NIH/dbGAP の動向

国立研究開発法人科学技術振興機構
バイオサイエンスデータベースセンター

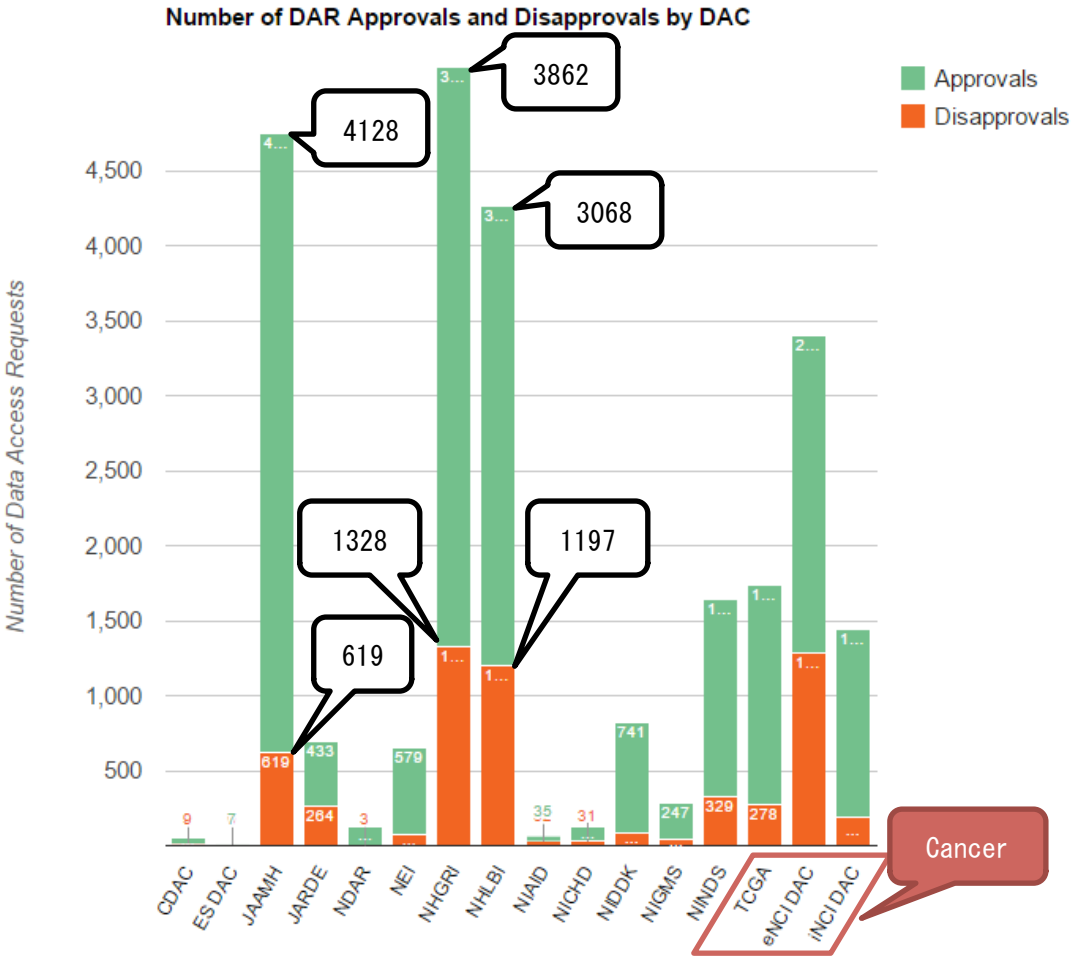
2015 年 5 月 19-21 日に NCBI にて開催された International Nucleotide Sequence Database Collaboration (INSDC)三極会議に DDBJ の真島さん、児玉さん等が参加し、5 月 22 日に dbGAP を訪問し、Dr. Michael Feolo、Dr. Lon Phan 等と打合せをした。その報告を受け、特に Data Access Committee に関与する部分について報告する。

