University of Massachusetts
UMASS Medical School

CRISPRseek Workshop

Design of target-specific guide RNAs in CRISPR-Cas9 genome-editing systems

Sept 10th 2014

Lihua Julie Zhu Michael Brodsky Jianhong Ou

INSTALLATION

- First install R 3.1.0
  - Windows: http://cran.fhcrc.org/bin/windows/base/
  - Mac OS X: http://cran.fhcrc.org/bin/macosx/
  - Source (Linux): http://cran.fhcrc.org/sources.html
- Then obtain Bioconductor 2.14 by starting R and entering the commands
  - **source("http://bioconductor.org/biocLite.R")**
  - **biocLite()**
- To install additional package, e.g., CRISPRseek, type:
  - **biocLite("CRISPRseek")**
- Installed at **ghpcc06 cluster** and

rstudio.umassmed.edu

AVAILABLE IN RSTUDIO SERVER AT UMASS

- CRISPRseek package has been installed in the Rstudio server at Umass.
- To access it, please send an email to the following email list.

  Tom.Weeks@umassmed.edu

  Alan.Ritacco@umassmed.edu

  UMWHelpdesk@umassmed.edu

  Also need to sign up for a HPCC account at https://ghpcc06.umassrc.org/hpc/index.php
- Once your access is granted, please validate that you indeed can access the Rstudio by login at rstudio.umassmed.edu using your email username and password in a web browser.

Run CRISPRseek in Your Own Installation

source("http://bioconductor.org/biocLite.R")

biocLite("CRISPRseek")

The human annotation packages need to be installed to run the example code in the vignette.

biocLite("BSgenome.Hsapiens.UCSC.hg19")

biocLite("TxDb.Hsapiens.UCSC.hg19.knownGene")

biocLite("org.Hs.eg.db")

GENOME SEQUENCE PACKAGES (68)

http://www.bioconductor.org/packages/release/BiocViews.html#___BSgenome

http://www.bioconductor.org/packages/release/bioc/vignettes/BSgenome/inst/doc/BSgenomeForge.pdf

BSgenome	Description
BSgenome.Hsapiens.UCSC.hg19	Full genome sequences for Homo sapiens (UCSC version hg19)
BSgenome.Hsapiens.UCSC.hg19.masked	Full masked genome sequences for Homo sapiens (UCSC version hg19)
BSgenome.Mmusculus.UCSC.mm10	Full genome sequences for Mus musculus (UCSC version mm10)
BSgenome.Mmusculus.UCSC.mm10.masked	Full masked genome sequences for Mus musculus (UCSC version mm10)
BSgenome.Dmelanogaster.UCSC.dm3	Full genome sequences for Drosophila melanogaster (UCSC version dm3)
BSgenome.Dmelanogaster.UCSC.dm3.masked	Full masked genome sequences for Drosophila melanogaster (UCSC version dm3)
BSgenome.Rnorvegicus.UCSC.rn5	Full genome sequences for Rattus norvegicus (UCSC version rn5)
BSgenome.Rnorvegicus.UCSC.rn5.masked	Full masked genome sequences for Rattus norvegicus (UCSC version rn5)
BSgenome.Drerio.UCSC.danRer7	Full genome sequences for Danio rerio (UCSC version danRer7)
BSgenome.Drerio.UCSC.danRer7.masked	Full masked genome sequences for Danio rerio (UCSC version danRer7)
BSgenome.Celegans.UCSC.ce10	Full genome sequences for Caenorhabditis elegans (UCSC version ce10)
BSgenome.Athaliana.TAIR.TAIR9	Full genome sequences for Arabidopsis thaliana (TAIR9)
BSgenome.Scerevisiae.UCSC.sacCer3	Saccharomyces cerevisiae (Yeast) full genome (UCSC version sacCer3)

