Notes 1 – The Technologies

Shannon Straub, Oregon State University

1. **Sanger Sequencing vs. Next-Generation Sequencing**
   The advantages of next-generation sequencing over Sanger sequencing for genome scale projects include in vitro library construction and template amplification, massive parallelization, low reagent cost, and low cost per base of sequence produced. The disadvantages of next-generation sequencing of shorter read lengths and lower accuracy are overcome by the huge amounts of data produced.

2. **Next-Generation Sequencing Steps**
   A. Library Preparation

   B. Template amplification
      1. Emulsion PCR

      2. Solid-Phase Amplification

   C. Sequencing
      1. by synthesis
         a. Cyclic reversible termination (CRT)

         b. Single nucleotide addition (SNA or pyrosequencing)

      2. by ligation

   D. Imaging
      1. 4-color imaging of single events

      2. 1-color imaging of single events

      3. Bioluminescence

3. The "Big Three" Platforms

A. Illumina
B. 454/Roche
C. SOLiD (ABI by Life Technologies)

4. Other Next-Generation and Third Generation Sequencing Platforms

A. Ion Torrent

1. Semiconductor sequencing by synthesis using pH change caused by H^+ release upon base incorporation and SNA.

B. Pacific Biosciences

1. Single molecule real-time sequencing by synthesis.

C. Oxford Nanopore Technologies

1. Single molecule sequencing by direct electronic analysis of DNA passing through a nanopore.

Table 1. Comparison of the three most commonly encountered next-generation sequencing platforms.

Platform	Template Type/ Template Preparation	Sequencing Method	Data Collection Method	Read Length	Sequence per Run	Accuracy	Most Common Error Type
Illumina	clonally amplified by solid-phase amplification	by synthesis – cyclic reversible termination	4-color imaging	~100 bp	200 Gb	>85-90%	substitutions
454/Roche	clonally amplified by emulsion PCR	by synthesis – single nucleotide addition	bioluminescence imaging	~400 bp	400 - 600 Mb	99%	indels
SOLiD	clonally amplified by emulsion PCR	by ligation	4-color imaging	~75 bp	300 Gb	99.9%	substitutions

Technology References:

Ansorge, W.J. 2009. Next-generation DNA sequencing techniques. *New Biotechnology* 25: 195-203.
Branton, D. *et al.* 2008. The potential and challenges of nanopore sequencing. *Nature Biotechnology* 26: 1146-1153.
Cockroft, S.L. *et al.* 2008. A single-molecule nanopore device detects DNA polymerase activity with single nucleotide resolution. *Journal of the American Chemical Society* 130: 818-820.
Mardis, E.R. 2008. Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics* 9: 387-402.
Metzker, M.L. 2010. Sequencing technologies – the next generation. *Nature Reviews Genetics* 11: 31-46.
Shendure, J. and H. Ji. 2008. Next-generation DNA sequencing. *Nature Biotechnology* 26: 1135-1145.

http://www.illumina.com/
http://www.454.com/
http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing.html
http://www.iontorrent.com/
http://www.pacificbiosciences.com/
http://www.nanoporetech.com/

Notes 2 – An Overview of Applications

Rich Cronn, USDA Forest Service

1. **An overview of the possibilities**:

	DNA	RNA
Sequencing Experiments	Genome Sequencing Microsatellite Discovery Polymorphism (SNP, Indel, CNV) discovery Structural Variation Analysis	Transcriptome Sequencing Polymorphism (SNP, Indel) discovery Structural Analysis Discovery of miRNA, siRNA Epigenomics
Counting Experiments	Metagenomics Population Genetics Mutation Rate Estimation Chromatin Immunoprecipitation Seq Methyl-Seq	Metagenomics Digital Gene Expression Profiling

2. **Sequence assembly methods.**
   A. Reference guided assembly methods. Examples: MACQ, BLAT
   B. De novo assembly methods. Examples: ABYSS
   C. Hybrid Methods: Velvet, YASRA

3. **Sample Multiplexing:** Multiplexing is crucial for efficient use of sequencing platforms; helps to scale sample complexity to sequencing capacity, significantly reduce costs.
   A. ***Commercially available*** for all sequencing platforms: 454 (multiplex identifiers, MIDs), Illumina (index sequencing adapters), and SOLiD (barcode adapters).
   B. ***Non-commercial solutions*** extend the utility of multiplexing:
      1. More adapter combinations
      2. Error detecting and error detecting/error correcting multiplex codes
   C. ***Examples***: "Routine" multiplex sequencing run on the Illumina; multiplex chloroplast genomes (Cronn et al. 2008); multiplex sequencing of amplicons (Craig et al. 2008); multiplex sequencing of 1,000s of individuals (1000X MID for 454).

