

25/07/2013 Traitement des séquences obtenues par pyroséquençage des produits en utilisant le JUNIOR au sein de la plateforme GenoSol (RUN du 23/07/2013).

ORGANISATION DES ECHANTILLONS

Ce RUN a pour objectif de tester et d'évaluer la reproductibilité et la bonne qualité de l'ajout des adaptateurs par PCR. En effet, habituellement, le protocole utilisé consiste en deux étapes de PCRs : une première PCR pour amplifier le fragment de gène recherché sur la matrice ADN environnementale, puis une seconde PCR dite « nested » en utilisant comme matrice le produit purifié de la première PCR pour ajouter les MIDs. Le produit obtenu est ensuite utilisé dans une étape de ligation pour ajouter les adaptateurs. Cependant, cette étape de ligation n'est pas très stable et reproductible. De plus, le produit obtenu est difficile à doser et à conserver.

La solution envisagée ici est d'ajouter les adaptateurs en même temps que les MIDs durant la seconde étape de PCR. Pour évaluer l'effet du nombre de cycles, mais aussi des différents MIDs utilisés, 8 échantillons avec 8 MIDs différents provenant d'une PCR de 7 cycles avec 7,5ng de matrice ont été évalués, mais aussi 8 échantillons avec 8 autres MIDs provenant d'une PCR de 9 cycles avec 5ng de matrice. Le mélange équimolaire a été réalisé. Enfin, ce mélange a été séquencé par pyroséquençage au sein de la plateforme GenoSol.

PRIMERS UTILISES:

PCR1

Oligo sens (F479): CAGCMGCGCNGTAANAC

Oligo antisens (R888) : CCGYCAATTCMTTTRAGT

PCR2

Oligo sens (16S_F479_ADAPT B) :

CCTATCCCCTGTGTGCCTTGGCAGTCGACTCAGCMGCGCNGTAANAC

Oligo antisens (16S_R888_MID_ADAPT A) :

CCATCTCATCCCTGCGTGTCTCCGACGACTCGTGTCTCTACCGYCAATTCMTTTRAGT

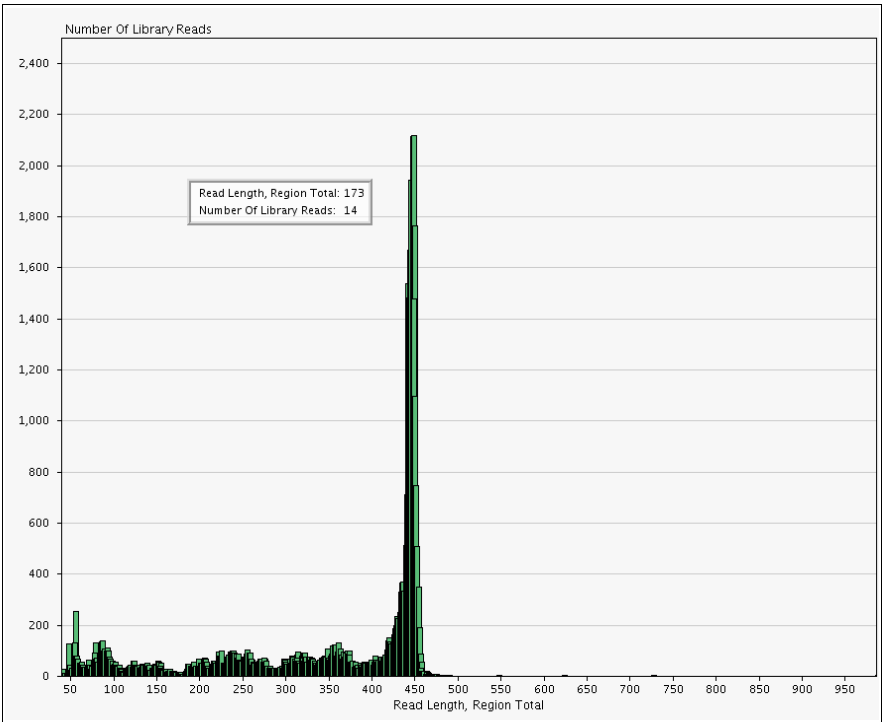
Légende : adaptateurs A et B (bleu), clé (gras), MID (rouge), primer (souligné).

DONNEES DU RUN

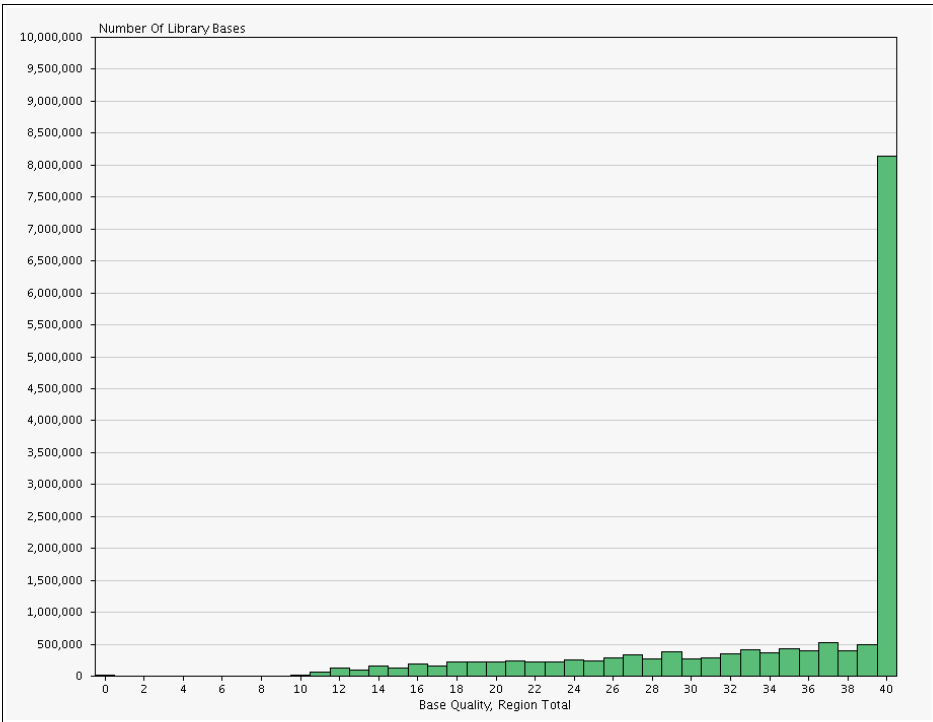
NOMBRE DE READS BRUTS

Library (GACT)			
Key Pass	Passed Filter (% Key Pass)	Average Length	Total Bases
319,665	45,890 (14.4%)	350.8 ± 126.5	16,097,885 bp
Control (ATGC)			
Key Pass	Passed Filter (% Key Pass)	Average Length	Total Bases
1,144	645 (56.4%)	373.9 ± 143.8	241,193 bp
Control (CATG)			
Key Pass	Passed Filter (% Key Pass)	Average Length	Total Bases
1,034	545 (52.7%)	307.9 ± 130.2	167,779 bp
Comments			

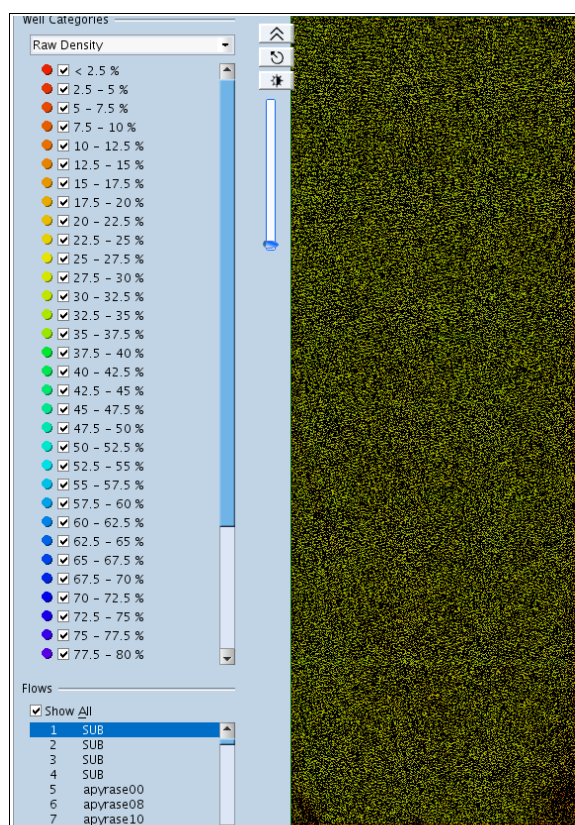
REPARTITION DE LA TAILLE DES READS DANS LA LIBRAIRIE



