The ClinGen Individual Level Database (ILDB)

Vision statement, outline and working definition

Draft 1
July 22, 2015

Snehit Prabhu
On behalf of
The ClinGen ILDB taskforce

Notes to Carlos, Heidi and Sharon:

1. This document outlines a personal vision for the ILDB project that I have pieced together after many discussions with team members and much teeth grinding on my own. I hope it does not differ significantly from each of your individual visions for this project.

2. I made an executive decision to rename it the ClinGen ILDB (individual > case). Please let me know if there is a reason we should keep the term "case".

3. On page 5, is a critical set of **"show-stopper" issues** that I can foresee, to the best of my current knowledge. Without having at least *some* due-diligence work to show against each item on this list, there will be a reasonably strong argument against to the project – either from regulators, or from a legal standpoint. If you can think of other critical show stoppers, please add in comments.

4. I've tried my best to fit the ILDB project as seamlessly as possible into the current ClinGen narrative. Please go through **figures 1 and 2** carefully, because they tell the story. I tried many alternate storylines, but this contrast worked best. Overall, I think pitching it in this manner the ILDB looks like an important piece of the ClinGen puzzle.

Aims and Objectives

The ClinGen ILDB project is an umbrella initiative to help patients and covered entities to contribute their clinical data towards disease-related genomic research. This will be done through the development of consent standards, data representation models, secure storage repositories, data submission and QC protocols, quantitative and qualitative privacy guarantee mechanisms, data usage guidelines, and lastly, regulatory standards and oversight measures for users of this data by the ILDB taskforce. The overarching aim of the ILDB taskforce will be to facilitate the transparent and seamless use of (and conversely, explicitly define the conditions for non-use of) clinical genomic, phenomic and health data – acquired in the context of delivering patient care – in medical research relevant to the disease.

Genomic information generated in clinical settings today is often locked away in data silos at their source institutions. Despite its obvious scientific value and immense potential, such data is unable to contribute meaningfully towards patient-oriented medical research – despite the often contrary intentions of those patients. A variety of causes lie at the root of this *status quo*, including (a) legal hurdles that discourage covered entities from contributing sensitive data, (b) regulatory hurdles that dis-incentivize their participation, and (c) technical hurdles pertaining to the ease with which such data might be securely deposited at a single site.

The existing state of affairs has given rise to several organic, inefficient and potentially error-prone mechanisms through which the medical professionals, each with access to a single institutional data silo, share their limited insights pertaining to the molecular mechanisms of disease. An illustration of one popular workflow surrounding clinical interpretation of genomic variants (pathogenicity assessment and actionability) in widespread use today is provided in figure 1. In contrast, we envision a new workflow by which patient data that is relevant to the task of interpretation can easily and efficiently be shared for such research (see figure 2). A framework that simplifies the use of such data has the potential to directly and positively impact patient care, with either minimal or no appreciable increase in patient risk.

