Paired-end versus *mate-pair*

Auteurs : Ségolène CABOCHE[1], Gaël EVEN[2]
[1] Docteur en informatique, Université de Lille2
[2] Ingénieur en bioinformatique, Gènes Diffusion

Version : 1.0

Dernière modification : 17/02/2012

Table des matières

- Genèse de la confusion : paired-end versus mate-pair
- Mate-pair
  - Généralités
  - Focus sur les technologies mate-pair
    - Illumina, Roche et SOLiD
    - Iontorrent : processus de préparation d'une librairie mate-pair
  - Récapitulatif - mate-pair
- Paired-end
  - Généralités
  - Focus sur les technologies paired-end
    - Illumina
    - SOLiD
    - Roche 454
    - IonTorrent
  - Paired-end : utilisation
    - Spécificité du de novo
  - Récapitulatif - paired-end
- Focus sur les logiciels
  - Exemple de solution : BOWTIE
  - Cas des reads chevauchants
- Résumé de l'étude
- Bibliographie
- Notes techniques

Genèse de la confusion : *paired-end* versus *mate-pair*

Il existe sur de nombreux supports (brochures techniques, publications, forums, …) une confusion entre les termes ***paired-end*** et ***mate-pair****.* Aucune définition claire ne semble être disponible pour ces deux approches.
En réalité, ils désignent des objets biologiques bien différents ; le terme *mate-pair* fait référence à une façon de préparer la librairie (protocole de *circularisation de l'ADN*) et génère une paire de reads éloignés de plusieurs kb (de 2 à 20kb), tandis que le *paired-end* est un type de séquençage qui génère une paire de reads éloignés de centaines de bp au maximum (< 600bp).
Si la technologie le permet, il est en théorie possible de préparer une librairie *mate-pair* et de choisir de séquencer en *paired-end* ou en *single-end*.

Il est aussi à noter que l'orientation des reads en *mate-pair* et *paired-end* est différente et varie selon la technologie de séquencage utilisée. Par exemple avec Illumina l'orientation est reverse-forward en *mate-pair* et forward-reverse en *paired-end* (voir tableau dans le focus sur les technologies).

La confusion entre les deux termes vient notamment du fait que les méthodes d'analyses bioinformatiques semblent assez similaires entre les analyses *paired-end* et *mate-pair* En effet, la taille de l'insert n'a pas beaucoup d'influence au niveau analytique et algorithmique.
Par exemple, le logiciel SOPRA [6] ne fait volontairement pas la distinction entre *paired-end* et *mate-pair* . D'autre logiciels nécessitent une "réorientation" des reads avant alignement, à l'instar de BWA.
D'autre part, la librairie *mate-pair* chez Illumina est synonyme de la librairie Roche *long paired-end library* (LPE), ce qui renforce la confusion.

Dans ce document nous nous attacherons à présenter les deux approches, leurs spécificités par rapport aux technologies de séquençage deuxième génération, leurs utilisations (particulièrement pour le *paired-end*) et la prise en compte dans les logiciels. Nous terminerons par une synthèse qui permettra de comprendre pourquoi il est important de rester prudent quant aux termes utilisés dans les textes.

