F1000Research

CrossMark
click for updates

OPINION ARTICLE

REVISED Finding small molecules for the 'next Ebola' [v2; ref status: indexed, http://f1000r.es/5lb]

Sean Ekins[1,2], Christopher Southan[3], Megan Coffee[4]

[1]Collaborations in Chemistry, 5616 Hilltop Needmore Road, Fuquay-Varina, NC, 27526, USA
[2]Collaborative Drug Discovery, 1633 Bayshore Highway, Suite 342, Burlingame, CA, 94010, USA
[3]IUPHAR/BPS Guide to PHARMACOLOGY, Centre for Integrative Physiology, University of Edinburgh, Hugh Robson Building, Edinburgh, EH8 9XD, UK
[4]Center for Infectious Diseases and Emergency Readiness, University of California at Berkeley, 1918 University Ave, Berkeley, CA, 94704, USA

Abstract
The current Ebola virus epidemic may provide some suggestions of how we can better prepare for the next pathogen outbreak. We propose several cost effective steps that could be taken that would impact the discovery and use of small molecule therapeutics including: 1. text mine the literature, 2. patent assignees and/or inventors should openly declare their relevant filings, 3. reagents and assays could be commoditized, 4. using manual curation to enhance database links, 5. engage database and curation teams, 6. consider open science approaches, 7. adapt the "box" model for shareable reference compounds, and 8. involve the physician's perspective.

This article is included in the Ebola channel.

Open Peer Review

Referee Status: ☑☑

	Invited Referees	
	1	**2**
REVISED version 2 published 07 Jul 2015	☑ report	
version 1 published 27 Feb 2015	☑ report	☑ report

1 **Qiaoying Zeng**, Gansu Agricultural University China

2 **Martin Zacharias**, Technical University Munich Germany

Discuss this article

Comments (0)

Corresponding author: Sean Ekins (ekinssean@yahoo.com)

Competing interests: SE is a consultant for CDD.

Introduction

The current Ebola virus (EBOV) epidemic points to opportunities for preparing for the next pathogen outbreak or newly identified infectious disease. While control measures and therapeutic strategies have certainly been learned from past outbreaks[1], they may be insufficient to control a new one. Given the rapid evolution of viruses and the inexorable global increase in human mobility, this quote from a recent Nature editorial seems prescient "*because one thing is clear: whether it is Ebola virus, another filovirus or something completely different, there will be a next time*"[2]. For comparison we can also look at infectious diseases we do have treatments for but still need to improve and/or circumvent drug resistance. For example in our experience in tuberculosis and malaria the patchiness of explicit chemistry connectivity between papers, patents and database entries impedes progress. In the case of EBOV with less than 1500 papers in PubMed and only 25 crystal structures in the Protein Data Bank (at the time of writing), mechanistic aspects that could open the way for therapeutic developments are still not elucidated. Like others[3–5] our focus is on small molecule interventions[6–8] and we have therefore considered various steps that might help prepare for future pathogens as follows:

Text mine the literature

The highest quality and density of information about pathogens resides in peer reviewed publications, patents and databases. In recent years, text mining in general and natural language processing in particular, has become the method of choice for the extraction and collation of facts from document corpora[9]. This could thus have a rapid payoff in mining for the similarities and differences between emergent versus known pathogens. In the case of EBOV we have immediately found antiviral medicinal chemistry basic recall searches (i.e. not authentic text mining) had specificity challenges. This was also observed with synonyms for EBOV, related isolates and phylogenetic neighbors (e.g. Marburg virus)[10]. A corollary of this is that full text of at least EBOV papers could be released for text mining outside pay-walls, by agreement with publishers. Several publishers have agreed to make publications freely accessible (http://www.oxfordjournals.org/en/our-journals/medicine-and-health/ebola.html; http://www.springer.com/biomed/virology/spotlight+on+ebola?SGWID=0-1771314-0-0-0) so this would not be too difficult.

