Technical Report

# From experimental setup to bioinformatics: An RNAi screening platform to identify host factors involved in HIV-1 replication

Kathleen Börner[1]*, *Johannes Hermle*[1]*, *Christoph Sommer*[2], *Nigel P. Brown*[3], *Bettina Knapp*[4], *Bärbel Glass*[1], *Julian Kunkel*[5], *Gloria Torralba*[6], *Jürgen Reymann*[7], *Nina Beil*[7], *Jürgen Beneke*[7], *Rainer Pepperkok*[8], *Reinhard Schneider*[3], *Thomas Ludwig*[5], *Michael Hausmann*[6], *Fred Hamprecht*[2], *Holger Erfle*[7], *Lars Kaderali*[4], *Hans-Georg Kräusslich*[1] *and Maik J. Lehmann*[1]

[1] Department of Infectious Diseases, Virology, University of Heidelberg, Heidelberg, Germany [2] Heidelberg Collaboratory for Image Processing (HCI), University of Heidelberg, Heidelberg, Germany [3] Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany [4] Viroquant Research Group Modelling, BioQuant Centre, University of Heidelberg, Heidelberg, Germany [5] Parallel and Distributed Systems, Institute for Informatics, University of Heidelberg, Heidelberg, Germany [6] Kirchhoff Institute of Physics, University of Heidelberg, Heidelberg, Germany [7] Viroquant-CellNetworks RNAi Screening Facility, BioQuant Centre, University of Heidelberg, Heidelberg, Germany [8] Cell Biology and Cell Biophysics Unit, European Molecular Biology Laboratory, Heidelberg, Germany

RNA interference (RNAi) has emerged as a powerful technique for studying loss-of-function phe-Notypes by specific down-regulation of gene expression, allowing the investigation of virus-host interactions by large-scale high-throughput RNAi screens. Here we present a robust and sensitive small interfering RNA screening platform consisting of an experimental setup, single-cell image and statistical analysis as well as bioinformatics. The workflow has been established to elucidate host gene functions exploited by viruses, monitoring both suppression and enhancement of viral replication simultaneously by fluorescence microscopy. The platform comprises a two-stage pro cedure in which potential host factors are first identified in a primary screen and afterwards re-test ed in a validation screen to confirm true positive hits. Subsequent bioinformatics allows the iden tification of cellular genes participating in metabolic pathways and cellular networks utilised by viruses for efficient infection. Our workflow has been used to investigate host factor usage by the human immunodeficiency virus-1 (HIV-1), but can also be adapted to other viruses. Importantly, we expect that the description of the platform will guide further screening approaches for virus host interactions. The ViroQuant-CellNetworks RNAi Screening core facility is an integral part of the recently founded BioQuant centre for systems biology at the University of Heidelberg and will provide service to external users in the near future.

**Keywords:** High-throughput screening · HIV · RNA interference · Small interfering RNA

**Correspondence**: Dr. Maik J. Lehmann, Department of Infectious Diseases, Virology, University of Heidelberg, Im Neuenheimer Feld 324, 69120 Heidelberg, Germany **E-mail**: maik.lehmann@med.uni-heidelberg.de **Fax**: +49-6221-565003 * These authors contributed equally to this work.

**Abbreviations: GFP**, green fluorescent protein; **HIV**, human immunode-ficiency virus; **RNAi**, RNA interference; **siRNA**, small interfering RNA

## 1 Introduction

Despite considerable advances in virological research over the last few decades, viruses continue to represent a major health risk, being responsible for millions of deaths worldwide each year. Like other viruses, HIV-1 has evolved the capability to successfully infect – and efficiently transmit between – human cells by recruiting various host proteins for each step of its life cycle [1–3]. Unravelling these critical cellular factors will not only improve our fundamental understanding of HIV-host interactions, but may eventually lead to novel anti-HIV therapeutics. Since the rate of mutations of cellular genes is substantially lower than for viral genomes, the particular benefit of targeting host factors is that it may provide a higher barrier to the generation of anti-drug resistance. A most powerful and versatile approach to identify such potential cellular interaction partners of HIV-1 are RNAi-based loss-of-function screens, as suggested by very recent reports [4–6].

Although each of the three high-throughput screens published thus far reported a large number of potential host cell factors, there is only little overlap between the different sets [7–9].This might be explained by differences in the individual experimental conditions, such as the use of distinct cell lines, siRNA libraries or virus strains, all of which could have significantly affected the results. Importantly, however, it may also be due to the use of different criteria for defining a "hit" or inconsistencies concerning the techniques applied to validate potential hits [9]. This highlights the need for comparable experimental conditions in further studies and for the selection of consistent analytical methods for future screening approaches.

In this report we describe a sensitive, automated microscopy-based siRNA screening platform, which has been designed to elucidate host factors utilised by a variety of viruses. This platform has been used for studying host cell interactions of infectious HIV-1. The detailed description of the experimental setup as well as guidance on subsequent image analysis, statistical methods and bioinformatics approaches provides essential information for establishing further screening platforms.

In our platform several sub-genomic siRNA libraries are tested in a primary screen and the identified potential hits are subsequently re-confirmed in a validation screen using different siRNAs. In both screening stages a non-silencing siRNA and an siRNA targeting the HIV-1-specific cell surface receptor CD4 are used as a negative and positive control, respectively. In addition, the knockdown efficacy and cytotoxicity of siRNAs are determined within the validation screen. Genes showing similar effects in both the primary and the validation screen are considered as validated. Bioinformatics and modelling approaches on the validated hits in combination with published HIV-1 host factors, known metabolic pathways and protein-protein interactions enable the identification of cellular networks and pathways involved in the replication of HIV-1. Further studies using this platform will characterise fundamental cellular functions of the identified hits and will shed light on their role in viral pathogenesis.

## 2 Experimental setup

The key elements of the experimental setup are the reverse transfection of commercially available siRNA libraries into HIV-permissive HeLa P4 cells (Fig. 1A). Following a 36 h incubation period to allow target knockdown, the cells are infected with HIV-1 virions encoding for GFP (Fig. 1B). This allows the straight-forward detection and quantitation of infected cells via a highly sensitive automated microscopy-based assay (Fig. 1C).

