

SOLiD, Illumina and Roche Sequencing Technologies



NGS Sequence Files

```
T.22312.2.0..333.0.33313020300..3.233330323.0313303
>2_22_906_F3
T.32021.3.2..333.3.23133203303..0.303322323.2330203
>2_23_207_F3
T.0332321.2..302.3.30133002201.02110110001311101000
>2_23_362_F3
T.2110001.3..032.1.01132200321.221310312301.2230000
>2_23_540_F3
T.0212131.2..321.2.12112222013.11111322230301211201
>2_23_561_F3
T.0123010.1..213.2.01221333331.32212211112210112122
>2_24_516_F3
T.2121220.1..003.1.32303222313.33123221111202112211
>2_24_902_F3
T.0101123.3..033.0.32103222032.02120302332233320100
>2_25_835_F3
```

Current Method for Read Statistics

```
my $inFastaFile=$ARGV[0];
my @inFiles=<$ARGV[0]/*.fa>;
push( @inFiles, <$ARGV[0]/*.fasta>);
open(OUT,">FastaStats.txt");
print OUT "InFileName\tNumberOfFastaSeqs\tLengthofSeqs\n";
my $grandTotal=0;
my $grandSeqs=0;
foreach my $file ( @inFiles ){
    $file=~s/\\/\\\\/g;
    print "\nParsing the input fasta file $file...";
    my $seqio = Bio::SeqIO->new(-file => $file, '-format' => 'Fasta');
    my ($noofFastaSeqs,$totalLength)=(0,0);

    while(my $seq = $seqio->next_seq) {
        my $seqid = $seq->id;
        my $desc = $seq->desc;
        #my $string = $seq->seq;
        my $len = $seq->length;
        $noofFastaSeqs++;
        #$totalLength+=length($string);
        $totalLength+=$len;
        #print "Found -> $seqid\t$desc\n";
        #print "\n$seqid\t$desc\t", length($string), "\n";
    } # end of while loop
    $grandTotal+=$totalLength;
    $grandSeqs+=$noofFastaSeqs;
    print OUT basename($file),"\t $noofFastaSeqs\t$totalLength \n";
}
```

Magic of Unix Command

find /labs/bioscience_customer/Arnab_Pain/Solid/solid0397_20110808_PE_BC_LF46 -name '*csfasta'| xargs grep -c "^>"

```
/labs/bioscience_customer/Arnab_Pain/Solid/solid0397_20110808_PE_BC_LF46/c_velia1/  
solid0397_20110808_PE_BC_LF46_bcSample1_F3_c_velia1.csfasta:24648458  
/labs/bioscience_customer/Arnab_Pain/Solid/solid0397_20110808_PE_BC_LF46/c_velia2/  
solid0397_20110808_PE_BC_LF46_bcSample1_F3_c_velia2.csfasta:5201106  
/labs/bioscience_customer/Arnab_Pain/Solid/solid0397_20110808_PE_BC_LF46/m_bovis/  
solid0397_20110808_PE_BC_LF46_bcSample1_F3_m_bovis.csfasta:6700023  
/labs/bioscience_customer/Arnab_Pain/Solid/solid0397_20110808_PE_BC_LF46/m_marinum/  
solid0397_20110808_PE_BC_LF46_bcSample1_F3_m_marinum.csfasta:3629682  
/labs/bioscience_customer/Arnab_Pain/Solid/solid0397_20110808_PE_BC_LF46/m_tuberculosis1/  
solid0397_20110808_PE_BC_LF46_bcSample1_F3_m_tuberculosis1.csfasta:16370665  
/labs/bioscience_customer/Arnab_Pain/Solid/solid0397_20110808_PE_BC_LF46/m_tuberculosis2/  
solid0397_20110808_PE_BC_LF46_bcSample1_F3_m_tuberculosis2.csfasta:6368279  
/labs/bioscience_customer/Arnab_Pain/Solid/solid0397_20110808_PE_BC_LF46/p_falciparum1/  
solid0397_20110808_PE_BC_LF46_bcSample1_F3_p_falciparum1.csfasta:9044872  
/labs/bioscience_customer/Arnab_Pain/Solid/solid0397_20110808_PE_BC_LF46/p_falciparum2/  
solid0397_20110808_PE_BC_LF46_bcSample1_F3_p_falciparum2.csfasta:16931047  
/labs/bioscience_customer/Arnab_Pain/Solid/solid0397_20110808_PE_BC_LF46/p_falciparum3/  
solid0397_20110808_PE_BC_LF46_bcSample1_F3_p_falciparum3.csfasta:9244234  
/labs/bioscience_customer/Arnab_Pain/Solid/solid0397_20110808_PE_BC_LF46/p_falciparum4/  
solid0397_20110808_PE_BC_LF46_bcSample1_F3_p_falciparum4.csfasta:3407158  
/labs/bioscience_customer/Arnab_Pain/Solid/solid0397_20110808_PE_BC_LF46/p_yoelii/  
solid0397_20110808_PE_BC_LF46_bcSample1_F3_p_yoelii.csfasta:4778936  
/labs/bioscience_customer/Arnab_Pain/Solid/solid0397_20110808_PE_BC_LF46/p_yoelii2/  
solid0397_20110808_PE_BC_LF46_bcSample1_F3_p_yoelii2.csfasta:17481517
```

-bash-3.2\$ find /labs/bioscience_customer/Christian_Voolstra/Solid/ -name '*csfasta'| xargs grep -c "^>"

```
/labs/bioscience_customer/Christian_Voolstra/Solid/Pool1/Chris_Voolstra_RNASeq_20111010_FRAG_BC_Chris_Voolstra_Pool1_F3_Pool1.csfasta:1938  
/labs/bioscience_customer/Christian_Voolstra/Solid/Pool2/Chris_Voolstra_RNASeq_20111010_FRAG_BC_Chris_Voolstra_Poo2_F3_Pool2.csfasta:2221  
/labs/bioscience_customer/Christian_Voolstra/Solid/Pool3/Chris_Voolstra_RNASeq_20111010_FRAG_BC_Chris_Voolstra_Poo3_F3_Pool3.csfasta:760  
/labs/bioscience_customer/Christian_Voolstra/Solid/Pool4/Chris_Voolstra_RNASeq_20111010_FRAG_BC_Chris_Voolstra_Poo4_F3_Pool4.csfasta:2232
```

Take Home Message

Multiple lines of a particular programming language script can be replaced with short and crisp unix command like this,

```
-bash-3.2$ find /labs/bioscience_customer/  
Christian_Voolstra/Solid/ -name '*csfasta'|  
xargs grep -c "^>"
```