BioConductor

- Agenda

Module 1 Introduction to bioinformatics

What is bioinformatics

Types of online databases and online tools

Types of publications & literature

Types of datasets (biocviews)

Module 2 Introduction to bioConductor

BioConductor website

installing BioConductor in R

Bioconductor packages, vignettes, tutorials

- Agenda

  Module 3 Databases in BioConductor

  Annotation Hub

  biomarts (online biomarts)

  Genomic (full genomes)

  GEOquery (NCBI Gene expression Omnibus)

  KEGG (Kyto Encyclopedia for Genes and Genomes)

  Gene libraries

- Agenda

  Module 4 Data files in BioConductor

  data containers

  3 types of data (Biobase package)

  rtracklayer package

  rsamtools package

  Module 5 ranges into BioConductor

  Iranges

  Granges

  Granges plotting

  Biostrings package

- Agenda

  Module 6 Next Generation Sequencing (NGS)

  what is NGS

  applications of NGS

  shortreads package

  Module 7 Microarrays (DNA Chip)

  what is Microarray technology

  applications of microarrays

  oligo package

  Module 8 Workflows (optional)

  what are workflows

  sequencing workflows

Module 1

Introduction to Bioinformatics

- What is bioinformatics ?

  Computational branch of molecular biology

  this includes analyzing DNA sequence, analyzing RNA sequences, analyzing protein sequences, working with protein 3D structures, working with entire genomes, medical publications and much more...

- Types on online databases

  Ensembl http://www.ensembl.org/ (human/mouse genome)

- USCS https://genome.ucsc.edu/ (human genome)
- Genbank http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide (nucleotide sequences)
- PubMed http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed (literature references)
- NR http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=protein (non redundant protein sequences)
- Swiss-Prot http://www.expasy.ch (protein sequences)
- InterPro htttp://www.ebi.ac.uk (protein domains)
- OMIM http://www.ncbi.nlm.nih.gov/entrez/
- Enzymes http://www.chem.qmul.ac.uk
- PDB http://www.rcsb.org/pdb/ (protein structures)
- KEGG http://www.genome.sd.jp (metabolic pathways)

- Types of online software

SRS http://www.srs.ebi.sc.uk (Database search)

BLAST http://www.ncbi.nlm.nih.gov/blast (homology search)

DALI http://www.ebi.ac.uk/dali (structure database)

ClustalW http://www.ebi.ac.uk/clustalw (multiple seq alignment)

MUSCLE http://phylogenomics.berkley.edu/muscle(multi seq)

GenScan http://genes.mit.edu (gene predict)

psiPred http://bioinf.cs.usl.ac.uk/psipred (protein predict)

Mfold http://www.bioinfo.rpi.edu/applications/mfold/ (RNA predict)

phylip http://bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html

(phylogeny – tree reconstruction)

PhyML http://atgc.lirmm.fr/phyml/ (phylogeny – tree reconstruct)

Jalview http://www.jalview.org (alignment editor)

- Types of online resources

  Expasy www.expasy.cn (proteins)

  ArrayExpress www.ebi.ac.uk/microarray/ (microarrays)

  Swbic ww.swbic.org/ (misc links)

  Pasteur bioweb-pasteur.fr/into-uk.html (many online tools)

  RNA www.imb-jena.de/RNA.html (RNA – related link)

  miRNAs microrna.sanger.ac.uk/sequences/index.shtml (miRNA links)

  phylip evolution.genetics.washington.edu/phylip/software.html (everything on phylogeny)

  NCBI primers www.ncbi.nlm.nih.gov/education

  Bieliefield bibiserv.techfak.uni-bielefeld.de/intro/dist.html (eCourse)

  Bio-informer www.ebi.ac.uk/Information/News/ (EBI news)

Module 2

Introduction to Bioconductor

- What is Bioconductor?

  Bioconductor is a free, open source and open development software project for the analysis and comprehension of genomic data generated by wet lab experiments in molecular biology.

- Bioconductor is based primarily on the statistical R programming language, It has two releases each year that follow the semiannual releases of R.

- there are a large number of genome annotation packages available that are mainly, but not solely, oriented towards different types of microarrays.

