

# 16S Metagenomics App

Introduction .....	3
Running 16S Metagenomics .....	5
16S Metagenomics Output .....	6
16S Metagenomics Methods .....	11
Technical Assistance .....	



This document and its contents are proprietary to Illumina, Inc. and its affiliates ("Illumina"), and are intended solely for the contractual use of its customer in connection with the use of the product(s) described herein and for no other purpose. This document and its contents shall not be used or distributed for any other purpose and/or otherwise communicated, disclosed, or reproduced in any way whatsoever without the prior written consent of Illumina. Illumina does not convey any license under its patent, trademark, copyright, or common-law rights nor similar rights of any third parties by this document.

The instructions in this document must be strictly and explicitly followed by qualified and properly trained personnel in order to ensure the proper and safe use of the product(s) described herein. All of the contents of this document must be fully read and understood prior to using such product(s).

FAILURE TO COMPLETELY READ AND EXPLICITLY FOLLOW ALL OF THE INSTRUCTIONS CONTAINED HEREIN MAY RESULT IN DAMAGE TO THE PRODUCT(S), INJURY TO PERSONS, INCLUDING TO USERS OR OTHERS, AND DAMAGE TO OTHER PROPERTY.

ILLUMINA DOES NOT ASSUME ANY LIABILITY ARISING OUT OF THE IMPROPER USE OF THE PRODUCT(S) DESCRIBED HEREIN (INCLUDING PARTS THEREOF OR SOFTWARE) OR ANY USE OF SUCH PRODUCT(S) OUTSIDE THE SCOPE OF THE EXPRESS WRITTEN LICENSES OR PERMISSIONS GRANTED BY ILLUMINA IN CONNECTION WITH CUSTOMER'S ACQUISITION OF SUCH PRODUCT(S).

© 2014 Illumina, Inc. All rights reserved.

**Illumina, IlluminaDx, 24sure, BaseSpace, BeadArray, BeadXpress, BlueFish, BlueFuse, BlueGnome, cBot, CSPro, CytoChip, DASL, DesignStudio, Eco, GAllx, Genetic Energy, Genome Analyzer, GenomeStudio, GoldenGate, HiScan, HiSeq, HiSeq X, Infinium, iScan, iSelect, MiSeq, MiSeqDx, NeoPrep, Nextera, NextSeq, NuPCR, SeqMonitor, Solexa, TruGenome, TruSeq, TruSight, Understand Your Genome, UYG, VeraCode, VeriSeq, the pumpkin orange color, and the Genetic Energy streaming bases design** are trademarks of Illumina, Inc. in the U.S. and/or other countries. All other names, logos, and other trademarks are the property of their respective owners.

## Introduction

The BaseSpace app 16S Metagenomics analyzes DNA from amplicon sequencing of prokaryotic 16S small subunit rRNA genes. Read classification is performed using a high performance version of the RDP Naïve Bayes taxonomic classification algorithm. Classification statistics are calculated – both by sample and in aggregate – and results are summarized in interactive visualizations. Raw classification data are also available for download.

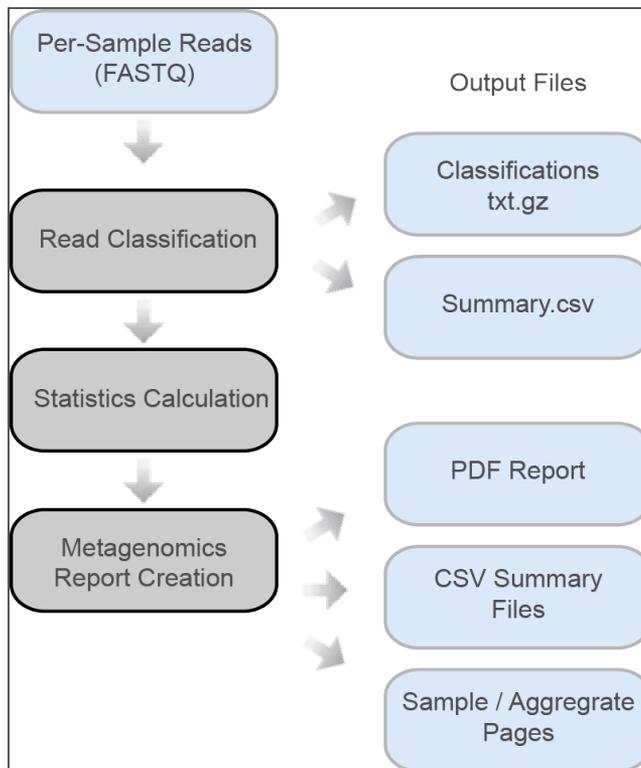
The main output files generated by 16S Metagenomics are:

- ▶ **Summary.csv** files, containing summarized classification statistics for each sample.
- ▶ **Txt.gz** compressed files, containing per-read classifications for each sample.
- ▶ **Aggregate\_Counts.csv** files, containing aggregate classification results by taxonomic level across all samples.

