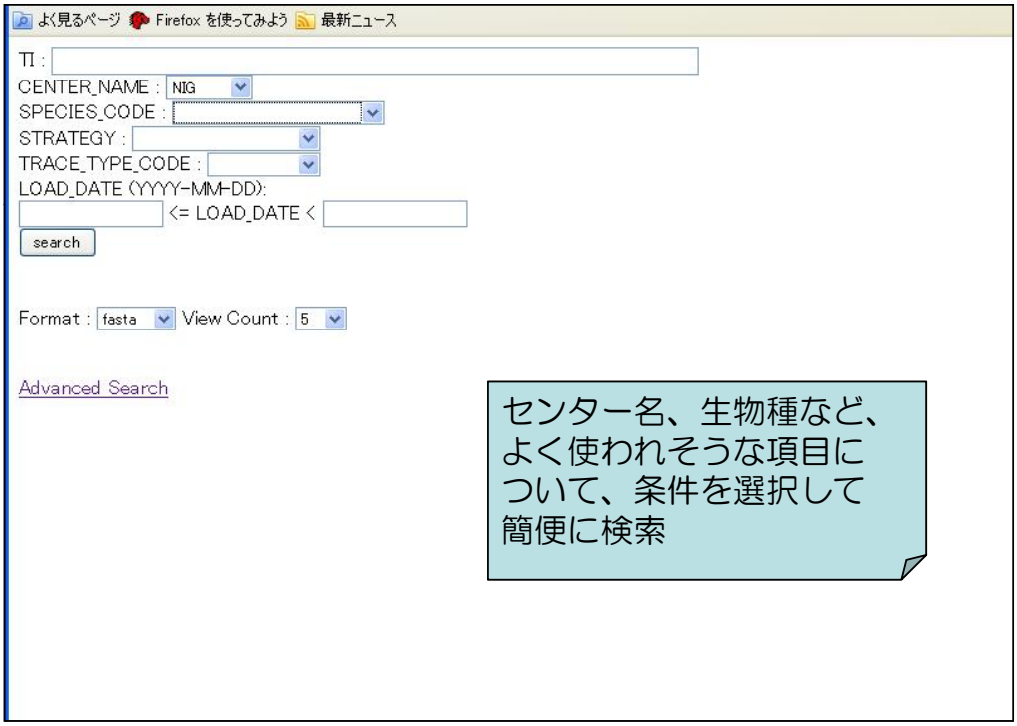
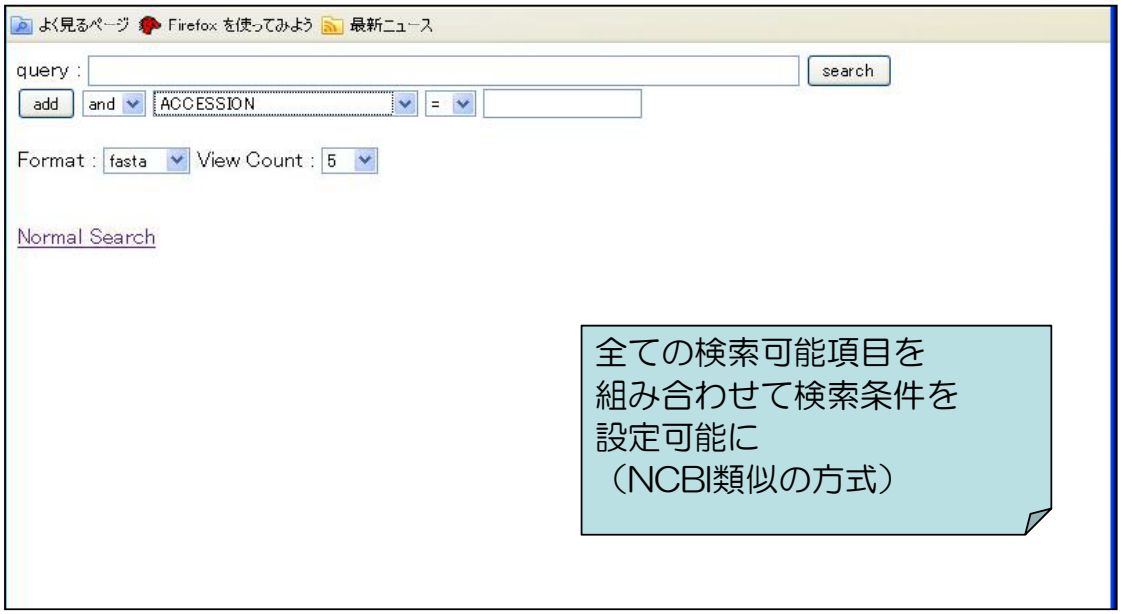
資料 2

