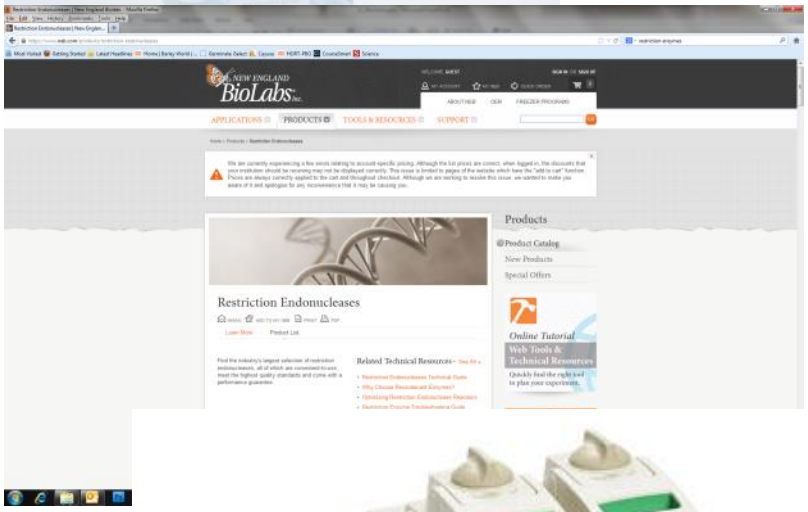
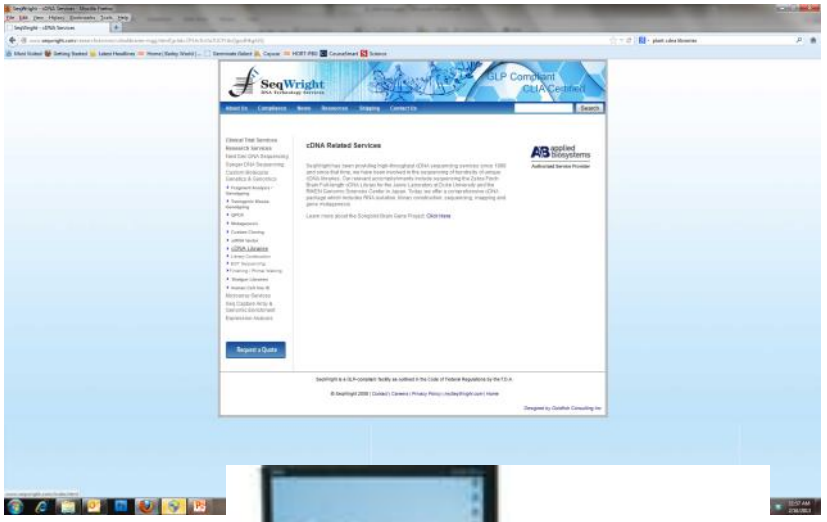
Molecular Tools





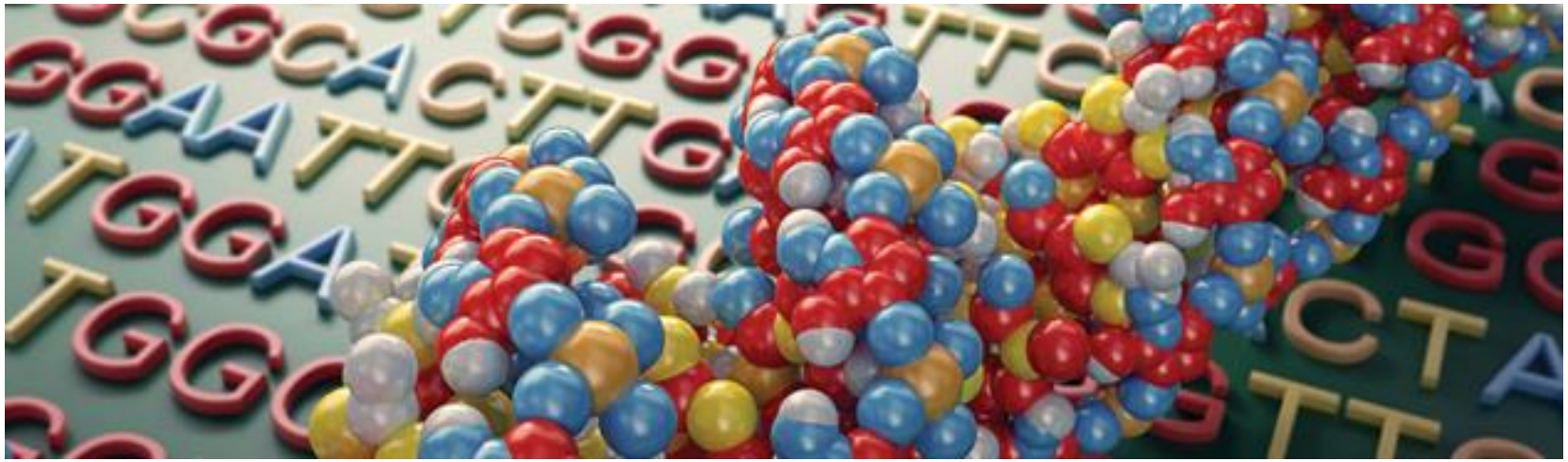
Steps in Genetic Analysis

1. Knowing how many genes determine a phenotype, and where the genes are located, is a first step in understanding the genetic basis of a phenotype

2. *A second step is determining the sequence of the gene, or genes, determining the phenotype and understanding how the expression of the genes is regulated at the transcriptional level*

3. Subsequent steps involve analysis of post-transcriptional events, understanding how the genes fit into metabolic pathways and how these pathways interact with the environment

Barley DNA Sequence



- Total sequence is 5,300,000,000 base pairs
  - 165 % of human genome
  - Enough characters for 11,000 large novels
- Expressed Genes - 60,000,000 base pairs
  - ~ 1% of total sequence, like humans
  - 125 large novels

Step 2 & Sequencing

1. Complete genome sequences are coming, but aren't yet available for many plants

2. *The trend is sequencing with multiple applications - e.g. whole genomes, specific targets within genomes, or genotyping by sequencing (GBS)*

3. Even when complete genome sequence information is available for every plant, there will always be reason to study allelic diversity and interactions at specific loci and to compare genome sequences of multiple individuals

Molecular Tools for Step 2

1. Getting DNA
   Can be a rate limiting step, unless automated

2. Cutting the DNA with *restriction enzymes*
   Reducing complexity

3. Managing the pieces of DNA in *vectors (*or *alternatives)*; collections of pieces are maintained in *libraries*

4. Selecting DNA targets via *amplification* and/or *hybridization*

5. Determining *nucleotide sequence* of the targeted DNA

Extracting DNA

Genomic DNA: Order your kit today!

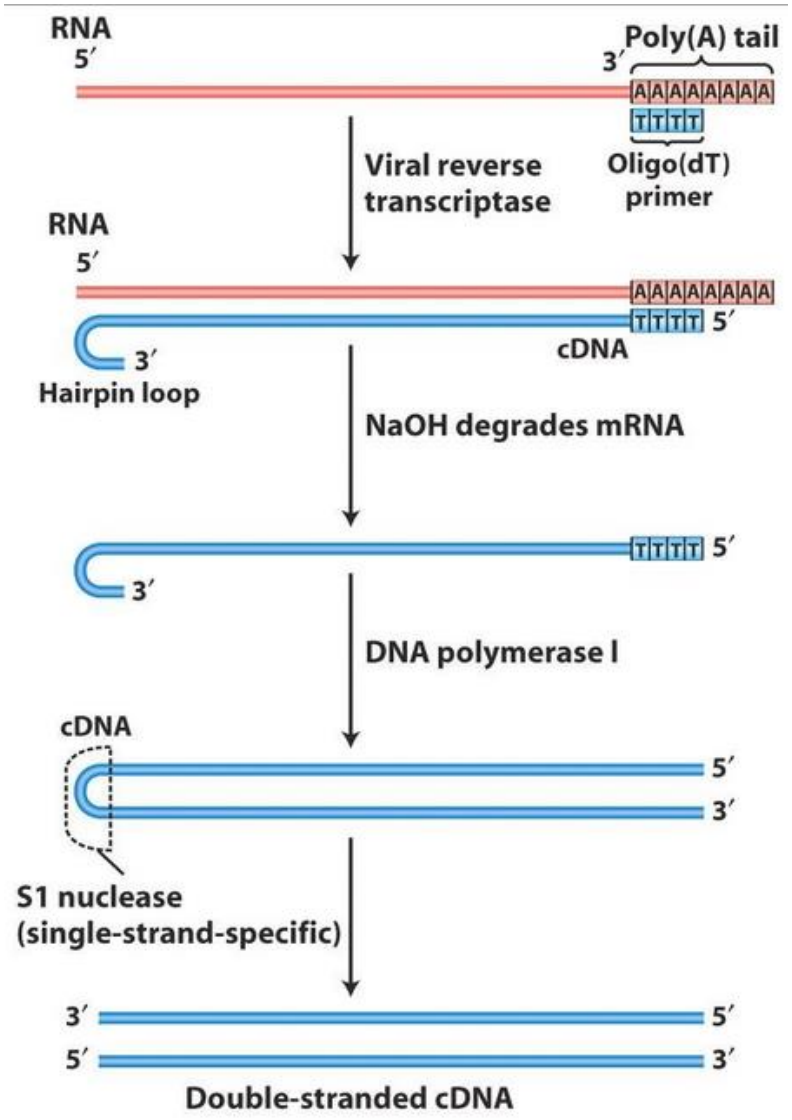
- One-by-one (artisanal) to high-throughput (DNA from seed chips + robotics)
- Leaf segments to cheek swabs

Key considerations are

- Concentration
- Purity
- Fragment size

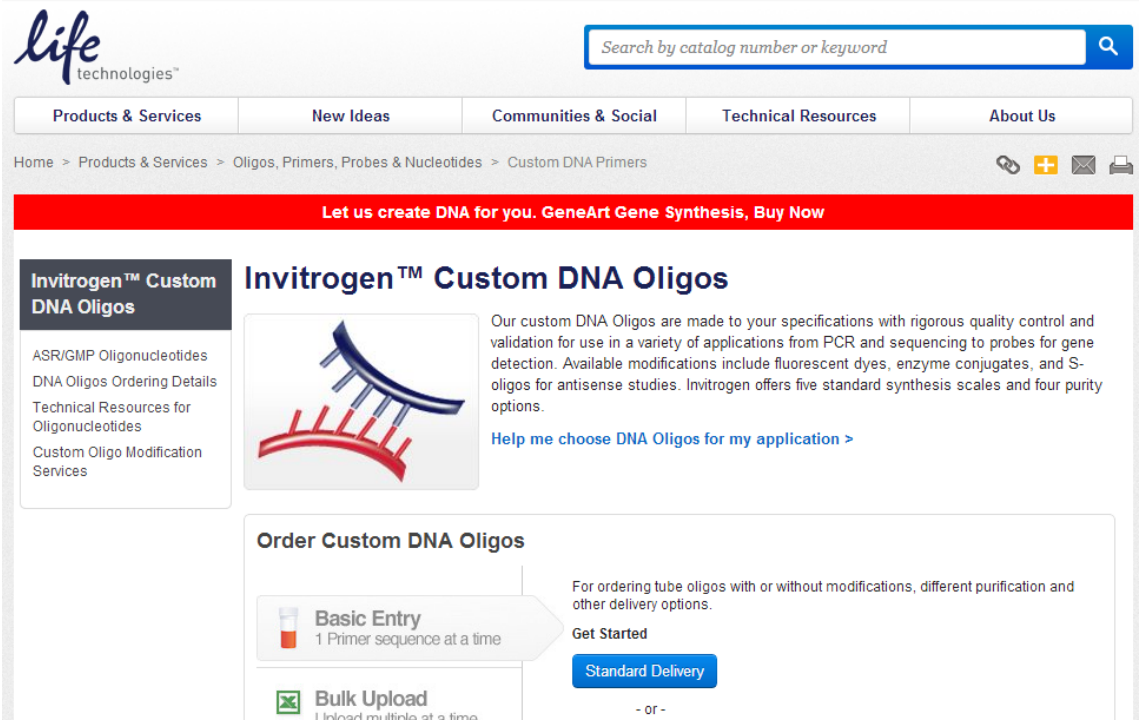
Making DNA – the cDNA Way

cDNA: From mRNA to DNA



Making synthetic DNA: oligonucleotides

Primers, adapters, and more …~$0.010 per bp...