Patent assignees and/or inventors should openly declare their relevant filings

Patents contain more published medicinal chemistry data than papers[11]. For example nearly 200 WO patents for HIV protease inhibitors can be retrieved by a simple word search. This information source also presents a paradox in being, on the one hand, difficult to extract structured data from because of varying degrees of obfuscation, but on the other, full-text is easier to access than papers. In addition, not only does PubChem contain over 18 million structures from patents but also SureChEMBL now automatically extracts the chemistry from newly published filings within days. Preliminary queries indicate the patent corpus covering direct EBOV entry or replication inhibitor chemistry (or for host processing proteases as targets) is small but would nonetheless be very important to access. The only way to make retrieval rapid and complete is for assignees to openly declare their relevant published patent titles and numbers that are inevitably missed by keyword searching. In addition extraction would be much more effective if they re-surfaced the data to make it more accessible. This could be as simple as just uploading an Excel sheet to Figshare (or other open repository) with a few hundred rows of structures (linked to PubChem CIDs where these are already out there), activity values and short assay descriptions, rather than leaving the community to grapple with a hundred page PDF. We realize this is unprecedented but as a type of emergency response has to be considered. Assignee organizations should also encourage their inventors to do exactly this. Logically, another 'precedent breaker' can be considered, namely that applicants publish or surface their anti-pathogen patent results effectively the day after filing. This may sound scary to some, but IP rights are conserved while, in community terms, 18 months are cut off the "information shadow" phase. Another move in the right direction has been shown by the World Intellectual Property Organization Re:Search Consortium. Their initiative to open up patents for neglected tropical disease research could also be extended to cover filoviruses and other viruses (http://www.wipo.int/research/en/about/index.html).

Reagents and assays could be commoditized

Research can be accelerated by the collaborative exchange of assay reagents and protocols between teams and this has other positive consequences beyond just speed. Crucially, it contributes to inter-lab reproducibility if assays are made robust enough to be transferred. In addition, structure activity relationship (SAR) results will have reduced variance and will thus be more comparable between laboratories. This reciprocity becomes particularly valuable if a pharmaceutical company or other organization (Molecular Libraries Screening Centre, or Euro Screen for example) engages to run a high throughput screen. Consequently, multiple collaborators can pick up the baton of analog expansion of confirmed hits via the same standardized assay. A good example in the EBOV case is a recent publication of PDB structures for small molecules that bind the filovirus VP35 protein and inhibit its polymerase cofactor activity[12]. Supplies of assay-ready VP35 (even from a reagent vendor) would thus be valuable to expand take-up by more screening centers.

Using manual curation to enhance database links

Once a search of various publications and databases is complete one should be able to navigate reciprocally from a molecule identifier to a structure or from a target to modulating chemistry. This is not always the case, for example where pharmaceutical company lead structures are obfuscated[13]. A relevant example of useful linkages is exemplified by "*Pyridinyl imidazole inhibitors of p38 MAP kinase impair viral entry and reduce cytokine induction by Zaire ebolavirus in human dendritic cells*"[14]. It certainly helped

that SB202190 is PubChem positive as CID 5353940 (NJNKPVPF-GLGHPA-UHFFFAOYSA-N). This is well-linked as a kinase inhibitor but not directly to these recent EBOV results. The essential role of facilitating such bioactive chemistry linkage is taken by curated databases[15–17]. These not only curate literature-extracted activity results into structured records but also merge this connectivity into PubChem. The overall data findability/linkability could be further enhanced if the MeSH system could fast-track new EBOV-specific indexing and shorten the lag time.

Engage database and curation teams

Data mining is expedited by activity results and associated chemical structures being captured in databases. But the time to publish a paper and index the contents into structured records may take years. However, major chemistry resources such as PubChem[18], ChEMBL[19] and ChemSpider[20] are all willing (by prior arrangement) to take direct submissions (e.g. from EBOV screening teams) possibly even pulled directly from electronic lab notebooks (ELNs) along with the crucial metadata. The same approach of data generators actively engaging with speeding up and improving the transfer of their own results into databases applies equally to the sequence side of things. For EBOV the bioinformatics and genomics communities appear demonstrably ahead of the game compared to the medicinal chemistry community. For example the ViralZone in Europe and the US Virus Pathogen Resource were quickly established as integrated knowledge portals[21] (http://www.viprbrc.org/brc/home.spg?decorator=vipr). Significantly, the related problems of terminology mapping for curators and retrieval specificity for users is already being addressed[22]. What might be less well known is that authors submitting sequences can apply this rule by engaging directly with database staff (e.g. including feedback for MeSH indexing) to ensure the rapid and precise annotation of new virus entries.

