



---

## FORMATION RADSEQ SOUS GALAXY

Intervenant(s) : Yvan Le Bras, Cyril Monjeaud



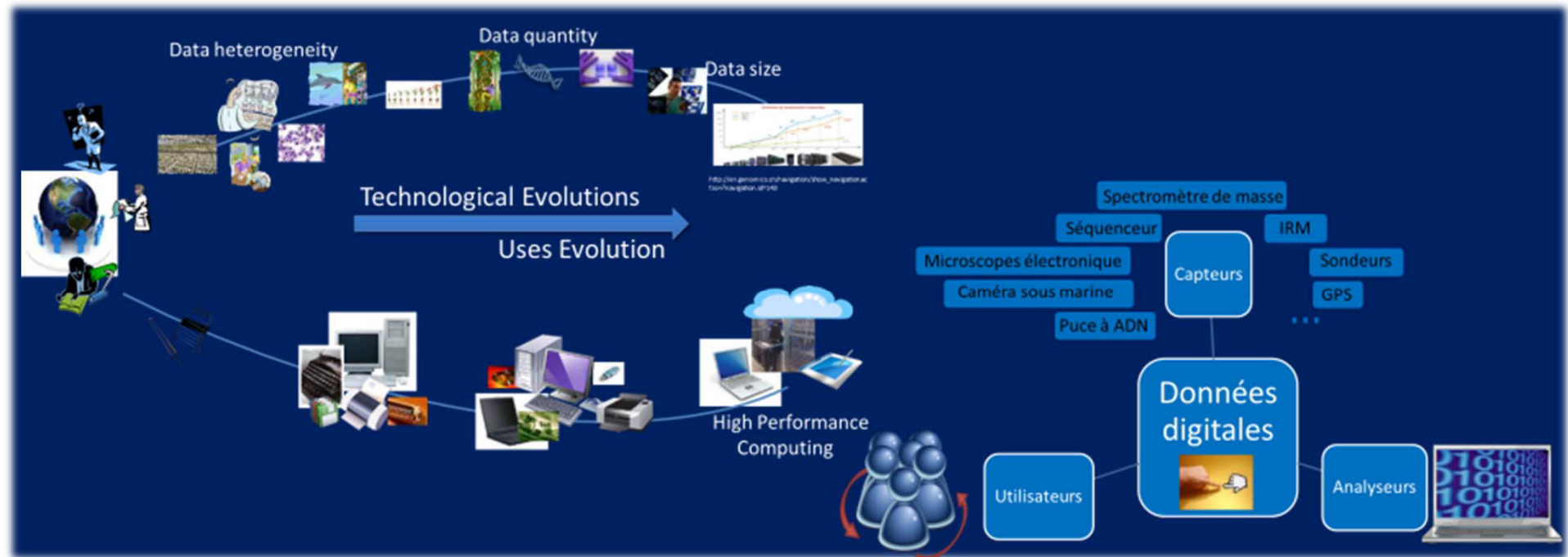
# E-BIOGENOUEST

---

Programme fédérateur Biogenouest co-financé par les Régions Bretagne et Pays de la Loire

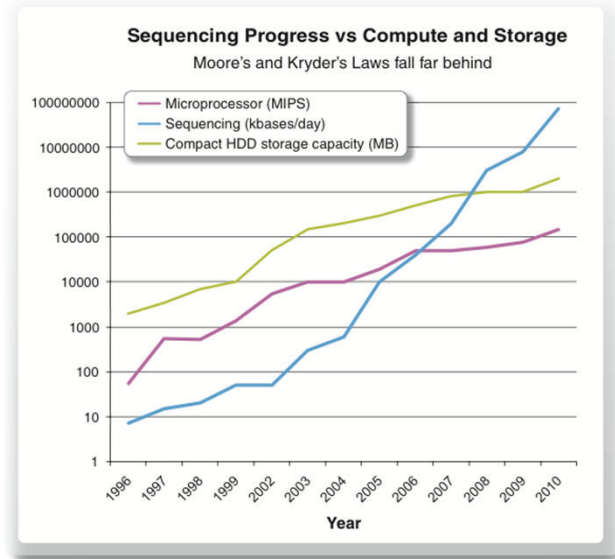
- 24 mois
- Lancé depuis Mai 2012
- Porteur : Olivier Collin (IRISA) – Animateur : Yvan Le Bras (IRISA)
  - Tester une approche e-Science en Sciences de la vie
  - Proposer une structuration e-Science dans le Grand Ouest

# Briser les silos



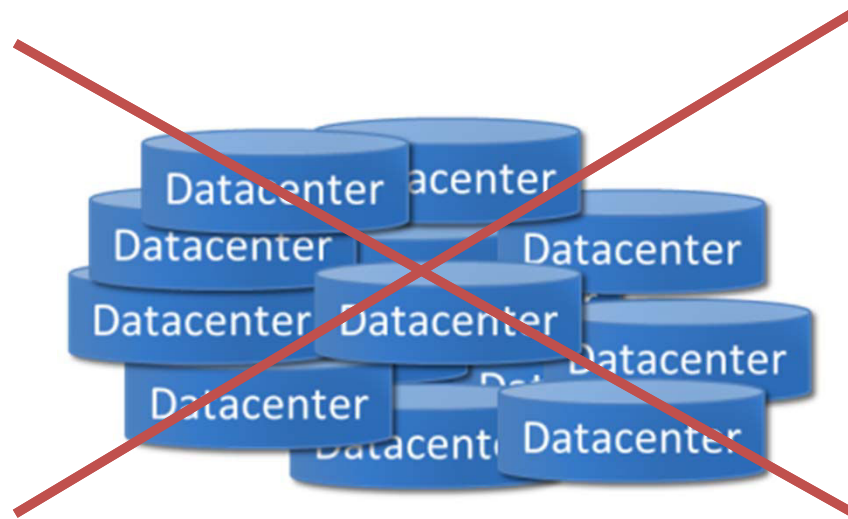
- Défis : Chaque domaine est particulier mais digital!
- Solution : Se centrer sur la donnée!
- Big Data relatif, collaboration, interdisciplinarité, mutualisation, solutions open source

# Optimiser / Standardiser



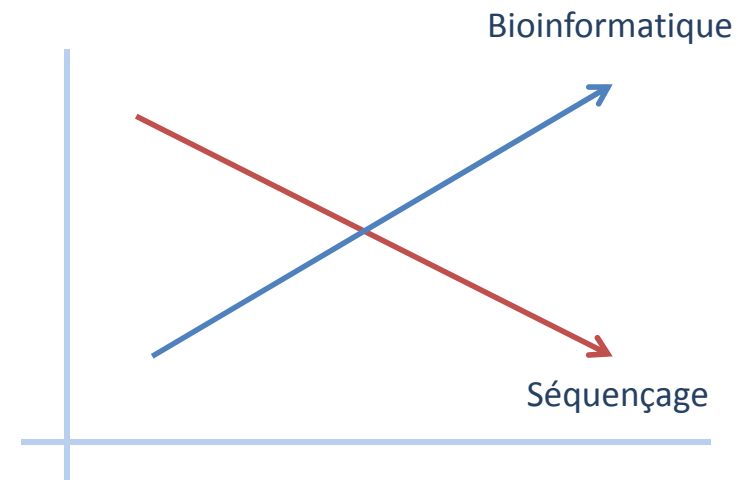
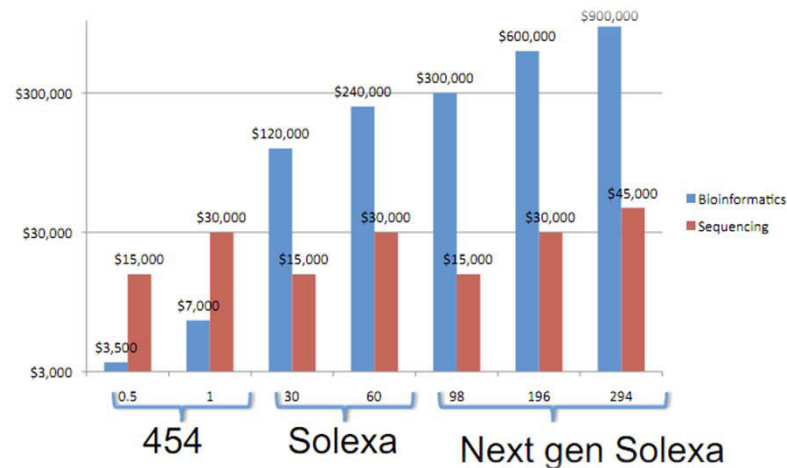
**Fig. 1.** A doubling of sequencing output every 9 months has outpaced and overtaken performance improvements within the disk storage and high-performance computation fields.