### 2.1 Preparation of the siRNA arrays

The platform was initially established for chambered coverglasses (LabTeks,NUNC,Thermo Fisher Scientific, Langenselbold, Germany), allowing the investigation of up to 384 individual spotted siRNAs per LabTek. LabTeks are characterised by easy cell and liquid handling. As only a few cells can be monitored per spot, we considered at least eight replicates necessary for a statistically reliable analysis. Later experiments showed that 384-well plates (BD FALCON 353962, BD Biosciences, Heidelberg, Germany) were more suitable for our high-throughput screening approach, as more images per well could be collected compared to LabTeks, resulting in a higher number of cells for statistical analysis. Therefore, 384-well plates have the advantage of providing stronger statistical power even with only two or three replicates, which is supported by the high correlation as shown in Fig. 2A. However, more sophisticated automated liquid handling devices are needed (for more information, see Section 2.3).

For validation screens, 96-well plates (Corning COSTAR 3603, Corning Life Sciences, Amsterdam, The Netherlands) were used, allowing collection of even more images per well compared to the 384 wells. All experiments were performed with HeLa P4 cells as they are well suited for culturing in all

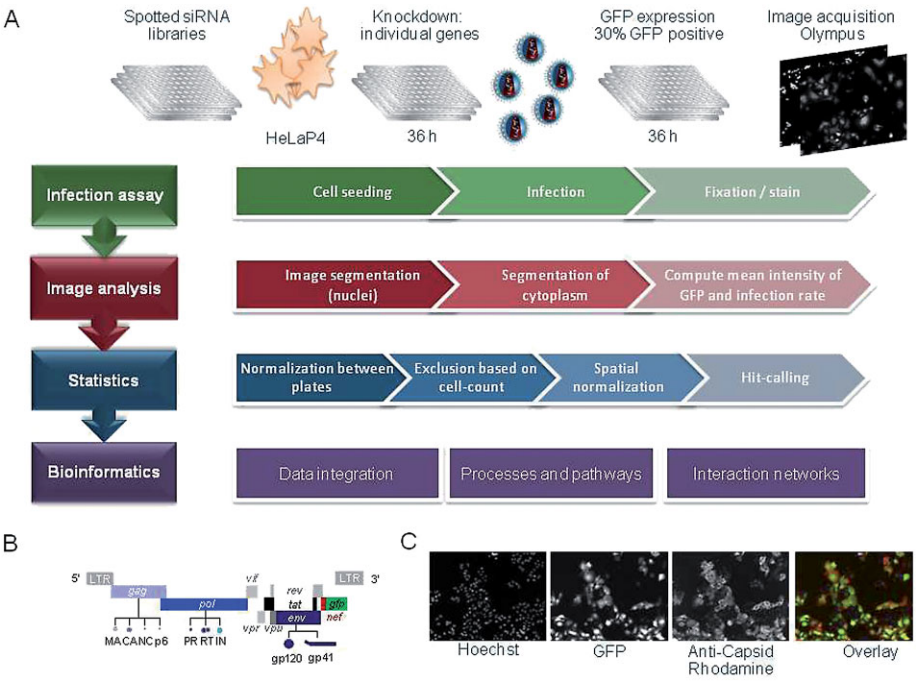


**Figure 1.** Overview of the siRNA screening platform. (A) Automated workflow consisting of the four main stages: infection assay, image analysis, statistics and bioinformatics (boxes at left, running vertically downwards). Key sub-activities of each stage are indicated (boxes to right). (B) Genome of the GFP-encoding infectious HIV-1 particles. A portion of the viral *nef* gene (red) is replaced by GFP (green) [13]. (C) Antibody staining against the viral capsid protein reveals GFP expression as a suitable readout for infection in HeLa P4 cells. From left to right: (i) Hoechst 33258 stain to identify individual cell nuclei. (ii) GFP expression after infection with GFP-encoding HIV-1 particles. (iii) Antibody stain against the viral capsid protein p24. (iv) Overlay of GFP-expression and anti-capsid stains with almost complete co-localisation, indicating that almost all GFP-expressing cells contain the viral capsid protein p24.

of the tested well plates and also highly transfectable with siRNAs (data not shown).

To keep replicates as reproducible as possible, the siRNA libraries were printed in a batch-wise manner using a previously described reverse transfection protocol [10,11].To this end,a mixture of the respective siRNA, the transfection reagent Lipofectamine 2000 (Invitrogen GmbH, Karlsruhe, Germany), fibronectin (Sigma-Aldrich Chemie GmbH, Taufkirchen, Germany), sucrose (USB Corporation, Ohio, USA) and gelatine (Sigma-Aldrich, Steinheim, Germany) were transferred automatically to a 384-well plate or any other cell culture format. After drying the well, the substrates can either be stored for up to 15 months without any loss of efficacy or can be directly used for knockdown studies [10]. This is a major advantage of the reverse transfection method compared to liquid transfection, as it permits better reproducibility and comparability between different plates of the same batch.

For the primary screen, various libraries containing *silencer* siRNAs (Ambion, Applied Biosystems,Austin,TX, USA) were used in 384-well plates or LabTeks, with one individual siRNA per well/spot and three distinct siRNAs per target gene. In contrast to individual siRNAs, "pooled siRNAs" are considered to achieve a more potent knockdown by the synergistic effects of combining several siRNAs targeting the same gene. However, less active siRNAs may interfere with the efficacy of highly active siRNAs and attenuate their effect.

Potential host factors identified in the primary screen as "hits" (for more information see Section 4) were subsequently re-tested within a validation screen with a minimum of five replicates in 96-well plates. Per gene, at least two novel chemically modified siRNAs (*silencer select* siRNAs, Ambion) were used to enhance specificity and minimise "off-target effects". Two siRNAs against CD4 were chosen as positive controls: for the primary screen: 5'-GAUCAAGAGACUCCUCAGUTT-3' [12] and for the validation screen: Ambion *silencer select* "s2579". The non-silencing controls were: Primary screen: 5'-AGGUAGUGUAAUCGCCUUGTT3' and for the validation screen:Ambion *silencer select* "negative control #1".

