Genomics 2012: Repetitorium

Lectures 1 and 8–15

Peter N Robinson

February 7, 2013

Repetitorium

Genomics 2012: Repetitorium

Peter N Robinson

VL1: Next-Generation Sequencing

VL8–9: Variant Calling

VL10–11: Structural Variants

VL12–13: RNA-seq

VL14: ChIP-seq

This is a review of the material from lectures 1 and 8–14. Note that the material from lecture 15 is not relevant for the final exam. Today we will go over the material of the lectures and point out some of the most important topics.

- We have covered a lot of material in this course, and could have gone into much more depth on many topics
- For the final exam, you are responsible for the material in the slides

Outline

VL1: Next-Generation Sequencing

VL8–9: Variant Calling

VL10–11: Structural Variants

VL12–13: RNA-seq

VL14: ChIP-seq

1. VL1: Next-Generation Sequencing
2. VL8–9: Variant Calling
3. VL10–11: Structural Variants
4. VL12–13: RNA-seq
5. VL14: ChIP-seq

VL1: Next-Generation Sequencing

Methods and Computational Analysis Important topics

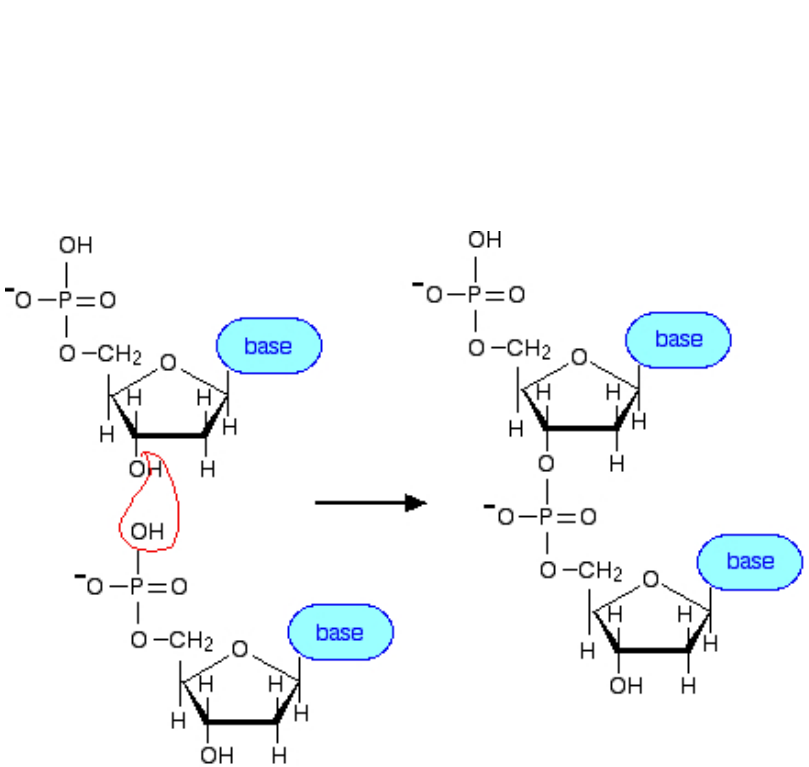
- Sanger sequencing (biochemistry at level shown in slides)
- Major steps of NGS procedure (important items shown in grey boxes)
- No questions about NGS technologies other than Illumina (don't worry about 454, Ion Torrent, Roche, nanopore)
- Formats: Know FASTQ, PHRED score
- Do not memorize SAM format, but understand basics of CIGAR

