WEBVTT

NOTE duration:"00:58:02.2830000"

NOTE recognizability:0.934

NOTE language:en-us

NOTE Confidence: 0.98753315

00:00:00.000 --> 00:00:00.940 Yeah.

NOTE Confidence: 0.9373763

00:00:03.180 --> 00:00:06.078 OK, so today's second part of our

NOTE Confidence: 0.9373763

00:00:06.078 --> 00:00:08.753 travel through the analysis of a

NOTE Confidence: 0.9373763

00:00:08.753 --> 00:00:11.459 single cell RNA seek data processing.

NOTE Confidence: 0.9373763

00:00:11.460 --> 00:00:14.348 Last time we started with the defining how

NOTE Confidence: 0.9373763

00:00:14.348 --> 00:00:17.697 single cell RNA sequencing works and the

NOTE Confidence: 0.9373763

00:00:17.697 --> 00:00:19.745 differences between different protocols.

NOTE Confidence: 0.9373763

00:00:19.750 --> 00:00:21.930 For example, coverage on jeans,

NOTE Confidence: 0.9373763

00:00:21.930 --> 00:00:24.975 how they isolate cells and so on.

NOTE Confidence: 0.9373763

00:00:24.980 --> 00:00:28.228 Today we we deal more with the with

NOTE Confidence: 0.9373763

00:00:28.228 --> 00:00:31.079 the real analysis of the data.

NOTE Confidence: 0.9373763

00:00:31.080 --> 00:00:33.380 So last time we arrived.

NOTE Confidence: 0.9373763

00:00:33.380 --> 00:00:35.726 At the point where we saw,

NOTE Confidence: 0.9373763

00:00:35.730 --> 00:00:37.338 these are starting steps,

NOTE Confidence: 0.9373763

00:00:37.338 --> 00:00:40.829 we saw that from the molecular point of view,

NOTE Confidence: 0.9373763

00:00:40.830 --> 00:00:43.672 this strategy is to link the original

NOTE Confidence: 0.9373763

00:00:43.672 --> 00:00:46.235 RNA molecule with an oligo nucleotide

NOTE Confidence: 0.9373763

00:00:46.235 --> 00:00:49.594 called the cell barcode that allow us to

NOTE Confidence: 0.9373763

00:00:49.594 --> 00:00:52.250 identify the cell of origin of the RNA

NOTE Confidence: 0.9373763

00:00:52.250 --> 00:00:56.300 and then another important part is the UMIA.

NOTE Confidence: 0.9373763

00:00:56.300 --> 00:00:58.485 Molecular identifier that is a

NOTE Confidence: 0.9373763

00:00:58.485 --> 00:01:01.249 random nucleotide that allow us to

NOTE Confidence: 0.9373763

00:01:01.249 --> 00:01:03.277 correct for amplification biases,

NOTE Confidence: 0.9373763

00:01:03.280 --> 00:01:05.968 so to keep only those duplicate

NOTE Confidence: 0.9373763

00:01:05.968 --> 00:01:08.320 reads that are belonging to

NOTE Confidence: 0.9373763

00:01:08.320 --> 00:01:10.720 different molecules in our cells,

NOTE Confidence: 0.9373763

00:01:10.720 --> 00:01:15.067 and we're not amplified during the PCR.

NOTE Confidence: 0.9373763

00:01:15.070 --> 00:01:18.174 So after this steps away and after the

NOTE Confidence: 0.9373763

00:01:18.174 --> 00:01:21.409 mapping we also cover these last time and

NOTE Confidence: 0.94731027

00:01:21.410 --> 00:01:24.080 we're sorry. I'm question yes.

NOTE Confidence: 0.94731027

00:01:24.080 --> 00:01:27.195 How can you have the same UMI

NOTE Confidence: 0.94731027

00:01:27.195 --> 00:01:30.039 and two different RNAs? Oh, I see

NOTE Confidence: 0.9228061

00:01:30.040 --> 00:01:31.930 2015. If you have the same,

NOTE Confidence: 0.9228061

00:01:31.930 --> 00:01:33.808 Umm I you collapse the reader.

NOTE Confidence: 0.9228061

00:01:33.810 --> 00:01:36.008 So if the read is the same,

NOTE Confidence: 0.9228061

00:01:36.010 --> 00:01:38.145 the UMI is the same and the

NOTE Confidence: 0.9228061

00:01:38.145 --> 00:01:39.780 cell barcode is the same.

NOTE Confidence: 0.9228061

00:01:39.780 --> 00:01:41.040 You collapse the read

NOTE Confidence: 0.9228061

00:01:41.040 --> 00:01:43.230 and you know I see I see

NOTE Confidence: 0.9228061

00:01:43.230 --> 00:01:46.020 the one I'm looking at I see so you could

NOTE Confidence: 0.9228061

00:01:46.097 --> 00:01:48.614 have the same cell barcode, the same.

NOTE Confidence: 0.9228061

00:01:48.614 --> 00:01:50.973 Umm I but a different sequence because

NOTE Confidence: 0.9228061

00:01:50.973 --> 00:01:53.587 you're in a different part of the same RNA.

NOTE Confidence: 0.9840043

00:01:54.640 --> 00:01:56.607 Well, in theory that depends on the

NOTE Confidence: 0.9840043

00:01:56.607 --> 00:01:58.340 protocol, because some of those are

NOTE Confidence: 0.9840043

00:01:58.340 --> 00:02:00.259 only three prime end and so this

NOTE Confidence: 0.9840043

00:02:00.260 --> 00:02:02.789 is I'm looking at numbers five and six there.

NOTE Confidence: 0.900073

00:02:06.450 --> 00:02:09.260 Five and six other reads you mean, yeah?

NOTE Confidence: 0.900073

00:02:09.260 --> 00:02:11.010 Well, uhm, yes in theory.

NOTE Confidence: 0.900073

00:02:11.010 --> 00:02:12.386 But in theory, yes.

NOTE Confidence: 0.900073

00:02:12.386 --> 00:02:14.450 So these would be a different

NOTE Confidence: 0.900073

00:02:14.522 --> 00:02:16.628 RNA could be a different gene,

NOTE Confidence: 0.900073

00:02:16.630 --> 00:02:18.736 but randomly they have the same.

NOTE Confidence: 0.900073

00:02:18.740 --> 00:02:23.540 Umm, I yeah. So in theory it can happen.

NOTE Confidence: 0.900073

00:02:23.540 --> 00:02:26.500 It depends on the length of the UMI,

NOTE Confidence: 0.900073

00:02:26.500 --> 00:02:28.350 because they are randomly generated.

NOTE Confidence: 0.900073

00:02:28.350 --> 00:02:29.076 For example,

NOTE Confidence: 0.900073

00:02:29.076 --> 00:02:32.420 if you have if they are 12 nucleotide long,

NOTE Confidence: 0.900073

00:02:32.420 --> 00:02:34.982 the probability to have two that are

NOTE Confidence: 0.900073

00:02:34.982 --> 00:02:37.598 identical is for elevated at the 12th,

NOTE Confidence: 0.900073

00:02:37.600 --> 00:02:40.216 so the longer they are the the lower

NOTE Confidence: 0.900073

00:02:40.216 --> 00:02:42.780 is the probability to have two.

NOTE Confidence: 0.900073

00:02:42.780 --> 00:02:45.600 Umm with the same sequence.

NOTE Confidence: 0.900073

00:02:45.600 --> 00:02:46.550 OK, yeah.

NOTE Confidence: 0.9335265

00:02:49.010 --> 00:02:51.116 Uhm, OK, so you're my abstract.

NOTE Confidence: 0.9335265

00:02:51.120 --> 00:02:53.238 Their strategy used to reduce amplification

NOTE Confidence: 0.9335265

00:02:53.238 --> 00:02:55.680 biases in order to correct for that.

NOTE Confidence: 0.9335265

00:02:55.680 --> 00:02:57.906 And in single sat there important

NOTE Confidence: 0.9335265

00:02:57.906 --> 00:03:00.093 because the low because of the

NOTE Confidence: 0.9335265

00:03:00.093 --> 00:03:01.995 low material we start with that.

NOTE Confidence: 0.9335265

00:03:02.000 --> 00:03:04.128 That is the content there any content

NOTE Confidence: 0.9335265

00:03:04.128 --> 00:03:07.036 of a single cell and also the elevated

NOTE Confidence: 0.9335265

00:03:07.036 --> 00:03:09.380 number of amplification cycles that are

NOTE Confidence: 0.9335265

00:03:09.380 --> 00:03:11.830 necessary in order to amplify the signal.

NOTE Confidence: 0.9335265

00:03:11.830 --> 00:03:14.198 So after the mapping of the reason we

NOTE Confidence: 0.9335265

00:03:14.198 --> 00:03:16.515 arrived at these gene expression matrix

NOTE Confidence: 0.9335265

00:03:16.515 --> 00:03:18.999 where you have each column represents.

NOTE Confidence: 0.9335265

00:03:19.000 --> 00:03:22.330 One of the cell of our sample and each

NOTE Confidence: 0.9335265

00:03:22.330 --> 00:03:25.829 row is a gene and already last time we

NOTE Confidence: 0.9335265

00:03:25.829 --> 00:03:29.593 saw this fact that if you compare bulk

NOTE Confidence: 0.9335265

00:03:29.593 --> 00:03:33.208 versus single cell matrix at the single cell,

NOTE Confidence: 0.9335265

00:03:33.208 --> 00:03:35.293 one is has lower numbers,

NOTE Confidence: 0.9335265

00:03:35.300 --> 00:03:37.918 lower counts and that means that we

NOTE Confidence: 0.9335265

00:03:37.918 --> 00:03:40.443 have a higher potential contribution of

NOTE Confidence: 0.9335265

00:03:40.443 --> 00:03:44.080 noise and also we have a several zeros.

NOTE Confidence: 0.9335265

00:03:44.080 --> 00:03:48.066 So like 60 to 80% of all the values

NOTE Confidence: 0.9335265

00:03:48.066 --> 00:03:49.488 will be 0.

NOTE Confidence: 0.9335265

00:03:49.490 --> 00:03:52.122 And the problem is that many of these

NOTE Confidence: 0.9335265

00:03:52.122 --> 00:03:54.389 zeros are not biologically true,

NOTE Confidence: 0.9335265

00:03:54.390 --> 00:03:57.099 so it doesn't mean that the gene

NOTE Confidence: 0.9335265

00:03:57.099 --> 00:03:59.290 is not expressed in the cell,

NOTE Confidence: 0.9335265

00:03:59.290 --> 00:04:01.330 but they are technical because

NOTE Confidence: 0.9335265

00:04:01.330 --> 00:04:03.370 they were not detected during

NOTE Confidence: 0.9335265

00:04:03.438 --> 00:04:05.410 our any capturing approaches.

NOTE Confidence: 0.9335265

00:04:05.410 --> 00:04:08.546 So that's the main difference in terms

NOTE Confidence: 0.9335265

00:04:08.546 --> 00:04:12.009 of number with respect to bug RNA seek.

NOTE Confidence: 0.9335265

00:04:12.010 --> 00:04:14.509 So the first step said that we

NOTE Confidence: 0.9335265

00:04:14.509 --> 00:04:17.289 cover all the preprocessing steps.

NOTE Confidence: 0.9335265

00:04:17.290 --> 00:04:19.490 So after the digital account

NOTE Confidence: 0.9335265

00:04:19.490 --> 00:04:21.250 matrix arriving at that,

NOTE Confidence: 0.9335265

00:04:21.250 --> 00:04:24.295 try to remove a basically put cells

NOTE Confidence: 0.9335265

00:04:24.295 --> 00:04:26.736 that are potentially low of low

NOTE Confidence: 0.9335265

00:04:26.736 --> 00:04:29.836 quality and and also gene set that are

NOTE Confidence: 0.9335265

00:04:29.836 --> 00:04:32.686 potentially irrelevant for our analysis.

NOTE Confidence: 0.9335265

00:04:32.690 --> 00:04:35.826 So the first step in the preprocessing.

NOTE Confidence: 0.9335265

00:04:35.830 --> 00:04:39.208 Is that we want to remove?

NOTE Confidence: 0.9335265

00:04:39.210 --> 00:04:42.738 Empty droplets or dying cells, so it could.

NOTE Confidence: 0.9335265

00:04:42.738 --> 00:04:45.354 It could happen that during the

NOTE Confidence: 0.9335265

00:04:45.354 --> 00:04:47.200 preparation of our libraries,

NOTE Confidence: 0.9335265

00:04:47.200 --> 00:04:48.272 some cells,

NOTE Confidence: 0.9335265

00:04:48.272 --> 00:04:51.488 some droplets are empty or filled

NOTE Confidence: 0.9335265

00:04:51.488 --> 00:04:55.059 there with the cells that are dying.

NOTE Confidence: 0.9335265

00:04:55.060 --> 00:04:57.985 So usually I wait to spot these are is

NOTE Confidence: 0.9335265

00:04:57.985 --> 00:05:01.247 a quality of the data so that we can.

NOTE Confidence: 0.9335265

00:05:01.250 --> 00:05:04.175 What we can do is we can count the

NOTE Confidence: 0.9335265

00:05:04.175 --> 00:05:06.861 number of reads or the number of UM

NOTE Confidence: 0.9335265

00:05:06.861 --> 00:05:09.510 eyes that we detect in each cell.

NOTE Confidence: 0.9335265

00:05:09.510 --> 00:05:12.009 That's the sum of the number of

NOTE Confidence: 0.9335265

00:05:12.009 --> 00:05:14.121 unique reads that are aligned for

NOTE Confidence: 0.9335265

00:05:14.121 --> 00:05:16.778 each cell and we we can rank the

NOTE Confidence: 0.9335265

00:05:16.778 --> 00:05:19.137 cells from the the one with more.

NOTE Confidence: 0.9335265

00:05:19.140 --> 00:05:21.948 Umm I with the one with less you MI

NOTE Confidence: 0.9335265

00:05:21.948 --> 00:05:25.069 and we have this sort of distribution.

NOTE Confidence: 0.9335265

00:05:25.070 --> 00:05:28.070 And then we can decide to remove the

NOTE Confidence: 0.9335265

00:05:28.070 --> 00:05:30.805 bottom cells that you see here in red,

NOTE Confidence: 0.9335265

00:05:30.810 --> 00:05:34.107 the one the cells where the UMI

NOTE Confidence: 0.9335265

00:05:34.107 --> 00:05:36.589 is number is very low.

NOTE Confidence: 0.9335265

00:05:36.590 --> 00:05:38.830 So this is a onesie strategy to

NOTE Confidence: 0.9335265

00:05:38.830 --> 00:05:40.480 remove ourselves where we don't

NOTE Confidence: 0.9335265

00:05:40.480 --> 00:05:42.045 have coverage of many genes.

NOTE Confidence: 0.9335265

00:05:42.050 --> 00:05:44.290 We don't have a lot of reads,

NOTE Confidence: 0.9335265

00:05:44.290 --> 00:05:47.050 and likely the it's.

NOTE Confidence: 0.9335265

00:05:47.050 --> 00:05:48.762 South of something wrong

NOTE Confidence: 0.9335265

00:05:48.762 --> 00:05:50.046 during the preparation.

NOTE Confidence: 0.9335265

00:05:50.050 --> 00:05:50.880 For example,

NOTE Confidence: 0.9335265

00:05:50.880 --> 00:05:53.370 the IT was the droplet was

NOTE Confidence: 0.9335265

00:05:53.370 --> 00:05:56.060 slow or this cell was dying.

NOTE Confidence: 0.9335265

00:05:56.060 --> 00:05:58.478 Another way to capture dying to

NOTE Confidence: 0.9335265

00:05:58.478 --> 00:06:01.054 remove dying cells is that usually

NOTE Confidence: 0.9335265

00:06:01.054 --> 00:06:03.700 dying cells are associated with a

NOTE Confidence: 0.9335265

00:06:03.700 --> 00:06:06.795 high number of reads that mapped to

NOTE Confidence: 0.9335265

00:06:06.795 --> 00:06:09.475 mitochondrial genes so they have dying.

NOTE Confidence: 0.9335265

00:06:09.475 --> 00:06:11.110 Cells have extensive

NOTE Confidence: 0.9335265

00:06:11.110 --> 00:06:12.200 mitochondrial contamination.

NOTE Confidence: 0.9863362

00:06:12.200 --> 00:06:14.672 And so one can quantify the number of

NOTE Confidence: 0.9863362

00:06:14.672 --> 00:06:17.099 reads that map to mitochondrial genes.

NOTE Confidence: 0.9863362

00:06:17.100 --> 00:06:21.834 I think there are 40 genes in the human.

NOTE Confidence: 0.9863362

00:06:21.840 --> 00:06:23.755 Human cells, they are associated

NOTE Confidence: 0.9863362

00:06:23.755 --> 00:06:25.287 with the mitochondrial chromosome,

NOTE Confidence: 0.9863362

00:06:25.290 --> 00:06:28.602 and if these numbers, if the number of

NOTE Confidence: 0.9863362

00:06:28.602 --> 00:06:30.647 mitochondrial reads is less than 5%,

NOTE Confidence: 0.9863362

00:06:30.650 --> 00:06:32.560 then you keep the cell.

NOTE Confidence: 0.9863362

00:06:32.560 --> 00:06:35.624 If it's higher than that 10 or 20%,

NOTE Confidence: 0.9863362

00:06:35.630 --> 00:06:38.045 then you remove the entire set because

NOTE Confidence: 0.9863362

00:06:38.045 --> 00:06:40.988 there is a high probability that the

NOTE Confidence: 0.9863362

00:06:40.988 --> 00:06:43.736 these high numbers contamination is due

NOTE Confidence: 0.9863362

00:06:43.811 --> 00:06:46.347 to the fact that the cell was dying.

NOTE Confidence: 0.972797

00:06:49.340 --> 00:06:51.356 Uhm then, on the other side

NOTE Confidence: 0.972797

00:06:51.356 --> 00:06:53.330 we want also to remove.

NOTE Confidence: 0.972797

00:06:53.330 --> 00:06:56.002 So all the technique is based on the

NOTE Confidence: 0.972797

00:06:56.002 --> 00:06:58.420 fact that we isolate single cell.

NOTE Confidence: 0.972797

00:06:58.420 --> 00:07:01.400 But sometimes this doesn't.

NOTE Confidence: 0.972797

00:07:01.400 --> 00:07:02.110 Happen properly,

NOTE Confidence: 0.972797

00:07:02.110 --> 00:07:04.950 so it means that it can happen that

NOTE Confidence: 0.972797

00:07:05.020 --> 00:07:07.407 two cells share the same barcode or

NOTE Confidence: 0.972797

00:07:07.407 --> 00:07:10.210 two cells were not physically separated,

NOTE Confidence: 0.972797

00:07:10.210 --> 00:07:13.266 so they were included in the same droplet.

NOTE Confidence: 0.972797

00:07:13.270 --> 00:07:16.166 For example, if we are using the droplet

NOTE Confidence: 0.972797

00:07:16.166 --> 00:07:18.943 approach and so we want to identify

NOTE Confidence: 0.972797

00:07:18.943 --> 00:07:20.923 possible doublets and remove those,

NOTE Confidence: 0.972797

00:07:20.930 --> 00:07:26.160 so a double letter is a we define doublet as.

NOTE Confidence: 0.972797

00:07:26.160 --> 00:07:28.476 A droplet or as a isolation

NOTE Confidence: 0.972797

00:07:28.476 --> 00:07:30.540 not of one single cell,

NOTE Confidence: 0.972797

00:07:30.540 --> 00:07:32.928 but of two or more sets.

