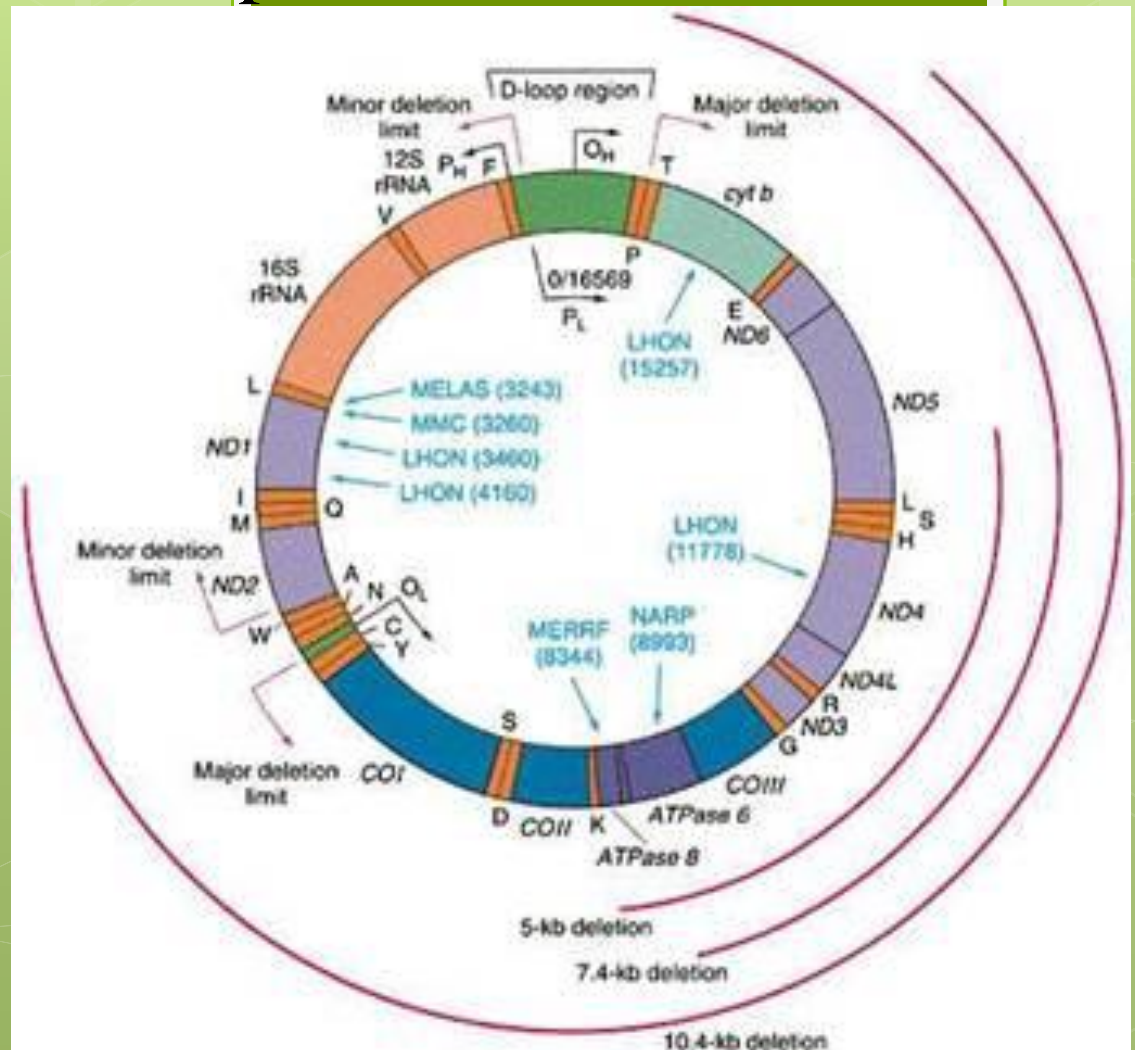
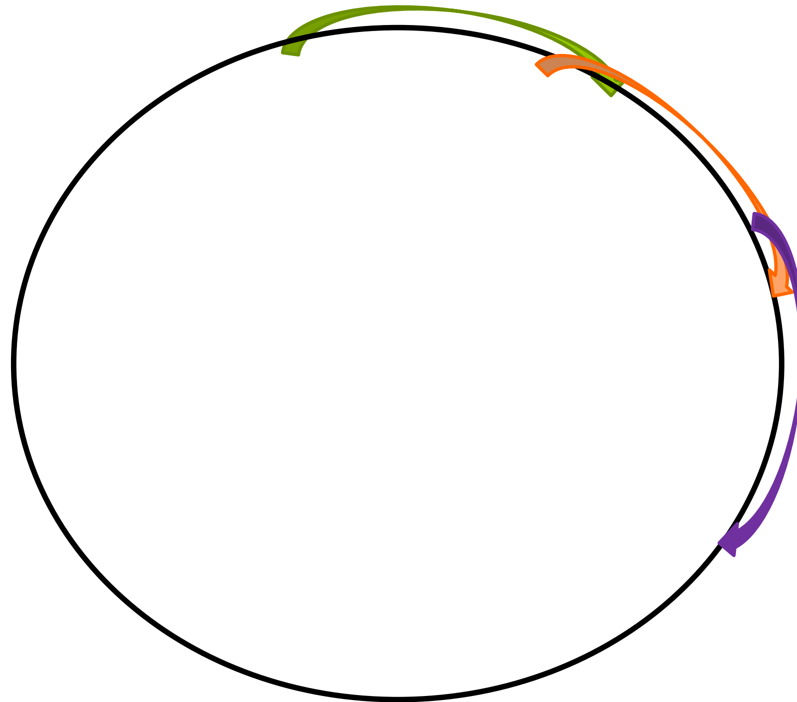
B3 Summer Science Camp 2016



The overlapping amplicon approach to CP genome recovery and sequencing



- The Chloroplast genome is 160,000bp long. The longest PCR we can carry out very well is ~20,000bp.
- The overlaps are to let me arrange the pieces correctly, since I cannot ‘number’ them.

