LECTURE NOTES

Summer School in Bioinformatics & NGS Data Analysis

CBiES, Jachranka, Poland
September 10-17, 2017

#NGSchool2017
`https://ngschool.eu/2017`

Organized by

- International Institute of Molecular and Cell Biology in Warsaw
- Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava
- Department of Information Technologies, Masaryk University in Brno
- Institute of Genetics, Hungarian Academy of Sciences in Szeged

Contributions

- Leszek Pryszcz: main organiser, `https://ngschool.eu/`, Lecture notes and more
- Organisation & scientific committee: Broňa Brejová, German Demidov, Alina Frolova, Katarzyna Kędzierska, Bogumił Konopka, Łukasz Kiełpiński & Tomáš Vinař
- IIMCB Admin, Grant, Finance & PR Units
- Centrum Badań i Edukacji Statystycznej GUS: accommodation & boarding
- PTBi: help with obtaining the grant from Polish Ministry of Science and Higher Education
- Ewa Ramotowska #NGSchool logo

Supporters This activity is financially supported by grants from **International Visegrad Fund** (Visegrad Grant No. 21710381), **Polish Ministry of Science and Higher Education** (842/P-DUN/2017) and **International Institute of Molecular and Cell Biology in Warsaw**.

Contents

Programme

We'll have **morning (9-13)** and **afternoon (14-18)** sessions with coffee breaks around 11:00 and 16:00. In the evenings, we'll have lecture-only sessions, shot-talks, discussions and some more relaxed activities. Breakfast will be served from 8:00, lunch at 13:00 and dinner around 19:00. Workshops last 4 hours and consists of theoretical introduction and practical exercises.

	Day 0: Sunday	
15:00	**Introduction to Bioinformatics & NGS**	*Broňa Brejová & Tomáš Vinař*
18:00	**Welcome, dinner & Shot talks #1**	*Leszek Pryszcz*
	Day1: Monday	
9:00	**Data visualisation**	*Przemysław Biecek*
14:00	**Molecular data integration**	*Jacek Marzec*
20:00	**Shot talks #2**	*Leszek Pryszcz*
	Day2: Tuesday	
9:00	**Workflows & pipelines**	*Alina Frolova*
14:00	**Genome & transcriptome assembly**	*Andrey Prjibelski*
20:00	**Beta & Bit games**	*Przemysław Biecek*
	Day3: Wednesday	
9:00	**Detection of structural variations**	*Tomasz Gambin*
14:00	**Introduction to Statistics**	*German Demidov*
14:00	**RNA Structure Probing**	*Łukasz Kiełpiński*
17:30	**Open science - discussion**	*Paweł Szczęsny*
20:00	**Chilling evening: BBQ #1**	
	Day4: Thursday	
9:00	**Functional genome annotation**	*Marina Marcet-Houben*
14:00	**ChIP-seq**	*Aliaksei Holik*
20:00	**Team-working**	*Panagiotis Theodorakis*
	Day5: Friday	
9:00	**Single-cell RNA-seq analysis**	*Davis McCarthy*
14:00	**Differential expression analysis**	*Davis McCarthy*
20:00	**Genome structure & function**	*Noam Kaplan*
	Day6: Saturday	
9:00	**Single-cell Hi-C data analysis**	*Aleksandra Galitsyna*
14:00	**ATAC-seq**	*Katarzyna Kędzierska*
14:00	**Microbial genomics**	*Adam Witney*
20:00	**Chilling evening BBQ #2**	
	Day7: Sunday	
10:00	**Recap & farewell**	*Leszek Pryszcz*

Introduction to Linux, Bioinformatics & NGS (Lecture & Workshop)

Broňa Brejová & Tomáš Vinař
Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava

Sunday, 15:00

In this workshop, we will go through several exercises which will allow participants to learn or improve basic skills in working with Linux command line and processing NGS data using bioinformatics software. Exercises will include genome assembly and read mapping, gene finding, and detection of positive selection. Participants can choose exercises according to their proficiency levels, advanced bioinformaticians are welcome to help beginners. In case of interest, exercises will be complemented by short lectures on basics of bioinformatics methods used. During this workshop, we will also help participants to install software needed later during the summer school.

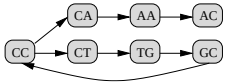
Genome assembly

Sequencing technologies currently in use

Technology	Read length	Error rate	Per day	Price per MB
1st generation				
Sanger	up to 1000 bp	$< 2\%$	3 MB	\$4000
2nd (next) generation (cca 2004)				
Illumina	$2\times$ 150bp	$< 1\%$	30 GB	$< \$1$
3rd generation (cca 2015)				
PacBio	6-25 Kbp	15%	1 GB	\$2
Oxford Nanopore	100kbp possible	15%	1 GB	\$1

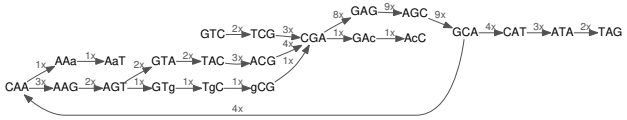
Assembling short reads: de Bruijn graphs

- Split reads to overlapping windows of length k
- **de Bruijn graph** of dimension k:
  - **vertices:** substrings of length k from all reads
  - **directed edges:** connect k-mers consecutive in at least one of the reads (overlapping by $k-1$)
- **Example:** $k = 2$, reads: CCTGCC, GCCAAC
  (typical k would be between 30 and 70)

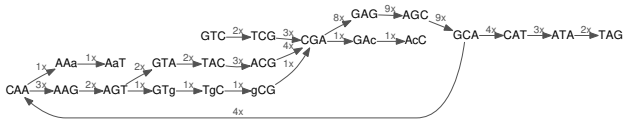


Example: set of reads and the resulting de Bruijn graph

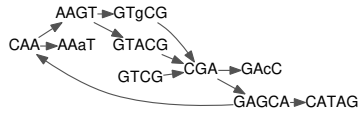
```
GTCGAGCAAGTACGAGCATAG
 TCGAGCA       AGCATAG
    AGCAAaT    AGCATAG
GTCGAcC  GTACGAG
GTCGAGC   TACGAGC
  CGAGCAA  ACGAGCA
        AGTgCGA
       CAAGTAC
     GCAAGTA GAGCAT
   GAGCAAG   GAGCATA
         TACGAGC
```



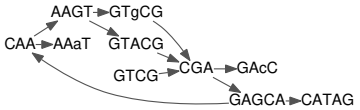
Example: simplifying de Bruijn graph



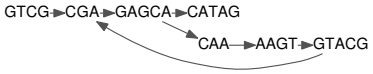
Collapse unique paths to nodes



Example: remove errors from de Bruijn graph



Remove errors (tips and bubbles with low coverage)



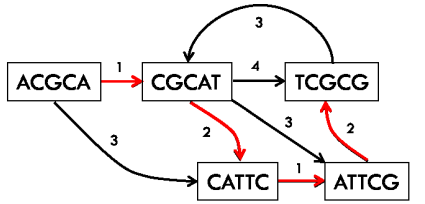
Collapse again, obtain 4 contigs (orig. `GT|CGAGCA|AGTA|CGAGCA|TAG`)

Software tools: Velvet, ALLPATHS-LG, SGA, Spades

Results of genome assembly

- It is unrealistic to expect complete chromosomes from a short read assembly
- Fragmented assemblies with long contigs can be used for downstream analysis in most cases
- Measures of assembly fragmentation:
  - Number of contigs
  - Maximum and average length of contigs
  - N50 value x: if we only take contigs of length $> x$, we will cover $> 50\%$ of the genome

Assembling long reads: Overlap graphs



Software tools: HGAP, miniasm, Canu

(figure from EE372 at Stanford)

Example: Assembling *Magnusiomyces capitatus* genome

(19.6 Mbp genome, 4 chromosomes + mtDNA)

Technology	Coverage	# contigs	largest	avg	N50
Illumina / Spades	250x	1102	172.6 Kbp	17.6 Kbp	62.0 Kbp
PacBio / Canu	37x	17	4.7 Mbp	1.2 Mbp	1.7 Mbp
PacBio + MinION	65x	11	4.4 Mbp	1.8 Mbp	2.0 Mbp

Sequence alignment / read mapping

Sequence alignment, homology search

Given two sequences / databases, find alignments of similar regions

```
CCCGACGAGAAGGCCATAATGACCTATGTGTCCAGCTTCTACCATGCCTTT
|| |||||||||| ||||    ||||| |||  || || ||| ||   ||||
CCGGACGAGAAGTCCAT---CACCTACGTGGTCACCTACTATCACTACTTT
```

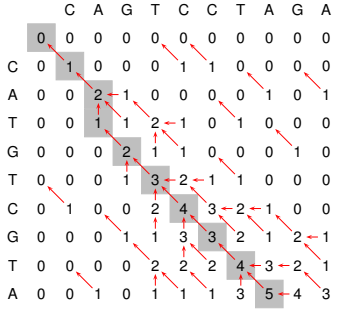
Alignments are needed in many bioinformatics tasks:

- Finding overlaps of (long) reads in genome assembly
- Mapping reads to a genome (variation, RNA-seq, ChIP-seq, etc)
- Finding homologs of a gene (BLAST)
- Comparing whole assemblies/genomes

Classical approach: dynamic programming

It can find **all** high-scoring alignments

It compares each pair of bases - running time scales quadratically



seq. length	time
100	0.0008s
1,000	0.08s
10,000	8s
100,000	13m (*)
1,000,000	22h (*)
10,000,000	3months (*)
100,000,000	25years (*)

Too slow for long sequences

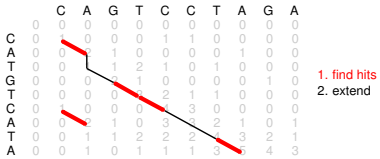
Seed and extend alignment algorithms

- Trade sensitivity for speed (some alignments not found)
- Reduce the search to "promising" parts of the matrix

Example: BLASTN [Altschul et al 1990]

- Find **short exact matches** of length w (**seeds**)
- Extend seeds to longer alignments
- Discard seeds that cannot be well extended

Example:
$w = 2$
(in practice $w \geq 10$)



How to find short exact matches?

- Build an index – a"dictionary" of words of length w from the first sequence
- Look up each word from the second sequence in the dictionary

Example: CAGTCCTAGA vs CATGTCATA

Dictionary:
AG 2, 8
CA 1
CC 5
CT 6
GA 9
GT 3
TA 7
TC 4

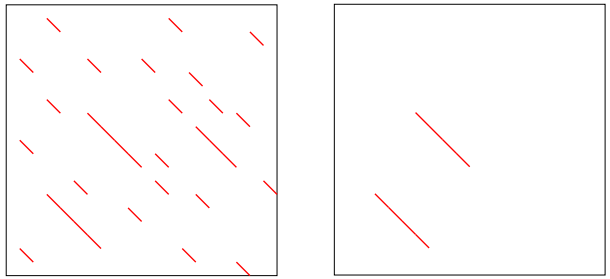
Lookup:
CA → 1
AT → -
TG → -
GT → 3
TC → 4
CA → 1
AT → -
TA → 7

Dictionary stored using various computer science techniques (hashing, Burrows-Wheeler transform)

Sensitivity vs. running time

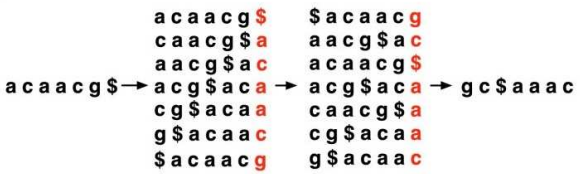
Small w
many spurious seeds, slow

Large w
many alignments not found



Burrows-Wheeler transform

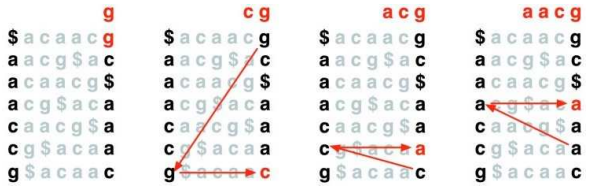
Encode acaacg, special symbol $ added as delimiter



Originally designed for compression, e.g. bzip2

Example from Langmead et al, Genome Biol. 2009

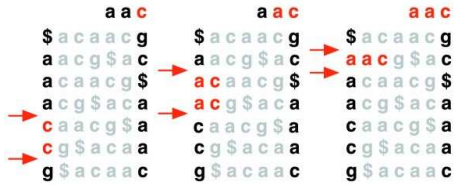
Decoding Burrows-Wheeler transform



The ith occurrence of X in the last column corresponds to the same text character as the ith occurrence of X in the first column

Searching in Burrows-Wheeler transform

Occurrences of aac

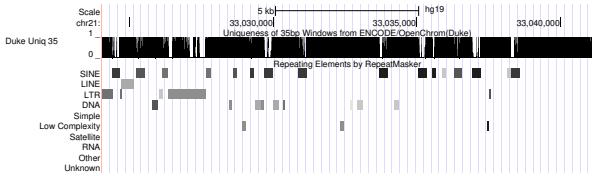


Add first letter a:
– find range containing a's
– add those a's that precede ac
– count a's in BWT column up to current position

Advantages of BWT: small index size,
fast search for short exact matches,
can be extended to inexact matches etc

Mappability

- Repeats and duplicated sequences cannot be distinguished with short reads (paired reads can help)
- Mapping quality: reflects probability that alignment is wrong (read comes from a different part of the genome)
- Report all high-quality alignments or choose one randomly (e.g. to estimate coverage in duplicated areas)



Basic file formats

FASTA

Used for storing DNA/RNA/protein sequences

Only name + sequence, name starts with >

```
>HWI-ST1218:80:D0VGUACXX:4:1101:1321:1960 1:N:0:CAGATC
NACTACTGTAGAAGTTCAAGATTATTCCACAGGATCATCATATTCATATGGAGATCAATCTGGTTC
>HWI-ST1218:80:D0VGUACXX:4:1101:1321:1960 2:N:0:CAGATC
ACTGCCAGCAGAAAAATCTCCCAATTTTTCACCGGATATTTTGCCGCCTTCAGACTGAATTGAAGT
```

```
>sp|P00410|COX2_YEAST Cytochrome c oxidase subunit 2
MLDLLRLQLTTFIMNDVPTPYACYFQDSATPNQEGILELHDNIMFYLLVILGLVSWMLYTIVMT
YSKNPIAYKYIKHGQTIEVIWTIFPAVILLIIAFPSFILLYLCDEVISPAMTIKAIGYQWYWKY
EYSDFINDSGETVEFESYVIPDELLEEGQLRLLDTDTSMVVPVDTHIRFVVTAADVIHDFAIPS
LGIKVDATPGRLNQVSALIQREGVFYGACSELCGTGHANMPIKIEAVSLPKFLEWLNEQ
>sp|P21534|COX2_SCHPO Cytochrome c oxidase subunit 2
MLFFNSILNDAPSSWALYFQDGASPSYLGVTHLNDYLMFYLTFIFIGVIYAICKAVIEYNYNSH
PIAAKYTTHGSIVEFIWTLIPALILILVALPSFKLLYLLDEVQKPSMTVKAIGRQWFWTYELND
FVTNENEPVSFDSYMVPEEDLEEGSLRQLEVDNRLVLPIDTRIRLILTSGDVIHSWAVPSLGIK
CDCIPGRLNQVSLSIDREGLFYGQCSELCGVLHSSMPIVVQGVSLEDFLAWLEENS
```


FASTQ

Used for storing reads including quality values for each base

@: read name, technology-specific format

```
@HWI-ST1218:80:D0VGUACXX:4:1101:1321:1960 1:N:0:CAGATC
NACTACTGTAGAAGTTCAAGATTATTCCACAGGATCATCATATTCATATGGAGATCAATCTGGTTC
+
#1=D?DDDDFHHHIGHIIIIIIIIIIIIIIIGIIIIIIGIIGIIIIIIIIIIIIIIIIIIIIIHII
@HWI-ST1218:80:D0VGUACXX:4:1101:1321:1960 2:N:0:CAGATC
ACTGCCAGCAGAAAAATCTCCCAATTTTTCACCGGATATTTTGCCGCCTTCAGACTGAATTGAAGT
+
CCCFFFFFHHHHFIGIIJJJJJJIJJJIJIHJJJHHIIIJJJJJJIIIIGFEIIIJJJJGJHFHHH
```

Base quality q: probability of error $10^{-q/10}$

Quality encoded as a single character with ASCII code $q + 33$

For example symbol + has ASCII value 43,

which means $q = 10$ and probability of error $10^{-1} = 10\%$

Base quality codes

	q	error		q	error
!	0	1	2	17	0.02
"	1	0.794	3	18	0.0158
#	2	0.631	4	19	0.0126
$	3	0.501	5	20	0.01
%	4	0.398	6	21	0.00794
&	5	0.316	7	22	0.00631
’	6	0.251	8	23	0.00501
(	7	0.2	9	24	0.00398
)	8	0.158	:	25	0.00316
*	9	0.126	;	26	0.00251
+	10	0.1	<	27	0.002
,	11	0.0794	=	28	0.00158
-	12	0.0631	>	29	0.00126
.	13	0.0501	?	30	0.001
/	14	0.0398	@	31	0.000794
0	15	0.0316	A	32	0.000631
1	16	0.0251	...		
			Z	57	2e-06

Older versions of Illumina use a different encoding!

SAM/BAM

Used for storing results of **read mapping** (alignments)

SAM: text format, (somewhat) human readable

BAM: binary format, less space, convert to SAM via samtools

Header plus one row for each read

Columns: read ID, flag, contig, position, mapping quality, CIGAR, 3 columns for paired reads, read sequence, read quality, optional fields

Flags: binary encoded, e.g. if forward or reverse strand

CIGAR: matches, insertions, deletions, introns, etc., e.g. 87M2D43M

One row of a SAM file:

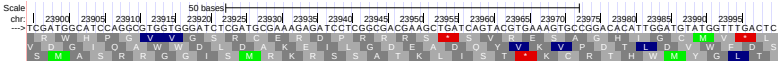
```
HWI-ST1300:156:H7599ADXX:1:1110:17800:29845     145     contig0004    \
10      50      41M     contig0346      339     0       \
CTATAGATCTTATTACCGCTTATCCAAAGCTGAAAGTGACT       \
IIIIIIIIIIIIIIFIIIIIIIIIIIIIIFFFFFFFFFFBBB      \
AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:41 YT:Z:UU NH:i:1
```

Gene finding

Gene finding in prokaryotic genomes

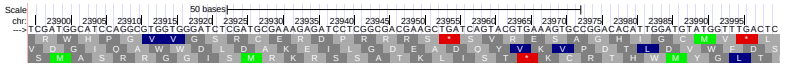
Goal: find all genes encoding proteins in a genome.
This gives us a catalogue of all proteins.

ORF: open reading frame
Sequence of codons that does not contain a stop codon
+ start codon at the beginning and stop codon at the end



It is very easy to write a program that finds all ORFs
(tricky details: 6 reading frames, differences in genetic code)

Gene finding in prokaryotic genomes



It is easy to find all ORFs in a genome.

But not every ORF is a gene:

- Purely random ORFs (particularly short ones)
- Pseudogenes
- Multiple possible start codons

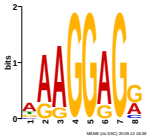
For example, in E. coli annotation between 1997 and 2005:
682 changes in start codon position, 31 genes excluded, 48 new added

Need to consider subtler features of a gene

- Typical frequencies of codon usage (in various organisms)
- Sequence motifs, e.g. ribosomal binding site, cca. 5-10bp before start codon

Codons for threonine

	Haemophilus influenzae codons	Escherichia coli codons	Escherichia coli genome
ACA	16	6	13
ACC	11	23	16
ACG	9	14	16
ACT	16	8	11



Combination using statistical models
e.g. hidden Markov models (HMMs)

Gene finding in eukaryotic genomes

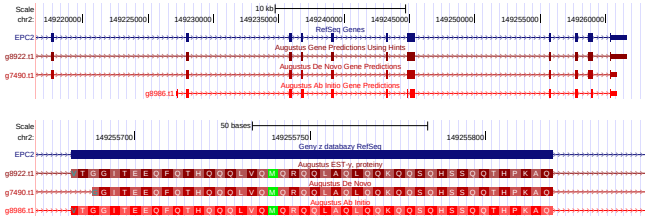
More difficult problem

- ORF interrupted by introns
- Average values in the human genome:
  - cca. 10 exons per gene
  - coding exons cover 1.2% of the genome

More complex statistical models:

– splicing signals at exon boundaries
– statistical properties of introns
– reading frame in adjacent exons agrees

Gene finders make mistakes



Best methods in 2005 on the human genome: [Guigo et al 2006]
20% of genes, 60% of exons correct using DNA only
70% of genes, 85% of exons correct with other information sources

Since then some progress in methods, much more additional data,
also many genomes easier than human
Beware: alternative splicing often not predicted at all

Practical considerations

Improving prediction accuracy using additional data

Expression data (RNA-seq)
Similarity to known proteins
Comparative genomics
Chromatin state (histone modifications)

Training parameters for a new species

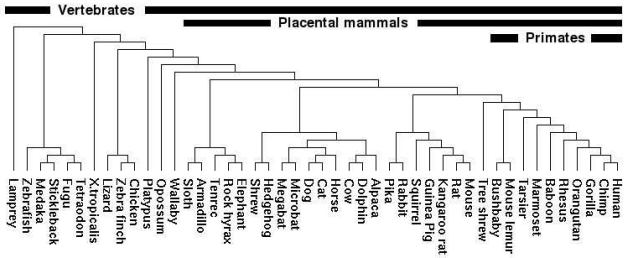
Many statistics used for gene finding vary among genomes
Ideally adjusted from a set of manually curated genes
or iterative training and prediction

Example of gene finding software:

Augustus has parameters for many eukaryotic genomes
Maker is a pipeline for training Augustus, using additional data

Comparative genomics

Comparative genomics



Nothing in biology makes sense except in the light of evolution
(Theodosius Dobzhansky, 1973)

Why do we need so many genomes?

- Common features of genomes: Which genes are responsible for basic biological functions?
- Differences between genomes: Which mutations are responsible for typical traits of individual species?
- Identify elusive functional regions.
  (RNA genes, regulatory regions, . . .)
- Study evolutionary mechanisms and their impact on genomes.

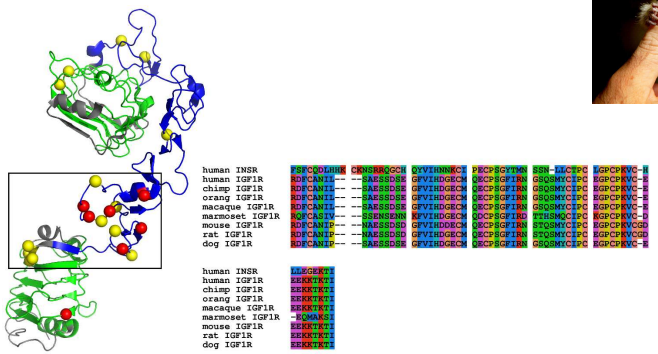
Whole-genome studies: positive selection in protein coding genes

Looking at patterns of mutation in protein coding genes:

- Synonymous: local "neutral" speed
  e.g. ACA (Thr) → ACT (Thr)
- Non-synonymous: possible functional changes
  e.g. ACA (Thr) → AAA (Lys)

High ratio of non-synonymous to synonymous changes (ω)
is a sign of positive selection

IGF1R: Example of a gene under positive selection



Marmoset Genome Consortium, Nature Genetics, 2014

Data visualisation (Lecture & Workshop)

Przemysław Biecek
Warsaw University of Technology

Monday, 9:00

Molecular data integration (Lecture & Workshop)

Jacek Marzec
Barts Cancert Institute, London

Monday, 14:00

The open access to huge volumes of genomic data stored in public data repositories, such as Gene Expression Omnibus (GEO), ArrayExpress or Sequence Read Archive (SRA), enables novel large-scale discovery studies. Indeed, integration of data produced across many studies is now well-recognised as a powerful approach that provides novel biological insights while allowing for the identification of numerous alterations contributing to a given phenotype not evident from single experiments. Multi-study data integration increases the statistical power to capture consistent molecular alterations that might be hampered by a limited sample size and experimental artefacts associated with individual datasets, and thus offers more accurate signatures. Moreover, cross-study data integration gives the opportunity for broader data overview and the potential to ask novel biological questions.
During this workshop you will integrate publicly available molecular data from independent microarray experiments on prostate cancer.

Overview

- Why integrate the data?
- Approaches
- Challenges
- Methodology

http://www.freeimages.com

Why integrate the data?

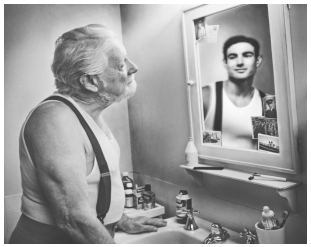
- Multitude of microarray and NGS technologies are used for gene expression profiling
- Produced data are stored in public repositories:
  - NCBI Gene Expression Omnibus (GEO)
  - EMBL-EBI ArrayExpress
  - NCBI/EBI Sequence Read Archive (SRA)
- …or international consortia data portals:
  - International Cancer Genome Consortium (ICGC)
  - The Cancer Genome Atlas (TCGA)

➢ This only makes sense if these data could be reused/integrated and further explored

http://www.isix.com

Why integrate the data?

- Lack of reproducibility and poor overlap of molecular signatures across studies
  - limited sample size
  - differing laboratory protocols and analysis pipelines

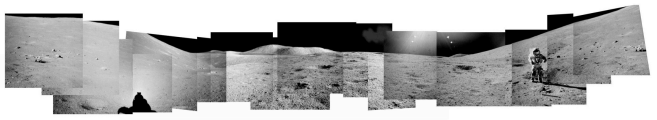


➢ These can be addressed by systematic integrative analysis performed on larger patient cohorts

http://narrativeincreativedirection.myblog.arts.ac.uk

Why integrate the data?

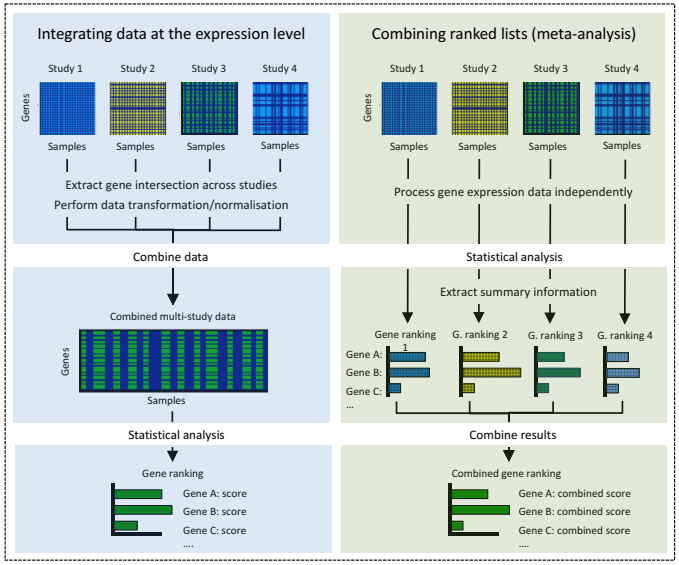
- Increased **statistical power** enables identification of alterations not evident from individual experiments
- **Compensate** for possible **errors** in individual studies
  (the more datasets the more control over study-specific variances)
- Potential to **estimate reproducibility** and develop more accurate gene signatures
- Opportunity for **broader data overview** and to gain answers to more questions



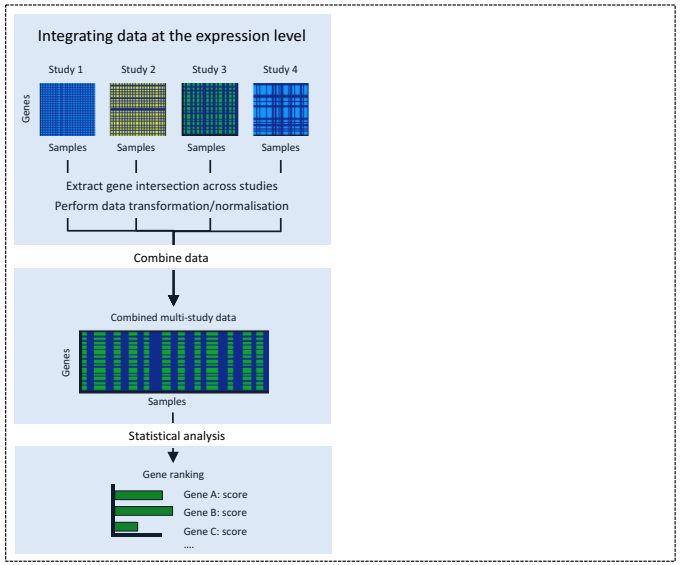
https://www.hq.nasa.gov

Approaches

1. Integrating data at the expression level
2. Combining ranked lists (meta-analysis)



Adapted from Sarmah CK, Samarasinghe S. Curr Bioinform. 2010; 5(4):281



Adapted from Sarmah CK, Samarasinghe S. Curr Bioinform. 2010; 5(4):281

Approaches

	(1) At the expression level	(2) Based on meta-analysis
+	• Preliminary data assessment and filtering • Application of single optimised analytical workflow	• Free from prior assumption about underlying data distributions
–	• Limited to genes intersection • Confounded by experimental variation • Limited number of studies with raw data and associated metadata	• Depends on careful selection of studies with good quality data • Variation in pre-processing and analysis methods across studies • Vulnerable to studies with small sample sizes

Approaches

	(1) At the expression level	(2) Based on meta-analysis
+	• Preliminary data assessment and filtering • Application of single optimised analytical workflow	• Free from prior assumption about underlying data distributions
–	• Limited to genes intersection	• Depends on careful selection of studies with good quality data
–	• Confounded by experimental variation	• Variation in pre-processing and analysis methods across studies
–	• Limited number of studies with raw data and associated metadata	• Vulnerable to studies with small sample sizes

Challenges

- Careful QC is vital

➢ The robustness of data integration depends on the quality of underlying data

http://www.grantthornton.com; http://www.dentaltech.com

Challenges

- Differences in technologies and related experimental parameters
- Systematic variations and noise among datasets

➢ These influence consistency and reliability of downstream analysis

➢ Need to minimise the variance caused by experimental factors

Methodology

- Language and environment for statistical computing and graphics
- Free (open-source) software
- Compiles and runs across all platforms (UNIX, Windows, Mac OS)
- Provides a wide variety of statistical and graphical techniques

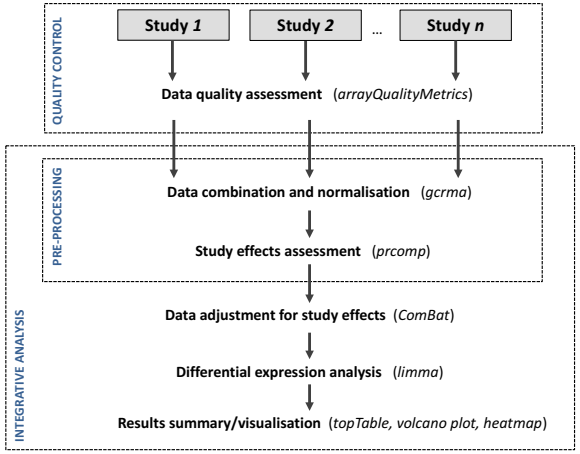
https://www.r-project.org/about.html

Methodology

- Relatively simple
- Large collection of tools for data analysis
- Effective data handling and storage facility
- Graphical facilities for data analysis and visualisation
- Highly extensible

https://www.r-project.org/about.html

Pipeline



Data QC

'*arrayQualityMetrics*'

- Provides access to a variety of QC metrics
- Handles most current microarray platforms
- Generates automated QC report
- Applicable to automated analytical pipelines

https://www.bioconductor.org/packages/devel/bioc/html/arrayQualityMetrics.html

QC: Quality control

Data normalisation

'*gcrma*' (Affymetrix platforms)

Combines three pre-processing steps

1. Background correction: corrects for noise in the data by adjusting for the effects of non-specific hybridisation
1. Normalisation: allows comparisons to be made between measurements from different samples
2. Summarisation: aggregates intensity values of multiple probes in a given probeset to a single expression value

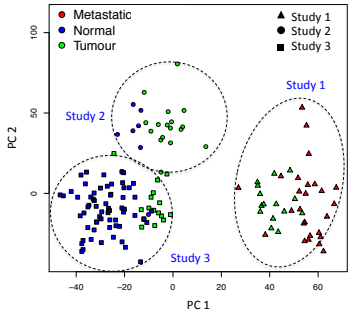
https://www.bioconductor.org/packages/release/bioc/html/gcrma.html

Study effects assessment

'*prcomp*' (PCA) (*stats* package)

- Facilitates identification of key components of variability in expression data derived from different studies

https://stat.ethz.ch/R-manual/R-devel/library/stats/html/prcomp.html



Study effects adjustment

'*ComBat*' (*sva* package)

- Adjusts data for known batches
- Applicable to microarray and NGS data
- Robustly manages high dimensional data with small sample sizes
- Superior to other methods

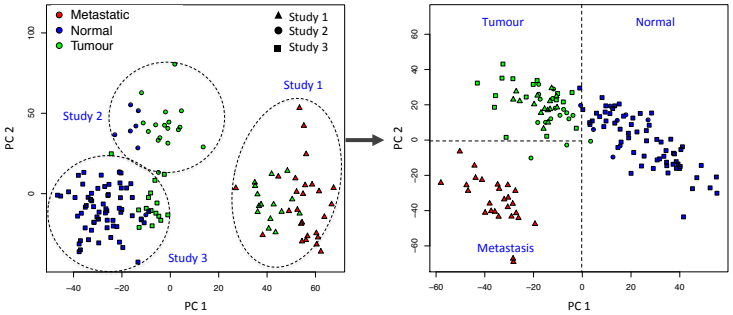
Chen C, Grennan K, Badner J *et al. PLoS ONE*, 2011;6(2):e17238

Müller C, Schillert A, Röthemeier C *et al. PLoS ONE*, 2016;11(6):e0156594

https://bioconductor.org/packages/release/bioc/html/sva.html

Study effects adjustment

'*ComBat*' (*sva* package)



Differential expression analysis

'*limma*' (*limma* package)

- Set of functions for differential expression analysis
- Based on linear modelling and empirical Bayes methods
- Applicable to microarray and NGS data
- Superior to other methods

Rapaport F, Khanin R, Liang Y *et al. Genome Biol*, 2013;14(9):R95

Seyednasrollah F, Laiho A, Elo LL. *Briefings in Bioinformatics*, 2015;16(1):59–70

Soneson C, Delorenzi M. *BMC Bioinformatics*, 2013;14(1):91

https://bioconductor.org/packages/release/bioc/html/limma.html

Results summary

'*topTable*' (*limma* package)

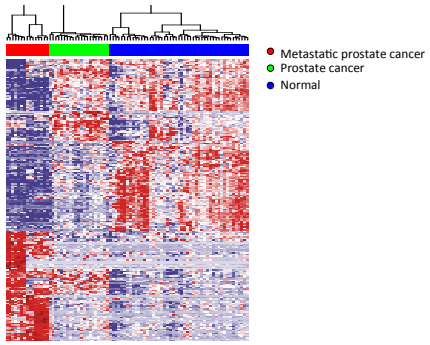
- Extracts table of the top-ranked genes from linear model fit

Gene symbol	Chr	Band	log2FC	p-value	p-value (BH)
DLX1	2	q31.1	-6.3	2.53E-76	2.41E-72
DLX2	2	q31.1	-5.6	1.07E-58	5.08E-55
HOXD13	2	q31.1	3.4	1.76E-54	5.60E-51
AMACR	5	p13.2	-4.1	1.90E-47	4.53E-44
NETO2	16	q12.1	-3.0	1.22E-45	2.33E-42
AOX1	2	q33.1	2.9	4.78E-45	7.59E-42
ZIC2	13	q32.3	-4.5	6.81E-45	9.27E-42
ROR2	9	q22.31	2.3	1.17E-40	1.39E-37
PPARGC1A	4	p15.2	2.3	1.81E-40	1.91E-37
ACSF2	17	q21.33	2.4	1.82E-39	1.74E-36
TCEAL2	X	q22.1	3.0	3.20E-39	2.77E-36
SLC16A5	17	q25.1	3.2	4.84E-39	3.84E-36
CYP3A5	7	q22.1	4.1	5.62E-39	4.12E-36
LUZP2	11	p14.3	-3.1	6.93E-39	4.72E-36

Results visualisation

'*heatmap*' (*heatmap.2* (gplots) or *heatmaply* (heatmaply) *functions*)

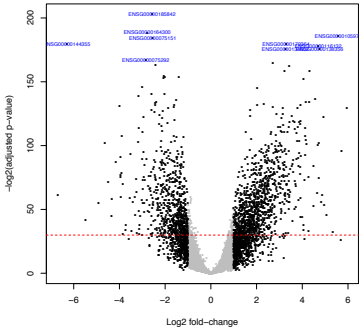
- Reflects gene expression values across samples representing various conditions



Results visualisation

'*volcano plot*' (*plot (graphics) or plot_ly (plotly)* functions)

- Plots significance (p-value) versus fold-change values
- Helps to quickly identify changes in large data sets



Workflows & pipelines (Lecture & Workshop)

Alina Frolova
IMBG, Kyiv

Tuesday, 9:00

Reproducible and efficient scientific pipelines are the core element of the successful and solid research. And the bioinformatics field is not an exception here, on the contrary, the seeming easiness of re-doing the "experiment" and therefore less thorough testing and intermediate results check-up may lead to mistakes in the initial steps and ruin your final conclusions. Putting aside errors in the calculations or results interpretation, bioinformaticians usually need to run the same tool multiple times with different parameters or need a series of consequent code testing while developing new software. And doing it manually is not an option.
To address mentioned issues number of tools have been developed, many of which originate from pure Computer Science like version control systems, package, dependency and environment management systems or integrated development environments. There are also bioinformatics workflow management systems — specialized form of workflow management systems designed specifically to compose and execute a series of computational or data manipulation steps, or a workflow, that relate to bioinformatics. Among the well-known examples are Galaxy, GenePattern, KNIME. However, there are even more flexible tools for precise pipelines constructions and fine-grained parameters tuning such as SnakeMake — a scalable bioinformatics workflow engine, or NextFlow — a domain specific language for parallel and scalable computational pipelines. During the workshop we will try to review key components of efficient workflows management to make the everyday life of bioinformatician easier and more productive.

