Utilisation de la fréquence relative des substitutions somatiques d'un résultat de séquençage exomique comme nouvel outil diagnostique des cancers d'origine pulmonaire

Stéphane Busca

UNIVERSITÉ DE LORRAINE 2018 FACULTÉ DE MÉDECINE DE NANCY

THÈSE

Pour obtenir le grade de

DOCTEUR EN MÉDECINE

Présentée et soutenue publiquement

Dans le cadre du troisième cycle de médecine spécialisée

D'anatomie et cytologie pathologiques

Par

Stéphane BUSCA

Le 21 décembre 2018

Utilisation de la fréquence relative des substitutions somatiques d'un résultat de séquençage exomique comme nouvel outil diagnostique des cancers d'origine pulmonaire.

Membres du jury :

Président et juge : M. le Professeur Guillaume GAUCHOTTE

Juges :

M. le Docteur Guillaume VOGIN

Mme le Docteur Céline BONNET

M. le Professeur Jean-Louis MERLIN

M. le Docteur Alexandre HARLE

Président de l'Université de Lorraine
Professeur Pierre MUTZENHARDT

Doyen de la Faculté de Médecine
Professeur Marc BRAUN

Vice-doyenne
Pr Laure JOLY

Assesseurs :
Premier cycle : Dr Julien SCALA-BERTOLA
Deuxième cycle : Pr Marie-Reine LOSSER
Troisième cycle : Pr Laure JOLY

Président de Conseil Pédagogique : Pr Louise TYVAERT
Président du Conseil Scientifique : Pr Jean-Michel HASCOET

Formation à la recherche : Dr Nelly AGRINIER

SIDES : Pr Laure JOLY
Relations Grande Région : Pr Thomas FUCHS-BUDER
CUESIM : Pr Stéphane ZUILY

Chargés de mission

Bureau de docimologie : Dr Guillaume VOGIN
Orthophonie : Pr Cécile PARIETTI-WINKLER
PACES : Dr Mathias POUSSEL
International : Pr Jacques HUBERT

==========

DOYENS HONORAIRES

Professeur Jean-Bernard DUREUX - Professeur Jacques ROLAND - Professeur Patrick NETTER - Professeur Henry COUDANE

==========

PROFESSEURS HONORAIRES

Etienne ALIOT - Jean-Marie ANDRE - Alain AUBREGE - Gérard BARROCHE - Alain BERTRAND - Pierre BEY
Marc-André BIGARD - Patrick BOISSEL – Pierre BORDIGONI - Jacques BORRELLY - Michel BOULANGE
Jean-Louis BOUTROY – Serge BRIANÇON - Jean-Claude BURDIN - Claude BURLET - Daniel BURNEL - Claude CHARDOT
Jean-François CHASSAGNE - François CHERRIER - Jean-Pierre CRANCE - Emile de LAVERGNE - Jean-Pierre DESCHAMPS
Jean DUHEILLE - Jean-Bernard DUREUX - Gilbert FAURE - Gérard FIEVE - Bernard FOLIGUET Jean FLOQUET - Robert FRISCH - Alain GAUCHER - Pierre GAUCHER - Jean-Luc GEORGE - Alain GERARD - Hubert GERARD
Jean-Marie GILGENKRANTZ - Simone GILGENKRANTZ -Gilles GROSDIDIER - Philippe HARTEMANN Gérard HUBERT - Claude HURIET - Christian JANOT - Michèle KESSLER - François KOHLER - Jacques LACOSTE - Henri LAMBERT
Pierre LANDES - Marie-Claire LAXENAIRE - Michel LAXENAIRE - Alain LE FAOU - Jacques LECLERE - Pierre LEDERLIN Bernard LEGRAS - Jean-Pierre MALLIÉ - Philippe MANGIN – Jean-Claude MARCHAL – Yves MARTINET Pierre MATHIEU Michel MERLE – Daniel MOLÉ - Pierre MONIN - Pierre NABET – Patrick NETTER - Jean-Pierre NICOLAS - Pierre PAYSANT Francis PENIN - Gilbert PERCEBOIS Claude PERRIN - Luc PICARD - François PLENAT - Jean-Marie POLU
Jacques POUREL - Jean PREVOT - Francis RAPHAEL - Antoine RASPILLER – Denis REGENT - Jacques ROLAND
Daniel SCHMITT - Michel SCHMITT - Michel SCHWEITZER - Daniel SIBERTIN-BLANC - Claude SIMON - Danièle SOMMELET Jean-François STOLTZ -Michel STRICKER - Gilbert THIBAUT - Gérard VAILLANT - Paul VERT - Hervé VESPIGNANI
Colette VIDAILHET - Michel VIDAILHET - Jean-Pierre VILLEMOT - Michel WEBER – Denis ZMIROU

==========

• PROFESSEURS ÉMÉRITES

Etienne ALIOT - Pierre BEY - Henry COUDANE - Serge BRIANÇON - Jean-Pierre CRANCE - Gilbert FAURE
Bernard FOLIGUET - Jean-Marie GILGENKRANTZ - Simone GILGENKRANTZ - Michèle KESSLER - François KOHLER
Alain LE FAOU - Jacques LECLERE - Yves MARTINET - Patrick NETTER - Jean-Pierre NICOLAS - Luc PICARD
François PLENAT - Jean-Pierre VILLEMOT

==========

PROFESSEURS DES UNIVERSITÉS - PRATICIENS HOSPITALIERS

(Disciplines du Conseil National des Universités)

42e Section : MORPHOLOGIE ET MORPHOGENÈSE

1re sous-section : ***(Anatomie)***
Professeur Marc BRAUN – Professeure Manuela PEREZ
2e sous-section : ***(Histologie, embryologie et cytogénétique)***
Professeur Christo CHRISTOV
3e sous-section : ***(Anatomie et cytologie pathologiques)***
Professeur Jean-Michel VIGNAUD – Professeur Guillaume GAUCHOTTE

43e Section : BIOPHYSIQUE ET IMAGERIE MÉDICALE

1re sous-section : ***(Biophysique et médecine nucléaire)***
Professeur Gilles KARCHER – Professeur Pierre-Yves MARIE – Professeur Pierre OLIVIER
2e sous-section : ***(Radiologie et imagerie médicale)***
Professeur René ANXIONNAT - Professeur Alain BLUM - Professeur Serge BRACARD - Professeure Valérie CROISÉ - Professeur Jacques FELBLINGER – Professeur Damien MANDRY - Professeur Pedro GONDIM TEIXEIRA

44e Section : BIOCHIMIE, BIOLOGIE CELLULAIRE ET MOLÉCULAIRE, PHYSIOLOGIE ET NUTRITION

1re sous-section : ***(Biochimie et biologie moléculaire)***
Professeur Jean-Louis GUEANT - Professeur Bernard NAMOUR - Professeur Jean-Luc OLIVIER
2e sous-section : ***(Physiologie)***
Professeur Christian BEYAERT - Professeur Bruno CHENUEL - Professeur François MARCHAL
3e sous-section ***(Biologie cellulaire)***
Professeure Véronique DECOT-MAILLERET
4e sous-section : ***(Nutrition)***
Professeur Didier QUILLIOT - Professeure Rosa-Maria RODRIGUEZ-GUEANT - Professeur Olivier ZIEGLER

45e Section : MICROBIOLOGIE, MALADIES TRANSMISSIBLES ET HYGIÈNE

1re sous-section : ***(Bactériologie – virologie ; hygiène hospitalière)***
Professeur Alain LOZNIEWSKI – Professeure Evelyne SCHVOERER
2e sous-section : ***(Parasitologie et Mycologie)***
Professeure Marie MACHOUART
3e sous-section : ***(Maladies infectieuses ; maladies tropicales)***
Professeur Bruno HOEN - Professeur Thierry MAY - Professeure Céline PULCINI - Professeur Christian RABAUD

46e Section : SANTÉ PUBLIQUE, ENVIRONNEMENT ET SOCIÉTÉ

1re sous-section : ***(Épidémiologie, économie de la santé et prévention)***
Professeure Nelly AGRINIER - Professeur Francis GUILLEMIN
4e sous-section : ***(Biostatistiques, informatique médicale et technologies de communication)***
Professeure Eliane ALBUISSON - Professeur Nicolas JAY

47e Section : CANCÉROLOGIE, GÉNÉTIQUE, HÉMATOLOGIE, IMMUNOLOGIE

1re sous-section : ***(Hématologie ; transfusion)***
Professeur Pierre FEUGIER
2e sous-section : ***(Cancérologie ; radiothérapie)***
Professeur Thierry CONROY - Professeur François GUILLEMIN - Professeur Didier PEIFFERT - Professeur Frédéric MARCHAL
3e sous-section : ***(Immunologie)***
Professeur Marcelo DE CARVALHO-BITTENCOURT - Professeure Marie-Thérèse RUBIO
4e sous-section : ***(Génétique)***
Professeur Philippe JONVEAUX - Professeur Bruno LEHEUP

48e Section : ANESTHÉSIOLOGIE, RÉANIMATION, MÉDECINE D'URGENCE, PHARMACOLOGIE ET THÉRAPEUTIQUE

1re sous-section : ***(Anesthésiologie-réanimation et médecine péri-opératoire)***
Professeur Gérard AUDIBERT - Professeur Hervé BOUAZIZ - Professeur Thomas FUCHS-BUDER
Professeure Marie-Reine LOSSER - Professeur Claude MEISTELMAN
2e sous-section : ***(Médecine intensive-réanimation)***
Professeur Pierre-Édouard BOLLAERT - Professeur Sébastien GIBOT - Professeur Bruno LÉVY
3e sous-section : ***(Pharmacologie fondamentale ; pharmacologie clinique ; addictologie)***
Professeur Pierre GILLET - Professeur Jean-Yves JOUZEAU
4e sous-section : ***(Thérapeutique-médecine de la douleur ; addictologie)***
Professeur Nicolas GIRERD - Professeur François PAILLE - Professeur Patrick ROSSIGNOL - Professeur Faiez ZANNAD

49e Section : PATHOLOGIE NERVEUSE ET MUSCULAIRE, PATHOLOGIE MENTALE, HANDICAP ET RÉÉDUCATION

1re sous-section : ***(Neurologie)***
Professeur Marc DEBOUVERIE - Professeur Louis MAILLARD - Professeur Sébastien RICHARD - Professeur Luc TAILLANDIER
Professeure Louise TYVAERT

2e sous-section : ***(Neurochirurgie)***
Professeur Jean AUQUE - Professeur Thierry CIVIT - Professeure Sophie COLNAT-COULBOIS - Professeur Olivier KLEIN

3e sous-section : ***(Psychiatrie d'adultes ; addictologie)***
Professeur Jean-Pierre KAHN – Professeur Vincent LAPREVOTE - Professeur Raymund SCHWAN

4e sous-section : ***(Pédopsychiatrie ; addictologie)***
Professeur Bernard KABUTH

5e sous-section : ***(Médecine physique et de réadaptation)***
Professeur Jean PAYSANT

50e Section : PATHOLOGIE OSTÉO-ARTICULAIRE, DERMATOLOGIE ET CHIRURGIE PLASTIQUE

1re sous-section : ***(Rhumatologie)***
Professeure Isabelle CHARY-VALCKENAERE - Professeur Damien LOEUILLE

2e sous-section : ***(Chirurgie orthopédique et traumatologique)***
Professeur Laurent GALOIS - Professeur Didier MAINARD - Professeur François SIRVEAUX

3e sous-section : ***(Dermato-vénéréologie)***
Professeure Anne-Claire BURSZTEJN - Professeur Jean-Luc SCHMUTZ

4e sous-section : ***(Chirurgie plastique, reconstructrice et esthétique ; brûlologie)***
Professeur François DAP - Professeur Gilles DAUTEL - Professeur Etienne SIMON

51e Section : PATHOLOGIE CARDIO-RESPIRATOIRE ET VASCULAIRE

1re sous-section : ***(Pneumologie ; addictologie)***
Professeur Jean-François CHABOT - Professeur Ari CHAOUAT

2e sous-section : ***(Cardiologie)***
Professeur Edoardo CAMENZIND - Professeur Christian de CHILLOU DE CHURET - Professeur Yves JUILLIERE
Professeur Nicolas SADOUL

3e sous-section : ***(Chirurgie thoracique et cardiovasculaire)***
Professeur Juan-Pablo MAUREIRA

4e sous-section : ***(Chirurgie vasculaire ; médecine vasculaire)***
Professeur Sergueï MALIKOV - Professeur Denis WAHL – Professeur Stéphane ZUILY

52e Section : MALADIES DES APPAREILS DIGESTIF ET URINAIRE

1re sous-section : ***(Gastroentérologie ; hépatologie ; addictologie)***
Professeur Jean-Pierre BRONOWICKI - Professeur Laurent PEYRIN-BIROULET

3e sous-section : ***(Néphrologie)***
Professeur Luc FRIMAT - Professeure Dominique HESTIN

4e sous-section : ***(Urologie)***
Professeur Pascal ESCHWEGE - Professeur Jacques HUBERT

53e Section : MÉDECINE INTERNE, GÉRIATRIE, CHIRURGIE GÉNÉRALE ET MÉDECINE GÉNÉRALE

1re sous-section : ***(Médecine interne ; gériatrie et biologie du vieillissement ; addictologie)***
Professeur Athanase BENETOS - Professeur Jean-Dominique DE KORWIN - Professeure Gisèle KANNY
Professeure Christine PERRET-GUILLAUME – Professeur Roland JAUSSAUD – Professeure Laure JOLY

2e sous-section : ***(Chirurgie générale)***
Professeur Ahmet AYAV - Professeur Laurent BRESLER - Professeur Laurent BRUNAUD – Professeure Adeline GERMAIN

3e sous-section : ***(Médecine générale)***
Professeur Jean-Marc BOIVIN – Professeur Paolo DI PATRIZIO

54e Section : DÉVELOPPEMENT ET PATHOLOGIE DE L'ENFANT, GYNÉCOLOGIE-OBSTÉTRIQUE, ENDOCRINOLOGIE ET REPRODUCTION

1re sous-section : ***(Pédiatrie)***
Professeur Pascal CHASTAGNER - Professeur François FEILLET - Professeur Jean-Michel HASCOET
Professeur Emmanuel RAFFO - Professeur Cyril SCHWEITZER

2e sous-section : ***(Chirurgie infantile)***
Professeur Pierre JOURNEAU - Professeur Jean-Louis LEMELLE

3e sous-section : ***(Gynécologie-obstétrique ; gynécologie médicale)***
Professeur Philippe JUDLIN - Professeur Olivier MOREL

4e sous-section : ***(Endocrinologie, diabète et maladies métaboliques ; gynécologie médicale)***
Professeur Bruno GUERCI - Professeur Marc KLEIN - Professeur Georges WERYHA

55e Section : PATHOLOGIE DE LA TÊTE ET DU COU
1re sous-section : ***(Oto-rhino-laryngologie)***
Professeur Roger JANKOWSKI - Professeure Cécile PARIETTI-WINKLER
2e sous-section : ***(Ophtalmologie)***
Professeure Karine ANGIOI - Professeur Jean-Paul BERROD
3e sous-section : ***(Chirurgie maxillo-faciale et stomatologie***
Professeure Muriel BRIX

==========

- **PROFESSEURS DES UNIVERSITÉS**

61e Section : GÉNIE INFORMATIQUE, AUTOMATIQUE ET TRAITEMENT DU SIGNAL
Professeur Walter BLONDEL

64e Section : BIOCHIMIE ET BIOLOGIE MOLÉCULAIRE
Professeure Sandrine BOSCHI-MULLER - Professeur Pascal REBOUL

65e Section : BIOLOGIE CELLULAIRE
Professeure Céline HUSELSTEIN

==========

PROFESSEUR ASSOCIÉ DE MÉDECINE GÉNÉRALE
Professeure associée Sophie SIEGRIST

==========

- **MAÎTRES DE CONFÉRENCES DES UNIVERSITÉS - PRATICIENS HOSPITALIERS**

42e Section : MORPHOLOGIE ET MORPHOGENÈSE
1re sous-section : ***(Anatomie)***
Docteur Bruno GRIGNON

43e Section : BIOPHYSIQUE ET IMAGERIE MÉDICALE
1re sous-section : ***(Biophysique et médecine nucléaire)***
Docteur Antoine VERGER

44e Section : BIOCHIMIE, BIOLOGIE CELLULAIRE ET MOLÉCULAIRE, PHYSIOLOGIE ET NUTRITION
1re sous-section : ***(Biochimie et biologie moléculaire)***
Docteure Shyue-Fang BATTAGLIA - Docteure Sophie FREMONT - Docteure Catherine MALAPLATE - Docteur Marc MERTEN - Docteur Abderrahim OUSSALAH
2e sous-section : ***(Physiologie)***
Docteure Silvia DEMOULIN-ALEXIKOVA - Docteur Mathias POUSSEL – Docteur Jacques JONAS

45e Section : MICROBIOLOGIE, MALADIES TRANSMISSIBLES ET HYGIÈNE
1re sous-section : ***(Bactériologie – Virologie ; hygiène hospitalière)***
Docteure Corentine ALAUZET - Docteure Hélène JEULIN - Docteure Véronique VENARD
2e sous-section : (Parasitologie et mycologie)
Docteure Anne DEBOURGOGNE

46e Section : SANTÉ PUBLIQUE, ENVIRONNEMENT ET SOCIÉTÉ
1re sous-section : ***(Epidémiologie, économie de la santé et prévention)***
Docteur Cédric BAUMANN - Docteure Frédérique CLAUDOT - Docteur Alexis HAUTEMANIÈRE
2e sous-section ***(Médecine et Santé au Travail)***
Docteure Isabelle THAON
3e sous-section ***(Médecine légale et droit de la santé)***
Docteur Laurent MARTRILLE

47e Section : CANCÉROLOGIE, GÉNÉTIQUE, HÉMATOLOGIE, IMMUNOLOGIE
1re sous-section : ***(Hématologie ; transfusion)***
Docteure Aurore PERROT – Docteur Julien BROSEUS – Docteure Maud D'AVENI (stagiaire)
2e sous-section : ***(Cancérologie ; radiothérapie)***
Docteure Lina BOLOTINE – Docteur Guillaume VOGIN
4e sous-section : ***(Génétique)***
Docteure Céline BONNET

48e Section : ANESTHÉSIOLOGIE, RÉANIMATION, MÉDECINE D'URGENCE, PHARMACOLOGIE ET THÉRAPEUTIQUE
1e sous-section : ***(Anesthésiologie-réanimation et médecine péri-opératoire)***
Docteur Philippe GUERCI (stagiaire)
2e sous-section : ***(Médecine intensive-réanimation)***
Docteur Antoine KIMMOUN
3e sous-section : ***(Pharmacologie fondamentale ; pharmacologie clinique ; addictologie)***
Docteur Nicolas GAMBIER - Docteure Françoise LAPICQUE - Docteur Julien SCALA-BERTOLA

50e Section : PATHOLOGIE OSTÉO-ARTICULAIRE, DERMATOLOGIE ET CHIRURGIE PLASTIQUE
1re sous-section : ***(Rhumatologie)***
Docteure Anne-Christine RAT
4e sous-section : ***(Chirurgie plastique, reconstructrice et esthétique ; brûlologie)***
Docteure Laetitia GOFFINET-PLEUTRET

51e Section : PATHOLOGIE CARDIO-RESPIRATOIRE ET VASCULAIRE
3e sous-section : ***(Chirurgie thoracique et cardio-vasculaire)***
Docteur Fabrice VANHUYSE
4e sous-section : ***(Chirurgie vasculaire ; Médecine vasculaire)***
Docteure Nicla SETTEMBRE (stagiaire)

52e Section : MALADIES DES APPAREILS DIGESTIF ET URINAIRE
1re sous-section : ***(Gastroentérologie ; hépatologie ; addictologie)***
Docteur Jean-Baptiste CHEVAUX – Docteur Anthony LOPEZ

53e Section : MÉDECINE INTERNE, GÉRIATRIE, CHIRURGIE GÉNÉRALE ET MÉDECINE GÉNÉRALE
2e sous-section : (Chirurgie générale)
Docteur Cyril PERRENOT

54e Section : DEVELOPPEMENT ET PATHOLOGIE DE L'ENFANT, GYNECOLOGIE-OBSTETRIQUE, ENDOCRINOLOGIE ET REPRODUCTION
4e sous-section : ***(Endocrinologie, diabète et maladies métaboliques ; Gynécologie médicale)***
Docteure Eva FEIGERLOVA (stagiaire)
5e sous-section : ***(Biologie et médecine du développement et de la reproduction ; gynécologie médicale)***
Docteure Isabelle KOSCINSKI

55e Section : PATHOLOGIE DE LA TÊTE ET DU COU
1re sous-section : ***(Oto-Rhino-Laryngologie)***
Docteur Patrice GALLET

==========

- **MAÎTRES DE CONFÉRENCES**

5e Section : SCIENCES ÉCONOMIQUES
Monsieur Vincent LHUILLIER

7e Section : SCIENCES DU LANGAGE : LINGUISTIQUE ET PHONETIQUE GENERALES
Madame Christine DA SILVA-GENEST

19e Section : SOCIOLOGIE, DÉMOGRAPHIE
Madame Joëlle KIVITS

64e Section : BIOCHIMIE ET BIOLOGIE MOLÉCULAIRE
Madame Marie-Claire LANHERS - Monsieur Nick RAMALANJAONA

65e Section : BIOLOGIE CELLULAIRE
Madame Nathalie AUCHET - Madame Natalia DE ISLA-MARTINEZ - Madame Ketsia HESS Monsieur Christophe NEMOS

66e Section : PHYSIOLOGIE
Monsieur Nguyen TRAN

69e Section : NEUROSCIENCES
Madame Sylvie MULTON

==========

- **MAÎTRES DE CONFÉRENCES ASSOCIÉS DE MÉDECINE GÉNÉRALE**

Docteur Cédric BERBE - Docteur Olivier BOUCHY - Docteur Jean-Michel MARTY

==========

- **DOCTEURS HONORIS CAUSA**

Professeur Charles A. BERRY (1982)
Centre de Médecine Préventive, Houston (U.S.A)
Professeur Pierre-Marie GALETTI (1982)
Brown University, Providence (U.S.A)
Professeure Mildred T. STAHLMAN (1982)
Vanderbilt University, Nashville (U.S.A)
Professeur Théodore H. SCHIEBLER (1989)
Institut d'Anatomie de Würtzburg (R.F.A)
Université de Pennsylvanie (U.S.A)
Professeur Mashaki KASHIWARA (1996)
Research Institute for Mathematical Sciences de Kyoto (JAPON)

Professeure Maria DELIVORIA-PAPADOPOULOS (1996)
Professeur Ralph GRÄSBECK (1996)
Université d'Helsinki (FINLANDE)
Professeur Duong Quang TRUNG (1997)
Université d'Hô Chi Minh-Ville (VIÊTNAM)
Professeur Daniel G. BICHET (2001)
Université de Montréal (Canada)
Professeur Marc LEVENSTON (2005)
Institute of Technology, Atlanta (USA)

Professeur Brian BURCHELL (2007)
Université de Dundee (Royaume-Uni)
Professeur Yunfeng ZHOU (2009)
Université de Wuhan (CHINE)
Professeur David ALPERS (2011)
Université de Washington (U.S.A)
Professeur Martin EXNER (2012)
Université de Bonn (ALLEMAGNE)

REMERCIEMENTS

A MONSIEUR LE PRESIDENT DU JURY

Monsieur le Professeur Guillaume GAUCHOTTE

Professeur des Universités - Praticien Hospitalier d'Anatomie et Cytologie Pathologiques.

Nous vous remercions de nous avoir fait l'honneur de présider ce jury. Votre présence représente pour nous l'opportunité de voir notre travail jugé du point de vue du pathologiste. Merci pour votre accessibilité, votre bienveillance, et votre sympathie au cours de mon cursus. Veuillez-trouver ici l'assurance de mon plus profond respect.

A MESIEURS ET MADAME LES MEMBRES DU JURY

A monsieur le Docteur Guillaume VOGIN

Maître de Conférences des Universités - Praticien Hospitalier de Radiothérapie.

Nous vous remercions de nous avoir fait l'honneur d'accepter de faire partie de ce jury. Votre présence représente pour nous l'opportunité de voir notre travail jugé du point de vue du clinicien. Veuillez trouver ici l'expression de nos sincères remerciements et de notre profond respect.

A madame le Docteur Céline BONNET

Maître de Conférences des Universités - Praticien Hospitalier de Génétique.

Nous vous remercions de nous avoir fait l'honneur d'accepter de faire partie de ce jury. Le semestre passé au laboratoire de génétique il y a deux ans nous a permis d'approfondir et d'élargir nos connaissances en génétique somatique et constitutionnelle, il nous a permis de mieux appréhender les obstacles rencontrés au cours de ce travail. Nous voudrions vous témoigner notre gratitude pour votre professionnalisme, votre sympathie et votre bienveillance.

A monsieur le Professeur Jean-Louis MERLIN,

Professeur des Universités à la Faculté de Pharmacie de Nancy – Pharmacien praticien hospitalier à l'Institut de Cancérologie de Lorraine.

Nous vous remercions de votre accueil, il y a 3ans, lors de ce semestre à l'Unité de Biologie des Tumeurs, nous avons pu y trouver les conditions nécessaires pour que l'idée de ce projet puisse germer. Votre présence représente pour nous l'opportunité de voir notre travail jugé du point de vue du spécialiste en biologie moléculaire des cancers. Nous vous remercions pour votre sympathie et votre bonne humeur au cours de nos rencontres à l'ICL durant les 3 semestres que nous avons passé dans le service de biopathologie.

A notre directeur de thèse, monsieur le Docteur Alexandre HARLE

Maître de conférences à la Faculté de Pharmacie de Nancy – Pharmacien praticien hospitalier à l'Institut de Cancérologie de Lorraine.

Nous vous remercions d'avoir accepté de nous encadrer pour ce projet. Merci de votre orientation, vos conseils et de votre aide apportée au cours de ce travail et dans la correction de ce manuscrit. Malgré la distance, nos efforts ont abouti à ce travail de thèse. Nous voudrions également vous témoigner notre gratitude pour votre confiance qui nous a été précieuse afin de mener notre travail à bon port.

AUTRES REMERCIEMENTS

A ceux qui ont contribué à ce travail :

-Merci à Julia SALLERON pour sa méthodologie, son expertise en biostatistiques, ses conseils, sa sympathie et son travail.

-Monsieur le Docteur Romain BOIDOT pour son aide, son important travail de recueil de données et sa réactivité dans nos échanges au cours de ces 2 dernières années.

A l'ensemble des équipes des laboratoires d'Anatomie et Cytologie Pathologiques de Brabois, de Central, de Nouméa, de l'Institut de Cancérologie de Lorraine, de l'Unité de Biologie des Tumeurs et du laboratoire de Génétique Médicale du CHU de NANCY,

Merci à tous les personnels qui ont participé à ma formation et/ou qui ont rendu mon quotidien agréable. Merci pour votre professionnalisme, votre gentillesse et votre bienveillance. Veuillez-trouver ici l'expression de ma sincère gratitude.