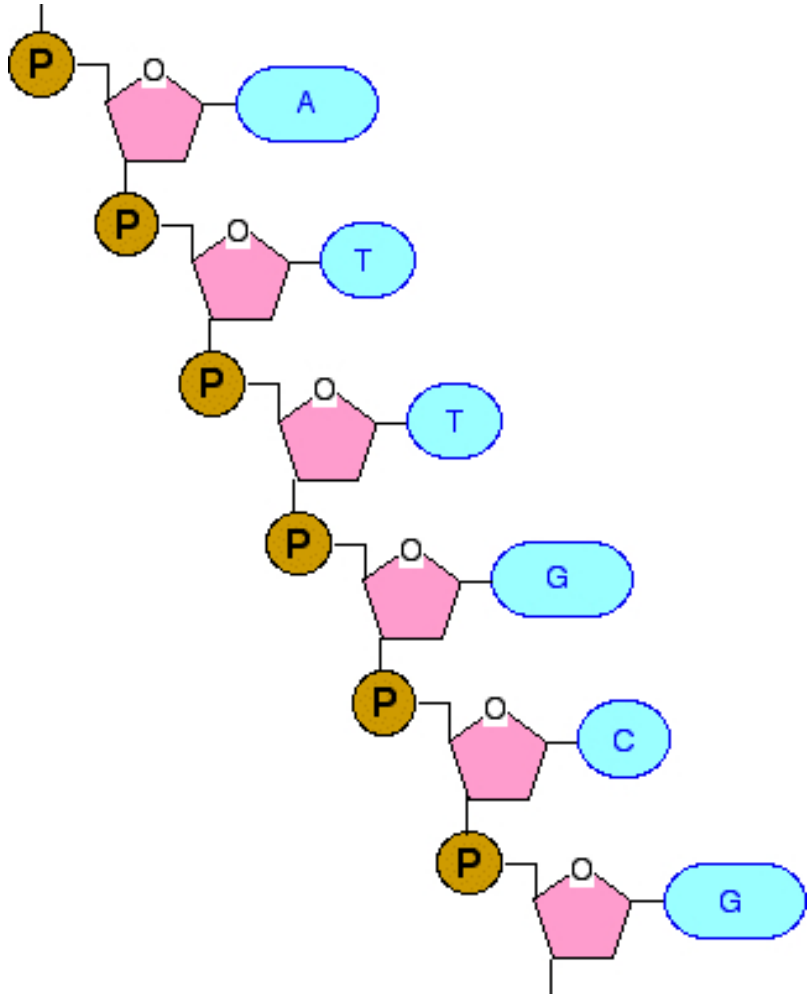
VL1:Q1

1) Explain the role of dideoxynucleotides in Sanger sequencing

(3 minutes)

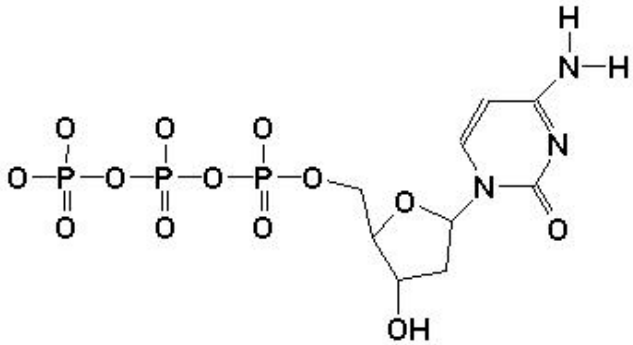
VL1: A1

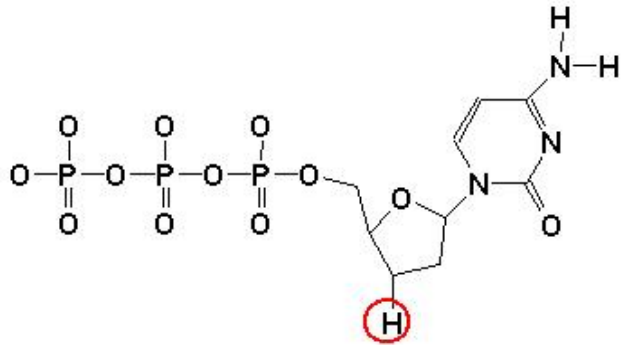




- Recall that DNA is extended from 5' to 3'

VL1: A1





Deoxycytosine (dCTP)

- A good answer would be

Chain-terminating precursors of DNA synthesis that block further polymerization when added to the end of the DNA strand by DNA polymerase. These nucleotides lack a 3'-OH (hydroxyl) group necessary for continued 5'-to-3' DNA synthesis. The length of a fragment whose synthesis has been terminated at a given nucleotide by a ddNTP incorporation has a defined length that can be resolved by electrophoresis.