Why do we need specific tools to organize our data analysis?

- Dozens of dependencies (binary tools, compilers, libraries, system tools, etc)
- Experimental nature of scientific workflows tends to be difficult to install, configure and deploy
- Heterogeneous executing platforms and system architecture (laptop→supercomputer)

CONDA
package manager

Some slides are adapted from Travis Oliphant and Kale Franz presentation

Reproducibility layers

- Code (github, bitbucket)
- Data (datproject, git LFS)
- Workflow (**SnakeMake**, NextFlow)
- Environment (**conda**, docker, vagrant)

Complex things are built out of simple things

- Fundamental principle of software engineering is “separation of concerns” (modularity)
- Reusability is enhanced when you “do one thing and do it well”
- To deploy you need to bring the pieces back together
- This all means you need a good packaging system

Packaging is a critical part of software

A **package manager** or **package management system** is a collection of software tools that automates the process of installing, upgrading, configuring, and removing computer programs for a computer's operating system in a consistent manner.

Poor packaging and deployment solutions are everywhere in open source and industry and lead to software engineering mistakes with poorly factored code:

- hard to test
- hard to debug
- hard to maintain

Packages typical metadata

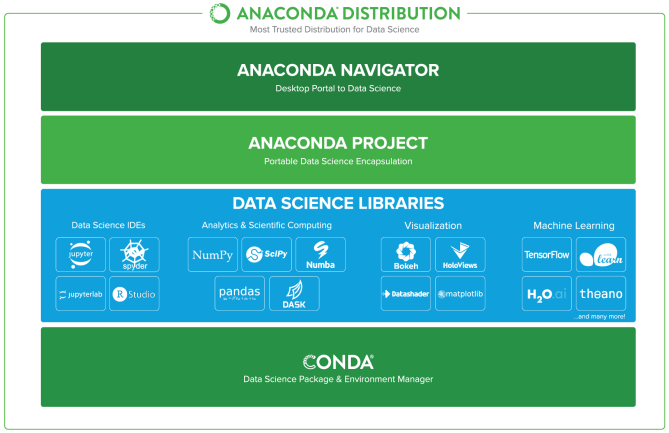
- the software's name
- description of its purpose
- version number
- vendor
- checksum
- list of dependencies necessary for the software to run properly

Package manager functions

- Extracting package archives
- Ensuring the integrity and authenticity of the package by verifying their digital certificates and checksums
- Looking up, downloading, installing or updating existing software from a software repository
- Grouping packages by function to reduce user confusion
- Managing dependencies to ensure a package is installed with all packages it requires, thus avoiding "**dependency hell**"

Typical software managing problems for data scientists

- I work on a server, I don’t have root access, how to compile/install locally (python, R packages)?!
- I need both python 2 and 3!
- I need to replace standard lib by compiling custom lib!
- I compiled and installed custom lib and everything stopped working!
- I updated the soft and custom lib was replaced again!
- I somehow configured working environment, but now I need to switch to another server!
- I try to install R Bioconductor package, but it says the package is missing for my R version!
- What the heck is JAVA_HOME, I don’t understand update-alternatives utility!
- Friend recommended me to try Gentoo linux, he said it is fun...
- **Your story?**



Conda allows to

- Manage project dependencies, including programming languages and libraries
- Isolate development and production environments with channels
- Share environments with a minimal footprint
- Support multiple versions of languages, storage systems, and packages
- Provide a common interface for building, installing, and sharing packages

Anaconda Distribution

- **Anacoda Navigator**: GUI that allows you to launch applications and easily manage conda packages, environments and channels without using a command prompt or terminal program
- **Anaconda Project**: automates setup steps such as installing the right packages, downloading files, setting environment variables and running commands; simplifies deployment to servers.
- **Conda**: cross-platform package manager, works from console

Enabling Environments

- **portability**
  system-level package management that's not tied to hard-coded system paths
- **multiple, composable environments**
  multiple instances of otherwise-colliding software, functionally isolated on the same system
- **preferential use of hard links**
  soft links are slower and problematic when working with linked shared libraries; hard links (or copies) more reliably keep the compiled-in relative paths intact)

Power and Flexibility for Users and Sysadmins Alike

- **natively multi-user**
  fully functional within the limited privileges of a non-privileged user

  admin/root users enabled with extensive configuration and enforcement capabilities

Enforcer of Safety and Correctness

- **pre-compiled packages only**
  will never require a compiler in production, or unexpectedly invoke one on you
- **environment integrity**
  disk-mutating operations are wrapped in a transaction, and rolled back in the event of errors
- **environment correctness**
  conda enforces compatibility of packages within environments

Channels for User Empowerment

- **the channel is a component of a package's identity**
  first-class citizen in package specifications
- **easy package building**
  with conda-build, a dedicated tool with engaged and dedicated code contributors
- **channels enable community**
  independent contributors, independent packaging communities, and corporations

BIOCONDA®

channel for the conda package manager specializing in bioinformatics software

a repository of > 2700 bioinformatics packages ready to use with conda install

Conda: Cross-Platform Package Manager

Linux

yum (rpm)
apt-get (dpkg)

macOS

macports
homebrew
fink

Windows

chocolatey
npackd

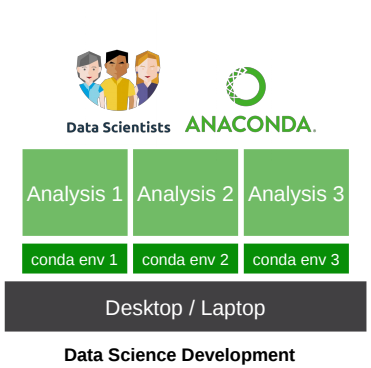
Cross-Platform

conda

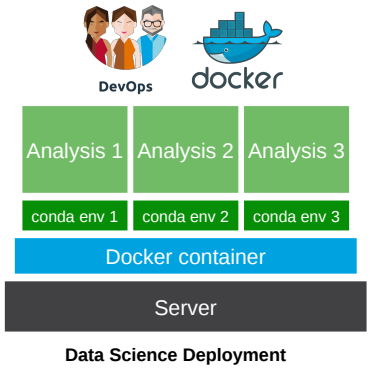
http://conda.pydata.org

Sophisticated light-weight environments included!

Conda + Docker: better together

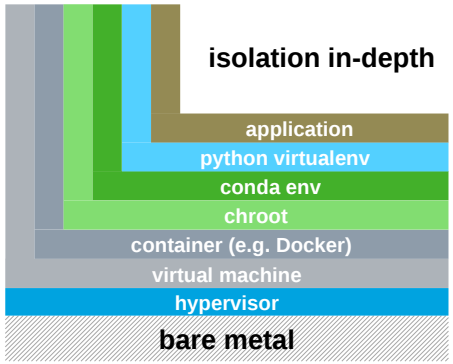


Data Science Development



Data Science Deployment

Layers of Process Isolation



pure-python isolation
functional isolation

more
secure isolation

How to start

DOWNLOAD Anaconda or miniconda and do test-drive
https://www.anaconda.com/download/
http://conda.pydata.org/docs/test-drive.html

GET the conda cheatsheet
https://conda.io/docs/_downloads/conda-cheatsheet.pdf

READ more about conda in the developer blog
https://www.anaconda.com/blog/developer-blog/

What is conda?	What is conda *not*?
packages	configuration management
environments	process management

SnakeMake
workflow manager

Some slides are adapted from Johannes Köster presentation

Advantages of workflows

efficiency	being automatic make researchers free from repetitive tasks and support “good practices”
reproducibiliy	analysis may be replicated over time, easily and effectively
reuse	both intermediate results and workflows can be reused
traceability	the workflow is enacted in a environment that allows tracing back results
separation of concerns	modulatization of the sub-tasks
multi-tasking	involving many pieces with a single command in a parallel manner
scaling	across nodes, clusters with heterogeneous environments
graphical user interfaces	some workflows automatically generate GUI (Galaxy)

Workflow systems examples

- Apache Taverna
- Galaxy
- Pegasus
- KNIME
- ExTasy
- MyExperiment
- **SnakeMake**

SnakeMake Concept

- Snakemake is a workflow management system that aims to **reduce the complexity** of creating workflows by providing a fast and comfortable execution environment
- Snakemake workflows are essentially **Python scripts** extended by declarative code to define rules.
- **Rules** describe how to create output files from input files.

Rules

```
rule myrule:
    input:
        "path/to/{sample}.txt"
    output:
        "path/to/{sample}.column1.txt"
    shell:
        "cut -f1 < {input} > {output}"
```

Rules

```
rule myrule:
    input:
        a="path/to/{sample}.txt"
    output:
        b="path/to/{sample}.column1.txt"
    shell:
        "cut -f1 < {input.a} > {output.b}"
```

Rules

```
rule myrule:
    input:
        a="path/to/{sample}.txt"
    output:
        b="path/to/{sample}.column1.txt"
    run:
        with open(output.b, "w") as out:
            for l in csv.reader(open(input.a)):
                print(l[0], file=out)
```

```
SAMPLES = "500 501 502 503".split()

# require a bam for each sample
rule all:
    input:
        expand("{sample}.bam", sample=SAMPLES)

# map reads
rule map:
    input:
        ref="reference.fasta",
        index="reference.bwt",
        reads="{sample}.fastq"
    output:
        "{sample}.bam"
    threads: 8
    shell:
        "bwa mem -t {threads} {input.ref} {input.reads} | "    # refer to threads and input
        "samtools view -Sbh - > {output}"    # refer to output files

# create an index
rule index:
    input:
        "reference.fasta"
    output:
        "reference.bwt"
    shell:
        "bwa index {input}"
```

Workflow execution

- disjoint paths in the DAG can be parallelized
- only outdated or missing files are created

```
# perform a dry-run
$ snakemake -n

# execute the workflow using 8 cores
$ snakemake --cores 8

# execute the workflow on a cluster (with up to 20 jobs)
$ snakemake --jobs 20 --cluster "qsub -pe threaded {threads}"

# execute the workflow on a cluster using the DRMAA API
$ snakemake --jobs 20 --drmaa
```

Workflow execution

- dependencies between rules are determined automatically
- directed acyclic graph (DAG) of jobs

```
# visualize the DAG of jobs
$ snakemake --dag | dot | display
```

Workflows interoperability

COMMON WORKFLOW LANGUAGE

- Common format for bioinformatics tool & workflow execution
- Community based standards effort
- Designed for clusters & clouds
- Supports the use of containers (e.g. Docker)
- Specify data dependencies between steps
- Scatter/gather on steps
- Nest workflows in steps

- Develop your pipeline on your local computer (optionally with Docker)
- Execute on your research cluster or in the cloud
- Deliver to users via workbenches

by Carole Goble

More to read about workflows

- https://www.biostars.org/p/91301/
- https://github.com/common-workflow-language/common-workflow-language/wiki/Existing-Workflow-systems
- https://github.com/pditommaso/awesome-pipeline
- https://academic.oup.com/bib/article/18/3/530/2562749/A-review-of-bioinformatic-pipeline-frameworks

Genome & transcriptome assembly from single-cell data (Lecture & Workshop)

Andrey Prjibelski
SPSU, St. Petersburg

Tuesday, 14:00

While the majority of projects related to human health use reference-based analysis of sequencing data, de novo analysis is essential when studying previously unsequenced organisms. De novo genome assembly from short reads is a challenging algorithmic problem, complexity of which highly depends on the genome size and structure. Although bacterial genomes are small and typically do not have complex repeats, there are multiple ways of sequencing bacteria: conventional (isolate) sequencing, metagenomics (sequencing whole bacterial community at once) and single-cell sequencing (implies whole genome amplification).
During the first part of this class we will cover common NGS processing tools for QC and filtering, learn about basic assembly algorithms and assemble conventional and single-cell bacterial dataset. In addition, we will talk about differences between these ways of sequencing bacteria in terms of quality of raw data and assembly.
The second part will be devoted to RNA-Seq data processing: QC, filtering, de novo assembly and it evaluation. Although size de novo transcriptome assembly seems to be a less challenging problem than a genome assembly, it is amplified by highly uneven coverage depth (due to different expression levels) and presence of alternative splicing in eukaryotic genomes. During this class we will also discuss similarities and differences between genomic and transcriptomic data processing, de novo assembly and assembly evaluation.

Basic NGS tools

Andrey Prjibelski
Center for Algorithmic Biotechnology
SPbU

From the very beginning

...AACCCGTACGTTTTGCAAACGACCGT...

From the very beginning

- Sequencing

From the very beginning

- Sequencing
- Coverage

From the very beginning

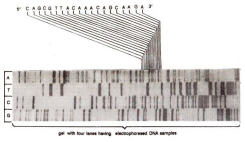
- Sequencing
- Coverage
- Errors
  - Mismatches

From the very beginning

- Sequencing
- Coverage
- Errors
  - Mismatches
  - Indels

Early days

- Sanger sequencing
  - Long reads (~900 bp)
  - Low coverage (< 10x)
  - Extreme cost



- Human genome project
  - 3 Gbp
  - 3 billion USD
  - 10 years

NGS

- Shorter reads (25-400bp)
- Low cost
- High coverage
- Huge amount of data
- Many new applications
- **Required new algorithms**

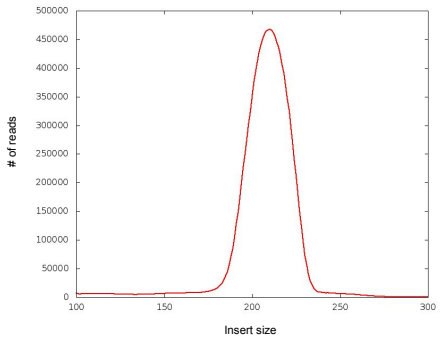
NGS technologies

	illumina	Roche 454 SEQUENCING	ion torrent	PACIFIC BIOSCIENCES	NANOPORE
Read length, bp	25-300	400-1100	200-400	1000-70000	5000-900000
Error rate	0.1-1%	1%	1-2%	10-20%	10-30%
Error type	Mismatches only	Indels & Mismatches	Indels & Mismatches	Indels & Mismatches	Indels & Mismatches
Comments	Error rate grows at the end of read	Problems with homopolymers	Problems with homopolymers	Errors distributed randomly	Typically several deletions in a row
$ per 1 Mbp	0.05 - 0.5	30	0.5 - 20	2+	0.1-0.5
Sequencer cost	100-500 K	100 K	80K	700 K	1 K

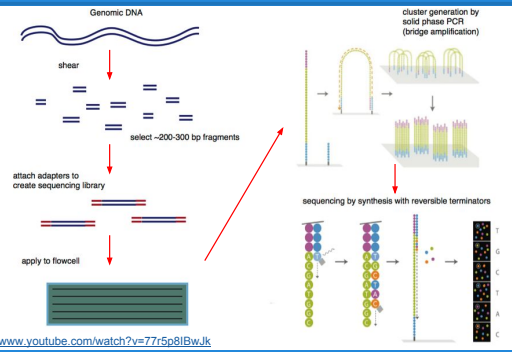
Paired reads

- Paired-end (< 1 kbp)
- Mate-pairs (1 - 20 kbp)

Insert size distribution

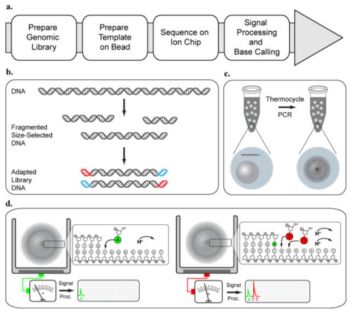


Illumina sequencing



http://www.youtube.com/watch?v=77r5p8IBwJk

IonTorrent sequencing



https://www.youtube.com/watch?v=WYBzbxIfuKs

FASTA/FASTQ

- FASTA

```
>EAS20_8_6_1_9_1972/1
ACCACCATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGC
>EAS20_8_6_1_163_1521/1
GCAGAAAACGTTCTGCATTTGCCACTGATGTACCGCCGAACTTCAACACTCGCA
```

- FASTQ

```
@EAS20_8_6_1_1477_92/1
ACCGTTACCTGTGGTAATGGTGATGGTGGTGGTAATGGTGGTGCTAATGCGTTT
+EAS20_8_6_1_1477_92/1
HHGHFHHHHHHHHHGFFHHHBG?GGC8DD9GF??=FFBCGBAF>FGCFHGHGGG
```

- Phred quality

$Q_{phred} = -10 \log_{10} e$

seqtk utility

- Subsampling
  sample
- Converting between interleaved/paired files
  mergepe, seq -1/-2
- fastq->fasta
  seq -A
- Quality trimming
- Shifting the quality
- Modifying names
- etc...

Quality Control

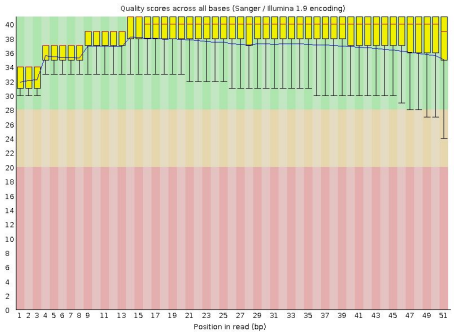
FastQC

- Easy and lightweight quality control for sequencing data
- Does not require reference genome

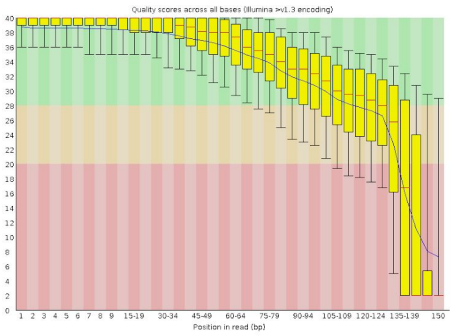
FastQC

- **fastqc** -h
- **mkdir** <output>
- **fastqc** <file1.fastq> <file2.fastq> …
  -o <output>

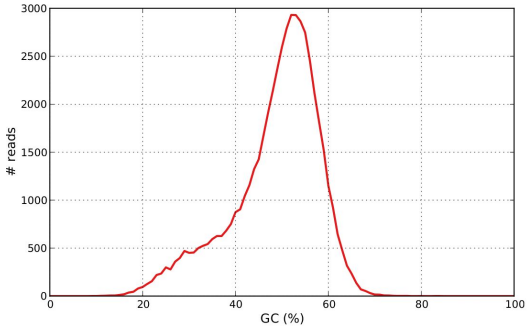
Per base sequence quality



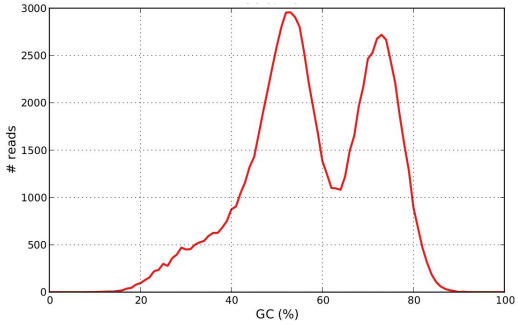
Per base sequence quality



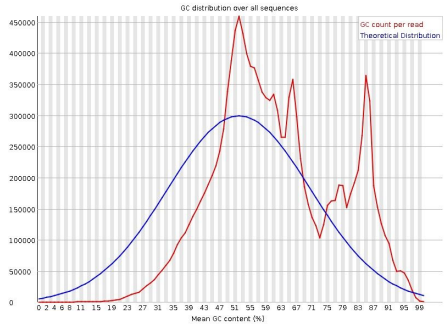
Per sequence GC content



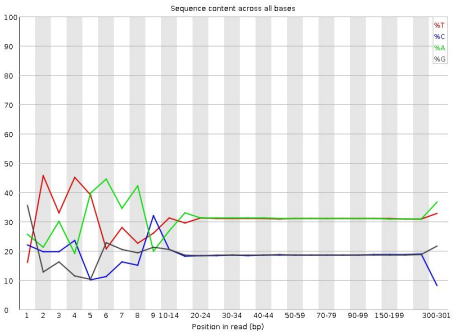
Per sequence GC content



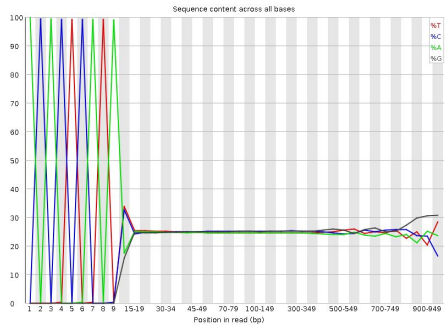
Per sequence GC content



Per base sequence content



Per base sequence content



Trimming and adapter removal

A lot of options

- Cutadapt
- Trimmomatic
- ...
- Skewer

Trimmomatic

java -jar trimmomatic-0.36.jar

Trimmomatic

- **SE** <input reads> <output reads> LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
- Remove leading low quality or N bases (below quality 3) (LEADING:3)
- Remove trailing low quality or N bases (below quality 3) (TRAILING:3)

Trimmomatic

- Scan the read with a 4-base wide sliding window, cutting when the average quality per base drops below 15 (SLIDINGWINDOW:4:15)
- Drop reads below the 36 bases long (MINLEN:36)

Trimmomatic

- **PE** <left reads> <right reads> <left paired> <left unpaired> <right paired> <right unpaired> OPTIONS
- ILLUMINACLIP:<path to adapters>
  - ILLUMINACLIP:TruSeq3-PE.fa

Short read alignment

Alignment

AACGCTAACGGTAA
AACCGCGAACTAA

Alignment

AACGCTAACGGTAA
AACCGCGAACTAA

AAC - GCTAACGGTAA
AACCGCGAAC - - TAA

Short read alignment

Find the read in the genome/assembly

Alignment applications

- Quality assessment
  - Error rate
  - Insert size distribution
  - Chimeric read/read-pairs
  - Genome fraction
- SNP calling
- Comparative analysis
  - CNVs
- Transcriptomics
  - Gene expression
  - Exon/intron detection

Short read alignment

- Challenges?

Short read alignment

- Challenges
  - Small length
  - Gigabytes of data
  - Different sequencing errors
  - SNPs
  - Genomic repeats
- Tools
  - Bowtie2, BWA-MEM
  - Bowtie2, BWA-SW, BWA-MEM (Multiple technologies)
  - TopHat, STAR (RNA-Seq)
  - and many more

Bowtie2

- **bowtie2** -h
- **bowtie2-build** <reference.fasta> <genome index>

Bowtie2

- **bowtie2 -x** <genome index> **-U** <single reads> **-S** <SAM output>

Bowtie2

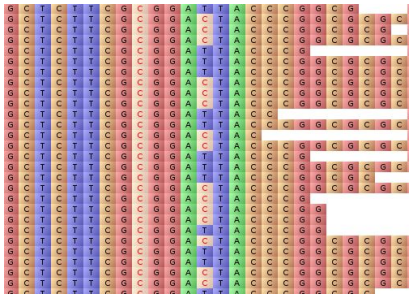
- **bowtie2 -x** <genome index> **-1** <left reads> **-2** <right reads> **-S** <SAM output>
- **bowtie2 -X** <max insert size> **-x** <genome index> **-1** <left reads> **-2** <right reads> **-S** <SAM output>

SAM files

- Read ID (QNAME)
- Reference ID (RNAME)
- Mapping position (POS)
- Mate reference ID (RNEXT)
- Mate position (PNEXT)
- Observed insert length (TLEN)
- Read sequence (SEQ)
- Read quality (QUAL)
- CIGAR string
  - 34M 1I 4M 2D 1X 3M

Alignment visualization with Tablet

https://ics.hutton.ac.uk/tablet/download-tablet/



Thank you!

Genome assembly with SPAdes

Introduction

Andrey Prjibelski
Center for Algorithmic Biotechnology
SPbU

Why to assemble?

Why to assemble?

- Sequencing data
  - Billions of short reads
  - Sequencing errors
  - Contaminants

Why to assemble?

- Sequencing data
  - Billions of short reads
  - Sequencing errors
  - Contaminants

Hard to perform analysis

Why to assemble?

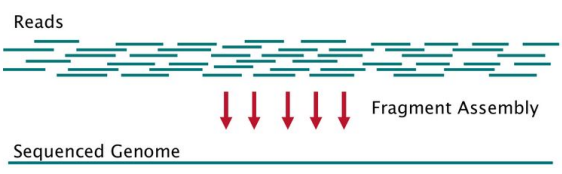
- Sequencing data
  - Billions of short reads
  - Sequencing errors
  - Contaminants

Hard to perform analysis

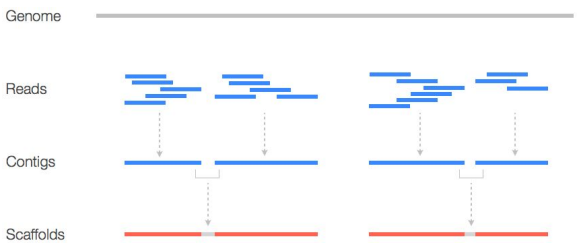
- Assembly
  - ✓ Corrects sequencing errors
  - ✓ Much longer sequences
  - ✓ Each genomic region is presented only once
  - × May introduce errors

Assembly basics

De novo whole genome assembly



De novo whole genome assembly

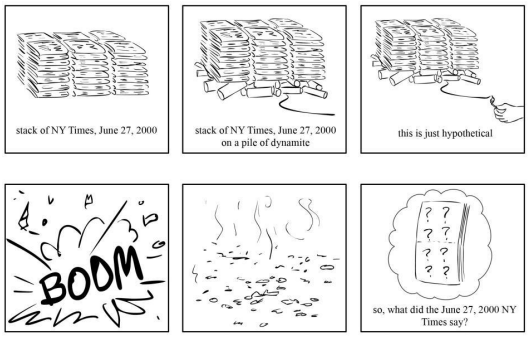


9

De novo whole genome assembly

10

De novo whole genome assembly



11

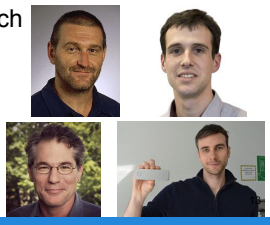
Early days

- Sanger sequencing
  - Long reads
  - Low coverage
- Overlap-Layout-Consensus (OLC)
  - Find overlaps between all reads (BLAST)
  - Order reads according to the overlaps
  - Merge reads into consensus string

12

NGS and OLC

- Overlap-Layout-Consensus is not applicable
  - Hard to find overlaps between short reads
  - Impossible to scale to such amount of reads
- De Bruijn graph approach
  (Pevzner et al., 2001)
  (Zerbino et al., 2008)
- String Graph approach
  (Meyers, 2005)
  (Simpson, Durbin 2011)



13

De Bruijn graph in a nutshell

He that mischief hatches, mischief catches

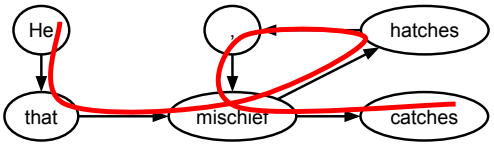
Sequencing

He that mischief
mischief hatches,
hatches, mischief
, mischief catches

14

De Bruijn graph in a nutshell

<u>, mischief catches</u>
<u>mischief hatches,</u>
<u>He that mischief</u>
<u>hatches, mischief</u>



15

De Bruijn graph

ACGTCCGTAA

16

De Bruijn graph

ACGTCCGTAA

k=2

AC

17

De Bruijn graph

ACGTCCGTAA

k=2

AC CG

18

De Bruijn graph

ACGTCCGTAA

k=2

AC CG GT

19

De Bruijn graph

ACGTCCGTAA

k=2

AC CG GT TC

20

De Bruijn graph

ACGTCCGTAA

k=2

AC CG GT CC TC

21

De Bruijn graph

ACGTCCGTAA

k=2

AC CG GT CC TC

22

De Bruijn graph

ACGTCCGTAA

k=2

AC CG GT CC TC

23

De Bruijn graph

ACGTCCGTAA

k=2

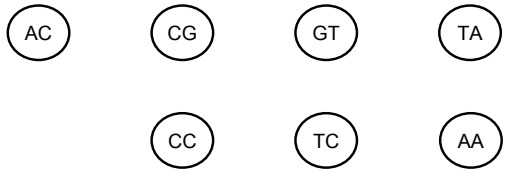
AC CG GT TA CC TC

24

De Bruijn graph

ACGTCCGTAA

k=2

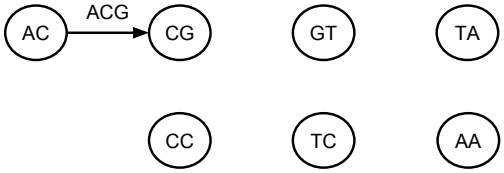


25

De Bruijn graph

ACGTCCGTAA

k=2

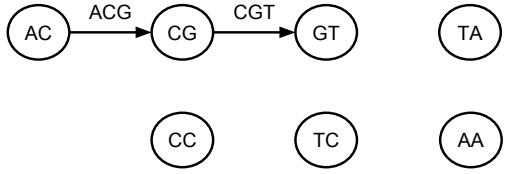


26

De Bruijn graph

ACGTCCGTAA

k=2

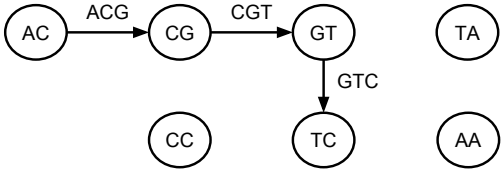


27

De Bruijn graph

ACGTCCGTAA

k=2

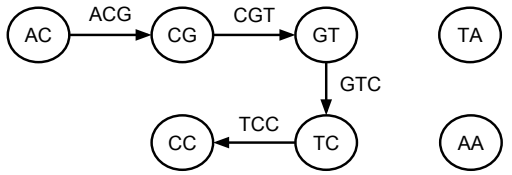


28

De Bruijn graph

ACGTCCGTAA

k=2

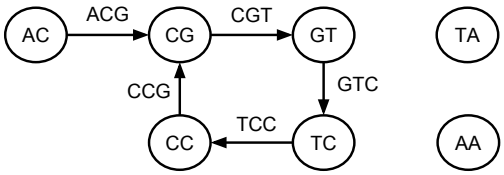


29

De Bruijn graph

ACGTCCGTAA

k=2

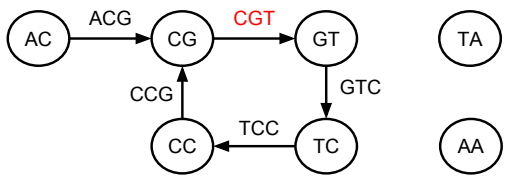


30

De Bruijn graph

ACGTCCGTAA

k=2

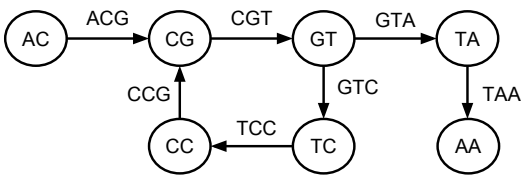


31

De Bruijn graph

ACGTCCGTAA

k=2

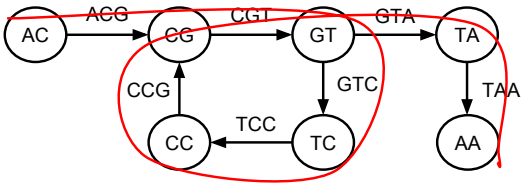


32

De Bruijn graph

ACGTCCGTAA

k=2

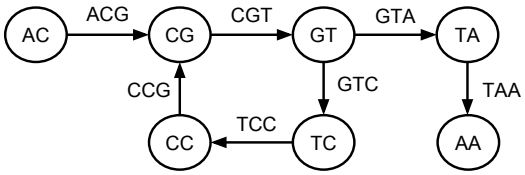


33

Condensed de Bruijn graph

ACGTCCGTAA

k=2

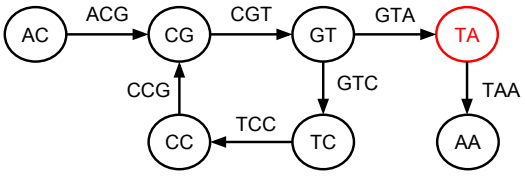


34

Condensed de Bruijn graph

ACGTCCGTAA

k=2

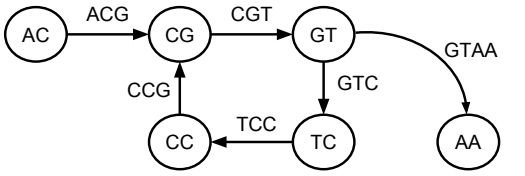


35

Condensed de Bruijn graph

ACGTCCGTAA

k=2

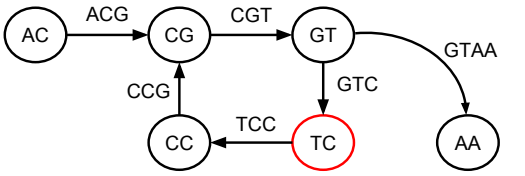


36

Condensed de Bruijn graph

ACGTCCGTAA

k=2

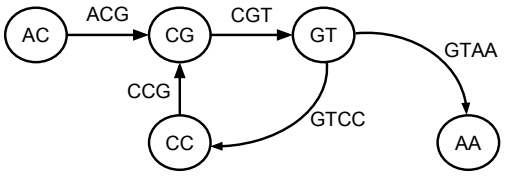


37

Condensed de Bruijn graph

ACGTCCGTAA

k=2

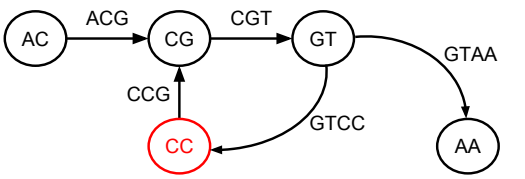


38

Condensed de Bruijn graph

ACGTCCGTAA

k=2

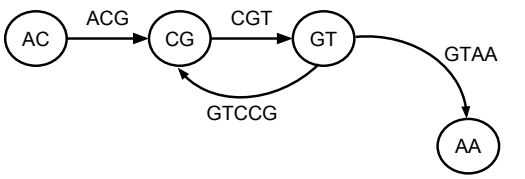


39

Condensed de Bruijn graph

ACGTCCGTAA

k=2

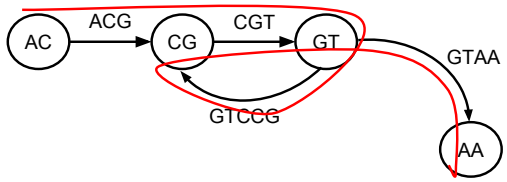


40

Condensed de Bruijn graph

ACGTCCGTAA

k=2



41

What about real data?

