National Bioinformatics Infrastructure Sweden (NBIS) and Introduction to NGS data analysis

Jeanette Tångrot
CLiC – Computational Life Science Cluster
NBIS – National Bioinformatics Infrastructure Sweden
jeanette.tangrot@umu.se / jeanette.tangrot@nbis.se

NBIS - National Bioinformatics Infrastructure Sweden - will from 2016 be the continuation of:

Bioinformatics Infrastructure for Life Sciences

Wallenberg Advanced Bioinformatics Infrastructure

Systems Biology Infrastructure for the Life Sciences

SciLifeLab Bioinformatics Platform

SciLifeLab Platforms and facilities

National facilities

Affinity Proteomics
Biobank Profiling
Cell Profiling
Fluorescence Tissue Profiling
Mass Cytometry
PLA Proteomics
Protein and Peptide Arrays
Tissue Profiling

Bioimaging
Advanced Light Microscopy
Fluorescence Correlation Spectroscopy

Chemical Biology Consortium Sweden
Laboratories for Chemical Biology Umeå (LCBU)
The Laboratories for Chemical Biology at Karolinska Institutet (LCBKI)
Uppsala Drug Optimization and Pharmaceutical Profiling (UDOPP)

Drug Discovery and Development
ADME (Absorption Distribution, Metabolism Excretion) of Therapeutics (UDOPP)
Biochemical and Cellular Screening
Biophysical Screening and Characterization
Human Antibody Therapeutics
In Vitro and Systems Pharmacology
Medicinal Chemistry – Hit2Lead
Medicinal Chemistry – Lead Identification
Protein Expression and Characterization

Functional Genomics
Eukaryotic Single Cell Genomics
Karolinska High Throughput Center (KHTC)
Microbial Single Cell Genomics
Single Cell Proteomics

Metabolomics
Swedish Metabolomics Centre (SMC)

National Bioinformatics Infrastructure Sweden (NBIS)
Bioinformatics Compute and Storage (UPPNEX)
Bioinformatics Long-term Support (WABI)
Bioinformatics Short-term Support and Infrastructure (BILS)
Systems Biology

National Genomics Infrastructure
NGI Stockholm (Genomics Applications)
NGI Stockholm (Genomics Production)
NGI Uppsala (SNP&SEQ Technology Platform)
NGI Uppsala (Uppsala Genome Center)

Next-Generation Diagnostics (NGD)
Clinical Biomarkers
Clinical Genomics
Clinical Sequencing
Integrative Clinical Genomics
Translational and Clinical Genomics

Structural Biology
Cryo-EM
Protein Science Facility
Swedish NMR Centre (SNC)

www.nbis.se


Support
Infrastructure
Training
NBIS
NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN
NBIS is a distributed national bioinformatics infrastructure, supporting life sciences in Sweden

Why bioinformatics infrastructure?

A continuous technical scale-up will provide an unprecedented amount of heterogeneous omics data

- Support, Tools, Training

System-level analyses in biomedical research will transform life science

- Strategic positioning in systems biology

Large-scale omics is will make a major leap into translational research and diagnostics

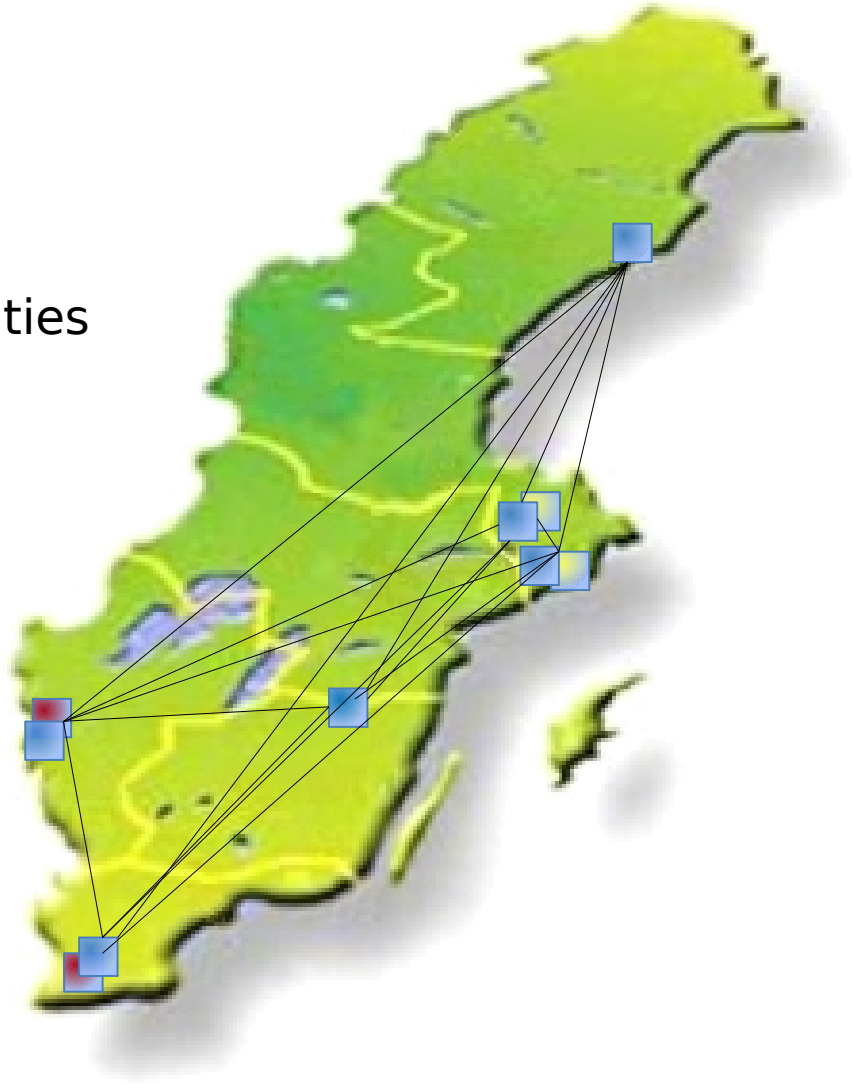
- Method adaptation and expert advice

NBIS

UiO

NBIS - National Bioinformatics Infrastructure Sweden

- NBIS nodes
- NGI
- Other sequencing facilities



NBIS - National Bioinformatics Infrastructure Sweden

SUPPORT: Distributed national infrastructure providing bioinformatics support to life science researchers in Sweden

