Booklet for the PiER

Contents

Chapter 1

Background

Figure 1.1: The logo for the web-based facilities - PiER that enables and automates genetics-led and network-based genetic target prioritisation.

Motivation

The field of target discovery has been advanced by genetics-led target prioritisation approaches. Integrative prioritisation for early-stage genetic target discovery has proven cost-effective in promoting the translational use of disease genetic associations, which is increasingly recognised in reducing drug attrition rate in late-stage clinical trials.

Design

Building on the verified Pi approach (see Nature Genetics 2019), here I introduce web-based servers/facilities called `PiER`. The PiER is free and open to all users and there is no login requirement, allowing the users to perform `ab initio` and `real-time` target prioritisation harnessing human disease genetics, functional genomics and protein interactions.

By analogy to the piano stave, the PiER consists of five horizontal lines, with three lines representing the elementary facility (`eV2CG`, `eCG2PG` and `eCrosstalk`),

each doing specific tasks on their own, and the rest two lines signifying the combinatory facility (`cTGene` and `cTCrosstalk`).

- eV2CG, linking variants to core genes; see Example Output
- eCG2PG, networking core genes to peripheral genes; see Example Output
- eCrosstalk, identifying the crosstalk between pathways; see Example Output
- cTGene, prioritising targets at the gene level; see Example Output
- cTCrosstalk, prioritising targets at the crosstalk level; see Example Output

Chapter 2

Facilities

The elementary facility supports three specific tasks, including three online tools: (i) `eV2CG`, utilising functional genomics to link disease-associated variants (including those located at the non-coding genome) to core genes likely responsible for genetic associations; (ii) `eCG2PG`, using knowledge of protein interactions to 'network' core genes with each other and with additional peripheral genes as well, producing a ranked list of core and peripheral genes; and (iii) `eCrosstalk`, exploiting the information of pathway-derived interactions to identify highly ranked genes that mediate the crosstalk between molecular pathways. By chaining together elementary tasks supported in the elementary facility, the combinatory facility enables the automation of genetics-led and network-based integrative prioritisation for genetic targets, both at the gene level (`cTGene`) and at the crosstalk level (`cTCrosstalk`). Notably, in addition to target crosstalk, the `cTCrosstalk` further supports target pathway prioritisation and crosstalk-based drug repurposing analysis (that is, repositioning approved drugs from original disease indications into new ones).

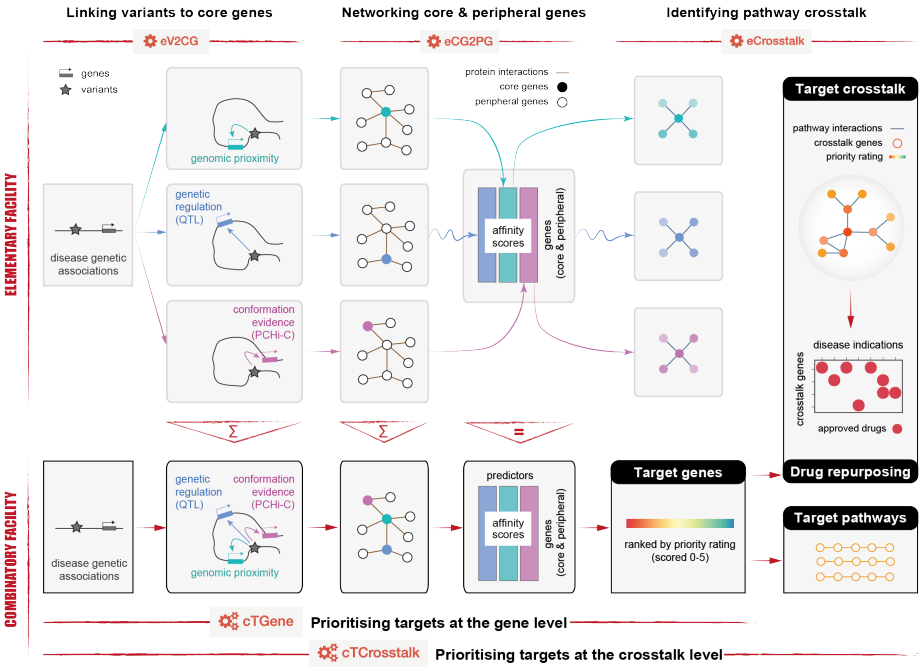

Figure 2.1: Schematic illustration of two facilities supported in the PiER.

Chapter 3

Compatibility

Table 3.1: A summary of the PiER website browser compatibility.

