Plateforme de recherche de mutations

L'identification des causes des maladies génétiques humaines n'est pas encore terminée.

Les méthodes de séquençage à haut débit permettent de séquencer des régions de taille très importantes

Les méthodes de « capture des séquences » permettent de cibler les régions pour les séquencer ultérieurement avec les séquenceurs à haut débit.

« Sequence Capture »

- Malgré le fait qu'on arrive a séquencer à haut débit –
- on n'arrive pas encore à séquencer la totalité du génome humain pour un prix raisonnable et dans un délai raisonnable.
- → il faut séquencer les régions ciblées

Plateforme

- **Les intervenants techniques:**

-Capture des séquences: Laboratoire de Ressources Génomiques (LRG).

-Séquençage: Laboratoire de Séquençage.

-Traitements et analyses des données: Laboratoire d'Analyse Bioinformatique des Séquences. (LABiS)

D’où viennent les projets?

•Appels de proposition « Maladies rares ».
- « Guichet ouvert ».
- Conseil et expertise de la plateforme.
- Passage devant le conseil scientifique désigné par l'INSERM.

•Appels de proposition « Génoscope ».

Plateforme mutations.

Fichier Édition Affichage Historique Marque-pages Outils ?

http://www.genoscope.cns.fr/spip/Plateforme-d-identification-de.html

Les plus visités NCBI HomePage Google Sytadin : l'état du traf... FT Privilege - Newspa... BiblioVie : Rechercher ... CEA_mail Ensembl Genome Bro... Genoscop

Centre National de Séquençage

GENOSCOPE

ACCUEIL | PRÉSENTATION | SÉQUENÇAGE | RECHERCHE | RESSOURCES | PUBLICATIONS

Ressources bioinformatiques
Plateforme Mutations
Serveurs externes

Rechercher

Plateforme d'identification de mutations humaines par séquençage à haut débit

Accueil du site > Ressources >

Modalités d'accès
Annexe scientifique

Lettre du 11 décembre 2008

L'identification des causes des pathologies humaines d'origine génétique est loin d'être achevée. Il reste en particulier à identifier les gènes les plus rarement impliqués. Ils n'en sont pas moins importants sur le plan scientifique et médical et cette identification reste une condition indispensable pour la compréhension des processus biologiques concernés. Les GIS Institut des maladies rares et IBiSA, les Instituts thématiques de l'Inserm de Génétique et

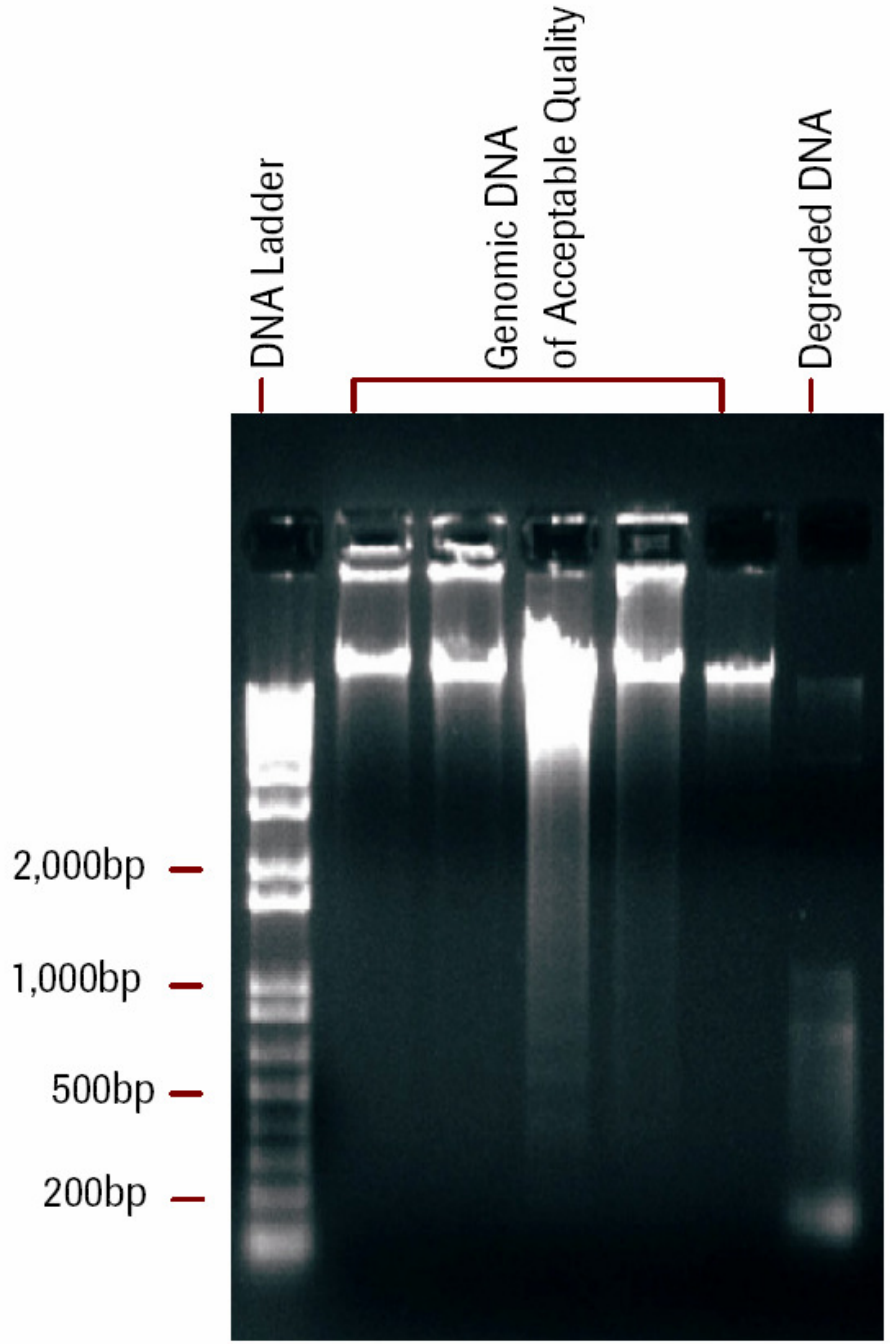
Terminé

« In put »

ADN de bonne qualité

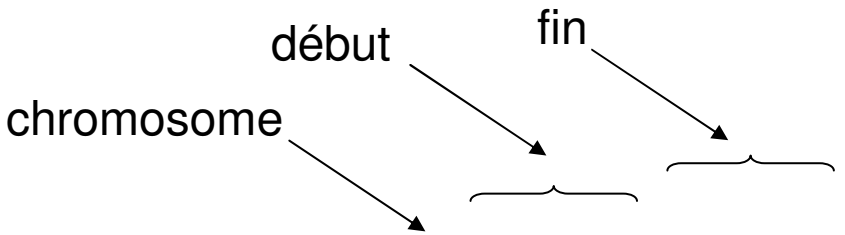
$D0_{268/280}$ >1,8

Quantité: 21 µg



Fichier de régions à capturer

Il faut préciser le numéro de « built »