１：データ利用申請数および承認数（2007 年以降）

利用申請数：30374、承認数：20272

2/3 程度が承認されている。データの『利用制限（制限事項)』に抵触することから否認されている。

（http://gds.nih.gov/20ComplianceStatistics_dbGap.html#FN_violation_01 のリスト参照）

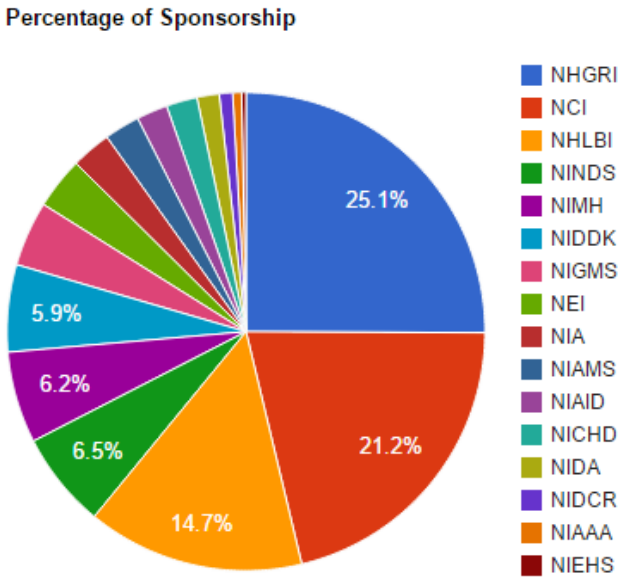


http://gds.nih.gov/19dataaccesscommitteereview_dbGaP.html

NIH には１６の DAC が存在する。それぞれの Institute や Center に DAC が存在することが多い。JAAMH の様に薬物中毒・アルコール中毒・エイジング・メンタルヘルス 4 機関で１つの DAC を所有していたり、The Cancer Genome Atlas (TCGA)の様にプロジェクトで DAC を所有することもある。

２：dbGAP に登録されているデータは、（１）National Human Genome Research Institute (NHGRI)、（２）National Cancer Institute (NCI)、（３）National Heart, Lung, and Blood Institute (NHLBI)の順で多い。341/350 研究（97.4%）は米国からの登録（日本からは 1 研究）。

NIH Institutes and Centers Sponsoring[2] dbGap Studies



http://gds.nih.gov/17summary_dbGaP_statistics.html

３:すでにデータの 2 次利用で 950 報以上の論文が Publish されている(Data Use Report の記載を参照)。
＞日本でも 2 次利用の論文を成果リストとして NBDC ヒト DB から公開していきたい。