A l'ensemble des personnes qui ont participé à ma formation depuis l'école primaire,

Qu'ils soient remerciés pour la qualité de leurs conseils, de leurs formations, de leurs enseignements et qu'ils trouvent ici l'expression de mes respectueux sentiments.

Je voudrais remercier en particulier,

-Mr Jean-Claude PERRIN qui fût mon professeur de physique-chimie au lycée J-V PONCELET (SAINT-AVOLD), et qui m'a fait aimer les sciences par sa pédagogie, sa sympathie et son humour.

-Mr le Professeur Denis REGENT, pour ces séances bi-hebdomadaires de cours auxquels j'ai assisté assidûment de janvier à mai 2014, qui m'ont permis d'appréhender la médecine et le diagnostic médicale en particulier d'une manière nouvelle. Vous avez fait germer en moi la passion du diagnostic bien fait et l'envie d'en connaître toujours plus sur les maladies, ce qui a orienté à 180° mon choix de carrière vers l'anatomie pathologique. Pour l'anecdote, le

lendemain du concours national (ECN 2014), soit le jeudi 1 mai 2014, je retournais au CHU pour assister à vos cours si passionnants, en oubliant même que ce jour était férié.

-Mr le Professeur Bernard FOLIGUET, merci pour le temps que vous m'avez accordé au microscope au cours de mon stage d'externe, vous avez démystifié la spécialité par votre pédagogie et votre sympathie.

-Mr le Docteur Xavier SASTRE-GARAU, merci pour vos conseils, votre simplicité au regard de votre immense carrière, votre pragmatisme, votre pédagogie et votre sympathie au cours des 3 semestres durant lesquels j'ai eu la chance et l'honneur de travailler avec vous. Vous aviez raison, c'est en se lançant dans son propre projet scientifique, et donc en étant confronté à tous les obstacles, que l'on est le plus à même de progresser. Je suis certain que sans avoir croisé votre route, j'aurai sagement attendu que l'on me propose un sujet de thèse en fin d'internat, et ce travail n'aurait pas vu le jour. Vous êtes un modèle pour moi dans bien des domaines.

À tous les médecins qui ont été pour moi une source d'inspiration,

Les docteurs Agnès LEROUX, Jacqueline CHAMPIGNEULLE, Didier MONCHY, Claude DEPARDIEU, et Mrs les Professeurs Jean Michel VIGNAUD et Philippe JONVEAUX.

Merci à ceux qui ont contribué à ma formation et que j'ai pu côtoyer au cours de mon cursus, chacun m'a appris quelque chose et m'a fait réfléchir d'une certaine manière,

Les Docteurs Claire BASTIEN, Charlène VIGOUROUX, Hélène BUSBY-VENER, Jean Matthieu CASSE, Léa LEGRAND, Antoine MAX, Charlie PIERRE, Clémence YGUEL, Eliovel CABRERA, Jacques THOMAS, Philippe LEROUX, Marc MULLER, Sarab LIZARD, Anne-Sophie DENOMMÉ-PICHON, Pauline GILSON, Béatrice MARIE, Catherine BARLIER.

A mes amis

Parfois notre lumière s'éteint, puis elle est rallumée par un autre être humain. Je remercie sincèrement ceux qui ont ravivé ma flamme au cours de ma vie.

A ma famille, mes autres amis, mes co-internes,

J'ai une pensée pour chacun d'entre vous, vous connaissez chacun ma pensée. Les bons moments restent gravés !

A mes 4 parents

Merci pour votre amour, votre soutien sans faille dans les étapes de ma vie, vos sacrifices sans lesquels aujourd'hui je ne pourrais pas voler de mes propres ailes et ce malgré la distance qui parfois nous a séparé. Je ne vous remercierai jamais assez. Je vous aime.

A celle que j'aime,

Tu m'accompagnes, m'écoutes et me conseilles depuis le début de ce travail. Merci de m'avoir montré d'autres horizons pour m'évader. Merci pour ton soutien infaillible et ces incroyables moments de complicité et de délire au quotidien. Tu es un modèle de valeurs humaines. Toi et ta maman vous pouvez être fiers de vous. Merci d'exister.

SERMENT

« Au moment d'être admis à exercer la médecine, je promets et je jure d'être fidèle aux lois de l'honneur et de la probité. Mon premier souci sera de rétablir, de préserver ou de promouvoir la santé dans tous ses éléments, physiques et mentaux, individuels et sociaux. Je respecterai toutes les personnes, leur autonomie et leur volonté, sans aucune discrimination selon leur état ou leurs convictions. J'interviendrai pour les protéger si elles sont affaiblies, vulnérables ou menacées dans leur intégrité ou leur dignité. Même sous la contrainte, je ne ferai pas usage de mes connaissances contre les lois de l'humanité. J'informerai les patients des décisions envisagées, de leurs raisons et de leurs conséquences. Je ne tromperai jamais leur confiance et n'exploiterai pas le pouvoir hérité des circonstances pour forcer les consciences. Je donnerai mes soins à l'indigent et à quiconque me les demandera. Je ne me laisserai pas influencer par la soif du gain ou la recherche de la gloire.

Admis dans l'intimité des personnes, je tairai les secrets qui me sont confiés. Reçu à l'intérieur des maisons, je respecterai les secrets des foyers et ma conduite ne servira pas à corrompre les mœurs. Je ferai tout pour soulager les souffrances. Je ne prolongerai pas abusivement les agonies. Je ne provoquerai jamais la mort délibérément.

Je préserverai l'indépendance nécessaire à l'accomplissement de ma mission. Je n'entreprendrai rien qui dépasse mes compétences. Je les entretiendrai et les perfectionnerai pour assurer au mieux les services qui me seront demandés. J'apporterai mon aide à mes confrères ainsi qu'à leurs familles dans l'adversité.

Que les hommes et mes confrères m'accordent leur estime si je suis fidèle à mes promesses ; que je sois déshonoré et méprisé si j'y manque ».

TABLE DES MATIERES

LISTE DES ABBREVIATIONS

A : Adénine

ADN : Acide désoxyribonucléique

ARN : Acide ribonucléique

ARNm : Acide ribonucléique messager

BaP : Benzo(a)pyrene

BPDE : Benzopyrène-diol-époxyde

C : Cytosine

CAPI : Carcinome de primitif inconnu

CD45 : Cluster de différenciation 45

CD56 : Cluster de différenciation 56

CK20 : Cytokératine 20

CK5/6: Cytokératine 5/6

CK7: Cytokératine 7

ctDNA: ADN tumoral circulant

CUP: Carcinoma of unknown primitive

CYP1A1: Cytochrome P450 1A1

DOG1: Delay of germination 1

EASILUNG-score: Exome And Signature LUNG score

EASILUNG-test: Exome And Signature LUNG test

EBV: Epstein Barr Virus

EGFR: Epidermal growth factor

EMA: Epithelial membrane antigen

ERBB2: Erb-B2 Receptor Tyrosine Kinase 2

ERG: ETS-related gene

FFPE: Formalin-fixed paraffin-embedded

FISH: Fluorescent in situ hybridization

G: Guanine

GATA-3: GATA binding protein 3

GCDFP-15: GATA binding protein 3

HAP: Hydrocarbure aromatique polycyclique

HepPar-1: Hepatocyte paraffin 1

HGVS: Human Genome Variation Society

HHV-8: Herpèsvirus humain type 8

HPV: Human papillomavirus

KRAS: Kirsten rat sarcoma viral oncogene homolog

NGS: Next generation sequencing

NNK: Nicotine-derived nitrosamine ketone

PAX-8: Paired box gene 8

PCR: Polymerase chain reaction

Polη : Polymérase êta

PSA : Prostate specific antigen

RO : Récepteur d'oestrogène

ROS : Reactive oxygen species

RP : Récepteur de progestérone

RT-PCR: Reverse transcriptase polymerase chain reaction

SALL4: Spalt like transcription factor 4

SATB2: SpecialAT-rich sequence-binding protein

T: Thymine

TCGA: The Cancer Genome Atlas

TP53: Tumor protein 53

TTF-1 : Thyroid transcription factor-1

WT1: Wilms tumor 1

AVANT-PROPOS

Les objectifs de ce travail étaient, d'une part, de montrer que les fréquences relatives des substitutions somatiques d'un résultat de séquençage exomique d'une tumeur permettent à elles seules de discriminer les cancers pulmonaires des autres cancers, et d'autre part d'établir un test diagnostique (EASILUNG-test) comportant un score validé (EASILUNG-score) permettant à partir des substitutions somatiques d'un résultat d'exome d'une tumeur donnée, et en aveugle de tous les autres paramètres, d'estimer s'il s'agit ou non d'un cancer pulmonaire. C'est un angle d'approche diagnostique inédit que nous proposons, ouvrant sur de nombreuses perspectives de recherches, mais aussi sur une application diagnostique concrète pour les patients atteints de cancer de primitif inconnu et chez qui le diagnostic de cancer pulmonaire et toujours en considération.

❖ Généralités sur la cancérogénèse moléculaire

Les mutations à l'origine des cancers sont soit transmises (mutations constitutionnelles) soit induites (mutations somatiques) par des facteurs environnementaux, mais elles peuvent également survenir suite à des erreurs lors de la réplication de l'ADN. Il a été montré qu'environ trois mutations se produisent à chaque division de cellules souches et qu'il existe une corrélation entre le nombre de divisions des cellules souches et l'incidence des cancers, indépendamment de l'environnement; ceci explique le rôle majeur du vieillissement en tant que facteur de risque de cancer[1]. Les agents mutagènes peuvent être endogènes (espèces réactives de l'oxygène, hormones...) ; ils sont alors appelés « facteurs de l'hôte ». Les mutagènes peuvent aussi être exogènes, il existe 3 catégories d'agents mutagènes exogènes[2], comprenant :

- les carcinogènes physiques, tels que les rayonnements ultraviolets et ionisants ;

- les carcinogènes chimiques, tels que l'amiante, les composants de la fumée de tabac, l'aflatoxine (un contaminant alimentaire) et l'arsenic (un contaminant de l'eau potable) ;

- les carcinogènes biologiques, tels que les infections par certains virus (ex : Human Papillomavirus, Epstein Barr Virus, HHV-8, …), bactéries ou parasites.

Dans les tumeurs solides, on observe en moyenne 33 à 66 gènes mutés qui vont donner une protéine dont la fonction est altérée[3]. Cette moyenne est augmentée (environ 200 mutations) dans le cas de cancers se développant du fait d'une importante exposition à des mutagènes exogènes physiques tels que les rayons ultraviolets (ex : mélanomes cutanés, carcinomes épidermoïdes cutanés, …) ou les carcinogènes exogènes chimiques du tabac (cancers du poumon). Les mutations somatiques identifiées dans les cancers sont principalement des substitutions de simple base (95%), mais on trouve également des petites délétions ou insertions d'une à quelques bases, des amplifications, des grandes délétions et des réarrangements chromosomiques.

- **Substitutions : transition et transversion**

Les bases azotées (Adénine, Guanine, Cytosine et Thymine) sont classées en 2 catégories : les bases puriques et les bases pyrimidiques. L'Adénine et la Guanine sont des purines, composées de deux cycles aromatiques. La Thymine et la Cytosine sont des pyrimidines composées d'un seul cycle. Lorsqu'une mutation d'une base purique vers une base pyrimidique se produit, on parle de transversion (ou alors lorsqu'une base pyrimidique vers une base purique). Lorsqu'une mutation d'une base purique vers une autre base purique se produit, on parle de transition (ou alors lorsqu'une base pyrimidique qui devient une autre base pyrimidique).

- **Nomenclature HGVS (Human Genome Variation Society) des substitutions**

Une substitution est une modification de séquence nucléotidique (en comparaison à la séquence de référence) dans laquelle un nucléotide est remplacé par un autre nucléotide.

Format : "nucléotide de référence" " >" "nouveau nucléotide"

Ex : G>T

"G"= "nucléotide de référence" = nucléotide initiale

">" = le type de changement est une substitution

"T" = "nouveau nucléotide" = nucléotide substitué

Selon la nomenclature internationale HGVS, une substitution s'écrit donc sous la forme "nucléotide de référence" ">" " nouveau nucléotide". Par exemple la substitution G>T signifie que la base de référence, la Guanine est remplacée par la Thymine. Si nous résumons les différents cas de figures possibles (Figure 1), il y a au total :

4 transitions : A>G, G>A, C>T, T>C

8 transversions : A>C, C>A, G>T, T>G, G>C, C>G, A>T, T>A

Rappelons que l'ADN est double brin avec une complémentarité des bases. L'Adénine est toujours en face d'une Thymine et la Guanine toujours en face d'une Cytosine. S'il y a une mutation sur un brin, par exemple un G>T, alors il y a sur le brin complémentaire un C>A. Ces deux notations sont souvent confondues sous la forme G>T/C>A pour une position de substitutions sur l'ADN sans que le brin ne soit précisé, si le brin est précisé (comme dans notre étude ou les substitutions sont situées sur le brin codant non-transcrit), on utilise plutôt G>T ou C>A.

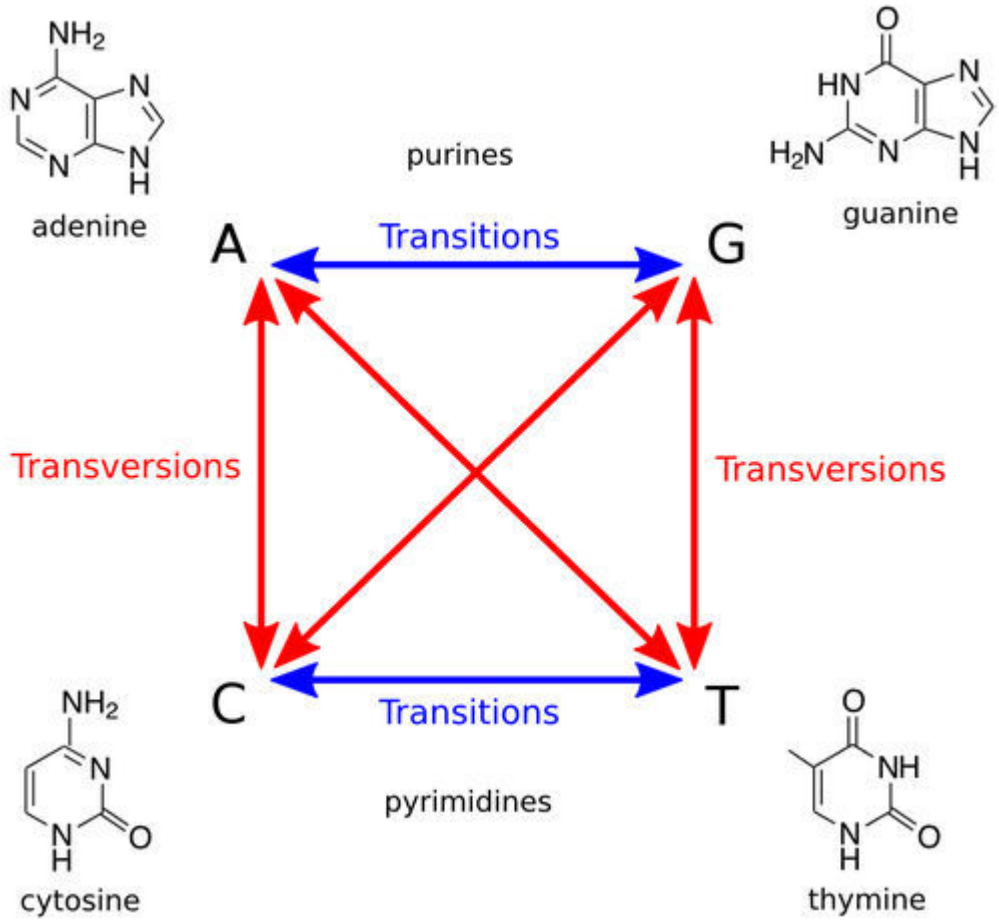


Figure1. Les différentes possibilités de substitution [4].

❖ Généralités sur les cancers broncho-pulmonaires

Les carcinomes à petites cellules (CPC) représentent 15-20 % des cas et les carcinomes bronchiques non à petites cellules (CBNPC) représente 80-85% des cas. Pour les CBNPC, le type histologique le plus fréquent est l'adénocarcinome (40-50%) [5]. Les principales anomalies moléculaires rencontrées sont les mutations de KRAS, de l'EGFR, et la translocation EML4-ALK, elles sont plus fréquentes dans les adénocarcinomes. D'autres anomalies

moléculaires, comme les mutations BRAF, de ERBB2 (HER2) ou de PI3K, les translocations de ROS1 ou KIF5B-RET sont également retrouvées. A l'inverse des mutations de l'EGFR, classiquement rencontré chez la femme asiatique et chez les non-fumeurs, les mutations de KRAS sont fortement associées au tabagisme. Le pronostic du cancer du poumon est sombre avec une survie tout stade confondu à 5 ans estimée à environ 15%, car une majorité de cancers broncho-pulmonaires (70%) sont diagnostiqués à un stade avancé ou métastatique. En 2018, il n'y a pas de dépistage organisé de ce type de cancer en France. Le plus souvent, le cancer broncho-pulmonaire est évoqué chez un patient avec une histoire d'exposition au tabagisme devant des symptômes respiratoires comme une hémoptysie, une toux, ou une dyspnée, voir d'autres signes moins spécifiques tels qu'une altération de l'état général, une maladie thromboembolique ou des signes d'atteintes métastatique (hypercalcémie, défaillance d'organe). La confirmation diagnostique passe par la réalisation de biopsies, de prélèvements cytologiques, ou par l'exérèse chirurgicale du foyer tumoral primitif ou d'une localisation métastatique avec examen anatomopathologique, qui permet, outre le diagnostic, d'apporter des informations à visées théranostiques tels que l'expression de biomarqueurs corrélées à la réponse à l'immunothérapie comme le PD-L1. Dans la carcinogenèse pulmonaire, le tabac joue un rôle mutagène prédominant, ainsi la fraction attribuable du tabagisme actif ou passif est de l'ordre de 80%.

❖ Mutagenèse induite par le tabac

- **Agents alkylants et agents oxydants de la fumée de cigarette**

Des cancérogènes comme les hydrocarbures aromatiques polycycliques (HAP), les aza-arènes, les nitrosamines spécifiques du tabac, par exemple le 4- (méthylnitrosamino) -1- (3-pyridyl) -1-butanone (également appelée NNK pour « nicotine-derived nitrosamine ketone »), le 1,3-butadiène, le carbamate d'éthyle, l'oxyde d'éthylène, le nickel, le chrome, le cadmium, le polonium-210, l'arsenic et l'hydrazine sont présents dans la fumée de cigarette. L'étude de ces composés identifiés dans la fumée de cigarette a démontré qu'ils pouvaient induire des tumeurs pulmonaires chez au moins une espèce animale. Les N-nitrosamines spécifiques de la combustion du tabac sont un groupe de cancérogènes dérivés des alcaloïdes et sont liés au développement de cancers du poumon, de l'œsophage, du pancréas et de la cavité buccale chez les personnes exposées à la fumée de cigarette. La formation d'adduits à l'ADN à partir de NNK a été étudiée de façon approfondie ; les NNK induisent également des cassures simple brin et

augmente les niveaux de 8-oxodésoxyguanosine dans l'ADN [6,7]. La NNK est la N-nitrosamine la plus abondante et la plus cancérogène issue de la combustion du tabac. L'une de ses principales caractéristiques est sa haute spécificité pour les poumons. La NNK induit systématiquement des tumeurs pulmonaires chez le rat, la souris et le hamster, quelle que soit la méthode d'administration : nourriture, eau potable, voie sous-cutanée ou intrapéritonéale[8].

Parmi les HAP, le benzo [a] pyrène (BaP) est le composé le plus étudié et sa capacité à induire des tumeurs pulmonaires est établie[9]. Des études *in vitro* ont démontré que les adduits à l'ADN formés à partir du métabolisme des BaP, des NNK et des espèces réactives oxygénées (ROS pour *reactive oxygen* species) peuvent tous conduire à des transversions G>T/C>A [10,11]. Ainsi, les transversions G>T/C>A du gène *TP53* sont très fréquentes dans les tumeurs pulmonaires des fumeurs, moins fréquentes chez les ex-fumeurs et rares chez les personnes n'ayant jamais fumé, ce type de transversion est très spécifique des cancers du poumon associés au tabagisme[12,13]. Par ailleurs, les particules fines de goudron sont déposées dans les alvéoles pulmonaires et entrent en contact avec les fluides pulmonaires qui en extraient les composants hydrosolubles. Bien que les radicaux de goudron de la cigarette ne se lient pas directement à l'ADN, le système semi quinone réduit l'oxygène en O2-, conduisant à la formation de radicaux hydroxyles extrêmement réactifs (HO•). Ces réactions peuvent avoir lieu à l'intérieur du noyau cellulaire et les radicaux HO• peuvent attaquer l'ADN pour produire un grand nombre de substitutions, de cassures de brins ou encore d'autres dommages oxydatifs dans ces cellules comme les adduits mutagènes 8-OHdG ou les cassures d'ADN simple brin[14,15]. Deux réactions d'oxydation sont fréquentes. La première implique la désoxyguanosine (dG), qui est oxydée en 8-oxo-désoxyguanosine (8-oxo-dG), facilement mésappariée avec une désoxyadénosine, entraînant une substitution G>T/C>A. La seconde implique la désoxy-5-osine (d5mC), le nucléotide présent dans les séquences CpG méthylées. Lors de l'oxydation, cette dernière forme initialement une base instable qui se désamine rapidement, produisant de la désoxythymidine glycol (dTg), il en résulte une substitution C>T/G>A.

- **Rôle du CYP1A1 dans la signature mutationnelle des cancers pulmonaires**

Après deux réactions d'oxydation successives induites par les enzymes du cytochrome P450 (CYP1A1 en grande partie), le BaP est converti en benzo [a] pyrènediolépoxyde (BPDE). Cette molécule hautement réactive est dénommée « carcinogène terminal » car, contrairement à son précurseur BaP, elle est capable d'attaquer

directement et de former des adduits covalents avec les bases de l'ADN, ce qui peut alors générer des mutations somatiques. Fait important, l'exposition d'une cellule aux HAP comme le BaP augmente la synthèse du CYP1A1, accélérant ainsi ces réactions[16]. Les carcinogènes terminaux ayant des durées de vie en tant que molécules libres de l'ordre de la seconde, la majeure partie des mutations somatiques qu'elles entraînent apparaissent dans les mêmes cellules où elles ont subi leur activation métabolique initiale. Ainsi, le BaP de la fumée de tabac est souvent activé dans les premières cellules où il pénètre, c'est à dire les cellules épithéliales des voies respiratoires, en particulier celles des poumons.

- **Rôle des polymérases dans la signature mutationnelle des cancers pulmonaires**

La polymérase êta (Polη) a été découverte à l'origine comme une polymérase de synthèse translésionnelle sans génération d'erreur, réparant les dimères de pyrimidine (TT, CC) induits par les ultraviolets (UV)[17,18]. Cependant, les premiers indices suggérant que cette polymérase serait impliquée dans la synthèse translésionnelle sujette aux erreurs dans la réparation des adduits de BaP provenait d'études biochimiques utilisant des matrices contenant un site spécifique (+) - trans-anti-BPDE-N2-dG. adduct[19]. La Polη purifiée est capable de reconnaître cette lésion volumineuse de l'ADN et insère principalement un A en regard de la lésion[19]. Des études similaires avec l'adduit (-) - trans-anti-BPDE-N2-dG ont montré que cette lésion était moins efficacement réparé par Polη que le produit d'addition (+) - trans-anti-BPDE-N2-dG[20,21]. La Polη participe à la synthèse translésionnelle sujette aux erreurs par insertion d'un A en face des adduits (+) - et (-) - trans-anti-BPDE-N2-dG, entraînant des substitutions G>T/C>A [22] (Figure 2). Par conséquent, contrairement à son rôle anti-mutagène en réponse aux lésions UV-induites sur l'ADN, Polη favorise la mutagénèse induite par les adduits (+) - et (-) - trans-anti-BPDE-N2-dG. En plus de l'insertion d'une base A, un autre mode majeur de synthèse translésionnelle mutagène est l'insertion d'un G en face des adduits (+) - et (-) - trans-anti-BPDE-N2-dG, conduisant à des substitutions G>C/A>T. Ainsi, les cancers pulmonaires des patients exposés aux fumées de tabac pourraient présenter une signature mutationnelle caractéristique qui serait due aux spécificités substitutionnelles des carcinogènes exogènes chimiques issues de la combustion du tabac.

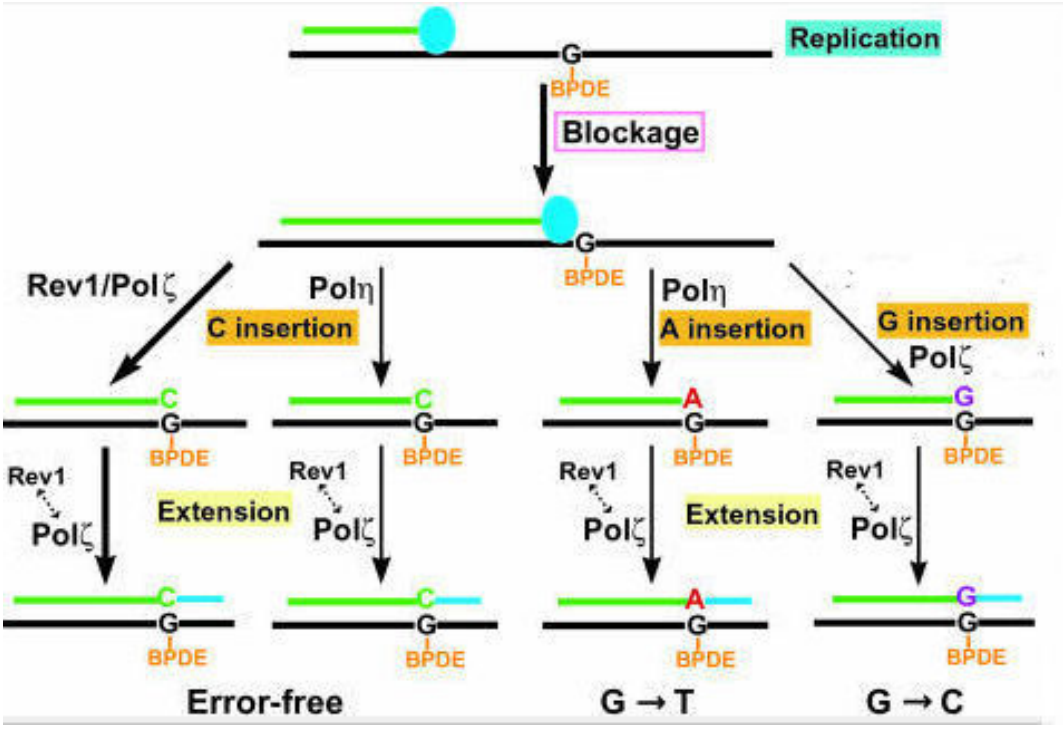


Figure 2. Adduit de BPDE et synthèse translésionnelle mutagène (G :A→T :A et G :A→C : T) et non-mutagène.[22]

❖ Carcinome de primitif inconnu

Un cancer sans primitif connu (CAPI en français et CUP en anglais pour Carcinoma of Unknown Primary origin) est un cancer métastatique histologiquement prouvé, pour lequel le bilan diagnostique initial n'a pas permis de trouver le site de la tumeur primitive. Leur incidence est faible (8/100 000 habitants/an, en France), mais il s'agit d'une cause de décès importante par cancer [23]. Ce sont des cancers hétérogènes qui continuent d'être un défi diagnostique et thérapeutique pour les médecins. La survie des patients atteints de CAPI est faible, avec une survie médiane de 8 à 11 mois et seulement 25% survivent à 1 an.[24] Les sites métastatiques peuvent varier, leur nombre également ainsi que la symptomatologie clinique associée, notamment l'état général des patients qui fait partie des facteurs pronostiques.