TRANSCRIPT PACKAGES (18)

http://www.bioconductor.org/packages/release/BiocViews.html#___AnnotationData

http://bioconductor.org/packages/release/bioc/html/GenomicFeatures.html

txdb	Description
TxDb.Athaliana.BioMart.plantsmart21	Athaliana genes
TxDb.Celegans.UCSC.ce6.ensGene	Worm genes
TxDb.Dmelanogaster.UCSC.dm3.ensGene	Fly genes
TxDb.Hsapiens.UCSC.hg18.knownGene	Human hg18 genes
TxDb.Hsapiens.UCSC.hg19.knownGene	Human hg19 genes
TxDb.Hsapiens.UCSC.hg19.lincRNAsTranscripts	Human hg19 lincRNAs
TxDb.Mmusculus.UCSC.mm10.knownGene	Mouse mm10 genes
TxDb.Mmusculus.UCSC.mm9.knownGene	Mouse mm9 genes
TxDb.Rnorvegicus.UCSC.rn4.ensGene	Rat rn4
TxDb.Rnorvegicus.UCSC.rn5.refGene	Rat rn5
TxDb.Scerevisiae.UCSC.sacCer2.sgdGene	Yeast Cer2
TxDb.Scerevisiae.UCSC.sacCer3.sgdGene	Yeaset Cer3

To create additional TxDb object

makeTranscriptDbFromUCSC and makeTranscriptDbFromGFF

Genome Annotation Packages (19)

http://www.bioconductor.org/packages/release/BiocViews.html#___OrgDb

Genome Annotation Package	Genome	orgAnn
org.At.tair.db	Arabidopsis	org.At.tairSYMBOL
org.Bt.eg.db	Bovine	org.Bt.egSYMBOL
org.Ce.eg.db	Worm	org.Ce.egSYMBOL
org.Cf.eg.db	Canine	org.Cf.egSYMBOL
org.Dm.eg.db	Fly	org.Dm.egFLYBASE2EG
org.Dr.eg.db	Zebrafish	org.Dr.egSYMBOL
org.Gg.eg.db	Chicken	org.Gg.egSYMBOL
org.Hs.eg.db	Human	org.Hs.egSYMBOL
org.Mm.eg.db	Mouse	org.Mm.egSYMBOL
org.Mmu.eg.db	Rhesus	org.Mmu.egSYMBOL
org.Pt.eg.db	Chimp	org.Pt.egSYMBOL
org.Rn.eg.db	Rat	org.Rn.egSYMBOL
org.Sc.sgd.db	Yeast	org.Sc.sgdGENENAME
org.Ss.eg.db	Pig	org.Ss.egSYMBOL
org.Xl.eg.db	Xenopus	org.Xl.egSYMBOL

Install Common annotation packages in Your Own Installation

- source("http://bioconductor.org/biocLite.R")
- biocLite("BSgenome.Hsapiens.UCSC.hg19")
- biocLite("BSgenome.Mmusculus.UCSC.mm10")
- biocLite("BSgenome.Rnorvegicus.UCSC.rn5")
- biocLite("BSgenome.Drerio.UCSC.danRer7")
- biocLite("BSgenome.Dmelanogaster.UCSC.dm3")
- biocLite("BSgenome.Celegans.UCSC.ce6")
- biocLite("TxDb.Hsapiens.UCSC.hg19.knownGene")
- biocLite("TxDb.Mmusculus.UCSC.mm10.knownGene")
- biocLite("TxDb.Rnorvegicus.UCSC.rn5.refGene")
- biocLite("TxDb.Dmelanogaster.UCSC.dm3.ensGene ")
- biocLite("TxDb.Celegans.UCSC.ce6.ensGene")
- biocLite("org.Hs.eg.db")
- biocLite("org.Mm.eg.db")
- biocLite("org.Dm.eg.db")
- biocLite("org.Ce.eg.db")
- biocLite("org.Dr.eg.db")
- biocLite("org.Rn.eg.db")

Request to Install Additional Packages In UMass Servers

- rstudio.umassmed.edu
  - Alper.Kucukural@umassmed.edu
- ghpcc06
  - hpcc-support@umassmed.edu