TRAINING: Educate users, mainly PhD students and post-docs

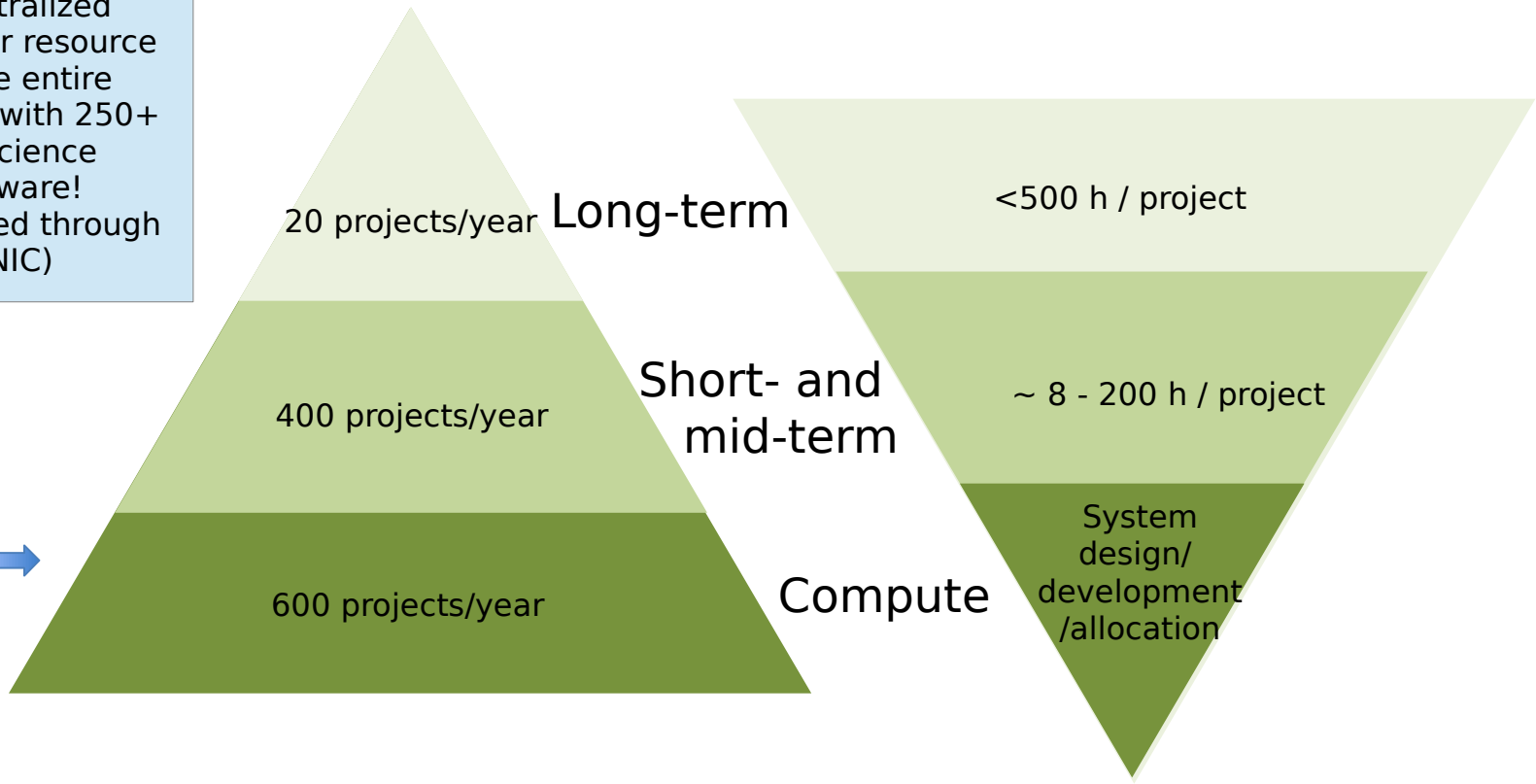
COMPUTE AND STORAGE: Develop systems and strategies for long-term large-scale storage of bioinformatics data (MS proteomics data, NGS sequence data, metabolomics). Provide high-performance computing (SNIC-UPPMAX) and a secure computing environment (MOSLER)

BIOINFORMATICS TOOLS: Provide more user friendly infrastructure (tools and databases) enabling researchers to perform more bioinformatics analyses on their own

“ELIXIR” NODE: Swedish contact point to the European infrastructure for biological information - ELIXIR

NBIS

A centralized computer resource for the entire country, with 250+ life science software! (Organized through SNIC)



Compute and Storage

UPPNEX

- free
- majority of hardware and system administration belongs to SNIC
- Apply: https://supr.snic.se
- Read more: http://www.uppmax.uu.se

Director

Hans Karlsson

Manager

Ola Spjuth

Short-term Support (Formerly known as BILS)

- When you have your data
- First come first serve
- ≤8h/PI/year for free
- >8h user fee, 800 SEK/hour
- Requests are reviewed every second week
  - Which scientific question do you want to answer?
  - What kind of data do you have?
  - What kind of help do you need?

Director

Bengt Persson

Technical coordinator

Mikael Borg

Proteomics coordinator

Fredrik Levander

Syst. dev. coodinator

Jonas Hagberg

Genomics coordinators

Magnus Alm-Rosenblad

Henrik Lantz

Dag Ahrén

Training coordinators

Sara Light

Jessica Lindvall

NBIS

Support request forms at nbis.se/support

Long-term Support

Wallenberg Advanced Bioinformatics Infrastructure
www.scilifelab.se/facilities/wabi/

Tailored solutions – high impact

- Scientific evaluation
- ≤500h, currently free
- Someone in the group must be assigned to work on the data
- Next deadline January 27th, 2017

Directors

Siv Andersson

Gunnar von Heijne

Managers

Björn Nystedt

Pär Engström

Swedens strongest unit for analyses of large-scale genomic data (~20 FTE)

NBIS

Support request forms at nbis.se/support

Criteria for accepted projects

Scientific level

A proposals evaluation committee with ***national delegates*** will score the scientfic level of the project.

Feasibility

The bioinformatics management will evaluate if the support team has the technical expertise needed for the project.

Involvement

The applying party must assign at least one scientist from their group to take part in the bioinformatics work to ensure efficient knowledge transfer and longevity of the project beyond the time of the granted support

Consultation

Consultation meetings (<3h, free)

- When you are in the planning stage

Drop-in sessions

biosupport.se

Support request forms at nbis.se/support

Expert teams

Assembly/annotation service

- part of Short-term Support
- (2 + 2 people, running)

Human WGS ToolBox

- Method implementation, community building
- https://wabi-wiki.scilifelab.se/display/SHGATG/
- (2+ people, running)

BigData/Integrative bioinformatics

- Method development, project support
- (4 people, hiring now, part of Long-term Support)

The Swedish Bioinformatics Advisory Program

A new teaching model, where PhD students get a senior bioinformatician as a personal advisor during 2 years of their PhD.

Overall aim: Great research in Sweden!

How?

- Strategic investment in PhD education
- Complementing PhD supervisors with technical expertise
- Catalyze transition to large-scale data analyses

Monthly project meetings + two grand meetings per year to aid networking and knowledge transfer. The PhD student is responsible to prepare and drive the monthly meetings

Last call, Nov 2016: 111 applicants for 15 places

www.scilifelab.se/education/mentorship/the-swedish-bioinformatics-advisory-program/

Next call Nov-December 2017

Bioinformatics Drop-In

Are you planning a project and need someone to discuss the bioinformatics analysis with?

Do you need bioinformatics support, but do not know who to turn to?

Are you stuck in your own bioinformatics project and need help?

Meet the NBIS staff at bioinformatics drop-in!

- Umeå:
  - Weekly on Tuesdays at 10 am
  - KBC cafeteria (uneven weeks) / Department of Molecular Biology lunchroom (even weeks)
- Similar activities in the other NBIS nodes/cities, e.g.:
  - Lund: Wednesdays at 10 AM, alternating Café Inspira / Café Marina
  - Stockholm: Tuesdays at 10.30 AM, SciLifeLab, gamma, level 6

NBIS representatives in Umeå

Short-term Support

Jeanette Tångrot
Genomics

Rui Pinto
Metabolomics and Chemometrics

Joakim Bygdell
mass spectrometry proteomics

Long-term Support

Allison Churcher
Genomics

NBIS Annual Symposium and User Meeting 2016

Meet with NBIS staff and listen to interesting bioinformatics presentations!

