LECTURE NOTES

Summer School in Bioinformatics & NGS Data Analysis

Dolný Smokovec, Slovakia, August 14-21, 2016

#NGSchool2016
`https://ngschool.eu/`

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
- Scientific committee: Broňa Brejová, Maciej Łapiński, Marina Marcet-Houben & Tomáš Vinař
- IIMCB Admin, Grant, Finance & PR Units
- Centrum Pod Lesom: arranged accommodation & boarding
- Ewa Ramotowska #NGSchool logo

Supporters This activity is financially supported by grant from International Visegrad Fund (Standard Grant No. 21610099) and International Institute of Molecular and Cell Biology in Warsaw.

Contents

Programme 4
Social activities 4
Lectures & workshops 5
Roman Cheplyaka: Introduction to Linux 5
Broňa Brejová & Tomáš Vinař: Introduction to Bioinformatics & NGS 12
Leszek Pryszcz: Genome & transcriptome assembly 29
Marina Marcet-Houben: Functional genome annotation 43
German Demidov: CNV detection 51
Leszek Pryszcz: Differential expression 73
Maciej Łapiński: ChIP-seq 84
Russell Hamilton: Bisulphite sequencing 96
Jacek Marzec: Molecular data integration 108
Sophia Derdak: NGS & Biomedicine 115
Lectures 124
Paulina Stachula: Library construction for next-generation sequencing: overview and potential troubleshooting 124
Bartek Wilczyński: From Chip-Seq to Hi-C: looking at higher order structure in chromatin 131

Programme

We'll have **morning (9-13)** and **afternoon (14-18)** sessions, with coffee two breaks at 11:00 and 16:00. Breakfast will be served from 8:00, lunch at 13:00 and dinner around 19:00.

	Day 0: Sunday	
18:00	**Welcome & Shot talks #1**	*Leszek Pryszcz*
	Day 1: Monday	
9:00	**Introduction to Linux**	*Roman Cheplyaka*
14:00	**Introduction to Bioinformatics & NGS**	*Broňa Brejová / Tomáš Vinař*
	Day 2: Tuesday	
9:00	**Genome & transcriptome assembly**	*Leszek Pryszcz / German Demidov*
14:00	**Functional genome annotation**	*Marina Marcet-Houben*
20:00	**Shot talks #2**	*Leszek Pryszcz*
	Day 3: Wednesday	
9:00	**CNV detection**	*German Demidov*
14:00	**Differential expression**	*Leszek Pryszcz*
18:00	**Library construction for NGS**	*Paulina Stachula*
	Day 4: Thursday	
9:00	**ChIP-seq**	*Maciej Łapiński*
12:00	**From ChIP-seq to Hi-C**	*Bartek Wilczyński*
14:00	**Bisulphite sequencing**	*Russell Hamilton*
	Day 5: Friday	
9:00	**Molecular data integration**	*Jacek Marzec*
16:00	**Press conference**	*Organising committee*
	Day 6: Saturday	
9:00	**NGS & Biomedicine**	*Sophia Derdak*
	Day 7: Sunday	
9:00	**Recap & farewell**	*Leszek Pryszcz*

Social activities

Social activities will take place during or after the dinner.
We'll go for hiking on Friday or Saturday, depending on the weather.
You can rent bikes from Galfy (`http://www.galfy.sk/sk/pozicovna-bicyklov`) in Horny Smokovec.

Introduction to Linux (Lecture & Workshop)

Roman Cheplyaka
Software engineer, Odessa, Ukraine

Monday, 9:00

What am I running?

```
ngsuser@ubuntu:~$ cat /etc/lsb-release
DISTRIB_ID=Ubuntu
DISTRIB_RELEASE=16.04
DISTRIB_CODENAME=xenial
DISTRIB_DESCRIPTION="Ubuntu 16.04 LTS"

ngsuser@ubuntu:~$ ps -p $$
  PID TTY          TIME CMD
   21 pts/0    00:00:00 bash
```

Where am I?

```
ngsuser@ubuntu:~$ pwd
/home/ngsuser

ngsuser@ubuntu:~$ ls

ngsuser@ubuntu:~$ ls -a
.   .bash_history  .bashrc   .tmux.conf
..  .bash_logout   .profile

ngsuser@ubuntu:~$ ls -al
total 20
drwxr-xr-x 2 ngsuser ngsuser   96 Jul 28 10:11 .
drwxr-xr-x 3 root    root      21 Jul 28 10:05 ..
-rw------- 1 ngsuser ngsuser    5 Jul 28 10:11 .bash_history
-rw-r--r-- 1 ngsuser ngsuser  220 Jul 28 10:05 .bash_logout
-rw-r--r-- 1 ngsuser ngsuser 3771 Jul 28 10:05 .bashrc
-rw-r--r-- 1 ngsuser ngsuser  655 Jul 28 10:05 .profile
-rw-r--r-- 1 root    root    1805 Jul 28 10:05 .tmux.conf
```

What is this command?

```
ngsuser@ubuntu:~$ type cat
cat is /bin/cat

ngsuser@ubuntu:~$ man cat

CAT(1)                    User Commands                    CAT(1)
NAME
       cat - concatenate files and print on the standard output
SYNOPSIS
       cat [OPTION]... [FILE]...
DESCRIPTION
       Concatenate FILE(s) to standard output.
       With no FILE, or when FILE is -, read standard input.
```

Find out what these are:

1. `vim`
2. `cd`
3. `for`

The Linux file system

```
/
|-- bin
|-- boot
|-- dev
|-- etc
|-- home
|   `-- ngsuser
|-- lib
|-- lib64
|-- media
|-- mnt
|-- opt
|-- proc
|-- root
|-- sbin
|-- sys
|-- tmp
|-- usr
`-- var
```

Creating directories

```
ngsuser@ubuntu:~$ mkdir ngschool
ngsuser@ubuntu:~$ mkdir ngschool/day1
ngsuser@ubuntu:~$ mkdir ngschool/day1/lecture1
```

Or:

```
ngsuser@ubuntu:~$ mkdir -p ngschool/day1/lecture1
```

Change directories

```
ngsuser@ubuntu:~$ cd ngschool/day1
ngsuser@ubuntu:~/ngschool/day1$ ls -l
total 0
drwxrwxr-x 2 ngsuser ngsuser 6 Jul 28 13:04 lecture1
```

These are equivalent:

```
ngsuser@ubuntu:~/ngschool/day1$ mkdir ../day2
ngsuser@ubuntu:~/ngschool/day1$ mkdir ~/ngschool/day2
ngsuser@ubuntu:~/ngschool/day1$ mkdir /home/ngsuser/ngschool/day2
```

Shortcuts

Go to ...	Command
Home directory	`cd`
Home directory (alt.)	`cd ~`
Subdirectory under home	`cd ~/ngschool`
Previous directory	`cd -`
Go up one level	`cd ..`
Go up two levels	`cd ../..`

Moving files around

Action	Command
Copy a file	`cp file1 file2`
Copy a file to another directory	`cp file1 ~/ngschool/`
Copy a directory	`cp -r ~/ngschool ~/ngschool2`
Rename a file/directory	`mv file1 file2`
Move a file/directory somewhere	`mv file1 ~/ngschool/`

Running things as root

```
ngsuser@ubuntu:~$ whoami
ngsuser

ngsuser@ubuntu:~$ sudo whoami
root

ngsuser@ubuntu:~$ sudo -i
root@ubuntu:~#
```

Practical: working with FASTQ files

Downloading a file

```
$ wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/\
DRR004/DRR004004/DRR004004.fastq.gz

2016-07-30 14:20:27 (7.75 KB/s) - 'DRR004004.fastq.gz' saved [15438]
```

Useful wget options:

- `-b`: download in background
- `-c`: continue an interrupted download
- `-i file.txt`: read URLs from a text file

Uncompressing a file

Uncompress:

```
$ gunzip DRR004004.fastq.gz
```

- Removes the original compressed file `DRR004004.fastq.gz`
- Creates an uncompressed file named `DRR004004.fastq`
- `16Kb` → `83Kb`
- Better to keep the file compressed

Compress an uncompressed file:

```
$ gzip DRR004004.fastq
```

Working with a compressed file

Some standard commands have analogues that work on gzipped files:

- `zcat`
- `zless`
- `zgrep`
- ... and a few others

To look at the compressed file:

```
$ zless DRR004004.fastq.gz
$ zcat DRR004004.fastq.gz | head
```

Task: count the reads

Every record starts with `@`, so let's count those:

```
$ zgrep -c '^@' DRR004004.fastq.gz
```

- `zgrep` searches for strings in compressed files
  - for uncompressed files, it's simply `grep`
- `-c` means count the occurrences
  - without `-c`, it would print the occurrences
- `^` means only match the beginning of the line
- quotes are to prevent shell interpreting special characters
  - are not necessary in this case (`^` and `@` are not special), but don't hurt

Task: count the reads

```
$ zgrep -c '^@' DRR004004.fastq.gz
```

gives the wrong number of reads. Why?

Exercise: find the lines that break our algorithm

Task: count the reads

Approach #2: rely on the fact that every read occupies 4 lines

1. Count the number of lines
2. Divide it by 4

Read the manpage for the `wc` command to learn how to count lines.

Task: count the reads

```
$ zcat DRR004004.fastq.gz | wc -l
472

$ echo $(( $(zcat DRR004004.fastq.gz | wc -l) / 4 ))
118
```

Task: extract read sequences

```
$ zcat DRR004004.fastq.gz | awk 'NR % 4 == 2 {print}'
```

or simply

```
$ zcat DRR004004.fastq.gz | awk 'NR % 4 == 2'
```

Task: write duplicate reads to the file `dups.txt`

Hint 1: use `uniq`.

Hint 2: you'll also need to use `sort`.

To redirect the output of a command to a file, do:

```
$ command > file.txt
```

Note that this overwrites the previous file contents!

Task: extract reads, replace ACG with ACT

```
$ zcat DRR004004.fastq.gz | awk 'NR % 4 == 2' | \
sed -e 's/ACG/ACT/g'
```

Writes to the standard output; use > to redirect to a file.

Task: find the GC content of all the reads

The GC content is defined as

$$\frac{N_G + N_C}{N} = \frac{N_G + N_C}{N_G + N_C + N_A + N_T}$$

1. Find $N_G + N_C$ and write it to a variable `N_GC`:

   ```
   $ N_GC=$(zcat DRR004004.fastq.gz | \
   awk 'NR % 4 == 2' | grep -o '[GC]' | wc -l)
   ```

   Note: quotes are not optional here!

2. To find N, replace the pattern `[GC]` with a dot (`.`).
   Write the result to a variable `N`.

3. Use the arithmetic expansion (see earlier examples) to compute `$N_GC / $N`.

Task: put each read into its own fastq file

Put the code into a file called `split.sh`:

```
#!/bin/bash
nline=0
zcat DRR004004.fastq.gz | while read line; do
  filename=$(printf read-%.3d.fastq $((nline / 4)))
  printf "%s\n" "$line" >> "$filename"
  nline=$((nline+1))
done
```

Make it executable:

```
$ chmod +x split.sh
```

Run it:

```
$ ./split.sh
```

Task: put each read into its own fastq file

High-level algorithm:

1. Read the fastq file line by line and append the line into the current file
2. Every 4 lines, change the name of the current file

Task: rename each file to its read identifier

E.g. `read-006.fastq` → `DRR004004.7.fastq`

Use a for loop to iterate over files:

```
for file in read-*.fastq; do
  ...
done
```

Inside the loop, use `head` and `grep` to extract the sequence name.

Introduction to Bioinformatics & NGS (Lecture & Workshop)

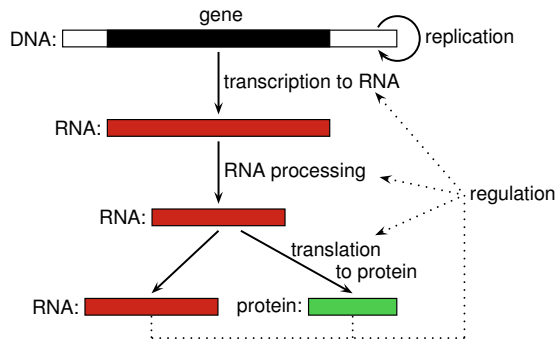
Broňa Brejová & Tomáš Vinař
Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava

Monday, 14:00

Outline

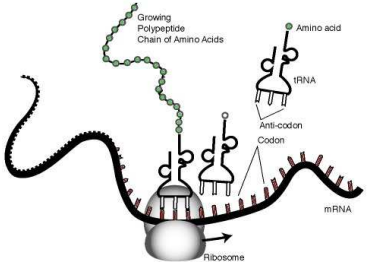
- Two-part lecture with brief introduction to NGS and some bionformatics topics
- **Lecture part (1):** next generation sequencing, genome assembly and genome projects, comparative genomics (Tomáš Vinař)
- **Lecture part (2):** sequence alignment and read mapping, more applications of NGS (Broňa Brejová)
- **Workshop for beginners:** NGS file formats (fastq, sam) and read mapping (BB)
- **More advanced workshop:** NGS and comparative genomics (TV, Matej Lexa)

Central dogma (Francis Crick 1958,1970)

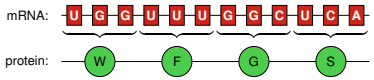


"The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred back from protein to either protein or nucleic acid."

Translation



Codon (triple of nucleotides) determines 1 amino acid



Genetic code

Alanine (A)
GC*

Cysteine (C)
TGC
TGT

Aspartic acid (D)
GAC
GAT

Glutamic acid (E)
GAA
GAG

Phenylalanine (F)
TTC
TTT

Glycine (G)
GG*

Histidine (H)
CAC
CAT

Isoleucine (I)
ATA
ATC
ATT

Lysine (K)
AAA
AAG

Leucine (L)
CT*
TTA
TTG

Methionine (M)
ATG

Asparagine (N)
AAC
AAT

Proline (P)
CC*

Glutamine (Q)
CAA
CAG

Arginine (R)
CG*
AGA
AGG

Serine (S)
TC*
AGT
AGC

Threonine (T)
AC*

Valine (V)
GT*

Tryptophan (W)
TGG

Tyrosine (Y)
TAC
TAT

Stop codon (*)
TAA
TAG
TGA

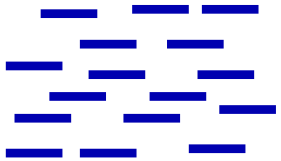
DNA sequencing

Sequencing technologies produce short **reads** from random locations in the DNA sample

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


DNA sequencing

Position of individual reads on the target DNA is not known



Solved by computational methods:

– **mapping** if target DNA is known

– **assembly** if it is not known

Overview of Sequencing Technologies

Technology	Read length	Errors	Output per day
1st generation			
Sanger	up to 900bp	$< 2\%$	3 MB
2nd (next) generation (cca 2004)			
454	400bp	$< 2\%$	400 MB
Illumina MiSeq	150bp	$< 1\%$	2 GB
Illumina HiSeq	150bp	$< 0.5\%$	85 GB
Ion Torrent	200bp	$< 2\%$	10 GB
3rd generation (now emerging)			
PacBio	up to 14kbp	15%	
Oxford Nanopore	up to 100kbp	30%	

Genome Sequencing Overview

1976	MS2 (RNA virus) 40 kB
1988	**Human genome sequencing project (15 years)**
1995	bacterium H. influenzae 2 MB, shotgun (TIGR)
1996	S. cerevisiae 10 MB, BAC-by-BAC (Belgium, UK)
1998	C. elegans 100 MB, BAC-by-BAC (Wellcome Trust)
1998	**Celera: human genome in three years!**
2000	D. melanogaster 180 MB, shotgun (Celera, Berkeley)
2001	**2x human genome 3 GB (NIH, Celera)**
after 2001	mouse, rat, chicken, chimpanzee, dog,. . .
2007	**Genomes of Watson and Venter (454)**

Revolution 1: Shotgun sequencing

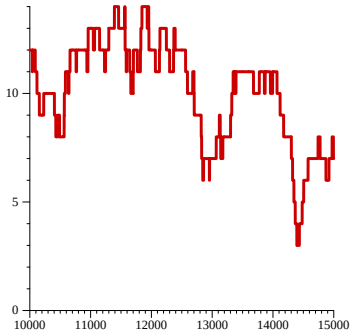
Shotgun sequencing:

- Sequence the whole genome
- Trust sequence assembler (provided we have high-enough coverage)

BAC-by-BAC:

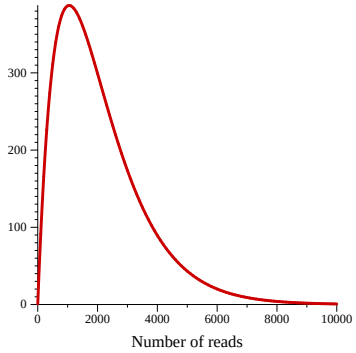
- Create BACs (approx. 100 kb)
- Genome mapping / select BACs that cover the whole genome with small overlaps
- Sequence BACs one by one

Why is coverage important?



Coverage of individual bases at $10\times$ genome coverage with reads of length 1,000.

Why is coverage important?



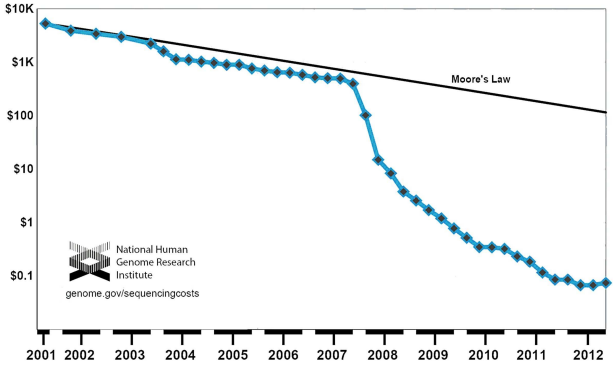
Expected number of contigs when genome of size 1,000,000 is covered by increasing number of segments of length 1,000

Genome Sequencing Overview

1976	MS2 (RNA virus) 40 kB
1988	**Human genome sequencing project (15 years)**
1995	bacterium H. influenzae 2 MB, shotgun (TIGR)
1996	S. cerevisiae 10 MB, BAC-by-BAC (Belgium, UK)
1998	C. elegans 100 MB, BAC-by-BAC (Wellcome Trust)
1998	**Celera: human genome in three years!**
2000	D. melanogaster 180 MB, shotgun (Celera, Berkeley)
2001	**2x human genome 3 GB (NIH, Celera)**
after 2001	mouse, rat, chicken, chimpanzee, dog,. . .
2007	**Genomes of Watson and Venter (454)**

Revolution 2: Next-generation sequencing

Cost per raw megabase of DNA sequence



Until 2007: Sanger sequencing

Starting in 2008: next-generation (454, Illumina, SOLiD)

Typical Results of NGS Assembly

- Many **short contigs** that can be further combined to **longer scaffolds** by using **paired reads**
- Some portions cannot be resolved due to **long repetitive sequences**

Example: Human chromosome 14, 88 Mbp, $70\times$ coverage
(source: GAGE)

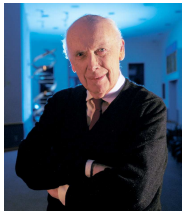
Method	Contigs	Errors	N50 corr
Velvet (basic de Bruijn)	>45000	4910	2.1 kbp
Velvet (with scaffolding)	3565	9156	27 kbp
AllPaths-LG	225	45	4.7 Mbp

N50: reads with this length or longer contain 50% of the genome
here N50 after error correction is shown

NGS pioneers

(first uses of NGS for mapping rather than sequencing novel genomes)

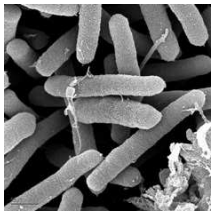
Wheeler, D. A. et al. (Nature 2008) Baylor College of Medicine
The complete genome of an individual by massively parallel DNA sequencing
– genome of James Watson
– sequenced by 454



NGS pioneers

(de novo assembly of a small genome is possible from hybrid data)

Reinhardt J.A. et al. (Genome Research 2009)
De novo assembly using low-coverage short read sequence data from the rice pathogen Pseudomonas syringae pv. oryzae
– genome size 6Mb, N50 scaffold size 500kb
– Illumina 26x coverage, some paired and unpaired 454 reads



NGS pioneers

(first Illumina-only vertebrate-size genome)

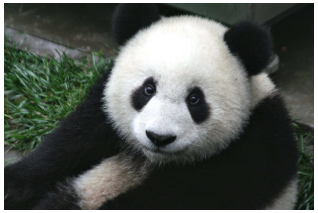
Li R. et al. (Nature 2010)

The sequence and de novo assembly of the giant panda genome

– genome size 2.3Gb, N50 contig size 40kb

– paired end Illumina 56x coverage

– different paired libraries up to 1kb, read length 52bp



For comparison: Sanger sequencing

Lindblad-Toh K. et al. (Nature 2005)

Genome sequence, comparative analysis and haplotype structure of the domestic dog

– genome size 2.4Gb, N50 contig size 180kb

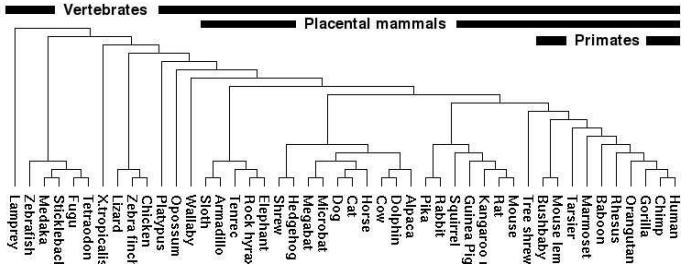
– Sanger 7.5x coverage

– paired reads from libraries 4kb-200kb

A typical modern genome paper

- One or more genomes finished to high quality (chromosomes)
- Additional genomes sequenced to draft quality (scaffolds)
- or: 10s of individuals from various population groups to map population diversity
- Basic genome description / annotation
- Advanced analyses (comparative genomics, population genetics, . . .)
- Biological function (individual genes or gene families responsible for specific traits)

Comparative genomics



Nothing in biology makes sense except in the light of evolution

(Theodosius Dobzhansky, 1973)

Vol 450 | 8 November 2007 | doi:10.1038/nature06340 nature

ARTICLES

Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures

Alexander Stark[1,2]*, Michael F. Lin[1]*, Pouya Kheradpour[1]*, Jakob S. Pedersen[1,3]*, Leopold Parts[1,4], Joseph W. Carlson[1], Madeline A. Crosby[5], Matthew D. Rasmussen[1], Sushmita Roy[6], Ameya N. Deoras[2], J. Graham Ruby[1,7,8], Julius Brennecke[1], Harvard FlyBase curators†, Berkeley Drosophila Genome Project†,

American Genetic Association

Journal of Heredity 2009:100(6):659–674
doi:10.1093/jhered/esp086
Advance Access publication November 5, 2009

Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10 000 Vertebrate Species

Genome 10K Community of Scientists*

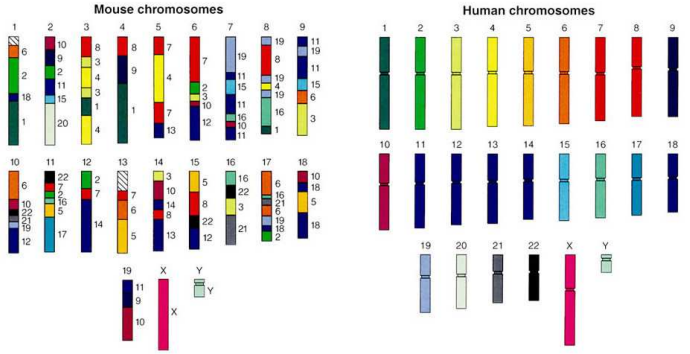
The 1KP Project

The 1000 plants (oneKP or 1KP) initiative is a public-private partnership generating large scale gene sequence information for 1000 different species of plants. Major

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.

Human-mouse comparison



Whole-genome alignments

For each section of reference genome (e.g. human)

find corresponding sections of other genomes.

```
        Human AGTGGCTGCCAGGCTG---GGATGCTGAGGCCTTGTTTGCAGGGAGGT
       Rhesus AGTGGCTGCCAGGCTG---GGTTGCTGAGGCCTTGTTTGCCGGGAGGT
       Mouse GGTGGCTGCCGGGCTG---GGTGGCTGAGGCCTTGTTGGTGGGGTGGT
         Dog AGTGGCTGCCCGGCTG---GGTGGCTGAGGCCTTATTTGCAGGGAGGT
       Horse GATGGCTGCCGGGCTG---GGCTGCCGAGGCCTTGTTCGTGGGGAGGT
   Armadillo AGTGGCTGCCGGGCTG---GGAGGCCAAGGCCTTGTTCGCGGGCAGGT
     Chicken AGTGGCTGCCAGTCTGCGCCGTGGCCGACGTCTTGCTCGGGGGAAGGT
X. tropicalis AATGGCTTCCATTTTGTGCCGCTGCTGAGGTCTTGTTCTGGGGAAGAT
```

Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes

W. James Kent*†, Robert Baertsch*, Angie Hinrichs*, Webb Miller‡, and David Haussler§

*Center for Biomolecular Science and Engineering and §Howard Hughes Medical Institute, Department of Computer Science, University of California, Santa Cruz, CA 95064; and ‡Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA 16802

Mutation rates vary depending on function

- Neutrally evolving sequence
- **Purification selection** (typical for functional element)
  ⇒ slower than neutral mutation rates (higher similarity)
- **Positive selection** (typical in regions responsible for novel functions)
  ⇒ faster than neutral mutation rates (greater divergence)

Human Accelerated Regions

We are looking for regions, which:

- evolved slowly for a long time (purification selection)
- in Humans evolved surprisingly fast (positive selection)


OPEN ACCESS Freely available online PLoS GENETICS

Forces Shaping the Fastest Evolving Regions in the Human Genome

Katherine S. Pollard[1¤a*], Sofie R. Salama[1,2], Bryan King[1,2], Andrew D. Kern[1], Tim Dreszer[1], Sol Katzman[1,2], Adam Siepel[1¤b], Jakob S. Pedersen[1], Gill Bejerano[1], Robert Baertsch[1], Kate R. Rosenbloom[1], Jim Kent[1], David Haussler[1,2]

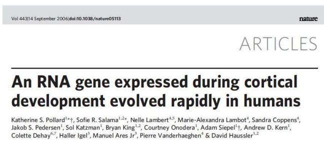
1 Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, California, United States of America, 2 Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, California, United States of America

Human Accelerated Regions: HAR1

Region of length 118

18 differences between human and chimp

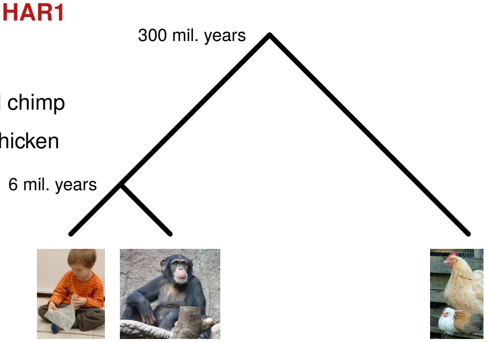
2 differences between chimp and chicken


Vol 443|14 September 2006|doi:10.1038/nature05113 nature

ARTICLES

An RNA gene expressed during cortical development evolved rapidly in humans

Katherine S. Pollard[1*†], Sofie R. Salama[1,2*], Nelle Lambert[4,5], Marie-Alexandra Lambot[4], Sandra Coppens[4], Jakob S. Pedersen[1], Sol Katzman[1], Bryan King[1,2], Courtney Onodera[1], Adam Siepel[1†], Andrew D. Kern[1], Colette Dehay[6,7], Haller Igel[3], Manuel Ares Jr[3], Pierre Vanderhaeghen[4] & David Haussler[1,2]




```
  Human C T G A A A T G A T G G G C G T A G A C G C A C G T C A G C G G C G G A A A T G G T T T C T A T
  Chimp C T G A A A T T A T A G G T G T A G A C A C A T G T C A G C A G T G G A A A T A G T T T C T A T
Gorilla C T G A A A T T A T A G G T G T A G A C A C A T G T C A G C A G T G G A A A T A G T T T C T A T
 Rhesus C T G A A A T T A T A G G T G T A G A C A C A T G T C A G C A G T G G A A A T A G T T T C T A T
  Mouse C T G A A A T T A T A G G T G T A G A C A C A T G T C A G C C G T G G A A A T G G T T T C T A T
    Cow C T G A A A T T A T A G G T G T A G A C A C A T G T C A G C A G T G G A A A C C G T T T C T A T
    Dog C T G A A A T T A T A G G T G T A G A C A C A T G T C A G C G G T G C A A A C A G T T T C T A T
Chicken C T G A A A T T A T A G G T G T A G A C A C A T G T C A G C A G T A G A A A C A G T T T C T A T
```

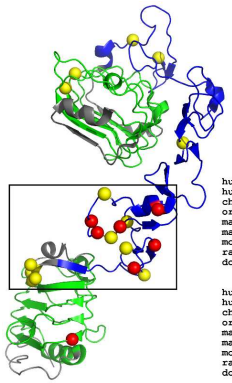

Whole-genome studies: positive selection in protein coding genes

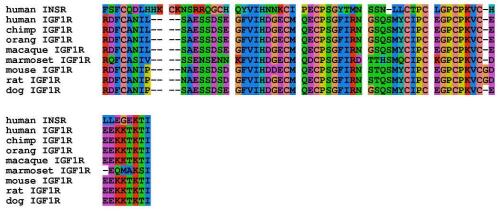
Looking at patterns of mutation in protein coding genes:

- **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

Summary

- Next-generation sequencing is a powerful tool for many applications
- Nowadays, it is “easy” to sequence genomes (draft)
- More difficult to obtain high quality genomes (assembled to the level of chromosomes) and annotate / analyze data
- Comparative genomics benefits from large availability of genomes (even draft quality is often useable if we have a good reference)
- Whole-genome comparative studies can sometimes yield unexpected results

Introduction to Bioinformatics and Next Generation Sequencing Data Part (2)

Broňa Brejová
Faculty of Mathematics, Physics and Informatics
Comenius University in Bratislava, Slovakia

Recall from part (1)

- NGS technologies produce large number of reads from a given DNA sample
- Some technologies produce short reads with low error rates
  others long reads with high error rates
- Reads can be used to assemble genomes
- In comparative genomics we compare groups of related genomes

Part (2)

- Sequence alignment and read mapping
- More applications of NGS

Sequence alignment, homology search

- Given two sequences/sequence databases,
  find regions of similarity between them

Input: two sequences

```
ggcccttggagttgactgtcctgctgctcctt  tgatgccgaggatgtgttcgtcgagcatccgg
gaggccattctcagagagaggaagtggcctca  acgagaagtccatcacctacgtggtcacctac
ttttaatccgcttcccacagccttgtcctttc  tatcactactttagcaaactcaagcaggagac
cagacccatgggagagggaggggctgagggtg  ggtgcagggcataagcgtatcggtaaggtggt
tggctgagcccacccaagtcacgcgtcactct  cggcattgccatggagaacgacaaaatggtcc
gcaggtccctctcccccaaggccgtggccttg  acgactacgagaacttcacaagcgatctgctc
ggagcccgtggatcccagtgagtgacgcctcc  aagtggatcgaaacgaccatccagtcgctggg
acccccccgccctactcgggcagtttaacccttt  cgagcgggagttcgaaaactcgctggccggcg
gttgttcacttgcagacatcgtgaacacggcc  tccaagggcagttggcccagttctccaactac
cggccccgacgagaaggccataatgacctatgt  cgcaccatcgagaagccgcccaagtttgtgga
gtccagcttctaccatgcctttcaggagcgc  aaagggcaacctcgaggtgctccttttcaccc
agaaggtaccgagcagggccaggcaggccctc  tgcagtccaagatgcgggccaacaaccagaag
ctcgccgccaccgcgcaatgccgccgctgcct  ccctacacacccaaagagggcaagatgatttc
ctcgcctcccgtgctcacctcatttctcttgc  ggacatcaacaaggcctgggagcgtctggaga
agacggcagtggcctctctccaactggaagcc  aggccgagcacgaacgcgaattggccctgcgc
acccccagctccct...                 gaggagctcatccg...
```


Output: similar regions in the form of alignments

```
ggcccttggagttgactgtcctgctgctcctt  tgatgccgaggatgtgttcgtcgagcatccgg
gaggccattctcagagagaggaagtggcctca  acgagaagtccatcacctacgtggtcacctac
ttttaatccgcttcccacagccttgtcctttc  tatcactactttagcaaactcaagcaggagac
cagacccatgggagagggaggggctgagggtg  ggtgcagggcataagcgtatcggtaaggtggt
tggctgagcccacccaagtcacgcgtcactct  cggcattgccatggagaacgacaaaatggtcc
gcaggtccctctcccccaaggccgtggccttg  acgactacgagaacttcacaagcgatctgctc
ggagcccgtggatcccagtgagtgacgcctcc  aagtggatcgaaacgaccatccagtcgctggg
acccccccgccctactcgggcagtttaacccttt  cgagcgggagttcgaaaactcgctggccggcg
gttgttcacttgcagacatcgtgaacacggcc  tccaagggcagttggcccagttctccaactac
cggcccgacgagaaggccataatgacctatgt  cgcaccatcgagaagccgcccaagtttgtgga
gtccagcttctaccatgcctttcaggagcgc  aaagggcaacctcgaggtgctccttttcaccc
agaaggtaccgagcagggccaggcaggccctc  tgcagtccaagatgcgggccaacaaccagaag
ctcgccgccaccgcgcaatgccgccgctgcct  ccctacacacccaaagagggcaagatgatttc
ctcgcctcccgtgctcacctcatttctcttgc  ggacatcaacaaggcctgggagcgtctggaga
agacggcagtggcctctctccaactggaagcc  aggccgagcacgaacgcgaattggccctgcgc
acccccagctccct...                 gaggagctcatccg...

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


Sequence alignment, homology search

- Given two sequences/sequence databases,
  find regions of similarity between them
- Display in the form of an **alignment**

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

Insert dashes (gaps) so that corresponding bases in the same column.

A good alignment has many aligned matching bases, few gaps.

What are alignments good for?

- **Read mapping:**
  From which part of the genome is a given read?
  Must be fast, mapping many reads.
- **Determine function (e.g. of a protein):**
  Similar sequences often have the same or similar function.
- **Comparative genomics/evolution:**
  Search for **homologs**, sequences which have evolved from the same common ancestor.
  Ideally, gaps correspond to insertions and deletions,
  aligned bases to conserved bases and substitutions.

Many types of alignment: local/global/...

What task we want to solve?

- **Global alignment:** Align two sequences across their whole length
  e.g. two homologous proteins
- **Local alignment:** Find similarities between shorter regions of long sequences
  E.g.: find homologous genomes between two related genomes
- **Read mapping:** Align entire read to a short region of the reference genome
- **Whole-genome alignment:** Select representative subset from all local alignments between two genomes

Many types of alignment: pairwise vs multiple

- **Pairwise alignment:** create alignments with two rows

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

- **Multiple alignment:** align more than 2 sequences
  Usually global or whole-genome alignments

