Electronic Supplementary Material (ESI) for ChemComm.

Supplementary Information

SENSI: Signal ENhancement by Spectral Integration for Analysis of Metabolic Mixtures

Takuma Misawa,[ab] Takanori Komatsu,[ab] Yasuhiro Date,[ab] and Jun Kikuchi*[abc]

a. Graduate School of Medical Life Science, Yokohama City University (YCU), 1-7-29 Suehirocho, Tsurumi-ku, Yokohama 230-0045, Japan.

b. RIKEN Center for Sustainable Resource Science, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 235-0045, Japan. E-mail: jun.kikuchi@riken.jp

c. Graduate School of Bioagricultural Sciences and School of Agricultural Sciences, Nagoya University, 1 Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan.

Materials and methods

Sample collection

The human urine and feces samples were obtained from eight volunteers under the conditions of no limit and control. We only requested that volunteers take the samples at their convenience. The details of the sample collection method and conditions of storage are described in a previous report. [1]

NMR spectroscopy

Lyophilized human feces (400 mg) were incubated with 5 ml of water at 50°C for 15 min in a Thermomixer Comfort (Eppendorf Japan Co., Ltd., Tokyo, Japan). The supernatant was collected before centrifugation (3,345 g, 4°C, 10 min), and this step was repeated three times. The combined supernatants (total ~15 ml) were dried in a centrifugal concentrator (EYELA CVE-200D; Tokyo Rikakikai, Tokyo, Japan) and redissolved in 1 ml of phosphate-buffered solution (0.1 M K_2HPO_4/KH_2PO_4, pH 7.0) containing 90% deuterium oxide and 1 mM sodium 2,2-dimethyl-2-silapentane-5-sulfonate as an internal standard for NMR spectroscopy. The NMR spectra were acquired at 298 K with a Bruker AVANCEIII-600 spectrometer equipped with a DCH cryoprobe (Bruker Biospin, Rheinstetten, Germany), which is optimized to detect ^{13}C. ^{13}C NMR spectra of feces sample were acquired with a 238.9226 ppm spectral width, 32,768 data points, 0.5-s relaxation delay, 1024 scans, and 30° pulse length. For the measurements, an inverse-gated pulse sequence was used; note that other pulse sequences that can be used for detecting ^{13}C signals were tested, with inferior results, including power gate and DEPT 45° (Figure S8). Quaternary carbons cannot be detected by DEPT 45° pulse sequences, which detect ^{13}C by transferring the magnetic charge from ^{1}H to ^{13}C. On the other hands, the results showed no difference between the power gate and inverse-gated pulse sequences. It is conceivable that insufficient magnetic charge was obtained from the ^{1}H owing to the Nuclear Overhauser effect (nOe) because the relaxation delay is very short due to the use of 30° pulses in ^{13}C NMR analysis. In the present study, inverse-gated pulse sequences were used to obtain the plain spectra without the nOe due to the existence of near ^{1}H. To confirm the accuracy of the annotations, heteronuclear single quantum coherence (HSQC) and HSQC–totally correlated spectroscopy (HSQC-TOCSY) were performed. The HSQC spectrum was acquired with 512 × 1024 (f1:^{13}C × f2:^{1}H) data points, 180 × 16.0221 ppm spectral width, 1-s relaxation delay, and 32 scans. The HSQC-TOCSY spectrum was acquired under the same conditions as those of HSQC, with the exception of the number of scans (64 scans in HSQC-TOCSY) (2D spectrum data not shown). Human urine (3, 12, 24, or 48 ml) samples were dried in a centrifugal concentrator (EYELA CVE-200D; Tokyo Rikakikai, Tokyo, Japan) and redissolved in 3 ml phosphate-buffered solution (0.1 M K_2HPO_4/KH_2PO_4, pH 7.0) containing 90% deuterium oxide and 1 mM sodium 2,2-dimethyl-2-silapentane-5-sulfonate as an internal standard for NMR spectroscopy. The NMR spectra were acquired at 298 K with a Bruker 500 MHz spectrometer equipped with a BBO probe (Bruker Biospin, Rheinstetten, Germany). ^{13}C NMR spectra of the urine samples were acquired with a 236.6369 ppm spectral width, 32,768 data points, 1.5-s relaxation delay, 6144 scans, and 30° pulse length.

For the measurements, an inverse-gated pulse sequence was used. The NMR spectra were processed using TopSpin 3.1 (Bruker Biospin, Rheinstetten, Germany). Furthermore, the peaks detected were annotated using databases such as the SpinAssign database and the Human Metabolomics Database.

Peak alignment

Peak alignment was performed on 181 spectra using the *i*coshift algorithm to correct the chemical shift errors caused by differences of pH, temperature, and non-uniformity of the magnetic field due to the presence of magnetic material in each sample. At first, we performed peak alignments without setting the interval (same co-shift algorithm). However, inaccurate alignments were frequently

observed due to the differences in the peak shapes and specific peaks. Thus, intervals were set in each region, including one or more peaks, which resulted in improved alignment quality.

Quantum chemistry calculations

Quantum chemistry calculations were performed using Spartan Software (Wavefunction Inc., California, USA) to confirm the accuracy of the annotations derived from the STOCSY results and to annotate the low-intensity peaks that emerged from the spectral integration. The method was in accordance with that of a previous study.[2]

Statistical analysis

All statistical analyses were performed using an original script that we wrote using R language on revolution R open (8.0.1 beta 64 bit). The computational speed is increased by using revolution R due to the multicore processing on a routine basis, as compared to standard R. First, probabilistic quotient normalization (PQN) was performed. [3] PQN enables high-quality normalization by choosing a reliable internal standard. PQN cannot be used on spectra involved in the error of chemical shifts because correction value between each spectra is calculated by each chemical shift. However, we avoided this weak point by performing peak alignment. Next, we integrated the aligned spectra that had PQN performed on it and calculated the coefficients of variation. To calculate the signal-to-noise ratio (S/N ratio), manual peak picking was performed on all the spectra and integrated spectra (by bootstrap sampling; n = 1, 20, 40, 60, 80, 100, 120, 140, 160, and 180; peak picking was repeated 20 times for each “n”).

