**Statistical analysis of sequins with Anaquin**

Tim Mercer, Wendy Chen, Ted Wong

Garvan Institute of Medical Research

**INTRODUCTION**

Anaquin is a C++ statistical analysis tool for sequin data-analysis. It covers RNA, DNA and metagenomics. In this tutorial, we will demonstrate the usage of the software on simulated data-sets and real data-sets.

Adding Standards to natural sample DNA/RNA standards can be added to a sample comprising a target DNA/RNA sequence to be determined. The source of target DNA/RNA can come from any known organism or environmental sample. For example, DNA/DNA standards can be added to natural RNA derived from animal (such as mammalian and human), plant (such as corn and rice), microbial (including bacteria and archaea) and environmental (soil samples, human stools, clinical samples such as infected wound fluid) sources. Because RNA/DNA standards have no homology to natural sequences, sequenced reads derived from RNA/DNA standards can be distinguished from sequenced reads derived from natural RNA/DNA sample (Figure 18). This enables DNA/RNA standards to be added to the RNA/DNA sample, prior to sequencing, and therefore undergo the same library preparation, sequencing, alignment and analysis as for the DNA/RNA sample of interest. RNA/DNA standards are combined with RNA/DNA samples to comprise only a fraction of the combined total. This fractional contribution (typically between 0.1 and 10% of total) varies according to the type of library preparations – rRNA removal, polyA or total RNA purification. Note that the fractional contribution of RNA/DNA standards is inversely proportion to the sequencing depth attributed to the RNA/DNA sample. Therefore, the fractional total should be the minimum amount required to sufficiently enable analysis of RNA/DNA standards.

# 1. Preparation

**Operating system**

This protocol assumes a unix operating system (I.e, Linux or MacOSX), with a bash terminal. All commands given here should run in a terminal window.

**Software**

* The Anaquin statistical software and its bioconductor package from [http://www.anaquin.org](http://www.anaquin.org/).
* The samtools package ([http://samtools.sourcegorget.net](http://samtools.sourcegorget.net/))
* A splice junction aligner for RNA reads. Tophat2 is covered here, but other softwares such as STAR, BBMap, SpliceMap, Subread or GSNAP are also possible.
* The cufflinks package (<http://cole-trapnell-lab.github.io/cufflinks>) for transcriptome assembly and differential expression analysis.
* The Picard
* The GATK package (<https://www.broadinstitute.org/gatk>) for variant discovery and genotyping. Other softwares such as FreeBayes are also possible.
* The velvet assembler (<http://www.ebi.ac.uk/~zerbino/velvet>) for sequence assembly for a single genome.
* The MetaVelvet assembler ([http://metavelvet.dna.bio.keio.ac.jp](http://metavelvet.dna.bio.keio.ac.jp/)) for sequence assembly for multiple genomes.
* The R statistical package ([http://www.r-project.org](http://www.r-project.org/)) for plotting and advanced statistical analysis.

**IMPORTANT** It is important to note the softwares demonstrating in this protocol can be replaced by their equivalents. For example, anything that produces a SAM/BAM file from RNA next-generation sequencing reads can be used. In fact, the goal of the project is to provide a framework for statistical analysis of bioinformatics softwares.

**Installation**

While there are many ways to install the softwares, we will demonstrate two popular package managers: homebrew (for OS-X) and APT (for Debian Linux). Please consider manual installation if either package manager is not option. Consult the documentation for further information if necessary.

Not everything can be installed by the package managers. We will show the commands if available.

1. Install homebrew (OS-X only)

ruby -e "$(curl -fsSL [https://raw.githubusercontent.com/Homebrew/install/master/install](https://raw.githubusercontent.com/homebrew/install/master/install))"

Refer to the documentation on [http://brew.sh](http://brew.sh/) for more installation options.

1. Install bowtie2

$ brew install bowtie2

$ sudo apt-get install bowtie2

Download the latest release from <http://bowtie-bio.sourceforge.net/bowtie2> and install it.