NOTE Confidence: 0.972797

00:07:32.930 --> 00:07:35.866 The most common event is that you have

NOTE Confidence: 0.972797

00:07:35.866 --> 00:07:38.900 two cells included in the same droplet.

NOTE Confidence: 0.972797

00:07:38.900 --> 00:07:41.276 So when you develop are they?

NOTE Confidence: 0.972797

00:07:41.280 --> 00:07:42.561 Single cell techniques?

NOTE Confidence: 0.972797

00:07:42.561 --> 00:07:45.123 Are there are experimental ways to

NOTE Confidence: 0.972797

00:07:45.123 --> 00:07:47.100 evaluate the probability to have

NOTE Confidence: 0.972797

00:07:47.100 --> 00:07:49.640 doubles and the approaches that we use?

NOTE Confidence: 0.972797

00:07:49.640 --> 00:07:51.480 There are spacious mixing,

NOTE Confidence: 0.972797

00:07:51.480 --> 00:07:54.240 so you combine for example population

NOTE Confidence: 0.972797

00:07:54.312 --> 00:07:56.538 of human cells and mouse cells.

NOTE Confidence: 0.972797

00:07:56.540 --> 00:07:59.762 And then use when you map the reads from

NOTE Confidence: 0.972797

00:07:59.762 --> 00:08:02.749 each cell you see or you see how many

NOTE Confidence: 0.972797

00:08:02.749 --> 00:08:06.419 for how many cells you have a double mapping.

NOTE Confidence: 0.972797

00:08:06.420 --> 00:08:09.100 So how for how many cells some of

NOTE Confidence: 0.972797

00:08:09.100 --> 00:08:11.906 your reads mapped to the human genome,

NOTE Confidence: 0.972797

00:08:11.910 --> 00:08:14.106 some of your reads mapped to

NOTE Confidence: 0.972797

00:08:14.106 --> 00:08:15.204 the mouse genome.

NOTE Confidence: 0.972797

00:08:15.210 --> 00:08:19.610 You see here in this plot the mapping of the.

NOTE Confidence: 0.972797

00:08:19.610 --> 00:08:20.272 The cells,

NOTE Confidence: 0.972797

00:08:20.272 --> 00:08:21.927 so on the human transcript

NOTE Confidence: 0.972797

00:08:21.927 --> 00:08:24.169 and on the mouse transcript.

NOTE Confidence: 0.972797

00:08:24.170 --> 00:08:26.837 So all these cells are means that

NOTE Confidence: 0.972797

00:08:26.837 --> 00:08:29.109 they contain only mouse a cells.

NOTE Confidence: 0.972797

00:08:29.110 --> 00:08:31.390 Here they contain only human cells.

NOTE Confidence: 0.972797

00:08:31.390 --> 00:08:34.050 What you see here is the identification

NOTE Confidence: 0.972797

00:08:34.050 --> 00:08:34.810 of doublets,

NOTE Confidence: 0.972797

00:08:34.810 --> 00:08:37.090 because here the content is mixed,

NOTE Confidence: 0.972797

00:08:37.090 --> 00:08:38.990 you have something from mouse,

NOTE Confidence: 0.972797

00:08:38.990 --> 00:08:40.127 something from human,

NOTE Confidence: 0.972797

00:08:40.127 --> 00:08:42.780 and this is likely to be because

NOTE Confidence: 0.972797

00:08:42.854 --> 00:08:44.960 one mouse and one human cells

NOTE Confidence: 0.972797

00:08:44.960 --> 00:08:47.349 were included in the same droplet.

NOTE Confidence: 0.972797

00:08:47.350 --> 00:08:49.720 So the comparison of these two.

NOTE Confidence: 0.972797

00:08:49.720 --> 00:08:52.058 Plot is something to say that the

NOTE Confidence: 0.972797

00:08:52.058 --> 00:08:53.515 probability to having doublets

NOTE Confidence: 0.972797

00:08:53.515 --> 00:08:55.460 obviously depends on the concentration

NOTE Confidence: 0.972797

00:08:55.460 --> 00:08:57.929 of your cells at the beginning.

NOTE Confidence: 0.972797

00:08:57.930 --> 00:08:59.522 That's why, for example,

NOTE Confidence: 0.972797

00:08:59.522 --> 00:09:02.670 here when you have 12.5 cells where we

NOTE Confidence: 0.972797

00:09:02.670 --> 00:09:05.382 call it are you have very few events,

NOTE Confidence: 0.972797

00:09:05.390 --> 00:09:07.250 only one droplet doublet event.

NOTE Confidence: 0.972797

00:09:07.250 --> 00:09:09.085 When you increase the contrast

NOTE Confidence: 0.972797

00:09:09.085 --> 00:09:10.553 concentration of cells probably

NOTE Confidence: 0.972797

00:09:10.553 --> 00:09:12.467 increase the efficiency of sequencing.

NOTE Confidence: 0.972797

00:09:12.470 --> 00:09:14.335 Are your single self because

NOTE Confidence: 0.972797

00:09:14.335 --> 00:09:16.200 you have less empty droplet,

NOTE Confidence: 0.972797

00:09:16.200 --> 00:09:18.205 but you also increase the

NOTE Confidence: 0.972797

00:09:18.205 --> 00:09:19.809 probability you have doublets.

NOTE Confidence: 0.972797

00:09:19.810 --> 00:09:21.390 That you see here.

NOTE Confidence: 0.972797

00:09:21.390 --> 00:09:23.365 So the number here increase.

NOTE Confidence: 0.972797

00:09:23.370 --> 00:09:25.340 So obviously this is possible.

NOTE Confidence: 0.972797

00:09:25.340 --> 00:09:27.265 This evaluation is possible because

NOTE Confidence: 0.972797

00:09:27.265 --> 00:09:29.690 you are mixing two species before,

NOTE Confidence: 0.972797

00:09:29.690 --> 00:09:31.665 but it's not always feasible

NOTE Confidence: 0.972797

00:09:31.665 --> 00:09:32.850 in our experiment,

NOTE Confidence: 0.972797

00:09:32.850 --> 00:09:35.498 so we need to have a way to

NOTE Confidence: 0.972797

00:09:35.498 --> 00:09:37.598 predict the possibility that a

NOTE Confidence: 0.972797

00:09:37.598 --> 00:09:40.748 cell was not really a single cell,

NOTE Confidence: 0.972797

00:09:40.750 --> 00:09:42.720 but it was a doublet,

NOTE Confidence: 0.972797

00:09:42.720 --> 00:09:44.815 so there are computational approaches

NOTE Confidence: 0.972797

00:09:44.815 --> 00:09:48.124 that try to evaluate for each of these

NOTE Confidence: 0.972797

00:09:48.124 --> 00:09:50.777 cells that we obtain the possibility that.

NOTE Confidence: 0.972797

00:09:50.780 --> 00:09:52.316 It's not really a single cell,

NOTE Confidence: 0.972797

00:09:52.320 --> 00:09:55.160 but it's a doublet.

NOTE Confidence: 0.972797

00:09:55.160 --> 00:09:57.670 So there are many progress,

NOTE Confidence: 0.972797

00:09:57.670 --> 00:09:58.157 many,

NOTE Confidence: 0.972797

00:09:58.157 --> 00:10:00.592 many procedures that are used

NOTE Confidence: 0.972797

00:10:00.592 --> 00:10:04.158 at a common approach is these in

NOTE Confidence: 0.972797

00:10:04.158 --> 00:10:06.194 silico simulation of tablets.

NOTE Confidence: 0.972797

00:10:06.200 --> 00:10:09.620 This means that you have your

NOTE Confidence: 0.972797

00:10:09.620 --> 00:10:11.900 matrix with digital counts

NOTE Confidence: 0.97385466

00:10:12.002 --> 00:10:13.628 with your cells.

NOTE Confidence: 0.97385466

00:10:13.630 --> 00:10:15.670 You simulate the doublet by

NOTE Confidence: 0.97385466

00:10:15.670 --> 00:10:17.710 selecting two random cells to

NOTE Confidence: 0.97385466

00:10:17.787 --> 00:10:20.077 random cells and combining them,

NOTE Confidence: 0.97385466

00:10:20.080 --> 00:10:23.520 meaning that for each of these two cells,

NOTE Confidence: 0.97385466

00:10:23.520 --> 00:10:25.396 you calculate the hypothetical

NOTE Confidence: 0.97385466

00:10:25.396 --> 00:10:28.210 cell that contains the sum of

NOTE Confidence: 0.97385466

00:10:28.289 --> 00:10:30.395 the reeds of the two cells.

NOTE Confidence: 0.97385466

00:10:30.400 --> 00:10:33.410 So this is an in silico tablet,

NOTE Confidence: 0.97385466

00:10:33.410 --> 00:10:36.446 so you generate thousands of these

NOTE Confidence: 0.97385466

00:10:36.446 --> 00:10:40.030 in silico tablets and you and the

NOTE Confidence: 0.97385466

00:10:40.030 --> 00:10:42.958 procedure is to mix these doubles

NOTE Confidence: 0.97385466

00:10:42.958 --> 00:10:45.349 together with the real cells.

NOTE Confidence: 0.97385466

00:10:45.350 --> 00:10:47.968 And so that they are analyzed together.

NOTE Confidence: 0.97385466

00:10:47.970 --> 00:10:50.130 So at some point of the

NOTE Confidence: 0.97385466

00:10:50.130 --> 00:10:52.460 analysis that we will see later,

NOTE Confidence: 0.97385466

00:10:52.460 --> 00:10:54.330 cells can be clustered together,

NOTE Confidence: 0.97385466

00:10:54.330 --> 00:10:57.050 and so for each of the original cell

NOTE Confidence: 0.97385466

00:10:57.050 --> 00:11:00.276 one can see how many in silico tablets

NOTE Confidence: 0.97385466

00:11:00.276 --> 00:11:02.930 are in the surrounding of the cell.

NOTE Confidence: 0.97385466

00:11:02.930 --> 00:11:05.698 So for each cell I can calculate how

NOTE Confidence: 0.97385466

00:11:05.698 --> 00:11:07.545 many neighbors in the neighborhood

NOTE Confidence: 0.97385466

00:11:07.545 --> 00:11:10.554 how many real cells there are and how

NOTE Confidence: 0.97385466

00:11:10.554 --> 00:11:12.649 many simulated tablets there are.

NOTE Confidence: 0.97385466

00:11:12.650 --> 00:11:15.359 And the principle is that the ratio.

NOTE Confidence: 0.97385466

00:11:15.360 --> 00:11:17.634 Between the simulated tablets and the

NOTE Confidence: 0.97385466

00:11:17.634 --> 00:11:20.639 real cells is a score that represents

NOTE Confidence: 0.97385466

00:11:20.639 --> 00:11:22.999 the possibility the probability of

NOTE Confidence: 0.97385466

00:11:22.999 --> 00:11:25.586 this cell to be a tablet itself.

NOTE Confidence: 0.97385466

00:11:25.590 --> 00:11:28.742 So the principle is that if my cell

NOTE Confidence: 0.97385466

00:11:28.742 --> 00:11:31.310 is surrounded by in silico tablets,

NOTE Confidence: 0.97385466

00:11:31.310 --> 00:11:34.360 then it's likely a tablet.

NOTE Confidence: 0.97385466

00:11:34.360 --> 00:11:36.640 If it's surrounded by if the

NOTE Confidence: 0.97385466

00:11:36.640 --> 00:11:39.348 tablets are all far from my cells,

NOTE Confidence: 0.97385466

00:11:39.350 --> 00:11:42.038 then probably these cells are not tablets.

NOTE Confidence: 0.98445684

00:11:44.190 --> 00:11:45.758 Was this step clear?

NOTE Confidence: 0.9851623

00:11:47.950 --> 00:11:49.386 Kind of sort of somehow

NOTE Confidence: 0.9851623

00:11:49.386 --> 00:11:51.060 you you teach it what a

NOTE Confidence: 0.9851623

00:11:51.127 --> 00:11:52.390 doublet looks like,

NOTE Confidence: 0.9851623

00:11:52.390 --> 00:11:54.609 and then it can find those things,

NOTE Confidence: 0.9851623

00:11:54.610 --> 00:11:56.829 or you teach it, but a double

NOTE Confidence: 0.9851623

00:11:56.829 --> 00:11:59.044 it looks like, and it says OK,

NOTE Confidence: 0.9851623

00:11:59.044 --> 00:12:00.624 I'm certain percentage should be

NOTE Confidence: 0.9851623

00:12:00.630 --> 00:12:02.849 doubled, yes, so you build the doublets,

NOTE Confidence: 0.9851623

00:12:02.850 --> 00:12:05.386 taking two random cells. After I got that

NOTE Confidence: 0.9851623

00:12:05.390 --> 00:12:07.430 part, I just don't understand how

NOTE Confidence: 0.9851623

00:12:07.430 --> 00:12:09.510 that helps you identify a real one.

NOTE Confidence: 0.9589141

00:12:10.820 --> 00:12:13.018 Yeah, so the idea is that yeah,

NOTE Confidence: 0.9589141

00:12:13.020 --> 00:12:15.491 yeah is is that real tablets will

NOTE Confidence: 0.9589141

00:12:15.491 --> 00:12:17.647 be surrounded by in silico tablets

NOTE Confidence: 0.9589141

00:12:17.647 --> 00:12:20.481 while a real cells will be far from

NOTE Confidence: 0.9589141

00:12:20.481 --> 00:12:23.430 the in silico tablets. OK OK, I have

NOTE Confidence: 0.96831703

00:12:23.430 --> 00:12:25.602 a related, maybe a related question

NOTE Confidence: 0.96831703

00:12:25.602 --> 00:12:28.068 'cause the idea of a doublet is

NOTE Confidence: 0.96831703

00:12:28.068 --> 00:12:30.273 that you have jeans from more than

NOTE Confidence: 0.96831703

00:12:30.347 --> 00:12:32.627 one cell that are being sequenced.

NOTE Confidence: 0.96831703

00:12:32.630 --> 00:12:34.580 We have this thing that happened

NOTE Confidence: 0.96831703

00:12:34.580 --> 00:12:37.228 and I'm I'm a basket in general,

NOTE Confidence: 0.96831703

00:12:37.230 --> 00:12:39.617 'cause I'm assuming it would be true

NOTE Confidence: 0.96831703

00:12:39.617 --> 00:12:42.018 for other people as well when we

NOTE Confidence: 0.96831703

00:12:42.018 --> 00:12:43.986 did parathyroid the cells that make

NOTE Confidence: 0.96831703

00:12:44.053 --> 00:12:45.948 parathyroid hormone have a humongous

NOTE Confidence: 0.96831703

00:12:45.948 --> 00:12:48.624 amount of PTH as their you know,

NOTE Confidence: 0.96831703

00:12:48.624 --> 00:12:49.398 main transcript.

NOTE Confidence: 0.96831703

00:12:49.398 --> 00:12:51.720 The ones that were negative ETH.

NOTE Confidence: 0.96831703

00:12:51.720 --> 00:12:53.040 All had some PTH,

NOTE Confidence: 0.96831703

00:12:53.040 --> 00:12:55.020 nothing on the order of like.

NOTE Confidence: 0.96831703

00:12:55.020 --> 00:12:57.330 Let's say we had 1000 for PTH.

NOTE Confidence: 0.96831703

00:12:57.330 --> 00:12:59.190 We'd have like three or one

NOTE Confidence: 0.96831703

00:12:59.190 --> 00:13:00.960 or two in the cells.

NOTE Confidence: 0.96831703

00:13:00.960 --> 00:13:03.290 That should have been negative.

NOTE Confidence: 0.96831703

00:13:03.290 --> 00:13:04.850 And it's hard to believe that

NOTE Confidence: 0.96831703

00:13:04.850 --> 00:13:06.268 every cell in the parathyroid

NOTE Confidence: 0.96831703

00:13:06.268 --> 00:13:08.164 actually has some RNA in it.

NOTE Confidence: 0.96831703

00:13:08.170 --> 00:13:09.458 For this parathyroid hormone,

NOTE Confidence: 0.96831703

00:13:09.458 --> 00:13:11.798 it's much more likely that the cell

NOTE Confidence: 0.96831703

00:13:11.798 --> 00:13:13.543 that looks like an endothelial

NOTE Confidence: 0.96831703

00:13:13.543 --> 00:13:14.939 cell really isn't endothelial

NOTE Confidence: 0.96831703

00:13:14.996 --> 00:13:16.766 cell in those three little reeds.

NOTE Confidence: 0.96831703

00:13:16.770 --> 00:13:17.248 We're wrong,

NOTE Confidence: 0.96831703

00:13:17.248 --> 00:13:18.682 but I don't know how that

NOTE Confidence: 0.96831703

00:13:18.682 --> 00:13:19.650 would have happened.

NOTE Confidence: 0.9832647

00:13:20.940 --> 00:13:23.635 Yeah, I don't know if that could

NOTE Confidence: 0.9832647

00:13:23.635 --> 00:13:28.170 be also like a contamination. Uhm?

NOTE Confidence: 0.9832647

00:13:28.170 --> 00:13:30.433 But if it's three instead of 3000, well,

NOTE Confidence: 0.9832647

00:13:30.433 --> 00:13:32.697 it that's a good signal to noise ratio.

NOTE Confidence: 0.9832647

00:13:32.700 --> 00:13:34.674 I would say I. I absolutely agree.

NOTE Confidence: 0.97211456

00:13:34.680 --> 00:13:36.766 I just thought there was maybe some

NOTE Confidence: 0.97211456

00:13:36.766 --> 00:13:38.236 general principle in single cells

NOTE Confidence: 0.97211456

00:13:38.236 --> 00:13:40.056 seek that we needed to look at,

NOTE Confidence: 0.97211456

00:13:40.060 --> 00:13:41.470 but that's not the case.

NOTE Confidence: 0.98215544

00:13:42.550 --> 00:13:44.182 No, the only things come coming to my

NOTE Confidence: 0.98215544

00:13:44.182 --> 00:13:46.039 mind is this for the possibility of it.

NOTE Confidence: 0.98215544

00:13:46.040 --> 00:13:48.536 Yeah there is. There are some.

NOTE Confidence: 0.98215544

00:13:48.540 --> 00:13:49.698 Possibility of supernatant

NOTE Confidence: 0.98215544

00:13:49.698 --> 00:13:52.014 contamination so that you get some

NOTE Confidence: 0.98215544

00:13:52.014 --> 00:13:53.837 early that is in the solution.

NOTE Confidence: 0.98215544

00:13:53.840 --> 00:13:55.954 For example, it could be something,

NOTE Confidence: 0.98215544

00:13:55.954 --> 00:13:57.370 especially if it's abundant,

NOTE Confidence: 0.98215544

00:13:57.370 --> 00:13:58.778 so it could be.

NOTE Confidence: 0.9823647

00:13:58.780 --> 00:14:01.678 Thank you that Diane maybe one other

NOTE Confidence: 0.9823647

00:14:01.678 --> 00:14:04.158 explanation for your finding is that.

NOTE Confidence: 0.9823647

00:14:04.160 --> 00:14:05.770 That those cells have some

NOTE Confidence: 0.9823647

00:14:05.770 --> 00:14:07.058 illegitimate transcription going on,

NOTE Confidence: 0.9823647

00:14:07.060 --> 00:14:11.110 and so you know that could be an explanation.

