Genome scan

Next Generation Sequencing

Bioinformatics for NGS projects

Guidelines

genomescan.nl

GenomeScan's Guidelines for Bioinformatics Services on NGS Data

Using our own proprietary data analysis pipelines

Dear customer,

As of the beginning of 2015 ServiceXS became a trademark of GenomeScan B.V. GenomeScan focuses exclusively on Molecular Diagnostics whereas our ServiceXS trademark is intended for your R&D projects.

GenomeScan is dedicated to help you design and perform Next Generation Sequencing (NGS) experiments that generate high quality results. This guide provides information for our data analysis services and resources and tools for further analysis of your sequencing data. NGS experiments result in vast amounts of data and therefore data analysis can be challenging. Our ability to assist in the analysis of your results can be the key factor leading to a successful project.

Our experience in the past years is that even state-of-the-art NGS software is not always able to fulfil the data analysis needs of our customers. To alleviate this problem our experienced team of bioinformaticians and molecular biologists can provide standard or custom bioinformatics solutions to get the most out of your project.

GenomeScan provides a comprehensive package of bioinformatics services for our NGS customers, which enable them to utilise all the applications that are possible with billions of bases of sequence data per run. GenomeScan can advise and assist you in every step of the data analysis. Do not hesitate to contact us if you have any questions after reading this guideline!

On behalf of the Bioinformatics team,

Thomas Chin-A-Woeng
Project Manager

Document Outline

Changes to Previous Version (1.0)
-Lay-out changes

Chapter 1 Introduction

Next-generation sequencing (NGS, also high-throughput sequencing) makes it possible to generate vast amounts of sequencing data in comparison to the Sanger sequencing method. The large amount of data generated and new types of applications based on this technology offer novel challenges to scientists and bioinformaticians. One run with a next-generation sequencer typically generates several megabases to hundreds of gigabases of DNA sequences.

NGS enables a broad range of new applications including large scale mutation screening and SNP or marker discovery, sequence based gene expression studies (digital gene expression), ChIP sequencing, whole transcriptome sequencing, methylome sequencing, etc. Due to the typical nature of the data (gigabases of data, relatively short reads, high redundancy, high error rates) it cannot be handled by conventional molecular biological software tools. Often, specialised high throughput data analysis software and hardware to account for the large memory requirements and computational speed is needed. The amount of data also requires new ways to summarise and visualise the data to the researcher.

NGS results in data that often need to be transformed to information and formats intelligible to researchers. Data analysis and interpretation of NGS data is therefore crucial for the success of a project. The major importance of data analysis is shown in our complete work flows starting from experimental design, followed by excellent performance of your experiments resulting in high quality data, to extensive support and assistance in the analysis and interpretation of the generated data.

Our team of skilled bioinformatics and molecular biologists ensures that your data is analyzed and interpreted the way you would like to have interpreted. GenomeScan has both the hardware and software required to perform these sometimes computationally intensive tasks. A choice from commercially available software packages, published and validated software tools, or in-house developed and validated software is made that best fits the job resulting in reliable and easily intelligible data. GenomeScan performs only in-house analysis and uses local copies of major publicly accessible databases, always ensuring that your data and results are treated in a confidential way.

GenomeScan would also like to draw your attention to our possibilities to validate the outcome of your NGS project. The broad GenomeScan portfolio of genomics services allows you to take your data across other genomic technologies to validate or expand your results. We are able to offer you next to experimental design consultancy, experiment performance and extensive data analysis, validation of your results using one of our other services. Our data processing pipelines are optimized to transform your experimental design and output data to and across multiple platforms such as array-based genotyping and gene expression analysis on Illumina or Affymetrix platforms.

GenomeScan has many years of experience in data analysis for its portfolio of services. As one of world's first service providers for NGS GenomeScan has gathered a vast amount of knowledge and tools to handle the whole process of transforming raw sequence data obtained from sequencers such as Illumina, Pacific Biosciences, and Life Technologies, into easily intelligible reports and interpretable file formats. Visit our website, download our detailed guidelines or contact us for more information.