Date: 2016-12-15

Time: 10:00 to 15:00

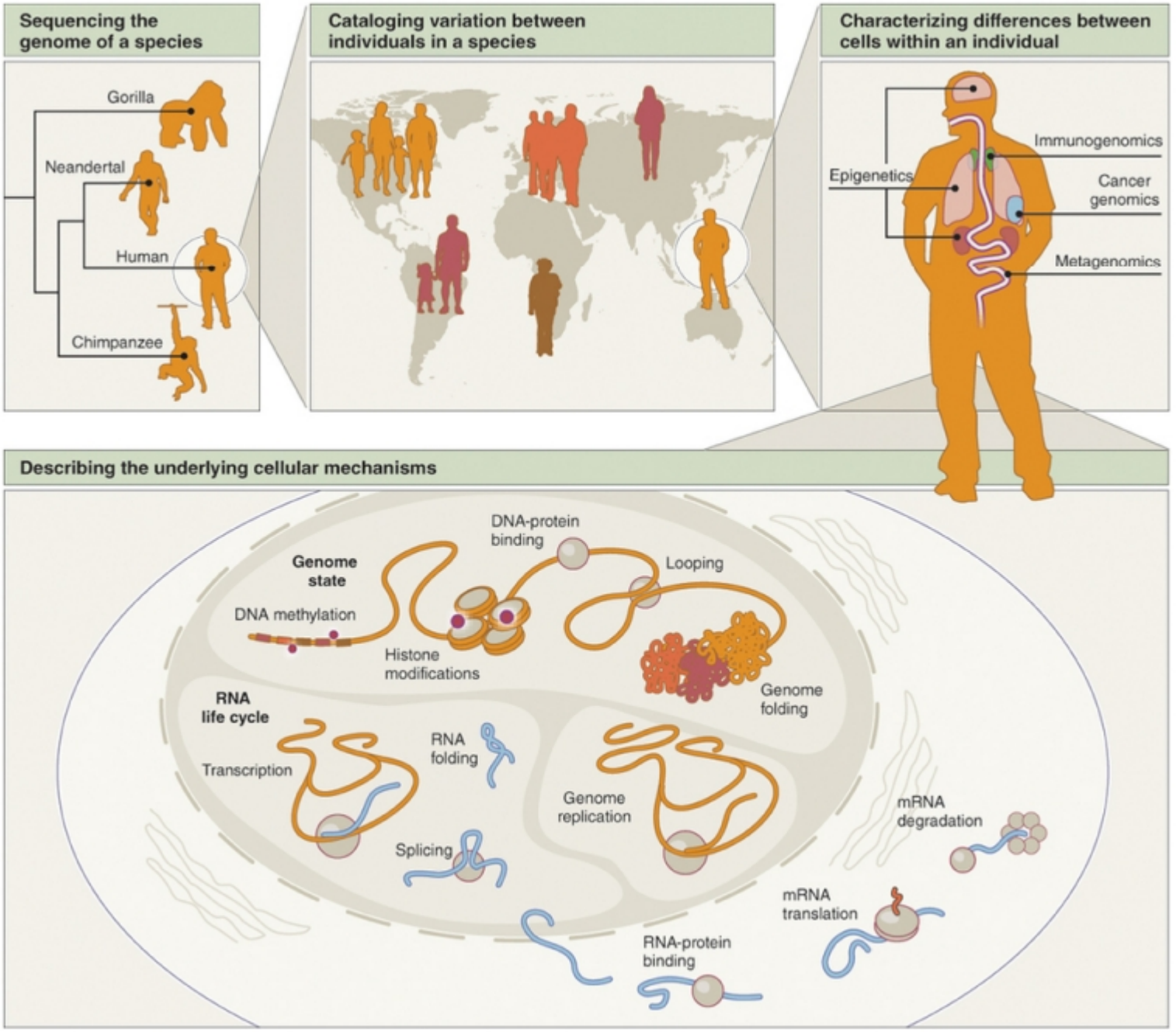
Location: KB.E3.03 (Stora Hörsalen), Umeå University

Register before Dec 9 at nbis.se

We're here for you!

Don’t be scared to contact us at any level

Just becuase you contacted us does not mean that you have to sign up for anything



Nature Biotechnology **30, 1084–1094 (2012)**

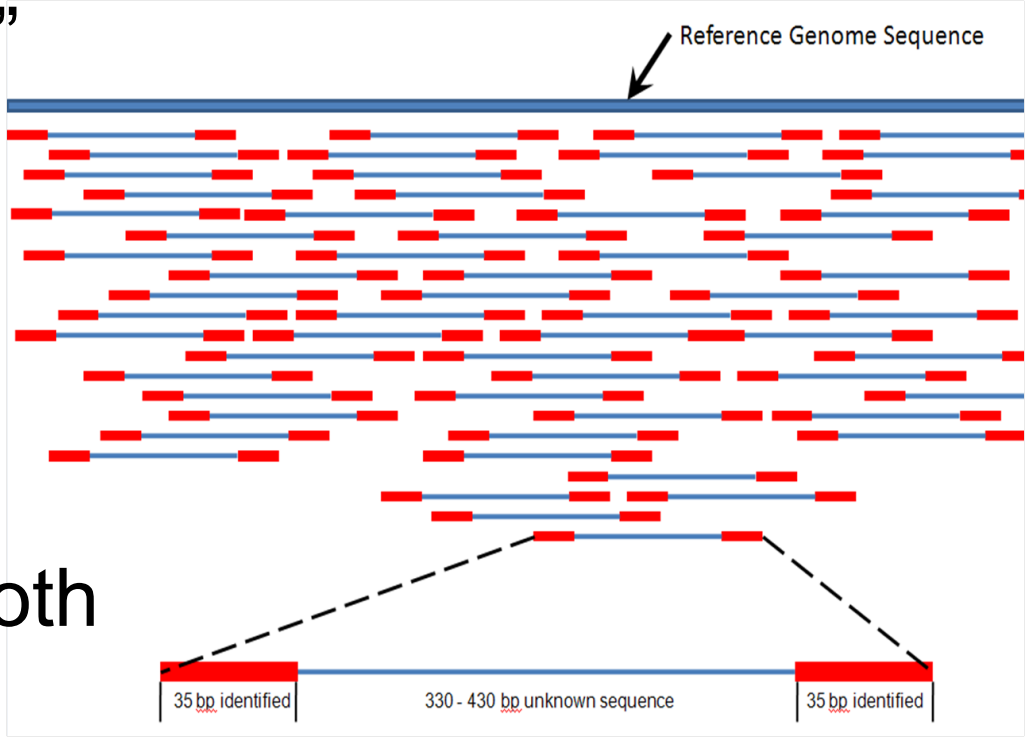
Bioinformatics of NGS data

NGS data analysis

- Obtain raw reads
  - basecalling, demultiplexing
  - quality control, read trimming
- Data processing
  - mapping/alignment
  - assembly
  - variant calling / expression values
- Data analysis
  - annotation
  - comparative genomics
  - variant filtering and variant annotation
  - multisample comparison – disease models
  - diagnosis suggestion / disease variant candidates
  - ...