VL1:Q2

2) Explain the major differences if Sequencing by synthesis (NGS sequencing with the Illumina protocol as explained in the slides) as compared to Sanger sequencing

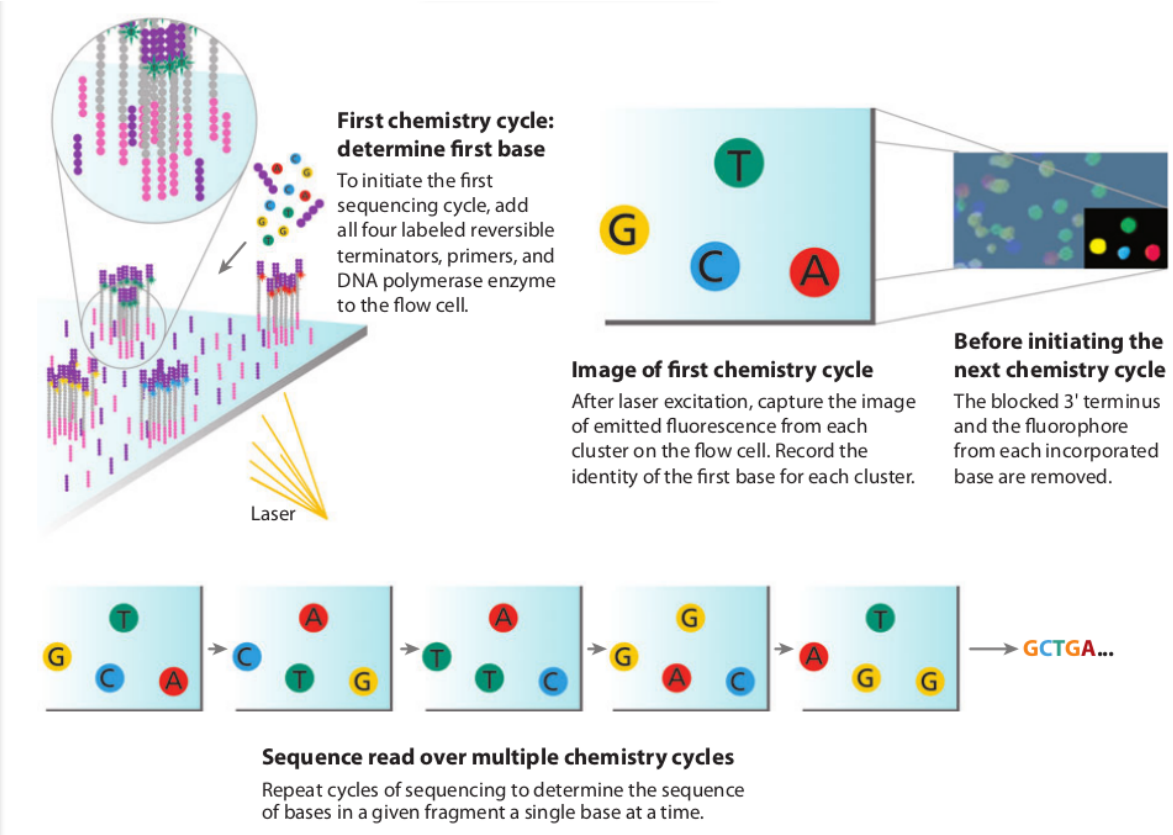
(3 minutes)

Sequencing by synthesis

- Sequencing by synthesis (SBS) is relatively simple to understand after all of the prep steps
- Basic idea:
  1. incorporate one base at a time using four differently marked, ddNTPs
  2. Difference to Sanger sequence: the ddNTPs have *reversible* terminators
  3. identify which base was incorporated in each cluster by measuring wavelength of incorporated ddNTP
  4. unblock the ddNTP and repeat cycle.