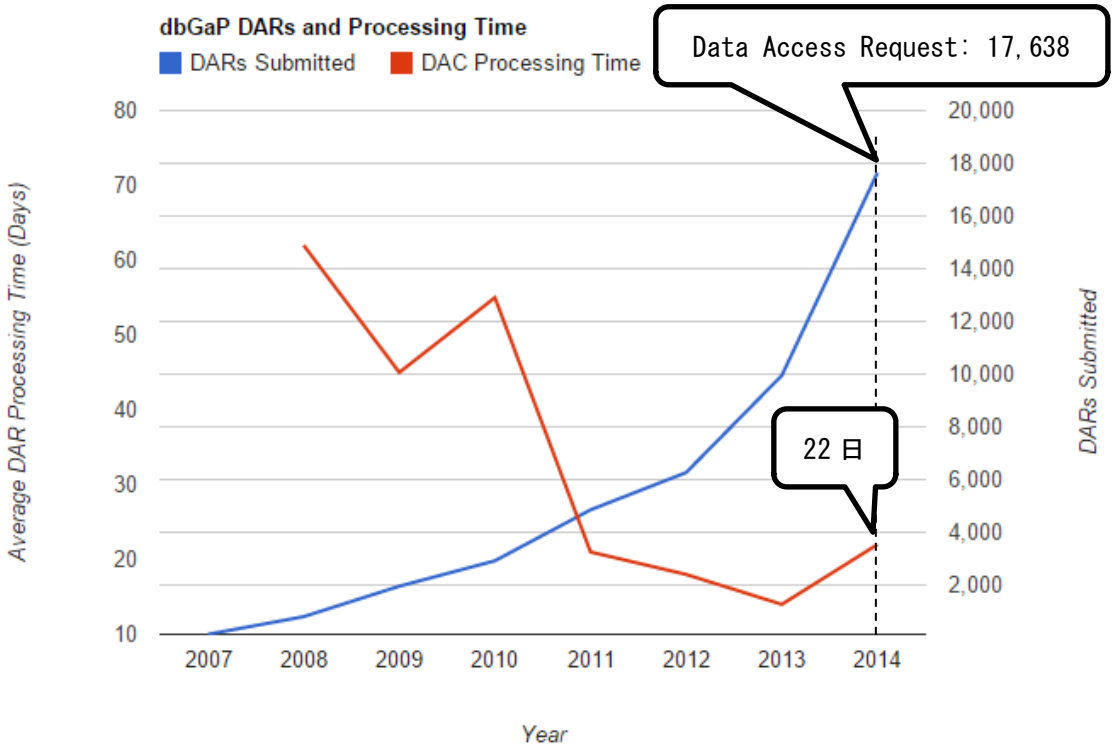
４：100 万人以上の被験者のデータが登録されている。データキュレーションを実施し、同一検体からのデータは同じ内部 ID が付けられている。新たに検出された変異は dbSNP や dbVar に登録している。
＞NBDC ヒト DB でも、同一検体からのデータかどうか等のデータキュレーションを実施していきたい。

５：GDS policy では、（１）NIH が Funding した全研究データをタイムリーに公開する（２）ヒトデータについては、Quality Control が終了次第、データを Submission し、6 か月以内にデータをリリースする（３）Publication Embargo 期間については、NIH による監視が難しく違反が起こりえるので、初めから設けないこととする、という運営を目指す。

６：NBDC ヒト DB ではアクセス制限の表記を NIH の Genome-wide Association Policy に合わせて、『Open Data』と『Controlled Access Data』と名付けていたが、NIH の GDS に従い、NBDC でも Open data を Unrestricted Access Data と名称変更することとする。
７：Controlled access data を使用する際は、Genomic Data User Code of Conduct を遵守する必要が

ある。Genomic Data User Code of Conduct は『再配布の禁止』といった基本的項目が記載されていて（７項目）、NBDC ヒトデータ共有ガイドラインの中に記載済みの内容であった。