Consider Open Science approaches

If Open Science approaches are considered, the advantages of this to pathogen drug discovery (including the abrogation of intellectual property generation) primary data sharing can then become instantaneous and global[23]. The consequent shortening of drug research stages can be dramatic. For example an InChIkey surfaced from an ELN (or other open instantiations such as Wikis or Figshare (http://figshare.com/)) means that chemistry becomes findable within hours of Google indexing[24]. It also frees teams from the 'tyranny of novelty' where new leads can be rationally optimized from pre-existing ones. In contrast the conventional IP-centric research model not only includes the years of delay to prepare a paper (i.e. 18 months after a patent application) but, even then, not all relevant antiviral chemistry flows from papers into public databases. This can also expedite the free exchange of reagents and protocols. The Open Science model can also leverage the "wisdom of the crowd" such that a global volunteer cadre of experienced chemists (industry and academic) can immediately participate both in SAR interpretation and in the design cycle. We would also suggest open sharing not only of small molecules or data sets from relevant assays, but also of a range of predictive (and sharable) models or hypotheses that can be used for virtual screening. Any SAR data from the literature

can be mined to understand physicochemical properties or molecular features important for antiviral activity. Such ligand-based computational models in turn could then be used for searching additional libraries of compounds (e.g. pharma companies might even implement this on their complete proprietary screening collections and share the results). Curated data sets could be used to construct "whole cell" virus specific machine learning models, similar to those for *Mycobacterium tuberculosis*[25]. Computational algorithms like Connectivity Map[26], SEA[27] and others could be implemented to enable fast querying of the data so that the most similar virus to an unknown could be found, and from there the most active compounds selected. Software like Euretos BRAIN, could be used to mine relationships between different biological terms and molecules that can then be used for target inference[28].

Adapt the "box" model for shareable reference compounds

The idea here is to create an openly available diverse set of compounds that are not likely to yield false positives, aggregators or other undesirable structural types commonly termed PAINS[29]. These "box" compounds would possess known antiviral and anti-pathogen activity plated out for wide availability. The notable precedents here are the MMV Malaria Box (http://www.mmv.org/malariabox), the forthcoming Pathogen Box (http://pathogenbox.org/) and the NIH Clinical Collection (http://www.nihclinicalcollection.com/). These could be hosted by a third party on behalf of NIAID, CDC etc. This model is flexible in terms of multiple "boxes" being possible. For example sets of ~1400 screening-ready FDA approved drugs (that physicians would have ready access to), would be the first logical pass for repurposing investigations. The next "box" could include the ~8000 structures in PubChem that include an International Nonproprietary Name (INN) designation and therefore in most cases have clinical testing. The advantages of what we can call 'virtuous circularity' of connectivity, apply exactly in this case. Specifically the major chemical databases can ensure a) they tag availability (i.e. a retrievable flag "this compound is in free box Y", b) the publications, patents and historical assay results are linked to the same entries and crucially c) new results (with appropriate provenance) from users of box compounds are promptly added back into the database records. By logical extension the computational modelling efforts can then loop through more rounds of improvement, new testing, leading to better hits that are put back into the 'box'.

Involve the physician's perspective

As we are seeing with EBOV, clinicians and healthcare providers are the first line of defense for the rest of the world. They are also at the greatest risk from the pathogen themselves. They are also clearly in the best position to decide how to treat their patients; the steps above should result in treatments that can actually be obtained[8] and tolerated by the patient. Physicians with experience of treating infectious diseases could be engaged to group treatments as 1. drugs which they would use in patients who are very ill; 2. those drugs which would be of concern as they may do more harm than good and 3. those drugs which might be used regardless (to explore if effective). In the case of a virulent pathogen with only palliative treatment options, from a physician's perspective,

anything that reduces mortality (even slightly) is crucial or which may even have other clinical endpoints (such as reduced hospitalization or reduced symptoms) even if mortality isn't affected. Another way to think of this is to increase the number of patients that can be treated.

In conclusion, what we are seeing now has a precedent in other viruses we were not "expecting" (e.g. HIV). Even decades on we have combination therapies to control the disease but no cure or vaccine. For EBOV we have had nearly 40 years to prepare. The cost effective suggestions above could be implemented to prepare for when the next new pathogen arrives, otherwise we will be in the same situation again. We propose that as new pathogens are identified we should be able to rapidly identify new antiviral drugs as well as establish where approved drugs that physicians have experience with, can be effective. This approach could be applicable to other infectious diseases beyond those which we currently know.

Author contributions
All authors contributed to the writing of the final manuscript.

