Research use only

Raw Data

Report

December 2018

Humanizing Genomics
macrogen

Project Information

Client Name	Carolina Figueroa
Company / Institution	Pontificia Universidad Catolica de Valparaiso
Order Number	1810KNO-0058
Type of Read	Paired-end
Read Length	101
Number of Samples	24
Library Kit	TruSeq Stranded mRNA LT Sample Prep Kit
Library Protocol	TruSeq Stranded mRNA Sample Preparation Guide, Part # 15031047 Rev. E
Type of Sequencer	Illumina platform

Table of Contents

1. Data Download Information

1. 1. Raw Data

Download link	File size	md5sum
1810KNO-0058.zip	60.2G	6172c0cbeb7862de5ee871902e712aff

- *.zip : Read1/Read2 of All Samples (- One compressed file of all samples)

Download link	File size	md5sum
CONTROL1_1.fastq.gz	1.1G	4b8ec2e0e1a718b628ddde7fbe40ce9f
CONTROL1_2.fastq.gz	1.1G	6fe4bc3000f61a26c4872cc252768c87
CONTROL2_1.fastq.gz	1.1G	5a1de9e0c93ea6671163536f01bb2607
CONTROL2_2.fastq.gz	1.1G	5a6544e97bf462eb0f1aaee8d84199fa
CONTROL3_1.fastq.gz	1.2G	b6e24de7bf5a10de001de03ad7ba5b9a
CONTROL3_2.fastq.gz	1.3G	d3e3bffdc79e3b2f328ee00685d3b848
SUSCALPSVAC1_1.fastq.gz	1.5G	d167db2b0a49ec6277ea5b864d164f5d
SUSCALPSVAC1_2.fastq.gz	1.6G	6830fffc68bba214d831072ee04af3a6
SUSCALPSVAC2_1.fastq.gz	1.6G	0ed4ce11cc3aac2d986c582fd9d6cf5d
SUSCALPSVAC2_2.fastq.gz	1.6G	b1c4526b4167b1c8bcdaa6000347706d
SUSCALPSVAC3_1.fastq.gz	1.1G	891ea4e805bac6637f1a9c1d64a840ac
SUSCALPSVAC3_2.fastq.gz	1.1G	9a86020d113ca4bd8a0a6a230b8860cb
SUSPSNVAC1_1.fastq.gz	1.2G	413273cf9bb45bbad67de391133e3c92
SUSPSNVAC1_2.fastq.gz	1.3G	afcd6401be6b14041e73ef5a672b5db9
SUSPSNVAC2_1.fastq.gz	1.2G	9cfe1c38f6fb2b3df7f5fa40dee990b8
SUSPSNVAC2_2.fastq.gz	1.3G	c9422cbbab3ec4a45e7ba414fc6dca87
SUSPSNVAC3_1.fastq.gz	1.2G	cad37830bc77e62429239cab68fc7cf5
SUSPSNVAC3_2.fastq.gz	1.2G	1b198a09e05f98ccfb14849d7c70778a
SUSPSVAC1_1.fastq.gz	1.3G	d52d3d25b3daefec28c9d67c1ca92d92
SUSPSVAC1_2.fastq.gz	1.3G	6e123daa3bb0bc6a1096ad2affa5ef59
SUSPSVAC2_1.fastq.gz	1.2G	2ddb4a9fe2400a774de02d3c7b0759f4
SUSPSVAC2_2.fastq.gz	1.3G	eb7f11ddc8967d51fc11f6e698a1a8dd
SUSPSVAC3_1.fastq.gz	1.2G	d1f732dec67664e9a24b07b1445cbe6c
SUSPSVAC3_2.fastq.gz	1.2G	f463fedde256c83a45509f34397bd511
RESCALPSVAC1_1.fastq.gz	1.0G	a00b0355f347a69313851249495c37eb
RESCALPSVAC1_2.fastq.gz	1.1G	7041883867223a0ee08f327245bbdda1
RESCALPSVAC2_1.fastq.gz	1.1G	b35c53c9ba90b92470c51c0a39d9825a
RESCALPSVAC2_2.fastq.gz	1.2G	9f22ac414987b7fa4b25566c33d104c4
RESCALPSVAC3_1.fastq.gz	1.3G	ea92afa97e3950ef7256a6a77f74a3b9