Supplementary figures

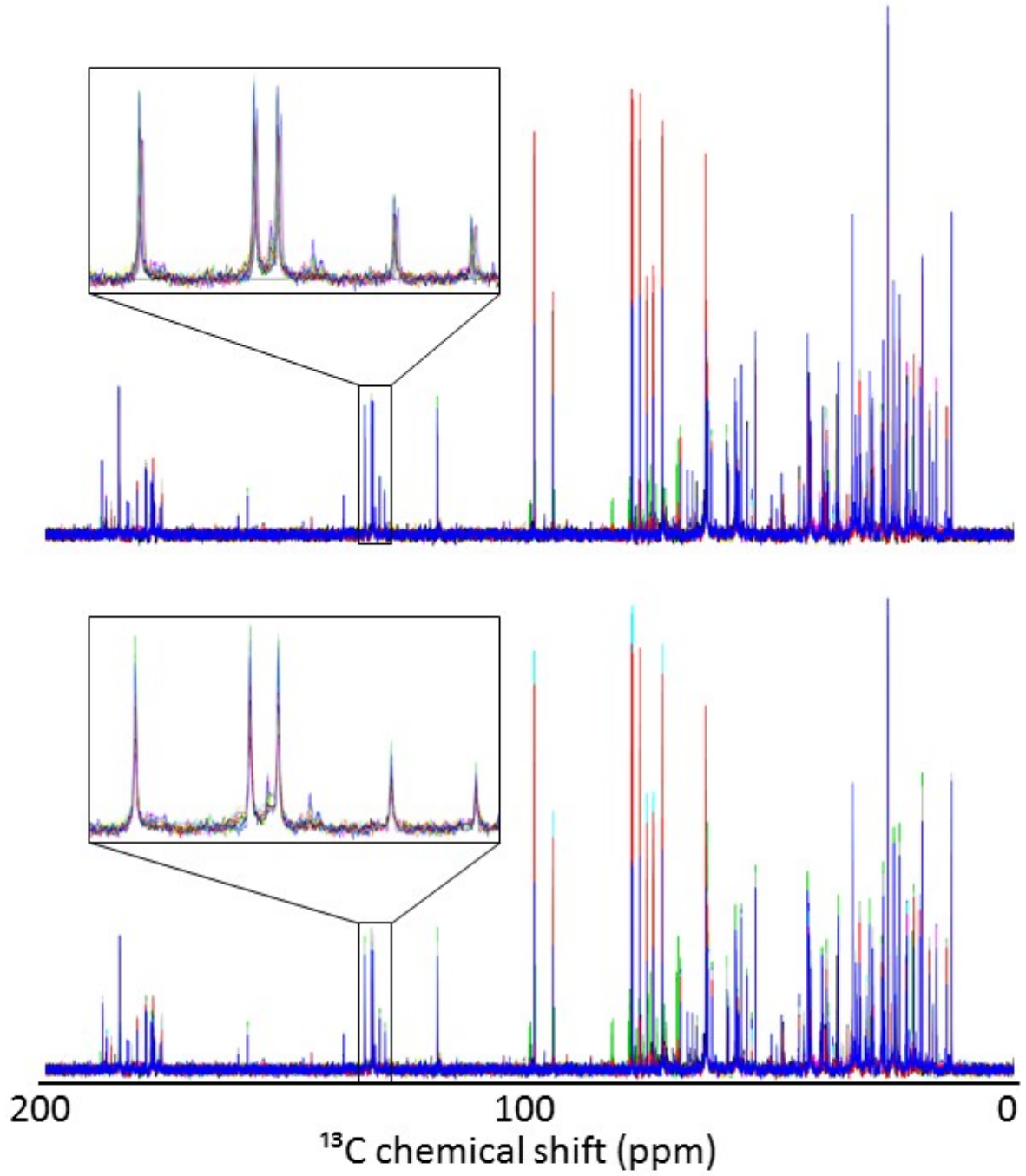


Figure S1. The spectra before alignment (upper) and after alignment (lower). Using the alignment process, the reliability of the statistical analysis using chemical shifts obtained by peak picking is increased because the peak tops are aligned.

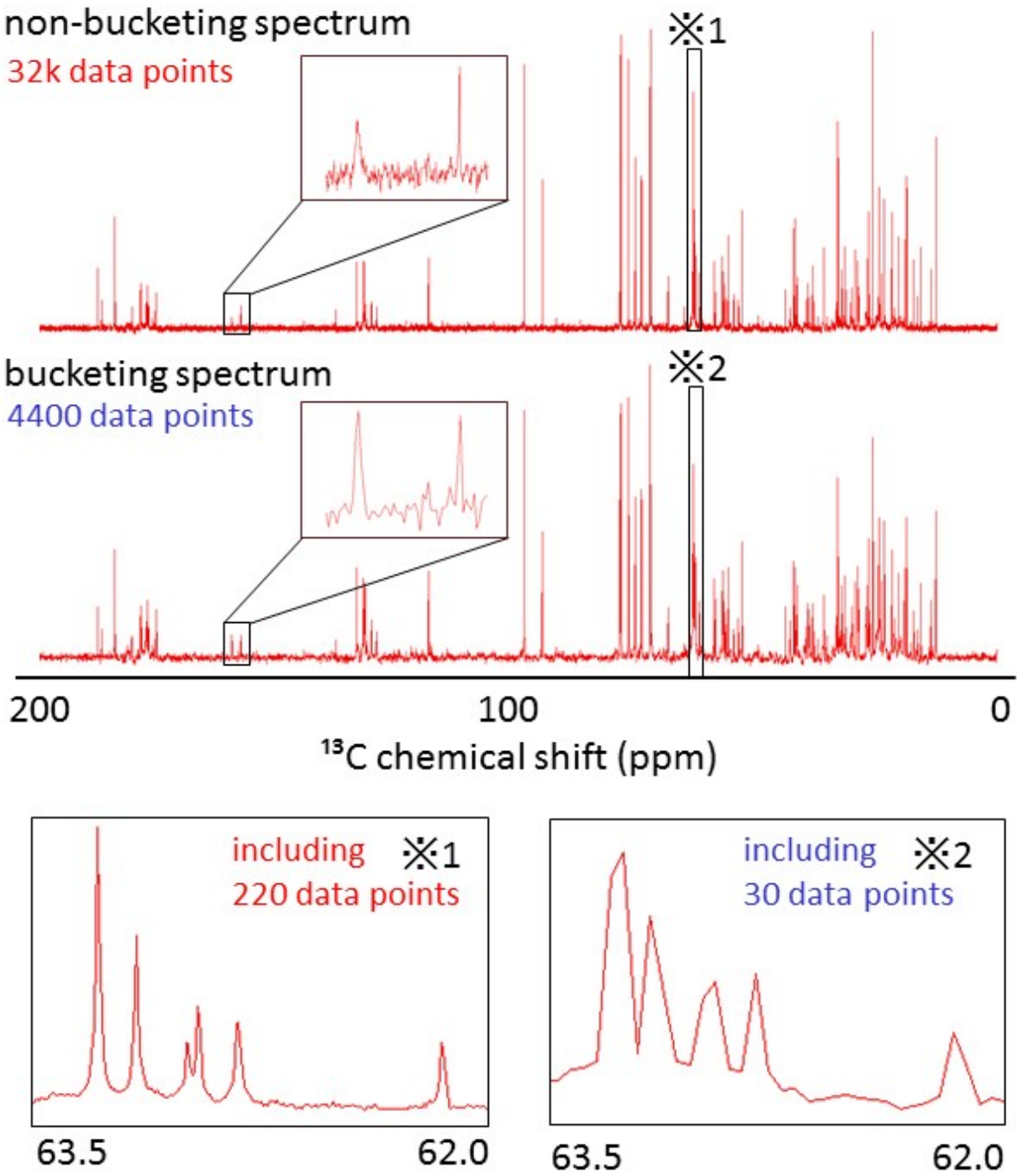


Figure S2. ^{13}C NMR spectra of a human feces sample. The upper panel shows a non-bucketing spectrum, while the lower panel shows a spectrum with binning performed on it by 0.05 ppm. The non-bucketing spectrum is not likely to be affected by the width of the signal and can keep the high resolution of ^{13}C NMR.

