DNA Sequence Bioinformatics Analysis with the Galaxy Platform

University of São Paulo, Brazil
28 July - 1 August 2014

Dave Clements
Johns Hopkins University

Robson Francisco de Souza
University of São Paulo

José Belizario
University of São Paulo

The Week's Agenda

Mon	Introductions: Cloud Computing, Nuvem Cloud, Basic Analysis in Galaxy
Tues	Workflows, Sharing, Quality Control, ChIP-Seq
Wed	ChIP-Seq cont., Genome Assembly, RNA-Seq
Thur	RNA-Seq continued, SNP and Variant Calling
Fri	Intro to Command Line, Genome Annotation using MAKER, CloudMan and AWS

bit.ly/gxyusp2014

Wednesday's Agenda

9:00 Differential Expression Analysis with Tuxedo Suite of Tools

20 minute Break at around 10:20

10:40 Differential Expression Alternatives

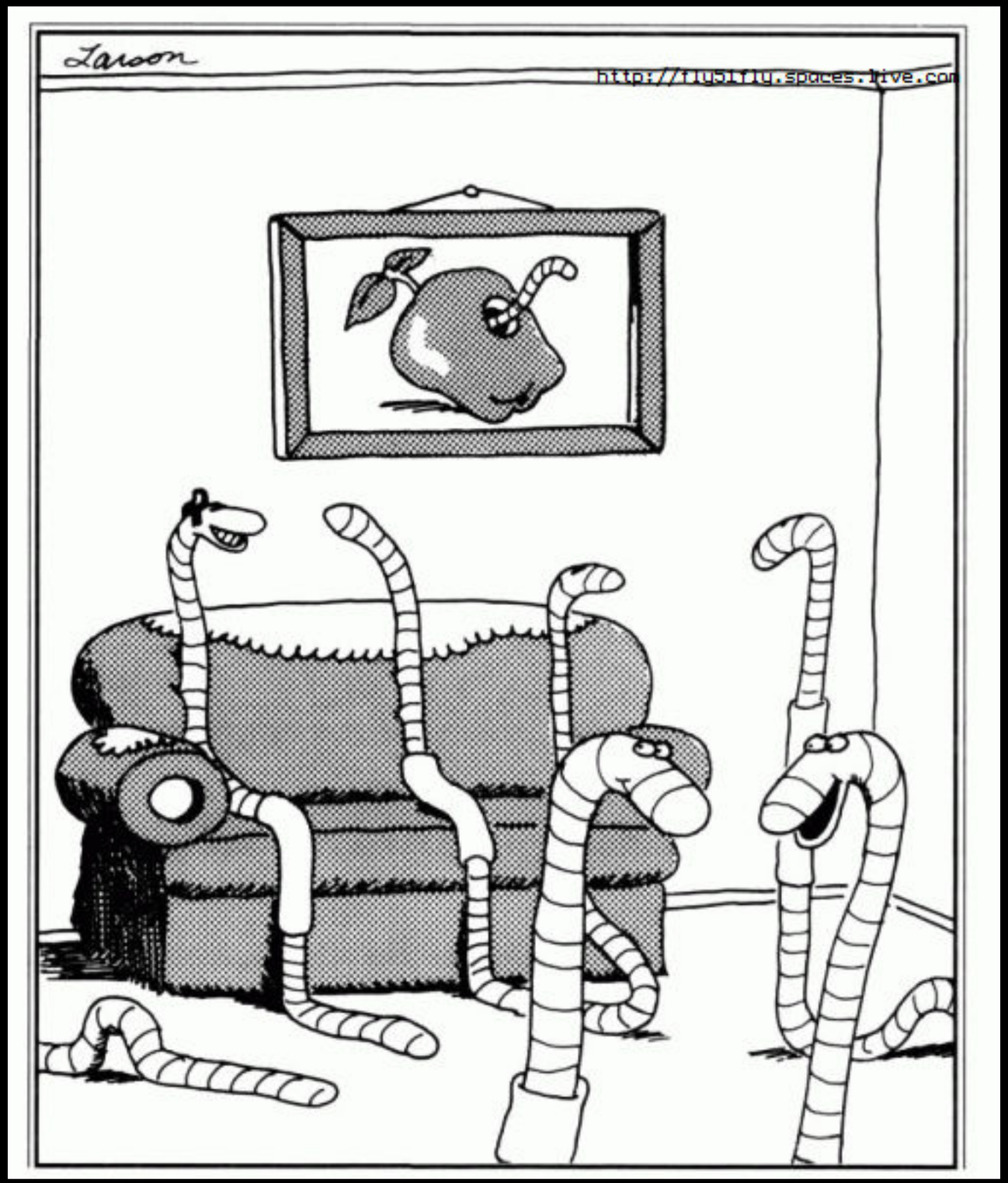
12:00 Lunch

2:00 Open Discussion and Q & A

2:15 SNP / Variant Analysis

20 minute Break at around 3:20

5:00 Done


Larson
http://flysfly.spaces.live.com

Tuxedo Suite: Some parts of the ensemble

Bowtie	Short read mapper. Bowtie2 can do gapped alignments and emphasizes reads > 50 bases
Tophat	Intron-aware mapper for RNA-Seq data. Works with Bowtie to find best mapping locations
Cufflinks	Construct transcript predictions from mapped reads (from Tophat output)
Cuffmerge	Merges multiple sets of transcript predictions into a unified set with one coherent set of IDs.
Cuffdiff	Differential expression analysis; Can work with Tophat output directly or Cufflinks/merge, if looking for novel genes/transcripts

Used already	Will discuss today	Will use today

Transcript Prediction: Cufflinks

- Cufflinks runs on Tophat output to assemble reads into transcripts
  - Tophat does not make any predictions about how the reads it mapped assemble together into transcripts.
- *Imagine pages and pages of discussion on the intricacies and pitfalls of RNA-seq transcript prediction here*

NGS: RNA Analysis → Cufflinks