- Install bioconductor

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

- Start bioconductor

  biocLite( ) → start installing packages.updating

- Checking packages

  biocValid ( ) → see if packages mist up to date

- Installing packages

  biocLite("GenomicRanges") → installs this package

- Loading the package

  library(GRanges)

- Reading documentation

  browseVignettes(package=”GenomicRanges”)

- To upgrade bioconductor version

  biocLite(“BiocUpgrade”)

Module 3

Querying databases using Bioconductor

- Types of databases

- **Annotation Hub** (different data resources)
  different genomes, cell lines etc
- **BiomaRt** (interface with online biomarts)
  containing different databases
- **BSGenome** (full genomes)
  list of available genomes
- **GEOquery** (interface with NCBI Gene omnibus)
  repository for public data
- **KEGGREST** (kyoto encyclopedia for genes/genomes)
  databases and images for metabolic pathways
- **Databases** (data on entire gene libraries)
  searchable w keys

Annotation Hub

- Install and load the database metadata

  biocLite("AnnotationHub")

  library(AnnotationHub)

- ?AnnotationHub (get documentation)

Annotation Hub : Metadata

- ah=AnnotationHub( )

  ah (view all the data)

- ah[1] (view just the first dataset)
- ah[[1]] (download the first dataset)

- unique(ah$dataproviders) (see dataproviders)
- unique(ah$species) (see species data)

Annotation Hub: query database

query(ah, “H3K4me3”)

- (look at data about this histone)

- query(ah, “H3K4me3”, “Gm12878”)
- (look at histone data in this cell type)

- **Annotation Hub: Display the data**

  (need shiny package)

- ah2=display(ah)
-
- Select the rows you are interested in, click “return rows to R session”

- Data will be stored in variable ah2

Challenge

- Use the Annotation Hub package to obtain data on “CpG islands” on human genome

- How many islands exists on autosomes
- How many islands exist on chromosome 5

- In the human genome reference build hg19, what is the length of chr 16?

- **Biomart**

- Install and load package

  biocLite(“biomaRt”)

  library(biomaRt)

- ??biomaRt

- **Biomart : querying databases and datasets**

- head(listMarts( ))

  (list down the first 6 datamarts)

- mart=useMart(“ensembl”)

  we use the emsembl data mart

  (www.ensembl.org)

- head(listDatasets(mart))

  (list down first 6 dataset)
- ensembl=useDataset(“hsapiens_gene_ensembl”, mart)
- (choose this dataset)

- **BSGenome**

- Install and load packages
- biocLite(“BSgenome”)
- library(BSGenome)

- ??BSGenome

- BSGenome – genome lists

- available.genomes( )

  (list of all available genomes in Bioconductor)

- Installed.genomes( )

  (list of genomes installed in your computer)

- **BSGenome** – installing genomes

- biocLite(“ <genome name>”)

  biocLite(“BSGenome.Scerevisiae.UCSC.sacCer1”)

  (installing yeast genome)

- library((“BSGenome.Scerevisiae.UCSC.sacCer1”)

  (loading genome in the library)

- **GEOquery**

- Installing and loading package

  biocLite("GEOquery")

  library(GEOquery)

  ??GEOquery

- Used to query gee expression (microarrays)

- **GEOquery** – get the data

- eList=getGEO(“GSE11675”)
  (download this microarray set)

- Look at the data
  length(eList)
  names(eList)

- eData=eList[[1]]
  eData

- **KEGG**

- biocLite(“KEGGREST”)
- library(“KEGGREST”)

- ??KEGGREST

- **KEGG** – view biological pathways

- library(png)
- library(grid)

- brpng=keggGet(“hsa05212”, “image”)
  grid.raster(brpng)
  (biological pathway of pancreatic cancer)

- **Databases** – gene libraries

- biocLite(“org.Hs.eg.db”)
- library(org.Hs.eg.db)
  (getting homo sapien library) - human

- class(org.Hs.eg.db)

- **Databases** – using keys to search library

  org.Hs.eg.db

  keytypes(org.Hs.eg.db)

  columns(org.Hs.eg.db)