Chapter 2 Sequencing Applications

After obtaining NGS data it has to be further processed or transformed before interpretation of the results can begin. Our own scientists determine whether the results meet the GenomeScan standards as well as quality criteria defined by the platform supplier before the data is send out. Our bioinformatics team provide a comprehensive bioinformatics service for our next-generation sequencing customers, which enable them to utilise all the applications that are possible.

GenomeScan offers standard workflows for:

- Quality assurance
- Data filtering
- Whole genome (re-)sequencing
- Targeted sequencing
- SNP and mutation analysis
- *de novo* assembly and annotation
- Whole transcriptome sequencing (mRNA-seq)
- Digital gene expression (small RNA, SAGE, CAGE)
- ChIP sequencing
- Methyl sequencing
- Copy number variation
- Functional annotation
- Visualisation and annotation of sequence data within a genome browser
- Performance and validation of custom data analyses

2.1 Quality Assurance of sequencing data

A major factor determining the quality of the final outcome of a NGS project is proper quality checking and filtering of the data. Based on these quality metrics we decide whether the data set and what part of the data set is optimal for further processing. Quality checks and quality scores at each level of the work flow ensures minimal errors in the final outcome. Summary information of the general quality and distribution of errors are generated in comprehensive charts with clear descriptions how to interpret the results.

Quality score analysis

Most of the current sequencers produce output in Sanger FASTQ format (see Chapter 5: File formats). The FASTQ files contain all the bases and a quality score attached to each base. These quality scores are in a certain format, in which a character represents its corresponding Q-score (see Table 2.1).

A quality score (or Q-score) expresses an error probability. In particular, it serves as a convenient and compact way to convey very small error probabilities.

Given an assertion, A, the probability that A is not true, P(~A), is expressed by a quality score, Q(A), according to the relationship: Q(A) =-10 log10(P(~A)) where P(~A) is the estimated probability of an assertion A being wrong.

The relationship between the quality score and error probability is demonstrated with the following Table 1.

Table 1. Relationship between quality score and error probability

Quality score, Q(A)	Error probability P(~A)
10	0.1
20	0.01
30	0.001

Quality metrics

After sequence files are generated, the statistics of the quality of the sequence data is summarised per sample. The following section describes some general quality metrics that can be used to evaluate and characterise the sequence data. Typical quality metrics are quality distribution, mean probabilities of error, base composition, read or sample GC content, number of N calls, and trim results. Fig. 1 shows some quality metrics. Not all metrics are relevant for the different sequencing platforms.

