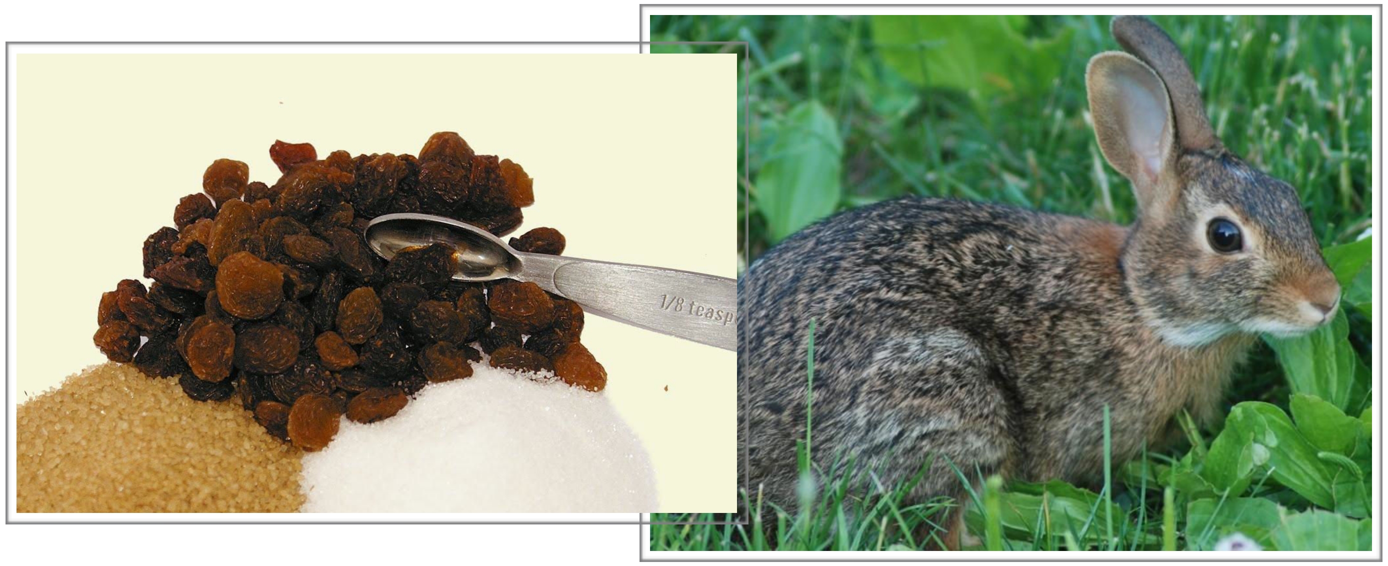
TRAINING DAY



GALAXY COMMUNITY CONFERENCE
BALTIMORE, MD | JUNE 30 - JULY 2, 2014

Raisins and Rabbit Turds

NGS Quality Control with Galaxy

Training Day, Galaxy Community Conference 2014, Johns Hopkins University

Monday, June 30th, 9-11:30 am, Charles Commons, Salon A Room 303

Tom Bair, Director of Iowa Institute for Human Genetics

Jennifer Hillman-Jackson, Galaxy Team at Penn State

Agenda

- Internet and CloudMan Access, Goals, WhoAreWe?
- Quality Assurance versus Quality Control
- Introduction to Common Tools
- Biobreak
- Data: Good, Bad, and …
- Hands-On Challenge

TRAINING DAY

GALAXY COMMUNITY CONFERENCE

BALTIMORE, MD || JUNE 30 - JULY 2, 2014

Agenda

- **Internet and CloudMan Access, WhoAreWe?, Goals**
- Quality Assurance versus Quality Control
- Introduction to Common Tools
- Biobreak
- Data: Good, Bad, and …
- Hands-On Challenge

TRAINING DAY

GALAXY COMMUNITY CONFERENCE

BALTIMORE, MD || JUNE 30 - JULY 2, 2014

Internet and CloudMan Access

Wi-Fi Access (connect first):

>GCC

> g@l@xycommittee

Galaxy CloudMan Instance for today (second):

> http://cloud[1-5].galaxyproject.org —please spread out

> Top **User Menu: Register** with email and password

write this down somewhere in case you log out

Shared Histories for Tutorials

On Galaxy CloudMan Instance:

> Top **Shared Data, Published Pages menu**

> On list, click on **Page** name: **Raisins**

*We won’t be importing yet, but this is the **source data** for the hands-on exercises. Remember how you got here.*

*This is a Galaxy »**Page**« that has been shared as »**Published**«. More about **Galaxy Objects** and **Sharing** later…*

Learning Goals

- applied QA and QC for experimental design
- explore Galaxy’s advanced web interface operations
- execute common QC tools and interpret the results
- extract, edit, and run a workflow
- accurately import SRA data of unknown provenance
- … pick the yummy raisins out for your next project

Who Are We All?

Biologist or Technologist or Both?

Wet lab or Computer?

Familiar with Galaxy? User interface, code, … ?

Other applications do you use?

What are your challenges as they relate to our goals?

Dogs name?

Agenda

- Internet and CloudMan Access, Goals, WhoAreWe?
- **Quality Assurance versus Quality Control**
- Introduction to Common Tools
- Biobreak
- Data: Good, Bad, and …
- Hands-On Challenge

TRAINING DAY

GALAXY COMMUNITY CONFERENCE

BALTIMORE, MD || JUNE 30 - JULY 2, 2014

Quality Assurance vs Quality Control

» ***Quality Assurance*** *is about* ***PROCESS*** »

- make sure you are doing the right things the right way

» ***Quality Control*** *is about* ***CONTENT*** »

- make sure your outputs are as expected

QC always occurs after **QA**

Quality Assurance Factors

Why it matters … and how to include

- Reproducibility
- Methods
- Timing
- Risk
- Balance
- Solutions

» it is always less painful to make QA part of the ongoing strategy than to try to engineer it back in at the end once things go wrong… and anyone who has done a lot of analysis will tell you, 'going wrong' is more common than magically 'going right' »

QA: Reproducibility

- Remember: QA is process. **How** the work is done. Exactly.

- **QA** is the the heart of reproducible science

- Without methods and materials that are known, executed consistently, and captured in some format where someone else can review and duplicate … your work is essentially a dead end (no one else can continue where you left off).

- Science is when conclusions/hypothesis can be tested and verified using ONLY reported methods and materials.

- If your work cannot be reproduced (tested) .. it becomes an opinion. That is not science. It is called philosophy.

QA: Methods

Practical places to implement QA strategy in NGS Informatics.

- did the data actually load completely and from the correct source
- is it of the correct size, datatype, and other macro metadata descriptions?
- logical connections makes sense in analysis (paired data really are pairs)?
- as tools are used, are inputs/outputs in correct formats, meeting expected criteria for »success« (mapping rates, skipped data)?
- technical failures can be for scientific reasons. technical successes can contain scientific failures. reading the manual, test runs, reviewing peer's results - all can help diagnose issues and help you to build up a collection of metrics that fit your data/workflow.

» *The best Methods include* ***metrics*** *that can quickly be scanned that PROVE what was* ***actually done*** *is what was* ***expected to be done****. Why waste your time?* »

QA: Timing and Risk

QA must be incorporated into the workflow to be successful. If you are working on a project where the results are time-bound or there is a deliverable due, then NOT using QA increases risk of failure dramatically.

- start with QA, execute throughout, and then end with it.
- if you are going to bother to capture QA metrics, then make sure you look. running FastQC is not enough, you must review and act on it.
- sound and absolutely correctly executed data is very glamorous right now (as it should be). and it gives you an edge. change a variable, know the effect, improve the results, publish higher quality work. you simply cannot do that with a noisy, uncertain, workflow.
- should you work in industry and someone is paying for your results, QA is mandatory for success. if you just want a job in industry later, or want that option open, this is THE skill to master first.

» Small risk: it *is too late to fix an incomplete uploaded file after Cuffdiff has run.* Large risk*: it is too late to publish your paper on the original deadline (or using original budget) after 100 Cuffdiff runs were executed incorrectly and the results are off because the quality scores were scaled wrong way back before even Tophat was run on the 1000+ replicates.*»

QA: Balance and Solutions

Balance is good. And if you have problems due to poor QA, they can of course be addressed.

