FLORAGENEX
Your partner from DNA to data

Floragenex Project File Readme

July 2012

Version 1.3

Overview

This document provides a description of folder content and files contained within your project archive. Please note that depending on the bioinformatics analysis used, not all of the files described below may be present. For example, a population genomics study would not have the same files included for a genetic mapping study. Please contact your Floragenex project manager with any questions.

Raw_Illumina_Sequence Directory

All raw sequence data generated for your project is contained within this folder. Sequence data is typically demultiplexed from individual Illumina GAIIx / HiSeq flowcells and is segregated into individual sequence files for each sample prior to analysis. Sequence data is provided in fastq format with Illumina 1.9+ quality score encoding (http://en.wikipedia.org/wiki/FASTQ_format). Files are named by appending the sample name the designation "_sequence_1.txt" indicating a single read sequence or "_sequence_2.txt" for a paired-end read. To facilitate downstream analysis the 5' multiplex index (barcode) has been stripped away from the sequence read and quality score. Please note the RAD-Seq restriction enzyme digestion site is retained in the read. An accompanying sequencing quality control report which details the sequencing metrics obtained for each sample is also included in this folder (Figure 1).

Genome Directory

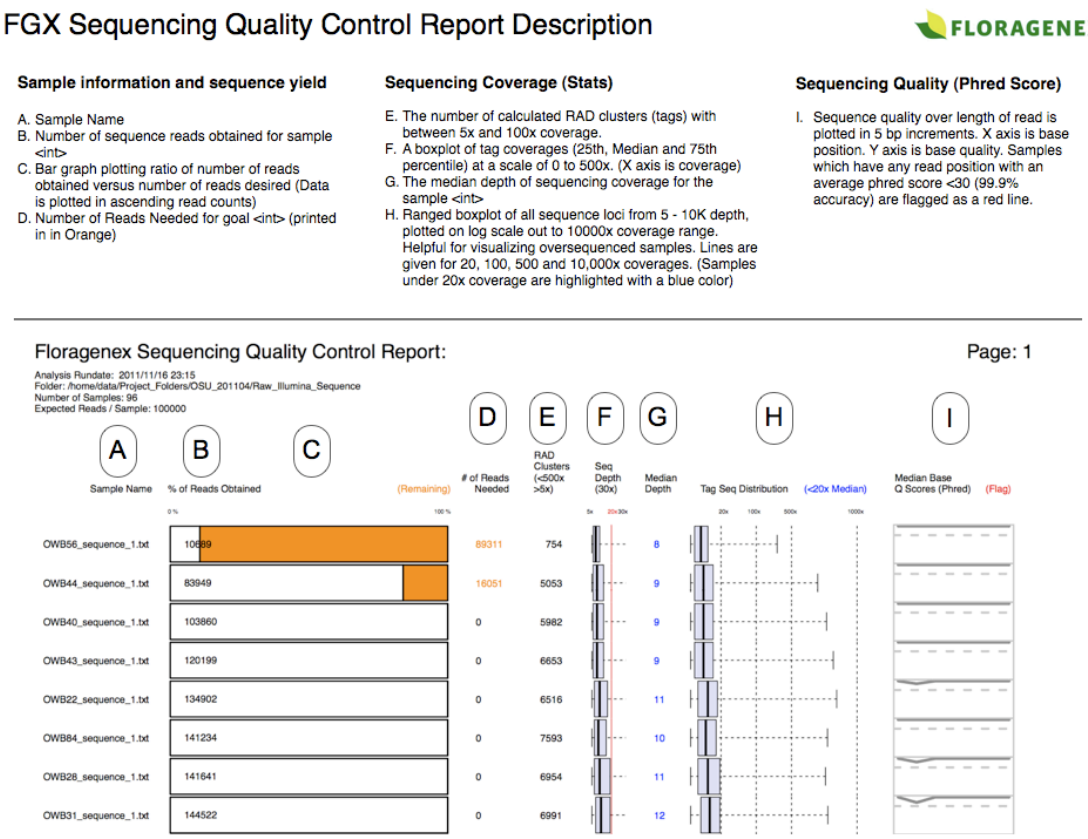
Depending on the specifics of your project, one or more of the following reference sequences may be present in this folder.

1. If the analysis uses or requires a pre-existing reference genome (i.e for Maize, the *Zea mays* B73 ref assembly), it will be included here in fasta format.
2. If your project involves single end RAD-Seq and a reference genome is not available, a RAD-Seq "unitag" assembly may be provided. The unitag assembly is a skeleton

framework of putative single dose / low copy RAD sequences generated from a single sample which are used for positioning and alignment of raw Illumina sequence reads from other samples.

3. If your project involves paired-end RAD-Seq, sequence contigs assembled using Velvet are provided. A .flif file (an intermediate file used during Velvet assembly) will be present, along with a filtered assembly that has been scrubbed of putative contaminant high copy sequences (ribosomal, prokaryotic, plastid). Filtering is handled through bwa alignment of contigs to a custom FGX sequence database. Contaminant sequences which are removed from the initial assembly are printed to the "contaminants" file for investigator reference. Basic descriptive statistics (total bp, N50, histogram) on the filtered assembly are printed to the Logs directory (see Logs below for more information).