REPARTITION DE LA QUALITE DES BASES DES READS DANS LA LIBRAIRIE



REPARTITION DANS LA PLAQUE DES READS



DETAILS RUN ET FILTRES APPLIQUES

GACT (Library)		Total
Raw Wells		326,812
Key Pass Wells		319,665
Failed	Dot	27,936
	Mixed	56,308
	Short Quality	187,25
	Short Primer	1
Passed Filter Wells		45,89
% Dot + Mixed		26.35
% Short		58.58
% Passed Filter		14.36

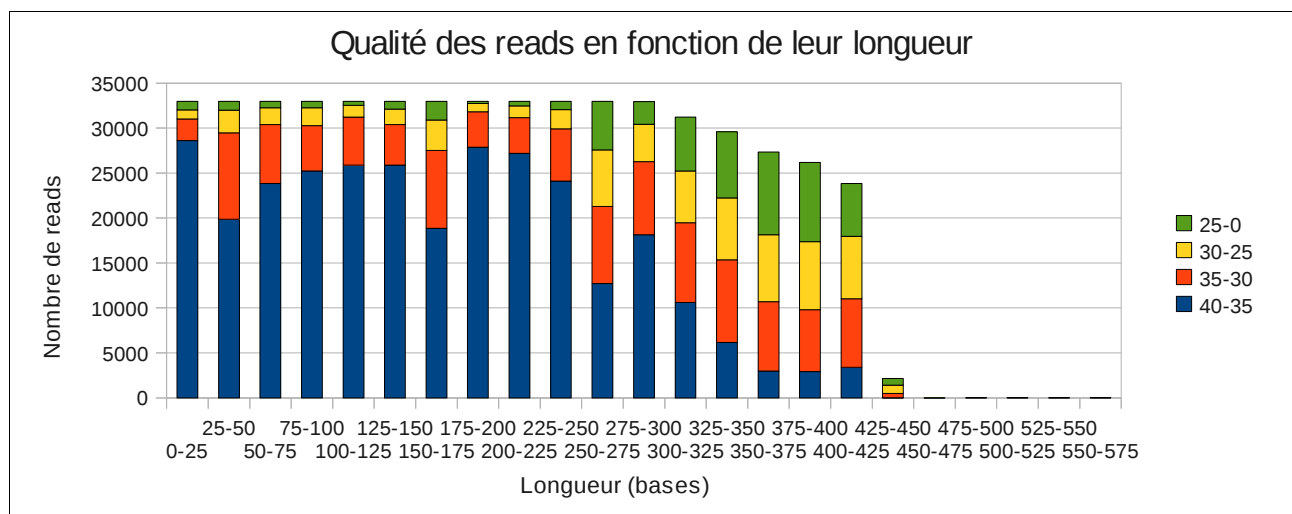
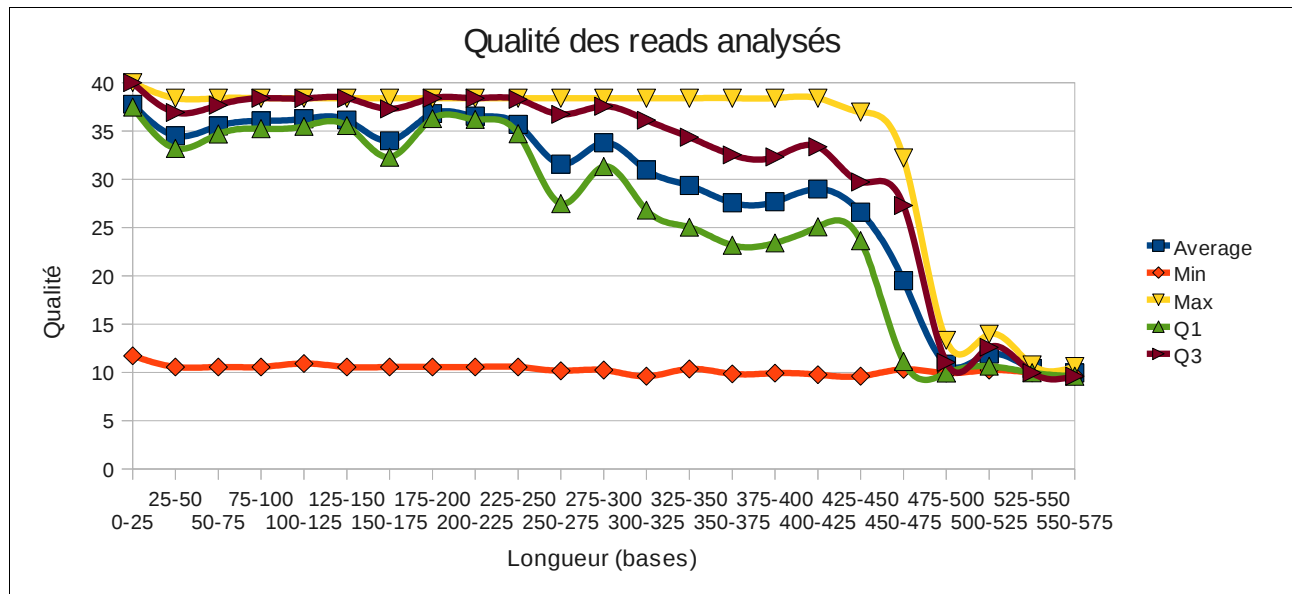
BILAN

Comme le montre ces premières données, la répartition des billes est homogène au sein de la PicoTiterPlate. Le nombre de billes détecté est très élevé (plus de 300 000 billes), mais ne correspond pas encore à un chargement excessif. Cependant, un grand nombre de ces billes est de type « Short Quality ». Cela vient probablement du fait que l'enrichissement obtenu après emPCR soit trop élevé (20%), et que le nombre de cpb (copies par bille) défini soit trop élevé pour un produit non ligué (confirmation de S. Ferreira de Genoscreen, qui prévoit entre 0.3 et 0.5 cpb avec ce type de produit). Les résultats de ce RUN restent analysables, car les séquences produites semblent de bonne qualité.

STEP 00 : Evaluation de la qualité des reads bruts obtenus avant preprocessing pour déterminer la taille des reads a utiliser, etc... (Programme PERL : *Eval_qual.pl* sur le serveur avec paramètre de taille à **300 bases minimum**).

Rappel : qualité élevée (entre 40 et 35), qualité moyenne (entre 25 et 35), en-dessous mauvaise qualité, avec risque d'erreur.

Nombre de reads : 45 890 reads bruts (32 989 reads analysés)



Bilan : Comme le montre ces graphiques, la qualité des données obtenues est assez importante (mais avec des baisses à certaines zones, liées à l'orientation défini du séquençage). On note par contre une baisse un peu plus importante du séquençage dès 300 bases. Ce RUN va pouvoir être analysé, au moins pour déterminer la reproductibilité de chaque échantillon en fonction de chaque MID et nombre de cycles.

