RNA-Seq expression analysis in Galaxy

From A to Z

Y. Hoogstrate[1,2] S. Hiltemann[1,2]

[1]Department of Bioinformatics & Department of Urology
ErasmusMC, Rotterdam [2]CTMM Translational Research IT (TraIT)

Galaxy Community Conference 2015, Norwich

Overview

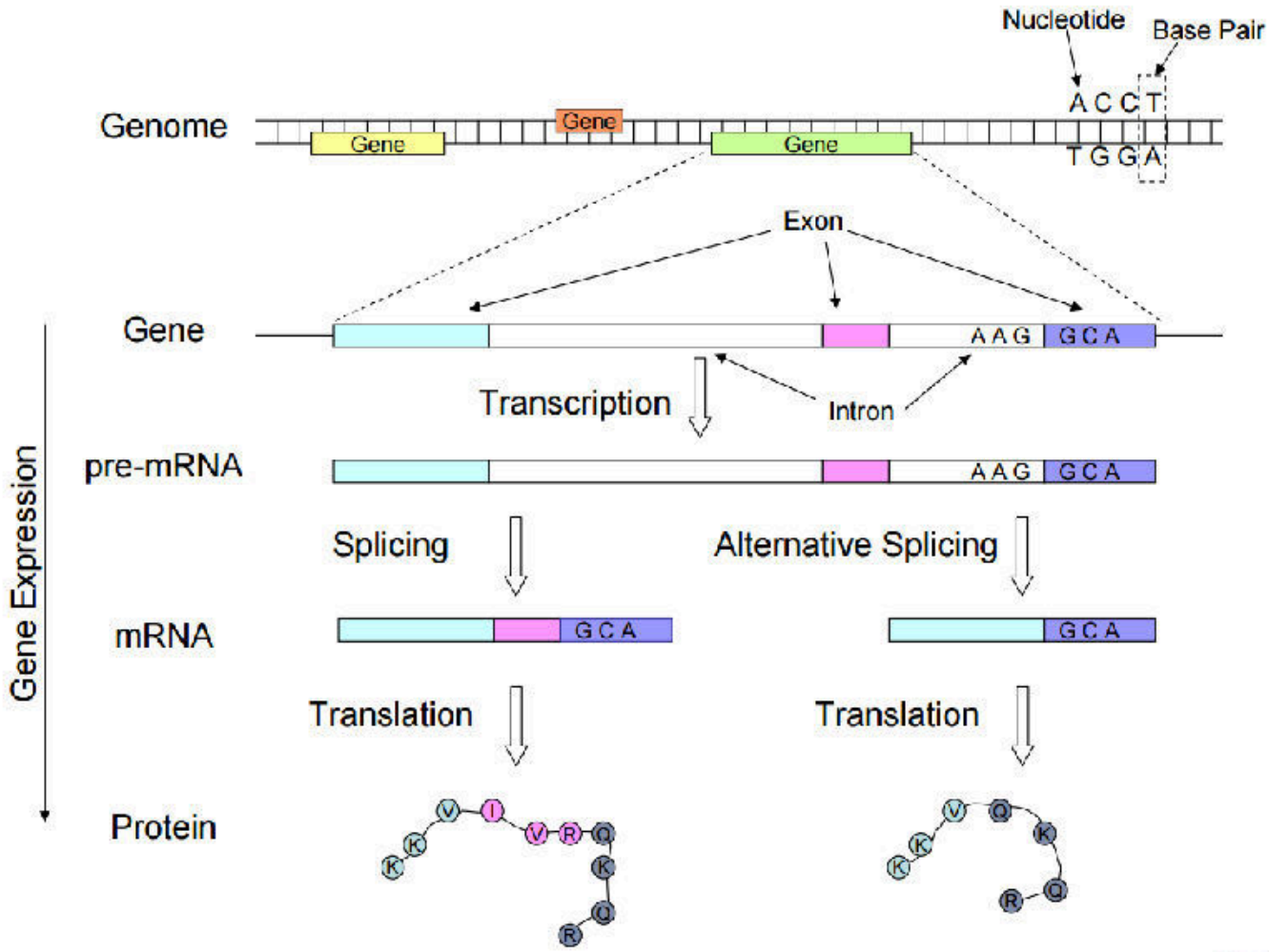
Introduction
- RNA-Seq

Raw data to alignment
- Raw data
  - Data acquisition
  - FASTQ
  - QA/QC
- Alignment
- Measure expression

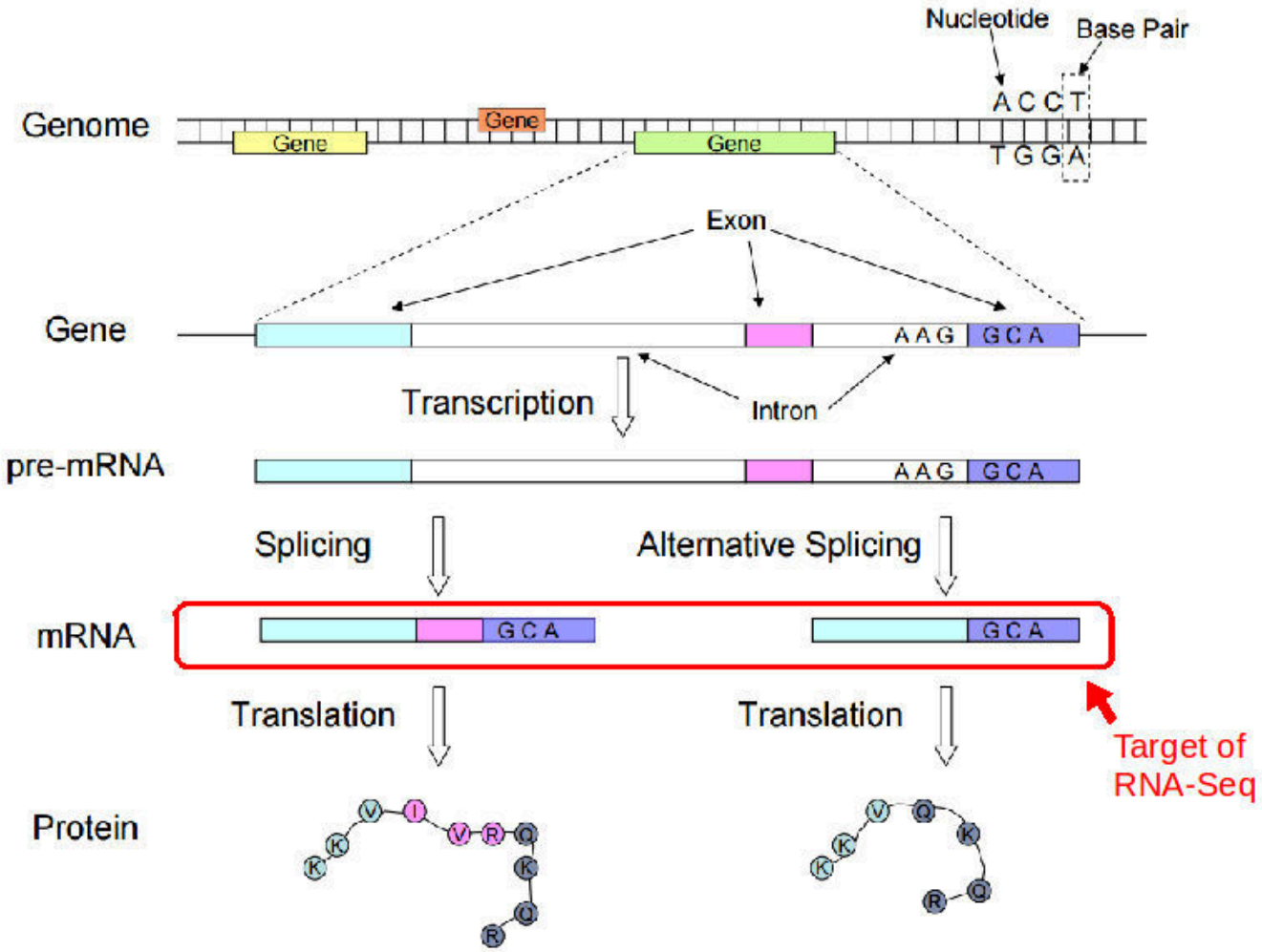
Differential Gene Expression (DGE) analysis
- Count data
- Expression and design matrix
- Replicates

