Donnelly Sequencing Centre

Guide to NGS Project Planning

K. Turner 2020-04-22

This guide is written to give a comprehensive overview of planning a next-generation sequencing (NGS) experiment. It is based on the services offered at the Donnelly Sequencing Centre, but is intended to help any lab, regardless of where services are conducted. Following the written explanation is an **abbreviated list of questions** to be answered when planning an NGS project. Contact dseqcentre@utoronto.ca with questions or feedback.

Project outline

An NGS project will involve the following steps:

1. Sample collection
2. Nucleic acid extraction
3. Sample quality control
4. Sequencing library preparation
5. Library quality control
6. Sequencing
7. Data processing and bioinformatic analysis

There are many different options for how to perform each step, determined by your organism and tissue type, as well as the questions you're trying to answer.

1. Sample collection

Sample collection (e.g. dissection, cell culturing) will depend largely on your organism and setup, but there are some considerations to keep in mind if you intend to use the material for next-generation sequencing; your collection and storage method may impact the quality of your nucleic acids, particularly if you're working with RNA.

- Make sure you collect enough material. Sequencing library preparation methods have minimum inputs, so make sure you have enough tissue/cells that you'll be able to extract sufficient nucleic acids.
- Be sure to work in clean conditions. This avoids contamination with foreign genetic material as well as nucleases.
- Work to preserve your nucleic acids intact. In most cases, it's ideal to proceed immediately from sample collection to nucleic acid extraction, though that's often not possible. If not,

freeze your samples immediately and keep them frozen until extraction. Samples for RNA extraction are best-preserved at -80°C. Avoid freeze-thaw cycles as this degrades your nucleic acids.

- Certain sample preservation methods, such as formalin-fixed paraffin-embedding, yield extremely degraded nucleic acid extracts. If you are collecting FFPE samples for histology, consider collecting separate subsamples for NGS.
- Certain other buffers or preservatives which you may collect samples into (e.g. RNAlater) may be incompatible with particular extraction methods. Plan your extraction method ahead of time, and check that it's compatible with your storage method.

2. Extraction methods

Common extraction chemistries, such as those manufactured by Qiagen or Zymo, will work on many organisms and tissue types. However, different methods are tailored to different material, so here are some key points to consider when selecting and using a kit for an NGS project:

- Again, work under clean conditions to avoid contamination.
- Select the right protocol for your organism and input material. Some kits are specialized for tissue, while others are better for blood, or cells. Some tissue types, especially from non-model organisms, may require specialized chemistry. Consult the literature for your organism and tissue type for guidance on which kit to choose.
- Choose a kit for your desired output. Different protocols are optimized to extract, for example, plasmids or high-molecular weight DNA. Others are optimized to get maximum output from tiny amounts of input material. Again, the literature will be a helpful guide.
- Components of the extraction protocol may interfere with library preparation reagents downstream. Many protocols include optional cleaning steps to wash out lingering chemicals—be sure to take these steps.
- Many extraction methods will elute into a buffer containing EDTA, which keeps nucleic acids stable by interfering with nucleases. However, EDTA also interferes with many enzymes involved in library preparation, so elute in Tris or water if possible. Storing your extract properly will reduce the need for stabilizers.
- Freeze your samples (-20 for DNA, -80 for RNA) and avoid freeze thaw cycles.

If you are extracting RNA for RNA-Seq methods, be sure to include a DNase treatment step to remove contaminating genomic DNA, as this can be processed by library preparation and sequenced, affecting your results. Many extraction kits include DNase treatment; if not, ensure that a separate DNase treatment is performed on your samples. The DSC recommends the ThermoFisher DNA-free kit, and offers treatment with this kit as a service. Be sure that samples for library prep are QC'd after DNase treatment.

3. Sample quality control

Sample QC for NGS projects will ideally involve 3 components:

1. Fluorometric quantification (e.g. Qubit or Quant-iT): These methods use fluorescent dyes targeted to specific nucleic acid types (e.g. dsDNA, RNA) to give accurate concentration measures. Accurate quantification is needed to input the appropriate amount into library preparation protocols.
2. Spectrometric measurement (e.g. NanoDrop): This method checks absorbance at different wavelengths, and can indicate the presence of chemical contaminants from the original sample or from the extraction method. Generally, a 260/280 of ~1.8 for DNA or ~2.0 for RNA and a 260/230 ratio of 2.0-2.2 for either type is considered 'pure'. Spectrometry also gives a concentration estimate, but this is based on 260 nm absorbance, which reflects not just your nucleic acid of interest but all DNA, RNA, free nucleotides, or contaminants (e.g. guanidine isothiocyanate, used in some RNA extraction protocols). Thus, it is not sufficiently reliable for NGS projects.
3. Nucleic acid integrity. For DNA, this will involve a check on a gel or TapeStation; for RNA, this will involve a gel, or preferably a Bioanalyzer or TapeStation run. Most library preparation methods begin with fragmentation; if samples are already degraded, fragmentation parameters must be changed to avoid over-fragmentation, or a completely different preparation method must be chosen. For RNA, integrity is measured with a score called RIN (RNA Integrity Number), or an analogous score. A minimum score of 7 or 8 is required for most standard methods.

The DSC offers Qubit, Quant-iT, Bioanalyzer, and TapeStation services.

4. Library preparation

Library preparation refers to the process of converting nucleic acids into a form compatible with a sequencing instrument (here we're discussing methods for sequencing on Illumina instruments). The process takes 1-3 days, plus any upstream steps (e.g. in IP-Seq or Methyl-Seq). In these methods, the starting nucleic acid is typically fragmented into shorter inserts, and then several steps add tails known as *adapters* (~120 bp total) to each end of the fragments to make them compatible with sequencing.

The method chosen depends on what your targeted type of nucleic acid is, and what questions you want to answer. Here's an overview of common prep types:

- DNA-Seq: In these methods, large DNA fragments are chopped up into shorter libraries for sequencing. These methods may be used for determining the sequence of whole genomes, plasmids, or long amplicons (greater than ~600 bp). There is no target selection, so what is in the extract pool is what will be sequenced. Common DNA-Seq protocols include

Nextera Flex and Nextera XT, which use transposase enzymes; NEBNext Ultra II FS, which uses enzymatic fragmentation; or NEBNext Ultra II, which would be paired with a mechanical fragmentation method such as Covaris. The DSC offers any of the listed DNA-Seq methods.

- IP-Seq: In the upstream steps for these methods, DNA- or RNA-interacting proteins are immunoprecipitated using specific antibodies, pulling down their associated nucleic acid fragments. The resulting pool of fragments is then prepared for sequencing in a very similar way to DNA-Seq or RNA-Seq, with the exception that the nucleic acids are already fragmented. These methods are used to determine which regions of a genome or transcriptome interact with the protein of interest. Such methods include ChIP-Seq, RIP-Seq, CLIP-Seq, and other variations. For projects at the DSC, immunoprecipitation steps are completed by the user, and QC and downstream library prep on the resulting fragments are completed at the DSC.
- Methyl-Seq: These methods are used to identify the sites of methylation throughout the genome. Illumina sequencing methods cannot directly identify 5-methylcytosine, and so instead, bisulfite conversion alters unmethylated cytosine to uracil. Then, the resulting sequencing data is compared with untreated genomic data: All C sites in the treated samples were methylated, and sites that are T in the treated data but C in the untreated data represent unmethylated C. The DSC offers reduced-representation bisulfite sequencing.
- Amplicon sequencing: In this method, only a specific region is targeted for sequencing, using specific primers in PCR. This method is a good substitute for researchers who have previously used Sanger sequencing, but has the advantage of allowing hundreds of samples to be sequenced at once. A common application of amplicon sequencing is amplifying the 16S region for bacterial identification, or ITS region for fungal identification, but it can be used in any application where researchers want to focus on only one or a handful of regions, rather than a whole genome. Amplicon sequencing involves two PCR amplifications: the first uses primers that target the region of interest, while also adding a known 'flagging' sequence to the amplicon; the second reaction then targets the flagging sequence and adds sequencing adapters. The first reaction's primers will be custom to each project, while the second reaction uses a primer type common to all projects. The DSC can perform the second reaction, or both reactions if users submit primer designs.
- RNA-Seq: In this family of methods, RNA is reverse-transcribed to cDNA, which is then used to generate a sequencing library in a way very similar to DNA-Seq. The most common application of RNA-Seq is *expression analysis*: very short reads (50-75 bp) are used to simply count how many transcripts originated from a given region, thus giving a measure of how gene expression differs between organisms, treatments, etc. When the full structure of transcripts is needed, e.g. transcriptome assembly or splicing/fusion analysis, modified RNA-Seq may be used, using longer reads and altering library prep to generate

longer inserts accordingly. The DSC offers this modified method.
In most cases, over 95% of your RNA extract will consist of ribosomal RNA, which means the vast majority of reads will be uninformative. Thus, there are several methods to selectively prepare only non-rRNA libraries:

- **ribo-depletion RNA-Seq:** This type of method uses probes targeted to known rRNA sequence, which then allow the rRNA to be pulled down with magnetic beads or degraded with enzymes. This eliminates rRNA and leaves all other RNA species (e.g. mRNA, tRNA, lncRNA, pseudogenes, etc), though short species will often still be eliminated by downstream steps. Because this method relies on known rRNA sequence, species-specific depletion kits must be used, though probes are often at least partially effective across wide phylogenetic distances.
- **mRNA-Seq:** This method uses oligo-d(T) probes to specifically bind to poly-A tails on mRNA and pull them down. This means mRNA is sequenced while all other species are excluded, including pre-mRNA in steps prior to polyadenylation. In some sample types (e.g. bacteria, plant mitochondria, chloroplasts), polyadenylation is transient and marks RNA for degradation; mRNA-Seq is not applicable to these types of samples.
  A specialized family of mRNA-Seq methods known as **3' mRNA-Seq** specifically targets the fragment containing the poly-A tail, meaning that only one fragment is sequenced per transcript. This can be an option to consider for count-only applications, as it reduces the amount of sequencing required, and can sometimes yield more accurate measures of expression.
- **small RNA-Seq:** Because small fragments are often lost in standard rRNA-depletion RNA-Seq or mRNA-Seq, researchers specifically interested in small RNA species should use specialized protocols, such as NEBNext Small RNA or TruSeq Small RNA.

Common RNA-Seq protocols include NEBNext Ultra II Stranded RNA-Seq or TruSeq Stranded. *Directional* or *stranded* protocols ensure that only the strand of cDNA corresponding to the original RNA is sequenced. This leads to a more accurate assessment of expression. The DSC no longer offers non-stranded RNA-Seq protocols.

- **Exome and other targeted sequencing:** This diverse family of preparation methods uses probes or primers designed against known sequences to specifically target sequences of interest in either a DNA-Seq or RNA-Seq project. This is conceptually similar to performing amplicon sequencing, but instead can target large regions on the order of tens of megabases. A common application is targeting the exome, but this may be used in any situation when researchers are interested in a specific known subset of the genome or transcriptome.

Within a given type of library prep, there are often a wide variety of available preparation methods. To assess your choices, consider data quality, price, and input requirements. The DSC

uses only library preparation methods known to produce high-quality data. Consult the literature for comparisons of these and other methods. Contact us for current pricing.

See the table below for input requirements of common library preparation methods currently offered at the DSC. Maximizing the amount of starting material for library prep will allow you to minimize the number of PCR cycles required during prep, reducing bias and error, so optimize your sample collection and extraction protocols. Methods that tolerate lower minimum inputs are typically more expensive.

Table I.

Preparation type		minimum ng (minimum ng/uL)	optimum ng (minimum ng/ul)
DNA-Seq			
	NEBNext Ultra II DNA-Seq	0.5 ng (0.01 ng/uL)	1000 ng (20 ng/uL)
	NEBNext Ultra II FS DNA-Seq	0.1 ng (0.004 ng/uL)	500 ng (20 ng/uL)
	Nextera Flex	1 ng (0.04 ng/uL)	500 ng (17 ng/uL)
	Nextera XT	1 ng (0.2 ng/uL)	1 ng (0.2 ng/uL)
ChIP-Seq			
	NEBNext Ultra II DNA-Seq	0.5 ng (0.01 ng/uL)	20 ng (0.4 ng/uL)
	TruSeq ChIP-Seq	5 ng (0.1 ng/uL)	10 ng (0.2 ng/uL)
Reduced-Representation Bisulfite Sequencing			
	Ovation RRBS Methyl-Seq	100 ng (12 ng/uL)	100 ng (12 ng/uL)
Amplicon sequencing (including 16S/ITS)			
	Illumina amplicon sequencing	1 ng (0.2 ng/uL)	5 ng (1 ng/uL)
rRNA-depletion RNA-Seq			
	NEBNext Ultra II Stranded RNA-Seq	5 ng (0.5 ng/uL)	1000 ng (84 ng/uL)
	TruSeq Stranded RNA-Seq	100 ng (12 ng/uL)	1000 ng (118 ng/uL)
	NEBNext Ultra II Stranded RNA-Seq, modified	2500 ng (208 ng/uL)	2500 ng (208 ng/uL)
	TruSeq Stranded RNA-Seq, modified	2500 ng (295 ng/uL)	2500 ng (295 ng/uL)
	NEBNext Ultra II Stranded RNA-Seq, FFPE samples	10 ng (0.9 ng/uL)	1000 ng (84 ng/uL)
mRNA-Seq			
	NEBNext Ultra II Stranded RNA-Seq	10 ng (0.2 ng/uL)	1000 ng (20 ng/uL)
	TruSeq Stranded RNA-Seq	100 ng (2 ng/uL)	1000 ng (20 ng/uL)
	NEBNext Ultra II Stranded RNA-Seq, modified	2500 ng (50 ng/uL)	2500 ng (50 ng/uL)
	TruSeq Stranded RNA-Seq, modified	2500 ng (50 ng/uL)	2500 ng (50 ng/uL)
	NEBNext Single-Cell/Low Input RNA-Seq	2 **pg** (0.0003 ng/uL)	200 ng (29 ng/uL)
Small RNA-Seq			
	NEBNext Small RNA-Seq	100 ng (17 ng/uL)	1000 ng (167 ng/uL)
Exome sequencing			
	Nextera Exome	50 ng (5 ng/uL)	50 ng (5 ng/uL)