NOTE Confidence: 0.9823647

00:14:11.110 --> 00:14:13.378 Yes, absolutely, but that would be real.

NOTE Confidence: 0.9823647

00:14:13.380 --> 00:14:15.030 That would suggest that endothelial

NOTE Confidence: 0.9823647

00:14:15.030 --> 00:14:17.036 cells in the parathyroid like to

NOTE Confidence: 0.9823647

00:14:17.036 --> 00:14:18.556 turn on some parathyroid hormone,

NOTE Confidence: 0.9823647

00:14:18.560 --> 00:14:21.470 which would be a little weird.

NOTE Confidence: 0.9823647

00:14:21.470 --> 00:14:23.150 But the definition of illegitimate

NOTE Confidence: 0.9823647

00:14:23.150 --> 00:14:24.830 transcription is expression of any

NOTE Confidence: 0.9823647

00:14:24.888 --> 00:14:26.448 gene transcripted any cell type.

NOTE Confidence: 0.9823647

00:14:26.450 --> 00:14:29.929 I mean that's fine, but you know.

NOTE Confidence: 0.9823647

00:14:29.930 --> 00:14:31.710 Is that that parathyroid tissue

NOTE Confidence: 0.9823647

00:14:31.710 --> 00:14:34.060 that was sequenced is not adenoma,

NOTE Confidence: 0.9823647

00:14:34.060 --> 00:14:35.930 it's normal, it's abnormal, OK?

NOTE Confidence: 0.9875988

00:14:38.590 --> 00:14:40.750 Diane were those cells like washed

NOTE Confidence: 0.9875988

00:14:40.750 --> 00:14:43.370 before they were put on the sequencer?

NOTE Confidence: 0.9875988

00:14:43.370 --> 00:14:44.450 'cause maybe there's.

NOTE Confidence: 0.9875988

00:14:44.450 --> 00:14:45.890 Maybe somehow some transcripts

NOTE Confidence: 0.9875988

00:14:45.890 --> 00:14:47.450 are just leaking through.

NOTE Confidence: 0.9875988

00:14:47.450 --> 00:14:49.474 If there's a lot of them they were.

NOTE Confidence: 0.9875988

00:14:49.480 --> 00:14:51.251 Yeah, yeah, that gets back to maybe

NOTE Confidence: 0.9875988

00:14:51.251 --> 00:14:52.504 the contamination. I don't know.

NOTE Confidence: 0.9875988

00:14:52.504 --> 00:14:54.540 I thought that the machine washed the cell,

NOTE Confidence: 0.9875988

00:14:54.540 --> 00:14:56.058 but I don't know specifically I.

NOTE Confidence: 0.9875988

00:14:56.060 --> 00:14:58.274 I'm sure that the person that kind of goes

NOTE Confidence: 0.9875988

00:14:58.274 --> 00:15:00.097 through all this plumbing to get there,

NOTE Confidence: 0.9875988

00:15:00.100 --> 00:15:01.365 so it's a little surprising

NOTE Confidence: 0.9875988

00:15:01.365 --> 00:15:03.900 that would happen, but maybe.

NOTE Confidence: 0.9875988

00:15:03.900 --> 00:15:04.948 All right, moving on.

NOTE Confidence: 0.9875988

00:15:04.948 --> 00:15:06.520 I thought it was maybe something

NOTE Confidence: 0.9875988

00:15:06.571 --> 00:15:07.636 we all needed to know,

NOTE Confidence: 0.9875988

00:15:07.640 --> 00:15:09.872 but that it seems to be a specific problem.

NOTE Confidence: 0.9875988

00:15:09.880 --> 00:15:10.872 Sorry, cells were definitely

NOTE Confidence: 0.9875988

00:15:10.872 --> 00:15:12.112 washed before they went on.

NOTE Confidence: 0.98769796

00:15:16.930 --> 00:15:18.688 OK, moving on but thank you.

NOTE Confidence: 0.8589663

00:15:20.710 --> 00:15:24.246 Uh, OK, so the next step after this,

NOTE Confidence: 0.8589663

00:15:24.250 --> 00:15:27.434 so these were to remove cells that we

NOTE Confidence: 0.8589663

00:15:27.434 --> 00:15:30.428 didn't want in the following analysis.

NOTE Confidence: 0.8589663

00:15:30.430 --> 00:15:33.088 The next step is the normalization.

NOTE Confidence: 0.8589663

00:15:33.090 --> 00:15:36.177 So the normalization, as in any experiment,

NOTE Confidence: 0.8589663

00:15:36.180 --> 00:15:38.922 has the aim of removing systematic

NOTE Confidence: 0.8589663

00:15:38.922 --> 00:15:40.750 differences in the quantification

NOTE Confidence: 0.8589663

00:15:40.823 --> 00:15:42.367 of genes between cells.

NOTE Confidence: 0.8589663

00:15:42.370 --> 00:15:45.247 So we saw the methods that are

NOTE Confidence: 0.8589663

00:15:45.247 --> 00:15:48.109 used for the bulk RNA secret.

NOTE Confidence: 0.8589663

00:15:48.110 --> 00:15:50.858 So the simplest approach is the.

NOTE Confidence: 0.8589663

00:15:50.860 --> 00:15:53.224 Library size normalization so that each

NOTE Confidence: 0.8589663

00:15:53.224 --> 00:15:56.283 cell so the the signal for each from

NOTE Confidence: 0.8589663

00:15:56.283 --> 00:15:59.141 each cell is divided for the total sum

NOTE Confidence: 0.8589663

00:15:59.141 --> 00:16:01.837 of the Council of the number of reads

NOTE Confidence: 0.8589663

00:16:01.840 --> 00:16:05.125 or Umm I across all genes for each cell.

NOTE Confidence: 0.8589663

00:16:05.130 --> 00:16:06.960 So this is simplest approach.

NOTE Confidence: 0.8589663

00:16:06.960 --> 00:16:08.424 Normalization for the size

NOTE Confidence: 0.8589663

00:16:08.424 --> 00:16:10.620 of the library for each cell.

NOTE Confidence: 0.8589663

00:16:10.620 --> 00:16:12.296 The questionable assumption of

NOTE Confidence: 0.8589663

00:16:12.296 --> 00:16:14.391 these approaches is that you're

NOTE Confidence: 0.8589663

00:16:14.391 --> 00:16:16.098 assuming that each cell should

NOTE Confidence: 0.8589663

00:16:16.098 --> 00:16:17.940 have the same number of reads.

NOTE Confidence: 0.8589663

00:16:17.940 --> 00:16:19.074 This is problematic.

NOTE Confidence: 0.8589663

00:16:19.074 --> 00:16:20.964 It's problematic in the biker.

NOTE Confidence: 0.8589663

00:16:20.970 --> 00:16:23.609 I have a secret to assume that

NOTE Confidence: 0.8589663

00:16:23.609 --> 00:16:25.807 all your samples should have

NOTE Confidence: 0.8589663

00:16:25.807 --> 00:16:28.753 approximately the same number of RNA.

NOTE Confidence: 0.8589663

00:16:28.760 --> 00:16:30.580 It's even more, uh,

NOTE Confidence: 0.8589663

00:16:30.580 --> 00:16:33.310 questionable for the single cell because

NOTE Confidence: 0.8589663

00:16:33.394 --> 00:16:35.998 we know some cells depending on the

NOTE Confidence: 0.8589663

00:16:35.998 --> 00:16:38.557 cell type can have different number

NOTE Confidence: 0.8589663

00:16:38.557 --> 00:16:41.287 of model of RNA molecules depending

NOTE Confidence: 0.8589663

00:16:41.287 --> 00:16:44.158 on the translation and transcript.

NOTE Confidence: 0.8589663

00:16:44.158 --> 00:16:47.168 Depending on the transcription activities.

NOTE Confidence: 0.8589663

00:16:47.170 --> 00:16:48.262 So the alternative,

NOTE Confidence: 0.8589663

00:16:48.262 --> 00:16:50.082 the main alternatives that are

NOTE Confidence: 0.8589663

00:16:50.082 --> 00:16:52.318 used to this simplest approach,

NOTE Confidence: 0.8589663

00:16:52.320 --> 00:16:56.058 is to use a spike in RNA.

NOTE Confidence: 0.8589663

00:16:56.060 --> 00:16:56.414 Uhm,

NOTE Confidence: 0.8589663

00:16:56.414 --> 00:16:57.830 there are many benches,

NOTE Confidence: 0.8589663

00:16:57.830 --> 00:16:59.954 many kids of spiking RNAs that

NOTE Confidence: 0.8589663

00:16:59.954 --> 00:17:01.016 are now available,

NOTE Confidence: 0.8589663

00:17:01.020 --> 00:17:03.510 and the assumption for this for

NOTE Confidence: 0.8589663

00:17:03.510 --> 00:17:05.601 when normalizing for this begin

NOTE Confidence: 0.8589663

00:17:05.601 --> 00:17:08.051 is that inside each cell there is

NOTE Confidence: 0.8589663

00:17:08.051 --> 00:17:10.646 the same amount of spike in RNA's.

NOTE Confidence: 0.8589663

00:17:10.650 --> 00:17:13.380 And then this suggestion that

NOTE Confidence: 0.8589663

00:17:13.380 --> 00:17:16.110 the common suggestion in the.

NOTE Confidence: 0.8589663

00:17:16.110 --> 00:17:18.987 Approach is that it's better to use

NOTE Confidence: 0.8589663

00:17:18.987 --> 00:17:21.691 a single cell specific methods and

NOTE Confidence: 0.8589663

00:17:21.691 --> 00:17:24.918 it's better not to use the methods

NOTE Confidence: 0.8589663

00:17:25.010 --> 00:17:27.428 that are commonly used in the

NOTE Confidence: 0.8589663

00:17:27.428 --> 00:17:29.495 bike and they seek normalization.

NOTE Confidence: 0.8589663

00:17:29.495 --> 00:17:33.135 The reason for this is that the bulk

NOTE Confidence: 0.8589663

00:17:33.135 --> 00:17:35.675 methods do not take into consideration

NOTE Confidence: 0.8589663

00:17:35.675 --> 00:17:39.380 the fact that most of the values are zeros,

NOTE Confidence: 0.8589663

00:17:39.380 --> 00:17:42.392 and so using by chronic normalization

NOTE Confidence: 0.8589663

00:17:42.392 --> 00:17:44.978 methods could lead that tool

NOTE Confidence: 0.8589663

00:17:44.978 --> 00:17:46.460 very stranger size.

NOTE Confidence: 0.8589663

00:17:46.460 --> 00:17:46.850 Factors.

NOTE Confidence: 0.8589663

00:17:46.850 --> 00:17:49.580 So all these single set specific methods

NOTE Confidence: 0.8589663

00:17:49.580 --> 00:17:51.611 somehow take into consideration this

NOTE Confidence: 0.8589663

00:17:51.611 --> 00:17:53.987 problem of the excessive zeros and

NOTE Confidence: 0.8589663

00:17:53.987 --> 00:17:56.839 they use different strategies to normalize.

NOTE Confidence: 0.8589663

00:17:56.840 --> 00:17:58.695 So there are many methods

NOTE Confidence: 0.8589663

00:17:58.695 --> 00:18:00.179 for the single cell.

NOTE Confidence: 0.8589663

00:18:00.180 --> 00:18:02.030 Some of those consider instead

NOTE Confidence: 0.8589663

00:18:02.030 --> 00:18:04.656 of all the single cells pools of

NOTE Confidence: 0.8589663

00:18:04.656 --> 00:18:06.756 cells so that they normalize that

NOTE Confidence: 0.8589663

00:18:06.756 --> 00:18:08.350 not each single cell,

NOTE Confidence: 0.8589663

00:18:08.350 --> 00:18:10.570 but the normalized groups of cells

NOTE Confidence: 0.8589663

00:18:10.570 --> 00:18:12.800 where the content is summed up,

NOTE Confidence: 0.8589663

00:18:12.800 --> 00:18:16.464 and this somehow reduces the number of zeros.

NOTE Confidence: 0.8589663

00:18:16.470 --> 00:18:18.936 And then another methodology is try

NOTE Confidence: 0.8589663

00:18:18.936 --> 00:18:21.119 to correct to normalize differently

NOTE Confidence: 0.8589663

00:18:21.119 --> 00:18:23.454 for different groups of genes

NOTE Confidence: 0.8589663

00:18:23.454 --> 00:18:25.950 depending on whether they are low.

NOTE Confidence: 0.8589663

00:18:25.950 --> 00:18:27.768 They have low,

NOTE Confidence: 0.8589663

00:18:27.768 --> 00:18:30.798 medium or high expression levels.

NOTE Confidence: 0.8589663

00:18:30.800 --> 00:18:34.288 Uh, so the key point here is that,

NOTE Confidence: 0.8589663

00:18:34.290 --> 00:18:35.595 uh, as usual,

NOTE Confidence: 0.8589663

00:18:35.595 --> 00:18:38.205 the normalization choices affect the results,

NOTE Confidence: 0.8589663

00:18:38.210 --> 00:18:40.716 so this is taken from a paper

NOTE Confidence: 0.8589663

00:18:40.716 --> 00:18:43.473 published last year that was comparing

NOTE Confidence: 0.8589663

00:18:43.473 --> 00:18:45.609 a different normalization methods

NOTE Confidence: 0.8589663

00:18:45.609 --> 00:18:48.677 developed for single cell RNA seek data.

NOTE Confidence: 0.8589663

00:18:48.680 --> 00:18:51.186 So here this is a simple data

NOTE Confidence: 0.8589663

00:18:51.186 --> 00:18:52.260 set of mouse

NOTE Confidence: 0.9192497

00:18:52.351 --> 00:18:54.399 embryonic data where you

NOTE Confidence: 0.9192497

00:18:54.399 --> 00:18:56.959 have two population of cells.

NOTE Confidence: 0.9192497

00:18:56.960 --> 00:19:00.964 Then Veronica stem cells and the method.

NOTE Confidence: 0.9192497

00:19:00.970 --> 00:19:03.256 They are colored according to this,

NOTE Confidence: 0.9192497

00:19:03.260 --> 00:19:05.546 to the two, to the two

NOTE Confidence: 0.9192497

00:19:05.546 --> 00:19:07.070 populations they belong to.

NOTE Confidence: 0.9192497

00:19:07.070 --> 00:19:09.751 So what you see is the result

NOTE Confidence: 0.9192497

00:19:09.751 --> 00:19:11.260 without normalization at all.

NOTE Confidence: 0.9192497

00:19:11.260 --> 00:19:13.990 So it seems to work quite fine

NOTE Confidence: 0.9192497

00:19:13.990 --> 00:19:15.450 even without normalizing it.

NOTE Confidence: 0.9192497

00:19:15.450 --> 00:19:18.156 All this simple normalization here is

NOTE Confidence: 0.9192497

00:19:18.156 --> 00:19:20.711 the library size normalization and also

NOTE Confidence: 0.9192497

00:19:20.711 --> 00:19:23.077 this seems to be working quite fine

NOTE Confidence: 0.9192497

00:19:23.077 --> 00:19:25.350 except for this cell here and then.

NOTE Confidence: 0.9192497

00:19:25.350 --> 00:19:27.750 You see six different methods that

NOTE Confidence: 0.9192497

00:19:27.750 --> 00:19:30.146 were developed only for single cell

NOTE Confidence: 0.9192497

00:19:30.146 --> 00:19:32.318 RNA seek and their divided in.

NOTE Confidence: 0.9192497

00:19:32.320 --> 00:19:35.584 Two groups are based on the fact that

NOTE Confidence: 0.9192497

00:19:35.584 --> 00:19:38.370 they require spiking RNAs to worker,

NOTE Confidence: 0.9192497

00:19:38.370 --> 00:19:41.362 and this is basic German Sam STRT or

NOTE Confidence: 0.9192497

00:19:41.362 --> 00:19:44.417 they do not require speaking RNAs.

NOTE Confidence: 0.9192497

00:19:44.420 --> 00:19:46.862 So the general message here is

NOTE Confidence: 0.9192497

00:19:46.862 --> 00:19:49.170 that depending of these methods,

NOTE Confidence: 0.9192497

00:19:49.170 --> 00:19:51.235 the separation of this population

NOTE Confidence: 0.9192497

00:19:51.235 --> 00:19:53.920 change a lot and different methods,

NOTE Confidence: 0.9192497

00:19:53.920 --> 00:19:57.518 so there is no method that works

NOTE Confidence: 0.9192497

00:19:57.518 --> 00:20:00.139 better for each data set.

NOTE Confidence: 0.9192497

00:20:00.140 --> 00:20:03.276 So I would say it's usually important to

NOTE Confidence: 0.9192497

00:20:03.276 --> 00:20:06.246 to try different methods and depending

NOTE Confidence: 0.9192497

00:20:06.246 --> 00:20:09.820 on whether you have spikings or not,

NOTE Confidence: 0.9192497

00:20:09.820 --> 00:20:14.580 the possibility are either limited or not.

NOTE Confidence: 0.9192497

00:20:14.580 --> 00:20:16.460 This is another excellent yes,

NOTE Confidence: 0.94115114

00:20:16.460 --> 00:20:17.956 sorry so it's alright.

NOTE Confidence: 0.94115114

00:20:17.956 --> 00:20:19.455 Yeah, it's actually quite

NOTE Confidence: 0.94115114

00:20:19.455 --> 00:20:21.330 interesting to see this result,

NOTE Confidence: 0.94115114

00:20:21.330 --> 00:20:23.583 you know, seems to the simple

NOTE Confidence: 0.94115114

00:20:23.583 --> 00:20:26.210 normalization is at best in this case.

NOTE Confidence: 0.94115114

00:20:26.210 --> 00:20:28.590 If simply just judging from how tight

NOTE Confidence: 0.94115114

00:20:28.590 --> 00:20:31.743 the same cell population is and how far

NOTE Confidence: 0.94115114

00:20:31.743 --> 00:20:34.080 away two distinct populations should be,

NOTE Confidence: 0.94115114

00:20:34.080 --> 00:20:37.617 but I would assume this is done by maybe

NOTE Confidence: 0.94115114

00:20:37.617 --> 00:20:40.079 something like a sort of a Euclidean

NOTE Confidence: 0.94115114

00:20:40.080 --> 00:20:41.636 distance based measurements, because

NOTE Confidence: 0.94115114

00:20:41.636 --> 00:20:44.300 if you simply normalize by library size.

NOTE Confidence: 0.94115114

00:20:44.300 --> 00:20:45.780 If you use a correlation,

NOTE Confidence: 0.94115114

00:20:45.780 --> 00:20:47.250 that wouldn't change anything, right?

NOTE Confidence: 0.94115114

00:20:47.250 --> 00:20:48.730 Because the correlation between the

NOTE Confidence: 0.94115114

00:20:48.730 --> 00:20:50.500 genes will still remain the same,

NOTE Confidence: 0.94115114

00:20:50.500 --> 00:20:52.306 or between cells will still remain

NOTE Confidence: 0.94115114

00:20:52.306 --> 00:20:53.510 the same regardless whether

NOTE Confidence: 0.94115114

00:20:53.560 --> 00:20:54.920 mobilized by library or not.

NOTE Confidence: 0.96328956

00:20:55.670 --> 00:20:59.070 Yeah, then here. So I didn't see it.

NOTE Confidence: 0.96328956

00:20:59.070 --> 00:21:00.360 So this anticipation.

