**Genome Architecture**

**Study guide and readings**

***Reading assignments:*** Pdfs are posted on Canvas – in the “Files” folder.

Required:

* Pennisi\_Dark matter.
* Shualev et al\_Strawberry

Recommended.

* + Fedoroff\_TE
  + Feschotte et al\_TE

1. When one reads that there are “29,000” gene in *Theobroma cacao*, what sort of genes are probably being referred to?
2. What is the importance of fragment size in DNA extraction? Why is bigger generally better?
3. What part of a mRNA is the starting point for making a cDNA? Name the two enzymes, and briefly describe their essential properties, that are required for making a cDNA.
4. If someone refers to a “20-mer oligo” what are they talking about?
5. Define palindrome and write out your favorite one.
6. What is the relationship between restriction enzyme recognition sequence length and the average frequency (in terms of base pairs) at which that enzyme will cut in a sample of genomic DNA?
7. Briefly summarize the cloning of a DNA insert into a plasmid vector.
8. If your goal is to study promoter and dark matter sequences, what type of library would you use – cDNA or gDNA?
9. What are the three cyclic steps in PCR and in what order do they occur? What happens at each step?
10. What makes TAQ polymerase so special and useful for PCR?
11. After 30 cycles of PCR, approximately how many copies of your target sequence do you expect: tens, hundreds, thousands, millions, billions?
12. What is the difference between detecting a target gene sequence using PCR vs. nucleic acid hybridization?
13. Why is dideoxy sequencing also known as chain termination sequencing?
14. In each of the three sequencing methods described in class (Cycle, Illumina, and PACBio), how are the four nucleotides distinguished?
15. What is the principle of sequencing by synthesis (e.g. Illumina) and how does it differ from chain termination (Sanger)?
16. What is a key difference between Illumina and PACBio sequencing in terms of the fluorescently labeled nucleotides?
17. Approximately how fast is PACBio sequencing reading sequence: 10 bp per second, 10 bp per minute, or 10 bp/hour?
18. According to PACBio, what will be the approximate cost of sequencing a human genome and how long will that take?
19. For de novo assembly of a genome sequence, would you rather have short reads or longer reads? Why?
20. Automated, next generation, massively parallel, and high throughput sequencing open new vistas for you as a human and as an agriculturalist/horticulturalist. Briefly outline what you see as some of the opportunities and challenges of easily obtaining DNA sequence information.
21. If someone were to ask why you would bother going to all the work to grow plants from seeds when you can simply synthesize a genome, what would you answer?

***The following questions are based on the assigned reading on Fragaria genome sequencing***

* Why was *Fragaria vesca* chosen for this work rather than the more economically important *Fragaria x. ananassa*?
* What do the authors means when the say the used short-read technologies and performed assembly without a physical genome reference? If they authors were to do the same work today, do you believe they would also use short read technologies? Why or why not?
* Of what use was a linkage map in the sequencing of the *F. vesca* genome?
* The *F. vesca* genome is reported as 209 Mb of sequence and ~240 Mb based on flow cytometry. Why the difference?
* Of what use is the synteny of *F. vesca* with *Prunus*?
* What did the authors conclude and transposons and *F. vesca* genome size?
* Were tTNA and/or rRNA genes discovered in the *F. vesca* genome sequence, and if so, what was their relative abundance?
* What, if any, practical uses do you see for the *F. vesca* genome sequence in applied horticulture?

1. Review the classic definition of a protein coding gene and be sure to know the names and functions of the essential pieces.
2. Does knowing the chromosome number of a species allow you to predict the number of genes and/or the genome size?
3. What simple calculations lead one to conclude that there must be more DNA in a genome than just protein coding genes?
4. What is the C-value paradox?

***The following questions are based on the assigned reading -* “Shining a light on the genome’s dark matter”**

* 1. Has genome sequencing revealed more genes than anticipated or more regulatory factors than anticipated?
  2. Protein coding genes account for, on average, 2%, 20%, 100% of a plant genome?
  3. What percentage of the average plant genome would you estimate is transcribed?
     1. 5%,
     2. 20%,
     3. 80%?
  4. Why does conservation of DNA sequence in non-coding DNA imply a functional role for the non-coding DNA?

1. What is the relationship between epigenetics and the C-value paradox?
2. Which of the following allows greater transcription of a gene?
   1. Methylation
   2. Acetylation
3. How does epigenetics related to facultative heterochromatin?
4. What are transposable elements?
5. What is a key difference between the two principal classes of transposable elements?
6. How can transposable elements cause mutations and how could this be “harnessed” to related genes to functions.
7. How do transposons related to the C-value paradox?
8. What does the relatively small genome size of *Fragaria vesca* have to do with transposable elements?
9. Give one example of a famous Mendelian gene where the mutant allele is related to a transposon.