DNA structure and function

Recombinant DNA technology

DNA amplification

DNA sequencing

Arthur Günzl, PhD
Dept. of Genetics & Genome Sciences
University of Connecticut Health Center

Bioinformatics and Computational Biology

Course (BME5800)

September 13, 2016
Deoxyribonucleic Acid (DNA)

The haploid human genome consists of ~3,000,000,000 bp
The combined α-helical length is ~2 m
A human nucleus’ diameter is ~10 μm, containing 46 chromosomes

From Alberts et al., Molecular Biology of the Cell
Which one is the reverse primer?

5’ - TTGGGAAGCTCCTTGTCA-3’
3’ - AACCCTTTGAGGAACAGT-5’

c. 5’ - TGACAAGGAGCTTCCCAA-3’

d. 5’ - GGTCCTGGGAGAAAAAGTCT-3’

e. None of the above
Recombination

Process in which DNA molecules are broken and the fragments are rejoined in new combinations

Recombinant DNA

Any DNA molecule formed by joining DNA segments from different sources
Restriction enzyme

One of a large number of endonucleases that can cleave a DNA molecule at any site where a specific short nucleotide sequence occurs.

5' - ...GTACCTAGAATTCCTAG... - 3'  
                                       ||||||||||           ||||||||
3' - ...CATGGATCTTAAGGATC... - 5'

\[\text{Eco RI} \text{ (sticky end)}\]

5' - ...GTACCTAGAATTCCTAG... - 3'  
                                       ||||            |||||
3' - ...CATGGATCTTAAGGATC... - 5'
<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Recognition Sequence</th>
<th>Microorganism</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>AluI</em></td>
<td>AG↓C*T</td>
<td><em>Arthrobacter luteus</em></td>
</tr>
<tr>
<td><em>BamHI</em></td>
<td>G↓GATC*C</td>
<td><em>Bacillus amyloliquefaciens H</em></td>
</tr>
<tr>
<td><em>BglI</em></td>
<td>GCCNNNN↓NGCC</td>
<td><em>Bacillus globigii</em></td>
</tr>
<tr>
<td><em>BglII</em></td>
<td>A↓GATCT</td>
<td><em>Bacillus globigii</em></td>
</tr>
<tr>
<td><em>EcoRI</em></td>
<td>G↓AA*TTC</td>
<td><em>Escherichia coli RY13</em></td>
</tr>
<tr>
<td><em>EcoRII</em></td>
<td>↓CC*(†)GG</td>
<td><em>Escherichia coli R245</em></td>
</tr>
<tr>
<td><em>EcoRV</em></td>
<td>GA*T↓ATC</td>
<td><em>Eschericia coli J62P7G74</em></td>
</tr>
<tr>
<td><em>HaeII</em></td>
<td>RGCGC↓Y</td>
<td><em>Haemophilus aegyptius</em></td>
</tr>
<tr>
<td><em>HaeIII</em></td>
<td>GG↓C*C</td>
<td><em>Haemophilus aegyptius</em></td>
</tr>
<tr>
<td><em>HindIII</em></td>
<td>A*↓AGCTT</td>
<td><em>Haemophilus influenzae R_d</em></td>
</tr>
<tr>
<td><em>HpaII</em></td>
<td>C↓C*GG</td>
<td><em>Haemophilus parainfluenzae</em></td>
</tr>
<tr>
<td><em>MspI</em></td>
<td>C*↓GG</td>
<td><em>Moraxella species</em></td>
</tr>
<tr>
<td><em>PstI</em></td>
<td>CTGCA*↓G</td>
<td><em>Providencia stuartii 164</em></td>
</tr>
<tr>
<td><em>PvuII</em></td>
<td>CAG↓C*TG</td>
<td><em>Proteus vulgaris</em></td>
</tr>
<tr>
<td><em>SalI</em></td>
<td>G↓TCGAC</td>
<td><em>Streptomyces albus G</em></td>
</tr>
<tr>
<td><em>TaqI</em></td>
<td>T↓CGA*</td>
<td><em>Thermus aquaticus</em></td>
</tr>
<tr>
<td><em>XhoI</em></td>
<td>C↓TCGAG</td>
<td><em>Xanthomonas holcicola</em></td>
</tr>
</tbody>
</table>