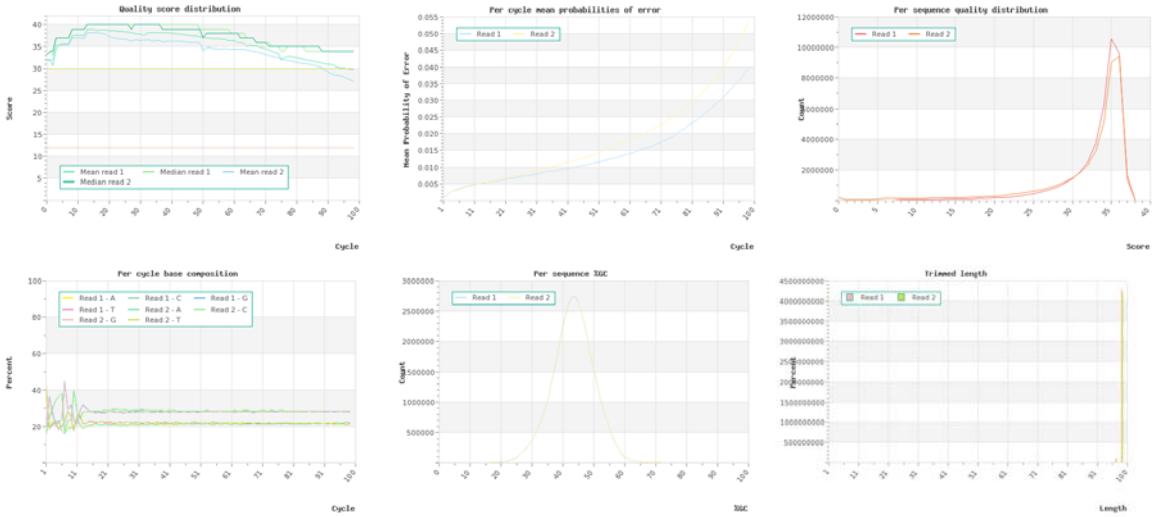

Fig. 1. Examples of quality metrics for NGS DNA sequencing

2.2 Data Filtering

Each next-generation sequencer generates data of different characteristics. Depending on the type of sequence data and application it may be necessary to further polish the sequencing data on top of the basic quality filter available via the manufacturer of sequencer. Often, the default sequencing software removes the worst sequences from the data set during primary analysis, but depending on the application (e.g. mutation analysis or expression analysis) additional quality filtering may be required. We perform quality assurance on sequencing data using in-house techniques and software tools while keeping quality measures for a particular purpose in mind. Data filtering may also differ per sequencing platform. E.g. Illumina/Solexa technology typically generates reads with a sudden quality drop after certain read lengths, while the Life Technologies Ion platform has problems sequencing homopolymeric tracts.

Quality filtering and read trimming

GenomeScan uses in-house tools to pre-process sequence reads prior to its downstream data analysis pipelines. Raw sequence files are read, checked for integrity and a new set of filtered data files is generated meeting the filtering criteria defined for a particular type of analysis. This also normalises data quality for different runs since this can differ between runs.

Reads with excessive low quality bases or uncalled bases (Ns) can be removed from your data set according certain filtering algorithms and thresholds by trimming reads at both the 5' and/or 3' end

or discarding the read. Reads can be discarded if they are too short, or when a percentage of quality scores is lower than the indicated threshold. More advanced trimming profiles allow separate thresholds for each quality score distribution (e.g.100%>Q10, 80%>Q20, 70%>Q30 for a read) to be defined.

Barcode sorting
When using custom bar coding reads can be demultiplexed into separate file based on the barcode in the sequence read or index read. Both standard barcoding as well as in-read barcodes and dual barcoding is supported.

Adapter clipping
Primer sequences due to adapter-adapter ligations or read-through of the insert can be trimmed from the reads. Our algorithms determine adapter sequences accurately at single base level assuring that not a single base of adapter sequence can interfere with downstream processing (e.g. assembly). Known adapter sequences can be removed from both 5' and 3' ends.

Size trimming
Fixed size trimming of reads at both 5' and 3' ends can be performed. Quality score-based trimming at 5' and 3' end leaving the longest part of the read of good quality is the most common means of filtering for our pipelines. Minimum lengths for the remaining sequence read and minimum base quality scores can be defined.

Error correction
Using coverage information errors in reads can be repaired. Roche 454/IonTorrent/IonProton reads contain unreliable polymeric tracts that can be repaired using this coverage information. DNA sequences can also be improved by combining multiple sequencing platforms.

Statistics
In addition to filtering the sequence files, general statistics are gathered and calculated in terms of the number of accepted or rejected reads and bases, quality and length distribution, and total counts. All is provided in a number of tables and charts. Summary information of the general quality and distribution of errors can be generated.

2.3 Whole genome (re-)sequencing of strains or related organisms

When re-sequencing an entire genome the reads can be aligned to a customer-provided or public reference. Any reference from any organism, bacterial, microbial, plant, or animal can be used. The alignment is provided in a standard indexed and sorted BAM file and optionally formatted to allow browsing and viewing in one of the commercial or public genome viewers. Alignment statistics to assess the quality of alignments for re-sequencing projects are provided.

Based on the alignments we can designate the major differences between your organism and the reference. Differences between wild-type and (mutant) derivative strains are listed in a comprehensive table showing the location and type of difference. Or we can simply compare your strain of interest to a known genome sequence. The differences can be coded into tracks that allow viewing and browsing in one of the commercially or publicly available genome browsers, providing you access to annotations and other knowledge databases. A new reference sequence of a derivative strain with liftover annotation can be generated.