	MacOS (Big Sur)	Windows (10)	Linux (Ubuntu)
Safari	14.1.2	N/A	N/A
Microsoft Edge	N/A	85.0.564.67	N/A
Google Chrome	96.0.4664.110	90.0.4430.93	96.0.4664.110
Firefox	95.0.2	95.0.2	95.0.2

Chapter 4

Runtime

Table 4.1: A summary of the estimated runtime.

Facilities	Tools	Runtime (Server + Client)
Elementary	eV2CG	(67 + 82) seconds
Elementary	eCG2PG	(15 + 70) seconds
Elementary	eCrosstalk	(53 + 71) seconds
Combinatory	cTGene	(90 + 91) seconds
Combinatory	cTCrosstalk	(143 + 97) seconds

Chapter 5

Frontpage

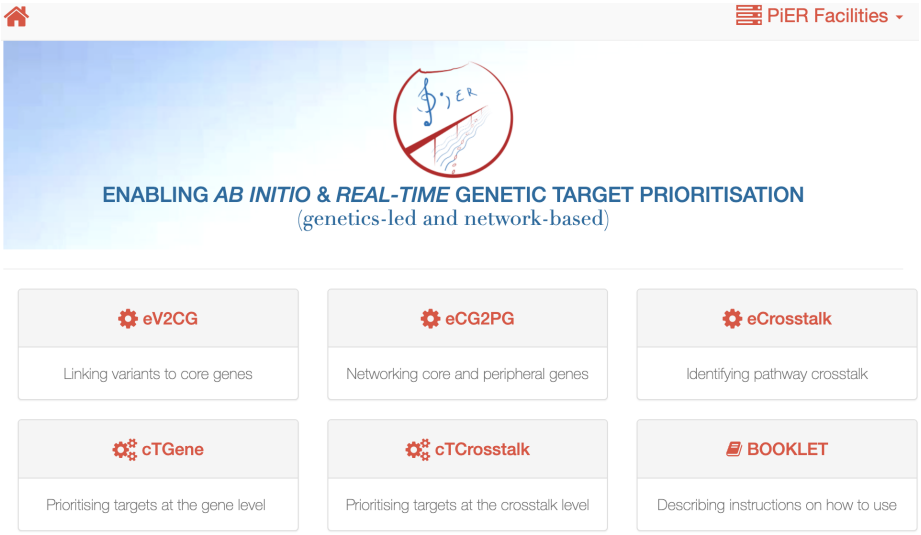

Figure 5.1: The landing frontpage of the PiER, featuring two facilities (elementary and combinatory). The elementary facility includes: (i) eV2CG, linking disease associated variants (particularly located at the non-coding genomic region) to (core) genes likely responsible for associations, based on either conformation evidence (that is, promoter-centered chromatin interactions), quantitative trait locus (QTL) mapping (that is, genetic regulation of gene expression or protein abundance), or simply genomic proximity; (ii) eCG2PG, using the knowledge of protein interactions to 'network' core genes to each other and to additional (peripheral) genes as well, generating a ranked list of core and peripheral genes; and (iii) eCrosstalk, exploiting the information of well-curated pathway-derived interactions to identify the subnetwork of highly ranked genes that mediate pathway crosstalk. Chaining together elementary functionalities above into pipelines provides the combinatory facility, enabling/automating genetics-led and network-based identification and prioritisation of drug targets: (iv) at the gene level (cTGene); and (v) at the crosstalk level (cTCrosstalk). Also included is the tutorial-like booklet (in an HTML format) describing step-by-step instructions on how to use.

Chapter 6

eV2CG

6.1 Interface

Input

- `Step 1`: a list of user-input SNPs, with 1st column for dbSNP rsIDs and 2nd column for significance info (p-values between 0 and 1). The error message will be displayed if the input is invalid. Example input data are shared genetic variants identified from cross-disease genome-wide association studies in inflammatory disorders; see Nature Genetics 2016.

Mechanism

- `Step 2`: includes SNPs in Linkage Disequilibrium (LD). By default, input SNPs with a typical threshold (p-value $<$ 5e$-$8) are considered, and additional SNPs in linkage disequilibrium (R2 $<$ 0.8) can be also included according to the European population.
- `Step 3`: uses genomic proximity, quantitative trait locus (QTL), or promoter capture Hi-C (PCHi-C) to identify core genes.
- `More Controls`: fine-tunes parameters involved in steps described above.

Output

- Example Output includes two interactive tables for core genes and evidence used, and a manhattan plot (illustrating scored core genes color-coded by chromosomes). A summary of input data and the runtime (computed on the server side) is also returned to the users for the reference.