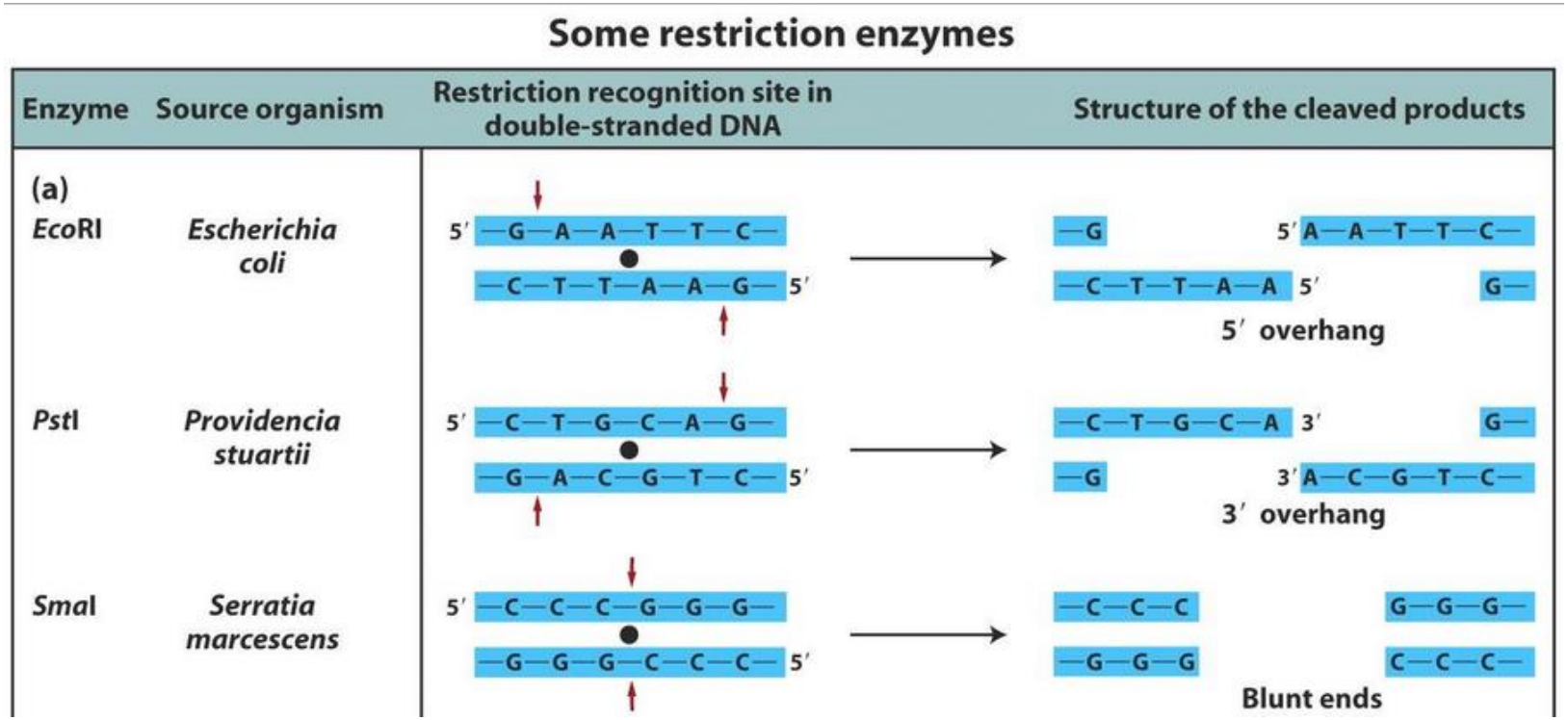
Synthetic organisms??? ~ 1million bp synthetic so far….

Cutting the DNA – Restriction Enzymes

Restriction enzymes make cuts at defined recognition sites in DNA

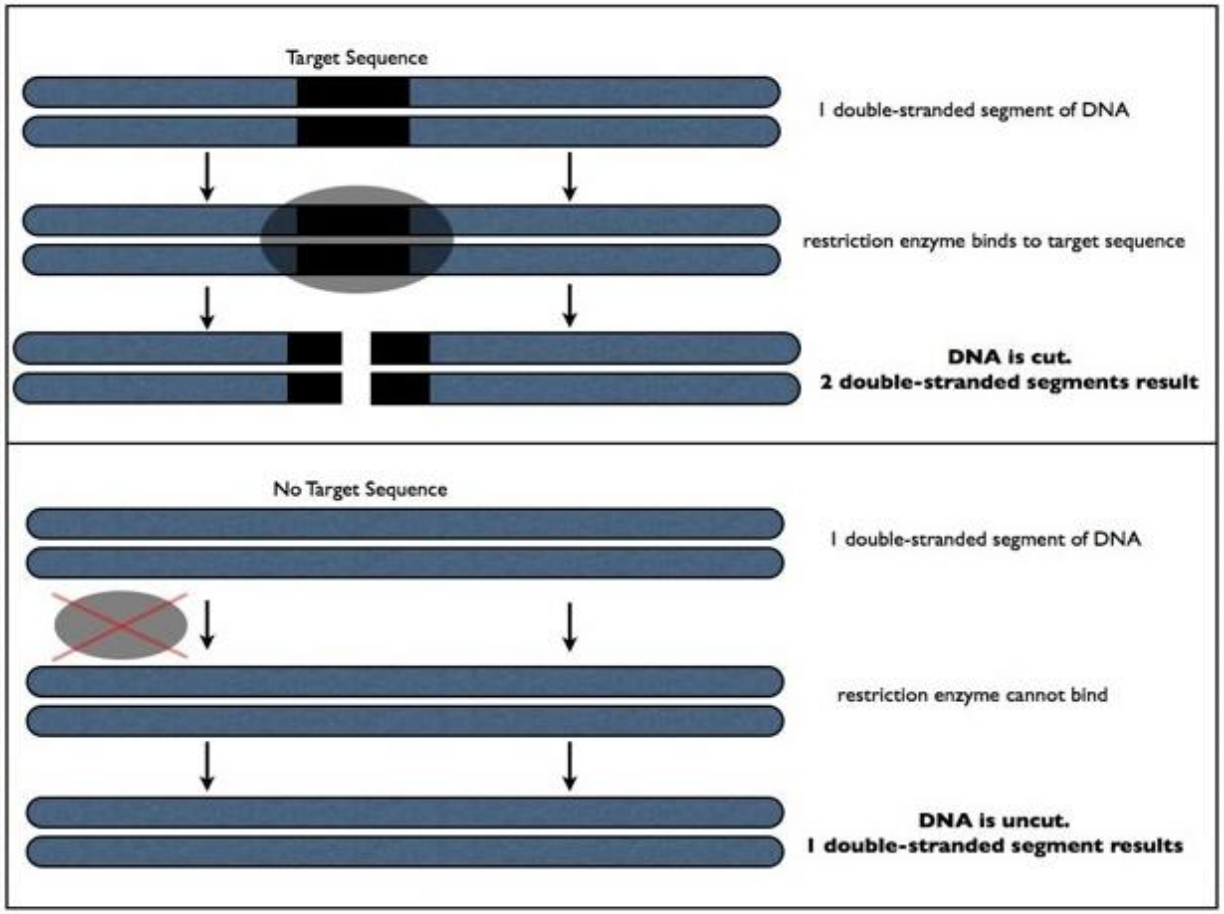
- A defense system for bacteria, where they attack and degrade the DNA of attacking bacteriophages
- The restriction enzymes are named for the organism from which they were isolated
- Harnessed for the task of systematically breaking up DNA into fragments of tractable size and for various polymorphism detection assays
- Each enzyme recognizes a particular DNA sequence and cuts in a specified fashion at the sequence

Cutting the DNA – Restriction Enzymes



Cutting the DNA – Restriction Enzymes

To test for a mutation using restriction endonucleases, the segment of the patient's DNA containing the target sequence is mixed with restriction enzyme. If the target sequence is present, the enzyme is able to cut the DNA (this in turn means that the patient is carrying the HNPCC mutation). If the target sequence is not present, the enzyme will not be able to cut the DNA (meaning the patient does not carry the HNPCC mutation).



Whether the DNA has been cut or not can be determined through a technique called gel electrophoresis.

Cutting the DNA – Restriction Enzymes

- An enzyme that has a four-base recognition site will cut approximately every 256 bp (4^4) and more frequently than one with a six base recognition site, which in turn will cut more often than one with an eight base recognition site
- Methylation sensitivity:
  - Avoid repetitive DNA in order to focus on coding regions
  - Target the epigenome

Cutting the DNA – Restriction Enzymes

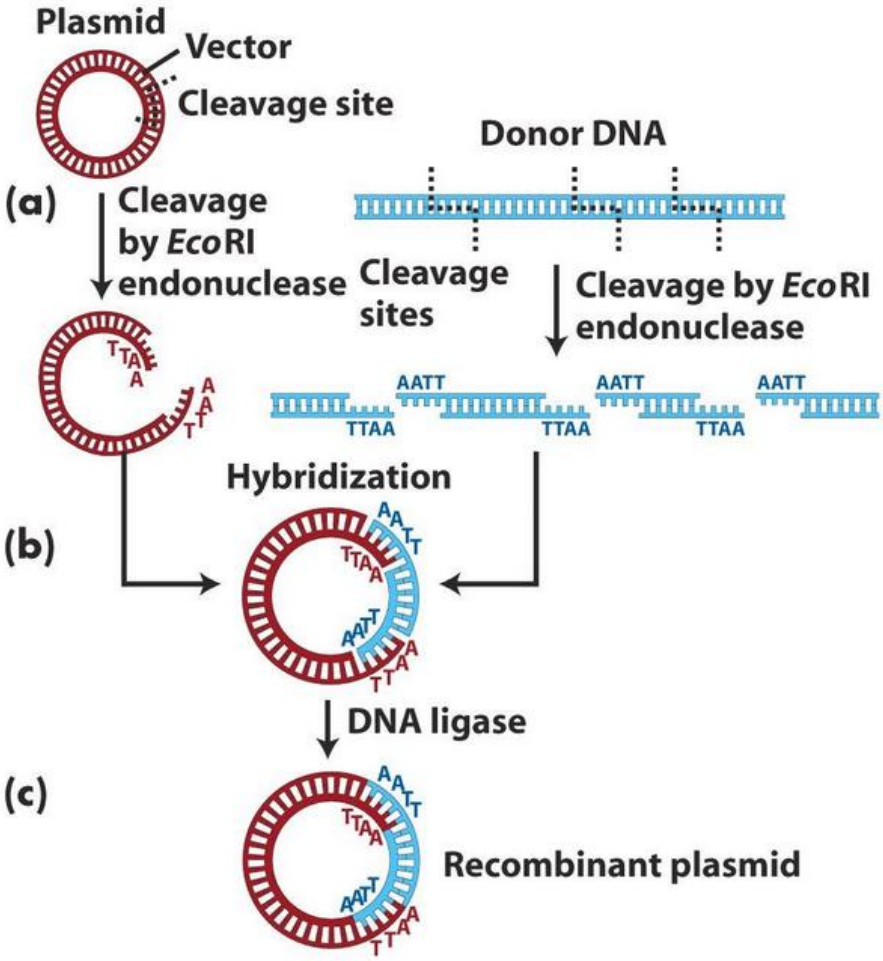
- *Palindrome recognition sites* – the same sequence is specified when each strand of the double helix is read in the opposite direction