Sequencing by synthesis

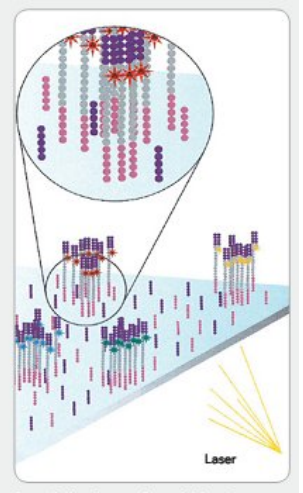
- Sequencing by synthesis:



Mardis E (2008) *Annu Rev Genomics Hum Genet* **9**:387402

Sequencing by synthesis

- Sequence "down" towards the flow cell
- All four ddNTPs added simultaneously
- Each ddNTP has reversible chemical block of 3' OH group
- Each ddNTP has unique fluorescent label



Outline

VL1: Next-Generation Sequencing

VL8–9: Variant Calling

VL10–11: Structural Variants

VL12–13: RNA-seq

VL14: ChIP-seq

1 VL1: Next-Generation Sequencing

2 VL8–9: Variant Calling

3 VL10–11: Structural Variants

4 VL12–13: RNA-seq

5 VL14: ChIP-seq

VL8–9: Variant Calling

Important topics

- Heterozygous and homozygous variants and the biological signals produced by them in NGS data
- Bayes theorem and the material on MAP
- EM: Do not memorize equations, but be familiar enough to understand them if you see them
- Beta distribution, hyperparameter (We will not include the Dirichlet distribution in the final)
- Maximum Likelihood (ML) Estimation vs. Maximum a posteriori (MAP) Estimation
- Annotation and variant nomenclature: Not on final exam.

Germline variants

From read mapping, we get a series of aligned columns of nucleotides and have information about

1. mapping quality for each read
2. base call quality for each position
3. A stack of nucleotides

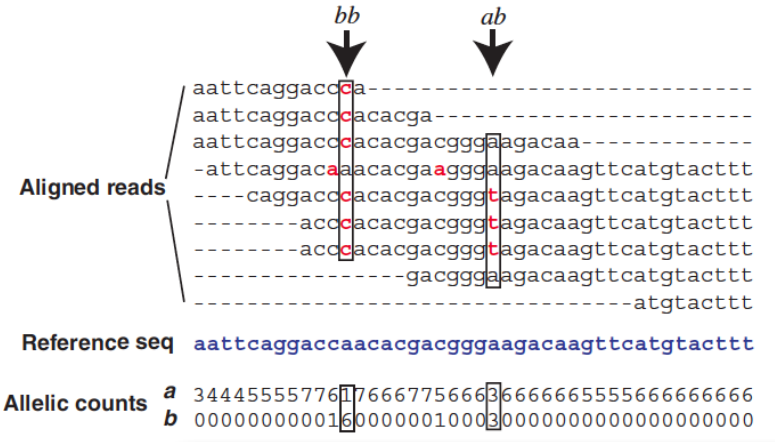
- Describe qualitatively how these attributes contribute to calling of single nucleotide variants in NGS

Germline variants

From read mapping, we get a series of aligned columns of nucleotides and have information about

1. mapping quality for each read
2. base call quality for each position
3. A stack of nucleotides

- k wildtype nucleotides a
- $n - k$ nucleotides b
- $a, b \in \{a, c, g, t\}$ and a $\neq$ b



MAQ: Mapping quality

We now calculate the posterior probability of the mapping at position u, $p_s(u|\mathcal{R}, z)$ using Bayes law

$$p_s(u|\mathcal{R}, z) = \frac{p(z|\mathcal{R}, u)p(u|\mathcal{R})}{\sum_v p(z|\mathcal{R}, v)p(v|\mathcal{R})}$$

Identify the posterior, likelihood, prior and normalization constant in this formula (1 minute)

Bayes Theorem

Bayes' theorem follows from the definition of the conditional probability and relates the conditional probability $P(A|B)$ to $P(B|A)$ for two events A and B such that $P(B) \neq 0$:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

- posterior
- likelihood
- prior
- normalization constant

MAQ: Mapping quality

We now calculate the posterior probability of the mapping at position u, $p_s(u|\mathcal{R}, z)$ using Bayes law

$$p_s(u|\mathcal{R}, z) = \frac{p(z|\mathcal{R}, u)p(u|\mathcal{R})}{\sum_v p(z|\mathcal{R}, v)p(v|\mathcal{R})}$$

If we assume a uniform prior distribution $p(u|\mathcal{R})$, then the read is equally likely to begin at any position of the reference. The sum then goes over all positions from 1 to $L - |z| + 1$, where L is the length of $\mathcal{R}$ and $|z|$ is the length of a read.