In addition, there are per-sample and aggregate PDF and HTML reports, and XML statistics files.

See *16S Metagenomics Methods* on page 11 and *16S Metagenomics Output* on page 6 for more information.

Figure 1 16S Metagenomics App Workflow



## Versions

The following module versions are used in the 16S Metagenomics app:

- ▶ Isis v2.5.35.6
- ▶ Illumina-curated version of May 2013 Greengenes taxonomic database

## Current Limitations

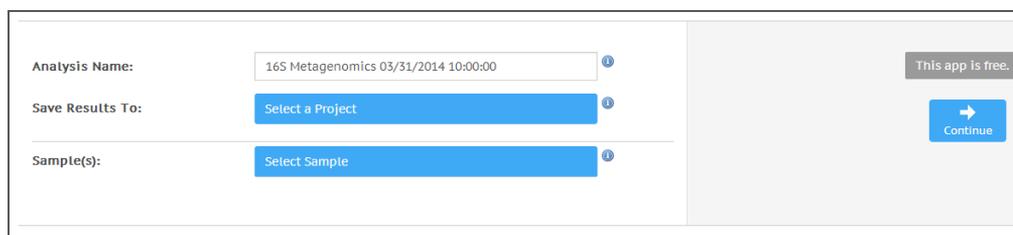
Before running the 16S Metagenomics app, be aware of the following limitations:

- ▶ Illumina-curated version of May 2013 Greengenes taxonomic database only
- ▶ Read length of at least 100 bp
- ▶ Data set size fewer than 50 gigabases
- ▶ Classification up to the species level is supported; subtype classification is not supported.

## Running 16S Metagenomics

- 1 Navigate to the project or sample you want to analyze.
- 2 Click the **Apps** button and select **16S Metagenomics** from the drop-down list.
- 3 If you see the End-User License Agreement and permissions, read them and click **Accept** if you agree.
- 4 Fill in the required fields in the 16S Metagenomics input form:
  - a **Analysis Name:** Provide the analysis name. Default name is the app name with the date and time the analysis was started.
  - b **Save Results To:** Select the project that stores the app results.
  - c **Sample(s):** Browse to the sample you want to analyze, and select the checkbox. You can analyze multiple samples.

Figure 2 16S Metagenomics Input Form



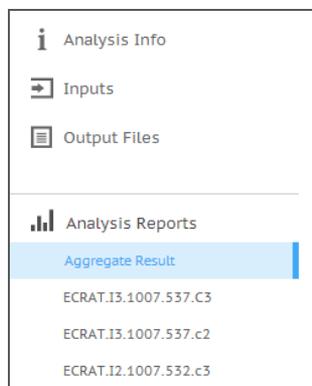
- 5 Click **Continue**

The 16S Metagenomics app now starts analyzing your samples. The status of the app session updates to show the app session progress. When completed, the status of the app session is set to *Complete*, and you receive an email.

# 16S Metagenomics Output

This chapter describes the output of the 16S Metagenomics app. To go to the results, click the **Projects** button, then the project, then the analysis.

**Figure 3** 16S Metagenomics Output Navigation Bar



When the analysis is completed, you can access your output through the left navigation bar, which provides the following:

- ▶ **Analysis Info:** an overview of the app session settings. See *Analysis Info* on page 8 for a description.
- ▶ **Output Files:** access to the output files, organized by sample. See *16S Metagenomics Output Files* on page 9 for descriptions.
- ▶ **Inputs:** overview of input settings, see *Inputs Overview* on page 9.
- ▶ **Aggregate Result:** access to analysis metrics for the aggregate results. The Aggregate Summary page is only displayed if multiple samples are analyzed. See *Aggregate Summary Page* on page 6.
- ▶ **Sample Pages:** access to analysis reports for each sample. See *Sample Summary Page* on page 7 for a description.