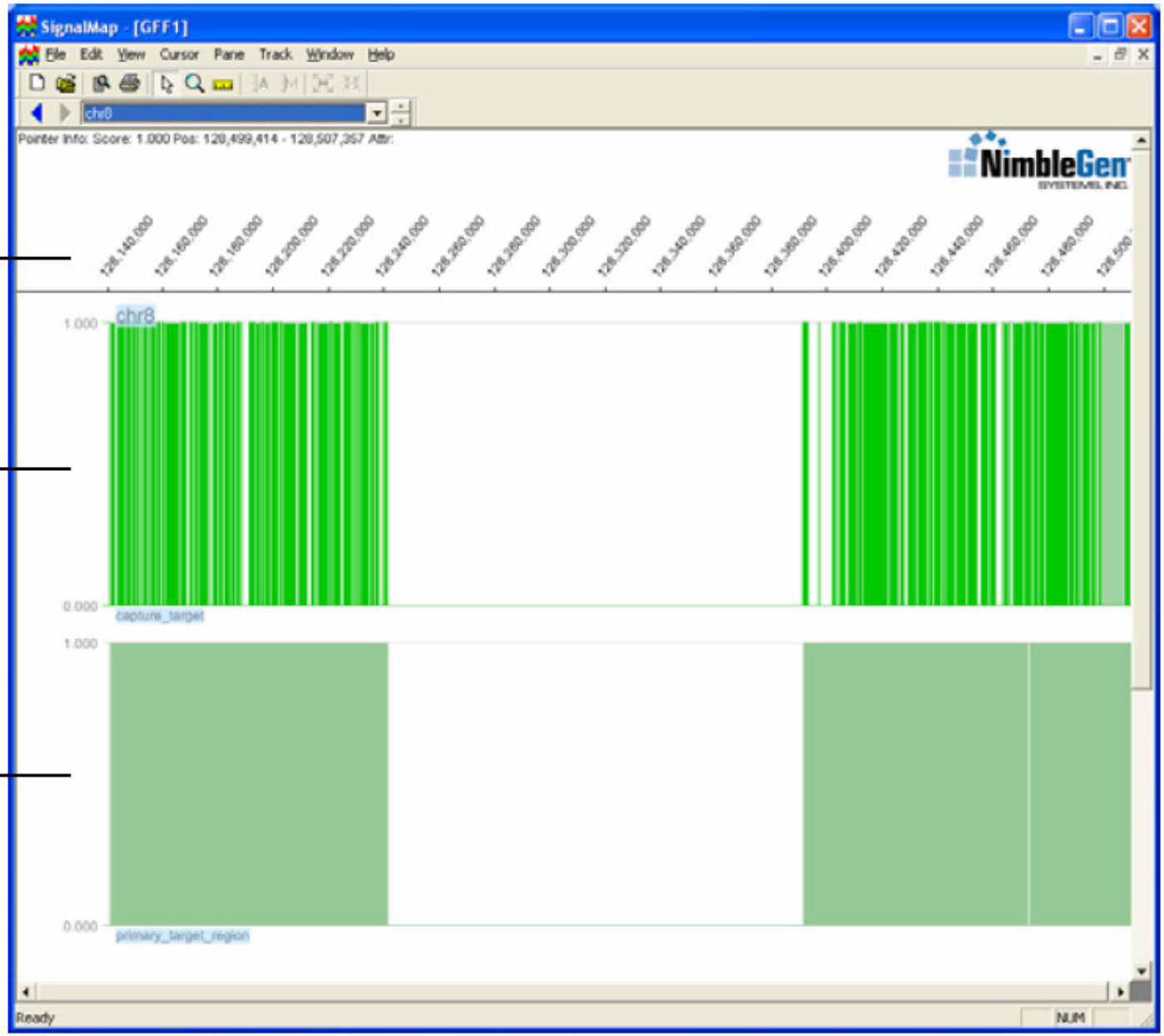

```
9 130069294 130069374
9 130069558 130069617
9 130070034 130070097
9 130070217 130070260
9 130070520 130070624
9 130075950 130076072
9 130077993 130078089
11 6694530 6695529
11 6695584 6697157
10 98747785 98751097
10 98751905 98752116
10 98752451 98752739
10 98753815 98753969
10 98754440 98754570
10 98756230 98756467
10 98760740 98760877
10 98763488 98763581
10 98763766 98763905
10 98768333 98768430
10 98768730 98768854
10 98770997 98771160
10 98780504 98780575
10 98781353 98781424
10 98784218 98784289
```