4. **Targeted Sequencing:** Many ways to obtain select targets for sequencing:
   A. ***Pooling of PCR amplicons.*** Efficient, stringent. Ideal for low coverage studies with large sample sizes; sample prep limitations; hard to match capacity of sequencers.
   B. ***Microdroplet PCR.*** New technology designed to amplify larger numbers of targets (low thousands) in parallel. Targets tend are smaller, emphasize 454 platform.
   C. ***Hybridization-based enrichment.*** Less stringent than PCR; ideal for high genomic coverage studies, large numbers of individuals; works on all major platforms; does not require exact matches to targets; well suited for plant organelle genomes.
   D. ***Examples***: PCR pooling for chloroplast genomes (Parks et al. 2009); sequencing of pooled short amplicons (Bundock et al. 2009); sequencing of pooled long amplicons (Bansal et al. 2010); microdroplet PCR (Tewhey et al. 2009); human exon hyb-seq

(Gnirke et al. 2009); plant chloroplast genome hyb-seq (Parks et al. Botany 2010); microsatellite hyb-seq (Cronn, unpublished); all methods (Mamanova et al. 2010).

5. **Anonymous Genomic Reduction:** Filtering genomic complexity for SNPs:
   A. ***Restriction-Site Associated DNA sequencing.*** Filtering genomes for restriction fragments of a defined size; filtering can also include methylation status. Methodologically facile, informatically challenging.
   B. ***Examples***: SNP discover using the Illumina (Baird et al. 2008) and 454 (Maughan et al. 2009).

6. **Genome Sequencing:** With small genomes, there's no need to select:
   A. ***Assembling targets from the genome.*** Depending on size of genome, organelles and repetitive DNAs can be easily assembled.
   B. ***Sources for genome references.***
   C. ***Examples***: assembling chloroplast genomes from total genomic DNA (Dempewolfe et al. 2010; Straub Botany 2010); scanning for adaptive variation in small genomes (Turner et al. 2010).

Application References:

Baird et al. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLOS One 3:e3376

Bansal et al. 2010. Accurate detection and genotyping of SNPs utilizing population sequencing data. *Genome Research* 20:537–545.

Bundock et al. 2009. Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing. *Plant Biotechnology Journal* 7:347–354.

Craig et al. 2008. Identification of genetic variants using bar-coded multiplexed sequencing. *Nature Methods* 5:887-893.

Cronn et al. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Research* doi:10.1093/nar/gkn502.

Dempewolf et al. 2010. Establishing genomic tools and resources for *Guizotia abyssinica* (L.f.) Cass.—the development of a library of expressed sequence tags, microsatellite loci, and the sequencing of its chloroplast genome. *Molecular Ecology Resources* doi: 10.1111/j.1755-0998.2010.02859.x

Gnirke et al. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology* 27:182-189.

Mamanova et al. 2010. Target-enrichment strategies for next-generation sequencing. *Nature Methods* 7:111-118.

Maughan et al. 2009. SNP discovery via genomic reduction, barcoding, and 454-pyrosequencing in Amaranth. The Plant Genome 2:260–270.

Parks et al. 2009. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biology* 7:84.

Parks et al., 2010. Striving for the First Finish Line: A Whole-Plastome Phylogeny for the Entire Genus *Pinus*. Botany 2010. Monday 2:15pm, Ballroom C, Session C2. Abstract 689.

Straub et al., 2010. Partial characterization of the *Asclepias syriaca* L. (common milkweed) genome using next-generation sequencing. Botany 2010. Wednesday 10:30 am, Room 551B, Session 53. Abstract 667.

Tewhey et al. 2009. Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nature Biotechnology* 27:1025-1031.

Turner et al., 2010. Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. *Nature Genetics* 42:260-264.

