OXFORD

# Reconstructing the temporal progression of HIV-1 immune response pathways

**Siddhartha Jain[1], Joel Arrais[2], Narasimhan J. Venkatachari[3], Velpandi Ayyavoo[3] and Ziv Bar-Joseph[4],***

[1]Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, USA, [2]Department of Computer Science, University of Coimbra, Coimbra, Portugal, [3]Department of Infectious Diseases, University of Pittsburgh, Pittsburgh, PA, USA and [4]Computational Biology and Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA

*To whom correspondence should be addressed.

## Abstract

**Motivation:** Most methods for reconstructing response networks from high throughput data generate static models which cannot distinguish between early and late response stages.

**Results:** We present TimePath, a new method that integrates time series and static datasets to reconstruct dynamic models of host response to stimulus. TimePath uses an Integer Programming formulation to select a subset of pathways that, together, explain the observed dynamic responses. Applying TimePath to study human response to HIV-1 led to accurate reconstruction of several known regulatory and signaling pathways and to novel mechanistic insights. We experimentally validated several of TimePaths' predictions highlighting the usefulness of temporal models.

**Availability and Implementation:** Data, Supplementary text and the TimePath software are available from http://sb.cs.cmu.edu/timepath

**Contact:** zivbj@cs.cmu.edu

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

High throughput data measuring various aspects of several biological systems is rapidly accumulating. These include RNA-Seq studies (Mortazavi *et al.*, 2008), profiling of microRNAs (Vergoulis *et al.*, 2012), ChIP-Seq, epigenetics studies (Gifford *et al.*, 2013), information about protein interactions within a cell (Prasad *et al.*, 2009) and information on interactions between host proteins and pathogen/environmental factors (Navratil *et al.*, 2009). Such datasets provide extensive information about the sets of genes that are activated, their regulation and their interactions both within a cell and between cellular proteins and the environment or pathogen. However, integrating these datasets to reconstruct a unified view of the networks and pathways that are activated in order to identify potential interventions that may lead to a desired response remains a major challenge.

Several methods have been proposed to integrate various biological datasets for this task (Huang *et al.*, 2009). However, the vast majority of these methods are aimed at obtaining static networks that do not provide temporal information making it hard to determine the various stages associated with the system being studied (for example, waves of expression changes (Chang *et al.*, 2013)) or the optimal time to apply an intervention. Consider the HIV-1 infection.

While the development of highly active antiretroviral therapy has made it possible to delay the progression of HIV infection, the persistence of the virus, rapid development of resistance and inability to completely eliminate the virus still pose major challenges for effective HIV-1 management (Shytaj and Savarino, 2013). HIV-1 infects a host cell by a sequential process involving several temporal events. These start with binding of the viral envelope protein to the host cell receptor followed by reverse transcription and integration of proviral DNA (early infection stage). Next, viral proteins are produced facilitating viral replication (intermediate stage). Finally, new viruses are released (late stage). While several studies have experimentally quantified the large scale changes and host-pathogen interactions for HIV-1 infection (Salgado *et al.*, 2011), to date no models exist to fully link these high throughput temporal datasets with the underlying dynamic networks that lead to the observed responses.

A small number of methods have been proposed for reconstructing dynamic interaction networks from high throughput data. These methods utilize the (relatively small number of) time series datasets to determine temporal information for the (mostly) static interaction datasets either directly (by projecting the time series data on the known interaction networks (de Lichtenberg *et al.*, 2005)) or indirectly (by looking at targets of transcription factors (TFs) and

associating temporal information for the interactions based on these targets (Ernst *et al.*, 2007; Schulz *et al.*, 2012). Since gene expression is the primary source of time series data these methods use, they have primarily focused on the reconstruction of regulatory networks (Bar-Joseph *et al.*, 2012). Signaling networks proved to be more challenging since much of the activity in these networks is post transcriptional (Filipowicz *et al.*, 2008) and often faster than regulatory networks which made it hard to use time series gene expression data to obtain temporal information about the activity of these networks.

Several other methods have been developed and evaluated for reconstructing regulatory networks using gene expression data (Haury *et al.*, 2012; Marbach *et al.*, 2012; Margolin *et al.*, 2006; Toepfer *et al.*, 2007). These methods utilize expression levels to determine regulatory interactions based on various statistical techniques including correlation, mutual information, regression, etc. While such methods can be successfully applied in some cases, they are less appropriate for modeling immune response dynamics since they cannot model post-transcriptional events (including the effects of virus–host and protein–protein interactions) which, as we show, play a major role in such responses.

To address these issues, two new methods have been proposed recently to jointly reconstruct dynamic signaling and regulatory networks by integrating static and time series data. SDREM (Gitter *et al.*, 2011) relies on a method for orienting protein interaction networks which are then combined with TFs and the networks they regulate using a separate input–output hidden Markov model (IOHMM). While SDREM has been successfully applied to study yeast and human response networks (Gitter and Bar-Joseph, 2013; Gitter *et al.*, 2013; Jain *et al.*, 2014) it does not provide temporal information about the pathways it finds. In SDREM, all pathways from source proteins (protein interacting with the environment/ pathogen) to TFs are assumed to be activated concurrently which does not explain expression waves and response phases. Further, SDREM does not optimize a single target function but rather two, separate, functions for different models (one for the IOHMM and the other for the combinatorial orientation algorithm) making it hard to determine optimal parameters for the networks. TimeXnet (Patil *et al.*, 2013) is another method for reconstructing such networks. It uses linear programming to formulate a max-flow problem imposing a constraint that the flow through expressed genes has to be greater than 0 so that they are accounted for in the networks identified. TimeXnet has been applied to study immune response in mice. However, TimeXnet does not directly consider the (often post-transcriptionally activated) source of the resulting response which may lead to missing important pathways. In addition, TimeXnet does not explain why some genes are activated early while others are only activated at a later stage.

Here, we present TimePath, a new method for reconstructing fully dynamic signaling and regulatory networks. TimePath uses a single Integer Programming (IP) based optimization function to jointly construct the networks. We initially select a large set of pathways that are rooted in source proteins and end in differentially expressed (DE) genes. This allows us to include sources that are only post-transcriptionally and/or post-translationally activated. Pathways for later DE genes are required to contain DE genes or miRNAs from earlier phases to explain their delayed response. Next, we use the IP to select a small subset of pathways that, together, explain the full set of DE genes. These selected pathways are analyzed to determine phase specific proteins and miRNAs and select those that are key to the response observed.

We applied TimePath to reconstruct dynamic models for HIV-1 immune response. As we show, the method accurately reconstructed the response networks identifying several known and novel pathways. We have performed experiments based on novel predictions made by TimePath several of which validated the ability of TimePath to determine a specific time for targeting a protein in order to reduce viral loads.

