Designing Highly Multiplex PCR Primer Sets with Simulated Annealing Design using Dimer Likelihood Estimation (SADDLE)

Nina G. Xie,[1] Michael X. Wang,[1] Ping Song,[1] Yifan Wang,[2] Yuxia Yang,[2] Junfeng Luo,[2] and David Yu Zhang[+1]

[1]*Department of Bioengineering, Rice University, Houston, TX*

[2]*NuProbe China, Shanghai, China*

(Dated: May 3, 2021)

S1. Experiment Protocols 1
S2. Data Analysis Workflow 3
S3. Analysis of 96 plex On-Target amplicons, Dimers and Non-specfic amplicons 5
S4. Results of libraries with adapter PCR-based protocol 11
S5. Clinical FFPE samples libraries prepared with PS3 14
S6. Results of 384-plex primer set NGS experiments 16
S7. Badness evaluation of PS2 and PS3 19
S8. Gene fusions detected and confirmed by Sanger Sequencing 20

S1. Experiment Protocols

In this paper, we performed validation of SADDLE optimized primer sets on NGS with 2 types of library preparation protocols to add adapters: ligation-based and PCR-based.

Library preparation

Sample preparation. Genomic DNA used in NGS experiments was NA18562 (Coriell Biorepository). All Genomic DNA was sheared using the Covaris M220 Focused-Ultrasonicator and Holder XTU Insert microTUBE 130L to achieve the product with the length peak at 150 nt, and quantified using Qubit 3.0 Fluorometer (Thermo Fisher Scientific) and Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific). Sheared genomic DNA was used as template input in both ligation-based and PCR-based library preparation protocol.

Ligation-based library preparation. As showed in Fig. S1, sheared genomic DNA was used for template input of first multiplex PCR. Multiplex PCR products were purified using DNA Clean & Concentrator-5 column (Zymo Research). Downstream end-prep ligation used NEBNext Ultra II DNA Library Prep Kit for Illumina, followed with column purification. The end product was quantified using qPCR and then amplified by index primers. Index PCR product was purified using AMPure XP (Beckman Coulter Life Sciences), and finally quantified using Qubit dsDNA HS Assay Kit.

PCR-based library preparation. As showed in Fig. S1, after the multiplex PCR and column purification same as the ligation-based protocol, second PCR was performed to add adapters, followed with column purification. In our project, the PCR cycles and PCR procedures may vary depends on different DNA polymerases selection and primer concentration. The product was quantified using qPCR and then sufficiently amplified with index primers. PCR amplicons were purified using AMPure XP, and finally quantified using Qubit dsDNA HS Assay Kit.

In both NGS library preparation protocols, we used PowerUp SYBR Green Master Mix (Thermo Fisher Scientific) for all qPCR quantitations; Phusion Hot Start Flex DNA Polymerase (New England Biolabs), Vent(exo-) DNA Polymerase (New England Biolabs) with standard protocol for PCR amplifications, DEPC water as dilution/elution buffer for all PCR mixture, column or AMPure XP for purification.

Comparison of two protocols. In the second step to add adapters, we compared the influence of 2 protocols: ligation and PCR. Ligation allows the enrichment of amplicons and dimers regardless of their directions; while PCR requires the amplicons to have one fP binding site and one rP binding site on each side. The reason is that when we ordered primers, fPs were attached with adapter 1 and rPs were attached with adapter 2. As the identification of amplicons in Illumina Miseq requires the attachment of p5 and p7 on each side, amplicons amplified by 2 fP or 2 rP in the multiplex PCR reaction won't be observed in final output NGS data. As shown in Fig. S2, species 2 and species 3 won't be observed in the FASTQ file when using PCR-based protocol. Utilizing PCR to add adapters will cause the missing of several species of dimers or non-specific amplicons, therefore it's not an ideal option to analyze dimers in this project.

FIG. S1: NGS library preparation protocol.

FIG. S2: Potential dimer species in an NGS library

S2. Data Analysis Workflow

1. Dimer Candidates Generation. As showed in Fig. S3, in our data analysis protocol, we did adapter trimming from raw FASTQ file and separated all the reads according to their length: we saved reads longer than 60 nt into long reads file, and reads shorter than 60 nt into short reads file. The choice of 60 nt reads length filter because of 2 reasons: 1), As in general primers are from 20 nt to 30 nt, therefore considering the overlap for at least 3 bp between two primers to dimerize, theoretically dimers are going to be shorter than 60 bp. 2), To increase the accuracy of Bowtie2 alignment[1]. We noticed that sometimes Bowtie2 can align the shorter dimer reads to the on-target reference inaccurately. As designed on-target amplicons were between 60 nt to 120 nt, therefore to prevent any potential inaccurate alignment, only the long reads files were aligned to reference sequences using Bowtie2.

2. Bowtie2 Alignment. Bowtie2 was used to do the paired-end alignment to the reference sequence. Unmapped long reads and short reads from previous steps together were termed as dimer candidates.