Raw data

- Raw data = “reads”
- Up to 6 billion reads/run
- 100 -300 bp read length (Illumina)
- Sequences from both ends of fragment



http://www.tutorgigpedia.com/ed/Next-generation_sequencing

Fastq-files

Fastq format:

```
@ILLUMINA-5C547F_0001:4:1:1043:19101#GATCAG/1
TTATTTATGCACTCCAAAAACAAACTTCTATTATAGATTTACCTGTATATTCATTTATAGATGCCTTTGTTACCGCAATATCTT
+
bbbbbbbbbbbbbbbbbbbb^]___bbbbbbbbbbbbbbbbbbbabbbbbbbabab_babb^bb_^bbbbbbbbbbbbZbbbbbb
@ILLUMINA-5C547F_0001:4:1:1043:13674#GATCAG/1
AATATGGTTCTCAAATAAGAGCACTTAAGCAAGGTGTAAAAGTTGTAGTTGGTACAACTGGTCGAGTAATGGATCATATTGAGA
+
b!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>CCCCCC65babC`babab_`bb_]b_b__b^[`Z
```

Quality control

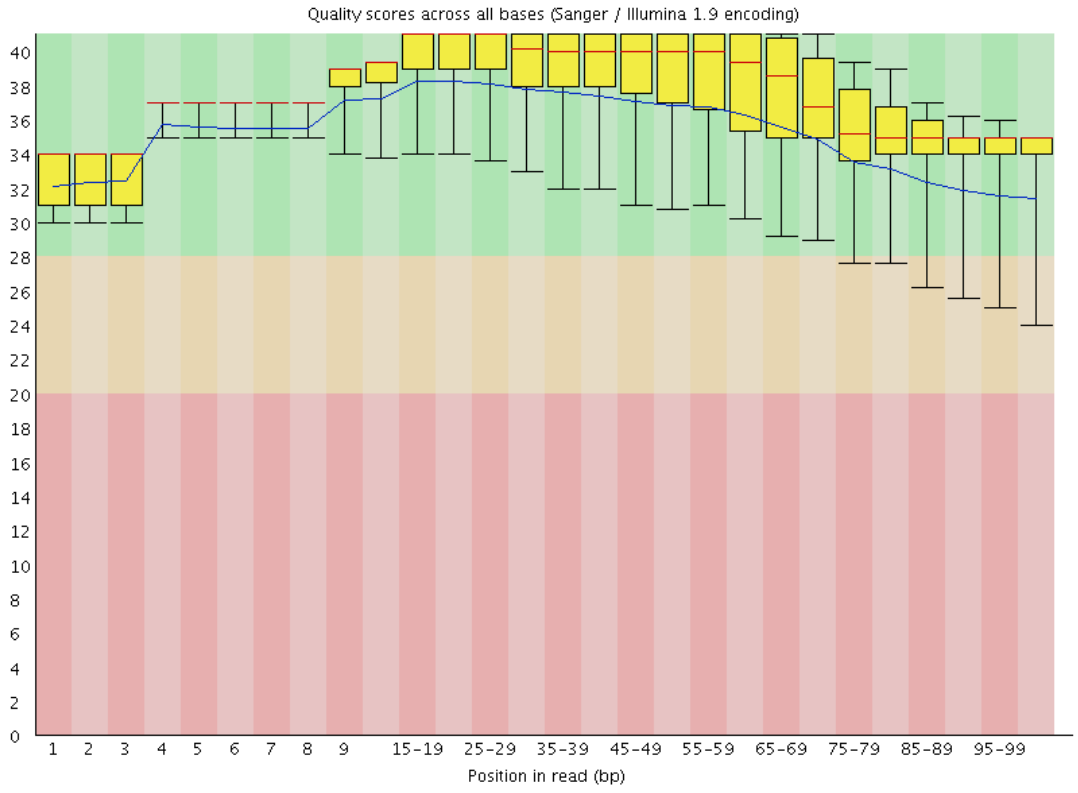
FastQC Report

Thu 29 Sep 2011
Sample_1_9_1.fastq.gz

Summary

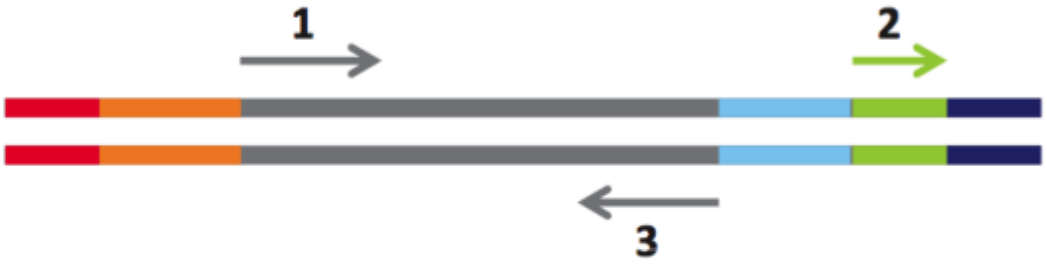
- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Per base sequence quality

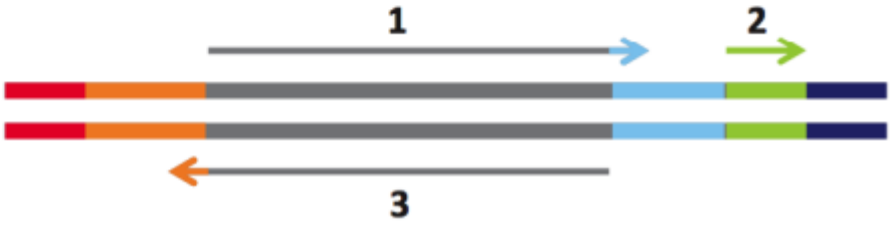


Adapter contamination

Only the gray part is your DNA-of-interest. So, ideally …



… single (1) or paired (1 & 3) reads are shorter than the fragment, and a separate *barcode* read (2) identifies it as belonging to a particular sample.