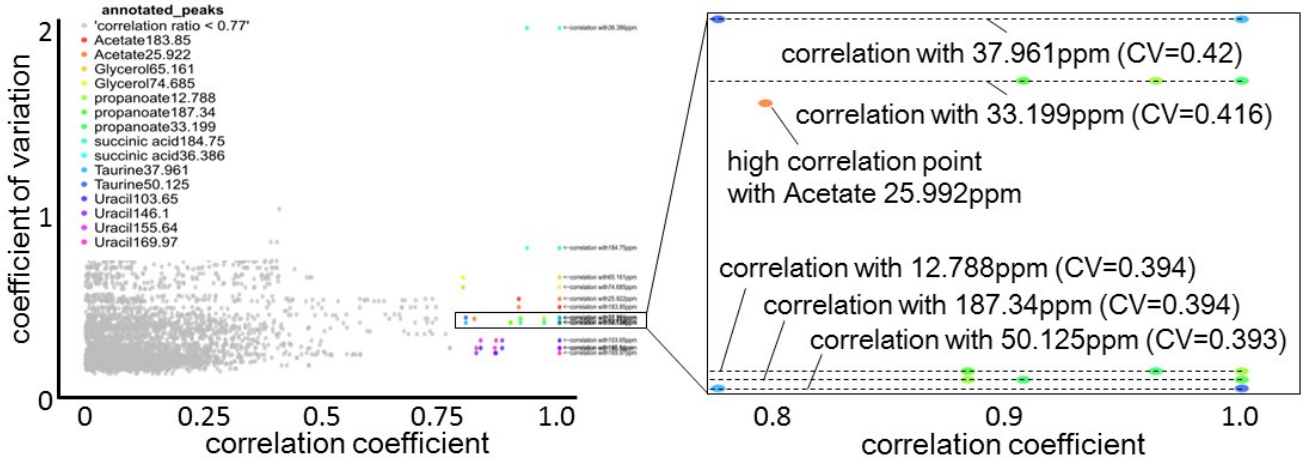


Figure S3. A scatter plot of the correlation coefficients vs the coefficients of variation (CV). It appears that the peaks that are possibly derived from the same metabolites have nearly identical CV values.

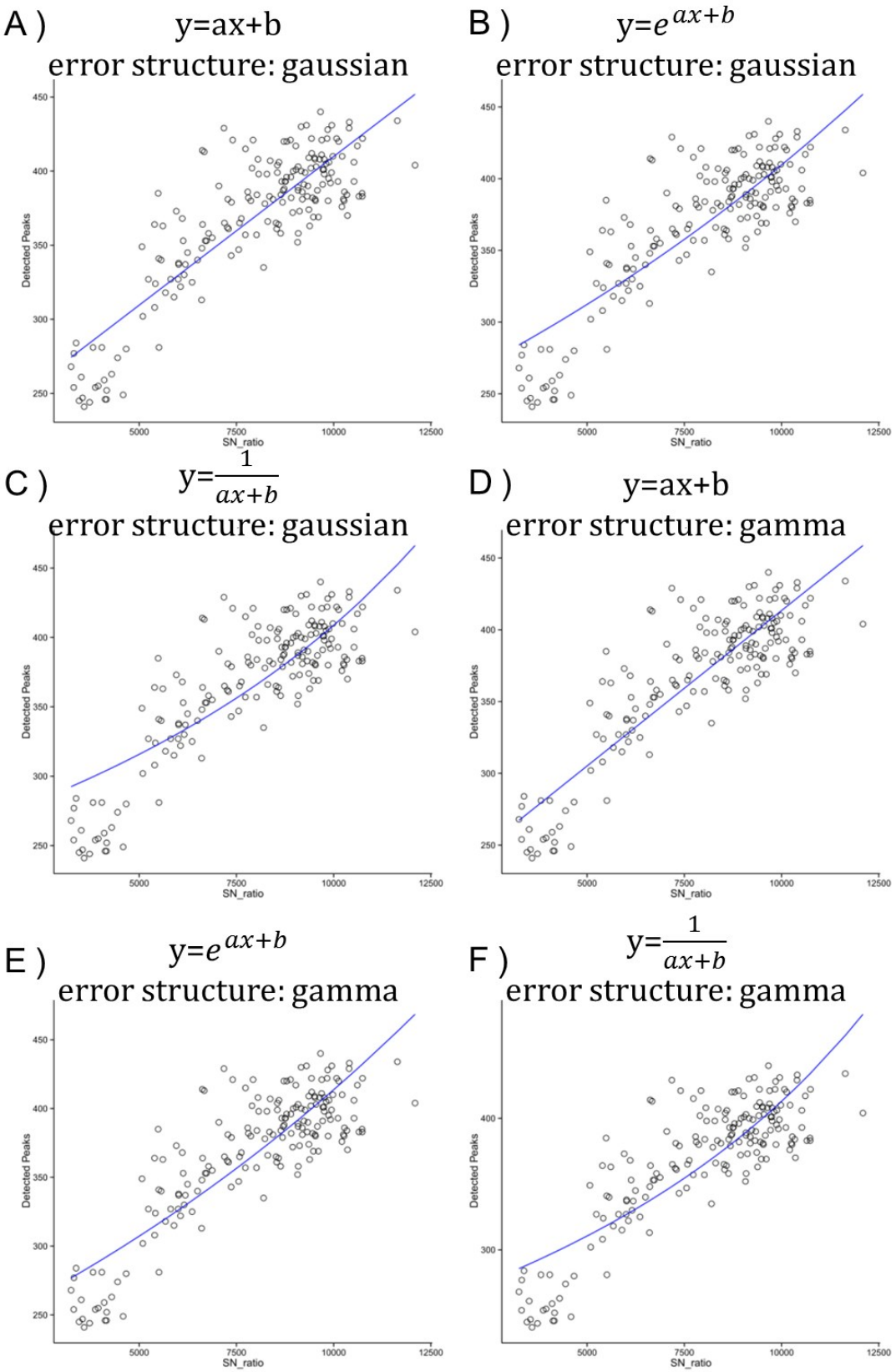


Figure S4. Scatter plots of the S/N ratios vs the numbers of detectable peaks and regression curves (A–O), and a table showing the results of the calculations of AIC and BIC for each regression model. The logarithmic model (O) has the best fit to the performing models. Unfortunately, correct prediction could not be performed due to the insufficiency of the S/N ratio depending on the number of spectra. However, this analysis has the potential to estimate the limit of the number of metabolites under the experimental conditions.