Kahn. On the future of genomic data. Science (2011) vol. 331 (6018) pp. 728-9



- Code, algorithmes, calcul, stockage, réseau, bonnes pratiques
- Prévoir interopérabilité, langage commun, vocabulaire contrôlé, ontologies

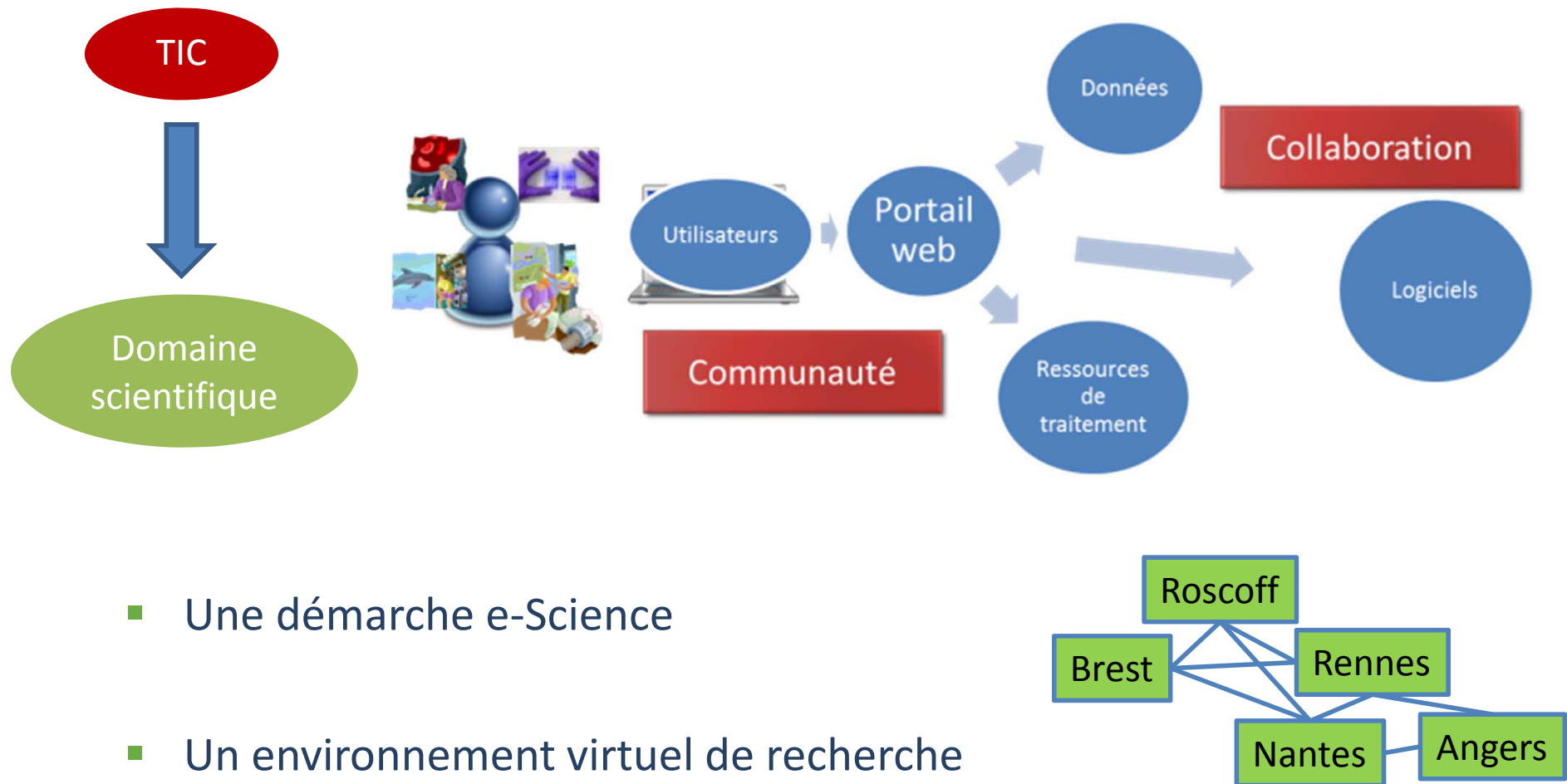
# People paradox



Using Clouds for Metagenomics: A Case Study Wilkening et al.  
IEEE cluster 2009

- Défis : Manque RH, Evolution des usages
- Solutions : mutualisation, science citoyenne, Environnement virtuel de recherche

# Solution : e-Science



# LE VRE

---

Un environnement virtuel de recherche en sciences de la vie

# Un début de structuration e-Science?

## Le HUB eBGO

HUBzero pour partager les connaissances et gérer les groupes et projets

### Informations

164 utilisateurs

83 projets

52 groupes

657 ressources

...





# Un début de structuration e-Science?

## eBGO : Collaboration

The screenshot displays the eBGO website interface. At the top, the eBGO logo is accompanied by navigation links: Discover, e-Infrastructure, Community, About, and Support. A user profile for Yvan Le Bras is visible in the top right corner.

The main banner features the text "Galaxy by GenOuest statistics" with the following data: 22729 jobs run to date, 15 jobs run this month, and 1 jobs run today. To the right, a dark box promotes the "Use e-Biogenouest Hub to" with the following actions: Create or join Groups and Projects for collaboration, Contribute Resources for education, training or knowledge sharing, Follow e-biogenouest project & e-Science point of view, and Manage or Analyse data.

Below the banner, a section titled "Welcome to ebgoHUB : The e-Science initiative of Biogenouest" is displayed. The page is divided into three main columns:

- Featured Resources:** Lists resources such as "iMarine Gateway - in Linked Resources", "Journal du CNRS Mai 2013 : Portail collaboratif du CNRS Core - in Linked Resources", and "Testons le blog avec Jaccard by Guillaume Collet - in Blogs". It includes a search bar and a section for Popular Tags (Galaxy, NGS, Data analysis, Training, Genomics, Data management, Population Genomics, Tutorial, Bio-informatics, Data, HUBzero, eBGO HUB, Tools, More tags).
- Utilize:** Promotes tutorials and documentation, including "HUBzero users documentation" and "eBGO HUB video tutorials".
- Community Polls:** Features a poll titled "How much time do you spend using Bioinformatics tools ?" with options: Less than 20%, From 20 to 40%, From 40 to 60%, From 60 to 80%, and More than 80%. It also includes a "Vote" button and a "Results..." link. Below the poll, a section titled "Who are you?" lists user roles: Graduate student, Undergraduate student, High school or elementary student, College faculty, and Professional scientist/engineer.

At the bottom, a footer links to various resources: Courses, Datasets, Downloads, Linked Resources, Manuals, Ph.D. Thesis, Presentation Slides, Publications, Reports, Seminars, Series, Tools, Tutorials, Workshops... All Categories.

# Un début de structuration e-Science?

## EMME

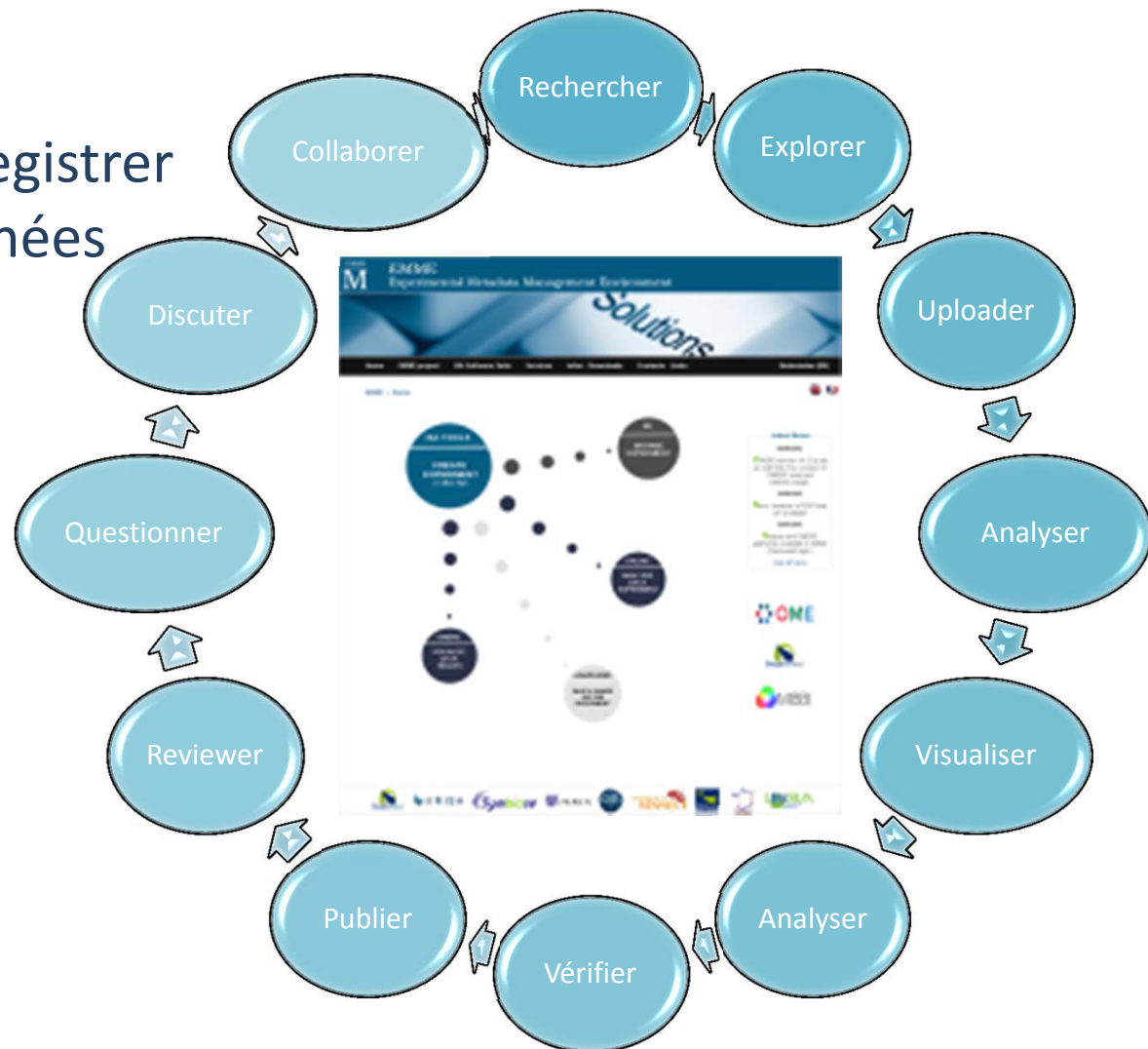
suite ISA tools pour enregistrer  
données et métadonnées