.GFF

genomic coordinates

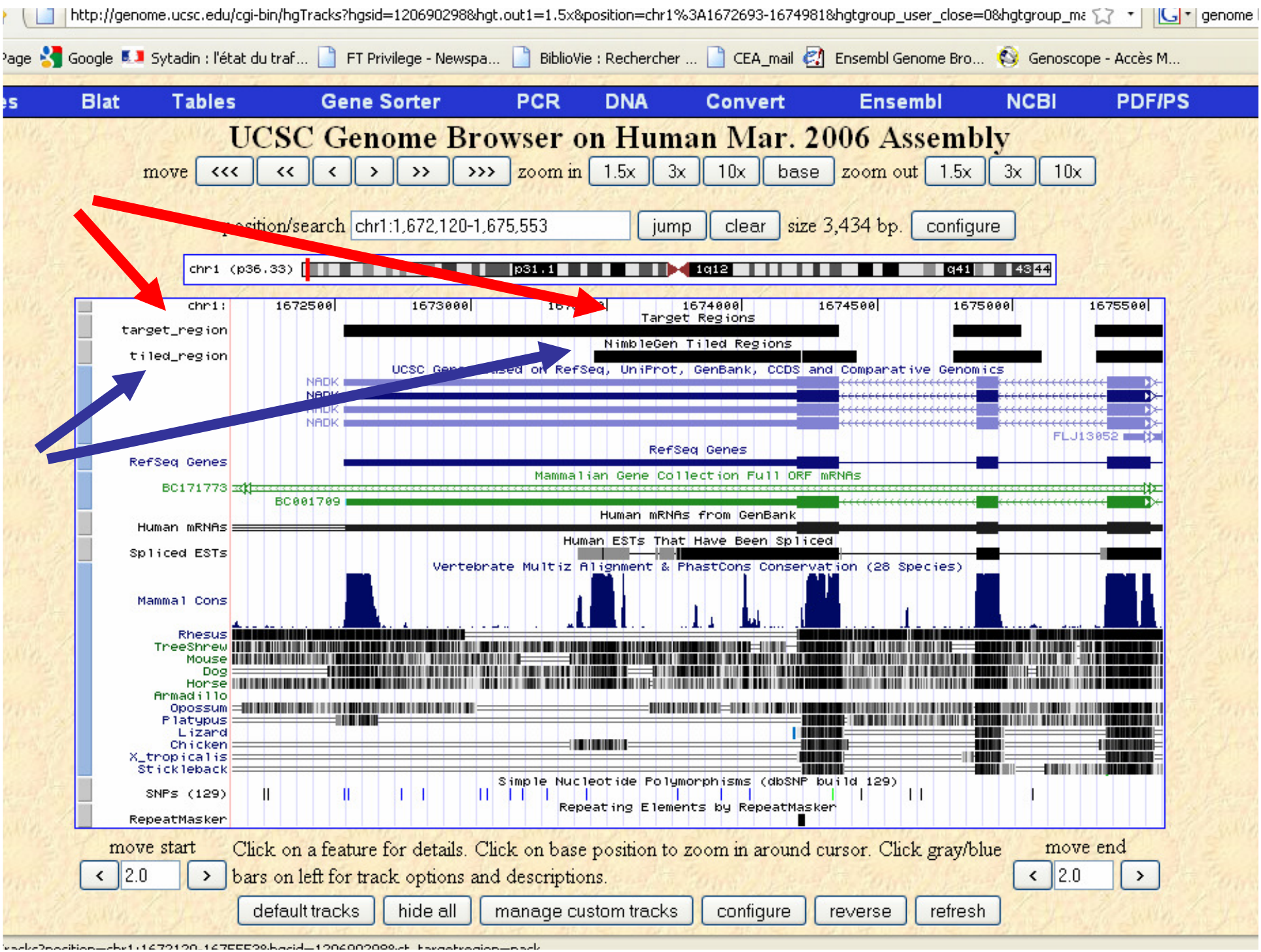
capture_target_ track

primary_target_ region track

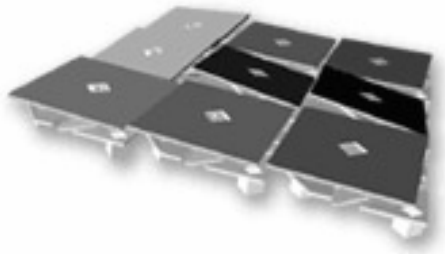


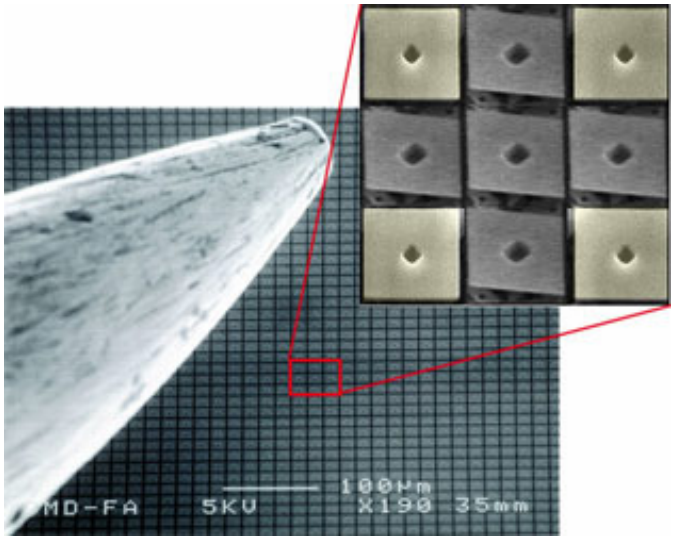
Utilisation le format .BED

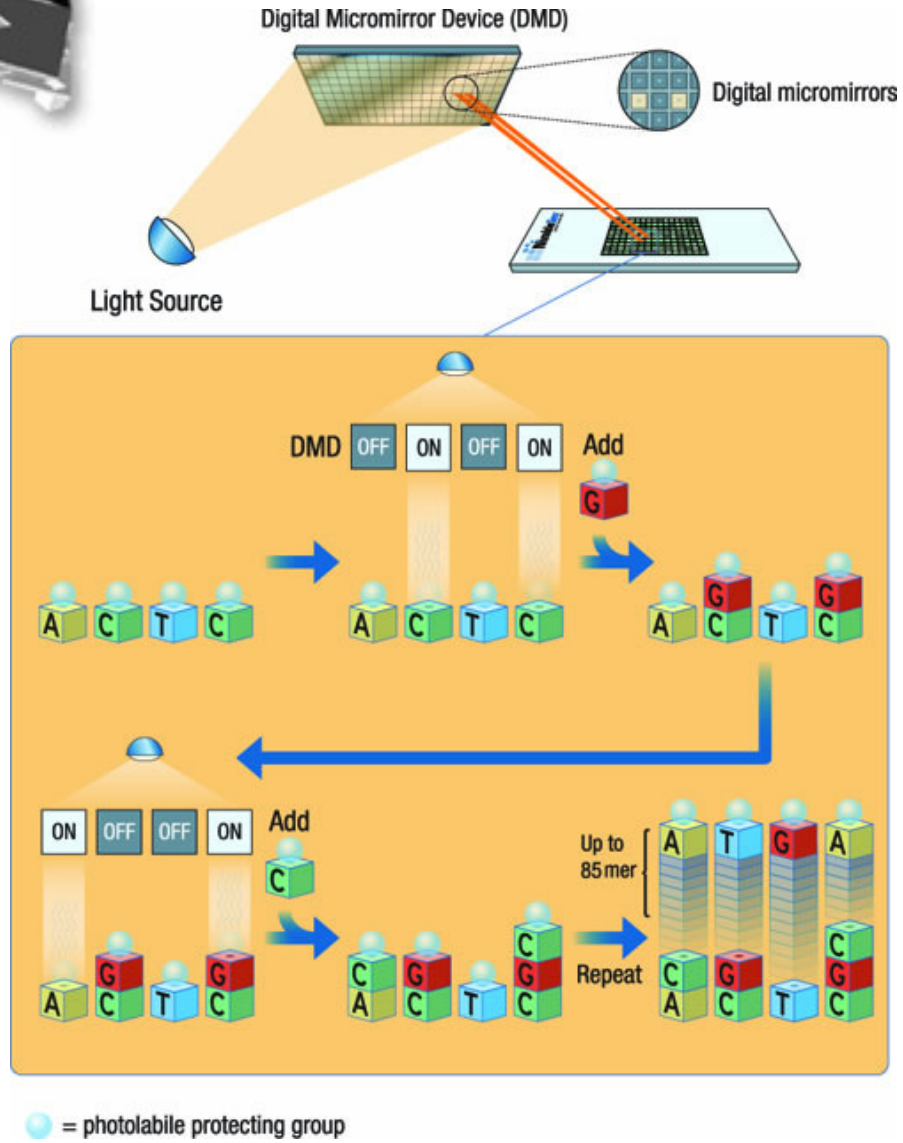
- Aller au site : www.genome.ucsc.edu

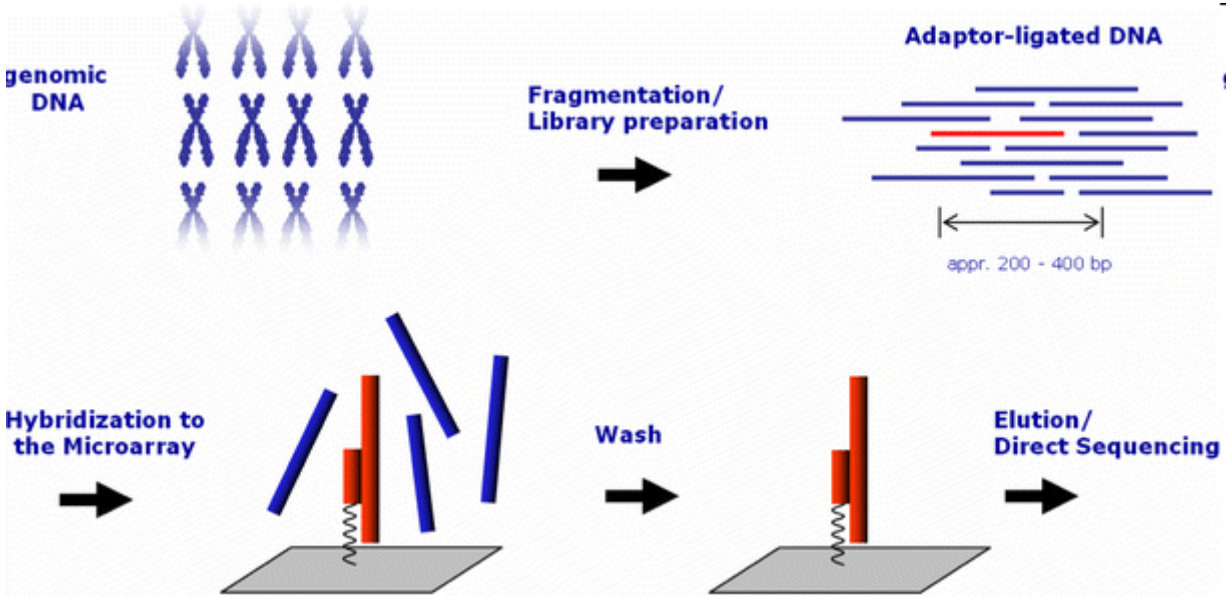

http://genome.ucsc.edu/cgi-bin/hgTracks?hgsid=120690298&hgt.out1=1.5x&position=chr1%3A1672693-1674981&hgtgroup_user_close=0&hgtgroup_ma
Google
Sytadin : l'état du traf...
FT Privilege - Newspa...
BiblioVie : Rechercher ...
CEA_mail
Ensembl Genome Bro...
Genoscope - Accès M...
Blat
Tables
Gene Sorter
PCR
DNA
Convert
Ensembl
NCBI
PDF/PS
UCSC Genome Browser on Human Mar. 2006 Assembly
move
<<<
<<
<
>
>>
>>>
zoom in
1.5x
3x
10x
base
zoom out
1.5x
3x
10x
position/search
chr1:1,672,120-1,675,553
jump
clear
size 3,434 bp.
configure
chr1 (p36.33)
p31.1
1q12
q41
43 44
chr1:
1672500
1673000
1674000
1674500
1675000
1675500
Target Regions
target_region
NimbleGen Tiled Regions
tiled_region
UCSC Genes based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
NADK
FLJ13052
RefSeq Genes
RefSeq Genes
Mammalian Gene Collection Full ORF mRNAs
BC171773
BC001709
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
Vertebrate Multiz Alignment & PhastCons Conservation (28 Species)
Mammal Cons
Rhesus
TreeShrew
Mouse
Dog
Horse
Armadillo
Opossum
Platypus
Lizard
Chicken
X_tropicalis
Stickleback
Simple Nucleotide Polymorphisms (dbSNP build 129)
SNPs (129)
Repeating Elements by RepeatMasker
RepeatMasker
move start
2.0
Click on a feature for details. Click on base position to zoom in around cursor. Click gray/blue bars on left for track options and descriptions.
move end
2.0
default tracks
hide all
manage custom tracks
configure
reverse
refresh