Les CAPI représentent environ 3% de toutes les tumeurs malignes, malgré les progrès de l'imagerie et de l'anatomo-pathologie [25], même après un bilan diagnostic complet, le site primaire d'un CAPI reste inconnu chez 20 à 50% des patients. Il s'agit d'un carcinome dans 95 % des cas, différents sous-types anatomopathologiques existent : le plus

souvent, il s'agit d'adénocarcinomes (80%) plus ou moins différenciés, mais il peut également s'agir de carcinomes épidermoïdes (15%) ou de carcinomes indifférenciés (5 %) qui peuvent alors être des tumeurs neuroendocrines ; une autopsie n'est pratiquée que chez une minorité de patients atteints de CAPI, mais même une évaluation approfondie à l'autopsie ne révèle que 55 à 85% des lésions primitives, qui sont généralement de petites tumeurs asymptomatiques, du poumon, pancréas, du tube digestif et du rein [26,27].

❖ Démarche anatomopathologique devant une métastase de primitif inconnu en pratique courante.

- **Diagnostic histologique et étude immunohistochimiques**

(À partir de « *référentiel ONCOLOR juillet 2011 » et de l'HISTOSÉMINAIRE DE LA SOCIÉTÉ FRANÇAISE DE PATHOLOGIE sur les « Carcinomes de site primitif inconnu. Le rôle du pathologiste en 2018 »*)

Toute métastase de primitif inconnu doit bénéficier d'un diagnostic anatomo-pathologique précis. Le tissu biopsié doit d'abord être évalué par microscopie optique après coloration standard (hématoxyline et éosine +/- safran). Il faut éliminer en premier lieu un diagnostic d'hémopathie de mélanome ou de carcinome. Une cytokératine à large spectre (AE1-AE3), la protéine S100 et un anticorps commun leucocytaire (CD45) sont à privilégier en premier lieu. Parmi les métastases d'une tumeur d'origine épithéliale, l'examen anatomo-pathologique permettra de distinguer :

- Un adénocarcinome (Figure 3)
- Un carcinome épidermoïde (Figure 4)
- Un carcinome peu différencié
- Un carcinome neuroendocrine

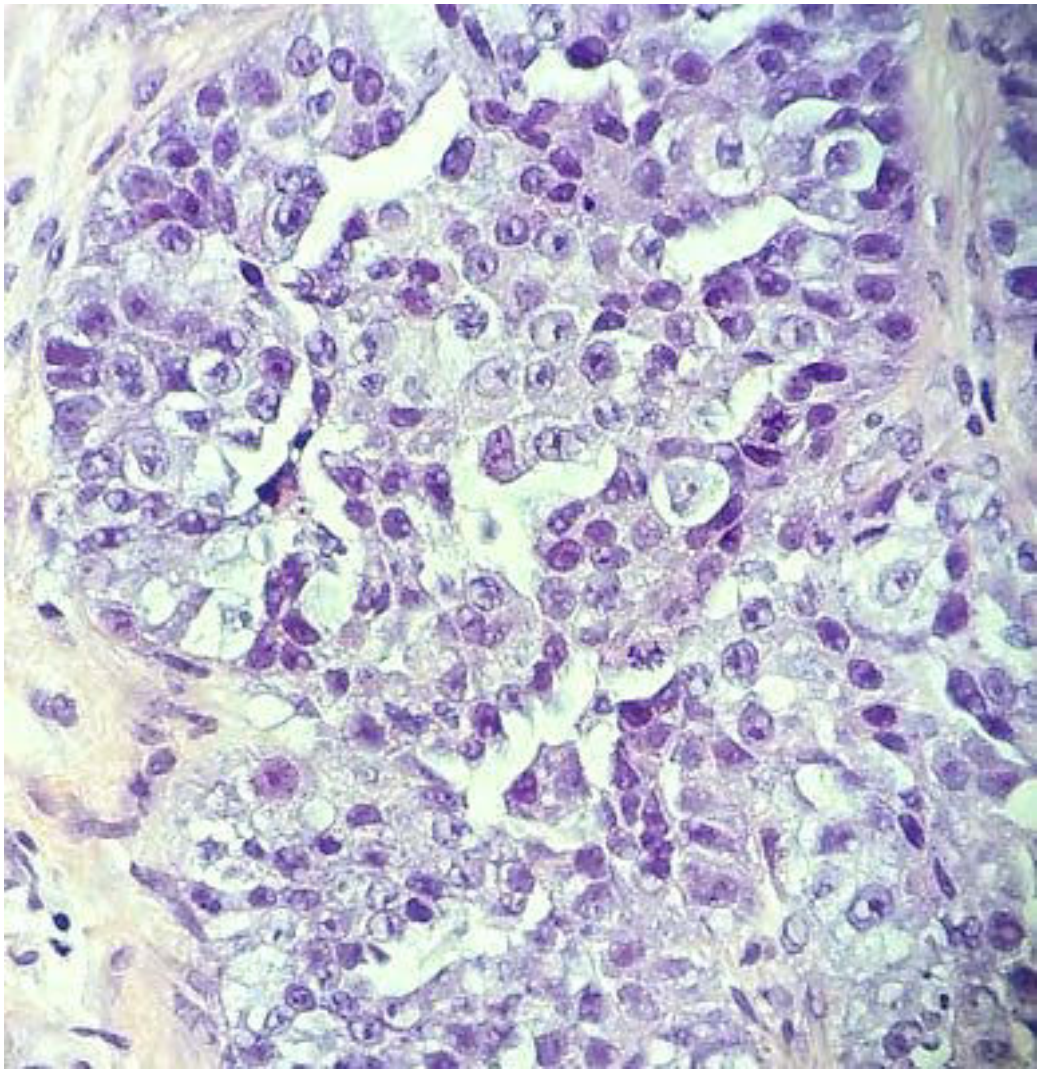


Figure 3. Prolifération tumorale dont les caractères sont ceux d'un adénocarcinome. La lésion est moyennement différenciée, d'architecture massive et tubulo-glandulaire. Les cellules sont de grande taille, au cytoplasme granuleux et aux noyaux modérément atypiques, nucléolés. L'index mitotique est moyennement élevé (1 à 2 mitoses/champ). (HESx20)

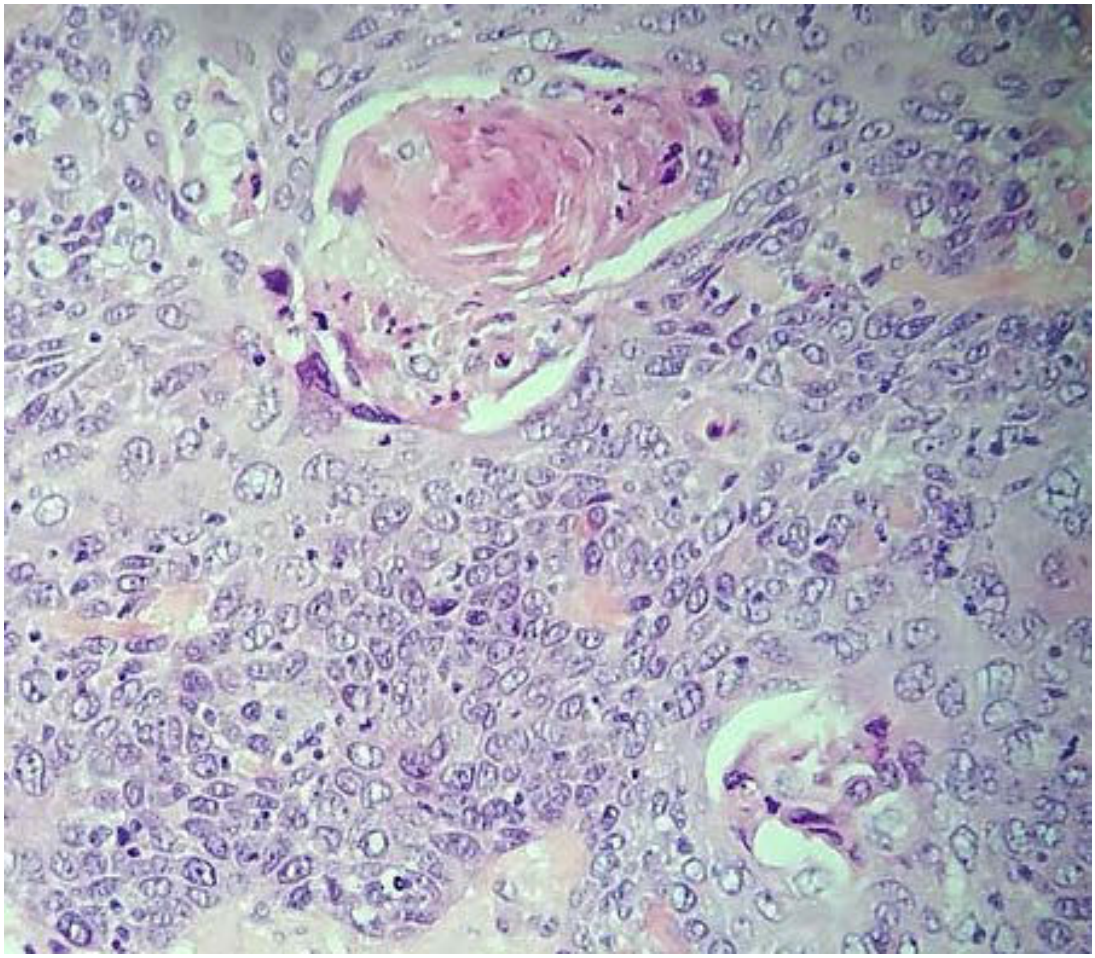


Figure 4 : Les carcinomes épidermoïdes bien différenciés présentent une kératinisation sous la forme d'un cytoplasme éosinophile abondant, de perles de kératine et/ou de ponts intercellulaires. Les tumeurs non kératinisantes ou faiblement différenciées nécessitent la démonstration immunohistochimique d'un ou plusieurs marqueurs épidermoïde (p40, p63, ou CK5/6). (HES x20)

Il est important d'identifier les tumeurs potentiellement curables et celles relevant de traitements spécifiques : lymphomes, tumeurs germinales, carcinomes pulmonaires non à petites cellules, carcinome prostatique, thyroïdien, ovarien, mammaire, neuroendocrine... Une étude immunohistochimique extensive est souvent nécessaire pour en préciser l'origine :

- Les cytokératines de faible poids moléculaire, notamment la cytokératine 7 (CK7) et la cytokératine 20 (CK20) sont de première importance et utilisées en routine pour tenter de déterminer l'origine d'un adénocarcinome.

- La CK7 est exprimée dans les tumeurs du poumon, de l'ovaire, de l'endomètre, des glandes salivaires et du sein alors que la cytokératine 20 s'exprime dans les carcinomes gastro-intestinaux ou urothéliaux.
- Le profil CK7 positif / CK20 négatif est fortement évocateur d'une origine pulmonaire, mammaire, thyroïdienne, biliaire, pancréatique ou ovarienne.
- Le profil CK7 négatif / CK 20 positif est suggestif d'une origine colorectale.
- Les adénocarcinomes gastriques et œsophagiens sont de phénotypes variables.
- La distinction entre un carcinome urothélial (CK20 et CK7 positifs) et un carcinome de prostate (CK 20 et CK 7 négatifs) est également importante.
- La p40 et la p63 sont exprimées dans les carcinomes épidermoïdes et les carcinomes urothéliaux.
- Le TTF1 (*Thyroïde Transcription Factor*) est caractéristique d'une origine pulmonaire ou thyroïdienne (thyroglobuline) d'un adénocarcinome (Figure 5).
- Les carcinomes neuroendocrines expriment le CD56, la synaptophysine, et pour les mieux différenciés, la chromogranine A.
- La thyroglobuline est spécifique des carcinomes thyroïdiens.
- Le PSA est spécifique des carcinomes prostatiques.
- DOG1 est présent dans les GIST (95 %)
- ERG est présent dans les tumeurs vasculaires (90%) : angiosarcome, hémangioendothéliome épithélioïde, hémangiome, sarcome de Kaposi, lymphangiome
- GATA3 dans les carcinomes mammaires (72 à 94 %) ; urothéliaux (67 à 93 %), les tumeurs annexielles cutanées (85 %) ; les choriocarcinomes et les tumeurs trophoblastiques gestationnelles (100 %) ; les paragangliomes (78 %) et phéochromocytomes (71 %) ; les tumeurs des glandes salivaires (51 % ; carcinome canalaire salivaire, 100 %) ; les tumeurs du rein : sous-type cellules claires et papillaire (76 %), chromophobe (51 %), oncocytome (17 à 19 %) ; dans les mésothéliomes, y compris sarcomatoïde et desmoplastique (58 %) ; les neuroblastomes (100 %) ; et le lymphome T périphérique sans autre précision (30 %)
- GCDFP-15 : Sein (60 à 80 %) Glandes salivaires (10 à 40 %) ; prostate (10 %) ; poumon adénocarcinome, 2 à 5 %)
- HepPar-1 : Hépatocellulaire (70 à 90 %) Estomac (5 à 15 %) ; œsophage (5 à 15 %) ; poumon (5 à 15 %)

- Inhibine A : Corticosurrénale (85 à 95 %) Tumeurs des cordons sexuels (90 %)
- Mammaglobine : Sein (60 à 80 %) Endomètre (10 à 30 %) ; glandes salivaires, glandes sudoripares (40 à 50 %)
- PAX8 : Rein (85 à 95 %), thyroïde (60 à 100 %) ; ovaire (non-mucineux, 90 à 100 % ; mucineux, 10 à 40 %) ; endomètre (90 à 100 %) ; col de l'utérus (90 à 100 %)
- SALL4 : Tumeurs germinales : séminome (100 %), carcinome embryonnaire (100 %), tumeur vitelline (100 %), choriocarcinome (100 %), tératome (45 %) Estomac (15 %) ; tumeurs rhabdoïdes malignes (80 %, système nerveux central ou extra-crâniennes)
- SATB2 : Colorectal (85 %, médullaire) Appendice (100 %) ; carcinome à cellules de Merkel (75 %) ; sinonasal (56 %) ; rein (26 %)
- Villine : Colorectal (80 à 90 %) Rein (60 à 80 %) ; estomac (60 à 80 %) ; endomètre (30 à 40 %) ; poumon (10 %) ; ovaire (10 %)
- WT1 : Ovaire (séreux, 88 à 100 %) Mésothéliome (76 %)

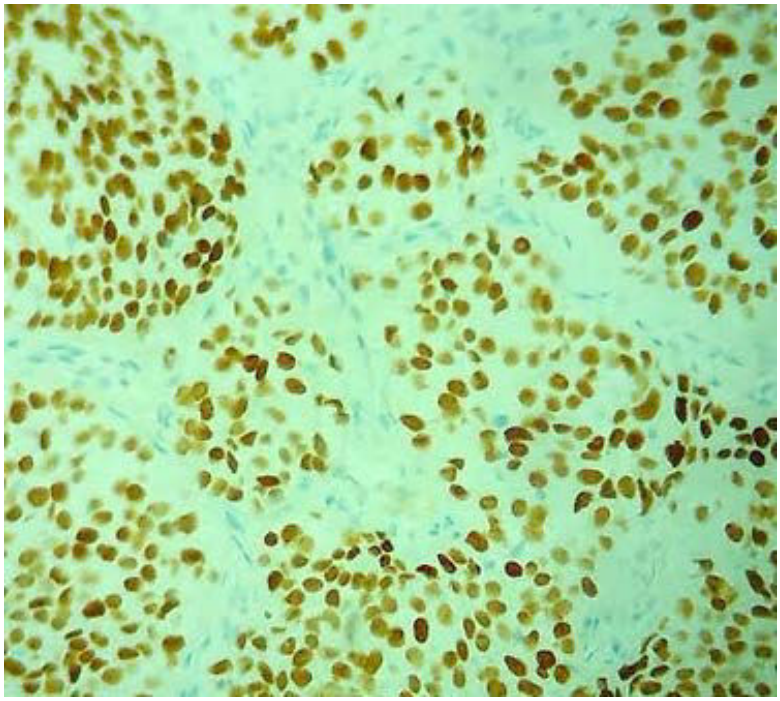
- 

Figure 5 : Expression immunohistochimique nucléaire du TTF-1 par les cellules carcinomateuses orientant vers une origine pulmonaire ou thyroïdienne. (X20)

❖ Objectif de notre étude

En dépit d'une étude histologique approfondie, dans un certain nombre de cas, il est difficile voire impossible de préciser avec certitude l'origine d'une métastase. Certaines tumeurs expriment en effet des immunophénotypes variables. Le diagnostic final se limite parfois à proposer un groupe de tumeurs présentant un profil comparable, et à des diagnostics d'exclusion de primitifs dont le profil immunophénotypique est incompatible. Ainsi, la détermination de l'origine pulmonaire d'une tumeur est parfois difficile, en particulier si l'on suspecte un cancer bronchique non à petites cellules (épidermoïde ou adénocarcinome), pour lequel des thérapies ciblées efficaces sont disponibles. Dans ce contexte, le but de notre travail était de développer un test moléculaire innovant capable de prédire l'origine pulmonaire d'une tumeur. Notre approche utilise les fréquences relatives des différents types de substitutions somatiques, avec l'hypothèse que ces caractéristiques pourraient être, à elle seules, discriminantes entre les cancers pulmonaires et les cancers d'une autre origine primitive sur un résultat d'exome.

ARTICLE

THE SOMATIC SUBSTITUTION SIGNATURE AS AN INNOVATIVE TOOL IN LUNG CANCER DIAGNOSIS

Stéphane BUSCA[1], Julia SALLERON[2], Romain BOIDOT[3,4], Jean-Louis MERLIN[1,5], Alexandre HARLE[1,5,*]

(1) Service de Biopathologie, Institut de Cancérologie de Lorraine F-54519, Vandœuvre-lès-Nancy, France

(2) Data Biostatistics Unit, Institut de Cancérologie de Lorraine F-54519, Vandœuvre-lès-Nancy, France

(3) Department of Tumor Biology and Pathology, Georges-Francois Leclerc Cancer Center – UNICANCER

(4) Institut National de la Santé et de la Recherche Médicale (Inserm), LNC UMR1231, Dijon, France.

(5) Université de Lorraine CNRS UMR 7039 CRAN, F-54519, Vandœuvre-lès-Nancy, France

Running title: Somatic substitutions signature for lungs cancer diagnosis **Keywords:** Lung Cancer, Diagnosis, somatic substitution, Score, Signature **Additional information:** Financial support: No funding * Corresponding author:
Dr. Alexandre Harlé

Service de Biopathologie,

Institut de Cancérologie de Lorraine

6 Avenue de Bourgogne, CS30519, 54519 Vandoeuvre-les-Nancy

France

Tel. +33 383 59 86 73

a.harle@nancy.unicancer.fr

Conflict of interest: The authors declare no conflicts of interest

❖ Abstract

Diagnosis of lung cancer can sometimes be challenging and is of major interest since effective molecular-guided therapies are available. To predict whether a tumor is of lung origin or not, we developed and validated the EASILUNG (Exome And SIgnature LUNG) test based on the relative frequencies of somatic substitutions on coding non-transcribed DNA strands from whole-exome sequenced tumors. Data from 7,796 frozen tumor samples (prior to any treatment) from 32 TCGA solid cancer groups were used for its development. External validation was carried out on a local dataset of 196 consecutive routine exome results. The EASILUNG test demonstrated good calibration and good discriminative power with a sensitivity of 80% and a specificity of 94.3%. Similar results were obtained from external validation. This innovative test may be helpful in medical decision-making in patients with unknown primary tumors potentially of lung origin and in the diagnosis of lung cancer.

SIGNIFICANCE

Our data indicate that relative frequencies of somatic substitutions on coding non-transcribed DNA strands from whole-exome sequenced tumors can predict whether a tumor is of lung origin or not. These parameters may be helpful in medical decision-making, especially in patients with unknown primary tumors.

❖ Introduction

Current cancer therapies are generally based on anatomical site. Identifying a tumor's origin is therefore crucial for treatment selection. Lung cancer is the most common major cancer diagnosed worldwide (1) and is mostly represented by two subtypes, small cell lung carcinoma and non-small cell lung carcinoma (NSCLC), accounting for 15% and 85% of all lung cancer, respectively (2). Squamous cell carcinoma (SCC) represents 25 to 30% of lung cancers and adenocarcinoma is the most common histological subtype of lung cancer, representing 40% of all lung cancers. A cancer diagnosed at a metastatic stage with no identifiable primary tumor is termed a cancer of unknown primary origin (CUP). A diagnosis of CUP is evocated when the clinical picture of metastatic disease is associated with one or more tissue biopsy results that are inconsistent with a determined primary tumor. Two to four percent of newly diagnosed cases of carcinoma are CUP (3); histological subtypes of these CUP are adenocarcinomas (40–60%), undifferentiated carcinomas (15–30%), and SCCs (15–20%) (4). Patients with CUP syndrome have a median survival of 8–11 months and only 25% survive at 1 year (5). The Human Genome Variation Society (HGVS) nomenclature defines 12 different somatic substitution classes (C>A, G>T, C>G, G>C, C>T, G>A, A>C, T>G, A>G, T>C, A>T, and T>A) that can be read on the coding non-transcribed DNA strand. Some childhood cancers, like acute myeloid leukemia (AML), are reported to carry very few somatic substitutions, whereas cancers such as lung carcinomas and malignant melanoma, which are related to chronic exposure to mutagens such as tobacco or ultraviolet light, present with a high number of somatic substitutions (6). The prevalence of somatic substitutions, quantified in substitutions per megabase (sub/Mb), can be calculated by dividing the number of somatic substitutions by the size of the DNA sequence targeted region in Mb; this is also known as the tumor mutational burden. A mutational signature is a characteristic combination of mutation types arising from specific mutagenesis processes, such as DNA replication infidelity, exposure to exogenous and endogenous genotoxins, defective DNA repair pathways, and DNA enzymatic editing (7). Tissues directly exposed to tobacco smoke (*e.g.*, lung), as well as some tissues not directly exposed (*e.g.*, bladder), show elevated levels of DNA adducts in smokers and thus evidence of exposure to carcinogenic components of tobacco smoke (8,9). Benzo[a]pyrene (BaP) present in tobacco smoke is the most extensively studied compound. After two successive oxidation reactions mediated by cytochrome P450 enzymes (largely CYP1A1), BaP is converted to benzo[a]pyrenediolepoxide (BPDE). BPDE is considered an ultimate

carcinogen because, unlike its BaP precursor, it has the ability to directly attack and form covalent adducts with DNA bases, which may then generate G>T/C>A somatic substitutions. Importantly, exposure of a cell to polycyclic aromatic hydrocarbons like BaP and other compounds of tobacco smoke increases synthesis of CYP1A1, thereby accelerating these reactions in cells (10,11), and may generate a specific substitutional signature in lung, which is the most exposed organ.

The aim of our study was to develop and validate an innovative genomics-based test, the EASILUNG (Exome And SIgnature LUNG) test, able to predict whether or not a tumor is of lung origin using only the prevalence and the relative frequencies of somatic substitutions from whole-exome sequencing (WES) results.

❖ Results

- **Score development**

The development dataset included 7,796 WES results from solid tumors from The Cancer Genome Atlas (TCGA). Among the 7,796 samples, 745 (9.56%) were tumors where the primary origin was lung, from the lung adenocarcinoma group ([LUAD]) or the lung SCC group ([LUSC]) and 7,051 (90.44%) were from the 30 other non-lung primary location groups (Supplementary Table 1). Among these 7,796 samples, 2,888 samples had a somatic substitution prevalence of < 1 sub/Mb, including 33 lung cancer cases (1.1%). We considered one lung tumor with a somatic substitution prevalence lower than 1 sub/Mb to be non-lung cancer (because reports of lung cancers with a < 1 sub/Mb prevalence are rare). Table 1 describes the 12 substitution classes for the 4,908 tumors with a somatic substitution prevalence of ≥ 1 sub/Mb, including 702 (14.3%) lung tumors and 4,206 (85.70%) non-lung tumors. The bivariate analyses demonstrated that all classes of substitutions were predictive of lung cancer (Table 2). A relative frequency of C>A (%CA) greater than 12% or of G>T (%GT) greater than 13% were associated with a greater risk of lung cancer (respectively with an odds ratio and 95%CI of 14.44 [11.95;17.46] and 9.17 [7.66;10.98]) whereas the presence of C>G or G>C were associated with a lower risk of lung cancer (respectively 0.45 [0.38;0.54] and 0.50 [0.42;0.60]). The full multivariate model to predict lung cancer including the 12 substitution classes is reported in Table 2. After backward selection with bootstrap validation, eight classes of substitutions were significantly associated with lung cancer and were retained for the EASILUNG test (Table 3). Table 4 provides two examples of how the EASILUNG score was calculated. To illustrate the application of the EASILUNG test (Example 1, Table 4),

when we considered a tumor with %CA=27, %GT=19, %CT=8, %GA=12, %AC=0, %TG=0, %AG=3, and %TC=5, the EASILUNG score was 93 and the estimated predicted probability that the tumor was lung cancer was 0.89, according to Fig. 1.