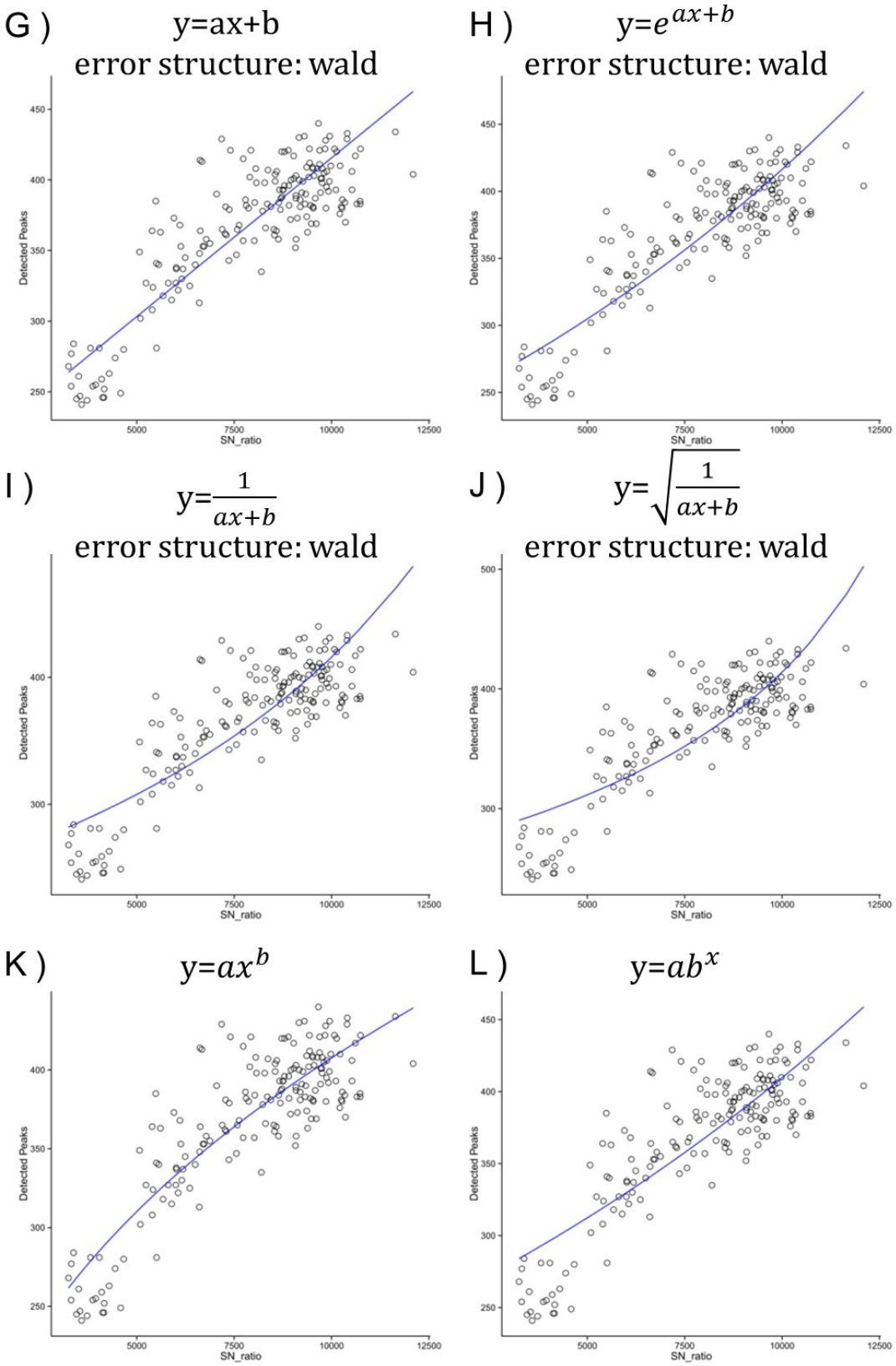
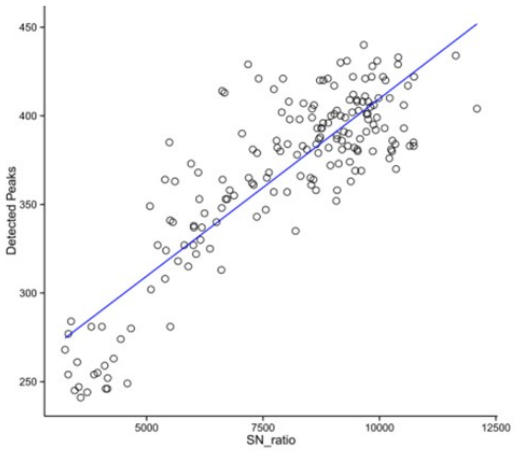


Figure S4 continue.

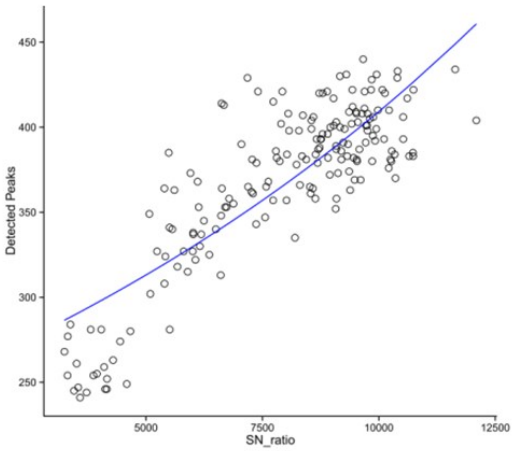
M)

$y=ab^x+c$



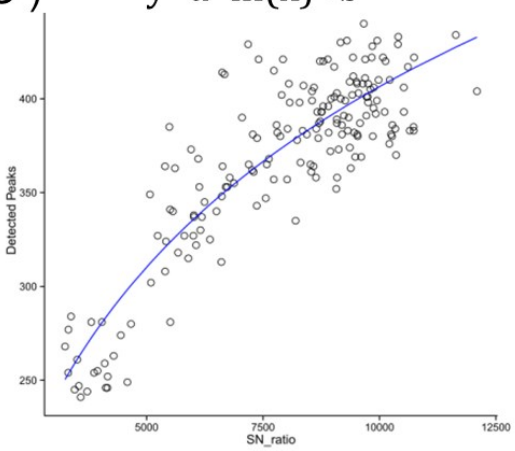
N)

$y=ab^{e^{-cx}}$



O)

$y=a\cdot\ln(x)+b$



models	error structure	a	b	c	AIC	BIC
A	gaussian	2.000.E-02	2.092.E+02	-	1683.0	1692.6
B	gaussian	5.420.E-05	5.470.E+00	-	1700.8	1710.4
C	gaussian	-1.440.E-07	3.800.E-03	-	1717.8	1727.4
D	gamma	2.164.E-02	1.969.E+02	-	1694.3	1703.9
E	gamma	5.940.E-05	5.430.E+00	-	1715.9	1725.5
F	gamma	-1.598.E-07	4.020.E-03	-	1736.2	1745.8
G	wald	2.246.E-02	1.910.E+02	-	1703.2	1712.8
H	wald	6.236.E-05	5.408.E+00	-	1726.5	1736.1
I	wald	-1.688.E-07	4.093.E-03	-	1748.6	1758.2
J	wald	-8.946.E-10	1.478.E-05	-	1768.7	1778.3
K	-	1.080.E+01	3.942.E-01	-	1655.3	1661.7
L	-	2.381.E+02	1.000.E+00	-	1698.8	1705.2
M	-	2.200.E+04	1.000.E+00	-2.179.E+04	1683.3	1692.9
N	-	1.217.E+01	2.010.E+01	-1.585.E-05	1704.0	1710.4
O	-	1.388.E+02	-8.716.E+02	-	1644.3	1653.9

Figure S4 continued.