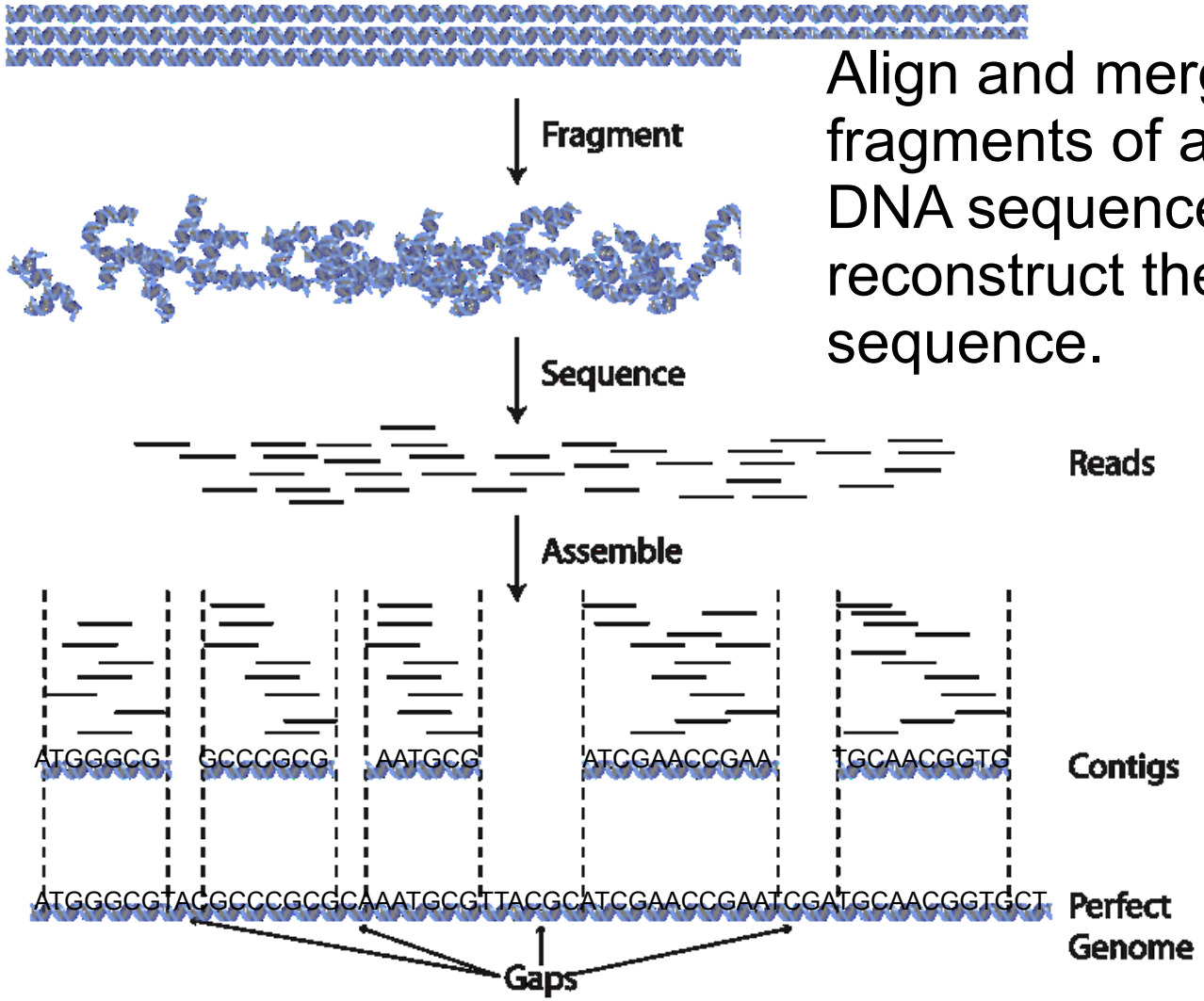
Trimming

- Trimming of data
  - Contamination removal
  - Adaptor cleaning
  - Quality trimming
- Can often be left to the alignment software deal with
- Trimming can rescue coverage and reduce noise
  - E.g. RNAseq, variant calling
- Trimming can also make the amount of data more manageable

NGS data analysis

- Obtain raw reads
  - basecalling, demultiplexing
  - quality control, read trimming
- Data processing
  - mapping/alignment
  - assembly
  - variant calling / expression values

De novo assembly



Align and merge short fragments of a much longer DNA sequence, in order to reconstruct the original sequence.

De novo assembly

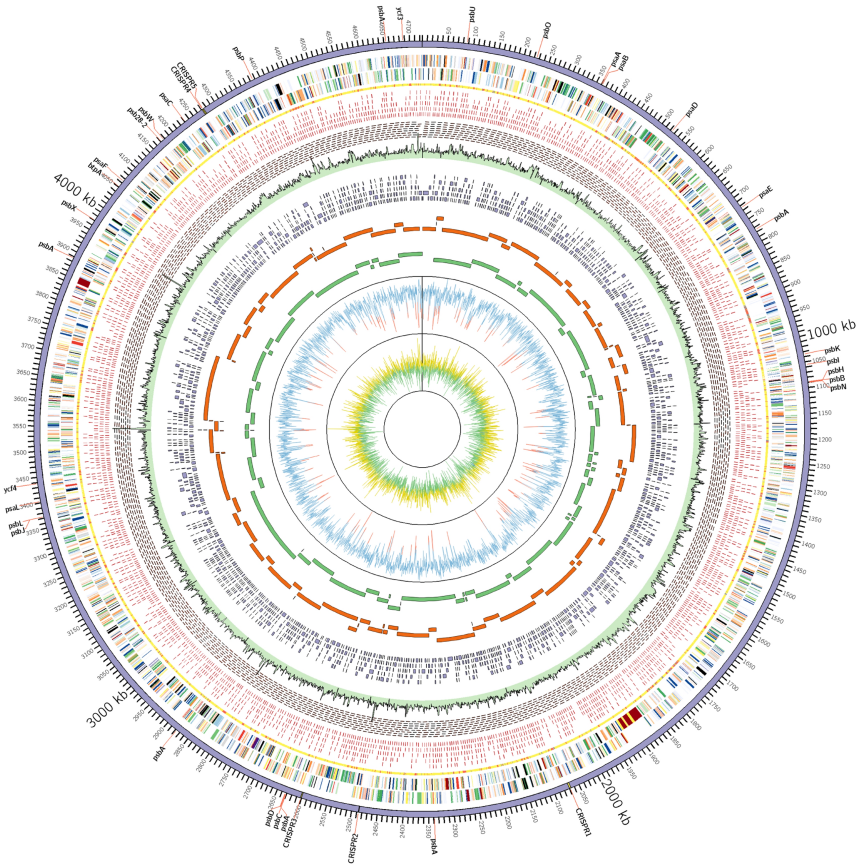
- Jigsaw puzzle from a pile of reads
- Find matches to other reads
- Challenges:
  - Sequence errors
  - Repeats
  - Polyploidy
  - GC content/complexity
  - A large amount of data
  - Contamination sequences

Novel genome analysis

- Genome assembly

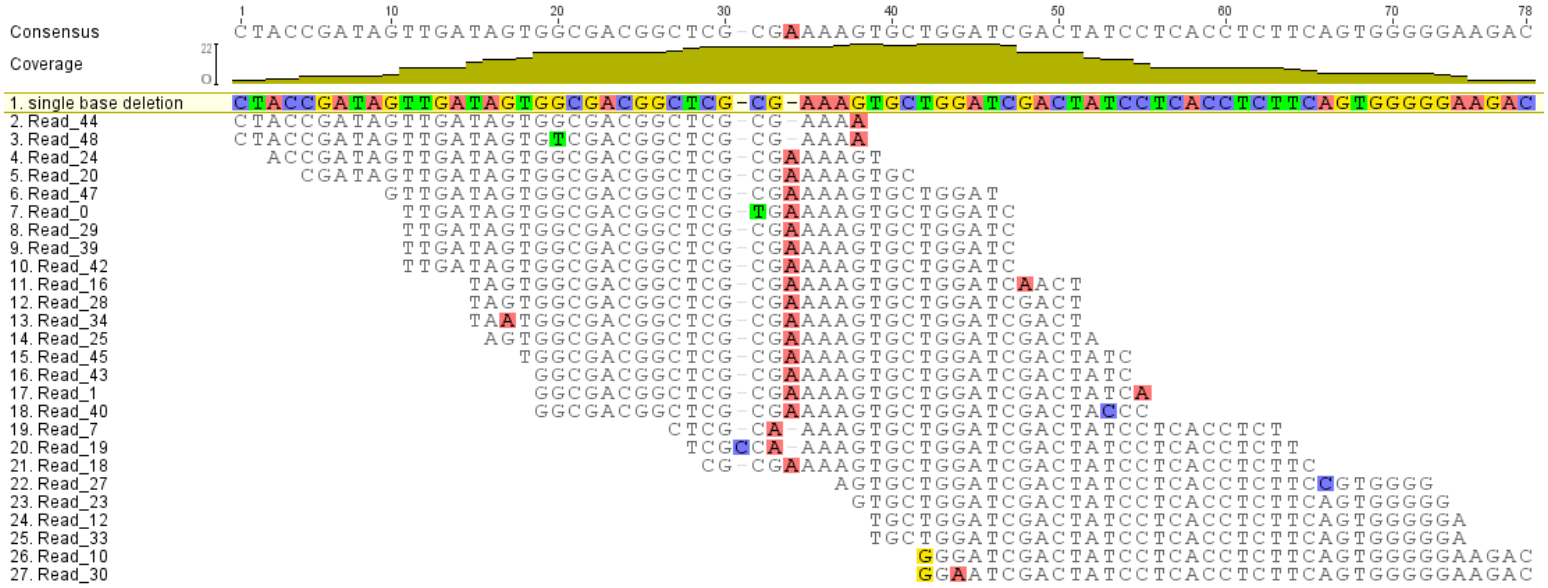
  (and finishing)
- Genome annotation
  - Find all functional elements