FIG. S3: NGS data analysis workflow for Dimer Candidates generation

3. Categorization of dimers and non-specific amplicons. We showed in Fig. S4 how we separated dimer candidates into dimers and non-specific amplicons. We first created a primer hash table, in which all the primers in a primer set are given a specific number according to their first 8 nt at 5' end. Therefore, this hash table saved all the first 8 nt information of all the primers. Noted that 8 nt at 5' end was an arbitrarily selected number. Considering the uniqueness of each primer, 8 nt was a relatively small value that we can identify each primer individually. If this number was too small, some primers shared the same first several bases couldn't be separated and would cause the collision; however, if this number was too large, the sequences would more likely to be involved in dimerization, as primer sequence approaching 3' end are more likely binding to each other.

After the creation of a hash table saving all the information of all the primers, we went over all the first 8 nt of paired-end dimer candidates reads and calculated them using the same function we created for the hash table. For an N-plex primer set, If both primers can be found in read 1 and read 2, either in the 2N×2N dimer matrix or the 2N×2N non-specific amplicon matrix, a particular position represented amplifications by two primers will add 1.

Importantly, once we successfully identified both primers in read 1 and read 2, we checked if there was any overlap between 2 primers. If an overlap more than 3 nt was between two primers, we identified this read to be a dimer, otherwise we identified it as a non-specific amplicon.

FIG. S4: Categorization of dimers and non-specific amplicons

[1] Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods, 9(4), 357.

S3. Analysis of 96 plex On-Target amplicons, Dimers and Non-specific amplicons

Experimental Protocols.

1. Multiplex PCR. We prepared PCR reaction mix by combining forward primer mix, reverse primer mix, ThermoPol Reaction Buffer (10X) , Vent(exo-) DNA Polymerase (NEB), dNTP, template solution and Milli-Q H2O. The total reaction volume was 50 μL in a 0.2mL Eppendorf PCR tube as shown in Table S1. The PCR tubes containing the reaction mixtures were placed into T100 Thermocycler (Bio-Rad), amplified following the PCR protocol listed in Table S2. Multiplex PCR products were purified using DNA Clean & Concentrator-5 (ZYMO Research).

Reagent	Working Stock Concentration	Final Concentration	Add in Volume(μL)
ThermoPol Reaction Buffer	10x	1x	5
Deoxynucleotide (dNTP)	10 mM	0.2mM	1
Forward Primer Mix	100nM	30nM	15
Reverse Primer Mix	100nM	30nM	15
DNA Template	5ng/μL	0.2ng/μL	2
Vent(exo-) DNA Polymerase	-	1 unit	0.5
Milli-Q H2O	-	-	11.5
Total Volume			50

TABLE S1: Multiplex PCR mixture formulation

Step	Temperature	Duration
1. Initial Denaturation	95°C	3 min
2. Denaturation	95°C	30 s
3. Annealing	60°C	3 min
4. Extension	72°C	30 s
Repeat Steps 2 to 4 for 17 times		
8. Final Extension	72°C	5 min
Hold at 4°C		

TABLE S2: Multiplex PCR reaction protocol

2. End-Repair and Ligation. We prepared the reaction mixture by combining NEBNext Ultra II End Prep Enzyme Mix, NEBNext Ultra II End Prep Reaction Buffer, purified multiplex PCR products and Milli-Q H2O. The total reaction volume was 60 μL in a 0.7mL Eppendorf PCR tube as shown in Table S3. The PCR tubes containing the reaction mixtures were placed into Personal Eppendorf MasterCycler Thermocycler, and the thermocycling started at 20°C for 30 min and 65 °C for 30 min, with the heated lid set to 80 °C. The reactions were then hold at 4°C.

Reagent	Add in Volume(μL)
NEBNext Ultra II End Prep Enzyme Mix	3
NEBNext Ultra II End Prep Reaction Buffer	7
Purified PCR Products	20
Milli-Q H2O	30
Total Volume	60

TABLE S3: End-Repair mixture formulation

End repair mixture was ligated with adapters using NEBNext® Ultra™ II Ligation Module (NEB). Each reaction was a mixture of previous end repair mixture, NEBNext Ultra II Ligation Master Mix, NEBNext Ligation Enhancer and NEBNext Adaptor for Illumina as shown in Table S5. Ligation was performed on a personal Eppendorf MasterCycler Thermocycler. Thermocycling started with the incubation at 20°C for 15 min with the heated lid off; after adding 3μl USER™ enzyme to the ligation mixture, the reactions were incubated at 37°C for

Reagent	Add in Volume(μL)
End Prep Reaction Mixture	60
NEBNext Ultra II Ligation Master Mix	30
NEBNext Ligation Enhancer	1
NEBNext Adaptor for Illumina	2.5
Total Volume	93.5

TABLE S4: Ligation mixture formulation

15 min with the heated lid set to 55°C. Ligation products were purified using DNA Clean & Concentrator Kits-5 (ZYMO Research).

3. Index PCR. Index PCR was performed to sufficiently enrich the ligated products. Each reaction was a mixture of p5 and p7 index primers using NEBNext® Multiplex Oligos for Illumina® (NEB), purified ligation products, PCR reagents including Vent(exo-) DNA polymerase, ThermoPol Reaction Buffer, dNTPs and Milli-Q H2O. The total volume of each reaction was 52μL in a 0.2mL Eppendorf PCR tube as shown in Table S5. Index PCR was conducted on a T100 Thermocycler or a C1000 Thermocycler (Bio-Rad), followed the protocol in Table S6. Noted that the N in Table S6 varied depend on the previous qPCR Ct value for quantitation of the ligated products. Index PCR products were purified using AMPure XP beads (Beckman Coulter). 1.8X beads were added to the Index PCR reaction mixture, 40μL Milli-Q water was used individually as elution buffer for each library.