## 2 Methods

### 2.1 Cell culture, HIV infection and reagents

Sup-T1 cell lines were obtained through the NIH AIDS Research and Reference Reagent Program, Division of AIDS, NIAID, NIH (A Sup-T1 from Dr. James Hoxie (Smith *et al.*, 2008) and were maintained in RPMI containing 10% FBS, 1% L-glutamine and 1% penicillin streptomycin (GIBCO). HIV-wt-EGFP reporter virus was obtained by transfecting HEK293 T cells ($2 \times 10^6$ per plate) with 10 µg of HIV-1 vpr(+)/EGFP proviral construct by Polyjet following manufacturers protocol. Forty-eight hours post transfection, the supernatants were collected, filtered through a 0.4-µm filter to remove cellular debris, and centrifuged at 22 000 rpm for 1 h. The virus pellets were resuspended in PBS and stored in aliquots at 80 °C for subsequent assays. Multiplicity of infection (MOI) for virus was calculated by TZM blue assay using the HIV-1 reporter cell line cMAGI (AIDS Research and Reference Reagent Program [RRRP], National Institutes of Health [NIH]). The Sup-T1 cells were infected at a MOI of 0.3 either in the presence or absence of specific inhibitor at indicated time points. Forty hours post infection, the cells were washed and fixed with 1% paraformaldehyde and the samples were analyzed using Fortessa (BD Biosciences) with 10 000 gated events acquired for each sample, and the results were analyzed using FlowJo software (Tree Star, Inc., OR). The infected cells were detected by the expression of reporter virus EGFP. Azidothymidine (AZT) obtained from Sigma–Aldrich was used as positive control. IKK2 inhibitor V, Dasatinib and Dinaciclib were obtained from CalBiochem. SP600125 and WP 1066 were obtained from Abcam Biochemicals and Enzo, respectively. SNS-032, Regorafenib, Carfilzomib and Veliparib, Olaparib were obtained from selleckchem.com. SAHA and 5-Azacytidine were obtained from Sigma–Aldrich. The viability of cells was estimated by Trypan blue staining. We conducted the experiments three times with duplicate wells for each experiment.

### 2.2 Data description

The overall goal of TimePath is to determine the dynamics of both the signaling and the regulatory events that take place as part of a cellular response process. For this, TimePath integrates time series gene expression data, static protein interaction data (both within and across species) and protein–DNA interaction data. We constructed a weighted, partially directed, protein interaction network using several databases including BIOGRID (Stark *et al.*, 2006), HPRD (Prasad *et al.*, 2009) and have also used Post-translational Modification Annotations from the HPRD. Protein–DNA interactions are based on data from (Schulz *et al.*, 2012). Sources (host proteins that interact with the HIV-1 proteins) were obtained from VirHostNet (Navratil *et al.*, 2009). Time series gene expression and miRNA expression data following HIV-1 infection in Sup-T1 was obtained from (Mohammadi *et al.*, 2013). See Supplementary Methods for complete details.

### 2.3 Candidate pathways

To reconstruct the dynamic set of signaling pathways that are activated we first divide the time series gene expression data into $K$

phases. Initial response is likely driven by host proteins that interact directly with virus proteins. However, later changes in expression data (for example, expression changes that only occur 10 h after infection) are likely driven by genes or TFs that have been activated as part of an earlier expression response. In general we assume that expression changes in phase $i$ can be partially explained by activation/repression of a gene(s) in phase $i - 1$. To guarantee that our reconstructed pathways satisfy this we impose the constraint that any pathway that explains differential gene expression for a gene in phase $i > 1$ has to include at least one gene that was DE in phase $i - 1$.

Based on these assumptions we initially select a subset of pathways that can be used to explain the DE genes as follows:

1. We divide the time series into $k$ phases each consisting of $T/k$ time points where $T$ is the total number of points. We use $k = 3$ for this paper.
2. We extract the top $N_1$ DE genes for each phase (we use $N_1 = 200$). How the significantly DE genes are extracted and ranked is explained in Supplementary Methods.
3. We then search for the highest scoring $N_2$ acyclic paths from the source proteins (host proteins interacting with the virus of drug) to the targets (DE genes) for each phase (we use $N_2 = 10$ million here). We use the edge weights to compute a score for each path (Supplementary Methods). We also guarantee that the following constraints are satisfied for each pathway:

   a. The last edge in the path has to be a protein–DNA interaction (i.e. we need a TF to activate/repress the gene) (Yeang *et al.*, 2004).
   b. A path to a phase $i > 1$ target has to contain a node that is a target for phase $i - 1$.

In general, searching for the top N2 acyclic paths in a graph is a #P-complete problem which is not considered to be solvable efficiently (Arora and Barak, 2009). We thus use a heuristic to compute the set of paths. See Supporting Methods for a detailed description of the above process.

## 2.4 Integer program to select subset of pathways

Given a set of top paths for each target, our next goal is to combine them to identify the actual pathways that are activated as part of the response. Consider two targets $g_1$ and $g_2$ in phase $k$ that are known to be bound by the same TF $A$. If we believe that $A$ explains the activation of $g_1$ in that phase it increases our belief that $A$ is also the TF activating $g_2$. More generally, our goal is to select a subset of these pathways that, together, would minimize the number of intermediate signaling and regulatory proteins that are used across all pathways while at the same time maximize the number of targets that can be explained.

To accomplish this we define a new Integer Programming (IP) problem which includes three sets of binary variables (bv)

1. bv for a path to indicate whether it is selected or not.
2. bv for a target to indicate whether there is at least one path ending at it.
3. bv for protein to determine whether it is part of a path selected.

Using these variables we maximize the following objective

$$\max \sum_{p \in P} w(p) \cdot b_p^P + \lambda_1 \sum_{k=1}^{K} \sum_{g \in T_k} f_g - \lambda_2 \sum_{g \in G} b_g^G \qquad (1)$$

with the constraints

$$\forall p \in P, \forall g \in p, b_g^G \geq b_p^P \qquad (2)$$

$$\forall p \in P, \sum_{g \in P} b_g^G \leq |p| - 1 + b_p^P \qquad (3)$$

$$\forall g \in G, \forall p \in P(:, g), f_g \geq b_p^P \qquad (4)$$

$$\forall g \in G, \sum_{p \in P(:, g)} b_p^P \geq f_g \qquad (5)$$

where $K$ is the number of phases, $T_k$ is the targets for phase $k$, $P$ is the set of all paths, $G$ is the set of all genes, $P(:, g)$ is the set of paths ending at gene $g$, $w(p)$ is the weight of path $p$. The score of each a pathway $p$ is defined as $\Pi_{e \in E_p} \mathbb{P}(e)$ where $E_p$ is the set of edges in pathway $p$ and $\mathbb{P}(e)$ is the edge score which is defined in Supplementary Methods, $b_p^P$ is whether path $p$ is selected or not, $f_g$ is whether gene $g$ has even one selected path ending at it, $b_g^G$ is whether gene $g$ is selected, $\lambda_{1-2}$ are the weights for balancing the minimization requirements in terms of intermediate nodes and the maximization requirements in terms of the number of targets. They are the parameters that decide in the end, how large of a network in terms of number of genes and edges will be chosen.

Note that setting $b_g^G = 0$ for a specific gene immediately implies that $b_p^P$ for a path containing that gene is 0 and similarly that $f_g$ is 0 for that gene and so these variables are not independent as the constraints above imply. We set $b_p^P = 1$ if and only if all the genes in the path are selected as enforced by constraints 1–2. $f_g$ is 1 if and only if there's at least one path with $b_p^P = 1$ ending at the gene $g$ as enforced by constraint 3.