STEP 01 : Preprocessing des séquences (recherche des MIDs). (Programmes PERL : tout d'abord le *preprocess_seqs_midonly.pl* disponible sur le serveur pour séparer les premiers échantillons en fonction de leurs MIDs)

Number of chosen MIDs chosen by the user : 16

MID021:	CGTAGACTAG	Corresponding name:	7L21
MID058:	CGTACAGTCA	Corresponding name:	7L58
MID090:	CACGTGTCGC	Corresponding name:	7L90
MID094:	CGTATGCGAC	Corresponding name:	7L94
MID127:	CACACGATAG	Corresponding name:	7L127
MID131:	CGACAGCGAG	Corresponding name:	7L131
MID161:	CGCGACATCT	Corresponding name:	7L161
MID170:	CATATCTCGT	Corresponding name:	7L170
MID013:	CATAGTAGTG	Corresponding name:	9L13
MID036:	CGACGTGACT	Corresponding name:	9L36
MID089:	CACGTAGATC	Corresponding name:	9L89
MID093:	CGAGACGCGC	Corresponding name:	9L93
MID160:	CGTGAGCTGA	Corresponding name:	9L160
MID164:	CGCTATCTAT	Corresponding name:	9L164
MID172:	CGAGCTAGCT	Corresponding name:	9L172
MID173:	CACTGATGTC	Corresponding name:	9L173

Number of raw reads found: 45890

Number of kept reads after MID filtering : 43003

Number of reads kept for the 7L21 sample :	3038
Number of reads kept for the 7L58 sample :	2856
Number of reads kept for the 7L90 sample :	2246
Number of reads kept for the 7L94 sample :	3066
Number of reads kept for the 7L127 sample :	2572
Number of reads kept for the 7L131 sample :	3016
Number of reads kept for the 7L161 sample :	2703
Number of reads kept for the 7L170 sample :	2894
Number of reads kept for the 9L13 sample :	2963
Number of reads kept for the 9L36 sample :	2273
Number of reads kept for the 9L89 sample :	2351
Number of reads kept for the 9L93 sample :	2631
Number of reads kept for the 9L160 sample :	2769
Number of reads kept for the 9L164 sample :	2917
Number of reads kept for the 9L172 sample :	2184
Number of reads kept for the 9L173 sample :	2524

Bilan : Le nombre de reads obtenu est assez homogène pour tous les échantillons, ce qui en soit est déjà rassurant. De plus un grand nombre de reads est conservé après recherche des MIDs. La majorité de ces séquences est donc de bonne qualité.

STEP 02 : Preprocessing des séquences (paramètres : N=0, length supérieure à 350, oligo sens : CGATAACGAACGAGACCT, oligo antisens : ANCCATTCAATCGGTANT, trois types de stringence évalués).

File name	Raw reads	After length filter	After ambiguities filter	STRINGENCY		
				HIGH	MEDIUM	LOW
7L127	2572	1619	1602	1268	1396	1558
7L131	3016	1894	1880	1455	1587	1803
7L161	2703	1771	1746	1402	1523	1681
7L170	2894	1758	1741	1362	1480	1674
7L21	3038	1915	1888	1468	1590	1791
7L58	2856	1750	1722	1291	1424	1628
7L90	2246	1547	1527	1165	1283	1453
7L94	3066	1976	1941	1483	1607	1823
9L13	2963	1809	1794	1385	1521	1735
9L160	2769	1759	1749	1378	1498	1687
9L164	2917	1818	1804	1419	1544	1723
9L172	2184	1416	1406	1111	1209	1349
9L173	2524	1651	1639	1263	1387	1566
9L36	2273	1464	1448	1111	1217	1364
9L89	2351	1602	1591	1255	1362	1512
9L93	2631	1775	1753	1377	1480	1659

Bilan : Ces résultats démontrent une assez bonne efficacité du séquençage, avec une perte assez importante de reads, tout de même, en terme de taille. On note ainsi une perte d'environ 50% après ce traitement initial. Et cela que ce soit avec une stringence élevée (recherche des deux primers parfaits complets), ou faible (un seul primer, le proximal, recherché parfait complet). Cependant, cette faible qualité est probablement liée à l'étape d'emPCR qui a été réalisées avec 10cpb, ce qui est trop élevé pour ce type de produit.

STEP 03 : Dé répliation stricte des séquences de chaque jeu afin de diminuer le nombre de reads à analyser tout en ne perdant pas d'information. (Programme PERL : *dereplicated_fasta_opt_serv.pl* sur le serveur).

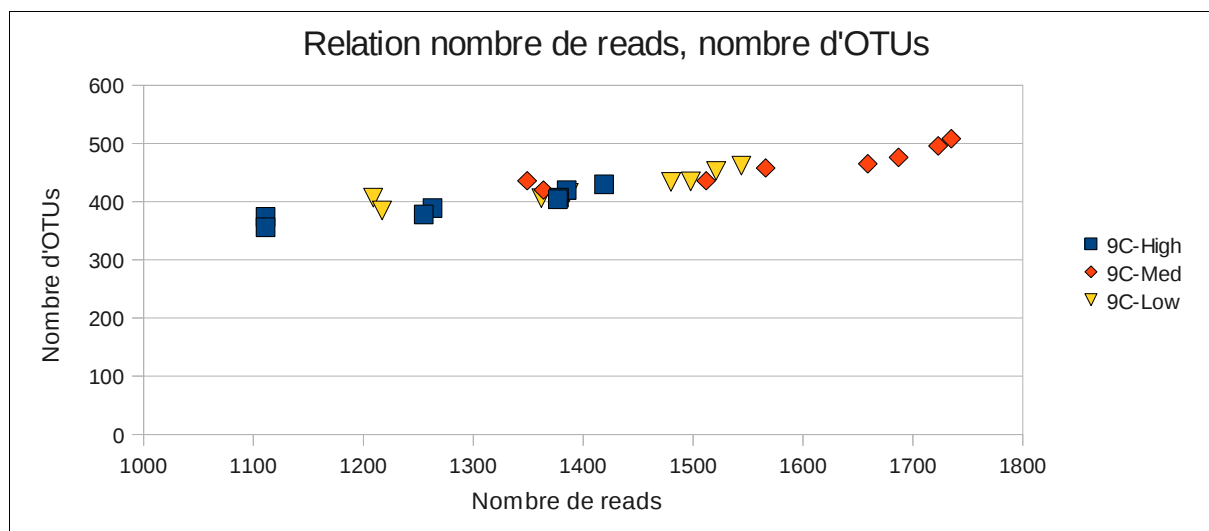
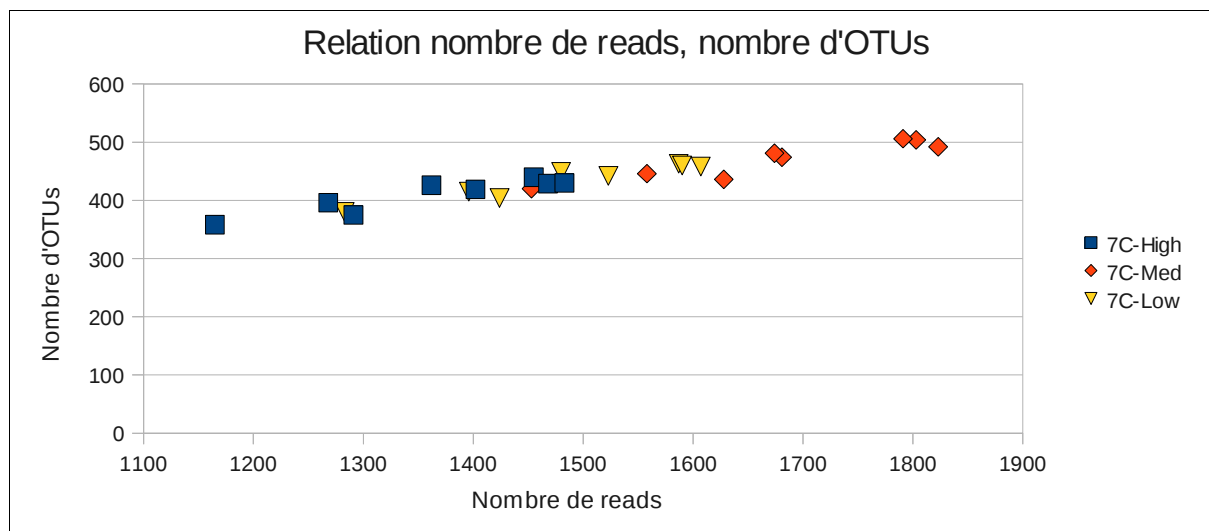
File Name	Totalnbreads	Totalnbdereplicatedreads	% Dereplication
Hpreprocess_7L127	1268	1175	7,33
Hpreprocess_7L131	1455	1310	9,97
Hpreprocess_7L161	1402	1254	10,56
Hpreprocess_7L170	1362	1234	9,40
Hpreprocess_7L21	1468	1304	11,17
Hpreprocess_7L58	1291	1150	10,92
Hpreprocess_7L90	1165	1059	9,10
Hpreprocess_7L94	1483	1320	10,99
Hpreprocess_9L13	1385	1246	10,04
Hpreprocess_9L160	1378	1230	10,74
Hpreprocess_9L164	1419	1260	11,21
Hpreprocess_9L172	1111	1018	8,37
Hpreprocess_9L173	1263	1157	8,39
Hpreprocess_9L36	1111	1025	7,74
Hpreprocess_9L89	1255	1130	9,96
Hpreprocess_9L93	1377	1200	12,85
Lpreprocess_7L127	1558	1449	7,00
Lpreprocess_7L131	1803	1640	9,04
Lpreprocess_7L161	1681	1516	9,82
Lpreprocess_7L170	1674	1529	8,66
Lpreprocess_7L21	1791	1613	9,94
Lpreprocess_7L58	1628	1476	9,34
Lpreprocess_7L90	1453	1335	8,12
Lpreprocess_7L94	1823	1649	9,54
Lpreprocess_9L13	1735	1586	8,59
Lpreprocess_9L160	1687	1527	9,48
Lpreprocess_9L164	1723	1547	10,21
Lpreprocess_9L172	1349	1244	7,78
Lpreprocess_9L173	1566	1446	7,66
Lpreprocess_9L36	1364	1264	7,33
Lpreprocess_9L89	1512	1377	8,93
Lpreprocess_9L93	1659	1466	11,63
Mpreprocess_7L127	1396	1289	7,66
Mpreprocess_7L131	1587	1425	10,21
Mpreprocess_7L161	1523	1362	10,57
Mpreprocess_7L170	1480	1337	9,66
Mpreprocess_7L21	1590	1412	11,19
Mpreprocess_7L58	1424	1273	10,60
Mpreprocess_7L90	1283	1165	9,20
Mpreprocess_7L94	1607	1434	10,77
Mpreprocess_9L13	1521	1373	9,73
Mpreprocess_9L160	1498	1340	10,55
Mpreprocess_9L164	1544	1373	11,08
Mpreprocess_9L172	1209	1106	8,52
Mpreprocess_9L173	1387	1268	8,58
Mpreprocess_9L36	1217	1117	8,22
Mpreprocess_9L89	1362	1227	9,91
Mpreprocess_9L93	1480	1290	12,84

