A Computing System for Discovering Causal Relationships among Human Genes to Improve Drug Repositioning

Enrico Blanzieri[†‡], Toma Tebaldi[†§], Valter Cavecchia[‡], Francesco Asnicar[*], Luca Masera[†], Gabriele Tomè[†], Eleonora Nigro[*], Enrica Colasurdo[*], Matteo Ciciani[*], Chiara Mazzoni[*] and Stefania Pilati[¶]
[*]CIBIO, University of Trento, Italy [†]DISI, University of Trento, Italy [‡]CNR-IMEM, Trento, Italy [§]Yale University School of Medicine, New Haven, USA [¶]Fondazione Edmund Mach, San Michele all'Adige, Trento, Italy

Abstract—The automatic discovery of causal relationships among human genes can shed light on gene regulatory processes and guide drug repositioning. To this end, a computationally-heavy method for causal discovery is distributed on a volunteer computing grid and, taking advantage of variable subsetting and stratification, proves to be useful for expanding local gene regulatory networks. The input data are purely observational measures of transcripts expression in human tissues and cell lines collected within the FANTOM project. The system relies on the BOINC platform and on optimized client code. The functional relevance of results, measured by analyzing the annotations of the identified interactions, increases significantly over the simple Pearson correlation between the transcripts. Additionally, in 82% of cases networks significantly overlap with known protein-protein interactions annotated in biological databases. In the two case studies presented, this approach has been used to expand the networks of genes associated with two severe human pathologies: prostate cancer and coronary artery disease. The method identified respectively 22 and 36 genes to be evaluated as novel targets for already approved drugs, demonstrating the effective applicability of the approach in pipelines aimed to drug repositioning.

Index Terms—distributed volunteer computing, gene regulatory network expansion, BOINC, prostate cancer, coronary artery disease

1 Introduction

Bioinformaticians and computational biologists have been experiencing an increased need for computational resources in order to extract knowledge from the ever-growing amount of information of the -omics data produced by the latest high-throughput technologies. This knowledge can be represented by gene regulatory networks, a synthetic and convenient way of representing as graphs the functional interactions of the genes of an organism [1]. The representation abstracts away from the details of the actual underlying chain of events producing an interaction between two genes, and draws it as an edge between the corresponding nodes. Inferring gene regulatory networks from transcriptomic data is a wide and active research area [2] whose scope includes also the discovery of causal relationships between the transcription levels of the genes. However, causal discovery techniques are not among the popular tools for gene network reconstruction. For example, a recent and extensive review [3] covers different kind of models (information theory, Boolean, ODE, Bayesian and neural) and the corresponding tools, without reporting any causal discovery technique. The viability and effectiveness of discovery causal effects from transcriptomic data has been already shown at least one decade ago [4], but so far it lacks the attention that its theoretical soundness deserves.

The work of Judea Pearl [5] established causality as a fundamental concept of statistics and computing. The notion of causality is receiving a growing attention by the data mining community [6] in applications where the focus is the discovery of cause-effect relationships from observational data. The PC algorithm [7] tackles, among others, this problem and it can be applied on transcriptome data to infer causal relationships [8] where it promises to improve upon pure-correlation approaches such as WGCNA [9]. However, the worst-case complexity of the algorithm makes this approach not directly suitable for data of complex organisms with tens of thousands of genes. These problems can be mitigated by subsetting the variables and restricting the inference to the expansion of known gene regulatory networks of interest [10], [11].

In this paper we present the approach, methods and results of the gene@home project [12]. Developed within a collaboration of Trento University with FEM and IMEM-CNR, the project aims to expand gene networks using transcriptomic datasets with the support of voluntary computation on the TN-GRID platform [12] based on the BOINC system. The project involved so far two thousand volunteers and thousands of computers with a current estimated power of 14 teraFLOPS. In particular for human data we intend to provide a public resource to navigate and combine the results by expanding each single human transcript. Such resource can have a substantial impact on biological and medical research, as we make evident in two case studies.

The TN-GRID platform, which runs on a virtual server in the Data Center of the University of Trento, hosted the NES^{2}RA algorithm [10], [11] which has been already used to calculate gene network expansions in plants and microor-

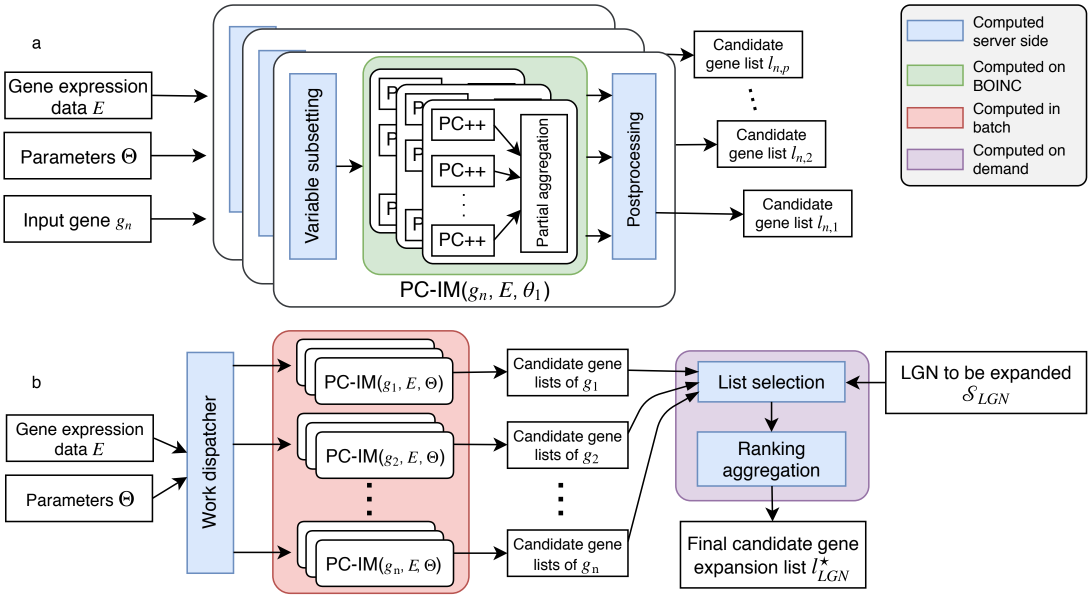


Fig. 1. Blocks scheme of the OneGenE architecture. Subfigure *a* shows the detail of the computation for a single gene g, while Subfigure *b* shows the overview of OneGenE, highlighting the two computational stages.

ganisms, such as *A. thaliana*, *E.coli* and *V. vinifera* [10], [13]. NES^{2}RA [11] outperformed the state of the art algorithm ARACNE [14] in the task of network expansion. Starting from a local gene network (LGN), based or hypothesized on previous biological knowledge, its expansion consists in a set of genes and a list of interactions which describe putative causal relationships with the genes in the LGN. The expansion is calculated on observational gene expression data, organized in a coherent normalized data matrix. Each expansion requires a few days to be carried out, even with the use of the BOINC distributed computing platform, thus presenting two main problems: the results cannot be provided to the user in real time and any expansion represents a unique elaboration, very unlikely to be submitted again given that the number of possible gene combinations is exponential in the size of a genome (5,000 - 30,000 genes).

To overcome these problems, we are now adopting an approach called OneGenE [15] which aims to expand each single gene in an organism. OneGenE's main idea is to calculate the list of gene expansions for each gene in an organism by systematically running single-gene NES^{2}RA expansions with fixed parameters, and then combine them afterwards to simulate LGN expansions. By doing so, these expansions can be used multiple times, thus effectively reducing the computational effort. With OneGenE, we will create a public database containing the expansions for each gene in an organism, thereby offering the possibility of building any LGN expansion in a very short time by combining the already calculated single expansions.

The infrastructure can be considered a socio-technical system that includes a server, the volunteers and their computers, and the biologists interested in the results. The communication of results at a scientific and popular level is essential both to expand the volunteer base and to establish a large and stable user base. To this end, we describe each application in the forum and pay particular attention to the requests of the volunteers.

The expansion of the human gene networks is based on the transcriptomic dataset provided by the FANTOM project [16]. FANTOM5 gene expression data come from sequencing of RNA extracted from 1,816 samples of different human tissues and cell lines and contains expression profiles of 201,802 gene isoforms (transcripts). It thus contains plenty of information on human gene transcriptional profiles in different biological contexts that can be exploited for data mining for different purposes. To our best knowledge ours is the first attempt to infer genome-scale regulatory information from FANTOM5 data. Few algorithms for network reconstruction scale to such large networks [3], and their benchmarking is usually done on far smaller networks. For example, a recent benchmark on networks of maximum 100 nodes [17] showed that ARACNE [14] (against which we already compared [11]) is still representative of the state of the art. The gold truth is necessary to compute the usual benchmarking metrics (precision, sensitivity, sensibility and ROC curves) but its absence prevents us from setting up big benchmarks with real data. Therefore we validate the results in terms of functional enrichment against public resources and present two in-depth case studies.

The two case studies here presented are focused on drug repositioning for two quite common pathologies, i.e. prostate cancer and cardiovascular diseases. Drug repositioning is an alternative approach for the discovery of new therapeutic opportunities for already approved medicines. Compared to traditional de novo drug development strategies, which have become increasingly expensive and time-consuming, this method, which relies on previous knowledge and speed up the approval procedure of the drug regulators, can represent a valuable approach. The biological opportunity of drug repositioning relies on one side on the fact that many diseases share common dysregulated pathways and proteins, and on the other side that medicines actually perturb multiple targets (off-targets interactions). Prostate cancer is currently one of the most common carcinomas

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TETC.2020.3031024, IEEE Transactions on Emerging Topics in Computing

among men, 3 million people are now affected by this disease in Europe [18]. The most common treatments include radical prostatectomy or radiotherapy, which often affect semen production and fertility. Other chemotherapy based medications exist, however most of them have serious side effects that drastically reduce the quality of life of patients. Therefore, drug repositioning could help identifying drugs with anti-cancer activity as well as reduced adverse effects for human health. Coronary artery disease (CAD), also known as coronary heart disease (CHD) or ischemic heart disease (IHD), is the most common type of heart disease. It is the leading cause of death in most of the Western countries as emerges from WHO data [19]. This condition leads to the formation of a waxy substance called plaque in the coronary arteries, decreasing the flow of oxygen-rich blood to the heart. CAD is a complex disease caused by multiple factors, including lifestyle and diet, thus different classes of drugs are actually used to treat it, among which aspirin is the most known one. Here, we present a data driven approach for drug repositioning based on NES^{2}RA ability to find genes causally-related to genes already known to be involved in these pathologies and propose them as novel drug targets for the treatment of CAD or prostate cancer.

The next two sections present method, input data, and the system. Section 4 is devoted to the validation of the OneGenE approach of single-gene NES^{2}RA expansions, whereas Sections 5 and 6 focus respectively on the application to drug repositioning for prostate cancer and coronary artery disease. The last section draws conclusions and future perspectives.