Since this is a problem with linear constraints, a linear objective and since the $b_g$ variables are binary, this is an IP and not an Linear Program (LP). The IP we are dealing with however is too large for standard IP solvers and we thus solve it using a greedy approach followed by a tabu search heuristic to escape local minimum. Briefly, we start with all the nodes selected. Then at each step, we search for a node whose addition or removal from network would increase the objective the most (this is accomplished by flipping the $b_n$ variable for that gene). Paths that contain a gene that is not in the current network are removed (i.e. their corresponding $b_p$ variable is 0). Once we find such a node, we add or remove it and keep going until we can find no node whose addition or removal will improve the objective. We randomly select nodes if there are ties between them. Thus the results can differ from one run to another—however, the actual genes selected by the network change little according to our experimental results. See Supplementary Results for details.

## 2.5 Ranking genes

After solving the IP we obtain a subset of the pathways that, combined, explain the observed expression response over time. While we attempt to minimize the number of proteins in these networks, we still end up with hundreds of proteins in the set of selected pathways. To identify key proteins for follow up analysis, we rank genes for each phase based on the 'path flow' going through them. The path flow $f$ through a node $n$ for phase $i$ is defined as follows.

$$f(n) = \sum_{p \in P} I(p) \cdot w(p)$$

where $P$ is the set of paths ending at a target in phase $i$ and containing node $n$. $I(p)$ is 1 when the path $p$ is selected and 0 otherwise. We further refine the phase specific genes for later phases to remove

those already identified by earlier phases. See Supporting Results for details.

## 3 Results and discussion

### 3.1 TimePath analysis of expression and interaction data

To identify the dynamic pathways that are activated by biological response processes, TimePath uses the time series gene expression to annotate the static protein–protein and protein–DNA interaction data (Fig. 1). This is achieved by formulating an Integer Programming (IP) optimization function that balances the ability to explain the DE genes at different time points with the requirement that relatively few of all possible pathways are used in each of the different responses. Following the application of IP to our modeling problem we obtain a sparse set of pathways, each associated with a specific experimental phase (or subset of time points), that together explain the expression profiles observed as part of the response. Each of these pathways is rooted at a source protein (a protein that directly interacts with the infecting agent or with the environment) allowing us to link the expression observed to its causes. In addition, to explain the different expression waves we require that pathways leading to genes that are DE at later time points contain at least one gene that is DE at an earlier time point. The complete set of pathways obtained by TimePath is then analyzed to identify key proteins and determine potential interventions that can block the response observed (Section 2).

### 3.2 TimePath analysis of HIV data

We used TimePath to examine cell response to HIV infection. Time series expression data for HIV-1 was obtained from Mohammadi et al. (2013) which profiled genes using SAGEseq every 2 h for 24 h after transfection with HIV-1 in Sup-T1 cell line. Expression data was normalized using DESeq (Anders and Huber, 2010). In addition to HIV expression data we obtained interaction data for HIV-1 proteins and host (human) proteins from VirHostNet (Navratil et al.,



**Fig. 1.** Overview of TimePath. (Left) Several time series and static datasets are used as input. (Top right) Based on these inputs an initial set of pathways is selected such that each starts at a source (a host protein interacting with a virus protein), ends in a target (a DE gene for one of the time points) and contains PPI and Protein–DNA edges. (Middle right) Next, Integer Programming (IP) is performed to select a subset of these pathways. (Bottom right) The resulting pathways explain the dynamics of the observed response including the different expression phases observed for different genes

2009). Of the 235 proteins in VirHostNet, 231 are present in our protein–protein interaction (ppi) network and were used as potential sources.

TimePath also uses general protein–protein interactions from BIOGRID (Stark et al., 2006) and HPRD (Prasad et al., 2009), Post-translational Modification Annotations from HPRD and Protein–DNA interaction data (Schulz et al., 2012) (Section 2).

To identify pathways for specific response phases we divided the time series expression into 3 phases (every 8 h) and extracted 200 targets (DE genes) for each phase (Section 2). We next used the static interaction data to identify a large number of potential pathways connecting sources and targets constraining potential pathways for *later targets* to contain a gene that is DE at an earlier phase. A subset of these pathways that, together, explain the observed response to HIV infection are then selected by the IP method. Pathways retained by the IP for this data included a total of 607 genes of which 319 are targets. We next ranked proteins in these pathways based on their importance to each phase (Section 2).

### 3.3 Pathways and proteins identified for HIV response

The resulting dynamic network is presented in Figure 2. Top ranked proteins for each of the three phases are presented in Table 1 and Supplementary Tables S1–S3.

The dynamics observed by TimePath provide important insights into the mechanisms used by the HIV-1 to proliferate and overcome host defenses. Several of the proteins identified as controlling the early phase response are related pathways that either promote viral replication (e.g. P53) or suppress immune response pathways (e.g. AP1B1, AP2B1, CALM3) are known to be key participants in HIV pathways (Greenway et al., 2002). Several of the signaling proteins and TFs controlling the later phases in the reconstructed TimePath model are also related to promotion of virus activity. These include proteins controlling virus elongation (GTF2H1), down regulated proteins for cell cycle arrest (CDC34, LCK (Strasner et al., 2008)) and several down regulated proteins that are involved in immune response (PTPN7, VAV1) and superinfection prevention (ACTB). However, later phases of the model also contain proteins that are part of the host defense response (much more than phase 1). These include immune response factors such as STAT and JUN (Phase 2) (Mak and Saunders, 2006) and the down regulation of apoptosis inhibitors (e.g. DDIT3. Thus, while the network reveals the strategy utilized by the virus to circumvent initial innate immune response, it also identifies the key factors of the cellular response networks that escape viral regulation and are utilize by the cell to respond to the infection. See Supplementary Tables S8–S10, Supplementary Results and Discussion below for more details.

### 3.4 Statistical validation of the reconstructed network and comparison with other methods

To more globally assess the ability of TimePath to accurately identify pathways and proteins, and to compare its performance with prior methods that were developed to reconstruct dynamic signaling and regulatory networks we used several complementary datasets to test the reconstructed pathways.