Reagent	Add in Volume(μL)
ThermoPol Reaction Buffer(10x)	5
dNTP(10mM)	1
i5 Index primer	2
i7 Index primer	2
DNA Template	5
Vent(exo-) DNA Polymerase	0.5
Milli-Q H2O	36.5
Total Volume	52

TABLE S5: Index PCR mixture formulation

Step	Temperature	Duration
1. Initial Denaturation	95°C	3 min
2. Denaturation	95°C	30 s
3. Annealing	60°C	30 s
4. Extension	72°C	30 s
Repeat Steps 2 to 4 for N times		
8. Final Extension	72°C	5 min
Hold at 4°C		

TABLE S6: Index PCR reaction protocol

4. Library quantitation and sizing measurement. All the libraries were quantified using the Qubit™ dsDNA HS Assay Kit (Thermo Fisher Scientific). Sizes of NGS libraries were measured using Bioanalyzer High Sensitivity DNA Assay (Agilent). DNA chip was run on the Agilent 2100 Bioanalyzer system. Products sizing of the PS1, PS2 and PS3 prepared libraries are shown in Fig. S5.

5. Next generation Sequencing. All the libraries were loaded on a Miseq Reagent V2 for obtaining pair-end reads and were sequenced on a Miseq (Illumina).

Analysis of 96 plex On-Target amplicons, Dimers and Non-specific amplicons.

Length distribution of reads in the 3 primer set libraries. The NGS reads length distributions of the three species in the PS1, PS2, and PS3 libraries were consistent with the Bioanalyzer results. In these 3 libraries, on-target amplicons ranged from 90-120 nt, as designed; dimers ranged from 22-64 nt, while non-specific amplicons had a broader distribution, ranging from 30-130 nt (Fig. S6).

FIG. S5: Bioanalyzer results of PS1, PS2 and PS3 after NGS library preparation

Uniformity of 96-plex on-target amplicons. As shown in Fig. S7, amplicons were sorted according to their reads count. X-axis is the sorted 96 plex, y-axis is the log10 (Reads + 0.1). All the reads were normalized to reads count per one million total reads. PS3 library showed not only the highest on-target rate but also the best uniformity among each plex. Noted that after primer concentration modification, there were still 4 Off-target plex in PS1 library and 1 Off-target plex in PS2 library.

GC ratio of on-target amplicons. GC ratios of the amplicons were calculated in PS1, PS2 and PS3. In Fig. S8, x-axis of the beeswarm plot is the 3 primer sets, y-axis is the GC contents. Because 3 primer sets were targeting approximately the same regions, there were no significant difference in amplicon's GC ratio among 3 primer sets. (EDIT HERE)

Heatmap of non-specific amplicons. We showed in Fig. S9 distributions of non-specific amplicons in PS1-PS3 libraries. As the most dominant species in PS1 library was dimer, non-specific amplicons in PS1 were much less than in PS2 and PS3. Non-specific amplicons in PS2 and PS3 were random and not comparable to each other. This is reasonable, as primers in PS2 and PS3 were different, and their combinations could potentially

FIG. S6: Length distribution of reads in the 3 primer set libraries

FIG. S7: 96-plex On-target amplicons in 3 primer sets prepared NGS libraries

FIG. S8: GC ratio of on-target amplicons in PS1, PS2 and PS3

amplify various kind of non-specific amplicons in genome. SADDLE doesn't include the elimination of non-specific amplicons, and we will try to do optimization of non-specific amplicons in the future.

FIG. S9: Heatmap of non-specific amplicons in 3 primer sets prepared NGS libraries

Heatmap of Dimers We showed in Fig. S10 distributions of dimers in PS1-PS3 libraries. Dimer reads formed by the 192 × 192 primers in the 3 primer set libraries are shown in which primers 1-96 are fP 1-96 and primers 97-192 are rP 1-96. It is obvious that PS1 prepared library had the most dimers, while PS3 prepared libraries had the least dimers.

FIG. S10: Heatmap of dimers in 3 primer sets prepared NGS libraries

Distribution of reads observed in NGS library constructed using PS1, PS2, and PS3 (before primer concentration modification). Before primer concentration modification, the primer concentration was 30 nM in the PCR reaction. On Target reads are defined as those that aligned to the intended amplicons; Dimer reads are defined as those whose insert lengths are smaller than the sum of the two primer lengths; all other reads were classified as Non-specific. The vast majority of Non-specific reads align to other regions of the human genome, via a non-cognate pair of forward and reverse primers. The fraction of NGS reads mapped to Dimers dramatically decreases from PS1 to PS2 to PS3.

FIG. S11: NGS-based characterization of amplicons generated from PS1, PS2, and PS3 prepared libraries before primer concentration modification. We separated NGS reads into 3 major species: on-target reads, dimer reads, and non-specific reads.

S4. Results of libraries with adapter PCR-based protocol

Experimental Protocols.