NOTE Confidence: 0.96328956

00:21:00.360 --> 00:21:02.080 So here the visualization

NOTE Confidence: 0.96328956

00:21:02.080 --> 00:21:04.623 of this cluster is based on

NOTE Confidence: 0.96328956

00:21:04.623 --> 00:21:06.203 this approach of dimensional

NOTE Confidence: 0.96328956

00:21:06.203 --> 00:21:08.419 reduction that is called Disney.

NOTE Confidence: 0.96328956

00:21:08.420 --> 00:21:10.550 So that could also affect.

NOTE Confidence: 0.96328956

00:21:10.550 --> 00:21:13.343 So these differences that you see here

NOTE Confidence: 0.96328956

00:21:13.343 --> 00:21:16.497 change also if you change the dimension.

NOTE Confidence: 0.96328956

00:21:16.500 --> 00:21:18.620 If you change the dimensionality

NOTE Confidence: 0.96328956

00:21:18.620 --> 00:21:19.468 reduction method.

NOTE Confidence: 0.97818357

00:21:22.130 --> 00:21:24.711 Used it to plot the results, but I agree,

NOTE Confidence: 0.97818357

00:21:24.711 --> 00:21:26.433 so here is the simple normalization.

NOTE Confidence: 0.97818357

00:21:26.440 --> 00:21:28.778 Seems to be one of the most

NOTE Confidence: 0.97818357

00:21:28.778 --> 00:21:30.684 effective in terms of separating

NOTE Confidence: 0.97818357

00:21:30.684 --> 00:21:32.759 the two clusters at least.

NOTE Confidence: 0.97818357

00:21:32.760 --> 00:21:34.962 This is another example with another

NOTE Confidence: 0.97818357

00:21:34.962 --> 00:21:37.850 data set of mouth longevity, real cells.

NOTE Confidence: 0.97818357

00:21:37.850 --> 00:21:40.195 So here you have more cluster of

NOTE Confidence: 0.97818357

00:21:40.195 --> 00:21:42.409 cells corresponding to different.

NOTE Confidence: 0.91594625

00:21:45.170 --> 00:21:46.625 Differentiation points so

NOTE Confidence: 0.91594625

00:21:46.625 --> 00:21:49.050 different stages of the embryo,

NOTE Confidence: 0.91594625

00:21:49.050 --> 00:21:51.829 a 14 and 1618 and then and

NOTE Confidence: 0.91594625

00:21:51.829 --> 00:21:54.914 then the green are the adults

NOTE Confidence: 0.91594625

00:21:54.914 --> 00:21:57.298 that cells epithelial cells,

NOTE Confidence: 0.91594625

00:21:57.300 --> 00:22:00.900 so also here they called the basic A

NOTE Confidence: 0.91594625

00:22:00.900 --> 00:22:05.009 take home message is that there is no

NOTE Confidence: 0.91594625

00:22:05.009 --> 00:22:07.969 consensus on which method is best,

NOTE Confidence: 0.91594625

00:22:07.970 --> 00:22:09.910 and different methods can

NOTE Confidence: 0.91594625

00:22:09.910 --> 00:22:11.850 lead to different results.

NOTE Confidence: 0.98387814

00:22:15.850 --> 00:22:19.009 So that in each, depending on the data set,

NOTE Confidence: 0.98387814

00:22:19.010 --> 00:22:21.467 the methods that perform the best changes.

NOTE Confidence: 0.8880952

00:22:24.530 --> 00:22:26.842 And what you don't have here is a

NOTE Confidence: 0.8880952

00:22:26.842 --> 00:22:28.999 methods that are taken from the back,

NOTE Confidence: 0.8880952

00:22:29.000 --> 00:22:30.770 so they were not considered in

NOTE Confidence: 0.8880952

00:22:30.770 --> 00:22:32.280 this in this comparison here.

NOTE Confidence: 0.90325

00:22:35.670 --> 00:22:38.646 OK, so this was for the preprocessing steps.

NOTE Confidence: 0.90325

00:22:38.650 --> 00:22:40.888 Then the post processing the steps.

NOTE Confidence: 0.90325

00:22:40.890 --> 00:22:43.530 At this the the after we have the

NOTE Confidence: 0.90325

00:22:43.530 --> 00:22:45.369 normalized our normalized data.

NOTE Confidence: 0.90325

00:22:45.370 --> 00:22:47.645 We can start the second part of

NOTE Confidence: 0.90325

00:22:47.645 --> 00:22:49.960 the analysis and the main steps

NOTE Confidence: 0.90325

00:22:49.960 --> 00:22:52.075 here are the dimensional reduction.

NOTE Confidence: 0.90325

00:22:52.080 --> 00:22:54.800 So we will see that these data since

NOTE Confidence: 0.90325

00:22:54.800 --> 00:22:57.464 they have a lot of rows and columns

NOTE Confidence: 0.90325

00:22:57.464 --> 00:23:00.289 there they have a high dimensionality.

NOTE Confidence: 0.90325

00:23:00.290 --> 00:23:02.460 This is problematic for the

NOTE Confidence: 0.90325

00:23:02.460 --> 00:23:04.196 code for the interpretation.

NOTE Confidence: 0.90325

00:23:04.200 --> 00:23:07.336 For the visualization and also for the,

NOTE Confidence: 0.90325

00:23:07.340 --> 00:23:09.990 uh, uh, running a computational

NOTE Confidence: 0.90325

00:23:09.990 --> 00:23:12.640 procedures because it can take

NOTE Confidence: 0.90325

00:23:12.727 --> 00:23:14.946 it can take a lot of time,

NOTE Confidence: 0.90325

00:23:14.950 --> 00:23:17.800 so the reduction to a medium

NOTE Confidence: 0.90325

00:23:17.800 --> 00:23:20.242 dimensional space is usually performed

NOTE Confidence: 0.90325

00:23:20.242 --> 00:23:23.392 or performed on the genes so that

NOTE Confidence: 0.90325

00:23:23.392 --> 00:23:25.983 instead of having a 10,000 genes

NOTE Confidence: 0.90325

00:23:25.983 --> 00:23:29.164 that we have at this point we have

NOTE Confidence: 0.90325

00:23:29.164 --> 00:23:31.873 1030 dimensions and we will see that

NOTE Confidence: 0.90325

00:23:31.873 --> 00:23:34.278 these dimensions can represent.

NOTE Confidence: 0.90325

00:23:34.280 --> 00:23:36.484 Combination of different genes.

NOTE Confidence: 0.90325

00:23:36.484 --> 00:23:41.460 But the key point is that you reduce the

NOTE Confidence: 0.90325

00:23:41.460 --> 00:23:45.114 number of dimensions from 10,000 to 10.

NOTE Confidence: 0.90325

00:23:45.120 --> 00:23:47.955 So this is the first important step,

NOTE Confidence: 0.90325

00:23:47.960 --> 00:23:51.614 so I will speak about this quite in detail.

NOTE Confidence: 0.90325

00:23:51.620 --> 00:23:54.476 So the problem is this curse of

NOTE Confidence: 0.90325

00:23:54.476 --> 00:23:57.005 dimensionality so that we have 2010

NOTE Confidence: 0.90325

00:23:57.005 --> 00:23:59.453 to 20,000 genes as features and

NOTE Confidence: 0.90325

00:23:59.453 --> 00:24:01.867 depending on our experiment we have

NOTE Confidence: 0.90325

00:24:01.867 --> 00:24:04.610 10,000 up to 1,000,000 of cells that

NOTE Confidence: 0.90325

00:24:04.610 --> 00:24:07.445 we want to analyze and to consider.

NOTE Confidence: 0.90325

00:24:07.450 --> 00:24:11.095 So we need to reduce the number of features,

NOTE Confidence: 0.90325

00:24:11.100 --> 00:24:13.536 in particular the number of genes,

NOTE Confidence: 0.90325

00:24:13.540 --> 00:24:14.326 the rational.

NOTE Confidence: 0.90325

00:24:14.326 --> 00:24:17.077 Is that there are two two points

NOTE Confidence: 0.90325

00:24:17.077 --> 00:24:18.580 for the rational.

NOTE Confidence: 0.90325

00:24:18.580 --> 00:24:21.464 The first is that not all the

NOTE Confidence: 0.90325

00:24:21.464 --> 00:24:22.700 genes are important.

NOTE Confidence: 0.90325

00:24:22.700 --> 00:24:25.484 If our aim is to classify cells according

NOTE Confidence: 0.90325

00:24:25.484 --> 00:24:28.059 to their differences in expression,

NOTE Confidence: 0.90325

00:24:28.060 --> 00:24:30.120 not all genes are important.

NOTE Confidence: 0.90325

00:24:30.120 --> 00:24:33.280 So for example for sure genes that are

NOTE Confidence: 0.90325

00:24:33.280 --> 00:24:35.468 never expressed are not important,

NOTE Confidence: 0.90325

00:24:35.470 --> 00:24:38.044 but also housekeeping genes that are

NOTE Confidence: 0.90325

00:24:38.044 --> 00:24:40.653 always expressed at the same level

NOTE Confidence: 0.90325

00:24:40.653 --> 00:24:43.101 are also not important in separating

NOTE Confidence: 0.90325

00:24:43.101 --> 00:24:45.420 the cells and we select these.

NOTE Confidence: 0.90325

00:24:45.420 --> 00:24:48.150 Jeans for this point that through

NOTE Confidence: 0.90325

00:24:48.150 --> 00:24:49.515 feature gene selection,

NOTE Confidence: 0.90325

00:24:49.520 --> 00:24:52.523 then the second point is that many

NOTE Confidence: 0.90325

00:24:52.523 --> 00:24:54.980 genes are correlated in expression,

NOTE Confidence: 0.90325

00:24:54.980 --> 00:24:57.272 so it's redundant to have two

NOTE Confidence: 0.90325

00:24:57.272 --> 00:24:59.461 genes that are highly correlated

NOTE Confidence: 0.90325

00:24:59.461 --> 00:25:01.797 as two separate information.

NOTE Confidence: 0.90325

00:25:01.800 --> 00:25:06.770 We can combine them into one dimension.

NOTE Confidence: 0.90325

00:25:06.770 --> 00:25:09.745 And these correlation is taking care during

NOTE Confidence: 0.90325

00:25:09.745 --> 00:25:11.730 the dimensionality reduction approaches.

NOTE Confidence: 0.90325

00:25:11.730 --> 00:25:13.534 So for this election,

NOTE Confidence: 0.90325

00:25:13.534 --> 00:25:15.338 for the first step,

NOTE Confidence: 0.90325

00:25:15.340 --> 00:25:19.516 selection of genes that are important.

NOTE Confidence: 0.90325

00:25:19.520 --> 00:25:21.960 The aim is to select the genes that

NOTE Confidence: 0.90325

00:25:21.960 --> 00:25:23.638 contain useful information about the

NOTE Confidence: 0.90325

00:25:23.638 --> 00:25:25.983 biology of the system and so they

NOTE Confidence: 0.90325

00:25:26.051 --> 00:25:28.865 are the genes that have difference in

NOTE Confidence: 0.90325

00:25:28.865 --> 00:25:30.434 expression between different cells

NOTE Confidence: 0.90325

00:25:30.434 --> 00:25:32.912 and we want to remove genes that

NOTE Confidence: 0.90325

00:25:32.912 --> 00:25:34.811 contain either only noise because

NOTE Confidence: 0.90325

00:25:34.811 --> 00:25:37.055 they have low expression level and

NOTE Confidence: 0.90325

00:25:37.055 --> 00:25:39.697 so all the variation is noise or the

NOTE Confidence: 0.90325

00:25:39.697 --> 00:25:42.400 jeans that do not have variation among genes.

NOTE Confidence: 0.90325

00:25:42.400 --> 00:25:44.376 So the housekeeping genes.

NOTE Confidence: 0.90325

00:25:44.376 --> 00:25:47.340 And the simplest approach to do

NOTE Confidence: 0.90325

00:25:47.434 --> 00:25:50.059 that is to calculate for each gene

NOTE Confidence: 0.90325

00:25:50.059 --> 00:25:53.644 a sort of measure that is a variance

NOTE Confidence: 0.90325

00:25:53.644 --> 00:25:55.540 corrected for the mean.

NOTE Confidence: 0.90325

00:25:55.540 --> 00:25:58.186 So we have seen something similar.

NOTE Confidence: 0.90325

00:25:58.190 --> 00:26:02.168 Also during the lesson for the bulk RNA seek,

NOTE Confidence: 0.90325

00:26:02.170 --> 00:26:05.257 because the approach is not so different.

NOTE Confidence: 0.90325

00:26:05.260 --> 00:26:08.796 So you rank GS, you build a model.

NOTE Confidence: 0.9336727

00:26:08.800 --> 00:26:11.446 Each dot. Here is a gene,

NOTE Confidence: 0.9336727

00:26:11.450 --> 00:26:15.058 and you expect the variance of the gene.

NOTE Confidence: 0.9336727

00:26:15.060 --> 00:26:17.100 To be proportional to the

NOTE Confidence: 0.9336727

00:26:17.100 --> 00:26:19.140 average expression of the gene,

NOTE Confidence: 0.9336727

00:26:19.140 --> 00:26:21.198 meaning that the more the gene

NOTE Confidence: 0.9336727

00:26:21.198 --> 00:26:23.625 is expressed that the more random

NOTE Confidence: 0.9336727

00:26:23.625 --> 00:26:26.080 fluctuation fluctuation you also expect.

NOTE Confidence: 0.9336727

00:26:26.080 --> 00:26:28.856 So you build a sort of model that

NOTE Confidence: 0.9336727

00:26:28.856 --> 00:26:30.578 captures a random variations

NOTE Confidence: 0.9336727

00:26:30.578 --> 00:26:33.416 that you expect in your genes,

NOTE Confidence: 0.9336727

00:26:33.420 --> 00:26:36.468 and then you see which genes are outliers

NOTE Confidence: 0.9336727

00:26:36.468 --> 00:26:39.607 so they show more variance than the

NOTE Confidence: 0.9336727

00:26:39.607 --> 00:26:42.725 baseline variance that is based on the

NOTE Confidence: 0.9336727

00:26:42.725 --> 00:26:45.322 noise or or on the random variation.

NOTE Confidence: 0.9336727

00:26:45.330 --> 00:26:47.598 In expression and those genes that

NOTE Confidence: 0.9336727

00:26:47.598 --> 00:26:50.310 are highly variants are the ones that

NOTE Confidence: 0.9336727

00:26:50.310 --> 00:26:52.185 you select for further analysis,

NOTE Confidence: 0.9336727

00:26:52.190 --> 00:26:54.906 because there are those genes where you

NOTE Confidence: 0.9336727

00:26:54.906 --> 00:26:57.168 don't have only technical variations

NOTE Confidence: 0.9336727

00:26:57.168 --> 00:27:00.240 or but you have biological variation.

NOTE Confidence: 0.9336727

00:27:00.240 --> 00:27:02.480 The questionable assumption here is

NOTE Confidence: 0.9336727

00:27:02.480 --> 00:27:04.720 that the biological variability is

NOTE Confidence: 0.9336727

00:27:04.793 --> 00:27:07.228 higher than the technical variability,

NOTE Confidence: 0.9336727

00:27:07.230 --> 00:27:09.804 because the assumption here is that

NOTE Confidence: 0.9336727

00:27:09.804 --> 00:27:12.531 all these outlier genes that show

NOTE Confidence: 0.9336727

00:27:12.531 --> 00:27:15.327 higher variance than the average are

NOTE Confidence: 0.9336727

00:27:15.327 --> 00:27:17.793 important because this higher variance

NOTE Confidence: 0.9336727

00:27:17.793 --> 00:27:20.228 is biological variance and obviously

NOTE Confidence: 0.9336727

00:27:20.228 --> 00:27:23.540 also here as in some balcony approach,

NOTE Confidence: 0.9336727

00:27:23.540 --> 00:27:26.168 you could have some methods that

NOTE Confidence: 0.9336727

00:27:26.168 --> 00:27:28.670 penalize jeans having high variance,

NOTE Confidence: 0.9336727

00:27:28.670 --> 00:27:29.768 but lo Mein.

NOTE Confidence: 0.9336727

00:27:29.768 --> 00:27:32.330 Because you don't trust them so much,

NOTE Confidence: 0.9336727

00:27:32.330 --> 00:27:34.496 but the assumption is that you

NOTE Confidence: 0.9336727

00:27:34.496 --> 00:27:36.825 calculate a measure of variance and

NOTE Confidence: 0.9336727

00:27:36.825 --> 00:27:39.273 you consider the top variant genes.

NOTE Confidence: 0.9336727

00:27:39.280 --> 00:27:41.296 And you remove the others from the analysis.

NOTE Confidence: 0.985337

00:27:44.890 --> 00:27:46.786 Then there is the dimensionality reduction,

NOTE Confidence: 0.985337

00:27:46.790 --> 00:27:49.310 so this is a family of approaches

NOTE Confidence: 0.985337

00:27:49.310 --> 00:27:51.439 that are using complex data.

NOTE Confidence: 0.985337

00:27:51.440 --> 00:27:54.590 To reduce the number of dimensions of

NOTE Confidence: 0.985337

00:27:54.590 --> 00:27:58.429 the data so this has a double purpose,

NOTE Confidence: 0.985337

00:27:58.430 --> 00:28:01.662 as I say that to help the analysis

NOTE Confidence: 0.985337

00:28:01.662 --> 00:28:03.531 downstream analysis because the

NOTE Confidence: 0.985337

00:28:03.531 --> 00:28:05.996 reducing the dimension speed the

NOTE Confidence: 0.985337

00:28:05.996 --> 00:28:08.553 calculation times and also to

NOTE Confidence: 0.985337

00:28:08.553 --> 00:28:10.545 help the visually visualization.

NOTE Confidence: 0.985337

00:28:10.550 --> 00:28:12.780 Especially when you report single

NOTE Confidence: 0.985337

00:28:12.780 --> 00:28:16.533 cell data they need is to show data

NOTE Confidence: 0.985337

00:28:16.533 --> 00:28:18.923 in simple and interpretable output.

NOTE Confidence: 0.985337

00:28:18.930 --> 00:28:22.290 So usually this is a 2D plot.

NOTE Confidence: 0.985337

00:28:22.290 --> 00:28:24.330 And so they mentioned I did.

NOTE Confidence: 0.985337

00:28:24.330 --> 00:28:27.221 Action are also used in order to

NOTE Confidence: 0.985337

00:28:27.221 --> 00:28:28.873 compress high dimensional information

NOTE Confidence: 0.985337

00:28:28.873 --> 00:28:31.888 so that it can be presented in a 2D

NOTE Confidence: 0.985337

00:28:31.961 --> 00:28:34.600 plot and the two are different needs.

NOTE Confidence: 0.985337

00:28:34.600 --> 00:28:36.156 There are multiple methodologies.

NOTE Confidence: 0.985337

00:28:36.156 --> 00:28:38.101 Each one has different advantages

NOTE Confidence: 0.985337

00:28:38.101 --> 00:28:39.210 and limitations,

NOTE Confidence: 0.985337

00:28:39.210 --> 00:28:41.870 so the classic example of a dimensional

NOTE Confidence: 0.985337

00:28:41.870 --> 00:28:44.912 reduction that we always have in mind

NOTE Confidence: 0.985337

00:28:44.912 --> 00:28:46.744 and possibly historically speaking,

NOTE Confidence: 0.985337

00:28:46.750 --> 00:28:50.730 is one of the oldest is when you have a

NOTE Confidence: 0.985337

00:28:50.832 --> 00:28:54.702 problem to draw a 2D map of the Earth.