Wrap up

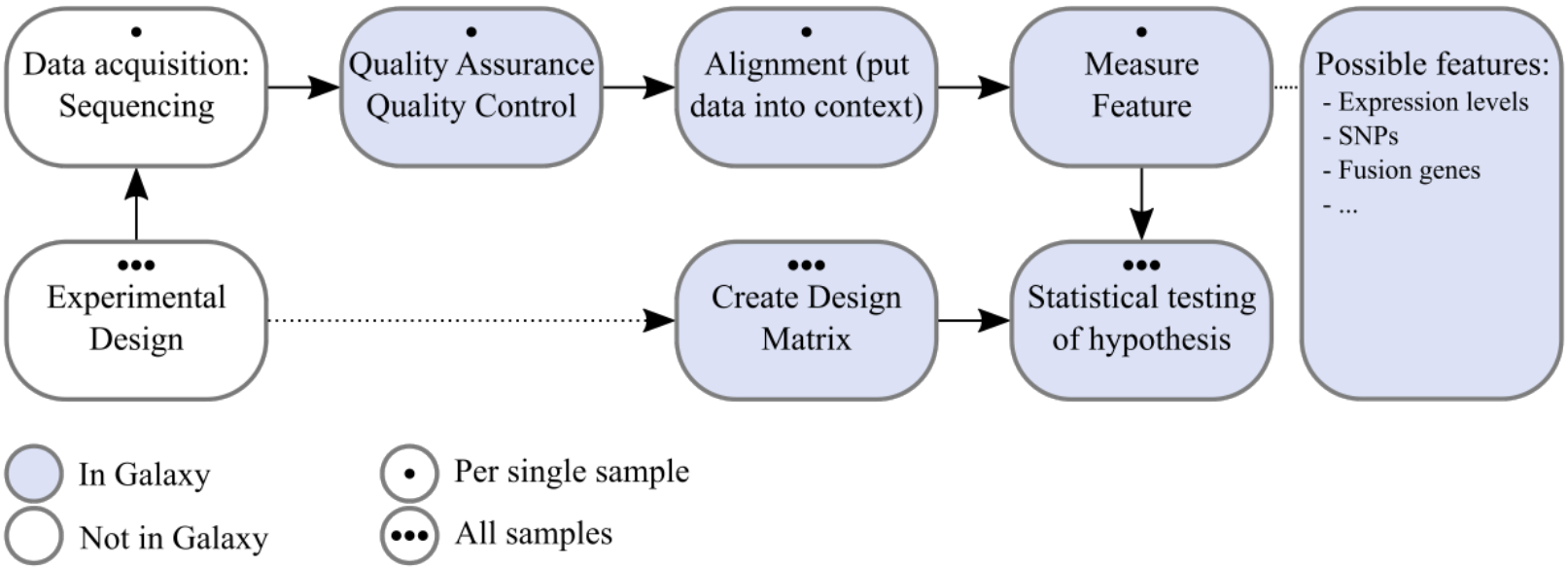
Central dogma



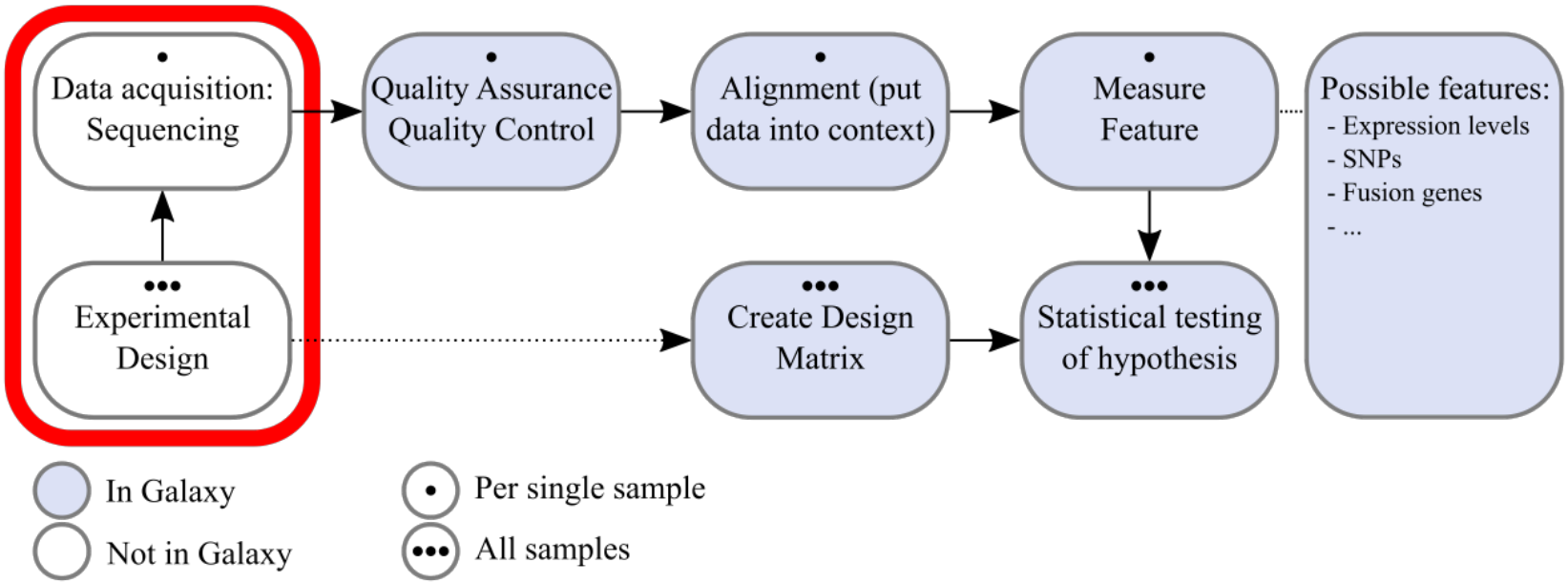
Central dogma



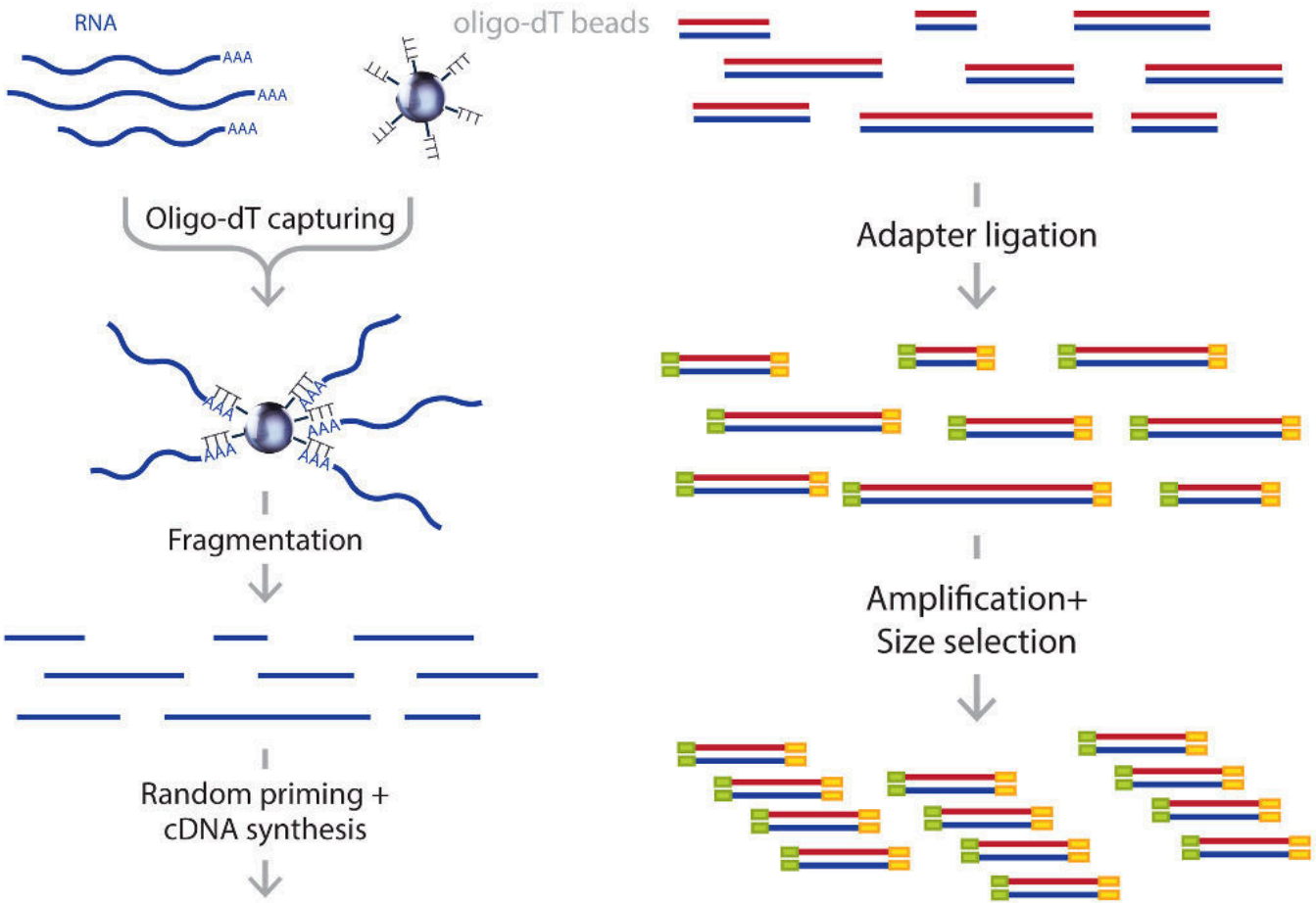
RNA-Seq experiment workflow



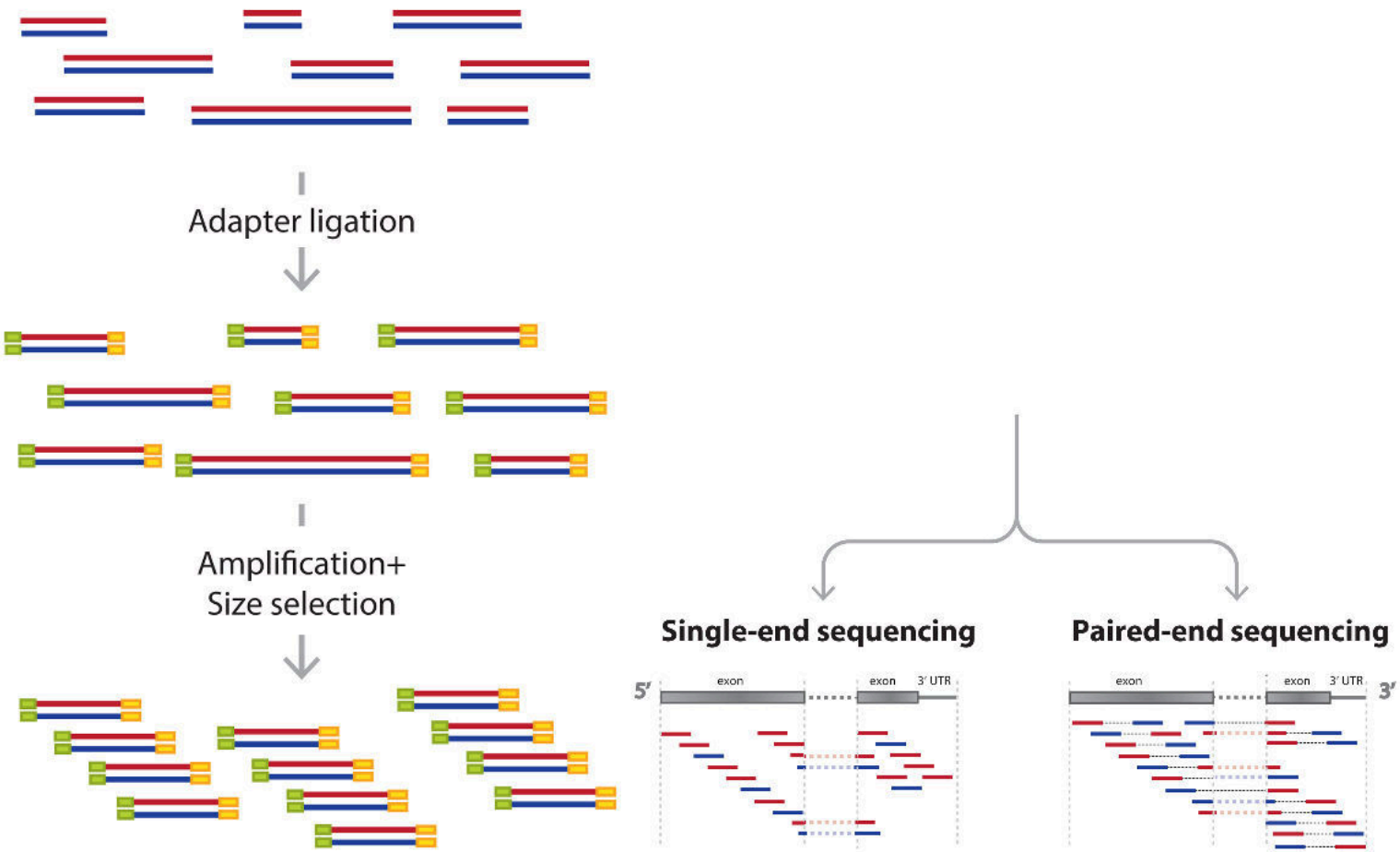
RNA-Seq experiment workflow



Library preparation



Library preparation



FASTQ file format

- Paired end data
  - Two corresponding files (often "R1", "R2")

Control_L7.D701_R1.fastq

```
1  @HISEQ:130607:C257AACXX:7:1101:1571:1959 1:N:0: ATTACTCG
2  GCCTTTTGTGACTGGCTTTTTTCACTCAGCATAATGTTTGCTATAGAATN
3  +
4  @@8DDDDD:C:ADBE:C:CFHBF99??FG>4CCF???BDF?:BAFF<?B#
5  @HISEQ:130607:C257AACXX:7:1101:1588:1971 1:N:0: ATTACTCG
6  CCAGGTCCATGGCTAATCATCATTTTGATGTTCTATTTCAAAGACAACAN
7  +
8  CC@FFFFFHHFBFGGEIGIIBHHFHHIIIJEHIIAGIJFIIJGIIJGIJ#
9  @HISEQ:130607:C257AACXX:7:1101:1957:1965 1:N:0: ATTACTCG
10 CATTACGTTATTGAATTCCACACACATCCTATGAGGTTATTATCCCCATN
11 +
12 @@@DFFFDHHFHHEIIGIIIIIIIIIGGIIGGGIIIIIIIIIGIHIIII#
13 @HISEQ:130607:C257AACXX:7:1101:2118:1955 1:N:0: ATTACTCG
14 GATCGGAAGAGCACACGTCTGAACTCCAGTCACATTACTCGATCTCGTAN
15 +
16 CCCFFFFFGHGHHJJJJJJJJHJJJJJJJHIHIJIJJJJJJIGGIIIIJ#
```