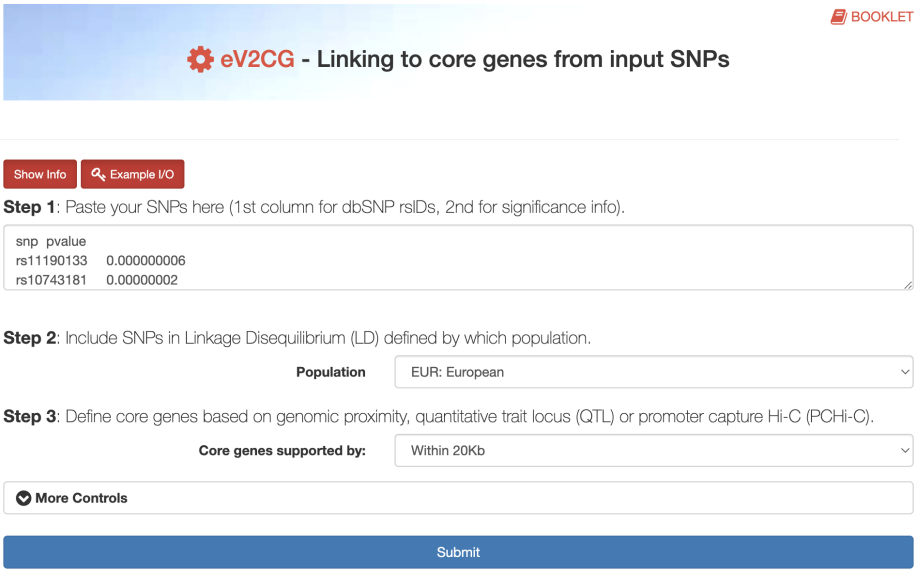

Figure 6.1: The interface of eV2CG, linking disease associated variants (particularly located at the non-coding genomic region) to (core) genes likely responsible for associations, based on either conformation evidence (that is, promoter-centered chromatin interactions), quantitative trait locus (QTL) mapping (that is, genetic regulation of gene expression or protein abundance), or simply genomic proximity. The Show/Hide Info toggle button contains the help information on how to use eV2CG, including input, output, mechanism, etc.

6.2 Linking results

- Under the tab `Output: core genes`, `Manhattan plot` illustrates scored core genes that are color-coded by chromosomes. Also provided is the downloadable PDF file.
- Under the tab `Output: core genes`, `An interactive table` lists core genes linked from the input SNPs, with scores quantifying the level of genes responsible for genetic associations (capped at 100). Genes are cross-referenced and hyperlinked to GeneCards. Also provided is the column Evidence used to define core genes.
- Under the tab `Output: core genes`, `Evidence table` for core genes, showing which SNPs (see the column `SNPs`) are used to define core genes (the column `Core genes`) based on which evidence (see the column `Evidence`). The column `SNP type` tells the SNP type (either `Input` for use-input SNPs or `LD` for LD SNPs). Notably, the column `Evidence` details datasets used: the prefix `Proximity_` indicative of SNPs in the proximity, the prefix `PCHiC_` for PCHi-C datasets, and the prefix `QTL_` for e/pQTL datasets.

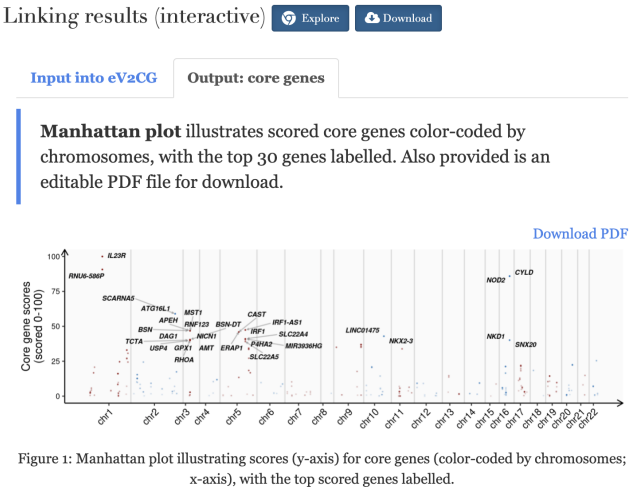

Figure 1: Manhattan plot illustrating scores (y-axis) for core genes (color-coded by chromosomes; x-axis), with the top scored genes labelled.

Figure 6.2: Interactive results for eV2CG. The user-input data are also returned for the exploration.

Chapter 7

eCG2PG

7.1 Interface

Input

- `Step 1`: a list of user-defined core genes, with 1st column for gene symbols, 2nd columns for weights (positive values), such as results from `eV2CG` above. The error message will be displayed if the input is invalid.