NOTE Confidence: 0.985337

00:28:54.710 --> 00:28:57.318 So Earth is 3D and you want that

NOTE Confidence: 0.985337

00:28:57.318 --> 00:29:00.560 2D map that keeps most of the

NOTE Confidence: 0.985337

00:29:00.560 --> 00:29:03.065 reliable information on the geography

NOTE Confidence: 0.985337

00:29:03.156 --> 00:29:05.326 on the geography of Earth.

NOTE Confidence: 0.985337

00:29:05.330 --> 00:29:07.886 So where the continents are placed,

NOTE Confidence: 0.985337

00:29:07.890 --> 00:29:10.865 their shapes, their areas and so on,

NOTE Confidence: 0.985337

00:29:10.870 --> 00:29:13.000 so there are different approaches,

NOTE Confidence: 0.985337

00:29:13.000 --> 00:29:15.215 many different approaches to convert

NOTE Confidence: 0.985337

00:29:15.215 --> 00:29:18.540 the 3D map of Earth into 2D maps,

NOTE Confidence: 0.985337

00:29:18.540 --> 00:29:20.766 and for example here you see one

NOTE Confidence: 0.985337

00:29:20.766 --> 00:29:23.502 of the most famous projections that

NOTE Confidence: 0.985337

00:29:23.502 --> 00:29:25.774 is called Mercator projections.

NOTE Confidence: 0.985337

00:29:25.780 --> 00:29:29.188 So this was developed in the 16th century,

NOTE Confidence: 0.985337

00:29:29.190 --> 00:29:32.438 and it's the one used by sailors because

NOTE Confidence: 0.985337

00:29:32.438 --> 00:29:35.497 it keeps the directions and shapes.

NOTE Confidence: 0.985337

00:29:35.500 --> 00:29:37.677 So it's a good map to know

NOTE Confidence: 0.985337

00:29:37.677 --> 00:29:39.710 what is North East or West.

NOTE Confidence: 0.985337

00:29:39.710 --> 00:29:42.815 And the problem is that there is a high

NOTE Confidence: 0.985337

00:29:42.815 --> 00:29:45.215 distortion in this map of areas so that

NOTE Confidence: 0.985337

00:29:45.215 --> 00:29:48.140 the the more you are far from the equator,

NOTE Confidence: 0.985337

00:29:48.140 --> 00:29:49.755 the more areas that seems

NOTE Confidence: 0.985337

00:29:49.755 --> 00:29:51.047 larger than they are.

NOTE Confidence: 0.985337

00:29:51.050 --> 00:29:53.192 And so for example here it seems

NOTE Confidence: 0.985337

00:29:53.192 --> 00:29:54.854 that Greenland is bigger than

NOTE Confidence: 0.985337

00:29:54.854 --> 00:29:56.559 the whole of South America.

NOTE Confidence: 0.985337

00:29:56.560 --> 00:29:59.144 That is not so. This is the distortion.

NOTE Confidence: 0.985337

00:29:59.150 --> 00:30:02.559 So other projections such as these two.

NOTE Confidence: 0.985337

00:30:02.560 --> 00:30:04.580 And they are projections where

NOTE Confidence: 0.985337

00:30:04.580 --> 00:30:06.600 the the area is preserved,

NOTE Confidence: 0.985337

00:30:06.600 --> 00:30:09.302 and so that this area corresponds really

NOTE Confidence: 0.985337

00:30:09.302 --> 00:30:11.848 to the smallest areas with respect.

NOTE Confidence: 0.985337

00:30:11.850 --> 00:30:13.870 For example to South America.

NOTE Confidence: 0.985337

00:30:13.870 --> 00:30:16.411 But these kind of maps do not

NOTE Confidence: 0.985337

00:30:16.411 --> 00:30:18.320 preserve shapes and direction,

NOTE Confidence: 0.985337

00:30:18.320 --> 00:30:21.480 so that the common point is is any

NOTE Confidence: 0.985337

00:30:21.480 --> 00:30:23.565 projection will be will distort

NOTE Confidence: 0.985337

00:30:23.565 --> 00:30:25.590 the some of the features.

NOTE Confidence: 0.985337

00:30:25.590 --> 00:30:27.625 So reduction of dimensionality is

NOTE Confidence: 0.985337

00:30:27.625 --> 00:30:30.619 always an approximation and it brings it

NOTE Confidence: 0.985337

00:30:30.619 --> 00:30:32.844 brings some distortions and deviations.

NOTE Confidence: 0.985337

00:30:32.850 --> 00:30:35.447 And as a for the Earth map,

NOTE Confidence: 0.985337

00:30:35.450 --> 00:30:37.220 we have different approaches also

NOTE Confidence: 0.985337

00:30:37.220 --> 00:30:39.831 for our single cell data we see

NOTE Confidence: 0.985337

00:30:39.831 --> 00:30:41.379 there are different techniques.

NOTE Confidence: 0.98599946

00:30:44.150 --> 00:30:45.440 Is this clear?

NOTE Confidence: 0.9871331

00:30:48.520 --> 00:30:50.448 Anyway, it's a very good analogy,

NOTE Confidence: 0.9871331

00:30:50.448 --> 00:30:53.750 so this is awesome feedback.

NOTE Confidence: 0.9871331

00:30:53.750 --> 00:30:56.547 So the first one that we will see with a real

NOTE Confidence: 0.9871331

00:30:56.547 --> 00:30:58.707 example is principal component analysis.

NOTE Confidence: 0.9871331

00:30:58.710 --> 00:31:00.942 So in our case we are studying cells

NOTE Confidence: 0.9871331

00:31:00.942 --> 00:31:03.050 based on the expression of genes.

NOTE Confidence: 0.9871331

00:31:03.050 --> 00:31:05.045 So in that in our simple example

NOTE Confidence: 0.9871331

00:31:05.045 --> 00:31:07.259 we will have six cells and since

NOTE Confidence: 0.9871331

00:31:07.259 --> 00:31:09.197 they are simple cells they have

NOTE Confidence: 0.9871331

00:31:09.264 --> 00:31:11.420 only they express or women age 25,

NOTE Confidence: 0.9871331

00:31:11.420 --> 00:31:12.482 only four genes.

NOTE Confidence: 0.9871331

00:31:12.482 --> 00:31:15.407 And So what you see here is the

NOTE Confidence: 0.9871331

00:31:15.407 --> 00:31:17.217 expression level of each gene

NOTE Confidence: 0.9871331

00:31:17.217 --> 00:31:20.119 from A to D in this six sets.

NOTE Confidence: 0.9871331

00:31:20.120 --> 00:31:23.224 So now we can use the expression levels,

NOTE Confidence: 0.9871331

00:31:23.230 --> 00:31:26.060 uh, so in in as a way to map cells

NOTE Confidence: 0.9871331

00:31:26.148 --> 00:31:28.746 and the expression level of each

NOTE Confidence: 0.9871331

00:31:28.746 --> 00:31:31.400 gene is a different dimension.

NOTE Confidence: 0.9871331

00:31:31.400 --> 00:31:34.576 So in this case we have a four

NOTE Confidence: 0.9871331

00:31:34.576 --> 00:31:36.200 dimensional space that obviously

NOTE Confidence: 0.9871331

00:31:36.200 --> 00:31:39.176 we cannot plot a on a 2D plot.

NOTE Confidence: 0.9871331

00:31:39.180 --> 00:31:41.392 So one simple so we could plot

NOTE Confidence: 0.9871331

00:31:41.392 --> 00:31:44.268 on a two diploid cells based on

NOTE Confidence: 0.9871331

00:31:44.268 --> 00:31:46.568 the expression of two genes,

NOTE Confidence: 0.9871331

00:31:46.570 --> 00:31:50.226 and so we can take jeanae engine be.

NOTE Confidence: 0.9871331

00:31:50.230 --> 00:31:52.925 And build these sort of map of

NOTE Confidence: 0.9871331

00:31:52.925 --> 00:31:55.343 these cells based on the expression

NOTE Confidence: 0.9871331

00:31:55.343 --> 00:31:58.476 level of gene eight that is our X

NOTE Confidence: 0.9871331

00:31:58.476 --> 00:32:01.149 axis and gene B that is our Y axis.

NOTE Confidence: 0.9871331

00:32:01.150 --> 00:32:04.062 And here you see where cells are

NOTE Confidence: 0.9871331

00:32:04.062 --> 00:32:06.236 located according to the expression

NOTE Confidence: 0.9871331

00:32:06.236 --> 00:32:08.768 of eight of these two genes.

NOTE Confidence: 0.9871331

00:32:08.770 --> 00:32:11.398 So the expression of each gene

NOTE Confidence: 0.9871331

00:32:11.398 --> 00:32:12.712 is a dimension.

NOTE Confidence: 0.9871331

00:32:12.720 --> 00:32:15.360 So now with this weekend plot,

NOTE Confidence: 0.9871331

00:32:15.360 --> 00:32:17.988 two genes in a 2D map,

NOTE Confidence: 0.9871331

00:32:17.990 --> 00:32:19.742 and so for performing

NOTE Confidence: 0.9871331

00:32:19.742 --> 00:32:21.056 principal component analysis,

NOTE Confidence: 0.9871331

00:32:21.060 --> 00:32:23.811 what is usually done at the beginning

NOTE Confidence: 0.9871331

00:32:23.811 --> 00:32:26.329 is to center the measurement,

NOTE Confidence: 0.9871331

00:32:26.330 --> 00:32:28.640 meaning that these genes here

NOTE Confidence: 0.9871331

00:32:28.640 --> 00:32:30.950 they have an average expression

NOTE Confidence: 0.9871331

00:32:31.026 --> 00:32:32.916 of seven for these jeans.

NOTE Confidence: 0.9871331

00:32:32.920 --> 00:32:35.920 This is the average of jinei

NOTE Confidence: 0.9871331

00:32:35.920 --> 00:32:37.920 across the six cells.

NOTE Confidence: 0.9871331

00:32:37.920 --> 00:32:39.078 And so on.

NOTE Confidence: 0.9871331

00:32:39.078 --> 00:32:41.780 So these gene B has an average

NOTE Confidence: 0.9871331

00:32:41.877 --> 00:32:44.413 of 4.5 GC of six and so on.

NOTE Confidence: 0.9871331

00:32:44.420 --> 00:32:46.694 So centering the data means that

NOTE Confidence: 0.9871331

00:32:46.694 --> 00:32:48.636 you calculated the mean expression

NOTE Confidence: 0.9871331

00:32:48.636 --> 00:32:51.142 of the gene across all these cells

NOTE Confidence: 0.9871331

00:32:51.142 --> 00:32:53.588 and you subtract the mean from all

NOTE Confidence: 0.9871331

00:32:53.588 --> 00:32:57.359 the values of the gene so that you.

NOTE Confidence: 0.9871331

00:32:57.360 --> 00:32:59.436 Switch from this matrix that is

NOTE Confidence: 0.9871331

00:32:59.436 --> 00:33:01.623 not centered to this matrix that

NOTE Confidence: 0.9871331

00:33:01.623 --> 00:33:03.087 is centered around 0.

NOTE Confidence: 0.9871331

00:33:03.090 --> 00:33:05.238 So I simply from the top

NOTE Confidence: 0.9871331

00:33:05.238 --> 00:33:06.670 row I subtracted seven,

NOTE Confidence: 0.9871331

00:33:06.670 --> 00:33:09.883 so 11 -- 7 is 4 and so on.

NOTE Confidence: 0.9871331

00:33:09.890 --> 00:33:12.130 From the second I subtract 4.5 and

NOTE Confidence: 0.9871331

00:33:12.130 --> 00:33:15.542 so on so that you see in the centered

NOTE Confidence: 0.9871331

00:33:15.542 --> 00:33:18.228 values are also negative and the common

NOTE Confidence: 0.9871331

00:33:18.228 --> 00:33:22.979 point is that the mean for each gene is 0.

NOTE Confidence: 0.9871331

00:33:22.980 --> 00:33:24.520 So usually before performing

NOTE Confidence: 0.9871331

00:33:24.520 --> 00:33:25.675 like I mentioned,

NOTE Confidence: 0.9871331

00:33:25.680 --> 00:33:26.832 I did action.

NOTE Confidence: 0.9871331

00:33:26.832 --> 00:33:29.136 This centering is is performed and

NOTE Confidence: 0.9871331

00:33:29.136 --> 00:33:32.066 it's also helpful in the visualization.

NOTE Confidence: 0.9871331

00:33:32.070 --> 00:33:33.096 So before centering,

NOTE Confidence: 0.9871331

00:33:33.096 --> 00:33:35.490 the cells were were looking like this.

NOTE Confidence: 0.9871331

00:33:35.490 --> 00:33:36.516 After the centering,

NOTE Confidence: 0.9871331

00:33:36.516 --> 00:33:38.226 these are the new coordinates,

NOTE Confidence: 0.9871331

00:33:38.230 --> 00:33:40.582 so nothing changed that it's only the

NOTE Confidence: 0.9871331

00:33:40.582 --> 00:33:43.210 origin of the axis and the position of

NOTE Confidence: 0.9871331

00:33:43.210 --> 00:33:45.750 the zero that are that are different.

NOTE Confidence: 0.9871331

00:33:45.750 --> 00:33:48.137 But if you look at the cells,

NOTE Confidence: 0.9871331

00:33:48.140 --> 00:33:51.002 the points are exactly in the

NOTE Confidence: 0.9871331

00:33:51.002 --> 00:33:52.910 same position as before.

NOTE Confidence: 0.9871331

00:33:52.910 --> 00:33:55.166 No one question that we can

NOTE Confidence: 0.9871331

00:33:55.166 --> 00:33:56.670 ask here is weather.

NOTE Confidence: 0.9871331

00:33:56.670 --> 00:33:59.694 So what you see here is that the

NOTE Confidence: 0.9871331

00:33:59.694 --> 00:34:01.558 difference here of the cells.

NOTE Confidence: 0.9411312

00:34:01.560 --> 00:34:04.032 We can capture the difference of

NOTE Confidence: 0.9411312

00:34:04.032 --> 00:34:06.122 these cells because they differ

NOTE Confidence: 0.9411312

00:34:06.122 --> 00:34:08.698 in the expression of A&B and one

NOTE Confidence: 0.9411312

00:34:08.698 --> 00:34:11.174 question we can ask is whether it

NOTE Confidence: 0.9411312

00:34:11.174 --> 00:34:13.965 is better to weather GTA or gene bit

NOTE Confidence: 0.9411312

00:34:13.965 --> 00:34:16.215 is better in separating these cells.

NOTE Confidence: 0.9411312

00:34:16.220 --> 00:34:18.180 So this corresponds to asking

NOTE Confidence: 0.9411312

00:34:18.180 --> 00:34:20.566 how much of the variability of

NOTE Confidence: 0.9411312

00:34:20.566 --> 00:34:22.690 the data is associated with this.

NOTE Confidence: 0.9411312

00:34:22.690 --> 00:34:25.150 Progression of GD or with

NOTE Confidence: 0.9411312

00:34:25.150 --> 00:34:27.610 the expression of Gene D.

NOTE Confidence: 0.9411312

00:34:27.610 --> 00:34:30.080 And so the question is,

NOTE Confidence: 0.9411312

00:34:30.080 --> 00:34:33.139 what is the variation of these six

NOTE Confidence: 0.9411312

00:34:33.139 --> 00:34:36.535 points that is associated with gene

NOTE Confidence: 0.9411312

00:34:36.535 --> 00:34:39.187 expression engine B expression?

NOTE Confidence: 0.9411312

00:34:39.190 --> 00:34:42.286 So there is a simple way to calculate

NOTE Confidence: 0.9411312

00:34:42.286 --> 00:34:44.140 the variation associated that

NOTE Confidence: 0.9411312

00:34:44.140 --> 00:34:47.626 corresponds to the formula of the variance.

NOTE Confidence: 0.9411312

00:34:47.630 --> 00:34:51.158 So this is an example to calculate the

NOTE Confidence: 0.9411312

00:34:51.158 --> 00:34:54.286 variation that is associated with gene 8.

NOTE Confidence: 0.9411312

00:34:54.290 --> 00:34:57.834 So here we're considering the X axis here,

NOTE Confidence: 0.9411312

00:34:57.840 --> 00:35:01.136 so I can draw a projection from each

NOTE Confidence: 0.9411312

00:35:01.136 --> 00:35:04.675 cell to these axis and calculate the

NOTE Confidence: 0.9411312

00:35:04.675 --> 00:35:08.329 distance from the origin to each cell.

NOTE Confidence: 0.9411312

00:35:08.330 --> 00:35:09.060 And basically,

NOTE Confidence: 0.9411312

00:35:09.060 --> 00:35:11.615 since here where we centered the data,

NOTE Confidence: 0.9411312

00:35:11.620 --> 00:35:13.080 the distance basically correspond

NOTE Confidence: 0.9411312

00:35:13.080 --> 00:35:14.540 to these expression level.

NOTE Confidence: 0.9411312

00:35:14.540 --> 00:35:17.816 So cell one has a distance of four cell,

NOTE Confidence: 0.9411312

00:35:17.820 --> 00:35:21.980 two or distance of five and so on.

NOTE Confidence: 0.9411312

00:35:21.980 --> 00:35:24.514 Now if we want to measure the

NOTE Confidence: 0.9411312

00:35:24.514 --> 00:35:26.959 variation with the variance formula,

NOTE Confidence: 0.9411312

00:35:26.960 --> 00:35:29.312 the variance formula is to take

NOTE Confidence: 0.9411312

00:35:29.312 --> 00:35:32.272 the square of each of the distance

NOTE Confidence: 0.9411312

00:35:32.272 --> 00:35:34.020 of these six distances.

NOTE Confidence: 0.9411312

00:35:34.020 --> 00:35:36.792 Some disc wears and then divide everything

NOTE Confidence: 0.9411312

00:35:36.792 --> 00:35:39.827 by the number of observation minus one.

NOTE Confidence: 0.9411312

00:35:39.830 --> 00:35:42.553 So this is how we calculate the

NOTE Confidence: 0.9411312

00:35:42.553 --> 00:35:45.220 variance of the expression of GD.

NOTE Confidence: 0.9411312

00:35:45.220 --> 00:35:48.125 So the formula here is the following.

NOTE Confidence: 0.9411312

00:35:48.130 --> 00:35:50.620 So we take the six distances,

NOTE Confidence: 0.9411312

00:35:50.620 --> 00:35:52.244 we square the distances.

NOTE Confidence: 0.9411312

00:35:52.244 --> 00:35:55.678 With some the results and we divide by 5.

NOTE Confidence: 0.9411312

00:35:55.680 --> 00:35:59.684 So the variance of jinei is 30.8.

NOTE Confidence: 0.9411312

00:35:59.690 --> 00:36:03.148 We can do the same for jinbe.

NOTE Confidence: 0.9411312

00:36:03.150 --> 00:36:06.153 And in order to have the variance

NOTE Confidence: 0.9411312

00:36:06.153 --> 00:36:08.525 associated with Gene B Now looking

NOTE Confidence: 0.9411312

00:36:08.525 --> 00:36:10.535 at this blotter A and C,

NOTE Confidence: 0.9411312

00:36:10.540 --> 00:36:13.652 it seems by I that gene A has

NOTE Confidence: 0.9411312

00:36:13.652 --> 00:36:14.430 more differences,

NOTE Confidence: 0.9411312

00:36:14.430 --> 00:36:16.380 higher variance than Gene B,

NOTE Confidence: 0.9411312

00:36:16.380 --> 00:36:19.292 and you can see these also by

NOTE Confidence: 0.9411312

00:36:19.292 --> 00:36:22.031 looking at the range of the axis

NOTE Confidence: 0.9411312

00:36:22.031 --> 00:36:25.710 minus 6 to 6 -- 4 -- 4 to four.