To determine a suitable time for efficient siRNA-mediated gene knockdown, the surface expression levels of the HIV-1 specific receptor CD4 were determined by FACS analysis after varying siRNA treatment times. CD4 mediates cell entry of HIV-1 and thus represents a pivotal host factor for HIV-1 infection. Our study revealed a gene knock

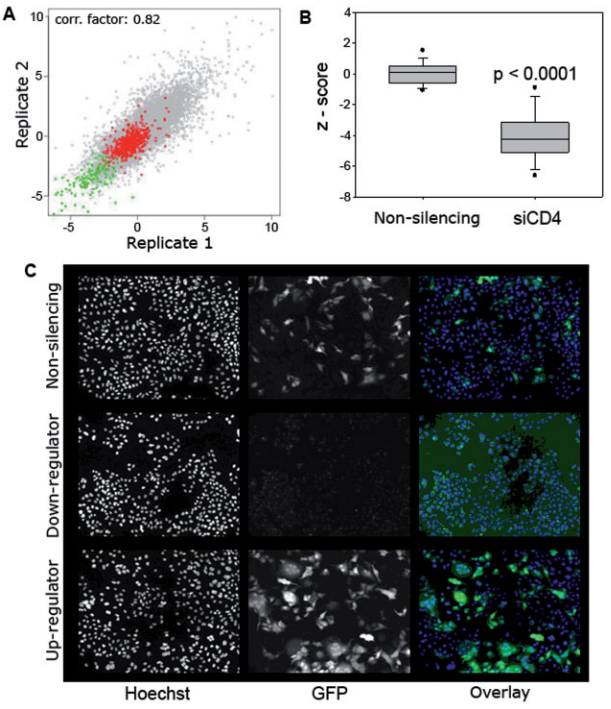


**Figure 2.** Enhancing or inhibiting effects of individual gene knockdowns on HIV-1-infection imaged by fluorescence microscopy. HeLa P4 cells were transfected with siRNAs and infected with HIV-1-AGFP particles after 36 h. After incubation for a further 36 h, the cells were fixed and GFP and Hoechst images acquired by high-throughput fluorescence microscopy. (A) Correlation between z–scores of two replicates of an RNAi screen investigating host cell functions in HIV-1 pathogenesis with individual siRNA-mediated gene knockdowns (grey), non-silencing control siRNAs (red) and CD4 positive controls (green). The Pearson correlation coefficient is 0.82. (B) Z-score distributions of the non-silencing control siRNA and the positive control CD4. The separation of the two distributions confirms CD4 as a significant down-regulator in our assay. (C) Hoechst stain, GFP-expression and the overlay for the non-silencing control siRNA and for representatives of down- and up-regulating gene function in HIV-1 replication.

down efficiency of approximately 90% within a period of 24–92 h after siRNA transfection (data not shown). Thus, in subsequent experiments cells were infected 36 h after transfection and fixed after an additional 36 h.

## 2.2 Virus constructs, culture and harvesting

In our screens, fully infectious HIV-1 particles encoding the green fluorescence protein (GFP) were used. The construct was derived from wild-type HIV-1 (pNL4-3), in which the *gfp* open reading frame was C-terminally fused to the first 16 amino acids of the *nef* gene (referred to as HIV-1-AGFP, Fig. 1B, [13]).

The GFP expression after infection allows a direct measurement of viral replication by fluorescence microscopy, thereby avoiding further timeand cost-intensive working steps like antibody staining or substrate-based indicator assays. Standard protocols for the production of HIV-1-AGFP yielded fully functional particles as described previously [13]. However, we found that particles lost the ability to induce cellular GFP expression over a period of several weeks, most likely due to high recombination rates. Thus, a commonly used procedure for large-scale virus production was not applicable and the protocol had to be adapted. The most important factor was to limit cell culture to ten passages. Cell culture and infection conditions were optimised to generate high titres of HIV-1-AGFP as follows. Human embryonic kidney 293T cells were transfected with HIV-1-AGFP proviral DNA. At 42 h after transfection, the virus containing supernatant was harvested and used for an initial infection of human C8166 cells. After one passage the virus was used to start a co-culture with MT4 cells for five passages in total. The virus-containing cell culture supernatant was enriched up to 100-fold using the crossflow filtration system VivaFlow 200 (Vivascience AG, Hannover, Germany) and a subsequent ultracentrifugation step.

A typical production round yielded 20 mL of virus preparation with an average concentration of 66 µg/mL viral capsid protein (p24), determined by in-house ELISA. Infectivity and functional integrity of produced HIV-1-AGFP particles were confirmed prior to high-throughput screening. Infectivity of the viral particles was verified by measurement of luciferase activity in TZM cells (data not shown). As the assay depends on a GFP reporter, it is crucial that GFP and viral protein expression are correlated. Accordingly, HeLa P4 cells were infected and stained for p24 with fluorescently tagged antibodies and subsequently monitored for co-localisation with GFP by fluorescence microscopy (Fig. 1C).

## 2.3 Infection assay

For siRNA transfection and subsequent infection HeLa P4 cells were seeded into siRNA-pre-coated 384-well plates (850 cells/well in 30 µL) or 96-well plates (2500 cells/well in 100 µL) using the reagent dispenser MultiDrop Combi (Thermo Fisher Scientific) or manually into LabTeks ($1.2 \times 10^5$ cells/ LabTek in 3 mL). Following a 36 h incubation for target gene knockdown, the cells are infected with HIV-1-AGFP in a BSL-3 facility, using the compact automated liquid handler Hydra DT (Matrix,Thermo Fisher Scientific) for well plates.As we were interested in detecting potential decreases as well as increases in GFP expression to identify down- and up-regulating genes in HIV-1 replication, we ad-justed the infection rate to approximately 30%

(384-well plate: 28 ng p24/well in 30 µL; 96-well plate: 80 ng p24/well in 50 µL; LabTek: 6.7 µg p24/LabTek in 2 mL). After an additional 36 h, cells are fixed for 90 min with 4% paraformaldehyde, removed from the BSL-3 facility and stained for 45 min with the dye Hoechst 33258 (330 ng/mL) for cell nuclei detection as further described in Section 3.