Notes 3 – Overview of Next-Generation Sequence Sample Preparation

Matt Parks

Basic process of Next-Gen Sequence prep:

1. Accurate quantification of input DNA concentration.
2. Fragmentation to desired size range.
3. Blunt-ending of fragmented DNA (and A-tailing for Illumina preps).
4. Ligation of adapters to fragmented, end-polished DNA.
5. Second size selection step by gel isolation.
6. PCR enrichment of adapter-ligated, size-selected DNA fragments.
7. Accurate Quantification of enriched DNA concentration.

Samples are cleaned between each step from step 2-7.

Notes on each step.

1. ***Quantification***: We recommend the Qubit fluorometer, as this equipment is relatively cheap, fast, very accurate, and consumes little of your sample.

2. ***Fragmentation***: Sonication is currently the best option, but the initial investment is very high ($20-40000). If sonication is not available, nebulization is probably the best second option, although sample loss can be quite high (up to 50% or more).

3 and 4. These steps are typically easy and rapid.

5. ***Gel isolation*** is time-consuming and sample recovery is not great, but it is still the most commonly used option. Other options include: dilutions of AMPure solution to select for a specific size range, skipping size selection altogether (appropriate for single-end rather than paired-end reads).

6. ***PCR enrichment*** is simple and fast. If you have high input DNA concentrations, you may be advised to keep cycles low (10) to decrease bias. If you have low initial input DNA or sample loss prior to this step, it may be necessary to use higher (18) cycles.

7. See step 1.

Purchasing Notes.

Before you jump into a sample prep, shop around for the best reagent costs. NEB offers reasonable prices on reagents put together specifically for Illumina, 454 and SOLiD preps. In addition, both primers and adapters are able to be custom-made, usually for around $75 per oligonucleotide pair. Sequence details can be found in the following papers:

Bentley DR, et al. **2008**. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53.

Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T. **2008**. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Research 36: e122.

Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, Corneveaux JJ, Pawlowski TL, Laub T, Nunn G, Stephan DA. **2008**. Identification of genetic variants using bar-coded multiplexed sequencing. Nature Methods 5: 887-893.

Notes 4 – Implementation and Bioinformatics

Aaron Liston

I. Where to sequence & how much will it cost?
   A. Sequencer locations (Google maps)
   B. Sequencing Services
      1. Illumina Genome Analyzer prices
      2. 454 – same price, 1/10 the data
      3. SOLiD – twice the price; twice the data
   C. Sample Submission

II. Bioinformatics
   A. Data Files
      1. Data Transfer
      2. FASTQ Format
      3. Read Mapping Formats
         a. Eland
         b. SAM / BAM
   B. Computing Hardware
      1. Desktop
      2. Server / Cluster
      3. CLC Genomics Server
      4. Cloud Computing
   C. Data Processing Options
      1. Commercial Packages
      2. Bioinformatics Expert
      3. Undergraduate Programmer
      4. Web-Services
         a. ComputePortalProject
         b. Galaxy
      5. Do it Yourself

a. Scripting Languages (Perl, Python, Ruby, Java, etc.)

b. Linux + Google

III. Genome Assembly

A. Sequence Reads

1. Sort Tags

2. Filter by Quality

B. De Novo Assembly

1. Overlap-layout-consensus approach (tiling or walking)

2. Eulerian path / de Bruijn graph

3. Paired Ends

C. Reference Guided Assembly

1. YASRA (yet another short read assembler)

a. Reference can be 80-90% divergent.

b. Closes gaps with overlap-layout consensus.

c. Creates a chimeric pseudo-reference.

d. Repeats the process until no additional improvement.

2. Velvet 1.0

a. Columbus extension released June, 2010

b. Uses previously mapped reads (SAM/BAM format)

D. (Chloroplast) Genome Finishing

1. Scaffold Assembly

a. Sanger software (Sequencher, CodonCode, BioEdit, etc.)

b. MULAN (Ovcharenko et al. , 2005)

c. post-YASRA Python scripts (Short Read Toolbox, Zach Foster)

2. Annotation

a. DOGMA (Wyman et al. 2004)

b. MAKER (Yandell, 2009)

c. Drop down annotation in BioEdit (Hall, 1999-2007)

3. Alignment

a. MAFFT (Katoh, 2002-2010)

b. Vista (Dubchak et al, 2003-2010)

c. Refinement

i. RASCAL (Thompson et al. 2003)

ii. AQUA (Muller, 2010)

An incomplete list of recent and important Next-Generation citations

Reviews

Ansorge WJ. 2009. Next-generation DNA sequencing techniques. New Biotechnology 25: 195-203.