NOTE Confidence: 0.9411312

00:36:25.710 --> 00:36:29.600 So we can so the variance of GA is 30.8.

NOTE Confidence: 0.9411312

00:36:29.600 --> 00:36:30.782 In this case,

NOTE Confidence: 0.9411312

00:36:30.782 --> 00:36:34.349 the variance of gene B is less in this.

NOTE Confidence: 0.9411312

00:36:34.350 --> 00:36:37.089 Cases in 8.3.

NOTE Confidence: 0.9411312

00:36:37.090 --> 00:36:38.815 So the calculation of the

NOTE Confidence: 0.9411312

00:36:38.815 --> 00:36:41.269 variance of Steam V is the same,

NOTE Confidence: 0.9411312

00:36:41.270 --> 00:36:42.658 but instead of projecting

NOTE Confidence: 0.9411312

00:36:42.658 --> 00:36:44.393 cells on the X axis,

NOTE Confidence: 0.9411312

00:36:44.400 --> 00:36:47.008 we project cells on the Y axis and

NOTE Confidence: 0.9411312

00:36:47.008 --> 00:36:49.620 that's how I come with these results.

NOTE Confidence: 0.9411312

00:36:49.620 --> 00:36:52.012 Now we can see that if we consider

NOTE Confidence: 0.9411312

00:36:52.012 --> 00:36:54.279 the global variance of our data

NOTE Confidence: 0.9411312

00:36:54.279 --> 00:36:55.875 along these two dimensions,

NOTE Confidence: 0.9411312

00:36:55.880 --> 00:36:58.728 we can say that.

NOTE Confidence: 0.9411312

00:36:58.730 --> 00:37:01.985 That that the expression of gene A

NOTE Confidence: 0.9411312

00:37:01.985 --> 00:37:04.932 contains 80% of the global variance.

NOTE Confidence: 0.9411312

00:37:04.932 --> 00:37:07.950 And the expression of Gene B

NOTE Confidence: 0.9411312

00:37:08.049 --> 00:37:11.022 contains 20% approximately of the

NOTE Confidence: 0.9411312

00:37:11.022 --> 00:37:14.646 whole variance where the whole the

NOTE Confidence: 0.9411312

00:37:14.646 --> 00:37:18.127 whole variance is just 30.8 + 8.3.

NOTE Confidence: 0.9411312

00:37:18.130 --> 00:37:20.560 So if now I have to select only one

NOTE Confidence: 0.9411312

00:37:20.560 --> 00:37:23.643 of these dimension based on the on the

NOTE Confidence: 0.9411312

00:37:23.643 --> 00:37:25.960 fact that variation is information,

NOTE Confidence: 0.9411312

00:37:25.960 --> 00:37:27.224 I would select Jeannie.

NOTE Confidence: 0.9411312

00:37:27.224 --> 00:37:30.141 So if I have to drop one of the

NOTE Confidence: 0.9411312

00:37:30.141 --> 00:37:32.353 genes I would drop Gene B because

NOTE Confidence: 0.9089681

00:37:32.426 --> 00:37:34.510 it contains less information,

NOTE Confidence: 0.9089681

00:37:34.510 --> 00:37:35.930 less variance than Jenny.

NOTE Confidence: 0.8130577

00:37:38.230 --> 00:37:41.074 Now the question for PCA is

NOTE Confidence: 0.8130577

00:37:41.074 --> 00:37:44.348 whether it is whether is there a

NOTE Confidence: 0.8130577

00:37:44.348 --> 00:37:47.330 line that is not jeanae origin B.

NOTE Confidence: 0.8130577

00:37:47.330 --> 00:37:50.823 It's not one of these that captures

NOTE Confidence: 0.8130577

00:37:50.823 --> 00:37:53.030 more variation that maximizes

NOTE Confidence: 0.8130577

00:37:53.030 --> 00:37:56.130 the variation that is captured.

NOTE Confidence: 0.8130577

00:37:56.130 --> 00:37:59.746 So the question is to try to calculate

NOTE Confidence: 0.8130577

00:37:59.746 --> 00:38:02.209 the variance that is associated

NOTE Confidence: 0.8130577

00:38:02.209 --> 00:38:05.137 with each of these possible lines

NOTE Confidence: 0.8130577

00:38:05.137 --> 00:38:08.200 in the same way as we did here.

NOTE Confidence: 0.8130577

00:38:08.200 --> 00:38:10.882 But the changing the line and

NOTE Confidence: 0.8130577

00:38:10.882 --> 00:38:12.670 so changing this calculation.

NOTE Confidence: 0.8130577

00:38:12.670 --> 00:38:16.606 So this is a problem of like minimization

NOTE Confidence: 0.8130577

00:38:16.606 --> 00:38:21.086 of the distances or maximization of the.

NOTE Confidence: 0.8130577

00:38:21.090 --> 00:38:23.349 Of the various.

NOTE Confidence: 0.8130577

00:38:23.350 --> 00:38:25.798 And so we can find that

NOTE Confidence: 0.8130577

00:38:25.798 --> 00:38:27.430 among all the possibilities,

NOTE Confidence: 0.8130577

00:38:27.430 --> 00:38:29.465 the line that maximizes the

NOTE Confidence: 0.8130577

00:38:29.465 --> 00:38:31.093 variance for our data.

NOTE Confidence: 0.8130577

00:38:31.100 --> 00:38:35.188 In this case this is the line.

NOTE Confidence: 0.8130577

00:38:35.190 --> 00:38:36.998 That maximizes the variance,

NOTE Confidence: 0.8130577

00:38:36.998 --> 00:38:40.236 and basically what we found is the

NOTE Confidence: 0.8130577

00:38:40.236 --> 00:38:42.924 principal component want of our data.

NOTE Confidence: 0.8130577

00:38:42.930 --> 00:38:44.814 So principal component principal

NOTE Confidence: 0.8130577

00:38:44.814 --> 00:38:47.640 component one is exactly that the

NOTE Confidence: 0.8130577

00:38:47.718 --> 00:38:49.823 that the dimension that maximizes

NOTE Confidence: 0.8130577

00:38:49.823 --> 00:38:52.463 the variance of data with respect

NOTE Confidence: 0.8130577

00:38:52.463 --> 00:38:54.748 to all the other possibilities

NOTE Confidence: 0.8130577

00:38:54.748 --> 00:38:57.258 toward the other possible lines.

NOTE Confidence: 0.8130577

00:38:57.258 --> 00:39:02.004 In this case that cross the origin.

NOTE Confidence: 0.8130577

00:39:02.010 --> 00:39:04.514 Now once we identify PC one PC two,

NOTE Confidence: 0.8130577

00:39:04.520 --> 00:39:06.215 so the second principle component

NOTE Confidence: 0.8130577

00:39:06.215 --> 00:39:08.658 is the line that is orthogonal to

NOTE Confidence: 0.8130577

00:39:08.658 --> 00:39:11.031 the first step and this is easy

NOTE Confidence: 0.8130577

00:39:11.031 --> 00:39:12.935 because we are in a case where

NOTE Confidence: 0.8130577

00:39:12.935 --> 00:39:16.410 we have only two dimension so.

NOTE Confidence: 0.8130577

00:39:16.410 --> 00:39:17.822 The second principle component

NOTE Confidence: 0.8130577

00:39:17.822 --> 00:39:20.425 is simply the the line that is

NOTE Confidence: 0.8130577

00:39:20.425 --> 00:39:22.505 orthogonal to the principal component,

NOTE Confidence: 0.8130577

00:39:22.510 --> 00:39:23.898 one that we found.

NOTE Confidence: 0.8130577

00:39:23.898 --> 00:39:26.700 So once we identify this principal component,

NOTE Confidence: 0.8130577

00:39:26.700 --> 00:39:28.818 now we can represent our data

NOTE Confidence: 0.8130577

00:39:28.818 --> 00:39:31.779 not from the point of view of our

NOTE Confidence: 0.8130577

00:39:31.779 --> 00:39:33.619 original jeans of the expression

NOTE Confidence: 0.8130577

00:39:33.619 --> 00:39:35.459 of our original jeans,

NOTE Confidence: 0.8130577

00:39:35.460 --> 00:39:37.833 but from the point of view of

NOTE Confidence: 0.8130577

00:39:37.833 --> 00:39:39.823 a principal component want and

NOTE Confidence: 0.8130577

00:39:39.823 --> 00:39:41.179 principal component tool.

NOTE Confidence: 0.8130577

00:39:41.180 --> 00:39:45.302 So this means that we are rotating the data.

NOTE Confidence: 0.8130577

00:39:45.310 --> 00:39:46.438 In this way,

NOTE Confidence: 0.8130577

00:39:46.438 --> 00:39:48.694 so that now our reference system

NOTE Confidence: 0.8130577

00:39:48.694 --> 00:39:51.577 system of reference is not given by

NOTE Confidence: 0.8130577

00:39:51.577 --> 00:39:54.358 our regional expression but by PC1 and PC2.

NOTE Confidence: 0.9277948

00:39:56.830 --> 00:39:59.668 But the data are always dissing.

NOTE Confidence: 0.9277948

00:39:59.670 --> 00:40:02.045 They didn't change their respective

NOTE Confidence: 0.9277948

00:40:02.045 --> 00:40:05.358 localization, so we just rotated the data.

NOTE Confidence: 0.9277948

00:40:05.360 --> 00:40:08.356 Now the advantage of doing this is

NOTE Confidence: 0.9277948

00:40:08.356 --> 00:40:11.783 that now if we calculate the variance

NOTE Confidence: 0.9277948

00:40:11.783 --> 00:40:15.320 associated with PC one and PC two,

NOTE Confidence: 0.9277948

00:40:15.320 --> 00:40:18.528 we can see that a difference with respect

NOTE Confidence: 0.9277948

00:40:18.528 --> 00:40:22.429 to our original to our regional dimensions.

NOTE Confidence: 0.9277948

00:40:22.430 --> 00:40:26.759 So we can see that PCA captures almost 100%.

NOTE Confidence: 0.9277948

00:40:26.760 --> 00:40:29.772 Of the variance of our data

NOTE Confidence: 0.9277948

00:40:29.772 --> 00:40:32.909 while PC two captures much less.

NOTE Confidence: 0.9277948

00:40:32.910 --> 00:40:35.438 And this is because.

NOTE Confidence: 0.9277948

00:40:35.440 --> 00:40:37.895 Exactly because PC one was

NOTE Confidence: 0.9277948

00:40:37.895 --> 00:40:40.350 selected because it was maximising

NOTE Confidence: 0.9277948

00:40:40.434 --> 00:40:42.849 my maximising this value here.

NOTE Confidence: 0.9277948

00:40:42.850 --> 00:40:45.132 So and here you see the difference

NOTE Confidence: 0.9277948

00:40:45.132 --> 00:40:46.828 between the variance with the

NOTE Confidence: 0.9277948

00:40:46.828 --> 00:40:48.784 original dimension gene and gene B,

NOTE Confidence: 0.9277948

00:40:48.790 --> 00:40:50.770 and with the new principal components.

NOTE Confidence: 0.9277948

00:40:50.770 --> 00:40:53.450 So the advantage of the technique is that

NOTE Confidence: 0.9277948

00:40:53.450 --> 00:40:56.706 now if I want to drop one of the dimension.

NOTE Confidence: 0.9277948

00:40:56.710 --> 00:40:59.038 So if we want to pass from 2

NOTE Confidence: 0.9277948

00:40:59.038 --> 00:41:00.669 dimensions to one dimension,

NOTE Confidence: 0.9277948

00:41:00.670 --> 00:41:02.320 if I select PC one,

NOTE Confidence: 0.9277948

00:41:02.320 --> 00:41:05.332 I lose a less than 5% of the information,

NOTE Confidence: 0.9277948

00:41:05.332 --> 00:41:07.930 while with the original gene and gene B,

NOTE Confidence: 0.9277948

00:41:07.930 --> 00:41:10.570 if I choose ginae I had to lose

NOTE Confidence: 0.9277948

00:41:10.570 --> 00:41:13.006 20% of the information in this way.

NOTE Confidence: 0.9277948

00:41:13.010 --> 00:41:14.192 I reduced dimension.

NOTE Confidence: 0.9277948

00:41:14.192 --> 00:41:16.950 It can reduce the dimension from 2

NOTE Confidence: 0.9277948

00:41:17.026 --> 00:41:19.606 dimensions to one that keeping almost

NOTE Confidence: 0.9277948

00:41:19.606 --> 00:41:22.258 all of the information of the data.

NOTE Confidence: 0.9277948

00:41:22.260 --> 00:41:25.662 And this is the trick used by

NOTE Confidence: 0.9277948

00:41:25.662 --> 00:41:27.120 principal component analysis.

NOTE Confidence: 0.9277948

00:41:27.120 --> 00:41:28.014 Ah, so.

NOTE Confidence: 0.9277948

00:41:28.014 --> 00:41:30.696 This is a more complex example,

NOTE Confidence: 0.9277948

00:41:30.700 --> 00:41:33.396 so this was an example with four dimensions.

NOTE Confidence: 0.9277948

00:41:33.400 --> 00:41:35.080 If you remember our regional,

NOTE Confidence: 0.9277948

00:41:35.080 --> 00:41:37.439 our original table was with four genes,

NOTE Confidence: 0.9277948

00:41:37.440 --> 00:41:40.473 so we can do the same with four jeans.

NOTE Confidence: 0.9277948

00:41:40.480 --> 00:41:42.676 With four dimensions we can calculate

NOTE Confidence: 0.9277948

00:41:42.676 --> 00:41:44.140 the original variance associated

NOTE Confidence: 0.9277948

00:41:44.195 --> 00:41:45.869 with each of the original jeans.

NOTE Confidence: 0.9277948

00:41:45.870 --> 00:41:48.228 So Gene age in BC and D expressed as

NOTE Confidence: 0.9277948

00:41:48.228 --> 00:41:50.918 a percentage of the entire variance.

NOTE Confidence: 0.9277948

00:41:50.920 --> 00:41:51.492 And again,

NOTE Confidence: 0.9277948

00:41:51.492 --> 00:41:53.780 if I had to choose the two genes

NOTE Confidence: 0.9277948

00:41:53.849 --> 00:41:55.979 containing most of the variance,

NOTE Confidence: 0.9277948

00:41:55.980 --> 00:41:58.091 I would choose jeanae engine, see.

NOTE Confidence: 0.9277948

00:41:58.091 --> 00:42:00.246 But still I would lose.

NOTE Confidence: 0.9277948

00:42:00.250 --> 00:42:02.824 10% of the variance associated with

NOTE Confidence: 0.9277948

00:42:02.824 --> 00:42:06.032 Gene B and 20% associated with Gene D.

NOTE Confidence: 0.9277948

00:42:06.032 --> 00:42:07.897 Like if I perform principal

NOTE Confidence: 0.9277948

00:42:07.897 --> 00:42:09.260 component transformation,

NOTE Confidence: 0.9277948

00:42:09.260 --> 00:42:11.462 I found I find four principal

NOTE Confidence: 0.9277948

00:42:11.462 --> 00:42:14.496 components in a way that the first

NOTE Confidence: 0.9277948

00:42:14.496 --> 00:42:16.976 step maximizes the explained variance.

NOTE Confidence: 0.9277948

00:42:16.980 --> 00:42:19.560 II is orthogonal to the first,

NOTE Confidence: 0.9277948

00:42:19.560 --> 00:42:21.276 and maximising maximizes the

NOTE Confidence: 0.9277948

00:42:21.276 --> 00:42:23.850 residual variance and and so on.

NOTE Confidence: 0.9277948

00:42:23.850 --> 00:42:27.058 So the advantages that now if I consider

NOTE Confidence: 0.9277948

00:42:27.058 --> 00:42:30.359 these two components and they remove it.

NOTE Confidence: 0.9277948

00:42:30.360 --> 00:42:34.490 These two I only lose that like 3 to 4%

NOTE Confidence: 0.9277948

00:42:34.490 --> 00:42:38.620 of the variance and I can keep more than 90%.

NOTE Confidence: 0.9277948

00:42:38.620 --> 00:42:40.680 While here I could keep

NOTE Confidence: 0.9277948

00:42:40.680 --> 00:42:44.009 only 70% of the variance.

NOTE Confidence: 0.9277948

00:42:44.010 --> 00:42:47.216 And if I consider only these two

NOTE Confidence: 0.9277948

00:42:47.216 --> 00:42:49.900 dimensions and I plot my data,

NOTE Confidence: 0.9277948

00:42:49.900 --> 00:42:53.516 my cells here, I can obtain this plotter.

NOTE Confidence: 0.9277948

00:42:53.520 --> 00:42:56.131 So these are are the original cells

NOTE Confidence: 0.9277948

00:42:56.131 --> 00:42:58.621 based on the expression of these

NOTE Confidence: 0.9277948

00:42:58.621 --> 00:43:01.617 four genes plotted in the first two

NOTE Confidence: 0.9277948

00:43:01.699 --> 00:43:04.011 principal component where dimension

NOTE Confidence: 0.9277948

00:43:04.011 --> 00:43:07.109 one explains 74% of the variance

NOTE Confidence: 0.9277948

00:43:07.109 --> 00:43:09.374 and dimension to explains 23%.

NOTE Confidence: 0.9277948

00:43:09.380 --> 00:43:13.568 This corresponds to these values here.

NOTE Confidence: 0.9277948

00:43:13.570 --> 00:43:15.730 And the advantagous PCA. So this.

NOTE Confidence: 0.9277948

00:43:15.730 --> 00:43:16.402 So again,

NOTE Confidence: 0.9277948

00:43:16.402 --> 00:43:18.754 the trick was to reduce the space

NOTE Confidence: 0.9277948

00:43:18.754 --> 00:43:20.768 from four to two dimensions,

NOTE Confidence: 0.9277948

00:43:20.770 --> 00:43:23.746 but keeping most of the information.

NOTE Confidence: 0.96314806

00:43:23.750 --> 00:43:26.108 And so they they new dimensions,

NOTE Confidence: 0.96314806

00:43:26.110 --> 00:43:28.777 dimension, PC one and PC two are

NOTE Confidence: 0.96314806

00:43:28.777 --> 00:43:30.812 combinations of linear combinations of

NOTE Confidence: 0.96314806

00:43:30.812 --> 00:43:33.266 the old dimensions and the advantage

NOTE Confidence: 0.96314806

00:43:33.266 --> 00:43:36.399 of PCA is that I can easily calculate

NOTE Confidence: 0.96314806

00:43:36.399 --> 00:43:38.872 how much the expression of the original

NOTE Confidence: 0.96314806

00:43:38.872 --> 00:43:41.140 jeans is important in each of the

NOTE Confidence: 0.96314806

00:43:41.213 --> 00:43:43.397 newly found principal components.

NOTE Confidence: 0.96314806

00:43:43.400 --> 00:43:46.144 For example, in a plot like this.

NOTE Confidence: 0.96314806

00:43:46.150 --> 00:43:49.420 And this is the this is a plot that shows

NOTE Confidence: 0.96314806

00:43:49.506 --> 00:43:51.941 that principal component one captures

NOTE Confidence: 0.96314806

00:43:51.941 --> 00:43:55.259 a lot of the expression of Gene 8.