One other major consideration when planning a library prep protocol is **indexing strategy**. In almost all cases, you will be including multiple samples on one sequencing run, and so the indexing step of a library prep adds a specific 6-10 bp string of bases to one or both ends of the library to

uniquely identify each library. Be sure that unique indices are assigned to each sample, and that there is enough distance between each index that one or two mutations or sequencing errors will not yield a different index from the same pool. A chart of sample IDs and their indices will be required in order to demultiplex (assign reads to the appropriate sample) after sequencing; it may be required before running the instrument, so be sure to have it prepared.

Sequencing also has a small risk of *index hopping*, where the index from one sample is assigned to a different sample. Certain sequencers are more susceptible than others. To avoid this risk, it is recommended to use *unique dual indices*, where each sample has a unique index on both ends—if a swap occurs, this will be identified as an unexpected pair and the read will be thrown out. The DSC will determine indexing strategy for libraries prepared in-house, and uses unique dual indexing in all cases where it's possible.

Certain library prep methods, such as single-cell RNA-Seq, may also include *unique molecular identifiers*, which give a unique molecular code to each library molecule pre-amplification, so that actual differences in fragment representation (expression, in the case of RNA-Seq) can be distinguished from random differences generated in PCR.

5. Library quality control

NGS library quality control will involve 2 components:

1. Quantification: Quantifying libraries is important for checking whether you've generated sufficient material for sequencing, and for pooling multiple samples at the desired proportions for sequencing. qPCR provides the most direct measure of abundance of libraries, but can be quite expensive and laborious for large projects. Some libraries, particularly those with strong base bias, can also amplify poorly with some qPCR kits. Instead, fluorometric quantification, as in sample QC, often provides sufficient accuracy.
2. Size distribution: The Bioanalyzer or TapeStation (or a gel, in a pinch) can be used to examine the size distribution of finished libraries. You're looking for two things:
    a. Is the size of the library what I expected? Most library prep protocols will indicate an expected size distribution. For custom methods, this will be the length of the insert plus the length of the adapters. A spread of a few hundred bp is not abnormal, though the mode should be close to expectation.
    b. Are there adapter or primer dimers in the finished libraries? These ligation and amplification artefacts will show up as distinct peaks somewhere between 35 and 170 bp, depending on prep method. These products can strongly compete for clusters on the sequencing instrument and lower data quality, and so if present they **must be removed**, by gel excision or SPRI bead cleanup.

Together, fluorescent quantification and sizing can be used to calculated library molarity:

nM = ([ng/uL] * 10^6) / ([avg length in bp] * 660)

Molarity is then used to **pool** libraries at the appropriate proportions to determine the amount of sequencing depth each will receive. Once pooled, it is advisable to measure the final pool using qPCR (if possible), in order to avoid over- or underloading the sequencer. For more on both of these points, see below.

The DSC offers Qubit/Quant-iT, Bioanalyzer/TapeStation, and qPCR, as well as bead cleanup.

6. Sequencing

There are two main parameters to determine for planning sequencing: **length** and **depth**. These will, along with project scale, determine which sequencing instrument to use. **Other considerations** affect the success of a sequencing run.