Mardis ER. 2008. Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics 9: 387-402.

Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. Trends in Genetics 24: 133-141.

Metzker ML. 2010. Sequencing technologies – the next generation. Nature Reviews Genetics 11: 31-46.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. Nature Biotechnology 26: 1135-1145.

Tucker T, Marra M, Friedman JM. 2009. Massively parallel sequencing: The next big thing in Genetic medicine. The American Journal of Human Genetics 85(2): 142-154.

Basic Applied

Bentley DR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Research 36: e105.

Hudson ME. 2008. Sequencing breakthroughs for genomic ecology and evolutionary biology. Molecular Ecology Resources 8: 3-17.

Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M. 2008. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature 456: 66-72.

Mardis ER, et al. 2009. Recurring mutations found by sequencing an acute myeloid leukemia genome. N Engl J Med: NEJMoa0903840.

Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ. 2008. A large genome center's improvements to the Illumina sequencing system. Nature Methods 5: 1005-1010.

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J. 2008. The diploid genome sequence of an Asian individual. Nature 456: 60-65.

Botany-Specific

Chung SM, Gordon VS, Staub JE. 2007. Sequencing cucumber (Cucumis sativus L.) chloroplast genomes identifies differences between chilling-tolerant and-susceptible cucumber lines. Genome 50: 215-225.

Greiner S, Wang X, Herrmann RG, Rauwolf U, Mayer K, Haberer G, Meurer J. 2008. The complete nucleotide sequences of the five genetically distinct plastid genomes of Oenothera, subsection Oenothera: II. A microevolutionary view using bioinformatics and formal genetic data. Molecular Biology and Evolution.

Greiner S, Wang X, Rauwolf U, Silber MV, Mayer K, Meurer J, Haberer G, Herrmann RG. 2008. The complete nucleotide sequences of the five genetically distinct plastid genomes of Oenothera, subsection Oenothera: I. Sequence evaluation and plastome evolution. Nucleic Acids Research 36: 2366.

Jansen RK, Wojciechowski MF, Sanniyasi E, Lee SB, Daniell H. 2008. Complete plastid genome sequence of the chickpea (Cicer arietinum) and the phylogenetic distribution of rps12 and clpP intron losses among legumes (Leguminosae). Molecular Phylogenetics and Evolution.

Jansen RK, Kaittanis C, Saski C, Lee SB, Tomkins J, Alverson AJ, Daniell H. 2006. Phylogenetic analyses of Vitis (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. BMC Evolutionary Biology 6: 32.

Jansen RK, Cai Z, Raubeson LA, Daniell H, dePamphilis CW, Leebens-Mack J, Muller KF, Guisinger-Bellian M, Haberle RC, Hansen AK. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci USA 104: 19369.

Kahlau S, Aspinall S, Gray JC, Bock R. 2006. Sequence of the Tomato Chloroplast DNA and Evolutionary Comparison of Solanaceous Plastid Genomes. Journal of Molecular Evolution 63: 194-207.

Kim JS, Jung JD, Lee JA, Park HW, Oh KH, Jeong WJ, Choi DW, Liu JR, Cho KY. 2006. Complete sequence and organization of the cucumber (Cucumis sativus L. cv. Baekmibaekdadagi) chloroplast genome. Plant Cell Reports 25: 334-340.

Mathews S. 2009. Phylogenetic relationships among seed plants: Persistent questions and the limits of molecular data. American Journal of Botany 96: 228-236.

Matsuoka Y, Yamazaki Y, Ogihara Y, Tsunewaki K. 2002. Whole chloroplast genome comparison of rice, maize, and wheat: implications for chloroplast gene diversification and phylogeny of cereals. Molecular Biology and Evolution 19: 2084-2091.

Moore MJ, Bell CD, Soltis PS, Soltis DE. 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proc Natl Acad Sci USA 104: 19363.

Moore MJ, Dhingra A, Soltis PS, Shaw R, Farmerie WG, Folta KM, Soltis DE. 2006. Rapid and accurate pyrosequencing of angiosperm plastid genomes. BMC Plant Biology 6: 17.