Cufflinks: Min Isoform Fraction

Cufflinks can predict many different transcripts for a gene.

One transcript is likely to dominate.

Min Isoform Fraction tells Cufflinks to ignore any isoforms that fall below this level of expression, *relative to the dominant isoform*.

Higher values: less noise; less likely to report/discover low-expression transcripts.

Cufflinks: Pre mRNA Fraction

From the Cufflinks Manual

> "Some RNA-Seq protocols produce a significant amount of reads that originate from incompletely spliced transcripts, and these reads can confound the assembly of fully spliced mRNAs. Cufflinks uses this parameter to filter out alignments that lie within the intronic intervals implied by the spliced alignments. The minimum depth of coverage in the intronic region covered by the alignment is divided by the number of spliced reads, and if the result is lower than this parameter value, the intronic alignments are ignored. The default is 15%."

Basically, sets your tolerance for noise / novel constructs in intronic regions.

Cufflinks: Normalization and Correction

How hard should Cufflinks work to do the right thing?

Quartile Optimization: Attempt to compensate for skew caused by highly expressed genes

Bias Correction: Attempt to compensate for known issues with use of random hexamers in library preparation.*

Multi-Read Correct: Try to make reads that mapped to multiple locations more useful**

These optimizations often improve results. However, sample datasets may be too small for these to be effective.

* see Kasper D. Hansen, Steven E. Brenner, Sandrine Dudoit, Biases in Illumina transcriptome sequencing caused by random hexamer priming Nucleic Acids Research, Volume 38, Issue 12 (2010)

** see http://cufflinks.cbcb.umd.edu/howitworks.html#hmul

Cufflinks: Reference Annotation

How biased should we be, based on what we already know?