Main Functions of CRISPRseek

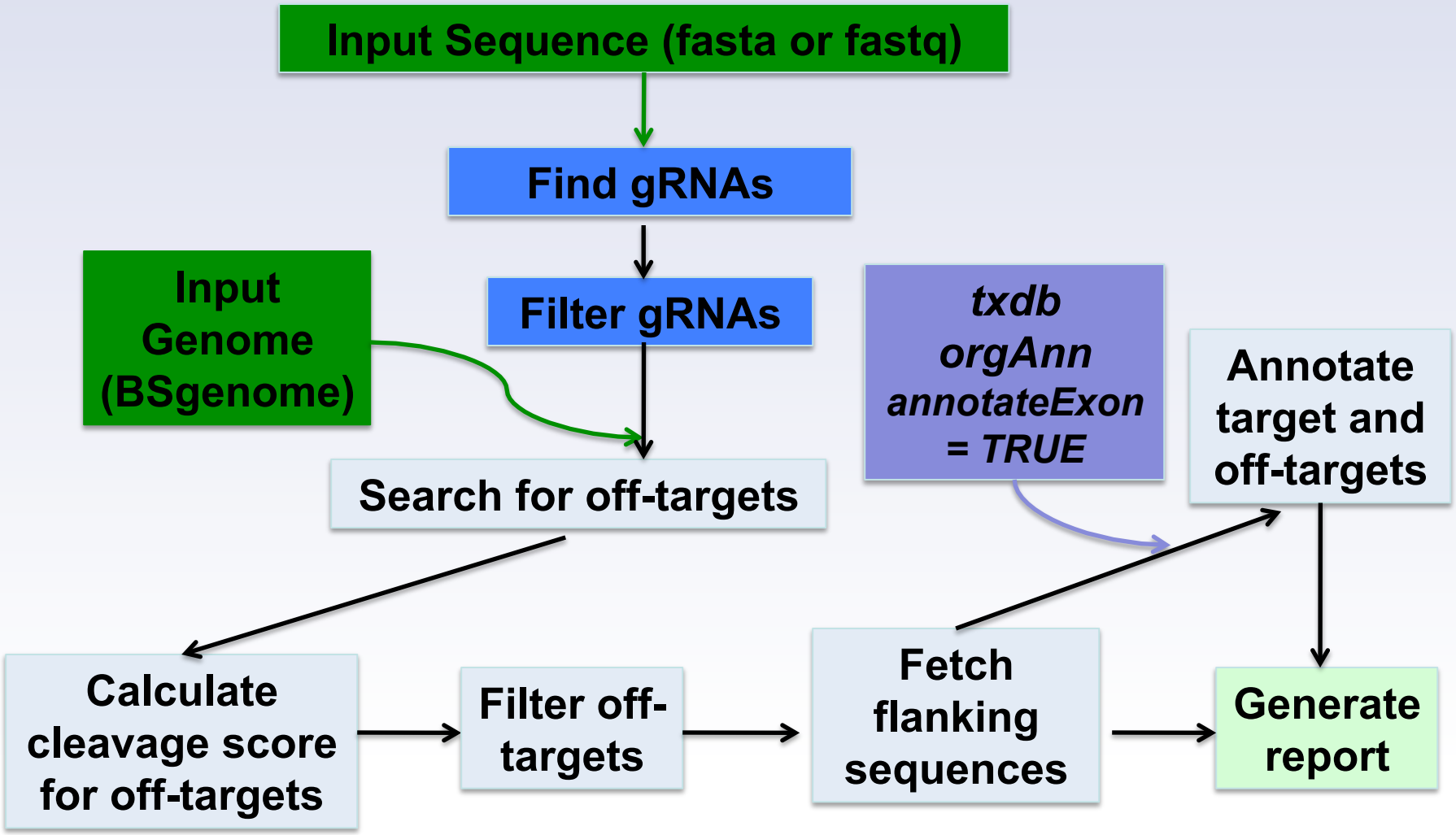
offTargetAnalysis workflow

- gRNA searching and off-target analysis for one or a set of input sequences

compare2Sequences workflow

- Identify gRNAs that specifically target one of the two input sequences or both

offTargetAnalysis



DEFAULT PARAMETERS *OFFTARGETANALYSIS*

- ***offTargetAnalysis****(inputFilePath, format = "fasta", findgRNAs = TRUE, exportAllgRNAs = c("all", "fasta", "genbank", "no"), findgRNAsWithREcutOnly = TRUE, REpatternFile, minREpatternSize = 6, overlap.gRNA.positions = c(17, 18), findPairedgRNAOnly = TRUE, min.gap = 0, max.gap = 20, gRNA.name.prefix = "gRNA", PAM.size = 3, gRNA.size = 20, PAM = "NGG", BSgenomeName, chromToSearch = "all", max.mismatch = 4, PAM.pattern = "N[A|G]G$", gRNA.pattern = "", min.score = 0.5, topN = 100, topN.OfftargetTotalScore = 10, annotateExon = TRUE, txdb, orgAnn, outputDir, fetchSequence = TRUE, upstream = 200, downstream = 200, weights = c(0, 0, 0.014, 0, 0, 0.395, 0.317, 0, 0.389, 0.079, 0.445, 0.508, 0.613, 0.851, 0.732, 0.828, 0.615, 0.804, 0.685, 0.583), overwrite = FALSE)*