1. Multiplex PCR. We prepared PCR reaction mix by combining forward primer mix, reverse primer mix, ThermoPol Reaction Buffer (10X) , Vent(exo-) DNA Polymerase, dNTP, template solution and Milli-Q H2O. The total reaction volume was 50 μL in a 0.2mL Eppendorf PCR tube as shown in Table S7. The PCR tubes containing the reaction mixtures were placed into T100 Thermocycler, amplified following the PCR protocol listed in Table S8. Multiplex PCR products were purified using DNA Clean & Concentrator Kits-5.

Reagent	Working Stock Concentration	Final Concentration	Add in Volume(μL)
ThermoPol Reaction Buffer	10x	1x	5
Deoxynucleotide (dNTP)	10 mM	0.2mM	1
Forward Primer Mix	100nM	30nM	15
Reverse Primer Mix	100nM	30nM	15
DNA Template	8ng/μL	1.6ng/μL	10
Vent(exo-) DNA Polymerase	-	1 unit	0.5
Milli-Q H2O	-	-	3.5
Total Volume			50

TABLE S7: Multiplex PCR mixture formulation

Step	Temperature	Duration
1. Initial Denaturation	95°C	3 min
2. Denaturation	95°C	30 s
3. Annealing	60°C	3 min
4. Extension	72°C	30 s
Repeat Steps 2 to 4 for 5 times		
8. Final Extension	72°C	5 min
Hold at 4°C		

TABLE S8: Multiplex PCR reaction protocol

2. Index PCR. Index PCR was performed to enrich the purified Multiplex PCR products. Each reaction was a mixture of p5 and p7 index primers using Nextera XT Index Kit (illumina), purified multiplex PCR products, and PCR reagents including Vent(exo-) DNA polymerase, ThermoPol Reaction Buffer, dNTPs and Milli-Q H2O. The total volume of each reaction was 50μL in a 0.2mL Eppendorf PCR tube as shown in Table S9. PCR amplification was conducted on a T100 Thermocycler or a C1000 Thermocycler, followed the protocol in Table S10. Index PCR products were purified using 1.2X ratio of AMPure XP beads, 40μL Milli-Q water was used individually as elution buffer for each library.

Reagent	Add in Volume(μL)
ThermoPol Reaction Buffer(10x)	5
dNTP(10mM)	1
i5 Index primer	2
i7 Index primer	2
DNA Template	5
Vent(exo-) DNA Polymerase	0.5
Milli-Q H2O	34.5
Total Volume	50

TABLE S9: Index PCR mixture formulation

Results. PCR-based library preparation to add adapters. We showed here the result of library using PCR to add adapters. In the heatmap of log10 (Reads count + 0.1) for all combinations of amplicons amplified

Step	Temperature	Duration
1. Initial Denaturation	95°C	3 min
2. Denaturation	95°C	30 s
3. Annealing	60°C	30 s
4. Extension	72°C	30 s
Repeat Steps 2 to 4 for 15 times		
8. Final Extension	72°C	5 min
Hold at 4°C		

TABLE S10: Index PCR reaction protocol

by 192×192 primers in PS1, primers 1-96 are fP 1-96, primers 97-192 are rP 1-96. Each pixel is the read count number of amplicons that were amplified by two specific primers.

A zoomed-in plot of the first quadrant of the figure is on the left. Orange-yellow slash indicated the 96-plex on-target amplicons, amplified by the 96 fPs and their corresponding 96 rPs. Apart from 96-plex on-target amplicons, there were several color points indicated a number of dimers and non-specific amplicons in the NGS reads.

Importantly, as we used PCR to add adapters, only amplicons that were amplified by 1 fP and 1 rP can be seen in the heatmap. All the fP & fP or rP & rP dimers couldn't be observed in this figure. This situation was the same for non-specific amplicons: only if the non-specific amplicons were amplified by 1 fP and 1 rP could they be seen in final data. The missing species of dimers and non-specific amplicons suggested that PCR to add adapters was not an ideal method to observe dimers compared with ligation.

FIG. S12: Heatmap of reads in PS1 library using PCR to add adapters

Dimers and non-specific amplicons distribution. We showed here heatmaps of dimers and non-specific amplicons in library using PCR to add adapters. In Fig. S13, log10 (Reads count + 0.1) of dimer and non-specific amplicons amplified by 192×192 primers in PS1, primers 1-96 are fP 1-96, primers 97-192 are rP 1-96. Each pixel is the read count number of amplicons that were amplified by two specific primers. Noted that the colorbar are the same between 2 heatmaps, indicated the majority of the off-target reads were dimers, which is consistent with the results of library prepared with ligation-based protocol.

FIG. S13: Heatmap of dimers and non-specific amplicons in PS1 library using PCR to add adapters

On-Target amplicons distribution. As shown in Fig. S14, on-target amplicons were sorted in a descending order. On-target reads were normalized to be on-target reads per 1 million reads.

FIG. S14: Sorted On-target amplicons in Adapter PCR-based library preparation protocol

S5. Clinical FFPE samples libraries prepared with PS3

Clinical FFPE samples libraries prepared with PS3.

We validated the optimized PS3 on 5 clinical samples (Origene). The cancer types of 5 samples included breast cancer, lung cancer and colorectal cancer, detailed information of sample types, patient age, patient gender and minimum stage grouping are shown in Table S11. FFPE blocks of each sample were extracted individually using GeneRead DNA FFPE Kit (Qiagen). The library preparation protocols for extracted DNA samples were the same as described in Section S3, with the input for individual reaction was 10 ng Qubit quantified DNA . Data analysis workflow was the same as described in Section S2.