Figure 1. Sequencing Quality Control Report



FGX Sequencing Quality Control Report Description

FLORAGENEX

Sample information and sequence yield

A. Sample Name
B. Number of sequence reads obtained for sample <int>
C. Bar graph plotting ratio of number of reads obtained versus number of reads desired (Data is plotted in ascending read counts)
D. Number of Reads Needed for goal <int> (printed in in Orange)

Sequencing Coverage (Stats)

E. The number of calculated RAD clusters (tags) with between 5x and 100x coverage.
F. A boxplot of tag coverages (25th, Median and 75th percentile) at a scale of 0 to 500x. (X axis is coverage)
G. The median depth of sequencing coverage for the sample <int>
H. Ranged boxplot of all sequence loci from 5 - 10K depth, plotted on log scale out to 10000x coverage range. Helpful for visualizing oversequenced samples. Lines are given for 20, 100, 500 and 10,000x coverages. (Samples under 20x coverage are highlighted with a blue color)

Sequencing Quality (Phred Score)

I. Sequence quality over length of read is plotted in 5 bp increments. X axis is base position. Y axis is base quality. Samples which have any read position with an average phred score <30 (99.9% accuracy) are flagged as a red line.

Floragenex Sequencing Quality Control Report: Page: 1

Analysis Rundate: 2011/11/16 23:15
Folder: /home/data/Project_Folders/OSU_201104/Raw_Illumina_Sequence
Number of Samples: 96
Expected Reads / Sample: 100000

A Sample Name	B % of Reads Obtained	C (Remaining)	D # of Reads Needed	E RAD Clusters (<500x >5x)	F Seq Depth (30x)	G Median Depth	H Tag Seq Distribution (<20x Median)	I Median Base Q Scores (Phred) (Flag)
OWB56_sequence_1.txt	10689		89311	754		8		
OWB44_sequence_1.txt	83949		16051	5053		9		
OWB40_sequence_1.txt	103860		0	5982		9		
OWB43_sequence_1.txt	120199		0	6653		9		
OWB22_sequence_1.txt	134902		0	6516		11		
OWB84_sequence_1.txt	141234		0	7593		10		
OWB28_sequence_1.txt	141641		0	6954		11		
OWB31_sequence_1.txt	144522		0	6991		12		

SAM_BAM_Pileups (or similar) Directory

Sorted binary alignment map (BAM) files generated from Bowtie alignment of raw Illumina data to the specified reference genome are provided for each project sample. (http://bowtie-bio.sourceforge.net/index.shtml) SAMtools pileups files are a versatile format for parsing of SNP and genotype data and are included in this directory (http://samtools.sourceforge.net/). Note pileups are used in downstream population genetics and genetic mappins analysis (PopGen / GenMap).

PopGen Directory

For SNP discovery and population genotyping efforts, marker and genotype files are contained in this folder. Sequence data for the target population is distilled into Variant Call Format (VCF) 4.1 format (for description see: http://www.1000genomes.org/wiki/Analysis/Variant Call Format/vcf-variant-call-format-version-41). VCF is a data-rich scheme that can be parsed on multiple parameters during allele mining (sequence coverage, MAF, heterozygosity, etc). Floragenex VCF format is further optimized for SNP discovery and genotyping by sequencing experiments.

Custom Floragenex VCF INFO Fields: Report metrics on overall population surveyed

- NS = Number of samples with genotype data
- DP = Total sequencing depth at locus across population
- AN = Number of non-reference alleles in population
- AS = SNP annotated with flanking genomic sequence NNN ... [N/N] ... NNN

Custom Floragenex VCF FILTER Fields:

- q10 =SNP quality below phred scaled likelihood of 10 (90% accuracy)
- s50 =Less than 50% of samples have data (high fraction of missing data)

Custom Floragenex VCF FORMAT Fields:

- GT =FGX consensus genotype (basic threshold model). Genotype reported in VCF numerical format
- DP = Sequencing depth observed for this sample
- GQ = Genotype quality (SAMtools bayesian framework)
- EC = Alternate (non-reference) allele counts observed in sample
- SG = SAMtools consensus genotype (bayesian, diploid model)

In SNP discovery studies where downstream genotyping design is a goal, annotated SNPs with flanking sequence can be located in the VCF INFO field with the designation AS="...[N/N]..." for each putative locus.

GenMap Directory

For genetic mapping studies with RAD-Seq, a common deliverable is distillation of sequence data into genotypes compatible with the genetic mapping software JoinMap. If present, this folder will contain one or more JoinMap 4.0 formatted genotype files, which have been produced from analysis of the sequence data.

Logs Directory

Analysis logs provide a record of analyses performed for your project and provide a large number of descriptive statistics generated during Floragenex bioinformatics processing. Please refer to the individual logs for more details or contact your project manager for more information on these files.