TraitMAGIC™ service

Bi parental Genetic map creation and QTL Analysis Service Package

Description:

high-resolution genetic map is the basis of advanced genomics applications such as assembly proofing, pseudo chromosome arrangement, trait associations and more. Utilizing heterozygous parents in a segregating population analysis poses a challenge since offspring contain up to four segregating haplotypes (two from each parent) and may need to be assigned to different linkage groups. NRGene has developed a unique approach that relies on the use of sequence-based haplotype markers that distinctively mark a single haplotype of a single parent. This service is designed for the analysis of bi parental mapping population of all types. The output of this analysis can be used to QC assemblies, construct chromosome level assemblies out of scaffolds and accurately map traits to specific genomic regions (QTL mapping). While this document mostly explains how the complex case of heterozygote F1 population are serviced it is also applicable to the simpler case of homozygous mapping populations (e.g. RILs, NILs, BCs, DH and more).

NRGene's unique service is an All-In-One package that saves the customer time and effort all the way from shipping the DNA samples to receiving QTL region for the trait of interest.

The service includes:

- DNA samples quality check (optional).
- Construction of genetic libraries (optional).
- Sequencing data production (optional).
- Sequencing data quality assurance procedures.
- Creation of a phased genetic map.
- Construction of Pseudo-chromosomes from existing scaffold level assembly made by NRGene (optional).
- QTL mapping associating haplotype to phenotype (optional).

Benefits:

This service utilizes an existing reference level assembly to provide the breeder with specific deliverables that are of high and unique value:

1. Dense coverage of informative dominant markers spanning the entire genome. These markers differentiate two parental samples from one another but can be used to differentiate other samples very efficiently. In a heterozygote cross the output markers will all be unique to a single parental haplotype.
2. For heterozygotes- Two genetic maps are provided, one for each parent, phasing all data into two distinct linkage groups per chromosome.
3. Complete phasing of the entire mapping population reconstructing the parental meiotic recombination events.
4. Based on the above- QTL mapping is done at the finest resolution representing the minimal biological units created by recombination. On the other hand, the False discovery rate is reduced extensively as the analysis is done not at the marker level but at the haplotype block level, providing better sensitivity and specificity at the same time.
5. As all the analysis is done on top of a physical map, drill down into gene content of mapped QTL regions is straight forward.

Input data (customer materials):

The customer is required to provide parental and progeny DNA or sequence data, phenotypic data for each genetic sample, as well as a chromosome level assembly. If a chromosome level assembly does not exist, NRGene can create a scaffold level assembly using DeNovoMAGIC technology and arrange the scaffolds into pseudo chromosomes using the genetic map as part of the deliverables.

The detailed list of customer materials is as follows:

Energin .R Technologies 2009 Ltd. (NRGENE), CN 514240779
5 Golda Meir St., Ness-Ziona 7403649 Israel | Tel: +972-77-414-4281 | Fax: +972-8-648-9897
info@nrgene.com | www.nrgene.com

1. Chromosome level reference assembly. This assembly will be used to give physical positions for all marker and QTL data. The parental produced sequences should be anchored to pseudo-chromosomes with at least 70% alignment of the haploid genome size (measured by 1C flow-cytometry of the assembled genome).
2. Parental genomic data: NRGene discovers markers distinguishing the two parents from one another. For this aim it can utilize 3 input types (either input can be used for one parent or both):
    a. A scaffold level assembly (preferably done by NRGene to ensure quality and compatibility) meeting these criteria:
        i. made from a specific individual used to parent the mapping population.
        ii. Heterozygous line must be phased (alleles differentiation between scaffolds)
        iii. Assembly size:
            1. For homozygote: should cover at least 80% of the haploid genome size (measured by 1C flow-cytometry of the assembled genome)
            2. For heterozygote: should cover at least 80% of the total genome size (measured by 2C flow-cytometry of the assembled genome)
    b. Raw ILMN Sequence data: At least X30 coverage of Illumina data (standard library type; PE150bp) as described in NRGene's standards [1].
    c. Extracted DNA meeting NRGene's requirements.
3. Progeny genomic data (at least 192 individuals of a segregating population) at least one of the following options:
    a. Raw Genotyping By Sequencing (GBS) data in accordance with NRGene's requirements.
    b. DNA samples- as described in NRGene's standards[2].
4. Phenotypic data: quantitative data describing each of the samples included. If the trait is not quantitative (e.g. resistant/ susceptible) a binary score or a discrete qualitative score (ranking each sample for the degree of resistance) can be used. The data should include the sample name and phenotypic score. Format:
   A tsv, csv or excel file with trait name, sample name/ID, discrete score. Should include data for all parental and progeny lines. Example:

Trait	Sample ID	Score

All digital data (GBS, sequencing data, assemblies) should be uploaded to a dedicated Amazon Web Services (AWS) S3 cloud bucket provided by NRGene and accompanied with QC report for GBS and Sequencing data (fastqc or equivalent).