RESCALPSVAC3_2.fastq.gz	1.3G	5aeb99d702c0b5c07c72b400749819b5
RESPSNVAC1_1.fastq.gz	1.3G	be8a05dbc0e0190837091665df6ade2f
RESPSNVAC1_2.fastq.gz	1.3G	bb402f438499d1df62d63bcdba6d3395
RESPSNVAC2_1.fastq.gz	1.1G	6670cc840b9e005ba97764f47da25713
RESPSNVAC2_2.fastq.gz	1.2G	7a1c9b5b39209123d08e611dd7350c17
RESPSNVAC3_1.fastq.gz	1.2G	71e1a20e0c16aa8e736ea5daf5c31d84
RESPSNVAC3_2.fastq.gz	1.2G	cb6280f330218832cab6e79a4710968b
RESPSVAC1_1.fastq.gz	1.3G	09aeb28d275f9e62b2133baaf471a00e
RESPSVAC1_2.fastq.gz	1.4G	d0b65851bcfffc0719288782d18d2814
RESPSVAC2_1.fastq.gz	1.3G	2efb2b29750323e34987910833f59adb
RESPSVAC2_2.fastq.gz	1.3G	32f6c198bcca4777ae6813b14b6fd5b0
RESPSVAC3_1.fastq.gz	1.5G	6f487288d5f6296043c8a4b74a3ede62
RESPSVAC3_2.fastq.gz	1.5G	021b12a00059ed27855cbeb783b7dfce
SUSCALPSNVAC1_1.fastq.gz	1.4G	45247556e2026e6d01bd8e61fdda6618
SUSCALPSNVAC1_2.fastq.gz	1.5G	6a80d067967bcc9e8c4a8f1b648cce7f
SUSCALPSNVAC2_1.fastq.gz	1.0G	30b3206c12d7864773a939d2f3841597
SUSCALPSNVAC2_2.fastq.gz	1.1G	158abfce9ef54b5be70172b246aefb47
SUSCALPSNVAC3_1.fastq.gz	1.1G	25bfb27e2f072897822e352335c39640
SUSCALPSNVAC3_2.fastq.gz	1.1G	17a6e227559ace24487b4a4652a67941

- fastq.gz : This is a zip file of raw data used in analysis.
- md5sum : In order to verify the integrity of files, md5sum is used. If the values of md5sum are the same, there is no forgery, modification or omission.

Your data will be retained in our server for 3 months. Should you wish to extend the retention period, please email (ngs@macrogen.com) or contact our sales team.

2. Experimental Methods and Workflow

2. 1. Experiment Overview

Fig1. Experiment overview

The Illumina NGS workflow includes 4 basic steps :

1) Sample Preparation

For library construction, DNA/RNA is extracted from a sample. After performing quality control (QC), qualified samples proceed to library construction.

2) Library Construction

The sequencing library is prepared by random fragmentation of the DNA or cDNA sample, followed by 5' and 3' adapter ligation. Alternatively, "tagmentation" combines the fragmentation and ligation reactions into a single step that greatly increases the efficiency of the library preparation process. Adapter-ligated fragments are then PCR amplified and gel purified.

3) Sequencing

For cluster generation, the library is loaded into a flow cell where fragments are captured on a lawn of surface-bound oligos complementary to the library adapters. Each fragment is then amplified into distinct, clonal clusters through bridge amplification. When cluster generation is complete, the templates are ready for sequencing.

Illumina SBS technology utilizes a proprietary reversible terminator-based method that detects single bases as they are incorporated into DNA template strands. As all 4 reversible, terminator-bound dNTPs are persent during each sequencing cycle, natural competition minimizes incorporation bias and greatly reduces raw error rates compared to other technologies. The result is highly accurate base-by-base sequencing that virtually eliminates sequence-context-specific errors, even within repetitive sequence regions and homopolymers.

4) Raw data

Sequencing data is converted into raw data for the analysis.

2. 2. Generation of Raw Data

The Illumina sequencer generates raw images utilizing sequencing control software for system control and base calling through an integrated primary analysis software called RTA (Real Time Analysis). The BCL (base calls) binary is converted into FASTQ utilizing illumina package bcl2fastq. Adapters are not trimmed away from the reads.