CCGTTG
TGCAGG
GTTGCA

k=3

42

What about real data?

CCG

CCGTTG
TGCAGG
GTTGCA

k=3

43

What about real data?

CCG CGT

CCGTTG
TGCAGG
GTTGCA

k=3

44

What about real data?

CCG CGT

GTT

CCGTTG
TGCAGG
GTTGCA

k=3

45

What about real data?

CCG CGT

GTT TTG

CCGTTG
TGCAGG
GTTGCA

k=3

46

What about real data?

CCG CGT

GTT TTG

TGC GCA

CAG AGG

CCGTTG
TGCAGG
GTTGCA

k=3

47

What about real data?

CCG CGT

GTT TTG

TGC GCA

CAG AGG

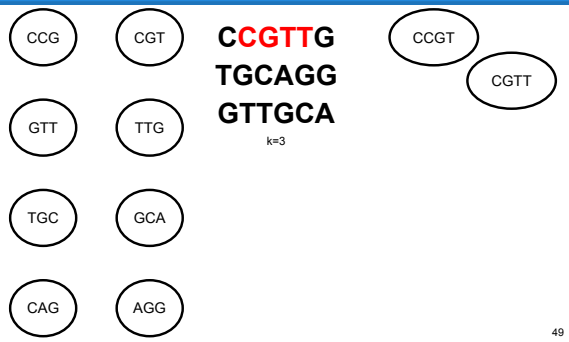
CCGTTG
TGCAGG
GTTGCA

k=3

CCGT

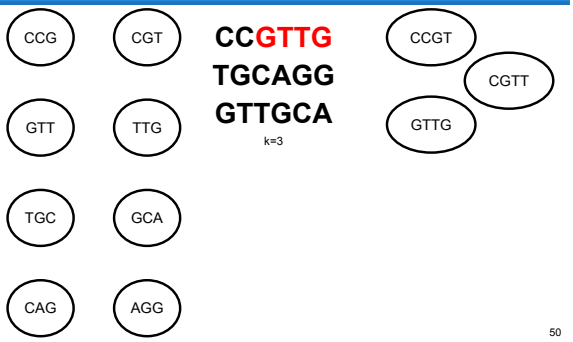
48

What about real data?



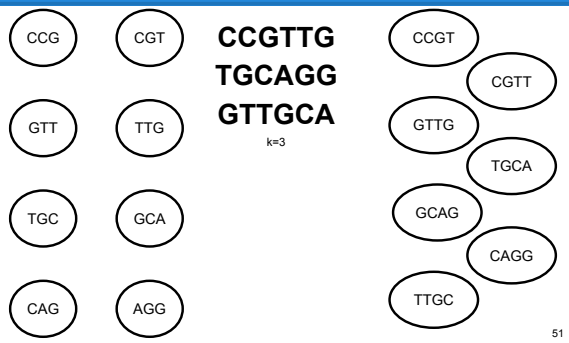
49

What about real data?



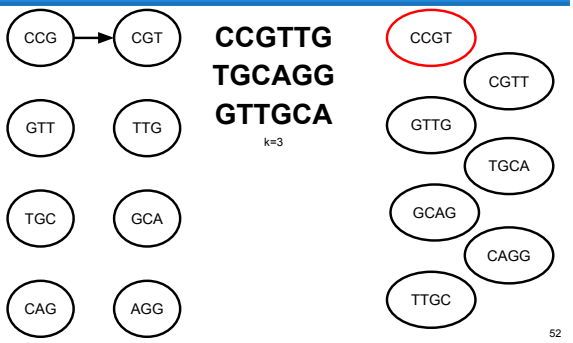
50

What about real data?



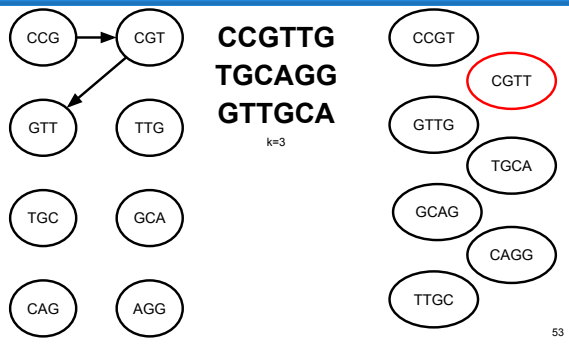
51

What about real data?



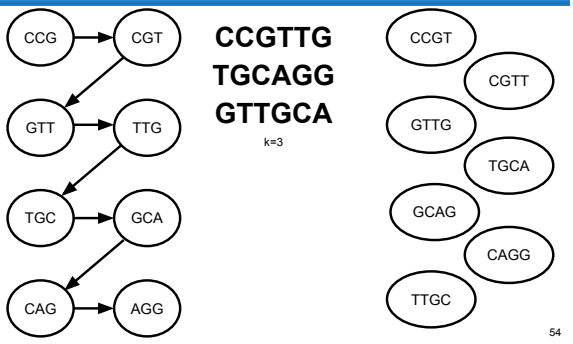
52

What about real data?



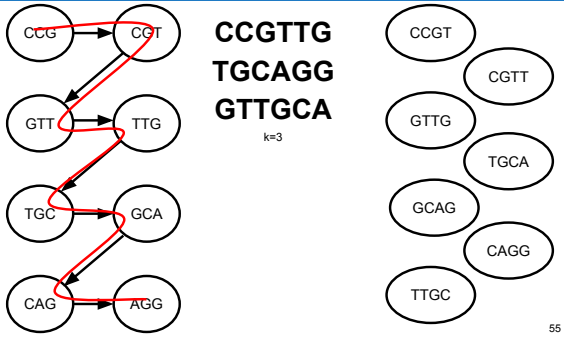
53

What about real data?



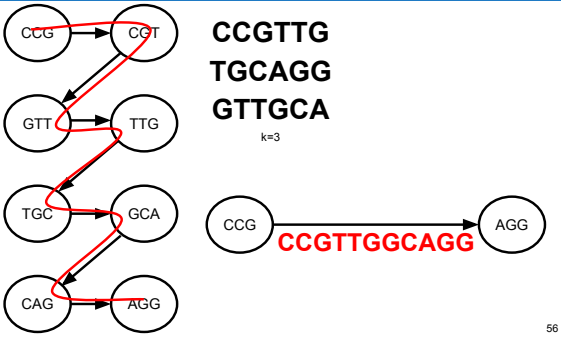
54

What about real data?



55

What about real data?

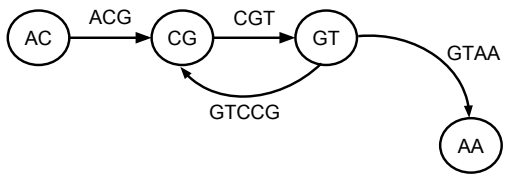


56

Repeats in de Bruijn graph

ACGTCCGTAA

k=2

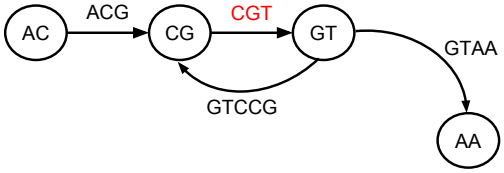


57

Repeats in de Bruijn graph

ACGTCCGTAA

k=2

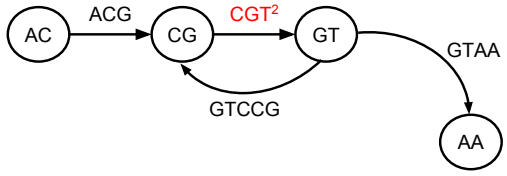


58

Repeats in de Bruijn graph

ACGTCCGTAA

k=2

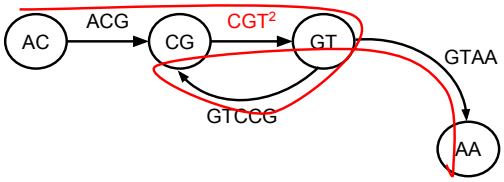


59

Eulerian path with multiplicities

ACGTCCGTAA

k=2



60

Oh, repeats...

- Ribosomal operons (5-8 kbp)
- ALU, SINEs
  - < 1 kbp, extremely high multiplicity
- LINEs
  - >> 1 kbp, high multiplicity
- Tandem repeats

61

Oh, repeats...

NCBI contains assemblies with 100K+ scaffolds!

"These are not the genomes I wanted you to assemble"

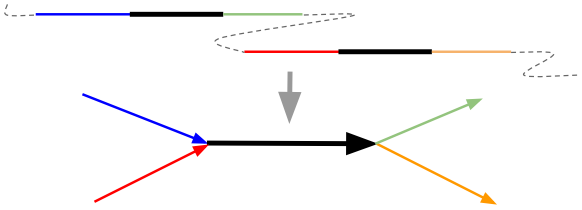
Gene Meyers

62

Resolving repeats

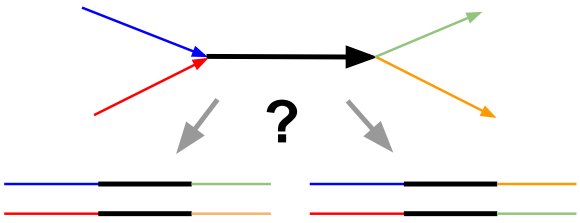
63

Resolving repeats



64

Resolving repeats

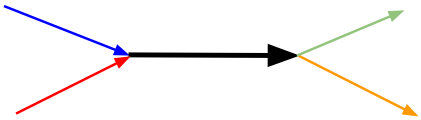


65

Paired reads

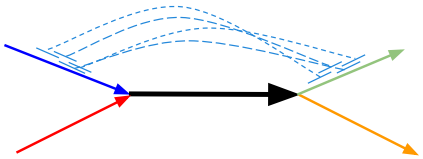
66

Resolving repeats



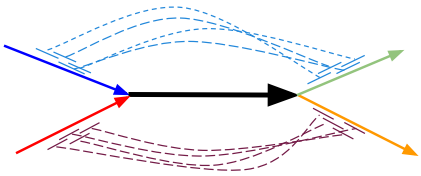
67

Resolving repeats



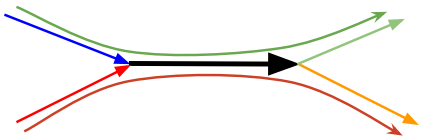
68

Resolving repeats



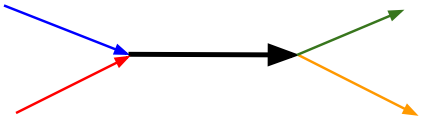
69

Resolving repeats



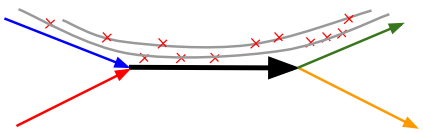
70

Long reads to the rescue



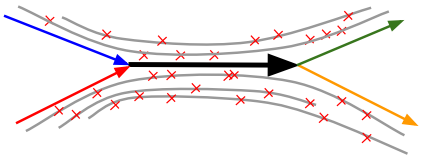
71

Long reads to the rescue



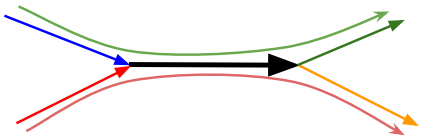
72

Long reads to the rescue



73

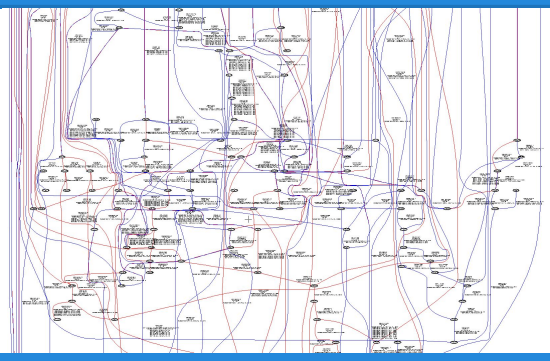
Long reads to the rescue



74

Real life

Part of *E.coli* genome, K = 99



75

SPAdes assembler

SPAdes genome assembler

- Data types
  - Small genomes (bacteria, fungi)
  - Standard data sets
  - MDA single-cell data sets
- Sequencing technologies
  - Illumina
  - IonTorrent

77

SPAdes genome assembler

- Hybrid assembly
  - Paired-end reads
  - Mate-pairs
  - PacBio
  - Oxford Nanopore
  - Sanger
  - Previous assemblies
- Works with high-coverage projects (1000x+)
- Fast & efficient
- User-friendly

78

SPAdes genome assembler

- Requirements
  - System
    - 64-bit Linux
    - Mac OS
  - Python 2.4 or higher

- bioinf.spbau.ru/spades/

79

SPAdes first steps

- spades.py

80

SPAdes first steps

- spades.py
- spades.py --help
- spades.py --test

81

SPAdes first steps

- spades.py
- spades.py --help
- spades.py --test
- -o **<output_dir>**

82

Input data formats

- FASTA: .fasta / .fa
- FASTQ: .fastq / .fq
- Gzipped: .gz
- Unmapped BAM (IonTorrent): .bam

83

Input data options

Unpaired reads

- Illumina unpaired
- IonTorrent

84

Input data options

Unpaired reads

- Illumina unpaired
- IonTorrent
- -s ***single.fastq***
- -s ***single1.fastq*** -s ***single2.fastq*** ...

85

Input data options

Paired-end reads

- Interlaced pairs in one file

```
>left_read_id
ACGTGCAGG…
>right_read_id
GCTTCGAGG…
```

- Separate files

file1.fastq

```
>left_read_id
ACGTGCAGG…
```

file2.fastq

```
>right_read_id
GCTTCGAGG…
```

86

Input data options

Paired-end reads

- Interlaced pairs in one file
  --pe1-12 ***file.fastq***

- Separate files
  --pe1-1 ***file1.fastq*** --pe1-2 ***file2.fastq***

87

Input data options

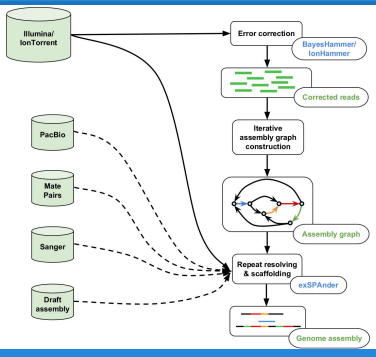
Paired-end reads

- Interlaced pairs in one file
  --pe1-12 ***file.fastq***

- Separate files
  --pe1-1 ***file1.fastq*** --pe1-2 ***file2.fastq***
  --pe1-s ***unpaired.fastq***

88

SPAdes pipeline



89

Pipeline options

- Run only error correction (I will use my own assembler)
  - --only-error-correction

90

Pipeline options

- Run only error correction (I will use my own assembler)
  - --only-error-correction
- Run only assembler (input reads are already corrected or quality-trimmed)
  - --only-assembler

91

Pipeline options

- Run only error correction (I will use my own assembler)
  - --only-error-correction
- Run only assembler (input reads are already corrected or quality-trimmed)
  - --only-assembler
- Run mismatch correction after the assembly
  - --careful

92

SPAdes performance options

- Number of threads

  -t ***N***
- Maximal available RAM (GB)
  - SPAdes will terminate if exceeded

  -m ***M***

93

Restarting SPAdes

- SPAdes / system crashed
  - --continue -o ***your_output_dir***

94

Restarting SPAdes

- SPAdes / system crashed
  - --continue -o ***your_output_dir***
- You forgot some options
  - --restart-from **check_point**
  - **ec** — error correction
  - **as** — assembly
  - **mc** — mismatch correction
  - **k##** --- specific k value

95

Why to create new assembler?

96

Single-cell sequencing

How to sequence bacteria?

98

How to sequence bacteria?

Sequencing requires a lot of DNA

99

How to sequence bacteria?

- Conventional sequencing

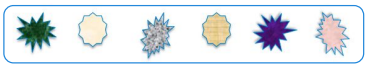
  - >99% of bacteria cannot be cultivated in the lab

100

How to sequence bacteria?

- Conventional sequencing

- Metagenomics



  - Reads from different genomes mixed in one data set
  - Hard to assemble and classify resulting sequences

101

Metagenomics

- Metagenomics: sequencing of whole bacterial community
  - Reads from dozens of different genomes mixed in one data set
  - Different coverage for different bacteria
  - Presence of different strains
  - Conservative genomic regions
- Hard to assemble and classify resulting sequences

102

How to sequence bacteria?

- Conventional sequencing

- Metagenomics

- Single cell

  - Needs whole genome amplification

103

Single-cell sequencing via MDA

Multiple Displacement Amplification

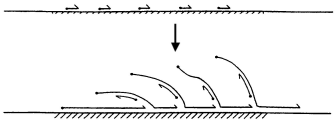
- Random hexamer primers

104

Single-cell sequencing via MDA

Multiple Displacement Amplification

- Random hexamer primers
- Phi29 DNA polymerase strand displacement

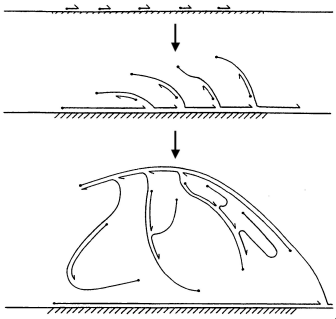


105

Single-cell sequencing via MDA

Multiple Displacement Amplification

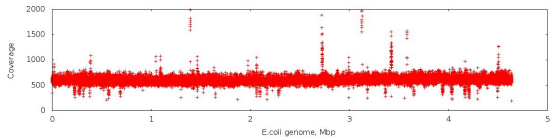
- Random hexamer primers
- Phi29 DNA polymerase strand displacement



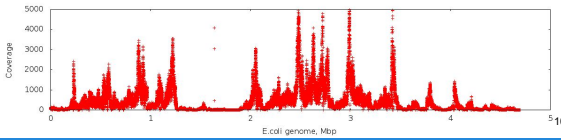
106

Challenges in single-cell assembly

- *E. coli* isolate dataset

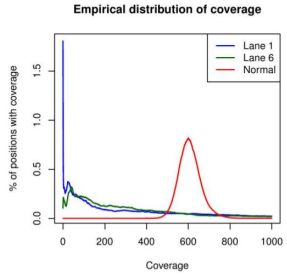


- *E.coli* single-cell dataset



107

Challenges in single-cell assembly

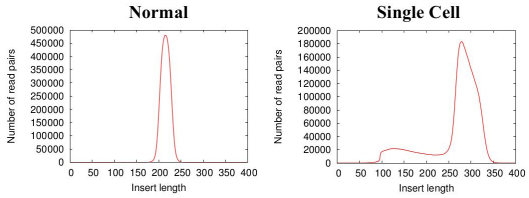


A cutoff threshold will eliminate about 25% of valid data in the single cell case, whereas it eliminates noise in the normal multicell case.

108

Challenges in single-cell assembly

- Insert size deviation



- Chimeric reads
  - Isolate dataset 0.01%
  - Single-cell dataset ~2%

109

Single-cell assembly techniques

How to select K?

- SPAdes is iterative assembler
  - Uses several K values iteratively
  - Output is from the last iteration
- K is selected automatically
  - Read length
  - Data type
- Setting K manually (not recommended)
  - -k ***21,33,55***

111

Why iterative?

- Small K

112

Why iterative?

- Small K
  - Tangled graph
  - Unresolved short repeats
  - Short contigs

ACGGATC

TTGGAAG

k = 2 **k = 4**

113

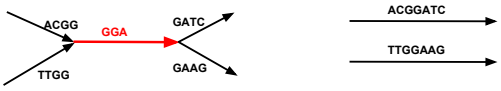
Why iterative?

- Small K
  - Tangled graph
  - Unresolved short repeats
  - Short contigs

ACGGATC

TTGGAAG

k = 2 **k = 4**



114

Why iterative?

- Small K
  - Tangled graph
  - Unresolved short repeats
  - Short contigs
- Large K

115

Why iterative?

- Small K
  - Tangled graph
  - Unresolved short repeats
  - Short contigs
- Large K
  - Many gaps

ACCGT GTAAT

k = 2 **k = 4**

116

Why iterative?

- Small K
  - Tangled graph
  - Unresolved short repeats
  - Short contigs
- Large K
  - Many gaps

ACCGT GTAAT

k = 2 **k = 4**

117

Why iterative?

- Small K
  - Tangled graph
  - Unresolved short repeats
  - Short contigs
- Large K
  - Many gaps

ACCGT GTAAT

k = 2 **k = 4**

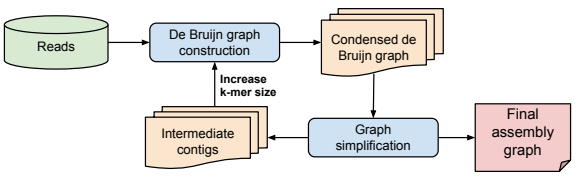
118

Why iterative?

- Small K
  - Tangled graph
  - Unresolved short repeats
  - Short contigs
- Large K
  - Many gaps
- Iterative run
  - Contigs constructed with small K re used as reads to close gaps
  - Last iteration has larger K to resolve short repeats

119

Iterative SPAdes run



- Smaller *k*-mer sizes are needed for reconstructing low-coverage regions
- Larger *k*-mer sizes are needed for resolving short repeats

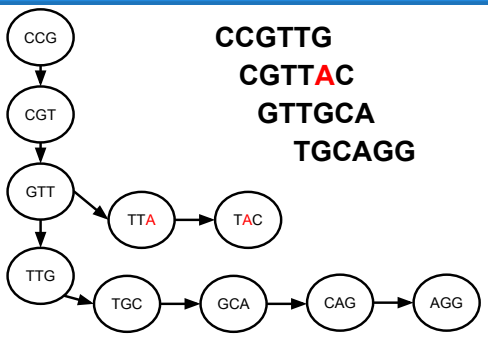
120

What about sequencing errors?

CCGTTG
CGTTAC
GTTGCA
TGCAGG

121

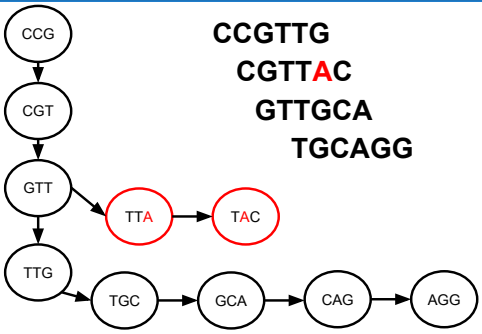
What about sequencing errors?



CCGTTG
CGTTAC
GTTGCA
TGCAGG

122

What about sequencing errors?

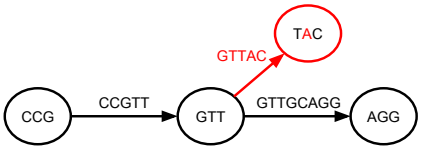


CCGTTG
CGTTAC
GTTGCA
TGCAGG

123

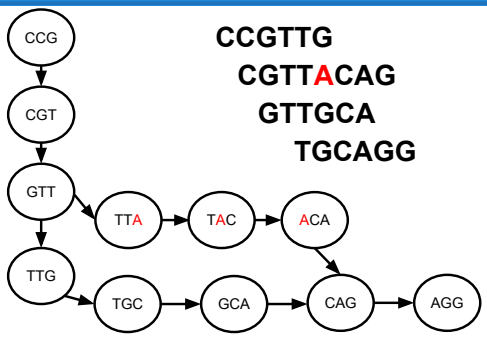
What about sequencing errors?

CCGTTG
CGTTAC
GTTGCA
TGCAGG



124

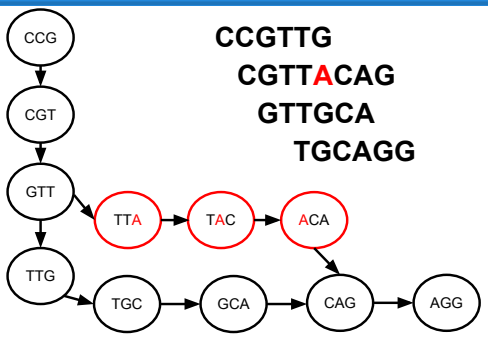
More about sequencing errors



CCGTTG
CGTTACAG
GTTGCA
TGCAGG

125

More about sequencing errors

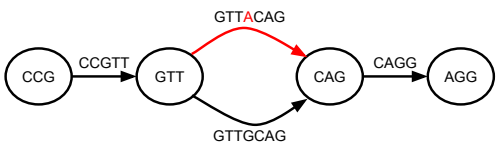


CCGTTG
CGTTACAG
GTTGCA
TGCAGG

126

More about sequencing errors

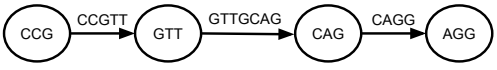
CCGTTG
CGTTACAG
GTTGCA
TGCAGG



127

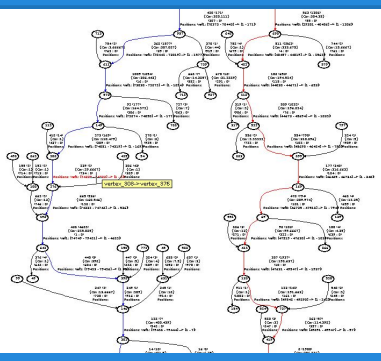
More about sequencing errors

CCGTTG
CGTTACAG
GTTGCA
TGCAGG



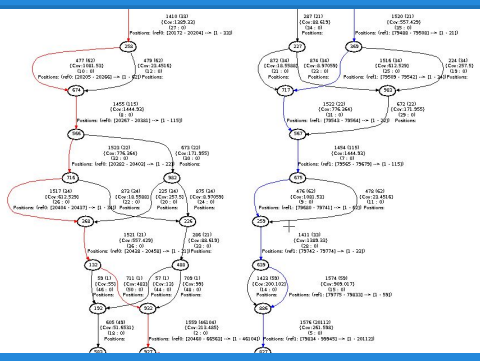
128

Real life



129

Real life

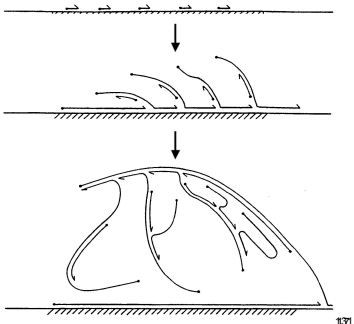


130

Single-cell sequencing via MDA

Multiple Displacement Amplification

- Random hexamer primers
- Phi29 DNA polymerase strand displacement



131

Chimeric junctions

CCGTTG
CGTTGC
GTTGCA
ATTTAA
TTAAAG
TAAAGG
TTGTAA

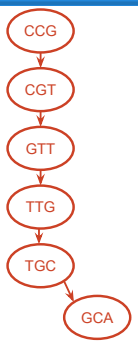
132

Chimeric junctions

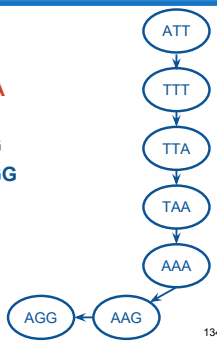
CCGTTG
CGTTGC
GTTGCA
ATTTAA
TTAAAG
TAAAGG
TTGTAA

133

Chimeric junctions

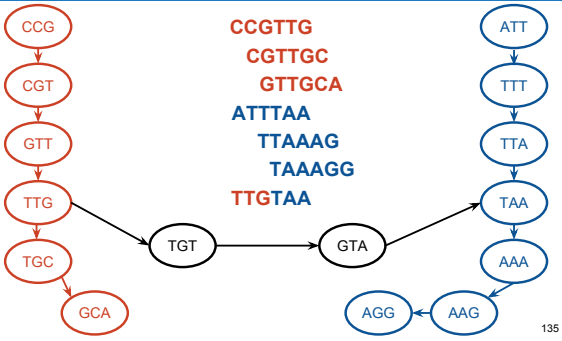


CCGTTG
CGTTGC
GTTGCA
ATTTAA
TTAAAG
TAAAGG
TTGTAA



134

Chimeric junctions



CCGTTG
CGTTGC
GTTGCA
ATTTAA
TTAAAG
TAAAGG
TTGTAA

135

Assembling single-cell data

- Add option

  --sc

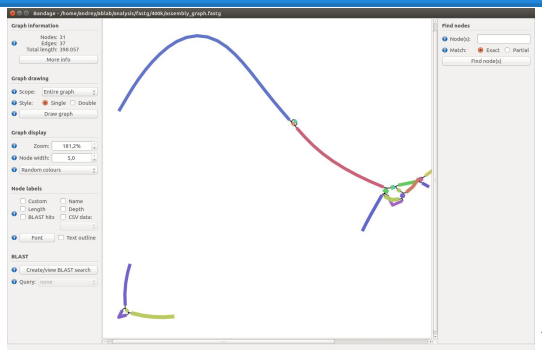
136

Visualizing assembly graph

- SPAdes outputs assembly graph and contigs paths in 2 formats
  - FASTG + .paths files
  - GFA
- https://github.com/rrwick/Bandage/

137

Bandage



138

Thank you!

Questions?

http://cab.spbu.ru/software/spades/

Cite

- Bankevich et al., SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. JCB, 2012.
- Nurk et al., Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. JCB, 2013.
- Prjibelski et al., ExSPAnder: a universal repeat resolver for DNA fragment assembly. Bioinformatics, 2014.

