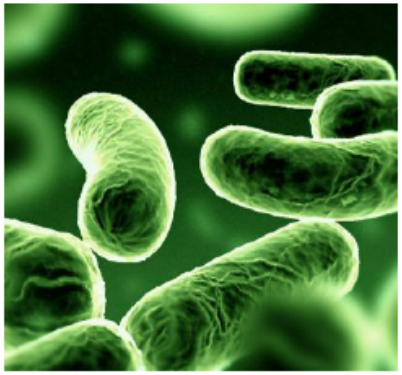
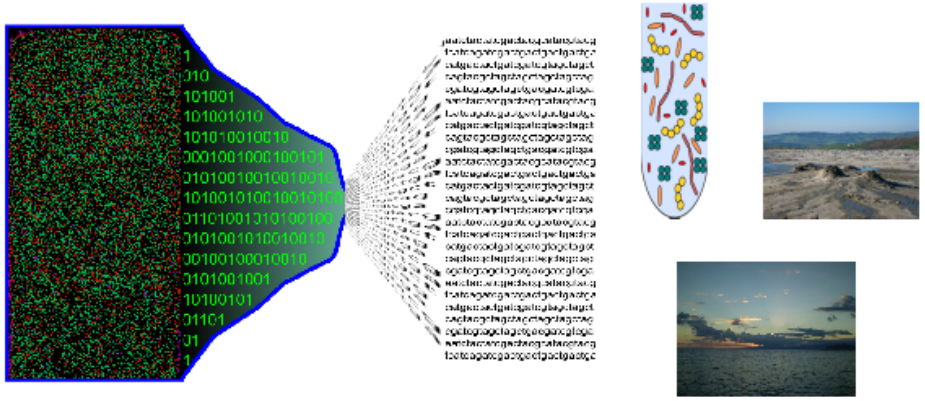
Introduction into the processing of raw data