### Protocoles existant

Métabolomique  
Protéomique  
Génomique  
Bio-informatique

### Protocoles sur mesure

Colorimétrie  
Dosage d'ions plasmatique  
...



# Un début de structuration e-Science?

## EMME suite ISA tools : isacreator configurator

The screenshot displays the eBGO website interface. At the top, there's a navigation bar with links: Discover, e-Infrastructure, Community, About, and Support. Below this, a menu structure shows 'Data-centered World' leading to 'Metadata management', which then leads to 'Data temporary repository'. Other options include 'Education' leading to 'Data analysis' and 'Processing resources' leading to 'Genocluster FTP'. A user profile for 'Yvan Le Bras' is visible in the top right corner.

The main content area is titled 'ISA' and contains descriptions for three tools:

- ISAconfigurator**: Java Application for curators to customize and regulate the fields displayed in ISAcreeator, by declaring some required fields and specifying allowed values (e.g. from a set of ontologies term).
- ISAcreeator**: friendly Java application with which the "experimentalists" can include or edit experimental metadata and validate the ISA-Tab files formatted according to the specified configuration (via the ISAconfigurator).
- ISAconverter**: Java application to validate and convert ISA-Tab files into several other tabular or XML formats to submit into public repositories such as ENA (genomics), PRIDE (proteomics) and ArrayExpress (transcriptomics).

Below the descriptions, there's a section for 'isacreator configurator' with a logo and instructions: 'These applications are launched by Java Web start. During the first launch, a directory named ISA\_metadata will be created in your home directory. T'.

Overlaid on the bottom right is a screenshot of the 'isacreator configurator' application window. The window has a menu bar (File, Mappings, Validation, Help) and a title bar 'isacreator configurator'. It shows a list of 'tables & forms' on the left, including 'cellcount\_flowcytometry', 'cellsorting\_flowcyt', 'clinical\_chemistry', 'copynumvariation\_micro', 'dnamethylation\_micro', 'dnamethylation\_seq', 'envgen\_survey\_seq', 'genome\_seq', 'genome\_seq\_mapping', 'genome\_seq\_quality', 'genome\_seq\_RAD', 'hematology', 'heterozygosity\_micro', 'histology', 'histonemodification\_seq', 'histonemodification\_seq\_ex', 'ICPL\_analysis', 'investigation', 'metaboliteprofiling\_ms', 'metaboliteprofiling\_nmr', 'metagenome\_seq', and 'ppi\_detection\_micro'. The main area displays 'fields' for 'Sample Name', 'Protocol REF', 'Extract Name', 'Protocol REF', 'Parameter Value[library layout]', 'Parameter Value[mid]', 'Protocol REF', 'Parameter Value[sequencing i', 'Parameter Value[base caller]', 'Parameter Value[quality score]', 'Assay Name', 'Comment[Export]', 'Raw Data File', 'Protocol REF', 'Normalization Name', 'Data Transformation Name', 'Derived Data File', and 'factors'. The 'field definition' panel on the right shows details for 'Sample Name', including its description, datatype (String), default value, and behavioural attributes (Required, Accepts file locations, Force ontology selection, Allow multiple instances, hidden?).

**Protocoles existant**  
Métabolomique, Protéomique, Génomique  
**Protocoles sur mesure**  
Bio-informatique, Protéomique ICPL, ...

# Un début de structuration e-Science?

**GALAXY by GenOuest**  
pour analyser et partager les  
données

## Informations

26325 jobs

136 utilisateurs

Plus de 800 outils

-NGS

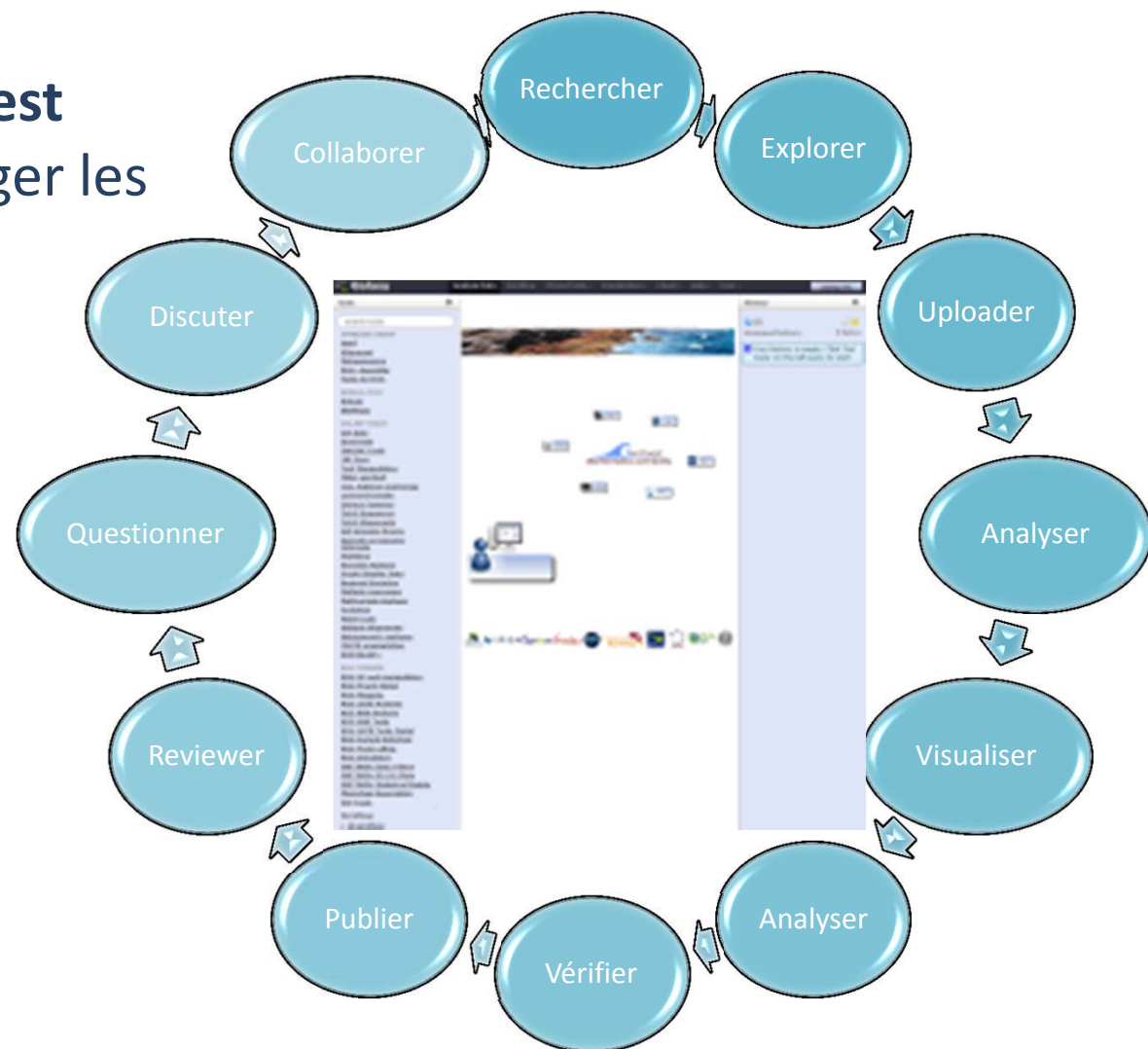
-Génomique des populations

-Génétique quantitative

-Protéomique

-Nombreux utilitaires

-...



