Genome Sequence Quality

When looking at sequence:

Caveat emptor

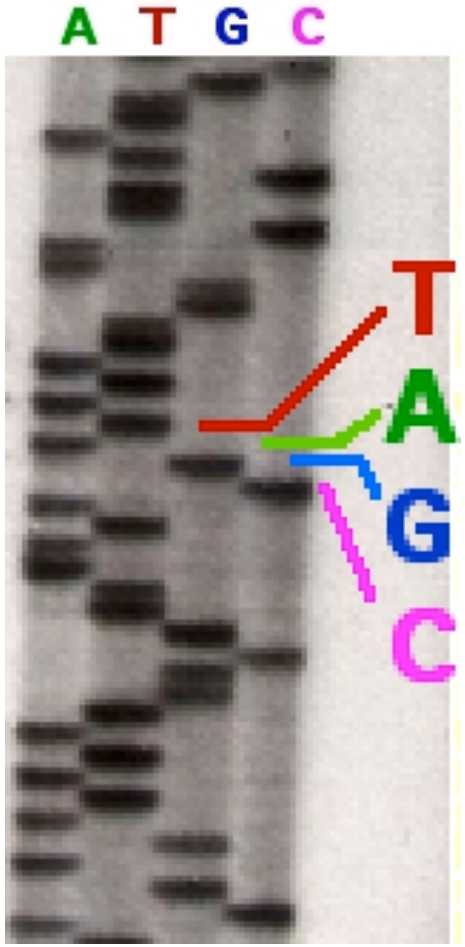
Caveat emptor

- A phrase in Latin. It is a commonly cited axiom or principle in commerce that the buyer alone is responsible for assessing the quality of a purchase before buying

- As a “consumer” of DNA sequence, one must realize that not all data points are equal and that some DNA sequences are less reliable than others

Genome Sequence Quality

- Most current sequences deposited in GenBank (and other repositories) has been derived from Sanger dideoxy (chain termination) sequencing that was developed in the late 70s
- Note at right, this is an autoradiogram of a DNA sequence using the chain termination method
- Radioactivity was the mechanism employed from the late 1970s through the late 1980s
- However, radioactivity is not amenable to high-throughput sequencing



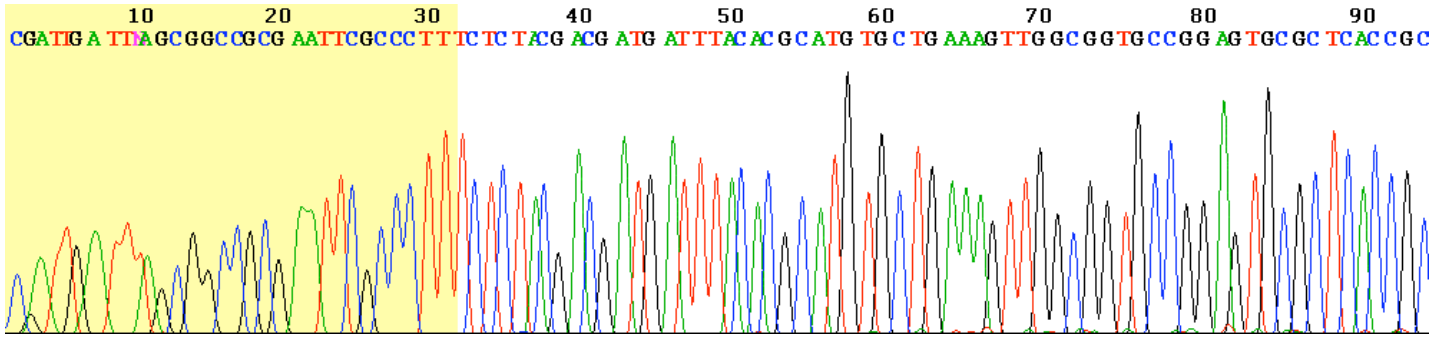
Reading an autoradiogram to determine the DNA sequence

Genome Sequence Quality

- During the late 1980s and early 1990s, the radioactive labeling was largely replaced by “dye terminator sequencing”