Giuseppe D'Auria

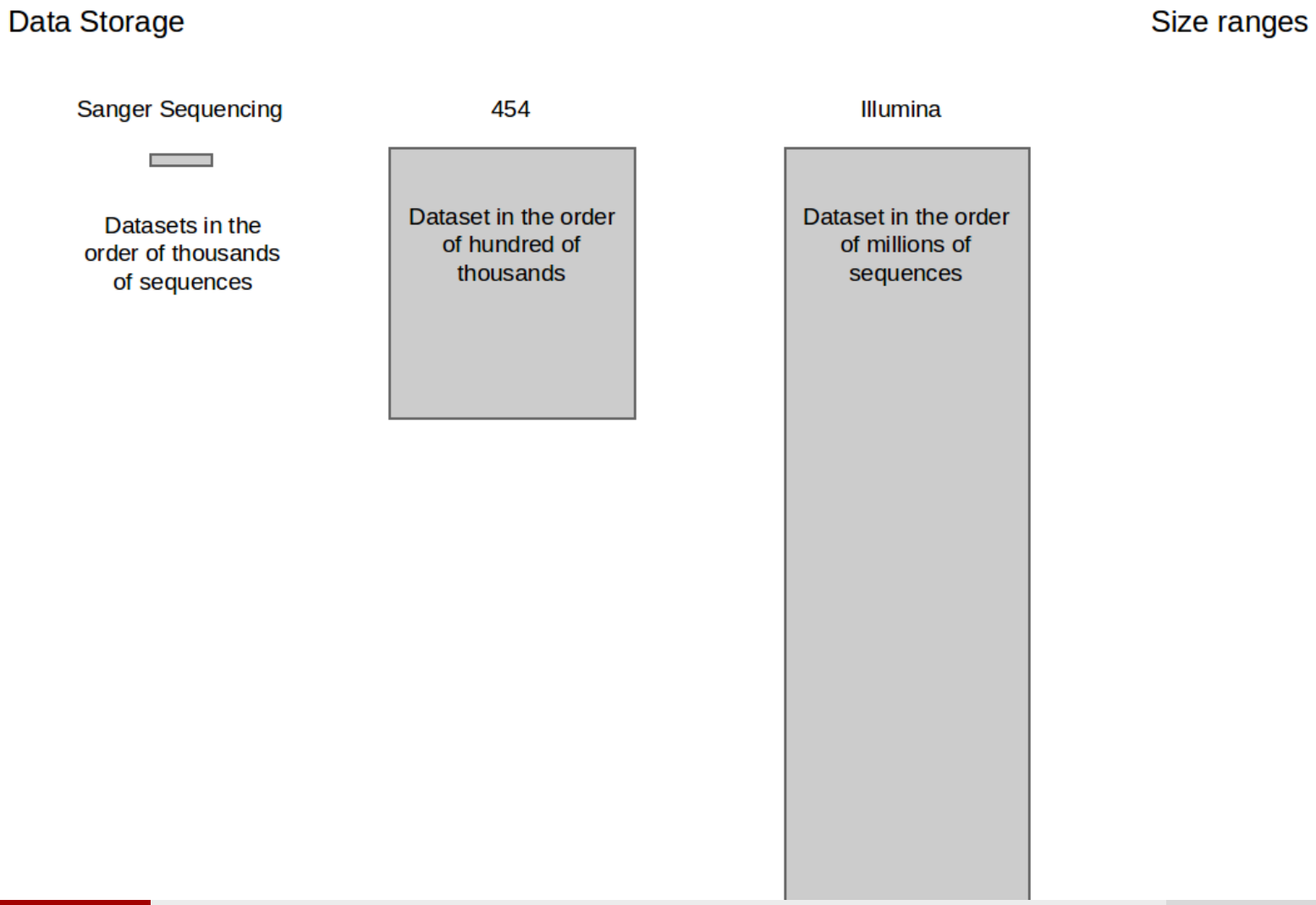




FISABIO, Valencia

Norwich 12-16 October 2015

Considering needed storage space by technology



Considering needed storage space by technology

Data Storage

Size ranges

Sanger Sequencing

454

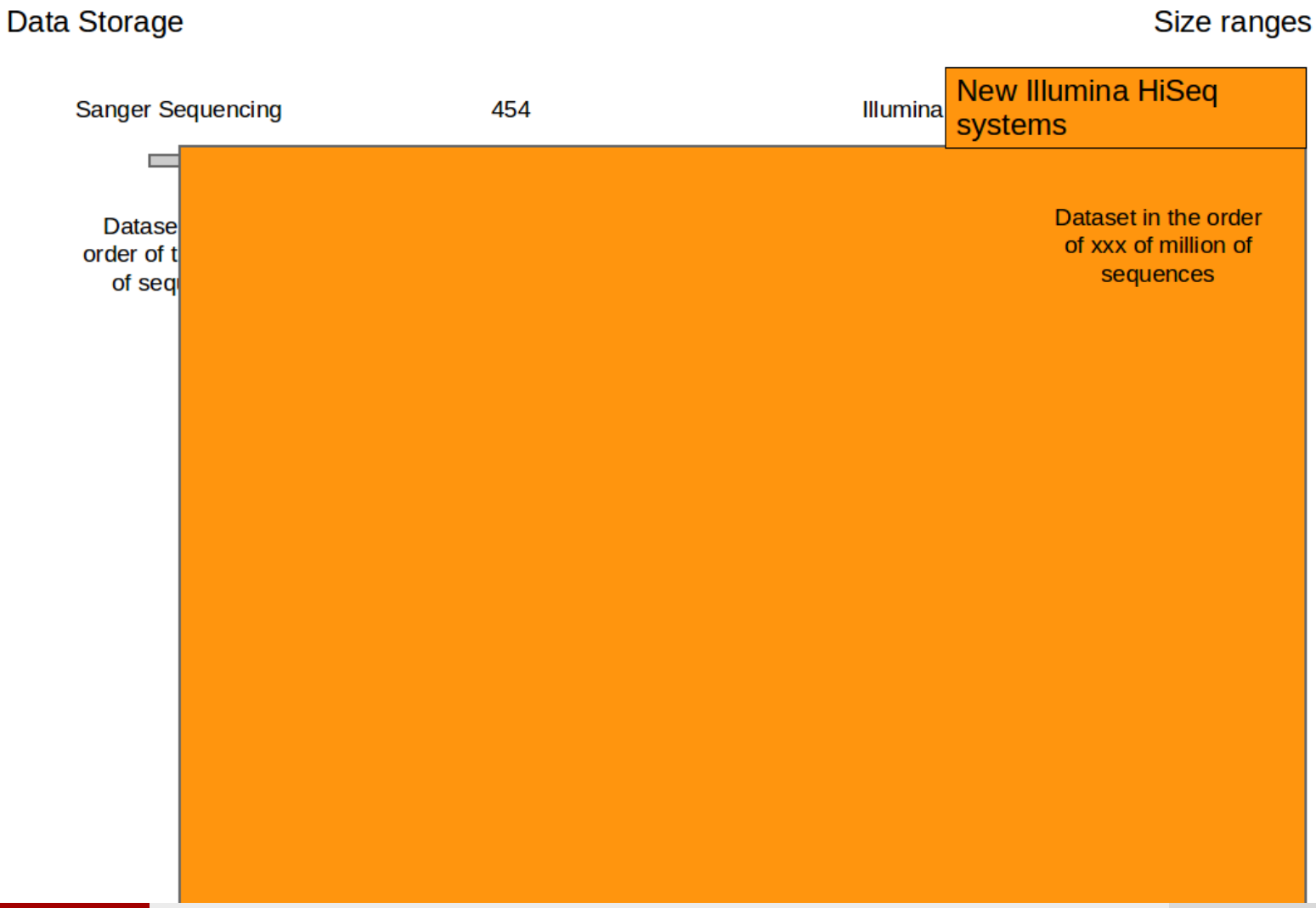
Illumina

Solid

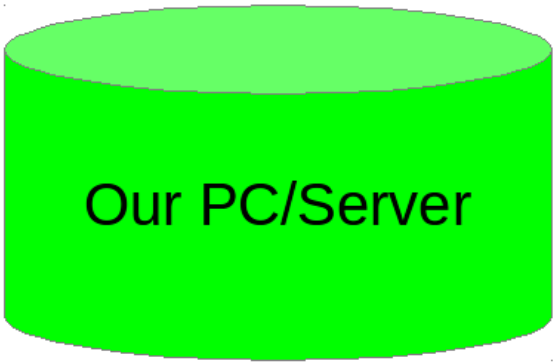
Datase
order of t
of seq

Dataset in the order
of xxx of million of
sequences

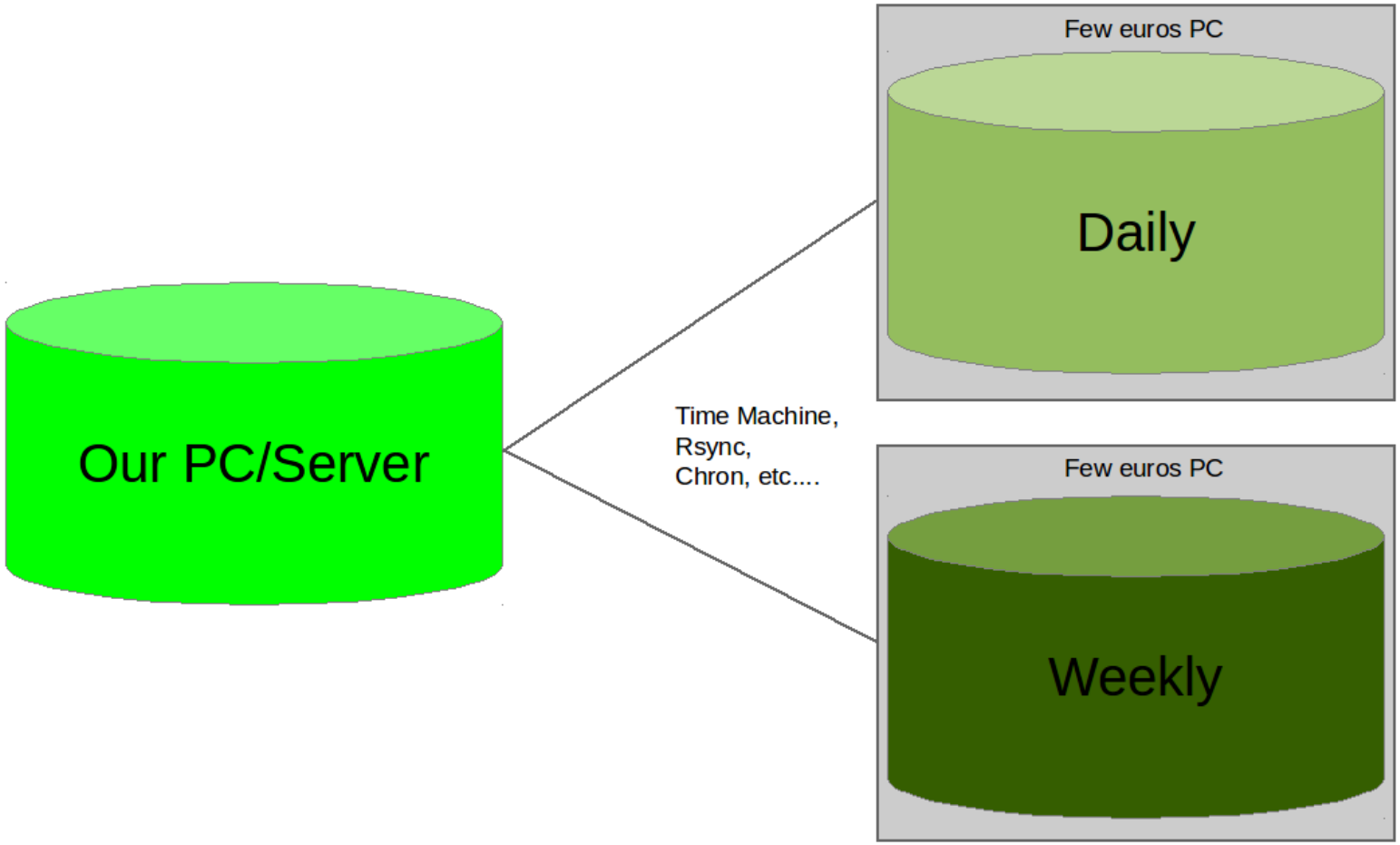
Considering needed storage space by technology



Considering backup



Considering backup



Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunitat Valenciana

Considering backup

Wide range of solutions

BIG Servers, NAS systems, etc..

Thousands of Euro/Dollars

We spend so much money for sequencing, we can save few of them for saving our data

Considering backup

Wide range of solutions

BIG Servers, NAS systems, etc..

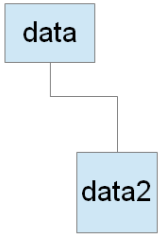
Thousands of Euro/Dollars

~ 50 Euros/Dollars

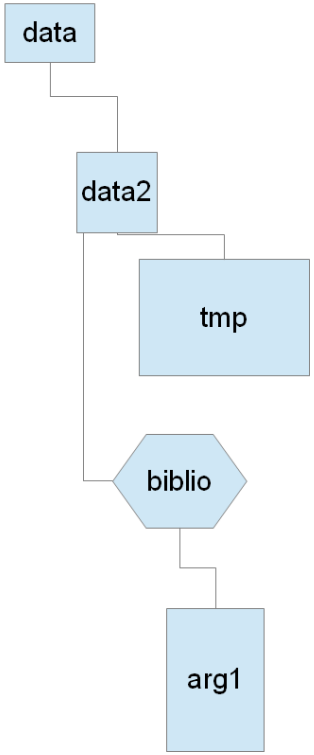
+ some tera of disks

We spend so much money for sequencing, we can save few of them for saving our data