Reference Annotation: Use the reference annotation as dogma. Only doing quantification of known transcripts

Reference Annotation as Guide: Take advantage of what we already know, but be open to novel transcripts, if there is sufficient evidence

No: Transcript prediction will be based entirely on mapped reads in this dataset.

Transcript Prediction: Cuffmerge

- Each Cufflinks run creates a set of transcript predictions.
- Cuffmerge unifies all those predictions into a single set.
- Makes this incredibly tedious task easy.

Cuffdiff

- Identifies differential expression between multiple datasets
- Widely used and widely installed on Galaxy instances

NGS: RNA Analysis → Cuffdiff

Cuffdiff

Cuffdiff uses FPKM/RPKM as a central statistic.

Total # mapped reads heavily influences FPKM/RPKM. Can lead to challenges when you have very highly expressed genes in the mix.

Cuffdiff supports several FPKM/RPKM normalization techniques

Cuffdiff

- Running with 2 conditions: MeOH and R3G
- Each group has 3 replicates each

Cuffdiff

- Which Transcript definitions to use?
  - Official
  - MeOH or R3G Cufflinks transcripts
  - Results of Cuffmerge on MeOH & R3G Cufflinks transcripts
- Depends on what you care about

NGS: RNA Analysis → Cuffdiff

Cuffdiff

- Produces many output files, all explained in doc
- We'll focus on "gene differential expression testing" file (4th from the top)
- Column 7 ("status") can be FAIL, NOTEST, LOWDATA or OK
- Filter and Sort → Filter
  - c7 == 'OK'
  - Column 14 ("significant") can be yes or no
- c14 == 'yes'

Cuffdiff: differentially expressed genes

Column	Contents
test_stat	value of the test statistic used to compute significance of the observed change in FPKM
p_value	Uncorrected P value for test statistic
q_value	FDR-adjusted p-value for the test statistic

Wednesday's Agenda

9:00 Differential Expression Analysis with Tuxedo Suite of Tools

20 minute Break at around 10:20

10:40 Differential Expression Alternatives

12:00 Lunch

2:00 Open Discussion and Q & A

2:15 SNP / Variant Analysis

20 minute Break at around 3:20

5:00 Done

Cuffdiff Alternatives

Rapaport, *et al*., "Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data."
Genome Biology 2013, 14:R95 doi:10.1186/gb-2013-14-9-r95

Reviews 7 packages

Cuffdiff itself supports 3 possible normalization methods
http://cufflinks.cbcb.umd.edu/manual.html#library_norm_meth

Each has it's own strengths and weaknesses.

What's a biologist to do?

Alternatives: What's a biologist to do?

Learn the strengths and weaknesses of the tools you have ready access to. Are they a good match for the questions you are asking?

If not, then research alternatives, identify good options and then work with your bioinformatics/systems people to get access to those tools.

Cuffdiff Alternatives: DESeq

DESeq is an R based differential expression analysis package where expression analysis is much more effectively isolated between features.

Cuffdiff Alternatives: DESeq

Takes a simple, tab delimited list of features and read counts across different samples.

First, have to create that list.

htseq-count

Is a tool that walks BAM files producing these lists

Cuffdiff Alternatives: DESeq

NGS: SAM Tools → htseq-count
once for each BAM file

Join the HTSeq datasets together on gene name

Cut out the duplicate gene name columns

OR, just use the 6x DESeq Prep workflow

DESeq → DESeq2

Cuffdiff Alternatives: DESeq

DESeq output is a list of genes,
sorted by adjusted P value,
with lowest P values listed first

How many genes have an adjusted P value < 0.05 ?