Control_L7.D701_R2.fastq

```
1  @HISEQ:130607:C257AACXX:7:1101:1571:1959 3:N:0: ATTACTCG
2  GCATTATGTCCAGTGGAAATTGAGGCTGTTAGCAATAAAAACAATTAAGG
3  +
4  @@<=BBDBAFH4DEEB<<<CDC?C:?FHCF+<CE9C9<?FCDFHEDCG@C
5  @HISEQ:130607:C257AACXX:7:1101:1588:1971 3:N:0: ATTACTCG
6  AAAATTTTTGTTTTACTTTTAGCTTTGTTTGTGTAAATTGTATAAGTATA
7  +
8  @@@DFFFFHDFHHJEHBGEIGIGIJIJHIIIIIGIIJIGIFDGHGGGIIB
9  @HISEQ:130607:C257AACXX:7:1101:1957:1965 3:N:0: ATTACTCG
10 CAAGCTTGGCTCCTTGCACTACCTGGAGGTGTAAGCTTTGGCAAGTCGCT
11 +
12 @@@?BDD>DHAFHDH@FHHGFGICAG@DH)?1CC>CCFHHGADBEGGGH6
13 @HISEQ:130607:C257AACXX:7:1101:2118:1955 3:N:0: ATTACTCG
14 GGGAAGGGGAAGGGGGGGGGGGGGGGGAGGGGGGGGGGAGAGGTGGGAT
15 +
16 ################################################
```

Sequence data raw format: FASTQ

- Sequence is given per char
  - Two corresponding files (often "R1", "R2")
  - Pairs linked by position in file (and name)

Control_L7.D701_R1.fastq

```
1  @HISEQ:130607:C257AACXX:7:1101:1571:1959 1:N:0: ATTACTCG
2  GCCTTTTGTGACTGGCTTTTTTCACTCAGCATAATGTTTGCTATAGAATN
3  +
4  @@8DDDDD:C:ADBE:C:CFHBF99??FG>4CCF???BDF?:BAFF<?B#
5  @HISEQ:130607:C257AACXX:7:1101:1588:1971 1:N:0: ATTACTCG
6  CCAGGTCCATGGCTAATCATCATTTTGATGTTCTATTTCAAAGACAACAN
7  +
8  CC@FFFFFHHFBFGGEIGIIBHHFHHIIIJEHIIAGIJFIIJGIIJGIJ#
9  @HISEQ:130607:C257AACXX:7:1101:1957:1965 1:N:0: ATTACTCG
10 CATTACGTTATTGAATTCCACACACATCCTATGAGGTTATTATCCCCATN
11 +
12 @@@DFFFDHHFHHEIIGIIIIIIIIIGGIIGGGIIIIIIIIIGIHIIII#
13 @HISEQ:130607:C257AACXX:7:1101:2118:1955 1:N:0: ATTACTCG
14 GATCGGAAGAGCACACGTCTGAACTCCAGTCACATTACTCGATCTCGTAN
15 +
16 CCCFFFFFGHGHHJJJJJJJJJHJJJJJJJJHIHIJIJJJJJJIGGIIIIJ#
```

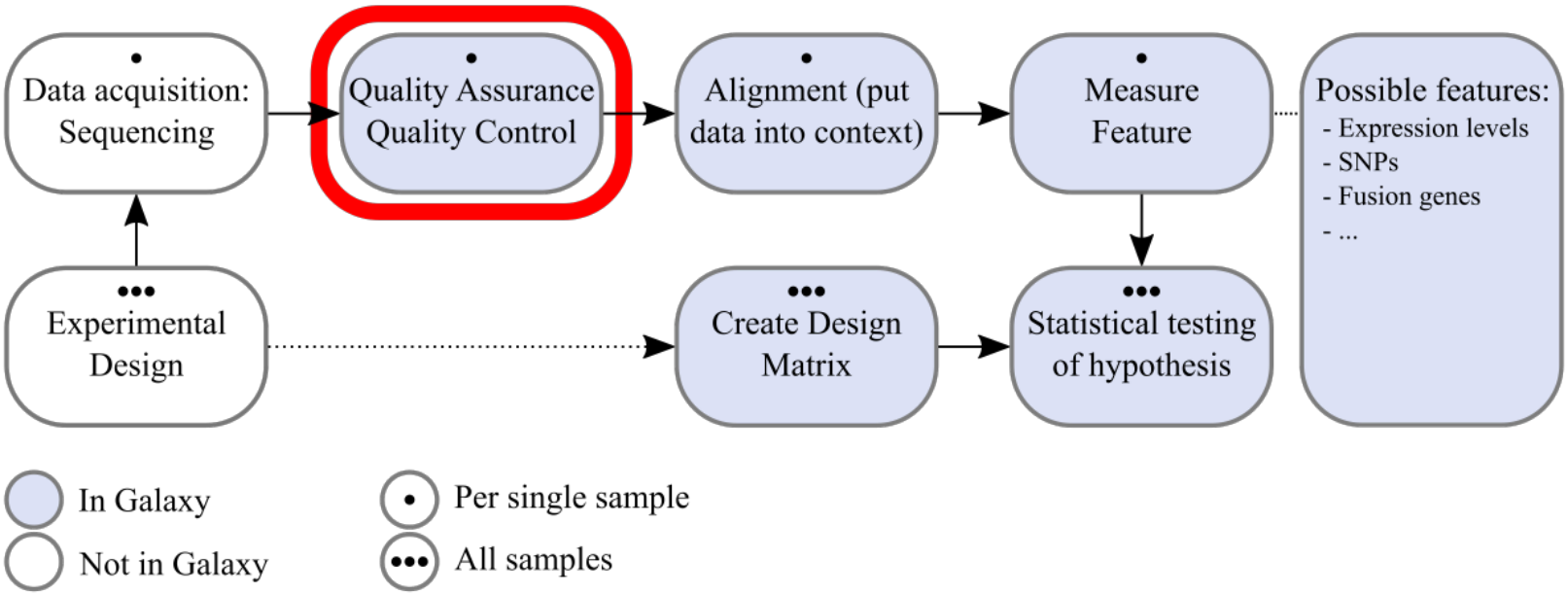
Control_L7.D701_R2.fastq

```
1  @HISEQ:130607:C257AACXX:7:1101:1571:1959 3:N:0: ATTACTCG
2  GCATTATGTCCAGTGGAAATTGAGGCTGTTAGCAATAAAAACAATTAAGG
3  +
4  @@<=BBDBAFH4DEEB<<<CDC?C:?FHCF+<CE9C9<?FCDFHEDCG@C
5  @HISEQ:130607:C257AACXX:7:1101:1588:1971 3:N:0: ATTACTCG
6  AAAATTTTTGTTTTACTTTTAGCTTTGTTTGTGTAAATTGTATAAGTATA
7  +
8  @@@DFFFFHDFHHJEHBGEIGIGIJIJHIIIIIGIIJIGIFDGHGGGIIB
9  @HISEQ:130607:C257AACXX:7:1101:1957:1965 3:N:0: ATTACTCG
10 CAAGCTTGGCTCCTTGCACTACCTGGAGGTGTAAGCTTTGGCAAGTCGCT
11 +
12 @@@?BDD>DHAFHDH@FHHGFGICAG@DH)?1CC>CCFHHGADBEGGGH6
13 @HISEQ:130607:C257AACXX:7:1101:2118:1955 3:N:0: ATTACTCG
14 GGGAAGGGGAAGGGGGGGGGGGGGGGAGGGGGGGGGGGAGAGGTGGGAT
15 +
16 ################################################
```