2.4 Targeted Re-sequencing

There are many ways to obtain selected genes or genomic regions. These include PCR, reduced representation libraries (RRLs), hybridisation capture etc. It is often useful to analyse the genomic

regions in multiple samples to, for example, detect known and new mutations in specific genes or regions. Large (hundreds) numbers of disease and control samples can be analysed with multiplexing. Targeted re-sequencing is particularly of interest for disease inheritance studies, mutation detection in inherited clinical disorders, deep sequencing of selected regions in family members with affected individual(s), targeted re-sequencing of QTL loci, and many more applications.

Each application may require a different handling of the raw sequence data and further processing. To view the enriched regions, files or formats required to visualise the enriched regions can be provided in a custom or software specific formats or as files loadable in standard genome browsers (e.g. UCSC Wiggle and BED file formats).

2.5 Genotyping, SNP, and Mutation Analyses

GenomeScan has a well-performing and validated pipeline for identification and enumeration of nucleotide variants and micro insertion/deletions (indels). SNV (Single Nucleotide Variant) calling can be based on a reference sequence alignment or if a reference is not available a novel reference from the PCR, RRL (reduced representation library), or whole genome sequencing data is generated prior to variant discovery.

SNVs are provided in the VCF standard and summarized in tables with quality scores. Subsets can be filtered and selected and validated with any of our other genomics services. Mutations or SNVs can be mapped back onto the genome and annotation of genes affected by the mutations allows the researcher to determine the biological importance of mutations in specific regions. Our reports provide information such as the reference nucleotide, read depth, percentages for each possible nucleotide as well as insertion and deletion percentages for each mutation.

Verification of the SNVs and mutations found on other platforms shows that NGS combined with our SNV discovery pipeline is a very efficient tool to find SNVs throughout entire genomes. Mutations or SNVs can be mapped back onto the genome and annotation of genes affected by the mutations allows the researcher to determine the biological importance of mutations in specific regions. Our pipeline generates a (custom-)filtered SNV table along with the reliability of the SNV calls or detailed mutation report with detailed information about reported mutations. The reports provide information such as the reference nucleotide, variants, observed alleles, genotypes, read depth, frequencies, and mapping qualities. Our SNV reports can compare mutation calls in two or more genotypes based on the same reference.

When you are ready to further validate or use the new SNP data our pipeline can make a selection of SNVs of interest, filter for SNV assay designability, and transform the SNP information along with the flanking region information into a format that can be fed in genotyping assay design tools such as for Illumina Infinium genotyping.

For more disease-related projects GenomeScan can provide additional structural variation information such as translocations, gene fusions, exon skipping, and readthroughs for transcriptome analysis) using the mate information of the sequences. Even for single-read data pseudo paired reads can be generated if reads are of sufficient length or elongated. The structural variation report lists areas of possible structural variations. Variations can be cross-referenced to already reported occurrences in disease or mutation databases.

Please see our small variants analysis guidelines for further details about the principles, methods, and file formats.

2.6 *de novo* assembly

Although *de novo* assembly of short sequencing reads into entire genomes remains a challenging task, assembly into a small number of contigs can provide valuable information for data mining and gene discovery, and comparative genomics. It allows the discovery of novel or foreign genes in an organism and allows metagenomics of all kinds of communities and environments, e.g. intestinal microbial flora, soil, marine ecosystems, and other micro-environments. Depending upon the type of data, quality and quantity (coverage), short read assembly techniques can accurately produce contigs of greater than 100 kb from short reads. The obtained assembly can be used for further analysis or simply blasted and annotated with information from available databases.

De novo assembly using short reads is challenging since many genomic structures (e.g. repeats, gene homologs) cannot be easily resolved due to the nature of the data. We utilise the commonly used de Bruijn graph assembly method for generating large contig assemblies. This method involves the use of short words that are used as indices to create a graph which reduces redundancy instead of using the entire read. When mate information is available this additional information can be used as additional information for assembly. To obtain the best result some optimisation and tweaking of assembly parameters is often required which is time-consuming and some experience is needed.