Bilan : Comme le montre ces résultats de dé répliation, les échantillons sont assez diversifié, et le nombre de séquences étant assez faible, le taux de dé répliation est lui aussi assez bas (entre 7 et

13% environ). Cela risque d'engendrer quelques difficultés dans la réalisation des clustering.

STEP 04 : Alignement des séquences dé répliquées avec INFERNAL disponible sur le pipeline de RDP.

STEP 05 : Clustering des différents jeux de données à 5,0% de dissimilarité (correspondant au niveau du genre sur une séquence de 400 bases) des séquences alignées à l'aide du programme de clustering développé en local (programme prenant en compte les erreurs d'homopolymères et les ignorant le cas échéant pour les clustering ainsi que les différences de tailles entre séquences, et ignorant la partie non commune aux deux séquences).



Bilan : Si l'on regarde la relation entre le nombre de reads et d'OTUs, que ce soit pour 7 ou 9 cycles, on voit clairement une relation linéaire entre les deux, confortant l'idée qu'il y n'a pas de biais liés aux différents MIDs utilisés. Cela sera bien sûr confirmé par d'autres analyses.

STEP 06 : Élimination des séquences dites de mauvaise qualité à l'aide d'un outil proche de celui développé par Richard Christen. Le clustering utilisé est celui réalisé à 5,0%.

Note: Cette étape consiste à éliminer les reads « rares » (appelés singletons) qui forment à eux seuls un OTU au niveau de dissimilarité défini (appelés donc single-singletons).

STEP 07 : Assignment taxonomique des données *via USEARCH* contre la base de données SILVA installée en local pour les séquences mises de côté dites 'del' (supprimées des jeux complets) et application d'un programme PERL qui, en analysant les assignments taxonomiques, réinjectent dans le jeu de données les séquences ayant une assignment au niveau du phylum supérieure à 90% d'identité.

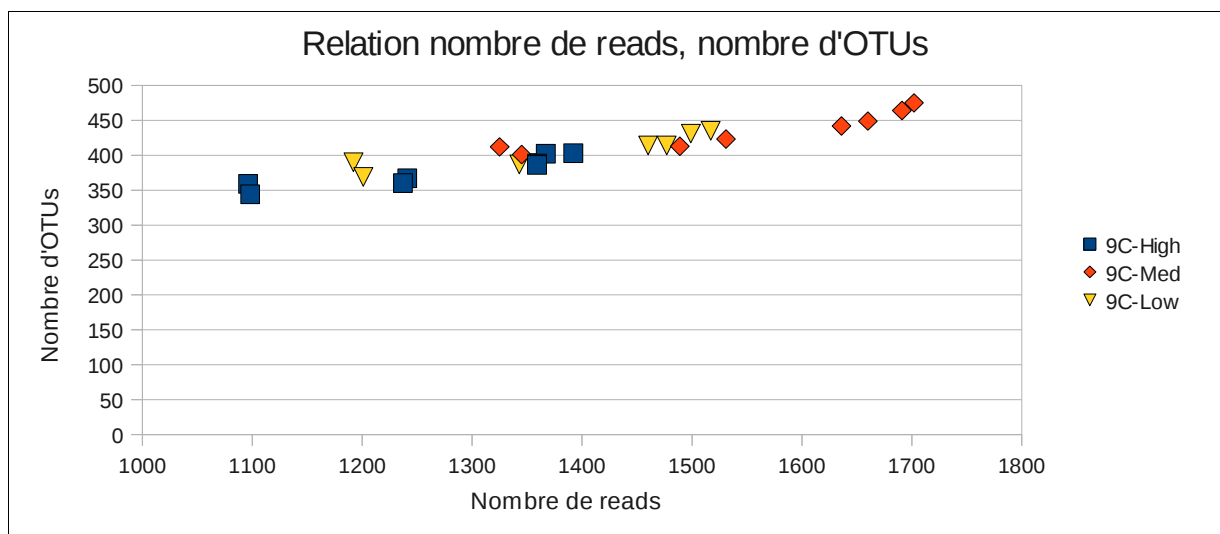
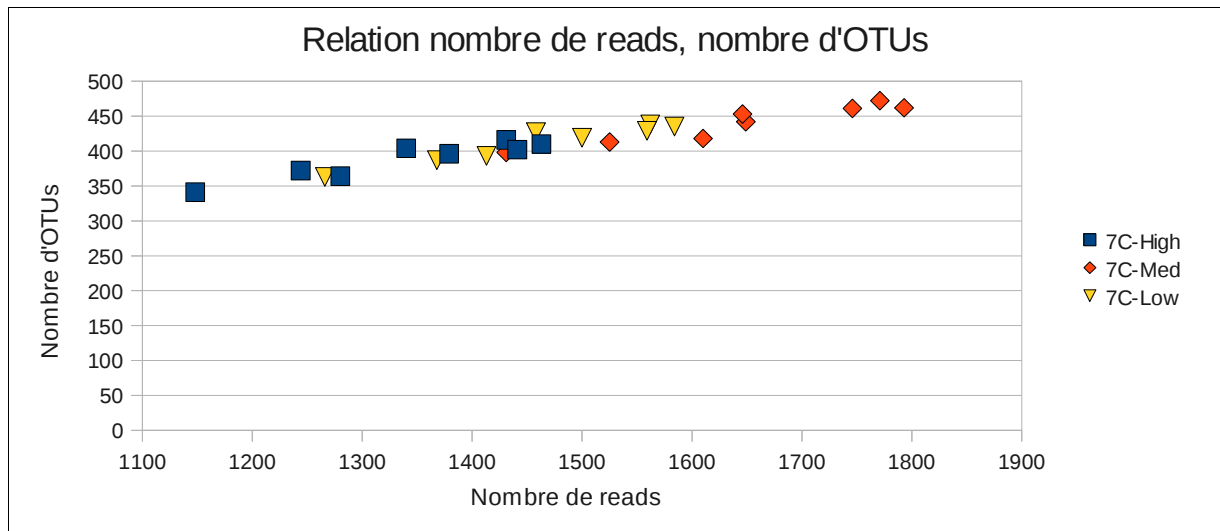
Filename	Nbreads	NbOTUs	Nbreads/NbOTUs	HUNTING			RECOVERING SILVA 16S	
				Nbreads deleted	Nbreads kept	% de délétion	Recovered (90%)	% de délétion finale
Hpreprocess_7L127	1268	396	3,20	249	1019	19,64	225	1,89
Hpreprocess_7L131	1455	440	3,31	286	1169	19,66	262	1,65
Hpreprocess_7L161	1402	419	3,35	270	1132	19,26	247	1,64
Hpreprocess_7L170	1362	426	3,20	278	1084	20,41	256	1,62
Hpreprocess_7L21	1468	429	3,42	271	1197	18,46	244	1,84
Hpreprocess_7L58	1291	375	3,44	223	1068	17,27	212	0,85
Hpreprocess_7L90	1165	358	3,25	216	949	18,54	199	1,46
Hpreprocess_7L94	1483	430	3,45	271	1212	18,27	251	1,35
Hpreprocess_9L13	1385	420	3,30	248	1137	17,91	230	1,30
Hpreprocess_9L160	1378	407	3,39	247	1131	17,92	228	1,38
Hpreprocess_9L164	1419	430	3,30	266	1153	18,75	239	1,90
Hpreprocess_9L172	1111	374	2,97	249	862	22,41	234	1,35
Hpreprocess_9L173	1263	389	3,25	236	1027	18,69	214	1,74
Hpreprocess_9L36	1111	356	3,12	219	892	19,71	206	1,17
Hpreprocess_9L89	1255	378	3,32	236	1019	18,80	218	1,43
Hpreprocess_9L93	1377	404	3,41	245	1132	17,79	227	1,31
Lpreprocess_7L127	1558	446	3,49	271	1287	17,39	238	2,12
Lpreprocess_7L131	1803	504	3,58	317	1486	17,58	285	1,77
Lpreprocess_7L161	1681	474	3,55	306	1375	18,20	274	1,90
Lpreprocess_7L170	1674	481	3,48	298	1376	17,80	270	1,67
Lpreprocess_7L21	1791	506	3,54	325	1466	18,15	280	2,51
Lpreprocess_7L58	1628	436	3,73	256	1372	15,72	238	1,11
Lpreprocess_7L90	1453	420	3,46	246	1207	16,93	224	1,51
Lpreprocess_7L94	1823	492	3,71	304	1519	16,68	274	1,65
Lpreprocess_9L13	1735	508	3,42	299	1436	17,23	266	1,90
Lpreprocess_9L160	1687	476	3,54	286	1401	16,95	259	1,60
Lpreprocess_9L164	1723	496	3,47	304	1419	17,64	272	1,86
Lpreprocess_9L172	1349	436	3,09	280	1069	20,76	256	1,78
Lpreprocess_9L173	1566	458	3,42	288	1278	18,39	253	2,23
Lpreprocess_9L36	1364	420	3,25	269	1095	19,72	250	1,39
Lpreprocess_9L89	1512	436	3,47	266	1246	17,59	243	1,52
Lpreprocess_9L93	1659	465	3,57	282	1377	17,00	259	1,39
Mpreprocess_7L127	1396	415	3,36	254	1142	18,19	226	2,01
Mpreprocess_7L131	1587	463	3,43	298	1289	18,78	273	1,58
Mpreprocess_7L161	1523	442	3,45	282	1241	18,52	259	1,51
Mpreprocess_7L170	1480	449	3,30	286	1194	19,32	264	1,49
Mpreprocess_7L21	1590	460	3,46	296	1294	18,62	265	1,95
Mpreprocess_7L58	1424	404	3,52	239	1185	16,78	228	0,77
Mpreprocess_7L90	1283	380	3,38	224	1059	17,46	207	1,33
Mpreprocess_7L94	1607	458	3,51	288	1319	17,92	265	1,43
Mpreprocess_9L13	1521	453	3,36	264	1257	17,36	242	1,45
Mpreprocess_9L160	1498	435	3,44	270	1228	18,02	249	1,40
Mpreprocess_9L164	1544	462	3,34	287	1257	18,59	260	1,75
Mpreprocess_9L172	1209	407	2,97	270	939	22,33	253	1,41
Mpreprocess_9L173	1387	415	3,34	258	1129	18,60	232	1,87
Mpreprocess_9L36	1217	385	3,16	244	973	20,05	228	1,31
Mpreprocess_9L89	1362	406	3,35	252	1110	18,50	233	1,40
Mpreprocess_9L93	1480	434	3,41	268	1212	18,11	248	1,35