2 Method

OneGenE [15] is a method to compute ranked candidate gene lists that expands known local gene networks given gene expression data. As its predecessor NES^{2}RA, OneGenE (Algorithms 1 and 3) is based on the systematic and iterative application of the skeleton function of the PC algorithm (Algorithm 2) on subsets of the input data. OneGenE aims to overcome the large latency of NES^{2}RA when applied to LGNs, by pre-computing partial results, namely single-gene NES^{2}RA expansions, on the BOINC platform.

Figure 1 shows the block scheme of the OneGenE architecture [15], highlighting the different platforms and the two computational stages. First, a pre-computation step: candidate expansion lists are pre-computed for each gene (or more generally transcript) of the target organism exploiting the BOINC platform. Second, a ranking aggregation step: the user provides as input a set of transcripts of interest (LGN) and chooses a specific procedure to aggregate the intermediate results pre-computed by OneGenE.

Data and Input. The pipeline starts with Algorithm 1 whose input consists of a $n \times m$ gene expression data matrix E, where $n = |\mathcal{S}|$ is the number of transcripts $\mathcal{S}$ and m the number of samples, and a set of parameter tuples $\Theta = \{\theta\} = \{(\alpha, d, i) \mid \alpha \in A,\ d \in D,\ i \in I\}$ where A, D and I are respectively the sets of alpha values, tile sizes (namely subset sizes), and number of iterations.

Nested Loop. For each transcript g in $\mathcal{S}$, (Algorithm 1) p instances of PC-IM (depicted in Figure 1) are executed on the BOINC platform, where $p = |\Theta| = |A| \cdot |D| \cdot |I|$. The internal loop shown in Algorithm 1 is a special case of the NES^{2}RA Ranking Procedure, that receives as input-LGN a single gene g with probability vector $\Pi = \mathbf{1}$, i.e. the gene g is present in each subset. The internal loop comprises the subsetting of the transcripts, the run of PC-skeleton (Algorithm 2) on each subset, and the computation of absolute frequencies, relative frequencies, and the corresponding order of the candidates list.

Pre-computation Result. Unlike NES^{2}RA, in OneGenE the ranking aggregation is postponed. Each PC-IM returns a candidate expansion list $l_{g,\theta}$ for each tuple of parameters $\theta = (\alpha, d, i)$ corresponding to the set of lists resulting from Algorithm 1. The candidate gene expansion lists are stored on a local server (an example in Supplementary Material), and, once all the results have been computed, OneGenE is ready to be queried by the user with input LGNs.

```
Data: $S$ set of $n$ transcripts, $E$ $n \times m$ matrix of expression data.
Input: $I$ set of values of number of iterations, $D$ set of values
       of the subset dimension, $A$ set of values of the
       significance level $\alpha$
Result: a set of ordered lists of candidate transcripts
$L \leftarrow \emptyset$                    // $L$ set of ordered lists
foreach $g \in S$ do
    foreach $\theta = (\alpha, d, i) \in A \times D \times I$ do
        // NES$^2$RA Ranking Procedure (RP) [20]
           call
        // $l_{g,\theta}$ ← RP($S$, {$g$}, $E$,$\mathbf{1}$, $i$, $d$, $\alpha$)
        // equivalent to:
        foreach $j \leq i$ do
            Randomly generate a minimal collection of
              subsets of dimension $d$ of $S$ such that $g$ is in
              every subset and each transcript is in at least one
              subset
        end
        foreach subset do
            Run the PC-skeleton function Algorithm 2 [21],
              [20] on the expression data $E$ restricted to the
              transcripts of the subset and generate a network.
        end
        foreach $\gamma$ adjacent to $g$ in the networks do
            // compute absolute frequency
            $f_\gamma \leftarrow$ #networks s.t. $\gamma$, $g$ are adjacent
            // compute relative frequency
            $f'_\gamma \leftarrow f_\gamma$/(#subsets that contain $\gamma$)
        end
        $l_{g,\theta} \leftarrow$ genes ordered with respect to $f'_\gamma$
        // NES$^2$RA RP ends, $g$ is in each subset
           with probability 1.
        // In this case RP takes the name PC-IM
           [22] and the call above is written
           $l_{g,\theta}$ ← PC-IM ($S$,{$g$},$E$,$i$,$d$,$\alpha$)
        $L \leftarrow L \cup l_{g,\theta}$
    end
end
return $L$
```

Algorithm 1: OneGenE: Pre-computation Step.

Ranking or list aggregation. Let $\mathcal{S}_{LGN}$ be the set of transcripts in an input LGN and $l_{g,\theta}$ is the candidate expansion list of the gene g with the parameter tuple θ. The final candidate gene expansion list is obtained by combining the set of partial results $\mathcal{L} = \{l_{g,\theta} \mid g \in \mathcal{S}_{LGN},\ \theta \in \Theta\}$ by means of a ranking aggregator [15]. Algorithm 3 shows possible alternatives (threshold on the relative frequency and fixed or variable cut-offs on the rank) that are relevant for the two case studies (Sections 5 and 6) where the expanded lists intervene in rather complex workflows. The higher the relative frequency the harder is to explain the correlation

```
Data: T, Set of transcripts, $E$ expression data
Input: Significance level $\alpha$
Result: An undirected graph with causal relationship between
        transcripts
Graph $G \leftarrow$ complete undirected graph with nodes in T;
$l \leftarrow -1$;
while $l < |G|$ do
    $l \leftarrow l + 1$;
    foreach $\exists u, v \in G$ s.t. $|Adj_G(u) \setminus \{v\}| \geq l$ do // $Adj_G(u)$
      adjacent nodes of u in G
        if $v \in Adj_G(u)$ then
            foreach $A \subseteq Adj_G(u) \setminus \{v\}$ s.t. $|A| = l$ do
                if $u, v$ are conditionally independent given $A$
                  w.r.t. $E$ with significance level $\alpha$ then
                    remove edge $\{u, v\}$ from G;
                end
            end
        end
    end
end
return G;
```

Algorithm 2: PC Algorithm: skeleton procedure [21].

```
Data: $L = \{l_{t,\theta}\}$ set of ordered lists of candidate transcripts
Input: $f'_{min}$ minimum relative frequency or $k$ maximum
       length of the lists or $K = \{k_t\}$ set of maximum lengths
       of each list
Result: ordered list of candidate transcripts
while true do
    $\mathcal{L}_{temp} \leftarrow \emptyset$ // $\mathcal{L}_{temp}$ set of ordered lists
    User enters/selects/edits $T$ set of transcripts
    // lists selection
    foreach $l_{t,\theta} \in L$ such that $t \in T$ do
        $\mathcal{L}_{temp} \leftarrow \mathcal{L}_{temp} \cup \{l_{t,\theta}\}$
    end
    // lists trimming
    Case frequency: $\mathcal{L}_{temp} \leftarrow$ fre($\mathcal{L}_{temp}$, $f'_{min}$) // trim
        each list in $\mathcal{L}_{temp}$ at threshold $f'_{min}$
    Case top k: $\mathcal{L}_{temp} \leftarrow$ top($\mathcal{L}_{temp}$, $k$) // trim each
        list in $\mathcal{L}_{temp}$ to the first $k$ elements
    Case top variable $k_t$:$\mathcal{L}_{temp} \leftarrow$ top($\mathcal{L}_{temp}$, $K$) // trim
        each list $l_{t,\theta} \in \mathcal{L}_{temp}$ to the first $k_t \in K$
        elements
    // final lists aggregation
    output $l^*_T$=list_aggregation($\mathcal{L}_{temp}$) // the final
        relative frequency in $l^*_T$ is computed as
        the mininum or the average
end
```

Algorithm 3: OneGenE: List Aggregation

between two genes in terms of other genes and this provides evidence of a putative direct causal relationship. Ranks, on the other hand, are useful for comparison between lists and prioritization.

2.1 Data and running parameters

The human transcriptome data used in this work have been downloaded from the repository of the FANTOM5 project (http://fantom.gsc.riken.jp/5/). FANTOM is an international research consortium that generates and shares high-quality transcriptome datasets (CC BY 4.0). The FANTOM5 dataset was generated by RNA sequencing using single molecule CAGE (Cap Analysis Gene Expression)[16]. Normalized expression values are estimated as transcripts per million (TPM). The raw FANTOM5 dataset amounts to 1829 samples, encompassing human cell lines (271), primary cells (564) and tissues (188) also part of time course experiments (785) and fractionations/perturbations (21). FANTOM5 identified 201802 distinct genomic transcription start site (TSS) locations, corresponding to bona-fide transcript isoforms. A first filtering has been applied on the dataset to exclude unknown transcripts, i.e. without an annotated HGNC symbol (https://www.genenames.org/). This step resulted in a collection of 87554 transcripts, associated with 18889 genes, constituting our full version of the dataset (FANTOM-full). In order to remove possible sources of redundancy and noise in the data, additional filters were applied to remove transcripts with absent or low expression values in almost all samples, and transcript isoforms corresponding to the same genetic locus and showing highly correlated expression profiles across the FANTOM dataset (> 0.7). After the application of these filters, 49727 transcripts, associated with 16356 genes, were retained in the small version of the dataset (FANTOM-small). The single-gene NES^{2}RA expansions of FANTOM-small were submitted on the BOINC platform, with a tile size of 1000 transcripts, 1000 iterations and 0.05 as alpha threshold. With the same parameters OneGenE is currently running the single-gene NES^{2}RA expansions on FANTOM-full.

For the analysis presented in Case Study 1, an additional dataset has been used, retrieved from the Tumor Cancer Genome Atlas (TCGA) [23], one of the most comprehensive patient-derived cancer databases collecting data about 33 different cancer types. The RNA expression dataset from the Prostate Adenocarcinoma (PRAD) project contains 551 samples and the expression values of 60,483 transcripts obtained by RNASeq, expressed in Fragments Per Kilobase Million (FPKM) [24]. Data was grouped inside a matrix and then filtered according to these criteria: i) samples having an average gene expression over the 0.975 quantile were discarded; ii) genes not expressed in more than 500 samples were discarded. The resulting dataset (TCGA-PRAD-s) contains 43128 transcripts for 537 samples. The single-gene NES^{2}RA expansion were submitted on the BOINC platform, with a tile size of 2000 transcripts, 2000 iterations and 0.05 as alpha threshold.

3 System and Implementation

Since year 2014 CNR-IMEM, in collaboration with the University of Trento, has run TN-Grid, a computing infrastructure based on the BOINC platform. TN-Grid is hosting gene@home, a project developed with the Edmund Mach Foundation [12] with the goal to expand genetic regulatory networks with putative causal relationships by analyzing gene expression data, using the NESRA and NES^{2}RA algorithms. The gene@home project is now also running OneGenE.

3.1 Performance

As shown in Section 2, the OneGenE application running on the gene@home BOINC server, has the following main sets of parameters: I (number of iterations), D (the subset dimensions, i.e. the tile sizes), and A (the set of α to be used in the statistical test of the PC algorithm). These parameters need to be carefully chosen by balancing the execution speed of the application, the accuracy of the results and the statistical errors, and their values depend on the expression dataset in input.