```
   Human  ctccatagcaatgt-cagagatagggcagagcggat------ggtggtgac
  Rhesus  ctccatggcaatgt-cagagatagggcagagcggat------gctggtgac
   Mouse  ttt--tgacaaca--tagagac-tgagatagaaaat-------atgctgac
     Dog  -tccccgctaatgtacaaagatggggcag-gaaga--a----tgtgctgaa
   Horse  -tccacggcaatac-tggagatggggcagagcaga--agat-ggtgatgaa
Armadillo ctgcatagaaatct-cagagatgggggaaagcaga------agacattcat
 Opossum  atccatggaaacat-cagaagtgggagaaatagaaga----tggcaatga-
Platypus  acccggggaagggg-aagaggaagggccggccg-----------------
```


Speed of alignment algorithms

- **Exact algorithms** find the best alignment possible
- **Pairwise:** dynamic programming algorithms.
  Running time grows with the product of the sequence lengths.
  Practical only for moderately long sequences (e.g. proteins, RNAs)
- **Multiple:**
  Running time grows exponentially with the number of sequences
  Practical only for few very short sequences

Speed of alignment algorithms

- **Exact algorithms** find the best alignment possible,
  but are often too slow in practice
- **Heuristic algorithms** may miss some alignments,
  but have more practical running time
- **Main trick: alignment seeds**
  BLAST starts by finding all exact matches of length 11 between input sequences; these are called hits
  Hits can be found fast (e.g. by hashing, BWT, . . .)
  Each hit is extended to a full alignment by slower methods

Speed vs. sensitivity

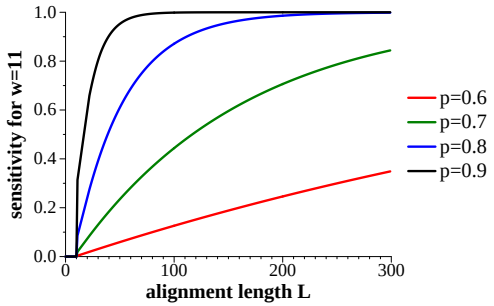
- **Alignment seed:** e.g. $w = 11$ consecutive matches
  Alignments without a seed are not found
  Sensitivity of a tool: what portion of real alignments are found

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

- **Increasing w:** fewer hits, faster program, less sensitive
  Decreasing w: more hits, slower program, more sensitive
- More complex alignment seeding used in many tools

Example: sensitivity in random alignments

Use $w = 11$, random alignments with probability p of match
(human-mouse: $p \approx 0.7$)
Sensitivity: probability that alignment contains 11 consecutive matches



Examples of popular alignment tools

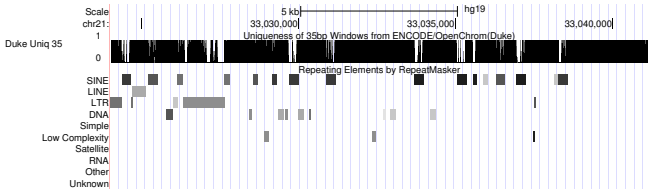
- Water: exact pairwise local alignment
- Needle: exact pairwise global alignment
- FASTA, BLAST, BLAT, Lastz, Last, . . . : heuristic pairwise local alignment
- Clustal, muscle, mafft, t-coffee, . . . : heuristic multiple global alignment
- TBA/multiz, Pecan/Enredo, . . . : multiple whole-genome alignment
- Bowtie, BWA, SHRiMP, SOAP, Last, . . . : heuristic read mapping
- New tools published all the time!

More on read mapping

- Different tools/settings depending on technology:
  for short low-error reads (e.g. Illumina) can use longer seeds
  for longer high-error reads (e.g. PacBio) needs shorter seeds, more similar to genome alignment
- Unified format for storing alignments: sam/bam
- Specialized tools for some applications (e.g. split mapping for structural variants and RNA-seq)
- What if a read has multiple matches?

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)



Example: Details of Bowtie mapper

B. Langmead et al. (Genome Biology 2009)
Ultrafast and memory-efficient alignment of short DNA sequences to the human genome

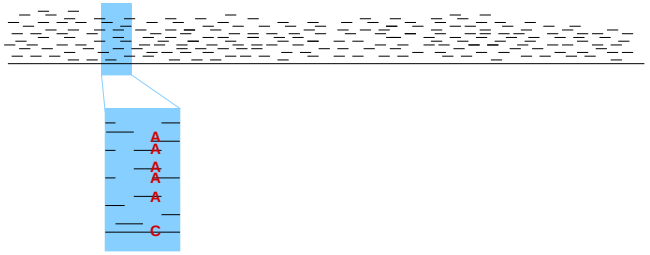
- builds index of reference in memory
- then aligns reads one after another
- optimized for high speed, low memory, but fewer features
- aligns 25 million 35-bp reads per hour
- 2.2 GB of memory for the human genome (2.9 GB for paired-end)

Indexing based on Burrows-Wheeler transform (BWT, 1994)
and Ferragina-Manzini index (FM index, 2000)

Genome resequencing, population genomics

Explore differences between individuals by sequencing multiple genomes from the same species

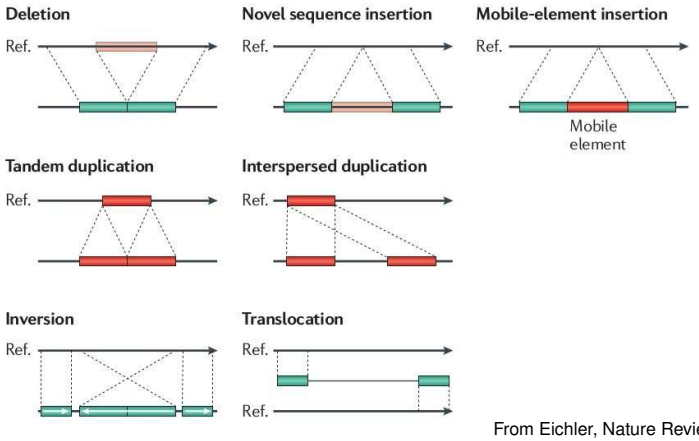
- Discover single nucleotide polymorphisms (SNPs)
  – map reads to reference, identify differences



- Discover structural variants
  – bigger differences, but harder to find

Discovering structural variants

Inconsistent read pairs, split alignments, coverage variation



From Eichler, Nature Reviews 2011

Genome resequencing, population genomics

Explore differences between individuals by sequencing multiple genomes from the same species

Association studies:

- Correlate SNPs with diseases/other traits
- Find causal variants

Population history:

- Study ancient population sizes, migrations, domestication etc.
- Purely from present day individuals
  or ancient DNA (up to hundreds thousands year old)

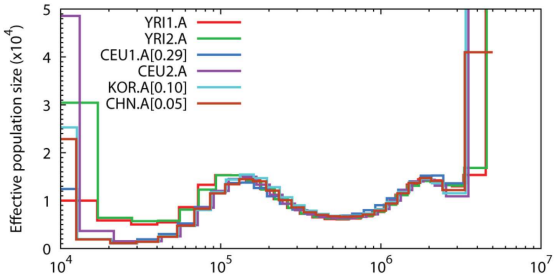
Example: effective population size of human populations

Li and Durbin (Nature 2011) Inference of human population history from individual whole-genome sequences

Compare the two copies of a chromosome from the same individual

Locate blocks with different density of heterozygous sites

Estimate time to most recent ancestor; longer times $\Rightarrow$ larger populations



Cancer genomes

Find mutations occurring in cancer cells

– causes of disease

– therapeutic targets

– disease variants (large variability)

Typical sample genetically heterogeneous

– clonal structure explored by single-cell sequencing

Large projects

The Cancer Genome Atlas (TCGA)

Cancer Cell Line Encyclopedia (CCLE)

Metagenomics/environmental genomics

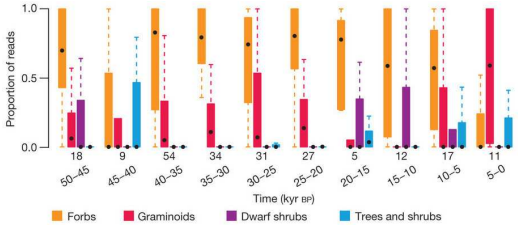
Sequence DNA of uncultured microbial community

– may have more than 10,000 species

Also sequence DNA fragments from environment (e.g. sediments, ice)

Explore diversity by read binning

assign reads to taxonomic groups by similarity to known genomes



From: Willerslev et al. (Nature 2014) Fifty thousand years of Arctic vegetation and megafaunal diet.

Human microbiome

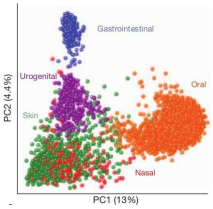
Human body: 10^{13} human cells, 10^{14} bacterial cells

The Human Microbiome Project Consortium (Nature 2012)

Structure, function and diversity of the healthy human microbiome

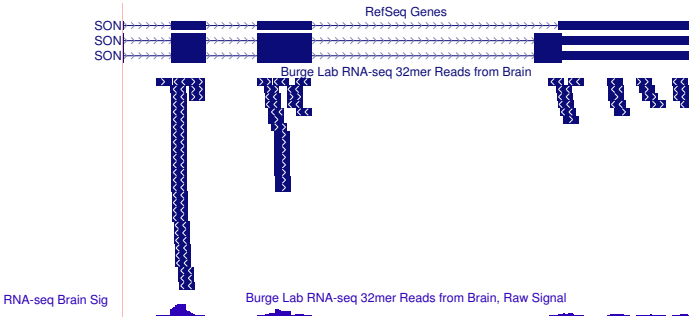
242 people, 15-18 sites per person, some repeat visits

Large variation within and between individuals



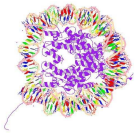
RNA-seq

- Sequence RNA extracted from a sample, map reads to genome
- Measure expression level of individual genes
- Discover new genes, splicing variants, polymorphisms, RNA editing, transcript fusion in cancer

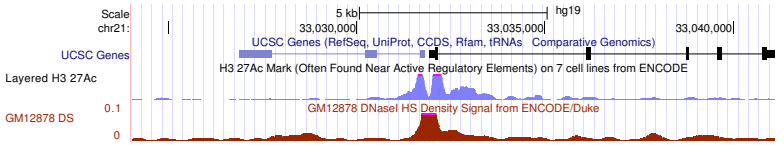


Epigenomics

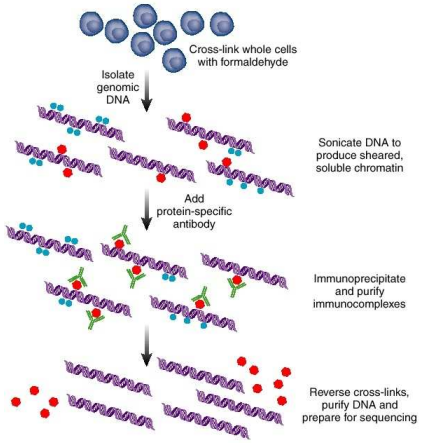
- Expression of genes is regulated by DNA methylation and histone modifications



- ChIP-seq with antibodies specific for individual modifications
- DNase-seq: DNase I Hypersensitivity Site footprinting
  – discovers open chromatin structure (potential regulatory sequences)



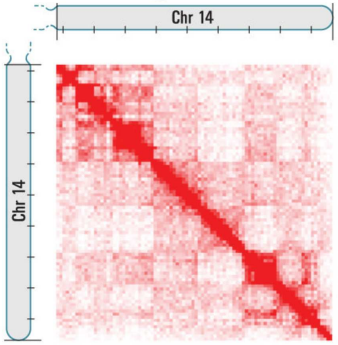
Chromatin immunoprecipitation (ChIP)



From E.R. Mardis (Nature Methods 2007)

Molecule interactions

- ChIP-seq: DNA-protein interactions (transcription factors)
- HITS-seq: RNA-protein interactions
- Hi-C, ChIA-PET: DNA-DNA interactions in the nucleus



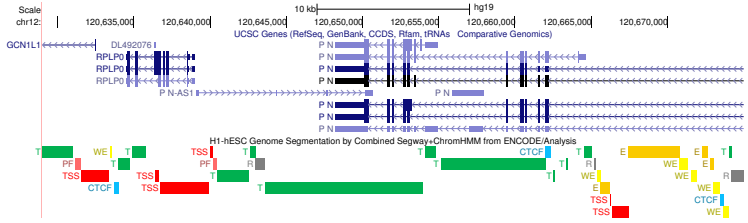
From: Lieberman-Aiden et al. (Science 2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome

ENCODE project (Encyclopedia Of DNA Elements)

Human genome, also some model organisms (mouse, drosophila, worm)

Many NGS technologies, e.g. RNA-seq, ChIP-seq in many tissues

Segmentation of genome to regions based on experimental data



TSS	Bright Red	Predicted promoter region including TSS
PF	Light Red	Predicted promoter flanking region
E	Orange	Predicted enhancer
WE	Yellow	Predicted weak enhancer or open chromatin cis regulatory element
CTCF	Blue	CTCF enriched element
T	Dark Green	Predicted transcribed region
R	Gray	Predicted Repressed or Low Activity region

Summary

- Next-generation sequencing is a powerful tool for many applications, including genome (re)sequencing, transcriptomics, epigenomics, interactions
- NGS data requires sophisticated bioinformatics analysis
- ZOO of tools for read mapping and alignment

Next: workshops

- **Workshop for beginners:** NGS file formats (fastq, sam) and read mapping (BB)
- **More advanced workshop:** NGS and comparative genomics (TV, Matej Lexa)

Common 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
```


Common formats: 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
+
CCCFFFFFHHHHFIGIIJJJJJJIJJJIJHJJJJHIIIJJJJJJIIIIGFEIIIJJJJGJHFHHH
```

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!

Common formats: 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:

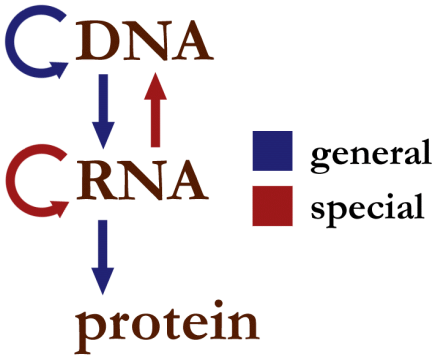
```
HWI-ST1300:156:H7599ADXX:1:1110:17800:29845     145     contig0004
10      50      41M     contig0346      339     0       \
CTATAGATCTTATTACCGCTTATCCAAAGCTGAAAGTGACT       \
IIIIIIIIIIIIIIFIIIIIIIIIIIIIIFFFFFFFFFFBBB      \
AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:41 YT:Z:UU NH:i:1
```

Genome & transcriptome assembly (Lecture & Workshop)

Leszek Pryszcz
International Institute of Molecular and Cell Biology in Warsaw

Tuesday, 9:00

Why do we care?



DNA & evolution of sequencing

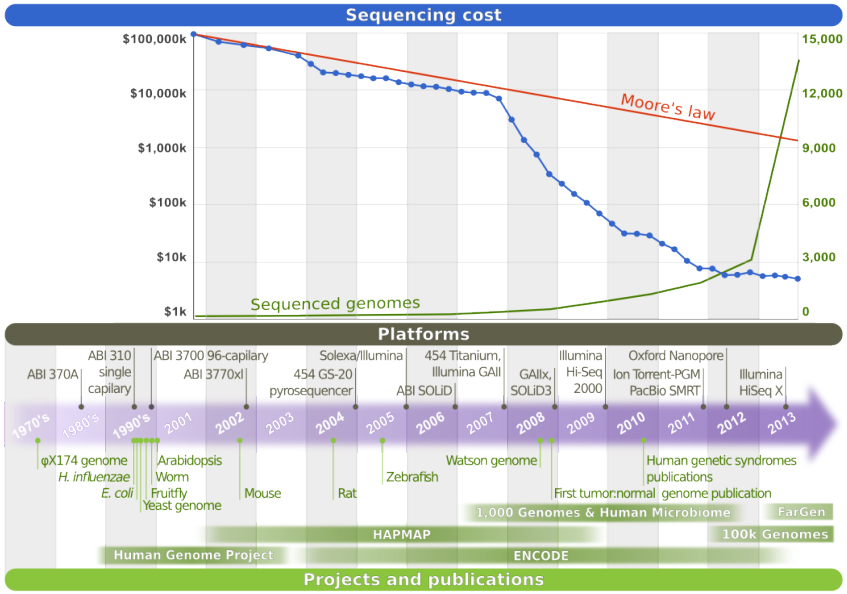
- 1869: DNA discovered by Friedrich Miescher
- 1953: double helix model by Watson, Crick & Franklin
- 1977: rapid DNA sequencing method by Sanger
  - bacteriophage φX174
- 1986: ABI 370A – automated sequencer
- 1995: *Haemophilus influenzae*
- 1996: *Saccharomyces cerevisiae*
- 2001: Human genome draft
- 2012: MinION

wikipedia.org

wikipedia.org

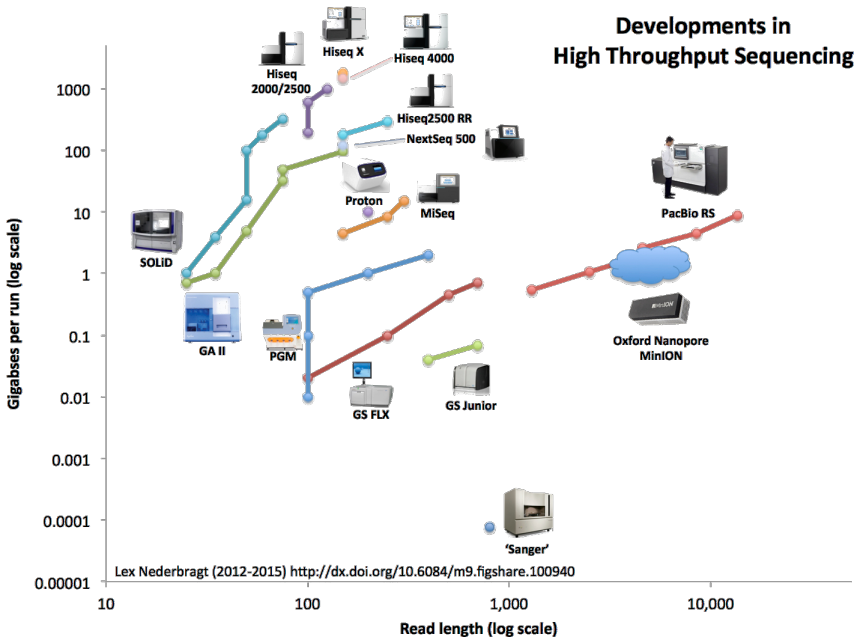
Human Genome Project

- 1990s': $0.75 / base
- Project:
  - Decode human genome
  - Tens of research groups around the world
  - 15 years
  - $3,000,000,000 budget
- 2001: draft published
- 2003: completed
  - Euchromatic regions (~90%)



nature.com

http://bit.ly/lpryszcz_thesis

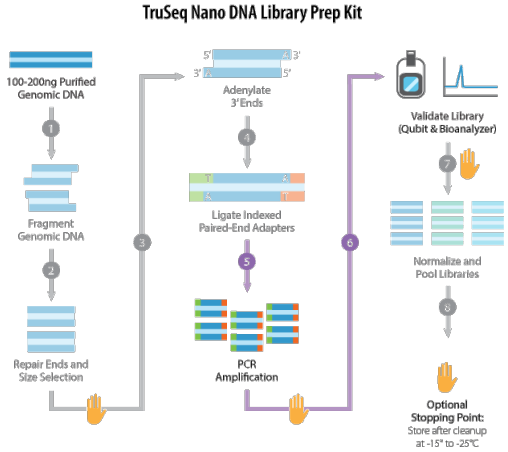


NextSeq 500

- Up to 120 Gb in ~29 h
- 4 lanes
- Multiplexing
- Modes
  - 1x 75 bp
  - 2x 75 bp
  - 2x 150 bp
- DNA-Seq, RNA-Seq, ChiP-Seq ...

illumina.com

Sample preparation



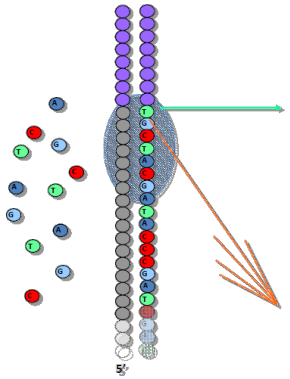
www.abmgood.com

Flowcell

illumina.com

Illumina sequencing

- Cycle 1:
  - Add sequencing reagents
  - First base incorporated
  - Remove unincorporated bases
  - Detect signal
  - Unprotect/remove dye
- Cycle 2-n:
  - Add sequencing reagents...
- All four labeled nucleotides in one reaction
- Base-by-base sequencing
  - No homopolymers



Paired-end, overlapping PE and mate-pairs?

- PE
  - 200-600 bp

- Overlapping PE
  - 150-350 bp

- MP
  - 2-10 kb

Reads

- FASTQ
  - 35bp – 250+ bp; single or paired; FR or RF
  - bases with various error probabilities