Bilan : Au final, cette étape ne permet d'éliminer qu'un faible nombre de reads sur tous les échantillons évalués. Il serait probablement intéressant de tester d'autres paramètres plus stringents, voir de faire l'analyse sans ces reads. A évaluer en fonction des besoins.

STEP 08 : Alignement des séquences dé répliquées et curées avec INFERNAL.

STEP 09 : Clustering des différents jeux de données à 5.0% de dissimilarité des séquences alignées (programme prenant en compte les erreurs d'homopolymères et les ignorant le cas échéant pour les clustering ainsi que les différences de tailles entre séquences, et ignorant la partie non commune aux deux séquences).

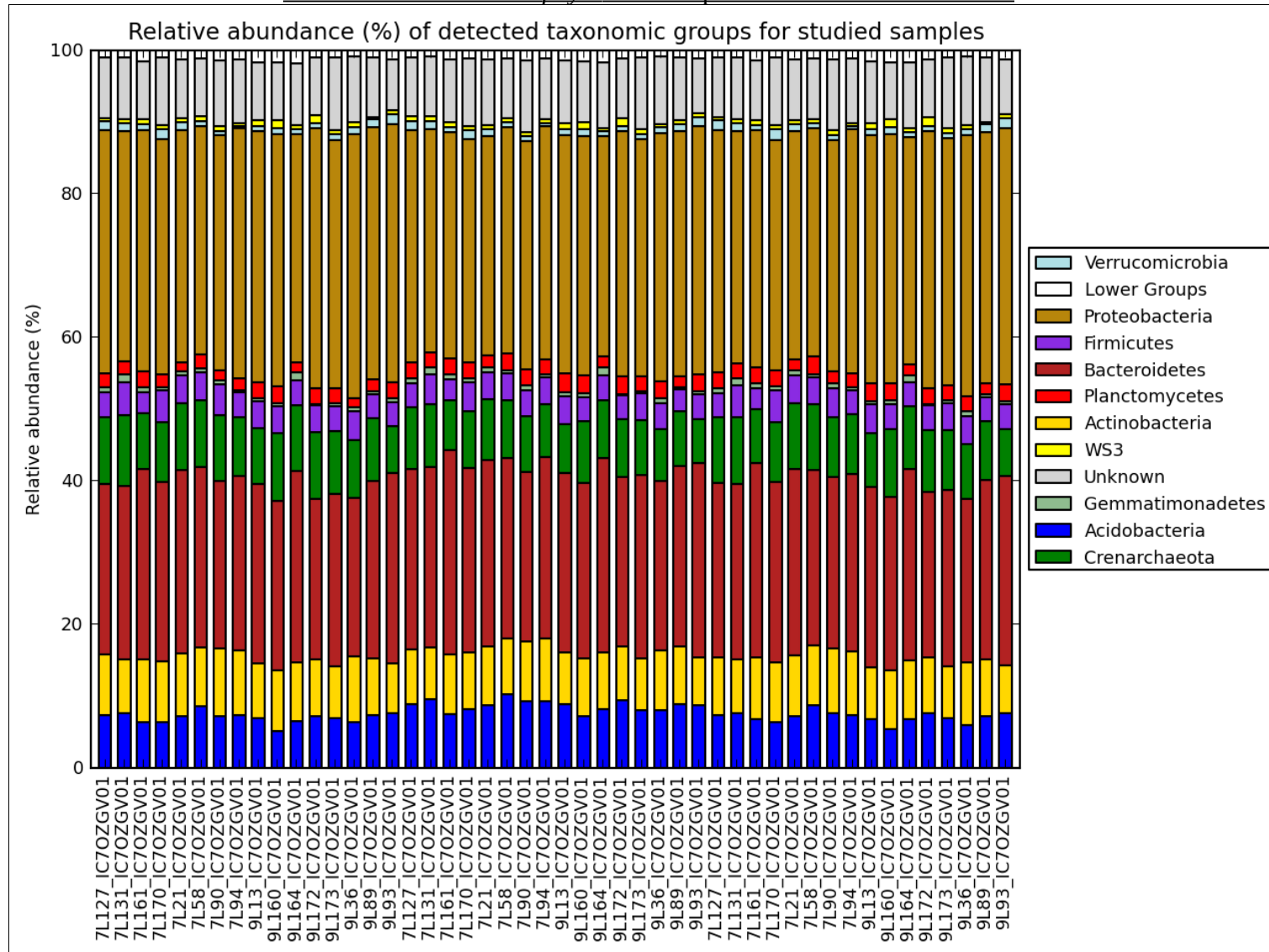
Filename	Nbreads	NbOTUs	Nbreads/NbOTUs
Hpreprocess_7L127	1244	372	3,34
Hpreprocess_7L131	1431	416	3,44
Hpreprocess_7L161	1379	396	3,48
Hpreprocess_7L170	1340	404	3,32
Hpreprocess_7L21	1441	402	3,58
Hpreprocess_7L58	1280	364	3,52
Hpreprocess_7L90	1148	341	3,37
Hpreprocess_7L94	1463	410	3,57
Hpreprocess_9L13	1367	402	3,40
Hpreprocess_9L160	1359	388	3,50
Hpreprocess_9L164	1392	403	3,45
Hpreprocess_9L172	1096	359	3,05
Hpreprocess_9L173	1241	367	3,38
Hpreprocess_9L36	1098	344	3,19
Hpreprocess_9L89	1237	360	3,44
Hpreprocess_9L93	1359	386	3,52
Lpreprocess_7L127	1525	413	3,69
Lpreprocess_7L131	1771	472	3,75
Lpreprocess_7L161	1649	442	3,73
Lpreprocess_7L170	1646	453	3,63
Lpreprocess_7L21	1746	461	3,79
Lpreprocess_7L58	1610	418	3,85
Lpreprocess_7L90	1431	398	3,60
Lpreprocess_7L94	1793	462	3,88
Lpreprocess_9L13	1702	475	3,58
Lpreprocess_9L160	1660	449	3,70
Lpreprocess_9L164	1691	464	3,64
Lpreprocess_9L172	1325	412	3,22
Lpreprocess_9L173	1531	423	3,62
Lpreprocess_9L36	1345	401	3,35
Lpreprocess_9L89	1489	413	3,61
Lpreprocess_9L93	1636	442	3,70
Mpreprocess_7L127	1368	387	3,53
Mpreprocess_7L131	1562	438	3,57
Mpreprocess_7L161	1500	419	3,58
Mpreprocess_7L170	1458	427	3,41
Mpreprocess_7L21	1559	429	3,63
Mpreprocess_7L58	1413	393	3,60
Mpreprocess_7L90	1266	363	3,49
Mpreprocess_7L94	1584	435	3,64
Mpreprocess_9L13	1499	431	3,48
Mpreprocess_9L160	1477	414	3,57
Mpreprocess_9L164	1517	435	3,49
Mpreprocess_9L172	1192	390	3,06
Mpreprocess_9L173	1361	389	3,50
Mpreprocess_9L36	1201	369	3,25
Mpreprocess_9L89	1343	387	3,47
Mpreprocess_9L93	1460	414	3,53