- **Databases** – querying

- Search pubMed for ORMDL3 gene

  select(org.Hs.eg.db, keys=”ORMDL3”, keytype=”SYMBOL”, columns=”PMID”)

- Seach info about the same gene

  select(org.Hs.eg.db, keys=”ORMDL3”, keytype=”SYMBOL”, columns=”GO”)

Challenge

- Obtain data from org.Hs.eg.db about

- BRCA1
- LSM1

- Get useful info like the gene name, etc

Module 4

Datafiles in Bioconductor

- **Expression set** – analysis using ALL dataset

- The data in an ExpressionSet is complicated, consisting of expression data from microarray experiments

- biocLite(“ExpressionSet”)

  biocLite(“ALL”)

  library(ALL)

  ?ALL

  ?ExpressionSet

- **Expression set** – summary and metadata

- slotNames(ALL) (look at the types of headers)
- data(ALL) (summary stats)
- phenoData (look at the phenotype data)

- **Summarized Experiment** – using airways dataset

- The SummarizedExperiment container contains one or more assays, each represented by a matrix-like object of numeric or other mode. The rows typically represent genomic ranges of interest and the columns represent samples

- biocLite(“SummarizedExperiment”)

  biocLite(“airways”)

  library(airways)

  ?airways

  ?SummarizedExperiment

- **Biobase package**

- biocLite(“Biobase”)
  library(“Biobase”)

  exprs( ) - genetic expression
  pData( ) - phenotype of samples
  fData( ) - genotype of samples

- **Rtracklayer package**

- Used to import unique datatype files (WIG, GTF, BED)

  install package and load library
- biocLite(“rtracklayer”)

  library(rtracklayer)

- **Rsamtools package**

  for sequencing files

  SAM (Sequence Alignment / Map), FASTA, binary variant call (BCF), BAM(Binary Alignment / Map)

- biocLite(“Rsamtools”)

  library(Rsamtools)

Module 5

Ranges in Bioconductor

- **Biobase package**

- biocLite(“Biobase”)
  library(“Biobase”)

  exprs( ) - genetic expression
  pData( ) - phenotype of samples
  fData( ) - genotype of samples

- **Iranges** – interval ranges

  biocLite("IRanges")
  library(IRanges)

- Define a range:
  need 2 out of the following to define the range
  start, width, end

  start( ) ; end ( ); width( );
  length( ) (see how long the Irange - bases)

- **Iranges** – functions

- shift( ) (shift the Irange)
- narrow( ) (define where to start/end range)
- flank( ) (get flanking sequences next to Irange)
- resize (readjust length of Irange – modify)

- range( ) - define total length of series of Iranges
- reduce( ) - base pairs covered w/o gaps
- gaps( ) - gaps in Iranges
- disjoin( ) - set of ranges with same coverage

- **Iranges** – more functions

- findOverlaps(ir1,ir2) (where ir1 overlaps ir2)
- queryHits( )
- countOverlaps(ir1,ir2) (count the overlaps)
- nearest(ir1,ir2) which Iranges in ir1 are close to ir2

- **Granges**

- biocLite(“GenomicRanges”)
- library(GenomicRanges)

  ??Granges

- Genomic ranges contain Iranges set typically on the same chromosome

  gr1=GRanges(seqnames, strand, ranges)

- **Granges** – functions and data

- grL=GRangesList(gr1,gr2)

- length(grL)
- grL[1] (subset - get the first range)
- gr$gr1 (same as above)

- gr[[1]] (get data for first range)

- **Granges** – functions

- start( )
- seqnames( )
- seqlevels( )
- seqinfo( )
- elementLengths( )
- promoters( )
- genes( )

- **Biostrings package**

- biocLite(“Biostrings”)
  library(Biostrings)

- DNAString( ) (keying in a DNA string)
- DNAStingSet( )

- **Biostrings package** – functions

- width( ) (how many bases in string)
- sort( ) (sort lowest to highest)
- rev( ) (reverse the order of the set)
- reverse( ) (reverse the sequence)
- translate( ) (translate DNA into aminoacid ID)

- **Biostrings package** – more functions

- alphabetFrequency( ) - how often DNA base occurs
- letterFrequency( ) -counting DNA string (eg AC)
- dinucleotideFrequency( )
- consensusMatrix( ) - generate matrix

Challenge

- Obtain data from H3K4me3 histone modification from the H1 cell line from Epigenomic roadmap using Annotation Hub

  Subset these regions to keep only regions mapped to autosomes (chr 1 to 22)

  how many bases do the regions cover?