```
@HWUSI-EAS1696_0025_FC:3:1:2892:17869/1
CAGCAAGTTGATCTCTCACCCAGAGAGAAGTGTTTCATGCTAAGTGGCAGTTTCTGGTGCAGAACAG
+
IH
@HWUSI-EAS1696_0025_FC:3:1:2892:17869/2
TGGCAGTTTCTGGTGCAGAACAGTTCTGCAATGAGGGAGGAGGCAGAAAACATAAGTGTGTAATAAG
+
GGHIIIHIIIIIIIBBIEDGGFHHEIHGIGEGHEBCHDBFC>CBCCECEEAAAAEEE:B@B@BBB;B@
```

ord('E') → 69
69 – 33 = 36
69 – 64 = 5

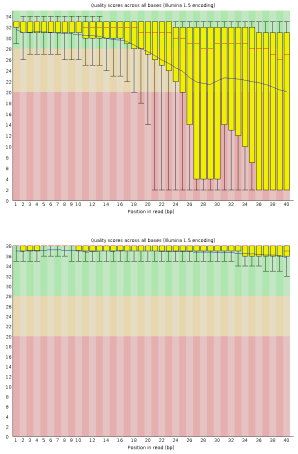
ord('B') → 66
66 – 33 = 33
66 – 64 = 2

- Quality encoding
  - Q phred = -10 * log10($e)
  - ASCII chars
    - offset 33: Sanger / Illumina 1.8+
    - offset 64: Illumina 1.3-1.7

http://en.wikipedia.org/wiki/FASTQ_format

Reads: Quality Control

- Basic Statistics
- Sequence quality & content
- GC content
- N content
- Length Distribution
- Duplication Levels
- Overrepresented sequences
- Kmer Content

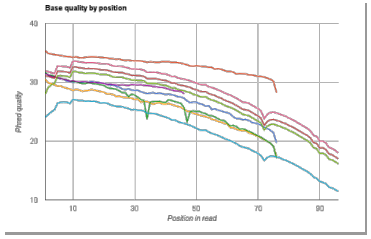


http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Preprocessing

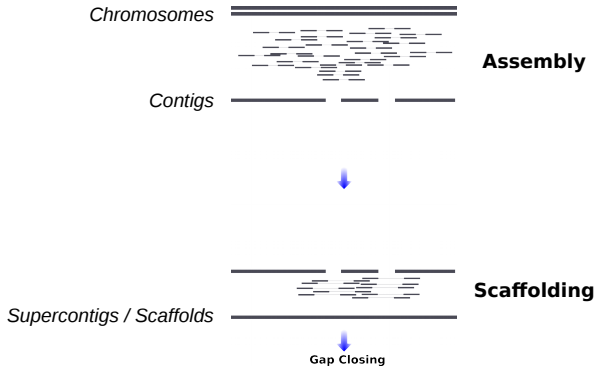
- Assembly affected by N's and mismatches
  - Remove adapters
  - Quality trimming & filtering
- FastQ quality filtering
  - filterReads.py
  - trimmomatic

PHRED quality	Error probability	Base call accuracy
10	10.00%	90.00%
20	1.00%	99.00%
30	0.10%	99.90%
40	0.01%	99.99%



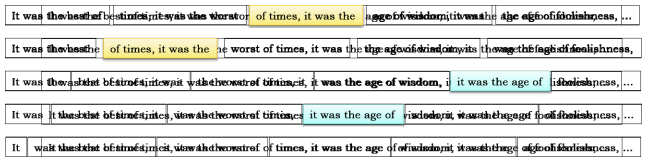
http://en.wikipedia.org/wiki/FASTQ_format

De novo genome assembly



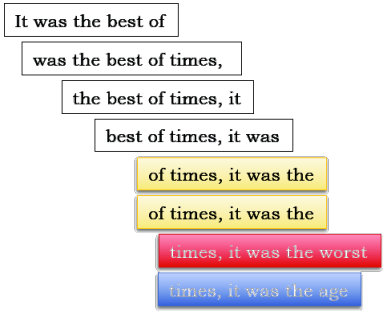
Reconstruction of shredded book

- Book accidentally shredded
  - 10 copies
- How can he reconstruct the text?
  - fragments are mixed
  - some are identical



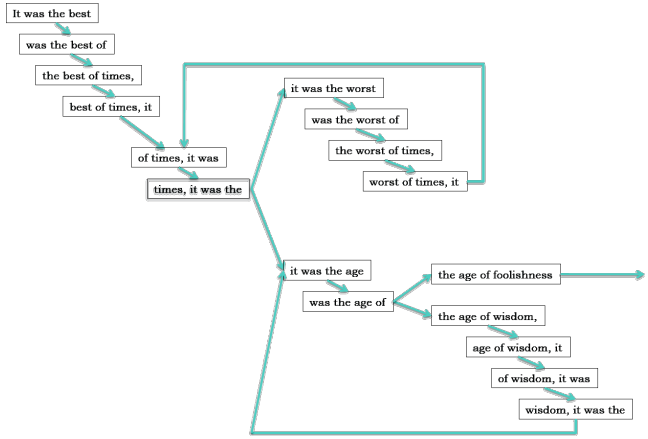
Courtesy of M. Schatz

Greedy algorithm



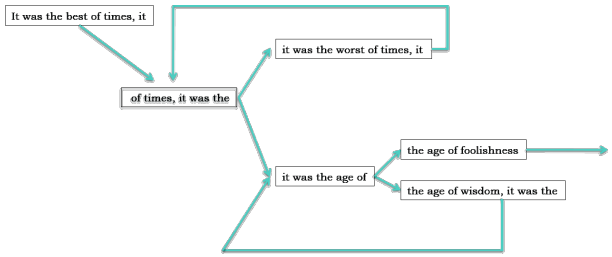
Courtesy of M. Schatz

de Bruijn Graph



Courtesy of M. Schatz

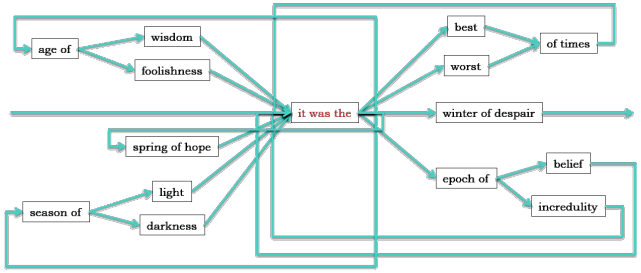
de Bruijn Graph



Courtesy of M. Schatz

de Bruijn Graph

… it was the best of times it was the worst of times …
… it was the age of wisdom it was the age of foolishness …
… it was the epoch of belief it was the epoch of incredulity …
… it was the season of light it was the season of darkness …
… it was the spring of hope it was the winder of despair …

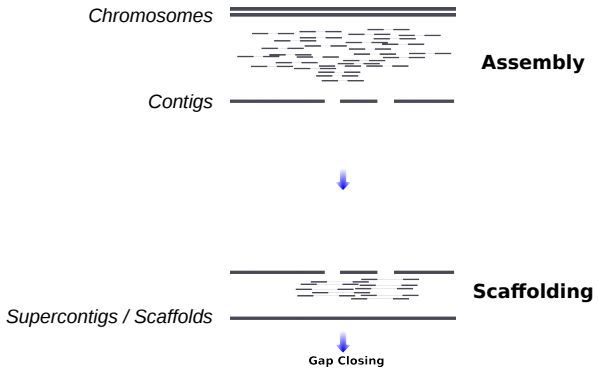


Courtesy of M. Schatz

de Bruijn vs Overlap graph

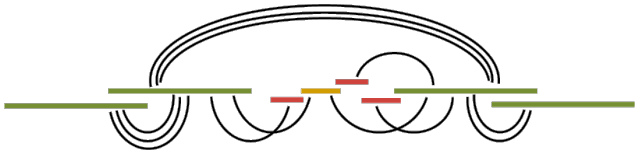
- Short read assemblers
  - Repeats depends on word length
  - Read coherency, placements lost
  - Robust to high coverage
- Long read assemblers
  - Repeats depends on read length
  - Read coherency, placements kept
  - Tangled by high coverage

De novo genome assembly

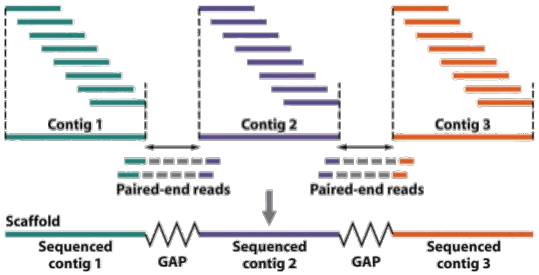


Scaffolding

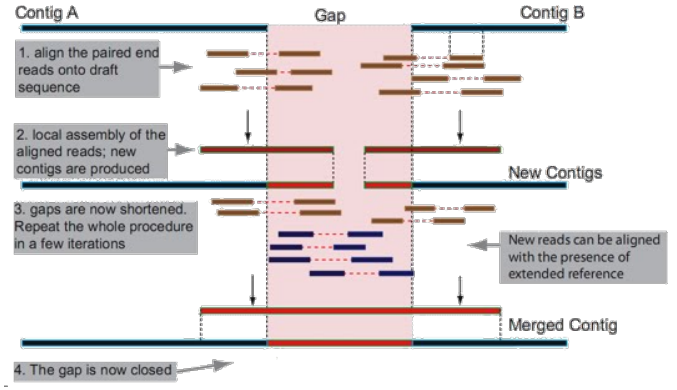
- Resolve longest, most unique contigs
- Assembly breaks
  - Coverage gaps
  - Conflicts



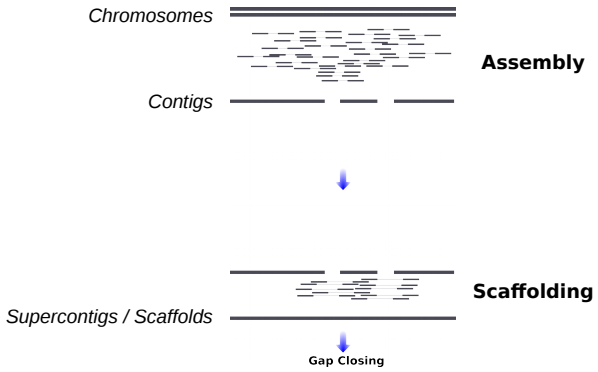
Scaffolding



Gap closing

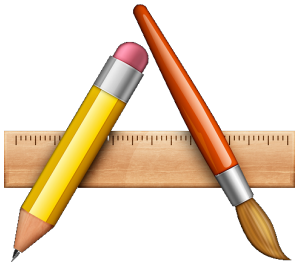


De novo genome assembly



Applications

- MIRA
- Celera
- Velvet
- SOAPdenovo
- Ray
- IDBA
- SPAdes
- ALLPATH-LG
- Platanus
- ...



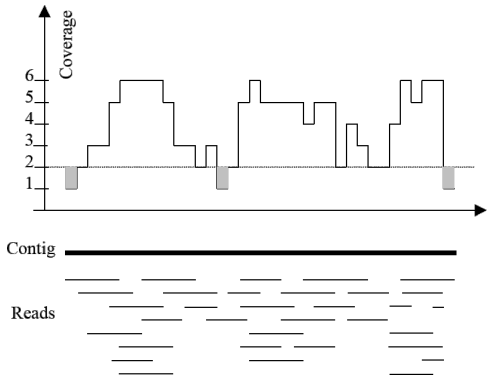
Some genomes are hard to assemble

- Biological
  - high genome size, ploidy, heterozygosity, repeats
- Technological
  - short reads, but good quality (Illumina)
  - long reads, but low quality (PacBio, Nanopore)
- Computational
  - Large & complex genomes, memory constrains
- Accuracy
  - difficult to assess correctness

Good assembly recipe

- Sample
  - haploid or homozygous is better
- Reads
  - longer reads for larger genomes
  - PE & MP libraries with varying insert sizes
  - low error rate
- Coverage
  - high and constant

Coverage vs completeness

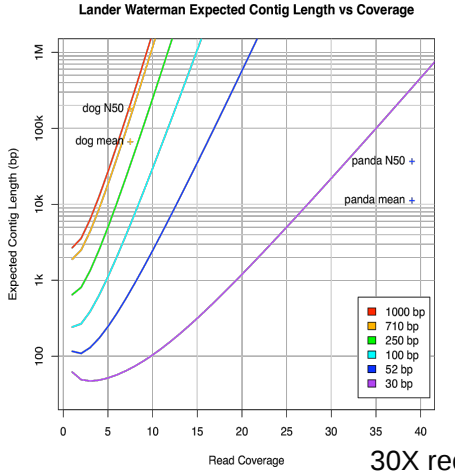


$$P_{(Y=y)} = (C^{y} \times e^{-C}) / y!$$

where: C = coverage; y=no. of times the base is read

$P_{(Y=3)} = (6.3^{3} \times e^{-6.3}) / 3! = 0.077$

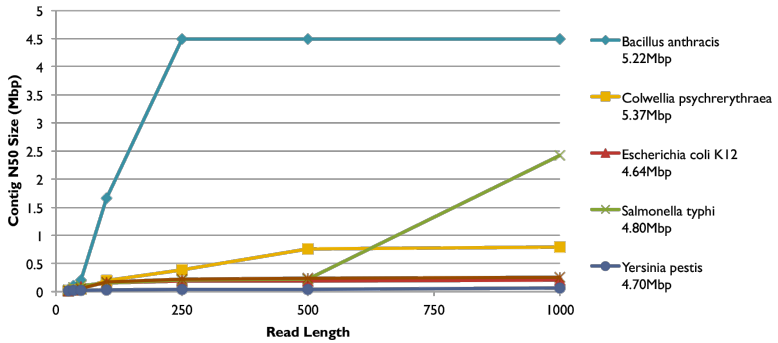
Read length vs contig length



30X recommended

100X for harder genomes

Repeats and read length



Courtesy of M. Schatz

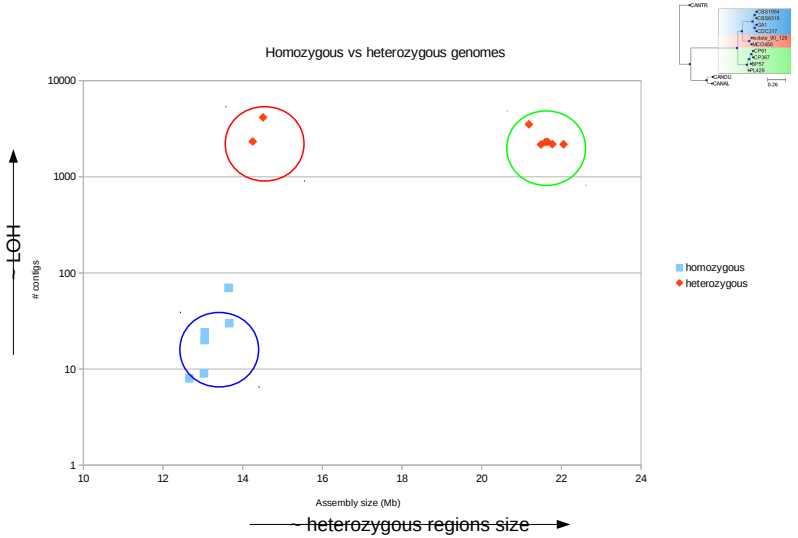
Hybrid sequencing

- Short reads
  - Cheaper
  - High coverage
  - Good quality
  - Correct long reads
- Long reads
  - Expensive
  - Low coverage
  - Poor quality
  - Assembly

Illumina
Sequencing by Synthesis

Pacific Biosciences
SMRT Sequencing

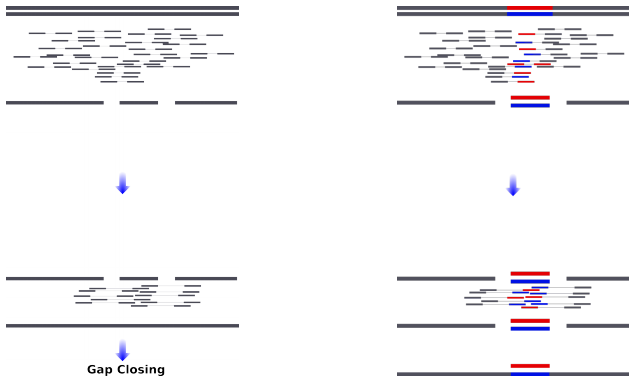
Dealing with heterozygous genomes

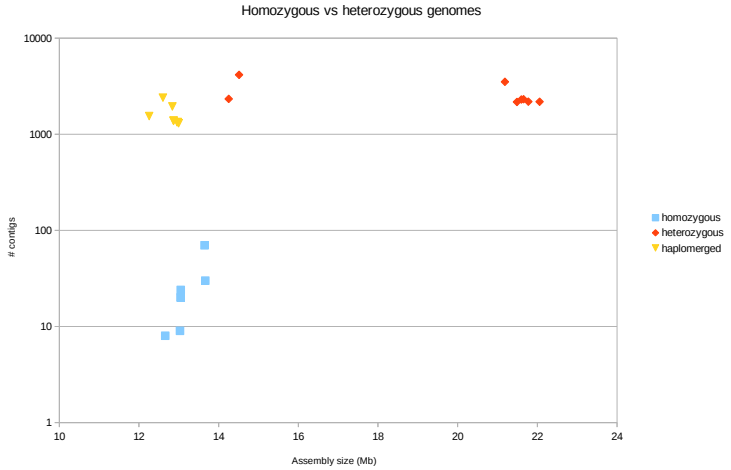


Expected genome: ~13Mb in 7-8 chromosomes

36

Heterozygous genome assembly

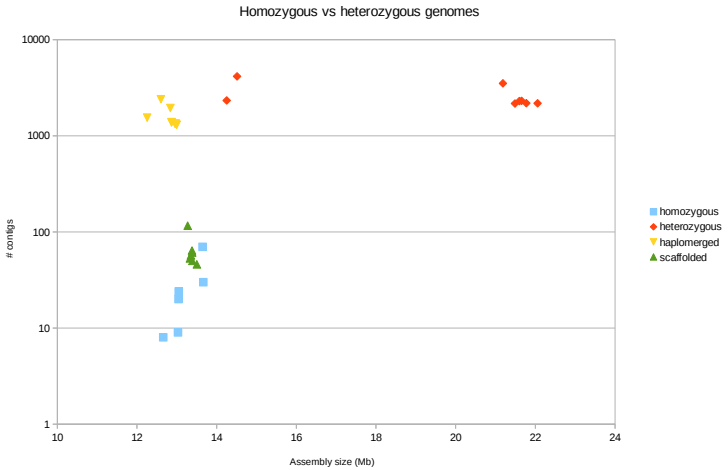




Expected genome: ~13Mb in 7-8 chromosomes

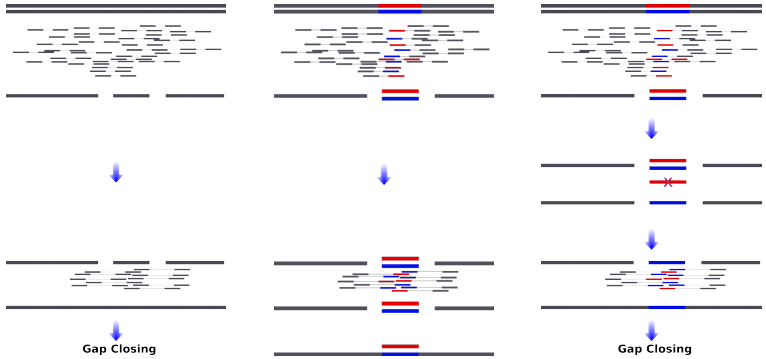
38

Pryszcz & Gabaldon (2016) NAR



Expected genome: ~13Mb in 7-8 chromosomes

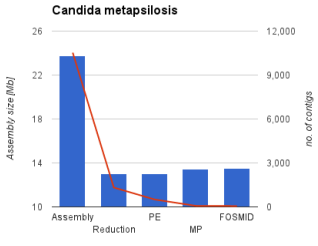
Heterozygous genome assembly



Pryszcz & Gabaldon (2016) NAR

Candida metapsilosis

- 11 strains sequenced
  - Paired-end
  - Mate-pairs
  - Overlapping 2x250bp
  - Fosmids
- *De novo* assembly
  - Fragmented & larger
- Redundans
  - ~ chromosomes

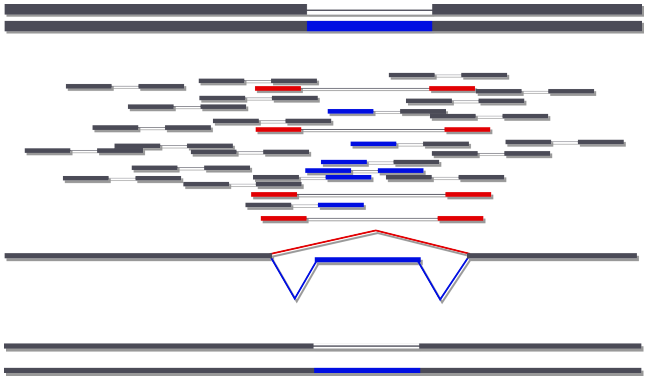


Pryszcz & Gabaldon (2015) Plos Genetics

De novo transcriptome assembly

GENOME ASSEMBLY

TRANSCRIPTOME ASSEMBLY



TRINITY: TRANSCRIPTS ASSEMBLY

- Multiple de Bruijn graphs
  - one per locus
  - processed independently
- Heavy computations
  - Uneven coverage
  - Multiple isoforms

- USAGE:
  - Trinity.pl # call main program
    - --seqType fa # sequence format
    - --JM 10G # virtual memory limit
    - --left reads/q20_1.fasta # left reads file(s)
    - --right reads/q20_2.fasta # right reads file(s)
    - --CPU 2 # number of threads
    - --output assembly # output directory

http://trinityrnaseq.sourceforge.net/

TRANSDECODER: ORF DETECTION

- Find coding regions in transcripts:
  - Min. length
  - Coding potential
  - Hits in PFAM
- Belong to Trinity package

- USAGE:
  - TransDecoder # call main program
    - -t Trinity.fasta # assembled transcripts
    - -G genetic_code # genetic code
    - -S # strand specific
    - --search_pfam Pfam-A.hmm # check for hits vs PFAM

http://trinityrnaseq.sourceforge.net/

Quality assessment

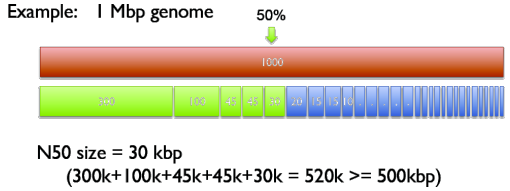
- Assembly statistics
- Whole genome aligment

Assembly quality metrics

- Total size
- Number of contigs
- N50, N90
- Gaps
- Longest contig

Basic statistics

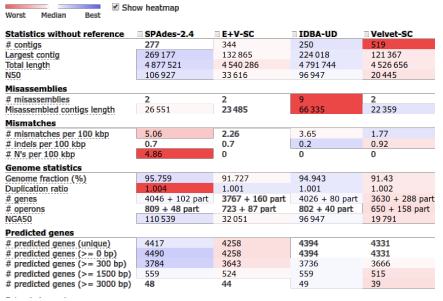
- N50:
  - 50% of the genome in contigs as large as N50 value
  - meaningful only if the same assembly size
- NG50
  - Normalised by genome size
- N90



Courtesy of M. Schatz

Assembly statistics

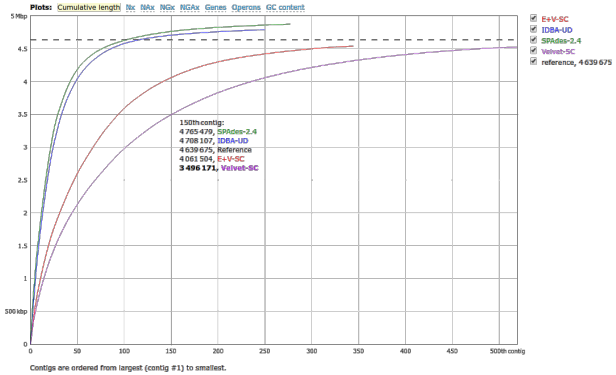
- fragmentation
  - N50
  - no. of contigs
- completeness
  - expected size
  - mappability
  - WGA
- correctness
  - WGA with close species
  - inconsistencies between mates



Statistics without reference	SPAdes-2.4	E+V-SC	IDBA-UD	Velvet-SC
# contigs	277	344	250	519
Largest contig	269 177	132 865	224 018	121 367
Total length	4 877 521	4 540 286	4 791 744	4 526 656
N50	106 927	33 616	96 947	20 445
Misassemblies				
# misassemblies	2	2	9	2
Misassembled contigs length	26 551	23 485	66 335	22 359
Mismatches				
# mismatches per 100 kbp	5.06	2.26	3.65	1.77
# indels per 100 kbp	0.7	0.7	0.2	0.92
# N's per 100 kbp	4.86	0	0	0
Genome statistics				
Genome fraction (%)	95.759	91.727	94.943	91.43
Duplication ratio	1.004	1.001	1.001	1.002
# genes	4046 + 102 part	3767 + 160 part	4026 + 80 part	3630 + 288 part
# operons	809 + 48 part	723 + 87 part	802 + 40 part	650 + 158 part
NGA50	110 539	32 051	96 947	19 791
Predicted genes				
# predicted genes (unique)	4417	4258	4394	4331
# predicted genes (>= 0 bp)	4490	4258	4394	4331
# predicted genes (>= 300 bp)	3784	3643	3736	3666
# predicted genes (>= 1500 bp)	559	524	559	515
# predicted genes (>= 3000 bp)	48	44	49	39

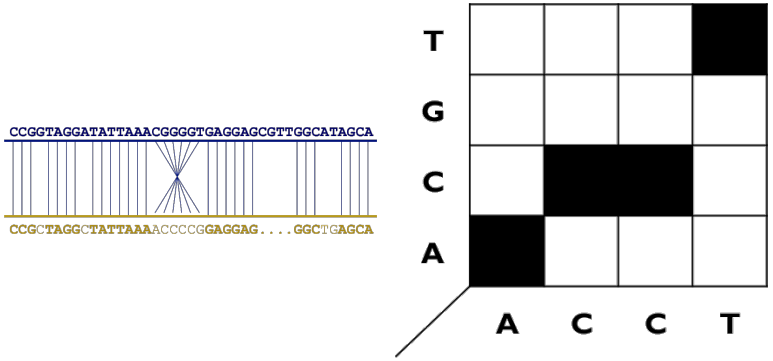
fasta_stats.py
http://quast.bioinf.spbau.ru/

Quast report

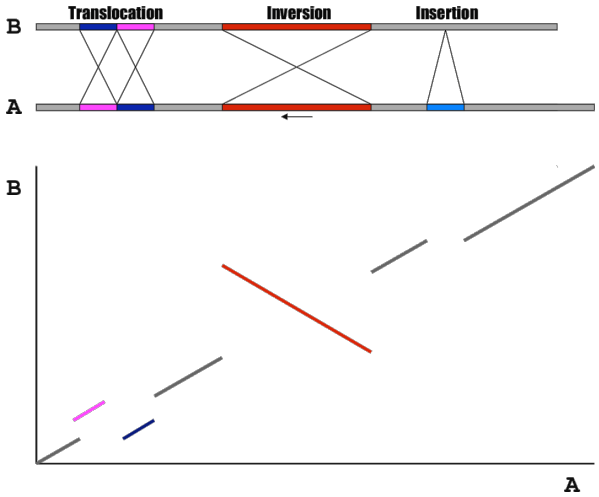


http://quast.bioinf.spbau.ru/

Whole genome alignment



Dotplot