Mechanism

- `Step 2`: networks core genes with each other and with additional (peripheral) genes based on the knowledge of protein interactions, generating a ranked list of core and peripheral genes. It is achieved using the random walk with restart (RWW) algorithm. By default, the restarting probability of 0.7 is set, empirically optimised for immune-mediated diseases; selecting a value smaller than 0.6 is not recommended as there is a higher chance to expect low performance.
- `More Controls`: fine-tunes parameters involved in steps described above.

Output

- Example Output includes an interactive table for core and peripheral genes, and a manhattan plot (illustrating scores for genes color-coded by chromosomes). A summary of input data and the runtime (computed on the server side) is also returned to the users for the reference.

7.2 Networking results

- Under the tab `Output: core and peripheral genes`, `Manhattan plot` illustrates affinity scores for genes that are color-coded by chromosomes.

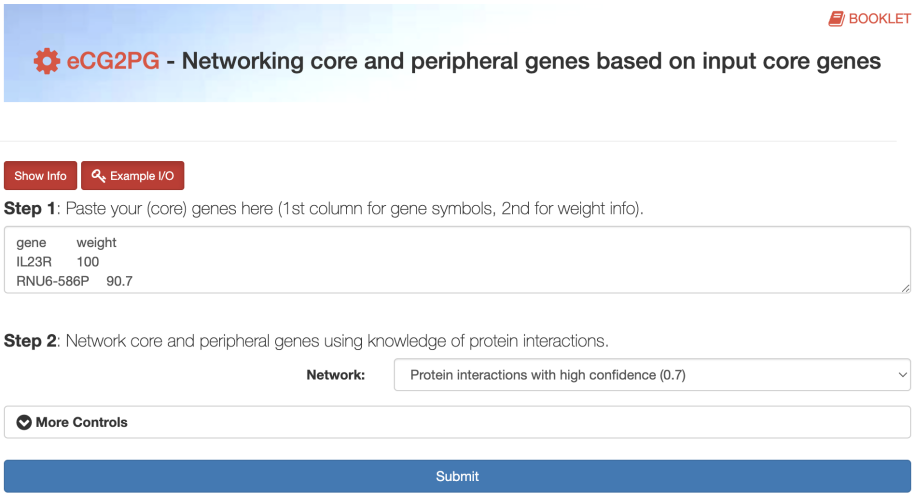

Figure 7.1: The interface of eCG2PG, using the knowledge of protein interactions to 'network' core genes to each other and to additional (peripheral) genes as well, generating a ranked list of core and peripheral genes. The Show/Hide Info toggle button contains the help information on how to use eCG2PG, including input, output, mechanism, etc.

Also provided is the downloadable PDF file.

- Under the tab `Output: core and peripheral genes`, `An interactive table` lists core and peripheral genes, with scores quantifying the affinity to core genes (sum up to 1). Genes are cross-referenced and hyperlinked to GeneCards.

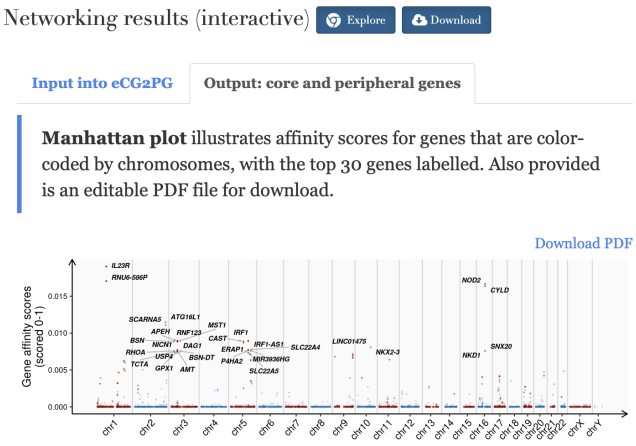

Figure 1: Manhattan plot illustrating scores (y-axis) for genes (color-coded by chromosomes; x-axis), with top scored genes labelled.

Figure 7.2: Interactive results for eCG2PG. The user-input data are also returned for the exploration.

Chapter 8

eCrosstalk

8.1 Interface

Input

- `Step 1`: a ranked list of genes, with 1st column for gene symbols, 2nd columns for scores (positive values), such as results from `eCG2PG` above. The error message will be displayed if the input is invalid.

Mechanism

- `Step 2`: identifies the subnetwork of highly-ranked genes that mediate the crosstalk between molecular pathways. The significance (p-value) of observing the identified crosstalk by chance is estimated by a degree-preserving node permutation test.

Output

- Example Output includes an interactive table for pathway crosstalk genes, and a network visualisation (illustrating the crosstalk between pathways).