### 2.4 Image acquisition

For high-throughput image acquisition a fully automated epifluorescence Scan$^{R}$ screening microscope equipped with the Scan$^{R}$ acquisition software (Olympus Biosystems GmbH, Münster, Germany) was used. Images were acquired with a 10x objective in 9 positions per well for 384-well plates, up to 16 positions for 96-well plates and in 1 position per spot on a LabTek. In each position images were acquired in the Hoechst and in the GFP channel using the corresponding excitation and emission filters.

### 2.5 Cytotoxicity and non-specific effects

Cytotoxic siRNAs in the primary screen were filtered out,as described later in Section 4.All validation-screen siRNAs were tested for potential confounding effects by the following methods. Cytotoxicity was assessed by the Toxilight Assay Kit (Lonza, Sales Ltd., Basel, Switzerland). The general influence of the siRNAs on cellular protein expression was determined using a stably GFP-expressing HeLa P4 cell line seeded onto siRNAcoated cell culture plates under screening conditions, but without viral infection. As variations in GFP intensities reveal a virus-independent siRNAinduced effect on the cellular expression machinery, siRNAs with significant effects on GFP expression were excluded from further investigations. Finally, the knockdown efficacy of the siRNAs used in the validation screen was tested by qRT-PCR (SYBR Green, Applied Biosystems, Darmstadt, Germany).

### 2.6 Concluding remarks

The specific viral construct (HIV-1-AGFP) was harvested and produced in sufficient amounts for high-throughput screening. Co-localisation of p24 and GFP demonstrated that virus-induced GFP expression was a suitable readout for infection. The siRNA against CD4 and a non-silencing siRNA were established as working positive and negative controls and could be clearly distinguished (Fig. 2B). The platform was shown to work on all tested cell culture formats. In summary, the screening assay is able to reveal both down- and up-regulating host genes that are modulating HIV-1 replication by altering the GFP-intensity and infection ratio (Fig. 2C).

## 3 Image analysis

Image analysis consisted of the following three successive steps: (i) Cell nuclei segmentation in the Hoechst channel, (ii) cell segmentation in the GFP channel, and (iii) quantification of cell characteristics. The inputs of the image analysis routine were the Hoechst and GFP channels, representing cell nuclei and cell cytoplasm (Fig. 3A). As output we computed the number of nuclei, the average signal intensity over all cells (in the GFP channel), and the proportion of cells that were infected. These calculations were performed per position in a well and served as input to the statistical analysis.

### 3.1 Cell nuclei segmentation

To segment cell nuclei we used the ‘marker-controlled watershed transform’ method [14]. This method consisted of two basic steps. First, the algorithm detected a unique initial ‘foreground marker’ for each cell nucleus (typically the marker that corresponds to the centre of the object). At the same time ‘background markers’ that lie in the dark region between cell nuclei were extracted (Fig. 3B). Second, the watershed transform expanded markers spatially to enclose the cell nuclei in the image. Finally, basins originating from foreground markers corresponded to masks of cell nuclei.

### 3.2 Cell nuclei marker extraction

A preliminary binarisation of the Hoechst image was found using a histogram adaptive threshold [15], leading to an initial nuclei *versus* background mask. In the case of an uneven illumination of the image, background correction could be applied before binarisation [16]. A Gaussian blurred version of the cell nuclei image was used to detect foreground markers (the variance for the Gaussian kernel is empirically adapted to cell nuclei size).All local maxima above the binarisation threshold were determined and dilated to suppress spurious markers (Fig. 3B). To extract background markers we computed the Euclidian distance map from the initial binary mask.This map contains the distance from each non-nucleic pixel to its nearest nucleic pixel in the binarised image. All ridges from the distance map were extracted to get a Voronoi-net

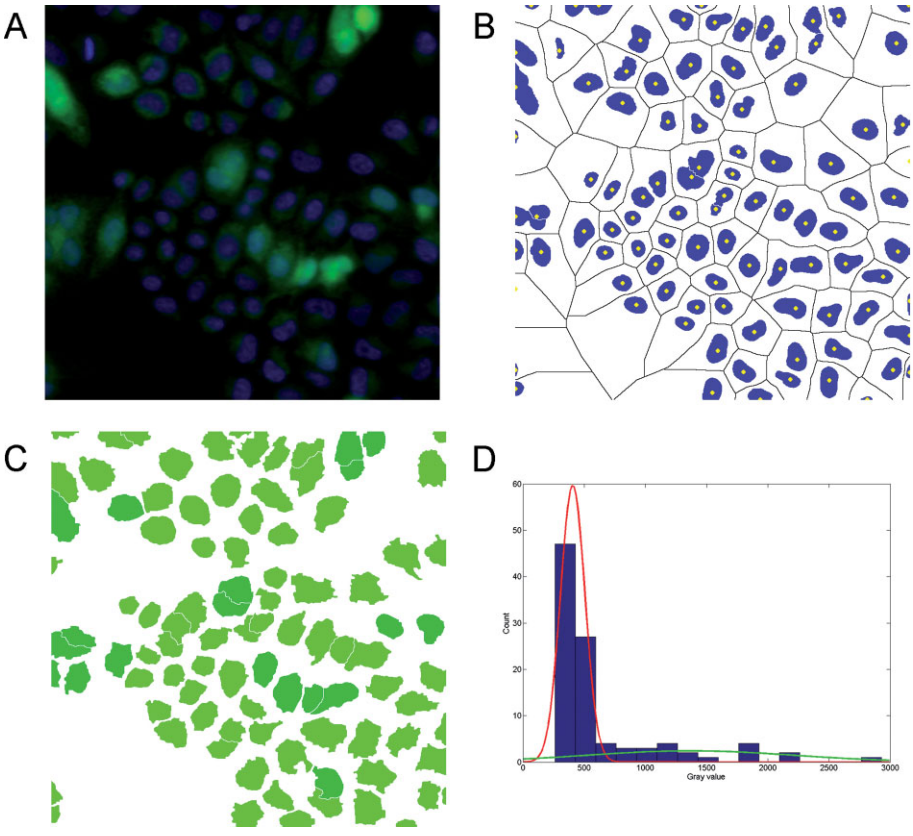


**Figure 3.** Image segmentation and infection cut-off. (A) Epifluorescence microscopy image of HIV-1-AGFP-infected GFP-expressing cells with cell nuclei (blue) and GFP-expressing cells (green). (B) Segmentation of cell nuclei (blue spots). (C) Segmentation of cells showing individual infected cell bodies (green) (D) Histogram of mean grey values with fitted curves used to define infection cut-off criterion.