Sit on a potato pan, Otis
Cigar? Toss it in a can, it is so tragic
UFO. tofu
Golf? No sir, prefer prison flog
Flee to me remote elf
Gnu dung
Lager, Sir, is regal
Tuna nut

DNA: Vectors and libraries

Vectors: The role of the vector is to propagate and maintain the DNA fragments generated by the restriction digestion

- Efficiency and simplicity of inserting and retrieving the inserted DNA fragments
- A key feature of the cloning vector is size of the DNA fragment insert that it can efficiently and reliably handle
- *Example: the principle of cloning a DNA fragment in a plasmid vector*



DNA Libraries – Vectors

Common vectors and approximate insert sizes

Vector	**Insert size(kb)**
Plasmid	~ 1
Lambda phage	~ 20
Bacterial Artificial Chromosomes (BAC)	~ 200

DNA Libraries

Libraries are repositories of DNA fragments cloned in their vectors **or** platform-specific oligonucleotide adapters subsequent use.

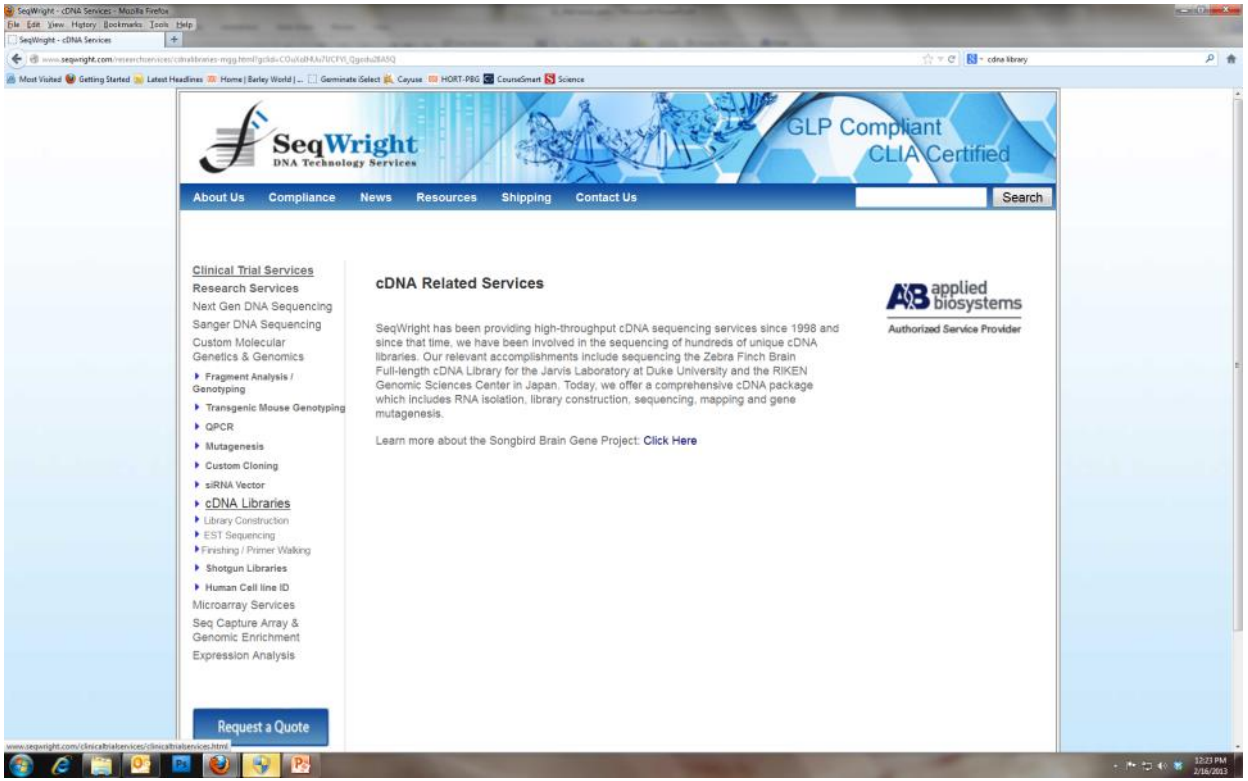
- Libraries can be classified based
  - on the cloning vector – e.g. plasmid, BAC
  - In terms of the source of the cloned DNA fragments – e.g. genomic, cDNA
  - In terms of intended use: next generation sequencing (NGS), genotyping by sequencing (GBS)

DNA Libraries – Genomic

- Total genomic DNA digested and the fragments cloned into an appropriate vector or system
- In principle, this library should consist of samples of all the genomic DNA present in the organism, including both coding and non-coding sequences
- Ideally, every copy of every gene (or a portion of every sequence) should be represented somewhere in the genomic library
- There are strategies for enriching genomic libraries for specific types of sequences and removing specific types of sequences – e.g. favoring unique vs. highly repetitive sequences

DNA Libraries – cDNA

- A cDNA (complementary DNA) library is generated from mRNA transcripts, using the enzyme *reverse transcriptase*, which creates a DNA complement to a mRNA template
- The cDNA library is based on mRNA: therefore the library will represent only the genes that are expressed in the tissue and/or developmental stage that was sampled

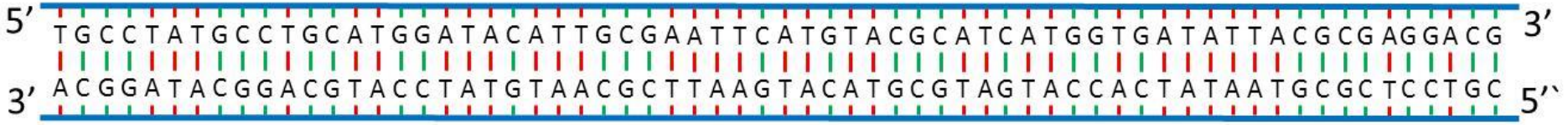


DNA Amplification: Polymerase Chain Reaction (PCR)

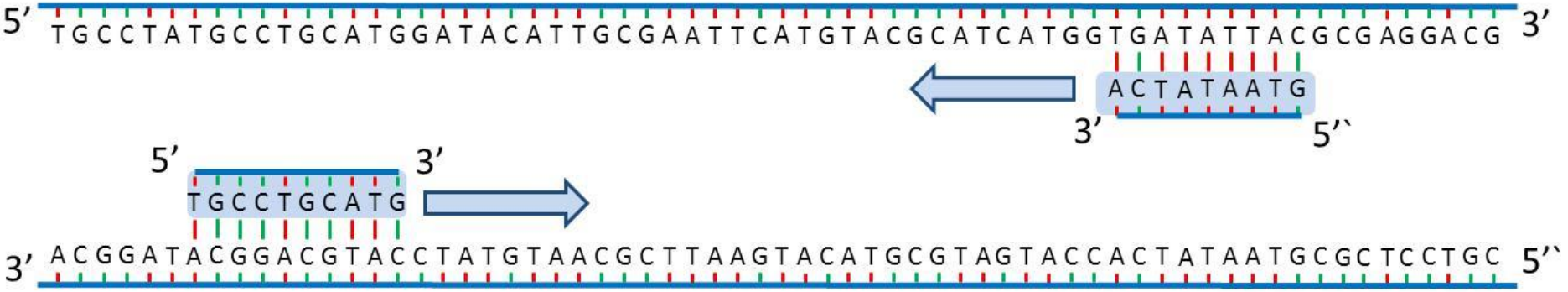
- Invented by K.B Mullis in 1983
- Allows in vitro amplification of ANY DNA sequence in large numbers

https://www.youtube.com/watch?v=2KoLnIwoZKU

DNA Amplification: Polymerase Chain Reaction (PCR)

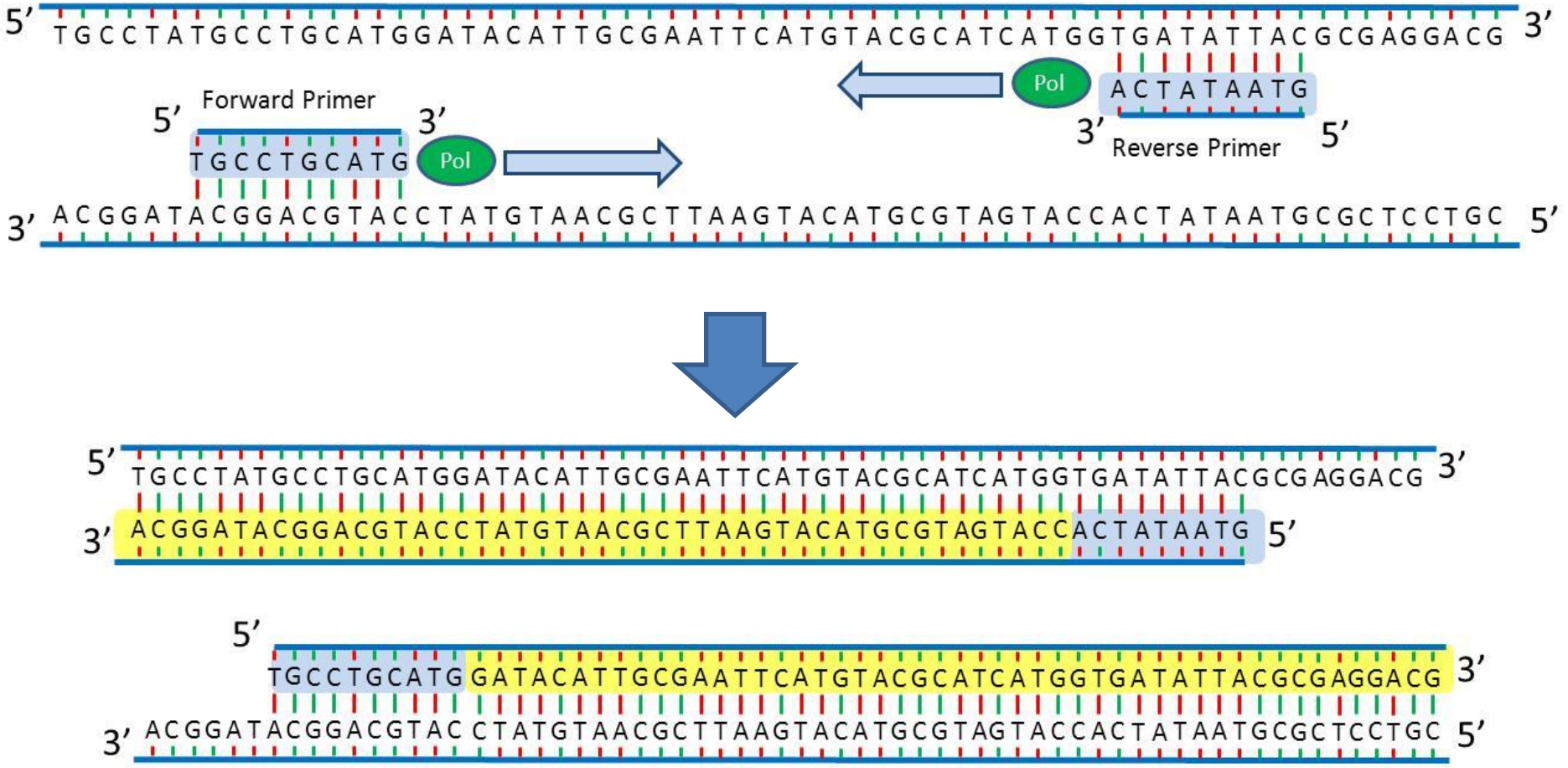


Design of two single stranded oligonucleotide primers complementary to motifs on the template DNA.