While several methods have been proposed for reconstructing biological networks [28], relatively few are focused on analyzing dynamic response networks. These include SDREM (Gitter and Bar-Joseph, 2013; Gitter et al., 2011), which combines a HMM method for modeling dynamic regulatory networks with a combinatorial algorithm for signaling network reconstruction and TimeXnet (Patil et al., 2013) which uses a linear programming (LP) formulation to
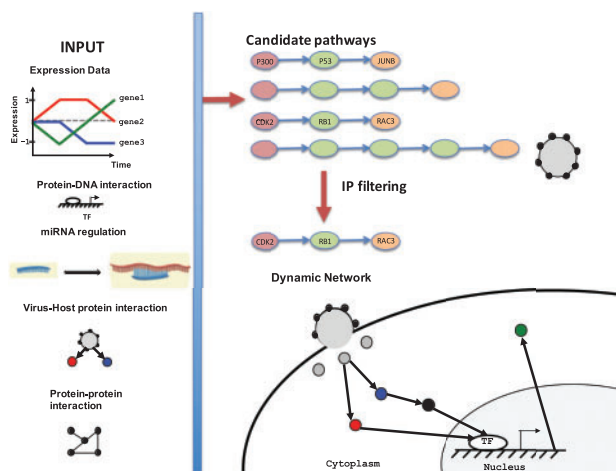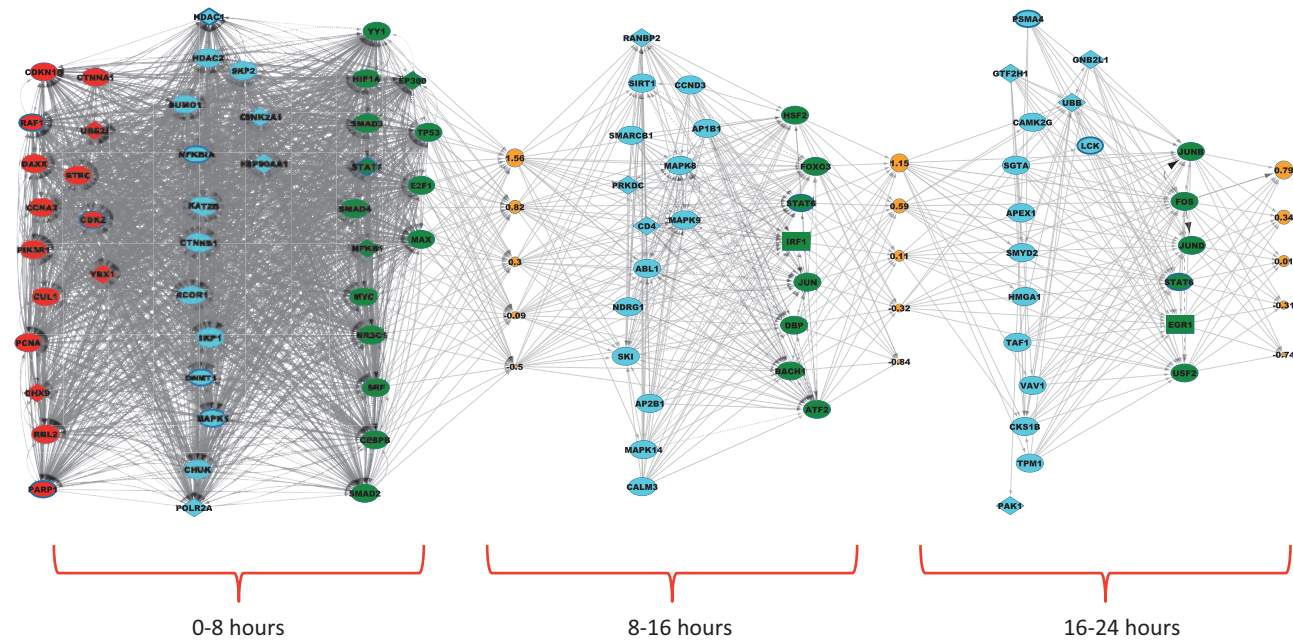
**Fig. 2**. Dynamic signaling and regulatory network for HIV-1 immune response. The red nodes are the host proteins that interact with the HIV-1 proteins (selected sources). Blue nodes are intermediate signaling proteins and green nodes are the TFs that are predicted to directly up/down-regulate the differential expression of target genes (targets not shown in figure, but the average levels of the regulated targets for each TF is presented by the yellow nodes while the size of each of the yellow nodes indicates how many genes belong to the cluster represented by the node). The figure displays the top predicted nodes for each of the three phases and also demonstrates is directly linked to the sources via the signaling proteins and DE genes in earlier phases. Diamond shaped nodes were identified as supported RNAi screen hits (text) and rectangular nodes are targets for the phase they are in. Nodes with bold blue border represent proteins we experimentally tested. Note that some intermediate proteins may also be TFs. The functional role in the network figure is based on the location of the protein in the selected paths based on the IP

find important genes. Note that neither of these methods uses miRNA expression data and so we constrained our comparison to TimePath models that do not utilize such data (models using miRNA expression data are discussed in Section 3.6).

In addition to comparing TimePath with prior methods that construct both signaling and regulatory networks, we have also compared the top ranked genes from TimePath to the top DE genes in the dataset (Supporting Methods) since several methods for analyzing gene expression data still focus on such DE genes (Rapaport *et al.*, 2013).

### 3.4.1 RNAi screen hits
First, we looked at RNAi screen experiments which test the impact of gene knockdown on HIV viral load. Three such experiments were conducted though a meta-analysis of the results determined that only three proteins were detected by all studies (Bushman *et al.*, 2009). We have filtered the combined list to select a subset of the hits that are supported by at least two lines of evidence (Supplementary Results) resulting in 389 supported hits, 364 of which were present in our initial network.

The results are in Table 2. We find that the pathways obtained by TimePath are significantly enriched for screen hits (P-value of $1.7 \times 10^{-17}$). This significant overlap also holds separately for each the subset of proteins identified for the three phases (Supplementary Tables S1–S3). We next compared these results to results from the other two network reconstruction methods and to the top DE genes. For this comparison we ranked the genes using path flow for TimePath and SDREM (Section 2) and used the TimeXnet output ranking for that method. The RNAi overlap is presented in Tables 2. As can be seen, rankings for all network reconstruction methods greatly outperforms the DE genes rankings highlighting the

importance of post-transcriptional and post-translational events in the response process. Further, both TimePath and SDREM significantly outperform TimeXnet in this analysis with almost a quarter of the top ranked genes supported by screen hits.

### 3.4.2 Analysis using GO and Reactome
To further analyze the pathways identified by TimePath we looked at the agreement between them and two complementary databases: The Gene Ontology (GO) and the set of HIV curated pathways in Reactome. GO analysis was performed on the top 100 genes (nodes) identified based on the path flow metric (Section 2) using FuncAssociate (Berriz *et al.*, 2003) while Reactome analysis was performed using the set of pathway edges. The results indicate that the pathways obtained by TimePath agree very well with known pathways involved in HIV response. The full list of enriched GO categories (corrected *P*-value $\leq 0.001$) is presented on the Supporting Website and includes 'toll-like receptor signaling pathway', an important component of innate immune response (Mak and Saunders, 2006), 'positive regulation of defense response', 'innate immune response-activating signal transduction', etc. We also find that TimePath achieves a higher number and a higher percentage of significantly enriched immune related categories compared to SDREM and TimeXnet 4 using the FuncAssociate (Berriz *et al.*, 2003) tool. We compared the % of significantly enriched GO categories that were immune response related (Supporting Methods). TimePath again has a both a slightly higher number and a higher percentage of significantly enriched immune related categories compared to SDREM and TimeXnet (Table 4).