1. Install bwa

$ brew install bwa

$ sudo apt-get install bwa

More information can be found on the project homepage at [http://bio-bwa.sourceforge.net](http://bio-bwa.sourceforge.net/).

1. Install tophat2

$ brew install tophat2

$ sudo apt-get install tophat2

Download the latest binary from <https://ccb.jhu.edu/software/tophat/index.shtml>.

1. Install cufflinks

$ brew install cufflinks

$ sudo apt-get install cufflinks

Download from <http://cole-trapnell-lab.github.io/cufflinks> for the latest executable.

1. Install samtools

$ brew install samtools

$ sudo apt-get install samtools

Download from [http://samtools.sourceforge.net](http://samtools.sourceforge.net/) for the latest executable.

1. Install picard & GATK

Download the latest release for picard and GATK respectively from <http://broadinstitute.github.io/picard> and <https://www.broadinstitute.org/gatk/download>.

1. Install velvet

$ brew install velvet

$ sudo apt-get install velvet

A useful reference manual to velvet is available on <https://www.ebi.ac.uk/~zerbino/velvet>. Download the latest release of velvet and install it.

1. Install metaVelvet

Download the latest release from [http://metavelvet.dna.bio.keio.ac.jp](http://metavelvet.dna.bio.keio.ac.jp/) and extract the archive to install it.

1. Install R and Bioconductor

$ brew install R

$ sudo apt-get install R

Download the latest release version of R from [http://cran.r-project.org](http://cran.r-project.org/) and install it. Consult the R installation and administration manual if necessary. A useful quick reference for R commands can be found at [http://cran.r-project.org/doc/contrib/Short-refcard.pdf](http://cran.r-project.org/doc/contrib/short-refcard.pdf).

1. Install Anaquin

Download the latest release from [http://www.anaquin.org](http://www.anaquin.org/). Extract the archive and check its version number with:

$ wget <http://www.anaquin.org/downloads/anaquin.tar>

$ tar -xf anaquin.tar

$ cd anaquin

$ ./anaquin -v

Version 1.0. Garvan Institute of Medical Research, 2015.

Chromosome: chrT version 1.0

Mixture A: version 1.0

Mixture B: version 1.0

The output shows the version number for the software and its data files.

To test the installation:

$ ./anaquins -t

**==================================**

All tests passed (300 assertions in 30 test cases)

The -t option runs internal test cases shipped with the software. All test cases must pass. Please contact Garvan Institute if any of the test fails.

To install the Bioconductor package, start R:

> source('http://www.bioconductor.org/biocLite.R')

> biocLite('BiocUpgrade')

> biocLite(c('anaquin'))

This retrieves an automatic installation tool (biocLite) and installs the version-matched packages. In addition, the installation tool will automatically download and install all other prerequisite packages.

Try the installation:

> library('anaquin')

1. Download the example data
2. Download the reference genome

Download the human reference genome in GZ format from UCSC. Browse the GRCh38 human genome at [http://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips](http://hgdownload.cse.ucsc.edu/goldenpath/hg38/bigzips). Only the soft-masked assembly sequence is required for this protocol:

$ wget [http://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz](http://hgdownload.cse.ucsc.edu/goldenpath/hg38/bigzips/hg38.fa.gz)

$ gunzip hg38.fa.gz

1. Download the in-silico chromosome

The chromosome is available at <http://www.anaquin.org/downloads>.

$ wget [http://www.anaquin.org/downloads/silico.tar](http://www.anaquin.org/downloads/chromosome.tar)

$ tar -xf silico.tar

This includes the chromosome sequence, gene model annotations, mixture files and other useful files that we will use later in the protocol.

The files can also be copied from the anaquin software directly:

$ anaquin -silico

Copied silico.fa

Copied rna\_silico.gtf

Copied rna\_mixture\_A.csv

Copied rna\_mixture\_B.csv

[… truncated …]

1. Build the reference index

Before reads can be aligned, an index for the genome reference must be constructed. Since we will use bowtie2 (for RNA) and bwa (for DNA), we will need to build an index for each aligner.