8.2 Crosstalk results

- Under the tab `Output: pathway crosstalk`, `A network visualisation` illustrates crosstalk genes color-coded by input scores. The significance (p-value) of observing the identified crosstalk by chance is estimated by a degree-preserving node permutation test. Also provided is the downloadable PDF file.
- Under the tab `Output: pathway crosstalk`, `An interactive table`: lists crosstalk genes together with input scores. Genes are cross-referenced and hyperlinked to GeneCards.

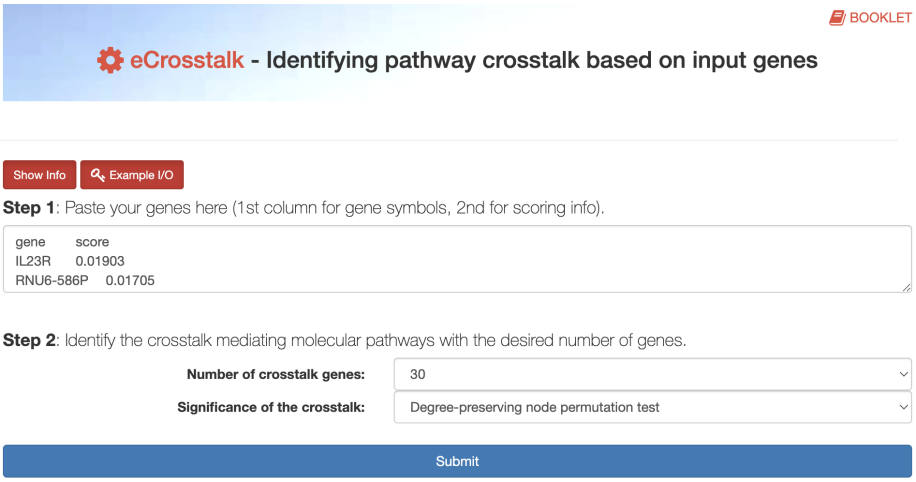

Figure 8.1: The interface of eCrosstalk, exploiting the information of well-curated pathway-derived interactions to identify the subnetwork of highly ranked genes that mediate pathway crosstalk. The Show/Hide Info toggle button introducing how to use eCrosstalk, including input, output, mechanism, etc.

Crosstalk results (interactive) Explore Download

Input into eCrosstalk | Output: pathway crosstalk

A network visualisation of pathway crosstalk, with genes/nodes color-coded by input gene scores. The significance (p-value) of observing the identified crosstalk by chance is 1.1e-37, as estimated by a degree-preserving node permutation test. Also provided is an editable PDF file for download.

Download PDF

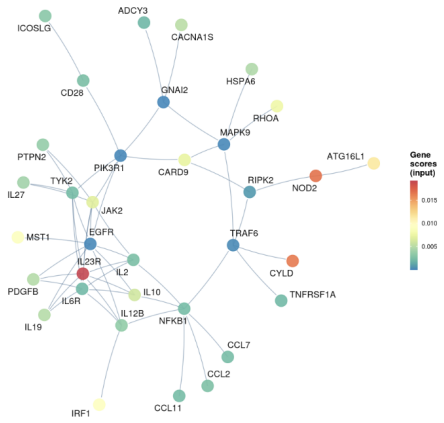

Figure 8.2: Interactive results for eCrosstalk. The user-input data are also returned for the exploration.

Chapter 9

cTGene

9.1 Interface

Input

- `Step 1`: a list of user-input SNPs, with 1st column for dbSNP rsIDs and 2nd column for significance info (p-values between 0 and 1). The error message will be displayed if the input is invalid. Example input data are shared genetic variants identified from cross-disease genome-wide association studies in inflammatory disorders; see Nature Genetics 2016.

Mechanism

- `Step 2`: includes SNPs in Linkage Disequilibrium (LD). By default, input SNPs with a typical threshold (p-value $<$ 5e$-$8) are considered, and additional SNPs in linkage disequilibrium (R2 $<$ 0.8) can be also included according to the European population.
- `Step 3`: uses functional genomic datasets, including genomic proximity, quantitative trait locus (QTL) and promoter capture Hi-C (PCHi-C), to identify core genes.
- `Step 4`: networks core genes with each other and with additional (peripheral) genes based on the knowledge of protein interactions, generating a ranked list of core and peripheral genes. It is achieved using the random walk with restart (RWW) algorithm. By default, the restarting probability of 0.7 is set, empirically optimised for immune-mediated diseases; selecting a value smaller than 0.6 is not recommended as there is a higher chance to expect low performance.
- `More Controls`: fine-tunes parameters involved in steps described above.