- 遺伝研補完課題
  - 進捗状況
    - 開発
      - 全般：プロトタイプを元に改良開発中
      - データベースの構築
        - 検索用DBの項目を決定
        - 仮の実装は動作可能レベル
        - 実データを投入して「試験⇒改良」のサイクル中
      - web検索システム
        - 検索、結果表示、個別データDLまで動作可能。
        - DBとあわせて「試験⇒改良」のサイクル中
        - 検索画面（通常）#1
        - 検索画面（詳細）#2
      - 波形表示システム
        - 既存プログラムに手を入れ、テスト中
        - 仮方針。後日の取替え、再開発も選択肢
      - 統計情報表示ページ
      - 統合DBセンター連携
        - 詳細#3
    - 運用
      - 登録～公開の運用
        - ワークフロー概要#4
        - TR登録・公開の実績#5
      - ショートリード
        - 既に対応開始
但し試行錯誤
          - NCBIも走りつつ対応中の模様
            - 1000人ゲノムと同時並行
          - データ構成が比較的複雑
            - データ構成の説明#6
          - 従って、1件の登録ごとに、登録者ならびにNCBIとの間で繰り返し手戻りが発生#7-8
        - SRA登録・公開の実績#9
        - 2月にNCBIを訪問、パイプを強化する予定
  - 来年度予定
    - 基本は
      - 開発してきたものの運用を開始
      - 実運用を進めながら改修と拡充
    - 新規取り組み
      - BoLやセルイノベーションのデータ登録！？
        - 参考：BoLとは・・#10
        - セルイノベーションの計画では・・
      - データ量の見通しに基づいて必要資源を見積もりたいが・・#11-14
      - スケーラビリティは？⇒今年度開発システムの見直しも必要か？
      - その他候補
        - ウェブサービス化
        - 相同性検索
          - 特にBoLには必要
          - 資源は？
        - ショートリード
        - 代理登録⇒自律ID発行可能に？
      - 随時状況を先読みしつつ柔軟な対応が必要

検索画面(通常インタフェース)



検索画面(詳細インタフェース)