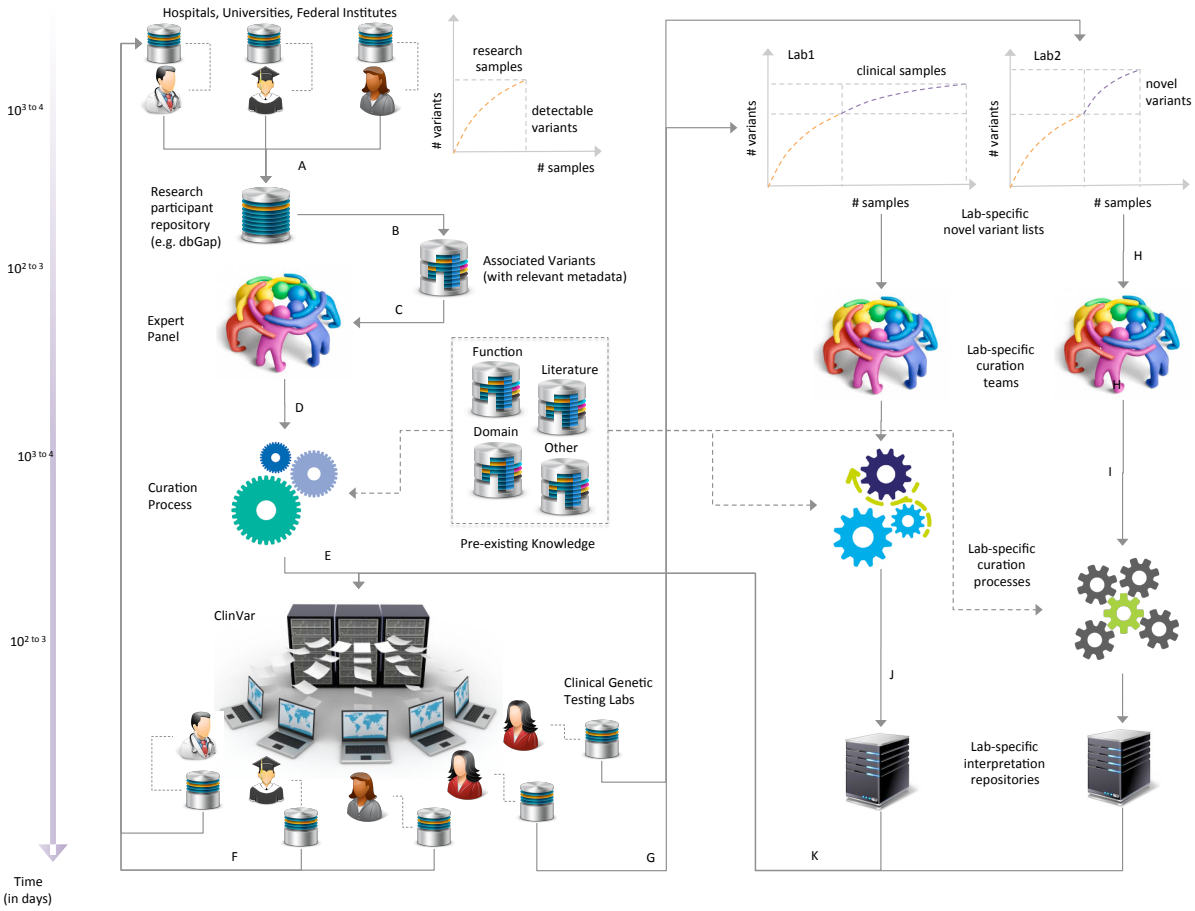

Figure 1. This info-graphic shows the flow of clinical and research-grade genomic data collected for a particular disease domain, the relevant entities handling such data, and how the results of their analysis trickle into a global scientific knowledgebase today. **(A)** A research study identifies and enrolls participants at the collaborating institutions' sites, often over a period of years. Participants' individual-level records (genotype data, phenotype data and relevant health data) typically have broad usage consent, and are subsequently aggregated into a central repository. **(B)** Disease-associated variants are revealed through genomic analyses on the dataset, **(C, D)** followed by a variant interpretation process, usually adjudicated by a panel of experts in the disease domain. **(E)** Variants judged to be clinically significant are then submitted into databanks link ClinVar. **(F)** Any additional samples collected in the interim at the study centers can provide a boost to statistical power, and potentially reveal more disease-associated variants. This necessitates a reiteration through the entire analysis (A though F). Clinical testing laboratories are usually unable to contribute relevant patient samples due to privacy-related safeguards, and are left out of the process thus far. **(G)** Often, novel but suggestive variants – with no existing clinical interpretations – are observed in the patient samples of these labs. The amount of such "suggestive" variation might differ significantly from lab to lab, due to *bona-fide* factors like ethnic representation of patients in the region or number of samples processed by the lab, but also potentially due to artefactual causes like type of genetic testing technology used and batch-effects due to undiagnosed errors in sample processing. **(H)** Absent any framework to share patient records (on their contents) with the research community, testing labs maintain internal curation teams who are assigned the same task as disease domain experts. **(I)** Curation teams at each testing lab perform triage studies to interpret their own novel variants. Although best-practice recommendations are usually observed, pathogenicity assessment might differ significantly from lab to lab. **(J)** These custom interpretations are subsequently deposited into an internal repository of the testing lab, and might become part of its intellectual property. Unlike the lab's overall curation process, individual interpretations rarely undergo rigorous peer-review. **(K)** Even if these interpretations are subsequently deposited into a community resource like ClinVar, without access to patient phenotype information and a better understanding of the process used to identify and annotate these variants at the depositing lab, significant hurdles remain in assessing their true merit.