like background marker (Fig. 3B). Both types of markers (background from distance map, foreground from Gaussian blur) served as the final markers for the watershed transform operating on the gradient image.

### 3.3 Cell segmentation

The extracted cell nuclei masks were passed as foreground markers to the cell segmentation process. As background markers we re-used the background markers from the cell nuclei segmentation (Fig. 3C). Note that background markers cannot be computed directly from the GFP channel, as only infected cells are visible in this channel.

### 3.4 Quantification

The main readouts of the image analysis were the average signal intensity over all cells in the image, the number of nuclei, and the infection ratio. Uninfected cells appeared dark, while infected cells showed rather high grey values over their area. A Gaussian mixture model with two components was fitted to the extracted mean grey values (Fig. 3D). One Gaussian component explains the variation of uninfected, while the other accounts for the variation of infected cells. If zero or all cells are infected, one of the two components should vanish; however, in practice, an intermediate number of infected cells produces two overlapping components. To remedy this, a prior is imposed on the position of the two Gaussian means, which essentially acts as a repelling force between the mixture components. Having found the parameters of the mixture model, an optimal threshold is computed to classify each cell as infected or not infected. Finally, the infection ratio is computed as the number of infected cells divided by the total cell count.

### 3.5 Implementation and pipeline

A fully automated pipeline was set up to process the large number of images produced by the screens. Image processing was implemented in Matlab and C++ and runs in a data-parallel fashion distributed over a 12 × 8-core Linux (64bit) compute cluster. For each well position, output is writ

ten to a virtual file system which automatically maps data into a relational database for convenient post-processing.

### 3.6 Concluding remarks

The use of parallel batch processing yields a major speed-up as a typical screen with 1.3 million images requiring 3.6 TB of storage is processed in approximately 36 h. Automated image quality control is being added to exclude foreign bodies and artefacts in images before quantification, resulting in improved measurements. A long-term goal is to create a framework for interactive statistical learning of segmentation tasks, allowing simple adaptation to novel cell types, sensor settings, chemical dyes and assay-specific phenotypes, without reprogramming the image processing routines.

## 4 Statistical analysis

Segmented images were processed further statistically, to identify up- and down-regulators of viral replication. Based on the number of nuclei per well and the average signal intensity over all cells in the GFP channel, our statistical analysis pipeline consisted of the following steps: (i) Log-transformation of raw data, (ii) normalisation between different plates, (iii) identification and removal of siRNAs with cytotoxic effects, (iv) removal of systematic effects of cell counts on signal intensity, (v) spatial normalisation within one plate, and (vi) hit calling. The entire workflow was complemented by strict quality control, monitoring at each step the quality of raw and transformed data. In case of failed quality control for individual wells or plates, these were either removed from further analysis or repeated.

### 4.1 Data normalisation

As a first step for the analysis of the data, a logtransformation of the raw data is carried out. Since the distribution of raw fluorescence intensities is heavily right-skewed, a logarithmic transformation of the data results in a more normal distribution.

In the primary screen, different plates are made comparable by subtracting the plate median from each intensity value, and dividing by the median absolute deviation (z–score). These two measures are more robust alternatives to the mean and the standard deviation. The normalisation steps are based on the assumptions that the siRNAs are randomly distributed over the plates, and that most siRNAs do not have an effect on viral replication. Clearly, for validation screens, this assumption is not valid. In this case, plates are made comparable by subtracting the median of the negative controls, and dividing by their median absolute deviation.

Since some knockdowns are cytotoxic and therefore interfere with the readout of viral replication, we excluded wells with the lowest 5% of cell counts in the entire screen. Furthermore, due to possible incorrect segmentation of images with very dense cell populations, wells with the highest 5% of cell counts were also removed. Locally weighted scatterplot smoothing [17] was used to de-correlate signal intensities and cell count, by adjusting the signal intensities for the effect of unequal cell numbers in wells.This was done for each plate individually, since effects may be different from plate to plate. Spatial effects within each plate were then corrected using B-score normalisation [18]. This method adjusts intensities using a median polish procedure on rows and columns separately, thus estimating row- and column-effects on each plate. These estimates were then used to correct each spot individually. The procedure is very effective at removing spatial artefacts such as edge effects on a plate, but rests on the assumptions that siRNAs are randomly placed on the plates, and that most siRNAs do not have an effect on viral replica-tion.This method must therefore not be used in the analysis of secondary/validation screens.

### 4.2 Statistical testing

The current standard practice to select hit genes from RNAi data is to select siRNA hits whose z-score-normalised intensity deviates from the bulk [19], but hits with smaller intensities will be missed using this method [20]. Thus, if enough replicates are available, a statistical approach is used that assigns a $p$ value to each siRNA. If the $p$ value is smaller than a given significance level $\alpha$, the null hypothesis $H_0$ that assumes no significant effect can be rejected. Since data in this study were more or less normally distributed, we used the one-sample two-sided Welch's $t$-test to compute $p$ values. If siRNAs are randomly distributed on a plate and if it can be assumed that most siRNAs have no effect, replicates in the test can be compared with the overall population, acting as a *de-facto* negative control. If this assumption is not valid, *e.g.* in a validation screen, the test is carried out against negative controls. For primary screens, where only two replicates are available and a statistical test is thus not feasible, hits are called if their mean z-score over replicates is $>2$ (or $< -2$). For validation screens, we combine the information given by z-score and $p$ value by requiring a $p$ value $<0.05$ and a z-score $>1.5$ (or $< -1.5$), respectively. The $p$ value

ensures that effects are repeatable over the replicates, while the z-score assures that they are sufficiently strong on average.