For longer read lengths whole genome assembly software or manufacturer specific algorithms can be used. These assemblies are often based on a maximum overlap method.

For complicated genomes, hybrid assemblies e.g. Illumina short reads combined with PacBio long reads) result in better assemblies. Since our portfolio consists of multiple types of sequencers we can complete the entire process of sequencing and assembly using the hybrid approach in a single project. Customers may also provide sequencing data they already have generated to be combined with the new sequencing results. Please see our sequence assembly guidelines for further details about the principles, methods, and file formats.

2.7 Whole Transcriptome Sequencing (mRNA-Seq)

Whole Transcriptome Shotgun Sequencing or mRNA-Seq refers to the use of high-throughput sequencing to sequence cDNA in order to get information about the RNA content of a sample. With deep coverage, NGS provides information on differential expression of genes, including gene alleles and differently spliced transcripts; non-coding RNAs; post-transcriptional mutations or editing; and gene fusions.

Transcriptome/RNA sequencing of organism can also be performed with limited reference or no reference information available. After assembling sequence reads into a transcriptome, gene expression levels can be calculated. Furthermore, annotation of the data set can be performed, premature stop codons, splice isoforms, genomic rearrangements, and gene/transcript read counts can be provided.
Please see our mRNA-Seq guidelines for further details about the principles, methods, and file formats.

2.8 Digital Gene Expression

cDNA, SAGE, CAGE, miRNA and sncRNA data can be aligned onto a reference and visualised in a genome browser. Digital gene expression reports are created to show the sequence of each tag, the coverage, gene names, and the location in the genome along with annotations and links to other resources. New gene tags that are not in the library are also reported as novel tags.

2.9 ChIP sequencing

Discover unidentified protein binding regions and map them to your annotated reference. We perform:

- Mapping of your read to the reference
- Peak detection for enriched regions
- Create visualisation of peaks in a genome browser (Wiggle or BED format)
- Finding genes related to or flanking the enriched region
- Provide summary information of the loci (exon, intron, etc)
- Motif mapping and novel motif discovery

2.10 Copy Number Variation

GenomeScan is able to predict copy number variation based on Illumina reads in comparison with a reference sample. Using a set of dedicated tools, CNV regions can be identified and reported in a comprehensive table along with visualisation of the hit regions.

2.11 Functional annotation

Using a list of genomic positions all types of annotation columns and records can be added with relation to biochemical function, biological function, regulation and interactions, and gene expression. Functional annotation clustering , BioCarta & KEGG pathway mapping, GO, gene-disease association, homologue match, ID translation, literature match, etc. are examples of the many possible way the data set can be enriched.

2.12 Visualisation and annotation of sequence data within a genome browser

If you prefer to visualise your reads and browse through the sequence manually to inspect regions of interest or inspect SNVs and mutations, we can align the reads to your reference and transform the alignment into a file or track that can be loaded in a visualisation tool or one of the genome browsers. These software tools allow you to browse through the data providing you with annotations, other information and links to other resources. Alignments or assembly can be annotated with publicly available information and presented as a table that can be browsed.

Depending on the project a specific way of visualisation of the data can be provided that suits the need of the customer.

2.13 Custom data analysis

On some occasion one of our standard solutions for data analysis do not, or only partly suit your needs. To accommodate the needs of particular research questions, GenomeScan' team of bioinformaticians and molecular biologist can perform custom data analysis. We can validate the solution for your problem extensively before the actual data is analysed. Discuss your data analysis needs with scientists from our bioinformatics team to see if they can present a solution to you.

Chapter 3 Sequencing platforms

3.1 Illumina/Solexa

Sequencing data generation

Sequencing is performed on the Illumina HiSeq sequencers. The machine produces raw data in the form of intensity files that are the input for the analysis pipeline. This pipeline, and its output, is described in this section.

After the Illumina sequencer generates the sequencing images, the entire processing pipeline consists of three steps: image analysis, base calling, and sequence analysis.