    (genes, ncRNA, ...)
- Comparative genomics
  - Copy Number Variants (CNVs)
  - Single Nucleotide Polymorphisms (SNPs)
  - structural rearrangements
  - large INDELs



Picture from Saw JHW et al. (2013) PLoS ONE 8(10): e76376.

Aligning reads to a reference genome / Mapping



- Mapping this large volume of short reads to a genome as large as human poses a great challenge!
- This is the first step in the data analysis of many NGS applications

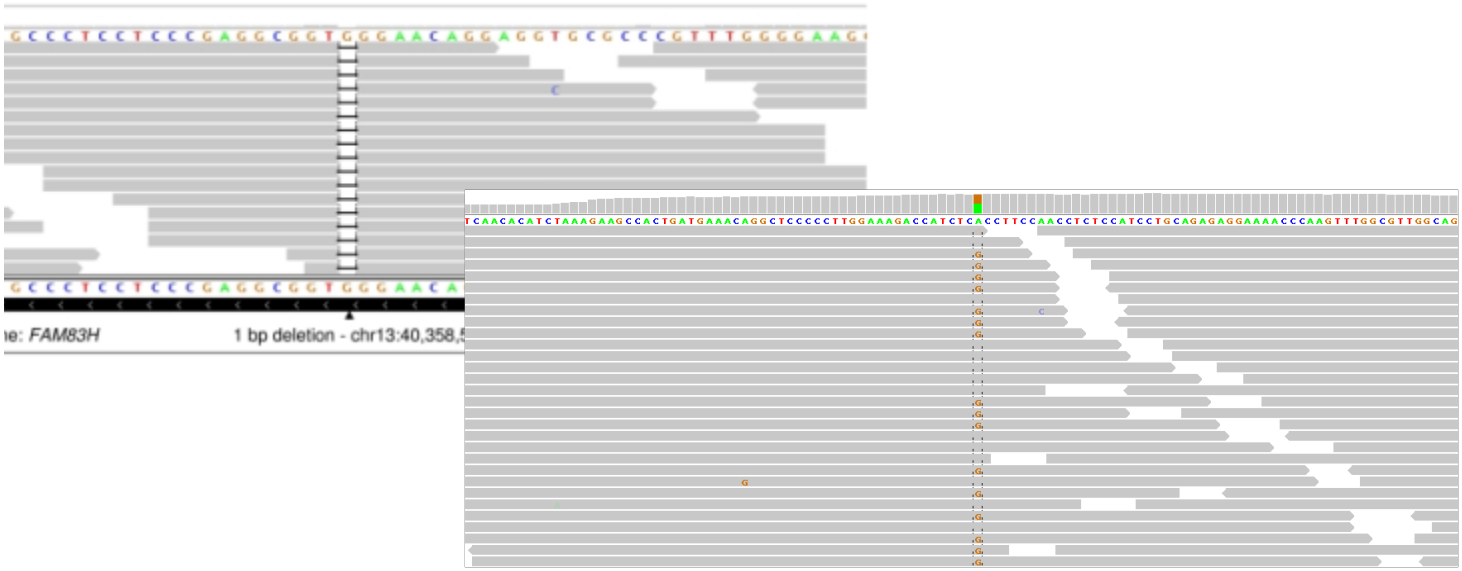
Variant detection

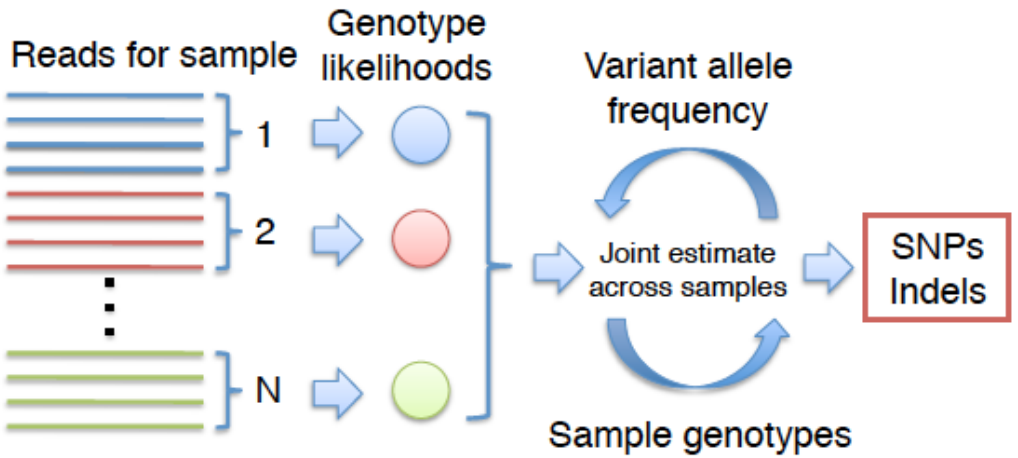
* Align reads to reference genome (BWA, Bowtie etc)

* Mark duplicates

    * Identify variations (e.g. GATK by the Broad institute)