- **EASILUNG test performance on an TCGA development dataset**

The median EASILUNG score was -485 (IQR from -582 to -353). The score was significantly higher for lung cancer: 0 (from -139 to 2) vs -511 (from -597 to -411) for non-lung cancer ($p < 0.0001$). Area under the ROC curve (AUC) of the final model was 0.931 [95%CI 0.920;0.942]. Descriptive characteristics of the EASILUNG score on the development dataset and the probability of lung cancer are provided in Figure 1. The calibration assessed by the Hosmer-Lemeshow chi-square test was equal to 10.89 (p=0.2079). The best EASILUNG score threshold to separate lung and non-lung cancer was set at -303: a score greater than -303 favored lung cancer. The associated sensitivity and specificity were respectively 85.7% [83.2%; 88.3%] and 89.6% [88.7%; 90.6%]. Among the 7,796 samples, when we considered the 2,888 samples with a somatic substitution prevalence of < 1 sub/Mb to be non-lung cancer, the respective overall sensitivity and overall specificity of the EASILUNG-test were 80.0% [77.1%;82.8%] (596/745) and 94.3% [93.7%;94.8%] (6,646/7,051). Performance of the EASILUNG test thresholded at -303 on the development dataset are presented for adenocarcinoma groups and subgroups, SCC groups and subgroups, and other solid cancer groups and subgroups in Supplementary Table 2.

- **Test performance on external validation dataset**

Characteristics of 196 samples and distribution of the relative frequencies of the 12 substitution classes are presented in supplementary tables 3 and 4. Among 196 available tumors with WES data, we considered one lung tumor with a somatic substitution prevalence lower than 1 sub/Mb to be non-lung cancer. The EASILUNG score was computed on the remaining 195 samples. When using the threshold of -303, the associated sensitivity and specificity were respectively 48.9% [34.6 %; 63.2%] and 95.3% [91.8%;98.7%]. After calibration on the validation dataset, the AUC of the EASILUNG score was 0.849 [0.781;0.917]. The calibration assessed by the Hosmer-Lemeshow chi-square test was equal to 8.63 (p=0.3745) with close agreement between predicted and observed risk of lung cancer, with no

apparent over- or underprediction (Supplementary Fig. 1). Descriptive characteristics of the EASILUNG score on the development dataset and the probability of lung cancer are provided in Fig. 2. When using the new threshold of -467 calculated on the validation dataset, the associated sensitivity and specificity were respectively 83.0% [72.2 %;93.7%] and 72.3% [65.1%;79.5%]. The associated performance of the EASILUNG test on the validation dataset is detailed in Supplementary Table 5.

❖ Discussion

EASILUNG is the first test to use the relative frequency of somatic substitutions from WES results as a differential diagnostic tool. Using TCGA data, Alexandrov *et al.* (6) published a comprehensive signature classification based on the different patterns of genomic substitutions in cancer. In their study, they describe signatures with 96 parameters (corresponding to the type of substitution in their trinucleotide context), and in particular a tobacco-induced signature. These 96-parameter signatures are descriptive and are not actually used clinically in diagnosis, because each is present in several classes of cancer (*e.g.* the signature proposed for tobacco is present in the same way in lung cancers, head and neck SCCs, and in liver hepatocellular carcinoma (6)) without specified thresholds and therefore without assessment performed for the sensitivity and specificity of these signatures. Based on only 8 variables (8 of the 12 somatic substitution possibilities that can be read on the nontranscribed coding strand), here we propose that can be easily applied to whole-exome results in cases where the somatic substitution prevalence is ≥ 1 sub/Mb.

We excluded by convention patients with a somatic substitution prevalence of < 1 sub/Mb and categorized them as having non-lung cancer, because few lung cancers with a < 1 sub/Mb somatic substitution prevalence have been described (6). Our approach of having 12 somatic substitution classes allowed consideration of transcriptional strand bias while analyzing a small number of parameters. Strand bias of tobacco-induced mutagenesis could be indicative of specific repair processes that are active in the cell; the mutational signature associated with lung cancer exhibited predominantly G>T/C>A substitutions with a transcriptional strand bias (more G>T transversions than complementary C>A somatic substitutions in the coding nontranscribed strand), suggesting the formation of bulky adducts on guanine and previously described by *Alexandrov et al.* (6). The efficiency of DNA damage and DNA maintenance processes differ between the transcribed and non-transcribed gene strands. A well-known cause of

this phenomenon is transcription-coupled nucleotide excision repair (NER) that operates predominantly on the transcribed strand of genes and is initiated by RNA polymerase II when it encounters bulky DNA helix-distorting lesions (12). The high rate of G>T/C>A somatic substitutions in lung cancer is probably an imprint of the bulky DNA adducts generated by polycyclic hydrocarbons found in tobacco smoke and their removal by transcription-coupled NER (13). Furthermore, in our study high C>A relative frequency (%CA >12%) was the most discriminative parameter, with the highest odds ratio, followed by high G>T relative frequency (%GT >13%). The higher prevalence of G>T substitutions on the non-transcribed strand compared to complementary C>A substitutions in lung cancer reflects this strand bias and confirms that bulky adduct damage to guanine may be the cause of the observed mutations. In contrast, no evidence of a transcriptional strand bias was observed for C>G/G>C, C>T/G>A, T>A/A>T, or T>C/A>G somatic substitutions in our study, as suggested by the extremely close odds ratios (and cut-offs when there were thresholds for variables) of these pairs of parameters in the EASILUNG test (Table 1).

We chose to make distinct TCGA histological subgroups, anatomical subgroups, and immunophenotypical subgroups (see Supplementary Table 2) relevant in the differential diagnosis, because some TCGA groups are very heterogeneous, especially histologically (cervical SCC and endocervical adenocarcinoma [CESC], cholangiocarcinoma [CHOL], esophageal carcinoma [ESCA], mesothelioma [MESO], pheochromocytoma and paraganglioma [PCPG], sarcoma [SARC], skin cutaneous melanoma [SKCM], testicular germ cell tumors [TGCT], thyroid carcinoma [THCA], thymoma [THYM], and uterine corpus endometrial carcinoma [UCEC] groups). We also specified anatomical subdivision in the head and neck SCC [HNSC] group because epidemiological and etiological heterogeneities exist in these subgroups (oral cavity, oropharynx, hypopharynx, and larynx) although the histological type is still SCC. The larynx is the anatomic region closer to the trachea than other anatomic regions of head and neck SCC. It is therefore clinically relevant to evaluate the results of our test in these different anatomical subgroups. Positive EASILUNG tests in SCCs localized in the larynx (30/61; 49%), which suggests this location is more closely similar to lung SCCs than other head and neck SCC locations (oral cavity (6/153; 4%), oropharynx (2/27; 7%), and hypopharynx (0/3; 0%)). The gradient in test positivity through the airways in the HNSC group may reflect the existence of an underlying, parallel gradient in the extent of exposure of cells of the respiratory tract to BaP and other tobacco smoke carcinogens. These results can be linked with a previous study (13) about a gradient in the prevalence of

TP53 G>T transversions in cancers of smokers, from low in the oral cavity, to intermediate in the larynx, and high in various histological types of lung cancers.

Besides smoking, expression of APOBEC (Apolipoprotein B mRNA editing catalytic polypeptide-like) family members, especially APOBEC3B, is a key source of mutations in NSCLC (14). The APOBEC enzyme family catalyzes the hydrolytic reaction of cytosine to uracil in single-stranded DNA, and is known to play a role in approximately half of all human cancers, with lung, breast, head and neck, bladder, and cervical cancers bearing the largest APOBECattributable mutation burdens with a high level of protein expression (15). Cytidine deamination converts cytosine to uracil, which usually results in C>T/G>A somatic substitutions (16), and contributes to the high absolute frequency of C>T (and complementary G>A) somatic substitutions in lung cancer. Interestingly, the relative frequencies of C>T (%CT) and complementary G>A (%GA) in the EASILUNG score have a negative coefficient, and therefore the increase in C>T/G>A tends to decrease our score (Table 2). Our results suggest that a high prevalence of G>T/C>A somatic substitutions induced by carcinogens from tobacco smoke are strongly specific to tumors of lung origin. In contrast, APOBEC-induced somatic substitutions (C>T/G>A) appeared not specific enough of lung origin to have a positive coefficient in their related parameters (%CT and %GA) in our test; they nevertheless contribute to the increase in the overall prevalence of somatic substitutions, which is a necessary condition (≥ 1 sub/Mb) both for the application of and a positive EASILUNG test.

Although in most cases the primary origin of a lung tumor is not questioned by clinicians or pathologists, some situations are more problematic because lung is both an extremely common site of origin of metastasis and a common site for metastases. Some RT-PCR molecular tests studying mRNA expression exist to help identify cases not resolved by immunohistochemistry and classified as CUP. More recently, an alternative strategy was to propose a treatment targeting the supposed primary tumor according to the molecular signature, based on gene expression analyzed by a 92-gene RT-PCR assay (17). In that study, the sensitivity was 63% for lung adenocarcinoma from frozen samples and 0% for formalinfixed paraffin-embedded (FFPE) samples. One of the hypotheses explaining the limitations of diagnostic methods studying mRNA expression is that a genetic signature predisposing to metastasis coexists in primary tumors with the tissue-specific genetic signature. This "metastatic" genetic program that transcends distinctions between the original tissues has been described and could lead to errors in the molecular classification of various tumors at primary sites such as breast or bladder (18). The selection of genes relies heavily

on the number and type of primary solid tumors studied. In most published series, the "training set" included fewer than 1000 tumor specimens (typically 100–400). In addition, gene expression profiles are heterogeneous in pulmonary adenocarcinomas (19). By studying DNA, our approach is not impacted by these limitations; this explains our good results and suggests the superiority of our approach in lung cancers, particularly in adenocarcinomas.

By analyzing lung adenocarcinomas and lung SCC, we cover the two most prevalent lung cancer histologies and our results suggest that the EASILUNG test remained acceptably sensitive and strongly specific on FFPE samples. The mutational artefact potentially generated by formalin fixation (20) was not a limitation for our test, as demonstrated by our local sample results. Indeed, formalin crosslinking with cytosine nucleotides on either strand can result in incorrect incorporation of adenine in place of guanosine, causing an artificial C>T/G>A mutation (21,22) that could impact negatively the score (Table 2). Thus, pulmonary tumors with a pronounced somatic substitutional signature would remain positive, while lung tumors and non-lung tumors with a borderline positive EASILUNG score (greater than but close to 303) would be negative.

We showed that the EASILUNG test can distinguish lung cancer from other groups of solid tumors. Unlike sensitivity and specificity, positive predictive values (PPV) and negative predictive values (NPV) are largely dependent on disease prevalence in the examined population. Therefore, predictive values from one study cannot be transferred to another setting with a different prevalence of the disease in the population. PPV increases and NPV decreases when a disease increases in a population. Prevalence of lung cancer in CUP, by definition, cannot be estimated, and nor can the prevalence in CUP of different groups and subgroups of cancers across the different datasets (development dataset, external validation dataset, and dataset of laboratories where the test will be assessed); for this reason, NPV and PPV are less relevant than sensitivity and specificity in our study. A known major issue of external validation studies is that they are often based on small and local datasets. This means that there is often heterogeneity in the performance of prediction models and that multiple external validation studies are needed to fully appreciate the generalizability of a model. In our study, results are validated on our local platform. Our validated calibration allows us to consider artefactual variation of data analysis and interpretation of biologic data across centers. However, multiple external validation studies and prospective validation studies on exome results are needed to fully realize the generalizability of our model. Clinical trials are required to evaluate its benefits to patients.

No somatic substitutions were described in 35 tumors (1 lung SCC and 34 non-lung tumors), representing approximately one tumor out of 1000 in TCGA. We assume that these tumors may have low allele frequency somatic mutations that may be excluded from the variants list (23,24). We thus decided to exclude these 35 tumors from our validation dataset, even though specificity would have been better, because all 35 tumors would have been considered negative by the EASILUNG test (< 1 sub/Mb). The AML TCGA cancer group was not included simply because there is no possible diagnostic confusion between AML and solid cancers in routine clinical practice. Furthermore, AML genomes have fewer mutations than most other adult cancers, with an average of only 13 mutations found (25). Because of the low density of somatic substitutions in AML (< 1 sub/Mb), to reject AML the specificity of our test would have be equal to 100%. Inclusion of AML would have caused inflation of the specificity of our test without any clinical relevance.

In conclusion, we developed an innovative genomics-based tool which may be helpful in diagnosis of lung cancer and in medical decision-making for patients with unknown primary tumors potentially of lung origin. The external validation of our test confirmed its great potential for clinical settings.

❖ Methods

- **Development dataset**

We used the full publicly available TCGA data where somatic mutations are provided as simple somatic mutations (SSM) from WES results. SSM consists of somatic substitutions, deletions of ≤ 200 bp, insertions of ≤ 200 bp, and multiple base substitutions (≥ 2 bp and ≤ 200 bp). Mutation annotation format (MAF) files containing open access SSM data of the cancer genomes used in this study were obtained from all 33 TCGA groups (February 1, 2016). SSM data from all genome sequencing center (GSC) platforms available were downloaded, corresponding to 130 MAF files containing somatic mutation data from 28,437 WES results from 9,522 distinct tumors sequenced by one to seven GSCs. All TCGA data were generated from frozen tumor samples prior to any treatment. The AML group (2 MAF files containing 394 WES results from 197 distinct samples), WES results with any substitutions (35 WES results), and MAF files that did not include the size of the targeted region of the exome (76 MAF files containing 16,514 WES results) were excluded from the development dataset; these conditions were not mutually exclusive. All WES results from all TCGA solid cancer groups (32/33 TCGA groups) from all GSC platforms that included the size of the targeted region of the exome in MAF files or in previously related TCGA publications were evaluated for the development dataset (26–43), corresponding to 54 of the 130 MAF files containing somatic mutations in 11,720 WES results from 7,796 of the 9,342 distinct tumors. Each TCGA tumor had a unique barcode; İf several GSCs had sequenced the same tumor and there were multiple WES results for the same tumor, a randomization was carried out to select only one WES result for each tumor. In total, the development dataset included 7,796 WES results from 7,796 distinct tumor samples (745 lung cancer and 7,051 non-lung cancer). All details are provided in Supplementary Methods.

- **Substitution classes and prevalence of somatic substitutions**

Somatic substitutions were analyzed on the coding non-transcribed strand and their absolute frequencies were classified into the 12 classes according to the HGVS nomenclature (C>A, G>T, C>G, G>C, C>T, G>A, A>C, T>G, A>G, T>C, A>T and T>A). Targeted region size (in Mb) and absolute frequencies of the 12 somatic substitution classes read on the coding nontranscribed strand were collected. Insertion, deletion, and multiple base substitutions were not

considered. The total number of somatic substitutions was obtained by adding absolute frequencies of the 12 somatic substitution classes. Prevalence of somatic substitutions for each tumor was quantified in sub/Mb and was calculated by dividing the total number of somatic substitutions by the size in Mb of the exome targeted region, which varied between 36 to 64 Mb across TCGA GSCs and groups of solid cancers (Supplementary Methods).

- **Relative frequencies of substitution classes**

Relative frequencies of somatic substitutions for each class (%CA, %GT, %CG, %GC, %CT, %GA, %AC, %TG, %AG, %TC, %AT, and %TA) were determined as the absolute frequency of each somatic substitution class (respectively C>A, G>T, C>G, G>C, C>T, G>A, A>C, T>G, A>G, T>C, A>T, and T>A) divided by the total number of somatic substitutions.

- **External validation dataset**

A total of 196 consecutive tumor samples sequenced between January 2015 and August 2017 available from the Molecular Tumor Board of Centre Georges-François Leclerc (CGFL, Dijon, France) was used for the external validation dataset. All patients gave their consent and the study was approved by the CGFL ethical board. All tumors had WES results and matched histopathological characteristics. All tumors and paired germline DNA were sequenced in 2 x 150 cycles on NextSeq500 using the Agilent SureSelect XT Human All exon v5 kit (targeted region = 50Mb). All samples were FFPE or snap-frozen human tumors (primary or metastasis) from different histological classes and localizations. Initial site and diagnoses were determined using histopathological information. For all samples, normal DNA from the same individuals had also been sequenced to establish the somatic origin of variants and exclude germline single nucleotide polymorphisms. For each tumor, a file containing two columns with the allele of reference and mutated allele for each single-base substitution read in the coding non-transcribed strand was sent to our institution (Institut de Cancérologie de Lorraine, France), blinded of all other parameters.

- **Identification of the predictive model and establishing the EASILUNG score**

Patients with a somatic substitution prevalence of < 1 sub/Mb were classified as having non-lung cancer. For patients with a somatic substitution prevalence of ≥ 1 sub/Mb, the EASILUNG score was computed using the 12 substitution classes. They were described by the median and the interquartile range, since the distributions of these parameters were not normal according to the Kolmogorov-Smirnov test. The log-linearity assumption of the logistic model was checked. This was carried out by categorizing each classes of substitution into 10 groups (corresponding to deciles) and by looking at the plot of the logit of observed percentages of lung cancer in each class. When this assumption was not verified, the continuous parameter was transformed into binary variables. This was performed by using the AUC. The threshold maximizing the Youden index was chosen.

A multivariable logistic regression was performed with the 12 substitution classes as independent parameters and the lung cancer as the dependent parameter (full model). The simplification of this full model to avoid overfitting was done using a multivariable logistic regression with backward selection at the level p=0.2. The stability of the selected model was investigated using the bootstrap resampling method (44). A multivariable logistic regression was performed using the substitution classes selected by the bootstrap resampling method (final model). Logistic coefficients of regression (β) of the final model were multiplied by 100 and rounded to the nearest integer. These integers were then added to obtain an overall lung cancer risk score for each patient, namely the linear predictor lp. The model's discriminative performance was assessed by the AUC, and the calibration by the Hosmer-Lemeshow chisquare test. The threshold of the predicted risk score maximizing the Youden index was chosen and the sensitivity and the specificity with their 95% confidence intervals were computed.

The overall performance of the classification was assessed by taking into account the patients with a somatic substitution prevalence of < 1 sub/Mb by classifying them as having non-lung cancer. The sensitivity and the specificity with their 95% confidence intervals were computed on the total population.

- **External validation**

The EASILUNG score was computed on the validation set. The discriminative performance of the EASILUNG test on the validation set was assessed by the sensitivity and the specificity using the threshold identified with the development dataset. To have a more accurate assessment of the probability of lung cancer in the validation set, the EASILUNG test was calibrated using a logistic regression. After the recalibration, the discriminative performance was assessed by the AUC and the calibration by the Hosmer-Lemeshow chi-square test. A new threshold maximizing the Youden index was chosen and the sensitivity and the specificity with their 95% confidence intervals were computed.

- **Data availability**

The local validation sample datasets analyzed during the current study are not fully publicly available: due to reasonable privacy and security concerns, the data are not easily redistributable to researchers other than those engaged in Institutional Review Board approved research collaborations with the named medical center.

- **Acknowledgements**

TCGA consortium. Centre Georges-François Leclerc (Dijon, France).

- **Author contributions**

S.B. conceived the original idea, extracted the TCGA data, and wrote the manuscript. J.S. conceived and wrote the statistical part of the manuscript, planned statistical analysis, and drafted figures. S.B, J.S, and A.H contributed with study design and revised the paper. R.B. contributed with acquisition of data from the external validation samples. A.H. directed the overall research.

❖ REFERENCES (ARTICLE)

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012: Globocan 2012. Int J Cancer. 2015;136:E359–86.

2. Sher T, Dy GK, Adjei AA. Small Cell Lung Cancer. Mayo Clin Proc. 2008;83:355–67.

3. Levi F, Te VC, Erler G, Randimbison L, La Vecchia C. Epidemiology of unknown primary tumours. Eur J Cancer Oxf Engl 1990. 2002;38:1810–2.

4. Zaun G, Schuler M, Herrmann K, Tannapfel A. CUP Syndrome—Metastatic Malignancy with Unknown Primary Tumor. Dtsch Ärztebl Int. 2018;115:157.

5. Pavlidis N, Briasoulis E, Pentheroudakis G, ESMO Guidelines Working Group. Cancers of unknown primary site: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol Off J Eur Soc Med Oncol. 2010;21 Suppl 5:v228-231.

6. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013;500:415–21.

7. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 2017;45:D777–83.

8. Hecht SS. DNA adduct formation from tobacco-specific N-nitrosamines. Mutat Res Mol Mech Mutagen. 1999;424:127–42.

9. Phillips DH, Venitt S. DNA and protein adducts in human tissues resulting from exposure to tobacco smoke. Int J Cancer. 2012;131:2733–53.

10. ADME: Atomic views of a human P450. Nat Rev Drug Discov. 2003;2:685–685.

11. Miller EC. Some current perspectives on chemical carcinogenesis in humans and experimental animals: Presidential Address. Cancer Res. 1978;38:1479–96.

12. Hanawalt PC, Spivak G. Transcription-coupled DNA repair: two decades of progress and surprises. Nat Rev Mol Cell Biol. 2008;9:958–70.

13. Pfeifer GP, Denissenko MF, Olivier M, Tretyakova N, Hecht SS, Hainaut P. Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. Oncogene. 2002;21:7435–51.

14. Burns MB, Temiz NA, Harris RS. Evidence for APOBEC3B mutagenesis in multiple human cancers. Nat Genet. 2013;45:977–83.

15. Swanton C, McGranahan N, Starrett GJ, Harris RS. APOBEC Enzymes: Mutagenic Fuel for Cancer Evolution and Heterogeneity. Cancer Discov. 2015;5:704–12.

16. Burns MB, Leonard B, Harris RS. APOBEC3B: pathological consequences of an innate immune DNA mutator. Biomed J. 2015;38:102–10.

17. Hainsworth JD, Rubin MS, Spigel DR, Boccia RV, Raby S, Quinn R, et al. Molecular gene expression profiling to predict the tissue of origin and direct site-specific therapy in patients with carcinoma of unknown primary site:

a prospective trial of the Sarah Cannon research institute. J Clin Oncol Off J Am Soc Clin Oncol. 2013;31:217–23.

18. Ramaswamy S, Ross KN, Lander ES, Golub TR. A molecular signature of metastasis in primary solid tumors. Nat Genet. 2003;33:49–54.

19. Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, et al. Diversity of gene expression in adenocarcinoma of the lung. Proc Natl Acad Sci U S A. 2001;98:13784–9.

20. Williams C, Pontén F, Moberg C, Söderkvist P, Uhlén M, Pontén J, et al. A High Frequency of Sequence Alterations Is Due to Formalin Fixation of Archival Specimens. Am J Pathol. 1999;155:1467–71.

21. Srinivasan M, Sedmak D, Jewell S. Effect of Fixatives and Tissue Processing on the Content and Integrity of Nucleic Acids. Am J Pathol. 2002;161:1961–71.

22. Chen G, Mosier S, Gocke CD, Lin M-T, Eshleman JR. Cytosine Deamination Is a Major Cause of Baseline Noise in Next-Generation Sequencing. Mol Diagn Ther. 2014;18:587– 93.

23. Balaparya A, De S. Revisiting signatures of neutral tumor evolution in the light of complexity of cancer genomic data. Nat Genet. 2018;

24. Williams MJ, Werner B, Heide T, Barnes CP, Graham TA, Sottoriva A. Reply to “Revisiting signatures of neutral tumor evolution in the light of complexity of cancer genomic data.” Nat Genet. 2018;

25. Network TCGAR. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. N Engl J Med. 2013;368:2059–74.

26. Collisson EA, Campbell JD, Brooks AN, Berger AH, Lee W, Chmielecki J, et al. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014;511:543–50.

27. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature. 2011;474:609–15.

28. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487:330–7.

29. Hammerman PS, Lawrence MS, Voet D, Jing R, Cibulskis K, Sivachenko A, et al. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012;489:519–25.

30. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490:61–70.

31. Cancer Genome Atlas Research Network, Kandoth C, Schultz N, Cherniack AD, Akbani R, Liu Y, et al. Integrated genomic characterization of endometrial carcinoma. Nature. 2013;497:67–73.

32. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature. 2013;499:43–9.

33. Brennan CW, Verhaak RGW, McKenna A, Campos B, Noushmehr H, Salama SR, et al. The somatic genomic landscape of glioblastoma. Cell. 2013;155:462–77.

34. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. Nature. 2014;507:315–22.

35. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. Nature. 2014;513:202–9.

36. Davis CF, Ricketts CJ, Wang M, Yang L, Cherniack AD, Shen H, et al. The Somatic Genomic Landscape of Chromophobe Renal Cell Carcinoma. Cancer Cell. 2014;26:319–
30.

37. Cancer Genome Atlas Research Network. Integrated genomic characterization of papillary thyroid carcinoma. Cell. 2014;159:676–90.

38. Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. 2015;517:576–82.

39. Cancer Genome Atlas Research Network, Brat DJ, Verhaak RGW, Aldape KD, Yung WKA, Salama SR, et al. Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. N Engl J Med. 2015;372:2481–98.

40. Cancer Genome Atlas Network. Genomic Classification of Cutaneous Melanoma. Cell. 2015;161:1681–96.

41. Cancer Genome Atlas Research Network, Linehan WM, Spellman PT, Ricketts CJ, Creighton CJ, Fei SS, et al. Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. N Engl J Med. 2016;374:135–45.

42. Cancer Genome Atlas Research Network. The Molecular Taxonomy of Primary Prostate Cancer. Cell. 2015;163:1011–25.

43. Zheng S, Cherniack AD, Dewal N, Moffitt RA, Danilova L, Murray BA, et al.
Comprehensive Pan-Genomic Characterization of Adrenocortical Carcinoma. Cancer Cell. 2016;29:723–36.

44. Sauerbrei W, Schumacher M. A bootstrap resampling procedure for model building:
application to the Cox regression model. Stat Med. 1992;11:2093–109.

❖ Tables

Table 1: Distribution of the relative frequencies of the 12 substitution classes in the development dataset for the 4,908 tumors with a somatic substitution prevalence of ≥ 1 sub/Mb

	All	*Other location*	*Lung cancer*	p-value
	(n=4908)	(n=4206)	(n=702)	
%CA >12	13.47% (661)	7.01% (295)	52.14% (366)	<.0001
%GT >13	15.14% (743)	9.49% (399)	49% (344)	<.0001
%CG ≥1	82.46% (4047)	84.38% (3549)	70.94% (498)	<.0001
%GC ≥1	81.32% (3991)	83.05% (3493)	70.94% (498)	<.0001
%CT, median (IQR)	22 (15 ;30)	24 (18 ;31)	11 (0 ;16)	<.0001
%GA, median (IQR)	22 (15 ;30)	25 (18 ;32)	11 (0 ;16)	<.0001
%AC, median (IQR)	1 (0 ;2)	1 (0 ;3)	1 (0 ;1)	<.0001
%TG, median (IQR)	1 (0 ;3)	1 (0 ;3)	0 (0 ;1)	<.0001
%AG >5	43.44% (2132)	47.81% (2011)	17.24% (121)	<.0001
%TC >5	42.97% (2109)	47.0% (1977)	18.8% (132)	<.0001
%AT >3	26.67% (1309)	24.82% (1044)	37.75% (265)	<.0001
%TA >3	26.55% (1303)	24.37% (1025)	39.6% (278)	<.0001

Results presented as percentage and frequency %(n) unless otherwise indicated. IQR: interquartile range;
%CA=relative frequency of C>A substitutions (%); %GT=relative frequency of G>T substitutions (%);
%CG=relative frequency of C>G substitutions (%); %GC=relative frequency of G>C substitutions (%);
%CT=relative frequency of C>T substitutions (%); %GA=relative frequency of G>A substitutions (%);
%AC=relative frequency of A>C substitutions (%); %TG=relative frequency of T>G substitutions (%);
%AG=relative frequency of A>G substitutions (%); %TC=relative frequency of T>C substitutions (%);
%AT=relative frequency of A>T substitutions (%); %TA=relative frequency of T>A substitutions (%).