## Aggregate Summary Page

The 16S Metagenomics App provides a comparison of all samples on the Aggregate Summary page. You can view interactive charts showing comparisons of sample results, or download aggregate result files.

### Aggregate Results

The aggregate results section contains summary files for download. The PDF Aggregate Summary Report contains tables showing the same information as in the BaseSpace report. The CSV files contain the number of reads classified to each classification by sample, organized by taxonomic level of the classifications.

### Sample Information

The Sample Information table provides the Sample Number, Sample ID, and statistics about genus-level classification rate and primary analysis for each sample.

Statistic	Definition
Number reads PF	The total number of reads with an index sequence matching this sample that passed quality filtering
% Reads PF Classified to Genus	$(\# \text{ Reads with Genus-level classification}) / (\# \text{ Reads PF}) * 100$

## Principal Coordinate Analysis (PCoA)

Principal coordinate analysis shows similarity between normalized relative abundance of all samples. The PCoA is generated using Classical MDS on a Pearson covariance distance matrix generated from per-sample normalized classification abundance vectors. You can select the maximum level at which to compare samples by adjusting the scrollbar above the chart. Points are labeled with a shortened version of the Sample ID, you can mouse over any point to view the full Sample ID.

## Hierarchical Clustering Dendrogram

The dendrogram shows a hierarchical clustering of samples based on genus-level classifications. Mouse over the bar charts to view taxonomic labels.

## Species Diversity Results

Statistic	Definition
Shannon Species Diversity	The species-level Shannon diversity value for the sample (see <a href="https://en.wikipedia.org/wiki/Shannon-Wiener_index">en.wikipedia.org/wiki/Shannon-Wiener_index</a> for details)
Number of species identified	The number of unique species classifications in the sample, only including reads classified to the species level.

## Sample Summary Page

The 16S Metagenomics App provides an overview of statistics per sample on the sample pages. You can also download the Summary Report as a PDF.

## Sample Results

The sample results section contains a PDF Summary Report that provides information about the classification rate and top sample classifications by taxonomic level. The sample results section also contains a CSV Classification Summary, which shows the aggregate classification counts for the sample.

## Sample Information

The Sample Information table provides basic primary analysis information for the sample.

Statistic	Definition
Number reads PF	The total number of reads with an index sequence matching this sample that passed quality filtering

## Classification Statistics

The Classification Statistics section provides information on the classification rate by taxonomic level.

Statistic	Definition
Reads PF Classified to Taxonomic Level	The total number of PF reads for this sample that were classified at the given taxonomic level
% Reads PF Classified to Taxonomic Level	$(\# \text{ Reads PF Classified to Taxonomic Level} / \# \text{ Reads PF}) * 100$

## Sunburst Classification Chart

The sunburst chart presents statistics about the relative abundance and taxonomic hierarchy of classifications in the sample.

Statistic	Definition
Total reads	# Reads PF
% Total reads	$(\# \text{ Reads PF Classified in taxonomic category} / \# \text{ Reads PF}) * 100$
% [parent category] reads	$(\# \text{ Reads PF classified in taxonomic category} / \# \text{ Reads PF classified in parent taxonomic category}) * 100$

## Top 20 Classification Results by Taxonomic Level

The column chart shows a comparison of the relative abundance of classifications at each taxonomic level, excluding reads unclassified at that taxonomic level.

Statistic	Definition
Total reads	# Reads PF
% Reads classified to level	$(\# \text{ Reads PF Classified in taxonomic category} / \# \text{ Reads PF classified to taxonomic level}) * 100$
% total reads	$(\# \text{ Reads PF classified in taxonomic category} / \# \text{ Reads PF}) * 100$

## Analysis Info

This app provides an overview of the analysis on the Analysis Info page.

A brief description of the metrics is below.

Row	Definition
Name	Name of the app session.
Application	App that generated this analysis.
Date started	Date and time the app session started.

Row	Definition
Date completed	Date and time the app session completed.
Duration	Duration of analysis.
Session Type	The number of nodes used.
Size	Total size of all output files.
Status	Status of the app session.

## Log Files

Clicking the **Log Files** link at the bottom of the Analysis Info page provides access to 16S Metagenomics app log files. Log files are located in a folder in the Output Files section.

- ▶ **Output-AppSessionID.log**: shows the raw console output from Isis and ClassifyReads
- ▶ **Spacedock-AppSessionID.log**: shows console output from the SpaceDock and BaseSpace communication and input/output file staging.