TABLE 1
Optimization history of the OneGenE application

version	SIMD	Dataset	Time (s)	Relative gain
0.09	—	*FANTOM-full*	101.93	—
0.10	sse2	*FANTOM-full*	37.00	2.75
0.11	sse2	*FANTOM-full*	23.37	1.58
1.10	sse2	*FANTOM-full*	15.19	1.54

Additionally other two parameters (*n_pc* and *cut_off*) have to be set in order to optimize server and client performance and the overall network bandwidth. The parameter *n_pc* is the number of PC algorithm executions collected in a single workunit, namely the BOINC unit of work that is distributed to the volunteers. The parameter *cut_off*, whose effect is shortening the size of the output file, is a threshold that controls the removal of the lesser present interactions from the ranked output list of a workunit.

Due to the large running times expected for completing a OneGenE run, it is crucial to identify the optimal parameter values before the beginning of the run and carefully monitor and adjust them in the very early stages of the computational experiment, as it will be seen in Section 3.2.

The OneGenE experiments are intrinsically slow. The core application, initially written in the R language, was first rewritten in C++ (obtaining a substantial execution speed increase), adapted to BOINC using its API, compiled for different computing platforms (Windows x32/x64, Linux x32/x64, Mac OS and Linux ARM), and made publicly available[1] to the volunteers and to anyone else. One of the BOINC volunteers (Daniel Frużyński, Motorola Solutions Systems Polska Sp.z.o.o), in his spare time, made significant performance improvements to the code, profiling it using Callgrind (Valgrind/Linux) [25], by

- removing two top bottlenecks in code (range check, better I/O performance);
- rewriting of the correlation function, focusing on unaligned load/store instructions and unnecessary memory writes;
- templated versions of most performance-critical functions;
- adding SIMD (Single instruction, multiple data), SSE-AVX-FMA hardware-related optimization;
- using Gray code (reflected binary code, RBC, [26]) for generating combinations (version 1.0).

An example of the achieved relative speed-ups, for versions 0.09-1.10 of the application on the *H. sapiens* FANTOM-full dataset ($\alpha = 0.05$, tile size = 1000) are shown in Table 1; version 1.10 is the one currently running on gene@home.

3.2 System benchmarks and tuning

Among the parameters of a BOINC workunit, an important one is the estimate of the computing time, which allows the server to perform efficient scheduling and assign the so-called *credits* (virtual rewards for the volunteers) in a fair way. If a volunteer participates in many BOINC projects, appropriate compute time estimate allows the BOINC client to better choose its scheduling parameters (cache, priority, resource shares), thus minimizing deadline misses. The calculation is usually done by running a small number of randomly generated PC algorithm executions on a benchmark machine, getting the execution time and calculating the needed FLOPS with the formula $running\_time * host\_flops * scale\_factor$.

The reference machine used for the benchmark is an Intel I7-4770k workstation running Linux, with 8GB RAM, hyperthreading enabled, and a theoretical computational power of 4374.07 MFLOPS and 16809.68 MIPS (average values given by the BOINC client using standard Whetstone/Dhrystone synthetic benchmark system). Benchmarks were run using only one thread, keeping the others free. The *Time* column of Table 2 shows the averaged time needed for completing five single PC++. The *ETA* column (calculated by assuming 1000 as number of iterations) gives an estimate of the time needed for completing a OneGenE pre-computation step for all $g \in \mathcal{S}$ using all the eight threads of the reference machine.

The speed of our flow-controlled *work generator* (one of the BOINC server key components, the one that actually builds the workunits, written in the Python language) depends only on the number of transcripts of the dataset and the *tile size*.

In the specific case of OneGenE on the FANTOM-full dataset the workunit *n_pc* parameter was set to the value of 600, this number was carefully chosen to minimize network bandwidth and computational errors (achieving relatively fast redistribution in such cases). The time requested for processing a single *n_pc* on the reference computer is ≈ 15.19 sec (as seen in Table 2). The workunit execution time is therefore $15.19 \cdot 600 = 9114$ sec (≈ 2.5 hours). The expansion of a transcript becomes made up by 294 workunits, to be sent twice (for homogeneous redundancy).

The FANTOM-full dataset size (compressed with gzip) is ≈ 132 Mb, this is a *sticky* file, i.e. it remains on the client after job is finished, it could be reused. The size of an input file (bz2 compressed) is ≈ 1.4 Mb, the output file (gzip) size is ≈ 8.6 Kb. The network bandwidth needed for processing the complete expansion of a transcript, in the optimal case, is $294 \cdot 2 \cdot (1.4 + 0.0086) \approx 828$ Mb (not counting the dataset).

Another BOINC scheduler's parameter needed to be set is the workunit's *deadline*, it has to be carefully tuned to speed-up job return time, to have a fast turnaround in case of abandoned jobs and to prevent job preemption. At the very beginning we set the parameter to 4 days, we increased it to 6 days after deploying applications for low performance devices such as ARM-based computers.

3.3 Computational power and drawbacks

The use of a volunteer-based distributed computing system, like the BOINC framework, has some drawbacks. Any workunit is sent to at least two different volunteers, with a deadline (4 to 6 days in our set–up). The returned results, i.e. the application output files containing the expansion list, are considered correct only if returned before the deadline and if at least two of them are bit-wise identical (*homogeneous redundancy*), if not they are sent to other volunteers. This procedure, while obviously minimizing client-side computational errors, practically halves our theoretical computing

[1] https://bitbucket.org/francesco-asnicar/pc-boinc

TABLE 2
OneGenE benchmarks, application version 1.1

Dataset	# of Transcripts	# of Samples	Tile size	α	Time (s)	ETA (years) $i = 1000$
FANTOM-full	87554	1829	100	0.05	0.4753	144.39
			200	0.05	1.2313	187.03
			500	0.05	5.0811	310.13
			1000	0.05	15.1895	463.56
			1500	0.05	32.6940	668.96
			2000	0.05	50.9953	778.14
FANTOM-small	49727	1829	100	0.05	0.6014	58.99
			200	0.05	2.4171	118.55
			500	0.05	7.9092	155.78
			1000	0.05	27.4379	270.22
			1500	0.05	54.6467	365.96
			2000	0.05	99.8513	491.68
TCGA-PRAD-s	43128	537	100	0.05	0.0421	3.11
			200	0.05	0.0723	2.67
			500	0.05	0.2744	4.08
			1000	0.05	1.0705	8.05
			1500	0.05	2.5450	12.61
			2000	0.05	4.7694	17.92

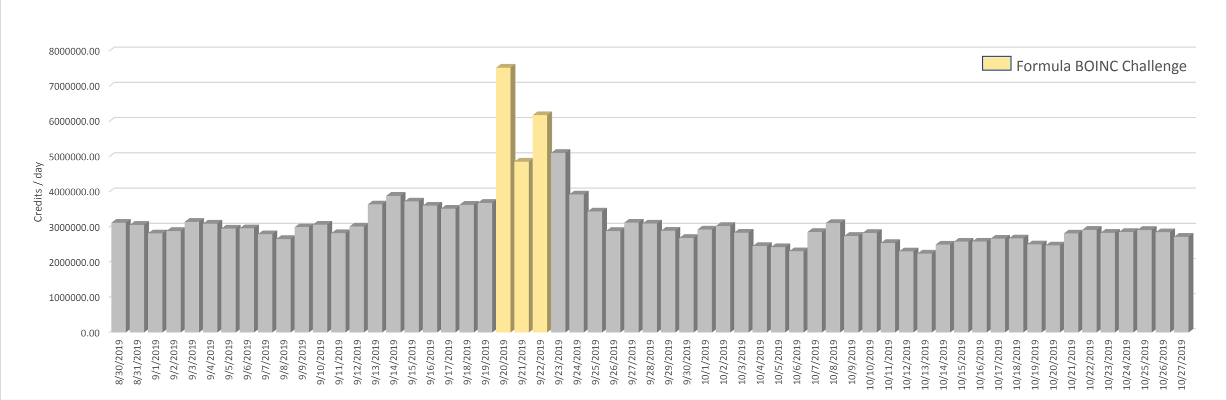
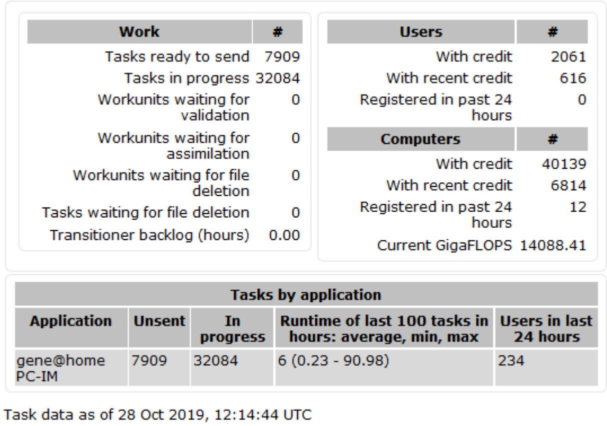


Fig. 2. Statistics for gene@home project, courtesy of Willy de Zutter (boincstats.com)



Work	#
Tasks ready to send	7909
Tasks in progress	32084
Workunits waiting for validation	0
Workunits waiting for assimilation	0
Workunits waiting for file deletion	0
Tasks waiting for file deletion	0
Transitioner backlog (hours)	0.00

Users	#
With credit	2061
With recent credit	616
Registered in past 24 hours	0
Computers	**#**
With credit	40139
With recent credit	6814
Registered in past 24 hours	12
Current GigaFLOPS	14088.41

Tasks by application

Application	Unsent	In progress	Runtime of last 100 tasks in hours: average, min, max	Users in last 24 hours
gene@home PC-IM	7909	32084	6 (0.23 - 90.98)	234

Task data as of 28 Oct 2019, 12:14:44 UTC

Fig. 3. Computing status page of the gene@home project

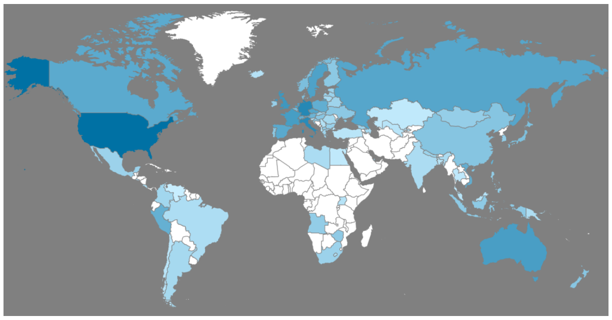
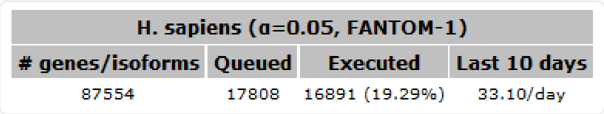


Fig. 4. gene@home distribution of computational power among countries

power. Moreover, for expanding a single transcript, the system distributes several workunits, and the slowest one to be completed determines the overall computation time.