Mate-pair

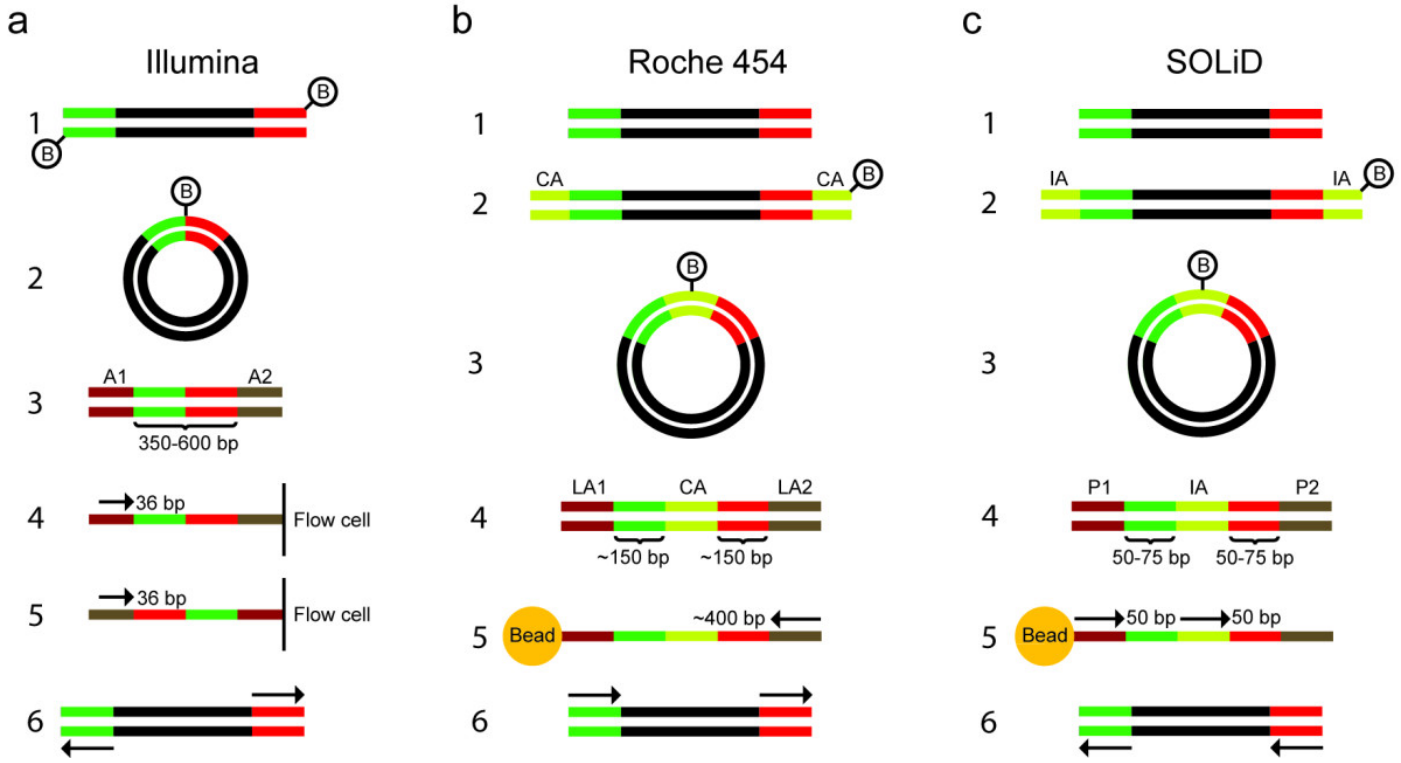
Généralités

Dans les librairies *mate-pair*, le but est de séquencer deux fragments, c'est-à-dire deux reads, séparés par une distance fixe. Pour ce faire, un fragment d'ADN est circularisé, ainsi les deux extrémités deviennent adjacentes et pourront être séquencées.

Focus sur les technologies *mate-pair*

Illumina, Roche et SOLiD

Il existe des spécificités pour chacune des technologies de séquençage de deuxième génération; le schéma sur la page suivante détaille le principe de construction de librairies *mate-pair* pour les trois technologies [8].



Principes de construction des librairies *mate-pair*: (a) Préparation des librairies *mate-pair* Illumina. Des nucléotides biotinylés sont ajoutés à l'extrémité des fragments (1). Après circularisation, les deux extrémités des fragments (en rouge et en vert) deviennent adjacentes (2). L'ADN circularisé est fragmenté et les fragments biotinylés sont purifiés par capture par affinité. Les adaptateurs de séquençage (A1 et A2) sont liés aux extrémités des fragments purifiés (3), et les fragments sont hybridés sur une *flow cell*, dans laquelle ils sont *bridge-amplifiés.* La séquence du premier read est obtenue avec l'adaptateur A2 lié à la *flow cell* (4). Le brin complémentaire est synthétisé et linéarisé avec l'adaptateur A1 lié à la *flow cell*, et la seconde séquence est obtenue (5). Les deux séquences des reads (représentés par des flèches) seront dirigées vers l'extérieur par rapport au fragment d'origine. **(b)** Préparation des librairies *paired-end* 454 de Roche (elles sont appelées *paired-end*, mais sont basées sur les mêmes principes que les librairies *mate-pair* des autres technologies). Des nucléotides non-marqués sont liés aux extrémités des fragments originaux (1) et des adaptateurs de circularisation (CA) marqués à la biotine sont liés à l'extrémité des fragments (2). Après circularisation (3), fragmentation et purification par affinité, les adaptateurs de la librairie (LA1 et LA2) sont liés à l'extrémité des nouveaux fragments (4) puis les fragments sont amplifiés sur bille par PCR par émulsion. La séquence d'un seul read qui couvre les deux extrémités d'origine et l'adaptateur interne est générée (5). La séquence de l'adaptateur est enlevée *in silico* et la séquence est séparée en deux reads qui ont tous les deux la même orientation (6). **(c)** Préparation de la librairie *mate-pair* de SOLiD. Les étapes 1 à 4 sont analogues à celles des librairies *paired-end* 454 de Roche, avec un adaptateur interne (IA) marqué à la biotine et deux adaptateurs de séquençage (P1 et P2). Le séquençage est effectué avec deux primers différents, respectivement complémentaires à l'adaptateur P1 et à l'adaptateur interne (5). Les reads résultants ont la même orientation.

Iontorrent : processus de préparation d'une librairie *mate-pair*

Comme pour les autres technologies, une librairie *mate-pair* sous Intorrent consiste en une paire de fragments d'ADN qui sont couplés car ils proviennent au départ du séquençage de deux extrémités d'un même fragment d'ADN génomique. Les deux extrémités sont connectées grâce à un adaptateur interne (MP Adaptor) pour former un cercle.

Pour une librairie *long mate-pair,* les deux figures ci-dessous illustrent le protocole de circularisation (Figure 2) et l'amplication grâce à des primers spécifiques des adaptateurs utilisés (Ion Adaptors) (Figure 4).
A noter que l'orientation des adaptateurs importe peu.