Table 2. Bivariate analysis and full multivariate model with the 12 substitution classes using logistic regression for the 4,908 tumors with a somatic substitution prevalence of ≥ 1 sub/Mb**.**

Relative frequencies of substitutions in tumors	BIVARIATE ANALYSES		FULL MUTIVARIATE MODEL	
	Odds ratio and 95% CI	p-value	Odds ratio and 95% CI	p-value
%CA >12 vs ≤12	14.44 [11.95;17.46]	<0.001	4.86 [3.59;6.58]	<0.001
%GT >13 vs ≤13	9.17 [7.66;10.98]	<0.001	1.62 [1.2;2.19]	0.001
%CG ≥1 vs 0	0.45 [0.38;0.54]	<0.001	1.52 [0.75;3.06]	0.24
%GC ≥1 vs 0	0.50 [0.42;0.60]	<0.001	1.58 [0.78;3.22]	0.20
%CT	0.84 [0.83;0.85]	<0.001	0.91 [0.88;0.93]	<0.001
%GA	0.84 [0.83;0.85]	<0.001	0.93 [0.91;0.95]	<0.001
%AC	0.67 [0.63;0.72]	<0.001	0.82 [0.76;0.90]	<0.001
%TG	0.65 [0.61;0.70]	<0.001	0.83 [0.76;0.90]	<0.001
%AG >5 vs ≤5	0.23 [0.19;0.28]	<0.001	0.43 [0.33;0.58]	<0.001
%TC >5 vs ≤5	0.26 [0.21;0.32]	<0.001	0.44 [0.33;0.59]	<0.001
%AT >3 vs ≤3	1.84 [1.55;2.17]	<0.001	0.99 [0.74;1.33]	0.96
%TA >3 vs ≤3	2.04 [1.72;2.40]	<0.001	1.15 [0.86;1.53]	0.36

CI=confidence interval; vs=versus; %CA=relative frequency of C>A substitutions (%); %GT=relative frequency of G>T substitutions (%); %CG=relative frequency of C>G substitutions (%); %GC=relative frequency of G>C substitutions (%); %CT=relative frequency of C>T substitutions (%); %GA=relative frequency of G>A substitutions (%); %AC=relative frequency of A>C substitutions (%); %TG=relative frequency of T>G substitutions (%); %AG=relative frequency of A>G substitutions (%); %TC=relative frequency of T>C substitutions (%); %AT=relative frequency of A>T substitutions (%); %TA=relative frequency of T>A substitutions (%).

- Table 3. Final logistic regression defining the EASILUNG score

	OR and 95% CI	β Coefficient	EASILUNG score points*
%CA >12 vs ≤12	5.78 [4.36;7.66]	1.7547	175
%GT >13 vs ≤13	1.86 [1.4;2.47]	0.6209	62
%CT	0.92 [0.9;0.94]	-0.0863	-9
%GA	0.94 [0.92;0.96]	-0.0631	-6
%AC	0.84 [0.77;0.92]	-0.1719	-17
%TG	0.84 [0.78;0.92]	-0.1704	-17
%AG >5 vs ≤5	0.48 [0.36;0.63]	-0.7433	-74
%TC >5 vs ≤5	0.48 [0.36;0.64]	-0.7317	-73

** β Coefficient*100 rounded to the nearest integer*
%CA=relative frequency of C>A substitutions (%); %GT=relative frequency of G>T substitutions (%); %CG=relative frequency of C>G substitutions (%); %GC=relative frequency of G>C substitutions (%); %CT=relative frequency of C>T substitutions (%); %GA=relative frequency of G>A substitutions (%); %AC=relative frequency of A>C substitutions (%); %TG=relative frequency of T>G substitutions (%); %AG=relative frequency of A>G substitutions (%); %TC=relative frequency of T>C substitutions (%); %AT=relative frequency of A>T substitutions (%); %TA=relative frequency of T>A substitutions (%).

- Table 4: Calculating the EASILUNG score with two examples from the TCGA.

EASILUNG scoring element	Substitution classes	Example 1 (1)		Example 2 (2)	
		Relative frequency	EASILUNG score points	Relative frequency	EASILUNG score points
+175 points if %CA>12	%CA	27	175	2	0
+62 points if %GT>13	%GT	19	62	2	0
-9*(%CT) points	%CT	8	-72	27	-243
-6*(%GA) points	%GA	12	-72	37	-222
-17*(%AC) points	%AC	0	0	4	-68
-17*(%TG) points	%TG	0	0	1	-17
-74 points if %AG>5	%AG	3	0	6	-74
-73 points if %TC>5	%TC	5	0	5	0
EASILUNG score			93		-624

(1) Example 1 /Tumor sample barcode= TCGA-75-7027-01A-11D-1945-08 (lung cancer).
(2) Example 2 /Tumor sample barcode= TCGA-75-7027-01A-11D-1945-08 (non-lung cancer).
%CA=relative frequency of C>A substitutions (%); %GT=relative frequency of G>T substitutions (%); %CT=relative frequency of C>T substitutions (%); %GA=relative frequency of G>A substitutions (%); %AC=relative frequency of A>C substitutions (%); %TG=relative frequency of T>G substitutions (%); %AG=relative frequency of A>G substitutions (%); %TC=relative frequency of T>C substitutions (%).

❖ Figures

Figure 1. Description of the EASILUNG score in the development dataset. This includes the 4,908 tumors with a somatic substitution prevalence of ≥ 1 sub/Mb and shows the estimated probability of lung cancer calculated from the EASILUNG score according to the values of the score. Logit (lung cancer) = 1.216+0.010 × EASILUNG score. Probability of lung cancer = 1/(1 + exp [–logit(lung cancer)]).

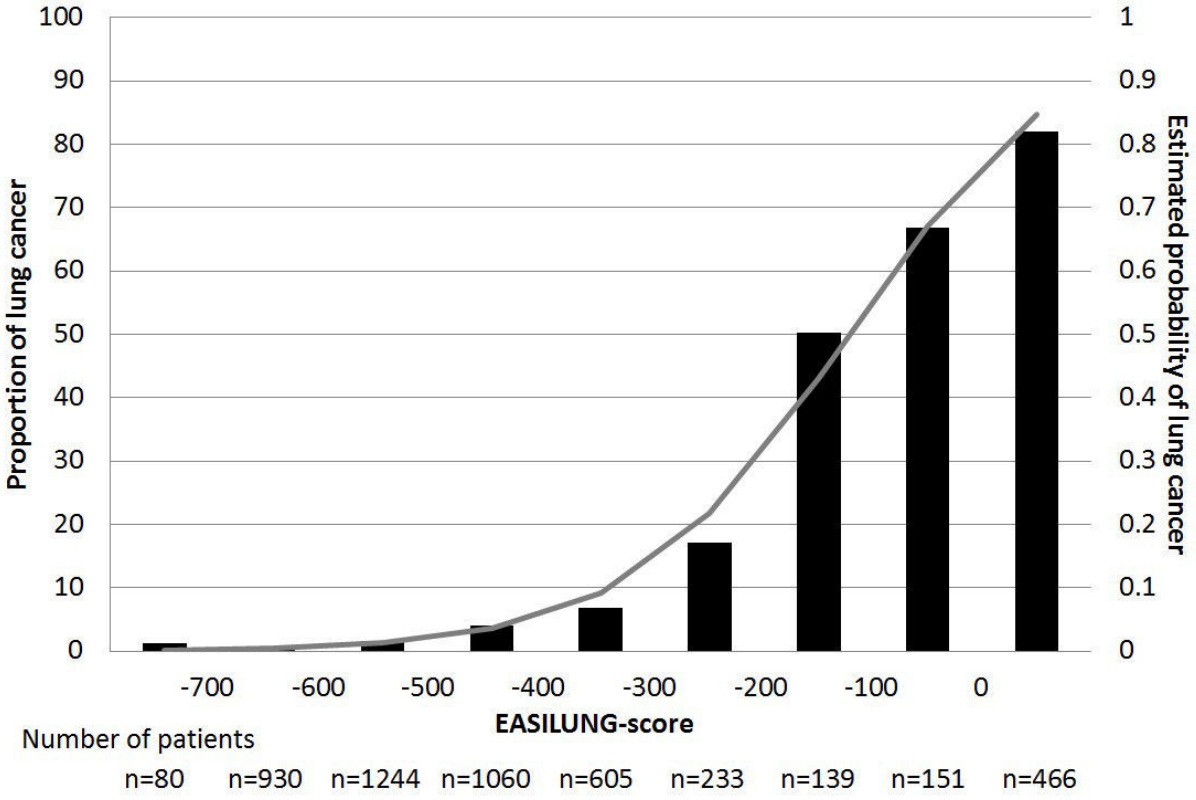
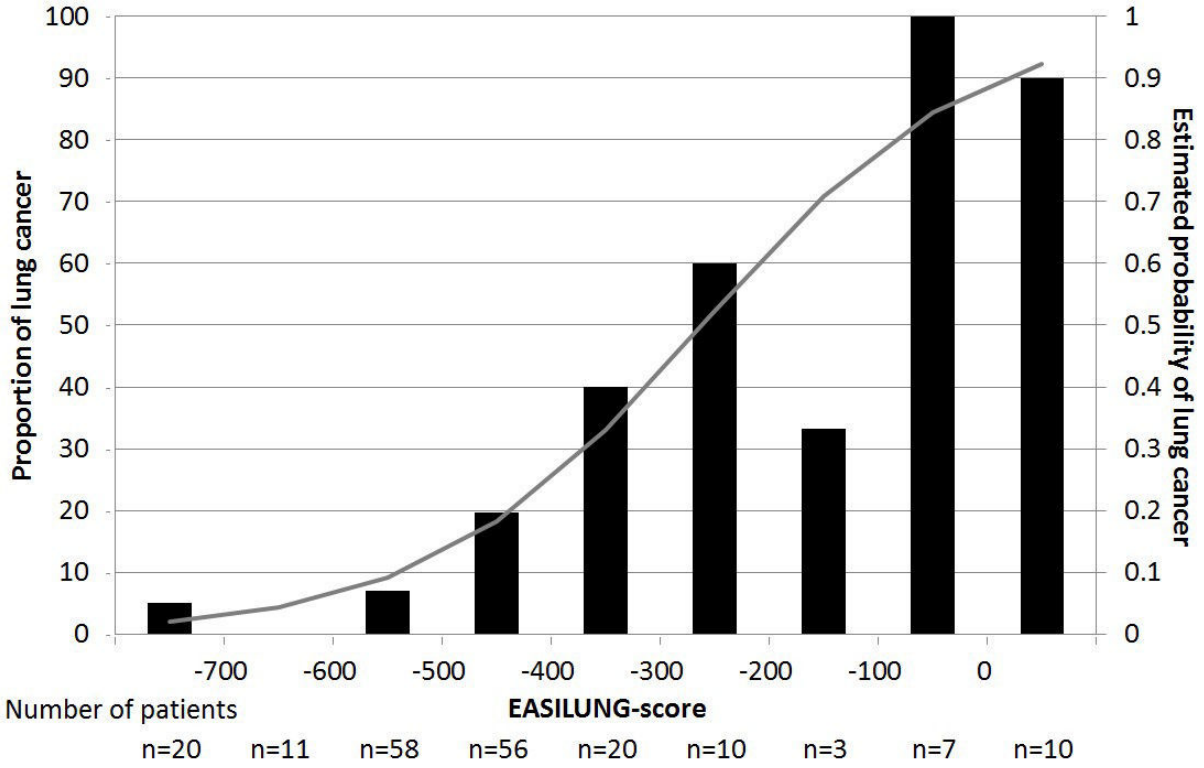


Figure 2. Description of the EASILUNG score in the validation dataset. This includes the 195 tumors with a somatic substitution prevalence of ≥ 1 sub/Mb and shows the estimated probability of lung cancer calculated from the EASILUNG score according to the values of the score. Logit (lung cancer) = 2.094+0.008 × EASILUNG score. Probability of lung cancer = 1/(1 + exp [–logit(lung cancer)]).



SUPPLEMENTARY INFORMATION

❖ Supplementary Methods

- **TCGA MAF files with targeted region size downloaded:**

Below is the table containing TCGA MAF files that have been downloaded for study.

MAF file name	File label	Exome kit targeted region size (Mb)
ACCbroadmitedu__IlluminaGA_curated_DNA_sequencing_level2.maf	ACC1	Not available
ACChgsc.bcm.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	ACC2	42
ACChgsc.bcm.edu__IlluminaGA_curated_DNA_sequencing_level2.maf	ACC3	Not available
ACChgsc.bcm.edu__Mixed_curated_DNA_sequencing_level2.maf	ACC4	Not available
BLCAbcgsc.ca__IlluminaHiSeq_automated_DNA_sequencing_level2.maf	BLCA5	Not available
BLCAbroad.mit.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	BLCA6	44
BLCAbroad.mit.edu__IlluminaGA_curated_DNA_sequencing_level2.maf	BLCA7	Not available
BLCAbroad.mit.edu__Illumina_Genome_Analyzer_DNA_Sequencing_level2.maf	BLCA8	Not available
BRCAgenome.wustl.edu__IlluminaGA_curated_DNA_sequencing_level2.maf	BRCA9	Not available
BRCAgenome.wustl.edu__IlluminaHiSeq_automated_DNA_sequencing_level2.maf	BRCA10	64
BRCAgenome.wustl.edu__Illumina_Genome_Analyzer_DNA_Sequencing_level2.maf	BRCA11	Not available
CESCbcgsc.ca__IlluminaHiSeq_automated_DNA_sequencing_level2.maf	CESC12	Not available
CESCbroad.mit.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	CESC13	Not available
CESCbroad.mit.edu__Illumina_Genome_Analyzer_DNA_Sequencing_level2.maf	CESC14	64
CESCgenome.wustl.edu__IlluminaGA_curated_DNA_sequencing_level2.maf	CESC15	Not available
CESCgenome.wustl.edu__IlluminaHiSeq_automated_DNA_sequencing_level2.maf	CESC16	Not available
CESCucsc.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	CESC17	Not available
CHOLbcgsc.ca__IlluminaHiSeq_automated_DNA_sequencing_level2.maf	CHOL18	Not available
CHOLbroad.mit.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	CHOL19	44
CHOLhgsc.bcm.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	CHOL20	Not available
CHOLhgsc.bcm.edu__Mixed_curated_DNA_sequencing_level2.maf	CHOL21	Not available
CHOLucsc.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	CHOL22	Not available
COADhgsc.bcm.edu__ABI_SOLiD_DNA_Sequencing_level2.maf	COAD23	36
COADhgsc.bcm.edu__Illumina_Genome_Analyzer_DNA_Sequencing_level2.maf	COAD24	44
DLBChgsc.bcm.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	DLBC25	42
ESCAbcgsc.ca__IlluminaHiSeq_automated_DNA_sequencing_level2.maf	ESCA26	Not available
GBMbroad.mit.edu__Illumina_Genome_Analyzer_DNA_Sequencing_level2.maf	GBM27	50
GBMmdanderson.org__Illumina_Genome_Analyzer_DNA_Sequencing_level2.maf	GBM28	50
GBMmdanderson.org__Mixed_automated_DNA_sequencing_level2.maf	GBM29	50
GBMucsc.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	GBM30	50
HNSCbcgsc.ca__IlluminaHiSeq_automated_DNA_sequencing_level2.maf	HNSC31	Not available

HNSCbroad.mit.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	HNSC32	44
HNSCbroad.mit.edu__IlluminaGA_curated_DNA_sequencing_level2.maf	HNSC33	44
HNSCbroad.mit.edu__Illumina_Genome_Analyzer_DNA_Sequencing_level2.maf	HNSC34	Not available
ESCAbroad.mit.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	HSCA35	44
ESCAgenome.wustl.edu__IlluminaHiSeq_automated_DNA_sequencing_level2.maf	HSCA36	64
ESCAhgsc.bcm.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	HSCA37	Not available
ESCAucsc.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	HSCA38	Not available
KICHbcgsc.ca__IlluminaHiSeq_automated_DNA_sequencing_level2.maf	KICH39	Not available
KICHbroad.mit.edu__Illumina_Genome_Analyzer_DNA_Sequencing_level2.maf	KICH40	Not available
KICHhgsc.bcm.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	KICH41	42
KICHhgsc.bcm.edu__Mixed_curated_DNA_sequencing_level2.maf	KICH42	42
KIRCbroad.mit.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	KIRC43	Not available
KIRChgsc.bcm.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	KIRC44	45,1
KIRChgsc.bcm.edu__Mixed_DNA_Sequencing_level2.maf	KIRC45	Not available
KIRPbroad.mit.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	KIRP46	42
KIRPbroad.mit.edu__IlluminaGA_curated_DNA_sequencing_level2.maf	KIRP47	42
KIRPbroad.mit.edu__Illumina_Genome_Analyzer_DNA_Sequencing_level2.maf	KIRP48	42
KIRPhgsc.bcm.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	KIRP49	42
KIRPhgsc.bcm.edu__IlluminaGA_curated_DNA_sequencing_level2.maf	KIRP50	42
KIRPhgsc.bcm.edu__Mixed_curated_DNA_sequencing_level2.maf	KIRP51	42
KIRPucsc.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	KIRP52	42
LAMLgenome.wustl.edu__IlluminaHiSeq_automated_DNA_sequencing_level2.maf	LAML53	44
LAMLgenome.wustl.edu__Illumina_Genome_Analyzer_DNA_Sequencing_level2.maf	LAML54	Not available
LGGbroad.mit.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	LGG55	44
LGGbroad.mit.edu__IlluminaGA_curated_DNA_sequencing_level2.maf	LGG56	44
LGGbroad.mit.edu__Illumina_Genome_Analyzer_DNA_Sequencing_level2.maf	LGG57	Not available
LGGhgsc.bcm.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	LGG58	Not available
LGGmdanderson.org__Illumina_Genome_Analyzer_DNA_Sequencing_level2.maf	LGG59	Not available
LGGmdanderson.org__Mixed_automated_DNA_sequencing_level2.maf	LGG60	Not available
LGGucsc.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	LGG61	Not available
LIHCbcgsc.ca__IlluminaHiSeq_automated_DNA_sequencing_level2.maf	LIHC62	Not available
LIHChgsc.bcm.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	LIHC63	Not available
LIHChgsc.bcm.edu__Mixed_curated_DNA_sequencing_level2.maf	LIHC64	42
LIHCucsc.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	LIHC65	Not available
LUADbroad.mit.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	LUAD66	50
LUADbroad.mit.edu__IlluminaGA_curated_DNA_sequencing_level2.maf	LUAD67	50
LUADbroad.mit.edu__Illumina_Genome_Analyzer_DNA_Sequencing_level2.maf	LUAD68	50
LUSCbroad.mit.edu__Illumina_Genome_Analyzer_DNA_Sequencing_level2.maf	LUSC69	50
MESObcgsc.ca__IlluminaHiSeq_automated_DNA_sequencing_level2.maf	MESO70	Not available
MESOsanger.ac.uk__IlluminaHiSeq_automated_DNA_sequencing_level2.maf	MESO71	50
OVbroad.mit.edu__Illumina_Genome_Analyzer_DNA_Sequencing_level2.maf	OV72	Not available
OVgenome.wustl.edu__IlluminaHiSeq_automated_DNA_sequencing_level2.maf	OV73	44
OVgenome.wustl.edu__Illumina_Genome_Analyzer_DNA_Sequencing_level2.maf	OV74	Not available
OVhgsc.bcm.edu__SOLiD_curated_DNA_sequencing_level2.maf	OV75	Not available
PAADbcgsc.ca__IlluminaHiSeq_automated_DNA_sequencing_level2.maf	PAAD76	Not available
PAADbroad.mit.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	PAAD77	44
PAADbroad.mit.edu__IlluminaGA_curated_DNA_sequencing_level2.maf	PAAD78	44

PAADbroad.mit.edu__Illumina_Genome_Analyzer_DNA_Sequencing_level2.maf	PAAD79	Not available
PAADhgsc.bcm.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	PAAD80	Not available
PAADucsc.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	PAAD81	Not available
PCPGbcgsc.ca__IlluminaHiSeq_automated_DNA_sequencing_level2.maf	PCPG82	Not available
PCPGbroad.mit.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	PCPG83	44
PCPGhgsc.bcm.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	PCPG84	Not available
PCPGucsc.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	PCPG85	Not available
PRADbroad.mit.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	PRAD86	44
PRADbroad.mit.edu__IlluminaGA_curated_DNA_sequencing_level2.maf	PRAD87	44
PRADhgsc.bcm.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	PRAD88	Not available
READhgsc.bcm.edu__ABI_SOLiD_DNA_Sequencing_level2.maf	READ89	36
READhgsc.bcm.edu__Illumina_Genome_Analyzer_DNA_Sequencing_level2.maf	READ90	44
SARCbcgsc.ca__IlluminaHiSeq_automated_DNA_sequencing_level2.maf	SARC91	Not available
SARCbroad.mit.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	SARC92	44
SARCgenome.wustl.edu__IlluminaHiSeq_automated_DNA_sequencing_level2.maf	SARC93	Not available
SARCucsc.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	SARC94	Not available
SKCMbroad.mit.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	SKCM95	44
SKCMbroad.mit.edu__Illumina_Genome_Analyzer_DNA_Sequencing_level2.maf	SKCM96	Not available
SKCMhgsc.bcm.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	SKCM97	Not available
STADbcgsc.ca__IlluminaHiSeq_automated_DNA_sequencing_level2.maf	STAD98	Not available
STADbroad.mit.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	STAD99	44
STADbroad.mit.edu__IlluminaGA_curated_DNA_sequencing_level2.maf	STAD100	Not available
STADbroad.mit.edu__Illumina_Genome_Analyzer_DNA_Sequencing_level2.maf	STAD101	Not available
STADgenome.wustl.edu__IlluminaHiSeq_automated_DNA_sequencing_level2.maf	STAD102	Not available
STADhgsc.bcm.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	STAD103	Not available
STADucsc.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	STAD104	Not available
TGCTbcgsc.ca__IlluminaHiSeq_automated_DNA_sequencing_level2.maf	TGCT105	Not available
TGCTbroad.mit.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	TGCT106	44
TGCThgsc.bcm.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	TGCT107	Not available
TGCThgsc.bcm.edu__Mixed_curated_DNA_sequencing_level2.maf	TGCT108	Not available
TGCTucsc.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	TGCT109	Not available
THCAbcgsc.ca__IlluminaHiSeq_automated_DNA_sequencing_level2.maf	THCA110	Not available
THCAbroad.mit.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	THCA111	44
THCAbroad.mit.edu__Illumina_Genome_Analyzer_DNA_Sequencing_level2.maf	THCA112	Not available
THCAhgsc.bcm.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	THCA113	Not available
THYMbcgsc.ca__IlluminaHiSeq_automated_DNA_sequencing_level2.maf	THYM114	Not available
THYMbroad.mit.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	THYM115	44
THYMgenome.wustl.edu__IlluminaHiSeq_automated_DNA_sequencing_level2.maf	THYM116	Not available
THYMhgsc.bcm.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	THYM117	Not available
THYMucsc.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	THYM118	Not available
UCECbroad.mit.edu__Illumina_Genome_Analyzer_DNA_Sequencing_level2.maf	UCEC119	Not available
UCECgenome.wustl.edu__IlluminaHiSeq_automated_DNA_sequencing_level2.maf	UCEC120	64
UCECgenome.wustl.edu__Illumina_Genome_Analyzer_DNA_Sequencing_level2.maf	UCEC121	Not available
UCSbcgsc.ca__IlluminaHiSeq_automated_DNA_sequencing_level2.maf	UCS122	Not available
UCSbroad.mit.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	UCS123	44
UCSbroad.mit.edu__IlluminaGA_curated_DNA_sequencing_level2.maf	UCS124	44
UCShgsc.bcm.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	UCS125	Not available

UVMbcgsc.ca__IlluminaHiSeq_automated_DNA_sequencing_level2.maf	UVM126	Not available
UVMbroad.mit.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	UVM127	44
UVMbroad.mit.edu__IlluminaGA_curated_DNA_sequencing_level2.maf	UVM128	44
UVMhgsc.bcm.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	UVM129	Not available
UVMucsc.edu__IlluminaGA_automated_DNA_sequencing_level2.maf	UVM130	Not available

- **Details of platforms and exome kit used for each cancer groups that have been included**

Adrenocortical carcinoma [ACC] n= **86** tumors sequenced by "BCM IlluminaGA DNASeq automated" platform using "NimbleGen VCRome 2.1" (42 MB). [Metadata associated file]

Bladder Urothelial Carcinoma [BLCA] n=**396** tumors sequenced by "BI IlluminaGA DNASeq automated" platform using "Agilent SureSelect Human All Exon v2 kit 44mb". [Metadata associated file]

Brain Lower Grade Glioma [LGG] n=**533** tumors sequenced by "BI IlluminaGA DNASeq automated" platform using "Agilent SureSelect Human All Exon v2 kit 44mb" and n=**286** tumors sequenced by "BI IlluminaGA DNASeq curated" platform using "Agilent SureSelect Human All Exon v2 kit 44mb". [Metadata associated file]

Breast invasive carcinoma [BRCA] n=**776** tumors sequenced by "WUSM IlluminaHiSeq DNASeq automated" platform using "Nimblegen SeqCap EZ Human Exome Library v3.0", the total size of the regions covered by probes is 64 Mb. [Metadata associated file]

Cervical squamous cell carcinoma and endocervical adenocarcinoma [CESC] n=**39** tumors sequenced by "BI IlluminaGA DNASeq" platform using "Nimblegen SeqCap EZ Human Exome Library v3.0 (64mb)". [Metadata associated file]

Cholangiocarcinoma [CHOL] n=**36** tumors sequenced by "BI IlluminaGA DNASeq automated" platform using "Agilent SureSelect Human All Exon v2 kit 44mb". [Metadata associated file]

Colon adenocarcinoma [COAD] n=**219** tumors sequenced by "BCM IlluminaGA DNASeq" platform using "NimbleGen SeqCap EZ Exome 2.0 Solution Probes" targeting ~44Mbs of sequence from ~30K genes, or "VCRome 2.1, HGSC design, NimbleGen" targeting 43 Mb of sequence from ~30K genes, according to the manufacturer's protocol. **53** others tumors were sequenced by "BCM SOLiD DNASeq" platform using "NimbleGen CCDS Solution Probes" which targets ~36 Mbs of sequence from ~17K genes, according to the manufacturer's protocol. [Supplementary information TCGA related publication][28]

Esophageal carcinoma [ESCA] n=**185** tumors sequenced by "BI IlluminaGA DNASeq" platform using "Agilent SureSelect Human All Exon v2 kit" (44Mb) [Metadata associated file] and **184** tumors sequenced by "WUSM

IlluminaHiSeq DNASeq automated" using "Nimblegen SeqCap EZ Human Exome Library v3.0" (64Mb). [Metadata associated file]

Glioblastoma multiforme **[GBM]** n=**282** tumors sequenced by "BI IlluminaGA DNASeq" platform, **316** tumors sequenced by "MDA IlluminaGA DNASeq" platform, **316** tumors sequenced by "MDA Mixed DNASeq automated" platform and **315** tumors sequenced by "UCSC IlluminaGA DNASeq automated platform". Massively Parallel Sequencing Exome capture was performed by using Agilent SureSelect Human All Exon 50 Mb according the manufacturer's instructions. [supplement TCGA related publication][29]

Head and Neck squamous cell carcinoma **[HNSC]** n=**270** tumors sequenced by "BI IlluminaGA DNASeq automated" platform using "Agilent SureSelect Human All Exon v2 kit" (44mb) [Metadata associated file] and **260** tumors sequenced by "BI IlluminaGA DNASeq curated" platform using "Agilent SureSelect Human All Exon v2 kit" (44mb) [Metadata associated file].