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide Search
Advanced Help

NCBI is phasing out sequence GI numbers in September 2016. Please use accession.version! Read more...

GenBank Send:

Castanea mollissima chloroplast, complete genome

NCBI Reference Sequence: NC_014674.1
FASTA Graphics

Go to:

```
LOCUS       NC_014674            160799 bp    DNA     circular PLN 06-JAN-2011
DEFINITION  Castanea mollissima chloroplast, complete genome.
ACCESSION   NC_014674
VERSION     NC_014674.1  GI:313183972
DBLINK      Project: 60297
            BioProject: PRJNA60297
KEYWORDS    RefSeq.
SOURCE      chloroplast Castanea mollissima (Chinese chestnut)
  ORGANISM  Castanea mollissima
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; Gunneridae;
            Pentapetalae; rosids; fabids; Fagales; Fagaceae; Castanea.
REFERENCE   1  (bases 1 to 160799)
  AUTHORS   Jansen,R.K., Saski,C., Lee,S.B., Hansen,A.K. and Daniell,H.
  TITLE     Complete plastid genome sequences of three Rosids (Castanea,
            Prunus, Theobroma): evidence for at least two independent transfers
            of rpl22 to the nucleus
  JOURNAL   Mol. Biol. Evol. 28 (1), 835-847 (2011)
   PUBMED   20935065
REFERENCE   2  (bases 1 to 160799)
  CONSRTM   NCBI Genome Project
  TITLE     Direct Submission
  JOURNAL   Submitted (18-NOV-2010) National Center for Biotechnology
            Information, NIH, Bethesda, MD 20894, USA
REFERENCE   3  (bases 1 to 160799)
  AUTHORS   Jansen,R.K., Saski,C., Lee,S.-B., Hansen,A.K. and Daniell,H.
  TITLE     Direct Submission
  JOURNAL   Submitted (28-SEP-2010) Integrative Biology, University of Texas at
            Austin, 1 University Station C0930, Austin, TX 78712, USA
COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
            NCBI review. The reference sequence is identical to HQ336406.
```

Change region shown

Customize view

Analyze this sequence
- Run BLAST
- Pick Primers
- Highlight Sequence Features
- Find in this Sequence

Related information
- BioProject
- Full text in PMC
- Gene
- Genome
- Identical GenBank Sequence
- Protein
- PubMed
- PubMed (Weighted)
- Taxonomy

LinkOut to external resources
- Dryad Digital Repository [Dryad Digital Repository]
- Dryad Digital Repository [Dryad Digital Repository]
- Dryad Digital Repository [Dryad Digital Repository]

```
159481 tgcatacatt tgtctcatac actatgggga tggtgagaaa agatatattt tacatcccag
159541 aggggctata attggagata ccattgtttc tggtacagaa gttcctataa aaatgggaaa
159601 tgccctacct ttgagtgcgg tttgaactat tgatttacgt aattggaagt aaccaattag
159661 gtttacaacg aaacctagaa atcgatcacc gatccaattt gagtacctct acaggataga
159721 cctcaacaga aaactgaaga gtaacggcag caagtgattg agttcagtag ttcctcatat
159781 aaaattattg actctagaga tatagtaata tggagaagac aaaattgttt caagcaccga
159841 caaaaccaga agcgcccctt gtttcaaaga gaggaggacg ggttattcac atttcatttg
159901 atggtcagag gcaaattgaa agctaagaag tggtaattct aaggattccc ccgggggaaa
159961 atagagatgt ctcctacgtt acccataata tgtggaagta tcgacgtaat ttcatagagt
160021 cattcggtct gaatgctaca tgaagaacat aagccagatg aaggaacggg aagacctagg
160081 atgtagaaga tcataacatt actgattcgg cagatttgga ttcctatata tcactcatgt
160141 ggtacttcat tatacgatat atataagaat tatacgatat atataagatc catctgtata
160201 gatatcatca tctacatcca gaaagccgta tgctttggaa gaagcttgta cagtttggga
160261 agaggttttg attgatcaaa aagaagaatc tacttcaacc gatatgccct taggcacggc
160321 catacataac atagaaatca cacttggaaa gggtggacaa ttagctagag cagcaggtgc
160381 tgtagcgaaa ctgattgcaa aagaggggaa atcggccaca ttaaaattac cttctgggga
160441 ggtccgtttg atatccaaaa actgctcagc aacagtcgga caagtgggta atgttggggt
160501 gaaccagaaa agtttgggta gagccggatc taaatgttgg ctaggtaagc gtcctgtagt
160561 aagaggagta gttatgaacc ctgtagacca tccccatggg ggtggtgaag ggagggcccc
160621 aattggtaga aaaaaacccg caacccottg gggttatcct gcacttggaa gaagaagtag
160681 aaaaaggaat aaatatagtg ataatttgat tcttcgtcgc cgtagtaaat aggataggag
160741 agaaaataga atttgtttct tcgtctttac aaaaaaaaaa gaaataggag taatgtgac
//
```


GETTING STARTED
NCBI Education
NCBI Help Manual
NCBI Handbook
Training & Tutorials
Submit Data

RESOURCES
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Variation

POPULAR
PubMed
Bookshelf
PubMed Central
PubMed Health
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

FEATURED
Genetic Testing Registry
PubMed Health
GenBank
Reference Sequences
Gene Expression Omnibus
Map Viewer
Human Genome
Mouse Genome
Influenza Virus
Primer-BLAST
Sequence Read Archive