Sequence data raw format: FASTQ

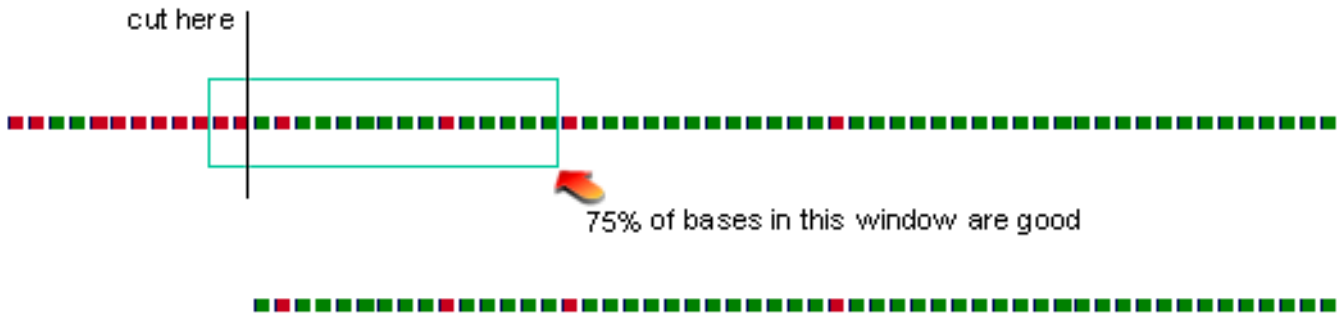
- Sequence is given per char
  - N means *sequencer doesn't know*
- Quality is encoded as a char
  - reflects probability of being called correctly
- Different encodings
  - `http://en.wikipedia.org/wiki/FASTQ_format#Encoding`
- RNA-Seq: data usually unstranded, but stranded does exist

RNA-Seq experiment workflow

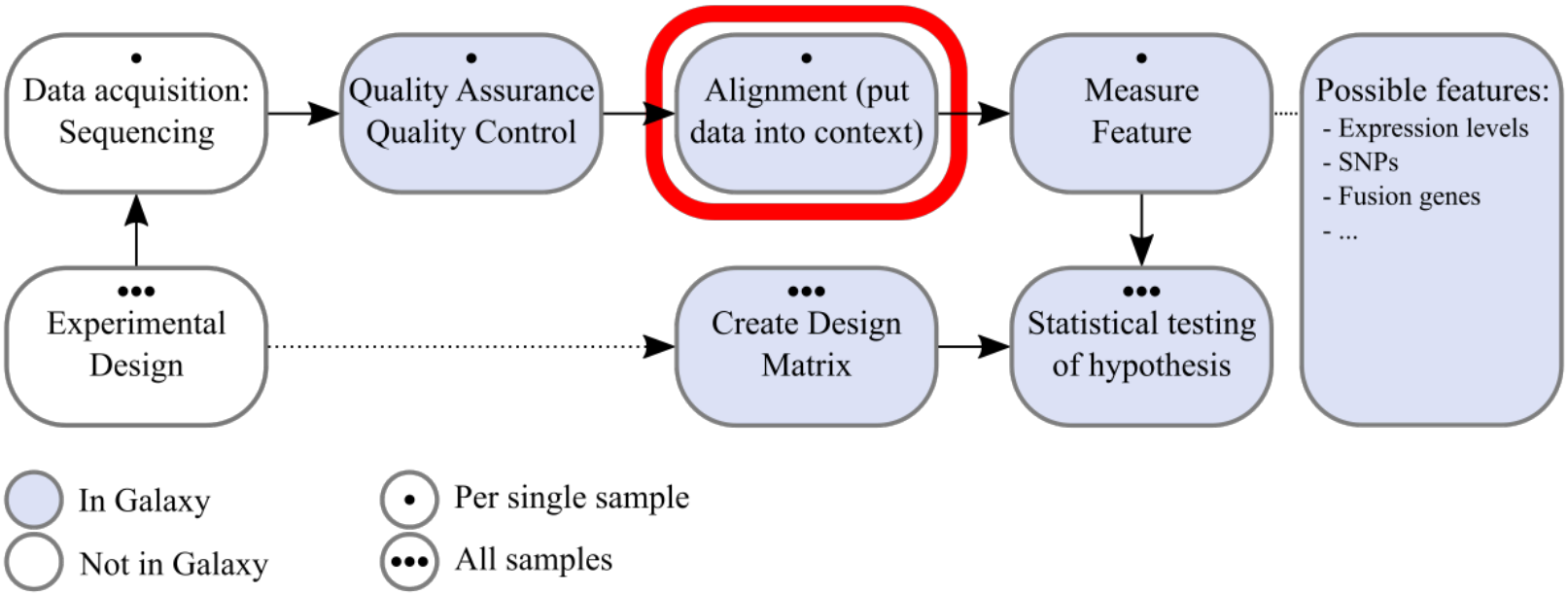


Quality assurance & quality control

- Adapter contamination
- Trim low quality bases from the ends
  - Be aware: in paired end data reads are linked by position in file
  - Proceed with trimmed reads