Default setting for CRISPR-cas9 system in S. pyogenes

guide sequence /gRNA in CRISPRseek

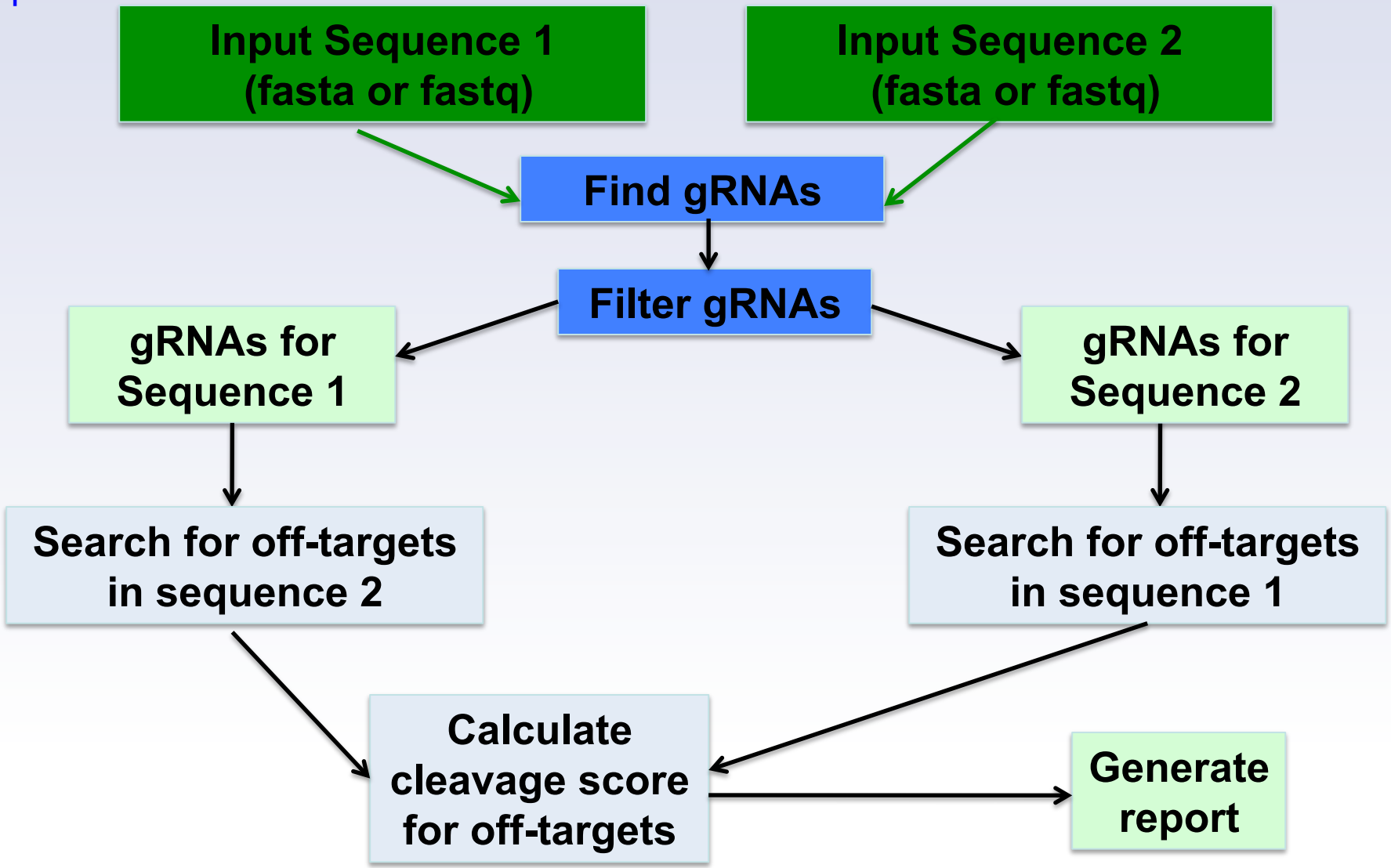
PAM sequence

CCACTGTGTGCACTTCATCCTGG

OUTPUT FILES OF OFFTARGETANALYSIS

- gRNAs (genbank, fasta)
- Restriction Site Overlap (tab delimited)
- Paired Sites (tab delimited)
- Off-Target sites (tab delimited)
- Summary (tab delimited)

compare2Sequences: Identify gRNAs that specifically target one of the two sequences or both



DEFAULT PARAMETERS COMPARE2SEQUENCES

- *compare2Sequences*(inputFile1Path, inputFile2Path, format = "fasta", findgRNAsWithREcutOnly = FALSE, REpatternFile, minREpatternSize = 6, overlap.gRNA.positions = c(17, 18), findPairedgRNAOnly = FALSE, min.gap = 0, max.gap = 20, gRNA.name.prefix = "gRNA", PAM.size = 3, gRNA.size = 20, PAM = "NGG", PAM.pattern = "N[A|G]G$", max.mismatch =4, outputDir, weights = c(0, 0, 0.014, 0, 0, 0.395, 0.317, 0, 0.389, 0.079, 0.445, 0.508, 0.613, 0.851, 0.732, 0.828, 0.615, 0.804, 0.685, 0.583), overwrite = FALSE)

args(compare2Sequences)

Output Files

- Scores For 2 Input Sequences (tab delimited)
- gRNAs (genbank, fasta)
- Restriction Site Overlap (tab delimited)
- Paired Sites (tab delimited)