Write an explicit formula for the normalization constant and simplify the equation where possible.

MAQ: Mapping quality

We now calculate the posterior probability of the mapping at position u, $p_s(u|\mathcal{R}, z)$ using Bayes law

$$p_s(u|\mathcal{R}, z) = \frac{p(z|\mathcal{R}, u)p(u|\mathcal{R})}{\sum_v p(z|\mathcal{R}, v)p(v|\mathcal{R})}$$

If we assume a uniform prior distribution $p(u|\mathcal{R})$, then the read is equally likely to begin at any position of the reference. The sum then goes over all positions from 1 to $L - |z| + 1$, where L is the length of $\mathcal{R}$ and $|z|$ is the length of a read.

$$p_s(u|\mathcal{R}, z) = \boxed{\frac{p(z|\mathcal{R}, u)}{\sum_{v=1}^{L-|z|+1} p(z|\mathcal{R}, v)}}$$

Outline

VL1: Next-Generation Sequencing

VL8–9: Variant Calling

VL10–11: Structural Variants

VL12–13: RNA-seq

VL14: ChIP-seq

1 VL1: Next-Generation Sequencing

2 VL8–9: Variant Calling

3 VL10–11: Structural Variants

4 VL12–13: RNA-seq

5 VL14: ChIP-seq

VL10–11: Structural Variants

Important topics

- The various kinds of signals used to detect structural variants (SVs) in NGS data
- The various kinds of SVs and what effects they have on NGS reads
- Basic steps in using read depth to identify copy number variations (CNVs)
- Poisson distribution and normal approximation
- GC bias
- empirical CDF

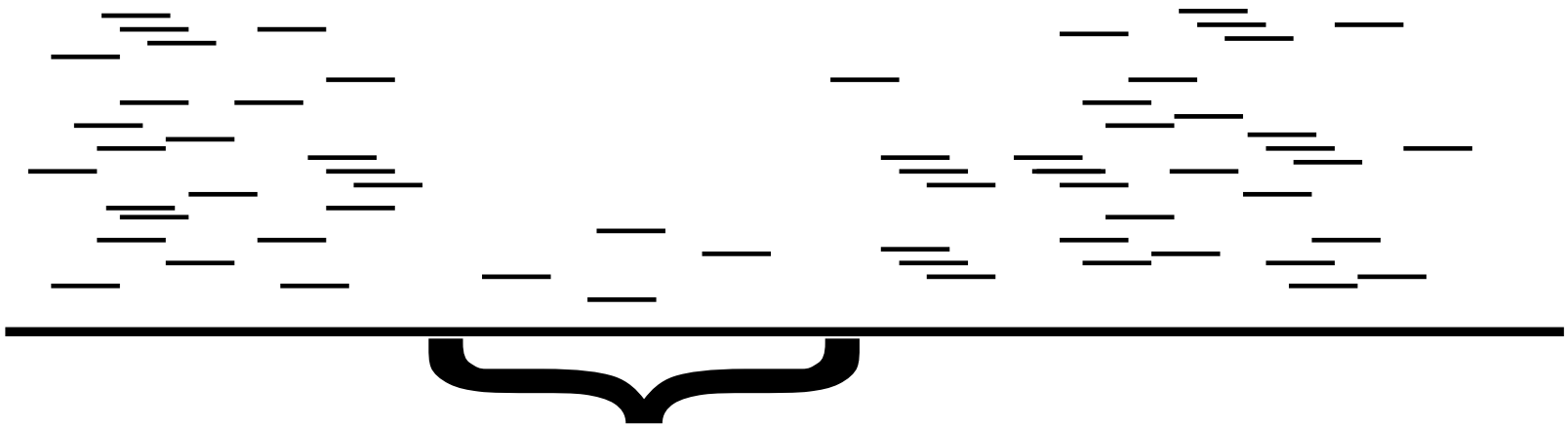
Read depth

Analysis of read depth is well suited to detect which class of structural variation and why? Name two potential sources of bias?