DNA Amplification - PCR

A Polymerase extends the 3’ end of the primer sequence using the DNA strand as a template.



DNA Amplification - PCR Principles

- The PCR reaction has the following steps:
  - Denaturing: raising the temperature to 94 C to make DNA single stranded
  - Annealing: lowering the temperature to 35 – 65 C the primers bind to the target sequences on the template DNA
  - Elongation: DNA polymerase extends the 3’ ends of the primer sequence. Temperature must be optimal for DNA polymerase activity.

DNA Amplification - PCR Principles

- Each cycle can be repeated multiple times if the 3’ end of the primer is facing the target amplicon. The reaction is typically repeated 25-50 cycles.

- Each cycle generates exponential numbers of DNA fragments that are identical copies of the original DNA strand between the two binding sites.

- The PCR reaction consists of:
  - A buffer
  - DNA polymerase (thermostable)
  - Deoxyribonucleotide triphosphates (dNTPs)
  - Two primers (oligonucleotides)
  - Template DNA
  - Labelling as required

DNA Amplification - PCR Principles

- Denaturing: raising the temperature to 94 C to make DNA single stranded
- Annealing: lowering the temperature to 35 – 65 C the primers bind to the target sequences on the template DNA
- Elongation: DNA polymerase extends the 3’ ends of the primer sequence. Temperature must be optimal for DNA polymerase activity.

DNA Amplification - PCR Priming

- The choice of what DNA will be amplified by the polymerase is determined by the **primers** (short pieces of synthesized DNA - *oligonucleotides*) that prime the polymerase reaction
- The DNA between the primers is amplified by the polymerase: in subsequent reactions the original template, plus the newly amplified fragments, serve as templates
- Steps in the reaction include **denaturing** the target DNA to make it single-stranded, addition of the **single stranded oligonucleotides**, **hybridization of the primers** to the template, and **primer extension**

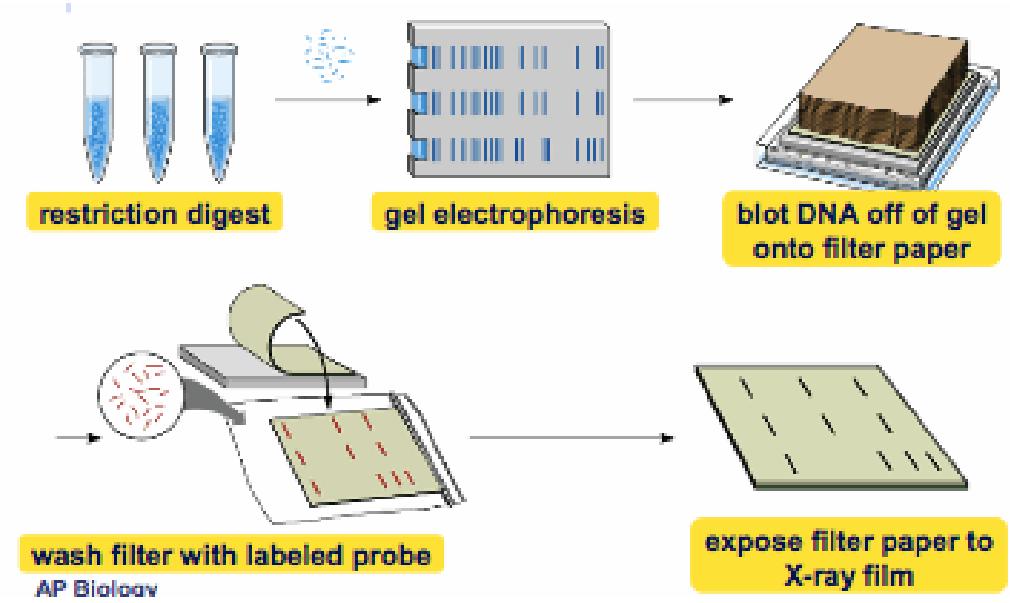
DNA Amplification - PCR Application

- The PCR process is repeated as necessary until the target fragment is sufficiently amplified that it can be isolated, visualized, and/or manipulated
- A key component of PCR is a thermostable polymerase, such as **TAQ polymerase**
- PCR can be used to amplify rare fragments from a pool of DNA, generate an abundance of a particular fragment from a single copy from a small sample (e.g. fossil DNA), generate samples of all DNA in a genome, and it is the foundation for many types of molecular markers

https://www.youtube.com/watch?v=2KoLnIwoZKU

DNA Hybridization

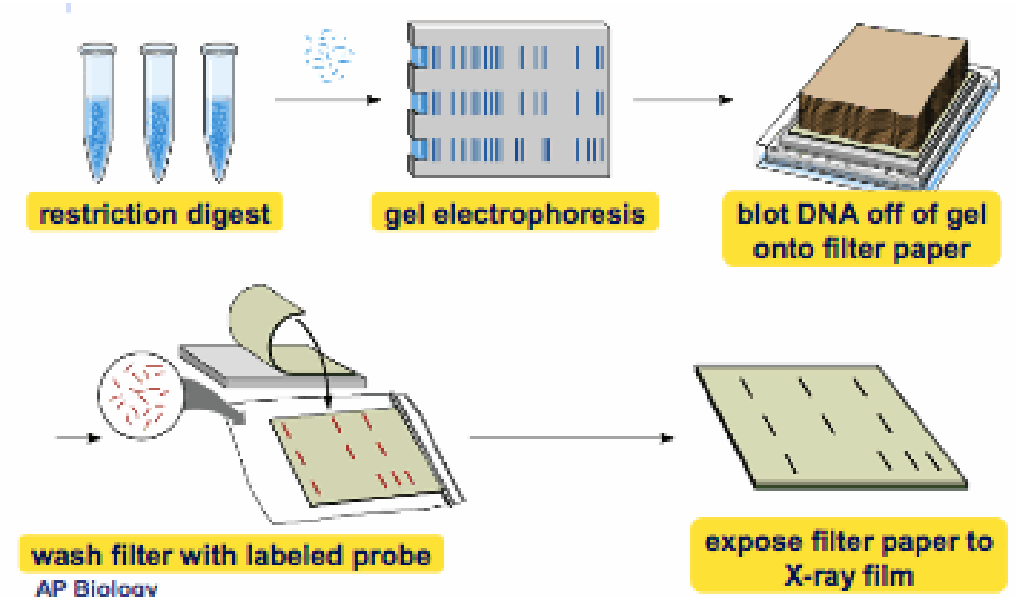
- Single strand nucleic acids have a natural tendency to find and pair with other single strand nucleic acids with a complementary sequence
- An application of this affinity is to label one single strand with a tag – radioactivity and fluorescent dyes are often used - and then to use this *probe* to find complementary sequences in a population of single stranded nucleic acids
- For example, if you have a cloned gene – either a cDNA or a genomic clone - you could use this as a probe to look for a homologous sequence in another DNA sample



DNA Hybridization

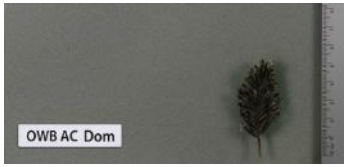
- By denaturing the DNA in the sample, and using your labeled single stranded probe you can search the sample for the complementary sequence
- Pairing of probe and sample can be visualized by the label – e.g. on X-ray film or by measuring fluorescence
- The principle of hybridization can be applied to pairing events involving DNA: DNA; DNA: RNA; and protein: antibody

Southern blot
Northern blot
Western blot



Sequencing the DNA

- Advances in technology have removed the technical obstacles to determining the nucleotide sequence of a gene, a chromosome region, or a whole genome
- The starting point for any sequencing project – be it of a single cloned fragment or of an entire genome - is a defined fragment of DNA



>**Hox1_101** [cultivar=Oregon Wolfe Barley Dominant (OWB-D)] [type=two-rowed] [haplotype=Vrs1.t]
AgcACTCAGCCGGAGCCCCGCCGGCGCTTGCCGTTGGGTACCTCTGCCACCTATTTATATTACCCCTAGGTCTCTCCCTGGAGACACGCACTC
CCCTCCTTCAACTAGTGCTTTGCGGCCCGTGGTCCTCCTCTCGATCCAGTTCCTGAGCACACCAACAGGCAACAGAACAACCTACCGTGTCTC
CCCTCCAATCTCCTCACGATCCCTTCTTTCCCTCAGATCCGAACCGAAAGCATGGACAAGCATCAGCTCTTTGGTTCATCCAACGTGGACACG
ACTTTCTTCGCGGCCAATGGTACACACGACGCCGCGCGCGCCCGGTCTTTGCGCATGCGATGATGCAGCTGCAGTAGCTTCAGTTTCACCGGC
CAGGACACGCATGTGATGACGTTTTTTCCATTCTGTGTTTGTATGTGCAGGCACGGCGCAGGGGGAGACCAGCAAGCAGAGGGCGCGGCGCAG
GCGGCGGAGGTCGGC-GAGGTGCGGCGGAGGGGATGGTGACGGTGGGGAGATGGACGGAGGAGGGGACCCCAAGAAGCGGCGGCTCACCGACG
AGCAGGCCGAGATTCTGGAGCTGAGCTTCCGGGAGGACCGCAAGCTGGAGACAGCCCGCAAGGTGTATCTGGCCGCCGAGCTCGGGCTGGACC
CCAAGCAGGTCGCCGTGTGGTTCCAGAACCGCCGCGCGCGCCACAAGAACAAGACGCTCGAGGAGGAGTTCGCGAGGCTCAAGCACGCCCACG
ACGCCGCCATCCTCCACAAATGCCACCTCGAGAACGAGGTATGCTTGCTCGCATACATTCACACTGGCTTACATATGGCGCTGCACATCTGCA
GTTCCTCTCCGTTCTTGAACATGCTTACTGACAAACATATGGCCAGCTGCTGAGGCTGAAGGAGAGACTGGGAGCGACTGAGCAGGAGGTGCG
GCGCCTCAGGTCGGCAGCTGGGAGCCACGGGGCATCTGTGGATGGCGGACACGCCGCTGGCGCCGTTGGCGTGTGCGGCGGGGGCCCGAGCTC
GTCCTTCTCGACGGGAACCTGCCAGCAGCAGCCGGGTTTCAGCGGGGCAGACGTGCTGGGGCGGGACGATGACCTGATGATGTGCGTCCCCGA
GTGGTTTTTAGCATGAATTAGAGTTTATGCTGGCTAAGCCGATAGCAGCGTGGTCGAGTGTTTTTTAGCATGAAAT