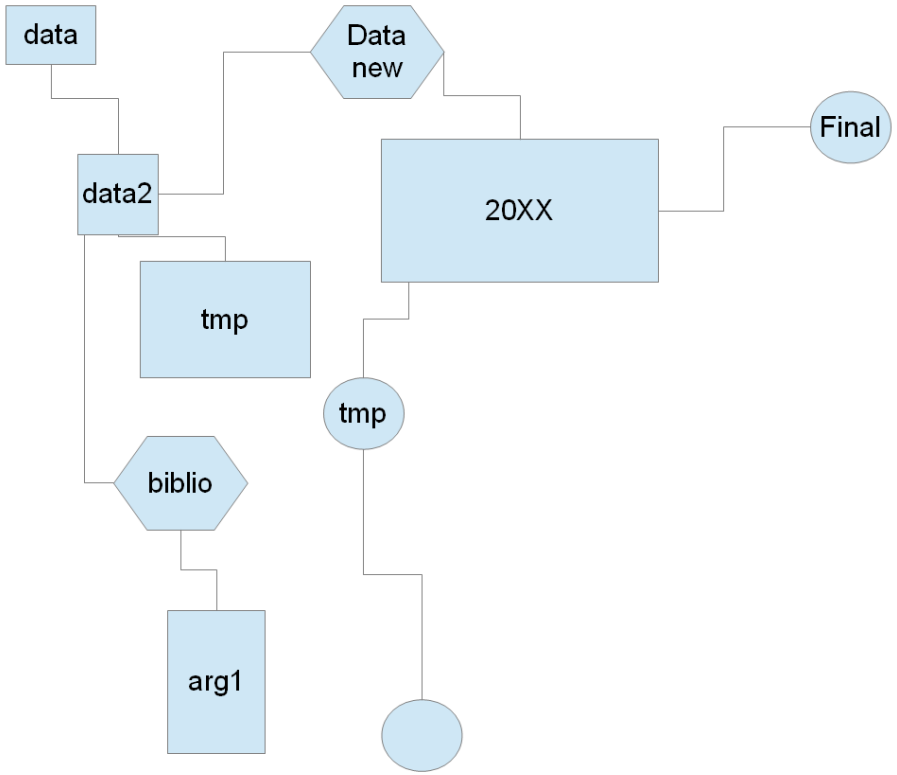
Data storage and disk structure



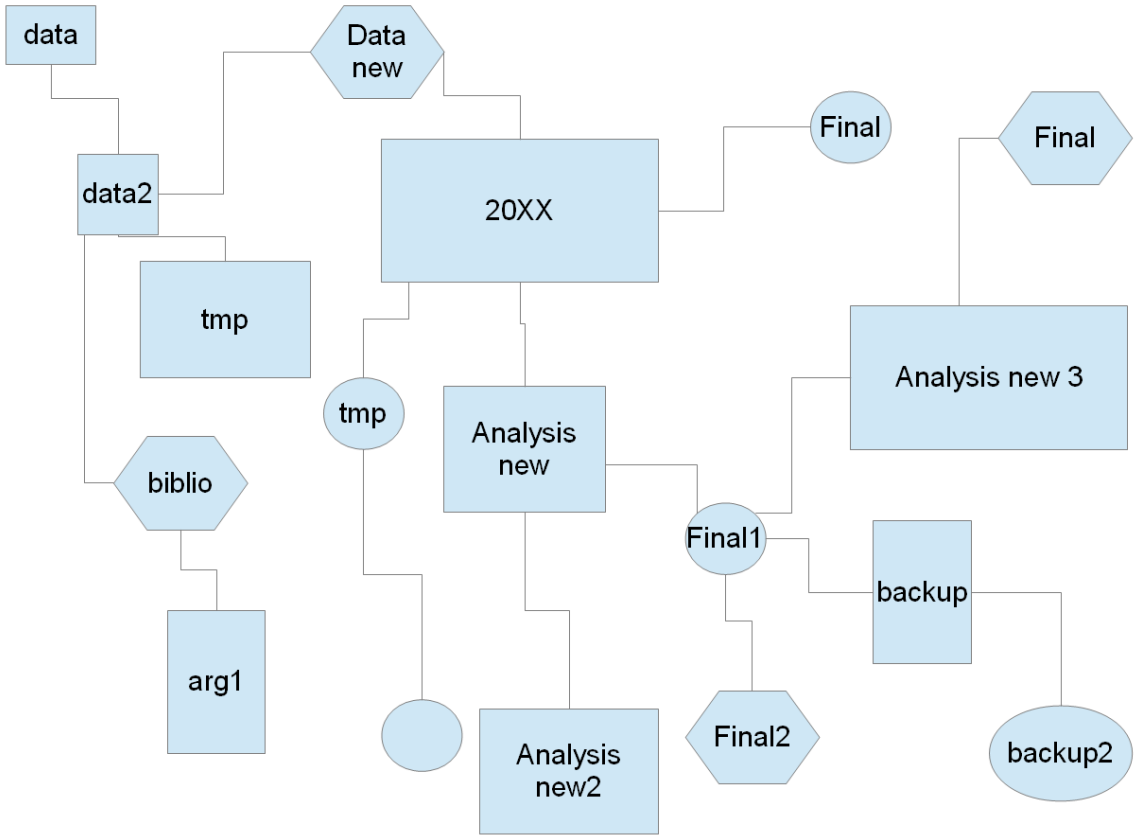
Data storage and disk structure



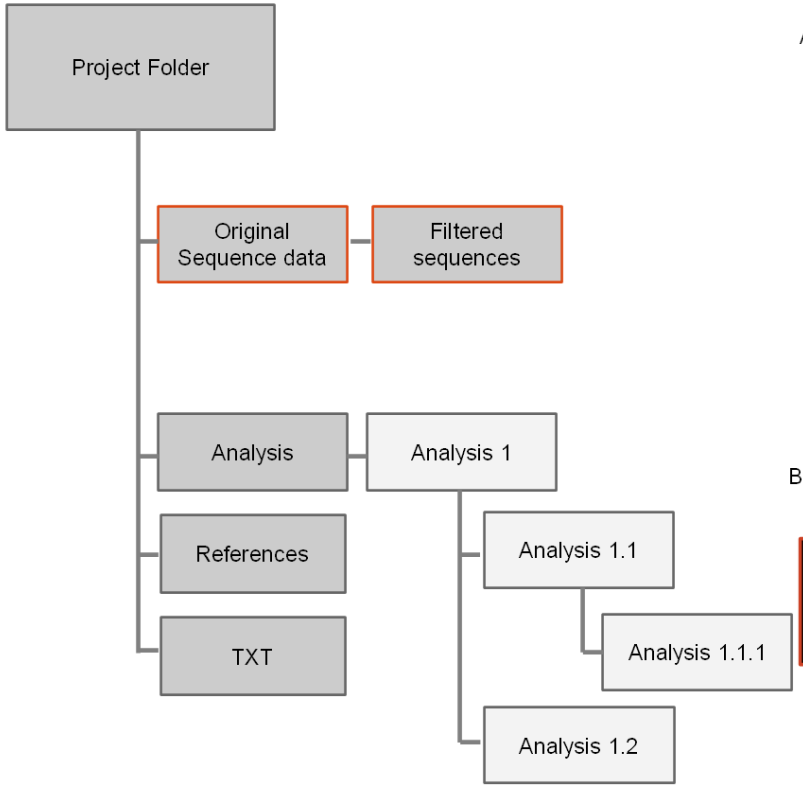
Data storage and disk structure



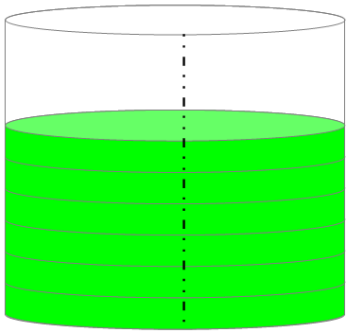
Data storage and disk structure



Data storage and disk structure



AVOID COPYING AND COPYING AND SECURITY COPYING AND AGAIN COPYING not useful data



Better using symbolic links, just pointing to the needed big data files

```
> ln -s TARGET LINK_NAME
```

The system

- Linux or Windows? 
- Both allow good bioinformatics analysis
- Linux is more stable for massive *data crunching* analysis and it is FREE
- Windows is not FREE
- Most of the software work in both systems but several are exclusively working on Linux.
- The best structure for bioinformatics (just my personal advice):
- A Linux Desktop system (Ubuntu – Fedora) +
- A virtual machine (Virtual Box)

Data format

Fasta and FastaQuality format

FASTA

```
>G12OEMT03CWVU1
AGAGTTTGATCATGGCTCAGGATGAACGCTAGCGGCAGGCCTAACACATGCAAGTCGAGGGAGGAG
CCTTCGGGCTTCGACCGGCGTACGGGTGCGTAACG
>G12OEMT03DH3XQ
AGAGTTTGATCATGGCTCAGTGCCAGCCGCCGCGGGAGCGCATTAG
>G12OEMT03DD28C
AGAGTTTGATCCTGGCTCAGGGTGGTCATATGTTTGGAATTGGTGCCAGCCGCCGCGGGAGCGCATT
AG
>G12OEMT03DGQ48
AGAGTTTGATCATGGCTCAGGAGGTGCCAGCAGCCGCGGAGCGCATTAG
>G12OEMT03C0MSF
AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTAGAACGCTGAA
GCTTGGCGCTTGCACCGAGCGGATG
```