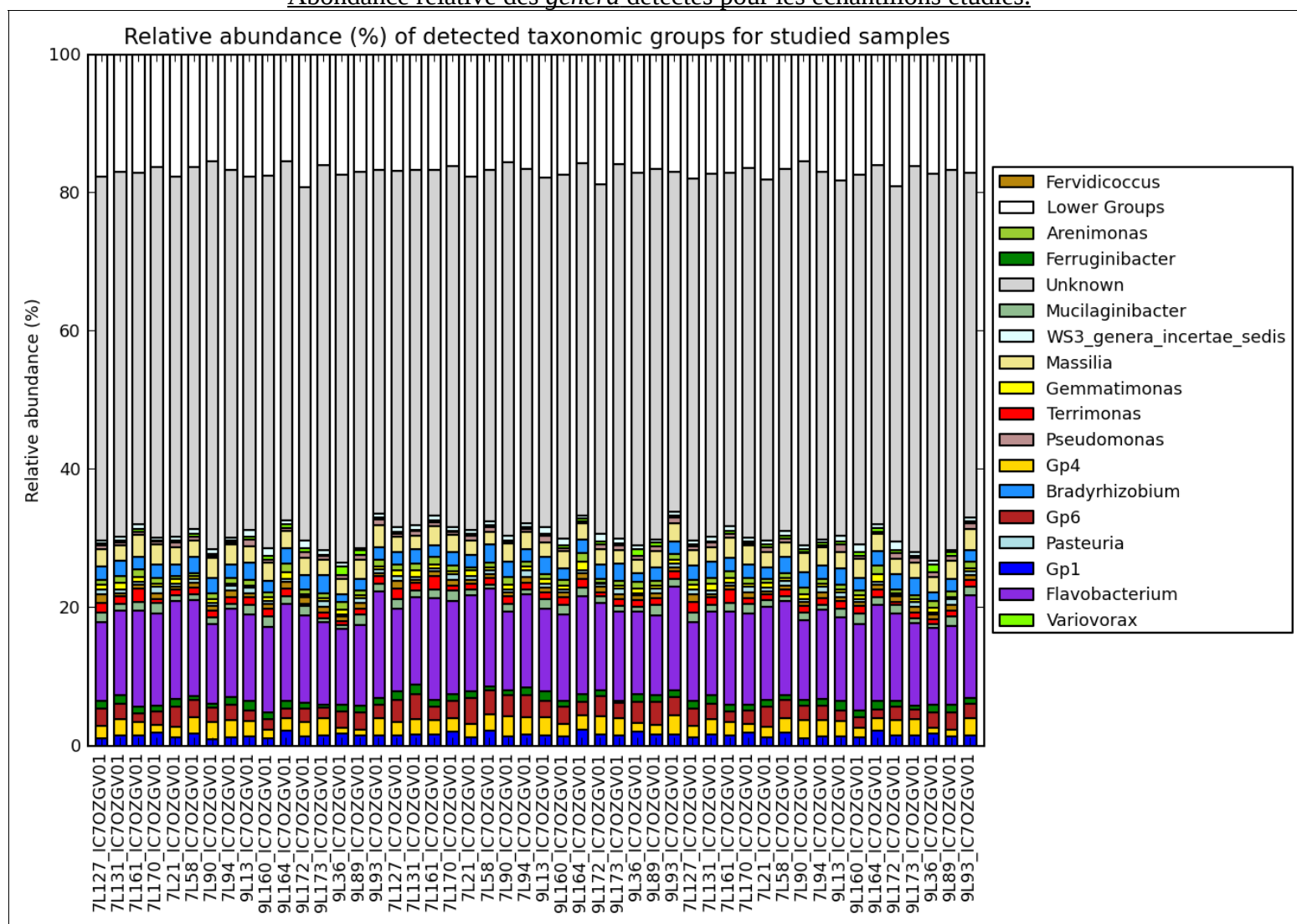
Bilan : Si l'on regarde la relation entre le nombre de reads et d'OTUs, que ce soit pour 7 ou 9 cycles, on voit clairement une relation linéaire entre les deux, confortant l'idée qu'il y n'a pas de biais liés aux différents MIDs utilisés. Cela sera bien sûr confirmé par d'autres analyses. Cette relation semble tout de même plus stable à 7 cycles qu'à 9 cycles, bien qu'il soit difficile avec si peu de séquences de conclure précisément.

STEP 10 : Assignment taxonomique des séquences dé répliquées et curées contre la base de données RDP installée en local. L'assignation taxonomique a été réalisée à 80% de similarité pour les séquences.

Abondance relative des *phyla* détectés pour les échantillons étudiés.