>**Hox1_102** [cultivar=Oregon Wolfe Barley Recessive (OWB-R)] [type=six-rowed] [haplotype=vrs1.a3]
AGCACTCAGCCGGAGCCCCGCCGGCGCTTGCCGTTGGGTACCTCTGCCACCTATTTATATTACCCCTAGGTCTCTCCCTGGAGACACGCACTCCC
CTCCTTCAACTAGTGCTTTGCGGCCCGTGGTCCTCCTCTCGATCCAGTTCCTGAGCACACCAACAGGCAACAGAACAACCTACCGTGTCTCCCCT
CCAATCTCCTCACGATCCCTTCTTTCCCTCAGATCCGAACCGAAAGCATGGACAAGCATCAGCTCTTTGGTTCATCCAACGTGGACACGACTTTC
TTCGCGGCCAATGGTACACACGACGCCGCGCGCGCCCGGTCTTTGCGCATGCGATGATGCAGCTGCAGTAGCTTCAGTTTCACCGGCCAGGACAC
GCATGTGATGACGTTTTTTCCATTCTGTGTTTGTATGTGCAGGCACGGCGCAGGGGGAGACCAGCAAGCAGAGGGCGCGGCGCAGGCGGCGGAGG
TCGGC-GAGGTGCGGCGGAGGGGATGGTGACGGTGGGGAGATGGACGGAGGAGGGGACCCCAAGAAGCGGCGGCTCACCGACGAGCAGGCCGAGA
TTCTGGAGCTGAGCTTCCGGGAGGACCGCAAGCTGGAGACAGCCCGCAAGGTGTATCTGGCCGCCGAGCTCGGGCTGGACCCCAAGCAGGTCGCC
GTGTGGTTCCAGAACCGCCGCGCGCGCCACAAGAACAAGACGCTCGAGGAGGAGTTCGCGAGGCTCAAGCACGCCCACGACGCCGCCATCCTCCA
CAAATGCCACCTCGAGAACGAGGTATGCTTGCTCGCATACACTCACACTGGCTTACATATGGCGCTGCACATCTGCAGTTCCTCTCCGTTCTTGA
ACATGCTTACTGACAAACATATGGCCAGCTGCTGAGGCTGAAGGAGAGACTGGGAGCGACTGAGCAGGAGGTGCGGCGCCTCAGGTCGGCAGCTG
GGAGCCACGGGGCATCTGTGGATGGCGGACACGCCGCTGGCGCCGTTGGCGTGTGCGGCGGGGGCCCGAGCTCGTCCTTCTCGACGGGAACCTGC
CAGCAGCAGCCGGGTTTCAGCGGGGCAGACGTGCTGGGGCGGGACGATGACCTGATGATGTGCGTCCCCGAGTGGTTTTTAGCATGAATTAGAGT
TTATGCTGGCTAAGCCGATAGCAGCGTGGTCGAGTGTTTTTTAGCATGAAAT

TATA box, two introns, a 5’ untranslated region and a 3’ untranslated regions are underlined.

Translated regions are in yellow

Start and Stop codons are in red

Sanger DNA Sequencing (old but still relevant)

1. Start with a defined fragment of DNA

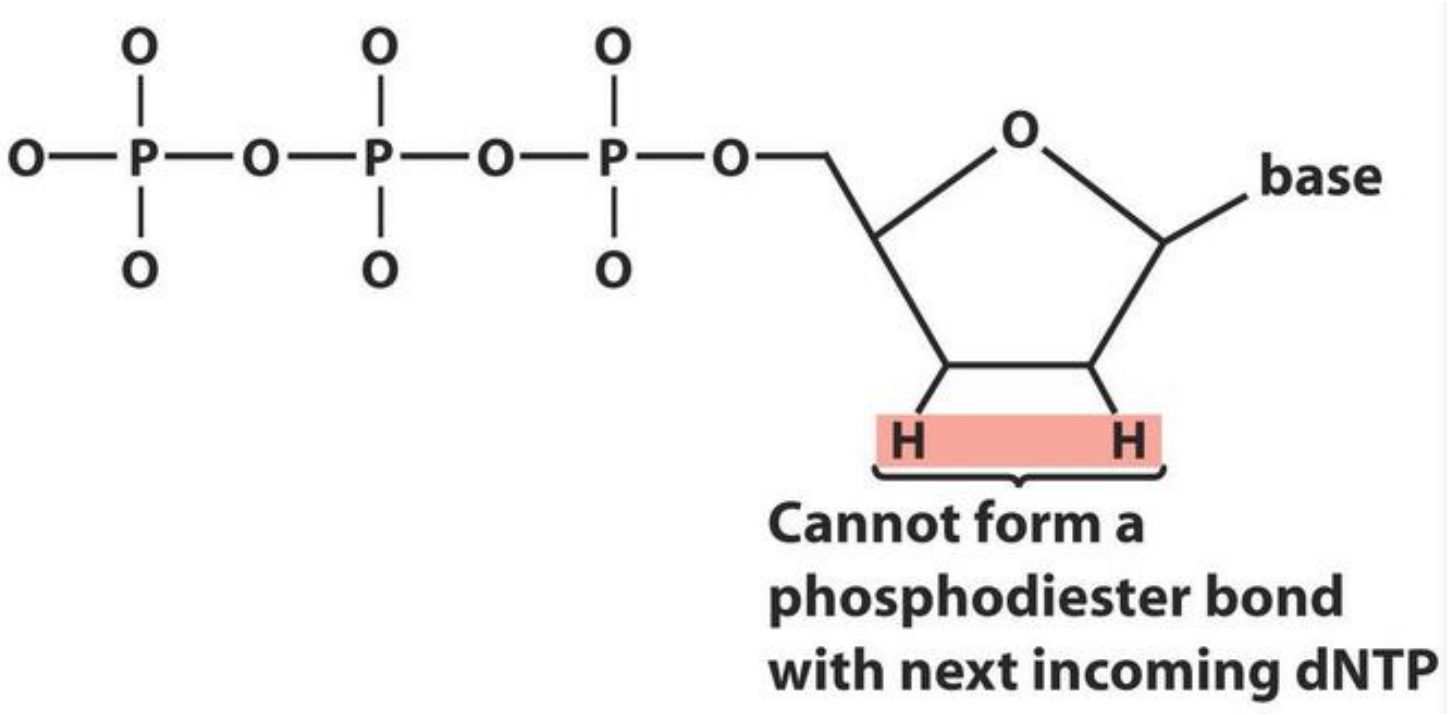
2. Based on this template, generate a population of molecules differing in size by one base of known composition

3. Fractionate the population molecules based on size

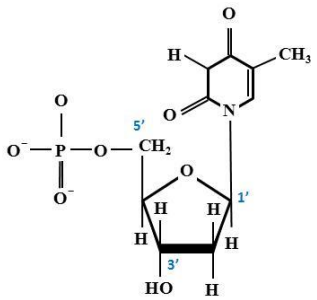
4. The base at the truncated end of each of the fractionated molecules is determined and used to establish the nucleotide sequence

Sanger Sequencing - ddNTPs

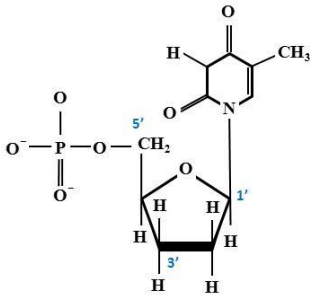
A dideoxy nucleotide lacks a 3' OH and once incorporated, it will terminate strand synthesis. L-1. No free 3' OH



Decoding DNA – Sanger Sequencing

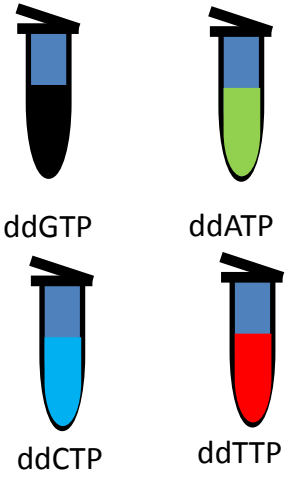


deoxinucleotyde (dNTP)



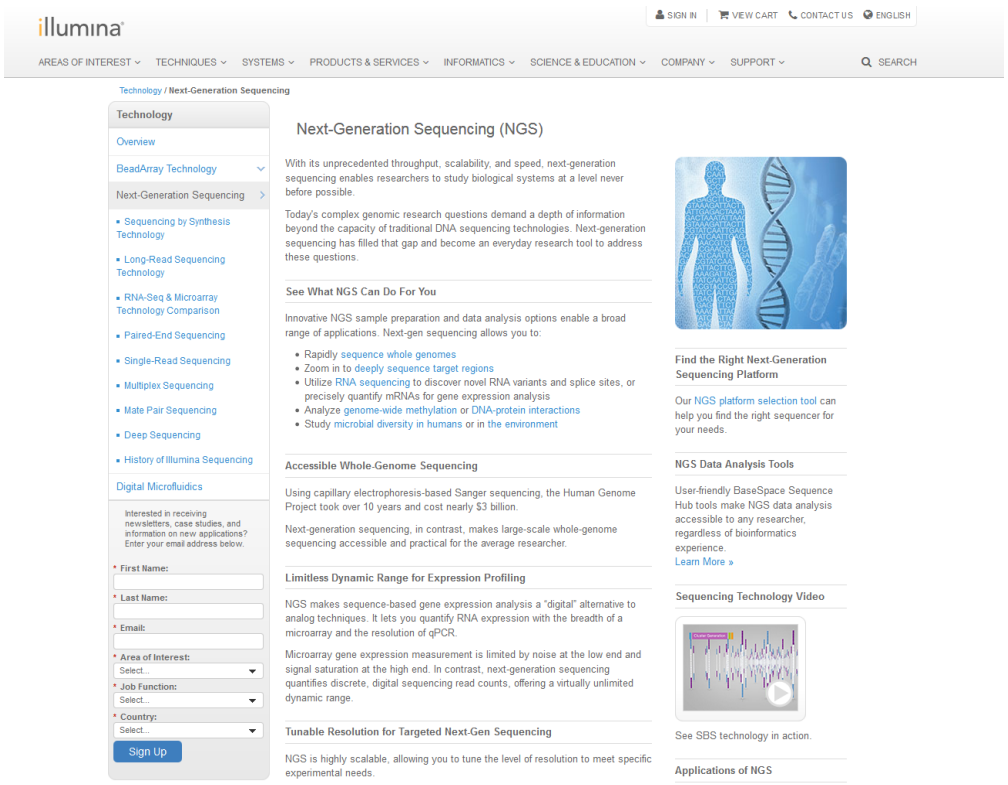
dideoxinucleotyde (ddNTP)