Kidney Chromophobe **[KICH]** n=**66** tumors sequenced by "BCM IlluminaGA DNASeq automated" platform and **66** tumors sequenced by "BCM Mixed DNASeq curated" platform. Massively parallel sequencing exome capture was performed using "NimbleGen (custom designed) VCRome 2.1" (42 Mb) according the manufacturer's instructions. [Metadata associated file]

Kidney renal clear cell carcinoma **[KIRC]** n=**349** tumors sequenced by "BCM IlluminaGA DNASeq automated" platform using "Nimblegen SeqCap EZ HGSC VCRome". The HGSC scientists worked with NimbleGen R&D scientists to optimize capture performance through empirical rebalancing. The name comes from the primary coding sequence (CDS) annotation sources which were Vega, CCDS, and RefSeq. Also included are miRNAs from miRBase. The design covers 23,585 genes and 189,028 non-overlapping exons (July 2014 Ensembl annotation). Design files are included for both the hg18 and hg19 genome assemblies. Total design size is 45.1 Mb capture target. [Metadata associated file]

Kidney renal papillary cell carcinoma **[KIRP]** n=**282** tumors sequenced by "BCM IlluminaGA DNASeq automated" platform, n=**168** tumors sequenced by "BCM IlluminaGA DNASeq curated" platform, n=**161** tumors sequenced by "BCM Mixed DNASeq curated" platform, 115/283 tumors sequenced by "BI IlluminaGA DNASeq" platform, **116** tumors sequenced by "BI IlluminaGA DNASeq automated" platform, **171** tumors sequenced by "BI IlluminaGA

DNASeq curated" platform, and n=**171** tumors sequenced by "UCSC IlluminaGA DNASeq automated" platform, using "NimbleGen HGSC VCRome 2.1 design" (42Mb) according to the manufacturer's protocol NimbleGen SeqCap EZ Exome Library SR User's Guide (Version 2.2). [Supplement appendix TCGA related article][30]

Liver hepatocellular carcinoma [LIHC] n=**198** tumors sequenced by "BCM Mixed DNASeq curated" Platform. Massively parallel sequencing exome capture was performed using NimbleGen (custom designed) VCRome 2.1" (42 MB) according the manufacturer's instructions. [Metadata associated file].

Lung adenocarcinoma [LUAD] n=**548** tumors sequenced by "BI IlluminaGA DNASeq" platform, n=**561** tumors sequenced by "BI IlluminaGA DNASeq automated" platform, and n=**230** tumors sequenced by "BI IlluminaGA DNASeq curated" platform. Exome capture was performed using the "Agilent SureSelect Human All Exon 50Mb kit". [supplement TCGA related publication][31]

Lung squamous cell carcinoma [LUSC] n=**178** tumors sequenced by "BI IlluminaGA DNASeq" platform. Exome capture was performed using "Agilent SureSelect Human All Exon 50 Mb" according to the manufacturers'instructions. [Metadata associated file].

Lymphoid Neoplasm Diffuse Large B-cell Lymphoma [DLBC] n= **48** tumors sequenced by "BCM IlluminaGA DNASeq automated" platform. Massively parallel sequencing exome capture was performed using "NimbleGen (custom designed) VCRome 2.1" (42 MB) according the manufacturer's instructions. [Metadata associated file]

Mesothelioma [MESO] n=**83** tumors sequenced by "SANGER IlluminaHiSeq DNASeq automated" platform using "Agilent SureSelect Human All Exon V3 kit" (50mb). [Metadata associated file]

Ovarian serous cystadenocarcinoma [OV] n=**88** tumors sequenced by "WUSM IlluminaHiSeq DNASeq automated" platform using "Agilent SureSelect Human All Exon 44 Mb v2". [Metadata associated file]

Pancreatic adenocarcinoma [PAAD] n=**185** tumors sequenced by "BI IlluminaGA DNASeq automated" platform using "Agilent SureSelect Human All Exon v2 kit" (44Mb) and n=150/186 tumors sequenced by "BI IlluminaGA DNASeq curated" platform using "Agilent SureSelect Human All Exon v2 kit" (44mb). [Metadata associated file]

Pheochromocytoma and Paraganglioma [PCPG] n=**184** tumors sequenced by "BI IlluminaGA DNASeq automated" platform using "Agilent SureSelect Human All Exon v2 kit" (44mB). [Metadata associated file]

Prostate adenocarcinoma **[PRAD]** n=**501** tumors sequenced by "BI IlluminaGA DNASeq automated" platform using "Agilent SureSelect Human All Exon v2 kit" (44mb) and n=**334** tumors sequenced by "BI IlluminaGA DNASeq curated" platform using "Agilent SureSelect Human All Exon v2 kit" (44mb). [Metadata associated file]

Rectum adenocarcinoma **[READ]** n=**81** tumors sequenced by "BCM IlluminaGA DNASeq" platform using "NimbleGen SeqCap EZ Exome 2.0 Solution Probes" targeting ~44Mbs of sequence from ~30K genes, or "VCRome 2.1 (HGSC design, NimbleGen)" targeting 43 Mb of sequence from ~30K genes, according to the manufacturer's protocol. **35** others tumors were sequenced by "BCM SOLiD DNASeq" platform using NimbleGen CCDS Solution Probes which targets ~36 Mbs of sequence from ~17K genes, according to the manufacturer's protocol. [Supplementary information TCGA related article][28]

Sarcoma **[SARC]** n=**259** tumors sequenced by "BI IlluminaGA DNASeq automated" using "Agilent SureSelect Human All Exon v2 kit" (44Mb). [Metadata associated file]

Skin Cutaneous Melanoma **[SKCM]** n=**472** tumors sequenced by "BI IlluminaGA DNASeq automated" platform using the "Agilent Sure-Select Human All Exon v2.0 kit" (44Mb). [Metadata associated file]

Stomach adenocarcinoma **[STAD]** n=**91** tumors sequenced by "BI IlluminaGA DNASeq automated" platform using "Agilent SureSelect Human All Exon v2 kit" (44mb) [Metadata associated file]

Testicular Germ Cell Tumors **[TGCT]** n=**156** tumors sequenced by "BI IlluminaGA DNASeq automated" platform using "Agilent SureSelect Human All Exon v2 kit" (44Mb). [Metadata associated file]

Thymoma **[THYM]** n=**123** tumors sequenced by "BI IlluminaGA DNASeq automated" platform using "Agilent SureSelect Human All Exon v2 kit" (44Mb). [Metadata associated file]

Thyroid carcinoma **[THCA]** n=**504** tumors sequenced by "BI IlluminaGA DNASeq automated" platform using "Agilent SureSelect Human All Exon v2 kit" (44Mb). [Metadata associated file]

Uterine Carcinosarcoma **[UCS]** n=**57** tumors sequenced by "BI IlluminaGA DNASeq automated" platform using "Agilent SureSelect Human All Exon v2 kit" (44Mb) and n=57/57 tumors sequenced by "BI IlluminaGA DNASeq curated" platform using "Agilent SureSelect Human All Exon v2 kit" (44Mb). [Metadata associated file]

Uterine Corpus Endometrial Carcinoma **[UCEC]** n=**248** tumors sequenced by "WUSM IlluminaHiSeq DNASeq automated" platform using "Nimblegen SeqCap EZ Human Exome Library v3.0" (64Mb). [Metadata associated file]

Uveal Melanoma **[UVM]** n=**80** tumors sequenced by "IlluminaGA DNASeq automated" platform using "Agilent SureSelect Human All Exon v2 kit" (44Mb) and n=80/80 tumors sequenced by "BI IlluminaGA DNASeq curated" platform using "Agilent SureSelect Human All Exon v2 kit" (44Mb). [Metadata associated file]

- **TCGA Clinical Data Elements**

bcr_patient_barcode / Tumor_Sample_Barcode

« bcr_patient_barcode » corresponds to a participant in the clinical data file « nationwidechildrens.org_clinical_patient_hnsc » (exemple for a tumor from [HNSC] group) and annotated « **TCGA-UV-WXYZ** ». This annotation can be found in the somatic mutation file (in this example HNSC maf file) in the column « Tumor_Sample_Barcode » as the prefix part of the barcode (« **TCGA-UV-WXYZ**-01A-01D-1434-08 » for example)

Histologic subgroups:

« histologic_diagnosis » refer to the histological type of a tumor, it is found in the clinical data file « nationwidechildrens.org_clinical_patient_XXXX» (XXXX refer to a TCGA group). We specify for TCGA groups (bolded) in this table. Histologic subgroups obtained are precised, unbolded, for each group:

Cervical squamous cell carcinoma and endocervical adenocarcinoma [CESC] n=39
Cervical Squamous Cell Carcinoma n=36
Endocervical Type of Adenocarcinoma n=3
Cholangiocarcinoma [CHOL] n=36
Cholangiocarcinoma; hilar/perihilar n=4
Cholangiocarcinoma; intrahepatic n=30
Cholangiocarcinoma; distal n=2
Colon adenocarcinoma [COAD] n=271*
Colon Adenocarcinoma n=236
Colon Mucinous Adenocarcinoma n=32
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma [DLBC] n=48
Diffuse large B-cell lymphoma (DLBCL) NOS (any anatomic site nodal or extranodal) n=41
Primary DLBCL of the CNS n=3
Primary mediastinal (thymic) DLBCL n=4
Oesophageal carcinoma [ESCA] n=186
Oesophagus Adenocarcinoma, NOS n=89
Oesophagus Squamous Cell Carcinoma n=97
Mesothelioma [MESO] n=82
Biphasic mesothelioma n=23
Diffuse malignant mesothelioma - NOS n=4
Epithelioid mesothelioma n=53

Sarcomatoid mesothelioma n=2
Ovarian serous cystadenocarcinoma [OV] n=88
High-grade serous carcinoma n=88
Pancreatic adenocarcinoma [PAAD] n=185*
Pancreas-Adenocarcinoma Ductal Type n=153
Pancreas-Undifferentiated Carcinoma n=1
Pancreas-Colloid (mucinous non-cystic) Carcinoma n=4
Pancreas-Adenocarcinoma-Other Subtype n=26
Pheochromocytoma and Paraganglioma [PCPG] n=184
Pheochromocytoma n=33
Paraganglioma n=151
Rectum adenocarcinoma [READ] n=115*
Rectal Adenocarcinoma n=100
Rectal Mucinous Adenocarcinoma n=10
Sarcoma [SARC] n=259
Dedifferentiated liposarcoma n=57
Desmoid Tumor n=2
Malignant Peripheral Nerve Sheath Tumors (MPNST) n=10
Myxofibrosarcoma n=25
Sarcoma; synovial; poorly differentiated n=2
Synovial Sarcoma - Biphasic n=2
Synovial Sarcoma - Monophasic n=6
Undifferentiated Pleomorphic Sarcoma (UPS) n=50**
Leiomyosarcoma (LMS) n=105
Skin Cutaneous Melanoma [SKCM] n=472
Acral lentiginous melanoma, malignant n=3
Amelanotic melanoma n=7
Epithelioid cell melanoma n=8
Lentigo maligna melanoma n=2
Malignant melanoma, NOS n=420
Mixed epithelioid and spindle cell melanoma n=1
Nodular melanoma n=22
Spindle cell melanoma, NOS n=4
Superficial spreading melanoma n=5
Testicular Germ Cell Tumors [TGCT] n=156* **
Seminoma; NOS n=66
Non-Seminoma; Yolk Sac Tumor n=3
Non-Seminoma; Embryonal Carcinoma n=16
Thyroid carcinoma [THCA] n=504
Thyroid Papillary Carcinoma - Classical/usual n=357
Thyroid Papillary Carcinoma - Follicular (≥ 99% follicular patterned) n=102
Thyroid Papillary Carcinoma - Tall Cell (≥ 50% tall cell features) n=36
Other, specify n=9
Thymoma [THYM] n=123**
Thymoma; Type C n=11
Uterine Corpus Endometrial Carcinoma [UCEC] n=248
Endometrioid endometrial adenocarcinoma n=200

Serous endometrial adenocarcinoma n=44
Mixed serous and endometrioid n=4

*Missing histology specification for some tumors in TCGA clinical data file explains differences between the number of tumors in the subgroups and the number of tumors in the TCGA group.

** Undifferentiated Pleomorphic Sarcoma (UPS) includes “Undifferentiated Pleomorphic Sarcoma (UPS)”, “Pleomorphic 'MFH' / Undifferentiated pleomorphic sarcoma” and” Giant cell 'MFH' / Undifferentiated pleomorphic sarcoma with giant cells”, due to the homogeneity of these entity, in agreement with last World Health Organisation Soft Tissue classification modification (4th edition, 2013).

***For [TCGT] group, tumors histology subgroups are unspecified because there were too many heterogeneous subgroups with low numbers of case, and low clinical relevance elsewhere. The subgroups in [TCGT] were performed to show the histologic heterogeneity of this group, the presence of seminomatous and non-seminomatous tumors, and the homogenous performance of our test on this group.

****For [THYM] group, only “Thymoma; Type C » histologic subgroup has been specified because it is the most aggressive subtype of thymoma, with a high metastatic potential. Result of our test in this subgroup is particularly relevant.

Anatomics subgroups in [HNSC] group

« anatomic_organ_subdivision » refer to the anatomical location of the tumor, it is specified in clinical data file «nationwidechildrens.org_clinical_patient_hnsc» for the [HNSC] group. Four anatomical subdivision were subgrouped:

- Oral Cavity (Alveolar Ridge + Buccal Mucosa + Floor of mouth + Hard Palate + Lip + Oral Tongue)
- Oropharynx (Base of tongue + Tonsil)
- Hypopharynx
- Larynx.

Tobacco smoking history indicator subgroups in [LUSC] and [LUAD] groups

"TOBACCO_SMOKING_HISTORY_INDICATOR" Tobacco smoking history was informed for [LUSC] and [LUAD] groups by referring to their respecting TCGA clinical data file "nationwidechildrens.org_clinical_patient_XXXX" (XXXX refer to TCGA groups, here “luad” and “lusc”)

« TOBACCO_SMOKING_HISTORY_INDICATOR » only has values from the first 1-5 conditions plus [Not Available] corresponding to:

- Lifelong Non-smoker (less than 100 cigarettes smoked in Lifetime) = 1
- Current smoker (includes daily smokers and non-daily smokers or occasional smokers) = 2
- Current reformed smoker for > 15 years (greater than 15 years) = 3
- Current reformed smoker for ≤15 years (less than or equal to 15 years) = 4
- Current reformed smoker, duration not specified = 5
- [Not Available]

Microsatellite instability status subgroups in Colorectal adenocarcinoma groups [COAD] and [READ]

Subgroups according to microsatellite instability immunohistochemistery status (Positive or Negative) was made using «microsatellite_instability» for Colorectal adenocarcinoma groups [COAD] and [READ], available in TCGA clinical data file « nationwidechildrens.org_clinical_patient_COAD» and « nationwidechildrens.org_clinical_patient_READ».

- **About local DATA (external validation):**

Eligible tumors for external validation were all consecutive tumors with somatic mutation exome data and matched clinicopathologic information available from Molecular Tumor Board of Centre Georges-François Leclerc in Dijon (France), sequenced since 2015 to august 2017 corresponding to 198 tumors. Two tumors were metastasis of an unknown primitive. All tumors and paired germline DNA were sequenced in 2 x 150 cycles on NextSeq500 using Agilent SureSelect XT Human All exon v5 kit (targeted region design = 50Mb). Samples consisting with FFPE or Snap-

frozen human tumor, spanning different tumor class (see Supplementary Table 3). Initial site and diagnoses were made by using histopathological information. Tumors were specimens from primary or metastasis site obtained before any treatment. The inclusion criteria were: consecutive whole exome sequenced tumors with available somatic single base substitutions, not otherwise specified. Selection were made by Dijon Molecular platform centre, in all examined samples, normal DNA from the same individuals had been sequenced to establish the somatic origin of variants. For each tumor barcode, a file containing two columns with allele of reference and mutated allele read on the coding non-transcribed strand for each single base substitution was sent in our department, in blind of all other parameters except the size of targeted region of exome (50Mb). The Microsoft Excel CONCATENATE function (CONCAT function since Excel 2016) allows joining two or more columns together. Substitution classes were extracted by joining the "reference allele" column with the next "mutated allele" column, corresponding to "C>A" (nucleotide C (reference allele) substituted by nucleotide A (mutated allele)), "C>T", "C>G", "G>C", "G>T", "G>A", "T>A", "T>G", "T>C", "A>T", "A>G" or "A>C"). To count the occurrence of each substitution possibilities appears in the resulting list, we use the Microsoft Excel COUNTIF statistical function. Supplementary Table 2 shows the extracted data.

❖ Supplementary Table(s)

- **Supplementary Table 1. Distribution TCGA cancer groups for development dataset.**

TCGA cancer groups	*All samples (n=7796)*	*Sample with prevalence of somatic substitutions ≥ 1 sub/Mb (n=4908)*
ACC	91 (1.17%)	61 (1.24%)
BLCA	396 (5.08%)	386 (7.86%)
BRCA	776 (9.95%)	198 (4.03%)
CESC	39 (0.5%)	37 (0.75%)
CHOL	36 (0.46%)	29 (0.59%)
COAD	271 (3.48%)	261 (5.32%)
DLBC	48 (0.62%)	46 (0.94%)
ESCA	186 (2.39%)	183 (3.73%)
GBM	362 (4.64%)	238 (4.85%)
HNSC	271 (3.48%)	259 (5.28%)
KICH	66 (0.85%)	37 (0.75%)
KIRC	349 (4.48%)	262 (5.34%)
KIRP	286 (3.67%)	224 (4.56%)
LGG	533 (6.84%)	224 (4.56%)
LIHC	198 (2.54%)	180 (3.67%)
LUAD	568 (7.29%)	526 (10.72%)
LUSC	177 (2.27%)	176 (3.59%)
MESO	82 (1.05%)	25 (0.51%)
OV	88 (1.13%)	67 (1.37%)
PAAD	185 (2.37%)	156 (3.18%)
PCPG	184 (2.36%)	4 (0.08%)
PRAD	499 (6.4%)	200 (4.07%)
READ	115 (1.48%)	111 (2.26%)
SARC	259 (3.32%)	198 (4.03%)

SKCM	472 (6.05%)	443 (9.03%)
STAD	91 (1.17%)	87 (1.77%)
TGCT	156 (2%)	38 (0.77%)
THCA	504 (6.46%)	43 (0.88%)
THYM	123 (1.58%)	19 (0.39%)
UCEC	248 (3.18%)	138 (2.81%)
UCS	57 (0.73%)	51 (1.04%)
UVM	80 (1.03%)	1 (0.02%)

- **Supplementary Table 2. Summary of EASILUNG test performance on TCGA cancer groups and subgroups (development dataset)**

TCGA CANCER GROUPS AND SUBGROUPS	TOTAL CASES	EASILUNG TEST POSITIVE
ADENOCARCINOMA CANCER GROUPS AND SUBGROUPS		
Lung adenocarcinoma [LUAD]	568	447(79%)
Current reformed smoker for ≤15 years [LUAD subgroup]	171	159(93%)
Current smoker (includes daily smokers and non-daily smokers or occasional smokers) [LUAD subgroup]	116	105(91%)
Current reformed smoker for > 15 years [LUAD subgroup]	136	104(76%)
Current reformed smoker, duration not specified [LUAD subgroup]	3	2(67%)
Lifelong Non-smoker (less than 100 cigarettes smoked in Lifetime) [LUAD subgroup]	77	29(38%)
Cases without clinical data [LUAD subgroup]	52	42(81(%)
Tobacco smoking history status [Not Available] [LUAD subgroup]	12	6(50%)
Tobacco smoking history status [Unknown] [LUAD subgroup]	1	0(0%)
Breast invasive carcinoma [BRCA]	776	31(4%)
RH-(ER- and PR-) Breast carcinoma [BRCA immunophenotypical subgroup]	150	14(9%)
Endocervical Type of Adenocarcinoma [CESC subgroup]	3	3(100%)
Cholangiocarcinoma [CHOL]	36	1(3%)
Cholangiocarcinoma; hilar/perihilar [CHOL subgroup]	4	0(0%)
Cholangiocarcinoma; intrahepatic [CHOL subgroup]	30	1(3%)
Cholangiocarcinoma; distal [CHOL subgroup]	2	0(0%)
Colon adenocarcinoma [COAD]	271	6(2%)
Colon Adenocarcinoma [COAD subgroup]	236	5(2%)
Colon Mucinous Adenocarcinoma [COAD subgroup]	32	0(0%)
With microsatellite instability [COAD immunophenotypical subgroup]	9	0(0%)
Esophagus Adenocarcinoma, NOS [ESCA subgroup]	89	10(11%)
Kidney Chromophobe [KICH]	66	0(0%)
Kidney renal clear cell carcinoma [KIRC]	349	35(10%)
Kidney renal papillary cell carcinoma [KIRP]	286	37(13%)
Liver hepatocellular carcinoma [LIHC]	198	45(23%)
Hepatitis B positive [LIHC virological subgroup]	19	4(21%)
Hepatitis C positive [LIHC virological subgroup]	32	7(22%)
Hepatitis B and Hepatitis C positive [LIHC virological subgroup]	3	1(33%)
Ovarian high-grade serous cystadenocarcinoma [OV]	88	14(16%)
Pancreatic adenocarcinoma [PAAD]	185	4(2%)
Pancreas-Adenocarcinoma Ductal Type [PAAD subgroup]	153	3(2%)
Pancreas-Undifferentiated Carcinoma [PAAD subgroup]	1	0(0%)
Pancreas-Colloid (mucinous non-cystic) Carcinoma [PAAD subgroup]	4	0(0%)
Prostate adenocarcinoma [PRAD]	499	10(2%)
Rectum adenocarcinoma [READ]	115	5(4%)
Rectal Adenocarcinoma [READ subgroup]	100	5(5%)
Rectal Mucinous Adenocarcinoma [READ subgroup]	10	0(0%)
Stomach adenocarcinoma [STAD]	91	0(0%)
Thyroid carcinoma [THCA]	504	5(1%)
Thyroid Papillary Carcinoma - Classical/usual [THCA subgroup]	357	5(1%)

Thyroid Papillary Carcinoma - Follicular ≥ 99% follicular patterned) [THCA subgroup]	102	0(0%)
Thyroid Papillary Carcinoma - Tall Cell (≥ 50% tall cell features) [THCA subgroup]	36	0(0%)
Uterine Corpus Endometrial Carcinoma [UCEC]	248	11(4%)
Endometrioid endometrial adenocarcinoma [UCEC subgroup]	200	10(5%)
Serous endometrial adenocarcinoma [UCEC subgroup]	44	1(2%)
SQUAMOUS CELL CARCINOMA CANCER GROUPS AND SUBGROUPS		
Lung squamous cell carcinoma [LUSC]	177	149(84%)
Current smoker (includes daily smokers and non-daily smokers or occasional smokers) [LUSC subgroup]	28	24(86%)
Current reformed smoker for ≤15 years [LUSC subgroup]	100	85(85%)
Lifelong Non-smoker (less than 100 cigarettes smoked in Lifetime) [LUSC subgroup]	6	5(83%)
Current reformed smoker for > 15 years [LUSC subgroup]	38	30(79%)
tobacco smoking history status [Not Available] [LUSC subgroup]	5	0(0%)
Cervical Squamous Cell Carcinoma [CESC subgroup]	36	34(94%)
Esophagus Squamous Cell Carcinoma [ESCA subgroup]	97	28(29%)
Head and Neck squamous cell carcinoma [HNSC]	271	39(14%)
Larynx [HNSC anatomical subgroup]	61	30(49%)
Hypopharynx [HNSC anatomical subgroup]	3	0(0%)
Oropharynx [HNSC anatomical subgroup]	27	2(7%)
Oropharynx p16+ [HNSC immunophenotypical subgroup]	6	0(0%)
Oral Cavity [HNSC anatomical subgroup]	153	6(4%)
NON-SQUAMOUS CELL AND NON-ADENOCARCINOMA CANCER GROUPS AND SUBGROUPS		
Adrenocortical carcinoma [ACC]	91	9(10%)
Bladder Urothelial Carcinoma [BLCA]	396	22(6%)
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma [DLBC]	48	0(0%)
Diffuse large B-cell lymphoma NOS [DLBC subgroup]	41	0(0%)
Primary DLBCL of the CNS [DLBC subgroup]	3	0(0%)
Primary mediastinal (thymic) DLBCL [DLBC subgroup]	4	0(0%)
Glioblastoma multiforme [GBM]	362	8(2%)
Brain Lower Grade Glioma [LGG]	533	6(1%)
Mesothelioma [MESO]	82	4(5%)
Biphasic mesothelioma [MESO subgroup]	23	1(4%)
Diffuse malignant mesothelioma - NOS [MESO subgroup]	4	0(0%)
Epithelioid mesothelioma [MESO subgroup]	53	3(6%)
Sarcomatoid mesothelioma [MESO subgroup]	2	0(0%)
Pheochromocytoma and Paraganglioma [PCPG]	184	0(0%)
Paraganglioma [PCPG subgroup]	151	0(0%)
Pheochromocytoma [PCPG subgroup]	33	0(0%)
Sarcoma [SARC]	259	25(10%)
Dedifferentiated liposarcoma [SARC subgroup]	57	6(11%)
Desmoid Tumor [SARC subgroup]	2	0(0%)
Leiomyosarcoma (LMS)	105	9(9%)
Malignant Peripheral Nerve Sheath Tumors [SARC subgroup]	10	1(10%)
Myxofibrosarcoma [SARC subgroup]	25	4(16%)
Sarcoma; synovial; poorly differentiated [SARC subgroup]	2	0(0%)
Synovial Sarcoma - Biphasic [SARC subgroup] n=2	2	0(0%)
Synovial Sarcoma - Monophasic [SARC subgroup]	6	1(17%)
Undifferentiated Pleomorphic Sarcoma [SARC subgroup]	50	4(8%)
Skin Cutaneous Melanoma [SKCM]	472	4(1%

Amelanotic melanoma [SKCM subgroup]	7	0(0%)
Epithelioid cell melanoma [SKCM subgroup]	8	0(0%)
Lentigo maligna melanoma [SKCM subgroup]	2	0(0%)
Malignant melanoma, NOS [SKCM subgroup]	420	4(1%)
Mixed epithelioid and spindle cell melanoma [SKCM subgroup]	1	0(0%)
Nodular melanoma [SKCM subgroup]	22	0(0%)
Spindle cell melanoma, NOS [SKCM subgroup]	4	0(0%)
Superficial spreading melanoma [SKCM subgroup]	5	0(0%)
Testicular Germ Cell Tumors [TGCT]	156	4(3%)
Seminoma; NOS [TGCT subgroup]	66	2(3%)
Non-Seminoma; Yolk Sac Tumor [TGCT subgroup]	3	0(0%)
Non-Seminoma; Embryonal Carcinoma [TGCT subgroup]	16	0(0%)
Thymoma [THYM]	123	2(2%)
Thymoma; Type C [THYM subgroup]	11	0(0%)
Uterine Carcinosarcoma [UCS]	57	3(5%)
Uveal Melanoma [UVM]	80	0(0%)

Subgroup are extracted from TCGA clinical data related to each TCGA group below : Breast invasive carcinoma [BRCA]; Cervical squamous cell carcinoma and endocervical adenocarcinoma [CESC]; Cholangiocarcinoma [CHOL]; Colon adenocarcinoma [COAD]; Oesophageal carcinoma [ESCA]; Kidney Chromophobe [KICH]; Kidney renal clear cell carcinoma [KIRC]; Kidney renal papillary cell carcinoma [KIRP]; Liver hepatocellular carcinoma [LIHC]; Lung adenocarcinoma [LUAD]; Ovarian serous cystadenocarcinoma [OV]; Pancreatic adenocarcinoma [PAAD]; Prostate adenocarcinoma [PRAD]; Rectum adenocarcinoma [READ]; Stomach adenocarcinoma [STAD]; Thyroid carcinoma [THCA]; Uterine Corpus Endometrial Carcinoma [UCEC].Cervical squamous cell carcinoma and endocervical adenocarcinoma [CESC]; Oesophageal carcinoma [ESCA]; Head and Neck squamous cell carcinoma [HNSC]; Lung squamous cell carcinoma [LUSC].Adrenocortical carcinoma [ACC]; Bladder Urothelial Carcinoma [BLCA]; Brain Lower Grade Glioma [LGG]; Glioblastoma multiforme [GBM]; Lymphoid Neoplasm Diffuse Large B-cell Lymphoma [DLBC]; Mesothelioma [MESO]; Pheochromocytoma and Paraganglioma [PCPG]; Sarcoma [SARC]; Skin Cutaneous Melanoma [SKCM]; Testicular Germ Cell Tumors [TGCT]; Thymoma [THYM]; Uterine Carcinosarcoma [UCS]; Uveal Melanoma [UVM].