Results for Reactome are presented in Table 2, Supplementary Figure S4 and Supplementary Table S6. As can be seen, we achieve a significant overlap between edges in the selected pathways and those

**Table 1.** Phase ranking for top genes

| Phase | Gene | R1 | R2 | R3 | Expression change direction |
|---|---|---|---|---|---|
| 1 | EP300 | 1 | 2 | 2 | Up |
| 1 | TP53 | 2 | 5 | 4 | Up |
| 1 | HDAC1 | 3 | 6 | 6 | Up |
| 1 | RELA | 4 | 20 | 10 | Up |
| 1 | RB1 | 5 | 4 | 3 | Up |
| 1 | BRCA1 | 6 | 8 | 11 | Up |
| 1 | PCNA | 7 | 11 | 9 | Up |
| 1 | SUMO1 | 8 | 9 | 8 | Up |
| 1 | HDAC2 | 9 | 22 | 14 | Up |
| 1 | CEBPB | 10 | 21 | 12 | Up |
| 1 | DNMT1 | 15 | 23 | 15 | Up |
| 1 | STAT1 | 27 | 25 | 33 | Up |
| 1 | RAF1 | 28 | 66 | 39 | Up |
| 1 | CDK2 | 29 | 59 | 41 | Up |
| 2 | JUN | NP | 1 | 1 | Up |
| 2 | ATF2 | 143 | 7 | 5 | No change |
| 2 | CALM3 | 127 | 10 | 106 | Up |
| 2 | CD4 | 136 | 12 | 109 | Up |
| 2 | STAT5B | 105 | 13 | 86 | Up |
| 2 | CCND3 | 91 | 14 | 100 | Up |
| 2 | SMARCB1 | 92 | 15 | 97 | Up |
| 2 | AP1B1 | 124 | 16 | 114 | Up |
| 2 | SKI | NP | 17 | 147 | Up |
| 2 | AP2B1 | 138 | 18 | 130 | Up |
| 3 | FOS | NP | NP | 7 | Up |
| 3 | PSMA4 | NP | NP | 23 | Up |
| 3 | DDIT3 | NP | NP | 25 | Up |
| 3 | GTF2H1 | NP | NP | 26 | No change |
| 3 | SGTA | NP | NP | 36 | Down |
| 3 | JUNB | 224 | 148 | 38 | Down |
| 3 | JUND | 276 | NP | 40 | Down |
| 3 | GNB2L1 | 118 | NP | 46 | No change |
| 3 | UBB | 113 | 155 | 47 | Down |
| 3 | VAV1 | 112 | 170 | 49 | Down |
| 3 | LCK | NP | NP | 291 | Down |

R1/2/3 indicates the rank of the gene in phase 1/2/3. If the rank is 'NP', that means the gene was not found to be present in the phase. Genes tested experimentally are colored red (see Supplementary Tables S1–S3 for complete rankings). For later phases we focused on genes that were ranked high for these phases compared to their rank in an earlier phase. Genes with absolute log fold change expression <2 are designated as not being differential expressed.

present in the HIV Reactome pathways. Comparison with the other methods clearly demonstrates the advantages of TimePath which is able to identify a much larger number of correct interactions than the other two network reconstruction methods. Note that Reactome comparison is not available for the DE gene list since it does not contain interactions.

We have also analyzed the usefulness of the various stages of TimePath. As can be seen in Table 3, each step in the TimePath method further improves the overlap with the screen hits. Initially, only 3.7% of the expressed genes are screen hits. The initial pathway extraction step increases the overlap to 10% while the overlap following IP increases to 14%.

Finally, we investigated the impact of the constraint imposed on later paths in our network to include a DE gene from an earlier phase. As we show in Table 5, we obtain almost three times as many edges in the overlap compared to the network without the time constraint with correspondingly better P-value.

**Table 2.** Overlap between RNAi screen hits and top 100 genes for the different dynamic network reconstruction methods and between edge list from Reactome (1265 edges in network) and the edges extracted by the different methods

| Method | Overlap with screen hits | P-value | Overlap with Reactome edges | P-value |
|---|---|---|---|---|
| TimePath | 23 | $1.7 \times 10^{-17}$ | 101/3203 | $7.9 \times 10^{-44}$ |
| SDREM | 21 | $3.2 \times 10^{-16}$ | 74/3203 | $3.9 \times 10^{-24}$ |
| TimeXnet | 16 | $4.9 \times 10^{-10}$ | 54/2585 | $3.9 \times 10^{-16}$ |
| DE ranking | 5 | 0.23 | NA | NA |

Comparison with a baseline ranking of the differentially expression (DE) genes is also presented.

**Table 3.** Overlap with HIV screen hits at various stages of the algorithm

| Stage | Overlap | Overlap % |
|---|---|---|
| Pre-algorithm | 364/16 671 | 2.1 |
| Unexpressed genes filtered | 246/6604 | 3.7 |
| After pathway search | 144/1374 | 10.4 |
| After IP | 85/607 | 14.0 |

'Pre-algorithm' is the initial overlap for all genes in the network. 'Unexpressed genes filtered' is when we remove all genes from our interaction network that are unexpressed. 'After pathway search' is that stage that uses all genes included in the initial top scoring set of pathways. 'After IP' is the final stage after the IP (and thus the whole algorithm) has run. As can be seen, the IP step seems to further improve the resulting set of genes indicating that the selection process indeed identifies HIV response pathways.

**Table 4.** GO comparison

| Method | % of immune-related categories | P-value |
|---|---|---|
| TimePath | 11.16 (72/645) | $2.074 \times 10^{-5}$ |
| SDREM | 8.04 (71/883) | 0.077 |
| TimeXnet | 10.44 (66/632) | $3.18 \times 10^{-4}$ |

We give the % of immune-related categories as well as the absolute number of immune related categories and total categories enriched for in parenthesis. The P-value cutoff for all categories was 0.05. The GO enrichment was performed on the top 100 genes as ranked by path flow (Section 2) using the FuncAssociate tool (Berriz et al., 2003).

**Table 5.** Validation for the time constraint.

| Method | Overlap | P-value |
|---|---|---|
| TimePath | 101/3203 | $7.9 \times 10^{-44}$ |
| TimePath without time constraint | 37/3203 | $3.6 \times 10^{-5}$ |