**! CRITICIAL** Make sure that the in-silico chromosome is added to the reference genome.

Nothing will be mapped to the in-silico chromosome if this step is not performed.

To add the in-silico chromosome to the human ch38 genome:

$ cp hg38.fa combined.fa

$ cat silico.fa > combined.fa

It is always a good idea to confirm before indexing:

$ grep chrT combined.fa

>chrT

To build a bowtie2 index, issue the command:

$ bowtie2-build combined.fa combined

A set of bt2 files will be produced, with names starting with combined\_ specified above.

To build a bwa index, issue the command:

$ bwa index combined.fa combined

The following files should be generated:

$ ls combined.fa.\*

combined.fa.amb combined.fa.ann combined.fa.bwt combined.fa.pac combined.fa.sa

This procedure needs to be run only once for each reference genome used.

# 2. RNA Sequencing

RNA sequencing is a commonly used method to assemble the structure of genes and profile gene expression. We can use RNA standards to assess the accuracy with which RNA sequencing assembles and quantifies the gene loci encoded within the in-silico chromosome. To demonstrate the use of RNA standards with a RNA sequencing analysis of a human RNA sample, we will employ an experiment on K562 cell gene expression.

K562 cells were cultured according to Coriell Cell Repositories growth protocols and standards. Brieftly, K562 cells were cultured in RPMI 1640 medium (Gibco) supplemented with 10% fetal bovine serum (FBS) at 37C under 5% CO2. Total RNA was extracted from K562 cells using TRIzol (Invitrogen) according to the manufacturer's instruction. Dnase treatment was subsequently performed on each sample with TURBO Dnase (Life Technologies) followed by a cleanup with the RNA Clean and Concentrator Kit (Zymo Research). Total RNA was run on a BioAnalyzer to check for integrity and to determine the concentration. Only RNA with a RNA integrity number (RIN) > 9.5 were used for library preparation.

**2.1 | Download data**

Create a working directory and download the sequencing data into it. It was a paired-end experiment (ie, R1 == first and R2 == second). Paired-end sequencing facilitates detection of genomic rearrangements and repetitive sequence elements. Each library corresponds to a biological replicate.

$ mkdir rna\_protocol

$ wget ????

$ unzip ????

$ ls \*.fastq

K562\_1A\_R1.fastq K562\_1A\_R2.fastq

K562\_2A\_R1.fastq K562\_2A\_R2.fastq

K562\_3A\_R1.fastq K562\_3A\_R2.fastq

**2.2 |** **RNA Alignment**

Align the sequence files to the reference genome. Please make sure the in-silico chromosome has been added (section ?).

$ tophat2 -p 4 -G rna\_silico.gtf -o K562\_1A combined \

K562\_1A\_R1.fastq K562\_1A\_R2.fastq

$ tophat2 -p 4 -G rna\_silico.gtf -o K562\_1A combined \

K562\_2A\_R1.fastq K562\_2A\_R2.fastq

$ tophat2 -p 4 -G rna\_silico.gtf -o K562\_1A combined \

K562\_3A\_R1.fastq K562\_3A\_R2.fastq

The option -G supply tophat2 with a gene model annotations to facilitates mapping reads to the in-silico chromosome. Only reads that have failed to map to the chromosome will then be mapped to the human genome. This option is recommended but not mandatory. -p specifies the number of threads to align reads, the default is 1. -o specifies the output directory.

**2.2 |** Check the mapping results

$ ls K562\_1A/accept\*.bam K562\_2A/accept\*.bam K562\_3A/accept\*.bam

K562\_1A/accepted\_hits.bam

K562\_2A/accepted\_hits.bam

K562\_3A/accepted\_hits.bam

**2.3 |** Examine the BAM/SAM files (Optionally)

Examine the mapping to ensure in-silico chromosome has been mapped. We need to convert the BAM files to SAM format because BAM is a binary format and thus unable to show on a terminal.