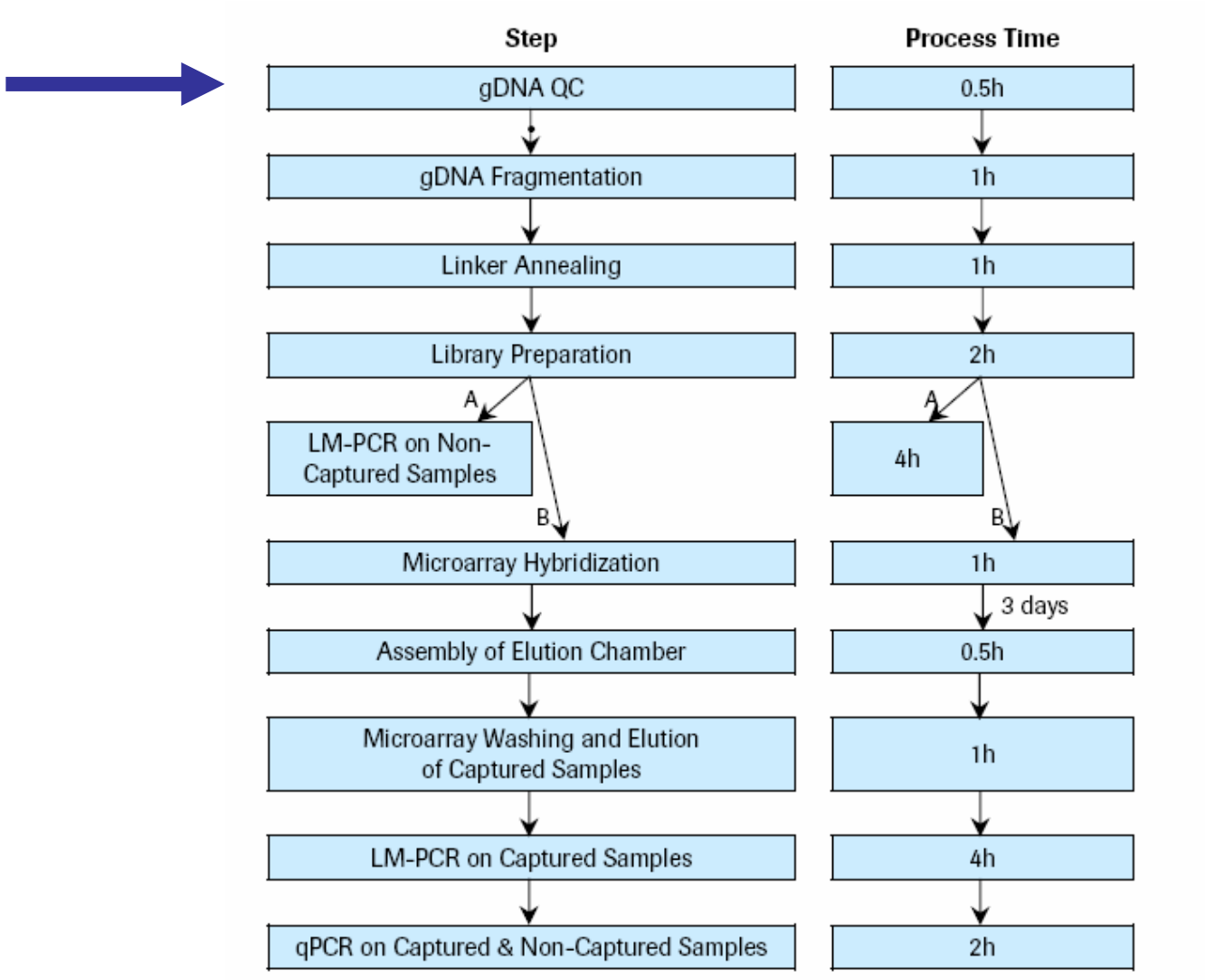
Digital Light Processing™ technology








genomic
DNA
Fragmentation/
Library preparation
Adaptor-ligated DNA
appr. 200 - 400 bp
Hybridization to
the Microarray
Wash
Elution/
Direct Sequencing

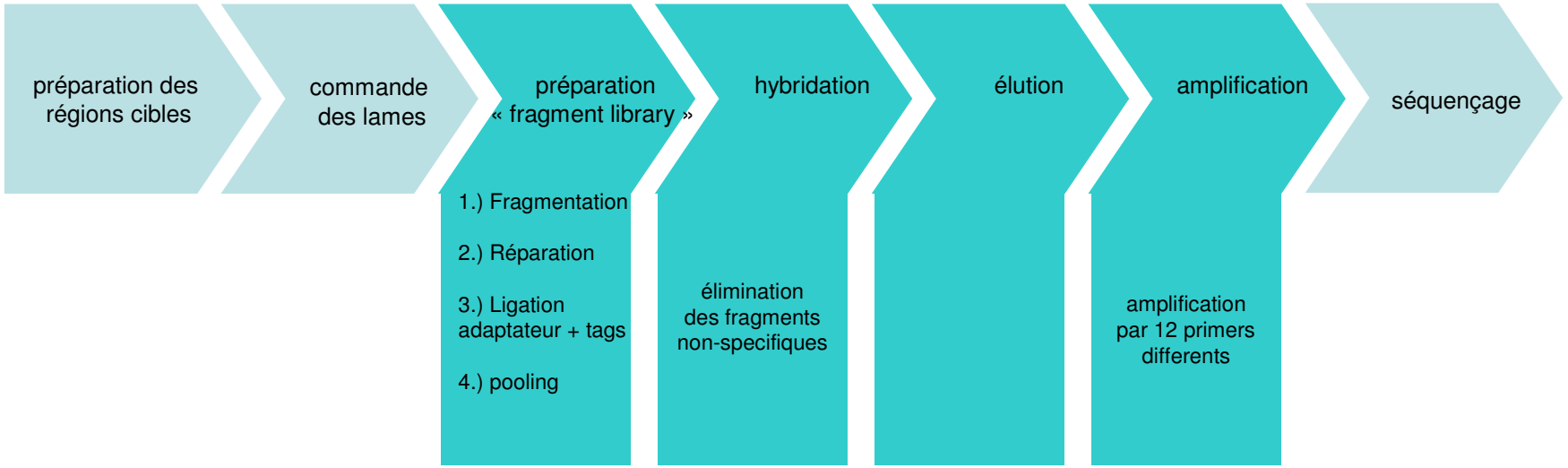

Step
Process Time
gDNA QC
0.5h
gDNA Fragmentation
1h
Linker Annealing
1h
Library Preparation
2h
A
A
LM-PCR on Non-Captured Samples
4h
B
B
Microarray Hybridization
1h
3 days
Assembly of Elution Chamber
0.5h
Microarray Washing and Elution of Captured Samples
1h
LM-PCR on Captured Samples
4h
qPCR on Captured & Non-Captured Samples
2h

Comparaison

- Approche classique
  - Taille de région limitéé
  - Amplification des régions individuellement (PCR)
  - Séquençage des régions individuellement
  - Grand nombre d'individus (séquencés individuellement)

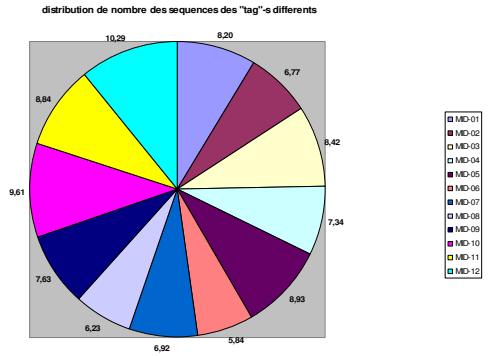
- Plateforme à haute débit
  - Taille de région plus grande
  - Capture des séquences de grandes régions.
  - Séquençage simultané des régions
  - Nbr. des individus limités (→ multiplexage)

Multiplexage



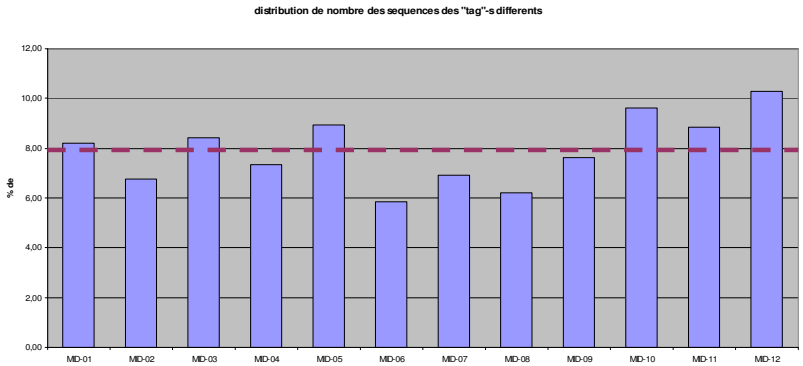
Adaptor Multiplex Identifier Sequence

MID-1	ACGAGTGCGT
MID-2	ACGCTCGACA
MID-3	AGACGCACTC
MID-4	AGCACTGTAG
MID-5	ATCAGACACG
MID-6	ATATCGCGAG
MID-7	CGTGTCTCTA
MID-8	CTCGCGTGTC
MID-9	TAGTATCAGC
MID-10	TCTCTATGCG
MID-11	TGATACGTCT
MID-12	TACTGAGCTA