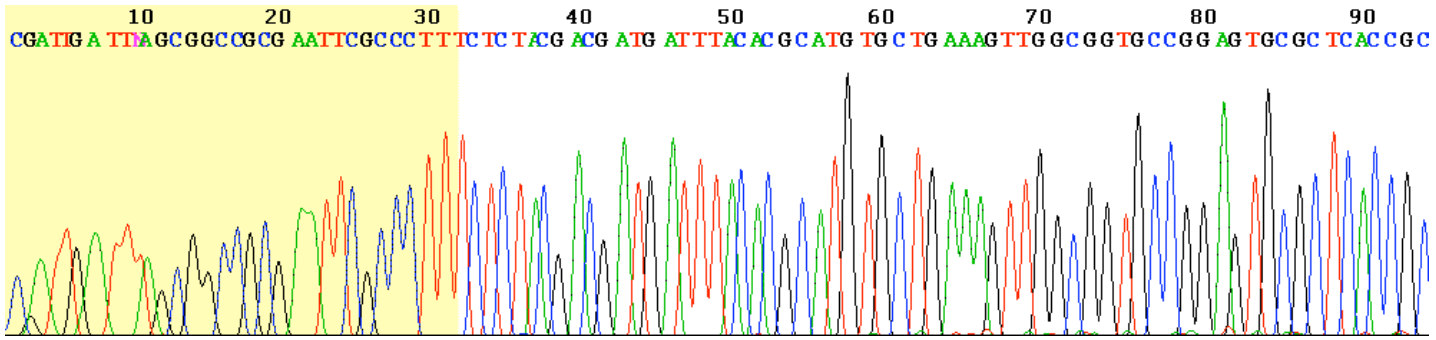
- This had several advantages:
  - No radioactivity meant no autoradiograms
  - The dyes were fluorescent which means that photons ejected post-excitation with a laser could be read by a scanner and the four dyes would eject photons at four different wavelengths
  - All four reactions for the four bases could be combined and run in a single lane
  - Computers would generate the chromatogram for the sequence eliminating a human having to interpret the gel

Example of a chromatogram



Genome Sequence Quality

- Beginnings of high-throughput sequencing
  - PCR became a standardized technology in the late 1980s
  - Computer chips were becoming powerful enough to handle real-time applications
  - Storage capacity was still minimal but becoming sufficient to hold sequence data
  - Laser techonology had allowed for scanning of the polyacrylamide gels and the simple conversion of relative photon wavelength distributions into a chromatogram



Note: The yellow shading is vector sequence that was determined from sequence comparisons to the vector seuqence and is done in a automated fashion. This sequence can be removed (or “trimmed”) automatically.

Genome Sequence Quality

- Early problems
  - *Taq* polymerase had to be improved for fidelity (source of polymorphism error)
  - Dyes that were involved in the chain termination had differing rates of incorporation by the polymerase
  - The length of reads (of high sequence quality) was hampered by this incorporation effect
  - How to develop a criteria for an accurate base call

Genome Sequence Quality

Early 1990s potential errors from this would be:

1. Single nucleotide polymorphisms (SNPs) that are introduced using the *Taq* polymerase

2. Misreads/shorted read length due to different rates of incorporation in the dyes could lead to SNPs

3. Not trimming sequences when the quality was low (towards the end of the reads)

It became important to identify the quality of the base call to to each base to ensure that reads from the chromatogram were of sufficient quality to be of biological use

Note: When using DNA sequences from the early 1990s, *caveat emptor*

Genome Sequence Quality

- With the advent of automated sequencing in the early 1990s:
  - cDNA libraries can be quickly sequenced to identify mRNA transcripts of genes
  - Sequence genomic clones of DNA completely for positional cloning, finding introns, promoters, etc.
  - Genomic clones can fragmented, sequenced and reassembled by homology
  - And so on….the possibilities were seemingly endless

The genomics revolution had begun!

Genome Sequence Quality

- Expressed sequence tags (ESTs) were developed in the early 1990s as a way to quickly catalog the expressed genes of an organism

- EST sequencing was originally implemented as “quick and dirty”
  - Read lengths were short (400-600 bp)
  - cDNAs were picked randomly
  - Error rate in sequencing was rather high (2-5%)
  - Single pass sequencing was employed to make it cheap

How is a full-length cDNA different than an EST???