- Repeat the exercise for H3K27me3 histone

Challenge

- What is the GC content of chr 22 in the “hg19” build in the human chromosome

- CpG islands are dense clusters of CpGs. What is the observed number of “CG” dinucleotides for CpG islands on chromosome 22

- A TATA box is a DNA element of the form “TATAAA”. Around 25% of genes should have TATA box in their promoters. How many TATA boxes are there on chr22 on hg19 build of human genome

Challenge

- It is possible for two promoters from different transcripts to overlap

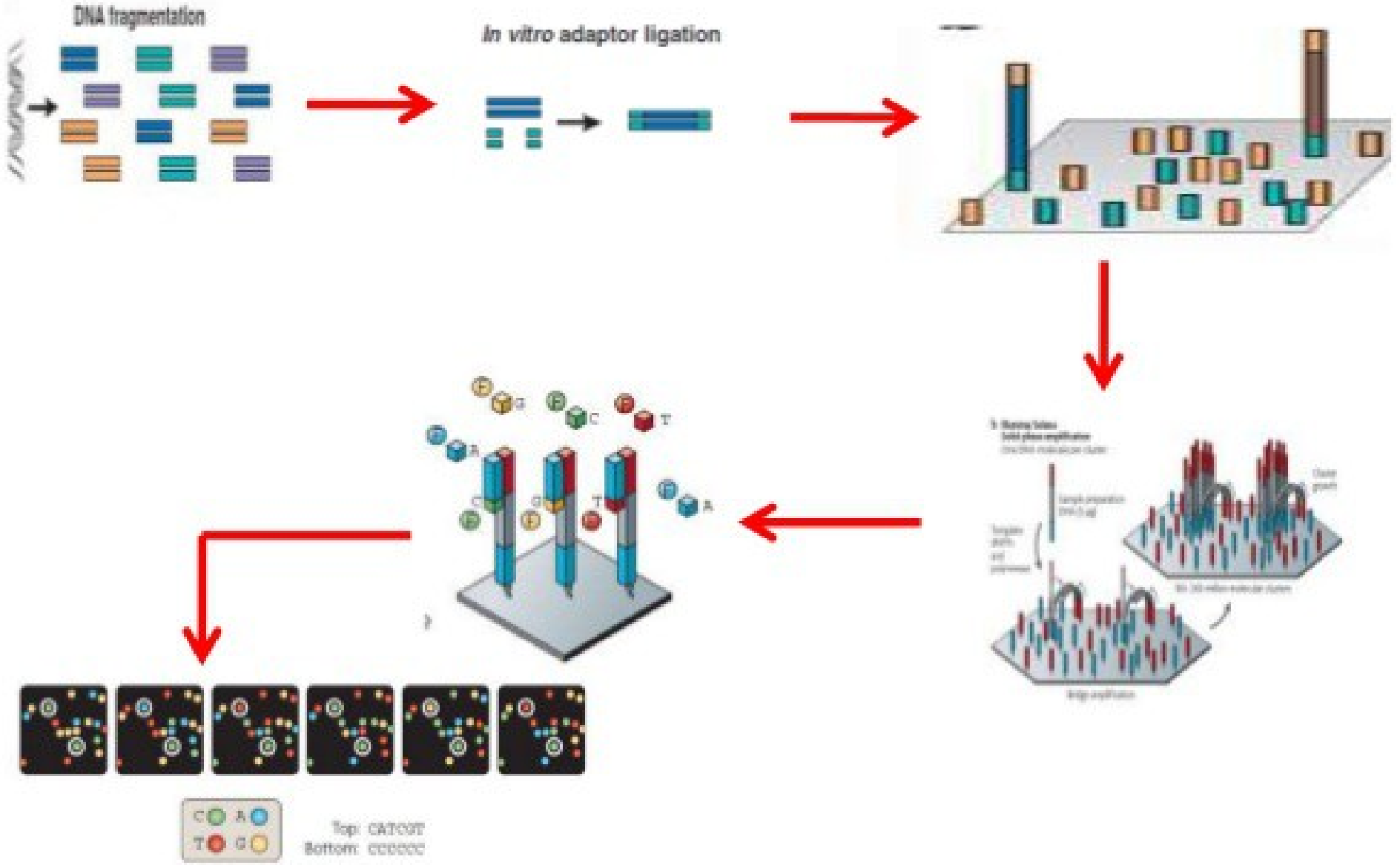
  how many bases in chr 22 are part of one or more of promoter of a coding region