Data format

Fasta and FastaQuality format

FASTA

```
>G12OEMT03CWVU1
AGAGTTTGATCATGGCTCAGGATGAACGCTAGCGGCAGGCCTAACACATGCAAGTCGAGGGAGGAG
CCTTCGGGCTTCGACCGGCGTACGGGTGCGTAACG
>G12OEMT03DH3XQ
AGAGTTTGATCATGGCTCAGTGCCAGCCGCCGCGGGAGCGCATTAG
>G12OEMT03DD28C
AGAGTTTGATCCTGGCTCAGGGTGGTCATATGTTTGGAATTGGTGCCAGCCGCCGCGGGAGCGCATT
AG
>G12OEMT03DGQ48
AGAGTTTGATCATGGCTCAGGAGGTGCCAGCAGCCGCGGAGCGCATTAG
>G12OEMT03C0MSF
AGAGTTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTAGAACGCTGAA
GCTTGGCGCTTGCACCGAGCGGATG
```

QUALITY

```
>G12OEMT03CWVU1
40 40 38 30 20 20 20 30 38 36 36 36 36 36 38 40 40 40 40 40 39 38 38 38 34 40 40 40 40 40 40 40 40
40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 39 39 39 39 34 34 35 39 40 40 40 36 39 39 40 40 40 39 39
39 39 40 40 40 39 39 39 40 40 40 40 40 40 40 40 39 39 39 40 40 40 40 39 39 38 35 32 35 40 40 40 40
40 40
>G12OEMT03DH3XQ
40 40 40 38 20 20 20 30 38 40 40 38 36 36 36 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40 40
30 30 30 40 40 40 40 35 35 34 34 39 35
>G12OEMT03DD28C
40 38 37 35 22 22 22 26 31 35 36 33 30 32 33 36 36 30 28 20 18 18 35 27 30 32 32 32 32 27 21 22 16
16 14 19 19 23 23 23 23 23 23 21 24 27 32 27 27 25 27 30 24 24 25 27 26 28 28 32 22 29 27 25 22 20
19 21 27
>G12OEMT03DGQ48
40 40 40 36 21 21 20 30 36 40 40 40 40 36 36 40 40 40 40 40 34 30 21 21 25 26 36 36 40 34 32 32 32
31 31 31 26 23 22 25 20 30 34 25 29 24 29 23 24
>G12OEMT03C0MSF
40 40 36 28 19 19 19 28 31 36 36 36 37 36 40 40 40 40 39 39 39 40 40 40 40 40 40 40 40 40 40 40 40
40 40 40 40 39 39 39 40 40 40 40 40 40 39 35 35 35 35 34 39 40 40 40 40 40 40 40 40 39 39 39 39 39
39 39 39 39 40 40 40 40 40 40 40 39 39 39 40 40 40 40 40 40 40 39 39 39 39
```