We do not have any control over our volunteer's computing resources, this implies that we likely are subjected to errors, because of misconfigured or faulty devices. However, the error rate is acceptable: a 7-days statistics shows, among 125989 received results, 782 invalid (0.62%) and 1033 compute errors (0.82%), most actually generated by a very small number of computers. BOINC implements different strategies to limit the impact of faulty hosts, like automatically decreasing the number of sent workunits.

Another issue is that BOINC is a *volatile* resource, only



H. sapiens (α=0.05, FANTOM-1)			
# genes/isoforms	Queued	Executed	Last 10 days
87554	17808	16891 (19.29%)	33.10/day

Fig. 5. Computational status of HS OneGenE at 2019-11-05

relying on volunteers, therefore the overall available computing power is difficult to predict and to maintain over time. However, being active in solving volunteer's problems and taking care of communications issues we were able to involve a large number of volunteers, many of them with a large number of powerful computers, more than 40000 different computes contacted the server to request workunits, around 6800 recently (a snapshot of the server's status page is shown in Figure 3). In the last year we achieved an average power of 14 TFLOPS (equivalent to about 350 of our reference computer), and this value could be further increased by upgrading our server resources. Figure 2 (credits are proportional to FLOPS) shows the computing power of TN-Grid in a 60 days range. The high peaks starting at 2019-09-19 16:00 (UTC) are due to a *competition* among volunteers (*Formula Boinc Sprint*, http://formula-boinc.org/sprint.py) that attracted a significant number of them to the project, just for its duration (3 days), the higher peak value is very close to our theoretical power limit, without using non-volatile resources, such as HPC. BOINC itself maintains useful real-time statistics and leaderboards: FLOPS per application (https://gene.disi.unitn.it/test/apps.php), top users, top computers, and CPU models (https://gene.disi.unitn.it/test/stats.php).

Other interesting statistics is the world-map distribution of credits (see Figure 4). Note that this is a subset of the total number of users, a lot of them do not specify the *country* optional field while registering.

3.4 Current state of OneGenE

The gene@home project is running the complete systematic expansion of the dataset FANTOM-full. Figure 5 shows the current state of the project as a snapshot from the site. Although at the current rate the completion will require almost six years it is worth to say that the system could be at least three times faster as seen during the competition. Moreover, we can prioritize the genes on-demand and we plan to do actions to increase the steady base of volunteers if and when the demand of expansion of genes will increase.

A possible concern is the environmental impact of such a heavy computation that takes years on thousands of computers. The actual carbon-footprint of the project depends on the nature of the energy sources available to the volunteers. However, the distributed nature of the computation allows for local optimization like the ones foreseen by the SUNBURN project [27] where the computation is performed directly on a solar panel when there is excess of production. This is a potential advantage with respect to use centralized computing resources.

4 Validation

The objective of this section is the evaluation of the biological pertinence of the single-gene NES^{2}RA expansions obtained by OneGenE. The experiment was performed using both the full FANTOM expression dataset (FANTOM-full) and the dataset where redundant or low expression isoforms were filtered away (FANTOM-small), using 307 OneGenE expansions involving genes of medical relevance for two large families of human pathologies: neuronal motor diseases and hematopoietic tumors. For each expansion, the top scoring 250 transcripts were considered. OneGenE expansions were benchmarked against a simple Pearson correlation analysis: starting from the same *seed* transcript, the top 250 correlated transcripts were considered and compared to the 250 transcripts identified by OneGenE.

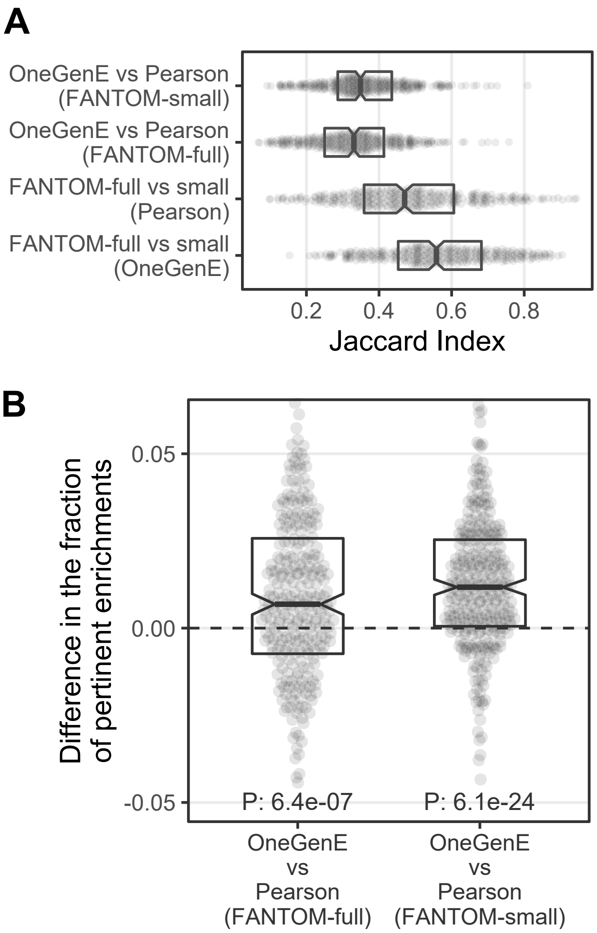


Fig. 6. Comparison of the gene similarity (**A**) (Jaccard Index) and the difference in biological pertinence (**B**) (fraction of significant functional enrichments that include the "seed" of the expansion) of the expansions obtained using the OneGenE method against an approach based on Pearson correlation. The comparison was performed for 307 expansions in the FANTOM-full and FANTOM-small datasets.

To quantify the overlap among the expansions obtained from the same transcript using different methods or different datasets, we calculated the *Jaccard Index*, defined as the number of items shared between two sets divided by the total number of items in both sets (shared and un-shared). As represented in Figure 6A, the distribution of Jaccard Indexes indicates that OneGenE expansions are largely populated by distinct genes with respect to the correlation approach (median Jaccard Index 0.33 and 0.35 for the FANTOM-full and FANTOM-small datasets, respectively). On the other hand, the application of the same method to different datasets

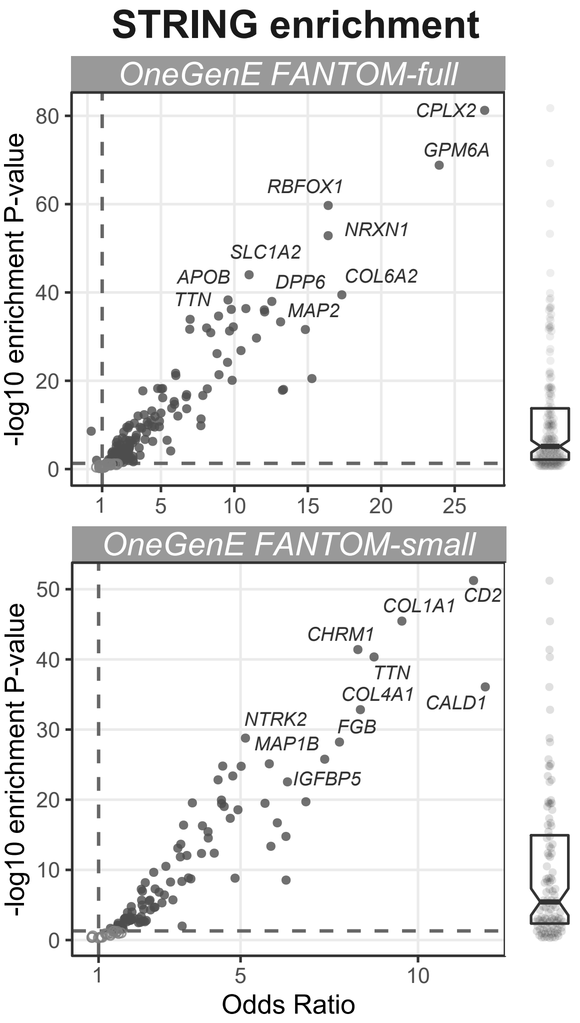


Fig. 7. Overlap between OneGenE expansions and protein-protein interactions annotated in STRING, using the datasets FANTOM-full (upper panel) and FANTOM-small (lower panel). For each gene, the scatter plot displays the Odds Ratio and the enrichment p-value. The ten most enriched genes are highlighted in the plot. The distribution of the p-values is also displayed as a box-whisker plot on the left.

yields more similar results (median Jaccard Index 0.56 and 0.47 for OneGenE and Pearson, respectively).

To evaluate the biological pertinence of an expansion, we performed functional enrichment analysis on the list of the top scoring 250 transcripts with the EnrichR resource (https://amp.pharm.mssm.edu/Enrichr/). Next, we quantified the *fraction of pertinent enrichments*, defined as the fraction of statistically significant enrichments (Fisher test P-value < 0.05) that also contain the *seed* transcript of the expansion. The higher the value of this measure, more the expanded transcripts are functionally related to the seed, based on existing annotations. As represented in Figure 6B, this analysis revealed that the single-gene NES^{2}RA expansions of OneGenE consistently achieves higher biological pertinence than the correlation approach, in both the FANTOM-full and the FANTOM-small datasets (P: 6.4e-07 and 6.1e-24 respectively, tested with a Mann Whitney U test).

Finally, in order to validate our results with published protein-protein interaction data, we performed a comparison between OneGenE expansions and interactions annotated in STRING v.11 [28]. We considered the NES^{2}RA expansions of 179 genes, selecting their most expressed isoform. For each expansion, we compared the list of genes identified with OneGenE with the list of known interactions in STRING (combined score > 0.15). Based on the number of genes in the OneGenE expansion, the number of annotated STRING interactions and the overlap between the two lists, we calculated odds ratio values and corresponding enrichment P-values (Fisher test), as represented in Figure 7 and ST1 (see Supplementary Material). We report that the odds ratio is > 1 in 96% and 94% of the expansions, with P-value < 0.05 in 84% and 82% of the expansions (FANTOM-small and FANTOM-full, respectively, Figure 7). This result further supports the biological pertinence of the single-gene NES^{2}RA expansions obtained by OneGenE.

5 Case Study 1: Prostate Cancer

Here, we present the application of the OneGenE method for prostate cancer drug repositioning. A selection of 22 genes known to be important for this cancer onset and development has been expanded as single-gene by NES^{2}RA on two transcriptomic datasets (see Section 2): FANTOM-small, including a wide collection of data from primary cell types, human tissues and cancer cells, and TCGA-PRAD-s, a dataset dedicated to prostate cancer. These lists have been initially used to analyze the direct interactions within the input genes and represent them as graphs. We then focused on two prostate specific networks and the expansion lists of the genes belonging to the networks were aggregated. A comparison analysis with STRING and functional enrichment analyses allowed to understand the nature and composition of these gene networks. Finally, after filtering out genes already known to be related to prostate cancer, a query against the Gene Drug Interaction database allowed to identify novel targets for this disease that can support drug repositioning. The pipeline of the procedure is depicted in Figure 8.