Length depends on application. Most projects that are counting-focused, like expression analysis or IP-Seq, will only require short reads. Projects that seek to obtain new information about the sequence or structure of transcripts or genetic regions, particularly *de novo* assembly, will require long reads. Analyses for amplicon applications like 16S or ITS barcoding will often also require longer reads.

Sequencing kits are available in a variety of lengths (expressed as "cycle number"), from 50 bp to 600 bp on Illumina instruments, though not all lengths are available on all instruments. In most cases, a kit can be run single-end (e.g. "1x100") or paired-end ("2x50")—the price will be the same, as the same reagents are used. Single-end mode is often sufficient for short count-based methods, while for more assembly-focused projects, paired-end yields slightly more information: you get the sequence at each end of the insert, plus the knowledge that those pairs are within a certain distance of each other.

Reads do not have to be the full length of the kit (e.g. you can run a 100-cycle kit for only 90 bp) and paired reads do not have to be even (e.g. a 40+60 bp run is possible). These can be useful for custom library designs; the total length must just be less than the total length of the kit. There are also limitations in the length of individual reads, as quality progressively falls off: 2x300 is possible, but 1x600 is not viable. All samples on one sequencing run must be sequenced via the same format (i.e. you cannot have half the samples running 1x75 and half 2x50), so if you're combining multiple project types, think of a sequencing format that works for all of them.

Whatever your desired read length, ensure it makes sense based on library length. If your insert is 100 bp (which would appear on the Bioanalyzer as a library around 220 bp), 2x150 bp reads will be wasted.

Depth is counted in *clusters,* referring to the number of spots on a sequencing flowcell where clonally-amplified libraries are sequenced by synthesis. You will often hear this referred to as *reads*, but we use "clusters" to avoid ambiguity: when a paper says "100 million paired-end 50 bp reads", sometimes they mean 100 million pairs of 50-bp reads, while other times they mean 100

million individual 50-bp reads, in 50 million pairs. Since both parts of a paired-end read take place within one individual cluster, from one original library molecule, we refer to both one million single-end reads or one million pairs of paired-end reads as one million clusters.

The number of clusters needed is largely a sampling issue. In RNA-Seq, shallower sequencing will suffice for highly-expressed genes, but deeper sequencing is required to accurately sample less-abundant genes, and to assemble novel transcripts. Similarly, in DNA-Seq, deeper sequencing ensures that each position in the genome is sequenced to a desired coverage. Greater coverage will ensure accurate calling of SNVs, indels, etc while distinguishing them from sequencing error.

Larger genomes (and larger transcriptomes) obviously require more sequencing to accomplish the same coverage. Required read depth is calculated from the Lander/Waterman equation:

clusters = ([desired coverage] * [assembly size])/[read length (bp)]

Here, assembly size is the size of the targeted genome, exome, IP-targeted regions, or transcriptome (though depth for transcriptome projects is typically determined by standards for a given application). Remember that read length includes both reads in paired-end reads (e.g. 2x150 is a total of 300 bp in length).

The most rigorous way to determine the appropriate depth for a sequencing project is to run a pilot study with a small number of representative samples subjected to fairly deep sequencing. Then, the data can be bioinformatically down-sampled (i.e. randomly selecting 50%, 25%, 10%) and re-analyzed to see how much you could reduce depth without changing the results.

However, it is often possible to go with existing standards from the literature. The table below summarizes typical read lengths and depths for a number of applications, from manufacturer and literature recommendations, as well as typical submissions to the DSC.

Table II.

Application	Typical read lengths	Recommended depth per sample (coverage in x or depth in million clusters)
DNA-Seq for CNV detection	2x300, 2x250, 2x150 [1]	1-8x [2]
DNA-Seq for genotyping/SNV calling	2x300, 2x250, 2x150 [1]	35x [2]
DNA-Seq for indel calling	2x300, 2x250, 2x150 [1]	60x [2]
DNA-Seq for *de novo* assembly	2x300, 2x250	100x
ChIP-Seq	1x50, 1x75, 1x100, 2x50	100x [3]
Reduced-representation bisulfite sequencing	1x50, 1x75, 1x100, 2x50	10x [2]
Amplicon sequencing: 16S rRNA (V3+V4) for bacterial metagenomics	2x300, 2x250 [4]	0.001-0.05 M (dependent on expected diversity) [5]
Amplicon sequencing: ITS for fungal metagenomics	2x300, 2x250 [6]	0.001-0.05 M (dependent on expected diversity) [5]