Output

- Example Output includes a manhattan plot (illustrating priority rating for target genes color-coded by chromosomes), and two tabular displays about prioritisation and evidence. A summary of input data and the runtime (computed on the server side) is also returned to the users for the reference.

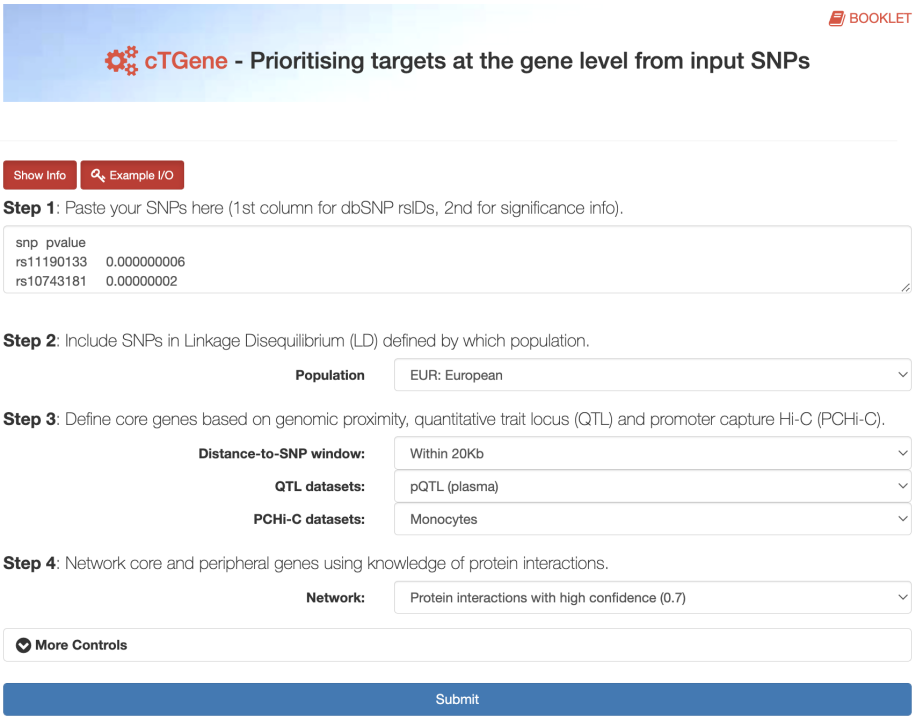

Figure 9.1: The interface of cTGene, enabling/automating genetics-led and network-based identification and prioritisation of drug targets at the gene level. The Show/Hide Info toggle button contains the help information on how to use cTGene, including input, output, mechanism, etc.

9.2 Prioritisation results

- Under the tab `Output: target genes`, `Manhattan plot` illustrates priority rating for target genes that are color-coded by chromosomes. Also provided is the downloadable PDF file.
- Under the tab `Output: target genes`, `Prioritisation table` lists all prioritised genes, each receiving 5-star priority rating (scored 0-5). Genes are cross-referenced and hyperlinked to GeneCards. The column `Type` tells the target gene type (either `Core` for core genes or `Peripheral` for peripheral genes). Also provided is a summary of evidence used to define

core genes, including columns `Proximity` (evidence of genomic proximity), `QTL` (e/pQTL evidence) and `PCHiC` (conformation evidence).

- Under the tab `Output: target genes`, `Evidence table` for core genes, showing which SNPs (see the column `SNPs`) are used to define core genes (the column `Core genes`) based on which evidence (see the column `Evidence`). The column `SNP type` tells the SNP type (either `Input` for use-input SNPs or `LD` for LD SNPs). Notably, the column `Evidence` details datasets used: the prefix `Proximity_` indicative of SNPs in the proximity, the prefix `PCHiC_` for PCHi-C datasets, and the prefix `QTL_` for e/pQTL datasets.