5.1 Prostate cancer genes selection and expansions with NES^{2}RA

Prostate cancer related genes were retrieved from the OpenTargets platform (release 3.8 2018-08-28) [29] which provides genes associated to clinical conditions and scores according to the reliability of this evidence. In this study, germinal and somatic genetic variations were equally considered and 22 genes, corresponding to 125 isoforms of the FANTOM-small dataset, were selected (ST2, ST3). They were subsequently expanded as single genes by NES^{2}RA using FANTOM-small and TCGA-PRAD-s expression datasets and thus producing 125 and 22 expansion lists, respectively.

5.2 Network analysis of the input genes

As a first step, the expansion gene lists computed on the two datasets were trimmed in order to have lists of comparable length. The relative frequency has been compared against a minimum threshold: 0.5 for the FANTOM-small lists, 0.1 for the TCGA-PRAD-s, producing lists of 250-400 genes. Then, these lists were analyzed to find the functional relationships

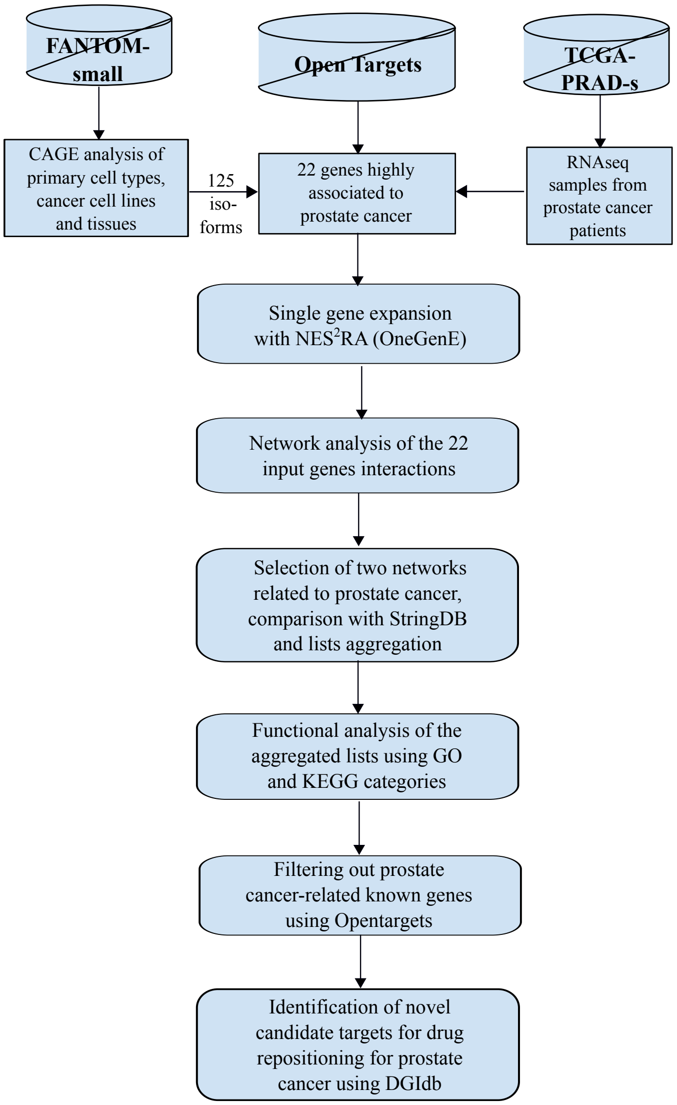

FANTOM-small
Open Targets
TCGA-PRAD-s
CAGE analysis of primary cell types, cancer cell lines and tissues
125 iso-forms
22 genes highly associated to prostate cancer
RNAseq samples from prostate cancer patients
Single gene expansion with NES²RA (OneGenE)
Network analysis of the 22 input genes interactions
Selection of two networks related to prostate cancer, comparison with StringDB and lists aggregation
Functional analysis of the aggregated lists using GO and KEGG categories
Filtering out prostate cancer-related known genes using Opentargets
Identification of novel candidate targets for drug repositioning for prostate cancer using DGIdb

Fig. 8. Case Study 1: Pipeline.

among the input genes. The mutual interaction among each pair (x, y) of input genes, defined as the presence of gene x into the expansion list of gene y and vice versa, was computed to reconstruct a gene network. In the case of the FANTOM-small dataset, as many isoforms have been measured for each gene, the expansion lists of the isoforms of each gene were combined by averaging their relative frequency. The strength of the mutual interaction was calculated as the average between the relative frequencies of gene x in the expansion list of the gene y and that of gene y in the list of gene x. The resulting networks are presented as graphs in Figure 9, where the 22 initial genes are the nodes and the edges represent the mutual interaction.

The graphs produced from the FANTOM-small (Figure 9A) and TCGA-PRAD-s (Figure 9B) datasets appeared quite different: in the former case 20 out of 22 genes were connected, whereas in the latter only 5 genes, forming 2 networks, were. In particular, these genes were CHEK2, TERT, PTEN, MXI1 and SPOP. This result likely reflected the different nature of the datasets, a generic one and a condition-specific one. To understand the way cancer affects and impairs cellular gene networks, we focused our attention onto these 5 genes. The network formed by CHEK2 (a kinase involved in cell cycle regulation) and TERT (a reverse transcriptase involved in telomere maintenance) was identified in both analyses. These genes are well-known to be strictly related to cell cycle and cancer development. Therefore, their interaction was expected in the prostate cancer dataset; nonetheless, by checking their expression profile in FANTOM5 using SSTAR [30], we observed that they were mainly expressed in tumor cell types also in FANTOM5. A different case is represented by PTEN (a phosphatidylinositol phosphatase that acts as a tumor suppressor) and MXI1 (a transcription factor that down-regulates the oncogenic c-myc gene): their interaction was identified in both analyses, but in FANTOM-small the isoforms responsible of the interaction (listed in ST4) are expressed mainly in brain tissues, so it seems that in prostate cancer cells, where they are likely to be mutated, an unusual interaction takes place. Finally, SPOP, a protein that mediates gene repression by binding to histone complexes and the most frequently point-mutated gene in prostate cancer, interacts with different genes in FANTOM-small and TCGA-PRAD-s. In the former, it is connected to CAPS8 and ATM, both involved in cell cycle regulation and apoptosis, while in the latter, it interacts with MXI1. This result is supported by previous findings, which reports that the mutated SPOP protein loses the ability to interact with the other proteins involved in DNA repair, causing spontaneous replication stress [31].

5.3 Comparison with STRING

Focusing now on the whole expansion gene lists computed by NES^{2}RA on the two datasets FANTOM-small and TCGA-PRAD-s, a first validation analysis was carried out. NES^{2}RA output was compared with the lists of interacting genes provided by STRING (v. 11) [28] similarly to what shown in Section 4. Results are shown in Table 5.3.

TABLE 3
Case Study 1: enrichment analysis between OneGenE expansions and interactions annotated in STRING

Gene	Dataset	Expansion genes	STRING overlap	Odds ratio	P_value
CHEK2	*FANTOM-small*	335	110	**3.13**	**2.61E-19**
CHEK2	*TCGA-PRAD-s*	213	81	**3.91**	**5.03E-19**
PTEN	*TCGA-PRAD-s*	244	87	**1.82**	**8.79E-06**
TERT	*FANTOM-small*	261	52	**2.02**	**1.28E-05**
PTEN	*FANTOM-small*	244	79	**1.57**	**6.81E-04**
SPOP	*FANTOM-small*	341	31	**1.85**	**1.37E-03**
TERT	*TCGA-PRAD-s*	137	19	**1.29**	1.69E-01
SPOP	*TCGA-PRAD-s*	171	6	0.66	1.94E-01
MXI1	*TCGA-PRAD-s*	239	11	**1.21**	2.48E-01
MXI1	*FANTOM-small*	258	12	**1.23**	2.55E-01

The overlap between STRING and OneGenE output is significant for 4 out 5 of the gene lists computed on FANTOM-small, while only for 2 out of 5 of those computed on TCGA-PRAD-s. This result was not unexpected, as STRING and FANTOM5 are both comprehensive and generic databases: the remarkable agreement of this comparison represented a good validation of OneGenE analysis. Conversely, the lower overlap observed between STRING and TCGA-PRAD-s analyses likely reflected the different behavior of some genes between the healthy and prostate cancer condition, as reported for SPOP [31].

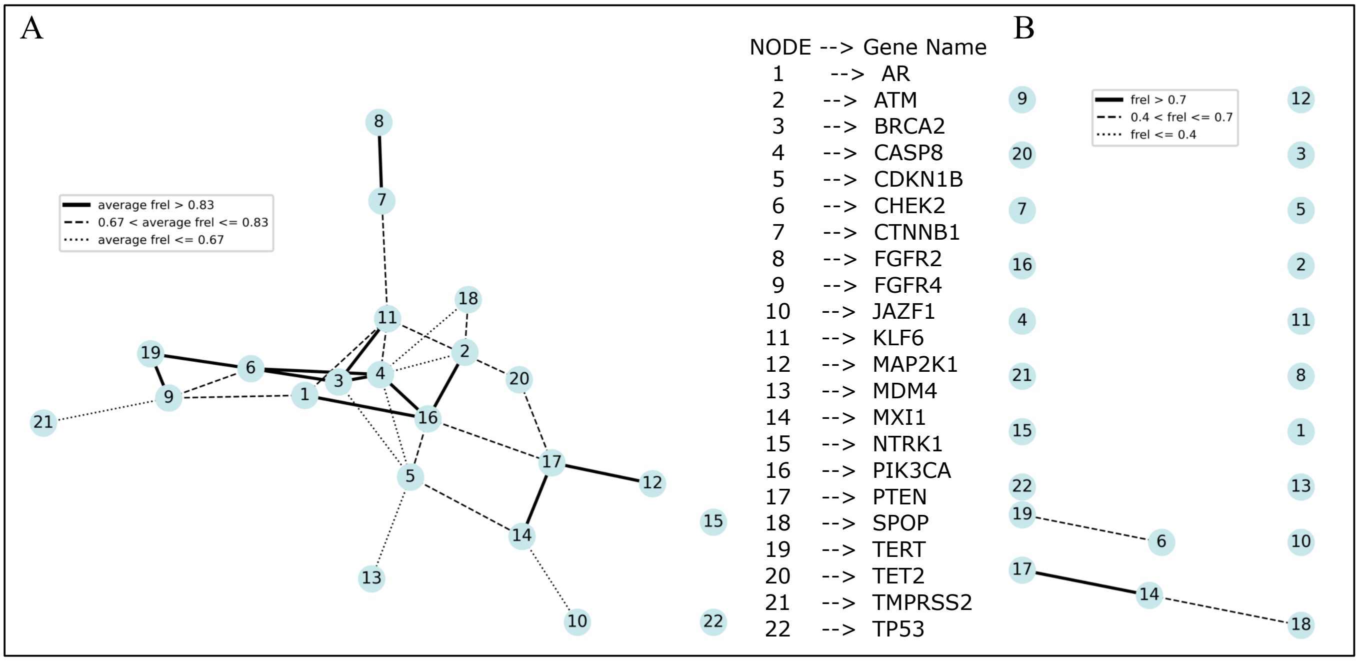


Fig. 9. Case Study 1. Network representation of the interactions among the 22 genes related to prostate cancer. The two graphs were produced with NetworkX Python pkg considering the mutual interaction computed on the expansion lists obtained with NES^{2}RA on the FANTOM-small (A) and TCGA-PRAD-s (B) datasets. Nodes represent genes as described in the legend; edges connect interacting genes. Mutual interaction has been calculated as the mean of the relative frequencies of the two genes and has been represented with three line styles.