Ruhlman T, Lee SB, Jansen RK, Hostetler JB, Tallon LJ, Town CD, Daniell H. 2006. Complete plastid genome sequence of Daucus carota: implications for biotechnology and phylogeny of angiosperms. BMC Genomics 7: 222.

Soltis DE, Albert VA, Savolainen V, Hilu K, Qiu YL, Chase MW, Farris JS, Stefanovic S, Rice DW, Palmer JD. 2004. Genome-scale data, angiosperm relationships, and ending incongruence: a cautionary tale in phylogenetics. Trends in Plant Science 9: 477-483.

Steele PR, Guisinger-Bellian M, Linder CR, Jansen RK. 2008. Phylogenetic utility of 141 low-copy nuclear regions in taxa at different taxonomic levels in two distantly related families of rosids. Molecular Phylogenetics and Evolution 48: 1013-1026.

Whittall JB, Syring J, Parks M, Buenrostro J, Dick C, Liston A, Cronn R. 2009. Finding a (pine) needle in a haystack: chloroplast genome sequence divergence in rare and widespread pines. Molecular Ecology 19(s1): 100-114.

Zou XH, Zhang FM, Zhang JG, Zang LL, Tang L, Wang J, Sang T, Ge S. 2008. Analysis of 142 genes resolves the rapid diversification of the rice genus. Genome Biology 9: R49.

Methodology

Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, Corneveaux JJ, Pawlowski TL, Laub T, Nunn G, Stephan DA. 2008. Identification of genetic variants using bar-coded multiplexed sequencing. Nature Methods 5: 887-893.

Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Research 36: e122.

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature Biotechnology 27: 182-189.

Harismendy O, Frazer KA. 2009. Method for improving sequence coverage uniformity of targeted genomic intervals amplified by LR-PCR using Illumina GA sequencing-by-synthesis technology. BioTechniques 46: 229-231.

Herman DS, Hovingh GK, Iartchouk O, Rehm HL, Kucherlapati R, Seidman JG, Seidman CE. 2009. Filter-based hybridization capture of subgenomes enables resequencing and copy-number detection. Nature Methods 6:507-510.

Parks, M. Cronn R., Liston A. 2009. Increasing phylogenetic resolution at low taxonomic levels using massively paralle sequencing of chloroplast genomes. BMC Biology 7:84.

Porreca GJ, et al. 2007. Multiplex amplification of large sets of human exons. Nature Methods 4: 931-936.

Phylogenomics and Bioinformatics

Cock PJ, Fields CJ, Goto N, Heuer ML, and Rice PM. 2009. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acid Research 38(6): 1767-1771.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nature Reviews Genetics 6: 361-375.

Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends in Genetics 22: 225-231.

Miller JR, Koren S, Sutton G. 2010. Assembly algorithms for next-generation sequencing data. Genomics 95: 315-327.

Pop M, Salzberg SL. 2008. Bioinformatics challenges of new sequencing technology. Trends in Genetics 24: 142-149.

Rokas A, Chatzimanolis S. 2008. From gene-scale to genome-scale phylogenetics: the data flood in, but the challenges remain. Methods in Molecular Biology 422: 1-12.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research 18: 821-829.

Third-Generation Sequencing Technology

Branton D, et al. 2008. The potential and challenges of nanopore sequencing. Nature Biotechnology 26: 1146-1153.

Gupta PK. 2009. Single-molecule DNA sequencing technologies for future genomics research. Trends in Biotechnology 26: 602-611.

Pushkarev D, Neff NF, Quake SR. 2009. Single-molecule sequencing of an individual human genome. Nature Biotechnology 27(9): 847-850.