### 3.5 Experimental results

To experimentally test the temporal predictions of TimePath we selected top ranking phase proteins for which we could obtain commercial inhibitors and examined the impact of blocking these proteins at various time points in the response (Fig. 3). Note that the RNAi knockdown screens discussed above were performed on a different cell type (Hela/TZM-bl and 293T) and so, while they are useful for statistical validation, they may not completely reflect pathways activated in Sup-T1 cells. More importantly, these screens
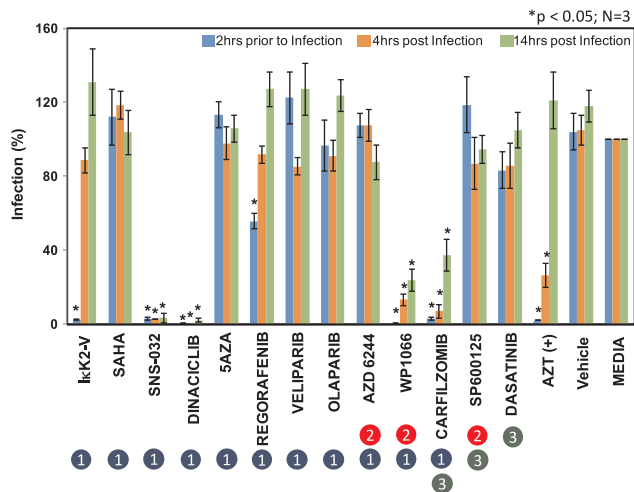
**Fig. 3.** Experimental validations. Relative infection after treatment with inhibitors. Significant changes in infection are highlighted with a *. The inhibitor names are given on the *X* axis and the target proteins of the inhibitors are given in parenthesis. See also supporting Figure S3 for the full list of targets for each inhibitor

do not provide information about the dynamics of the response while our experiments are aimed at testing not just the predictions regarding top ranked proteins but also their phase specific assignment. We performed experiments in which we varied the time of applying the inhibitors w.r.t the infection time. For each of the proteins tested, inhibitors were applied 2 h prior to infection (phase 1), 4 h (phase 2) and 14 h (phase 3) post infection. Amount of infection was determined at 40 h post infection for all experiments. We concurrently measured cell viability to test the toxicity of the inhibitor (Supplementary Fig. S1).

The results are presented in Figure 3. As can be seen, for five of the inhibitors we tested (targeting 11 of the 22 proteins tested) we observed a significant impact on viral load as predicted by TimePath. Note that the screen results indicate that less than 1.5% of all proteins lead to decreased viral load, and so such a high validation rate is a strong indication for the accuracy of TimePath. Importantly, several of the time specific predictions were validated in these experiments. We expected that inhibiting proteins that are ranked at the top for all phases or for phase 3, at any time, would lead to reduction in viral load since even early inhibition prevents them from being activated at a later stage. We indeed see this effect for the STAT inhibition (ranked in the top 30 for all phases) and for PSMA4 (ranked at the top only for phase 3). In contrast, for proteins ranked high in phase 1 and lower at the next phases we expected to see a much greater impact for the early treatment vs. later ones since their impact may have already been exerted by the time of the later treatments. This is exactly what we see for two of these proteins. For both NFKB1 (ranked 14 in the first phase but dropping to 50 in the 2nd) and for Raf1 (dropping from 28 to 66) we see significant response when treated early but a much lower impact on viral load when treated at later stages strongly supporting TimePath's predictions. Published studies suggest that NF-kB has a major role in HIV-1 transcription due to it is binding sites in HIV-1 LTR and TAR-RNA (Kwon *et al.*, 1998; Takada *et al.*, 2002; Tareq Hassan Khan *et al.*, 2012; Williams *et al.*, 2007; Wires *et al.*, 2012). Results from our analyses predicted a role for NF-kB during the early phase (phase 1) and blocking this TF inhibited virus replication only in pretreatment (2 h) and did not affect virus replication when treated at the later stages and this effect is independent of cellular toxicity.

Similarly, another protein Raf1, predicted as early phase response to HIV-1 also exhibited similar phase dependent inhibition. Though Raf1 is known to interact with HIV-1 Nef and perturb T cell signaling and activation pathway (Hodge *et al.*, 1998), the mechanisms by which Raf1 exerts its effects is unclear. It is possible to predict that blocking Raf1 might have an effect on the function of HIV-1 early protein Nef, thus altering T cell signaling and virus infection. Another phase 1 protein, CDK2 (dropping from 29 to 59) also showed strong impact when treated at the early time point but unlike the other phase 1 predictions, later treatments continued to have a significant impact on viral loads. CDK is known to play a role in HIV-1 transcription by the viral transactivator, Tat (Cujec *et al.*, 1997), thus there is a direct correlation predicted by TimePath. However, blocking CDK using inhibitors blocked both at the early and late phase suggest that these inhibitors might have direct and indirect effect on virus replication.

PSMA41 is part of the proteasomal complex and so inhibiting this protein with Carfilzomib not only blocks the proteasomal pathway, but could also alter additional cellular processes such as sumoylation, ubiquitination and Cul1 activity. These results are further supported by the early time points predictions that identified SUMO1, UBE2I and CUL1 in Phase 1. Sumoylation of HIV-1 integrase is essential for efficient viral replication (Zamborlini *et al.*, 2011) and cullin ligases are recruited by HIV-1 viral proteins to overcome host viral restriction factors, HIV-1 Vif degrades APOBEC proteins (Goila-Gaur *et al.*, 2008) and HIV-1 Vpr induces degradation of UNG and SMUG uracil-DNA glycosylases (Schröfelbauer *et al.*, 2005). Also HIV-1 Vpr is known to interact with damaged DNA binding protein 1 (DDB1) to induce G2/M arrest which contributes to efficient viral replication (Hakata *et al.*, 2014). Indeed, many of the factors predicted for the early stage response (Phase 1: 0–8 h) are related to DNA modification and chromatin remodeling (HDAC1, HDAC2, DNMT1, KAT2B) and cell cycle (CTNNB1, CSNK2A1, CDK2, E2F1). Also there is an enrichment of transcription factors (P53, RELA, NFKB1, NR3C1, Stat1, MYC, RAF1, TBP, YY1), which have binding sites on HIV-1 LTR. These factors may have a role in integration of proviral DNA and regulation of HIV-1 transcription.

### 3.5.1 The role of late activated pathways
A key advantage of TimePath is its ability to highlight key pathways that are only activated later in the response. As we show in Supplementary Results and in Supplementary Table S12 some of these pathways also support the ability of the virus to promote its replication and dissemination by facilitating elongation of viral transcripts, preventing reinfection, promoting survival of infected cells and by immune evasion. Previous results have shown that viral proteins Nef, Env and Vpu modulate the surface expression of critical immune molecules such as CD4, CD28, MHC class I and others through protein–protein interaction (Haller *et al.*, 2014). Our results show that the virus can also regulate the expression levels of CD4, AP1 subunits and other related genes. This suggests that the virus has additional mechanisms to prevent super infection. However, later phase regulators identified by TimePath also contain several proteins related to host defense mechanisms which are activated to suppress virus infection (Supplementary Table S12). Similar to the pathways discussed above, which were ranked based on the network scoring technique (Section 2), when we classified the expression changes of individual genes we observed that the majority of the changes that occur during early stages tend to facilitate virus infection, whereas the host defense responses are observed largely at the later phases. Though multiple HIV-1 repressor genes are down regulated early on, there are certain TF genes, including TP53, RELA, NR3C1 which can bind sites on HIV LTR, that are also repressed.

This suggests that the virus has evolved ways to differentially regulate specific TFs to enhance virus production while at the same time trying to prevent a boost to the immune response, as many of these TFs are also binding in front of immune response genes. The dynamic network also explains how genes that are differentially regulated at later stages of virus infection (Supplementary Table S12) result from changes observed in earlier Phases. TimePath analysis underlines the temporal relation of the well-established ability of virus to exploit the host machinery.