RNA-Seq experiment workflow

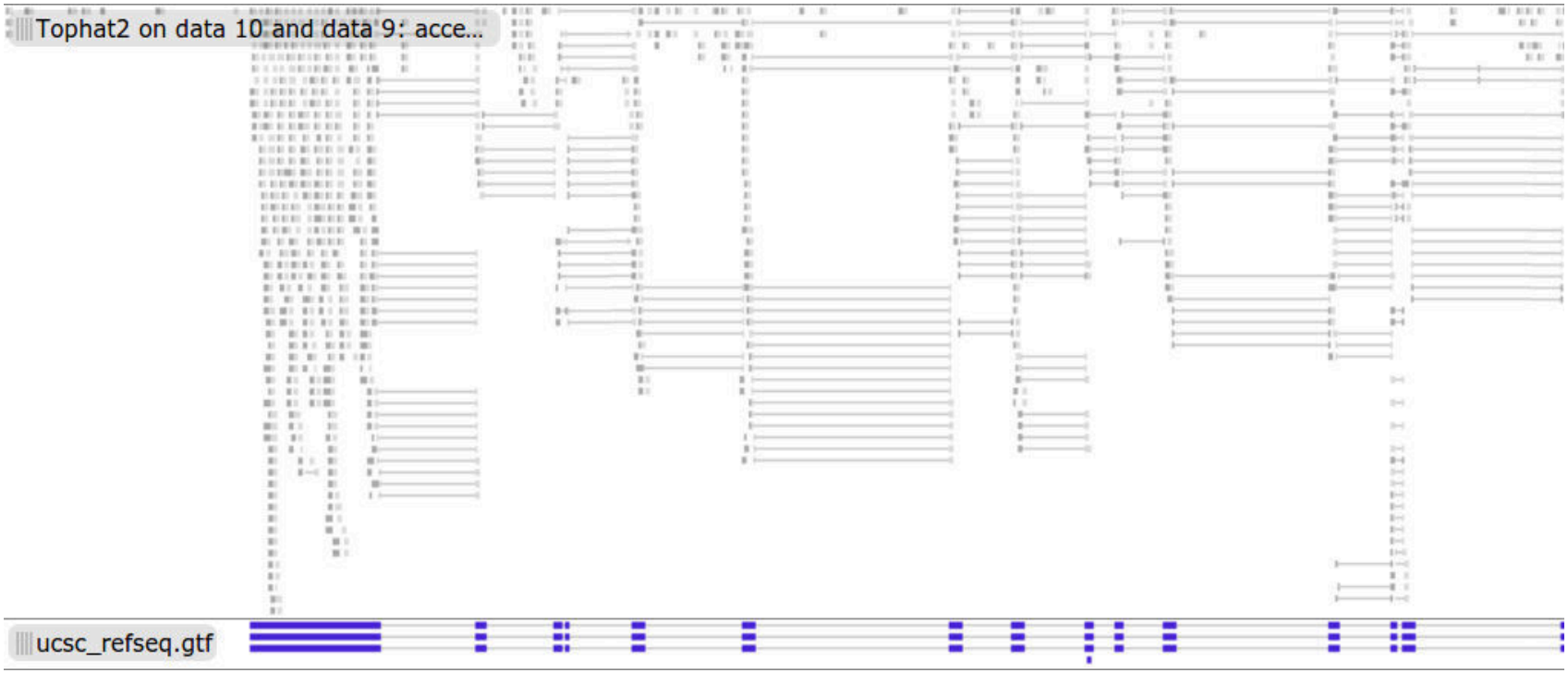


Single Nucleotide Polymorphisms in RNA-Seq

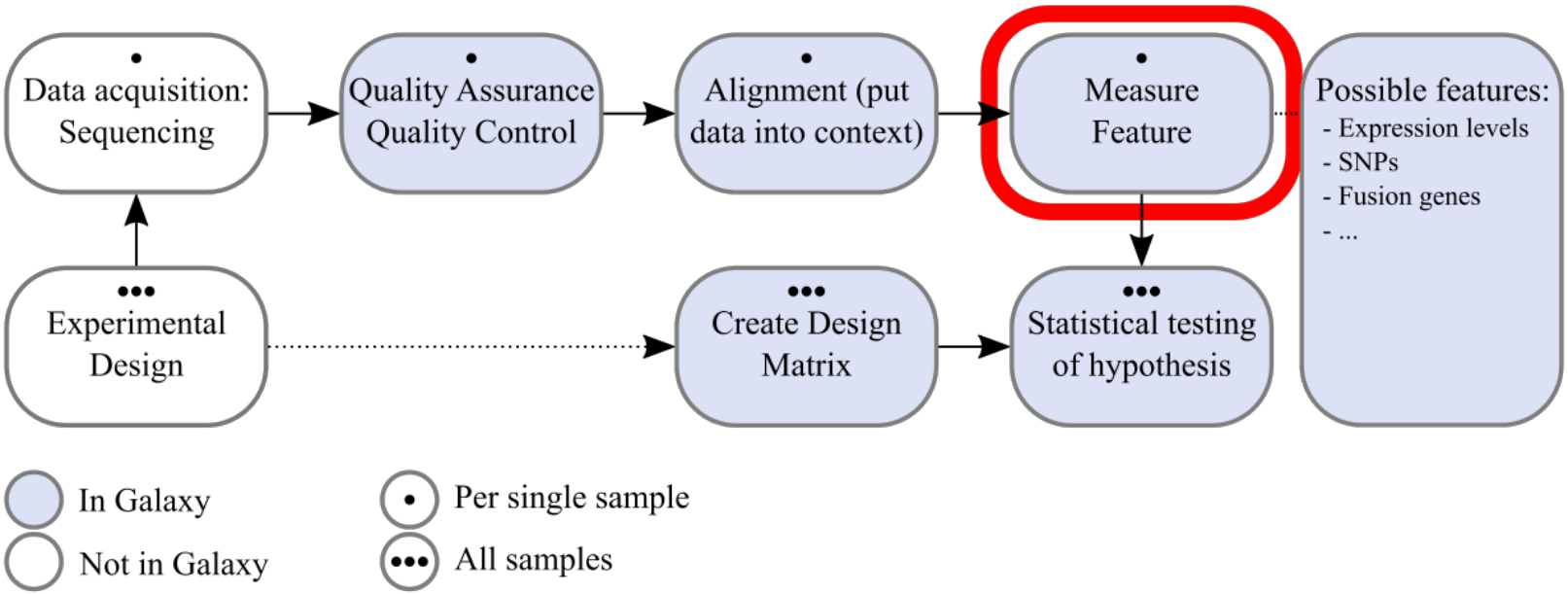
Covered examples during hands-on

- Biological interpretation: map reads to reference genome
- mRNA: spliced
  - Aligning: low/no penalty for gaps near introns
- mRNA: expressed
  - Only reads in expressed regions
- Requires specialized (slower) aligners

Typical RNA-Seq alignment



RNA-Seq experiment workflow

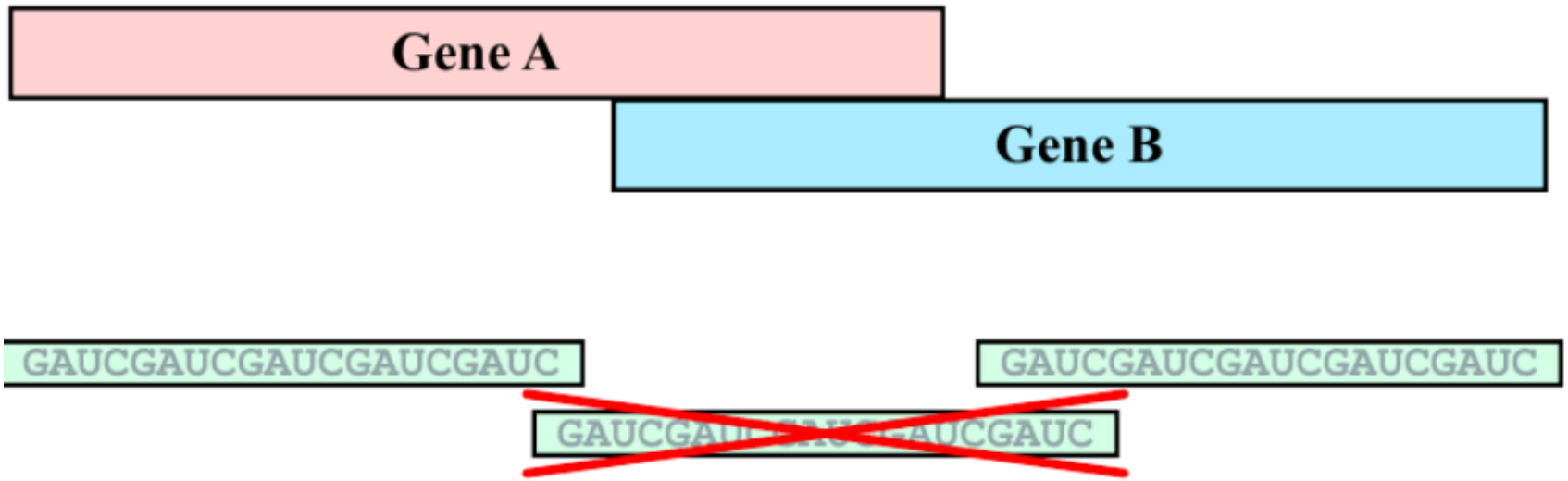


What information does RNA-Seq contain?

- Expression levels
  - Gene level
  - Transcript level Measure expression levels in RNA-Seq data(splice variants)
- Variants
  - SNPs, SNVs
  - Structural variants: fusion gene, conjoined genes, deletions
- Non-reference transcripts
  - Novel genes
  - Viral/bacterial RNA
  - Insertions
- Theoretically
  - Allele specific expression
  - RNA-editting
  - Intron retention time, RNA-stability

Measure expression levels in RNA-Seq

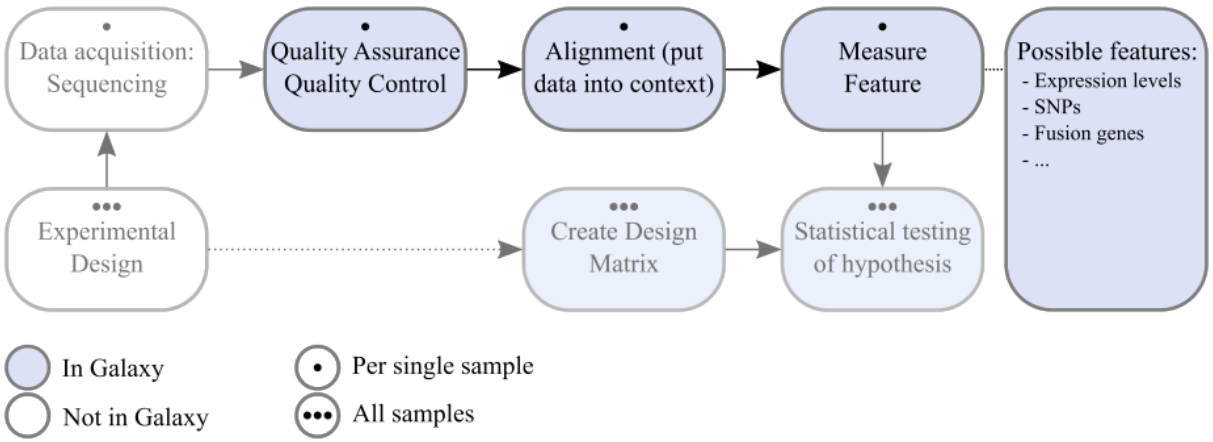
- Basic principle: count aligned reads in alignment
- Statistical independence (ensure a read belongs to **only** that gene)
  - Skip reads aligned to multiple places ('multi-mappers')
  - Skip overlapping gene annotations
  - Only look in exons