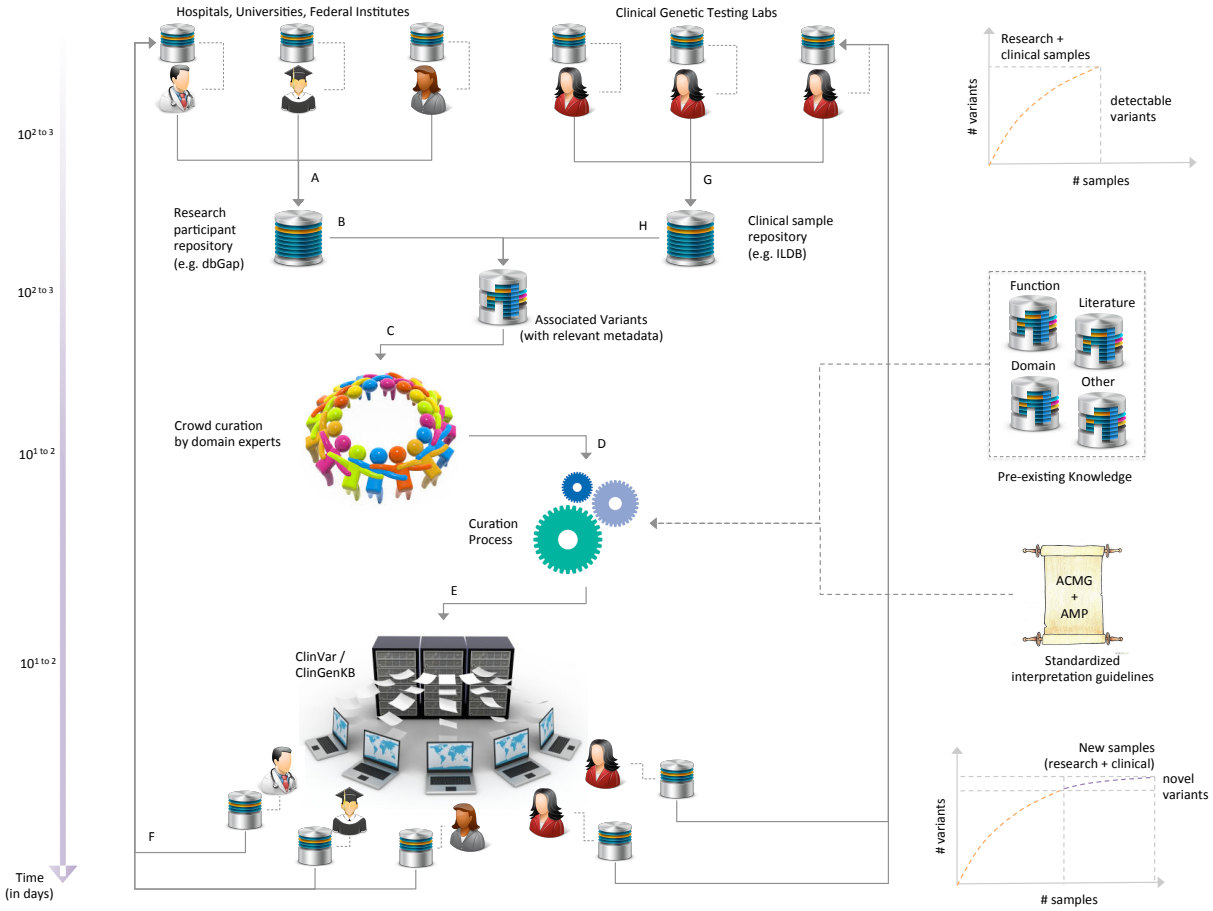

Figure 2. This info-graphic shows our new, proposed workflow for genomic data through all the entities relevant to disease-related genomic analysis. **(A, G)** Research participant samples are deposited separately as before, but the ILDB resource now allows clinical testing labs to deposit their patient samples (those with appropriate consent) and be included in the process from the start. The combination of patient and participant records will provide increased statistical power to detect disease associated genomic variation – as evidenced by the chart to the top-right corner. **(B, H)** While incorporation of research samples into analysis stays unchanged, appropriate processes to hide/transform the relevant fields in patient data need to be developed. Subsequent downstream analysis will be restricted to "variant" and "gene" level information, with curators having no access to patient-specific phenotype or health data **(C, D)** The ClinGen crowd curation infrastructure (phone-app, web-interface) will bring together teams of domain experts for scalable interpretation of the much larger pool of associated variants. The curation and triage process for each disease domain will adhere to common standards, and remove lab-specific biases in interpretation. **(E, F)** Variant interpretations will be deposited in the new ClinGenKB repository, with any additional patient records, research samples or associated variants being iterated through the entire process (A through F), as before.

Innovations required for the use of clinical data in genomic research.

In order for hitherto-protected clinical data from patient records to be leveraged, The ILDB task force will need to provide solutions to the list of key issues. These are:

1. **Consent standards**
   [Text]
   [Contributors: Snehit Prabhu. Kelly Ormond? Heidi Rehm? Sharon Plon?]

2. **Data representation and storage models**
   [Text]
   [Contributors: Snehit Prabhu. Larry Babb? Sandy Aronson?]

3. **Data submission and QC protocols**
   [Text]
   [Contributors: Snehit Prabhu. Sam Baxter? Samuel Aronson? Larry Babb?]

4. **Patient privacy and security frameworks**
   [Text]
   [Contributors: Snehit Prabhu. Sharon Plon? Sam Baxter?]

5. **Data usage guidelines and oversight/regulation of users**
   [Text]
   [Contributors: Snehit Prabhu. Heidi Rehm?]

Through the research and development of these mechanisms, the ILDB taskforce will provide a comprehensive solution through which appropriate clinical data might be identified and incorporated into research studies. The ILDB will provide a superior alternative to numerous *ad-hoc* solutions in use today, through which clinical data is manually "scrubbed" to protect patient identity before being shared with researchers.