- Too much QA is just as bad as too little. Knowing how much to add, and where, and when - not just when in the workflow, but when in that workflows lifecycle - is critical. Start small and obvious.

- There are entire specialties around QA for almost every endeavor and many of the theories can be applied to scientific applications. Let risk be your guide, just like the larger data projects do (UCSC, Ensembl, SRA, etc. all have very strong scientific QA groups).

- There are a plentitude of examples of very important work done without adequate QA or method capture. I won't write then down but may tell you of a few or maybe you can tell me.

- Where to start if you have problems? At the most critical junctures: transfer points between groups or sources, versioning of data and tools and environments, proper methods written down somewhere (completely - exactly) then work on how to capture the execution details when implemented. This applies to anyone working alone - you are your own method-policemen at first, but the journals and later anyone reading your paper will be.

» *We'll walk through a solution to a very common problem, datatype assignment, as an exercise later in this tutorial. It could and should probably be used in some version by everyone doing any analysis in Galaxy.* »

Quality Control Factors

Why it matters … and how to include

- Content
- Conclusions
- Action
- Fit

» just like QA, it is always less painful to make QC part of the ongoing strategy than to try to engineer it back in at the end once things go wrong. »

QC: Content and Conclusions

- Remember: QC is content. **What** the work represents.

- **QC** is the the heart of quality science

- If you don’t know what you have going into an experiment, then it is very difficult to know what is coming out. Again, your work is essentially a dead end (no one else can continue where your **conclusions** left off).

- QC always occurs after QA. It is a waste of your time to do it first. If you do QA but then skip QC, then expect GIGO.

- The same rules about reproducibility for QA apply here for QC - capture methods and materials. All steps in an experiment matter.

QC: Action and Fit

- Good QC methods provide you with **actionable metrics**.

- The same action, after QC, is not appropriate for every experiment, even given the same inputs. The action must **fit** the experimental goals.

- Example A: RNA-seq data intended for expression analysis should have very little done to it beside basic artifact removal to preserve original abundance counts and maximize paired mappings.

- Example B: RNA-seq data intended for variant calling analysis should be screened more thoroughly. Quality of individual bases matters very much. Matched paired going into the tools matters very much.

» Know you data, your QC tools, your intended analysis tools, and what you need to do to the inputs to maximise the quality of the results. How? **That is what we'll be covering NEXT!! »**

Critical things to check off when starting an experiment

- Do you understand the biology and experimental design
    - Ideally you would have been involved in the experimental design
- Do you understand/have worked with before this sequencing technology
    - ABI is not Illumina is not PacBio
    - Do you know what to expect in terms of quality and quantity of data
- Do you have a good understanding of the consequences of doing or not doing “routine” steps
    - should you trim this data
    - what aligner is appropriate
    - should you remove duplicates
    - what about multiple alignments

What to do with “bad” data

- You have gotten some data it is labeled as a fastqc.gz data.

- You run FastQC on the data and it looks normal

- You run bowtie and get very low alignment percentages 52-68%

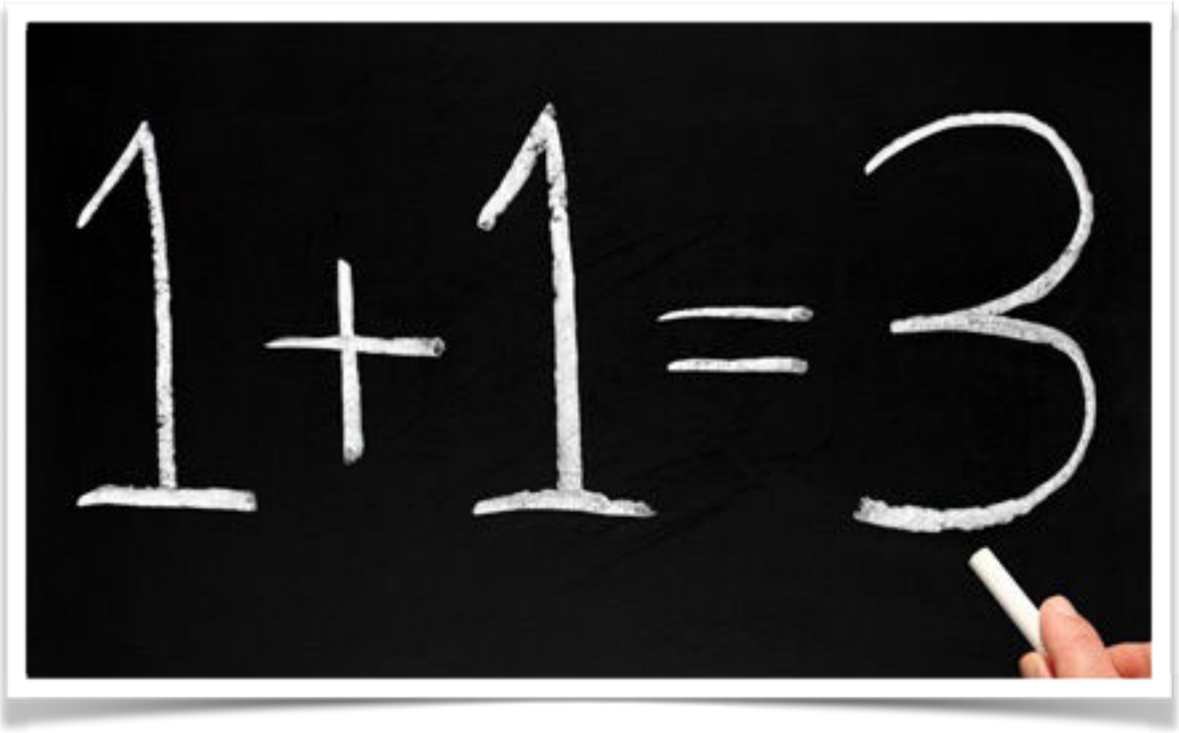
- What could be wrong? What should you do?