Data format

Standard Flowgram Format

```
>G12OEMT03CWZL8
 Run Prefix:   R_2011_05_03_06_02_36_
 Region #:     3
 XY Location:  1078_3006

 Run Name:       R_2011_05_03_06_02_36_FLX07090549_Administrator_RUN19
 Analysis Name:  D_2011_05_04_03_47_28_NAVELINA_signalProcessingAmplicons
 Full Path:      /data/R_2011_05_03_06_02_36_FLX07090549_Administrator_RUN19/D_2011_05_04_03_47_28_NAVELINA_signalProcessingAmplicons/

 Read Header Len:  32
 Name Length:      14
 # of Bases:       518
 Clip Qual Left:   16
 Clip Qual Right:  397
 Clip Adap Left:   0
 Clip Adap Right:  0

Flowgram:  0.04 0.00 0.11 1.01 0.03 1.01 1.02 0.08 0.90 1.16 0.97 0.99 0.06 0.98 0.09 0.95 0.89 1.02 0.09 1.06 0.06 1.05 0.96 0.08 1.13 0.00 1.94 0.07 0.09 1.02 0.11 0.03 3.02 0.07 0.06 0.83 0.15 0.93 0.07
 0.05 1.94 0.10 0.96 0.96 0.10 1.84 0.17 0.09 1.02 0.15 0.07 0.91 1.01 0.03 1.00 0.16 1.07 0.00 0.16 0.94 2.05 0.02 0.10 2.08 0.16 0.02 0.99 1.04 1.03 0.93 0.09 2.02 0.14 0.90 0.11 0.14 3.03 0.11 0.97 2.04
0.12 0.86 0.14 1.06 0.04 0.95 0.12 0.91 0.07 0.13 0.99 0.13 0.09 1.04 1.02 0.91 3.02 0.08 0.09 0.95 0.15 0.01 0.88 1.04 0.08 0.86 0.15 0.13 0.98 1.07 0.95 1.05 0.14 0.10 1.03 1.07 0.91 1.00 0.20 0.12 0.95
0.10 0.97 0.13 0.95 0.00 0.19 0.97 0.16 0.00 0.95 0.15 0.98 0.00 0.19 1.00 0.14 0.00 1.00 0.17 0.93 0.02 1.99 1.04 0.15 0.06 1.15 1.97 0.09 3.10 0.16 1.09 0.16 2.00 0.19 0.19 3.33 4.92 2.13 2.09 0.93 0.16
0.16 1.07 0.16 2.85 0.16 0.18 2.00 1.09 1.04 1.01 0.17 0.15 1.01 0.18 0.11 0.94 0.14 2.14 0.10 0.93 0.10 0.18 1.02 0.13 0.11 1.00 1.22 0.03 0.13 1.00 0.13 0.05 1.05 0.98 1.13 0.09 0.17 1.08 0.16 1.94 0.13
1.02 0.07 0.99 0.06 1.12 0.10 2.08 0.09 0.15 0.91 0.22 1.09 0.15 1.14 0.15 0.15 1.07 0.09 0.91 0.15 1.01 0.09 1.95 0.18 0.11 4.37 0.26 0.94 0.17 0.20 3.10 0.19 1.15 0.16 2.00 0.20 0.10 0.96 0.13 1.07 0.07
3.14 1.05 0.15 1.07 0.23 0.98 1.02 0.21 0.16 1.15 0.18 0.11 1.03 0.14 0.22 1.89 1.14 1.96 2.09 0.15 0.16 2.08 0.22 0.11 1.02 0.22 1.01 0.07 1.01 0.14 1.03 0.07 0.20 2.04 0.22 0.12 0.90 2.10 1.06 0.16 1.06
0.20 0.19 1.91 0.11 0.15 0.95 0.16 0.18 0.98 0.16 0.14 1.10 0.13 0.11 0.89 0.07 0.08 1.98 0.14 2.06 0.08 0.94 0.10 0.20 1.09 0.13 0.13 1.09 0.03 0.17 1.09 0.16 1.92 0.19 0.11 0.85 0.11 1.18 0.16 5.13 0.15
0.20 1.18 0.08 0.12 1.11 0.16 2.05 0.23 0.93 0.17 0.94 0.05 0.17 1.10 0.16 0.14 1.17 0.05 0.18 0.95 0.12 2.13 0.16 0.12 1.09 0.12 1.13 0.98 0.18 0.11 2.79 0.00 0.14 0.99 0.15 3.20 0.15 1.95 0.20 0.02 1.03
0.10 0.13 0.99 0.10 1.09 0.14 0.05 2.17 0.06 1.02 0.12 0.08 1.94 1.04 0.12 0.11 1.89 0.12 0.04 1.13 0.12 0.08 1.09 0.18 0.17 1.06 0.10 1.26 0.13 0.09 1.21 0.04 0.21 1.16 0.00 2.07 0.16 0.02 2.29 0.14 0.09
1.15 0.08 0.12 1.01 0.09 0.07 1.14 2.10 0.95 1.06 0.08 1.15 4.43 0.02 1.21 0.18 0.21 1.04 0.08 1.05 0.18 0.03 1.11 0.15 1.16 1.22 0.14 0.15 1.35 0.08 0.16 1.03 4.11 0.99 0.19 0.14 1.17 1.10 0.18 0.18 1.04
0.12 1.21 0.15 1.28 0.05 0.14 0.95 0.22 1.09 0.11 0.21 1.11 0.34 1.12 2.00 0.14 3.94 0.10 0.16 1.22 0.73 0.17 0.15 1.04 0.32 0.16 0.94 0.14 1.02 0.14 1.00 1.02 1.19 0.16 0.04 1.00 2.76 0.14 1.16 1.04 0.99
0.16 0.11 0.93 0.24 0.94 1.01 1.16 0.15 0.79 0.14 1.16 0.16 0.17 0.93 1.89 0.26 0.11 0.74 0.23 1.94 0.96 0.23 2.13 0.05 0.81 0.14 0.10 1.44 0.10 1.08 0.43 3.43 0.26 0.11 2.14 0.93 0.11 0.08 1.92 0.38 0.89
1.30 1.11 3.09 0.14 0.04 1.18 0.07 0.15 2.08 0.55 1.18 0.16 0.16 5.06 1.17 0.17 0.15 0.98 0.25 0.18 1.05 1.44 0.14 0.83 0.24 1.08 1.40 1.01 0.89 0.56 1.02 0.13 0.17 2.25 1.24 0.98 0.30 0.99 0.14 0.20 2.10
0.63 1.17 0.19 0.07 4.36 1.20 0.09 0.36 0.83 1.02 1.13 3.12 0.54 1.12 0.17 0.06 1.32 0.11 0.90 0.21 1.11 1.33 0.88 0.09 0.32 0.97 0.19 1.09 0.22 2.04 0.21 0.13 1.24 0.27 0.91 0.35 0.16 1.19 0.17 1.13 0.43
1.10 0.21 1.85 1.89 0.57 0.21 0.72 0.20 4.48 0.85 0.30 0.53 0.84 0.20 0.98 2.67 0.31 0.09 0.89 0.33 0.29 0.92 0.29 1.05 0.15 0.10 1.21 0.46 1.06 0.21 0.13 3.15 0.14 0.23 1.18 0.25 0.16 0.93 0.74 0.24 0.89
0.12 1.17 0.31 1.07 0.17 0.04 1.05 0.15 0.32 1.13 0.98 0.16 1.57 0.17 0.28 1.04 0.07 0.21 1.26 0.04 0.87 0.26 0.13 1.04 0.18 0.16 1.16 0.23 0.15 1.06 0.20 0.16 0.83 0.06 0.31 0.80 0.18 1.05 0.10 0.97 0.17
0.13 1.09 0.23 0.22 0.83 0.21 1.64 0.19 0.09 2.20 0.34 0.87 1.03 0.81 1.07 0.14 0.12 1.17 0.05 0.97 0.20 0.15 1.27 0.18 0.23 1.10 0.93 0.09 0.15 1.10 0.17 1.17 0.18 1.06 0.34 0.09 0.88 0.44 2.04 0.26 0.20
2.24 0.15 0.74 0.14 0.98 0.15 0.20 0.90 1.99 1.19 0.37 0.21 1.16 0.12 0.79 2.04 0.10 0.47 1.17 0.01 0.46 2.01 1.91 1.19 0.56 0.69 0.10 0.33 3.14 1.50 1.26 1.77 0.14 0.66 0.20 0.08 1.47 0.36 0.23 1.11 0.28
1.09 0.98 0.18 1.74 1.01 0.83 0.36 3.47 0.12 0.21 1.10 3.04 1.07 0.31 0.19 1.84 0.09 1.01 0.77 0.69 0.38 1.10 0.64
Flow Indexes:  4 6 7 9 10 11 12 14 16 17 18 20 22 23 2527 27 30 33 33 33 36 38 41 41 43 44 46 46 49 52 5355 57 60 61 61 64 64 67 68 69 70 72 72 74 77 77 7779 80 80 82 84 86 88 91 94 95
96 97 97 97 100 103 104106 109 110 111 112 115 116 117 118 121 123 125 128 131 133 136 139141 143 143 144 147 148 148 150 150 150 152 154 154 157 157 157 158158 158 158 158 159 159 160
160 161 164 166 166 166 169 169 170 171172 175 178 180 180 182 185 188 189 192 195 196 197 200 202 202 204206 208 210 210 213 215 217 220 222 224 226 226 229 229 229 229 231234 234 234
236 238 238 241 243 245 245 245 246 248 250 251 254 257260 260 261 262 262 263 263 266 266 269 271 273 275 278 278 281 282282 283 285 288 288 291 294 297 300 303 303 305 305 307 310
313 316318 318 321 323 325 325 325 325 325 328 331 333 333 335 337 340 343346 348 348 351 353 354 357 357 357 360 362 362 362 364 364 367 370372 375 375 377 380 380 381 384 384 387 390
393 395 398 401 403 403406 406 409 412 415 416 416 417 418 420 421 421 421 421 423 426 428431 433 434 437 440 441 441 441 441 442 445 446 449 451 453 456 458461 463 464 464 466 466 466
466 469 470 473 476 478 480 481 482 485486 486 486 488 489 490 493 495 496 497 499 501 504 505 505 508 510510 511 513 513 515 518 520 522 522 522 525 525 526 529 529 531 532533 534 534
534 537 540 540 541 542 545 545 545 545 546 549 552553 555 557 558 559 560 561 562 565 565 566 567 569 572 572 573 574577 577 577 578 581 582 583 584 584 585 586 589 591
593 594595 598 600 602 602 605 607 610 612 614 616 616 617 617 618 620 622622 622 622 623 625 626 628 629 629 629 632 635 637 640 642 645 645645 648 651 652 654 656 658 661 664 665 667
667 670 673 675 678 681684 687 690 692 694 697 700 702 702 705 705 707 708 709 710 713 715718 721 722 725 727 729 732 734 734 737 737 739 741 744 745 745 746749 751 752 752 755 758 758
759 759 760 761 762 765 765 765 766 766767 768 768 770 773 776 778 779 781 781 782 783 785 785 785 788 789789 789 790 793 793 795 796 797 799 800
Bases:
gactacgagtagactCCATTTGATTCGAATGTCTGTTGGCGTAGGATTTCGGAGAGCACGTTTGCGATACGCGTATCTGCTGCTCCGCGGAAAGAATTTAAAAACCGGTGAAATTACGCAGGATGTGCGTGAAGAGAATCTGAGAATTTT
CAAAGAATCTTTAGACATGGTAACCAATCTCAATAACTGGCATGCCTTCATGAATCTTTTTGCTTCTGCAGGCTATTTGAAAGGCAGCCTGGTGGCATCATCCAATGCGGTAGTTTTCAGCTATGTTTTATATCTGATCGGAAAATATGA
GTATAAAGTATCGTCTGTTGAACTTCAGAAATTATTCGTAAATGGTATTTTTATGTCTACGTATTACTGGTATTTTATACGGGTATCTACAGAATCAgaggttagaaaactagtttgctgatttgcgagatgccatcatgcagatgaattcgtatcatatctgaatttctgttat
cggcaaccgtatttaacggatgacttacttgtttattcgtcg
Quality Scores:  37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 3740 40 40 39 39 39 39 40 40 40 40 40 40 40 40 40 40 4040 40 40 40 40 40 40 40 40 40 40 40 40 40 40 39 39 3940 40 40 40 40 40 40 37 37
37 37 37 37 37 37 37 3737 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 35 30 30 30 35 23 21 18 18 18 2020 18 18 18 32 33 33 35 37 37 35 34 34 37 37 37 3737 37 37 37
37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 37 35 32 30 30 31 32 32 23 23 15 15 15 15 1923 23 23 29 24 25 32 32 28 30 30 37 37 37 34 3434 37 37 37 37 37 37 37 37 37 37 37 37 37 37
37 3737 37 35 35 35 37 37 37 37 37 37 37 35 35 35 35 35 3532 32 32 30 20 20 20 20 20 28 32 33 33 33 33 33 3333 33 33 33 33 33 20 20 20 35 25 25 25 25 32 32 35 3737 37 37 37 37 37 37 37 37 35
32 32 30 30 30 32 3228 27 27 28 28 26 29 30 24 13 13 13 13 13 18 22 2825 18 18 18 25 19 21 21 21 33 21 32 28 30 30 32 3028 28 32 35 28 28 26 26 30 21 30 30 35 35 35 35 3520 20 20 25 32
27 33 33 32 27 27 27 27 23 23 23 2121 27 26 21 21 13 13 13 13 13 17 17 22 21 16 18 1821 21 23 26 31 15 15 14 19 16 16 20 17 17 28 13 2217 19 19 19 21 22 17 17 15 15 15 22 20 15 15 15
1811 11 11 11 11 11 20 22 16 11 11 11 17 17 22 21 2221 21 26 24 24 18 18 18 15 15 20 11 11 11 11 1111 11 11 11 11 9 18 11 11 11 15 18 22 18 18 17 2118 21 21 19 21 19 22 24 21 15 17 17
17 17 22 22 2222 22 22 22 22 17 17 17 17 17 17 21 23 25 22 2222 22 22 23 25 20 21 21 21 17 17 17 20 25 23 19 1715 17 21 21 19 21 22 19 13 11 11 11 11 11 11 11 1111 11 15 13 15 15 15
19 19 19 13 11 12 12 12 17 1517 12 18 12 12 18 17 12 12 12
```