### 3.6 Incorporating miRNAs to TimePath models

Similar to TFs, miRNAs have also been shown to regulate the expression of mRNAs. In most cases the effects of this regulation is inhibitory including interference with translation (Alberts *et al.*, 2007), and binding to mRNAs to expedite their degradation (Eulalio *et al.*, 2009). We can easily extend TimePath to include miRNAs by incorporating into the analysis pipeline two new types of edges: (i) Edges representing interactions between from miRNAs and their targets and (ii) edges representing the regulation of miRNAs themselves by TFs (Wang *et al.*, 2010). To determine the set of edges we used the TargetScan database (Grimson *et al.*, 2007) and ENCODE data (Section 2). Using these new edges miRNAs can be treated as any other node (genes) in our model with two exceptions: (i) Unlike TFs, which can be post-transcriptionally regulated, miRNAs are only regulated transcriptionally and so we only include a miRNA in our model if it is DE in at least one of the phases and (ii) We require that gene targets of miRNAs be anti-correlated with the expression of the miRNA regulating them (to model the inhibitory impact). See Supplementary Methods for details.

The miRNA genome locations were taken from miRbase (Griffiths-Jones *et al.*, 2006), release 20 which consists of 4446 miRNAs, 499 of which are present in the RNA-seq expression dataset we used. The raw sequence counts of the miRNAs are given in the expression dataset for the same timepoints as the genes and are normalized in the same fashion as the gene expression counts are. We re-ran TimePath using these modifications and the additional miRNA expression data. As can be seen in Supplementary Figure S2, the resulting network contained most of the proteins/genes that were included in the original network for the different phases. However, the network also identified a number of miRNAs as controlling the expression of phase 1 gene expression. Specifically, of the 499 miRNAs we analyzed 16 were selected for the final network (5 are shown in Supplementary Figure S2).

To determine the relevance of the selected miRNAs, we evaluated the list by using Ingenuity pathway analysis (2015) (Supporting Methods). Five of the 17 validated HIV miRNAs in Ingenuity were within the group of 16 miRNAs selected by TimePath which is almost 10 times more than expected by random chance (*P*-value $8.5 \times 10^{-5}$). The full list of 16 is presented in Supplementary Table S13. These include miR-148a which was shown to significantly control the HIV virus via its regulation of HLA-C expression (Kulkarni *et al.*, 2011, 2013), miR-27a and miR-27b which are part of a class of miRNAs that have been found to affect HIV infection (Chiang *et al.*, 2012) and miR-214 which is known to exhibit broad antiviral activity (Hayes *et al.*, 2011).

## 4 Discussion

Since most of the high throughput data used to reconstruct cellular response networks is static, current models based on these data are often unable to provide specific *temporal* hypotheses regarding the effects of perturbations and drugs on cellular responses. Here, we formulated a new Integer Programming (IP) optimization function to connect observed temporal responses (from gene expression data) with the underlying sources, to further identify the pathways and transcription factors that activate them. We then use the pathways and their predicted time to reconstruct the full response network leading to insights regarding the propagation of cellular responses, key proteins controlling the responses and testable hypothesis regarding the effects of perturbing proteins at various time points following infection.

Applying TimePath to model HIV response networks led to the identification of known and novel proteins and miRNAs for the HIV response pathways. The reconstructed network explains the roles of several HIV screen hits, the function of TFs and miRNA controlling expression levels and is enriched for functional categories related to immune and viral responses.

The pathways identified can be divided to those induced by the virus to promote survival/replication and those induced by the host to curtail virus infection and promote cellular survival. Our temporal regulatory model indicates that these can also be divided based on their dynamics.

Follow up experiments using inhibitors confirmed the prediction of TimePath, where 11 of the 22 predicted proteins (that were evaluated in the experiment) were identified to have a role in HIV infection. NFKB and related genes are exclusively essential for virus infection in the initial phase as predicted by TimePath, similarly, RAF1 was also confirmed to have an important role in the initial phase. As predicted by TimePath, these genes may either be required for virus infection during the initial phase, or the changes triggered by these genes in the initial phase can temporally affect downstream events that are essential for virus infection. It is also noted that CDKs, STATs and proteasomal machinery are essential during all phases of HIV infection, and TimePath had predicted a role for these genes starting with phase 1 (CDKs) and/or a combination of phases—phase 1 and phase 2 (STATs) or phase 3 (proteasomal machinery and related processes). Though TimePath identifies the role for these genes or processes in specific phase, it suggests that the event occurs at the identified phase; however, it does not rule out that the events are continuing over time and have a role in later stages too.

Unlike other methods that attempt to link treatments to disease stages (for example, in cancer which uses pathological analysis to determine tumor grades) TimePath is fully based on the molecular data, thus could be applied to much shorter time scales. This approach enables the programme to obtain a more fine resolution of the disease stage, which cannot be observed by other methods. With higher resolution, it may be possible to use TimePath to tailor appropriate treatment options to treat infected individuals.

## References

Alberts,B. *et al.* (2007) *Molecular Biology of the Cell*, 5th ed. Garland Science, New York.

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

Arora,S. and Barak,B. (2009) *Computational Complexity: A Modern Approach*. Cambridge University Press, Cambridge, UK.

Bar-Joseph,Z. *et al*. (2012) Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet*., **13**, 552–564.

Berriz,G.F. *et al*. (2003) Characterizing gene sets with FuncAssociate. *Bioinformatics*, **19**, 2502–2504.

Bushman,F.D. *et al*. (2009) Host cell factors in HIV replication: meta-analysis of genome-wide studies. *PLoS Pathogens*, **5**, e1000437.

Chiang,K. *et al*. (2012) Regulation of cyclin t1 and HIV-1 replication by micrornas in resting CD4+ t lymphocytes. *J. Virol*., **86**, 3244–3252.

Chang,K.N. *et al*. (2013) Temporal transcriptional response to ethylene gas drives growth hormone cross-regulation in arabidopsis. *Elife*, **2**, 00675.

Cujec,T.P. *et al*. (1997) The HIV transactivator tat binds to the cdk-activating kinase and activates the phosphorylation of the carboxy-terminal domain of RNA polymerase II. *Genes Devel*., **11**, 2645–2657.

de Lichtenberg,U. *et al*. (2005) Dynamic complex formation during the yeast cell cycle. *Science*, **307**, 724–727.

Ernst,J. *et al*. (2007) Reconstructing dynamic regulatory maps. *Molecular Systems Biology*, **3**, 74.

Eulalio,A. *et al*. (2009) Deadenylation is a widespread effect of miRNA regulation. *RNA*, **15**, 21–32.

Filipowicz,W. *et al*. (2008) Mechanisms of post-transcriptional regulation by micrornas: are the answers in sight? *Nat. Rev. Genet*., **9**, 102–114.

Gifford,C.A. *et al*. (2013) Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. *Cell*, **153**, 1149–1163.

Gitter,A. and Bar-Joseph,Z. (2013) Identifying proteins controlling key disease signaling pathways. *Bioinformatics*, **29**, i227–i236.

Gitter,A. *et al*. (2011) Discovering pathways by orienting edges in protein interaction networks. *Nucleic Acids Res*., **39**, e22–e22.