Genome Sequence Quality

- Sources of polymorphisms in EST sequencing
  - As noted, the sequencing is done quickly in order to maximize number of clones rather than the highest quality of each base
  - ESTs may come from a range of isolates/cultivars/subspecies which have innate polymorphisms in their alleles (allelic variation)
  - cDNAs libraries (particularly older ones) were made with AMV (error rate: 1 base every 17,000) and MMLV (1 error every 30,000) reverse transcriptase which are inherently error prone and will introduce SNPs
  - The PCR reactions run for sequencing were using older versions of *Taq* which was error prone which will introduce SNPs

Genome Sequence Quality

- During the mid 1990s, the reagents and machinery were being optimized for higher quality and longer reads
- During the mid to late 1990s, the machinery was optimized to minimize volumes and maximize number and lengths of reads in 96 well and 384 well formats
- New vectors had been developed to clone DNA (e.g. P1 and BACs – Bacterial Artificial Chromosomes) that were amenable to high-throughput sequencing
- Protocols for efficient sequencing of BACs via a shotgun strategy were developed
- By the late 1990s, sequencing of cDNA and genomic libraries had become largely automated and extensive improvement in quality and length of reads

Genome Sequence Quality

- Phred is a base calling program for DNA sequence traces
- Originally developed by Drs. Phil Green and Brent Ewing at U of Washington
- Phred uses the chromatogram/trace data to make a highly accurate, base-specific quality score
- Phred examines the peaks around each base call to assign a quality score
- These scores range from 4 to 60 with higher scores corresponding to higher quality
- Phred scores for each base make them ideal to evaluate the quality of a sequence (overall and for trimming)
- The threshold for high quality sequence has been traditionally set at 20 or more

Genome Sequence Quality

Phred quality score	Probability that the base is called wrong	Accuracy of the base call
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Genomic Sequence Quality

But where to store, organize and distribute all of this sequence data??

NCBI – National Center for Biotechnology Information is one of several organizations (DDBJ and EMBL are two othes) created for this purpose

Within NCBI, there are number of databases that hold different types of sequence and sequence related data of varying levels of quality

Some of those databases will be highlighted in this talk

Genome Sequence Quality

- Whole Genome Shotgun (WGS) strategy
  - First proposed by Fred Sanger in 1982
  - In eukaryotes, shotgun reads and assemblies can be difficult to map to a chromosome unless they contain a marker previously known to map to the genome
  - What happens in larger, more complex genomes when one tries to assemble shotgun reads into contigs?

Genome Sequence Quality

1. Coverage is a factor….lower coverage of the genome will lead to shorter contigs

2. Transposable elements (Repeats) confound the ability to assemble reads since they occur multiple times in the genome

3. Duplicated genes in a genome can make assembly difficult if there is a high level of homology and particularly in tandem repeats

Genome Sequence Quality

At NCBI – These sorts of whole genome shotgun data are stored at the following site

http://www.ncbi.nlm.nih.gov/Genbank/wgs.html

Each shotgun sequencing project has a stable four letter identifier that is assigned to it which is the project ID

Each project will have a version number that will increase as the project matures

For example, if the project ID is ABCD:

The first assembly version would be: ABCD01000000

The first contig of the first version would be: ABCD01000001

(Note the last six digits of the ID will identify the contig)

Genome Sequence Quality

WGS submissions are only contigs…there are no supercontigs allowed (two contigs cannot be bridged by a series of N's)

WGS submissions may also have annotation associated with them, but the unannotated sequence files can have a separate project ID from the annotated files (just keep in mind)

The rice indica sequence assembly is maintained in the WGS database at NCBI (AAAAxxxxxxxx)

Genome Sequence Quality

- There are two complementary strategies to WGS that have been employed recently in maize to obtain genomic sequence
  - Methylation Filtration
  - High $C_o t$

Genome Sequence Quality

- One variation on whole genome shotgun strategy
  - Methylation filtration:
    - This is based upon the principle that repeated (i.e. transposable element) sequences are often heavily methylated in eukaryotic genomes
    - Genomic DNA is sheared and cloned into vectors that are then transformed into *E. coli* strains that cleave methylated DNA
    - This strategy will enrich for hypomethylated DNA which is often euchromatic and gene dense and remove hypermethylated DNA that is generally heterochromatin and gene poor
    - Assemblies generated from MF strategy are often short in genomes that have a large content of repeats
    - Agnes will discuss this strategy in her talk on maize genomic resources