- **Supplementary Table 3: Characteristics of the validation dataset (196 samples)**

PRIMITIVE SITE	
Lung	24.49% (48)
Breast	19.39% (38)
Colon	11.73% (23)
Pancreas	10.71% (21)
Ovary	8.16% (16)
Cervix	3.06% (6)
Rectum	3.06% (6)
Endometrium	2.55% (5)
Stomach	2.55% (5)
Bile duct	1.53% (3)
Prostate	1.53% (3)
Anus	1.02% (2)
Brain	1.02% (2)
Gall bladder	1.02% (2)
Liver	1.02% (2)
Muscle	1.02% (2)
Renal	1.02% (2)
Uterus	1.02% (2)
Bladder	0.51% (1)
Oesophagus	0.51% (1)
Ovary or Peritoneum	0.51% (1)
Pelvis	0.51% (1)
Peritoneum	0.51% (1)
Soft tissue (Muscle)	0.51% (1)
Testicle	0.51% (1)

Vagina	0.51% (1)
TCGA RELATED GROUP	
BRCA	21.11% (38)
LUAD	19.44% (35)
COAD	12.22% (22)
PAAD	11.67% (21)
OV	6.67% (12)
LUSC	5.56% (10)
CESC	3.89% (7)
READ	3.33% (6)
UCEC	3.33% (6)
STAD	2.78% (5)
SARC	2.22% (4)
CHOL	1.67% (3)
GBM	1.11% (2)
LIHC	1.11% (2)
PRAD	1.11% (2)
BLCA	0.56% (1)
ESCA	0.56% (1)
KIRC	0.56% (1)
MESO	0.56% (1)
TGCT	0.56% (1)
TYPE OF SAMPLE	
FFPE	82.14% (161)
Frozen	17.86% (35)

- **Supplementary Table 4: Distribution of the relative frequencies of the 12 substitution classes in the validation dataset**

	ALL (N=196)	***OTHER LOCATION*** (N=148)	***LUNG CANCER*** (N=48)	**P-VALUE**
%CA >12	20.9%(41)	10.8%(16)	52.1%(25)	<.0001
%GT >13	17.9%(35)	6.8%(10)	52.1%(25)	<.0001
%CG ≥1	99.5%(195)	99.3%(147)	100%(48)	1
%GC ≥1	100%(196)	100%(148)	100%(48)	1
%CT, median (IQR)	13(10 ;17)	14(10 ;18)	12(10 ;16)	.042
%GA, median (IQR)	15(11 ;19)	16(12 ;20)	13(11 ;15)	.004
%AC, median (IQR)	4(2 ;7)	4(3 ;8)	3(1 ;4)	<.0001
%TG, median (IQR)	5(2 ;8)	5(3 ;10)	2(1 ;5)	<.0001
%AG >5	60.2%(118)	60.1%(89)	60.4%(29)	.972
%TC >5	53.1%(104)	53.4%(79)	52.1%(25)	.976
%AT >3	49.5%(97)	44.6%(66)	64.6%(31)	<.0001
%TA >3	71.9%(141)	68.2%(101)	83.3%(40)	.016

Results presented as percentage and frequency %(n) unless otherwise indicated. IQR: interquartile range; %CA=relative frequency of C>A substitutions (%); %GT=relative frequency of G>T substitutions (%); %CG=relative frequency of C>G substitutions (%); %GC=relative frequency of G>C substitutions (%); %CT=relative frequency of C>T substitutions (%); %GA=relative frequency of G>A substitutions (%); %AC=relative frequency of A>C substitutions (%); %TG=relative frequency of T>G substitutions (%); %AG=relative frequency of A>G substitutions (%); %TC=relative frequency of T>C substitutions (%); %AT=relative frequency of A>T substitutions (%); %TA=relative frequency of T>A substitutions (%).

- **Supplementary Table 5. Summary of EASILUNG test performance on external validation dataset (EASILUNG score thresholded at -467)**

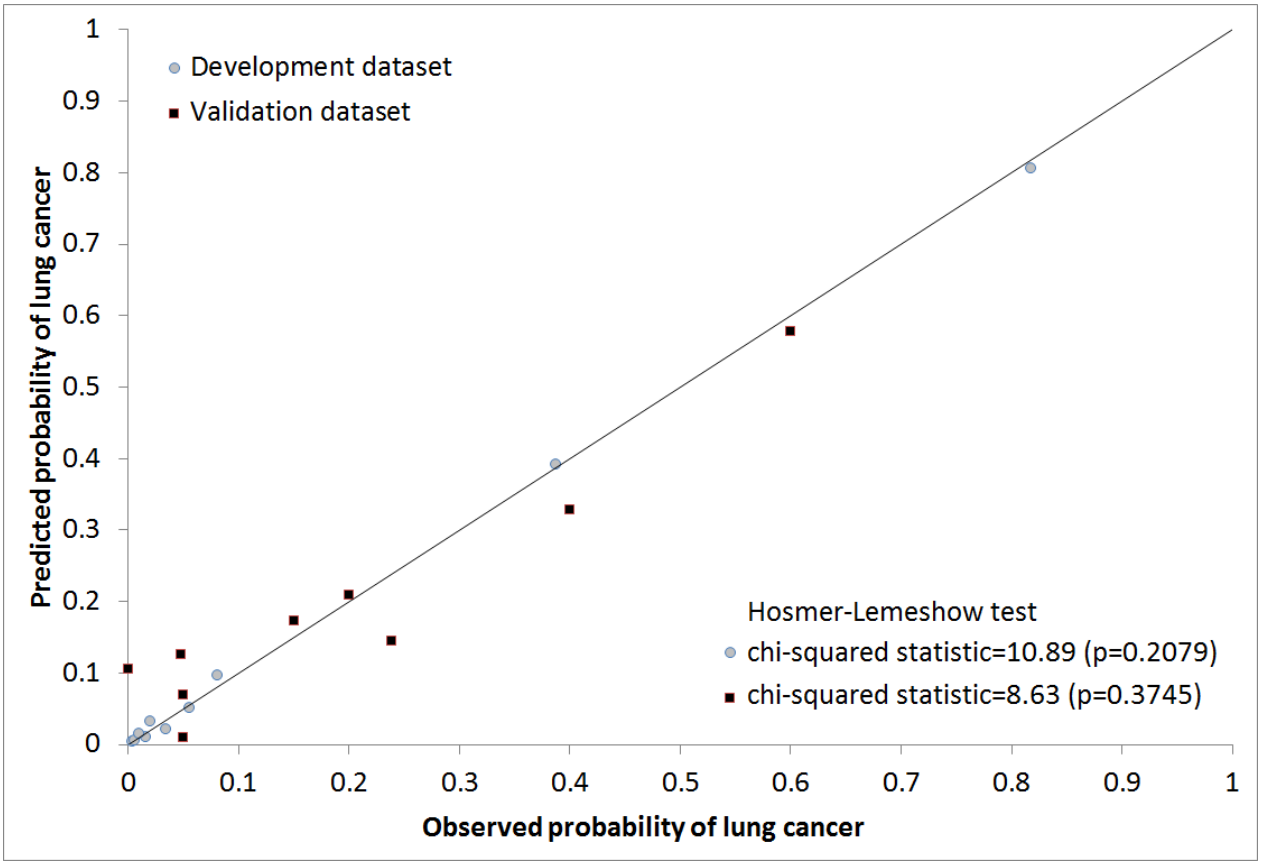
PRIMITIVE SITE	HISTOLOGY	TCGA RELATED GROUP	TYPE OF SAMPLE	TOTAL CASES	EASILUNG TEST POSITIVE
Lung	**All**		**FFPE**	**34**	**30(88%)**
Lung	**All**		**Frozen**	**14**	**9(64%)**
Lung	Adenocarcinoma	LUAD	FFPE	23	19(83%)
Lung	Adenocarcinoma	LUAD	Frozen	11	7(64%)
Lung	Large-Cell Neuroendocrine Carcinoma		Frozen	1	1(100%)
Lung	Neuroendocrine carcinoma		Frozen	1	1(100%)
Lung	Small cell carcinoma		FFPE	1	1(100%)
Lung	Squamous cell carcinoma	LUSC	FFPE	10	10(100%)
Lung	Squamous cell carcinoma	LUSC	Frozen	1	0(0%)
Not Lung	**All**		**FFPE**	**127**	**36(28%)**
Not Lung	**All**		**Frozen**	**21**	**4(19%)**
Anus	Squamous cell carcinoma		FFPE	1	1(100%)
Anus	Squamous cell carcinoma		Frozen	1	0(0%)
Bile duct	Cholangiocarcinoma	CHOL	FFPE	2	0(0%)
Bile duct	Cholangiocarcinoma	CHOL	Frozen	1	0(0%)
Bladder	Urothelial carcinoma	BLCA	FFPE	1	1(100%)
Brain	Glioblastoma	GBM	FFPE	2	1(50%)
Breast	Adenocarcinoma	BRCA	FFPE	23	6(26%)
Breast	Adenocarcinoma	BRCA	Frozen	12	3(25%)
Breast	Carcinoma (NOS)	BRCA	FFPE	2	0(0%)
Breast	Carcinoma (NOS)	BRCA	Frozen	1	0(0%)
Cervix	Squamous cell carcinoma	CESC	FFPE	6	4(67%)
Colon	Adenocarcinoma	COAD	FFPE	22	2(9%)
Endometrium	Adenocarcinoma (NOS)	UCEC	FFPE	1	1(100%)
Endometrium	Endometrioid carcinoma	UCEC	FFPE	3	2(67%)
Endometrium	Serous carcinoma	UCEC	FFPE	1	1(100%)
Gall bladder	Adenocarcinoma	CHOL	FFPE	1	0(0%)
Liver	Hepatocellular carcinoma	LIHC	FFPE	2	0(0%)
Oesophagus	Signet-ring cell carcinoma	ESCA	FFPE	1	0(0%)
Ovary	Adenocarcinoma		FFPE	1	0(0%)
Ovary	Clear cell carcinoma		FFPE	2	0(0%)
Ovary	High-grade serous carcinoma	OV	FFPE	11	3(27%)
Ovary	High-grade serous carcinoma	OV	Frozen	1	0(0%)
Ovary	Mixed serous (50%)/ endometrioid (50%) ovarian carcinoma		FFPE	1	0(0%)
Ovary or Peritoneum	High-grade serous carcinoma		FFPE	1	0(0%)
Pancreas	Adenocarcinoma	PAAD	FFPE	18	6(33%)
Pancreas	Adenocarcinoma	PAAD	Frozen	2	0(0%)

Pancreas	Adenosquamous carcinoma		FFPE	1	0(0%)
Pelvis	Malignant peripheral nerve sheath tumor (MPNST)		FFPE	1	1(100%)
Peritoneum	Epithelioid Mesothelioma	MESO	FFPE	1	0(0%)
Peritoneum	Malignant Perivascular Epithelioid Cell Neoplasm (PEComa)		FFPE	1	0(0%)
Prostate	Adenocarcinoma	PRAD	FFPE	2	0(0%)
Prostate	Large-Cell Neuroendocrine Carcinoma		FFPE	1	1(100%)
Rectum	Adenocarcinoma	READ	FFPE	5	3(60%)
Rectum	Adenocarcinoma	READ	Frozen	1	0(0%)
Renal	Clear cell renal cell carcinoma	KIRC	FFPE	1	1(100%)
Renal	Sarcomatoid Carcinoma		Frozen	1	1(100%)
Soft tissue	Embryonal Rhabdomyosarcoma		FFPE	1	0(0%)
Soft tissue	Leiomyosarcoma	SARC	FFPE	2	1(50%)
Soft tissue	Undifferentiated high-grade spindle cell sarcoma		FFPE	1	1(100%)
Stomach	Adenocarcinoma	STAD	FFPE	5	0(0%)
Testicle	Malignant Leydig cell tumor		Frozen	1	0(0%)
Uterus	Endometrioid carcinoma	UCEC	FFPE	1	0(0%)
Uterus	Squamous cell carcinoma		FFPE	1	0(0%)
Vagina	Squamous cell carcinoma		FFPE	1	0(0%)
Unknown	Adenocarcinoma		FFPE	1	0(0%)
Unknown	Carcinoma		FFPE	1	0(0%)

TCGA related group refer to TCGA group below : Bladder Urothelial Carcinoma [BLCA]; Breast invasive carcinoma [BRCA]; Cervical squamous cell carcinoma and endocervical adenocarcinoma [CESC]; Cholangiocarcinoma [CHOL]; Colon adenocarcinoma [COAD]; Oesophageal carcinoma [ESCA]; Glioblastoma multiform [GBM]; Kidney renal clear cell carcinoma [KIRC]; Liver hepatocellular carcinoma [LIHC]; Lung adenocarcinoma [LUAD]; Lung squamous cell carcinoma [LUSC]; Mesothelioma [MESO]; Ovarian serous cystadenocarcinoma [OV]; Pancreatic adenocarcinoma [PAAD]; Prostate adenocarcinoma [PRAD]; Rectum adenocarcinoma [READ]; Sarcoma [SARC]; Stomach adenocarcinoma [STAD]; Uterine Corpus Endometrial Carcinoma [UCEC].

❖ Supplementary Figure

Supplementary Figure 1. Calibration plot of observed versus predicted probability of lung cancer. Hosmer-Lemeshow chi-squared statistic is shown for the risk score in both development and validation dataset. The points and squares indicate the observed frequencies by decile of predicted probability



DISCUSSION ET CONCLUSION

La pratique du diagnostic anatomopathologique repose sur l'évaluation des caractéristiques macroscopiques et microscopiques des tissus, intégrés à l'histoire clinique, radiologique et à d'autres données biologiques, pour atteindre un diagnostic définitif, attribuer un pronostic, et assister le clinicien pour choisir la thérapie la plus appropriée. L'examen histologique des tissus a occupé un rôle central dans la détermination de la malignité, la classification des tumeurs et la détermination de leur grade. Cette approche diagnostique a depuis longtemps fait ses preuves. En outre, les procédures sont bien établies, l'acquisition de données nécessaire au diagnostic peut être rapide (exemple de l'examen extemporané sur coupes congelées) et reste relativement peu coûteuse. Pourtant, il existe des limites incontestables à l'analyse morphologique seule, par manque de caractéristiques morphologiques utiles pour distinguer certains sous-groupes de cancers ayant des réponses cliniques différentes aux traitements. L'immunohistochimie joue un rôle complémentaire dans le diagnostic anatomopathologique de routine, facilitant la distinction entre les tumeurs présentant des caractéristiques histologiques qui se chevauchent et permettant une reclassification plus fine de certains cancers en fonction de l'expression de protéines spécifiques. L'analyse génétique moléculaire des tissus est la dernière innovation à avoir eu une incidence sur le diagnostic anatomopathologique final. Certaines des techniques dans ce domaine sont déjà appliquées en routine (FISH, PCR, panel de gènes ciblés en NGS,), tandis que d'autres (séquençage exomique, séquençage génomique) offrent la possibilité de conduire à la découverte de nouveaux marqueurs et éventuellement à de nouvelles méthodes diagnostiques. Notre travail s'inscrit dans cette dernière optique, relative à la découverte de nouvelles méthodes diagnostiques, puisque de façon innovante, nous utilisons des résultats de séquençage exomique (fréquences relatives des substitutions somatiques) pour discriminer les cancers d'origine pulmonaire de nombreux autres groupes de cancer. En plus d'être innovante, elle trouve une application potentielle dans une problématique complexe et cliniquement pertinente, celle des CAPI, puisque ranger une présentation clinique en CAPI ou faire une hypothèse diagnostique est plus qu'une querelle sémantique, cela a des conséquences thérapeutiques.

- **Diversité et hétérogénéité des CAPI**

La diversité des sous-types histopathologiques et de leur localisation primitive fait des CAPI une entité hétérogène. Nous avons cherché à être représentatif de cette diversité par la variété des origines anatomiques (diversité

intergroupe) et dans la variété des sous-groupes (variété intra-groupe) par la présence dans notre étude de nombreux sous-types histologiques, sous-groupes immunophénotypiques (ex : cancers du sein n'exprimant pas les récepteurs hormonaux, cancers du côlon et du rectum avec instabilité des microsatellites) et sous-groupes anatomiques (cancers de la tête et du cou). Le nombre important de tumeurs analysées (environ 8000) répartis dans les groupes et sous-groupes d'intérêt représente une des forces principales de notre étude, puisque c'est un gage important de variété intra-groupe et de représentativité.

- **Les CAPI, un problème diagnostique en carence de solution.**

Les CAPI posent au pathologiste une problématique complexe, car il doit porter le diagnostic le plus précis possible sur un matériel généralement exigu. En outre, la multiplication des anticorps et des techniques de biologie moléculaire en cas de cancer éligible aux thérapies ciblées complexifie la démarche diagnostique. Cette démarche diagnostique classique des CAPI a pour but essentiel d'identifier les entités anatomocliniques de CAPI sensibles à un traitement connu. En dehors de ces entités anatomocliniques, la recherche de la tumeur primitive n'a pas de conséquence pronostique ou thérapeutique et un bilan exhaustif systématique est inutile et onéreux [32]. Cependant, le choix thérapeutique et les quelques cas de meilleur pronostic justifient une précision histopathologique qu'apporte souvent l'étude immunohistochimique, et plus récemment, l'analyse moléculaire [33,34]. Parmi les métastases d'une tumeur d'origine épithéliale, l'étude histopathologique permettra de distinguer trois sous-types histopathologiques principaux : un adénocarcinome (environ 50 % des cas), un carcinome peu différencié (environ 30 % des cas), ou un carcinome épidermoïde (environ 15 % des cas). Concernant les adénocarcinomes, le profil immunohistochimique observé avec les cytokératines CK7 et CK20 permet, dans un second temps, une orientation grossière vers une origine primitive. Le profil immunohistochimique CK7+/CK20- d'un adénocarcinome est fortement évocateur d'une origine pulmonaire [35], biliaire [36], mammaire [37], endocervicale ou endométriale [38], œsophagienne [39], salivaire [40], thyroïdienne [41] ou encore mésothéliale [42]. L'expression du TTF1 est classique et spécifique des adénocarcinomes d'origine pulmonaire, TTF1 est également exprimé dans les carcinomes thyroïdiens, bien que l'immunophénotype (PAX8+/thyroglobuline+) et la morphologie des carcinomes papillaires thyroïdiens permettent aisément de différencier ces deux entités en pratique. La complexité du diagnostic peut être aggravée par les réactivités croisées de certains anticorps liés à des épitopes communs à des sites antigéniques de cellules différentes ; par exemple, des antigènes épithéliaux (cytokératines AE1/AE3, EMA) peuvent être exprimés par des

sarcomes, et exceptionnellement par des mélanomes ou même des lymphomes, ce qui peut égarer le diagnostic. D'autre part, certains anticorps caractéristiques d'une tumeur primitive peuvent s'épuiser dans les métastases ou ne pas être présente comme attendu classiquement. Si un site important de la protéine est muté, l'expression ou la réaction antigène-anticorps permettant le démasquage par un chromogène pourra être altéré, et donc l'absence d'expression immunohistochimique pourra induire en erreur le pathologiste. Ainsi, environ 12,8% des adénocarcinomes pulmonaires sont TTF1-, en particulier chez les gros fumeurs [43]. De façon intéressante, c'est dans les sous-groupes d'adénocarcinomes pulmonaires chez les patients fumeurs que notre test marche le mieux, cela s'explique par une accumulation plus importante de mutations et donc une probabilité plus forte de retrouver l'empreinte substitutionnelle des carcinogènes du tabac au sein de ces tumeurs. Les cancers du sein n'exprimant ni les récepteurs aux œstrogènes (RO-) ni les récepteurs à la progestérone (RP-) en immunohistochimie sont aisément discriminé par notre test, ce qui pourrait s'avérer d'autant plus intéressant puisqu'ils ont un profil CK7+/CK20- similaire au poumon, conduisant dans certains cas le pathologiste vers une impasse diagnostique entre un cancer du poumon et un cancer du sein (adénocarcinome CK7+/CK20-/TTF1-/RH- chez une femme fumeuse par exemple) qui sont par ailleurs les cancers parmi les plus fréquents. Une étude immunohistochimique extensive complétée par des anticorps plus ou moins spécifiques de certains types tumoraux est souvent nécessaire pour en préciser l'origine. Ces dernières années, de nouveaux anticorps ont été développés, essentiellement dirigés contre des facteurs de transcription, et qui ont apporté pour certains, en combinaison avec les anticorps plus « classiques », une amélioration de la spécificité des propositions diagnostiques du pathologiste [44, 45]. Cependant, il n'y a pas de morphologie, de coloration ou d'examen immunohistochimique intéressant qui soit utile pour déterminer l'origine d'un carcinome épidermoïde (poumon, tête et cou, col de l'utérus, peau, vulve ou canal anal) en dehors du cas particulier de la métastase ganglionnaire cervical pour laquelle le profil p16/EBV/HPV peut orienter vers un carcinome du nasopharynx p16-, EBV+ ou de l'oropharynx (amygdale) p16+, HPV+. En immunohistochimie, les anticorps p40, p63 et CK5/6 sont positifs dans la plupart des carcinomes épidermoïdes et l'anticorps dirigé contre la p16 a une utilité très limitée dans la détermination du site primitif des carcinomes épidermoïdes [46], puisque non strictement spécifique à la présence de HPV. Par ailleurs, d'autres situations diagnostiques sont tout aussi problématiques car le cancer du poumon est non seulement un grand pourvoyeur de métastases, mais aussi un site commun de drainage des métastases. Les grandes séries d'autopsies de patients atteints de tumeurs malignes extra-thoraciques révèlent des métastases pulmonaires chez 20 à 54% des patients [47, 48, 49]. Parmi les cas d'autopsie, le

sein, le côlon, le rein, l'utérus et les cancers ORL sont les origines les plus fréquemment retrouvées. Les choriocarcinomes, les sarcomes, les tumeurs testiculaires, les mélanomes et les carcinomes thyroïdiens métastasent aussi dans les poumons, bien que la fréquence de ces tumeurs primitives soit relativement faible [50, 51]. De manière intéressante notre test est très précis pour discriminer les cancers du poumon de la plupart des entités citées dans ce paragraphe, tous comme il est très puissant pour les discriminer de la vaste majorité des groupes d'adénocarcinomes et de bien d'autres groupes de cancers analysés dans notre étude.

- **Immunohistochimie versus Biologie moléculaire ?**

L'immunohistochimie, si elle est utilisée de manière pertinente, apporte une orientation diagnostique dans la vaste majorité des tumeurs malignes indifférenciées, bien que parfois cela soit au terme d'un algorithme décisionnel morpho-immunohistochimique fastidieux et onéreux, cette approche reste le gold standard en première intention, notamment pour des raisons de délais : 1 jour de délai technique en routine pour une demande d'examen immunohistochimique, contre plusieurs jours voire semaines pour les examens de biologie moléculaire. L'errance diagnostique engendrée par une démarche diagnostique inadaptée ou une tumeur de diagnostic difficile peut prolonger ce délai de réponse par la multiplication des séries de demandes immunohistochimiques, ce qui est souvent le cas dans les CAPI. De façon générale, le diagnostic de CAPI est porté sur des prélèvements de petite taille pour lesquels il est indispensable de s'assurer de la disponibilité suffisante de matériel pour l'ensemble des examens immunohistochimiques complémentaires et de pouvoir correctement réaliser une analyse moléculaire. De plus, la multitude de marqueurs immunohistochimiques nécessaires pour orienter le diagnostic oblige à s'interroger sur leur disponibilité et l'impact budgétaire dans chaque laboratoire d'anatomie pathologique et potentiellement sur la mise en place de « centres experts » pour pouvoir réaliser ces examens. Dans notre étude, la sensibilité du test EASILUNG était étroitement liée au tabagisme dans le groupe des adénocarcinomes pulmonaires du TCGA, renforçant la causalité entre notre signature substitutionnelle pulmonaire et le tabagisme, et son intérêt potentiel dans le diagnostic des adénocarcinomes pulmonaires chez les fumeurs dont la tumeur n'exprime pas TTF1. Il permettrait, en pratique, de rattraper le diagnostic dans ce sous-groupe de patients chez qui une origine pulmonaire n'est pas écartée. Notre test diagnostique basé sur les résultats de séquençage exomique n'est donc pas à opposer à la démarche classique utilisant l'IHC, au contraire, elle s'intègre dans une démarche diagnostique globale et

intégrative, dans laquelle le pathologiste joue un rôle central dans la sélection des cas pour lesquels des analyses moléculaires complémentaires seraient pertinentes.