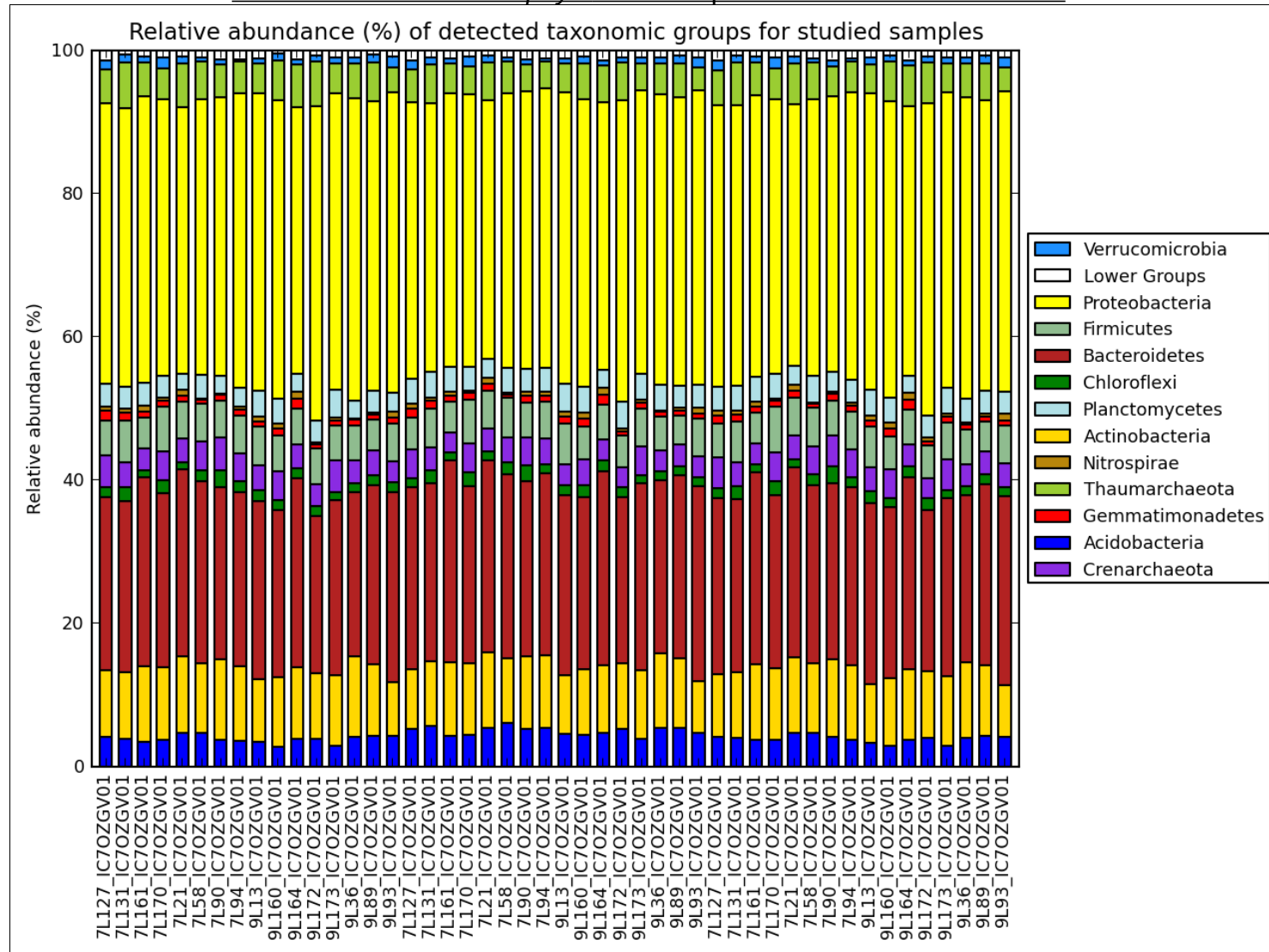
Abondance relative des *genera* détectés pour les échantillons étudiés.



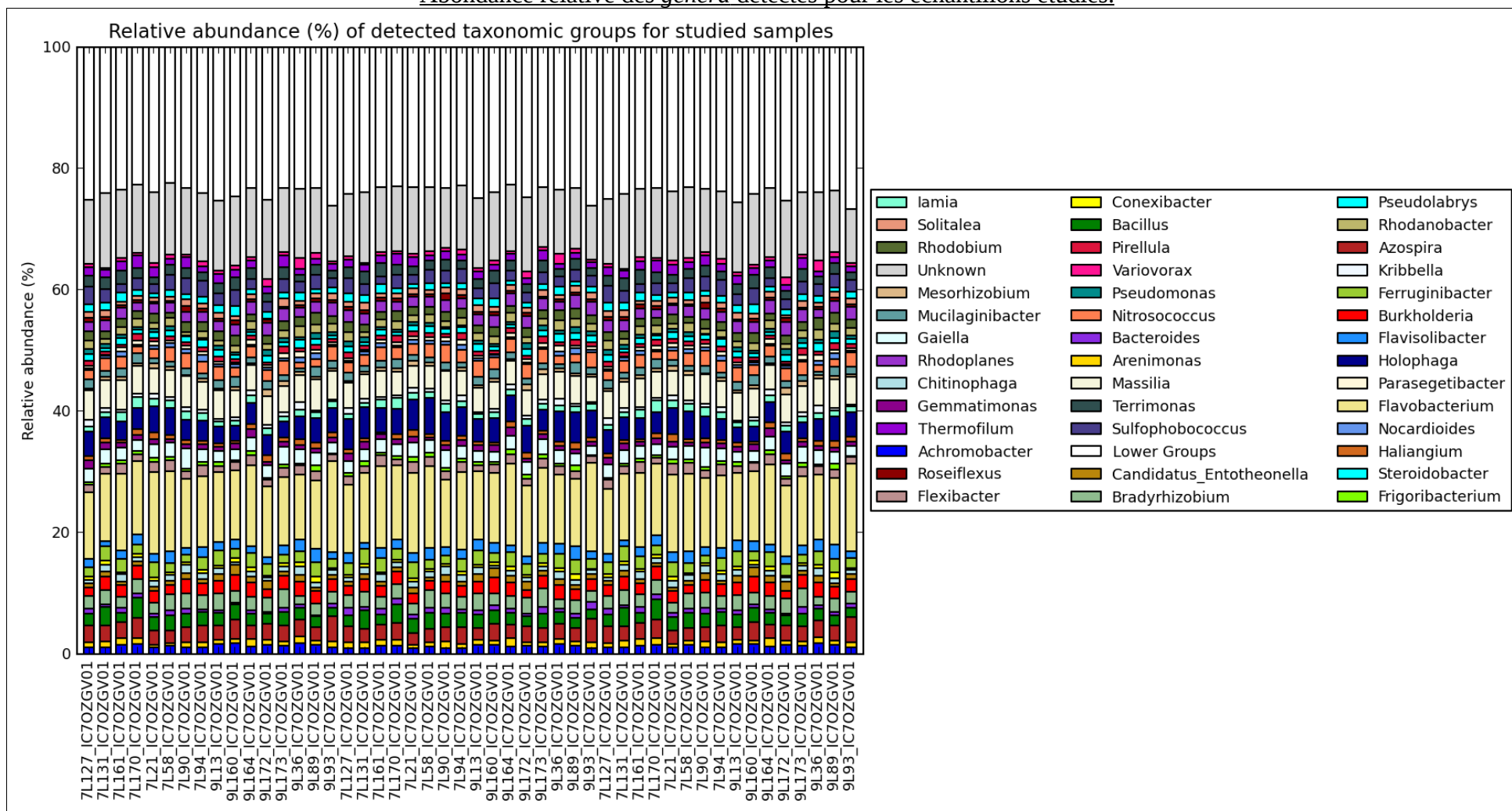
Bilan : Comme le montre ces résultats, il y a une très forte homogénéité des résultats, à tous les niveaux taxonomiques ou de stringence, avec 7 ou 9 cycles. Il n'y a donc que peu ou pas d'effet MID sur les données obtenues, surtout avec un aussi faible nombre de séquence par échantillon.

STEP 10a : Assignment taxonomique des séquences dé répliquées et curées contre la base de données SILVA installée en local avec l'outil USEARCH. L'assignation taxonomique a été réalisée à 80% de similarité pour les séquences.

Abondance relative des *phyla* détectés pour les échantillons étudiés.