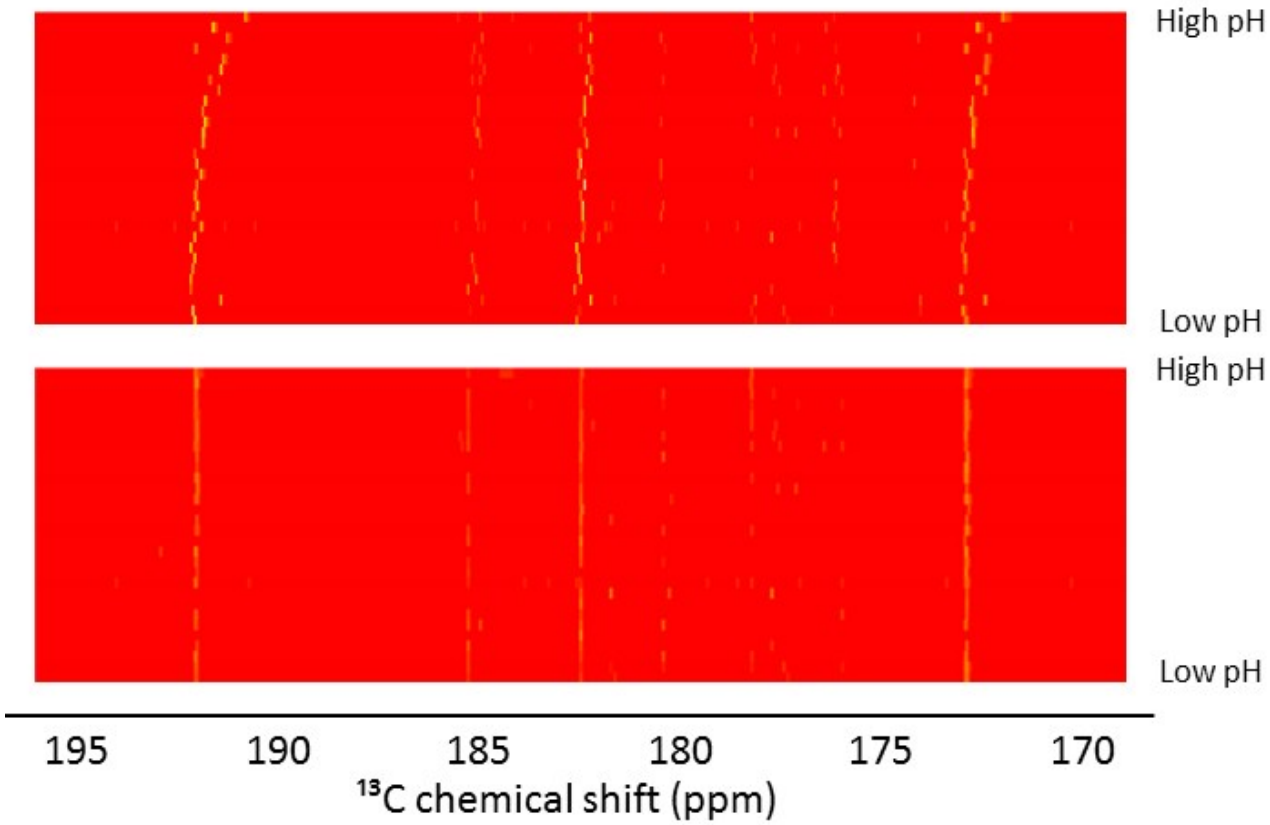


Figure S5. Peak alignment of NMR spectra obtained from expanded low field region of urine samples visualized as a heatmap. The figures showed the heatmap derived from spectral data of urine before (upper) after (lower) peak alignment. Vertical axis represents pH.

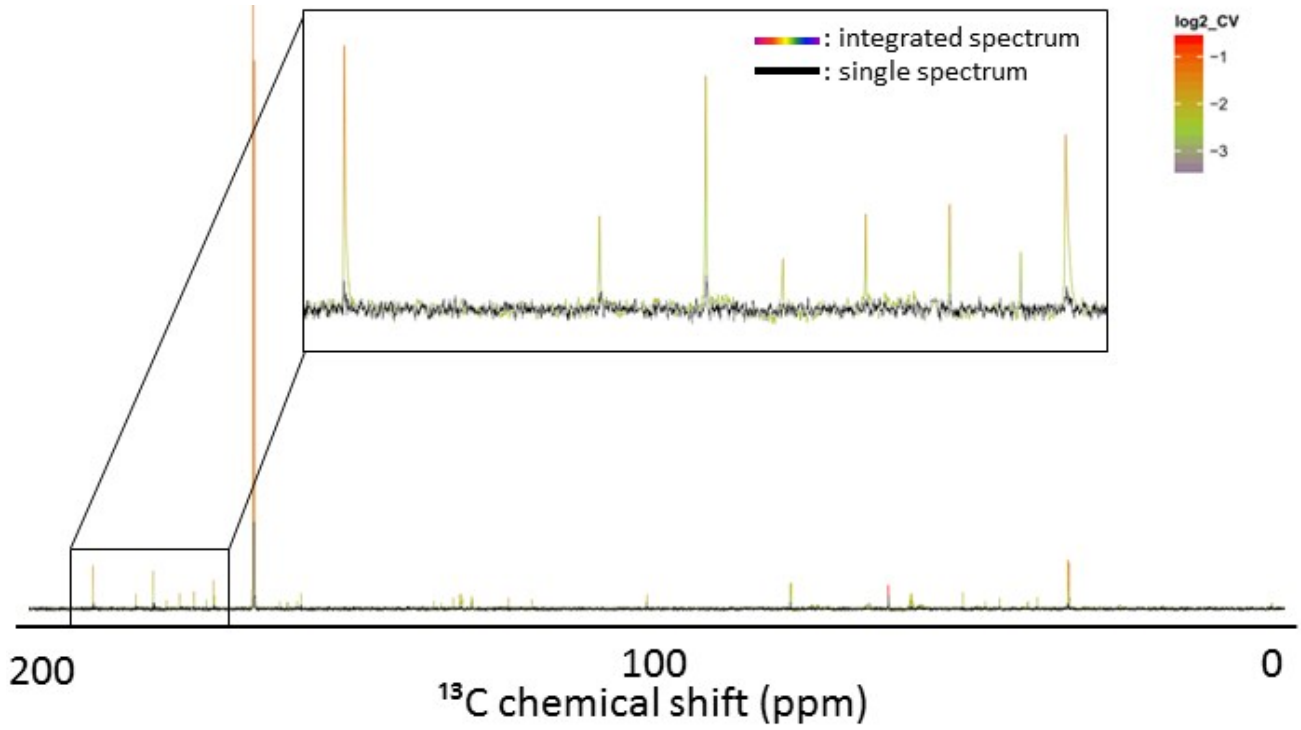


Figure S6. Comparison of single spectrum (black) versus 30 integrated SENSI data (colored) obtained from a human urine sample. Each color in the spectrum reflects the coefficient of variation (CV) in each chemical shift (high: red, low: blue). The expanded spectral region from 195 ppm to 170 ppm is shown in top. The black spectrum was normalized by calculating the difference between the maximum and minimum values of the noise (185–200 ppm).