5.4 Functional analysis of the TCGA-PRAD-s lists

The expansion lists computed on the TCGA-PRAD-s dataset were aggregated according to the network analysis presented in 5.2. We thus named CHEK2-TERT network the list of 567 genes coming from the union of their respective expansion lists; similarly, the PTEN-MXI1-SPOP network is formed by 810 genes. By doing so, we focused our attention of the interactions taking place in the perturbed cancer situation, that is the real condition in which drugs should be effective. To explore the gene composition and functional involvement of the two networks, enrichment analyses considering KEGG pathways and Gene Ontology biological process categories were performed (Table 5.4 and ST5).

TABLE 4
Case Study 1: Enrichment analysis. Top enriched categories obtained with EnrichR considering KEGG pathways 2019 on the aggregated lists of the two networks, CHEK2-TERT and PTEN-MXI1-SPOP.

KEGG Pathways 2019	P-value	Adj P-value
CHEK2-TERT network		
DNA replication	2.73E-09	8.40E-07
Cell cycle	2.37E-06	3.64E-04
Mismatch repair	3.96E-04	4.07E-02
PTEN-MXI1-SPOP network		
Signaling pathways regulating pluripotency of stem cells	4.28E-03	1.00E+00
Protein processing in endoplasmic reticulum	7.31E-03	7.51E-01
Tight junction	9.40E-03	5.79E-01
Oxytocin signaling pathway	2.19E-02	1.00E+00
Ras signaling pathway	2.72E-02	1.00E+00
Gastric cancer	4.00E-02	1.00E+00

As expected, the CHEK2-TERT network was enriched in genes involved in DNA replication, cell cycle and DNA damage repair. Conversely, the PTEN-MXI1-SPOP gene network was populated by heterogeneous groups of genes involved in signaling cascades (such as RAS, RAP1, WNT pathways), tight junction, protein transport from the ER to the plasma membrane and many cancer types, and did not produce significantly enriched categories.

5.5 Selection of novel drug targets

The two lists were then filtered by removing those genes that were present in the OpenTargets list as prostate cancer related: 159 genes (28%) for the CHEK2-TERT network, and 249 (31%) for the other network. By querying the DGIdb database [32] of gene-drug interactions with these filtered lists (408 and 561 genes, respectively), we obtained a selection of 12 FDA-approved drugs for the CHEK2-TERT network and 23 to the PTEN-MXI1-SPOP one (ST6). Finally, the presence of these genes in the expansion lists obtained from FANTOM-small has been checked in order to exclude targets that might raise more cytotoxic response, due to their interaction with the input genes also in healthy conditions. The genes GAL, UST and LIAS were thus not considered in the list of targets.

5.6 Discussion of Case Study 1

Here, we show that starting from 22 genes related to prostate cancer and expanding them with the OneGenE algorithm, two small cancer specific networks have been identified. About 30% of these genes were already classified as prostate cancer related in Opentargets, thus supporting the validity of the method. Among the others, novel putative targets for drug repositioning are proposed (Table 5.5). Three subunits of the Calcium-voltage-gated channel have been found, targeted by GABA-analogue anti-convulsant drugs. Despite being commonly associated with physiological processes

TABLE 5
Case Study 1: target genes.

Gene	Gene description	Drug	Class	Disease (DrugBank)
CACNG4 CACNA1E CACNB2	calcium voltage-gated channel subunits	PREGABALIN GABAPENTIN	Anticonvulsants, Analgesic Anticonvulsants, Analgesic	neuropathic pain neuropathic pain
GALR1	galanin receptor 1	METHOXAMINE TESTOSTERONE TEMAZEPAM	Antihypotensive agents Hormone replacement agents Hypnotics and sedatives	Acute hypotensive state Hypogonadism Short-term insomnia
GNG2	G protein subunit gamma 2	HALOTHANE	Anesthetics, inhalation	Anesthetic
KCNJ12	potassium inwardly rectifying channel subfamily J member 12	DOFETILIDE	Anti-arrhythmia agents	Cardiac action, antiarrhythmic agent
RPE65 IL1R1	retinoid isomerohydrolase RPE65 interleukin 1 receptor type 1	TRETINOIN	Keratolytic agents antineoplastic agents	Skin conditions, promyelocytic leukemia Skin conditions, promyelocytic leukemia

such as neurotransmitter release, excitation-contraction coupling or hormone secretion, they play an important role in cell proliferation and apoptosis in many cancer types and represent therapeutical targets already under investigation [33]. The genes GALR1, GNG2 and KCNJ12 have been found as functionally correlated and involved in tumor suppression in other tissues, representing novel target for prostate tumor [34]. Finally, IL1R1 and RPE65, targeted by tretinoin, the former involved in mammary and pulmonary cancer [35], could be considered for further characterization in prostate cancer.

6 Case Study 2: Coronary Artery Disease

In this second case study, the OneGenE approach of single-gene NES^{2}RA expansions was applied to identify novel putative drug targets for Coronary Artery Disease (CAD), by considering genes genetically associated with this disease which can be used for drug repositioning. The pipeline and step-by-step results have been summarized in Figure 10.

Open Targets is a platform which integrates public domain data to enable drug target identification and prioritization [29]. The list of 643 genes genetically associated with CAD was retrieved (Open Targets release 3.8 (2018-08-28)) and ranked based on the genetic association score. We focused on the first 46, which correspond to 183 isoforms of the FANTOM-small dataset (ST7). These were expanded as single genes by NES^{2}RA, obtaining the corresponding expansion lists.

6.1 Score aggregation and filtering

In order to retain the genes most related to CAD, the expansion lists were trimmed according to the following criteria: For each expansion list, only the first N_l isoforms were selected, where $N_l = max\{5; 100 - R_l + 1\}$, and R_l is the rank, based on genetic association derived from Open Targets, of the starting gene. The resulting lists of isoforms were aggregated summing the relative frequencies computed by NES^{2}RA. Then, the list of isoforms was ranked and converted into a ranked list of 2043 genes. The rank of a gene was obtained as the minimum of the ranks of its corresponding isoforms.

6.2 Selection of target genes and functional analysis

To have an insight on the biological function of these genes, ToppGene [36] was used to perform enrichment analysis against the Biological Processes of the Gene Ontology (ST8).

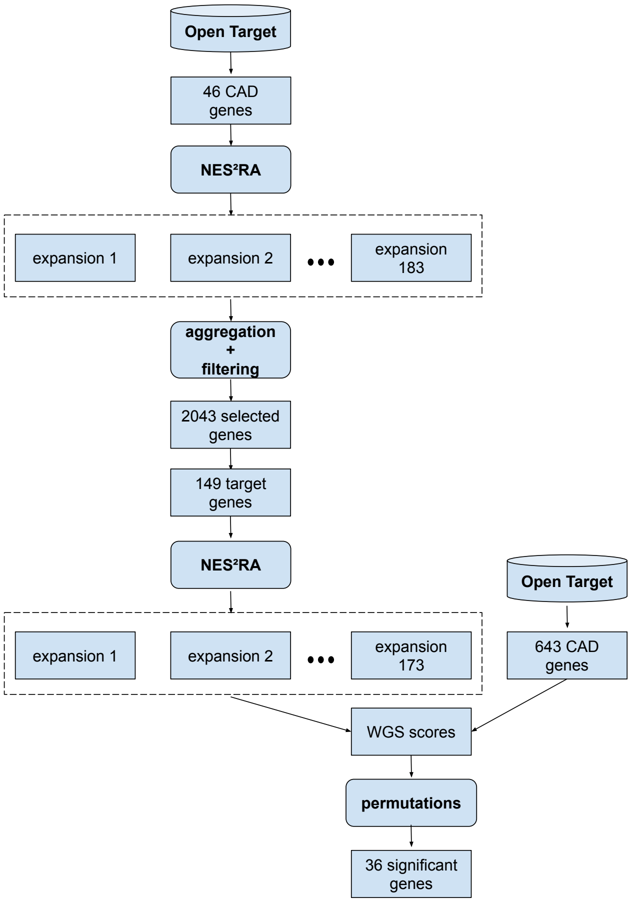


Fig. 10. Case Study 2: Pipeline.

The most significant terms are reported in Table 6.2. The Open Targets Python API was used to query which of these genes were already targets of clinically approved drugs. A list of 149 target genes was obtained, which were then considered for the following analyses (ST9).

6.3 Single-gene NES^{2}RA expansion of the target genes

In order to identify a subset of genes with a strong association with CAD, 173 isoforms corresponding to the 149 target genes were used as starting genes for single-gene NES^{2}RA expansions. For each gene, only isoforms that were present in the filtered and aggregated list described above were considered.

TABLE 6
Case study 2: Selection of top enriched GO Biological processes from the list of 2043 genes resulting from step 3 of the pipeline. The q-values result from Benjamini-Hochberg correction.

Biological Process	**q-value**
organic substance catabolic process	*4.40E-11*
regulation of intracellular signal transduction	*1.20E-10*
positive regulation of gene expression	*4.10E-10*
response to oxygen-containing compound	*4.60E-10*
positive regulation of RNA metabolic process	*1.60E-09*
regulation of response to stress	*2.00E-08*
vasculature development	*4.60E-08*
cardiovascular system development	*8.40E-08*
regulation of cell proliferation	*1.80E-07*
lipid metabolic process	*3.30E-07*
blood vessel development	*6.30E-07*
regulation of kinase activity	*9.10E-07*
response to endogenous stimulus	*1.10E-06*
cell cycle	*1.10E-06*
symbiont process	*1.10E-06*

6.4 Comparison of expanded lists with CAD genes

In order to quantify the overlap between the ranked list of genes genetically associated with CAD (obtained from Open Targets) and the ranked expansion lists obtained in the previous step as output of NES^{2}RA, we used the Weighted Jaccard Similarity (WJS) [37], which we define below.

Definition. Given two weighted lists of items, ρ and σ, of length N, the Weighted Jaccard Similarity, WJS(ρ, σ), is defined as:

$$\text{WJS}(\rho, \sigma) = \frac{\sum_{i=1}^{N} min(\rho_i, \sigma_i)}{\sum_{i=1}^{N} max(\rho_i, \sigma_i)}$$

where ρ_i and σ_i are the weights corresponding to the same item i.