## 16S Metagenomics Status

For single or multiple samples, the status of the 16S Metagenomics app can have the following values:

- 1 Initializing
- 2 PendingExecution
- 3 Running
- 4 Downloading Application Container
- 5 Launching Isis
- 6 Classifying Samples
- 7 Calculating Statistics
- 8 Creating Metagenomics Report
- 9 Uploading Results
- 10 Application completed successfully

## Inputs Overview

The 16S Metagenomics app provides an overview of the input samples and settings that were specified when setting up the 16S Metagenomics run.

## 16S Metagenomics Output Files

The output files link provides access to the output files. See the following topics for descriptions.

Per-Sample output:

- ▶ *Summary.csv* on page 10
- ▶ *Txt.gz* on page 10

- ▶ *Report.pdf* on page 10
- ▶ *Report.html* on page 10

#### Aggregate output:

- ▶ *MetagenomicsAggregateReport.html* on page 10
- ▶ *MetagenomicsAggregateReport.pdf* on page 10
- ▶ *Aggregate\_Counts.csv* on page 10

### Summary.csv

The *summary.csv* files show summarized counts of how reads were classified in the sample. Each unique classification is a row, and the number of hits and % of total hits are recorded for each row. Partial classifications (e.g. where a Class-level classification was made, but not an Order-level classification) have unclassified entries left blank.

### Txt.gz

The *txt.gz* files provide raw classifier output. Entries in the file contain the following:

- ▶ A read identifier from the input FASTQ file.
- ▶ The classification assigned to the read with a confidence value in the range 0 to 1 for each taxonomic level of the classification.

### Report.pdf

The per-sample PDF reports contain summary information about the sample, similar to the information presented in the BaseSpace per-sample report. All statistics have the same definitions as in the BaseSpace per-sample report. The PDF report does not contain the interactive visualizations.

### Report.html

The per-sample HTML report contains the same information as presented in the BaseSpace per-sample report. There are some formatting differences between the BaseSpace and standalone versions; however the statistics and results are identical.

### Aggregate\_Counts.csv

The aggregate counts *csv* files provide per-level aggregate counts for all samples. The files are split by taxonomic level, as indicated in the file name (e.g. *Class\_Level\_Aggregate\_Counts.csv*). Each row represents a unique classification that occurred in one or more samples. Each sample is assigned a column, labeled with the sample ID in the header row. Entries in the table represent the number of reads with the row classification label in the column sample.

### MetagenomicsAggregateReport.pdf

The aggregate PDF report contains the same statistics tables as presented in the BaseSpace aggregate report. Definitions and values of the statistics are identical to the statistics reported in the BaseSpace report. The aggregate PDF report does not contain interactive visualizations.

### MetagenomicsAggregateReport.html

The aggregate HTML report contains the same information as presented in the BaseSpace aggregate report. There are some formatting differences between the BaseSpace and standalone versions. The statistics and results are identical.

## 16S Metagenomics Methods

This chapter describes the methods that are used in the 16S Metagenomics app.

### Read Classification

The classification step uses ClassifyReads, a high-performance implementation of the Ribosomal Database Project (RDP) Classifier described in Wang Q. et al., 2007 ([dx.doi.org/10.1128/2FAEM.00062-07](https://doi.org/10.1128/2FAEM.00062-07)). The original RDP classifier algorithm used 8-base words due to implementation constraints. ClassifyReads uses more efficient data structures and is able to use 32-base words – giving each word more specificity for each species. Below, we provide information on the classification steps used in ClassifyReads.

#### Word-specific Priors

Let  $W = \{w_1, w_2, \dots, w_d\}$  be the set of all possible 32-base words. From the corpus of  $N$  sequences, let  $n(w_i)$  be the number of sequences containing subsequence  $w_i$ . The word-specific prior estimate of the likelihood of observing word  $w_i$  in a sequence over the entire corpus can be calculated with the following formula:

$$P_i = \frac{n(w_i) + 0.5}{N + 1}$$

#### Taxonomic Conditional Probabilities

In the original RDP classifier, classifications were performed to the genus level. In ClassifyReads, we have extended this behavior to the species level. In the sections below, we denote  $T$  to represent classifications at a given taxonomic level (genus, species, or subtype). By default, we classify at the species level.