$ samtools view -h -o K562\_1A/RK1A.sam K562\_1A/RK1A.bam

$ samtools view -h -o K562\_2A/RK2A.sam K562\_2A/RK2A.bam

$ samtools view -h -o K562\_3A/RK3A.sam K562\_3A/RK3A.bam

Count number of reads mapped to the in-silico chromosome.

$ grep chrT K562\_1A/RK1A.sam | wc

$ grep chrT K562\_2A/RK1B.sam | wc

$ grep chrT K562\_3A/RK1C.sam | wc

Approximate the dilution (amount of mixture spiked relative to the overall sample).

$ grep -v chrT K562\_1A/RK1A.sam | wc

$ grep -v chrT K562\_2A/RK1B.sam | wc

$ grep -v chrT K562\_3A/RK1C.sam | wc

The approximate dilution for the replicates are ??, ?? and ??.

**2.4 |** Calculate the alignment statistic

Small sequenced reads are first aligned to a reference genome. The alignment of reads to a large reference genome is a computationally intensive task that can be performed in numerous ways, providing differential outcomes for speed, sensitivity and accuracy. Furthermore, sequenced reads that traverse introns are required to be aligned to the reference genome in a split or non-contiguous manner, further increasing this challenge.

Sequins are designed to emulate splicing of introns and exons and can therefore be used to assess the split alignment of reads across introns. Split reads derived from the RNA standards are aligned to both the in silico and natural chromosome. Split alignments on the in silico chromosomes are then be compared to known gene models to assess the sensitivity and specificity with which reads align across introns.

Anaquin takes an alignment file (SAM/BAM) to compare it with a reference in-silico chromosome. It calculates the specificity, sensitivity and sensitivity at the exon level, intron level and base level. Please refer to (?) for the definitions.

The rna flag specifies the RNA mode. -align indicates we want to calculate statistics for RNA alignment. The -o flag specifies the output directory for the calculations, default is spike\_out if not specified. The last argument is the alignment file. Anaquin supports both SAM and BAM file format.

$ anaquin rna -o rna\_output -align K562\_1A/RK1A.**bam**

$ anaquin rna -o rna\_output -align K562\_2A/RK2A.**bam**

$ anaquin rna -o rna\_output -align K562\_3A/RK3A.**bam**

OR

$ anaquin rna -o rna\_output -align K562\_1A/RK1A.**sam**

$ anaquin rna -o rna\_output -align K562\_2A/RK2A.**sam**

$ anaquin rna -o rna\_output -align K562\_3A/RK3A.**sam**

The output files are saved in the rna\_ouput folder. We should get the following files:

$ ls rna\_output

ralign\_base.stats

ralign\_exons.stats

ralign\_intron.stats

The files correspond to metric result at each level.

$ cat spike\_out/ralign\_exon.stats

The file lists the specificity (sn), sensitivity (sp) and sensitivity (ss). The sensitivity is about 0.31, this is the lowest quantity of abundance that can still be detected. The file also lists number of counts for each sequin.

$ cat spike\_out/ralign\_intron.stats

$ cat spike\_out/ralign\_base.stats

Similar interpretation applies for the intron and base level.