National Center for Biotechnology Information, U.S. National Library of Medicine
8600 Rockville Pike, Bethesda MD, 20894 USA

attatactcc tgactatgaa accaaagata Find No matches found

```
60061 caatatcatc catattattt atcgaattcg aaccggaact ctatttctat ttacgattca
60121 ttttttctat cttattggtg cccttatctc atttcttagt taatatttca gcctatcgat
60181 ttatgcctag cctatttttt ttttacctat accctttagt tttagagttt tagtttcatg
60241 ctttgctttt catggacgaa ctcctcatat tttcacatct aggatttaca tatacaacat
60301 atatcactgt caagagttaa tttcttatta tttagatatt ccaatggggg gcgggggtaa
60361 gagattagaa acttgagaaa acaaggattg ggttgcgcca tacatatgaa agagtataca
60421 ataatggtgt atttggcgaa tcaaatatca tggtctaata acgaaccatt ctaattagtt
60481 gataattttt tgaaaggatt ccttgaaagg tttcattaac tcctaattca tgtcgagtag
60541 accttgttgt tgcgaaaatt cttaattcat gagttgtagg gagggactta tgtcaccaca
60601 aacagagact aaagcaagtg ttggattcaa agctggtgtt aaagattata aattgactta
60661 ttatactcct gactatcaaa ccaaagatac tgatatcttg gcagccttcc gagtaactcc
60721 tcaacctgga gttccgccgg aggaagcagg ggccgcggta gctgctgaat cttccactgg
60781 gacatggaca actgtgtgga ctgacgggct taccagtctt gatcgttaca aaggacgatg
60841 ctaccacatc gagccggttg ctggagaaga aaatcaattt attgcttatg tagcttaccc
60901 cttagacctc tttgaagaag gttctgttac taacatgttt acttccattg tgggtaatgt
60961 atttggattc aaggccctgc gcgctctacg tctggaggat ttgcgaatcc ctacttctta
61021 ttctaaaact ttccaaggtc cgcctcatgg catccaagtt gagagggata aattaaacaa
61081 gtatggccgc cccctattag gatgtactat taaacctaaa ttgggattat ccgctaagaa
61141 ttacggtaga gcagtttatg aatgtctccg cggtgggctt gattttacca aagatgatga
61201 gaacgttaat tcccaaccat ttatgcgttg gagagaccgt ttcctatttt gtgccgaagc
61261 aatttataaa gcgcaggctg aaacaggtga aatcaagggg cattacttga atgctactgc
61321 aggtacatgc gaagaaatga tcaaaagggc tgtatttgcc agagaactgg gagctcctat
61381 cgtaatgcat gactacttaa ccgggggtt cactgcaaat actagcttgg ctcattattg
61441 ccgagataat ggtctacttc ttcacatcca tcgtgcaatg catgcagtta ttgatcgaca
61501 gaagaatcat ggtatacact ttcgtgtact agctaaagca ttacgtatgt ctggtggaga
61561 tcatattcat gccggtaccg tagtaggtaa acttgaaggg gaaagagaaa tcactttagg
61621 ctttgttgat ttactacgtg atgattatat tgaaaaagat cgaagccgcg gtatttattt
61681 cactcaagat tgggtctctt taccaggtgt tctgcccgtg gcttcagggg gtattcacgt
61741 ttggcatatg cctgctctga ccgaaatctt tggagatgat tccgtactac aatttggcgg
61801 aggaactttg ggcacccctt ggggaaatgc acctggtgcc gtagctaatc gagtagctct
61861 agaagcatgt gtacaagctc gtaatgaggg acgtgatctt gctcgtgagg gtaatgaaat
61921 tattcgtgag gctgctaaat ggagtcctga gctagctgct gcttgtgaag tatggaaaga
61981 gatcaaattt gaattcccag caatggatac tttgtaatcc agtaattacc gctcgttctc
62041 ttaattgaat tgtaattaaa ctcggcccaa tcttttacta aaaaaaggat tgagccgaat
62101 acaaaacaaa gatactattg tatctatttt acagatacat acttatctag atatacgtat
62161 acaagatctt aaataaatat ataaaacgaa acaacttaag cgtttctatt gttgtgttag
62221 gtccacaatt aattctatgg atcctaaggg ttggttggta tcttctttta tatcccgtag
62281 tttccgatca tggatcgagt catatatgag agcttcttct acccatcccg tatattgtcc
62341 ctttcgttct gtgttggact agaaccttat tagtagacga gattttacga aaaacgttct
62401 ttctttatat tcataggata gaacaattat ttcttttttc gatttgaaat agctagtgat
62461 ccccccgcgt gtgttgtatc aaattcaaat cgaaaaaaga aataattgaa ataattaaac
62521 aacacctact tgtttttttt gttaatagaa ccctagtgat tgtatttgga tgcgtattag
62581 gatagtaaat gaaatattca aatgattttt tatcgaatga ctattcatct attgtatttt
62641 tcatgtaaat atgtgcaaca aggctttatg gaaaaagggt ggttcaactc gatgttgtcc
62701 aaaaaaaagg agttagaata cgggtatggg ctaagtaaat caatgggcag ccttggttct
62761 attgaaaata ccagtgtaag tgaagacccg attagaaatg ata[illegible] [illegible]
62821 agtgggagcg atagtgacaa ttctagttac agtaatgttg atcatttagt cggcgttaga
```

