MiSeq: Imaging and Base Calling

Page	Narration
Welcome	Welcome to MiSeq: Imaging and Base Calling. This course takes 35 minutes to complete. Click Next to continue.
Navigation	Please take a moment to familiarize yourself with the navigation for this course. You can access these instructions by clicking Help in the top right corner of the page.
Presenter Introduction	This course is presented by Jeremy Peirce, Staff Field Applications Scientist for Illumina.
MiSeq Sequencing Workflow	Sequencing on the MiSeq, as with other Illumina Sequencers, begins with sample preparation. A DNA library is prepared with adapters A and B, one on either end, of the insert to be sequenced. Following this, clusters are formed and a flow cell is placed on the MiSeq. Images are then taken using the MiSeq image control software and images are processed to extract intensities and then base calls. Following extraction of base calls, the MiSeq provides additional data analysis tools using the MiSeq Data Reporter software.
MiSeq Sequencing Workflow	Again, the MiSeq Sequencing workflow begins with image processing and intensity extraction from those images followed by base calling based on the intensities. The latter two pieces of this process are controlled by RTA (Real Time Analysis) software. Which is a program launched automatically by the MiSeq Control Software. RTA has two functions analysis of images and calling of bases. The input to image analysis is of course the tiff images and the outputs are intensities and cluster positions encoded in cluster intensity files and compressed location files. Base callings input are the compressed intensity files and the outputs are base call files which are a binary representation of base calls and quality scores.
Image Analysis	Image Analysis
Images Generated on the Instrument	To begin with images are generated on the instrument. MCS the MiSeq Control Software controls image generation for the MiSeq and images are generated with four images per tile per cycle. One cycle includes the chemical addition and imaging of one base for each cluster on the flow cell. For imaging purposes the MiSeq flow cell is broken-up into areas or tiles. For each tile in each cycle an image is taken for every base.

Image Analysis Input and Output	Image analysis inputs are the tiff images and the RunInfo.xml, which contains information about the run. These are processed through image analysis to extract intensities and locations of clusters.
Clusters are Bright Spots in the Image	Clusters are bright spots in the image. They represent approximately a thousand copies of the same DNA strand in a one micron spot. Each image then represents the florescence in one of four channels representing the four possible nucleotides; G, A, T or C. Spots that produce signal in the C channel, for instance, which represents the C nucleotides in a given cycle will also have some signal in the A channel due to a phenomenon known as spectral overlap. As can be seen in the diagram at right, the A and the C channels emissions spectral are somewhat overlapping. So that an image taken in the A channel will have some signal from the C channel and vice versa.
Multi-Cycle Detection of Cluster Positions	We use multi-cycle detection to differentiate cluster positions over a series of 4 cycles, the first four cycles imaged on the MiSeq. Multi-cycle cluster detection improves our positioning and resolution of clusters. In the example here, it is difficult to resolve the cluster pictured in one cycle only because when there are overlapping clusters with the same base the cluster will simply look more like an oval cluster rather than two clusters with the same base. However, if we use multiple cycles it is easy to see the difference between the two clusters when they do not have the same base in the same cycle. By detecting clusters positions in multiple cycles then, there is a better chance of resolving overlapping clusters. Since they are less likely to be of the same base over the four cycles used to detect the cluster positions.
Cluster Detection Algorithm	The cluster detection algorithm looks for the maximum intensity area of the cluster and also takes a measure of the threshold background noise level on the flow cell. When there a two clusters close to each other integrating over the entire area of the cluster, it would make it difficult to distinguish between the two clusters. So we take a thinner slice from the center of the maximum intensity region of the cluster. We use sub-pixel interpolation to generate the position of the best signal for the cluster. In this case, the interpolated position for best signal is between 77 and 78 pixels. If we were to choose 77 or 78 pixels, rather than interpolating between them we might not get the best intensity for that cluster. Accounting for sub-pixel interpolation results in more detected clusters and more accurate error rates.
Cluster Detection Algorithm	After all clusters are detected in this way the cluster intensities are extracted and stored.

Intensity Extraction Algorithm	Viewed another way the intensity extraction algorithm defines an integration area. As a top view, you can see this integration area is essentially at the center of a given cluster in the x and y. The intensity is then extracted from this area for all four bases from the four noted images. In this case A has a much higher intensity than the three other possibilities. Note that C is a little higher than G or T because the A and C emissions spectra are somewhat closer so there is some overlap in the emission spectra. When the real base is A, there will be some additional background in C.
Process of Extracting Intensity	The process of extracting intensities, first computes the background for each cluster and signal for each cluster. Then subtracts the background from the signal for each cluster. As noted, we are collecting background from a distance outside the cluster and signal is collected from the very center; the most intense region of the cluster.
Base Calling	Base Calling
Base Calling Input and Output	The inputs to base calling are the compressed intensity files and the compressed location files. After intensity is extracted, it is put through intensity normalization and cross talk estimation. To account for differences in dye deficiencies and the differences in the amount of cross talk between bases. These will be discussed in a moment. Bases are then called and quality scores are determined and the results are stored in fastq files as the output of base calling.
Base Calling Important Caveat	There is an important caveat to base calling on the MiSeq; the intensity correction process requires a reasonable base composition to work well. This balance does not have to be perfect and indeed we have sequenced both Rhodobacter at 69% GC and B. Cereus at 36% GC and both give good results. Intensity correction is calculated across the first 12 base pairs and since the MiSeq does not have a separate control lane samples without base composition balance in the first 12 bases may be challenging. These samples could be run on the HiSeq or GA which have the ability to use control lanes to make up for this requirement. However, it is important to notice that we have demonstrated between 36 and 69% GC balanced given good results and it probable that the good results are available outside this range.
Intensity Normalization	Intensity normalization takes the raw peak intensities and normalizes them for all four bases; G, A, T, and C. So that they have a common mean to reduce the effects from different dye channels efficiencies. At the left, there is a simplified example for two nucleotides that initially have different intensities per molecule. Post normalization these intensities would be the same.