```
ACTAAGGGTTCAACTACTAAATTAATTAAATATTGTAAACCTCTGTCTTTAAGAGTAGGAATCCCTAT
AGGTCTAAGTTTCTCATTTCCTTTTGGTATTCAAACTCTTTTTACCGGATCACTCTTATATAACTTAG
GATGTTTTACTGTTGTTCCTAGTCATTCGACTATTTCCACGAGTAAGGCTCTATCCTTATCTACTCCT
TCAATACATACATTATCCACTCCAGATGAAGGGTTTAATGCATTTTTCTACTCTTTGGTGTCGAAAGA
TACTCAAGAAATGTATTATTCCACAAAGAGACAAGCATTCCTAGTTGACCAAGGTTATGCATTCAAAG
TCATTACACATTTGAGTGGCATGGAGCAATTACCTGACTTGGCGTATGCATCAGCTAGAGAACGTCGA
GAATTGTTGCAGCAAGTGTTGTTGAAGAATGAAGATGCTGCCGGTTTGGAAATTGGTGATGATGTATT
ATCAGTCAAACTCAAATCCTCACCACCATCATCATCCAGTGAACCTTTCGACTTATCACCATCATCAT
CATCATCATCGACACCCCCATTCAGCACCGTATCTTGATCGAATATTTGAGCCCATTGATCATAATAA
TAGATACGAATTTCATTTCCCGAGAGCAAAGACAAAGGATGGACAATAGCAAACAAAGTGGTGCTACT
CGATCGGCTGGTTCGTTGGCTGGGTTGGCTGGTGGTGAAGATATGGCTTATATTGAGTATAGCAGGAA
CAAGAATAAGGAGTTGGGAGCGCATCACCCATTGATTCAAAAGATGTATTACAAGAAGAAGCAACAGC
AACAACAACAACTGCAAGAGCAACAACTGCAGCTGCTGCAACAACATCAAGGACGCAGATAGAAATTC
TATAATTTACAATTGTCATTTACTTACTTAA_THANK_TAAGTCACCAAAATCATCGTCTTCTTCTTC
TTCACTCTCTGAGCTTTGACCAGCTAGTGTTT_YOU_TCTATCGGCTGTTTCATAGTTATTTTCATTA
TCAGTCAAACTCAAATCCTCACCACCATCATCATCCAGTGAACCTTTCGACTTATCACCATCATCATC
ATCATCATCGACACCCCCATTCAGCACCGTATCTTGATCGAATATTTGAGCCCATTGATCATAATAAT
ACTGTTGCTGTAAATGATCATATCTCATGTTATAATTCTCTAAATGCTCAAGAAACTCTTCTGAGTCA
ATATCATGTCTGAAACGTAGCAAGAGGGCAATCTGTGCCACTTTTGTCGCATCATCAAGGTTGTGAAG
ATCGCTACCGCCTCCTGTCCCTACTTCCTCTTCTTCAACACTGTCGGTTATTAGTGGATGGTACCCTG
AATTGGGCAATTGTACAATAACCCGAAACCATTGAGTTGACACAATACCCAAATCACGTCAAATAGGA
AATTAATGTTTTAATATACGGAGGAGGACCTATTGAATAAATCACCTCTGCGATTGAACCATCGATTG
AATCATTTTACGGTAGTCCATCAAGTGTCACATCTCAGTAATCAGGGGTAGAGCATATTGGGACGGCA
ATCCACATGATAAATCAGCGTTGGATGAGTAATAGCCGTATCATGAGCTACTCAACCCGCTCCTGTCC
CTACTTCCTCTTCTTCAACACTGTCGGTTATTAGTGGATGGTACCCTGAATTGGGCAATTGTACAATA
ACCCGAAACCATTGAGTTGACACAATACCCAAATCACGTCAAATAGGAAATTAATGTTTTAATATACG
GAGGAGGACCTATTGAATAAATCACCTCTGCGATTGAACCATCGATTGAATCATTTTACGGTAGTCCA
TCAAGTGTCACATCTCAGTAATCAGGGGTAGAGCATATTGGGACGGCAATCCACATGATAAATCAGCG
TTGGATGAGTAATAGCCGTATCATGAGCTACTCAACCCGGGGACGGCAATCCACATGATAAATCAGCG
```

Functional genome annotation (Lecture & Workshop)

Marina Marcet-Houben
Centre for Genomic Regulation, Barcelona

Tuesday, 14:00

Well, we have a genome, and now what?

Gene prediction

Marina Marcet-Houben
NGSchool 2016
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.

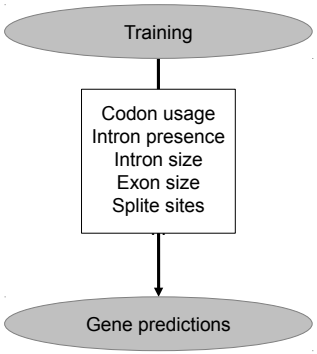
Group of predicted sequences (fasta format) ⟷ Genome

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

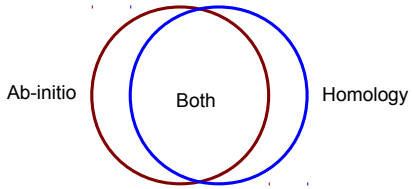
geneid HomePage

DOGMA
Dual Organellar GeneMe 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

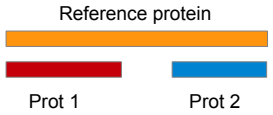
1.- How well assembled is your genome?

2.- Do you have the core proteins?

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

CEGMA
(discontinued)

3.- Do you have many split genes?



Exercise 1.- Local blast search

Exercise 2.- Annotate a prokaryotic genome using on-line tools: glimmer and genemark.

Exercise 3.- Annotate a eukaryotic genome using exonerate

Exercise 4.- Annotate a eukaryotic genome using Augustus

Well, we have a proteome, and now what?

Functional prediction

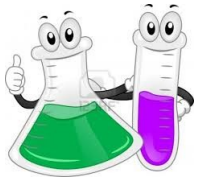
How can we know what our protein does?

1.- Experimental evidence

2.- Homology

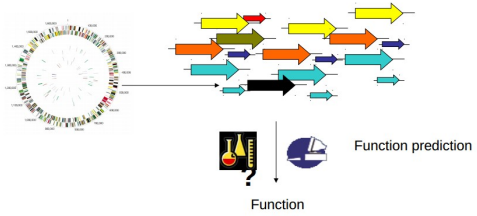
3.- Orthology

4.- Others



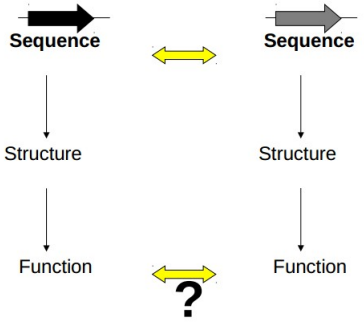
Doing experiments for each element we have identified to discover their function is impossible right now.

Experimental evidence



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

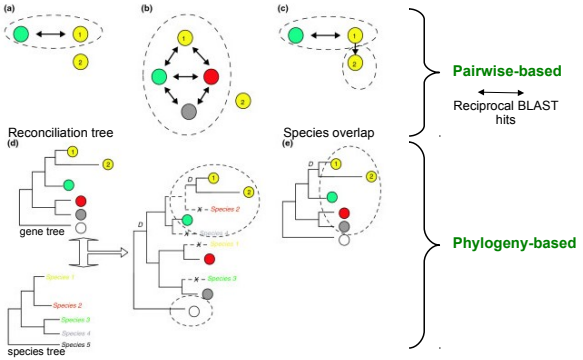
Classic method: function prediction by homology



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

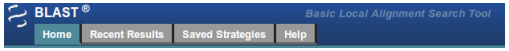
Two proteins can be more similar than two others, but never more homologous.

Best reciprocal hits, mcl, orthoMCL, eggNOG, Inparanoid



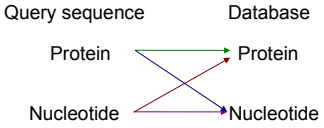
Comparison of whole proteins (Similarity search)

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:

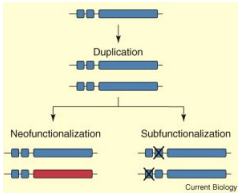


http://www.uniprot.org/

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.

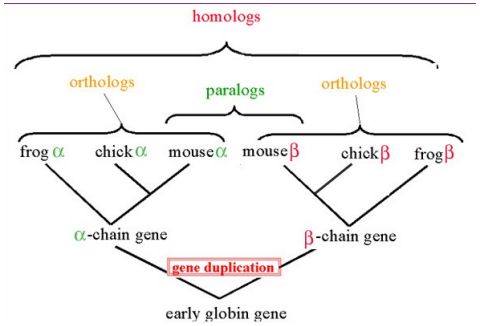


Subfunctionalization: Each paralog performs part of the original function.

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.

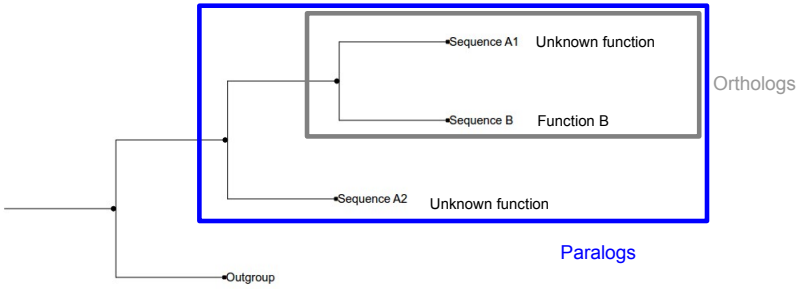
Identification of orthologs and paralogs.



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

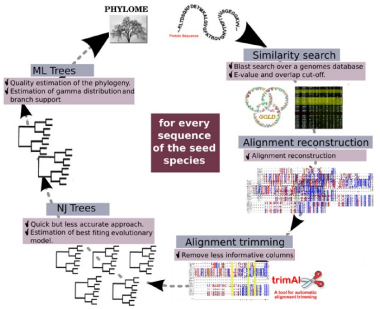
Paralogs are less likely to keep the same function than orthologs.

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

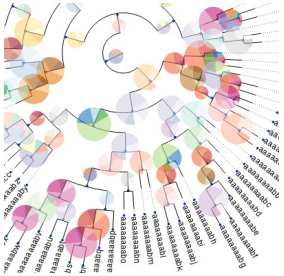
Reviewed by Jaime Huerta-Cepas,[1] Joaquín Dopazo,[2] and Toni Gabaldón[1]

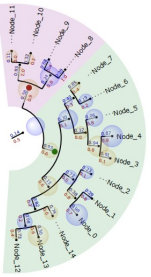
E.T.E.
A python Environment for Tree Exploration

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/

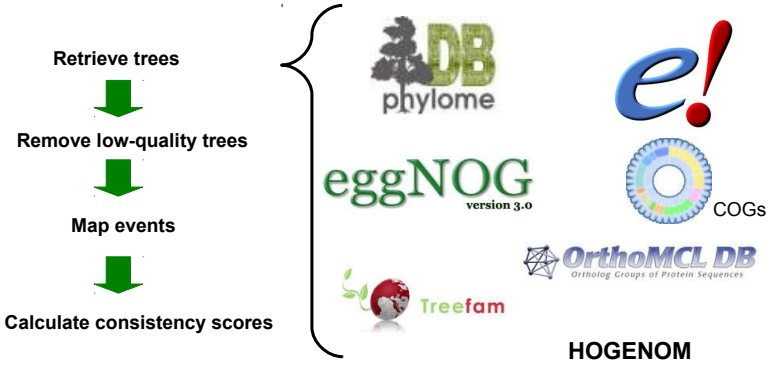
PhylomeDB uses the ETE visualization modules to show the trees.



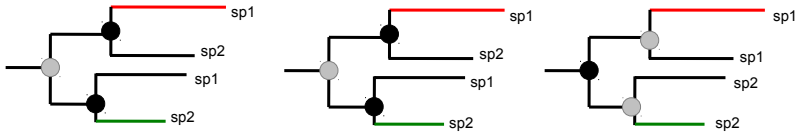


MetaPhors:

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



Orthology and paralogy prediction



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

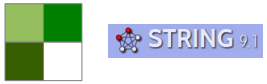
Other sources of information

There is a high possibility that the homologs or orthologs to your protein are annotated as unknown. What then?

Search for conserved protein domains

Search for metabolic information

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

CNV detection
(Lecture & Workshop)

German Demidov
Centre for Genomic Regulation, Barcelona

Wednesday, 9:00

Outline

Introduction to the problem
- Mechanisms of variation
- What is a CNV/CNA
- Importance of CNV/CNA detection
- Current methods
- NGS-based methods: Problems

NGS-based methods: Review of approaches
- WGS vs. WES vs. targeted enrichment
- Paired-end mapping (WGS)
- Split-read-based methods (WGS)
- Read-depth methods
- Read-depth methods (WGS)
- Read-depth methods (WES and hybridization-based panels)
- Read-depth methods (Amplicon Capture)
- De novo assembly (WGS)
- B-Allele Frequency

Disclaimer

- There is no "silver bullet" for CNVs detection
- The successful variants' detection is only possible with the right understanding of the situation and your needs
- There is a huge pool of methods for CNV detection, but the very best and reliable mean of selection and verification of its results is your knowledge and common sense

CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

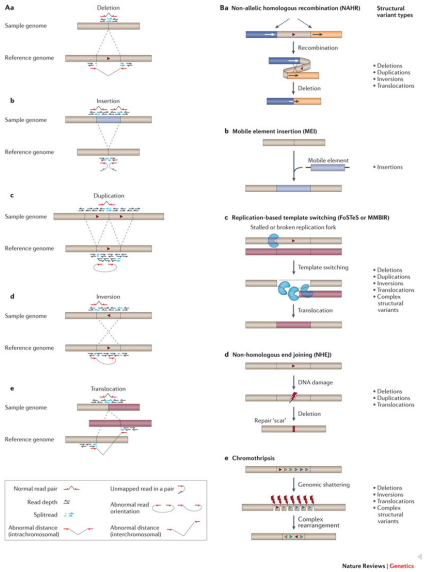
Outline

Introduction to the problem
- Mechanisms of variation
- What is a CNV/CNA
- Importance of CNV/CNA detection
- Current methods
- NGS-based methods: Problems

NGS-based methods: Review of approaches
- WGS vs. WES vs. targeted enrichment
- Paired-end mapping (WGS)
- Split-read-based methods (WGS)
- Read-depth methods
- Read-depth methods (WGS)
- Read-depth methods (WES and hybridization-based panels)
- Read-depth methods (Amplicon Capture)
- De novo assembly (WGS)
- B-Allele Frequency

CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Structural variation (SV)

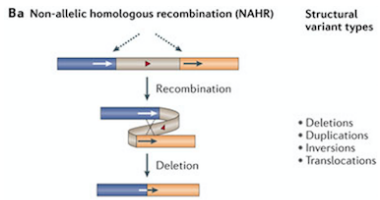


CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Structural variation (SV)

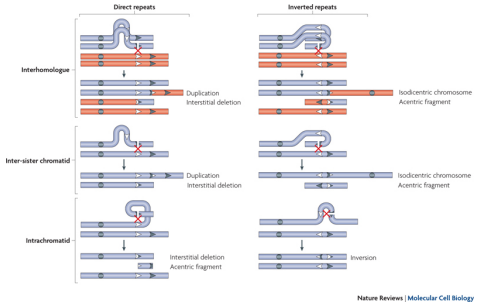
Comprise unbalanced copy-number variations ≥ 50 bp, including deletions, insertions and duplications, as well as balanced variants such as inversions and translocations.

Molecular mechanisms leading to structural variant formation



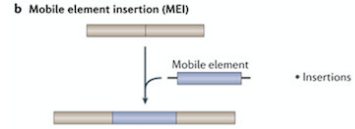
Recurrent structural variants often result from non-allelic homologous recombination (NAHR) which involves recombination between long highly similar low-copy-number repeats.

Molecular mechanisms leading to structural variant formation



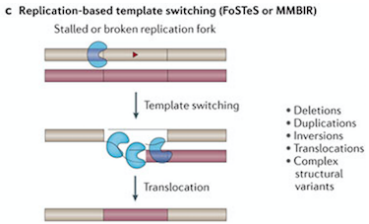
Homologous chromosomes are shown in blue and red, and sister chromatids are depicted in the same colour. Low-copy repeats (LCRs, SDs) – white and black arrows.
Genome destabilization by homologous recombination in the germ line, Sasaki et al., Nature Reviews Molecular Cell Biology 11, 182-195 (March 2010)

Molecular mechanisms leading to structural variant formation



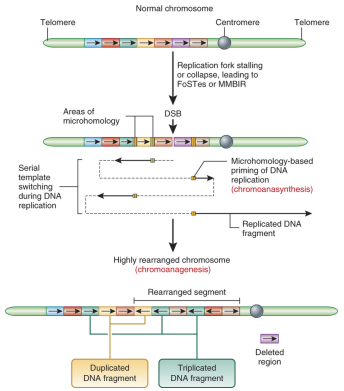
Genomic insertions can involve mobile element insertions of transposable elements by retrotransposition.

Molecular mechanisms leading to structural variant formation



DNA-replication-associated template-switching events, involving the fork-stalling and template switching (FoSTeS) and microhomology-mediated break-induced replication (MMBIR) mechanisms.

Molecular mechanisms leading to structural variant formation



Chromoanagenesis and cancer: mechanisms and consequences of localized, complex chromosomal rearrangements, Holland et al, Nature Medicine 18, 2012.

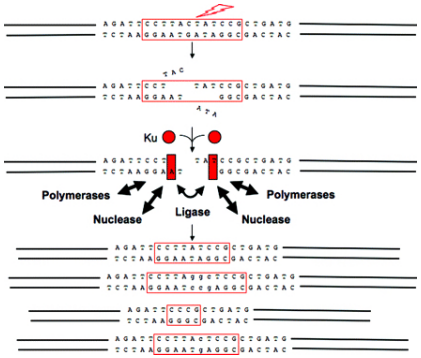
CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Molecular mechanisms leading to structural variant formation

Non-homologous end joining (NHEJ) is a process that repairs DNA double-strand breaks in the absence of extensive sequence homology and is often accompanied by the addition or deletion of several nucleotides in the form of a 'repair-scar'.

CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

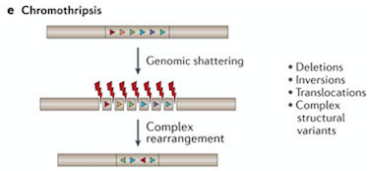
Molecular mechanisms leading to structural variant formation



NHEJ is an emergency repair mechanism which involves a "repair or die" chance.
Chris from biology.stackexchange.com .

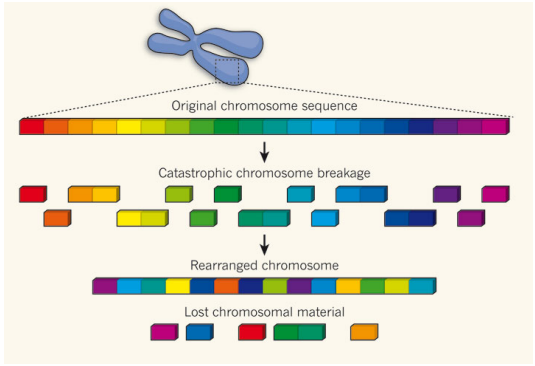
CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Molecular mechanisms leading to structural variant formation



Chromothripsis is a phenomenon that seems to involve chromosome shattering leading to numerous breakpoints, followed by error-prone DNA repair.
In both cancer and congenital diseases.

Molecular mechanisms leading to structural variant formation



Cancer: When catastrophe strikes a cell. Tubio and Estivill, Nature 470, 476477 (24 February 2011)

Outline

Introduction to the problem
- Mechanisms of variation
- What is a CNV/CNA
- Importance of CNV/CNA detection
- Current methods
- NGS-based methods: Problems

NGS-based methods: Review of approaches
- WGS vs. WES vs. targeted enrichment
- Paired-end mapping (WGS)
- Split-read-based methods (WGS)
- Read-depth methods
- Read-depth methods (WGS)
- Read-depth methods (WES and hybridization-based panels)
- Read-depth methods (Amplicon Capture)
- De novo assembly (WGS)
- B-Allele Frequency

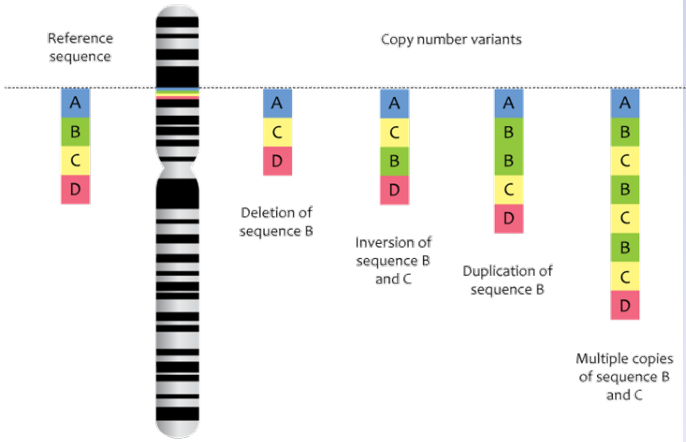
Aneuploidy

It is not a CNV.

1. (45, X) - Turner syndrome.
2. In uniparental disomy, both copies of a chromosome come from the same parent (with no contribution from the other parent).
3. Trisomy 21, Trisomy 18, Trisomy 13 - Down, Edwards, Patau. (47, XXX), (47, XXY), (47, XYY).
4. XXXX, XXYY, XXXXX, XXXXY and XYYYY.

What is a CNV/CNA

Different definitions



CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

What is a CNV/CNA

Different definitions

- Copy number alterations and copy number aberrations are **synonyms**.
- Copy number alterations/aberrations **(CNAs)** are changes in copy number that have arisen in **somatic** tissue (for example, just in a tumor), copy number variations **(CNVs)** originated from changes in copy number in **germline** cells (and are thus in all cells of the organism).
- However some articles seem to use copy number alterations/aberrations **(CNAs)** as the term encompassing both germline and somatic copy number changes and use somatic copy number alterations/aberrations **(SCNAs)** as the term for somatic copy number changes.

CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Segmental Duplication

Hotspot regions in the genome where copy number variations are four times more enriched. Range from 1 to 400 kb in length and occur at more than one site within the genome.

- Segmental duplications (SDs), low copy repeats, are large continuous stretches of DNA that can be mapped to multiple locations on the genome and share $> 90\%$ nucleotide similarity with each other.
- These hotspot regions have an increased rate of chromosomal rearrangement.
- The higher frequencies of SDs within the human population suggest that they are shared duplications that have been fixed in the population rather than being recurrent structural mutations.

CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

What is a CNV/CNA

When CNV and indel become different

- CNVs: "a segment of DNA that is 1 kb or larger and is present at a variable copy number in comparison with a reference genome". However, the cutoff of 1 kb is completely arbitrary.
- Based on a functional definition, it may be better to choose an average exon size ($\sim$ 100 bp) as a parameter for defining CNV.
- Recent observations in the Watson and Venter genomes clearly indicate that the CNV size distributions show a marked enrichment in the range of 300 to 350 bp owing to the known retrotransposition-based Alu polymorphisms.

CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Outline

Introduction to the problem
- Mechanisms of variation
- What is a CNV/CNA
- Importance of CNV/CNA detection
- Current methods
- NGS-based methods: Problems

NGS-based methods: Review of approaches
- WGS vs. WES vs. targeted enrichment
- Paired-end mapping (WGS)
- Split-read-based methods (WGS)
- Read-depth methods
- Read-depth methods (WGS)
- Read-depth methods (WES and hybridization-based panels)
- Read-depth methods (Amplicon Capture)
- De novo assembly (WGS)
- B-Allele Frequency

Importance of CNV/CNA detection

Structural variants account for 1.2% of the variation among human genomes while single nucleotide polymorphisms (SNPs) represent 0.1% (Pang et al., 2010).

Is there a lot of CNVs?

Table 1 | **Summary of copy number variation in the genome based on the inclusive and stringent maps**

Copy number variation measures	All variants		Gains		Losses	
	Inclusive map	Stringent map	Inclusive map	Stringent map	Inclusive map	Stringent map
Total genome variable (%)*	9.5	4.8	3.9	2.3	7.5	3.6
Total genome variable (Mb)	273	136.6	111.5	64.7	215	102.4
Median interval length of CNVRs (bp)	981	1,237	3,334	9,741	956	1,137
Mean interval length of CNVRs (bp)	11,362	11,647	35,581	55,370	9,181	8,883
Number of CNVRs	24,032	11,732	3,132	1,169	23,438	11,530

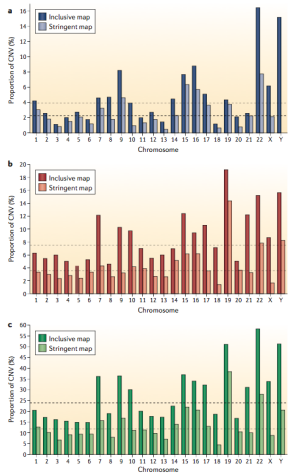
CNVR, copy number variable region. *Numbers listed are based on the upper boundary size estimates of CNVRs. Average boundary sizes of the total genome include: all variants in the inclusive map (8.8%) and stringent map (4.1%), gains in the inclusive map (3.5%) and stringent map (1.9%), and losses in the inclusive map (6.9%) and stringent map (3.1%).

(Zarrei et al, Nature, 2015)

Are the CNVs disease-causing?

- Most of CNVs age benign variants that will not directly cause disease.
- Some CNVs found in the general population can be millions of bases in size, affecting numerous genes, yet they have no observable consequence.
- Was estimated that 4.8–9.5% of the genome contributes to CNV and found approximately 100 genes that can be completely deleted without producing apparent phenotypic consequences. (Zarrei et al, Nature, 2015)
- There are several instances where CNVs that affect critical developmental genes do cause disease. For example, recent reviews have listed 17 conditions of the nervous system alone – including Parkinsons Disease and Alzheimers Disease – that can result from CNV. (http://www.gene-quantification.de/cnv-faq.pdf)

Are CNVs distributed uniformly?



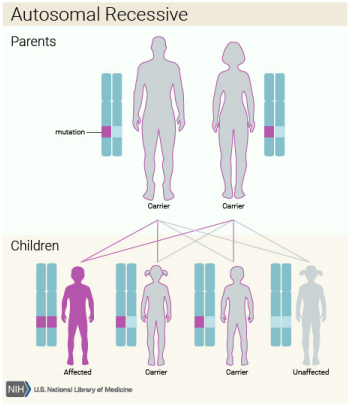
(Zarrei et al, Nature, 2015)

Are the CNVs population-specific?

As with all types of genetic variation, CNVs can vary in frequency and occurrence between populations telling us something of our shared history. As a result of our recent common origin in Africa, the vast majority of copy-number variation around 89% is shared among the diverse human populations studied.

CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Importance of CNV/CNA detection: Mendelian disease



CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Importance of CNV/CNA detection: Complex traits

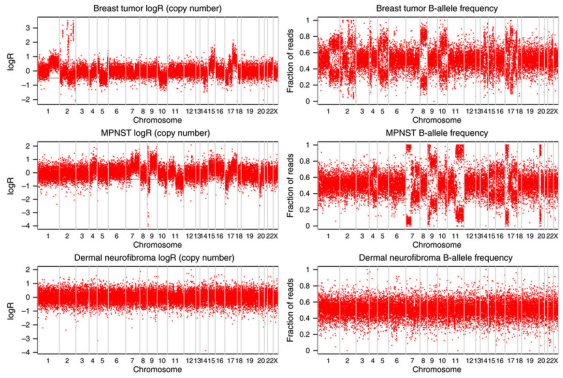
Four examples of CNVs associated with complex traits:

1. a 20-kb deletion upstream of the IRGM gene with Crohns disease,
2. a 45-kb deletion upstream of NEGR with body mass index,
3. a 32-kb deletion that removes two late-cornified envelope genes with psoriasis,
4. a 117-kb deletion of UGT2B17 with osteoporosis.

(Conrad et al, 2013, Nature)
Also well known connections: Parkinson disease, Alzheimer, Mental retardation, Autism, Schizophrenia, etc.

CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Importance of CNV/CNA detection: Cancer



Whole-exome sequencing of breast cancer, malignant peripheral nerve sheath tumor and neurofibroma from a patient with neurofibromatosis type 1, Cancer Medicine, 2015, John R McPherson et al.

Importance of CNV/CNA detection: Evolution

Purifying Selection

CNVs are preferentially located outside of genes and ultraconserved elements in the human genome and that a significantly lower proportion of deletions than duplications overlaps with disease-related genes and RefSeq genes.
Zhang et al, Annu Rev Genomics Hum Genet., 2009.

CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Importance of CNV/CNA detection: Evolution

Gene duplication and Positive Selection

Gene duplication has long been thought to be a central mechanism driving long-term evolutionary changes.
Selection has also been shown to shape the architecture of segmental duplications during human genome evolution.
CNVs encompassing functional genes can be evolutionally favored because of their adaptive benefits.
Zhang et al, Annu Rev Genomics Hum Genet., 2009.

CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

CNA/CNV databases

- Several databases e.g., the Database of Genomic Variants archive which reports structural variation identified in healthy control samples (DGVa) have been created for the collection of SVs data (Lappalainen et al., 2013).
- Public data resources have been developed with the purpose of supporting the interpretation of clinically relevant variants, e.g., dbVar, or collecting known disease genes (Online Mendelian Inheritance in Man, OMIM) hit by SVs.

CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Outline

Introduction to the problem
- Mechanisms of variation
- What is a CNV/CNA
- Importance of CNV/CNA detection
- Current methods
- NGS-based methods: Problems

NGS-based methods: Review of approaches
- WGS vs. WES vs. targeted enrichment
- Paired-end mapping (WGS)
- Split-read-based methods (WGS)
- Read-depth methods
- Read-depth methods (WGS)
- Read-depth methods (WES and hybridization-based panels)
- Read-depth methods (Amplicon Capture)
- De novo assembly (WGS)
- B-Allele Frequency

Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

What does it mean to detect a CNV?

Before choosing the tool for CNV detection, the researcher should understand what does he/she wants to detect:

- exons with CN change
- regions with CN change (not necessarily protein-coding)
- panel of genes of interest with CN change
- etc.

CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

What does it mean to detect a CNV?

One can be interested in:

- populational CNVs
- rare CNVs
- CNVs/CNAs that happen only in a subclone of the sequenced cells (cancer or prenatal diagnostics)

CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

What does it mean to detect a CNV?

The goal can be

- identify CN of genes
- to find breakpoints of CNVs
- to find just some regions with CNVs for further analysis (i.e., tumor purity and clonal structure)
- to identify CNVs associated with traits (for example, level of RNA expression)
- etc.

CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Locus-specific CNV detection

- MLPA
- fusion amplicon formation
- qPCR, dPCR
- FISH-hybridization
- paralog-ratio testing
- molecular copy number counting
- RFLP (restriction fragment length polymorphism) followed by Southern blot analysis
- long-range PCR
- Sanger sequencing of the fragment of interest

Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Locus specific CNV detection

Common features:

- reliable
- takes a lot of time and money
- often allow to detect breakpoints of the CNV

CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Genome-wide CNV detection

- NanoString nCounter (a lot of genes at once, but not whole genome)
- aCGH and SNP arrays
  "The Agilent Human Genome CGH Microarray is a dual color array containing 60-mer oligonucleotide probes, Distinct Biological Features: 963,029, Probe Spacing: 2.1 KB overall median probe spacing (1.8 KB in Refseq genes)"
  "The SNP array platform includes $\sim$ 900000 SNP probes and 900000 non-SNP oligonucleotide probes at an average distance of 0.7Kb"
- NGS-based methods

CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Array-based CNV detection

Common features:

- quite cheap
- has comparatively low resolution
- difficult to find exact breakpoints

CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

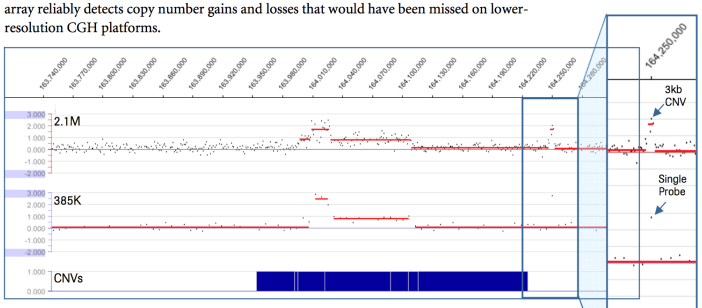
Resolution

- aCGH: from 100 kilobases (according to wiki)
- 25-50 kbps according to Affymetrix website OncoScan FFPE Assay Kit: 50-100 kb copy number resolution in $\sim$ 900 cancer genes, 300 kb genome-wide copy number resolution outside of the cancer genes
  In Practice: "To reduce the number of false positives, parameters were set to consider only imbalances $>$ 75Kb encompassing at least 80 probe sets." Bernardini, 2010, Eur J Hum Genet.
- NGS-based methods can potentially detect 1 kbps events and even less. NGS and SNP-based arrays have troubles with duplications' detection in comparison with aCGH.

Resolution: Important remark

NimbleGen CGH HD2 Analysis

With 2.1 million probes on a single slide, the NimbleGen CGH 2.1M Whole-Genome Tiling v2.0D array reliably detects copy number gains and losses that would have been missed on lower-resolution CGH platforms.



Test and reference gDNA samples were independently labeled with fluorescent dyes, co-hybridized to a NimbleGen Human CGH 2.1M (HD2) or 385K Whole-Genome Tiling array, and scanned using a 5µm scanner. Log_2-ratio data from the two platforms are displayed in Roche NimbleGen SignalMap software alongside an annotation track showing known copy number variants (CNVs) from the Database of Genomic Variants (http://projects.tcag.ca/variation/). The increased probe density on the CGH HD2 array (1.1kb vs. 6kb median probe spacing) enabled the detection of a novel ~3kb CNV that was detected by only a single probe on the 385K platform (see arrows above). In addition, fine structure of a previously reported CNV region was further elucidated using the CGH HD2 array.

April 2008 Roche NimbleGen

Resolution: Important remark

Structural variations of DNA greater than 1 kilobase in size account for most bases that vary among human genomes, but are still relatively under-ascertained. Here we use tiling oligonucleotide microarrays, comprising 42 million probes, to generate a comprehensive map of 11,700 copy number variations (CNVs) greater than 443 base pairs, of which most (8,599) have been validated independently.
Origins and functional impact of copy number variation in the human genome, Conrad et al, Nature, 2013.

Resolution: Important remark

- Experimental strategy to discover CNVs greater than 500 base pairs (bp) in individuals with European or West African ancestry
- 20 NimbleGen arrays (median spacing of 56 bp)
- 800 comparative genome hybridization (CGH) experiments with female lymphoblastoid cell-line DNA competed against a common male European reference sample (NA10851)
- The female test DNAs comprised 19 CEU European HapMap individuals, 20 YRI (Yoruba, Nigeria)-West Africans, and a Polymorphism Discovery Resource individual (NA15510)
- It was estimated that 40 samples would provide 95% power to sample variants with minor allele frequencies of 5% in either population

Origins and functional impact of copy number variation in the human genome, Conrad et al, Nature, 2013.

Resolution: Important remark II, Venter's genome

Table 1
Structural variants detected by different methods

Method	Type	Number	Minimum size (bp)	Median size (bp)	Maximum size (bp)	Total size (bp)
Assembly comparison [9]	*Homo. insertion*	*275,512*	*1*	*2*	*82,711*	*3,117,039*
	Homo. deletion	*283,961*	*1*	*2*	*18,484*	*2,820,823*
	Hetero. insertion	*136,792*	*1*	*1*	*321*	*336,374*
	Hetero. deletion	*99,814*	*1*	*1*	*349*	*250,300*
	Inversion	*88*	*102*	*1,602*	*686,721*	*1,627,877*
Mate-pair	Insertion	780	346	3,588	28,344	3,880,544
	Deletion	1,494	340	3,611	1,669,696	10,531,345
	Inversion	105	368	3,121	2,026,495	8,068,541
Split-read	Insertion	8,511	11	16	414	224,022
	Deletion	11,659	11	18	111,714	1,764,522
Agilent 24 M	Duplication	194	445	1,274	113,465	1,065,617
	Deletion	319	439	1,198	852,404	2,779,880
NimbleGen 42 M	Duplication	366	448	4,665	836,362	11,292,451
	Deletion	358	459	2,460	359,736	3,861,282
Affymetrix 6.0	Duplication	17	8,638	42,798	640,474	2,011,557
	Deletion	21	2,280	13,145	856,671	1,978,028
Illumina 1 M	Duplication	3	11,539	22,148	87,670	121,357
	Deletion	9	8,576	32,199	145,662	431,131
Custom Agilent 244 k	Duplication	44	219	1,356	8,737	98,529
	Deletion	7	170	332	2,258	4,130
Non-redundant total [b]	Insertion/duplication	417,206	1	1	836,362	19,981,062
	Deletion	390,973	1	2	1,669,696	19,539,369
	Inversion	167	102	1,249	2,026,495	9,257,035

[a]We used an italicized font to distinguish the results from the Levy *et al.* [1] study. Moreover, from that previous study, we included all homozygous indels, heterozygous indels, indels embedded within simple, bi-allelic, and non-ambiguously mapped heterozygous mixed sequence variants, and only those inversions whose size is at most 3 Mb. [b]Complete data are presented in Additional files 19, 20 and 21. Non-redundant variation size distribution is presented in Figure 2a.

Towards a comprehensive structural variation map of an individual human genome, Pang et al, 2010, Genome Biology.

Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Outline

Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

NGS-based methods: Problems

"Copy-number variants (CNVs) are considerably more difficult to find – at least using NGS.
Why? In a word, length.
Todays NGS technologies produce millions upon millions of sequence reads, but they're mostly relatively short, measuring a few hundred bases in size. Its difficult using such data to piece together the subtle structural variations that distinguish one individual from another, simply because individual reads often are too short to span the variant regions of the genome." - Jeffrey M. Perkel, biocompare.com

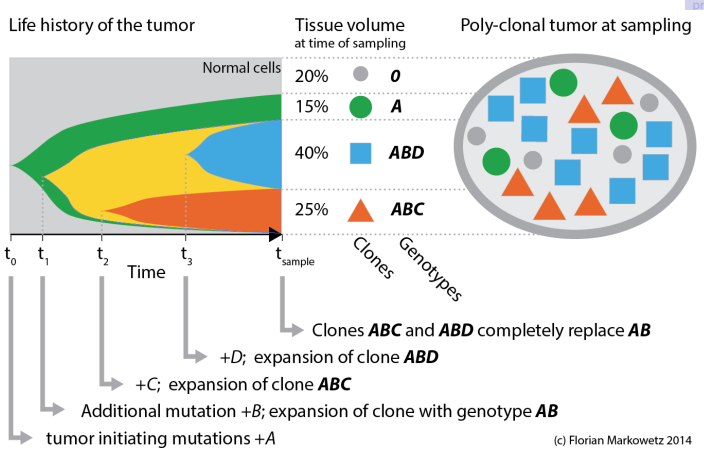
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

NGS-based methods: General Problems

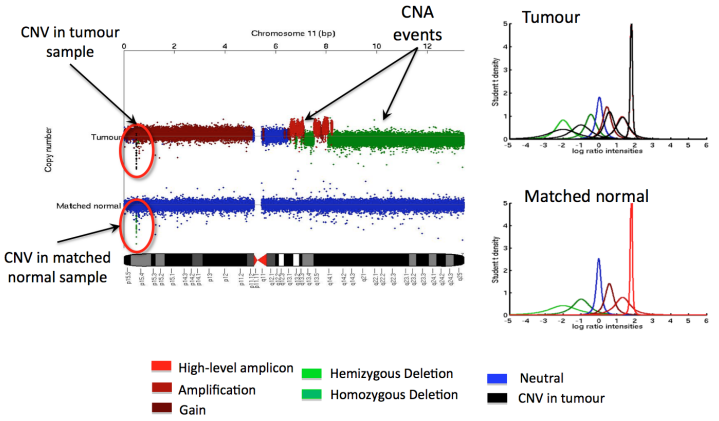
- Low coverage and high level of noise
- Events that look like CNVs but they are not (i.e., alignment artifacts)
- Batch effects
- Bad quality of extracted DNA
- Wrong preparation of experiment
- Low-complexity regions

Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

NGS-based methods: Problem in CNA



NGS-based methods: Problem in CNA



NGS-based methods: Problem in CNA

Oesper *et al. Genome Biology* 2013, **14**:R80
http://genomebiology.com/2013/14/7/R80

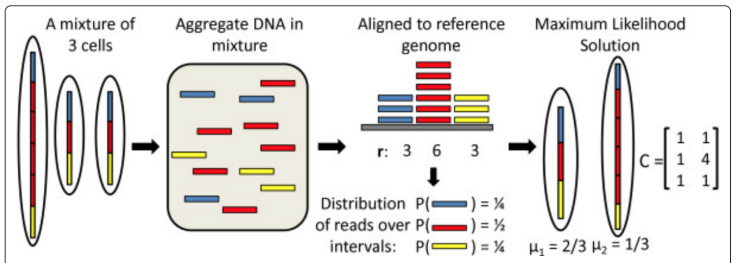
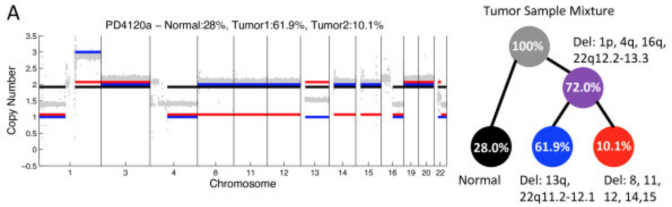


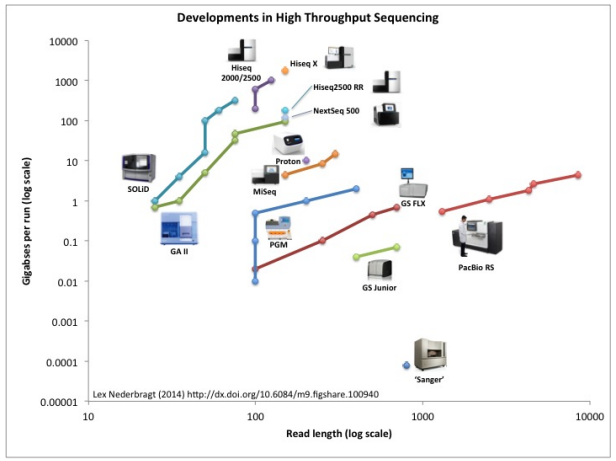
Figure 1 Algorithm overview. A mixture of three subpopulations with two distinct genomes: a normal genome (represented here with one copy of each interval for simplicity), and an aneuploid genome with a duplication of one interval (red). If reads are distributed uniformly over the aggregate DNA in the sample, then the observed distribution of reads over the blue, red and yellow intervals will follow a multinomial distribution with parameter $C\mu$. Here **C** is the interval count matrix giving the integral number of copies of each interval in each genome in the mixture, and μ is the genome mixing vector giving the proportion of each subpopulation in the mixture. We find the pair ($\mathbf{C}$, μ) that maximizes the likelihood of the observed read depth vector **r**.

NGS-based methods: Problem in CNA



Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

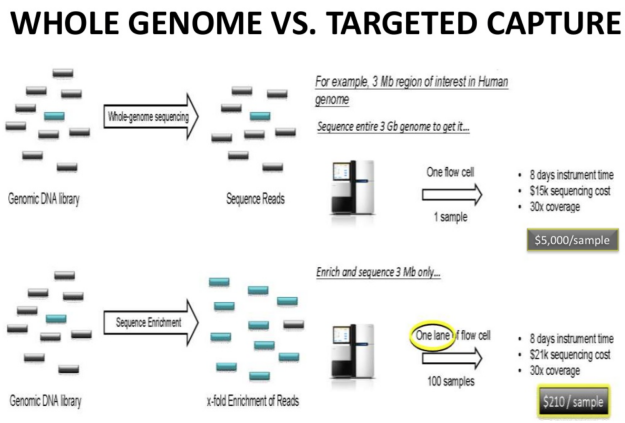
Does sequencing technology influence the CNV detection protocol?



Outline

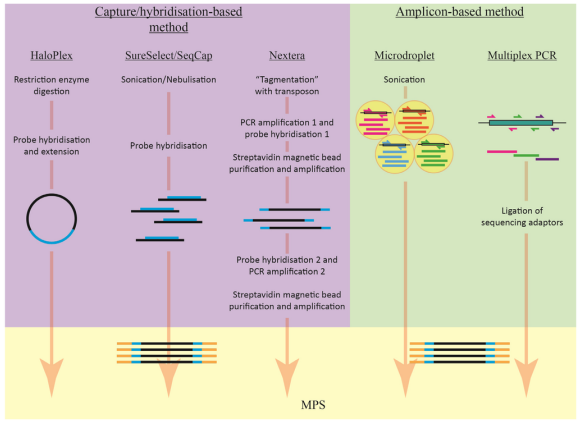
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Targeted enrichment



Hybridization vs AmpliSeq

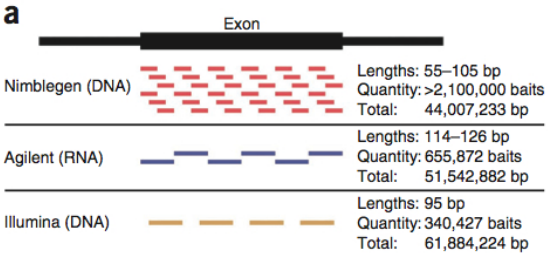
Q: What type of biases may arise?



(from http://www.mdpi.com)

Design of the panel

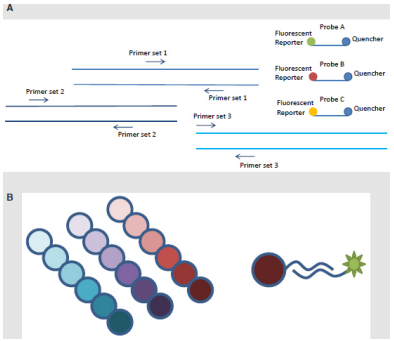
Q: What type of biases may arise due different designs?



Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Pooling strategy

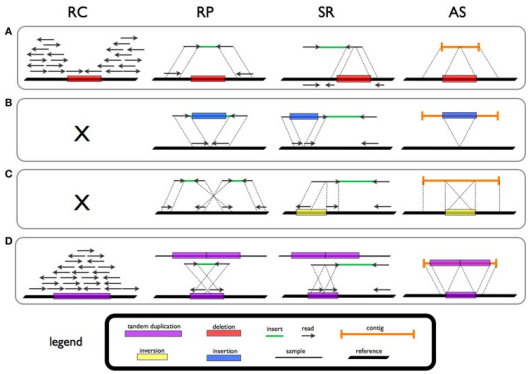
Q: What type of biases may arise due different designs?



(from BioWatch PCR Assays: Building Confidence, Ensuring Reliability; Abbreviated Version (2015))

Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Overview of the approaches



Deletion (A), novel sequence insertion (B), inversion (C), and tandem duplication (D) in read count (RC), read-pair (RP), split-read (SR), and de novo assembly (AS) methods. Tattini et al, Front. Bioeng. Biotechnol., 25 June 2015

Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Outline

Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Read-pair

- Two different strategies have been used in PEM-based tools to detect SVs/CNVs, namely the clustering approach and the model-based approach.
- The difference lies in that the clustering approach employs a predefined distance to identify discordant reads, while the model-based approach adopts a probability test to discover the unusual distance between read pairs in comparison to the distance distribution in genome.

Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Read-pair: Conclusion

- This method is quite reliable.
- It can not detect exact copy numbers.
- RP algorithms cannot detect the signatures of novel sequence insertions larger than the average insert size.

CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Outline

Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Split-read-based methods (WGS)

- SR methods were conceived for Sanger sequencing reads.
- SR methods start from read pairs in which one read from each pair is aligned to the reference genome uniquely while the other one fails to map or only partially maps to the genome.
- Those unmapped or partially mapped reads potentially provide accurate breaking points at the single base pair level for SVs/ CNVs. SR methods split the incompletely mapped reads into multiple fragments. The first and last fragments of each split read are then aligned to the reference genome independently.

CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Split-read-based methods (WGS): Conclusion

- The SR-based approach heavily relies on the length of reads and is only applicable to the unique regions in the reference genome.
- It provides super high resolution (1 bp), however it is highly rely on the coverage and these divided reads do not necessarily show the CNV/SV (depending on probe preparation).
- It can not detect exact copy numbers.

CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Outline

Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Read-depth methods

- The underlying hypothesis of RD-based methods is that the depth of coverage in a genomic region is correlated with the copy number of the region, e.g., a gain of copy number should have a higher intensity than expected.
- Compared to PEM and SR-based tools, RD-based methods can detect the exact copy numbers, which the former approaches are lacking because PEM/SR methods only use the position information.
- RD-based methods can detect large insertions and CNVs in complex genomic region classes, which are difficult to detect using PEM and SR methods.

CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Read-depth methods

- Generally, RD-based tools can be classified into three categories depending on the study design: single samples, paired case/control samples (somtimes trios), and a large population of samples.

Q: How the copy number detection is different between these cases?

CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

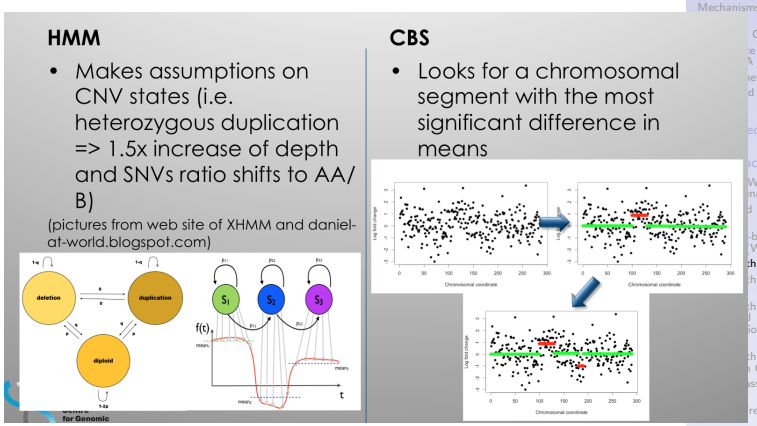
Read-depth methods

- Basically, RD-based methods follow a four-step procedure to discover CNVs: mapping, normalization, estimation of copy number, and segmentation.

Q: How would you do the normalization? Which are important facts that need to be taken into account?

Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Circular Binary Segmentation and HMM



Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Others

- Mean Shift-Based
- Shifting Level Model
- Poisson modelling
- a lot of crazy algorithms

Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Outline

Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Read-depth methods (WGS)

- Generally, RD-based tools define non-overlapping genomic windows, calculate read depths for these windows, and estimate copy numbers for each of them.

Q: How would you do the segmentation into windows?

CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Outline

Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

WES-based CNV detection

- The full spectrum of CNVs and breakpoints may not be completely characterized.
- Cross-chromosome events may not be detected.
- In contrast to WGS, WES data have higher depth for targeted regions, which is ideal for more accurate CNVs using an RD-based calling approach.

CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

WES-based CNV detection

- Due to differing capture efficiency, the depth from different genomic regions may vary substantially and should be considered in the downstream analysis of CNV calling.
- Due to inconsistent capture efficiency, there might be regions that are poorly sequenced, which requires pre-processing for WES data.
- The assumption of normal distribution may no longer be valid due to the biases regarding read depth distribution.
- Due to the discontinuation of genomic regions, most CNV breakpoints could not be detected.

CNV detection
G. Demidov
Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

WES-based CNV detection: Conclusion

- Cheap
- Not reliable

Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Outline

Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Amplicon Capture

- Amplicon sequencing data show different biases in respect of WES data.
- Protocols involved in the preparation of amplicon libraries result in high depth of coverage at the expense of coverage homogeneity.

Q: PCR duplicates are typically removed before CNV detection in WES data. Should they be removed from AmpliSeq data?

Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Outline

Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

De novo assembly

- By comparing the assembled contigs to the reference genome, the genomic regions with discordant copy numbers are then identified
- This direct assembly of short reads without using a reference is called de novo assembly.
- Assembly can also use a reference genome as a guide to improve its computational efficiency and contig quality.

Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

De novo assembly: Conclusion

- Time consuming
- Requires quite high coverage
- Has problems with non-unique positions in the genome

Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

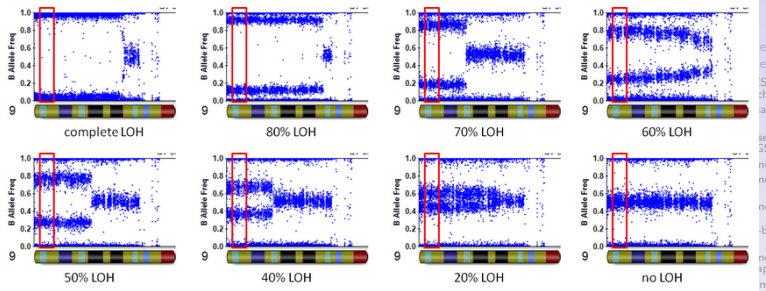
Outline

Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

B-Allele Frequency

- Used in cancer-specific field mainly



Introduction to the problem
Mechanisms of variation
What is a CNV/CNA
Importance of CNV/CNA detection
Current methods
NGS-based methods: Problems
NGS-based methods: Review of approaches
WGS vs. WES vs. targeted enrichment
Paired-end mapping (WGS)
Split-read-based methods (WGS)
Read-depth methods
Read-depth methods (WGS)
Read-depth methods (WES and hybridization-based panels)
Read-depth methods (Amplicon Capture)
De novo assembly (WGS)
B-Allele Frequency

Differential expression (Lecture & Workshop)

Leszek Pryszcz
International Institute of Molecular and Cell Biology in Warsaw

Wednesday, 14:00

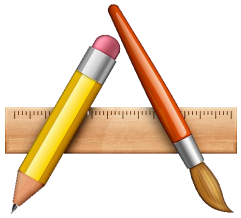
Why is RNAseq so popular?

WTSS: Whole Transcriptome Shotgun Sequencing

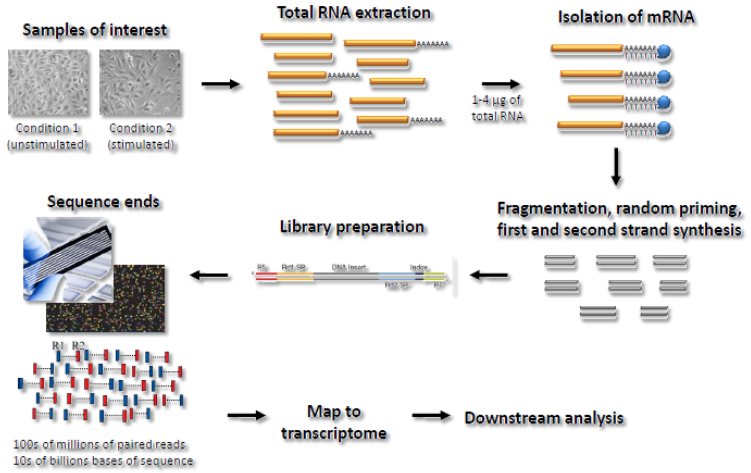
- Reference free
- High sensitivity
- High throughput
- High dynamic range
- Endless applications
- Relatively cheap

RNAseq applications

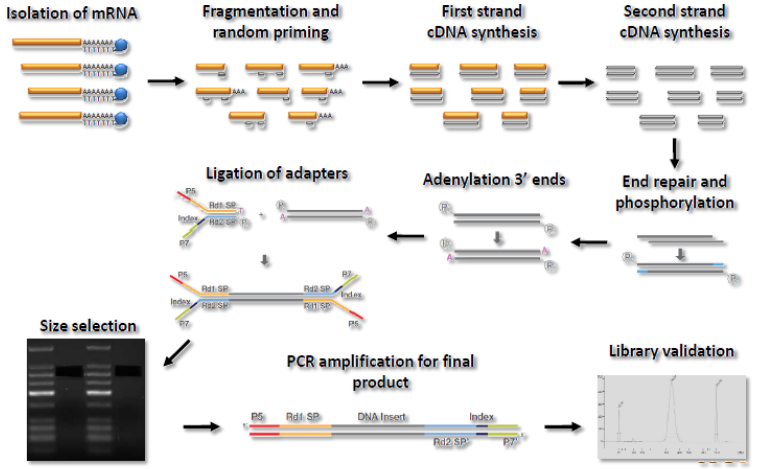
- Gene structure
- Alternative splicing
- Transcription Start Sites (TSS)
- Novel transcripts
- *De-novo* transcriptome assembly
- Differential expression
- Gene fusions



RNA sequencing

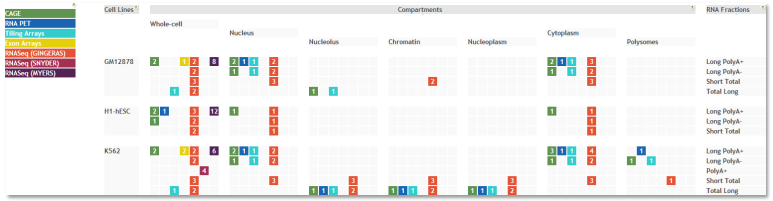


Library preparation (Truseq)



RNAseq methodologies

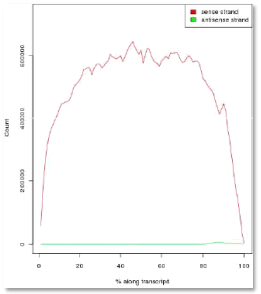
- Non-directional vs directional (strand-specific)
- Whole cell, nuclear, cytosolic…
- Small vs long RNAs
- poly-A$^+$ vs poly-A$^-$
- Dual RNA-Seq: host + pathogen
- Enrichment RNA-Seq

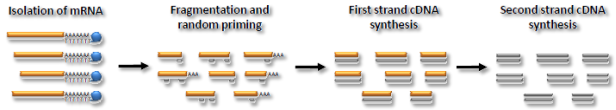


http://genome.crg.es/encode_RNA_dashboard/hg19/

Strand specific RNAseq

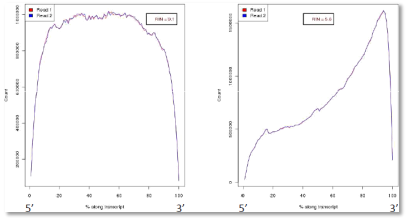
- Different sequences of 5' and 3' adapters
  - directionality is maintained
  - ie. miRNA always directional

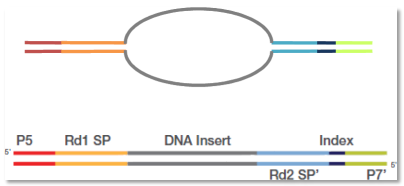




Sources of problems

- RNA is fragile
  - easily degrade
- Reverse-transcription
  - mismatches
- Amplification
  - GC bias
- Primer dimers
  - ~100bp
- Heteroduplex
- rRNA in Bacteria
  - Ribozero or ds nuclease





Differential expression

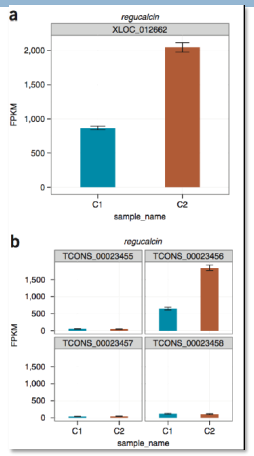
Splice-aware alignment

Transcripts quantification

Calling differentially expressed genes

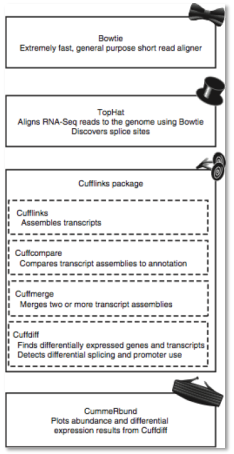
Differential Expression Analysis

- Differential expression of genes or isoforms
- Between two or more experimental conditions
  - Biological **replicates** needed
- Statistical models to assess the significance
  - Correction for **multiple testing**



Cufflinks pipeline

- Tophat
  - Align reads to the genome and ascertains splice sites
- Cufflinks
  - Assemble the reads into transcripts and estimates transcript abundance
- Cuffdiff
  - Call differentially expressed genes
- CummeRbund
  - Visualization of Cuffdiff results



Trapnell *et al.* (2012). *Nat Protoc.* 7:562-78.

Splice-aware alignment

- Reads can span introns
  - segments mapping to different locations in the genome
- Solution
  - Splice-sites detection
  - Large gaps enabled
- Tophat: splice-junction mapper
- Bowtie for alignment
  - large gaps not allowed
- Multiple iterations:
  - Unligned reads broken up and smaller segments aligned
  - Spice-sites infered

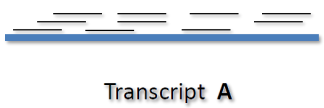
http://ccb.jhu.edu/software/tophat/index.shtml

Splice-aware aligners

STAR	Tophat
Index: 16GB	Index: 1.6GB
89% aligned	69% aligned
74% unique	56% unique
Time: 0:04:49	Time: 0:35:49
223M reads / h	<20M reads / h

Quantifying Transcripts Expression

- Predicted transcripts
  - Reference annotation sometimes not enough
- Normalization
  - Gene length bias
  - Differences in sequencing yields
  - GC bias
  - Differences in read composition
  - Technical artifacts
- More than just counting



Quantifying Transcripts Expression

- Expression given in FPKM
  - **F**ragments **P**er **K**ilobase of exon per **M**illion fragments mapped
- Transcript abundance directly proportional to the number of reads generated from it
- Two normalisation steps
- Example:
  - 10M reads aligned in total
  - 100 aligned to transcript1
  - transcript1 is 1 kb

10 FPKM: 100 (fragments aligned to transcript) / 1 (transcript length in kb) / 10 (M aligned reads)

Mortazavi *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature Methods*

RNA-Seq normalisation

- Normalization based on transcript length

- Normalization based on machine yields

cuffdiff

- Estimates FPKM values per replicate and per condition
- Test for differential expression using statistical model

- Usage:

```
$ cuffdiff
    -p 8                                    # use 8 cores
    -L Ctrl,Treat                           # labels for conditions
    Gene_annotation_file                    # gene annotation file
    [sample1_rep1.bam,..,sample1_repN.bam]  # bam files condition 1
    [sample2_rep1.bam,..,sample2_repN.bam]  # bam files condition 2
```

- Output
  - gene and transcript expression level changes with P values

Visualization

- Tons of data
  - Viewing and comparing is cumbersome

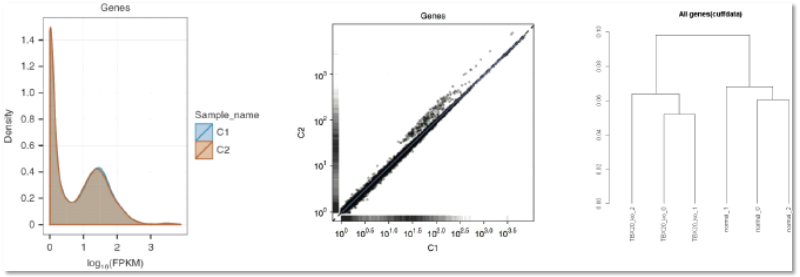
cummeRbund

- R/Bioconductor package
- Help exploring the output of Cuffdiff
- Simple to create figures!

- Usage:
  - $ R # start R from shell
  - > library(cummeRbund) # load library
  - > cuff_data <- readCufflinks('cuffdiff_outdir') # load data

CummeRbund

```
> csDensity(genes(cuff_data))                    # Density plot
> csScatter(genes(cuff_data), 'C1', 'C2')        # Scatter plot
> csDendro(genes(cuff_data), replicates=TRUE)    # Sample clustering
```



Cummerbund: Explore Single Genes

Select a gene

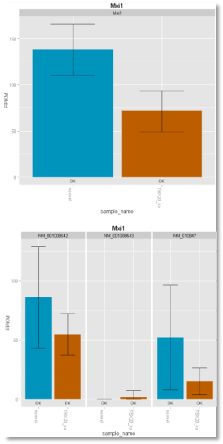
```
> mygene <- getGene(cuff_data, "clock3")
```

Show its gene expression levels

```
> expressionBarplot(mygene)
```

Show isoforms expression

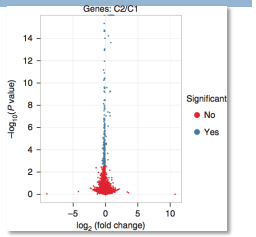
```
> expressionBarplot(isoforms(mygene))
```



Cummerbund: Differential Expression

Volcano plot

```
> csVolcano(genes(cuff_data), 'C1', 'C2')
```

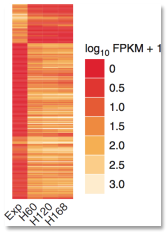


Get list of differentially expressed genes

```
> gene.diff.data <- diffData(genes(cuff_data))
> sig.genes <- subset(gene.diff.data, significant=="yes")
> geneIdList <- sig.genes$gene_id
```

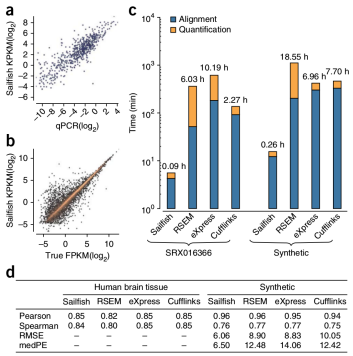
Plot a heatmap for the significant genes

```
> sig.genes.data <- getGenes(cuff_data, geneIdList)
> csHeatmap(sig.genes.data, clustering="row", +
    labRow=F)
```



Alignment-free isoform quantification

- Sailfish
  - alignment-free
- Salmon
  - lightweight alignment
  - wicked-fast
  - superior accuracy

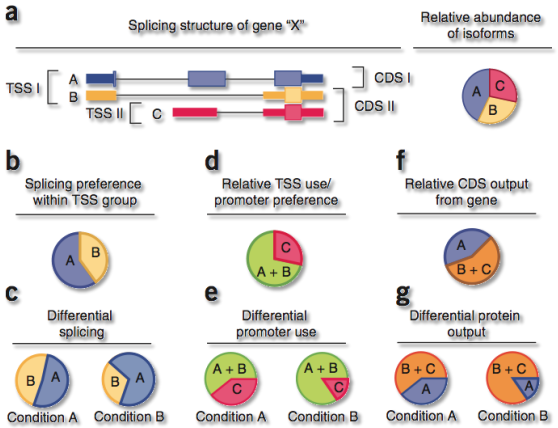


	Human brain tissue				Synthetic			
	Sailfish	RSEM	eXpress	Cufflinks	Sailfish	RSEM	eXpress	Cufflinks
Pearson	0.85	0.82	0.85	0.85	0.96	0.96	0.95	0.94
Spearman	0.84	0.80	0.85	0.85	0.76	0.77	0.77	0.75
RMSE	–	–	–	–	6.06	8.90	8.83	10.05
medPE	–	–	–	–	6.50	12.48	14.06	12.42

```
# index
salmon index -t transcripts.fa -i transcripts.index
# quantify
s=RZE020
salmon quant -p 4 -l SF -i transcripts.index -r <(zcat $s.fq.gz) -o salmon/$s
```

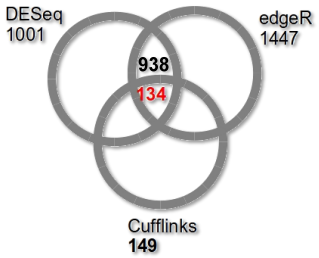
Patro, R., Mount, S. M., & Kingsford, C. (2014). Nature Biotechnology, 32(5), 462–4. doi:10.1038/nbt.2862

Follow-up analyses



Problems with cufflinks

- No flexibility
  - Recomputing each time some condition changes
- Results differ significantly
  - from version to version
    - Cuffdiff 1.1: 145
    - Cuffdiff 1.2: 69
    - Cuffdiff 1.3: 50
  - from other tools (!)



http://gettinggeneticsdone.blogspot.com/2012/04/rna-seq-methods-march-twitter-roundup.html

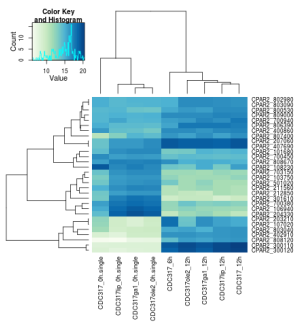
deseq

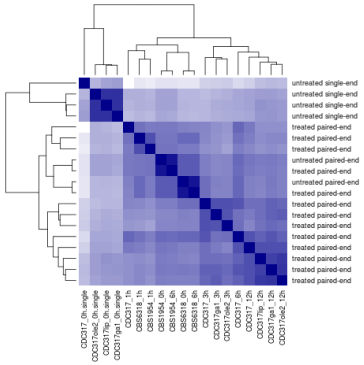
- R/Bioconductor package
- Test for differential expression
- Highly flexible

- Usage:
  - $ R — start R from shell
  - library(DESeq) — import library
- Input:
  - raw reads counts per gene per replica per condition
- Output:
  - differentially expressed genes at given FDR
  - bunch of usefull figures

http://www.biostars.org/p/46427/ http://bioconductor.org/packages/2.13/bioc/html/DESeq.html

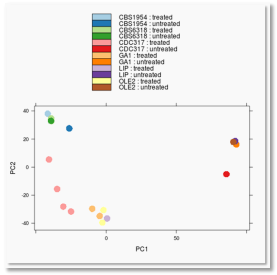
Deseq: heatmaps

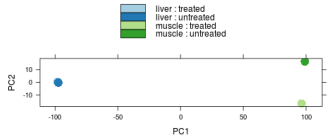




Deseq: PCA

- Multiple samples shown in the 2D plane spanned by their first two principal components
  - largest possible variance





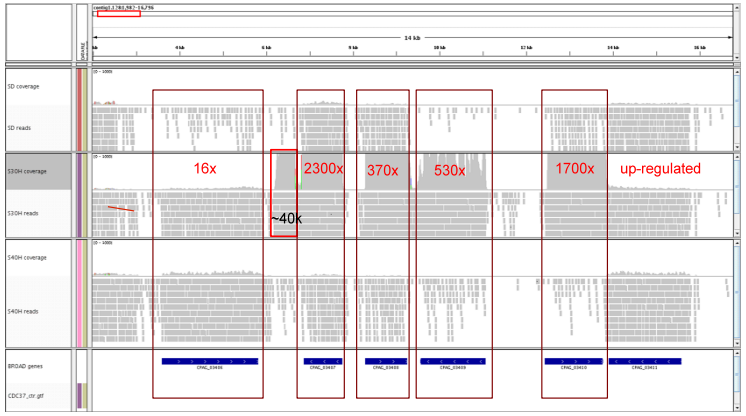
Discussion

Other RNAseq applications

- Detection of novel transcripts
- *De novo* transcriptome assembly
- Dual RNA-Seq
- Polysome profilling
- RNA editing detection
- Parental expression

Sources of biases in NGS

Detection of novel transcripts



Detection of novel transcripts

- Assemble individual transcripts from reads aligned to the genome
- $ cufflinks
  - -p 8 # use 8 threads
  - -G annotation_file # annotation file GTF/GFF
  - alignment_file.bam # read alignments
- Input: Aligned reads in SAM/BAM
- Output: Detected transcripts (transcripts.gtf)
- Cuffmerge:
  - Combine predicted transcripts with known annotation
  - Select novel transcripts

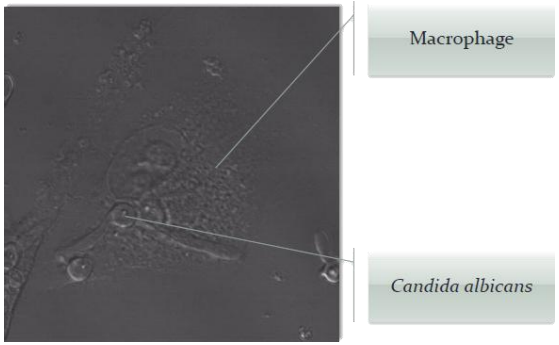
De novo transcriptome assembly

- Multiple de Bruijn graphs
  - one per locus
  - processed independently
- Heavy computations
  - Uneven coverage
  - Multiple isoforms

- USAGE:
  - Trinity.pl
    - --seqType fa # sequence format
    - --JM 10G # virtual memory limit
    - --left reads/q20_1.fasta # left reads file(s)
    - --right reads/q20_2.fasta # right reads file(s)
    - --CPU 2 # number of threads
    - --output assembly # output directory

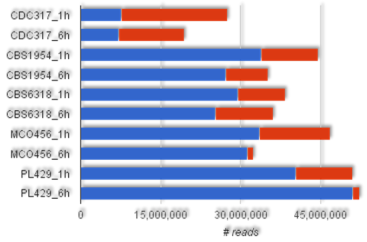
http://trinityrnaseq.sourceforge.net/

DUAL RNAseq

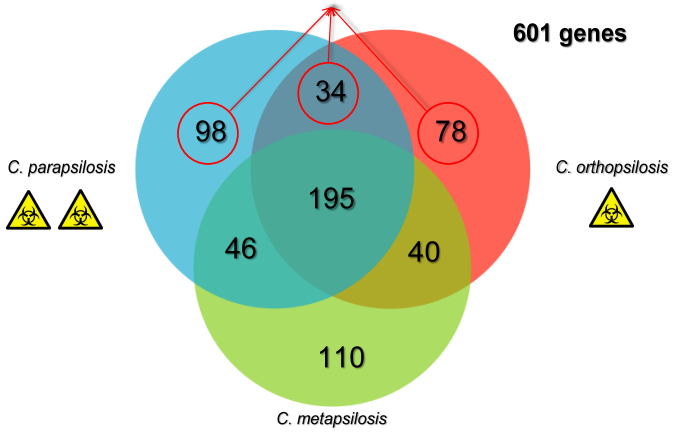


Dual RNAseq

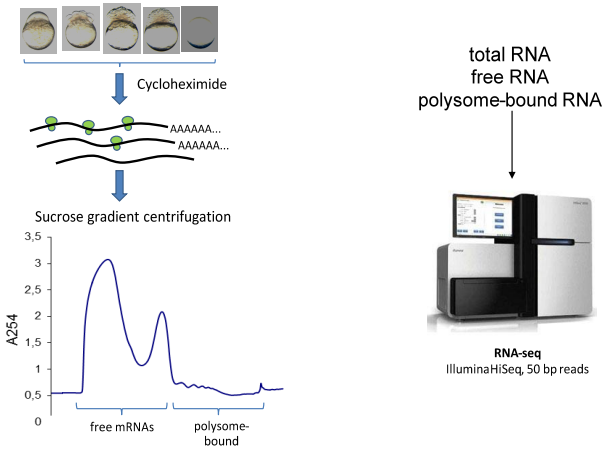
- Measure gene expression of host and pathogen simultaneously
- Two separate controls
  - host
  - pathogen
- Time course experiment
- Difficulties
  - experimental design
  - uneven host:pathogen ratio
  - mapping: divergence >1%
  - expression quantification



DUAL RNAseq



Polysome profiling



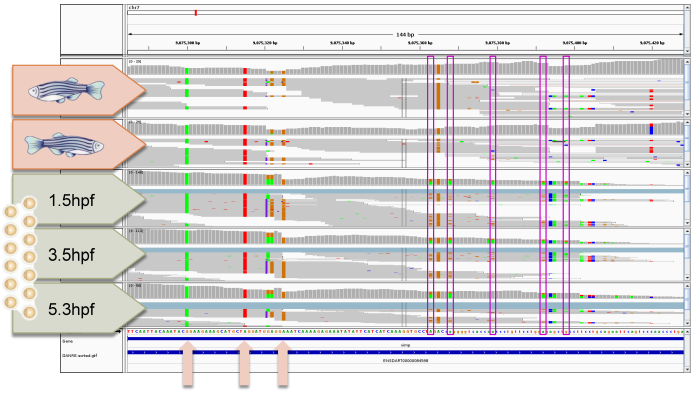
RNA editing detection

38

```
Gene (DNA):      ATTGCTCAT     ATTGCTCAT
Transcript (RNA): AUUGCUCAU     AUUGCUCIU
Protein:         I  A  H       I  A  R
```

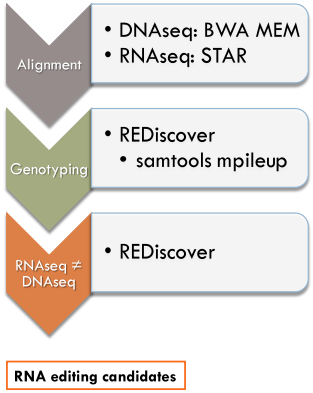
RNA editing detection

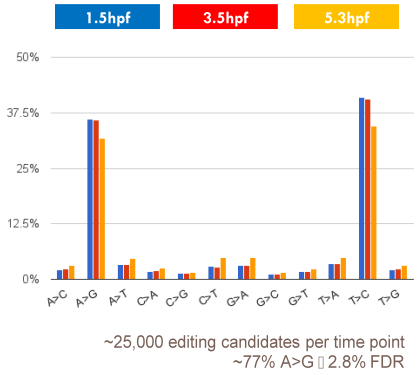
39



https://www.broadinstitute.org/igv/

RNA editing detection



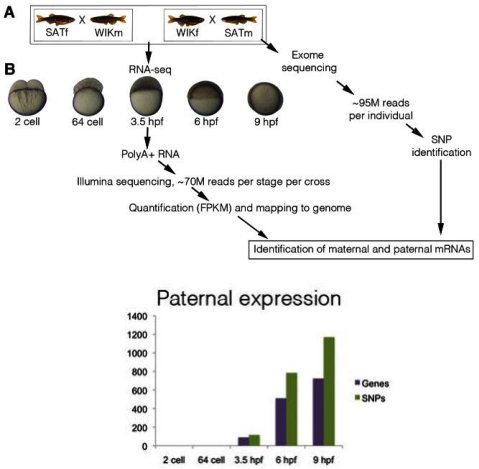


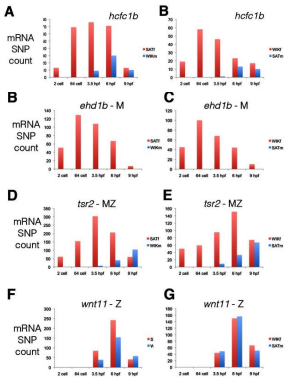
~25,000 editing candidates per time point
~77% A>G
2.8% FDR