# I have a dream...

- Pour les scientifiques en sciences de la vie et environnement
  - **Optimiser** son temps (programmation vs compréhension)
  - **Préserver** (données et processus analytiques)
  - **Accéder, partager et visualiser** de n'importe où
  - Une aide à la **gestion** de projet ...
- Pour les développeurs
  - Améliorer l'utilisation des algo et outils : Bioinfo **Recherche** ➡ **Service**
  - Accélérer la mise en production
- Pour la gestion des infrastructures
  - **Optimiser** l'utilisation (stockage, calcul et réseaux...)
- Infrastructure **pour** la donnée ➡ infrastructure **de** donnée

# LE RADSEQ

---

Voir la présentation de Karim Gharbi du 30/01/2014  
edinburgh genomics / University of Edinburgh



# Buts

THE UNIVERSITY of EDINBURGH

**edinburgh  
genomics.**

## **RAD sequencing: next-generation tools for an old problem**



**Karim Gharbi**  
Karim.Gharbi@ed.ac.uk



**RAD Workshop (Rennes)**

**K Gharbi 30/01/14**

**I**

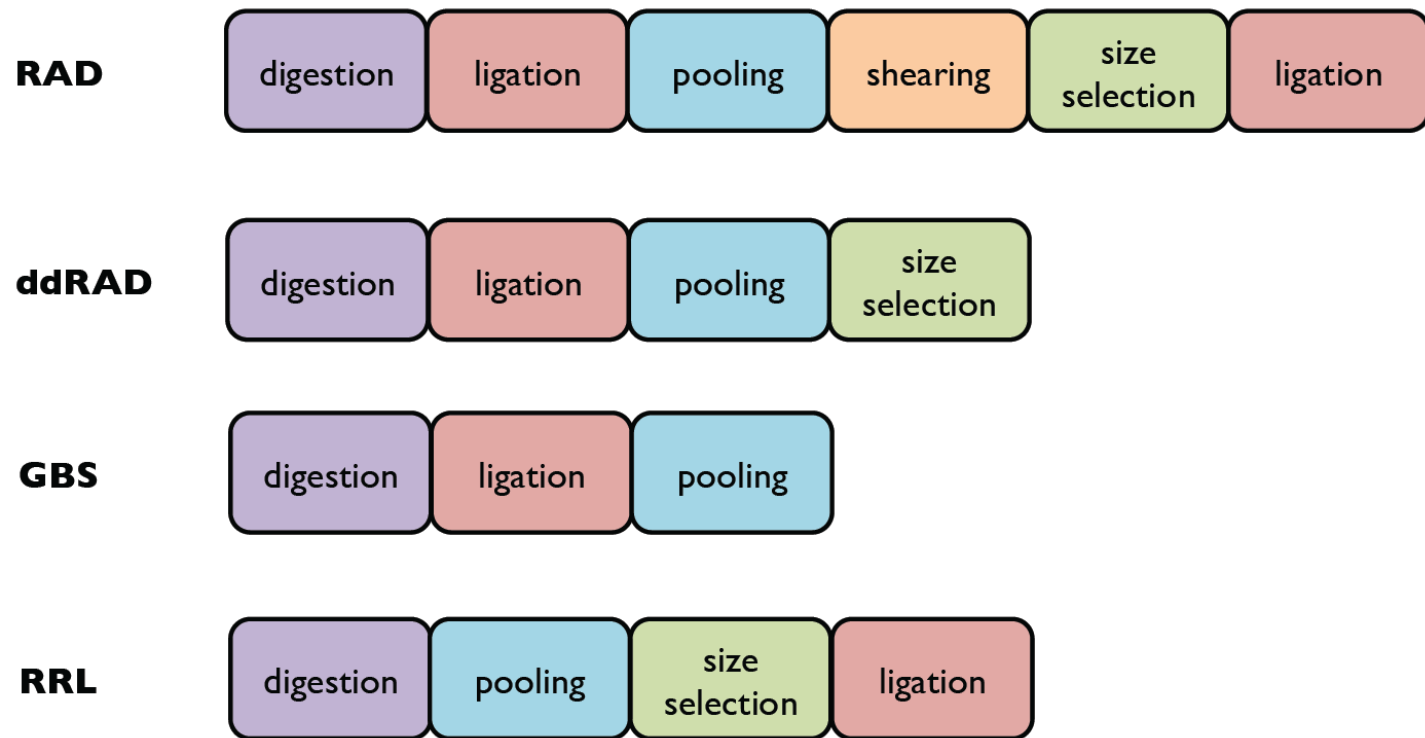
# L'arrivée du NGS

- Allozymes, RAPD, AFLP, Microsatellites, Puces SNPs... NGS
- Baisse du coût de séquençage ... pas de l'analyse ;)
  - Le séquençage de beaucoup d'individus + bcp de marqueurs reste onéreux
  - Besoins de sous-échantillonner le génome



# Réduction de génome

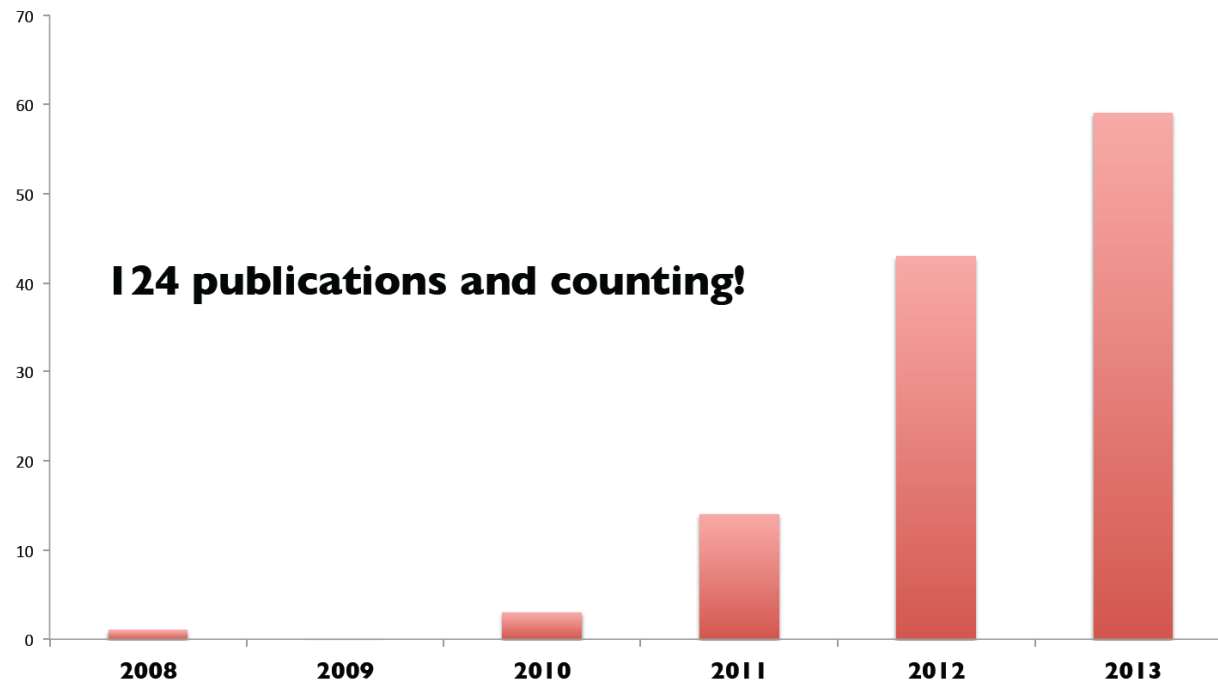
- Fragmentation de l'ADN (**étape très couteuse!**)
- Sélection des fragments par taille
- Ligation d'adaptateurs pour le séquençage