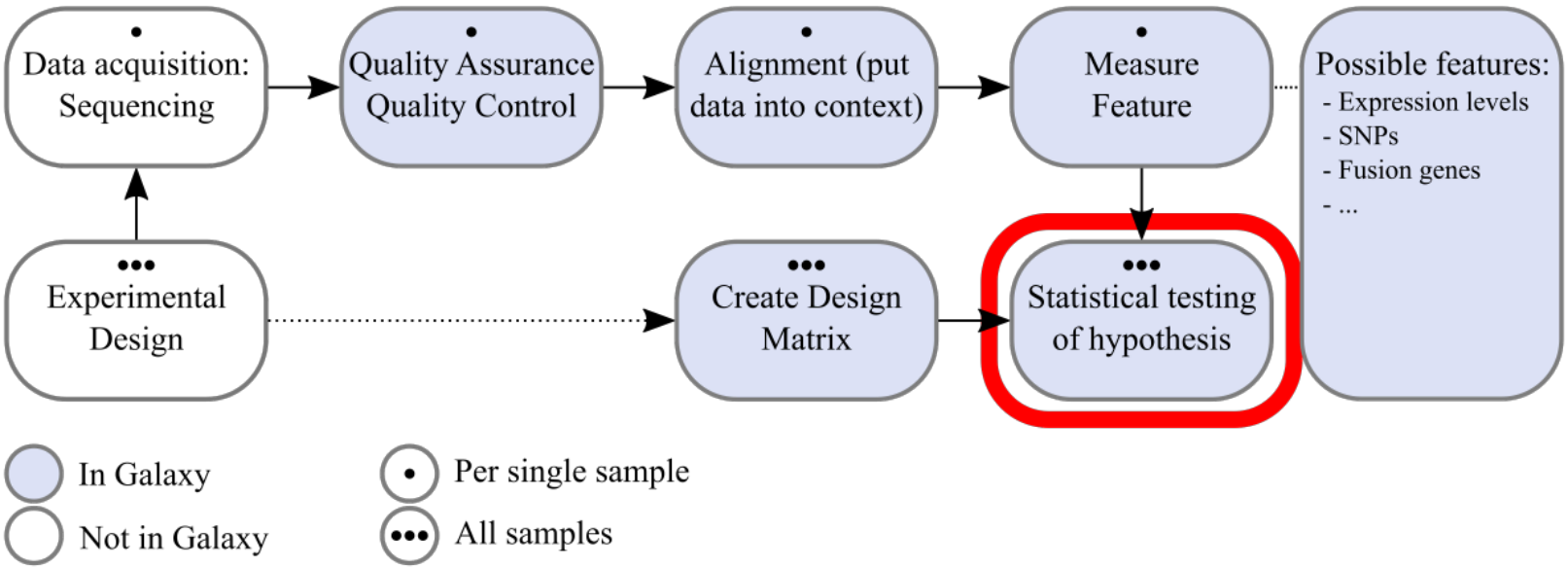
Practical part 1

From raw data to expression levels

- Artificial small dataset
- Start galaxy!

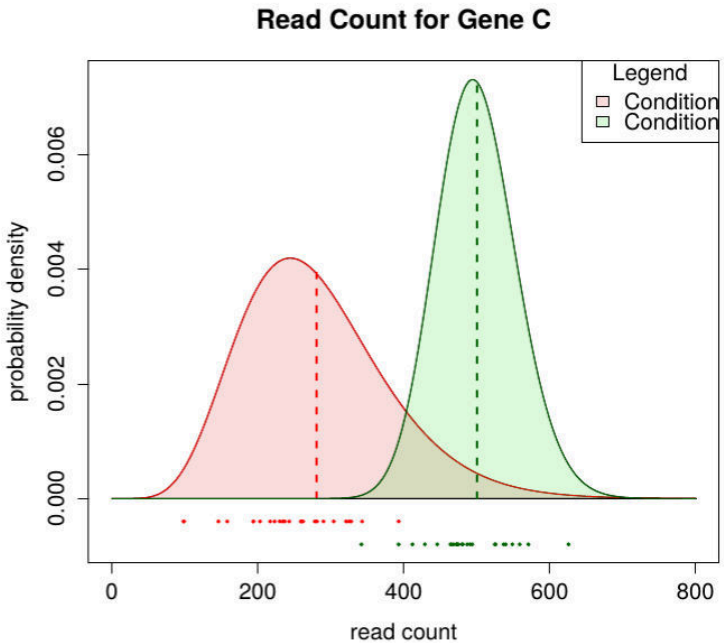


RNA-Seq experiment workflow



Differential gene expression

- RNA-Seq: count-data
- Not normal-distributed, negative binomial
  - Read counts of 1.45 and -42 don't exist!
  - Special tests for count data

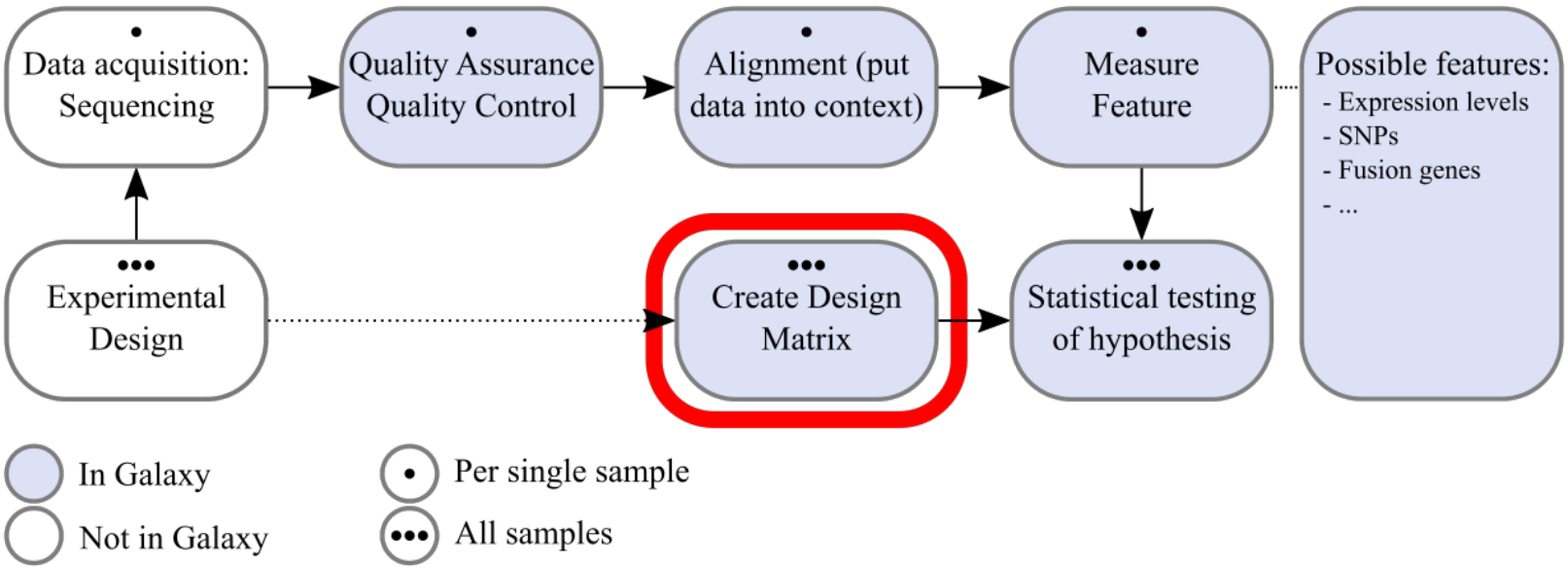


Expression matrix

- Rows: one candidate gene per row
- Columns: read counts, per gene, per sample

	Sample-1	Sample-2	Sample-3	Sample-4	Sample-5	Sample-6	Sample-7	Sample-8
Gene-1	112	4	10	21	8	16	584	59
Gene-2	173	10	39	38	12	24	949	157
Gene-3	152	123	177	155	113	355	536	673
Gene-4	46	36	132	49	52	124	206	366
Gene-5	51	19	40	27	20	51	101	282
Gene-6	23	28	34	13	7	12	47	128
Gene-7	48	105	125	56	49	68	254	408
Gene-22,000	38	1155	68	60	10	43	155	381