```
REDiscover.py -d DNAseq*.bam -r RNAseq*.bam > RNAediting.tsv
```

https://github.com/lpryszcz/REDiscover

Parental expression

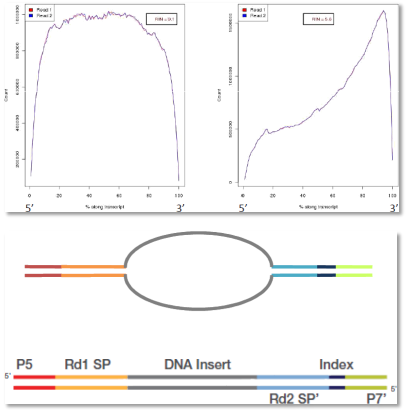




http://www.ncbi.nlm.nih.gov/pubmed/23720042

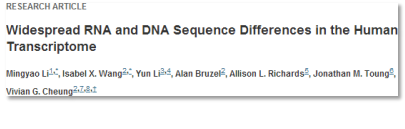
Library prep/sequencing problems

- RNA is fragile
  - easily degrade
- Reverse-transcription
  - mismatches
- Amplification
  - GC bias
- Primer dimers
  - ~100bp
- Heteroduplex
- rRNA in Bacteria
  - Ribozero or ds nuclease
- Contaminations
  - ΦX174, multiplexing



RNA-Seq difficulties

- Spliced-alignment
- Wrong base calls
  - RT + PCR
- Paralogous copies
- Non-directed RNA-Seq
  - Sense or anti-sense?
- Multiple isoforms
- Stranded RNAseq is enrichment


RESEARCH ARTICLE

Widespread RNA and DNA Sequence Differences in the Human Transcriptome

Mingyao Li[1-3], Isabel X. Wang[2-4], Yun Li[3-5], Alan Bruzel[6], Allison L. Richards[6], Jonathan M. Toung[6], Vivian G. Cheung[2,7,8]

http://www.genomesunzipped.org/2012/03/questioning-the-evidence-for-non-canonical-rna-editing-in-humans.php

ChIP-seq
(Lecture & Workshop)

Maciej Łapiński
International Institute of Molecular and Cell Biology in Warsaw

Thursday, 9:00

Walther Flemming
(21 April 1843 – 4 August 1905)

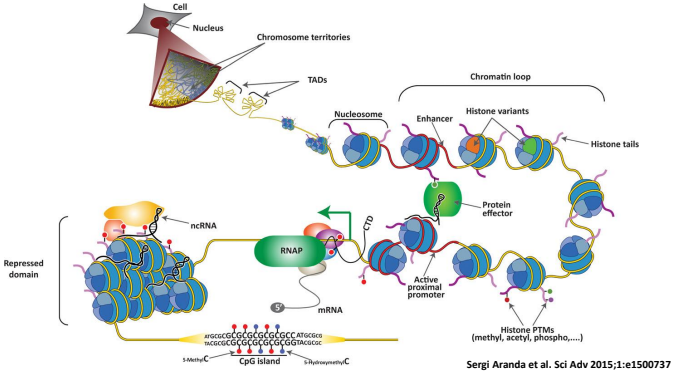
„Therefore, we will designate as chromatin that substance, in the nucleus, which upon treatment with dyes known as nuclear stains does absorb the dye."

Zellsubstanz, Kern und Zelltheilung (1882)

Chromatin immunoprecipitation followed by massively paralel sequencing

Technology that can identify, in an unbiased manner, all DNA segments in the genome physically associated with a specific DNA-binding protein.

Deciphering transcriptional regulatory network



Sergi Aranda et al. Sci Adv 2015;1:e1500737

ChIP-Seq- Applications

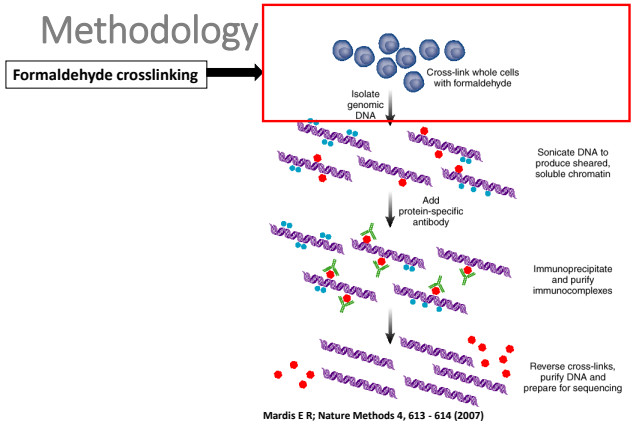
Map the chromosomal locations of:

- transcription factors,
- nucleosomes,
- histone modifications,
- chromatin remodeling enzymes,
- chaperones,
- polymerases

ChIP-Seq- Applications

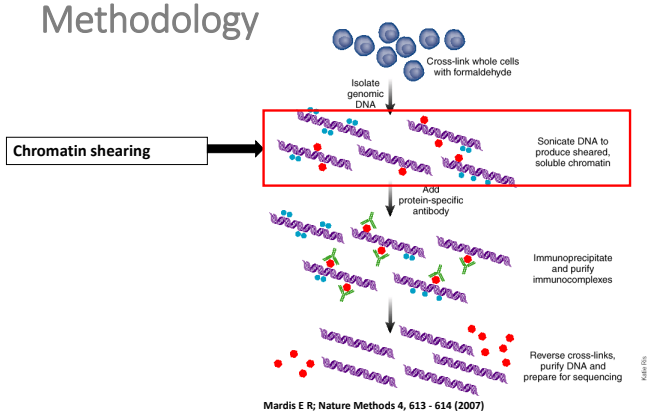
- Identification of precise regulatory sites across the genome
- Computation of recognized DNA sequence motifs
- Determination of downstream targets of transcription factors
- Clustering of multiple regulatory proteins at specific genomic positions
- Determination of chromatin states and DNA modifications

Methodology



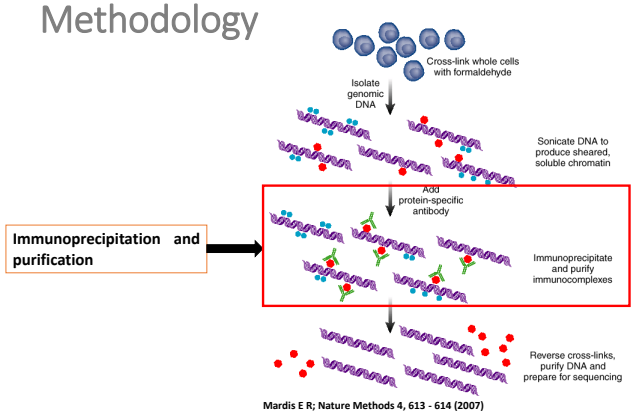
Mardis E R; Nature Methods 4, 613 - 614 (2007)

Methodology



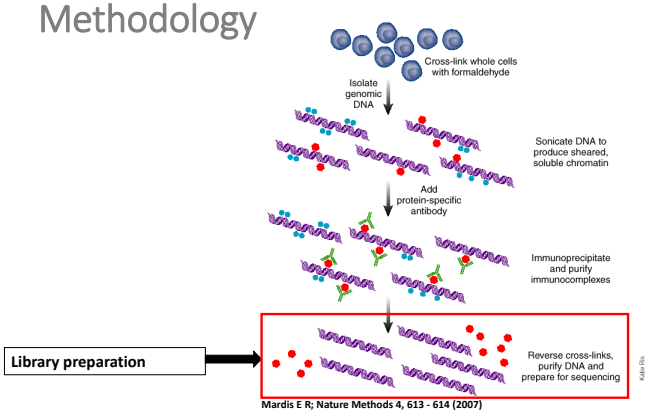
Mardis E R; Nature Methods 4, 613 - 614 (2007)

Methodology



Mardis E R; Nature Methods 4, 613 - 614 (2007)

Methodology



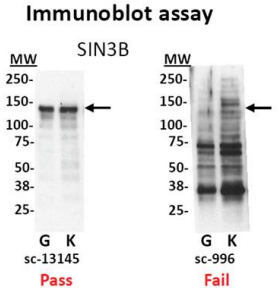
Mardis E R; Nature Methods 4, 613 - 614 (2007)

The importance of controls

Antibody and immunoprecipitation specificity

Antibody flaws:
- Poor reactivity against a chosen target protein
- Cross-reactivity with other DNA-binding proteins

ENCODE standard: primary reactive band contains >**50%** of the signal observed on the blot.



Landt SG et al. **ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia**. Genome Res. 2012 Sep;22(9):1813-31. doi:10.1101/gr.136184.111

The importance of controls

Control sample

Critical for the experiment due to:

- Non-uniform chromatin shearing- regions of open chromatin are preferentially represented
- Repetitive regions with tendency to accumulate more sequencing tags
- Platform-specific efficiency bias (eg. GC-bias)

The importance of controls

Control sample

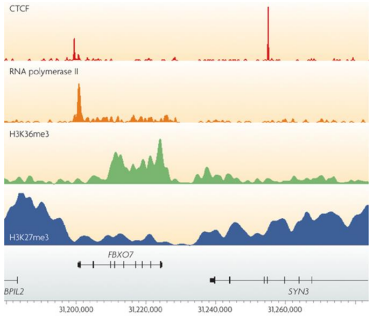
Different approaches:

1. „Input" DNA control- cross-linked and fragmented under the same conditions as the immunoprecipitaded sample
2. „Mock" IP control- using antibody against non-nuclear antigen

Experimental design

Distinct modes of interaction of DNA-binding proteins:

- *Point source* factors: sequence specific TFs, cofactors, transcription start site or enhancer-associated histone marks
- *Broad source* factors: chromatin marks and chromatin proteins associated with transcriptional elongation or repression
- *Mixed-source* factors



Park PJ, Nature Reviews Genetics 10, 669-680 (2009) Nature Reviews | Genetics

Experimental design

Sequencing depth

- *Point source*: mammals: mammals: at least 20 million reads per factor in two replicates, ENCODE./worms and flies: minimum of 2 million uniquely mapped reads per replicate, ENCODE/
- *Broad source*: mammals: up to 60 million reads / worms and flies: 5 million uniquely mapped reads per replicate, ENCODE/

Nakato R, Shirahige K**. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation**. Brief Bioinform. 2016 Mar 15

Experimental design

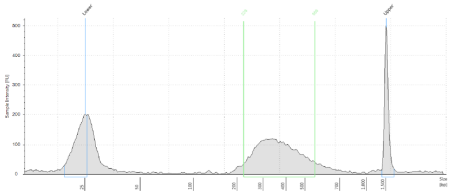
Read Length

36-100 bases

Library Type

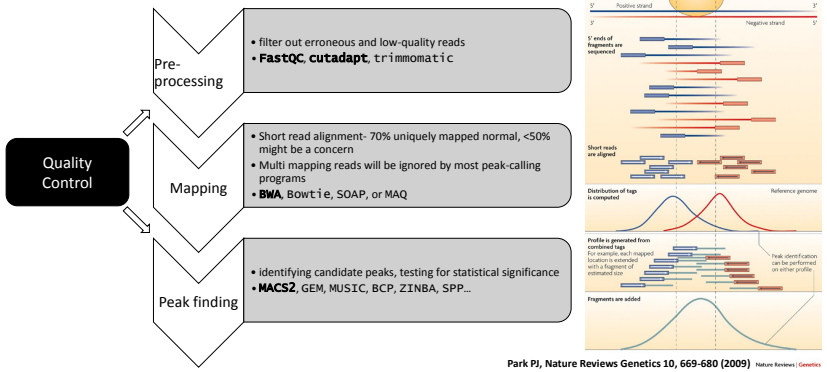
Single-end

(if particularly interested in improving the signal strength in repetitive regions paired-end reads recommended)



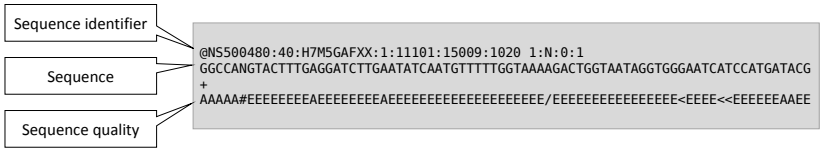
Nieścierowicz K, Unpublished

ChIP-Seq analysis workflow

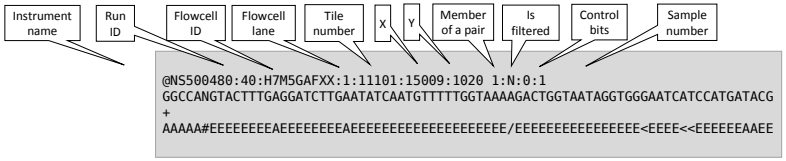


Park PJ, Nature Reviews Genetics 10, 669-680 (2009) Nature Reviews | Genetics

File formats: FASTQ



File formats: FASTQ



Quality:

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

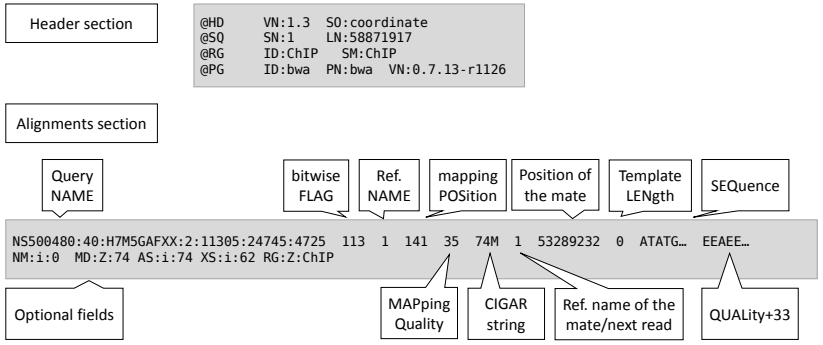
e- the estimated probability of an incorrect base

ASCII encoding:

Sanger/Illumina 1.8+: characters Q_{phred} + 33

Solexa/Illumina 1.3-1.7: Q_{phred} + 64

File formats: SAM



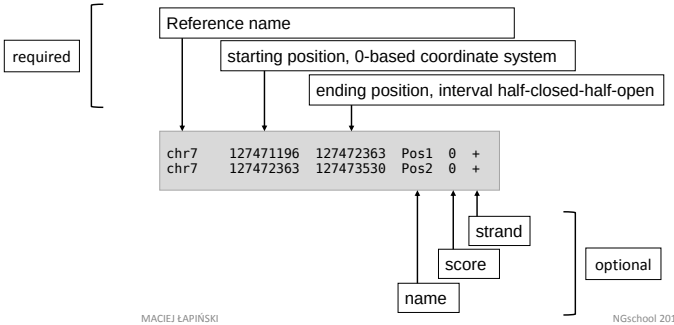
File formats: BAM

Compressed, binary version of sam.

- Space efficiency: BGZF compression- block compression on top of standard gzip file format
- Efficient random access for the indexed queries

File formats: BED

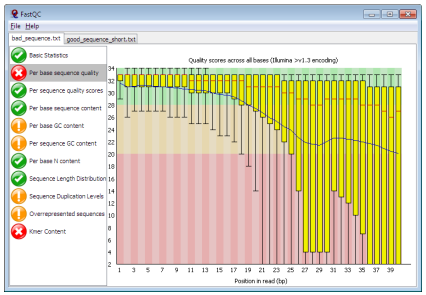
Browser Extensible Data format



Sequencing quality metrics

FastQC- quality control of raw sequence data

cutadapt- adaper and quality trimming, read filtering



Short read alignment

Maping low-divergent sequences against a large reference genome.

- Burrows-Wheeler Alignment tool (BWA)
- String matching using Burrows–Wheeler Transform (BWT)
- Same principle: SOAPv2, Bowtie
- Inexact matching algorithm
- Capable of gapped alignment of single reads

Li H. and Durbin R. (2009) **Fast and accurate short read alignment with Burrows-Wheeler transform**. Bioinformatics, 25, 1754-1760.

Resource

ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia

Stephen G. Landt,[1,26] Georgi K. Marinov,[2,26] Anshul Kundaje,[3,26] Pouya Kheradpour,[4](...)

Genome Res. 2012 Sep

Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation

Ryuichiro Nakato and Katsuhiko Shirahige

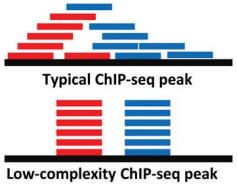
Brief Bioinform. 2016 Mar 15

Library complexity

- Measured as the fraction of nonredundant mapped reads (**NRF**) in a data set

$$NRF = \frac{N_{nonred}}{N_{all}}$$

- NRF decreases with sequencing depth
- ENCODE: NRF ≥0.8 for 10M uniquely mapped reads

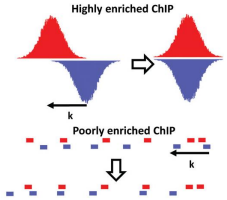


Cross-correlation analysis

Highquality ChIP-seq experiment produces significant clustering of enriched DNA sequence tags at locations bound by the protein of interest, and that the sequence tag density accumulates on forward and reverse strands centered around the binding site.

Sequence tags are positioned at a distance from the binding site center that depends on the fragment size distribution

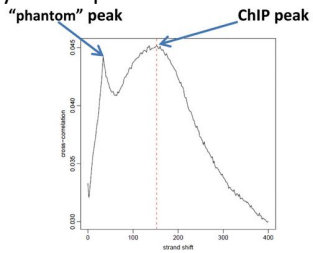
Lack of pattern of shifted stranded tag densities



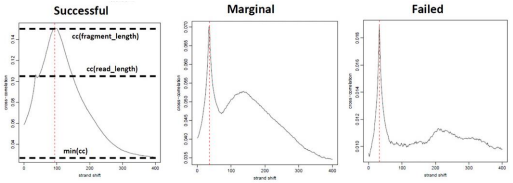
Cross-correlation analysis

Correlation between genome-wide stranded tag densities is computed as the Pearson linear correlation between the Crick strand and the Watson strand after shifting Watson by *k* base pairs.

Typically this produces two peaks when cross-correlation is plotted against the shift value: a peak of enrichment corresponding to the predominant fragment length and a peak corresponding to the read length (''phantom'' peak).



Cross-correlation analysis



$$NSC = \frac{cc(fragment\ length)}{min(cc)} \qquad RSC = \frac{cc(fragment\ length) - min(cc)}{cc(read\ length) - min(cc)}$$

Normalized strand coefficient (**NSC**)- the normalized ratio between the fragment-length crosscorrelation peak and the background cross-correlation

Relative strand correlation (**RSC**)- the ratio between the fragment length peak and the read-length peak

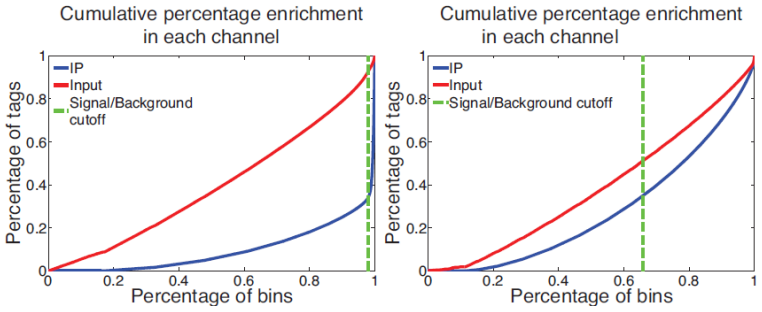
ENCODE: NSC >**1.05** and RSC >**0.8** for *point source* TFs

Signal extraction scaling

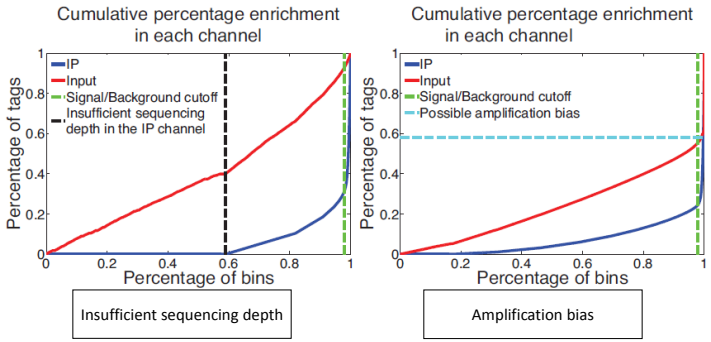
1. Partition the genome into non-overlaping windows
2. Count the number of alignments that fall within a given window
3. Order the windows based on the tag count
4. Plot the percentage of tags that fall into the given percentage of ordered windows

Diaz A, Nellore A, Song JS. **CHANCE: comprehensive software for quality control and validation of ChIP-seq data**. Genome Biology. 2012;13(10):R98. doi:10.1186/gb-2012-13-10-r98.

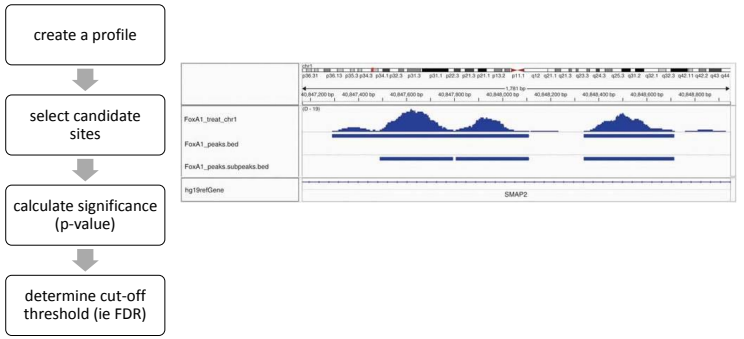
Signal extraction scaling



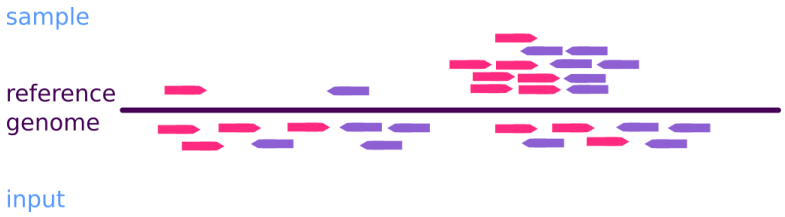
Signal extraction scaling



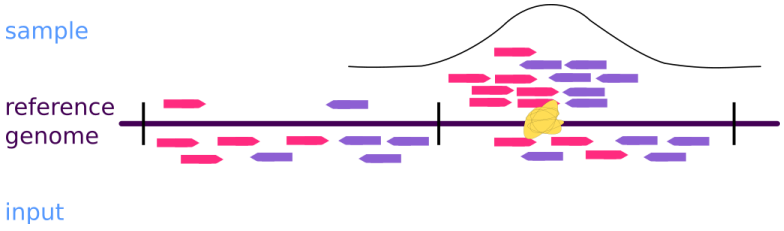
Peak calling



Peak calling



Peak calling



Peak calling

Selecting the best peak caller for a given application:

- *Point source* factors: **GEM**, MACS2, BCP, SPP, ZINBA…
- *Broad source factors*: **BCP, ZINBA**, **MUSIC**, BroadPeak, MACS2, ZINBA, SPP…

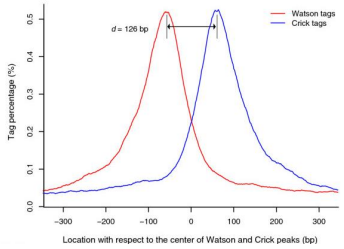
Reuben Thomas, Sean Thomas, Alisha K. Holloway, and Katherine S. Pollard **Features that define the best ChIP-seq peak calling algorithms**. Brief Bioinform 2016 : bbw035v1-bbw035.

MACS2

Decide the fragment length ***d***

Empirical modeling of '*d*' and tag shifting by d/2 to putative protein-DNA interaction site.

'*d*' is used to extend the ChIP-sample and compute ChIP coverage



MACS2

Build local bias track from control

Tag distribution along the genome is modeled by a Poisson distribution, described by the average numer of events:

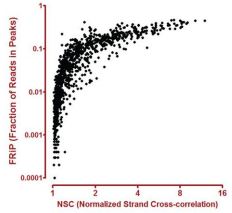
$$\lambda_{local} = \max(\lambda_{BG}, [\lambda_{1k},] \lambda_{5k}, \lambda_{10k}),$$

where λ_{1k}, λ_{5k} and λ_{10k} are λ estimated from the 1 kb, 5 kb or 10 kb window centered at the peak location in the control sample

Global ChIP enrichment (FRiP)

- Fraction of Reads in Peaks
- Positive and linear correlaton with the number of called regions
- ENCODE: **> 1%**, most successful point-source factor- FRiP values of 0.2–0.5 and NSC/RSC values of 5–12.

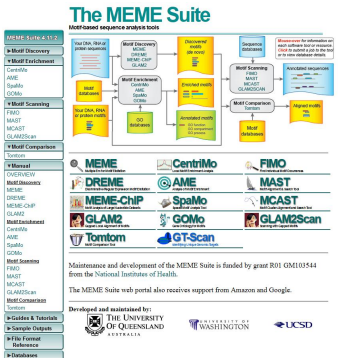
Correlatio between FRiP an NSC for 1052 human ChiP-Seq experiments



MEME- novel motifs discovery

- Discover novel sequence motifs
- Scan your sequences with a given motif
- Analyse the similarity to known motifs

http://meme-suite.org



Gene set enrichment analysis

- **GREAT**- annotate non-coding regions, http://bejerano.stanford.edu/great/public/html/
- **DAVID**- functional annotation, https://david.ncifcrf.gov/

Thank you!

Bisulphite sequencing (Lecture & Workshop)

Russell Hamilton
Centre for Trophoblast Research, University of Cambridge

Thursday, 14:00

1 Introduction

1.1 What is methylation?

The most common form of methylation occurs at the 5' position of cytosines through the addition of a methyl group. In mammals this is most common at CpG sites, but also occurs at other cytosine positions. Some prokaryotes, such as *E. coli*, have pentameric methylation sites.

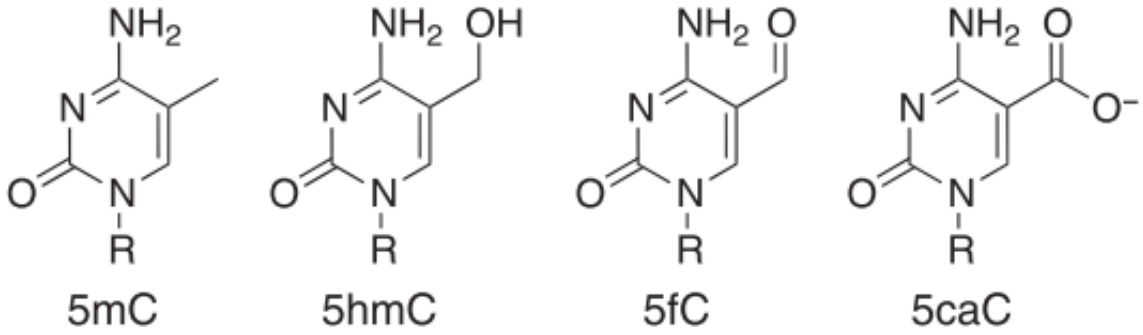


Figure 1: Cytosine Modifications. Structures of 5'-methyl-cytosine (5-mC), 5'-hydroxymethyl-cytosine (5-hmC), 5'-formyl-cytosine (fC) and 5'carboxy-cytosine(caC). [Figure from (Booth *et al.*, 2013)]

1.2 Identifying methylation

There are several techniques for assaying single base methylation levels. Cost and tissue availability are often limiting factors therefore reduced genome methods are very popular.

- Introduce bisulfite bismark (Krueger and Andrews, 2011)
- Introduce RRBS
- Introduce 450K / EPIC
- Introduce hmC calling BS/ox-bs subtraction
- Introduce TAB-Seq

1.3 What is hydroxymethyl-cytosine?

Introduce hmC and its biological significance, also highlight coverage requirement very costly 30x as subtraction required

1.4 What is formyl-cytosine and carboxy-cytosine?

Introduce fC, caC and their biological significance

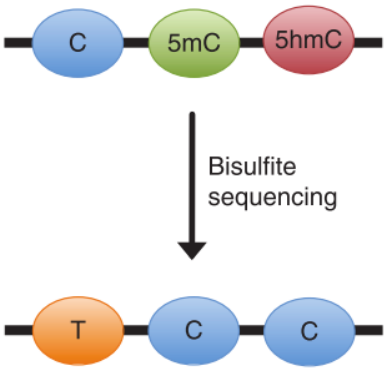


Figure 2: Bisulfite Sequencing. Cytosines are converted to thymine, 5-methyl-cytosine and 5-hydroxymethyl-cytosine are protected from conversion and are read as cytosine in sequencing. [Figure from (Booth *et al.*, 2013)]

2 Analysis

The recommendation for bisulfite sequencing and oxidative bisulfite sequencing to be able to call single base resolution 5-hmC is 30x. Therefore this is going to be an expensive experiment, requiring careful consideration of samples, replicates and tissue types.

2.1 Pipeline

To processing a large number of samples in a consistent, documented and reproducable manner it is advisable to use a pipeline system. Pipelines can be custom bash scripts, docker containers or specific pipeline tools such as clusterflow.io. Exact versions and command line options should be recorded in log files.

```
#fastqc
#trim_galore
        #bismark_align
                #bismark_deduplicate
                      #preseq_lc_extrap
                      #preseq_bound_pop
                      #qualimap_bamqc
                      #picard_insert_size_metrics
                      #featureCounts
                      #bismark_methXtract
                                #bismark_report
                                        >bismark_summary_report
>multiqc
```

Listing 1: Bismark Clusterflow Pipeline

```
$ cf --genome <PATH/TO/GENOME/INDEX> pipeline_name *.fq.gz
```

Listing 2: Calling the Bismark Clusterflow Pipeline

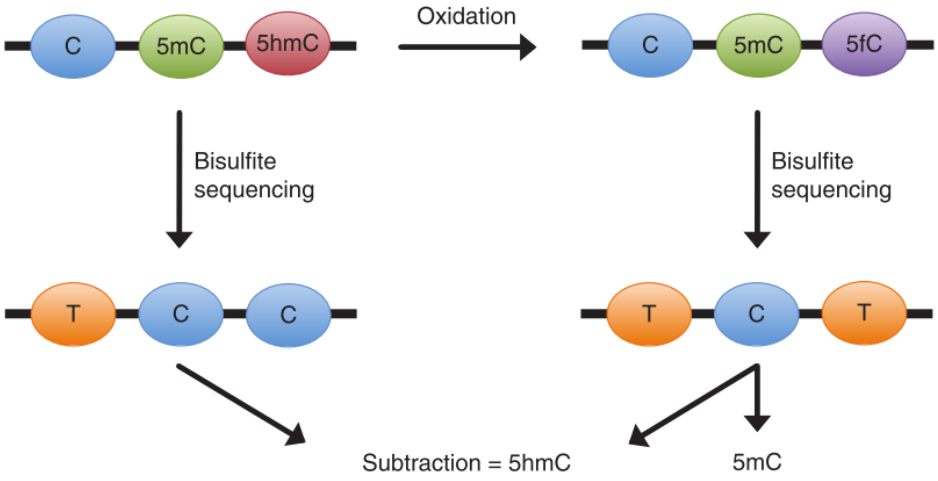


Figure 3: Oxidative bisulfite sequencing. Cytosines are converted to thymine, 5-methyl-cytosine is protected from conversion and is read as cytosine in sequencing. 5-hydroxymethyl-cytosine is converted to 5-formyl-cytosine during oxidation and then to thymine in bisulfite. A subtraction is required to read the 5-hydroxymethyl-cytosine component.[Figure from (Booth *et al.*, 2013)]

2.2 Pre Alignment Quality Control

Fastqc for the fastq files provides a comprehensive assessment of the sequencing quality and the adapter contamination.

2.3 Alignment

The alignment of bisulfite treated samples is broadly the same for whole genome, targetted and reduced-representation bisulfite sequencing (RRBS). There are specific options in bismark worth noting (RRBS, pbat, directional).

Bismark alignment of PBAT samples:

```
$ bismark /path/to/reference/GRCm38_Lambda/ --pbat \
    -1 read_1.fq.gz -2 read_2.fq.gz
```

Listing 3: Bismark alignment: pbat samples

The `--pbat` option is the only non-default option used as this is required as the samples were prepared with post-bisulfite adapter tagging. This is an attempt to reduce the number of fragments lost due to fragmentation, but adding the adapters after bisulfite sequencing (Miura *et al.*, 2012).

Example bismark Alignment Report:

```
Final Alignment report
======================
Sequence pairs analysed in total:   23456857
Number of paired-end alignments with a unique best hit:     16059532
Mapping efficiency: 68.5%

Sequence pairs with no alignments under any condition:      6471470
```

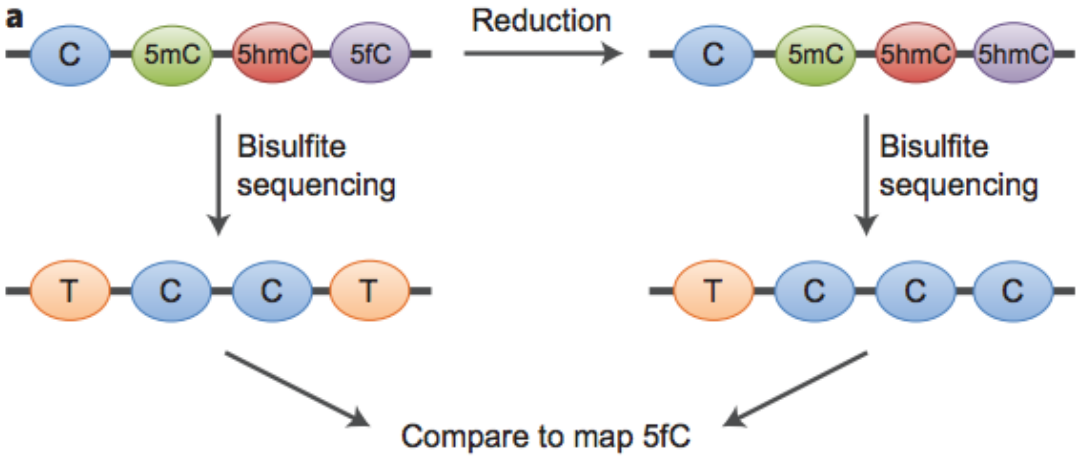


Figure 4: Reduced bisulfite sequencing. Cytosines are converted to thymine, 5-methyl-cytosine and 5-hydroxymethyl-cytosine are protected from conversion ands read as cytosine in sequencing. 5-formyl-cytosine is converted to thymine in bisulfite treatment, and converted to 5-hydroxymethyl-cytosine during reduction, read as cytosine. A subtraction is required to read the 5-formylmethyl-cytosine component. [Figure from (Booth *et al.*, 2014)]

```
Sequence pairs did not map uniquely:        925855
Sequence pairs which were discarded because genomic sequence could not
be extracted:        5

Number of sequence pairs with unique best (first) alignment came from the
bowtie output:
CT/GA/CT:   0       ((converted) top strand)
GA/CT/CT:   7983330 (complementary to (converted) top strand)
GA/CT/GA:   8076197 (complementary to (converted) bottom strand)
CT/GA/GA:   0       ((converted) bottom strand)

Final Cytosine Methylation Report
=================================
Total number of C's analysed:       896797899

Total methylated C's in CpG context:        25511027
Total methylated C's in CHG context:        1667723
Total methylated C's in CHH context:        6550281
Total methylated C's in Unknown context:    0

Total unmethylated C's in CpG context:      10166967
Total unmethylated C's in CHG context:      211368803
Total unmethylated C's in CHH context:      641533098
Total unmethylated C's in Unknown context:  22

C methylated in CpG context:        71.5%
C methylated in CHG context:        0.8%
C methylated in CHH context:        1.0%
C methylated in unknown context (CN or CHN):        0.0%
```

Listing 4: Example bismark Alignment Report

Bismark alignment of Lux control samples. As they are short, and with known methylation state, the alignments must be intollerant to mismatches.

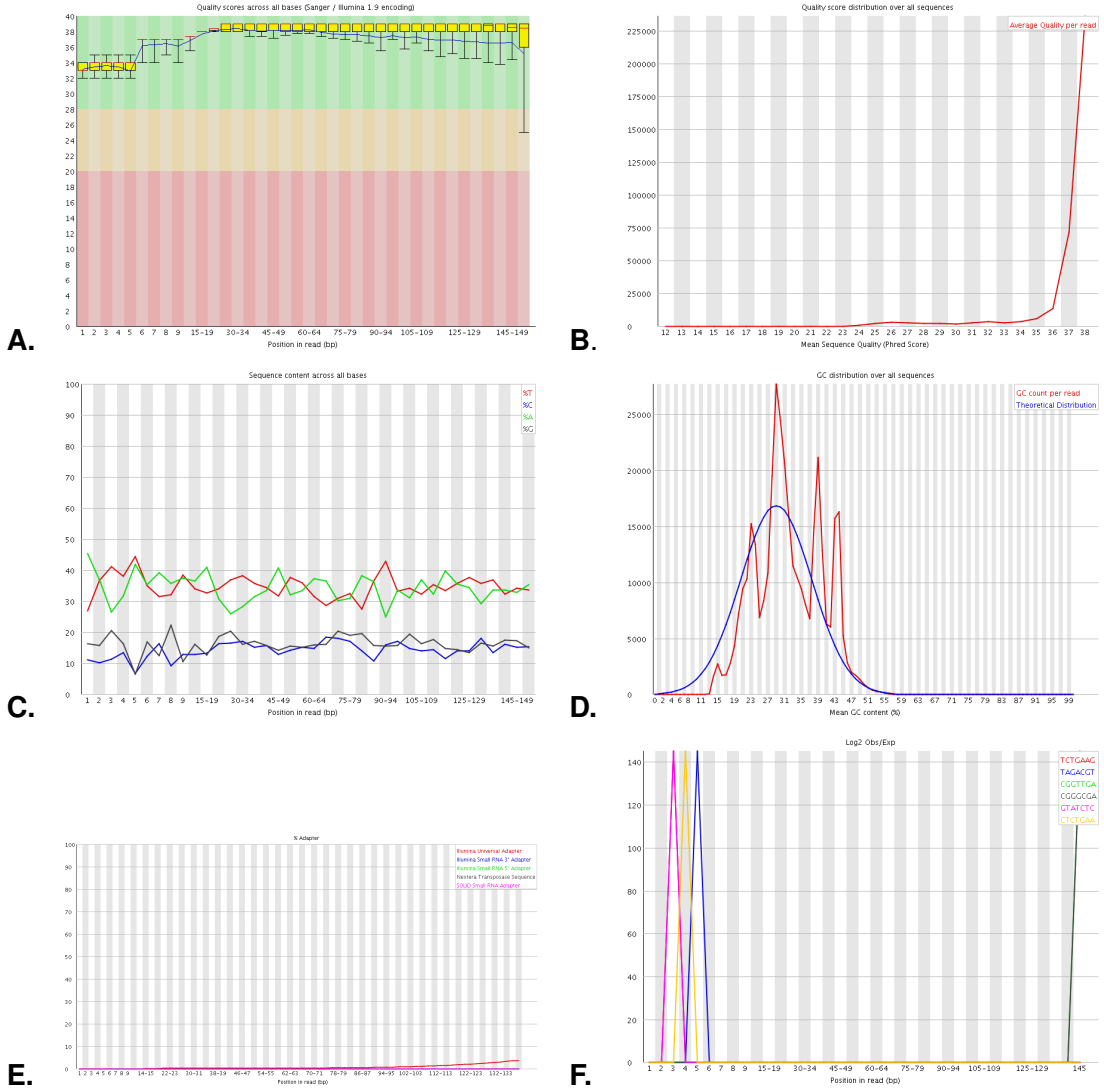


Figure 5: FastQC Metrics. A. Per base quality scores. B. Quality scores. C. Per base sequence content. D. GC content. E. Adapter content. F. Kmer content

```
$ bismark /path/to/reference/GRCm38_Lambda/ -I 0 -X 2000 -N 0 \
    --non_directional -1 read_1.fq.gz -2 read_2.fq.gz
```

Listing 5: Bismark alignment: lux controls

- `-I 0` : minimum insert size of zero - no overlapping R1 / R2
- `-X 2000` : maximum insert size of 2000nt (default is 500nt)
- `-N 0` : number of allowed mismatches
- `--non_directional` : selected for non directional library preps (not current illumina protocols)

Typically a bismark alignment of approx 70% is an to be expected. Below this there could be issues such as adapter contamination. The duplication rate and fragmentation are factors influencing the alignment rate.

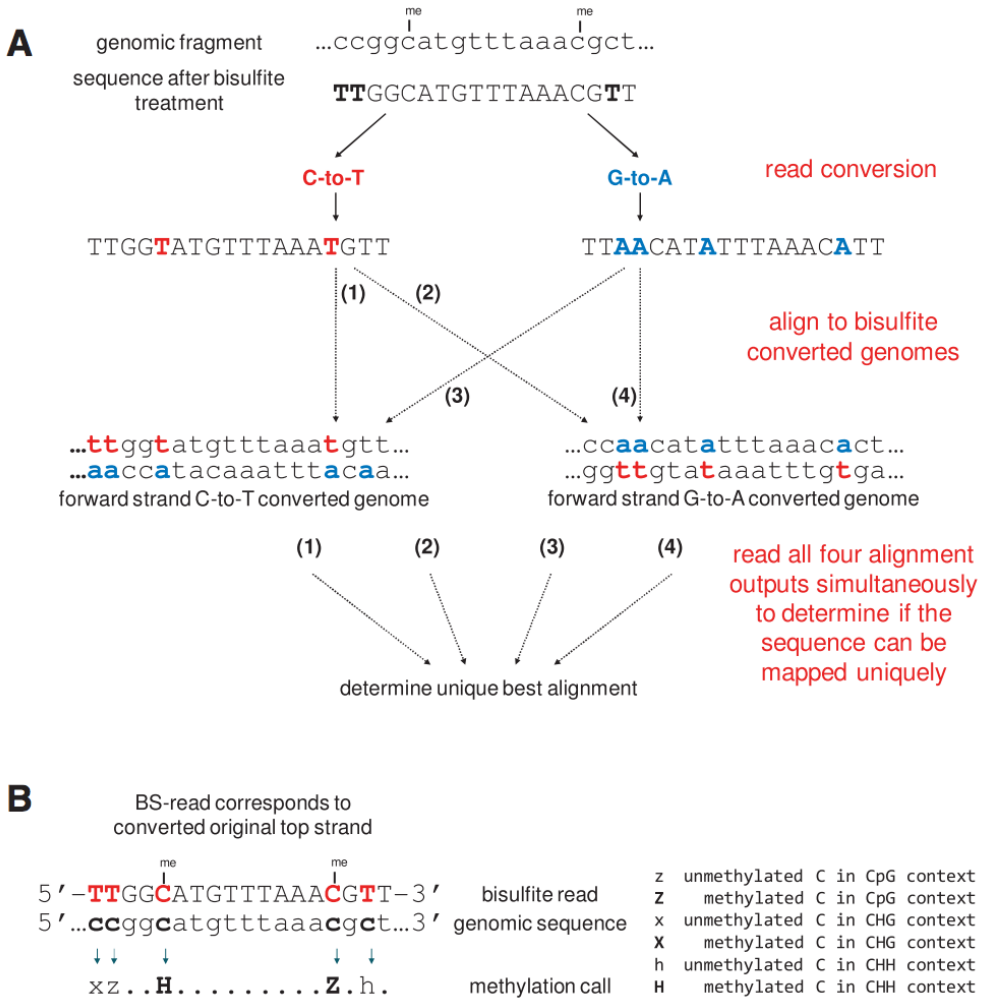


Figure 6: Aligning Bisulfite Sequencing Reads with Bismark. [Figure from (Krueger and Andrews, 2011)]

2.4 Post Alignment Quality Control

Bismark produces name sorted bam files to be compatible with the methylation extractor. To perform the post alignment QC, most analysis tools require coordinate sorted so an extra re-sort step is required.