GLOSSARY

Methods:

pyrosequencing: Sequencing DNA using chemiluminescent enzymatic reactions. As each nucleotide is added, a flash of light is emitted that enables the identification of its base. The Roche 454 method.

sequencing by synthesis: Sequencing by catalyzing the synthesis of the complement of a single strand of DNA and simultaneously determining which nucleotide is added at each position. The Illumina (Solexa) method.

sequencing by ligation: A single-stranded target DNA molecule is interrogated with fluorescently-labeled oligonucleotides using DNA ligase. Multiple rounds of ligation of the target with various oligonucleotide pools and starting positions on the DNA strand result in determination of the sequence with high accuracy due to the sensitivity of DNA ligase and repeated interrogation of bases. The ABI SOLiD method.

flow cell: A glass slide divided into eight lengthwise lanes in which the sequencing process takes place during Illumina sequencing. One lane is typically reserved for a sequencing control.

genomic library: A collection of DNA fragments from an organism that represent all or part of the genetic information present in one or more of its genomes (nuclear, mitochondrial, chloroplast). Such DNA may be directly extracted from the tissue of an organism, or may be reverse-transcribed from mRNA (a cDNA library).

enrichment: Using PCR or a combination of hybridization and PCR to enrich for specific components of a genomic library or sample prep.

micro-read: Sequences 25-125 bases long, obtained from next-generation sequencing.

single-end: Sequencing in which fragmented DNA molecules are sequenced from one end only.

paired-end: Sequencing in which fragmented DNA molecules are sequenced first from one end, then from the other end. This allows read pairs to be aligned to a reference genome with a known, approximated distance in between, resulting in more effective mapping. Typical 'insert' sizes are several hundred base pairs long, but may be over a thousand base pairs long.

mate-pair: Sequencing both ends of long (> 1 kbp) DNA molecules. In Illumina sequencing, the reads are oriented ←…→ in mate-pairs instead of →…← as in paired ends.

adapters: Short dsDNA molecules that are ligated onto both ends of fragmented DNA during Next-Generation sample prep. Adapters carry a sequence complementary to PCR primers used in a subsequent enrichment step, and may also carry short identifying sequences called barcodes (see multiplex).

multiplex: A sequencing reaction that contains DNA from more than one accession in the same sequencing pool. In such cases, the DNA of each accession is recognized by the use of barcoded adapters (see adapters).

transcriptome (RNA-Seq): The process of converting RNA extracted from a tissue into a cDNA library, and subsequently sequencing the cDNA library (or select parts of it).

targeted selection: Isolating specific regions of an organism's genome. This is typically done either through targeted PCR or hybridization with targeted probes followed by PCR.

exome: The collection of exons found in an organism's genome. This may refer solely to protein-coding exons, or may include any region of the genome that codes for an RNA product.

Bioinformatics:

quality score: A measure of confidence in a base call at a specific position in a sequence read. The quality score (Q) is related to the probability of error (incorrect base call) such that $Prob_{error} = 10^{Q/-10}$. Typically, quality scores are either quantified using the Phred scale or are converted to the Phred scale prior to use.

de novo assembly: Assembly of sequence reads into larger contigs based solely on their shared/overlapping sequences, i.e. without the aid of a reference genome.

reference-guided (or -assisted) assembly: Assembly of sequence reads into contigs by mapping them onto the sequence of a divergent reference genome (e.g. from a different species).

read-mapping: Mapping of sequence reads on to a reference when limited divergence is expected. Used to find SNPs in organisms that have a well-characterized genome sequence.

contig: A sequence of DNA formed from the assembly or continuous mapping of shorter sequence reads.

masking: Removing the putative identity of one or more positions in a sequencing read, contig or alignment. This typically involves changing the identity of such positions to 'N' or lower case, and may be done based on presence of repeated sequences, quality, coverage depth, or issues with alignment.

SNP: Single nucleotide polymorphism. DNA sequence variations that occur when a single nucleotide in the sequence of a genome is altered.

N50: The N50 is determined by sorting a set of contigs from longest to shortest and sequentially summing their lengths until the total is at least 50% of the sum of all contigs. The length of the shortest contig in this set is the N50 value for the assembly.

coverage: The proportion of a reference genome for which homologous sequence is identified among contigs assembled from a short read sequencing pool.

depth: The number of reads that support a call at a given position in a de novo contig or reference-guided assembly. This is often called "coverage", but "depth" is preferable.

cluster: A group of linked computers, working together closely so that in many respects they form a single computer.

node: A single computer in a cluster. Some nodes will be optimized for data processing, while others are optimized for data storage (file servers).

cloud: A computing cluster in which the data processing and storage are provided over the Internet.

Linux: A freely-distributable implementation of the UNIX operating system (OS) that runs on a number of different hardware platforms. The most widely used OS in bioinformatics.