### 4.3 Implementation

The statistical analysis described here was implemented in the free statistical environment R (www.R-project.org), and used the cellHTS [19] and RNAither [21] Bioconductor packages (www.bioconductor.org).The pipeline was integrated into the full-screening workflow, and generated user-readable HTML webpages for the assessment of data quality and normalisation results. Z-scores and *p* values were also written to text files for further processing (see Section 5).

### 4.4 Concluding remarks

Statistical analysis is important to robustly identify "hit" genes and avoid errors. Data normalisation should be based on controls.This procedure is only valid if sufficient numbers of positive and negative controls have been spotted per plate. This can normally only be achieved in validation screens with a custom-designed layout. Normalisation on plate medians is more robust, but requires random spotting of siRNAs on plates and is feasible only for primary screens. Considerable spatial effects are observed in cell arrays, but to a lower extent also on well plates, and must be corrected during the normalisation. As the correlation between cell numbers and signal intensity is often non-linear, sophisticated methods such as Lowess normalisation have to be used. With all these considerations the raw data produced by the image analysis is analysed and the outcome is represented by z-scores.

## 5 Bioinformatics

High-throughput screening of many hundreds or thousands of genes requires integration of correspondingly large amounts of data, both from the screens themselves and from external data sources. External data may comprise information about the genes and gene products including crossreferences to other data sources, as well as relationships between the genes, such as participation in metabolic pathways, protein-protein interaction and gene regulatory networks.

The principal aims of bioinformatics are (i) to integrate data about biological entities of interest into descriptive or qualitative models, and (ii) to use this as a basis for comparisons and predictions to guide further experimentation.

### 5.1 Data integration

The suppliers for our siRNA libraries designate target genes using NCBI Entrez GeneID [22] and RefSeq [23] identifiers. Thus the range of possible genes and their associated identifiers in our human screens is NCBI-centric, rather than being based on the overlapping EMBL-EBI/Sanger Ensembl

[24] collection. Entrez and RefSeq are continually being updated by NCBI as gene and protein product entries are refined, old records are withdrawn and new records are added, so that the identifiers and data for the gene targets referenced in the libraries, and in published studies, are subject to change. Symbolic gene names are particularly unreliable, so that, when reporting the results of a study, the Entrez and/or RefSeq identifiers as well as the database release from which they were obtained should be reported.The bioinformatics processing stage must track changes in gene accession numbers/identifiers and maintain cross-references to other external data sources, also having potentially changing identifiers.

We use a relational database to store human protein-encoding gene identifiers obtained primarily from NCBI Entrez and cross-referenced to the Human Genome Organisation Nomenclature Committee (HGNC [25]) and NCBI RefSeq collections. For our purposes, we group related identifiers into a functional 'locus' referring to a set of related gene or protein product identifiers, which are considered to represent the same protein or set of transcripts. Each 'locus' is assigned a unique, stable internal numeric identifier in the database so that all identifiers associated with a given locus, even completely different historical Entrez GeneID, HGNC and RefSeq identifiers, can be united and crossreferenced to older (or newer) identifiers found in other studies. Additional information mined from Entrez GeneID, HGNC and RefSeq collections are associated with each locus. These include current and obsolete gene symbols and synonyms, official gene product names, chromosome and map locations, UniProt [26] protein sequence identifiers, enzyme classification (EC) numbers, PubMed literature identifiers and OMIM (http://www.ncbi.nlm. nih.gov/omim) identifiers for disease associations. This forms the core of the database around which other types of data can be assembled. As new data are added to the database, they can be cross-referenced to the appropriate locus (for a single gene or gene product) or multiple loci (for relationships such as protein-protein interactions).

## 5.2 Processes and pathways

A basic problem facing the researcher is to understand the contents of the screening library, since individual gene product names and descriptions are often too terse or abstruse. Systematised collections of general gene product descriptions exist in the form of ontologies, the best known being the Gene Ontology project (GO [27]). GO associates curated terms with genes according to three ontologies: molecular function, biological process, and cellular component. The terms form three different networks (actually directed graphs), which are stratified from the general to the specific, but allow multiple paths rather than enforcing a strict hierarchy. In our workflow, hit genes are automatically mapped to GO annotations using RNAither [21], for all three ontologies. Geneset enrichment analysis is then used to identify categories and pathways containing more hits than would randomly be expected, using topGO [28], to find molecular functions, biological processes and cellular compartments that may be particularly important for viral replication.

Another ontology collection that is specialised for human, mouse, rat and the fruit fly is the PANTHER Classification System [29], which is intended for use in high-throughput analysis and is simplified accordingly. PANTHER biological processes cover almost half (about 11 300 genes) of the human protein coding loci, although more coverage is obtainable by looking at molecular function without placement in any particular process. The designers of PANTHER have defined a set of about 30 top-level biological processes dealing with very general (overlapping) cellular activities such as 'amino acid metabolism' or 'cell adhesion', and we incorporate these into our core database to give a simplified overview, both of the genes in the screening libraries and of the hits.

A second problem, having performed a screen and identified known pathways of interest, is to visualise the hits in these pathways. High-quality pathway information is obtained from the Kyoto Encyclopaedia of Genes and Genomes (KEGG, [27]) in the form of navigable schematic diagrams of metabolic and regulatory pathways, based on extensive literature mining across multiple organisms.The maps are interactive and KEGG provides a simple API (Application Programming Interface) for several common programming languages allowing selective manipulation of the diagrams from locally written client applications communicating with the KEGG server. Note, however, that the KEGG pathway maps cover considerably less of the human genome than PANTHER processes, accounting for slightly less than 4800 unique genes.

Clearly, no single process or pathway data source is sufficient to address every need, and several need to be combined in an analysis.