Répartition des « tag »-s

			114260	séquences				
MID-01	ACGAGTGCGT		9365	8,20	%			
MID-02	ACGCTCGACA		7730	6,77	%			
MID-03	AGACGCACTC		9620	8,42	%			
MID-04	AGCACTGTAG		8390	7,34	%			
MID-05	ATCAGACACG		10204	8,93	%			
MID-06	ATATCGCGAG		6674	5,84	%			
MID-07	CGTGTCTCTA		7902	6,92	%			
MID-08	CTCGCGTGTC		7116	6,23	%			
MID-09	TAGTATCAGC		8721	7,63	%			
MID-10	TCTCTATGCG		10981	9,61	%			
MID-11	TGATACGTCT		10097	8,84	%			
MID-12	TACTGAGCTA		11757	10,29	%			
		=	108557	95,01	%			
?	unmapped		5242	4,59	%			
?	too short		2245	1,96	%			
?	mapped (partial + full)		70037	61,30	%			
			moyenne:	7,92	±	1,37	±	17,29 %

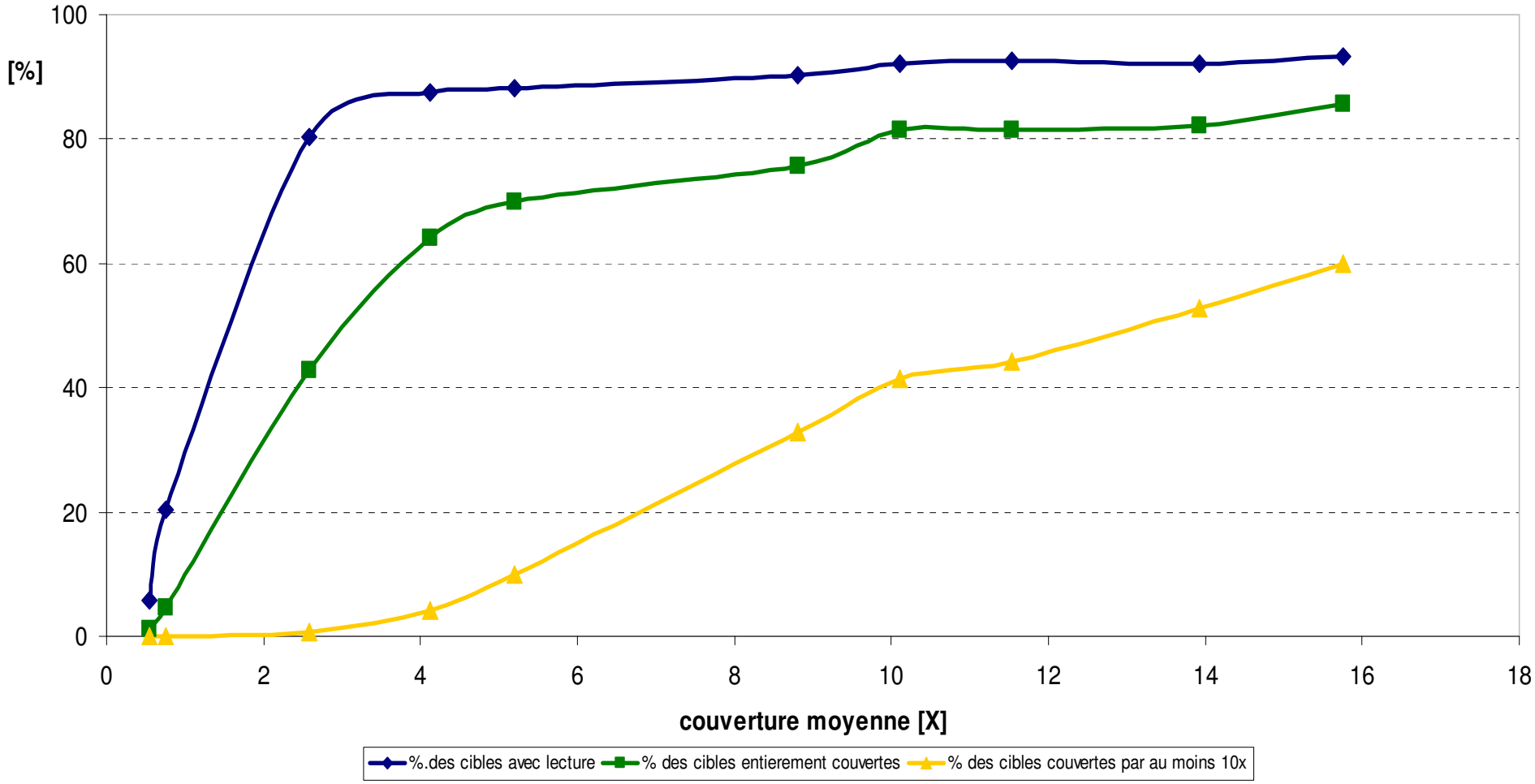


Projet pilote

- Séquençage de 13,315 exons
- Total « target size »: 3,957 Mb
- Total « tiled » size: 10,718 Mb

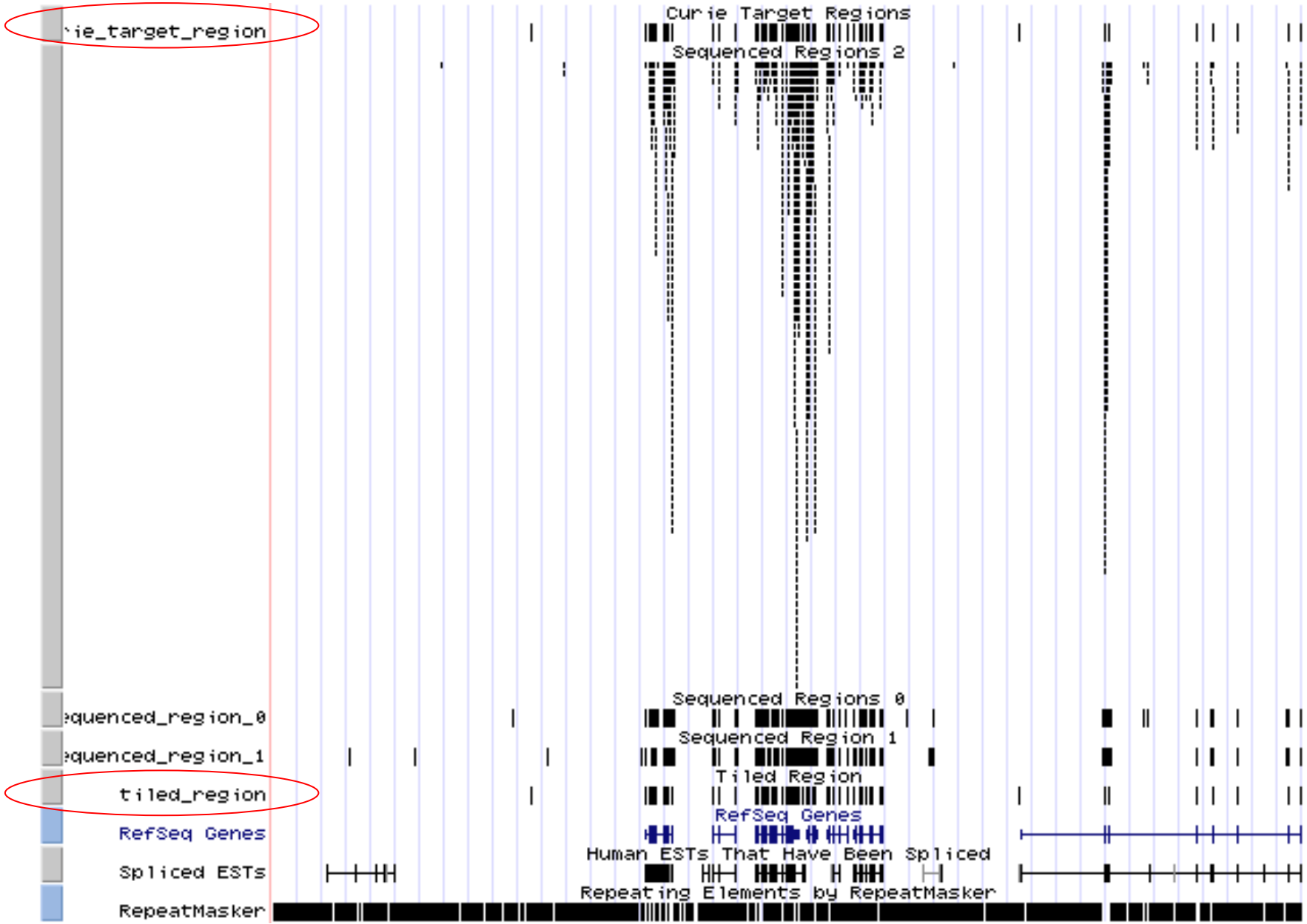
Library	# of reads	# of mapped reads	# of mapped reads which overlap enriched regions	# of mapped reads included in enriched regions	# of targets with reads	# of target regions entirely covered	Avg. coverage	# of targets with a coverage ratio of 10X at least
B	**740642**	649017	450267	220646	12275	10932	**13,92X**	7026
B	100%	**87,63%**	**69,38%**	**49,00%**	**92,19%**	**82,10%**		**52,77%**
C	**964866**	822999	525778	260185	12434	11405	**15,76**	7985
C	100%	**85,30%**	**63,89%**	**49,49%**	**93,38%**	**85,66%**		**59,97%**
D	**602719**	564580	353295	175029	12796	11091	**12,73X**	7123
D	100%	**93,67%**	**62,58%**	**49,54%**	**96,10%**	**83,30%**		**53,50%**
E	**601841**	531657	348492	160027	12325	10856	**11,54X**	5886
E	100%	**88,34%**	**65,55%**	**45,92%**	**92,56%**	**81,53%**		**44,21%**
F	**683096**	607093	269594	119974	10574	8610	**10,47X**	4097
F	100%	**88,87%**	**44,41%**	**44,50%**	**79,41%**	**64,66%**		**30,77%**
H	**480811**	431060	297016	131609	12261	10862	**10,1**	5502
H	100%	**89,65%**	**68,90%**	**44,31%**	**92,08%**	**81,58%**		**41,32%**