- First step: Image analysis—Uses the raw files to locate clusters on the image, sharpens and enhances clusters through image filtering, removes background noise, detects clusters based on morphological features on the image, and outputs the cluster intensity, X,Y positions, and an estimate of the noise for each cluster. The output from image analysis provides the input for base calling. Image analysis is performed by the instrument control software's Real Time Analysis (RTA).
- Second step: Base calling—Deconvolves cluster intensities and noise estimates and applies correction for cross-talk, phasing, and prephasing. It outputs the sequence of bases read from each cluster, a confidence level for each base, and whether the read passes filtering. Base calling is performed by the instrument control software's Real Time Analysis (RTA) or the Off-Line Basecaller (OLB).
  - *Frequency cross-talk*—The Genome Analyzer uses two lasers and four filters to detect four dyes attached to the four types of nucleotide, respectively. The emission spectra of these four dyes overlap so that the four images are not independent. The frequency cross-talk is deconvolved using a frequency cross-talk matrix.
  - *Phasing/Prephasing*—Depending on the efficiency of the fluidics and chemistry of the sequencing reactions, a small number of molecules in each cluster may run ahead of (prephasing) or fall behind (phasing) the current incorporation cycle (Fig. 1). This effect is mitigated by applying corrections during the base calling step.
- Third step: The third step – data analysis – exists of several steps.
  - *Bcl conversion*—Converts *.bcl files into *.fastq.gz files (compressed FASTQ). *.bcl files are the primary sequence input. Multiplexed samples are demultiplexed during this step. The output is organised in Project and Sample folders (based on the sample sheet).
  - *Demultiplexing*— Multiplexed sequencing allows you to run multiple samples per lane. The samples are identified by index sequences (barcodes) that are attached to the template during sample preparation. Multiplexed sequencing allows you to run up to 96 individual samples in one lane, for a total of 384 samples. The samples are identified by an index sequence (barcode) that was attached to the template during sample preparation.
  - *Processing of the FASTQ-files*-The FASTQ-files are separated from each other over the different sample folders. After that, per sample, an statistics report is generated as well as a summary of the flowcell statistics. An example of the graphs that are generated in these reports can be found below.

Sequence quality

During the sequencing run, quality statistics are collected. This is summarised in the run folder under Data/Status_files/Summary.htm. In this file there are 6 tabs: Run info, Tile status, Charts, Summary, Cluster Density and Data by cycle.

Characteristics of the data

One of the characteristics typical for Illumina data is the decreased quality closer to the 3' end. A number of factors can cause the quality of base calls to be low at the end of a read. For example, phasing artefacts can degrade signal quality in some reads, and the affected portions of these reads have high error rates and unreliable base calls. Typically, the increase in phasing causes quality scores to be low in these regions, and thus these unreliable bases are scored correctly. However, the occurrence of phasing artefacts may not always correlate with segments of high miscall rates and biased base calls, and therefore these low quality segments are not always reliably detected by our current quality scoring methods. Illumina therefore marks all reads that end in a segment of low quality, even though not all marked portions of reads will be equally error prone. The read segment quality control metric identifies segments at the end of reads that may have low quality, and unreliable quality scores. If a read ends with a segment of mostly low quality (Q15 or below), then all of the quality values in the segment are replaced with a value of 2 (Q2), while the rest of the quality values within the read remain unchanged. Illumina flags these regions specifically because the initially assigned quality scores do not reliably predict the true sequencing error rate. This Q2 indicator does not predict a specific error rate, but rather indicates that a specific final portion of the read should not be used in further analyses.

Output file formats

Data from the Illumina HiSeq sequencer is provided in Sanger FASTQ format. Files are filtered.

Chapter 4 File Formats

4.1 GNU zip

The FASTQ sequence files output by the Illumina sequencers are saved compressed in the commonly used GNU zip format. This is indicated by the .gz file extension. Most downstream data analysis tools automatically decompress the files when used as input as well a most decompression software packages can inflate this format.