Deliverables (output data):

GenoMAGIC™ genetic map construction service deliverables include:

1. Raw sequence data[3] .fastq files
2. Statistical report of the genetic map and QTL mapping results
3. Marker sequence, position, and phased genotype per parent
4. Two or one Genetic maps (for hetero or homo- zygous populations respectively) including haplotype blocks per progeny marking the sites of recombination.
5. Chromosome level .fatsa file (optional)
6. QTL mapping results, showing physical location, confidence intervals, trait effect and statistical significance per QTL (optional)

Specific description for each deliverable is given below.

[1] Requirements are detailed in the document titled "DeNovoMAGIC Sequencing data preparation-standard v10 050618"

[2] Requirements are detailed in the document titled "GBS and skimSeq DNA Extraction Standard-Quality and Handling Procedures v3.0 for 250918"

3 Provided the sequencing has been done by NRGene™.

Energin .R Technologies 2009 Ltd. (NRGENE), CN 514240779
5 Golda Meir St., Ness-Ziona 7403649 Israel | Tel: +972-77-414-4281 | Fax: +972-8-648-9897
info@nrgene.com | www.nrgene.com

2. Statistical report

The statistical report is generated once the analysis is complete. The report will include an overview of the project describing its content and motivations, followed by description of the methods used. The report will then show highlights of the different analyses and deliverables which may vary for a given service/ project. The different items which may be included are:

1. Marker discovery (variant calling).
2. Genetic map construction.
3. Pseudo chromosome ordering (optional).
4. QTL mapping (optional).

Each item will include broad statistics (e.g. number of markers) and some visualizations (usually screenshots from GenoMAGIC™ visualization software).

3. Markers (Parental variants)

The genetic map of each parent is generated by analyzing the segregation of parental variants in the progenies. The variants (markers) are provided in a tab delimited file described below. The files are slightly different in heterozygote or homozygote populations.

Heterozygote markers

A tab-delimited file for each parent is delivered:

P1.markers.txt
P2.markers.txt

The format of each file is (three rows are given as examples):

Chromosome	Start position	End position	Reference allele	Alternative allele	Phase
1	290	291	T	C	1
1	6997	6998	A	T	2
1	7307	7308	A	ATTT	0

The fifth column titled 'Phase' indicates which of the two parental haplotypes (designated '1' or '2') does this marker signify. In cases where the location is a homozygous region (where the two parental haplotypes are the same) the underlying markers will receive phase '0'.
The phasing of the parental line's markers is defined by the segregation among progenies. See details below in the genetic map deliverable.

Homozygote markers

For homozygote we provide a tab-delimited file for both parents with the variants used.

population.markers.txt

The format of the file is (three rows are given as examples):

Chromosome	Start position	End position	Reference allele	Alternative allele	Parent
1	290	291	T	C	1
1	6997	6998	A	T	2
1	7307	7308	A	ATTT	0

The fifth column titled 'Parent' indicates which of the two parental haplotypes (designated '1' or '2') does this marker signify. In cases where the location is a homozygous region (where the two parental haplotypes are the same) the underlying markers will receive phase '0'.
The phasing of the parental line's markers is defined by the segregation among progenies. See details below in the genetic map deliverable.

Energin .R Technologies 2009 Ltd. (NRGENE), CN 514240779
5 Golda Meir St., Ness-Ziona 7403649 Israel | Tel: +972-77-414-4281 | Fax: +972-8-648-9897
info@nrgene.com | www.nrgene.com

4. Genetic maps

The genetic map of each parent is provided in a Variant Call Format (VCF4) file. This format allows visual exploration of genetic maps using NRGene's adapted version of the Integrative Genomics Viewer (IGV) tool5. as briefly explained in the figure below.

Figure 1 shows a genome-wide segregation of the genetic maps of two parents, as shown in NRGene's adapted version of the IGV tool. Each parent has its own genetic map, divided into two colors (one for each chromatid/ haplotype). Each row in the figure represents a single progeny and the color indicates which haplotype it inherited in each locus (the horizontal axis reflects the physical position along the chromosome). Recombination events are reflected in positions where the haplotype color of a progeny switches between the two colors of one of the parents.