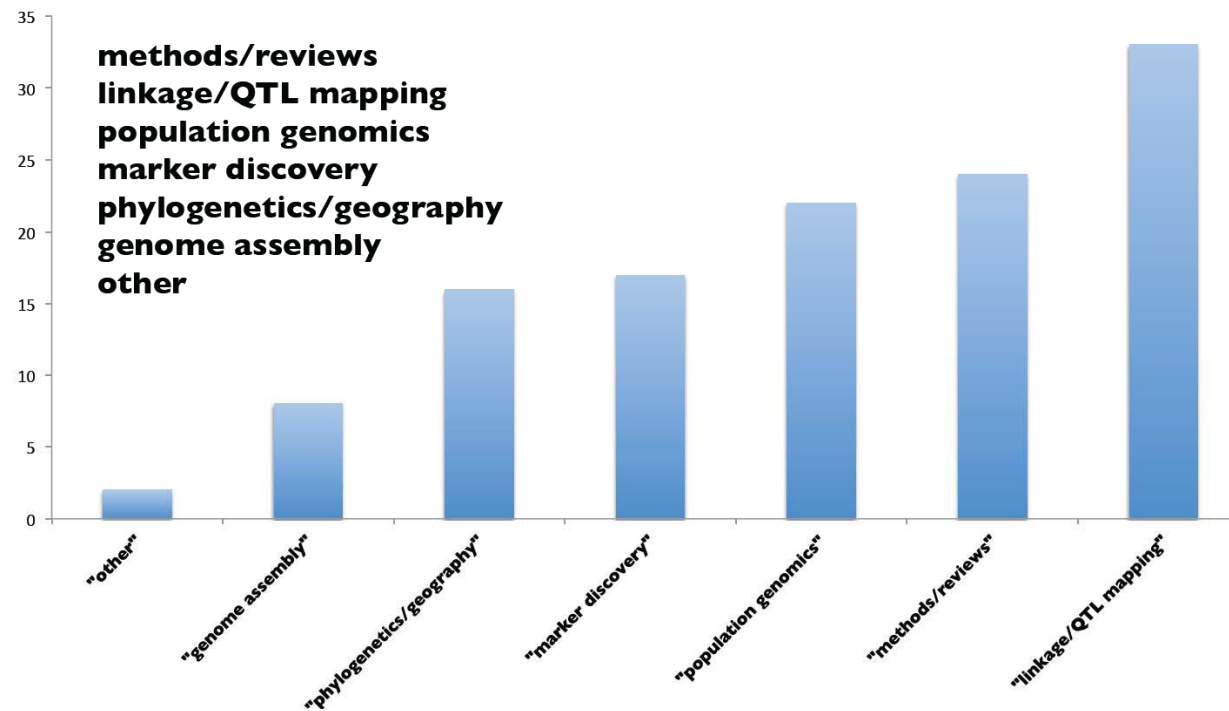
Utilisé au Roslin institute pour faire puce SNP

# Réduction de génome

- Publi initiale : Eric Johnsson, 2008
- Article d'Hohenlohe dans les 5 premiers



# Application



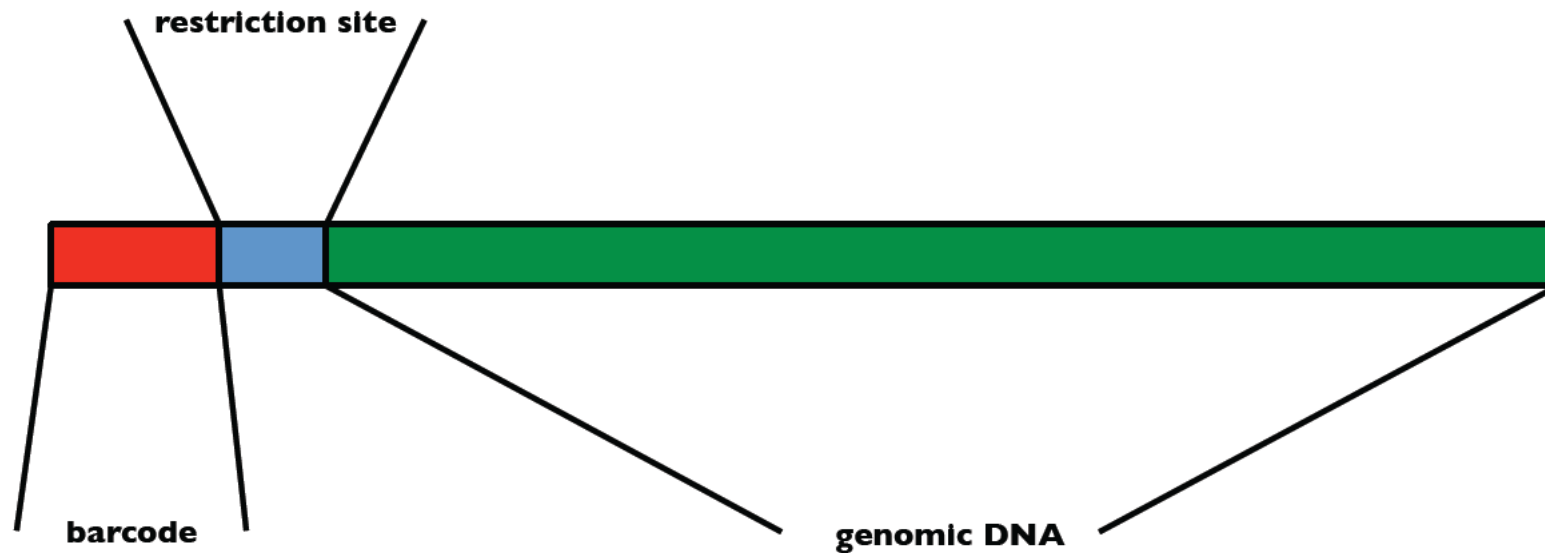
# Single-end RAD



 *restriction site*  
 *Illumina read (30-300 bp)*

**1 restriction site = 2 RAD tags**

# Single-end RAD



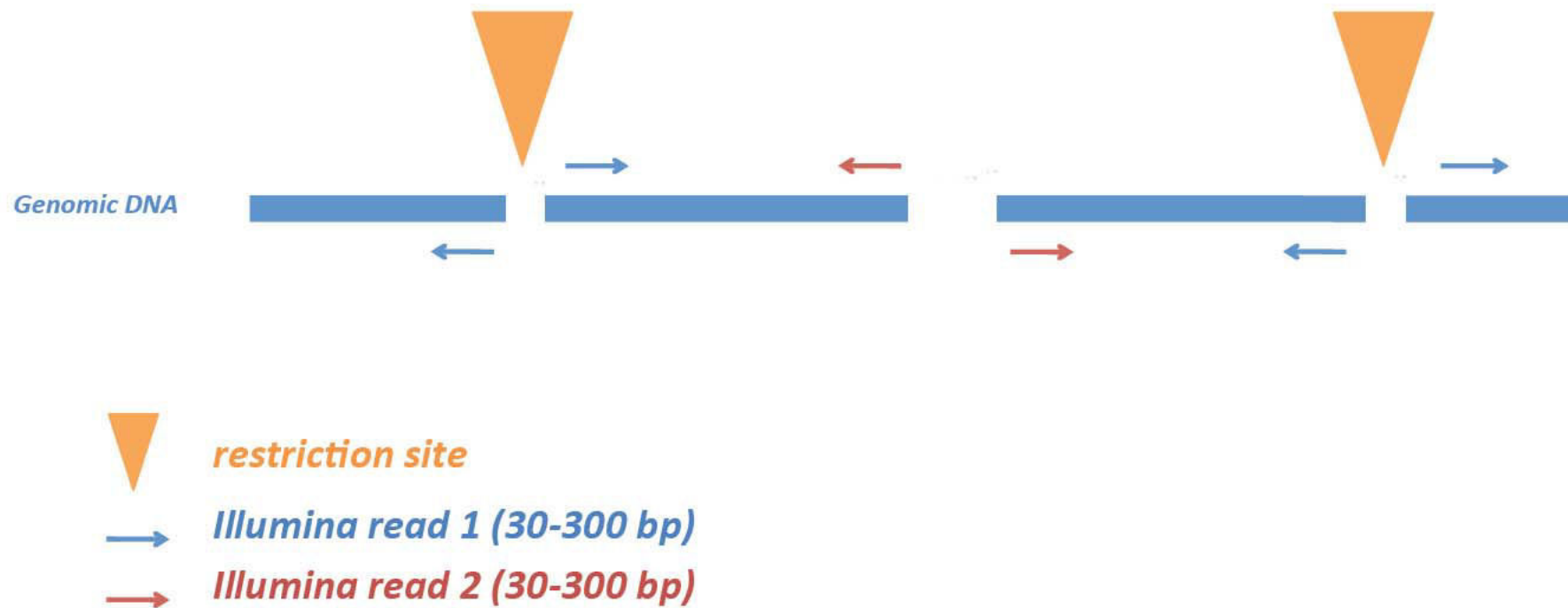
@M00689:44:000000000-A1N97:1:1101:11642:2590 1:N:0:1