８：データ利用申請に係る審査期間は平均 22 日



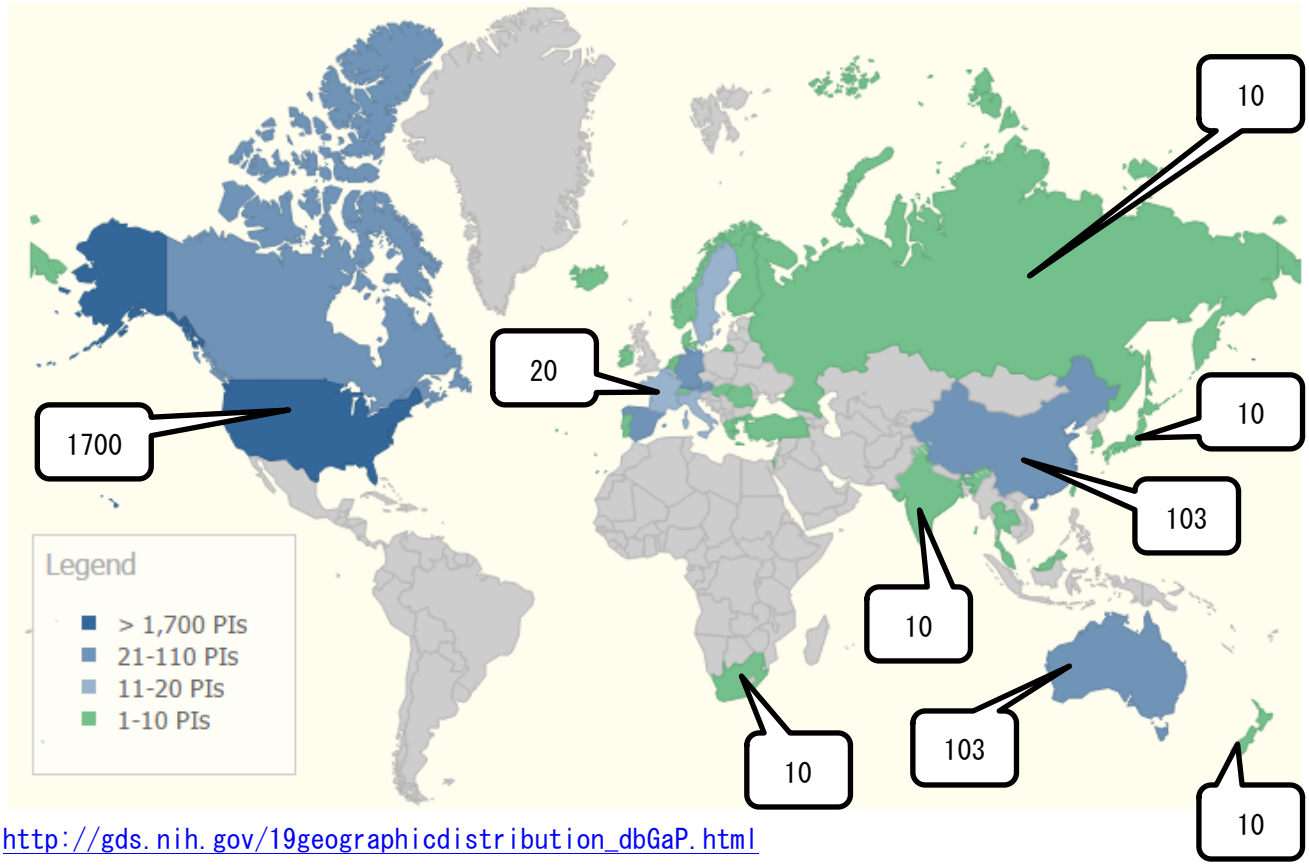
http://gds.nih.gov/19timeprocessing_dbGaP.html

９：データの質やデータ共有方法について NIH 基準を満たす組織は、Trusted Partner になることができる。NIH の Institute や Center と契約をすることで Trusted Partner になれる。ただ Data Access に関する審査は NIH の Institute や Center の DAC が実施する。
既に Trusted Partner となった組織：Cancer Genomic Hub, Bionimbus, Cancer Genomics Data Commons, Cancer Genomics Cloud Pilot Project の４つ。

１０：2015 年 3 月 23 日、クラウドサービスを利用した制限公開データの利用（解析や保管）に関する GDS policy への記載が公表された。クラウドサービス提供者ではなく、Controlled access data をクラウド環境で利用する者が所属する組織が責任を持つ、とした。データ利用時には、PI、組織の長、組織の IT 管理者がサインし、セキュアなクラウドを利用しているかどうかの責任は IT 管理者にある。クラウド環境によるデータ利用を希望する際は、利用申請時にクラウドを使用する旨を表明し、研究過程でどのような使用を予定しているか、また、どのプロバイダーのサービスを利用するのかについても報告しなければならない。現在、29 件利用している（大学・研究機関：17、民間：９、非営利：１、NIH：２）。プロバイダーは以下の通り。Amazon：11、Google：５、Private：５、DNAnexus：４、Microsoft：２、Seven Bridges：２
１１：データ利用申請時に DAC が確認していることは『どんな研究に使うか』と『Data Use Restriction』

が一致しているかどうか。Data Use Restrictionについては参考資料(Standard Data Use Limitations.pdf)参照

１２：データ利用者数（PI 数）3597 PIs (41 countries)



http://gds.nih.gov/19geographicdistribution_dbGaP.html