NOTE Confidence: 0.96314806

00:43:55.260 --> 00:43:58.530 B&C while Gindi is not very

NOTE Confidence: 0.96314806

00:43:58.530 --> 00:44:00.710 important in principal component

NOTE Confidence: 0.96314806

00:44:00.803 --> 00:44:03.619 one while principal components,

NOTE Confidence: 0.96314806

00:44:03.620 --> 00:44:06.308 who is mainly capturing

NOTE Confidence: 0.96314806

00:44:06.308 --> 00:44:09.668 the expression of Gene D.

NOTE Confidence: 0.96314806

00:44:09.670 --> 00:44:11.006 And in this example,

NOTE Confidence: 0.96314806

00:44:11.006 --> 00:44:13.776 a explanation of this is if you look

NOTE Confidence: 0.96314806

00:44:13.776 --> 00:44:15.864 at the original values that gene

NOTE Confidence: 0.96314806

00:44:15.864 --> 00:44:18.228 AB&C are very highly correlated,

NOTE Confidence: 0.96314806

00:44:18.230 --> 00:44:20.105 so they're highly expressed in

NOTE Confidence: 0.96314806

00:44:20.105 --> 00:44:22.320 the first three cells and low.

NOTE Confidence: 0.96314806

00:44:22.320 --> 00:44:24.504 They have low expression in the

NOTE Confidence: 0.96314806

00:44:24.504 --> 00:44:27.148 in the four to the 6th cells,

NOTE Confidence: 0.96314806

00:44:27.150 --> 00:44:29.382 while gindi is a little bit

NOTE Confidence: 0.96314806

00:44:29.382 --> 00:44:31.333 different because gene is highly

NOTE Confidence: 0.96314806

00:44:31.333 --> 00:44:34.035 expressed in cell 24 and five and

NOTE Confidence: 0.96314806

00:44:34.035 --> 00:44:36.077 low expression in 1/2 and three.

NOTE Confidence: 0.96314806

00:44:36.080 --> 00:44:39.500 So this means that Gene D is not correlated.

NOTE Confidence: 0.96314806

00:44:39.500 --> 00:44:41.908 With the expression of the other genes,

NOTE Confidence: 0.96314806

00:44:41.910 --> 00:44:44.630 so that's why using PCA I can capture

NOTE Confidence: 0.96314806

00:44:44.630 --> 00:44:47.082 the correlated expression of these three

NOTE Confidence: 0.96314806

00:44:47.082 --> 00:44:49.632 genes in the first principal component.

NOTE Confidence: 0.96314806

00:44:49.640 --> 00:44:52.352 And and the the.

NOTE Confidence: 0.96314806

00:44:52.352 --> 00:44:56.018 Uhm, expression of Gene D that is

NOTE Confidence: 0.96314806

00:44:56.018 --> 00:44:59.410 different and not correlated with the other.

NOTE Confidence: 0.96314806

00:44:59.410 --> 00:45:01.374 Using the second component.

NOTE Confidence: 0.96314806

00:45:01.374 --> 00:45:02.847 The second dimension,

NOTE Confidence: 0.96314806

00:45:02.850 --> 00:45:05.796 obviously in the real case scenario

NOTE Confidence: 0.96314806

00:45:05.796 --> 00:45:07.269 we start from,

NOTE Confidence: 0.96314806

00:45:07.270 --> 00:45:10.216 if we start from 3000 genes,

NOTE Confidence: 0.96314806

00:45:10.220 --> 00:45:13.160 we start from 3000 of dimensions.

NOTE Confidence: 0.96314806

00:45:13.160 --> 00:45:16.076 But if you look at PCA

NOTE Confidence: 0.96314806

00:45:16.076 --> 00:45:18.560 plots up sometimes you can.

NOTE Confidence: 0.96314806

00:45:18.560 --> 00:45:21.070 You can always find also

NOTE Confidence: 0.96314806

00:45:21.070 --> 00:45:23.078 the percentage of variance.

NOTE Confidence: 0.96314806

00:45:23.080 --> 00:45:25.810 That is explained from each dimension

NOTE Confidence: 0.96314806

00:45:25.810 --> 00:45:29.905 so you can see how much of the entire

NOTE Confidence: 0.96314806

00:45:29.905 --> 00:45:32.666 information of the data can be

NOTE Confidence: 0.96314806

00:45:32.666 --> 00:45:35.071 explained only using two dimensions

NOTE Confidence: 0.96314806

00:45:35.071 --> 00:45:37.930 and how much you are missing.

NOTE Confidence: 0.8432129

00:45:40.080 --> 00:45:41.860 Uh, no PCA is, uh,

NOTE Confidence: 0.8432129

00:45:41.860 --> 00:45:43.888 it was worth explaining because he's

NOTE Confidence: 0.8432129

00:45:43.888 --> 00:45:46.840 still one of the most used at techniques.

NOTE Confidence: 0.8432129

00:45:46.840 --> 00:45:48.976 Also in single cell data analysis.

NOTE Confidence: 0.8432129

00:45:48.980 --> 00:45:52.076 But you don't see PCA offer in the

NOTE Confidence: 0.8432129

00:45:52.076 --> 00:45:54.219 visualization of single cell data.

NOTE Confidence: 0.8432129

00:45:54.220 --> 00:45:56.225 And that's because the principal

NOTE Confidence: 0.8432129

00:45:56.225 --> 00:45:58.296 component analysis, as I said,

NOTE Confidence: 0.8432129

00:45:58.296 --> 00:46:00.828 has the advantage of being highly

NOTE Confidence: 0.8432129

00:46:00.828 --> 00:46:02.508 interpretable because from the

NOTE Confidence: 0.8432129

00:46:02.508 --> 00:46:04.866 components I can go back quite

NOTE Confidence: 0.8432129

00:46:04.866 --> 00:46:07.449 easily to the to the original jeans,

NOTE Confidence: 0.8432129

00:46:07.450 --> 00:46:10.730 so I can establish which genes are important

NOTE Confidence: 0.8432129

00:46:10.730 --> 00:46:13.469 are important in each of the dimensions.

NOTE Confidence: 0.8432129

00:46:13.470 --> 00:46:15.206 It is computationally efficient,

NOTE Confidence: 0.8432129

00:46:15.206 --> 00:46:18.218 but when I want to visualize a

NOTE Confidence: 0.8432129

00:46:18.218 --> 00:46:21.780 single cell RNA seek data, it's.

NOTE Confidence: 0.8432129

00:46:21.780 --> 00:46:23.550 It's lesser it's not very

NOTE Confidence: 0.8432129

00:46:23.550 --> 00:46:24.966 appealing to the eye,

NOTE Confidence: 0.8432129

00:46:24.970 --> 00:46:27.794 so and the reason for this is again,

NOTE Confidence: 0.8432129

00:46:27.800 --> 00:46:30.624 that the data in single cell are nonlinear.

NOTE Confidence: 0.8432129

00:46:30.630 --> 00:46:32.748 They have an excess of zeros,

NOTE Confidence: 0.8432129

00:46:32.750 --> 00:46:35.228 and so if you plot the principle,

NOTE Confidence: 0.8432129

00:46:35.230 --> 00:46:37.000 the first two principal components,

NOTE Confidence: 0.8432129

00:46:37.000 --> 00:46:39.124 often you don't have a clear

NOTE Confidence: 0.8432129

00:46:39.124 --> 00:46:40.186 separation of cells,

NOTE Confidence: 0.8432129

00:46:40.190 --> 00:46:42.668 and that's what you want to show,

NOTE Confidence: 0.8432129

00:46:42.670 --> 00:46:46.499 especially if you want if you're generating.

NOTE Confidence: 0.8432129

00:46:46.500 --> 00:46:49.884 Figure that is going to represent your data.

NOTE Confidence: 0.8432129

00:46:49.890 --> 00:46:51.582 So for this reason,

NOTE Confidence: 0.8432129

00:46:51.582 --> 00:46:53.274 mainly for the visualization,

NOTE Confidence: 0.8432129

00:46:53.280 --> 00:46:56.248 not for the analysis of the data,

NOTE Confidence: 0.8432129

00:46:56.250 --> 00:46:58.889 come in the first year of single

NOTE Confidence: 0.8432129

00:46:58.889 --> 00:47:01.520 cell analysis of the most employed

NOTE Confidence: 0.8432129

00:47:01.520 --> 00:47:03.880 approach was called the Disney,

NOTE Confidence: 0.8432129

00:47:03.880 --> 00:47:06.430 so it's at least a caustic

NOTE Confidence: 0.8432129

00:47:06.430 --> 00:47:07.280 neighborhood embedding.

NOTE Confidence: 0.8432129

00:47:07.280 --> 00:47:09.818 So this approach is not linear,

NOTE Confidence: 0.8432129

00:47:09.820 --> 00:47:11.512 as principal component is

NOTE Confidence: 0.8432129

00:47:11.512 --> 00:47:13.204 based on graph methods.

NOTE Confidence: 0.8432129

00:47:13.210 --> 00:47:16.507 So on this I will not spend.

NOTE Confidence: 0.8432129

00:47:16.510 --> 00:47:19.765 A lot in explaining how it works,

NOTE Confidence: 0.8432129

00:47:19.770 --> 00:47:23.760 but basically it's a random procedure and

NOTE Confidence: 0.8432129

00:47:23.760 --> 00:47:27.760 being nonlinear it means that it corrects.

NOTE Confidence: 0.8432129

00:47:27.760 --> 00:47:29.930 The original data using the

NOTE Confidence: 0.8432129

00:47:29.930 --> 00:47:31.666 nonlinear equation equation and

NOTE Confidence: 0.8432129

00:47:31.666 --> 00:47:34.202 the advantage is that it's better

NOTE Confidence: 0.8432129

00:47:34.202 --> 00:47:36.232 in showing clusters of cells,

NOTE Confidence: 0.8432129

00:47:36.240 --> 00:47:37.449 so that's it.

NOTE Confidence: 0.8432129

00:47:37.449 --> 00:47:39.867 It's able to retain the local

NOTE Confidence: 0.8432129

00:47:39.867 --> 00:47:42.729 structure of the data in low dimension

NOTE Confidence: 0.8432129

00:47:42.729 --> 00:47:44.847 where the low structure local

NOTE Confidence: 0.8432129

00:47:44.847 --> 00:47:47.132 structure means cluster of cells

NOTE Confidence: 0.8432129

00:47:47.132 --> 00:47:50.229 that are very similar to each other.

NOTE Confidence: 0.8432129

00:47:50.229 --> 00:47:52.767 The disadvantage is that it's a

NOTE Confidence: 0.8432129

00:47:52.767 --> 00:47:55.538 stochastic method so that each iteration

NOTE Confidence: 0.8432129

00:47:55.538 --> 00:47:57.803 can produce a different result.

NOTE Confidence: 0.8432129

00:47:57.810 --> 00:47:59.270 That's not true for PCA.

NOTE Confidence: 0.8432129

00:47:59.270 --> 00:48:01.300 It has a long time to run,

NOTE Confidence: 0.8432129

00:48:01.300 --> 00:48:03.328 especially when you increase the number

NOTE Confidence: 0.8432129

00:48:03.328 --> 00:48:06.290 of cells and it's considered to be bad then.

NOTE Confidence: 0.8432129

00:48:06.290 --> 00:48:09.618 In keeping the global structure of the data,

NOTE Confidence: 0.8432129

00:48:09.620 --> 00:48:12.525 and I have an example of design,

NOTE Confidence: 0.8432129

00:48:12.530 --> 00:48:15.858 so this is a data set with a.

NOTE Confidence: 0.8432129

00:48:15.860 --> 00:48:18.392 I think it's balcony seeker samples

NOTE Confidence: 0.8432129

00:48:18.392 --> 00:48:20.430 from different cancer from the.

NOTE Confidence: 0.8432129

00:48:20.430 --> 00:48:23.111 So each color here is a sample

NOTE Confidence: 0.8432129

00:48:23.111 --> 00:48:25.566 from a different cancer type and

NOTE Confidence: 0.8432129

00:48:25.566 --> 00:48:28.415 then the same data where run twice

NOTE Confidence: 0.8432129

00:48:28.505 --> 00:48:31.145 with the Disney approach and these

NOTE Confidence: 0.8432129

00:48:31.145 --> 00:48:34.618 are the two outputs so you can see

NOTE Confidence: 0.8432129

00:48:34.618 --> 00:48:36.753 that that something is conserved.

NOTE Confidence: 0.8432129

00:48:36.760 --> 00:48:38.448 Between the two run.

NOTE Confidence: 0.8432129

00:48:38.448 --> 00:48:41.718 Uh, so the number of cluster and they come,

NOTE Confidence: 0.8432129

00:48:41.720 --> 00:48:44.933 the size of the class and probably they they.

NOTE Confidence: 0.8432129

00:48:44.940 --> 00:48:46.890 The assignment of each sample to

NOTE Confidence: 0.8432129

00:48:46.890 --> 00:48:48.860 each cluster has been conserved.

NOTE Confidence: 0.8432129

00:48:48.860 --> 00:48:51.261 Them and also the shape of this

NOTE Confidence: 0.8432129

00:48:51.261 --> 00:48:52.790 single clastres somehow closer.

NOTE Confidence: 0.8432129

00:48:52.790 --> 00:48:55.558 But if you look at the organization of

NOTE Confidence: 0.8432129

00:48:55.558 --> 00:48:58.498 the whole cluster of the of the classes,

NOTE Confidence: 0.8432129

00:48:58.500 --> 00:48:59.574 that is different.

NOTE Confidence: 0.8432129

00:48:59.574 --> 00:49:00.290 For example,

NOTE Confidence: 0.8432129

00:49:00.290 --> 00:49:02.205 these orange cluster here in

NOTE Confidence: 0.8432129

00:49:02.205 --> 00:49:04.120 run one is in the

NOTE Confidence: 0.9128054

00:49:04.204 --> 00:49:06.997 middle and while uh and and green

NOTE Confidence: 0.9128054

00:49:06.997 --> 00:49:09.448 is opposite to read the wild.

NOTE Confidence: 0.9128054

00:49:09.450 --> 00:49:12.969 Red and green are very near to each other,

NOTE Confidence: 0.9128054

00:49:12.970 --> 00:49:15.707 so for the capturing the class time,

NOTE Confidence: 0.9128054

00:49:15.710 --> 00:49:17.270 visualizing the class set

NOTE Confidence: 0.9128054

00:49:17.270 --> 00:49:18.830 this method it's good.

NOTE Confidence: 0.9128054

00:49:18.830 --> 00:49:21.259 But then if I start the interpreting

NOTE Confidence: 0.9128054

00:49:21.259 --> 00:49:23.530 the distance between different clusters,

NOTE Confidence: 0.9128054

00:49:23.530 --> 00:49:26.260 these methods is not any more valuable,

NOTE Confidence: 0.9128054

00:49:26.260 --> 00:49:28.962 so it's not reliable because it depending

NOTE Confidence: 0.9128054

00:49:28.962 --> 00:49:32.126 on the initial random step of the analysis,

NOTE Confidence: 0.9128054

00:49:32.130 --> 00:49:34.356 it could lead to different maps

NOTE Confidence: 0.9128054

00:49:34.356 --> 00:49:36.430 and that's the main reason.

NOTE Confidence: 0.9662275

00:49:36.430 --> 00:49:38.778 Yeah, I'm sorry, yeah.

NOTE Confidence: 0.9662275

00:49:38.780 --> 00:49:41.131 I was just wondering so like is there

NOTE Confidence: 0.9662275

00:49:41.131 --> 00:49:43.566 any value in like sort of running the

NOTE Confidence: 0.9662275

00:49:43.566 --> 00:49:45.660 program like a bunch of times, right?

NOTE Confidence: 0.9662275

00:49:45.660 --> 00:49:47.035 Like just iteratively and then

NOTE Confidence: 0.9662275

00:49:47.035 --> 00:49:48.950 taking the average of the distances?

NOTE Confidence: 0.9662275

00:49:48.950 --> 00:49:50.954 There's some truth that emerges there

NOTE Confidence: 0.9662275

00:49:50.954 --> 00:49:53.317 when you like repeat it a whole bunch

NOTE Confidence: 0.9662275

00:49:53.317 --> 00:49:55.230 of times or it's just not useful.

NOTE Confidence: 0.83529186

00:49:56.620 --> 00:49:59.875 Uhm, I don't think I can answer

NOTE Confidence: 0.83529186

00:49:59.875 --> 00:50:03.274 to that personally, so I wouldn't

NOTE Confidence: 0.83529186

00:50:03.274 --> 00:50:07.282 know the answer to this question.

NOTE Confidence: 0.83529186

00:50:07.290 --> 00:50:09.817 Uhm, I don't know if anyone tried,

NOTE Confidence: 0.83529186

00:50:09.820 --> 00:50:11.815 so there is a way to reproduce

NOTE Confidence: 0.83529186

00:50:11.815 --> 00:50:13.686 the analysis to performing so

NOTE Confidence: 0.83529186

00:50:13.686 --> 00:50:15.586 called of pseudorandom analysis,

NOTE Confidence: 0.83529186

00:50:15.590 --> 00:50:16.790 meaning random analysis.

NOTE Confidence: 0.83529186

00:50:16.790 --> 00:50:19.964 When you run a program is based on

NOTE Confidence: 0.83529186

00:50:19.964 --> 00:50:22.085 a seed that is a random number,

NOTE Confidence: 0.83529186

00:50:22.090 --> 00:50:23.900 but it can be kept.

NOTE Confidence: 0.83529186

00:50:23.900 --> 00:50:26.406 It can be remembered during the different

NOTE Confidence: 0.83529186

00:50:26.406 --> 00:50:28.970 iterations and if you keep the seed

NOTE Confidence: 0.83529186

00:50:28.970 --> 00:50:30.755 constant you can reproduce results,

NOTE Confidence: 0.83529186

00:50:30.760 --> 00:50:31.909 but that's not.

NOTE Confidence: 0.83529186

00:50:31.909 --> 00:50:34.207 So that's a way to keep

NOTE Confidence: 0.83529186

00:50:34.207 --> 00:50:36.428 consistent the program if you run.

NOTE Confidence: 0.83529186

00:50:36.430 --> 00:50:38.320 The program in different on

NOTE Confidence: 0.83529186

00:50:38.320 --> 00:50:39.834 different machines, for example,

NOTE Confidence: 0.83529186

00:50:39.834 --> 00:50:42.096 but I don't know if that.

NOTE Confidence: 0.9837538

00:50:44.550 --> 00:50:45.990 If this has been done,

NOTE Confidence: 0.9837538

00:50:45.990 --> 00:50:47.901 maybe so, but I don't know what

NOTE Confidence: 0.9837538

00:50:47.901 --> 00:50:49.720 would be the result of that.

NOTE Confidence: 0.9837538

00:50:49.720 --> 00:50:51.624 So to run it a lot of times

NOTE Confidence: 0.9837538

00:50:51.624 --> 00:50:53.499 and trying to capture a sort

NOTE Confidence: 0.9837538

00:50:53.499 --> 00:50:55.169 of stability in the distances,

NOTE Confidence: 0.9837538

00:50:55.170 --> 00:50:59.042 sure sure. So in the field,

NOTE Confidence: 0.9837538

00:50:59.042 --> 00:51:01.724 are the fact that if, for example,

NOTE Confidence: 0.9837538

00:51:01.724 --> 00:51:05.170 if you look at the publication so you can,

NOTE Confidence: 0.9837538

00:51:05.170 --> 00:51:07.788 you can like data the analysis according

NOTE Confidence: 0.9837538

00:51:07.788 --> 00:51:10.920 to the method that they used to visualize.