RNA-Seq experiment workflow



Design matrix

- Rows: one sample per row
- Columns: mutually exclusive conditions

	Condition
Sample-1	tumor
Sample-2	tumor
Sample-3	tumor
Sample-4	tumor
Sample-5	normal
Sample-6	normal
Sample-7	normal
Sample-8	normal

Biological question

- Biological question (difference between conditions)

	Condition
Sample-1	tumor
Sample-2	tumor
Sample-3	tumor
Sample-4	tumor
Sample-5	normal
Sample-6	normal
Sample-7	normal
Sample-8	normal

Design matrix

	Sample-1	Sample-2	Sample-3	Sample-4	Sample-5	Sample-6	Sample-7	Sample-8
Gene-1	112	4	10	21	8	16	584	59
Gene-2	173	10	39	38	12	24	949	157
Gene-3	152	123	177	155	113	355	536	673
Gene-4	46	36	132	49	52	124	206	366
Gene-5	51	19	40	27	20	51	101	282
Gene-6	23	28	34	13	7	12	47	128
Gene-7	48	105	125	56	49	68	254	408
Gene-22,000	38	1155	68	60	10	43	155	381

Expression matrix

Biological replicates

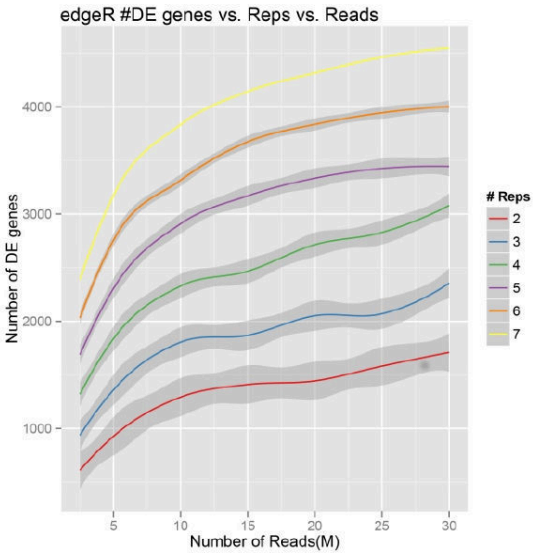
- 2 class problem (tumor normal)
  - Scenario 1
    - Sequence 100M reads
    - 3 replicates
    - 100M * 3 = 300M reads
  - Scenario 2
    - Sequence 10M reads
    - 30 replicates
    - 10M * 30 = 300M reads
- Question: "*more sequence or more replication?*"

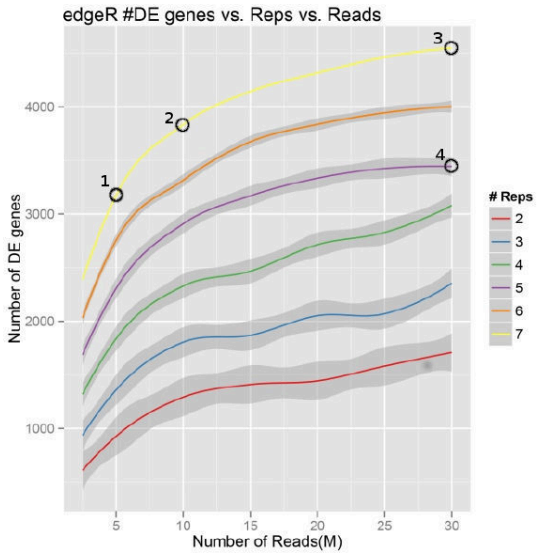
Practical part 2: more sequence or more replication?

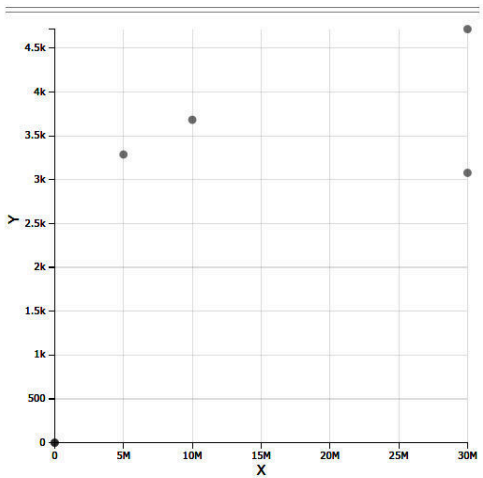
- `http://www.ncbi.nlm.nih.gov/pubmed/24319002`
  - MCF7 cell line
  - 2 conditions: treated and untreated with hormone
  - n^o DE genes reflects statistical power
- Practical: complete table:

Replicates	Seq. depth (million)	DE genes
0	0	0
7	5	**?**
7	10	**?**
7	30	**?**
5	30	**?**

Wrap up

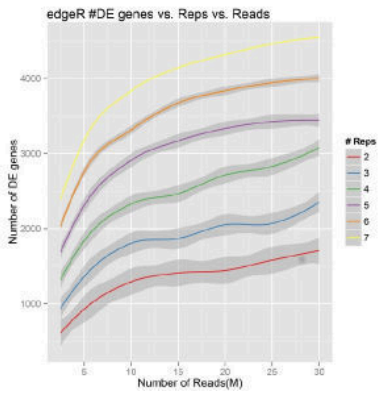

edgeR #DE genes vs. Reps vs. Reads
Number of DE genes
Number of Reads(M)
Reps
2
3
4
5
6
7

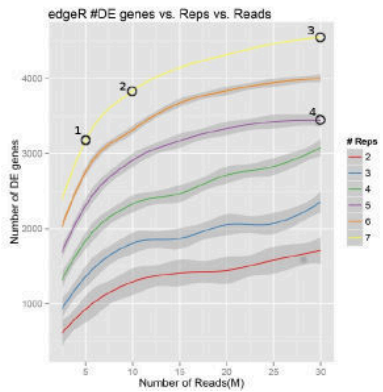

edgeR #DE genes vs. Reps vs. Reads
Number of DE genes
Number of Reads(M)
Reps
2
3
4
5
6
7
1
2
3
4

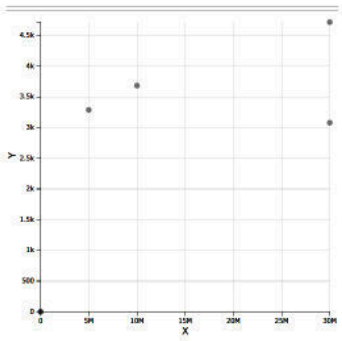

Y
X

Wrap up

- `http://bioinformatics.oxfordjournals.org/content/30/3/301.long`
  - "*In the human cell line MCF7, adding more sequencing depth after 10M reads gives diminishing returns on power to detect DE genes*"
  - Using 5 or 7 replicates still makes a difference







Useful links

- https://usegalaxy.org/u/jeremy/p/galaxy-rna-seq-analysis-exercise
- https://testtoolshed.g2.bx.psu.edu/view/yhoogstrate/edger_with_design_matrix
- http://bioinformatics.oxfordjournals.org/content/30/3/301.long
- https://bioinf-galaxian.erasmusmc.nl/galaxy/
- https://github.com/ErasmusMC-Bioinformatics/galaxy-tools
- http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0103207
- http://www.bioinformatics.babraham.ac.uk/training/RNA-Seq_analysis_course.pptx