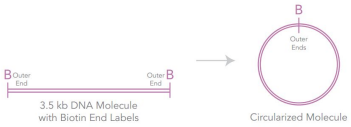
139

Extra slides

Mate-pairs

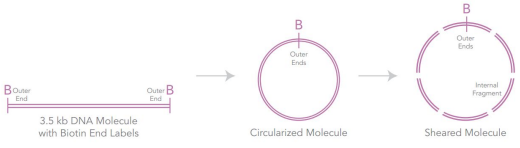
141

Mate-pairs



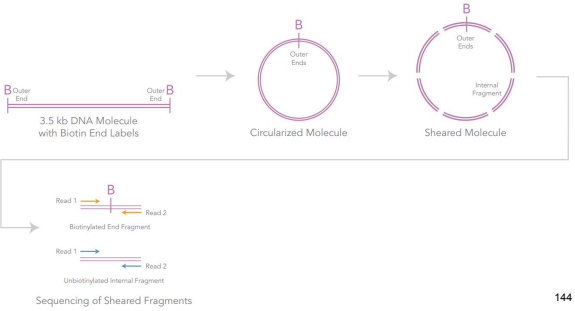
142

Mate-pairs



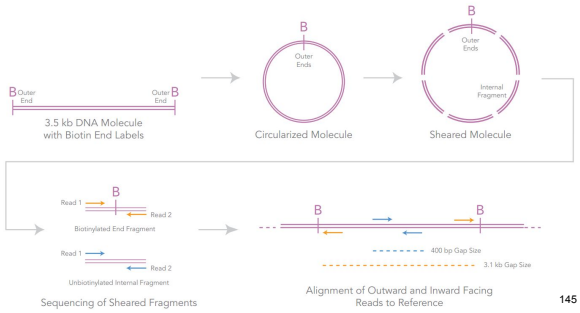
143

Mate-pairs



144

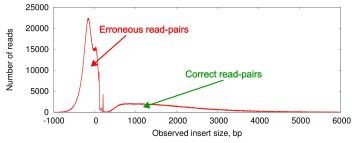
Mate-pairs



145

Mate-pairs

Conventional mate-pairs:



146

Input data options

Mate-pair reads

- Cannot be used separately
- Interlaced pairs in one file

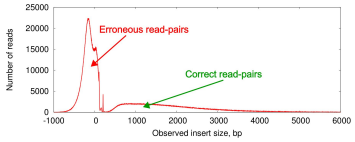
  --mp1-12 ***mp.fastq***
- Separate files

  --mp1-1 ***mp1.fastq*** --mp1-2 ***mp2.fastq***

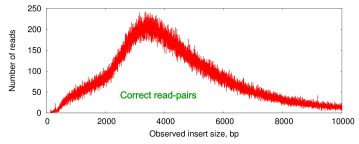
147

Nextera mate-pairs

Conventional mate-pairs:



Illumina Nextera mate-pairs:



148

Input data options

High-quality mate-pair reads (e.g. Nextera MP)

- Can be used separately
- Interlaced pairs in one file

  --hqmp1-12 ***hqmp.fastq***
- Separate files

  --hqmp1-1 ***hqmp1.fastq***

  --hqmp1-2 ***hqmp2.fastq***

149

Input data options

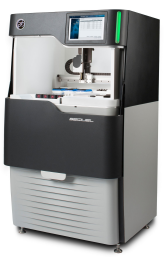
Read-pairs orientation

- Forward-reverse →←
  - Default for paired-end and high quality mate-pair
- Reverse-forward ←→
  - Default for conventional mate-pairs
- Forward-forward →→
- --pe1-**fr** / --pe1-**rf** / --pe1-**ff**
- --mp2-**fr** / --mp2-**rf** / --mp2-**ff**

150

Pacific Biosciences

- Up to 70 kbp long
- Much cheaper than Sanger
- **10-20% error rate**



Oxford NanoPores

- In 2010 announced whole genome sequencing
- Sequencer as small as USB stick
- Longest reported read — 200 kbp
- **15-30% error rate**

Hybrid assembly options

- PacBio CLR
  - --pacbio ***pb.fastq***
- Oxford Nanopore reads
  - --nanopore nanopore_reads.fastq
- Sanger reads
  - --sanger ***sanger.fastq***
- Additional contigs
  - --trusted-contigs ***contigs.fa***
  - --untrusted-contigs ***contigs.fa***

153

PacBio only assembly

Thm:
Perfect assembly possible iff
a) errors random
b) sampling is Poisson
c) reads long enough 2 solve repeats.
Note: e-rate not needed

Gene Meyers' twitter

PacBio only assembly

- *D. melanogaster*
- Assembled with Illumina
  - N50 = 100 kbp
- Assembled with PacBio P5
  - N50 = 21 Mbp
  - Assembled new highly repetitive regions

Input data types

- Standard Illumina
  - default
- Single-cell data sets
  - --sc
- IonTorrent data
  - --iontorrent

156

Compiling SPAdes

- Requirements
  - 64-bit Linux based OS
  - g++ 4.7+
  - cmake 2.12+
  - zlib
  - bzlib
- ./spades_compile.sh
- PREFIX=**/usr/local** ./spades_compile

157

SPAdes warnings and errors

- Provided at the end of log
- spades.support@cab.spbu.ru
- Don't forget to attach spades.log and params.txt

158

SPAdes for Cloud Platforms

SPAdes runs on:

- Illumina BaseSpace
- DNAnexus
- TorrentServer
- Galaxy (available from Galaxy Tool Shed)

OLC

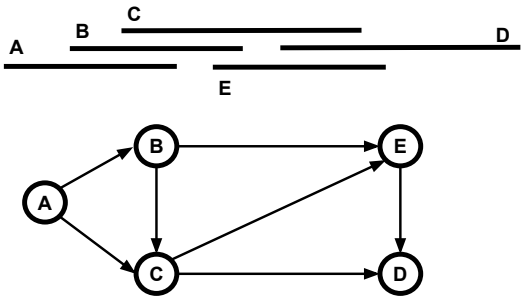
159

Early days

- Sanger sequencing
  - Long reads
  - Low coverage
- Overlap-Layout-Consensus (OLC)
  - Find overlaps between all reads
  - BLAST and similar algorithms
  - Ignore "insufficient" overlaps
    - At least 40bp
    - >94% similarity

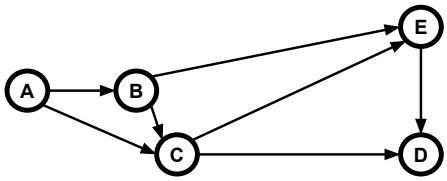
161

Overlap graph



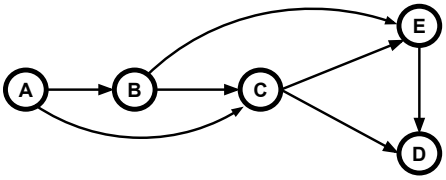
162

Layout



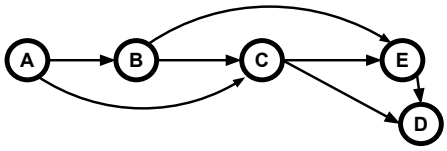
163

Layout



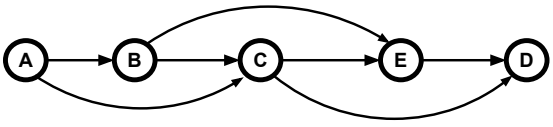
164

Layout



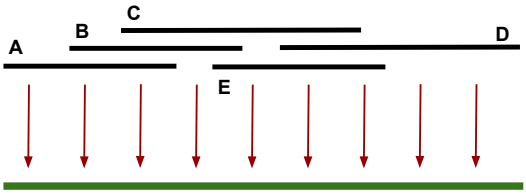
165

Layout



166

Consensus

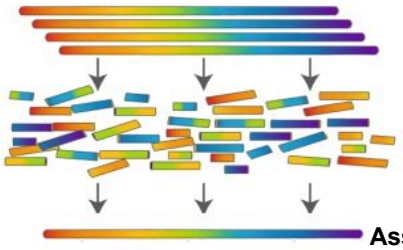


167

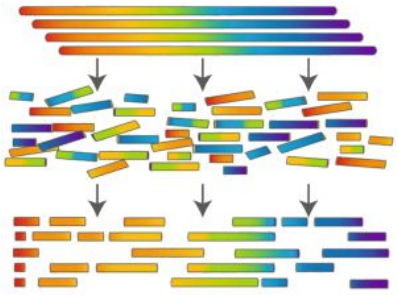
Genome assembly evaluation with QUAST

Andrey Prjibelski
Center for Algorithmic Biotechnology
SPbU

In a perfect world



In reality



Which assembler to use?

- ABySS
- ALLPATHS-LG
- CLC
- IDBA-UD
- MaSuRCA
- MIRA
- Ray
- SOAPdenovo
- SPAdes
- Velvet
- and many more...

Which assembler to use?

- Assemblathon 1 & 2
  - Simulated and real datasets
  - More than 30 teams competing
- Independent studies
  - Papers (GAGE, GAGE-B, GABenchToB)
  - Web-sites (nucleotid.es, …)
  - Surveys
- Genome assembly evaluation tools
  - QUAST
  - GAGE

There is no best assembler

Which assembler to use?

- Different technologies (Illumina, 454, IonTorrent, ...)
- Genome type and size (bacteria, insects, mammals, plants, ...)
- Type of prepared libraries (single reads, paired-end, mate-pairs, combinations)
- Type of data (multicell, metagenomic, single-cell)

Who needs assembly evaluation?

Who needs assembly evaluation?

- Assembler developers
  - To find weak points
  - To publish papers

Who needs assembly evaluation?

- Assembler developers
  - To find weak points
  - To publish papers
- Researchers
  - To choose best assembler for their data
  - To perform assembly QC

Assembly evaluation

- Basic evaluation
  - No extra input
  - Very quick
- Reference-based evaluation
  - A lot of metrics
  - Very accurate
- *De novo* evaluation
  - Advanced analysis of *de novo* assemblies

Basic statistics

- Only assemblies are needed (no additional input)
- Very fast to compute

Contig sizes

- Number of contigs

Contig sizes

- Number of contigs
- Number of large contigs (i.e. > 1000 bp)

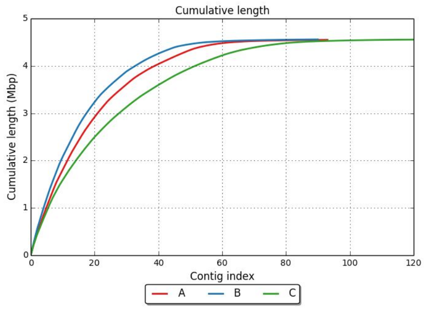
Contig sizes

- Number of contigs
- Number of large contigs (i.e. > 1000 bp)
- Largest contig length

Contig sizes

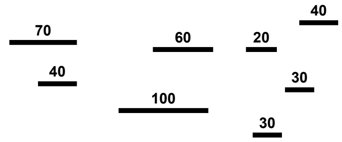
- Number of contigs
- Number of large contigs (i.e. > 1000 bp)
- Largest contig length
- Total assembly length

Cumulative length plot



N50

The maximum length **X** for which the collection of all contigs of length >= **X** covers at least **50**% of the assembly

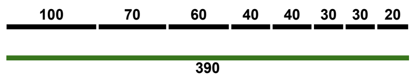


N50

The maximum length **X** for which the collection of all contigs of length >= **X** covers at least **50**% of the assembly

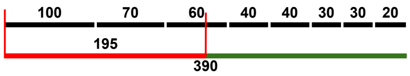
N50

The maximum length **X** for which the collection of all contigs of length >= **X** covers at least **50**% of the assembly



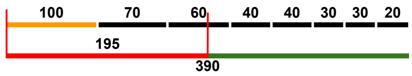
N50

The maximum length **X** for which the collection of all contigs of length >= **X** covers at least **50**% of the assembly



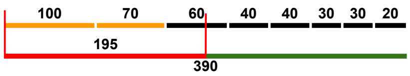
N50

The maximum length **X** for which the collection of all contigs of length >= **X** covers at least **50**% of the assembly



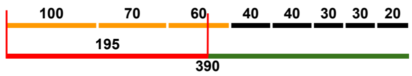
N50

The maximum length **X** for which the collection of all contigs of length >= **X** covers at least **50**% of the assembly



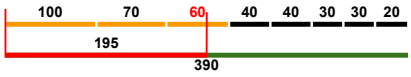
N50

The maximum length **X** for which the collection of all contigs of length >= **X** covers at least **50**% of the assembly



N50

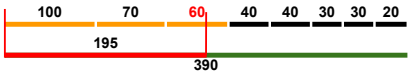
The maximum length **X** for which the collection of all contigs of length >= **X** covers at least **50**% of the assembly



N50 = 60

L50

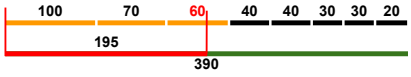
The minimum number **X** such that **X** longest contigs cover at least **50**% of the assembly



L50 =

L50

The minimum number **X** such that **X** longest contigs cover at least **50**% of the assembly



L50 = 3

N50-variations

- N25, N75
- L25, L75

N25 = , N75 =
L25 = , L75 =

N50-variations

- N25, N75
- L25, L75

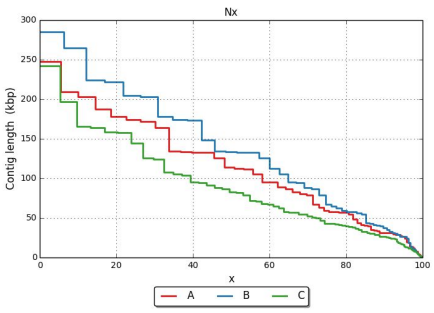
N25 = 100, N75 = 40
L25 = 1, L75 = 5

N50-variations

- N25, N75
- L25, L50, L75

N50-variations

- N25, N75
- L25, L50, L75
- Nx, Lx



Other

- Number of N's per 100 kbp

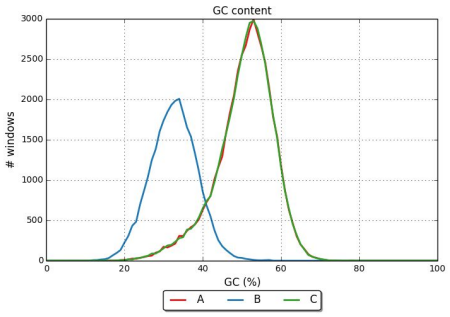
Other

- Number of N's per 100 kbp
- GC %

Other

- Number of N's per 100 kbp
- GC %
- Distributions of GC % in small windows:

Other



Reference-based metrics

- A lot of metrics
- Accurate assessment

Basic reference statistics

- Reference length
- Reference GC %
- Number of chromosomes

Basic reference statistics

- Reference length
- Reference GC %
- Number of chromosomes
- Number of genes/operons

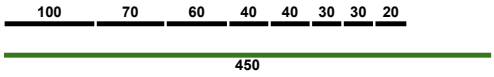
Basic reference statistics

- Reference length
- Reference GC %
- Number of chromosomes
- Number of genes/operons

- NGx, LGx

Basic reference statistics

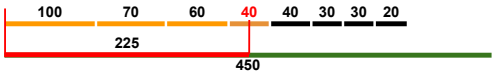
- NGx, LGx



NG50 =
LG =

Basic reference statistics

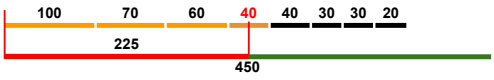
- NGx, LGx



NG50 =
LG =

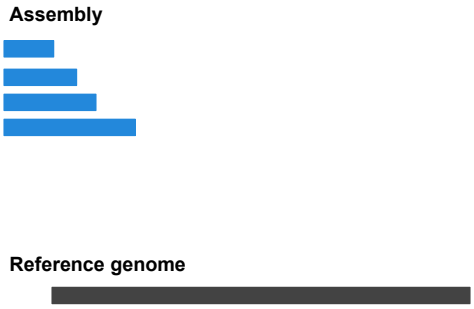
Basic reference statistics

- NGx, LGx



NG50 = 40
LG = 4

Alignment statistics



Alignment statistics

Alignment statistics

- Genome fraction %

Alignment statistics

- Genome fraction %
- Duplication ratio

Alignment statistics

- Genome fraction %
- Duplication ratio
- Number of gaps

Alignment statistics

- Genome fraction %
- Duplication ratio
- Number of gaps
- Largest alignment length

Alignment statistics

- Genome fraction %
- Duplication ratio
- Number of gaps
- Largest alignment length
- Number of unaligned contigs (full & partial)

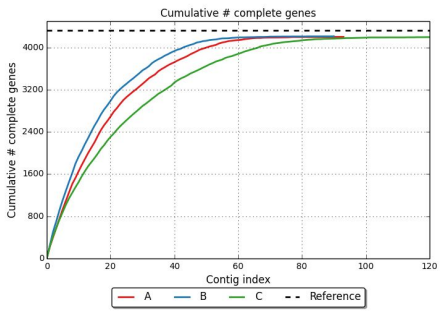
Alignment statistics

- Genome fraction %
- Duplication ratio
- Number of gaps
- Largest alignment length
- Number of unaligned contigs (full & partial)
- Number of mismatches/indels per 100 kbp

Alignment statistics

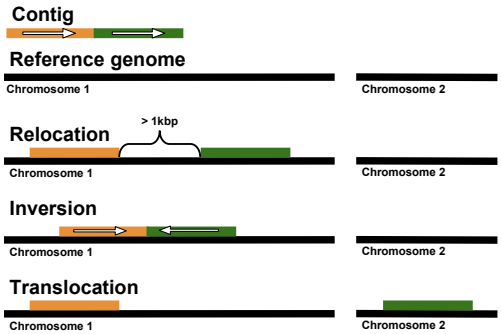
- Genome fraction %
- Duplication ratio
- Number of gaps
- Largest alignment length
- Number of unaligned contigs (full & partial)
- Number of mismatches/indels per 100 kbp
- Number of genes/operons (full & partial)

Alignment statistics



Misassemblies

Misassemblies

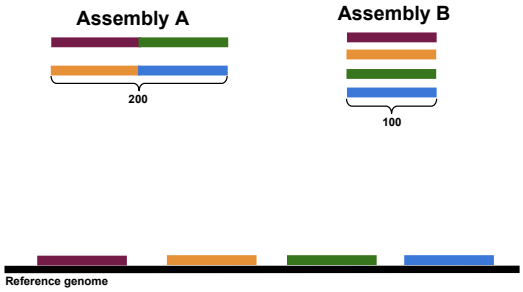


NB!

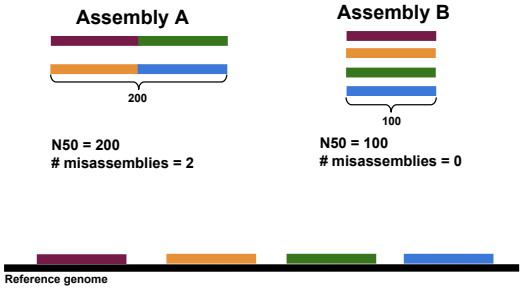
There is **no** best metric

NA50

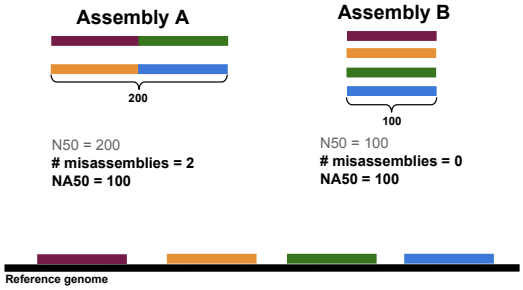
NA50



NA50



NA50



QUAST

Quast

QUality ASsesment Tool
for Genome Assemblies

- Assembly statistics
  - Basic statistics
  - Reference-based evaluation
  - Simple *de novo* evaluation
- Available as a web-based and a command line tool
- quast.sf.net

QUAST: console tool

quast.sf.net/manual

- Installation
- Options & Input data
- Metrics & Plots
- Output

QUAST: console tool

- quast.py
- quast.py --help

QUAST basics

- quast.py
- quast.py --help
- quast.py **contigs.fasta**
- quast.py *[options]* **contigs.fasta**
- quast.py -o **out_dir contigs.fasta**

Reference options

- Reference genome
  - -R ***reference.fasta***
- Gene annotation
  - -G ***genes.gff***
- Operon annotation
  - -O ***operons.gff***

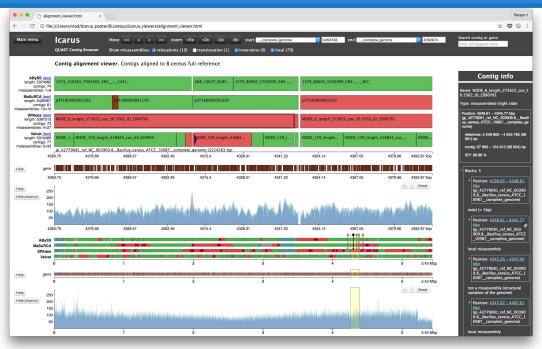
QUAST output

- Reports in different formats
  - Plain text table
  - Tab separated values (Excel, Google Spreadsheets)
  - Interactive HTML
- Plots (**PDF**/PNG/SVG)
  - Nx, NGx, NAx
  - Genes
  - Cumulative length
- Interactive contig viewers (Icarus)
  - Contig alignment viewer
  - Contig size viewer

Contig alignment viewer

- All alignments for each contig
- Misassembly details
- Contig ordering along the genome
- Overlaps / gaps

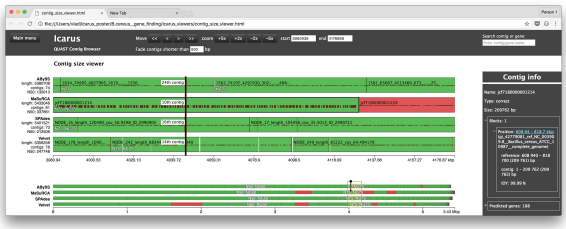
Contig alignment viewer



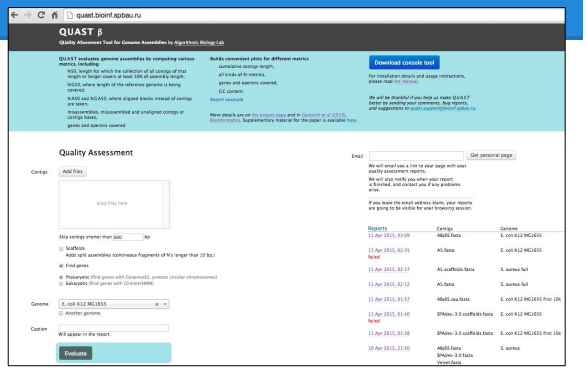
Contig size viewer

- Contigs ordered from longest to shortest
- N50, N75 (NG50, NG75)
- Filtration by contig size
- Gene prediction results
- Available without a reference

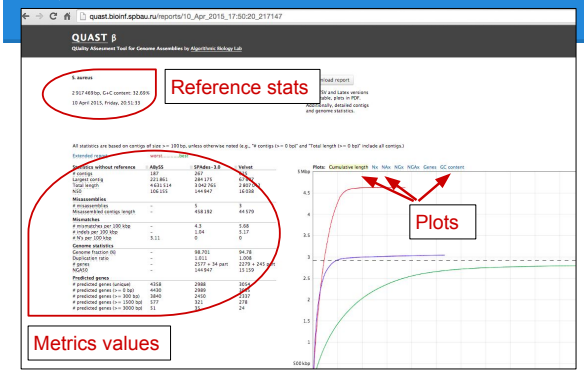
Contig size viewer



QUAST: website



QUAST: website



QUAST: web and console tools interconnection

- console tool generates HTML report <output_dir>/**report.html**
- website supplies full console tool output

De novo evaluation

Read-based statistics

- Number of aligned/unaligned reads
- % of assembly covered by reads

Read-based statistics

- Number of aligned/unaligned reads
- % of assembly covered by reads

- Points with low coverage
- Points with multiple read clipping
- Points with incorrect insert sizes

Read-based statistics

- Number of aligned/unaligned reads
- % of assembly covered by reads

- Points with low coverage
- Points with multiple read clipping
- Points with incorrect insert sizes

Will be added to QUAST v.**5**

Annotation-based statistics

- Number of ORFs

Annotation-based statistics

- Number of ORFs
- Number of gene/operon-like regions
  - **GeneMarkS** (Borodovsky *et al.*)
  - **GlimmerHMM** (Majoros *et al.*)

Annotation-based statistics

- Number of ORFs
- Number of gene/operon-like regions
  - **GeneMarkS** (Borodovsky *et al.*)
  - **GlimmerHMM** (Majoros *et al.*)
- Number of conservative genes
  - **BUSCO** (Simão *et al*.)
  - CEGMA (Korf *et al.,* no longer supported)

Thank you! Questions?

Useful links

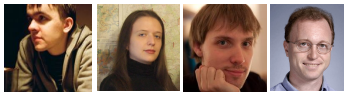
- cab.spbu.ru/software/quast
- quast.sf.net

Cite

- Mikheenko *et al.*, Icarus: visualizer for *de novo* assembly evaluation. *Bioinformatics*, 2016
- Mikheenko *et al.*, MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics*, 2016
- Gurevich *et al.*, QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 2013.

Acknowledgements

- QUAST team
  - Vlad Saveliev, SPbU
  - Alla Mikheenko, SPbU
  - Nikolay Vyahhi, Bioinformatics Institute / Stepik
  - Dr. Glenn Tesler, UCSD



De novo transcriptome assembly

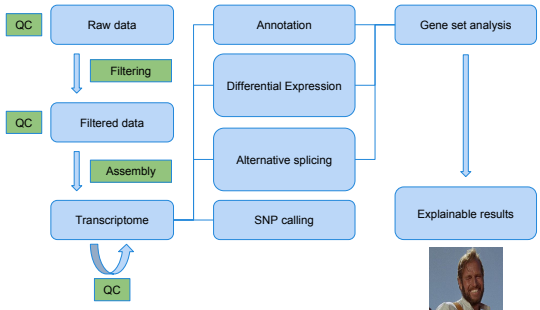
Andrey Prjibelski
Center for Algorithmic Biotechnology
SPbU

Overview

- RNA-Seq QC
- Read alignment and QC using reference genome
- *De novo* transcriptome assembly
- Transcriptome assembly evaluation

2

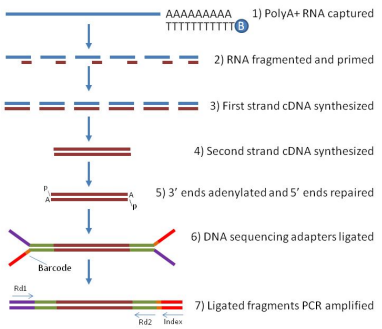
Typical workflow



3

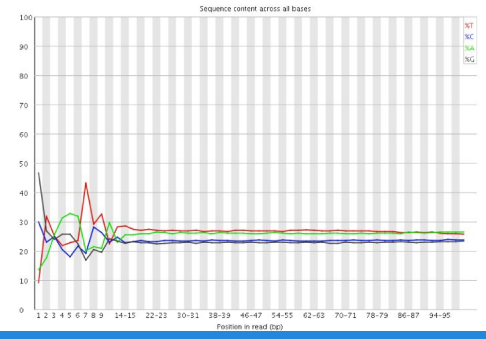
RNA-Seq QC

RNA-Seq



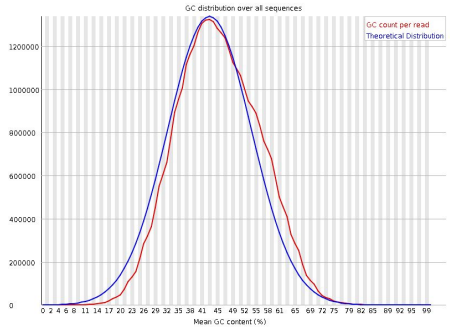
5

RNAseq as seen in FastQC (good data)



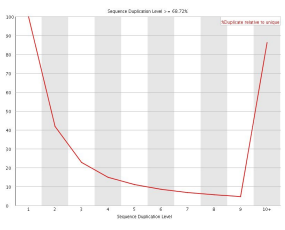
6

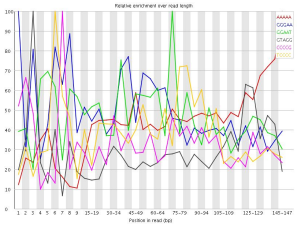
RNAseq as seen in FastQC (good data)



7

RNAseq as seen in FastQC (good data)





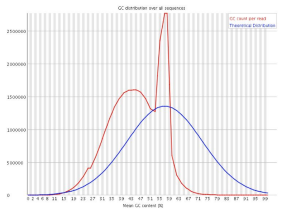
8

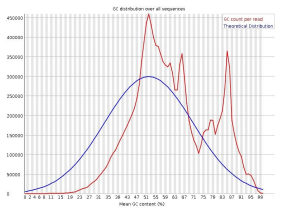
Possible problems with RNAseq data

- rRNA
- mtDNA
- Adapters
- Platform-specific problems
- Short fragment length in paired-end sequencing

9

Possible problems with RNAseq data

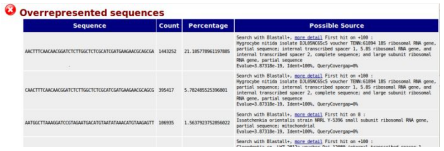




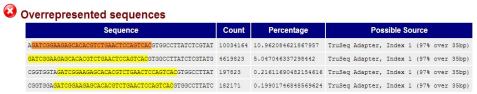
10

Possible problems with RNAseq data

rRNA



Adapters



11

Solutions

- Read filtering (bbduk)
- Read trimming (Trimmomatic, cutadapt)

12

Read filtering with bbtools

- Download rRNA database (e.g. from rnacentral.org)
- Run bbduk

13

Read filtering with bbtools

- Download rRNA database (e.g. from rnacentral.org)
- Run bbduk

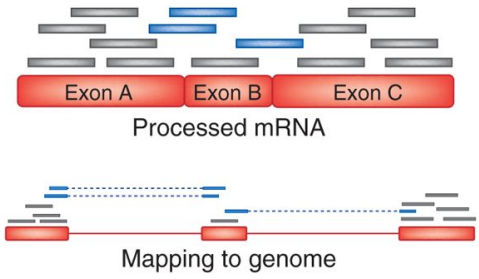
bbduk.sh in1=<left reads> in2=<right reads>
out1=<left unmatched> out2=<right unmatched>
outm1=<left matched> outm2=<right matched>
ref=<reference> k=31 hdist=1 stats=stats.txt

- Unmatched = good
- Matched = rRNA reads

14

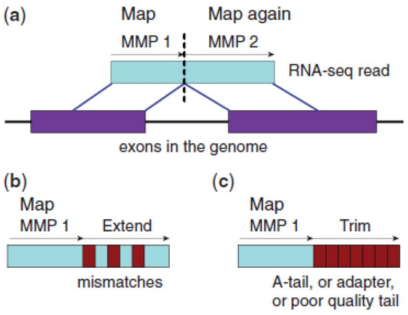
Reference based QC

Mapping RNA-Seq to the genome



16

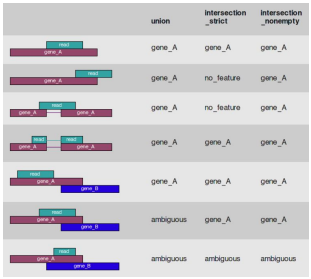
STAR aligner



17

Simple quantification (htseq-count)

- Simply count reads number
- Ignore multiple alignments
- Ignore ambiguous
- Do not differentiate between isoforms (gene counts only)



18

Alignment with STAR

- Create index

mkdir *index*

STAR --runMode genomeGenerate --runThreadN 2 --genomeFastaFiles *reference.fasta* --genomeDir *index/*

- Map reads

STAR --runThreadN 2 --genomeDir *index* --readFilesIn *1.fastq 2.fastq* --outSAMtype BAM Unsorted

19

QC with qualimap

qualimap rnaseq -bam *Aligned.out.bam* -gtf *genes.gtf* -oc gene_counts.txt -pe

Output:

- Text summary
- HTML report
- Gene counts (only with -oc option)

20

RNA-Seq assembly

RNAseq assembly features

- Widely varying coverage
- Alternative splicing leading to multiple isoforms
- Gene overlapping potentially resulting in chimeric transcripts
- Paralogous genes
- Sequencing biases leading to non-uniform coverage within transcripts

22

Reference-based assembly

- Reads are aligned to reference, splice junctions are spanned by reads or fragments
- Depends on alignment quality and annotation of genome

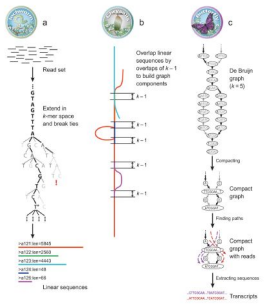
23

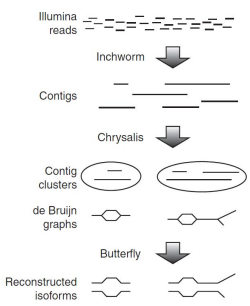
Reference-free assembly

- Reads are assembled without alignment to any reference
- Applicable if we don't have any reference at all, have a reference of inferior quality (fragmented, gapped, significantly altered)

24

De novo assembly workflow (Trinity example)



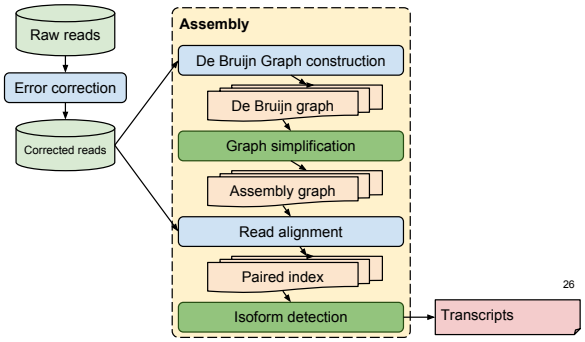


DOI:10.1038/nbt.1883.

25

doi:10.1038/nprot.2013.084

rnaSPAdes pipeline



26

Run rnaSPAdes

rnaspades.py --help

Strand-specific data

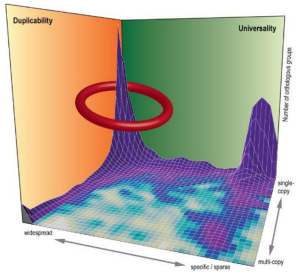
- **--ss-fr**
- **--ss-rf**

RNA-Seq assembly QC

27

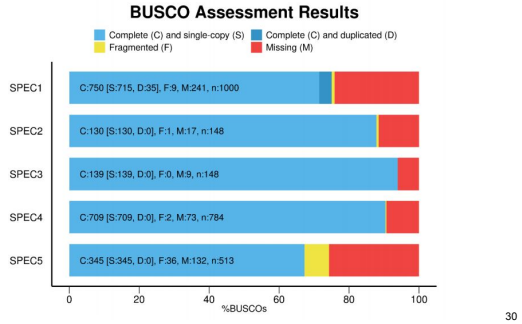
BUSCO

Benchmarking Universal Single-Copy Orthologs



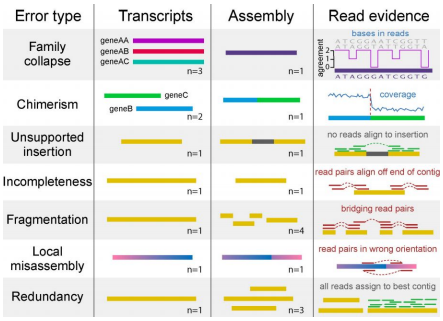
29

BUSCO



30

Types of assembly errors



DOI: 10.1101/gr.196469.115

31

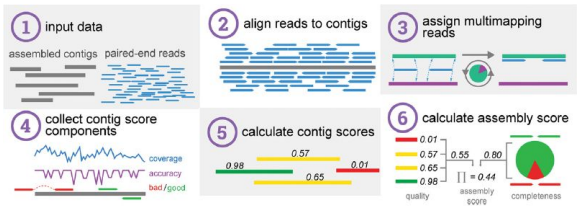
Transrate metrics

$s(C_{nuc})$	The proportion of nucleotides in the mapped reads that are the same as those in the assembled contig
$s(C_{cov})$	The proportion of nucleotides in the contig that have no supporting read data
$s(C_{ord})$	The extent to which the order of the bases in the contig are correct by analyzing the pairing information in the mapped reads
$s(C_{seq})$	The probability that the coverage depth of the transcript is univariate

DOI: 10.1101/gr.196469.115

32

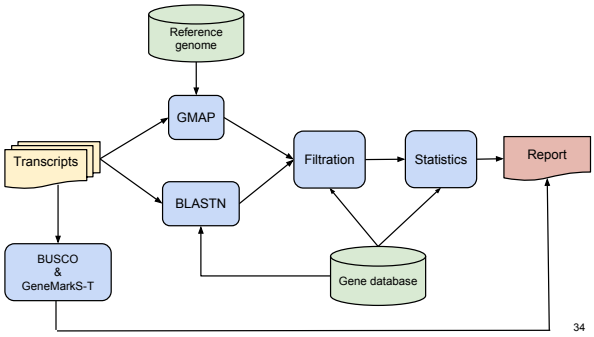
Transrate workflow



DOI: 10.1101/gr.196469.115

33

rnaQUAST pipeline

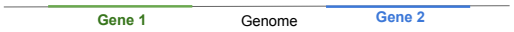


34

Which metrics are important?

35

Which metrics are important?



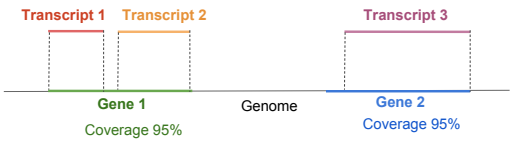
36

Which metrics are important?

37

Which metrics are important?

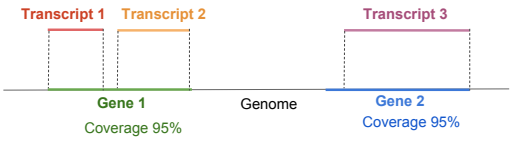
- Covered genes



38

Which metrics are important?