More info

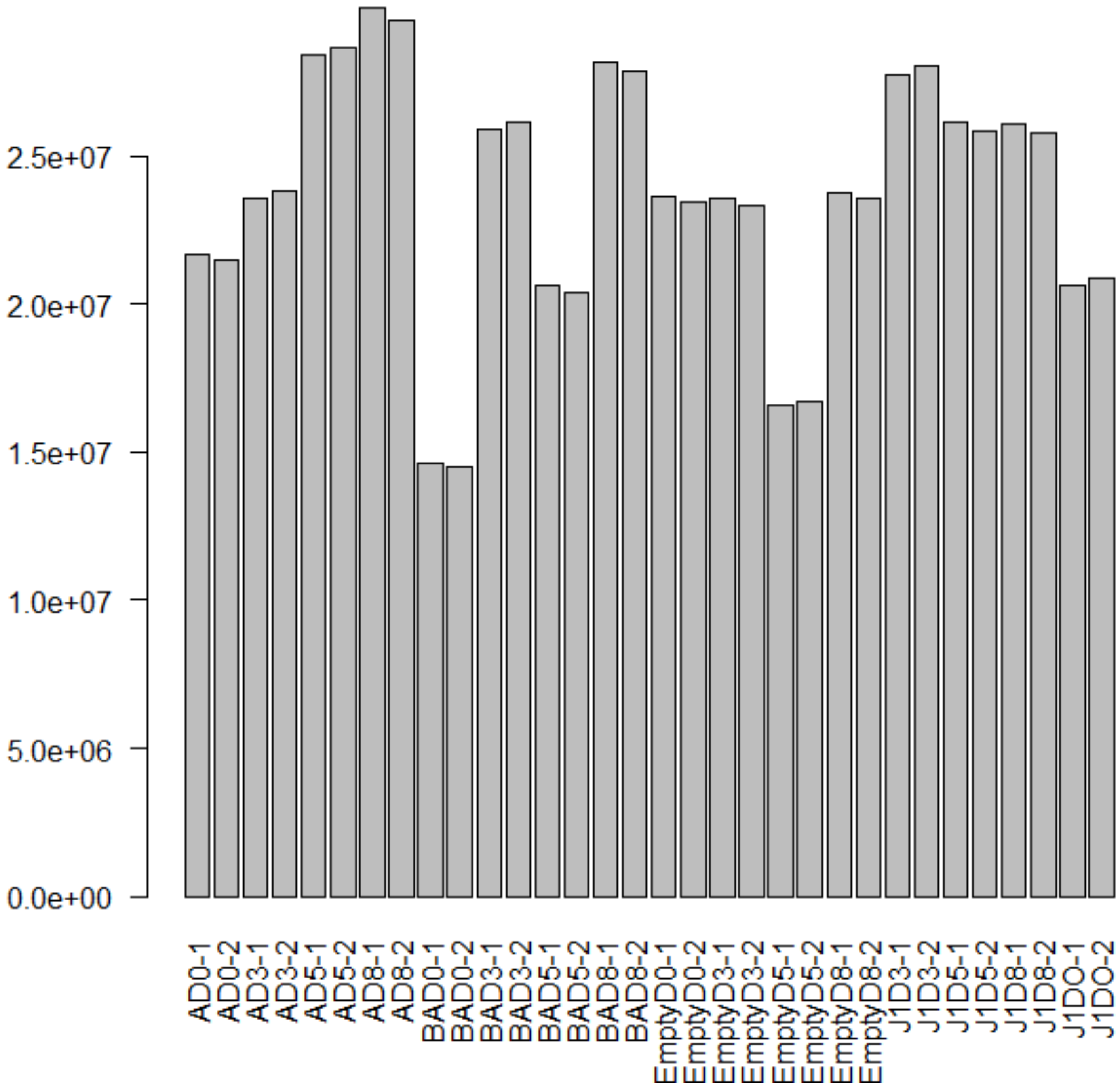
- The data is from an ion torrent machine…
- Re-running with bwa-mem and TMAP both give percentages in the 95+%
- Much longer reads than bowtie was designed for
  - more homopolymeric mis-calls

An “easy” RNA-Seq experiment

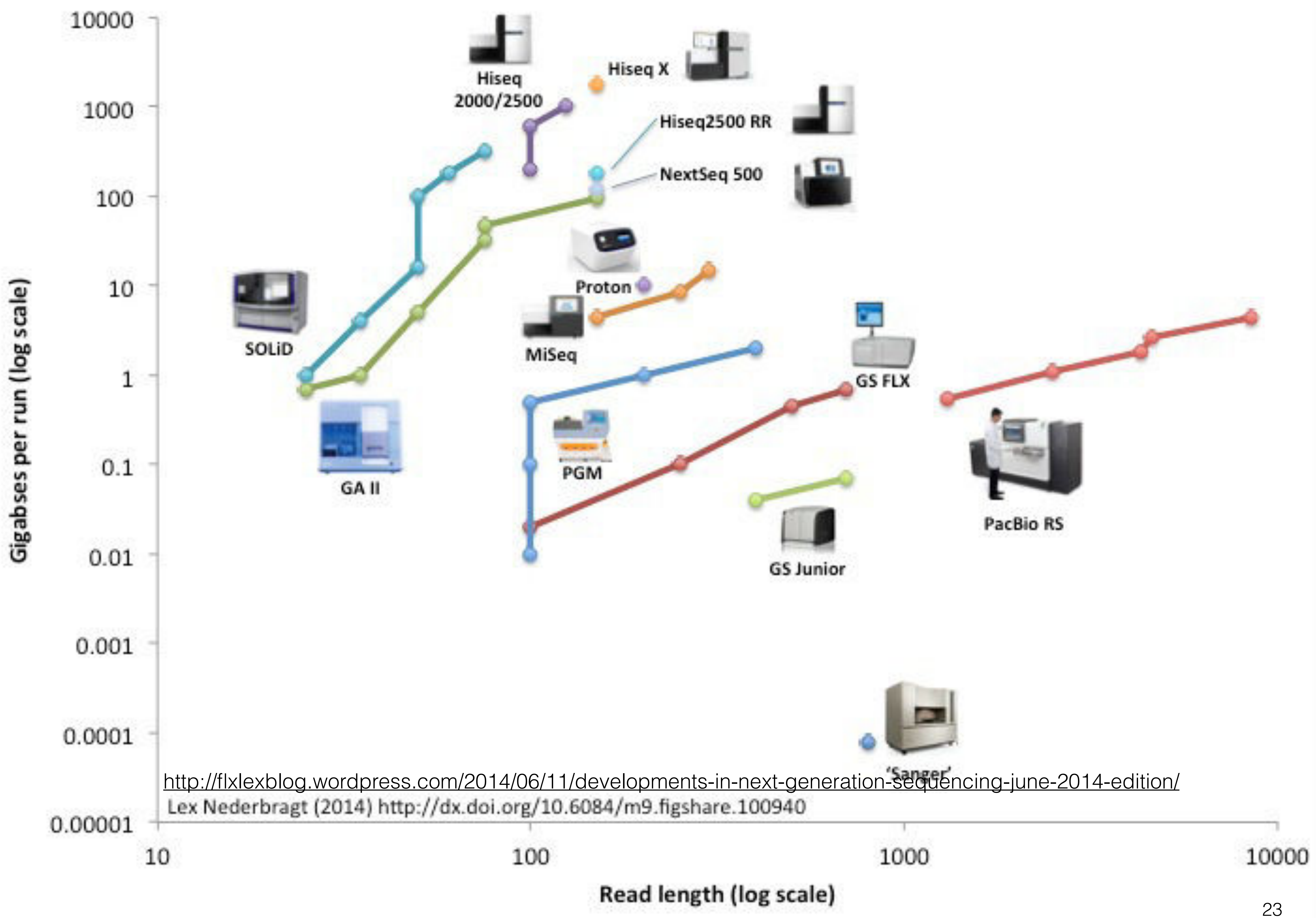
Sample names are obvious from the file names
Simple time-course design
Two replicates at each point



Here are some replicates…



Developments in High Throughput Sequencing



http://flxlexblog.wordpress.com/2014/06/11/developments-in-next-generation-sequencing-june-2014-edition/

