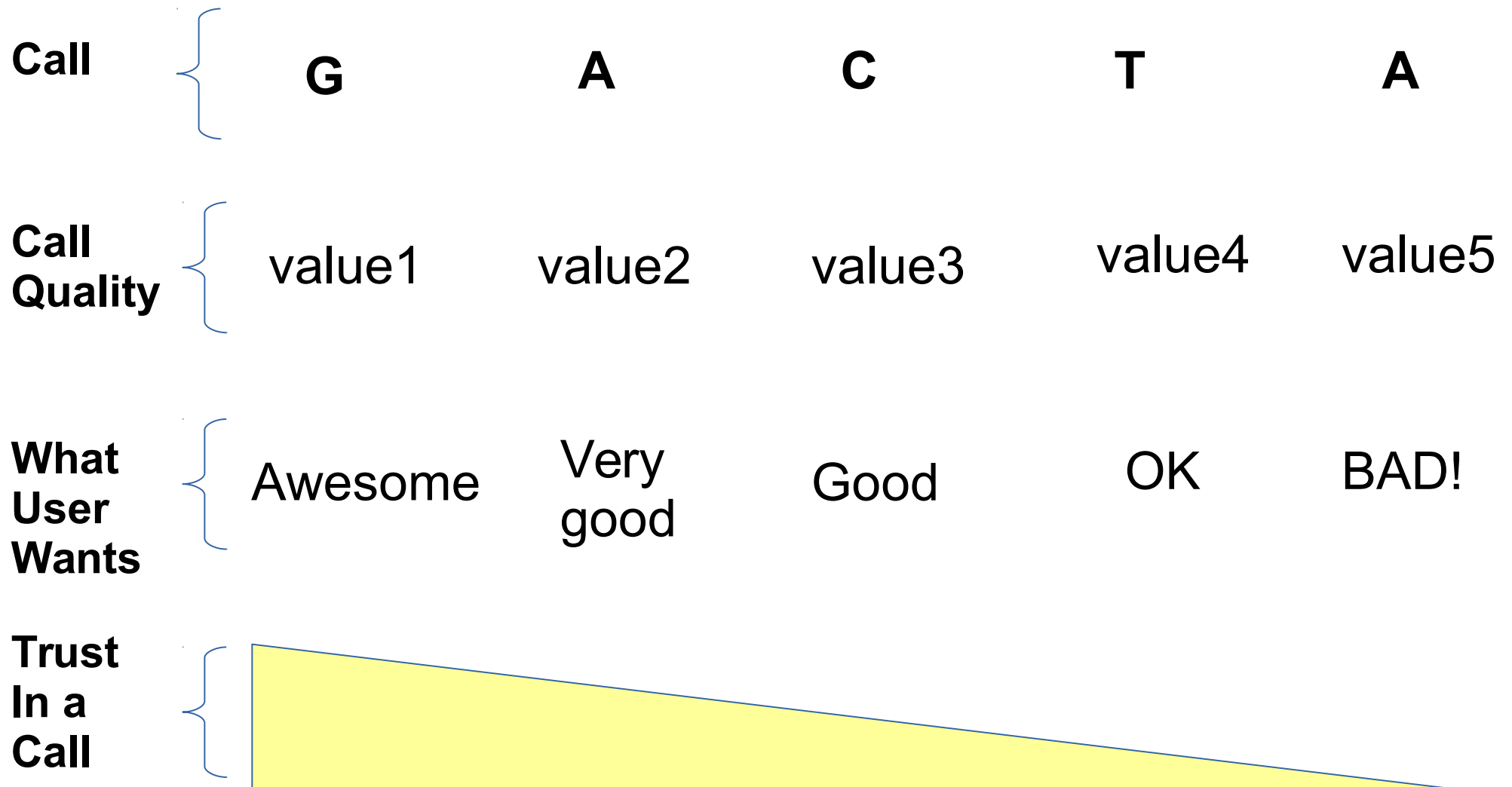




# Interpretation of call quality

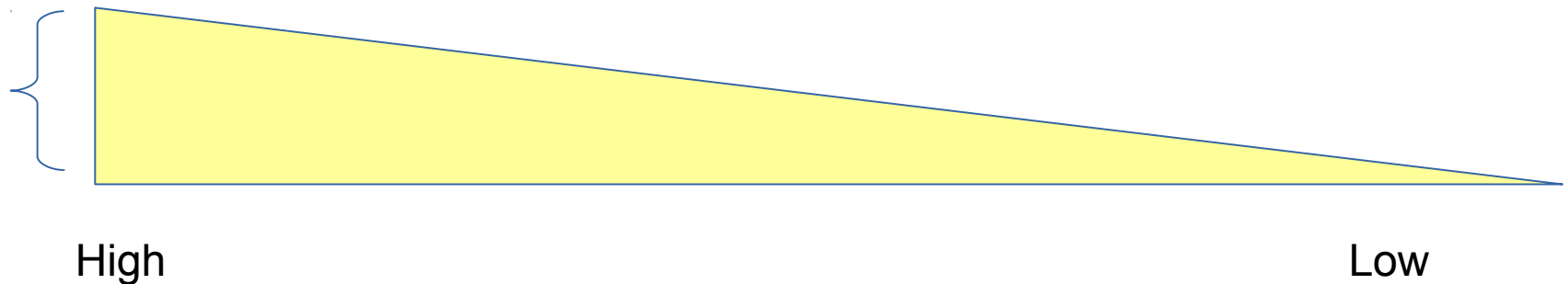
- The hardware detects and measures the color intensity
- The machine is not perfect so every call has a quality!
- The software makes an assessment of quality
- The Biologist/End\_User has to make sense of the call quality
- Call Quality will govern the number sequencing mistakes the machine makes
- Mistakes will influence downstream processing
- **Quality must be conveyed to the end user in a simple ,understandable and a quantitative framework**

# Trust in a call



# Quantifying trust in a call

Trust  
In a  
Call



## Value

- Higher the value higher the trust
- **Higher the value higher the probability that call is correct**
- Amenable to statistical and probabilistic methods
- Common across all studies/platforms/machines
- Universally accepted
- Easily encoded/printed in a file
- Easily understood by the biologist

# PHRED Scores

- Denoted by letter Q
- $Q = -10 \log_{10} P$  #P is the probability of error or the call being wrong
- $P = 10^{(-Q/10)}$

**Phred quality scores are logarithmically linked to error probabilities**

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

# Problem

- You are 100% certain that the base call is Wrong. What is the Phred score ?
- Call1 has a phred score of twice as call2. By how many times is the trust in call1 better than call2 when