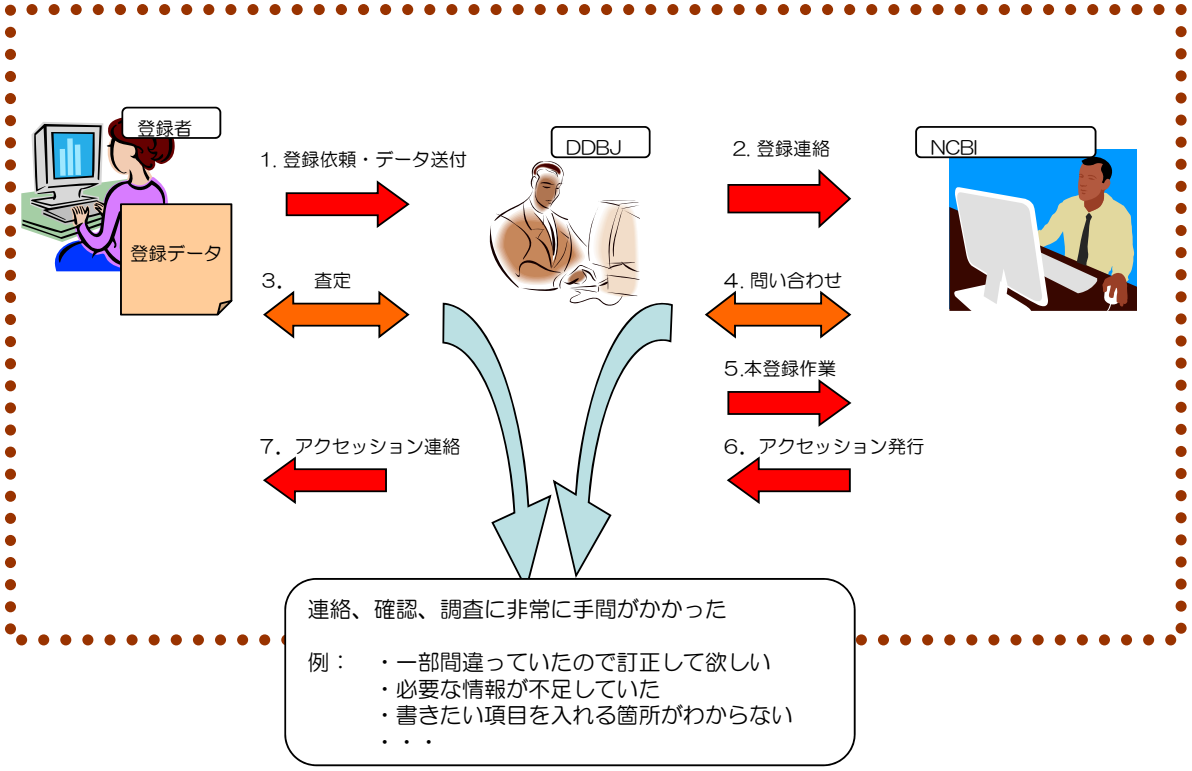


統合DBセンター連携

- 内容の協議
  - DDBJ：池尾准教授、DBCLS：西川教授
  - 昨年９月、今年1月とこれまで2回

- 協力決定事項
  - 「横断検索」向けメタデータの提供
    - DDBJ側でメタデータ作成、提供を予定
    - 今後の運用でも継続的に作成して行く
    - 提供の単位は各登録（プロジェクト）ごと
    - 横断検索側では、検索結果からDDBJのFTPへリンク
  - 「ダウンロードサイト」向けメタデータの提供
    - まずは「トレースアーカイブ」全体について
    - DDBJ側から、必要な項目の情報を提供予定

Trace Archive登録ワークフロー



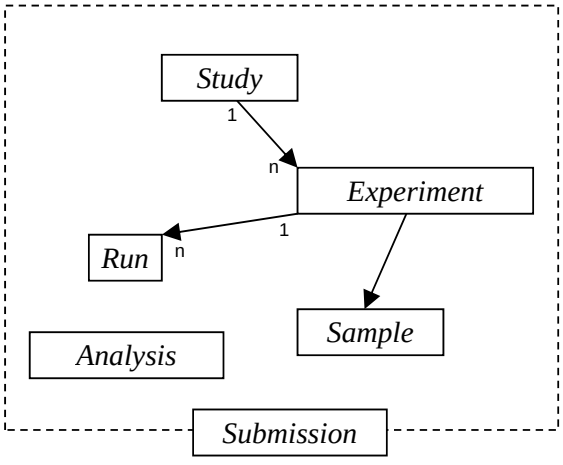
Trace Archive登録実績

項番	ステータス	機関名	生物種	機種	件数	容量(圧縮時)	TI発行日	作業期間	状況
1	完了	遺伝研（NIG）小原研	メダカ	ABI	約１４８万件	約５０ＧＢ	7月上旬	約４ヶ月	登録、公開完了
2	完了	東大新領域（UTCOB）服部研	ヒトメタゲノム	ABI	約１０６万件	約４０ＧＢ	7月上旬	約３ヶ月半	登録、公開完了
3	問合せ段階	非公開	非公開	ABI	BAC数で2147件	不明	未完了		9/30にご案内送信、返信待ち