Data format

FastQ format

```
@AAII-ZZ123:123:ABCDEFGHT:4:1101:1885:2240 1:N:0:ATTTCT
ATCTGACCGCCGCATTTGATGCAGTAAATTATTTATATGAGCAAGGGCATA
+
@@@FFFBBDFHHHGHIICBFHIIIGGIIGGGHIGCHGHIDHGIIIIIIIGI
@AAII-ZZ123:123:ABCDEFGHT:4:1101:1969:2247 1:N:0:ATTCCT
TAAACGCCCGCAGTTGCGATCCCAGGTGCATGACAGAGGCAATAAACCCGA
+
@CCFFFFFHHHHHJJJJJIJJJJIJIFHHIIJIJIJIJIIIIIJIJJIEHH
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
GGAGTTTCATTACAATTTATATATTTAAAGAGGNNNANGNNNNNGACTGAA
+
CCCFFFFFHGHHFIJIJJBHHIDHJIFHEFEEG###1#1#####00?DGFH
@AAII-ZZ123:123:ABCDEFGHT:4:1101:2226:2183 1:N:0:ATTCCT
TTCAGTTTGTGATGTGCGACGATGGTTCGCTCANGCGNCTNNNGTTCTGCG
+
CCCFFEFFHHHHHGHGGIIIJIJJJGIJIIJIJ#07B#-7###--;CHIJH
@AAII-ZZ123:123:ABCDEFGHT:4:1101:2094:2194 1:N:0:ATTCCT
CTCCACACTAACAATACCGTTCCCCAGGTGGTATCGCCAGNNCAGTAGAGC
+
<?@D?DDDFFHHBDGDCBGIIDFCDGDC??D:C@F??GHF##07;;CB@@F
@AAII-ZZ123:123:ABCDEFGHT:4:1101:2544:2173 1:N:0:ATTCCT
GCCGCCCAGCTGAAAAACATCATCATGCTGATCNNNANTNNNNNAGGCAGA
```

Data format

FastQ format

```
@AAII-ZZ123:123:ABCDEFGHT:4:1101:1885:2240 1:N:0:ATTTCT
ATCTGACCGCCGCATTTGATGCAGTAAATTATTTATATGAGCAAGGGCATA
+
@@@FFFBBDFHHHGHIICBFHIIIGGIIGGGHIGCHGHIDHGIIIIIIIGI
@AAII-ZZ123:123:ABCDEFGHT:4:1101:1969:2247 1:N:0:ATTCCT
TAAACGCCCGCAGTTGCGATCCCAGGTGCATGACAGAGGCAATAAACCCGA
+
@CCFFFFFHHHHHJJJJJIJJJJIJIFHHIIJIJIJIJIIIIIJIJJIEHH
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
GGAGTTTCATTACAATTTATATATTTAAAGAGGNNNANGNNNNNGACTGAA
+
CCCFFFFFHGHHFIJIJJBHHIDHJIFHEFEEG###1#1#####00?DGFH
@AAII-ZZ123:123:ABCDEFGHT:4:1101:2226:2183 1:N:0:ATTCCT
TTCAGTTTGTGATGTGCGACGATGGTTCGCTCANGCGNCTNNNGTTCTGCG
+
CCCFFEFFHHHHHGHGGIIIJIJJJGIJIIJIJ#07B#-7###--;CHIJH
@AAII-ZZ123:123:ABCDEFGHT:4:1101:2094:2194 1:N:0:ATTCCT
CTCCACACTAACAATACCGTTCCCCAGGTGGTATCGCCAGNNCAGTAGAGC
+
<?@D?DDDFFHHBDGDCBGIIDFCDGDC??D:C@F??GHF##07;;CB@@F
@AAII-ZZ123:123:ABCDEFGHT:4:1101:2544:2173 1:N:0:ATTCCT
GCCGCCCAGCTGAAAAACATCATCATGCTGATCNNNANTNNNNNAGGCAGA
```

SequenceID

Sequence

Optional

Quality

Data format

FastQ format - ID explanation

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
GGAGTTTCATTACAATTTATATATTTAAAGAGGNNNANGNNNNNGACTGAA
+
CCCFFFFFHGHHFIJIJJBHHIDHJIFHEFEEG###1#1#####00?DGFH
```

Unique instrument name

Run id

Flowcell id

Flowcell lane

Tile number within the flowcell lane

'x'-coordinate of the cluster

'y'-coordinate of the cluster

The mate member of a pair

Y if the read fails filter (read is bad), N otherwise

Control bits

Index sequence

@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG

SequenceID

FastQ format - Quality scores

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
GGAGTTTCATTACAATTTATATATTTAAAGAGGNNNANGNNNNNGACTGAA
+
CCCFFFFFHGHHFIJIJJBHHIDHJIFHEFEEG###1#1####00?DGFH
```

CCCFFFFFHGHHFIJIJJBHHIDHJIFHEFEEG###1#1#####00?DGFH

Quality

$Q_{phred} = -10 \log_{10}(e)$

e = estimated probability of a base being wrong

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
                          XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
                               IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
                                 JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                              |                    |
33                        59   64       73                            104                  126
0........................26...31.......40
                         -5....0........9.............................40
                               0........9.............................40
                                  3.....9.............................40
0........................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