  - Q of call 1 is 10
  - when Q of call 1 is 30?

$$Q = - 10 \log_{10} P$$

$$P = 10^{(-Q/10)}$$



# Phred Scores

- Easy to understand
- Fits in a probabilistic framework
- Easy to do computations on
- Easy to do compare two calls
- Covers wide range of error probabilities (Is log scaled)
- **The calling software will calculate it for you :)**

# Encoding Phred Scores

- Encoding ~ printing the phred scores along with base calls in a file.
- Nucleotides are typically available as a fasta file

# Fasta Format

- Each sequence starts with a > character followed by seq id and/or name
- Sequence starts in the next line
- Multiple sequence in each file

```
>BC1G_00449
ATGTCAGCACGGAGAGTACACATCGCAC
GCCACAGAAGGATGTTGGAACATTCGAT
ACCCGAGGGATCAAAGTTTGACAGCACA
ATTGTCAAATTCAGCATGTATCGAGCAT
>BC1G_00062
ATGGCAGCCGTCGCAGCACAAGCACCAA
CTGGTGAACAAGTTGACTTGTCAACCAT
TCCAGTCTCGCCTGAAGGGGCCAGCAAT
TCAAATGAGAACACCAAGCCAGTCGCAG
AGGGTGATCAAATTACCGTTTTCCATGA
>BC1G_00372
ATGTCGTCCGCCTCTCGACCTCGCCCCA
ACCGACGGGTGAACAGTCTGGCAAATAG
TGAGCGCATGTCCATATCTCGAAATAGC
TCTCGATCACGACCGAAAGCTATTGGGT
ATACCGATGCATATACATTCGCCCTCAG
GGTTGCATACTTGCACCACCTTCTCCAA
CCGCGACGAAAGACCAAACAATATGTCC
CCGCAGAGAGAAAGATTATTTCAAGGAA
```

# Problem

There is a fasta file test.fa in folder Lect\_02 in your home directory. How many sequences does it contain ? Print only the id of each sequence. Hint: Use the commands introduced previously.