CTGATGCTTGCAGGACGCACCTCCCCGCGGTGCGGCTAATGTCCTCGCAGC

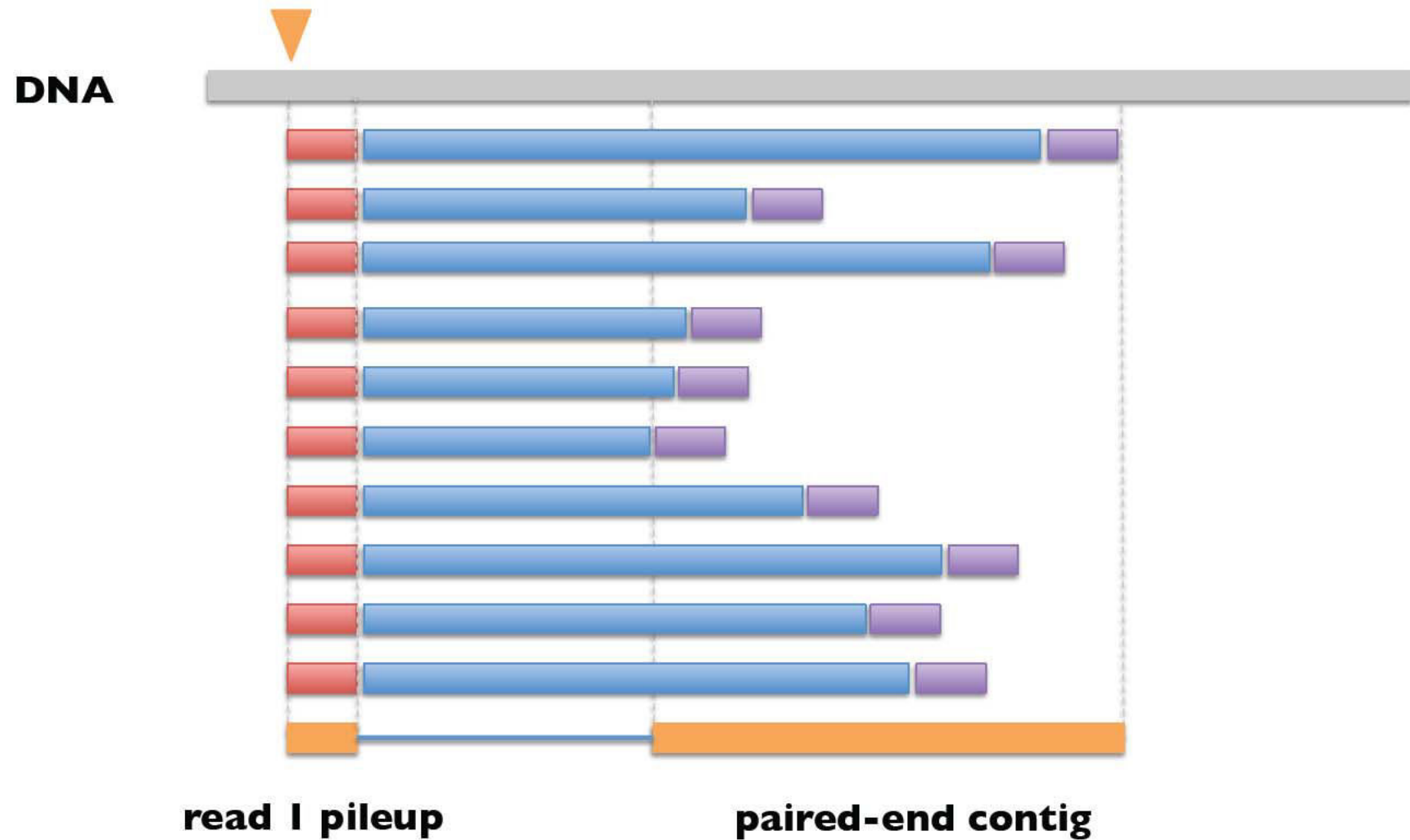
+

AAAAABBBDDDDDDDDGGGGGGGIIHHHHHEHHHHHBHHIIIIIIHHH@E

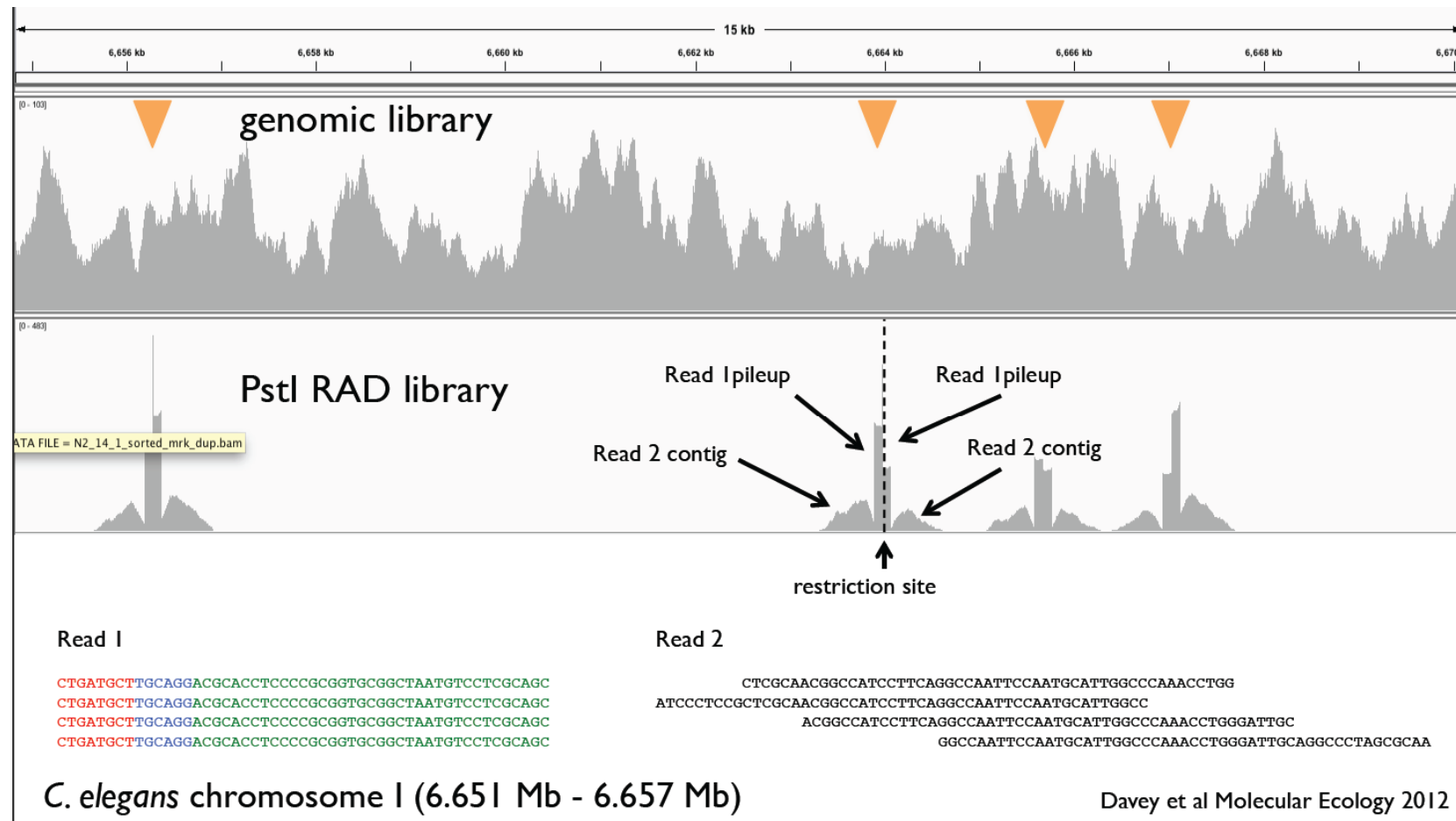
# Paired-end RAD



# Paired-end RAD



# Paired-end RAD



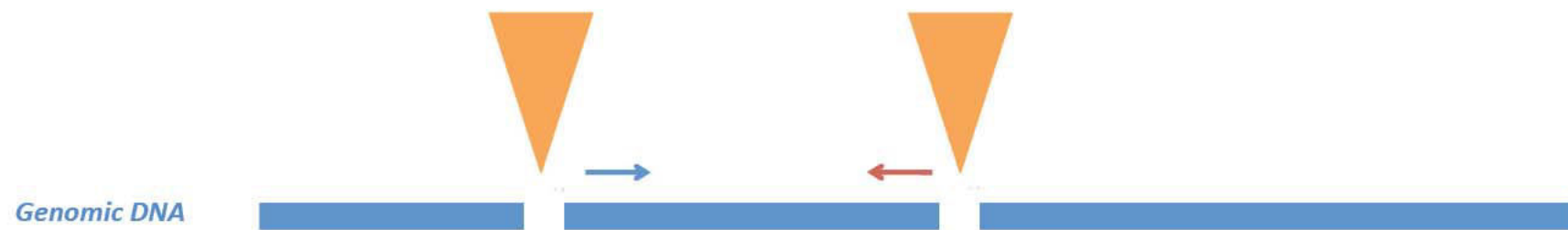


# Single vs Paired-end RAD



|                                     | single-end | paired-end    |
|-------------------------------------|------------|---------------|
| <b>Library preparation</b>          | =          | =             |
| <b>Costs</b>                        | +          | ++            |
| <b>Bioinformatics</b>               | +          | ++            |
| <b>Bases per tag</b>                | up to 300  | up to 300-500 |
| <b>Design of genotyping assays</b>  | limited    | good          |
| <b>Filtering of duplicate reads</b> | no         | yes           |
| <b>Paralog resolution</b>           | +          | ++            |