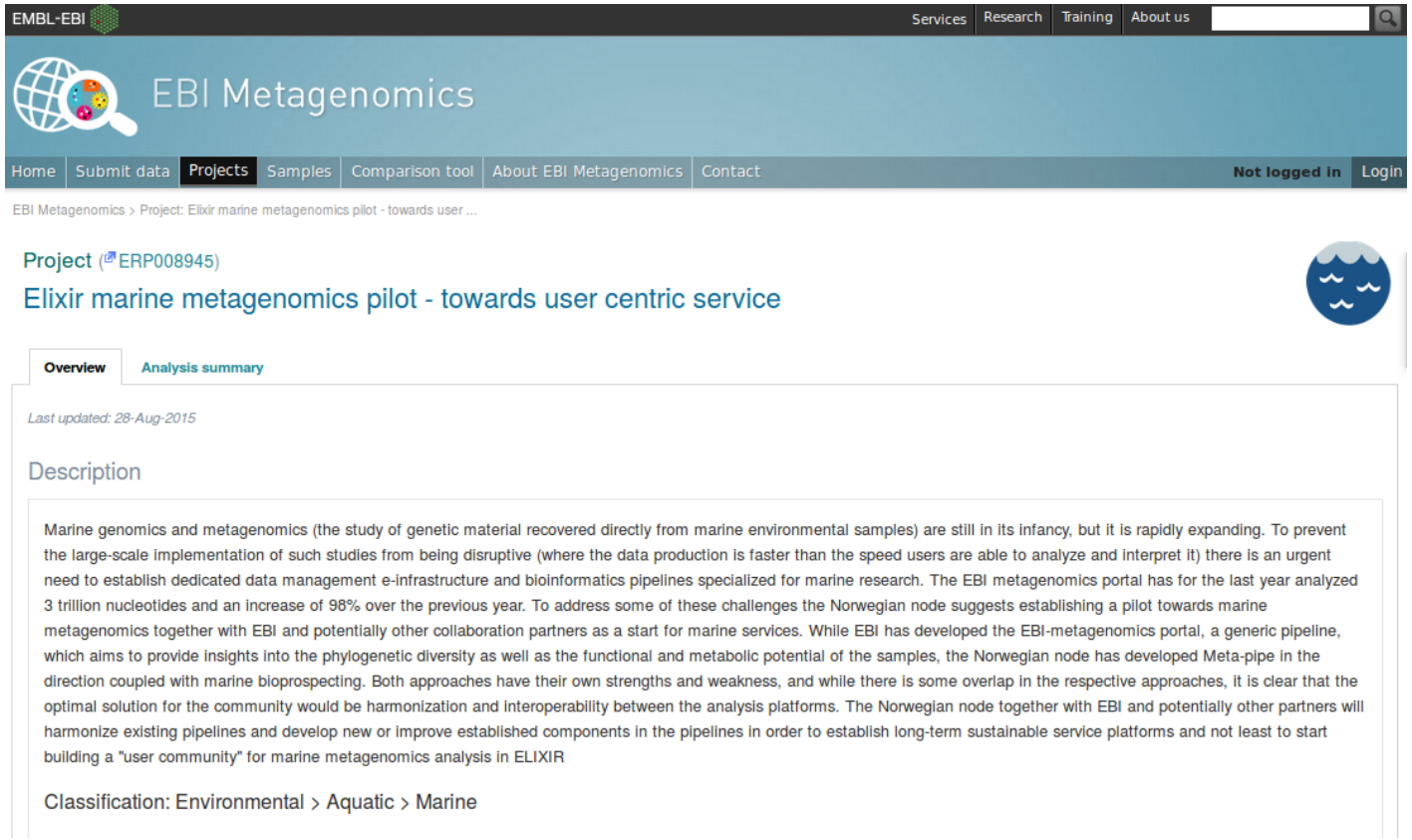
Have a look at the project

Where we can find Metagenomics data

https://www.ebi.ac.uk/metagenomics/

Have a look at the project

Just to start we will work on real metagenomics data downloaded from public data on EMBL-EBI Metagenomics



https://www.ebi.ac.uk/metagenomics/projects/ERP008945

Have a look at the project

Contact details

Institute: UIT
Name: Espen Robertsen
Email: Espen.m.robertsen@uit.no

Associated runs

Sample Name	Sample ID	Run ID	Experiment type	Version	Analysis results
Muddy	ERS624612	ERR695596	Metagenomic	1.0	Taxonomy | Function |
		ERR695596	Metagenomic	2.0	Taxonomy | Function |
Sandy	ERS624613	ERR695597	Metagenomic	1.0	Taxonomy | Function |
		ERR695597	Metagenomic	2.0	Taxonomy | Function |

Submitting Centre	Run Date	Platform	Model	Read Count	Base Count
University of Tromso, NORWAY		ILLUMINA	Illumina MiSeq		

Library Layout	Library Strategy	Library Source	Library Selection	Library Name
PAIRED	WGS	METAGENOMIC	size fractionation	unspecified

Navigation | Read Files

This table contains the files for run ERR695596
Download files

Download: 1 - 1 of 1 results in TEXT

Select columns

Showing results 1 - 1 of 1 results

Study accession	Secondary study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	Fastq files (ftp)	Fastq files (galaxy)	Submitted files (ftp)	Submitted files (galaxy)	CRAM index files (ftp)	CRAM index files (galaxy)
PRJEB7944	ERP008945	SAMEA3168559	ERS624612	ERX640077	ERR695596	412755	marine sediment metagenome	Illumina MiSeq	PAIRED			Fastq file 1 Fastq file 2	Fastq file 1 Fastq file 2		

Creating project folder structure

Operative folder

```
# go to practice folder.. linux is case sensitive
cd Metagenomic

# have a look at the folder
ls -ltr

# have a look at the tree
tree
```

Creating project folder structure

Creating project folder structure

```
# Create project folder
mkdir project

# Go to project folder
cd project

# have a look, it should be empty.
ls -ltr
```

Creating project folder structure

Creating project folder structure

```
# Create project folder
mkdir project

# Go to project folder
cd project

# have a look, it should be empty.
ls -ltr
```

Linking original data file - 454 sff file

```
# Make a symbolic link FROM - TO
ln -s ../original_data/*.fastq ./
```

which means:

make a symbolic links to all files from one folder up, original data, everything ending with ".fq.gz" in this folder.

Cleaning/trimming original data

Illumina files (as we explained previously) are usuali gzip-ped

```
# just in case we need to unzip them before cleaning
# gunzip file.fastq.gz
```

Prinseq-lite for cleaning datasets.

Always, always, always READ the Manual...

```
# to have an idea of the program.... read the manual
prinseq-lite.pl -h

# execute prinseq-lite.pl with a bunch of parameters
prinseq-lite.pl -fastq sample_R1.fastq -fastq2 sample_R2.fastq -out_format 3 -out_good cleaned \
  -min_len 50 -trim_qual_right 20 -trim_qual_type mean -trim_qual_window 20 -out_bad null

# have a look at the output
ls -ltr
```

FASTQ STATISTICS BY FASTQ-STATS FROM EA-UTILS

FASTQ-STATS FROM EA-UTILS

Have a look at the input/output

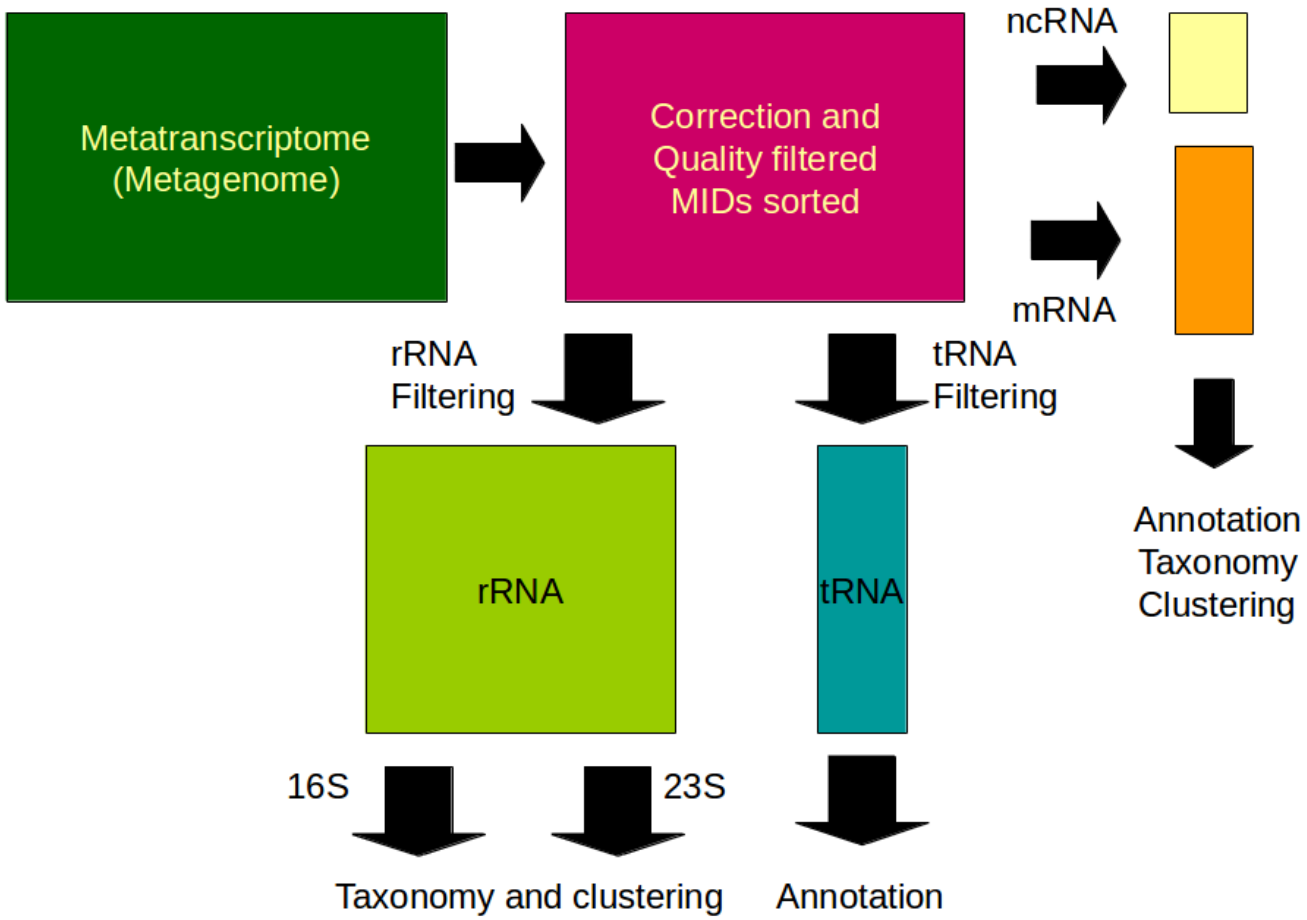
```
ls -ltr

# have a look at the help
fastq-stats -h

# check fastq file after
fastq-stats sample_R1.fastq

# check fastq file after
fastq-stats cleaned_1.fastq
```