# Encoding Phred Scores

- Encoding ~ printing the phred scores along with base calls in a file.
- Nucleotides are typically available as a fasta file
- Quality scores could be added to the fasta file?
- Cumbersome and space consuming

```
>read1
ATGC
>read1
10 20 30 40
```

# Better Solution

- Put calls and quality scores and one below another

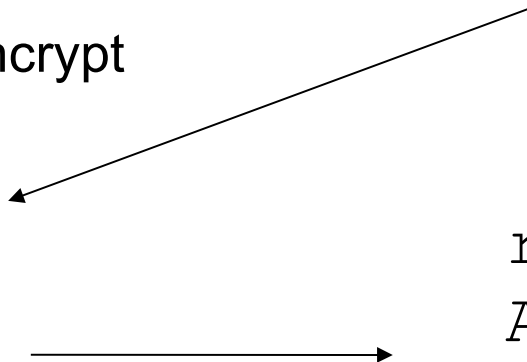
```
read1
ATGC
10 20 30 40
```

- A lot of space is wasted as space
- Most Q scores are 2 characters long
- What if 2 characters in Q score are compressed to one

Encode ~ Encrypt

```
10 = +
20 = 5
30 = ?
40 = I
```

```
read1
ATGC
+5?I
```



# ASCII codes

- American Standard Code for Information Interchange
- Total 128 characters
- Each character is assigned a number
- Easier for binary conversion
- Examples
  - A = 65
  - B = 66
  - ? = 63 ... and so on

# Problem

- Look at your ascii chart (Only columns decimal and char)
- Write the word “CAT” in ascii
- Write the word “cat” in ascii
- Create a file called 'name' in your home directory and write your name in ascii.



# Phred to ASCII

- Depends on encoding
- Sanger Encoding
  - Add 33 to the phred score and convert the number to character
  - Subtract 33 from the ascii code of the character
- Illumina encoding  $< 1.8$  add 64
- Illumina encoding  $1.8+$  add 33
- **Softwares like FASTQC will tell you the encoding**

# Problem

Convert following phred scores to sanger encoding. Hint: Sanger encoding is Phred+33

```
read1
ATGC
10 20 30 40
```

# FASTQ Format

- Store calls (ATGC ... )
- Store Phred scores (Encoded)
- Store Machine make/ID
- Store Flowcell id for each spot
- Store coordinates of the spot
- Store additional info (Seq names .... )
- Easily parsed and stored.

# FASTQ Format

- Each read is **4** lines
- Read starts with a character **@** followed by the read descriptor
- Sequence follows in the second line
- Third line is reserved for additional info
- Fourth line is the Phred score encoding
- Read pairs are typically in different files

```
@61CFUAAXX:7:1:1206:12900
ACCAATAGGGCAAAACGCATCGGNGCCAGTGAAC TTGGGGTCATCCANANACGCAGTGTCTAACATCCC
+
BCCCCCACACC?CCBCCCC?A??%>>>;>9;7%9%/%%%/0%%%/0%/%/0/;7;9?BB9ABB<>BB
```

# FASTQ Format

```
@HWUSI-EAS100R:6:73:941:1973#0/1
```

<b>HWUSI-EAS100R</b>	the unique instrument name
<b>6</b>	flowcell lane
<b>73</b>	tile number within the flowcell lane
<b>941</b>	'x'-coordinate of the cluster within the tile
<b>1973</b>	'y'-coordinate of the cluster within the tile
<b>#0</b>	index number for a multiplexed sample (0 for no indexing)
<b>/1</b>	the member of a pair, /1 or /2 ( <i>paired-end or mate-pair reads only</i> )

Source: Wikipedia

# FastQ Format

With Casava 1.8 the format of the '@' line has changed:

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

<b>EAS139</b>	the unique instrument name
<b>136</b>	the run id
<b>FC706VJ</b>	the flowcell id
<b>2</b>	flowcell lane
<b>2104</b>	tile number within the flowcell lane
<b>15343</b>	'x'-coordinate of the cluster within the tile
<b>197393</b>	'y'-coordinate of the cluster within the tile
<b>1</b>	the member of a pair, 1 or 2 ( <i>paired-end or mate-pair reads only</i> )
<b>Y</b>	Y if the read is filtered, N otherwise
<b>18</b>	0 when none of the control bits are on, otherwise it is an even number
<b>ATCACG</b>	index sequence

Source: Wikipedia

# Problem

In the folder Lect\_02 in your home dir there are two fastq files named R1.fastq and R2.fastq. Find

- How many reads are there in each file ?
- What is the name of the machine ?
- How many flowcells the data comes from ?
- Which file contains the first read in pair and which one the second.

# Problem

In the folder Lect\_02 there is a file called 'test.fastq'. Find

- How many reads ?
- How many flowcells ?
- Is there a problem with one flowcell? If yes which one ?