## 5.3 Interaction networks

For many genes, the pathway data are insufficient to identify any relationship between the hits from a screen. We then turn to protein-protein interaction networks and try to embed hits and search for clusters.

A number of protein-protein interaction databases are publicly available, most of them containing interaction data for multiple organisms. For a review of these databases focusing on human-specific interactions, see [30]. These datasets vary according to the quality of the data source and curation that they offer. Interactions may be predicted computationally or obtained experimentally. Experimental observations may derive from highthroughput analyses such as yeast two-hybrid, or from co-immunoprecipitation assays, with varying degrees of confidence in the result. Similarly, the level of curation ranges from automatic literature mining, through manual literature mining by biologists, to direct submission of results by the investigator. One of the most comprehensive collections of human-only, experimental interaction data is the Human Protein Reference Database (HPRD, [31], http://www.hprd.org), which is manually curated and includes details of the type of experiment, the protein domains involved and post-translational modifications.

In general, the reliability of an interaction increases as more corroboratory evidence is found. The STRING [32] database provides a merged multi-organism collection assembled from the public interaction databases, together with other predictive information such as pathways and expression profiles that are common across organisms. Each interaction is assigned an edge weighting to indicate the degree of confidence in that interaction. We currently use the STRING interactive web interface (http://string.embl.de) to visualise small networks as graphs of nodes (proteins) and edges (interactions). Larger networks, such as the HPRD dataset, are visualised using Cytoscape [33], which is a cross-platform Java program (http://www. cytoscape.org).The graph layout can be manipulated directly and the software offers many different automatic layout schemes.

There are broadly two strategies to identifying interesting genes in these networks for further study: First, given the known interactions of the pathogen with the host proteins, candidate hits that lie within one or a few steps away from these

„front-line“ host genes on the known host protein-protein interaction network are of interest, as they deepen our understanding of known pathogen/ host dependencies into the rest of the network. One can also look for interaction neighbours of other published modulators, such as those found in other screens, to extend the boundaries of these clusters. Second, one can search for novel, statistically significant clusters of hit genes co-occurring in the known interaction network.

Much published information already exists for well-studied organisms or pathogen/host interactions such as HIV/human. Known information on HIV/host gene interactions is available from the NIAID HIV interaction dataset [34].This is also incorporated into our database, allowing us to see which HIV proteins are known to be associated with any of our screening hits, together with the nature of the interaction and the evidence for it. Similarly, the collections of host factors described in previous published HIV screens [4–6] have been mined from the papers and added to the database, allowing direct comparison of the degree of overlap and gene composition of the various studies with each other and with our screens.

### 5.4 Implementation

All bioinformatics processing is performed under Linux. Regular updates of material from NCBI and other data sources are automatically transferred and post-processed locally into suitable form for subsequent stages. The relational database uses PostgreSQL. Applications are written in Python, Perl or using UNIX scripting tools.

### 5.5 Concluding remarks

Known or suspected interaction networks around single identified genes can be mined from bioinformatics databases to provide a connected subnetwork, into which multiple hits from the screen can be embedded and correlated with phenotypes of interest. A central component of the bioinformatics strategy therefore is a database system to store and process all this data effectively, since the amount and variety of types of data that need to be managed, cross-referenced and queried is potentially huge and changing. Finally, the system is also used to help selecting interesting and functionally related candidates from the primary screen for validation screening, and for subsequent detailed experimental analysis. In the case of a virus-host interaction screen as discussed in this paper, interesting hits need to be mapped not only into host biological processes, but also, if possible, to subcellular compartments and to associated stages of the virus life cycle.

## 6 Conclusion

We have described a modular and flexible microscopy-based RNAi screening platform for investigation of host factors involved in virus-host interaction. This uses a two-stage procedure (primary and subsequent validation screen) comprising four main steps: experimental assay, image analysis, statistical analysis and bioinformatics, each of which has been presented in detail. The platform was demonstrated in stably CD4-expressing HeLa P4 cells using a modified infectious HIV-1 strain carrying a GFP reporter (HIV-1-AGFP). The procedure was shown to be suitable for robust and sensitive detection of host cell factors involved in HIV-1 replication using different cell culture formats.

Design and testing of standardised experimental setups for production of sufficient amounts of virus were found to be critical to obtaining reliable and comparable datasets. Dedicated image processing procedures were developed to process the very large amounts of high-throughput image data (tens of terabytes over the lifetime of a full genome screen) in reasonable time. Problems associated with the need for different data normalisation approaches in primary and validation screens were addressed in the statistical analysis step, which also deals with multiple potential sources of error in the readout. The bioinformatics step integrates experimental results with data mined from public databases, allowing screening hits to be functionally classified and embedded into known pathways and protein-protein interaction networks.

Future extensions will include automated classification of images based on cell morphologies, so that knockdown phenotypes can be classified in more detail and better correlated with biological processes. Comparative analyses using other viruses on the same platform will elucidate commonly used cellular host factors exploited by different viruses, which may serve as novel drug targets for ‘broad spectrum’ antivirals. Finally, we would like to encourage efforts at standardisation of RNAi screening procedures and offer our experience with this platform as a robust basis on which to build new systems.

We thank Christiane Jost and Dirk Grimm for critical reading of the manuscript and Marc Hemberger for IT support. This study was supported by the

BMBF-funded project ViroQuant (0313923). H.G.K. and L.K. are members of the excellence cluster Cell-Networks (EXC81). M.J.L. is supported by the Sonderforschungsbereich 638 (Dynamics of macromolecular complexes in biosynthetic transport).The ViroQuant-CellNetworks RNAi Screening core facility would like to acknowledge funding within the Forsys-ViroQuant consortium (0313923) as well as by the Federal Ministry of Education and Research (BMBF). This work was supported by the cluster of excellence CellNetworks at the University of Heidelberg (EXC81).

*The authors have declared no conflict of interest.*

## 7 References

[1] Carter, C. A., Ehrlich, L. S., Cell biology of HIV-1 infection of macrophages. *Annu. Rev. Microbiol.* 2008, *62*, 425–443.