Lex Nederbragt (2014) http://dx.doi.org/10.6084/m9.figshare.100940

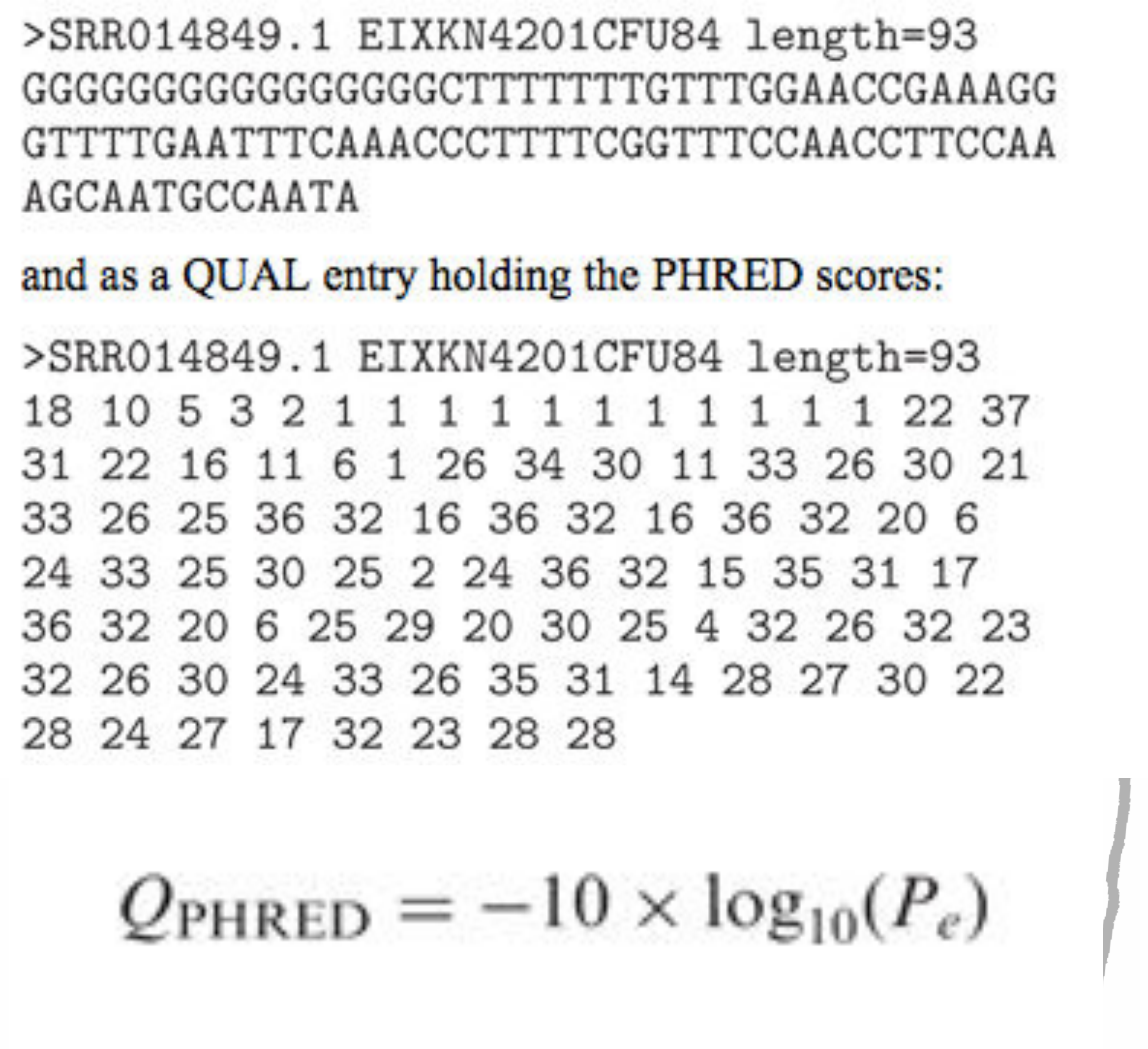
FASTQ file formats

- Great review at Nucleic Acids Res. Apr 2010; 38(6): 1767–1771.

- Just because it has the same file extension does not mean you can treat it the same…fastqillumina doesn’t mean what you think it means

What is a PHRED score?

- Started way back with Sanger sequencing
- Gives the confidence in the base call 1:10 = 10,1:100 = 20, 1:1000 = 30, 1:10000 = 40
- Kinda clunky, two characters
  - hard to parse
  - big files

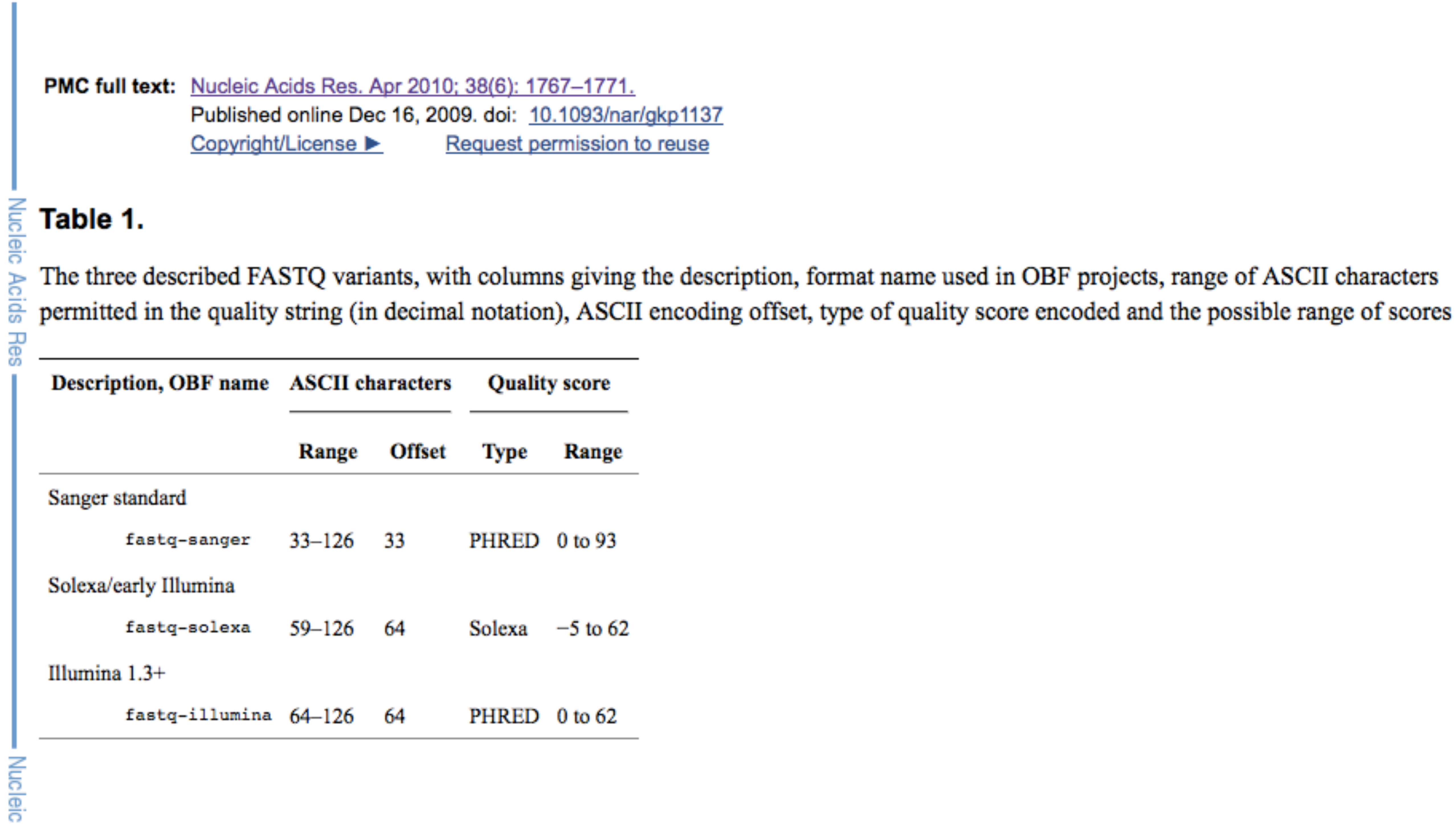


```
>SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGG
GTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAA
AGCAATGCCAATA
```

and as a QUAL entry holding the PHRED scores:

```
>SRR014849.1 EIXKN4201CFU84 length=93
18 10 5 3 2 1 1 1 1 1 1 1 1 1 1 1 22 37
31 22 16 11 6 1 26 34 30 11 33 26 30 21
33 26 25 36 32 16 36 32 16 36 32 20 6
24 33 25 30 25 2 24 36 32 15 35 31 17
36 32 20 6 25 29 20 30 25 4 32 26 32 23
32 26 30 24 33 26 35 31 14 28 27 30 22
28 24 27 17 32 23 28 28
```