3. Summary of Produced Data

3. 1. Raw Data Statistics

The total number of bases, reads, GC (%), Q20 (%), and Q30 (%) are calculated for the 20 samples. For example, in CONTROL1, 44,919,326 reads are produced, and total read bases are 4.5G bp. The GC content (%) is 48.941% and Q30 is 95.650%.

✔ **The following table only shows maximum of 20 samples. If your samples are more than 20, please refer to the attached excel file. View full table** : 1810KNO-0058_RawData_Stat.xlsx

Table 1. Raw data Stats (maximum 20 samples)

Sample ID	Total read bases (bp)	Total reads	GC(%)	AT(%)	Q20(%)	Q30(%)
CONTROL1	4,536,851,926	44,919,326	48.941	51.06	98.564	95.650
CONTROL2	4,554,153,832	45,090,632	48.920	51.08	98.657	95.884
CONTROL3	5,084,885,804	50,345,404	48.877	51.12	98.477	95.454
SUSCALPSVAC1	6,299,258,294	62,368,894	48.577	51.42	98.439	95.336
SUSCALPSVAC2	6,365,381,378	63,023,578	47.961	52.04	98.462	95.335
SUSCALPSVAC3	4,448,686,198	44,046,398	48.294	51.71	98.601	95.713
SUSPSNVAC1	5,059,734,582	50,096,382	48.870	51.13	98.559	95.656
SUSPSNVAC2	5,077,878,424	50,276,024	48.678	51.32	98.579	95.661
SUSPSNVAC3	4,837,345,106	47,894,506	48.778	51.22	98.342	95.147
SUSPSVAC1	5,221,345,490	51,696,490	48.964	51.04	98.577	95.662
SUSPSVAC2	5,098,944,600	50,484,600	49.149	50.85	98.339	95.083
SUSPSVAC3	4,742,422,680	46,954,680	49.192	50.81	98.652	95.860
RESCALPSVAC1	4,212,306,202	41,706,002	48.407	51.59	98.442	95.367
RESCALPSVAC2	4,658,276,146	46,121,546	48.630	51.37	98.434	95.331
RESCALPSVAC3	5,261,670,750	52,095,750	48.796	51.2	98.448	95.329
RESPSNVAC1	5,222,984,720	51,712,720	48.801	51.2	98.473	95.382
RESPSNVAC2	4,611,961,384	45,662,984	48.893	51.11	98.474	95.418
RESPSNVAC3	4,873,954,576	48,256,976	49.130	50.87	98.359	95.197
RESPSVAC1	5,488,260,614	54,339,214	49.109	50.89	98.600	95.704
RESPSVAC2	5,332,148,752	52,793,552	48.951	51.05	98.567	95.685

- Sample ID : Sample name.
- Total read bases : Total number of bases sequenced.
- Total reads : Total number of reads. For Illumina paired-end sequencing, this value refers to the sum of read 1 and read 2.
- GC(%) : GC content.
- AT(%) : AT content.
- Q20(%) : Ratio of bases that have phred quality score of over 20.
- Q30(%) : Ratio of bases that have phred quality score of over 30.