[2] Malim, M. H., Emerman, M., HIV-1 accessory proteins – Ensuring viral survival in a hostile environment. *Cell Host Microbe* 2008, *3*, 388–398.

[3] Martin, N., Sattentau, Q., Cell-to-cell HIV-1 spread and its implications for immune evasion. *Curr. Opin. HIV AIDS* 2009, *4*, 143–149.

[4] Brass,A.L.,Dykxhoorn,D.M.,Benita,Y.,Yan,N. *et al.*, Identification of host proteins required for HIV infection through a functional genomic screen. *Science* 2008, *319*, 921–926.

[5] Konig, R., Zhou, Y., Elleder, D., Diamond, T. L. *et al.*, Global analysis of host-pathogen interactions that regulate earlystage HIV-1 replication. *Cell* 2008, *135*, 49–60.

[6] Zhou,H.,Xu,M.,Huang,Q.,Gates,A.T. *et al.*, Genome-scale RNAi screen for host factors required for HIV replication. *Cell Host Microbe* 2008, *4*, 495–504.

[7] Bushman, F. D., Malani, N., Fernandes, J., D'Orso, I. *et al.*, Host cell factors in HIV replication: meta-analysis of genome-wide studies. *PLoS Pathog.* 2009, *5*, e1000437.

[8] Goff, S. P., Knockdown screens to knockout HIV-1. *Cell* 2008, *135*, 417–420.

[9] Prudencio, M., Lehmann, M. J., Illuminating the host – How RNAi screens shed light on host-pathogen interactions. *Biotechnol. J.* 2009, *4*, 826–837.

[10] Erfle, H., Neumann, B., Liebel, U., Rogers, P. *et al.*, Reverse transfection on cell arrays for high content screening microscopy. *Nat. Protoc.* 2007, *2*, 392–399.

[11] Erfle, H., Neumann, B., Rogers, P., Bulkescher, J. *et al.*, Work flow for multiplexing siRNA assays by solid-phase reverse transfection in multiwell plates. *J. Biomol. Screen.* 2008, *13*, 575–580.

[12] Novina, C. D., Murray, M. F., Dykxhoorn, D. M., Beresford, P. J. *et al.*, siRNA-directed inhibition of HIV-1 infection. *Nat. Med.* 2002, *8*, 681–686.

[13] Welker,R.,Harris,M.,Cardel,B.,Krausslich,H.G.,Virion incorporation of human immunodeficiency virus type 1 Nef is mediated by a bipartite membrane-targeting signal: Analysis of its role in enhancement of viral infectivity. *J. Virol.* 1998, *72*, 8833–8840.

[14] Beucher, S., *The watershed transformation applied to image segmentation,* Scanning Microscopy International, Chicago 1992, pp. 299–314.

[15] Otsu, N., A threshold selection method from grey level histograms. *IEEE Trans. Systems Man Cybernetics* 1979, *9*, 62–66.

[16] Lindblad, J., Bengtsson, E., A comparison of methods for estimation of intensity nonuniformities in 2D and 3D microscope images of fluorescence stained cells. *Proceedings of the 12th Scandinavian Conference on Image Analysis (SCIA)* 2001, 264–271.

[17] Cleveland, W. S., Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.* 1979, 829–836.

[18] Brideau, C., Gunter, B., Pikounis, B., Liaw, A., Improved statistical methods for hit selection in high-throughput screening. *J. Biomol. Screen.* 2003, *8*, 634–647.

[19] Boutros, M., Bras, L. P., Huber, W., Analysis of cell-based RNAi screens. *Genome Biol.* 2006, *7*, R66.

[20] Malo, N., Hanley, J. A., Cerquozzi, S., Pelletier, J., Nadon, R., Statistical practice in high-throughput screening data analysis. *Nat. Biotechnol.* 2006, *24*, 167–175.

[21] Rieber, N., Knapp, B., Eils, R., Kaderali, L., RNAither, an automated pipeline for the statistical analysis of highthroughput RNAi screens. *Bioinformatics* 2009, *25*, 678–679.

[22] Maglott,D.,Ostell,J.,Pruitt,K.D.,Tatusova,T.,Entrez Gene: Gene-centered information at NCBI. *Nucleic Acids Res.* 2005, *33*, D54–58.

[23] Pruitt, K. D., Tatusova, T., Maglott, D. R., NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 2007, *35*, D61–65.

[24] Hubbard, T. J., Aken, B. L., Ayling, S., Ballester, B. *et al.*,Ensembl 2009. *Nucleic Acids Res.* 2009, *37*, D690–697.

[25] Eyre, T. A., Ducluzeau, F., Sneddon, T. P., Povey, S. *et al.*,The HUGO Gene Nomenclature Database, 2006 updates. *Nucleic Acids Res.* 2006, *34*, D319–321.

[26] The universal protein resource (UniProt). *Nucleic Acids Res.* 2008, *36*, D190–195.

[27] Kanehisa, M., Goto, S., KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000, *28*, 27–30.

[28] Alexa, A., Rahnenfuhrer, J., Lengauer, T., Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* 2006, *22*, 1600–1607.

[29] Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H. *et al.*, PANTHER:A library of protein families and subfamilies indexed by function. *Genome Res.* 2003, *13*, 2129–2141.

[30] Mathivanan, S., Periaswamy, B., Gandhi, T. K., Kandasamy, K. *et al.*, An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics* 2006, *7 Suppl 5*, S19.

[31] Mishra, G. R., Suresh, M., Kumaran, K., Kannabiran, N. *et al.*, Human protein reference database – 2006 update. *Nucleic Acids Res.* 2006, *34*, D411–414.

[32] Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S. *et al.*, STRING 8 – A global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* 2009, *37*, D412–416.

[33] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S. *et al.*,Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003, *13*, 2498–2504.

[34] Ptak,R.G.,Fu,W.,Sanders-Beer,B.E.,Dickerson,J.E. *et al.*, Cataloguing the HIV type 1 human protein interaction network. *AIDS Res. Hum. Retroviruses* 2008, *24*, 1497–1502.