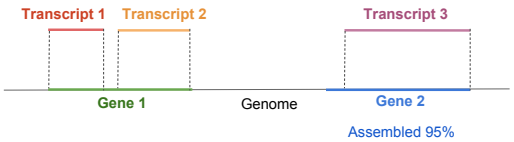
- Covered genes
  - 95%-covered genes: 2
  - 50%-covered genes: 2



39

Which metrics are important?

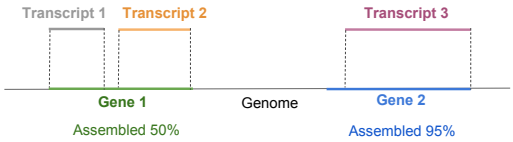
- Assembled genes



40

Which metrics are important?

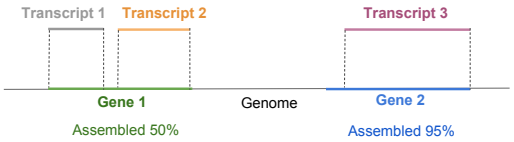
- Assembled genes



41

Which metrics are important?

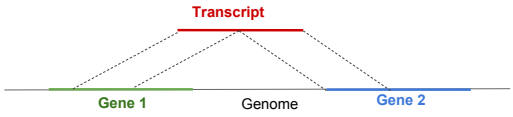
- Assembled genes
  - 95%-assembled genes: 1
  - 50%-assembled genes: 2



42

Which metrics are important?

- Missassemblies



43

Which metrics are important?

- Number of assembled/covered genes
- Number of assembled/covered isoforms
- Misassemblies
- Database coverage
- Number of transcripts, number of transcripts > 500 bp
- Mismatch rate
- etc

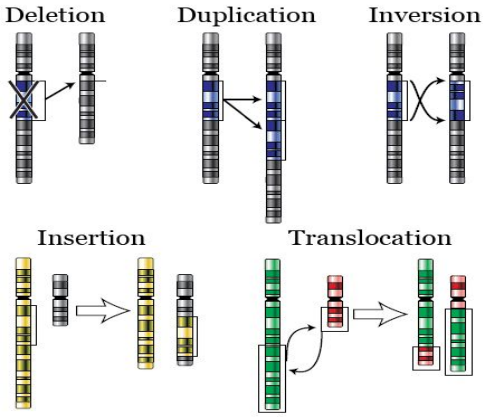
44

Detection of structural variations (Lecture & Workshop)

Tomasz Gambin
Warsaw University of Technology

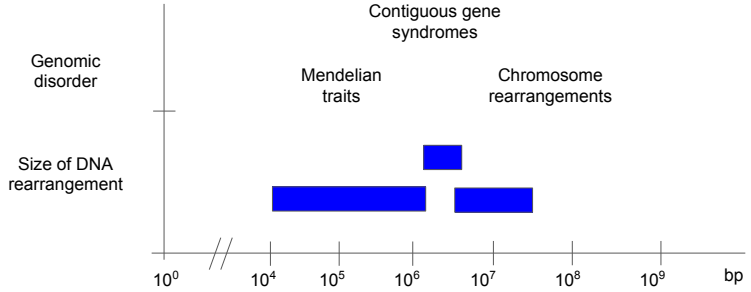
Wednesday, 9:00

Types of Structural Variants (SVs)

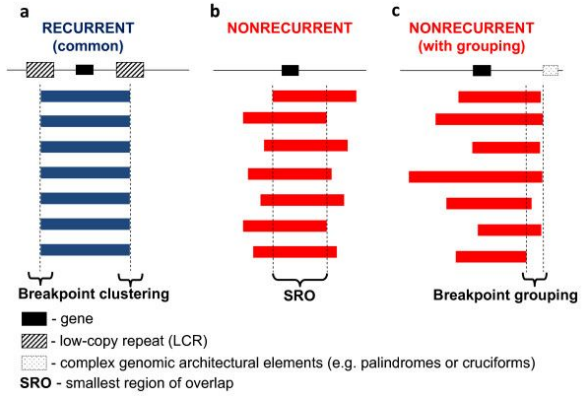


- Unbalanced (copy number variants, CNVs)
  - Deletions
  - Duplications
- Balanced
  - Inversions
  - Translocations
- Other
  - Absence of Heterozygosity (AOH)
  - Uniparental disomy (UPD)

Size of structural variants



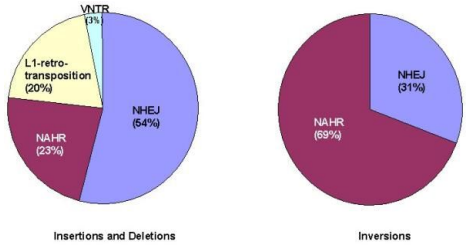
Recurrent vs. nonrecurrent



Gu et al., 2008

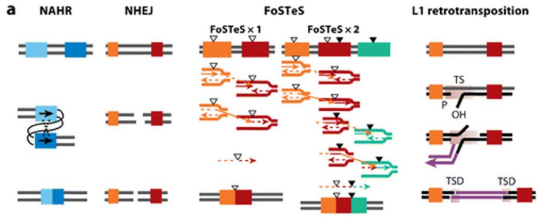
Formation mechanisms

- Recombination based
  - Nonallelic homologous recombination (**NAHR**)
  - Nonhomologous end joining (**NHEJ**)
- Replication based
  - Fork stalling and template switching (**FoSTeS or MMBIR**) - complex rearrangements
- L1-mediated retrotransposition



Cooper et al., 2011

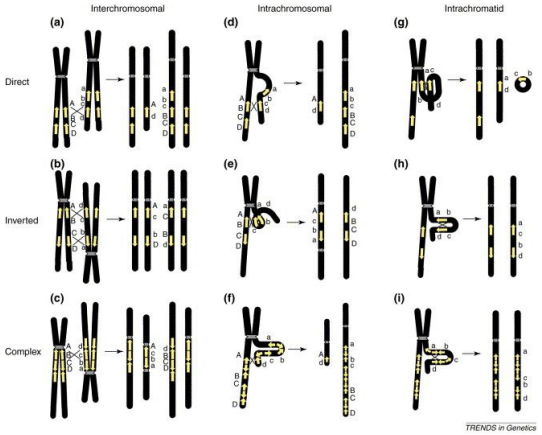
Formation mechanisms



b	NAHR	NHEJ	FoSTeS	Retrotransposition
Structural variation type	dup, del, inv	dup, del	dup, del, inv, complex	ins
Homology flanking breakpoint (before rearrangement)?	Yes (LCR/SD, *Alu*, L1, or pseudogene)	No	No	No
Breakpoint	Inside homology	Addition or deletion of basepairs, or microhomology	Microhomology	No specification
Sequence undergoing SV	Any	Any	Any	Transcribed sequences

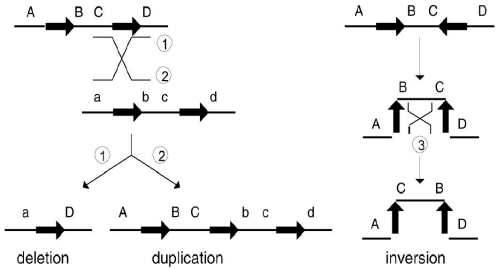
Zhang et al., 2009

Types of NAHR



Stankiewicz and Lupski, 2002

Non-allelic homologous recombination (NAHR)

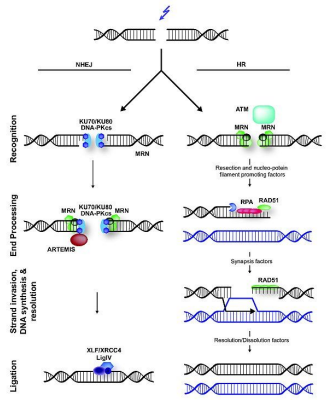


Gu et al., 2008

- NAHR leads to **recurrent rearrangements** and is usually mediated by low copy repeats (LCRs, also called segmental duplications, SDs)
- LCRs are defined as directly or inversely oriented highly similar repeats of genome (> 10kb in size and > 95% of DNA sequence identity)

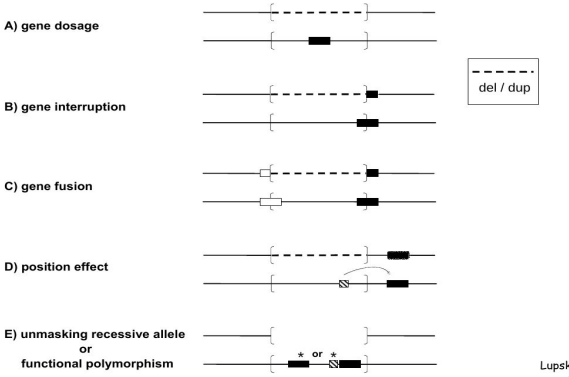
Non-homologous end joining (NHEJ)

- NHEJ proceeds in four steps
  - detection of Double Strand Break (DSB)
  - molecular bridging of both broken DNA ends
  - modification of the ends to make them compatible and ligatable
  - final ligation step
- No LCRs are obligatorily required for NHEJ
- NHEJ leaves an 'information scar' at the rejoining site



https://en.wikipedia.org/wiki/Non-homologous_end_joining

Phenotypic consequences of genomic re



Lupski and Stankiewicz (2005)
PLoS Genetics 1:e49

Importance of CNV detection

- CNVs account for more variation in the human genome than SNVs, i.e. 1.2% vs. 0.1% (Pang et al., 2010)
- CNVs can represent benign polymorphic variations or convey clinical phenotypes by mechanisms such as altered gene dosage and gene disruption.
- Locus-specific *de novo* mutation rates for CNV can be 100 to 10,000 times more frequent than for SNP

Copy number variation (CNV) versus single nucleotide polymorphism (SNP)

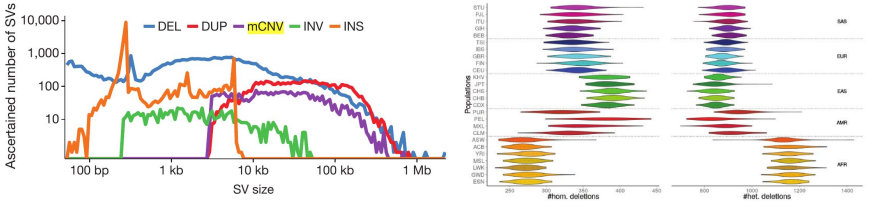
	CNV (Database of Genomic Variants, http://projects.tcag.ca/variation/)	SNP (dbSNP, http://www.ncbi.nlm.nih.gov/SNP/)
Total number	38,406[a] (Mar 11, 2009)	14,708,752 (Build 129)
Size	100 bp to 3 Mb	Mostly 1 bp
Type	Deletion, duplication, complex	Transition, transversion, short deletion, short insertion
Effects on genes	Gene dosage, interruption, etc.	Missense, nonsense, frameshift, splice site
Percentage of the reference genome covered	29.74%[b]	<1%

Zhang et al., 2009

SVs identified in 1000 genomes data (2,504 individuals)

Table 1 | Phase 3 extended SV release

SV class	No. sites	Median size of SV sites (bp)	Median kbp per individual	Median alleles per individual
Deletion (biallelic)	42,279	2,455	5,615	2,788
Duplication (biallelic)	6,025	35,890	518	17
mCNV	2,929	19,466	11,346	340
Inversion	786	1,697	78	37
MEI	16,631	297	691	1,218
NUMT	168	157	3	5.3

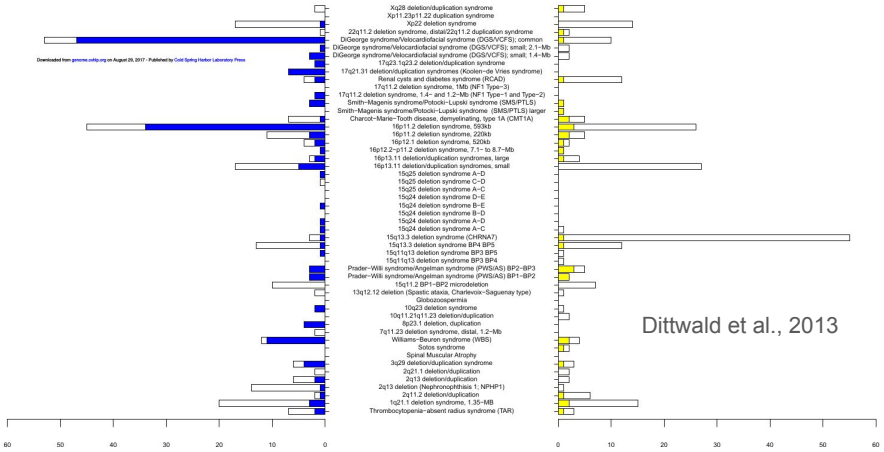


Sudmant et al., 2015

Clinically relevant CNVs

- ClinVar variants
  - 48,243 pathogenic or likely pathogenic SNVs
  - 4,299 pathogenic or likely pathogenic CNV deletions or duplications
- The overall detection rate for genomic rearrangements in children with DD/MR +/- multiple congenital anomalies is **~12–18%**
- CNV also implicated in many complex neurological and psychiatric phenotypes, including Autism Spectrum Disorder, schizophrenia

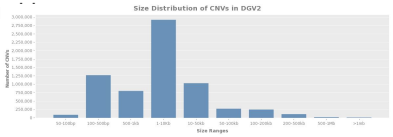
52 known disease-associated NAHR regions



Dittwald et al., 2013

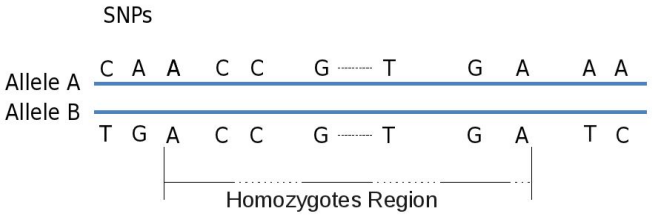
CNV databases

- Database of Genomic Variants (DGV)
  - Includes 1000 genomes data
  - Contains 0.5 M different CNVs and > 6M on the sample level
- DECIPHER
  - DatabasE of genomiC varIation and Phenotype in Humans using Ense Resources
  - ~30,000 CNVs + phenotype observations
- Copy Number Variation Morbidity Map of Developmental Delay in UCSC genome browser (Coe et al. 2014)



Absence of heterozygosity (AOH)

- Also called runs of homozygosity (ROH); in somatic cells we also observe loss of heterozygosity (LOH)
- Segments from both parents are the same in the offspring



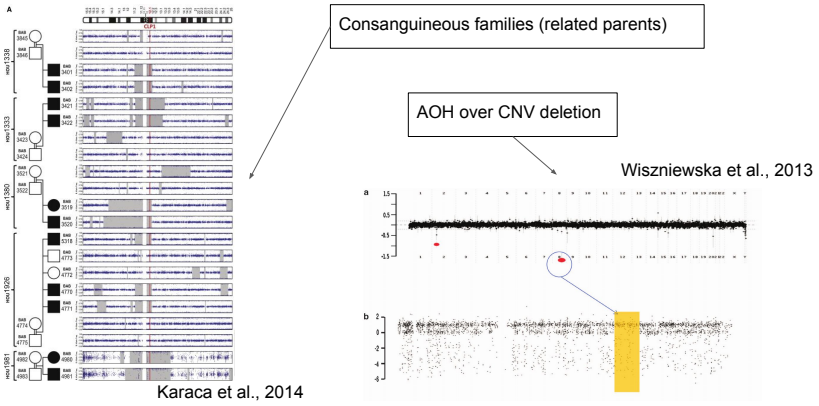
How exactly we define AOH ?

- Consecutive runs of homozygous SNPs
- There is no agreement on exactly what length of homozygotes can be called an AOH

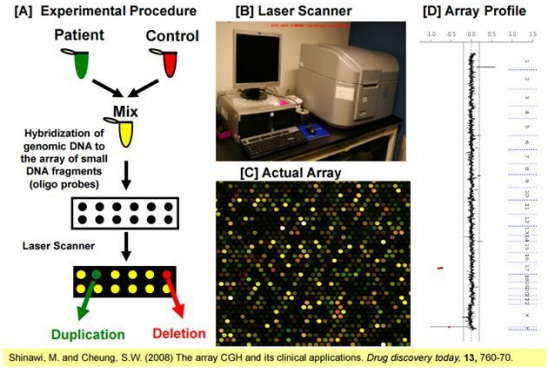
Parameter		Value
Homozygous SNP number	>=	50-100
Missing allowed	eq	2-5
Heterozygotes	<=	1-2%
Length	>	500 kb - 1 Mbp
Density (per SNP)	<=	50 kb

Chee seng ku et al., 2011

Origin of AOH regions



Array-based Comparative Genomic Hybridization (aCGH)



Detection of Structural Variants

- Locus specific methods
  - FISH
  - MLPA
  - PCR, dPCR
  - …
- Genome wide analysis
  - CGH and SNP arrays
  - NGS

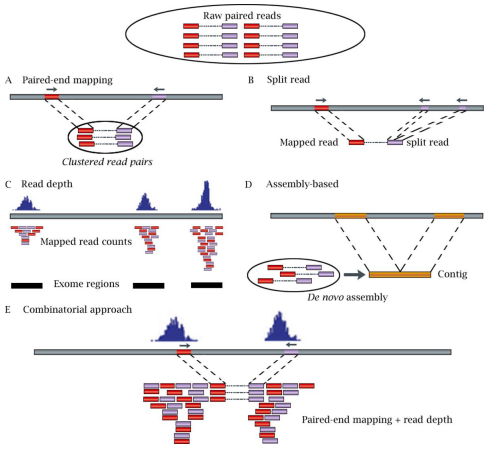
aCGH in comparison to NGS based methods

- Pros
  - Still cheaper than WGS/WES
  - Better in detection of duplications
  - Customized design allows to increase resolution in certain regions (e.g. exons)
  - Single exon deletions/duplications can be easily detected, which is hard using read depth approaches and WES data
- Cons
  - Do not allow to detect inversions and balanced translocation
  - Do not allow to detect small CNVs
  - Limited breakpoint resolution
  - If WGS/WES is performed anyway, aCGH is an additional cost

Detection of Structural Variants using NGS

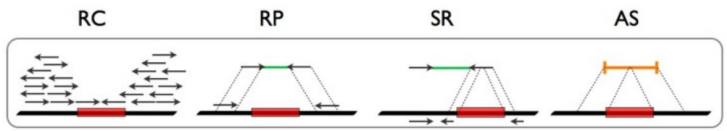
- In principle it should be possible to use new NGS technologies to identify all forms of SV by combining paired end read analyses with read-depth analyses
- In reality, this is still quite an analytical challenge, and not robust enough
- SV detection is not included in standard pipelines and best practices

Detection of SV using NGS - different approaches



Zhao et al., 2013

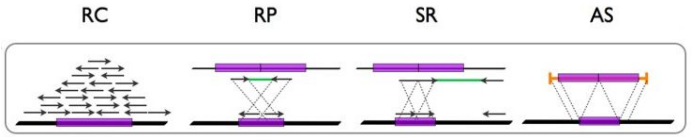
Detection of CNV deletions



Tattini et al., 2015

Method	Sign of variation
Read count (RC)	Decreased number of reads
Read pair (RP)	Increased inter-pair distance
Split read (SR)	Single read is merged from two segments surrounding deletion
Assembly (AS)	Assembled sequence show a gap

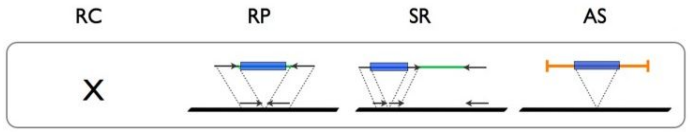
Detection of CNV duplications



Tattini et al., 2015

Method	Sign of variation
Read count (RC)	Increased number of reads
Read pair (RP)	Aberrant mapping (<---> instead of >---<) and inter-pair distance
Split read (SR)	Single read is split into end of one duplicated block followed by beginning of next block
Assembly (AS)	Assembled sequence with duplicated sequence

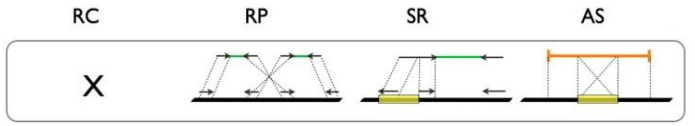
Detection of insertions



Tattini et al., 2015

Method	Sign of variation
Read count (RC)	Not applicable
Read pair (RP)	Decreased inter-pair distance
Split read (SR)	Single read is split into two segments surrounding novel insertion sequence
Assembly (AS)	Assembled sequence with inserted novel sequence

Detection of inversions



Tattini et al., 2015

Method	Sign of variation
Read count (RC)	Not applicable
Read pair (RP)	Aberrant mapping (>---> instead of >---<) and inter-pair distance
Split read (SR)	Single read is split into two segments one of which is inverted
Assembly (AS)	Assembled sequence with inverted sequence

Comparison of different approaches

Method	Pros	Cons
Read count (RC)	- **The only method applicable for WES/targeted NGS data** - Can detect exact copy number	- Limited breakpoint resolution - Usually requires a control data set - Cannot detect inversions, translocations
Read pair (RP)	- Sensitive and reliable	- Cannot detect exact copy number - There are problems in detection of large insertions or CNVs with breakpoints within LCRs
Split read (SR)	- Great method for PacBio/ Nanopore data - Very high breakpoint resolution	- Less reliable for short-read data - Cannot detect exact copy number
Assembly (AS)	- Can detect insertions of unmapped sequences	- Time consuming - Requires high coverage - Problems in non-unique sequence

SV detection using WES / targeted sequencing data

- Only read depth based approaches are applicable to targeted sequencing
  - Deletions, duplications can be detected
  - but not balanced translocations, inversions
- Depth of coverage is higher than in WGS but there are many more fluctuations due to inconsistent capture efficiency among targeted regions
- Exact CNV breakpoints location cannot be identified due to the lack of continuous coverage
- Most tools are capable to detect CNVs that encompass at least 3 target regions
- Tools are usually fine-tuned to detect rare CNVs; detection of common CNVs is harder
- Most tools require additional set of control samples
- B-allele frequency information can be used to identify potential AOH regions

Other challenges in SV analysis

- Annotation
  - Allele frequency information from external and local CNV databases
  - genes, exons
  - dosage sensitive AD genes
  - haploinsufficiency scores
  - known microdeletion and microduplication syndromes
- Filtration
  - by size, mappability (LCRs), allele frequency, gene content
- Visualization
  - Genome wide
  - Locus specific
- Validation by orthogonal experimental approach (aCGH, PCR, ddPCR, MLPA, FISH)
- Interpretation
  - **Simultaneous analysis of SNV and SV data**
  - Genotype-phenotype correlation

Introduction to Statistics (Lecture & Workshop)

German Demidov
Centre for Genomic Regulation, Barcelona

Wednesday, 14:00

Massive Parallel Sequencing-based RNA Structure Probing (Lecture & Workshop)

Łukasz Kiełpiński
Hoffmann-La Roche, Frederiksberg, DK

Wednesday, 14:00

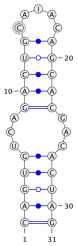
RNA molecules are central to conveying and regulating gene expression. They exist as three-dimensional entities with the structure largely determined by their base-pairing pattern (secondary structure). Function of many non-coding RNAs, such as ribosome, tRNAs or riboswitches, among many others, directly depends on their structure. Messenger RNA molecules, apart from coding for proteins, also carry a structural layer of information, which can influence interactions with other molecules leading to multitude of downstream effects, such as modulating splicing, polyadenylation, translation efficiency, RNA modifications or interactions with microRNAs and RNA binding proteins. Moreover, understanding mRNA structure is important for the design of RNA drugs as it can affect the siRNA and antisense oligonucleotides knockdown efficiencies. The structure of an RNA molecule can be to a certain extent predicted using many different computational approaches, whose accuracy can be largely increased using evolutionary or experimentally defined constraints. Strong evolutionary signal is considered a gold-standard for proving a certain functional structure, but the data is often not available. Traditional experimental methods are low throughput and labor intensive. The advent of the massive-parallel sequencing allowed to simultaneously probe RNA structures of large sets of molecules in a single experiment. Published methods for high-throughput RNA structure probing will be discussed. During the workshop we will use DMS-Seq as an example (Rouskin et al. 2014). We will start from processing raw sequencing reads, and we will calculate the normalized structure scores, all using publicly available tools such as bowtie2 (Langmead et al. 2012) and RNAprobR (Kielpinski et al. 2015). We will compare those scores to the annotated structure of the mitochondrial ribosome to evaluate the performance of the DMS-Seq method and visualize the data.

Why RNA Structure is important?

- Provides functional basis for many non-coding RNAs,
  - e.g. tRNA, ribosome, ribozymes, micro-RNA precursors, riboswitches, RNA thermometers, viral elements (e.g. IRES)
- Modulates function of mRNAs, e.g. it impacts:
  - interactions with RNA binding proteins or other RNA molecules (e.g. microRNAs) leading to multitude of downstream effects
  - splicing, polyadenylation and translation
- Modulates activity of antisense oligonucleotides and siRNAs
- …

RNA Structure

Secondary



Tertiary

Dot-bracket notation:

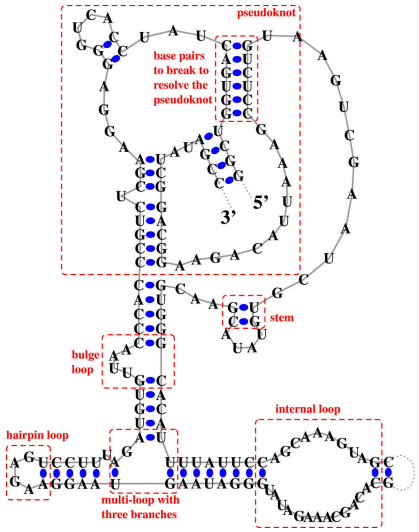
(((((. . . (((((. . . . .))))) . . .)))))

Rules governing RNA Structure

- RNA structure depends on:
  - Sequence,
  - Solvent properties,
  - Molecular interactors,
  - History of the molecule,
  - …
- RNA structure is hierarchical:
  - secondary structure contributes to the majority of the negative free energy, hence
  - it determines the tertiary structure

Rules governing RNA Structure

- Major energetic contributors are
  - Hydrogen bonds
    - Canonical Base pairing: A:U, G:C, G:U
  - Stacking
  - Loops



Traditional approaches for RNA structure determination

- Computational prediction
  - Energy based, e.g.: Mfold, RNAStructure, RNAfold (Vienna package)
  - Probabilistic, e.g.: CONTRAfold, Pfold
- Structure probing experiments:
  - By detection method:
    - RNA cleavage
    - reverse transcription (RT) termination
  - By the nature of the probe:
    - small chemicals (e.g. SHAPE, DMS, in-line, hydroxyl radical)
    - nucleases (e.g. S1, V1, P1)
  - By probing conditions:
    - *in vivo*
    - *in vitro*

Traditional approaches for RNA structure determination, cont.

- Biophysical methods:
  - NMR,
  - X-ray crystallography
  - SAXS
- Comparative structure models

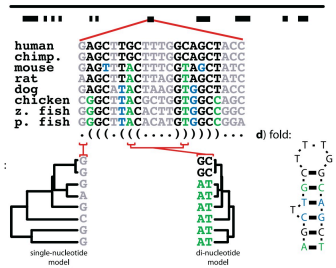


Figure reproduced under CC-BY license from Pedersen,J. et al (2006) *PLoS Comp Bio*, **2**, e33.

Examples of NGS-based RNA Structure Probing

- 1st generation
  - PARS, FragSeq
  - nuclease probing
  - ligation to the cleaved site (5' phosphate)
- 2nd generation
  - SHAPE-Seq, structure-seq*, HRF-Seq
  - RT termination at the modified (cleaved) site,
  - ligation to cDNA 3' end

*demonstrated in vivo

Examples of NGS-based RNA Structure Probing

- 3rd generation
  - DMS-Seq*, icSHAPE*, SHAPES
  - Detection of RT termination
  - Various strategies for signal enrichment
- 4th generation
  - SHAPE-MaP, DMS-MaPseq*
  - No cleavage, no intended RT termination
  - Probing event encoded in the read-body as a mutated base
  - No ligation bias
  - Potential for covariance analysis

*demonstrated in vivo

Examples of biological insights gained from the NGS-based RNA structure probing

- mRNA structure around the start codon reduces translational efficiency
- miRNA binding sites are preferentially unstructured
- mRNAs *in vivo* are less structured than *in vitro*
- +/- approaches allowed to detect RNA-protein interactions or RNA modifications *in vivo*
- SNPs that change the RNA structure (riboSNitches) are depleted from specific regions, such as microRNA or RBP binding sites

DMS-Seq

Get the

"*Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo*" by Rouskin et al. Nature 2014

Freely available at:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3966492/

Understand the method

- Read the first paragraph, analyze Figure 1
  - If you have time, check out the "Library Generation" paragraph in the 'Methods' section
- Discuss the method with your partner
  - What in your opinion is the biggest limitation of DMS as a RNA probing RNA reagent?
  - How is the structural signal deduced from the NGS output?
- Discuss with your partner a suitable design of the data analysis workflow

Doing now what patients need next

Functional genome annotation (Lecture & Workshop)

Marina Marcet-Houben
Centre for Genomic Regulation, Barcelona

Thursday, 9:00

One of the natural steps that needs to be done after assembling a new genome is to predict which genes are encoded in it and have an idea about their functionality. Discovering genes in prokaryotic genomes is a relatively simple matter due to their lack of introns and clear promoter-sequences. For eukaryotic genomes the process is more complex due in large part to the presence of long introns. The first part of the class will show different kinds of programs to use to predict gene in a genome and how to decide which program is the best one for our kind of data.
Once genes are predicted we are also interested in knowing their function. Experimental analysis are expensive and time-consuming so it is a good idea to have a general idea of the function of an unknown protein before we start working with it. Additionally, while blast searches are the universal tools used to assign function to a protein, there are times that this transference is incorrect or that it does not provide us with any information. We will explore and understand the limitations of blast, how to go around them and which other tools we can use when blast searches fail.

Well, we have a genome, and now what?

Gene prediction

Marina Marcet-Houben
NGSchool 2017
mmarcet@crg.es

What kind of information can we look for?

- **Genes**
- RNA
- Binding sites
- Conserved motifs
- Repetitive regions
- Transposable elements

...

What does this tell us?

We need to find the information encoded within this piece of DNA else we still will know nothing

Starting codon:
ATG - Metionin

Is there a better way to show all this information?

The GFF3 format

It is a standardized format that aims to describe elements encoded in a genome in a comprehensive way. It is mainly used to gene annotation but can be used for any element found in a genome.

```
##gff-version 3
# date: 2016-03-10 14:18:29
Oe6_s00002	CNAG	gene	46801	48266	.	+	.	ID=OE6A002689
Oe6_s00002	CNAG	transcript	46801	48266	.	+	.	ID=OE6A002689T1;Parent=OE6A002689;Name=OE6A002689T1;product=OE6A002689P1
Oe6_s00002	CNAG	exon	46801	47032	.	+	.	Parent=OE6A002689T1;ID=OE6A002689T1.exon1;Name=OE6A002689T1
Oe6_s00002	CNAG	exon	47161	48266	.	+	.	Parent=OE6A002689T1;ID=OE6A002689T1.exon2;Name=OE6A002689T1
Oe6_s00002	CNAG	CDS	46801	47032	.	+	0	Parent=OE6A002689T1;ID=OE6A002689C1;Name=OE6A002689C1
Oe6_s00002	CNAG	CDS	47161	48266	.	+	2	Parent=OE6A002689T1;ID=OE6A002689C1;Name=OE6A002689C1
Oe6_s00002	CNAG	gene	48456	49157	.	+	.	ID=OE6A044524
Oe6_s00002	CNAG	transcript	48456	49157	.	+	.	ID=OE6A044524T1;Parent=OE6A044524;Name=OE6A044524T1;product=OE6A044524P1
Oe6_s00002	CNAG	exon	48456	48649	.	+	.	Parent=OE6A044524T1;ID=OE6A044524T1.exon1;Name=OE6A044524T1
Oe6_s00002	CNAG	exon	48755	49157	.	+	.	Parent=OE6A044524T1;ID=OE6A044524T1.exon2;Name=OE6A044524T1
Oe6_s00002	CNAG	CDS	48456	48649	.	+	0	Parent=OE6A044524T1;ID=OE6A044524C1;Name=OE6A044524C1
Oe6_s00002	CNAG	CDS	48755	49157	.	+	1	Parent=OE6A044524T1;ID=OE6A044524C1;Name=OE6A044524C1
Oe6_s00002	CNAG	gene	49442	49924	.	+	.	ID=OE6A057168
Oe6_s00002	CNAG	transcript	49442	49924	.	+	.	ID=OE6A057168T1;Parent=OE6A057168;Name=OE6A057168T1;product=OE6A057168P1
Oe6_s00002	CNAG	exon	49442	49812	.	+	.	Parent=OE6A057168T1;ID=OE6A057168T1.exon1;Name=OE6A057168T1
Oe6_s00002	CNAG	exon	49897	49924	.	+	.	Parent=OE6A057168T1;ID=OE6A057168T1.exon2;Name=OE6A057168T1
Oe6_s00002	CNAG	CDS	49442	49812	.	+	0	Parent=OE6A057168T1;ID=OE6A057168C1;Name=OE6A057168C1
Oe6_s00002	CNAG	CDS	49897	49924	.	+	1	Parent=OE6A057168T1;ID=OE6A057168C1;Name=OE6A057168C1
Oe6_s00002	CNAG	gene	72128	73332	.	-	.	ID=OE6A057440
Oe6_s00002	CNAG	transcript	72128	73332	.	-	.	ID=OE6A057440T1;Parent=OE6A057440;Name=OE6A057440T1;product=OE6A057440P1
Oe6_s00002	CNAG	exon	72128	72394	.	-	.	Parent=OE6A057440T1;ID=OE6A057440T1.exon3;Name=OE6A057440T1
Oe6_s00002	CNAG	exon	72897	73031	.	-	.	Parent=OE6A057440T1;ID=OE6A057440T1.exon2;Name=OE6A057440T1
Oe6_s00002	CNAG	exon	73141	73332	.	-	.	Parent=OE6A057440T1;ID=OE6A057440T1.exon1;Name=OE6A057440T1
Oe6_s00002	CNAG	CDS	72128	72394	.	-	0	Parent=OE6A057440T1;ID=OE6A057440C1;Name=OE6A057440C1
Oe6_s00002	CNAG	CDS	72897	73031	.	-	0	Parent=OE6A057440T1;ID=OE6A057440C1;Name=OE6A057440C1
Oe6_s00002	CNAG	CDS	73141	73332	.	-	0	Parent=OE6A057440T1;ID=OE6A057440C1;Name=OE6A057440C1
```

Contig | Annotation technology | Feature | Start position | End position | Score | Frame | Phase | Attributes

How to predict genes in a newly sequenced genome

MAKER

Glimmer

JIGSAW: gene prediction using multiple sources of evidence

geneid 1.2

GeneWise

Augustus

Exonerate

SoftBerry

fgenesh

EuGene: Integrative Gene Finding

BLAT search

GeneMark

GenomeScan
webserver at MIT

The GENSCAN Web Server at MIT

GlimmerHMM
Eukaryotic Gene-Finding System

EVM

GeneZilla

Before doing the gene prediction: to mask or not to mask?