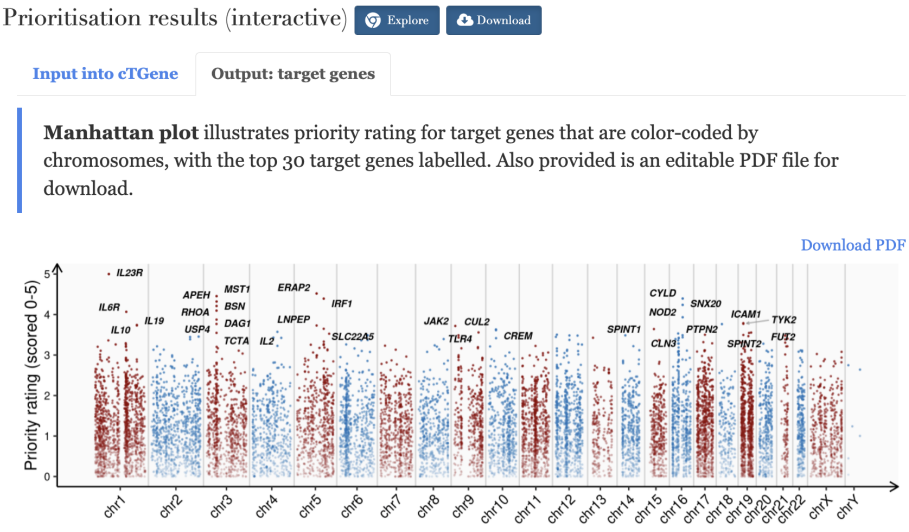

Figure 9.2: Interactive results for cTGene. The user-input data are also returned for the exploration.

Chapter 10

cTCrosstalk

10.1 Interface

Input

- `Step 1`: a list of user-input SNPs, with 1st column for dbSNP rsIDs and 2nd column for significance info (p-values between 0 and 1). The error message will be displayed if the input is invalid. Example input data are shared genetic variants identified from cross-disease genome-wide association studies in inflammatory disorders; see Nature Genetics 2016.

Mechanism

- `Step 2`: includes SNPs in Linkage Disequilibrium (LD). By default, input SNPs with a typical threshold (p-value $<$ 5e$-$8) are considered, and additional SNPs in linkage disequilibrium (R2 $<$ 0.8) can be also included according to the European population.
- `Step 3`: uses functional genomic datasets, including genomic proximity, quantitative trait locus (QTL) and promoter capture Hi-C (PCHi-C), to identify core genes.
- `Step 4`: networks core genes with each other and with additional (peripheral) genes based on the knowledge of protein interactions, generating a ranked list of core and peripheral genes. It is achieved using the random walk with restart (RWW) algorithm. By default, the restarting probability of 0.7 is set, empirically optimised for immune-mediated diseases; selecting a value smaller than 0.6 is not recommended as there is a higher chance to expect low performance.
- `Step 5`: identifies the subnetwork of highly-ranked genes that mediate the crosstalk between molecular pathways. The significance (p-value)

of observing the identified crosstalk by chance is estimated by a degree-preserving node permutation test.

- `More Controls`: fine-tunes parameters involved in steps described above.

Output

- Example Output includes target genes, target pathways, targets at the crosstalk level, and crosstalk-based drug repurposing. A summary of input data and the runtime (computed on the server side) is also returned to the users for the reference.

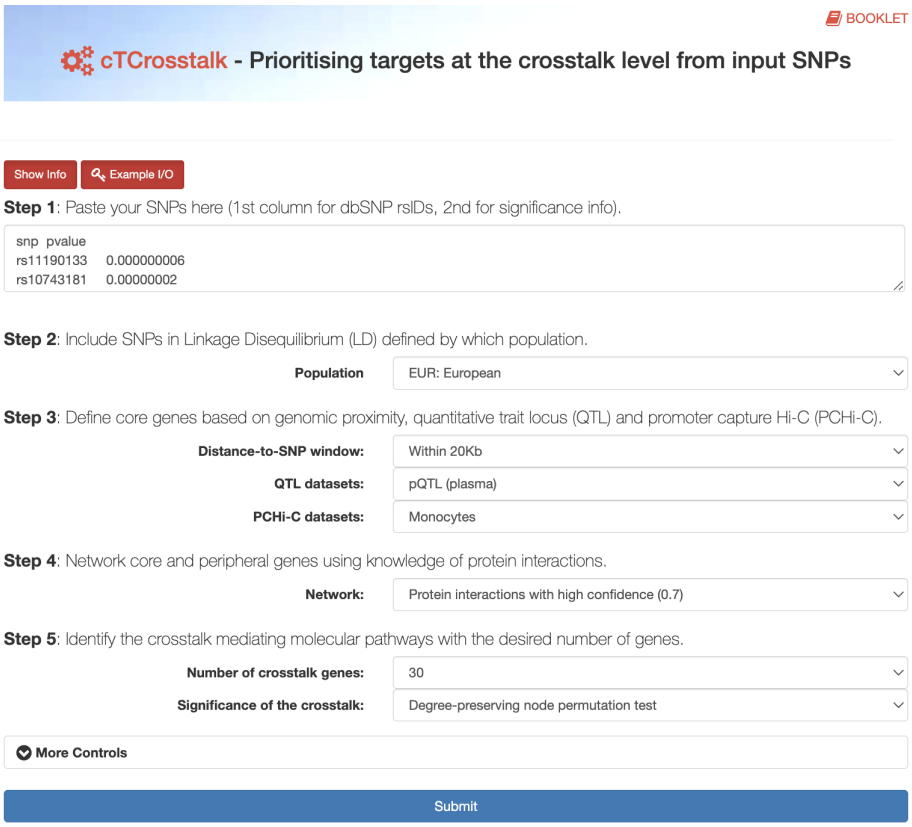

Figure 10.1: The interface of cTCrosstalk, enabling/automating genetics-led and network-based identification and prioritisation of drug targets at the crosstalk level. The Show/Hide Info toggle button contains the help information on how to use cTCrosstalk, including input, output, mechanism, etc.