- **Tests moléculaires et CAPI.**

Des tests moléculaires basés sur l'expression de gènes sont en cours d'expérimentation et pourraient apporter une réponse à certains cas non résolus par l'étude immunohistochimique. Ils sont basés sur la comparaison du profil d'expression génique de la tumeur qui est comparé à des résultats de série de tumeurs caractérisées par cette même approche. Ces tests utilisent l'étude de l'expression des ARNm par RT-PCR. L'ARN tumoral peut être extrait à partir de blocs inclus en paraffine. Certains tests sont actuellement commercialisés. Dans ces études moléculaires, les chercheurs se sont appuyés sur des profils d'expression multigéniques de groupes de tumeurs solides de primitifs connus. Grâce à ces groupes de tumeurs dont l'origine primitive est connue, chaque type de tumeur s'est vu attribuer une signature multigénique spécifique « unique ». La spécificité de chaque signature moléculaire a été confirmée par son application dans un second ensemble indépendant de tumeurs solides de primaire connu (validation externe), afin de tester le succès de la prédiction du tissu d'origine de chaque tumeur. L'hypothèse sur laquelle les investigateurs se sont appuyés était que chaque tumeur tout au long de son évolution et de sa dissémination métastatique conservait sa signature moléculaire distincte, hypothèse sur laquelle nous nous sommes également appuyés dans notre étude. L'étape finale impliquait la comparaison du modèle d'expression multigénique d'un cas de CAPI à chacun des motifs spécifiques du type de tumeur et ainsi l'attribution du primitif auquel la signature moléculaire du CAPI ressemblait le plus. Dans ces modèles, plusieurs milliers de gènes ont été analysés pour établir l'expression différentielle dans divers types de tumeurs solides. Des approches de modélisations mathématiques complexes ont été utilisées afin d'identifier un ensemble de gènes pouvant discriminer les tumeurs originaires de différents tissus. Les approches mathématiques utilisées n'étaient pas standardisées dans les études publiées et constituent donc une source significative d'hétérogénéité, en plus des plateformes techniques et des tumeurs utilisées. Nous avons, dans notre étude, choisi de travailler sur l'ADN en utilisant uniquement les données du TCGA afin de conserver un certain degré d'homogénéité. Une des hypothèses expliquant les limites avec les méthodes diagnostiques utilisant l'étude de l'expression des ARNm par RT-PCR est qu'une signature transcriptionnelle prédisposant aux métastases coexiste dans certaines tumeurs, ce programme génétique « métastatique » qui transcende les distinctions entre les tissus d'origine a été décrit et pourrait mener à

une erreur de classification moléculaire (tumeurs classées en primitif du sein ou de la vessie de façon erronée [52]). Cette approche évaluant l'expression génique présente plusieurs autres inconvénients. On ne sait pas quels gènes sont pertinents pour évaluer le comportement biologique d'une tumeur donnée et la sélection des gènes repose fortement sur le nombre et le type de tumeurs solides primitives étudiées [53]. Dans les séries publiées, les échantillons d'entraînement (« training set ») incluaient moins de 1000 échantillons tumoraux (typiquement 100 à 400). Par ailleurs, on sait que les adénocarcinomes pulmonaires sont hétérogènes sur la base des études de profil d'expression génique [54]. Plus récemment, une stratégie alternative est de proposer un traitement ciblant le primitif supposé selon la signature moléculaire (expression génique en RT-PCR sur un panel d'ARNm de 92 gènes) [55]. Sur ce panel de référence, la sensibilité était de 63% sur les adénocarcinomes pulmonaires issus de tissu congelé et tombait à 0% sur tissus fixés au formol et inclus en paraffine (FFPE), ce qui peut s'expliquer par une dégradation importante des ARNm dans les tissus FFPE. Ces facteurs expliquent pourquoi de nombreux gènes qui orientent vers un tissu d'origine ne sont pas les mêmes selon les différentes études. D'autre part, ces tests sont relativement complexes et on ne sait pas encore si l'analyse devrait être limitée aux cellules néoplasiques (microdissection par capture laser) ou inclure un stroma non néoplasique [52]. Des données montrent que les cellules stromales portent une expression génique caractéristique différente selon les tumeurs et contribuant à leur survie des tumeurs malignes [56]. En étudiant l'ADN tumoral (et non les ARNm), notre approche n'est pas impactée par ces contraintes, expliquant nos bons résultats et suggérant la supériorité de notre approche dans les cancers du poumon, en particulier concernant les adénocarcinomes. En ayant analysé les adénocarcinomes et les carcinomes épidermoïdes pulmonaires, nous couvrons les deux histologies les plus répandues du cancer du poumon. Nous avons montré que le test EASILUNG était suffisant seul pour distinguer les cancers du poumon de la plupart des autres groupes de cancers solides. Par ailleurs, nos résultats suggèrent que le test EASILUNG a conservé de bonnes performances sur les échantillons FFPE. L'artéfact mutationnel potentiellement généré par la fixation du formol [57] n'était pas une limitation pour notre test, comme le suggèrent nos résultats de validation externe sur série locale. Le problème majeur connu des études de validation externes est qu'elles sont souvent basées sur de petits ensembles de données locales. Cela implique qu'il y a souvent une hétérogénéité dans la performance du modèle, et que de multiples études de validation externes sont nécessaires pour apprécier pleinement les possibilités concernant la généralisation d'un modèle prédictif. C'est aussi une des forces de notre étude, que d'avoir analysé des résultats de tumeurs issues de blocs FFPE, bien que nous ne connaissions pas le type d'échantillon initialement (congelé ou FFPE)

puisque nous étions en aveugle de ce paramètre. Les tumeurs présentes dans les bases de données (TCGA ou autres) sont, à notre connaissance et en 2018, quasi exclusivement constituée d'échantillons issu de tissu congelé, puisque c'est le gold standard en termes de conservations, que ces banques de tumeurs se sont constituées à visé de recherche, et que le coût des analyses réalisés sur ces échantillons (séquençage génomique, séquençage exomique, étude du protéome, de l'ARNome, du méthylome) justifie une qualité d'échantillon optimale. Or, les tissus FFPE présentent une qualité altérée des constituants cellulaires, notamment les ARNs, et représentent donc un facteur limitant de la qualité d'une analyse globale d'un groupe de tumeurs. Bien que les tissus FFPE soient exclus de ces bases de données, les blocs FFPE représentent en pratique quotidienne le mode de conservation des tissus tumoraux le plus répandu, puisque l'examen histologique et immunohistochimiques nécessaires au diagnostic initial d'un cancer sont réalisés dessus. Même si de plus en plus de laboratoires d'anatomie pathologique et de biopathologie se dotent de banques de tumeurs et réalisent une congélation d'échantillons tumoraux de plus en plus systématiquement, l'analyse moléculaire sur bloc FFPE devrait rester, pour quelques années encore, un incontournable de la routine diagnostique, puisque beaucoup de laboratoire sont dépourvus de tumorothèques et que tous les prélèvements n'ont pas d'indication à être congelés (très petites tumeurs, tumeurs non vues, échantillons parvenus fixés au laboratoire,…).

- **Biais de brin**

On parle de biais de brin (« strand bias » en anglais) lorsque l'efficacité de la réparation des erreurs et du maintien de l'intégrité de l'ADN au cours de la réplication diffère entre le brin transcrit et le brin non-transcrit. Notre approche analysant les 12 classes de substitutions somatiques lues sur le brin codant non-transcrit permet de considérer le biais de brin présent dans la mutagénèse induite par le tabac, et déjà décrit par *Alexandrov et al* [58]. C'est le modèle utilisant le minimum de classes de substitutions (12 classes) permettant d'analyser ce biais de brin. Les seuils de notre score EASILUNG indique que les cancers du poumon ont un biais de brins concernant le couple G>T/C>A avec une prédominance de transversions G>T comparée aux transversions complémentaires C>A, témoignant indirectement de la formation d'adduits volumineux de guanine qui ont été réparés au cours de la réplication de l'ADN, de manière plus efficace sur le brin transcrit. Une des causes bien étudiée est la réparation par excision de nucléotides couplée à la transcription (« NER » pour *nucléotide excision repair* en anglais) qui opère principalement sur le brin transcrit et est recrutée par l'ARN polymérase II lorsqu'elle rencontre des lésions

volumineuses distordant l'hélice de l'ADN, comme celles engendrées par le BPDE [59]. Le paramètre « fréquence relative élevée de C>A (%CA >12%) » est le paramètre le plus discriminant, avec l'*odd ratio* le plus élevé dans notre étude, suivie du paramètre « fréquence relative élevée de G>T (%GT >13%) ». La prévalence plus élevée de mutations G>T sur le brin non-transcrit comparé aux mutations C> A dans le cancer du poumon reflète ce biais de brin (G>T est mieux réparé sur le brin transcrit complémentaire) et confirme que des lésions volumineuses de la guanine peuvent être à l'origine du pattern de substitutions observée. En revanche, ce biais de brin n'a pas été observé pour les couples de substitutions somatiques C>G/G>C, C>T/G>A, T>A/A>T ou T>C/A>G dans notre étude, comme le suggèrent les valeurs très proches (ou identiques) des seuils et des *odds ratio* (pour les paramètres non seuillées) de ces couples de paramètres, confirmant la spécificité de ce biais de brin pour le couple G>T/C>A.

- **ADN tumoral circulant (ctDNA)**

De façon intéressante, notre test pourrait permettre la détection du cancer du poumon sur ctDNA. L'ADN tumoral circulant est un biomarqueur prometteur pour l'évaluation non invasive du cancer. Dans une étude précédente, de l'ADN tumoral circulant a été détecté chez 100% des patients atteints de cancer bronchique non à petites cellules de stade II à IV et chez 50% des patients atteints d'un stade I [60]. L'analyse exomique de l'ADN tumoral circulant pourrait compléter les approches actuelles de biopsies invasives pour diagnostiquer le cancer du poumon ou détecter le cancer du poumon dans une future stratégie de dépistage utilisant le ctDNA. Actuellement, il est possible de rechercher dans le sang des mutations (*KRAS, TP53, EGFR*, ...) que l'on retrouve dans de nombreux cancers mais avec une spécificité d'organe faible (poumon, pancréas, côlon...). Notre approche innovante basée sur la proportion relative des substitutions pourrait enrichir en informations les résultats d'analyse de l'ADN tumoral circulant en apportant une dimension de spécificité d'organe inédite.

❖ Conclusion

Le carcinome de primitif inconnu illustre qu'en médecine, lorsque le diagnostic fait défaut, le traitement ne peut être optimisé. Nous proposons une nouvelle méthode de discrimination des cancers basée sur la signature substitutionnelle pulmonaire des carcinogènes spécifique des cancers du poumon. Elle représente une alternative puissante pour identifier avec précision l'origine pulmonaire d'une tumeur en aveugle de tous les autres paramètres. Elle peut donc avoir une utilité clinique potentielle pour les patients ayant un CAPI, en particulier lorsque le

diagnostic de cancer du poumon est toujours en considération. La partie validation externe du test sur cohorte locale, utilisant les résultats d'exomes d'échantillons de routine (congelé et FFPE), a confirmé les bons résultats du test EASILUNG, même sur des groupes de cancers non présents dans les échantillons du TCGA, le rendant potentiellement intéressant pour une application clinique. Des études de validation externes et de validations prospectives multiples sur résultats d'exomes sont nécessaires pour déterminer si ce modèle de prédiction peut être généralisé et des essais cliniques sont également nécessaires pour évaluer les bénéfices pour les patients. Francis Crick écrivait : « L'ADN est une molécule si importante qu'il est presque impossible d'en apprendre trop sur elle » [61]. Aujourd'hui, plus de 60 ans après l'établissement de la structure chimique primaire de l'ADN [62], il reste nécessaire de développer notre compréhension de l'ADN et ses potentialités. Notre approche utilise, en ce sens, les avancées permises par les progrès de la biologie moléculaire pour les mettre au service du diagnostic, et donc des patients.

❖ BIBLIOGRAPHIE (Hors article principal)

1. Tomasetti, C., Li, L. & Vogelstein, B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science* **355**, 1330–1334 (2017).
2. Weinberg, R. A. *The biology of cancer*. (Garland Science, 2014).
3. Vogelstein, B. *et al.* Cancer Genome Landscapes. *Science* **339**, 1546–1558 (2013).
4. Transition et transversion dans le génome humain. Available at: http://dridk.me/transition_transversion.html.
5. Sher, T., Dy, G. K. & Adjei, A. A. Small Cell Lung Cancer. *Mayo Clin. Proc.* **83**, 355–367 (2008).
6. Hecht, S. S. DNA adduct formation from tobacco-specific N-nitrosamines. *Mutat. Res. Mol. Mech. Mutagen.* **424**, 127–142 (1999).
7. Hecht, S. S. Cigarette smoking and lung cancer: chemical mechanisms and approaches to prevention. *Lancet Oncol.* **3**, 461–469 (2002).
8. Akopyan, G. & Bonavida, B. Understanding tobacco smoke carcinogen NNK and lung tumorigenesis. *Int. J. Oncol.* **29**, 745–752 (2006).
9. Phillips, D. H. Fifty years of benzo(a)pyrene. *Nature* **303**, 468–472 (1983).

10. You, M., Candrian, U., Maronpot, R. R., Stoner, G. D. & Anderson, M. W. Activation of the Ki-ras protooncogene in spontaneously occurring and chemically induced lung tumors of the strain A mouse. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 3070–3074 (1989).
11. Belinsky, S. A., Devereux, T. R., Foley, J. F., Maronpot, R. R. & Anderson, M. W. Role of the alveolar type II cell in the development and progression of pulmonary tumors induced by 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone in the A/J mouse. *Cancer Res.* **52**, 3164–3173 (1992).
12. Ruggeri, B. *et al.* Benzo[a]pyrene-induced murine skin tumors exhibit frequent and characteristic G to T mutations in the p53 gene. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 1013–1017 (1993).
13. Greenblatt, M. S., Bennett, W. P., Hollstein, M. & Harris, C. C. Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis. *Cancer Res.* **54**, 4855–4878 (1994).
14. Leanderson, P. & Tagesson, C. Cigarette smoke-induced DNA-damage: role of hydroquinone and catechol in the formation of the oxidative DNA-adduct, 8-hydroxydeoxyguanosine. *Chem. Biol. Interact.* **75**, 71–81 (1990).
15. Stone, K. K., Bermúdez, E. & Pryor, W. A. Aqueous extracts of cigarette tar containing the tar free radical cause DNA nicks in mammalian cells. *Environ. Health Perspect.* **102 Suppl 10**, 173–178 (1994).
16. Miller, E. C. Some current perspectives on chemical carcinogenesis in humans and experimental animals: Presidential Address. *Cancer Res.* **38**, 1479–1496 (1978).
17. Johnson, R. E., Prakash, S. & Prakash, L. Efficient bypass of a thymine-thymine dimer by yeast DNA polymerase, Poleta. *Science* **283**, 1001–1004 (1999).
18. Masutani, C. *et al.* Xeroderma pigmentosum variant (XP-V) correcting protein from HeLa cells has a thymine dimer bypass DNA polymerase activity. *EMBO J.* **18**, 3491–3501 (1999).
19. Zhang, Y. *et al.* Error-prone lesion bypass by human DNA polymerase eta. *Nucleic Acids Res.* **28**, 4717–4724 (2000).
20. Zhang, Y. *et al.* Two-step error-prone bypass of the (+)- and (-)-trans-anti-BPDE-N2-dG adducts by human DNA polymerases eta and kappa. *Mutat. Res.* **510**, 23–35 (2002).
21. Chiapperino, D. *et al.* Preferential misincorporation of purine nucleotides by human DNA polymerase eta opposite benzo[a]pyrene 7,8-diol 9,10-epoxide deoxyguanosine adducts. *J. Biol. Chem.* **277**, 11765–11771 (2002).

22. Zhao, B., Wang, J., Geacintov, N. E. & Wang, Z. Poleta, Polzeta and Rev1 together are required for G to T transversion mutations induced by the (+)- and (-)-trans-anti-BPDE-N2-dG DNA adducts in yeast cells. *Nucleic Acids Res.* **34**, 417–425 (2006).
23. Wood, A. J. J., Hainsworth, J. D. & Greco, F. A. Treatment of Patients with Cancer of an Unknown Primary Site. *N. Engl. J. Med.* **329**, 257–263 (1993).
24. Pavlidis, N., Briasoulis, E., Pentheroudakis, G. & ESMO Guidelines Working Group. Cancers of unknown primary site: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **21 Suppl 5**, v228-231 (2010).
25. Levi, F., Te, V. C., Erler, G., Randimbison, L. & La Vecchia, C. Epidemiology of unknown primary tumours. *Eur. J. Cancer Oxf. Engl. 1990* **38**, 1810–1812 (2002).
26. Didolkar, M. S., Fanous, N., Elias, E. G. & Moore, R. H. Metastatic carcinomas from occult primary tumors. A study of 254 patients. *Ann. Surg.* **186**, 625–630 (1977).
27. Massard, C., Loriot, Y. & Fizazi, K. Carcinomas of an unknown primary origin--diagnosis and treatment. *Nat. Rev. Clin. Oncol.* **8**, 701–710 (2011).
28. Network, T. C. G. A. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
29. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
30. The Cancer Genome Atlas Research Network. Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. *N. Engl. J. Med.* **374**, 135–145 (2016).
31. Collisson, E. A. *et al.* Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
32. Ross, J. S. *et al.* Comprehensive Genomic Profiling of Carcinoma of Unknown Primary Site: New Routes to Targeted Therapies. *JAMA Oncol.* **1**, 40–49 (2015).
33. Pavlidis, N. & Pentheroudakis, G. Cancer of unknown primary site. *Lancet Lond. Engl.* **379**, 1428–1435 (2012).
34. Bahrami, A., Truong, L. D. & Ro, J. Y. Undifferentiated tumor: true identity by immunohistochemistry. *Arch. Pathol. Lab. Med.* **132**, 326–348 (2008).
35. Ikeda, S. *et al.* Combined immunohistochemistry of β-catenin, cytokeratin 7, and cytokeratin 20 is useful in discriminating primary lung adenocarcinomas from metastatic colorectal cancer. *BMC Cancer* **6**, (2006).

36. Cabibi, D., Licata, A., Barresi, E., Craxì, A. & Aragona, F. Expression of Cytokeratin 7 and 20 in Pathological Conditions of the Bile Tract. *Pathol. - Res. Pract.* **199**, 65–70 (2003).
37. Tot, T. Patterns of distribution of cytokeratins 20 and 7 in special types of invasive breast carcinoma: a study of 123 cases. *Ann. Diagn. Pathol.* **3**, 350–356 (1999).
38. Castrillon, D. H., Lee, K. R. & Nucci, M. R. Distinction between endometrial and endocervical adenocarcinoma: an immunohistochemical study. *Int. J. Gynecol. Pathol. Off. J. Int. Soc. Gynecol. Pathol.* **21**, 4–10 (2002).
39. Taniere, P. *et al.* Cytokeratin expression in adenocarcinomas of the esophagogastric junction: a comparative study of adenocarcinomas of the distal esophagus and of the proximal stomach. *Am. J. Surg. Pathol.* **26**, 1213–1221 (2002).
40. Lee, J.-H., Lee, J. H., Kim, A., Kim, I. & Chae, Y.-S. Unique expression of MUC3, MUC5AC and cytokeratins in salivary gland carcinomas. *Pathol. Int.* **55**, 386–390 (2005).
41. Bejarano, P. A., Nikiforov, Y. E., Swenson, E. S. & Biddinger, P. W. Thyroid transcription factor-1, thyroglobulin, cytokeratin 7, and cytokeratin 20 in thyroid neoplasms. *Appl. Immunohistochem. Mol. Morphol. AIMM* **8**, 189–194 (2000).
42. Tot, T. The value of cytokeratins 20 and 7 in discriminating metastatic adenocarcinomas from pleural mesotheliomas. *Cancer* **92**, 2727–2732 (2001).
43. Zhang, Y. *et al.* Negative Thyroid Transcription Factor 1 Expression Defines an Unfavorable Subgroup of Lung Adenocarcinomas. *J. Thorac. Oncol. Off. Publ. Int. Assoc. Study Lung Cancer* **10**, 1444–1450 (2015).
44. Lin, F. & Liu, H. Immunohistochemistry in undifferentiated neoplasm/tumor of uncertain origin. *Arch. Pathol. Lab. Med.* **138**, 1583–1610 (2014).
45. Conner, J. R. & Hornick, J. L. Metastatic carcinoma of unknown primary: diagnostic approach using immunohistochemistry. *Adv. Anat. Pathol.* **22**, 149–167 (2015).
46. Pereira, T. C., Share, S. M., Magalhães, A. V. & Silverman, J. F. Can We Tell the Site of Origin of Metastatic Squamous Cell Carcinoma? An Immunohistochemical Tissue Microarray Study of 194 Cases: *Appl. Immunohistochem. Mol. Morphol.* **19**, 10–14 (2011).
47. Willis, R. *Secondary tumours of the lungs. In: The spread of tumors in the human body.* (Butterworths & Co, London, 1973).
48. Spencer H. *Pathology of the lung*. (Pergamon, 1977).

49. Crow, J., Slavin, G. & Kreel, L. Pulmonary metastasis: a pathologic and radiologic study. *Cancer* **47**, 2595–2602 (1981).

50. Coppage, L., Shaw, C. & Curtis, A. M. Metastatic disease to the chest in patients with extrathoracic malignancy. *J. Thorac. Imaging* **2**, 24–37 (1987).

51. Libshitz, H. I. & North, L. B. Pulmonary metastases. *Radiol. Clin. North Am.* **20**, 437–451 (1982).

52. Ramaswamy, S., Ross, K. N., Lander, E. S. & Golub, T. R. A molecular signature of metastasis in primary solid tumors. *Nat. Genet.* **33**, 49–54 (2003).

53. Simon, R., Radmacher, M. D., Dobbin, K. & McShane, L. M. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J. Natl. Cancer Inst.* **95**, 14–18 (2003).

54. Garber, M. E. *et al.* Diversity of gene expression in adenocarcinoma of the lung. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 13784–13789 (2001).

55. Hainsworth, J. D. *et al.* Molecular gene expression profiling to predict the tissue of origin and direct site-specific therapy in patients with carcinoma of unknown primary site: a prospective trial of the Sarah Cannon research institute. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **31**, 217–223 (2013).

56. Zalatnai, A. Molecular aspects of stromal-parenchymal interactions in malignant neoplasms. *Curr. Mol. Med.* **6**, 685–693 (2006).

57. Williams, C. *et al.* A High Frequency of Sequence Alterations Is Due to Formalin Fixation of Archival Specimens. *Am. J. Pathol.* **155**, 1467–1471 (1999).

58. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).

59. Pfeifer, G. P. *et al.* Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene* **21**, 7435–7451 (2002).

60. Newman, A. M. *et al.* An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. *Nat. Med.* **20**, 548–554 (2014).

61. Crick, F. H., Wang, J. C. & Bauer, W. R. Is DNA really a double helix? *J. Mol. Biol.* **129**, 449–457 (1979).

62. Todd, A. R. CHEMICAL STRUCTURE OF THE NUCLEIC ACIDS. *Proc. Natl. Acad. Sci. U. S. A.* **40**, 748–755 (1954).

RÉSUMÉ DE LA THÈSE

Introduction : Les produits de combustions du tabac laissent une forte empreinte mutationnelle sur les cancers pulmonaires, ils engendrent un grand nombre de substitutions somatiques avec un biais dans le type de substitutions. Le diagnostic de l'origine primitive pulmonaire d'un cancer peut parfois être difficile alors qu'il revêt un intérêt majeur car des thérapies efficaces sont disponibles même au stade métastatique. L'objectif de notre étude était de développer un test moléculaire innovant capable de prédire l'origine pulmonaire d'une tumeur en utilisant uniquement les fréquences relatives des différents types de substitutions somatiques.

Matériel et Méthodes : Pour prédire si une tumeur est d'origine pulmonaire ou non, nous avons développé et validé le test EASILUNG (Exome And SIgnature LUNG) basé sur la fréquence relative des 12 types de substitutions somatiques possibles sur le brin d'ADN codant de tumeurs ayant bénéficié d'un séquençage exomique, en aveugle de tous les autres paramètres. Les données de 7796 échantillons de tumeurs congelés de primitif connu (sans traitement néoadjuvant) issues des 32 groupes de cancers solides du TCGA ont été utilisées pour son développement. La validation externe a été réalisée sur 196 résultats consécutifs de séquençage exomique de routine (cohorte local).

Résultats : Le test EASILUNG a démontré une bonne calibration et un bon pouvoir discriminant avec une sensibilité de 80% et une spécificité de 94,3%. Des résultats similaires ont été obtenus sur les données de validation externe.

Conclusion : C'est un test innovant et performant mais surtout un angle d'approche diagnostique inédit que nous proposons, ouvrant sur de nombreuses perspectives de recherches, mais aussi sur une application diagnostique concrète pour les patients atteints de cancer de primitif inconnu et chez qui le diagnostic de cancer pulmonaire et toujours en considération. Des études de validations externes et de validations prospectives multiples sur résultats d'exomes sont souhaitables pour déterminer si ce modèle de prédiction peut être généralisé et des essais cliniques sont nécessaires pour évaluer les bénéfices pour les patients.

TITRE EN ANGLAIS: Relative frequency of somatic substitution classes from a whole exome sequencing result as an innovative tool in lung cancer diagnosis.

THÈSE D'ANATOMIE ET CYTOLOGIE PATHOLOGIQUES – ANNÉE 2018.

MOTS CLES : Cancer du poumon – diagnostic – substitutions somatiques – score - signature

INTITULÉ ET ADRESSE :

UNIVERSITÉ DE LORRAINE

Faculté de Médecine de Nancy

9, avenue de la Forêt de Haye

54505 VANDOEUVRE LES NANCY Cedex