Relation entre la couverture générale et couverture individuelle de « cible »

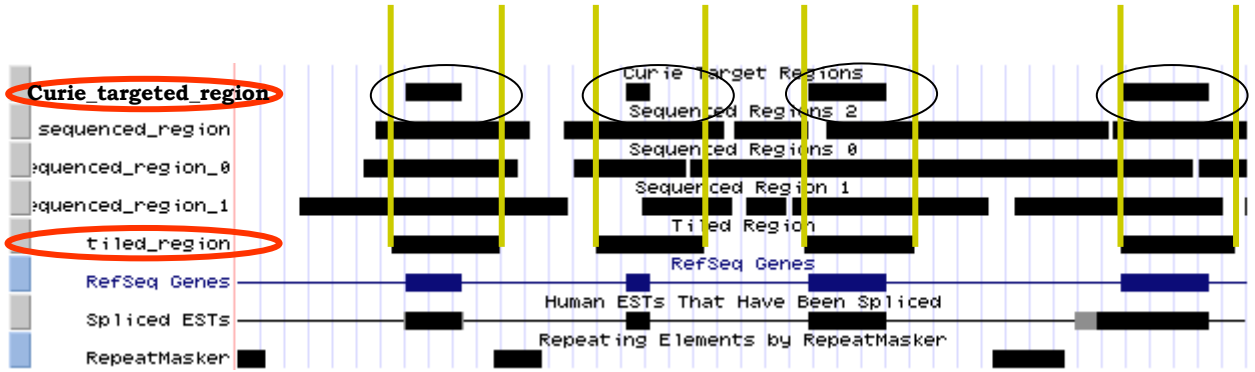


- Spécificité de la capture
- Répartition des tailles →WGA
- Uniformité de représentation
- Analyse de données
  - Statistiques de capture
  - Représentation des différences
  - Soustraction des témoins

Size of this region
519,722 bp.



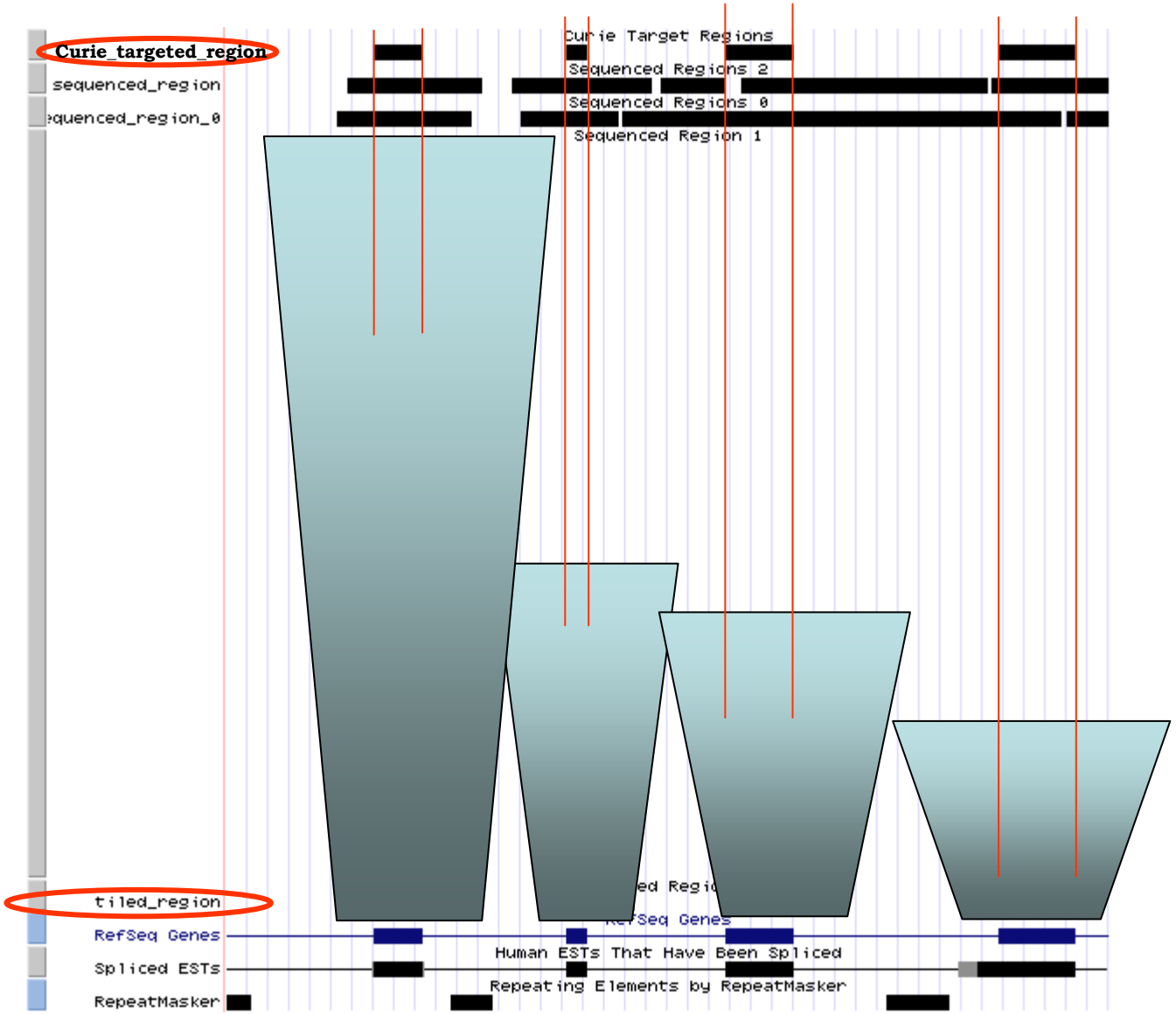
Taille de la région
2,951 bp.