Short Read Archiveのデータ構成

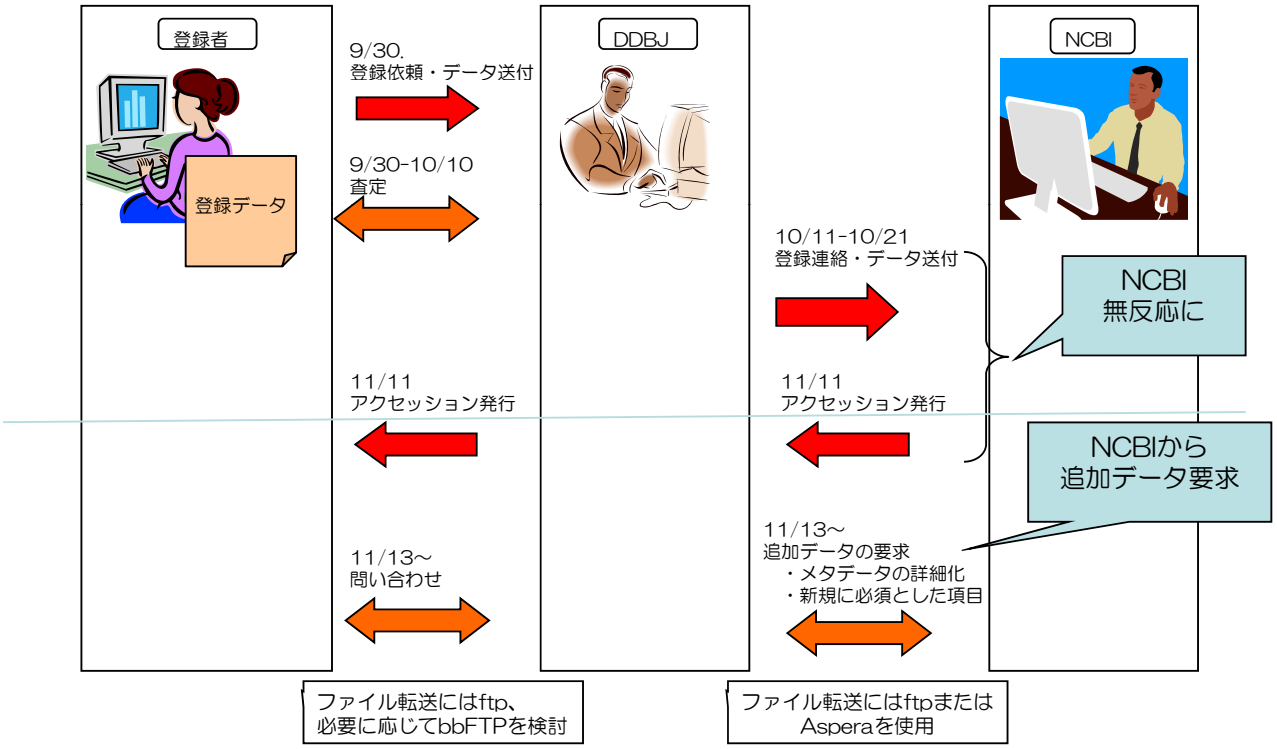
- **Study** – Identifies the sequencing study or project and contains multiple experiments.
- **Sample** – Identifies the organism, isolate, or individual being sequenced.
- **Experiment** – Specifies the sample, sequencing protocol, sequencing platform, and data processing that will result one or more runs.
- **Run** – Identifies run data files, the experiment they are contained in, and any runtime parameters gathered from the sequencing instrument.
- **Analysis** – Packages data associated with short read objects that are intended for downstream usage or that otherwise needs an archival home. Examples include assemblies, alignments, spreadsheets, QC reports, and read lists.
- In addition, all details concerning submissions are contained in a separate document called **Submission**, which contains center specific submitting information, contacts, actions for the archive, and a file manifest.