NOTE Confidence: 0.9837538

00:51:10.920 --> 00:51:13.424 So if you look at some plot and

NOTE Confidence: 0.9837538

00:51:13.424 --> 00:51:15.580 it's at Disney probability analysis

NOTE Confidence: 0.9837538

00:51:15.580 --> 00:51:18.954 is before a 2018 2018 because in

NOTE Confidence: 0.9837538

00:51:19.035 --> 00:51:21.731 2018 the what is now the the the

NOTE Confidence: 0.9837538

00:51:21.731 --> 00:51:23.820 most used approach to visualize

NOTE Confidence: 0.9837538

00:51:23.820 --> 00:51:26.640 the single cell data has been

NOTE Confidence: 0.9837538

00:51:26.640 --> 00:51:28.970 presented and that's the you map.

NOTE Confidence: 0.9837538

00:51:28.970 --> 00:51:29.981 The human method.

NOTE Confidence: 0.9837538

00:51:29.981 --> 00:51:34.029 So if you see a plot that is uses as

NOTE Confidence: 0.9837538

00:51:34.029 --> 00:51:36.177 a dimensionality reduction technique,

NOTE Confidence: 0.9837538

00:51:36.180 --> 00:51:38.416 the human deaths from

NOTE Confidence: 0.9837538

00:51:38.416 --> 00:51:41.770 2018 to now more or less.

NOTE Confidence: 0.9837538

00:51:41.770 --> 00:51:44.158 So the humor is also another

NOTE Confidence: 0.9291115

00:51:44.160 --> 00:51:46.958 question is also common, sort of, uh,

NOTE Confidence: 0.9291115

00:51:46.958 --> 00:51:49.744 so I've actually noticed that if you,

NOTE Confidence: 0.9291115

00:51:49.750 --> 00:51:50.938 for example Disney,

NOTE Confidence: 0.9291115

00:51:50.938 --> 00:51:54.140 if you just have a completely random data,

NOTE Confidence: 0.9291115

00:51:54.140 --> 00:51:57.796 so let's say, OK, I generate the a

NOTE Confidence: 0.9291115

00:51:57.796 --> 00:52:00.790 computer generated completely random data.

NOTE Confidence: 0.9291115

00:52:00.790 --> 00:52:02.644 Supposedly it will be a fuzzy

NOTE Confidence: 0.9291115

00:52:02.644 --> 00:52:04.300 ball in this PC plot,

NOTE Confidence: 0.9291115

00:52:04.300 --> 00:52:07.152 but if you do in the Disney it will

NOTE Confidence: 0.9291115

00:52:07.152 --> 00:52:09.329 become some kind of patterns you can

NOTE Confidence: 0.9291115

00:52:09.329 --> 00:52:11.318 start to see emerging from that.

NOTE Confidence: 0.9291115

00:52:11.320 --> 00:52:13.114 I just wonder for things like

NOTE Confidence: 0.9291115

00:52:13.114 --> 00:52:15.150 you map and other things. Is

NOTE Confidence: 0.9728567

00:52:15.150 --> 00:52:17.172 this also the same problem or

NOTE Confidence: 0.9728567

00:52:17.172 --> 00:52:19.806 it's a yeah it's the same problem

NOTE Confidence: 0.9728567

00:52:19.806 --> 00:52:22.224 so all these methods because all

NOTE Confidence: 0.9728567

00:52:22.224 --> 00:52:24.828 these methods like try to maximize.

NOTE Confidence: 0.9728567

00:52:24.830 --> 00:52:27.116 The separation of these objects and

NOTE Confidence: 0.9728567

00:52:27.116 --> 00:52:29.927 the the problem is that the boss 40s

NOTE Confidence: 0.9728567

00:52:29.927 --> 00:52:32.599 and also for your map you have noise.

NOTE Confidence: 0.9728567

00:52:32.600 --> 00:52:34.712 So if your differences mainly driven

NOTE Confidence: 0.9728567

00:52:34.712 --> 00:52:37.190 by noise, UM, they tried, they they.

NOTE Confidence: 0.9728567

00:52:37.190 --> 00:52:39.296 They basically create patterns from noise,

NOTE Confidence: 0.9728567

00:52:39.300 --> 00:52:43.100 yeah. And this is a less a problem

NOTE Confidence: 0.9728567

00:52:43.100 --> 00:52:45.429 with the PCA. That's right, yeah.

NOTE Confidence: 0.9728567

00:52:45.430 --> 00:52:47.614 So this is not solved by the human,

NOTE Confidence: 0.9728567

00:52:47.620 --> 00:52:49.924 So what it seems to be solved by

NOTE Confidence: 0.9728567

00:52:49.924 --> 00:52:52.467 the UMAP is mainly that it's faster.

NOTE Confidence: 0.9728567

00:52:52.470 --> 00:52:54.756 Uh, it's faster than his name,

NOTE Confidence: 0.9728567

00:52:54.760 --> 00:52:58.189 so it can be applied in a reasonable time.

NOTE Confidence: 0.9728567

00:52:58.190 --> 00:53:01.232 So when the data set is very high in

NOTE Confidence: 0.9728567

00:53:01.232 --> 00:53:04.017 terms of number of cells and also it

NOTE Confidence: 0.9728567

00:53:04.017 --> 00:53:06.912 seems to be better in a preserving

NOTE Confidence: 0.9728567

00:53:06.912 --> 00:53:09.992 this global structure of the data so

NOTE Confidence: 0.9728567

00:53:09.992 --> 00:53:12.224 that also the distances between the

NOTE Confidence: 0.9728567

00:53:12.224 --> 00:53:14.570 different clusters are more like reliable,

NOTE Confidence: 0.9728567

00:53:14.570 --> 00:53:17.412 I think there is actually a parameter

NOTE Confidence: 0.9728567

00:53:17.412 --> 00:53:20.658 where you kind of tune how much you are.

NOTE Confidence: 0.9728567

00:53:20.660 --> 00:53:22.780 You give weight to the.

NOTE Confidence: 0.9728567

00:53:22.780 --> 00:53:25.370 Local structure or to the global structure,

NOTE Confidence: 0.9728567

00:53:25.370 --> 00:53:27.897 but it's generally considered to be more

NOTE Confidence: 0.9728567

00:53:27.897 --> 00:53:30.918 reliable on the global structure of the data,

NOTE Confidence: 0.9728567

00:53:30.920 --> 00:53:33.356 so it's considered to be like a

NOTE Confidence: 0.9728567

00:53:33.356 --> 00:53:35.947 trade off a good tradeoff between

NOTE Confidence: 0.9728567

00:53:35.947 --> 00:53:38.863 the PCA and the Disney approach.

NOTE Confidence: 0.9728567

00:53:38.870 --> 00:53:40.040 Barca, it hasn't.

NOTE Confidence: 0.9728567

00:53:40.040 --> 00:53:42.380 So both these methods have problems.

NOTE Confidence: 0.9728567

00:53:42.380 --> 00:53:44.330 For example in interpret ability.

NOTE Confidence: 0.9728567

00:53:44.330 --> 00:53:47.678 So PCA is easy to go back to the

NOTE Confidence: 0.9728567

00:53:47.678 --> 00:53:50.163 original jeans in Disney and

NOTE Confidence: 0.9728567

00:53:50.163 --> 00:53:52.683 you map it's very problematic.

NOTE Confidence: 0.9728567

00:53:52.690 --> 00:53:53.370 And uhm,

NOTE Confidence: 0.9728567

00:53:53.370 --> 00:53:53.710 yeah.

NOTE Confidence: 0.9728567

00:53:53.710 --> 00:53:56.090 And also you map is a random

NOTE Confidence: 0.9728567

00:53:56.175 --> 00:53:57.600 so different runs,

NOTE Confidence: 0.9728567

00:53:57.600 --> 00:53:59.868 so give you slightly different results.

NOTE Confidence: 0.98754835

00:54:02.190 --> 00:54:04.828 Can I ask you another quick question?

NOTE Confidence: 0.98754835

00:54:04.830 --> 00:54:07.467 So when you say the difference in

NOTE Confidence: 0.98754835

00:54:07.467 --> 00:54:09.731 time for like processing the data,

NOTE Confidence: 0.98754835

00:54:09.731 --> 00:54:12.370 what is the scale of that time?

NOTE Confidence: 0.98754835

00:54:12.370 --> 00:54:15.009 Are you saying like hours or days?

NOTE Confidence: 0.8509196

00:54:15.910 --> 00:54:18.106 Ah well, it did, but it's,

NOTE Confidence: 0.8509196

00:54:18.110 --> 00:54:19.578 uh, well it diverted.

NOTE Confidence: 0.8509196

00:54:19.578 --> 00:54:21.413 Depending on the number of,

NOTE Confidence: 0.8509196

00:54:21.420 --> 00:54:23.508 uh, cells, so it could be

NOTE Confidence: 0.8509196

00:54:23.508 --> 00:54:25.820 that if you have 100 cells,

NOTE Confidence: 0.8509196

00:54:25.820 --> 00:54:27.650 you don't notice the difference,

NOTE Confidence: 0.8509196

00:54:27.650 --> 00:54:29.858 but scaling. So adding data you.

NOTE Confidence: 0.955195369999999

00:54:32.190 --> 00:54:33.862 You delete a lot, so for this day

NOTE Confidence: 0.955195369999999

00:54:33.862 --> 00:54:35.843 I know that there are a lot of

NOTE Confidence: 0.955195369999999

00:54:35.843 --> 00:54:37.515 variation of these names that have

NOTE Confidence: 0.955195369999999

00:54:37.515 --> 00:54:39.045 been working on the efficiency,

NOTE Confidence: 0.955195369999999

00:54:39.050 --> 00:54:42.292 so they are faster. Uh, but, uh.

NOTE Confidence: 0.955195369999999

00:54:42.292 --> 00:54:45.589 I guess it's also problem of memory,

NOTE Confidence: 0.955195369999999

00:54:45.590 --> 00:54:47.956 so personally I never run an analysis

NOTE Confidence: 0.955195369999999

00:54:47.956 --> 00:54:50.428 on a sample that was more than

NOTE Confidence: 0.955195369999999

00:54:50.428 --> 00:54:52.498 20 thirty thousand cells and so

NOTE Confidence: 0.955195369999999

00:54:52.575 --> 00:54:54.771 personally I don't know how problematic

NOTE Confidence: 0.955195369999999

00:54:54.771 --> 00:54:57.660 it is to work with Disney with a

NOTE Confidence: 0.955195369999999

00:54:57.660 --> 00:54:59.790 large data set of 1,000,000 cells.

NOTE Confidence: 0.955195369999999

00:54:59.790 --> 00:55:01.920 But the problem is that the

NOTE Confidence: 0.955195369999999

00:55:01.920 --> 00:55:03.340 more cells you are,

NOTE Confidence: 0.955195369999999

00:55:03.340 --> 00:55:05.470 the more you gain in time.

NOTE Confidence: 0.955195369999999

00:55:05.470 --> 00:55:08.958 Using you map against at least basic Disney.

NOTE Confidence: 0.955195369999999

00:55:08.960 --> 00:55:12.660 Sure. OK, so this is only technical.

NOTE Confidence: 0.955195369999999

00:55:12.660 --> 00:55:13.914 It's not really.

NOTE Confidence: 0.955195369999999

00:55:13.914 --> 00:55:18.171 So the the key point also is that depending

NOTE Confidence: 0.955195369999999

00:55:18.171 --> 00:55:21.867 on the dimensionality choice you make.

NOTE Confidence: 0.955195369999999

00:55:21.870 --> 00:55:24.078 There is an answer look different,

NOTE Confidence: 0.955195369999999

00:55:24.080 --> 00:55:27.040 so this is the same data set up,

NOTE Confidence: 0.955195369999999

00:55:27.040 --> 00:55:28.880 so normalization is the same.

NOTE Confidence: 0.955195369999999

00:55:28.880 --> 00:55:31.100 The input data was the same,

NOTE Confidence: 0.955195369999999

00:55:31.100 --> 00:55:32.940 it's from a mouse brain.

NOTE Confidence: 0.955195369999999

00:55:32.940 --> 00:55:35.154 So here you see some populations

NOTE Confidence: 0.955195369999999

00:55:35.154 --> 00:55:36.630 that correspond to neurons,

NOTE Confidence: 0.955195369999999

00:55:36.630 --> 00:55:38.838 different types of neurons and microglia,

NOTE Confidence: 0.955195369999999

00:55:38.840 --> 00:55:39.761 and solar cells.

NOTE Confidence: 0.955195369999999

00:55:39.761 --> 00:55:41.910 And you see the representation of these

NOTE Confidence: 0.955195369999999

00:55:41.967 --> 00:55:44.377 datasets using principal component analysis.

NOTE Confidence: 0.955195369999999

00:55:44.380 --> 00:55:46.768 So cells here are colored according

NOTE Confidence: 0.955195369999999

00:55:46.768 --> 00:55:49.667 to the cell type and and and you

NOTE Confidence: 0.955195369999999

00:55:49.667 --> 00:55:52.200 see that you can see the classes.

NOTE Confidence: 0.955195369999999

00:55:52.200 --> 00:55:54.426 But you but but for example,

NOTE Confidence: 0.955195369999999

00:55:54.430 --> 00:55:56.290 points within the same clusters.

NOTE Confidence: 0.955195369999999

00:55:56.290 --> 00:55:58.150 That kind of spread around,

NOTE Confidence: 0.955195369999999

00:55:58.150 --> 00:56:00.424 so that's why for the visualization

NOTE Confidence: 0.955195369999999

00:56:00.424 --> 00:56:02.726 of the cluster this is less

NOTE Confidence: 0.955195369999999

00:56:02.726 --> 00:56:04.850 clear than the other two methods.

NOTE Confidence: 0.955195369999999

00:56:04.850 --> 00:56:07.886 So these two methods try basically

NOTE Confidence: 0.955195369999999

00:56:07.886 --> 00:56:09.910 maximizes the completeness of

NOTE Confidence: 0.955195369999999

00:56:09.994 --> 00:56:12.209 the data inside the cluster.

NOTE Confidence: 0.955195369999999

00:56:12.210 --> 00:56:16.218 And using different two different approaches.

NOTE Confidence: 0.9799622

00:56:19.440 --> 00:56:21.790 And then again, for interpreting these data,

NOTE Confidence: 0.9799622

00:56:21.790 --> 00:56:23.465 it's more about just seeing

NOTE Confidence: 0.9799622

00:56:23.465 --> 00:56:25.469 like how cells are similar to

NOTE Confidence: 0.9799622

00:56:25.470 --> 00:56:27.150 each other within a cluster,

NOTE Confidence: 0.9799622

00:56:27.150 --> 00:56:28.640 like how they cluster separately

NOTE Confidence: 0.9799622

00:56:28.640 --> 00:56:30.544 as opposed to like the distances

NOTE Confidence: 0.9799622

00:56:30.544 --> 00:56:32.329 between the cluster being able

NOTE Confidence: 0.9799622

00:56:32.329 --> 00:56:34.163 to infer any relationship from

NOTE Confidence: 0.9799622

00:56:34.163 --> 00:56:35.520 that distance, right? Well,

NOTE Confidence: 0.9799622

00:56:35.520 --> 00:56:37.865 here you could be interested also well,

NOTE Confidence: 0.9799622

00:56:37.870 --> 00:56:39.880 for sure if you use Disney,

NOTE Confidence: 0.9799622

00:56:39.880 --> 00:56:42.560 the distances between them as you see here,

NOTE Confidence: 0.9799622

00:56:42.560 --> 00:56:46.070 the distances tend to be like.

NOTE Confidence: 0.9799622

00:56:46.070 --> 00:56:48.190 More or less the same,

NOTE Confidence: 0.9799622

00:56:48.190 --> 00:56:50.445 so their equally distributed button

NOTE Confidence: 0.9799622

00:56:50.445 --> 00:56:53.950 here in new map you can have a

NOTE Confidence: 0.9799622

00:56:53.950 --> 00:56:56.266 distances that are that have a

NOTE Confidence: 0.9799622

00:56:56.266 --> 00:56:58.786 range so low to high distances.

NOTE Confidence: 0.9799622

00:56:58.790 --> 00:57:01.328 So here it can be informative.

NOTE Confidence: 0.9799622

00:57:01.330 --> 00:57:03.450 But basically when you use

NOTE Confidence: 0.9799622

00:57:03.450 --> 00:57:05.146 this for the visualization,

NOTE Confidence: 0.9799622

00:57:05.150 --> 00:57:07.838 what you want to communicate is

NOTE Confidence: 0.9799622

00:57:07.838 --> 00:57:10.588 that you identify the clusters of

NOTE Confidence: 0.9799622

00:57:10.588 --> 00:57:13.612 different cells and you want to be

NOTE Confidence: 0.9799622

00:57:13.612 --> 00:57:16.670 able to see them where they are in.

NOTE Confidence: 0.9799622

00:57:16.670 --> 00:57:17.942 Which relation they are?

NOTE Confidence: 0.9799622

00:57:17.942 --> 00:57:20.390 How many cells belong to each cluster,

NOTE Confidence: 0.9799622

00:57:20.390 --> 00:57:21.359 and so on.

NOTE Confidence: 0.9799622

00:57:21.359 --> 00:57:23.297 So usually these then are annotated

NOTE Confidence: 0.9799622

00:57:23.297 --> 00:57:25.546 with the name of the cluster based

NOTE Confidence: 0.9799622

00:57:25.546 --> 00:57:27.588 on marker genes and so obviously

NOTE Confidence: 0.9799622

00:57:27.588 --> 00:57:29.423 these for the visualization are

NOTE Confidence: 0.9799622

00:57:29.423 --> 00:57:31.541 better if you want to label

NOTE Confidence: 0.9799622

00:57:31.541 --> 00:57:33.226 your clusters and then this.

NOTE Confidence: 0.9799622

00:57:33.230 --> 00:57:35.596 But PCA is still the common tool,

NOTE Confidence: 0.9799622

00:57:35.600 --> 00:57:37.553 one of the most common tools that

NOTE Confidence: 0.9799622

00:57:37.553 --> 00:57:39.990 are run in the downstream analysis,

NOTE Confidence: 0.9799622

00:57:39.990 --> 00:57:41.820 meaning that Disney new map I

NOTE Confidence: 0.9799622

00:57:41.820 --> 00:57:43.489 really used mainly for visualization

NOTE Confidence: 0.9799622

00:57:43.489 --> 00:57:45.399 of data but nothing else.

NOTE Confidence: 0.9799622

00:57:45.400 --> 00:57:47.220 Specially it isn't but PCA.

NOTE Confidence: 0.9799622

00:57:47.220 --> 00:57:49.768 Is the basis for the clustering and

NOTE Confidence: 0.9799622

00:57:49.768 --> 00:57:52.220 for the trajectory analysis and so on.

NOTE Confidence: 0.9799622

00:57:52.220 --> 00:57:55.140 So still all the pipelines many of the

NOTE Confidence: 0.9799622

00:57:55.140 --> 00:57:57.570 old departments are still using the PCA,

NOTE Confidence: 0.9799622

00:57:57.570 --> 00:57:59.940 it's just for the visualization that

NOTE Confidence: 0.9799622

00:57:59.940 --> 00:58:02.283 they use these alternative approaches.