Figure 2 Mate-paired library preparation workflow.

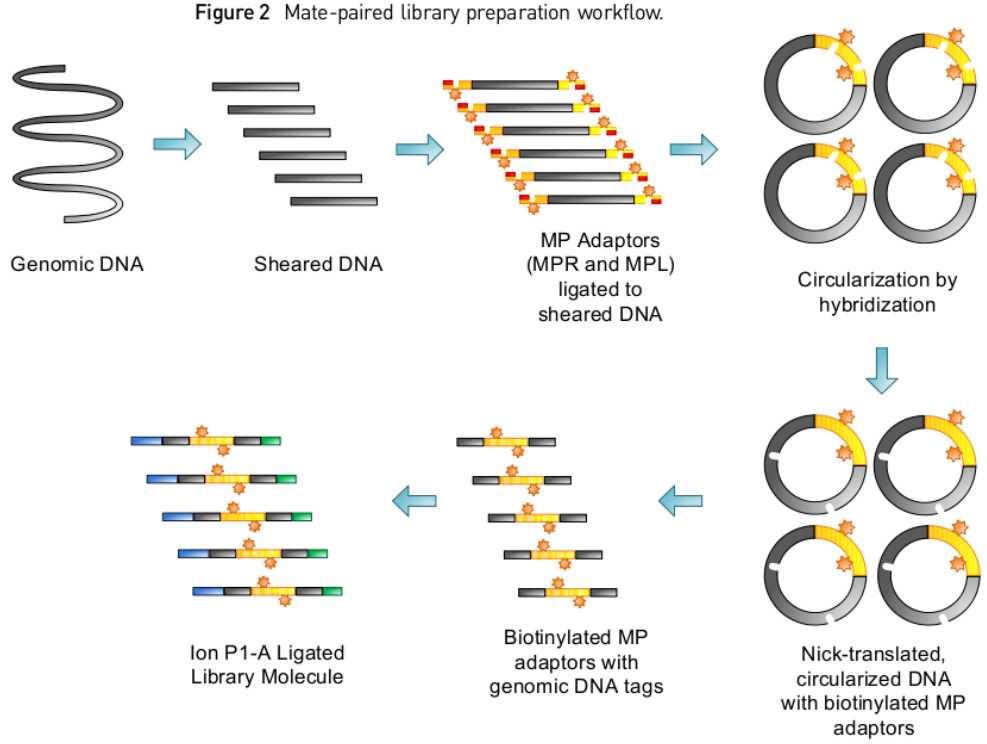
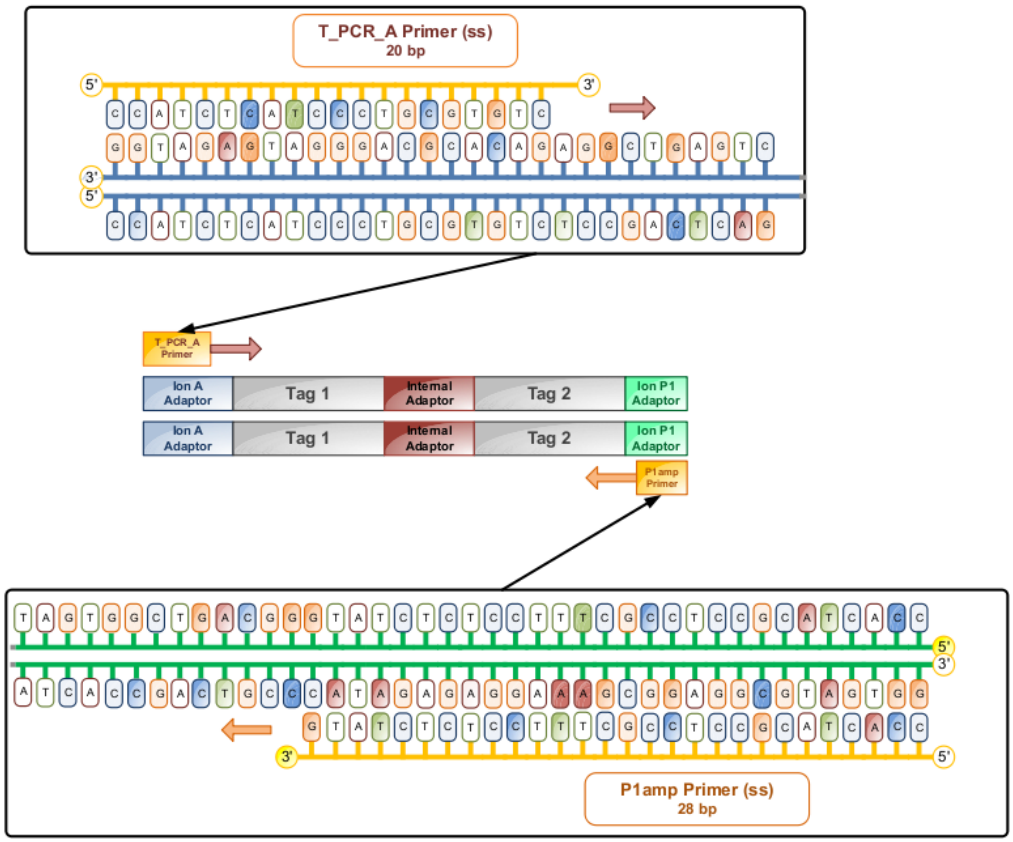


Figure 4 Mate-paired library amplification design.