For species  $T$  with a training set consisting of  $M$  sequences, let  $m(w_i)$  be the number of sequences containing word  $w_i$ . The conditional probability that a member of  $T$  contains  $w_i$  was estimated with the equation:

$$P(w_i|T) = \frac{m(w_i) + P_i}{M + 1}$$

The joint probability of observing from species  $T$  a partial sequence,  $S$ , containing a set of words,  $V = \{v_1, v_2, \dots, v_f\}$  ( $V \subseteq W$ ), was estimated as:

$$P(S|T) = \prod P(v_i|T)$$

#### Naïve Bayesian Assignment

By Bayes' theorem, the probability that an unknown query sequence,  $S$ , is a member of species  $T$  is:

$$P(T|S) = \frac{P(S|T)P(T)}{P(S)}$$

$P(T)$  is the prior probability of a sequence being a member of  $T$   
 $P(S)$  is the overall probability of observing sequence  $S$  (from any species).

Assuming all species are equally probable (equal priors), the constant terms  $P(T)$  and  $P(S)$  can be ignored. We classify the sequence as a member of the genus giving the highest probability score.

## Example

Here are the results from a typical classification (results are in log space):

```
Methanocaldococcaceae;MethanocaldococcusP(T|S) = -367.7  
Methanococcaceae;MethanothermococcusP(T|S) = -2936  
Thermofilaceae;ThermofilumP(T|S) = -2963.2
```

The natural log-transformed probabilities are small due to product multiplication during the joint probability calculation. The huge difference between the best hit and the second best hit is notable. A difference of 3 would indicate  $20\times$  higher probability for the best hit. Here the difference is 2568.3 (i.e.  $2.5 \times 10^{115}$  higher probability).

## Removing Taxonomic Levels

If the difference between the two best hits is small: 0.05 in log space (within  $1.6\times$  higher probability), we prune the taxonomic levels of the best hit until it is the same as the second best hit.

For example, consider the following results:

```
Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Prevotellaceae  
;Prevotella;melaninogenica - P(T|S) = -200.00  
Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Prevotellaceae  
;Prevotella;histicola - P(T|S) = -200.04
```

In this case, the difference between the top two hits is less than 0.05. Therefore we prune the away the taxonomic levels from the bottom (species) to the top (kingdom) until both hits are identical. In this case, we stop pruning when we reach the *Prevotella* (genus) classification.

The pruning procedure reflects that we have strong evidence that the read comes from the *Prevotella* genus, but that we do not have enough evidence to deduce which species.

The removal threshold (0.05) was determined by analyzing misclassifications during loop-back classification of the Greengenes database.

## Gaining Species Level Classification

ClassifyReads uses the full Greengenes database. Some of the entries in that database are only classified down to the genus level, while other entries are classified down to the species and subtype level. In some instances, the top results can look like this example:

```
Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Prevotellaceae  
;Prevotella; - P(T|S) = -200.000  
Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Prevotellaceae  
;Prevotella;histicola - P(T|S) = -200.000  
Bacteria;Firmicutes;Bacilli;Gemellales;Gemellaceae;Gemella;haem  
olysans - P(T|S) = -1846
```

In this case, the best hit is classified down to the genus level and the second best hit is classified down to the species level. If the difference between the two best hits is near identical (difference less than 0.001 in log space), we use the species level classification instead.

## Classification Confidence

The RDP classifier used a bootstrapping method of randomly subsampling the words in the sequence to determine the classification confidence.

In ClassifyReads, we no longer perform this bootstrapping procedure. This change is primarily due to performance reasons (bootstrapping slowed the algorithm down by 20–50×) and weak correlation between the resulting confidence estimate and the actual classification accuracy.

At the moment, confidence is statically assigned based on the overall accuracy of our classification algorithm at different taxonomic levels:

Taxonomic Level	Accuracy
Kingdom	100%
Phylum	100%
Class	100%
Order	99.98%
Family	99.97%
Genus	99.65%
Species	98.24%

## Taxonomic Database

The taxonomic database used is an Illumina-curated version of the May 2013 release of the Greengenes Consortium Database ([greengenes.secondgenome.com/downloads](http://greengenes.secondgenome.com/downloads)).