※SRA Submission Guidelinesより抜粋



5種の「オブジェクト」それぞれに
アクセション番号が発行される：
Submission (SRA), Study(SRP),
Experiment(SRE), Sample (SRS),
Run(SRR)

Short Read登録処理の実際の例(1)



Short Read登録処理の実際の例(2)

具体的な作業および登録者やNCBIとのやり取りの実例

(9/30) 登録依頼のメールを受領。メタデータはエクセルで受領。生データはFTPで受領。

(9/30) ファイル転送、内容確認、再圧縮等の作業

(10/1)登録者の依頼でメタデータの一部を修正

(10/2)データの追加分を受領

(10/6)追加データのチェックの結果、ファイルの破損があることが判明。登録者に連絡

(10/8)破損したファイルを再度受領。

(10/10)メタデータの修正項目についての確認。

(10/10)メタデータの修正方針に問題ない旨のご連絡を頂く。

(10/11)NCBIへ登録受領を連絡。データをFTPでNCBIに転送開始。転送失敗で再送数回。

(10/21)NCBIへ転送完了を連絡。

➔NCBI担当者からの連絡途絶（多忙のため）

(10/21-11/10)NCBIへ何度も確認や催促。登録者から数回の進捗の確認。

(11/10)NCBIからアクセッション番号の通知。

(11/11)登録者へアクセッション番号を連絡

➔NCBI担当者を増

(11/13)NCBIからstudy_title等のメタデータの修正依頼あり。

(11/14)登録者へ、NCBIからの修正依頼があった旨を連絡。

(11/15)NCBIから、sampleデータの詳細化要求。

(11/17)登録者から、study_titleの修正結果を受領。DDBJで査定の結果、再修正を依頼。

(11/18)登録者から、study_titleの再修正結果を受領。NCBIに連絡。

(11/21)NCBIへsampleの詳細と、訂正事項を送付。不足事項について登録者に問い合わせ。

(12/2）登録者から不足事項の連絡あり。

(12/4)NCBIへsampleのattributeの詳細を送付

(12/5)登録者へproject_id取得のためのフォームを送付

(12/17)登録者からproject_id取得のための情報を確認した旨の返信あり