(3 minutes)

Read depth

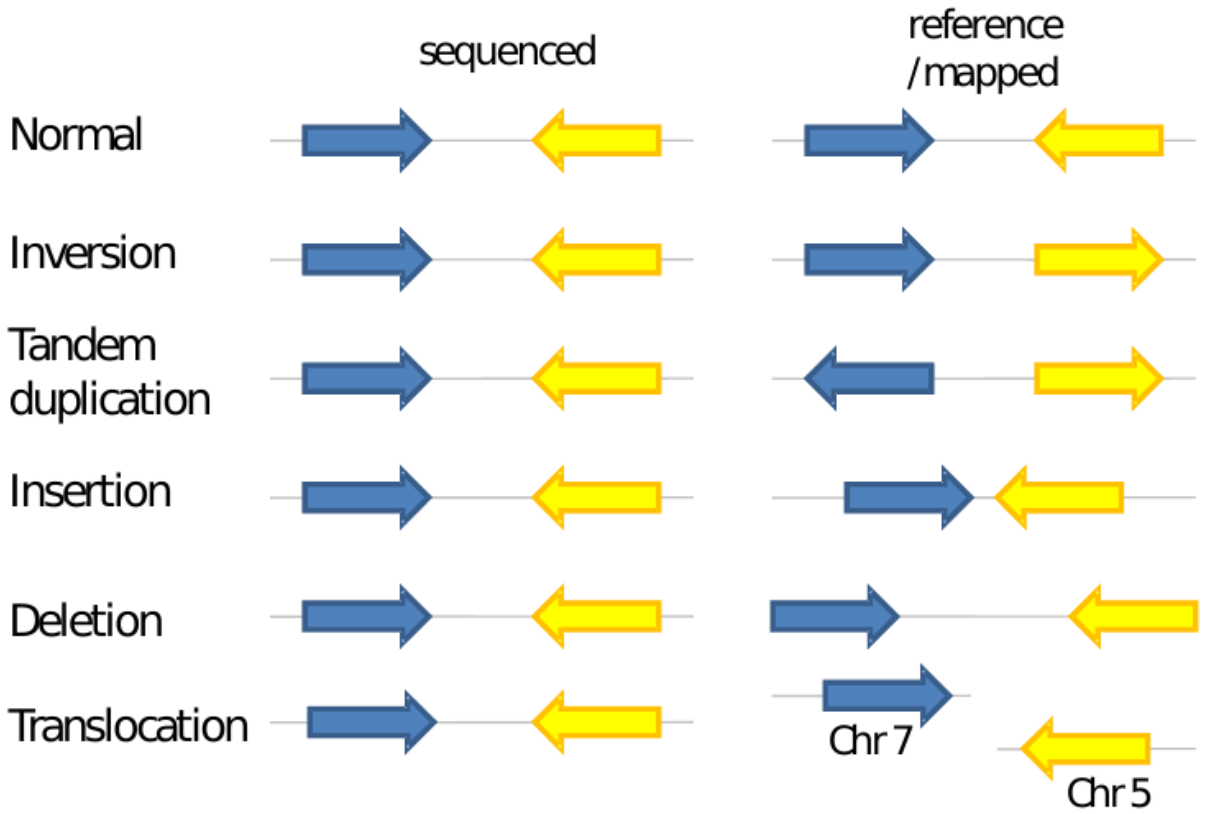
Analysis of read depth can identify deletion/duplications



Heterozygous Deletion?
Mappability Issue?
Poor "sequencability"?