target region ←→ tiled region

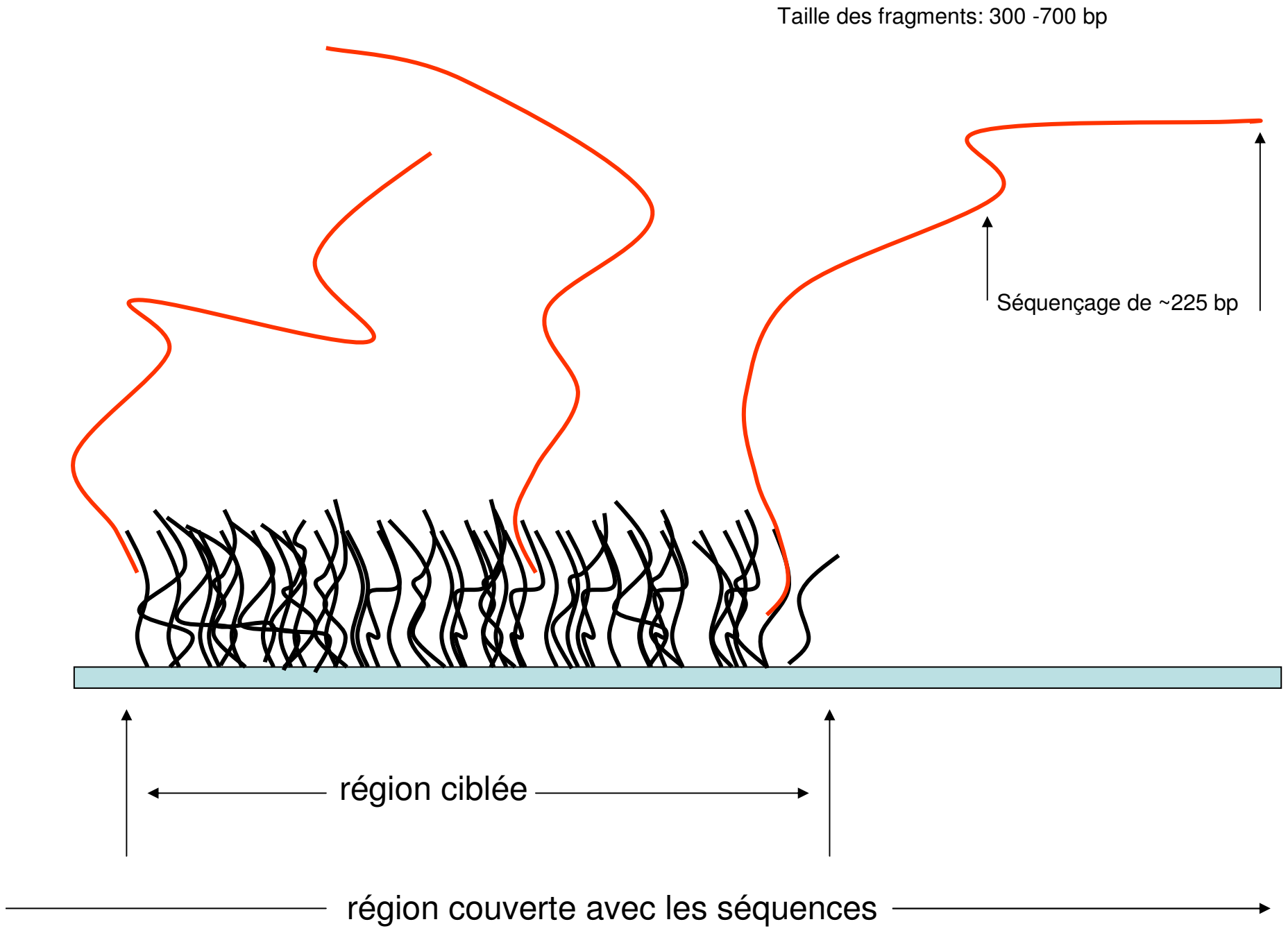
Taille de la région
2,951 bp.

Nombre de lectures:
189 955



Taille de la région
2,951 bp.

Nombre de lectures:
601 841

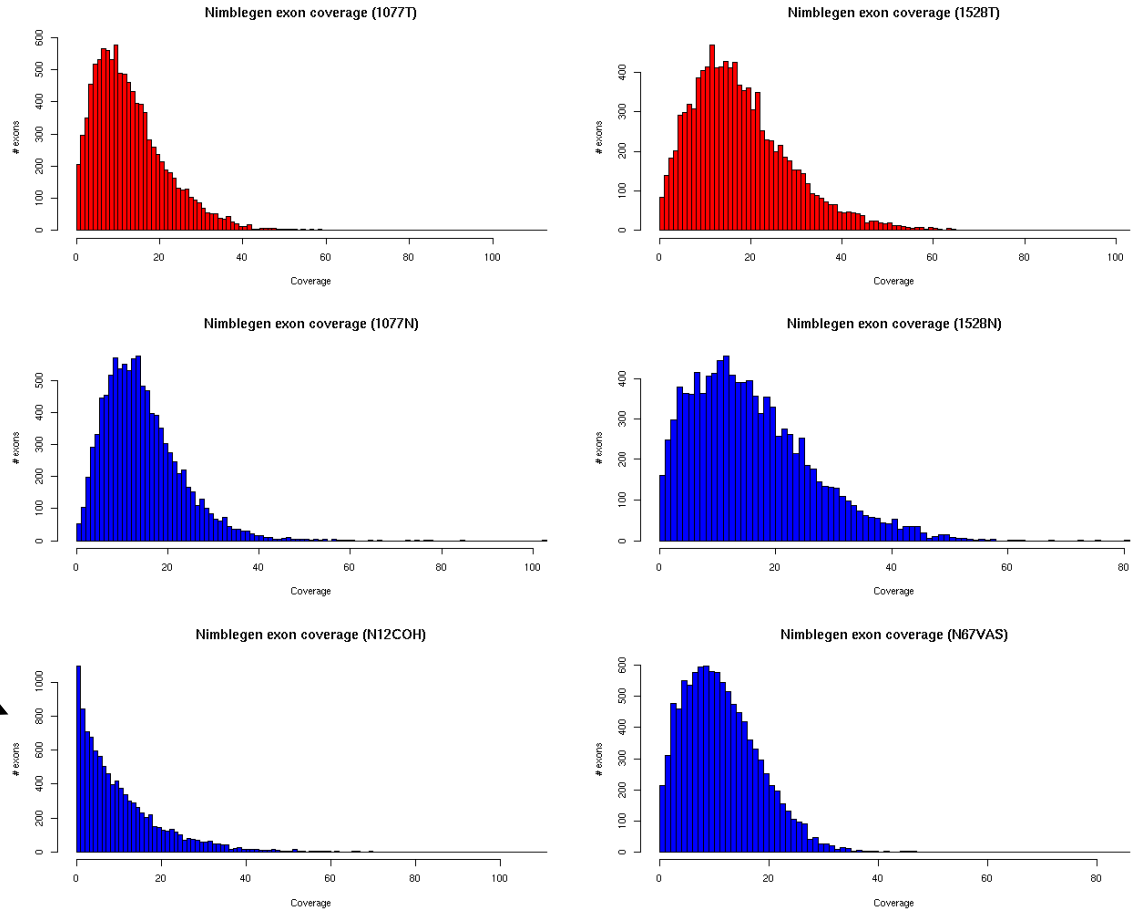

Taille des fragments: 300 -700 bp
Séquençage de ~225 bp
région ciblée
région couverte avec les séquences

Effet d'intégrité des régions sur le spécificité

Library name	# of reads	# of reads mapped uniquely to the genome	# of reads overlapping the selection targets	# of reads mapped into the target region
B	216 700	171 713 **(79%)**	129 514 **(75%)**	79 428 **(46%)**
C	172 955	136 030 **(79%)**	96 584 **(71%)**	58 035 **(43%)**
E	199 867	173 712 **(87%)**	125 813 **(72%)**	69 453 **(40%)**
Chr_8_test	59 446	43 516 **(73,2%)**	40 234 **(92,5%)**	40 150 **(92,3%)**

- Spécificité de la capture
- Répartition de la couverture →WGA
- Uniformité de représentation
- Analyse de données
  - Statistiques de capture
  - Représentation des différences
  - Soustraction des témoins