10.2 Prioritisation results

- `Output: target genes`: includes `Manhattan plot` illustrating priority rating for target genes that are color-coded by chromosomes. Also provided is the downloadable PDF file. It also includes `Prioritisation table` listing all prioritised genes, each receiving 5-star priority rating (scored 0-5), and `Evidence table` for core genes showing which SNPs are used to define core genes based on which evidence. Genes are cross-referenced and hyperlinked to GeneCards.
- `Output: target pathways`: includes a dot plot and a prioritisation table for target pathways. Also provided is the downloadable PDF file.
- `Output: targets at the crosstalk level`: includes `A network visualisation` illustrating the crosstalk between pathways, with genes colored by priority rating and labelled in the form of `rating@rank`, `Prioritisation table` listing crosstalk genes, each receiving 5-star priority rating (scored 0-5), and `Evidence table` for pathway crosstalk genes, showing which SNPs are used to crosstalk genes based on which evidence. Genes are cross-referenced and hyperlinked to GeneCards.
- `Output: crosstalk-based drug repurposing`: includes `A heatmap-like illustration` showing drug repurposing analysis of approved drugs (licensed medications) based on pathway crosstalk genes, with crosstalk genes on y-axis, disease indications on x-axis, red dots indexed in number and referenced beneath in the table where the information on approved drugs and mechanisms of action is detailed. It also includes `An interactive table` of crosstalk genes (the column `Crosstalk genes`), disease indications (the column `Disease indications`), approved drugs and mechanisms (the column `Approved drugs [mechanisms of action]`), and drug index (the column `Index`) shown above within the dot plot.

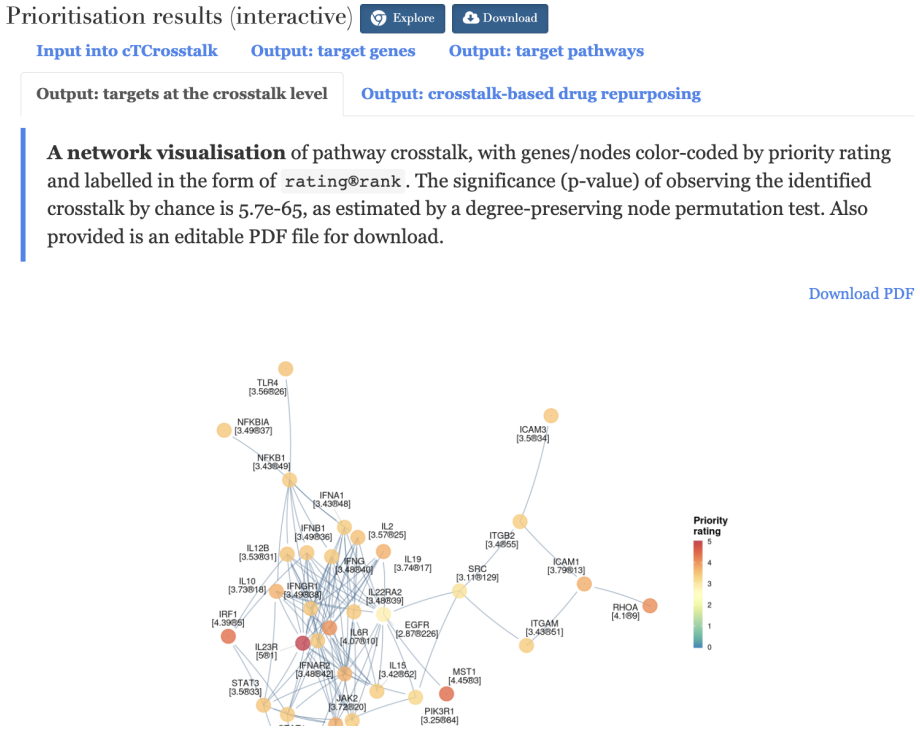

Figure 10.2: Interactive results for cTCrosstalk. The user-input data are also returned for the exploration.