Principe de préparation de la librairie mate pair avec ionTorrent issue du site officiel

Récapitulatif - *mate-pair*

MATE-PAIR	***Orientations des reads***	***Tailles d'inserts annoncées***
Illumina	*Reverse-Forward*	***2 - 5 kb [9]***
SOLID	*Forward-Forward*	***0.6 - 6kb***
Roche 454	*Forward-Forward*	***3 - 20kb [9]***
Ion Torrent	*Reverse-Forward*	***2 - 10kb***

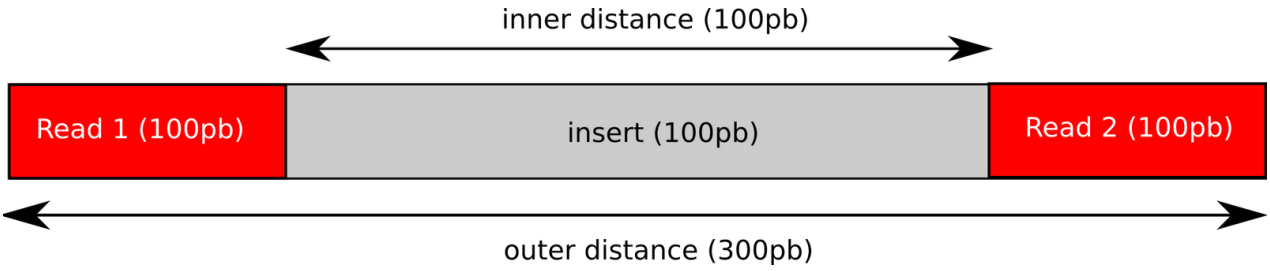
Orientations des reads et tailles d'inserts autorisés pour les principales technologies de séquençage seconde génération

Paired-end

Généralités

Dans le cas général, le *paired-end* consiste à séquencer les reads par paires, ces reads étant séparés par une distance connue (*insert-size* - bp). On distingue deux types de distances :

- la distance **inner** est la distance qui sépare les deux reads, sans inclure la taille des reads dans le calcul de la distance.
Par exemple, pour des runs paired-end avec des fragments de 300bp, où les deux reads ont une taille de 100pb, la distance inner sera de 100pb.



- la distance **outer** inclue la taille des reads dans le calcul de la distance.

Si la distance **outer** est égale à la taille des reads (par exemple 100bp) alors la paire de reads sera en chevauchement **total.**

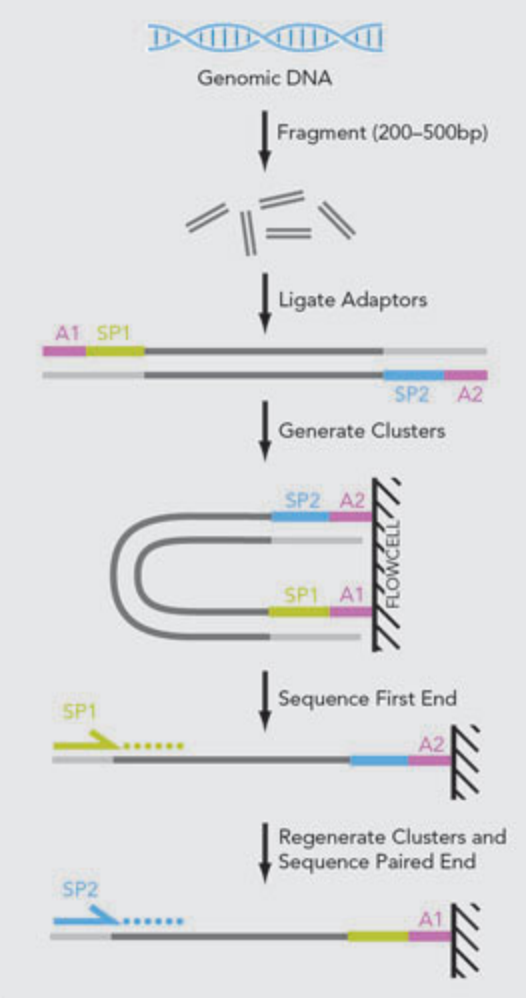
Dans la littérature, la grande majorité des études créent du *paired-end* non chevauchant avec une distance fixe (par ex : 100bp avec Illumina).
On trouve des exceptions avec des *paired-end* chevauchants en Illumina [3] : si la taille de la librairie est inférieure à la somme des tailles des reads forward et reverse séquencés alors on obtiendra des séquences chevauchantes.

Toutes ces études *paired-end* concernent la génération de petits reads. Récemment IonTorrent affirme avoir adapté la technologie *paired-end* pour les reads longs [7].