Differential Expression: Reading & Resources

Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data
by Rapaport, *et al.*

DESeq Reference Manual

DESeq Galaxy Wrapper
by Nikhil Joshi

htseq-count Galaxy Wrapper
by Lance Parsons

Wednesday's Agenda

9:00 Differential Expression Analysis with Tuxedo Suite of Tools

20 minute Break at around 10:20

10:40 Differential Expression Alternatives

12:00 Lunch

2:00 Open Discussion and Q & A

2:15 SNP / Variant Analysis

20 minute Break at around 3:20

5:00 Done

Wednesday's Agenda

9:00 Differential Expression Analysis with Tuxedo Suite of Tools

20 minute Break at around 10:20

10:40 Differential Expression Alternatives

12:00 Lunch

2:00 Open Discussion and Q & A

2:15 SNP / Variant Analysis

20 minute Break at around 3:20

5:00 Done

Wednesday's Agenda

9:00 Differential Expression Analysis with Tuxedo Suite of Tools

20 minute Break at around 10:20

10:40 Differential Expression Alternatives

12:00 Lunch

2:00 Open Discussion and Q & A

2:15 SNP / Variant Analysis

20 minute Break at around 3:20

5:00 Done

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Cloud | Help | User

Published Pages | galaxyproject | Galaxy 101 NGS: Introduction to Polymorphism Detection via Variant Analysis

Galaxy 101 NGS: Introduction to Polymorphism Detection via Variant Analysis

Heteroplasmy: Mother-Child mtDNA Variant Polymorphism

• heteroplasmy • ismb2010-demo

This tutoria

• Import Sequence
• Interpret FASTQ t
• Execute a series
• Execute one com
• Filter key results.

What is Hete

The heteroplastom
mtDNA heteroplast
research in the fiel

Being heteroplasto

• Go ahead, start w

Experiment

Import child and m
(*where Q20 indica
Convert the resulti
and FreeBayes. Not
Variant Annotator

• Explore the resul

• Questions:

  • Can you ident
  • Can you ident
    How could thi
  • Are any SNPs
    Do any polym
  • How would yo
    mtDNA (per ir

* Source at Illumina

Input NGS Dat

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Cloud | Help | User

Published Pages | aun1 | heteroplasmy

Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study

Hiroki Goto[1], Benjamin Dickins[2], Enis Afgan[3,5], Ian M. Paul[4], James Taylor[3,5], Kateryna D. Makova[1], and Anton Nekrutenko[2,5]

Published in Genome Biology on June 23, 2011

Correspondence should be addressed to KDM, JT, or AN.

1. How to

This document is
with them by re-
hassle-free proce

access our datase
re-use workflows
view and import
In addition, we cr

Watch the analysi
Watch how the co
If you experience

2. Access

All datasets discu

A Galaxy Library
An S3 bucket on

From there these
family is "F4", "F7
replicate 1 from
table = count of

F1

F2 M9 blood 6,106,21

Goto *et al. Genome Biology* 2011, **12**:R59
http://genomebiology.com/2011/12/6/R59