4.2 Sequence files

Sanger FASTQ

The Sanger FASTQ sequence format is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. It usually has the extension .fq or .fastq. Quality score are encoded with a single ASCII character. Each sequence entry in a FASTQ file consists of four lines in the following order:

Table 2. FASTQ file layout.

Line	Requirements	Description
1	@ + text	Sequence identifier (starting with a @)
2	DNA sequence	Sequence
3	'+' + text	Quality score identifier (starting with a +)
4	Text	Quality score

An example of a valid entry is as follows:
@contig000001
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
+contig000001
BBBBCCCC?<A?BC?7@@???????DBBA@@@@A@@

The quality score identifier may be repeated from the sequence header or abbreviated to '+' only. Quality scores are encoded into a compact form in FASTQ files which uses only one byte per quality value. In this encoding the quality score is represented as the character with an ASCII code equal to its value + 33. The following Table 3 demonstrates the relationship between the encoding character, the ASCII code of the character, and the quality score represented.

Table 3. ASCII table and corresponding quality scores

Symbol	ASCII value	Q-score	Symbol	ASCII value	Q-score	Symbol	ASCII value	Q-score
!	33	0	/	47	14	=	61	28
“	34	1	0	48	15	>	62	29
#	35	2	1	49	16	?	63	30
$	36	3	2	50	17	@	64	31
%	37	4	3	51	18	A	65	32
&	38	5	4	52	19	B	66	33
'	39	6	5	53	20	C	67	34
(	40	7	6	54	21	D	68	35
)	41	8	7	55	22	E	69	36
*	42	9	8	56	23	F	70	37
+	43	10	9	57	24	G	71	38
,	44	11	:	58	25	H	72	39
-	45	12	;	59	26	I	73	40
.	46	13	<	60	27	J	74	41

Illumina FASTQ (v1.9)

As of CASAVA v1.8/OLB v1.9 Illumina has structured the information in their FASTQ files in the following way. Each sequence identifier, the line that precedes the sequence and describes it, is in the following format:

@<instrument>:<run number>:<flowcell ID>:<lane>:<tile>:<xpos>:<y-pos> <read>:<is filtered>:<control number>:<index sequence>

The elements are described in Table 4.

Table 4. Illumina v1.9 FASTQ sequence header elements.

Element	Requirements	Description
@	@	Each FASTQ sequence identifier line starts with @
<instrument>	Characters allowed: a-z, A-Z, 0-9 and underscore	Instrument ID
<run number>	Numerical	Run number on instrument
<flowcell ID>	Characters allowed: a-z, A-Z, 0-9	
<lane>	Numerical	Lane number
<tile>	Numerical	Tile number
<x_pos>	Numerical	X coordinate of cluster
<y_pos>	Numerical	Y coordinate of cluster
<read>	Numerical	Read number. 1 can be single read or read 2 of paired-end
<is filtered>	Y or N	Y if the read is filtered, N otherwise
<control number>	Numerical	0 when none of the control bits are on, otherwise it is an even number
<index sequence>	ACTG	Index sequence

Illumina FASTQ (prior to v1.9)

Prior to CASAVA v1.8/OLB v1.9, Illumina uses a non-Sanger FASTQ format with different quality score encoding. Quality values are calculated by subtracting an offset value of 64 instead of 33 for Sanger encoding. Each sequence identifier, the line that precedes the sequence and describes it, is in the following format:
@<machine_id>:<lane>:<tile>:<x_coord>:<y_coord>#<index>/<read_#>
The elements are described in Table 5.

Table 5. Illumina prior v1.9 FASTQ sequence header elements.