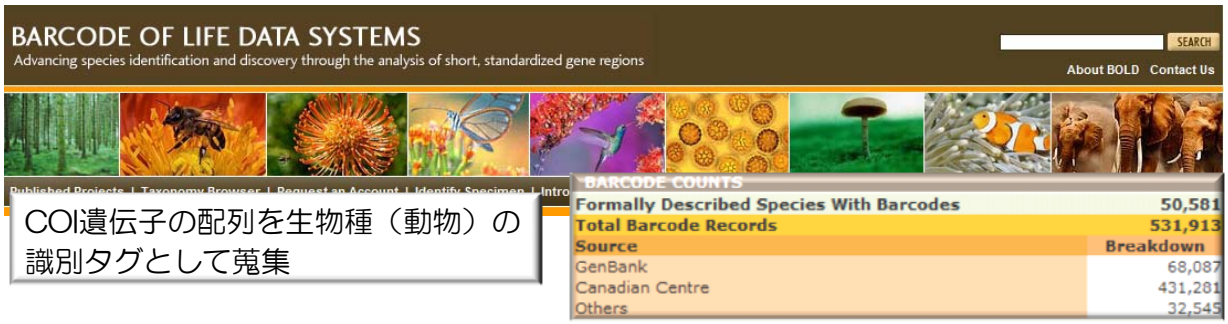
・・・さらに継続中

Short Read Archive登録実績

項番	ステータス	機関名	生物種	機種	件数	容量(圧縮)	SRA発行(※1)	作業期間(※2)	状況
1	進行中(SRA発行済)(公開要請)	東京大学新領域創成科学研究科菅野・鈴木研	寄生虫6種と媒介昆虫2種	Solexa	約1億件	約300GB	2008/9/14	約2ヶ月	NCBIへ送付完了 NCBIからのメタデータ追加の要求に対応中（※3）
2	進行中(SRA発行済)(公開要請)	理研林崎研	ヒト	Solexa	20万件	約37MB（fastQ形式）	2008/11/12	約2ヶ月半	NCBIへ送付完了 NCBIからのメタデータ追加の要求に対応中
3	問い合わせ段階	（非公開）	(非公開)	Solexa	200万件	不明	未完了		9/4にご案内を送信、返信待ち
4	進行中(SRA発行済)	（非公開）	(非公開)	Solexa	約4000万件？	約23GB	2008/11/11	約2ヶ月	NCBIへ送付完了 NCBIからのメタデータ追加の要求に対応中
5	進行中(SRA発行済)	(非公開)	(非公開)	Solexa	不明	約450GB	2008/11/11	約1ヶ月半	NCBIへ送付完了 NCBIからのメタデータ追加の要求に対応中
6	キャンセル	(非公開)	(非公開)	454	不明	不明	キャンセル	キャンセル	キャンセル
7	完了（非公開）	（非公開）	(非公開)	454	1 run分（454のbinary ）	約800MB	2008/12/16	約1ヶ月半	完了。HUPデータにつき未公開
8	進行中	（非公開）	(非公開)	Solexa	不明		未完了		データ受領済、査定作業中
9	問い合わせ段階	（非公開）	（非公開）	Solexa	不明	不明	未完了	未完了	12/5にご案内を送信、返信待ち

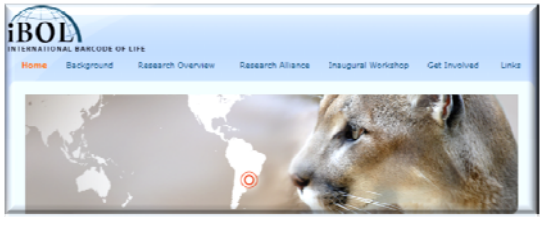
※1 登録時にSubmissionのAccession番号（SRA）が発行されるが、データ（Study、Experiment、Run等）に対するAccession番号は、その後の登録処理を経て発行される。
※2 最初の連絡から、SRA発行までに要した時間。
※3 NCBIでの処理の遅れから、登録後にデータの詳細化や項目追加を要請され対応中。