# ddRAD



*restriction site*

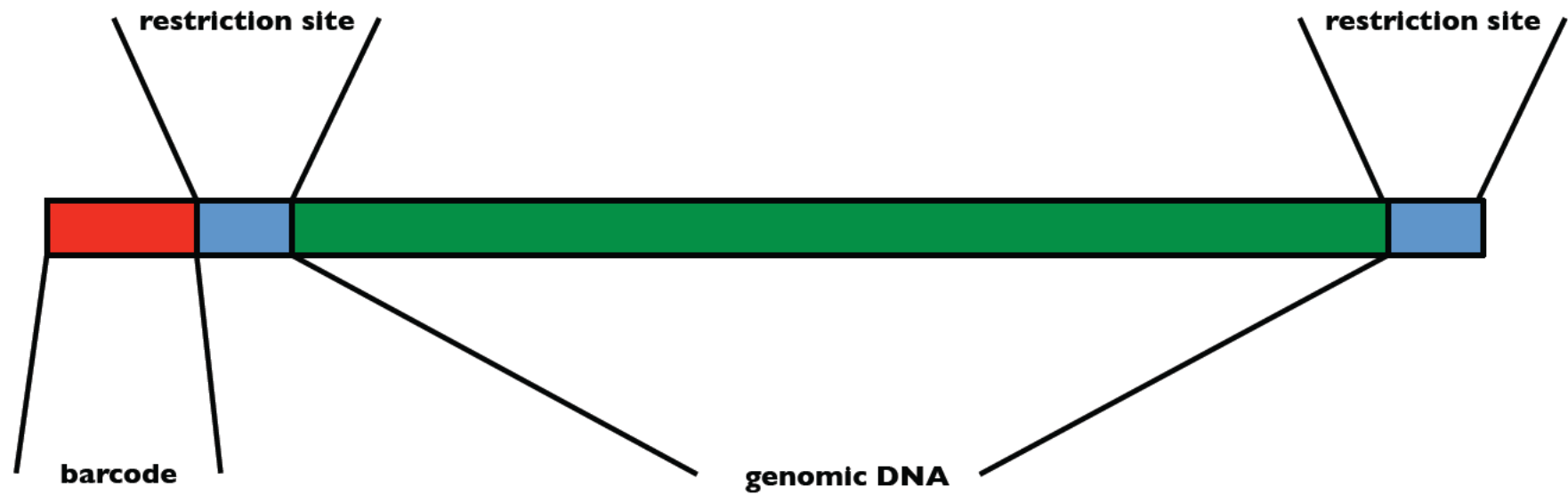


*Illumina read 1 (30-250 bp)*

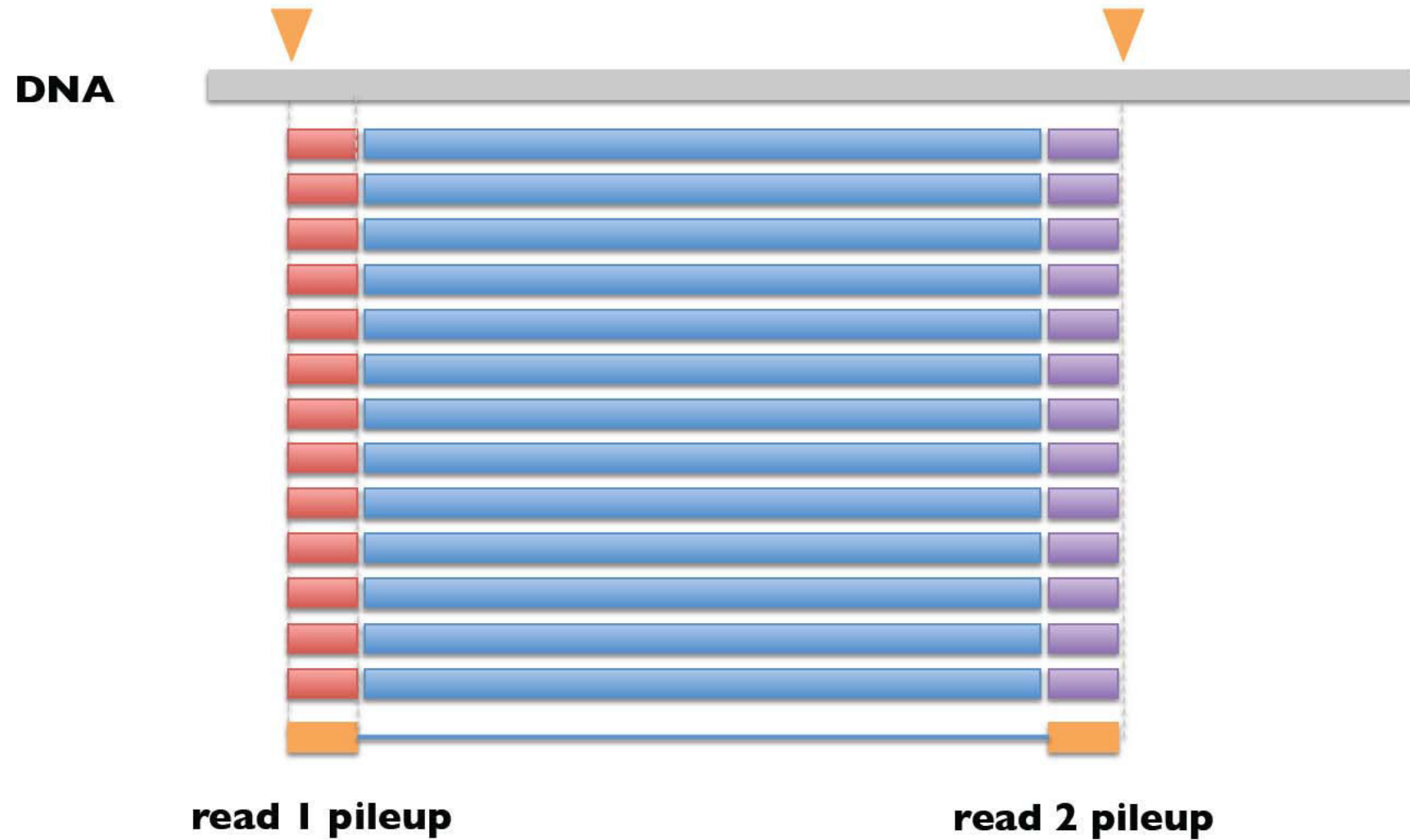


*Illumina read 2 (30-250 bp)*

# ddRAD



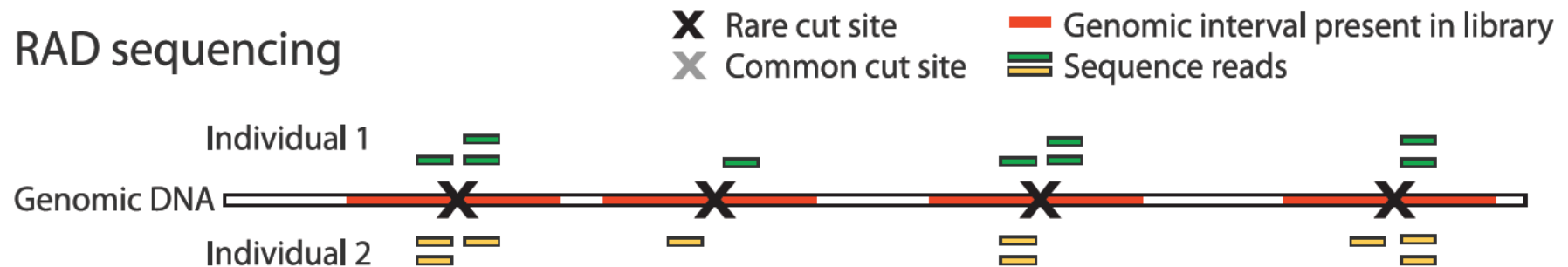
# Paired-end ddRAD



# RAD vs ddRAD

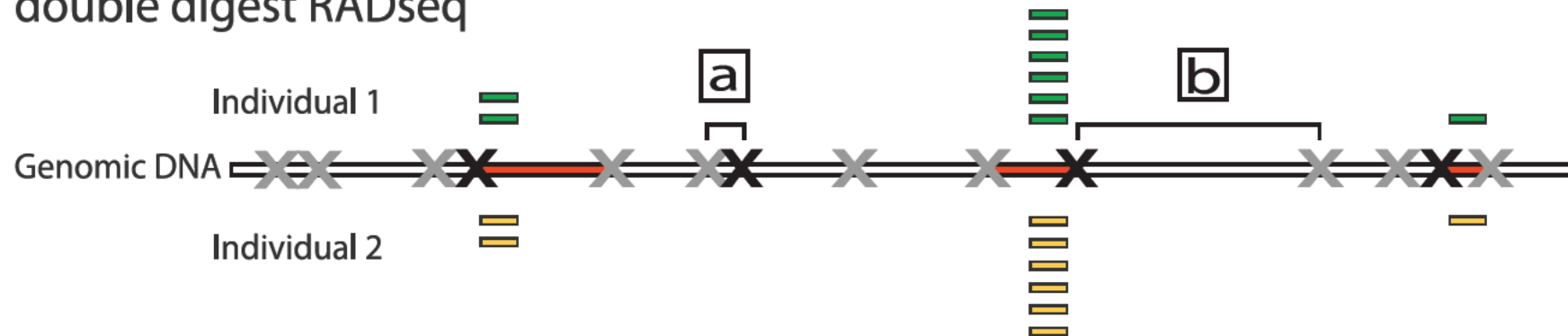
A

RAD sequencing

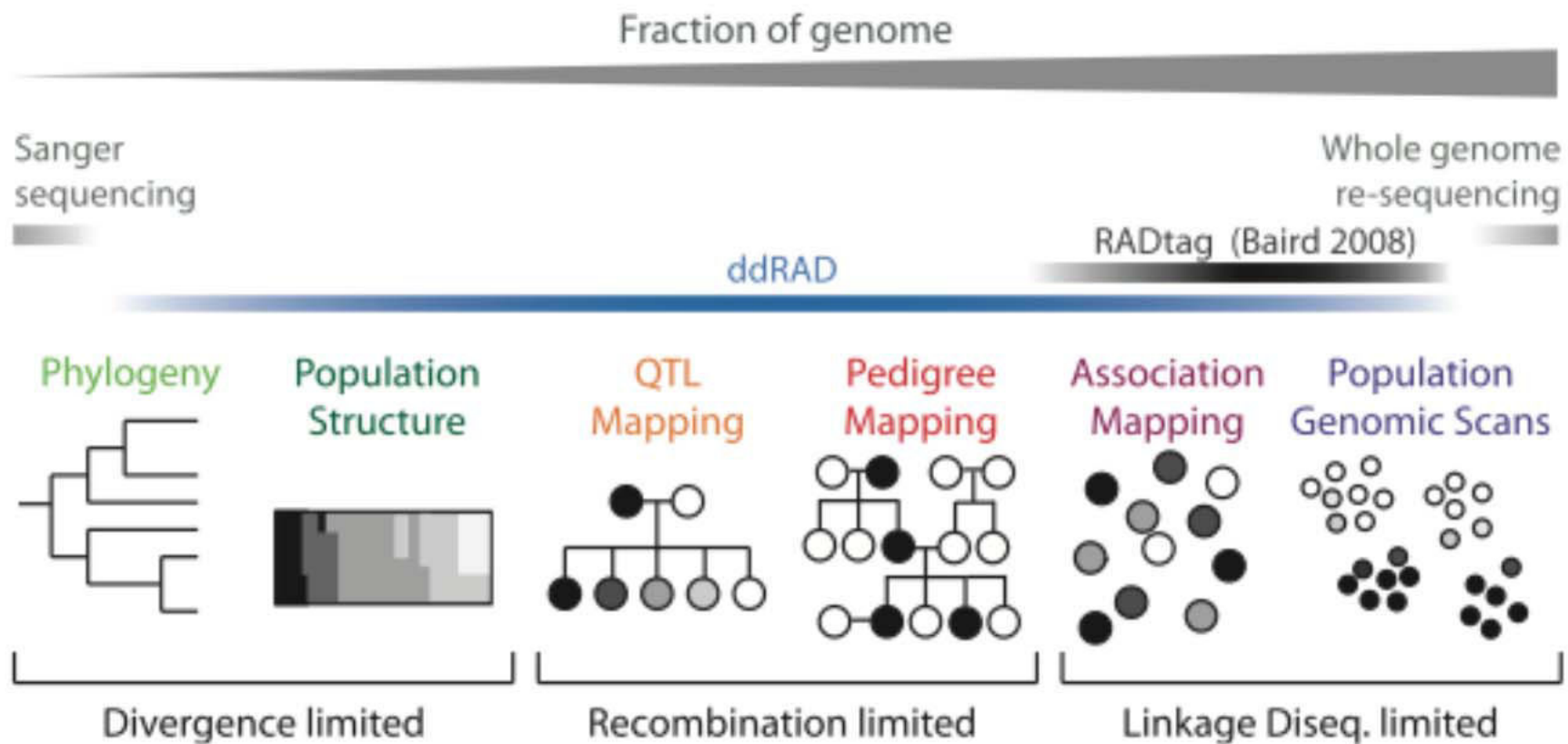


B

double digest RADseq



# Ne rester pas en RAD, ni à sec....



# Retours

- Paired-end / single
  - Préparation de banque identique
  - SE : 300pb / PE 300pb + 300 à 500pb
  - Coût très supérieur en PE
  - Single end limité pour trouvé SNP et faire puce à ADN
  - Plus il y a de séquences, plus il est facile de filtrer paralogues....
  - Filtrage de duplication de lectures possible en PE, pas en SE
- RAD simple / ddRAD
  - RAD simple : séquençage entre site de restriction et cassure par sonication
  - ddRAD : séquençage entre 2 sites de restriction donc plus de flexibilité sur les fragments générés, répartirions des lectures
  - En ddRAD, le tout adaptateur + barcode + site de restriction + ADN + adaptateur ~500pb

# Retours

- Biais connus
  - Pb analyse image quand même nucléotide dans toute les séquences à la même position. Surtout sur HiSeq, moins sur MiSeq
  - Mieux vaut utiliser plein de barcodes différents .... Reste pb du site de restriction! Il vaut mieux alors également mélanger les expériences!
- infos
  - 250 ng ADN nécessaire, 1 µg demandé à edinburgh genomics
  - Il faut une qualité au top! Pas d'ADN dégradé. Cela peut induire de ne pas utiliser 30 à 40% des données
  - PCR 12 à 14 cycles
  - Profondeur différente en fonction des tailles de fragments!