3. 2. Total Read Bases

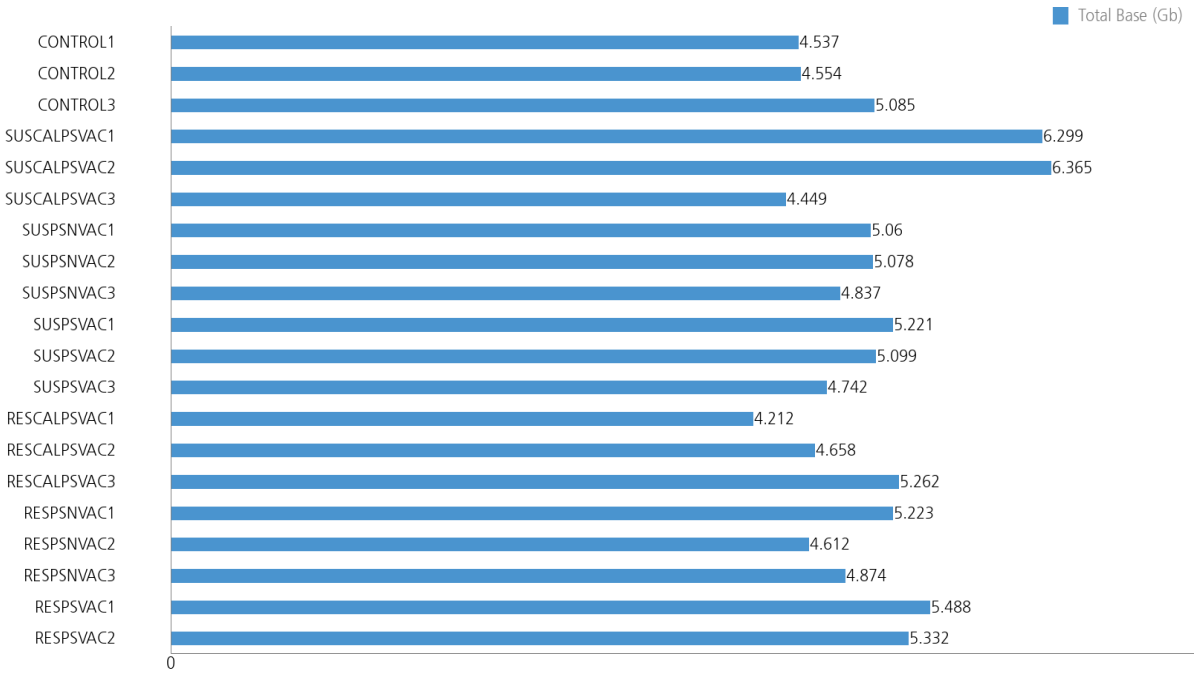

Figure 2.Throughput of Raw data

3. 3. Total Reads

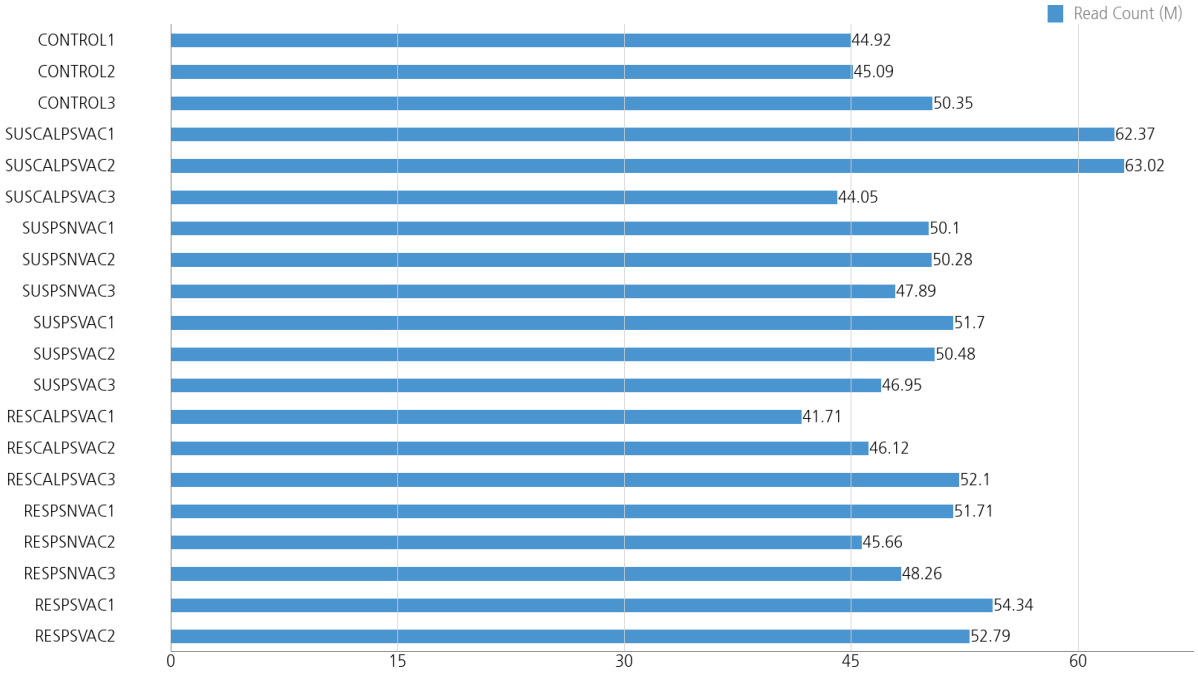