Read depth

Characteristic signatures of paired-end sequences



- *Understand what is happening in this slide*

graphic credit: Victor Guryev

Outline

VL1: Next-Generation Sequencing

VL8–9: Variant Calling

VL10–11: Structural Variants

VL12–13: RNA-seq

VL14: ChIP-seq

1 VL1: Next-Generation Sequencing

2 VL8–9: Variant Calling

3 VL10–11: Structural Variants

4 VL12–13: RNA-seq

5 VL14: ChIP-seq

VL12–13: RNA-seq

Important topics

- Alternative splicing
- tophat
- partially ordered set
- cufflink & the graph algorithms as described
- RPKM and length bias
- Poisson
- likelihood ratio test

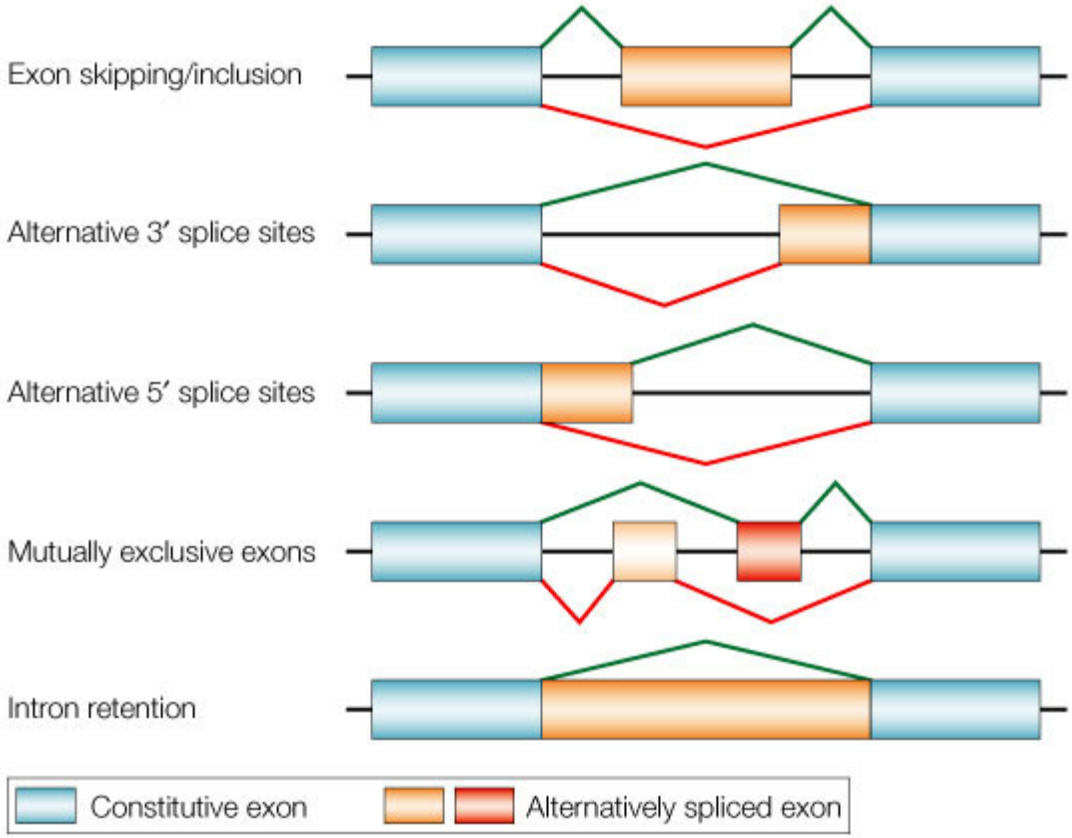
VL12–13: RNA-seq

Name two classes of alternative splicing events, and show with a sketch how a split read could distinguish between two isoforms of a gene with such an event.

(5 minutes)

Alternative splicing

Several different classes of alternative splicing events



Cartegni L et al. (2002) *Nature Reviews Genetics* **3**:285–298

Outline

VL1: Next-Generation Sequencing

VL8–9: Variant Calling

VL10–11: Structural Variants

VL12–13: RNA-seq

VL14: ChIP-seq

1 VL1: Next-Generation Sequencing

2 VL8–9: Variant Calling

3 VL10–11: Structural Variants

4 VL12–13: RNA-seq

5 VL14: ChIP-seq

VL14: ChIP-seq

Important topics

- the basic steps (wetlab) of Chip-seq
- 5' Read bias and how it is exploited for analysis of peaks
- mappability
- XSET algorithm
- Poisson
- MACS algorithm