In our analysis, the weight of a feature i (gene or isoform) in a ranked list ρ is computed as $length(\rho) - rank(i) + 1$. This allows us to assign large weights to high ranking features. In principle, the lists of features do not contain the same number of elements. To solve this issue, we included the missing elements of one list to the other (and vice versa) as ties in the last position of the ranking. To compute the WJS between the CAD genes and each expansion list, we converted NES^{2}RA isoforms into genes, ranking them by the largest relative frequency of their isoforms.

The scores obtained from the WJS are not directly comparable, since they depend on the length of the lists. In order to obtain values that can be compared directly, we used a permutation approach to estimate a set of score distributions. For each length present, we generated 2000 random lists of genes or isoforms, we computed the WJS and we generated the score distribution associated to that length. Then, we used these distributions to compute empirical p-values. The Benjamini Hochberg correction was used to adjust p-values for multiple hypothesis testing [38]. A significance level of 0.05 was used to identify statistically significant WJS scores.

After the permutation approach 36 genes were obtained (Table 7 and ST10). The tables show the possible target genes, ordered by their q-value. 19 of them are already associated with CAD (column "CAD") and 4 of these (PDE4D, PDE4B, NDUFA4L2 and INSR) already present a drug for the disease: Dyripimadole (Phase IV) and Pentoxyfylline (Phase II) for PDE4D and PDE4B, Metformin (Phase IV) for NDUFA4L2, Insulin Human (Phase IV), Insulin Aspart (Phase IV), Insulin Lispro (Phase IV), Insulin Glargine (Phase IV) for INSR (data retrieved by Open Targets). Drug Bank (Version 5.1.1, released July 03, 2018) [39] was used to obtain the drugs in Phase IV that have the 36 genes as targets, the disease and the molecule type (respectively columns "Drug", "Disease" and "Molecule Type" in the table). The drugs reported in the table are the new putative drugs that can be further investigated.

6.5 Discussion of Case Study 2

NES^{2}RA one gene expansions are meant to explore putative causal relationships between genes. Since NES^{2}RA retrieves up to several thousands genes for each expansion, we have to apply aggregation and filtering techniques in order to extract meaningful biological information from the expansion lists. One of such techniques is the comparison of expansion lists of putative targets with the list of genes already known to be associated with CAD, which allows one to identify target genes that are more likely to interact directly with CAD genes. It is worth mentioning that this approach focuses only on the identification of a subset of all the possible putative targets, namely those that are correlated to CAD genes at the transcriptional level. In order to critically evaluate the biological relevance of the results of NES^{2}RA, we perform enrichment analysis. We expect to find Gene Ontology terms related to the cardiovascular systems. To assess whether the subset of 149 target genes is related to the cardiovascular system, we performed an enrichment analysis that retrieved a larger number of terms associated with the circulatory system, such as blood circulation, circulatory system process and angiogenesis. This suggests that NES^{2}RA was able to retrieve a group of genes that directly interact with CAD genes and appear to be involved in the cardiovascular system. The 36 genes identified with our pipeline are involved in metabolic diseases, autoimmune diseases (particularly skin and joints), skin diseases, diseases of female reproductive organs, cardiovascular diseases and kidney diseases. We decided to analyze in further detail the relationship of some of these diseases with CAD, using the literature. The relation between diabetes mellitus (DM) and cardiovascular diseases is well known. For DM, CAD is a major determinant of the long-term prognosis among patients. DM is associated with a 2 to 4-fold increased mortality risk from heart disease [40]. Metformin, a drug targeting NDUFAL4L2 is an example of a drug used to treat both CAD and DM [41]. Considering this result, a proposal could be to repurpose drugs used to treat DM for CAD. From our analyses, the gene GAA seems to be a promising target for drug repurposing. GAA (Glucosidase Alpha, Acid) is a protein coding gene essential for the degradation of glycogen in lysosomes. This gene is not associated with CAD, but it was identified by our pipeline. It is targeted by Miglitol, a drug that acts by inhibiting the ability of the patient to breakdown complex carbohydrates into glucose. Another identified gene, PDE4B, already has a phase IV drug for CAD (namely, Dypiridamole), alongside approved drugs for some skin conditions, such as atopic eczema and alopecia. This link suggests that the several drugs acting on skin conditions, targeting the gene RARA (one of our 36 genes), could be investigated for CAD. RARA is very weakly associated with CAD, therefore any positive therapeutic confirmation would be particularly significant

TABLE 7
Case Study 2: 6 of the 36 putative new target genes for CAD, together with their Phase IV associated drugs (DrugBank (Version 5.1.1, released July 03, 2018)). For the complete list, see ST10 in the supplementary material.

Target gene	q-value	CAD gene	Drug (retrieved by Drug Bank)	Disease	Molecule Type
PDE4D	0	yes	PENTOXIFYLLINE	Hepatitis, Alcoholic	Small molecule
			DYPHYLLINE	obstructive lung disease	Small molecule
			ROFLUMILAST	chronic obstructive pulmonary disease	Small molecule
PDE4B	0	yes	DIPYRIMADOLE	coronary artery disease	Small molecule
			AMLEXANOX	obstructive lung disease	Small molecule
			FLAVOXATE	pain	Small molecule
			THEOPHYLLINE	asthma	Small molecule
			APREMILAST	Alopecia	Small molecule
			CRISABOROLE	atopic eczema	
NDUFA4L2	0	yes	METFORMIN	Obesity	Small molecule
INSR	0	yes	INSULIN HUMAN	Type II diabetes mellitus	Protein
RARA	0	yes	TRETINOIN	neoplasm	Small molecule
			ADAPALENE	acne	Small molecule
			ETRETINATE	psoriasis	Small molecule
			ALITRETINOIN	neoplasm	Small molecule
			ISOTRETINOIN	acne	Small molecule
			TAZAROTENE	acne	Small molecule
			ACITRETIN	psoriasis vulgaris	Small molecule
GAA	4.4e-02	no	MIGLITOL	type II diabetes mellitus	Small molecule

for the assessment of the efficacy of NES^{2}RA for drug repurposing. Also Pentoxifylline, currently a phase II drug for CAD disease, is reported in the literature to be a drug with wide spectrum applications in dermatology, although it has not been investigated thoroughly for these applications [42]. This could potentially foster the connection between CAD and skin conditions.

7 Conclusion

We presented the activity of systematic discovery of causal relationships between the transcripts of human genes and its application to prostate cancer and coronary artery disease with the goal of drug repositioning. The activity is performed within the gene@home project that relies on volunteer computation using BOINC. The distributed computation provides for several executions of the PC algorithm, a popular causality discovery algorithm on subsets of data with a high number of variables. In particular the system runs on data about the transcription of human genes. Validation shows that this approach has an edge w.r.t. the pure correlation for finding relevant and functionally related genes, and that results are significantly enriched in known protein-protein interactions. Case studies on prostate cancer and coronary artery disease show that the method can be effectively inserted in pipelines aimed at drug repositioning/repurposing.

Causality discovery and inference from purely observational data is a central topics in statistics, biostatistics and data mining. In this work we show that, with a big computational effort mitigated by the use of BOINC and the help of the BOINC users community, the task to discover putative causal relationships can be done for variables in the order of 50000 and more. This permits to tackle applications in sensitive domains like drug repurposing for important and severe diseases.

Drug repurposing is a challenging task for which computational tools could provide a good starting point, allowing to start from a relatively short list of putative drugs to explore, hence permitting to save resources and time. The aim of the two case studies is to show the potential of One-GenE applied to this complex task for two different diseases: prostate cancer (Case Study 1) and coronary artery disease (Case Study 2). In both cases, as a starting point, a list of 22 and 46 genes, known to be genetically associated with prostate cancer and CAD respectively, has been selected. In Case Study 1 the pipeline produced a list of 22 genes likely involved in prostate cancer that are therapeutic targets for already FDA-approved drugs currently used to treat other pathologies. Case Study 2 resulted in a list of 36 target genes for different diseases. Of these 36 genes, 19 emerged to be associated with CAD and 4 (PDE4D, PDE4B, NDUFA4L2 and INSR) to have a drug already used to treat it.

Both case studies retrieved genes that are related to their disease of interest, which shows the potential to improve drug repositioning. However, in order to draw conclusions on the validity of the targets and of the drugs individuated in the two case studies experimental and clinical validations are required.

Supplementary Material

The file **SupplTables.xlsx** contains supplementary tables ST1-10 referred to in Sections 4, 5 and 6 and **Code.zip** collects the related source code. The file **ExampleofOneGeneExpansion.xlsx** contains, as an example, the output of the expansion of the first isoform of gene CHEK2 (p1@CHEK2).

Acknowledgment

FA, LM, and VC conceived, implemented and run the system and the benchmarking. TT selected the FANTOM dataset, and conceived and performed the validation. LM preprocessed the FANTOM data. SP and GT performed and discussed the case study 1 on prostate cancer. MC, EC, CM, EN conceived, performed, and discussed the case study 2 on coronary artery disease. EB and SP conceived and coordinated the overall research. All the authors contributed to the discussion of the results and to the writing. Thanks to Giordano Alvari, Alberto Boldrini, Stefano Flor and Davide Golzato who initially conceived case study 1 and selected the input genes, for TCGA-PRAD data selection, and preprocessing. The authors wish to thank also Enrico Domenici, Daniel Frużyński, Willy de Zutter, Lorenzo Mastrogiacomi, Veronica Rizzi and all the BOINC volunteers.

References

[1] J. Hasty, D. McMillen *et al.*, "Computational studies of gene regulatory networks: in numero molecular biology," *Nature Reviews Genetics*, vol. 2, no. 4, p. 268, 2001.

[2] V. A. Huynh-Thu and G. Sanguinetti, *Gene Regulatory Network Inference: An Introductory Survey*. New York, NY: Springer New York, 2019, pp. 1–23. [Online]. Available: https://doi.org/10.1007/978-1-4939-8882-2_1

[3] F. M. Delgado and F. Gómez-Vela, "Computational methods for gene regulatory networks reconstruction and analysis: A review," *Artificial Intelligence in Medicine*, vol. 95, pp. 133 – 145, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0933365718303865

[4] M. H. Maathuis, D. Colombo *et al.*, "Predicting causal effects in large-scale systems from observational data," *Nature Methods*, vol. 7, no. 4, pp. 247–248, 2010.

[5] J. Pearl, *Causality*. Cambridge university press, 2009.

[6] T. D. Le, K. Zhang *et al.*, "Preface: The 2018 ACM SIGKDD workshop on causal discovery," in *Proceedings of 2018 ACM SIGKDD Workshop on Causal Discovery*, 2018, pp. 1–3.

[7] P. Spirtes and C. Glymour, "An algorithm for fast recovery of sparse causal graphs," *Social Science Computer Review*, vol. 9, no. 1, pp. 62–72, 1991.