We compared the on-target amplicons, dimers and non-specific amplicons in 5 NGS libraries. The highest on-target rate among the 5 clinical FFPE sample libraries was 85%, while the lowest was 29%. The variation of on-target rate may have been due to difference of FFPE samples' quality, because the FFPE samples have different amplifiable molecules even though we quantified each to be 10 ng by Qubit.

FFPE ID	Cancer type	Patient Age	Patient Gender	Minimum Stage Grouping
FFPE 1	Breast Cancer	54	Female	IIA
FFPE 2	Lung Cancer	57	Male	Not Reported
FFPE 3	Lung Cancer	75	Male	IB
FFPE 4	Colorectal Cancer	89	Female	IIA
FFPE 5	Colorectal Cancer	71	Male	IV

TABLE S11: Summary of Clinical FFPE Samples.

FIG. S15: Sorted on-target amplicons of 5 FFPE sample libraries

FIG. S16: Dimers of 5 FFPE sample libraries

FIG. S17: Non-specific amplicons of 5 FFPE sample libraries

S6. Results of 384-plex primer set NGS experiments

Experimental Protocols.

1. Multiplex PCR. We prepared PCR reaction mix by combining primer mix, Phusion buffer(5x), Phusion® Hot Start Flex DNA Polymerase (NEB), dNTP, template solution and Milli-Q H2O. The total reaction volume was 50 μL in a 0.2mL Eppendorf PCR tube as shown in Table S12. The PCR tubes containing the reaction mixtures were placed into T100 Thermocycler, amplified following the PCR protocol listed in Table S13. Multiplex PCR products were purified using DNA Clean & Concentrator Kits-5.

Reagent	Working Stock Concentration	Final Concentration	Add in Volume(μL)
Phusion HF Buffer	5x	1x	10
Deoxynucleotide (dNTP)	10 mM	0.2mM	1
Primer Mix	100nM	30nM	15
DNA Template	8ng/μL	0.8ng/μL	5
Phusion Hot Start Flex DNA Polymerase	-	1 unit	0.5
Milli-Q H2O	-	-	18.5
Total Volume			50

TABLE S12: Multiplex PCR mixture formulation

Step	Temperature	Duration
1. Initial Denaturation	98°C	1 min
2. Denaturation	98°C	10 s
3. Annealing	60°C	3 min
4. Extension	72°C	30 s
Repeat Steps 2 to 4 for 20 times		
8. Final Extension	72°C	5 min
Hold at 4°C		

TABLE S13: Multiplex PCR reaction protocol

2. End-Repair and Ligation. End-repair and ligation protocols for multiplex PCR products were the same as described in Section S3. Ligation product was purified using DNA Clean & Concentrator Kits-5.

3. Index PCR. Index PCR was performed to enrich the purified ligation products. Purified ligation product was diluted 100-fold. Each reaction was a mixture of p5 and p7 index primers using NEBNext® Multiplex Oligos for Illumina®, diluted purified ligation products, PCR reagents including Phusion buffer(5x), Phusion® Hot Start Flex DNA Polymerase, dNTPs and Milli-Q H2O. The total volume of each reaction was 52μL in a 0.2mL Eppendorf PCR tube as shown in Table S14. PCR amplification was conducted on a T100 Thermocycler or a C1000 Thermocycler, followed the protocol in Table S15. Index PCR reaction mixture was purified using 1.2X AMPure XP beads individually, 40μL Milli-Q water was used as elution buffer for each library.

Reagent	Add in Volume(μL)
Phusion HF Buffer (5x)	10
dNTP(10mM)	1
i5 Index primer	2
i7 Index primer	2
DNA Template	5
Phusion Hot Start Flex DNA Polymerase	0.5
Milli-Q H2O	31.5
Total Volume	52

TABLE S14: Index PCR mixture formulation

4. Library quantitation and sizing measurement. All the libraries were quantified using the Qubit™ dsDNA HS Assay Kit. Sizes of NGS libraries were measured using Bioanalyzer High Sensitivity DNA Assay, and

Step	Temperature	Duration
1. Initial Denaturation	98°C	1 min
2. Denaturation	98°C	10 s
3. Annealing	60°C	30 s
4. Extension	72°C	30 s
Repeat Steps 2 to 4 for 8 times		
8. Final Extension	72°C	5 min
Hold at 4°C		

TABLE S15: Index PCR reaction protocol

FIG. S18: Bioanalyzer results of 384-plex primer set prepared library.

DNA chips were run on the Agilent 2100 Bioanalyzer system. Products sizing of the 384-plex primer set prepared library is shown in Fig. S18.

Results.

Reads distribution in 384-plex primer set prepared libraries. We showed here the results of 384-plex primer set prepared NGS library. Within the total reads that passed QC and after trimmed adapter, the on-target rate of this library is 43%, the dimer rate is 1% and the non-specific amplicons rate is 56%. Heatmaps of dimers and non-specific amplicons amplified by 768 * 768 primers are shown in Fig. S19, primers 1-384 were fP 1-384, primers 385-768 were rP 385-768. Each pixel is the read count number of amplicons that were amplified by two primers. Because ligation was used to add adapters in library preparation protocol, all the fP & fP and rP & rP species can be observed.