*SmaI* CCC↓GGG *Serratia marcescens*
Recombinant DNA Technology

Key Enzymes

DNA ligase

Enzyme that joins the ends of two DNA strands together with a covalent bond to make a continuous DNA strand

\[
\begin{align*}
5'&-\ldots G T A C C T A C C C -3' \quad 5' - p G G G C T A G T G \ldots -3' \\
3'&-\ldots C A T G G A T G G G p -5' \quad 3' - C C C G A T C A C \ldots -5'
\end{align*}
\]

T4 DNA ligase + ATP

\[
\begin{align*}
5'&-\ldots G T A C C T A C C C G G G C T A G T G \ldots -3' \\
3'&-\ldots C A T G G A T G G G C C C G A T C A C \ldots -5'
\end{align*}
\]
Recombinant DNA Technology

plasmid cloning vector

pUC18
2686 bp

Ampicillin resistance
LacZα
Multiple Cloning Sites

HindIII
SphI
PstI
SalI
XbaI
BamHI
SmaI
Asp718I
KpnI
SacI
EcoRI
Gene cloning

plasmid vector → RESTRICTION DIGEST → DNA LIGASE (ATP) → recombinant plasmid
Transformation of *E. coli*

The bacteria will not take up plasmid DNA from its environment. It has to be mistreated to do so, e.g. the membrane has to be damaged to allow for DNA entry.

*E. coli* is made *competent* for transformation by

- CaCl$_2$ / heat shock
- electroporation

Transformed *E. coli* are selected via expression of a selectable marker (*Amp*$_r$)
Plasmid DNA does not integrate into the bacterial genome. It replicates autonomously typically in high copy numbers. Due to its small size (~5 kbp) it can be separated from genomic DNA (~3,700 kbp) and efficiently purified.
How to obtain DNA fragments for cloning?

(Note: Historic slide)

 Libraries:

Genomic DNA – fragmentation (mechanical, sonication, enzymatic) → cloning

mRNA -> cDNA → cloning

Nowadays, genome sequences are known or relatively easy to obtain. Thus, it is possible to amplify a DNA sequence of interest and either use it directly in an assay or clone it into a plasmid construct.
DNA is synthesized by DNA polymerase

\[
5' - \ldots \text{GpApApTpT}_{\text{OH}} \quad P_\gamma P_\beta P_\alpha C_{\text{OH}}
\]

\[
3' - \ldots \text{CpTpTpApApGpApA} \ldots -5'
\]

\[
5' - \ldots \text{GpApApTpTpC}_{\text{OH}}
\]

\[
3' - \ldots \text{CpTpTpApApGpApA} \ldots -5'
\]
Amplification of DNA fragment of interest by the Polymerase Chain Reaction (PCR)

Heating denatures DNA (separates strands)
Primers anneal to complimentary strands
DNA polymerase elongates from primer annealing

Multiple reaction cycles result in an exponential increase of specific amplification product

Kary B Mullis obtained the Nobel prize for Chemistry in 1993 for inventing PCR
[real-time] quantitative (q)PCR

Semi-quantitative (RT-)PCR

Ct (cycle threshold) is defined as the number of cycles required for the fluorescent signal to cross the threshold
SYBR Green binds to DNA. The resulting DNA-dye-complex absorbs blue light ($\lambda_{\text{max}} = 497$ nm) and emits green light ($\lambda_{\text{max}} = 520$ nm). The stain preferentially binds to double-stranded DNA.

The TaqMan probe principle relies on the 5′–3′ exonuclease activity of Taq polymerase to cleave a quenched dual-labeled probe that binds to the target sequence. This method is highly specific and, due to the availability of different fluorophores, it can be applied to multichannel applications that measure more than one amplification in the same reaction.
DNA Sequencing

1st and “Next” generation sequencing (NGS) is predominantly based on the “chain termination” ("enzymatic", "dideoxy") method invented by Frederick Sanger *et al.* (1977, PNAS 74, pp 5463-5467).