Amplicon sequencing: other	Amplicon length-dependent	0.001-0.03 M (dependent on expected variation)
RNA-Seq for expression analysis	1x50, 1x75, 1x100, 2x50 [1]	10-25 M [2]
RNA-Seq for allele-specific expression analysis	1x50, 1x75, 1x100, 2x50 [1]	50-100 M [2]
RNA-Seq for alternative splicing, fusion (modified prep protocol)	2x150	50-100 M [2]
RNA-Seq for *de novo* transcriptome assembly (modified prep protocol)	2x150	>100 M [2]
Small RNA-Seq for expression analysis	1x50, 1x75 [1]	1-2 M [2]
Small RNA-Seq for discovery	1x50, 1x75, 1x100, 2x50 [1]	5-8 M [2]
Exome sequencing	2x150 [1]	100x [3]

1- https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/read-length.html
2- https://genohub.com/recommended-sequencing-coverage-by-application/
3- https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/coverage.html
4- https://support.illumina.com/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf
5- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4837688/
6- https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/metagenomic/fungal-metagenomic-demonstrated-protocol-1000000064940-01.pdf

Together, length and depth determine which sequencing instrument and flowcell are appropriate. Each flowcell type has a typical output of a certain number of clusters, with actual output varying about ± 25%. A given flowcell is in turn only compatible with certain sequencing length kits. Here is the chart of possible sequencing modes for sequencers available at the DSC:

Table III.

instrument	MiSeq				NextSeq		NovaSeq			
flowcell >	Nano v2	Micro v2	v2	v3	Mid-Output	High-Output	SP	S1	S2	S4
clusters >	1 M	4 M	15 M	25 M	130 M	400 M	800 M	1.6 B	4.1 B	10 B
50 cycles			X							
75 cycles						X				
100 cycles							X	X	X	
150 cycles				X	X	X				
200 cycles							X	X	X	X
300 cycles	X	X	X		X	X	X	X	X	X
500 cycles	X		X				X			
600 cycles				X						

When selecting an appropriate instrument and flowcell, choose the instrument whose output is closest to the total number of clusters required for your project, while also ensuring that it supports a cycle number close to what you desire. Longer-read kits (greater cycle numbers) are more expensive, and larger flowcells (more clusters) are more expensive, though the cost per cluster is significantly lower on larger flowcells. Thus, it's most cost-effective to use larger flowcells; some sequencing centres (including the DSC) will allow users to split large flowcells with other projects in order to take advantage of these savings. However, index sequences between the projects must not overlap, so this requires planning in advance. Further, requesting part of a flowcell can increase project turnaround time as you wait for a compatible submission to "fill" the remainder of the flowcell.

The amount of output each sample in a sequencing run receives is determined by how the samples are pooled before loading on the sequencer:

1. If Sample A is put on a NextSeq High-Output flowcell alone, it would be expected to output 400 M clusters.
2. If equal moles (based on the molarity measured by qPCR or calculated as described in section 5) of Sample A, B, C, and D are loaded, then each would be expected to output 100 M clusters.
3. If 2x moles of sample A, x moles of sample B, and x moles of sample C are loaded, then you would expect 200 M clusters for A, and 100 each for samples B and C.

The *absolute* concentration of each sample in the pool does not determine the sequencing output—this is determined by the *relative* concentration, where the proportions of moles of each library in the pool determine the split of reads from the total output of the flowcell.

However, absolute concentration of the pool is important for whether sequencing can proceed: typically, **at least 2-4 nM** is required before the pre-loading steps of sequencing. Typical absolute minimum volumes required (at minimum concentration) are as follows:

- MiSeq: 5 uL
- NextSeq: 10 uL
- NovaSeq SP/S1: 100 uL
- NovaSeq S2: 150 uL
- NovaSeq S4: 310 uL

As with library prep, it is best to have far above the minimum volume required, for quality control, backup, and loading flexibility. If working with a sequencing centre, don't accidentally over-dilute: submit at the highest concentration available and let them quantify and dilute.

Accurate quantification of the final pool to be loaded onto the sequencer is essential. It's best to use qPCR, as long as the library is compatible. Accurate quantification ensures that the right concentration is loaded onto the sequencer. Loading below the recommended concentration leads to "underclustering", reducing the total output of the sequencing run, and leading to run

failure in extreme cases. Overloading leads to "overclustering", where clusters are packed too close together. This can lead to run failure, a large number of clusters being excluded from quality filtering (low pass-filter %), and low-quality base calls (low Q30).