Genome Biology

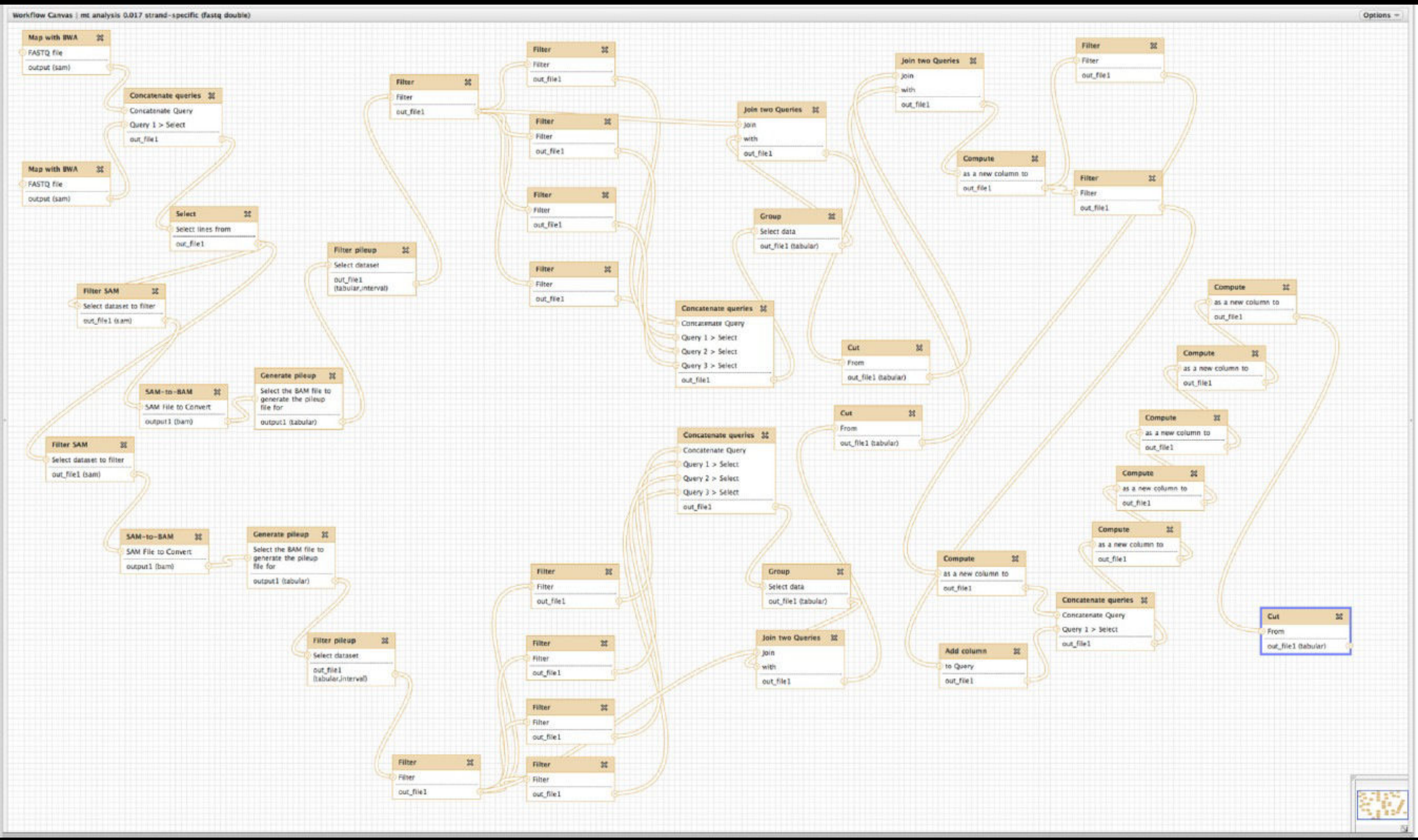
RESEARCH Open Access

Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study

Hiroki Goto[1†], Benjamin Dickins[2†], Enis Afgan[3], Ian M Paul[4], James Taylor[3*], Kateryna D Makova[1*] and Anton Nekrutenko[2*]

Abstract

Background: Originally believed to be a rare phenomenon, heteroplasmy - the presence of more than one mitochondrial DNA (mtDNA) variant within a cell, tissue, or individual - is emerging as an important component of