Graphics

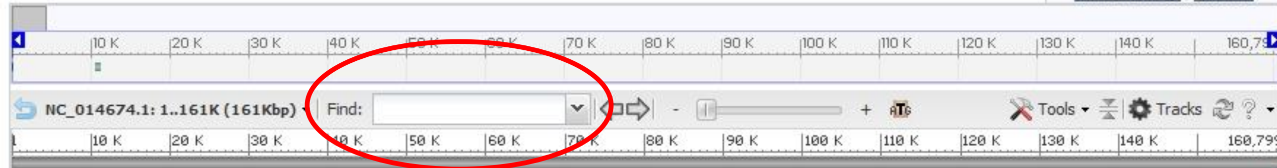
Send:

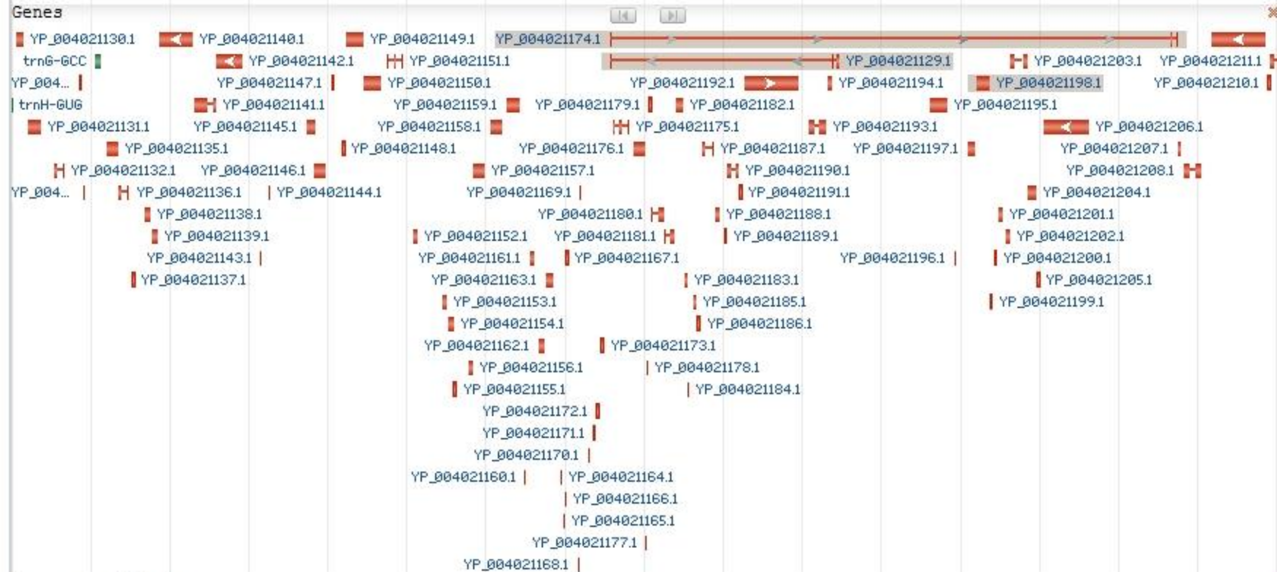
Castanea mollissima chloroplast, complete genome

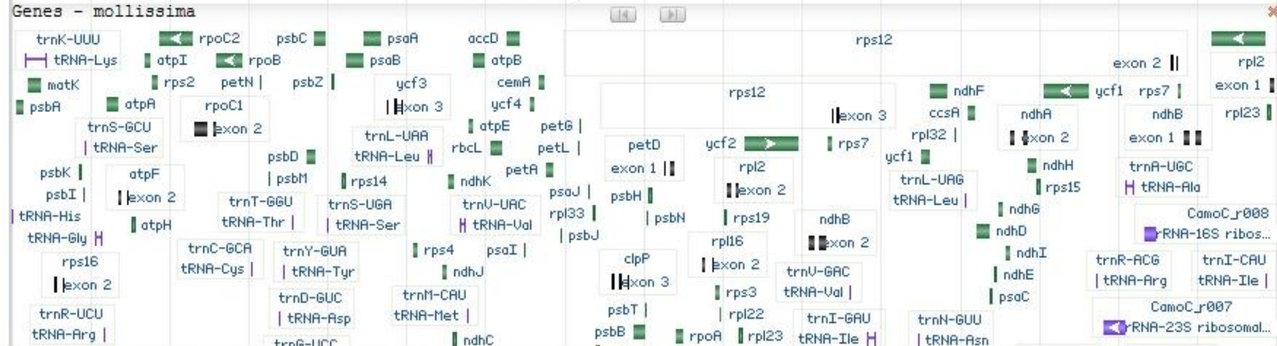
NCBI Reference Sequence: NC_014674.1

GenBank FASTA

Link To This Page | Feedback







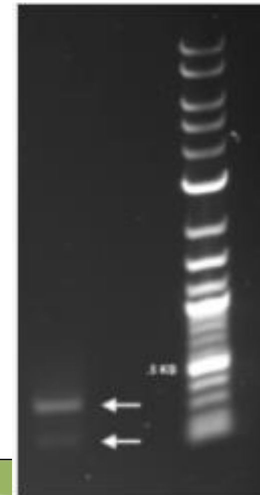
Gene Annotation in Graphics View

PCR Assay Design Constraints

- Specificity: did primers amplify an unintended locus? What if your genome has a SNP where the primer binds?
  - If the length is correct: sequence or restrict the product