Competing interests
SE is a consultant for CDD.

Grant information

References

1. Dhillon RS, Srikrishna D, Sachs J: **Controlling Ebola: next steps.** *Lancet.* 2014; **384**(9952): 1409–11.
   PubMed Abstract | **Publisher Full Text**

2. Anon: **Call to action.** *Nature.* 2014; **514**(7524): 535–536.
   PubMed Abstract | **Publisher Full Text**

3. Picazo E, Giordanetto F: **Small molecule inhibitors of ebola virus infection.** *Drug Discov Today.* 2014; **20**(2): 277–286.
   PubMed Abstract | **Publisher Full Text**

4. De Clercq E: **Ebola virus (EBOV) infection: Therapeutic strategies.** *Biochem Pharmacol.* 2015; **93**(1): 1–10.
   PubMed Abstract | **Publisher Full Text**

5. Southan C: **Anti-Ebola medicinal chemistry - time for crowd sourcing?** 2014.
   Reference Source

6. Ekins S, Freundlich JS, Coffee M: **A common feature pharmacophore for FDA-approved drugs inhibiting the Ebola virus [v2; ref status: indexed, http://f1000r.es/4wt].** *F1000Res.* 2014; **3**: 277.
   PubMed Abstract | **Publisher Full Text** | **Free Full Text**

7. Litterman NK, Lipinski CA, Ekins S: **Small molecules with antiviral activity against the Ebola virus [v1; ref status: indexed, http://f1000r.es/523].** *F1000Res.* 2015; **4**: 38.
   Publisher Full Text

8. Ekins S, Coffee M: **FDA approved drugs as potential Ebola treatments [v1; ref status: approved 1, http://f1000r.es/53k].** *F1000Res.* 2015; **4**: 48.
   Publisher Full Text

9. Rebholz-Schuhmann D, Oellrich A, Hoehndorf R: **Text-mining solutions for biomedical research: enabling integrative biology.** *Nat Rev Genet.* 2012; **13**(12): 829–39.
   PubMed Abstract | **Publisher Full Text**

10. Southan CD: **Anti-Ebola medicinal chemistry - time for crowd sourcing?** 2014.
    Reference Source

11. Southan CD: **Expanding opportunities for mining bioactive chemistry from patents.** *Drug Discov Today Technol.* 2014.
    Publisher Full Text

12. Brown CS, Lee MS, Leung DW, et al.: **In silico derived small molecules bind the filovirus VP35 protein and inhibit its polymerase cofactor activity.** *J Mol Biol.* 2014; **426**(10): 2045–58.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

13. Southan C, Williams AJ, Ekins S: **Challenges and recommendations for obtaining chemical structures of industry-provided repurposing candidates.** *Drug Discov Today.* 2013; **18**(1–2): 58–70.
    PubMed Abstract | **Publisher Full Text**

14. Johnson JC, Martinez O, Honko AN, et al.: **Pyridinyl imidazole inhibitors of p38 MAP kinase impair viral entry and reduce cytokine induction by Zaire ebolavirus in human dendritic cells.** *Antiviral Res.* 2014; **107**: 102–9.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

15. Liu T, Lin Y, Wen X, et al.: **BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities.** *Nucleic Acids Res.* 2007; **35**(Database issue): D198–201.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

16. Bento AP, Gaulton A, Hersey A, et al.: **The ChEMBL bioactivity database: an update.** *Nucleic Acids Res.* 2014; **42**(Database issue): D1083–90.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

17. Pawson AJ, Sharman JL, Benson HE, et al.: **The IUPHAR/BPS Guide to PHARMACOLOGY: an expert-driven knowledgebase of drug targets and their ligands.** *Nucleic Acids Res.* 2014; **42**(Database issue): D1098–106.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

18. Wang Y, Bolton E, Dracheva S, et al.: **An overview of the PubChem BioAssay resource.** *Nucleic Acids Res.* 2010; **38**(Database issue): D255–66.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

19. Gaulton A, Bellis LJ, Bento AP, et al.: **ChEMBL: a large-scale bioactivity database for drug discovery.** *Nucleic Acids Res.* 2012; **40**(Database issue): D1100–7.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

20. Pence HE, Williams AJ: **ChemSpider: An Online Chemical Information Resource.** *J Chem Educ.* 2010; **87**(11): 1123–1124.
    Publisher Full Text