- Which of the following sequences are most common on chr 11? “ATG”,”TGA”,”TAA”,”TAG”

Module 6

Next Generation Sequencing

TECHNOLOGY: NEXT GENERATION SEQUENCING



- **Applications of NGS**
- All read mapped to reference genome

- **#1 variant detection**

  find new SNPs

  take a sample of DNA and align to genome, see if person is heterozygote at a particular bp location
- **#2 RNA-sequencing**

  gene expression

  Take RNA, convert to DNA, align all reads

  a lot of reads in sampleA align to gene1 than gene2, therefore gene1 is highly expressed

- **Applications of NGS**

- **#3 Binding sites for Transciption factors**

  align reads to DNA, which location has enough read (peak detection), we believe that protein was bound to that site

- **Shortread package**

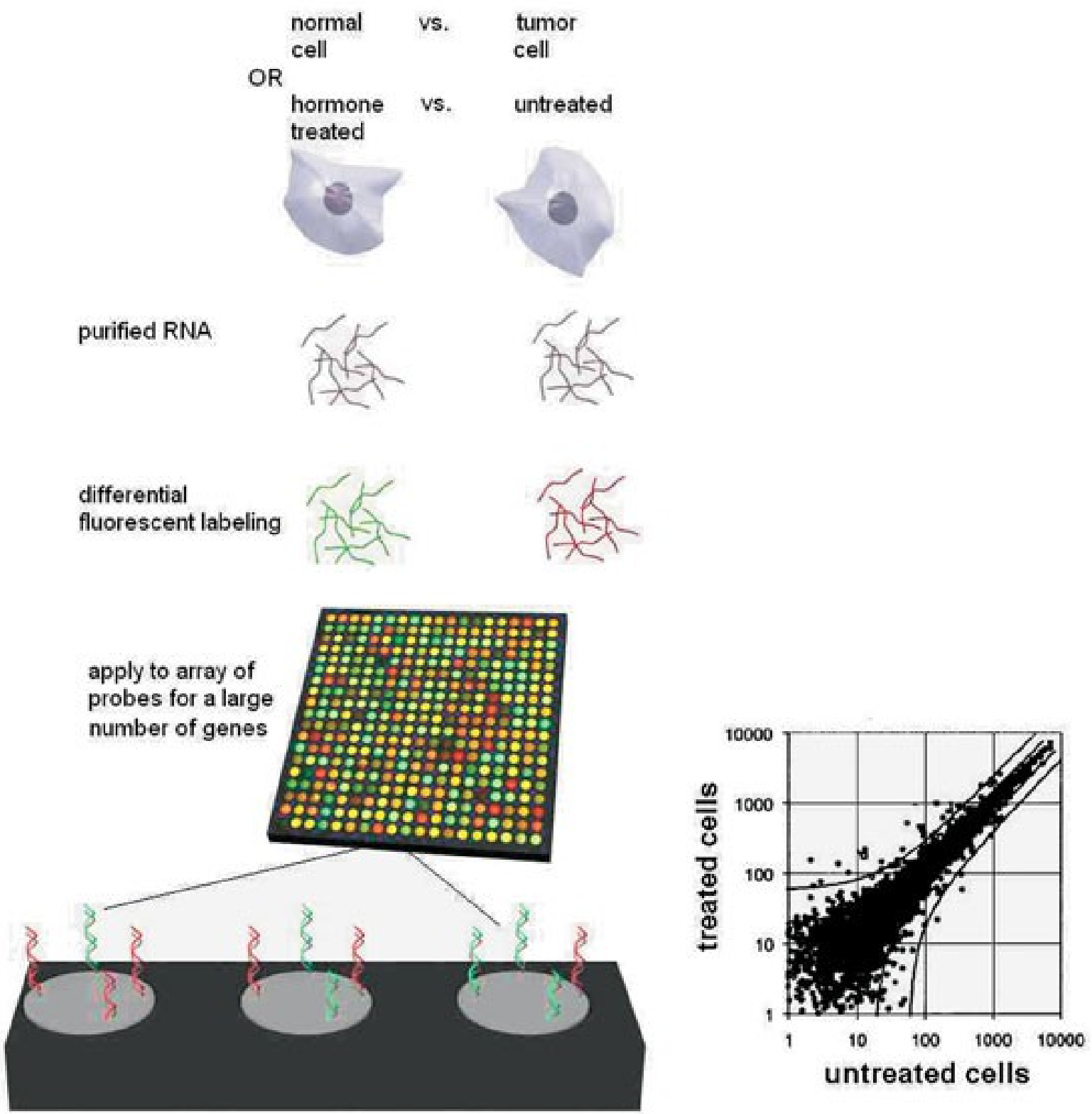
- Used for reading raw sequence reads, alignment