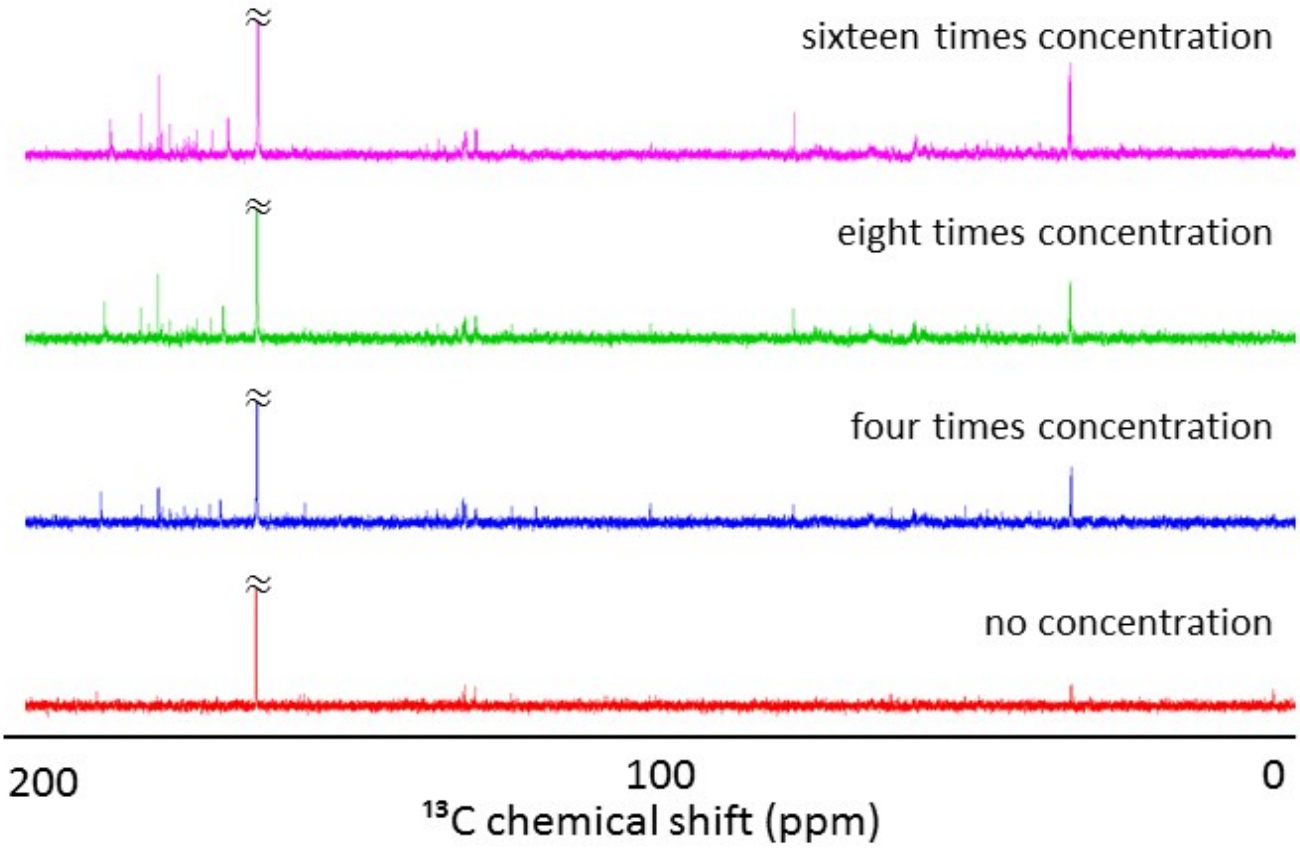


Figure S7. ^{13}C NMR spectra of urinary samples based on different concentration conditions.

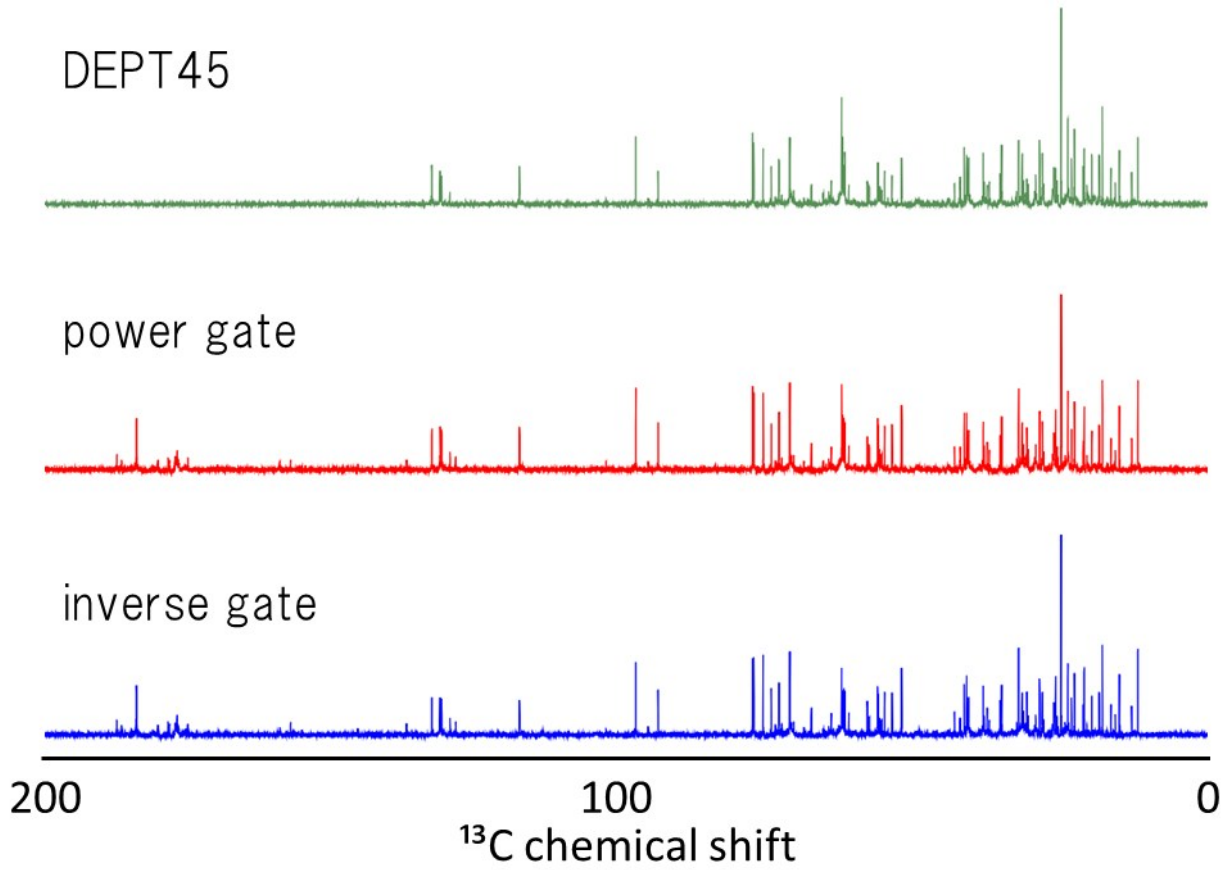


Figure S8. The spectra were acquired using different pulse sequences. Some signals included quaternary carbons, which cannot be detected by DEPT 45° (green). The result shows no differences between the power gate (red) and inverse-gated (blue) pulse sequences. Figure S8 is not cited in the main text.