Figure 3. Total read count of Raw data

3. 4. GC/AT Content

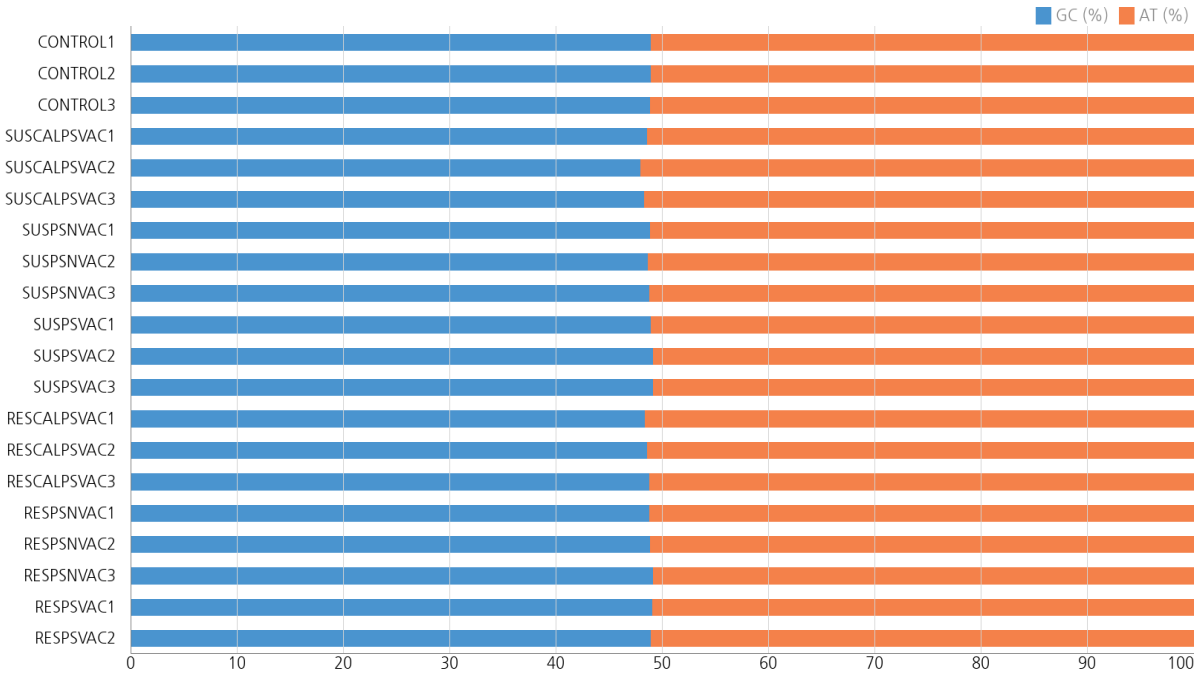

Figure 4. GC/AT Content of Raw data

3. 5. Q20/Q30 (%)