BoLとは・・

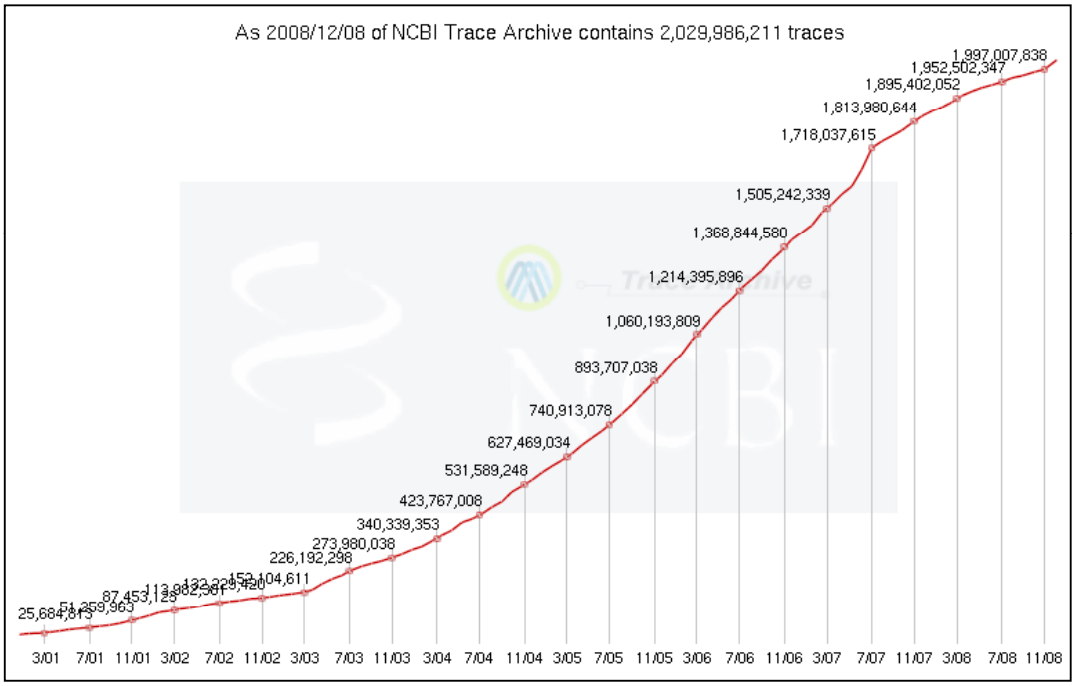


COI遺伝子の配列を生物種（動物）の識別タグとして蒐集

iBoL：5ヵ年で
50万種、
500万標本の調査予定

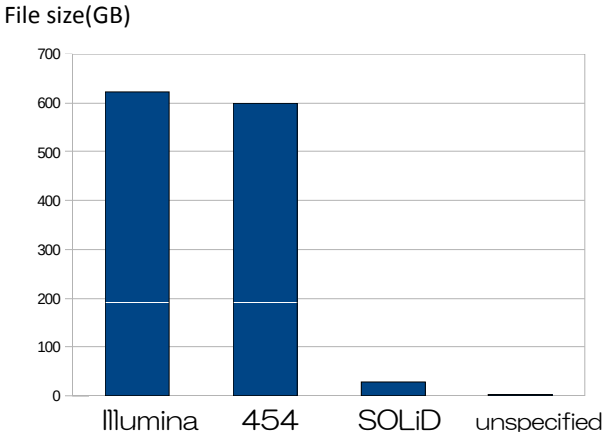


NCBI Trace Archiveデータ推移



全体のファイルサイズ：約67TB（推定）

NCBI Short Read Archiveデータサイズ



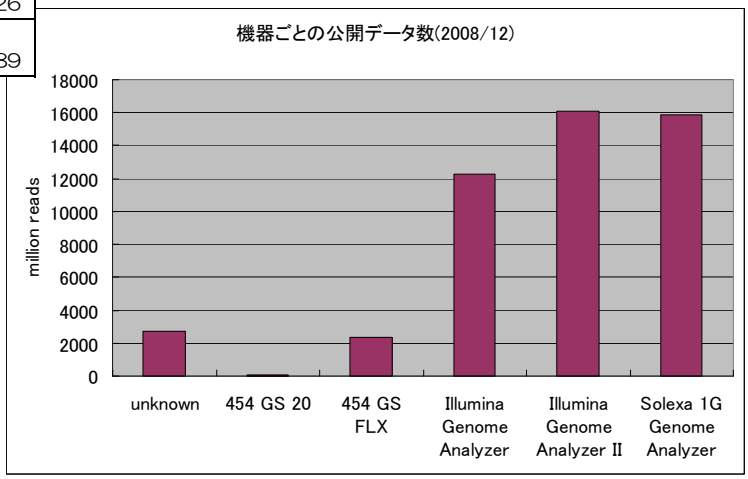
2008/12時点（実質2008/6まで）

- 現在NCBI SRAからＦＴＰ ダウンロードできるファイルのサイズ。
- 2008/06までのエントリ数（プロジェクト数）は358
- 2008/07以降のデータはほとんどFTPサイト上で未公開（アクセッション番号のみ公開）
- 2008/07-2008/12のエントリ数は2,542（ほぼ1000人ゲノム由来）
- SRAの登録は2007/06より。

NCBI Short Read Archiveデータ件数

	runs	reads
unknown	340	2,727,697,394
454 GS 20	215	106,046,331
454 GS FLX	2,026	2,320,781,032
Illumina Genome Analyzer	885	12,276,263,288
Illumina Genome Analyzer II	1,073	16,110,342,026
Solexa 1G Genome Analyzer	886	15,890,127,389

・Webサイトで公開されている情報を集計
・FTP公開されていないが閲覧可能なものも多数
・2008/12時点



新世代シーケンサ導入状況

機関・グループ	機器
東大新領域	454, Solexa
遺伝研	454, Solexa
岡崎：長谷部グループ	Solid
沖縄：佐藤グループ	454
産総研-沖縄	Solid
理研	454, Solexa, Solid
農水動物遺伝研	Solexa
農水STAFF研	・・・
・・・	・・・