  - Peut poser pb surtout si mutation au niveau d'un site de restriction chez certains individus
  - Licence / Brevet QiaGen?
  - Voir *Davey et al, Molecular Ecology (2012)*



# LE RADSEQ SOUS GALAXY

---

Utilisation du pipeline Stacks

# Programme

- Détection de SNP
  - Etude des parents d'une famille
- Cartographie Génétique
  - Etude d'une famille avec 93 descendants
- Construction de mini-contig
  - Données pairées
- Génomique des population
  - Sans génome de référence
  - Avec génome de référence

# Buts

- Se familiariser avec l'utilisation de données NGS à partir de librairies réduites (RRL)
- S'initier à l'utilisation
  - de Galaxy
  - du pipeline STACKS
- Apprendre
  - Préparation de données brutes Illumina RAD
  - Alignement de lectures sur un génome de référence
  - Assembler des loci RAD
  - Détecter des SNP, déterminer génotypes et haplotypes
  - Calculer des statistiques en génétique des populations

# Jeux de données et outils

- Ceux proposés par Julian dans ces formations
- Jeux de données épinoche de l'article d'*Hohenlohe et al. 2010*
- Nettoyage et l'analyse des données via Galaxy, le pipeline STACKS et BWA.
- Les jeux de données seront tous produits via des séquenceurs de type Illumina GAII ou HiSeq2000.
- Les logiciels sont tous open source

# Merci de votre attention

La plate-forme Bio-informatique GenOuest



Le groupe Symbiose IRISA/INRIA  
*GenOuest-Dyliss-Genscale*



eBGO HUB (*collaboration*)

<http://www.e-biogenouest.org/>

EMME portal (*data management*) <http://emme.genouest.org/>

Galaxy instance (*data analysis*) <http://galaxy.genouest.org/>

GO4Bioinformatics (*education*) <http://go4bioinformatics.genouest.org/>

Cyril Monjeaud



Olivier Collin