Sequencing performance is also affected by the diversity of the sample. Problems will arise if a large proportion of libraries have the same base at a given position in the first several bases of the first read, particularly the first 4-7, when clusters are identified and distinguished. The sequencer will fail to correctly distinguish different clusters, leading to many clusters being excluded from quality filtering. Low diversity typically arises in the following cases:

1. Amplicon sequencing, when the majority of the entire library is identical.
2. When the organism sampled has a strong base bias (high or low GC content).
3. When the library prep method generates fragments starting with the same sequence, e.g. RAD-Seq or CRISPR screening.

Several strategies can be employed to ameliorate low diversity:

1. Mixing the problematic library type with a high-diversity library (e.g. RNA-Seq).
2. Lowering the loading concentration of the pool to intentionally undercluster.
3. Using an instrument that uses four colour channels to image libraries (e.g. MiSeq), as opposed to two channels (e.g. NextSeq, NovaSeq).
4. Using 'dark cycles' at the start of a run, which are cycles of synthesis occuring through the low-diversity region that are not imaged. Thus cluster identification is delayed until a higher-diversity region.
5. Increasing the percentage of PhiX spike-in on a run. PhiX is a pre-generated viral library included in most runs as a positive control of known sequence, to score accuracy of a run. It is typically spiked in at 0.5-2% representation, but since it's high-diversity, it can be included at higher percentages to compensate for low-diversity libraries: >5% on MiSeq or NovaSeq, and 10-50% on NextSeq*. Note that PhiX is not available if custom sequencing primers are used instead of Illumina defaults.

   *https://support.illumina.com/bulletins/2017/02/how-much-phix-spike-in-is-recommended-when-sequencing-low-divers.html

7. Data processing and bioinformatic analysis

Data from an Illumina run can either be uploaded to Illumina's BaseSpace cloud platform, or manually processed and distributed. Initial processing typically involves trimming adapter sequences, demultiplexing (using index sequences to assign reads to samples), and compressing the resulting .fastq files for distibution. These processing steps are available at the DSC.

There are numerous pipelines for downstream analysis, and several GUI-based tools (including BaseSpace) that can be used for common analyses. Consult the literature and bioinformatics

experts for the latest recommendations, and ensure that your sequencing plan and any initial processing steps are compatible with the intended analysis method.

Sequencing runs also output a number of quality metrics that allow you to assess the run and re-adjust for subsequent runs. Here are some key metrics given by the Sequencing Analysis Viewer for the entire run and per lane:

1. Density: The number of clusters per area. Optimal cluster density is 1000-1200 K/mm^2 for MiSeq v2, 1200-1400 for MiSeq v3, and 170-220 for NextSeq. Being under or over this range will indicate under- or overclustering, respectively. The NovaSeq uses patterned flowcells and thus has a fixed density.
2. Clusters PF (%): The percentage of clusters which "passed filter", i.e. have a pure signal (not multiple clusters mixed together). This is measured from the ratio of strongest to second-strongest signal at each base in the first 20-25 cycles: if more than one base fails, the cluster does not pass. Above 80% is typical for MiSeq or NextSeq; lower values often point to overclustering. Clusters PF is calculated differently for patterned flowcells, and so somewhat lower scores (e.g. 70%) are acceptable on NovaSeq.
3. Phasing/prephasing: What % of the reads appear to have jumped ahead or fallen behind in cycles, contributing to noise in the data. Optimal values are below 0.2-0.5%, depending on platform. High values may indicate base bias (high/low GC%) or issues with reagents or the flowcell.
4. % >= Q30: The percentage of reads that have a Q-score of 30, i.e. an error proportion of 0.001 or below. The expected percentage is dependent on chemistry and read length, but is typically in the 70-90% range. See Illumina's specifications for each sequencer and run mode. Q30 can also be viewed on a per-cycle basis. It typically falls off towards the end, though gradually, not abruptly. Low overall Q30 % or a strange pattern in Q30 across cycles may indicate issues with library preparation or reagents.
5. Aligned (%): The percentage of reads that aligned to the known genome of the PhiX spike-in. This should match the percentage of PhiX loaded. If aligned % is low, this indicates that other libraries were overloaded; if it's high, other libraries were underloaded.
6. Error Rate (%): The accuracy of PhiX reads aligned to its known genome, used as a proxy for error rate in samples. This is broken down across different portions of the read length, and can be viewed on a per-cycle basis. It typically increases along the length of a read. Aligned (%) and Error Rate (%) are not available if PhiX is not loaded or custom sequencing primers are used instead of default Illumina primers.
7. Other output of the Sequencing Analysis Viewer can help diagnose exactly what issues may have arisen.