Cross Talk Estimation	In addition, we estimate cross talk values. Cross talk values are the values for signal overlap between the emissions spectra between the dyes used. The MiSeq uses two wavelengths to excite the fluors but reads four emission spectra, one for each base. For overlapping emissions spectra, it is possible to define and remove overlaps in signal to achieve to purer read.
Graphical View of Cross Talk Correction	As a graphical view of cross talk correction, on the left, you can see that signal in one channel pre-cross talk correction effects signal in the other channel. On the right hand side, this cross talk has been eliminated and signal in one channel is independent of signal in the second channel. So the channels themselves are parallel to one or the other axes.
Background Noise Correction	We also improved signal purity by correcting for the small number of strands at each cycle that do not add the base pair expected. In this case, within a single cluster of about a thousand strands the vast majority will have added C as expected. However, a very small number of cycle of strands will be phase—that is say they will be one base pair behind the current base pair addition. And an additional small number will be prephased—that is to say they will be at least one base pair ahead of the expected position. Correction for phasing and prephasing requires unbiased base content of template that is used for the calculation as discussed previously. It is worth noting that the last base of a read cannot be corrected for prephasing and so typically has a slightly increased amount of noise relative to previous cycle.
Base Calling	After phasing, prephasing, normalization and cross talk correction are taken into account, base calling is performed by looking at the highest intensity signal. The base then with the highest intensity is the one called except for base positions where all bases are very low intensity—where we may call no base. In this case, the C intensity is much higher than the A, G, T and the base called is C.
Quality Filtering of Clusters	Once the first 25 bases have been called, the CHASTITY filter is calculated for each cluster over the first 25 bases of sequence. CHASTITY is a rough measure of signal purity over the 25 bases of sequence. It is the ratio of the highest intensity to the sum of the highest and second highest intensities. As shown in the diagram, the highest intensities is C and the second highest is intensity is G. These are noted as I-a and I-b. CHASTITY, C, is the ratio of I-a over the sum of I-a plus I-b.

Quality Filtering of Clusters	For perfect signal, the second highest signal is zero, CHASTITY is one because the formula reduces to I-a over I-a. For very poor signal where the highest and second highest signals are equal, I-a equal I-b, the equation reduces to zero point 5, one over two, or one half. To pass filter all but one of the first 25 bases of sequence for each cluster must have a CHASTITY of at least point six. This quality filter removes overlapping and low intensity clusters.
Base Calls Quality Scores	Quality scoring is then performed on base calls. Quality scoring predicts the probability of an error in base calling and is calculated using a model that matches observed values of quality predictors to a pre-calculated table. This method is very similar to Phred quality scores for Sanger sequencing. Although since the chemistry is different, the predictors of course are also different. Base call quality scores are calculated after quality filtering when CHASTITY is performed. They are generated at the beginning of cycle 25.
Base Call Quality Scores	Quality scores are usually expressed as 1 error in 10 to the q over the 10 base calls of Q quality. For instance, if were to take a base quality score of 30 which is typically accounted as good sequence quality. This would represent one error in 10 to the 30 over 10 or 10 to the 3 or a thousand base calls. This would equate to a base call accuracy of 99.9% and is often referred to as bases of Q-30 quality.
Important Quality Score Caveats	There are a number of important caveats, however to the quality score. Quality scoring is based on a model and can be very accurate when conditions used to build the quality table are similar to the data whose quality is to be predicted. At Illumina, our Q tables are created at standard cluster densities and very well characterized DNA samples and are accurate for: • Standard library types similar to genomic samples that we train our data on • They are also accurate, or most accurate at standard cluster densities and • In supported read lengths Quality score specifications for the percent Q30 apply when the above conditions are true. That is to say the library types are similar to genomic samples and the sequencing run is at standard cluster densities and at supported read lengths.

*_fastq	After generation, data is saved in a “fastq” format, which is an industry standard. This fastq format contains: • Information about the read • The sequence of the read • As well as the Q scores for each base in the read sequence—all in a convenient file format.
After Base Calling: MiSeq Reporter	After base calling is completed and fastq files are generated, the MiSeq Reporter or third-party software can be used to perform additional data analysis. The MiSeq Reporter is an integrated data analysis package that facilitates a number of workflows that are common to many next generation sequence analyses. These include: • Resequencing • Amplicon Sequencing • Library QC • smRNA Analysis • *De Novo* assembly • 16S Metagenomics analyses MiSeq Reporter operations will be covered in a separate module.
Course Complete	This concludes the presentation. Congratulations! You have completed this course.