- Buffer
- DNA polymerase
- dNTPs
- Labeled primer
- Target DNA



https://www.dnalc.org/view/15923-Cycle-sequencing.html

Next Generation Sequencing - Illumina

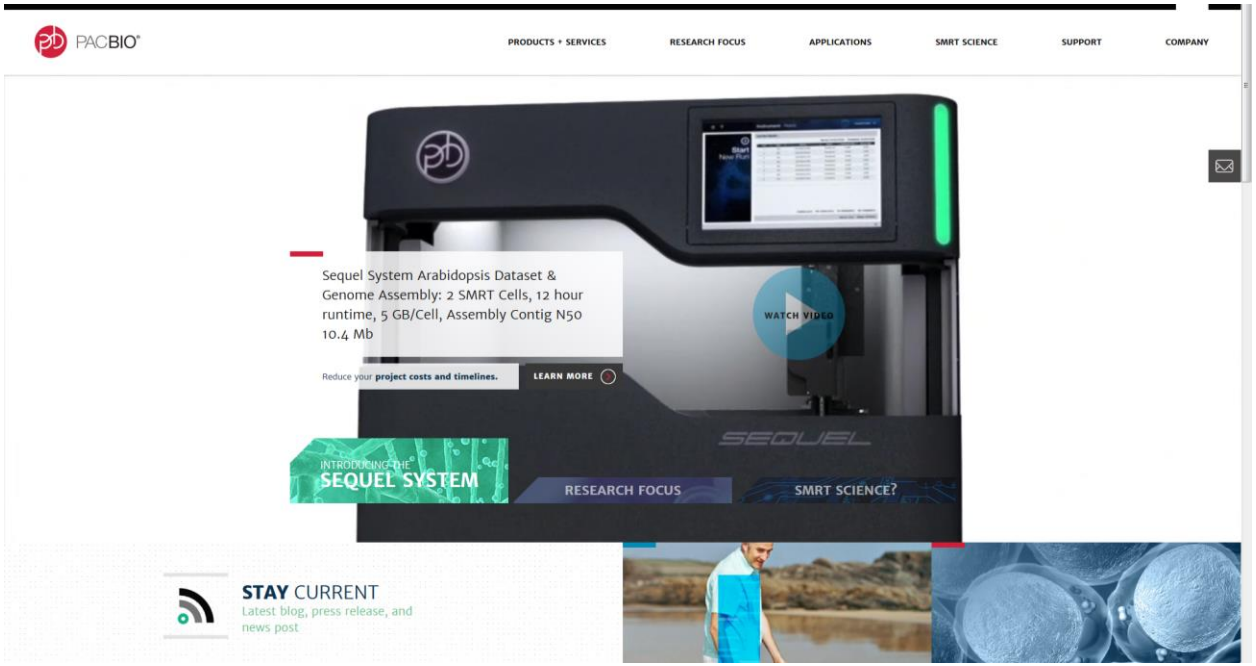


https://www.illumina.com/technology/next-generation-sequencing.html

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

PAC Bio

Single Molecule Real Time Sequencing



https://www.youtube.com/watch?v=v8p4ph2MAvI

Sequencing considerations

Method	Read length	Accuracy	Reads per run	Time per run	Cost per 1 million bases (in US$)	Advantages	Disadvantages
Single-molecule real-time sequencing (**Pacific Bio**)	5,500 bp to 8,500 bp avg (10,000 bp); maximum read length >30,000 bases	99.999% consensus accuracy; 87% single-read accuracy	50,000 per SMRT cell, or ~400 megabases	30 minutes to 2 hours	$0.33–$1.00	Longest read length. Fast.	Moderate throughput. Equipment can be very expensive.
Sequencing by synthesis (**Illumina**)	50 to 300 bp	98%	up to 3 billion	1 to 10 days, depending upon sequencer and specified read length	$0.05 to $0.15	Potential for high sequence yield, depending upon sequencer model and desired application.	Equipment can be very expensive. Requires high concentrations of DNA.
Chain termination (**Sanger** sequencing)	400 to 900 bp	99.9%	N/A	20 minutes to 3 hours	$2400	Long individual reads.	More expensive and impractical for larger sequencing projects.

Genome Size and whole genome sequencing

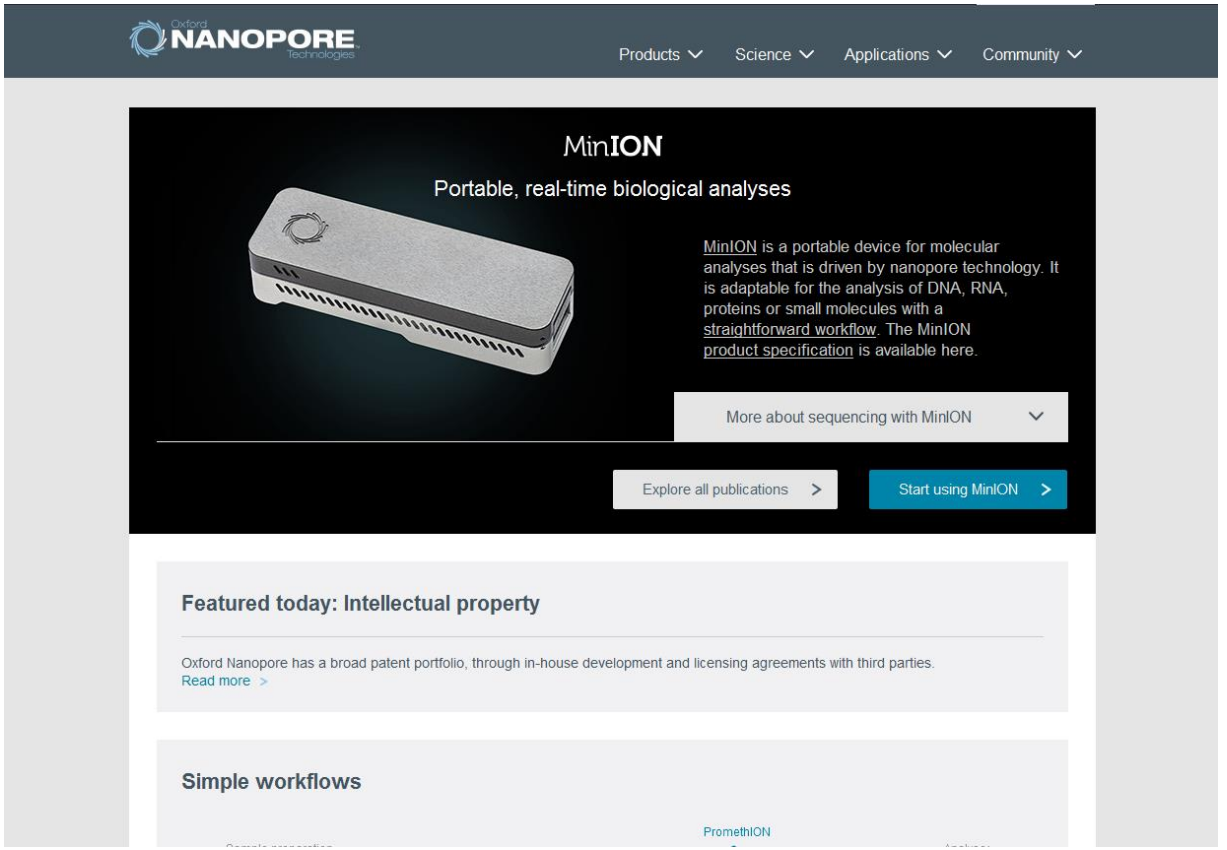
- *Arabidopsis thaliana* has the smallest genome known in the plant kingdom (135 Mb) and for this reason has become a favorite of plant molecular biologists
- *Psilotum nudum* (the "whisk fern") is a far simpler plant than Arabidopsis (it has no true leaves, flowers, or fruit) and has a genome size is 2.5 x 10^{11} Mb
- Dealing with the C value paradox and whole genome sequencing…technology, time, and $

Sequencing Developments

Technologies for whole genome sequencing are evolving very rapidly and too fast for us to compare and contrast in this class

- Key considerations
  - Cost
  - Speed
  - Read length
  - Assembly

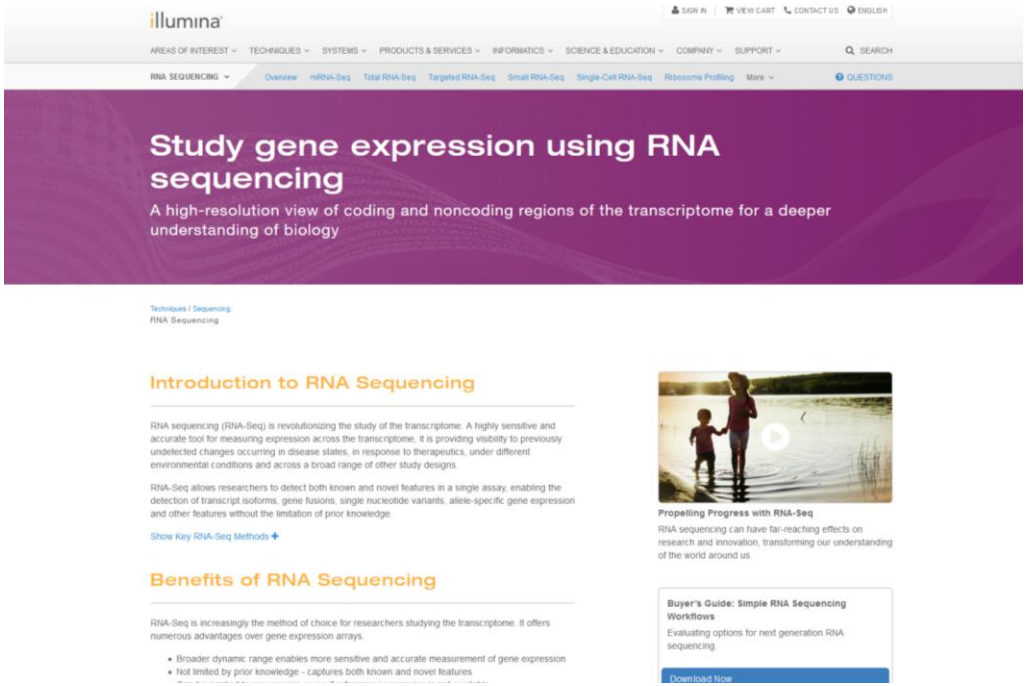
https://nanoporetech.com/



RNA seq
Target the transcriptome rather than the genome

http://rnaseq.uoregon.edu/

https://www.illumina.com/techniques/sequencing/rna-sequencing.html



OSU Sequencing Resources

OSU Oregon State University

Search this site

Center for Genome Research and Biocomputing

Core Facilities

Home | About | People | Affiliated Faculty | Core Facilities | Bioinformatics | Biocomputing | Training | Events | Industry | Contact Us

Home » Core Facilities

Sequencing

Sanger Sequencing

Illumina HiSeq 3000

Illumina MiSeq

Core Facilities

- Calendar
- Web Order
- Sequencing
  - Sanger Sequencing
  - Illumina HiSeq 3000
  - Illumina MiSeq
- Genotyping
- DNA Size Selection
- DNA/RNA Extraction Service
- Quantitation & Analysis
- Microscopy & Imaging
- Protein Simple Western Service
- Custom Services
- Other Resources
- Service Fees