R script

```
##first row: chemical shift
##the others: signal intensity

SENSI<-function(spectra_data){
##read_library
library(pracma)
library(ggplot2)

##original_function
##looplot
looplot<-function(spectra,x=c(NULL,NULL),y=c(NULL,NULL),main_title){
      main_title<-main_title
      spectra<-spectra

      for(i in 2:length(spectra[,1])){
         plot(spectra[i,],type="l",xlim=x,ylim=y,col=i,main=main_title)

         if(i!=length(spectra[,1])){
         par(new=T)
         }
      }
      legend("topright",legend=1:(length(spectra[,1])-1),col=2:length(spectra[,1]),pch=15)
   }

PPlist_ref<-function(spectra_data,noise_start_col_number,noise_end_col_number,chemical_shift,color_number){
      ##preparation
      bnmr<-spectra_data
      noise_start<-noise_start_col_number
      noise_end<-noise_end_col_number
      row_number<-length(bnmr[,1])
      chemical_shift<-chemical_shift
      color_number<-color_number
      ##peak pick
      peaklist_all<-1
      add_peak_res<-NULL
      for(i in 1){

         noise_max<-max(bnmr[i,noise_start:noise_end])
```

```
        noise_min<-min(bnmr[i,noise_start:noise_end])
        noise_max<-as.numeric(noise_max)
        noise_min<-as.numeric(noise_min)
        noise_SE<-sd(bnmr[i,noise_start:noise_end])/sqrt(length(noise_start:noise_end))

        noise_max_2<-noise_max+noise_SE
        c_val<-1
        limx<-c(max(chemical_shift),min(chemical_shift))
        limy<-NULL
        targetspectrum<-as.numeric(bnmr[i,])
        repeat{
           peaklist<-findpeaks(targetspectrum,nups=1,minpeakdistance=1,threshold=noise_max_2*c_val,npeaks=2000,sortstr=F)
           title<-"Auto pick mode on integrated spectrum"

           plot(chemical_shift,bnmr[i,], type="l", col=color_number,xlim=limx,ylim=limy,main=title)

           grid()
           points(chemical_shift[peaklist[, 2]], peaklist[, 1], pch=20, col="maroon")
           ANS<-readline("
-------Auto peak pick phase-------
OK? type 'yes' or 'no'.
If you want to fix the x or y-axis, please type the 'x' or 'y'
" )
           if(ANS=="yes"){
              break
              }else{
              if(ANS=="x"){
                 repeat{
                    umx<-readline("type the max value of x-axis        ")
                    lmx<-readline("type the min value of x-axis        ")
                    umx<-as.numeric(umx)
                    lmx<-as.numeric(lmx)
                    if((is.na(umx)!=T)&&(is.na(lmx)!=T)){
                       break
                    }
                 }
                 limx<-c(umx,lmx)
              }else{
                 if(ANS=="y"){
                    repeat{
```

```
                umy<-readline("type the max value of y-axis        ")
                lmy<-readline("type the min value of y-axis        ")
                umy<-as.numeric(umy)
                lmy<-as.numeric(lmy)
                if((is.na(umy)!=T)&&(is.na(lmy)!=T)){
                   break}
              }
              limy<-c(lmy,umy)
            }else{
              repeat{
                c_val<-readline("threshold*n type the correction value          ")
                c_val<-as.numeric(c_val)
                if(is.na(c_val)!=T){
                break}
              }
            }
          }
        }    #auto pick 1cycle#
        all_region<-data.frame(chemical_shift,bnmr[1,])
        peaklist<-as.matrix(all_region[peaklist[,2],])
        colnames(peaklist)<-NULL
        ###################################
        ####manual pick#####
        repeat{
          MANUANS<-readline("Do you want to pick more peaks by manual?
Please type 'yes' or 'no'.------>        ")

          if(MANUANS=="yes"){
            ##pick##
            title2<-"Manual pick mode on integrated spectrum"

            #manu loop#
            repeat{

              plot(chemical_shift,bnmr[i,], type="l", col=color_number,xlim=limx,ylim=limy,main=title2)

              grid()
              points(peaklist[,1], peaklist[, 2], pch=20, col="maroon")
```

```
                ACTION<-readline("
please select the next action
'x'-----> fixing of x-axis
'y'-----> fixing of x-axis
'pick'-----> peak pick in this region
'exit'-----> end of manual pick mode
")
                if(ACTION=="x"){
                repeat{
                   umx<-readline("type the max value of x-axis        ")
                   lmx<-readline("type the min value of x-axis        ")
                   umx<-as.numeric(umx)
                   lmx<-as.numeric(lmx)
                   if((is.na(umx)!=T)&&(is.na(lmx)!=T)){
                   break}
                }
                limx<-c(umx,lmx)
                }else{
                   if(ACTION=="y"){
                      repeat{
                         umy<-readline("type the max value of y-axis        ")
                         lmy<-readline("type the min value of y-axis        ")
                         umy<-as.numeric(umy)
                         lmy<-as.numeric(lmy)
                         if((is.na(umy)!=T)&&(is.na(lmy)!=T)){
                         break}
                      }
                      limy<-c(lmy,umy)
                   }else{
                      if(ACTION=="pick"){

                         manupick_frag_x<-NULL
                         manupick_frag_y<-NULL

                         repeat{
                            add_peak<-locator(1)
                            add_peak2<-add_peak
                            if((length(add_peak$x)!=1)&&(length(add_peak$y)!=1)){
                               #manu pick loop end#
```

```
                POINTFIX<-readline("ok?    'ok' or 'no'            ")
                if(POINTFIX=="ok"){
                if((length(manupick_frag_x)>=1)&&(length(manupick_frag_y)>=1)){
                  add_x<-as.matrix(manupick_frag_x)
                  add_y<-as.matrix(manupick_frag_y)

                  add_peak_res<-cbind(add_x,add_y)
                  add_peak_res<-as.matrix(add_peak_res)
                  colnames(add_peak_res)<-NULL
                  peaklist<-rbind(peaklist,add_peak_res)
                  break##manuloop region fix
                }
                }else{
                  break##manuloop region cancel
                }
              }
              ##locator_correct##
              target_region<-all_region[all_region$chemical_shift>=(add_peak$x-0.1),]##minus_correct_value
              target_region<-target_region[target_region$chemical_shift<=(add_peak$x+0.1),]##plus_correct_value