- biocLite(“ShortRead”)
  library(ShortRead)

- sread( )
- quality( )

Module 7

Microarray
(DNA gene chip)


normal cell
vs.
tumor cell
OR
hormone treated
vs.
untreated
purified RNA
differential fluorescent labeling
apply to array of probes for a large number of genes
10000
1000
100
10
1
1
10
100
1000
10000
treated cells
untreated cells

- **Applications of gene-CHIP**

- **#1 measure gene expression**

  for every gene, creates probes for several locations in each transcipt

  label RNA

  if transcipt is abundant, lot of hybridization, high reading for all the probes

- **Applications of gene-CHIP**

- **#2 genotyping (SNP) array**

  different people have different alleles

  AA, AG, GG persons in SNPs

  create probes to hybridize to A, another to hybridize to G, millions of probes and genotype millions of SNPs at the same time

  genotype people w certain disease and those w/o see the difference

- **Applications of gene-CHIP**

- **#3 detection of trasnciption factor binding sites**

- Where in the genome a specific protein is bound
- Fragmenting the DNA

  sonicate, use antibody to bind to protein bound DNA isolate and seperate from other strands
- Remove the proteins, amplify the DNA and hybrdize with chip
- Total genome (w/o protein bound) is also hybridized for comparison

- **Oligo package**

- Package for handling microarray geneExp and snp data from Affymetrix

- biocLite(“oligo”)
- library(oligo)

Module 8 (optional)

Workflows

- **Workflows**

- Enables scientists to perform multi-step operations using sereral packages at a time for common work routines

- Source("http://bioconductor.org/workflows.R")
  workflowInstall("sequencing")

  (this will install the entire human genome in your computer ~ about 650MB) – take note

- Lots of info on the bioconductor website

- **Miscellaneous** - Useful datasets

- Microarrays from Affymetrix

  hu6800.db

  hgu95a.db

  hgu133a.db

  mgu74a.db

  rgu34a.db

- **Miscellaneous** – packages to read data files

- Biostrings – FASTA files
- Shortread – FASTQ files
- GenomicAlignments – BAM files
- VariantAnnotation – VCF files

- **Miscellaneous** – useful bioconductor packages

- **General** – Biobase, rhdf5, tkWidgets, reposTools
- **Annotation** – annotate, AnnBuilder
- **Graphics** – geneplotter, hexbin
- **Preprocessing for Affymetrix ligo chip data** -

  affy, CDF packages
- **Preprocessing cDNA microarry data** -

  marrayClasses, myarrayInput, myarrayNorm, myarrayPlots
- **Differential gene expression** -

  edd, genefilter, multtest, ROC

- **Miscellaneous** – useful bioconductor packages

- 2 color spotted arrays – limma
- RNA-seq – qProject
- NGS – DESeq2