VL14: ChIP-seq

Genomics 2012: Repetitorium

Peter N Robinson

VL1: Next-Generation Sequencing

VL8–9: Variant Calling

VL10–11: Structural Variants

VL12–13: RNA-seq

VL14: ChIP-seq

Explain the three major steps of the CHip-seq wetlab experiment

(5 minutes)

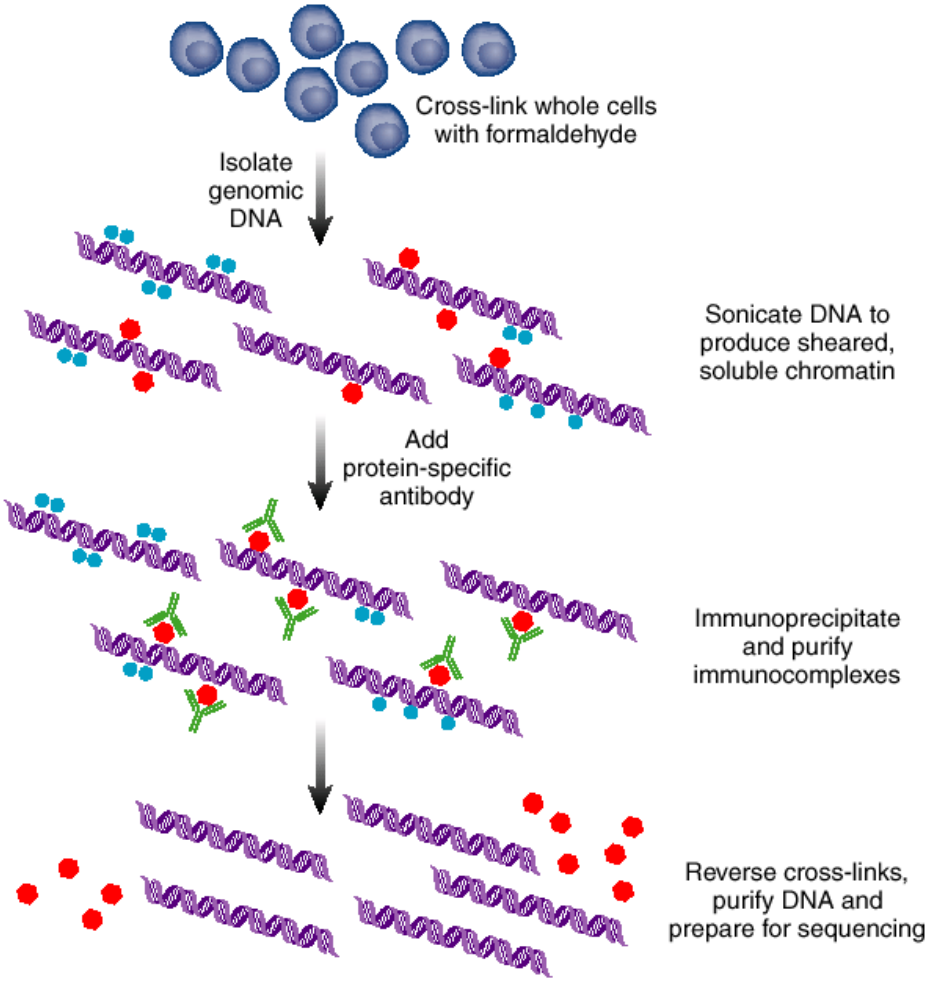
ChIP-Seq

Chromatin Immunoprecipitation following by next generation sequencing (**ChIP-seq**) is used to analyze protein interactions with DNA.

Three basic steps:

1. covalent cross-links between proteins and DNA are formed, typically by treating cells with formaldehyde
2. an antibody specific to the protein of interest is used to selectively coimmunoprecipitate the protein-bound DNA fragments that were covalently cross-linked.
3. the immunoprecipitated protein-DNA links are reversed and the recovered DNA is assayed to determine the sequences bound by that protein

ChIP-Seq: Workflow



Elaine R Mardis ER (2007) ChIP-seq: welcome to the new frontier. *Nature Methods* **4**:613–614

The Klausur

The final exam will take place on 12.02.2013 (Tuesday)

Good luck !

[1] often comes to those who are prepared for it!