[8] C. Glymour, K. Zhang, and P. Spirtes, "Review of causal discovery methods based on graphical models," *Frontiers in Genetics*, vol. 10, p. 524, 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fgene.2019.00524

[9] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC bioinformatics*, vol. 9, no. 1, p. 559, 2008.

[10] F. Asnicar, L. Erculiani *et al.*, "Discovering candidates for gene network expansion by distributed volunteer computing," in *2015 IEEE Trustcom/BigDataSE/ISPA*, vol. 3, Aug 2015, pp. 248–253.

[11] F. Asnicar, L. Masera *et al.*, "NES^2RA: Network expansion by stratified variable subsetting and ranking aggregation," *The International Journal of High Performance Computing Applications*, vol. 32, no. 3, pp. 380–392, aug 2016.

[12] F. Asnicar, N. Sella *et al.*, "TN-grid and gene@home project: Volunteer computing for bioinformatics," in *Ivashko, E. (ed.) Second International Conference BOINC-based High Performance Computing: Fundamental Research and Development (BOINC:FAST 2015)*. Petrozavodsk, Russia: Petrozavodsk: Russian Academy of Sciences, September 14-18 2015.

[13] G. Malacarne, S. Pilati *et al.*, "Discovering causal relationships in grapevine expression data to expand gene networks. a case study: Four networks related to climate change," *Frontiers in Plant Science*, vol. 9, p. 1385, 2018.

[14] A. A. Margolin, I. Nemenman *et al.*, "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," in *BMC Bioinformatics*, vol. 7, no. S1. Springer, 2006, p. S7.

[15] F. Asnicar, L. Masera *et al.*, "OneGenE: Regulatory gene network expansion via distributed volunteer computing on boinc," in *2019 27th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*, Feb 2019, pp. 315–322.

[16] S. Noguchi, T. Arakawa *et al.*, "FANTOM5 CAGE profiles of human and mouse samples," *Scientific Data*, vol. 4, p. 170112, 08 2017.

[17] S. Chen and J. C. Mar, "Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data," *BMC Bioinformatics*, vol. 19, no. 1, p. 232, 2018.

[18] E. Crocetti, "Epidemiology of prostate cancer in Europe," 2015. [Online]. Available: https://ec.europa.eu/jrc/en/publication/epidemiology-prostate-cancer-europe

[19] A. N. Nowbar, M. Gitto *et al.*, "Mortality from ischemic heart disease: Analysis of data from the World Health Organization and coronary artery disease risk factors from NCD risk factor collaboration," *Circulation: Cardiovascular Quality and Outcomes*, vol. 12, no. 6, p. e005375, 2019.

[20] F. Asnicar, L. Masera *et al.*, "NES^{2}RA: Network expansion by stratified variable subsetting and ranking aggregation," *The International Journal of High Performance Computing Applications*, vol. 32, no. 3, pp. 380–392, aug 2016.

[21] M. Kalisch and P. Bühlmann, "Estimating high-dimensional directed acyclic graphs with the pc-algorithm," *J. Mach. Learn. Res.*, vol. 8, pp. 613–636, May 2007.

[22] E. Coller, "Analysis of the PC algorithm as a tool for the inference of gene regulatory networks: evaluation of the performance, modification and application to selected case studies." Ph.D. dissertation, University of Trento, 2013.

[23] Hoadley *et al.*, "Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer," *Cell*, vol. 173, no. 2, pp. 291–304.e6, apr 2018.

[24] A. Abeshouse, J. Ahn *et al.*, "The molecular taxonomy of primary prostate cancer," *Cell*, vol. 163, pp. 1011–1025, 11 2015.

[25] V. Developers, "Callgrind: a call-graph generating cache and branch prediction profiler," 2010.

[26] G. Frank, "Pulse code communication," Mar. 17 1953, uS Patent 2,632,058.

[27] J. K. Nurminen, T. Niemi *et al.*, "Sunburn—using excess energy of small-scale production for distributed computing," *Energy Efficiency*, vol. 11, no. 1, pp. 97–119, Jan 2018. [Online]. Available: https://doi.org/10.1007/s12053-017-9552-1

[28] D. Szklarczyk, A. L. Gable *et al.*, "STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Research*, vol. 47, no. D1, pp. D607–D613, 11 2018. [Online]. Available: https://doi.org/10.1093/nar/gky1131

[29] D. Carvalho-Silva, A. Pierleoni *et al.*, "Open Targets Platform: new developments and updates two years on," *Nucleic Acids Research*, vol. 47, no. D1, pp. D1056–D1065, jan 2019. [Online]. Available: https://academic.oup.com/nar/article/47/D1/D1056/5193331

[30] I. Abugessaisa, H. Shimoji *et al.*, "FANTOM5 transcriptome catalog of cellular states based on semantic mediawiki," *Database*, vol. 2016, 2016.

[31] K. Hjorth-Jensen, A. Maya-Mendoza *et al.*, "SPOP promotes transcriptional expression of dna repair and replication factors to prevent replication stress and genomic instability," *Nucleic acids research*, vol. 46, no. 18, pp. 9484–9495, 2018.

[32] K. C. Cotto, A. H. Wagner *et al.*, "DGIdb 3.0: a redesign and expansion of the drug–gene interaction database," *Nucleic Acids Research*, vol. 46, no. D1, pp. D1068–D1073, jan 2018. [Online]. Available: https://academic.oup.com/nar/article/46/D1/D1068/4634012

[33] G. Shapovalov, R. Skryma, and N. Prevarskaya, "Calcium channels and prostate cancer," *Recent patents on anti-cancer drug discovery*, vol. 8, no. 1, pp. 18–26, 2013.

[34] T. Kanazawa, K. Misawa *et al.*, "G-protein-coupled receptors: Next generation therapeutic targets in head and neck cancer?" *Toxins*, vol. 7, no. 8, pp. 2959–2984, 2015.

[35] M. Dagenais, J. Dupaul-Chicoine *et al.*, "The interleukin (IL)-1r1 pathway is a critical negative regulator of PyMT-mediated mammary tumorigenesis and pulmonary metastasis," *OncoImmunology*, vol. 6, no. 3, p. e1287247, Feb. 2017. [Online]. Available: https://doi.org/10.1080/2162402x.2017.1287247

[36] J. Chen, E. E. Bardes *et al.*, "ToppGene Suite for gene list enrichment analysis and candidate gene prioritization," *Nucleic Acids Research*, vol. 37, 05 2009.

[37] S. Ioffe, "Improved consistent sampling, weighted minhash and l1 sketching," in *2010 IEEE International Conference on Data Mining*, Dec 2010, pp. 246–255.

[38] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995. [Online]. Available: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1995.tb02031.x

[39] D. S. Wishart, Y. D. Feunang *et al.*, "DrugBank 5.0: a major update to the DrugBank database for 2018," *Nucleic Acids Research*, vol. 46, no. D1, pp. D1074–D1082, 11 2017. [Online]. Available: https://doi.org/10.1093/nar/gkx1037

[40] A. Doron and E. R. Edelman, "Coronary artery disease and diabetes mellitus," *Cardiology Clinics*, vol. 57, no. 1, pp. 439–455, 2014.

[41] R. Pryor and F. Cabreiro, "Repurposing metformin: an old drug with new tricks in its binding pockets," *Biochemical Journal*,

vol. 471, no. 3, pp. 307–322, 10 2015. [Online]. Available: https://doi.org/10.1042/BJ20150497

[42] S. Kulcu Cakmak, A. Cakmak *et al.*, "Pentoxifylline use in dermatology," *Inflammation & Allergy - Drug Targets (Formerly Current Drug Targets - Inflammation & Allergy)*, vol. 11, no. 6, pp. 422–432, 2012. [Online]. Available: https://www.ingentaconnect.com/content/ben/iadt/2012/00000011/00000006/art00002

PLACE
PHOTO
HERE

Enrico Blanzieri received a Laurea degree (cum laude) in electronic engineering from the University of Bologna, Italy, and the PhD in cognitive science from the University of Turin, Italy, in 1992 and 1998, respectively. Since 2012, he is Associate Professor at the Department of Information Engineering and Computer Science (DISI), University of Trento, Italy where he works on machine learning and bioinformatics.

PLACE
PHOTO
HERE

Toma Tebaldi is a computational biologist currently working as Associate Research Scientist at Yale School of Medicine. His research aims at understanding the RNA molecular mechanisms underlying dysregulation in human pathologies, by combining experimental and computational approaches.

PLACE
PHOTO
HERE

Valter Cavecchia holds an MSc degree in Physics from the University of Trento, Italy. He is working as research assistant at the Institute of Materials for Electronics and Magnetism of the National Research Council of Italy in Trento. His research interests include computer graphics algorithms, code optimization techniques and volunteer-based distributed systems.

PLACE
PHOTO
HERE

Francesco Asnicar received the MSc in Computer Science in 2014 and the PhD in Communications and Information Technology in 2019 at the University of Trento. Currently he is a postdoc fellow at the Segata Lab Computational Metagenomics, University of Trento. His main research interests are the study of the relationship of the gut microbiome with diet and the development of new software tools for metagenomic analyses, mainly in the field of computational phylogenetics.

Luca Masera received his PhD in Communications and Information Technology in 2019 at the University of Trento, DISI. Currently he is working as machine learning engineer at Nexoya, Zürich (Switzerland). His research interests include causality discovery and prediction of structured data.

PLACE
PHOTO
HERE

Gabriele Tomè received the BSc degree in Computer Science at the University of Trento in 2020 where he developed his final project at DISI in collaboration with personnel of Fondazione E. Mach. He is currently master student in Quantitative and Computational Biology at the same university.

Eleonora Nigro received the MSc in Quantitative and Computational Biology, CIBIO, University of Trento in 2019 and she is now a predoctoral fellow at the Segata Lab Computational Metagenomics, University of Trento. She is now working on the characterization of specific families of under-represented gut microbes using metagenomic sequencing.

PLACE
PHOTO
HERE

Enrica Colasurdo is master student in Quantitative and Computational Biology, CIBIO, University of Trento, where she previously received a BS in Physics. She is currently interning in computational biophysics, working on enhanced sampling techniques for molecular simulations.

Matteo Ciciani received the BSc degree in molecular biology from the University of Padova, Padova, Italy, in 2017 and he is currently a master student in Quantitative and Computational Biology at the University of Trento, CIBIO, Trento, Italy. His research interests include metagenomics, classification of expression profiles, machine learning approaches to biological data analysis and CRISPR/Cas9-mediated genome editing.

Chiara Mazzoni received the MSc in Quantitative and Computational Biology, CIBIO, University of Trento in 2019. In 2020 she did a Post-Graduation Internship at The Hebrew University of Jerusalem.

PLACE
PHOTO
HERE

Stefania Pilati holds a PhD in Agro-industrial Biotechnology from the University of Verona, Italy. She is a researcher in Plant biology at Fondazione E. Mach, Italy, where her main interest is on hormonal and oxidative stress signalling pathways in grapevine physiology. Her experience in biochemistry and omics data analysis brought her attention to biological data mining.