Quick Links

Web Order

Calendar

Sequencing a plant genome

nature genetics

The genome of woodland strawberry (*Fragaria vesca*)

Vladimir Shulaev[1*], Daniel J Sargent[2], Ross N Crowhurst[3], Todd C Mockler[4,5], Otto Folkerts[6], Arthur L Delcher[7], Pankaj Jaiswal[4], Keithanne Mockaitis[8], Aaron Liston[4], Shrinivasrao P Mane[9], Paul Burns[10], Thomas M Davis[11], Janet P Slovin[12], Nahla Bassil[13], Roger P Hellens[3], Clive Evans[9], Tim Harkins[14], Chinnappa Kodira[14], Brian Desany[14], Oswald R Crasta[6], Roderick V Jensen[15], Andrew C Allan[3,16], Todd P Michael[17], Joao Carlos Setubal[9,18], Jean-Marc Celton[19], D Jasper G Rees[19], Kelly P Williams[9], Sarah H Holt[20,21], Juan Jairo Ruiz Rojas[20], Mithu Chatterjee[22,23], Bo Liu[11], Herman Silva[24], Lee Meisel[25], Avital Adato[26], Sergei A Filichkin[4,5], Michela Troggio[27], Roberto Viola[27], Tia-Lynn Ashman[28], Hao Wang[29], Palitha Dharmawardhana[4], Justin Elser[4], Rajani Raja[4], Henry D Priest[4,5], Douglas W Bryant Jr[4,5], Samuel E Fox[4,5], Scott A Givan[4,5], Larry J Wilhelm[4,5], Sushma Naithani[30], Alan Christoffels[31], David Y Salama[22], Jade Carter[8], Elena Lopez Girona[2], Anna Zdepski[17], Wenqin Wang[17], Randall A Kerstetter[17], Wilfried Schwab[32], Schuyler S Korban[33], Jahn Davik[34], Amparo Monfort[35,36], Beatrice Denoyes-Rothan[37], Pere Arus[35,36], Ron Mittler[1], Barry Flinn[21], Asaph Aharoni[25], Jeffrey L Bennetzen[29], Steven L Salzberg[7], Allan W Dickerman[9], Riccardo Velasco[27], Mark Borodovsky[10,38], Richard E Veilleux[20] & Kevin M Folta[22,23]

OSU in the lead with plant genome sequencing
And the Beaver too

nature genetics

The genome of woodland strawberry (*Fragaria vesca*)

Vladimir Shulaev[1*], Daniel J Sargent[2], Ross N Crowhurst[3], Todd C Mockler[4,5], Otto Folkerts[6], Arthur L Delcher[7], Pankaj Jaiswal[4], Keithanne Mockaitis[8], Aaron Liston[4], Shrinivasrao P Mane[9], Paul Burns[10], Thomas M Davis[11], Janet P Slovin[12], Nahla Bassil[13], Roger P Hellens[3], Clive Evans[9], Tim Harkins[14], Chinnappa Kodira[14], Brian Desany[14], Oswald R Crasta[6], Roderick V Jensen[15], Andrew C Allan[3,16], Todd P Michael[17], Joao Carlos Setubal[9,18], Jean-Marc Celton[19], D Jasper G Rees[19], Kelly P Williams[9], Sarah H Holt[20,21], Juan Jairo Ruiz Rojas[20], Mithu Chatterjee[22,23], Bo Liu[11], Herman Silva[24], Lee Meisel[25], Avital Adato[26], Sergei A Filichkin[4,5], Michela Troggio[27], Roberto Viola[27], Tia-Lynn Ashman[28], Hao Wang[29], Palitha Dharmawardhana[4], Justin Elser[4], Rajani Raja[4], Henry D Priest[4,5], Douglas W Bryant Jr[4,5], Samuel E Fox[4,5], Scott A Givan[4,5], Larry J Wilhelm[4,5], Sushma Naithani[30], Alan Christoffels[31], David Y Salama[22], Jade Carter[8], Elena Lopez Girona[2], Anna Zdepski[17], Wenqin Wang[17], Randall A Kerstetter[17], Wilfried Schwab[32], Schuyler S Korban[33], Jahn Davik[34], Amparo Monfort[35,36], Beatrice Denoyes-Rothan[37], Pere Arus[35,36], Ron Mittler[1], Barry Flinn[21], Asaph Aharoni[25], Jeffrey L Bennetzen[29], Steven L Salzberg[7], Allan W Dickerman[9], Riccardo Velasco[27], Mark Borodovsky[10,38], Richard E Veilleux[20] & Kevin M Folta[22,23]

The woodland strawberry, *Fragaria vesca* ($2n = 2x = 14$), is a versatile experimental plant system. This diminutive herbaceous perennial has a small genome (240 Mb), is amenable to genetic transformation and shares substantial sequence identity with the cultivated strawberry (*Fragaria* × *ananassa*) and other economically important rosaceous plants. Here we report the draft *F. vesca* genome, which was sequenced to ×39 coverage using second-generation technology, assembled *de novo* and then anchored to the genetic linkage map into seven pseudochromosomes. This diploid strawberry sequence lacks the large genome duplications seen in other rosids. Gene prediction modeling identified 34,809 genes, with most being supported by transcriptome mapping. Genes critical to valuable horticultural traits including flavor, nutritional value and flowering time were identified. Macrosyntenic relationships between *Fragaria* and *Prunus* predict a hypothetical ancestral Rosaceae genome that had nine chromosomes. New phylogenetic analysis of 154 protein-coding genes suggests that assignment of *Populus* to Malvidae, rather than Fabidae, is warranted.

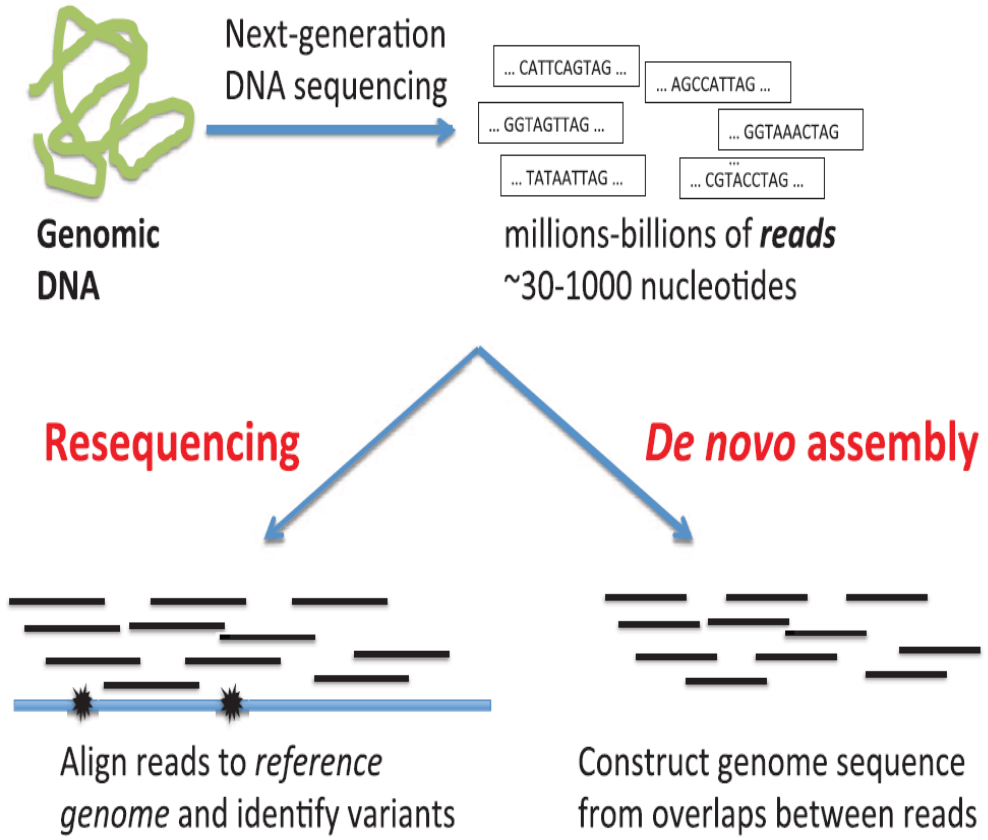
Fragaria vesca

Herbaceous, perennial
2n=2x=14 240 Mb

- Reference species for Rosaceae
- Genetic resources
- Forward and reverse genetics

Fragaria x ananassa: 2n=8x=56. The youngest crop? 250 years.

Why sequence the "weed" when you could sequence the crop???



Short reads
No physical reference
De novo assembly
Open source

Would they follow the same path in 2017?

Benjamin J. Raphael*

Department of Computer Science and Center for Computational Molecular Biology, Brown University, Providence, Rhode Island, United States of America

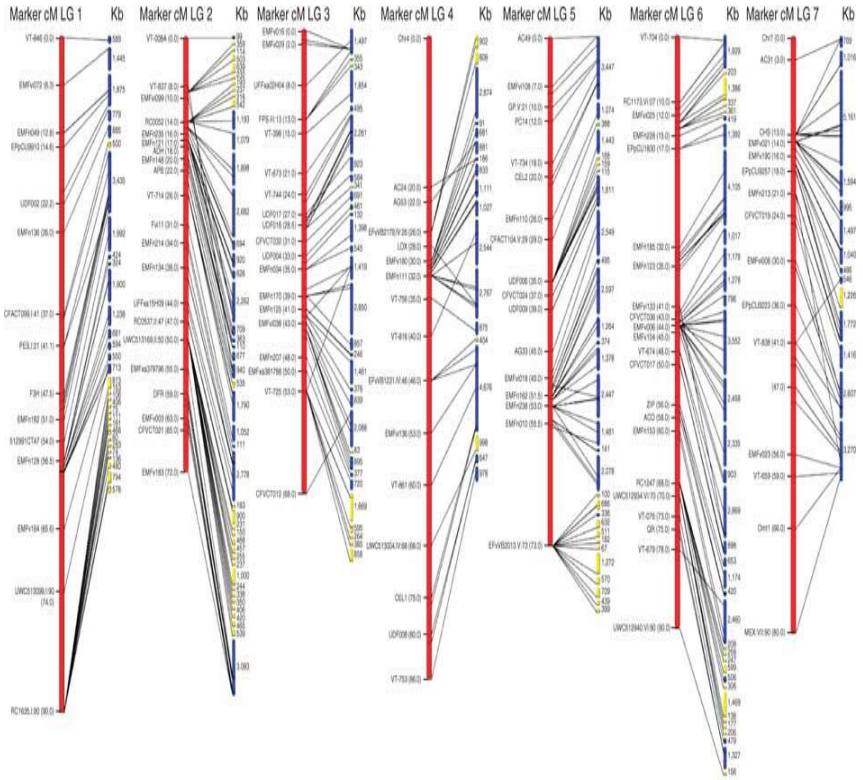
Genome Sequencing and Assembly

- 3 next-gen platforms (new in the 20-10's – old school in the 20-teens)
- X39 coverage (number of reads including a given nucleotide)
- Contigs (overlapping reads) assembled into scaffolds (contigs + gaps)