Répartition de couverture de 13315 régions



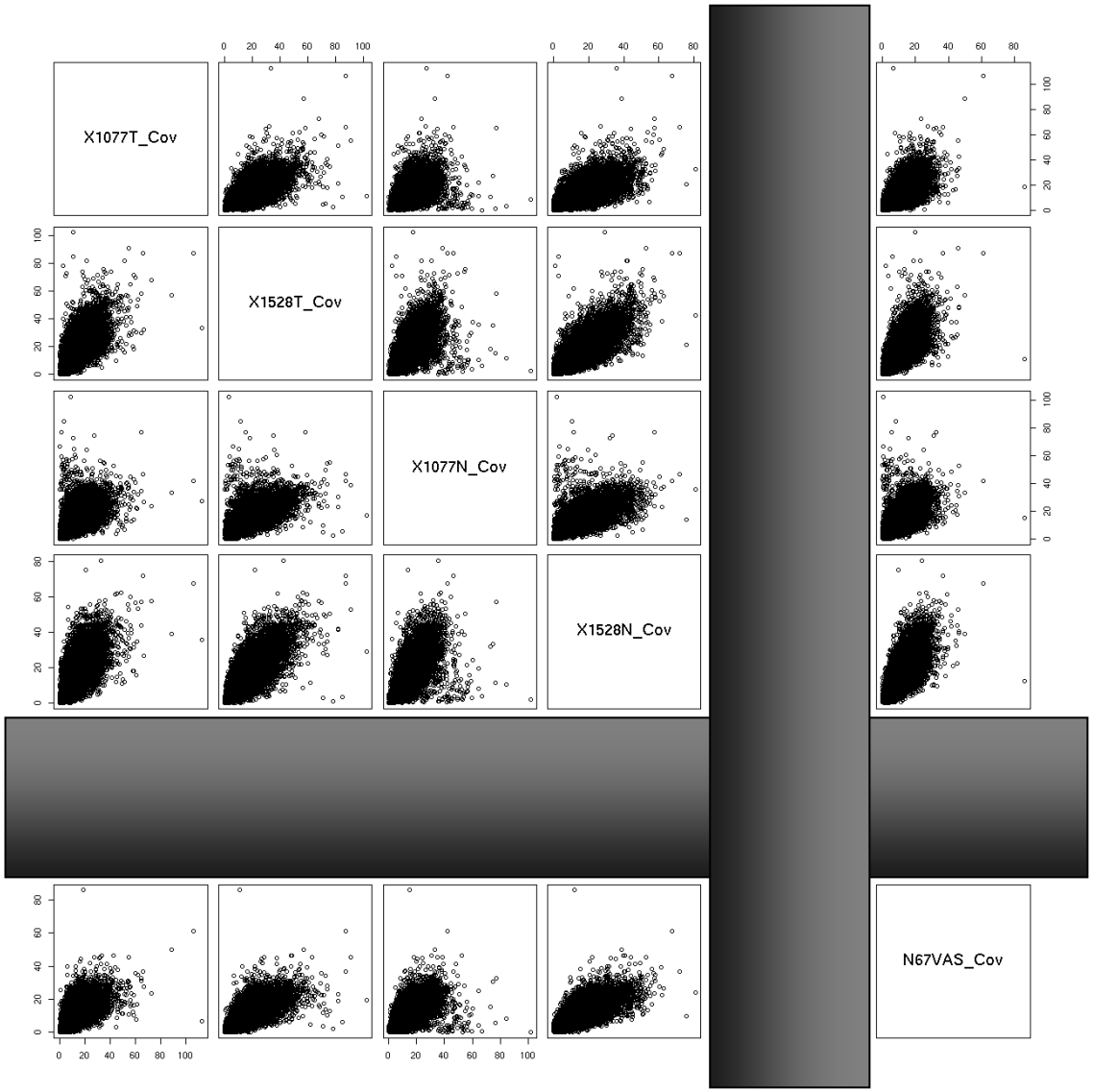
tumeur

normal

WGA

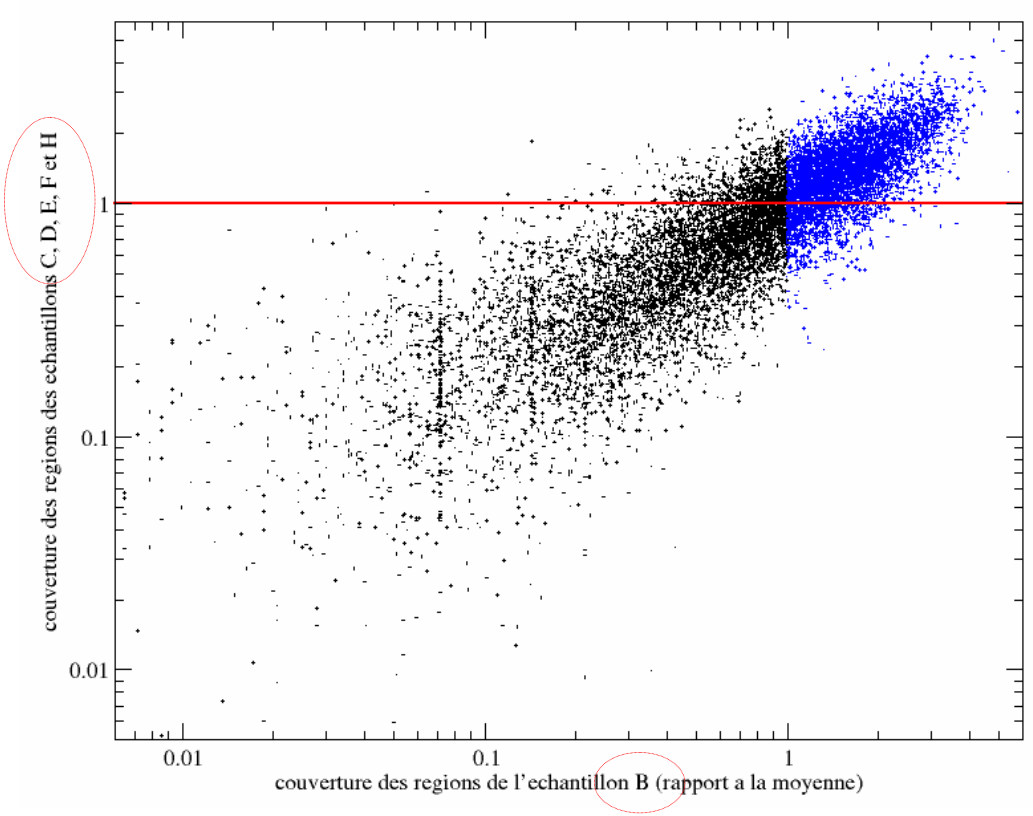
- Spécificité de la capture
- Répartition de la couverture →WGA
- Uniformité de représentation
- Analyse de données
  - Statistiques de capture
  - Représentation des différences
  - Soustraction des témoins

Comparaison deux à deux des répartition de couverture de 13315 régions



Comparaison de la répartition de la couverture des exons de 1077T et les autres échantillons

Les régions faiblement couvertes sont souvent communes à différents échantillons => biais de capture



- Spécificité de la capture
- Répartition de la couverture →WGA
- Uniformité de représentation
- Analyse de données
  - Statistiques de capture
  - Représentation des différences
  - Soustraction des témoins

Recherche des differences

Nombre de différences

Nombre d'exons:	13,315
Total « target size »:	3,957 Mb
Total « tiled » size:	10,718 Mb

Nom_individue	All_diff	HC_Diff
B	19 567	4 523
C	23 012	5 325
E	15 793	3 826
H	13 731	3 706
D	15 915	4 490
F	13 710	3 460
moyenne	**16 955**	**4 222**

Projet cancer de la vessie

- localisation des variations (exon, intron, intergénique) et comparaison entre échantillon
- Sélection des variations :
    - de haute qualité présentes dans les échantillons tumoraux
    - absentes des échantillons normaux (apparié et autres)
    - absentes des bases de données de variations connues
    - localisées dans les exons ou à proximité (50pb)

Exemple sur un échantillon tumoral:

couverture	# HC_diff	# HC_diff – des diff. dans d'autres échantillons	# HC_diff – des diff. dans d'autres échantillons – des diff. dans des base de données	Dont ceux qui se trouvent dans des exons
2	3092	406	360	244
3	3092	406	360	244
4	2721	218	183	132
5	2439	141	120	96
6	2202	110	93	77
10	**1457**	**53**	**45**	**40**

Projet cancer de la vessie

- Initialement environ 20.000 variations de haute qualité
- Une 50aine de variations à valider par re-séquençage après classification et sélection
- Les critères de sélection importants :
    - qualité de la variation (profondeur de séquence)
    - localisation de la variation
    - comparaison entre échantillon et avec les variations connues

Conclusions

- La plateforme a terminé les projets pilotes et commence à accueillir les premiers projets.
- Les outils d'analyse sont sélectionnés
- Couverture de 10X environ nécessaire pour une détection optimale des mutations
- Le nombre de lectures à séquencer dépend des régions ciblées (nombre, taille) et de la qualité du run (taille moyenne des lectures, qualités)
- Nécessité de classifier les variations pour réduire l'espace de recherche et de valider chaque variation