The length distribution of reads in this library is shown in Fig. S20 on-target amplicon lengths were 60-120 nt as designed, and the dimer lengths were 26-67 nt, while the non-specific amplicon lengths were from 31-123 nt in three libraries.

FIG. S19: Heatmap of dimers and non-specific amplicons in 384-plex primer set prepared NGS libraries

FIG. S20: Length distribution of reads in the 384-plex primer set prepared NGS libraries.

S7. Badness evaluation of PS2 and PS3

Evaluation of Badness prediction in PS2 and PS3. For the best prediction of sensitivity and specificity of 36864 data points in PS2, true positive was 11, true negative was 35486, false positive was 1363 and false negative was 4. We plotted Receiver operator characteristic (ROC) accuracy of experimental dimers versus predicted badness in the determination of the dimer formation in the PS2, the AUC was 0.8146.

For the best prediction of sensitivity and specificity of 36864 data points in PS3, true positive was 4, true negative was 29477, false positive was 7377 and false negative was 6. We plotted ROC accuracy of experimental dimers versus predicted badness in the determination of the dimer formation in the PS3, the AUC was 0.5417.

The badness predictions of PS2 and PS3 were not ideal as when we optimize the dimer formation from PS1, we eliminated most possible formed dimers. However, there was still several experiment dimers that are unexpected and couldn't be predicted by the algorithm. We can look into the false positive dimers in PS2 and PS3 to improve our algorithm.

FIG. S21: Evaluation of Badness prediction. **(a)** Comparison of experimental dimers and predicted Badness in PS2. All the counts of dimer reads were normalized to be dimer reads per 1000,000 reads. Orange line was a reference of the average on-target amplicons in this library, which was also set as the fixed Threshold for Reads (TR), purple line was the flexible Threshold for Badness (TB). **(b)** Receiver Operator Characteristic (ROC) accuracy of experimental dimers versus predicted Badness in the determination of dimerization in PS2. **(c)** Comparison of experimental dimers and predicted Badness in PS3. **(d)** ROC accuracy of experimental dimers versus predicted Badness in the determination of dimerization in PS3.

S8. Gene fusions detected and confirmed by Sanger Sequencing

List of Gene Fusions detected.

Fusion Variant ID	Fusion Type	Fusion Variant ID	Fusion Type
1	EML4 exon 13 – ALK exon 20(EML4-ALK Variant 1)	29	SDC4 exon 4 – ROS1 exon 34
2	EML4 exon 20 – ALK exon 20(EML4-ALK Variant 2)	30	SLC34A2 exon 4 – ROS1 exon 32
3	EML4 exon 6 – ALK exon 20(EML4-ALK Variant 3a)	31	SLC34A2 exon 4 – ROS1 exon 34
4	EML4 exon 6 – ALK exon 20(EML4-ALK Variant 3b)	32	SLC34A2 exon 13 del – ROS1 exon 32
5	EML4 exon 6 – ALK exon 20(EML4-ALK Variant 3c)	33	SLC34A2 exon 13 del – ROS1 exon 34
6	EML4 exon 15 – ALK exon 20(EML4-ALK Variant 4)	34	CD74 exon 6 – ROS1 exon 34
7	EML4 exon 18 – ALK exon 20(EML4-ALK Variant 5)	35	TPM3 exon 8 – ROS1 exon 35
8	EML4 exon 2 – ALK exon 20(EML4-ALK Variant 5a)	36	TPM3 exon 2 – ROS1 exon 36
9	EML4 exon 14 – ALK exon 20	37	GOPC exon 7 – ROS1 exon 35
10	KIF5B exon 15 – ALK exon 20	38	GOPC exon 4 – ROS1 exon 36
11	KIF5B exon 17 – ALK exon 20	39	EZR exon 10 – ROS1 exon 34
12	KIF5B exon 24 – ALK exon 20	40	LRIG3 exon 16 – ROS1 exon 35
13	TFG exon 4 – ALK exon 20	41	KIF5B exon 15 – RET exon 11
14	HIP1 exon 28 – ALK exon 20	42	KIF5B exon 15 – RET exon 12
15	KLC1 exon 9 – ALK exon 20	43	KIF5B exon 16 – RET exon 12
16	STRN exon 3 – ALK exon 20	44	KIF5B exon 22 – RET exon 12
17	ATIC exon 7 – ALK exon 20	45	KIF5B exon 23 – RET exon 12
18	CARS exon 17 – ALK exon 20	46	KIF5B exon 24 – RET exon 8
19	CLTC exon 31 – ALK exon 20	47	KIF5B exon 24 – RET exon 11
20	DCTN1 exon 26 – ALK exon 20	48	CCDC6 exon 1 – RET exon 12
21	MSN exon 11 – ALK exon 20	49	NCOA4 exon 6 – RET exon 12
22	NPM1 exon 5 – ALK exon 20	50	TRIM33 exon 16 – RET exon 12
23	TPM3 exon 7 – ALK exon 20	51	TPM3 exon 7 – NTRK1 exon 10
24	TPM4 exon 7 – ALK exon 20	52	LMNA exon 2 – NTRK1 exon 10
25	VCL exon 16 – ALK exon 20	53	NACC2 exon 4 – NTRK2 exon 13
26	SDC4 exon 2 – ROS1 exon 32	54	QKI exon 6 – NTRK2 exon 16
27	SDC4 exon 2 – ROS1 exon 34	55	ETV6 exon 4 – NTRK3 exon 15
28	SDC4 exon 4 – ROS1 exon 32	56	ETV6 exon 5 – NTRK3 exon 15

TABLE S16: NSCLC-related gene fusions

Quantitation of 10% Fusion Variant samples. We define the Variant Allele Frequency (VAF) as fusion gene sequence / wildtype sequence, noted that wildtype sequence is the sequence of the 5' fused gene and its original adjacent exon sequence. The reason we used this sequence as wildtype sequence is that the expression level of fusions is more likely to be similar to its 5' fused gene, as they share the same promoter.