Genome Sequencing and Assembly

- ~ 3,200 scaffolds N50 of 1.3 Mb (weighted average length)
- Over 95% (209.8 Mb) of total sequence is represented in **272** scaffolds
- Flow cytometry to measure genome size: ~240 Mb ***Why is this > than the number of Mb sequenced (209)?***

Anchoring the genome sequence to the genetic map



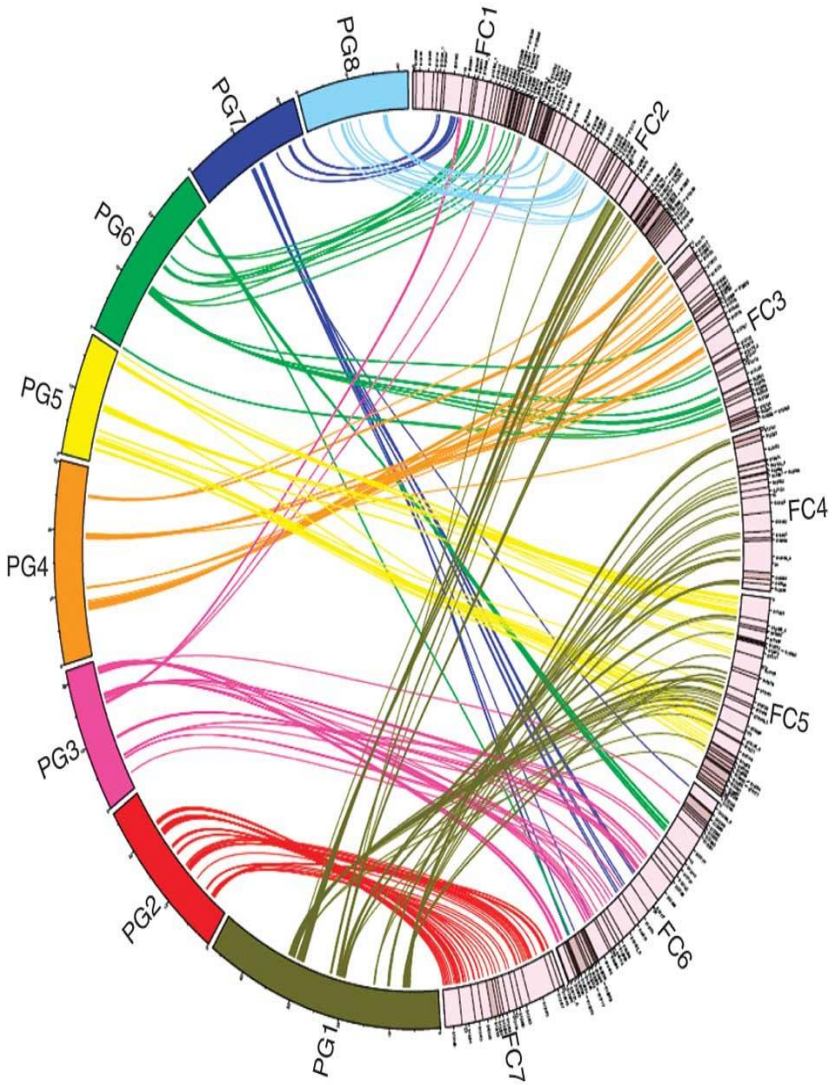
94% of scaffolds anchored to the diploid *Fragaria* reference linkage map using 390 genetic markers

Pseudochromosomes ~ linkage groups ~ karyotype

Of what use is a linkage map when you can have a whole genome sequence?

Synteny

Prunus and *F. vesca*



Homologs

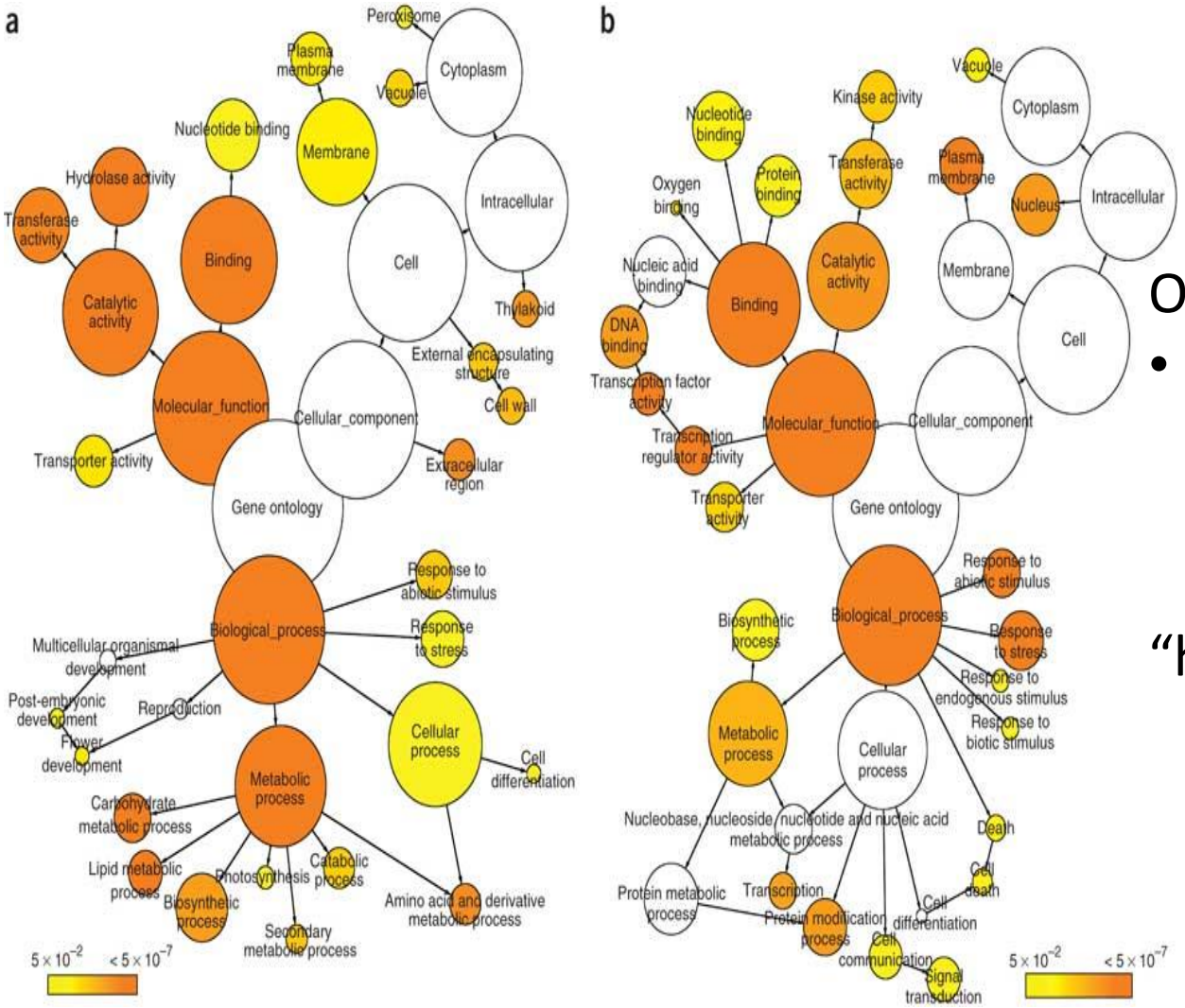
- Orthologs
- Paralogs

Nice picture,
but of what use?

The small genome size (240 Mb)

- Absence of large genome duplications
- Limited numbers of transposable elements, compared to other angiosperms - ***the driver of small genome size?***

Transcriptome sequence (cDNAs)



Organ specificity

- Fruits and roots – different types of genes

“host gene deserts”

Figure 3 Gene ontology mapping and functional annotation of strawberry genes. Overrepresented gene ontology categories in fruit (**a**) and root (**b**) expressed genes. The circles are shaded based on significance level (yellow, false discovery rate < 0.05), and the radius of each circle denotes the number of genes in each category.

Genetics accounting

Gene prediction

- 34,809 nuclear genes
  - flavor, nutritional value and flowering time
  - 1,616 transcription factors

RNA genes

- 569 tRNA, 177 rRNA, 111 spliceosomal RNAs, 168 small nuclear RNAs,

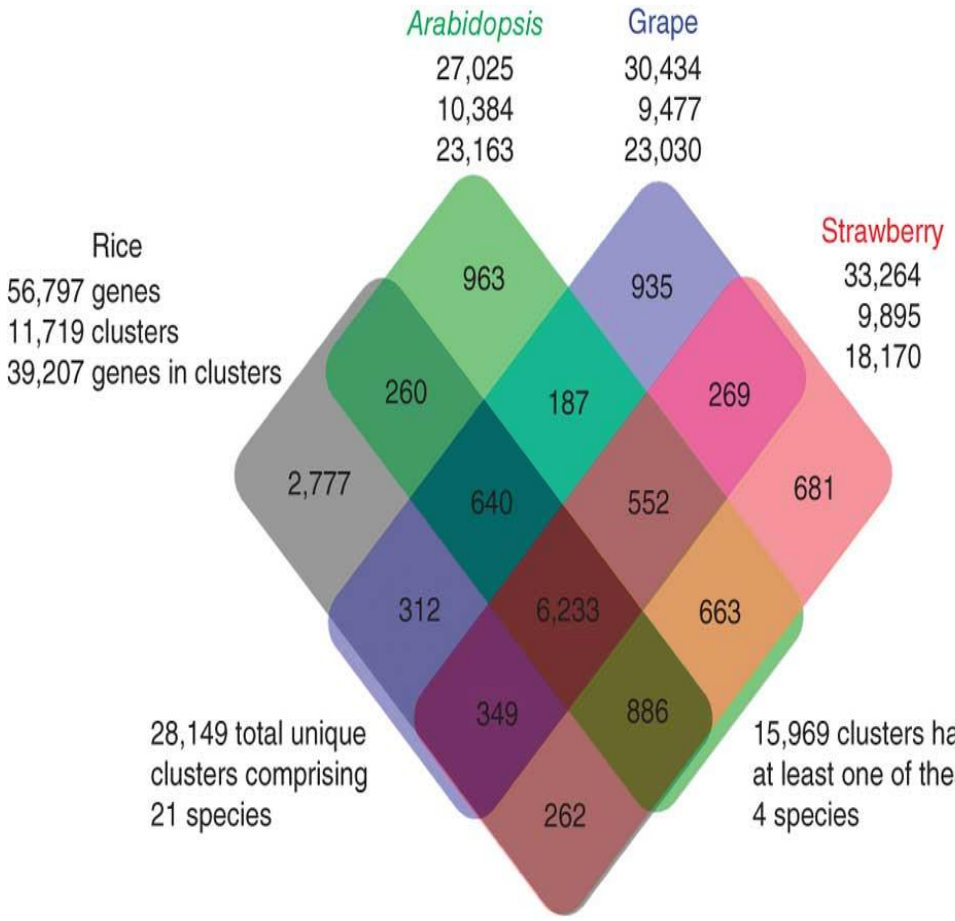
76 micro RNA and 24 other RNAs

Are these results expected?

Chloroplast genome

- 155,691 bp encodes 78 proteins, 30 tRNAs and 4 rRNA genes
- Evidence of DNA transfer from plastid genome to the nuclear genome

Strawberry unique gene clusters



Utility in practical horticulture?

Acknowledgements: Thanks to Merve Sekerli for the slide preparation