Focus sur les technologies *paired-end*

Illumina

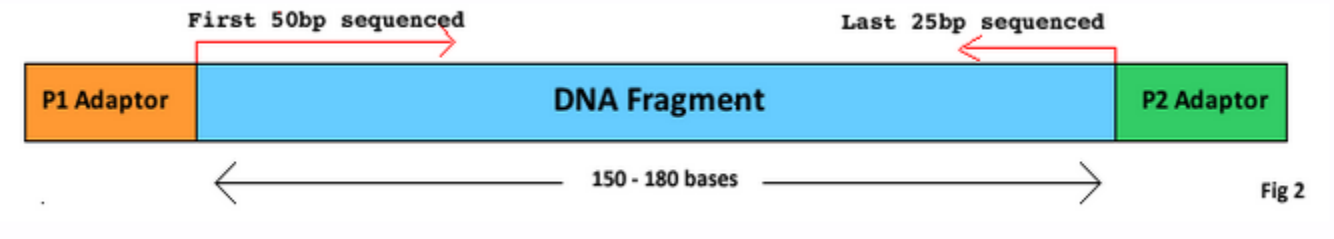
Une petite modification à la préparation de la librairie d'ADN *single-read* permet la lecture sur le brin forward et le brin reverse du fragment d'ADN (voir figure ci-dessous).
Le protocole permet à l'utilisateur de choisir la taille d'insertion (insert size) de 200 à 500bp.
Un run classique en Illumina *paired-end* permet d'obtenir des reads de 2 x 75bp pour une quantité d'environ 200 millions.



Protocole de séquençage *paired-end* avec Illumina
(source:http://www.illumina.com/technology/paired_end_sequencing_assay.ilmn**)**[11]

SOLiD

Une librairie *paired-end* SOLiD contient une insertion (i*nsert size*) d'environ 150-180 bp flanquée de deux adaptateurs (P1 et P2 - voir figure ci-dessous).
Les 50 premières bases, à partir de l'adaptateur P1 et les 35 dernières bases, à partir de l'adaptateur P2, sont séquencées.



Protocole de séquençage *paired-end* avec SOLiD
(source: http://nextgenseq.onc.jhmi.edu/nextgenseq/technology.html)[12]

Roche 454

Il n'existe pas de protocole *paired-end* en 454, le protocole *long reads paired-end* correspond a un protocole *mate-pair* classique.

IonTorrent

Paired-end sur des reads longs :

Courant octobre 2011, Ion Torrent annonce la sortie d'un kit de séquençage permettant de faire du *paired-end* sur des reads longs (500bp). Cette prise en compte des reads longs le différencie d'Illumina qui travaille sur des reads courts.
A noter qu'il est possible de générer des *paired-end* avec chevauchement complet, ce qui maximise la qualité de la séquence générée. Par déduction, la détection de zones répétées devrait être facilitée avec des reads longs.