```
$ samtools sort -o sample.coordsrt.bam sample.bam
```

Listing 6: Samtools coordinate sort

2.4.1 QualiMap

In particular, check the GC content, we are losing a base so important to check conversion accuracy.

```
$  qualimap bamqc -bam sample.bam -outfile result.pdf
```

Listing 7: Qualimap bamqc

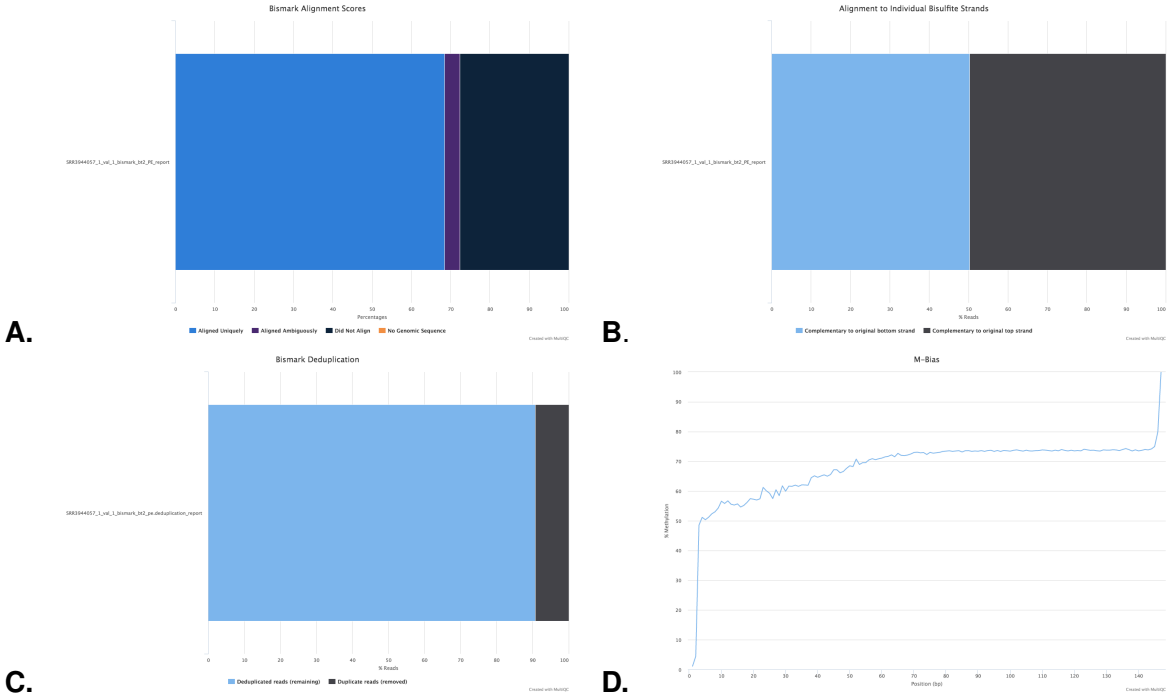


Figure 7: Bismark Alignment Metrics. A. Percentage alignment. B. Strand bias. C. Bismark Deduplication. D. Bismark M-Bias

2.4.2 Preseq

Saturation curves give an indication of how much sequencing is required in light of the bisulfite induced fragmentation.

```
$ preseq lc_extrap -e 1000000000 -P -l 999999999999 -v -Q -B sample.bam
```

Listing 8: Preseq library complexity estimation

2.4.3 Picard Insert Size Metrics

Due to bisulfite treatment causing fragmentation it is crucial to check the PE insert sizes. Picard, like Qualimap, calculates insert sizes of paired end data.

```
$ java -jar picard.jar CollectInsertSizeMetrics \
     I=sample.bam O=insert_size_metrics.txt \
     H=insert_size_histogram.pdf
```

Listing 9: Picard insert size metrics

2.5 Deduplication

Bismark includes tools for deduplication, based on identical genomic mapping. The advantage of using this tool rather than e.g. `samtools dedup` is that is is fully compatible with the bismark name sorted alignments.

```
$ deduplicate_bismark -p --bam sample_1_val_1_bismark_bt2_pe.bam
```

Listing 10: Bismark deduplication

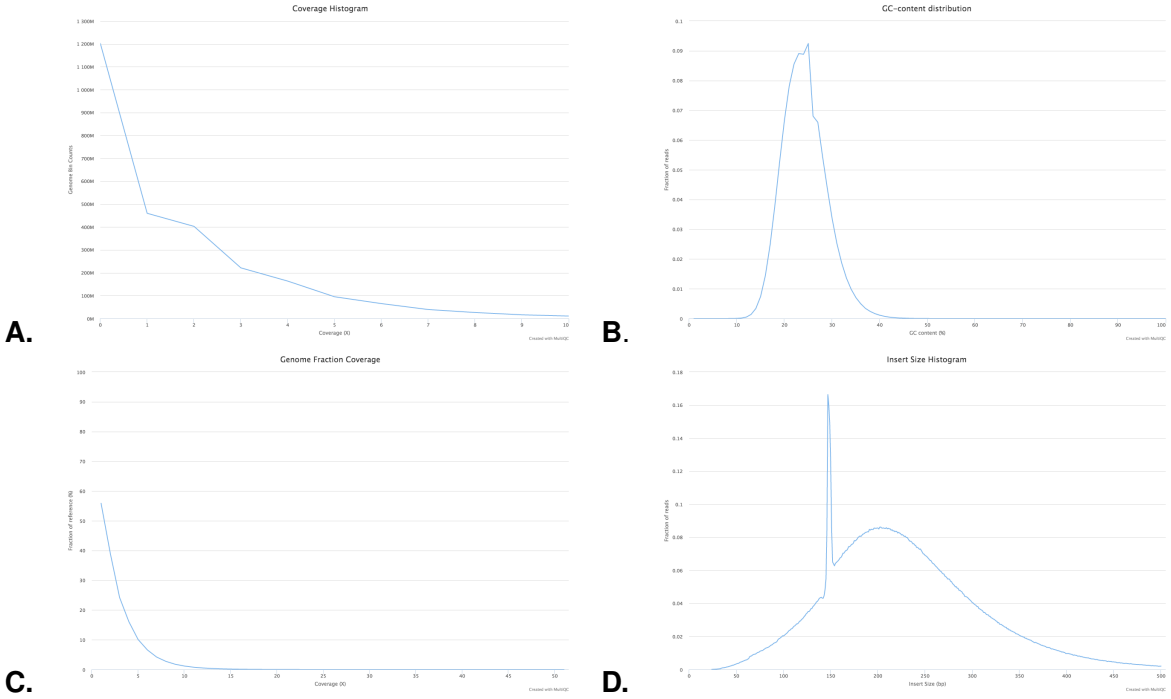


Figure 8: Qualimap bamqc metrics. A. Coverage Histogram. B. GC-Content. C. Geneome Fraction Coverage. D. Insert Size

2.6 Spike In Controls

To assess the conversion efficiency of the bisulfite, oxidation and reduction treatments it is advisable to include synthetic spike in controls with known methylation states in to the sequencing. These can either be designed by the individual research group, or companies like Cambridge Epigenetix include controls in the kits they sell. Analysing the controls, often due to their short length, can be non trivial so CEGX provide analysis software for assessing the methylation conversion rates of their controls.

In the Lux paper, (Äijö *et al.*, 2016), they designed their own controls (See table below)

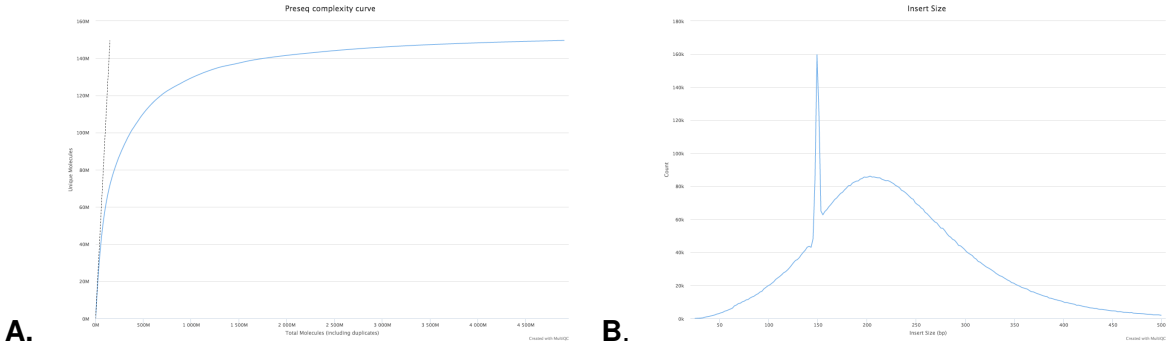


Figure 9: Preseq estimate library complexity and Picard insert size metrics. A. Percentage alignment. B. Strand bias.

Software	*Version*	*Link*
FastQC	v0.11.5	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
trim_galore	v0.4.1	http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
samtools	1.3.1	http://www.htslib.org/download/
bowtie2	recent	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
bismark	0.16.1	https://github.com/FelixKrueger/Bismark/releases
Qualimap	2.2	http://qualimap.bioinfo.cipf.es/
preseq	2.0	http://smithlabresearch.org/software/preseq/
multiqc	v0.8dev	pip install git+https://github.com/ewels/MultiQC.git
R-Studio	recent	https://www.rstudio.com/products/RStudio/
R	recent	https://www.r-project.org/
Lux	recent	https://github.com/tare/Lux/
methyl-kit	v0.99.2	https://github.com/al2na/methylKit
picard	recent	https://broadinstitute.github.io/picard/

Table 1: Prerequisite Software

2.7 Methylation Calling

2.7.1 Bismark methylation extractor

Bismark comes packaged with its own methylation extractor, with the ability to call methylation in all cytosine environments, not just CpG. Also a variety of reports, and levels of verbosity can be specified.

```
$ bismark_methylation_extractor --multi 4 --ignore_r2 1 \
   --ignore_3prime_r2 2 --bedGraph --counts --gzip -p \
   --no_overlap --report sample.deduplicated.bam
```

Listing 11: Bismark methylation extraction

2.8 DMR Calling

- Introduce principals od DMR calling
- Focus on Methyl-Kit (Akalin *et al.*, 2012)

3 Prerequisites

3.1 Software Required

3.2 Reference Genome

The reference genome needs to be prepared for bisulfite sequencing. In the case of bismark the top and bottom strands need to be converted C to T and G to A and then indexed with bowtie2. In the example data sets, both are from mouse and in the Lux data Lamba derived spike in controls are used. Therefore for this practical a custom genome of GRCm38 with Lambda is used.