Here are the current statistics for that database:

Taxonomic Level	# of classifications
Kingdoms	3
Phyla	33
Classes	74
Orders	148
Families	321
Genera	1086
Species	6466

To get taxonomies down to the species level, we used the Greengenes SQL database files (gg\_13\_5.sql.gz). Specifically our database started off with everything contained in the Greengenes clones, isolates, and symbionts tables. From there, we apply a set of filters:

- 1 Filter all entries where the 16S sequence length was below 1250 bp.
- 2 Filter all entries that had more than 50 wobble bases (i.e. M, R, W, S, Y, K, V, H, D, B, N)
- 3 Filter all entries that were only partially classified (no classification for genus or species)

The Greengenes database had a number of classifications placed in the wrong field. i.e. improper genus or species names, placing clone or strain IDs in the species field, etc. We developed a program to help identify and clean up these entries.

Ambiguous epithets and classifications (sp, aff, cf, genosp, genomsp) were removed, because they effectively mean the same thing as an empty taxonomic level.

*Listeria monocytogenes* (GenBank entry X56153.1), *Listeria innocua* (GenBank entry FJ774235.1), and PhiX (NCBI reference sequence: NC\_001422) were added to the database to support internal research projects.

## Aggregate PCoA Chart

The principal coordinates analysis (PCoA) chart in the aggregate report is generated using classical multidimensional scaling (MDS) on normalized classification vectors for each sample. An overview of the steps of the algorithm is presented in this section.

### Normalized Classification Matrix

A normalized classification matrix is created for a range of taxonomic levels. The first matrix includes only kingdom classifications, the second includes kingdom and phylum, and so on, until the full range of taxonomic levels is considered in the species-level classification matrix. These ranges correspond to the user-selectable levels in the aggregate report.

The set of classifications present within the current range of taxonomic levels across all samples is collected. Then a label vector is created by placing each unique classification at a unique index in the vector. Classifications for each sample within the current taxonomic range are collected, and `unclassified` classifications are discarded. A vector is created for each sample, which is the projection of the sample classifications within the current taxonomic range onto the label vector of all non-empty classifications at each index. Each sample vector is then L-1 normalized by multiplying every index by the inverse of the sum of the sample vector.

The resulting vectors form a classification matrix, in which each row represents a unique non-empty classification present within the current taxonomic range. Each column represents one sample L-1 normalized projection of non-empty classifications onto this space.

### Pearson Correlation Distance Matrix

Pearson correlation is calculated for each pair of L-1 normalized sample classification vectors. A distance matrix is then calculated as  $1 - r$ , where  $r$  is the Pearson correlation

measure between two samples normalized classification vectors.

### Classical MDS

Classical MDS is performed on the distance matrix output from the previous step. The MDS is implemented as described in [steep.inrialpes.fr/~Arnaud/indexation/mds03.pdf](http://steep.inrialpes.fr/~Arnaud/indexation/mds03.pdf).

### Aggregate Dendrogram Chart

A per-sample classification vector is created for each sample using the same method described in *Aggregate PCoA Chart* on page 14. The difference is that the current taxonomic range is defined as genus only (only genus-level, non-empty classifications are considered). We then calculate a Pearson Correlation Distance Matrix, as described in *Pearson Correlation Distance Matrix* on page 14. Finally, we use standard UPGMA hierarchical clustering to generate the dendrogram.

## Notes

## Technical Assistance

For technical assistance, contact Illumina Technical Support.

**Table 1** Illumina General Contact Information

<b>Illumina Website</b>	www.illumina.com
<b>Email</b>	techsupport@illumina.com

**Table 2** Illumina Customer Support Telephone Numbers

<b>Region</b>	<b>Contact Number</b>	<b>Region</b>	<b>Contact Number</b>
North America	1.800.809.4566	Italy	800.874909
Austria	0800.296575	Netherlands	0800.0223859
Belgium	0800.81102	Norway	800.16836
Denmark	80882346	Spain	900.812168
Finland	0800.918363	Sweden	020790181
France	0800.911850	Switzerland	0800.563118
Germany	0800.180.8994	United Kingdom	0800.917.0041
Ireland	1.800.812949	Other countries	+44.1799.534000

### Safety Data Sheets

Safety data sheets (SDSs) are available on the Illumina website at [support.illumina.com/sds.ilmn](http://support.illumina.com/sds.ilmn).

### Product Documentation

Product documentation in PDF is available for download from the Illumina website. Go to [support.illumina.com](http://support.illumina.com), select a product, then click **Documentation & Literature**.



Illumina

San Diego, California 92122 U.S.A.

+1.800.809.ILMN (4566)

+1.858.202.4566 (outside North America)

[techsupport@illumina.com](mailto:techsupport@illumina.com)

[www.illumina.com](http://www.illumina.com)