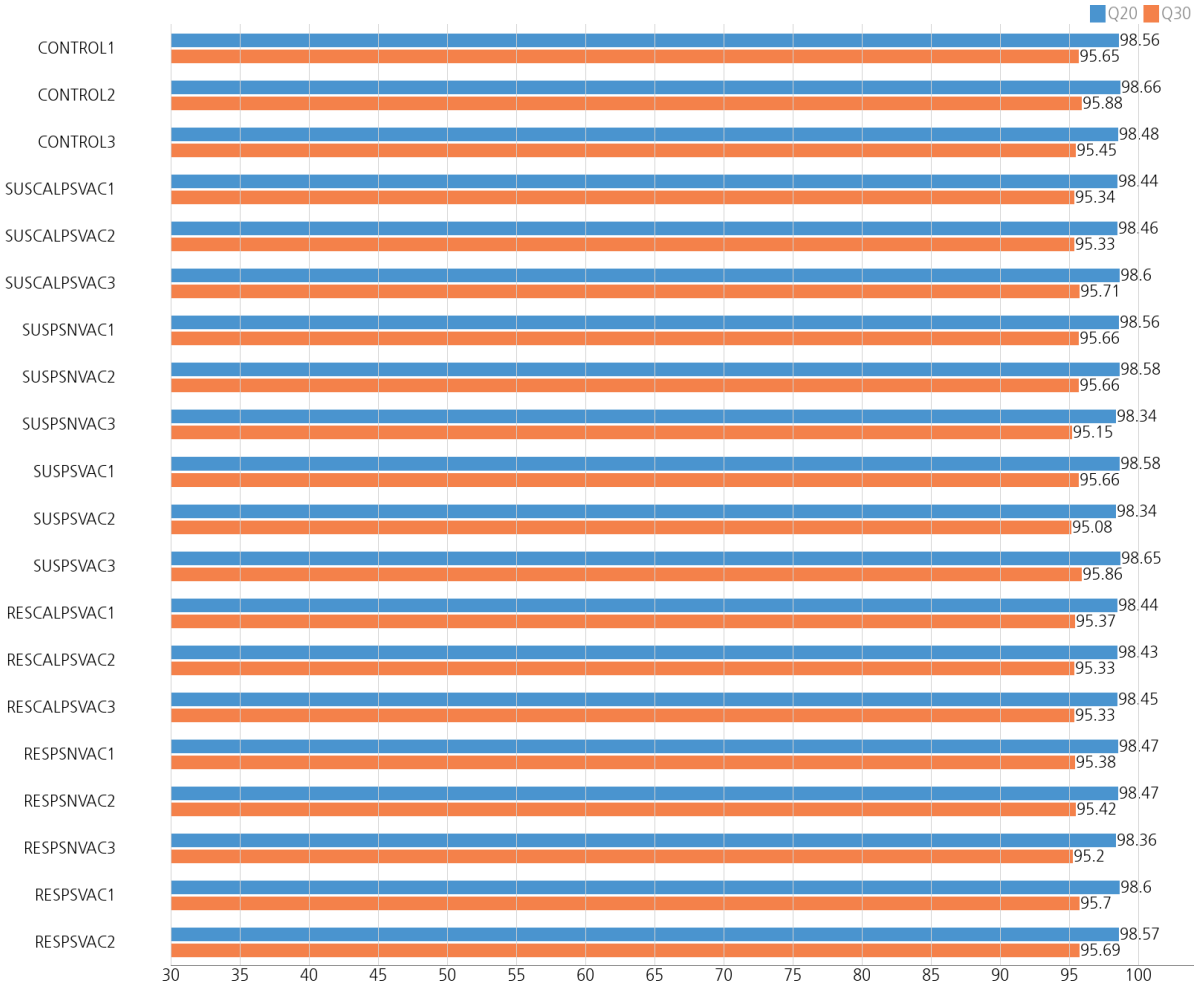

Figure 5. Q20/Q30 scores of Raw data

4. Appendix

4. 1. FAQ

Q: I want to see the produced data. How can I open the files?

A: As the large size zip files provided by our company are hard to process in the Windows environment, we highly recommend using Linux environment for a smoother operation.

4. 2. FASTQ File

Example of FASTQ

```
@HISEQ-MFG:501:HB0TFADXX:1:1101:1247:2183 1:N:0:
CTCAGCTAAATACTTTGACACCNGTANNANNNNNNNNNNTNNNNNNNNNNNN
+
@@@BDDDDHHHHFHIIIIIIII#3AC##############################
```

FASTQ file is composed of four lines.
Line 1 : ID line includes information such as flow cell lane information.
Line 2 : Sequences line.
Line 3 : Separator line (+ mark).
Line 4 : Quality values line about sequences.

4. 3. Phred Quality Score Chart

Phred quality score numerically expresses the accuracy of each nucleotide. Higher Q number signifies higher accuracy. For example, if Phred assigns a quality score of 30 to a base, the chances of having base call error are 1 in 1000.

Phred Quality Score Q is calculated with $-10\log_{10}P$, where P is probability of erroneous base call.

Quality of phred score	Probability of incorrect base call	Base call accuracy	Characters
10	1 in 10	90%	!"#$%&'()*+
20	1 in 100	99%	,-./012345
30	1 in 1000	99.9%	6789:;h=i?
40	1 in 10000	99.99%	@ABCDEFGHIJ

- Encoding : Sanger Quality (ASCII Character Code=Phred Quality Value + 33)

Humanizing Genomics
macrogen

Macrogen Inc. | 08511 10F, 254 Beotkkot-ro, Geumcheon-gu, Seoul, Rep of Korea

Contact us
Tel:+82-2-2180-7016
Site: www.macrogen.com | http://dna.macrogen.com