Genomic Sequence Quality

- Another whole genome shotgun strategy
  - High C_ot genomic libraries
    - High C_ot is a strategy whereby you can enrich for low complexity DNA through hybridization kinetics
    - C_ot analysis of DNA was common in the 1970s to assess the repeat content in a given genome using a spectrophotometer and knowledge of reassociation kinetics
    - Principle is that highly repeated DNA will hybridize quickly to itself, moderately repeated DNA will take longer to reanneal and low complexity DNA will take the longest to reanneal
    - Single and double stranded DNA can be separated on columns to enrich for particular types of DNA that one is interested in
    - Low complexity DNA isolated by this method can be used as a complement to the MF generated libaries to obtain DNA that is gene rich as well as DNA that flanks gene rich regions and is a useful complement to MF libraries
    - As with MF libraries, in complex eukaryotic genomes, the assemblies from these reads can be short due to islands of transposable elements
    - Agnes will discuss this in her talk on maize genomic resources

Genome Sequence Quality

At NCBI – These sorts of modified shotgun data are stored in the Genome Survey Sequences (GSS) Database

http://www.ncbi.nlm.nih.gov/dbGSS

GSS projects include MF and HiC$_o$t sequences but also include:

- -BAC end sequencing
- -Alu PCR sequences
- -Transposon tagged sequences

Presently, maize has >2 million GSS reads, which is the greatest number of any species and sorghum has >600,000 reads

GSS reads are kept as single reads and are not represented as contigs

For MF and HiCot assemblies of the reads, these are maintained and updated at their respective sites (Agnes will discuss maize and Kevin will discuss sorghum)

Genome Sequence Quality

- ESTs that were discussed earlier are stored at dbEST
- http://www.ncbi.nlm.nih.gov/projects/dbEST
- dbEST contains sequence information on ‘single pass’ cDNA sequences
- Note: Queries can be written that will allow you to download all EST sequences from dbEST for a given species of interest
- Summary statistics are given dbEST at regular updates for all of the available ESTs for all species
- ESTs in dbEST can be from varying ages

Genome Sequence Quality

- The japonica subspecies of rice (cultivar Nipponbare) was sequenced using a BAC-by-BAC approach
- Luke has described the BAC-based seqeuencing strategy in detail earlier
- BAC sequencing goes through three stages as it progresses from contigs to finished sequence
  - Phase I
  - Phase II
  - Phase III

Genomic Sequence Quality

- Phase 1
  - This indicates the sequences of the BAC (can also be YAC, PAC, etc) have been assembled into contigs
  - The order of these contigs is not known relative to the BAC
  - The size of the gaps is not known
  - The sequences in phase 1 are noted with a WARNING in the comment line with a summary of the status of the contigs

Genome Sequence Quality

- Phase 2
  - As the sequencing effort on the BAC continues, the coverage becomes greater
  - This allows the contigs to (often) coalesce as well as using library information, the order of the contigs can be determined and their orienation is correct
  - Gaps may exist and can be of an unknown size
  - In some cases, the complete sequence of the BAC may exist but the quality scores are too low for some regions to promote the record to phase 3

Genome Sequence Quality

- Phase 3
  - This sequence is considered “finished”
  - The gaps between the ordered and oriented contigs from phase 2 have been filled
  - The base quality scores are all sufficient
  - This may (or may not) have annotation associated with it

  Phase 3 sequences are then deposited into the appropriate organismal division of Genbank

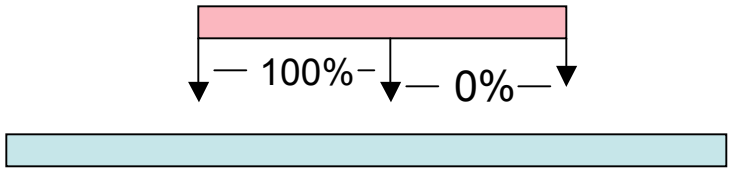
  For rice, this is the PLN database which covers plant and fungal species

Genome Sequence Quality

Exercise 1: What if a whole genome shotgun (WGS) sequence assembly matches at 100% identity over 50% of its length to a BAC based sequence and the other 50% has no identity??