Tuning of Maximum Mismatch Allowed

- ***compare2Sequences*** might output no match or more than one match to the alternative input sequence for each gRNAs identified for each input sequence depending on max.mismatch (default is 4 mismatches allowed)
  - Solution
    - Sort the output by gRNAplusPAM and scoreDiff to examine possible multiple off-target sites in the alternative sequence, if you aim to identify gRNAs to target one of the two input sequences only.
    - Tune max.mismatch

REFERENCE AND HELP

- http://www.bioconductor.org/packages/release/bioc/vignettes/CRISPRseek/inst/doc/CRISPRseek.pdf
- http://www.bioconductor.org/help/course-materials/2014/BioC2014/CRISPRseek-forBioc2014.pdf
- http://www.bioconductor.org/help/course-materials/2014/BioC2014/CRISPRdemo.Rmd
- In a R session
  - browseVignettes("CRISPRseek")
  - ?offTargetAnalysis
  - ?compare2Sequences
- Zhu LJ*, Holmes BR, Aronin N and Brodsky MH*. (2010) [* denotes co-corresponding author] CRISPRseek: a Bioconductor package to identify target-specific guide RNAs for CRISPR-Cas9 genome-editing systems. PloS One (In press)
- Email: bioconductor bioconductor@stat.math.ethz.ch

DEMO & EXERCISE