Workflow Canvas | mt analysis 0.017 strand-specific (fastq double)
Options
Map with BWA
FASTQ file
output (sam)
Concatenate queries
Concatenate Query
Query 1 > Select
out_file1
Map with BWA
FASTQ file
output (sam)
Select
Select lines from
out_file1
Filter SAM
Select dataset to filter
out_file1 (sam)
SAM-to-BAM
SAM File to Convert
output1 (bam)
Generate pileup
Select the BAM file to generate the pileup file for
output1 (tabular)
Filter pileup
Select dataset
out_file1 (tabular,interval)
Filter SAM
Select dataset to filter
out_file1 (sam)
SAM-to-BAM
SAM File to Convert
output1 (bam)
Generate pileup
Select the BAM file to generate the pileup file for
output1 (tabular)
Filter pileup
Select dataset
out_file1 (tabular,interval)
Filter
Filter
out_file1
Filter
Filter
out_file1
Filter
Filter
out_file1
Filter
Filter
out_file1
Filter
Filter
out_file1
Filter
Filter
out_file1
Filter
Filter
out_file1
Filter
Filter
out_file1
Join two Queries
Join
with
out_file1
Group
Select data
out_file1 (tabular)
Concatenate queries
Concatenate Query
Query 1 > Select
Query 2 > Select
Query 3 > Select
out_file1
Concatenate queries
Concatenate Query
Query 1 > Select
Query 2 > Select
Query 3 > Select
out_file1
Group
Select data
out_file1 (tabular)
Join two Queries
Join
with
out_file1
Cut
From
out_file1 (tabular)
Cut
From
out_file1 (tabular)
Join two Queries
Join
with
out_file1
Filter
Filter
out_file1
Compute
as a new column to
out_file1
Filter
Filter
out_file1
Compute
as a new column to
out_file1
Compute
as a new column to
out_file1
Compute
as a new column to
out_file1
Compute
as a new column to
out_file1
Compute
as a new column to
out_file1
Compute
as a new column to
out_file1
Add column
to Query
out_file1
Concatenate queries
Concatenate Query
Query 1 > Select
out_file1
Cut
From
out_file1 (tabular)

Variant analysis

Goal is to find variation in these individuals' mitochondria

On your Nuvem cloud instance

Start a new history

Shared Data → Data Libraries → Polymorphism ...

Import both Child datasets and mother dataset 1

then

Import Mother dataset 2

paired end datasets from mother and child

Variant analysis

Could use FastQC to look at quality

But, that would merely alarm us!

Map mother and child paired end datasets using

NGS Mapping → Map with BWA for Illumina

Variant analysis

Filter the SAM files to keep only properly mapped pairs

Flag 1: Read is Paired: Yes

Flag 2: Read is mapped in a proper pair: Yes

Convert filtered SAM files to BAM files

Variant analysis

add in read groups with Picard:

NGS SAM Tools -> Add or Replace Groups

Read group ID (ID tag):	child	mother
Read group sample name (SM tag):	child	mother
Read group library (LB tag):	child	mother
Read group platform (PL tag):	illumina	illumina
Read group platform unit:	bc	bc

Variant analysis

Merge the two BAM Files

NGS SAM Tools -> Merge BAM Files

We now have all mapped reads in one BAM file.
Can distinguish mother and child by read group.

Let's do some Variant analysis

NGS Variant Analysis -> Naive Variant Analysis

10 Minimum number of reads need to consider a REF/ALT
20 Minimum base quality
1 Ploidy
x Only write out positions with possible alternate alleles
x Report counts by strand

http://bit.ly/naiveVariants

Let's do some Variant analysis

NGS Variant Analysis -> Variant Annotator

1.0	Minor Allele Frequency threshold (in percent)
10	Coverage threshold (in reads per strand)
False	Do not filter sites or alleles
True	Output stranded base counts
True	Write header line

Let's do some Variant analysis

NGS Variant Analysis -> FreeBayes

Reference from Local cache
Sample BAM: Child-Mother
Limit to region: chrM 1000-10000

Simple Diploid calling

Let's do some Variant analysis

Filter and Sort -> Filter

c16 >= 0.02
Skip 1 header line

Wednesday's Agenda

9:00 Differential Expression Analysis with Tuxedo Suite of Tools

20 minute Break at around 10:20

10:40 Differential Expression Alternatives

12:00 Lunch

2:00 Open Discussion and Q & A

2:15 SNP / Variant Analysis

20 minute Break at around 3:20

5:00 Done

Thanks

Dave Clements

Galaxy Project

Johns Hopkins University

outreach@galaxyproject.org

Variant analysis

SAM to BAM

http://bit.ly/1xEcR90

http://www.ncbi.nlm.nih.gov/nuccore/251831106?report=fasta