Whole genome shotgun assembly

BAC based phase 3 assembly



Some points to consider:

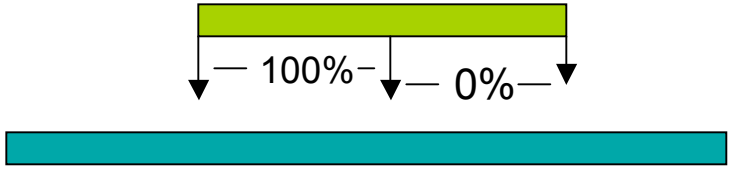
1. How was the WGS assembly created?
2. How was the phase 3 BAC sequence generated?
3. Are these two sequences derived from the same genotype?

Genome Sequence Quality

Exercise 2: What if a whole genome shotgun (WGS) sequence assembly matches at 100% identity over 50% of its length to a cDNA sequence??

cDNA sequence

Whole genome shotgun sequence



Some points to consider:

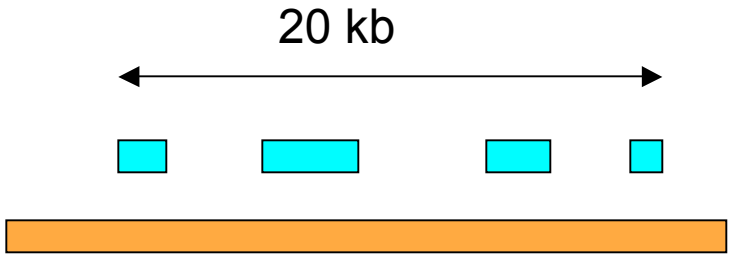
1. How was the WGS assembly created?
2. How was the EST created?
3. Are these from the same genotype?
4. Can ESTs be chimeric?

Genome Sequence Quality

Exercise 3: What if a full-length cDNA matches a phase 3 BAC sequence at 100% identity over 100% of the length but the alignments are scattered over 20kb?

cDNA sequence

Whole genome shotgun sequence



Some points to consider:

1. Is this a valid alignment?
2. What might cause the gaps?

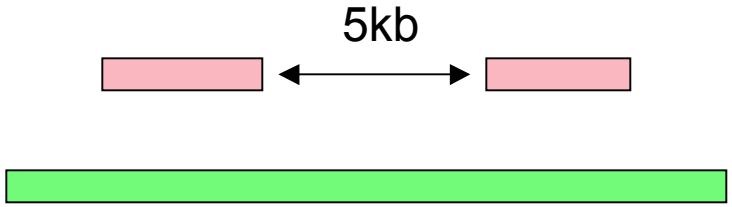
Genome Sequence Quality

Exercise 4: What if a Whole Genome Shotgun sequence assembly from one genotype matches a phase 3 BAC sequence of another genotype at 100% identity over 100% of the length but the alignment is divided into two separate pieces that are separated by 5 kb?

Whole genome shotgun sequence

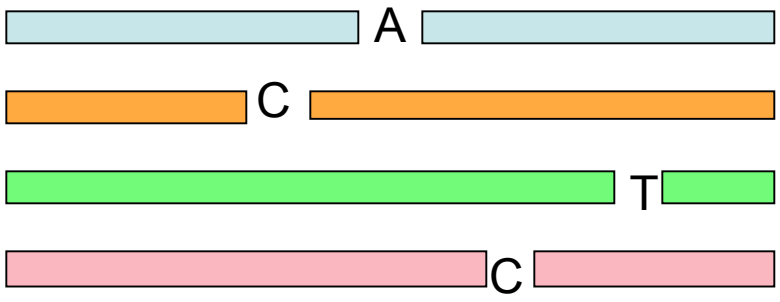
Phase 3 BAC sequence

What might be the cause of the gap?



Genome Sequence Quality

Exercise 5: What if we have ESTs for the same gene from four different genotypes in the same species that have small differences in their sequence. What might be the cause of the differences?



Genome Sequence Quality

Exercise 6: What if we have 3 ESTs for the same gene that match >99% identity and >99% of the length with respect to a phase 3 BAC (and are 100% identical to each other) where the difference is due to a single bp insertion in the ESTs. The ESTs and BAC sequence come from the same genotype. Which is right?