Masking your genome consists in turning repetitive regions or low complexity regions into Ns.

There is not a correct answer, it will depend on your data.

INSTITUTE FOR Systems Biology RepeatMasker

RepeatMasker is the most used tool to detect repeats and mask genomes.

It uses pre-defined repeats to detect regions in your genome that should be masked

An *ab initio* tool can be used to find repeats and then the genome can be masked using tools such as maskfasta from the bedtools package

Nucleic Acids Res. 2008 Apr; 36(7): 2284–2294.
Published online 2008 Feb 20. doi: 10.1093/nar/gkn064

Empirical comparison of *ab initio* repeat finding programs

Surya Saha,[1,2,3] Susan Bridges,[1,3] Zenaida V. Magbanua,[2,3,4] and Daniel G. Peterson[2,3,4,*]

Author information ▶ Article notes ▶ Copyright and License information ▶

Homology based programs.

They use as input a set of genes that have been previously predicted, usually in a different, closely related, species.

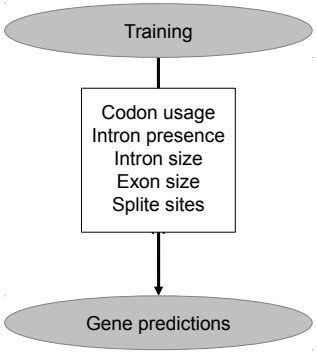
Group of predicted sequences (fasta format) ⟷ Genome

Blast
Blat
Exonerate
geneWise

Drawbacks:
- Will not find proteins that have not been predicted before (orphan proteins)
- Often they will not provide the complete protein, it may miss the beginning or the end.
- In Prokaryotes you will need to consider the real possibility of HGT.

Prediction of genes *ab initio*

It will use descriptive parameters to find genes in a genome sequence.



Many programs to choose from

Augustus

Glimmer

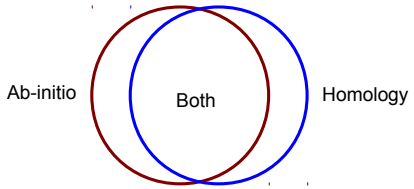
geneid HomePage

DOGMA
Dual Organellar GenoMe Annotator

It may fail for genes with very unusual characteristics (i.e. mitochondrial genes)

Some of them already have parameters from some of the species, it is important to know which species you are working with

Combined methods



\+

Other information:

- EST data
- RNAseq data

MAKER

EVM

JIGSAW: gene prediction using multiple sources of evidence

They render the **most reliable annotations**, though they may miss some of them. In the end, protein annotation can be a very difficult process which needs a lot of intervention, specially for complex organisms (i.e. plants, animals,...). More simple species can be annotated with simpler methods.

How to choose which is the best program for my genome

1.- Am I working with a Eukaryote or with a Prokaryote?

2.- Which kind of program do I want to use?
- *ab initio* prediction
- Homology based prediction
- Program that combines multiple sources of information

3.- Do I have enough data to run the kind of program I chose above?

4.- Do I have any additional information that I want to use to make the predictions?

5.- Can I run it locally or will I need to use a website?

6.- How reliable do my gene predictions need to be?

Once you have answered these questions then you can go through a list of programs and choose those that are better adapted to your needs

How do I know whether my genome is well annotated?

You don't, but you can make estimates

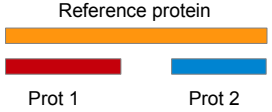
1.- How well assembled is your genome?

2.- Do you have the core proteins?

BUSCO
QUEST FOR QUALITY
"BUSCO CALIDAD"
"BUSCO QUALIDADE"

CEGMA
(discontinued)

3.- Do you have many split genes?



Well, we have a proteome, and now what?

Functional prediction

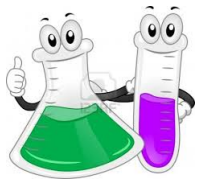
How can we know what our protein does?

1.- Experimental evidence

2.- Homology

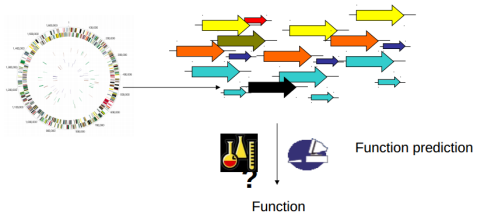
3.- Orthology

4.- Others



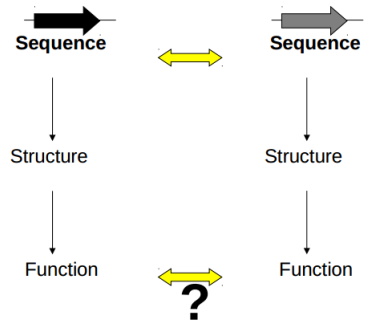
Doing experiments for each element we have identified to discover their function is impossible right now.

Experimental evidence



- E. coli, the most intensively studied organism: only 1924 genes (~43%) have been (partially) experimentaly characterized.

Classic method: function prediction by homology

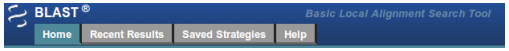


Homology: they have a common evolutionary origin. Two proteins are either homologous or not, there are no degrees.

Two proteins can be more similar than two others, but never more homologous.

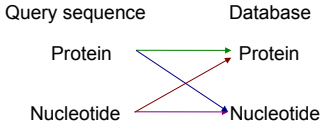
Comparison of whole proteins (Similarity search)

It is used to transfer the annotation of a “known” protein to an unknown one.



http://blast.ncbi.nlm.nih.gov/Blast.cgi

Different kind of blast searches:

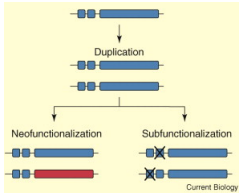


http://www.uniprot.org/

Yet, using only similarity searches can produce miss-annotations when the species are very distantly related or the evolution of the protein family is very complex.

Why can blast fail when you have complex evolutionary histories?

After duplication, genes can change their original function.

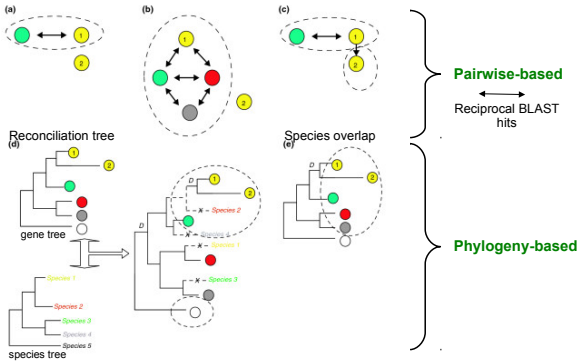


Subfunctionalization: Each paralog performs part of the original function.

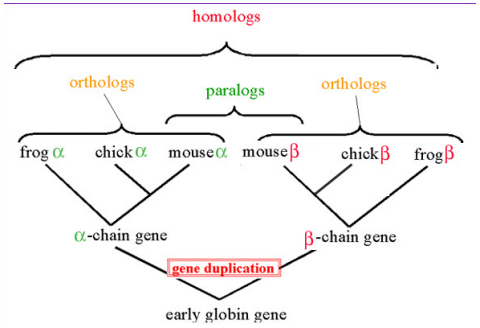
Neofunctionalization: One of the paralog obtains a new function.

We should not transfer function between two paralogs, as they are more likely to not share the function. Blast cannot properly identify between paralogs in complex evolutionary histories.

Best reciprocal hits, mcl, orthoMCL, eggNOG, Inparanoid



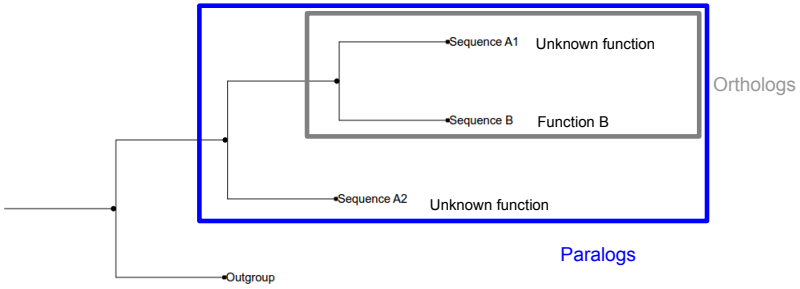
Identification of orthologs and paralogs.



Orthologs: Proteins that are derived from a speciation point.

Paralogs: Proteins that are derived from a duplication point.

More reliable: function prediction by orthology



Blast results:

Sequence A1 → Function B

Sequence A2 → Function B

Orthology prediction:

Sequence A1 → Function B

Sequence A2 → Unknown

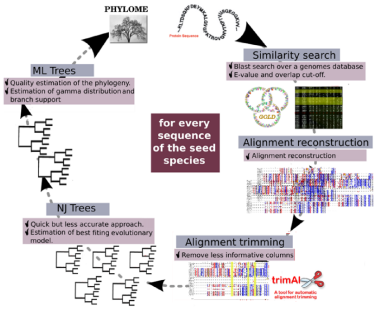
Paralogs are less likely to keep the same function than orthologs.

How can we obtain orthologs?

1.- Reconstruct your own phylogenetic tree

2.- Obtain orthologs from tree-based databases:

- EnsemblCompara: http://www.ensembl.org/info/genome/compara/index.html
- TreeFam: http://www.treefam.org/
- PhylomeDB & MetaPhors:
  http://http://phylomedb.org/
  http://betaorthology.phylomedb.org/



How do we find ANYTHING with so many trees?

BMC Bioinformatics. 2010; 11: 24.
Published online 2010 January 13. doi: 10.1186/1471-2105-11-24

ETE: a python Environment for Tree Exploration

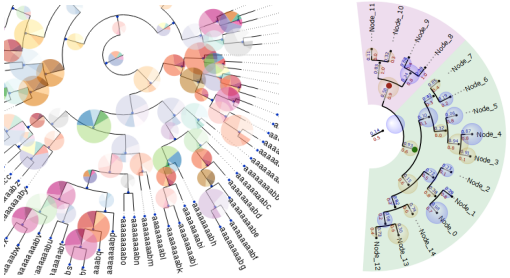
Reviewed by Jaime Huerta-Cepas,[1] Joaquín Dopazo,[2] and Toni Gabaldón[1]

E.T.E.
A python Environment for Tree Exploration

ETE is a python module to work with phylogenetic trees. It allows the user to work with thousands and thousands of trees with little effort.

http://ete.cgenomics.org/

PhylomeDB uses the ETE visualization modules to show the trees.



MetaPhors:

A meta-method to predict orthology and paralogy from multiple phylogenetic evidence:

Retrieve trees

↓

Remove low-quality trees

↓

Map events

↓

Calculate consistency scores

eggNOG version 3.0

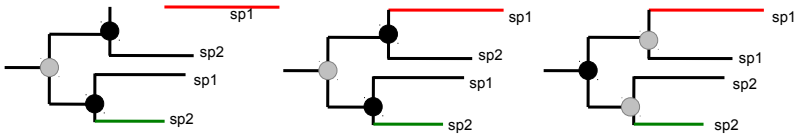
COGs

OrthoMCL DB
Ortholog Groups of Protein Sequences

Treefam

HOGENOM

Orthology and paralogy prediction



PhylomeDB Ensembl EggNOG

duplication
speciation

CSo= To / (To + Tp) = 1 / 1+2 = 0.33

CSp = Tp / (To + Tp)= 2 / 1+2 = 0.67

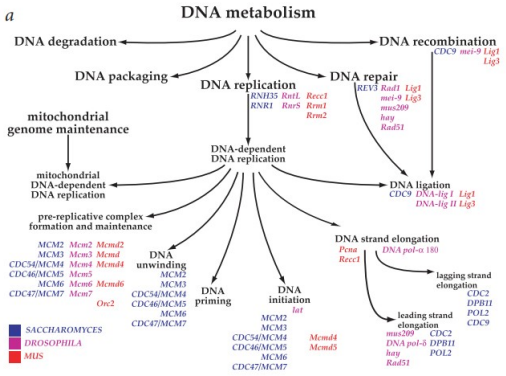
Evidence level: 1

By default CSo threshold for orthology prediction is 0.5

Gene Ontology (GO)

http://geneontology.org

The GO defines concepts/classes used to describe gene function, and relationships between these concepts. It provides a controlled vocabulary of terms for describing gene product characteristics and gene product annotation.



Terms are divided in three different categories:

- Molecular function
- Cellular component
- Biological process

Annotation based on conserved protein domains

Proteins have regions that are highly conserved across other proteins and other species that have orthologs to those proteins. These conserved domains often carry the functionality of the protein and as such, identifying those domains can give us an idea of the function of the protein.

Pfam http://pfam.xfam.org/

InterPro
Protein sequence analysis & classification
http://www.ebi.ac.uk/interpro/

HMMER can be used to search domains in your proteins

HMMER http://hmmer.org/

How can I assign GO terms to my proteins?

1.- Transference of annotation between orthologs.

2.- Obtain annotations from webpages. For instance Uniprot provides GO annotations for all its proteins.

3.- Use external tools to annotate your proteins.

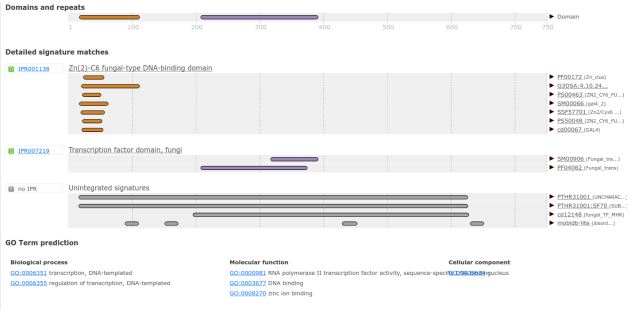
List of GO annotation tools:
ftp://ftp.geneontology.org/go/www/GO.tools.annotation.shtml

InterProScan

http://www.ebi.ac.uk/interpro/search/sequence-search

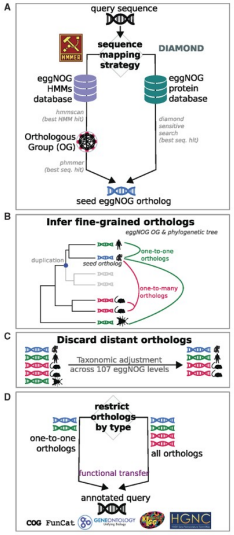
Interproscan is a tool that provides annotations of a protein based on domain conservation using multiple datasets for the annotation. In one single run it can provide, when present, Pfam annotations, Interpro annotations, GO terms, PROSITE, metabolic annotations such as KEGG annotation.

It can be run locally or you can use the website tool



Eggnog-Mapper

http://eggnogdb.embl.de/#/app/emapper



Practical differences between eggnog-mapper and interproscan:

1.- Interproscan provides a wider variety of annotations (GO terms, Pfam domains, PANTHER, HAMAP, TIGRFAM, KEGG,...)

2.- Interproscan is faster for few sequences. For many sequences eggnog-mapper is much faster

3.- Interproscan annotates more proteins, but eggnog-mapper assigns more GO terms. The sets of annotated proteins do not overlap completely

4.- Interproscan is a JAVA program and eggnog-mapper is based on python

5.- Both have a command-line and web tool option

What do I do then when I have no idea about the function of my protein?

Search for conserved protein domains

Search for metabolic information

Eggnog-mapper

Search for interactions with other proteins

Search for cellular localization

You will still not know which protein you have, but you will have an idea.

Search for conserved protein domains

Protein of unknown function (DUF1093)

Search for metabolic information

Search for interactions with other proteins

Search for cellular localization

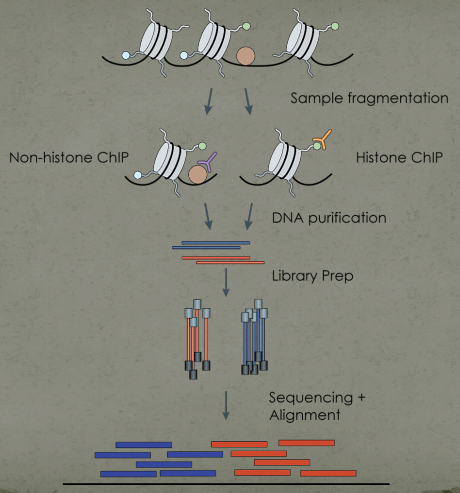
Sadly, there's also the possibility that you will still know little after doing the analysis

ChIP-seq
(Lecture & Workshop)

Aliaksei Holik & Maciej Łapiński
Centre for Genomic Regulation, Barcelona& International Institute of Molecular and Cell Biology in Warsaw

Thursday, 14:00

ChIP-seq workflow



Typical pre-processing workflow

Read alignment
BWA, Bowtie, subread, STAR

↓

Read sorting
Picard, samtools

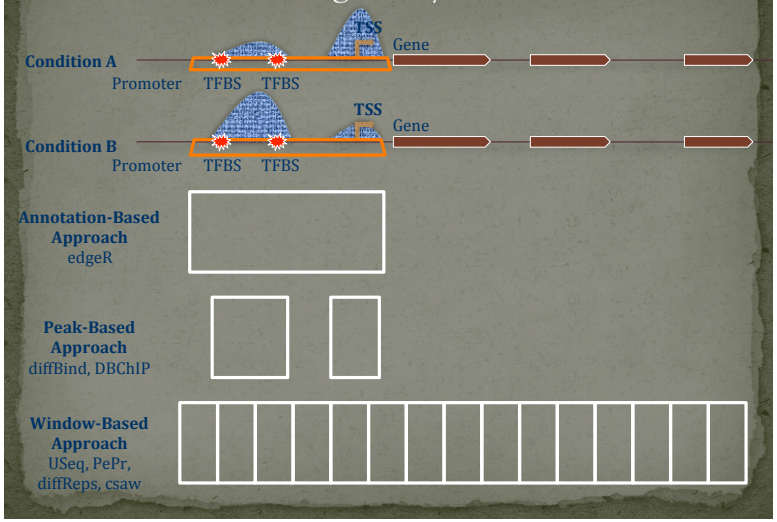
↓

Duplicate marking
Picard, samtools

↓

bam file indexing
samtools

Differential Binding Analysis – main flavours



Differential Binding Analysis workflow

Count reads into defined regions

⌄

Filter

⌄

Normalise

⌄

Compute the statistics
(negative binomial framework with edgeR)

⌄

Annotate and visualise

Single-cell RNA-seq & Differential expression (Lecture & Workshop)

Davis McCarthy
EBI, Hinxton

Friday, 9:00 & 14:00

Today it is possible to obtain genome-wide transcriptome data from single cells using high-throughput sequencing (scRNA-seq). The main advantage of scRNA-seq is that the cellular resolution and the genome wide scope makes it possible to address issues that are intractable using other methods, e.g. bulk RNA-seq or single-cell RT-qPCR. However, to analyze scRNA-seq data, novel methods are required and some of the underlying assumptions for the methods developed for bulk RNA-seq experiments are no longer valid.
In a short space of time, many methods have been developed to address important questions that can be address with scRNA-seq and key aspects of scRNA-seq analysis: pre-processing and quality control, visualisation, clustering, differentiation trajectories and differential expression. This workshop provides an introduction to these key topics and demonstrates the use of a set of open-source R packages to solve frequently-encountered problems in scRNA-seq analysis.
The workshop will use material developed with Martin Hemberg, Tallulah Andrews and Vlad Kiselev, which is available here: `https://hemberg-lab.github.io/scRNA.seq.course/index.html`. The course is taught through the University of Cambridge Bioinformatics training unit (`http://training.csx.cam.ac.uk/bioinformatics/`), but the material found on these pages is meant to be used for anyone interested in learning about computational analysis of scRNA-seq data. The course is taught twice per year and the material here is updated prior to each event.

Single-cell Hi-C data analysis (Lecture & Workshop)

Aleksandra Galitsyna
MSU, Moscow

Saturday, 9:00

ARTICLE 2013

doi:10.1038/nature12593

Single-cell Hi-C reveals cell-to-cell variability in chromosome structure

Takashi Nagano[1]*, Yaniv Lubling[2]*, Tim J. Stevens[3]*, Stefan Schoenfelder[1], Eitan Yaffe[2], Wendy Dean[4], Ernest D. Laue[3], Amos Tanay[2] & Peter Fraser[1]

[1]Nuclear Dynamics Programme, The Babraham Institute, Cambridge CB22 3AT, UK. [2]Department of Computer Science and Applied Mathematics and Department of Biological Regulation, Weizmann Institute, Rehovot 76100, Israel. [3]Department of Biochemistry, University of Cambridge, Cambridge CB2 1GA, UK. [4]Epigenetics Programme, The Babraham Institute, Cambridge CB22 3AT, UK.

LETTER 2017

doi:10.1038/nature21711

Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition

Ilya M. Flyamer[1,2,3]*†, Johanna Gassler[1]*, Maxim Imakaev[4,5]*, Hugo B. Brandão[6], Sergey V. Ulianov[2,3], Nezar Abdennur[7], Sergey V. Razin[2,3], Leonid A. Mirny[4,5,6]§ & Kikuë Tachibana-Konwalski[1]§

[1]IMBA - Institute of Molecular Biotechnology of the Austrian Academy of Sciences, Vienna Biocenter (VBC), Dr Bohr-Gasse 3, 1030 Vienna, Austria. [2]Institute of Gene Biology, Russian Academy of Sciences, Moscow 119334, Russia. [3]Faculty of Biology, Lomonosov Moscow State University, Moscow 119234, Russia. [4]Institute for Medical Engineering and Science, Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts 02139, USA. [5]Department of Physics, Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts 02139, USA. [6]Harvard Program in Biophysics, Harvard University, Cambridge, Massachusetts 02138, USA. [7]Computational and Systems Biology Program, Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts 02139 USA. †Present address: MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK.

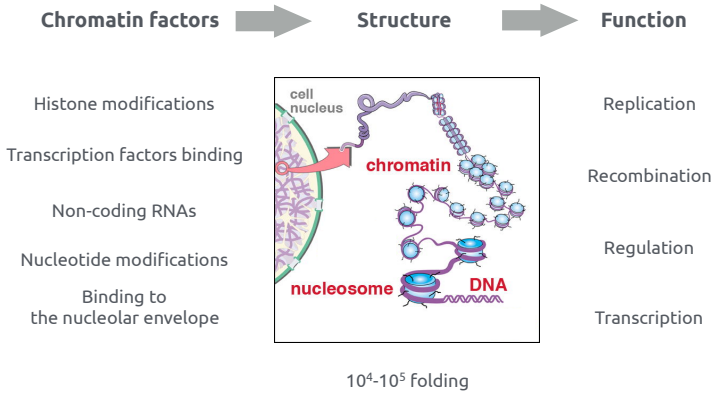
Single-cell Hi-C research in a large collaboration

- Kikuë Tachibana-Konwalski's lab from Institute of Molecular Biotechnology of the Austrian Academy of Sciences, VBC, Vienna, Austria
- Prof. Leonid Mirny's lab from MIT, Cambridge, Massachusetts, USA
- Prof. Sergey Razin's lab from Institute of Gene Biology of Russian Academy of Sciences (IGB RAS), Moscow, Russia
- Prof. Mikhail Gelfand's lab from Institute for Information Transmission Problems of Russian Academy of Sciences (IITP RAS), Moscow, and Skolkovo Institute of Science and Technology (Skoltech), Skolkovo, Russia

Outline

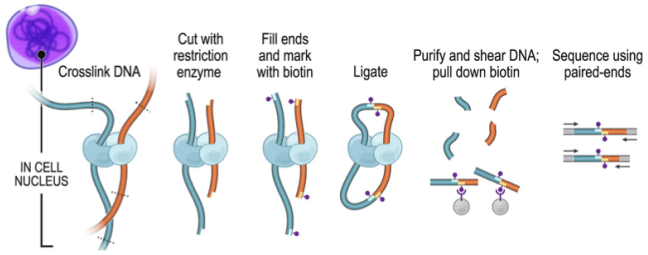
- Introduction
  - Eukaryotic chromatin structure
  - Hi-C and chromatin interaction map
  - Interaction map features: TADs, compartments, loops
  - Single-cell Hi-C
- From theory to practice: Hi-C data processing workflow
  - Reads mapping
  - Binning & filtering
  - Matrix balancing
  - TADs and compartments calling
  - Single-cell data analysis
- Workshop overview

1.1 Introduction: Eukaryotic chromatin structure

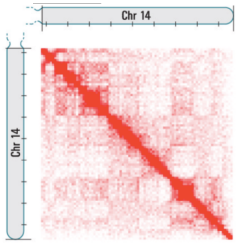


1.2 Hi-C: high-throughput chromosomes conformation capture

Procedure:



Resulting interactions heatmap:

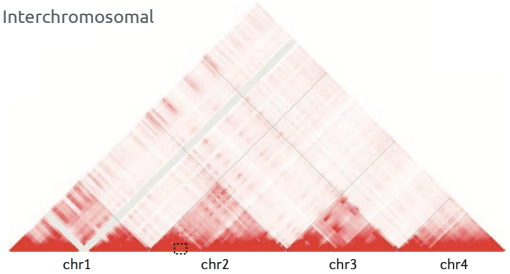


Lieberman-Aiden et al. *Science* 2009

1.3 Interaction map features: Chromosome territories

- At the highest level of spatial organization, trans-interactions are rare.
- Individual chromosomes occupy distinct territories within the nucleus.

Interchromosomal



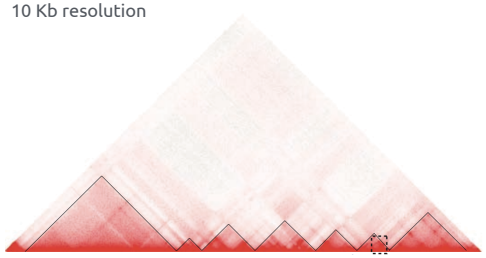
Map is rotated by 90°, upper triangle visualized

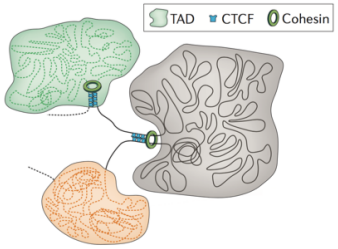
Bonev et al. *Nature Reviews* 2016

1.3 Topologically-associating domains (TADs)

- Chromosomes are further spatially segregated into sub-megabase scale domains, or TADs.

10 Kb resolution



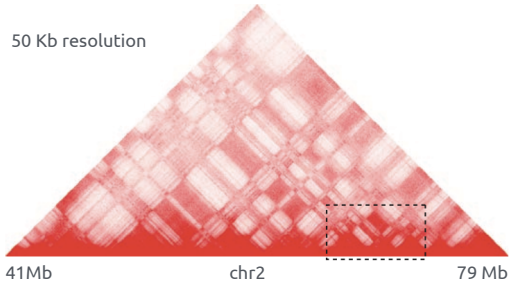


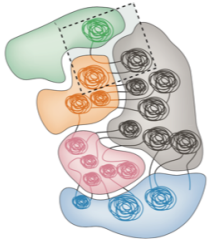
Bonev et al. *Nature Reviews* 2016

1.3 Chromatin compartments

- TADs have preferential long-range contacts with each other, forming two types of compartments, A and B (domains in compartment A interact mostly with other type A domains, and vice versa).
- Two major compartments can be further subdivided into six different subcompartments.

50 Kb resolution

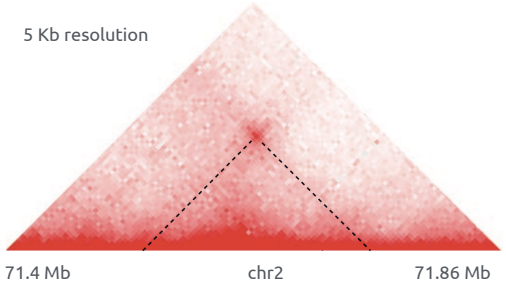


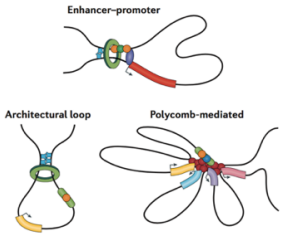


Bonev et al. *Nature Reviews* 2016

1.3 Chromatin loops

- Cis-regulatory elements of vertebrates, such as enhancers, are separated by relatively long distances and can be brought into close spatial proximity with its target through the formation of chromatin loops.
- There are also other cases of loops (e.g. between co-regulated genes, between Polycomb-repressed genes).

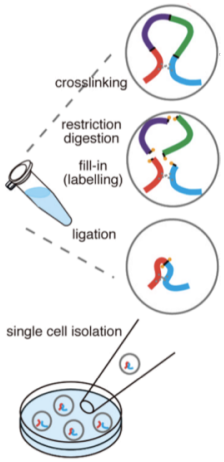




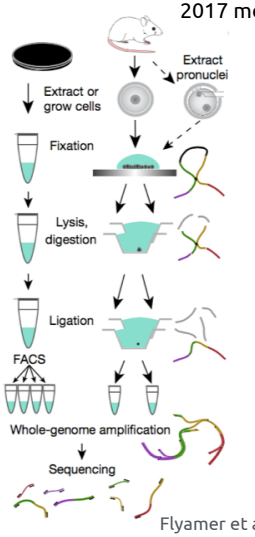
Bonev et al. *Nature Reviews* 2016

1.4 Single-cell Hi-C

2013 method:



2017 method:



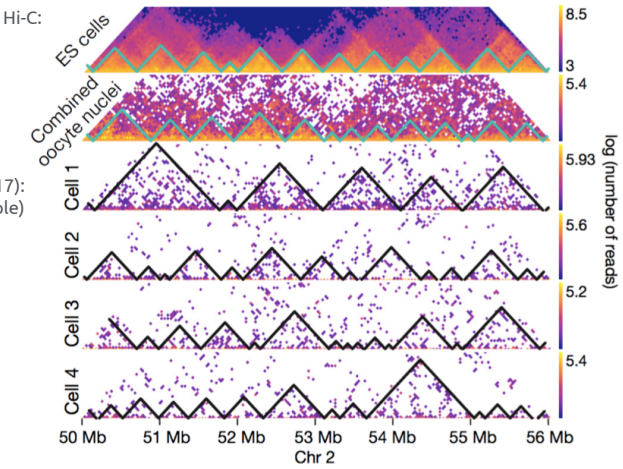
Nagano et al. *Nature* 2013

Flyamer et al. *Nature* 2017

1.4 Single-cell Hi-C

Ensemble (bulk) Hi-C:

Single-cell Hi-C (2017): (multiple)



Flyamer et al. *Nature* 2017

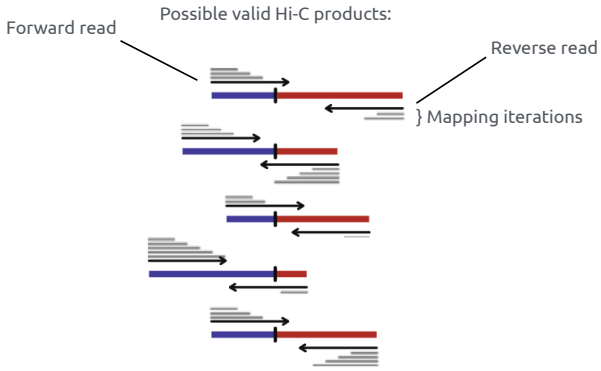
2. From theory to practice: Hi-C processing workflow

2. Hi-C processing workflow

1. Reads mapping: paired-end mode is not used, iterative mapping.
2. Filtering & binning
   - Fragment assignment: the mapped read is assigned according to its 5' mapped position, mapped read positions should fall close to a restriction site
   - Fragment filtering: multiple mapping, PCR duplicates, undigested restriction sites
   - Binning
   - Bin level filtering: remove 1% low signal rows/columns
3. Balancing: correction for technical biases
4. Features calling (TADs, compartments, loops, etc.)

2.1 Reads mapping

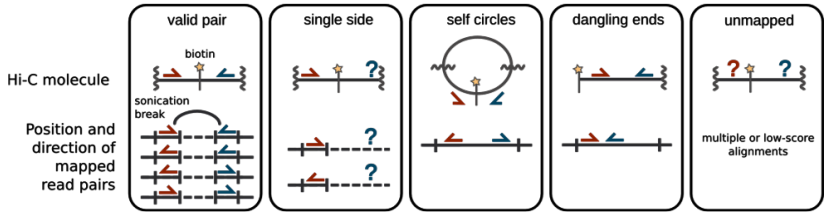
- Iterative or split reads mapping is required.