Element	Requirements	Description
@	@	Each FASTQ sequence identifier line starts with @
<machine_id>	Characters allowed: a-z, A-Z, 0-9 and underscore	Machine ID
<lane>	Numerical	Lane number
<tile>	Numerical	Tile number
<x_coord>	Numerical	X coordinate of cluster
<y_coord>	Numerical	Y coordinate of cluster
#<index>	0 or 1	0 means no index, 1 means indexed sample
/<read_#>	Numerical	Read number. 1 can be single read or read 2 of paired-end

An example of a valid entry is as follows:
@HWI-EAS216:1:2:1:2017#0/1
GCCATGCTCAGGAACAAAGAAACGCGGCACAGAATG
+HWI-EAS216:1:2:1:2017#0/1
a_aa`^aaaaa`_aa`YW`__a`__`___aa`____

Illumina sequence .txt format

These files contain sequences produced by the Illumina Bustard base caller represented as tab separated text files where each row corresponds to a single read from the sequencing machine. The first column is in the format lane:tile:x_coordinate_on_tile:y_coordinate_on_tile. The second column is the sequence corresponding to the read. The third column is the quality string in symbolic numeric format. There is a numeric score for each base, each separated by a single space. These values are non-alignment-normalized (aka raw) Solexa/Illumina (not Sanger) quality scores. The fourth column is a single boolean value where 'Y' means the read passed quality filtering, and 'N' means that it failed to pass the filter.

FastA format

A sequence in FastA format begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than ('>') symbol. The word following the '>' symbol is the identifier of the sequence, and the rest of the line is the description (both are optional). There should be no space between the '>' and the first letter of the identifier. It is recommended that all lines of text be shorter than 80 characters. The sequence ends if another line starting with a ">" appears; this indicates the start of another sequence.

An example of a file in FastA format:
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
There is no standard file extension for a text file containing FastA formatted sequences. The most commonly used are summarised in Table 6.

Table 6. Commonly used file extensions for FastA files

Extension	Description	Comments
.fa, .fasta, .seq, .fsa	Generic FastA	
.fna	Nucleotide FastA	For coding regions of a specific genome, use ffn, but otherwise fna is useful for generically specifying nucleic acids.
.ffn	FASTA nucleotide coding regions	Contains coding regions for a genome.
.faa	Amino acid FastA	Contains amino acids. A multiple protein fasta file can have the more specific extension .mpfa
.frn	FASTA non-coding RNA	Contains non-coding RNA regions for a genome, in DNA alphabet e.g. tRNA, rRNA

4.3 Alignment formats

SAM specification
The current definition of the SAM format is at http://samtools.sourceforge.net/SAM1.pdf.

BAM files
The BAM Format is a binary format for storing sequence data. BAM and SAM formats are designed to contain exactly the same information. The SAM format is more human readable, and easier to process by conventional text-based processing programs, such as awk, sed, cut, and so on. The BAM format provides the binary equivalent and is designed to be more compact, randomly accessible, and faster to manipulate. To allow fast random access to the data, many applications require BAM file to be indexed. This can be done with tools such as 'samtools'.

4.4 Feature/annotation files

GTF/GFF
The GFF (General Feature Format) format consists of one line per feature, each containing 9 columns of data, plus optional track definition lines. The GTF (General Transfer Format) is identical to GFF version 2. The documentation for GFF3 can be found at http://www.sequenceontology.org/gff3.shtml.

Genbank

GenBank format (GenBank Flat File Format) consists of an annotation section and a sequence section. The start of the annotation section is marked by a line beginning with the word "LOCUS". The start of sequence section is marked by a line beginning with the word "ORIGIN" and the end of the section is marked by a line with only "//".

EMBL

The EMBL flat format is a rich format for storing sequences and their associated meta-information, feature coordinates, and annotations. It shares details with the GenBank sequence format.

Notes

Appendix

Terms and definitions

SAM	Sequence Alignment/Map format
BAM	Binary compressed SAM format
FastA	Standard format for storing sequence data
FASTQ	Standard format for storing sequence data with quality information
GTF	General Transfer Format, tab-delimited format used to hold information about gene structure
GFF	General Feature Format, tab-delimited format used to hold information about gene structure

ServiceXS™ | GenomeScan B.V. | Plesmanlaan 1D | 2333 BZ Leiden | The Netherlands
Telephone: +31 (0)71 568 1050 | Fax: +31 (0)71 568 1055
info@genomescan.nl | info@servicexs.com | www.genomescan.nl | www.servicexs.com

Caring for your future