The principle is based on dideoxy nucleotides in which the 3’ hydroxyl group, which is essential for formation of the next phosphodiester bond, is replaced by a hydrogen.

These nucleotides are so-called chain terminators because they prevent the incorporation of the next nucleotide and cause DNA synthesis to stop.

Frederick Sanger and Walter Gilbert obtained the Nobel prize for Chemistry in 1980 for *Determination of base sequences in nucleic acids*.
Ribose sugars

ribose
deoxyribose
di-deoxyribose
DNA Sequencing

oligonucleotide primer for DNA polymerase

5′-...GCTACCTGCATGGGAA-3′

3′-...CGATGGACGTACCTCTCTGAAGCG...-5′

normal deoxyribonucleotides
dGTP, dATP, dCTP, dTTP

rare incorporation of one ddATP by DNA polymerase blocks further growth of the DNA molecule

single-stranded DNA molecule (template DNA) to be sequenced
DNA Sequencing

From Alberts et al., Molecular Biology of the Cell, p 505
Polyacrylamide Gel Electrophoresis (PAGE)

PAGE apparatus

Direction of DNA migration

autoradiography
Modern Automatic Sequencing works in one reaction with fluorescent-labeled terminators
Next (2\textsuperscript{nd}) generation sequencing technology:

454 sequencing (pyrosequencing)
illumina solexa sequencing

Advantages

\textbf{fast} (Hi-seq 4000: per 3.5 day run: 12 human genomes or 1.5 Tb of sequence data (~5 billion reads).

\textbf{cost-effective} (reduced cost per base)

\textbf{without cloning} (PCR integrated in the method)

Disadvantages

\textbf{short reads} (454: 1000 bp/read)/Solexa: up to 250 bp/read)

intense on bioinformatics
“next generation sequencing”

Sequencing the nuclear genome of the extinct woolly mammoth

Webb Miller¹, Daniela I. Drautz¹, Aakrosh Ratan¹, Barbara Pusey¹, Ji Qi², Arthur M. Lesk¹, Lynn P. Tomchick², Michael D. Packard¹, Fangqing Zhao¹, Andrei Sher²,³, Alexei Tikhonov³, Brian Raney⁴, Nick Patterson⁵, Kerstin Lindblad-Toh⁵, Eric S. Lander⁵, James R. Knight⁵, Gerard P. Irzyk⁶, Karin M. Fredrikson⁷, Timothy T. Sharon Sheridan⁷, Tom Pringle⁸ & Stephan C. Schuster¹


A Draft Sequence of the Neandertal Genome

Richard E. Green,²,⁴†† Johannes Krause,²,††³‡‡ Udo Stenzel,³‡‡ Martin Kircher,⁴‡‡ Nick Patterson,⁵ Carsten Russ,² Nathaniel North⁵, Christine Verna,²¹ Pavao Rudan,¹⁰ Dejana Banić,²¹ Vladimir B. Dorničevič,¹² Liubov V. Golovanova,¹² Javier Fortea,¹³† Antonio Rosas,³² Ralf W. Schmitz,¹³ Daniel Falush,¹⁹† Evan Birney,†† James C. Mullikin⁵,‡‡ Janet Kelso,¹†† Michael Lachmann,¹†† David Reich,²,¹⁰⁺⁺ Svante Pääbo²,⁴†

Neandertals, the closest evolutionary relatives of present-day humans, lived in large parts of Europe and western Asia before disappearing 30,000 years ago. We present a draft sequence of the Neandertal genome composed of more than 4 billion nucleotides from three individuals. Comparisons of the Neandertal genome to the genomes of five present-day humans from different parts of the world identify a number of genomic regions that may have been affected by positive selection in ancestral modern humans, including genes involved in metabolism and in cognitive and skeletal development. We show that Neandertals shared more genetic variants with present-day humans in Eurasia than with present-day humans in sub-Saharan Africa, suggesting that gene flow from Neandertals into the ancestors of non-Africans occurred before the divergence of Eurasian groups from each other.