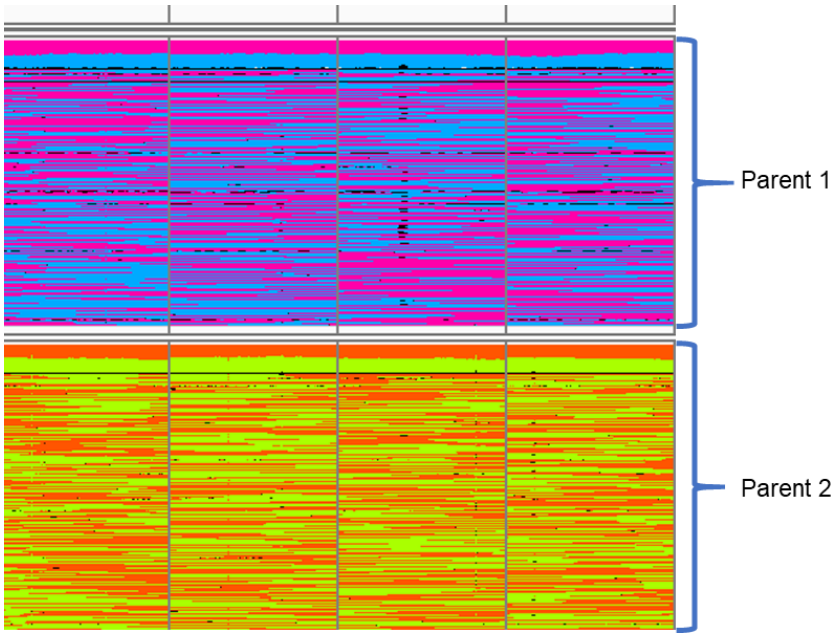

Figure 1 genome-wide genetic maps of two parental lines of a segregating population. Each row represents a different F1 progeny and the horizontal axis represents the physical coordinates. The four parental haplotypes are given distinct colors (pure blue and pure pink for parental line 1 and pure green and pure orange for parental line 2). The maternal and paternal maps of the same progenies are separated (each progeny is represented in both maps). Recombination events are clearly visible through the change of colors in each chromosome. Different chromosomes are separated with vertical lines (four autosomes are shown in the above example). Regions where the haplotype of a progeny cannot be determined are colored with black.

In addition to the visual exploration of the data, the representation of the genetic maps in a VCF file also allows exploration of the data in other ways (e.g., combine it with additional genomic information etc.), and thus allow for a more advanced downstream data analysis of the results.

The segregation of the two parental haplotypes in the progenies divides the entire genome into informative windows: Two consecutive regions are divided to separate and distinct informative windows if and only if the segregation of the haplotypes in the two regions is different (i.e., at least one progeny has two different haplotypes in the two consecutive regions). Thus, some progenies may have the same color between two consecutive windows, but at least one progeny

[4]General information on VCF format, which was developed to store DNA polymorphism data from various types (i.e., SNVs, INDELs, SVs), together with additional annotations about them, can be found here: https://doi.org/10.1093/bioinformatics/btr330; https://en.wikipedia.org/wiki/Variant_Call_Format

[5] (http://software.broadinstitute.org/software/igv/)

Energin .R Technologies 2009 Ltd. (NRGENE), CN 514240779
5 Golda Meir St., Ness-Ziona 7403649 Israel | Tel: +972-77-414-4281 | Fax: +972-8-648-9897
info@nrgene.com | www.nrgene.com

has switched color between the two windows (either because of recombination event or because its haplotype could not be determined in that region) and therefore they were split into different windows.
In a standard VCF files, each line represents a single genetic marker. NRgene's genetic map vcf follows this format to represent each informative window (as defined above) in a different line. While the full format of the VCF file is kept for compatibility (as explained above), not all fields are informative. Below are the tab-delimited columns of the genetic map VCF (similar to regular VCF), with a short description.