    * Filter results





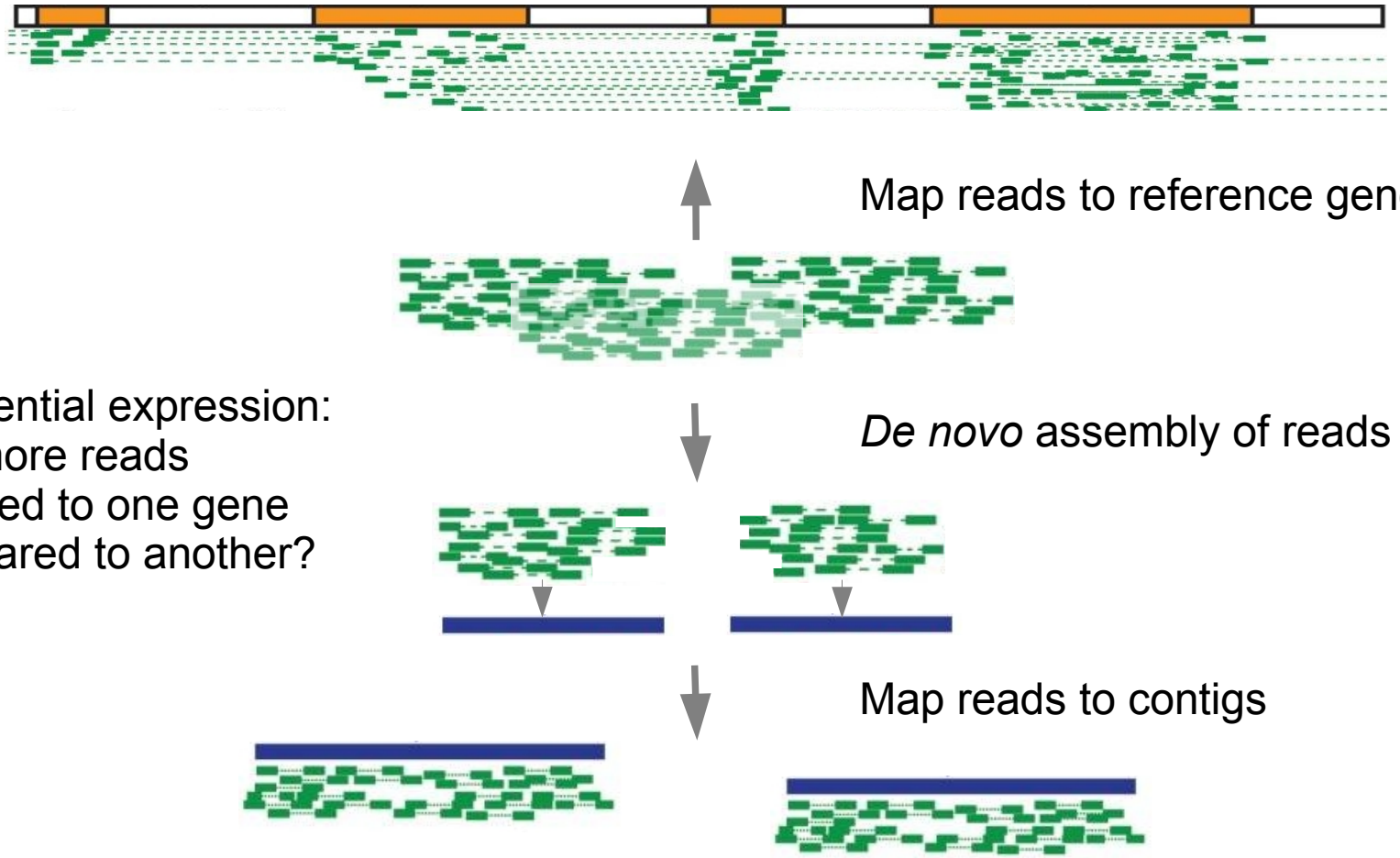
Re-sequencing

- Single Nucleotide Polymorphisms (SNPs)
- Small INDELs
- Structural variation
  - Copy Number Variants (CNVs)
  - Structural rearrangements
  - Large INDELs
- Tumour mutations

RNA-seq

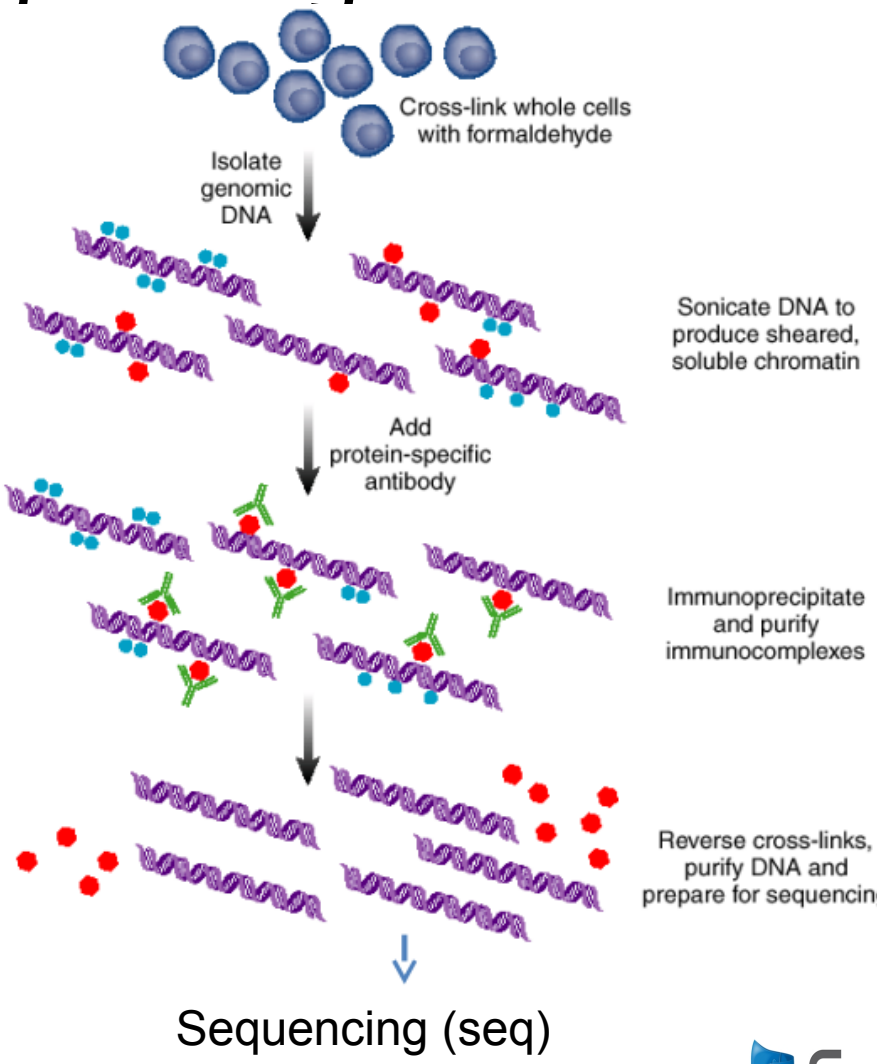
- Differential gene expression analysis
  - Healthy vs. diseased
  - Time course experiments
  - Different genotypes
- Transcriptional profiling
  - Tissue-specific expression
- Novel gene identification/transcriptome assembly
- Identification of splice variants
- SNP finding
- RNA editing