**Box ? | Metric Definitions**

* **Sensitivity:** measures the proportion of actual positives which are correctly identified as such.
* **Specifically:** measures the proportion of negatives which are correctly identified as such.
* **Detection limit**: the highest abundance RNA standard that is not reliably detected within the sequenced library and is without overlapping alignments.
  + - **Reads to Genome/In silico Chromosome:** The number of reads that align to the in silico chromosome (Reads To ChrT) and the human genome (Reads to Hg19). For K562, we aligned 1,091,683 reads to the in silico chromosome and 65,778,796 reads to the human genome sequence.
    - **Fraction Dilution:** Fraction of reads aligning to the in silico chromosome relative to the genome indicates the dilution of the standards relative to the sample library (Fraction Dilution). For K562 sample, 1.63% of library aligns to the in silico chromosome, indicating a 61-fold dilution factor.
    - **Alignment Sensitivity:** The number of in silico gene bases of the gene loci encoded on the in silico chromosome with alignments (true positive) divided by the total number of in silico gene bases. For K562 samples, we observe an alignment sensitivity of 0.81
    - **Alignment Specificity:** The number of in silico gene bases with alignments divided by the total number of bases with alignments. For K562 samples (Figure 4.2.6), we observe an alignment specificity of 0.83.
    - **Spliced Alignment Sensitivity:** The number of in silico gene introns with correct split alignments divided by the total number of in silico gene introns. For K562 samples, we observe an alignment sensitivity of 0.86.
    - **Spliced Alignment Specificity:** The number of in silico gene introns matching split alignments divided by the number unique split alignments. For K562 samples, we observe an alignment specificity of 0.85 (Figure 22A,B).
    - **Detection Limit:** The highest abundance RNA standard that is not reliably detected within the sequenced library and is without overlapping alignments. We determine a lower limit of detection at 0.005 attamoles/ul (the highest abundance RNA standard R\_8\_2 (SEQ ID NOS 47,48) not detected multiplied by dilution factor). Isoforms within the corresponding K562 RNA sample that are below this concentration may not be represented or detected within the sequencing library, and library sequencing has not proceeded to total saturation.

Alternative splicing, transcription initiation and termination generate a range of isoforms from single gene loci. We can use RNA standards to assess the accuracy with which spliced and unspliced alignments are assembled into full-length transcript models. Firstly, we assemble full-length transcript isoforms from overlapping read alignments on both the *in silico* and natural chromosomes. We can then compare the structure of RNA transcripts assembled on the *in silico* chromosome against known gene models to assess the sensitivity and specificity with which transcript assembly has occurred. This assessment can then inform the assembly of gene models in the accompanying natural sample.

We perform the following to assess novel gene discovery and transcript assembly: DNA/RNA standards can be added to a sample comprising a target DNA/RNA sequence to be determined. The source of target DNA/RNA can come from any known organism or environmental sample. For example, DNA/DNA standards can be added to natural RNA derived from animal (such as mammalian and human), plant (such as corn and rice), microbial (including bacteria and archaea) and environmental (soil samples, human stools, clinical samples such as infected wound fluid) sources.

Because RNA/DNA standards have no homology to natural sequences, sequenced reads derived from RNA/DNA standards can be distinguished from sequenced reads derived from natural RNA/DNA sample (**Figure 18**). This enables DNA/RNA standards to be added to the RNA/DNA sample, prior to sequencing, and therefore undergo the same library preparation, sequencing, alignment and analysis as for the DNA/RNA sample of interest.

RNA/DNA standards are combined with RNA/DNA samples to comprise only a fraction of the combined total. This fractional contribution (typically between 0.1 and 10% of total) varies according to the type of library preparations – rRNA removal, polyA or total RNA purification. Note that the fractional contribution of RNA/DNA standards is inversely proportion to the sequencing depth attributed to the RNA/DNA sample. Therefore, the fractional total should be the minimum amount required to sufficiently enable analysis of RNA/DNA standards.

**2.5 |** Create transcripts

We use the cufflinks package to generate transcripts from BAM files:

$ cufflinks -p 8 -G r\_silico.gtf -o K562\_1A /

K562\_1A/accepted\_hits.bam

$ cufflinks -p 8 -G r\_silico.gtf -o K562\_2A /

K562\_2A/accepted\_hits.bam

$ cufflinks -p 8 -G r\_silico.gtf -o K562\_3A /

K562\_3A/accepted\_hits.bam

**2.6 |** Check assembly results

$ ls K562\_1A/transcripts.gtf

$ ls K562\_2A/transcripts.gtf

$ ls K562\_3A/transcripts.gtf

**2.7 |** Calculate the assembly statistic

$ anaquin rna -o stats\_1A -assembly K562\_1A/transcripts.gtf

$ anaquin rna -o stats\_2A -assembly K562\_2A/transcripts.gtf

$ anaquin rna -o stats\_3A -assembly K562\_3A/transcripts.gtf

Once again, the output files are saved in the directories specified by the -o flag. Three files are generated for each replicate:

$ ls stats\_1A

assembly.base.stats

assembly.exons.stats

assembly.intron.stats

Similar to the alignment stage, the files correspond to each level of calculation: exon, intron and base.