Field	**Description**
CHROM	Chromosome name
POS	Start position of the window on the chromosome (1-based)
ID	"." (not informative)
REF	"N" (not informative)
ALT	a comma separated list of possible haplotypes in the window. In each window, there are five possible haplotypes. Four of them represent the four possible parental haplotypes: two haplotypes of the first parental line (<HAP1>,<HAP2>) and two haplotypes of the second parental line (<HAP3>,<HAP4>). In addition, there is also <NAHAP> which indicates that the haplotype in the window is not similar to any of the two parental haplotypes (i.e., because of missing data, complex rearrangement event etc.).
QUAL	"1" (not informative).
FILTER	"PASS" (not informative).
INFO	This field provides information on the end of the window. It is in the format of: HS;END=<numeric end position of the variant in this record>, where HS stands for "Haplotype Similarity".
FORMAT	List of fields (separated with ":") describing the samples. In the genetic map VCF, the default format is: GT:CO1:LN, which reflects the following information on the samples: GT – Genotype. A numerical value which defines the haplotype of the sample (based on the list of haplotypes defined in the ALT field). CO1 – Color of the haplotype in RGB (in decimal format of ddd,ddd,ddd). LN – Length of the window in bp (numeric; identical for all samples for a given window)

Following these fields are *n* additional fields, each in the GT:CO1:LN format, one for each one of the *n* samples, describing its haplotype in that window.

In practice, we divide each sample to two phases, where each phase contains the markers the sample inherited from one parent. This is done also for the two parental lines. Each phase is represented alone.

For illustration, assume we have six samples (1-6): two parental lines (1,2), and four progenies (2,3,4,5,6).

After resolving the two different phases (haplotypes) of each of the parental lines, we divide the samples. Sample 1 (parental line 1) is divided into two samples: 1_p1 and 1_p2. Similarly, sample 2 (parental line 2) is divided to two samples: 2_p1 and 2_p2. In addition, each of the progenies (samples 3-6) are also divided into two samples. For example, sample 3 is divided into two samples: 3_p1 (which includes all the markers it inherited from parental line 1, either from 1_p1 or 1_p2 phases) and s_p2 (which includes all the markers it inherited from parental line 2, either from 2_p1 or 2_p2 phases).

In the genetic map of parental line 1 we present, in addition to the parental lines themselves (1_p1,1_p2,2_p1,2_p2), all the haplotypes that the progenies inherited from parental line 1 (3_p1,4_p1,5_p1,6_p1). Similarly, in the genetic map of parental line 2 we present, in addition to the parental lines themselves (1_p1,1_p2,2_p1,2_p2), all the haplotypes that the progenies inherited from parental line 2 (3_p2,4_p2,5_p2,6_p2).

Consider a genetic map VCF of parental line 1 in the above illustration. In this line we have information on the following samples: 1_p1,1_p2,2_p1,2_p2,3_p1,4_p1,5_p1,6_p1.

Energin .R Technologies 2009 Ltd. (NRGENE), CN 514240779
5 Golda Meir St., Ness-Ziona 7403649 Israel | Tel: +972-77-414-4281 | Fax: +972-8-648-9897
info@nrgene.com | www.nrgene.com

The header of the file will include the following line, which provides information on the format (as explained above), with the list of samples whose information is provided in the file. Consider one of the lines in the file that captures information on a single informative window:

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT 1_p1 1_p2 2_p1 2_p2 3_p1 4_p1 5_p1 6_p1

1 239801 . N <HAP1>,<HAP2>,<HAP3>,<HAP4>,<NAHAP> 1 PASS HS;END=242000 GT:CO1:LN 1|1:0,170,255:2200 2|2:255,0,170:2200 3|3:170,255,0:2200 4|4:255,85,0:2200 2|2:255,0,170:2200 1|1:0,170,255:2200 2|2:255,0,170:2200 1|1:0,170,255:2200

The values of each field in this line are as follows:

Field	**Value**
CHROM	1
POS	23801
ID	.
REF	N
ALT	<HAP1>,<HAP2>,<HAP3>,<HAP4>,<NAHAP>
QUAL	I
FILTER	PASS
INFO	HS;END=25200
FORMAT	GT:CO1:LN
1_p1	1|1:0,170,255:1400
1_p2	2|2:255,0,170:1400
2_p1	3|3:170,255,0:1400
2_p2	4|4:255,85,0:1400
3_p1	2|2:255,0,170:1400
4_p1	1|1:0,170,255:1400
5_p1	2|2:255,0,170:1400
6_p1	1|1:0,170,255:1400

This line describes a window in chromosome 1 in position 1:23801-25200 (total length: 1400 bps). There are five possible haplotypes: <HAP1>, <HAP2>,<HAP3>,<HAP4>,<NHAP>. <HAP1> (RGB color: 0,170,255; pure blue) represents the first phase of parental line 1 (as reflected in 1_p1 status) while <HAP2> (RGB color: 255,0,170; pure pink) represents the second phase of that parental line (reflected in 1_p2 status). Similarly, <HAP3> (RGB color: 170,255,0; pure green) represents the first phase of parental line 2 (as reflected in 2_p1 status) while <HAP4> (RGB color: 255,85,0; pure orange) represents the second phase of that parental line (reflected in 2_p2 status).
As this is the genetic map of parent 1, only phases p1 are presented for the progenies and their values can be <HAP1>, <HAP2> or <NHAP> (but not <HAP3> and <HAP4> which are reported in the genetic map of parental line 2). In this window, samples 4 and 6 (4_p1 and 6_p1) inherited the first phase of parental line 1 (<HAP1>) while samples 3 and 5 (3_p1 and 5_p1) inherited the second phase of parental line 2 (<HAP2>).