Science 2010, 328:710-722

ARTICLE

Single-cell RNA-seq reveals dynamic paracrine control of cellular variation

Alex K. Shalek,¹,²,³⁺⁺ Rahul Satija,⁴⁺⁺ Joe Shuga,⁴⁺⁺ John J. Trombetta,³ Dave Guns et al., Diana Liu², Peilin Chen¹, Rona S. Gartner¹,², Jellert T. Gamblin,²,³ Rachel Schwartz,³ Brian Fowler,⁴ Suzanne Weaver⁴, Jing Wang⁴, Xiaohui Wang⁴, Ruifeng Ding²,³, Rakatina Raychowdhury², Nair Friedman², Nair Hacohen²,³, Hongkun Park²,³, Andrew P. May² & Aviv Regev³,⁴

Nature 2014, 510:363-369
Solexa technology (http://www.illumina.com)

step 1 - prepare genomic DNA sample

Prepare genomic DNA sample
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments
454 pyrosequencing

The GSflx can sequence 100 megabases of raw DNA sequence per 4.5-hour run.

1. Attach DNA fragments to beads
2. Amplify the DNA
3. Place beads in microscopic wells (44 μm in diameter / PicoTiterPlate)
   - A mix of bead-bound enzymes are also packed into the well.

The plate is flowed successively with oligonucleotides and light flashes are measured.
454 pyrosequencing

light signal

- Polymerase
- Annealed primer
- Sulfurylase
- Luciferase
- ATP
- Luciferin
- Light + Oxy Luciferin
- APS
- PPi
The solexa technology
step 2 - immobilize DNA on flow cell surface

Attached DNA to surface
Bind single stranded fragments randomly to the inside surface of the flow cell channels.
Solexa technology
step 3 - amplify DNA to immobilized clusters

Bridge amplification
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.
Solexa technology
step 3 - amplify DNA to immobilized clusters

Completion of amplification
On completion, several million dense clusters of double stranded DNA are generated in each channel of the flow cell.
Solexa technology
step 4 - sequence one nucleotide at a time

First chemistry cycle:
determine first base
Wash off all unincorporated reagents

Laser
Solexa technology
step 4 - sequence one nucleotide at a time
DNA polymerization requires the 3’ hydroxyl group
Next generation sequencing has enabled many innovative high throughput approaches

RNA-seq: Transcriptome; quantification of transcript levels, discovery of novel transcripts and transcript isoforms (alternative splicing).
Next generation sequencing has enabled many innovative high throughput approaches.

Polysome and ribosome profiling: Translatome

From Piccirillo et al., 2014, *Nature Immunology*
Next generation sequencing has enabled many innovative high throughput approaches

Chromatin Immunoprecipitation (ChIP)-seq: Genome-wide binding sites of a histone, transcription factor, RNA polymerase, DNA repair enzyme, etc.

ChIP-seq revealed that open chromatin marks (modified histones and histone variants) identified regions of transcription initiation in *Trypanosoma brucei*.
Third generation (single molecule) sequencing
Pacific Biosciences Single molecule real time (SMRT) technology

Reads: 3-5 kb, up to >20 kb
no DNA amplification
Highly accurate
Can detect minor variants, <1%
Modified nucleotide detection

See video http://www.pacb.com/smrt-science/smrt-sequencing/
Nanopore sequencing

A SOLID STATE NANOPORE ARTICULATED WITH PROBES

An electrode measures variations in the current

No polymerization
No library preparation
No light detection
Sequence accuracy not high
Not yet enough reads for transcriptomes

Oxford Nanopore Technologies®
THE END

Any questions - contact me!

Arthur Gunzl

office R1144 (400 Farmington Ave)
phone 8878
E-mail gunzl@uchc.edu