Join paired reads

fastq-join from ea-utils

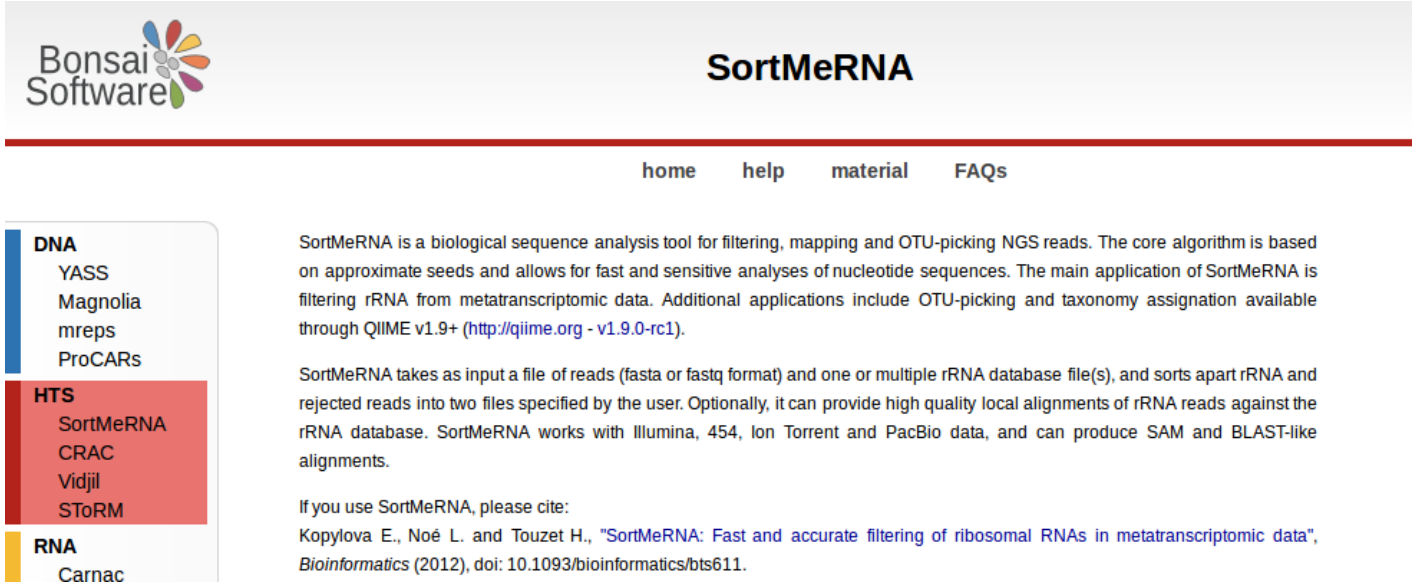
```
# so easy as reading the manual
fastq-join -h

# and calling it
fastq-join -v ' ' cleaned_1.fastq cleaned_2.fastq -o sample.%.fastq

# Have a look at the output
ls -ltr
```

WE HAVE OUR DATASET!!!!



Searching for ribosomal RNAs

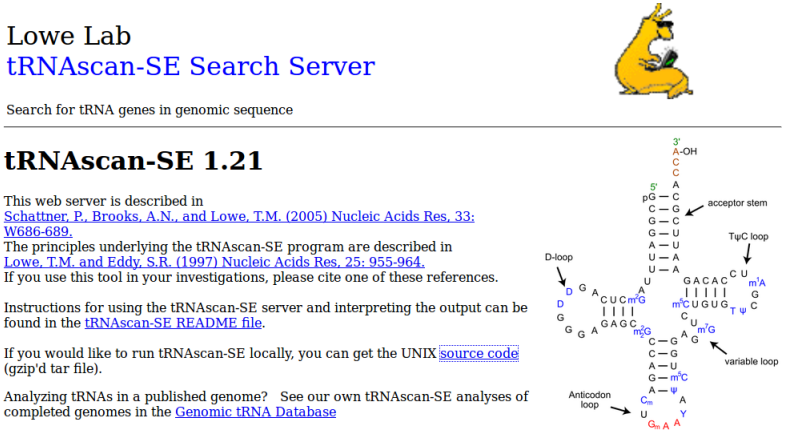
We can use *SortMeRNA* which searches among provided databases.




```
# we can format our database(s)
# indexdb_rna --ref ../rRNA_databases/silva-bac-16s-id90.fasta,./index/silva-bac-16s-db:\
# ./rRNA_databases/silva-bac-23s-id98.fasta,./index/silva-bac-23s-db
# I already did it

sortmerna --ref ../rRNA_databases/silva-bac-16s-id90.fasta,../index/silva-bac-16s-db:\
../rRNA_databases/silva-bac-23s-id98.fasta,../index/silva-bac-23s-db \
--reads sample.join.fastq --sam --num_alignments 1 --fastx --aligned sample_rRNA \
--other sample_not_rRNA --log -v
```

Searching for tRNAs

From previously file **sample_non_rRNA** we can sort out tRNAs sequences using tRNAScanSE




```
# first of all we need to convert fastq to fasta (FASTX-Toolkit)
fastq_to_fasta -Q 33 -i sample_not_rRNA.fastq > sample_not_rRNA.fasta

# executing tRNAscan-SE on general models (three kingdoms)
tRNAscan-SE -G -o tRNAs.txt sample_not_rRNA.fasta

# have a look at the results
less tRNAs.txt

# Extract the first column of the results (IDs column) skipping the fist 4 lines
tail -n +4 tRNAs.txt | awk '{print $1}' > tRNAsIDs.txt
```

Searching for tRNAs

We have now to filter out **tRNAs.txt** reads in the IDs table from the **sample_non_rRNA.fastq**

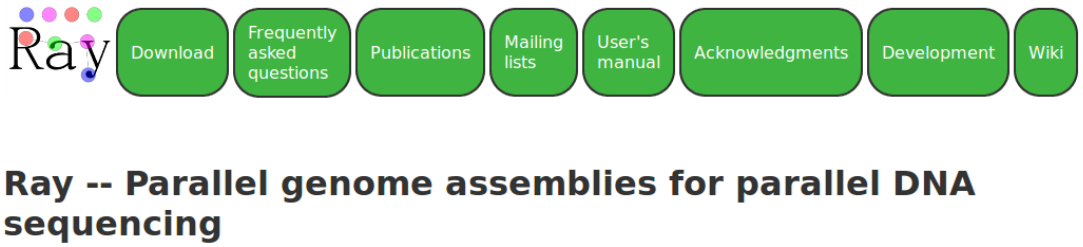
THIS SCRIPT DOES NOT WORK!! See the good one in /home/training/bin/RExtractFastqFromIdList.R

```
args<-commandArgs(TRUE)
if(args[1] == ""){
  print("usage: RExtractFastqFromIdList.R tabIdFile IN-FastqFile OUT-FastqFile");
  print("library ShortRead is required");
  print("Try again......")
  q()
}
suppressMessages(library("ShortRead"))

# Read table
ta<-read.table(args[1], sep="\t")

# Read fastq file
fqi<-readFastq(args[2])
```

```
RExtractFastqFromIdList.R tRNAsIDs.txt sample_not_rRNA.fastq sample_no_trnas.fastq
```

Assembling metagenome

Now we are ready to assemble our filtered metagenome. We will use **Ray**




```
# Execute SPAdes on sample_no_trnas.fastq
spades.py -s sample_no_trnas.fastq -o spades_out

tree -f

less .....
```

Searching for ORFs.

For ORFs search we will use **prodigal** software

We will search orfs within spades_out/contigs.fasta

```
prodigal -i spades_out/contigs.fasta -a orfs.faa -d orfs.fna -f gff -o orfs.gff -p meta
```