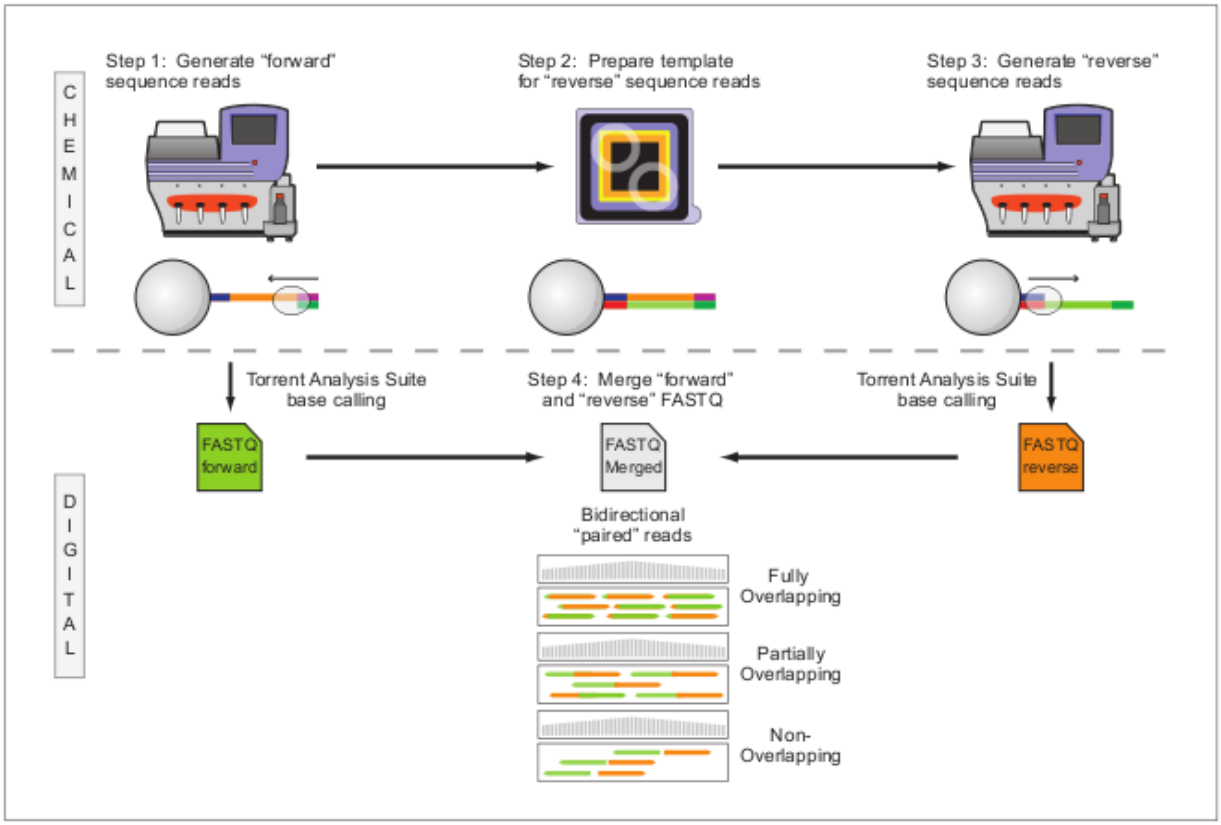


Figure 3. Only Ion Torrent Personal Genome Machine™ (PGM™) semiconductor sequencing technology delivers the unique combination of long reads and PES. This figure depicts the chemical and digital workflow of the novel PES method developed by Ion Torrent. Forward and reverse reads are denoted by green and orange colors, respectively. The degree of paired-read overlap is shown aligned to a reference genome.

Paired-end IonTorrent - issue de www.iontorrent.com [7]

Paired-end : utilisation

Nous présentons ici les spécifificités de l'utilisation du *paired-end* chevauchant et du *paired-end* non chevauchant.

Le *paired-end* non chevauchant permet de repérer les variations structurelles (insertion, délétion, translocation...) en comparant la distance observée entre la paire de reads et la distance théorique. [4][5]

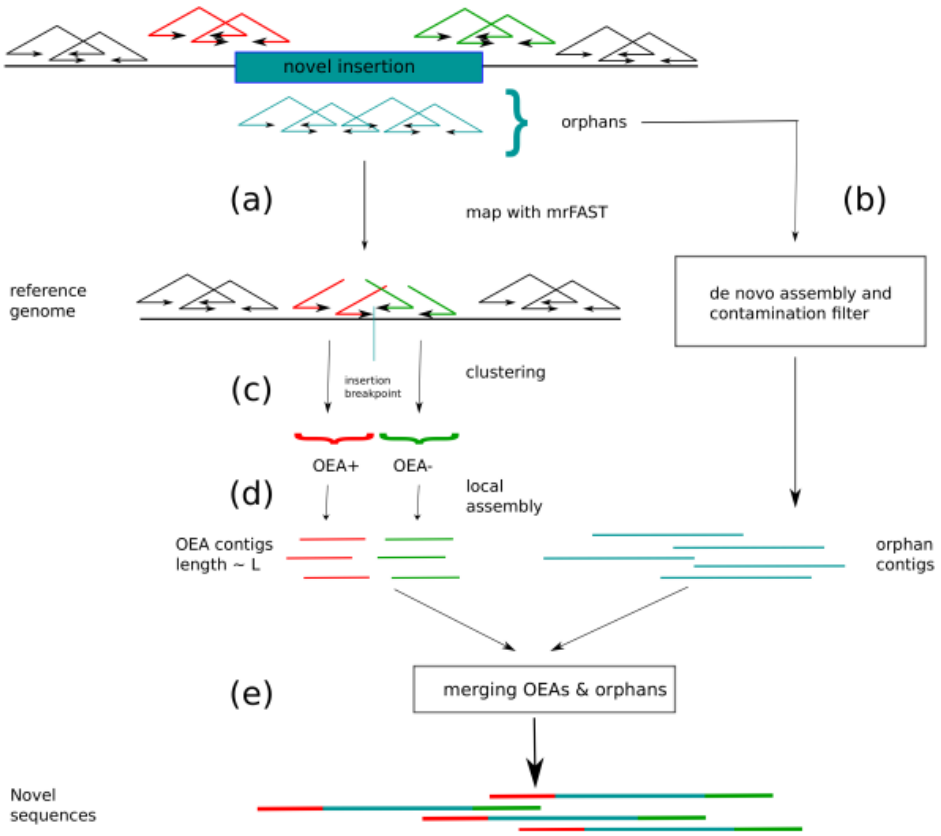


Fig. 1. The overall approach of the NovelSeq pipeline. (a) We start by mapping the paired-end reads to the reference genome and then classify the paired-end reads as OEA and orphan reads. (b) We then assemble the orphan paired-end reads using available *de novo* assembly algorithms and screen the contigs for possible contamination. (c) We cluster the OEA reads into groups and nd the insertion locus supported by each OEA cluster. (d) We assemble the unmapped end-read in each OEA cluster (the OEA reads with different orientation of mapping should be assembled independently). (e) Finally, we merge the orphan contigs and OEA contigs to anchor the orphan contigs to the reference genome.