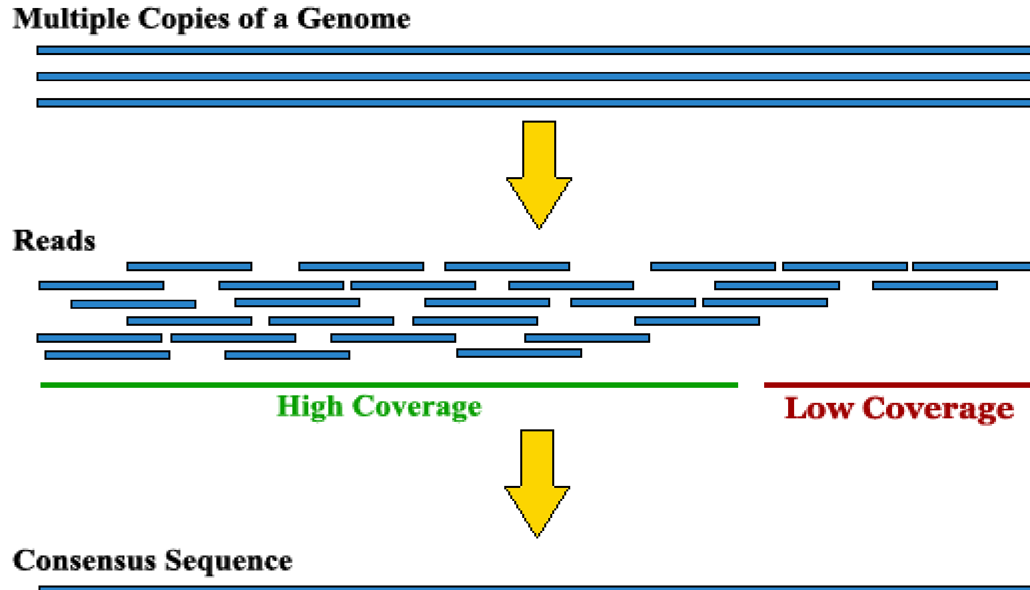
```
        1          11         21         31         41         51         61         71         81         91
          VF2
  1     aaaagtgcca cctgacgtct aagaaaccat tattatcatg acattaacct ataaaaatag gcgtatcacg aggcagaatt tcagataaaa aaaatcctta

101     gctttcgcta aggatgattt ctggaattcg cggccgcttc tagagccagg catcaaataa aacgaaaggc tcagtcgaaa gactgggcct ttcgttttat

201     ctgttgtttg tcggtgaacg ctctctacta gagtcacact ggctcacctt cgggtgggcc tttctgcgtt tatatactag agacctgtag gatcgtacag

301     gtttacgcaa gaaaatggtt tgttatagtc gaataaatac tagtagcggc cgctgcagtc cggcaaaaaa acgggcaagg tgtcaccacc ctgccctttt

401     tctttaaaac cgaaaagatt acttcgcgtt atgcaggctt cctcgctcac tgactcgctg cgctcggtcg ttcggctgcg gcgagcggta tcagctcact

501     caaaggcggt aatacggt
                     VR
```

```
VF2: 5' T     C-C-A-C-C-T-G-A-C-G-T-C-T
          G                             A-A-G-A-A 3'
3'...G-C-A-T-G-T-C-C-A-A-A-T-G-C-G-T-T-C-T-T...5'
```



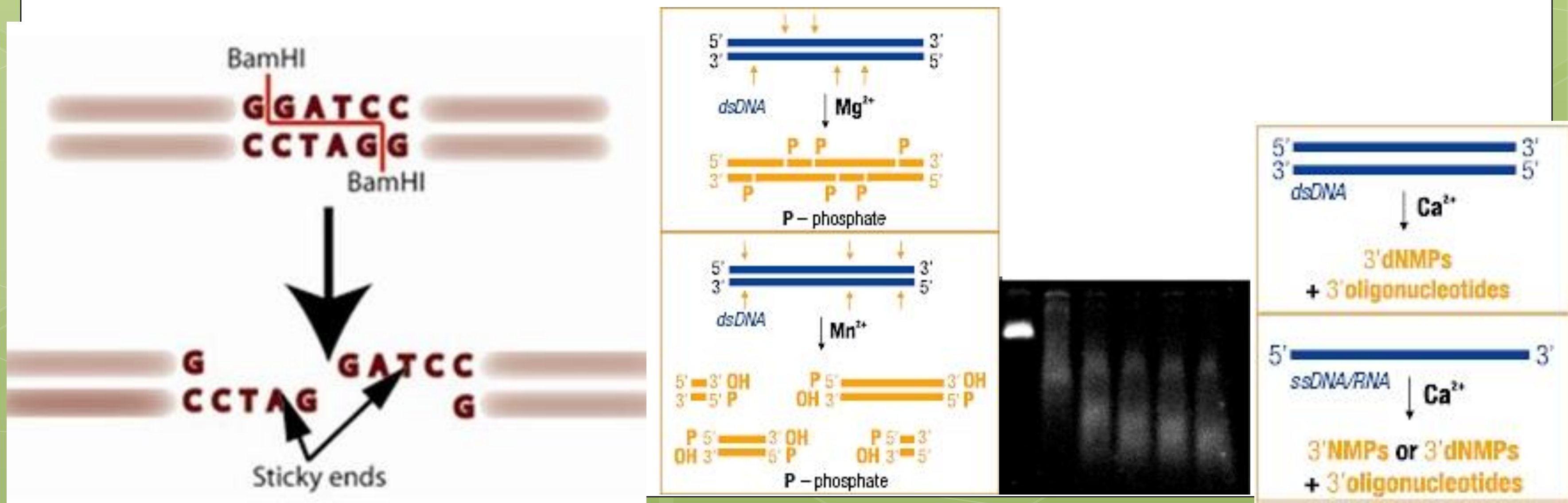
Fragmenting the long amplicons

- The sequencer cannot handle fragments that are longer than ~600bp long. So once I have made the long amplicons I have to break them into much smaller pieces.
- The sequencer requires known ends on every fragment – one is to act as an anchor to hold the fragment in place, and the other lets a primer bind so DNA polymerase can attach and make a copy of the fragment – this is what gives us the sequence.