$$Q_{\text{PHRED}} = -10 \times \log_{10}(P_e)$$

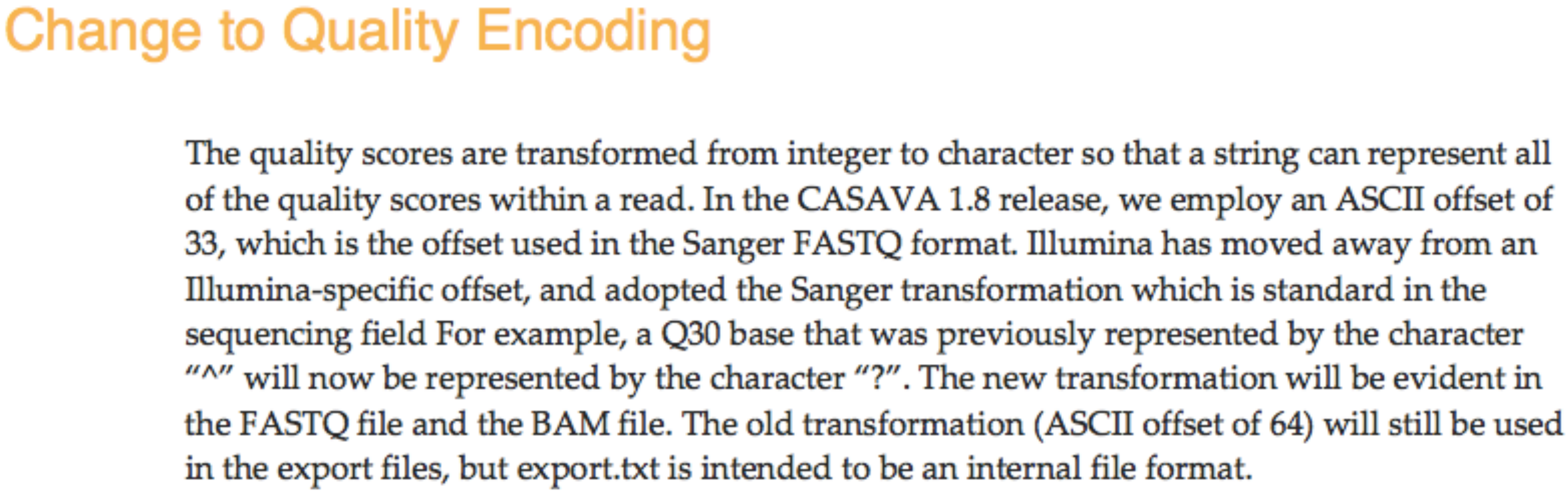


PMC full text: Nucleic Acids Res. Apr 2010; 38(6): 1767–1771.
Published online Dec 16, 2009. doi: 10.1093/nar/gkp1137
Copyright/License ▶ Request permission to reuse

Table 1.

The three described FASTQ variants, with columns giving the description, format name used in OBF projects, range of ASCII characters permitted in the quality string (in decimal notation), ASCII encoding offset, type of quality score encoded and the possible range of scores

Description, OBF name	ASCII characters		Quality score	
	Range	Offset	Type	Range
Sanger standard				
`fastq-sanger`	33–126	33	PHRED	0 to 93
Solexa/early Illumina				
`fastq-solexa`	59–126	64	Solexa	−5 to 62
Illumina 1.3+				
`fastq-illumina`	64–126	64	PHRED	0 to 62



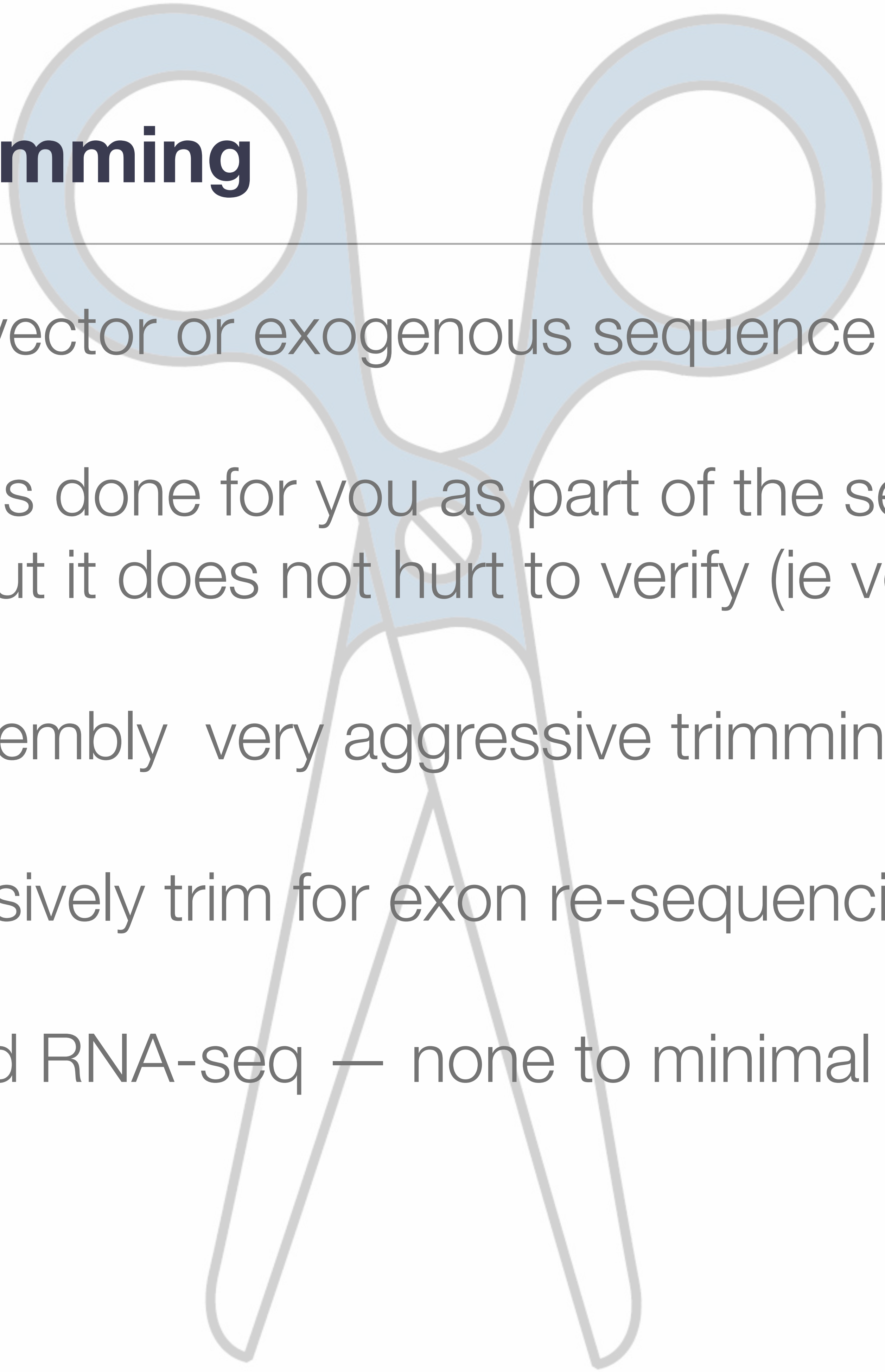
Change to Quality Encoding

The quality scores are transformed from integer to character so that a string can represent all of the quality scores within a read. In the CASAVA 1.8 release, we employ an ASCII offset of 33, which is the offset used in the Sanger FASTQ format. Illumina has moved away from an Illumina-specific offset, and adopted the Sanger transformation which is standard in the sequencing field For example, a Q30 base that was previously represented by the character "^" will now be represented by the character "?". The new transformation will be evident in the FASTQ file and the BAM file. The old transformation (ASCII offset of 64) will still be used in the export files, but export.txt is intended to be an internal file format.