RNA-seq

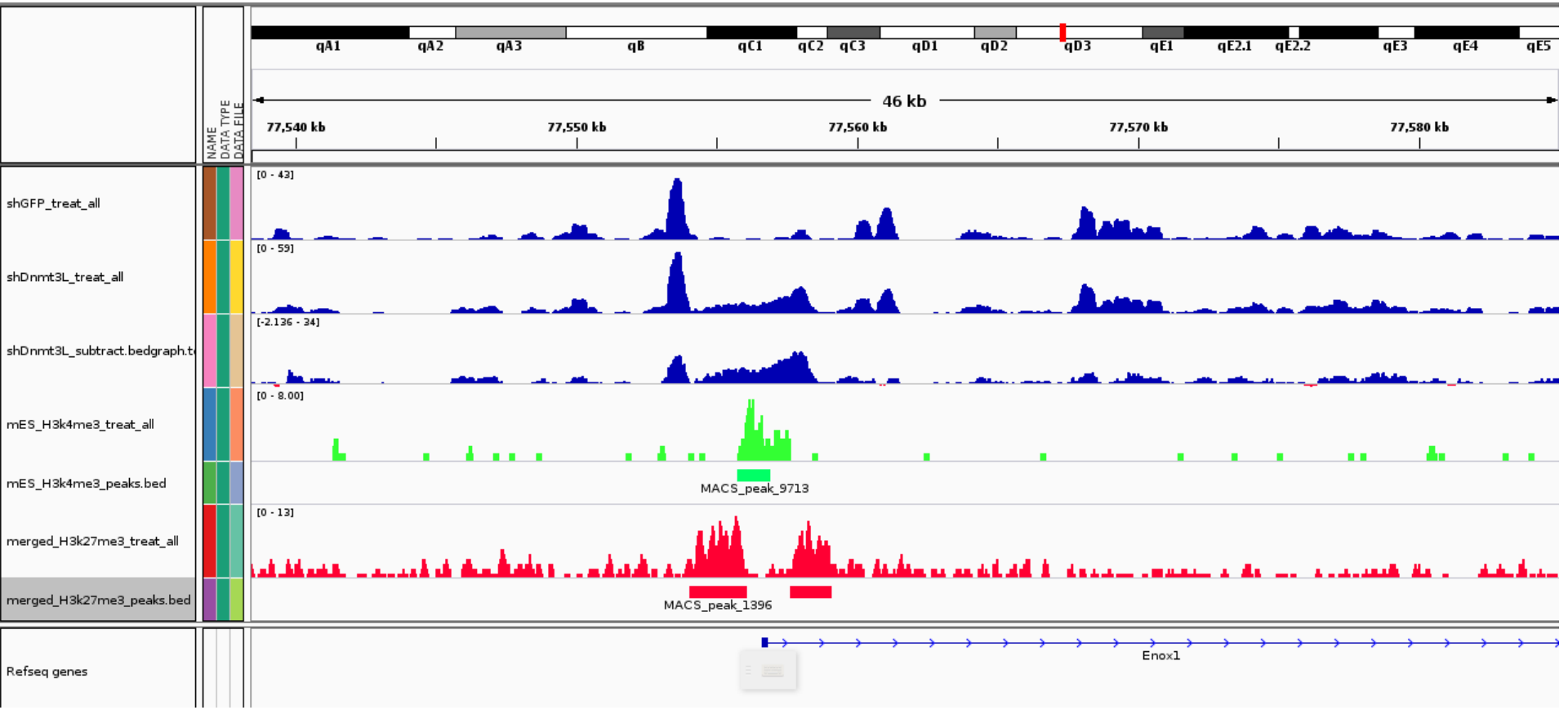

Map reads to reference genome
Differential expression:
Are more reads
mapped to one gene
compared to another?
De novo assembly of reads
Map reads to contigs

Sequencing to study gene regulation

- **ChIP-seq**: combines chromatin immunoprecipitation with sequencing to identify the binding sites of DNA-associated proteins
- **MeDIP-seq**: combines methylated DNA immunoprecipitation with sequencing.

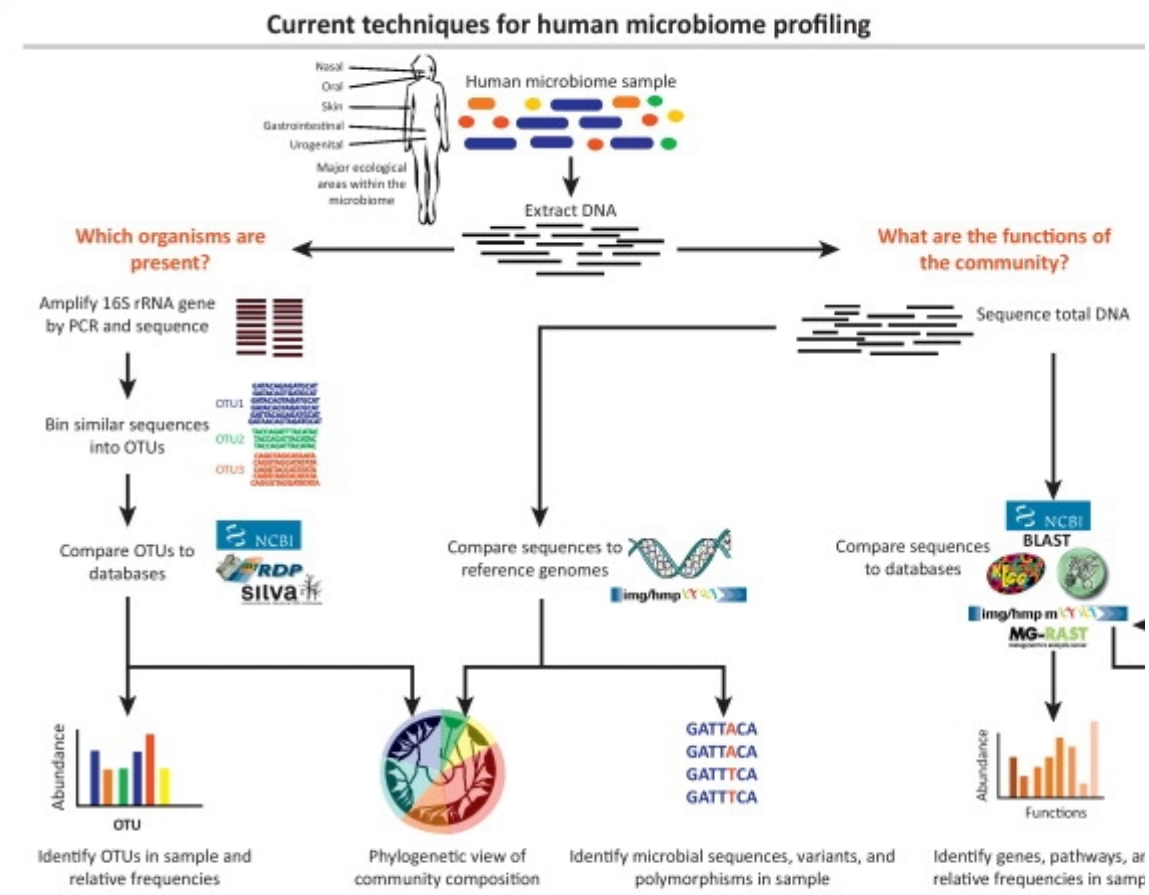


Mapping to reference and finding peaks



Picture from http://crazyhottommy.blogspot.se/2014/01/medip-seq-and-histone-modification-chip.html

Metagenomics

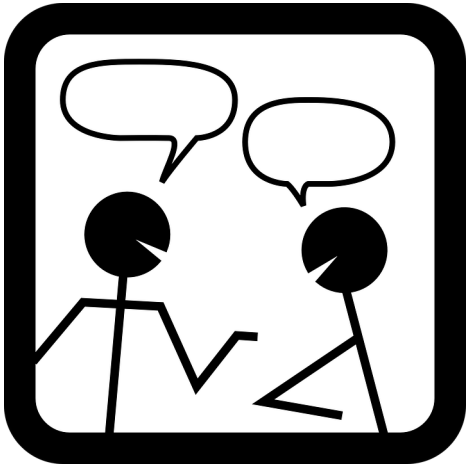


Morgan, Xochitl C. et al. Trends in Genetics 29:1, 51-58

NBIS

UiO

Bioinformatics Drop-in Support



Every Tuesday at 10:00!

The bioinformatics experts in CLiC/NBIS are available to discuss your bioinformatics needs in the **Department of Molecular Biology lunchroom** or the **KBC cafeteria** on alternating Tuesdays.