Gitter,A. *et al*. (2013) Linking the signaling cascades and dynamic regulatory networks controlling stress responses. *Genome Res*., **23**, 365–376.

Goila-Gaur,R. and Strebel,K. (2008) HIV-1 vif, apobec, and intrinsic immunity. *Retrovirology*, **5**, 10–1186.

Greenway,A.L. *et al*. (2002) Human immunodeficiency virus type 1 nef binds to tumor suppressor p53 and protects cells against p53-mediated apoptosis. *J. Virol*., **76**, 2692–2702.

Griffiths-Jones,S. *et al*. (2006) mirbase: microrna sequences, targets and gene nomenclature. *Nucleic Acids Res*., **34**, D140–D144.

Grimson,A. *et al*. (2007) Microrna targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, **27**, 91–105.

Hakata,Y. *et al*. (2014) Interactions with dcaf1 and ddb1 in the crl4 e3 ubiquitin ligase are required for vpr-mediated g2 arrest. *Virol. J*., **11**, 1–11.

Haller,C. *et al*. (2014) HIV-1 nef and vpu are functionally redundant broad-spectrum modulators of cell surface receptors, including tetraspanins. *J. Virol*., **88**, 14241–14257.

Haury,A.C. *et al*. (2012) Tigress: trustful inference of gene regulation using stability selection. *BMC Syst. Biol*., **6**, 145.

Hayes,A.M. *et al*. (2011) Tat RNA silencing suppressor activity contributes to perturbation of lymphocyte miRNA by HIV-1. *Retrovirology*, **8**, 36.

Hodge,D.R. *et al*. (1998) Binding of c-raf1 kinase to a conserved acidic sequence within the carboxyl-terminal region of the HIV-1 Nef protein. *J. Biol. Chem*., **273**, 15727–15733.

Huang,SS, and Fraenkel,E. (2009) Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. *Sci. Signal*., **2**, ra40.

Jain,S. *et al*. (2014) Multitask learning of signaling and regulatory networks with application to studying human response to flu. *PLoS Comput. Biol*., **10**, e1003943.

Ingenuity pathway analysis. (2015) http://www.ingenuity.com/products/ipa.

Kulkarni,S. *et al*. (2011) Differential microRNA regulation of HLA-C expression and its association with HIV control. *Nature*, **472**, 495–498.

Kulkarni,S. *et al*. (2013) Genetic interplay between HLA-C and MIR148A in HIV control and Crohn disease. *Proc. Natl. Acad. Sci*., **110**, 20705–20710.

Kwon,H. *et al*. (1998) Inducible expression of iκbα repressor mutants interferes with nf-κb activity and HIV-1 replication in Jurkat t cells. *J. Biol. Chem*., **273**, 7431–7440.

Mak,T.W. and Saunders,M.E. (2006) *The Immune Response: Basic and Clinical Principles*, vol. 1. Academic Press, Cambridge, MA.

Marbach,D. *et al*. (2012) Wisdom of crowds for robust gene network inference. *Nat. Methods*, **9**, 796–804.

Margolin,A.A. *et al*. (2006) Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7**, S7.

Mohammadi,P. *et al*. (2013) 24 hours in the life of HIV-1 in a t cell line. *PLoS Pathogens*, **9**, e1003161.

Mortazavi,A. *et al*. (2008) Mapping and quantifying mammalian transcriptomes by rna-seq. *Nat. Methods*, **5**, 621–628.

Navratil,V. *et al*. (2009) Virhostnet: a knowledge base for the management and the analysis of proteome-wide virus–host interaction networks. *Nucleic Acids Res*., **37**, suppl 1 D661–D668.

Patil,A. *et al*. (2013) Linking transcriptional changes over time in stimulated dendritic cells to identify gene networks activated during the innate immune response. *PLoS Comput. Biol*., **9**, e1003323.

Prasad,T.K. *et al*. (2009) Human protein reference database 2009 update. *Nucleic Acids Res*., **37**, D767–D772.

Rapaport,F. *et al*. (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol*., **14**, R95.

Salgado,M. *et al*. (2011) Characterization of host genetic expression patterns in HIV-infected individuals with divergent disease progression. *Virology*, **411**, 103–112.

Schröfelbauer,B. *et al*. (2005) Human immunodeficiency virus type 1 Vpr induces the degradation of the UNG and SMUG uracil-DNA glycosylases. *J. Virol*., **79**, 10978–10987.

Schulz,M.H. *et al*. (2012) Drem 2.0: improved reconstruction of dynamic regulatory networks from time-series expression data. *BMC Syst. Biol*., **6**, 104.

Shytaj,I.L. and Savarino,A.A. (2013) cure for aids: a matter of timing. *Retrovirology*, **10**, 145.

Smith,J.A. *et al*. (2008) Evidence that the Nijmegen breakage syndrome protein, an early sensor of double-strand DNA breaks (dsb), is involved in HIV-1 post-integration repair by recruiting the ataxia telangiectasia-mutated kinase in a process similar to, but distinct from, cellular DSB repair. *Virol. J*., **5**, 11.

Stark,C. *et al*. (2006) Biogrid: a general repository for interaction datasets. *Nucleic Acids Res*., **34**, suppl 1 D535–D539.

Strasner,A.B. *et al*. (2008) The SRC kinase LCK facilitates assembly of HIV-1 at the plasma membrane. *J. Immunol*., **181**, 3706–3713.

Takada,N. *et al*. (2002) Rela-associated inhibitor blocks transcription of human immunodeficiency virus type 1 by inhibiting nf-κb and sp1 actions. *J. Virol*., **76**, 8019–8030.

Tareq Hassan Khan,M. *et al*. (2012) Structure-based analysis of the molecular recognitions between HIV-1 tar-RNA and transcription factor nuclear factor-kappab (nfkb). *Curr. Top. Med. Chem*., **12**, 814–827.

Toepfer,S. *et al*. (2007) The netgenerator algorithm: reconstruction of gene regulatory networks. In: *Knowledge Discovery and Emergent Complexity in Bioinformatics*. Springer, Berlin, pp. 119–130.

Vergoulis,T. *et al*. (2012) Tarbase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Res*., **40**, D222–D229.

Wang,J. *et al*. (2010) Transmir: a transcription factor–microRNA regulation database. *Nucleic Acids Res*., **38**, suppl 1 D119–D122.

Williams,S.A. *et al*. (2007) Sustained induction of nf-κb is required for efficient expression of latent human immunodeficiency virus type 1. *J. Virol*., **81**, 6043–6056.

Wires,E.S. *et al*. (2012) Methamphetamine activates nuclear factor kappa-light-chain-enhancer of activated b cells (nf-κb) and induces human immunodeficiency virus (hiv) transcription in human microglial cells. *J. Neurovirol*., **18**, 400–410.

Yeang,C.H. *et al*. (2004) Physical network models. *J. Comput. Biol*., **11**, 243–262.

Zamborlini,A. *et al*. (2011) Impairment of human immunodeficiency virus type-1 integrase SUMOylation correlates with an early replication defect. *J. Biol. Chem*., **286**, 21013–21022.