21. Hulo C, de Castro E, Masson P, et al.: **ViralZone: a knowledge resource to understand virus diversity.** *Nucleic Acids Res.* 2011; **39**(Database issue): D576–82.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

22. Kuhn JH, Andersen KG, Bào Y, et al.: **Filovirus RefSeq entries: evaluation and selection of filovirus type variants, type sequences, and names.** *Viruses.* 2014; **6**(9): 3663–82.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

23. Robertson MN, Ylioja PM, Williamson AE, et al.: **Open source drug discovery - a limited tutorial.** *Parasitology.* 2014; **141**(1): 148–57.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

24. Southan C: **InChI in the wild: an assessment of InChIKey searching in Google.** *J Cheminform.* 2013; **5**(1): 10.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

25. Ekins S, Freundlich JS, Reynolds RC: **Are bigger data sets better for machine learning? Fusing single-point and dual-event dose response data for Mycobacterium tuberculosis.** *J Chem Inf Model.* 2014; **54**(7): 2157–65.
    PubMed Abstract | **Publisher Full Text**

26. Lamb J, Crawford ED, Peck D, et al.: **The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease.** *Science.* 2006; **313**(5795): 1929–35.
    PubMed Abstract | **Publisher Full Text**

27. Besnard J, Ruda GF, Setola V, et al.: **Automated design of ligands to polypharmacological profiles.** *Nature.* 2012; **492**(7428): 215–20.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

28. van Haagen HH, Hoen PA, Mons B, et al.: **Generic information can retrieve known biological associations: implications for biomedical knowledge discovery.** *PLoS One.* 2013; **8**(11): e78665.
    PubMed Abstract | **Publisher Full Text** | **Free Full Text**

29. Baell JB, Holloway GA: **New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays.** *J Med Chem.* 2010; **53**(7): 2719–2740.
    PubMed Abstract | **Publisher Full Text**

Open Peer Review

Current Referee Status: ☑ ☑

Version 2

Referee Report 17 July 2015

☑ Qiaoying Zeng

College of Veterinary Medicine, Gansu Agricultural University, Lanzhou, China

The authors have addressed my concerns, and I am happy to confirm my 'Approved' status.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Version 1

Referee Report 14 April 2015

☑ Martin Zacharias

Physics Department, Technical University Munich, Garching bei München, Germany

The recent Ebola virus outbreak came as a surprise and the authors suggest possible steps to improve the situation in case of another Ebola epidemic outbreak. The main focus is on cost effective easy implementable steps that could be taken to accelerate drug development or other therapeutic approaches. I think the suggested strategies and cost effective approaches are relevant not only in case of Ebola but could also be useful in case of other infectious diseases. The paper is well and clearly written and of interest for a broad readership in the area of pharmaceutical research, medicine and other biomedical disciplines.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Author Response 03 Jul 2015
Sean Ekins, Collaborations in Chemistry, USA

Thank you for this review.

Competing Interests: No competing interests were disclosed.

Referee Report 13 April 2015

Qiaoying Zeng
College of Veterinary Medicine, Gansu Agricultural University, Lanzhou, China

The article proposed an integrative strategy in anticipation of the next outbreak of an emerging or reemerging infectious disease like Ebola. The authors suggested seven cost effective steps that help to establish a synergistic mechanism for a fast discovery of small molecule therapeutics. Under this mechanism, the information and data from publications, patents, and database, even some reagents and assays in different labs, could be shared instantaneously and efficiently among scientific and clinical communities around the globe. The direct experience of physicians is also a great plus. These cost effective approaches could be implemented to prepare in advance for the next pathogen outbreak or newly identified infectious disease that is definite no matter when or where it starts. The article is interesting and will be beneficial for an efficient battle against the future new outbreaks of infectious diseases.

Minors: Some sentences are confused and needed to be more concise or clarified:
1. In the case of EBOV we immediately found antiviral medicinal chemistry basic recall searches (i.e. not authentic text mining) had specificity challenges, even just associated with synonyms for EBOV, related isolates and phylogenetic neighbors (e.g. Marburg virus).

2. The advantages of this to pathogen drug discovery (including the abrogation of intellectual property generation) primary data sharing can then become instantaneous and global.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Author Response 03 Jul 2015
Sean Ekins, Collaborations in Chemistry, USA

Thank you for these suggestions which have now been addressed in the latest version.

Competing Interests: No competing interests were disclosed.