Figure 1 : exemple de pipeline de détection d'insertion grâce au *paired-end* non-chevauchant [5]

Il permet également de détecter et de corriger les erreurs d'alignement en comparant ces mêmes distances [2] .
Dans l'article [2], les auteurs démontrent que le *paired-end* non-chevauchant améliore la **sensibilité** et la **spécificité** de l'alignement par rapport au *single-end*. De même, le gain de détection de **SNP est significatif** en *paired-end*. Ceci est dû au fait que les zones répétées sont mieux repérées, le nombre de matériel présent augmente (plus de reads uniques) et les

zones avec de trop nombreuses erreurs d'alignement sont évitées (alignements invalides).

Le *paired-end* est également utile pour la prise en compte des zones répétées. Lorsqu'un read est positionné dans une région répétée il est difficile de replacer ce read sur le génome de référence ou en assemblage *de novo*. En séquençage sur référence comme en *de novo*, en observant la distance théorique (*insert-size)* entre les paires de reads, il est possible d'estimer l'éloignement attendu entre les reads dans l'assemblage final (*de novo*) / l'alignement (sur référence). On considère qu'au moins un read peut être mappé à une positon unique, ainsi il est possible d'assigner une position approximative du read associé, même si celle-ci se situe dans une zone répétée. Toutefois, il semble difficile de repérer des zones répétées quand la taille du read est plus petite que la région répétée [1].
Si on combine l'étude de la profondeur moyenne et l'observation de ces paires de reads, on peut obtenir un bon assemblage/alignement au niveau des régions répétées du génome.[1]

Le *paired-end* chevauchant permet de réduire les erreurs dues au séquençage : l'information de la séquence chevauchante est de meilleure qualité [3].
Dans le cas de short reads Illumina [3] (mais on peut également généraliser aux autres technologies) : pour ces paires de reads chevauchants l'identification correcte des adaptateurs est facilitée en maximisant l'autocorrélation entre les 2 reads par le fait qu'il est nécessaire d'avoir des positions identiques pour les débuts d'adaptateurs. Cette statégie de chevauchement est donc plus puissante que les approches *single-end* pour identifier le début d'adaptateur, qui souvent supprime des bouts de séquences en fin de read matchant l'adaptateur par chance ou qui n'identifie pas la vraie séquence de l'adaptateur à cause des erreurs de séquençage en fin de read.
Dans cette même étude, **Kircher et al** [3] concluent que sur les séquences chevauchantes, on constate une réduction d'un facteur 5 du taux d'erreur ; c'est ce qu'annonce également Iontorrent [7]

Spécificité du *de novo*

Le *paired-end* non chevauchant est utilisé pour connecter les contigs entres eux. Si on repère plusieurs paires de reads présents sur deux contigs (un read sur un contig et l'autre read sur un autre contig), on peut considérer que ces reads font la jointure entre les deux contigs. Ce principe se retrouve dans les algorithmes dits de *scaffolding* [1]. On considère tout de même que 10-20% des paires de reads servant à faire la jointure sont fausses. Il faut donc plusieurs paires de reads pour valider une jonction. A titre d'exemple, le logiciel *SOAPDenovo* considère un minimum de trois paires couvrant deux contigs pour considérer la jonction comme valide, tandis que d'autres comme *ABySS* considèrent 5 paires. [1]

Les séquences répétées peuvent mener à des erreurs de jointure des contigs ; en appliquant le même principe qu'au paragraphe précédent il est possible de détecter avec le *paired-end* une grande partie de ces "fausses" jointures.

Récapitulatif - *paired-end*

PAIRED-END	***Orientations des reads***
Illumina	*Forward-Reverse*
SOLID	*Forward-Reverse*
Roche 454	*X*
Ion Torrent	*Forward-Reverse*

PAIRED-END : Orientations des reads pour les principales technologies de séquençage seconde génération

Focus sur les logiciels

La confusion demeure entre *mate-pair* et *paired-end* dans beaucoup de solutions (cf. [6]). Cela vient du fait que, à part l'orientation et la taille de l'insert, les deux définitions sont très proches au niveau analytique.

La notion de distance est également ambiguë. Par exemple Bowtie, ne permet pas de définir une distance impliquant le chevauchement complet. CLCbio lui, ne déclare que la distance inner, ce qui ne permet pas chevauchement entre les reads.

Exemple de solution : BOWTIE

Bowtie permet d'aligner des reads *paired-end* quand les fichiers contenant les reads respectent une nomenclature précise (en utilisant les options -1 et -2).
Les alignements *paired-end* dans lesquels un read est entièrement contenu dans le read couplé sont considérés comme invalides.
La technologie IonTorrent permet de créer des reads *paired-end* avec chevauchement complet ; ce type de reads n'est pas analysable (dans l'état) avec Bowtie.

