

Genome sequencing

Miler T. Lee

28 January 2005
revised 19 February 2005

1 Hierarchical

- Top-down, map-based, clone-by-clone strategy to genome sequencing.
 - byproduct was creation of high resolution physical and genetic maps
 - collaborative effort
1. Create **DNA libraries** consisting of genomic clones in chunks of ~50-200 kb
 - Digest or sonically shear genomic DNA - usually individual chromosomes
 - ligate fragments into vectors - BACs, PACs (P1 phage-derived), cosmids
 - 5-10x redundancy
 2. Form a **tiling path** by aligning overlapping fragments
 - **Hybridization** with radioactive probe containing a sequence of interest will identify which clones have the common sequence
 - **Fingerprint** the group of clones with a restriction digest and align sequences with corresponding fingerprints, to form Mb-length contigs
 - **End-sequence** both ends of the clones to span remaining gaps - i.e., if you can align one end of the clone with one assembled contig, the other end will with high probability link to an adjacent contig or extend into a gap
 3. Sequence individual clones along the tiling path
 - shear the clone into smaller fragments using sonication (ensures unique ends) and subclone - M13 phagemid for ~1kb fragments, plasmids for ~2-3 kb fragments, though you have to sequence from both ends due to length
 4. Assemble the draft genome
 - **Filter** to remove vector-contaminated fragments, misannotations, recombinations
 - Assemble the **layout** of the contigs for each clone using mapping information (e.g., *in silico* digest of the inferred contig compared with the real restriction fingerprint)
 - Merge contigs by aligning overlaps, incorporating information such as known mRNA sequences
 5. Finishing
 - Draft sequence consists of sequence-contig scaffolds with 5-10x coverage, with gaps every few hundred kb
 - finishing efforts include filling gaps, increasing quality of sequence reads to error rate < 1/10,000

2 Whole-genome shotgun

- Computer-aided large-scale assembly of contigs based mostly on sequence overlaps
 - All sequencing is done ahead of time using PCR products, to form shotgun libraries of sequence reads
1. Screener
 - repeat-mask
 2. Overlapper
 - compare each sequence read against every other read to find overlaps of >40 bp with <6 base differences (to account for heterozygosity, sequencing errors)
 - p-value on the order of 1E-17
 3. Unitigger
 - to distinguish between true alignment and alignment due to repetitive sequences (repetitive is not necessarily low-complexity)
 - a **unitig** is a contig formed from overlapping unambiguously unique sequences
 - each unitig is statistically analyzed to determine whether two different sequences (wrt genomic location) have been **overcollapsed** into a single unitig; this is based on how much coverage you would expect a given location to have as compared to the actual number of reads that appear to overlap that location
 4. Scaffolder
 - links **U-unitigs** (Unitigger-filtered unitigs) into scaffold contigs
 - mate-pair information is used: mate-pair libraries are generated for fragments of different lengths consisting of end-sequences
 - mate-pair information for 2-10 kb clones are used to pair U-unitigs, with >2 mate-pairs required for accurate linking
 - 50 kb mate-pairs can be used to verify the assembly; disagreements are called **break-points**
 5. Repeat Resolver
 - unitigs failing the unitigger check can be used to fill gaps if supported by >2 mate-pairs to form **rocks**
 - unitigs supported by a single mate-pair are inserted to form **stones**
 - any unitigs disagreeing with the proposed alignment are discarded
 - remaining gaps are walked across *in silico* or by sequencing PCR products using primers from the flanking unitigs
 6. Verification
 - chromosomes are assigned by matching the assembly to physical maps
 - accuracy is assessed based on STS and EST information
 - corrections are made based on BAC clone fingerprints

3 Human genome

1. Celera effort was actually a compartmentalized shotgun assembly
 - 15 Gb of Celera sequence reads at 5x coverage + 4.4 Gb of public sequence shredded to 3x coverage
 - Contiguous chromosome segment chosen based on physical maps / public draft assembly
 - Isolate Celera scaffolds matching to that region and manually check
 - Shred public bactigs from that region and reassemble with Celera reads
2. Weber & Myers' argument for WGS (1997)
 - Computer simulations indicated feasibility
 - contigs/scaffolds would be built that spanned adjacent STSs (200-500 bp), far smaller than the size of bacterial genomes that were successfully sequenced using WGA
 - polymorphisms would be detected due to the multi-donor approach, which can be distinguished from sequencing errors based on phred score and base distribution
 - elimination of artifacts arising from BAC constructions
 - less expensive
3. Green's rebuttal
 - clone-based sequencing is modular, can target specific regions, allows immediate data quality assessment since you can go back to the subclones to do additional sequencing, single haplotype
 - WGS throws out clone data (no "clone tracking")
 - human genome is much harder and less well understood than bacterial
 - sequencing all at one is more expensive, less reliable, can't take advantage of future technologies
 - higher rate of false joins because there are more ways of overlapping
 - errors not detected until assembly phase