Adopted from Lajoie et al., The Hitchhiker's guide to Hi-C analysis: Practical guidelines. *Methods* 2015

2.2 Filtering at the level of fragments

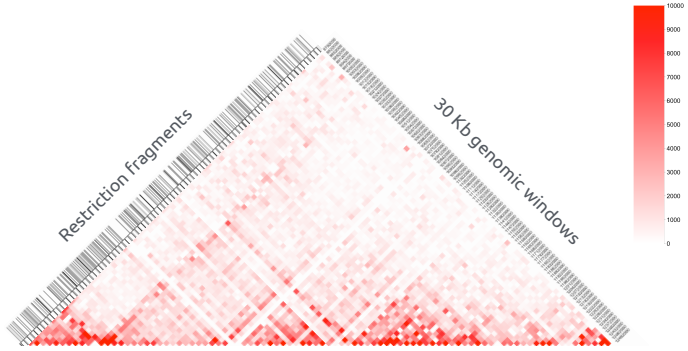
- Possible Hi-C mapping results:



Imakaev et al. *Nature Methods* 2012

2.2 Binning

- Hi-C restriction fragments are assigned to bins (sequential same size genomic windows) and aggregated by taking the sum:



Binned interactions heatmap

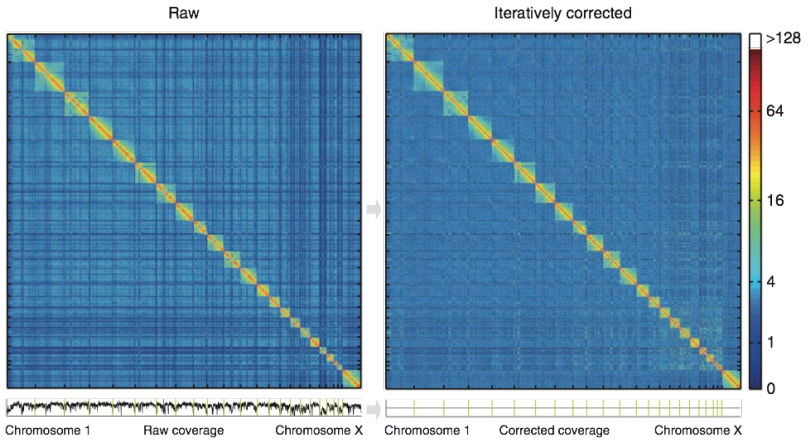
2.3 Matrix balancing

- Balancing is the procedure of correction of systematic technical bias in data.
- Major balancing methods and two general types of balancing:

Approach	Type	Model assumption	Implementation	Computational speed
Yaffe and Tanay	Explicit	Restriction enzyme fragment lengths, GC content and sequence mappability are three major systematic biases in Hi-C	Perl and R	Slow
HiCNorm			R	Fast
Iterative correction (ICE)	Implicit	All the bias is captured by the sequencing coverage of each bin, equal visibility	Python	Fast
Knight and Ruiz			JAVA	Fast
HiC-Pro			Python and R	Very fast

Adopted from Schmitt et al. *Nature Reviews* 2016

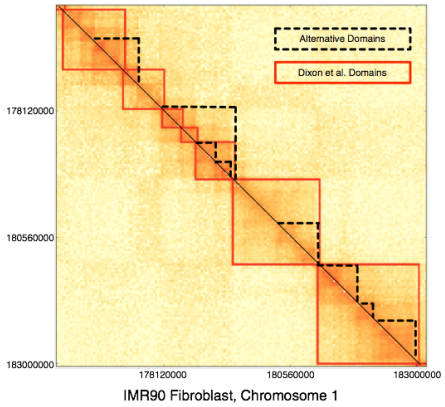
2.3 Iterative correction



Imakaev et al. *Nature Methods* 2012

2.4 TADs calling

- TADs are hierarchical, there is no gold standard for TADs selection:

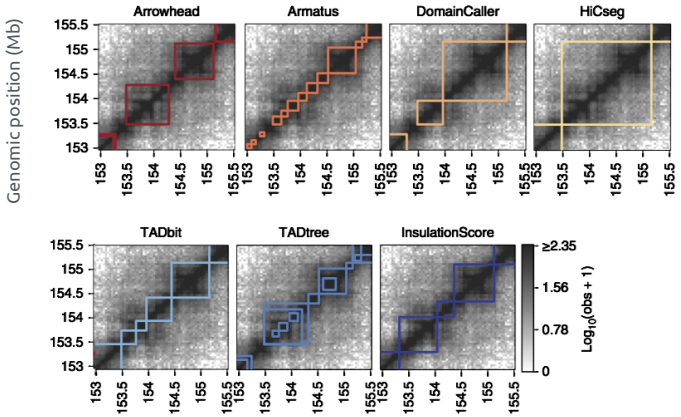


For example, Armatus algorithm is based on dynamic programming and has variable parameter, gamma.

Filippova et al. *Algorithms for Molecular Biology* 2014

2.4 TADs calling

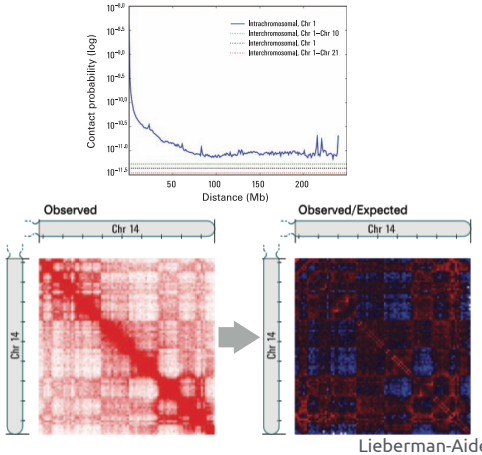
- A recent comparison of multiple TADs calling tools:



Forcato et al. *Nature Methods* 2017

2.4 Compartments calling

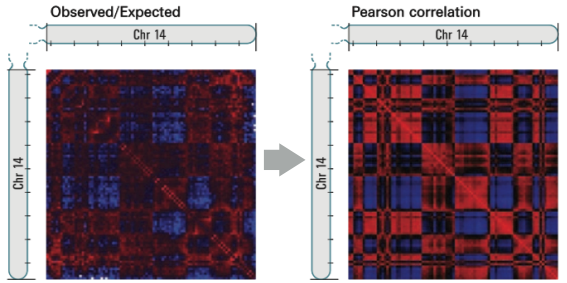
- Method from Lieberman-Aiden, 2009:
  1. Normalization of interaction matrix by expected interactions:



Lieberman-Aiden et al. *Nature* 2009

2.4 Compartments calling

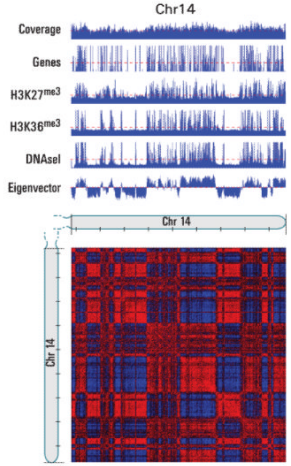
- Method from 2009:
  2. Calculation of Pearson correlation



Lieberman-Aiden et al. *Nature* 2009

2.4 Compartments calling

- Eigenvector decomposition:
  3. Eigenvector expansion (PCA, principal component analysis)



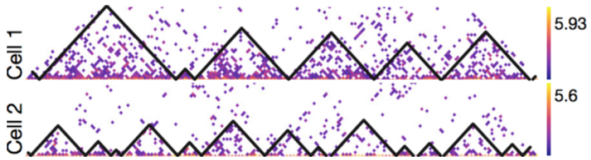
Lieberman-Aiden et al. *Nature* 2009

2.5 Single-cell data analysis

- Generally the same processing workflow, except:
  - Stringent amplification duplicates filtering.

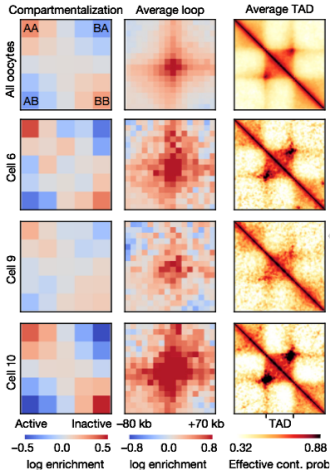
Example elimination of counting the same ligation junction many times (Flyamer et al. *Nature* 2017): if two reads map to the same strand, and each side of the read is within 500 bp of any side of the other read, only one copy of the read is retained.

  - Iterative correction and normalization are not applicable due to data sparsity.



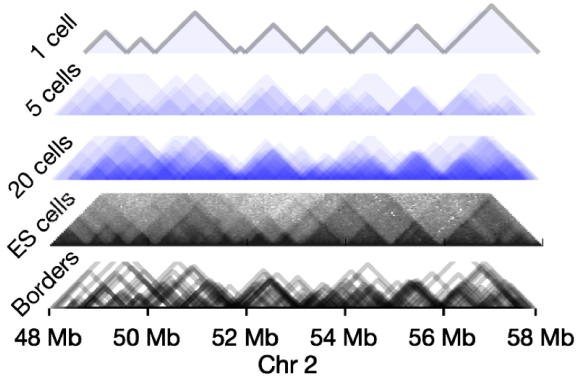
Flyamer et al. *Nature* 2017

- Indirect detection of compartments, TADs and loops due to data sparsity:



Flyamer et al. *Nature* 2017

- Still, TAD-like structures ("contact domains") could be found directly:



Flyamer et al. *Nature* 2017

3. From theory to practice: workshop overview

Workshop overview

- Single-cell and bulk Hi-C raw datasets from Flyamer et al. *Nature* 2017 (GEO: GSE80006)
- Data processing with hiclib (one of the best Hi-C data practices since 2012):
  - Iterative mapping of reads with bowtie2
  - Data filtering
  - Binning
  - Data visualization
  - TADs calling
  - Comparison of single-cell and bulk Hi-C experiments
  - Compartments detection
  - …
- Powered by: python docker

ATAC-seq (Lecture & Workshop)

Katarzyna Kędzierska
University of Virginia, Charlottesville, USA

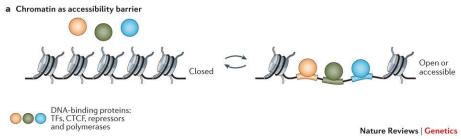
Saturday, 14:00

Assay for Transposase Accessible Chromatin followed by Next Generation Sequencing (ATAC-seq) is rapid, sensitive and efficient method method for mapping chromatin accessibility genome-wide. The method requires no more than 50k or as little as 500 cells as input and the straight-forward protocol comes to isolating the nuclei and in vitro transposition of sequencing adaptors into native chromatin. Samples can be ready for sequencing in less than 3 hours following cell harvest.

Some of the many advantages of the method include: ATAC-seq doesn't require sonication or the phenol-chloroform extraction (FAIRE-seq), no antibodies needed (ChIP-seq), no sensitive enzymatic digestion (MNase-seq or DNA-seq), and significant reduction of the required input material and time needed to process the samples.

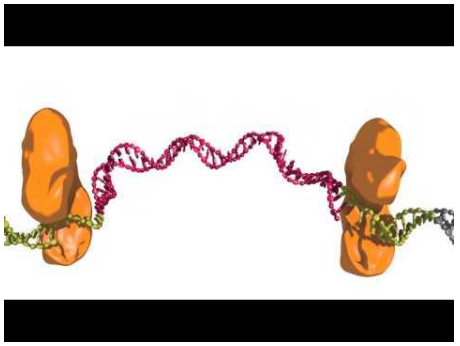
The workshop will focus on hands-on analysis of the already published data. We would work through experimental design, pre-processing the data and the analysis. Topics covered by the workshop include motif search, nucleosome positioning and TFs footprinting.

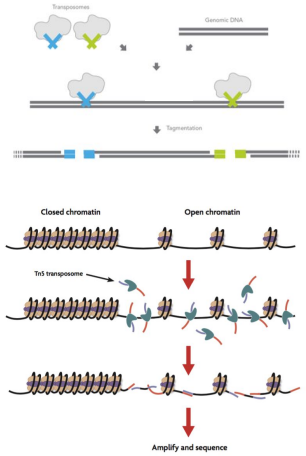
Accessible chromatin



If chromatin is open there's place for DNA-binding proteins, like TFs or polymerases to bind.

Transposition





Research questions

generate epigenomic profiles

map accessible chromatin across tissues or conditions

retrieve nucleosome positions

identify important transcription factors

generate occupancy profiles of TFs (footprinting)


ARTICLE
doi:10.1038/nature18606

The landscape of accessible chromatin in mammalian preimplantation embryos


SCIENTIFIC REPORTS


OPEN

Cell Type-Specific Epigenomic Analysis Reveals a Uniquely Closed Chromatin Architecture in Mouse Rod Photoreceptors

Received: 11 October 2016
Accepted: 19 January 2017
Published: 03 March 2017

Andrew E. O. Hughes, Jennifer M. Enright, Connie A. Myers, Susan Q. Shen & Joseph C. Corbo

Advantages of ATAC-seq

doesn't require sonication or phenol-chloroform extraction (FAIRE-seq)

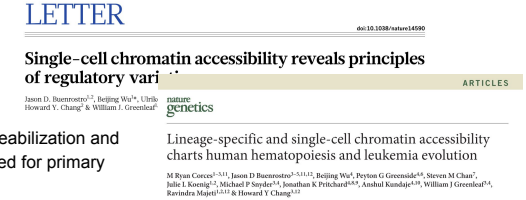
no antibodies needed (ChIP-seq)

no sensitive enzymatic digestion (MNase-seq or DNA-seq)

and significant reduction of the required input material and time needed to process the samples

Variation of the method

scATAC-seq - single cell ATAC-seq


LETTER

doi:10.1038/nature14590

Single-cell chromatin accessibility reveals principles of regulatory vari

ARTICLES

Jason D. Buenrostro[1,2], Beijing Wu[1]*, Ulrik Howard Y. Chang[2] & William J. Greenleaf[1]

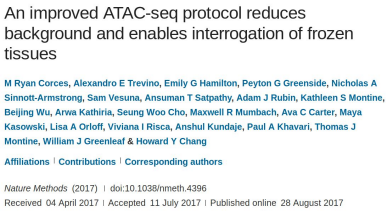
nature genetics

FastATAC - one-step membrane permeabilization and transposition, requires 5k cells; optimized for primary blood cells

Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution

M Ryan Corces[1–3,11], Jason D Buenrostro[2–5,11,12], Beijing Wu[4], Peyton G Greenside[6,8], Steven M Chan[7], Julie I. Koenig[1,2], Michael P Snyder[3,4], Jonathan K Pritchard[6,8,9], Anshul Kundaje[4,10], William J Greenleaf[3,4], Ravindra Majeti[1,2,11] & Howard Y Chang[3,4,12]

OmniATAC - modified version of ATAC protocol, published on Aug 28th 2017.


An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues

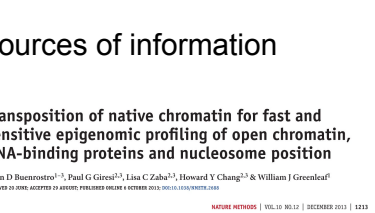
M Ryan Corces, Alexandro E Trevino, Emily G Hamilton, Peyton G Greenside, Nicholas A Sinnott-Armstrong, Sam Vesuna, Ansuman T Satpathy, Adam J Rubin, Kathleen S Montine, Beijing Wu, Arwa Kathiria, Seung Woo Cho, Maxwell R Mumbach, Ava C Carter, Maya Kasowski, Lisa A Orloff, Viviana I Risca, Anshul Kundaje, Paul A Khavari, Thomas J Montine, William J Greenleaf & Howard Y Chang

Affiliations | Contributions | Corresponding authors

Nature Methods (2017) | doi:10.1038/nmeth.4396

Received 04 April 2017 | Accepted 11 July 2017 | Published online 28 August 2017

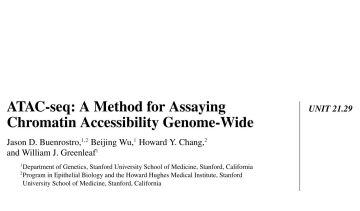
Sources of information


Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position

Jason D Buenrostro[1–3], Paul G Giresi[2,3], Lisa C Zaba[2,3], Howard Y Chang[2,3] & William J Greenleaf[1]

RECEIVED 20 JUNE; ACCEPTED 29 AUGUST; PUBLISHED ONLINE 8 OCTOBER 2013; DOI:10.1038/NMETH.2688

NATURE METHODS | VOL.10 NO.12 | DECEMBER 2013 | 1213


ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide

UNIT 21.29

Jason D. Buenrostro,[1,2] Beijing Wu,[1] Howard Y. Chang,[2] and William J. Greenleaf[1]

[1]Department of Genetics, Stanford University School of Medicine, Stanford, California [2]Program in Epithelial Biology and the Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California

ATAC-seq forum:
https://sites.google.com/site/atacseqpublic/home

Encode guidelines:
https://www.encodeproject.org/atac-seq/ - official pipeline currently in beta tests

Bioconductor support, Biostars forum.


LETTER

doi:10.1038/nature14590

Single-cell chromatin accessibility reveals principles of regulatory variation

Jason D. Buenrostro[1,2], Beijing Wu[1]*, Ulrika M. Litzenburger[3]*, Dave Ruff[3], Michael L. Gonzales[3], Michael P. Snyder[1], Howard Y. Chang[2] & William J. Greenleaf[1,3]

ATAC-seq experimental procedure

Input: crude nuclei or fixed tissue


ATAC-see reveals t
transposase-media

Xingqi Chen[1], Ying Shen[1], Will Drape
Ansuman T Satpathy[1], Ava C Carter[1],
William J Greenleaf[1,2,3], Jan T Liphar

RECEIVED 23 AUGUST; ACCEPTED 19 SEPTEMBER; PUBLISHED

Published online 28 November 2016

Nucleic Acids Research, 2017, Vol. 45, No. 6 e41
doi: 10.1093/nar/gkw1179

Combining ATAC-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes

Zefu Lu[1], Brigitte T. Hofmeister[2], Christopher Vollmers[3], Rebecca M. DuBois[3] and Robert J. Schmitz[1,*]

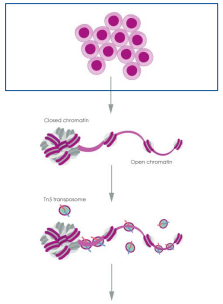
[1]Department of Genetics, University of Georgia, Athens, GA 30602, USA, [2]Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA and [3]Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA

Received August 05, 2016; Revised November 03, 2016; Editorial Decision November 11, 2016; Accepted November 15, 2016

human	$3.3 * 10^9$
mouse	$2.7 * 10^9$
zebrafish	$1.5 * 10^9$
fruit fly	$1.2 * 10^8$
A. thaliana	$1.4 * 10^8$

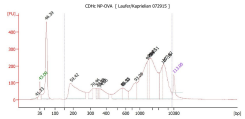
Amount of nuclei depends on **genome** size.

(500, 50k or more) // human and mouse



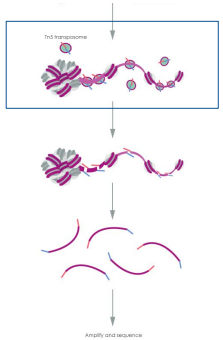
ATAC-seq experimental procedure

Transposition reaction
default: 30 min in 37°C



3. **Question: How do I know how many cells to add to the transposition reaction?**

Answer: Assuming cells are happy, the biggest source of failure comes from variations in cell number. We see biggest differences in the requirement of the number of cells between species; however, variation exists between cell types as well. If desired, a good way to troubleshoot or improve signal-to-noise for your particular application is to do a titration of cells, and if your cheap like I am, I would scale the reaction down 10x and titrate using 5,000 cells and 5uL transposition reactions. When you find a sample that best matches the gel above, then simply scale up to the 50uL reaction.



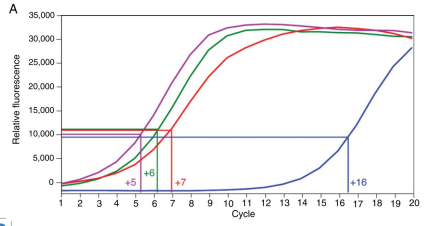
Over-transposition:
- Increase number of nuclei
- Decrease enzyme volume (non linear)

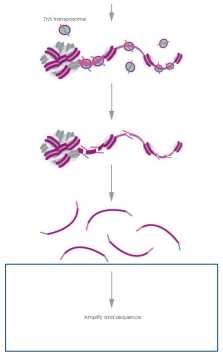
Under-transposition:
- Decrease number of nuclei
- Increase time of reaction

ATAC-seq experimental procedure

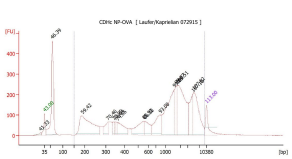
Perform PCR - 5 cycles -> take aliquot and do qPCR to calculate how many additional cycles need to be run.

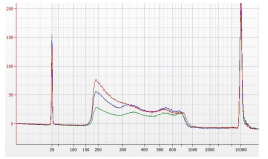
# of additional cycles: ⅓ of the max fluorescence intensity





ATAC-seq experimental procedure





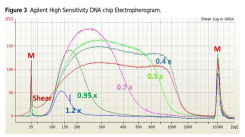
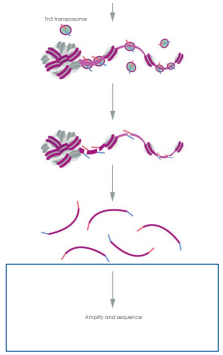


Figure 3 Agilent High Sensitivity DNA chip Electropherogram.

M = upper and lower markers for High Sensitivity DNA chip.
Shear = 1 µL of 20 ng/µL input control sample in water.
1.2x to 0.4x = 1 µL of shear, size selected with given ratio of **SPRIselect** volume to sample volume.

SPRIselect User Guide, Beckman Coulter

size selection: double sided / right side beads selection



Experimental design

Control:

experiment dependent, no need for “input”

Replicates:

experiment dependent, at least **two** biological replicates - if there is high variability among samples I would recommend more

Library type:

paired-end, with single-end only some analysis can be performed

Sequencing depth:

depends on the genome size, assuming 70% mappability ratio to satisfy encode standard 70 milion reads would be needed

Analysis workflow

Reads processing quality assessment filtering and trimming if necessary	fastqc with multiqc trim galore
Alignment	bowtie / bwa mem
Alignment processing quality filtering filtering out blacklisted regions shifting alignments	samtools bedtools R ATACseqQC
Peak calling	macs2

Before you start

There are two types of people: those who backup, and … those who will backup.

Keep raw, unprocessed data until your experiment is safely deposed in the database (ENA, GenBank or DDBJ).

Reads processing - fastq file format

```
@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATATGACAAGACAGAATC
+
<>;##=><9=AAAAAAAAA9#:<#<;<<<????#=
```

sequence identifier (always starts with @)
sequence [ATCG]
quality score identifier
quality score

Sequence identifier always starts with @ but how it is constructed depends on the source of files (sequencing platform or database)

<is filtered>	N - means that read passed the filtering Y - means that read did not pass Illumina filtering

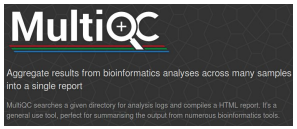
Sample of the reads filtering

```
zcat sample.raw.fastq.gz
grep -A 3 '^@.*[^:]*:N:[0-9]*:'
grep -v '^\-\-$'
gzip
> sample.filtered.fastq.gz
```

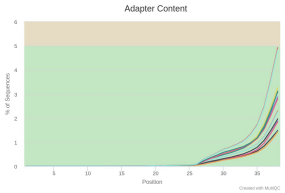
- ➔ Open and print to stdout reads in gzipped fastq files
- ➔ Find all lines having the desired pattern (:N:) and print that line and 3 following
- ➔ Don't print lines with only -
- ➔ Gzip output
- ➔ Save output to this file

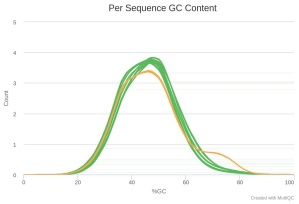
```
zcat sample.raw.fastq.gz  | grep -A 3 '^@.*[^:]*:N:[0-9]*:' | grep -v '^\-\-$' |
gzip > sample.filtered.fastq.gz
```


Reads processing



http://multiqc.info/





Trim Galore!
wrapper, cutadapt + fastQC

trims poor quality 3' bases, trims adapters (either specified or found in first 1 milion sequences) and discards too short reads

https://github.com/FelixKrueger/TrimGalore

Alignment

Aligning short reads to reference genome

Bowtie2

```
bowtie2 \
        -x ${reference_genome} \
        -1 <(zcat ./cleaned/${sample}_R1.fastq.gz) \
        -2 <(zcat ./cleaned/${sample}_R2.fastq.gz)  \
        -p ${n_threads} \
        --very-sensitive \
        -X 2000
```

bwa mem (or bwa aln)

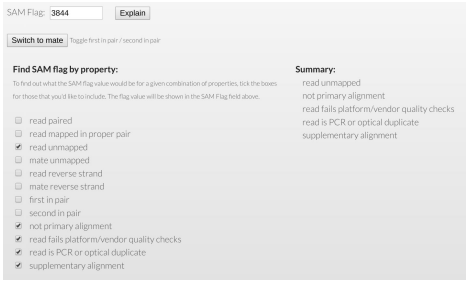
```
bwa mem \
        -v 3 \
        -t ${n_threads} ${reference_genome} ./cleaned/${sample}_R1.fq.gz ./cleaned/${sample}_R2.fq.gz 2> ./bwa/${sample}.log |
                samtools view -b -@ ${compression_threads} -o ./bwa/${sample}_raw.bam
```


Alignment - SAM

```
@HD     VN:1.3  SO:coordinate
@SQ     SN:1    LN:195471971
@SQ     SN:10   LN:130694993
@SQ     SN:11   LN:122082543
@SQ     SN:12   LN:120129022
@RG     ID:id   SM:sample LB:lib
@PG     ID:bwa  PN:bwa  VN:0.7.15-r1142-dirty   CL:bwa mem -v 3 -t 16 -R @RG\tID:id\tSM:sample\tLB:lib ./ref/Mdna.toplevel.fa ./sample_R1_trimmed.fq.gz
./sample_R2_trimmed.fq.gz
HWI-ST1309F:284:C8KYNANXX:2:2110:2778:8477   99      1       3000081 40      35M     =       3000141 79
CCCATCTGGTCCTGGGCTTTTTTTTTTTTTTTTT              BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF     NM:i:0  MD:Z:35 AS:i:35 XS:i:35
```

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,254}	Query template NAME
2	FLAG	Int	[0,2^{16}-1]	bitwise FLAG
3	RNAME	String	*|[!-()+-<>-~][!-~]*	Reference sequence NAME
4	POS	Int	[0,2^{31}-1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,2^{8}-1]	MAPping Quality
6	CIGAR	String	*|([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	*|=|[!-()+-<>-~][!-~]*	Ref. name of the mate/next read
8	PNEXT	Int	[0,2^{31}-1]	Position of the mate/next read
9	TLEN	Int	[-2^{31}+1,2^{31}-1]	observed Template LENgth
10	SEQ	String	*|[A-Za-z=.]+	segment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

Alignment - Flags



https://broadinstitute.github.io/picard/explain-flags.html

samtools view

-q INT Skip alignments with MAPQ smaller than INT [0].

-f INT Only output alignments with **all bits** set in INT present in the FLAG field.

-F INT **Do not** output alignments with **any bits** set in INT present in the FLAG field.

-G INT **Do not** output alignments with **all bits** set in INT present in the FLAG field. This is the opposite of -f such that -f12 -G12 is the same as no filtering at all.

Alignment processing

Blacklisted regions - regions having high signal / read counts independent of cell line or experiment type.
Available for human, mouse, worm and fruit fly.
https://sites.google.com/site/anshulkundaje/projects/blacklists

```
#n - number of chromosomes in a given organism
chromosomes=$(echo $(for i in {1..n}; do echo "chr"$i; done | xargs) "chrX" "chrY");

bedtools intersect -v -abam ./bwa/${sample}.bam -b ${blacklisted_regions} |
     samtools view -h -b -F 3844 -f 2 -q 5 ${sample}_filtered.bed ${chromosomes} |
          samtools sort -n -T ${sample}_tmp -o ${sample}_sorted.bed -@ ${n_threads};
```

Alignment processing - quality check

Collect statistics

```
samtools flagstat ${sample}.bam > ${sample}.txt
```

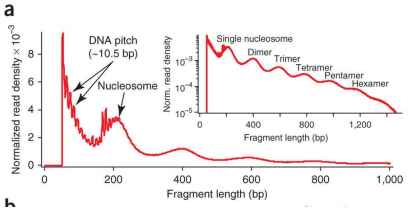
ENCODE standards:

at least 25 milion non-duplicate, non-mitochondrial reads

more than 80%, preferable more than 95% mapped reads

Non-Redundant Fraction (i.e. # of non-duplicate reads / total # of reads) > 0.9

Alignment processing - quality check



Check fragment size distribution.

Experimental design (size selection)

Peak calling - MACS2

```
macs2 callpeak \
        --verbose 3 \
        --treatment ${sample}_sorted.bam \
        -g hg \
        -B \
        -q 0.05 \
        --extsize 200 \
        --nomodel \
        --shift -100 \
        --nolambda \
        --keep-dup all \
        -f BAM \
        --outdir ./peaks/${sample}\
        --call-summits
```


Method

Open Access

Model-based Analysis of ChIP-Seq (MACS)

Yong Zhang[*], Tao Liu[*], Clifford A Meyer[*], Jérôme Eeckhoute[†], David S Johnson[‡], Bradley E Bernstein[§¶], Chad Nusbaum[¶], Richard M Myers[§], Myles Brown[†], Wei Li[#] and X Shirley Liu[*]

Published: 17 September 2008
Genome Biology 2008, 9:R137 (doi:10.1186/gb-2008-9-9-r137)
The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2008/9/9/R137
Received: 4 August 2008
Revised: 3 September 2008
Accepted: 17 September 2008

From the MACS2 manual

Here are some examples for combining --shift and --extsize:

1. To find enriched cutting sites such as some DNAse-Seq datasets. In this case, all 5' ends of sequenced reads should be extended in both direction to smooth the pileup signals. If the wanted smoothing window is 200bps, then use '--nomodel --shift -100 --extsize 200'.

2. For certain nucleosome-seq data, we need to pileup the centers of nucleosomes using a half-nucleosome size for wavelet analysis (e.g. NPS algorithm). Since the DNA wrapped on nucleosome is about 147bps, this option can be used: '--nomodel --shift 37 --extsize 73'.

README for MACS: https://github.com/taoliu/MACS

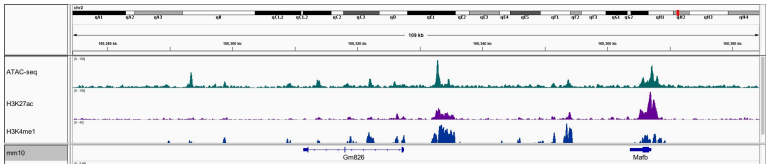
Peak calling - file formats

BED (Browser Extensible Data) - 3 columns (chrom, chromStart, chromEnd) required
BED6, BED6+4 - BED files with additional columns
bedGraph - version of BED file used for visualisation

MACS2 outputs:

1. **NAME_peaks.xls** - header with run description; chrom, chromStart, chromEnd, length, summitPosition (absolute), pileup (at summit), -log10(pvalue), fold_enrichment, -log10(qvalue), name
2. **NAME_peaks.narrowPeak** (BED6+4) - chrom, chromStart, chromEnd, name, score, strand, integer score, fold-change, -log10pvalue, -log10qvalue, summitPosition (from peak start)
3. **NAME_summits.bed** (BED) - location of summits
4. **NAME_treat_pileup.bdg** (bedGraph) - chrom, chromStart, chromEnd, signal

Peak calling - quality control

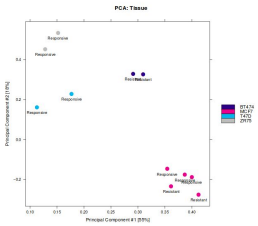


Fraction of reads in peaks (FRiP) - Fraction of all mapped reads that fall into the called peak regions, i.e. usable reads in significantly enriched peaks divided by all usable reads.

FRiP should be >0.3, though values greater than 0.2 are acceptable

Peak calling - quality control

Visualize for example by PCA plot



http://setosa.io/ev/principal-component-analysis/

Transcription Start Site (TSS) Enrichment Score

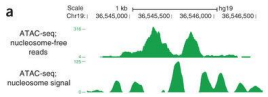
Annotation used	Value	Resulting Data Status
hg19 Refseq TSS annotation	<6	Concerning
	6-10	Acceptable
	>10	Ideal
GRCh38 Refseq TSS annotation	<5	Concerning
	5-7	Acceptable
	>7	Ideal
mm9 GENCODE TSS annotation	<5	Concerning
	5-7	Acceptable
	>7	Ideal
mm10 Refseq TSS annotation	<10	Concerning
	10-15	Acceptable
	>15	Ideal

Accessible chromatin ≠ open chromatin

Open chromatin - can be defined as nucleosome free region

Accessible chromatin - regions of chromatin that are accessible for transposase

It all depends on what you need and how you call peaks!

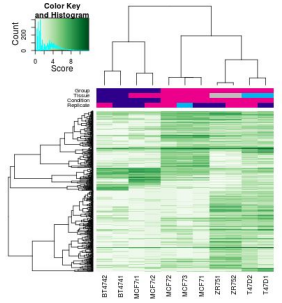


Differentially accessible regions

Task: identify the sites that are accessible in one, but not the other sample

Diffbind package in R based on DESeq2

1. Scan peaksets and merge them creating consensus.
2. Create matrix of counts, peaks x samples.
3. Calculate the library size, normalize.
4. Apply statistical tests to assess which sites are differentially open.



Gene set enrichment

Enrichr

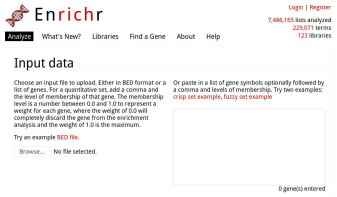
can use both gene symbols or BED file

http://amp.pharm.mssm.edu/Enrichr/

GREAT

"GREAT assigns biological meaning to a set of non-coding genomic regions by analyzing the annotations of the nearby genes."

http://bejerano.stanford.edu/great/public/html/




GREAT improves functional interpretation of *cis*-regulatory regions

Cory Y McLean[1], Dave Bristor[1,2], Michael Hiller[2], Shoa L Clarke[3], Bruce T Schaar[2], Craig B Lowe[4], Aaron M Wenger[1] & Gill Bejerano[1,2]

NATURE BIOTECHNOLOGY VOLUME 28 NUMBER 5 MAY 2010 495

Motif search

Identify transcription factors bound to the chromatin

HOMER (v4.9, 2-20-2017)

Software for motif discovery and next generation sequencing analysis

The MEME Suite

Motif-based sequence analysis tools

Footprinting

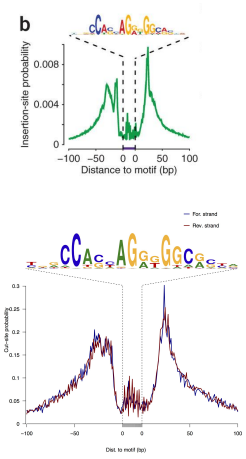
CENTIPEDE

integrates histone modifications or DNase I cleavage patterns with genomic information such as gene annotation and evolutionary conservation to generate genome-wide map of transcription factor binding sites

Pique-Regi, R., Degner, J., & Pai, A. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Research, 3, 447–455. https://doi.org/10.1101/gr.112623.110.Freely

ATACseqQC

doesn't use the conservation (PhyloP)



Workshop outline

Workshop will cover:

1. Checking alignment quality
2. Shifting and splitting the reads with R package ATACseqQC
3. Calling peaks with MACS2
4. Motif search with HOMER
5. Identifying differentially bound sites with R package Diffbind
6. Basic enrichment analysis with GREAT and Enrichr
7. Transcription Factor footprinting with R package ATACseqQC

Software

Reads processing	Alignment	Alignment processing	Peak calling
cutadapt Trim Galore! Trimmomatic	bwa bowtie2	Samtools bedtools Picard Sambamba	MACS2 Hotspot Homer ZINBA

MEME suite

CENTIPEDE - tutorial msCENTIPEDE R ATACseqQC pyDNase

NucleoATAC chromVAR

R DiffBind R ChIPseeker R ChIPQC

grep, awk

Summary

1. Optimize the procedure and analysis for a given experiment.
2. Design the experiment.
   a. Choose proper controls;
   b. Consider tissue and sample type;
   c. Set your goals.
3. Keep unprocessed data until you depose it in a database.
4. Carefully read software documentation before using it.
5. Do quality checks and follow guidelines.
   a. Check raw and processed (filtering, trimming) reads;
   b. Filter alignment (blacklisted, uncanonical, low quality);
   c. Check fragment size distribution, mappability ratio, NRF;
   d. Calculate FRiP and TSS enrichment.