Cas des reads chevauchants

La plupart des logiciels ne gère pas le cas des reads chevauchants. Dans ce cas précis, il existe des solutions qui permettent d'effectuer un pré-traitement des reads pairés. Le principe est simple : il s'agit d'assembler les reads *paired-end* chevauchants dans un seul contig. Cela permet d'exécuter la suite de l'analyse en considérant les reads comme *single-end* (mais de meilleure qualité sur la partie chevauchante).

```
reads A 5' ================ ==================================> 3'
reads B                 3' <================================== ===========  5'
contig C 5' ================ ================================== ===========> 3'
                                <--------- "the middle" ---------->
```

Exemple de création d'un contig (C) à partir de deux reads paired-end (A et B) - logiciel STITCH

A titre d'exemple, le logicel STITCH permet de générer ce type de jeux de données (plus d'information sur https://github.com/audy/stitch)

Résumé de l'étude

- *Mate-pair* et *paired-end* sont biologiquement différents mais analytiquement proches : *mate-pair* est une façon de préparer la librairie, et *paired-end* est un type de séquençage. Ils ne sont pas différents, en effet, si on utilise la technologie Illumina, une librairie *mate-pair* nécessitant un séquençage *paired-end*.
- Le *mate-pair* grâce à ces reads séparés par une longue distance est utile pour l'assemblage d*e novo*, et pour repérer des variations structurelles sur le génome. Le *mate-pair* n'est pas utilisé fréquemment en projet de reséquencage, du fait qu'il nécessite initalement une grande quantité d'ADN [8].
- Le *paired-end* existe en short reads (majoritairement Illumina) et maintenant en long reads (Iontorrent)
- On rencontre du *paired-end* non chevauchant, partiellement chevauchant et complètement chevauchant:
    - Le chevauchement partiel ou complet permet d'améliorer la qualité de la séquence en augmentant le nombre de reads et en présentant une meilleure qualité de séquençage (augmentation du nombre de reads redondants (chevauchement complet), amélioration du séquençage des fins de reads...).
    - Le non-chevauchement est utile pour détecter des variations structurelles sur le génome et pour l'assemblage *de novo* grâce à la comparaison distance observée/distance théorique. Il est très utile pour l'assemblage des reads qui se trouvent dans des zones répétées. Il améliore également la qualité du séquençage (augmentation du nombre de reads mappés uniques). Analytiquement, on est très proche du *mate-pair*.
- Certaines subtilités comme l'orientation des reads par défaut ou la prise en compte du chevauchement demeurent dans l'utilisation des logiciels : il faut donc être prudent dans l'interprétation des mots *paired-end* et *mate-pair* et dans la définition des distances.
- Il est possible de prendre en compte les reads chevauchants en transformant les paires de reads en simple contigs assemblés.

Bibliographie

[1] Konrad Paszkiewicz and David J. Studholme
"De novo assembly of short sequence reads"
Brief Bioinform (2010) 11(5): 457-472 first published online August 19, 2010 doi:10.1093/bib/bbq020

[2]Heng Li and Nils Homer
"A survey of sequence alignment algorithms for next-generation sequencing"
Brief Bioinform (2010) 11(5): 473-483 first published online May 11, 2010 doi:10.1093/bib/bbq015

[3] Martin Kircher, Patricia Heyn and Janet Kelso
"Addressing challenges in the production and analysis of illumina sequencing data"
BMC Genomics 2011, ***12****:382 doi:10.1186/1471-2164-12-382*

[4] Ali Bashir et al.
"Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer."
PLoS Comput Biol. 2008 Apr 25 ;4(4):e1000051.

[5]Hajirasouliha I, Hormozdiari F, Alkan C, Kidd JM, Birol I, Eichler EE, Sahinalp SC.
"Detection and characterization of novel sequence insertions using paired-end next-generation sequencing."
Bioinformatics. 2010 May 15 ;26(10):1277-83. Epub 2010 Apr 12 .

[6]Adel Dayarian, Todd P Michael and Anirvan M Sengupta
"SOPRA: Scaffolding algorithm for paired reads via statistical optimization"
BMC Bioinformatics 2010, **11**:345 doi:10.1186/1471-2105-11-345

[8] Eva C Berglund, Anna Kiialainen and Ann-Christine Syvänen
"Next-generation sequencing technologies and applications for human genetic history and forensics"
Investig Genet. 2011; 2: 23.

[9] Filip Van Nieuwerburgh, Ryan C. Thompson, Jessica Ledesma, Dieter Deforce, Terry Gaasterland,Phillip Ordoukhanian, and Steven R. Head
"Illumina mate-paired DNA sequencing-library preparation using Cre-Lox recombination"
Nucl. Acids Res. first published online November 29, 2011 doi:10.1093/nar/gkr1000

Notes techniques

[7]Life technologie
"Ion semiconductor sequencing uniquely enables both accurate long reads and paired-end sequencing"
http://www.iontorrent.com/lib/images/PDFs/pe_appnote_v12b.pdf
Application note, octobre 2011

[10] Illumina
"Mate pair sequencing assay"
http://www.illumina.com/technology/mate_pair_sequencing_assay.ilmn

[11] Illumina
"Paired-end Sequencing Assay"
http://www.illumina.com/technology/paired_end_sequencing_assay.ilmn

[12] SOLiD
"Library preparation"
http://nextgenseq.onc.jhmi.edu/nextgenseq/technology.html