All the assumed 10% variant samples were the mixture of diluted gBlocks (Integrated DNA Technologies) and human cDNA were Random-primed qPCR Human Reference cDNA (Takara Bio). Each sample was quantitated separately. In Fig. S22, we showed the workflow to quantitate VAF of a sample: 1 universal forward primer, 1 wildtype(WT) reverse primer and 1 fusion reverse primer were designed to amplify both the variant sequence and WT sequence simultaneously, as variant and WT share the same sequence of gene 1. Thus, for each variant sample, we used 3 primers with adaptor sequences attached targeting both Variant and WT simultaneously for library preparation. The efficiency of different primers can be neglected as the first PCR was only 2 cycles and the primer concentration were much higher than the template concentration. Index PCR and purification were conducted further for library preparation. Note that this quantitation process can be multiplexed.

As shown in Fig. S22b, we used Unique molecular identifiers (UMI) attached in front of fP to quantitate variant reads and WT reads. We calculated Relative Quantity of gBlock as variant reads ×Plex Number/WT reads. Sample Variant Ratio was further calculated as $\frac{RelativeQuantity(gBlock)}{(RelativeQuantity(gBlock)+1)}$.

We presented the quantitation results of our presumed 10% Variant sample in Fig. S22c. Noted that, our presumed value was according to our experience on gBlocks and genomic DNA, and it may be different with cDNA. As in genomic DNA, the copy number of different genes are in general the same, while the expression level of different genes vary, and this may lead to different copies of cDNA sequence [1].

The x axis is the sorted 56 plex, the y axis is the quantitated VAF. The real VAF range from 1% to 90%, which showed that the pre-quantitation was necessary as the expression level of different genes varies, thus the number of different genes in cDNA varies. After the quantitation of 10% VAF variant sample, we further did a serial dilution to 1% VAF sample and ready for use to test the design of our multiplex PCR-Sanger assay.

FIG. S22: Quantitation of 10% Fusion Variant samples. **(a)** NGS workflow to quantitate synthetic fusion variant samples. **(b)** Calculation of synthetic sample's VAF. **(c)** NGS quantitation results of 56-plex presumed 10% fusion variants. **(d)** Further dilution to 1% VAF sample from NGS quantitation results.

Validation on Clinical samples. EV RNA was extracted separately from 10 NSCLC de-identified clinical plasma samples using exoRNeasy Serum/Plasma Kit (Qiagen). 10 EV RNA samples were reverse transcribed to cDNA using SuperScript™ IV Reverse Transcriptase (Thermo Fisher Scientific). NEED to make sure which kit.

We prepared PCR reaction mix by combining multiplex primer mix, iTaq polymerase supermix , cDNA template solution and Milli-Q H2O.

The total reaction volume was 50μL in a 0.2mL Eppendorf PCR tube as shown in Table S17. Final mastermix were triplicated to be 15 μL/well to a hard-shell 96-well PCR plate (Bio-rad).

Reagent	Working Stock Concentration	Final Concentration	Add in Volume(μL)
2x iTaq polymerase supermix	2x	1x	25
Primer mix	250nM	25 nM	5
template	-	-	5
Milli-Q H2O	-	-	15
Total Volume	-	-	50

TABLE S17: PCR reaction mixture formulation

The 96-well PCR plate containing the reaction mixtures was placed into T100 Thermocycler (Bio-Rad), amplified following the PCR protocol listed in Table S18.

Step	Temperature	Duration
1. Initial Denaturation	95°C	3 min
2. Denaturation	95°C	15 s
3. Annealing	60°C	1min
Repeat Steps 2 to 3 for 40 times		

TABLE S18: PCR reaction mixture formulation

PCR product was purified using ExoSAP-IT™ PCR Product Cleanup Reagent (Thermo Fisher Scientific). 15 μL PCR products and 6 μL ExoSAP-IT were pooled together in a 0.2 mL Eppendorf PCR tube, and the PCR tubes containing the reaction mixtures were placed into T100 Thermocycler and incubated at 37°C for 4 min and 80°C for 1 min, then hold at 4°C.

Cycle sequencing was conducted on purified PCR products. A mixture of purified PCR products, 5 μM sequencing primer, BigDye™Terminator v1.1 Ready Reaction Mix and BigDye™ Terminator v1.1 5X Sequencing Buffer (Thermo Fisher Scientific) and Milli-Q H2O were pooled together in a 0.2 mL Eppendorf PCR tube. The total reaction volume was 10 μL as shown in Table S19. The amplification PCR protocol is shown in Table S20.