Energin .R Technologies 2009 Ltd. (NRGENE), CN 514240779
5 Golda Meir St., Ness-Ziona 7403649 Israel | Tel: +972-77-414-4281 | Fax: +972-8-648-9897
info@nrgene.com | www.nrgene.com

5. Chromosome level assembly

The final delivery may also include the construction of pseudo-chromosomes, given a scaffold level assembly done by NRGene is provided.

For Heterozygous genomes the delivery also includes analysis of complete scaffolds phasing.
The output will be a '.fasta' file where the number of sequences is the number of linkage groups (chromosomes). Unassigned scaffold will be under 'chromosome 0'.

Further analysis and genome-to-genome comparisons are also available by NRGene™'s PanMAGIC™ services or purchasing a license for NRGene™'s GenoMAGIC™ software package.

6. QTL analysis results

Based on the genetic maps above, QTL analysis is preformed, and calculated for each window the correlation between the phenotype and haplotype segregation among the progenies.
We provide the correlation values (r^2) in each region between the phenotypes and the segregation among the progenies for each parental genetic map as a bedGraph file:

QTL_ANALYSIS_P1.bedGraph
QTL_ANALYSIS_P2.bedGraph

In the following format:

Chromosome	Start position	End position	r^2

The BedGraph file can be viewed in a genome browser to see picks of statistical significance as well as detect possible false hits (e.g. due to assembly errors or structural variation between the reference used and the parental lines)
The statistical report will also contain a summary table for QTL found in this format (for heterozygotes):

QTL ID (parental origin)	Chr	Start	End	Peak start	Peak end	P-value	Average trait Score Phase1	Average trait Score Phase2
QTL-1 (Parent 1)	1	29654801	36555201	32795801	32802001	7.04E-07	0.68	0.34
QTL-2 (Parent 2)	1	31783201	33829401	32565801	32566001	6.81E-07	0.66	0.32
QTL-3 (Parent 2)	5	20230401	54668401	33785201	33787001	0.001046	0.38	0.61

The table denotes which heterozygote parent segregates in which location, sets the general locus interval and also the region of highest significance. A statistical p value (after correction to account for multiple comparisons) and also distinguishes which the trait effect of each phase.
A table listing homozygote populations will look as follows:

QTL ID	Chr	Start	End	Peak start	Peak end	P-value	Average trait Score Parent 1	Average trait Score Parent 2
QTL-1	1	29654801	36555201	32795801	32802001	7.04E-07	0.68	0.34
QTL-2	2	31783201	33829401	32565801	32566001	6.81E-07	0.66	0.32
QTL-3	5	20230401	54668401	33785201	33787001	0.001046	0.38	0.61

Energin .R Technologies 2009 Ltd. (NRGENE), CN 514240779
5 Golda Meir St., Ness-Ziona 7403649 Israel | Tel: +972-77-414-4281 | Fax: +972-8-648-9897
info@nrgene.com | www.nrgene.com

The phases used in this table are the same ones used in the 'marker.txt' deliverable. The customer can therefore extract phase or parent specific markers for the QTLs identified to design a trait marker assay for industrial application of Marker Aided Breeging.

Delivery Schedule:

- Raw sequencing data and QC report, for each DNA samples set or raw sequencing data set, shall be provided within 3 weeks from the sequencing data sets QC approval and within 10 weeks[6] from DNA samples QC approval.
- All other Deliverables are to be provided within 6 months[7] from **ALL** the raw sequencing data sets QA approval.

[6] Sequencing schedule is based on estimation and excludes any holidays. The period might extend due to workload limitations and subject to NRGene's instructions.
[7] Customary delivery time might be down to 4 months in certain cases and subject to NRGene's prior approval.

Energin .R Technologies 2009 Ltd. (NRGENE), CN 514240779
5 Golda Meir St., Ness-Ziona 7403649 Israel | Tel: +972-77-414-4281 | Fax: +972-8-648-9897
info@nrgene.com | www.nrgene.com