Acknowledgments

Aakrosh Ratan
Center for Public Health Genomics
University of Virginia

Microbial genomics (Lecture & Workshop)

Adam Witney
SGUL, London

Saturday, 14:00

NGS is transforming clinical microbiology by enabling the prediction of drug resistance profiles in a much faster and robust way, leading to more appropriate, personalised treatment plans for patients. The result is better clinical outcomes for the patients, potentially reducing the risk of the development of antimicrobial resistant and reducing overall healthcare costs. In addition the high resolution obtained by NGS enables accurate tracking of infectious disease isolates, thus enabling doctors and public health officials to track and intervene rapidly in outbreak scenarios. In this workshop you will hear of some real examples where NGS has been used to direct patient care and track active outbreaks. You will then perform the exact analysis yourselves, using sequence analysis tools to predict the drug resistance profiles of Tuberculosis isolates from a set of patients and then predict the transmission network between these patients.

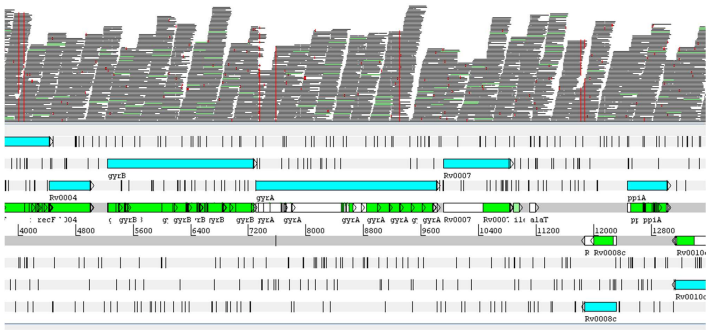
Overview

- Drug resistance prediction
  - *Mycobacterium tuberculosis*
- Outbreak investigation
  - *Mycobacterium tuberculosis*
  - *Pseudomonas aeruginosa*

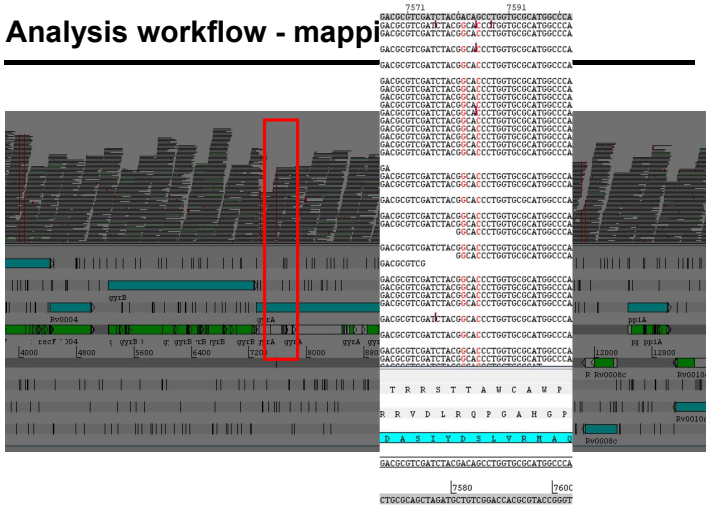
Drug resistance prediction

- *Mycobacterium tuberculosis*
  - Major public health priority world wide
  - MDR and XDR phenotypes widespread
  - slow growing organism
  - Current diagnostics
    - Chest X-ray
    - Sputum sample for microscopy and culture
    - specific molecular tests (e.g. GeneXpert)
  - Drug resistance profile – 4-8 weeks

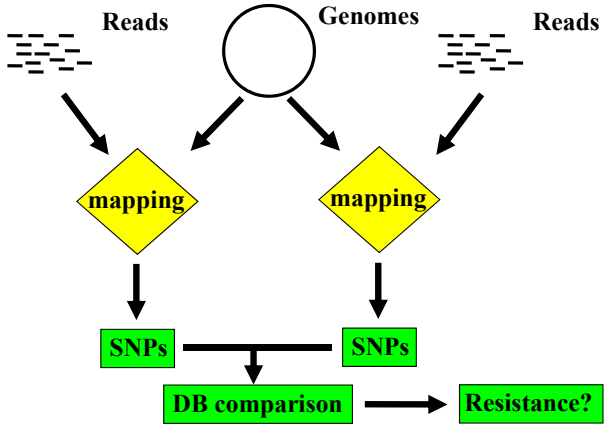
Analysis workflow - mapping



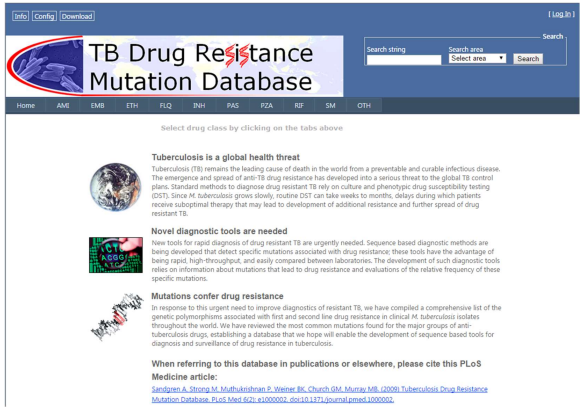
Analysis workflow - mappi



Analysis workflow - resistance

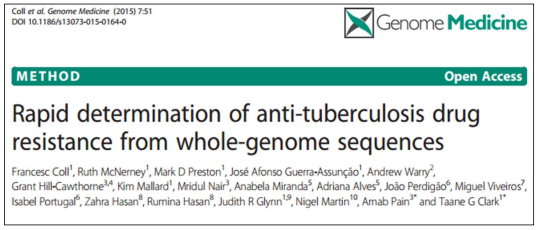


Resistance databases - TBDreamDB



https://tbdreamdb.ki.se/Info/

Resistance databases



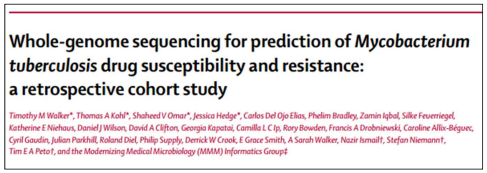
Coll et al. Genome Medicine (2015) 7:51
DOI 10.1186/s13073-015-0164-0

Genome Medicine

METHOD Open Access

Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences

Francesc Coll[1], Ruth McNerney[1], Mark D Preston[1], José Afonso Guerra-Assunção[1], Andrew Warry[2], Grant Hill-Cawthorne[3,4], Kim Mallard[1], Mridul Nair[3], Anabela Miranda[5], Adriana Alves[5], João Perdigão[6], Miguel Viveiros[7], Isabel Portugal[6], Zahra Hasan[8], Rumina Hasan[8], Judith R Glynn[1,9], Nigel Martin[10], Arnab Pain[3*] and Taane G Clark[1*]



Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study

Timothy M Walker*, Thomas A Kohl*, Shaheed V Omar*, Jessica Hedge*, Carlos Del Ojo Elias, Phelim Bradley, Zamin Iqbal, Silke Feuerriegel, Katherine E Niehaus, Daniel J Wilson, David A Clifton, Georgia Kapatai, Camilla L C Ip, Rory Bowden, Francis A Drobniewski, Caroline Allix-Béguec, Cyril Gaudin, Julian Parkhill, Roland Diel, Philip Supply, Derrick W Crook, E Grace Smith, A Sarah Walker, Nazir Ismail†, Stefan Niemann†, Tim E A Peto†, and the Modernizing Medical Microbiology (MMM) Informatics Group‡

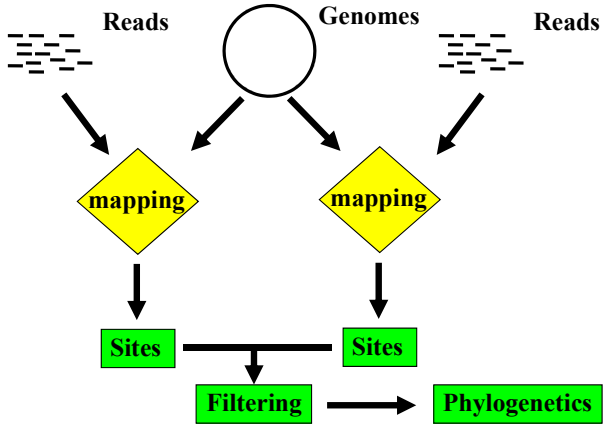
Drug resistance prediction

- ***Mycobacterium tuberculosis***
  - **Highly monomorphic genome**
  - **Relatively straight forward to predict resistance**
- **Gram negatives more difficult**
  - **e.g. *Escherichia coli, Pseudomonas aeruginosa, Klebsiella pneumoniae***
  - **Resistance determinants on mobile elements, e.g. plasmids**
  - **Chromosomal efflux pumps**

Outbreak investigation

- How to identify outbreak clusters?
  - *Is there a link between patients e.g. same ward in hospital?*
  - *Antibiotic resistance profiles*
  - *Strain Typing*
    - *Serotyping – serological testing of capsule locus*
    - *VNTR – Variable Number Tandem Repeat*
    - *MLST – Multi Locus Sequence Typing*

Analysis workflow – phylogenetics



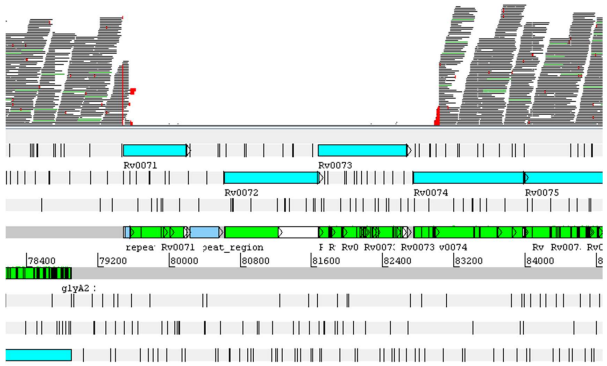
Site calling – phylogenetics

- Build consensus for all sites for each isolate
- Filter sites
  - Depth of coverage

```
A   GATGGTA
B   GATCGTC
C   GATCATA
D   GCTCGTA

R   GATCGTA
```

Site calling - depth filtering



Site calling - phylogenetics

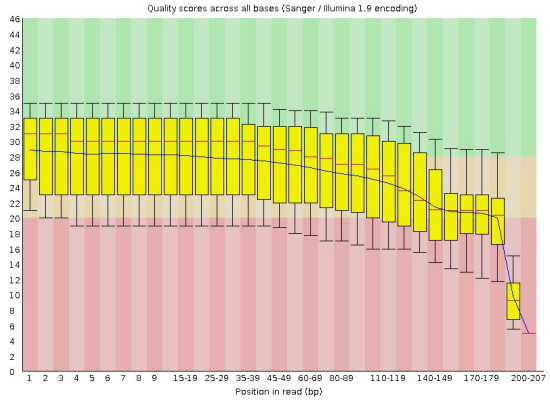
- Build consensus for all sites for each isolate
- Filter sites
  - Depth of coverage
  - Base quality score

```
A   GATGGTA
B   GATCGTC
C   GATCATA
D   GCTCGTA

R   GATCGTA
```

Site calling - quality filtering

Q



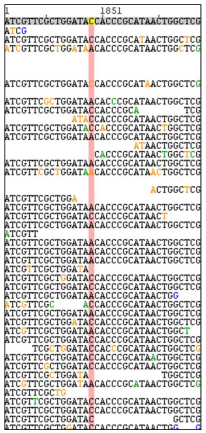
Site calling - phylogenetics

- Build consensus for all sites for each isolate
- Filter sites
  - Depth of coverage
  - Base quality score
  - Heterogenous sites

```
A   GATGGTA
B   GATCGTC
C   GATCATA
D   GCTCGTA

R   GATCGTA
```

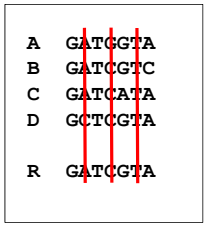
Site calling - heterogenous



C	72 reads	67%
A	35 reads	33%

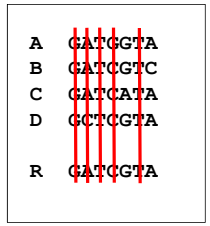
Site calling - phylogenetics

- Build consensus for all sites for each isolate
- Filter sites
  - Depth of coverage
  - Base quality score
  - Heterogenous sites



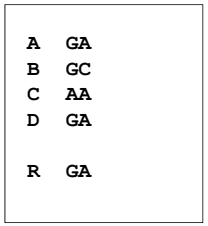
Site calling - phylogenetics

- Build consensus for all sites for each isolate
- Filter sites
  - Depth of coverage
  - Base quality score
  - Heterogenous sites
- Remove filtered sites from all isolates
- Remove non-variant sites
- Concatenate remaining variant sites

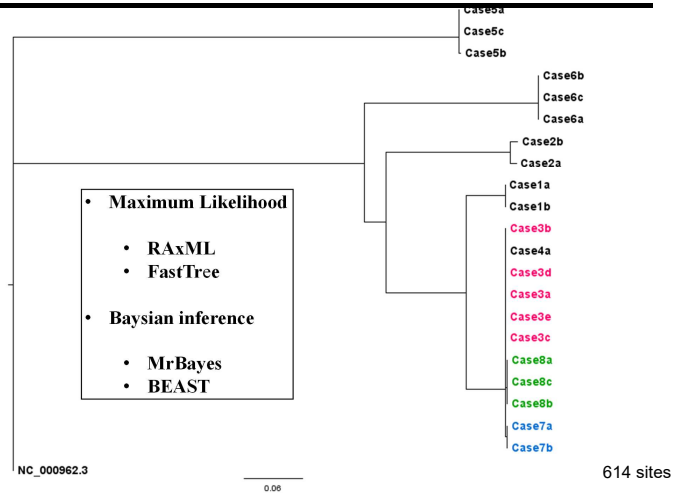


Site calling - phylogenetics

- Build consensus for all sites for each isolate
- Filter sites
  - Depth of coverage
  - Base quality score
  - Heterogenous sites
- Remove filtered sites from all isolates
- Remove non-variant sites
- Concatenate remaining variant sites
- Maximum Likelihood estimation (RAxML)

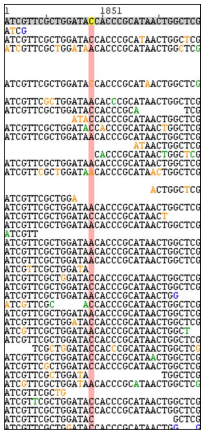


Analysis workflow - phylogenetics



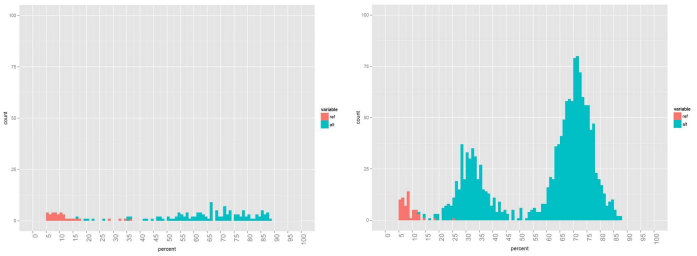
614 sites

Heterogenous sites – mixed sites



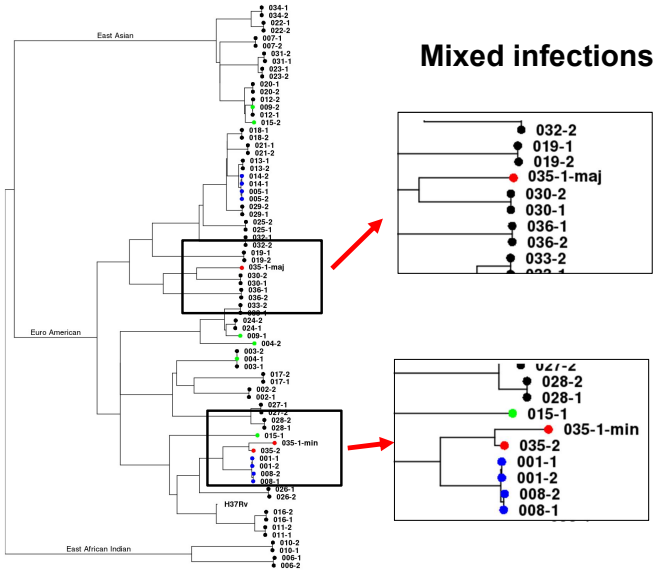
C	72 reads	67%
A	35 reads	33%

Mixed infections



126 sites 1449 sites

Mixed infections



Outbreak investigation

- **PHE sequenced 90 Pseudomonas aeruginosa ST111 isolates from across the UK**

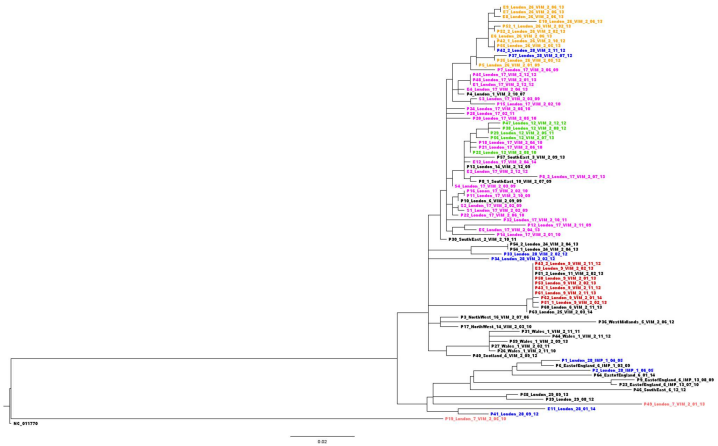
JCM

High-Resolution Analysis by Whole-Genome Sequencing of an International Lineage (Sequence Type 111) of *Pseudomonas aeruginosa* Associated with Metallo-Carbapenemases in the United Kingdom

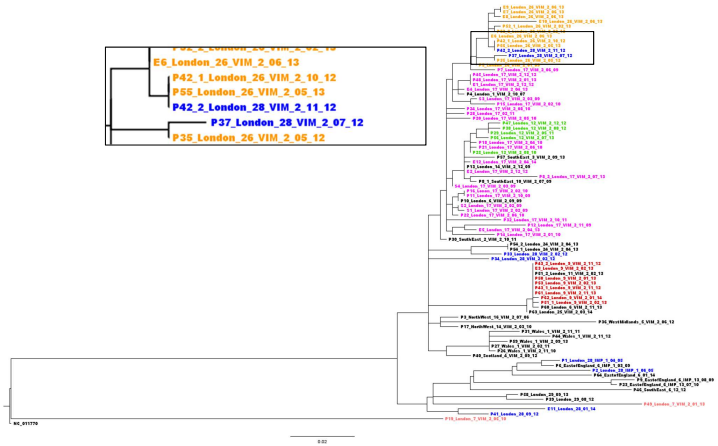
Jane F. Turton,[a] Laura Wright,[a] Anthony Underwood,[b] Adam A. Witney,[a] Yuen-Ting Chan,[c] Ali Al-Shahib,[b] Catherine Arnold,[c] Michel Doumith,[a] Bharat Patel,[a] Timothy D. Planche,[a,d] Jonathan Green,[b] Richard Holliman,[a,e,f] Neil Woodford[a]

Antimicrobial Resistance and Healthcare Associated Infections (AMRHAI) Reference Unit,[a] Infectious Disease Informatics,[b] and Genomic Services and Development Unit,[c] Public Health England, Colindale, London, United Kingdom; Public Health Laboratory London, Whitechapel, London, United Kingdom[d]; Institute of Infection and Immunity, St George's University of London, London, United Kingdom[e]; Department of Microbiology, St George's Healthcare NHS Trust, London, United Kingdom[f]

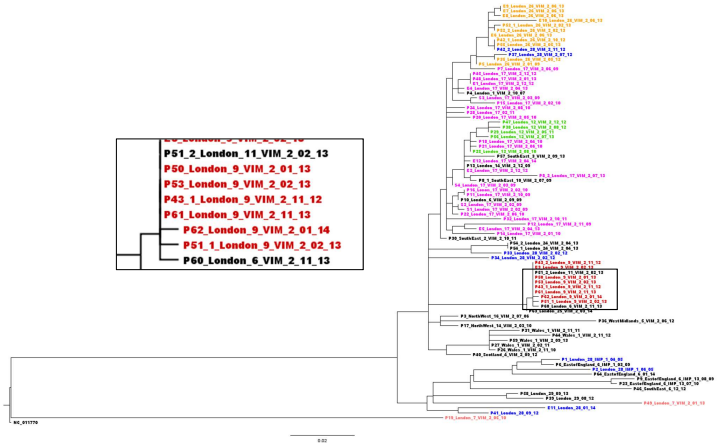
P. aeruginosa ST111 in the UK



P. aeruginosa ST111 in the UK

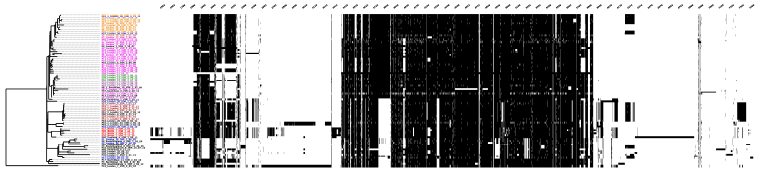


P. aeruginosa ST111 in the UK



P. aeruginosa ST111 in the UK

- **Accessory elements**



Summary

- **Identify resistance associated SNPs**
- **Identify circulating accessory elements**
- **Suggest possible therapies**
- **Identify transmission chains**
- **Inform infection control strategies**
- **Inform public health strategies**

Open science (Discussion)

Paweł Szczęsny
IBB, Warsaw

Wednesday, 17:30

Team-working (Lecture & Discussion)

Panagiotis Theodorakis
Institute of Physics, PAS, Warsaw

Thursday, 20:00

Every step forward in science and our society is a result of team-work

Examples:

1) Sharing of ideas (books, journals, discussions)
2) Sharing work load
3) Sharing emotions
4) Building up on previous work
5) This summer school is a TEAM WORK!!!!

In the end of the day you would like your scientific output to be useful for other people and also appreciated for yourself!
So, at least be kind to other people!!!!

Good Team

Innovative

Trust

Work effortlessly

Supportive

Motivation

Participation

Good communication

Team-work is the cornerstone of every successful Institution

Together Everyone Achieves More

TEAM:

- Allows us to achieve uncommon results
- Environment allows everyone to go beyond their limitation

Benefits of team work

Improve productivity

Distribution of work load

Diversity of ideas

Better decision

Quick solution

Motivation

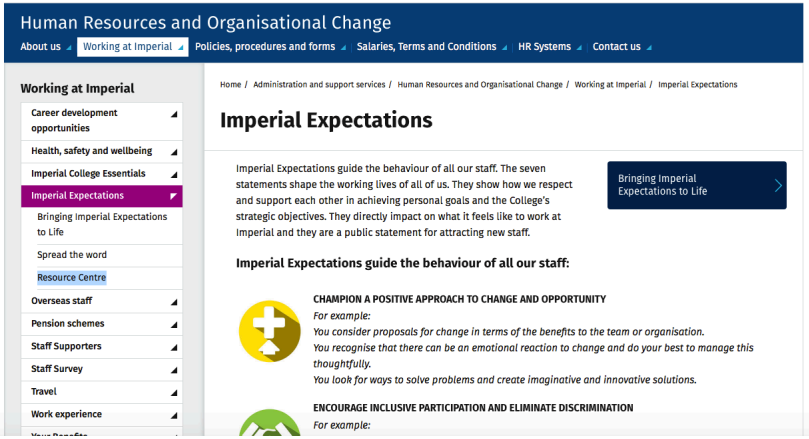
Learning

Role of Project Leader:

- Control
- Inspire
- Adjust
- Update
- Advice
- Consult

Our example: Imperial College London

A look at Imperial College's website. Sharing my personal experience.
http://www.imperial.ac.uk/human-resources/working-at-imperial/imperial-expectations/



No matter what is our role, we should all champion the following approaches!!!!

CHAMPION A POSITIVE APPROACH

- You consider proposals for change in terms of benefits to the team and the organisation
- You recognise that there can be an emotional reaction to change and do your best to manage this thoughtfully
- You look for ways to solve problems and create imaginative and innovative solutions
- Stimulating frank and open discussion about the major changes that need to be made.
- Challenging political interests and functional barriers to achieve improvements in how we operate.
- Establishing stretching goals to rethink how we operate and raise everyone's expectations of what can be achieved.

- Establishing a clear and coherent game plan to indicate what we need to do to sustain and build future success.
- Consulting and involving others fully in planning any future changes within the work area.
- Anticipate change, plan for it and manage expectations
- Lead by example – maintain a professional and positive outlook to change even if you don't agree with it
- Be willing and flexible to accept and adapt to change
- Discuss what changes might occur in PRDPs and explore what opportunities this might offer to each individual
- Highlight the opportunities for growth that often come with change – embrace it
- Ensure fair access to opportunities

ENCOURAGE INCLUSIVE PARTICIPATION AND ELIMINATE DISCRIMINATION

- You treat individuals with respect.
- You support team working and involve others in team activities – formal and informal.
- You challenge behaviour, actions and words that do not support the promotion of equality and diversity.
- You comply with legal requirements and organisational policies
- Be a good role model – champion equalities
- Encourage Equality & Diversity training for all
- Provide support
- Champion teamwork and provide team building
- Be open to feedback and ideas from your staff
- Be clear about what is acceptable behaviour and what isn't.
- Create a culture that supports inclusion for all

COMMUNICATE REGULARLY AND EFFECTIVELY WITHIN, AND ACROSS, TEAMS

- *You use communication styles appropriate to different people and situations.*
- *You present information clearly, concisely and accurately to promote understanding.*
- *You ask questions and listen with care in order to understand better*
- Have regular check-ins/meetings in the team and across the department
- Have regular catch-ups/1:1s with team members
- Address miscommunications
- Promote feedback, discussion, queries, critical reflection and continually incorporate it into next stage/meeting
- Ensure clear channels of information and how it will cascade

- Challenge unacceptable behaviours
- Operate fair, transparent recruitment and selection
- Creating an environment in which everyone feels confident to express their ideas.
- Being objective and even handed in dealing with any difficult interpersonal issues.
- Tackling immediately any suggestion of discriminatory or inappropriate behaviour.
- Managing the dynamics of diversity within the work group positively to develop a productive work environment.
- Involving others fully in decision making, encouraging their input and considering their opinions.

- Be approachable and available
- Have a communication plan
- Ensure timely communication
- Adopting a collaborative approach in my dealings with colleagues in other work areas.
- Building and drawing on the full talents and energies of everyone in the team in working towards shared goals.
- Tackling openly any conflict with other work areas in a positive and constructive way to achieve "win-win" outcomes.
- Adapting my interpersonal approach easily to respond to the demands of different situations and audiences.
- Facilitating effective team briefings in which the key issues are raised and resolved constructively.
- Using a Coaching style for team meetings

CONSIDER THE THOUGHTS AND EXPECTATIONS OF OTHERS

- You ask for and listen to other people's views and ideas.
- You take other people's views into account when planning and setting deadlines
- You discuss and agree what is expected of others and what they can expect of you
- Ask for and pay attention to feedback from staff
- Understand the individuals in your team
- Respond to and follow up concerns people have
- Take into account people's previous experience
- Be empathic, compassionate and considerate while remaining professional
- Listen and keep the channels open
- Be clear and discuss expectations of individuals
- Be clear about standards in operation – yours and the ones from the College

DELIVER POSITIVE OUTCOMES

- You deliver results and consistently seek to improve your performance
- You monitor the quality of your work and progress against plans and take appropriate corrective action, where necessary
- You look for ways to contribute to the group or team's success.
- Achieve tasks within deadline
- Produce and get papers published
- Create a happy working atmosphere
- Share the vision and goals of the team
- Establish clarity about the task and plan accordingly
- Fit skills to tasks with discussion when you distribute work
- Ensure effective communication
- Support and reward achievements

- Recognising how my own plans affect work colleagues and taking their priorities into account.
- Listening actively to others with a genuine interest to find out what matters to them.
- Making it easy for others to give open and frank feedback on my impact and effectiveness
- Committing quality time to encourage each individual in the team to feel an important and valued part of our overall efforts
- Giving others clear and consistent feedback on the individual performance and contribution

- Give constructive feedback and development opportunities
- Cutting through any irrelevant discussions quickly to tackle the practicalities of what needs to be done
- Responding with speed and urgency to the issues which others pass for my attention
- Implementing robust systems and practices to manage work and achieve desired outcomes
- Conducting regular reviews into the way things are done to identify major improvements in effectiveness and efficiency.
- Supporting and translating promising ideas into agreed and well defined goals that focus individual and team effort

DEVELOP AND GROW SKILLS AND EXPERTISE

- *You take advantage of opportunities to learn and develop*
- *You ask for feedback from others and use it to develop*
- *You encourage others to learn and develop*
- Encourage people to take up learning and development/training opportunities
- Give positive and constructive feedback
- Take PRDPs seriously and invest time in the discussions about development
- Take one to one meetings seriously
- Discuss expectations and career aspirations
- Pass on information about relevant courses to staff

- Being approachable to make it easy for others to discuss any concerns and worries.
- Recognising where others' real skills like and giving them the opportunity to develop their talents to the full.
- Reviewing and agreeing personal objectives and priorities with all my staff on a regular basis.
- Providing the practical coaching and career guidance my staff need to help maximise their effectiveness.
- Delegating work in a structured way to encourage others to take on additional responsibility

WORK IN A PLANNED AND MANAGED WAY

- *You make best use of your time at work*
- *You plan and prioritise tasks*
- *You are mindful of other's work priorities when working with others*
- Have task based meetings and clear project timelines
- Have a schedule of meetings set up in advance
- Set clear objectives with realistic deadlines
- Plan in advance to give people warning about things and time to achieve the task
- Discuss and agree priorities and time table
- Forecasting – what will be needed in the future
- Communicate

- Ensuring my plans and objectives for the work area provide clear and specific priorities for others.
- Juggling working pressures to manage time productively across a number of different activities
- Clarifying accountabilities to overcome any confusion over responsibilities across different work activity
- Allocating work in a fair and even handed way to give staff the resources they need to perform at their best
- Responding with speed and urgency to meet the expectations of my customers/clients/colleagues

All this matters. For this reason institutions put in place a plethora of different actions and strive to improve their staff strengthening tights within and across teams.

Additional Resources

Coaching at Imperial

Imperial launched the Coaching Academy in 2009 and we are now able to offer coaching widely as an additional support to development for College staff, either linked to talent programmes or the Imperial Leadership and Management programme, or through individual requests for coaching support.

What is coaching?

Coaching is a 1-1 process that helps individuals think through their options in relation to a range of situations they may be facing in the workplace. Coaching is a confidential and voluntary process, involvingup to 4 meetings over a period of 3-6 months.

A coach uses a combination of observation, questioning, listening and feedback to create a conversation that's rich in insight and learning for the coachee, who will be able to develop a greater awareness and appreciation of their own circumstances. They will also create new ways to resolve issues, and develop skills and strategies so they feel empowered to take action.

Register Interest

To register interest in having a coach, please complete an interest form. You will be asked for further information about the kind of topics you want to explore, which will help with the coach matching process. Your assigned coach will then contact you to arrange an initial meeting to discuss how you can both take the coaching forward. If you have any questions about coaching, contact Judy Barnett, Talent Development Manager

Interested in Having a coach?

Learning Development Centre (LDC) – career development provision

what's available?

Workshops and Coaching

Career Planning Workshop
Career Strategies Workshop
Networking Skills Workshop
Interview Skills Workshop
Coaching Support

READ MORE

Online Resources

Web links
Videos
Online Training
Articles
Apps

READ MORE

Imperial Resources and Opportunities

Work Shadowing
Co-Action
Staff Networks
PDF Guides and Logs

READ MORE

Benefits of coaching:

- Improved sense of direction and focus
- Increased self-awareness
- Improved ability to influence and relate to others
- Increased personal effectiveness
- Increased resourcefulness/resilience such as ability to handle change
- Help them handle the transition into a new role more effectively
- Enhance their ability to lead, engage and develop their team

Personal Review and Development Plan (PRDP)

The College is committed to creating a supportive, inclusive and highly motivated staff community across all disciplines, functions and activities. One of the ways we do this is through the annual Personal Review and Development Plan process (PRDP).

A PRDP is a conversation that focuses on the previous year's work, plans and objectives for the forthcoming year, and includes the preparation of an individual development plan. It is recommended that PRDPs take place on a regular annual cycle, in line with local requirements.

The conversation in a PRDP is broader in focus than one-to-ones. Time should be specifically allocated to cover a review of the previous year, and to discuss and agree the specific aims and objectives for the next year, which should include details of personal and career development and aspirations.

Imperial Expectations

PRDPs support the application of Imperial Expectations. When done well, PRDPs:

- Celebrate achievements
- Enhance performance
- Help staff to develop careers
- Identify individual development plans

Leadership Inventory

Online Resources

Marshall Goldsmith's book *What got you here, won't get you there* includes in it a Leadership Inventory tool. This is a useful reflective exercise to carry out, thinking about your own strengths and areas where you could be more effective.

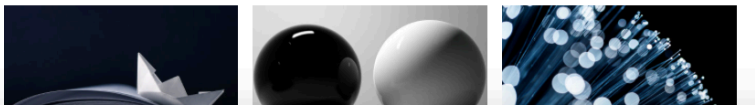
Leadership area	Reflections – What are your strengths? What might you need to do to be more effective?
Thinking Globally 1. Recognises the impact of globalisation on our business 2. Demonstrates the adaptability required to succeed in the global environment. 3. Strives to gain the variety of experiences needed to conduct global business. 4. Makes decisions that incorporate global considerations. 5. Helps others understand the impact of globalisation.	

Imperial College London

1

Postdoc and Fellows Development Centre

The Postdoc and Fellows Development Centre provides tailored support for postdocs and fellows working at Imperial College London



Courses and Workshops

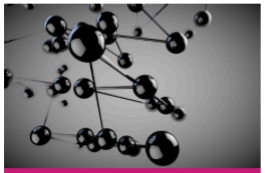
A-Z of all courses offered by the Postdoc Development Centre, including pop-up workshops and Pop-up+ sessions.

Individual Support

One-to-one Support
Mock Interviews

Fellows

Imperial College Research Fellows
Resources for Fellows
Fellow's Forum



Working with Departments

Postdoc Reps Network
Reps Organised Events
PFDC Fund

Fellowships

Early Career Fellowships
Funders Showcase
Imperial College Research Fellowship

Online Resources

Tip sheets
Postdoc Profiles
FAQs
Publications
PFDC Fund
What is a Postdoc
Useful links

Educational Development Unit
About us
Workshops
Programmes
Networks and events
For new lecturers
For postdocs
For research students
Consultancy
Introduction to
A practical guide to
Focus on
Faculty of Medicine
By-request
STAR Framework
Book online
Workshop calendar
TOP LINKS
Perspectives in Education
For new lecturers
Workshops
Imperial STAR Framework
Contact us
With support from the Faculty of Medicine, the EDU is able to offer a workshop strand catered to those Imperial and NHS staff who teach Imperial College undergraduate medical students. The workshops are generally free of charge and normally carry CPD approval from the Royal College of Physicians.
5 of 13

Genome structure & function (Keynote lecture)

Noam Kaplan
Technion, Haifa, IL

Friday, 20:00