Reagent	Add in Volume(μL)
Purified products	1
Sequencing primer (5 μM)	1
BigDye™Terminator v1.1 Reaction Mix	1
BigDye™ Terminator v1.1Sequencing Buffer	1.5
Milli-Q H2O	5.5
Total Volume	10

TABLE S19: Cycle Sequencing mixture formulation

Step	Temperature	Duration
1.	96°C	1 min
2.	96°C	10 s
3.	50°C	5 s
4.	60°C	4 min
Repeat Steps 2 to 4 for 30 times		
8.	4°C	-

TABLE S20: PCR product purification

Purification of cycle sequencing products were conducted by pooled together with 45 μL SAM solution and 10 μL BigDye XTerminator™ bead solution in a MicroAmp™ Optical 96-Well Reaction Plate (Thermo Fisher Scientific). The plate was sealed using Microseal 'B' PCR Plate Sealing Film (Bio-rad) and was vortexed on a Digital Vortex-Genie™ 2 at 1,800 rpm for 30 minutes. The plate was then centrifuged at 1,000 x g for 2 minutes and ready to be sequenced.

Results. As shown in Table S21, among 10 clinical samples, 3 were identified positive, 7 were not identified by Sanger sequencing. 2 identified by Sanger sequencing were further confirmed as EML4 exon20-ALK exon20, and one was identified as EML4 exon 15-ALK exon 20.

FIG. S23: Example sanger traces of clinical EV samples.

Sample ID	Sanger Results	Sample ID	Sanger Results
EV 1	Negative	EV 6	Negative
EV 2	Negative	EV 7	EML4 exon 20 – ALK exon 20
EV 3	EML4 exon 20 – ALK exon 20	EV 8	Negative
EV 4	EML4 exon 15 – ALK exon 20	EV 9	Negative
EV 5	Negative	EV 10	Negative

TABLE S21: Results of clinical NSCLC samples

[1] Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. Genome biology, 17, 13.

ALK Fusion Sequence

EML4 exon 13 -- ALK exon 20(EML4-ALK Variant 1)

EML4 exon 20 -- ALK exon 20(EML4-ALK Variant 2)

EML4 exon 6 -- ALK exon 20(EML4-ALK Variant 3a)

EML4 exon 6 -- ALK exon 20(EML4-ALK Variant 3b)

EML4 exon 6 -- ALK exon 20(EML4-ALK Variant 3c)

EML4 exon 15 -- ALK exon 20(EML4-ALK Variant 4)

ALK Fusion Sequence

EML4 exon 18 -- ALK exon 20(EML4-ALK Variant 5)

EML4 exon 2 -- ALK exon 20(EML4-ALK Variant 5a)

EML4 exon 14 -- ALK exon 20

KIF5B exon 15 -- ALK exon 20

KIF5B exon 17 -- ALK exon 20

KIF5B exon 24 -- ALK exon 20

ALK Fusion Sequence

TFG exon 4 -- ALK exon 20

HIP1 exon 28 -- ALK exon 20

KLC1 exon 9 -- ALK exon 20

STRN exon 3 -- ALK exon 20

ATIC exon 7 -- ALK exon 20

CARS exon 17 -- ALK exon 20

ALK Fusion Sequence

CLTC exon 31 -- ALK exon 20

DCTN1 exon 26 -- ALK exon 20

MSN exon 11 -- ALK exon 20

NPM1 exon 5 -- ALK exon 20

TPM3 exon 7 -- ALK exon 20

TPM4 exon 7 -- ALK exon 20

VCL exon 16 -- ALK exon 20

ROS1 Fusion Sequence

SDC4 exon 2 -- ROS1 exon 32

SDC4 exon 2 -- ROS1 exon 34

SDC4 exon 4 -- ROS1 exon 32

SDC4 exon 4 -- ROS1 exon 34

SLC34A2 exon 4 -- ROS1 exon 32

SLC34A2 exon 4 -- ROS1 exon 34

SLC34A2 exon 13 del -- ROS1 exon 32

SLC34A2 exon 13 del -- ROS1 exon 34

CD74 exon 6 -- ROS1 exon 32

TPM3 exon 8 -- ROS1 exon 35

TPM3 exon 2 -- ROS1 exon 36

GOPC exon 7 -- ROS1 exon 35

GOPC exon 4 -- ROS1 exon 36

EZR exon 10 -- ROS1 exon 34

LRIG3 exon 16 -- ROS1 exon 35

RET Fusion Sequence

KIF5B exon 15 -- RET exon 11

KIF5B exon 15 -- RET exon 12

KIF5B exon 16 -- RET exon 12

KIF5B exon 22 -- RET exon 12

KIF5B exon 23 -- RET exon 12

KIF5B exon 24 -- RET exon 8

KIF5B exon 24 -- RET exon 11

CCDC6 exon 1 -- RET exon 12

NCOA4 exon 6 -- RET exon 12

TRIM33 exon 16 -- RET exon 12

NTRK Fusion Sequence

TPM3 exon 7 -- NTRK1 exon 10

LMNA exon 2 -- NTRK1 exon 10

NACC2 exon 4 -- NTRK2 exon 13

QKI exon 6 -- NTRK2 exon 16

NTRK Fusion Sequence

ETV6 exon 4 -- NTRK3 exon 15

ETV6 exon 5 -- NTRK3 exon 15