Enzymatic Fragmentation of Nucleic acids

- Some degree of sequence specificity results.
  - REs: specificity is as precise as the recognition site
  - dsDNA is the required substrate
  - With DNAaseI, MNase, etc there is sequence *preference* but not specificity.



Transposase method (Nextera)

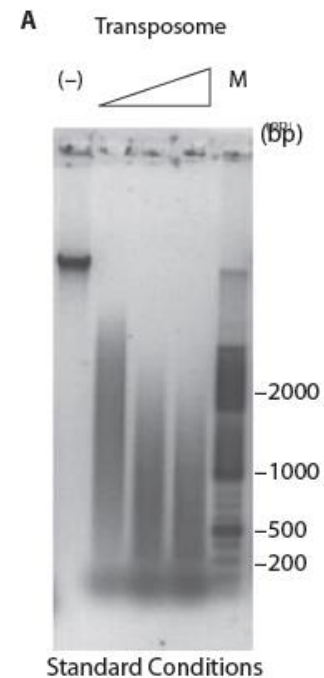
Transposome™ complexes containing free transposon ends

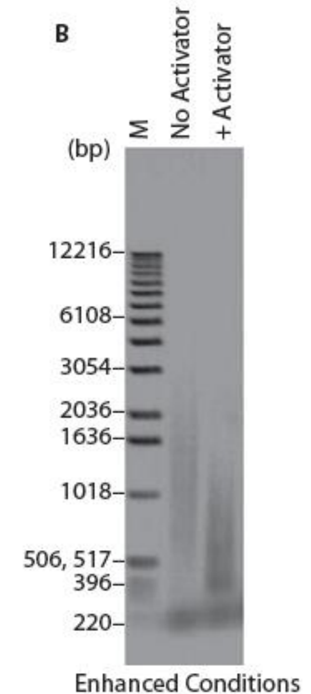
Add target dsDNA

Fragmented and tagged DNA

Add bPCR-compatible sites by PCR (+/- bar codes)

bPCR Input

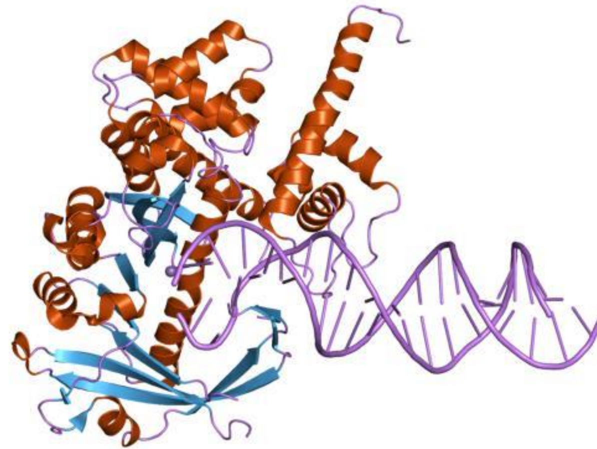




From: EpiBio.con

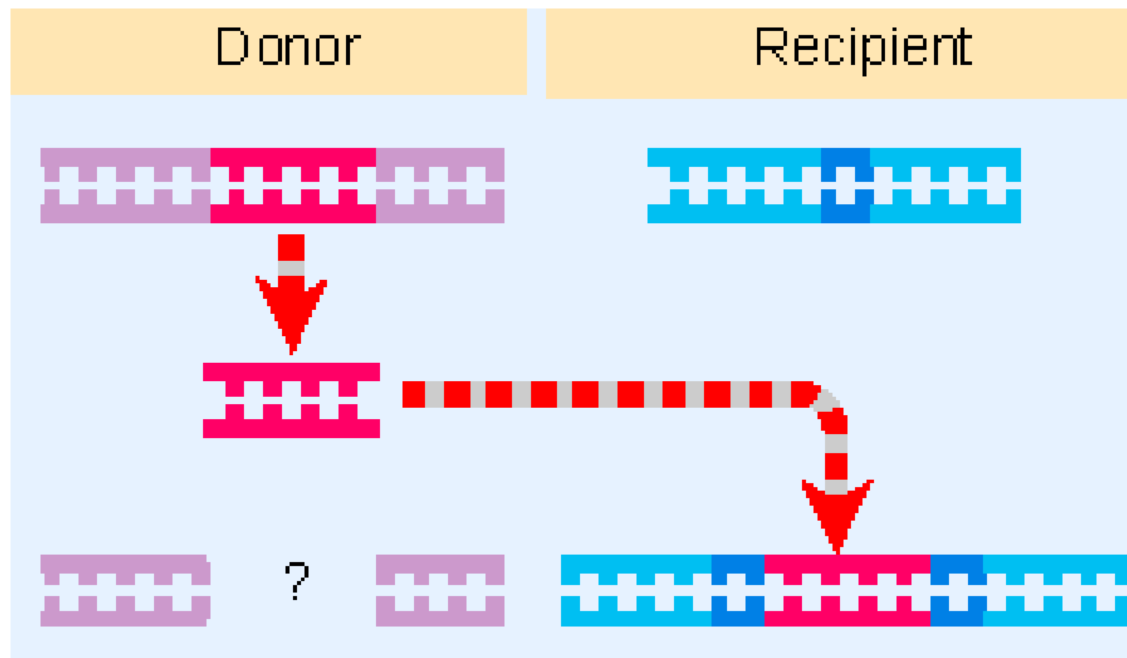
Transposase

- An enzyme that binds to the end of a piece of DNA called a transposon, and helps it insert the DNA into the middle of another piece of DNA.