$ cat spike\_out/assembly.exons.stats

The metrics are very similar to the RNA alignment. Sensitivity measures the proportion of exons that have been correctly predicted as exon. Specificity measures the proportion of predicted exons that are actually correct. Counts for each RNA sequin are also included.

**2.7 |** Generate differential expression

We use the cuffcompare package to generate differential expression:

**2.8 |** Calculate the abundance statistic

To assess the accuracy of quantification accuracy, we plot the measured gene abundance (in RPKM) relative to the known gene concentration (in attamoles/ul) of each in-silico gene (ie. Where alternative isoforms are combined). To provide an indication of quantitative accuracy, we report

* **Correlation** (Pearson's r) provides a bounded measurement for the linear relationship between observed and known gene abundance.
* **Slope** indicates the linear proportionality of observed compared to known abundance across the dynamic range of the RNA standards.

**! Important** While correlation and slope both measure relationship between the variables, they are no identical. However, the values are expected to be positively correlated in this context.

Calculate for the genes and isoforms:

$ anaquin rna -o stats\_1A -abundance K562\_1A/genes.fpkm\_tracking

$ anaquin rna -o stats\_2A -abundance K562\_2A/genes.fpkm\_tracking

$ anaquin rna -o stats\_3A -abundance K562\_3A/genes.fpkm\_tracking

OR

$ anaquin rna -o stats\_1A -abundance K562\_1A/isoforms.fpkm\_tracking

$ anaquin rna -o stats\_2A -abundance K562\_2A/isoforms.fpkm\_tracking

$ anaquin rna -o stats\_3A -abundance K562\_3A/isoforms.fpkm\_tracking

Examine the results (genes and isoforms):

$ cat stats\_1A/abundance.stats

$ cat stats\_2A/abundance.stats

$ cat stats\_3A/abundance.stats

r s r2 ss

0.83499 1.17772 0.691702 0.0190735

The correlation (r) , slope, R2 (r2) and sensitivity (ss) are reported. Anaquin also generates a R script that plots the variables. Run the script and generate a PDF file of the plot:

$ R CMD BATCH stats\_1A/abundance.R

$ R CMD BATCH stats\_2A/abundance.R

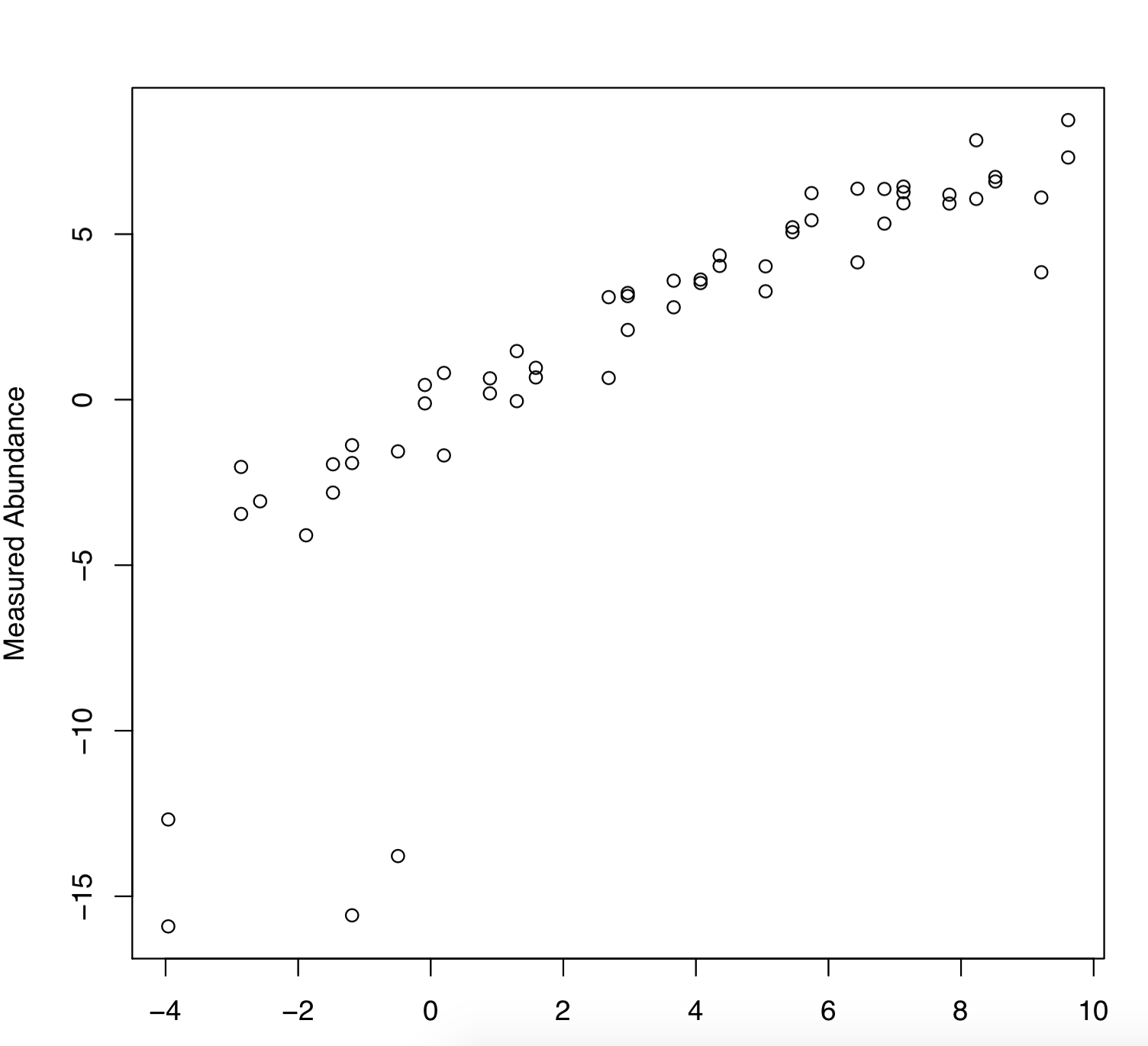
$ R CMD BATCH stats\_3A/abundance.R

$ ls stats\_1A/Rplots.pdf

$ ls stats\_2A/Rplots.pdf

$ ls stats\_3A/Rplots.pdf

Open the PDF with a PDF reader:



There is a good but not perfect linear relationship. The inaccuracy likely caused by insufficient sequencing coverage.

**2.9 |** Perform a differential expression test

Group the replicates in two sample groups to find significant changes in transcript expression.

$ cuffdiff -G rna\_silico.gtf \

K562\_1A/accepted\_hits.bam,\

K562\_2A/accepted\_hits.bam,\

K562\_3A/accepted\_hits.bam \

K562\_1B/accepted\_hits.bam,\

K562\_2B/accepted\_hits.bam,\

K562\_3B/accepted\_hits.bam

Compare with the known log fold-changes:

$ anaquin rna -o g\_output -diffs genes.fpkm\_tracking

$ anaquin rna -o i\_output -diffs isoforms.fpkm\_tracking

Examine differential expression:

$ cat g\_output/gene\_exp.diff

$ cat i\_output/isoform\_exp.diff

Anaquin also generates a R script for the plotting:

$ R CMD BATCH g\_output/differential.R

$ R CMD BATCH i\_output/differential.R

$ ls g\_output/Rplots.pdf

$ ls i\_output/Rplots.pdf

Open the PDF with a PDF reader:

# 3. DNA Sequencing

**3.1 |** Download the files for the tutorial

$ wget DNA\_Variant

$ unzip DNA\_Variant.zip

You should find

$ ls ./DNA\_Variant

simultated.bam

realGM12878.bam

simultated.vcf

realGM12878.vcf

simultated.mix

realGM12878.mix

**3.4 |** Calculate the alignment statistic

$ ./anaquin dna -al DNA\_Variant/tutorial\_sim.bam

Here is an explanation of metrics the anauqin output provides:

• **Reads to Genome/In silico Chromosome:** The number of reads that align to the in silico chromosome and the human genome. For example, for the GM12878 sample, we aligned 2,029,597 reads to the in silico chromosome and 458,521,347 reads to the human genome sequence.

**• Fraction Dilution**: The fraction of reads aligning to the in silico chromosome relative to the genome indicates the dilution of the standards relative to the sample library (Fraction Dilution). For GM12878 sample, 0.4% of library aligns to the in silico chromosome, indicating a 250-fold dilution factor.

**• Alignment Sensitivity:** The size of in silico DNA standard bases with overlapping alignments (true positive) divided by the total number of in silico DNA standard bases (true positive and false negative). For GM12878 samples, we observe a base-wise alignment sensitivity of 0.849.

**• Alignment Specificity**: The number of in silico DNA standard bases with overlapping alignments (true positive) divided by the total number of bases with overlapping alignments (true and false positive). For GM12878 samples, we observe a base-wise alignment specificity of 0.961.

• Detection Limit: The highest abundance DNA standard that are without read alignments and not reliably detected within the sequenced library. For GM12878 we observe a detection limit of 0.0037 attamoles/ul.

**3.4 |** Calculate the variant detection statistic

$ ./anaquin dna -var DNA\_Variant/tutorial\_sim.bam

Here is an explanation of metrics the anauqin output provides:

• **Variants Covered:** The proportion of genetic variation with alignment coverage. For example, alignments overlap 490 (88%) of variation instances in the DNA standards accompanying the GM12878 DNA sample.

• **Variant Sensitivity:** The number of variants correctly identified (true positive) divided by the total number of variants represented within the DNA standards (true + false negative). This depends both sequencing depth and variant detection. For example, for GM12878 sample, we achieve a variation sensitivity of 0.65.

• **Variant Detection:** The Variation Sensitivity divided by Variants Covered provides a measure of variant detection independent to sequencing depth or coverage. For example, for GM12878 sample, we achieve a variant efficiency of 0.73

• **Variant Specificity** the number of variants correctly identified (true positive) divided by the total number of variants detected (true positive + false negative). For example, for GM12878 sample, we achieve a variant specificity of 0.57.

• Median Quality Score A quality score, defined as the Phred scaled probability that a variant exists at this site, can be assigned to each identified variant. For the GM12878 sample, the median quality score for correct variant calls is 1,803, whilst the median quality score for erroneous variant calls is 61 (Figure 28E).

**COMMENT -** Given that the erroneous variant calls on the in silico chromosome exhibit lower quality score than correct calls (Figure 30A), this may indicate that some variant calls within the GM12878 genome are erroneous.

Descriptive statistics can be restricted to specific subsets of the variation represented within the DNA standards. For example, we can determine the sensitivity for detecting insertions within the DNA standards.

**3.5 |** Quantify allele frequencies:

$ ./anaquin dna -al DNA\_Variant/tutorial\_sim.vcf

Note that the correct DNA mixture file must be employed for accurate quantitication. Corresponding mixture files can be downloaded from …

# 4. Metagenome Sequencing

**4.1 |** Download the files for the tutorial

$ wget META

$ unzip META.zip

You should find

$ ls ./DNA\_Variant

vcf file

bam files