              if(length(which.max(target_region[,2]))!=1){
                add_peak_element<-target_region[which.max(target_region[,2]),]
                add_peak_element<-add_peak_element[which(abs(add_peak_element[,1]-
add_peak2$x)==min(abs(add_peak_element[,1]-add_peak2$x))),]
              }else{
                add_peak_element<-target_region[which.max(target_region[,2]),]
              }

              add_peak$x<-add_peak_element[1,1]
              add_peak$y<-add_peak_element[1,2]

              ##locator_corrected##
              points(add_peak, pch=20, adj=0.5,col="cyan")
              manupick_frag_x<-append(manupick_frag_x,add_peak$x)
              manupick_frag_y<-append(manupick_frag_y,add_peak$y)

            }
          }else{
            if(ACTION=="exit"){
            break##manuloop
```

```
                        }
                      }
                    }
                  }
                }
              }else{
              if(MANUANS=="no"){
              break##manual yes no
              }
              }
            }
            ############################
        }
        return(peaklist)
    }

refpick<-function(spectra_data,type="sum",noise_start_col_number,noise_end_col_number,filename){

    nmr<-spectra_data
    type<-type
    noise_start<-noise_start_col_number
    noise_end<-noise_end_col_number

    if(!is.character(filename)==TRUE){
      stop('type the filename (例;"name")')
      }
    filename<-paste(filename,".csv",sep="")

    nmr<-as.matrix(nmr)
    mode(nmr)<-"numeric"
    ppm<-nmr[1,]
    nmr_for_ref<-nmr[-1,]

    refdata<-t(as.matrix(apply(nmr_for_ref,2,sum)))

    refpeak<-PPlist_ref(refdata,noise_start,noise_end,ppm,1)

    peak_region<-sort(refpeak[,1],decreasing=T)
    detect_nmr<-NULL
    nmr2<-t(nmr)
```

```
  for(i in 1:length(peak_region)){
    detect_nmr<-rbind(detect_nmr,subset(nmr2,nmr2[,1]==peak_region[i]))
    }

  detect_nmr<-t(detect_nmr)

  ##write results
  write.table(detect_nmr,filename,sep=",",row.names=F,col.names=F)
  return(detect_nmr)

  }

nmr<-spectra_data
nmr<-as.matrix(nmr)
mode(nmr)<-"numeric"
chemical_shift<-nmr[1,]
anmr<-nmr[-1,]
anmr_chem<-rbind(chemical_shift,anmr)
anmr<-anmr+abs(min(anmr))##base_up

data_sum<-apply(anmr,2,sum)
data_ave<-apply(anmr,2,mean)
data_sd<-apply(anmr,2,sd)
data_cv<-data_sd/data_ave
data_cv_log2<-log2(data_cv)

data_sum<-t(as.matrix(data_sum))
data_ave<-t(as.matrix(data_ave))
data_sd<-t(as.matrix(data_sd))
data_cv<-t(as.matrix(data_cv))
data_cv_log2<-t(as.matrix(data_cv_log2))
plot_mid<-mean(data_cv_log2)

data<-rbind(chemical_shift,data_sum,data_ave,data_sd,data_cv,data_cv_log2)
dataDF<-as.data.frame(t(data))

colnames(dataDF)<-c("chemical_shift","sum_spectrum","average","sd","CV","log2_CV")

chem_max<-xmax<-dataDF$chemical_shift[1]
```

```
chem_min<-xmin<-dataDF$chemical_shift[length(dataDF$chemical_shift)]

int_max<-ymax<-NaN
int_min<-ymin<-NaN

plotbase<-ggplot(dataDF,aes(x=chemical_shift,y=sum_spectrum))
plotbase2<-plotbase+geom_line(aes(colour=log2_CV),size=0.5)+ggtitle("Integrated spectrum")
plotbase3<-plotbase2+scale_colour_gradient2(low="blue",mid="olivedrab3",high="red",midpoint=plot_mid)
plotbase4<-plotbase3+theme_classic()
plotbase5<-plotbase4+xlab("chemical shift")+ylab("sum intensity")
plotbase6<-plotbase5+theme(axis.title.x=element_text(size=15),axis.title.y=element_text(size=15))
plotbase7<-plotbase6+theme(axis.text.x=element_text(size=15),axis.text.y=element_text(size=15))

repeat{

   plotbase8<-plotbase7+xlim(chem_max,chem_min)+ylim(int_min,int_max)

   plot(plotbase8)

   repeat{
      ANS<-readline(
"If you want to fix the x-axis -> type    'x'
If you want to fix the y-axis -> type    'y'
If you want to print this figure    -> type    'print'
If you want to pick the peak    -> type    'pick'
If you want to finish display mode    -> type    'end'
")
      if((ANS=="x")||(ANS=="y")||(ANS=="end")||(ANS=="print")||(ANS=="pick")){
         break
      }
      }

   if(ANS=="x"){
      chem_max<-readline("type the max chemicalshift->")
      chem_min<-readline("type the min chemicalshift->")
      chem_max<-as.numeric(chem_max)
      chem_min<-as.numeric(chem_min)
      }else{
      if(ANS=="y"){
         int_max<-readline("type the max intensity->")
```

```
        int_min<-readline("type the min intensity->")
        int_max<-as.numeric(int_max)
        int_min<-as.numeric(int_min)
        }else{
        if(ANS=="print"){
           fig_output_name<-readline("Type the filename        ex. 'figure1.pdf'
           -> ")
           fig_output_name<-as.character(fig_output_name)
           dev.copy2pdf(file=fig_output_name)
        }else{
           if(ANS=="pick"){

              for(i in 1:length(anmr[,1])){
                 plot(anmr[i,],ylim=c(min(anmr),max(anmr)),type="l",ylab="column number of x-axis",col=i,main="select noise area")
                 if(i!=length(anmr[,1])){
                    par(new=T)
                    }
              }
              repeat{
                 noise_start<-readline("noise start column->")
                 noise_end<-readline("noise end column->")
                 noise_start<-as.numeric(noise_start)
                 noise_end<-as.numeric(noise_end)
                 if(!(is.na(noise_start))&&!(is.na(noise_end))){
                    break
                    }
                 }
              pick_data<-refpick(anmr_chem,type="sum",noise_start,noise_end,"peak pick data using SENSI")
              print("-----peak pick data was created in current directory!!-----")
              }else{
              if(ANS=="end"){
              break
                 }
              }
           }
        }
     }
  }
dataDFt<-t(dataDF)
return(dataDFt)##intagrated_data
```

```
}

##Usage

##SENSI_res<-SENSI(aligned_spectra_data)
```

Acknowledgements

The authors thank Y. Shiokawa (YCU), S. Tsutsui (YCU), and A. Shino (RIKEN) for their advice and encouragement during this study and Enago (www.enago.jp) for the English language review. The RIKEN Junior Research Associate Program also supported this study.

References

1. T. Misawa, Y. Date and J. Kikuchi, *J. Proteome Res.*, 2015, **14**, 1526-1534.
2. H. Takekawa, K. Tanaka, E. Fukushi, K. Matsuo, T. Nehira and M. Hashimoto, *J. Nat. Prod.*, 2013, **76**, 1047-1051.
3. F. Dieterle, A. Ross, G. Schlotterbeck and H. Senn, *Anal. Chem.*, 2006, **78**, 4281-4290.