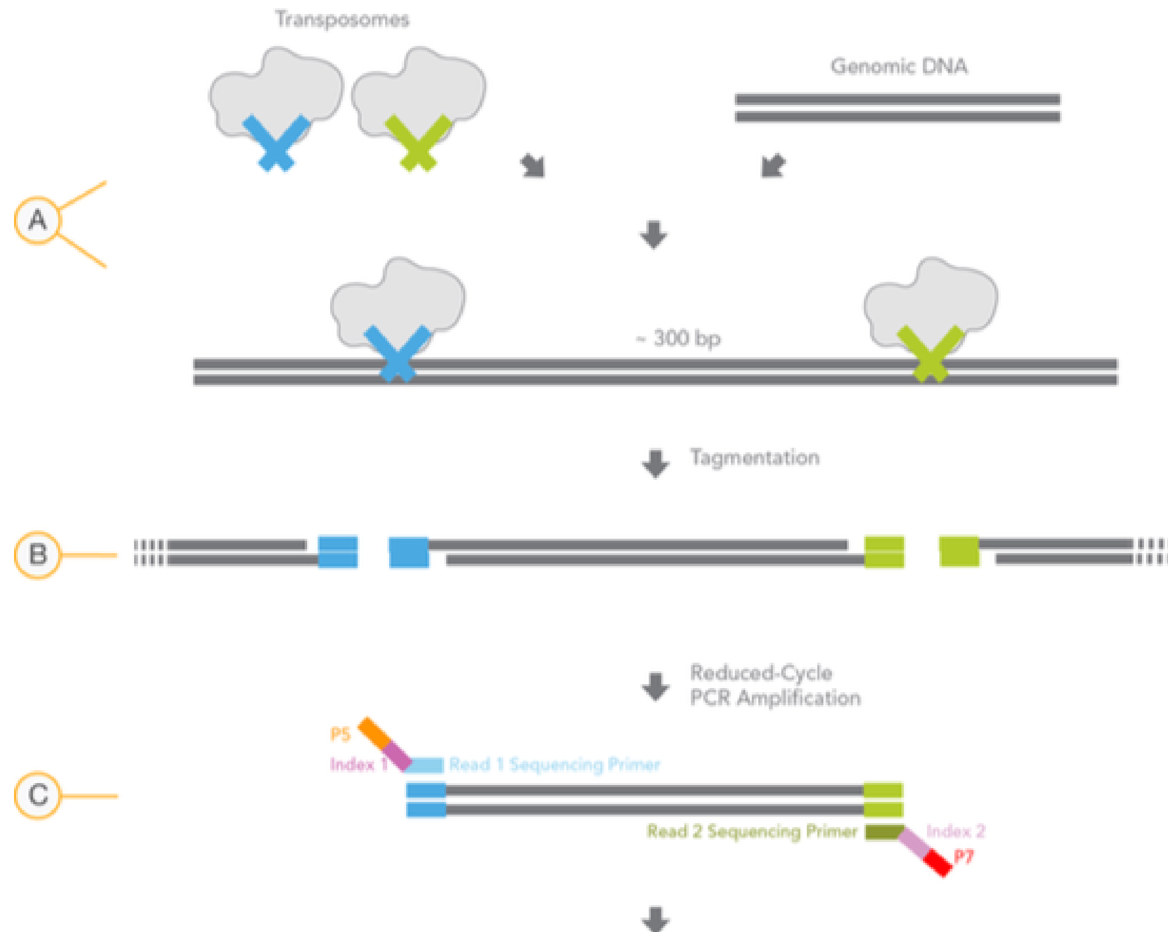


Transposase

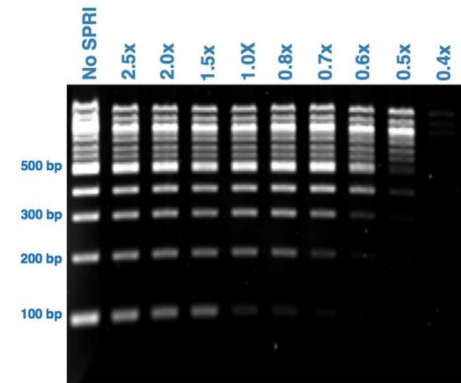
- The transposing element moves as a physical entity directly from one site to another, and is conserved.



Transposase – modified for sequencing



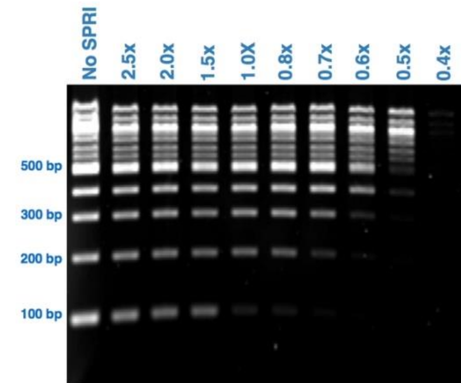
Notes on size selection – the bead lab



- After cutting up the DNA with a restriction enzyme – how do we select just the ones in the size range that works for sequencing?
- Cut a band out of the gel, try to extract the DNA
- Use selective precipitation of DNA onto Ampure XP magnetic beads.
- Beads should be at room temperature and ethanol wash solution should be made fresh.
  - Remove the adaptors and things that are too short: now increase the Beads/PEG in the Supernatant. 1:8 means fragments >100bp are ON THE BEADS – keep these.
- Elute the DNA from the beads and quantify the remainder.