```
$ bismark_genome_prepare --bowtie2 /path/to/GRCm38_Lambda
```

Listing 12: Bismark genome preparation

	Replicate	Treatment	Sample Name
Female Aged 3 months	1	mkBS	SRR3944074:M4_1813
	2	mkBS	SRR3944075:M5_1815
	3	mkBS	SRR3944076:M6_1817
	1	oxBS	SRR3944064:M4_1813
	2	oxBS	SRR3944065:M5_1815
	3	oxBS	SRR3944066:M6_1817
Female Young 24 months	1	mkBS	SRR3944057:M1_1801
	2	mkBS	SRR3944058:M2_1805
	3	mkBS	SRR3944069:M3_1808
	1	oxBS	SRR3944061:M1_1801
	2	oxBS	SRR3944062:M2_1805
	3	oxBS	SRR3944063:M3_1808
Male Aged 24 months	1	mkBS	SRR3944080:M3_1813
	2	mkBS	SRR3944059:M4_1815
	3	mkBS	SRR3944060:M6_1817
	1	oxBS	SRR3944071:M3_1813
	2	oxBS	SRR3944072:M4_1815
	3	oxBS	SRR3944073:M6_1817
Male Young 3 months	1	mkBS	SRR3944077:M1_1801
	2	mkBS	SRR3944078:M2_1805
	3	mkBS	SRR3944079:M5_1808
	1	oxBS	SRR3944067:M1_1801
	2	oxBS	SRR3944068:M2_1805
	3	oxBS	SRR3944070:M5_1808

Table 2: Data from (Hadad *et al.*, 2016)

3.3 Data Sets Required

3.3.1 Aging Data

Paper (Hadad *et al.*, 2016) DOI:http://dx.doi.org/10.1186/s13072-016-0080-6

3.3.2 Amplicon Data

Amplicons from the Lux paper (Äijö *et al.*, 2016) DOI:http://dx.doi.org/10.1186/s13059-016-0911-6
Other sequences are often spike into the prep to increase library diversity due to the loss of cytosines - in this case lambda, but PhiX also used. Spike in controls with known methylation status are also used.

Note: Align to GRCm38 and Lambda simultaneously

4 Discussion / Concluding Remarks

Due to fragmentation and the statistical power required for bs/oxBS subtraction samples need to be sequenced at very high depth (30x). Also due to the dual treatment, double the

Sample	Replicate	Treatment	Sample Name
Tet2 KO	1	mkBS	SRR2009038:Tet2_KO_mkbs
	2	mkBS	SRR2009039:Tet2_KO_mkbs
	3	mkBS	SRR2009040:Tet2_KO_mkbs
	1	oxBS	SRR2009041:Tet2_KO_oxbs
	2	oxBS	SRR2009042:Tet2_KO_oxbs
	3	oxBS	SRR2009043:Tet2_KO_oxbs
v6.5 KD	1	mkBS	SRR2009044:v6.5_KD_mkbs
	2	mkBS	SRR2009045:v6.5_KD_mkbs
	3	mkBS	SRR2009046:v6.5_KD_mkbs
	1	oxBS	SRR2009047:v6.5_KD_oxbs
	2	oxBS	SRR2009048:v6.5_KD_oxbs
	3	oxBS	SRR2009049:v6.5_KD_oxbs

Table 3: Data from (Äijö *et al.*, 2016)

amount of sample is required, this may be a limiting factor in cases with low tissue availability (FFPE). Bisulfite and oxidative bisulfite treatments to get single base resulution methylcytosine and/or hydrodymethylcytosine are therefore expensive experiment. Some of the reduced genome approaches may therefore be more appropriate e.g. RRBS, 450K/EPIC. However these too heve their limitations (number of CpGs covered).

Molecular data integration (Lecture & Workshop)

Jacek Marzec
Barts Cancert Institute, London

Friday, 9:00

Overview

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

http://www.freeimages.com

Why integrate the data?

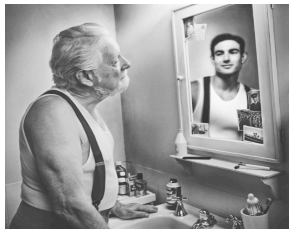
- Multitude of microarray and NGS technologies are used for gene expression profiling
- Produced data are stored in public repositories:
  - NCBI Gene Expression Omnibus (GEO)
  - EMBL-EBI ArrayExpress
  - NCBI/EBI Sequence Read Archive (SRA)
- …or international consortia data portals:
  - International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

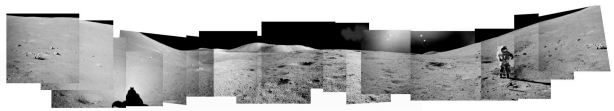


➢ These can be addressed by systematic integrative analysis performed on larger patient cohorts

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



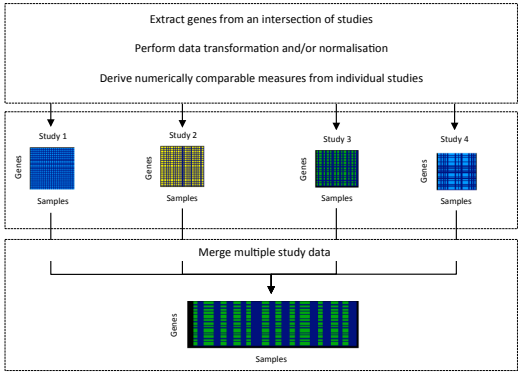
https://www.hq.nasa.gov

Approaches

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

Approaches

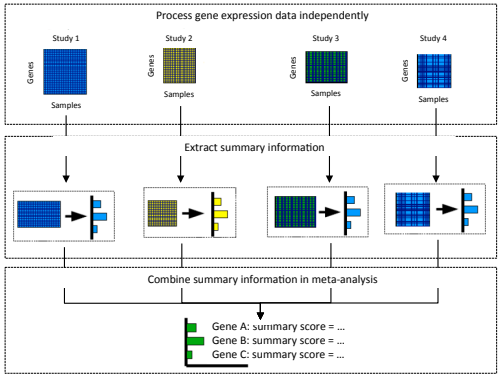
1. Integrating data at the expression level



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

Approaches

2. Combining ranked lists (meta-analysis)



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

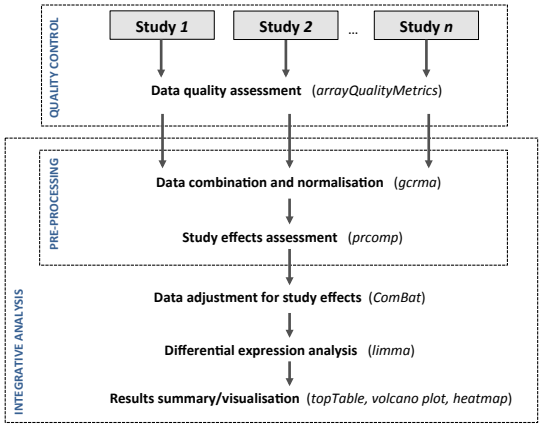
https://www.r-project.org/about.html

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
2. Normalisation: allows comparisons to be made between measurements from different samples
3. Summarisation: aggregates intensity values of multiple probes in a given probeset to a single expression value

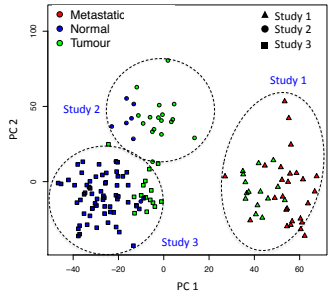
https://www.bioconductor.org/packages/release/bioc/html/gcrma.html

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

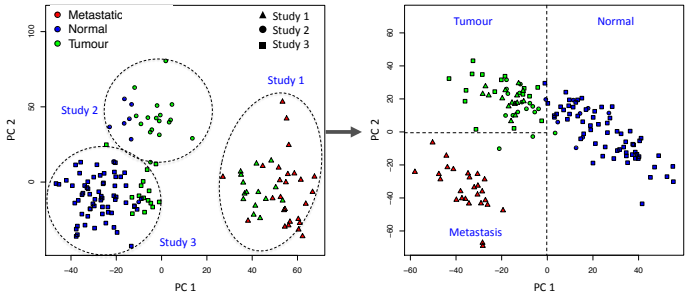
Chen C, Grennan K, Badner J *et al*. *PLoS ONE*, 2011;6(2):e17238

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)

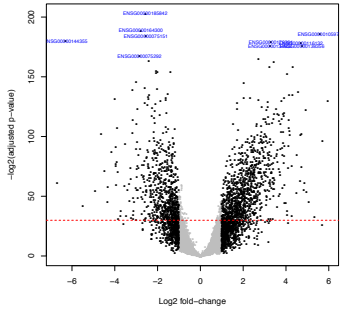
- Extracts table of the top-ranked genes from linear model fit

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

'*volcano plot*' (*plot* function in *graphics* package)

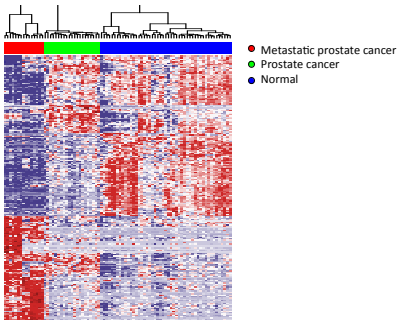
- Plots significance (p-value) versus fold-change values
- Helps to quickly identify changes in large data sets



Results visualisation

'*heatmap.2*' (*gplots* package)

- Reflects gene expression values across samples representing various conditions



NGS & Biomedicine (Lecture & Workshop)

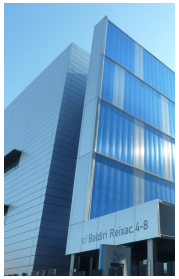
Sophia Derdak
Centro Nacional de Análisis Genómico, Barcelona

Saturday, 9:00

cnag centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

The CNAG

- Situated in the Parc Científic de Barcelona
- Staff > 60, > 50% informatics
- Directed by Ivo Gut
- Agilent 's Certified Service Provider in Spain – Target Enrichment System (since 2014)
- Illumina Certified Service Provider (since 2013)
- ISO 9001 quality certification (since 2014)
- ISO 17025 accreditation (in process)



Mission

Carry out projects in genome analysis that will lead to significant improvements in people's health and quality of life, in collaboration with the Spanish, European and International Research Community.

Research interests

- Cancer Genomics
- Disease Gene Identification and Infectious Diseases
- Agrogenomics and Model Organisms

Partners in International and National Projects

- ICGC, GEUVADIS, BLUEPRINT, SYBARIS, RD-CONNECT, ESGI, AirPROM, EVA, READNA, ...
- Citrus, Melon, Olive, Lynx, Primates, Mouse, Drosophila,...

What are genomic variants?

Variants in the diploid genome

Identification of genetic differences in comparison to a reference

Reference (haploid) TGGACCATCTGGTTGAGCATGTGGGGGTCAACTCCCACATTCCCAGGGAGCCCCCGG

The true diploid genome of the sample

TGGACCATCTGGTTGAGCATGTGGGGGTCAACTTCCACATTCCCAGGGAGGCCCCGG
TGGACCATCTGGTTGAGCACGTGGGGGTCAACTTCCACATTCCCAGGGAGCCCCCGG

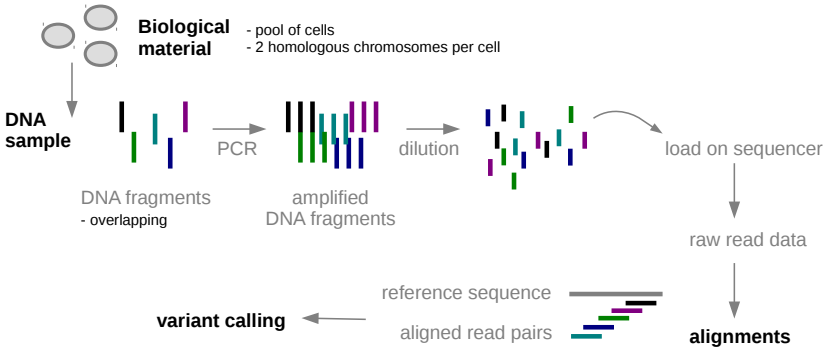
ref/ref	ref/alt	alt/alt	ref/alt
0/0	0/1	1/1	0/1
homozygous reference	heterozygous	homozygous alternative	heterozygous

>99% of the genomic positions are **not** variant positions

~3.700.000 variant positions / 3.200.000.000 base position human genome

<100.000 variant positions / 51.000.000 base position human **exome**

Genome sequencing: the experimental workflow

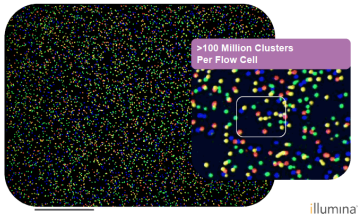


Genomic Variants in Biomedicine NGSchool 2016, August 20, 2016 Sophia Derdak, CNAG

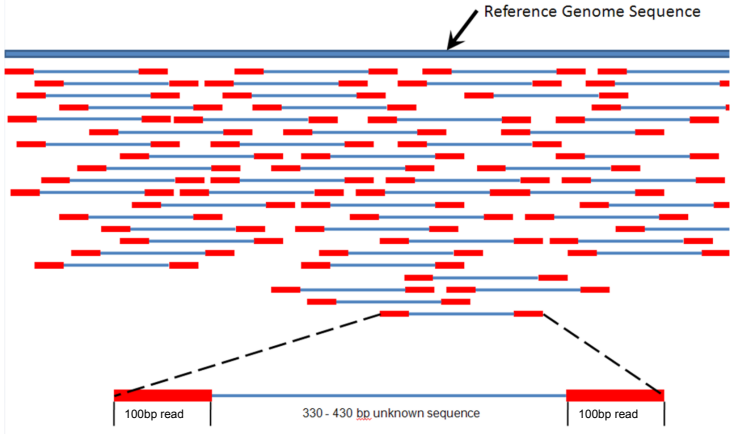
Sequencing-by-synthesis

From flowcell to computer: Base calling

- The sequence of colors read for each cluster in each cycle is translated to nucleotide sequence



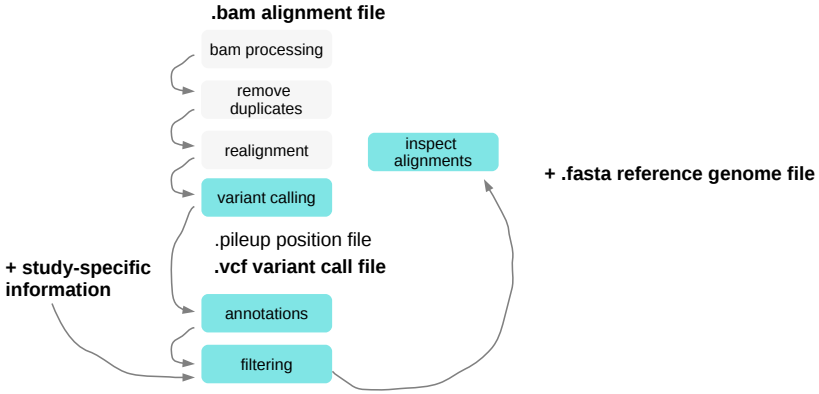
Mapping of reads to the reference sequence



(adapted from wikipedia)

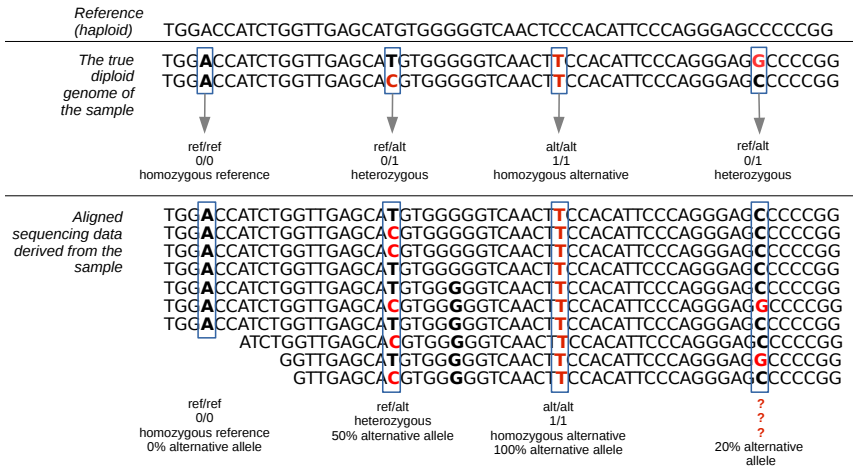
Genomic Variants in Biomedicine NGSchool 2016, August 20, 2016 Sophia Derdak, CNAG

From alignments to variants: the bioinformatic workflow



Variant calling: Variants in the sequencing data

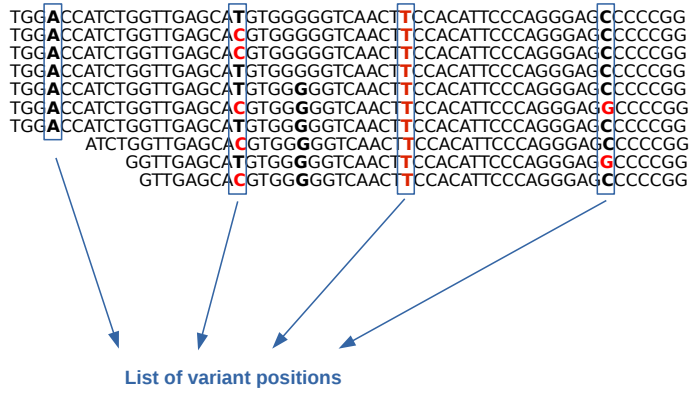
Identification of genetic differences in comparison to a reference



Genomic Variants in Biomedicine NGSchool 2016, August 20, 2016 Sophia Derdak, CNAG

Variant calling: Extract variants only

Aligned sequencing data derived from the sample

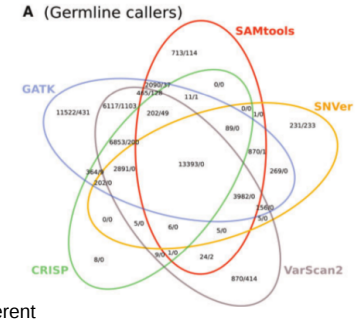


List of variant positions

Altmann A et al. Hum Genet 2012: A beginners guide to SNP calling from high-throughput *DNA-sequencing data.*

Bioinformatic tools for single nucleotide variant calling

- samtools + bcftools (Sanger Institute, UK, and Broad Institute, US)
- Genome Analysis Tool Kit (GATK) (Broad Institute, US)
- VarScan (Washington University)
- Platypus (Welcome Trust Center, UK)
- freebayes (Boston College, US)



Keep in mind that different software use different algorithms and thresholds and results may vary **A LOT**.

Pabinger S et al. Briefings in Bioinformatics 2013: *A survey of tools for variant analysis of next-generation genome sequencing data.*

Benchmarks for variant calling

- NA12878 50x Whole Genome FASTQs from Illumina Platinum Genomes analyzed with the pipeline:
http://www.illumina.com/platinumgenomes/

- Results compared independently for SNPs and INDELs agains NIST reference set:
Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Zook et al. Nat Biotechnol. 2014 Mar;32(3):246-51.

- Results (on reliably callable region = 70% of the genome);

Feature	Mapper	Variant Caller	TP	FP	FN	Specificity	Sensitivity
SNVs	GEM3	GATK-HC	2738414	7009	2318	0.9974	0.9992
Deletions	GEM3	GATK-HC	84783	1349	1176	0.9843	0.9863
Insertions	GEM3	GATK-HC	83189	784	1394	0.9907	0.9835

cnag

S. Laurie, S. Derdak, R. Tonda, S. Beltran

CRG

The coverage

represents the number of times a base of the sample genome (or target region) is read during sequencing.

A higher coverage provides higher power for data analysis.

How to get a higher coverage:

- mainly by loading more sequencing units (indexes, lanes, entire flowcells) with the same library preparation

Typical coverage numbers (in CNAG projects):

- whole genome: 30x
- exome: 50-100x
- custom gene panel capture: >1000x

Sequencing data analysis is all probabilities!

"I believe that we do not know anything for certain, but everything probably."
Christiaan Huygens

- base calling (base qualities in the fastq files)
- contig order in the reference assembly
- reference sequence (not yet...)
- **read alignment (mapping quality)**
- **variant position (variant and genotype quality)**

Plato, ~400 BC

- p-values
- probability likelihoods
- PHRED scores

Single sample variant results

raw vcf file ("all variants")

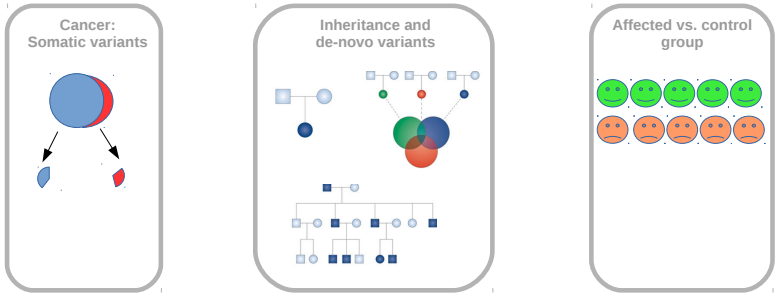
mostly experiment-independent
technical and quality filtering (well-covered positions with confident alternative allele)

filtered vcf file ("good quality variants")

CHR	POS	REF	ALT	GT
1	148588972	G	C	0/1
1	154284894	A	G	0/1
1	203923829	A	G	0/1
1	243329075	T	C	0/1
2	102968362	T	C	0/1
2	122096456	G	A	0/1
2	242612151	C	T	1/1
3	56591283	TAAGCAGGGG	TAAGCAGGGGGAAGCAGGGG	0/1
4	146297387	CAAAAAAAA	CAAAAAAAAAA	0/1
6	116263181	T	C	1/1
8	96070181	T	C	1/1
9	129831659	T	A	1/1
9	131454120	C	T	1/1
10	29834095	G	A	1/1
11	18159254	A	G	1/1
11	35274829	A	G	0/1
12	21628320	C	T	0/1
15	42619508	C	T	1/1
15	75336729	A	G	0/1
16	84035844	T	C	0/1

CHR chromosome
POS position on the chromosome
REF sequence in the reference genome
ALT alternative sequence detected in the sample
GT genotype in the (diploid) sample

Variant calling: multi-sample analyses



Compare two samples of the same individual (e.g. tumor-normal)

vcf file ("good quality variants - all genotypes")

Definition of "somatic variant", consider sample purity information
Select variants with genotype 0/0 in the normal and 0/1 in the tumor sample
Additionally, select alternative allele frequency thresholds for normal and tumor sample using AC

filtered vcf file ("somatic variants")

CHR	POS	REF	ALT	GT_normal	AC_normal	GT_tumor	AC_tumor
1	1421916	T	C	0/0	37,1	0/1	44,7
1	179528803	A	C	0/0	53,0	0/1	53,16
1	59096853	C	T	0/0	19,1	0/1	21,7
2	132236963	C	T	0/0	11,0	0/1	21,5
2	166756497	T	A	0/0	20,1	0/1	23,8
3	53910122	G	A	0/0	28,0	0/1	29,7
7	151962062	A	G	0/0	12,0	0/1	19,5
9	136083801	A	G	0/0	10,0	0/1	16,5
11	89407177	C	T	0/0	15,0	0/1	31,6
12	129298780	G	A	0/0	19,1	0/1	30,6
15	20454042	A	T	0/0	22,1	0/1	24,5
15	23113683	T	C	0/0	12,0	0/1	4,6
17	15468718	G	A	0/0	11,0	0/1	16,5
17	15468728	T	C	0/0	11,0	0/1	19,5
17	36365191	A	T	0/0	11,0	0/1	18,4
17	66195635	T	G	0/0	12,0	0/1	12,5
19	43783125	C	A	0/0	10,0	0/1	24,5
19	43783146	A	T	0/0	13,0	0/1	29,8
20	29628070	T	C	0/0	13,0	0/1	22,4
21	11181025	G	A	0/0	39,0	0/1	36,10

CHR chromosome
POS position on the chromosome
REF sequence in the reference genome
ALT alternative sequence detected in the sample
GT genotype in the (diploid) sample, per sample
AC allele count, number of (ref, alt) bases, per sample

Compare three samples of a pedigree

vcf file (“good quality variants - all genotypes”)

Apply model of inheritance: e.g. autosomal recessive
Select variants with genotype 0/1 in the parents and 1/1 in the daughter

filtered vcf file (“recessively inherited variants”)

CHR	POS	REF	ALT	GT_daughter	GT_father	GT_mother
1	200827638	A	G	1/1	0/1	0/1
1	22158157	A	G	1/1	0/1	0/1
2	171256597	A	C	1/1	0/1	0/1
2	208976955	A	C	1/1	0/1	0/1
4	48496368	A	G	1/1	0/1	0/1
7	14017007	C	T	1/1	0/1	0/1
8	143310815	G	A	1/1	0/1	0/1
8	41517860	G	A	1/1	0/1	0/1
11	47437403	C	T	1/1	0/1	0/1
11	59837097	C	T	1/1	0/1	0/1
12	9833628	C	T	1/1	0/1	0/1
13	36699762	G	A	1/1	0/1	0/1
13	52523808	C	T	1/1	0/1	0/1
14	38256944	T	C	1/1	0/1	0/1
15	79026001	C	A	1/1	0/1	0/1
16	10788129	G	T	1/1	0/1	0/1
16	1498197	A	G	1/1	0/1	0/1
19	49640002	G	T	1/1	0/1	0/1
20	10026357	T	C	1/1	0/1	0/1
22	23657980	G	A	1/1	0/1	0/1

CHR chromosome

POS position on the chromosome

REF sequence in the reference genome

ALT alternative sequence detected in the sample

GT genotype in the (diploid) sample, per sample

Compare three samples of a pedigree

vcf file (“good quality variants - all genotypes”)

Apply model of inheritance: e.g. de-novo
Select variants with genotype 0/0 in the parents and 0/1 in the daughter

filtered vcf file (“de-novo variants”)

CHR	POS	REF	ALT	GT_daughter	GT_father	GT_mother
1	13365778	A	G	0/1	0/0	0/0
1	144676632	T	C	0/1	0/0	0/0
1	144853029	C	G	0/1	0/0	0/0
1	16891333	C	T	0/1	0/0	0/0
3	143697451	T	G	0/1	0/0	0/0
3	72311749	G	T	0/1	0/0	0/0
3	9057481	C	A	0/1	0/0	0/0
5	39002519	C	T	0/1	0/0	0/0
6	33060143	A	G	0/1	0/0	0/0
6	37845185	C	A	0/1	0/0	0/0
7	1586741	G	T	0/1	0/0	0/0
8	49987965	A	C	0/1	0/0	0/0
9	39888209	C	A	0/1	0/0	0/0
12	92562268	G	T	0/1	0/0	0/0
12	94034134	G	T	0/1	0/0	0/0
13	19042019	G	T	0/1	0/0	0/0
15	84855648	C	A	0/1	0/0	0/0
16	46427389	T	C	0/1	0/0	0/0
17	10550780	A	G	0/1	0/0	0/0
22	20643742	C	A	0/1	0/0	0/0

CHR chromosome

POS position on the chromosome

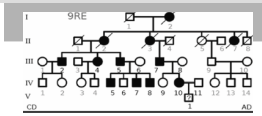
REF sequence in the reference genome

ALT alternative sequence detected in the sample

GT genotype in the (diploid) sample, per sample

Genomic Variants in Biomedicine NGSchool 2016, August 20, 2016 Sophia Derdak, CNAG

Causative variant for inherited retinal dystrophy?



chr17

… a real world success story of finding the causative variant

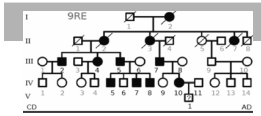
Discard variants because:

- they have low technical quality
- they are frequent polymorphisms in the general population
- they do not have a protein-coding effect

de Castro-Miró M et al. PLOS One 2014: *Combined Genetic and High-Throughput Strategies for Molecular Diagnosis of Inherited Retinal Dystrophies.*

Genomic Variants in Biomedicine NGSchool 2016, August 20, 2016 Sophia Derdak, CNAG

Causative variant for inherited retinal dystrophy?



chr17:7918347, T>C, 0/1

chr17

annotations at gene level	GUCY2D	• ENSEMBL Functional annotations: genes, transcripts, coding sequences
	Retina	• UCSC genome browser, GeneCards... Tissue specificity of gene function

de Castro-Miró M et al. PLOS One 2014: *Combined Genetic and High-Throughput Strategies for Molecular Diagnosis of Inherited Retinal Dystrophies.*

Causative variant for inherited retinal dystrophy?

chr17:7918347, T>C, 0/1

chr17

annotations at gene level	GUCY2D	• ENSEMBL Functional annotations: genes, transcripts, coding sequences
	Retina	• UCSC genome browser, GeneCards... Tissue specificity of gene function

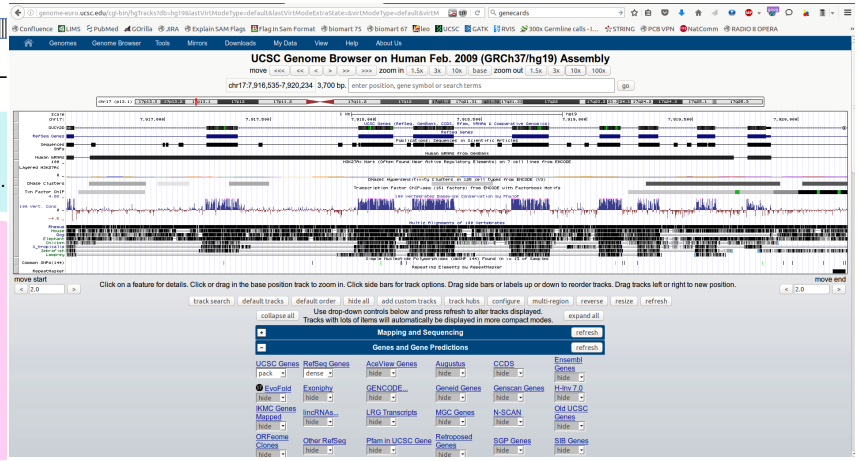
annotations at position level	c.2747T>C p.I916T	• base change • amino acid change
	Variant not annotated	• ExAC (> 60.000 exomes) general population frequency
	damaging probably damaging	• Deleteriousness predictions: · SIFT · PolyPhen2 · CADD

de Castro-Miró M et al. PLOS One 2014: *Combined Genetic and High-Throughput Strategies for Molecular Diagnosis of Inherited Retinal Dystrophies.*

Genomic Variants in Biomedicine NGSchool 2016, August 20, 2016 Sophia Derdak, CNAG

Causative variant for inherited retinal dystrophy?

Look up a gene in a Genome Browser:



Genomic Variants in Biomedicine NGSchool 2016, August 20, 2016 Sophia Derdak, CNAG

More genomics resources

Variants **inside candidate genes or genomic regions** are interesting variants

HGMD:: Human Gene Mutation Database (Cardiff University and Biobase GmbH)

OMIM :: Online Mendelian Inheritance in Man (John Hopkins University)

Orphanet :: The portal for rare diseases and orphan drugs (INSERM, France)

ClinVar :: Information about relationships among variation and human health (NCBI)

Disease-specific databases and publications (e.g. COSMIC database for cancer)

Genetic linkage studies

→ Helpful, when studying a case with a previously described disease phenotype

The OMIM database is available and may be queried at: http://omim.org/

The Orphanet database is available at: http://www.orpha.net/consor/cgi-bin/index.php

ClinVar is available at: http://www.ncbi.nlm.nih.gov/clinvar/

The COSMIC Catalogue for somatic mutations in cancer is available at: http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/

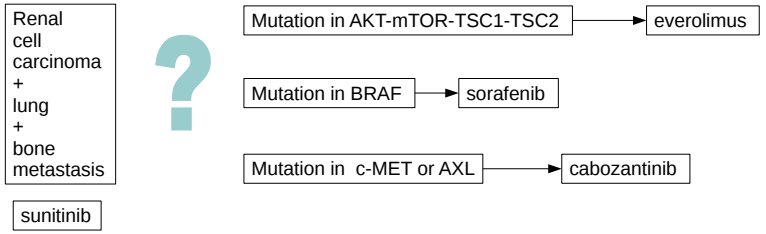
What else can genomic variants tell us?

Genomic Variants in Biomedicine NGSchool 2016, August 20, 2016 Sophia Derdak, CNAG

Somatic variants in cancer

… may help to decide on one therapy or another

Example from one of my collaborations (not clinical routine yet):



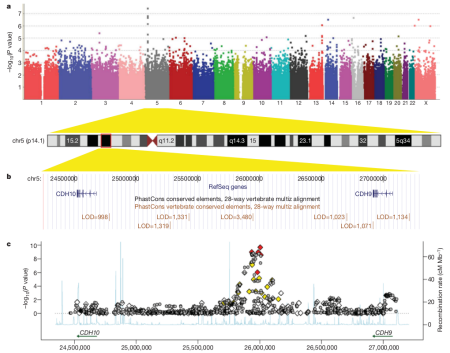
Bellmunt J et al. Clin Genitourin Cancer 2014: *Sequential targeted therapy after pazopanib therapy in patients with metastatic renal cell cancer: efficacy and toxicity.*

Complex diseases

… more complex than coding effect and inheritance

One of the methods to assess complex disease is GWAS – Genome Wide Association Studies.

- Look for genetic polymorphisms (not necessarily coding!) that associate with the trait
- in 1000's of samples: cases and controls, perform statistical tests
- Results can be single position or hotspot region around a position: "Manhattan plots"



e.g. Autism spectrum disorders

← highly associated polymorphisms on chr5

← zoom in

← the hotspot is in the intergenic region

← close-up of the hotspot

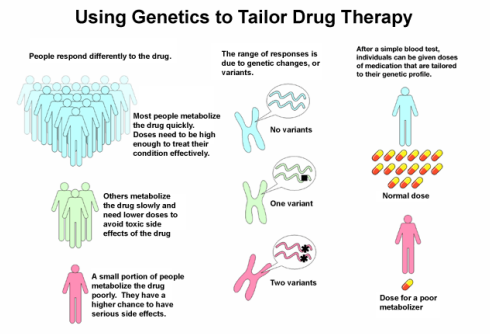
Wang K et al. Nature 2009: *Common genetic variants on 5p14.1 associate with autism spectrum disorders.*

Genomic Variants in Biomedicine NGSchool 2016, August 20, 2016 Sophia Derdak, CNAG
Genomic Variants in Biomedicine NGSchool 2016, August 20, 2016 Sophia Derdak, CNAG

Pharmacogenomics

Pharmacogenomics is an emerging field that combines genetics with pharmacokinetics and pharmacodynamics of drugs.

- to understand genetic polymorphisms among patients
- to study the effect of these polymorphisms on the activity of the enzyme metabolizing the drug
- to develop more accurate drug dosing in order to avoid intoxication or insufficient drug action.



drug	*Genes with variants affecting drug action*
Warfarin (inhibitor of blood coagulation)	VKORC1 and CYP2C9
Irinotecan (cancer)	UGT1A1
Thiopurine drugs (autoimmune disorders)	TPMT and ITPA

Lee JW et al. Clin Genet 2014: *The emerging era of pharmacogenomics: current successes, future potential,and challenges.*

Incidental findings

- originally coined in the field of radiology

A clinically relevant incidental DNA variation can be defined as a verified DNA variation that has a proven medically relevant phenotype not directly related to the condition being studied for research.

It is an unforeseen clinical finding relevant to the individual research participant involved (and possibly to the family of the participant).

- to be discussed in the field of bioethics

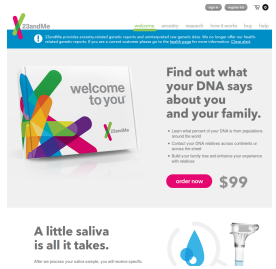
Should the participant (or the participant's physician) be informed about the incidental finding?

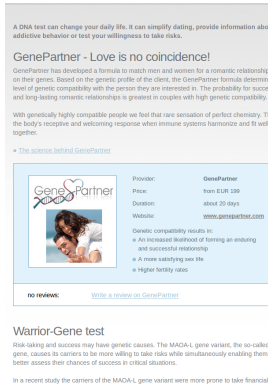
Does it make a difference whether the incidentally discovered genetic variant points at a disease with a therapy available or not?

Properly informed consent for the study participants must explain the possibility of finding an incidental DNA variation (especially in whole genome sequencing).

Krier JB and Green RC. Curr Protoc Hum Genet. 2013: *Management of Incidental Findings in Clinical Genomic Sequencing.*

Recreational genotyping?





~~Health reports~~

- ancestry-related genetic reports
- uninterpreted raw genetic data
- oddities:

Does fresh cilantro taste like soap to you?
Yes
No
Not sure

Eriksson N et al. arXiv 2012: *A genetic variant near olfactory receptor genes influences cilantro preference.*

Library construction for next-generation sequencing: overview and potential troubleshooting (Lecture)

Paulina Stachula
International Institute of Molecular and Cell Biology in Warsaw

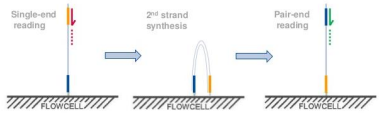
Wednesday, 18:00

Outline of the talk

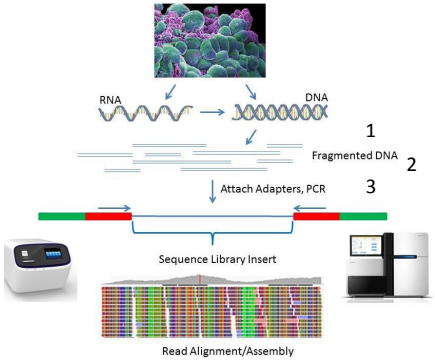
1. What is the NGS library and how to prepare it
2. From library preparation to computer analysis
3. Some problems occuring during library construction

What to consider when preparing library for NextGenerationSequencing ?

1. **Sample preparation: The amount of starting material** whether the application is for resequencing or de novo sequencing
2. **Quality of the material.**
3. **Problematic genomes with high or low GC content**
4. **Library preparation**
5. **Paired-end or single-end reads**



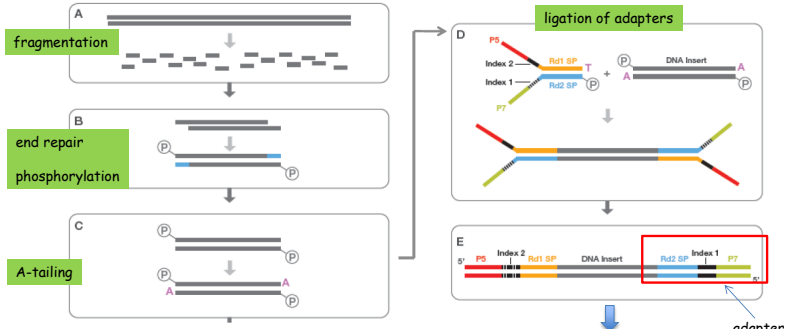
The core steps in library preparation



Head et al., 2014

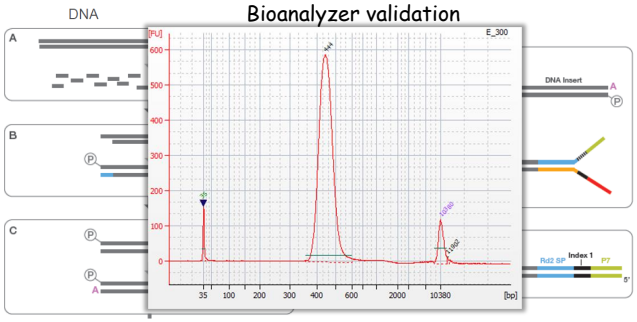
1. Fragmenting and/or sizing the target sequences to a desired length
2. Converting target to double-stranded DNA,
3. Attaching oligonucleotide adapters to the ends of target fragments,
4. Quantitating the final library product for sequencing.

Illumina library preparation workflow



- Purifies ligation products
- Removes unligated adapters and adapter dimers
- Selects library size-range for cluster generation

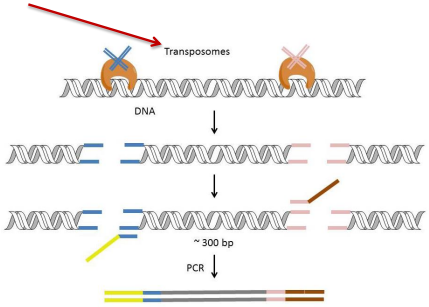
Standard DNA library workflow



starting material: genomic DNA, Chip-seq, amplicons.
amount: 100ng-1ug

Nextera DNA libraries

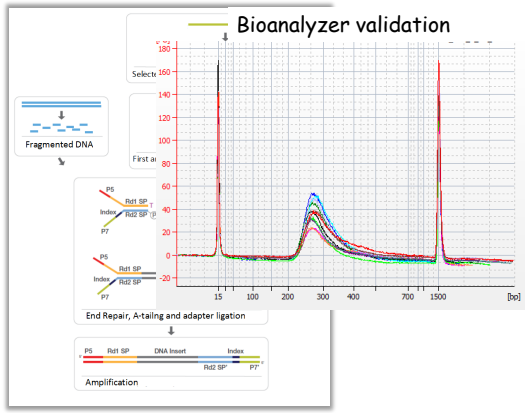
transposase enzyme simultaneously fragments and inserts adapter sequences into dsDNA.



Head et al., 2014

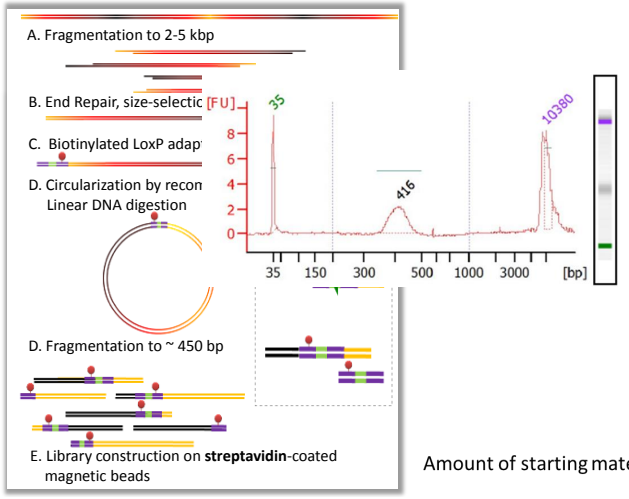
starting material: genomic DNA, Chip-seq, amplicons.
amount: 50ng

Standard RNA library workflow



starting material: different fraction of RNA
amount: 2ug before rRNA removal, 100ng

RNA mate-pair libraries



Amount of starting material: 10 µg

IIMCB

Library validation

Quality by Agilent Bioanalyzer/Tape Station

http://gc.nci.nih.gov/TapeStation.html

✓ Accurate determination of fragment size distribution
✗ Less reliable quantitation
✗ Requires expensive equipment

NanoDrop

✓ Low cost, as most laboratories already have access to UV/vis spectrophotometers
✗ Not specific for DNA
✗ Results can be skewed by RNA or protein contamination
✗ Cannot determine fragment size

✓ Low cost, as most laboratories already have access to fluorimeters
✓ Can quantitate specifically dsDNA, ssDNA, RNA or protein
✗ Quantitates all nucleic acid present in sample, not just molecules to be sequenced
✗ Cannot determine fragment sizes

Quantity with qPCR

http://www.bio-rad.com/en-cn/product/cfx96-touch-real-time-pcr-detection-system

✓ By using primers specific to the illumina **universal adapters** in a qPCR reaction containing library template, only cluster-forming templates will be amplified and quantified

✓ Most accurate quantitation method
✗ More expensive
✗ Cannot determine fragment sizes

IIMCB

The Illumina NGS workflows include 4 basic steps

Library Preparation | Cluster Formation | Sequencing | Computer Analysis

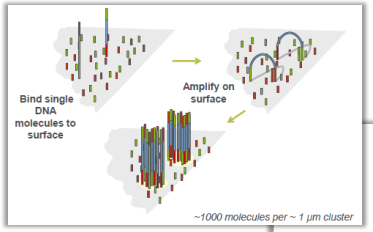
http://www.illumina.com/techniques/sequencing/ngs-library-prep.html

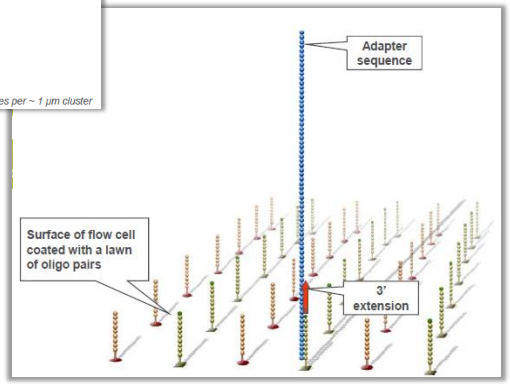
Library validation

http://www.illumina.com/systems/nextseq-sequencer/applications.html

IIMCB

Cluster generation

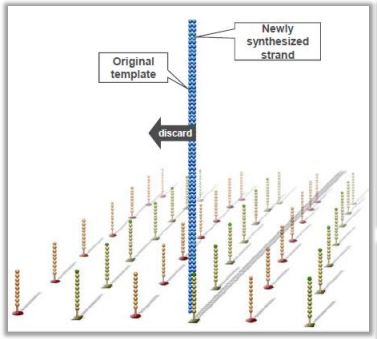


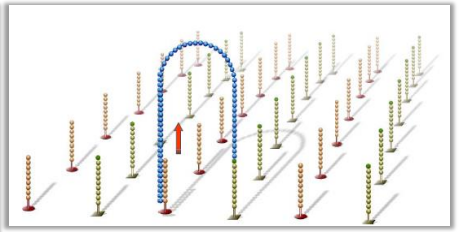


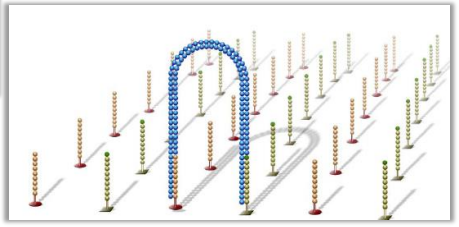
- Hybridization
- Amplification
- Linearization
- Blocking
- Primer hybridization

IIMCB

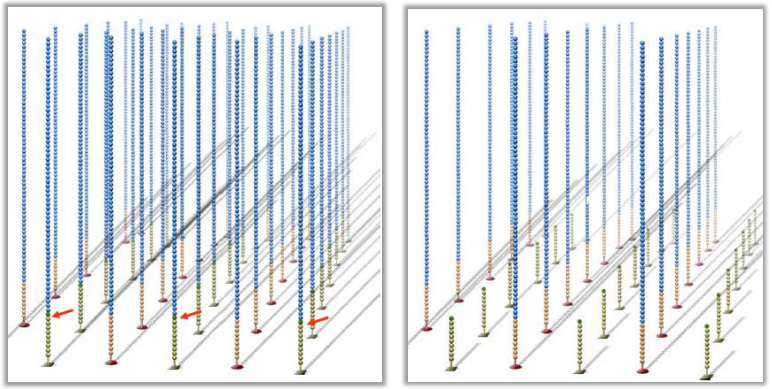
Cluster generation





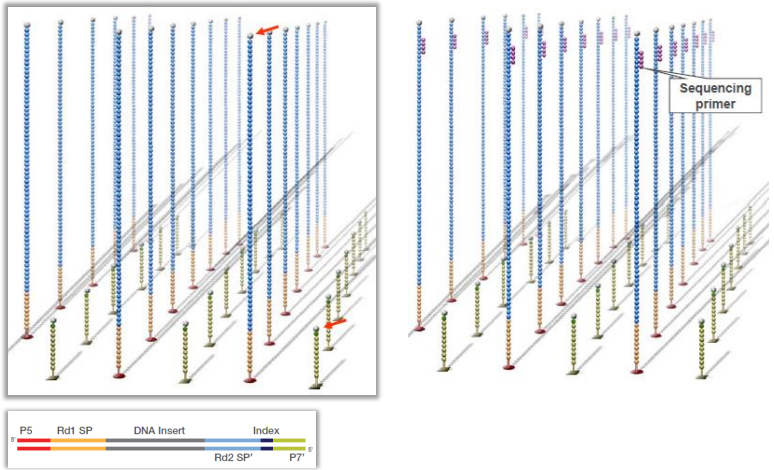


Preparation cluster for sequencing



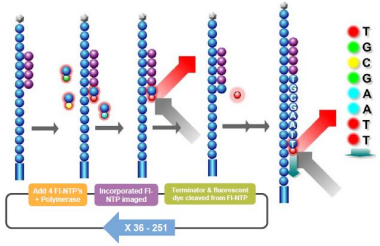
IIMCB

Preparation cluster for sequencing



IIMCB

Sequencing by synthesis

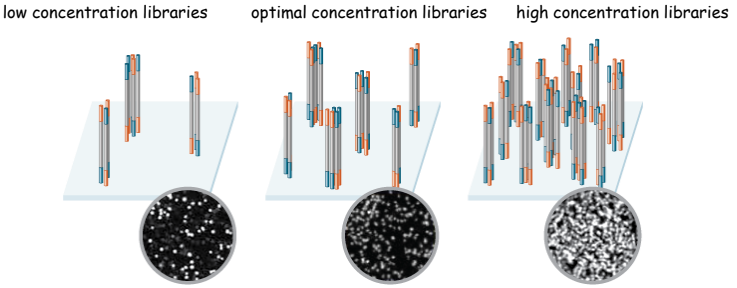


Illumina SBS technology

http://www.youtube.com/embed/HMyCqWhwB8E?iframe&rel=0&autoplay=1

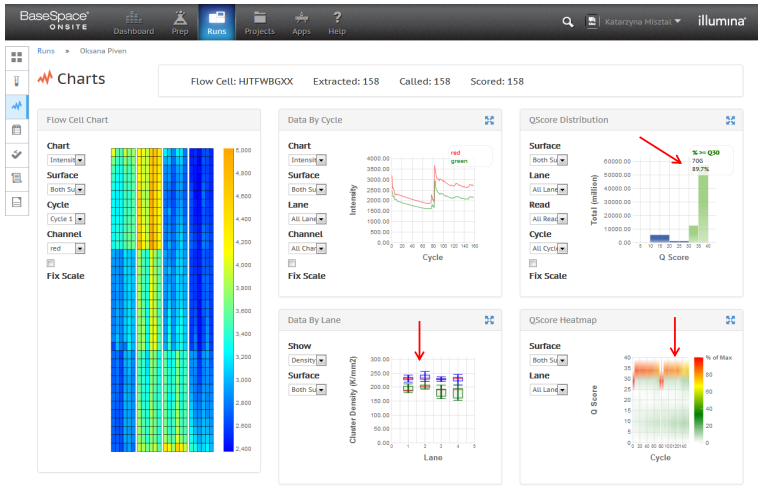
IIMCB

Optimal cluster density enables efficient and accurate quantitation

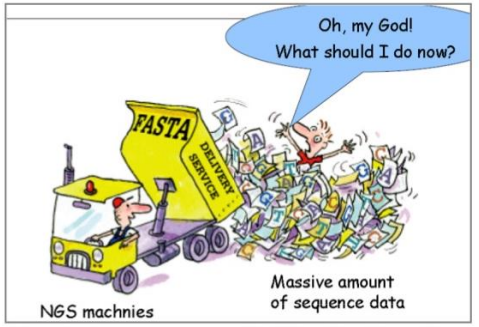


https://www.neb.com/tools-and-resources/feature-articles/the-quantitation-question-how-does-accurate-library-quantitation-influence-sequencing

Evaluation of clustering of your libraries can be done after the sequencing run



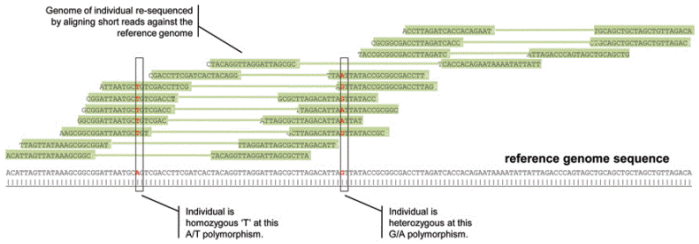
Computer analysis



http://ueb.ir.vhebron.net/NGS

Computer analysis

aligment and data analysis

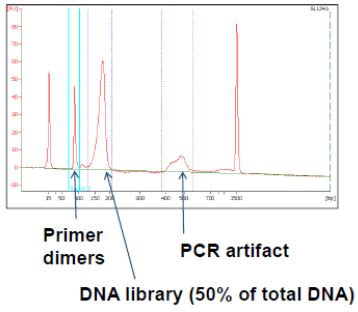


http://bio.cwchen.tw/ngs/2015/03/29/next-generation-sequencing/

Troubleshooting in library construction

Primer dimers;
error or omission of DNA gel cut (proper cut should remove <150 bp
Improper AMPure beads ratio (proper ration should remove <150 bp
Inefficient ligation
Too little input DNA
Primers <100bp cannot cluster or be sequenced, but can bind to flowcell

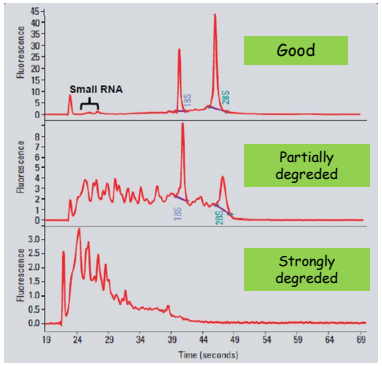
Library of unexpected sizes;
Recommended gel conditions not used
Other types of gel and staining methods can shift size ranges leading to incorrect band excision



Agilent's Bioanalyzer Applications for Next generation Sequencing: Updates and Tips March1, 2011

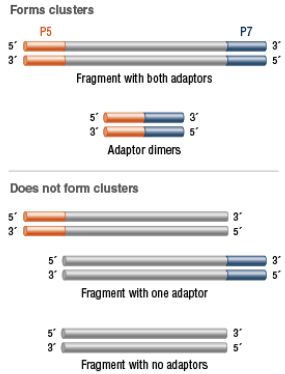
Troubleshooting in library construction

Degradation of the material



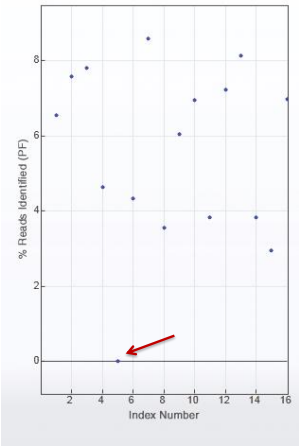
Troubleshooting in library construction

properly attached addapters
is the way for succesful sequencing



https://www.neb.com/tools-and-resources/feature-articles/the-quantitation-question-how-does-accurate-library-quantitation-influence-sequencing

Troubleshooting in library construction



uneven pooling of libraries yields uneven sequence coverage

Inadequate or uneven pooling of libraries can result in suboptimal data, and even lead to the need for library resequencing, as seen with library #5

https://www.neb.com/tools-and-resources/feature-articles/the-quantitation-question-how-does-accurate-library-quantitation-influence-sequencing

From Chip-Seq to Hi-C: looking at higher order structure in chromatin (Lecture)

Bartek Wilczyński
Faculty of Mathematics, Informatics and Mechanics, University of Warsaw

Thursday, 12:00

Next generation sequencing is frequently used to interrogate the state of chromatin through different biochemical protocols followed by sequencing. These new methods give us an unprecedented insight into the workings of eukaryotic nucleus. Protocols such as Chip-Seq, DNAseI-Seq, Mnase-Seq, ATAC-Seq, ChIP-Exo-Seq, GRO-Seq and Hi-C are showing us different facets of chromatin biology, frequently inconsistent with the textbook model of gene regulation. I will discuss some of these techniques and how they can be utilized together to give us better understanding of the processes taking place in the nucleus mostly focusing on those related with gene regulation. I will discuss some of the more interesting published results as well as share some yet unpublished data. I will include some of my own experiences with difficulties of integration of experimental results of different type with the current state of the art.