The demultiplexing quality file will additionally break down several metrics, e.g. % >= Q30, for each individual sample. It will additionally give a count of PF Clusters per sample and % of the lane occupied, which will indicate how accurately samples were pooled, and how each individual library performed. This can be used for a conservative sequencing strategy where libraries are pooled

and sequenced at low depth, then re-pooled to correct for sequencing performance before a full sequencing run.

The "Undetermined" bin indicates reads that could not be assigned to any of the indices given in the sample sheet used by the demultiplexing pipeline. Completely missing or very low PF Clusters counts for some samples, along with high counts in the Undetermined bin, may indicate that the wrong index was listed for those samples. Thankfully, it is possible to simply correct the index list on the sample sheet and re-run the demultiplexing pipeline on the existing data.

To assess quality on a per-read basis, .fastq files will provide not only the bases in a sequence, but a quality score (a Q-score, encoded as an ASCII character as listed here: https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/QualityScoreEncoding_swBS.htm). Higher scores indicate lower error probabilities. This score will typically be used by analysis pipelines to exclude low-quality portions of reads.

NGS planning questions

1. How am I collecting and preserving my samples? Is it NGS-friendly (i.e. avoids contamination and keeps nucleic acids intact)? Is it compatible with the extraction method I'm planning on using?
2. What is my extraction method? Is it optimized for my organism, tissue type, and desired extract? Does it yield enough for my desired library prep method?
3. Is my extraction optimized to avoid interfering with library prep chemistry, by including all wash steps and avoiding EDTA in elution buffer?
4. For RNA-Seq, am I including a DNase treatment step, or will I be outsourcing that? The DSC offers DNase treatment service.
5. Do I have access to quality control instruments: fluorometric quantification (e.g. Qubit/Quant-iT), fragment sizing (e.g. gel/Bioanalyzer/TapeStation), and spectrophotometric reading for contamination (e.g. NanoDrop?) The DSC provides fluorometric quantification and Bioanalyzer/TapeStation fragment sizing on submissions by default.
6. What library prep type am I planning to use? Which manufacturer and method? See Table I for methods supported at the DSC. The DSC can choose which manufacturer and method is best for your application.
7. Does my desired library prep method include upstream processing or selection steps? The DSC offers poly-A selection or ribo-depletion for RNA-Seq, and bisulfite processing for RRBS workflows. Immunoprecipitation for ChIP/RIP/etc must be performed by customers.
8. Do my samples meet the quality and quantity minimums of my desired library prep method? See Section 3 and Table I.

9. How will my samples be indexed? If I'm indexing myself, have I coordinated with any other projects that will sequence along with mine? For library preps done at the DSC, staff will coordinate indexing, using unique dual indexing when available.
10. Do I have access to quality control instruments (as in question 5) to assess whether my libraries are high quality and sufficient concentration? The DSC will QC all libraries prepared in-house, as well as customer-prepared libraries submitted individually. Customers may also QC their individual libraries themselves and submit them pooled.
11. Have I eliminated any primer or adapter dimers (peaks 35-170 bp on Bioanalyzer/TapeStation traces) in my libraries, which will interfere with sequencing? The DSC will by default use bead purification to remove these from libraries.
12. What is my desired sequencing depth per sample? See Table II.
13. What are the sequencing lengths I could use for my samples? See Table II.
14. What instrument and flowcell type would be appropriate for my project? See Table III. The DSC can choose which sequencing parameters are appropriate for your needs.
15. Do I have sufficient material for my desired sequencer? Am I able to pool enough of each sample to yield my desired distribution of sequencing output (clusters)? See Section 6.
16. Are my libraries low-diversity? What is my mitigation strategy? See Section 6. The DSC can advise on the best mitigation strategy, if alerted.
17. Have any past sequencing runs from a similar sample and library type had issues with over- or underclustering? Alert the DSC if this is the case and we will adjust sequencing accordingly.
18. What is my planned bioinformatic analysis strategy? Are my sequencing parameters and the initial bioinformatic processing steps compatible with this plan? By default, the DSC conducts adapter trimming and demultiplexing, followed by tar compression. Let us know if this is not desired.

Thank you for reading the DSC's Guide to NGS Project Planning. Please contact dseqcentre@utoronto.ca for additional questions or advice. We offer free project consultations in person or remotely. We're also available for guest lectures and workshops.

Best of luck with your NGS project!