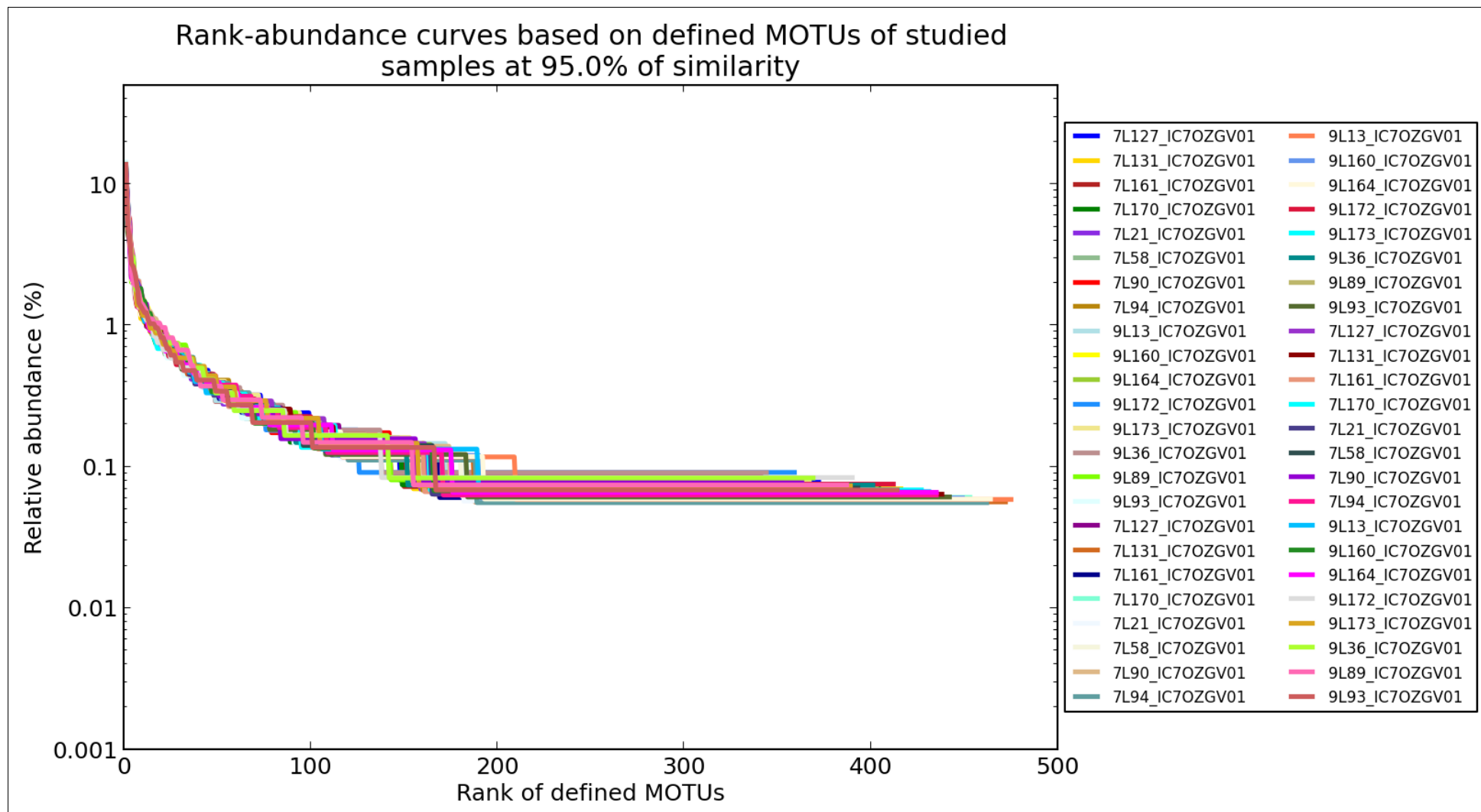


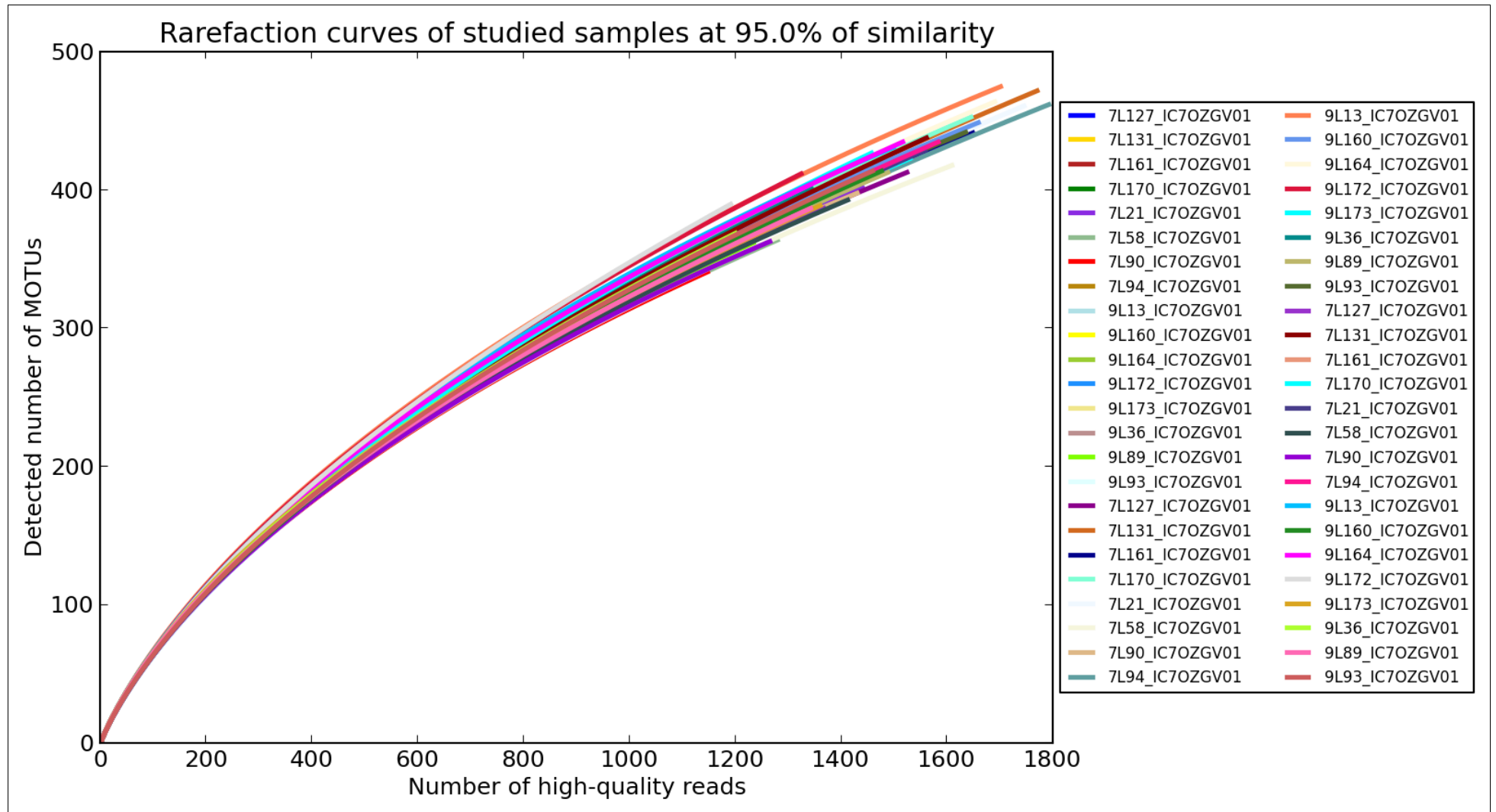
Abondance relative des *genera* détectés pour les échantillons étudiés.



Bilan : Comme le montre ces résultats, il y a une très forte homogénéité des résultats, à tous les niveaux taxonomiques ou de stringence, avec 7 ou 9 cycles. Il n'y a donc que peu ou pas d'effet MID sur les données obtenues, surtout avec un aussi faible nombre de séquence par échantillon. Cependant, par rapport à précédemment, on note qu'une grande partie des séquences est plus facilement annotée, mais aussi différemment annotée.

STEP 11 : Calculs des courbes de raréfaction, et de d'abondance des OTUs, à 95 % de similarité.





Bilan : comme le montre les courbes obtenues, très peu de différences existent entre les différents échantillons analysés, quel que soit la stringence ou le nombre de cycles utilisé, et cela même avec un faible nombre de reads analysés.

Il est donc visible qu'il n'y a pas d'effet MID sur ces 16 MIDs testés, et que le nombre de cycles PCRs ne semblent pas jouer sur les profils des communautés bactériennes obtenues.