Finally

http://supportres.illumina.com/documents/myillumina/354c68ce-32f3-4ea4-9fe5-8cb2d968616c/casava1_8_changes.pdf

Notes on trimming



- Always trim vector or exogenous sequence data
  - Often this is done for you as part of the sequencing pipeline, but it does not hurt to verify (ie verify)
- Genome assembly very aggressive trimming
- Fairly aggressively trim for exon re-sequencing
- Chip-seq and RNA-seq — none to minimal trimming

CloudMan Login: Double Check

Wi-Fi Access (connect first):

> *name_here*

> *password_here*

Galaxy CloudMan Instance for today (second):

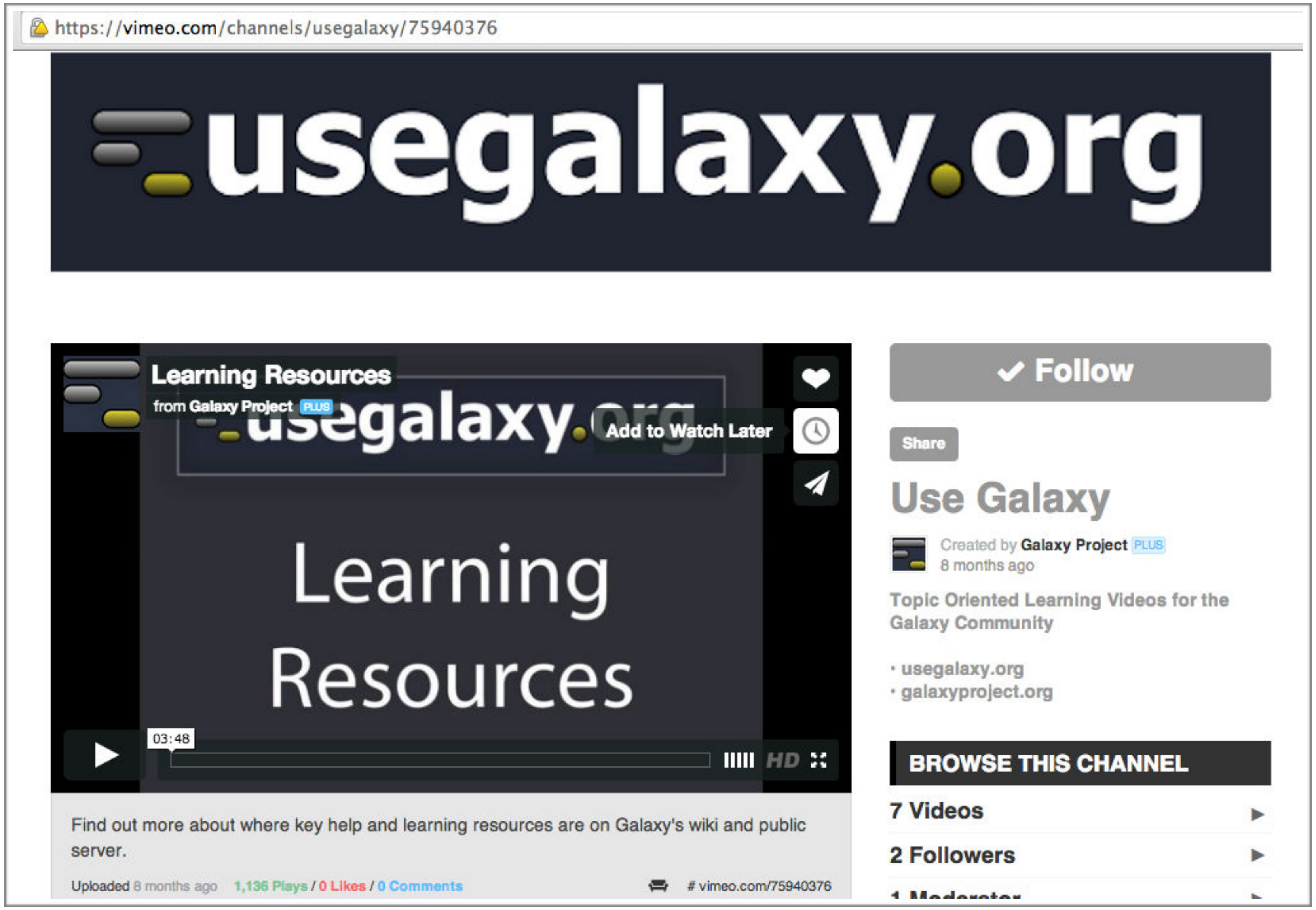
> url_for_cloudman

> Top **User Menu: Register** with email and password

write this down somewhere in case you log out

Galaxy orientation

- Histories
- Tools
- Datasets
- Metadata
- Galaxy Objects
- Resources
- ***Vimeo (more!):***
  http://vimeo.com/galaxyproject/learning
  http://vimeo.com/galaxyproject/datasets1



Find out more about where key help and learning resources are on Galaxy's wiki and public server.

Uploaded 8 months ago 1,136 Plays / 0 Likes / 0 Comments # vimeo.com/75940376

Use Galaxy

Created by Galaxy Project PLUS
8 months ago

Topic Oriented Learning Videos for the Galaxy Community

- usegalaxy.org
- galaxyproject.org

BROWSE THIS CHANNEL

7 Videos ▶
2 Followers ▶
1 Moderator ▶

Agenda

- Internet and CloudMan Access, Goals, WhoAreWe?
- Quality Assurance versus Quality Control
- **Introduction to Common Tools**
- Biobreak
- Data: Good, Bad, and …
- Hands-On Challenge

TRAINING DAY
GALAXY
COMMUNITY CONFERENCE
BALTIMORE, MD | JUNE 30 - JULY 2, 2014

Agenda

- Internet and CloudMan Access, Goals, WhoAreWe?
- Quality Assurance versus Quality Control
- Introduction to Common Tools
- **Biobreak - 10 minutes**
- Data: Good, Bad, and …
- Hands-On Challenge

TRAINING DAY

GALAXY COMMUNITY CONFERENCE

BALTIMORE, MD || JUNE 30 - JULY 2, 2014

Agenda

- Internet and CloudMan Access, Goals, WhoAreWe?
- Quality Assurance versus Quality Control
- Introduction to Common Tools
- Biobreak
- **Data: Good, Bad, and …**
- Hands-On Challenge