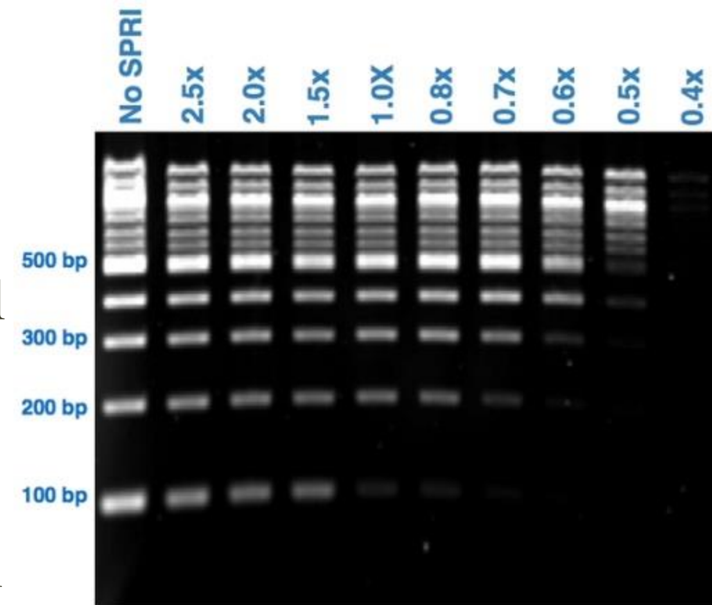
Prep notes

- Solubility is very temperature dependent
  - Beads should be at room temperature
  - Ethanol will absorb water from the air, which changes the concentration over time – make it up fresh each day.
  - Ethanol and water change each other's volume when mixed – when you make the solution measure each separately and then combine them.

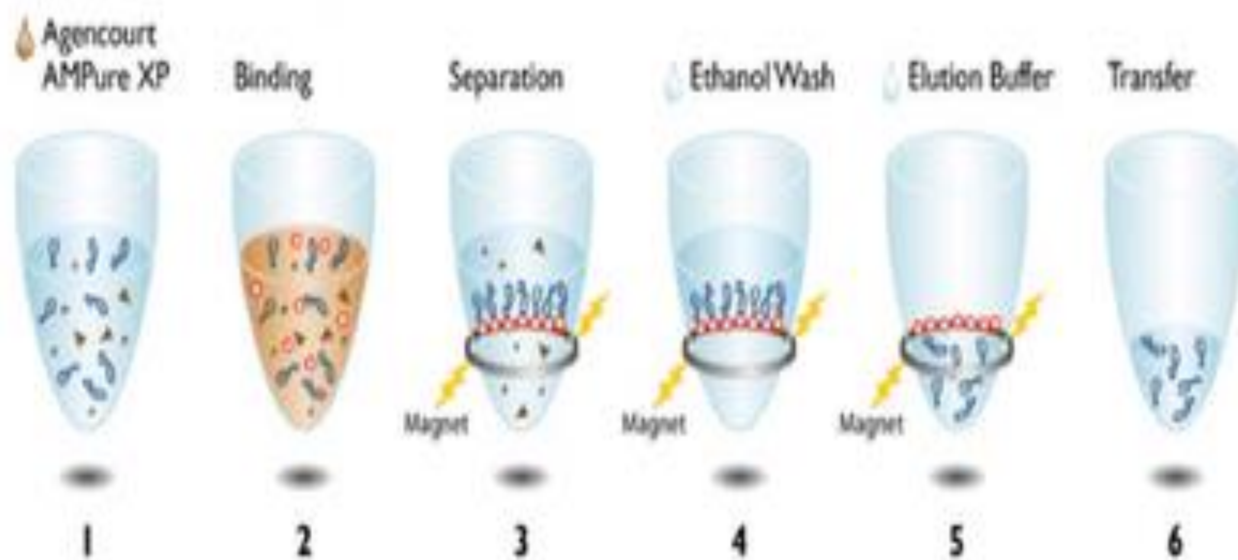


Prep notes

- Polyethylene glycol is used to ‘crowd’ the DNA out of solution – the beads provide a convenient surface for the DNA to land.
  - At a high ratio of PEG, most DNA will be crowded onto the beads – only small things are in the solution.
  - At a low ratio of beads only very big DNA is crowded onto the beads, medium and small DNA fragments will be in solution.
- To get back what is on the beads: remove the PEG and add water – the DNA dissolves back off the bead (this is called elution).



AMPure XP beads



- The bead creates a large surface area for DNA to precipitate onto
  - The bead is coated with carboxyl groups (net negative charge) - this means the DNA won't stick so hard it does not come off.
- The beads are in a solution of 20% polyethylene glycol (PEG) which has been used to precipitate DNA since 1975 (Lis and Schleif) by crowding it out of solution.

AMPure XP beads

- Higher mw DNA precipitates at lower conc of PEG.
  - They used 5% - 15% PEG in 10mM Tris pH 7.5 with 2mM EDTA.
  - At 0.35M NaCl high mw DNA was ppt by >12% PEG.
  - At 0.55M the >700bp DNA was ppt in 6.5% PEG and at 1.1M NaCl DNA> 375bp was ppt by 6.5% PEG.
  - At 12% PEG, 0.55M ppt >125bp and at 1.1M NaCl >80bp.

- Conclusion: small polynucleotides require a higher [salt] for strong binding to the beads, so salt can be used to manipulate the release of the DNA.

The beads

- XP beads are paramagnetic (magnetite sealed wit a polymer that has the carboxylic groups on it) sc they will stick to magnets.
- The buffer used has polyethylene glycol (PEG) 8000 in it (20%) and the salt is 2.5M NaCl.
- Binding capacity is supposed to be ~2ug/100ug of beads, but you don't usually get to know how many beads you have, instead you use a volume of bead suspension.
- A free agent noted that using 10% PEG 8000 with 1.25M NaCl and 10mM MgCl2 gave ppt results similar to the yield with Ampure XP beads.
  - If 7% PEG was used a lot of 50 and 100bp DNA was lost but the 200bp product (and greater) was nearly quantitative.