TRAINING DAY

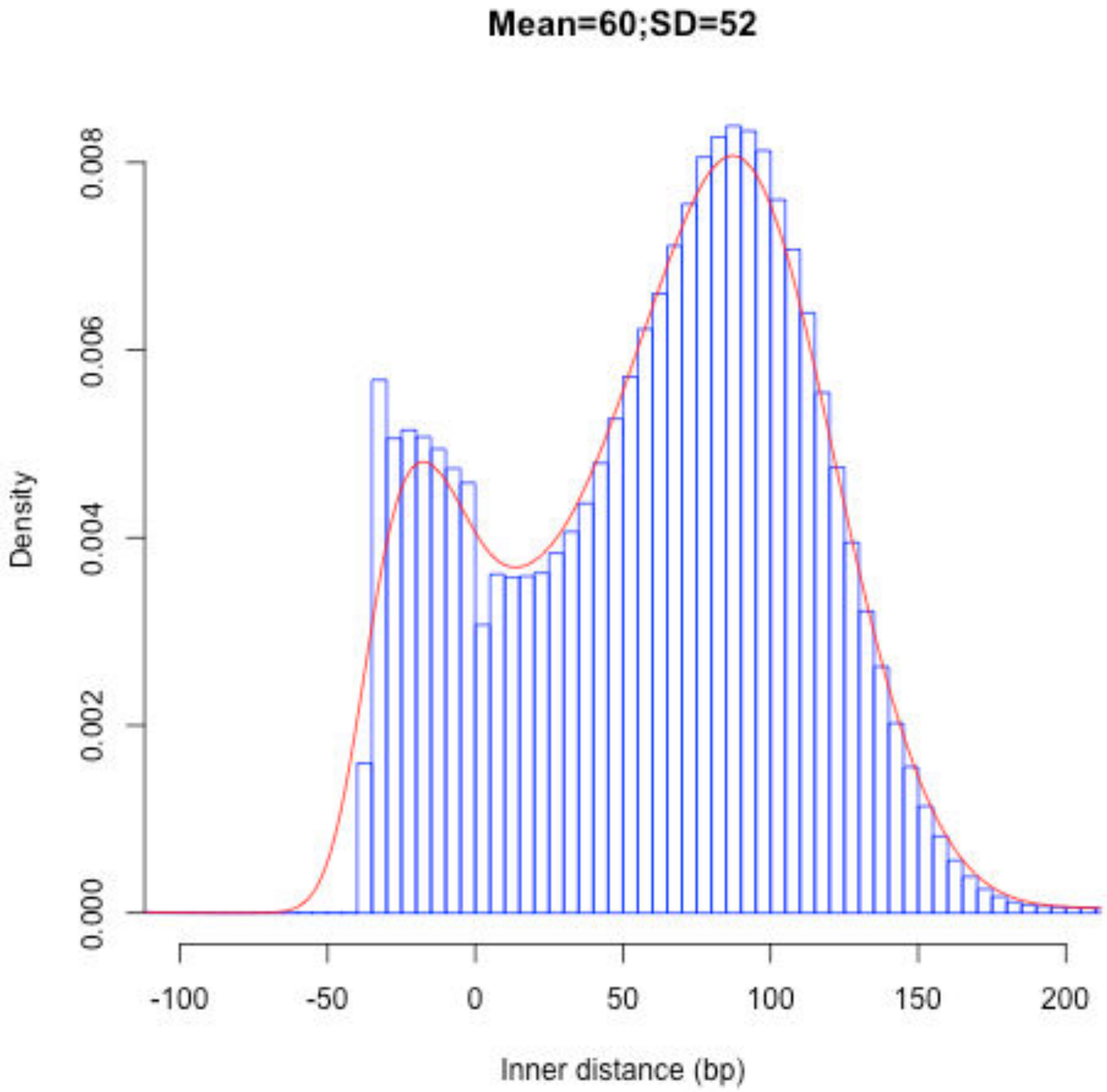
GALAXY COMMUNITY CONFERENCE

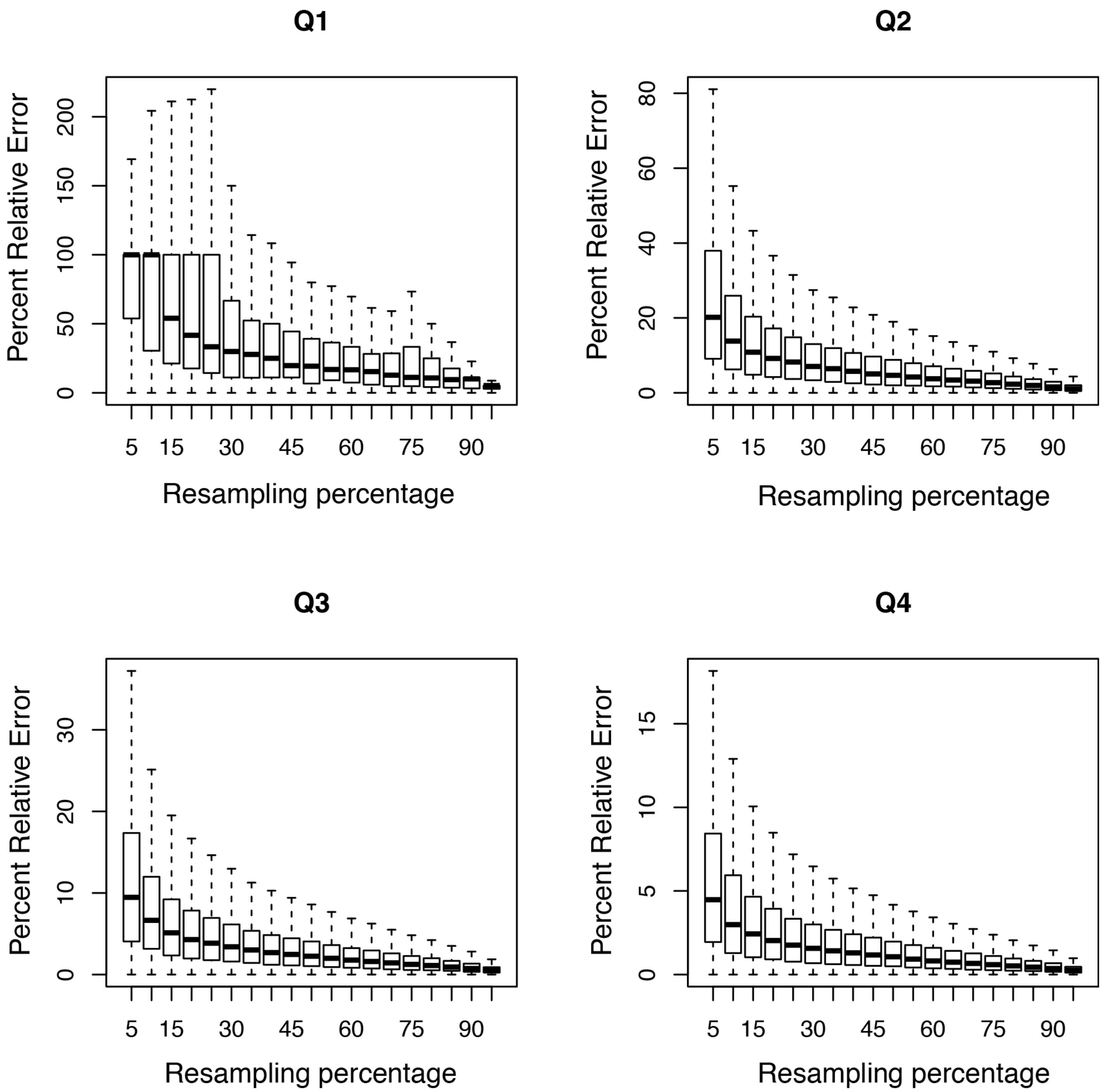
BALTIMORE, MD || JUNE 30 - JULY 2, 2014

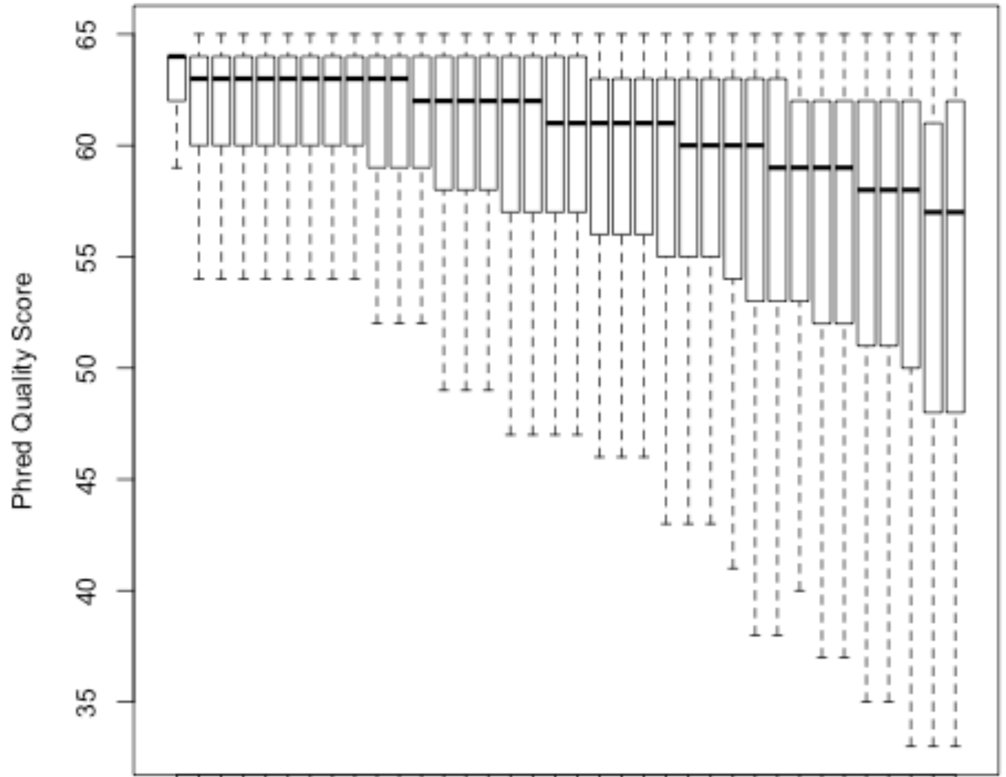
RSeQC: An RNA-seq Quality Control Package

Right tool for the job

http://rseqc.sourceforge.net/
Focused around RNA-Seq not genomic shotgun sequences




Q1
Q2
Q3
Q4
Percent Relative Error
Resampling percentage
0
50
100
150
200
20
40
60
80
10
30
5
15
5
15
30
45
60
75
90


Phred Quality Score
35
40
45
50
55
60
65

QA Practical: Datatype assignment

- Load datasets
- Run FastQC
- Interpret datatype
- Modify datatype if needed or directly assign datatype

Start by loading the History on our Tutorial »Page« named:

FASTQ dataset QA

More resources:

http://vimeo.com/galaxyproject/fastqprep
http://wiki.galaxyproject.org/Support#FASTQ_Datatype_QA

Acknowledgements

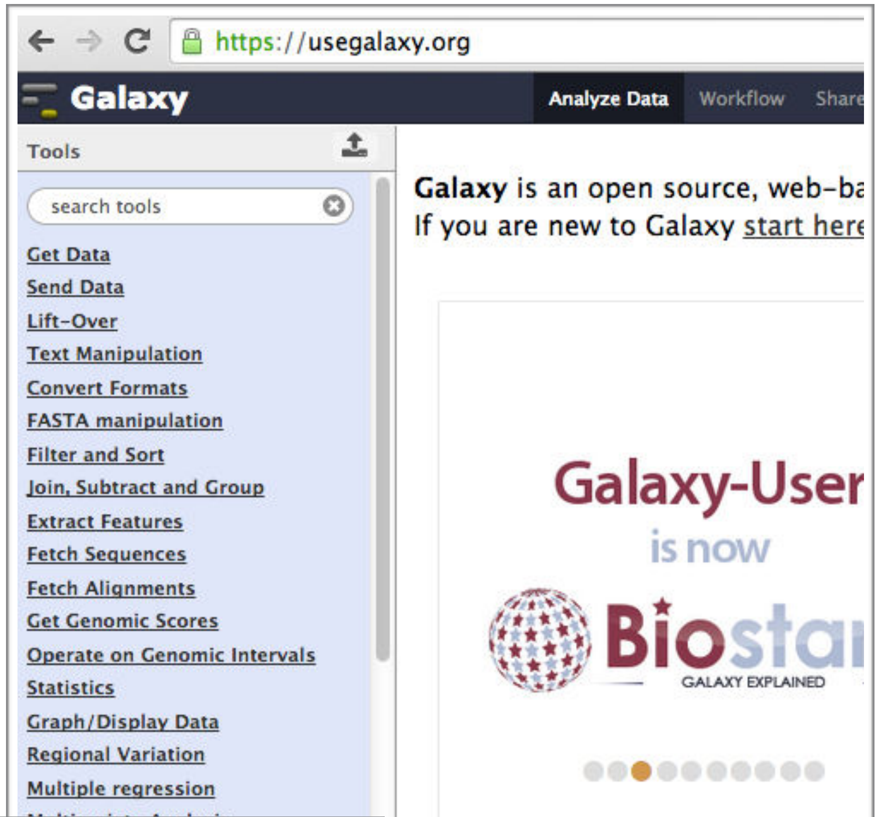
- The GCC 2014 Organizing Committee and our hosts Johns Hopkins with a special thanks to **Dave Clements** and **Stacey Hooker**
- IIHG group in particular **Richard Smith** for the title
- The Galaxy team, especially **Anton Nekrutenko** and **James Taylor** (our awesome PIs) plus **Dannon Baker** for exceptional CloudMan development, set-up, and support
- All of **YOU** for coming to **GCC**!! Working together as a *community* is what makes Galaxy special.

So… Learn, Share, and **HAVE A GOOD TIME**!!!

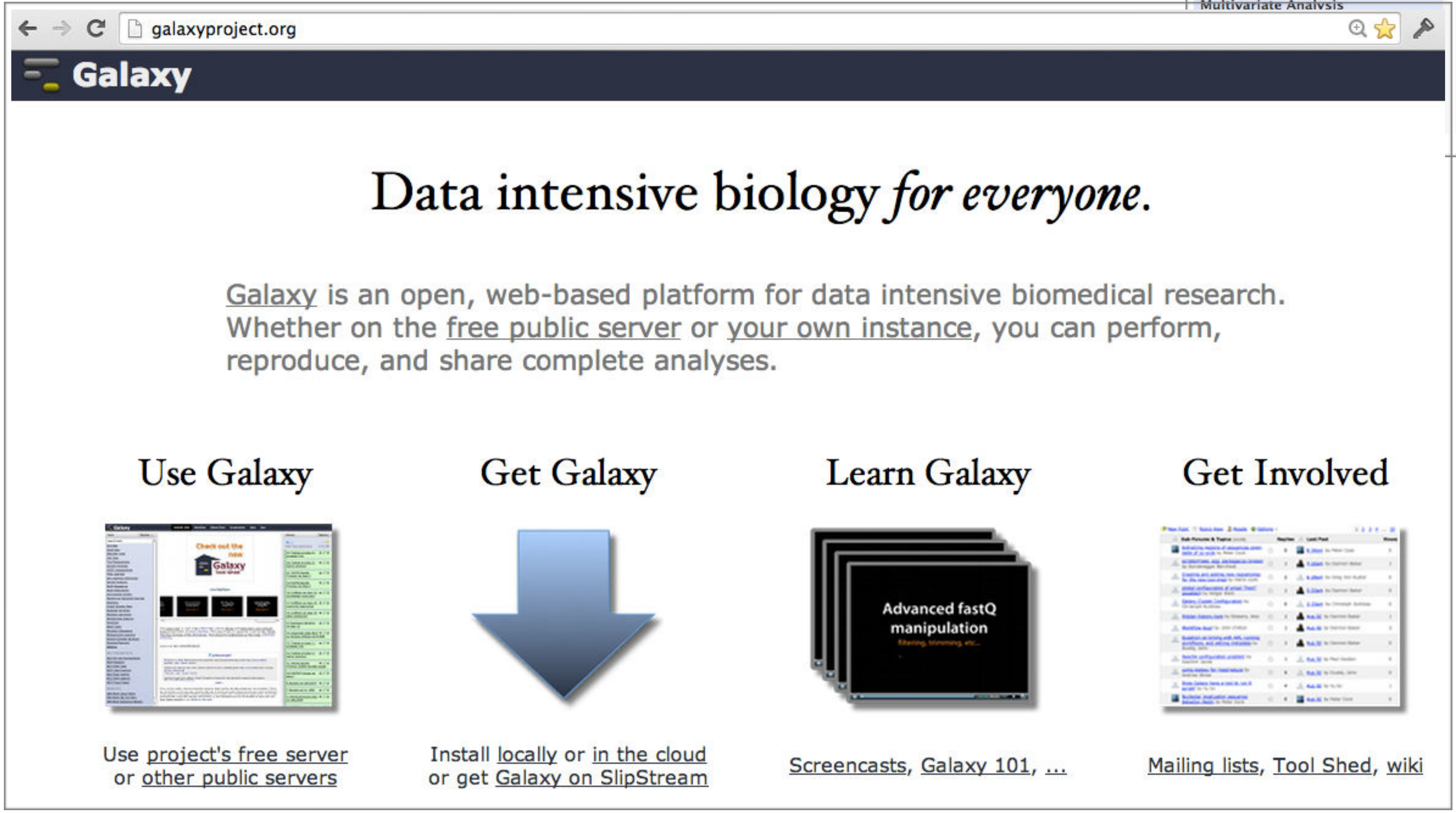
All you really need to remember ….

http://usegalaxy.org

http://galaxyproject.org



Galaxy is an open source, web-ba
If you are new to Galaxy start here



Data intensive biology *for everyone.*

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on the free public server or your own instance, you can perform, reproduce, and share complete analyses.

Use Galaxy

Use project's free server or other public servers

Get Galaxy

Install locally or in the cloud or get Galaxy on SlipStream

Learn Galaxy

Screencasts, Galaxy 101, ...

Get Involved

Mailing lists, Tool Shed, wiki

TRAINING DAY

GALAXY COMMUNITY CONFERENCE

BALTIMORE, MD | JUNE 30 - JULY 2, 2014

Homework

»*There's so much time, and so little to do!*«

- Willy Wonka

- Locate the »**Raisins**« Page on CloudMan during GCC
- The Page will be migrated **UseGalaxy.org** after GCC along with the example datasets and histories.
- Questions can be sent through regular support resource channels. **Galaxy Biostar** would be a good choice. Linked from any tool on UseGalaxy.org, more info is at: http://wiki.galaxyproject.org/Support#Biostar